Introduction to Biomedical Informatics

Contents

1 의생명정보학 개론 **9**

2 서론 **10**
- 2.1 의생명정보학의 정의와 범위 . . . 10
- 2.2 생명과학의 발전과 생명정보학의 등장 . . . 11
- 2.3 체학 . . . 12
  - 2.3.1 “-체(-ome)”과 “-체학(-omics)”의 개념 . . . 12
  - 2.3.2 인간 지놈 프로젝트와 기술 발전 . . . 12
  - 2.3.3 후성유전체학 . . . 17
- 2.4 이미지 데이터 . . . 17
- 2.5 단백질 구조 데이터 . . . 20
- 2.6 의료정보학적 데이터 . . . 24
  - 2.6.1 전자의무기록(Electronic Medical Record, EMR) . . . 24
  - 2.6.2 국제질병분류(International Classification of Diseases, ICD) . . . 24
  - 2.6.3 의료영상 및 통신 표준(DICOM) . . . 24
  - 2.6.4 비디지털 데이터와 그 처리 . . . 27
  - 2.6.5 통계 데이터 . . . 27
- 2.7 현대 생명정보학의 동향 . . . 27

3 유전체학 기초 **29**
- 3.1 유전체와 유전체학의 기본 개념 . . . 29
  - 3.1.1 유전체의 정의와 표준 유전체 . . . 29
  - 3.1.2 모델 생물 . . . 29
  - 3.1.3 유전체 브라우저 . . . 32
- 3.2 유전적 다양성과 변이 . . . 32
  - 3.2.1 핵형과 배수성 . . . 32
  - 3.2.2 유전형과 표현형 . . . 32
  - 3.2.3 유전적 변이 . . . 36
- 3.3 인구집단 유전학 연구와 데이터베이스 . . . 43
  - 3.3.1 1000 지놈 프로젝트(1000 Genome Project) . . . 43
  - 3.3.2 유전체 통합 데이터베이스(gnomAD) . . . 46
  - 3.3.3 dbSNP . . . 46
  - 3.3.4 유전적 질환 관련 데이터베이스 . . . 47
  - 3.3.5 희귀 변이 및 국가별 연구 . . . 47
- 3.4 생물학적 경로와 단백질 구조 . . . 47
  - 3.4.1 생물학적 경로 . . . 47
  - 3.4.2 유전자 집합 농축 분석 . . . 50
  - 3.4.3 단백질 구조 데이터베이스 . . . 50
- 3.5 생명정보학 종합 리소스: NCBI . . . 50
- 3.6 생명정보학 데이터베이스의 활용 . . . 53

4 차세대 시퀀싱 **53**

4.1 차세대 시퀀싱의 개요 . . . 53
4.2 주요 NGS 플랫폼의 이해 . . . 53
4.2.1 Illumina 시퀀싱 기술 . . . 53
4.2.2 Long Read 시퀀싱 기술: Nanopore와 PacBio . . . 56
4.3 Illumina 시퀀싱 시스템의 구조 . . . 58
4.3.1 플로우 셀과 시퀀싱 칩의 이해 . . . 58
4.4 NGS 시퀀싱 라이브러리의 구조 . . . 59
4.4.1 라이브러리 어댑터 구성요소 . . . 59
4.5 시퀀싱 리드의 특성과 유형 . . . 59
4.5.1 페어드 엔드와 싱글 엔드 시퀀싱 . . . 63
4.6 Fastq 파일 형식과 구조 . . . 63
4.6.1 시퀀싱 품질 정보의 이해 . . . 66
4.7 NGS 데이터의 정렬과 알고리즘 . . . 68
4.7.1 주요 정렬 알고리즘의 원리 . . . 69
4.7.2 유전체 정렬 기술 . . . 71
4.7.3 SAM 파일 형식과 구조 . . . 73

5 변이와 진화 **74**
5.1 변이 찾기 (Variant Calling) . . . 74
5.2 VCF 포맷 . . . 74
5.3 Phasing . . . 75
5.4 Contamination . . . 77
5.5 개인 유전체 분석 과정 . . . 77
5.6 Hardy–Weinberg Equilibrium . . . 77
5.7 Genetic Drift . . . 79
5.8 진화 . . . 80
5.9 다중서열정렬과 분자 진화 . . . 85
5.9.1 병원체 진화 추적 . . . 85
5.9.2 COVID-19 사례를 통한 감염병 전파 추적 . . . 86
5.10 고유전학 . . . 86
5.10.1 고유전학의 기술적 도전 . . . 86
5.10.2 고유전학의 응용 분야 . . . 86
5.11 결론 . . . 87

6 유전질환과 GWAS **87**
6.1 유전질환 . . . 87
6.2 유전질환의 유형 . . . 88
6.2.1 멘델 질환(Mendelian disorders) . . . 88
6.2.2 다유전자 질환(Polygenic disorders) . . . 88
6.3 다유전자 형질(Polygenic Traits) . . . 88
6.4 전장유전체 연관분석(Genome-wide Association Study, GWAS) . . . 89
6.4.1 GWAS 설계 예시 . . . 89
6.4.2 P-값의 의미 . . . 89
6.4.3 맨해튼 플롯(Manhattan plot) . . . 93
6.5 연관 불균형(Linkage Disequilibrium, LD) . . . 93
6.5.1 해플로타입(Haplotype) . . . 96
6.5.2 HapMap 프로젝트와 연관 불균형 . . . 96
6.5.3 태그 SNP를 이용한 효율적 분석 . . . 98
6.6 다유전자 위험 점수(Polygenic Risk Score, PRS) . . . 98
6.6.1 PRS의 기본 개념 . . . 98
6.6.2 PRS의 활용과 한계 . . . 99
6.6.3 PRS 계산 도구 . . . 99
6.7 과잉 표현 분석(Over-Representation Analysis, ORA) . . . 99
6.7.1 유전자 온톨로지(Gene Ontology, GO) . . . 99

6.7.2 다양한 경로 데이터베이스 . . . 100
6.7.3 초기하 분포(Hypergeometric Distribution) . . . 100
6.8 결론 . . . 105

7 암 107
7.1 암이란 무엇인가? . . . 107
7.2 암은 왜 발생하는가? . . . 107
7.2.1 단세포 생물의 유전자 유산 . . . 107
7.2.2 종양 억제자(Tumor suppressors) . . . 107
7.3 암 관련 신호 전달 경로(Pathways) . . . 107
7.3.1 MAPK-ERK 경로 . . . 107
7.3.2 RAS 유전자의 종간 보존 . . . 107
7.4 원암유전자와 암유전자 . . . 111
7.5 돌연변이 . . . 111
7.5.1 BCR-ABL . . . 111
7.5.2 복제수 변이(Copy Number Variation, CNV) . . . 111
7.6 대표적인 암유전자와 종양 억제자 . . . 113
7.6.1 종양 억제자 . . . 113
7.7 드라이버 돌연변이와 패신저 돌연변이 . . . 113
7.7.1 드라이버 돌연변이(Driver mutations) . . . 113
7.7.2 패신저 돌연변이(Passenger mutations) . . . 113
7.8 체세포 돌연변이(Somatic mutations) . . . 113
7.8.1 작은(또는 단순한) 돌연변이: SNV, Indel . . . 119
7.8.2 Rainfall plot . . . 120
7.8.3 CNV(Copy Number Variation) . . . 120
7.8.4 큰(또는 복잡한) 돌연변이: 대규모 Indel, 구조적 변이(SV) . . . 120
7.9 암 변이 데이터베이스 . . . 123
7.9.1 주요 암 데이터베이스 . . . 123
7.9.2 데이터 시각화 도구 . . . 123
7.9.3 Oncoprint . . . 124
7.10 변이 주석(Variant Annotation) . . . 124
7.11 종양 순도(Tumor Purity) . . . 124
7.12 암 치료를 위한 정밀 의학 . . . 124
7.13 암 유전체의 특성 . . . 124
7.13.1 종양 이질성(Tumor Heterogeneity) . . . 124
7.13.2 클론 진화(Clonal Evolution) . . . 127
7.13.3 변이 시그니처(Mutational Signatures) . . . 127
7.13.4 마이크로위성체 불안정성(Microsatellite Instability, MSI) . . . 127
7.13.5 종양 변이 부담(Tumor Mutational Burden, TMB) . . . 129
7.14 액체생검과 순환종양DNA . . . 129
7.14.1 액체생검(Liquid Biopsy)의 개념 . . . 129
7.14.2 순환종양DNA(Circulating Tumor DNA, ctDNA) . . . 129
7.14.3 액체생검의 임상 적용 . . . 129
7.15 유전체학 기반 정밀의학 . . . 129
7.15.1 표적치료 . . . 131
7.15.2 면역치료 . . . 131
7.15.3 차세대 치료 기술 . . . 134
7.16 암과 노화 . . . 137
7.17 결론 . . . 137

8 비코딩 유전체와 발현 조절 139
8.1 인간 유전체의 구성 . . . 139
8.2 GWAS 변이의 분포와 비코딩 영역의 중요성 . . . 141
8.2.1 질병별 GWAS 변이 분포 . . . 141

8.3 발현 수량적 형질 유전자좌(eQTL) . . . 141
8.3.1 QTL과 eQTL의 개념 . . . 141
8.3.2 eQTL의 분류 . . . 141
8.3.3 eQTL 분석의 의의 . . . 144
8.3.4 GTEx 프로젝트 . . . 144
8.4 트랜스포존과 유전체 진화 . . . 145
8.5 비코딩 RNA의 종류와 기능 . . . 145
8.6 비코딩 RNA의 의생명정보학적 분석 . . . 146
8.7 치료적 응용과 전망 . . . 146
8.8 결론 . . . 146

9 후성유전체학 146
9.1 크로마틴 구조와 후성유전학적 조절 . . . 146
9.1.1 크로마틴의 기본 구조 . . . 146
9.1.2 후성유전학적 조절 요인 . . . 148
9.1.3 질병과의 연관성 . . . 148
9.2 DNA 메틸화의 분자적 기전 . . . 148
9.2.1 DNA 메틸화의 기본 개념 . . . 148
9.2.2 메틸화와 유전자 발현의 관계 . . . 148
9.2.3 메틸화의 생화학적 특성 . . . 148
9.3 전장 바이설파이트 시퀀싱(WGBS) . . . 148
9.3.1 WGBS의 원리 . . . 148
9.3.2 바이설파이트 변환의 화학적 원리 . . . 148
9.3.3 WGBS 실험 과정 . . . 150
9.3.4 WGBS 데이터의 해석 . . . 150
9.4 후성유전체 연관 분석(EWAS) . . . 150
9.4.1 EWAS의 정의와 목적 . . . 150
9.4.2 EWAS와 GWAS의 차이점 . . . 150
9.4.3 EWAS의 주요 발견들 . . . 151
9.5 크로마틴 면역침강 시퀀싱(ChIP-seq) . . . 151
9.5.1 ChIP-seq의 기본 원리 . . . 151
9.5.2 ChIP-seq 실험 과정 . . . 151
9.5.3 면역침강의 생화학적 원리 . . . 153
9.5.4 ChIP-seq 데이터 분석 . . . 153
9.5.5 ChIP-seq의 응용 분야 . . . 154
9.6 표적 시퀀싱과 캡처 기술 . . . 154
9.6.1 표적 시퀀싱의 필요성 . . . 154
9.6.2 엑솜 시퀀싱과 표적 후성유전체 분석 . . . 154
9.7 ATAC-seq: 접근 가능한 크로마틴 분석 . . . 154
9.7.1 ATAC-seq의 기본 원리 . . . 154
9.7.2 Tn5 태그멘테이션의 분자적 기전 . . . 154
9.7.3 ATAC-seq 실험 과정 . . . 158
9.7.4 ATAC-seq 데이터 분석 . . . 158
9.8 크로마틴 형태 포획(3C 및 Hi-C) . . . 158
9.8.1 크로마틴 형태 포획의 개념 . . . 158
9.8.2 3C 기술의 원리 . . . 158
9.8.3 Hi-C: 유전체 전체 크로마틴 상호작용 . . . 158
9.9 피크 호출 방법론 . . . 161
9.9.1 피크 호출의 기본 원리 . . . 161
9.9.2 피크 형태와 특성 . . . 161
9.9.3 피크 호출 방법론 . . . 163
9.10 결론 . . . 164

10 전사체학 기초 164

10.1 전사체학 개요 . . . 164
10.2 마이크로어레이 기술 . . . 164
10.2.1 마이크로어레이의 기본 원리 . . . 164
10.2.2 마이크로어레이의 장단점 . . . 167
10.2.3 마이크로어레이 데이터 분석 . . . 167
10.3 정량적 중합효소연쇄반응(qPCR) . . . 167
10.3.1 qPCR의 기본 원리 . . . 167
10.3.2 qPCR의 장단점과 응용 . . . 168
10.4 Bulk RNA 시퀀싱 . . . 168
10.4.1 실험 과정 . . . 168
10.4.2 데이터 정량화와 정규화 . . . 168
10.4.3 유전자 발현의 확률적 모델 . . . 171
10.4.4 데이터 변환과 정규화 . . . 174
10.5 배치 효과와 보정 . . . 174
10.5.1 배치 효과의 원인 . . . 174
10.5.2 정규화 기법 . . . 177
10.6 차등 발현 분석 . . . 177
10.6.1 개념 . . . 177
10.6.2 주요 분석 도구 . . . 177
10.6.3 핵심 통계 지표 . . . 177
10.6.4 화산 플롯(Volcano plot) . . . 177
10.7 RNA-seq 데이터 분석 파이프라인 . . . 179
10.7.1 데이터 전처리 . . . 179
10.7.2 다운스트림 데이터 분석 . . . 179
10.7.3 생물학적 해석 . . . 180
10.8 결론 . . . 180

11 단일 세포 전사체학 **180**
11.1 단일세포 RNA 시퀀싱의 혁신 . . . 180
11.1.1 세포 분화와 발생학적 이해 . . . 181
11.1.2 유전자 발현의 확률적 특성 . . . 181
11.1.3 단일세포 데이터의 클러스터링 . . . 181
11.2 플로우 사이토메트리 . . . 182
11.2.1 기술의 물리적 원리 . . . 182
11.2.2 형광 활성화 세포 분류(FACS) . . . 182
11.2.3 다중 매개변수 분석 . . . 182
11.3 미세유체학과 단일세포 포획 기술 . . . 185
11.3.1 미세유체학의 물리적 특성 . . . 185
11.3.2 Drop-seq 기술의 원리 . . . 186
11.3.3 라이브러리 제작과 상용화 플랫폼 . . . 186
11.4 전사체 데이터 분석 방법론 . . . 187
11.4.1 통계적 검정과 다중 비교 보정 . . . 187
11.4.2 차원 축소 기법 . . . 187
11.4.3 클러스터링과 품질 평가 . . . 187
11.5 단일세포 데이터 분석 파이프라인의 개요와 중요성 . . . 187
11.6 단일세포 데이터 분석의 핵심 단계 . . . 189
11.6.1 데이터 정량화와 품질 관리 . . . 189
11.6.2 배치 효과 교정과 데이터 정규화 . . . 189
11.6.3 차원 축소와 그래프 기반 클러스터링 . . . 194
11.7 세포 유형 식별과 마커 유전자 분석 . . . 194
11.7.1 마커 유전자 발견 . . . 194
11.7.2 세포 유형 주석 과정 . . . 195
11.7.3 자동화된 세포 유형 주석 . . . 195
11.8 대규모 단일세포 연구와 세포 아틀라스 프로젝트 . . . 198

11.8.1 Human Cell Atlas (HCA) 프로젝트 . . . 198
11.8.2 모델 생물 아틀라스 프로젝트 . . . 198
11.8.3 참조 매핑과 전이 학습 . . . 201
11.9 세포 디콘볼루션: 벌크 데이터에서 세포 구성 추정 . . . 201
11.9.1 참조 기반 디콘볼루션 방법 . . . 202
11.9.2 참조 프리 디콘볼루션 방법 . . . 202
11.9.3 디콘볼루션 결과의 검증과 해석 . . . 203
11.10 임상 응용과 미래 전망 . . . 204
11.10.1 정밀의학에서의 역할 . . . 204
11.10.2 기술 발전 방향 . . . 204
11.10.3 도전 과제와 해결 방안 . . . 204

12 차원 축소와 데이터 분석 **205**
12.1 고차원 데이터의 문제점과 차원 축소의 필요성 . . . 205
12.1.1 차원의 저주(Curse of Dimensionality) . . . 206
12.1.2 희소성(Sparsity) 문제 . . . 206
12.1.3 밀집(Dense)과 희소(Sparse) 데이터의 처리 전략 . . . 206
12.2 주성분 분석(Principal Component Analysis, PCA) . . . 206
12.2.1 PCA의 수학적 원리 . . . 206
12.2.2 최적 주성분 개수 결정 . . . 207
12.2.3 의생명정보학에서의 PCA 응용 사례 . . . 207
12.3 비선형 차원 축소 . . . 207
12.3.1 확산 지도(Diffusion Maps) . . . 212
12.3.2 t-SNE . . . 212
12.3.3 UMAP . . . 212
12.4 기법 비교 . . . 213
12.4.1 Swiss Roll 예제 . . . 213
12.4.2 MNIST 손글씨 . . . 213
12.5 클러스터링 기법 . . . 213
12.5.1 K-means . . . 214
12.5.2 네트워크 기반 클러스터링 . . . 214
12.6 주요 도구 . . . 214
12.7 해석 시 주의점 . . . 219
12.7.1 t-SNE . . . 219
12.7.2 UMAP 매개변수 . . . 219
12.7.3 2차원 시각화의 한계와 왜곡의 필연성 . . . 219
12.8 결론 . . . 222

13 조직 병리학과 공간체학 **222**
13.1 공간 전사체학과 조직 병리학의 통합 . . . 222
13.1.1 조직학적 염색법의 원리 . . . 222
13.1.2 형광 이미징과 다중 표지 . . . 225
13.1.3 공간 전사체학의 임상적 의의 . . . 225
13.2 공간전사체학 기술의 발전과 응용 . . . 225
13.2.1 공간전사체학의 기본 원리 . . . 225
13.2.2 Visium 기술과 발전 방향 . . . 229
13.2.3 형광 기반 공간 분석 기술 . . . 229
13.2.4 제자리 시퀀싱 기술 . . . 232
13.2.5 공간전사체학 데이터 분석의 특수성 . . . 232
13.2.6 단일세포 수준에서의 공간 전사체학 . . . 243
13.3 통합 분석과 다중 모달 접근법 . . . 245
13.3.1 단일세포와 공간전사체학 데이터의 통합 . . . 245
13.3.2 다중 오믹스 통합 . . . 245
13.4 기술의 한계와 도전 과제 . . . 245

13.4.1 기술적 한계 . . . 245
13.4.2 분석상의 도전 과제 . . . 248
13.5 임상 응용과 미래 전망 . . . 248
13.5.1 질병 연구에서의 응용 . . . 248
13.5.2 정밀의학으로의 발전 . . . 248
13.5.3 기술 발전 방향 . . . 248
13.6 결론 . . . 248

14 단백체학 249
14.1 단백체학 개요 . . . 249
14.2 단백체학 기법 . . . 250
14.2.1 질량분석법 기반 단백체학 . . . 250
14.2.2 항체 기반 단백체학 . . . 252
14.3 공간 단백체학 . . . 252
14.3.1 질량분석 기반 공간 이미징 . . . 252
14.3.2 형광 기반 공간 이미징 . . . 256
14.4 단일세포 단백체학 . . . 257
14.4.1 질량 세포분석법 (CyTOF) . . . 257
14.4.2 다중 오믹스 단일세포 분석 . . . 257
14.5 단백체학 데이터 분석 및 해석 . . . 257
14.5.1 질량분석 데이터 처리 . . . 257
14.5.2 단일세포 단백체학 데이터 분석 . . . 257
14.5.3 정량적 분석 및 통계 . . . 257
14.5.4 기능적 주석 및 경로 분석 . . . 258
14.6 임상 단백체학의 응용 . . . 258
14.6.1 질병 진단을 위한 단백체학 활용 . . . 258
14.6.2 공간 단백체학의 임상 적용 . . . 258
14.6.3 단백체학 데이터 분석의 도전과 해결방안 . . . 258
14.6.4 통합적 단백체학 접근법의 미래 . . . 259
14.7 결론 . . . 259

