Human Genetics

Contents

1 Introduction **14**
- 1.1 About This Book . . . 14
- 1.2 Table of Contents . . . 15
- 1.3 Part 1. The Human Genome Project and NGS . . . 15
- 1.4 Part 2. Mendelian Genetics in NGS . . . 15
- 1.5 Part 3. Polygenic and Genetic Architecture . . . 15
- 1.6 Part 4. Recombination, Linkage and Population . . . 15
- 1.7 Part 5. Variant to System . . . 16

2 Getting Started **16**
- 2.1 Why I wrote this book for undergraduate students . . . 16
- 2.2 From consuming knowledge to creating it . . . 18
- 2.3 Why this course feels different . . . 18
- 2.4 Learning the Language of Science . . . 19
  - 2.4.1 Types of Scientific Papers . . . 19
  - 2.4.2 Practical Tips for Reading Scientific Papers . . . 19
- 2.5 Using AI to Read Papers More Effectively . . . 20
  - 2.5.1 Using AI with This Textbook . . . 20
  - 2.5.2 Using AI to Read Scientific Papers . . . 21
  - 2.5.3 Recommended AI Tools for This Course . . . 21
  - 2.5.4 Writing Good Prompts: A Conversation Example . . . 22
  - 2.5.5 Understanding AI's Limitations . . . 22
  - 2.5.6 The Real Goal: Learning to Think, Not Just to Know . . . 22
- 2.6 Welcome to Human Genetics . . . 23

3 Part 1. The Human Genome Project and Next-Generation Sequencing: A Revolution in Modern Biology **23**
- 3.1 From Invisible Factors to Readable Genomes . . . 23
- 3.2 What This Chapter Is About . . . 23
  - 3.2.1 The Foundation: Reference Genomes . . . 23
  - 3.2.2 The Technology: Sequencing Platforms . . . 24
  - 3.2.3 The Strategy: What to Sequence . . . 24
  - 3.2.4 The Interpretation: Making Sense of Data . . . 24
- 3.3 Connecting to Mendel: From Ratios to Sequences . . . 24
- 3.4 The Democratization of Genomics . . . 25
- 3.5 How to Approach This Chapter . . . 25
  - 3.5.1 Key Themes to Watch For . . . 25
  - 3.5.2 Focus on Understanding . . . 26
- 3.6 The Paradigm Shift . . . 26
  - 3.6.1 Before HGP and NGS . . . 26
  - 3.6.2 After HGP and NGS . . . 26
- 3.7 Learning Objectives . . . 26

3.8 Let's Begin . . . 27

4 Chapter 1. The Human Genome Project 27
4.1 What Did the HGP Set Out to Do? . . . 27
4.1.1 1. Sequence the Entire Genome . . . 27
4.1.2 2. Identify All Human Genes . . . 27
4.1.3 3. Develop New Technologies . . . 28
4.1.4 4. Map Genetic Variation . . . 28
4.2 How Did They Do It? The HGP Journey . . . 28
4.2.1 Key Milestones . . . 28
4.3 The Big Surprise: Only 20,000 Genes? . . . 29
4.4 How the HGP Changed Genetics . . . 29
4.4.1 From Gene-by-Gene to Genome-Wide Analysis . . . 29
4.4.2 A Common Reference for Comparing Genomes . . . 30
4.4.3 Technological Revolution . . . 30
4.4.4 Open Data Sharing: The Bermuda Principles . . . 30
4.5 What the HGP Couldn't Do: The 8% That Remained . . . 30
4.6 The Foundation for Everything That Followed . . . 31
4.7 Timeline at a Glance . . . 31

5 Chapter 2. The Telomere-to-Telomere Genome Project 31
5.1 What Made T2T-CHM13 Possible? . . . 32
5.1.1 The Technologies . . . 32
5.1.2 String Graphs: A Tool for Assembly . . . 32
5.2 The Special Cell Line: CHM13 . . . 33
5.3 The First Complete Human Chromosome: X (2020) . . . 34
5.3.1 Why Start with X? . . . 34
5.3.2 Finishing the Breaks . . . 34
5.3.3 The Innovation: Marker-Assisted Polishing . . . 34
5.3.4 Validating the X Centromere . . . 37
5.3.5 Achievements of the Complete X . . . 37
5.3.6 Methylation Discovery . . . 37
5.3.7 Impact . . . 39
5.4 What's New in T2T-CHM13? (2022) . . . 39
5.4.1 Region 1: Centromeric Satellite Arrays . . . 41
5.4.2 Region 2: Segmental Duplications . . . 41
5.4.3 Region 3: Short Arms of Acrocentric Chromosomes . . . 41
5.5 Completing the Y Chromosome (2023) . . . 42
5.5.1 What Makes the Y Chromosome Special? . . . 42
5.5.2 The Y Centromere . . . 44
5.5.3 The Heterochromatic Region (Yq12) . . . 44
5.6 Quick Reference: Complex Genomic Regions . . . 46
5.7 Summary: Why These Regions Matter . . . 47
5.8 Looking Forward . . . 47

6 Chapter 3. The CHM13 Cell Line 47
6.1 What Is a Complete Hydatidiform Mole? . . . 48
6.1.1 Normal Fertilization vs CHM Formation . . . 48
6.1.2 Normal Fertilization . . . 49
6.1.3 What Happens in a Complete Hydatidiform Mole . . . 49
6.1.4 Understanding "Functionally Haploid" . . . 49
6.1.5 Nearly Uniform Homozygosity . . . 50
6.1.6 Why This Is Perfect for Sequencing . . . 50
6.2 The hTERT Modification: Making Cells Immortal . . . 51

6.2.1 The Telomere Problem . . . 51
6.2.2 Enter Telomerase . . . 52
6.2.3 The hTERT Gene . . . 52
6.3 Genetic Stability and Quality Control . . . 53
6.3.1 Karyotype Analysis . . . 53
6.3.2 Sequence-Level Validation . . . 53
6.3.3 Why These Checks Matter . . . 53
6.4 Ancestry and Genetic Background . . . 53
6.4.1 Does Ancestry Matter? . . . 55
6.5 Why CHM13 Specifically? . . . 55
6.5.1 1. Availability and Characterization . . . 55
6.5.2 2. Karyotype Stability . . . 55
6.5.3 3. Favorable Karyotype (46,XX) . . . 56
6.5.4 4. Successful hTERT Immortalization . . . 56
6.5.5 5. High-Quality DNA . . . 56
6.5.6 6. Community Support . . . 56
6.6 Why CHM13 Made T2T Possible . . . 56
6.6.1 1. Functionally Haploid Genome . . . 56
6.6.2 2. Immortalized Cells . . . 56
6.6.3 3. Genetic Stability . . . 57
6.7 Comparing Haploid and Diploid Approaches . . . 57
6.8 The Limitations of CHM13 . . . 58
6.8.1 1. Represents Only One Genome Source . . . 58
6.8.2 2. Ancestry Bias . . . 58
6.8.3 3. No Heterozygosity Information . . . 58
6.8.4 4. Chimeric Reference (X and Y from Different Sources) . . . 59
6.8.5 5. Cannot Study Allelic Imbalance . . . 59
6.9 Moving Beyond CHM13 . . . 59
6.10 Summary . . . 59
6.11 Key Takeaways . . . 60

7 Chapter 4. From Single Reference to Pangenome 60
7.1 What Is a Pangenome? . . . 60
7.1.1 The Human Pangenome Reference Consortium (HPRC) . . . 61
7.2 What the Pangenome Revealed . . . 61
7.2.1 1. We Were Missing A Lot of DNA . . . 61
7.2.2 2. Genes Come in Different Copy Numbers . . . 61
7.2.3 3. Better Detection of Structural Variants . . . 63
7.3 How Was the Pangenome Built? . . . 64
7.3.1 Key Technologies . . . 64
7.3.2 The Challenge: Phasing . . . 64
7.4 Real-World Impact: Why This Matters . . . 64
7.4.1 1. Clinical Diagnostics . . . 64
7.4.2 2. Drug Development . . . 64
7.4.3 3. Disease Research . . . 64
7.5 Current Limitations and Next Steps . . . 65
7.5.1 What's Still Challenging . . . 65
7.5.2 Future Expansion . . . 65
7.6 Comparing Genome References: The Evolution . . . 65
7.7 The Path Forward: From Reference to Resources . . . 66
7.8 Why This Matters for You . . . 66
7.9 Summary: One Map to Many Maps . . . 66

8 Chapter 5. The Emergence of Next-Generation Sequencing Technologies 67

8.1 The Bottleneck of Sanger Sequencing . . . 67
8.2 What Makes NGS "Next-Generation"? . . . 67
8.3 Understanding the Basics: Key Concepts in NGS . . . 69
8.3.1 Sequencing Reactions and Reads . . . 69
8.3.2 Coverage and Depth . . . 69
8.3.3 Base Quality Scores . . . 69
8.3.4 Libraries and Adaptors . . . 69
8.3.5 Alignment and Variant Calling . . . 70
8.4 The NGS Technology Landscape . . . 71
8.4.1 Illumina: The Workhorse of Genomics (Second-Generation) . . . 71
8.4.2 PacBio Revio: Long Reads with High Accuracy (Third-Generation) . . . 71
8.4.3 Oxford Nanopore: Ultra-Long Reads and Portable Sequencing (Third-Generation) . . . 73
8.5 Comparing the Technologies . . . 75
8.6 How NGS Changed Genetics . . . 75
8.6.1 From Gene-by-Gene to Genome-Wide . . . 75
8.6.2 The Variant Catalog . . . 75
8.6.3 Clinical Diagnostics . . . 76
8.6.4 Completing the Genome . . . 76
8.7 The Economics of Sequencing . . . 76
8.8 Practical Considerations . . . 76
8.9 Looking Forward . . . 77

9 Chapter 6. NGS Application for Genomics 77
9.1 The Fundamental Question: Why Not Always Sequence Everything? . . . 77
9.2 Understanding the Exome: Where Most Disease Variants Hide . . . 78
9.3 Head-to-Head Comparison: WES vs WGS . . . 78
9.4 Whole-Exome Sequencing (WES): The Focused Approach . . . 79
9.4.1 How It Works . . . 79
9.4.2 The Historic Success: Miller Syndrome (2010) . . . 79
9.4.3 Clinical Impact of WES . . . 80
9.5 Whole-Genome Sequencing (WGS): The Comprehensive Approach . . . 80
9.5.1 How It Works . . . 80
9.5.2 The Power of Seeing Everything: Real Examples . . . 81
9.5.3 Non-Coding Variants: The Frontier . . . 81
9.6 The Technical Process: From Sample to Variants . . . 82
9.6.1 Step 1: Library Preparation . . . 82
9.6.2 Step 2: Sequencing . . . 84
9.6.3 Step 3: Variant Calling – From Reads to Biological Meaning . . . 85
9.6.4 Putting It All Together: A Complete Workflow . . . 86
9.7 The Clinical Decision Tree: Which Test to Order? . . . 86
9.7.1 Start with WES if: . . . 86
9.7.2 Move to WGS if: . . . 86
9.7.3 A Real Decision-Making Case . . . 86
9.8 The Economics and Practicalities . . . 87
9.8.1 The Cost Trajectory . . . 87
9.8.2 Storage and Computing . . . 87
9.8.3 The Analysis Bottleneck . . . 87
9.9 The Future: Is WES Becoming Obsolete? . . . 87
9.9.1 The "WGS Now, Interpret Exome First" Strategy . . . 87
9.9.2 What's Driving the Shift? . . . 87
9.9.3 But WES Isn't Dead Yet . . . 88
9.10 A Glimpse of the Long-Read Future . . . 88
9.11 Summary: The Bottom Line . . . 88

10 Chapter 7. Annotation and Databases for Genetic Variants **88**
10.1 The Data Deluge Problem . . . 88
10.2 What is Variant Annotation? . . . 89
10.3 Annotation Tools . . . 89
10.4 Variant Databases: The Three-Tier System . . . 89
10.4.1 Tier 1: dbSNP (Variant Catalog) . . . 89
10.4.2 Tier 2: gnomAD (Population Reference) . . . 89
10.4.3 Tier 3: ClinVar (Clinical Interpretation) . . . 90
10.5 Predicting Variant Pathogenicity . . . 90
10.6 Genomic Annotation: The Reference Framework . . . 90
10.6.1 GENCODE vs. RefSeq . . . 90
10.7 Biobanks: Connecting Genotype to Phenotype . . . 91
10.7.1 Major Biobanks Comparison . . . 91
10.8 Clinical Case: Integrating All Resources . . . 91
10.9 Why This Matters . . . 92
10.10Key Takeaways . . . 92

11 Part 2. Mendelian Genetics in Next-Generation Sequencing: From Peas to Patients **93**
11.1 Introduction: The Same Laws, New Tools . . . 93
11.2 What This Chapter Is About . . . 93
11.2.1 1. Genetic Variants . . . 93
11.2.2 2. Transmission Through Families . . . 93
11.2.3 3. Dominant Inheritance: Molecular Mechanisms . . . 93
11.2.4 4. Haploinsufficiency and Gene Constraint . . . 93
11.2.5 5. Recessive Inheritance and Consanguinity . . . 94
11.2.6 6. Structural Variations . . . 94
11.2.7 7. Integration: From Variant to Diagnosis . . . 94
11.3 From Abstract Rules to Molecular Reality . . . 94
11.4 How to Approach This Chapter . . . 94
11.4.1 Focus on Mechanisms, Not Memorization . . . 94
11.4.2 Connect Multiple Evidence Types . . . 94
11.4.3 Recognize What We Don't Know . . . 95
11.5 What Makes This Different From Classical Genetics . . . 95
11.6 Learning Objectives . . . 95
11.7 Let's Begin . . . 96

12 Chapter 8. Genetic Variants **96**
12.1 The Molecular Basis of "Different" . . . 96
12.2 Terminology: Quick Reference . . . 96
12.3 Variant Classification Framework . . . 97
12.3.1 Quick Overview Table . . . 97
12.4 Variants by Size . . . 97
12.4.1 Single Nucleotide Variants (SNVs) . . . 97
12.4.2 Insertions and Deletions (Indels) . . . 97
12.4.3 Structural Variants (SVs) . . . 98
12.5 Variants by Location and Effect . . . 99
12.5.1 The 98/2 Rule . . . 99
12.6 Coding Variants: Five Main Types . . . 99
12.6.1 Coding Variant Comparison Table . . . 99
12.6.2 Key Insights . . . 99
12.6.3 Visual: Frameshift vs In-Frame . . . 100
12.7 Non-Coding Variants: More Complex, Less Understood . . . 100
12.7.1 1. Splice Site Variants: Breaking the Cut-and-Paste . . . 100
12.7.2 2. Regulatory Variants: Affecting Gene Expression . . . 100

12.7.3 3. Deep Intronic and Intergenic Variants . . . 101
12.8 Clinical Interpretation: Which Variants Matter? . . . 101
12.8.1 The Clinical Filtering Pipeline . . . 101
12.8.2 Evidence Integration Framework . . . 101
12.9 Practical Example: Interpreting a Novel Variant . . . 101
12.10 WGS vs WES: What Can You Detect? . . . 102
12.11 Key Takeaways . . . 102
12.12 Looking Ahead . . . 103

13 Chapter 9. Transmission: How Variants Pass Through Generations 103
13.1 Two Sources of Your Variants . . . 103
13.2 Part 1: Inherited Variants . . . 103
13.2.1 Mendelian Segregation at the DNA Level . . . 103
13.2.2 Identity by Descent (IBD): Relatedness in DNA . . . 103
13.2.3 Recombination: Shuffling Between Generations . . . 104
13.2.4 Common vs. Rare Inherited Variants . . . 104
13.3 Part 2: De Novo Variants . . . 105
13.3.1 What Are They? . . . 105
13.3.2 Why They Matter . . . 105
13.4 The Paternal Age Effect . . . 105
13.4.1 The Biology . . . 105
13.5 Quantifying the Effect: Landmark Studies from Iceland . . . 105
13.5.1 Kong et al. (2012): The First Direct Measurement . . . 105
13.5.2 Jónsson et al. (2017): Expanded Analysis with 1,548 Families . . . 106
13.5.3 Porubsky et al. (2025): Long-Read Revolution . . . 107
13.6 Clinical Implications . . . 107
13.6.1 For Genetic Counseling . . . 107
13.6.2 For Clinical Diagnosis . . . 108
13.7 Identifying De Novo Variants: The Trio Approach . . . 108
13.7.1 Trio Sequencing Strategy . . . 108
13.7.2 Quality Control Steps . . . 108
13.8 Evolutionary Perspective . . . 109
13.8.1 Mutation as the Engine of Evolution . . . 109
13.8.2 Population-Level Effects . . . 109
13.9 Summary Table: Inherited vs. De Novo . . . 109
13.10 Key Takeaways . . . 109
13.11 Beyond Individual Mutations: An Evolutionary Perspective . . . 110
13.11.1 From Individuals to Species: Time Scales . . . 110
13.11.2 Three Mechanisms Drive Long-Term Evolution . . . 110
13.11.3 Case Study: FOXP2 and Speech Evolution . . . 111
13.11.4 The Connection: Micro to Macro . . . 111
13.12 Looking Ahead: From Transmission to Pattern . . . 111

14 Chapter 10. Dominant Allele in Human Disorders 111
14.1 Kabuki Syndrome: PTVs in KMT2D . . . 112
14.1.1 Finding KMT2D Mutations with WES . . . 112
14.1.2 How It Causes Disease . . . 113
14.2 Autism and Seizures: Diverse SCN2A Variants . . . 113
14.2.1 Finding SCN2A Variants with WES . . . 113
14.2.2 How They Cause Disease . . . 115
14.3 Congenital Heart Disease: Multiple Dominant Genes . . . 115
14.3.1 Finding CHD Genes with WES and Targeted Sequencing . . . 117
14.3.2 How They Cause Disease . . . 118
14.4 Neurodevelopmental Disorders: Non-Coding Variants in RNU4-2, RNU5B-1, and RNU5A-1 118

14.4.1 Finding snRNA Variants with Genome Sequencing . . . 118
14.4.2 How They Cause Disease . . . 119
14.5 Genetic Variant as Allele Leading to Phenotypic Changes . . . 119
14.5.1 Variants as Alleles in the Central Dogma . . . 119
14.5.2 Codon Changes in Coding Regions . . . 121
14.5.3 Non–Coding Variants and Splicing Disruption . . . 121
14.6 Dominant Alleles Shape Phenotypes . . . 121

15 Chapter 11. Haploinsufficiency 122
15.1 Haploinsufficient Genes vs. Haplosufficient Genes . . . 122
15.2 Natural Selection and Purifying Selection . . . 123
15.3 Measuring Haploinsufficiency with pLI and LOEUF . . . 123
15.3.1 The ExAC Study and pLI . . . 124
15.3.2 The gnomAD Study and LOEUF . . . 126
15.4 Summary . . . 129

16 Chapter 12. Recessive Alleles in Human Disorders 129
16.1 Alleles, Genotypes, and Variants . . . 130
16.1.1 Biallelic Variation . . . 130
16.2 Consanguinity and Autozygosity . . . 131
16.3 Recessive Inheritance Patterns in Human Pedigrees . . . 134
16.4 Recessive Alleles in Mendelian and Complex Diseases . . . 134
16.4.1 Mendelian Disorders . . . 134
16.4.2 Complex Diseases . . . 134
16.5 Population Genetics and Selection . . . 135
16.6 Summary . . . 136

17 Chapter 13. Structural Variations in Human Genomes 137
17.1 What Are Structural Variants? . . . 137
17.2 Classification of SVs . . . 137
17.2.1 Copy Number Altering SVs . . . 137
17.2.2 Copy Number Neutral SVs . . . 138
17.3 Distribution of SVs . . . 138
17.3.1 How Many SVs Does Each Person Have? . . . 138
17.3.2 Size Matters . . . 140
17.3.3 Rare Variants Dominate . . . 140
17.3.4 Population Structure . . . 140
17.4 Mechanisms of SV Formation . . . 140
17.4.1 Non–Allelic Homologous Recombination (NAHR) . . . 142
17.4.2 Non–Homologous End Joining (NHEJ) . . . 142
17.4.3 Mobile Element Insertion (MEI) . . . 142
17.4.4 Replication–Based Mechanisms (Fork Stalling and Template Switching) . . . 142
17.4.5 Chromothripsis . . . 143
17.5 Impact of SVs . . . 143
17.5.1 Gene Dosage Effects . . . 143
17.5.2 Structural Rearrangements . . . 143
17.5.3 Selection Pressure . . . 143
17.6 Summary . . . 145

18 Part 3. Polygenic Genetics and Complex Traits: From Single Genes to Genomic Landscapes 146
18.1 Introduction: When Mendel's Laws Meet Thousands of Genes . . . 146
18.2 What This Chapter Is About . . . 147
18.2.1 1. Additive vs. Dominant Alleles . . . 147
18.2.2 2. Heritability: What It Means and What It Doesn't . . . 147
18.2.3 3. Variance Partitioning in Practice . . . 147

18.2.4 4. Genome-Wide Association Studies (GWAS) . . . 147
18.2.5 5. Genetic Architecture of Complex Diseases . . . 147
18.2.6 6. From Statistics to Mechanism . . . 147
18.3 From Pea Ratios to Polygenic Scores . . . 147
18.4 How to Approach This Chapter . . . 148
18.4.1 Think Additively, Not Categorically . . . 148
18.4.2 Understand What Heritability Does and Doesn't Tell You . . . 148
18.4.3 Connect Multiple Evidence Streams . . . 148
18.4.4 Recognize the Spectrum . . . 148
18.5 What Makes This Different From Mendelian Genetics . . . 148
18.6 Learning Objectives . . . 149
18.7 Let's Begin . . . 149

19 Chapter 14. Additive and Dominant Alleles 150
19.1 Dominant Alleles and Achondroplasia . . . 150
19.1.1 How It Works . . . 150
19.1.2 The Genetic Culprit: FGFR3 . . . 150
19.1.3 Why This Mutation Causes Problems . . . 151
19.1.4 Why It's Dominant . . . 153
19.2 Dominant Alleles and Marfan Syndrome . . . 153
19.2.1 The Genetic Culprit: FBN1 . . . 153
19.2.2 Why FBN1 Mutations Cause Tall Stature . . . 153
19.3 Additive Alleles and Normal Variation in Human Height . . . 154
19.3.1 The Genetic Basis of Height Variation . . . 154
19.3.2 How Additive Alleles Work . . . 155
19.4 Comparing Dominant and Additive Alleles . . . 155

20 Chapter 15. Polygenic Model in Traits 156
20.1 Most Traits Are Influenced by Additive Alleles . . . 156
20.2 How Scientists Measure Additive and Dominant Effects . . . 158
20.2.1 Measuring Additive Effects . . . 158
20.2.2 Measuring Dominant Effects . . . 159
20.2.3 Explaining "Orthogonalized Dominance" Without Math . . . 159
20.2.4 Measuring the Overall Contribution to Traits . . . 160
20.2.5 Why Additive Effects Dominate . . . 160

21 Chapter 16. Heritability 160
21.1 Fisher's 1918 Paper and the Polygenic Model . . . 161
21.1.1 Fisher's Unification of Mendel and Biometry . . . 161
21.1.2 Partitioning Phenotypic Variance . . . 161
21.1.3 Additive Alleles and the No-Dominance Assumption . . . 162
21.2 Understanding Heritability . . . 162
21.2.1 What Heritability Means (and Doesn't Mean) . . . 162
21.2.2 Narrow-Sense and Broad-Sense Heritability . . . 163
21.3 How Heritability Is Estimated . . . 163
21.3.1 Covariance and Relatedness . . . 163
21.3.2 Twin Studies . . . 164
21.3.3 SNP-Based Heritability . . . 164
21.4 Why Heritability Changes Across Traits and Environments . . . 165
21.5 Fisher's Model in Modern Context . . . 165
21.6 Putting It All Together . . . 166

22 Chapter 17. Partitioned Variance: Learn from Real-World Examples 166
22.1 The Microbiome and Autism . . . 167
22.1.1 What They Found . . . 167

22.1.2 Why This Matters for Variance Partitioning . . . 167
22.2 Familial Confounding and Additive Factors—Acetaminophen and Autism . . . 168
22.2.1 What They Found at First . . . 168
22.2.2 The Sibling Control Analysis . . . 168
22.2.3 What This Means . . . 168
22.2.4 Connecting This to Fisher's Model . . . 169
22.3 Connecting the Two Examples . . . 169
22.4 What We Learn from These Examples . . . 170

23 Chapter 18. Introduction to Genome-Wide Association Studies 170
23.1 What is GWAS? . . . 171
23.2 How GWAS Works . . . 171
23.2.1 Understanding the Null Hypothesis . . . 171
23.2.2 The Multiple Testing Problem . . . 172
23.3 Visualizing GWAS Results: Manhattan and QQ Plots . . . 172
23.3.1 Manhattan Plot . . . 172
23.3.2 Quantile–Quantile (QQ) Plot . . . 174
23.4 Biological Interpretation: Pathways, Cell Types, and Functional Convergence . . . 174
23.5 Polygenic Scores (PGS) . . . 176
23.5.1 Example: Dilated Cardiomyopathy (DCM) . . . 176
23.6 Application of Polygenic Scores . . . 178
23.7 Ethical Considerations in the Use of Polygenic Scores . . . 178
23.8 Putting It All Together . . . 179

24 Chapter 19. Genetic Architecture of Traits and Disorders 179
24.1 What is Genetic Architecture? . . . 180
24.2 Example 1: Dilated Cardiomyopathy (DCM) . . . 180
24.2.1 Rare High-Impact Variants . . . 180
24.2.2 Polygenic Risk . . . 181
24.2.3 The Plot Thickens: Modifiers and Multi-Hit Models . . . 181
24.2.4 Limited Yield of Clinical Panels . . . 181
24.2.5 Functional Convergence . . . 181
24.3 Example 2: Alzheimer's Disease . . . 182
24.3.1 Rare Mutations in Early-Onset Familial AD . . . 182
24.3.2 The APOE $\varepsilon 4$ Allele . . . 182
24.3.3 Polygenic Background . . . 182
24.3.4 Converging Pathways: Microglial Efferocytosis . . . 183
24.3.5 Composite Model of Risk . . . 183
24.4 Example 3: Autism Spectrum Disorder (ASD) . . . 186
24.4.1 De Novo Variants . . . 186
24.4.2 Inherited Rare Variants . . . 186
24.4.3 Polygenic Risk . . . 186
24.4.4 Syndromic Autism . . . 186
24.4.5 Sex Differences and the Female Protective Effect . . . 186
24.4.6 Functional Convergence . . . 187
24.5 Genetic Heterogeneity and Biological Convergence . . . 189
24.5.1 Features of Complex Genetic Architecture . . . 189
24.6 Why This Matters . . . 190
24.7 Putting It All Together . . . 190

25 Part 4. Recombination, Linkage and Population: From Individual Genomes to Collective Variation 191
25.1 When Genes Refuse to Assort Independently . . . 191
25.2 What This Chapter Is About . . . 191
25.2.1 Six Interconnected Topics . . . 191

25.3 What Makes This Different . . . 192
25.4 How to Approach This Chapter . . . 192
25.4.1 Think in Three Scales . . . 192
25.4.2 Connect History to Biology . . . 192
25.4.3 Master the Data Format . . . 193
25.4.4 Key Principles to Watch . . . 193
25.5 Why This Matters . . . 193
25.6 Learning Objectives . . . 193
25.7 A Note on Terminology . . . 194
25.8 Let's Begin . . . 194

26 Chapter 20. Allele Frequency 194
26.1 Definition of Allele Frequency . . . 195
26.2 How to Calculate Allele Frequency . . . 195
26.3 Minor Allele Frequency (MAF) . . . 196
26.4 Variant Categories by Frequency . . . 196
26.4.1 Commonly Used Classification Schemes . . . 196
26.4.2 Why Do Cutoffs Vary? . . . 197
26.4.3 General Framework . . . 197
26.4.4 Key Insights by Category . . . 198
26.5 How Selection Shapes Allele Frequency . . . 198
26.5.1 The Frequency–Function Relationship . . . 199
26.5.2 Selection Coefficient and Population Size . . . 199
26.5.3 The Allele Frequency–Effect Size Spectrum . . . 199
26.6 Summary . . . 200
26.6.1 Core Concepts . . . 200
26.6.2 Why This Matters . . . 200
26.6.3 Looking Ahead . . . 200

27 Chapter 21. Population Structure 201
27.1 What Is Population Structure? . . . 201
27.2 Race, Ethnicity, Ancestry, and Population—Different Concepts . . . 202
27.3 How Population Structure Arises . . . 202
27.3.1 Geographic Isolation and Drift . . . 202
27.3.2 Founder Effects . . . 203
27.3.3 Population Bottlenecks . . . 203
27.3.4 Admixture . . . 203
27.3.5 Selection and Local Adaptation . . . 203
27.4 Detecting Population Structure . . . 203
27.4.1 The Idea of Distance and Similarity . . . 203
27.4.2 Principal Component Analysis (PCA) . . . 204
27.4.3 Uniform Manifold Approximation and Projection (UMAP) . . . 204
27.4.4 Mapping Human Genetic Diversity . . . 204
27.4.5 Why Dimensionality Reduction Matters . . . 207
27.5 Allele Frequency Differences Among Populations . . . 207
27.5.1 Lactose Tolerance and Dairy Farming . . . 207
27.5.2 Malaria Resistance: Sickle Cell and Beyond . . . 207
27.5.3 High-Altitude Adaptation in Tibetans . . . 207
27.5.4 Population-Specific Rare Variants . . . 208
27.6 Quantifying Genetic Differentiation: FST . . . 208
27.7 Why Population Structure Matters . . . 208
27.7.1 Correcting Confounding in Association Studies . . . 208
27.7.2 Tracing Human History . . . 208
27.7.3 Improving Medical Genetics . . . 209

27.7.4 Ensuring Diversity in Research . . . 209
27.8 Summary . . . 209
27.8.1 Core Concepts . . . 209
27.8.2 Why This Matters . . . 209
27.8.3 Looking Ahead . . . 209

28 Chapter 22. From Mendel to Morgan: Discovery of Linkage 210
28.1 Early Clues: Something Doesn't Add Up . . . 210
28.2 Morgan's Fly Room: Where Genetics Got Physical . . . 211
28.3 The Discovery of Linkage . . . 213
28.4 Sturtevant's Stroke of Genius: The First Genetic Map . . . 213
28.5 How Crossing Over Works . . . 214
28.5.1 Double Crossing Over . . . 215
28.6 Linkage in Humans . . . 217
28.7 Putting It All Together . . . 217
28.8 From Flies to Genomes . . . 218

29 Chapter 23. Recombination, Linkage, and Haplotype 218
29.1 Two Flavors of Recombination: Cross-Over and Non-Cross-Over . . . 219
29.2 What the Study Revealed . . . 219
29.3 How CO and NCO Form: The Molecular Dance . . . 220
29.4 Linkage: Alleles That Travel Together . . . 221
29.5 Regional Differences: Not All Chromosomes Recombine Equally . . . 222
29.6 Haplotype: Blocks of Inherited DNA . . . 224
29.7 Why Recombination Matters . . . 225
29.8 Putting It All Together . . . 225

30 Chapter 24. Data Type for Alleles and Population 226
30.1 From Sequencing to Variant Calling . . . 226
30.2 VCF File Structure . . . 226
30.2.1 Core Columns . . . 227
30.3 The INFO Field: Population-Level Statistics . . . 227
30.4 The FORMAT Field and Genotypes . . . 228
30.4.1 Genotype Codes . . . 228
30.5 Multi-Sample VCF: Population Data . . . 229
30.6 Joint Genotyping: Complete Data . . . 229
30.7 Storage Strategies: Dense vs. Sparse . . . 229
30.7.1 Dense vs. Sparse VCF Comparison . . . 230
30.8 The Future of VCF: Scaling for a Genomic Revolution . . . 230
30.8.1 The Scaling Challenge . . . 230
30.8.2 GA4GH's Approach to Scaling . . . 230
30.8.3 Three Approaches to Scaling . . . 231
30.8.4 Why VCF Scaling Matters . . . 231
30.9 Why VCF Matters . . . 231

31 Part 5. From Variant to System: Experimental Genetics and Gene Regulation 232
31.1 When Finding the Gene Wasn't Enough . . . 232
31.2 What This Chapter Is About . . . 232
31.2.1 Five Interconnected Topics . . . 233
31.3 What Makes This Different . . . 233
31.4 How to Approach This Chapter . . . 233
31.4.1 Think in Complementary Directions . . . 233
31.4.2 Connect Association to Mechanism . . . 234
31.4.3 Master the Experimental Logic . . . 234
31.4.4 Integrate Data Types . . . 234

31.5 Why This Matters . . . 234
31.6 Learning Objectives . . . 234
31.7 What Makes This Different from Classic Genetics . . . 235
31.8 Let's Begin . . . 235

32 Chapter 25. Forward Genetics: From Phenotype to Gene 236
32.1 The evolution of gene hunting . . . 236
32.2 Genome-wide association studies (GWAS) . . . 237
32.2.1 What GWAS tells you . . . 238
32.2.2 What GWAS doesn't tell you . . . 239
32.3 Burden tests: hunting rare variant effects . . . 239
32.3.1 How burden tests work . . . 240
32.3.2 Burden test output . . . 241
32.4 GWAS vs burden tests: two sides of the same coin . . . 241
32.4.1 Example: Schizophrenia genetic architecture . . . 242
32.5 From association to causation . . . 243

33 Chapter 26. Reverse Genetics: From Gene to Function 243
33.1 The logic of reverse genetics . . . 244
33.2 Classic reverse genetic approaches . . . 244
33.2.1 Loss of function: what happens when the gene is gone? . . . 244
33.2.2 Gain of function: what happens when you add more? . . . 244
33.2.3 Conditional and inducible systems: controlling when and where . . . 244
33.3 CRISPR: a unified platform for reverse genetics . . . 245
33.3.1 CRISPR knock-outs: cutting DNA . . . 245
33.3.2 Base editing and prime editing: precision changes . . . 245
33.3.3 CRISPRa and CRISPRi: controlling gene expression . . . 246
33.3.4 Cas13: targeting RNA . . . 246
33.4 High-throughput reverse genetics: screening thousands of genes at once . . . 246
33.4.1 Pooled CRISPR screens . . . 246
33.4.2 Perturb-seq: linking perturbations to transcriptomes . . . 247
33.4.3 MPRA: testing regulatory sequences . . . 247
33.5 Functional rescue: proving causality . . . 249
33.5.1 1. Gene complementation: adding back a normal copy . . . 249
33.5.2 2. Gene correction: fixing the mutation . . . 250
33.5.3 3. Allelic activation: boosting the remaining copy . . . 250
33.6 Putting it all together: from variant to mechanism . . . 250
33.7 The future: integrating forward and reverse genetics . . . 251

34 Chapter 27. CRISPRa Therapy for SCN2A Haploinsufficiency 251
34.1 SCN2A and the haploinsufficiency problem . . . 251
34.2 Not all SCN2A variants are the same . . . 252
34.3 What is CRISPRa, and how does it work? . . . 253
34.4 Rescuing SCN2A haploinsufficiency with CRISPRa . . . 255
34.4.1 Results in mouse neurons . . . 255
34.4.2 Results at the circuit and behavioral level . . . 256
34.4.3 Results in human cells . . . 258
34.5 Why this matters: dosage restoration therapy . . . 258
34.6 How does this compare to other approaches? . . . 259
34.6.1 Gene replacement therapy . . . 259
34.6.2 Base editing or prime editing . . . 259
34.6.3 Small molecule drugs . . . 259
34.6.4 CRISPRa . . . 260
34.7 What does "rescue" mean? . . . 260

34.8 Looking ahead: cis–regulatory gene therapy . . . 261
34.9 Key takeaways . . . 262
34.10 Key terms . . . 262

35 Chapter 28. Gene Regulation: How Cells Read the Same Book Differently 262
35.1 The central dogma, revisited . . . 263
35.2 Promoters: where transcription starts . . . 264
35.3 Enhancers: long–range control . . . 265
35.3.1 Key properties of enhancers . . . 265
35.3.2 They work at a distance . . . 266
35.3.3 They integrate multiple signals . . . 266
35.3.4 They're marked by specific histone modifications . . . 266
35.3.5 They evolve rapidly but functionally important ones are conserved . . . 267
35.4 How enhancers contact promoters: 3D chromatin looping . . . 267
35.4.1 A real example: IGF2/H19 imprinting . . . 269
35.5 Key takeaways . . . 270
35.6 Key terms . . . 271

36 Chapter 29. Gene Regulation: Methods and Applications 271
36.1 Measuring gene expression: RNA–seq . . . 271
36.2 Bulk vs single–cell RNA–seq . . . 273
36.3 Measuring gene regulation: chromatin profiling . . . 276
36.3.1 Chromatin states and histone modifications . . . 276
36.3.2 ChIP–seq: mapping protein–DNA interactions . . . 278
36.3.3 ATAC–seq: mapping chromatin accessibility . . . 279
36.3.4 CUT&RUN: high–resolution chromatin profiling . . . 282
36.3.5 Putting it all together: integrated epigenomics . . . 284
36.4 Why this matters for genetics . . . 286
36.5 Key takeaways . . . 287
36.6 Key terms . . . 287

37 Chapter 30. QTLs: Connecting alleles to Molecular Traits 287
37.1 What is a QTL? . . . 288
37.2 How QTL mapping works . . . 288
37.2.1 Encoding genotypes . . . 288
37.2.2 The molecular phenotype matrix . . . 288
37.2.3 Linear regression . . . 288
37.2.4 Cis vs trans QTLs . . . 289
37.2.5 Multiple testing . . . 289
37.3 eQTLs: genetic effects on RNA . . . 289
37.4 pQTLs: genetic effects on protein . . . 289
37.5 Example 1: Developmental plasma pQTLs in children . . . 290
37.6 Example 2: Parent–of–origin effects in pQTLs . . . 290
37.7 Example 3: Multi–ancestry brain pQTLs . . . 292
37.8 Interpreting QTLs: what they tell you (and what they don't) . . . 294
37.8.1 Association is not causation . . . 294
37.8.2 Tissue and cell–type specificity . . . 294
37.8.3 eQTLs don't always predict pQTLs . . . 294
37.8.4 Effect sizes are usually small . . . 294
37.8.5 Statistical challenges . . . 295
37.9 Integrating QTLs with GWAS: finding causal genes . . . 295
37.10 QTLs in systems biology: building regulatory networks . . . 295
37.11 Summary . . . 295

1 Introduction



Figure 1: Front Cover

Welcome to **Human Genetics**

Author: Joon-Yong An, Korea University

Version: Draft (Under Construction)

1.1 About This Book

This textbook introduces the foundation and cutting edge of human genetics — from Mendel's early concept of heritable factors to the modern era of genome sequencing and large-scale cohort discovery. It is designed for undergraduate students in biology and biomedical sciences, particularly those interested in pursuing careers as data scientists in large-scale genomics for human disorders.

The field of genetics has been rapidly transformed by next-generation sequencing, large-scale biobank cohorts, and international data sharing initiatives. These advances now allow us to revisit traditional genetic concepts with fresh perspectives and test long-standing models and principles against comprehensive

empirical evidence. This textbook takes a broad approach to cover the essential topics that undergraduate students need to master, integrating classical genetic theory with contemporary discoveries and methodologies.

Rather than focusing primarily on model organisms or molecular genetics from a molecular biology perspective, this book centers on the key concepts and emerging insights specific to human genetics. Students will not only learn the facts of genetics, but also develop the evolving perspectives that have shaped — and continue to reshape — our understanding of human heredity and disease.

Please note that chapters are currently being written and improved. The complete version is expected to be finished by January 2026.

1.2 Table of Contents

- Getting started

1.3 Part 1. The Human Genome Project and NGS

- Introduction
- Chapter 1. The Human Genome Project
- Chapter 2. T2T Genome Project
- Chapter 3. CHM13 Cell Line
- Chapter 4. Human Pangenome Reference Consortium
- Chapter 5. Next-Generation Sequencing
- Chapter 6. NGS Application
- Chapter 7. Annotation & Database

1.4 Part 2. Mendelian Genetics in NGS

- Introduction
- Chapter 8. Genetic Variant
- Chapter 9. Transmission
- Chapter 10. Dominant Allele
- Chapter 11. Haploinsufficiency
- Chapter 12. Recessive Allele
- Chapter 13. Structural Variation

1.5 Part 3. Polygenic and Genetic Architecture

- Introduction
- Chapter 14. Additive and Dominant Allele
- Chapter 15. Polygenic Model
- Chapter 16. Heritability
- Chapter 17. Partioned Variance
- Chapter 18. GWAS
- Chapter 19. Genetic Architecture

1.6 Part 4. Recombination, Linkage and Population

- Introduction
- Chapter 20. Allele Frequency
- Chapter 21. Population Structure
- Chapter 22. Discovery of Linkage

- Chapter 23. Linkage in NGS era
- Chapter 24. Data Type for Alleles and Population

1.7 Part 5. Variant to System

- Introduction
- Chapter 25. Forward Genetics
- Chapter 26. Reverse Genetics
- Chapter 27. CRISPRa in human disorder
- Chapter 28. Gene Regulation
- Chapter 29. Methods and Applications
- Chapter 30. QTL

2 Getting Started

2.1 Why I wrote this book for undergraduate students

If you learned genetics in Korean high school, you probably remember learning about "유전자" (gene) when studying Mendel's pea plant experiments. Later, you might have encountered "대립유전자" (allele) and thought, "Isn't that just another type of 유전자?" This confusion isn't your fault—it's a symptom of how biological concepts have evolved faster than the language we use to teach them. The two Korean terms sound like they're related (one just has 대립- added to the front), but they actually refer to completely different biological concepts. A gene is a functional unit of DNA that codes for a protein or RNA, like a recipe in a cookbook. An allele is a specific sequence variant at a position in your genome, like having version A versus version B of that recipe. The distinction matters profoundly when you move from textbook exercises to real research.

Many of you will graduate and enter research labs and companies where you'll analyze NGS (Next-Generation Sequencing) data. There, you'll encounter terms like "genetic variants," "allele frequencies," and "SNVs"—concepts that seem disconnected from what your textbooks taught. This gap exists because both high school and university textbooks often teach genetics from perspectives developed decades ago, while modern genomics operates with fundamentally different frameworks. But this disconnect reveals something deeper about how science actually works: it is not about memorizing terms, but about how our perspectives evolve. Biology has advanced not because new words were invented, but because scientists began to see the same phenomena differently.

Genetics shows this transformation more clearly than perhaps any other field. Over the past 150 years, what we now call an "allele" evolved from an invisible "factor" to a measurable genetic variant. Each generation of discovery didn't erase the previous one—it expanded what "allele" could mean, revealing new layers of how inheritance actually works.

Consider what Mendel himself wrote in 1866. In his original German paper, he never used words like "gene" or "allele." Instead, he wrote about Factoren—factors:

> "Wir müssen es daher als nothwendig erachten, dass auch bei Erzeugung der constanten Formen an der Hybridpflanze vollkommen gleiche Factoren zusammenwirken." Mendel, G. (1866). Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV, 3–47."
>
> (We must therefore consider it necessary that perfectly equal factors also work together in the production of the constant forms on the hybrid plant.)

Mendel never saw these Factoren. He inferred their existence purely from mathematical ratios—3:1, 9:3:3:1—in pea plants. He proposed that hybrids must produce different types of reproductive cells, each carrying different combinations of these invisible factors. This was radical: reducing inheritance to countable, combinable units that followed predictable laws, like equations in physics.

What Mendel called Factoren became the conceptual ancestor of both "gene" and "allele"—but these two descendants now mean very different things. The table below traces how scientific understanding evolved, and how what began as a single abstract concept gradually split into distinct ideas.

Era	Conceptual View	Key Idea
1860s	**Mendel's Factoren (Gregor Mendel)**	Invisible hereditary units determine observable traits through predictable ratios
1900s	**Chromosome Theory of Inheritance (Walter Sutton, Theodor Boveri)**	Mendel's hereditary factors correspond to specific parts of chromosomes
1910s	**Linkage and Recombination (Thomas Hunt Morgan)**	Factors are arranged linearly on chromosomes and can recombine during meiosis
1940s	**One Gene–One Enzyme Hypothesis (George Beadle & Edward Tatum)**	Each factor (now "gene") encodes a specific enzyme → gene as functional unit
1950s	**Central Dogma (James Watson, Francis Crick, Sydney Brenner)**	DNA → RNA → Protein; genetic information flows from DNA sequence to biological function
1960s–1970s	**Regulatory Unit (François Jacob, Jacques Monod, Barbara McClintock)**	Genes act as dynamic regulatory units; variation in genes creates different outcomes
2000s–Now	**Genome as a Collection of Variants (NGS Era)**	Genetic individuality is defined by sequence variation → allele as genetic variant

This journey—from invisible Factoren to molecular information, and now to measurable sequence variation—marks one of the most profound transformations in biology. What began as Mendel's mathematical inference became a physical chromosome segment in the early 1900s, then a DNA sequence in the 1950s, and finally variants we can read directly from your genome at single-nucleotide resolution. Along the way, that original concept of Factoren split into two related but fundamentally different ideas: the gene as a functional unit (a DNA segment encoding a protein or RNA), and the allele as a sequence variant (version A versus version B at a specific genomic position). When you analyze NGS data in a research lab, you're not studying "genes" in Mendel's sense—you're identifying alleles, the specific DNA variants that differ between individuals and populations.

To study genetics is not simply to learn definitions, but to follow how scientists have redefined what it means to inherit information. Many students arrive at university after years of exam preparation, where memorization often dominated science education. But modern biology evolves rapidly as new data emerge daily. Because biology interprets natural phenomena, our understanding shifts as new evidence accumulates. What matters most is not remembering every term, but learning to interpret data and think about the principles that best explain life right now—and recognizing that those principles will themselves evolve as new technologies and discoveries emerge.

In this textbook, we'll examine essential concepts through key scientific papers and explore how they're applied in real research today. By connecting historical discoveries like Mendel's Factoren to current genomic approaches like variant calling and population genetics, you'll see how biological understanding grows through evidence and reinterpretation. You'll learn not just what scientists currently think, but why they think it—and how new data might change that thinking tomorrow.

Science advances when perspectives evolve—and your own research may one day become part of that evolution.

2.2 From consuming knowledge to creating it

The confusion between "gene" and "allele" isn't just a translation problem—it reflects something deeper about how biological knowledge evolves. Every generation of scientists inherits not only data, but also a **paradigm**: a shared framework for interpreting the world. A paradigm defines what counts as a meaningful question, what methods are trusted, and what kinds of explanations make sense.

Consider the historical shifts we traced earlier. When Mendel's invisible Factoren became visible as chromosomes under the microscope, biologists didn't just gain new facts—they gained a **new way of asking questions**. Before chromosomes were seen, the question "Where in the cell are hereditary factors located?" seemed almost philosophical. After chromosomes, it became an experimental question with testable answers. When Watson and Crick revealed DNA's structure, the paradigm shifted again: genes became **information** encoded in sequence, not just physical substances. The Human Genome Project shifted it once more: suddenly, biology became a data science where computational analysis mattered as much as laboratory experiments. These weren't just discoveries—they were **changes in how we recognize and organize knowledge itself**. Each technological advance didn't simply add information; it transformed what "understanding genetics" actually meant.

In Korean education, success has traditionally meant absorbing established knowledge and reproducing it accurately. This approach works well when knowledge is stable and well-defined. But as you transition into research, your role fundamentally changes—from knowledge consumer to knowledge producer. Most research questions have **no predefined answers**. More importantly, they often don't even have predefined language yet.

When researchers first encountered millions of DNA sequence variants from the 1000 Genomes Project, existing terminology struggled to keep up. Were these all "mutations"? "Polymorphisms"? "Alleles"? The field had to negotiate new vocabulary because the old categories didn't capture what sequencing revealed. You'll face similar moments in your career—situations where you must interpret patterns that don't yet have names, or where established terms don't quite fit what you observe.

This is why learning to express ideas **in your own language** becomes crucial. Science advances not through repetition, but through reinterpretation. Someone must look at familiar data through a new lens, articulate what they see differently, and convince others that this new perspective reveals something important. That "someone" could be you.

In this book, we'll focus on the **logic of discovery**, not just conclusions. You'll read historical milestones and modern papers not to memorize their results, but to understand **how ideas shift when technology lets us see differently**. Recognizing these patterns will help you navigate future paradigm shifts—and perhaps contribute to them.

Your scientific journey begins not with knowing all the answers, but with learning to ask: What do we believe today, and why? What might the next generation see differently?

2.3 Why this course feels different

Some students tell me this course feels too computational or mathematical for biology. They're right—and that reflects how genetics itself has transformed. When I was a student in the early 2000s, understanding genetics meant knowing key mechanisms: how transcription works, how proteins fold, how genes regulate cell behavior. Today, the questions have fundamentally changed. We don't ask "What does this gene do?" but "How do thousands of genetic variants collectively influence disease risk?" We don't just describe what genes are expressed, but build models to predict expression patterns from DNA sequence using machine learning.

This shift happened because our ability to measure life exploded. We went from studying individual genes to analyzing millions of variants simultaneously. From microscope observations to computational predictions. Biology became a data science. That's why this textbook looks different. When we discuss alleles, we'll also examine allele frequencies and Hardy–Weinberg equilibrium—not as abstract formulas, but as tools for interpreting real population genomic data. When we explore gene expression, we'll see how RNA–seq data are quantified, because understanding data generation shapes interpretation.

You're entering biology at a remarkable moment: the boundary between wet–lab experiments, computation, and theory has never been thinner. Some concepts from earlier courses may look different here—not because your previous education was wrong, but because you're now seeing genetics through the lens contemporary researchers actually use. This course prepares you not just to understand today's genetics, but to help create tomorrow's.

2.4 Learning the Language of Science

Scientific papers often feel difficult not because of their content, but because they use a special dialect—what we call **scientific jargon**. Every field, every era, and even every research group develops its own preferred terms and expressions. When you read a paper, remember that a single phrase or word is **not a fixed dogma**, but the author's best attempt to describe a complex idea in the language available at that time. Learning to read science means learning to translate these expressions—to understand the author's intention rather than memorizing the phrasing.

2.4.1 Types of Scientific Papers

Not all papers serve the same purpose. Recognizing the type helps you approach it with the right expectations:

Paper Type	Purpose	When to Read
Article (Research Article)	Reports original experimental findings—new data, new discoveries	To see how researchers test hypotheses and generate evidence
Article (Resource)	Presents large datasets or computational tools, common in genomics and omics fields	When you need access to data or methods for your own analysis
Review	Summarizes and synthesizes existing research on a topic	To understand the big picture and catch up on a field (no new data)
Perspective	Offers an author's viewpoint on a field's direction or interpretation	To understand debates about what questions matter most or how to interpret conflicting results
Methods	Describes a new experimental or computational technique	When you need to understand how something is done, not what was discovered

2.4.2 Practical Tips for Reading Scientific Papers

1. **Start with the big picture.** Read the **title**, **abstract**, and **figure legends** first. Try to understand what question the authors asked and why it matters—not every technical detail.

2. **Focus on the question, not the methods.** Beginners often get lost in experimental procedures. Instead, ask: "What were they trying to test?" and "Did their data support that idea?" The logic matters more than the jargon.
3. **Read the figures before the text.** Figures are the core of a paper. Even if you don't understand all of them, try to guess what story the figure is telling—it trains scientific reasoning.
4. **Don't trust AI summaries blindly.** Tools that summarize papers can be helpful, but they often miss nuances, misinterpret complex arguments, or oversimplify critical details. Always read the original paper yourself—especially the figures and key results sections—to verify what the AI told you.
5. **Avoid secondary sources for cutting-edge genomics.** For the latest genomic and genetic research, **Korean-language resources simply don't exist**—not on Naver, not on Tistory blogs, not even in translated news articles. These sources are often outdated, oversimplified, or outright incorrect. If you're serious about understanding current research, you must read the original English papers.
6. **Wikipedia vs. research papers.** Wikipedia offers consensus knowledge—a stable background. Papers reflect active discovery—sometimes uncertain or even contradictory. The difference between them is what makes science dynamic.
7. **Annotate as you read.** Write short notes in the margins or keep a glossary of terms. Over time, you'll build your own understanding of the field's language.
8. **Check citation formats.** et al. means "and others." It's used to cite studies with multiple authors, e.g., Kim et al. (2024). When you write, always include the year and keep the format consistent.

Reading scientific papers is like learning a new language—you start by recognizing patterns, then understanding meaning, and finally gaining fluency. It's less about decoding every word and more about **joining an ongoing conversation**: learning how scientists reason, argue, and refine their understanding through shared curiosity and a willingness to keep interpreting.

2.5 Using AI to Read Papers More Effectively

Artificial intelligence can be a powerful companion in learning to read scientific papers—**and yes, in reading this textbook too**. Many of you will use AI to summarize chapters, clarify concepts, or review material before exams. This is understandable, even expected. But AI is most useful when you use it to think with, not to think for you.

2.5.1 Using AI with This Textbook

Let's be honest: you will likely ask AI to summarize sections of this book. That's fine—but understand what you gain and what you lose.

What AI summaries can do: – Give you a quick overview before reading in detail – Help you review key points after you've studied – Clarify a confusing paragraph you've already tried to understand

What AI summaries cannot do: – Replace the experience of working through an explanation step-by-step – Teach you how concepts connect across chapters – Help you develop the intuition that comes from seeing examples worked out in full

A summary might tell you "Hardy-Weinberg equilibrium describes allele frequencies in a non-evolving population." But this textbook shows you why that matters, how real populations deviate from it, and what those deviations teach us about evolution and disease. The summary gives you words to memorize; the textbook gives you a way of thinking.

A better approach: Use AI as a conversation partner while you read. For example:

```
I just read the section on linkage disequilibrium in Chapter 3. The textbook
says LD decays over generations due to recombination, but I'm confused about
why some regions have high LD even after many generations. Can you help me
think through this?
```

[Claude discusses factors that maintain LD: selection, population structure, low recombination rates]

```
So if I'm looking at a genomic region with high LD in a GWAS, that could mean
either low recombination OR recent selection? How would I tell the difference?
```

[Conversation continues, building understanding beyond what a simple summary provides]

This kind of dialogue helps you think with the material, not just consume it passively.

2.5.2 Using AI to Read Scientific Papers

When you encounter a difficult paper, you can use AI to: – **Summarize** sections in simpler language and check if you understood them correctly. – **Ask "why" and "how" questions**, not just "what." For example: "Why did the authors use RNA-seq instead of microarrays?" or "How does this figure support their main claim?" – **Compare interpretations.** You can ask the AI how two papers differ in their conclusions or methods—a skill essential for critical reading. – **Translate specialized language.** If a term or phrase feels confusing, let AI explain it in everyday language, then return to the paper and see how that meaning fits in context.

However, remember these rules:

1. **AI cannot replace your judgment.** It can explain what something means, but not decide whether it makes sense. Always verify claims by checking the data and figures in the paper itself.
2. **Avoid using AI-generated summaries as your only source.** Summaries are interpretations, not neutral facts. Always read the original paper—especially the figures and key results—to see what nuances were lost.
3. **Be critical of citations.** When AI mentions a reference, check whether that study really exists and whether it supports the point being made. This practice trains you to validate information—a core scientific skill.
4. **Use AI to practice paraphrasing.** Try explaining a result in your own words, then ask the AI to review whether your version keeps the meaning. This is how real understanding develops.

2.5.3 Recommended AI Tools for This Course

You can use several large language models (LLMs) to study and discuss scientific papers and textbook material. While many tools exist, **I particularly recommend two for this course:**

Claude (Anthropic) — My top recommendation for this course. Claude excels at: – Long-form reasoning and interpreting complex scientific papers – Having extended conversations where you build understanding iteratively – Handling technical genetics concepts with nuance – Reading uploaded PDFs of papers and textbook chapters directly

NotebookLM (Google) — Excellent for: – Uploading and organizing multiple papers or textbook chapters – Creating study guides and summaries from your uploaded materials – Asking questions across several documents simultaneously – Generating practice questions based on course content

Other tools you might encounter: – **ChatGPT (OpenAI)** — balanced general reasoning and academic reading ability – **Gemini (Google)** — strong integration with Google search and datasets

All of these tools can assist in reading, summarizing, and questioning research papers and textbooks. The goal is not to let the AI read for you, but to **learn to read** with **the AI**.

2.5.4 Writing Good Prompts: A Conversation Example

How you ask determines what you learn. Rather than asking isolated questions, treat the AI like a study partner and build your understanding through dialogue. Here's an example of how a student might work through understanding linkage disequilibrium (LD) by conversing with Claude:

```
I'm reading a GWAS paper and they keep mentioning "linkage disequilibrium."
I know it has something to do with alleles being inherited together, but I don't
really understand why it matters. Can you explain it like I'm a second-year
biology student?
```

[Claude explains LD as non-random association of alleles at different loci]

```
Okay, that makes sense. But why does LD matter for GWAS? The paper says they
"pruned SNPs for LD" before analysis-what does that mean and why would you
want to do that?
```

[Claude explains that nearby SNPs are often correlated, and pruning removes redundant information]

```
Wait, so if two SNPs are in high LD, they're basically giving you the same
information? But then how do you know which one is actually causal for the
disease?
```

[Claude explains the difference between causal variants and tag SNPs, and the challenge of fine-mapping]

```
So in the paper, when they find a significant SNP, it might not be the actual
causal variant-it could just be tagging the real one nearby? How do researchers
figure out which is which?
```

[Conversation continues, building deeper understanding through follow-up questions]

Notice how the conversation evolves: starting with a basic concept, then connecting it to the paper's methods, then questioning the interpretation, and finally thinking critically about causality. This is how real understanding develops—not through a single explanation, but through **iterative questioning**.

2.5.5 Understanding AI's Limitations

Even the best AI models make mistakes. Sometimes they **misread parts of a paper** or provide **incomplete interpretations**. This happens because language models have a limit on the number of tokens (the pieces of text) they can read and generate at once—so they may only "skim" a long article instead of processing it fully.

In addition, AI systems still suffer from **hallucination**—producing confident but incorrect statements or fabricated references. These errors arise because the model predicts plausible text patterns rather than verifying factual truth.

Therefore: - Always **cross-check** AI explanations with the original paper or trusted databases. - Do not rely on AI to do your thinking—use it to accelerate your own. - Remember that AI is a **tool to make your work smoother**, not a shortcut to reduce your study time by a hundredfold.

2.5.6 The Real Goal: Learning to Think, Not Just to Know

Here's the truth: you can use AI to pass this course without deeply learning the material. You can ask it to summarize every chapter, answer practice questions, even explain concepts just well enough to get through exams.

But genetics in our days is not about memorizing facts—it's about **thinking like a geneticist**. It's about seeing a GWAS result and immediately wondering about population structure. Reading an RNA-seq analysis and questioning the normalization method. Encountering a new disease gene and asking how its function connects to phenotype.

That kind of thinking only develops through **engagement**, not summarization. AI can help you engage more effectively—by answering your questions, challenging your assumptions, and walking you through complex ideas. But it can't do the thinking for you.

Use AI boldly, but never let convenience replace curiosity. The goal is not to read faster, but to **see more deeply**—to evolve your scientific perspective.

2.6 Welcome to Human Genetics

This course will be challenging. There will be moments when a paper feels impenetrable, when a concept refuses to click, when you wonder if you're cut out for this kind of thinking. Those moments are not signs of failure—they're signs you're learning something genuinely new. Every researcher, including your instructors, still experiences them regularly. Approach this course with curiosity, not caution. Use every tool available, but never let convenience replace genuine thinking.

Welcome. Let's begin.

3 Part 1. The Human Genome Project and Next-Generation Sequencing: A Revolution in Modern Biology

3.1 From Invisible Factors to Readable Genomes

In the 1860s, Gregor Mendel discovered that traits like pea flower color don't blend—they remain discrete, passing from generation to generation as distinct units he called **"factors."** He couldn't see these factors or measure them directly, but by counting offspring and analyzing patterns, he deduced they must follow specific rules.

Mendel was doing genetics without seeing genes—like discovering planets by observing gravitational effects without ever seeing them through a telescope.

For over a century, genetics remained this way. Scientists could map genes to chromosomes and study mutations, but they couldn't directly read the genetic information itself. Then came the molecular biology revolution: Watson and Crick revealed DNA's structure in 1953, Sanger sequencing in the 1970s made it possible to read hundreds of DNA bases, and by the 1990s, automated sequencing could read thousands.

But reading an entire human genome—3.2 billion letters long—seemed impossible.

The **Human Genome Project** said: let's do it anyway.

3.2 What This Chapter Is About

This chapter tells the story of how genetics transformed from an observational science into a digital one—how we went from inferring genes to directly reading them, base by base.

You'll learn about the technological and conceptual advances that made this revolution possible:

3.2.1 The Foundation: Reference Genomes

The Human Genome Project (2003) provided the first map of human DNA, revealing where genes are located and creating a reference sequence for comparing individual genomes.

Reference genome evolution progressed from the initial GRCh37/hg19 to GRCh38/hg38, culminating in the complete, gapless **T2T-CHM13** genome in 2022—finally filling all the gaps.

The pangenome addresses a fundamental limitation: a single reference can't represent human diversity. The Human Pangenome Project builds a collection of genomes from diverse populations to capture the full spectrum of human variation.

3.2.2 The Technology: Sequencing Platforms

Next-Generation Sequencing (NGS) dropped sequencing costs from $300 million per genome to under $1,000 in two decades—100,000-fold reduction, far exceeding Moore's Law.

Platform comparison: - **Illumina**: Short reads (150–300 bp), high accuracy, dominates clinical applications - **PacBio**: Long reads (10–30 kb), excellent for repetitive regions and structural variants - **Oxford Nanopore**: Ultra-long reads (100+ kb), portable, real-time sequencing

3.2.3 The Strategy: What to Sequence

WGS vs. WES: Should you sequence the entire genome or just protein-coding regions? Each approach has distinct advantages:

Approach	Coverage	Cost	Best For
WES	~2% of genome (exons)	Lower	Mendelian diseases, clinical diagnostics
WGS	Entire genome	Higher	Structural variants, regulatory regions, research

3.2.4 The Interpretation: Making Sense of Data

Variant annotation adds biological context: Is this variant in a gene? Does it change the protein? Is it common or rare? Has it been seen in disease?

Databases provide the knowledge: dbSNP catalogs variants, gnomAD shows population frequencies, ClinVar links variants to disease.

3.3 Connecting to Mendel: From Ratios to Sequences

Mendel's "factors" are DNA sequence variants. When he crossed purple and white pea plants: - **Mendel saw**: 3 purple : 1 white ratio in offspring - **Modern interpretation**: One DNA sequence produces pigment (purple), another has a mutation breaking the gene (white)

Today, instead of inferring genes from phenotype ratios, we read the DNA directly:

```
Mendel's approach:      Modern approach:
Purple × White          Sequence both plants
     ↓                       ↓
Count offspring        Compare DNA sequences
     ↓                       ↓
Infer factors         Find the exact G→T mutation
     ↓                       ↓
3:1 ratio             Same 3:1 ratio explained molecularly
```

The patterns Mendel discovered remain valid. We've just moved from observing effects to observing causes.

3.4 The Democratization of Genomics

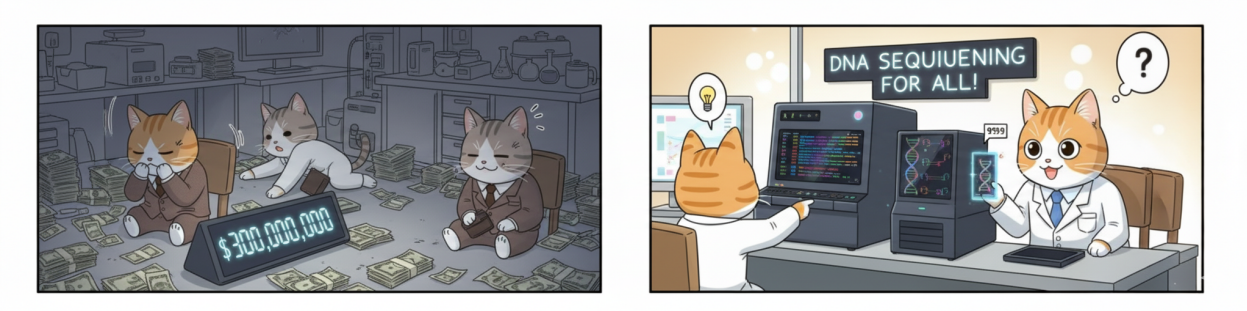


Figure 2: NGS Price Drop

2001: First human genome cost ~$100 million, took months
2024: Genome sequencing costs ~$600–1,000, takes days

This 100,000–fold cost reduction has transformed genomics from elite research into a clinical tool:

Medicine: – Diagnose rare genetic diseases (ending years of diagnostic odyssey) – Guide cancer therapy based on tumor mutations – Predict disease risk for early intervention – Enable pharmacogenomics (right drug, right dose)

Research: – Trace human evolution and migration – Discover new disease genes – Understand gene function – Study population diversity

But this revolution required more than cheaper sequencing. It needed reference genomes to compare against, computational tools to analyze billions of bases, databases to interpret variants, and new ways of thinking about genetics as a data science.

3.5 How to Approach This Chapter

This chapter differs from traditional textbooks—it tells a story of technological evolution rather than presenting concepts to memorize.

3.5.1 Key Themes to Watch For

1. Technology enables new questions
Once sequencing became affordable, impossible questions became routine. This pattern repeats throughout the chapter.

2. Better references, better answers
The same sequencing data yields more accurate results with improved reference genomes (GRCh37 → GRCh38 → T2T–CHM13 → Pangenome).

3. No single "right" approach
WGS vs. WES? Short reads vs. long reads? It depends on your question, budget, and sample. Modern genetics requires choosing the right tool for the job.

4. We're still learning
Even with complete genomes, we don't fully understand what most DNA does. Interpreting non–coding variants remains challenging. The field evolves rapidly.

3.5.2 Focus on Understanding

Rather than memorizing technical details, focus on: – **What problem** each technology solved – **What became possible** that wasn't before – **What limitations** remain – **When to use** each approach

---

3.6 The Paradigm Shift

3.6.1 Before HGP and NGS

- Study individual genes or small regions
- Use indirect methods (linkage analysis)
- Work with model organisms (human genetics too slow)
- Focus on simple Mendelian traits

3.6.2 After HGP and NGS

- Study entire genomes simultaneously
- Directly measure genetic variation
- Routine human genome sequencing in clinics
- Understand complex multi-gene traits
- Integrate genomics with other biological data

This shift—from studying genes one at a time to studying complete genomes—is as profound as the shift from studying individual stars to mapping entire galaxies.

Mendel's laws still apply. But now we can see those genes, read their sequences, track variants across populations, and understand how sequence differences create trait differences.

3.7 Learning Objectives

By the end of this chapter, you should be able to:

1. **Explain how Mendel's "factors" connect to modern DNA variants**, understanding how the same biology can be studied at different levels
2. **Describe the Human Genome Project's impact**, including what it achieved, what remained unfinished, and why reference genomes matter
3. **Compare reference genome versions** (GRCh37 → GRCh38 → T2T-CHM13 → Pangenome) and explain why diverse references are needed
4. **Distinguish NGS platforms** (Illumina, PacBio, Nanopore) by their principles, strengths, limitations, and appropriate applications
5. **Compare WGS and WES**, knowing when each approach is appropriate for research or clinical use
6. **Explain the sequencing workflow** from DNA sample to variant calls, including library prep, sequencing, alignment, and variant calling
7. **Understand variant interpretation** using databases (dbSNP, gnomAD, ClinVar) and why population data distinguishes benign from pathogenic variants
8. **Think critically about genomic data**, recognizing that raw sequences require quality control, appropriate references, and careful interpretation

---

3.8 Let's Begin

In 1990, scientists embarked on what seemed impossible: reading all 3 billion letters of the human genome. They didn't know exactly how they'd do it, how long it would take, or whether existing technology could achieve it.

But they knew it needed to be done.

What followed transformed biology—turning genetics from an observational science into a data science, from studying inheritance patterns to reading the inherited information directly.

This is the story of that transformation.

4 Chapter 1. The Human Genome Project

Imagine trying to understand how a car works by only looking at individual parts—a spark plug here, a piston there—without ever seeing the complete engine or having a manual. That's essentially how genetics worked before 1990.

Scientists knew about individual genes. They could identify genes responsible for specific diseases, like the gene for cystic fibrosis or sickle cell anemia. They understood that DNA carried genetic information. But they had no comprehensive map showing where all the genes were located, how many genes humans had, or what most of our DNA actually did.

Each discovery required years of painstaking work to isolate and characterize a single gene. If you wanted to study a new gene, you had to start almost from scratch, using indirect methods to figure out roughly which chromosome it was on, then narrowing down its location bit by bit. It was like searching for a specific house in a city without having a map or knowing the address.

The **Human Genome Project (HGP)**, launched in 1990, aimed to change this fundamentally. The goal was audacious: sequence all 3 billion DNA base pairs in the human genome and create a complete reference map that any scientist could use. This would be biology's equivalent of the periodic table in chemistry or the standard model in physics—a foundational resource that would accelerate all future research.

4.1 What Did the HGP Set Out to Do?

The Human Genome Project had several interconnected goals:

4.1.1 1. Sequence the Entire Genome

The primary objective was to determine the exact order of all ~3 billion nucleotides (A, T, G, C) in human DNA. This would create a reference sequence—a standard "text" that scientists could use to locate genes, compare individual genomes, and understand genetic variation.

Think of it like creating the first complete, accurate map of an unexplored continent. Once you have that map, every future explorer can use it to navigate, mark new discoveries, and share their findings with others.

4.1.2 2. Identify All Human Genes

Scientists wanted to catalog every gene in the human genome. At the time, estimates ranged wildly—some predicted 50,000 genes, others guessed as many as 100,000. The thinking was simple: humans are complex organisms, so we must have many more genes than simpler creatures like worms or flies.

As we'll see, this assumption turned out to be spectacularly wrong—and that surprise taught us something profound about how biology works.

4.1.3 3. Develop New Technologies

In 1990, DNA sequencing was slow, expensive, and labor-intensive. Sequencing even a small gene could take months. Sequencing 3 billion base pairs with existing technology would have taken centuries and cost billions of dollars.

The HGP needed to drive innovation in sequencing technology, robotics, and computational methods for assembling and analyzing huge amounts of data. This technological push was just as important as the scientific goals.

4.1.4 4. Map Genetic Variation

Beyond the reference sequence, the HGP aimed to create genetic and physical maps showing landmarks along chromosomes—like mile markers on a highway. These landmarks, particularly **single nucleotide polymorphisms (SNPs)**—single-letter DNA differences between individuals—would help scientists link specific genes to diseases or traits.

This mapping laid the groundwork for understanding how humans differ genetically, which became crucial for personalized medicine.

4.2 How Did They Do It? The HGP Journey

The HGP was a massive international collaboration involving sequencing centers in the United States, United Kingdom, France, Germany, Japan, China, and other countries. It was originally planned to take 15 years (1990–2005), but technological advances and intense collaboration—plus some competitive pressure—accelerated the timeline significantly.

4.2.1 Key Milestones

Early Years (1990–1998): Learning by Doing

The project didn't jump straight into sequencing human DNA. Instead, scientists first practiced on simpler organisms. They sequenced the genomes of bacteria, yeast, and the roundworm *C. elegans* (completed in 1998)—creatures with much smaller genomes. These projects served as training exercises, helping researchers refine their methods and develop the computational tools needed to assemble millions of short DNA fragments into complete genome sequences.

First Success (1999): Chromosome 22

In 1999, researchers achieved their first major milestone: the complete sequence of human chromosome 22, one of the smallest chromosomes. This proved that sequencing an entire human chromosome was feasible and provided a test case for the methods that would be used on the rest of the genome.

The Race (2000): Two Competing Efforts

By 2000, the public HGP faced an unexpected competitor: **Celera Genomics**, a private company led by Craig Venter. Celera used a different, faster approach called "shotgun sequencing" and claimed it could complete the genome more quickly and cheaply than the public consortium.

This created both rivalry and synergy. In June 2000, both groups announced they had completed "draft" sequences covering roughly 90% of the genome. President Bill Clinton, along with British Prime Minister Tony Blair, held a joint press conference celebrating this achievement as a historic moment for humanity.

Publication (2001): Two Versions

In February 2001, both the HGP and Celera published their draft sequences—the public consortium in *Nature* and Celera in *Science*. These weren't finished sequences; they contained gaps and errors, particularly

in repetitive regions. But they represented enormous progress: for the first time, scientists had a working draft of the entire human genome.

Completion (2003): The "Finished" Genome

In April 2003—exactly 50 years after Watson and Crick published the structure of DNA—the HGP announced completion of the human genome sequence. About 99% of the gene-containing portions (called **euchromatic regions**) had been sequenced to high accuracy, with an error rate of less than 1 mistake per 100,000 bases.

This "finished" genome, later refined and released as **GRCh37** (also called hg19) and then **GRCh38** (hg38), became the reference that researchers worldwide would use for the next two decades.

4.3 The Big Surprise: Only 20,000 Genes?

One of the HGP's most striking discoveries was how few genes humans actually have. The final count came to approximately **20,000–25,000 protein-coding genes**—fewer than many scientists expected, and not dramatically more than much simpler organisms:

- Humans: ~20,000 genes
- Roundworms (*C. elegans*): ~20,000 genes

- Fruit flies: ~14,000 genes
- Rice plants: ~40,000 genes

Wait—rice has more genes than humans? How can that be?

This discovery forced scientists to rethink what makes organisms complex. It's not just the number of genes, but how those genes are regulated, combined, and expressed. Human genes can produce multiple proteins through processes like **alternative splicing** (where one gene can be read in different ways to make different proteins). Gene regulation—when and where genes are turned on or off—is also far more sophisticated in humans than in simpler organisms.

The HGP also revealed that only about 1.5% of the human genome actually codes for proteins. The remaining ~98.5% includes regulatory sequences, structural elements, and repetitive DNA. Much of this "non-coding" DNA, once dismissed as "junk DNA," is now recognized as functionally important for controlling gene expression.

4.4 How the HGP Changed Genetics

The Human Genome Project didn't just produce a reference sequence—it fundamentally transformed how biological research is conducted. As geneticist Richard Gibbs observed, "Its success should be measured by how this project transformed the rules of research, the way of practicing biological discovery, and the ubiquitous digitization of biological science."

Let's examine these transformations:

4.4.1 From Gene-by-Gene to Genome-Wide Analysis

Before the HGP, genetics meant studying one gene at a time. Researchers would spend years isolating and characterizing a single gene. After the HGP, scientists could look at all genes simultaneously. Want to know which genes are active in cancer cells versus normal cells? Compare the expression of all 20,000 genes at once. This shift from individual genes to whole-genome analysis opened entirely new research approaches.

4.4.2 A Common Reference for Comparing Genomes

The reference genome provided a standard coordinate system—like longitude and latitude for the genome. Now when scientists discover a genetic variant associated with a disease, they can specify its exact location: "chromosome 7, position 117,559,593." Other researchers can immediately look up that location, see what genes are nearby, and replicate the finding.

This standardization accelerated disease gene discovery. Before the HGP, finding a disease gene could take a decade or more. After the HGP, researchers could conduct **genome-wide association studies (GWAS)**, comparing DNA from thousands of people with and without a disease to identify genetic risk factors. This has led to discoveries of genetic contributions to diabetes, heart disease, schizophrenia, and countless other conditions.

4.4.3 Technological Revolution

The HGP drove massive improvements in sequencing technology. In 1990, sequencing a single human genome would have cost billions of dollars. By 2003, the cost had dropped to about $300 million. Today, thanks to technologies developed after the HGP, sequencing a genome costs less than $1,000 and takes just hours.

This cost reduction, far outpacing Moore's Law (which describes the exponential improvement in computer chips), made genomics practical not just for research but also for clinical medicine. Genome sequencing is now used to diagnose rare genetic diseases, guide cancer treatment, and predict drug responses.

4.4.4 Open Data Sharing: The Bermuda Principles

Perhaps the HGP's most important legacy was establishing a culture of open data sharing in genomics. In 1996, project leaders met in Bermuda and agreed that all sequence data would be released publicly within 24 hours of generation—no patents, no paywalls, no waiting for publication.

This was revolutionary. In most scientific fields, researchers guard their data until they publish papers. The **Bermuda Principles** established a different model: the genome belongs to everyone, and its sequence should be freely available to accelerate discoveries.

This open-access approach became a hallmark of modern genomics, influencing projects like the 1000 Genomes Project, the Cancer Genome Atlas, and the Telomere-to-Telomere consortium that produced T2T-CHM13.

4.5 What the HGP Couldn't Do: The 8% That Remained

Despite its achievements, the HGP left significant gaps. About 8% of the genome—roughly 240 million base pairs—remained unsequenced or poorly sequenced. These gaps occurred mainly in highly repetitive regions:

- **Centromeres**: The repetitive DNA at chromosome centers
- **Telomeres**: The repetitive caps at chromosome ends

- **Ribosomal DNA arrays**: Hundreds of copies of genes needed for making ribosomes
- **Segmental duplications**: Large blocks of nearly identical DNA

These regions aren't genetic "junk." Centromeres are essential for cell division. Variations in ribosomal DNA copy number may affect protein synthesis and health. Segmental duplications are hotspots for genetic variation and disease-causing mutations.

The HGP's sequencing technology—which relied on reading short DNA fragments—couldn't reliably distinguish between nearly identical repeats. It's like trying to assemble a book where several pages are identical—you know those pages exist, but you can't tell where each one goes.

These limitations were addressed two decades later by the Telomere-to-Telomere project, which used long-read sequencing and the CHM13 cell line to finally complete these challenging regions, producing the first truly gapless human genome in 2022.

4.6 The Foundation for Everything That Followed

The Human Genome Project created the infrastructure for modern genetics. Every genetic discovery since 2003—from identifying disease genes to understanding human evolution to developing personalized cancer treatments—builds on the reference genome the HGP produced.

It transformed genetics from a science focused on individual genes into a data-driven discipline where researchers routinely work with billions of data points. It established that genomic data should be a public resource, freely shared to benefit all of humanity.

And perhaps most importantly, it showed us that humans are far more complex than the simple sum of our genes. Those ~20,000 genes, working together in intricate regulatory networks, create all the diversity and complexity of human life.

The HGP asked: What does our genome say? The projects that followed—like T2T-CHM13 and the Human Pangenome—are asking: How does it vary? What does that variation mean? These questions continue to drive genomics forward, building on the foundation the HGP laid more than two decades ago.

4.7 Timeline at a Glance

Year	Milestone
1990	Human Genome Project officially launches with 15-year timeline
1995–1998	Practice sequencing: model organism genomes completed (E. coli, yeast, C. elegans)
1999	First complete human chromosome sequenced (chromosome 22)
2000	Draft genome announced by both public HGP and private Celera Genomics (~90% coverage)
2001	Draft sequences published in Nature and Science
2003	"Finished" sequence released: 99% of gene-containing regions sequenced with <1/100,000 error rate; ~20,000–25,000 genes identified
2003–2013	Reference genome refined (GRCh37/hg19, then GRCh38/hg38), but ~8% gaps remain
2022	T2T-CHM13 completes the remaining gaps, producing first truly gapless human genome

5 Chapter 2. The Telomere-to-Telomere Genome Project

When the Human Genome Project (HGP) announced its completion in 2003, it was celebrated as a historic milestone. But here's something surprising: the "complete" human genome wasn't actually complete.

About 8% of our genome remained unsequenced—not because scientists didn't want to sequence it, but because the technology at the time simply couldn't handle certain regions.

Think of it like trying to assemble a jigsaw puzzle where some pieces are nearly identical. Early sequencing technologies read short fragments of DNA (around 100–200 base pairs), which worked well for most of the genome. But in highly repetitive regions—where the same sequence appears over and over—these short reads couldn't tell one repeat from another. It's like having dozens of identical puzzle pieces and not knowing where each one goes.

The missing regions weren't random leftovers. They included some of the most functionally important parts of our chromosomes: **centromeres** (the central structures that help separate chromosomes during cell division), **telomeres** (the protective caps at chromosome ends), and **ribosomal DNA arrays** (genes that make the machinery for protein production). These gaps limited our understanding of chromosome stability, cell division, and even how genetic diseases develop.

In 2022, the **Telomere-to-Telomere (T2T) Consortium** changed this. They produced the first truly complete human genome sequence, called **T2T-CHM13**, covering all 3.055 billion base pairs from one end (telomere) to the other end (telomere) of every chromosome—hence the name. This achievement added about 200 million base pairs of new sequence and revealed 1,956 previously unknown gene predictions, 99 of which are predicted to be protein coding (Nurk et al. 2022, Science).

5.1 What Made T2T-CHM13 Possible?

Two major advances enabled the T2T project to succeed where earlier efforts had failed: **new sequencing technologies** and **a special cell line**.

5.1.1 The Technologies

The T2T Consortium combined several cutting-edge technologies:

- **PacBio HiFi sequencing**: This technology reads much longer DNA fragments—around 20,000 base pairs (20 kb)—with high accuracy. These longer reads can span multiple repeats, making it possible to distinguish one repeat from another.
- **Oxford Nanopore ultra-long-read sequencing**: These reads can exceed 100,000 base pairs, allowing scientists to read through even longer repetitive regions in a single pass.
- **Supporting methods**: Additional techniques like **Illumina short-read sequencing** (for error correction), **Hi-C** (which maps how DNA folds in 3D space), **Bionano optical mapping** (which creates a physical map of the genome), and **Strand-seq** (which tracks which DNA strands came from each parent) all helped ensure accuracy.

5.1.2 String Graphs: A Tool for Assembly

The assembly process used a **high-resolution string graph**—a powerful computational approach that helped resolve the most complex repetitive regions (see Figure 2 in Nurk et al., 2022).

Figure: CHM13 String Graph. This diagram shows how the T2T-CHM13 genome was assembled using a string graph approach. Each line represents a DNA sequence, and connections show where sequences overlap. The tangled regions reveal highly repetitive areas like the ribosomal DNA arrays and centromeric satellites. Source: Nurk, S. et al. (2021). The complete sequence of a human genome. bioRxiv. https://doi.org/10.1101/2021.05.26.445798. License: CC-BY 4.0.

Unlike traditional linear reference genomes that represent a single sequence, a string graph represents DNA as a network where nodes are DNA sequences and edges show how they connect. Think of it like a

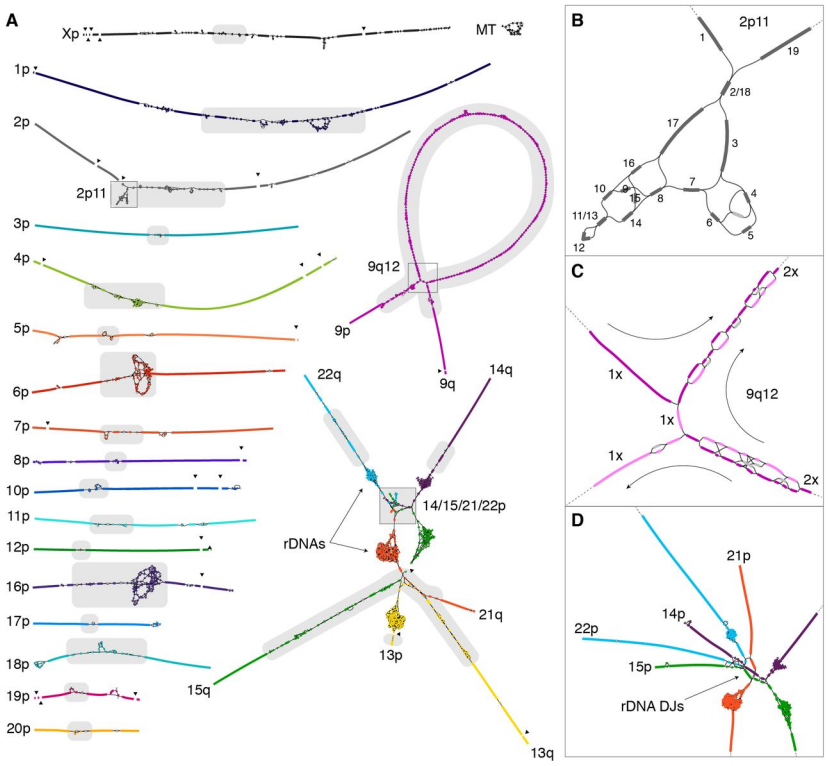


Figure 3: Nurk et al. 2021, bioRxiv – Figure 2

subway map showing multiple routes between stations—the graph displays all possible connections, then long reads help scientists determine the correct path through repetitive regions.

This approach was critical for handling repeats: when you have many identical DNA sequences (like in centromeres), a linear approach gets confused about which piece connects where. The graph shows all possibilities, and ultra-long reads spanning the repeats reveal the correct connections. The final T2T-CHM13 assembly achieved an exceptionally high accuracy with an error rate of just **1 mistake per 10 million bases**.

While T2T-CHM13 is presented as a linear sequence (because it represents one individual), the future of genomics is moving toward **pangenome references** that better capture human genetic diversity across populations.

5.2 The Special Cell Line: CHM13

The T2T project used DNA from a unique cell line called CHM13, which came from a **complete hydatidiform mole**—a rare type of tissue that forms when an egg without genetic material is fertilized by a sperm that duplicates its own genome. This means CHM13 has two identical copies of each chromosome from the father and none from the mother.

Why does this matter? In a typical human genome, you inherit one set of chromosomes from each parent, making assembly more complex because you have to distinguish between similar but not identical sequences. CHM13's duplicated genome simplified assembly—it's like solving a puzzle where you know matching pieces should be identical. However, CHM13 lacked a Y chromosome, which remained the last piece of the puzzle.

5.3 The First Complete Human Chromosome: X (2020)

While the full T2T-CHM13 genome was announced in 2022, the project achieved its first major milestone in **2020 by completing the human X chromosome**—the first chromosome to be assembled telomere-to-telomere with no gaps (Miga et al. 2020, Nature).

5.3.1 Why Start with X?

The research team chose the X chromosome as their first target for several strategic reasons: - **High initial continuity**: In the whole-genome assembly, the X chromosome was broken in only **three places**—the centromere and two large segmental duplications - **Well-characterized centromere**: The X centromere (DXZ1) had been extensively studied, providing validation benchmarks - **Medical importance**: The X chromosome is disproportionately involved in genetic diseases - **Technical advantage**: Using the CHM13 cell line (46,XX), they had two copies of X but no Y, making X more accessible

Figure: Initial CHM13 X Chromosome Assembly. (See Figure 1b) The X chromosome was initially broken at three locations: the centromere (artificially collapsed in the assembly), a 120-kb segmental duplication (DMRTC1B), and a 134-kb segmental duplication with a paralogue on chromosome 2. Black bars indicate gaps in the GRCh38 reference, and red bars show known segmental duplications. Source: Miga, K.H. et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585, 79-84. https://doi.org/10.1038/s41586-020-2547-7. License: CC-BY 4.0.

5.3.2 Finishing the Breaks

The two segmental duplications were resolved by finding **ultra-long reads that completely spanned the repeats** and were uniquely anchored on both sides. This allowed confident placement in the assembly.

The centromere presented a greater challenge—it's a **3.1 Mb array of highly repetitive alpha satellite DNA** where standard polishing methods failed. Initial attempts to polish the centromeric assembly actually **decreased quality** because reads were incorrectly placed due to the extreme repetitiveness.

5.3.3 The Innovation: Marker-Assisted Polishing

The team developed a novel **marker-assisted polishing** strategy to finish the large repetitive regions:

1. Catalogue unique markers: They identified short (21 bp), unique sequences present only once in the genome. Remarkably, even within the DXZ1 centromeric array, there was enough variation between repeat copies to create unique markers at semi-regular intervals.

```
Genome-wide: average spacing = 66 bp between markers
DXZ1 array:  average spacing = 2.3 kb between markers
Longest gap in DXZ1: 42 kb
```

2. Anchor reads precisely: These markers guided the correct placement of long reads during polishing, preventing the quality degradation that occurred with standard methods.

3. Iterative polishing: Multiple rounds of polishing were performed with Nanopore, then PacBio, then Illumina data, each improving accuracy.

Figure: Marker-Assisted Polishing Improves Assembly Quality. (See Figure 1d) Example from the GAGE locus showing coverage depth before (top) and after (bottom) marker-assisted polishing. Black dots indicate primary allele coverage, red dots show secondary alleles. The uniform coverage and elimination of secondary alleles after polishing demonstrates the dramatic quality improvement achieved by this method. Source: Miga, K.H. et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585, 79-84. https://doi.org/10.1038/s41586-020-2547-7. License: CC-BY 4.0.

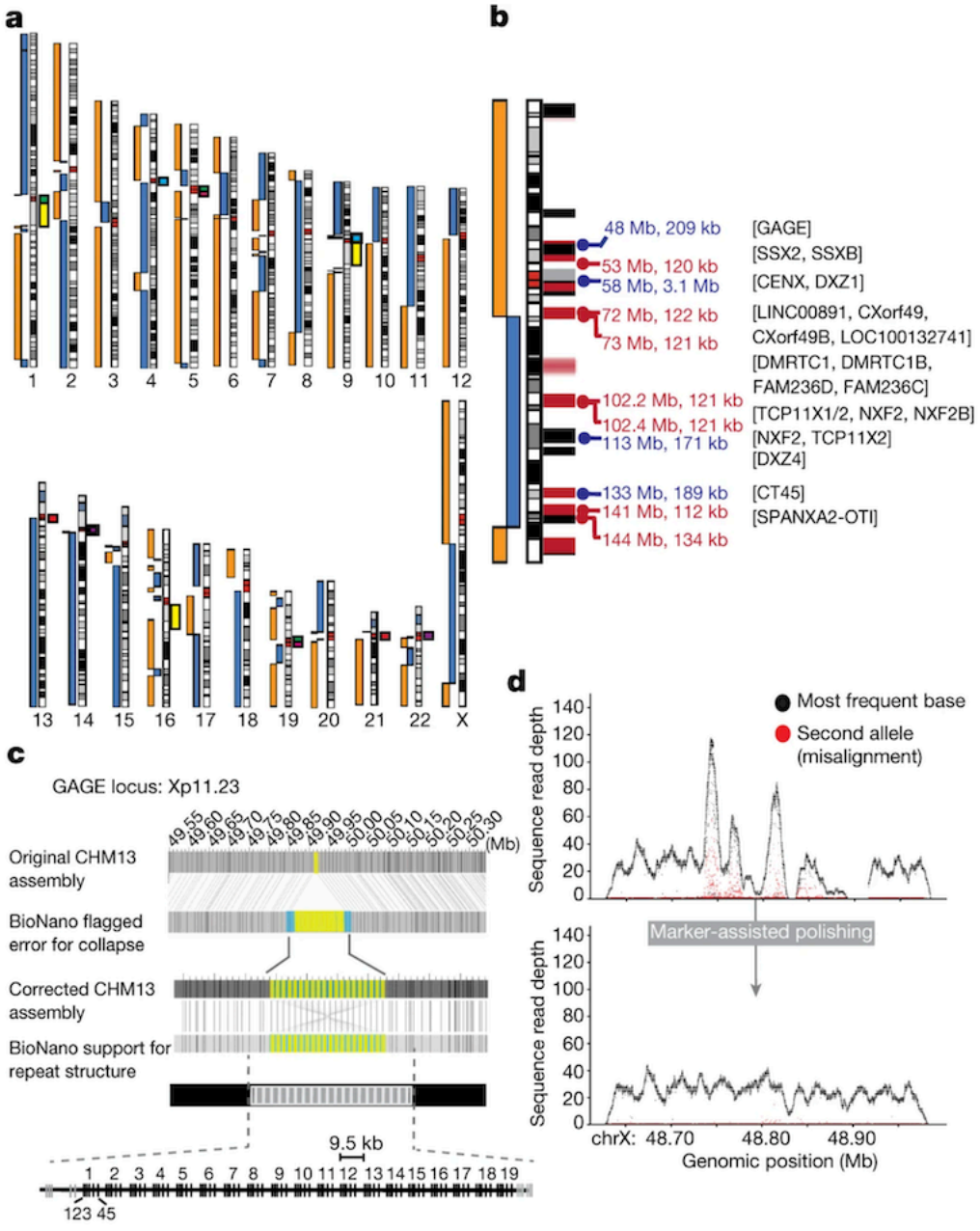


Figure 4: Miga et al. 2020, Nature – Figure 1b

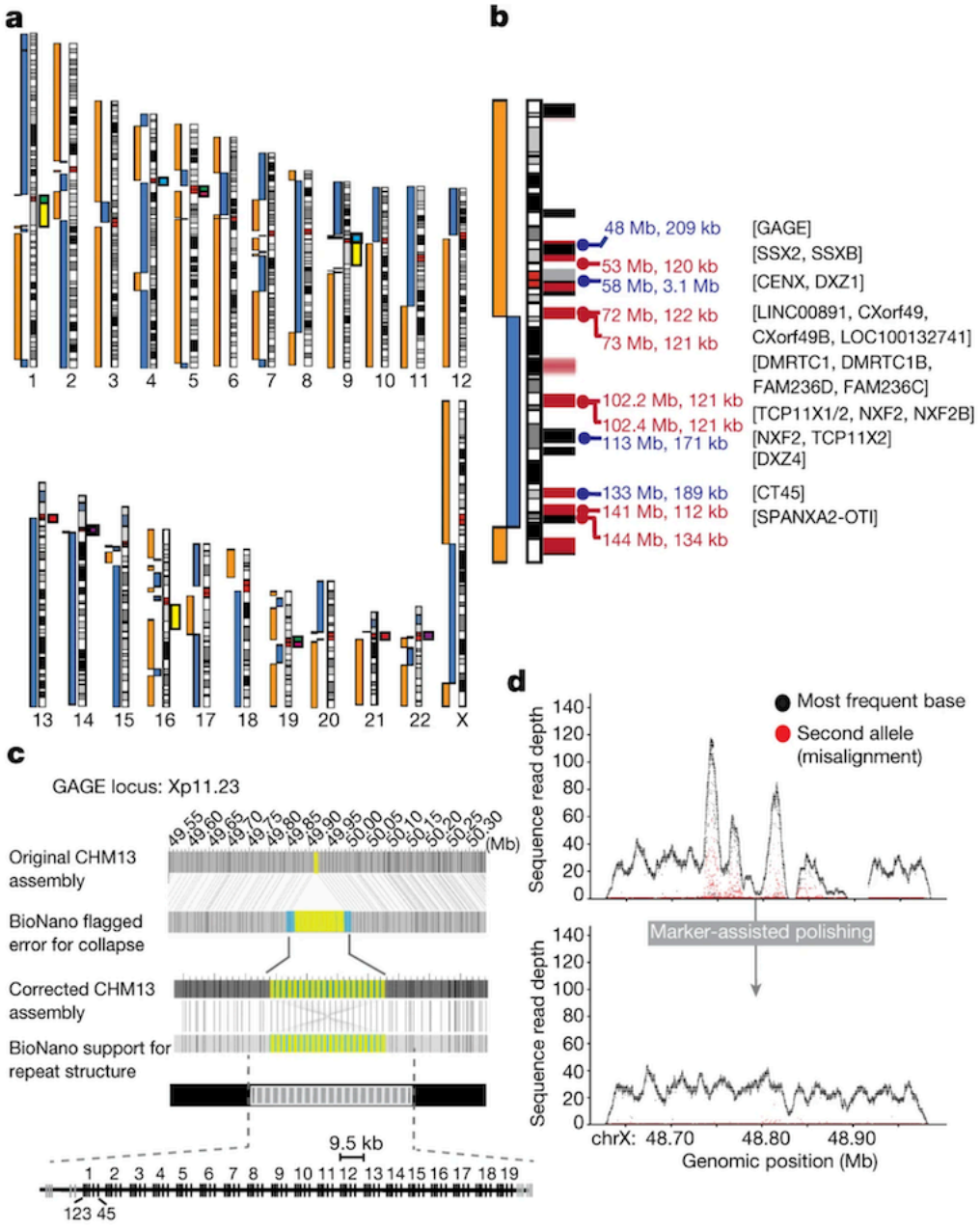


Figure 5: Miga et al. 2020, Nature – Figure 1d

5.3.4 Validating the X Centromere

The 3.1 Mb DXZ1 centromeric array consists of approximately 1,408 copies of a ~2,057–bp "higher–order repeat" (HOR) unit. This canonical repeat is made of 12 divergent alpha satellite monomers (each ~171 bp) arranged in a specific order. The team validated this structure through multiple independent methods:

Experimental validation: – **Pulsed–field gel electrophoresis (PFGE)**: Confirmed array size of 2.87 Mb ± 0.16 – **Droplet digital PCR (ddPCR)**: Counted 1,408 ± 41 copies of the 2,057–bp repeat unit – **Optical mapping**: Confirmed the restriction enzyme patterns matched predictions

Sequence–based validation: – **PacBio HiFi reads**: Showed uniform coverage across the entire array – **Ultra–long Nanopore reads**: Many reads spanned 〉100 kb of the centromere – **Completeness check**: All HiFi reads matching DXZ1 were explained by the assembly

Figure: Complete Structure of the 3.1 Mb X Centromere. The DXZ1 array consists of approximately 1,408 copies of a ~2–kb higher–order repeat unit. (a) Predicted restriction map showing the array structure. (b) Experimental PFGE Southern blots matching the in silico prediction. (c) ddPCR copy number estimates across multiple cell lines. (d) Catalogue of 33 structural variants identified within the array. Source: Miga, K.H. et al. (2020). Telomere–to–telomere assembly of a complete human X chromosome. Nature, 585, 79–84. https://doi.org/10.1038/s41586–020–2547–7. License: CC–BY 4.0.

The assembly catalogued **33 different structural variants** within the DXZ1 array, including: – 1,475 canonical 12–mer units (98% of all units) – Various altered units (11–mers, 22–mers, 18–mers, etc.) – One 8.1–kb LINE/L1 insertion

This level of detail was impossible with previous technologies and provides insights into centromere evolution and function.

5.3.5 Achievements of the Complete X

The complete X chromosome assembly: – **Closed 29 gaps** from GRCh38, totaling 1,147,861 bp of previous N–bases – Achieved **99.991% accuracy** (estimated from BAC sequences) or **99.995%** (from Illumina data) – Met the original **Bermuda Standards for finished genomic sequences** (〉99.99% accuracy, 〈1 error per 10,000 bases) – Added complete sequences of medically important gene families: – **GAGE** genes: 19 complete + 2 partial copies (9.5 kb each) – **CT45** genes: 6 copies – **CT47** genes: 7 copies – Revealed the complete structure of the **pseudoautosomal regions (PAR1 and PAR2)** – Provided the first complete view of cancer–testis gene families that were previously fragmented

5.3.6 Methylation Discovery

The precisely anchored ultra–long reads enabled **chromosome–wide methylation mapping** at single–base resolution—even across complex repeats that were previously invisible to methylation studies. This revealed several surprising patterns:

- **Hypomethylation in PAR regions** (consistent with their escape from X–inactivation)
- **A 93–kb hypomethylated region within the DXZ1 centromere** (chrX: 59,217,708–59,279,205)—potentially marking the site where kinetochore proteins bind during cell division
- **Distinct methylation bands within the DXZ4 macrosatellite array**—the repeat that partitions X chromosome topology into two large superdomains

Figure: Methylation Patterns Across the Complete X Chromosome. Nanopore sequencing captured methylation patterns across the entire chromosome. (a) Hypomethylation in PAR1, with detailed view showing unmethylated bases (blue) and methylated bases (red). (b) A 93–kb hypomethylated region within the DXZ1 centromere. (c) The DXZ4 macrosatellite array showing a transition from methylated to unmethylated regions. Source: Miga, K.H. et al. (2020). Telomere–to–telomere assembly of a complete human X chromosome. Nature, 585, 79–84. https://doi.org/10.1038/s41586–020–2547–7. License: CC–BY 4.0.

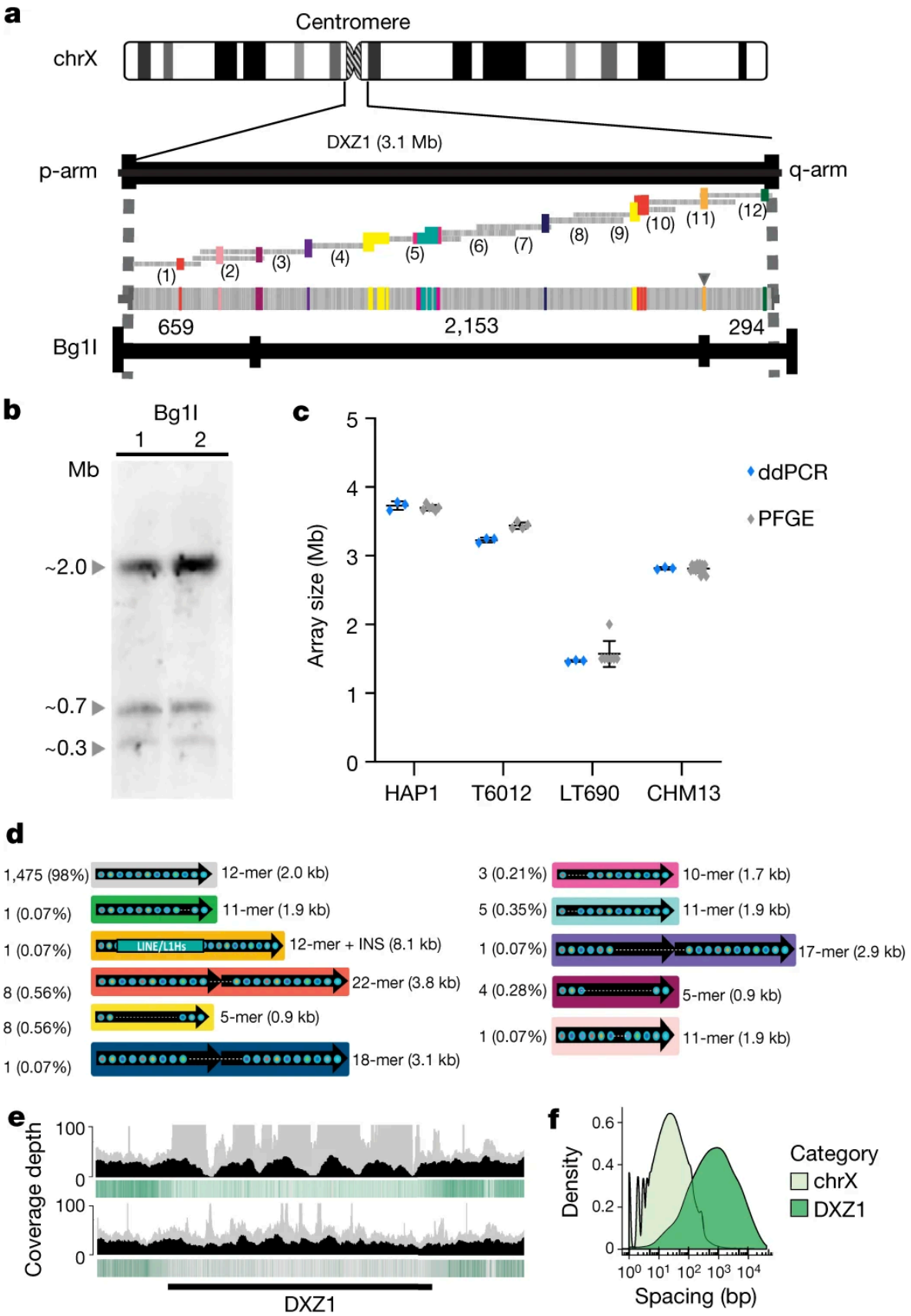


Figure 6: Miga et al. 2020, Nature – Figure 2

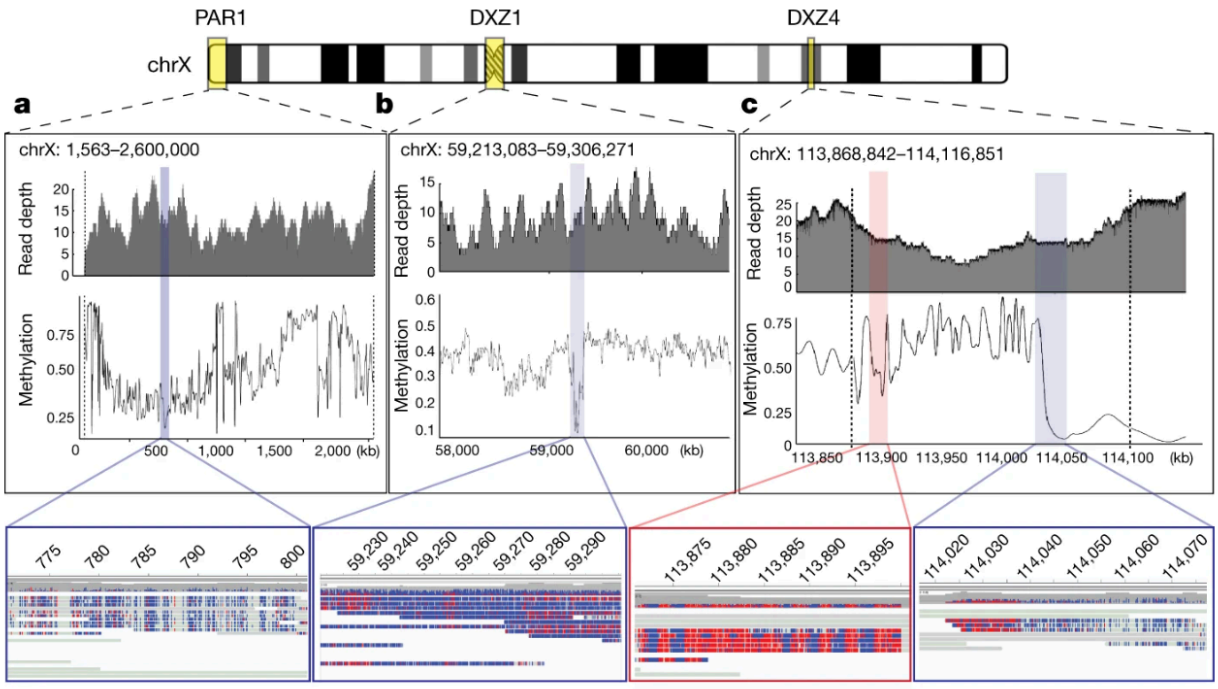


Figure 7: Miga et al. 2020, Nature – Figure 3

The hypomethylated centromeric region was particularly intriguing. To test whether this was unique to the X or a general centromeric feature, the team manually assembled the centromere of chromosome 8 (D8Z2, ~2.02 Mb) and found a **similar hypomethylated region**, suggesting this may be a conserved feature marking functional centromeres.

5.3.7 Impact

This 2020 achievement demonstrated that **completing an entire human chromosome was possible** and established the methods that would be used to complete all other chromosomes:

1. **Ultra-long-read sequencing** can span even megabase-scale repeats
2. **Marker-assisted polishing** is essential for accurate assembly of large tandem repeats
3. **Multiple orthogonal validation methods** (PFGE, ddPCR, optical mapping, HiFi reads) ensure accuracy
4. **Nanopore methylation calling** works even in highly repetitive regions

The complete X chromosome became the foundation for finishing the rest of the human genome.

5.4 What's New in T2T-CHM13? (2022)

Building on the success of the X chromosome, the T2T Consortium completed the remaining chromosomes to produce the full T2T-CHM13 genome. This assembly includes several types of regions that were missing or incomplete in the previous reference genome (GRCh38):

Figure: T2T-CHM13 Assembly Ideogram. (See Figure 1a) This ideogram shows what's new in T2T-CHM13 compared to GRCh38. Red areas highlight newly added regions, including complete centromeres and the short arms of five acrocentric chromosomes (13, 14, 15, 21, and 22). The track at the top shows that CHM13 has primarily European genetic ancestry. Source: Nurk, S. et al. (2021). The complete sequence of a human genome. bioRxiv. https://doi.org/10.1101/2021.05.26.445798. License: CC-BY 4.0.

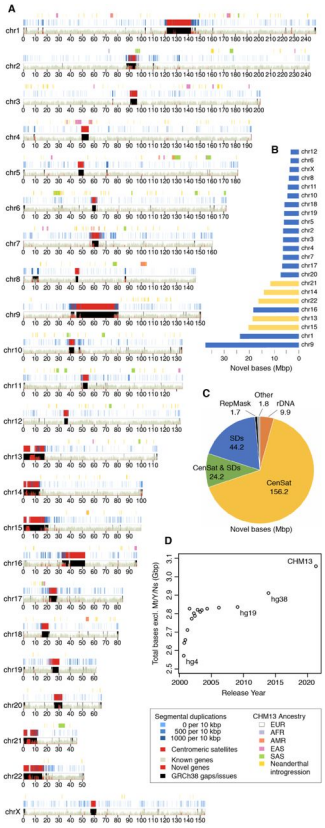


Figure 8: Nurk et al. 2021, bioRxiv − Figure 1

Let's explore the main types of regions that T2T–CHM13 finally completed. Each was challenging to sequence for different reasons, and each plays an important role in how our cells work.

5.4.1 Region 1: Centromeric Satellite Arrays

5.4.1.1 What Are They? Imagine each chromosome as having a "waist"—a pinched middle section. This is the **centromere**, and it's made of highly repetitive DNA called **satellite arrays**. In humans, the most common type is **alpha satellite DNA**, where a short sequence (about 171 base pairs) repeats hundreds or thousands of times in a row.

5.4.1.2 Why Do We Need Them? During cell division, chromosomes need to be pulled apart and distributed equally to daughter cells. The centromere acts as a handle where proteins called **kinetochores** attach. These proteins connect to spindle fibers—molecular ropes that pull chromosomes to opposite ends of the dividing cell.

If centromeres don't work properly, chromosomes can be distributed incorrectly, leading to cells with too many or too few chromosomes (called **aneuploidy**). This can cause serious problems: most aneuploid embryos don't survive, and in cells that do survive, aneuploidy is linked to cancer and genetic disorders like Down syndrome.

5.4.1.3 What Did T2T–CHM13 Reveal? T2T–CHM13 fully sequenced all centromeric regions, showing their complete structure for the first time. Each centromere has unique features: – Sizes range from ~366 kb (chromosome Y) to several megabases – Different chromosomes use different alpha satellite variants – Some centromeres (like those on chromosomes 1, 5, and 19) share similar sequences, suggesting recent exchanges of DNA between them

5.4.2 Region 2: Segmental Duplications

5.4.2.1 What Are They? **Segmental duplications** are large chunks of DNA—often thousands to millions of base pairs long—that appear in multiple places in the genome. These copies are usually 90–99% identical to each other. Think of them as long paragraphs that have been copied and pasted elsewhere in a document, with only minor edits.

5.4.2.2 Why Do We Need Them? Segmental duplications are major drivers of genetic diversity and evolution. Because they're nearly identical, they can accidentally pair up during DNA replication or recombination, creating **structural variants**—duplications, deletions, or rearrangements of DNA.

Some structural variants are beneficial. For example, having extra copies of the **AMY1** gene (which makes an enzyme that digests starch) helps people digest starchy foods better—populations that historically ate more starch tend to have more AMY1 copies.

But structural variants can also cause disease. Certain duplications are linked to **facioscapulohumeral muscular dystrophy (FSHD)**, a muscle-weakening disorder, and many other genetic conditions.

5.4.2.3 What Did T2T–CHM13 Reveal? T2T–CHM13 added substantial amounts of segmental duplication sequence and corrected many errors in GRCh38. The complete assembly revealed that about **6.61% of the genome** consists of segmental duplications (201.93 Mb), compared to only 5.00% that could be properly identified in GRCh38. This includes many medically relevant genes that were previously incomplete or incorrectly assembled.

5.4.3 Region 3: Short Arms of Acrocentric Chromosomes

5.4.3.1 What Are They? Five of our chromosomes—numbers 13, 14, 15, 21, and 22—are called **acrocentric** because their centromeres are located very close to one end. This creates a short arm (called the "p" arm) and a much longer arm (the "q" arm).

These short arms contain clusters of **ribosomal DNA (rDNA)**—genes that encode **ribosomal RNA (rRNA)**, a key component of ribosomes. Ribosomes are the molecular machines that read messenger RNA (mRNA) and build proteins. Each rDNA unit is about 45,000 base pairs long and is repeated dozens to hundreds of times.

5.4.3.2 Why Do We Need Them? Without ribosomes, cells can't make proteins, and without proteins, life stops. The rDNA arrays on acrocentric chromosomes cluster together to form the **nucleolus**—a structure inside the nucleus where ribosomes are assembled.

Interestingly, the number of rDNA copies varies between individuals. CHM13 has about 400 copies, but other people might have more or fewer. Scientists are still investigating whether this variation affects how efficiently cells make proteins, and whether it influences traits like growth, aging, or susceptibility to diseases like cancer.

5.4.3.3 What Did T2T-CHM13 Reveal? T2T-CHM13 provided the first complete view of the short arms of all five acrocentric chromosomes, totaling **66.1 Mb** of new sequence. These regions follow a similar structure: from telomere to centromere, they contain distal repeat arrays, the rDNA array, and proximal repeat arrays including various satellite sequences.

Remarkably, these short arms show about **98.7% identity** to each other, suggesting frequent exchange of DNA between them. This high similarity is probably because these chromosomes cluster together in the nucleolus during interphase.

5.5 Completing the Y Chromosome (2023)

Although T2T-CHM13 was a major achievement, it had one limitation: CHM13 is a 46,XX cell line, meaning it has no Y chromosome. The Y chromosome had been notoriously difficult to sequence because of its complex repeat structure, including long palindromes (sequences that read the same forwards and backwards), tandem repeats, and segmental duplications. In fact, more than half of the Y chromosome was missing from GRCh38.

In 2023, the T2T Consortium completed this final piece of the puzzle by sequencing the Y chromosome from a different genome, HG002, which is commonly used for benchmarking (Rhie et al. 2023, Nature). The resulting assembly, called **T2T-Y**, is 62,460,029 base pairs long with no gaps—adding over 30 million base pairs of sequence compared to GRCh38-Y.

Figure: Complete Structure of the Human Y Chromosome. (See Figure 1a) This comprehensive view shows alignment of T2T-Y to GRCh38-Y, locations of protein-coding genes with clusters of ampliconic genes highlighted, organization of palindromic sequences and inverted repeats, and detailed structure of centromeric and satellite DNA regions. The bottom panel shows a dotplot revealing the highly repetitive nature of this chromosome. Source: Rhie, A. et al. (2022). The complete sequence of a human Y chromosome. bioRxiv. https://doi.org/10.1101/2022.12.01.518724. License: CC0 (US Government work).

5.5.1 What Makes the Y Chromosome Special?

The Y chromosome carries genes critical for male development and fertility:

1. **SRY**: The master gene that determines male sex

2. **Ampliconic genes**: These genes exist in multiple copies and are important for sperm production: – **TSPY**: The largest gene family on the Y, with **45 protein-coding copies** in HG002 (compared to only 7 in GRCh38-Y) – **DAZ**: Four copies involved in sperm production – **RBMY**: 34 copies, also critical for spermatogenesis

3. **AZF regions** (Azoospermia Factors): Three regions (AZFa, AZFb, AZFc) where deletions can cause male infertility

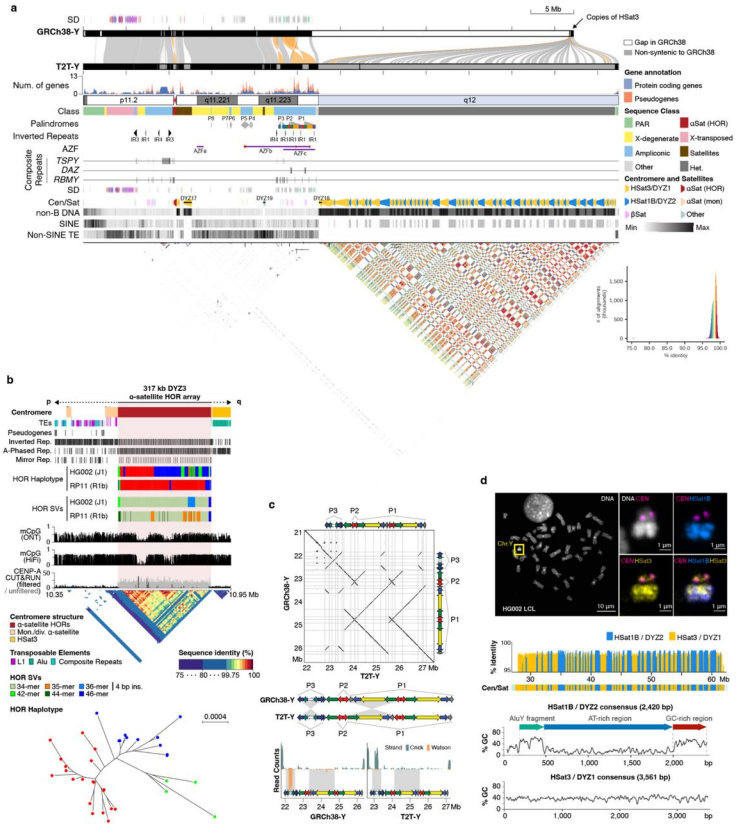


Figure 9: Rhie et al. 2022, bioRxiv – Figure 1a

5.5.2 The Y Centromere

Like other chromosomes, the Y has a centromere, but T2T-Y revealed unique features:

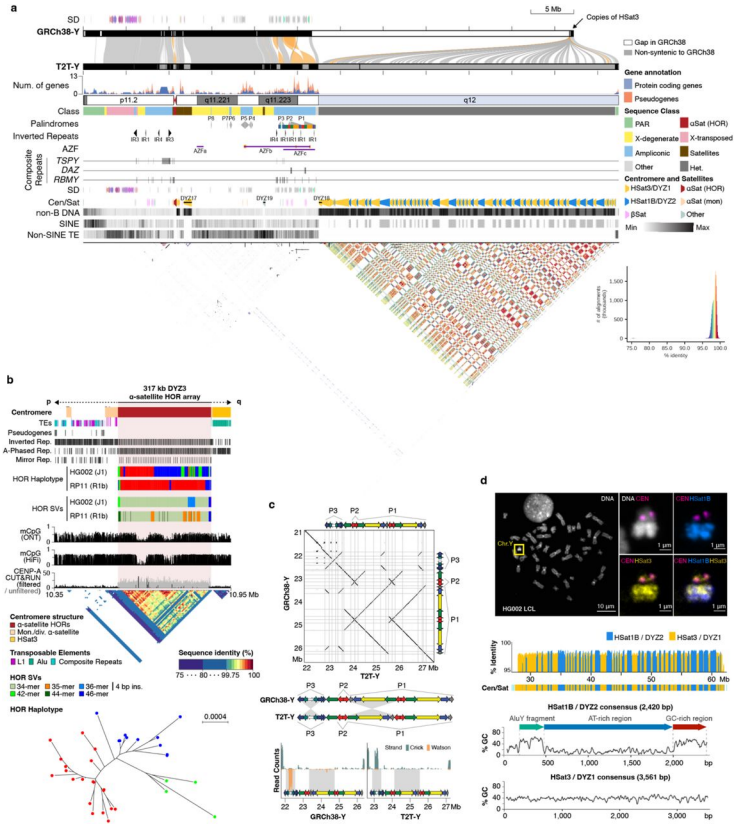


Figure 10: Rhie et al. 2022, bioRxiv – Figure 1b

Figure: Structure of the T2T-Y Centromere. (See Figure 1b) The Y centromere spans 366 kb and consists of the DYZ3 alpha satellite array with three different variants, no transposable elements within the main array, two regions of hypomethylation where CENP-A proteins bind, and a dotplot showing that the repeat units are 99.5–100% identical, demonstrating extreme homogeneity. Source: Rhie, A. et al. (2022). The complete sequence of a human Y chromosome. bioRxiv. https://doi.org/10.1101/2022.12.01.518724. License: CC0 (US Government work).

The centromere spans 366 kb and consists of highly similar alpha satellite repeats. Interestingly, the T2T-Y centromere shows **two distinct hypomethylated regions** where kinetochore proteins bind—a pattern also seen in some other chromosomes.

5.5.3 The Heterochromatic Region (Yq12)

Perhaps the most mysterious part of the Y chromosome is Yq12, the large heterochromatic region on the long arm that was almost entirely missing from GRCh38 (represented as a single 30+ megabase gap). T2T-Y finally revealed what's inside: over **30 million base pairs** of alternating blocks of two satellite families:

- **DYZ1 (HSat3)**: Blocks ranging from 80 kb to 1,600 kb (median 370 kb)
- **DYZ2 (HSat1B)**: Blocks ranging from 20 kb to 1,200 kb (median 230 kb)

Figure: The Mysterious Yq12 Region Revealed. (See Figure 1d) Fluorescence microscopy images show the Y chromosome with different colored probes. A map showing 86 large blocks alternating between DYZ2 and DYZ1, with nearly all repeat units being over 98% identical to consensus. Sequence composition

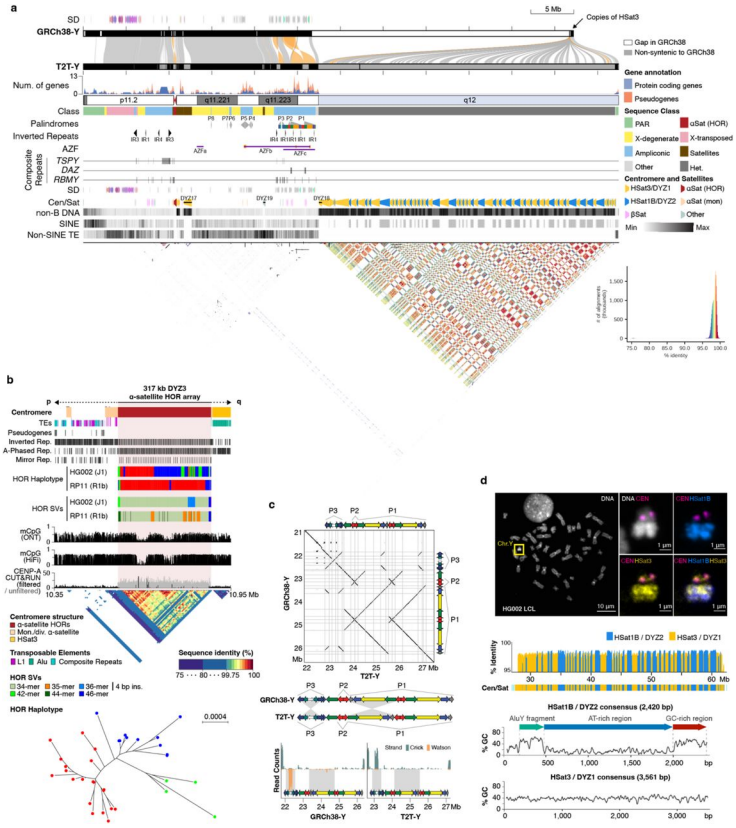


Figure 11: Rhie et al. 2022, bioRxiv − Figure 1d

shows DYZ2 contains an ancient AluY fragment. A phylogenetic tree shows that AluY fragments in HSat1B cluster together, suggesting this satellite family originated on the Y. Source: Rhie, A. et al. (2022). The complete sequence of a human Y chromosome. bioRxiv. https://doi.org/10.1101/2022.12.01.518724. License: CC0 (US Government work).

These satellite blocks show evidence of recent duplication events—some duplications span up to 5 megabases and include multiple DYZ1 and DYZ2 blocks. Interestingly, **HSat1B is almost unique to the Y chromosome** and the short arms of acrocentric chromosomes, while HSat3 is found on many chromosomes.

5.6 Quick Reference: Complex Genomic Regions

Region	What It Is	Why It Matters	Why It Was Hard to Sequence	Size in T2T
Centromeric Satellite Arrays	Repetitive DNA (171–bp alpha satellite repeats) at the chromosome's central "waist"	Acts as a handle for proteins to pull chromosomes apart during cell division; errors can cause aneuploidy, cancer, or developmental disorders	Short reads couldn't distinguish between identical repeats; long reads (20–100+ kb) finally resolved them	Varies by chromosome; Y centromere is 366 kb; X centromere is 3.1 Mb
Segmental Duplications	Large DNA chunks (thousands to millions of bp) copied across the genome with 90–99% similarity	Drive genetic diversity through structural variants; can affect traits (e.g., starch digestion) or cause disease (e.g., muscular dystrophy)	Near–identical sequences confused short–read technology; long reads distinguished them and revealed true variation	6.61% of genome (201.93 Mb) vs 5.00% in GRCh38
Short Arms of Acrocentric Chromosomes	Repetitive rDNA arrays (45,000–bp units repeated dozens to hundreds of times) on chromosomes 13, 14, 15, 21, 22	Encode ribosomal RNA for ribosomes; essential for protein synthesis; copy number varies between people (~400 in CHM13)	Repetitive arrays were indistinguishable with short reads; long reads and specialized assembly algorithms mapped complete copies	Total 66.1 Mb across all five chromosomes
Y Chromosome Heterochromatin (Yq12)	Alternating blocks of HSat1B/DYZ2 and HSat3/DYZ1 satellite repeats	Unknown function, but shows recent structural rearrangements; highly variable between individuals	Extremely long tandem repeats spanning 30+ Mb were impossible to sequence with short reads	Over 30 Mb (was a single gap in GRCh38)

5.7 Summary: Why These Regions Matter

Before the T2T project, these regions—centromeric satellite arrays, segmental duplications, acrocentric short arms, and the Y chromosome—were often called genomic "dark matter" because they remained largely invisible to sequencing technologies. They left massive gaps in earlier reference genomes like GRCh37 (hg19) and GRCh38 (hg38).

By using long-read sequencing technologies and innovative computational methods like marker-assisted polishing, the T2T Consortium finally illuminated these regions:

- **T2T X chromosome (2020)**: First complete human chromosome, established key methods
- **T2T-CHM13 (2022)**: Completed 22 autosomes and chromosome X, adding about 200 million base pairs
- **T2T-Y (2023)**: Completed the Y chromosome, adding over 30 million base pairs

Together, these assemblies were combined to create **T2T-CHM13v2.0** (also called T2T-CHM13+Y), providing the first truly complete sequence of a human genome—all 24 chromosomes from telomere to telomere with no gaps.

These advances improve our understanding of: - **Chromosome function**: How centromeres ensure accurate cell division - **Genetic diversity**: How segmental duplications and structural variants contribute to differences between individuals - **Cellular machinery**: How rDNA arrays support protein synthesis and how variation in these arrays might affect health - **Male fertility**: How ampliconic genes on the Y chromosome function and how rearrangements can cause infertility - **Evolution**: How satellite sequences spread between chromosomes and how the Y chromosome has changed over time - **Epigenetics**: How methylation patterns organize across centromeres and other complex regions

These insights are already improving **variant calling** (identifying genetic differences) in large-scale studies, supporting **precision medicine** by identifying disease-causing mutations, and helping us understand **male infertility** and other Y-linked conditions.

5.8 Looking Forward

The T2T assemblies represent major advances, but the work continues. Because these genomes represent specific individuals (CHM13 and HG002), they don't capture the full spectrum of human genetic diversity. That's where projects like the **Human Pangenome Reference Consortium (HPRC)** come in—by assembling genomes from hundreds of individuals from different populations, researchers are building references that reflect the true diversity of humanity.

The complete human genome isn't an ending—it's a new beginning. With these complete references, we can now: - **Detect variants we previously missed**: Many structural variants in centromeres and segmental duplications were invisible before - **Study chromosome function**: Complete centromeres allow us to understand how they work and why they sometimes fail - **Improve clinical diagnostics**: Complete gene sequences help identify disease-causing mutations more accurately - **Understand human diversity**: Comparing complete genomes from different populations reveals how humans have adapted to different environments

As sequencing technologies continue to improve and become more affordable, we'll be able to study these complex regions in even more detail, uncovering new insights into human health, evolution, and what makes us human.

6 Chapter 3. The CHM13 Cell Line

To understand why the T2T project used the CHM13 cell line, we need to understand a fundamental challenge in genome sequencing: **heterozygosity**.

In a typical human genome, you inherit one set of chromosomes from your mother and another from your father. These two sets are similar but not identical—they differ at millions of positions, creating what we call **heterozygous variants**. For most of the genome, this isn't a problem. But in highly repetitive regions like centromeres or ribosomal DNA arrays, these small differences create a serious headache for genome assembly.

Here's why: when you sequence DNA, you break it into millions of fragments and then use computers to reassemble them, like putting together a jigsaw puzzle. If you have two slightly different copies of a repetitive region—one from mom, one from dad—the computer can't always tell which fragments belong to which copy. Should this read go with the maternal chromosome or the paternal one? In repetitive regions where the same sequence appears dozens or hundreds of times, this ambiguity makes accurate assembly nearly impossible.

The earlier human reference genome, GRCh38, was constructed from **bacterial artificial chromosomes (BACs)**—large pieces of DNA (100–200 kb) cloned into bacteria and then sequenced using Sanger technology. These BACs were opportunistically assembled from multiple individuals, creating a **mosaic of haplotypes** (different genetic variants from different people). This approach resulted in about **151 million base pairs of unknown sequence (gaps)**, mostly in repetitive regions where the heterozygous variants and structural differences between individuals made assembly too complex with available technology.

The T2T Consortium needed a different approach. They needed DNA where both chromosome copies were identical—eliminating the confusion caused by heterozygous variants. That's where CHM13 comes in.

6.1 What Is a Complete Hydatidiform Mole?

The CHM13 cell line comes from a rare biological phenomenon called a **complete hydatidiform mole (CHM)**. To understand what this is, let's first review normal fertilization.

6.1.1 Normal Fertilization vs CHM Formation

Here's how normal fertilization compares to the formation of a complete hydatidiform mole:

```
Normal Fertilization:

   Egg (23)        +      Sperm (23)
  (Maternal)              (Paternal)

                    ↓

            Zygote (46 chr)
           Diploid, Heterozygous
           (23 from mom,
            23 from dad)


Complete Hydatidiform Mole Formation:

  Empty Egg        +      Sperm (23)
    (0 chr)               (Paternal)

                    ↓

           Sperm DNA duplicates
```

```
(endoreduplication)

         ↓

   CHM (46 chr)
 Diploid, Homozygous
(46 from dad only,
 duplicated)
```

6.1.2 Normal Fertilization

When a sperm fertilizes an egg, the resulting cell—the zygote—has 46 chromosomes: 23 from the mother (in the egg) and 23 from the father (in the sperm). Each chromosome from mom pairs with the corresponding chromosome from dad (chromosome 1 from mom with chromosome 1 from dad, and so on). This creates a **diploid** genome with two versions of every chromosome.

Because mom and dad have different genetic variants, these chromosome pairs aren't identical. You might inherit brown-eye variants from dad and blue-eye variants from mom. At the DNA sequence level, humans typically differ at about **1 in every 1,000 base pairs** (0.1% of the genome), which translates to approximately **4–5 million heterozygous variants** between any two chromosome copies in a diploid genome. This genetic diversity is normal and healthy—it's why siblings look similar but not identical.

6.1.3 What Happens in a Complete Hydatidiform Mole

A complete hydatidiform mole occurs when something goes wrong during fertilization. The egg loses its nucleus (and therefore all maternal DNA), leaving essentially an empty egg with cytoplasm but no genetic material. When a sperm fertilizes this empty egg, there's only paternal DNA present.

To compensate for the missing maternal contribution, **the sperm's DNA duplicates itself** through a process called **endoreduplication**. The single set of 23 paternal chromosomes replicates to create 46 chromosomes—but they're all identical copies from the father.

The result is a cell with 46 chromosomes—a normal number—but all of them are identical copies from the father. There are two copies of dad's chromosome 1, two copies of dad's chromosome 2, and so on. **No maternal DNA at all.**

In most cases, complete hydatidiform moles have a **46,XX karyotype** because the duplicated sperm either carried an X chromosome that replicated, or two sperm (both carrying X) fertilized the empty egg and fused. Complete moles with 46,XY karyotypes are extremely rare because YY combinations are not viable—humans need at least one X chromosome to survive.

CHM13 has a **46,XX karyotype**, with all chromosomes derived from a single father's genome.

6.1.4 Understanding "Functionally Haploid"

This creates what geneticists call a **functionally haploid** genome. Let's clarify this potentially confusing term:

Term	Number of Chromosomes	Are They Identical?	Heterozygous Variants	Example
True Haploid	23 (one copy of each)	N/A (only one copy)	0 (no pairs to compare)	Sperm, Egg
Diploid (Normal)	46 (two copies of each)	**No** (maternal ≠ paternal)	~4–5 million	You, Me, Most People

Term	Number of Chromosomes	Are They Identical?	Heterozygous Variants	Example
Functionally Haploid (CHM13)	46 (two copies of each)	**Yes** (both from duplicated paternal DNA)	~Few thousand (⟨0.01%)	CHM13

"Functionally haploid" means that even though there are two copies of each chromosome (making it diploid in chromosome number), both copies are nearly identical—so it behaves like a haploid genome for sequencing purposes. There are almost no heterozygous variants to complicate assembly.

6.1.5 Nearly Uniform Homozygosity

It's important to note that CHM13 is not perfectly homozygous—no biological system is perfect. When scientists sequenced CHM13, they found:

- **A few thousand heterozygous variants** scattered across the genome
- **One megabase-scale heterozygous deletion** within the rDNA array on chromosome 15

These likely arose from: – Rare errors during the endoreduplication process that created the mole – Somatic mutations that occurred as the cell line was cultured over decades

However, these variants represent **less than 0.01% of the genome**. For comparison: – Normal diploid genome: ~4–5 million heterozygous variants – CHM13: ~few thousand heterozygous variants – **Reduction in heterozygosity: ⟩99.99%**

So while CHM13 isn't perfectly homozygous, it's close enough—the assembly is simplified by more than 99.99% compared to a typical diploid genome, which was exactly what the T2T project needed to resolve repetitive regions.

Figure: Karyotyping of CHM13. The image shows two types of chromosome analysis: (a) Spectral karyotyping (SKY), where each chromosome type is labeled with a different fluorescent color, and (b) G-banding, which reveals chromosome structure through staining patterns. Both confirm that CHM13 has a normal 46,XX karyotype with no structural abnormalities such as translocations, deletions, or duplications. Source: Miga, K.H. et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585, 79–84. https://doi.org/10.1038/s41586-020-2547-7. License: CC-BY 4.0.

6.1.6 Why This Is Perfect for Sequencing

Because both chromosome copies in CHM13 are nearly identical, there are virtually no heterozygous variants to confuse the assembly process. When sequencing reads come from a repetitive region, the computer doesn't have to distinguish which chromosome copy they belong to—they're essentially the same either way.

Think of it this way: imagine you're assembling two jigsaw puzzles simultaneously.

In a normal diploid genome: You have pieces from two similar but not identical puzzles mixed together—some pieces fit only in puzzle A, some only in puzzle B, and in repetitive regions (like a solid blue sky in both puzzles), many pieces might fit in either puzzle. The assembler must figure out which pieces belong to which puzzle, which becomes nearly impossible when you have hundreds of nearly identical pieces.

In CHM13: Both puzzles are essentially identical, so it doesn't matter which piece goes where—they fit the same pattern. This makes the assembly dramatically simpler, especially for the repetitive regions that plagued earlier sequencing efforts.

This advantage is **critical** for regions like centromeres, where a 2-kb repeat might be copied 1,400 times. In a diploid genome, if mom's centromere has 1,387 copies and dad's has 1,421 copies, with slight sequence

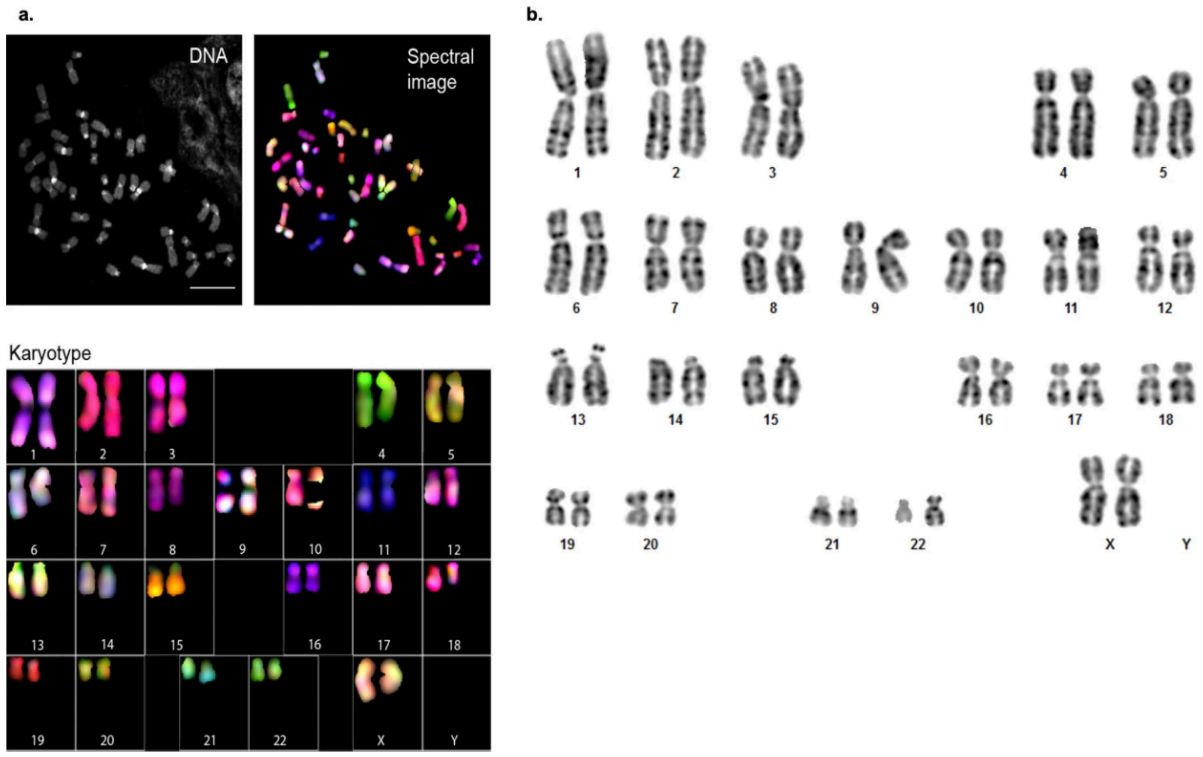


Figure 12: Miga et al. 2020, Nature – Extended Data Figure 1

variations between them, the assembler faces an impossible task of separating and reconstructing both. In CHM13, both copies have exactly the same number of repeats with virtually identical sequences.

6.2 The hTERT Modification: Making Cells Immortal

The full name of the cell line is **CHM13hTERT**. The "hTERT" part refers to a genetic modification that was added to make the cells **immortal**—capable of dividing indefinitely in culture.

6.2.1 The Telomere Problem

Normal human cells can only divide a limited number of times—typically **40 to 60 divisions** (known as the **Hayflick limit**)—before they stop. This limit exists because of structures called **telomeres**, which are repetitive DNA sequences (TTAGGG repeated thousands of times) that cap the ends of chromosomes, like the plastic tips on shoelaces.

```
Telomere structure:
5'-TTAGGGTTAGGGTTAGGGTTAGGG...(repeated 1,000-2,500 times)-3'
```

Every time a cell divides and copies its DNA, the telomeres get slightly shorter. This happens because of the **end-replication problem**: the DNA copying machinery (DNA polymerase) can't fully replicate the very ends of linear chromosomes. Each cell division results in the loss of approximately **50–200 base pairs** from the telomeres.

```
Telomere Shortening Over Cell Divisions:

Division 1:  TTAGGGTTAGGGTTAGGGTTAGGG... (10,000 repeats)
```

```
Division 2:  TTAGGGTTAGGGTTAGGG... (9,950 repeats)
Division 3:  TTAGGGTTAGGG... (9,900 repeats)
    ...
Division 50: TTAGGG... (5,000 repeats)
Division 60: TTA... (CRITICALLY SHORT!)
             ↓
         SENESCENCE
         (Cell stops dividing)
```

Eventually, after many divisions, telomeres become so short that the cell recognizes them as damaged DNA and stops dividing—a state called **senescence** (cellular aging). If cells try to divide past this point, they can undergo programmed cell death (apoptosis).

This built-in limit is actually a **protective mechanism** that prevents damaged cells from dividing uncontrollably and potentially becoming cancerous. But for research purposes, it's a problem. If you're trying to grow cells for extensive sequencing experiments that require billions of cells over multiple years, you need a large, consistent supply of cells with identical DNA. Running out of cells halfway through your project isn't an option.

6.2.2 Enter Telomerase

The solution is an enzyme called **telomerase**, which adds new TTAGGG repeats to telomeres, compensating for the bases lost during each cell division and preventing telomere shortening.

Telomerase consists of two key components: - **TERT (Telomerase Reverse Transcriptase)**: The catalytic protein that adds DNA repeats - **TERC (Telomerase RNA Component)**: An RNA template that specifies the TTAGGG sequence

Telomerase is naturally active in: - **Germ cells** (sperm and eggs): Need to divide many times and pass on full-length telomeres to offspring - **Stem cells**: Need to replenish tissues throughout life - **Some highly proliferative tissues**: Such as cells in the bone marrow or gut lining

But telomerase is **turned off in most adult somatic cells**, which is why they have limited lifespans and eventually senesce.

Importantly, approximately **85–95% of cancer cells reactivate telomerase**, allowing them to divide without limit—one key reason tumors can grow indefinitely. The remaining 5–15% of cancers use an alternative mechanism called ALT (Alternative Lengthening of Telomeres).

6.2.3 The hTERT Gene

The CHM13 cells were genetically modified to express **human telomerase reverse transcriptase (hTERT)**, the key catalytic component of telomerase. This modification was done by introducing the TERT gene into the cells using a viral vector, causing the cells to produce telomerase continuously.

```
With Telomerase (hTERT):

Division 1:  TTAGGGTTAGGGTTAGGGTTAGGG... (10,000 repeats)
Division 2:  TTAGGGTTAGGG... (9,950 repeats)
             ↓ Telomerase adds repeats
             TTAGGGTTAGGGTTAGGG... (10,000 repeats restored)
Division 3:  TTAGGGTTAGGGTTAGGGTTAGGG... (10,000 repeats)
    ...
Division ∞:  Still healthy, still dividing!
```

This modification makes the cells **immortal**—they can divide indefinitely while maintaining stable telomeres and consistent DNA. Critically, **the hTERT modification only affects telomere maintenance**; it doesn't

change the sequence of the 24 chromosomes (22 autosomes + X + X) that the T2T project was trying to sequence.

For the T2T project, this meant: – Researchers could grow as many CHM13 cells as needed over the multi–year project – They could extract large quantities of identical DNA for multiple rounds of sequencing – The cells remain **genetically stable** through many passages – The DNA sequenced at the beginning of the project matches the DNA sequenced years later

The original CHM13 cell line was developed in 1993 from a complete hydatidiform mole and has been used in genomics research for decades. The hTERT modification was added later to ensure long–term availability for research projects like T2T.

6.3 Genetic Stability and Quality Control

One major concern when using any cell line for reference genome sequencing is **genetic stability**. Cells grown in culture for many generations can accumulate mutations, chromosomal rearrangements, or other changes that would compromise the reference genome.

The T2T Consortium performed extensive quality control to ensure CHM13's stability:

6.3.1 Karyotype Analysis

G–banding and **spectral karyotyping (SKY)** were performed to visualize all 46 chromosomes and check for: – Correct chromosome number (should be 46) – Proper chromosome structure (no translocations, deletions, or duplications) – No evidence of aneuploidy (extra or missing chromosomes)

Results confirmed a **normal, stable 46,XX karyotype** with no observable chromosomal abnormalities (Extended Data Fig. 1 from Miga et al. 2020).

6.3.2 Sequence–Level Validation

The assembly itself provided evidence of genetic stability: – **Uniform read coverage** across all chromosomes (no large deletions or duplications) – **Low heterozygosity** confirmed throughout the genome (~few thousand variants, as expected) – **Consistent sequence quality** across multiple sequencing runs over several years

6.3.3 Why These Checks Matter

Unlike cancer cell lines (which also divide indefinitely through telomerase or ALT), CHM13hTERT maintains: – **Normal chromosome structure**: No translocations or aneuploidy – **Genetic stability**: Minimal mutations accumulating over time – **Reproducible results**: The same DNA sequence across multiple years and experiments

This stability is crucial for creating a reference genome. If the cells were accumulating mutations or chromosomal changes, the reference would be unreliable.

6.4 Ancestry and Genetic Background

The CHM13 genome doesn't represent all of humanity—it comes from a single source of DNA (one father's duplicated genome). Genetic ancestry analysis was performed using **maximum likelihood admixture analysis** to understand the genetic background of CHM13.

Figure: Genetic Ancestry Analysis of CHM13. Admixture analysis reveals the genetic ancestry composition of the CHM13 cell line. The analysis shows CHM13 has predominantly European ancestry, with

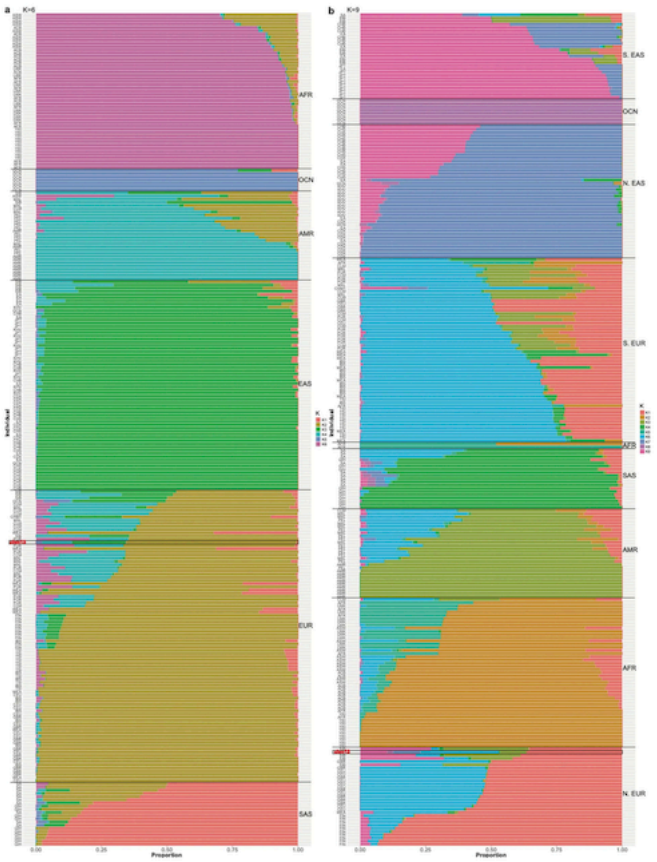


Figure 13: Miga et al. 2020, Nature – Extended Data Figure 2

smaller contributions from other ancestral populations. This ancestry information is important for understanding potential biases in using CHM13 as a reference genome and for comparative genomic studies across diverse human populations. Source: Miga, K.H. et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585, 79–84. https://doi.org/10.1038/s41586-020-2547-7. License: CC-BY 4.0.

The analysis revealed: - **Primarily European ancestry** (approximately 70–80% of the genome) - **Small amounts of admixture**: Possible South Asian, East Asian, or Native American ancestry - **Traces of Neanderthal DNA**: Like most non-African modern humans, CHM13 carries approximately **1–2% Neanderthal DNA**, inherited from interbreeding that occurred roughly 50,000–60,000 years ago

6.4.1 Does Ancestry Matter?

For the **T2T project's goals**, the specific ancestry of CHM13 doesn't significantly limit its usefulness. Here's why:

What matters about CHM13: - **Lack of heterozygosity** → enables complete assembly of repetitive regions - **Complete structural template** → shows where everything is located - **Gap-free reference** → resolves previously unsequenced regions

What doesn't matter as much: - **Specific genetic variants** → these vary between individuals anyway - **Population-specific alleles** → the pangenome captures these - **Ancestry composition** → any single genome has limitations

Think of it like creating the first complete map of a city's streets. It doesn't matter who drew the map or where the cartographer came from—what matters is that now we can see all the streets, buildings, and landmarks that were previously unmapped. Once we have that complete map, we can study how traffic patterns (genetic variants) differ in different neighborhoods (populations).

The key advantage of CHM13 isn't its specific genetic variants, but rather its **homozygous structure**, which enabled complete assembly of complex repetitive regions. Once you have that complete structural map, you can study how different variants exist in different populations—which is exactly what the **Human Pangenome Reference Consortium (HPRC)** is doing.

Important context: The limitation isn't CHM13's European ancestry specifically, but rather that any single genome can't represent all of humanity. Even if we had chosen a complete genome from an African, Asian, or South American individual, it still wouldn't capture human diversity. That's why we need pangenome references.

6.5 Why CHM13 Specifically?

You might wonder: if complete hydatidiform moles create homozygous genomes, why not use any CHM cell line? Why CHM13 specifically?

Several factors made CHM13 the ideal choice:

6.5.1 1. Availability and Characterization

CHM13 was already well-characterized and available to researchers. It had been used in previous genomics studies since the 1990s, including: - Segmental duplication analysis - As a benchmark genome for variant calling (part of the Genome in a Bottle consortium) - Validating assembly methods

6.5.2 2. Karyotype Stability

Unlike some cell lines that accumulate chromosomal abnormalities in culture, CHM13 maintained a **stable, normal 46,XX karyotype** through many passages over decades. This stability is unusual and valuable.

6.5.3 3. Favorable Karyotype (46,XX)

Having two X chromosomes (rather than XY) meant the project could immediately work on completing the X chromosome, which: – Had been extensively studied – Was chosen as the first target (completed in 2020) – Provided proof of concept for the methods

6.5.4 4. Successful hTERT Immortalization

The hTERT modification worked well in this line, ensuring: – Unlimited cell availability – Genetic stability despite unlimited passages – No chromosomal abnormalities from the immortalization

6.5.5 5. High-Quality DNA

CHM13 cells produce **high-quality, high-molecular-weight DNA** suitable for: – Ultra-long-read Nanopore sequencing (reads 〉100 kb, with some 〉1 Mb) – PacBio HiFi sequencing – Multiple complementary technologies

6.5.6 6. Community Support

By 2017, when the T2T project began seriously, CHM13 was already: – Used by multiple research groups – Part of the Genome in a Bottle reference materials – Well-documented with shared protocols

Other CHM cell lines exist, but CHM13 had the right combination of properties, existing characterization, stability, and research community support to serve as the foundation for the T2T project.

6.6 Why CHM13 Made T2T Possible

Let's bring all these pieces together. The T2T Consortium needed to sequence regions that had defeated earlier efforts—centromeres (multi-megabase satellite arrays), ribosomal DNA arrays (hundreds of tandem repeats), segmental duplications (〉90% sequence identity), and heterochromatic regions like Yq12 (〉30 Mb of satellites). These regions are highly repetitive, and in a normal diploid genome, small differences between the two chromosome copies make accurate assembly nearly impossible.

The **CHM13hTERT cell line** solved this problem through three key features:

6.6.1 1. Functionally Haploid Genome

Both chromosome copies are nearly identical (duplicated paternal DNA), eliminating 〉99.99% of heterozygous variants that confuse assembly algorithms in repetitive regions. This single feature enabled assembly of: – All 24 centromeres completely (ranging from 366 kb to several Mb) – Short arms of acrocentric chromosomes (66.1 Mb total) – Complex segmental duplications (6.61% of genome vs 5.00% in GRCh38) – Y chromosome heterochromatin (〉30 Mb, added from HG002 in v2.0)

The few thousand remaining heterozygous variants (~0.01% of genome) were manageable compared to the 4–5 million variants in a normal diploid genome.

6.6.2 2. Immortalized Cells

The hTERT modification allows unlimited cell division, providing a consistent, stable supply of DNA for extensive sequencing. The T2T project required enormous amounts of high-quality DNA: – **30× PacBio HiFi sequencing** (~20 kb reads, high accuracy) – **50× Oxford Nanopore ultra-long-read sequencing** (〉100 kb reads, up to 1+ Mb) – **100× Illumina short-read sequencing** (for polishing) – **Additional data**: Hi-C (chromatin conformation), Bionano optical mapping (physical map), Strand-seq (detecting inversions)

Generating this much data across multiple technologies over several years required a cell line that could provide unlimited, consistent DNA—which CHM13hTERT delivered.

6.6.3 3. Genetic Stability

Karyotype analysis confirmed a **normal 46,XX chromosome structure** with no abnormalities, ensuring the cells remain stable through many passages. Unlike some cancer cell lines (which also divide indefinitely), CHM13hTERT maintains: – Normal chromosome number – No translocations or large structural variants – Minimal accumulation of mutations over time

This stability was critical for generating consistent data across the entire multi-year project timeline.

When combined with advanced long-read sequencing technologies and innovative computational methods (marker-assisted polishing, string graph assembly), CHM13's unique properties enabled the first complete human genome assembly—**all 3.055 billion base pairs with no gaps**, covering the 22 autosomes and the X chromosome. The Y chromosome was added later from a different genome (HG002, a male genome commonly used for benchmarking) to create T2T-CHM13v2.0.

This achievement wouldn't have been possible with DNA from a typical diploid individual. Even with today's best long-read technologies, assembling a diploid human genome to telomere-to-telomere completeness remains challenging because of the need to separate and assemble two similar but distinct haplotypes in repetitive regions.

CHM13 provided the genomic simplicity needed to finally see what was hiding in the genome's most complex regions.

6.7 Comparing Haploid and Diploid Approaches

The table below summarizes how CHM13's functionally haploid genome compares to the approaches used in earlier reference genomes like GRCh38.

Feature	CHM13 (Functionally Haploid)	GRCh38 (Diploid, Mosaic)
Source	One complete hydatidiform mole (duplicated paternal DNA)	Multiple individuals, BAC clones from different donors
Chromosome Copies	Two nearly identical copies from duplicated paternal DNA	Mosaic of sequences from multiple individuals; represents different haplotypes
Heterozygous Variants	~Few thousand (⟨0.01% of genome)	~4–5 million per individual; structural differences between BAC clone donors
Assembly Complexity	Simple—minimal ambiguity in repetitive regions	Complex—heterozygous variants and structural differences create ambiguity, especially in repeats
Gaps in Assembly	**Zero gaps**—complete 3.055 billion bp coverage (with Y from HG002 in v2.0)	~151 million bp of unknown sequence (N-bases), mostly in repetitive regions
Centromeres	Fully assembled—all 24 centromeres complete, base-to-base	Largely absent or represented by small consensus sequences (~few kb placeholders)
Segmental Duplications	6.61% of genome (201.93 Mb) accurately assembled	5.00%—many regions collapsed, incorrectly assembled, or missing
Acrocentric Short Arms	Complete—66.1 Mb including all rDNA arrays	Almost entirely absent—represented as gaps
Y Chromosome	Complete in v2.0 from HG002 (62.46 Mb, no gaps)	⟩50% missing in GRCh38-Y (represented as one 30+ Mb gap)

Feature	CHM13 (Functionally Haploid)	GRCh38 (Diploid, Mosaic)
Accuracy	99.99% (1 error per 10 million bases) in unique regions; ~99.3% in largest satellite arrays	Very high in non-repetitive regions; undefined/not measurable in gaps
Technology	Long-read sequencing (PacBio HiFi, ONT ultra-long) + polishing	BAC cloning + Sanger sequencing (technology of early 2000s)
Assembly Approach	String graphs, marker-assisted polishing, ultra-long reads	Clone-by-clone BAC sequencing, short-read overlap assembly
Strengths	• Complete structural template• Enables studies of centromeres, satellites, and complex repeats• No reference gaps• High accuracy even in difficult regions	• Represents some genetic diversity (though limited)• Historical milestone• Mosaic includes some variant information• Well-validated in non-repetitive regions
Limitations	• Single genome source (one duplicated paternal genome)• Doesn't capture population diversity• No heterozygosity information• Chimeric (X from CHM13, Y from HG002)	• Incomplete in repetitive regions• Mosaic of different individuals• Reference bias toward source populations• Major gaps in functionally important regions
Best Use Cases	• Structural reference for complete genomic architecture• Centromere and satellite DNA studies• Complete gene annotations• Understanding complex repeat organization	• Standard variant calling in non-repetitive regions• Clinical diagnostics (with known limitations)• Many established bioinformatics workflows

6.8 The Limitations of CHM13

While CHM13 was essential for creating the first complete human genome, it's important to understand its limitations:

6.8.1 1. Represents Only One Genome Source

CHM13 is derived from a single father's duplicated DNA. It doesn't represent: - **Genetic diversity across human populations**: Different populations have different variants, structural variants, and gene copy numbers - **The typical diploid state**: Most humans have two different copies (maternal and paternal) of each chromosome - **Maternal contributions**: All sequences come from one father - **Common structural variants**: Many people carry structural variants (inversions, duplications, deletions) that aren't in CHM13

6.8.2 2. Ancestry Bias

With primarily European ancestry, CHM13 may not optimally represent: - **Variants common in African populations**: Africa has the most genetic diversity - **Asian-specific structural variants**: Important variants in East Asian, South Asian, or Native American populations - **Adaptive variants**: Genetic changes that helped populations adapt to different environments - **Population-specific gene families**: Some genes show copy number variation between populations

6.8.3 3. No Heterozygosity Information

Because both chromosome copies are nearly identical, CHM13 can't inform: - **Phasing of variants**: Which variants appear together on the same chromosome (linkage) - **Compound heterozygous effects**: Different mutations on each chromosome copy - **Heterozygous structural variants**: Many structural variants are heterozygous in populations

6.8.4 4. Chimeric Reference (X and Y from Different Sources)

Because CHM13 is 46,XX: - The **Y chromosome in T2T-CHM13v2.0 comes from HG002** (a different male genome) - This creates a **chimeric reference** for researchers working with both X and Y - The Y chromosome represents a different genetic background than the rest of the genome

6.8.5 5. Cannot Study Allelic Imbalance

In normal diploid genomes, the two chromosome copies can have different expression levels or functions. CHM13 can't inform: - **Allele-specific expression**: Whether maternal or paternal alleles are expressed differently - **Parent-of-origin effects**: Whether inheriting a variant from mom vs dad matters - **X-inactivation patterns**: Normal females have one X inactivated; CHM13's pattern may differ

---

6.9 Moving Beyond CHM13

While CHM13 was essential for creating the first complete human genome, it represents only one genome source—a single father's duplicated DNA. Human genetic diversity is vast, with populations differing in thousands of structural variants, gene copy numbers, and sequence variations.

The key lesson: The limitation isn't CHM13's specific ancestry, but rather that any single genome can't represent all of humanity. Even a complete genome from an African, Asian, or South American individual wouldn't capture human diversity.

This is why the field is now moving toward **pangenome references**—collections of many complete, high-quality genomes that represent human diversity. The Human Pangenome Reference Consortium (HPRC) is building on the T2T foundation by sequencing hundreds of diverse individuals, creating graph-based references that show all the different "paths" through variable regions of the genome. We'll explore this in detail in the next chapter.

For now, understand that CHM13 provided something essential: the first complete structural template of a human genome. It showed us what was hiding in the 151 million base pairs of gaps. Now we can study how these complex regions vary across populations—but first, we needed to see the complete structure.

Think of it like cartography: Before mapping how every city differs, you first need to know what a complete city map looks like—including the streets that were previously unmapped. CHM13 gave us that complete map.

6.10 Summary

The CHM13hTERT cell line made the T2T project possible through three unique features:

1. Functionally haploid genome: Nearly identical chromosome copies (〉99.99% homozygous) eliminated the heterozygosity problem that prevented assembly of repetitive regions.

2. Immortalized cells: The hTERT modification provided unlimited, genetically stable DNA for the multi-year, multi-technology sequencing effort.

3. Genetic stability: Normal 46,XX karyotype with no chromosomal abnormalities, confirmed through extensive validation.

These properties enabled the first complete human genome—all 3.055 billion base pairs with zero gaps—revealing ~200 million base pairs of previously unknown sequence in centromeres, rDNA arrays, segmental duplications, and heterochromatic regions.

CHM13's value and limitations: – **Value**: Provides the first complete structural template of human chromosomes – **Limitation**: Represents only one genome; doesn't capture human diversity – **Solution**: Pangenome projects build on this foundation to represent diversity

CHM13 wasn't the endpoint of genome sequencing—it was the essential foundation that makes studying human genetic diversity at complete resolution now possible.

6.11 Key Takeaways

1. **Heterozygosity prevents assembly** of repetitive regions—CHM13's near-homozygosity solved this fundamental problem.
2. **"Functionally haploid"** means diploid in number (46 chromosomes) but homozygous in sequence (both copies nearly identical).
3. **Complete hydatidiform moles** form when an empty egg is fertilized by a sperm whose DNA then duplicates through endoreduplication.
4. **The hTERT modification** enabled unlimited cell growth without compromising genetic stability.
5. **CHM13 represents one genome**, not human diversity—but provided the complete structural map needed to study diversity.
6. **The first complete human genome** filled ~151 million base pairs of gaps and revealed the hidden biology of our most complex genomic regions.

7 Chapter 4. From Single Reference to Pangenome

T2T-CHM13 is a remarkable achievement—the first complete, gapless human genome. But it has a fundamental limitation: it represents only one genome source. As we discussed in Chapter 3, this isn't unique to CHM13—any single genome can't capture human genetic diversity.

Why does this matter? Because humans are incredibly diverse genetically. If you compare your genome to someone else's, you'll find millions of differences—small changes in single DNA letters (called **SNPs**, or single nucleotide polymorphisms), along with larger changes where chunks of DNA are inserted, deleted, duplicated, or rearranged. These larger changes are called **structural variants (SVs)**.

Here's why a single reference is problematic:

The medication response problem: Imagine a drug that works by binding to a specific protein. The gene for this protein has a common structural variant in East Asian populations—a small insertion that slightly changes the protein's shape. If your reference genome doesn't include this variant, you might miss why the drug works differently in these patients.

The diagnosis problem: A patient of African ancestry has a large deletion in a gene region. Is this a disease-causing mutation, or a normal variant common in their population? A single reference genome based on European ancestry can't tell you—but a diverse pangenome can.

The solution? Don't use just one reference genome. Use many.

7.1 What Is a Pangenome?

A **pangenome** is a collection of multiple complete, high-quality genome sequences from diverse individuals, designed to capture the full range of genetic variation in a population or species. Instead of comparing

everyone's DNA to a single reference, you compare it to a set of genomes representing different ancestries.

The textbook analogy: If a single reference genome is like having one textbook, a pangenome is like having a library. Different books might explain the same concept in different ways, include different examples, or contain chapters the others don't have. By consulting multiple books, you get a more complete understanding.

7.1.1 The Human Pangenome Reference Consortium (HPRC)

In 2023, the **Human Pangenome Reference Consortium (HPRC)** released the first draft human pangenome (Liao et al. 2023, Nature). It includes complete, high-quality genome sequences from **47 individuals** carefully selected to represent global genetic diversity:

- **African populations**: 51% (Yoruba, Gambian, Mende)
- **American populations**: 34% (Puerto Rican, Peruvian, Colombian, Mexican)
- **Asian populations**: 13% (Han Chinese, Japanese, Punjabi, Bengali)
- **European populations**: 2% (British, Finnish, Iberian)

Because humans are diploid (we have two copies of each chromosome—one from each parent), these 47 individuals provide **94 haplotypes**—94 distinct chromosome sequences.

Figure: Genetic Diversity in the HPRC Pangenome. (See Figure 1a) Each dot represents one individual, positioned based on genetic similarity. Dots closer together are more genetically similar. The distribution shows that the pangenome samples were carefully selected to represent major human populations: African (51%), American (34%), Asian (13%), and European (2%). Source: Liao, W.W. et al. (2023). A draft human pangenome reference. Nature, 617, 312–324. https://doi.org/10.1038/s41586-023-05896-x. License: CC-BY 4.0.

7.2 What the Pangenome Revealed

7.2.1 1. We Were Missing A Lot of DNA

When the HPRC team assembled these 47 genomes, they made a striking discovery: collectively, these genomes contain about **119 million base pairs** of DNA sequence that isn't in the GRCh38 reference genome at all. That's roughly **4% more DNA** than the single reference showed us.

This "missing" DNA isn't junk—it includes: – **Gene sequences**: Complete genes that were missing from GRCh38 – **Regulatory regions**: DNA that controls when and where genes are turned on – **Structural variants**: Large insertions, deletions, or rearrangements

About **90 million base pairs** of this new sequence comes from structural variants—regions where different people have different DNA structures, not just different letters in the same sequence.

7.2.2 2. Genes Come in Different Copy Numbers

One of the most important findings was the discovery of **1,115 gene duplications** that vary between individuals. Some people have multiple copies of certain genes, while others have just one or two.

Figure: Gene Copy Number Variation in the Pangenome. (See panels b and c) Panel b shows most genes exist in 1–2 copies, but some genes have many more copies—up to 20+ in some individuals. Panel c shows different populations display different patterns of gene duplication. Each bar represents one genome, colored by ancestry: AMR (American, red), AFR (African, blue), EAS (East Asian, teal), SAS (South Asian, purple), and CHM13 (pink). The number and diversity of duplicated genes varies across populations, revealing variation that was invisible in single reference genomes. Source: Liao, W.W. et al. (2023). A

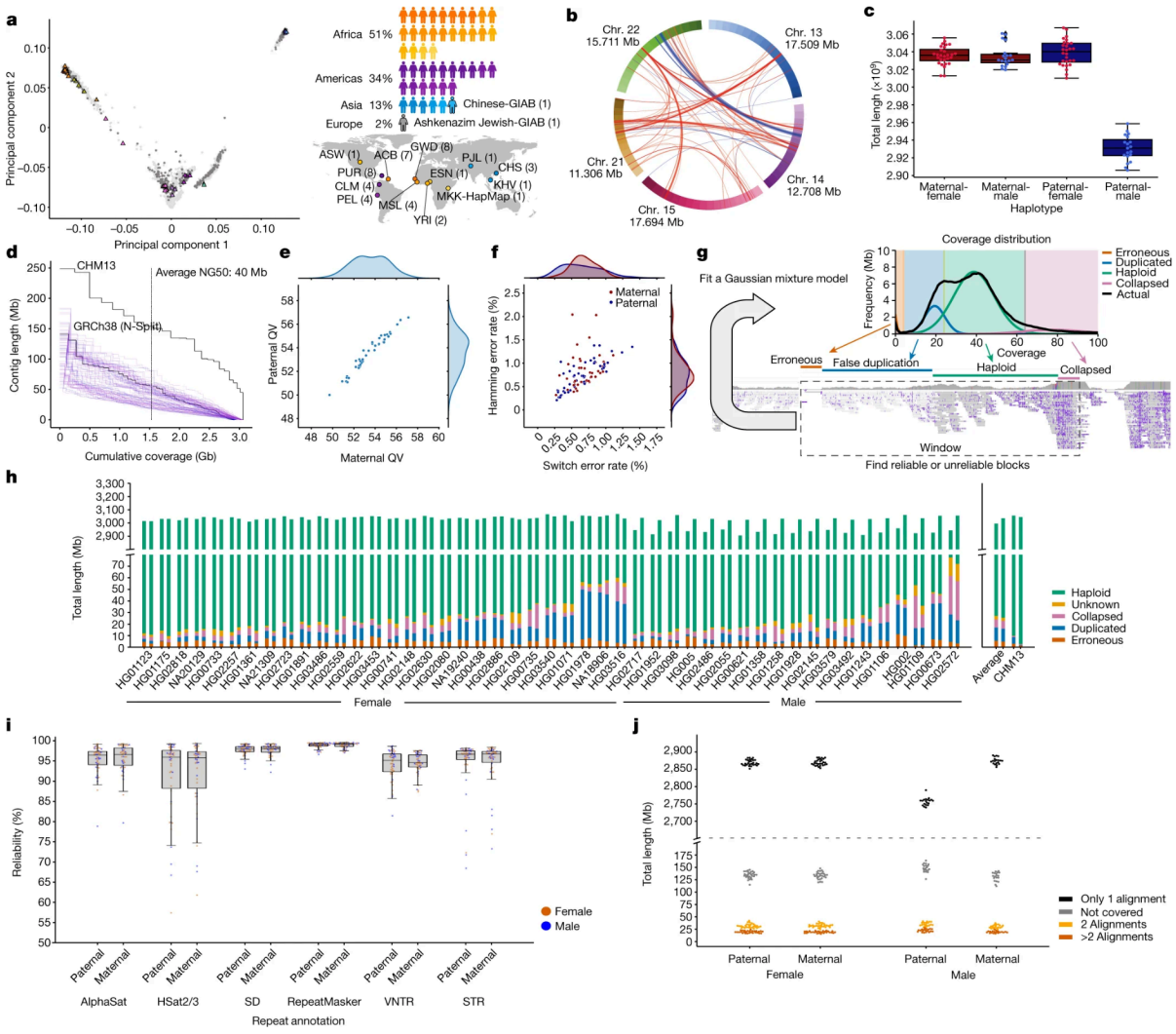


Figure 14: Liao et al. 2023, Nature – Figure 1a

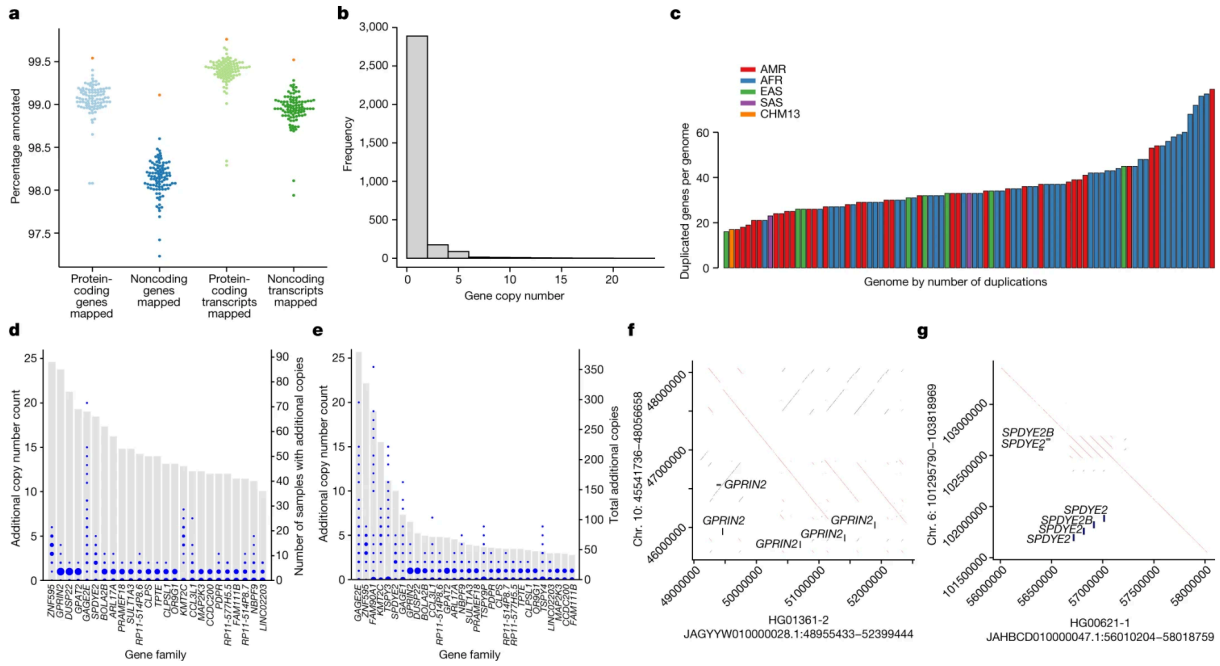


Figure 15: Liao et al. 2023, Nature – Figure 2b,c

draft human pangenome reference. Nature, 617, 312–324. https://doi.org/10.1038/s41586-023-05896-x. License: CC-BY 4.0.

Why does copy number matter? Having more copies of a gene typically means more of its protein product gets made. This can affect:

Starch digestion: The AMY1 gene makes an enzyme that breaks down starch. People whose ancestors ate starch-rich diets tend to have more AMY1 copies (up to 15+), helping them digest starches more efficiently. People from populations that historically ate less starch might have only 2–4 copies.

Immune response: Many immune system genes show copy number variation. Having more copies might provide better protection against certain infections, while having fewer copies might reduce the risk of autoimmune diseases.

Drug metabolism: Genes that break down medications can vary in copy number. This affects how quickly drugs are cleared from your body—crucial for determining the right dose.

The single reference genome assumed everyone had the same number of copies. The pangenome shows the truth is much more complex.

7.2.3 3. Better Detection of Structural Variants

Structural variants are large changes in DNA structure: – **Deletions**: A chunk of DNA is missing – **Insertions**: Extra DNA is added – **Duplications**: A region is copied one or more times – **Inversions**: A DNA segment is flipped backwards

These are important because: – **Many diseases are caused by structural variants**, not just single-letter changes – **Structural variants affect more DNA** than all SNPs combined – **They're harder to detect** with a single reference

The pangenome dramatically improves structural variant detection: – **34% fewer errors** when identifying small variants (SNPs and small insertions/deletions) – **104% improvement** in detecting structural variants—more than doubling accuracy

Why such improvement? When you compare someone's DNA to multiple references, you're more likely to find the right match. If someone has a structural variant common in their population, it's probably already in one of the 94 haplotypes.

7.3 How Was the Pangenome Built?

Building high-quality genomes for 47 individuals required the same technologies used for T2T-CHM13:

7.3.1 Key Technologies

- **PacBio HiFi sequencing**: Long, accurate reads (~20 kb)
- **Oxford Nanopore sequencing**: Ultra-long reads (>100 kb)
- **Hi-C**: Maps how DNA folds in 3D space, helping separate maternal and paternal chromosomes
- **Illumina sequencing**: Short, highly accurate reads for error correction

7.3.2 The Challenge: Phasing

Unlike CHM13 (where both chromosome copies are identical), these 47 individuals are diploid—each has two different versions of every chromosome. The assembly must:

1. **Separate the two haplotypes**: Keep maternal and paternal chromosomes distinct
2. **Maintain accuracy**: Ensure each haplotype is correctly assembled
3. **Resolve structural variants**: Correctly place large insertions, deletions, and rearrangements

The HPRC achieved **phased assemblies**—meaning they successfully separated the two haplotypes for each individual. The quality is impressive: the assemblies are nearly as complete as T2T-CHM13, with most gaps only in the most repetitive regions (centromeres and rDNA arrays).

7.4 Real-World Impact: Why This Matters

7.4.1 1. Clinical Diagnostics

Before (single reference): A patient of African ancestry is found to have a 50-kb deletion in a gene. The clinical report flags it as "likely pathogenic" because it differs from GRCh38.

With pangenome: The same deletion is found in 30% of the African haplotypes in the pangenome. It's reclassified as "benign variant, common in African populations." The patient is spared unnecessary worry and invasive testing.

7.4.2 2. Drug Development

Pharmaceutical companies test drugs on diverse populations, but analyze the results using a single reference genome. If a drug works differently in certain populations due to structural variants not in the reference, those variants might be missed. The pangenome helps identify these genetic factors, leading to: - Better dose recommendations for different populations - Understanding why some people have side effects - Identifying who will benefit most from a treatment

7.4.3 3. Disease Research

Many diseases linked to structural variants were hard to study with single references: - **Autism spectrum disorders**: Often caused by duplications or deletions - **Schizophrenia**: Associated with copy number variants - **Cardiovascular disease**: Some forms linked to structural variants in lipid metabolism genes

The pangenome's improved detection of these variants accelerates research into their causes and potential treatments.

7.5 Current Limitations and Next Steps

The 2023 draft pangenome is a major advance, but it's not yet complete:

7.5.1 What's Still Challenging

The most repetitive regions (about 4.4% of the genome): – Centromeric satellite arrays – Ribosomal DNA arrays
– Some heterochromatic regions

These regions vary dramatically between individuals and are still difficult to align and compare. This isn't a failure—it reflects the genuine biological complexity of these regions.

7.5.2 Future Expansion

The HPRC's long-term goal is to sequence **350 individuals** from diverse global populations, providing **700 haplotypes**. This expanded pangenome will:

- **Cover more rare variants**: Better represent genetic diversity within populations
- **Include underrepresented populations**: Current genetic databases are biased toward European ancestry; expansion will improve equity
- **Enable finer-scale studies**: Understand variation at the population and subpopulation level

Additionally, other species are getting pangenomes: – **Plant crops** (wheat, rice, maize): To improve breeding and understand disease resistance – **Livestock** (cattle, pigs, chickens): For agricultural improvement – **Model organisms** (mice, fruit flies): For biomedical research

---

7.6 Comparing Genome References: The Evolution

Let's see how genome references have evolved:

Reference	Year	Coverage	Key Achievement	Main Limitation
GRCh38	2013	~95%	Most widely used; established standard for clinical and research genomics	~151 Mb of gaps; represents mosaic of mostly European individuals; reference bias in diverse populations
T2T-CHM13	2022	100%	First complete genome; no gaps; resolved all centromeres, telomeres, and rDNA arrays	Single genome source; doesn't capture human diversity
HPRC Pangenome (draft)	2023	〉99% per haplotype	94 haplotypes from diverse populations; captures structural variants; reduces reference bias	Some gaps in most repetitive regions; 47 individuals (goal: 350)

Reference	Year	Coverage	Key Achievement	Main Limitation
HPRC Pangenome (future)	~2026–27	〉99% per haplotype	700 haplotypes; comprehensive diversity; improved tools for clinical use	Most repetitive regions may remain challenging to compare

---

7.7 The Path Forward: From Reference to Resources

The future of genomics isn't about choosing one reference—it's about using the right reference(s) for your question:

For studying genome structure: T2T-CHM13 is ideal—it shows the complete architecture with no gaps.

For clinical diagnostics: The pangenome reduces false positives from reference bias and improves variant interpretation across populations.

For population genetics: The pangenome reveals patterns of variation that were invisible with single references.

For evolutionary studies: Comparing human pangenomes to those of other primates reveals what makes us uniquely human.

Think of it as having different maps for different purposes: - A **detailed street map** (T2T-CHM13): Shows every street, building, landmark—the complete structure - An **atlas with many maps** (pangenome): Shows how different cities differ, capturing diversity - A **historical map series**: Shows how things have changed over time (evolution)

7.8 Why This Matters for You

If you pursue research or medicine involving genetics, understanding these references is essential:

For researchers: The pangenome enables more accurate studies of genetic variation, disease associations, and trait evolution across diverse populations.

For clinicians: As genome sequencing becomes routine in medicine, using references that capture diversity ensures accurate diagnoses and personalized treatments for all patients, not just those of European ancestry.

For equity: Using diverse references helps ensure that genomic medicine benefits everyone. Historically, genetic research has focused on populations of European ancestry, creating disparities in the quality of genetic medicine available to other populations. The pangenome is a step toward fixing this inequity.

For understanding humanity: The pangenome shows us how humans are similar and different at the most fundamental level—our DNA. It's not just about medicine; it's about understanding what makes us human and how we adapt to different environments.

7.9 Summary: One Map to Many Maps

The journey from single reference to pangenome represents a fundamental shift in how we think about genomes:

GRCh38: One map with gaps—useful but incomplete and biased toward certain populations

T2T-CHM13: One complete map with no gaps—shows the full structure but only one genome

HPRC Pangenome: Many complete maps—captures human diversity and reduces bias

The pangenome doesn't replace these earlier references; it complements them. T2T-CHM13 showed us what a complete human genome looks like. The pangenome shows us how humans differ.

Together, they provide the foundation for precision medicine, equitable healthcare, and a deeper understanding of human biology. As the pangenome expands to 700 haplotypes and beyond, it will become an increasingly essential resource for genomics research and clinical practice.

The future isn't one reference genome—it's a rich, diverse collection that reflects the genetic reality of humanity.

8 Chapter 5. The Emergence of Next-Generation Sequencing Technologies

8.1 The Bottleneck of Sanger Sequencing

When the Human Genome Project finished in 2003, it was an extraordinary achievement. But let's be honest about what it took: over a decade of work, contributions from sequencing centers around the world, and approximately $300 million. All of that effort produced one reference genome.

The method used was **Sanger sequencing**, developed in the 1970s by Frederick Sanger (who won his second Nobel Prize for it). Sanger sequencing is remarkably accurate—error rates of less than 0.1%—but it has a fundamental limitation: it reads one DNA fragment at a time.

Figure: Sanger sequencing. This diagram illustrates how Sanger sequencing works. Source: Microbe Notes

Imagine you're trying to transcribe a million books, but you can only read and copy one sentence at a time before moving to the next. That's essentially what Sanger sequencing does. For the Human Genome Project, this meant reading 3 billion letters of DNA code one fragment at a time. Even with hundreds of sequencing machines running 24/7, it took years.

After the HGP, it became clear that genomics couldn't progress at this pace. Scientists wanted to sequence not just one genome, but thousands—comparing genomes from healthy and sick individuals, studying genetic variation across populations, identifying disease-causing mutations in patients. At $300 million and a decade per genome, this was impossible.

The field needed a revolution: faster sequencing, cheaper sequencing, more sequencing. The answer came with **Next-Generation Sequencing (NGS)**.

8.2 What Makes NGS "Next-Generation"?

The key innovation of NGS is simple but powerful: **parallelization**. Instead of reading one DNA fragment at a time, NGS reads millions of fragments simultaneously.

Let's return to our book transcription analogy. Instead of one person copying sentences sequentially, imagine you have a million people, each transcribing a different sentence at the same time. Suddenly, what took years can be done in days.

That's NGS. By 2007, the first NGS platforms were commercially available, and by 2014, the cost of sequencing a human genome had dropped to around $1,000—a 300,000-fold reduction in just over a decade. Today, you can sequence a genome for about $600-$1,000 in a matter of hours or days.

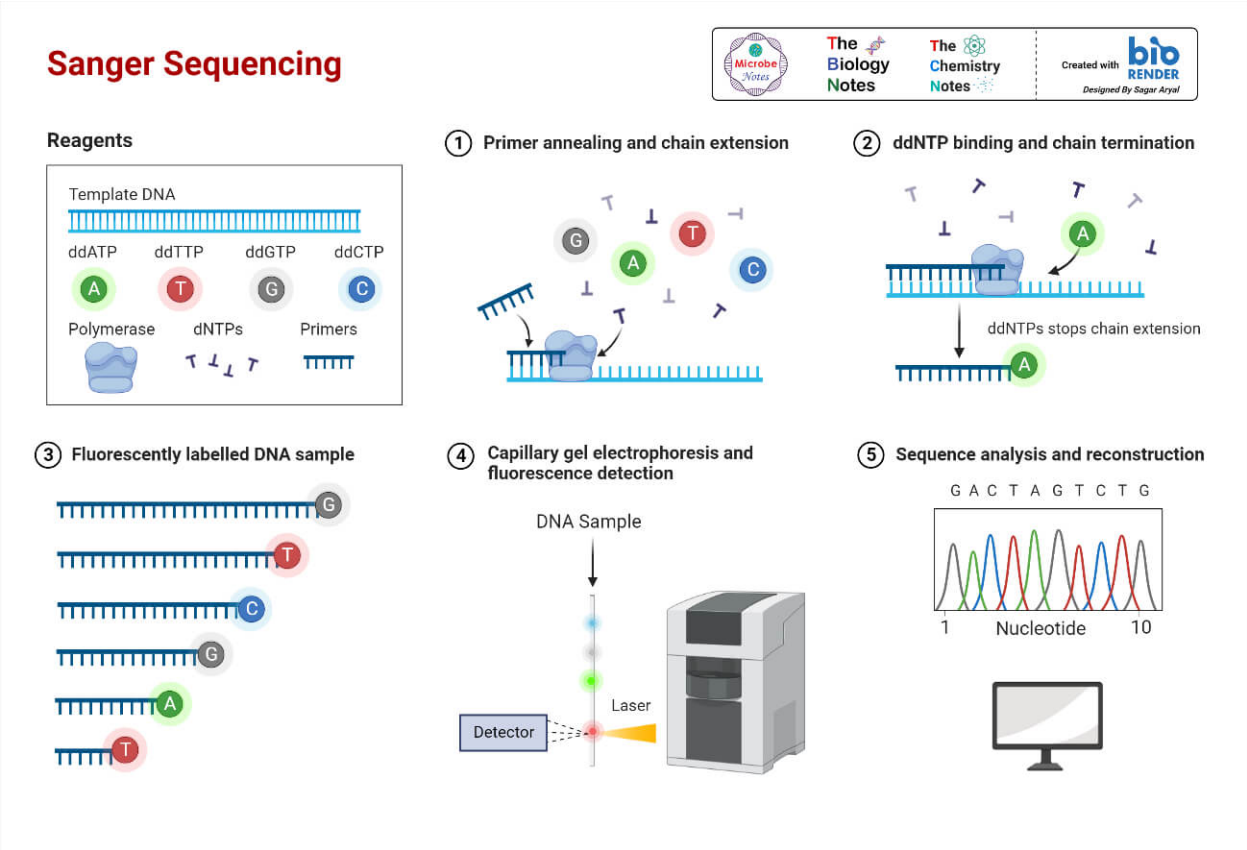


Figure 16: Sanger sequencing

This dramatic cost reduction, which has outpaced even Moore's Law (the exponential improvement in computer technology), made genomics practical for large-scale research and clinical medicine.

8.3 Understanding the Basics: Key Concepts in NGS

Before we dive into specific technologies, let's clarify some terms you'll encounter throughout this chapter and in genomics research.

8.3.1 Sequencing Reactions and Reads

When we sequence DNA, we break it into fragments and determine the order of bases (A, C, G, T) in each fragment. Each successfully sequenced fragment produces a **read**—literally, a "reading" of that fragment's sequence.

Different technologies produce different **read lengths**: – Illumina (a common NGS platform) produces reads of 50–300 base pairs – PacBio Revio produces reads of ~15,000–20,000 base pairs – Oxford Nanopore can produce reads exceeding 1 million base pairs

8.3.2 Coverage and Depth

Two related but distinct concepts measure how thoroughly we've sequenced a genome:

Coverage tells us what percentage of the target DNA (like a genome or a specific gene region) we've sequenced at all. For example, 95% coverage means we successfully read 95% of the genome, but 5% has no data—usually in very repetitive regions that are hard to sequence.

Depth (or coverage depth) tells us how many times, on average, each base was sequenced. If we sequence a genome at 30× depth (pronounced "30-times depth"), each base position is covered by an average of 30 different reads.

Why does depth matter? Because sequencing isn't perfect. Every technology makes occasional errors, and biological samples can contain multiple DNA molecules with slight differences. By sequencing each position multiple times, we can distinguish real genetic variants from sequencing errors through statistical confidence. A variant supported by 20 out of 30 reads is much more believable than one supported by 1 out of 2 reads.

For most applications, 30× depth is considered the minimum for reliable variant detection in humans.

8.3.3 Base Quality Scores

Every base call (identifying whether a position is A, C, G, or T) comes with a **quality score** that reflects the sequencer's confidence. These are usually expressed as Q scores:

- Q20 = 99% confidence (1% error rate)
- Q30 = 99.9% confidence (0.1% error rate)
- Q40 = 99.99% confidence (0.01% error rate)

High-quality bases (Q30 or higher) are critical for reliable variant detection, especially when looking for rare mutations.

8.3.4 Libraries and Adaptors

Before sequencing, DNA must be prepared into a **library**—a collection of fragments with short, known sequences called **adaptors** attached to both ends. These adaptors serve as handles that allow the sequencing machine to grab and amplify the DNA fragments. Different NGS platforms require different library preparation methods, but the concept is universal.

8.3.5 Alignment and Variant Calling

Once you have millions of reads, you need to figure out where they came from in the genome. This is called **alignment** or **mapping**—computationally comparing each read to a reference genome (like GRCh38) to determine its original location.

After alignment, you can identify **variants**—positions where the sequenced DNA differs from the reference. Variants include: – **Single nucleotide polymorphisms (SNPs)**: one–letter changes (e.g., A→G) – **Insertions and deletions (indels)**: added or missing bases – **Structural variants**: large–scale changes like duplications or rearrangements

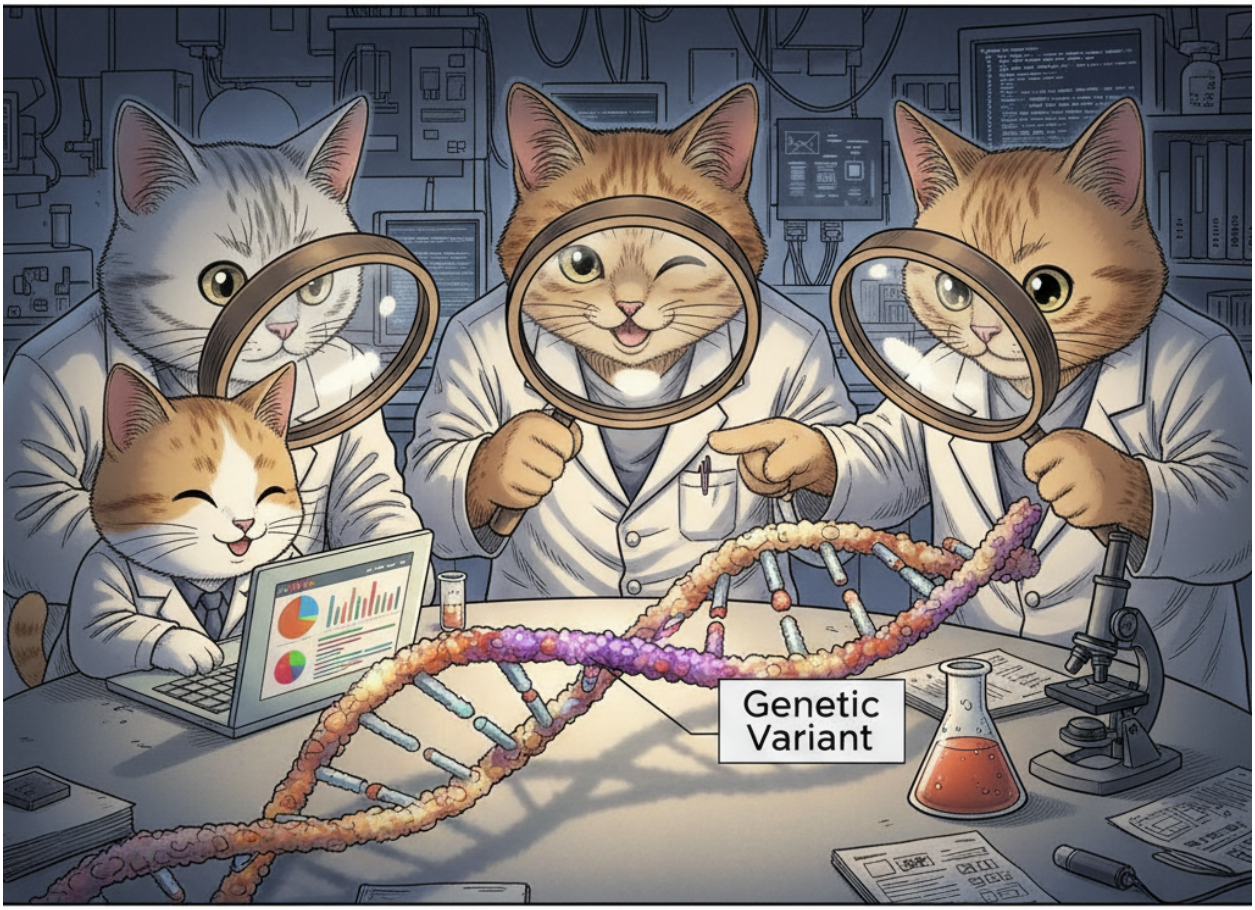


Figure 17: Cats for variants

8.4 The NGS Technology Landscape

NGS platforms fall into two main categories: **second-generation** (short-read, high-throughput) and **third-generation** (long-read, single-molecule). Each has distinct strengths and uses.

8.4.1 Illumina: The Workhorse of Genomics (Second-Generation)

How It Works:

Illumina sequencing uses a clever method called **sequencing by synthesis**. Here's the process simplified:

1. DNA fragments are attached to a flat surface called a **flow cell**, which contains millions of tiny spots
2. Each fragment is amplified into a cluster of identical copies (imagine making thousands of photocopies of each fragment)
3. The sequencing machine adds fluorescently labeled nucleotides one at a time. Each of the four bases (A, C, G, T) has a different colored fluorescent tag
4. A camera takes a picture after each addition, recording which color (and therefore which base) appeared at each spot
5. The fluorescent tag is removed, and the process repeats for the next base

This happens simultaneously for millions of clusters, generating billions of bases of sequence data in a single run.

Figure: Illumina Sequencing by Synthesis (SBS). This diagram illustrates how Illumina sequencing works. (Top) DNA fragments are attached to a flow cell surface and amplified into clusters through bridge amplification. (Middle) During sequencing, fluorescently labeled nucleotides are added one at a time, and a camera captures the color signal from each cluster. (Bottom) The fluorescent tag is cleaved off, allowing the next nucleotide to be added. This cycle repeats to build up the complete sequence. Source: Microbe Notes

Strengths: – **High throughput**: Can generate hundreds of billions of bases per run – **Low cost**: About $600–$1,000 for a complete human genome at 30× coverage – **High accuracy**: Error rates around 0.1%, comparable to Sanger sequencing – **Well-established**: Widely used with mature analytical tools

Limitations: – **Short reads**: Typically 150–300 base pairs, making it difficult to resolve repetitive regions or complex structural variants – **PCR amplification**: The clustering step can introduce biases

When to Use It:

Illumina is the default choice for most genomics applications: whole-genome sequencing (WGS), whole-exome sequencing (WES—sequencing just the protein-coding regions), identifying SNPs and small indels, RNA sequencing, and clinical diagnostics. If you need to sequence many samples affordably and accurately, Illumina is usually the answer.

8.4.2 PacBio Revio: Long Reads with High Accuracy (Third-Generation)

How It Works:

Pacific Biosciences (PacBio) uses **Single-Molecule Real-Time (SMRT) sequencing**. Unlike Illumina, which sequences amplified clusters, PacBio sequences individual DNA molecules in real time.

Here's the concept:

1. A single DNA polymerase enzyme (the molecular machine that copies DNA) is placed at the bottom of a tiny well called a **zero-mode waveguide (ZMW)**
2. A single DNA molecule threads through this polymerase
3. As the polymerase adds bases to the growing DNA strand, fluorescently labeled nucleotides flash briefly as they're incorporated
4. A detector at the bottom of the well records these flashes in real time, determining the sequence



Figure 18: Illumina Sequencing Principle

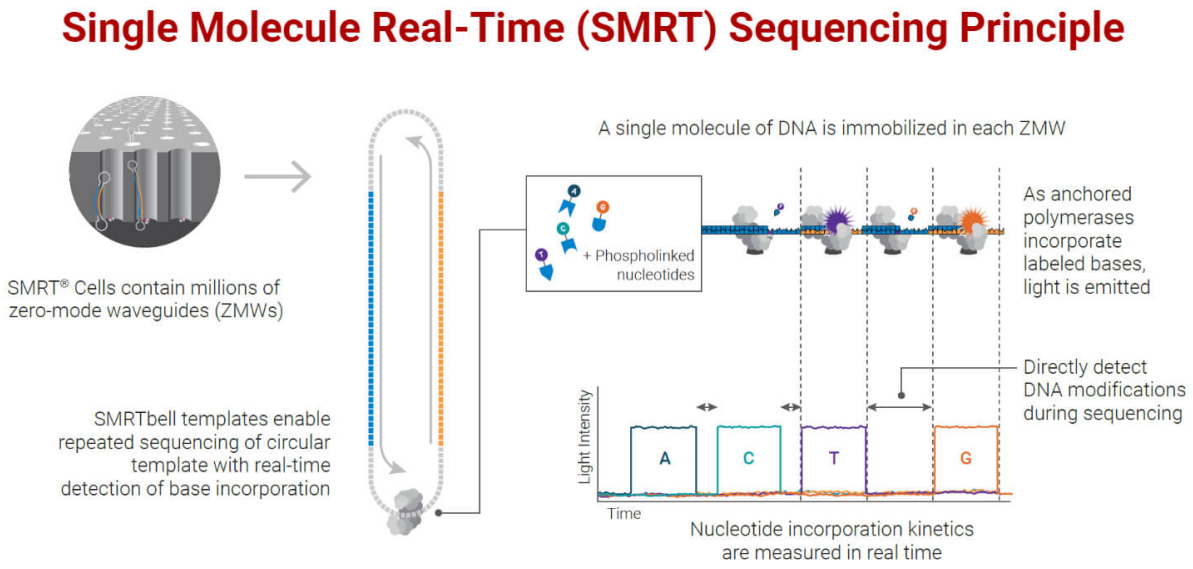


Figure 19: PacBio SMRT Sequencing Principle

Figure: PacBio Single-Molecule Real-Time (SMRT) Sequencing. This diagram shows the key components of PacBio sequencing. A DNA polymerase is fixed at the bottom of a zero-mode waveguide (ZMW)—a tiny well that allows detection of just the molecules at the bottom. As the polymerase incorporates fluorescently labeled nucleotides into the growing DNA strand, brief flashes of light are detected and recorded in real time. The circular DNA template (SMRTbell) allows the polymerase to sequence the same molecule multiple times, improving accuracy. Source: Microbe Notes

The newest PacBio system, **Revio** (released in 2022), contains 25 million ZMWs per chip—25 million DNA molecules being sequenced simultaneously. The system uses improved chemistry (called SPRQ) that requires less starting DNA and achieves higher accuracy.

PacBio produces what are called **HiFi reads** (High-Fidelity reads). By sequencing each DNA molecule multiple times in a circular fashion (the DNA is formed into a loop), the technology achieves both long read length (~15,000–20,000 base pairs) and high accuracy (>99.9%, or Q30).

Strengths: – **Long, accurate reads**: Resolves repetitive regions that defeat short-read sequencing – **Structural variant detection**: Can span and accurately sequence large insertions, deletions, or rearrangements – **Epigenetic information**: Can detect DNA methylation patterns while sequencing – **No PCR amplification**: Sequences native DNA molecules, avoiding amplification biases

Limitations: – **Higher cost**: About $500 per genome, more expensive than Illumina – **Lower throughput**: Can't sequence as many genomes simultaneously as Illumina – **More DNA required**: Library preparation typically needs more starting material

When to Use It:

PacBio excels at de novo genome assembly (building a genome from scratch without a reference), detecting structural variants, sequencing through repetitive regions (like centromeres and telomeres), and studying complex genetic regions. The T2T-CHM13 genome relied heavily on PacBio HiFi reads to complete previously inaccessible regions.

8.4.3 Oxford Nanopore: Ultra-Long Reads and Portable Sequencing (Third-Generation)

How It Works:

Oxford Nanopore takes a completely different approach. Instead of using chemical or optical detection, it uses electrical current.

The concept:

1. A protein structure called a **nanopore** (a tiny hole) is embedded in a membrane
2. An electrical current flows through the pore
3. A single strand of DNA is threaded through the nanopore
4. As each base passes through, it disrupts the current in a characteristic way
5. The pattern of current changes reveals the DNA sequence

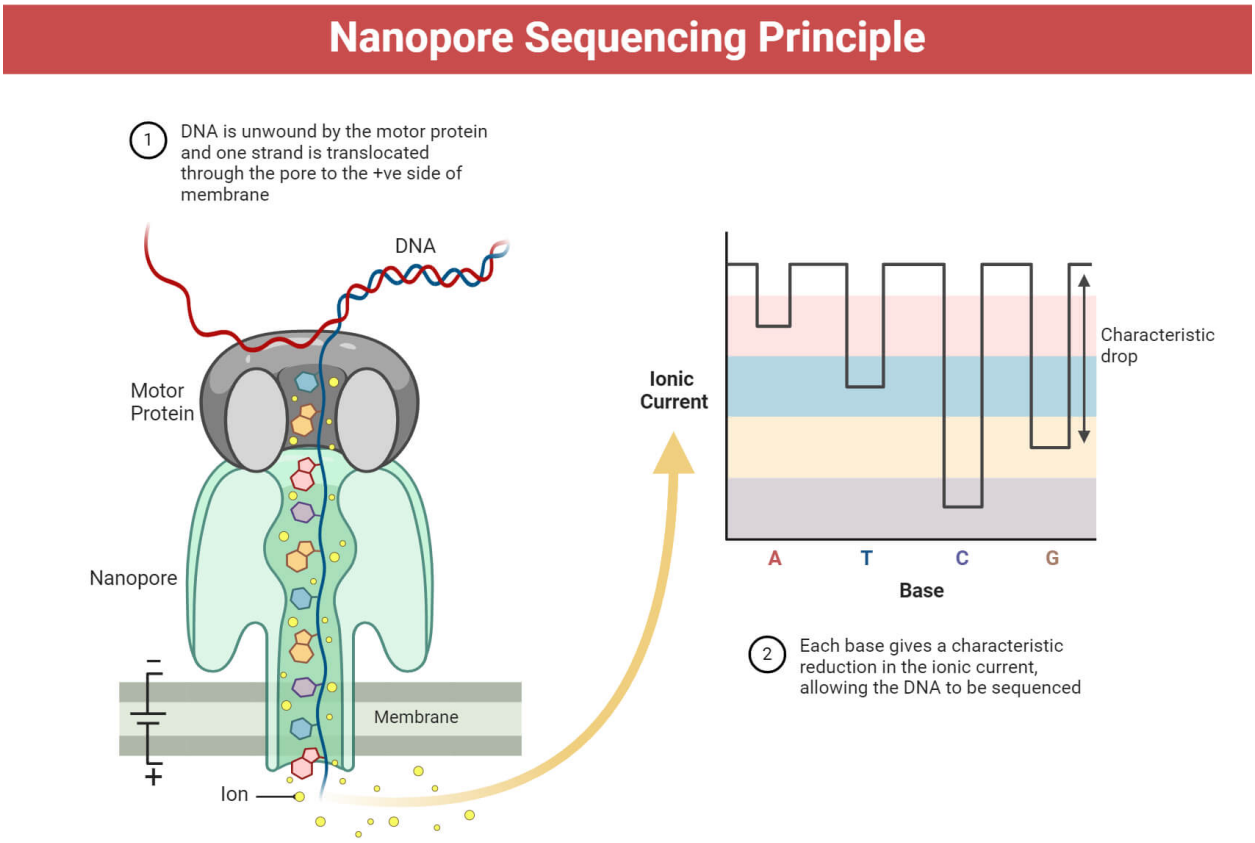


Figure 20: Oxford Nanopore Sequencing Principle

Figure: Oxford Nanopore Sequencing Technology. This illustration shows how nanopore sequencing works. (Left) A nanopore protein is embedded in a membrane with an electrical current flowing across it. Motor proteins (shown in purple) control the movement of a DNA strand through the nanopore. (Right) As different nucleotides pass through the nanopore, they cause characteristic disruptions to the electrical current. These current changes (shown as different colored signals) are measured in real time and decoded by software to determine the DNA sequence. Source: Microbe Notes

Strengths: – **Ultra-long reads**: Can produce reads exceeding 1 million base pairs—the longest of any sequencing technology – **Real-time analysis**: Data streams in as sequencing happens, allowing immediate analysis – **Portable devices**: The MinION sequencer is about the size of a USB drive and can sequence anywhere (used in Ebola outbreak monitoring, Antarctic research, even on the International Space Station) – **No amplification or special preparation**: Can sequence native DNA with minimal processing

Limitations: - **Higher error rate**: Historically 5–10%, though newer base-calling algorithms using machine learning have improved this significantly - **Less mature technology**: Still rapidly evolving compared to Illumina or PacBio

When to Use It:

Nanopore shines in situations requiring rapid results, ultra-long reads, or field sequencing—outbreak monitoring, rapid clinical diagnostics, environmental sampling, resolving extremely complex genomic regions, and research where sample preparation facilities are limited. The T2T project used ultra-long Nanopore reads to span entire repetitive regions and provide scaffolding for the assembly.

8.5 Comparing the Technologies

Platform	Read Length	Accuracy	Cost per Genome	Key Strength	Best For
Illumina	150–300 bp	~99.9% (Q30)	$600–$1,000	High throughput, low cost, proven reliability	Routine WGS/WES, variant calling, most clinical applications
PacBio Revio	15,000–20,000 bp	~99.9% (Q30) HiFi	~$500	Long reads + high accuracy	Genome assembly, structural variants, repetitive regions
Oxford Nanopore	Up to >1 million bp	~95–99% (improving)	Variable	Ultra-long reads, real-time, portable	Complex regions, rapid diagnostics, field sequencing

---

8.6 How NGS Changed Genetics

8.6.1 From Gene-by-Gene to Genome-Wide

Before NGS, sequencing a single gene in multiple patients was a major research project. NGS made it routine to sequence all ~20,000 genes (whole-exome sequencing) or the entire 3 billion base pairs (whole-genome sequencing) in thousands of individuals.

This enabled **genome-wide association studies (GWAS)**, where researchers compare genetic variants across thousands of people with and without a disease to identify risk factors. Diseases like diabetes, heart disease, schizophrenia, and hundreds of others now have known genetic risk variants identified through NGS-based GWAS.

8.6.2 The Variant Catalog

NGS enabled systematic cataloging of human genetic variation. The **dbSNP** database now contains over 1.1 billion recorded variants—a comprehensive map of how humans differ genetically. This redefined our

concept of genes: rather than viewing genes as fixed sequences, we now understand them as loci with many possible alleles (variant forms).

8.6.3 Clinical Diagnostics

NGS brought genomic medicine from research to the clinic. Whole-exome sequencing (WES) is now a standard diagnostic tool for rare genetic diseases, successfully identifying the cause in 25–50% of cases.

A famous early example: in 2010, researchers used WES to identify the cause of Miller syndrome in just four siblings by comparing their exomes to their healthy parents. The gene **DHODH** had never been linked to the disease before. What might have taken decades with traditional methods took months with NGS.

PacBio's long reads have further improved diagnostics by revealing structural variants and mutations in complex genomic regions that short-read sequencing misses.

8.6.4 Completing the Genome

Perhaps most dramatically, NGS—specifically long-read sequencing from PacBio and Nanopore—finally completed the human genome. The T2T-CHM13 project used these technologies to fill the 8% of the genome the HGP couldn't sequence, resolving centromeres, telomeres, and other highly repetitive regions.

8.7 The Economics of Sequencing

One way to appreciate NGS's impact is through cost:

- **2001** (HGP draft): ~$300 million per genome
- **2007** (first NGS platforms): ~$10 million per genome
- **2010**: ~$50,000 per genome
- **2014**: ~$1,000 per genome
- **2024**: ~$500 per genome (Illumina), ~$2000 (PacBio for research)

Figure: The Cost of Sequencing a Human Genome Over Time. This graph from the National Human Genome Research Institute shows the dramatic decrease in sequencing costs from 2001 to 2022. The blue line shows the actual cost per genome, while the dotted line represents Moore's Law (the rate at which computer technology improves). Notice that after 2008, when NGS technologies became widely available, sequencing costs dropped far faster than Moore's Law—a remarkable achievement. The vertical axis uses a logarithmic scale, so each step down represents a 10-fold cost reduction. Source: NHGRI Genome Sequencing Program

This exponential cost reduction—far faster than the improvement in computer technology (Moore's Law)—transformed genomics from an elite research endeavor to a clinical reality.