15 고급 유전체학 기법들 259
15.1 군유전체학(Metagenomics) . . . 259
15.1.1 군유전체학의 정의와 의의 . . . 259
15.1.2 군유전체학 연구 접근법 . . . 261
15.1.3 16S rRNA 유전자 분석 . . . 261
15.1.4 미생물 다양성 지표 . . . 262
15.1.5 군유전체학의 응용 사례 . . . 263
15.2 식물 유전체학 . . . 267
15.2.1 배수성 . . . 267
15.2.2 배수성의 진화적 의의 . . . 268
15.2.3 재배종과 야생종의 비교: 딸기의 사례 . . . 268
15.2.4 배수성과 현대 농업 . . . 268
15.3 유전체 편집 기술 . . . 273
15.3.1 유전체 편집 기술의 발전사 . . . 273
15.3.2 CRISPR/Cas 시스템의 구조와 원리 . . . 275
15.3.3 염기 편집 기술 . . . 276
15.3.4 프라임 편집 기술 . . . 276
15.3.5 CRISPR 기반 진단 기술 . . . 276
15.3.6 오프타겟 효과와 예측 도구 . . . 277
15.3.7 세포 내 소기관 유전체 편집 . . . 279
15.3.8 CRISPR를 이용한 이종장기이식 . . . 279
15.4 결론 . . . 279

16 유전체 데이터 보안 **281**
16.1 유전체 데이터 보안의 필요성과 윤리적 고려사항 . . . 281
16.1.1 개인 유전체 공개로 인한 위험성 . . . 283
16.2 유전체 데이터 익명화 기술 . . . 284
16.2.1 메타데이터 개요 . . . 284
16.2.2 메타데이터 보안 처리 . . . 287
16.2.3 메타데이터 기반 재식별 위험 . . . 288
16.3 통제 접근 저장 시스템 . . . 288
16.3.1 dbGaP (Database of Genotypes and Phenotypes) . . . 288
16.3.2 EGA (European Genome-phenome Archive) . . . 289
16.4 기타 보안 강화 기술 . . . 290
16.4.1 표준화 기반 보안 강화 기술 . . . 290
16.4.2 블록체인 기반 보안 기술 . . . 290
16.5 윤리적 고려사항과 미래 전망 . . . 290
16.5.1 현대의 유전체 데이터 보안 기술 . . . 290
16.6 결론 . . . 292

17 인공지능과 의생명정보학의 미래 **292**
17.1 인공지능과 기계학습의 의생명정보학 응용 . . . 292
17.1.1 딥러닝의 기본 원리와 의생명정보학에서의 역할 . . . 292
17.1.2 차원 축소와 잠재 공간 모델링 . . . 294
17.1.3 베이지안 접근법과 딥러닝의 확률적 원리 . . . 295
17.1.4 딥러닝 모델의 학습 원리와 손실 함수 . . . 295
17.2 딥러닝 기반 의생명 데이터 분석 . . . 296
17.2.1 오토인코더 . . . 296
17.2.2 합성곱 신경망(CNN) . . . 296
17.2.3 확산 모델(Diffusion Models) . . . 299
17.2.4 순환 신경망(RNN) . . . 299
17.2.5 트랜스포머와 대규모 언어 모델 . . . 301
17.2.6 트랜스포머를 넘어선 새로운 아키텍처 . . . 304
17.3 인공지능을 이용한 단백질 구조 예측 . . . 304
17.3.1 단백질 구조 예측의 중요성과 역사 . . . 304
17.3.2 AlphaFold의 혁신 . . . 307
17.3.3 최신 단백질 구조 예측 모델들 . . . 308
17.3.4 단백질 구조 예측의 응용과 미래 . . . 308
17.4 AI 에이전트의 등장과 의생명정보학 데이터 분석 . . . 309
17.4.1 AI 에이전트의 구조와 작동 원리 . . . 309
17.4.2 의생명정보학 분야의 AI 에이전트 활용 사례 . . . 309
17.5 인공지능과 의생명정보학의 미래 . . . 309
17.5.1 기술 발전의 방향 . . . 309
17.5.2 실용화와 접근성 . . . 311
17.5.3 도전 과제와 해결 방향 . . . 311
17.6 결론 . . . 311

18 컴퓨터 기반 신약 개발과 임상시험 **311**
18.1 서론 . . . 311
18.2 신약 개발 과정의 개요 . . . 312
18.2.1 전통적인 신약 개발 파이프라인 . . . 312
18.3 컴퓨터 기반 신약 설계 (CADD) . . . 313
18.3.1 표적 발견 기술 . . . 313
18.3.2 스크리닝 방법 . . . 313
18.3.3 분자 도킹 (Molecular Docking) . . . 314
18.3.4 리드 최적화 . . . 314
18.3.5 ADMET 예측 . . . 317

18.4 현대적 치료제 및 신약 개발 기술 . . . 318
18.4.1 항체-약물 접합체(ADC) . . . 318
18.4.2 CAR-T 세포 치료 . . . 318
18.4.3 PROTAC 기술 . . . 321
18.4.4 면역 체크포인트 억제제 . . . 321
18.5 인공지능 기반 신약 발견 . . . 321
18.5.1 AI 기반 신약 발견 워크플로우 . . . 321
18.5.2 최신 AI 기법의 활용 . . . 322
18.5.3 AI 기반 신약 발견의 성공 사례 . . . 325
18.6 임상시험의 설계와 방법론 . . . 325
18.6.1 임상시험의 개요 . . . 325
18.6.2 혼동 요인과 임상시험에서의 통제 . . . 328
18.6.3 임상시험의 단계 . . . 329
18.6.4 임상시험 설계 원칙 . . . 331
18.6.5 통계적 분석 방법 . . . 331
18.6.6 COVID-19 백신 임상시험 사례 . . . 333
18.6.7 임상시험 연구 설계의 중요성 . . . 333
18.6.8 멘델리안 무작위배정 . . . 336
18.7 결론 . . . 340

19 리눅스 **340**
19.1 리눅스의 중요성 . . . 340
19.2 리눅스의 역사와 철학 . . . 340
19.2.1 자유 소프트웨어 재단 (Free Software Foundation, FSF) . . . 340
19.2.2 GNU 프로젝트 . . . 340
19.2.3 리눅스 커널의 탄생 . . . 341
19.2.4 GNU/Linux 시스템 . . . 341
19.3 리눅스 배포판 . . . 342
19.3.1 배포판의 개념 . . . 342
19.3.2 주요 배포판 . . . 342
19.3.3 패키지 매니저 . . . 342
19.4 리눅스 기본 개념 . . . 342
19.4.1 루트 (Root) . . . 342
19.4.2 디렉토리 . . . 342
19.4.3 쉘 (Shell) . . . 343
19.5 가상 환경 구축: Conda . . . 343
19.5.1 Conda 설치 . . . 343
19.5.2 가상환경 생성 예제 . . . 343
19.6 생명정보학 문제 해결: Rosalind . . . 344
19.7 리눅스 필수 명령어 . . . 344
19.7.1 파일 시스템 탐색 . . . 344
19.7.2 파일 및 디렉토리 관리 . . . 344
19.7.3 시스템 모니터링 . . . 344
19.7.4 파일 조작 . . . 346
19.8 생명정보학에서의 리눅스 활용 . . . 346

1 의생명정보학 개론

본 강의노트는 부산대학교 의생명융합공학부에서 개설되는 “의생명정보학 개론” 강의를 위한 것입니다.

질문 또는 수정사항을 발견하신 경우 부산대학교 의생명융합공학부 박정빈 교수 (jeongbin.park@pusan.ac.kr) 에게 연락 부탁드립니다.

2 서론

2.1 의생명정보학의 정의와 범위

의생명정보학은 생명과학, 의학, 정보학이 융합되어 탄생한 학제간 분야이다. 의생명정보학은 다음과 같은 세 가지 주요 학문의 교차점에 위치한다: - **의학**: 인체의 구조, 기능, 질병에 관한 연구 - **생명과학**: 생명 현상과 생물체의 특성에 관한 연구 - **정보학**: 데이터의 저장, 검색, 처리 및 분석에 관한 연구

이러한 학문 분야의 교차점에서 파생된 세부 영역은 다음과 같다: - **의료정보학**: 의학과 정보학의 교차점으로, 의료 데이터의 관리와 분석에 중점 - **생명정보학**: 생명과학과 정보학의 교차점으로, 유전체, 단백질체 등의 생명과학적 데이터 분석에 중점 - **중개의학(Translational Medicine)**: 의학과 생명과학의 교차점으로, 기초 생명과학적 발견을 임상 응용으로 전환하는 과정에 중점

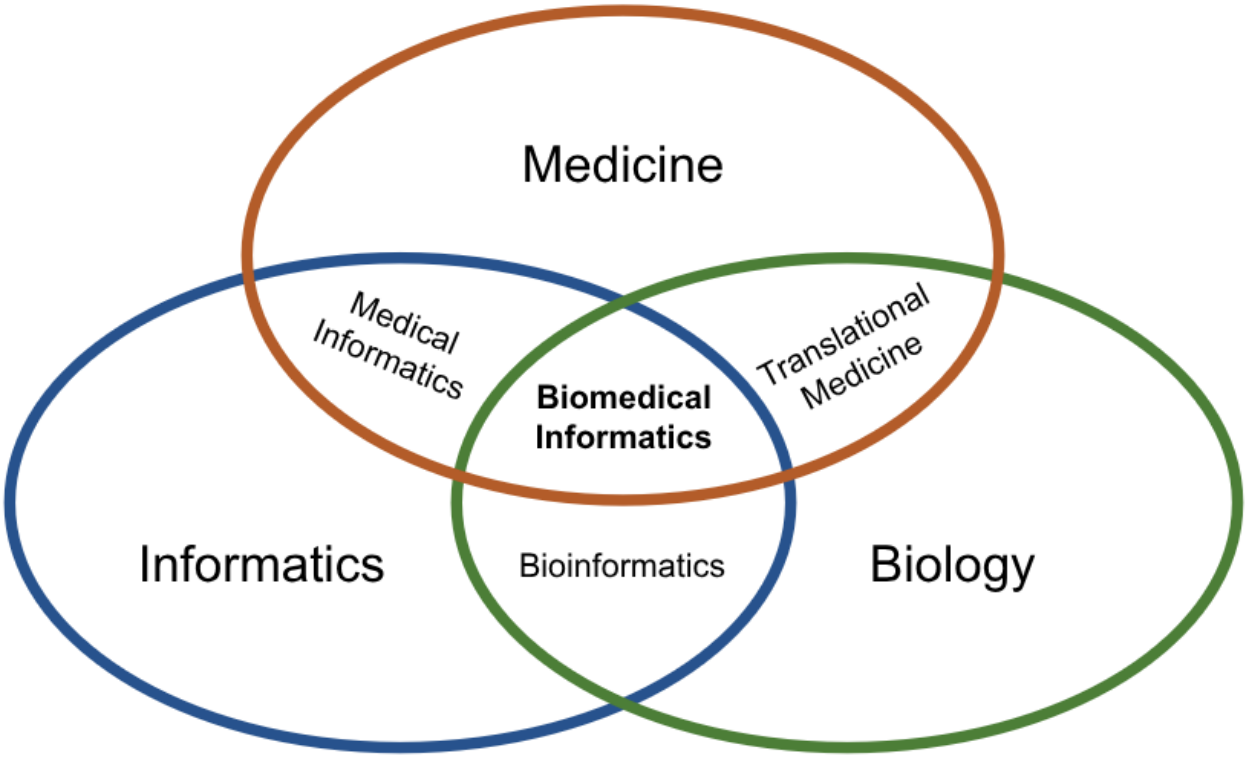


Figure 1: 의생명정보학의 학문적 위치

Figure 1.1 의생명정보학의 학문적 위치 - 의학, 생명과학, 정보학의 융합

따라서 이들의 교차점에 있는 의생명정보학이란 "의학적 문제를 해결하기 위해 생명과학적 데이터와 정보학적 방법을 통합적으로 활용하는 학문" 으로 정의내릴 수 있으며, 이는 곧 중개의학적 성격을 지닌 생명정보학이라고도 할 수 있을 것이다. 따라서 생명정보학에 대해 잘 아는 것이 매우 중요하다.

생명정보학은 그 역사가 다른 학문에 비해 짧으며, 따라서 그 경계 또한 모호한 편이다. 그러나, 많은 경우 생명과학적 데이터, 특히 대규모의 복잡한 데이터셋을 이해하기 위한 방법론과 소프트웨어 도구를 개발하는 학문으로 이해되고 있다. 간단히 말해, 생명과학 "빅데이터"를 다루기 위한 방법론과 도구를 개발하는 분야라고 할 수 있다.

Figure 1.2 생명정보학의 정의 (출처: Wikipedia)

다만 여기서 문제는 "생명과학" 이라는 단어에 있다. 생명이라는 현상은 매우 복잡하며, 단순하지 않다. 따라서 이를 잘 이해하기 위해서는 다양한 학문적 배경이 요구된다.

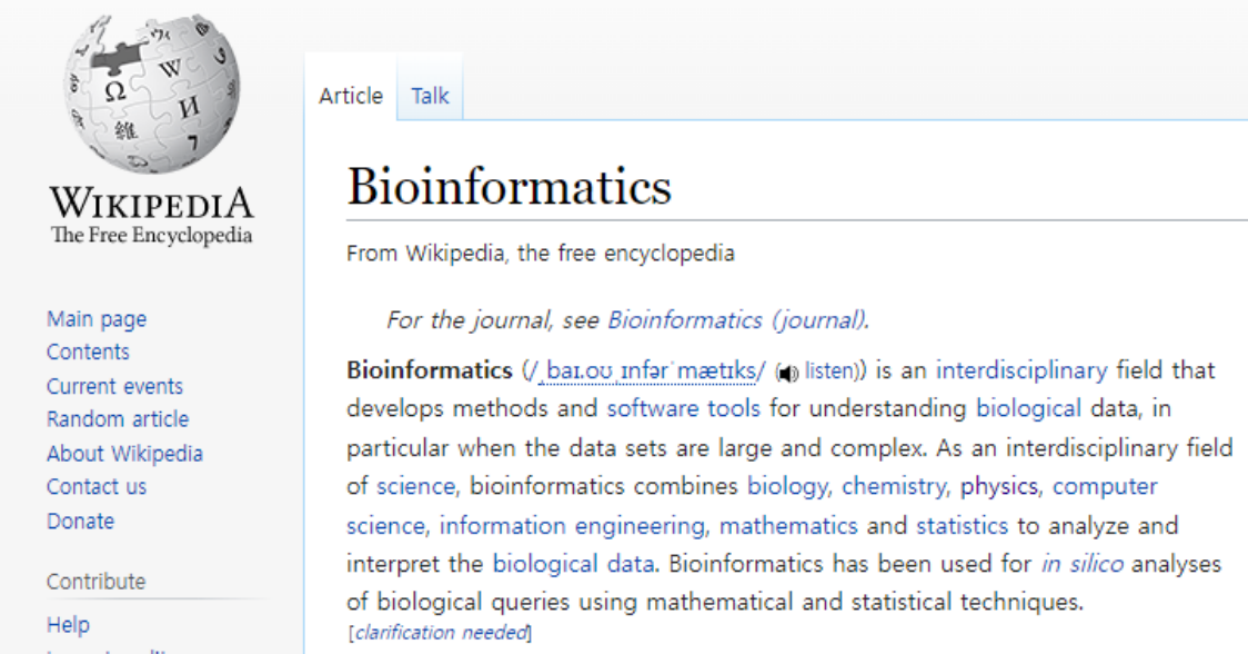


Figure 2: 생명정보학의 정의 – Wikipedia

2.2 생명과학의 발전과 생명정보학의 등장

생명과학은 오랜 역사를 가지고 있다. 고대에는 오감을 통한 생물 분류가 주를 이루었다. 다음은 중국의 황전(□銓, 903~965)이 그린 '사생진금도(□生珍禽□)' 라는 그림인데, 이는 희귀한 생물을 직접 관찰하고 그린 작품이며, 당시의 생명과학적 접근 방식을 보여준다.

Figure 1.3 사생진금도(□生珍禽□) – 고대 중국의 생물 분류 접근법 (황전, 903–965)

시간이 흘러 현미경이 발명되고, 전자현미경 등이 등장하며 현대 생명과학의 근간이 되는 중요한 발견들이 이루어졌다. 각 발견은 현대 생명과학을 이루는 다양한 학문들의 시발점이 되었다.

Figure 1.4 로버트 훅이 사용한 현미경 (17세기)

- **세포생물학**: 현미경으로 코르크를 관찰함으로써 '세포(cell)' 가 처음 발견되었다. 추후 모든 생물이 세포로 이루어져 있음이 알려졌다. 이를 통해 **세포가 생명의 기본 단위**라는 개념이 확립되었다.

Figure 1.5 로버트 훅의 코르크 관찰 – 세포(cell)의 최초 발견 (1665년)

Figure 1.6 세포 구조 – 진핵세포와 원핵세포의 기본 구조

- **분자생물학**: 세포의 내부를 관찰함으로써 DNA, RNA, 단백질 등의 분자를 발견하였다. DNA에서 RNA로, 그리고 단백질로 유전정보가 이어진다는 중심원리(Central Dogma)가 확립되었다. 이는 생명 현상을 이해하는 기본 틀이 되었다.

Figure 1.7 Central Dogma – DNA에서 RNA, 단백질로 이어지는 유전정보의 흐름

Figure 1.8 Central Dogma의 세부 메커니즘 – 전사와 번역 과정

DNA가 생명의 설계도라는 것이 알려진 이후, 컴퓨터를 기반으로 이를 분석함으로써 생명을 이해해보려 하는 시도가 생겨났다. 이는 곧 현대 생물학의 중요한 관심사로 떠오르게 되었으며, "생명정보학" 이 등장하게 되는 계기가 되었다.

특히 인간 지놈 프로젝트 이후 시퀀싱 기술의 발전으로 유전체 분석 비용이 급격히 감소하여, 개인 맞춤 의학이 현실화되고 있다. 동시에 생명과학 분야에서는 데이터의 폭발적 증가가 나타나고 있다.

Figure 1.9 생명과학 분야 데이터 축적량의 급격한 증가 (출처: NIH/NLM)

Figure 3: 사생진금도 – 고대 생물 분류의 예

2.3 체학

2.3.1 "–체(–ome)"과 "–체학(–omics)"의 개념

생명정보학에서는 생명 현상을 총체적으로 이해하기 위해 "–체(–ome)"과 "–체학(–omics)"이라는 개념이 중요하게 사용된다:

- **"–체(–ome)"**: "모든 대상의 집합"을 의미한다.
  - 유전체(Genome): 유전자(Gene) + ome = 모든 유전자의 집합 (일반적으로 DNA 전체를 의미)
  - 전사체(Transcriptome): 전사(Transcription) + ome = 모든 메신저 RNA 분자의 집합
  - 단백체(Proteome): 단백질(Protein) + ome = 모든 단백질의 집합
  - 대사체(Metabolome): 대사물질(Metabolite) + ome = 모든 대사물질의 집합
- **"–체학(–omics)"**: "대상 집합(–ome)에 대한 정량적 분석"을 의미한다.
  - 유전체학(Genomics): 유전체에 대한 연구
  - 전사체학(Transcriptomics): 전사체에 대한 연구
  - 단백체학(Proteomics): 단백체에 대한 연구
  - 대사체학(Metabolomics): 대사체에 대한 연구

2.3.2 인간 지놈 프로젝트와 기술 발전

인간 지놈 프로젝트(Human Genome Project) 는 의생명정보학 발전의 중요한 이정표이다. 1990년 10월부터 2003년 4월까지 진행된 이 프로젝트는 인간의 표준 유전체를 완성하는 데 13년이 걸렸으며, 약 30억 달러가 소요되었다.

그러나 현재는 기술의 발전으로 인해 유전체 해독이 몇 시간에서 며칠 내에 가능해졌으며, 비용도 500~1500 달러로 크게 감소했다. 이로 인해 1000 Genome Project, UK10k 등 대규모 지놈 프로젝트가 가능해졌고, 정밀의학(Precision Medicine)의 토대가 마련되었다. 최근 국내에서도 "울산 10만명 지놈 프로젝트" 등 시도가 이어지고 있다.