8.8 Practical Considerations

When you encounter NGS data in research or the clinic, consider:

What sequencing platform was used? This determines read length, error characteristics, and what types of variants can be reliably detected.

What's the coverage depth? Higher depth (30×+) gives more confidence in variant calls, especially for rare variants or somatic mutations in cancer.

What's the target? Whole-genome, whole-exome, or targeted panels? Each approach has different coverage patterns and sensitivity.

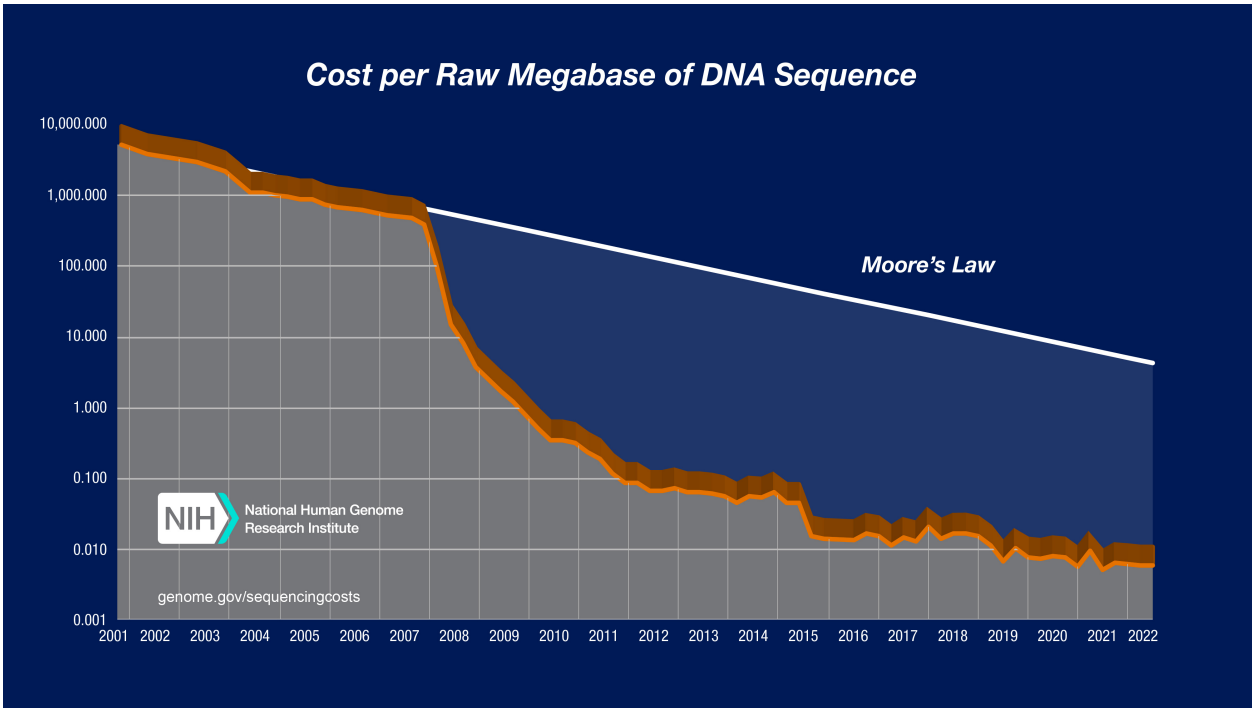


Figure 21: DNA Sequencing Cost Per Genome

What reference genome was used for alignment? Results aligned to GRCh38 might differ from those using T2T-CHM13 or the pangenome, especially in complex regions.

Understanding these details will make you a more sophisticated consumer and producer of genomic data—critical skills as NGS becomes ubiquitous in modern biology and medicine.

8.9 Looking Forward

Sequencing technologies continue to improve. Illumina is developing longer reads and more accurate chemistry. PacBio is improving throughput and reducing costs. Oxford Nanopore is dramatically improving accuracy through better machine learning algorithms.

The next frontiers include: - **Single-cell sequencing**: Sequencing individual cells to understand tissue heterogeneity - **Long-read RNA sequencing**: Capturing full-length transcripts - **Epigenetic profiling**: Mapping DNA modifications alongside sequence - **Real-time clinical sequencing**: Same-day genome sequencing for critically ill patients

NGS didn't just make sequencing faster and cheaper—it fundamentally changed what questions we can ask and answer about genetics, disease, and human diversity. Every chapter that follows in this book builds on the foundation NGS created.

9 Chapter 6. NGS Application for Genomics

9.1 The Fundamental Question: Why Not Always Sequence Everything?

If sequencing technology can read all 3.2 billion base pairs of the human genome, why would we choose to sequence only part of it? It's a bit like asking: if you're looking for a lost key in your house, why not

search every room, every drawer, every pocket? The answer is simple: time, cost, and practicality.

In genomics, we face a similar trade-off. You can sequence:

1. **The whole genome** – all 3.2 billion bases, every gene and every space between genes
2. **Just the exome** – only the ~30–50 million bases that code for proteins (about 1–2% of the genome)

The choice between these approaches has shaped how we discover disease genes, diagnose patients, and conduct research. Let's understand what each approach offers and when to use which one.

9.2 Understanding the Exome: Where Most Disease Variants Hide

Before we compare WGS and WES, we need to understand what the **exome** is and why it's special.

Your genome contains about 20,000 protein-coding genes. But genes aren't continuous stretches of DNA. Instead, each gene is split into:

- **Exons**: the parts that code for proteins (typically 50–200 bases each)
- **Introns**: the non-coding parts between exons (often thousands of bases long)

When cells make proteins, they transcribe the entire gene into RNA, then cut out the introns and splice the exons together. The final messenger RNA (mRNA) contains only exonic sequence, which is then translated into protein.

The collection of all exons in the genome is called the **exome**. It makes up only 1–2% of your total DNA—about 30–50 million base pairs out of 3.2 billion. But here's why it's important:

About 85% of known disease-causing mutations occur in exons.

Why? Because exons directly code for proteins. A single base change in an exon can swap one amino acid for another (missense), create a premature stop signal (nonsense), or shift the reading frame (frameshift). Any of these can break a protein's function, potentially causing disease. Changes in introns or intergenic regions can affect gene regulation or RNA splicing, but they're less likely to have dramatic effects.

This high concentration of disease variants in exons created an opportunity: if most pathogenic variants are in just 1–2% of the genome, we could focus our sequencing there.

9.3 Head-to-Head Comparison: WES vs WGS

Before diving into details, let's see how these approaches compare:

Feature	Whole-Exome Sequencing (WES)	Whole-Genome Sequencing (WGS)
Coverage	~1–2% of genome (exons only)	100% of genome
Target size	30–50 million bases	3.2 billion bases
Data per sample	~6 GB	~90–100 GB
Cost (2024)	~$400–500	~$600–1,000
Typical depth	100–150×	30–40×
SNVs/indels in exons	☐ Excellent	☐ Excellent
Structural variants	☐ Mostly missed	☐ Detected
Non-coding variants	☐ Missed	☐ Detected

Feature	Whole-Exome Sequencing (WES)	Whole-Genome Sequencing (WGS)
Regulatory regions	☐ Missed	☐ Detected
Repeat expansions	☐ Missed	☐ Detected (especially with long reads)
Analysis complexity	Lower—fewer variants	Higher—millions of variants
Interpretation	Easier—focus on coding	Harder—non-coding interpretation uncertain
Diagnostic yield (rare disease)	25–50%	30–55% (slightly higher)
Best for	Known coding disorders, cost-sensitive studies	Complex cases, structural variants, discovery

---

9.4 Whole-Exome Sequencing (WES): The Focused Approach

9.4.1 How It Works

WES doesn't actually sequence only exons—that would be technically challenging. Instead, it uses a clever trick called **target capture** or **enrichment**:

1. Extract DNA and break it into fragments (just like WGS)
2. Add special "bait" molecules—short DNA sequences complementary to exons, labeled with biotin
3. These baits bind to exonic fragments through complementary base pairing
4. Add streptavidin-coated magnetic beads that bind tightly to biotin
5. Use a magnet to pull down the beads with captured exons attached
6. Wash away non-exonic DNA (introns, intergenic regions get discarded)
7. Release and sequence only the captured exonic fragments

Think of it like using a magnet to pick out just the metal pieces from a mixed pile of materials.

Figure: WES Process with Target Capture. This diagram shows the key step that distinguishes WES from WGS: exome enrichment. After DNA extraction and fragmentation, biotinylated probes (baits) specifically bind to exonic sequences. These are then captured using streptavidin-coated magnetic beads, physically separating exonic fragments from the rest of the genome. This targeted approach concentrates sequencing effort on the ~1–2% of the genome that codes for proteins. Source: Microbe Notes

Commercial capture kits from companies like Agilent SureSelect, Illumina TruSeq, and Twist Bioscience each target slightly different exon sets with different capture efficiencies. Not all exons capture equally well—regions with extreme GC content, repetitive sequences, or complex secondary structures may have coverage gaps.

9.4.2 The Historic Success: Miller Syndrome (2010)

WES made its breakthrough in 2010 with a paper that became a model for the approach. Researchers were studying four siblings with **Miller syndrome**—a rare developmental disorder causing facial abnormalities. Traditional gene-hunting methods had failed.

The strategy was simple but revolutionary: 1. Sequence all exons in the four affected siblings 2. Look for variants they share but their healthy parents don't have 3. Focus on rare variants (not common polymorphisms)

They found variants in a gene called **DHODH** that none of the healthy family members carried. DHODH had never been linked to any human disease before. This discovery took months instead of years and cost

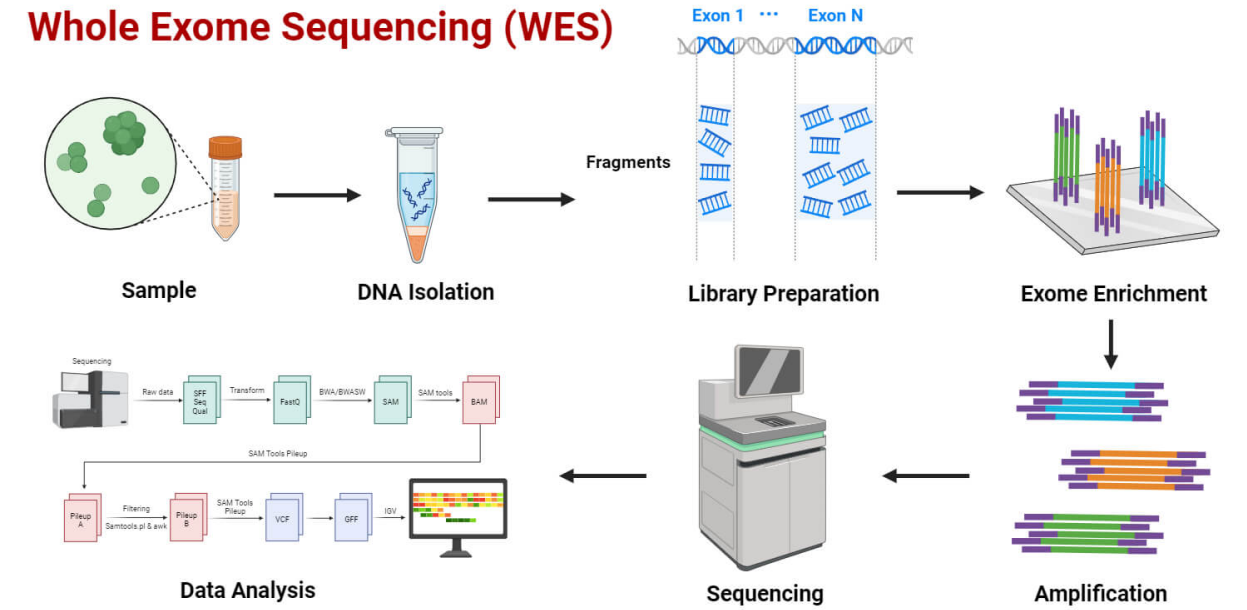


Figure 22: Whole Exome Sequencing Process

thousands instead of millions.

This paper launched the WES era. Suddenly, researchers could identify disease genes for rare disorders by sequencing just a few affected individuals.

9.4.3 Clinical Impact of WES

WES quickly became the first-line genetic test for diagnosing rare diseases, with a diagnostic yield of 25–50%—meaning for patients with suspected genetic disorders, WES identifies the genetic cause in about one-third to half of cases. This is remarkable considering these patients had often undergone years of testing with no diagnosis.

When WES works well: – Rare Mendelian disorders (single-gene diseases) – Diseases known to be caused by coding variants – When you need to screen many genes at once – When cost is a consideration

When WES comes up short: – No variant found in known disease genes – Patient's disease might be caused by non-coding variants – Structural variants might be involved (WES cannot detect large deletions, duplications, or rearrangements) – Coverage gaps in technically challenging exons

---

9.5 Whole-Genome Sequencing (WGS): The Comprehensive Approach

9.5.1 How It Works

WGS is conceptually simpler than WES—you just sequence everything:

1. Extract DNA and break it into fragments
2. Add adapters for sequencing
3. Sequence all fragments
4. Align to the reference genome

No target capture step. No enrichment. Just sequence the whole thing.

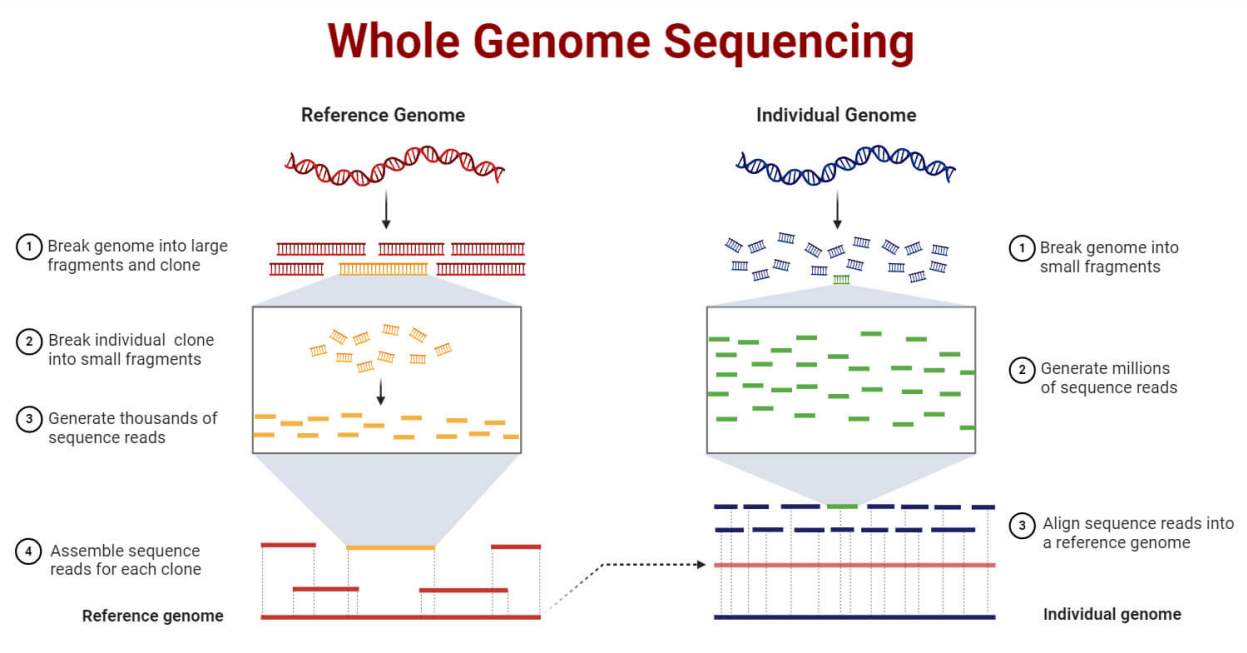


Figure 23: Whole Genome Sequencing Workflow

Figure: Whole-Genome Sequencing (WGS) Workflow. This diagram illustrates the complete process of WGS from sample to analysis. Unlike WES, this process captures all genomic regions including coding exons, introns, regulatory elements, and intergenic sequences. Source: Microbe Notes

9.5.2 The Power of Seeing Everything: Real Examples

Structural variants: Consider a patient with developmental delay and seizures. WES found nothing. With WGS, researchers discovered a large deletion removing three exons of a neurodevelopmental gene. WES had missed this because target capture requires intact DNA—when exons are deleted, there's nothing for the baits to capture. WGS, by sequencing everything, detected the deletion through absence of reads in that region. This happens in about 5–10% of cases where WES fails.

Regulatory variants: Some forms of beta-thalassemia (reduced hemoglobin production) are caused by mutations in the regulatory regions of the HBB gene, not in the coding sequence. These would be missed by WES but are detectable with WGS.

Deep intronic variants: While WES captures exon-intron boundaries, deep intronic variants can create new splice sites or destroy existing ones. These "cryptic splice variants" can cause disease but lie beyond WES's reach.

9.5.3 Non-Coding Variants: The Frontier

WGS opens up the study of non-coding variants, which make up 98% of the genome. Most non-coding sequence is probably functionally neutral, but some regions are critically important. Enhancers and promoters control when and where genes are expressed—a mutation in an enhancer can affect a gene hundreds of thousands of bases away, causing disease without touching the gene itself.

The challenge: We're still learning which non-coding variants matter. The human genome contains millions of non-coding variants, and distinguishing functional from neutral ones is difficult. This is an active area of research.

9.6 The Technical Process: From Sample to Variants

Whether you're sequencing a whole genome or just the exome, the basic workflow is similar—with one key difference for WES: an extra step to capture only the exons. This section walks through the laboratory and computational process that transforms a blood sample into a list of genetic variants.

9.6.1 Step 1: Library Preparation

Library preparation is the process of converting your DNA sample into a form that sequencing machines can read. The workflow differs significantly between short-read platforms (Illumina) and long-read platforms (PacBio, Nanopore), each optimized for their specific sequencing chemistry.

9.6.1.1 A. Illumina Library Preparation (Short-Read Sequencing) Illumina library prep is the most common approach for both WGS and WES, optimized for generating billions of short, highly accurate reads.

□ **Video Resource:** For a detailed walkthrough of Illumina library preparation, watch this Illumina expert tutorial which covers best practices for Nextera library prep.

Overview of the Workflow:

1. DNA Extraction and QC – Input: ~1 μg for WGS, ~200–500 ng for WES – Quality matters: degraded DNA leads to poor results – Check using spectrophotometry (UV absorbance) and fluorometry

2. Fragmentation – **Why fragment?** Illumina sequencers can't read long molecules – **Methods:** Sonication (ultrasound) or enzymatic digestion – **Target size:** 150–500 bp fragments – Result: DNA broken into manageable pieces for short-read sequencing

3. End Repair and Adapter Ligation – DNA fragment ends are "ragged" after fragmentation – Enzymes trim to create clean, blunt ends – **Adapters** (short synthetic DNA sequences) are ligated to both ends: – Provide primer binding sites – Enable attachment to flow cell – Include barcodes for multiplexing (mixing multiple samples)

4. PCR Amplification – Amplify library to ~12–15 cycles – Creates billions of copies for sequencing – **Trade-off:** Introduces some bias (GC-rich/poor regions amplify differently)

5. Target Capture (WES Only!) – This extra step distinguishes WES from WGS – Uses biotinylated probes complementary to exons – Magnetic beads capture probe-bound fragments – Non-exonic DNA is washed away – See WES section above for detailed explanation

6. QC Check – Bioanalyzer/TapeStation: verify fragment size – qPCR/fluorometry: measure concentration – For WES: confirm exon enrichment

Key Advantages: High accuracy (〉99.9%), massive throughput (billions of reads), well-established protocols

Limitations: Short reads struggle with repetitive regions and structural variants

---

9.6.1.2 B. PacBio Library Preparation (Long-Read HiFi Sequencing) PacBio library prep creates circular DNA templates that enable multiple reads of the same molecule for high-accuracy long reads.

□ **Video Resource:** For PacBio and Nanopore library preparation workflows, watch this comparative tutorial.

Overview of the Workflow:

1. **DNA Extraction and QC** – Input: ~500 ng of high–molecular–weight DNA – **Critical:** DNA must be high quality and intact—fragmentation ruins long–read potential – Use gentle extraction methods to preserve long molecules

2. **Fragmentation (Gentle!)** – **Target size:** 15,000–20,000 bp (100× longer than Illumina!) – Uses **controlled shearing** to maintain longer fragments – Can skip fragmentation entirely for ultra–long molecules

3. **End Repair and Hairpin Adapter Ligation** – Clean up fragment ends – **Key difference:** Add **hairpin adapters** to create **SMRTbell templates** – SMRTbell = circular DNA molecule (template loops back on itself) – This circular structure allows the polymerase to read the same molecule multiple times

4. **No PCR Amplification (Usually!)** – PacBio HiFi often sequences **native DNA** directly – Skipping PCR eliminates bias and preserves DNA modifications (methylation) – Enables detection of epigenetic marks

5. **QC Check** – Verify fragment length distribution (critical for long reads) – Check SMRTbell integrity – Quantify library concentration

The HiFi Advantage: The polymerase circles the SMRTbell template 10–20 times, reading the same DNA sequence repeatedly. A consensus is computed from these multiple passes, achieving 〉99.9% accuracy despite long read lengths.

Key Advantages: Long reads (10–20 kb), high accuracy, no PCR bias, detects DNA modifications

Limitations: Lower throughput than Illumina, higher cost per base

---

9.6.1.3 **C. Oxford Nanopore Library Preparation (Ultra–Long Read Sequencing)** Nanopore library prep is the simplest and fastest, enabling ultra–long reads and even portable sequencing.

□ **Video Resource:** See the comparative tutorial mentioned above for Nanopore workflows.

Overview of the Workflow:

1. **DNA Extraction and QC** – Input: Variable (can work with very small amounts) – **Goal:** Extract ultra–high–molecular–weight DNA – Some protocols achieve 〉100 kb, even 〉1 Mb intact molecules

2. **Minimal or No Fragmentation** – Nanopore can sequence **native, unfragmented DNA** – Longer molecules = longer reads – Some protocols skip fragmentation entirely

3. **Adapter Ligation** – **Two approaches:** – **Ligation–based:** Adapters ligated to DNA ends (more accurate) – **Transposase–based (Rapid Kit):** Faster prep, slight bias in AT–rich regions – Adapters contain **motor proteins** that feed DNA through nanopores

4. **No PCR Amplification (PCR–Free)** – Nanopore sequences **native DNA** directly – Preserves all biological information including base modifications – Fastest library prep among all platforms (~10 minutes possible with Rapid Kit)

5. **Minimal QC** – Quick concentration check – Ready to load on sequencer

The Nanopore Advantage: Simplest library prep, real–time sequencing (see data as it's generated), ultra–long reads (〉100 kb, sometimes 〉1 Mb), portable devices (MinION is USB–sized).

Key Advantages: Ultra–long reads, fastest library prep, portable sequencing, real–time results, native DNA sequencing

Limitations: Higher per–base error rate than Illumina/PacBio (though improving rapidly), lower throughput per run

---

9.6.1.4 Comparison: Which Platform for Which Application?

Feature	Illumina	PacBio HiFi	Oxford Nanopore
Library Prep Time	4–8 hours	3–6 hours	10 min – 2 hours
Read Length	150–300 bp	10–20 kb	10–100+ kb
Accuracy	〉99.9%	〉99.9%	95–99%
PCR Amplification	Yes (usually)	No (usually)	No
Best For	WES, high–throughput WGS	Structural variants, phasing	Ultra–long reads, genome assembly
Fragmentation	Required	Controlled	Minimal/none
Special Feature	Target capture for WES	SMRTbell circular templates	Motor proteins, portability

Bottom line: Choose your platform based on your scientific question: – Need billions of cheap, accurate reads? → Illumina – Need to resolve structural variants or phase haplotypes? → PacBio – Need ultra–long reads or portable/rapid sequencing? → Nanopore

---

9.6.2 Step 2: Sequencing

9.6.2.1 Illumina Sequencing (Most Common for WGS and WES) The library is loaded onto a **flow cell**—a glass slide with millions of tiny spots where sequencing happens.

Cluster generation: Library fragments bind to oligonucleotides on the flow cell surface, then through "bridge amplification" create dense clusters of ~1,000 identical copies. Single molecules don't produce enough signal to detect—clusters amplify the signal.

Sequencing by Synthesis (SBS): 1. All four nucleotides (A, C, G, T), each with a different fluorescent dye, are added with DNA polymerase 2. Polymerase adds one base to each growing strand (these are **reversible terminators** that block further addition) 3. A camera records which color (and therefore which base) was added at each cluster 4. The fluorescent tag and terminator are chemically removed 5. Repeat for 150–300 cycles

This produces millions of short reads simultaneously—typically 2×150 bp (paired–end sequencing, reading both ends of each fragment).

Typical output: – WES: ~50–100 million reads (~6–10 GB of data) – WGS: ~1 billion reads (~90–100 GB of data)

9.6.2.2 PacBio Revio Sequencing (For WGS, Especially Structural Variants) PacBio uses **SMRT Cells** containing 25 million tiny wells called **zero–mode waveguides (ZMWs)**. Each well holds a single DNA polymerase molecule.

Single–Molecule Real–Time (SMRT) sequencing: – A single SMRTbell template (circular DNA) is loaded into each ZMW – As the polymerase incorporates bases, it flashes brief pulses of light – The polymerase circles around the template multiple times, reading the same sequence repeatedly – **HiFi reads:** By reading the same molecule 10–20 times and taking a consensus, PacBio achieves 〉99.9% accuracy

Read lengths: 15,000–20,000 bp average, with some exceeding 100,000 bp. These long, accurate reads can span entire genes, resolve structural variants, phase variants, and sequence through repetitive regions.

9.6.2.3 Oxford Nanopore Sequencing (For Ultra-Long Reads) Nanopore sequencing threads DNA through protein pores in a membrane. As bases pass through, they disrupt an electrical current in characteristic ways, enabling ultra-long reads (〉100,000 bp, sometimes 〉1 million bp), real-time results, and portable devices like the USB-sized MinION.

9.6.3 Step 3: Variant Calling – From Reads to Biological Meaning

After sequencing, you have millions or billions of reads. Now comes the computational challenge: figuring out what your genome looks like and how it differs from the reference.

9.6.3.1 3.1 Quality Control of Raw Reads Each base call comes with a quality score (Q score): – Q20 = 99% accuracy (1 in 100 chance of error) – Q30 = 99.9% accuracy (1 in 1,000 chance of error) – Q40 = 99.99% accuracy (1 in 10,000 chance of error)

FastQC checks per-base quality, sequence duplication levels, adapter contamination, and GC content distribution. Low-quality bases (usually at read ends) and adapter sequences are trimmed.

9.6.3.2 3.2 Alignment to Reference Genome Your reads need to be mapped back to their original genomic positions.

Alignment tools: – **BWA (Burrows-Wheeler Aligner)**: Standard for Illumina short reads – **Minimap2**: Designed for long reads (PacBio, Nanopore)

The aligner compares each read to the reference genome (e.g., GRCh38 or T2T-CHM13), finds the best-matching location, and produces a **BAM file** (Binary Alignment Map).

Challenges: Repetitive regions create ambiguity for short reads. Structural variants can cause incorrect alignment or prevent alignment entirely.

9.6.3.3 3.3 Removing PCR Duplicates PCR created many identical copies of some original DNA molecules. Tools like **Picard MarkDuplicates** identify and mark duplicates based on identical mapping positions and sequences. Duplicates inflate coverage artificially and can bias variant calling.

9.6.3.4 3.4 Variant Detection Sophisticated software tools use statistical models to distinguish real variants from sequencing errors.

Popular tools: – **GATK (Genome Analysis Toolkit)**: The gold standard from the Broad Institute – **FreeBayes**: Bayesian variant caller – **DeepVariant**: Uses deep learning/AI from Google

How it works: For each genome position, the caller stacks up all covering reads, counts bases, calculates likelihood of a real variant versus error, considers quality scores, checks depth (20–30+ reads give confidence), and applies filters.

Example: – Position chr1:12345 in reference: G – Your sample: 28 reads show A (all Q30+), 2 reads show G (Q15, Q20) – Conclusion: Likely a real **homozygous SNP** (A/A instead of G/G)

Output: A **VCF file (Variant Call Format)** listing all detected variants with position, reference/alternate bases, quality score, genotype, and read depth.

9.6.3.5 3.5 Filtering Variants Initial variant calls contain false positives. Filtering removes these based on: – Minimum depth (≥10–20 reads) – Quality threshold (≥20 or ≥30) – Strand bias (if all supporting reads come from one DNA strand, likely an artifact) – Allele balance (for heterozygous variants, expect ~50% of reads showing each allele)

Result: A high-confidence variant call set—thousands to millions of variants depending on WGS versus WES.

9.6.3.6 3.6 Variant Annotation **Annotation tools** (VEP, ANNOVAR, SnpEff) add biological context:

Genomic location: Is this in a gene? Which one? In an exon, intron, or intergenic region?

Functional impact: – **Synonymous**: Changes DNA but not amino acid – **Missense**: Changes amino acid – **Nonsense**: Creates a stop codon – **Frameshift**: Insertion or deletion shifts the reading frame – **Splice site**: Affects exon joining

Population frequency: How common is this in databases like gnomAD? Common variants (>1%) are usually benign.

Clinical significance: Is it in ClinVar? What's the classification (pathogenic, benign, VUS)?

Prediction scores: CADD, REVEL scores indicate likely deleteriousness and pathogenicity.

Output: An annotated VCF file, often converted to a spreadsheet for easier interpretation.

9.6.4 Putting It All Together: A Complete Workflow

Day 1–3: DNA extraction and library preparation (WES includes exome capture)

Day 4–5: Sequencing on NovaSeq 6000, generating ~80 million read pairs for WES

Day 6–7: Bioinformatics—QC, alignment, duplicate marking, variant calling and filtering

Day 8–10: Interpretation—filter for rare variants, focus on genes related to symptoms, prioritize high-impact variants, check ClinVar, validate with Sanger sequencing

Result: Genetic diagnosis made in ~10 days, compared to months or years with older approaches.

9.7 The Clinical Decision Tree: Which Test to Order?

9.7.1 Start with WES if:

1. **The patient has features of a known Mendelian disorder** where most causative variants are in coding regions
2. **You need to screen many genes simultaneously** (e.g., hearing loss can involve >100 genes)
3. **Cost matters** for large studies or resource-limited settings
4. **The phenotype suggests a coding variant** like loss of protein function

9.7.2 Move to WGS if:

1. **WES was negative** but you still suspect a genetic cause (WGS finds answers in ~10–30% of WES-negative cases)
2. **You suspect a structural variant** (developmental delay with dysmorphic features, multiple congenital anomalies)
3. **The phenotype is complex or atypical** and might involve regulatory variants
4. **You need comprehensive screening** to rule out all possible genetic causes
5. **You want future-proof data** for reanalysis as we learn more about non-coding regions

9.7.3 A Real Decision-Making Case

Patient: 8-year-old boy with intellectual disability, autism, and dysmorphic facial features

Testing strategy: 1. **Chromosomal microarray** (cheap, quick) → Normal 2. **WES** → Several variants of uncertain significance, nothing diagnostic 3. **WGS** → 150 kb deletion removing first 5 exons of a neurodevelopmental gene

This deletion was missed by WES (no exons to capture) and too small for microarray. WGS provided the answer after other methods failed.

9.8 The Economics and Practicalities

9.8.1 The Cost Trajectory

The cost gap between WES and WGS is shrinking rapidly:

Year	WES Cost	WGS Cost	Difference
2010	~$5,000	~$50,000	10×
2015	~$1,000	~$5,000	5×
2020	~$500	~$1,000	2×
2024	~$400–500	~$600–1,000	1.5×

As costs converge, the argument for WES weakens. Soon, WGS may cost the same as WES.

9.8.2 Storage and Computing

Storage needs: – WES: 6 GB per sample (1,000 samples = 6 TB) – WGS: 90 GB per sample (1,000 samples = 90 TB)

For large biobanks, this 15× difference matters. The UK Biobank sequenced 500,000 genomes—at 90 GB each, that's 45 petabytes requiring serious infrastructure.

9.8.3 The Analysis Bottleneck

Sequencing is fast. Interpretation is slow.

- **WES**: ~20,000 variants per person, mostly coding with known functional impact, can analyze in days
- **WGS**: 4–5 million variants per person including millions of non-coding variants with uncertain interpretation, can take weeks

As one geneticist put it: "We went from being starved for data to drowning in it."

9.9 The Future: Is WES Becoming Obsolete?

There's an interesting argument emerging: **WES might be a transitional technology**.

9.9.1 The "WGS Now, Interpret Exome First" Strategy

You can do WGS but initially analyze only the exonic regions—getting WES-equivalent results while keeping the full dataset for later. This gives you: 1. Immediate WES-level analysis (just look at exons) 2. Option to expand analysis to non-coding regions if exons are negative 3. Future-proofing: As we learn more about non-coding variants, reanalyze the same data

9.9.2 What's Driving the Shift?

1. **Cost convergence**: WGS approaching WES pricing
2. **Improved long-read sequencing**: PacBio and Oxford Nanopore handle structural variants and repeats better

3. **Better interpretation tools**: AI and machine learning helping interpret non-coding variants
4. **Clinical evidence**: WGS diagnostic yield consistently 5-10% higher than WES

9.9.3 But WES Isn't Dead Yet

WES still has advantages: - Large population studies where storage and cost matter at scale - Focused disease studies involving known coding variants - Lower-income settings where every dollar counts - High coverage needs—WES provides deeper coverage of exons for the same cost

---

9.10 A Glimpse of the Long-Read Future

Most WGS and WES today uses short-read sequencing (Illumina), which struggles with repetitive regions, structural variants, and phasing.

Long-read sequencing (PacBio HiFi, Oxford Nanopore) is changing this with reads of 15,000-20,000 bp or 〉100,000 bp that can: - Span entire genes in single reads - Reveal structural variants directly - Phase variants naturally - Sequence through repetitive regions

The T2T-CHM13 genome (the first complete human genome) was built using long reads. As long-read WGS becomes more affordable, it will likely replace short-read approaches.

9.11 Summary: The Bottom Line

Whole-Exome Sequencing: - Focused, cost-effective approach targeting disease-rich coding regions - Great for known Mendelian disorders - Misses structural and regulatory variants - Still widely used clinically but likely a transitional technology

Whole-Genome Sequencing: - Comprehensive, unbiased approach detecting all variant types - More expensive and complex but increasingly preferred as costs drop - The future of clinical and research sequencing

The trend is clear: We're moving toward universal WGS. But for now, both approaches have their place, and the choice depends on your specific question, resources, and needs. The key insight: WES isn't a subset of WGS in practice—it's a different experimental design with different strengths and weaknesses. Understanding when to use each approach is an essential skill for modern geneticists and clinicians.

10 Chapter 7. Annotation and Databases for Genetic Variants

10.1 The Data Deluge Problem

When you sequence someone's genome using NGS, you get millions of differences compared to the reference genome. A typical person has:

- About 4-5 million **single nucleotide variants (SNVs)**
- Around 400,000-500,000 small **insertions and deletions (indels)**
- Thousands of larger **structural variants**

The overwhelming question is: **which ones matter?**

Finding those needle-in-a-haystack variants requires **annotation** (adding meaning to each variant) and **databases** (comparing your findings to what's already known).

10.2 What is Variant Annotation?

Variant annotation adds biological and clinical context to each variant. For every difference you find, annotation helps answer four key questions:

```
THE FOUR ANNOTATION QUESTIONS

1. LOCATION: Gene? Exon? Regulatory region?
2. FUNCTION: Protein change? Splicing effect?
3. POPULATION: Common or rare?
4. CLINICAL: Known disease association?
```

10.3 Annotation Tools

Three major tools automate variant annotation. Each has distinct strengths:

Tool	Primary Strength	Best Used For	Output Example
VEP (Ensembl)	Comprehensive location mapping, regulatory regions	Whole-genome analysis including non-coding regions	"CFTR exon 11, stop-gain, p.Gly542*"
ANNOVAR	Multi-database integration, flexible filtering	Prioritizing variants from large datasets	Adds gnomAD frequencies, conservation scores, disease links
SnpEff	Automatic impact classification	Quick screening for disease-causing variants	HIGH (stop-gain), MODERATE (missense), LOW (synonymous)

In practice, researchers often use multiple tools for cross-validation. For example, finding a CFTR stop-gain mutation: - VEP identifies the precise location and protein change - ANNOVAR adds population frequency data - SnpEff flags it as HIGH impact for prioritization

---

10.4 Variant Databases: The Three-Tier System

Variant databases form a hierarchical information structure: known variants → population frequency → clinical significance.

10.4.1 Tier 1: dbSNP (Variant Catalog)

What it is: The comprehensive catalog of human genetic variation maintained by NCBI since 1998.

Key features: - Over 1.1 billion variant sites - Each variant gets a unique rs number (e.g., rs429358) - Question answered: "Has this variant been documented?"

Usage: Check if your variant has an rs number. If common (⟩5% frequency), likely benign.

10.4.2 Tier 2: gnomAD (Population Reference)

What it is: High-quality sequencing data from 140,000+ individuals across diverse populations.

Key features: - Provides **allele frequencies** by population - Question answered: "How common is this variant?" - Critical logic: Common variants (>1%) can't cause severe childhood diseases

Example: A novel variant in a developmental disorder patient: - In gnomAD at 1% → probably not the cause (too common) - Absent from all 140,000+ individuals → suspicious, investigate further

Why diversity matters: What's common in one population may be rare in another. Population-specific data prevents misclassification.

10.4.3 Tier 3: ClinVar (Clinical Interpretation)

What it is: Clinical variant database linking variants to diseases with evidence-based classifications.

Classification system:

```
Pathogenic ↔ Likely Pathogenic ↔ VUS ↔ Likely Benign ↔ Benign

Causes disease            Uncertain significance      Harmless
```

Clinical workflow: 1. Patient with suspected genetic disease → sequence relevant genes 2. Find variant → check ClinVar classification 3. If "Pathogenic" + clinical match → diagnosis confirmed 4. If "VUS" → additional evidence needed

The VUS challenge: Many variants lack sufficient evidence for classification. As data accumulates, VUSs get reclassified—patients often recheck results annually.

10.5 Predicting Variant Pathogenicity

For variants not yet in ClinVar, computational tools predict likely effects:

Tool	Scoring System	Interpretation Threshold	Best For
CADD	Phred-like scale	>20 = top 1% deleterious>30 = top 0.1%	General variant impact
REVEL	0-1 scale	>0.5 = likely pathogenic	Missense variants in Mendelian diseases
AlphaMissense	AI-based classification	Benign/Ambiguous/Pathogenic	Structure-based predictions for 71M variants

Important limitation: These are predictions, not proof. Functional studies or clinical evidence needed for confirmation.

10.6 Genomic Annotation: The Reference Framework

Before annotating variants, we need a map of the genome showing where genes, exons, and regulatory regions are located.

10.6.1 GENCODE vs. RefSeq

Feature	GENCODE	RefSeq
Coverage	Comprehensive: ~45,000 genes (coding + non-coding)	Conservative: ~20,000 protein-coding genes
Curation	Manual + automated	Primarily manual
Identifiers	ENSG/ENST	NM_/NP_
Best for	Research, non-coding RNA studies	Clinical diagnostics
Updates	Frequent	Less frequent, more stable

Clinical example: BRCA1 founder mutation reported as "NM_007294.3:c.5266dup" using RefSeq coordinates—the standard in clinical settings.

10.7 Biobanks: Connecting Genotype to Phenotype

Biobanks are large collections of DNA samples linked to detailed health information. They answer: "Do people with this variant have higher disease rates?"

10.7.1 Major Biobanks Comparison

Biobank	Scale	Population	Key Strength	Representative Finding
UK Biobank	500,000	UK adults	NHS health records + imaging	Gene-environment interactions (coffee, sleep, heart disease)
All of Us	245,000+	US diversity focus	Addressing health disparities	Population-specific drug metabolism variants
KOVA	5,305	Korean	East Asian variant reference	22.8% better filtering vs European databases
FinnGen	500,000+	Finnish	Founder effect for rare variants	Rare heart/metabolic disorder variants
ToMMo	150,000	Japanese (3 generations)	Gene-environment over time	Post-disaster stress/trauma genetics

Clinical impact: Population-specific databases prevent misclassification. A variant common in East Asians but absent from European databases could be wrongly labeled pathogenic without appropriate reference data.

10.8 Clinical Case: Integrating All Resources

Patient: 4-year-old with developmental delays, seizures, abnormal brain MRI.

Diagnostic workflow:

```
25,000 variants found (whole-exome sequencing)
         ↓
    [Annotation: VEP + ANNOVAR]
         ↓
    ~200 rare variants affecting protein function
         ↓
    [Filter: ClinVar (remove benign) + gnomAD (<0.5%)]
```

```
     ↓
40 candidate variants
     ↓
[Prioritize: Genes causing similar symptoms]
     ↓
3 variants in relevant genes
     ↓
[Predict: CADD, REVEL, AlphaMissense]
     ↓
SCN1A variant identified:
• CADD: 32 (top 0.1%)
• REVEL: 0.89 (pathogenic)
• gnomAD: Not in 140,000+ people
• De novo (new mutation, not inherited)
     ↓
DIAGNOSIS: Dravet syndrome
```

Outcome: – Diagnostic odyssey ended – Treatment optimized (specific anti-seizure medications) – Low recurrence risk for future children – Family connected to support resources

---

10.9 Why This Matters

These databases and tools enable:

Research: – Discover new disease genes – Identify drug targets – Study human evolution

Medicine: – Diagnose rare genetic diseases – Guide cancer treatment – Predict drug response – Assess disease risk

Global Health: – Combat health disparities through diverse reference data – Build collective knowledge that benefits all patients – Understand that most genetic variation is benign—diversity is normal

As sequencing becomes widespread, every genome sequenced adds to our collective knowledge, making the next diagnosis faster and more accurate.

10.10 Key Takeaways

1. **Annotation adds meaning** to millions of variants through four key questions: location, function, population frequency, and clinical significance
2. **Three-tier database system**: dbSNP (catalog) → gnomAD (frequency) → ClinVar (clinical interpretation)
3. **Multiple tools complement each other**: Use VEP, ANNOVAR, and SnpEff together; combine CADD, REVEL, and AlphaMissense for predictions
4. **Population diversity is essential**: Reference data must match patient ancestry to avoid misclassification
5. **Integration is key**: Clinical diagnosis requires combining annotation tools, databases, pathogenicity prediction, and biobank data into a coherent workflow

11 Part 2. Mendelian Genetics in Next-Generation Sequencing: From Peas to Patients

11.1 Introduction: The Same Laws, New Tools

In Part 1, you learned how the Human Genome Project and Next-Generation Sequencing made it possible to read entire genomes—transforming genetics from an observational science into a data science. We can now directly see the DNA sequences that Mendel could only infer from breeding patterns.

But having sequencing data raises new questions: **Which variants cause disease? How are they inherited? Why does one mutant copy sometimes cause disease while two copies are needed for others?**

This chapter explores how Mendel's 150-year-old principles apply to modern genomic data. The laws of segregation and independent assortment still govern inheritance, but now we observe them at the molecular level—tracking specific DNA variants through families, understanding why alleles are dominant or recessive based on protein function, and predicting pathogenicity from population data.

The shift is profound: Mendel saw 3:1 ratios in flower colors and inferred invisible factors. We sequence families, see the exact DNA variants, and watch them segregate according to Mendel's laws—validating his insights while revealing the molecular mechanisms he couldn't observe.

11.2 What This Chapter Is About

This chapter has seven interconnected sections that explain how we interpret variants in the NGS era:

11.2.1 1. Genetic Variants

What types of DNA changes exist (SNVs, indels, structural variants)? How do we distinguish disease-causing mutations from the millions of benign differences using databases like gnomAD and ClinVar? What tools predict pathogenicity when variants are novel?

11.2.2 2. Transmission Through Families

How do we track variants from parents to children in sequencing data? What do Mendel's laws look like when observed directly in sequence reads? How do we identify de novo mutations that weren't inherited from either parent?

11.2.3 3. Dominant Inheritance: Molecular Mechanisms

Why does one mutant copy cause disease in some cases? We'll explore three mechanisms through clinical examples: - **Haploinsufficiency**: One gene copy doesn't produce enough protein (Kabuki syndrome) - **Gain-of-function**: The mutant does something harmful (autism, seizures) - **Dominant-negative**: The mutant interferes with the normal protein (heart disease)

11.2.4 4. Haploinsufficiency and Gene Constraint

Why are some genes sensitive to losing one copy while others tolerate it? How do pLI and LOEUF scores measure constraint using population-scale data? Why does natural selection eliminate loss-of-function variants in critical genes?

11.2.5 5. Recessive Inheritance and Consanguinity

What happens when both gene copies are disrupted? How does marriage between relatives increase autozygosity, revealing recessive alleles? How has studying consanguineous populations accelerated disease gene discovery?

11.2.6 6. Structural Variations

What about large DNA rearrangements—deletions of entire genes, duplications causing dosage imbalance, inversions disrupting gene function? How do they form and why are they harder to detect than SNVs?

11.2.7 7. Integration: From Variant to Diagnosis

How do we combine all these concepts—variant type, inheritance pattern, gene constraint, population frequency, functional prediction—to determine whether a variant explains a patient's disease?

11.3 From Abstract Rules to Molecular Reality

Consider cystic fibrosis, a recessive disorder. The traditional genetic view: – Two carrier parents (Aa × Aa) – 25% unaffected (AA), 50% carriers (Aa), 25% affected (aa) – Classic 3:1 ratio (Mendel's law)

The NGS molecular view: – Sequence the CFTR gene in patient and parents – Find the ⊿F508 deletion (3–bp loss) in the patient's reads: homozygous mutant – Parents' reads show: heterozygous (one normal, one deletion allele) – Mechanism: ⊿F508 removes one amino acid, causing protein misfolding – Confirmation: Both alleles must be disrupted for disease (recessive)

This is the same inheritance pattern Mendel described, but now we see: – The exact molecular lesion (3–bp deletion) – Why it's recessive (protein needs both copies functional) – How it was transmitted (one allele from each parent) – The sequence reads proving the genotypes

This chapter focuses on interpretation: once you have variant calls from NGS data, how do you determine which variants matter and why they cause disease?

11.4 How to Approach This Chapter

11.4.1 Focus on Mechanisms, Not Memorization

Don't memorize that "KMT2D causes Kabuki syndrome." Instead, understand: – Why it's dominant (haploinsufficiency) – How we know (pLI = 1.0, de novo variants in patients) – What this predicts (loss–of–function variants are pathogenic)

11.4.2 Connect Multiple Evidence Types

Interpreting variants requires integrating: – **Frequency**: Rare in gnomAD? (Yes → possibly pathogenic) – **Gene constraint**: High pLI? (Yes → intolerant to loss–of–function) – **Inheritance**: De novo in affected child? (Yes → strong evidence) – **Function**: Predicted damaging? (Yes → CADD 〉30, REVEL 〉0.5) – **Clinical match**: Phenotype fits known gene function? (Yes → diagnosis likely)

No single piece of evidence is sufficient. Pathogenicity is a weight–of–evidence determination.

11.4.3 Recognize What We Don't Know

- ~40–50% of variants remain uncertain significance (VUS)
- Non-coding variants are challenging to interpret
- Incomplete penetrance isn't fully understood
- Novel mechanisms continue to be discovered

The field is evolving rapidly. What's uncertain today may be resolved tomorrow as databases grow.

11.5 What Makes This Different From Classical Genetics

Classical Approach	NGS-Era Approach
Observe phenotype ratios in crosses	Read actual DNA sequences in families
Infer genotypes from breeding patterns	Directly see genotypes in sequence data
Map genes to chromosomes via linkage	Identify causal variants in all genes simultaneously
"This gene is dominant" (observed)	"This variant causes haploinsufficiency" (mechanistic)
Study common Mendelian diseases	Solve rare diseases, understand complex traits
Years per gene	Weeks for all genes

Mendel's laws remain valid—we've just moved from observing their effects to understanding their molecular basis.

11.6 Learning Objectives

By the end of this chapter, you should be able to:

1. **Classify genetic variants** by type (SNV, indel, structural variant) and explain detection/annotation approaches for each
2. **Trace variant transmission** through families using NGS data, identifying inherited vs. de novo patterns
3. **Explain three dominant mechanisms** (haploinsufficiency, gain-of-function, dominant-negative) with disease examples and predict pathogenicity based on mechanism
4. **Interpret gene constraint metrics** (pLI, LOEUF) and understand how natural selection shapes allele frequencies
5. **Recognize recessive inheritance patterns** in sequencing data, explain how consanguinity reveals recessive alleles, and understand biallelic requirements
6. **Describe structural variant mechanisms** (NAHR, mobile elements) and explain their phenotypic impacts
7. **Integrate evidence** (frequency, constraint, inheritance, function, phenotype) to assess variant pathogenicity
8. **Apply Mendelian principles to genomic data**, connecting abstract genetic rules to concrete molecular observations

---

11.7 Let's Begin

Part 1 showed you how to read genomes. Part 2 shows you how to interpret them—how to look at millions of variants and determine which ones explain disease, how they were inherited, and why they have the effects they do.

Mendel discovered the rules by counting pea plants. We validate those same rules by reading DNA sequences. The principles haven't changed. The resolution has.

Let's see how Mendel's laws manifest in modern genomic data.

12 Chapter 8. Genetic Variants

12.1 The Molecular Basis of "Different"

When Mendel crossed purple and white pea plants, he observed predictable patterns but didn't know what caused the difference. Today, we can answer his question at the molecular level: the difference is a change in DNA sequence—a **genetic variant**.

When you sequence a genome, you don't get one string of As, Ts, Gs, and Cs. You get millions of differences from the reference genome. Understanding these differences is the foundation of modern genetics. This chapter explains what variants are, how we classify them, and what effects they have.

12.2 Terminology: Quick Reference

The terms **mutation**, **variant**, and **polymorphism** are often used inconsistently. Here's a practical guide:

Term	Definition	Connotation	When to Use
Variant	Any DNA sequence difference	Neutral	Default term—safe for any difference
Mutation	DNA change, especially rare/new	Often implies disease	Rare changes causing disease
Polymorphism	Common variant (>1% frequency)	Usually benign	Common variation like blood types
SNP	Single Nucleotide Polymorphism	Common, benign	Pronounced "snip"—common single-base changes

The relationship:

```
        All Variants
       /            \
  Mutations      Polymorphisms
 (rare, often    (common, usually
  harmful)         benign)
```

Examples: - "CFTR mutation" = rare disease-causing change (cystic fibrosis) - "ABO polymorphism" = common blood type variants - "Novel variant" = newly discovered difference (unknown significance)

In this chapter, we use "variant" as the default neutral term.

12.3 Variant Classification Framework

Understanding variants requires multiple perspectives. The same variant can be described by its size, location, and effect:

12.3.1 Quick Overview Table

Classification	Types	Detection	Clinical Importance
By Size	SNV, indel, structural variant	SNVs easiest; SVs need long reads	Size affects detection and impact
By Location	Coding, non-coding, regulatory	WES sees coding; WGS sees all	Location determines interpretability
By Effect	LoF, missense, synonymous	Coding effects predictable; non-coding uncertain	Effect determines pathogenicity
By Frequency	Common (>1%), rare (<1%)	All methods	Common variants rarely cause severe disease

---

12.4 Variants by Size

12.4.1 Single Nucleotide Variants (SNVs)

Definition: One base differs from the reference.

```
Reference: ...ATGCGATCG...
Your DNA:  ...ATGCTATCG...
                 ↑
              G→T change
```

Frequency: ~4–5 million per genome (one every 600–800 bases)

Detection: Easiest to detect with both WGS and WES—the "bread and butter" of variant calling

12.4.2 Insertions and Deletions (Indels)

Definition: Small pieces of DNA added or removed (typically 1–50 bp)

Deletion example:

```
Reference: ...ATGCGATCG...
Variant:   ...ATG---TCG... (CGA deleted)
```

Insertion example:

```
Reference: ...ATGCGATCG...
Variant:   ...ATGCAAAGATCG... (AAA inserted)
```

Critical factor—divisible by 3? – Yes (3, 6, 9 bp): In-frame indel—adds/removes whole amino acids – **No (1, 2, 4, 5 bp):** Frameshift—scrambles everything downstream

Frequency: ~400,000–500,000 per genome

Detection: Slightly harder than SNVs, especially in repetitive regions

12.4.3 Structural Variants (SVs)

Definition: Large-scale changes (>50 bp, often >1 kb)

Types: – Large deletions (can remove entire genes) – Duplications (extra gene copies) – Inversions (DNA segment flipped) – Translocations (DNA moved between chromosomes) – Copy number variants (CNVs)

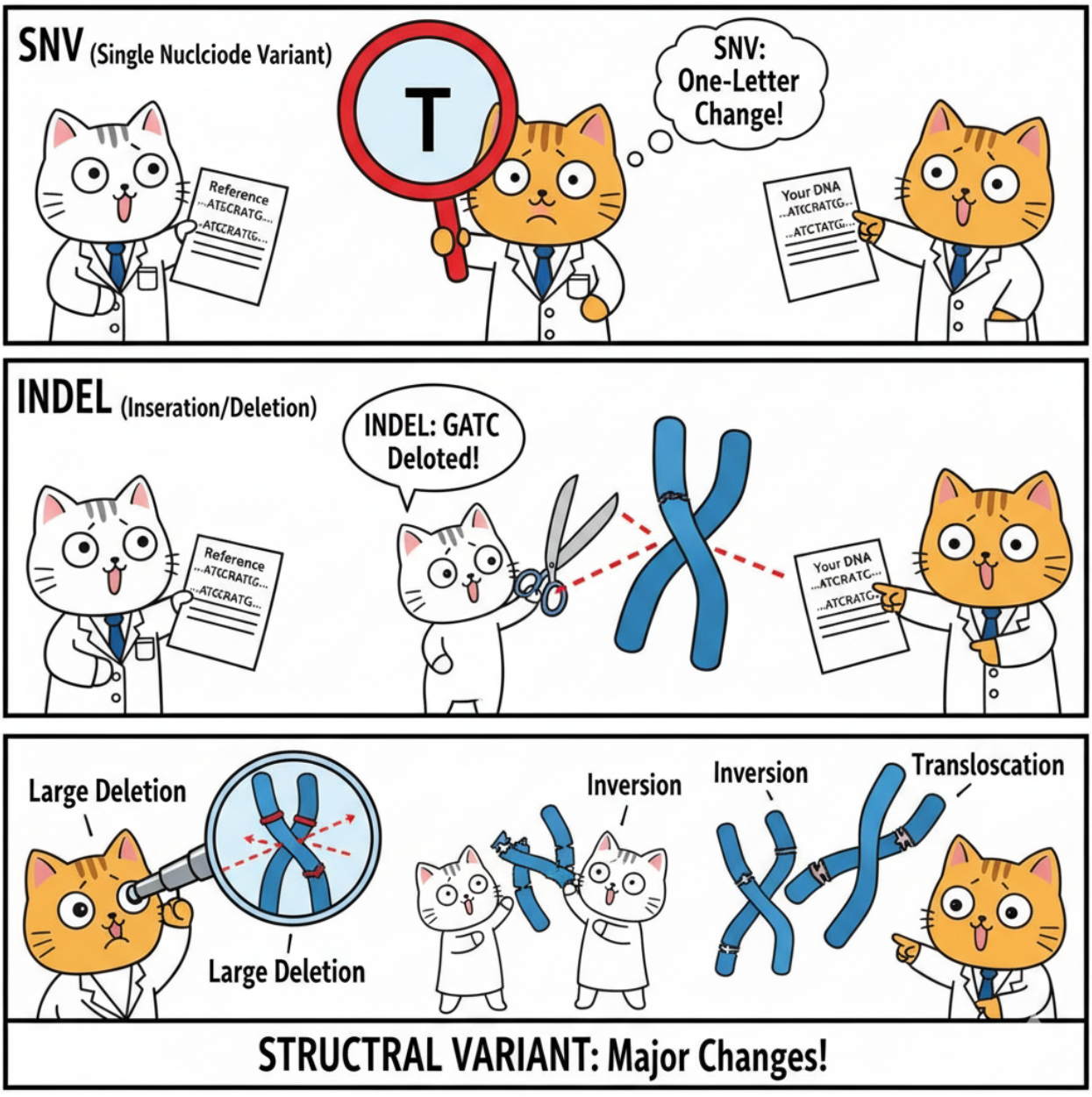


Figure 24: Variant Types

Impact: ~1,000–2,000 SVs per person, affecting more total bases than all SNVs combined

Detection challenge:

Technology	SV Detection Capability
Short-read (Illumina)	Poor—can infer but not confirm
Long-read (PacBio/Nanopore)	Excellent—reads span breakpoints
WES	Very poor—capture requires intact DNA
WGS	Good with long reads

12.5 Variants by Location and Effect

The genome isn't uniform. A variant's effect depends critically on where it lands.

12.5.1 The 98/2 Rule

- **2% coding (exons):** Most known disease variants occur here—effects on proteins are predictable
- **98% non-coding:** Effects on gene regulation are subtle and harder to interpret

Why coding variants are easier: - Direct protein consequence (change amino acid X to Y) - Predictable functional impact - Well-studied in databases

Why non-coding variants are harder: - Indirect effects on gene expression - Tissue-specific or context-dependent - Limited database information

---

12.6 Coding Variants: Five Main Types

All coding variants affect how DNA is translated into protein. Understanding requires knowing the genetic code reads in 3-base units (codons):

```
DNA:     ATG CAT GCA TTG AAA
Protein: Met-His-Ala-Leu-Lys
```

12.6.1 Coding Variant Comparison Table

Type	DNA Change	Protein Effect	Pathogenicity	Example Disease
Synonymous	GAA→GA**G**	Glu→Glu (same)	Usually benign	Rarely disease-causing
Missense	GAA→G**C**A	Glu→Ala (different)	**Variable**	Sickle cell (HBB)
Nonsense	CAG→**T**AG	Gln→**Stop**	Almost always harmful	Many CFTR variants
Frameshift	ATG|CAT|GCA → ATG|CA_|GCA	Scrambles downstream	Almost always harmful	Duchenne muscular dystrophy
In-frame indel	ATG|CAT|GCA|TTG → ATG|CAT|TTG	Removes/adds amino acid(s)	Variable	EGFR in lung cancer

12.6.2 Key Insights

1. Synonymous (Silent): - ~25% of coding changes due to codon redundancy - Usually benign but can affect splicing or translation efficiency

2. Missense (Most Complex): - Effect depends on amino acid properties and location - Conservative changes (similar amino acids) often benign - Radical changes (charged ↔ hydrophobic) often harmful - **Challenge:** Many classified as VUS (Variant of Uncertain Significance)

3. Loss-of-Function (LoF) = Nonsense + Frameshift: - Both produce truncated/absent proteins - Almost always pathogenic (with exceptions for specific genes) - Subject to nonsense-mediated decay (mRNA destroyed)

4. In-frame indels: - Effect depends on what's added/removed and where - Example: EGFR deletions in lung cancer make receptor hyperactive

12.6.3 Visual: Frameshift vs In-Frame

```
Normal:       ATG|CAT|GCA|TTG|AAA
              Met-His-Ala-Leu-Lys

Frameshift    ATG|CA_|GCA|TTG|AAA  (1 base deleted)
(1 bp del):   ATG|CAG|CAT|TGA|AA...
              Met-Gln-His-STOP ← Wrong amino acids, early stop

In-frame      ATG|CAT|___|TTG|AAA  (3 bases deleted)
(3 bp del):   ATG|CAT|TTG|AAA
              Met-His-Leu-Lys ← One amino acid missing, frame preserved
```

12.7 Non-Coding Variants: More Complex, Less Understood

12.7.1 1. Splice Site Variants: Breaking the Cut-and-Paste

Background: Genes have introns (removed) and exons (kept). Splicing happens at precise sequences: - 5' splice site (donor): GT - 3' splice site (acceptor): AG

Canonical splice site variants (GT→AT or AG→AA): - Splicing usually fails catastrophically - Intron retained, exon skipped, or cryptic sites used - Usually causes loss of function

Example: BRCA2 splice site mutations → non-functional protein → hereditary breast cancer

Near-splice variants: - Within ~10-20 bp of splice sites - Can weaken splicing without breaking it - Harder to predict effects

12.7.2 2. Regulatory Variants: Affecting Gene Expression

Types and locations:

Element	Location	Effect	Example
Promoter	Near gene start	Affects transcription initiation	HFE promoter → hemochromatosis
Enhancer	Can be 100+ kb away	Increases gene expression	MYC enhancer → cancer
5' UTR	Before start codon	Affects translation efficiency	Various
3' UTR	After stop codon	Affects mRNA stability, miRNA binding	TP53 3'UTR → cancer risk

Challenge: Unlike coding variants (direct amino acid change), regulatory effects are: - Tissue-specific - Context-dependent - Difficult to quantify (how much expression change matters?) - Hard to connect to distant genes

12.7.3 3. Deep Intronic and Intergenic Variants

Most are neutral but exceptions exist: – Cryptic splice site creation – Hidden regulatory elements – Long-range chromatin interactions

Bottom line: These are hardest to interpret—active research area.

12.8 Clinical Interpretation: Which Variants Matter?

When you sequence a genome and find millions of variants, you need a systematic filtering approach.

12.8.1 The Clinical Filtering Pipeline

```
~4-5 million variants detected
         ↓
Filter 1: Frequency (remove common >1%)
         ↓
~50,000-100,000 rare variants
         ↓
Filter 2: Location (focus on coding + splice sites)
         ↓
~5,000-10,000 protein-affecting variants
         ↓
Filter 3: Gene relevance (phenotype match)
         ↓
~100-500 candidates
         ↓
Filter 4: Effect prediction (LoF, pathogenic missense)
         ↓
~5-20 strong candidates
         ↓
Filter 5: Inheritance pattern + family data
         ↓
1-3 likely causal variants
```

12.8.2 Evidence Integration Framework

Strong evidence for pathogenicity: – □ Nonsense or frameshift in disease-relevant gene – □ Rare (<0.01%) or absent in population databases – □ High constraint gene (pLI >0.9) – □ Segregates with disease in family – □ Matches known inheritance pattern – □ Functional studies confirm effect

Uncertain significance (VUS): – ? Missense in relevant gene – ? Moderate frequency (0.1–1%) – ? Conflicting computational predictions – ? Limited family data – ? Novel variant, no prior reports

Likely benign: – □ Common in population (>1%) – □ Synonymous in non-splice region – □ Multiple computational predictions: benign – □ Seen in healthy individuals with same phenotype

---

12.9 Practical Example: Interpreting a Novel Variant

Case: Child with developmental delays. WES finds a novel variant in KMT2D gene:

Step 1—Check frequency: – gnomAD: Absent from 140,000+ individuals □ (rare)

Step 2—Check variant type: – Nonsense variant (creates premature stop) □ (loss-of-function)

Step 3—Check gene: - KMT2D: Known Kabuki syndrome gene □ (phenotype match) - pLI = 1.0 □ (extremely intolerant to LoF)

Step 4—Check inheritance: - De novo (not in either parent) □ (consistent with dominant disease) - Kabuki is autosomal dominant □ (fits pattern)

Step 5—Check databases: - ClinVar: Other LoF variants in KMT2D classified as pathogenic □ - Literature: Many de novo KMT2D LoF variants cause Kabuki □

Conclusion: Pathogenic variant—likely explains patient's phenotype

---

12.10 WGS vs WES: What Can You Detect?

Variant Type	WGS	WES	Notes
Coding SNVs/indels	□□	□□	Both excellent
Splice sites (canonical)	□□	□□	WES captures exon boundaries
Deep intronic	□□	□	WES misses most introns
Regulatory (promoter/enhancer)	□□	□	WES misses non-coding
Structural variants	□	□/?	WES poor; WGS with long reads best

Clinical decision: - **Start with WES** for suspected Mendelian diseases (coding variants) - **Use WGS** if WES negative or suspect regulatory/structural causes

---

12.11 Key Takeaways

1. **Variants exist at multiple scales:** SNVs (1 bp), indels (1–50 bp), structural variants (〉50 bp)—each requires different detection methods
2. **Location determines interpretability:** Coding variants are easier to interpret than non-coding; 98% of genome is non-coding
3. **Effect on protein predicts pathogenicity:** LoF variants (nonsense/frameshift) almost always harmful; missense variable; synonymous usually benign
4. **Frequency matters:** Common variants (〉1%) rarely cause severe disease due to natural selection
5. **Multiple evidence types needed:** No single criterion sufficient—integrate frequency, effect, gene constraint, inheritance, and functional data
6. **Many variants remain uncertain:** VUS (Variant of Uncertain Significance) is common, especially for missense and non-coding variants
7. **Sequencing strategy matters:** WES excellent for coding variants; WGS needed for comprehensive analysis

---

12.12 Looking Ahead

Understanding variant types is foundational. The next sections explore: – How variants are transmitted through families – Why some variants are dominant and others recessive – How population genetics shapes variant frequencies – How to use variant data to diagnose genetic diseases

All builds on this framework: variants are the molecular basis of alleles, and alleles follow Mendel's rules—but now we see them directly in DNA sequence.

13 Chapter 9. Transmission: How Variants Pass Through Generations

13.1 Two Sources of Your Variants

When you were conceived, you received ~3.2 billion base pairs of DNA—half from each parent. Within that DNA are ~4–5 million variants that differ from the reference genome.

Where did they come from?

Two sources:

1. **Inherited variants** (~99.98%): Already in your parents' genomes, passed through sperm and egg, traveling through populations for generations
2. **De novo variants** (~0.02%): Brand new mutations—occurred for the first time in you, not present in either parent

Understanding this distinction is crucial for genetics, medicine, and evolution. This chapter explores both.

13.2 Part 1: Inherited Variants

13.2.1 Mendelian Segregation at the DNA Level

You received one chromosome copy from each parent at every position. At heterozygous sites (where a parent has two different alleles), you randomly got one or the other.

Example—CFTR gene position:

```
Mother: A/G (heterozygous)    You inherit: Either A or G (50% each)
Father: G/G (homozygous)      You inherit: G (100%)

Your possible genotypes:
→ A/G (50% chance)
→ G/G (50% chance)
```

This is **Mendelian segregation** directly observable in DNA sequences.

13.2.2 Identity by Descent (IBD): Relatedness in DNA

When you inherit a DNA segment, you get an exact copy—**identity by descent (IBD)**. This reveals relationships:

Relationship	% IBD Shared	Why
Parent–child	50%	You got one chromosome copy from them at every position
Siblings	~50% (varies)	Each randomly got one of two parental chromosomes
Grandparent–grandchild	~25%	Two generations of 50% transmission
First cousins	~12.5%	Four generations from common ancestor

Sibling variation explained:

At each position, you and a sibling each randomly received one parental chromosome: – 25%: Same from both parents (identical) – 50%: Same from one parent (half–identical)
– 25%: Different from both (non–identical)

Average: 50% IBD, but actual sharing varies (45–55% typical range)

13.2.3 Recombination: Shuffling Between Generations

Chromosomes don't pass intact. During meiosis, **recombination** (crossing over) exchanges segments between the two parental chromosomes before transmission.

Result: Your mother's chromosome 1 is a mosaic:

```
Position 1-50 Mb:    From maternal grandmother
Position 50-200 Mb:  From maternal grandfather
Position 200-250 Mb: From maternal grandmother
```

Impact: – Creates novel allele combinations – Explains sibling diversity – ~1–2 crossovers per chromosome per generation

13.2.4 Common vs. Rare Inherited Variants

Feature	Common Variants (〉1%)	Rare Variants (〈1%)
Age	Ancient (thousands–millions of years)	Recent (hundreds–thousands of years)
Selection	Survived selection → usually benign	May be deleterious, not yet eliminated
Clinical	Complex trait risk factors	Often cause Mendelian diseases
Example	ApoE4 (~15% frequency, Alzheimer's risk)	Family–specific BRCA1 nonsense mutations

13.3 Part 2: De Novo Variants

13.3.1 What Are They?

De novo variants are mutations that occurred for the first time in you—not present in either parent.

Where they arise: – **Mostly in gametogenesis:** During sperm/egg formation – **Sometimes postzygotic:** In early embryonic divisions (→ mosaicism)

How many: ~50–200 per person, depending on sequencing technology and what's counted

13.3.2 Why They Matter

1. Source of new genetic variation
All diversity started as de novo mutations. Evolution in action.

2. Cause genetic diseases
Many severe developmental disorders: – Autism (often de novo in neurodevelopmental genes) – Achondroplasia (almost always de novo—affected individuals rarely reproduce) – Schizophrenia (increased risk with older fathers)

3. Reveal DNA replication biology
Studying where and when mutations occur teaches us about replication errors and repair mechanisms.

13.4 The Paternal Age Effect

Key finding: Father's age at conception is the dominant factor in de novo mutation rate.

13.4.1 The Biology

Why males contribute more mutations:

Female (Oogenesis)	Male (Spermatogenesis)
All eggs formed before birth	Sperm production starts at puberty
Cells arrested until ovulation	Continuous production throughout life
~23 cell divisions total	Divisions every ~16 days
Minimal age effect	**Strong age-dependent effect**

The math: – Father age 20: ~150 sperm cell divisions – Father age 30: ~230 divisions – Father age 40: ~330 divisions – Father age 50: ~430 divisions

Each division = opportunity for replication errors.

13.5 Quantifying the Effect: Landmark Studies from Iceland

13.5.1 Kong et al. (2012): The First Direct Measurement

Direct mutation counting in 78 Icelandic families revealed fundamental insights into human mutation rates (Kong et al. 2012, Nature).

Core findings:

1. **Mutation rate:** 1.20 × 10□⁸ per base per generation

- Translates to ~63 de novo SNVs per child (father age ~30)

2. **Paternal age effect: +2 mutations per year of father's age**

```
Father age 20: ~55 mutations
Father age 30: ~75 mutations
Father age 40: ~95 mutations
Father age 50: ~115 mutations
```

3. **Parent-of-origin ratio:** 4:1 (father:mother)
    - Father: ~55.4 mutations
    - Mother: ~14.2 mutations (appeared age-independent in this study)
4. **Mutation spectrum:**
    - CpG sites: 18× higher mutation rate
    - Transitions (A↔G, C↔T): 67.8%
    - Transversions: 32.2%
5. **Disease associations:**
    - Schizophrenia risk ↑ with paternal age ($P = 2\times10^{-5}$)
    - Autism risk ↑ with paternal age ($P = 5.4\times10^{-4}$)

13.5.2 Jónsson et al. (2017): Expanded Analysis with 1,548 Families

A more comprehensive study from the same Icelandic research group analyzed **1,548 parent-child trios** (20-fold larger than Kong 2012), identifying 108,778 de novo mutations (Jónsson et al. 2017, Nature). This larger dataset confirmed the paternal age effect and revealed important new findings about maternal contributions.

Updated findings:

1. **Paternal mutations:** +1.51 per year of father's age (95% CI 1.45–1.57)
    - Slightly lower than Kong 2012's estimate of +2 per year, but consistent within confidence intervals
2. **Maternal mutations:** +0.37 per year of mother's age (95% CI 0.32–0.43)
    - **Important discovery:** Maternal age effect exists but is ~4× weaker than paternal effect
    - Kong 2012 reported maternal mutations as age-independent, but this larger study revealed a small but significant maternal age effect
3. **Mutation spectrum changes with maternal age:**
    - CpG>TpG mutations: decrease by 0.26% per year of maternal age
    - C>G mutations: increase by 0.33% per year of maternal age
    - These spectral changes are not uniform across the genome
4. **Regional variation:**
    - A 20 Mb region on chromosome 8p shows up to **50-fold higher maternal C>G mutation rate**
    - Similar enrichment found on chromosomes 2p, 7p, 9p, 16p, and 16q
    - This regional pattern is not limited to specific genes but affects entire chromosomal regions
5. **Parent-of-origin ratio:** 75–81% paternal (3–4:1 ratio, consistent with Kong 2012)

For a visual comparison of paternal vs. maternal age effects across all mutation types, see Figure 1e in Jónsson et al. (2017).

Key implications:

The discovery of maternal age effects and regional mutation hotspots fundamentally changed our understanding of human mutation. While paternal age remains the dominant factor (due to continuous cell divisions in spermatogenesis), maternal age also contributes—particularly for certain mutation types and

in specific genomic regions. The regional C⟩G enrichment suggests distinctive mutational processes in aging oocytes, possibly related to double-strand break repair mechanisms.

13.5.3 Porubsky et al. (2025): Long-Read Revolution

Recent study measured the most accurate rate of de novo variants (including all variant types) using long read sequencing technologies (Porubsky et al. 2025, Nature).

New findings:

1. **Higher total count:** 98–206 mutations/child
    - 74.5 SNVs in unique DNA
    - 65.3 mutations in repetitive DNA (previously invisible)
    - 7.4 small indels
    - 12.4 Y chromosome mutations (per male)
2. **75–81% paternal contribution** (confirmed with more precision)
3. **16% postzygotic mutations** (arose after fertilization)
    - No parent-of-origin bias
    - Can cause mosaicism
4. **Mutation hotspots:** 32 specific regions with recurrent mutations
    - Tandem repeats (polymerase slippage)
    - Centromeres (satellite DNA)
    - Segmental duplications
5. **No recombination–mutation link** (independent processes)

Porubsky et al. 2025, Nature – Figure 2

Figure: De novo mutations in CEPH 1463 pedigree. The figure shows: (a) Number of de novo germline mutations, postzygotic mutations (PZMs), and indels in G2 parents and G3 children, (b) Allele balance comparison showing germline SNVs cluster near 0.50 while postzygotic SNVs show lower allele balance (⟨0.25), (c) Strong paternal age effect for germline de novo SNVs (+1.55 DNMs per year) but not for PZMs, (d) Estimated SNV DNM rates by genomic region, showing significant excess in repetitive regions including centromeres and segmental duplications. Source: Porubsky, D. et al. (2025). Human de novo mutation rates from a four-generation pedigree reference. Nature. https://doi.org/10.1038/s41586-025-08922-2. License: CC-BY 4.0.

13.6 Clinical Implications

13.6.1 For Genetic Counseling

Paternal age considerations:

Father's Age	Additional Risk
⟨30	Baseline mutation rate
30–40	+20 mutations (~35% increase)
40–50	+40 mutations (~70% increase)
⟩50	+60+ mutations (⟩100% increase)

Most mutations are harmless, but risk of specific conditions increases in certain disorders.

13.6.2 For Clinical Diagnosis

When a child has a disorder not present in parents:

1. Look for de novo mutations:

```
Sequence trio (child + both parents)
         ↓
Identify variants in child absent from both parents
         ↓
Prioritize: LoF or damaging missense in disease genes
         ↓
Validate: Check for mosaicism in parents
```

2. Consider sequencing technology: - Short-read (Illumina): Misses repetitive regions - Long-read: Comprehensive, includes centromeres, tandem repeats

3. Check for mosaicism: - Low-level parental mosaicism can explain apparent de novo variants - Important for recurrence risk counseling

---

13.7 Identifying De Novo Variants: The Trio Approach

Standard workflow:

13.7.1 Trio Sequencing Strategy

```
Sequence:
- Proband (affected child)
- Mother
- Father

Compare:
- Variants in child
- Variants in parents

De novo candidates:
- Present in child
- Absent in BOTH parents
- Pass quality filters
```

13.7.2 Quality Control Steps

Challenge: Sequencing errors can look like de novo mutations.

Filters applied:

Filter	Purpose	Typical Threshold
Read depth	Ensure reliable calls	⟩20× coverage
Allele balance	Check heterozygous ratio	30–70% variant reads
Mapping quality	Avoid mismapped reads	MAPQ ⟩30
Parental absence	Confirm not inherited	0 variant reads in parents
Validation	Confirm by Sanger or PCR	For clinical variants

False positive sources: – Sequencing errors (especially at low coverage) – Parental mosaicism (mutation in some parental cells) – Poor reference alignment in complex regions

---

13.8 Evolutionary Perspective

13.8.1 Mutation as the Engine of Evolution

Short-term (per generation): – Each person: ~50–200 new mutations – Most neutral or slightly deleterious – Rare beneficial mutations

Long-term (over millennia): – Beneficial mutations increase in frequency (positive selection) – Deleterious mutations decrease (negative selection)
– Neutral mutations drift randomly

Today's polymorphisms = Yesterday's de novo mutations

The common SNPs distinguishing populations were once de novo mutations in individuals thousands of years ago.

13.8.2 Population-Level Effects

As average paternal age increases:

```
1900s Iceland: Average father age = 34.9 years
1980s Iceland: Average father age = 27.9 years
2010s Iceland: Average father age = 33.0 years
```

Consequence: Population-wide mutation rate fluctuates with demographics.

Debate: Does delayed reproduction increase disease burden? – Pro: More mutations → more genetic disease – Con: Most mutations harmless; modern medicine compensates

---

13.9 Summary Table: Inherited vs. De Novo

Feature	Inherited Variants	De Novo Variants
Frequency	~4–5 million per genome	~50–200 per genome
Source	Parents' genomes	New mutations during reproduction
Age	Can be ancient (millions of years)	Brand new (this generation)
Detection	Standard variant calling	Trio sequencing required
Inheritance	Follow Mendelian segregation	Not inherited (absent in parents)
Clinical	Common → complex traitsRare → Mendelian diseases	Severe developmental disordersDominant de novo conditions
Selection	Already filtered by selection	Haven't faced selection yet

---

13.10 Key Takeaways

1. **Two sources of variants:** Inherited (from parents, following Mendel's laws) and de novo (new mutations in you)
2. **Paternal age effect dominates:** Father's age is the main determinant of de novo mutation rate (~1.5 mutations/year), though maternal age also contributes (~0.4 mutations/year)

3. **Males contribute more:** 75–81% of mutations due to continuous sperm production with repeated cell divisions
4. **Maternal mutations have unique patterns:** Mutation spectrum changes with maternal age, with regional hotspots showing up to 50–fold enrichment
5. **Technology matters:** Long–read sequencing reveals mutations in repetitive DNA invisible to short reads, nearly doubling the total mutation count
6. **Clinical importance:** De novo mutations cause many severe developmental disorders; trio sequencing identifies them
7. **Evolutionary engine:** All genetic diversity started as de novo mutations; today's polymorphisms were once new mutations

---

13.11 Beyond Individual Mutations: An Evolutionary Perspective

We've focused on de novo variants in individuals, but these mutations are also the raw material for evolution. Over long timescales, individual mutations accumulate to create species differences and even entirely new genes.

13.11.1 From Individuals to Species: Time Scales

Scale	Timeframe	Example
Individual	This generation	Your ~50–200 de novo variants
Population	100–10,000 generations	ABO blood types, lactose tolerance
Species	100,000+ generations	Human–chimp fixed differences (~35M SNVs)
New genes	1–10 million years	Gene duplications, de novo gene birth

13.11.2 Three Mechanisms Drive Long–Term Evolution

1. Accumulation of point mutations

The same replication errors causing your de novo variants, accumulated over millions of years, create species boundaries (Montañés et al. 2023, Molecular Biology and Evolution). Humans and chimpanzees differ by ~35 million single–nucleotide changes, each starting as a de novo mutation in some ancestor.

2. Gene duplication creating new functions

Genes can be duplicated by unequal crossing–over or retrotransposition. Usually the duplicate copy degrades, but occasionally it evolves a new function. Classic example: the globin gene family (myoglobin, α–globin, β–globin, fetal γ–globin) all arose from ancient duplications of a single ancestral gene. Each duplication allowed functional specialization—fetal hemoglobin has higher oxygen affinity for placental transfer, while adult hemoglobin optimizes oxygen delivery to tissues.

3. De novo gene birth from non–coding DNA

Entirely new genes can emerge from previously non–coding sequences. Recent comparative genomics studies show many species–specific genes arose this way, though most (~80%) are rapidly lost. Survivors start small and gradually acquire function under selection (Montañés et al., 2023, Molecular Biology and Evolution).

13.11.3 Case Study: FOXP2 and Speech Evolution

The **FOXP2** gene illustrates how point mutations can drive major phenotypic changes. Discovered in a British family with severe speech disorder, FOXP2 encodes a transcription factor critical for speech motor control (Atkinson et al. 2018, Cell).

Human FOXP2 differs from chimpanzee by only two amino acids (at positions 303 and 325). Originally thought to be a recent selective sweep enabling modern human language, genomic studies revealed these changes were already present in Neanderthals, indicating an older origin (⟩300,000 years ago). While no evidence exists for recent selection, these ancient changes likely contributed to vocal motor learning in the human lineage (Atkinson et al., 2018, Cell).

Recent work identified additional speech–related genes (CHD3, SETD1A, WDR5) through whole–genome sequencing of children with speech disorders, revealing that speech acquisition involves a regulatory network of genes co–expressed in the developing brain.

13.11.4 The Connection: Micro to Macro

Your de novo mutations: – ~50–200 new changes this generation – Most neutral or slightly harmful – Will likely disappear within a few generations

Your ancestors' de novo mutations: – Some accumulated to distinguish humans from other primates – Some created new genes through duplication – Some enabled uniquely human traits like speech

The DNA replication errors adding ~1.5 mutations per year of your father's age (and ~0.4 per year of your mother's age) are the same molecular process that, over evolutionary time, created all genetic diversity—from individual differences to the origin of new species.

13.12 Looking Ahead: From Transmission to Pattern

You now understand where variants come from—inherited or arising de novo—and how they connect to evolution across all timescales. The next chapters explore:

Chapter 10—Dominant Inheritance: Why does one mutant copy sometimes cause disease? You'll see three molecular mechanisms explaining dominance.

Chapter 11—Haploinsufficiency: Why are some genes sensitive to losing one copy? Population–scale metrics predict gene constraint.

Chapter 12—Recessive Inheritance: What happens when both copies are mutated? How does family structure reveal recessive alleles?

Understanding transmission is the foundation. Now we examine the patterns those transmitted variants create—the dominant and recessive inheritance that Mendel observed, now explained at the molecular level.

14 Chapter 10. Dominant Allele in Human Disorders

When you hear "dominant allele" in genetics, you might think back to Mendel's peas—where one version of a gene masks the other. In medical genetics, dominance means something more specific: **a single mutated copy of a gene is enough to cause disease**, even when the other copy is perfectly normal. This contrasts with recessive diseases, where you need mutations in both copies to see a phenotype. Why does one bad copy matter so much? Usually, the mutation disrupts the gene's function in a way that one good copy can't fully compensate for. **WES** and targeted sequencing have made it much easier to find these dominant alleles by letting us scan protein–coding regions (or specific non–coding regions) for rare variants that cause disease. This section walks through how dominant alleles cause disease, focusing on two main

types of mutations: **protein–truncating variants (PTVs)** that cut proteins short, and **missense variants** that swap one amino acid for another. We'll look at real examples from Kabuki syndrome, autism, seizures, congenital heart disease, and neurodevelopmental disorders caused by mutations in small RNA genes. Each example shows how sequencing, combined with smart filtering for rare variants and predictions of which changes are harmful, helps us connect a dominant allele to a disease. ## Molecular Mechanisms of Dominant Alleles So how does a single mutant copy—in what we call a **heterozygous** genotype (one mutant, one normal)—cause disease? Two main mechanisms: **Protein–Truncating Variants (PTVs)**: These are mutations that chop proteins short. **Nonsense mutations** introduce a premature stop sign in the genetic code, while **frameshift mutations** shift the reading frame, scrambling everything downstream. Both produce incomplete, usually nonfunctional proteins. The result is often **haploinsufficiency**—one good copy can't make enough protein to keep things running smoothly. Think of a developmental gene: if you don't have enough of its product, critical processes during embryonic development can go wrong, leading to birth defects. **Missense Variants**: These swap out a single amino acid in the protein. Depending on where and what gets swapped, this can have dramatic effects. Some missense mutations create a protein with a new, harmful function—a **gain–of–function**. Others produce a defective protein that gums up the works by interfering with the normal version—a **dominant–negative effect**. For instance, a missense mutation in an ion channel gene might disrupt the channel's ability to open and close properly, throwing off cellular signaling in neurons and causing neurological problems. These mechanisms explain why dominant alleles are, well, dominant: one mutant copy can overpower the normal one, affecting everything from embryonic development to brain function to heart formation. Let's see how this plays out in specific diseases.

14.1 Kabuki Syndrome: PTVs in KMT2D

Kabuki syndrome is a developmental disorder that shows us PTVs in action. **WES** was used to figure out what was going wrong at the genetic level, revealing that mutations in the KMT2D gene (formerly called MLL2) were the culprit—and these mutations behave as dominant alleles (Ng et al 2010, Nature Genetics).

14.1.1 Finding KMT2D Mutations with WES

The study started with ten people who had Kabuki syndrome but were unrelated to each other. WES was used to sequence all the protein–coding parts of their genomes—about 1–2% of the total genome, but the part where most disease–causing mutations hide. To zero in on disease–causing variants, **allele frequency filtering** was applied: any variant that was common in databases like dbSNP129 and the 1000 Genomes Project was removed. The logic? If a variant causes a serious developmental disorder, it shouldn't be common in the general population. This filtering pointed straight to KMT2D, where novel mutations were found in nine out of ten individuals. To confirm this wasn't a fluke, KMT2D was sequenced in 43 more people with Kabuki syndrome, with mutations found in 26 of them. That's pretty convincing evidence that KMT2D is a major player in this disease.

What kinds of mutations were found? Mostly **nonsense** and **frameshift** mutations—both PTVs. Nonsense mutations turn a "keep going" signal into a "stop here" signal, while frameshifts shift the three–letter reading frame of the genetic code, producing gibberish. The result in both cases: a truncated, useless protein instead of the full–length histone methyltransferase that KMT2D normally encodes.

See Figure 1 from Ng et al., which maps out where these mutations sit in the KMT2D gene. The gene is huge—54 exons—with several important functional domains. Of the 32 mutations shown, 20 are nonsense mutations and 7 are frameshifts. Notice how most of them cluster before the SET domain at the end of the gene, which is critical for the enzyme's methylation activity. If you cut the protein off before you get there, you lose function.

Interestingly, seven of the nine mutations in the initial group were **de novo**—they appeared spontaneously in the affected individuals and weren't present in either parent. The other two were inherited from an affected parent. Both scenarios make sense for a dominant disorder: you only need one mutant copy to have the disease.

14.1.2 How It Causes Disease

The mechanism here is **haploinsufficiency**. KMT2D encodes a histone methyltransferase—an enzyme that adds methyl groups to histones, the proteins that DNA wraps around. This modification is like a dimmer switch for genes: it helps control which genes are turned on or off during development.

When you have a PTV in one copy of KMT2D, you lose about half of the enzyme. That doesn't sound too bad, but development is incredibly sensitive to gene dosage. One good copy just can't make enough enzyme to maintain normal gene expression patterns. The result? Abnormal development, particularly affecting the face, skeleton, heart, and brain—the hallmarks of Kabuki syndrome. Look at Figure 1 again: most PTVs hit before the SET domain, meaning the truncated protein is missing the key part that does the methylation. Even if the cell tries to make protein from the mutant allele, it's useless. You're left with just the output from one functioning allele, and that's not enough.

This is textbook dominant inheritance: one mutant copy is sufficient to cause disease. The prevalence of de novo mutations explains why most cases are sporadic (parents are unaffected, but the child has the disorder). When mutations are inherited, it's from an affected parent, confirming the dominant pattern. WES, combined with careful filtering and validation, made it possible to nail down KMT2D as the culprit gene, showing how dominant PTVs can derail development by disrupting critical molecular processes.

14.2 Autism and Seizures: Diverse SCN2A Variants

The SCN2A gene gives us a more complex story. It encodes NaV1.2, a sodium channel protein that's essential for neurons to fire properly. Mutations in SCN2A can cause several different neurological problems—autism spectrum disorder (ASD), benign infantile familial seizures (BIFS), and epileptic encephalopathy (EE)—depending on exactly how the mutation affects the protein. Again, we're talking about dominant alleles: one mutant copy in a heterozygous individual is enough to cause disease. **WES** was used to find these mutations, and their effects were then tested in cells to understand how they work (Ben-Shalom et al 2017, Biological Psychiatry).

14.2.1 Finding SCN2A Variants with WES

The study used WES on people with ASD, BIFS, and EE, focusing on protein-coding regions where disease-causing mutations are likely to hide. To filter out noise, **allele frequency filtering** was applied: any variant that showed up more than 10 times per million alleles (frequency $> 10^{-5}$) in the ExAC database was excluded. This meant the focus was on extremely rare variants—the kind that might cause severe disease. For **missense variants** (which change one amino acid), prediction tools like SIFT and PolyPhen were used to figure out which ones were likely to be harmful. These tools look at how important each amino acid is based on evolution—if an amino acid has been conserved for millions of years across species, changing it is probably bad news.

What was found? In ASD, about half the mutations were PTVs (52%) and half were missense variants (48%), and almost all were **de novo**—new mutations in the affected child. In BIFS and EE, the vast majority (95-96%) were missense variants. BIFS variants were often inherited from a parent who also had seizures, while EE variants were mostly de novo.

These mutations map onto the NaV1.2 protein structure in a revealing pattern. The protein has four repeating domains (I-IV), each with six transmembrane segments. ASD variants (PTVs in blue, missense in red) cluster around the pore loop—the channel's ion selectivity filter. BIFS variants (yellow) and EE variants (green) cluster in the voltage sensor region (segments 4-5), which controls whether the channel opens or closes. The named variants like C959X (a nonsense mutation causing ASD) and R1319Q (a missense mutation causing BIFS) show how different types of mutations in different locations lead to different diseases.

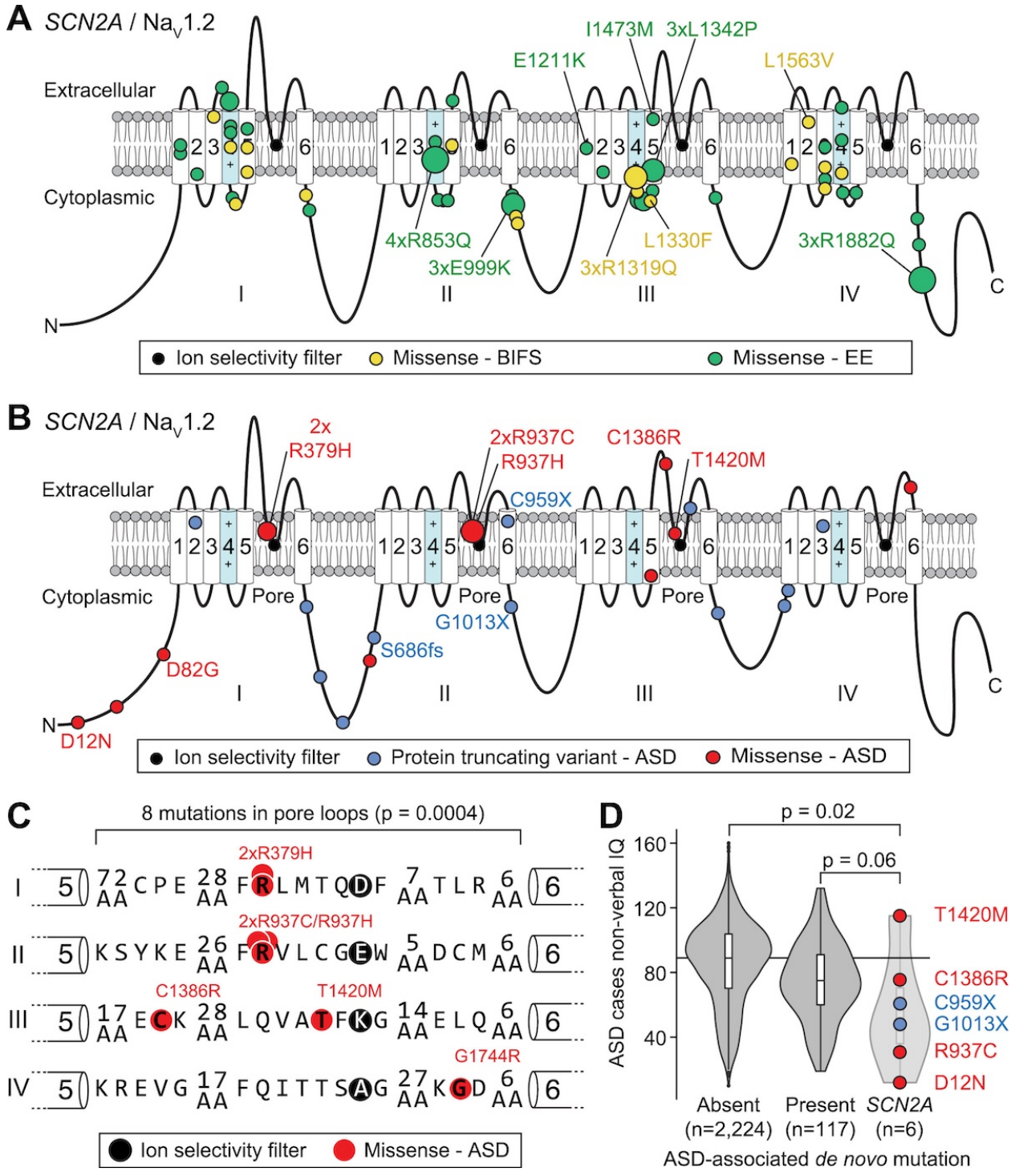


Figure 25: Ben-Shalom et al. 2017, Biological Psychiatry – Figure 1

Figure: SCN2A/NaV1.2 genotypes and phenotypes. This figure maps SCN2A variants onto the NaV1.2 sodium channel protein structure. PTVs (blue) and missense variants (red) associated with ASD, missense variants for BIFS (yellow) and EE (green) are shown across the four protein domains (I–IV). Circle size indicates the number of individuals carrying each variant. ASD variants cluster in the pore loop region (the channel's ion selectivity filter), while BIFS and EE variants cluster in the voltage sensor segments. This location difference explains why they cause different diseases—ASD variants reduce channel function, while BIFS/EE variants increase it. Source: Ben-Shalom, R. et al. (2017). Opposing Effects on NaV1.2 Function Underlie Differences Between SCN2A Variants Causing Intellectual Disability or Infantile Seizures. Biological Psychiatry. https://www.biologicalpsychiatryjournal.com/article/S0006-3223(17)30041-0/fulltext. License: CC-BY-NC-ND.

14.2.2 How They Cause Disease

Here's where it gets interesting. The study didn't just find the mutations—it tested them in cells to see what they actually do to the channel. In ASD, PTVs like C959X cause **haploinsufficiency**: you lose half your functional sodium channels because one allele makes a truncated, useless protein. But the ASD missense variants (like R379H in the pore loop) also reduce channel function, just through a different mechanism—**loss-of-function**. The mutant channel either doesn't conduct sodium ions properly or doesn't get to the cell surface. Either way, neurons don't fire as readily as they should during development, which apparently contributes to ASD.

In contrast, BIFS and EE missense variants (like R1319Q in the voltage sensor) cause **gain-of-function**: the mutant channel is hyperactive. It opens too easily or stays open too long, making neurons fire excessively. The result? Seizures. The difference between BIFS (relatively mild seizures in infancy that usually go away) and EE (severe, treatment-resistant seizures) seems to relate to how much the channel is hyperactivated and when in development the mutation exerts its effects.

These opposing effects are illustrated schematically below. In ASD (top), PTVs and loss-of-function missense variants reduce sodium current, decreasing neuronal excitability. In BIFS and EE (bottom), gain-of-function missense variants increase sodium current, boosting excitability and causing seizures.

Figure: Schematic of NaV1.2 channel function in SCN2A variants. In ASD, PTVs and missense variants reduce or eliminate sodium current (haploinsufficiency or loss-of-function), decreasing neuronal excitability. In BIFS and EE, missense variants increase sodium current (gain-of-function), making neurons hyperexcitable and causing seizures. Same gene, opposite functional effects, different diseases. Source: Ben-Shalom, R. et al. (2017). Opposing Effects on NaV1.2 Function Underlie Differences Between SCN2A Variants Causing Intellectual Disability or Infantile Seizures. Biological Psychiatry. https://www.biologicalpsychiatryjournal.com/article/S0006-3223(17)30041-0/fulltext. License: CC-BY-NC-ND.

Why are these dominant alleles? Because one mutant copy is enough to cause disease. In ASD, you lose half your channel function from PTVs, or the mutant missense protein disrupts things even in the presence of the normal protein. In BIFS and EE, the hyperactive mutant channels throw off the balance even though you still have normal channels from the other allele. The mutant allele dominates the cellular phenotype.

The prevalence of de novo mutations in ASD and EE explains why these are usually sporadic cases—the parents don't have the mutation. In BIFS, inherited mutations show classic dominant inheritance: an affected parent passes it to a child. WES, combined with functional testing in cells, revealed how the same gene can cause different diseases depending on where and how it's mutated—a great example of genetic complexity.

14.3 Congenital Heart Disease: Multiple Dominant Genes

Congenital heart disease (CHD)—structural heart defects present at birth—can be caused by mutations in many different genes. A massive study looked at over 11,000 people with CHD using **WES** and targeted

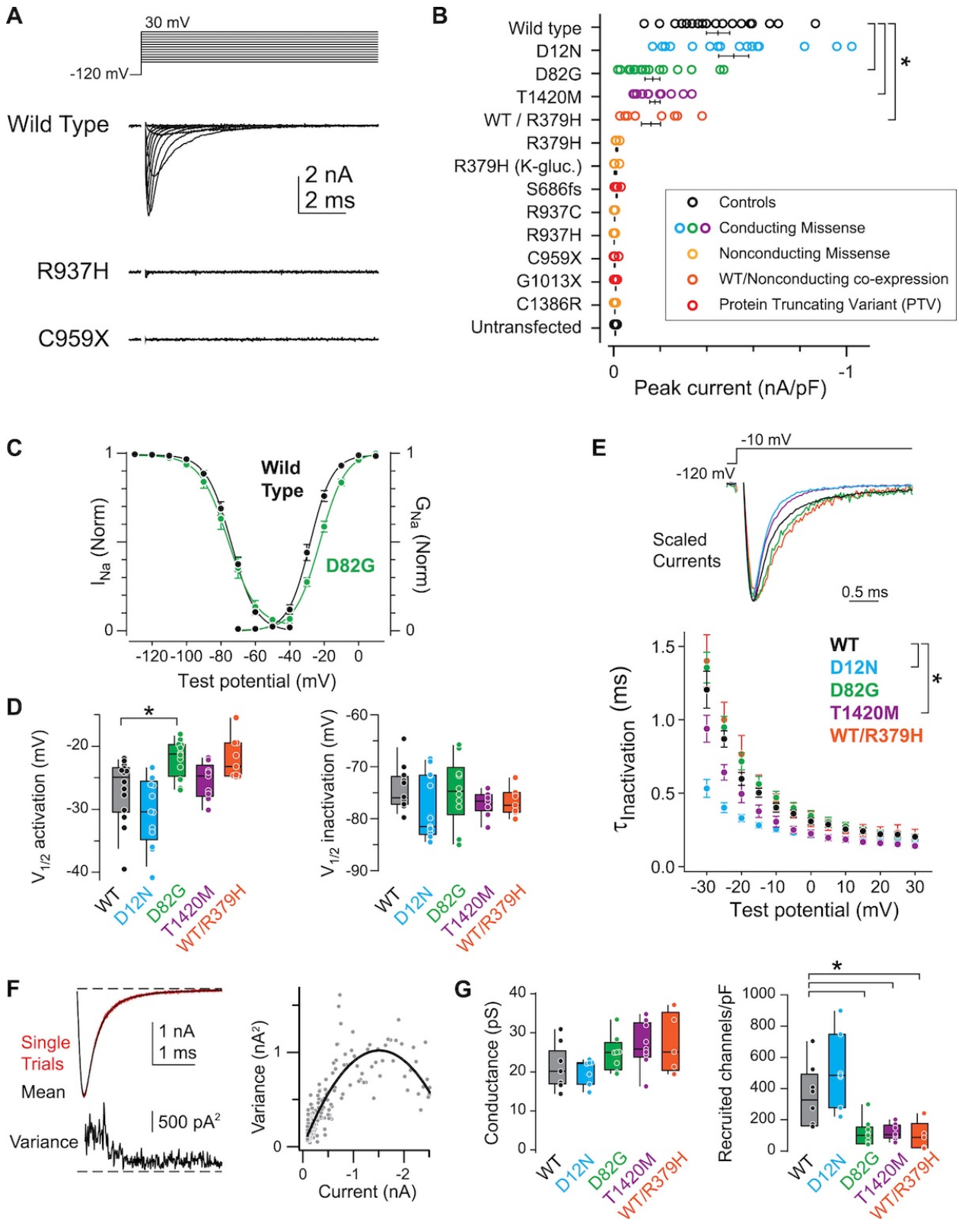


Figure 26: Ben-Shalom et al. 2017, Biological Psychiatry - Figure 2

sequencing, identifying 60 genes where rare variants significantly increased disease risk (Sierant et al 2025, PNAS).

14.3.1 Finding CHD Genes with WES and Targeted Sequencing

The study analyzed 11,555 people with CHD, using WES to scan the whole exome and targeted sequencing to zoom in on 248 genes already suspected of causing heart defects. **Allele frequency filtering** focused on rare variants—those with frequencies below 10^{-5} (one in 100,000 or rarer) in gnomAD and BRAVO databases. For missense variants, SIFT and PolyPhen were used to predict which ones were likely damaging. This effort identified 60 genes with significant burdens of PTVs and damaging missense variants, explaining about 10% of CHD cases. About half of the damaging variants were **de novo** (new mutations in the child), and half were inherited from a parent. Three genes stood out as major contributors: NOTCH1, KMT2D (the same Kabuki syndrome gene—it also causes heart defects), and MYH6. NOTCH1 missense variants were particularly common in tetralogy of Fallot, a specific type of CHD. KMT2D PTVs caused both heart and neurodevelopmental problems. MYH6 missense variants often caused isolated heart defects without affecting the brain.

The relationship between CHD and neurodevelopmental disorders (NDD) varies dramatically across these 60 genes. Some genes, like MYH6, rarely cause NDD (about 4% risk)—the mutation mainly affects the heart. Others, like CHD7, almost always cause NDD in addition to heart defects (95% risk). NOTCH1 and KMT2D fall in between. This spectrum makes sense: some genes are expressed mainly in the heart, while others are critical for both heart and brain development.