Figure 1.10 인간 유전체 해독 비용의 급격한 감소 (2001–2021)

이외에도 기술의 진보는 다양한 –체학 분야에서 대량의 데이터 생산을 가능하게 하고 있다: – **유전체학/후성유전체학**:

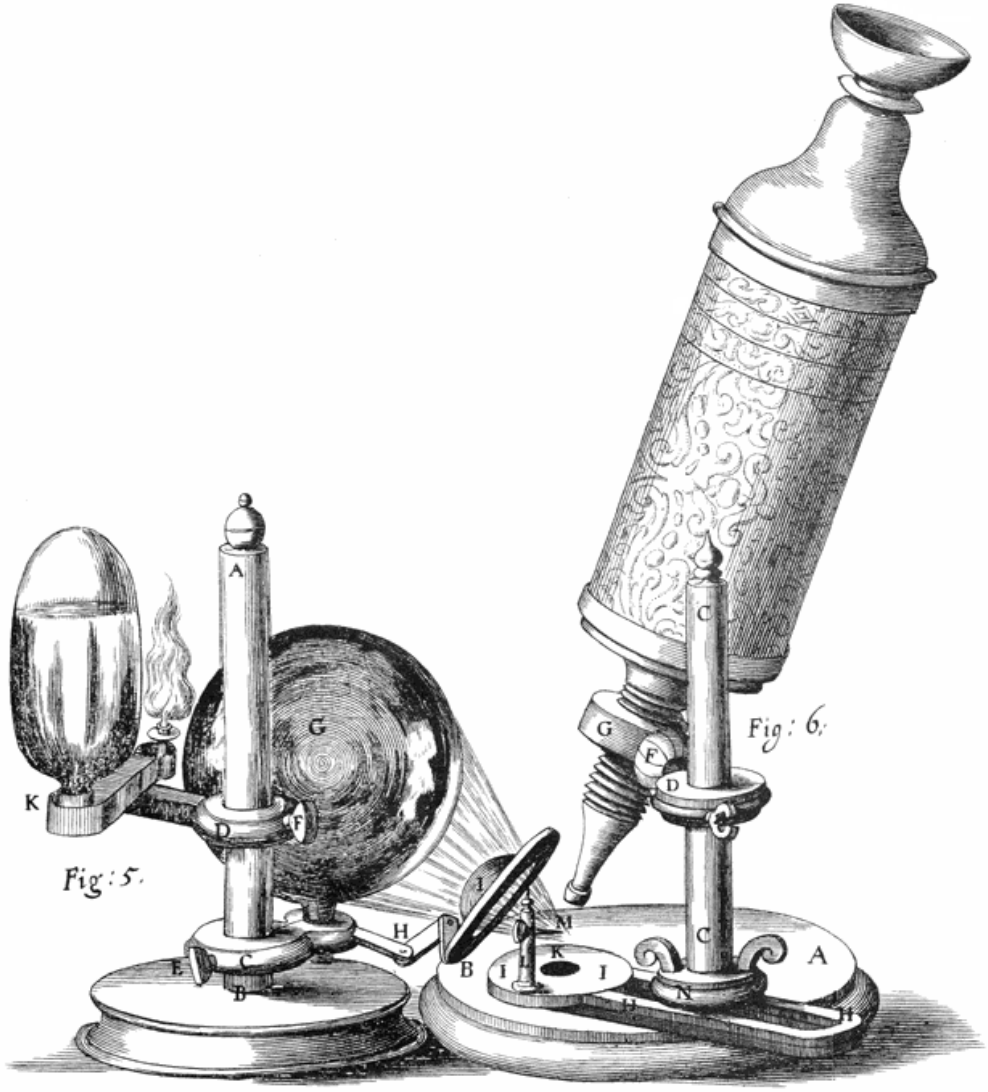


Figure 4: 로버트 훅의 현미경

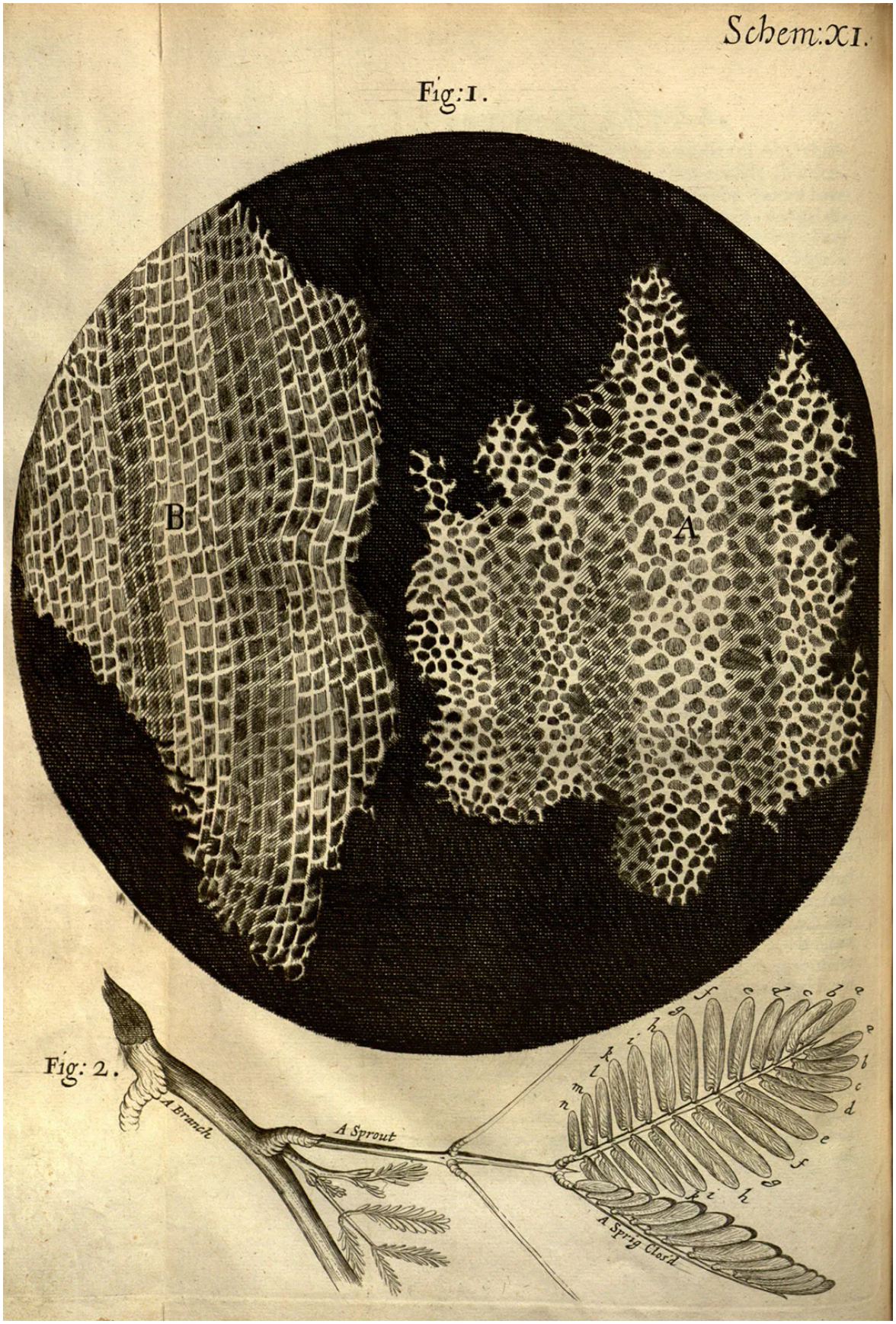


Figure 5: 로버트 훅의 코르크 관찰 – 세포의 최초 발견

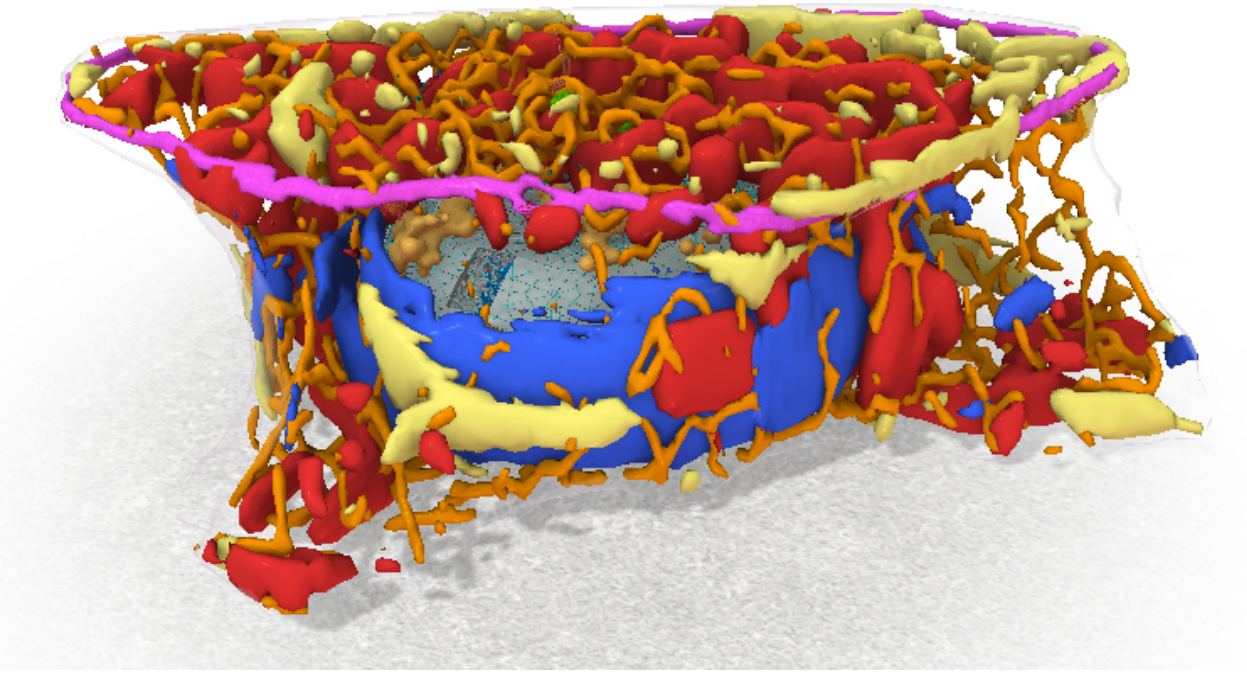


Figure 6: 세포 구조

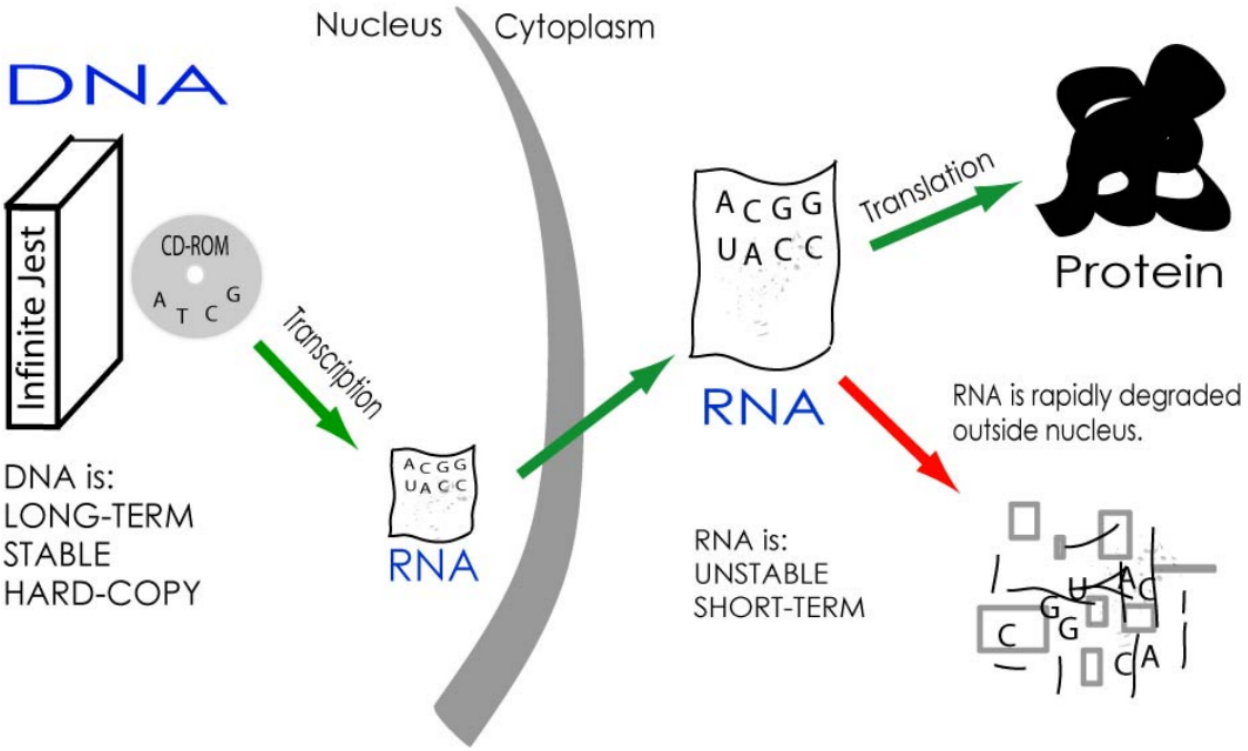


Figure 7: Central Dogma – 유전정보의 흐름

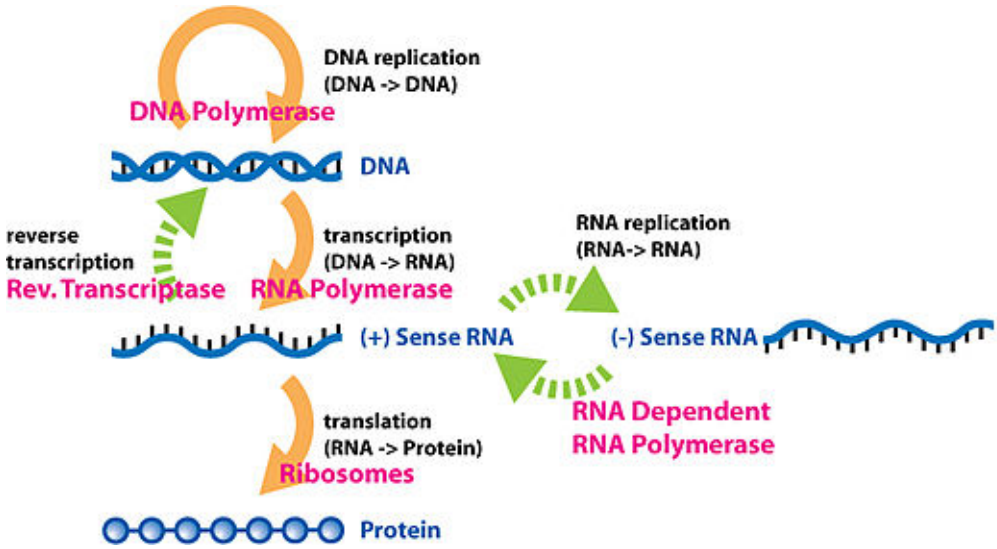


Figure 8: Central Dogma의 세부 메커니즘

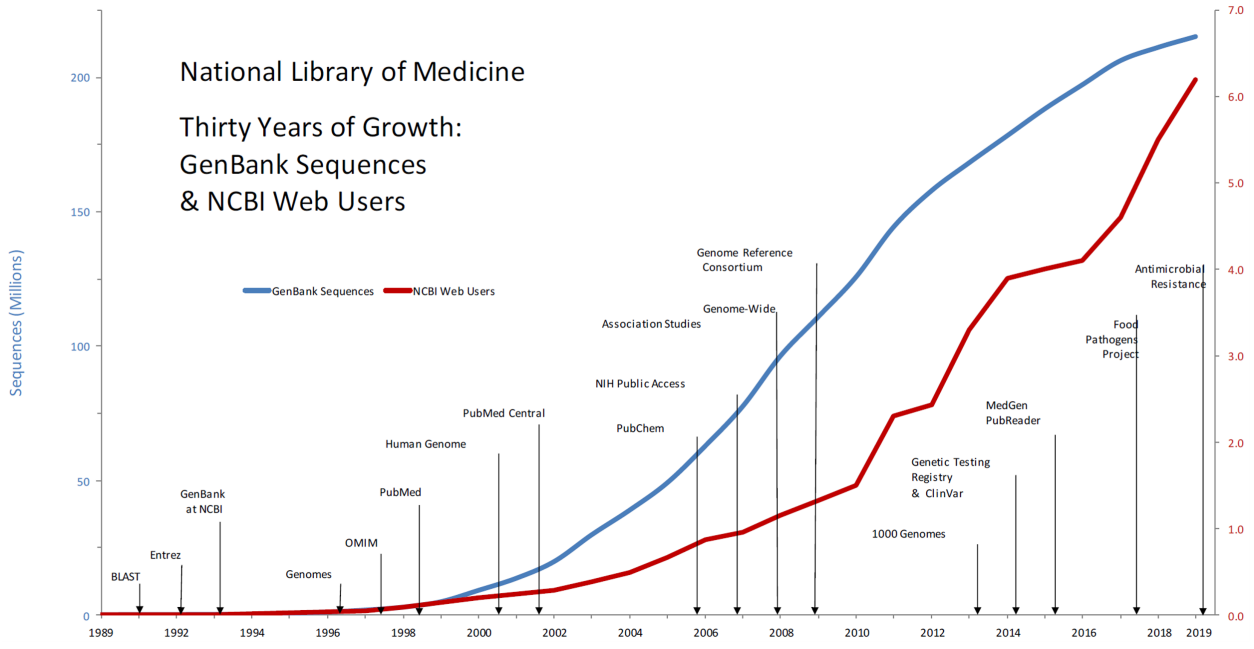


Figure 9: 데이터 축적량의 증가

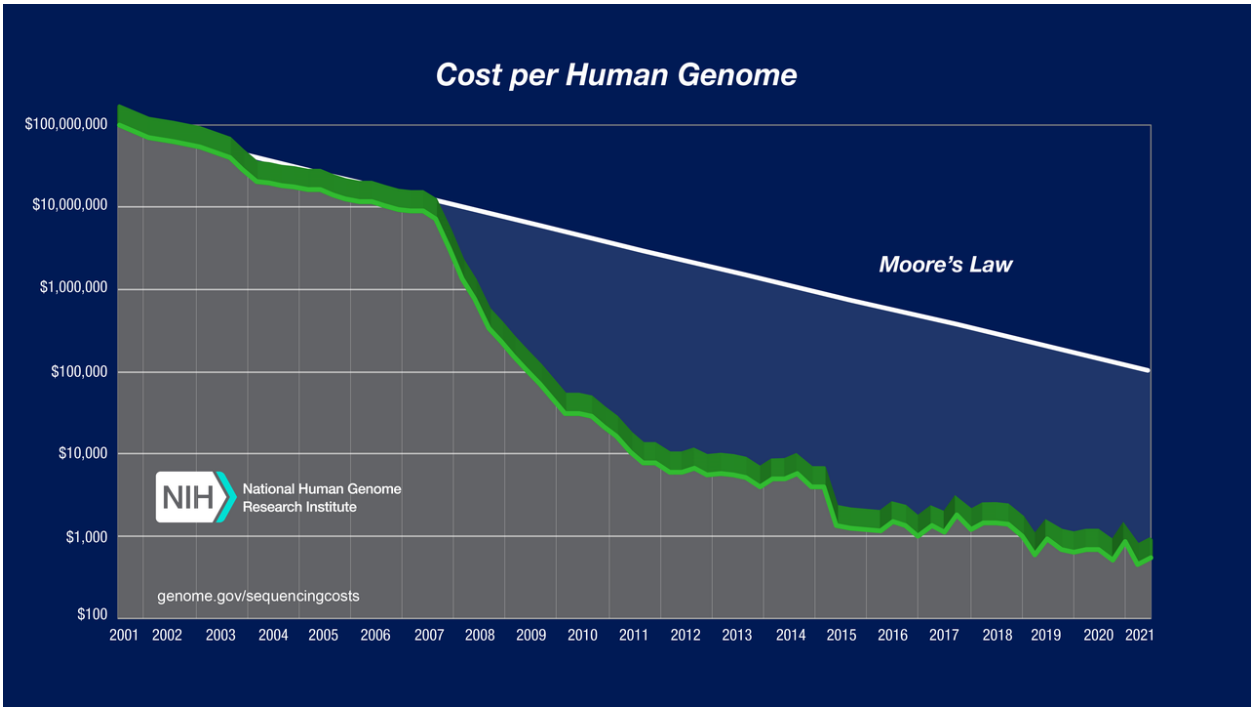


Figure 10: 인간 유전체 해독 비용의 급격한 감소

차세대 시퀀싱(Next Generation Sequencing) 기술의 발전 – **전사체학**: 마이크로어레이(Microarray) 기술의 발전 – **단백체학**: 질량 분석법(Mass Spectrometry)의 발전

2.3.3 후성유전체학

후성유전체학은 DNA 서열 자체의 변화 없이 유전자 발현을 조절하는 메커니즘을 연구하는 분야이다. 이 중에서 중요한 개념인 CpG island는 암 연구에서 특히 중요한 역할을 한다. CpG island는 DNA 상에서 시토신(C)과 구아닌(G) 염기가 인접하게 위치하는 부분이 집중적으로 나타나는 부위를 말한다.

Figure 1.11 CpG Island 메틸화와 유전자 발현 조절 메커니즘

이러한 CpG island의 메틸화(methylation) 패턴은 유전자 발현 조절에 중요한 역할을 하며, 비정상적인 메틸화는 암의 발생과 밀접한 연관이 있다. 특히, 종양 억제 유전자의 프로모터 부위에 있는 CpG island가 과메틸화(hypermethylation)되면 해당 유전자의 발현이 억제되어 암 발생 위험이 증가할 수 있다. 반대로, 암 유전자(oncogene)의 프로모터 부위에 있는 CpG island가 저메틸화(hypomethylation)되면 해당 유전자의 발현이 증가하여 암 발생 위험이 증가할 수 있다.

Figure 1.12 정상 대장 상피세포와 대장암에서의 메틸화 패턴 차이

2.4 이미지 데이터

의생명정보학에서는 이미지 데이터 분석 또한 중요한 영역을 차지한다.

Figure 1.13 주요 이미징 기술 – 밝은 시야 이미징, 위상차 이미징, 형광 이미징

주요 이미징 기술로는 다음과 같은 것들이 있다:

- **밝은 시야 현미경법(Bright Field Microscopy)**: 가장 기본적인 현미경 기술로, 빛이 시료를 통과하여 명암 대비를 통해 세포 구조를 관찰할 수 있다. 염색 기법과 함께 사용되어 특정 세포 구조를 시각화할 수 있다.

Figure 1.14 밝은 시야 현미경법으로 관찰한 식물 세포 구조

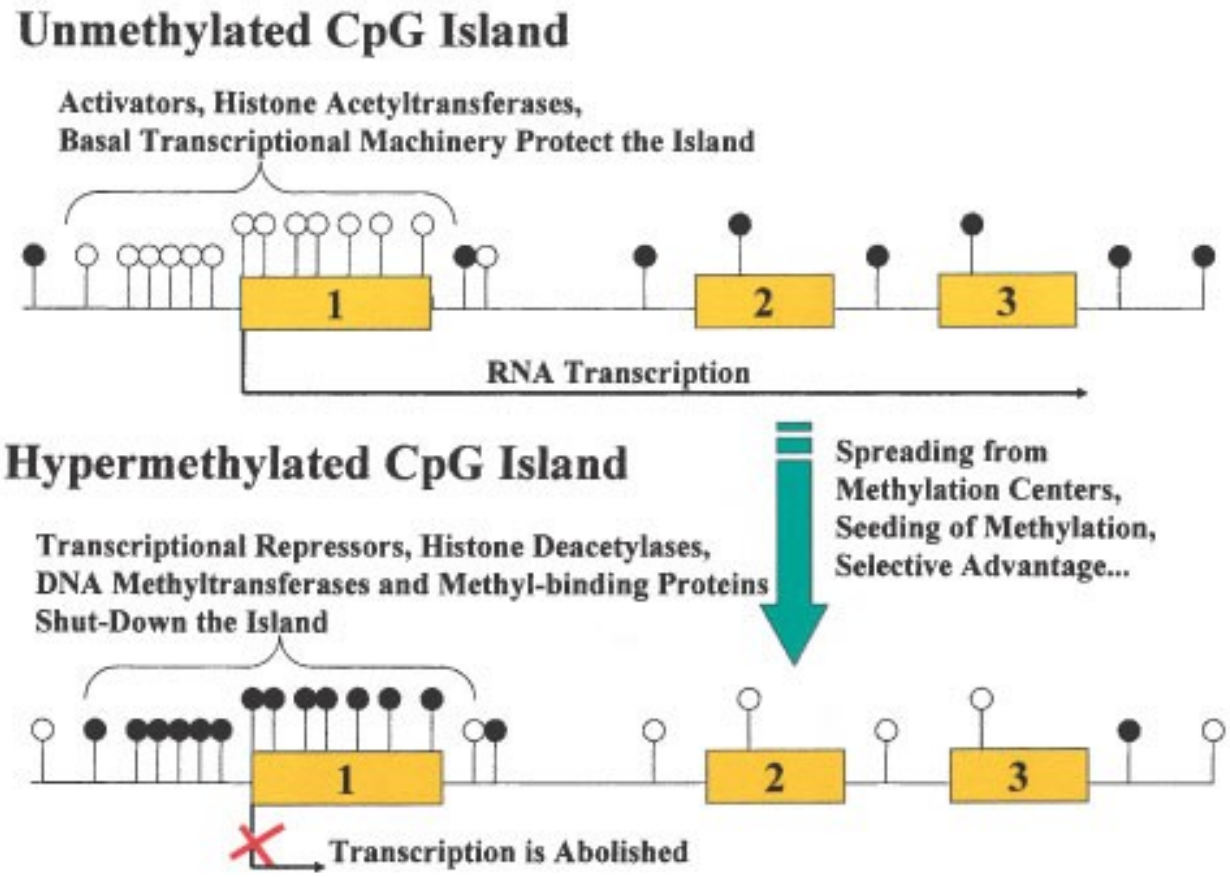


Figure 11: CpG Island 메틸화와 유전자 발현 조절

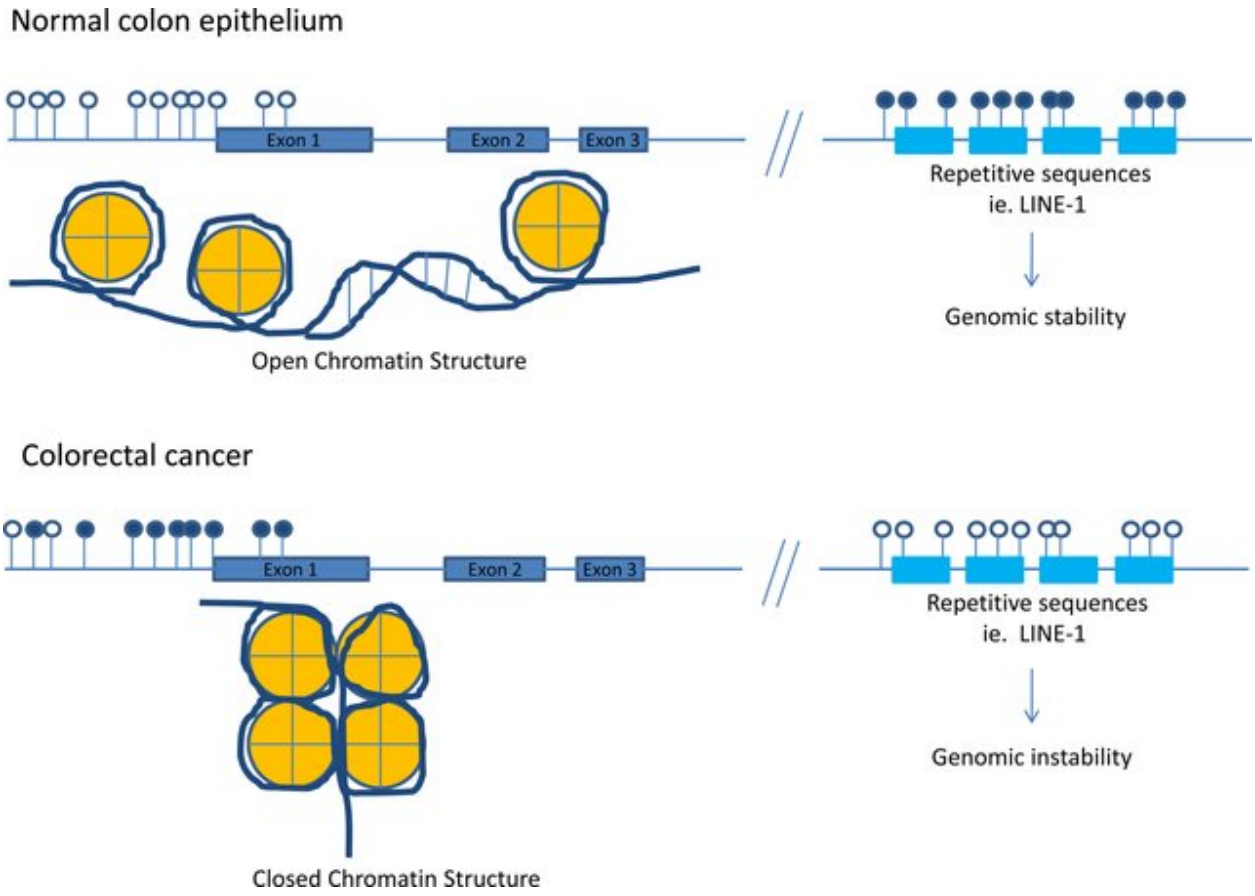


Figure 12: 정상 세포와 암세포에서의 메틸화 패턴 차이

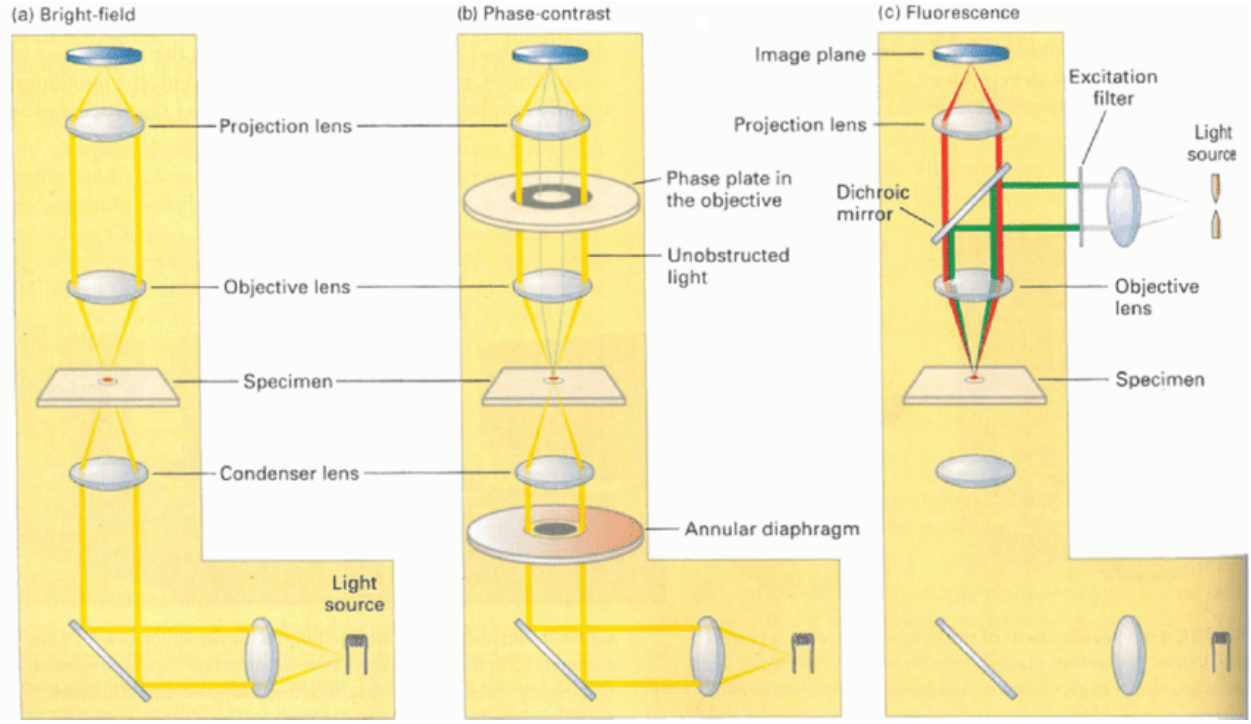


Figure 13: 주요 이미징 기술

- **위상차 현미경법(Phase Contrast Microscopy)**: 투명한 생체 시료의 위상차를 이용하여 염색 없이도 세포 구조를 선명하게 관찰할 수 있는 기술이다. 살아있는 세포의 동적 변화를 실시간으로 관찰하는 데 특히 유용하다.

Figure 1.15 위상차 현미경법으로 관찰한 살아있는 세포

- **형광 현미경법(Fluorescence Microscopy)**: 형광 염료나 단백질을 사용하여 특정 세포 구조나 단백질을 표지하고, 특정 파장의 빛을 조사하여 형광 신호를 관찰하는 기술이다. 이를 통해 특정 단백질의 위치, 상호작용, 발현 양상 등을 시각화할 수 있다.

Figure 1.16 형광 현미경을 이용한 다중 채널 이미징 (Nucleus, Vimentin, Actin)

이러한 이미지 데이터는 세포 및 조직 수준에서의 생물학적 현상을 이해하는 데 중요한 정보를 제공하며, 최근에는 딥러닝과 같은 인공지능 기술을 활용한 이미지 분석이 활발히 연구되고 있다.

2.5 단백질 구조 데이터

단백질의 3차원 구조를 분석하는 것은 단백질의 기능을 이해하고 약물 개발에 활용하는 데 필수적이다. 주요 단백질 구조 분석 기술로는 다음과 같은 것들이 있다:

- **X선 결정학(X-ray Crystallography, XRD)**: 단백질 결정에 X선을 조사하여 회절 패턴을 분석함으로써 원자 수준에서의 단백질 구조를 결정하는 기술이다. 가장 오래된 방법이지만 여전히 높은 해상도의 구조 정보를 제공한다.

Figure 1.17 X선 결정학 과정 – 결정에서 원자 모델까지

- **극저온 전자현미경법(Cryo-Electron Microscopy, Cryo-EM)**: 단백질 시료를 초저온으로 급속 동결시킨 후 전자현미경으로 관찰하는 기술이다. 최근 기술의 발전으로 원자 수준의 해상도를 달성할 수 있게 되었으며, 결정화가 어려운 큰 단백질 복합체 구조 연구에 특히 유용하다. 이 기술의 혁신성을 인정받아 2017년 노벨 화학상을 수상했다.