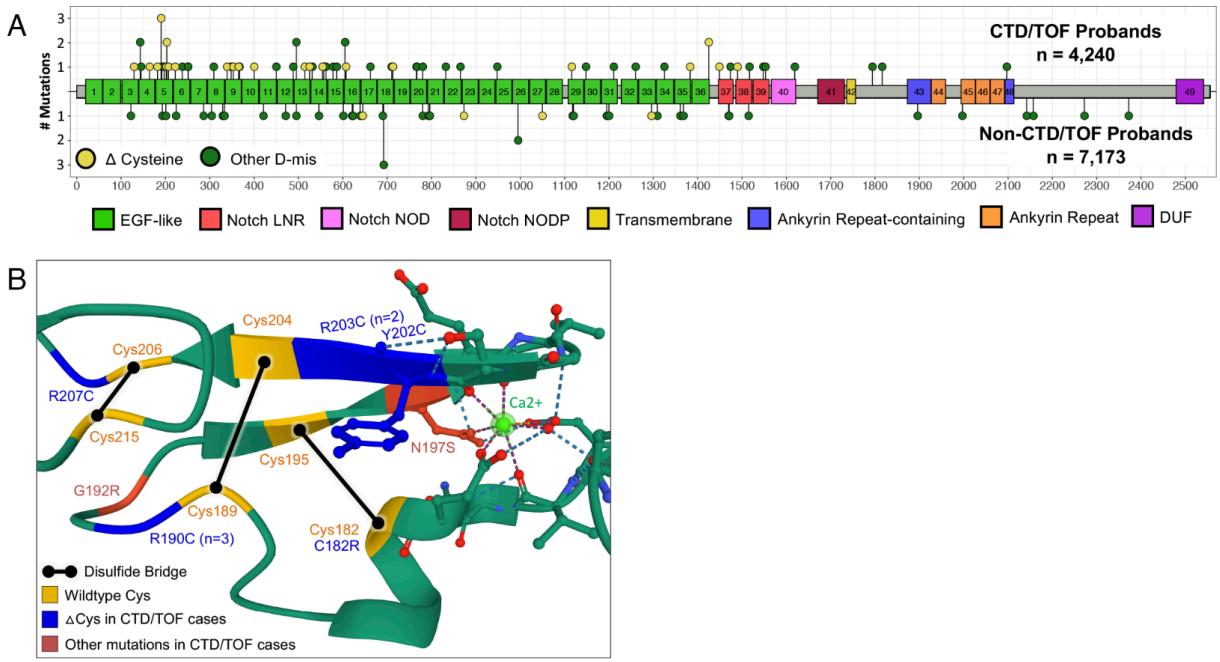


Figure 27: Sierant et al. 2025, PNAS – Figure 3

Figure: CHD genes and neurodevelopmental disorder risk. This figure shows the relationship between the 60 identified CHD genes and the risk of neurodevelopmental disorders (NDD). Genes like MYH6 (around 4% NDD risk) primarily affect the heart through missense variants that disrupt cardiac myosin function. At the other end, genes like CHD7 (around 95% NDD risk) cause syndromic CHD with severe neurodevelopmental problems through PTVs that impair chromatin regulation. NOTCH1 (missense variants in signaling domains) and KMT2D (PTVs causing haploinsufficiency) fall in between, affecting both cardiac and neural development. This demonstrates how dominant alleles in different genes can have varying effects

beyond the heart, depending on where and when the gene is normally active. Source: Sierant, M.C. et al. (2025). Congenital heart disease genetics uncovers context-dependent organization between genome and phenome. PNAS. https://www.pnas.org/doi/10.1073/pnas.2420343122. License: CC-BY.

14.3.2 How They Cause Disease

The three main mechanisms mirror what we've seen before:

Haploinsufficiency (KMT2D): PTVs in KMT2D produce truncated, nonfunctional histone methyltransferases. One working copy can't make enough enzyme to properly regulate the genes needed for heart development. KMT2D mutations also frequently cause neurodevelopmental problems because the same chromatin regulation is critical in the developing brain.

Dominant-negative effects (NOTCH1): Many NOTCH1 missense variants, especially those affecting cysteine residues in EGF-like domains, produce defective signaling proteins. These mutant proteins interfere with normal NOTCH signaling, which is essential for cardiac cell differentiation during heart formation. Even with normal NOTCH1 protein from the other allele, the defective protein disrupts the pathway enough to cause structural heart defects like tetralogy of Fallot.

Altered protein function (MYH6): Missense variants in MYH6, which encodes cardiac myosin, likely alter the motor protein's ability to contract properly. The exact mechanism might involve gain-of-function (abnormal contractile activity) or dominant-negative effects (interfering with normal myosin), but either way, one mutant copy is enough to disrupt heart muscle development and function.

The statistical enrichment of PTVs and damaging missense variants across these 60 genes confirms that a single mutant allele significantly increases CHD risk. In a heterozygous individual, one bad copy of KMT2D, NOTCH1, or MYH6 is sufficient to disrupt heart development. De novo mutations explain many sporadic CHD cases, while inherited mutations (especially in MYH6) show classic dominant inheritance patterns. This study demonstrates the power of large-scale sequencing: by analyzing over 11,000 individuals, dozens of dominant CHD genes could be statistically validated, improving genetic diagnosis and revealing the molecular diversity underlying congenital heart defects.

14.4 Neurodevelopmental Disorders: Non-Coding Variants in RNU4-2, RNU5B-1, and RNU5A-1

All the examples so far have been in protein-coding genes. But what about mutations outside of exons? A striking example comes from neurodevelopmental disorders caused by mutations in small nuclear RNA (snRNA) genes like RNU4-2, RNU5B-1, and RNU5A-1. These genes don't encode proteins—they encode RNA molecules that are part of the spliceosome, the cellular machine that cuts and pastes pre-mRNA to make mature mRNA. Genome sequencing (not WES, since these genes are non-coding) was used to identify these mutations in 23,649 people with rare disorders (Nava et al 2025, Nature Genetics).

14.4.1 Finding snRNA Variants with Genome Sequencing

Since WES only covers protein-coding regions, full genome sequencing was needed to examine snRNA genes. **Allele frequency filtering** focused on rare variants—specifically, those with 10 or fewer alleles in the gnomAD database, making them extremely rare in the general population. For **pathogenicity** assessment, traditional protein prediction tools like SIFT don't work because there's no protein. Instead, conservation scores (is this nucleotide unchanged across species?) and structural analysis (does the mutation disrupt the RNA's folded structure?) were used. snRNAs fold into specific 3D shapes to function properly, and mutations that disrupt these structures are likely harmful.

The study found 145 people with pathogenic variants in RNU4-2 (causing something called ReNU syndrome) and 21 with variants in RNU5B-1 or RNU5A-1. Almost all were **de novo** mutations—new in the affected child. The variants were mostly single nucleotide changes or small insertions. The pathogenic

variants in RNU4-2 map onto the U4 snRNA secondary structure, clustering in critical regions like stem III and the T-loop/quasi-pseudoknot—parts of the RNA that are essential for the spliceosome to work. One insertion, n.64_65insT, was found repeatedly in 78% of RNU4-2 cases, suggesting it's a hotspot.

Figure: Locations of RNU4-2 variants in U4 snRNA. This figure maps pathogenic variants in RNU4-2 onto the U4 snRNA's secondary structure (how the RNA folds). The recurrent insertion n.64_65insT and other variants cluster in stem III and the T-loop/quasi-pseudoknot—regions critical for the U4/U6 interaction needed for splicing. These structural disruptions explain how non-coding variants act as dominant alleles by breaking the spliceosome machinery. Source: Nava, C. et al. (2025). Pathogenic variants in spliceosomal snRNA genes cause neurodevelopmental disorders. Nature Genetics. https://www.nature.com/articles/s41588-025-02184-4. License: CC-BY-NC-ND 4.0.

14.4.2 How They Cause Disease

These variants disrupt RNA structure and function directly. The mutant snRNA is incorporated into spliceosomes, where it interferes with pre-mRNA splicing. For example, the RNU4-2 n.64_65insT insertion alters the T-loop structure, disrupting how U4 interacts with U6 (another snRNA component). This impairs recognition of splice sites in pre-mRNA, leading to mis-splicing across the genome. The result? Thousands of genes produce abnormal mRNA isoforms, many of which encode nonfunctional proteins or no protein at all.

RNA sequencing showed that people with RNU4-2 variants have altered splicing patterns, particularly at 5′ splice sites. Similarly, RNU5B-1 and RNU5A-1 variants disrupt U5's role in aligning exons during splicing. Unlike the haploinsufficiency we saw with KMT2D or the gain-of-function in SCN2A, this is more like throwing a wrench into a complex machine. The mutant snRNA acts as a dominant allele not because there isn't enough functional RNA (there's still one good copy), but because the defective RNA actively disrupts spliceosome function. The result is global splicing errors that particularly affect the developing brain, causing neurodevelopmental disorders like ReNU syndrome. Why dominant? One mutant copy of RNU4-2 produces enough defective snRNA to mess up splicing, even in the presence of normal snRNA from the other allele. The prevalence of de novo mutations explains why these disorders are usually sporadic. The recurrent n.64_65insT insertion shows that this particular change is especially potent—one bad allele is all it takes. This example expands our understanding of dominant alleles beyond proteins to include non-coding RNAs, showing that WES alone isn't enough for comprehensive genetic diagnosis. Genome sequencing is necessary to catch these non-coding culprits.

14.5 Genetic Variant as Allele Leading to Phenotypic Changes

Let's zoom out for a moment and connect this back to what Mendel discovered with his peas. Mendel figured out that traits are determined by **alleles**—different versions of a gene. In his experiments, a purple flower allele dominated over a white flower allele. In modern terms, those alleles are **genetic variants**—specific DNA sequences at a genomic location that differ from a reference sequence or from other versions in the population. A genetic variant is an allele. In dominant disorders, one mutant allele in a **heterozygous** individual (one mutant, one normal) is sufficient to cause disease. According to the **central dogma** of molecular biology (DNA → RNA → protein → cellular function → tissue function → organism phenotype), changes in DNA propagate through this cascade to produce observable traits or diseases. This section traces how dominant alleles travel through the central dogma, from variant to phenotype, using our examples.

14.5.1 Variants as Alleles in the Central Dogma

A genetic variant acts as an allele by changing the DNA sequence, which then affects everything downstream. In dominant disorders, one mutant allele disrupts this molecular cascade enough to cause disease, even with a normal allele present. In Kabuki syndrome, a mutant KMT2D allele alters the DNA sequence,

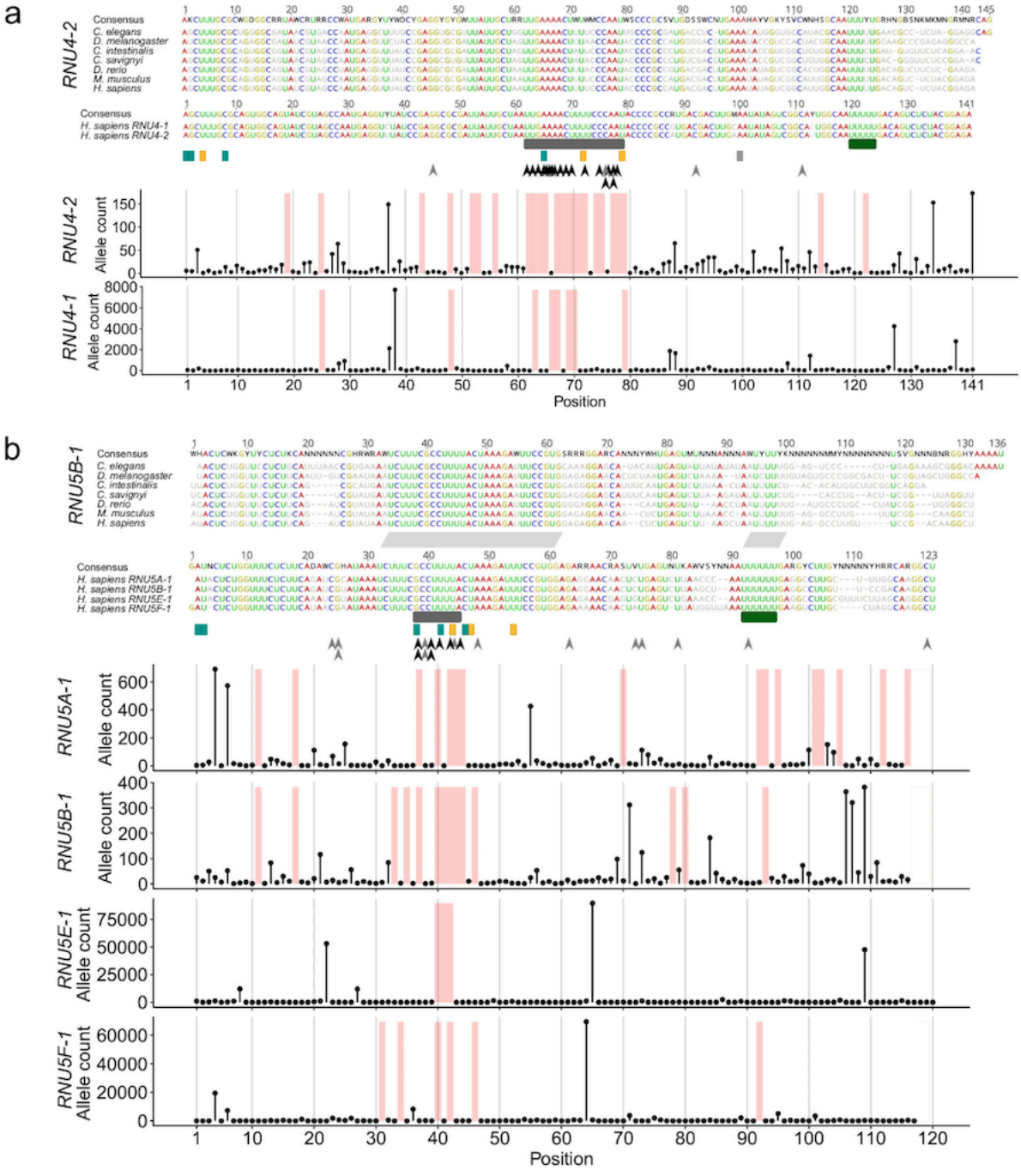


Figure 28: Nava et al. 2025, Nature Genetics – Extended Data Figure 1

producing defective mRNA (with a premature stop codon or frameshifted sequence), which is translated into a truncated, nonfunctional protein. Without enough functional histone methyltransferase, chromatin regulation goes awry in developing cells and tissues, leading to the developmental defects characteristic of Kabuki syndrome. In ASD, a mutant SCN2A allele affects the sodium channel protein, altering how neurons fire during brain development. In CHD, mutant NOTCH1 alleles produce defective signaling proteins that disrupt heart formation. In ReNU syndrome, mutant RNU4-2 alleles create defective snRNA, causing global splicing errors that affect neuronal proteins and brain development.

14.5.2 Codon Changes in Coding Regions

When a variant occurs in an exon (protein-coding region), it changes the **codon**—the three-nucleotide unit that specifies an amino acid. These changes affect RNA and protein in specific ways:

Nonsense Variants: A codon for an amino acid (like CAG for glutamine) is changed to a stop codon (like TAG). This produces an mRNA with a premature termination signal, leading to a truncated protein. The result is usually **haploinsufficiency**: the cell makes half as much functional protein as it should. In Kabuki syndrome, nonsense variants in KMT2D create mRNAs that encode truncated methyltransferases. These shortened proteins lack the critical SET domain and can't methylate histones. One functioning allele can't produce enough enzyme to maintain normal gene expression during development, leading to facial, skeletal, cardiac, and neurological abnormalities.

Missense Variants: One codon is changed to another, swapping one amino acid for a different one (like GAA for glutamic acid to GCA for alanine). Depending on where this happens and which amino acids are involved, the effect can range from harmless to devastating. In CHD, NOTCH1 missense variants often alter cysteines in EGF-like domains. These cysteines form disulfide bonds that stabilize the protein's structure. Changing them produces a misfolded protein that can't signal properly—a **dominant-negative effect**. The mutant protein interferes with normal NOTCH1 signaling, even though cells still make normal protein from the other allele, disrupting cardiac cell differentiation during heart development.

Frameshift Variants: Insertions or deletions that aren't multiples of three nucleotides shift the reading frame downstream of the mutation. This scrambles the amino acid sequence and usually introduces a premature stop codon shortly after. In ASD, SCN2A frameshift variants produce mRNAs encoding aberrant NaV1.2 proteins that are non-functional. This reduces the number of working sodium channels in developing neurons, decreasing excitability and contributing to neurodevelopmental abnormalities.

14.5.3 Non-Coding Variants and Splicing Disruption

Not all variants affect codons. Non-coding variants in regulatory regions can also act as dominant alleles by disrupting RNA processing. In neurodevelopmental disorders, variants in RNU4-2 alter the DNA sequence of an snRNA gene. The transcribed snRNA is defective—its secondary structure is disrupted. When this mutant snRNA is incorporated into spliceosomes, it can't properly recognize splice sites in pre-mRNA. This causes widespread mis-splicing across the genome, producing aberrant mRNAs and dysfunctional proteins, particularly in neurons. For example, the RNU4-2 insertion n.64_65insT alters the T-loop structure, disrupting the U4/U6 interaction critical for splicing. Even though the cell has normal U4 snRNA from the other allele, the mutant snRNA disrupts enough spliceosomes to cause widespread splicing errors, leading to ReNU syndrome. This is a dominant-negative effect at the RNA level.

14.6 Dominant Alleles Shape Phenotypes

In dominant disorders, a single mutant allele drives disease by disrupting the flow from DNA to phenotype:

In Kabuki syndrome,

```
KMT2D nonsense variants → defective mRNA → truncated protein →
    impaired chromatin regulation in developing cells → abnormal gene expression in tissues →
```

```
        developmental defects (facial features, heart structure, brain development).
```

In SCN2A-related disorders,

```
missense or PTV variants → altered or absent NaV1.2 protein →
    disrupted sodium channel function in neurons →
        abnormal brain cell excitability →
            ASD or seizures, depending on whether channels are under- or over-active.
```

In CHD,

```
NOTCH1 missense variants → misfolded signaling protein →
    disrupted cardiac cell differentiation → abnormal heart tissue formation →
        structural heart defects.
```

In ReNU syndrome,

```
RNU4-2 variants → defective snRNA →
 impaired spliceosome function → widespread mis-splicing →
   dysfunctional neuronal proteins → brain development abnormalities →
    neurodevelopmental disorder
```

These examples show how dominant alleles—whether PTVs, missense variants, or non-coding variants—cause disease by disrupting the molecular cascade from DNA to phenotype. One mutant copy is sufficient because the mutation either prevents adequate protein production (haploinsufficiency), creates a harmful new function (gain-of-function), or actively interferes with normal cellular processes (dominant-negative). Understanding this molecular path helps us connect genetic variants to disease and explains why sequencing technologies like WES and genome sequencing are so powerful for diagnosis and research.

15 Chapter 11. Haploinsufficiency

Here's a fundamental question: if we have two copies of each gene—one from Mom, one from Dad—why do we need both? For many genes, the answer is we don't. One working copy makes plenty of protein to keep things running smoothly. But some genes are pickier. They're called **haploinsufficient** genes, and they need both copies churning out protein to maintain normal function. Break one copy with a **protein-truncating variant (PTV)**, and you don't make enough protein—leading to problems like developmental disorders or cancer.

Not all genes are this sensitive, though. Some can lose one or even both copies without causing any noticeable harm. They're **loss-of-function tolerant**, acting like backup systems in our genome. How do scientists figure out which genes fall into which category? They use tools called **pLI** and **LOEUF** that essentially count how many PTVs show up in large populations compared to how many we'd expect by chance. If a gene has way fewer PTVs than expected, that's a sign that evolution has been weeding them out—a process called **purifying selection**—because they reduce **fitness** (the ability to survive and have kids).

This section explains haploinsufficiency, why some genes tolerate loss better than others, and how we measure these differences using population-scale data.

15.1 Haploinsufficient Genes vs. Haplosufficient Genes

Let's start with definitions. **Haploinsufficiency** happens when one copy of a gene gets knocked out (usually by a PTV), and the remaining copy can't produce enough protein to maintain normal function. The result? Disease or abnormal traits.

Think of it like baking a cake that requires two cups of sugar. If you only have one cup, you might still get something cake-like, but it won't taste right and the texture will be off. Haploinsufficient genes are like that recipe—they really need both copies to work.

In contrast, **haplosufficient** genes are fine with one functional copy. It's like a recipe that works perfectly well with one cup of sugar because it's more forgiving. Some genes go even further—they're **loss-of-function (LoF) tolerant**, meaning you can completely inactivate one or both copies without causing health problems. These genes often have backup copies elsewhere in the genome, or they're simply not critical for survival and reproduction.

Haploinsufficient genes are exquisitely sensitive to **gene dosage**—the amount of protein being produced. Cut that in half, and things go wrong. But here's the interesting part: even though PTVs are serious mutations that destroy gene function, healthy people walk around carrying about 100 PTVs each. Analysis of 185 genomes found that most of these PTVs sit in LoF-tolerant genes that don't affect health (MacArthur et al 2012, Science). So the mere presence of a PTV doesn't mean disaster—it depends entirely on which gene got hit.

See Figure 1 from MacArthur et al 2012 for a comparison of PTV frequencies across different filtering criteria in healthy adults. The data shows how high-confidence PTVs (after filtering out sequencing errors and other artifacts) are much rarer than the unfiltered set. This is partly because many apparent PTVs are actually mistakes, but also because real PTVs often occur in LoF-tolerant genes that don't matter much for health. The genes that really matter—the haploinsufficient ones—rarely carry PTVs in healthy people.

15.2 Natural Selection and Purifying Selection

Why are PTVs in haploinsufficient genes so rare? The answer is **natural selection**—the process that favors traits improving **fitness** (survival and reproduction) while weeding out harmful traits over generations.

Think of it as a very slow-motion race. If you have a mutation that helps you survive and have kids—say, better disease resistance—that mutation spreads through the population. If you have a mutation that reduces your chances—say, a PTV in a gene critical for brain development—you might not survive to reproduce, and that mutation dies with you. This is how evolution shapes which variants stick around.

Purifying selection is a specific type of natural selection focused on removing harmful mutations. It's the genome's quality control system. PTVs in haploinsufficient genes are harmful because they cause diseases, reduce survival, or impair reproduction. Purifying selection actively removes these variants from the population, which is why they're rare when we look at large datasets of healthy people.

On the flip side, PTVs in LoF-tolerant genes are more common because they don't harm fitness. Take olfactory receptor genes—we have hundreds of them for detecting smells. Losing one doesn't matter much, so PTVs in these genes are relatively frequent. No selective pressure, no removal.

This interplay between mutation (which constantly generates new PTVs) and purifying selection (which removes harmful ones) is key to understanding why some genes appear "intolerant" to loss-of-function mutations while others don't care.

15.3 Measuring Haploinsufficiency with pLI and LOEUF

So how do scientists actually identify haploinsufficient genes? The trick is to look at large populations and ask: "How many PTVs does this gene have compared to how many it should have?" If a gene has far fewer PTVs than expected, purifying selection has probably been removing them because they're harmful. That gene is likely haploinsufficient.

Two massive datasets—the **Exome Aggregation Consortium (ExAC)** and the **Genome Aggregation Database (gnomAD)**—developed scoring systems called **pLI** and **LOEUF** to quantify this intolerance. Let's walk through each.

15.3.1 The ExAC Study and pLI

Exome sequences from 60,706 adults were analyzed to identify patterns of loss-of-function intolerance (Lek et al 2016, Nature). Crucially, these were adults without severe developmental disorders—essentially a healthy population baseline. The exome covers the protein-coding parts of the genome, where most disease-causing mutations hide.

15.3.1.1 What is pLI? pLI stands for **probability of being loss-of-function intolerant**. It's a score from 0 to 1 that tells you how sensitive a gene is to PTVs. A pLI ≥ 0.9 means the gene is highly intolerant—it has way fewer PTVs than expected, suggesting it's haploinsufficient and critical for normal function. A pLI near 0 means the gene tolerates PTVs just fine, like LoF-tolerant genes.

15.3.1.2 How is pLI Calculated? The calculation starts with an expectation. Based on a gene's length and the background mutation rate, statisticians can predict roughly how many PTVs that gene should accumulate in a population of 60,706 people. This expected number accounts for the fact that bigger genes naturally acquire more mutations.

Then they compare this to the observed number of PTVs actually found in the ExAC data. If the observed number is dramatically lower than expected—say, near zero—that's strong evidence that purifying selection has been removing PTVs because they reduce fitness. The study identified 3,230 genes with pLI ≥ 0.9, indicating they're highly intolerant to loss of function. These genes tend to be involved in critical processes like ribosome assembly, chromatin regulation, and cell cycle control.

For example, SCN2A—a sodium channel gene critical for brain development—has a pLI of 1.0, the maximum score. The gene is expected to carry about 165.8 PTVs in a large population, but only 18 are actually observed. This dramatic depletion (o/e = 0.11) reflects strong purifying selection removing harmful PTVs because they cause epilepsy and neurodevelopmental disorders.

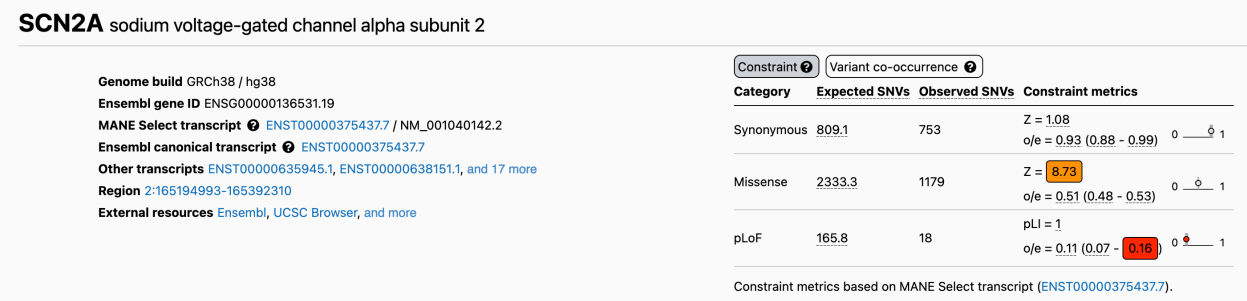


Figure 29: gnomAD constraint metrics for SCN2A

Figure: Constraint metrics for the SCN2A gene from gnomAD. The table shows expected versus observed variant counts across different categories. For protein-truncating variants (pLoF), SCN2A expects 165.8 variants but only 18 are observed, yielding an observed/expected (o/e) ratio of 0.11 and a pLI score of 1.0. This extreme constraint indicates that PTVs in SCN2A are strongly selected against because they cause disease. The missense Z-score of 8.73 also shows significant depletion of missense variants, reflecting the gene's overall intolerance to functional disruption. Source: gnomAD Browser. https://gnomad.broadinstitute.org/gene/ENSG00000136531?dataset=gnomad_r4.

Figure: Proportion of haploinsufficient genes across functional categories. This bar chart breaks down different gene categories by their pLI scores. Known haploinsufficient genes (those already linked to

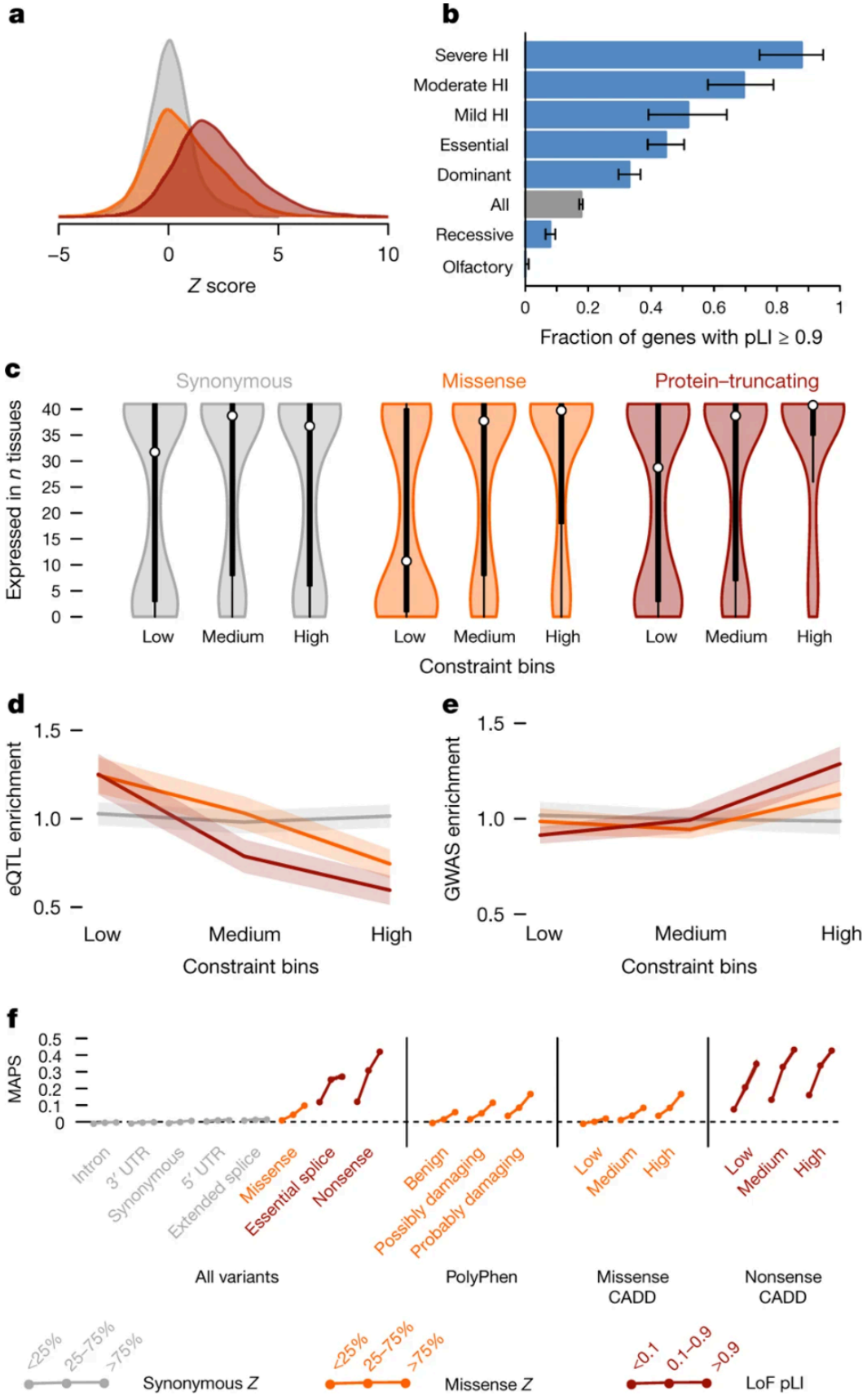


Figure 30: Lek et al. 2016, Nature – Figure 3

dominant disorders) overwhelmingly have high pLI scores (≥0.9), while recessive disease genes are more spread out. This validates pLI as a tool for identifying genes where one broken copy causes problems. The chart also shows that certain functional categories—like transcription factors and chromatin modifiers—are enriched for high pLI genes, reflecting their critical roles. Source: Lek, M. et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature. https://pmc.ncbi.nlm.nih.gov/articles/PMC5018207/. License: CC-BY 4.0.

15.3.1.3 Why Focus on Adults Without Developmental Delays? This is a smart design choice. Haploinsufficient genes often cause developmental disorders when disrupted. By studying adults who haven't been diagnosed with these conditions, the study ensured they were looking at a population where harmful PTVs are rare because affected individuals either didn't survive or didn't reproduce. It's like inspecting a library's intact books to figure out which ones are essential—you won't find torn-up copies of the critical ones because those were removed from circulation.

Figure: Allele frequency distribution of protein-truncating variants. Most PTVs in the ExAC dataset are extremely rare—the vast majority appear only once (singleton variants). This rarity pattern reflects purifying selection at work: harmful PTVs don't stick around long enough to become common. The contrast between PTVs in haploinsufficient genes (almost all singletons) and LoF-tolerant genes (more spread across frequencies) illustrates how selection shapes variant frequency. Source: Lek, M. et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature. https://pmc.ncbi.nlm.nih.gov/articles/PMC5018207/. License: CC-BY 4.0.

15.3.2 The gnomAD Study and LOEUF

In 2020, the **Genome Aggregation Database (gnomAD)** expanded on ExAC's work by analyzing 141,456 individuals—more than double the sample size (Karczewski et al 2020, Nature). This dataset included both exome and whole-genome sequences, providing even more comprehensive coverage. Like ExAC, it focused on adults without severe developmental disorders.

gnomAD introduced a refined metric called **LOEUF** that improves on pLI in important ways.

15.3.2.1 What is LOEUF? LOEUF stands for **Loss-of-Function Observed/Expected Upper-bound Fraction**. Instead of a binary "intolerant or not" classification, LOEUF provides a continuous score that ranks genes on a sliding scale of PTV sensitivity. A **low LOEUF score** (typically ⟨0.35) indicates a gene with very few PTVs—likely haploinsufficient and critical for fitness. A high LOEUF score means the gene tolerates PTVs well.

15.3.2.2 How is LOEUF Calculated? The calculation follows a similar logic to pLI: compare observed PTVs to expected PTVs based on gene size and mutation rate. But gnomAD adds sophistication through the **LOFTEE** (Loss-of-Function Transcript Effect Estimator) tool, which filters out false-positive PTVs—things that look like protein-truncating variants but aren't really, like sequencing errors or variants that occur in alternative transcripts where they don't actually truncate the protein.

After identifying 443,769 high-confidence PTVs across the dataset, the observed-to-expected ratio for each gene was calculated, adjusted for statistical uncertainty (hence "upper-bound fraction"). Genes with low LOEUF scores are depleted for PTVs far beyond what chance alone would explain. These tend to be genes critical for DNA repair, cell division, metabolism, and development—functions where losing one copy has serious consequences.

15.3.2.3 Why Continuous Scoring? pLI is essentially a yes/no test: a gene either meets the threshold for intolerance (pLI ≥ 0.9) or it doesn't. LOEUF is more like a dimmer switch—it shows how intolerant a gene is on a spectrum. This granularity helps identify genes with moderate intolerance that pLI might miss, and it's particularly useful for genes involved in less severe disorders or for prioritizing variants in clinical diagnosis.

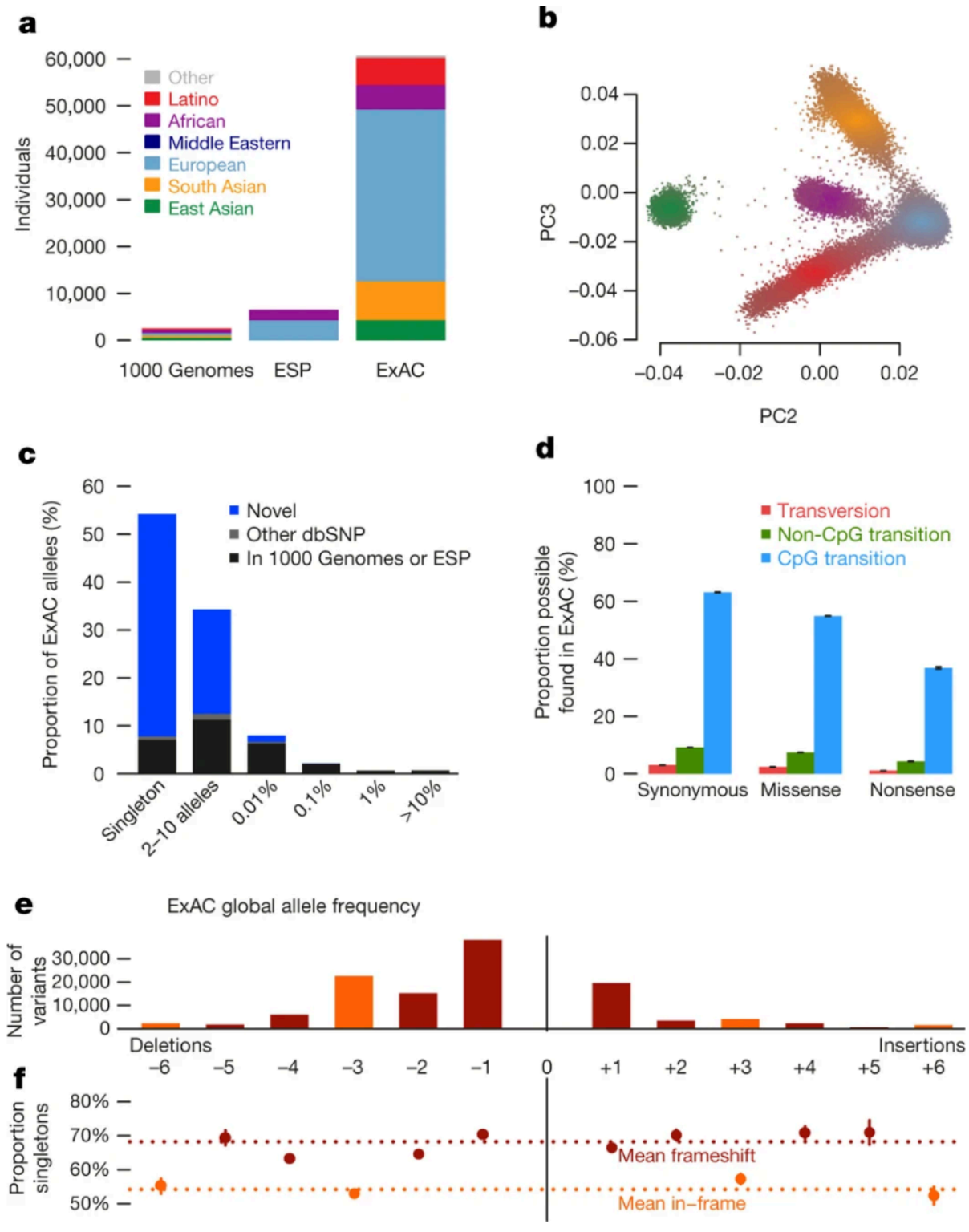


Figure 31: Lek et al. 2016, Nature – Figure 1

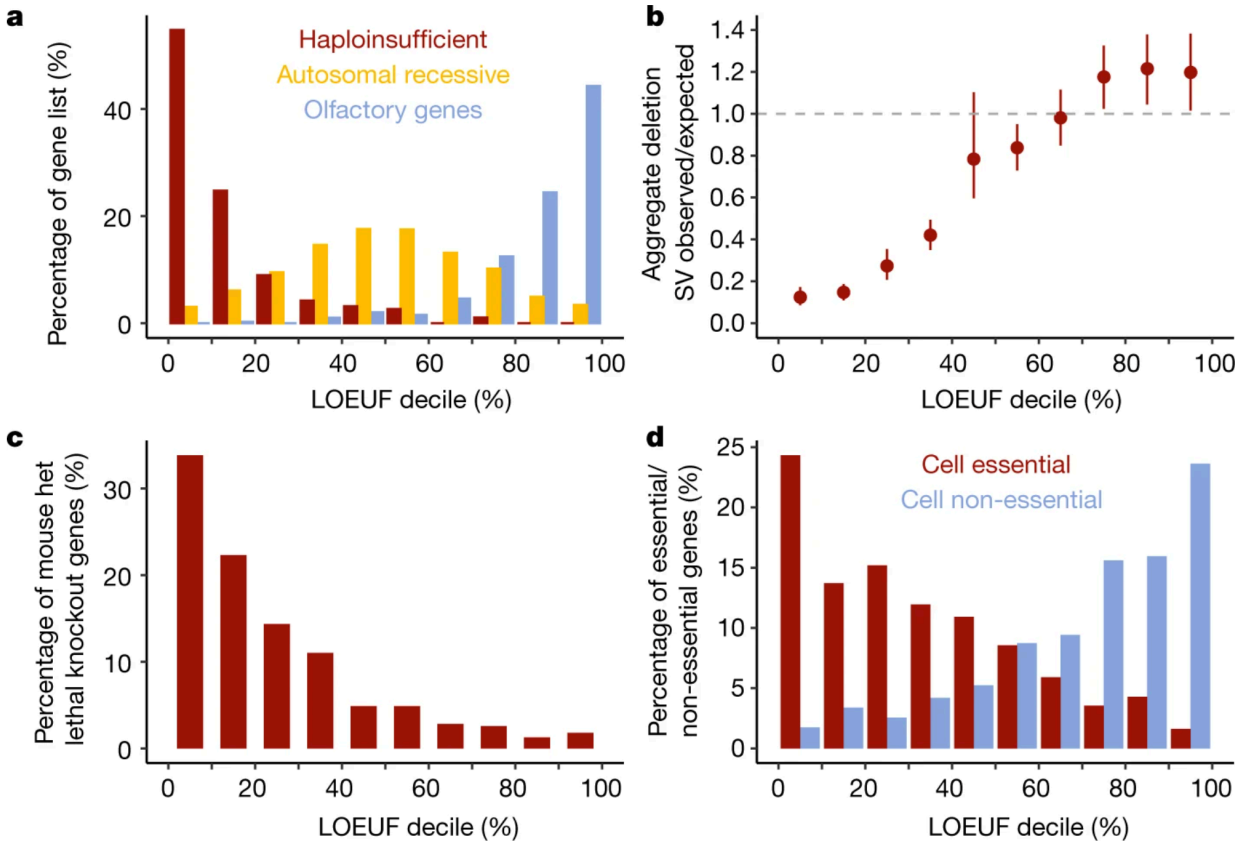


Figure 32: Karczewski et al. 2020, Nature – Figure 3

Figure: Distribution of LOEUF scores across gene categories. This plot shows how LOEUF scores correlate with other measures of gene importance, like whether knocking out the gene kills mice or causes developmental abnormalities. Genes with low LOEUF scores (left side) are overwhelmingly essential in mouse models, while high-LOEUF genes (right side) can often be knocked out without major consequences. This external validation confirms that LOEUF captures something real about biological importance and links nicely to the concept of fitness—genes you can't lose without reducing survival or reproduction. Source: Karczewski, K.J. et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. https://pmc.ncbi.nlm.nih.gov/articles/PMC7334197/. License: CC-BY 4.0.

15.4 Summary

Haploinsufficiency reveals a fundamental principle about how genes work: **not all genes are created equal when it comes to dosage sensitivity**. Some genes—like those encoding transcription factors, chromatin modifiers, and structural proteins—operate on a tight budget. They need both copies producing protein at full capacity to maintain normal cellular function. Cut that production in half with a PTV, and the cell can't compensate. The result is disease, often manifesting as developmental disorders because these genes are critical during embryonic growth when cells are dividing rapidly and precise gene expression matters most.

Other genes, by contrast, have built-in redundancy. Maybe they have backup copies elsewhere in the genome, or maybe the cell simply doesn't need that much of their protein product. These LoF-tolerant genes can absorb PTVs without consequence, which is why healthy people carry around 100 of them without even knowing it.

The difference comes down to **evolutionary pressure**. Haploinsufficient genes are under constant surveillance by purifying selection. Any PTV that reduces fitness—by causing death, developmental delay, or reproductive problems—gets filtered out of the population over generations. This is why PTVs in these genes are vanishingly rare in healthy adults. LoF-tolerant genes, meanwhile, face no such pressure. PTVs accumulate freely because they don't affect survival or reproduction.

Tools like **pLI** and **LOEUF** translate this evolutionary signal into practical numbers. By comparing observed versus expected PTVs across tens of thousands of genomes, they identify which genes are haploinsufficient and which aren't. This has immediate clinical applications. When a patient with a genetic disorder undergoes exome sequencing and a rare PTV is identified, clinicians can check the pLI or LOEUF score to assess whether that gene is a likely culprit. A PTV in a gene with pLI = 1.0 or LOEUF = 0.1? That's highly suspicious and warrants further investigation. A PTV in a gene with pLI = 0 or LOEUF = 1.5? Probably not the cause.

There are even cases where PTVs improve health. PTVs in *PCSK9* lower cholesterol levels and reduce heart disease risk. Drug companies developed cholesterol-lowering medications that mimic this effect—essentially "PTVs in a pill." This shows how understanding LoF tolerance can guide therapeutic development.

In summary, haploinsufficiency links gene dosage to disease through the lens of evolution. Some genes tolerate loss; others don't. Purifying selection has spent millions of years sorting out which is which, and by reading that signal in population data, we can now predict disease risk, prioritize diagnostic variants, and even design new therapies.

16 Chapter 12. Recessive Alleles in Human Disorders

Dominant disorders, as we've seen, need only one broken gene copy to cause problems. Recessive disorders play by different rules: you need **both alleles at a locus** to be disrupted before disease appears. Remember, we're diploid organisms—we carry two copies of every autosomal gene, one from Mom and

one from Dad. If both copies fail to produce functional protein, the cell has nothing to fall back on, and disease results.

The catch? People with just one bad copy—**carriers**—are typically perfectly healthy. That single working allele makes enough protein to keep everything running normally. This is why recessive disorders often surprise families: two healthy carrier parents can have an affected child who inherited the bad allele from both sides.

NGS technologies like WES and WGS have dramatically improved our ability to find recessive alleles, especially in populations where they're more likely to show up. Studies in British Pakistani and Bangladeshi communities, where marriage between relatives is culturally common, have uncovered hundreds of recessive disease associations that would be nearly impossible to detect in outbred populations (Narasimhan et al 2016, Science; Bilgüvar et al 2010, Nature; Heng et al 2025, Am J Hum Genet).

16.1 Alleles, Genotypes, and Variants

Since we're diploid, every locus contains two alleles—one maternal, one paternal. Let's break down the possibilities at any given locus, using a hypothetical gene we'll call Gene X:

Homozygous reference means both alleles match the reference genome—the "standard" sequence we use as a comparison point. You'd write this as Xref/Xref. This is what most people have at most loci.

Heterozygous means one allele is reference and the other carries a variant—some change from the reference sequence. Written as Xref/Xalt, where "alt" stands for alternative. This is the carrier state for recessive disorders: you have one working copy and one altered copy.

Homozygous alternative means both alleles carry the same variant: Xalt/Xalt. If that variant disrupts gene function—say, it's a nonsense mutation that truncates the protein—you've lost function from both copies. This is when recessive disease appears.

The "variant" itself is usually a **single-nucleotide variant (SNV)**—one DNA base swapped for another—or a small **insertion or deletion (indel)** of a few bases. Let me give you concrete examples.

For an SNV, imagine a position where the reference genome has the base G. A **homozygous reference** individual has G on both chromosomes (G/G). A **heterozygous** individual has G on one chromosome and a variant base—say T—on the other (G/T). A **homozygous alternative** individual has T on both chromosomes (T/T). If that T disrupts the gene—maybe it creates a stop codon—the homozygous alternative person loses function from both alleles.

For an indel, let's say the reference sequence is ATG. A **homozygous reference** individual has ATG/ATG. A **heterozygous** individual might have ATG on one chromosome and ATGT on the other (a one-base insertion: ATG/ATGT). A **homozygous alternative** individual has ATGT/ATGT. If that extra T shifts the reading frame and scrambles the downstream protein sequence, both alleles are now producing nonsense.

16.1.1 Biallelic Variation

Biallelic variation—affecting both alleles—is the key to recessive disorders. This can happen two ways:

Homozygous: The exact same variant appears on both alleles. You inherit the same mutant allele from Mom and from Dad. This is especially common when parents are related (consanguineous), because they're more likely to both carry the same ancestral mutation. In a study of 3,222 British Pakistani adults—many with parents who were second cousins or closer relatives—94.9% of rare homozygous loss-of-function (rhLOF) variants occurred within **autozygous segments**: stretches of DNA where both chromosomes are identical because they descended from a recent common ancestor (Narasimhan et al 2016, Science).

Compound heterozygous: Two different variants, one on each allele, both disrupt the gene. You might inherit variant A from Mom and variant B from Dad, and both knock out function. This is less common in consanguineous populations (where homozygous variants dominate) but shows up in outbred populations. Studies in Iceland and the ExAC database found 167 and 299 genes, respectively, with compound heterozygous LOF genotypes.

For this discussion, we'll focus mainly on homozygous variants, since they're more prevalent in the consanguineous cohorts that have powered recent recessive disease discoveries.

16.2 Consanguinity and Autozygosity

Consanguinity—marriage between close relatives like first or second cousins—dramatically increases the chance of recessive disease because it boosts **autozygosity**: regions of the genome where both your chromosomes are identical by descent (IBD) from a recent common ancestor.

Here's why this matters. Imagine your parents are second cousins. They share a pair of great-grandparents. Any rare mutation those great-grandparents carried has a decent chance of being passed down both family lines to your parents, and then to you—giving you two copies of the same mutant allele. It's like inheriting the same broken tool from both sides of the family.

This effect is particularly pronounced in British Pakistani and Bangladeshi communities in the UK, where about 32% of individuals are offspring of second cousins or closer. In a study of 3,222 Pakistani-heritage adults, an average of 5.6% of the coding genome was autozygous—both chromosomes identical by descent. That's orders of magnitude higher than in outbred European populations, where almost no autozygosity exists (Narasimhan et al 2016, Science).

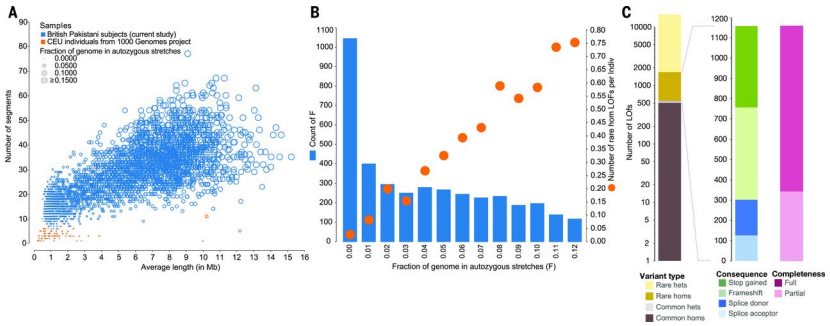


Figure 33: Narasimhan et al. 2016, Science – Figure 1

Figure: Autozygosity and rare homozygous loss-of-function variants in Pakistani-heritage individuals. Panel A compares autozygous segments between Pakistani-heritage subjects in the UK (blue circles) and European individuals from the 1000 Genomes Project (orange dots, CEU population). Each circle represents an individual, with the x-axis showing the average length of autozygous segments and the y-axis showing the number of such segments. Pakistani individuals show dramatically more and longer autozygous segments due to consanguinity—many have over 30 segments averaging 5–10 Mb in length, while Europeans have almost none. Panel B reveals the relationship between autozygosity and rare homozygous loss-of-function (rhLOF) genotypes. Blue bars show the number of individuals binned by the fraction of their genome that is autozygous (x-axis from 0 to 0.12, meaning 0% to 12%). Orange circles track the average number of rhLOF genotypes per individual (right y-axis). As autozygosity increases, individuals carry more rhLOF genotypes—those with ~6% autozygosity (typical for first-cousin offspring) average about 0.5 rhLOF genotypes, while those with 10% have nearly 0.8. Panel C breaks down the 1,111 rhLOF genotypes discovered across 781 genes by variant properties: allele frequency (rare variants dominate), genotype (mostly rare homozygotes, shown in dark colors), predicted protein consequence (frameshift, stop gained, splice donor/acceptor mutations), and whether the LOF prediction applies to all

gene transcripts (full) or only some (partial). This shows how consanguinity creates a natural experiment—revealing which genes can tolerate complete knockout and which cannot. Source: Narasimhan, V.M. et al. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. Science 352(6284):aac8624. https://pmc.ncbi.nlm.nih.gov/articles/PMC4985238/. License: CC-BY.

The consequence? These 3,222 individuals collectively carried 1,111 rare homozygous LOF genotypes across 781 genes. That's a treasure trove for geneticists, because it reveals which genes cause disease when completely knocked out—information that's nearly impossible to get from outbred populations where homozygous knockouts are vanishingly rare.

In a larger study of 44,190 British Pakistani and Bangladeshi individuals, this autozygosity enabled the detection of 185 genetic loci with recessive associations to 898 common diseases. Many of these associations were completely novel, only detectable because of the unique genetic structure of this cohort (Heng et al 2025, Am J Hum Genet).

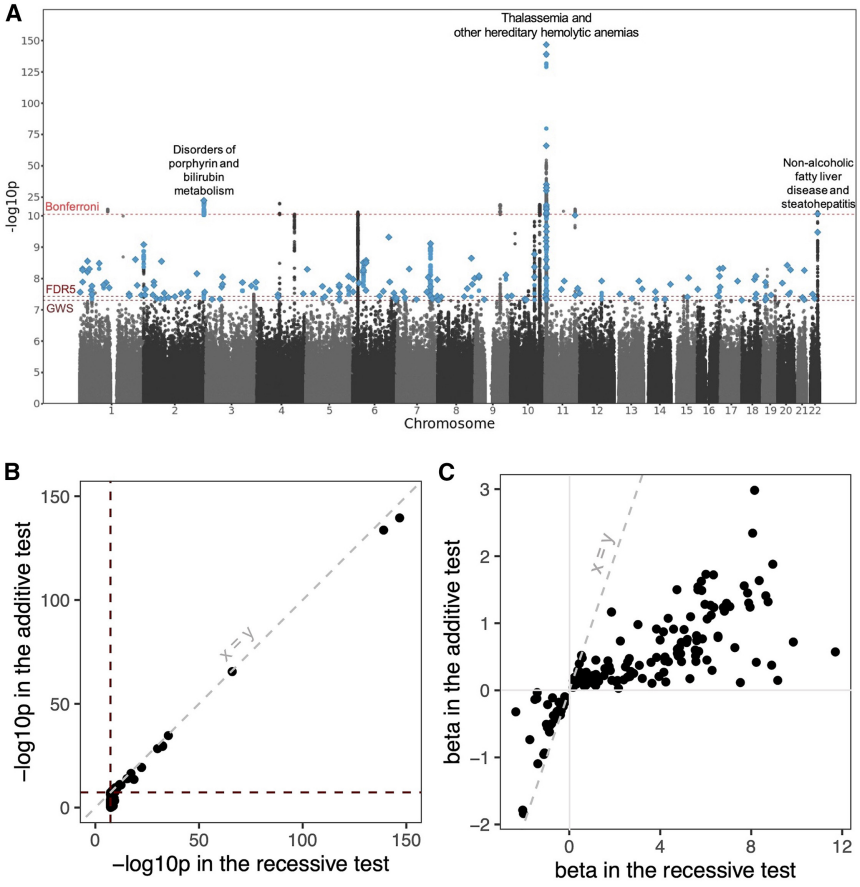


Figure 34: Heng et al. 2025, Am J Hum Genet – Figure 1A

Figure: Genome-wide landscape of recessive associations in British Pakistanis and Bangladeshis. This Manhattan plot displays the results from recessive association testing across the genome. Each point represents a genetic variant tested for association with one of 898 diseases from electronic health records. The y-axis shows the strength of the statistical evidence (−log10 p-value), and the x-axis shows the chromosomal position. The dashed horizontal lines mark different significance thresholds: genome-wide significance (GWS; $p < 5 \times 10^{-8}$, the standard cutoff for GWAS), false discovery rate of 5% (FDR5; $p < 3.7 \times 10^{-8}$), and Bonferroni-corrected threshold ($p < 5.5 \times 10^{-12}$) which accounts for multiple phenotypes tested. Blue points highlight variants that reached genome-wide significance and had stronger evidence under a recessive model than an additive model. Diamonds mark the lead variants—the most significant variant at each independent locus. Three clusters of Bonferroni-significant

associations are labeled: variants associated with non-alcoholic fatty liver disease (NAFLD) on chromosome 22, disorders of porphyrin and bilirubin metabolism on chromosome 2, and thalassemia and hereditary hemolytic anemias on chromosome 11. This visualization demonstrates how increased homozygosity in consanguineous populations reveals recessive disease architectures that remain hidden in outbred cohorts. Source: Heng, T.H. et al. (2025). Widespread recessive effects on common diseases in a cohort of 44,000 British Pakistanis and Bangladeshis with high autozygosity. Am J Hum Genet. https://pmc.ncbi.nlm.nih.gov/articles/PMC12256797/. License: CC-BY.

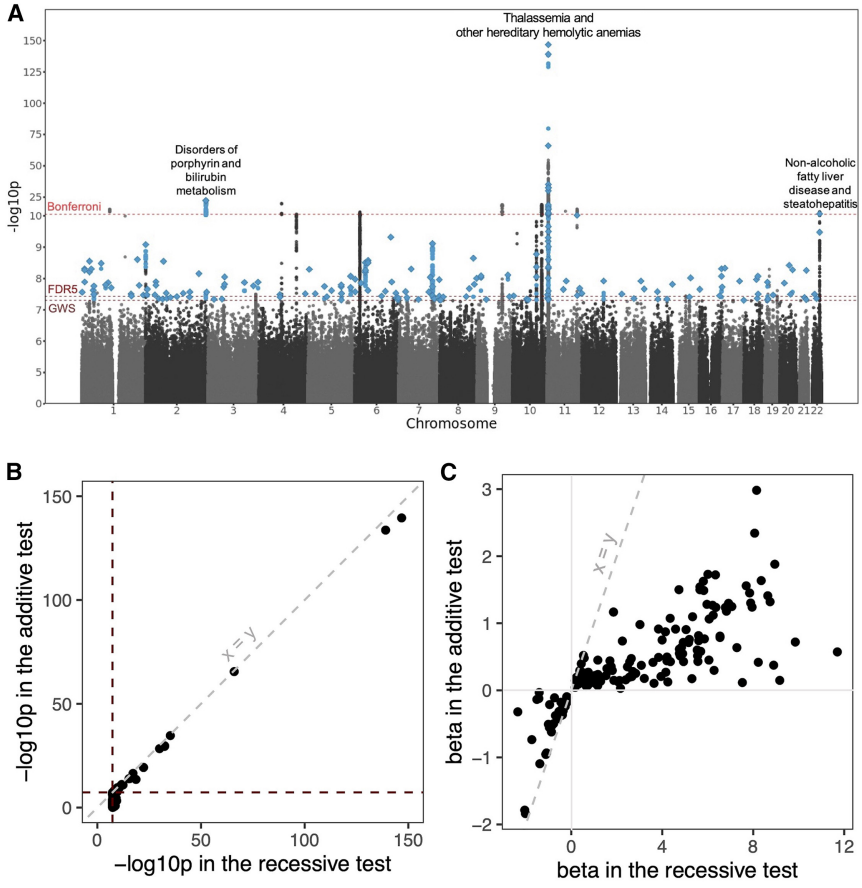


Figure 35: Heng et al. 2025, Am J Hum Genet – Figure 1A

Figure: Genome-wide landscape of recessive associations in British Pakistanis and Bangladeshis. This Manhattan plot displays the results from recessive association testing across the genome. Each point represents a genetic variant tested for association with one of 898 diseases from electronic health records. The y-axis shows the strength of the statistical evidence (-log10 p-value), and the x-axis shows the chromosomal position. The dashed horizontal lines mark different significance thresholds: genome-wide significance (GWS; $p < 5 \times 10^{-8}$, the standard cutoff for GWAS), false discovery rate of 5% (FDR5; $p < 3.7 \times 10^{-8}$), and Bonferroni-corrected threshold ($p < 5.5 \times 10^{-12}$) which accounts for multiple phenotypes tested. Blue points highlight variants that reached genome-wide significance and had stronger evidence under a recessive model than an additive model. Diamonds mark the lead variants—the most significant variant at each independent locus. Three clusters of Bonferroni-significant associations are labeled: variants associated with non-alcoholic fatty liver disease (NAFLD) on chromosome 22, disorders of porphyrin and bilirubin metabolism on chromosome 2, and thalassemia and hereditary hemolytic anemias on chromosome 11. This visualization demonstrates how increased homozygosity in consanguineous populations reveals recessive disease architectures that remain hidden in outbred cohorts. Source: Heng, T.H. et al. (2025). Widespread recessive effects on common diseases in a cohort of 44,000 British Pakistanis and Bangladeshis with high autozygosity. Am J Hum Genet.

https://pmc.ncbi.nlm.nih.gov/articles/PMC12256797/. License: CC-BY.

16.3 Recessive Inheritance Patterns in Human Pedigrees

See Figure 2 from Bilgüvar et al 2010, Nature, showing recessive inheritance of WDR62 mutations in six families (NG26, NG30, NG190, NG294, NG339, NG537). The pedigrees illustrate classic autosomal recessive inheritance, with each chromatogram showing the mutant allele (left, red bases) and the normal sequence (right, green bases). Notice the variety of mutation types—frameshifts (like V1402GfsX12), nonsense mutations (E526X, Q470X), missense variants (W224S, E526K), and small deletions (G1280AfsX21)—all occurring independently in different families. This illustrates genetic heterogeneity: different mutations in the same gene can cause the same disease. All affected individuals are homozygous for their respective mutations, consistent with consanguinity increasing the likelihood of inheriting the same recessive allele from both parents (Bilgüvar et al 2010, Nature).

16.4 Recessive Alleles in Mendelian and Complex Diseases

Traditionally, we think of recessive disorders as **Mendelian diseases**—single-gene conditions with dramatic phenotypes and clear inheritance patterns. Classic examples include cystic fibrosis, sickle cell disease, and Tay-Sachs. Break both copies of the gene, get the disease, with high penetrance (nearly everyone with the genotype shows symptoms).

But NGS studies have revealed that recessive alleles also contribute to **complex diseases**—conditions like hypertension, diabetes, and liver disease where multiple genes and environmental factors interact. The effects are subtler, but they're real and clinically important.

16.4.1 Mendelian Disorders

A great example comes from WES that identified homozygous recessive mutations in the WDR62 gene across several consanguineous families. Affected individuals had severe brain malformations: microcephaly (abnormally small heads), pachygyria (unusually thick, simplified brain folds), and lissencephaly (smooth brain, lacking normal folding) (Bilgüvar et al 2010, Nature).

The mutations varied—one family had a 4-base-pair deletion causing a frameshift, another had a nonsense mutation, others had missense variants—but all resulted in complete loss of WDR62 function. This protein is critical for neural progenitor cells, which divide to build the developing brain. Without functional WDR62, these cells can't proliferate properly, leading to the severe structural defects seen in affected children. The homozygous nature of these mutations, revealed by the consanguineous pedigrees, definitively proved recessive inheritance.

16.4.2 Complex Diseases

The 44,190-person British Pakistani and Bangladeshi cohort uncovered something more surprising: 185 loci with recessive associations to common diseases. These aren't dramatic Mendelian disorders—they're everyday health conditions (Heng et al 2025, Am J Hum Genet).

For instance, a missense variant in SGLT4 (rs61746559) showed a protective effect against hypertension. People homozygous for this variant had an 80% lower risk of high blood pressure (odds ratio = 0.2, $p = 3 \times 10^{-8}$). That's a huge effect size for a common disease. The variant likely alters how the kidney handles sodium, affecting blood pressure regulation through the renin-angiotensin-aldosterone system.

Another example: a recessive variant in PNPLA3 (rs66812091) was strongly associated with non-alcoholic fatty liver disease (odds ratio = 1.3, $p = 2 \times 10^{-12}$). Homozygous individuals had elevated liver enzyme

levels and increased disease risk. This variant disrupts fat metabolism in liver cells, causing lipid accumulation.

These findings highlight how recessive alleles can influence complex traits through more modest effects. Traditional GWAS studies, which typically assume additive genetic models, often miss these recessive associations. Only by specifically testing for recessive effects—made possible by the high autozygosity in consanguineous populations—do they become visible.

16.5 Population Genetics and Selection

Recessive alleles, especially harmful ones, face **natural selection** when they become homozygous. If a recessive mutation causes severe disease or death, individuals homozygous for it won't survive to reproduce, and that allele gets removed from the gene pool. This is why truly deleterious recessive alleles remain rare, even in consanguineous populations.

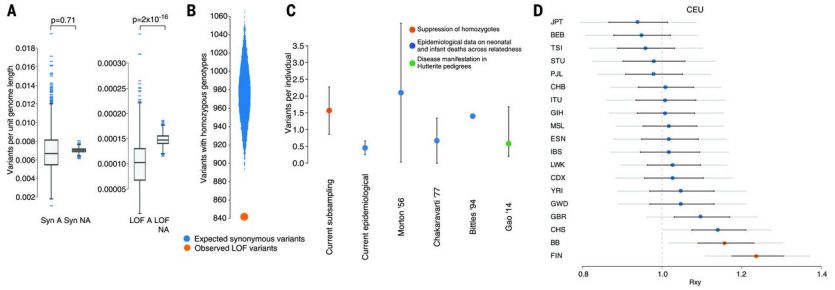


Figure 36: Narasimhan et al. 2016, Science – Figure 2

Figure: Natural selection against recessive loss-of-function variants. This figure quantifies how natural selection removes harmful recessive variants. Panel A shows that autozygous regions (where variants become homozygous) have significantly fewer LOF variants per unit length compared to non-autozygous regions—a depletion visible when comparing "LOF A" (LOF in autozygous) to "LOF NA" (LOF in non-autozygous). In contrast, synonymous variants (which don't affect protein function) show no such difference, confirming this is selection against harmful variants, not a technical artifact. Panel B demonstrates a 13.7% deficit of rare homozygous LOF genotypes. The study compared 16,708 rare LOF variants to frequency-matched synonymous variants: only 842 rare LOF variants had at least one homozygote (orange circle), versus an expected 975.5 synonymous variants with homozygotes (blue distribution). This missing 13.7% likely represents lethal or severely deleterious variants that prevented carriers from surviving to adulthood. Panel C compares different estimates of recessive lethal load—how many lethal-equivalent mutations each person carries. The direct estimate from rhLOF variants in this study (orange) finds ~1.6 lethal equivalents per person, matching epidemiological estimates from infant mortality data (blue circles) and data from Hutterite pedigrees (green). Panel D shows that Pakistani-heritage populations (BB) have slightly reduced LOF allele burden compared to Europeans (CEU), but less reduction than the Finnish population (FIN), which experienced a severe bottleneck. The y-axis represents the relative frequency of population-specific derived LOF alleles (under neutrality, this ratio should be 1.0). Error bars show standard errors, with significant differences marked by orange circles. This demonstrates that consanguinity increases the efficiency of purifying selection against recessive lethals. Source: Narasimhan, V.M. et al. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. Science 352(6284):aac8624. https://pmc.ncbi.nlm.nih.gov/articles/PMC4985238/. License: CC-BY.

The Pakistani-heritage study quantified this effect by comparing the observed number of rare homozygous LOF genotypes to the expected number based on allele frequencies and consanguinity rates. A 13.7% deficit was found—there were fewer homozygous knockouts than expected. What happened to the missing ones? They likely caused embryonic or fetal loss, preventing those individuals from reaching adulthood and being counted in the study (Narasimhan et al 2016, Science).

By modeling this deficit, the study estimated that each person carries approximately 1.6 recessive lethal-equivalent LOF variants. These are mutations that, when homozygous, cause death before birth or severe disease that prevents reproduction. In consanguineous populations, where these variants become homozygous more often, this "genetic load" manifests as increased rates of pregnancy loss and infant mortality.

But here's the nuance: not all recessive knockouts are harmful. The same study found a healthy mother who was homozygous for a LOF mutation in PRDM9, a gene involved in meiotic recombination (how chromosomes swap segments during sperm and egg formation). She had three healthy children, indicating that complete loss of PRDM9 has minimal health impact in humans, even though it's essential in mice.

This shows that "loss of function" doesn't automatically mean "disease." Some genes are dispensable, or have redundant backups, or only matter in specific contexts. The PRDM9 example reminds us that the effects of recessive alleles span a spectrum from lethal to neutral to occasionally even beneficial.

16.6 Summary

Recessive alleles reveal fundamental insights about human biology and disease across a spectrum from rare Mendelian disorders to common complex traits. By studying consanguineous populations where both copies of genes are more likely to be knocked out, we've learned several key lessons about human genetics.

Gene essentiality varies widely. Not all knockouts are harmful. The discovery of a healthy, fertile woman with complete PRDM9 knockout demonstrates that some genes essential in mice are dispensable in humans (Narasimhan et al 2016, Science). Her three healthy children and alternative meiotic recombination patterns show that humans have compensatory mechanisms not present in other mammals. This teaches us to be cautious when interpreting predicted loss-of-function variants—context matters, and human biology doesn't always follow animal models.

Selection shapes the genetic landscape. Each person carries approximately 1.6 recessive lethal-equivalent variants. Natural selection efficiently removes the most harmful recessive alleles when they become homozygous, explaining the 13.7% deficit of expected rare homozygous knockouts in adult populations. This "genetic load" manifests as increased embryonic lethality and infant mortality in consanguineous families, but the exact variants differ between individuals, spreading the burden across the genome.

Recessive effects extend beyond Mendelian disease. The identification of 185 loci with recessive associations to common diseases like hypertension and fatty liver disease challenges the additive model that dominates GWAS. Traditional studies miss these effects because they're only detectable when variants are homozygous—rare in outbred populations but common where autozygosity is high. This suggests we've been missing a significant component of disease architecture.

Clinical applications are emerging. Severe recessive disorders like WDR62 brain malformations enable genetic counseling and prenatal diagnosis for at-risk families. But the real excitement lies in protective variants like SGLT4 for hypertension. If knocking out a gene lowers disease risk without side effects, that gene becomes a drug target. This is "human genetics-guided drug development"—using natural knockouts as proof-of-concept for therapeutic interventions, the same logic behind PCSK9 inhibitors for cholesterol.

Diverse populations are essential for discovery. The focus on European-ancestry cohorts has created blind spots in our understanding of human genetics. British Pakistani and Bangladeshi populations, with their unique demographic history, have uncovered hundreds of recessive associations invisible in other cohorts. This isn't just about ancestry—it's about studying populations with different evolutionary histories and demographic structures to build a complete picture of human genetic architecture.

As whole-genome sequencing becomes routine, we'll continue identifying natural human knockouts across diverse populations. Each discovery teaches us which genes are truly essential, which have re-

dundant backups, and which harbor hidden recessive effects on common diseases. The key insight: two broken copies teach us different lessons than one broken copy, and both are essential for understanding human health and designing better therapies.

17 Chapter 13. Structural Variations in Human Genomes

So far, we've been talking mainly about small changes in DNA—single base swaps (SNVs) or tiny insertions and deletions (indels). But the genome can undergo much bigger rearrangements called **structural variations (SVs)**. These involve chunks of DNA that are 50 nucleotides or longer—sometimes stretching into the thousands or even millions of base pairs.

Think of the genome as a book. SNVs are like single-letter typos: "cat" becomes "bat." But SVs? They're entire paragraphs being cut out, duplicated, flipped upside down, or moved to a different chapter. These large-scale changes can have dramatic effects on gene function, human genetic diversity, and disease risk.

This chapter introduces SVs using insights from a massive study that analyzed 14,891 human genomes to create the gnomAD-SV reference—the most comprehensive catalog of structural variation to date (Collins et al 2020, Nature). We'll walk through what SVs are, how they're classified, where they occur in the genome, how they form, and what impact they have on health and evolution.

17.1 What Are Structural Variants?

Structural variants are DNA rearrangements that can include insertions (new DNA added), deletions (DNA removed), duplications (DNA copied), inversions (DNA flipped), or translocations (DNA moved between chromosomes). The key difference from SNVs is scale: while an SNV changes one letter, an SV can affect thousands to millions of letters at once.

These rearrangements can seriously disrupt genes or the regulatory elements that control them. Delete a critical exon, and the gene stops working. Duplicate a gene, and you might produce too much of its protein. Flip a regulatory region, and a gene might get turned on at the wrong time or in the wrong tissue. Understanding SVs is essential for understanding both normal genetic variation and disease.

17.2 Classification of SVs

SVs can be grouped based on whether they change the total amount of DNA in the genome—what we call **copy number**. Some SVs add or remove DNA, while others just rearrange it without changing the total amount.

17.2.1 Copy Number Altering SVs

These variants change how much DNA you have:

Deletions (DEL) remove a segment of DNA. If you delete part of a gene, you lose that genetic information, reducing gene dosage—the amount of protein that gene can make. For example, if a deletion removes one copy of a gene in a heterozygous individual, you're left with only one functional copy instead of two.

Duplications (DUP) create an extra copy of a DNA segment. Now you have more of that sequence than you should, increasing gene dosage. Some duplications are tandem (the extra copy sits right next to the original), while others insert elsewhere in the genome.

Insertions (INS) add new DNA that wasn't there before. Often these come from **mobile elements**—chunks of DNA that can copy themselves and jump to new locations. The three main types are: - **Alu elements**:

Short sequences, about 300 base pairs, that are the most abundant mobile elements in our genome – **LINE1 (Long Interspersed Nuclear Element 1)**: Longer sequences, around 6,000 base pairs – **SVA (SINE–VNTR–Alu)**: Composite elements, about 2,100 base pairs, that combine features of other repetitive sequences

These mobile elements leave distinctive signatures in the genome, which we'll see later in the data.

Multiallelic Copy–Number Variants (MCNVs) are regions where different people have different numbers of copies. You might have 2 copies of a region, your friend might have 3, and someone else might have 4. These show complex patterns of dosage variation across populations.

17.2.2 Copy Number Neutral SVs

These variants rearrange DNA without changing the total amount:

Inversions (INV) flip a DNA segment 180 degrees. The same bases are there, but they're now in reverse orientation. Imagine reading a sentence backwards—the letters are all present, but the meaning changes.

Translocations (BND) swap DNA segments between different chromosomes. You're not losing or gaining DNA overall; it's just being moved around. The abbreviation BND stands for "breakend," referring to where the DNA breaks to make the swap.

Complex SVs involve multiple types of rearrangements happening together. For example, you might have an inversion flanked by copy number changes, or **chromothripsis**—a catastrophic event where a chromosome shatters and gets pieced back together in a chaotic way. Depending on how the pieces reassemble, these can be copy number neutral or altering.

Figure: Classification of structural variant types. Structural variants fall into two major categories based on whether they alter copy number. Copy number altering variants include deletions (DEL), which remove DNA segments; duplications (DUP), which create extra copies; insertions (INS), which add new sequences often from mobile elements; and multiallelic copy number variants (MCNV), where different individuals carry different numbers of copies at the same locus. Copy number neutral variants include inversions (INV), which flip DNA segments 180 degrees without changing content; translocations (BND for "breakend"), which swap segments between chromosomes; and complex SVs combining multiple rearrangement types. Each SV type has distinct molecular signatures and functional consequences. For instance, deletions and duplications directly alter gene dosage—too little or too much protein product—while inversions can disrupt genes at breakpoints or alter their regulatory context without changing the amount of DNA present. The schematic uses different colors to represent distinct genomic segments, with arrows indicating orientation. Understanding these categories is essential for interpreting how large–scale genomic rearrangements affect human health and evolution. Source: Collins, R.L. et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444–451. https://www.nature.com/articles/s41586–020–2287–8. License: CC–BY 4.0.

17.3 Distribution of SVs

The gnomAD–SV dataset analyzed 14,237 genomes from diverse populations to understand how SVs are distributed across humanity. The cohort was remarkably diverse: 54% of individuals were non–European, including African, Latino, East Asian, and European ancestry groups. This diversity matters because SV patterns differ across populations, reflecting their unique evolutionary histories (Collins et al 2020, Nature).

17.3.1 How Many SVs Does Each Person Have?

The study identified 433,371 different SVs across all genomes analyzed. But how many does a typical person carry? The median was **7,439 SVs per genome**—that's a lot more than the 3,441 SVs per genome

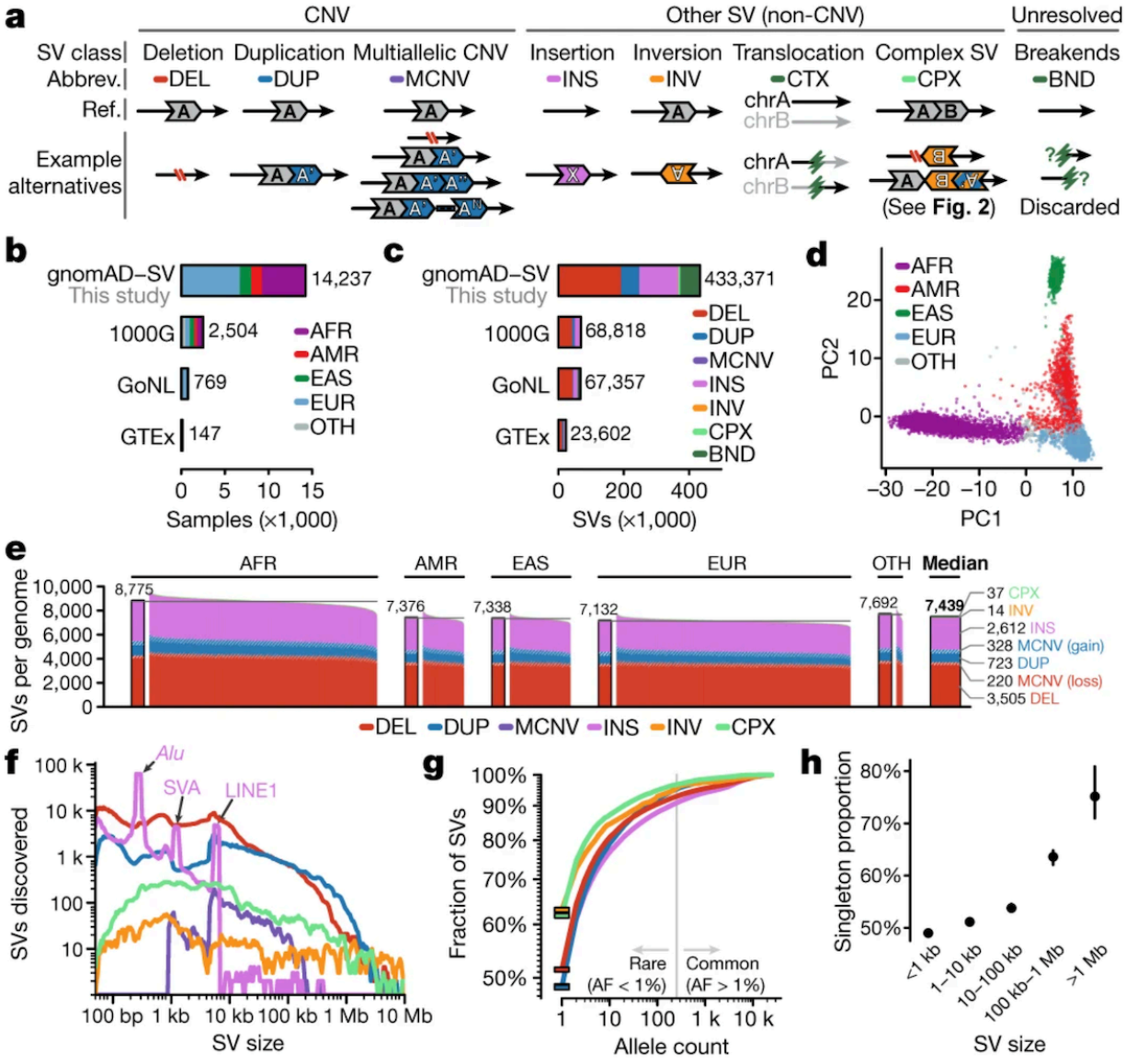


Figure 37: Collins et al. 2020, Nature – Figure 1

estimated by the earlier 1000 Genomes Project, which used lower-coverage sequencing that missed many variants.

Think about that number: each of us carries over 7,000 structural rearrangements in our genome. Most don't cause problems—they're just part of normal human genetic variation. But a small fraction can contribute to disease.

17.3.2 Size Matters

Most SVs are relatively small—under 10,000 base pairs—but there's interesting structure in the size distribution. When you plot the number of SVs at each size, distinct peaks appear at specific lengths. These peaks correspond to mobile element insertions:

- A peak around **300 base pairs** from Alu insertions
- A peak around **2,100 base pairs** from SVA insertions
- A peak around **6,000 base pairs** from LINE1 insertions

These mobile elements are constantly creating new insertions in human genomes, and their characteristic sizes show up clearly in the data.

17.3.3 Rare Variants Dominate

Here's a striking finding: **49.8% of all SVs are singletons**—they appear in only one person out of 14,237. Why are so many SVs this rare? Partly because they're recent mutations that haven't had time to spread, and partly because many large SVs disrupt important genes. If an SV knocks out a gene you need for normal development or survival, you're less likely to pass it on. Natural selection keeps these harmful variants rare.

The pattern is clear in the data: larger SVs are rarer than smaller ones. A deletion removing 100 base pairs might be tolerable, but a deletion removing 100,000 base pairs that spans multiple genes? That's more likely to cause problems and less likely to stick around in the population.

17.3.4 Population Structure

When a **principal component analysis** was performed on 15,395 common SVs (variants found in multiple individuals), people clustered by genetic ancestry—African, European, East Asian, and Latino groups separated in the analysis. This tells us that SV patterns differ across populations, shaped by unique demographic histories like population bottlenecks, expansions, and migrations. Some SVs might be common in one population but rare or absent in another.

17.4 Mechanisms of SV Formation

How do these large-scale rearrangements actually happen? Several molecular mechanisms can produce SVs, each leaving characteristic signatures in the genome.

Figure: Five major mechanisms generate structural variants in human genomes. NAHR (green) occurs when similar sequences like Alu elements misalign during meiosis, producing deletions on one chromosome and duplications on the other. NHEJ (orange) is a "quick and dirty" repair of double-strand breaks that glues DNA ends together without requiring sequence homology, often introducing small insertions or deletions. MEI (yellow) uses a copy-and-paste mechanism where mobile elements are transcribed to RNA, reverse-transcribed to DNA, and inserted at new locations, creating target site duplications and characteristic size peaks. FoSTeS (purple) results from stalled replication forks switching between DNA

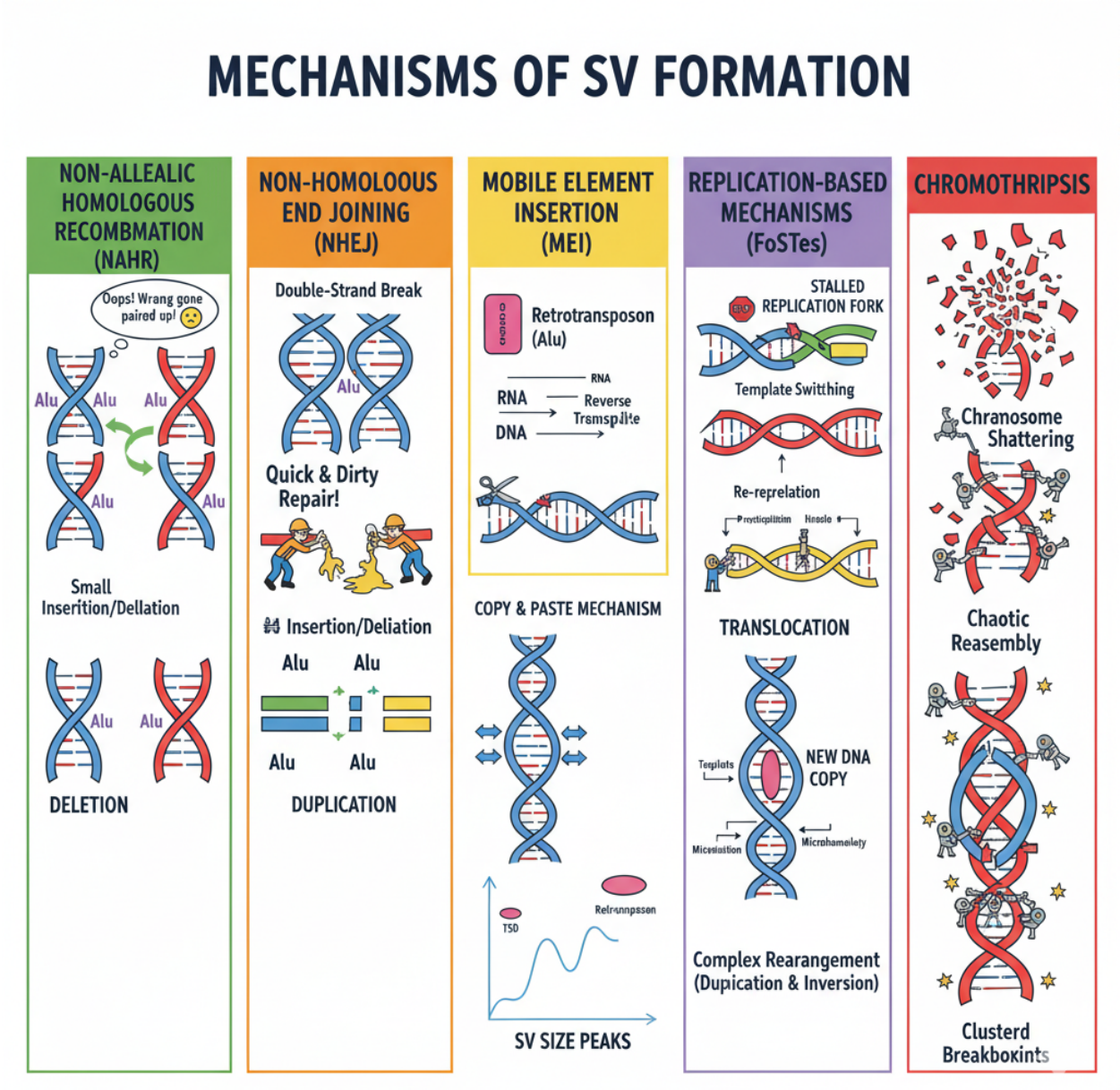


Figure 38: Mechanisms of SV Formation

templates, generating complex rearrangements with microhomology signatures. Chromothripsis (red) involves catastrophic chromosome shattering into many fragments followed by chaotic reassembly, creating clustered breakpoints—rare in healthy individuals but common in cancer.

17.4.1 Non-Allelic Homologous Recombination (NAHR)

During meiosis—the cell division that produces sperm and eggs—chromosomes pair up and exchange segments in a process called recombination. Usually this works perfectly: matching sequences on each chromosome (homologous sequences) line up correctly and swap corresponding pieces.

But sometimes things go wrong. If two repetitive sequences that aren't actually at the same position—say, two Alu elements separated by thousands of base pairs—mistakenly pair up, recombination between them produces a deletion on one chromosome and a duplication on the other. This is **non-allelic homologous recombination** (NAHR): homologous sequences that shouldn't recombine do anyway, because they're similar enough to trick the cellular machinery.

NAHR happens most often in regions with **segmental duplications**—large blocks of highly similar sequence scattered throughout the genome. The breakpoints (where the DNA breaks and rejoins) fall within these homologous regions, and the resulting SVs are typically deletions, duplications, or inversions.

17.4.2 Non-Homologous End Joining (NHEJ)

What happens when a chromosome suffers a double-strand break—both strands of the DNA double helix snap? Cells have repair mechanisms, but they're not always precise. **Non-homologous end joining** (NHEJ) is a quick-and-dirty repair method: the cell just glues the broken ends back together without looking for matching sequences to guide the repair.

This can work fine if the break is clean. But NHEJ often introduces small errors at the junction—tiny insertions or deletions. Sometimes the cell joins the wrong ends together, creating translocations (if pieces from different chromosomes get joined) or complex rearrangements.

NHEJ-mediated SVs show characteristic features: **microhomology** (just 1–10 matching base pairs) or completely blunt junctions at the breakpoints. Depending on what gets joined to what, these SVs can be copy number neutral (like a simple translocation) or altering (if pieces are lost or duplicated).

17.4.3 Mobile Element Insertion (MEI)

We mentioned Alu, LINE1, and SVA elements earlier. These **retrotransposons** create SVs through a "copy-and-paste" mechanism. Here's how it works:

1. The mobile element gets transcribed into RNA
2. That RNA is reverse-transcribed back into DNA
3. The new DNA copy inserts into a new location in the genome

The result is a copy number gain—you've inserted new DNA that wasn't there before. These insertions create **target-site duplications** (TSDs): short repeated sequences (2–20 base pairs) flanking the insertion, formed when the insertion machinery cuts the target DNA.

Mobile element insertions are common—we each carry hundreds of them—and they're constantly creating new genetic variation. The distinctive size peaks in the SV distribution we saw earlier come directly from these insertions.

17.4.4 Replication-Based Mechanisms (Fork Stalling and Template Switching)

DNA replication isn't always smooth. Sometimes the replication fork—the Y-shaped structure where DNA is being copied—stalls. Maybe it hits a difficult-to-replicate sequence, or there's damage in the DNA. When this happens, the replication machinery might switch to a different DNA template to finish copying, then switch back. This is called **fork stalling and template switching (FoSTeS)**.

If the machinery switches to the wrong template or switches multiple times, the result is a complex SV with multiple breakpoints and rearrangements. These can be copy number neutral (like inversions) or altering (like duplications), and they tend to occur in **fragile sites**—regions of the genome that are prone to replication stress.

17.4.5 Chromothripsis

Chromothripsis is the most dramatic mechanism—the name literally means "chromosome shattering." In a single catastrophic event, a chromosome or chromosomal region breaks into dozens or hundreds of pieces. The cell's repair machinery then frantically tries to reassemble the pieces, but not always in the right order. The result is a complex SV with multiple breakpoints clustered in one region, often combining deletions, duplications, and inversions.

Chromothripsis is rare in healthy individuals but more common in cancer, where it can drive tumor evolution by disrupting multiple genes at once. When it does occur, it leaves a distinctive signature: many breakpoints concentrated in a small genomic region, with evidence of chaotic reassembly.

17.5 Impact of SVs

SVs don't just passively exist in the genome—they have functional consequences that can affect health and evolution. The impact depends on what type of SV it is and where it occurs.

17.5.1 Gene Dosage Effects

Copy number altering SVs—deletions and duplications—directly affect **gene dosage**. Deletions and duplications contribute to **25–29% of rare protein-truncating events** (Collins et al 2020, Nature). That's comparable to the impact of nonsense SNVs! When you delete an entire exon or gene, the result is complete loss of function from that allele, just as devastating as a premature stop codon.

Duplications can be harmful too. More isn't always better—having three copies of a gene instead of two can disrupt tightly regulated processes. Genes intolerant to loss-of-function SNVs (those with high pLI scores that we learned about earlier) are also intolerant to copy number changes. Both too little and too much gene product can be problematic.

17.5.2 Structural Rearrangements

Copy number neutral SVs—inversions and translocations—don't change gene dosage, but they can still cause problems. An inversion might disrupt a gene by breaking it in the middle, or it might alter gene regulation by moving a gene away from its normal regulatory elements or placing it next to new ones. These effects are generally subtler than complete loss or gain of a gene, and the selection against copy number neutral SVs is weaker.

17.5.3 Selection Pressure

Figure: Natural selection acts differently on structural variants based on their functional impact. This analysis reveals how evolutionary forces shape SV frequencies in human populations. The plot shows selection coefficients (a measure of fitness impact) for different types of SVs across various genomic contexts. Copy number altering SVs—particularly deletions and duplications—that affect protein-coding genes experience strong negative selection, especially in genes intolerant to loss-of-function mutations (high pLI genes). These variants appear at significantly reduced frequencies because they disrupt gene dosage: too little protein from deletions or too much from duplications both impair fitness. In contrast, SVs in noncoding regions such as introns and intergenic sequences face much weaker selection pressure, since disrupting these areas is less likely to affect gene function critically. Copy number neutral SVs like inversions show

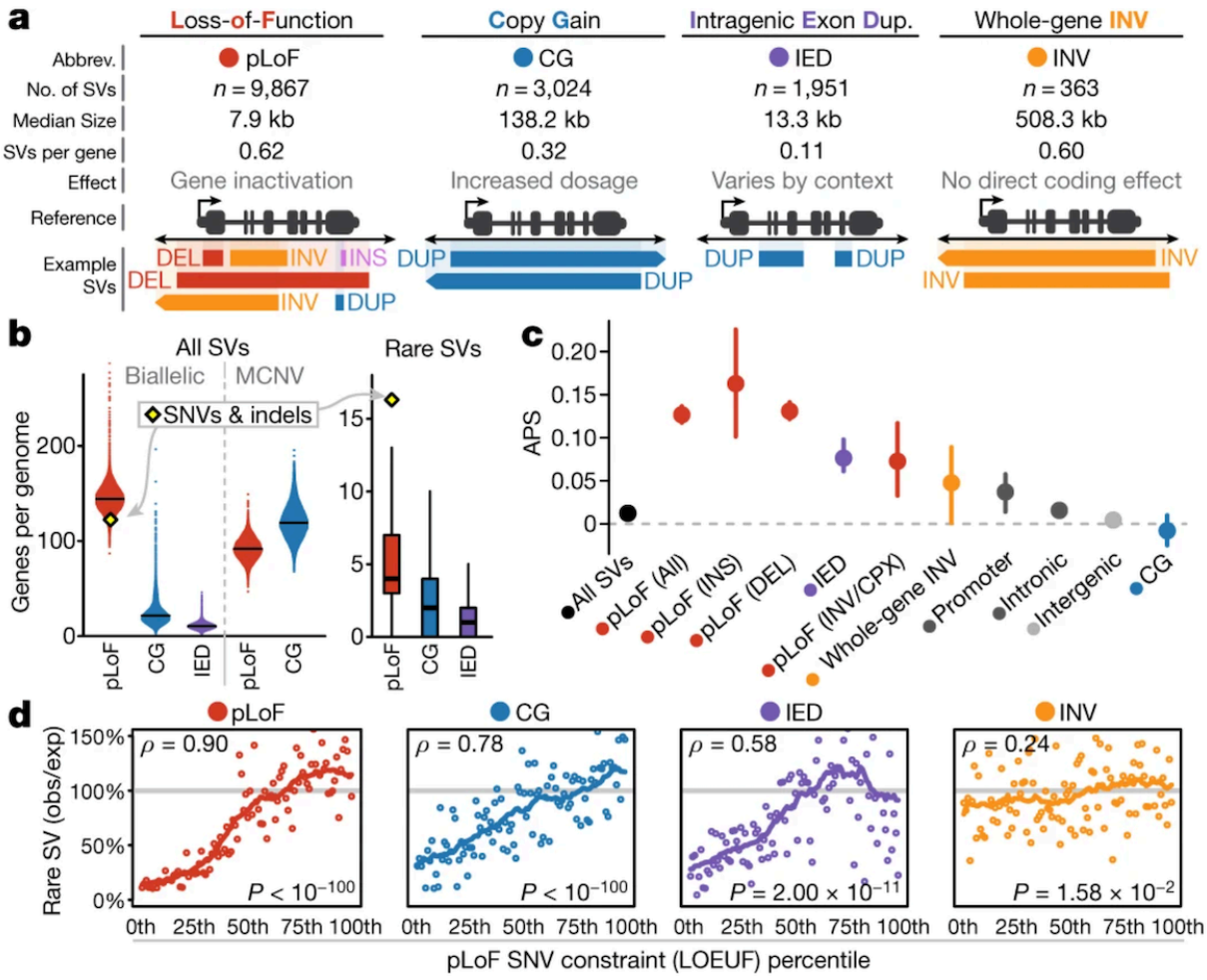


Figure 39: Collins et al. 2020, Nature – Figure 4

intermediate selection pressures—they can disrupt genes at breakpoints or alter regulatory landscapes, but the effects are often subtler than complete gene loss or gain. The pattern mirrors what we observe with single nucleotide variants: genes that cannot tolerate protein-truncating SNVs also cannot tolerate copy number changes, because both ultimately result in abnormal gene function. This demonstrates that whether you disrupt a gene with a point mutation or delete it entirely, natural selection acts to remove variants that eliminate essential genes from the population. The data come from comparing observed SV frequencies to those expected under neutral evolution across 14,237 diverse human genomes. Source: Collins, R.L. et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444–451. https://www.nature.com/articles/s41586-020-2287-8. License: CC-BY 4.0.

Copy number altering SVs in protein-coding genes face strong negative selection—they're kept rare in the population because they reduce fitness. SVs in noncoding regions (like introns or intergenic sequences) face milder selection, since disrupting these regions is less likely to affect gene function.

Interestingly, the pattern mirrors what we see with SNVs: genes that are intolerant to SNVs are also intolerant to SVs. This makes sense—whether you disrupt a gene with a point mutation or by deleting it entirely, the result is loss of function, and selection weeds out variants that eliminate essential genes.

17.6 Summary

Structural variations represent a major source of genetic diversity and disease risk in human populations. Unlike SNVs that alter single nucleotides, SVs involve large-scale rearrangements affecting 50 to millions of base pairs—deletions removing DNA, duplications copying it, inversions flipping it, insertions adding new sequences (often mobile elements), and translocations moving chunks between chromosomes.

The gnomAD-SV reference, built from 14,891 diverse genomes, revealed that each person carries a median of 7,439 SVs. Nearly half are singletons found in just one individual, reflecting both recent mutation and purifying selection against harmful variants. The size distribution shows distinctive peaks at 300, 2,100, and 6,000 base pairs—signatures of Alu, SVA, and LINE1 mobile element insertions that continuously generate new variation.

Multiple molecular mechanisms create SVs. Non-allelic homologous recombination (NAHR) occurs when similar but non-corresponding sequences mistakenly pair during meiosis, producing deletions and duplications. Non-homologous end joining (NHEJ) repairs double-strand breaks by gluing ends together without template guidance, sometimes joining wrong ends to create translocations. Mobile element insertions use a copy-and-paste mechanism via RNA intermediates. Replication fork stalling can trigger template switching, generating complex rearrangements. Chromothripsis—catastrophic chromosome shattering—creates clustered breakpoints with chaotic reassembly, common in cancer but rare in healthy genomes.

SVs have profound functional impacts. Copy number altering SVs contribute 25–29% of rare protein-truncating events, matching the impact of nonsense SNVs. Deletions and duplications alter gene dosage—both too little and too much protein can be harmful. Genes intolerant to loss-of-function SNVs are equally intolerant to copy number changes. Natural selection strongly suppresses SVs affecting essential genes, keeping them rare, while SVs in noncoding regions face weaker selection.

Understanding SVs is crucial for medical genetics—many developmental disorders, intellectual disabilities, and cancers involve large deletions, duplications, or complex rearrangements. SVs also shape population genetic diversity, with different patterns across ancestries reflecting unique demographic histories. As sequencing improves and more genomes are analyzed, catalogs like gnomAD-SV enable better interpretation of patients' genomes and deepen our understanding of how large-scale rearrangements shape human biology and disease.

18 Part 3. Polygenic Genetics and Complex Traits: From Single Genes to Genomic Landscapes

18.1 Introduction: When Mendel's Laws Meet Thousands of Genes

In Part 2, you learned how Mendel's laws apply to single genes—how one mutation in CFTR causes cystic fibrosis, how KMT2D haploinsufficiency causes Kabuki syndrome, how recessive inheritance requires two mutant alleles. These are clean stories: one gene, clear mechanism, predictable inheritance.

But most traits don't work this way. Height isn't controlled by one gene. Neither is heart disease, diabetes, or intelligence. These traits show continuous variation—smooth bell curves, not discrete categories. They run in families, suggesting genetic influence, yet don't follow Mendel's 3:1 ratios.

This paradox puzzled early geneticists. On one side were the Mendelians, who studied discrete traits like pea color that came in clear categories and followed simple ratios. On the other side were the biometricians, who measured continuous traits like height that showed smooth variation and no obvious Mendelian patterns. The two groups thought they were studying fundamentally different phenomena.



Figure 40: Figure: Two perspectives on inheritance

Left: Mendelian genetics explains discrete traits (green vs yellow peas) with clear ratios. Right: Continuous traits like height show smooth parent–offspring correlations without obvious categories—leaving early geneticists puzzled.

In 1918, statistician R.A. Fisher resolved this paradox with a simple insight: **continuous variation can arise from many genes, each following Mendel's laws, each contributing a small additive effect**. If hundreds of loci affect height, and you inherit random combinations from your parents, the result is a bell curve—not because inheritance has changed, but because you're summing many independent Mendelian factors.

Fisher couldn't prove this directly. He worked with mathematics and trait correlations between relatives, deducing that many genes must exist because that's what the data required. For 80 years, his model remained elegant theory.

Then came genome-wide association studies (GWAS). We scanned millions of variants across hundreds of thousands of people and found Fisher was right: height is influenced by **over 12,000 genetic variants** across the genome. Type 2 diabetes involves hundreds of loci. Schizophrenia has 287 associated regions. Each variant has a tiny effect—shifting phenotype by milligrams or millimeters—but together they explain substantial heritability.

This chapter shows how Mendel's laws scale from single genes to genomic complexity, how we find and interpret thousands of small-effect variants, and why the same disease can arise through completely different genetic routes that converge on shared biology.

18.2 What This Chapter Is About

This chapter has six interconnected sections explaining how we interpret complex genetic architecture:

18.2.1 1. Additive vs. Dominant Alleles

How do alleles combine their effects? Using height as our model, we'll contrast dominant disorders (achondroplasia, Marfan syndrome) where one mutation has a large effect, with normal variation where thousands of additive variants each contribute tiny amounts. You'll understand why most traits show predominantly additive effects—and what "orthogonalized dominance" means when testing this.

18.2.2 2. Heritability: What It Means and What It Doesn't

What does "height is 68% heritable" actually mean? This section tackles genetics' most misunderstood concept. Using Fisher's framework and modern twin/SNP studies, you'll learn why heritability measures population variance (not individual determination), why it changes across environments, and why high heritability doesn't mean traits are fixed or that environment doesn't matter.

18.2.3 3. Variance Partitioning in Practice

How does partitioning genetic and environmental variance answer biomedical questions? We'll examine two recent studies: one showing the gut microbiome doesn't cause autism (it reflects dietary preferences), another showing prenatal acetaminophen doesn't cause autism (the association was due to familial confounding). Both demonstrate why properly accounting for shared genetic and environmental factors is essential for causal inference.

18.2.4 4. Genome-Wide Association Studies (GWAS)

How do we actually find the thousands of variants influencing complex traits? You'll learn how GWAS works, how to read Manhattan and QQ plots, what genome-wide significance ($p < 5\times10^{-8}$) means and why it's necessary, and how researchers translate statistical associations into biological understanding through pathway and cell-type analyses. We'll also cover polygenic scores—combining thousands of tiny effects into risk predictions.

18.2.5 5. Genetic Architecture of Complex Diseases

Why do different people develop the same disease through different genetic routes? Using dilated cardiomyopathy, Alzheimer's disease, and autism as examples, you'll see how diseases have composite genetic architecture—rare high-impact mutations in some patients, polygenic risk in others, or combinations of both. Despite genetic heterogeneity, variants often converge on shared biological pathways.

18.2.6 6. From Statistics to Mechanism

How do we connect GWAS loci to causal biology? This section shows how statistical associations point to specific genes, pathways, and cell types. You'll understand the difference between a GWAS locus (a chromosomal region) and the actual causal variant, why most associations are in non-coding regions, and how functional studies validate predicted mechanisms.

18.3 From Pea Ratios to Polygenic Scores

Consider type 2 diabetes—a complex metabolic disease. The traditional view: - Runs in families (genetic component) - Influenced by diet and exercise (environmental component) - No clear Mendelian pattern (not a single-gene disorder) - Mechanism unclear (polygenic? environmental? both?)

The modern genomic view: – GWAS identifies ~400 associated loci across the genome – Each variant increases risk by 1–10% (tiny individual effects) – Polygenic score combining all loci predicts ~5–fold risk difference between top and bottom deciles – Rare monogenic forms exist (MODY, neonatal diabetes) but represent ⟨5% of cases – Common and rare variants affect same pathways: insulin secretion, glucose metabolism, beta–cell function – Confirmation: most diabetes is polygenic, but specific genes/pathways are identifiable

This is still Mendelian inheritance at each locus—alleles segregate according to Mendel's laws. But now we observe: – Thousands of loci simultaneously – How tiny effects combine into substantial risk – Why family history predicts risk (shared polygenic background) – How rare and common variants interact – Which biological pathways matter

This chapter focuses on complexity: moving from single–gene thinking to genomic–scale interpretation.

18.4 How to Approach This Chapter

18.4.1 Think Additively, Not Categorically

Most genetic effects are additive. Two copies of a risk allele □ twice the effect of one copy. This seems simple, but it has profound implications: – Traits show continuous variation (bell curves) – Many genes of small effect can produce large phenotypic differences – Family resemblance reflects shared polygenic background – Dominance effects are rare and weak for most traits

18.4.2 Understand What Heritability Does and Doesn't Tell You

"Height is 68% heritable" means: – 68% of height variation in this population is due to genetic variation – NOT: 68% of your height is genetic (that's meaningless) – NOT: height is 68% determined by genes (ignores gene–environment interplay) – NOT: environment only matters 32% (heritability changes with environment)

Heritability is a population statistic, not an individual property.