Figure 1.18 2017년 노벨 화학상 수상자들 – Cryo-EM 기술 개발

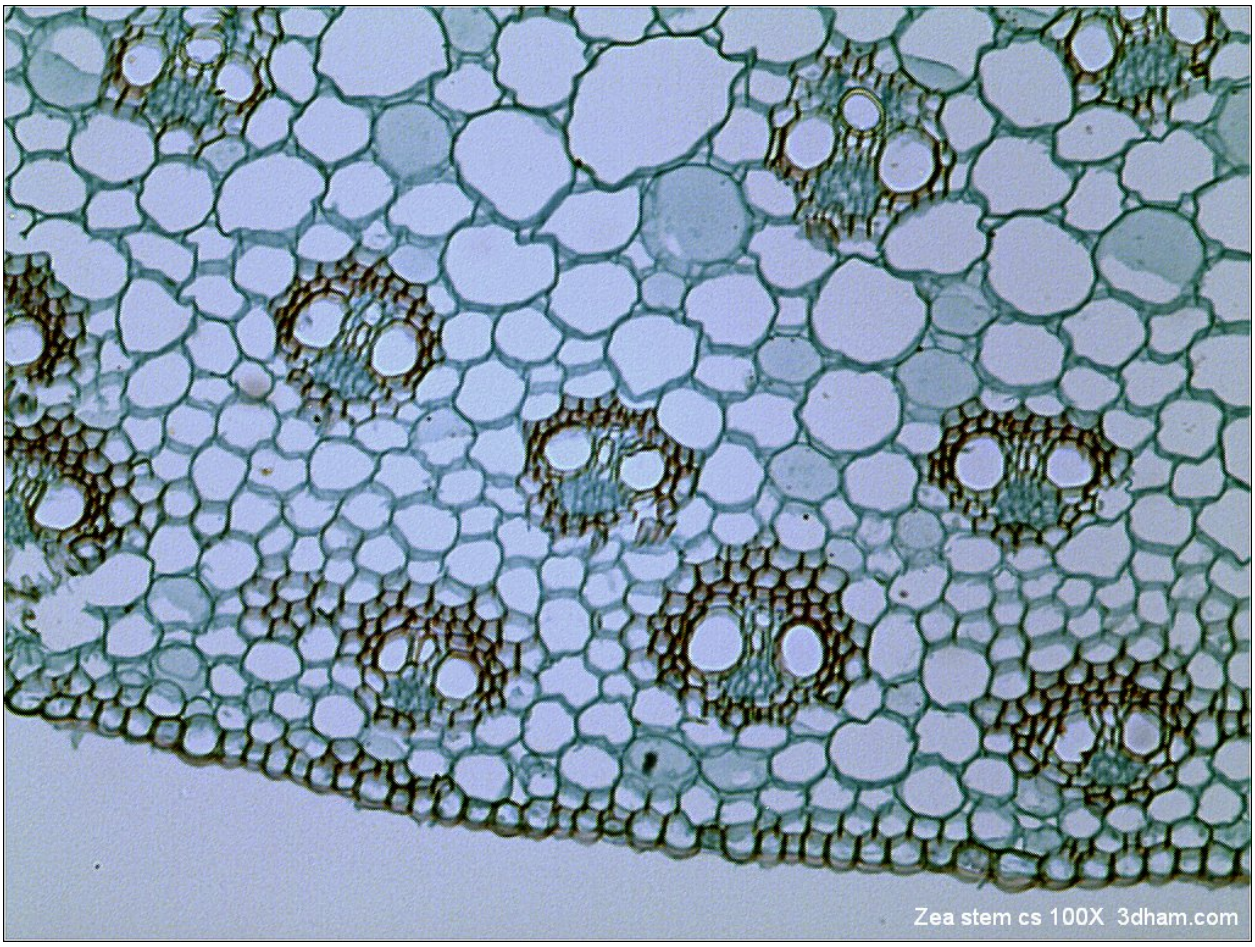


Figure 14: 밝은 시야 현미경으로 관찰한 세포 구조

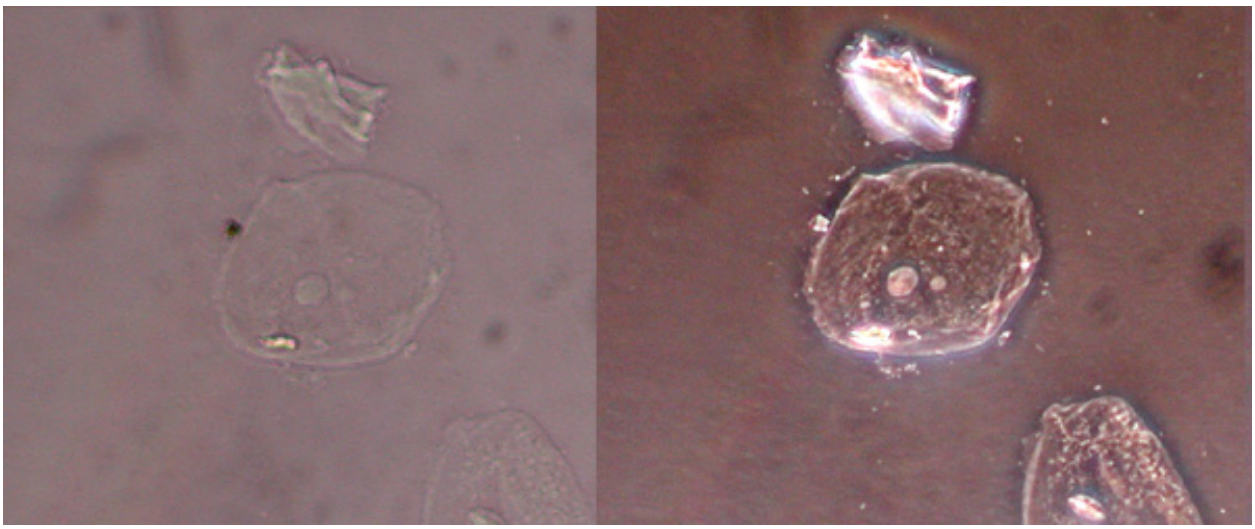


Figure 15: 위상차 현미경으로 관찰한 세포

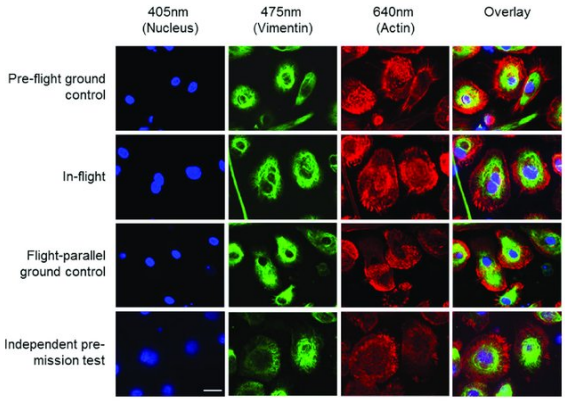


Figure 16: 형광 현미경을 이용한 다중 채널 이미징

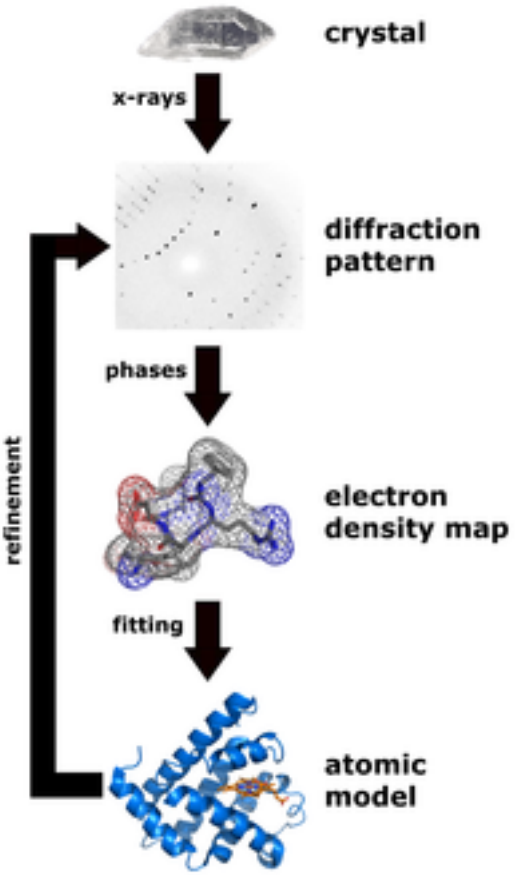


Figure 17: X선 결정학 과정 – 결정에서 원자 모델까지

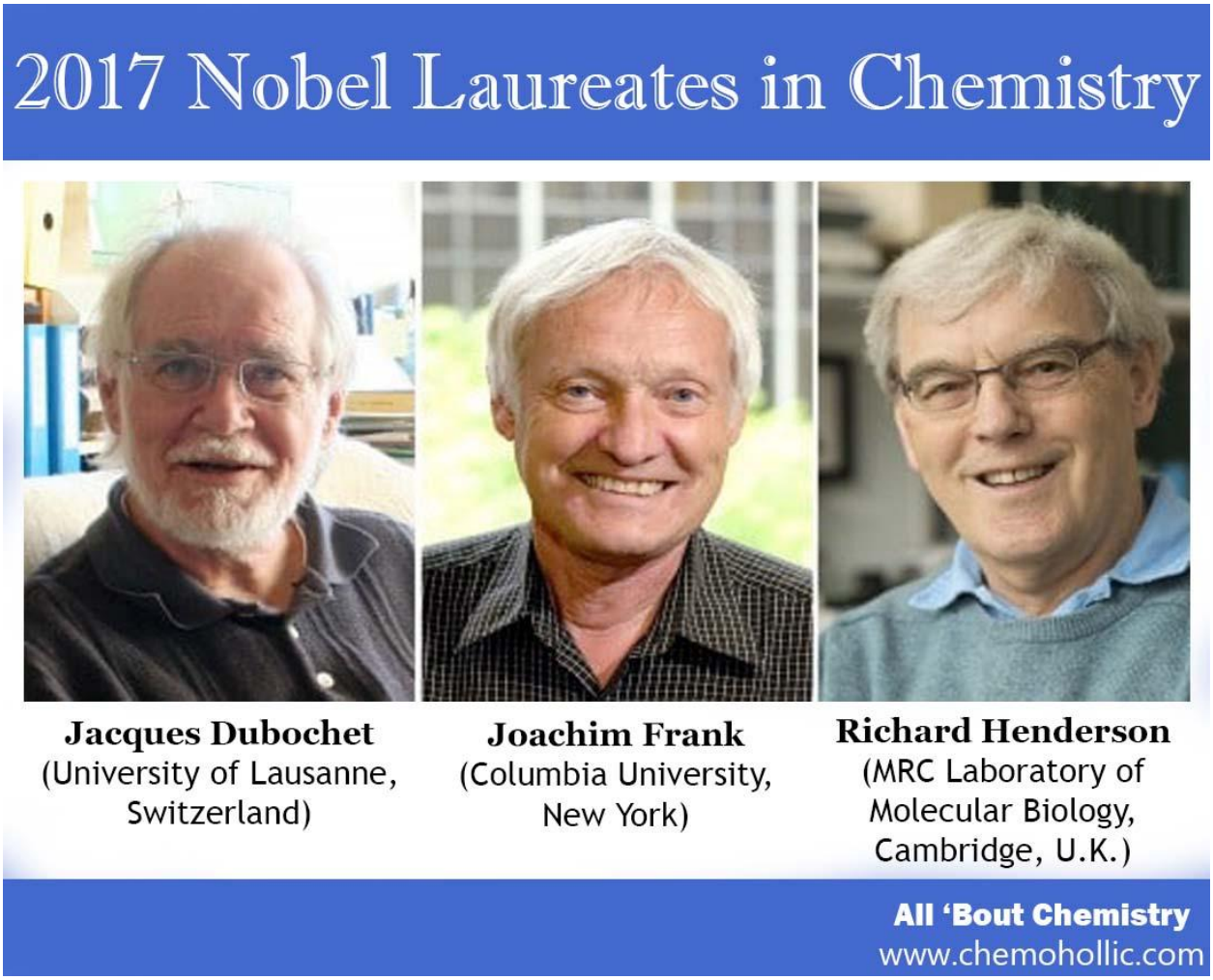


Figure 18: 2017년 노벨 화학상 수상자들 – Cryo-EM 기술 개발

- **단백질 접힘 예측(Protein Folding)**: 컴퓨터 알고리즘을 이용하여 단백질의 아미노산 서열로부터 3차원 구조를 예측하는 기술이다. 최근 DeepMind의 AlphaFold2와 같은 인공지능 기반 접근법이 획기적인 성과를 보이며 단백질 구조 생물학 분야에 혁명을 일으키고 있다.

Figure 19: AlphaFold – 50년간의 단백질 구조 예측 문제 해결

Figure 1.19 AlphaFold – 50년간의 단백질 구조 예측 문제 해결

이러한 구조 데이터는 PDB(Protein Data Bank)와 같은 데이터베이스에 저장되어 연구자들이 공유하고 활용할 수 있다.

2.6 의료정보학적 데이터

의료정보학은 의료 현장에서 생산된 빅데이터를 분석하는 방법론과 도구를 개발하는 분야로, 의생명정보학의 중요한 부분을 차지한다. 의료정보학적 데이터에는 다음과 같은 것들이 포함된다:

2.6.1 전자의무기록(Electronic Medical Record, EMR)

전자의무기록은 환자의 진료 정보를 디지털 형태로 저장한 데이터로, 다음과 같은 정보를 포함한다: – 진찰 기록 – 환자 이력 (처방력, 술/담배 이력 등) – 정량적 데이터 (키, 몸무게 등) – 영상 및 음성 데이터

EMR은 의료 데이터의 핵심으로, 환자 케어의 연속성을 보장하고 의료 의사결정을 지원하는 데 중요한 역할을 한다.

Figure 1.20 전자의무기록(EMR) 시스템 인터페이스

2.6.2 국제질병분류(International Classification of Diseases, ICD)

ICD는 질병 진단을 위한 국제 표준 코드 시스템이다: – 세계보건기구(World Health Organization, WHO)에서 관리한다. – 현재 최신 버전은 ICD-11이다. – 한국에서는 한국표준질병사인분류(Korean Standard Classification of Diseases, KCD)라는 한국 버전의 ICD를 사용한다.

Figure 1.21 ICD-11 질병 분류 시스템 (결핵 관련 코드 예시)

이 표준화된 코드 시스템은 의료 데이터의 일관성을 유지하고, 국제적인 데이터 공유와 비교를 가능하게 한다.

2.6.3 의료영상 및 통신 표준(DICOM)

DICOM(Digital Imaging and Communications in Medicine)은 표준 의료영상/기록 포맷이다: – 주로 병원 내 'PACS(Picture, Archive, Communications)' 시스템에 의해 처리된다. – 현재 거의 대부분의 의료영상기록이 DICOM 포맷으로 저장된다. – X-ray, CT, MRI 등 다양한 의료 영상 데이터를 포함한다.

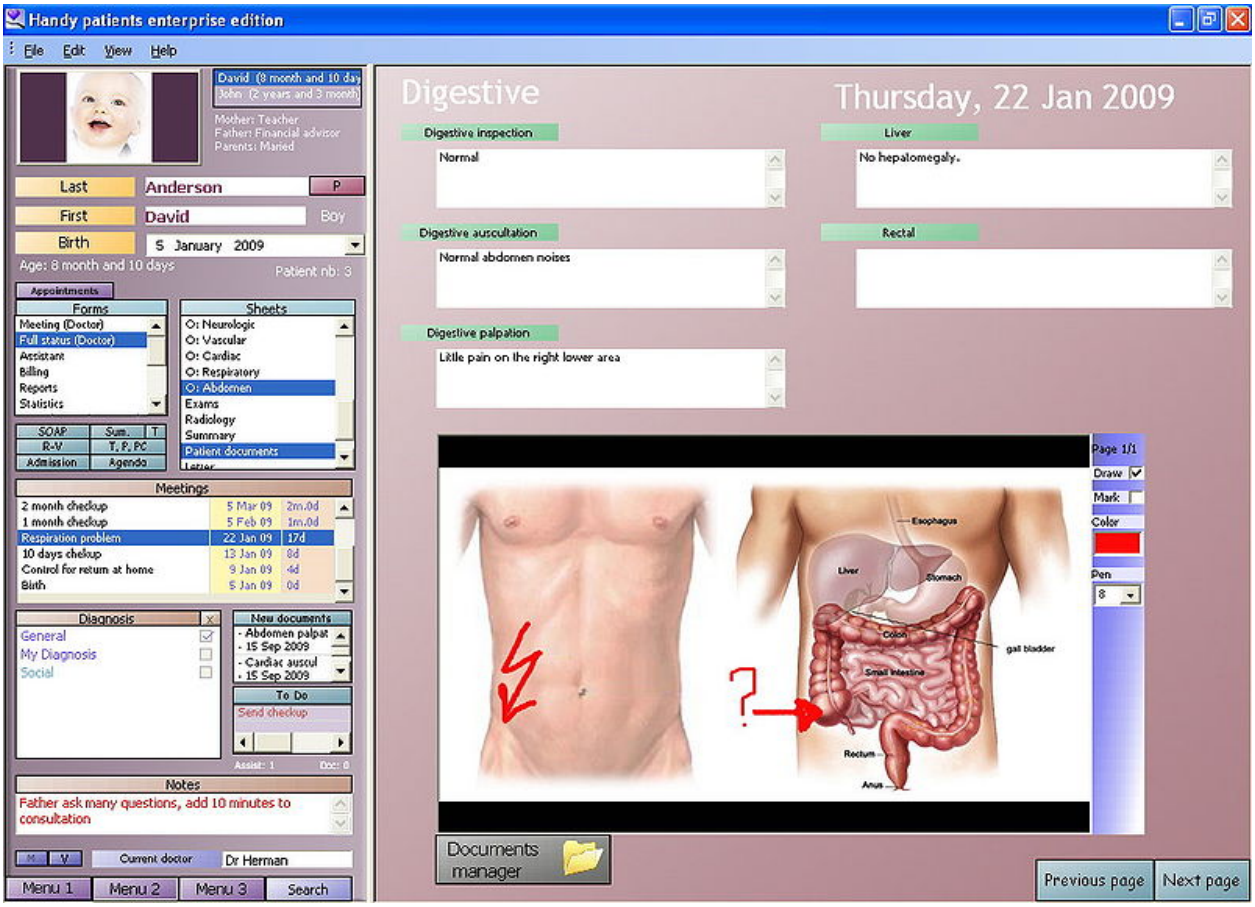


Figure 20: 전자의무기록(EMR) 시스템 인터페이스

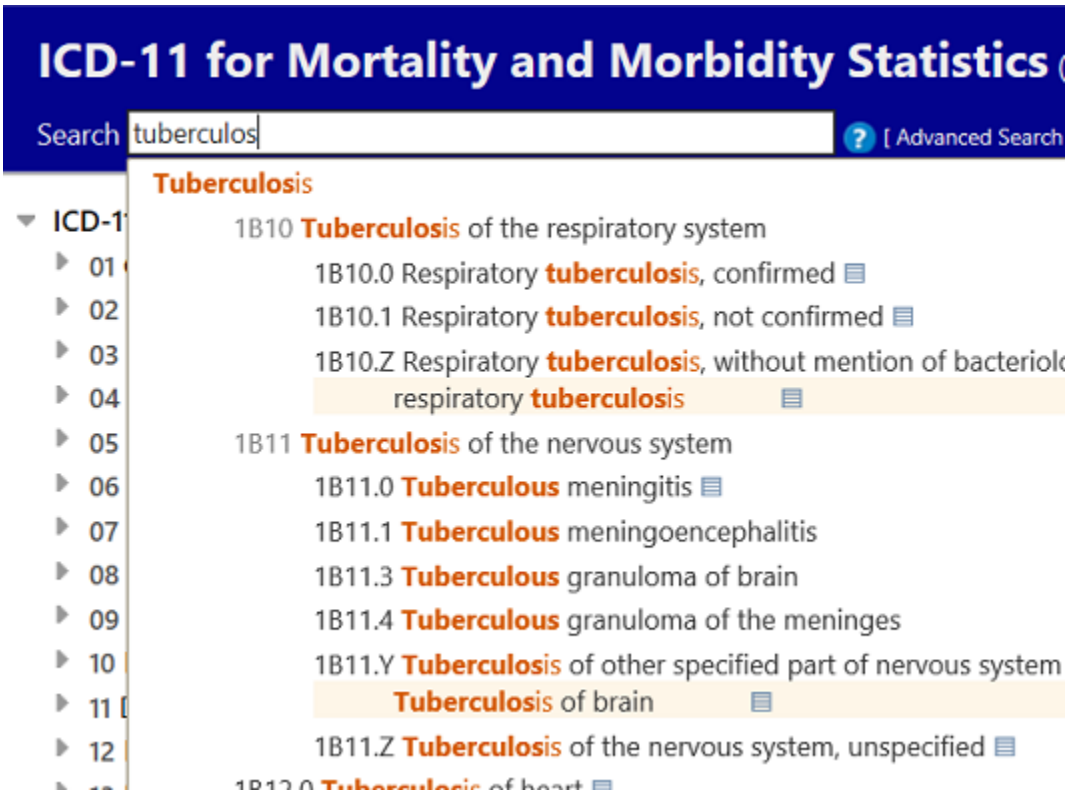


Figure 21: ICD-11 질병 분류 시스템

Figure 22: X-ray 영상

Figure 1.22 X-ray 영상 – 의료 영상의 기본적인 형태

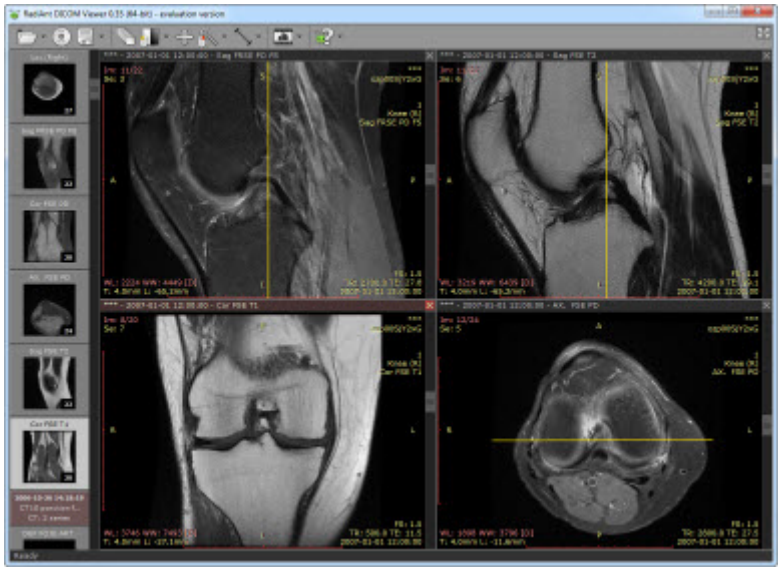


Figure 23: DICOM 뷰어를 통한 의료 영상 데이터 관리

Figure 1.23 DICOM 뷰어를 통한 의료 영상 데이터 관리

2.6.4 비디지털 데이터와 그 처리

그러나 모든 의료 데이터가 디지털 포맷으로 저장되어 있지는 않다: – 수기 및 필름 데이터가 여전히 존재한다.

Figure 1.24 의사의 손글씨 진료 기록 – 비디지털 의료 데이터의 예 – 이러한 데이터를 일일이 디지털 포맷으로 변환하는 것은 매우 어렵다. – 이를 해결하기 위해 영상처리/필기인식 기술과 자연어 처리 기술이 활용되고 있다.

2.6.5 통계 데이터

의료 분야에는 다양한 통계 데이터가 존재한다: – 임상시험 데이터 – 역학(Epidemiology) 데이터 – 메타분석 데이터

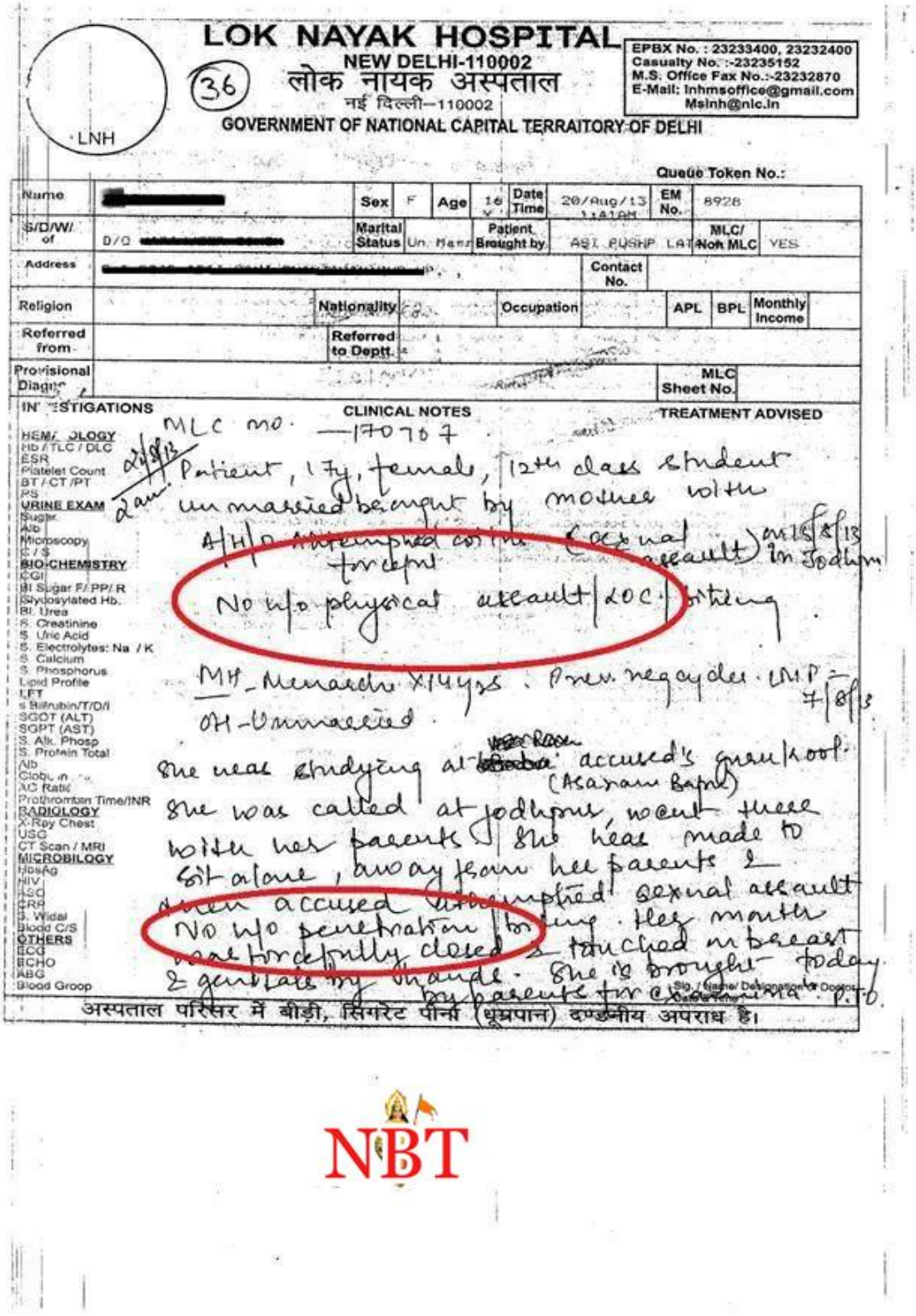
이러한 통계 데이터는 치료 효과 검증, 질병 패턴 분석, 공중 보건 정책 수립 등에 활용된다.

2.7 현대 생명정보학의 동향

이렇듯 컴퓨터를 활용한 체학 분석이 중요해지면서, 현대 생명과학의 연구에 있어서도 생명정보학이 점차 주도적 위치를 차지하고 있다. 이는 생명과학적 발견이 실험실 위주에서 데이터 분석 중심으로 변화하고 있음을 의미한다.

생명정보학 데이터 분석은 단순한 도구 사용을 넘어, 생명과학적 문제를 해결하기 위한 통합적 접근을 요구한다. 초창기에는 생명과학적 지식이 부족한 컴퓨터 과학자나 수학자들이 생명과학자들에게 보조적인 역할을 제공해 주는 식으로 연구가 이루어졌으나, 최근에는 생명과학적 지식, 통계적 방법론, 프로그래밍 기술을 모두 어느정도 갖추고 있는 생명정보학자들이 주체적으로 연구를 수행하는 형태로 점차 변화하고 있다.

Figure 1.25 계산생물학자의 역할 변화 – 과거 보조적 역할에서 현재 연구 주도적 역할로



LOK NAYAK HOSPITAL
NEW DELHI-110002
लोक नायक अस्पताल
नई दिल्ली–110002
GOVERNMENT OF NATIONAL CAPITAL TERRAITORY OF DELHI

EPBX No. : 23233400, 23232400
Casualty No. :-23235152
M.S. Office Fax No.:-23232870
E-Mail: lnhmsoffice@gmail.com
Msinh@nic.in

LNH

36

Queue Token No.:

Name	[redacted]	Sex	F	Age	16 Y	Date Time	20/Aug/13 11:41AM	EM No.	8928
S/D/W/ of	D/O [redacted]	Marital Status	Un. Marr.	Patient Brought by		ASI PUSHP LAT		MLC/ Non MLC	YES
Address	[redacted]						Contact No.		
Religion		Nationality		Occupation			APL	BPL	Monthly Income
Referred from		Referred to Deptt.							
Provisional Diagnosis							MLC Sheet No.		

INVESTIGATIONS

HEMATOLOGY
Hb / TLC / DLC
ESR
Platelet Count
BT / CT /PT
PS
URINE EXAM
Sugar
Alb
Microscopy
C / S
BIOCHEMISTRY
CGI
Bl Sugar F/PP/R
Glycosylated Hb.
Bl. Urea
S. Creatinine
S. Uric Acid
S. Electrolytes: Na / K
S. Calcium
S. Phosphorus
Lipid Profile
LFT
s Bilirubin/T/D/I
SGOT (ALT)
SGPT (AST)
S. Alk. Phosp
S. Protein Total
Alb
Globulin
A:G Ratio
Prothrombin Time/INR
RADIOLOGY
X-Ray Chest
USG
CT Scan / MRI
MICROBILOGY
HbsAg
HIV
VDRL
CBP
S. Widal
Blood C/S
OTHERS
ECG
ECHO
ABG
Blood Group

CLINICAL NOTES

MLC no. — 170707

24/8/13 2am. Patient, 17y, female, 12th class student unmarried brought by mother with A/H/O Attempted coitus (sexual assault) on 18/8/13 in Jodhpur [illegible] No H/o physical assault / LOC / bleeding.

MH - Menarche X14yrs. Prev. regular. LMP - 7/8/13

OH - Unmarried.

She was studying at [illegible] accused's room / roof (Asaram Bapu)

She was called at Jodhpur, w/ out there with her parents & she was made to sit alone, away from her parents & then accused attempted sexual assault N/O H/O penetration [illegible] Her mouth was forcefully closed & touched in breast & genitals by hands. She is brought today by parents for [illegible]

TREATMENT ADVISED

Sig. (Name/ Designation & Doctor's Code No.)

अस्पताल परिसर में बीड़ी, सिगरेट पीना (धूम्रपान) दण्डनीय अपराध है।

NBT

Figure 24: 의사의 손글씨 진료 기록

Table 1 Shifting roles of computational biologists

	Past	Current
Role in research	Supportive	Driver of research
A feeling for the biology	Computer science-centered	Biology- and computer science-centered
Environment	Isolated	Integrated
Data generation	Constrained	Resourceful
Data exploration	Largely limited to hypothesis testing	Both exploratory and hypothesis testing

Figure 25: 계산생물학자의 역할 변화

3 유전체학 기초

3.1 유전체와 유전체학의 기본 개념

3.1.1 유전체의 정의와 표준 유전체

유전체(Genome)는 한 생물체가 가지고 있는 모든 유전 정보의 총체이다. 생명정보학 연구에서는 개별 생물체의 유전체를 분석하기 위한 기준으로 표준 유전체(Reference Genome)를 사용한다. 표준 유전체는 한 종의 대표적인 유전체 서열로, 국제적인 협력을 통해 구축되고 지속적으로 개선된다.

Figure 2.1 Central Dogma의 기본 원리 – DNA에서 RNA, 단백질로 이어지는 유전정보의 흐름

Figure 2.2 DNA의 방향성 – 5'에서 3' 방향으로의 정보 흐름

유전체 참조 컨소시엄(Genome Reference Consortium, GRC)은 표준 유전체의 구축과 관리를 담당하는 국제 조직으로, 다음과 같은 기관들로 구성되어 있다: – 미국 국립생물정보센터(The National Center for Biotechnology Information, NCBI) – 워싱턴 대학교 맥도넬 유전체 연구소(The McDonnell Genome Institute at Washington University, MGI) – 유럽 생명정보학 연구소(The European Bioinformatics Institute, EBI) – 웰컴 생거 연구소(The Wellcome Sanger Institute) – 제브라피시 모델 생물 데이터베이스(The Zebrafish Model Organism Database) – 시궁쥐 유전체 데이터베이스(The Rat Genome Database)

GRC에서 관리하는 주요 생물종의 표준 유전체는 다음과 같다: – 인간(Homo Sapiens): GRCh38.p14 – 생쥐(Mus Musculus): GRCm39 – 제브라피시(Danio Rerio): GRCz11 – 닭(Gallus Gallus): GRCg6a – 시궁쥐(Rattus Norvegicus): GRCr8

Figure 2.3 GRC 기관들 – 유전체 참조 컨소시엄을 구성하는 주요 기관들

3.1.2 모델 생물

생명과학 연구에서는 다양한 모델 생물이 사용되며, 각 모델 생물에 대한 전문 데이터베이스가 존재한다: – WormBase: 예쁜꼬마선충(C. Elegans) 관련 데이터베이스 – FlyBase: 초파리(Drosophila Melanogaster) 관련 데이터베이스 – The Arabidopsis Information Resource (TAIR): 애기장대(Arabidopsis Thaliana) 관련 데이터베이스 – Xenbase: 개구리(Xenopus Laevis) 관련 데이터베이스 – Saccharomyces Genome Database (SGD): 효모(Saccharomyces Cerevisiae) 관련 데이터베이스

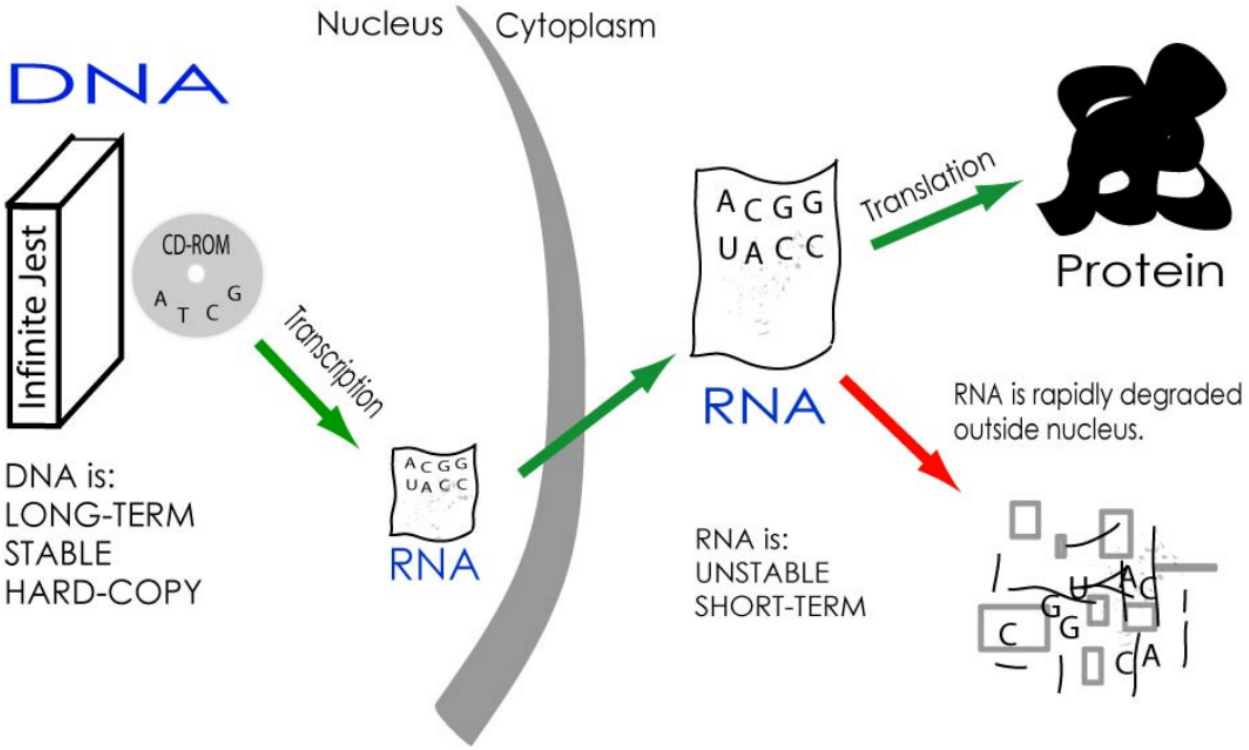


Figure 26: Central Dogma의 기본 원리

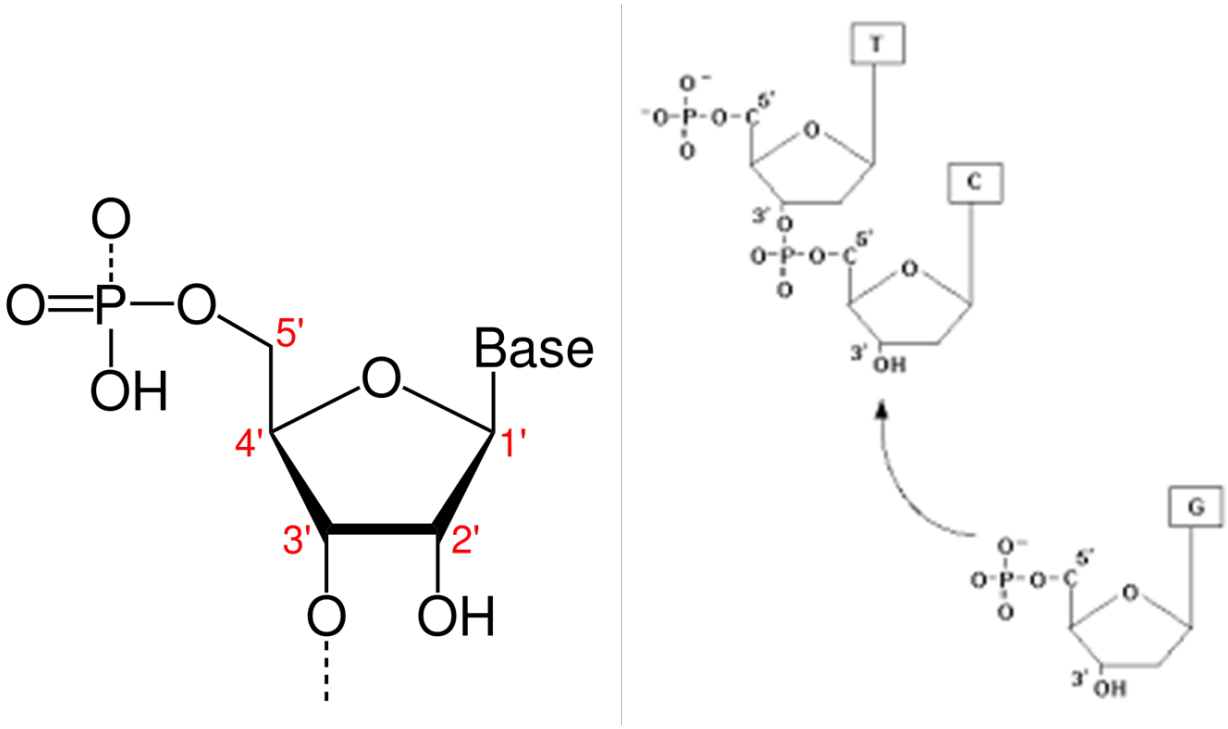


Figure 27: DNA의 방향성

Figure 28: GRC 기관들

위에서 언급한 생쥐와 제브라피시도 중요한 모델 생물이다.

3.1.3 유전체 브라우저

유전체 데이터를 시각화하고 분석하기 위한 도구로 유전체 브라우저가 사용된다. 대표적인 유전체 브라우저는 다음과 같다:

- Ensembl Genome Browser: 유럽 생명정보학 연구소(EBI)에서 개발한 브라우저

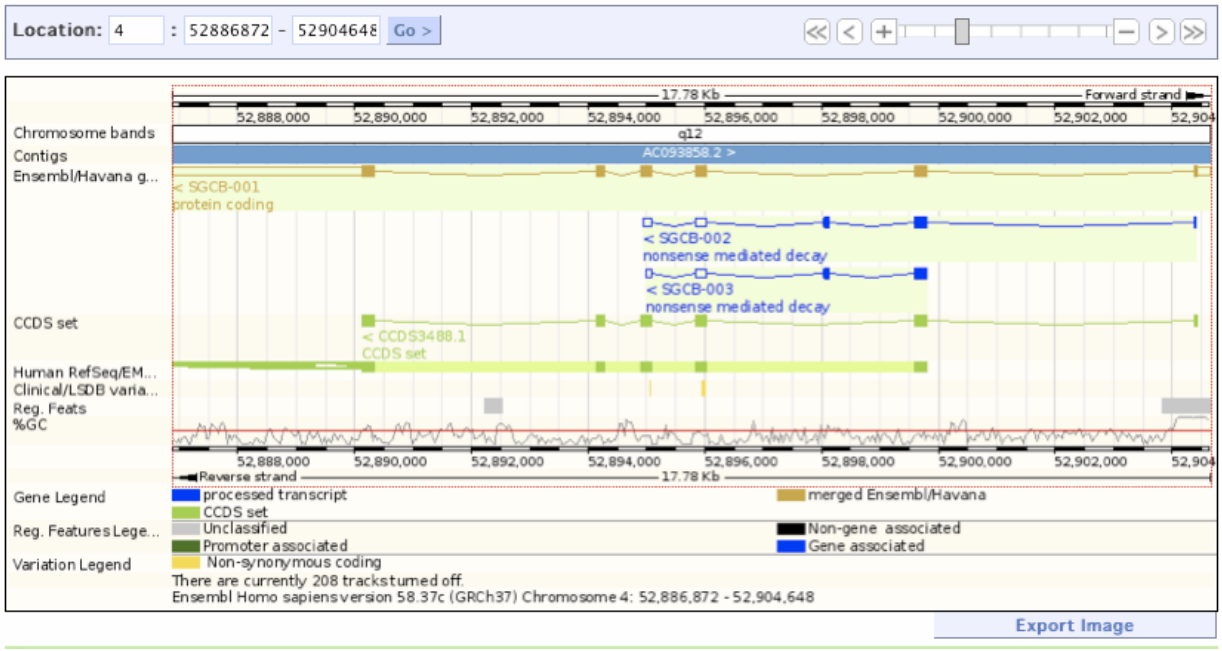


Figure 29: Ensembl Genome Browser

Figure 2.4 Ensembl Genome Browser – 유전체 데이터 시각화 인터페이스

- UCSC Genome Browser: 미국 캘리포니아 산타크루즈 대학에서 개발한 브라우저

Figure 2.5 UCSC Genome Browser – 통합 유전체 정보 시스템

3.2 유전적 다양성과 변이

3.2.1 핵형과 배수성

핵형(Karyotype)은 개체가 가지고 있는 염색체의 집합을 의미한다. 염색체의 수, 크기, 모양 등을 포함한다.

Figure 2.6 핵형 분석 – 인간 염색체 핵형 분석 결과

배수성(Ploidy)은 세포 내 염색체 세트의 수를 나타낸다:

Figure 2.7 배수성 – 염색체 세트 수에 따른 분류 – 단배체(Haploid): 1개의 염색체 세트 – 이배체(Diploid): 2개의 염색체 세트 – 삼배체(Triploid): 3개의 염색체 세트 – 사배체(Tetraploid): 4개의 염색체 세트

예를 들어, 인간 세포는 이배체로, 각 염색체의 2개 복사본(부모로부터 각각 하나씩)을 가지고 있다.

3.2.2 유전형과 표현형

유전형(Genotype)과 표현형(Phenotype)은 유전학의 기본 개념이다:

- 표현형: 관찰 가능한 특성의 집합(예: 키, 몸무게, 색상 등)