18.4.3 Connect Multiple Evidence Streams

Interpreting complex traits requires integrating: – **GWAS**: Which loci are associated? (statistical evidence) – **Functional studies**: What do these genes do? (biological mechanism) – **Pathway analysis**: Do variants converge on shared biology? (functional coherence) – **Cell–type enrichment**: Where do these genes act? (tissue specificity) – **Polygenic scores**: What's the cumulative effect? (predictive power) – **Rare variants**: Are there high–impact mutations too? (genetic architecture)

No single analysis suffices. Complex traits require comprehensive approaches.

18.4.4 Recognize the Spectrum

Traits aren't purely Mendelian or purely polygenic—they exist on a continuum: – Height: mostly polygenic, some monogenic cases (achondroplasia, Marfan) – Autism: mostly polygenic, ~10–20% monogenic (CHD8, SCN2A, fragile X) – Alzheimer's: mostly polygenic, rare early–onset forms (APP, PSEN1/2) – Coronary artery disease: mostly polygenic, rare monogenic (familial hypercholesterolemia)

The same trait can follow different genetic rules in different individuals.

18.5 What Makes This Different From Mendelian Genetics

Mendelian Approach	Polygenic Approach
Single gene, large effect	Thousands of genes, tiny effects
Discrete categories (affected/unaffected)	Continuous variation (bell curves)
Clear inheritance patterns (3:1 ratios)	Familial clustering without simple ratios
High penetrance (mutation → disease)	Low penetrance (risk alleles increase probability)
Study rare families with clear segregation	Study large populations with subtle associations
One mutation explains phenotype	Thousands of variants contribute to phenotype
Predict from single genotype	Predict from polygenic score

Both frameworks are valid—they describe different ends of the genetic architecture spectrum.

18.6 Learning Objectives

By the end of this chapter, you should be able to:

1. **Explain Fisher's insight** that continuous traits arise from many genes with small additive effects, and understand how this unified Mendelian and biometric perspectives
2. **Distinguish additive from dominant alleles** using molecular mechanisms and population data, and understand why most effects are additive
3. **Interpret heritability correctly** as a population statistic measuring variance components, not as individual genetic determination
4. **Apply variance partitioning** to distinguish correlation from causation in observational studies, recognizing how shared genetic and environmental factors create confounding
5. **Read and interpret GWAS results** including Manhattan plots, QQ plots, pathway enrichment, and cell-type analyses
6. **Calculate and understand polygenic scores**, including their construction, interpretation, limitations, and ethical implications
7. **Describe genetic architecture** of complex diseases, recognizing that rare and common variants can contribute to the same phenotype through convergent pathways
8. **Connect GWAS associations to biological mechanisms** by integrating statistical evidence with functional studies, pathway analyses, and cell-type data

---

18.7 Let's Begin

Part 2 showed you how to interpret single variants—determining whether a mutation causes disease, understanding dominance mechanisms, applying Mendel's laws to sequence data.

Part 3 shows you how to interpret thousands of variants simultaneously—understanding how tiny effects combine, why traits show continuous variation, and how different genetic routes can produce the same disease.

Fisher predicted in 1918 that many genes of small effect must exist. GWAS proved him right a century later. Now you'll learn how to find those genes, interpret their effects, and understand the complex genetic architecture underlying most human traits and diseases.

Mendel gave us the rules for single genes. Fisher showed how those rules scale to thousands of genes. Modern genomics makes both frameworks visible in sequence data.

Let's see how genetic complexity emerges from simple principles.

19 Chapter 14. Additive and Dominant Alleles

So far, we've talked about how mutations can disrupt genes and cause disease. But here's something interesting: not all genetic variants work the same way. Some variants are like a light switch—flip it once, and the whole room changes. Others are more like dimmer switches—you need to adjust many of them to see a real difference.

Let's use height as our example. Height is something we can all relate to, and it turns out genetics influences it in two very different ways. We'll look at **dominant alleles**—the "light switch" kind—that cause dramatic changes with just one mutated copy. Then we'll explore **additive alleles**—the "dimmer switch" kind—where hundreds or even thousands of genetic variants each nudge your height up or down by tiny amounts.

To see dominant alleles in action, we'll examine two genetic disorders: Achondroplasia (which causes much shorter stature) and Marfan syndrome (which causes much taller stature). Then we'll contrast these with how height works in most people, where common and rare variants team up to create the familiar bell curve of human heights. All heights here are in centimeters (cm), the standard unit for medical measurements.

19.1 Dominant Alleles and Achondroplasia

Achondroplasia is the most common form of dwarfism, affecting about 1 in every 15,000 to 25,000 babies born. The name literally means "without cartilage formation," which is a bit misleading. People with achondroplasia can form cartilage just fine—the problem is what happens next.

19.1.1 How It Works

Think of your bones as having two phases of development. First, your body lays down a cartilage template (cartilage is that tough but flexible tissue you can feel in your nose and ears). Then, during childhood and adolescence, that cartilage gradually converts into hard bone through a process called ossification. In achondroplasia, this conversion process gets disrupted, especially in the long bones of your arms and legs.

The result? Without treatment, adult males with achondroplasia average about 131 cm tall (roughly 4 feet 4 inches), while adult females average 124 cm (about 4 feet 1 inch). Compare that to average heights of around 175 cm for men and 162 cm for women in many populations.

But height isn't the only feature. People with achondroplasia typically have: – An average–sized torso (their trunk is normal length) – Short arms and legs, with the upper arms and thighs especially shortened—doctors call this "rhizomelia" – Limited ability to fully straighten their elbows – A larger–than–average head (macrocephaly) with a prominent forehead – Short fingers, where the ring and middle fingers sometimes spread apart, giving the hand a "trident" (three–pronged) appearance

Here's something that might surprise you: about 80% of achondroplasia cases pop up out of nowhere. The parents have completely normal stature, and the mutation happens brand new (we call these de novo or sporadic mutations) in the egg or sperm that made that child (Ornitz et al. 2017, Dev Dyn).

19.1.2 The Genetic Culprit: FGFR3

Achondroplasia is caused by mutations in a gene called **FGFR3**, which sits on the short arm of chromosome 4. FGFR3 stands for "Fibroblast Growth Factor Receptor 3"—quite a mouthful, but the name tells you what it does. This gene makes a receptor protein that sits on the surface of cartilage cells (chondrocytes) in your growth plates. Growth plates are the zones of growing tissue near the ends of your long bones—they're what make you taller during childhood.

Now here's the kicker: nearly everyone with achondroplasia has the exact same mutation. It's a single amino acid swap—glycine gets replaced by arginine at position 380 of the protein. Scientists write this as **p.Gly380Arg** (or G380R for short). This one-letter change in the genetic code shows up in achondroplasia patients across different populations worldwide. It's like finding the same typo in books printed on different continents—it tells us this particular mutation is the main troublemaker.

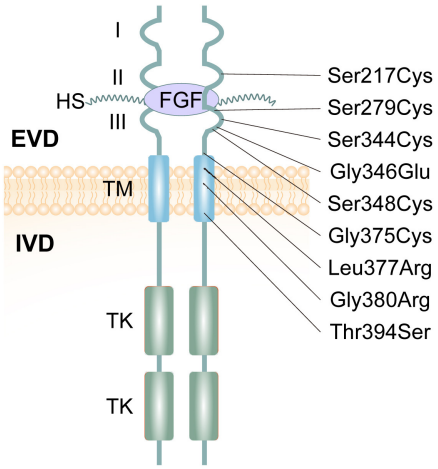


Figure 41: Sobreira2024-HealthMed-Fig1

Figure: FGFR3 receptor structure and achondroplasia mutation sites. The FGFR3 receptor spans the cell membrane. Outside the cell, the extracellular domain (ECD) contains three immunoglobulin-like regions (I, II, III) that can bind to growth factors and heparan sulfate (HS). The transmembrane domain (TM) anchors the receptor in the membrane—this is where the G380R mutation is located. Inside the cell, the intracellular domain (ICD) contains tyrosine kinase (TK) regions that send signals when the receptor is activated. Most achondroplasia mutations cluster in the TM region, right where G380R sits. Source: Sobreira, N. et al. (2024). Achondroplasia. HealthMedicine. https://www.sciencedirect.com/science/article/pii/S2352304224002332. License: Creative Commons.

19.1.3 Why This Mutation Causes Problems

Normally, FGFR3 acts like a brake on bone growth. When growth factors bind to the receptor, two receptor molecules come together (dimerize), activate each other, and send a signal into the cell saying "slow down bone growth." This is actually healthy—it prevents your bones from growing too much.

But the G380R mutation changes everything. This mutation sits in the transmembrane domain—the part of the receptor buried in the cell membrane. The swap from glycine (a small, flexible amino acid) to arginine (a much larger one with a bulky side chain) makes the receptor behave differently. It promotes dimerization even when no growth factor is around (Hartl et al. 2022, bioRxiv).

Think of it like a car alarm that goes off even when nobody's trying to break in. The receptor is stuck in the "on" position, constantly telling growth plate chondrocytes to slow down—even when they should be growing normally. The receptor activates signaling pathways (MAPK and STAT1) that suppress chondrocyte proliferation and maturation. Fewer dividing chondrocytes and less cartilage production means less bone growth, especially in the long bones.

Figure: G380R mutation increases FGFR3 signaling activity. This graph shows how strongly different FGFR3 variants activate downstream signaling, measured by GRB2 (an adaptor protein) recruitment. Higher bars indicate stronger signaling. The G380R variant shows elevated activation even without adding FGF ligand—the receptor signals constitutively, which doesn't occur with the wild-type receptor. This ligand-independent activation explains why one mutated copy is sufficient to cause achondroplasia.

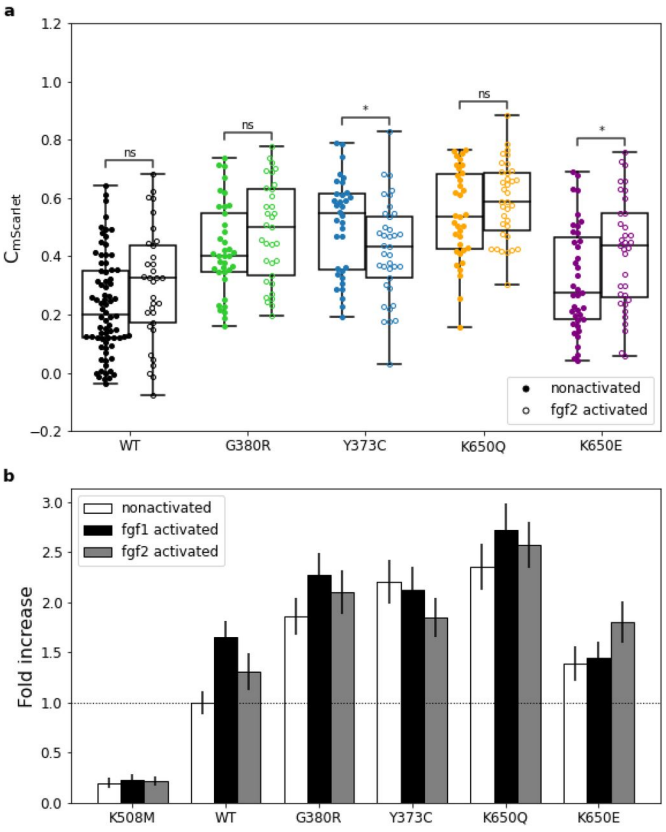


Figure 42: Hartl2022-bioRxiv-Fig4

Source: Hartl, D. et al. (2022). Molecular mechanisms underlying pathogenic FGFR3 mutations. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.04.11.487861v1.full License: CC-BY-NC-ND 4.0.

19.1.4 Why It's Dominant

Here's the key: achondroplasia is a **dominant disorder**. You only need one mutated copy of FGFR3 to have the condition. Even though you also have one perfectly normal copy inherited from your other parent, that normal copy can't compensate. The mutant receptor is so overactive that it dominates the situation—hence "dominant allele."

19.2 Dominant Alleles and Marfan Syndrome

Let's look at the flip side: a dominant disorder that makes people taller instead of shorter. Marfan syndrome is an autosomal dominant disorder affecting connective tissue—the material that holds your body together. People with Marfan syndrome are usually much taller than average: adult males often exceed 190 cm (about 6 feet 3 inches), while females typically reach 175–180 cm (5 feet 9 inches to 5 feet 11 inches) or more.

But tall stature is just one piece of the puzzle. Marfan syndrome comes with a distinctive set of features:
- Long, slender limbs and fingers (doctors call the long fingers "arachnodactyly"—literally "spider fingers")
- Lens dislocation in the eye (ectopia lentis)—the lens can slip out of place - A predisposition to thoracic aortic aneurysms—dangerous bulges in the body's main artery that can rupture

Unlike achondroplasia, which shortens the limbs while keeping the trunk normal, Marfan syndrome elongates body proportions overall (Sakai et al. 2016, Genes).

19.2.1 The Genetic Culprit: FBN1

Marfan syndrome is caused by mutations in the **FBN1** gene on chromosome 15. FBN1 encodes a protein called **fibrillin-1**, which is a structural protein. Think of fibrillin-1 as microscopic ropes that form the scaffolding of connective tissue. These "ropes" (microfibrils) provide strength and elasticity to tissues like the aorta, the ligaments that hold your eye lens in place, and the connective tissue in your bones.

Unlike achondroplasia—where nearly everyone has the same mutation—Marfan syndrome is caused by over **1,000 different mutations** in FBN1. These include missense mutations (amino acid swaps), nonsense mutations (premature stop signals), splice-site mutations (errors in how the gene is processed), and small insertions or deletions. Many of these mutations hit either the **calcium-binding EGF-like domains** or the **8-cysteine domains** of fibrillin-1, disrupting the protein's structure or how it folds (Learn.Genetics Utah).

See Figure: Fibrillin-1 structure from Learn.Genetics Utah. Fibrillin-1 proteins assemble into long, strong fibers that reinforce connective tissue throughout the body. When there is insufficient fibrillin-1, or when the protein is misshapen, the scaffolding is weak—this is what happens in Marfan syndrome.

19.2.2 Why FBN1 Mutations Cause Tall Stature

FBN1 mutations are **dominant alleles**—one bad copy is enough to cause Marfan syndrome. But how do defective microfibrils lead to excessive height?

It turns out fibrillin-1 has a second job beyond structural support: it regulates a signaling molecule called **TGF-β** (transforming growth factor beta). Normally, fibrillin-1 microfibrils capture and sequester TGF-β, keeping it in check. When fibrillin-1 is defective, TGF-β escapes this regulation and becomes overactive.

Excessive TGF-β signaling promotes abnormal growth in connective tissues, including bones. It drives long bone elongation beyond normal limits, resulting in the tall stature and elongated limbs characteristic of Marfan syndrome. So even though you have one normal FBN1 copy, the defective copy disrupts the scaffolding and unleashes too much TGF-β—enough to override the normal copy's function. That's why it's dominant.

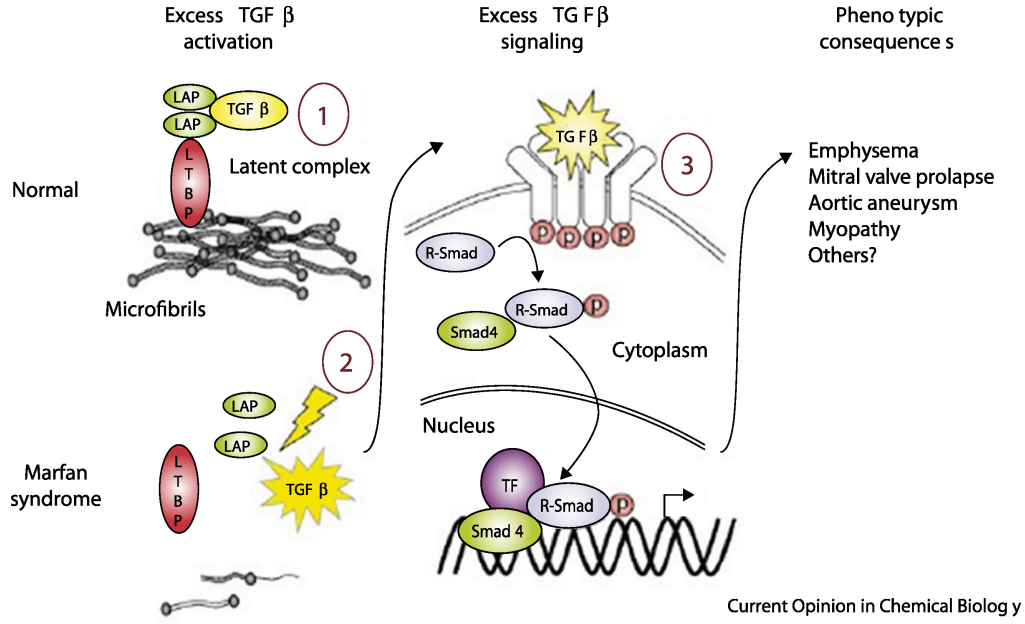


Figure 43: Coelho2020-CellSignal-Fig4

Figure: TGF-β dysregulation in Marfan syndrome. Left: In healthy tissue, fibrillin-1 microfibrils sequester TGF-β by binding to latency-associated peptide (LAP) and latent TGF-β binding protein (LTBP), keeping TGF-β signaling in check. Right: When microfibrils are deficient (as in Marfan syndrome), TGF-β is released and activates its receptors. This triggers excessive signaling through Smad proteins and transcription factors, driving abnormal tissue growth and the skeletal features of Marfan syndrome. Source: Coelho, S.G. & Almeida, A.G. (2020). Marfan syndrome revisited: From genetics to the clinic. Cellular Signalling. https://www.sciencedirect.com/science/article/pii/S217420492030115X. License: Creative Commons.

19.3 Additive Alleles and Normal Variation in Human Height

Now let's shift gears. We've just seen how single dominant mutations in FGFR3 or FBN1 can dramatically alter height. But here's the thing: most people don't have achondroplasia or Marfan syndrome. For the vast majority of us, height is influenced by **additive alleles**—lots of genetic variants, both common and rare, each contributing a tiny nudge up or down.

Instead of one mutation flipping a switch, imagine thousands of little dials all adjusting your height slightly. Some dial you a bit taller, others a bit shorter. The cumulative effect of all these dials determines where you land on the height spectrum. This is why human height follows a **normal distribution**—a bell curve—with most people clustering around the average and fewer people at the very short or very tall extremes.

19.3.1 The Genetic Basis of Height Variation

Let's look at some numbers to appreciate just how polygenic (literally "many genes") height really is.

A massive genome-wide association study (GWAS) analyzed **5.4 million people** from diverse ancestries (Yengo et al. 2022, Nature). The study identified **12,111 independent single-nucleotide polymorphisms (SNPs)**—single-letter DNA variants—associated with height. Almost all of these SNPs are **common variants**, meaning they appear in more than 1% of the population (we abbreviate this as MAF $\rangle$ 1%, where MAF stands for "minor allele frequency").

These 12,111 SNPs are scattered across **7,209 different genomic regions**, covering about 21% of the

entire genome. Together, they explain most of the height variation we can detect using common SNPs. Interestingly, genes like **FGFR3** and **FBN1** show up in this list—but as common variants with subtle effects, not the severe mutations causing achondroplasia or Marfan syndrome. Other genes in the list regulate growth-plate function and connective-tissue development, just like the Mendelian examples.

But common variants aren't the whole story. A complementary study used whole-genome sequencing on **25,465 unrelated Europeans** to explore the role of **rare variants**—those with frequencies below 0.1% (Wainschtein et al. 2022, Nat Genet). SNP-based heritability of height was estimated at **0.68** (with a standard error of 0.10), meaning about 68% of height variation can be attributed to genetic differences.

Here's where it gets interesting. Rare variants contribute substantially, but how they contribute depends on something called **linkage disequilibrium (LD)**. LD measures how tightly linked a variant is to nearby variants. **Low-LD rare variants**—those that aren't correlated with nearby common variants—accounted for **0.31** (31%) of the phenotypic variance. In contrast, **high-LD rare variants**—those tightly linked to common variants already captured by GWAS—contributed only **0.03** (3%).

What does this mean? Low-LD rare variants provide new, independent genetic information that GWAS can't easily pick up. High-LD variants, on the other hand, are essentially "tagging along" with common variants we already knew about, so they don't add much new information. These rare variants are often protein-altering changes under negative selection (evolution weeds them out because they're slightly harmful), but they still contribute **additively**, not dominantly.

19.3.2 How Additive Alleles Work

Let's make this concrete with an example. Imagine you inherit a height-increasing SNP that adds 0.2 cm to your height. You also inherit a height-decreasing SNP that subtracts 0.15 cm. Another SNP adds 0.1 cm. Another subtracts 0.05 cm. And so on, thousands of times.

Your final height is the net sum of all these tiny increments and decrements, plus environmental factors like nutrition and health during childhood. Because each individual SNP has such a small effect, you need to add up thousands of them to explain someone's height.

This additive model creates the familiar bell curve. Most people inherit a mix of height-increasing and height-decreasing variants that roughly balance out, putting them near the population average. A few people inherit more height-increasing variants than average and end up taller. A few inherit more height-decreasing variants and end up shorter. But nobody's height is determined by a single dominant mutation—unless they have a Mendelian disorder like achondroplasia or Marfan syndrome.

Rare variants enhance this picture by capturing additional variance that common SNPs miss. But crucially, they still work additively. Even if you have a rare protein-altering variant in, say, a growth-related gene, it doesn't override your other variants the way a dominant mutation would. It just adds its small effect to the pile.

19.4 Comparing Dominant and Additive Alleles

Let's put everything together. We've looked at three scenarios:

Achondroplasia (Dominant): Caused by mutations in the **FGFR3** gene on chromosome 4, almost always the p.Gly380Arg substitution. This single mutation is enough to cause short stature—about 131 cm in adult males and 124 cm in females. The mutation makes the FGFR3 receptor overactive, suppressing growth plate chondrocyte proliferation. It's a **dominant allele**: one copy causes the full effect, and penetrance is complete (everyone with the mutation has the disorder).

Marfan syndrome (Dominant): Caused by over 1,000 different mutations in the **FBN1** gene on chromosome 15. These mutations disrupt fibrillin-1 microfibrils, weakening connective tissue and unleashing excessive TGF-β signaling. The result is tall stature—often above 190 cm in males—plus other features like lens dislocation and aortic aneurysms. Like achondroplasia, it's a **dominant allele**: one bad copy is enough, and penetrance is complete.

Normal height variation (Additive): Influenced by thousands of variants across multiple genes—including FGFR3, FBN1, and many others. The 2022 GWAS identified **12,111 common SNPs** (MAF 〉 1%) spread across 7,209 genomic regions. Rare variants (MAF 〈 0.1%), especially those in low LD, contribute an additional 31% of the phenotypic variance. Each variant nudges height up or down by a tiny amount—fractions of a centimeter. Your height is the cumulative sum of all these effects. Most people end up somewhere between 157 cm and 183 cm (roughly 5 feet 2 inches to 6 feet), depending on ancestry and sex. These are **additive alleles**: small effects that add up, with variable penetrance depending on how many you inherit.

The contrast is striking. Dominant alleles like those in achondroplasia and Marfan syndrome act like master switches, overpowering the normal copy and producing a predictable phenotype. Additive alleles, on the other hand, are like a committee—each member gets a tiny vote, and the outcome depends on the majority decision across thousands of votes.

20 Chapter 15. Polygenic Model in Traits

In the last section, we saw how dominant alleles like those causing achondroplasia and Marfan syndrome can dramatically alter height with just one mutated copy. But we also learned that most people's height comes from thousands of additive alleles, each nudging the trait up or down by tiny amounts.

Now let's zoom out and ask: Is height unusual, or do most human traits work this way? And how do scientists actually measure whether a trait is shaped more by additive effects or dominant effects?

To answer these questions, we'll look at a massive study from the **UK Biobank**. The study analyzed over 1,000 different human traits—everything from height and weight to disease risks—in more than 360,000 people. The findings confirm what we suspected: most traits are influenced by many genetic variants acting additively, while dominant effects are surprisingly rare.

20.1 Most Traits Are Influenced by Additive Alleles

Let's start with a quick refresher on how additive alleles work. Imagine you have a genetic variant with two versions: allele A and allele B. Since you inherit one copy from each parent, you can have three possible genotypes: AA, AB, or BB.

With **additive alleles**, each B allele adds a consistent small effect. Let's use height as an example: - If you have **AA**, that's your baseline height—let's say 170 cm. - If you have **AB** (one copy of B), you might be 170.5 cm—half a centimeter taller. - If you have **BB** (two copies of B), you'd be 171 cm—a full centimeter taller than AA.

Notice the pattern: each additional B allele adds the same amount (0.5 cm). That's what makes it additive. It's like stacking blocks—each block adds the same height.

Figure: Three modes of allelic effects on traits. Panel A shows additive effects, where trait values increase linearly with genotype dosage (0, 1, 2 copies). The effect of having two copies is exactly twice the effect of having one copy. Panel B shows dominance, where heterozygotes (AB) deviate from the additive expectation—having one copy produces most or all of the effect. Panel C shows overdominance, where heterozygotes have more extreme trait values than either homozygote. Each panel shows examples with different allele frequencies (p) and effect sizes (q). The blue–red bars indicate the relative contribution of additive (blue) versus dominance (red) variance. For most human traits, additive effects dominate. Source: Palmer, C. et al. (2021). Quantifying the genetic contribution of dominance in complex traits. bioRxiv. https://www.biorxiv.org/content/10.1101/2021.08.15.456387v2.full. License: CC-BY-NC-ND 4.0.

Now, here's the key finding. The study examined **361,194 individuals** across **1,060 different traits**—physical measurements like height and BMI, blood biomarkers, disease outcomes, and more. The analysis

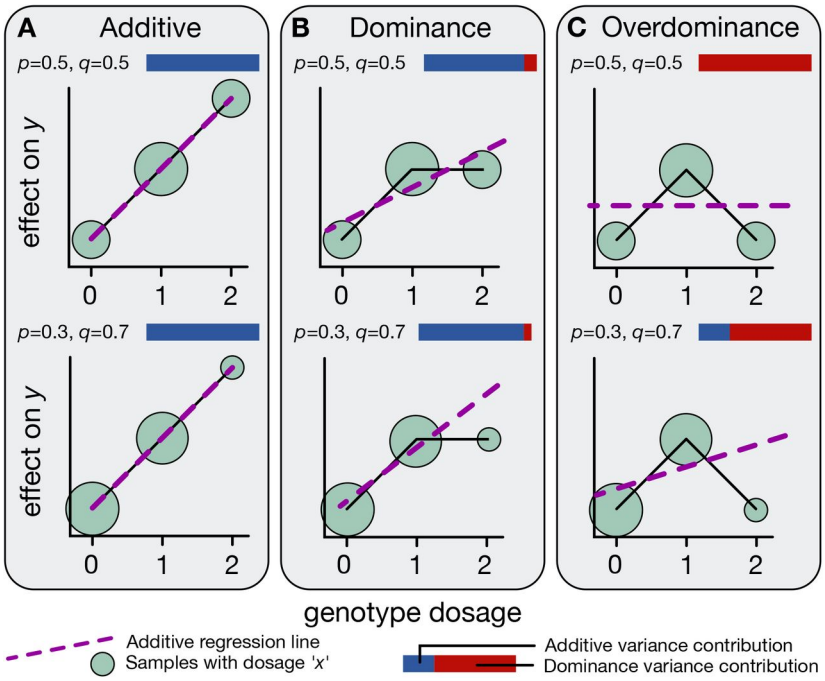


Figure 44: Palmer2021-bioRxiv-Fig1

tested 13.7 million genetic variants to see which traits showed additive versus dominant effects (Palmer et al. 2023, Science).

The result? About **700 of the 1,060 traits** showed detectable additive heritability. In other words, for roughly two-thirds of the traits examined, genetic variation could be explained by additive alleles acting together. This confirms that most human traits are **polygenic**—shaped by many small-effect variants rather than one or two big ones.

But what about dominant effects? Remember, **dominant alleles** have a strong impact with just one copy. We've seen this with achondroplasia (FGFR3) and Marfan syndrome (FBN1), where a single mutated allele overrides the normal one and causes extreme changes.

The study found **183 genomic loci** with detectable dominant effects. One example is **MC1R**, a gene affecting hair color—having just one copy of certain MC1R variants is enough to influence your hair color noticeably. But here's the kicker: even though 183 loci were identified, dominant effects explained only about **0.5% of the additive genetic variance** on average. They're there, but they're rare and weak compared to the cumulative power of additive effects.

Figure: Genome-wide distribution of additive and dominance effects across traits. This Manhattan-style plot shows the locations of genetic loci with significant additive (bottom, pink/red) and dominance (top, green/blue) effects across all chromosomes. Each vertical line represents a genomic region associated with one or more traits from different categories (cognitive function, physical traits, health biomarkers, etc.). The vast majority of associations show additive effects (bottom panel is densely populated), while dominance effects (top panel) are sparse—only 183 loci detected across all traits examined. Notable genes with dominance effects are labeled (MC1R, APOE, etc.). This visualization confirms that additive genetic architecture dominates human complex traits. Source: Palmer, C. et al. (2021). Quantifying the genetic contribution of dominance in complex traits. bioRxiv. https://www.biorxiv.org/content/10.1101/2021.08.15.456387v2.full. License: CC-BY-NC-ND 4.0.

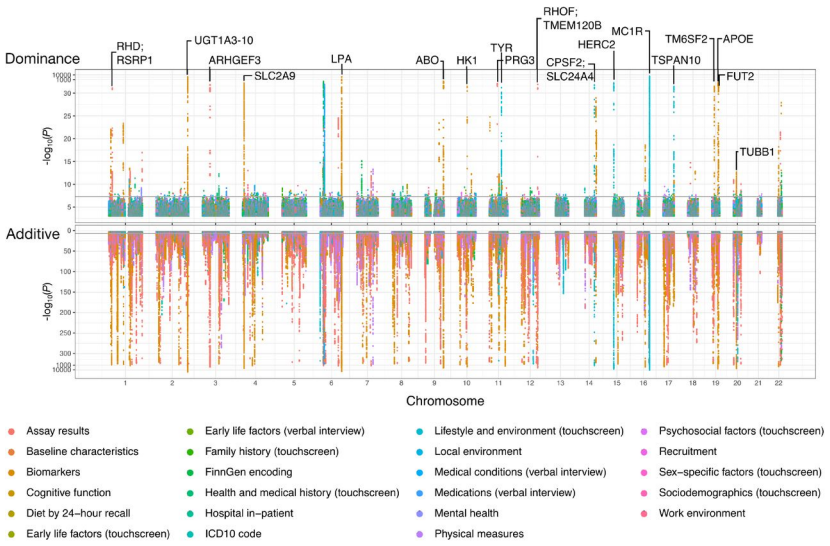


Figure 45: Palmer2021−bioRxiv−Fig2

20.2 How Scientists Measure Additive and Dominant Effects

The UK Biobank is one of the largest genetic datasets in existence. It includes genetic information and health records for over 500,000 people in the United Kingdom. The study used **361,194 participants** and examined **13.7 million genetic variants** to understand how these variants influence traits.

The approach builds on a method called **linkage disequilibrium score regression (LDSC)**. LDSC is a statistical technique that estimates how much of a trait's variation comes from genetics, without needing to identify every single causal variant. The method was extended to separately measure **additive** and **dominance** components of heritability. This new approach was called **dominance LDSC (d−LDSC)**, and it allowed researchers to tease apart how much extra variation dominance effects add beyond what additive effects already explain.

Let's break down how each type of effect is measured.

20.2.1 Measuring Additive Effects

Measuring additive effects is pretty straightforward. You look at how the trait changes as you go from AA to AB to BB. If the trait increases in a steady, predictable way, that's additive.

Let's use a concrete example. Suppose we're measuring height: − **AA genotype**: You have two copies of allele A. Let's call this your baseline—170 cm. − **AB genotype**: You have one A and one B. If B is a height−increasing allele, you might be 170.5 cm. − **BB genotype**: You have two copies of B. You'd be 171 cm.

We can encode this numerically as 0, 1, 2—counting the number of B alleles. AA gets 0, AB gets 1, BB gets 2. If the trait value increases proportionally (0 → 1 → 2), the effect is additive.

The study found that for most traits, this additive pattern captures nearly all the measurable genetic influence. The trait value scales smoothly with the number of effect alleles you carry.

20.2.2 Measuring Dominant Effects

Dominance is trickier to measure. By definition, dominance means that having one copy of a variant (AB) is enough to produce a noticeable effect—you don't need two copies.

Let's think about what complete dominance looks like. Imagine a trait where the B allele is completely dominant over A: - **AA**: Baseline trait value—let's say 0. - **AB**: The trait changes—let's say to 1. - **BB**: The trait stays at 1 (same as AB).

So the pattern is 0–1–1. Having one B is enough; a second B doesn't add anything extra.

But here's the problem: this 0–1–1 pattern overlaps with the additive pattern 0–1–2. When you look at AB, you can't immediately tell whether it's following a dominant pattern or just the halfway point of an additive pattern. The AB genotype could be showing dominance, or it could just be intermediate between AA and BB.

To separate these two possibilities, the study used a trick called **orthogonalized dominance encoding**. This mathematical technique isolates the heterozygote effect so it doesn't overlap with the additive signal.

Here's how it works in practice: - **Additive encoding**: AA = 0, AB = 1, BB = 2 (the steady increase) - **Orthogonalized dominance encoding**: AA = 0, AB = 1, BB = 0 (the special heterozygote effect)

Notice the dominance encoding: it's 0–1–0. This captures whether being a heterozygote (AB) gives you something special—something that can't be explained by simply counting alleles. If AB deviates from the linear trend between AA and BB, that's dominance.

The key word here is "orthogonalized." In math, orthogonal means "independent" or "perpendicular"—two things that don't overlap. By orthogonalizing the dominance encoding, the analysis ensures that additive and dominance effects are statistically independent. Each one captures its own unique contribution to the trait.

20.2.3 Explaining "Orthogonalized Dominance" Without Math

Okay, that probably still sounds abstract. Let's try a different analogy.

Imagine you're listening to music through headphones with an equalizer. The equalizer has two controls: - **Volume slider**: This is like the additive effect. Turn it up, and everything gets louder. The change is smooth and predictable. - **Bass boost button**: This is like the dominance effect. It adds a special quality—extra bass—that you can't get by just turning up the volume.

If you adjust both controls at once, it's hard to tell how much of the sound change is from volume versus bass boost. So you need to separate them. You set the volume to a baseline level, then test the bass boost independently.

That's what orthogonalization does. It removes the overlap between additive (volume) and dominance (bass boost) so you can measure each one separately.

In genetic terms: - The **additive effect** measures the steady increase as you go from 0 to 1 to 2 copies of an allele. - The **dominance effect** measures the special boost you get from being a heterozygote (AB), after accounting for the additive trend.

So when the analysis uses orthogonalized dominance encoding (0–1–0), it's asking: "Does being AB give you something extra that can't be explained by just having one B allele?" If yes, that's dominance.

20.2.4 Measuring the Overall Contribution to Traits

Using **additive LDSC** and **dominance LDSC (d-LDSC)**, the analysis quantified how much of each trait's variation comes from additive versus dominant effects.

Here's what was found: - The **additive component** explained a detectable proportion of variance in about **700 of the 1,060 traits**. For most traits, additive effects accounted for a substantial chunk of the genetic influence. - The **dominance component** was detected at **183 genomic loci**, but it accounted for only about **0.5% of the additive variance** on average.

In other words, dominance effects are real—they exist—but they're rare and weak compared to additive effects. For the vast majority of traits, what matters is the cumulative sum of thousands of small additive variants, not a handful of dominant ones.

One more thing: the additive and dominance components were **orthogonal**—statistically independent. This means they captured genuinely different aspects of genetic architecture. Additive effects tell you how many effect alleles you carry. Dominance effects tell you whether being a heterozygote gives you something special.

20.2.5 Why Additive Effects Dominate

So why do additive alleles explain most genetic variation, while dominant effects are so rare?

There are a few reasons. First, additive effects accumulate. Even if each individual variant has a tiny effect, thousands of them together can explain a lot of variation. It's like compound interest—small amounts add up over time (or in this case, across the genome).

Second, dominant effects require a very specific genetic architecture: the heterozygote has to deviate from the additive expectation. This is less common than simple additive effects, where each allele contributes independently.

Third, evolution plays a role. Strongly dominant variants—especially harmful ones—tend to get weeded out by natural selection. If having one copy of a variant is enough to cause a problem, that variant won't stick around in the population for long. The exceptions are dominant disease alleles that arise frequently by new mutation (like the FGFR3 mutation causing achondroplasia) or that have late-onset effects (like Huntington's disease), which selection can't remove efficiently.

What we're left with is a genome full of small-effect additive variants. These variants shape the continuous variation we see in traits like height, BMI, and disease risk. Dominant effects exist—183 loci were found—but they're the exception, not the rule.

This study from the UK Biobank gives us a powerful confirmation: **most human traits are polygenic**, shaped by many additive alleles acting together. Dominant effects are rare and weak. When you look at someone's height, or their risk of diabetes, or their cholesterol levels, you're seeing the cumulative result of thousands of tiny genetic nudges—not one or two big switches.

21 Chapter 16. Heritability

We've been talking about how thousands of genetic variants shape traits like height. Each variant nudges the trait up or down by a tiny amount, and the cumulative effect determines where you land on the spectrum. But here's a question: if we add up all these genetic effects, how much of the variation we see between people is actually due to genetics versus environment?

This question leads us to one of the most important—and most misunderstood—concepts in genetics: **heritability**. It's a number that tells us what proportion of trait variation in a population comes from genetic

differences. But it doesn't tell us whether a trait is "fixed" by genes or how much environment matters. We'll get to why that distinction is so important.

To understand heritability, we need to go back over a century to 1918, when a statistician named R.A. Fisher figured out how to connect Mendel's discrete genetic laws with the continuous variation we see in traits like height. Fisher's framework—called the **polygenic model**—is still the foundation of modern genetics. We'll explore his insights and then connect them to recent studies on human height, BMI, and disease risk.

21.1 Fisher's 1918 Paper and the Polygenic Model

21.1.1 Fisher's Unification of Mendel and Biometry

In the early 1900s, biology had a problem. Two groups of scientists were arguing about how traits are inherited, and they couldn't agree.

On one side were the **Mendelians**. They studied traits that come in discrete categories—like the pea plants Mendel worked with. Round or wrinkled seeds. Tall or short stems. These traits seemed to follow simple rules: you inherit one allele from each parent, and the trait appears in a predictable way.

On the other side were the **biometricians**. They measured traits like human height or weight, which don't come in neat categories. Instead, these traits show continuous variation—a smooth spectrum from short to tall, light to heavy. The biometricians argued that Mendelian laws couldn't explain continuous traits. They thought something else must be going on.

In 1918, R.A. Fisher figured out that both groups were right. His insight was simple but profound: continuous variation can arise from the combined small effects of many Mendelian genes.

Think of it like baking a cake. Each ingredient—flour, sugar, eggs, butter—contributes a little bit to the final taste. No single ingredient determines whether the cake is good or bad. But together, in the right proportions, they create something complex. That's how genes work for traits like height. Hundreds or thousands of genes, each following Mendel's rules, combine to produce the continuous variation we observe.

This 1918 paper laid the foundation for **quantitative genetics**, the field that studies complex traits without needing to know the specific genes involved (Visscher et al. 2019, Genetics). Fisher's framework is still used today, from twin studies to genome-wide association studies (GWAS) to polygenic risk scores.

21.1.2 Partitioning Phenotypic Variance

Fisher's next move was to mathematically break down trait variation into components. He asked: if we measure a trait like height in a population, where does all that variation come from?

He showed that total trait variation—what we call **phenotypic variance (VP)**—can be split into several pieces:

```
VP = VA + VD + VE + VEnvironment
```

Let's unpack each component:

VA (Additive variance) is the variation due to additive alleles—the kind we've been talking about throughout this chapter. Each allele contributes a small, independent effect. If you have more height-increasing alleles, you're taller. If you have fewer, you're shorter. Additive variance is the major source of heritable variation in most human traits.

VD (Dominance variance) captures the extra variation that comes from dominance—when having one copy of an allele (AB) produces an effect that's not simply halfway between AA and BB. We saw in the

last section that dominance effects are rare and weak in most traits. They contribute a bit of variance, but it's usually small.

VE (Epistatic variance) refers to gene–gene interactions—when the effect of one gene depends on which alleles you have at another gene. Epistasis definitely exists, but at the population level, it explains very little of the variance in most traits. Fisher recognized this early on.

VEnvironment is everything non–genetic. Nutrition, lifestyle, chance events during development—all the factors that make you different from someone else even if you have identical genomes. For height, environment includes childhood nutrition, illnesses, and other factors that affect bone growth.

Fisher emphasized that **additive genetic variance (VA)** is the key component. Why? Because additive effects determine how much a trait can respond to selection (evolution) and how well we can predict offspring traits from parent traits. Most of the heritability we measure today reflects additive variance.

And here's the cool part: modern studies confirm Fisher's model. Remember the 2022 GWAS on height? The study found **12,111 SNPs** spread across **7,209 genomic segments**, almost all with additive effects (Yengo et al. 2022, Nature). That's Fisher's model in action, nearly 100 years later.

21.1.3 Additive Alleles and the No–Dominance Assumption

Fisher made a simplifying assumption that turned out to be surprisingly accurate: he assumed that most alleles act additively, without dominance. Each additional copy of an allele adds a fixed amount to the trait value.

Let's use height again. Imagine you have an allele called B that increases height. Under the no–dominance assumption: – If you have **AA** (no B alleles), that's your baseline height—let's say 170 cm. – If you have **AB** (one B allele), you gain 0.5 cm—now you're 170.5 cm. – If you have **BB** (two B alleles), you gain another 0.5 cm—now you're 171 cm.

Notice that the effect is linear. Each B allele adds exactly 0.5 cm, regardless of how many you already have. There's no dominance—AB is exactly halfway between AA and BB.

Fisher knew this was a simplification. Dominance and epistasis do exist. But at the population level, their contributions to overall variance are small. The no–dominance assumption captures most of what's going on, which is why it's still used in many genetic models today.

21.2 Understanding Heritability

21.2.1 What Heritability Means (and Doesn't Mean)

Now we're ready to define **heritability (h^2)**. It's the proportion of phenotypic variance in a population that can be attributed to genetic differences.

For example, if height has an h^2 of 0.68, that means **68% of the variation in height among people** is explained by genetic differences. The other 32% comes from environmental differences—nutrition, health, and so on.

But here's the critical part, and it's where most people get confused: **heritability describes variation within a population, not how fixed or predetermined a trait is for an individual.**

Let me illustrate with an example. Imagine you plant 100 seeds from the same plant variety in two different fields. Field A has rich, well–watered soil. Field B has poor, dry soil.

In Field A, the plants grow to different heights—some are 150 cm, others are 180 cm. Because all the plants get the same excellent conditions, most of that variation is due to genetic differences between seeds. You'd calculate a high heritability—say, $h^2 = 0.80$.

In Field B, the plants also vary in height, but they're all shorter overall—maybe 100 to 130 cm. Again, because they all experience the same poor conditions, most of the variation within Field B is genetic. You'd calculate a similar $h^2 = 0.80$.

Now here's the key insight: even though heritability is high in both fields, the **difference between the two fields is entirely environmental**. The plants in Field A aren't genetically superior—they just had better soil. Heritability tells you nothing about why Field A plants are taller on average. It only tells you why plants within each field differ from each other.

This is why heritability doesn't mean a trait is "fixed" or that environment doesn't matter. High heritability just means that **within the specific environment you're studying**, most of the remaining variation between individuals comes from genetic differences. Change the environment, and heritability can change too.

For human height, $h^2 \approx 0.68$ in modern populations with good nutrition. But if we compared populations with vastly different access to food, environmental variance would increase, and heritability would drop—even though the genes haven't changed at all.

21.2.2 Narrow-Sense and Broad-Sense Heritability

There are actually two types of heritability, and it's worth knowing the difference.

Narrow-sense heritability (h^2) measures the fraction of variance due to additive genetic effects. It's calculated as:

$$h^2 = V_A / V_P$$

where V_A is additive variance and V_P is total phenotypic variance. Narrow-sense heritability is what matters for predicting resemblance between parents and offspring. If you want to know how tall your children will be based on your height, narrow-sense heritability is the relevant number.

Broad-sense heritability (H^2) includes all genetic effects—additive, dominance, and epistatic. It's calculated as:

$$H^2 = (V_A + V_D + V_E) / V_P$$

Broad-sense heritability captures the total genetic influence on a trait, but it's less useful for prediction. Why? Because dominance and epistatic effects don't reliably pass from parent to offspring. When you make a baby, you're shuffling alleles and creating new combinations. Additive effects pass on predictably, but dominance and epistatic effects get scrambled.

In practice, most studies focus on narrow-sense heritability because it's more informative for understanding inheritance and evolution.

21.3 How Heritability Is Estimated

So how do scientists actually calculate heritability? There are several methods, but they all rely on the same basic idea: **closer relatives share more genes and thus have more similar traits**.

21.3.1 Covariance and Relatedness

The key concept here is **covariance between relatives**. Covariance is a statistical measure of how two things vary together. If you and your sibling are both tall, or both short, your heights covary. If one is tall and the other is short, there's less covariance.

Genetic relatives share DNA, so their traits tend to covary. The more DNA they share, the stronger the covariance. Here's what that looks like in practice:

Identical twins share about 100% of their DNA (they're genetic clones). If you measure height in identical twins, you'll find they're very similar—not exactly the same, because environment still plays a role, but very close.

Siblings share about 50% of their DNA on average. Each sibling inherits a random half of each parent's genome. Siblings are more similar in height than unrelated people, but less similar than identical twins.

First cousins share about 12.5% of their DNA. They're more similar than strangers, but much less similar than siblings.

The pattern is clear: **closer relatives, stronger covariance**. This relationship forms the mathematical basis for estimating heritability. If you know how much DNA two relatives share and how similar their traits are, you can work backward to estimate how much of the trait variation is due to genetics.

21.3.2 Twin Studies

Twin studies are one of the classic methods for estimating heritability. The idea is to compare two types of twins:

Monozygotic (MZ) twins—also called identical twins—share about 100% of their DNA. They come from a single fertilized egg that split into two embryos.

Dizygotic (DZ) twins—also called fraternal twins—share about 50% of their DNA, just like regular siblings. They come from two separate fertilized eggs.

Here's the logic: if MZ twins are more similar in a trait than DZ twins, the difference reflects genetic influence. If MZ and DZ twins are equally similar, environment must be the main factor.

For height, MZ twins are much more similar than DZ twins. This gives an h^2 estimate of about 0.80 for height from twin studies—pretty high, meaning genetics plays a strong role.

But there's a caveat. Twins, especially MZ twins, often share very similar environments. They grow up in the same household, eat the same food, and may even be dressed alike. This shared environment can make them appear more similar, slightly inflating heritability estimates. Modern studies try to account for this by comparing twins raised together versus twins raised apart, but such cases are rare.

21.3.3 SNP-Based Heritability

Modern genetics offers a more direct approach: estimate heritability from genetic data itself. Instead of inferring shared DNA from pedigrees, we can now sequence people's genomes and measure how genetically similar they are at millions of SNPs.

Here's how it works. Take a large sample of unrelated individuals—say, 50,000 people. Measure a trait like height. Then compare genetic similarity (based on SNPs) to trait similarity. If people who are slightly more genetically similar also tend to have more similar heights, that tells you genetics influences the trait.

One popular method is **LDSC (linkage disequilibrium score regression)**, which we encountered in the previous section. LDSC estimates heritability by looking at how strongly trait-associated SNPs correlate with nearby SNPs due to linkage disequilibrium.

For height, SNP-based heritability comes out to about $\mathbf{h^2 \approx 0.68}$. That's a bit lower than the twin-study estimate of 0.80, and there are a few reasons why. SNP-based heritability typically uses common SNPs (MAF $>$ 1%), so it misses some rare variants. Also, it doesn't capture structural variants or other genetic differences that SNP chips don't detect well.

Interestingly, when rare variants are included, the estimate goes up. A 2022 study using whole-genome sequencing found that rare variants in low linkage disequilibrium contribute about 0.31 of the phenotypic variance in height (Wainschtein et al. 2022, Nat Genet). Add that to the common SNP component, and you're getting closer to the twin-study estimate.

21.4 Why Heritability Changes Across Traits and Environments

Heritability is not a fixed property of a trait. It can differ across populations, time periods, or environments. Here's why.

Remember the formula:

$h^2 = VA / VP$

Heritability depends on both the numerator (genetic variance) and the denominator (total phenotypic variance). If environmental variance changes, the denominator changes, and so does heritability—even if genetic variance stays the same.

Let's look at a few examples:

Height has an h^2 of about 0.68 in well-nourished populations. Why not higher? Because even in genetically similar people, environmental factors like childhood nutrition, illness, and growth conditions create variation. But in populations with severe malnutrition, environmental variance would be much larger, and heritability would drop.

BMI (body mass index) has an h^2 of about 0.30—much lower than height. Why? Because BMI is strongly influenced by diet, exercise, and lifestyle. Genetic differences still matter, but environmental differences dominate. If you took the same population and put everyone on identical diets and exercise routines (which you can't do in real life, but imagine), environmental variance would shrink and heritability would appear to increase.

Blood pressure has an h^2 of about 0.25, also relatively low. Diet, stress, physical activity, and medication all have large effects. Genetic differences contribute, but environment is a major player.

Here's the key takeaway: **high heritability does not mean environment is unimportant**. It simply means genetic differences explain most of the remaining variation within that specific population and environment. Change the environment, and the picture can shift dramatically.

21.5 Fisher's Model in Modern Context

More than a century after Fisher's 1918 paper, his framework is still the backbone of modern genetics. Let's see how recent findings fit with his model.

Thousands of common SNPs contribute small additive effects. The 2022 GWAS on height found 12,111 SNPs spread across 7,209 genomic regions (Yengo et al. 2022, Nature). Almost all of these effects are additive, just as Fisher predicted. Each SNP nudges height up or down by a fraction of a centimeter, and together they explain most of the genetic variance.

Dominance effects are rare and weak. The 2023 UK Biobank study found dominance effects at 183 loci, but they explained less than 1% of the additive variance on average (Palmer et al. 2023, Science). Fisher was right to focus on additive variance—it's where most of the action is.

Epistasis exists but contributes little to population-level variation. A 2019 review emphasizes this point: gene-gene interactions happen, but they don't explain much of the trait variance we observe across populations (Visscher et al. 2019, Genetics). Additive effects dominate.

Fisher's model has proven remarkably robust. It underlies everything from **twin studies** to **GWAS** to **polygenic risk scores**—those scores that combine information from thousands of SNPs to predict traits or disease risks.

As noted in a 2019 review, Fisher's 1918 framework still guides how we study genetic architecture today (Visscher et al. 2019, Genetics). It explains **trait variation**, not individual destiny. And that's an important distinction. Knowing that height has high heritability doesn't tell you how tall you will be—it tells you why people in a population differ from each other.

21.6 Putting It All Together

Let's recap what we've learned about heritability:

Fisher unified Mendelian genetics with continuous trait variation back in 1918. He showed that many genes, each with small additive effects, could produce the smooth variation we see in traits like height. His insight gave birth to quantitative genetics.

Phenotypic variance can be partitioned into components: **Additive (VA)** + **Dominance (VD)** + **Epistatic (VE)** + **Environment (VEnvironment)**. Additive variance drives most heritability and evolutionary response. Dominance and epistasis exist, but they explain little variance at the population level.

Heritability (h^2) measures variation within a population, not genetic "determination" of individuals. An h^2 of 0.68 for height means 68% of the differences among people come from genetic differences, not that your height is 68% genetic. Your height is 100% the product of genes and environment working together.

Heritability depends on environmental conditions and can change over time. If nutrition improves and everyone gets adequate food, environmental variance shrinks, and heritability can appear to increase—even though genes haven't changed. Heritability is a snapshot of a specific population in a specific environment.

We estimate heritability through **covariance** between relatives. Twin studies compare identical and fraternal twins. SNP-based studies measure genetic similarity directly. Both approaches confirm that traits like height have substantial genetic influence, while traits like BMI and blood pressure show stronger environmental effects.

Fisher's model still underlies modern GWAS and polygenic prediction. More than 100 years later, his framework remains the foundation for understanding complex traits. The details have gotten more sophisticated—we can now identify specific SNPs and measure their effects—but the basic model is Fisher's.

22 Chapter 17. Partitioned Variance: Learn from Real-World Examples

We've spent the last few sections building up theory—Fisher's model, heritability, additive versus dominant effects. Now let's see how all of this actually works in real biomedical research. How do scientists use the concept of **partitioned variance** to answer messy, complicated questions about health and disease?

The idea is simple but powerful. When you measure a trait in a population, the variation you see comes from multiple sources: genes, environment, behavior, and their interactions. If you can statistically separate these sources—partition the variance—you can figure out what's really driving the trait. More importantly, you can avoid mistaking correlation for causation.

In this section, we'll look at two recent studies that beautifully illustrate this approach:

1. **Yap et al. (2021, Cell)** asked: Does the gut microbiome cause autism? They partitioned variance across biological layers—genome, diet, and microbiome—to find out.
2. **Ahlqvist et al. (2024, JAMA)** asked: Does taking Tylenol during pregnancy increase autism risk? They partitioned variance into familial (shared) and environmental (unique) components using siblings.

Both studies reveal something surprising: associations that seem real at first glance often disappear when you properly account for where the variance comes from. Let's dive in.

22.1 The Microbiome and Autism

You've probably heard claims about gut bacteria affecting everything from mood to brain development. One popular idea is that the gut microbiome might play a role in autism spectrum disorder (ASD). Some studies have found differences in gut bacteria between autistic and non-autistic children, and this has led to speculation—and even commercial products—claiming that changing the microbiome could help treat autism.

But here's the question: do those microbial differences actually cause autism, or are they just a reflection of something else?

Yap et al. (2021) set out to answer this question using a clever approach. They studied 247 children, including those with ASD, and measured three things: autism diagnosis, dietary patterns, and gut microbiome composition. Then they used a statistical method to estimate how much of the variation in each trait could be explained by differences in the microbiome.

They created a metric called the **microbiome-association index (b^2)**, which works just like heritability (h^2), but for microbes instead of genes. If the microbiome explains a lot of variance in a trait, b^2 is high. If it explains little, b^2 is low.

22.1.1 What They Found

Here's what the numbers showed:

The **microbiome explained almost none of the variance in ASD diagnosis**. The b^2 was somewhere between 0% and 9%—essentially zero. In other words, knowing someone's gut microbiome composition tells you almost nothing about whether they have autism.

But the microbiome did explain a large chunk of variance in **dietary patterns and stool consistency**. For these traits, b^2 was around 40–64%. That makes sense—what you eat directly affects which bacteria thrive in your gut and how your digestion works.

So what's going on? If the microbiome doesn't explain autism, why do some studies find differences in gut bacteria between autistic and non-autistic kids?

The answer is **indirect pathways**. Here's how it works:

Autism affects behavior, including eating behavior. Many autistic children have selective eating patterns—they prefer certain textures, flavors, or routines around food. These dietary preferences shape the gut microbiome. If you eat a limited variety of foods, your microbiome becomes less diverse. If you eat mostly processed foods, certain bacteria thrive while others don't.

So the chain of causation goes like this: – Autism → selective eating → altered diet → different microbiome composition

The microbiome isn't causing autism. It's reflecting behavioral traits associated with autism. The microbiome is like a mirror—it shows you something about the person, but it's not the original cause.

22.1.2 Why This Matters for Variance Partitioning

This study is a great example of how variance partitioning extends beyond just genes and environment. Yap and colleagues effectively divided up the variance into layers:

Genetic factors (A or D) affect neurodevelopment and behavior. Some of the genes that increase autism risk also influence sensory processing and eating preferences.

Behavioral factors mediate the connection between genes and other traits. Selective eating is a behavior, shaped by both genetics and experience.

Environmental factors, including the microbiome, respond to those behaviors. Your gut bacteria are part of your environment, but they're shaped by what you do—especially what you eat.

Gene–by–environment interactions (G × E) might also play a role. Maybe certain gut bacteria affect physiology differently depending on your genetic background. But even if those interactions exist, they don't mean the microbiome is causing autism.

By partitioning variance across these biological layers—host genetics, behavior, diet, and microbiome—Yap's team showed that the apparent link between microbiome and autism is mostly a story about behavior, not microbial causation.

22.2 Familial Confounding and Additive Factors—Acetaminophen and Autism

Let's look at a different kind of question. Does taking acetaminophen (Tylenol) during pregnancy increase the risk of autism or ADHD in children?

This question matters because acetaminophen is one of the most commonly used medications during pregnancy. It's generally considered safe for pain and fever. But over the past decade, some studies have suggested a link between prenatal acetaminophen use and neurodevelopmental disorders. Understandably, this has worried a lot of pregnant people.

Ahlqvist et al. (2024) tackled this question using a massive dataset: **2.48 million children born in Sweden**. They had access to detailed records of medication use during pregnancy and diagnoses of autism and ADHD in childhood.

22.2.1 What They Found at First

When the researchers ran a standard population–level analysis, they found a weak association. Children whose mothers used acetaminophen during pregnancy had a slightly higher risk of autism and ADHD—about 5–7% higher (hazard ratio □ 1.05–1.07).

At first glance, this looks concerning. But here's where variance partitioning comes in.

22.2.2 The Sibling Control Analysis

Ahlqvist and colleagues did something clever: they compared **siblings born to the same parents**. In some families, the mother took acetaminophen during one pregnancy but not another. By comparing those siblings, the researchers could control for all the factors that siblings share—genetic background, parental health, socioeconomic status, home environment, and so on.

When they ran this sibling control analysis, the association completely disappeared. The hazard ratio dropped to **1.0** for all outcomes—autism, ADHD, everything. No link at all.

22.2.3 What This Means

The initial association wasn't due to acetaminophen causing neurodevelopmental problems. It was due to **familial confounding**—shared factors within families that influence both medication use and child outcomes.

Think about it: which families are more likely to use acetaminophen during pregnancy? Maybe families where the mother has more health problems, or more stress, or different health–seeking behaviors. Those same factors might also affect the child's neurodevelopment, either through genetics or through the prenatal environment.

Here's the chain of hidden causation: – Shared family factors → both acetaminophen use and child neurodevelopment

The acetaminophen isn't causing autism. It's just correlated with autism because both are influenced by the same underlying family–level factors.

By comparing siblings, the researchers effectively controlled for **additive familial effects (A)** and shared environment. Siblings share about 50% of their DNA and grow up in the same household. When you compare siblings, you're holding constant all those shared factors. What's left is the **unique environmental effect (E)**—the things that differ between pregnancies for the same mother.

And that unique effect of acetaminophen? Zero.

22.2.4 Connecting This to Fisher's Model

This is a textbook example of variance partitioning in action. Remember Fisher's equation:

VP = A + D + E + Environment

In the population–level analysis, acetaminophen use was confounded with A (additive genetic factors) and shared environment. The association looked real because all these factors were tangled together.

In the sibling control analysis, the researchers effectively subtracted out A and shared environment. What remained was the unique environmental effect of acetaminophen during that specific pregnancy. And that effect was zero.

This demonstrates why **additive variance (A)** matters so much in real–world research. Additive effects explain why family members resemble each other, and they can create spurious associations in observational studies. If you don't account for shared familial factors, you can easily mistake correlation for causation.

22.3 Connecting the Two Examples

Let's step back and see what these two studies have in common.

Yap et al. focused on the microbiome and behavioral mediation. They asked: Does the gut microbiome explain autism? Their approach was to partition variance using b^2 (the microbiome analog of heritability). They found that the apparent ASD–microbiome link was really a story about diet and behavior. The microbiome reflects dietary patterns, which in turn reflect autism–related eating behaviors. Once you account for behavior, the direct microbiome effect on autism vanishes.

Ahlqvist et al. focused on prenatal drug exposure and familial confounding. They asked: Does acetaminophen cause autism? Their approach was a sibling control analysis, which partitions variance into shared family factors versus unique environmental effects. They found that the apparent Tylenol–ASD link was really a story about shared family characteristics. Once you compare siblings (controlling for additive familial factors), the acetaminophen effect vanishes.

Despite studying completely different questions, both studies followed the same logic: 1. Observe an association in the population. 2. Partition variance to separate different sources of variation. 3. Discover that the association is driven by confounding factors, not direct causation.

In the Yap study, the confounding factor was **behavior** (selective eating). In the Ahlqvist study, the confounding factor was **shared family background** (genes and environment). But the principle is the same: don't assume correlation means causation until you've partitioned the variance.

Here's a side–by–side comparison:

Yap et al. (Cell, 2021): They studied the microbiome and behavioral mediation. Their approach used variance partitioning with b^2 (microbiome analog of heritability). The key finding was that the ASD–microbiome link was explained by diet and behavior, not direct microbial effects. The main takeaway: environmental variation (microbiome) can reflect behavior and gene–by–environment interactions, not causation.

Ahlqvist et al. (JAMA, 2024): They studied prenatal drug exposure and familial confounding. Their approach used sibling control analysis (within–family adjustment). The key finding was that the Tylenol–ASD link was

explained by shared family factors, not the drug itself. The main takeaway: additive familial factors must be accounted for to reveal true environmental effects.

Both studies show how **partitioning variance**—into additive, dominance, environmental, and behavioral components—helps answer complex biological questions. By properly accounting for shared influences and environmental interactions, researchers can distinguish genuine causal effects from correlations shaped by family background or behavior.

22.4 What We Learn from These Examples

Let's pull out the lessons that apply beyond these specific studies.

Variance partitioning isn't limited to genetics. Fisher's framework started with genes, but the logic applies to any biological or environmental variable. Yap's study shows you can partition variance between microbiome, diet, and behavior. The same approach could work for hormones, metabolites, or brain activity patterns. The key is to separate correlated factors and figure out which ones actually drive the trait.

Apparent associations often disappear when you control for confounding factors. Both studies found initial associations that looked real—microbiome differences in autistic kids, higher autism risk with Tylenol use. But once they partitioned variance properly, the associations vanished. This is a reminder to be skeptical of observational correlations, especially in complex traits.

Distinguishing between A (additive) and E (environmental) variance is essential for causal inference. If you don't account for familial factors (A), you'll mistake shared family characteristics for environmental causes. The sibling control design is one powerful way to separate these components. Twin studies and polygenic scores are others.

Understanding how variance is distributed across biological and social layers prevents misinterpretation. Autism research is full of claims about potential causes—vaccines, diet, pollution, microbiome, you name it. Many of these claims come from studies that found correlations but didn't partition variance properly. When you do partition variance, most of these associations either shrink dramatically or disappear.

Fisher's framework remains a universal tool. Whether you're studying genome, microbiome, or family-level effects, the logic is the same: VP = A + D + E + Environment. Partition the variance into components, model each one separately, and see which ones actually matter. More than 100 years after Fisher introduced this approach, it's still the gold standard for analyzing complex traits.

These real-world examples show why we spent so much time learning theory. The math and concepts aren't just abstract—they're tools for answering the questions that matter. Can we prevent autism by changing gut bacteria? Should pregnant people avoid Tylenol? These are high-stakes questions, and getting the answers right requires careful partitioning of variance.

The next time you see a headline claiming "X causes Y," ask yourself: Did they account for shared familial factors? Did they consider behavioral pathways? Did they partition the variance? If not, take the claim with a grain of sal

23 Chapter 18. Introduction to Genome-Wide Association Studies

We've been building up the theory of complex traits for several sections now. We know that most traits are polygenic—shaped by thousands of genetic variants, each with a small additive effect. We know how to partition variance into genetic and environmental components. But here's the practical question: how do we actually find those thousands of variants?

That's where **Genome-Wide Association Studies (GWAS)** come in. GWAS is the workhorse method for discovering genetic variants associated with complex traits. Whether you're studying height, heart disease, autism, or anything else, GWAS is usually the first step. It's like a genome-wide search, scanning millions of positions in the DNA to find which ones correlate with the trait you care about.

In this chapter, we'll learn how GWAS works, how to read its results using **Manhattan plots** and **QQ plots**, and how to translate statistical associations into biological understanding through **pathway and cell-type analyses**. We'll also explore **polygenic scores**—a way to combine all those small genetic effects into a single number that predicts an individual's risk. Finally, we'll discuss the ethical challenges that come with using genetic information in medicine.

Let's dive in.

23.1 What is GWAS?

GWAS stands for Genome-Wide Association Study. At its core, GWAS asks a simple question: are certain genetic variants more common in people with a trait than in people without it?

The variants we're usually talking about are **single nucleotide polymorphisms (SNPs)**—single-letter changes in the DNA sequence. Remember, these are common variants that appear in at least 1% of the population. They're not rare mutations that cause Mendelian disorders. They're the everyday genetic differences that make people different from each other.

Here's how GWAS works in practice. You collect two groups of people: a **case group** (people who have the trait or disease) and a **control group** (people who don't). Then you genotype millions of SNPs across their genomes and compare allele frequencies between the two groups. If a particular SNP allele is more common in cases than controls, that suggests the SNP is associated with the trait.

Before GWAS became possible in the mid-2000s, genetic research mostly focused on rare mutations in families with Mendelian disorders—things like cystic fibrosis or sickle cell disease, where one mutation causes the disease. GWAS changed the game by showing that common variants with small additive effects also matter. It confirmed Fisher's 1918 polygenic model on a scale that would have been unimaginable in his time.

23.2 How GWAS Works

A typical GWAS tests **hundreds of thousands to millions of SNPs**. Modern genotyping arrays can measure around 500,000 to 1 million SNPs directly, and statistical methods called imputation can infer several million more based on patterns of linkage disequilibrium.

For each SNP, a statistical test is run to see whether it's associated with the trait. The output is a **p-value**—a number between 0 and 1 that tells you how surprising the result would be if the SNP had no real effect.

Let's unpack what that means, because p-values confuse a lot of people.

23.2.1 Understanding the Null Hypothesis

In statistics, we start by assuming there's no effect—this is called the **null hypothesis (H_0)**. For a GWAS, the null hypothesis says: "This SNP has no true effect on the trait. Any difference between cases and controls is just random noise."

The **alternative hypothesis (H_1)** is the opposite: "This SNP is associated with the trait. The difference we see is real, not just chance."

Now here's the key: we can't prove H_1 is true. All we can do is calculate how unlikely the data would be if H_0 were true. That's what the p-value tells us.

Think of it like a courtroom. The defendant (the SNP) is assumed innocent (no effect) until proven guilty (associated with the trait). The p-value is like the strength of the evidence. If the p-value is very small—say, 0.00001—that means the evidence against innocence is strong. If the SNP really had no effect, you'd almost never see data this extreme. So we conclude the SNP is probably guilty—I mean, associated.

But here's the catch: "rejecting H_0" does **not** prove causation. It just means the observed pattern is unlikely to have arisen by chance alone. The SNP might be associated because it directly affects the trait, or because it's linked to another variant that does, or because of some confounding factor.

23.2.2 The Multiple Testing Problem

Now imagine you're doing a GWAS that tests 1 million SNPs. If you use the standard significance threshold of $p < 0.05$ (which is common in many sciences), you'd expect about **50,000 false positives** just by chance. That's a disaster—you'd have no idea which findings are real.

To prevent this, GWAS uses a much more stringent threshold:

$p < 5 \times 10$

(that's 0.00000005, or 5 in 100 million). This is roughly equivalent to a Bonferroni correction for one million independent tests. It's conservative, but it keeps the false positive rate low.

So when you see a GWAS result with $p < 5 \times 10^{-8}$, you can be fairly confident it's a real association, not just noise.

23.3 Visualizing GWAS Results: Manhattan and QQ Plots

GWAS generates millions of p-values—one for each SNP. How do you make sense of all that data? Two types of plots are used: Manhattan plots and QQ plots.

23.3.1 Manhattan Plot

A **Manhattan plot** is the iconic visualization of GWAS results. It's called a Manhattan plot because the peaks look like the skyscrapers of the Manhattan skyline. Each dot represents a SNP, arranged along the x-axis by chromosomal position (chromosome 1, then 2, then 3, and so on). The y-axis shows the statistical significance as $-\log_{10}(\text{p-value})$.

Why $-\log_{10}$? Because p-values are tiny numbers like 0.000001, and it's easier to work with them on a log scale. A p-value of 0.05 becomes $-\log_{10}(0.05) \approx 1.3$. A genome-wide significant p-value of 5×10^{-8} becomes $-\log_{10}(5 \times 10^{-8}) \approx 7.3$. Bigger numbers on the y-axis mean more significant associations.

The horizontal dashed line at $y \approx 7.3$ marks the genome-wide significance threshold. SNPs above this line are worth paying attention to. They're the "skyscrapers" that stand out from the background noise.

Let's look at a real example. In a large GWAS of dilated cardiomyopathy (DCM)—a heart condition where the heart muscle becomes weak and enlarged—**14,256 cases** and **1,199,156 controls** were analyzed. The study identified **80 genomic loci** associated with DCM. Many of these loci clustered near genes like **MAP3K7**, **NEDD4L**, and **SSPN**, which are involved in heart muscle function (Zheng et al. 2024, Nat Genet).

Figure: Manhattan plot of DCM GWAS identifying novel and previously reported genomic loci. Each dot represents a SNP. The y-axis shows $-\log_{10}$(p-value)—higher dots indicate stronger associations. Red peaks mark newly discovered loci, while orange peaks show previously known loci. The significant signals cluster at specific chromosomal regions. These clusters often contain genes involved in the same biological pathways, like muscle structure or cell adhesion. This visualization immediately

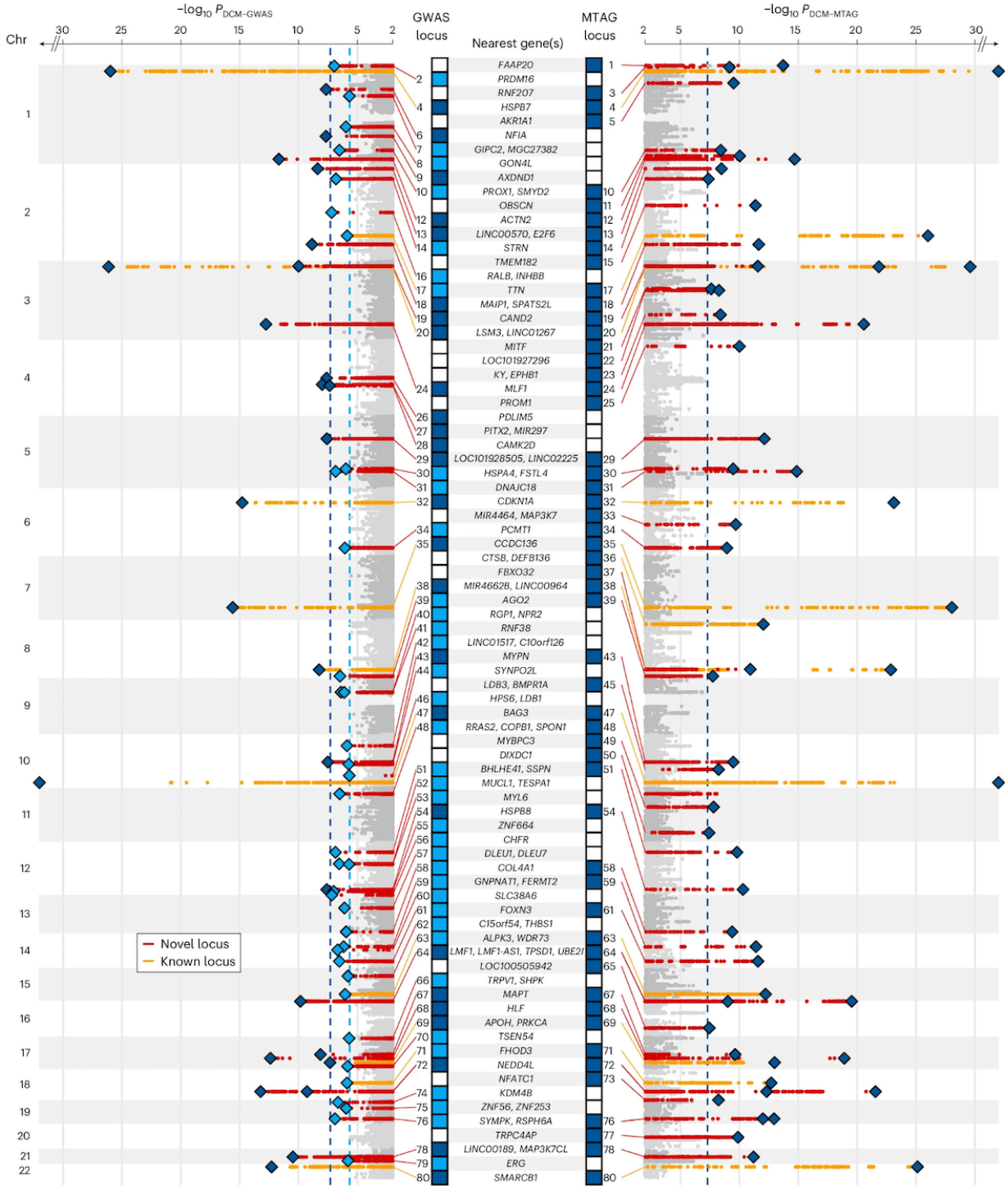


Figure 46: Zheng2024–NatGenet–Fig2

shows where in the genome the strongest disease associations are located. Source: Zheng, S.L. et al. (2024). Multi-ancestry genome-wide association study of dilated cardiomyopathy. Nature Genetics. https://www.nature.com/articles/s41588-024-01952-y. License: Creative Commons.

23.3.2 Quantile-Quantile (QQ) Plot

The **QQ plot** is a quality-control tool. It checks whether the distribution of observed p-values matches what you'd expect under the null hypothesis.

Here's how it works. If none of the SNPs had real effects, all the p-values would follow a uniform distribution—equal numbers of p-values at every level from 0 to 1. On a QQ plot, you plot the **expected $-\log_{10}(p)$** (what you'd see if all SNPs were null) on the x-axis and the **observed $-\log_{10}(p)$** (what you actually see) on the y-axis.

If everything is null, the points fall on a diagonal line (y = x). If there are true associations, you'll see upward deviation at the right tail—the most significant SNPs are more significant than you'd expect by chance. That's good. It means you found real signals.

But if the entire distribution is shifted upward—not just the tail—that's a problem. It suggests **confounding**, often from something called **population stratification**. This happens when your cases and controls come from different ancestral backgrounds, and the genetic differences between populations create false associations. Modern GWAS methods correct for this using principal components or linear mixed models.

In the DCM study, the QQ plot looked good. Most points fell on the diagonal, with upward deviation only at the tail where the 80 significant loci lived. This confirms the associations are real, not artifacts of confounding.

Together, the Manhattan plot tells you where the associations are, and the QQ plot tells you whether they're trustworthy.

23.4 Biological Interpretation: Pathways, Cell Types, and Functional Convergence

Finding 80 associated loci is exciting, but what does it mean biologically? How do all these scattered SNPs across the genome contribute to heart disease?

This is where **pathway enrichment analysis** and **cell-type analysis** come in. The DCM study prioritized **62 candidate genes** near the 80 DCM-associated loci and asked: Do these genes share common functions?

The answer was yes. The 62 genes were enriched in pathways related to: – **Muscle structure** (sarcomeres, the contractile units of muscle cells) – **Cell adhesion** (how heart muscle cells stick together) – **Extracellular matrix (ECM) organization** (the scaffolding that holds heart tissue together)

All of these are core components of heart muscle integrity. If the ECM is weak, or if cells can't adhere properly, the heart can't contract efficiently—leading to DCM.

Single-nucleus RNA sequencing was also used to see which cell types express these genes. The analysis found that several risk genes are highly expressed in specific **cardiomyocyte states** (the muscle cells of the heart) and **fibroblasts** (cells that produce ECM). This cell-type specificity tells you where in the heart the genetic risk plays out.

This is an example of **functional convergence**. Even though the 80 loci are scattered across the genome, they participate in shared biological pathways. It's like having 80 different entry points into the same building—they all lead to the same place.

Figure: Enrichment of prioritized DCM genes across pathways and cell types. This figure shows which biological pathways and cell types are enriched for DCM candidate genes. Pathways related to cardiac structure and ECM dominate, and specific cardiomyocyte and fibroblast states light up. Despite the

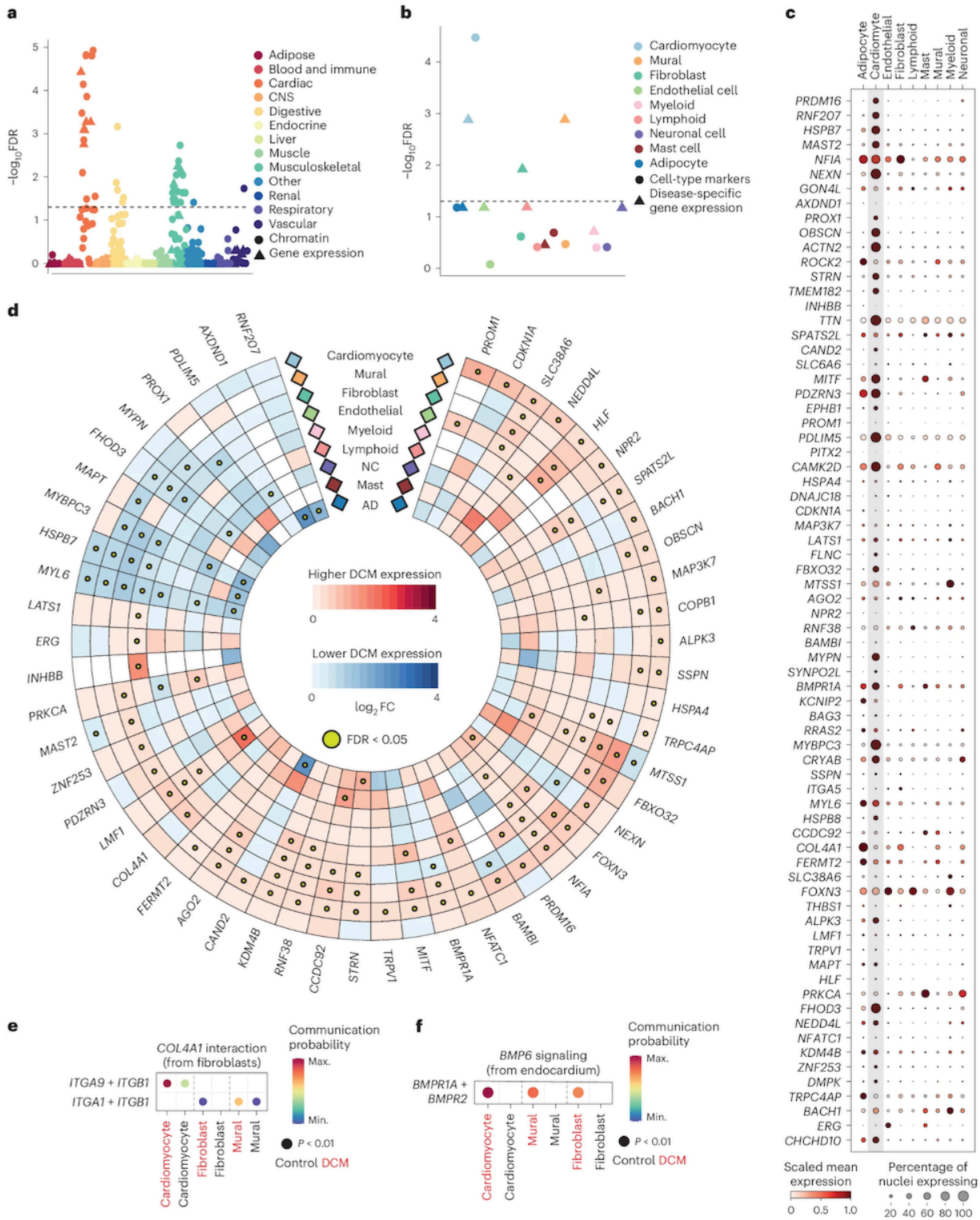


Figure 47: Zheng2024–NatGenet–Fig5

complexity—many loci, many genes—the biological story is coherent: these genes converge on mechanisms that maintain heart muscle structure. That's how a polygenic trait produces a unified disease phenotype. Source: Zheng, S.L. et al. (2024). Multi-ancestry genome-wide association study of dilated cardiomyopathy. Nature Genetics. https://www.nature.com/articles/s41588-024-01952-y. License: Creative Commons.

23.5 Polygenic Scores (PGS)

We've found 80 loci associated with DCM. Each locus contributes a tiny bit to disease risk. How do we combine all that information into something useful for an individual person?

That's where **polygenic scores (PGS)** come in. A polygenic score—also called a polygenic risk score or genomic score—quantifies your cumulative genetic risk by summing up the effects of all your risk alleles.

Here's how it works. For each of the 80 DCM-associated SNPs, the GWAS tells you the effect size—how much that SNP increases risk. You check which alleles the person carries at those SNPs, multiply by the effect sizes, and add them all up. The result is a single number representing that person's genetic liability for DCM.

Polygenic scores are especially valuable for complex traits, where no single mutation determines the outcome. Instead of looking at one gene, you're integrating information from thousands of variants across the genome.

23.5.1 Example: Dilated Cardiomyopathy (DCM)

In the DCM study, a **PGS for DCM** was calculated using the 80 significant loci. Then the score was tested in the UK Biobank—a huge dataset of about 500,000 people with genetic and health records.

Here's what was found:

Individuals in the **top 10%** of PGS had a **2.8-fold higher risk** of developing DCM compared to those in the lowest 10%. That's nearly three times the risk, just from common genetic variants.

This elevated risk held **even in people without rare pathogenic mutations**. In other words, your polygenic background matters even if you don't carry a known disease-causing mutation.

But here's the really interesting part: among people who do carry pathogenic variants—like mutations in **TTN** or **LMNA**, which are known causes of cardiomyopathy—those with high PGS were **significantly more likely** to develop DCM than carriers with low PGS.

This shows that polygenic background modifies penetrance. Having a rare mutation doesn't guarantee you'll get the disease. Your polygenic score acts as a modifier, tipping you toward or away from clinical disease.

Figure: DCM PGS is associated with DCM disease status in the UK Biobank. The prevalence of DCM increases steadily as you move from the lowest PGS decile (left) to the highest (right). People in the top 10% have about 2.8 times the risk of those in the bottom 10%. The dark bars represent carriers of rare pathogenic variants, while lighter bars represent non-carriers. Notice that even among pathogenic variant carriers, higher PGS predicts higher disease prevalence. This demonstrates that polygenic background affects penetrance—even in what we used to think of as "monogenic" diseases. Source: Zheng, S.L. et al. (2024). Multi-ancestry genome-wide association study of dilated cardiomyopathy. Nature Genetics. https://www.nature.com/articles/s41588-024-01952-y. License: Creative Commons.

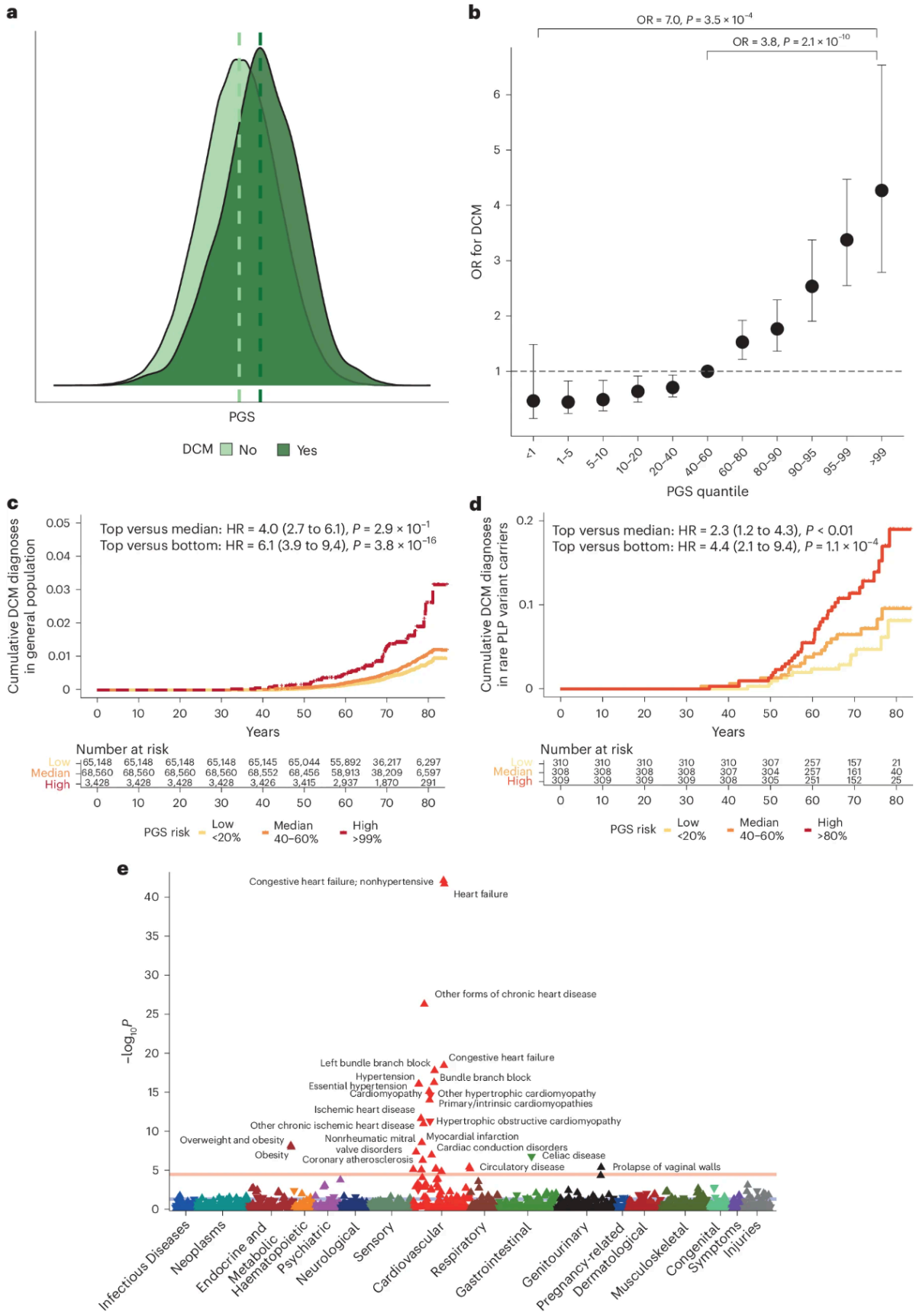


Figure 48: Zheng2024–NatGenet–Fig6

23.6 Application of Polygenic Scores

Polygenic scores are moving from research tools to clinical applications. Here are some of the ways they're being used:

Risk prediction and early detection. If you have a high PGS for heart disease, you might benefit from earlier cardiac screening—echocardiograms, stress tests, and so on. You can also take preventive measures like improving diet, exercising more, or monitoring blood pressure closely. The goal is to catch problems before they become serious.

Precision medicine. Combining PGS with rare variant data gives a more complete picture of genetic risk. If you carry a pathogenic mutation and have a high PGS, your doctor knows you're at especially high risk and can adjust treatment accordingly. Genetic counselors can use this information to give more accurate risk estimates.

Population research. PGS lets studies investigate how genetic architecture varies across populations. For example, does the genetic basis of height differ between Europeans and East Asians? How do genetic risk factors interact with environmental exposures like diet or pollution? PGS provides a quantitative measure of genetic liability that can be compared across groups.

Drug discovery. Genes contributing to polygenic risk often point to biological pathways that could be targeted with drugs. If many DCM risk genes are involved in ECM organization, maybe drugs that strengthen the ECM could help. GWAS and PGS can guide where to look for new therapies.

These applications show how GWAS findings extend beyond discovery. They're becoming tools for actionable insights in clinical and translational genomics.

23.7 Ethical Considerations in the Use of Polygenic Scores

Polygenic scores hold great promise, but they also raise serious ethical questions. We need to think carefully about how we use this technology.

Equity and ancestry bias. Most GWAS datasets are heavily European. As of 2024, more than 80% of GWAS participants have European ancestry. This means polygenic scores trained on European data often don't work well in other populations. If you apply a European-derived PGS to someone of African or East Asian ancestry, the predictions can be wildly inaccurate. This creates health disparities—people from underrepresented populations don't benefit equally from genomic medicine. The solution is to include more diverse participants in GWAS, but that's still a work in progress.

Privacy and data protection. Genetic data are uniquely identifying. Your genome is like a fingerprint, but even more personal—it reveals information about your relatives too. If genetic data leak or are misused, the consequences can be severe. Strict safeguards are needed to protect against breaches, especially as genomic databases grow.

Interpretation and communication. A high polygenic score means increased probability of disease, not certainty. If your PGS puts you in the top 10% for heart disease risk, that doesn't mean you'll definitely get heart disease. It means your baseline risk is higher than average. Clinicians need to communicate this distinction clearly. Patients need to understand that genetics is one factor among many—lifestyle, environment, and luck all matter too.

Psychological and social impacts. Learning you have high genetic risk can be distressing. It might affect your mental health, self-perception, or life decisions. Would you want to know if you're at high risk for Alzheimer's disease, knowing there's currently no cure? Different people answer that question differently. There's also a risk of genetic stigma—being treated differently because of your genes, whether by insurers, employers, or even family members.

Regulatory oversight. As polygenic scores move into clinical use, we need clear policies governing how they're used. Should insurance companies be allowed to use PGS in risk assessment? Should employers

have access to genetic data? Most countries have laws against genetic discrimination, but enforcement and scope vary. We need transparent, fair policies that balance the benefits of genomic medicine with protection of individual rights.

These considerations remind us that science doesn't exist in a vacuum. Polygenic scores are powerful tools, but they must be applied responsibly. The technology is advancing faster than our ability to handle its ethical implications, so ongoing dialogue between scientists, clinicians, policymakers, and the public is essential.

23.8 Putting It All Together

Let's recap what we've learned about GWAS and polygenic scores.

GWAS identifies common genetic variants associated with traits by testing millions of SNPs across the genome. It compares allele frequencies between cases and controls and uses stringent statistical thresholds to avoid false positives. The result is a list of genomic loci that contribute to the trait.

We visualize GWAS results using Manhattan plots (showing where the associations are) and QQ plots (validating that they're real). These plots help us make sense of millions of data points and catch potential confounding.

Pathway and cell-type analyses reveal functional convergence—showing that dispersed loci often act through shared biological mechanisms. For DCM, 80 scattered loci converged on pathways related to muscle structure, cell adhesion, and ECM. This tells us how genetic variation leads to disease.

Polygenic scores quantify additive genetic risk by combining the effects of many variants into a single number. They enable practical applications in risk prediction, precision medicine, and drug discovery. The DCM example showed that PGS can identify high-risk individuals and modify penetrance even in carriers of rare pathogenic mutations.

But with great power comes great responsibility. Ethical challenges—ancestry bias, privacy, interpretation, psychological impacts, and regulatory oversight—must be addressed as genomics moves from research to real-world implementation. We need to ensure that genomic medicine is not only scientifically valid but also socially fair.

GWAS has transformed our understanding of complex traits. It confirmed Fisher's century-old polygenic model and revealed the genetic architecture of thousands of traits and diseases. As datasets grow and methods improve, GWAS will continue to drive discoveries—but we must remain vigilant about how we apply those discoveries in society.

24 Chapter 19. Genetic Architecture of Traits and Disorders

We've learned that GWAS can identify hundreds or even thousands of genetic variants associated with a trait. But here's a question that might be puzzling you: if we've found all these variants, why can't we fully predict who will get a disease? Why do some people with "high-risk" genes stay healthy, while others with seemingly lower genetic risk develop the condition?

The answer lies in something called **genetic architecture**—the full structure of genetic influences on a trait or disorder. It's not just about counting up risk alleles. It's about understanding how different types of variants—rare and common, large-effect and small-effect—work together. It's about recognizing that genes act in specific tissues, at specific times, and in specific contexts. And it's about appreciating that the same disease can arise through many different genetic routes.

In this chapter, we'll explore the genetic architecture of three complex disorders: dilated cardiomyopathy (DCM), Alzheimer's disease, and autism spectrum disorder (ASD). Each one tells us something different about how genetics shapes disease. By the end, you'll understand why different people can develop the

same disorder through completely different genetic mechanisms—and why those diverse mechanisms often converge on the same biological pathways.

24.1 What is Genetic Architecture?

Think of genetic architecture like the blueprint of a city. A city isn't just buildings—it's roads, utilities, parks, neighborhoods, all interconnected in complex ways. Some structures are huge and obvious (like a downtown skyscraper). Others are small but numerous (like individual houses). And the way the city functions depends not just on the parts, but on how they interact.

Genetic architecture is the same idea, but for traits and diseases. It encompasses all the genetic contributors to a phenotype, including:

Variant types: Not all genetic changes are the same. You might have single nucleotide variants (SNVs—single-letter changes), copy-number variants (CNVs—duplications or deletions of DNA segments), indels (small insertions or deletions), or larger structural variants. Each type can affect genes in different ways.

Effect sizes: Some variants have large, dominant effects—like the FGFR3 mutation causing achondroplasia that we discussed earlier. Others have tiny additive effects that only matter when combined with hundreds of other variants. The genetic architecture includes both ends of this spectrum.

Allele frequency: Are we talking about rare variants that appear in less than 0.1% of people, or common variants that show up in 10% or more? Rare variants tend to have bigger effects (otherwise natural selection would weed them out). Common variants usually have smaller effects but are easier to study because more people carry them.

Interactions: Genes don't act in isolation. Gene-gene interactions (epistasis) happen when the effect of one variant depends on which alleles you have at another gene. Gene-environment interactions (G×E) occur when genetic risk is modified by environmental factors like diet, stress, or toxins.

Context dependencies: When does the gene act? During fetal development? In childhood? Only in old age? Which tissues matter? Does it work differently in males versus females? All of this context shapes how genetic variants manifest.

Modern genomics—whole-genome sequencing (WGS), large GWAS cohorts with hundreds of thousands of people—reveals that most traits and diseases follow a **composite architecture**. That means they're shaped by a mixture of rare and common variants, with modifiers and interactions layered on top. There's no single "disease gene" that explains everything. Instead, you have a complex landscape of genetic influences.

Let's look at three examples to see how this plays out.

24.2 Example 1: Dilated Cardiomyopathy (DCM)

Dilated cardiomyopathy (DCM) is a heart condition where the left ventricle—the main pumping chamber—becomes enlarged and weakened. The heart stretches out, losing its ability to contract efficiently. Over time, this leads to heart failure.

For decades, doctors knew DCM could run in families, suggesting a genetic component. But the genetic architecture turned out to be more complex than anyone expected.

24.2.1 Rare High-Impact Variants

About **10–25% of DCM patients** carry pathogenic mutations in key heart genes like TTN, LMNA, DSP, or FLNC (Zheng et al. 2024, Nat Genet; Hershberger et al. 2021, Circulation). These are rare variants—

each mutation might appear in only one family—but they have big effects. If you carry a pathogenic TTN mutation (TTN encodes titin, a giant protein that forms the scaffold of heart muscle), your risk of developing DCM is high.

But here's the thing: even among people with these pathogenic mutations, not everyone develops DCM. Some carriers stay healthy their whole lives. That tells us the rare mutation isn't the whole story.

24.2.2 Polygenic Risk

A massive GWAS compared over **14,000 DCM cases** with more than **a million controls** and found **80 genomic loci** associated with DCM (Zheng et al. 2024, Nat Genet). These are common variants with small effects—the kind we've been talking about throughout this chapter.

Individually, each of these 80 loci barely nudges your risk. But collectively, they add up. When a polygenic score (PGS) for DCM was calculated, people in the **top 10% of PGS had about 2.8 times the risk** of those in the bottom 10%—even without any rare pathogenic mutations.

24.2.3 The Plot Thickens: Modifiers and Multi-Hit Models

Here's where it gets really interesting. Among people who carry rare pathogenic variants, those with high PGS were much more likely to develop DCM than carriers with low PGS.

In other words, your polygenic background modifies the penetrance of rare mutations. The rare mutation doesn't act alone—it's influenced by the cumulative effect of common variants across the genome.

Some DCM patients also carry **multiple rare variants** in different cardiomyopathy-relevant genes (Hershberger et al. 2021, Circulation). It's not one big hit; it's several smaller hits that accumulate. Each variant on its own might not be enough to cause disease, but together they cross a threshold.

24.2.4 Limited Yield of Clinical Panels

When the evidence for about 51 candidate genes often tested in clinical genetic panels was curated, only about **19 genes had strong or moderate evidence** for causing monogenic DCM (Hershberger et al. 2021, Circulation). This means that even our best-known "DCM genes" explain only a fraction of cases.

The rest? A mix of polygenic risk, rare variants in genes we don't yet fully understand, and environmental factors like viral infections or alcohol use.

24.2.5 Functional Convergence

Despite all this genetic diversity, there's a unifying theme. The genes identified in both the rare-variant studies and the GWAS converge on specific biological pathways: **muscle structure** (sarcomeres and cytoskeleton), **cell adhesion** (how heart cells stick together), and **extracellular matrix (ECM)** (the scaffolding that holds the heart together).

Many of these genes are highly expressed in **cardiomyocytes** (heart muscle cells) and **fibroblasts** (cells that produce the ECM). So even though patients might have different genetic variants, those variants often disrupt the same core mechanisms.

It's like many roads leading to the same destination. Different starting points, but they all converge on heart muscle dysfunction.

24.3 Example 2: Alzheimer's Disease

Alzheimer's disease (AD) is the most common cause of dementia, characterized by progressive memory loss, cognitive decline, and brain pathology including amyloid plaques and neurofibrillary tangles. Its genetic architecture is similarly complex.

24.3.1 Rare Mutations in Early-Onset Familial AD

A small fraction of Alzheimer's cases—less than 1%—are caused by **rare, highly penetrant mutations** in three genes: APP, PSEN1, and PSEN2. These mutations cause early-onset familial AD, where symptoms appear before age 65, sometimes as early as the 30s or 40s. If you carry one of these mutations, your risk is nearly 100%. But these cases are exceptional.

24.3.2 The APOE ε4 Allele

The biggest genetic risk factor for late-onset AD—the common form that affects people in their 60s, 70s, and beyond—is the **APOE ε4 allele**. APOE comes in three major versions: ε2, ε3, and ε4. The ε4 allele increases AD risk by about **3-fold if you have one copy** and **12-fold if you have two copies**.

APOE ε4 is common—about 15% of people of European ancestry carry at least one copy. Unlike the rare APP or PSEN mutations, ε4 is not a guarantee of disease. Plenty of ε4 carriers never develop Alzheimer's, and plenty of non-carriers do. It's a strong risk factor, not a deterministic cause.

24.3.3 Polygenic Background

Large GWAS meta-analyses have identified many additional loci with smaller effects. These variants implicate pathways involved in **immune function** (especially microglia, the brain's immune cells), **lipid metabolism** (how the brain handles cholesterol and other fats), and **synaptic function** (how neurons communicate) (Bellenguez et al. 2022, Nat Genet; Logue et al. 2023, Nat Genet).

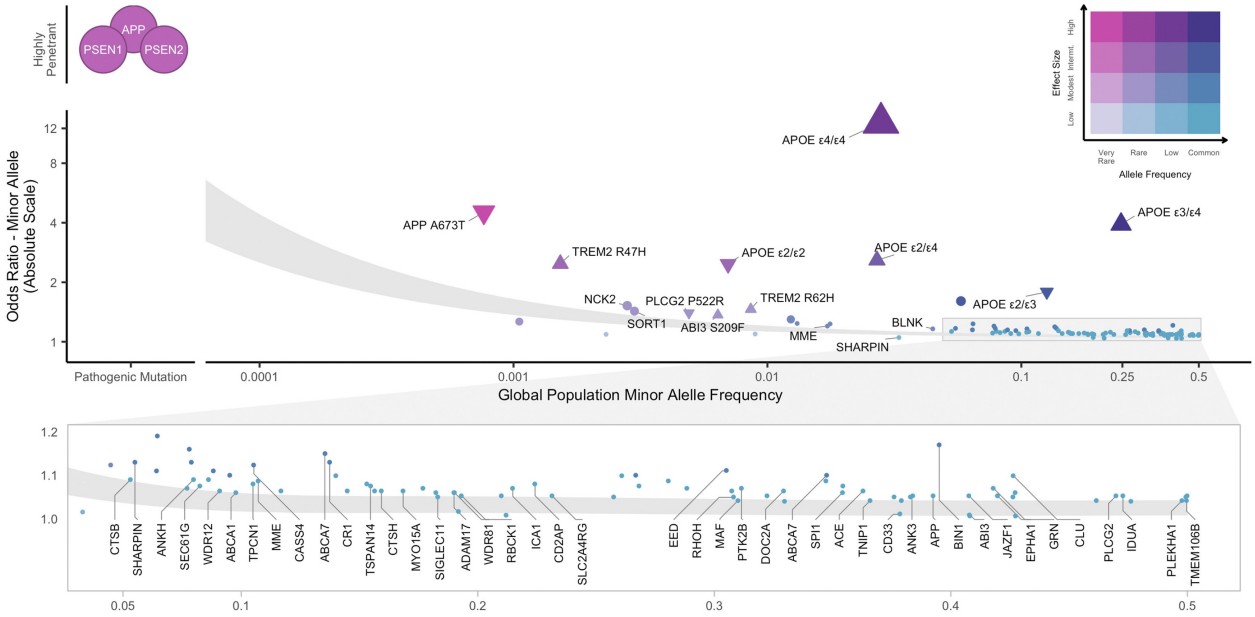


Figure 49: Andrews2023-EBioMed-Fig1

Figure: Alzheimer's genetic landscape across the genome. This Manhattan plot shows GWAS results for Alzheimer's disease and dementia. Each dot represents a SNP, and tall peaks indicate strong associations. Notice the peaks are scattered across many chromosomes—this is genetic heterogeneity

in action. No single gene dominates. Instead, risk is distributed across the genome, with multiple loci contributing small effects that collectively shape disease susceptibility. Source: Andrews, S.J. et al. (2023). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. EBioMedicine. https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00077-0/fulltext. License: CC-BY.

24.3.4 Converging Pathways: Microglial Efferocytosis

Despite the genetic complexity, many of the implicated genes converge on a specific biological process: **microglial efferocytosis**. This is the process by which microglia—the brain's resident immune cells—recognize and clear away dead cells and debris.

Think of microglia as the brain's garbage collectors. They patrol the brain, looking for damaged neurons, protein aggregates (like amyloid), and other cellular trash. When they find it, they engulf it, digest it, and recycle the components. This process is called efferocytosis (literally "eating of dead cells").

Here's the problem: many AD risk genes—both rare and common—affect this process. TREM2, ABCA1, APOE, LPL, and others are all involved in different stages of efferocytosis: 1. Detecting "eat-me" signals on damaged cells (like phosphatidylserine) 2. Engulfing debris into phagosomes 3. Digesting material through lysosomal fusion 4. Processing metabolic byproducts like cholesterol

If this clearance system is impaired, amyloid and other toxic proteins accumulate, driving neurodegeneration.