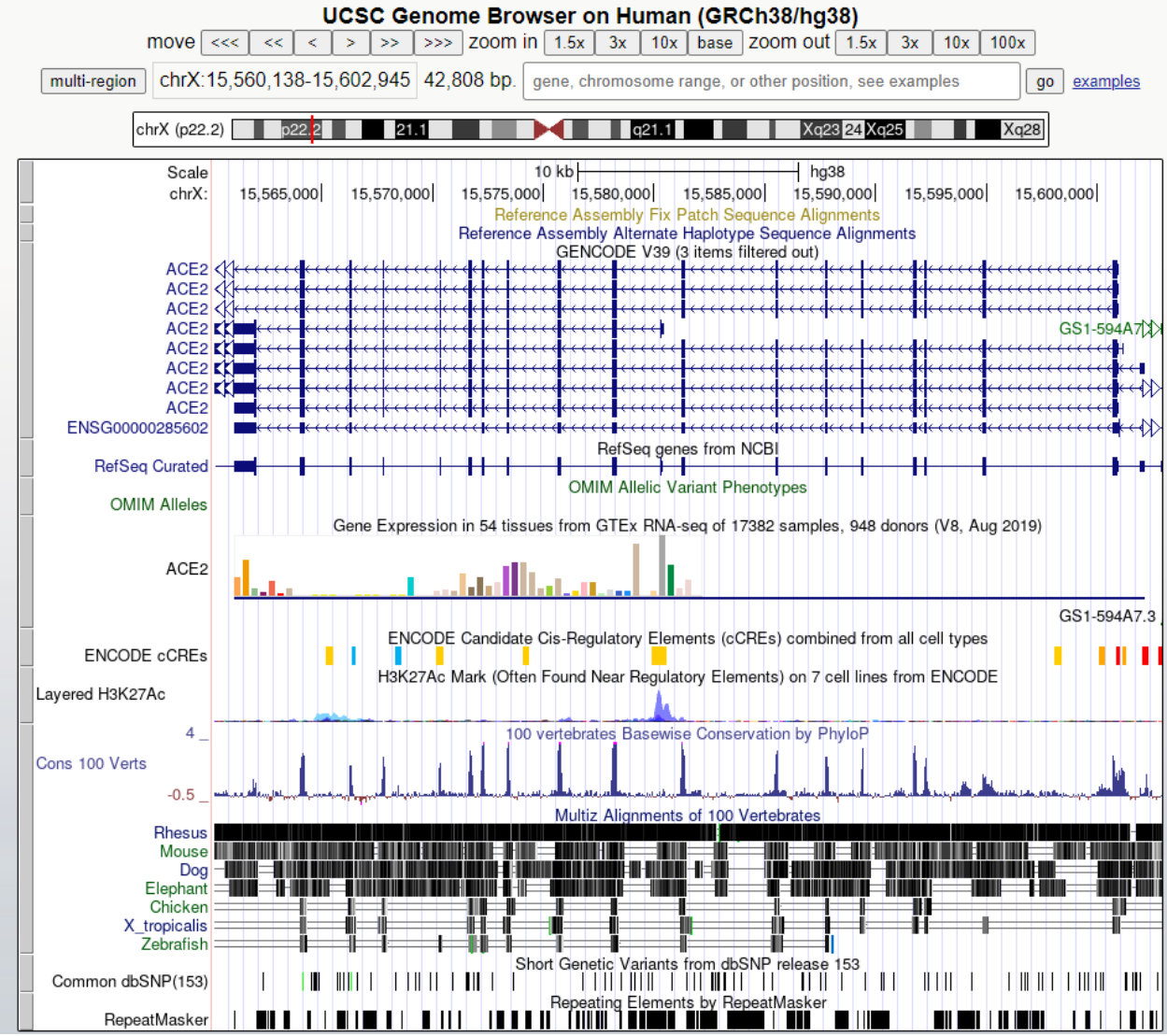


Figure 30: UCSC Genome Browser

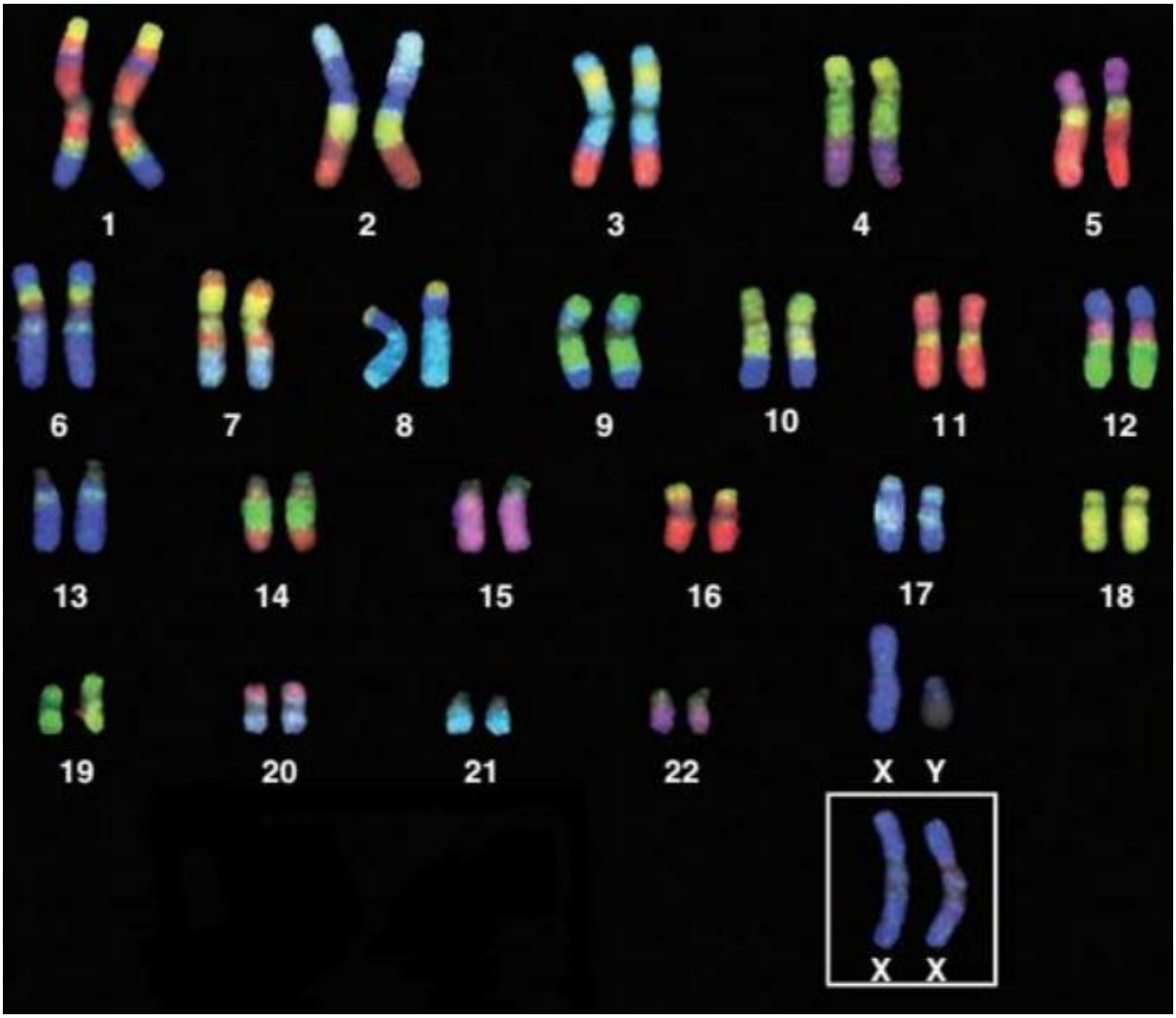


Figure 31: 핵형 분석

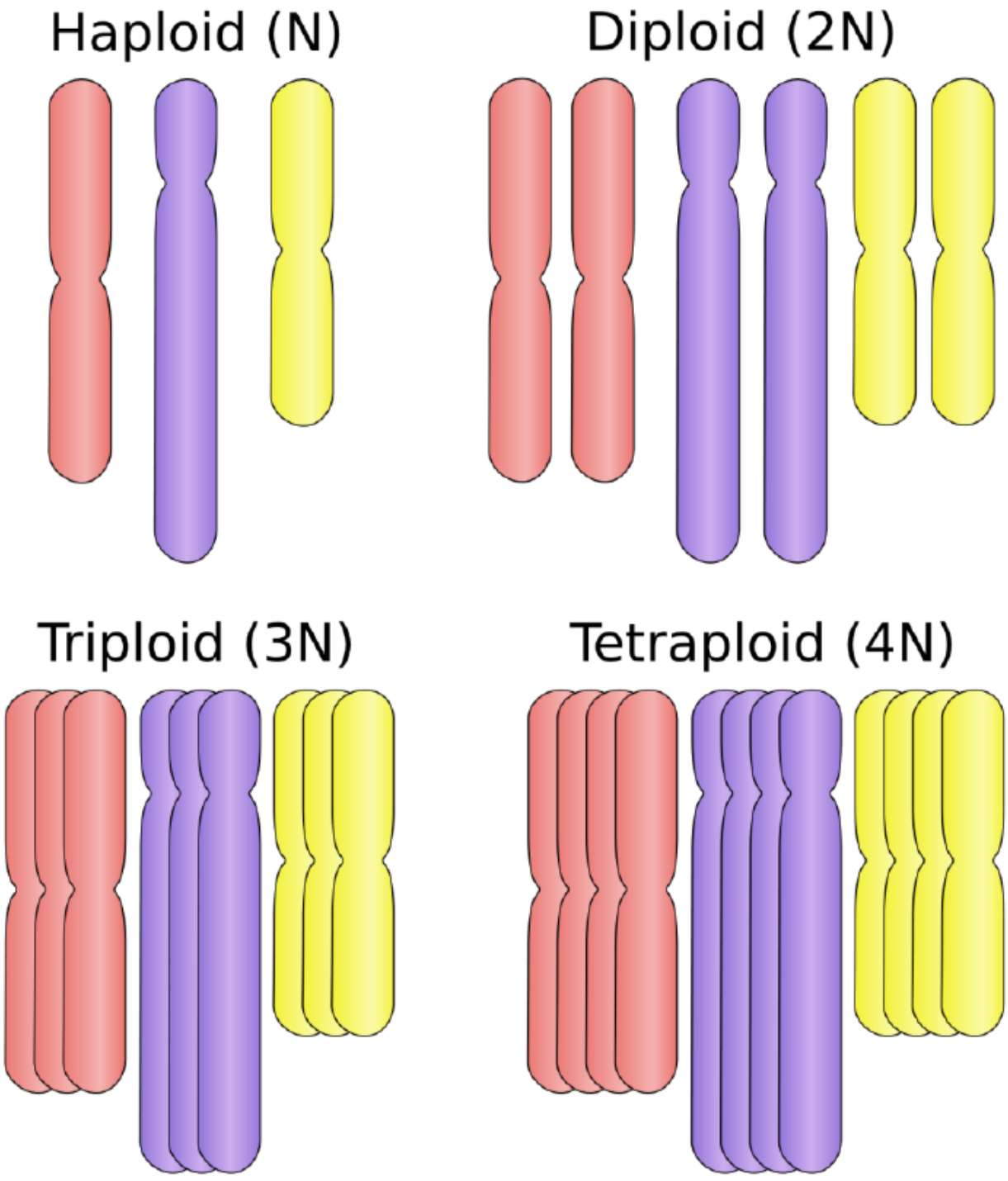


Figure 32: 배수성

- 유전형: 표현형에 기여하는 유전자 그룹
  - 대립유전자(Allele): 특정 유전자의 대안적 형태

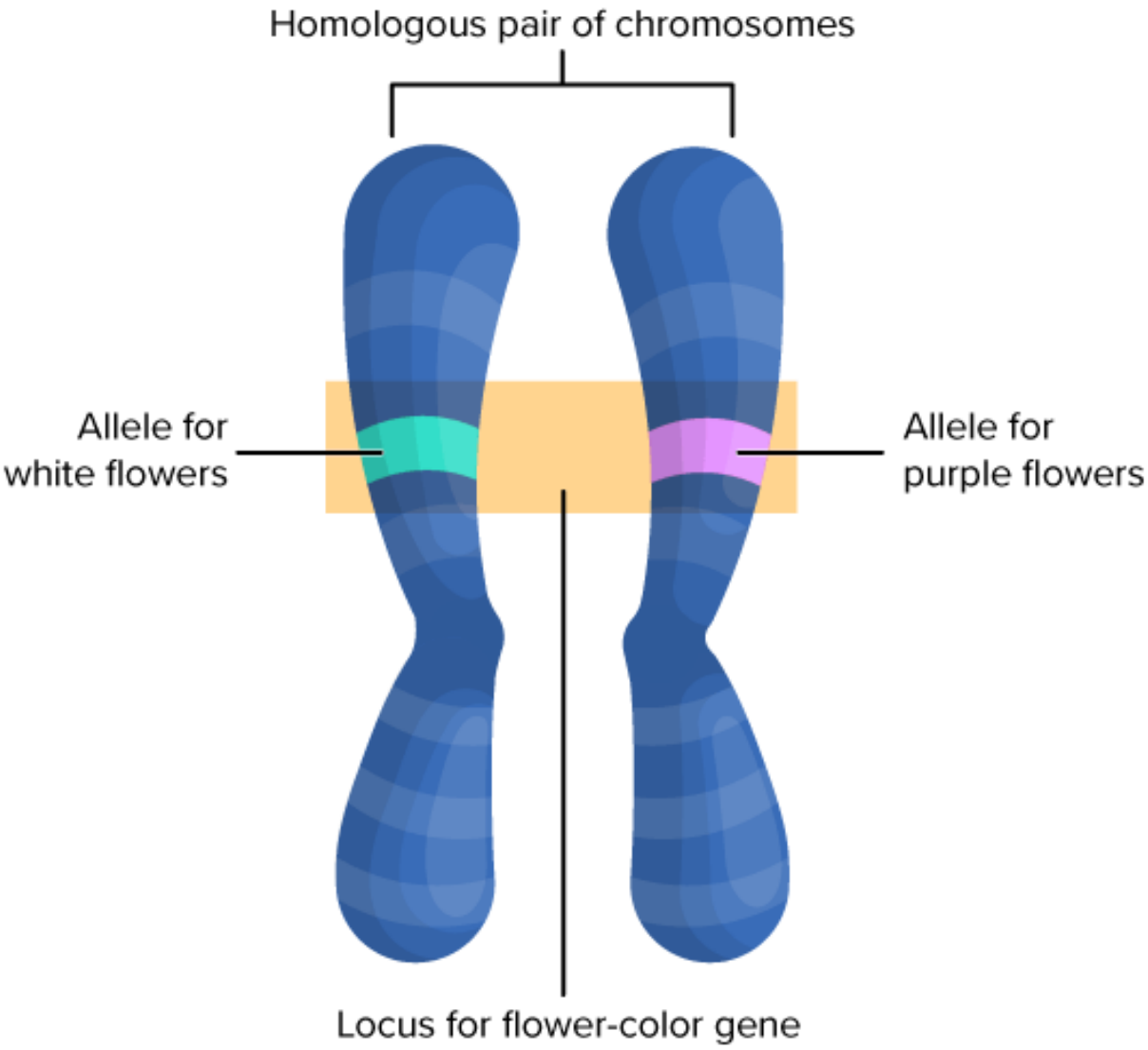


Figure 33: 대립유전자

Figure 2.8 대립유전자 – 같은 유전자의 다른 형태

만약 우리가 유전형으로부터 표현형을 완벽하게 예측해낼 수 있다면, 이는 곧 데이터로부터 생명이라는 현상을 완전히 이해하게 되는 것으로 이해할 수 있다. 이는 사실상 생명정보학이 달성하고자 하는 궁극적인 목표라고도 볼 수 있을 것이다.

3.2.3 유전적 변이

유전적 변이(Variation)는 개체 간의 유전적 차이를 의미한다. 주요 유전적 변이 유형은 다음과 같다:

- 단일 염기 변이(Single Nucleotide Variations, SNVs): 단일 염기의 변화

Figure 2.9 단일염기변이(SNV) – 단일 염기의 치환

Figure 2.10 이형접합자 SNV – 두 개의 다른 대립유전자를 가진 단일염기변이

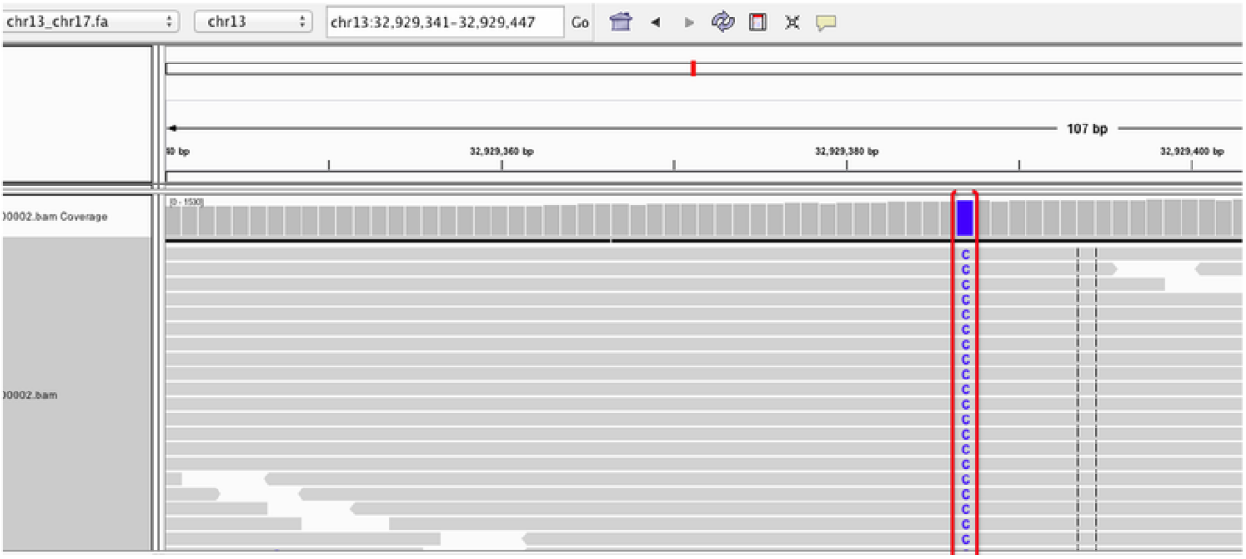


Figure 34: 단일염기변이

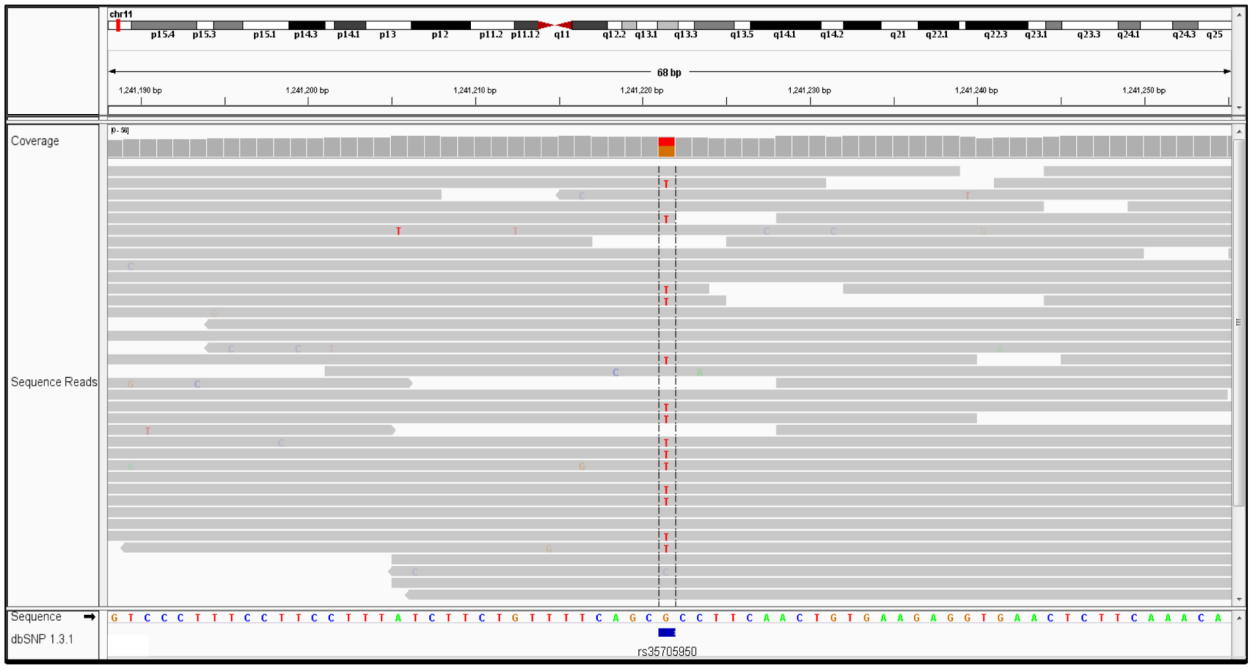


Figure 35: 이형접합자 SNV

- 삽입 및 결실(Insertions and Deletions, INDELs): DNA 서열의 삽입 또는 결실

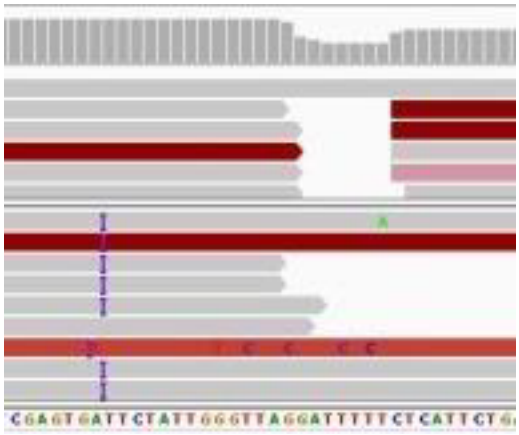


Figure 36: 삽입결실변이

Figure 2.11 삽입결실변이(INDEL) – 작은 서열의 삽입 또는 결실

Figure 2.12 결실변이 – DNA 서열의 결실

- 구조적 변이(Structural Variations, SVs):
  - 복제수 변이(Copy Number Variations, CNVs): DNA 서열의 복사 수 변화
  - 대규모 삽입 및 결실

Figure 2.13 대규모 중복 – 큰 DNA 서열의 복사수 증가

- 전위(Translocations), 역위(Inversions) 등

Figure 2.14 염색체 전좌 – 염색체 간 서열 이동

Figure 2.15 염색체 역위 – 염색체 내 서열 방향 뒤바뀜

유전적 변이는 발생 위치에 따라 다음과 같이 분류할 수 있다:

- 생식계 변이(Germline variants): 생식세포에 존재하는 변이로, 전체 유기체에서 발견됨
- 체세포 변이(Somatic variants): 특정 부위에만 영향을 미치는 변이

Figure 2.16 체세포변이와 생식계변이 – 변이 발생 위치에 따른 분류

단일 염기 다형성(Single Nucleotide Polymorphism, SNPs)은 인구 집단에서 흔하게 발견되는 생식계 SNV를 의미한다. 이와 관련된 중요한 개념으로 소수 대립유전자 빈도(Minor Allele Frequency, MAF)가 있으며, 이는 인구 집단 내에서 소수 변이 대립유전자의 빈도를 나타낸다.