Figure: Prioritized AD risk genes converge on microglial efferocytosis. Many Alzheimer's risk genes—both rare and common—converge on microglial efferocytosis, the process by which microglia clear dead cells and debris. Different genes affect different stages: recognizing targets (TREM2), engulfing debris, digesting it (lysosomes), and recycling components (cholesterol metabolism via ABCA1, APOE, LPL). This functional convergence suggests that impaired microglial clearance is a key hub in Alzheimer's pathology, linking immune dysfunction, lipid metabolism, and neurodegeneration. Source: Andrews, S.J. et al. (2023). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. EBioMedicine. https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00077-0/fulltext. License: CC-BY.

24.3.5 Composite Model of Risk

Late-onset Alzheimer's is influenced by a composite model: rare variants (like APOE $\varepsilon 4$ homozygosity acts almost like a rare variant in terms of effect size), common polygenic risk, and environmental/vascular factors (hypertension, diabetes, cardiovascular health).

A recent study from the K-ROAD consortium showed that individuals with **multiple genetic risk factors**—for example, APOE $\varepsilon 4$ plus high polygenic score plus rare noncoding variants or structural variants—have the highest amyloid burden and the worst cognitive outcomes. Risk accumulates (K-ROAD consortium 2025, Nat Commun).

Figure: Genetic risk accumulates in Alzheimer's disease. This violin plot shows how the number of genetic risk factors affects Alzheimer's disease outcomes. Individuals are grouped by their risk burden: non-carriers (no risk factors), single-factor carriers (rare noncoding variants only, structural variants only, high polygenic risk score only, or APOE $\varepsilon 4$ only), and those with 2 or 3 risk factors combined. As the number of risk factors increases from left to right, several patterns emerge: amyloid-β (Aβ) burden increases, cognitive function declines (lower K-MMSE scores), dementia severity worsens (higher CDR-SB scores), and both verbal and visual memory deteriorate. All comparisons show statistically significant trends (p-values displayed above each panel). This demonstrates that genetic risk is cumulative—having multiple risk factors (whether common variants, rare variants, structural variants, or APOE $\varepsilon 4$) compounds disease severity more than any single factor alone. This is composite genetic architecture in action. Source: K-ROAD consortium (2025). Integrated analysis of common and rare genetic variants identifies multiple loci associ-

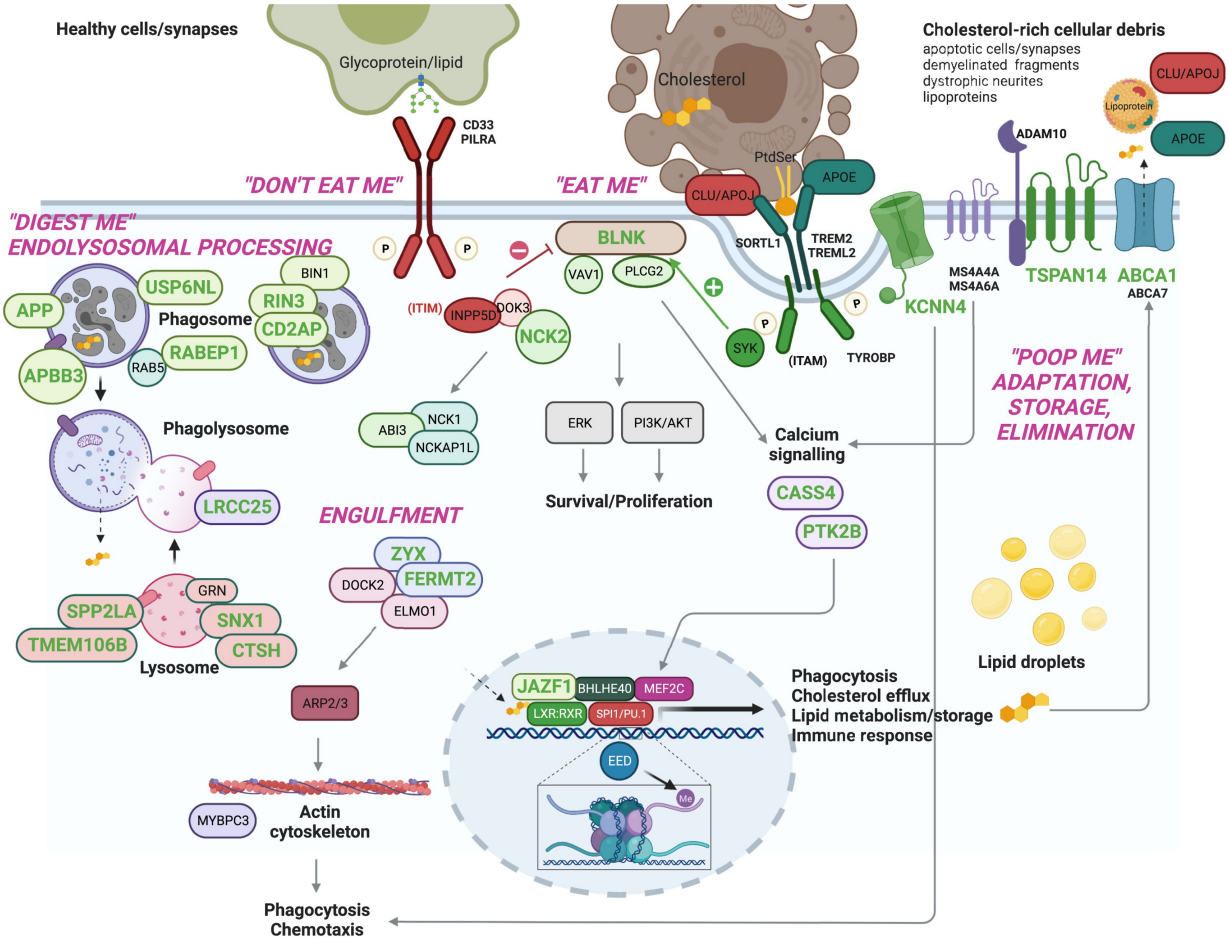


Figure 50: Andrews2023-EBioMed-Fig3

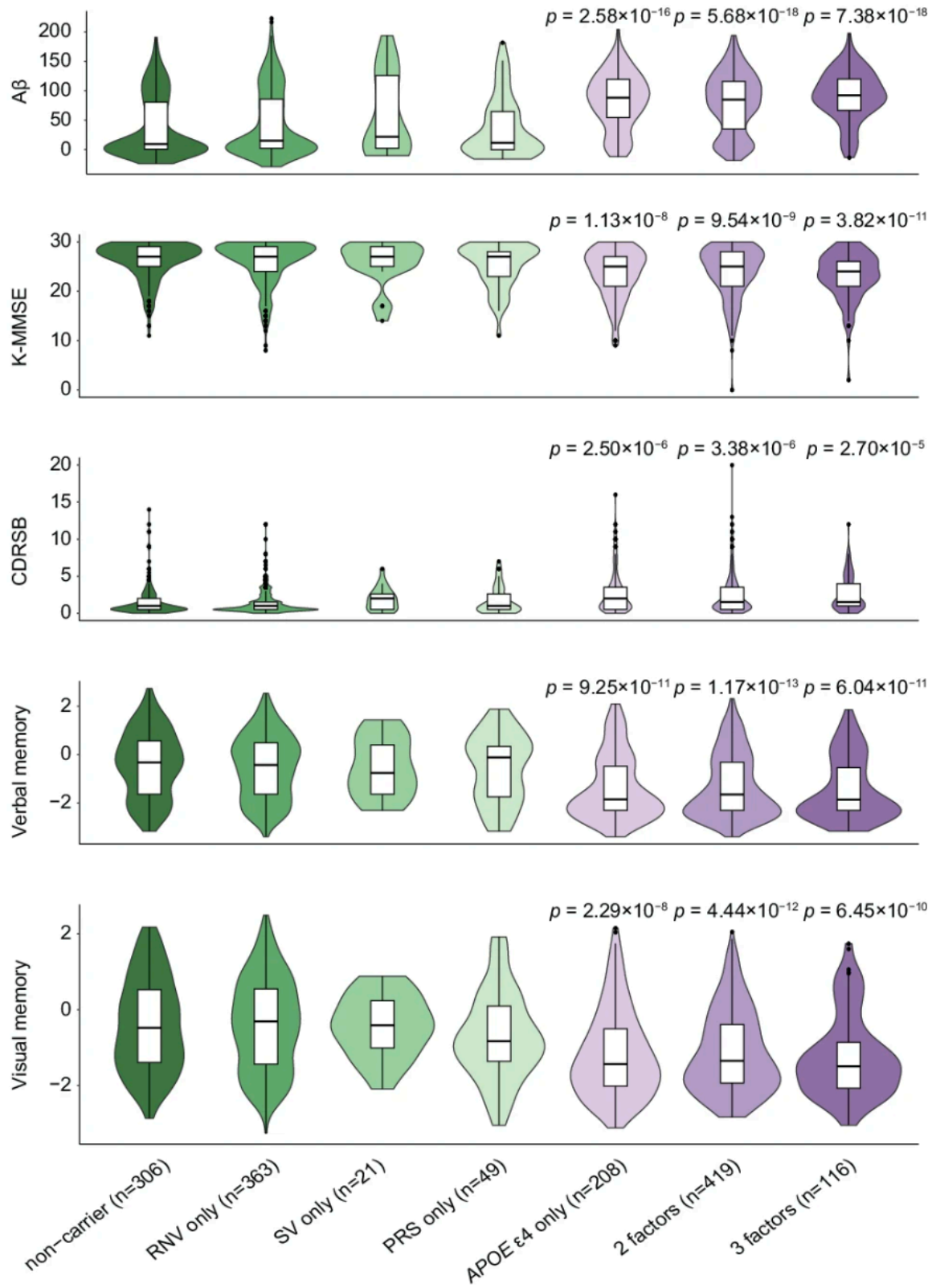


Figure 51: KROAD2025-NatCommun-Fig5

ated with Alzheimer's disease. Nature Communications. https://www.nature.com/articles/s41467-025-59949-y. License: CC-BY.

24.4 Example 3: Autism Spectrum Disorder (ASD)

Autism spectrum disorder (ASD) is perhaps the most genetically heterogeneous condition we'll discuss. It's not one disorder—it's a spectrum of neurodevelopmental differences with diverse genetic causes.

24.4.1 De Novo Variants

Many cases of autism—especially those classified as level 3 (requiring very substantial support)—are associated with **de novo variants**: new mutations that weren't inherited from either parent. These include: – **De novo protein-truncating variants (dnPTV)**: mutations that create premature stop codons – **De novo copy-number variants (dnCNV)**: large deletions or duplications

These variants often hit genes involved in brain development, especially during fetal and early postnatal periods. Each individual mutation is rare, but collectively they account for a meaningful fraction of autism cases.

24.4.2 Inherited Rare Variants

Many people with autism carry **inherited rare variants**—missense mutations or regulatory variants passed from parents who may not have autism themselves. Each variant might slightly increase risk without being sufficient to cause the disorder on its own.

This is sometimes called the "additive rare-variant model." You inherit a few risk variants from each parent, they add up, and if you cross a threshold, you develop ASD.

24.4.3 Polygenic Risk

Common SNPs explain about **30–40% of population-level liability** for autism (Al-Beltagi et al. 2024, World J Clin Pediatr). These are the typical GWAS variants—thousands of common alleles with tiny effects. As with the other disorders, polygenic risk modulates the impact of rare variants.

24.4.4 Syndromic Autism

Some autism cases occur as part of known genetic syndromes: – **Fragile X syndrome** (mutation in FMR1) – **Rett syndrome** (mutation in MECP2) – **Tuberous sclerosis** (mutations in TSC1 or TSC2)

In these cases, autism features are just one part of a broader clinical picture. These syndromic cases are easier to diagnose genetically, but they're a minority of all ASD.

24.4.5 Sex Differences and the Female Protective Effect

Autism is diagnosed about **4 times more often in males** than in females. One explanation is the **female protective effect**: females appear to require a higher genetic burden to develop autism. In other words, females can carry more risk variants without crossing the threshold into clinical diagnosis. When females do develop autism, they often carry more rare variants or a higher polygenic score than affected males.

This sex difference might be biological (X-linked genes, hormonal effects) or diagnostic (females with autism might be underdiagnosed because they present differently).

24.4.6 Functional Convergence

Despite involving hundreds of different genes, autism-associated variants converge on specific biological processes: - **Neuronal development** (how neurons form and migrate during brain development) - **Synaptic signaling** (how neurons communicate at synapses) - **Chromatin regulation** (how DNA is packaged and which genes are turned on or off)

Many of these genes are especially active during **mid-fetal cortical development**—the period when the cortex is forming. This suggests that autism often arises from disruptions during a specific developmental window (Yuen et al. 2024, Cell Genom).

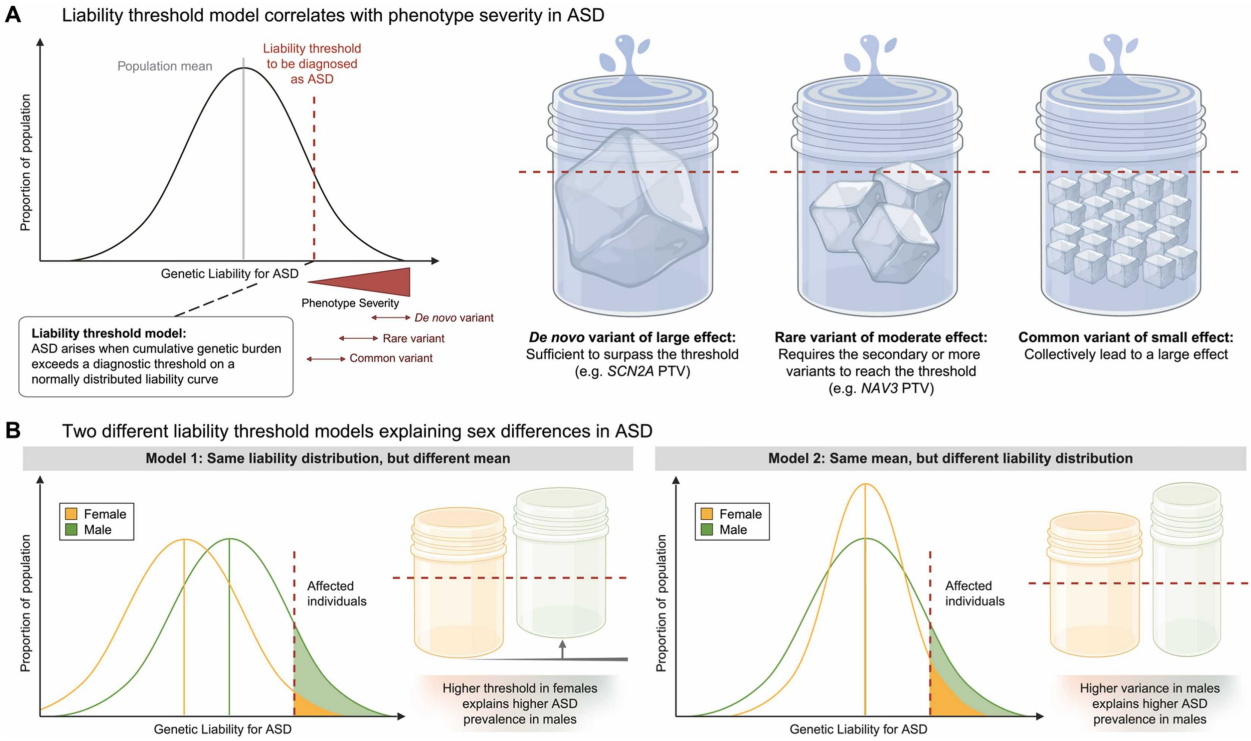


Figure 52: Kim2025-MolCells-Fig3

Figure: Multiple genetic contributors to autism risk. This figure shows how different types of genetic variants each contribute to ASD liability. De novo variants (both CNVs and PTVs), inherited rare variants, polygenic risk from common SNPs, and X-linked variants all add to the total genetic burden. No single category explains all cases—autism has a truly composite architecture where multiple variant types combine to determine individual risk. Source: Kim, Y.S. et al. (2025). Autism spectrum disorder: An evolutionary perspective. Molecules and Cells. https://doi.org/10.1016/j.mocell.2025.100248 License: CC-BY.

Figure: Network of ASD-associated genes converging on synaptic and chromatin pathways. Despite genetic heterogeneity, most ASD-associated genes interact in common functional networks. The major hubs are synapse formation (how neurons connect), chromatin regulation (how genes are turned on and off), and synaptic transmission (how neurons communicate). Different genes disrupt different nodes in these networks, but they all affect the same core processes. Many roads, but they lead to the same destination. Source: Satterstrom, F.K. et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. https://www.cell.com/cell/fulltext/S0092-8674(19)31398-4. License: CC-BY.

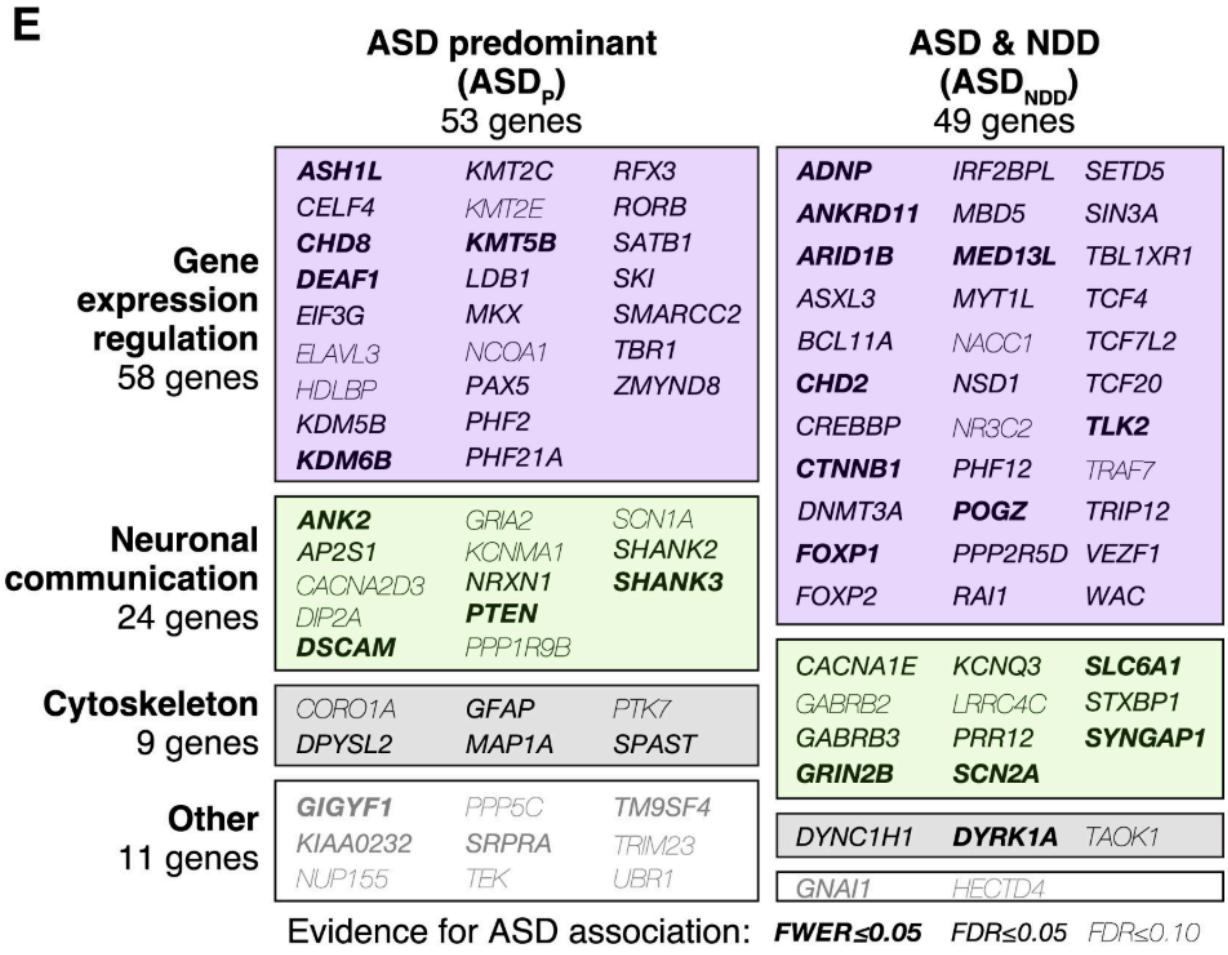


Figure 53: Satterstrom2020-Cell-Fig4

24.5 Genetic Heterogeneity and Biological Convergence

Let's step back and see the pattern. All three disorders—DCM, Alzheimer's, and autism—show **genetic heterogeneity**: different patients have different causal genes or variants leading to clinically similar outcomes.

In DCM, one patient might have a TTN mutation, another might have high polygenic risk, and a third might have multiple rare variants in different genes. Yet they all end up with a dilated, weakened heart.

In Alzheimer's, one patient might carry APOE $\varepsilon 4/\varepsilon 4$, another might have a rare TREM2 variant, and a third might have high polygenic risk. But they all accumulate amyloid plaques and lose neurons.

In autism, one child might have a de novo CNV, another might inherit multiple rare variants, and a third might have high polygenic risk. But they all show differences in social communication and repetitive behaviors.

This is the paradox of complex diseases: **many genetic routes, but shared clinical outcomes**.

How is this possible? Because of **biological convergence**. Despite the genetic diversity, the variants funnel into common pathways: – In DCM: sarcomere structure, cell adhesion, ECM – In Alzheimer's: microglial clearance, lipid metabolism, amyloid processing – In Autism: synaptic signaling, chromatin regulation, neuronal development

Think of it like a river system. Many small streams start in different places—some in mountains, some in forests, some in plains. But they all flow into the same major rivers, which eventually reach the ocean. The starting points are heterogeneous, but the end point converges.

24.5.1 Features of Complex Genetic Architecture

Let's summarize the key features that make genetic architecture complex:

Locus heterogeneity: Different genes can cause the same phenotype. In DCM, mutations in TTN, LMNA, DSP, or FLNC all lead to heart failure. The genes are different, but the outcome is similar.

Allelic heterogeneity: Different mutations in the same gene can have different effects. A missense mutation in TTN might cause mild DCM, while a truncating mutation causes severe early-onset disease. Even within one gene, there's variability.

Temporal context: When does the gene act? Some genes matter during fetal development (like many autism genes). Others act in adulthood (like Alzheimer's risk genes). The timing shapes the phenotype.

Sex differences: Males and females can have different genetic susceptibilities. The female protective effect in autism is one example. Sex hormones, X-chromosome dosage, and other factors modulate genetic risk.

Penetrance and modifiers: Having a pathogenic mutation doesn't guarantee disease. Penetrance—the probability that a genotype leads to the phenotype—is influenced by polygenic background, other rare variants, and environmental factors.

Environmental interactions: Genetics doesn't act in a vacuum. Diet, infections, toxins, stress, and lifestyle all interact with genetic risk. In Alzheimer's, vascular health modifies genetic risk. In DCM, viral infections or alcohol use can trigger disease in genetically susceptible people.

Pathway convergence: This is the flip side of genetic heterogeneity. Diverse variants converge on shared biological pathways. It's what allows clinically similar outcomes despite genetic diversity.

So **genetic architecture** isn't a simple blueprint. It's a dynamic landscape where multiple factors interact. Different people reach the same disease endpoint through different paths, and the same genetic variant can have different effects depending on the context.

24.6 Why This Matters

Understanding genetic architecture is not just an academic exercise. It has real implications for how we interpret genetic tests, counsel patients, and develop treatments.

Interpreting genetic tests: If a patient has DCM and you find a TTN mutation, you can't just say "case closed." You also need to consider their polygenic risk. A TTN carrier with low PGS might never develop symptoms, while one with high PGS might develop severe disease early. The mutation is one piece of the puzzle, not the whole answer.

Genetic counseling: When you tell someone they carry a risk variant, context matters. Are they in a high polygenic risk group? Do they have other rare variants? What's their family history and environment? All of this shapes their actual risk.

Drug development: If multiple genetic variants converge on the same pathway—like microglial efferocytosis in Alzheimer's—that pathway becomes an attractive drug target. You don't need to fix every individual variant. You just need to boost the impaired pathway.

Precision medicine: Understanding genetic architecture allows personalized risk prediction and treatment. If someone has high polygenic risk for DCM, you can monitor them closely and intervene early. If someone has a rare variant plus high polygenic risk, they might need more aggressive treatment.

24.7 Putting It All Together

Let's recap what we've learned about genetic architecture.

Most traits and disorders have **composite genetic architectures** that combine rare variants, common variants, and modifying factors. It's not one gene, one disease. It's many genes, many variants, interacting in complex ways.

The same clinical diagnosis can result from **different genetic causes**. This is genetic heterogeneity. In DCM, some patients have rare pathogenic mutations, others have high polygenic risk, and still others have multiple rare variants. Yet they all present with similar symptoms.

Despite this heterogeneity, many genetic variants converge on **shared biological pathways**. In DCM, it's muscle structure and ECM. In Alzheimer's, it's microglial clearance and lipid metabolism. In autism, it's synaptic signaling and chromatin regulation. This functional convergence explains how diverse genetic routes produce similar clinical outcomes.

Effect size, allele frequency, sex, developmental timing, and environment all influence disease expression. A variant's impact depends on context. When does it act? In which tissues? In males or females? Against what polygenic background? With what environmental exposures? All of these factors modulate genetic risk.

Grasping this complexity is essential for interpreting genomic results and advancing personalized medicine. We can no longer think in simple Mendelian terms for most diseases. We need to embrace the complexity—the heterogeneity, the interactions, the context-dependence—to truly understand how genetics shapes health and disease.

Genetic architecture is messy. But it's also beautiful, because it shows us how the same biological systems can be disrupted in many different ways, yet still funnel into coherent disease mechanisms. That gives us hope: if we understand the convergent pathways, we can develop therapies that work across genetically diverse patients. We don't need to fix every mutation—we just need to fix the shared pathway.

25 Part 4. Recombination, Linkage and Population: From Individual Genomes to Collective Variation

25.1 When Genes Refuse to Assort Independently

In 1910, Thomas Hunt Morgan discovered something that violated Mendel's law of independent assortment. Tracking fruit fly traits in his cramped Columbia laboratory, he noticed that eye color and wing shape—both on the X chromosome—traveled together through generations far more often than expected. Parental combinations stuck together. The 9:3:3:1 ratios were nowhere to be seen.

Morgan realized the genes weren't independent—they were **physically linked** on the same chromosome. When chromosomes segregated during meiosis, linked genes traveled together. But occasionally, the linkage broke. Something was shuffling the genetic deck even within a single chromosome.

That something was **recombination**—the physical exchange of DNA between paired chromosomes during meiosis. His student Alfred Sturtevant recognized that recombination frequency could measure distance: genes that rarely recombined were close together; genes that often recombined were far apart. In one inspired night, the 19-year-old constructed the first genetic map, turning recombination from a curiosity into a ruler.

This shifted genetics profoundly. Genes became physical objects with locations on chromosomes that could be mapped and measured. But another shift was happening: geneticists were moving from studying individual families to studying **populations**—understanding how genetic variation distributes across groups and changes over time.

Then in 2025, Palsson and colleagues sequenced over 5,000 Icelandic families and discovered recombination is far more complex than realized. The dramatic chromosome swaps Morgan studied are just part of the picture. **Non-cross-overs** (gene conversion)—small DNA patches copied from one chromosome to another—are far more common and shape diversity in ways we're only beginning to understand.

This chapter lives at the intersection of these perspectives: how genes are physically linked, how recombination shuffles them, how variation distributes across populations, and how we encode all this as data.

25.2 What This Chapter Is About

This chapter explores genetic variation from three complementary angles: physical organization on chromosomes, distribution across populations, and representation as data.

25.2.1 Six Interconnected Topics

1. Allele Frequency
How common or rare is a variant? You'll learn to calculate allele frequency, interpret what different frequencies mean (common, low-frequency, rare, private), and understand how frequency helps distinguish benign from pathogenic variants using the gnomAD database of 141,456 people.

2. Population Structure
Why do allele frequencies differ between populations? We'll explore how human populations separated by geography and history developed detectable genetic differences. You'll learn to interpret PCA and UMAP plots, and understand the crucial distinctions: race (social construct), ethnicity (cultural identity), ancestry (genetic lineage), and population (breeding group).

3. Discovery of Linkage
How did we learn that genes are connected? We'll trace Morgan and Sturtevant's discoveries—from white-eyed flies to the first genetic map—understanding how recombination frequency became a molecular ruler measured in centiMorgans (cM).

4. Recombination, Linkage, and Haplotype
How does DNA get shuffled during meiosis? Building on Palsson et al. 2025, we'll explore both cross-overs (large DNA swaps) and non-cross-overs (small gene conversions), why NCOs are far more common, how maternal and paternal recombination differ, and how haplotypes—blocks of DNA inherited together—record recombination history.

5. Data Type for Alleles and Population
How do we turn genetic variation into data? You'll learn to read VCF (Variant Call Format) files—the standard format for storing variants. We'll decode genotypes (0/0, 0/1, 1/1), understand INFO fields (AC, AN, AF), and see how multi-sample VCFs organize data from thousands of people.

6. From Statistics to Mechanism
How do population genetics principles apply to real problems? We'll see how linkage maps guide disease gene discovery, how population structure correction prevents false GWAS associations, and how haplotype analysis enables everything from ancestry testing to precision medicine.

25.3 What Makes This Different

Individual Genetics (Parts 2–3)	Population Genetics (Part 4)
What does this variant do in one person?	How common is this variant across populations?
Single genotype (0/0, 0/1, 1/1)	Allele frequency (5%, 50%, 95%)
Family pedigrees, Mendelian ratios	Population databases, frequency distributions
Variant effect and penetrance	Evolutionary forces (drift, selection, migration)
Clinical interpretation of individual variants	Population context for variant interpretation

The shift: From studying what variants do to understanding where variants are, how common they are, how they're inherited together, and how we represent this as data.

25.4 How to Approach This Chapter

25.4.1 Think in Three Scales

1. **Physical scale**: Genes have locations on chromosomes. Proximity creates linkage. Two variants 1 kb apart rarely recombine. Two variants 10 Mb apart often recombine.
2. **Population scale**: When you study one person, you see genotype (0/1). When you study a population, you see allele frequency (35%). Both perspectives are essential.
3. **Data scale**: Individual genomes versus biobank-scale datasets require different representations. VCF format is the universal language, but sparse formats are needed at scale.

25.4.2 Connect History to Biology

Every pattern in genomic data has a historical explanation: – Population structure reflects thousands of years of migration and isolation – Recombination hotspots reflect conserved molecular mechanisms – Allele frequencies carry signatures of drift, selection, and admixture – Haplotype blocks record where recombination has and hasn't happened

25.4.3 Master the Data Format

Understanding VCF isn't just a technical detail—it's essential infrastructure. When a methods section says "we filtered variants with MAF < 0.01," you need to know what that means and why it matters.

25.4.4 Key Principles to Watch

Recombination creates and constrains diversity: It generates new haplotypes by mixing grandparental chromosomes, but rare recombination between nearby variants creates haplotype blocks—the units of inheritance between individual SNPs and whole chromosomes.

Math must connect to biology: Allele frequency 0.05 means 5% of chromosomes carry this variant. Recombination frequency 0.10 reflects ~10 cM genetic distance. PCA components are mathematical but correlate with ancestry. Always translate math into biological meaning.

Scale changes everything: Analyzing one genome differs from analyzing millions. At biobank scale, you need sparse formats, efficient algorithms, and careful quality control. But biological principles remain the same.

25.5 Why This Matters

Modern genomics requires understanding:

Clinical interpretation: When you find a variant in a patient, you check gnomAD to see if it's common (probably benign) or rare (possibly pathogenic). Population frequency is critical for clinical decision-making.

Study design: GWAS without population structure correction yields false associations. Understanding ancestry and correcting for it is essential.

Disease gene mapping: Linkage analysis guided discovery of genes for cystic fibrosis, Huntington's disease, and hundreds of Mendelian disorders. Modern recombination maps make this mapping even more precise.

Evolutionary insights: Allele frequencies and population structure reveal human migration patterns, bottlenecks, and admixture. Most genetic variation (85–95%) exists within populations, not between them—a finding with profound implications.

Practical applications: From ancestry testing to forensics to HLA matching for transplantation, these principles underlie diverse applications across medicine, agriculture, and conservation.

25.6 Learning Objectives

By the end of this chapter, you should be able to:

1. **Calculate and interpret allele frequency**, distinguishing MAF of 0.001% versus 5%, and connecting frequency to evolutionary forces
2. **Explain population structure** and distinguish race (social construct), ethnicity (culture), ancestry (genetic lineage), and population (breeding group)
3. **Interpret PCA and UMAP plots** showing genetic similarity and population clusters
4. **Describe linkage discovery** from Mendel through Morgan to modern recombination maps
5. **Distinguish cross-overs from non-cross-overs**, explaining mechanisms, frequencies, and sex differences
6. **Read VCF files**, decoding genotypes and INFO fields

7. **Define haplotypes** as blocks of linked variants recording recombination history
8. **Explain why population context matters** for variant interpretation, study design, and disease mapping

---

25.7 A Note on Terminology

Key distinctions you'll encounter:

- **Linkage** (physical connection on chromosome) vs. **linkage disequilibrium** (statistical association in population)
- **Cross-over** (swaps flanking regions) vs. **non-cross-over** (copies small patch)
- **Haplotype** (phased alleles on one chromosome) vs. **genotype** (unphased allele combination)
- **Allele frequency** (counts chromosomes) vs. **genotype frequency** (counts individuals)
- **Reference allele** (0, matches reference genome) vs. **alternate allele** (1, the variant)—reference doesn't mean "normal"

Context will clarify these as you read.

25.8 Let's Begin

Morgan spent years breeding flies in cramped bottles, discovering that genes have locations and that recombination measures their distance. Today, you can download VCF files with frequencies from 141,456 people and recombination maps from 5,000 Icelandic families. The scale is unimaginable compared to Morgan's work.

But you're studying the same thing: how genetic material is organized, transmitted, and varied. The tools have changed from flies to files, from pedigrees to populations, from phenotypes to sequences.

The questions remain: How do genes travel through generations? How do they vary across individuals? How do we make sense of that variation?

In Part 2, you learned how individual variants cause disease. In Part 3, you learned how many variants combine to create complex traits. Now you'll learn how those variants are organized on chromosomes, distributed across populations, shuffled through recombination, and encoded as data.

Let's see how Morgan's century-old insight scales to modern genomics.

26 Chapter 20. Allele Frequency

Here's a question that might seem obvious but turns out to be profound: is a genetic variant rare or common?

If you sequence one person's genome and find a variant that's not in the reference sequence, what does that tell you? Not much, actually. Maybe it's a harmless difference that appears in millions of people worldwide. Maybe it's a disease-causing mutation that just arose in this person. Maybe it's somewhere in between. You can't tell from looking at one genome.

Genetic variation is observed in individuals, but its meaning emerges only when viewed within a **population**. When a DNA variant occurs in a single person, it's a private event—we don't know if it's important or just noise. But when it's shared among many individuals, it reflects the combined forces of **mutation, selection, and genetic drift** that have shaped the genetic landscape over thousands of generations.

To describe how common or rare a variant is, geneticists use the concept of **allele frequency**. It's a simple idea—what proportion of all copies of a gene carry this particular variant?—but it turns out to be remarkably

powerful. Allele frequency tells us about evolutionary history, disease risk, and population structure all at once.

Recent large-scale genome analyses have given us an unprecedented view of allele frequency. **141,456 people** from diverse populations were sequenced in the gnomAD database, cataloging hundreds of millions of variants (Karczewski et al. 2020, Nature). This dataset is large enough to detect variants that appear in only one or two people out of hundreds of thousands. It's revolutionized how we think about genetic variation.

Let's start with the basics and build up to why this matters.

26.1 Definition of Allele Frequency

Allele frequency is the proportion of a specific allele among all alleles at a given genetic locus in a population. Think of it as asking: if we collected all the copies of this gene from everyone in the population, what fraction would have this particular variant?

Here's where being human (and diploid) matters. Each person has two copies of every autosomal gene—one from mom, one from dad. So each person contributes two alleles to the population count.

Let's work through a concrete example. Suppose we're looking at a specific position on chromosome 21: chr21:2232323. At this position, most people have an A (adenine), but occasionally someone has a T (thymine) instead.

Now imagine we've sequenced **141,456 people**—that's the size of the gnomAD dataset. We find that **one person** is heterozygous for this variant—they have one A and one T. Everyone else is AA.

How do we calculate the allele frequency of the T allele?

First, count the total number of alleles in the dataset:

```
141,456 people × 2 alleles per person = 282,912 alleles
```

Second, count how many T alleles we observed. Our one heterozygote has one T allele, so:

```
1 T allele out of 282,912 total alleles
```

The allele frequency is:

```
1 ÷ 282,912 = 0.0000035
```

We usually express this as a percentage: **0.00035%**. That's extremely rare—about 3.5 in a million.

26.2 How to Calculate Allele Frequency

The general formula is simple:

Allele frequency = (number of variant alleles) ÷ (2 × number of individuals)

The key is remembering to count both alleles from homozygotes and one allele from heterozygotes. Let's look at a few scenarios in the gnomAD dataset (141,456 people = 282,912 alleles):

Scenario	Heterozygotes (AT)	Homozygotes (TT)	T alleles	Calculation	Frequency
Very rare	1	0	1	1 ÷ 282,912	0.00035%
Rare	2	3	8	8 ÷ 282,912	0.0028%
Low frequency	10	0	10	10 ÷ 282,912	0.0035%

Scenario	Heterozygotes (AT)	Homozygotes (TT)	T alleles	Calculation	Frequency
More common	0	10	20	20 ÷ 282,912	0.0071%

Counting rule: Each heterozygote contributes 1 variant allele, each homozygote contributes 2 variant alleles.

26.3 Minor Allele Frequency (MAF)

At most positions in the genome, we find two alleles in the population. One is more common—we call this the **major allele**. The other is less common—the **minor allele**. The frequency of the minor allele is called the **Minor Allele Frequency (MAF)**.

For example, suppose at a certain position: – 92% of alleles are A (the major allele) – 8% of alleles are T (the minor allele)

The MAF is 0.08, or 8%.

Why do we care about the minor allele specifically? Because in population genetics and GWAS, we're usually interested in the less common variant. The major allele is often just the "reference" or "normal" state. The minor allele is what varies, and variation is what we study.

MAF is also useful for categorizing variants. A variant with MAF 〉 5% is common—lots of people carry it. A variant with MAF 〈 1% is rare—few people have it. These categories matter because common and rare variants behave differently in studies and have different evolutionary implications.

26.4 Variant Categories by Frequency

Geneticists divide variants into categories based on allele frequency. However, **the exact cutoffs vary considerably between studies and research contexts**. There is no universal consensus, and different fields use different thresholds depending on their specific questions and datasets.

26.4.1 Commonly Used Classification Schemes

Multiple classification systems exist in the literature, reflecting different research needs:

Study	Common	Low-frequency	Rare	Ultra-rare/Singleton
1000 Genomes (Byrska-Bishop et al. 2022, Cell)	AF 〉 1%	0.5% ≤ MAF 〈 5%	AF ≤ 1%	Singletons (AC = 1)
Schizophrenia genetics (Akingbuwa et al. 2022, Mol Psychiatry)	MAF 〉 1%	–	AF 〈 0.1% (excluding singletons)	Singletons only

Study	Common	Low-frequency	Rare	Ultra-rare/Singleton
Gene expression heritability (Glassberg et al. 2019, Nat Genet)	–	–	–	Singletons; gnomAD MAF 〈 0.01%
AD/dementia GWAS (Andrews et al. 2023, EBioMedicine)	MAF 〉 5%	1–5%	MAF 〈 1%	–
COVID-19 severity (Fallerini et al. 2021, Hum Genet)	MAF ≥ 5%	1% ≤ MAF 〈 5%	0.1% ≤ MAF 〈 1%	MAF 〈 0.1%

Key observation: As noted in a comprehensive review, "the threshold of MAF in rare variants has not yet been clearly defined" and varies between 0.1% and 5% across studies (Momozawa & Mizukami 2021, J Hum Genet). The cutoffs are somewhat arbitrary and should be chosen based on the specific research question, disease penetrance, and available sample size.

26.4.2 Why Do Cutoffs Vary?

The variation in frequency thresholds reflects different research priorities:

Disease-specific considerations: For highly penetrant neuropsychiatric disorders, researchers often focus on ultra-rare variants or singletons that likely represent recent, highly deleterious mutations. In schizophrenia genetics, ultra-rare variants (singletons only) showed the strongest association with disease, while "rare" variants (AF 〈 0.1% excluding singletons) showed no significant enrichment (Akingbuwa et al. 2022, Mol Psychiatry). This suggests that the most deleterious variants are truly ultra-rare, supporting stringent thresholds for some disorders.

Methodological and technological constraints: The 1000 Genomes Project with high-coverage sequencing (30X) defined multiple frequency categories: very rare (MAF 〈 0.5%), rare (0.5% ≤ MAF 〈 5%), and common (AF 〉 1%), reflecting their ability to robustly detect variants across the entire frequency spectrum (Byrska-Bishop et al. 2022, Cell). Array-based GWAS historically focused on MAF 〉 5% due to imputation limitations, while whole-genome sequencing enables finer distinctions.

Statistical and biological considerations: Studies examining gene expression found that 90% of singleton heritability derives from variants with population frequency 〈 0.01% in gnomAD (Glassberg et al. 2019, Nat Genet). This suggests that functionally important regulatory variants are kept at extremely low frequencies by selection, motivating even stricter definitions of "ultra-rare" for specific biological questions.

When reading genetic studies, always check how the authors define their frequency categories—it matters for interpretation.

26.4.3 General Framework

Despite variation in exact thresholds, most studies follow this general pattern:

Category	Typical MAF Range	Typical Allele Age	Selection Pattern	Study Implications
Common	〉5%	Thousands to tens of thousands of years	Usually neutral or weakly selected	Useful for GWAS; high statistical power
Low-frequency	1–5%	Intermediate	Mild selection or recent spread	Moderate power; specialized designs needed
Rare	0.1–1% (or 〈 1%)	Recent (hundreds to thousands of years)	Often under purifying selection; may be functional	Large effects but hard to study; low power
Ultra-rare	〈 0.1%	Very recent	Strong selection; often deleterious	Detected mainly by sequencing; often private
Private (singletons)	~0%	Very recent (1–10 generations); often de novo	Not yet filtered by selection	Uncertain significance; most will be lost

26.4.4 Key Insights by Category

Common variants have persisted through many generations because they don't harm fitness much. Most are in non-coding regions or have minimal functional impact. They're old enough that selection has had plenty of time to act—if they were harmful, they'd be gone by now. Their high frequency makes them ideal for GWAS, where you need many people carrying the variant to detect associations.

Rare variants tell a different story. Here's a striking pattern from the gnomAD data: when loss-of-function variants were examined—mutations that completely break a gene by introducing a premature stop codon or disrupting splicing—most were found to be **singletons** (Karczewski et al. 2020, Nature). Each variant appeared in only one person out of 141,456. Why? Because breaking an important gene reduces fitness, and natural selection removes these variants before they can spread. Rare variants are often functional and frequently deleterious—they're kept rare by selection constantly pushing them down.

The definition of "rare" varies by context. For highly penetrant Mendelian diseases where even heterozygotes show symptoms, researchers often use stricter thresholds (MAF 〈 0.01% or even 〈 0.001%) to filter out benign variants (Gambin et al. 2018, Eur J Hum Genet). For complex diseases with reduced penetrance, more relaxed thresholds (MAF 〈 1% or 〈 5%) are common. The choice depends on the expected disease prevalence, penetrance, and genetic heterogeneity.

Private variants are too new to tell us much about selection. They haven't been tested by evolutionary forces yet. Most arose recently—perhaps de novo in the individual, or inherited from a parent who got a new mutation. Most will disappear in the next few generations through random chance or selection, but a tiny fraction might persist or even spread if they happen to be neutral or beneficial.

The distinction between these categories has become sharper as datasets have grown. With over 140,000 people sequenced, we can confidently distinguish variants that appear in hundreds of people from those that appear in just one or two—something impossible with smaller sample sizes.

26.5 How Selection Shapes Allele Frequency

Allele frequency isn't just a descriptive statistic—it's a window into evolutionary history. The frequency of a variant tells you something about how natural selection has acted on it.

Think of natural selection as a filter. Beneficial mutations pass through easily and spread. Neutral mutations drift through randomly—some make it, some don't. Harmful mutations get caught and removed. The frequency of a variant reflects which of these processes has been dominant.

26.5.1 The Frequency–Function Relationship

There's a striking pattern in genomic data: **the more severe a variant's functional impact, the rarer it tends to be.**

- **Synonymous variants** (don't change the amino acid) → often common
- **Missense variants** (change one amino acid) → intermediate frequencies
- **Nonsense variants** (truncate the protein) → almost always rare

This gradient reflects selection intensity. Variants that break genes completely are removed quickly. Variants with subtle effects take longer to remove, or may drift to moderate frequency if their effects are small enough.

26.5.2 Selection Coefficient and Population Size

The fate of a new variant depends on two things: 1. **Selection coefficient (s)**: How much it affects fitness 2. **Effective population size (Ne)**: How large the population is

Here's the critical relationship:

Selection is only effective when s ⟩ 1/Ne

If your selection coefficient is smaller than this threshold, the variant behaves like it's neutral—its fate is determined by random chance (genetic drift) rather than by selection.

In large populations: – Ne is large → 1/Ne is small – Even tiny fitness effects matter – Selection efficiently removes harmful variants – Few deleterious alleles persist

In small populations: – Ne is small → 1/Ne is large – Only large fitness effects overcome drift – Mildly harmful variants behave neutrally – More deleterious alleles persist by chance

Human populations have moderate effective sizes (Ne □ 10,000–50,000, depending on ancestry), so selection can act on variants with effect sizes down to about 0.01–0.001%. But that still means many mildly deleterious variants persist at low frequency, drifting around until selection eventually removes them or they disappear by chance.

This explains why loss-of-function variants are overwhelmingly rare but not completely absent. They keep arising through mutation, and in a finite population, it takes time for selection to purge them. Some stick around for a few generations before disappearing. A few get lucky and drift to higher frequency before selection catches up.

26.5.3 The Allele Frequency–Effect Size Spectrum

This inverse relationship between allele frequency and effect size creates a distinctive pattern in genetic disease. Variants across the genome occupy different positions along two dimensions: how common they are (minor allele frequency) and how much they affect disease risk (effect size).

Andrews2023-EBioMed-Fig1

Figure: Genetic architecture of Alzheimer's disease showing the relationship between allele frequency and effect size. Variants are conceptualized along two dimensions: population minor allele frequency (x-axis, log scale) and effect size as odds ratio (y-axis, log scale). Highly penetrant mutations in APP, PSEN1, and PSEN2 causing autosomal dominant AD are extremely rare (MAF ⟨ 0.0001) with large effect sizes (OR ⟩ 5). APOE genotypes show intermediate patterns with APOE $\varepsilon 4/\varepsilon 4$ at ~2% frequency conferring substantial risk (OR ~12). Most variants discovered by GWAS are common to low-frequency (MAF ⟩ 0.01) with small effect sizes (OR ⟨ 1.5). The gray shaded area illustrates statistical power to detect associations at different effect sizes and frequencies with current sample sizes. This pattern reflects natural selection: variants with strong detrimental effects are kept rare by purifying selection, while common variants that have persisted must have weak or neutral effects on fitness. The same inverse relationship between frequency and effect size appears across many complex diseases. Source: Andrews, S.J. et al. (2023). The

complex genetic architecture of Alzheimer's disease: novel insights and future directions. EBioMedicine 90, 104511. https://doi.org/10.1016/j.ebiom.2023.104511. License: CC BY-NC-ND 4.0.

26.6 Summary

Allele frequency—the proportion of a variant in a population—is the fundamental unit of population genetics. It's calculated simply (count variant alleles, divide by total alleles), but this single number encodes evolutionary history, functional impact, and disease risk.

26.6.1 Core Concepts

Frequency categories vary by study context, but generally span common (MAF 〉 5%), low-frequency (1–5%), rare (〈 1%), and ultra-rare/private variants. These categories reflect variant age and selection history: common variants are old and survived selection, rare variants are recent and often functional, private variants are brand new and untested by evolution.

Selection shapes frequency: The gnomAD dataset revealed that loss-of-function variants are overwhelmingly singletons—direct evidence of purifying selection removing harmful mutations. The inverse relationship between effect size and frequency appears across diseases: variants that strongly affect fitness cannot reach high frequency.

Population size modulates selection efficiency: In large populations like humans (Ne □ 10,000–50,000), even mildly deleterious variants face selection pressure. But drift still matters—many rare variants persist temporarily before selection removes them or chance makes them disappear.

26.6.2 Why This Matters

Allele frequency is essential for: – **Variant interpretation**: Distinguishing pathogenic mutations (rare) from benign polymorphisms (common) – **Study design**: Powering GWAS (need common variants) versus rare variant studies (need large samples) – **Understanding disease architecture**: Common variants explain much heritability despite small individual effects

26.6.3 Looking Ahead

This chapter established what allele frequencies are and why they vary. But a critical question remains: why do the same variants have different frequencies in different populations?

The next chapter on **Population Structure** will explore how human history—migration, isolation, admixture—created frequency differences between populations. We'll learn to distinguish genetic ancestry from social categories, interpret PCA plots showing population relationships, and understand why population structure matters for genomic medicine.

Then we'll see how variants don't just vary in frequency—they travel together on chromosomes. **Linkage and Recombination** will explain how physical proximity creates correlations between variants, how recombination breaks these correlations, and how haplotypes—blocks of variants inherited together—record this history.

Finally, we'll learn to work with this data. **VCF format** encodes allele frequencies, genotypes, and annotations in a standardized way, making modern population-scale genomics possible.

Allele frequency is where population genetics begins. Get comfortable with it—you'll see it in every chapter that follows.

27 Chapter 21. Population Structure

Imagine you're standing in a busy airport, watching people from all over the world walk by. You notice that some people look similar to each other—maybe they share certain facial features, skin tones, or body builds. Without knowing anything about where they're from, you might guess that people who look similar probably come from the same region or share common ancestry.

Genetics tells a similar but much more precise story. When we look at DNA instead of physical appearance, we see that people from the same geographic region tend to share certain genetic variants more often than people from distant regions. These patterns aren't sharp boundaries—they're gradual gradients, like colors blending into each other on a map. But they're measurable, and they tell us about human history: ancient migrations, population bottlenecks, and the mixing of groups over thousands of years.

This is what geneticists call **population structure**—systematic differences in allele frequencies among groups. It's not about dividing humans into categories. It's about recognizing that genetic variation isn't randomly distributed. History has left traces in our genomes, and we can read those traces.

Understanding population structure matters for both scientific and ethical reasons. Scientifically, we need to account for it in genetic studies to avoid false conclusions. If a variant is common in one population and rare in another, and you're studying a disease that's also more common in that first population, you might mistakenly think the variant causes the disease when it's just correlated with ancestry. Ethically, we need to be clear that genetic differences among populations are subtle gradients, not fundamental divisions—and that the concept of biological "races" has no basis in genetics.

27.1 What Is Population Structure?

Population structure refers to the presence of subgroups within a species that differ in their allele frequencies. In humans, we see small frequency differences between continental groups—African, European, East Asian, South Asian, and so on. These differences aren't huge—most genetic variation exists within populations, not between them—but they're detectable and meaningful.

How did these differences arise? The answer is history.

Humans evolved in Africa and began migrating out roughly 60,000 to 100,000 years ago. As groups moved to different continents, they became partially isolated from each other. Geographic distance, mountains, oceans, and deserts limited gene flow—the exchange of genetic material through reproduction. When gene flow is restricted, populations accumulate differences over time through several mechanisms:

Genetic drift: Random fluctuations in allele frequency from generation to generation. In small populations, drift can be strong. An allele might increase or decrease in frequency just by chance—who happens to have more children, who survives, who doesn't. Over many generations, drift causes populations to diverge even if they start out identical.

Mutation: New variants arise randomly in each population. Most mutations are neutral or slightly deleterious and stay rare. But occasionally, a new variant drifts to moderate frequency in one population while remaining absent in others. This creates population-specific variants.

Natural selection: Different environments favor different alleles. Selection can drive allele frequency differences between populations when environmental pressures differ.

The result is that populations separated by geography and time develop distinct—but overlapping—patterns of allele frequencies. These patterns form subtle genetic gradients across the globe. There are no sharp boundaries. If you sample people along a transect from West Africa through the Middle East to Europe, you see smooth changes in allele frequencies, not sudden jumps.

This is why geneticists emphasize that **population structure does not reflect biological "races."** Race is a social category based on visible traits like skin color. Genetically, human variation is continuous. Populations

blend into each other. Any attempt to draw sharp lines is arbitrary.

27.2 Race, Ethnicity, Ancestry, and Population—Different Concepts

In everyday conversation, people use words like "race," "ethnicity," and "ancestry" interchangeably. But in genetics, these terms have distinct meanings, and it's essential to keep them separate.

Concept	Definition	Nature	Example
Race	Socially defined category based on perceived physical traits	Social construct; varies by culture and time	"Black," "White," "Asian"
Ethnicity	Cultural identity including language, nationality, traditions	Cultural; may correlate weakly with ancestry	Han Chinese, Yoruba, Navajo
Ancestry	Genetic lineage reflecting where ancestors lived	Biological and historical; probabilistic	"50% European, 50% East Asian ancestry"
Population	Group interbreeding more with each other than outsiders	Biological; defined by gene flow patterns	"European population," "East Asian population"

Race has no precise biological definition. There's no set of genes that cleanly separates humans into racial groups. Skin color, for example, is controlled by a handful of genes that vary continuously across populations. Two people classified as "White" might be more genetically different from each other than either is from someone classified as "Black." Because of this imprecision—and because racial categories have been misused historically to justify discrimination—geneticists avoid using race as a biological concept.

Ethnicity is primarily about culture, not DNA. You can be ethnically French without having French ancestry, or vice versa.

Ancestry is a probabilistic description of where your DNA comes from, not a box you fit into. These percentages reflect which populations contributed to your genetic makeup based on allele frequency patterns.

Modern human genetics focuses on **ancestry** and **population**, not race or ethnicity. This distinction matters both scientifically and ethically. Understanding this helps researchers correct for ancestry-related bias in genetic studies and appreciate human diversity as a product of shared evolutionary history, not rigid divisions.

27.3 How Population Structure Arises

Population structure emerges through distinct evolutionary mechanisms, each leaving characteristic signatures in the genome.

27.3.1 Geographic Isolation and Drift

When populations live far apart—separated by oceans, deserts, or mountain ranges—gene flow between them is limited. Over time, genetic drift and mutation cause their allele frequencies to diverge. Even if the populations start out genetically identical, isolation ensures they'll become different. The longer they're separated, the more different they become.

27.3.2 Founder Effects

Founder effects happen when a new population is established by a small number of individuals who carry only a fraction of the genetic diversity from their original population. Some alleles might be overrepresented by chance, others might be missing entirely. As the new population grows, it retains this reduced diversity.

The Finnish population exemplifies founder effects and bottlenecks. The population descended from approximately 4,000 founders about 4,000 years ago, followed by a major bottleneck in the 17th century when one-third of the population perished during famine (Peltonen et al. 1999, Human Molecular Genetics). These events created the Finnish Disease Heritage—36 mostly recessive genetic disorders that are far more common in Finland than elsewhere (Norio 2003, Human Genetics). Each disease is typically caused by a single founder mutation that drifted to high frequency during population expansion. This demonstrates how founder effects combined with isolation can dramatically reshape a population's genetic landscape.

27.3.3 Population Bottlenecks

Population bottlenecks occur when a population's size drops sharply—due to disease, famine, or migration—and then recovers. During the bottleneck, rare alleles are lost just by chance, because only a few individuals survive to reproduce. The recovered population has less genetic diversity than before. Bottlenecks leave signatures in the genome: fewer rare variants, longer stretches of linkage disequilibrium, and reduced overall diversity.

27.3.4 Admixture

Admixture is the mixing of previously separated populations. When two groups that have been isolated for thousands of years start interbreeding, their offspring inherit a mosaic of alleles from both ancestries. For example, many African Americans have both West African and European ancestry due to historical admixture. Latino populations often show a mix of Indigenous American, European, and African ancestries.

27.3.5 Selection and Local Adaptation

Selection drives allele frequency differences when environments differ. The lactose tolerance variant (rs4988235 in LCT) shows one of the strongest signals of recent positive selection in Europeans. This variant maintains lactase production into adulthood and reaches frequencies of 70–90% in Northern Europe but remains under 10% in East Asian populations (Tishkoff et al. 2007, Nature Genetics). The frequency gradient across Europe correlates with historical reliance on dairy farming, demonstrating how cultural practices can create selective pressures that shape genetic variation.

27.4 Detecting Population Structure

We know populations differ in allele frequencies, but how do we actually measure and visualize those differences?

27.4.1 The Idea of Distance and Similarity

Start by imagining we're comparing two people's genomes. We look at millions of SNPs across their chromosomes and count how many alleles they share. If they share most of their alleles, they're genetically close. If they differ at many positions, they're genetically distant.

Now scale this up. In a dataset with over 140,000 people and millions of SNPs, you can calculate the genetic distance between every pair of people. People from the same population will be close together. People from different continents will be farther apart.

But here's the problem: you can't visualize millions of dimensions. Each SNP is like a dimension in space. We need a way to simplify this—to collapse all those dimensions down to two or three so we can plot them and see patterns.

That's where dimensionality reduction comes in.

27.4.2 Principal Component Analysis (PCA)

PCA is a mathematical method for simplifying high-dimensional data. It finds new axes—called **principal components (PCs)**—that capture the largest sources of variation in the data.

Think of it like finding the best vantage points to photograph a crowd. PCA identifies angles from which you can see the most variation. The first principal component (PC1) is the direction that shows the biggest differences. The second (PC2) is the next biggest, independent of the first.

Figure: How PCA finds new axes. This simple example shows data points scattered in 2D space. The original axes are horizontal (x) and vertical (y). But the data doesn't spread evenly along these axes—it follows a diagonal pattern. The red arrow shows PC1, the direction of maximum variance (where the data spreads most). The green arrow shows PC2, perpendicular to PC1, capturing the remaining variance. By rotating to these new axes, we can see the data's true structure more clearly. If we wanted to reduce this to 1D, projecting onto PC1 (red arrow) would preserve much more information than projecting onto either original axis. (Image: Wikimedia Commons, CC BY 4.0)

In genetics, PC1 might separate African from non-African ancestry because that's the biggest split in human history (the out-of-Africa migration). PC2 might separate European from East Asian ancestry. PC3 might capture finer substructure within continents.

When you plot individuals on PC1 vs PC2, you get a scatterplot. Each dot is a person. People with similar ancestry cluster together. People with different ancestry are far apart.

27.4.3 Uniform Manifold Approximation and Projection (UMAP)

PCA works well, but it's linear. It finds straight-line directions that capture variance. Sometimes the structure in your data isn't linear—it might have curves or complex shapes.

UMAP is a more flexible method. It preserves both global structure (large-scale patterns, like continents) and local structure (fine-scale patterns, like subpopulations within continents). UMAP maps your high-dimensional data into 2D or 3D space in a way that keeps nearby points close and distant points far.

27.4.4 Mapping Human Genetic Diversity

Over 141,000 people from diverse backgrounds were sequenced and their genetic data analyzed using UMAP, revealing the global landscape of human genetic variation (Karczewski et al. 2020, Nature).

Figure: Global patterns of human genetic diversity. Each point represents one individual's genome plotted in 2D space using UMAP based on millions of genetic variants. Nearby points indicate people with similar allele frequencies. Major ancestral groups form visible clusters: African/African American (purple), European (blue), East Asian (pink), South Asian (green), Latino/Admixed American (orange), and others. Critically, there are no sharp boundaries—populations blend into each other through continuous gradients. The Latino cluster spreads between European, African, and Native American ancestries, demonstrating admixture. The African cluster shows the greatest internal diversity (most spread out), reflecting Africa's status as the ancestral human population with the longest history. Other clusters are tighter, bearing signatures of founder bottlenecks during the out-of-Africa migration. This visualization captures how human genetic variation is structured by history—migration, isolation, and admixture—rather than discrete categories. Source: gnomAD official blog

Look at the clusters. They're distinct but not isolated. There's overlap at the edges—people of mixed ancestry sit between clusters. The shape of each cluster also tells a story. The African cluster is the most

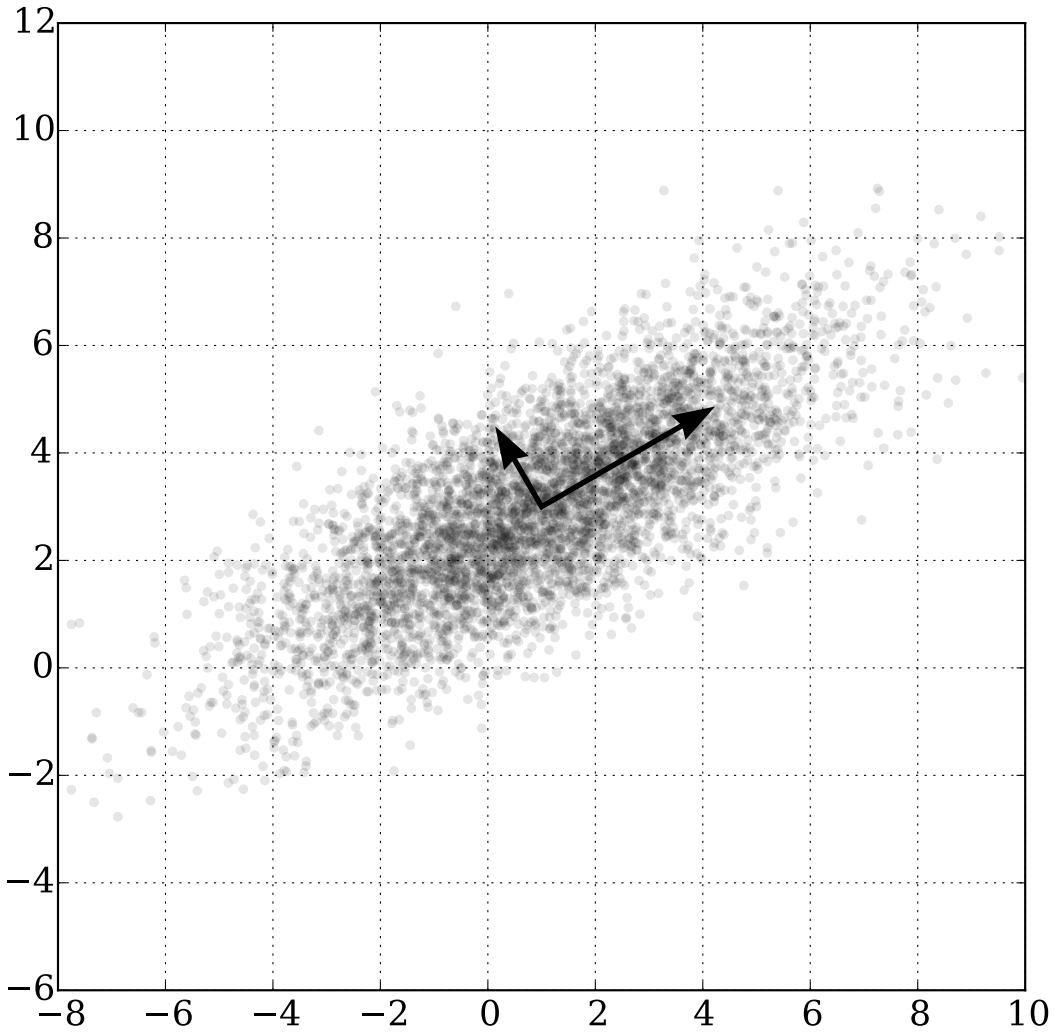


Figure 54: PCA visualization

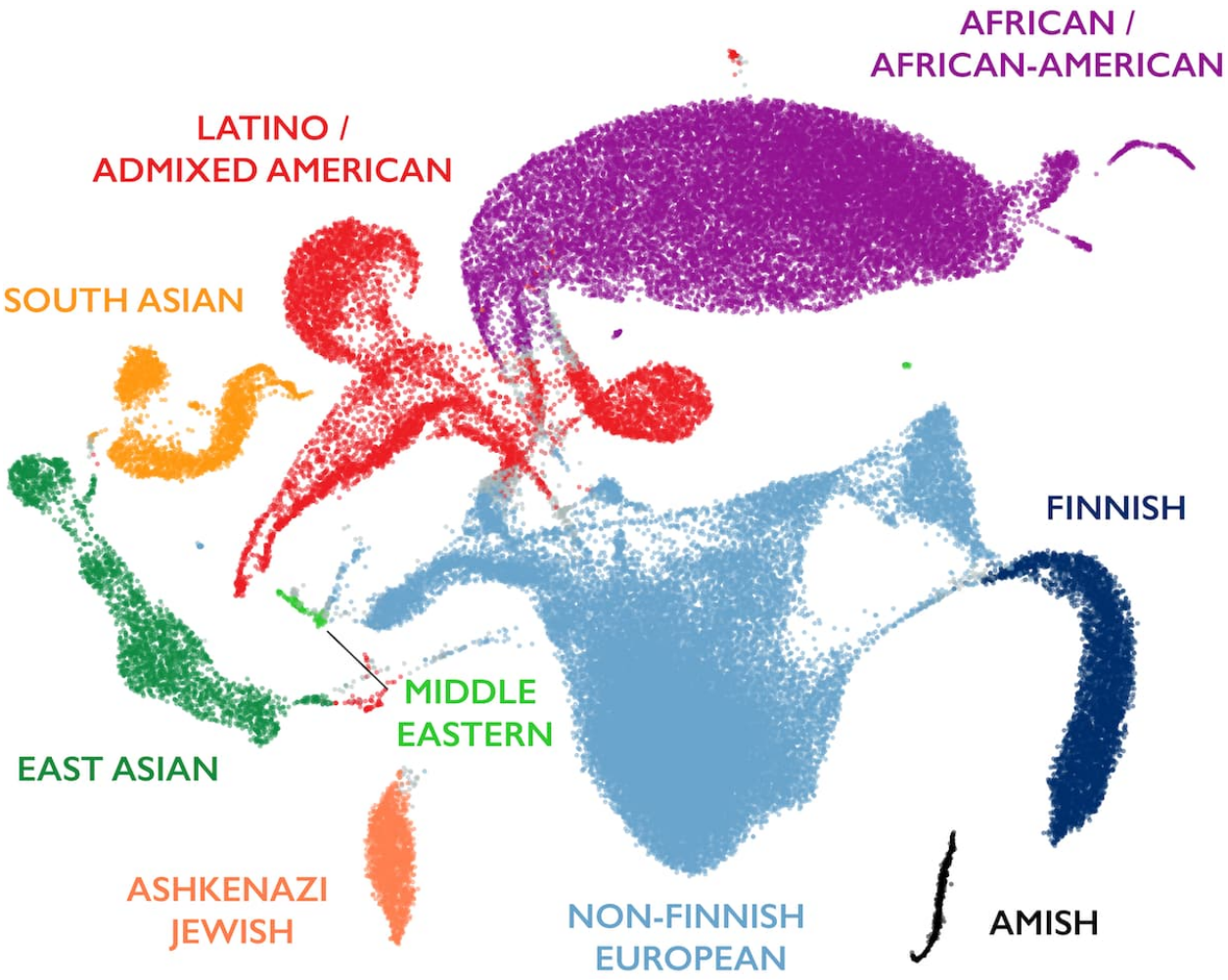


Figure 55: gnomAD UMAP

diverse because African populations are the oldest and have had more time to accumulate variation. The other clusters are tighter because they descend from small founder populations that left Africa.

27.4.5 Why Dimensionality Reduction Matters

PCA and UMAP make genetic relationships visible. These tools also let researchers correct for population structure in genetic studies. If you're doing a GWAS and you don't account for ancestry differences between cases and controls, you'll get false positives. By including PCs as covariates in your analysis, you adjust for ancestry-related allele frequency differences and focus on true disease associations.

They also demonstrate something philosophically important: **human genetic variation is continuous, not categorical**. The clusters aren't rigid boxes. They're fuzzy regions that blend into each other.

27.5 Allele Frequency Differences Among Populations

Let's examine specific examples of how evolution has shaped allele frequency differences.

27.5.1 Lactose Tolerance and Dairy Farming

Adult lactose tolerance represents one of the clearest examples of recent human evolution. The LCT gene produces lactase, which digests milk sugar. The regulatory variant rs4988235, located 14kb upstream of LCT, keeps lactase production active into adulthood. This variant reaches frequencies of 70–90% in Northern European populations but remains below 10% in East Asian populations, creating a stark frequency gradient (Bersaglieri et al. 2004, American Journal of Human Genetics). The allele shows one of the strongest signals of recent positive selection in the human genome, estimated to have been under selection for approximately 7,500 years in European populations where dairy farming was established (Tishkoff et al. 2007, Nature Genetics). Interestingly, lactose tolerance evolved independently in African pastoral populations through different genetic variants, demonstrating convergent evolution in response to similar selective pressures.

27.5.2 Malaria Resistance: Sickle Cell and Beyond

The sickle cell allele (HbS) in the HBB gene exemplifies balanced polymorphism. In sub-Saharan Africa, allele frequencies reach 10–20% in malaria-endemic regions but are nearly absent elsewhere (Piel et al. 2010, Nature Communications). Heterozygotes (HbAS) gain substantial protection against severe Plasmodium falciparum malaria while avoiding the severe disease seen in homozygotes (HbSS). Evidence shows that malaria continues to exert strong selective pressure maintaining HbS at high frequencies in affected regions (Gong et al. 2015, PNAS). The geographic distribution of HbS closely mirrors historical malaria endemicity, providing strong support for the malaria hypothesis of balanced selection. Multiple other hemoglobin variants (HbC, HbE) and G6PD deficiency alleles show similar patterns, demonstrating that malaria has been one of the strongest selective forces in recent human evolution.

27.5.3 High-Altitude Adaptation in Tibetans

Tibetan highlanders living above 4,000 meters face severe oxygen scarcity. Variants in EPAS1 (endothelial PAS domain protein 1, also called HIF-2α) and EGLN1 (involved in oxygen sensing) show the strongest selection signals in the Tibetan genome (Beall et al. 2010, PNAS; Yi et al. 2010, Science). These variants are associated with lower hemoglobin concentrations in Tibetans, avoiding the polycythemia (excess red blood cells) that causes altitude sickness in lowlanders. The EPAS1 haplotype is present in ~87% of Tibetans but only ~9% of Han Chinese. Remarkably, this haplotype was likely inherited from Denisovans (archaic hominins) through ancient admixture, representing adaptive introgression—the acquisition of beneficial alleles from another hominin species (Huerta-Sánchez et al. 2014, Nature). This demonstrates how population structure, admixture, and selection can interact to facilitate rapid adaptation.

27.5.4 Population-Specific Rare Variants

Many rare variants discovered in large-scale sequencing are population-specific, appearing in only one or a few individuals within a particular ancestry group. These variants are recent—they arose within the last few thousand years and haven't spread elsewhere because gene flow between populations remains limited. As datasets grow, we continue discovering more population-specific variants, enriching our understanding of recent human evolution and local adaptation.

27.6 Quantifying Genetic Differentiation: FST

How do we measure how different two populations are? One widely used statistic is **FST** (the Fixation Index). It quantifies how much allele frequencies differ between populations relative to total variation.

FST ranges from 0 to 1: - **FST = 0**: Populations have identical allele frequencies (no differentiation) - **FST = 1**: Populations are completely different (no shared variation)

For human populations, FST typically ranges from 0.05 to 0.15 between major continental groups. This means that about **85-95% of human genetic variation exists within populations, not between them**. If you randomly pick two people from the same population, they'll differ at millions of SNPs. If you pick two people from different continents, they'll differ at only slightly more SNPs.

This is a striking finding. It tells us that humans are genetically very similar despite visible diversity. The differences we perceive—skin color, facial features—are controlled by a tiny fraction of the genome. The vast majority of our DNA is shared.

FST also varies across the genome. Some regions have low FST (similar across populations), while others have high FST (more differentiated). High-FST regions often point to local adaptation—selection acting differently in different environments. For example, genes involved in skin pigmentation show high FST between African and European populations because skin color adapted to different levels of UV radiation.

27.7 Why Population Structure Matters

Understanding population structure isn't just academic. It has real implications for genetics research and medicine.

27.7.1 Correcting Confounding in Association Studies

This is the biggest practical concern. Imagine you're running a GWAS to find variants associated with type 2 diabetes. If your cases include more people of one ancestry than your controls, any variant that differs in frequency between ancestries will appear "associated" with diabetes—even if it has nothing to do with the disease. It's a false positive caused by population structure.

Researchers include principal components (PCs) as covariates in their analysis to adjust for ancestry-related allele frequency differences. Without this correction, GWAS results would be full of spurious findings.

27.7.2 Tracing Human History

Genetic patterns reveal migration routes, admixture events, and ancient population splits. By comparing allele frequencies and genomic segments, researchers can reconstruct who migrated where and when. Population structure is a record of human history written in DNA.

27.7.3 Improving Medical Genetics

Allele frequencies determine which variants are likely pathogenic versus benign. A variant that's common in healthy individuals of one ancestry is probably benign, even if it's rare overall. Knowing population-specific background frequencies prevents false clinical interpretations.

27.7.4 Ensuring Diversity in Research

Most early GWAS and sequencing studies focused on people of European ancestry. This creates two problems. First, polygenic scores and genetic risk predictions trained on European data don't work well in other populations. Second, it's inequitable—people of non-European ancestry don't benefit equally from genomic medicine. Expanding to multi-ancestry cohorts improves both accuracy and equity.

27.8 Summary

Population structure—systematic differences in allele frequencies among groups—emerges from human history: migration, isolation, bottlenecks, admixture, and local adaptation over tens of thousands of years.

27.8.1 Core Concepts

Distinct but overlapping concepts: Race is a social construct without biological precision. Ethnicity is cultural identity. Ancestry reflects genetic lineage probabilistically. Population describes groups with shared gene flow. Modern genetics focuses on ancestry and population, not race.

Multiple mechanisms create structure: Geographic isolation and drift cause divergence over time. Founder effects (Finnish Disease Heritage) and bottlenecks reduce diversity. Admixture creates mosaic genomes. Natural selection drives local adaptation (lactose tolerance in Europeans, high-altitude variants in Tibetans, malaria resistance in Africans).

Dimensionality reduction reveals patterns: PCA and UMAP compress millions of SNPs into visualizable dimensions, showing that human genetic variation forms continuous gradients, not discrete categories. The gnomAD UMAP demonstrates this—populations cluster but blend at edges.

Most variation is within, not between populations: FST between continental groups is typically 0.05–0.15, meaning 85–95% of genetic variation exists within any population. Humans are remarkably similar genetically despite visible diversity.

27.8.2 Why This Matters

Study design: PCs correct for ancestry-related confounding in GWAS, preventing false associations. Without correction, population structure creates spurious findings.

Clinical interpretation: Population-specific allele frequencies distinguish pathogenic from benign variants. A variant common in healthy individuals of one ancestry is likely benign overall.

Historical reconstruction: Allele frequency patterns record human migration, admixture, and adaptation—genetics as an archaeological record.

Equity in genomics: Multi-ancestry research ensures genomic medicine benefits everyone, not just European-ancestry populations. It also improves accuracy of risk prediction across populations.

27.8.3 Looking Ahead

This chapter showed how populations differ and why those differences arose. But variants don't just vary in frequency—they travel together on chromosomes. Physical proximity on a chromosome creates correlations between variants that persist until recombination breaks them apart.

The next chapter, **Linkage and Recombination**, will explore how genes refuse to assort independently when they're close together, how chromosome swaps during meiosis shuffle the genetic deck, and how these processes create haplotypes—blocks of variants inherited as units. We'll see how Morgan's century-old discovery of genetic linkage scales to modern genomics, revealing recombination hotspots and explaining why some gene pairs stay together through generations while others split apart.

Then we'll learn how all this information—allele frequencies, population structure, haplotypes—is encoded in the universal language of genomics: **VCF format**. Understanding VCF files is essential for working with population-scale data and interpreting genomic variation in research and medicine.

Population structure connects allele frequency (Chapter 20) to the physical architecture of chromosomes (Chapter 22) and the data formats that make modern genomics possible (Chapter 23). Together, these concepts form the foundation for understanding how genetic variation is organized, inherited, and distributed across humanity.

28 Chapter 22. From Mendel to Morgan: Discovery of Linkage

Remember Mendel's experiments with pea plants? He discovered two laws of inheritance. The first—segregation—says that each parent contributes one allele to offspring. The second—**independent assortment**—says that different gene pairs segregate independently during gamete formation.

Independent assortment is why, when you cross a plant that's round and yellow (RRYY) with one that's wrinkled and green (rryy), the F_2 generation shows a neat **9:3:3:1 ratio**. Nine are round and yellow, three are round and green, three are wrinkled and yellow, and one is wrinkled and green. The seed shape (R vs r) and color (Y vs y) seem to be inherited completely independently, as if they don't know about each other.

Mendel thought this was a universal rule. And for the traits he studied, it worked perfectly. But here's the catch: Mendel got lucky. The seven traits he picked in pea plants happen to be on different chromosomes (peas have seven chromosome pairs), so they really do assort independently. If he'd picked two traits on the same chromosome, he would have gotten very different results—and genetics history might have taken a different path.

Independent assortment happens because of how chromosomes behave during **meiosis I**. When your cells make eggs or sperm, homologous chromosome pairs line up at the cell's equator. Which chromosome of each pair goes to which side? It's random. Chromosome 1 from your mom might go left while chromosome 2 from your dad goes left. Or both might go right. The combinations shuffle randomly, so genes on different chromosomes assort independently.

But what about genes on the same chromosome? They should travel together, right? They're physically attached to the same DNA molecule. And that's exactly what scientists started discovering in the early 1900s. This discovery—that Mendel's second law isn't always true—opened up an entirely new understanding of how genes are organized.

28.1 Early Clues: Something Doesn't Add Up

In 1906, two British geneticists were studying sweet pea plants, looking at flower color and pollen shape—two different traits. If Mendel's law of independent assortment held true, they should have seen the classic 9:3:3:1 ratio in the F_2 generation.

But they didn't.

Instead, something strange emerged: certain trait combinations appeared together far more often than expected. When the purple flower allele was present, it tended to come with the long pollen allele. When

the red flower allele showed up, it usually brought the round pollen allele along with it. It was as if the alleles were traveling in pairs, sticking together through generations.

The researchers called this pattern **"coupling"**—traits inherited together. When traits separated more often than expected, they called it **"repulsion"**. They knew they'd found an exception to Mendel's law, but they didn't know why. They couldn't explain the mechanism. They just documented the pattern and called it peculiar (Punnett 1911, Journal of Genetics).

What had been discovered was evidence that some genes are **physically connected**. They're on the same chromosome, so they tend to be transmitted together to offspring. But in 1906, nobody understood chromosomes well enough to make that connection. The answer would have to wait a few more years.

28.2 Morgan's Fly Room: Where Genetics Got Physical

Enter a skeptical embryologist at Columbia University who wasn't entirely convinced that chromosomes had anything to do with heredity. To test the idea, breeding experiments with fruit flies—Drosophila melanogaster—began in milk bottles in a cramped room at Columbia that would become legendary as "the Fly Room."

In 1910, an extraordinary discovery emerged. Among thousands of normal red-eyed flies, one male had white eyes. This was a spontaneous mutation—a new variant that hadn't existed in the fly stock before. The value was immediately recognized, and crosses began: this white-eyed male was bred with normal red-eyed females.

The F_1 generation? All red-eyed. Red was dominant, just like Mendel would have predicted. But in the F_2 generation, something unexpected happened: white eyes appeared only among the **males**. About half the males had white eyes, while all the females had red eyes. This wasn't Mendel's 3:1 ratio. This was something tied to sex.

An explanation emerged based on sex chromosomes. Females have two X chromosomes (XX), while males have one X and one Y (XY). What if the eye-color gene was on the X chromosome? That would explain the pattern perfectly:

- The original white-eyed male had the white allele on his only X chromosome (X^w Y).
- F_1 females inherited one X with the white allele from dad and one X with the red allele from mom (X^w X^+). They looked red because red is dominant.
- But F_1 males inherited the red X from mom and the Y from dad (X^+ Y). They all looked red.
- In the F_2 generation, some males inherited the X^w from their F_1 mothers. Those males were white-eyed because they had no second X to mask it.

Figure: X-linked inheritance in Drosophila. The cross between white-eyed males (X^w Y) and red-eyed females (X^w+ X^w+) demonstrates sex-linked inheritance. In the P generation, white-eyed males are crossed with red-eyed females. The F_1 generation produces all red-eyed offspring (both females X^w+ X^w and males X^w+ Y), since red is dominant. When F_1 individuals are crossed, the F_2 generation shows a characteristic sex-linked pattern: red-eyed females (½ X^w+ X^w+ and ½ X^w+ X^w), red-eyed males (½ X^w+ Y), and white-eyed males (½ X^w Y). Notice that white eyes appear only in males in the F_2 generation, revealing that the eye color gene resides on the X chromosome. Source: Morgan, T. H. (1910). Sex-limited inheritance in Drosophila. Science, 32:120–122. License: Public domain (pre-1923 publication).

This experiment showed that a specific gene resides on a specific chromosome—the X. This was the first time anyone had connected Mendel's abstract "factors" to actual physical structures you could see under a microscope. Genes weren't just mathematical entities. They were real things, located on chromosomes (Morgan 1910, Science).

This discovery kicked off the era of **chromosomal genetics**. Genes had addresses. They lived on chromosomes. And if you knew which chromosome a gene was on, you could predict inheritance patterns based on how that chromosome behaved during meiosis.

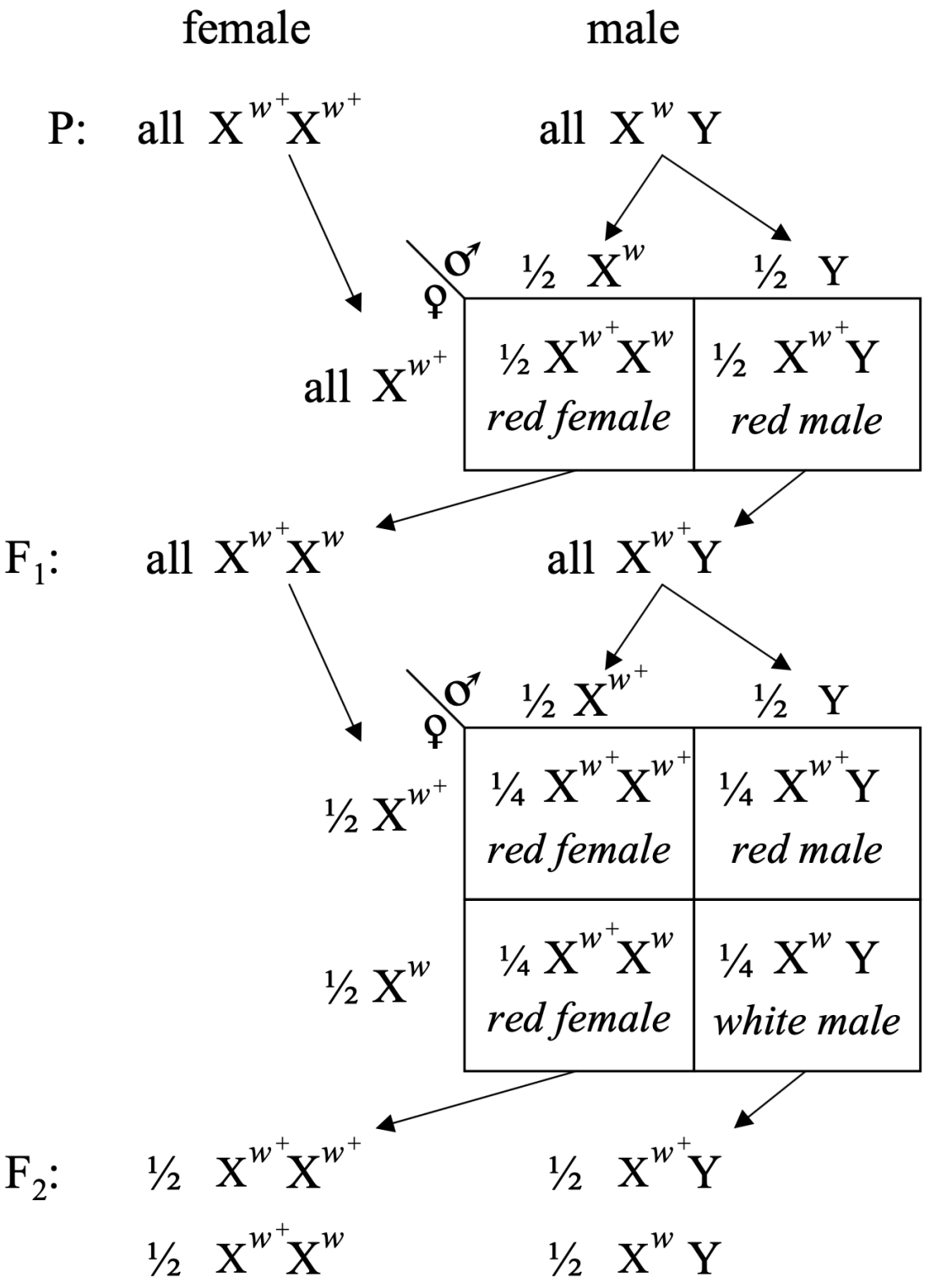


Figure 56: Morgan1910–Science–PostscriptDiagram.png

28.3 The Discovery of Linkage

The work didn't stop with one X-linked gene. More were found—genes affecting wing shape, body color, and other traits. And when flies with two X-linked traits at once were crossed, something fascinating emerged.

The traits didn't assort independently. They were linked.

For example, suppose you cross flies that differ in both eye color and wing shape, where both genes are on the X chromosome. If Mendel's law held, you'd expect all four possible combinations in equal frequencies: red eyes with normal wings, red eyes with miniature wings, white eyes with normal wings, white eyes with miniature wings.

But that's not what happened. Instead, **parental combinations appeared far more often**. If the original parents were "red eyes, normal wings" and "white eyes, miniature wings," then those same combinations dominated the offspring. Red with normal stayed together. White with miniature stayed together. Recombinant types—red with miniature, or white with normal—were rare.

The explanation was elegant. Genes on the same chromosome are **physically linked**—they're part of the same DNA molecule. During meiosis, when chromosomes line up and segregate, linked genes travel together. They're like passengers on the same bus. If two genes are on the same chromosome, they generally arrive at the same destination.

But occasionally—and here's the key insight—**crossing over** happens. During meiosis, homologous chromosomes pair up and can exchange DNA segments. It's like two trains running parallel, briefly touching, and swapping a few cars. When a crossover occurs between two linked genes, it breaks the linkage and creates recombinant offspring.

The closer two genes are on a chromosome, the less likely a crossover will land between them. If they're right next to each other, they almost always stay linked. If they're far apart, crossovers are more likely to separate them. This relationship—distance affects recombination—would become the foundation for genetic mapping (Morgan 1911, Journal of Experimental Zoology).

28.4 Sturtevant's Stroke of Genius: The First Genetic Map

One undergraduate student was listening to discussions about linkage and recombination frequencies when inspiration struck: **recombination frequency could measure distance**.

Think about it. If genes A and B are close together on a chromosome, crossovers between them are rare, so recombination frequency is low. If genes B and C are farther apart, crossovers between them are more common, so recombination frequency is higher. You could use recombination as a **genetic ruler**.

The realization emerged that by taking several genes on the same chromosome, measuring how often each pair recombined, and using those frequencies, the linear order of genes along the chromosome could be determined. It's like having several cities and knowing the distances between some pairs. You can reconstruct which cities are in between others by comparing distances.

That night, the problem was worked out. By morning, the first **genetic linkage map** had been constructed—a diagram showing the relative positions of six genes on the Drosophila X chromosome. The unit of genetic distance was defined: **1% recombination equals 1 map unit**, later called a **centiMorgan (cM)**.

Here's how it works. Suppose you're studying three genes: A, B, and C, all on the same chromosome. You cross flies and count recombinant offspring:

- **Genes A and B**: Recombination is found in 10% of offspring. That means they're 10 cM apart.
- **Genes B and C**: Recombination occurs in 5% of offspring. They're 5 cM apart.

- **Genes A and C**: If A and C are on opposite sides of B, they should be about 15 cM apart (10 + 5). And when you test it, that's roughly what you find.

So the gene order is A — B — C, with B in the middle. The linear arrangement matches the recombination distances perfectly.

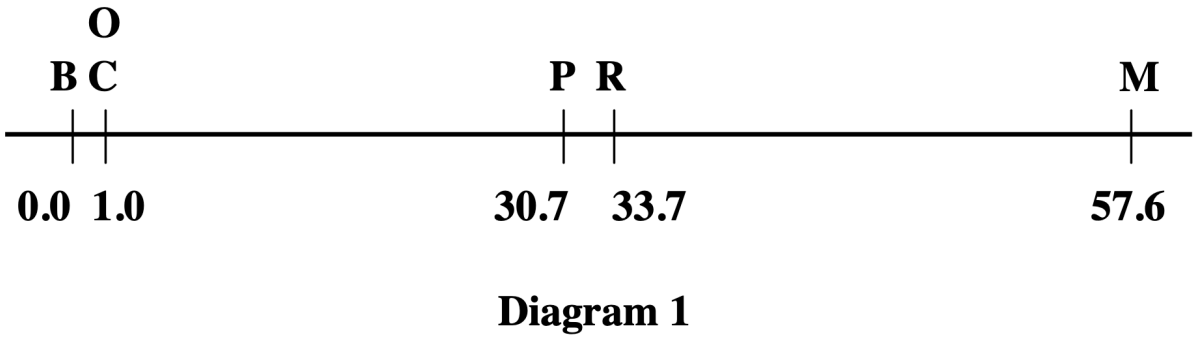


Figure 57: Sturtevant1913-JExpZool-Diagram1.png

Figure: The first genetic linkage map. This diagram shows the linear arrangement of six sex-linked genes on the Drosophila X chromosome, with distances measured in map units (now called centiMorgans). B = body color (yellow when recessive), C = color factor, O = eosin eye color, P = vermilion eye color, R = rudimentary wing, M = miniature wing. The numbers (0.0, 1.0, 30.7, 33.7, 57.6) represent genetic distances based on recombination frequencies—1% recombination equals 1 map unit. Note that C and O are positioned at the same location (1.0) because they showed complete linkage—no recombination was detected between them in thousands of flies. This map proved that genes are arranged linearly along chromosomes and that recombination frequency reflects physical distance. Source: Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology, 14:43-59. License: Public domain (pre-1923 publication).

This was monumental. It proved that genes are arranged linearly along chromosomes, like beads on a string. It showed that recombination frequency reflects physical distance. And it gave geneticists a tool to map genes even when they couldn't see them under a microscope. You didn't need to know what the genes did or where they were physically. You just needed to count recombinant offspring (Sturtevant 1913, Journal of Experimental Zoology).

28.5 How Crossing Over Works

Let's zoom in on what's actually happening during crossing over. This process occurs during **meiosis I**, when homologous chromosomes pair up.

Each chromosome has already replicated, so it consists of two sister chromatids joined at the centromere. When homologous chromosomes pair, they align side by side. At various points along their length, the chromosomes physically touch and form structures called **chiasmata** (singular: chiasma). At each chiasma, non-sister chromatids exchange DNA segments.

Think of it like two ribbons lying side by side. You cut both ribbons at the same position and swap the cut ends. Now each ribbon has a segment from the other. That's crossing over.

Figure: Crossing over during meiosis. Two homologous chromosomes (blue and red), each consisting of two sister chromatids, pair up during meiosis I. A chiasma forms where non-sister chromatids (one blue, one red) exchange segments. After the crossover, two chromatids remain parental (all blue or all red), while two become recombinant (part blue, part red). If genes are located on opposite sides of the

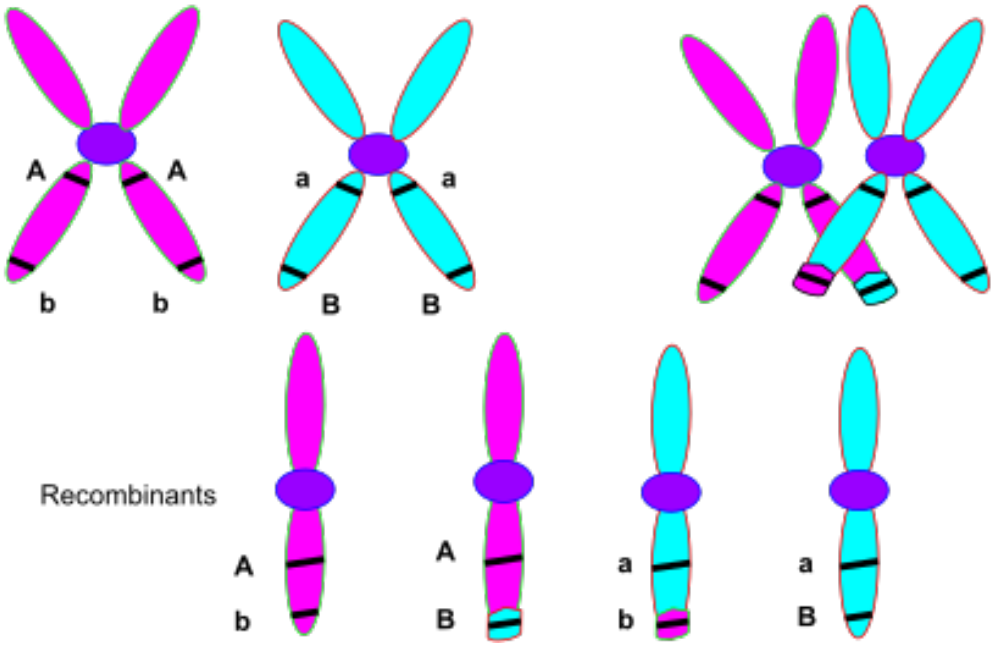


Figure 58: Figure. Chromosomal Crossover

crossover point, they'll be separated and end up in different gametes—that's what creates recombinant offspring. The farther apart two genes are, the more likely a crossover will occur between them. Source: Wikimedia Commons. License: CC BY-SA 3.0.

After crossing over, four gametes are produced. Two carry parental combinations (the original allele combinations from each parent), and two carry recombinant combinations (new mixtures of alleles). The proportion of recombinant gametes is what we measure as recombination frequency.

Genes that are very close together are rarely separated by crossovers—their recombination frequency approaches 0%, and they're said to be tightly linked. Genes that are far apart, or on different chromosomes, recombine about 50% of the time, and they assort independently—just like Mendel observed.

Each chromosome acts as a **linkage group**—a set of genes physically connected and usually inherited together. Humans have 23 linkage groups (22 autosomes plus the X), corresponding to our 23 chromosome pairs.

28.5.1 Double Crossing Over

Here's a complication: sometimes two crossovers occur in the same chromosome region. When this happens between three genes in a row (say A, B, and C), a crossover between A and B plus another between B and C can make it look like A and C never recombined at all—even though two exchanges actually happened.

Figure: Evidence for double crossing over. These tables show crosses involving three or four sex-linked genes simultaneously, revealing that double crossovers do occur but are rarer than expected. The diagrams illustrate no crossing over (parental combinations), single crossing over (one recombination event), and double crossing over (two recombination events). In Table 4, among 10,495 offspring, most showed no crossing over (6,972) or single crossing over (3,454), while only 69 showed double crossovers. Notably, when one crossover occurred, a second crossover in the same chromosome became less likely—a phenomenon called interference. For instance, where B and CO did not separate, the ratio for CO and R crossovers was about 1:2, but when B and CO did separate, it dropped to about 1:6.5. This interference explains why long-distance recombination frequencies are lower than predicted by simply adding

Table 4

No crossing over	Single crossing over		Double crossing over
B CO R	B CO R	B CO R	B CO R
6972	3454	60	9

Table 5

No crossing over	Single crossing over		Double crossing over
O P R	O P R	O P R	O P R
194	102	11	1

Table 6

No crossing over	Single crossing over		Double crossing over
B O M	B O M	B O M	B O M
278	160	1	0

Table 7

B O P R	B O P R	B O P R	B O P R	B O P R	B O P R	B O P R	B O P R
393	203	19	6	2	1	1	0

Figure 59: Sturtevant1913–JExpZool–Tables4–7.png

shorter distances. Source: Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology, 14:43–59. License: Public domain (pre–1923 publication).

The data revealed something else important: one crossover makes another nearby crossover less likely. This is called **interference**. It's as if the chromosome has a memory—once it's been cut and repaired at one spot, it resists being cut again nearby. This means that for genes far apart on a chromosome, the observed recombination frequency is actually less than you'd predict by adding up the distances between intermediate genes (Sturtevant 1913, Journal of Experimental Zoology).

28.6 Linkage in Humans

The same principles apply to humans, though it took longer to discover because you can't do controlled crosses with people (for obvious ethical reasons). Instead, geneticists had to study families and look for patterns.

In the 1930s, researchers noticed that **nail–patella syndrome** (a disorder affecting nails and kneecaps) and the **ABO blood group** were inherited together in some families more often than expected. This was one of the first demonstrations of linkage in humans. The genes turned out to be on the same chromosome (chromosome 9).

As more genetic markers became available—first blood groups, then protein variants, then microsatellites, and eventually SNPs—geneticists built detailed linkage maps of the human genome. These maps were essential for locating disease genes.

Here's how it works. Suppose a disease runs in a family. You don't know which gene causes it, but you have genetic markers scattered across the genome. You check whether any marker is inherited along with the disease more often than expected. If so, the disease gene is probably on the same chromosome, close to that marker. You've narrowed the search from the whole genome (3 billion base pairs) to a chromosomal region (maybe 10 million base pairs).

This approach led to the discovery of genes for **cystic fibrosis** (chromosome 7), **Huntington's disease** (chromosome 4), and many other disorders. Before the Human Genome Project, linkage mapping was one of the few ways to find disease genes.

The logic is always the same: **genes that are close together are rarely separated**. If a disease gene and a marker show low recombination (say, 1–5%), they're close. If they show high recombination (approaching 50%), they're far apart or on different chromosomes.

28.7 Putting It All Together

Let's review the key concepts using everyday language.

Linkage means that genes on the same chromosome tend to be inherited together. Think of them as neighbors traveling on the same bus. Unless they get off at different stops, they arrive together.

Crossing over is the exchange of DNA between homologous chromosomes during meiosis. It's like two ribbons swapping segments. This process can separate linked genes, creating new allele combinations.

Recombination frequency is the fraction of offspring that carry these new combinations. It's a genetic ruler: 1% recombination equals 1 cM of genetic distance. The farther apart two genes are, the more likely they'll be separated by crossing over, and the higher their recombination frequency.

A **linkage group** is all the genes on one chromosome. They're physically connected, so they form a group that tends to be inherited together. Humans have 23 linkage groups—one for each chromosome.

Here are a few examples to make this concrete:

Very close genes (recombination frequency near 0%): These are tightly linked. Crossovers between them are rare, so they almost always travel together. If you find one allele, you can predict the other with high confidence.

Distant genes (recombination frequency near 50%): These are on opposite ends of a long chromosome or on different chromosomes entirely. They recombine about half the time, so they assort independently—just like Mendel saw. You can't predict one from the other.

Intermediate distances (recombination frequency between 1–40%): These genes are on the same chromosome but separated by enough distance that crossovers sometimes land between them. The exact recombination frequency tells you how far apart they are.

28.8 From Flies to Genomes

The work in the Fly Room laid the groundwork for everything that came after. The discoveries showed that:

- Genes are physical entities located on chromosomes
- Genes on the same chromosome are linked
- Crossing over can break linkage
- Recombination frequency measures genetic distance
- You can build maps showing gene order and spacing

Today, we don't rely on crossing flies and counting recombinants. We sequence entire genomes and directly measure recombination by tracking which DNA segments are inherited together across thousands of people. But the underlying logic is exactly the same.

Modern recombination maps are built from millions of SNPs genotyped in large populations. We can see which alleles at different loci are inherited together (low recombination) versus independently (high recombination). We can build haplotype maps showing common blocks of linked variants. We can study **linkage disequilibrium**—the statistical association between alleles at different loci—which reflects both recombination history and population history.

All of this traces back to the white-eyed fly and the late-night insight about using recombination as a genetic ruler. What started as a curiosity—why don't these traits follow Mendel's law?—became the foundation for genetic mapping, genome assembly, and our understanding of how DNA is organized and transmitted.

Mendel gave us the laws of inheritance. The early 1900s showed us where those laws break down—and in doing so, revealed something even more interesting: the physical reality of genes and the dynamic reshuffling of genetic material through recombination.

29 Chapter 23. Recombination, Linkage, and Haplotype

Here's a question you've probably wondered about: why do siblings look different? You and your sister both got half your DNA from your mom and half from your dad, just like everyone else. So why don't all siblings look identical?

Part of the answer is which specific half you got. Your mom has two copies of each chromosome—one from her mother (your maternal grandmother) and one from her father (your maternal grandfather). When she makes an egg cell, she passes on one of those two copies at each chromosome. You might get mostly grandma's versions, while your sister might get mostly grandpa's versions. That alone creates variation.

But there's more to the story. Your mom doesn't pass on her mother's chromosome exactly as she received it. During the formation of egg cells—a process called **meiosis**—matching chromosomes pair up and actually **exchange DNA segments**. It's like shuffling a deck of cards, but with chromosomes. A piece of grandma's chromosome gets swapped with the matching piece from grandpa's chromosome. The result? The chromosome your mom passes to you is a mosaic—part grandma, part grandpa, mixed in new combinations.

This process is called **recombination**, and it's one of nature's most elegant mechanisms for generating diversity. Without recombination, you'd inherit entire chromosomes intact from each grandparent. With recombination, you inherit pieces from all four grandparents, shuffled into novel combinations. That's why you might have your maternal grandfather's nose but your maternal grandmother's eye color, even though those traits are on the same chromosome.

Recombination is also why genetic mapping works, why GWAS can find disease genes, and why we can trace ancestry. Understanding recombination is essential for understanding modern genomics.

29.1 Two Flavors of Recombination: Cross-Over and Non-Cross-Over

When geneticists first studied recombination, they focused on **cross-overs (COs)**—the dramatic events where large chunks of DNA get swapped between chromosomes. Cross-overs are relatively easy to detect because they create long stretches of DNA from one grandparent where you'd expect DNA from the other.

Think of cross-overs like this: imagine two ropes, one red and one blue, lying side by side. You cut both ropes at the same position and swap the cut ends. Now you have one rope that's red on the left and blue on the right, and another that's blue on the left and red on the right. That's a cross-over—a clean exchange of large DNA segments.

But in 2025, something remarkable was published. By analyzing whole-genome sequences from over **5,000 Icelandic families**, it was discovered that cross-overs are just part of the story. There's another type of recombination that's been hiding in plain sight: **non-cross-overs (NCOs)**, also called **gene conversion** (Palsson et al. 2025, Nature).

Non-cross-overs work differently. Instead of swapping large segments, a small stretch of DNA—typically just a few hundred to a few thousand base pairs—gets copied from one chromosome to the other. It's like patching a small section of the red rope with blue thread, but leaving everything else red. The overall structure doesn't change much, but locally, you've altered the sequence.

Here's what makes NCOs fascinating: **they're far more common than cross-overs**. For every cross-over that happens, there might be five or ten non-cross-overs. We just didn't realize it before because NCOs are small and hard to detect without very high-resolution sequencing data.

Both types of recombination start the same way: with a **double-strand break (DSB)**—a deliberate cut in both strands of the DNA double helix. Your cells make these breaks on purpose during meiosis. It sounds dangerous (and it is if it goes wrong), but it's a controlled process. The break is repaired using the matching chromosome as a template, and depending on how the repair proceeds, you get either a cross-over or a non-cross-over.

29.2 What the Study Revealed

The **first complete human recombination map** that includes both cross-overs and non-cross-overs was created. Previous maps only tracked COs, so they were missing most of the action. By sequencing entire families—parents, children, and often grandparents—it was possible to trace which DNA segments came from which grandparent and spot every place where the pattern switched (Palsson et al. 2025, Nature).

The findings are striking:

Non-cross-overs vastly outnumber cross-overs. On average, each meiosis produces about 40–60 cross-overs across all 23 chromosome pairs. But it also produces several hundred non-cross-overs. That's an order of magnitude more events, and they'd been largely invisible until now.

Mothers and fathers recombine differently. This has been known for cross-overs—maternal meiosis produces more COs than paternal meiosis. But the study found that mothers also produce **longer NCOs** than fathers. A typical maternal NCO might convert 1,000–2,000 base pairs, while a paternal NCO might convert only 300–500 base pairs. Why? We don't know yet, but it suggests that the molecular machinery works differently in eggs versus sperm.

Maternal age increases non-cross-overs. As women age, the number of NCOs in their eggs goes up, especially outside the programmed recombination hotspots. This is different from cross-overs, which don't increase much with age. The extra NCOs might be compensatory—a way to ensure that chromosomes pair correctly even when the normal CO machinery becomes less efficient with age.

Centromeres suppress cross-overs but allow non-cross-overs. Near the **centromere**—the pinched region in the middle of a chromosome where spindle fibers attach during cell division—cross-overs are rare. That makes sense: a CO near the centromere could disrupt chromosome segregation and cause errors. But NCOs happen readily in these regions. They might help chromosomes recognize each other and pair correctly without the risks of a full cross-over.

These patterns tell us that recombination isn't just random DNA shuffling. It's carefully regulated, with different mechanisms operating in different regions and at different life stages.

29.3 How CO and NCO Form: The Molecular Dance

Let's look at what happens at the molecular level during recombination. It starts with a **double-strand break**—both strands of the DNA double helix are cut. This happens on one of the two paired chromosomes, let's say the red one.

The broken DNA end is processed to create a 3' overhang—a single-stranded tail. This tail then "invades" the matching region on the blue chromosome. It's like the red chromosome reaching over, finding the corresponding sequence on blue, and using it as a template.

Now the cell has a choice. How does it resolve this structure?

Path 1: Synthesis-Dependent Strand Annealing (SDSA) leads to a non-cross-over. The invading strand copies some information from the blue chromosome, then pulls back and re-anneals to its original partner. The result? A small patch of DNA has been converted from the red sequence to the blue sequence, but the chromosomes don't swap large segments. It's a local, short-range change.

Path 2: Double Holliday Junction (dHJ) formation leads to a cross-over. Instead of pulling back, the invading strand stays attached, and DNA synthesis extends the connection. This creates two junctions where the chromosomes are physically joined. When these junctions are resolved—cut and rejoined—you get a cross-over. Large flanking regions get swapped.

Why does the cell sometimes choose one path and sometimes the other? That's one of the big mysteries in meiosis research. There seem to be proteins that promote one outcome versus the other, and the local chromatin context (how tightly the DNA is packaged) probably matters too.

Figure: From DNA breaks to recombination maps—how the study detected and analyzed recombination events. Panel (a) illustrates the molecular mechanism of meiotic recombination. Red and blue represent the two homologous chromosomes (one from each grandparent). The process begins when a double-strand break (DSB) is deliberately made in one chromosome (red). The broken end is processed to create a single-stranded tail, which then invades the partner chromosome (blue) to search for the matching DNA sequence. From this point, two different pathways can occur. In the top pathway, a second strand also invades, forming a structure called a double Holliday junction (dHJ), which can be resolved to produce either

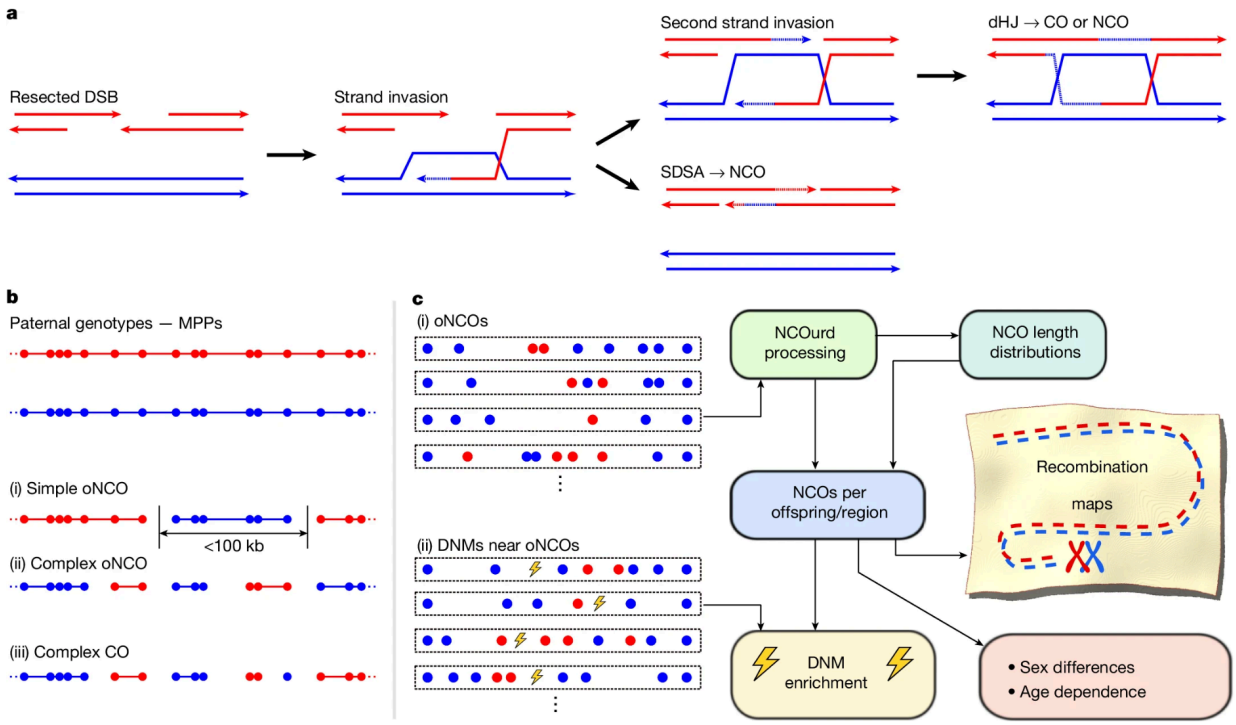


Figure 60: Palsson et al. 2025, Nature – Figure 1

a cross-over (CO) or a non-cross-over (NCO). In the bottom pathway, called synthesis-dependent strand annealing (SDSA), the invading strand copies a small segment of information from the blue chromosome, then pulls back and rejoins its original red partner—this always produces an NCO. Panel (b) shows how recombination events were detected in real family data. The top shows parental genotypes marked as red and blue dots (representing DNA variants from each grandparent). Below are examples of detected events: (i) a simple NCO where a short stretch switches from red to blue, (ii) a complex NCO with multiple switches in a small region, and (iii) a complex CO involving larger-scale exchanges. Panel (c) outlines the computational pipeline used to identify and characterize these events across thousands of families, ultimately producing complete recombination maps and revealing patterns like sex differences and maternal age effects. Source: Palsson, B. et al. (2025). Meiotic recombination shapes the landscape of genome variation. Nature. https://www.nature.com/articles/s41586-024-08450-5. License: CC-BY-NC-ND 4.0.

29.4 Linkage: Alleles That Travel Together

Now let's connect recombination back to the concept of linkage we learned about in the previous section. **Linkage** describes the tendency of alleles located close together on the same chromosome to be inherited together.

Why does this happen? Because recombination—whether CO or NCO—is a random process with respect to position (though it's biased toward hotspots, as we'll see). The closer two genetic variants are, the less likely that a recombination event will occur between them. And if no recombination separates them, they "travel as a pair" from parent to child.

We quantify linkage using the **recombination fraction (r)**. This is the probability that two loci will be separated by recombination in a single meiosis:

- **r = 0** means the loci are completely linked. They're so close together that recombination never

separates them, so they always travel together.
- **r = 0.5** means the loci are unlinked. They recombine half the time, which is what you'd expect for loci on different chromosomes or very far apart on the same chromosome. This is independent assortment—Mendel's second law.
- **r = 0.1** means the loci recombine 10% of the time. They're close enough to show linkage, but far enough apart that occasional recombination separates them. This corresponds to about 10 cM of genetic distance.

The smaller the recombination rate, the stronger the linkage. Two SNPs that are only 1 kb apart might have r = 0.0001—they're almost never separated. Two SNPs that are 10 Mb apart might have r = 0.15—they're separated fairly often.

Linkage is why genetic mapping works. If a disease gene is located near a particular DNA marker, that marker will often be inherited along with the disease allele. By tracking which markers co-segregate with the disease in families, you can narrow down where the disease gene must be. This principle underlies **linkage analysis** for Mendelian diseases and, in a modified form, **genome-wide association studies (GWAS)** for complex traits.

The study improved our understanding of linkage in two ways. First, by including NCOs in the recombination map, it captured the full picture of how often alleles get separated. NCOs can break linkage too, even if they don't swap large segments. Second, it showed that recombination rates vary by sex and age, which affects how strong linkage is depending on whether you're tracing maternal or paternal transmission (Palsson et al. 2025, Nature).

29.5 Regional Differences: Not All Chromosomes Recombine Equally

If recombination were evenly distributed across the genome, genetic mapping would be simple. But it's not. Some regions are **hotspots** where recombination happens frequently. Others are **coldspots** where it's rare. And some regions, like centromeres, have their own special rules.

Let's look at the patterns:

Telomeres (chromosome ends) are rich in cross-overs. When you plot CO frequency along a chromosome, you see peaks near the telomeres. This makes sense: COs near the ends don't disrupt the chromosome's structure much, and they help ensure proper segregation during meiosis. If chromosomes don't have at least one CO, they might fail to separate properly, leading to aneuploidy (wrong number of chromosomes).

Centromeres (chromosome centers) suppress cross-overs but allow non-cross-overs. A CO near the centromere could interfere with how spindle fibers attach and pull chromosomes apart. So evolution has selected against COs in these regions. But NCOs still happen, and they might serve a function—perhaps helping the chromosomes recognize each other and pair up during early meiosis.

Hotspots are scattered across the genome. These are short (1–2 kb) regions where recombination—both CO and NCO—occurs much more frequently than average. Hotspots are marked by a DNA-binding protein called PRDM9, which recognizes specific sequence motifs and recruits the machinery that makes double-strand breaks. Most recombination events cluster at these hotspots.

Interestingly, hotspot locations vary between populations. A hotspot active in Europeans might be inactive in East Asians, and vice versa. This is because PRDM9 itself is highly variable—different versions recognize different DNA motifs. Over evolutionary time, hotspots can move around the genome as PRDM9 variants spread through populations.

Figure: Recombination shows clear patterns: it differs between moms and dads, varies across the chromosome, and changes with maternal age. This figure uses chromosome 19 to illustrate key patterns in recombination. Panel (a) shows how frequently non-cross-overs (NCOs) occur at different positions along the chromosome. Notice that paternal NCOs (blue line) are much more frequent near the chromosome

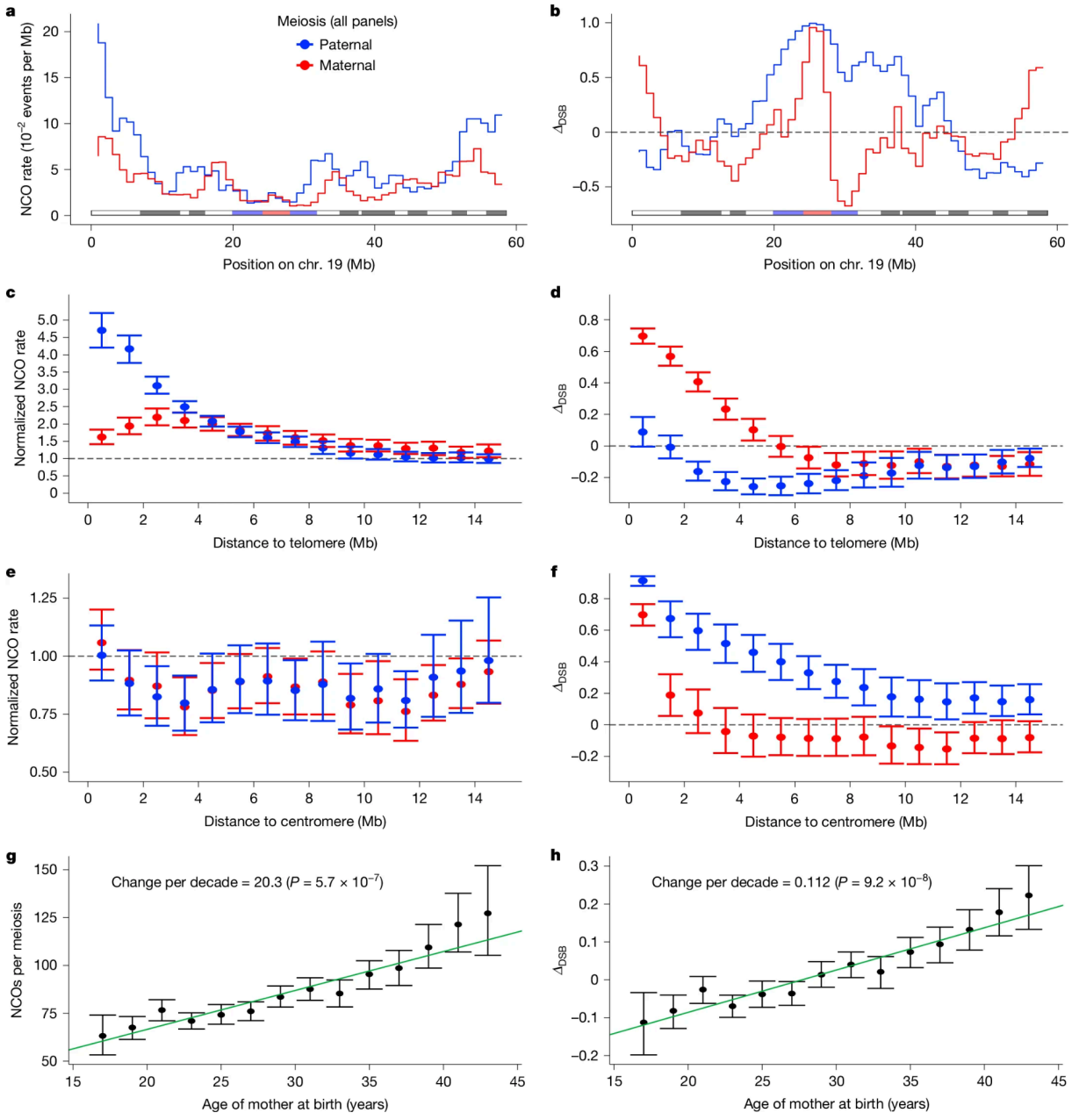


Figure 61: Palsson et al. 2025, Nature – Figure 2

ends, while maternal NCOs (red line) show lower overall rates. The gray bar marks the centromere location. Panel (b) displays a metric called ⊿csB that indicates whether recombination events resemble cross-overs (positive values) or non-cross-overs (negative values). Maternal meiosis tends to produce more CO-like events (red line above zero), while paternal meiosis produces more NCO-like events (blue line below zero). Panels (c-d) zoom in on regions near telomeres (chromosome ends): paternal meiosis produces many more NCOs here than maternal meiosis, and this sex difference is especially pronounced. Panels (e-f) examine regions near the centromere (chromosome center): here the NCO rates are more similar between sexes, though the CO/NCO balance still differs. Panels (g-h) reveal a striking maternal age effect: as mothers get older, they produce significantly more NCOs per meiosis (about 20 additional NCOs per decade of age), and their recombination pattern shifts to become more CO-like. These findings demonstrate that recombination is tightly regulated and influenced by sex, chromosomal location, and age. Source: Palsson, B. et al. (2025). Meiotic recombination shapes the landscape of genome variation. Nature. https://www.nature.com/articles/s41586-024-08450-5. License: CC-BY-NC-ND 4.0.

The study also revealed age-related changes. As women age, their eggs accumulate extra NCOs, particularly outside the normal hotspots. This suggests that when the programmed hotspot machinery starts to fail (which happens with aging), a backup mechanism kicks in to ensure some recombination still happens. It's not perfect—these extra NCOs might contribute to age-related increases in chromosomal abnormalities—but it's better than nothing (Palsson et al. 2025, Nature).

29.6 Haplotype: Blocks of Inherited DNA

Now we come to a concept that ties everything together: **haplotype**. A haplotype is a set of DNA variants (alleles) that are inherited together from a single parent. Because recombination only shuffles chromosomes occasionally, nearby variants tend to stay linked for many generations, forming "blocks."

Let's use an analogy. Imagine a chromosome as a beaded necklace where each bead represents a DNA variant (a SNP, for example). Your maternal grandmother has one necklace, and your maternal grandfather has another. The beads on each necklace are in a specific order—grandma's might be red-blue-blue-green-red, while grandpa's is blue-red-green-green-blue.

When your mother inherits one of these necklaces from each parent, she doesn't get them intact. During meiosis, pieces of the necklaces get swapped. Maybe the first three beads come from grandma (red-blue-blue), then a recombination breakpoint occurs, and the rest come from grandpa (green-blue). Now your mom's necklace is a mosaic: red-blue-blue-green-blue. That specific combination—that pattern of beads—is a haplotype.

When your mother passes her chromosome to you, you inherit that haplotype. If there's no recombination in your mother's meiosis, you get the whole mosaic intact. If there is recombination, you get a new mosaic with its own pattern.

Haplotypes are incredibly useful for tracing inheritance. By looking at which haplotype blocks a child inherited, geneticists can figure out where recombination happened. Here's how:

Suppose you sequence a child and both parents and grandparents. You look at a region of chromosome 1. For a stretch of 5 Mb, the child's sequence matches the maternal grandmother's pattern. Then suddenly, at position 147,500,000, it switches to match the maternal grandfather's pattern. That switch point is a **recombination breakpoint**.

If the switch is large—the child has grandma's pattern for millions of base pairs on one side and grandpa's pattern for millions on the other side—it's likely a cross-over. If the switch is small—just a few thousand base pairs change before switching back—it's likely a non-cross-over (gene conversion).

Exactly this logic was used to identify tens of thousands of CO and NCO events across Icelandic families. By comparing children to all four grandparents, every place where haplotypes changed could be spotted and classified as CO or NCO based on the size of the changed region (Palsson et al. 2025, Nature).

29.7 Why Recombination Matters

Recombination isn't just an abstract molecular process. It has real consequences for evolution, health, and medicine.

Genetic diversity. Recombination is one of the main engines of diversity in sexually reproducing organisms. Without it, each chromosome would be passed down intact through generations. Beneficial alleles on one chromosome and beneficial alleles on another could never be combined in the same individual. With recombination, new combinations form every generation, giving evolution more raw material to work with.

Disease genetics. Understanding recombination helps us find disease genes. If we know where recombination hotspots are, we can interpret linkage patterns more accurately. If we know that NCOs can break linkage too, we'll build better maps. The complete recombination map improves the resolution of genetic studies, helping researchers pinpoint causal variants more precisely.

Reproductive health. Errors in recombination cause problems. If cross-overs don't happen where they should, chromosomes might not segregate properly, leading to aneuploidy—an egg or sperm with the wrong number of chromosomes. This is a major cause of miscarriage and disorders like Down syndrome (trisomy 21). Understanding how recombination is regulated, and how it changes with age, could lead to interventions that improve fertility outcomes.

Human evolution. Haplotype blocks carry signatures of past evolutionary events. Long haplotype blocks suggest recent selection or recent admixture. Short blocks suggest old, well-mixed populations. By analyzing haplotype structure across populations, researchers can reconstruct migration patterns, population splits, and adaptive sweeps. Recombination is what shapes these blocks, so understanding its rate and distribution is essential for interpreting human history.

29.8 Putting It All Together

Let's tie together the three concepts in this chapter: recombination, linkage, and haplotypes.

Recombination is the process that breaks and reshapes haplotypes. It happens during meiosis when homologous chromosomes exchange DNA. There are two types: cross-overs, which swap large segments, and non-cross-overs, which copy small patches. Non-cross-overs are much more common than we used to think. Maternal recombination differs from paternal recombination, and maternal age increases NCO frequency.

Linkage describes how tightly alleles stay together before being separated by recombination. The closer two variants are on a chromosome, the lower their recombination rate, and the stronger their linkage. Linkage is quantified by the recombination fraction (r), which ranges from 0 (complete linkage) to 0.5 (independent assortment). Linkage is the basis for genetic mapping and GWAS. The complete recombination map, which includes both CO and NCO, gives us the most accurate picture yet of linkage across the human genome.

Haplotypes are blocks of alleles inherited together from one parent. They're the footprints of past recombination events. Long haplotype blocks indicate low recombination or recent origin. Short blocks indicate high recombination or old age. By tracing haplotypes through families, we can detect every recombination event and classify it as CO or NCO. Haplotypes also carry information about population history, natural selection, and disease risk.

These three concepts are interconnected. Recombination breaks haplotypes, creating new allele combinations. Linkage describes how often alleles are separated by recombination. Haplotypes preserve the outcome of past recombination events, recording where chromosomes were shuffled and where they stayed intact.

Together, they explain a fundamental paradox: how our genomes both preserve ancestry (through linked haplotypes) and generate diversity (through recombination). Every child is a mosaic of their four grandparents, with the seams between mosaic pieces marking where recombination occurred. Understanding these seams—where they happen, how often, and why—is essential for understanding genetics, evolution, and human health.

30 Chapter 24. Data Type for Alleles and Population

Here's a question that might not seem obvious at first: how do you turn a genome into data?

When you sequence someone's DNA, you get billions of short reads—fragments of sequence, each maybe 100–150 bases long. Those reads need to be assembled, aligned to a reference genome, and compared to figure out where this person's DNA differs from the reference. Once you've identified all those differences—the SNPs, the small insertions and deletions, maybe some larger variants—how do you write them down? How do you store them in a file that computers can read, that other scientists can use, that databases can search?

The answer is a standardized file format called **VCF (Variant Call Format)**. It's the universal language for describing genetic variation. Every variant caller—the software that finds differences between your sequence and the reference—outputs VCF files. Every genetic database stores data in VCF or a closely related format. Every analysis tool that works with variants expects VCF as input.

Understanding VCF is essential for modern genomics. It's how we turn biological variation—the actual differences in DNA molecules—into structured data that can be analyzed, compared, and shared across the world.

30.1 From Sequencing to Variant Calling

The path from DNA sample to variant list goes through several steps. You extract DNA and sequence it using a technology like Illumina, obtaining millions or billions of short reads—raw sequence fragments.

Next, you align those reads to a **reference genome**. For humans, that's usually GRCh38, the current standard reference assembly. Alignment means figuring out where each read came from in the genome using software like BWA or Bowtie2. The output is a BAM file (or its more compressed version, CRAM)—essentially a sorted list of all your reads and where they map.

Now comes **variant calling**. You compare the aligned reads to the reference genome and identify positions where your sample differs. If the reference has an "A" at position 12345 on chromosome 1, but your sample's reads consistently show a "T" instead, you've found a variant—an A→T substitution.

Variant calling software (like GATK, DeepVariant, or FreeBayes) scans through every position in your BAM file, looks at the reads covering that position, and decides: is this a variant or not? It considers read quality, mapping quality, strand bias, and other factors. The output is a VCF file—a text file listing every detected variant.

30.2 VCF File Structure

A VCF file is surprisingly simple in concept. It's a plain text file with one line per variant. Each line describes where the variant is, what the reference allele is, what the alternate allele is, and how confident we are about it.

30.2.1 Core Columns

CHROM tells you which chromosome (chr1, chrX, chrM, etc.)

POS is the genomic position—the exact base pair coordinate where the variant occurs (1-based coordinate system)

ID is an optional identifier. If catalogued in dbSNP, it gets an "rs" number like rs12345678. If novel, this field is just a dot (.)

REF is the reference allele—what the reference genome has at this position

ALT is the alternate allele—what your sample has instead

QUAL is a quality score. Higher numbers mean more confidence. QUAL of 30 means roughly 99.9% confidence; 10 means 90% confidence

FILTER indicates whether the variant passed quality filters. "PASS" is good—the variant looks legitimate. Other values like "LowQual" or "StrandBias" suggest potential problems

INFO holds metadata about the variant: read support, allele frequency, functional annotations, and more. It's a semicolon-separated list of key=value pairs

FORMAT describes the structure of the per-sample data that follows—a colon-separated list of field names like "GT:DP:AD"

Sample columns come after FORMAT. One person = one column; 1,000 people = 1,000 columns. Each gives that person's genotype and related metrics at this variant site

Here's a concrete example:

```
chr1  1234567  rs12345  A  T  99.0  PASS  AC=3;AN=6;AF=0.5  GT:DP  0/1:32
```

This says: On chromosome 1 at position 1,234,567, the reference genome has an A, but we've found a T variant (rs12345). We're very confident (quality 99). It passed filters. There are 3 alternate alleles out of 6 total alleles (50% frequency). The sample shown has genotype 0/1 (heterozygous) with 32 reads covering this site.

30.3 The INFO Field: Population-Level Statistics

The INFO field is where population-level statistics live.

AC (Allele Count) tells you how many copies of the alternate allele were found across all samples. If you sequenced 100 people (200 alleles, since humans are diploid), and 10 of those alleles are the alternate, AC=10.

AN (Allele Number) is the total number of alleles successfully called at this site. Ideally, this is 2 × number of samples. Sometimes sequencing quality is poor for some individuals at some sites, so those samples get marked as missing (./.), and AN is reduced.

AF (Allele Frequency) is simply AC ÷ AN—the proportion of alternate alleles. If AC=10 and AN=200, then AF=0.05 (5%).

These three fields are fundamental for population genetics. They tell you how common a variant is, which helps determine whether it's likely pathogenic (rare variants in disease genes often are) or benign (common variants usually are).

30.4 The FORMAT Field and Genotypes

After the INFO field comes FORMAT, which defines what information appears for each sample. The most common FORMAT fields are:

GT (Genotype) is the individual's allele combination, coded as numbers separated by "/" (unphased) or "|" (phased). The numbers refer to alleles: 0 = reference allele, 1 = first alternate, 2 = second alternate (if multiple), etc.

30.4.1 Genotype Codes

Genotype	Type	Phasing	Meaning
`0/0`	Homozygous reference	Unphased	Both alleles match the reference—you don't have the variant
`0/1`	Heterozygous	Unphased	One reference allele and one alternate allele—you're a carrier
`1/0`	Heterozygous	Unphased	Same as 0/1 (order doesn't matter when unphased)
`1/1`	Homozygous alternate	Unphased	Both alleles carry the variant
`./.`	Missing	Unphased	The genotype couldn't be determined, usually because sequencing coverage was too low or read quality was poor

You can use `0|1` for heterozygous variants or `0|0`, `1/1` for homozygous genotypes but it has another: – **/** (slash) = unphased genotypes. We know which alleles are present but not which came from mom vs. dad – **|** (pipe) = phased genotypes. We know which allele came from which parent (or at least which alleles are on the same chromosome)

If there are multiple alternate alleles (e.g., REF=A, ALT=T,G), you might see 0/2 (reference + second alternate), 1/2 (first alternate + second alternate), or 2/2 (both second alternate).

DP (Depth) is read depth—how many sequencing reads covered this position for this sample. Higher is generally better. DP=5 might not be trustworthy; DP=30 or higher provides more confidence.

AD (Allelic Depth) breaks down the read depth by allele. If DP=30 and AD=15,15, that means 15 reads support the reference and 15 support the alternate—a balanced heterozygote consistent with 0/1. If AD=30,0, all reads support the reference (consistent with 0/0). If AD=0,30, all support the alternate (consistent with 1/1).

Example:

```
chr1  1003215  .  A  T  99  PASS  .  GT:DP:AD  0/1:32:16,16
```

This sample is heterozygous (0/1), with 32 total reads, split evenly between the A reference allele (16 reads) and the T alternate allele (16 reads). That balanced ratio supports the heterozygous call.

30.5 Multi-Sample VCF: Population Data

The real power of VCF comes when analyzing multiple samples together in a **multi-sample VCF (msVCF)**. Each variant site is one row, and each sample is one column. The INFO field summarizes population-level statistics, while each sample column gives that individual's genotype.

Imagine we've sequenced three people—Alice, Bob, and Carol. At position chr1:1234567, the reference is A and we've found a T variant:

```
#CHROM  POS      ID       REF  ALT  QUAL  FILTER  INFO              FORMAT  Alice  Bob  Carol
chr1    1234567  rs12345  A    T    99    PASS    AC=3;AN=6;AF=0.5  GT      0/1    1/1  0/0
```

- Alice: **0/1** (heterozygous) = one A, one T
- Bob: **1/1** (homozygous alternate) = both alleles are T
- Carol: **0/0** (homozygous reference) = both alleles are A

Population statistics in INFO: – **AC=3**: Alice contributes 1 T allele, Bob contributes 2 T alleles, Carol contributes 0. Total: 3 T alleles – **AN=6**: Three people × 2 alleles each = 6 total alleles – **AF=0.5**: 3 ÷ 6 = 50% allele frequency

The key insight: **variant** and **genotype** are different concepts. A **variant** is a site where an alternate allele exists in the population—"there's a T at this position in some people." A **genotype** is an individual's allele combination—"Alice has one A and one T."

One variant, many genotypes. The variant is population-level. The genotypes are individual-level.

30.6 Joint Genotyping: Complete Data

When analyzing multiple samples, you could call variants on each sample independently and then merge the VCFs. But that creates a problem: sites that are variant in one person but reference in another might not be recorded consistently, leading to missing data and inaccurate allele frequencies.

Joint genotyping solves this by analyzing all samples together, simultaneously. The variant caller looks at reads from everyone at once and makes genotype calls for every sample at every site—even sites where most people are homozygous reference.

Think of it like taking attendance. If you call out names individually and only write down who answered "present," you won't know who's absent. But if you have a roster and mark everyone as either present or absent, you have complete information.

Joint genotyping ensures that every site in the VCF has a genotype entry for every sample. If Alice is 0/1, Bob is 1/1, and Carol is 0/0, all three genotypes appear in the file. That way, when you calculate AC and AN, you're using complete data.

Without joint genotyping, Carol's 0/0 genotype might not be recorded, leading to incorrect allele frequency calculations. With joint genotyping, Carol's 0/0 is explicit, so you correctly calculate AF=0.5 (3 out of 6 alleles).

Joint genotyping is standard practice for population-scale sequencing projects like gnomAD, UK Biobank, and All of Us.

30.7 Storage Strategies: Dense vs. Sparse

As sequencing projects grow to hundreds of thousands of genomes, VCF files become enormous. The UK Biobank has sequenced about 500,000 people. If you store every variant site for every person, that's billions of lines, each with 500,000 genotype columns. The file size is staggering.

Most of those genotypes are 0/0—homozygous reference. For rare variants (which are the majority), maybe 499,990 people are 0/0 and only 10 are 0/1 or 1/1. Storing 499,990 "0/0" entries is wasteful.

30.7.1 Dense vs. Sparse VCF Comparison

Feature	Dense VCF	Sparse VCF
Storage approach	Stores every genotype for every sample at every site	Stores only non-reference genotypes (0/1, 1/1, etc.)
0/0 genotypes	Explicitly written	Omitted (inferred)
File size	Very large, grows superlinearly with sample count	Much smaller, more efficient storage
Best for	Small projects (tens to hundreds of samples)	Large-scale projects (thousands to millions of samples)
Population statistics	Stored in INFO field	Stored in INFO field (no loss of information)
Example	500,000 samples with rare variant = 500,000 genotype entries	500,000 samples with rare variant = ~10–100 genotype entries
Redundancy	High (most entries are 0/0 for rare variants)	Low (only variant carriers recorded)
Computational feasibility	Impractical at biobank scale	Enables large-scale genomics

Sparse formats drastically reduce file size. A variant present in 10 people out of 500,000 only needs 10 genotype entries instead of 500,000. Population statistics (AC, AN, AF) are still stored in the INFO field, so you don't lose information—you just avoid redundancy.

Modern biobanks use sparse formats. GATK has gVCF (genomic VCF) that stores both variant and non-variant sites for each individual, which can then be combined into a sparse multi-sample VCF. BCFtools has sparse VCF options. These formats make large-scale genomics computationally feasible.

30.8 The Future of VCF: Scaling for a Genomic Revolution

The explosive growth of genetic variation data has created unprecedented challenges for VCF. National initiatives like Europe's 1+ Million Genomes Initiative, the US All of Us Research Program, and Australian Genomics are generating data from millions of individuals. This ballooning of data means that ways of representing genetic variation must also scale.

30.8.1 The Scaling Challenge

VCF faces a fundamental scaling problem: when new samples are added to a multi-sample VCF, you need to add rows for newly-observed variants, but you also need to update all existing rows to indicate that old samples did not contain the new variants. These updates expand the data held by all other samples, making file size grow superlinearly—faster than just linearly with sample count.

For example, exporting data from gnomAD (the Genome Aggregation Database) to VCF is in the petabyte range. Doubling the size of gnomAD would make export and interchange in VCF format completely infeasible. This isn't just an inconvenience—it makes workflows slower and sharing files more time-consuming and resource-expensive.

30.8.2 GA4GH's Approach to Scaling

The Global Alliance for Genomics and Health (GA4GH) maintains the VCF standard through its Large-Scale Genomics Work Stream. The Future of VCF Working Group, formed in 2019, meets monthly to discuss

VCF scaling challenges and review different approaches that have been developed to solve them. The group's main goal is to assess scaling solutions, make standards recommendations, and ensure that VCF keeps up with increasing data size and complexity.

Scaling VCF is important to maintain interoperability with other GA4GH standards, including htsget (for downloading variation data from specific genomic regions), Beacon (for querying genomic variants), and CRAM/SAM/BAM file formats (for representing sequencing read data).

30.8.3 Three Approaches to Scaling

Current scaling solutions can be broadly considered as three related approaches.

First: VCF specification modifications that make it scale linearly or sublinearly. Examples include: – Sparse Allele Vectors (SAV): a format for storing very large sets of genotypes and haplotype dosages that produces small file sizes and is optimized for fast association analysis – Scalable Variant Call Representation (SVCR): a format which takes advantage of reference block compression to guarantee linear scaling with the number of samples – Sparse Project VCF (spVCF): works by selectively reducing high-entropy quality control information and run-length encoding reference coverage information

This solution is relatively easy to accomplish and could gain adoption quickly, but adoption depends on the community coalescing around a chosen approach.

Second: New binary file format engineered to support large dataset sizes and different operations—for example, querying all data from one sample, or querying one region across all samples. SAV is an extension to BCF (the binary form of VCF). This approach could be challenging for community adoption, as it requires multiple community libraries to implement support.

Third: API-based solution with a protocol for how to retrieve data and the format in which data are returned, which could be any VCF format. This is a hybrid approach that exploits VCF specification improvements in parallel with development of protocols that interact with VCF, and is considered the most promising path forward.

30.8.4 Why VCF Scaling Matters

Successful adoption of VCF scaling solutions is vital for the scientific community. These solutions will enable the reuse of large population-based genomics and health datasets. The method of virtual cohorts—combining data from multiple studies to increase statistical power—will particularly benefit from VCF scaling solutions. This is especially important for studying rare diseases, where research often relies on smaller populations from multiple studies conducted over time.

The GA4GH Future of VCF Working Group offers an ideal collaborative setup to tackle the scaling challenges facing researchers and data service providers, providing a unique forum for exchanging ideas and experience that might otherwise remain siloed.

30.9 Why VCF Matters

VCF is more than just a file format—it's the data structure that makes modern genomics possible.

Before VCF, every lab and every tool had its own way of representing variants. Comparing data across studies was a nightmare. Sharing data required custom converters. Reproducibility was hard.

VCF standardized variant representation. Now, a variant called in one lab can be directly compared to variants from another lab, even if they used different sequencing technologies or variant callers. Databases like gnomAD and ClinVar use VCF to distribute millions of variants worldwide. Clinical labs use VCF to report results. Every major analysis tool accepts VCF as input.

VCF also scales—though with the challenges outlined above. The same format that works for one person's exome can work for a million people's whole genomes with appropriate scaling solutions. The logic remains identical—just more lines, more columns, and sophisticated storage strategies.

When you read a paper that says "a rare variant in gene X was identified," that variant came from a VCF file. When you see allele frequencies in gnomAD, those came from VCF INFO fields. When you run a GWAS or do variant annotation, you're manipulating VCF files.

The VCF file is how genomes become data—how biological molecules become numbers that computers can process, compare, and analyze. From a single A→T change to biobank-scale datasets with millions of variants across hundreds of thousands of people, VCF is the language we use to describe variation.

It's a simple text file. But it encodes something profound: the genetic diversity of our species, written in a compact digital language of **0/0**, **0/1**, and **1/1**.

31 Part 5. From Variant to System: Experimental Genetics and Gene Regulation

31.1 When Finding the Gene Wasn't Enough

In 1989, Francis Collins and Lap-Chee Tsui identified the gene responsible for cystic fibrosis—CFTR—after four years of painstaking work. The discovery made headlines worldwide. Families celebrated. Researchers declared it a breakthrough.

But then came the hard part.

Knowing that CFTR mutations caused cystic fibrosis didn't explain how. What did the CFTR protein actually do? Why did losing it fill lungs with mucus? Could you fix the disease by restoring the protein? And which of the 2,000+ different CFTR mutations could be treated the same way?

These questions required a completely different toolkit. Not sequencing or statistics, but **experimental manipulation**: knocking out genes in cells, measuring what changed, testing whether restoring function rescued the disease. It took another decade of reverse genetics—creating cell lines and mouse models with precise CFTR mutations—before researchers understood that CFTR was a chloride channel, that its loss disrupted ion transport, and that different mutations affected protein folding, trafficking, or function in distinct ways.

This pattern repeats across genetics. GWAS finds hundreds of type 2 diabetes loci—but which genes at each locus actually matter? Burden tests implicate loss-of-function mutations in SCN2A in epilepsy—but can you treat it by boosting the remaining gene copy? Variant discovery is just the beginning. Understanding mechanism requires **moving from variant to system**.

Then in 2024, Lenk and colleagues used CRISPR activation to treat SCN2A haploinsufficiency in mice—not by fixing the mutation, but by increasing expression of the normal copy. Neurons fired properly again. Seizures decreased. The approach worked because they understood not just which gene was broken, but how gene regulation could be manipulated to restore normal function.

This chapter lives at the intersection of discovery and mechanism: how we find disease genes, how we test their function experimentally, how gene regulation controls which genes are active, and how genetic variants affect molecular traits.

31.2 What This Chapter Is About

This chapter explores the complete path from genetic association to biological mechanism, integrating experimental manipulation with molecular measurement.

31.2.1 Five Interconnected Topics

1. Forward Genetics: Phenotype to Gene
How do we find disease genes in human populations? You'll learn how GWAS tests millions of common variants across huge cohorts, how burden tests aggregate rare variants to find genes where loss-of-function causes disease, and why association isn't causation—understanding the crucial difference between finding a locus and identifying the causal gene.

2. Reverse Genetics: Gene to Function
How do we test what genes actually do? We'll explore the experimental toolkit: CRISPR-based editing, knockouts, overexpression, and high-throughput screens. You'll learn how to design functional experiments, interpret phenotypic rescue, and understand why reverse genetics provides the causality that association studies can't.

3. CRISPRa Therapy for SCN2A Haploinsufficiency
How does this work in practice? Using SCN2A epilepsy as a case study, you'll see how researchers moved from identifying loss-of-function mutations to developing CRISPR activation therapy. We'll trace the complete pipeline: genetic discovery → functional validation → mechanism → therapeutic strategy → preclinical testing.

4. Gene Regulation: The Control Layer
How do cells decide which genes to express? You'll learn how promoters initiate transcription, how enhancers control it from a distance (sometimes megabases away), and how 3D chromatin looping brings regulatory elements together. We'll cover the key technologies—RNA-seq, ChIP-seq, ATAC-seq, Hi-C—that map the regulatory landscape and reveal where most disease variants actually act.

5. QTLs: Connecting Genotype to Molecular Phenotype
How do genetic variants affect gene expression and protein levels? We'll explore expression QTLs (eQTLs) connecting variants to RNA abundance, protein QTLs (pQTLs) connecting variants to protein levels, and how integrating QTL data with GWAS identifies causal genes and mechanisms. You'll see how the same variant can affect multiple genes in different tissues, and why this matters for disease.

31.3 What Makes This Different

Discovery (Parts 1–4)	Mechanism (Part 5)
Where is the disease gene?	What does the gene do?
Statistical association	Experimental causation
GWAS p-values and odds ratios	Knockout phenotypes and rescue experiments
Population-level patterns	Cellular and molecular mechanisms
Finding candidate genes	Testing candidate genes
Variants affect disease risk	Understanding how variants affect biology

The shift: From identifying genetic associations to understanding biological mechanisms. From knowing that a variant matters to knowing how it works and what we can do about it.

31.4 How to Approach This Chapter

31.4.1 Think in Complementary Directions

Forward genetics (phenotype → gene) and **reverse genetics** (gene → phenotype) aren't competing approaches—they're two halves of the discovery loop. GWAS finds candidates. Knockouts test them. QTLs connect them to molecular changes. Each method has blind spots that the others cover.

31.4.2 Connect Association to Mechanism

Every genetic association has a mechanistic explanation: – A GWAS locus contains the causal gene (find it with fine-mapping and QTLs) – The variant affects gene regulation (measure it with eQTLs and ATAC-seq) – Changed expression alters cellular function (test it with CRISPR) – Altered function causes disease (validate it with rescue experiments)

Never stop at association. Always ask: what's the mechanism?

31.4.3 Master the Experimental Logic

Understanding reverse genetics means thinking about controls: – Knockout shows the gene is necessary – Overexpression shows it's sufficient – Rescue proves causality – Dose-response reveals quantitative relationships

Every experiment answers a specific question. Know which question you're asking.

31.4.4 Integrate Data Types

Modern genetics combines multiple layers: – **Genetic**: Which variants associate with disease? – **Transcriptomic**: How do variants affect RNA levels? – **Proteomic**: How do RNA changes translate to protein? – **Functional**: What happens in cells when you manipulate the gene? – **Organismal**: Does the cellular phenotype cause disease?

Integration across these layers reveals mechanism.

31.5 Why This Matters

Modern genomics requires understanding:

Clinical interpretation: Finding a *de novo* loss-of-function mutation isn't enough. You need to know whether the gene is dosage-sensitive, whether the protein has essential domains, whether model organisms show relevant phenotypes. Functional evidence determines pathogenicity.

Drug development: Most genetic associations don't lead to drugs because we don't understand mechanism well enough. The exceptions—PCSK9 inhibitors for cholesterol, SGLT2 inhibitors for diabetes—came from deep mechanistic work showing that reducing gene function *treats* disease.

Precision medicine: Two patients with *CFTR* mutations might need different therapies depending on whether the mutation affects protein folding (treat with correctors) or channel function (treat with potentiators). Mechanism determines treatment.

Therapeutic strategies: When you find disease genes, you can: restore function (gene therapy), reduce toxic gain-of-function (antisense oligonucleotides), modulate expression (CRISPRa/i), or target downstream pathways (small molecules). Each strategy requires understanding mechanism.

Gene regulation: Most disease variants don't change protein sequence—they affect *when* and *where* genes are expressed. Understanding enhancers, promoters, and chromatin is essential for interpreting non-coding variants, which constitute 93% of GWAS hits.

31.6 Learning Objectives

By the end of this chapter, you should be able to:

1. **Distinguish forward and reverse genetics**, explaining when each approach is needed and how they complement each other in the discovery process

2. **Interpret GWAS and burden test results**, understanding what association means, what it doesn't mean, and why fine-mapping and functional studies are essential
3. **Design reverse genetics experiments** using CRISPR, explaining the logic of knockout, activation, and rescue experiments
4. **Explain the CRISPRa strategy** for treating haploinsufficiency, distinguishing it from gene therapy and describing when dosage restoration is appropriate
5. **Describe gene regulation mechanisms**, distinguishing promoters from enhancers, and explaining how 3D chromatin structure brings distant elements together
6. **Interpret regulatory genomics data** from RNA-seq, ChIP-seq, and ATAC-seq experiments
7. **Explain QTL analysis**, distinguishing eQTLs from pQTLs, and describing how QTL integration with GWAS identifies causal genes
8. **Connect variants to mechanisms** by integrating genetic association, gene regulation, and functional data

---

31.7 What Makes This Different from Classic Genetics

Classic Gene Discovery	Modern Mechanistic Genetics
One gene causes disease	Networks of genes contribute
Code for protein	Mostly regulate other genes
Mutation disrupts protein function	Variant changes expression timing/location
Study one gene at a time	Screen thousands simultaneously
Test in model organisms	Validate in human cells and organoids
Years to make a knockout mouse	Days to edit cells with CRISPR

The transformation: Speed, scale, and sophistication have exploded. Technologies that took decades now take days. But the logic remains: perturb the system, measure what changes, infer mechanism.

31.8 Let's Begin

Collins and Tsui spent four years finding CFTR. Today you can run GWAS on 500,000 people in hours, identify hundreds of loci, and use QTL databases to nominate causal genes at each locus. Then you can order CRISPR reagents online, knock out the gene in patient-derived cells, measure transcriptome-wide changes, and test rescue with small molecules—all in weeks.

The scale is unimaginable compared to 1989.

But you're asking the same fundamental question: how do genes cause disease, and what can we do about it?

In Parts 1–3, you learned how to find variants and understand their effects. In Part 4, you learned how variants are organized and distributed. Now you'll learn how to move from genetic discovery to biological understanding—how to test mechanism, manipulate regulation, and translate findings into potential therapies.

The tools have changed from chromosomes to CRISPR, from pedigrees to QTLs, from candidate genes to genome-wide screens.

But the goal remains: understand how genetic variation shapes biology, and use that understanding to treat disease.

Let's see how modern experimental genetics bridges the gap between finding variants and understanding systems.

32 Chapter 25. Forward Genetics: From Phenotype to Gene

You've sequenced a patient's genome and found 4 million variants. The patient has a severe developmental disorder. Which variant is responsible?

This is the central challenge of forward genetics: starting from an observable trait or disease and working backward to find the genetic cause. It's detective work—you know something went wrong, but you need to figure out which gene is the culprit.

Forward genetics has been around since Mendel crossed pea plants in the 1860s, but the tools we use today are unrecognizable from those early days. Instead of breeding flies or counting wrinkled peas, we now scan entire genomes across thousands of people, looking for statistical patterns that link specific variants to disease risk.

32.1 The evolution of gene hunting

In the early 20th century, geneticists like Thomas Hunt Morgan worked forward from visible traits—like white eyes in fruit flies—to map genes on chromosomes. They didn't know what DNA was, but they could trace inheritance patterns through generations.

By the 1980s and 1990s, human geneticists were using **linkage analysis** to find disease genes in families. If a disease consistently travels with a chromosomal region across multiple generations, that region probably contains the causal gene. This approach identified hundreds of single-gene disorders, including cystic fibrosis (CFTR), Huntington's disease (HTT), and Duchenne muscular dystrophy (DMD).

But linkage analysis has limits. It works well for rare, high-penetrance Mendelian diseases that run in families. It struggles with common diseases like diabetes or schizophrenia, where hundreds of variants each contribute a tiny amount of risk and no single family can reveal the full picture.

```
Timeline of Gene Hunting Approaches
===================================

1860s-1900s: Classical Genetics
Mendel → Morgan → Chromosome mapping
• Method: Cross breeding, pedigree analysis
• Scale: Single families, visible traits
• Output: Inheritance patterns

1980s-1990s: Linkage Analysis
• Method: Track chromosomal regions in families
• Scale: Dozens of families, microsatellite markers
• Output: Chromosomal regions (cM resolution)
• Success: CFTR, HTT, DMD identified

2000s-2010s: GWAS Era
• Method: Genotype millions of SNPs in thousands of people
• Scale: 10K-1M individuals, common variants (>1% frequency)
• Output: Genomic loci (kb resolution)
```

```
• Success: 100+ loci per complex trait

2010s-Present: Sequencing Era
• Method: Whole-exome/genome sequencing + burden tests
• Scale: 10K-100K individuals, all variants
• Output: Specific genes and mutations
• Success: Rare variant architecture revealed
```

Enter the genomic era. In the 2000s, microarrays made it possible to genotype hundreds of thousands of variants across the genome in thousands of people simultaneously. This gave birth to **genome-wide association studies (GWAS)**, which revolutionized how we find genes for complex traits. At the same time, next-generation sequencing allowed discovery of ultra-rare variants that only appear in a handful of individuals—leading to the development of **burden tests** that aggregate rare variants within genes.

Today, forward genetics in humans relies on two complementary strategies:

```
Forward Genetics Workflow
=========================

Observable Phenotype (Disease/Trait)
           ↓
    [Choose Strategy]
           ↓

    ↓             ↓
  GWAS      Burden Tests
    ↓             ↓
Common variants  Rare variants
Small effects   Large effects
    ↓             ↓

           ↓
   Candidate Genes
           ↓
   Reverse Genetics
   (Functional Testing)
```

Let's see how each one works.

32.2 Genome-wide association studies (GWAS)

GWAS asks a simple question: across many people, are certain genetic variants more common in individuals with a disease compared to those without?

Here's how it works. Imagine you recruit 50,000 people with type 2 diabetes and 50,000 people without diabetes. You genotype about 1 million common SNPs across the genome in each person. Then, for every SNP, you calculate: Is this variant more frequent in cases than controls?

```
GWAS Process Flow
=================

Step 1: Recruit cohort
  Cases: 50,000 individuals with disease
  Controls: 50,000 unaffected individuals
```

```
Step 2: Genotype
  ~1 million SNPs per person (genome-wide)

Step 3: Test association
  For each SNP: frequency in cases vs. controls

Step 4: Statistical testing
  Most SNPs: no difference (noise)
  Some SNPs: significant enrichment (signal)

Step 5: Multiple testing correction
  Genome-wide significance: p < 5×10

Output: List of associated loci
  Typically 10-400+ loci per trait
```

Most variants show no difference—they're just random noise. But some variants pop up significantly more often in people with diabetes. Those are your association signals.

GWAS typically identifies dozens to hundreds of genomic regions (called **loci**) associated with a trait. For example, the largest GWAS of type 2 diabetes to date found over 600 risk loci across 1.4 million participants (Suzuki et al. 2024, Nature). Each individual locus has a tiny effect—maybe increasing disease risk by 5% or 10%—but together they point to biological pathways involved in insulin secretion, glucose metabolism, and pancreatic beta-cell function.

32.2.1 What GWAS tells you

GWAS results are usually summarized as **summary statistics**: for each variant, you get an effect size (how much it changes disease risk), a p-value (how confident you are it's real), and an allele frequency (how common it is in the population).

```
GWAS Summary Statistics Format
==============================

CHR    POS        SNP          A1  A2   FREQ   BETA    SE      P
-------------------------------------------------------------------
1      12345678   rs123456     T   C    0.35   0.082   0.015   4.2e-08
6      89012345   rs789012     G   A    0.12   0.145   0.023   2.1e-10
10     45678901   rs345678     C   T    0.48   0.063   0.012   1.5e-07

CHR  = Chromosome
POS  = Position (bp)
SNP  = Variant identifier
A1   = Effect allele
A2   = Reference allele
FREQ = Effect allele frequency
BETA = Effect size (log odds ratio)
SE   = Standard error
P    = P-value
```

These summary statistics are incredibly useful beyond the original study:

Application	Description	Example
Polygenic risk scores (PRS)	Combine effects of thousands of variants into a single genetic risk estimate	Top 10% of T2D PRS → 3–4× higher risk than median
Genetic correlation	Compare GWAS across traits to identify shared genetic factors	Schizophrenia and bipolar disorder show strong correlation
Mendelian randomization	Use variants as natural experiments to test causality	LDL–raising variants → heart disease implies causation
Fine–mapping	Narrow down causal variants within loci	Identify credible sets of likely causal SNPs
Gene prioritization	Integrate with eQTL data to identify target genes	Connect GWAS locus to regulated gene

32.2.2 What GWAS doesn't tell you

GWAS is powerful for finding where genetic risk lies, but it's frustratingly vague about what those variants actually do.

GWAS Limitation	Challenge	Solution Needed
Non–coding variants	93% of GWAS hits are outside genes	Regulatory annotation, eQTLs, functional screens
Linkage disequilibrium	Multiple correlated variants—which is causal?	Fine–mapping, statistical genetics
Distant regulation	Variant may regulate gene 500kb away	Chromatin looping (Hi–C), enhancer mapping
Common variants only	Misses rare, high–impact mutations	Sequencing + burden tests
Population–specific	LD structure differs across ancestries	Multi–ancestry GWAS, functional validation

First, most GWAS hits aren't in genes at all—they're in regulatory regions between genes. A variant might be 500 kilobases away from the nearest gene. Which gene does it regulate? You'll need additional experiments to figure that out.

Second, GWAS only detects **common** variants (usually 〉1% frequency in the population). If a disease is caused by hundreds of different ultra–rare mutations scattered across a gene, GWAS will miss them entirely.

That's where burden tests come in.

32.3 Burden tests: hunting rare variant effects

Many genetic diseases—especially severe developmental disorders or early–onset conditions—are caused by rare or even **de novo** (brand new) mutations. These variants are so uncommon that you might see each one in only a single family, or even a single person.

Here's the problem: if a variant appears in just one individual out of 50,000, you can't do statistics on it. The sample size is literally 1.

The solution? Instead of testing variants one by one, test **genes**. Specifically, ask: Do individuals with this disease carry more damaging mutations in a particular gene than expected by chance?

This is called a **burden test**, because you're measuring the total mutational burden within a gene.

32.3.1 How burden tests work

```
Burden Test Process
==================

Step 1: Sequence cohort
  Cases: 10,000 individuals with disease
  Controls: 10,000 unaffected individuals

Step 2: Variant calling & filtering
  Identify all variants in coding regions
  Filter to damaging variants:
      Loss-of-function (LoF): nonsense, frameshift, splice
      Missense: predicted damaging (CADD, PolyPhen)
      Filter by frequency: MAF < 0.1%

Step 3: Aggregate by gene
  Count carriers per gene (not individual variants)

Step 4: Statistical test
  For each gene: cases vs. controls burden

Step 5: Multiple testing correction
  Genome-wide significance: p < 2.5×10  (20,000 genes)

Output: Ranked list of intolerant genes
```

Let's say you're studying autism. You sequence the exomes of 10,000 individuals with autism and 10,000 unaffected controls. For every gene, you count how many people carry a loss-of-function (LoF) mutation—nonsense, frameshift, or splice-site variants that knock out the gene.

In most genes, you'll see about the same number of LoF carriers in cases and controls. But in a few genes—like CHD8, SCN2A, or SYNGAP1—you'll see a striking excess in autism cases (Satterstrom et al. 2020, Cell). For example, 30 autism cases might carry a CHD8 LoF mutation, while only 2 controls do. That imbalance is statistically significant and tells you CHD8 is an autism risk gene.

Burden tests are especially powerful for:

Use Case	Description	Why It Works
De novo mutations	Brand-new mutations not inherited from parents	Even single occurrences are informative if recurrent across unrelated cases
Ultra-rare variants	Each mutation unique to one family	Aggregating across gene overcomes individual variant rarity
Recessive diseases	Requires two hits in same gene	Test for enrichment of compound heterozygotes or homozygotes
Gene intolerance	Identify genes critical for development	LoF-intolerant genes show depletion in controls (gnomAD pLI)

32.3.2 Burden test output

Burden test results are typically presented as a list of genes ranked by statistical significance:

```
Example Burden Test Results (Autism)
====================================

Gene      Cases_LoF  Controls_LoF  Odds_Ratio  P-value      Interpretation
-------------------------------------------------------------------------------
CHD8      35         2             18.5        1.2e-25      High-confidence risk gene
SCN2A     28         3             9.8         4.5e-20      High-confidence risk gene
SYNGAP1   22         1             23.1        8.1e-18      High-confidence risk gene
GRIN2B    18         4             4.7         2.3e-10      Moderate-confidence
SHANK3    15         2             7.8         6.7e-09      Moderate-confidence
ADNP      12         1             12.5        3.2e-08      Suggestive

Interpretation thresholds:
• p < 2.5×10 : Genome-wide significant (Bonferroni for ~20K genes)
• Odds ratio > 5: Strong effect size
• Cases > 10: Sufficient observations for confidence

Source: Satterstrom et al. 2020, Cell
```

These results tell you which genes are intolerant to mutation and likely play a key role in the disease.

32.4 GWAS vs burden tests: two sides of the same coin

GWAS and burden tests aren't competing methods—they're complementary. They detect different types of genetic risk.

Feature	GWAS	Burden Tests
Variant type	Common (MAF 〉 1%)	Rare/ultra-rare (MAF 〈 0.1%)
Effect size	Small (OR 1.05–1.3)	Large (OR 2–50)
Variant location	Mostly non-coding (93%)	Mostly coding (exons)
Mechanism	Gene regulation	Protein disruption
Metaphor	Volume knob	Light switch
Sample size needed	50K–1M people	5K–50K people
Per-variant power	Low (must be common)	High (aggregate effect)
Population frequency	Present in many individuals	Private to few families
Clinical utility	Polygenic risk scores	Diagnostic testing

GWAS finds common variants that individually have small effects but are shared across many people. These variants tend to sit in regulatory regions and fine-tune gene expression levels. Think of them as turning a volume knob: they don't break the system, but they shift it slightly up or down. Because each variant has a modest effect, you need large sample sizes (tens or hundreds of thousands of people) to detect them reliably.

Burden tests find rare variants that individually have large effects but are seen in only a few families. These variants tend to disrupt protein-coding sequences and knock out or severely damage gene function. Think of them as flipping a light switch off: the effect is dramatic, but each variant is unique. Because the effects are big, you can detect them with smaller sample sizes—though you still need thousands of individuals to find enough mutation carriers.

32.4.1 Example: Schizophrenia genetic architecture

```
Schizophrenia: Integrated Genetic Architecture
==============================================

GWAS Findings (Common Variants)
  287 genome-wide significant loci
  Each variant: OR 1.03-1.15 (3-15% risk increase)
  Collectively: ~20% of liability variance explained
  Location: 95% non-coding (regulatory regions)
  Mechanism: Fine-tune gene expression
   Source: Trubetskoy et al. 2022, Nature

Burden Test Findings (Rare Variants)
  10 high-confidence risk genes
  Each LoF mutation: OR 3-50 (3-50× risk increase)
  Frequency: 0.01-0.1% of cases carry each mutation
  Location: Protein-coding exons
  Mechanism: Knock out gene function
   Source: Singh et al. 2022, Nature

Key Risk Genes from Burden Tests:
• SETD1A: Histone methyltransferase (chromatin regulation)
• GRIN2A: NMDA receptor subunit (glutamate signaling)
• GRIA3: AMPA receptor subunit (glutamate signaling)
• RB1CC1: Autophagy regulator
• SP4: Transcription factor
• TRIO: Rho guanine nucleotide exchange factor
• CACNA1G: Voltage-gated calcium channel
• CUL1, HERC1: Ubiquitin ligases
• XPO7: Nuclear export protein

Biological Convergence:
Both common and rare variants converge on:
→ Synaptic function (glutamate signaling)
→ Chromatin regulation (gene expression control)
→ Neuronal development pathways
```

Here's a concrete example. In **schizophrenia**, GWAS has identified 287 common risk loci that collectively explain about 20% of the genetic variance in liability (Trubetskoy et al. 2022, Nature). Most of these variants are in non-coding regions and probably affect gene regulation. But whole-exome sequencing and burden tests have also found that ultra-rare LoF mutations in 10 genes—like SETD1A and GRIN2A—dramatically increase schizophrenia risk (Singh et al. 2022, Nature). Each mutation is found in maybe 0.01% of cases, but when you carry one, your risk goes up 3–50 fold.

So the full genetic architecture of schizophrenia includes:

- **Hundreds of common variants** (GWAS), each shifting risk by a few percent
- **Ten rare variant genes** (burden tests), each shifting risk by several hundred percent

Both matter. GWAS tells you which biological pathways are involved at the population level. Burden tests tell you which genes, when severely disrupted, cause or strongly predispose to disease. Together, they give you a comprehensive map of genetic risk.

32.5 From association to causation

Forward genetics gets you from phenotype to a list of candidate genes. But association isn't causation. Just because a gene shows up in a GWAS or burden test doesn't mean you understand what it does or how it contributes to disease.

```
The Forward → Reverse Genetics Loop
====================================

FORWARD GENETICS                    REVERSE GENETICS
(Discovery)                         (Validation)

Phenotype                           Manipulate gene
   ↓                                      ↓
GWAS/Burden test                    Measure phenotype
   ↓                                      ↓
Candidate genes    →→→→→→→→         Confirm causation
   ↓                                      ↓
Association signal                  Mechanism
   ↓                                      ↓
"This gene matters"    ←←←←←←       "This is how it works"

Key Questions Answered by Reverse Genetics:
• Is this gene sufficient to cause the phenotype?
• Is this gene necessary for normal function?
• What cellular processes does it affect?
• Can we rescue the phenotype by restoring function?
• Which variants are gain vs. loss of function?
```

That's where the next step comes in: **reverse genetics**—deliberately manipulating those candidate genes to see what happens. We'll cover that in the next chapter, where we'll explore how CRISPR, CRISPRa, base editing, and high-throughput screens let you move from gene lists to functional mechanisms.

For now, remember the key principle of forward genetics: **start with the biology (the phenotype), and let the data lead you to the genes**. It's the foundation of modern human genetics—and the starting point for everything that follows.

33 Chapter 26. Reverse Genetics: From Gene to Function

You've done a burden test and found that SCN2A is enriched for loss-of-function mutations in individuals with epilepsy. You're pretty confident this gene matters. But how does it matter? What does SCN2A actually do in neurons? What happens when you lose it? Can you fix the problem by restoring its function?

These are the questions that reverse genetics answers. Instead of starting from a phenotype and hunting for genes, you start from a known gene—or even a specific variant—and deliberately manipulate it to see what happens.

Forward genetics got us from disease to gene. Reverse genetics takes us from gene to mechanism.

33.1 The logic of reverse genetics

Reverse genetics is essentially controlled experimentation on genes. You pick a gene, you change it (remove it, activate it, edit it), and you observe the result. The goal is to establish causality: does this gene cause the phenotype we're interested in?

In model organisms like yeast, flies, or mice, reverse genetics has been around for decades. Scientists could delete genes, overexpress them, or swap in mutant versions to understand their function. But in humans, we obviously can't do those experiments directly. Instead, we use:

1. **Cell models** (human iPSCs, patient-derived cells, organoids)
2. **Animal models** (mice, zebrafish, sometimes non-human primates)
3. **Computational predictions** (when we can't do wet-lab experiments)

The core idea is the same: perturb the gene and measure what changes.

33.2 Classic reverse genetic approaches

Before we talk about CRISPR (which has taken over the field), let's quickly review the fundamental types of genetic manipulation. These strategies form the conceptual backbone of reverse genetics, even if the tools have changed.

33.2.1 Loss of function: what happens when the gene is gone?

The simplest reverse genetic experiment is to remove a gene and see what breaks.

Knock-out means completely eliminating the gene—either by deleting it, disrupting its coding sequence, or introducing a premature stop codon. If knocking out a gene causes lethality, that tells you the gene is essential. If it causes a specific defect—like neurons failing to fire action potentials—that tells you what the gene normally does.

Knock-down means reducing (but not eliminating) gene expression, often using RNA interference (RNAi). Small RNA molecules bind to the gene's mRNA and trigger its degradation, so less protein gets made. Knock-down is useful when complete knock-out is lethal, or when you want to test partial loss of function.

For example, SCN2A knock-out in mouse neurons eliminates sodium currents and prevents action potential generation, revealing that SCN2A encodes a sodium channel required for neuronal excitability.

33.2.2 Gain of function: what happens when you add more?

The flip side is to increase a gene's activity and see if that's sufficient to drive a phenotype.

Overexpression means cranking up gene expression—usually by putting the gene under a strong promoter that forces the cell to make lots of mRNA and protein. If overexpressing a gene causes a phenotype (say, excessive cell division), that tells you this gene is sufficient to trigger that process.

Ectopic expression means making a gene active in tissues or developmental stages where it's normally silent. For instance, expressing a neural transcription factor in skin cells might reprogram them into neurons, showing that the factor is sufficient to specify neural identity.

Both approaches test sufficiency: Can this gene, by itself, produce a particular outcome?

33.2.3 Conditional and inducible systems: controlling when and where

Sometimes you need precise control over timing and location. If a gene is required early in development, knocking it out might kill the embryo before you can study its role in adults. The solution? Conditional systems that let you turn genes on or off at will.

Inducible expression uses external triggers (like adding a drug) to activate or silence a gene. For example, the tetracycline (tet-on/tet-off) system uses doxycycline to control gene expression. Add the drug, the gene turns on; remove the drug, it turns off. This lets you study time-dependent effects.

CRE-lox recombination is a clever genetic trick used mostly in mice. You flank a gene with loxP sites (short DNA sequences), then introduce CRE recombinase enzyme in specific tissues. CRE cuts out everything between the loxP sites, deleting the gene—but only in tissues where CRE is expressed. This gives you tissue-specific knock-outs.

These tools have been workhorses of mouse genetics for 30 years. They're still widely used, but they're now often combined with CRISPR-based methods for even greater precision.

33.3 CRISPR: a unified platform for reverse genetics

CRISPR changed everything. Instead of needing different tools for different tasks—one system for knock-outs, another for overexpression, another for epigenetic modification—CRISPR provides a single, programmable platform that can do all of it (Zhang et al 2025, Genomics Proteomics & Bioinformatics).

The core mechanism is simple: a guide RNA directs a Cas protein to a specific DNA sequence. What happens next depends on which version of Cas you use.

Zhang2025-GenProtBio-Figure2.jpeg

Figure: Overview of CRISPR-Cas system technological innovations. The CRISPR toolkit has expanded dramatically beyond simple DNA cutting. Base editors use catalytically inactive Cas9 (nCas9) fused to deaminase enzymes to directly convert specific nucleotide bases without creating double-strand breaks. Prime editors combine nCas9 with reverse transcriptase to enable precise insertions, deletions, or base substitutions using a specialized guide RNA template. For epigenome editing, catalytically dead Cas9 (dCas9) is fused to DNA methylases or histone-modifying enzymes to alter gene expression without changing the DNA sequence itself—enabling both gene silencing (CRISPRoff) and activation (CRISPRon). CRISPRi and CRISPRa systems use dCas9 fused to transcriptional repressors (like KRAB domain) or activators (like VP64 domain) to reversibly control gene expression. Finally, RNA-targeting Cas13 systems enable direct editing of RNA sequences, providing transient modifications without permanent genomic changes. Together, these tools provide unprecedented precision in manipulating genetic information at multiple levels. Source: Zhang, X. et al. (2025). CRISPR Technology and Its Emerging Applications. Genomics, Proteomics & Bioinformatics. https://doi.org/10.1093/gpbjnl/qzaf034. License: CC-BY 4.0.

33.3.1 CRISPR knock-outs: cutting DNA

The original CRISPR application uses **Cas9**, which cuts both strands of DNA at a target site. The cell tries to repair the break, but often makes mistakes—inserting or deleting a few bases. If this happens in a coding region, it shifts the reading frame and creates a premature stop codon, knocking out the gene.

This is now the standard way to create knock-outs in human cells. You design a guide RNA targeting your gene of interest, deliver it along with Cas9, and wait a few days. Most cells will have frameshift mutations, and you can select the ones where the gene is fully inactivated.

For example, SCN2A knock-out in human iPSC-derived neurons using CRISPR confirmed loss of sodium current—validating the mouse experiments.

33.3.2 Base editing and prime editing: precision changes

Sometimes you don't want to destroy a gene—you want to introduce a specific mutation, like changing a single nucleotide.

Base editors are modified versions of Cas9 fused to enzymes that chemically convert one DNA base into another. There are two main types:

- **Cytosine base editors (CBE)** convert C → T (or G → A on the opposite strand)
- **Adenine base editors (ABE)** convert A → G (or T → C)

This is incredibly useful for modeling disease mutations. If a patient has a C–to–T mutation in SCN2A, you can use a base editor to introduce that exact change in cultured neurons and study how it affects channel function.

Prime editors go even further. They can make small insertions, deletions, or any base–to–base change without cutting the DNA. Think of prime editing as "find and replace" for the genome—you specify the target sequence and the desired replacement, and the editor writes in the new sequence. It's more flexible than base editing but also more complex.

33.3.3 CRISPRa and CRISPRi: controlling gene expression

What if you don't want to edit the DNA at all? What if you just want to turn a gene up or down?

Enter **dCas9**—a "dead" version of Cas9 that can't cut DNA but can still bind to target sequences. By itself, dCas9 does nothing useful. But if you fuse it to other proteins, it becomes a programmable gene regulator.

CRISPRa (activation) uses dCas9 fused to transcriptional activators. Guide the complex to a gene's promoter, and it recruits the transcriptional machinery, turning the gene on. This is essentially targeted overexpression—you're boosting a gene's output without changing its DNA sequence.

CRISPRi (inhibition) uses dCas9 to block transcription. The dCas9 complex sits on the promoter like a roadblock, preventing RNA polymerase from transcribing the gene. It's a reversible, tunable knock–down.

These tools are powerful for studying gene regulation, especially in contexts where permanent DNA changes would be problematic. For instance, if a disease is caused by haploinsufficiency (losing one copy of a gene), you could use CRISPRa to boost expression from the remaining copy and potentially rescue the phenotype.

33.3.4 Cas13: targeting RNA

Most CRISPR systems work on DNA. But there's also **Cas13**, which targets RNA instead.

Cas13 can be used to knock down specific mRNA transcripts without touching the genome. This is like RNAi but more programmable and specific. It's especially useful for studying RNA biology—splicing, localization, stability—or for temporary gene silencing when you don't want permanent edits.

Cas13 is still a newer tool and less commonly used than DNA–targeting systems, but it's expanding the reach of CRISPR into the transcriptome.

33.4 High–throughput reverse genetics: screening thousands of genes at once

Traditional reverse genetics tests one gene at a time. But what if you want to know which of 20,000 genes affect a particular phenotype? That's where high–throughput screens come in.

33.4.1 Pooled CRISPR screens

The idea is simple but powerful: create a library of guide RNAs that target every gene in the genome (or every gene in a pathway). Deliver this library to a large population of cells so that each cell gets one guide RNA and knocks out one gene. Then apply a selective pressure—maybe you treat the cells with a drug, or you sort them for a specific marker—and sequence which guide RNAs are enriched or depleted.

For example, you might screen for genes required for cancer cell survival. Knock out 20,000 genes across millions of cells, then grow them for a few weeks. Guides targeting essential genes will disappear (those

cells died), while guides targeting non-essential genes will persist. Sequence the surviving guides, and you have a ranked list of genes required for cell survival.

Pooled screens are now routine and incredibly powerful. They've been used to identify:

- Drug resistance genes (which genes, when lost, make cancer cells sensitive to therapy?)
- Host factors for viral infection (which genes does a virus need to replicate?)
- Developmental regulators (which genes control differentiation into specific cell types?)

33.4.2 Perturb-seq: linking perturbations to transcriptomes

Pooled screens tell you which genes affect a phenotype, but not how. Enter **Perturb-seq**, which combines CRISPR perturbations with single-cell RNA sequencing.

Here's the workflow: knock out thousands of genes using a pooled CRISPR library, then do single-cell RNA-seq on the entire population. Each cell's transcriptome tells you (1) which gene was knocked out (because the guide RNA is barcoded) and (2) how the knockout changed the cell's gene expression program.

Figure: In vivo AAV-Perturb-seq experimental design. AAV-Perturb-seq enables direct in vivo CRISPR screening by combining viral delivery with single-nucleus RNA sequencing. The workflow begins with systemic injection of AAV particles carrying pooled gRNA libraries into mice expressing Cas9 or dCas9-KRAB. After allowing time for viral transduction and gene perturbation, brain tissue is harvested and nuclei are isolated. Single-nucleus RNA-seq is then performed using a specialized 5' capture method that simultaneously sequences both the cellular transcriptome and the gRNA identity. This approach reveals which gene was perturbed in each cell while capturing the resulting transcriptional changes. The key innovation is the ability to perform multiplexed genetic screens directly in complex tissues like the adult brain, avoiding the limitations of in vitro systems or developmental-stage restrictions of lentiviral methods. By pairing perturbation identity with phenotypic readout at single-cell resolution, Perturb-seq enables systematic mapping of genotype-phenotype relationships in vivo. Source: Santinha, A.J. et al. (2023). Transcriptional linkage analysis with in vivo AAV-Perturb-seq. Nature. https://doi.org/10.1038/s41586-023-06570-y.

This is reverse genetics at scale. You can see not just whether knocking out a gene affects a phenotype, but which downstream genes and pathways are affected. For instance, if knocking out a transcription factor causes 500 genes to change expression, you've just mapped its regulatory network (Santinha et al 2023, Nature).

33.4.3 MPRA: testing regulatory sequences

CRISPR screens focus on genes. But what about the 98% of the genome that doesn't code for proteins? How do you test whether a non-coding variant in an enhancer region affects gene regulation?

Massively parallel reporter assays (MPRA) let you test thousands of DNA sequences simultaneously. You synthesize a library of short DNA fragments—each one containing a different variant or regulatory element—and clone them upstream of a reporter gene (like GFP or a barcode). Transfect the library into cells, measure reporter expression, and you can see which sequences drive high or low expression.

Figure: Large-scale lentiMPRA experimental strategy. Massively parallel reporter assays enable functional testing of thousands of regulatory sequences simultaneously. The process starts with designing a library of 230-nucleotide oligonucleotides representing candidate regulatory elements (cCREs) including both promoters and enhancers. Through PCR amplification, each sequence is linked to a minimal promoter and tagged with random barcodes. These constructs are cloned into a lentiviral vector containing a reporter gene (EGFP). The element-barcode associations are determined by deep sequencing before viral packaging. Cells are then infected with the lentiviral library, allowing genomic integration of the reporter constructs. After three days of culture, both DNA and RNA are extracted from the cells and barcodes are sequenced. By comparing RNA barcode counts (reflecting transcriptional activity) to DNA barcode counts (reflecting integration efficiency), researchers can calculate normalized activity scores for each regulatory

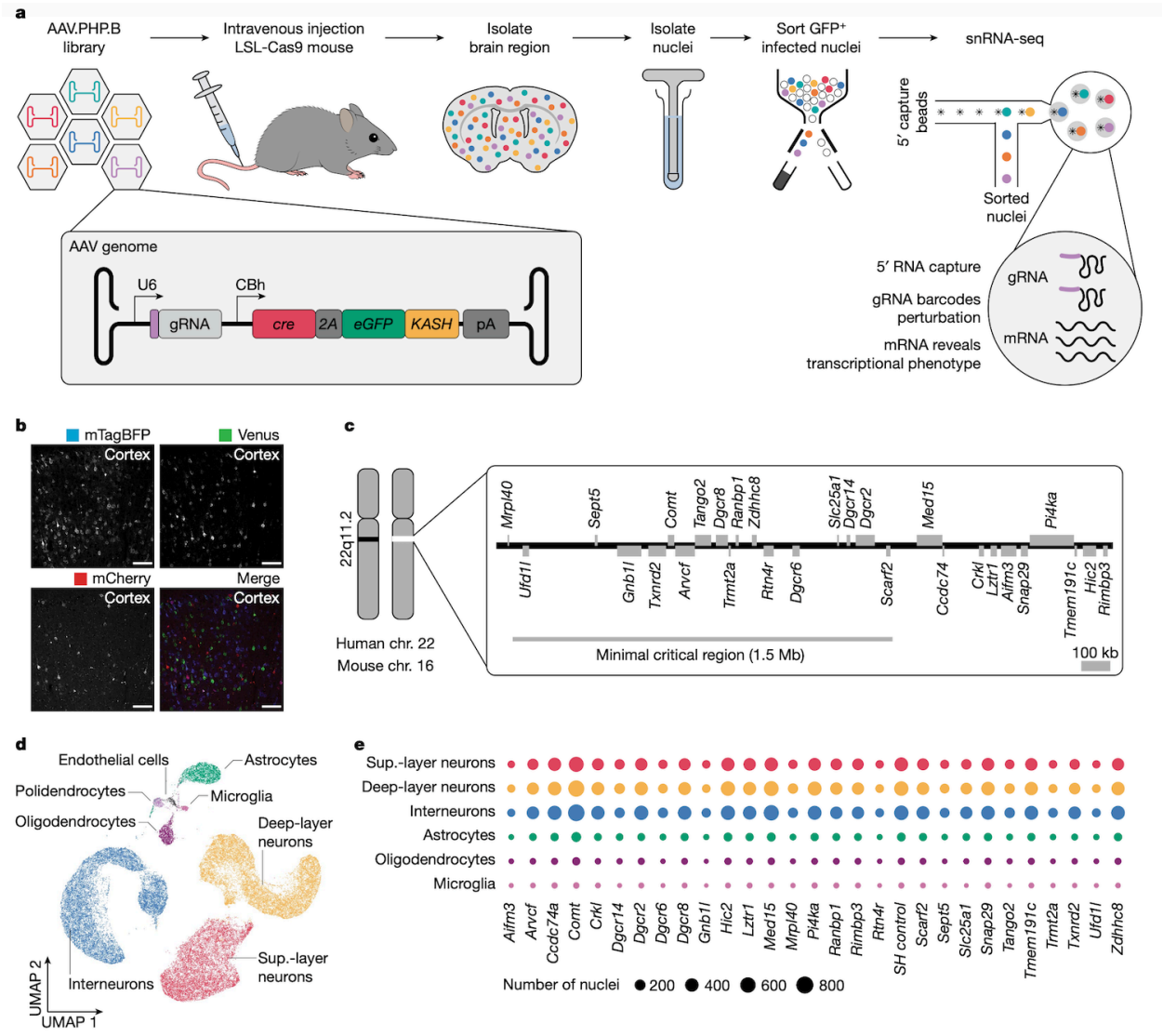


Figure 62: Santinha2023–Nature–Figure1.png

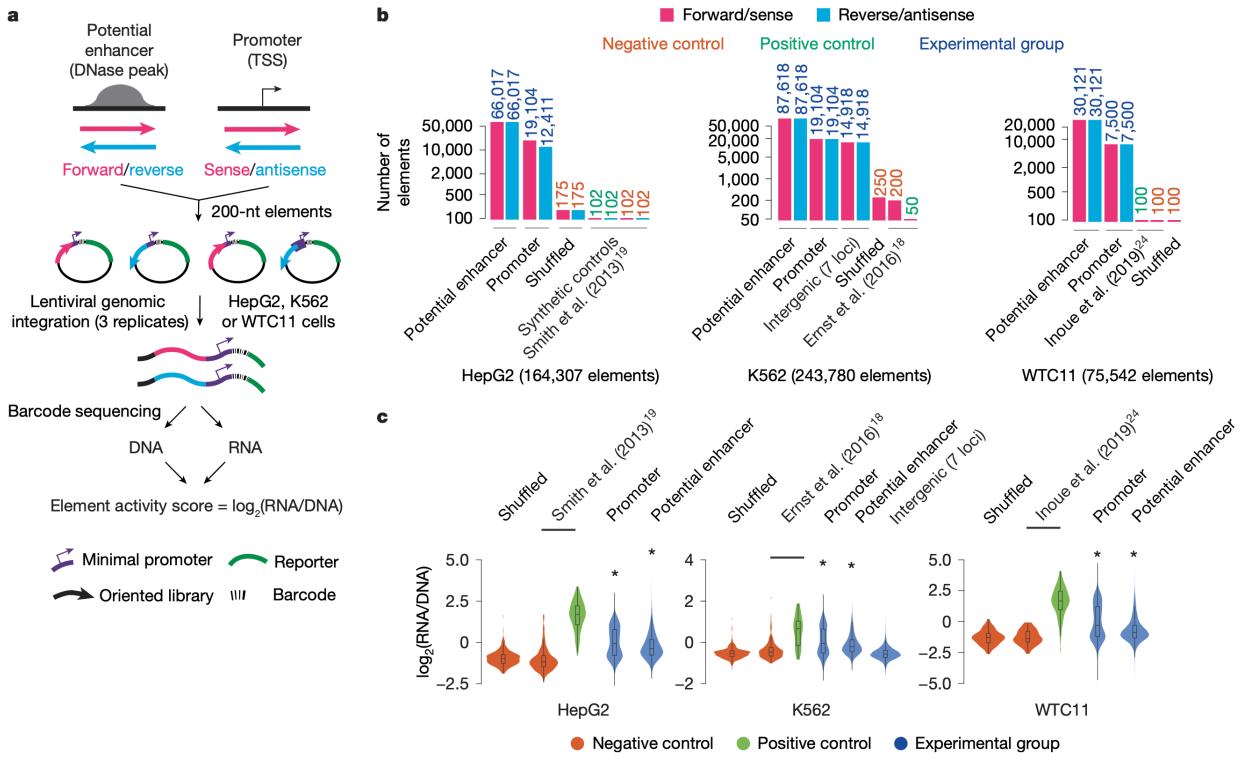


Figure 63: Agarwal2025-Nature-Figure1.png

element. This "in-genome" lentiviral approach provides more physiologically relevant measurements than transient transfection-based assays and enables testing of regulatory sequences in difficult-to-transfect cell types. Source: Agarwal, V. et al. (2025). Massively parallel characterization of transcriptional regulatory elements. Nature. https://doi.org/10.1038/s41586-024-08430-9.

For example, if a GWAS identifies a variant in an enhancer near a diabetes-risk gene, you can use MPRA to test whether the risk allele increases or decreases enhancer activity. This connects non-coding variants to functional effects (Agarwal et al 2025, Nature).

33.5 Functional rescue: proving causality

Once you've shown that losing or altering a gene causes a phenotype, the gold standard is to do a **rescue experiment**: restore the normal gene function and see if the phenotype goes away.

Rescue experiments prove causality. They show that the gene isn't just correlated with the phenotype—it's actually responsible for it.

There are three main rescue strategies:

33.5.1 1. Gene complementation: adding back a normal copy

The simplest rescue is to introduce a wild-type version of the gene into mutant cells or organisms.

For example, if zebrafish embryos with a scn2a mutation show abnormal brain development, you can inject wild-type SCN2A mRNA into the embryos. If the brain now develops normally, you've shown that SCN2A is necessary and sufficient for that developmental process.

In human cells, you can do the same thing—knock out a gene with CRISPR, then stably re-introduce a functional copy (often integrated at a safe harbor locus). If the phenotype disappears, the gene is confirmed as causal.

33.5.2 2. Gene correction: fixing the mutation

Instead of adding a new copy, you can use CRISPR to directly repair the mutated DNA sequence.

Base editors and prime editors are perfect for this. If a patient has a C-to-T mutation causing disease, you can use a base editor to convert the T back to C. If the cells now function normally, you've proven that this specific nucleotide change is responsible for the phenotype.

This approach is especially powerful because it corrects the mutation in its native genomic context—same promoter, same regulatory elements, same expression level. It's the cleanest possible test of causality.

33.5.3 3. Allelic activation: boosting the remaining copy

For haploinsufficiency disorders—where losing one gene copy causes disease because the remaining copy can't produce enough protein—you can use CRISPRa to upregulate the intact allele.

For instance, SCN2A haploinsufficiency causes developmental and epileptic encephalopathy. If you use CRISPRa to boost expression of the remaining SCN2A allele in patient-derived neurons, and the cells regain normal sodium current, you've demonstrated that simply increasing dosage is sufficient to rescue function.

This has therapeutic implications: instead of replacing a gene with gene therapy, maybe you can just turn up the remaining copy.

33.6 Putting it all together: from variant to mechanism

Reverse genetics is the experimental engine that turns genetic discoveries into biological understanding.

Here's the full workflow:

1. **Forward genetics identifies a candidate gene** (GWAS or burden test)
2. **Reverse genetics tests its function** (knock-out, overexpression, CRISPR screen)
3. **Functional rescue confirms causality** (complementation, correction, activation)
4. **Mechanistic studies reveal how it works** (Perturb-seq, live imaging, electrophysiology)

At each step, you're moving from association to causation, from correlation to mechanism.

For example, let's trace the journey of CHD8, an autism risk gene:

- **Forward genetics**: Burden tests found that autism cases are enriched for CHD8 loss-of-function mutations
- **Reverse genetics**: CRISPR knock-out of CHD8 in human neurons caused abnormal gene expression and synaptic defects
- **Rescue**: Adding back wild-type CHD8 restored normal function
- **Mechanism**: Perturb-seq showed that CHD8 regulates hundreds of genes involved in chromatin remodeling and neuronal differentiation

This is how you go from a statistical signal to a mechanistic understanding of disease.

33.7 The future: integrating forward and reverse genetics

We've come a long way from Mendel's pea plants. Forward genetics now operates at the scale of millions of genomes. Reverse genetics operates at the scale of genome-wide perturbation screens. And increasingly, the two are merging.

Imagine a future where every GWAS hit is immediately followed by a CRISPR screen to test which genes in the locus actually matter. Or where every rare variant discovered in a patient triggers an automated base-editing experiment in iPSC-derived cells to confirm its pathogenicity.

We're not there yet, but we're close. The tools are ready. The challenge now is integration—building pipelines that seamlessly connect variant discovery to functional validation at scale.

That's the promise of modern human genetics: not just finding genes, but understanding what they do and how to fix them when they break.

34 Chapter 27. CRISPRa Therapy for SCN2A Haploinsufficiency

Imagine you're a neuron trying to fire an action potential, but you only have half the sodium channels you need. You can still work, but you're sluggish—your signals are weak, your timing is off, and the whole circuit struggles to function normally. That's what happens in SCN2A haploinsufficiency, where losing one functional copy of the gene leaves neurons unable to generate normal electrical activity.

For decades, the standard approach to this kind of genetic problem was either gene replacement therapy (add a new working copy) or gene editing (fix the broken copy). But a 2025 study took a different approach: instead of fixing or replacing the mutant gene, the study simply turned up the expression of the remaining normal copy using **CRISPRa**—CRISPR activation (Tamura et al. 2025, Nature).

The results were striking. Boost SCN2A expression by about 50%, and neurons regain their firing capacity. Mice show fewer seizures, improved social behavior, and normalized brain activity. It's a proof-of-concept that sometimes the best therapy isn't to rewrite the genome—it's just to adjust the volume.

34.1 SCN2A and the haploinsufficiency problem

SCN2A encodes NaV1.2, a voltage-gated sodium channel that's critical for neuronal excitability. When an action potential starts, sodium channels open and let Na□ ions rush into the cell, depolarizing the membrane and propagating the electrical signal. NaV1.2 is especially important in early brain development and in certain types of excitatory neurons.

When you lose one working copy of SCN2A—through a nonsense mutation, a frameshift, or a splice-site variant—you're left with only 50% of the normal amount of NaV1.2 protein. This is called **haploinsufficiency**: one copy isn't enough to maintain normal function.

```
Normal vs. Haploinsufficient State
==================================

Normal (two functional alleles):
Allele 1: SCN2A → 50% NaV1.2 protein
Allele 2: SCN2A → 50% NaV1.2 protein
Total:           100% protein → Normal neuronal firing

Haploinsufficiency (one functional allele):
Allele 1: SCN2A → 50% NaV1.2 protein
Allele 2: Loss-of-function → 0% protein
Total:           50% protein → Hypoexcitability
```

```
Clinical consequences:
• Developmental delay
• Epilepsy (paradoxical seizures)
• Autism spectrum disorder
• Intellectual disability
```

The consequences? Neurons become hypoexcitable. They fire less frequently, generate weaker action potentials, and struggle to participate in fast, coordinated network activity. Clinically, SCN2A haploinsufficiency causes developmental and epileptic encephalopathy, autism spectrum disorder, and intellectual disability.

But here's the key point: the remaining SCN2A allele is perfectly normal. It's just not producing enough protein to compensate for the missing copy. So the question becomes: Can we make the good copy work harder?

34.2 Not all SCN2A variants are the same

Before we dive into CRISPRa, it's worth clarifying that SCN2A variants come in different flavors, and they cause very different problems.

Variant Class	Mechanism	Channel Activity	Neuronal Effect	Clinical Phenotype	Examples
Loss-of-function (LoF)	Premature stop, frameshift, splice defect, destabilizing missense	Reduced or absent	**Hypoexcitability** (neurons can't fire properly)	Developmental delay, ASD, paradoxical seizures	Nonsense, frameshift indels, critical splice sites
Gain-of-function (GoF)	Increased activation, impaired inactivation	Excessive or prolonged	**Hyperexcitability** (neurons fire too much)	Infantile spasms, early-onset epilepsy	Missense in voltage sensor or inactivation gate
Partial LoF	Reduced conductance, altered gating	Intermediate	**Mild hypoexcitability**	Variable—depends on remaining function	Some missense variants

Loss-of-function (LoF) variants reduce or eliminate channel activity. These include: - Nonsense mutations that introduce a premature stop codon - Frameshift indels that disrupt the reading frame - Splice-site mutations that skip essential exons - Some missense variants that destabilize the protein or block its ability to conduct sodium

LoF variants cause **hypoexcitability**—neurons can't fire properly, leading to developmental delays and sometimes paradoxical seizures (because inhibitory circuits are also affected).

Gain-of-function (GoF) variants, on the other hand, make channels open too easily or stay open too long. These cause **hyperexcitability**—neurons fire too much, leading to infantile spasms and early-onset epilepsy (Ben-Shalom et al. 2017, Biological Psychiatry).

The study focused on SCN2A haploinsufficiency caused by LoF variants. Electrophysiology experiments confirmed that neurons with only one functional SCN2A allele show reduced sodium currents and impaired firing. Combined with haploinsufficiency (only one allele makes protein), you get a severe reduction in total sodium current.

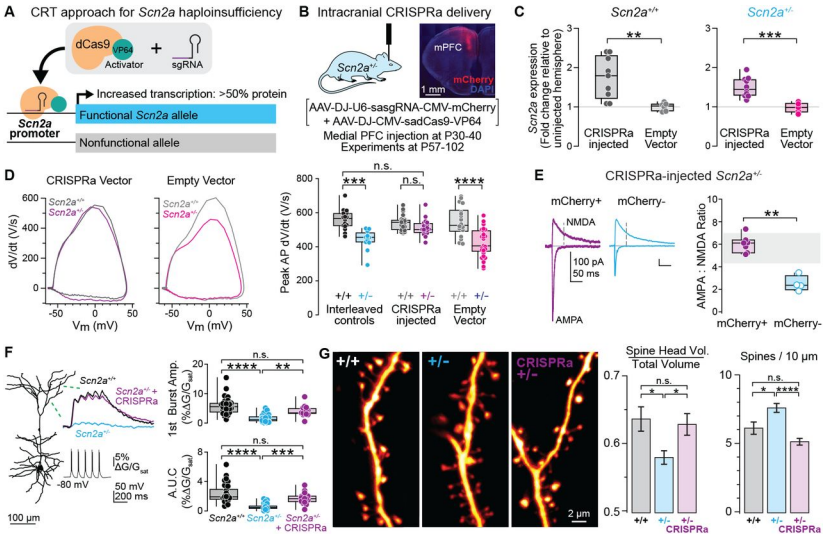


Figure 64: Tamura et al. 2022 bioRxiv – Figure 2

Figure: CRISPRa restores neuronal firing in SCN2A haploinsufficient neurons. The top traces show action potentials (electrical spikes) recorded from neurons when stimulated with current. Wild-type neurons (WT, black) fire many action potentials rapidly—they're healthy and excitable. Haploinsufficient neurons (Scn2a□/□, cyan) fire much less—they're "tired" with only half the normal sodium channels. After CRISPRa treatment (Scn2a□/□ + CRISPRa, purple), firing is restored to nearly normal levels. Each spike represents sodium channels opening—more spikes mean more functional channels. The graph at bottom quantifies this: CRISPRa-treated neurons fire as frequently as wild-type, while untreated haploinsufficient neurons show severely reduced firing. This is direct proof that boosting SCN2A expression fixes the fundamental cellular problem. Source: Tamura et al. (2022). CRISPRa rescues Scn2a haploinsufficient neurons. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.03.30.486483v1. License: CC-BY-NC-ND 4.0.

34.3 What is CRISPRa, and how does it work?

CRISPRa stands for **CRISPR activation**. To understand it, let's start with regular CRISPR-Cas9.

In standard genome editing, Cas9 is a molecular scissor. You give it a guide RNA that tells it where to cut, and Cas9 makes a double-strand break in the DNA. The cell tries to repair the break, and you can use that repair process to introduce or correct mutations.

But Cas9 doesn't have to cut. If you mutate the catalytic residues in Cas9, you get **dCas9** (dead Cas9)—a protein that can still bind to DNA at the guide RNA's target site, but it can't cut. By itself, dCas9 is useless. But if you fuse it to other proteins, it becomes a programmable platform for manipulating gene activity without editing the DNA sequence.

```
CRISPR Systems Comparison
==========================

Standard CRISPR-Cas9 (Genome Editing)
```

```
   Guide RNA → Cas9 → DNA target
             ↓
   Double-strand break (DSB)
             ↓
   DNA repair machinery
             ↓
   Mutation introduced

   Result: Permanent DNA sequence
           change


CRISPRa (Transcriptional Activation)

   Guide RNA → dCas9-VPR → Promoter
             ↓
   Recruit transcription factors
             ↓
   RNA polymerase activation
             ↓
   Increased transcription

   Result: More mRNA/protein
           (DNA unchanged)


CRISPRi (Transcriptional Repression)

   Guide RNA → dCas9-KRAB → Promoter
             ↓
   Block RNA polymerase access
             ↓
   Transcription blocked
             ↓
   Decreased transcription

   Result: Less mRNA/protein
           (DNA unchanged)


Key difference:
• CRISPR-Cas9 = Permanent genetic change
• CRISPRa/i = Reversible expression change
```

CRISPRa uses dCas9 fused to transcriptional activators—proteins like VP64, p65, or VPR (a combination of VP64, p65, and Rta). When you guide this complex to a gene's promoter, it recruits the transcriptional machinery and boosts RNA polymerase activity. The result? More mRNA gets made from that gene, and protein levels go up.

Think of it as a volume knob. The DNA sequence doesn't change, but the gene's output increases. You're essentially amplifying the signal from the normal allele.

CRISPRi (CRISPR interference) does the opposite. It uses dCas9 fused to repressor proteins (like KRAB) to block transcription. The dCas9 complex sits on the promoter like a roadblock, preventing RNA polymerase from accessing the gene. Expression goes down.

Both CRISPRa and CRISPRi are **epigenetic tools**. They don't rewrite the genetic code—they just tune how much it's read.

34.4 Rescuing SCN2A haploinsufficiency with CRISPRa

Here's what the 2025 study demonstrated (Tamura et al. 2025, Nature).

A mouse model of SCN2A haploinsufficiency was generated carrying one wild-type Scn2a allele and one mutant allele with a LoF variant. The mice showed reduced sodium currents in cortical neurons, abnormal EEG patterns, and behavioral deficits—essentially a mouse version of the human disorder.

```
Experimental Design
===================

Step 1: Generate haploinsufficient mice
  Genotype: Scn2a /  (one WT, one mutant allele)
  Phenotype: Reduced Na  current, seizures, social deficits
  Baseline measurements: Electrophysiology, EEG, behavior

Step 2: Design CRISPRa system
  Target: Scn2a promoter region
  Components: dCas9-VPR + guide RNA
  Delivery: AAV9 (adeno-associated virus)

Step 3: Administer therapy
  Route: Systemic (tail vein or retro-orbital injection)
  Timing: Adolescent mice (P30-P60)
  Wait 2-4 weeks for expression

Step 4: Measure outcomes
  Molecular: Scn2a mRNA and protein levels
  Cellular: Sodium currents, action potential firing
  Circuit: EEG recordings, seizure susceptibility
  Behavioral: Social interaction tests

Step 5: Test in human cells
  iPSC-derived neurons carrying SCN2A LoF variant
```

A guide RNA targeting the promoter region of Scn2a was designed. This guide RNA along with dCas9–VPR (a strong transcriptional activator) was packaged into an adeno-associated virus (AAV), a common gene therapy delivery vehicle.

The AAV was delivered systemically (via tail vein or retro-orbital injection) to adolescent mutant mice. Several weeks later, the effects were measured.

34.4.1 Results in mouse neurons

Measurement	Haploinsufficient (untreated)	After CRISPRa	Recovery
SCN2A mRNA level	~50% of WT	~80–90% of WT	☐ Nearly normalized
Sodium current amplitude	~50% of WT	~80–85% of WT	☐ Nearly normalized

Measurement	Haploinsufficient (untreated)	After CRISPRa	Recovery
Action potential firing frequency	Reduced	Restored to WT	☐ Fully normalized
Neuronal excitability	Hypoexcitable	Normal	☐ Phenotype rescued

- SCN2A mRNA levels increased, bringing total expression close to wild-type levels
- Sodium current amplitude in cortical neurons was restored to nearly normal
- Neuronal firing frequency—measured by patch-clamp electrophysiology—recovered
- The neurons looked and behaved like wild-type cells

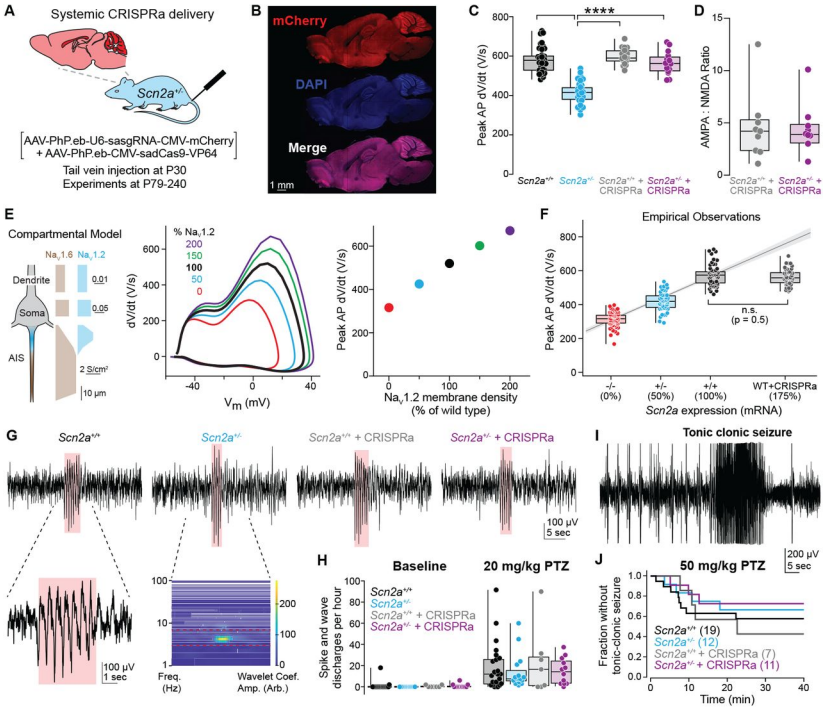


Figure 65: Tamura et al. 2022 bioRxiv – Figure 3

Figure: CRISPRa increases SCN2A expression and sodium current. (Left panel) Quantitative PCR measurements show that haploinsufficient mice (Het, cyan bar) have only ~50% of normal SCN2A mRNA compared to wild-type (WT, black bar). After CRISPRa treatment (CRISPRa, purple bar), expression increases to ~80–90% of wild-type levels—nearly complete restoration. (Right panel) This increase in mRNA translates directly to functional sodium channels: voltage-clamp recordings measure the sodium current amplitude (the amount of Na☐ ions flowing into neurons), which is restored from ~50% to ~85% of normal. Think of it like this: if a neuron needs 100 sodium channels to work properly, haploinsufficient neurons only have 50, but CRISPRa brings it up to 85—enough to function normally. Each dot represents one neuron or one mouse; the boxes show the median and spread of data. This demonstrates the complete chain: more gene expression → more protein → more functional channels. Source: Tamura et al. (2022). CRISPRa rescues Scn2a haploinsufficient neurons. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.03.30.486483v1. License: CC-BY-NC-ND 4.0.*

34.4.2 Results at the circuit and behavioral level

Measurement	Haploinsufficient (untreated)	After CRISPRa	Outcome
Seizure susceptibility	Elevated (to chemoconvulsants)	Protected	☐ Therapeutic benefit
Synaptic transmission	Impaired excitatory input	Restored	☐ Circuit-level rescue
Behavioral phenotype	Disease state	Improved	☐ Behavioral rescue
Overall phenotype	Disease state	Near-normal	☐ Compre-hensive rescue

- Mice treated with CRISPRa showed protection against chemoconvulsant-induced seizures (4-aminopyridine challenge)
- Synaptic deficits in neocortical pyramidal cells were corrected
- Behavioral improvements were observed
- Overall, the mice showed phenotypic rescue even when treatment started in adolescence

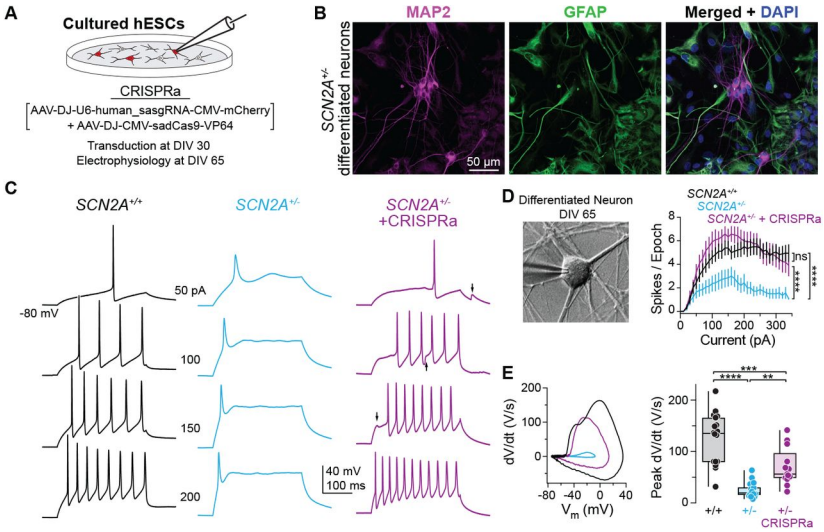


Figure 66: Tamura et al. 2022 bioRxiv – Figure 4

Figure: CRISPRa protects against chemically-induced seizures. Mice were injected with 4-aminopyridine (4-AP), a drug that triggers seizures by blocking potassium channels—this tests how vulnerable the brain is to seizure activity. The graph shows survival curves over 60 minutes after 4-AP injection. Wild-type mice (WT, black line) survive well because their neurons have enough sodium channels to maintain proper excitatory-inhibitory balance. Haploinsufficient mice (Scn2a□/□, cyan line) are much more vulnerable—about 60% die from severe tonic-clonic seizures within an hour. But remarkably, CRISPRa-treated haploinsufficient mice (Scn2a□/□ + CRISPRa, purple line) are strongly protected: their survival curve is nearly identical to wild-type. This isn't just a lab measurement—it demonstrates that CRISPRa treatment provides real, clinically meaningful protection against seizures, the most dangerous symptom of SCN2A haploinsufficiency in human patients. The dramatic difference between cyan and purple lines shows how effective dosage restoration can be. Source: Tamura et al. (2022). CRISPRa rescues Scn2a haploinsufficient neurons. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.03.30.486483v1. License: CC-BY-NC-ND 4.0.

34.4.3 Results in human cells

The study also tested CRISPRa in human iPSC-derived neurons carrying SCN2A haploinsufficiency. After transduction with the CRISPRa system, SCN2A expression increased, sodium currents improved, and the neurons regained normal excitability.

This is important because it shows the approach isn't mouse-specific—it works in human cells too.

34.5 Why this matters: dosage restoration therapy

This study demonstrates a new therapeutic concept: **dosage restoration therapy**. Instead of adding a new gene or fixing the mutant allele, you upregulate the endogenous normal allele to compensate for the loss.

Here's why that's powerful:

Advantage	Explanation	Clinical Benefit
No DNA cutting	CRISPRa doesn't introduce double-strand breaks	Lower risk of off-target mutations or chromosomal rearrangements
Endogenous regulation	Normal allele remains under native promoter/enhancers	Expression can be modulated by developmental and physiological cues
Tunable and reversible	Effect is transient; stops when dCas9 expression ceases	Can adjust dose over time or discontinue if side effects occur
Mutation-agnostic	Works for any LoF mutation as long as one allele is intact	Single approach for diverse mutation types (nonsense, frameshift, splice)
Preserves native protein	Using patient's own gene, not introducing foreign sequence	No immunogenicity concerns from novel epitopes
Late intervention possible	Works even when treatment starts in adolescence	Therapeutic window extends beyond early development

1. No DNA cutting
CRISPRa doesn't introduce double-strand breaks, so there's less risk of off-target mutations or chromosomal rearrangements. You're not permanently altering the genome—you're just changing how much it's used.

2. Endogenous regulation
The normal SCN2A allele remains under its native promoter and enhancers. It's still regulated by the cell's normal signaling pathways, so expression can be modulated in response to developmental or physiological cues. This is much cleaner than inserting a transgene with a strong constitutive promoter, which might overexpress the protein at the wrong time or in the wrong cells.

3. Tunable and reversible
CRISPRa is transient. If you stop expressing dCas9 and the guide RNA, gene expression drifts back toward baseline. This means you can potentially adjust the dose over time, or turn the system off if there are side effects. Compare this to a permanent gene edit, which is irreversible.

4. Applicable to many haploinsufficiency disorders
SCN2A is just one example. There are hundreds of genes where losing one copy causes disease—SHANK3, CHD8, MECP2, FOXP1, and many others. CRISPRa could, in principle, be applied to any of them, as long as boosting the normal allele is sufficient to restore function.

5. Late intervention is possible
Remarkably, the study showed that CRISPRa treatment works even when started in adolescent mice—long after early brain development. This suggests a therapeutic window that extends beyond the critical early developmental period, offering hope for patients who weren't diagnosed until later.

34.6 How does this compare to other approaches?

Let's put CRISPRa in context by comparing it to alternative strategies for treating SCN2A haploinsufficiency.

Strategy	Mechanism	Advantages	Disadvantages	Best For
Gene replacement therapy	AAV delivers functional SCN2A transgene	Well-established method, no need for precise targeting	Loss of endogenous regulation, AAV size limits (~4.7kb, SCN2A is ~6kb), potential overexpression	Diseases with small genes
Base editing	Direct correction of point mutation	Fixes root cause, restores mutant allele to WT	Requires delivery to all cells, off-target editing risk, works only for specific mutation types	Single-nucleotide mutations
Prime editing	Corrects small indels or point mutations	More flexible than base editing	Large editor size, delivery challenges, efficiency varies	Small insertions/deletions
Small molecule drugs	Pharmacological modulation of channel activity	Non-invasive, reversible, can cross BBB	Non-specific (affects all sodium channels), dosing challenges, limited efficacy	Gain-of-function variants (e.g., NBI-921352 for SCN8A)
CRISPRa	Boost expression of normal allele	No DNA cutting, preserves regulation, tunable, mutation-agnostic, works in adolescence	Requires viral delivery, large payload (dCas9 ~4kb), potential off-target activation, long-term persistence unknown	Haploinsufficiency with intact second allele

34.6.1 Gene replacement therapy

This is the classic approach: use a virus (like AAV) to deliver a functional copy of SCN2A into cells. The new gene integrates randomly or remains as an episome, and produces NaV1.2 protein.

Pros: Well-established method, doesn't require targeting a specific genomic locus
Cons: The transgene isn't under endogenous regulation, so you risk overexpression or expression in the wrong cell types. AAV has packaging size limits (~4.7 kb), and SCN2A is about 6 kb, so you'd need a smaller promoter or a split-AAV approach.

34.6.2 Base editing or prime editing

If the mutation is a point mutation or small indel, you could use base editors or prime editors to directly correct it.

Pros: Fixes the root cause, restoring the mutant allele to wild-type
Cons: Requires delivery of the editor plus guide RNA to every affected cell. Off-target editing is a concern. Works only if the mutation is correctable (not all LoF variants are single-base changes).

34.6.3 Small molecule drugs

Some sodium channelopathies can be treated with drugs that modulate channel activity.

Pros: Non-invasive, reversible, can cross the blood-brain barrier
Cons: Drugs typically act on all sodium channels, not just NaV1.2, leading to off-target effects. Hard to achieve precise dosage.

34.6.4 CRISPRa

Pros: Boosts the normal allele without cutting DNA, preserves endogenous regulation, tunable, works regardless of mutation type (as long as one allele is intact), effective even with late intervention
Cons: Requires viral delivery, dCas9 is large (~4 kb) so packaging can be challenging, potential for off-target activation if guides aren't specific, long-term persistence and safety need evaluation

Each approach has trade-offs. CRISPRa shines when you have a haploinsufficiency disorder with an intact second allele and you want to avoid permanent genome edits.

34.7 What does "rescue" mean?

You'll often see the term **"rescue"** in genetics papers. It's worth clarifying what it means.

In molecular biology, a rescue experiment tests whether restoring a missing function can reverse a mutant phenotype. If you knock out a gene and observe a defect, then you add back the normal gene and the defect disappears—that's a rescue. It's strong evidence that the gene is causally responsible for the phenotype.

```
Logic of Rescue Experiments
===========================

Step 1: Establish mutant phenotype
Genotype: Loss-of-function mutation
Phenotype: Disease state observed
Conclusion: Gene disruption → phenotype

Step 2: Restore function
Intervention: Add back functional gene/protein
                OR
           Boost remaining normal allele

Step 3: Measure outcome
If phenotype disappears → RESCUE (gene is causal)
If phenotype persists → NOT RESCUED (gene not sufficient)

Example in SCN2A study:

Mutation:      Scn2a haploinsufficiency
Phenotype:     Weak neuronal firing + seizure susceptibility
Intervention:  CRISPRa to boost normal Scn2a allele
Result:        Normal firing + seizure protection
Conclusion:      RESCUED → Proves SCN2A dosage
               is the primary problem
```

In the study by Tamura and colleagues, "rescue" means:

- Neurons that were firing weakly (because of SCN2A haploinsufficiency) now fire normally (after CRISPRa upregulates SCN2A)
- Mice that had increased seizure susceptibility now show protection against chemoconvulsants
- Synaptic and behavioral deficits improved

- The phenotype was reversed by restoring gene dosage

So "rescue" isn't just a metaphor—it's experimental proof that boosting SCN2A expression is sufficient to fix the problem.

34.8 Looking ahead: cis-regulatory gene therapy

The study by Tamura and colleagues is part of a broader trend toward **cis-regulatory gene therapy**—treatments that work by modulating the activity of a patient's own genes rather than adding new genetic material (Tamura et al. 2025, Nature).

```
Cis-Regulatory Gene Therapy Landscape
=====================================

Approach: Modulate endogenous gene expression
Goal: Restore balance without introducing foreign DNA

Current Applications:

  CRISPRa (Activation)
  • SCN2A haploinsufficiency (Tamura 2025)
  • SCN1A in Dravet syndrome (multiple studies)
  • MECP2 insufficiency (Rett syndrome)
  • Tumor suppressor reactivation
  • SIM1 in obesity (Matharu 2019)

  CRISPRi (Inhibition)
  • Huntington's disease (silence HTT repeat)
  • Toxic gain-of-function mutations
  • Oncogene silencing in cancer

  Epigenetic Editors
  • DNA methylation editing
  • Histone modification editing
  • Chromatin remodeling


Principle: Adjust gene expression to restore balance
• Too little expression? → CRISPRa
• Too much expression? → CRISPRi
• Wrong epigenetic state? → Epigenetic editors
```

Other examples include:

- Using CRISPRa to boost SCN1A expression in Dravet syndrome (Colasante et al. 2020, Molecular Therapy)
- Using CRISPRa to boost SIM1 in obesity caused by haploinsufficiency (Matharu et al. 2019, Science)
- Using CRISPRi to silence toxic repeat expansions in Huntington's disease
- Using epigenetic editors to reactivate silenced tumor suppressor genes in cancer

The unifying principle is: adjust gene expression to restore balance. Sometimes that means turning a gene up (CRISPRa). Sometimes it means turning a gene down (CRISPRi). Either way, you're fine-tuning the system rather than rewriting it.

CRISPRa for SCN2A haploinsufficiency is a proof-of-concept that this strategy can work for neurological

disorders caused by gene dosage imbalance. The next steps—moving toward clinical trials, optimizing delivery, ensuring long-term safety—are already underway.

And if it works for SCN2A, it might work for dozens of other haploinsufficiency disorders too.

34.9 Key takeaways

- **Haploinsufficiency** means one gene copy isn't enough—neurons need more protein than a single allele can provide
- **CRISPRa** uses dead Cas9 fused to activators to boost gene expression without cutting DNA
- Boosting SCN2A expression restores neuronal firing and protects against seizures in mice (Tamura et al. 2025, Nature)
- This approach is **tunable, reversible, and preserves endogenous regulation**—advantages over traditional gene therapy
- **Treatment works even when started in adolescence**, suggesting a broad therapeutic window
- **Cis-regulatory therapy** represents a new paradigm: instead of replacing or editing genes, adjust their expression to restore balance

---

34.10 Key terms

- **Loss-of-function (LoF)**: A mutation that reduces or eliminates protein activity
- **Gain-of-function (GoF)**: A mutation that increases or alters protein activity inappropriately
- **Haploinsufficiency**: A condition where one functional gene copy is not enough for normal function
- **CRISPRa (CRISPR activation)**: A system using dead Cas9 to increase gene transcription
- **CRISPRi (CRISPR interference)**: A system using dead Cas9 to decrease gene transcription
- **dCas9 (dead Cas9)**: Catalytically inactive Cas9 that binds DNA but doesn't cut
- **Rescue**: An experiment showing that a mutant phenotype can be reversed by restoring the missing function
- **Cis-regulatory therapy**: Treatment that modulates gene expression by targeting regulatory sequences
- **Dosage restoration therapy**: Therapeutic strategy that upregulates the normal allele to compensate for haploinsufficiency
- **VPR**: A transcriptional activator domain (VP64-p65-Rta fusion) used in CRISPRa

35 Chapter 28. Gene Regulation: How Cells Read the Same Book Differently

Your liver cells and your neurons have exactly the same DNA. So do your immune cells, your skin cells, and every other cell in your body. They all carry the same 20,000 protein-coding genes. Yet a liver cell looks nothing like a neuron, and a T cell behaves completely differently from a muscle cell.

How is this possible? The answer is **gene regulation**—the process that controls which genes are turned on or off in each cell type. Think of the genome as a massive library. Every cell has access to the same collection of books, but each cell type reads a different subset. Neurons read the synapse chapter. Liver cells read the metabolism chapter. Immune cells read the defense chapter.

Gene regulation determines what a cell is and what it does. It's not just about whether a gene is on or off—it's about how much, when, and where. And understanding gene regulation is the key to understanding how genetic variants affect biology, because most disease-associated variants don't change protein sequences—they change when and how much genes are expressed.

35.1 The central dogma, revisited

The classic central dogma says: DNA → RNA → protein. That's true, but incomplete. The interesting question isn't can DNA be transcribed into RNA—it's is it transcribed, and how much?

Gene expression has multiple control points:

```
Gene Expression Control Points
==============================

DNA (Genome)
    ↓
1. TRANSCRIPTIONAL REGULATION  (Primary control)

      Promoters (local control)
      Enhancers (long-range control)
      Silencers (repression)
      Chromatin state (accessibility)
    ↓
pre-mRNA
    ↓
2. RNA PROCESSING

      Alternative splicing (exon selection)
      5' capping
      3' polyadenylation
      RNA editing
    ↓
mature mRNA
    ↓
3. mRNA STABILITY

      3' UTR regulatory elements
      RNA-binding proteins
      microRNA targeting
    ↓
mRNA in cytoplasm
    ↓
4. TRANSLATION

      Ribosome recruitment
      Translation initiation factors
      Codon usage
    ↓
Protein
    ↓
5. PROTEIN STABILITY

      Post-translational modifications
      Ubiquitination (degradation signal)
      Protein folding/chaperones
    ↓
Functional Protein
```

```
= Most fundamental control point
   (If not transcribed, nothing downstream matters)
```

The first step—transcriptional regulation—is the most fundamental. If a gene isn't transcribed, nothing downstream matters. And transcriptional regulation is mostly controlled by two things: **promoters** (local control) and **enhancers** (long-range control).

35.2 Promoters: where transcription starts

A **promoter** is the DNA region immediately upstream of a gene where RNA polymerase II and its associated transcription factors assemble to begin transcription. Think of it as the on-ramp to the gene.

Promoters contain short DNA sequences (motifs) that serve as binding sites for the transcriptional machinery:

Promoter Element	Location	Sequence	Function	Notes
TATA box	~30 bp upstream of TSS	TATAAA	Recruits TATA-binding protein (TBP)	Classic element, but not universal (~10–20% of genes)
Initiator (Inr)	Overlaps TSS	PyPyA+1NT/APyPy	Positions RNA polymerase	Found in TATA-less promoters
CpG islands	−200 to +200 bp around TSS	High CG content	Marks housekeeping genes	Usually unmethylated when active
DPE (Downstream Promoter Element)	+28 to +32 bp	RGWYV	Alternative to TATA box	Common in Drosophila
BRE (TFIIB Recognition Element)	Flanking TATA box	G/C-rich	TFIIB binding site	Modulates transcription level

For transcription to start, RNA polymerase II doesn't just land on the DNA and start copying. It needs help from a set of **general transcription factors (GTFs)**—proteins like TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. Together, they form the **pre-initiation complex (PIC)** at the promoter:

```
Pre-Initiation Complex Assembly
================================

Step 1: TFIID (contains TBP) binds TATA box
    ↓
Step 2: TFIIB binds, recognizing BRE elements
    ↓
Step 3: RNA Polymerase II + TFIIF arrive
    ↓
Step 4: TFIIE joins the complex
    ↓
Step 5: TFIIH arrives (has helicase activity)
    ↓
```

```
Step 6: TFIIH unwinds DNA double helix
    ↓
Step 7: RNA Pol II begins transcription
    ↓
Step 8: PROMOTER-PROXIMAL PAUSING (~20-60 bp downstream)

      Paused Pol II waits for signal
      P-TEFb phosphorylates Pol II CTD
      Pause released → productive elongation
    ↓
Transcription proceeds into gene body
```

Interestingly, RNA polymerase often pauses shortly after it starts transcribing—about 20–60 base pairs downstream of the TSS. This **promoter-proximal pausing** is a regulatory checkpoint. Paused polymerase is like a runner in the starting blocks, ready to go but waiting for a signal. Factors like P-TEFb can release the pause and allow the polymerase to proceed into the gene body. This pause-release mechanism allows rapid gene activation in response to signals—important for genes like FOS and JUN that need to be turned on quickly during stress or immune responses.

35.3 Enhancers: long-range control

Promoters are local regulators—they sit right next to the gene. But most of the interesting regulatory complexity comes from **enhancers**—DNA elements that can be located hundreds of thousands of base pairs away from the genes they control.

Enhancers are short (50–1500 bp) sequences that serve as binding platforms for transcription factors (TFs). When multiple TFs bind to an enhancer, they recruit co-activator proteins—like the **Mediator complex** and histone acetyltransferases like **p300/CBP**—that physically loop through space to contact the promoter and boost transcription.

35.3.1 Key properties of enhancers

Property	Description	Example
Distance-independent	Can be 1 Mb away, upstream/downstream/intronic	SHH gene regulated by ZRS enhancer >1 Mb away
Orientation-independent	Work in both forward and reverse orientation	Unlike promoters which are directional
Combinatorial logic	Require multiple TFs binding simultaneously	Muscle enhancers: MYOD + MEF2 + SRF
Cell-type specific	Active only in cells expressing right TF combination	Neuron enhancers: NEUROD + REST + SOX
H3K27ac marked	Active enhancers have this histone acetylation	Universal signature across cell types
H3K4me1 marked	All enhancers (active and poised) have this mark	Distinguishes enhancers from promoters
Produce eRNAs	Active enhancers transcribe short enhancer RNAs	Bidirectional, unstable transcripts
Rapidly evolving	Sequence changes faster than coding regions	50–60% human–mouse similarity
Functionally conserved	Important developmental enhancers deeply conserved	HOX enhancers similar across vertebrates

35.3.2 They work at a distance

An enhancer can be 1 megabase away from its target gene. It doesn't matter if it's upstream, downstream, or even inside an intron. As long as it can physically contact the promoter via chromatin looping, it can activate transcription.

For example, the SHH (sonic hedgehog) gene—essential for limb and brain development—is regulated by an enhancer called ZRS that's located over 1 million base pairs away. Mutations in ZRS cause polydactyly (extra fingers) in humans because the enhancer fails to properly activate SHH in developing limbs.

35.3.3 They integrate multiple signals

Enhancers typically bind multiple transcription factors simultaneously. This creates **combinatorial logic**: the enhancer is only active when the right combination of factors is present.

```
Enhancer Combinatorial Logic Example: Muscle Gene
==================================================

Enhancer DNA: [MYOD site] [MEF2 site] [SRF site]
                    ↓           ↓          ↓

In Muscle Cells (all 3 factors present):
    MYOD + MEF2 + SRF → Enhancer ACTIVE
              ↓
    Recruit Mediator + p300
              ↓
    Chromatin loops to promoter
              ↓
    Gene TRANSCRIBED

In Liver Cells (factors absent):
    No MYOD, No MEF2, No SRF → Enhancer INACTIVE
              ↓
    No looping
              ↓
    Gene SILENT

This is why the same DNA produces different cell types:
• Same genome
• Different transcription factor combinations
• Different active enhancers
• Different gene expression programs
```

For example, muscle-specific enhancers require binding of **MYOD**, **MEF2**, and **SRF**. Neuron-specific enhancers need **NEUROD**, **REST**, and **SOX** factors. This combinatorial binding allows precise spatial and temporal control—muscle enhancers are active only in muscle cells because only muscle cells express the right combination of TFs.

35.3.4 They're marked by specific histone modifications

Active enhancers are marked by **H3K27ac** (histone H3 lysine 27 acetylation) and **H3K4me1** (histone H3 lysine 4 monomethylation). These are molecular signatures that say "this is an active regulatory region." Chromatin profiling studies can identify thousands of enhancers across the genome by looking for regions with these marks.

Interestingly, many active enhancers also produce short RNAs called **enhancer RNAs (eRNAs)**. These are

often unstable and don't code for proteins, but their production correlates with enhancer activity. Whether eRNAs are just a byproduct or actually help activate transcription is still debated.

35.3.5 They evolve rapidly but functionally important ones are conserved

Enhancer sequences evolve much faster than protein-coding genes. You can align human and mouse coding sequences and see 85–90% identity, but many enhancers are only 50–60% similar. This makes sense: changing an enhancer might alter when or where a gene is expressed without breaking the protein itself.

Despite this rapid evolution, some enhancers are deeply conserved across hundreds of millions of years. These **conserved noncoding elements (CNEs)** usually regulate essential developmental genes. For example, enhancers controlling HOX genes (which define body segment identity) are remarkably similar across vertebrates—fish, mice, and humans share recognizable HOX enhancers.

At the same time, species-specific enhancers drive evolutionary diversity. Changes in enhancer activity can rewire gene regulatory networks without changing proteins. This is thought to underlie much of human brain evolution—not new genes, but new patterns of gene expression driven by enhancer changes.

35.4 How enhancers contact promoters: 3D chromatin looping

Enhancers are far away in linear DNA sequence, but they get physically close to promoters through **chromatin looping**. DNA isn't a straight line in the nucleus—it's folded into complex 3D structures.

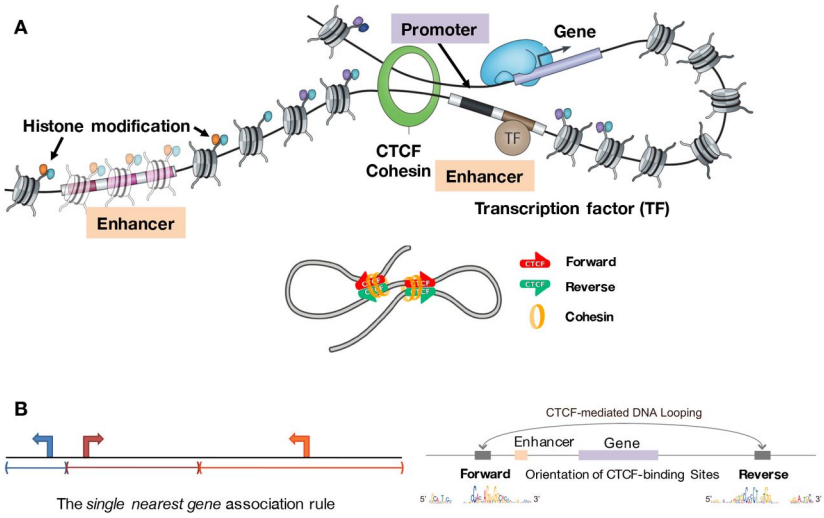


Figure 67: Osato 2019, bioRxiv

Figure: CTCF motif orientation determines chromatin loop formation. Panel A shows how enhancers and promoters are brought together through CTCF- and cohesin-mediated chromatin looping. The cohesin complex (orange ring) physically embraces DNA and extrudes loops until it encounters CTCF proteins (green = reverse orientation; red = forward orientation) bound at specific genomic sites. Transcription factors (TF, brown) bind to enhancers, which are marked by specific histone modifications (colored dots on nucleosomes). When an enhancer and promoter are brought into proximity by looping, gene transcription is activated. The lower schematic demonstrates that CTCF orientation matters: loops form preferentially when CTCF binding sites are in convergent orientation (forward facing reverse, indicated by red-green arrows pointing toward each other). Panel B illustrates the "single nearest gene association rule" showing how CTCF boundaries define regulatory domains. In the linear genome view (left), genes may be

equidistant from an enhancer. However, the 3D chromatin structure (right) determines function: convergently oriented CTCF sites create a loop that brings the enhancer to a specific gene promoter, while genes outside this loop are excluded from regulation by that enhancer. The CTCF motif sequences (shown at bottom) reveal the asymmetric DNA-binding preference that enables directional loop formation. This convergent orientation rule is fundamental to proper genome organization—over 90% of chromatin loops occur at convergent CTCF sites. Source: Osato, D. (2019). Discovery of biased orientations of human DNA motif sequences affecting enhancer–promoter interactions and transcription of genes. bioRxiv 290825. https://www.biorxiv.org/content/10.1101/290825v12.full. License: CC-BY 4.0.

```
Components of Chromatin Architecture:

CTCF: DNA-binding protein
• Acts as insulator and loop anchor
• Recognizes specific asymmetric DNA sequence motifs
• ~40,000-60,000 sites per genome
• ORIENTATION MATTERS: N-terminus blocks cohesin
• Convergent orientation (><) creates stable loops

Cohesin: Ring-shaped protein complex
• Physically holds DNA loops together
• Extrudes loops bidirectionally until stopped by CTCF
• Can slide along DNA ("loop extrusion")
• Stopped by CTCF's N-terminus (directional blocking)

TADs (Topologically Associating Domains):
• Size: 100 kb to few Mb
• Function: Insulated neighborhoods
• Enhancers contact promoters WITHIN same TAD
• Rare interactions ACROSS TAD boundaries
• Bordered by convergent CTCF sites (>90% of cases)

The Convergent Orientation Rule:
• >< (convergent) = LOOP FORMS (cohesin stops at both ends)
• <> (divergent) = NO LOOP (cohesin passes through both)
• >> or << (tandem) = WEAK LOOP (cohesin stops at one end)
```

The main architectural proteins involved are:

- **CTCF**: A DNA-binding protein that acts as an insulator and anchor point for chromatin loops
- **Cohesin**: A ring-shaped protein complex that holds DNA loops together

Together, CTCF and cohesin create loop structures that bring enhancers and promoters into contact. These loops occur within larger genomic neighborhoods called **Topologically Associating Domains (TADs)**—regions of the genome (typically 100 kb to a few megabases) where DNA interactions are frequent. Enhancers and promoters within the same TAD can contact each other, but interactions across TAD boundaries are rare.

This structure prevents enhancers from accidentally activating the wrong genes. For instance, if an enhancer controlling a developmental gene is in TAD A, and an oncogene is in TAD B, the enhancer won't activate the oncogene—they're in separate neighborhoods.

But sometimes things go wrong. **Enhancer hijacking** occurs when a chromosomal rearrangement (like a translocation or deletion) brings an enhancer into a new TAD:

```
Enhancer Hijacking in Cancer
============================
```

```
NORMAL:
TAD A: [Enhancer] controls [Developmental Gene]
                              (appropriate target)
TAD B:                        [MYC oncogene]
                              (normally low expression)

AFTER CHROMOSOMAL REARRANGEMENT:
TAD A:                        [Developmental Gene]
                              (lost control, silenced)
Broken TAD: [Enhancer]    [MYC oncogene]
            (super-enhancer)  (massive overexpression)
                                     ↓
                              T-cell leukemia

Example: T-ALL (T-cell Acute Lymphoblastic Leukemia)
• t(8;14) translocation
• Brings T-cell receptor enhancer near MYC
• MYC overexpression drives cancer
```

A classic example: in some T-cell leukemias, a rearrangement brings a super-enhancer near the MYC oncogene, causing massive overexpression and driving cancer.

35.4.1 A real example: IGF2/H19 imprinting

The IGF2/H19 locus provides a beautiful example of how 3D chromatin architecture controls gene expression (see Figure above). This locus contains two genes—IGF2 (insulin-like growth factor 2) and H19 (a long noncoding RNA)—that are regulated by genomic imprinting, meaning they're expressed differently depending on whether they came from mom or dad.

Both genes share the same distal enhancer, but CTCF-mediated looping determines which gene gets activated:

Paternal chromosome (IGF2 ON, H19 OFF): - The DMR (differentially methylated region) between the genes is methylated - CTCF cannot bind to methylated DNA - Without CTCF insulation, the enhancer loops directly to the IGF2 promoter - IGF2 is expressed, H19 is silenced

Maternal chromosome (IGF2 OFF, H19 ON): - The DMR is unmethylated - CTCF binds and creates an insulator boundary - This blocks the enhancer from reaching IGF2 - The enhancer instead loops to the H19 promoter - H19 is expressed, IGF2 is silenced

This demonstrates a crucial principle: **the same DNA sequence can produce opposite outcomes depending on 3D chromatin structure**. The enhancer's activity doesn't change—what changes is which promoter it can physically contact.

Chromatin loops are dynamic. They form and dissolve in response to developmental cues or environmental signals. For example, when neurons are activated, enhancer-promoter loops rapidly form to turn on immediate early genes like FOS. This allows fast transcriptional responses without having to rebuild the entire regulatory architecture.

Figure: 3D chromatin architecture brings distant regulatory elements together at the IGF2/H19 locus. The imprinted IGF2/H19 locus illustrates how chromatin loops create functional regulatory neighborhoods. In the linear genome (bottom), the IGF2 and H19 promoters (dark blue) are separated by a differentially methylated region (DMR, red insulator) and a distal enhancer (light blue). The 3D organization differs between parental alleles: On the paternal allele (left loop), the DMR is methylated and CTCF cannot bind, inactivating the insulator. This allows the enhancer to loop to and activate the IGF2 promoter through cohesin-mediated contacts (indicated by Mediator and Nipbl-cohesin complexes, yellow dots). On the maternal allele (right loop), the unmethylated DMR binds CTCF, which blocks enhancer-IGF2 communica-

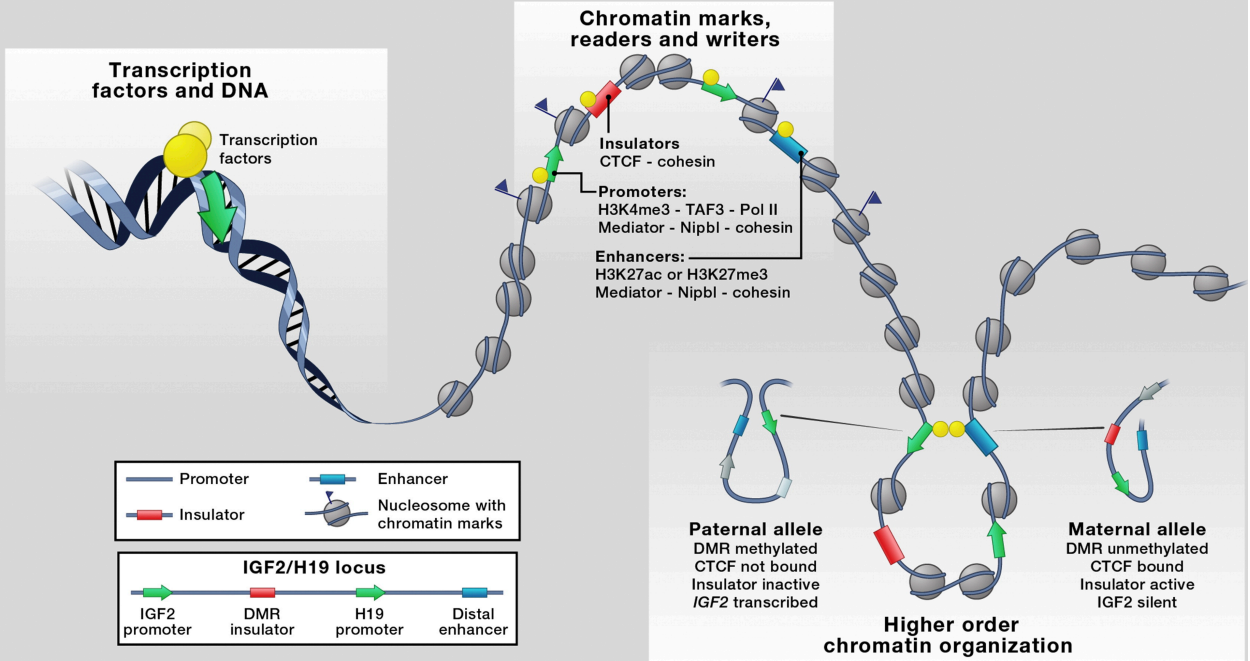


Figure 68: Merkenschlager & Odom 2013, Cell

tion. Instead, CTCF creates an insulated boundary that redirects the enhancer to activate H19, silencing IGF2. The loops are stabilized by cohesin complexes (gray circles representing nucleosomes with chromatin marks) that physically hold enhancers and promoters together in 3D space. This demonstrates how chromatin architecture—not just DNA sequence—determines which genes are expressed. Source: Merkenschlager, M. & Odom, D.T. (2013). CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. Cell 152(6):1285–1297.

35.5 Key takeaways

- **Cell identity is defined by gene expression, not DNA sequence**—every cell has the same genome, but different cells express different genes
- **Promoters** control local transcription initiation, while **enhancers** provide long-range, combinatorial regulation
- Enhancers work through **chromatin looping**, bringing distant regulatory elements into physical contact with promoters
- **TADs** (Topologically Associating Domains) create genomic neighborhoods that constrain enhancer–promoter interactions
- **Histone modifications** mark active enhancers (H3K27ac + H3K4me1) and distinguish them from promoters
- Enhancers use **combinatorial logic**—requiring multiple transcription factors to bind simultaneously for activation
- **Enhancer hijacking** through chromosomal rearrangements can cause cancer by activating oncogenes

---

35.6 Key terms

- **Gene regulation**: Control of gene expression in response to developmental, environmental, or cellular cues
- **Promoter**: DNA region upstream of a gene where transcription initiates
- **Enhancer**: Distal regulatory element that increases gene expression, often from great distances
- **Chromatin looping**: 3D folding that brings distant DNA elements into physical contact
- **CTCF**: CCCTC-binding factor—architectural protein that creates chromatin loop anchors
- **Cohesin**: Ring-shaped protein complex that holds chromatin loops together
- **TAD (Topologically Associating Domain)**: Genomic neighborhood where chromatin interactions are frequent
- **H3K27ac**: Histone modification marking active enhancers and promoters
- **H3K4me1**: Histone modification marking all enhancers (active and poised)
- **eRNA**: Enhancer RNA—short transcripts produced from active enhancers
- **Combinatorial logic**: Requirement for multiple transcription factors to activate an enhancer
- **Pre-initiation complex (PIC)**: Assembly of RNA polymerase II and general transcription factors at promoters

36 Chapter 29. Gene Regulation: Methods and Applications

In the previous chapter, we learned how cells with identical genomes create different identities through gene regulation—using promoters, enhancers, and 3D chromatin architecture. Now we'll explore how scientists actually measure and study these regulatory mechanisms in living cells.

This chapter covers the experimental toolkit that has revolutionized our understanding of gene regulation: methods to measure what genes are expressed, where regulatory elements are located, and which regions of chromatin are accessible. These technologies have transformed genomics from reading static DNA sequences to understanding dynamic gene regulation across cell types, developmental stages, and disease states.

36.1 Measuring gene expression: RNA-seq

If gene regulation determines what a cell is, then measuring gene expression tells you what's actually being produced. The workhorse method for this is **RNA sequencing (RNA-seq)**.

Here's how it works:

```
RNA-seq Workflow
================

Step 1: RNA Extraction
  Isolate total RNA from cells/tissue
  Can select: total RNA, mRNA (polyA+), or specific RNAs
  Quality check (RIN score for RNA integrity)

Step 2: Library Preparation
  Reverse transcribe RNA → cDNA
  Fragment cDNA to 200-500 bp
  Add sequencing adapters (for cluster generation)
  PCR amplification (if needed)

Step 3: Sequencing
  Load library onto flow cell
```

```
  Cluster generation (bridge amplification)
  Sequence-by-synthesis
  Generate millions of short reads (50-150 bp)

Step 4: Bioinformatics
  Quality control (FastQC)
  Align reads to reference genome (STAR, HISAT2)
  Count reads per gene (featureCounts, HTSeq)
  Normalize (CPM, FPKM, TPM)

Step 5: Analysis
  Differential expression (DESeq2, edgeR)
  Visualization (heatmaps, volcano plots)
  Pathway enrichment (GO, KEGG)
```

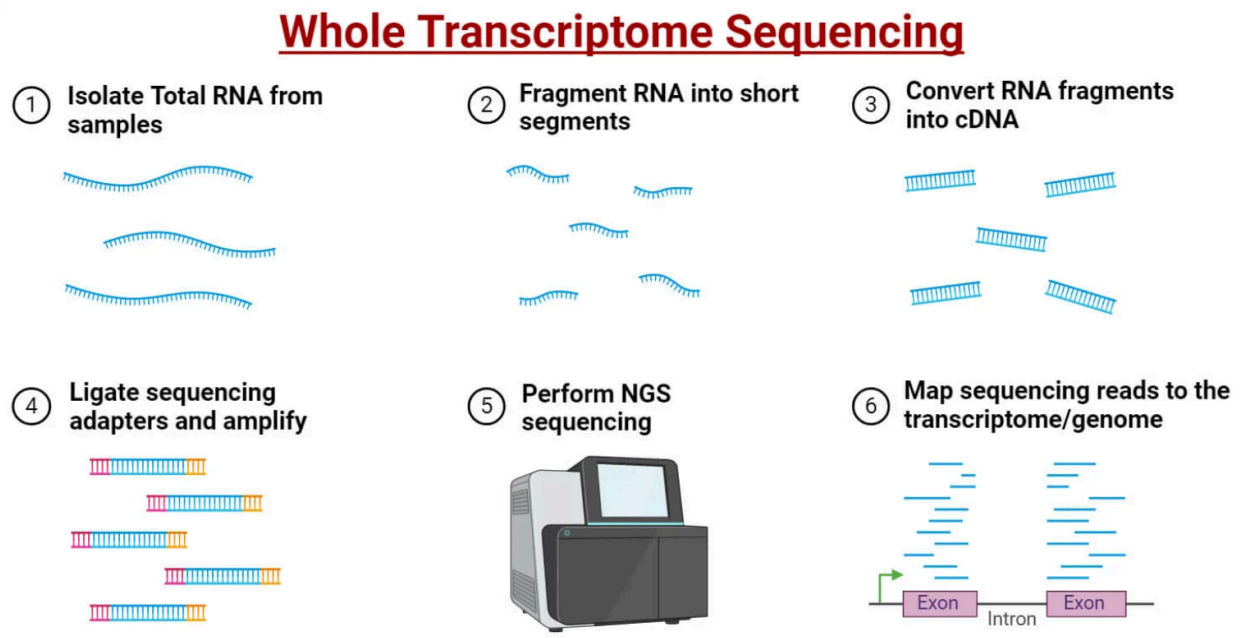


Figure 69: Bulk RNA-seq workflow

Figure: Bulk RNA-seq workflow for measuring gene expression. RNA is extracted from a population of cells (left)—this could be tissue, cultured cells, or sorted cell populations. The RNA is converted to complementary DNA (cDNA) through reverse transcription, then fragmented into short pieces. Sequencing adapters are added to these fragments, allowing them to bind to the flow cell for sequencing. Millions of short sequence reads (typically 50–150 base pairs) are generated and aligned back to the reference genome to identify which genes they came from. The number of reads mapping to each gene reflects its expression level—more reads mean more mRNA, indicating higher gene activity. This gives you an **average** expression profile across all cells in the sample—you can't see differences between individual cells, only the population average. This is powerful for comparing tissues (liver vs. brain) or conditions (healthy vs. diseased), but it masks cellular heterogeneity within the sample. Source: Microbe Notes. Whole Transcriptome Sequencing.

The output is a **count matrix**: for each gene, you get a number representing how many sequencing reads mapped to that gene. More reads = more mRNA = higher expression.

But raw counts aren't directly comparable across genes or samples because: – Longer genes generate more reads (simply because there's more RNA to sequence) – Some samples might have been sequenced

deeper than others

So we normalize the data:

Normalization Method	Formula	What It Corrects	When To Use
CPM (Counts Per Million)	(Counts / Total reads) × 10^6	Sequencing depth	Comparing samples from same experiment
RPKM/FPKM (Reads/Fragments Per Kilobase per Million)	(Counts × 10^9) / (Gene length × Total reads)	Depth + gene length	Comparing genes within one sample
TPM (Transcripts Per Million)	(Counts / Gene length) / (Sum of all Counts/Length) × 10^6	Depth + length (better scale)	Cross-sample comparisons, preferred over FPKM
DESeq2 normalization	Median-of-ratios method	Depth + RNA composition	Differential expression analysis

- **CPM (Counts Per Million)**: Scales counts by total sequencing depth. If sample A has 10 million reads and sample B has 20 million, CPM accounts for that difference.
- **FPKM/RPKM (Fragments/Reads Per Kilobase per Million)**: Normalizes for both sequencing depth and gene length. A 10 kb gene will naturally produce more reads than a 1 kb gene even if they're expressed at the same level, so FPKM corrects for that.
- **TPM (Transcripts Per Million)**: Similar to FPKM but rescaled so that all TPM values in a sample sum to 1 million. This makes cross-sample comparisons more intuitive.

Once you have normalized expression values, you can compare conditions. Tools like **DESeq2** and **edgeR** perform statistical tests to identify **differentially expressed genes (DEGs)**—genes that are significantly up- or downregulated between two conditions (e.g., disease vs healthy, or cell type A vs cell type B).

RNA-seq can also measure **alternative splicing**—the process where different combinations of exons are joined together to create multiple mRNA isoforms from a single gene. Some genes produce dozens of isoforms, each potentially coding for a slightly different protein. By counting reads that span exon junctions, you can quantify which isoforms are present.

36.2 Bulk vs single-cell RNA-seq

Traditional RNA-seq (bulk RNA-seq) sequences RNA from thousands or millions of cells at once. You get an average expression profile for the population. This is great for comparing tissues—liver shows high expression of metabolic genes like ALB (albumin) and CYP3A4, while skeletal muscle shows high expression of structural genes like ACTA1 (actin).

But bulk RNA-seq hides cellular heterogeneity. If you sequence a piece of brain tissue, you're mixing neurons, astrocytes, microglia, oligodendrocytes, and endothelial cells. The resulting profile is an average that doesn't represent any single cell type.

```
Bulk vs Single-Cell RNA-seq
===========================

BULK RNA-SEQ:
Input: 1,000,000 mixed cells
```

```
       ↓
Extract total RNA (average across all cells)
       ↓
Sequence
       ↓
Output: One average profile
       (Hides which cells express which genes)

Example brain tissue result:
• Gene A: 50 TPM ← but which cells express it?
• Gene B: 100 TPM ← neurons? glia? both?
• Can't distinguish cell types

SINGLE-CELL RNA-SEQ:
Input: 1,000,000 mixed cells
       ↓
Isolate individual cells → Barcode each cell
       ↓
Sequence (with cell-specific barcodes)
       ↓
Output: 1,000,000 individual profiles
       (Each cell's transcriptome preserved)

Example brain tissue result:
• Cluster 1 (neurons): Gene A high, Gene B low
• Cluster 2 (astrocytes): Gene A low, Gene C high
• Cluster 3 (microglia): Gene D high
• Can identify 10-50+ cell types/states
```

Single-cell RNA-seq (scRNA-seq) solves this by profiling individual cells. Here's the workflow:

```
Single-Cell RNA-seq Workflow (10x Genomics Droplet Method)
============================================================

Step 1: Cell Suspension
  Dissociate tissue into single cells

Step 2: Droplet Generation
  Encapsulate each cell in oil droplet
  Each droplet contains:
    One cell
    One gel bead with unique barcode
    Lysis buffer + RT reagents
  Result: ~10,000 droplets/run

Step 3: In-Droplet Reactions
  Cell lyses, releasing mRNA
  mRNA captured by oligo-dT on gel bead
  Reverse transcription → cDNA
  Cell barcode + UMI attached to every cDNA molecule

Step 4: Pool and Sequence
  Break droplets, pool all cDNA
  PCR amplification
  Sequence (Illumina)
```

```
  Barcodes trace reads back to cells of origin

Step 5: Computational Analysis
  Demultiplex by cell barcode
  Count UMIs (unique molecular identifiers)
  Quality control (filter empty droplets, doublets)
  Normalization
  Dimensionality reduction (PCA, UMAP)
  Clustering (identify cell types)
  Differential expression (between clusters)

Key Innovation: BARCODING
• Each cell gets unique 16-bp barcode
• Enables multiplexing thousands of cells
• Can trace each sequencing read to its cell
```

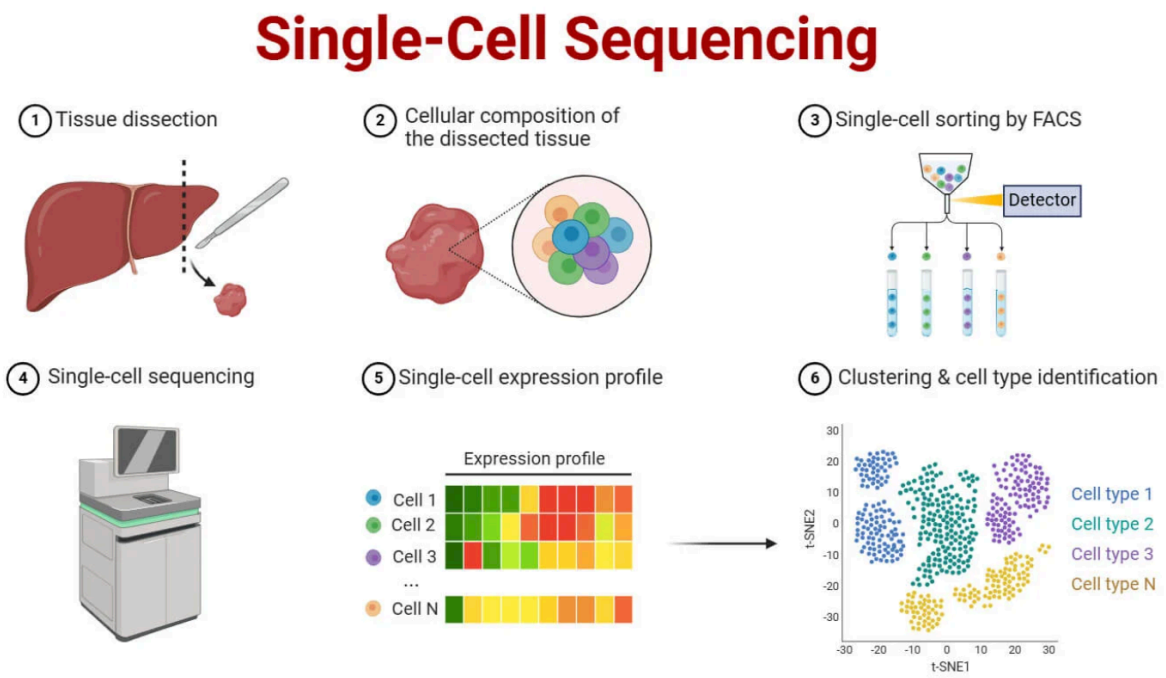


Figure 70: Single-cell RNA-seq workflow

Figure: Single-cell RNA-seq captures individual cellular identities. Unlike bulk RNA-seq, single-cell approaches isolate individual cells before sequencing. In droplet-based methods (like 10x Genomics, shown here), each cell is encapsulated in a tiny oil droplet along with a gel bead carrying a unique barcode sequence. When the cell lyses inside the droplet, its mRNA molecules are captured and tagged with that cell's specific barcode during reverse transcription. This means that even though all cells are pooled together for sequencing, every sequencing read can be traced back to its cell of origin using the barcode. The result is a **cell-by-gene matrix** where each row represents one cell's complete transcriptome. This reveals cellular heterogeneity that's completely hidden in bulk RNA-seq—you can discover rare cell types, transitional states, and cell-to-cell variability that bulk sequencing averages away. For example, in brain tissue, single-cell RNA-seq can distinguish dozens of neuronal subtypes, glial populations, and vascular cells that all look identical when bulk-sequenced. Source: Microbe Notes. Single Cell Sequencing.

The output is a **cell-by-gene matrix**: each row is a cell, each column is a gene, and each value is the expression level in that cell.

By clustering cells with similar expression profiles, you can identify distinct cell types. For example, scRNA–seq of blood reveals T cells (high CD3, CD4, CD8), B cells (high CD19, MS4A1), monocytes (high CD14, LYZ), and NK cells (high NCAM1, NKG7).

In the brain, scRNA–seq distinguishes excitatory neurons (high SLC17A7), inhibitory neurons (high GAD1), astrocytes (high GFAP), oligodendrocytes (high MBP), and microglia (high CX3CR1). All these cells have the same genome, but they express completely different gene sets.

Single–cell data are often visualized using dimensionality reduction—methods like **UMAP** or **t–SNE** that project thousands of genes down to 2D space while preserving similarities. Cells that are transcriptionally similar cluster together, revealing the diversity of cell states within a tissue.

The takeaway: **cell identity is defined by gene expression, not by DNA sequence**. What makes a neuron a neuron isn't a unique genome—it's the specific combination of genes that are active at any given time.

36.3 Measuring gene regulation: chromatin profiling

RNA–seq tells you what is expressed. But to understand how expression is controlled, you need to measure the regulatory landscape—where transcription factors bind, which regions of chromatin are open, and which genes are wrapped in repressive or active histone marks.

36.3.1 Chromatin states and histone modifications

DNA in eukaryotic cells is wrapped around histone proteins to form nucleosomes. Each nucleosome is ~147 bp of DNA wound around a protein core made of histones H2A, H2B, H3, and H4. This packaging organizes the genome but also restricts access to transcription factors and RNA polymerase.

Cells control gene expression by modifying histones through chemical tags—acetylation, methylation, phosphorylation, and others. These modifications recruit different proteins that either open chromatin (making it accessible) or compact it (silencing genes).

Histone Mark	Location	Function	Chromatin State	Biological Meaning
H3K4me3	Promoters	Active transcription start site	Open, active	Gene is being transcribed
H3K27ac	Promoters & Enhancers	Active regulatory regions	Open, active	Enhancer/promoter is ON
H3K4me1	Enhancers	Poised or active enhancers	Variable	Marks all enhancers (add H3K27ac = active)
H3K36me3	Gene bodies	Elongating transcription	Open	RNA Pol II is actively transcribing
H3K27me3	Regulatory regions	Polycomb repression	Closed, repressed	Developmental genes silenced
H3K9me3	Heterochromatin	Constitutive silencing	Closed, compacted	Permanent repression (centromeres, repeats)
H3K9me2	Facultative heterochro–matin	Reversible silencing	Closed	Temporarily silenced genes
H3K79me2	Gene bodies	Transcription elongation	Open	Active transcription marker

Some key histone marks explained:

H3K4me3 (histone H3, lysine 4, trimethylated): Marks active promoters. If you see a sharp peak of H3K4me3 at a gene's TSS, that gene is likely being transcribed.

H3K27ac (histone H3, lysine 27, acetylated): Marks active enhancers and promoters. Acetylation neutralizes the positive charge on histones, loosening their grip on DNA and making it more accessible. H3K27ac is the signature of active regulatory regions.

H3K4me1 (histone H3, lysine 4, monomethylated): Marks enhancers. By itself, it indicates a "poised" enhancer—one that's ready to be activated but not yet fully active. Add H3K27ac, and it becomes an active enhancer.

H3K36me3 (histone H3, lysine 36, trimethylated): Found in gene bodies during active transcription. It's a mark of elongating RNA polymerase and helps recruit RNA splicing machinery.

H3K27me3 (histone H3, lysine 27, trimethylated): A repressive mark deposited by Polycomb complexes. It silences developmental genes that aren't needed in a particular cell type. For example, in neurons, genes involved in liver development are marked with H3K27me3 to keep them off.

H3K9me3 (histone H3, lysine 9, trimethylated): Marks constitutive heterochromatin—permanently silenced regions like centromeres and repetitive DNA. This is the most repressive chromatin state.

These marks create a **histone code**—a combinatorial system where different patterns of modifications define different chromatin states:

```
Chromatin States Based on Histone Marks
=======================================

ACTIVE PROMOTER:
H3K4me3 (high) + H3K27ac + Open chromatin
→ Gene is being transcribed

ACTIVE ENHANCER:
H3K4me1 + H3K27ac + Open chromatin
→ Enhancer is driving gene expression

POISED ENHANCER:
H3K4me1 + NO H3K27ac + Accessible chromatin
→ Enhancer ready but not yet active

POISED PROMOTER:
H3K4me3 + H3K27me3 (bivalent domain)
→ Developmental gene ready for activation

POLYCOMB-REPRESSED:
H3K27me3 (high) + Closed chromatin
→ Gene silenced but can be reactivated

CONSTITUTIVE HETEROCHROMATIN:
H3K9me3 + HP1 binding + Very closed
→ Permanently silenced (repeats, centromeres)

TRANSCRIBED GENE BODY:
H3K36me3 + H3K79me2
→ RNA Pol II is elongating

INACTIVE/NEUTRAL:
```

```
Low all marks + Closed chromatin
→ Gene not expressed, no regulatory activity
```

36.3.2 ChIP–seq: mapping protein–DNA interactions

ChIP–seq (Chromatin Immunoprecipitation Sequencing) identifies where specific proteins—transcription factors or modified histones—bind to the genome.

Here's the protocol:

```
ChIP-seq Protocol
=================

Step 1: CROSSLINKING
  Add formaldehyde to living cells
  Creates covalent bonds between proteins and DNA
  Freezes protein-DNA interactions in place

Step 2: CELL LYSIS & CHROMATIN FRAGMENTATION
  Lyse cells, isolate nuclei
  Sonicate or use MNase to fragment chromatin
  Result: 200-500 bp DNA fragments with bound proteins

Step 3: IMMUNOPRECIPITATION (IP)
  Add antibody specific to target protein
   (e.g., anti-H3K27ac, anti-CTCF, anti-p53)
  Antibody binds protein of interest
  Pull down with magnetic beads (Protein A/G)
  Wash away unbound fragments

Step 4: REVERSE CROSSLINKS
  Heat sample to 65°C overnight
  Breaks formaldehyde crosslinks
  Protein-DNA complexes dissociate

Step 5: DNA PURIFICATION & SEQUENCING
  Remove proteins (proteinase K)
  Purify DNA
  Build sequencing library
  Sequence (typically 25-50M reads)

Step 6: BIOINFORMATICS
  Align reads to genome
  Call peaks (identify enriched regions)
  Annotate peaks (promoter? enhancer? intergenic?)
  Visualize as genome browser tracks
```

Figure: ChIP–seq maps where proteins bind across the genome. The technique starts with living cells, where proteins are chemically crosslinked to the DNA they're bound to using formaldehyde—this "freezes" the protein–DNA interactions so they survive the subsequent harsh extraction steps. The chromatin is then fragmented into small pieces (typically 200–500 base pairs) using sonication or enzymatic digestion. An antibody specific to your protein of interest—whether that's a transcription factor like p53, a histone modification like H3K27ac, or an architectural protein like CTCF—is added to the fragmented chromatin. The antibody binds to its target protein, and magnetic beads coated with Protein A or G are used to

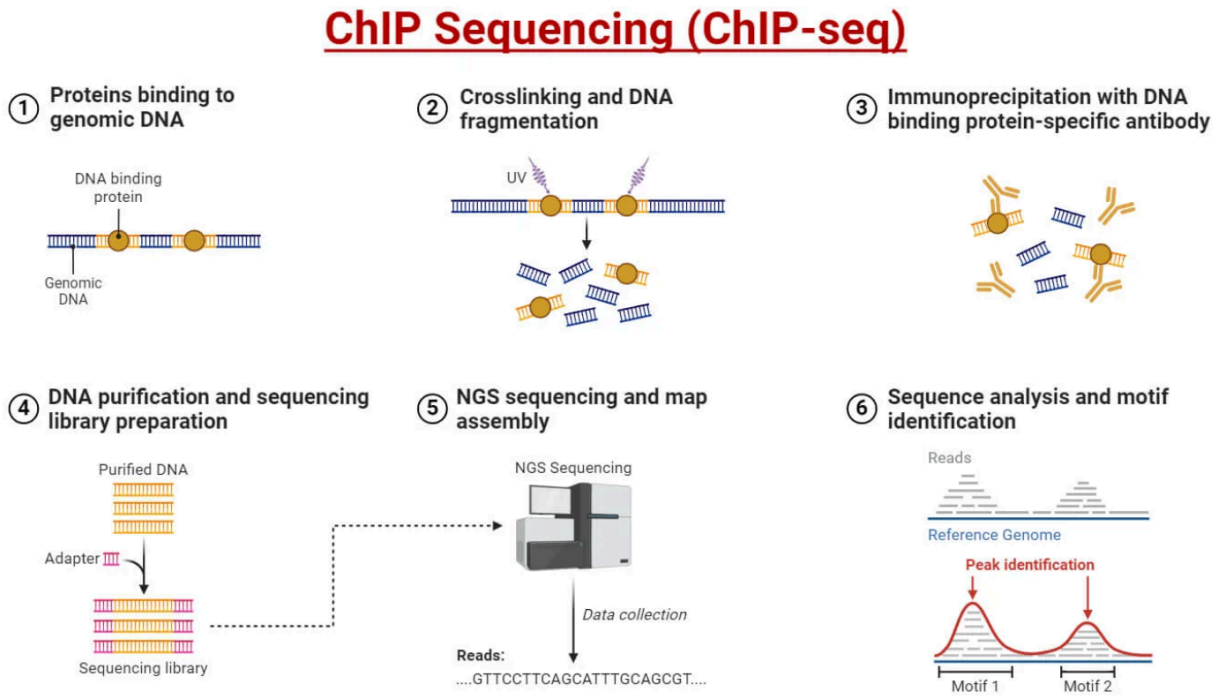


Figure 71: ChIP–seq workflow

pull down the antibody–protein–DNA complexes. After washing away everything else, the crosslinks are reversed and the DNA is purified and sequenced. **The key step is the immunoprecipitation**: only DNA fragments that were bound by your target protein are enriched and sequenced, giving you a genome–wide map of where that protein acts. The output is a "peak profile"—genomic regions with high read counts indicate where the protein was bound. For example, H3K27ac ChIP–seq reveals all active enhancers and promoters in a cell type; CTCF ChIP–seq shows all insulator sites; transcription factor ChIP–seq maps where that factor regulates genes. Source: Microbe Notes. ChIP Sequencing (ChIP–seq.

The output is a **peak profile**—regions with high read density indicate where the protein was bound. For example:

- ChIP–seq for H3K27ac identifies active enhancers and promoters
- ChIP–seq for H3K4me3 marks active promoters
- ChIP–seq for CTCF reveals insulator sites and loop anchors
- ChIP–seq for a transcription factor like p53 shows where p53 binds across the genome

By integrating multiple ChIP–seq datasets, you can build **epigenomic maps** that annotate the function of every region in the genome—promoter, enhancer, silencer, insulator, or repressed heterochromatin.

36.3.3 ATAC–seq: mapping chromatin accessibility

ATAC–seq (Assay for Transposase–Accessible Chromatin) measures how open or closed chromatin is at each genomic position.

The method uses a hyperactive transposase enzyme called **Tn5**. Tn5 cuts DNA and simultaneously inserts sequencing adapters, but it can only do this where chromatin is accessible—where nucleosomes are absent or loosely bound.

`ATAC-seq Protocol`

```
================

Step 1: CELL PREPARATION
  Start with fresh or frozen cells
  Permeabilize cell/nuclear membrane
  Keep nuclei intact (do NOT lyse completely)

Step 2: TAGMENTATION
  Add Tn5 transposase (loaded with adapters)
  Tn5 binds to accessible chromatin
  Simultaneously:
    Cuts DNA (at open regions only)
    Inserts sequencing adapters
  Duration: 30 minutes at 37°C

Step 3: DNA PURIFICATION
  Stop reaction (SDS, EDTA)
  Purify tagmented DNA
  Result: Library-ready fragments

Step 4: PCR AMPLIFICATION
  Amplify using adapter primers
  Add sample barcodes (for multiplexing)
  Typically 5-12 cycles (minimal bias)

Step 5: SIZE SELECTION
  Remove very short fragments (<100 bp)
  Remove very long fragments (>1000 bp)
  Enrich for nucleosome-size fragments

Step 6: SEQUENCING
  Paired-end sequencing (2×50 bp or 2×75 bp)

Step 7: ANALYSIS
  Align to genome
  Call peaks (accessible regions)
  Measure fragment size distribution:
    Nucleosome-free: <147 bp (TF binding sites)
    Mononucleosome: ~200 bp
    Dinucleosome: ~400 bp
  Footprinting: identify TF binding sites

Key Advantages:
• FAST: ~3 hours from cells to library
• LOW INPUT: Works with 500-50,000 cells
• NO ANTIBODY: Unlike ChIP-seq
• MAPS: Promoters, enhancers, insulators
```

Figure: ATAC-seq reveals open chromatin regions across the genome. The assay uses a hyperactive transposase called Tn5 that's pre-loaded with sequencing adapter sequences. When added to permeabilized nuclei, Tn5 binds to accessible (open) chromatin—regions where DNA is not tightly wrapped around nucleosomes or blocked by other proteins. In a single "tagmentation" reaction, Tn5 simultaneously cuts the DNA and inserts sequencing adapters, preparing the library in one step. Critically, **Tn5 can only access open chromatin**—it cannot cut DNA that's wrapped tightly around nucleosomes or buried in condensed

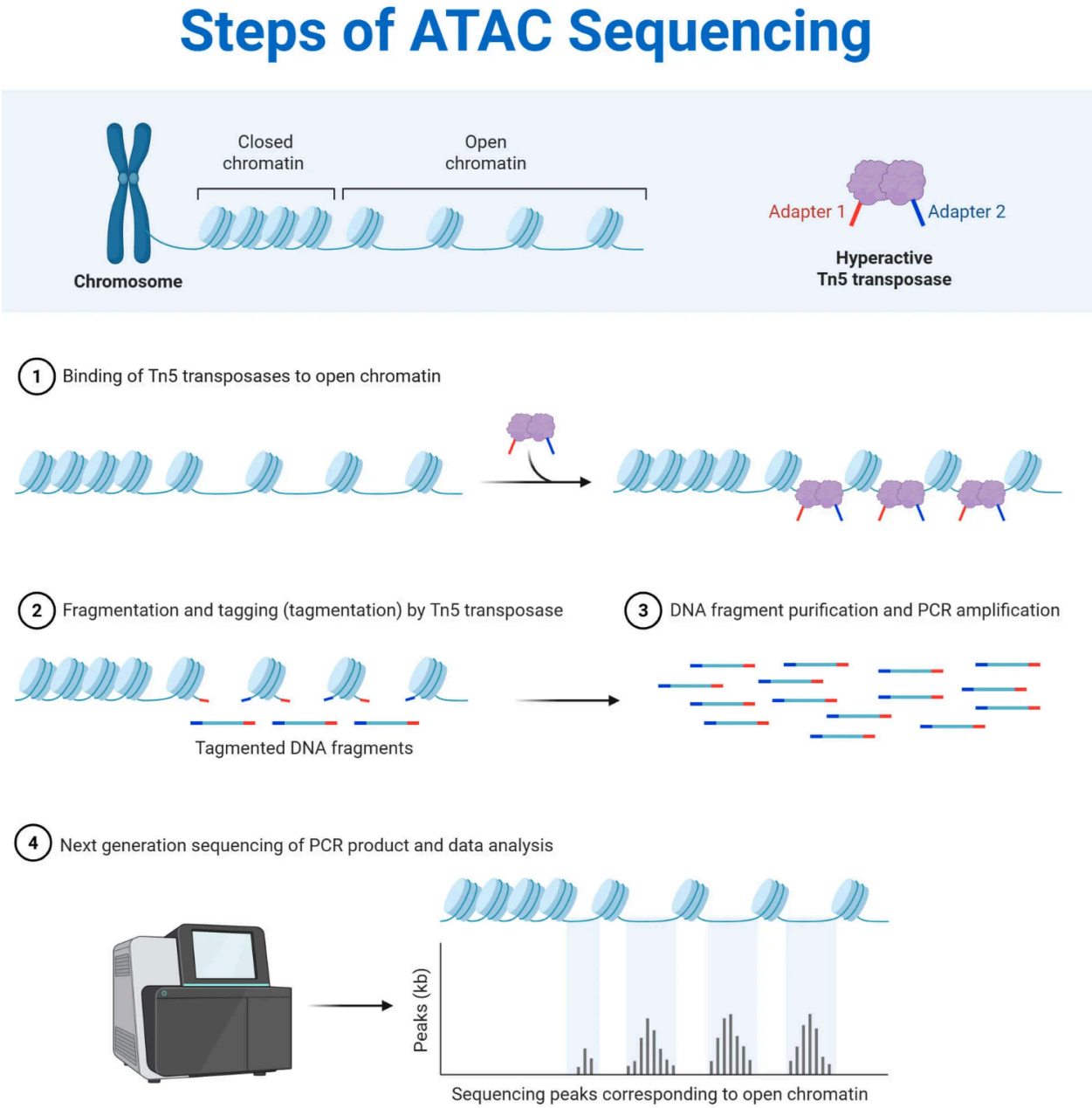


Figure 72: ATAC-seq workflow

heterochromatin. This selectivity is what makes ATAC–seq work: you sequence only the accessible regions, creating a genome–wide map of open vs. closed chromatin. Sharp peaks in the data indicate open promoters, active enhancers, and transcription factor binding sites. Valleys indicate closed, inaccessible chromatin. The beauty of ATAC–seq is its simplicity and speed: the entire library preparation takes about 3 hours, requires very few cells (as little as 500), and doesn't need antibodies like ChIP–seq does. This makes it ideal for profiling rare cell types or doing single–cell chromatin accessibility profiling (scATAC–seq). Source: Microbe Notes. ATAC–Seq (Assay for Transposase–Accessible Chromatin using Sequencing.

The output is an **accessibility map**:

- Sharp peaks indicate open promoters and enhancers
- Valleys indicate closed chromatin
- The width of peaks can tell you about nucleosome positioning

ATAC–seq is fast, requires very little input (as few as 500 cells), and works well for rare cell types. It's now widely used in single–cell form (**scATAC–seq**) to profile chromatin accessibility in thousands of individual cells, revealing cell–type–specific regulatory landscapes.

By looking for enriched transcription factor motifs within ATAC–seq peaks, you can infer which TFs are likely driving gene expression in that cell type—even if you don't have ChIP–seq data for those factors.

36.3.4 CUT&RUN: high–resolution chromatin profiling

CUT&RUN (Cleavage Under Targets and Release Using Nuclease) is a newer method that combines the best features of ChIP–seq and ATAC–seq while avoiding some of their limitations.

Instead of crosslinking and fragmenting chromatin, CUT&RUN uses a different strategy:

```
CUT&RUN Protocol
================

Step 1: CELL BINDING
  Bind cells to magnetic beads (Concanavalin A)
  Keeps cells immobilized during washes
  Cells remain intact (no lysis)

Step 2: PERMEABILIZATION
  Permeabilize cells (digitonin)
  Allows antibodies to enter nucleus

Step 3: ANTIBODY INCUBATION
  Add primary antibody (e.g., anti-H3K27ac)
  Antibody binds target protein
  Overnight at 4°C

Step 4: pA-MNase BINDING
  Add Protein A-Micrococcal Nuclease fusion
  pA binds to antibody Fc region
  Positions MNase right at binding site

Step 5: ACTIVATION & CLEAVAGE
  Add Ca²⁺ to activate MNase
  MNase cuts DNA immediately adjacent to protein
  2-hour incubation on ice
  Only small DNA fragments released
```

```
Step 6: DNA EXTRACTION
  DNA fragments diffuse out of nuclei
  Collect supernatant
  Purify DNA (no crosslink reversal needed!)

Step 7: LIBRARY PREP & SEQUENCING
  Add sequencing adapters
  PCR amplification
  Sequence

Key Advantages over ChIP-seq:
• NO CROSSLINKING (no artifacts)
• LOW BACKGROUND (targeted cleavage)
• SHARP PEAKS (precise cutting)
• LOW INPUT (fewer cells needed)
• WORKS FOR TFs (better than ChIP for many factors)
```

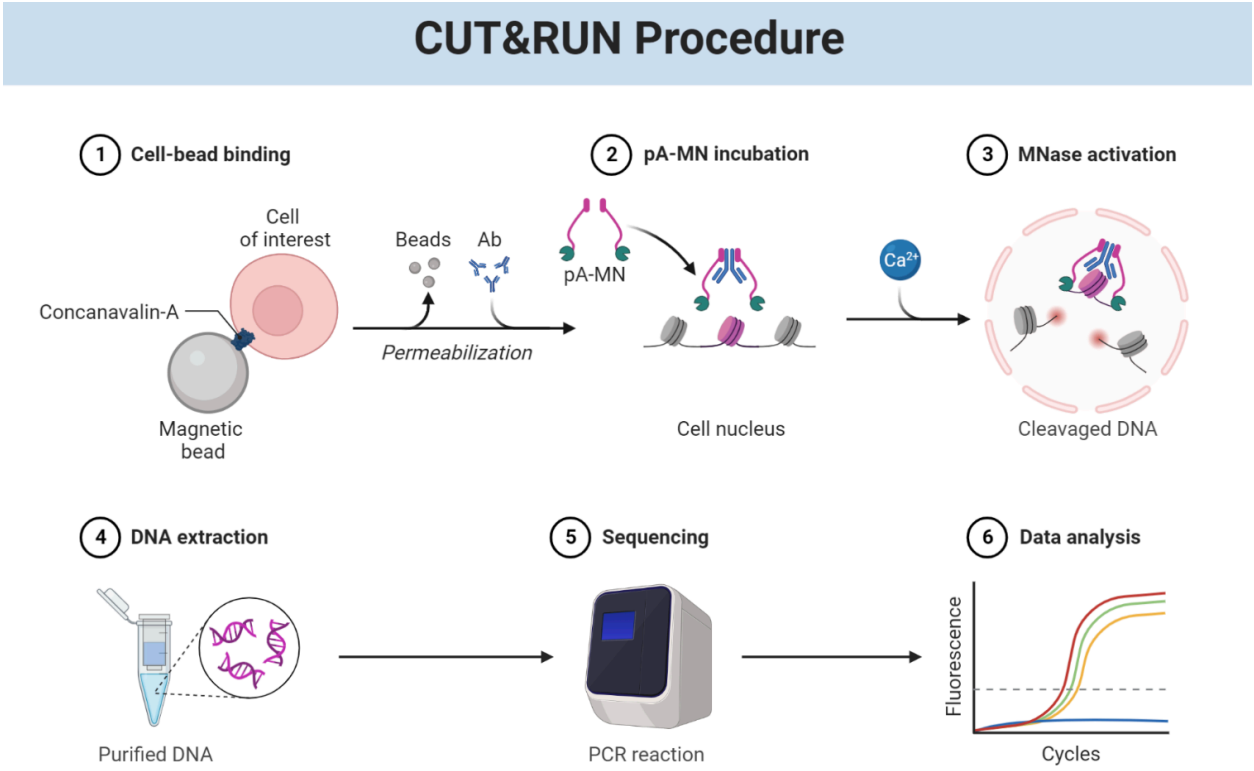


Figure 73: CUT&RUN workflow

Figure: CUT&RUN provides high-resolution maps of protein-DNA binding. Unlike ChIP-seq, CUT&RUN doesn't require formaldehyde crosslinking or sonication. Instead, cells are gently permeabilized and incubated with an antibody targeting your protein of interest (for example, a histone modification like H3K27ac or a transcription factor like CTCF). A Protein A-Micrococcal Nuclease (pA-MNase) fusion protein is then added, which binds to the antibody's Fc region, positioning the nuclease right at the site where your target protein sits on the DNA. When calcium is added to activate the MNase, it cuts the DNA immediately adjacent to the binding site—typically within 50-100 base pairs. These small DNA fragments diffuse out of the nucleus into the surrounding buffer, where they're collected and sequenced. **The key innova-**

tion is targeted cleavage: instead of fragmenting all the chromatin and then selecting bound fragments (like ChIP–seq), CUT&RUN only cuts DNA right where your protein binds, giving you extremely sharp, high–resolution peaks with very low background noise. The lack of crosslinking also means no artifacts from protein–protein crosslinks, and the method works with far fewer cells than ChIP–seq (as few as 100 cells). This makes CUT&RUN particularly powerful for studying rare cell types, difficult–to–ChIP transcription factors, and generating clean, publication–quality chromatin profiles. Source: BioRender. CUT&RUN Procedure.

Because the nuclease cuts right at the binding site, CUT&RUN produces very sharp peaks with high signal–to–noise ratio. It requires less material than ChIP–seq, doesn't need crosslinking (which can introduce artifacts), and works well for both histone marks and transcription factors.

36.3.5 Putting it all together: integrated epigenomics

Each technique provides a different layer of regulatory information:

Method	What It Measures	Output	Best For	Input Needed
RNA–seq	Gene expression (mRNA levels)	Transcript counts per gene	Identifying which genes are ON/OFF	100 ng RNA (~1M cells)
scRNA–seq	Single–cell gene expression	Cell–by–gene matrix	Cell type identification, heterogeneity	500–50K single cells
ChIP–seq	Protein–DNA binding	Peak locations (where protein binds)	Mapping TF binding, histone marks	10–50M cells
ATAC–seq	Chromatin accessibility	Open vs. closed regions	Identifying regulatory elements	500–50K cells
CUT&RUN	High–res protein–DNA binding	Sharp binding peaks	Precise TF/histone localization	100–10K cells
Hi–C	3D chromatin structure	Contact frequency matrix	TAD boundaries, enhancer–promoter loops	1–25M cells

- **RNA–seq**: What genes are expressed?
- **ChIP–seq**: Where are transcription factors and histone marks located?
- **ATAC–seq**: Which regions of chromatin are open?
- **CUT&RUN**: High–resolution mapping of specific chromatin features

By combining these datasets, you can reconstruct the regulatory logic of a cell:

```
Integrated Epigenomic Analysis
==============================

Example: Identifying an Active Enhancer

Data Integration:

  ATAC-seq:         (open chromatin)
```

```
  H3K27ac:          (active mark)
  H3K4me1:          (enhancer mark)
  RNA Pol II:       (not at enhancer)
  Gene 500kb:       (high expression)

          ↓
    INTERPRETATION:
    Active Enhancer
    • Open chromatin = accessible
    • H3K27ac = active
    • H3K4me1 = enhancer identity
    • Likely controls nearby gene

Example: Identifying Polycomb-Repressed Gene

Data Integration:

  ATAC-seq:         (closed)
  H3K27me3:         (repressive)
  H3K4me3:          (no activation)
  Gene expr:        (silent)

          ↓
    INTERPRETATION:
    Polycomb-Silenced Gene
    • Closed chromatin
    • H3K27me3 repression
    • Developmental gene OFF in this cell type

Example: Identifying Active Promoter

Data Integration:

  ATAC-seq:         (open)
  H3K4me3:          (sharp peak at TSS)
  H3K27ac:          (active)
  RNA Pol II:       (at promoter)
  RNA-seq:           (high expression)

          ↓
    INTERPRETATION:
    Actively Transcribed Gene
    • All marks of active transcription
    • Pol II present
    • mRNA being produced
```

For example:

- A region with open chromatin (ATAC peak), bound by RNA polymerase II (ChIP peak), and marked with H3K27ac (CUT&RUN peak) is likely an **active enhancer or promoter**
- A region with closed chromatin and marked with H3K27me3 is **repressed** by Polycomb
- A region with H3K4me1 but no H3K27ac is a **poised enhancer**—ready to be activated but not yet active

When you integrate these data with RNA-seq, you can link chromatin state to gene expression: genes near active enhancers are transcribed; genes in repressed chromatin are silent.

This integrated view reveals how the same genome produces different cell types. Neurons have open chromatin at neuron-specific enhancers and closed chromatin at liver enhancers. Liver cells have the opposite pattern. The DNA is the same, but the regulatory architecture is completely different.

36.4 Why this matters for genetics

Most disease-associated genetic variants identified by GWAS fall in non-coding regions—often in enhancers. A single nucleotide change can disrupt a transcription factor binding site, weakening or eliminating enhancer activity and reducing expression of a nearby gene.

```
How Non-Coding Variants Cause Disease
=====================================

Normal Enhancer:
[TF1 site][TF2 site][TF3 site]
    ↓       ↓       ↓
  TF1 + TF2 + TF3 bind
         ↓
   Active enhancer
         ↓
   Gene expressed
         ↓
   Normal function

SNP in Enhancer (breaks TF2 site):
[TF1 site][TF2-X][TF3 site]
    ↓       X       ↓
  TF1 +    + TF3 bind
         ↓
   Weak enhancer
         ↓
   Reduced gene expression (50%)
         ↓
   Insufficient protein
         ↓
   DISEASE

Example: Type 2 Diabetes GWAS hits
• Most are in pancreatic islet enhancers
• Reduce insulin gene expression by 10-20%
• Small effect per variant
• But 100+ variants add up → disease risk
```

For example, many type 2 diabetes risk variants are in enhancers that control pancreatic beta-cell genes. The variants don't break the genes—they just turn down the volume, reducing insulin production slightly. Over time, that small reduction contributes to disease risk.

Understanding gene regulation is also essential for interpreting rare variants. If you find a de novo variant in a patient and it's in a conserved non-coding region marked by H3K27ac, there's a good chance it's disrupting an enhancer. Functional validation (using reporter assays or CRISPR deletion of the enhancer) can confirm whether the variant is causal.

Finally, gene regulation is a therapeutic target. If a disease is caused by insufficient gene expression, maybe you can use CRISPRa to boost expression of the normal allele—as we saw with SCN2A in previous chapters. If a disease is caused by overexpression, maybe you can use CRISPRi to turn it down. Understanding the regulatory landscape tells you where and how to intervene.

36.5 Key takeaways

- **RNA-seq** measures gene expression genome-wide; **single-cell RNA-seq** reveals cellular heterogeneity within tissues
- **Normalization** (CPM, TPM, FPKM) is essential for comparing expression across genes and samples
- **Bulk RNA-seq** gives average expression across millions of cells; **scRNA-seq** profiles individual cells to identify distinct cell types and states
- **Histone modifications** create a combinatorial code that defines chromatin states (active, poised, repressed)
- **ChIP-seq** maps where specific proteins bind to DNA using antibody-based enrichment
- **ATAC-seq** measures chromatin accessibility using Tn5 transposase—fast, low input, no antibodies needed
- **CUT&RUN** provides high-resolution protein-DNA binding maps with minimal background
- **Integration of multiple epigenomic datasets** reveals the complete regulatory logic of cell types
- Most disease-associated variants affect **gene regulation, not protein sequence**—understanding the regulatory code is essential for interpreting genetic variation
- Cell identity is defined by **which enhancers are active**, not by DNA sequence

---

36.6 Key terms

- **RNA-seq**: RNA sequencing—quantifies gene expression genome-wide
- **Single-cell RNA-seq (scRNA-seq)**: RNA-seq performed on individual cells to reveal cellular heterogeneity
- **CPM, TPM, FPKM**: Normalization methods for RNA-seq data
- **Differential expression**: Statistical comparison to identify genes with changed expression between conditions
- **Chromatin**: DNA wrapped around histone proteins; can be open (accessible) or closed (repressed)
- **Histone modifications**: Chemical marks (e.g., acetylation, methylation) that regulate chromatin state
- **ChIP-seq**: Chromatin Immunoprecipitation Sequencing—maps where proteins bind to DNA
- **ATAC-seq**: Assay for Transposase-Accessible Chromatin—measures chromatin accessibility
- **CUT&RUN**: Cleavage Under Targets and Release Using Nuclease—high-resolution method to profile chromatin features
- **H3K4me3**: Histone mark indicating active promoters
- **H3K27ac**: Histone mark indicating active enhancers and promoters
- **H3K4me1**: Histone mark indicating all enhancers (active or poised)
- **H3K27me3**: Repressive histone mark deposited by Polycomb complexes
- **Epigenomic map**: Integrated dataset showing chromatin states across the genome

37 Chapter 30. QTLs: Connecting alleles to Molecular Traits

You've learned about GWAS, which connects genetic variants to diseases. You've learned about gene regulation, which controls how much RNA and protein gets made. Now let's connect those two worlds.

Here's the question: if you carry a particular genetic variant, does it change how much of a gene gets expressed? Does it change how much protein is made? And if so, by how much?

This is what **Quantitative Trait Locus (QTL)** analyses answer. QTLs measure whether genetic variants affect molecular traits—like RNA levels (**expression QTLs** or **eQTLs**) or protein levels (**protein QTLs** or **pQTLs**). They're the bridge between genotype and molecular phenotype, and they're essential for understanding how DNA variation actually works.

37.1 What is a QTL?

A QTL is a genetic variant that's statistically associated with variation in a quantitative trait. That trait could be anything you can measure numerically—height, blood pressure, cholesterol levels. In this chapter, we're focusing on **molecular traits**: gene expression (mRNA levels) or protein abundance.

Here's the basic idea. You have 100 people. You measure the expression of a gene—let's say APOE—in their liver tissue using RNA-seq. Some people have high APOE expression, some have low expression. Now you genotype a nearby SNP. You notice that people with the **A** allele have higher APOE expression on average, while people with the **G** allele have lower expression.

That SNP is an **eQTL** for APOE. It's associated with variation in APOE mRNA levels across individuals.

37.2 How QTL mapping works

The mechanics are straightforward. You need three things:

1. **Genotype data** for a bunch of genetic variants (usually millions of SNPs across the genome)
2. **Molecular phenotype data**—RNA-seq for eQTLs, mass spectrometry for pQTLs
3. **Statistical tests** to see if genotype predicts phenotype

For each gene, you test whether any nearby (or distant) SNPs are associated with its expression or protein level.

37.2.1 Encoding genotypes

Genotypes are usually coded as 0, 1, or 2, representing the number of copies of the alternate allele:

- **0** = homozygous for the reference allele (e.g., AA)
- **1** = heterozygous (e.g., AG)
- **2** = homozygous for the alternate allele (e.g., GG)

This is called an **additive model** (or **additive encoding**) because it assumes each copy of the alternate allele has an additive effect. If one copy raises expression by 10%, two copies raise it by 20%.

37.2.2 The molecular phenotype matrix

Expression or protein data are organized as a matrix: rows are genes (or proteins), columns are individuals, and each cell contains a normalized expression value (like TPM for RNA or normalized intensity for proteins).

For example, if you measured 20,000 genes in 500 people, you'd have a 20,000 × 500 matrix. Each row is a gene's expression profile across those 500 people.

37.2.3 Linear regression

The simplest QTL test is linear regression:

```
Expression =   +   × Genotype + error
```

Where: – β_1 is the effect size—how much expression changes per copy of the alternate allele – β_0 is the baseline expression level

If β_1 is significantly different from zero, you've found a QTL.

In practice, you add covariates to control for confounding factors—things like sex, age, sequencing batch, or population structure. These can create spurious associations if not accounted for.

37.2.4 Cis vs trans QTLs

QTLs come in two flavors:

Cis–QTLs are variants located near the gene they regulate—usually within ±1 megabase. These are the easiest to detect and interpret. If a SNP is 50 kb upstream of a gene and affects its expression, it's probably in a promoter or enhancer.

Trans–QTLs are variants that affect genes far away—on different chromosomes or many megabases distant. These are much weaker and harder to find. Trans–QTLs often work through intermediate factors, like transcription factors or signaling molecules. For example, a variant in a transcription factor gene might affect the expression of hundreds of downstream target genes.

37.2.5 Multiple testing

When you test millions of SNPs against thousands of genes, you're doing billions of statistical tests. Most will be noise. To avoid false positives, **False Discovery Rate (FDR)** correction or permutation–based thresholds are used to determine which associations are real.

37.3 eQTLs: genetic effects on RNA

eQTLs are the most common type of QTL. They tell you whether a genetic variant changes how much mRNA is produced from a gene.

Most disease–associated variants from GWAS are actually eQTLs. They don't change protein sequences—they change gene expression. A single nucleotide change in an enhancer can reduce expression by 20–30%, and that's enough to shift disease risk.

For example, many type 2 diabetes risk variants fall in enhancers that regulate genes in pancreatic beta cells. The variants don't break the genes—they just turn down the volume slightly, reducing insulin production over time.

eQTLs are tissue–specific. A variant might affect expression in the liver but not in the brain, because different tissues use different enhancers and have different chromatin states. The GTEx project mapped eQTLs across 54 human tissues and found that most eQTLs are active in only a subset of tissues (GTEx Consortium et al 2020, Science).

37.4 pQTLs: genetic effects on protein

pQTLs take things one step further. Instead of measuring RNA, you measure protein levels—usually in blood plasma using mass spectrometry.

Proteins are closer to phenotype than RNA is. RNA tells you what the cell is trying to make; protein tells you what it actually made. And because proteins are subject to additional layers of regulation—translation efficiency, protein stability, secretion, degradation—pQTLs capture effects that eQTLs might miss.

Interestingly, only about **40–60% of eQTLs** have a corresponding pQTL for the same gene. This discordance happens because:

- **Translation efficiency varies**: Some mRNAs are translated more efficiently than others
- **Protein half-life differs**: Some proteins are quickly degraded after synthesis
- **Post-translational regulation**: Proteins can be modified, sequestered, or secreted in ways that don't depend on RNA levels

So eQTLs tell you about transcriptional control, while pQTLs tell you about the realized protein output after all regulatory layers.

37.5 Example 1: Developmental plasma pQTLs in children

Plasma protein levels were measured in **2,147 children and adolescents** using mass spectrometry, identifying 1,216 high-confidence proteins. About **70% of these proteins** were influenced by at least one factor—genetics, age, sex, or BMI (Niu et al 2025, Nature Genetics).

Genetic effects were widespread: **one-third of all proteins** had at least one significant pQTL. Some variants changed protein levels by just a few percent; others caused up to **30-fold changes**.

The pQTLs mapped to key biological pathways:

- **Growth**: Variants affecting IGF1 and IGFALS (part of the growth hormone axis)
- **Inflammation**: Variants affecting CRP, complement proteins, and coagulation factors
- **Metabolism**: Variants affecting ANGPTL3, APOE, and LBP (lipopolysaccharide-binding protein)

Using Mendelian randomization—a statistical method that uses genetic variants as natural experiments—the study linked **41 causal genes** to **33 cardiometabolic traits** like cholesterol, blood sugar, and inflammatory markers.

The key finding? Genetic control of protein levels is established early in life and persists into adulthood. When the analysis was replicated in adults, the same pQTLs were still active. This means that genetic variants shaping metabolism and immunity in childhood continue to influence lifelong health.

Figure: Genome-wide pQTLs identified in children and adolescents. Panel (a) displays primary pQTLs across all chromosomes, where each point represents a genetic variant tested for association with protein levels—the higher the point, the stronger the association. Panel (b) shows where these pQTLs are located relative to the genes encoding their target proteins, revealing that most pQTLs cluster near their corresponding genes (cis-pQTLs). Panel (c) categorizes pQTLs by variant type—most are non-coding variants in introns or intergenic regions, with only 3% being missense mutations that alter protein sequences. Panel (e) distinguishes between cis-pQTLs (variants within ±1 Mb of the gene) and trans-pQTLs (variants on different chromosomes or far from the gene), showing that local regulation dominates. This demonstrates the pervasive genetic control of the plasma proteome, where genetic variants—mostly in regulatory regions—shape protein abundance throughout development. Source: Niu, L. et al. (2025). Plasma proteome variation and its genetic determinants in children and adolescents. Nature Genetics. https://www.nature.com/articles/s41588-025-02089-2. License: CC-BY 4.0.

37.6 Example 2: Parent-of-origin effects in pQTLs

Most genetic analyses assume that a variant has the same effect whether it comes from your mother or your father. But that's not always true. Some genes are **imprinted**, meaning only one parental copy is expressed.

UK Biobank data were used to test whether pQTLs show **parent-of-origin effects (POEs)**—cases where the maternal and paternal alleles have different effects on protein levels or disease risk (Hofmeister et al

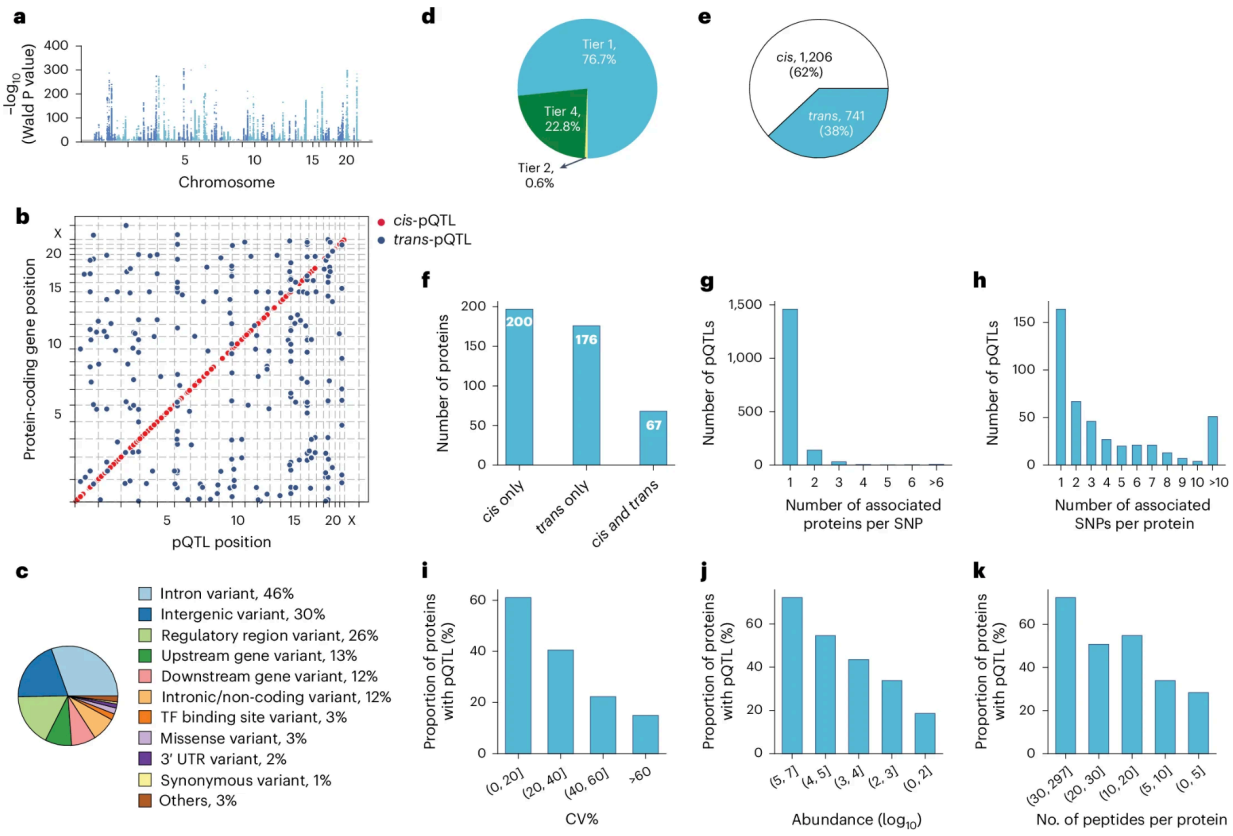


Figure 74: Niu2025–NatGenet–Fig3.png

2025, Nature).

The study analyzed ~109,000 participants and tested over 14,000 known pQTLs for parent-specific effects. More than **30 significant POEs** were identified, including classic imprinted loci like:

- **H19–IGF2**: Involved in growth; the paternal allele of IGF2 is expressed while the maternal allele is silenced
- **KCNQ1**: A potassium channel gene implicated in diabetes; shows parent-specific regulation
- **KLF14/MEST**: Two nearby genes with opposite imprinting patterns—KLF14 is maternally expressed, MEST is paternally expressed

What's striking is that **40% of these POEs showed antagonistic effects**—the maternal and paternal alleles pushed traits in opposite directions. For example, at one locus, the maternal allele increased triglycerides while the paternal allele decreased them.

This matters because standard additive models miss these effects. If you average over maternal and paternal contributions, you underestimate the true variance explained by genetics.

The findings were replicated in the Estonian Biobank (~85,000 people) and Norwegian MoBa cohort (~42,000 offspring), with **~87% of associations confirmed**.

37.7 Example 3: Multi-ancestry brain pQTLs

Most QTL studies have been done in European-ancestry populations, but genetic architecture can differ across ancestries. Brain pQTLs were mapped in a multi-ancestry cohort: **1,362 brain donors** from African-American, Hispanic/Latino, and non-Hispanic White backgrounds (Wingo et al 2025, Nature Genetics).

The study quantified ~11,750 proteins and identified **858 fine-mapped causal pQTLs**. By integrating these pQTLs with GWAS data for neurological and psychiatric diseases, **119 pQTL-protein-trait triads** were found linking genetic variants to protein levels to disease risk.

The causal proteins fell into several functional categories:

- **Synaptic signaling**: Proteins involved in neurotransmitter release and receptor trafficking
- **Immune and complement pathways**: Inflammation in the brain, especially relevant for Alzheimer's disease
- **Mitochondrial metabolism**: Energy production and oxidative stress

About **29% of causal pQTLs** were in coding regions (missense or nonsense variants that directly change the protein). The rest were in **promoters, UTRs, or enhancers**, showing that both transcriptional and post-transcriptional regulation shape brain protein levels.

Importantly, **ancestry-specific signals** were found. For example, novel loci for alcohol use disorder appeared only in African-ancestry data. This highlights that genetic architecture isn't universal—some regulatory variants are population-specific.

Predictive models trained on one ancestry generally worked well in others, but not perfectly. This means multi-ancestry studies are essential for capturing the full picture of genetic regulation.

Figure: Population-specific and shared brain pQTLs across ancestries. These LocusZoom plots illustrate genetic regulation patterns across populations. Panel (a) shows the DDX58 locus on chromosome 9, where a pQTL (purple diamond) is specific to African Americans—the variant strongly associates with DDX58 protein levels in the AA population but not in Hispanic or non-Hispanic white populations. Panel (b) shows the PPIL3 locus on chromosome 2, where a pQTL is shared across all three populations, demonstrating multi-ancestry regulation. Each point represents a genetic variant, with the y-axis showing association strength (-log P value) and the x-axis showing genomic position. Point colors indicate linkage disequilibrium (r^2) with the lead causal variant (purple diamond), helping distinguish true causal signals from correlated variants. This comparison reveals that while most causal pQTLs are shared across ancestries, some regulatory

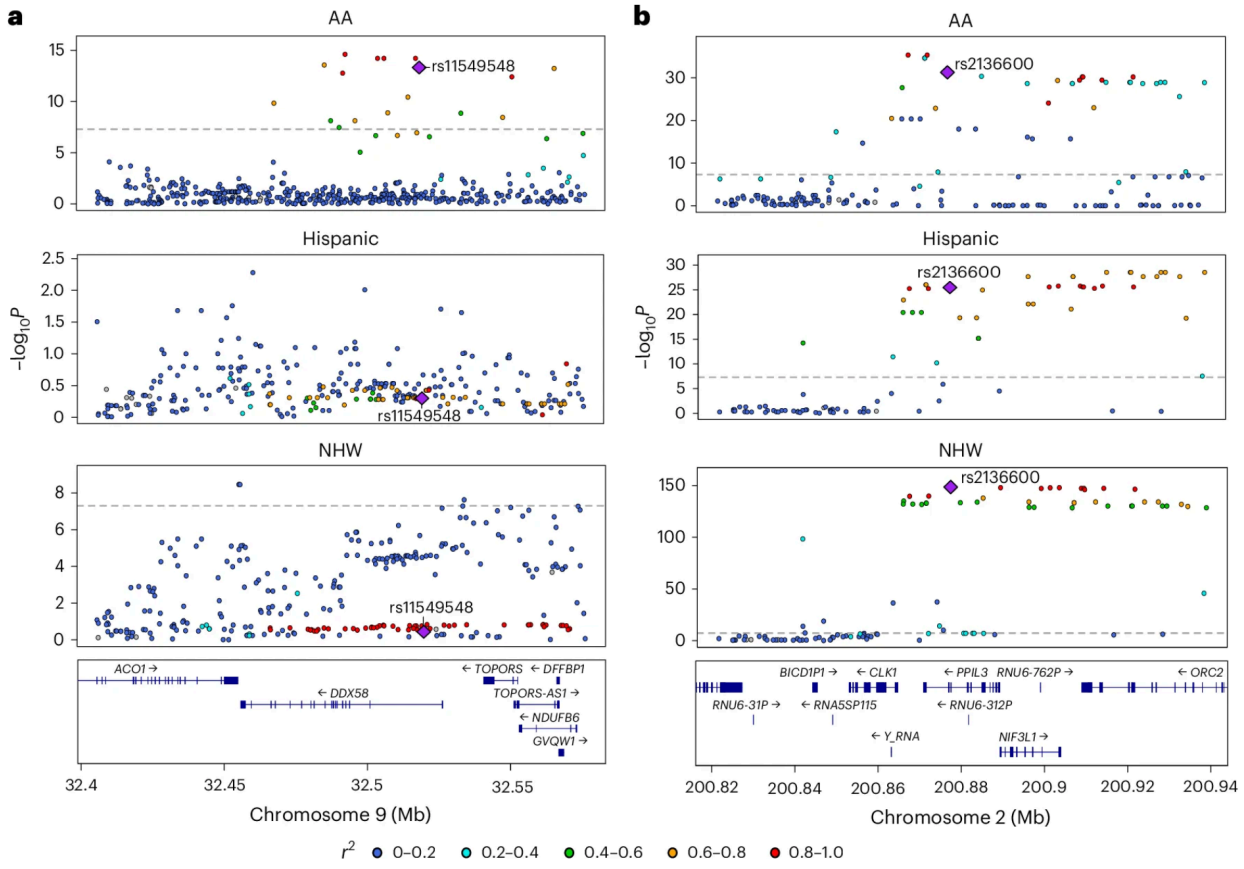


Figure 75: Wingo2025-NatGenet-Fig2.png

variants show population-specific effects, underscoring the importance of multi-ancestry studies for understanding protein regulation. Source: Wingo, T. S. et al. (2025). Multiancestry brain pQTL fine-mapping and integration with genome-wide association studies of 21 neurologic and psychiatric conditions. Nature Genetics. https://www.nature.com/articles/s41588-025-02291-2. License: CC-BY 4.0.

37.8 Interpreting QTLs: what they tell you (and what they don't)

QTLs are powerful, but they come with caveats.

37.8.1 Association is not causation

A significant QTL means a variant correlates with a molecular trait. But correlation doesn't prove the variant directly causes the change. The variant might be in linkage disequilibrium with the true causal variant, or it might act through an intermediate factor.

For example, the famous FTO locus is associated with obesity. But the causal mechanism doesn't involve FTO itself—it involves a distant enhancer that regulates IRX3, a gene involved in energy metabolism. The FTO variants are just markers tagging the real functional element.

To prove causality, you need functional validation: CRISPR perturbations, reporter assays, or colocalization with disease GWAS signals.

37.8.2 Tissue and cell-type specificity

Most QTL effects are context-dependent. A variant might regulate expression only in specific tissues or cell types.

For example: - An APOE eQTL in the liver affects cholesterol metabolism - An IL2RA eQTL in T cells affects immune activation, but only after cytokine stimulation - A KLF14 pQTL in adipose tissue regulates insulin sensitivity, but only in females

This is why eQTL databases like GTEx provide tissue-specific data. If you want to understand a diabetes risk variant, you should look at pancreatic islet eQTLs, not brain eQTLs.

37.8.3 eQTLs don't always predict pQTLs

RNA and protein levels are correlated, but not perfectly. Many eQTLs have no corresponding pQTL because:

- **Post-transcriptional regulation**: mRNA splicing, stability, or localization can change without affecting protein
- **Translation efficiency**: Some mRNAs are translated more efficiently than others
- **Protein stability**: Some proteins are rapidly degraded after synthesis

So eQTLs tell you about potential protein changes, while pQTLs tell you about realized protein output.

37.8.4 Effect sizes are usually small

Most QTLs have modest effects. A typical cis-eQTL explains 5-20% of variance in gene expression. Trans-QTLs explain less than 1%. pQTLs tend to be even smaller—1-5% per locus—because additional regulatory layers dilute the genetic signal.

But small effects matter. Complex traits are shaped by thousands of small-effect variants, and the cumulative burden adds up. That's the essence of polygenic architecture.

37.8.5 Statistical challenges

QTL mapping involves testing millions of SNP–gene pairs, so multiple testing correction is essential. Population structure can also create spurious associations if not controlled. And measurement noise—batch effects in RNA-seq, variability in proteomics—can obscure real signals or inflate false ones.

Large sample sizes help. For trans-QTLs and rare-variant effects, you need tens of thousands of samples to achieve statistical power.

37.9 Integrating QTLs with GWAS: finding causal genes

The real power of QTLs emerges when you combine them with GWAS.

Here's the workflow:

1. **GWAS identifies a disease-associated locus**—let's say a region on chromosome 19 linked to Alzheimer's disease
2. **eQTL/pQTL mapping shows that the same variants affect** APOE **and** CLU **expression** in astrocytes
3. **Colocalization analysis tests whether the GWAS signal and the QTL signal are driven by the same causal variant**
4. **If they colocalize, you have strong evidence that dysregulation of** APOE **and** CLU **mediates Alzheimer's risk**

This approach has identified causal genes for hundreds of GWAS loci. Without QTL data, GWAS can only point to a genomic region—it can't tell you which gene in that region actually matters.

Mendelian randomization takes this further. It uses genetic variants as instrumental variables to test whether changing a molecular trait causally affects disease risk. For example, if raising APOE protein levels reduces Alzheimer's risk, that suggests APOE is a therapeutic target.

37.10 QTLs in systems biology: building regulatory networks

QTLs don't act in isolation. They're part of larger regulatory networks where transcription factors, signaling pathways, and chromatin modifiers interact.

By integrating eQTLs, pQTLs, and protein–protein interaction data, you can map how genetic variants propagate through molecular layers:

```
DNA variant → RNA expression (eQTL) → Protein level (pQTL) → Pathway activity → Cellular phenotype → Di
```

This systems view reveals: – **Master regulators**: Variants in transcription factors that affect hundreds of downstream genes – **Shared regulation**: Gene modules that are co-regulated by the same variants – **Pathway-level effects**: How variants influence entire biological processes, not just individual genes

For example, variants in immune genes often form trans-eQTL hotspots—single loci affecting dozens of immune response genes. This suggests they regulate a shared immune pathway rather than acting on each gene independently.

37.11 Summary

QTLs bridge the gap between DNA variation and molecular traits, transforming human genetics from a static "gene → disease" model to a dynamic "variant → regulation → molecular trait → disease" framework. eQTLs measure genetic effects on RNA levels while pQTLs measure genetic effects on protein levels, revealing that most disease-associated GWAS variants work by changing gene expression rather

than protein sequences. These effects are highly tissue–specific and context–dependent, with only 40–60% of eQTLs showing corresponding pQTLs—demonstrating that RNA and protein regulation operate as distinct regulatory layers. Individual QTL effect sizes are usually small, explaining just 1–20% of variance in molecular traits, but thousands of these small effects accumulate across the genome to shape complex traits and disease risk.

The power of QTLs lies in their ability to explain GWAS hits, reveal tissue–specific disease mechanisms, guide drug development, and connect biological layers. Most GWAS variants are non–coding and work through gene regulation, making QTLs essential for identifying which genes they actually regulate and in which cell types the action occurs. When a pQTL shows that raising a protein level reduces disease risk, that protein becomes a therapeutic target. By measuring both RNA and protein, QTLs reveal how genetic effects propagate through regulatory networks, and inter–individual variation in drug response can be traced to eQTLs and pQTLs affecting drug–metabolizing enzymes and targets—enabling precision medicine approaches.

Key discoveries from QTL studies include parent–of–origin effects where imprinted genes show maternal versus paternal allele–specific regulation, ancestry–specific regulatory variants that differ across populations, and developmental timing effects where genetic control established in childhood persists throughout life. Integrating QTLs with GWAS through colocalization and Mendelian randomization has identified causal genes for hundreds of disease loci—without QTL data, GWAS can only point to broad genomic regions, but QTLs pinpoint which genes actually matter and how to target them therapeutically.