Figure 2.17 단일염기다형성(SNP) – 집단에서 흔한 유전적 변이

3.2.3.1 변이 표기법

유전적 변이는 표준화된 표기법을 사용하여 기술된다:

DNA 수준의 변이 표기: – 단일 염기 치환: C>T (시토신이 티민으로 변화) – 삽입: A>ATCG (아데닌 다음에 TCG 서열 삽입) – 결실: ATCG>A (TCG 서열이 결실되어 A만 남음)

단백질 수준의 변이 표기: DNA 변이가 단백질 코딩 영역에 발생하면 아미노산 변화를 일으킬 수 있으며, 이는 다음과 같이 표기된다:

- p.Arg47His (47번째 위치의 아르지닌이 히스티딘으로 변화)
- p.Trp24* (24번째 위치의 트립토판이 정지코돈으로 변화, *는 정지코돈)
- p.Met1? (1번째 위치의 메티오닌에 변화가 있으나 정확한 결과 불명)

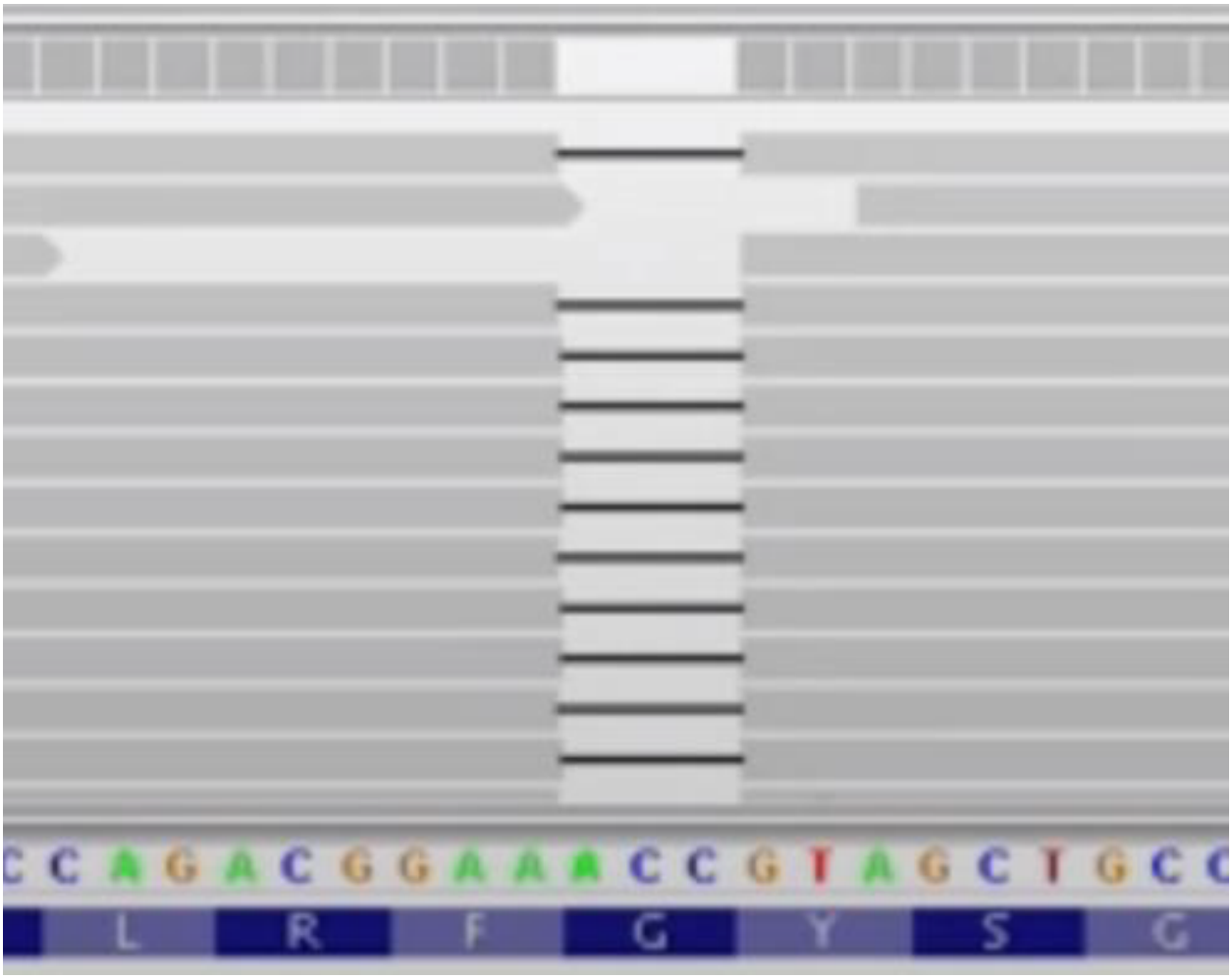


Figure 37: 결실변이

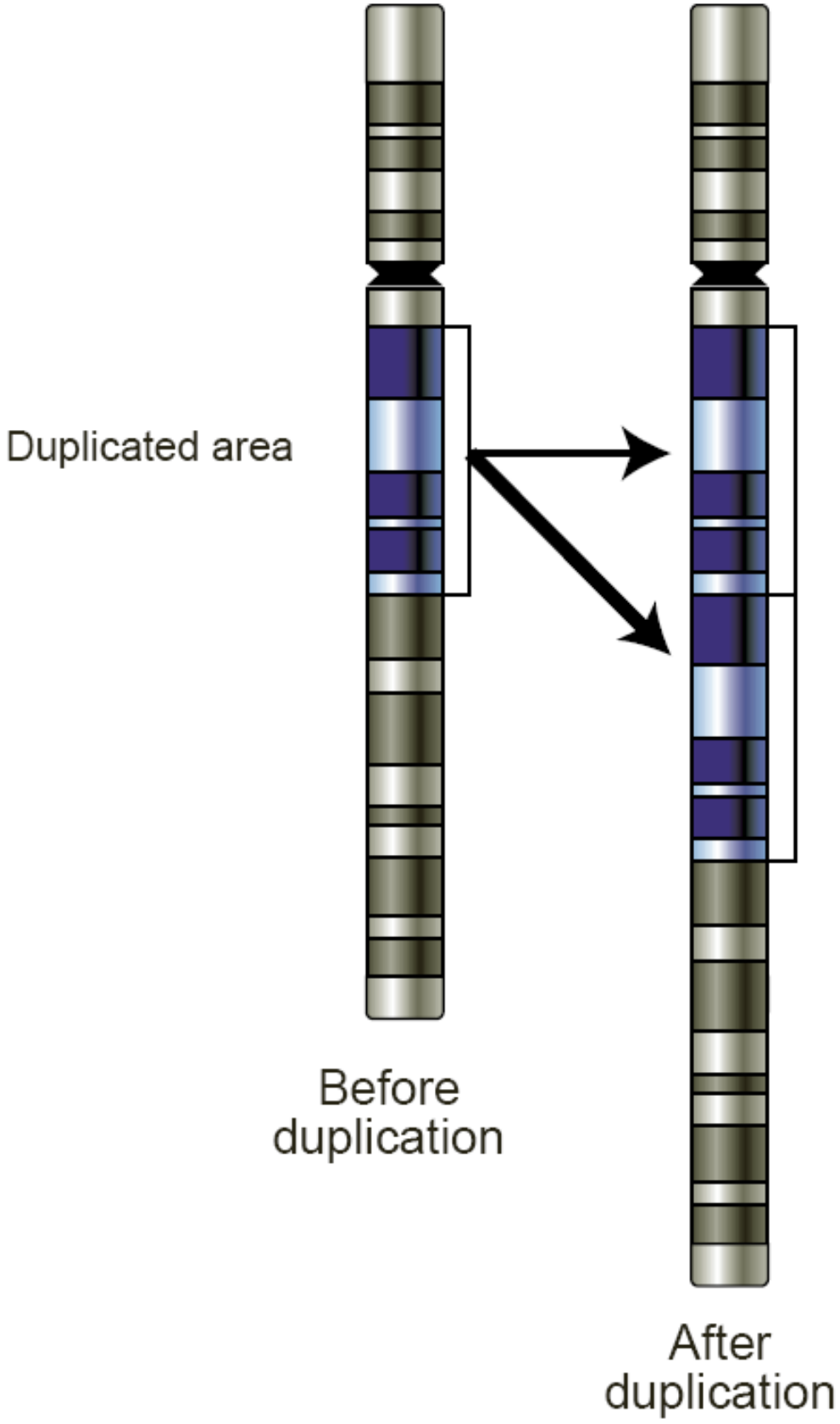


Figure 38: 대규모 중복

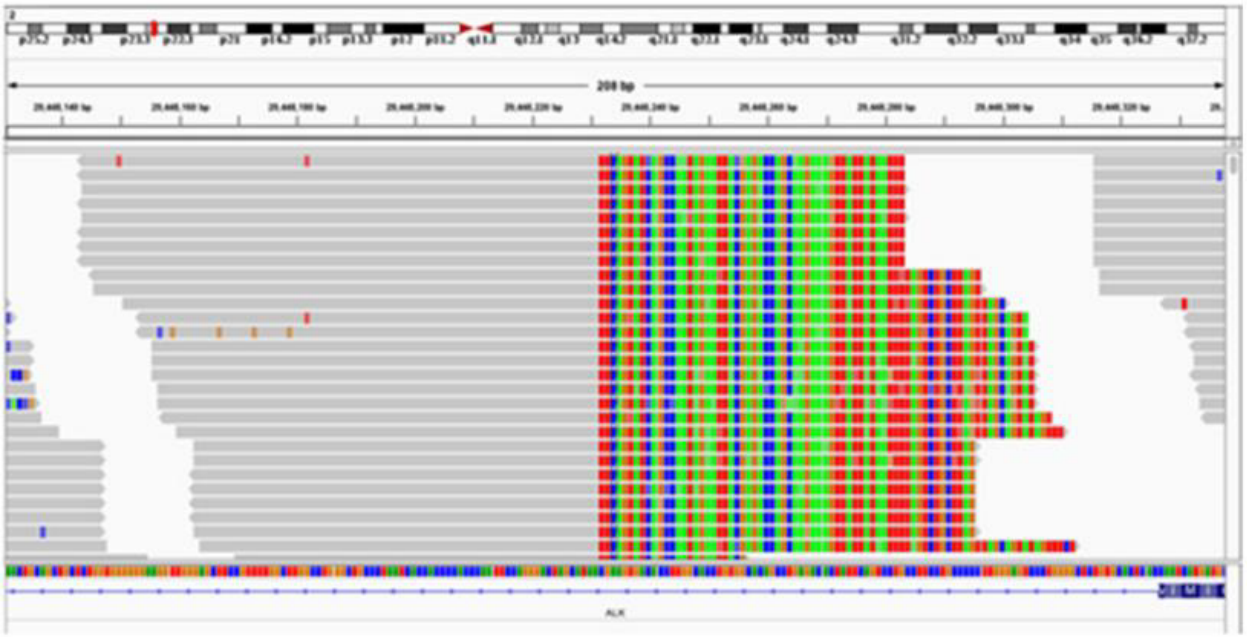


Figure 39: 염색체 전좌

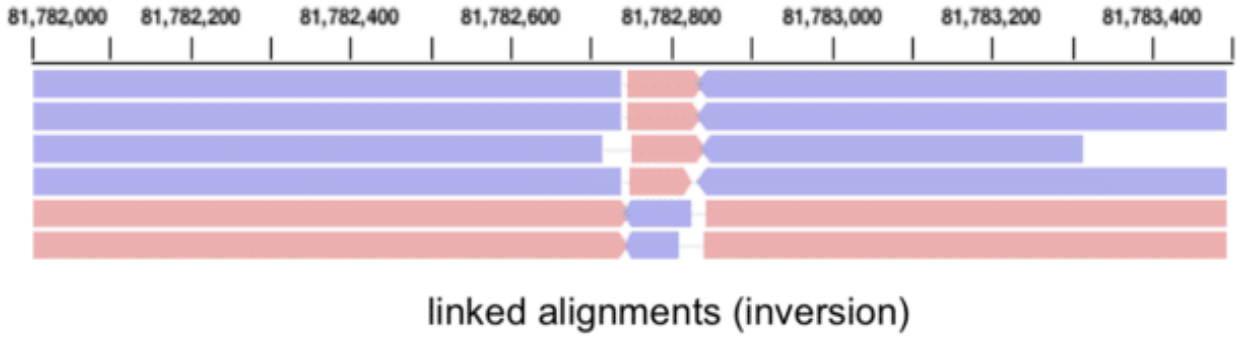


Figure 40: 염색체 역위

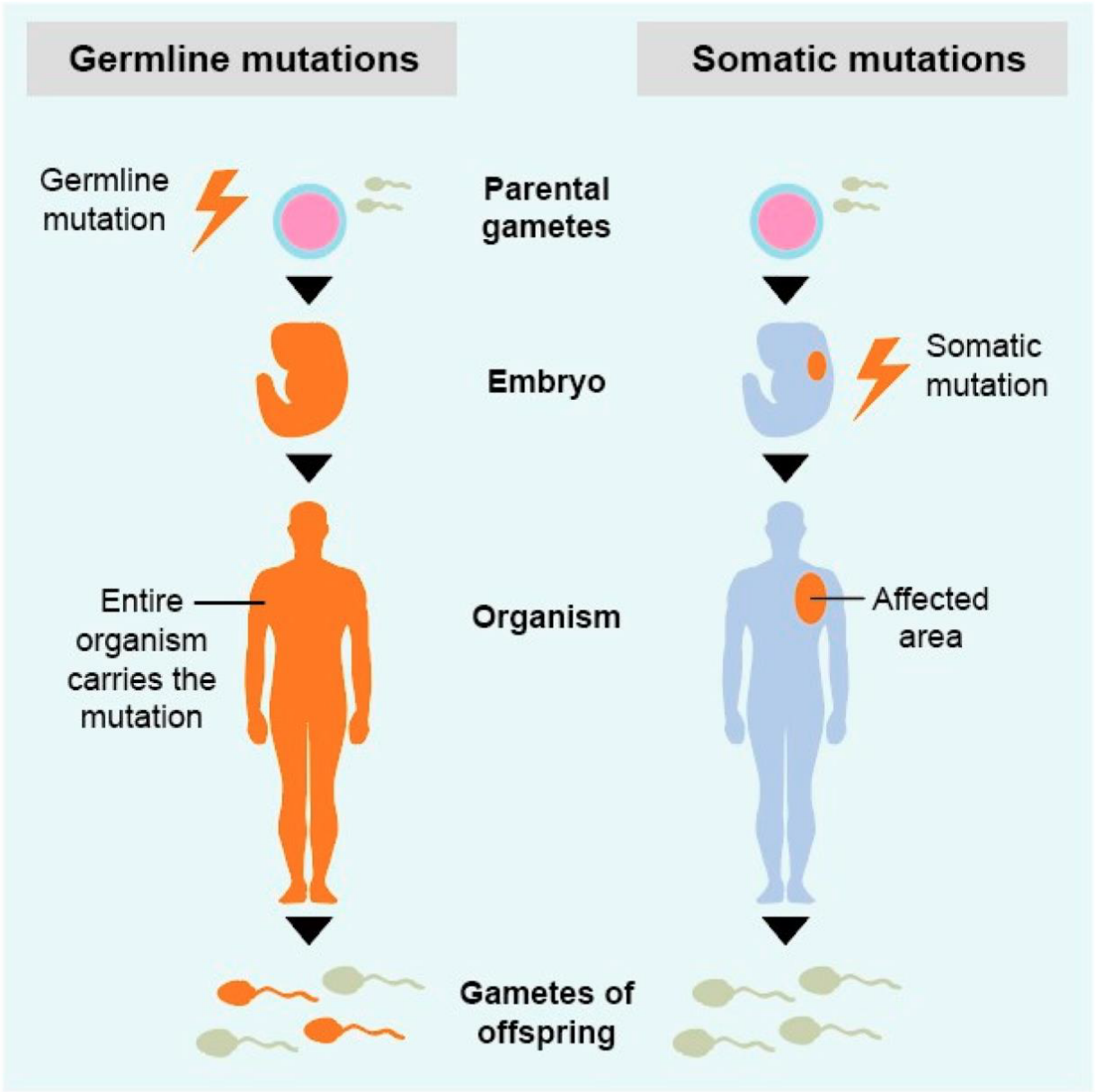


Figure 41: 체세포변이와 생식계변이

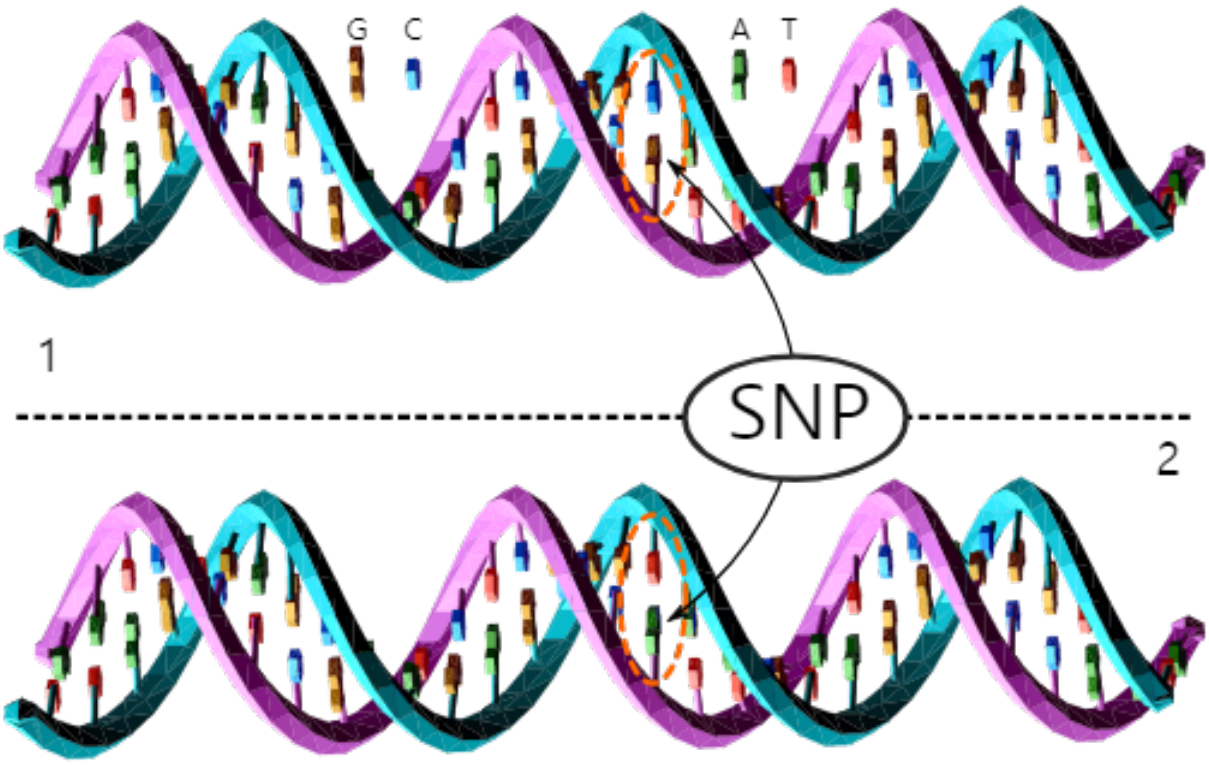


Figure 42: 단일염기다형성

변이의 기능적 영향에 따른 분류: - 동의 변이(Synonymous variant): 아미노산 변화 없음 - 비동의 변이(Missense variant): 다른 아미노산으로 변화 - 무의미 변이(Nonsense variant): 정지코돈으로 변화 - 프레임시프트 변이(Frameshift variant): 삽입/결실로 인한 읽기틀 변화

코돈

코돈(Codon)은 하나의 아미노산을 지정하는 3개의 연속된 염기서열을 말한다. 64개의 가능한 코돈 중 61개는 아미노산을 암호화하고, 3개는 번역 종료를 나타내는 정지코돈이다. DNA 변이는 코돈의 변화를 유발할 수 있다.

Figure 2.18 코돈 - 코돈의 개념

Figure 2.19 유전 코돈표 - 유전 암호와 아미노산 대응

3.3 인구집단 유전학 연구와 데이터베이스

3.3.1 1000 지놈 프로젝트(1000 Genome Project)

1000 지놈 프로젝트는 2008년 1월 미국, 영국, 중국의 협력으로 시작되었다. 초기에는 14개 인구집단에서 약 1,000명의 개인 유전체를 시퀀싱했으며(https://www.nature.com/articles/nature11632), 현재는 25개 인구집단에서 약 13,000명의 유전체가 시퀀싱되었다(https://www.internationalgenome.org/data-portal/sample). 이 프로젝트를 통해 변이 호출 포맷(Variant Call Format, VCF)이 탄생했다.

Figure 2.20 1000 Genomes Project - 인구집단 유전체 다양성 연구

Figure 2.21 VCF 형식 - 변이 정보 표준 포맷

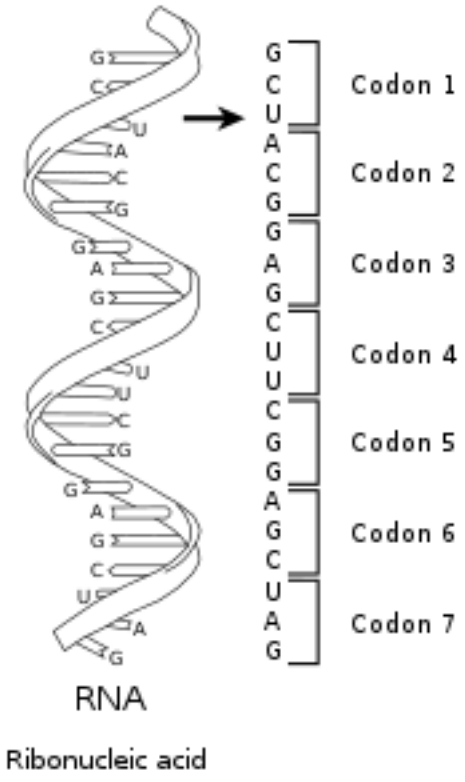


Figure 43: 코돈

Standard genetic code[1][10]

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine (np)	UCU	(Ser/S) Serine (p)	UAU	(Tyr/Y) Tyrosine (p)	UGU	(Cys/C) Cysteine (p)	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine (np)	UCA		UAA	Stop (*Ochre*) *[note 2]	UGA	Stop (*Opal*) *[note 2]	A
	UUG ⇒		UCG		UAG	Stop (*Amber*) *[note 2]	UGG	(Trp/W) Tryptophan (np)	G
C	CUU		CCU	(Pro/P) Proline (np)	CAU	(His/H) Histidine (b)	CGU	(Arg/R) Arginine (b)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine (p)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine (np)	ACU	(Thr/T) Threonine (p)	AAU	(Asn/N) Asparagine (p)	AGU	(Ser/S) Serine (p)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine (b)	AGA	(Arg/R) Arginine (b)	A
	AUG ⇒	(Met/M) Methionine (np)	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine (np)	GCU	(Ala/A) Alanine (np)	GAU	(Asp/D) Aspartic acid (a)	GGU	(Gly/G) Glycine (np)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid (a)	GGA		A
	GUG ⇒		GCG		GAG		GGG		G

Figure 44: 유전 코돈표

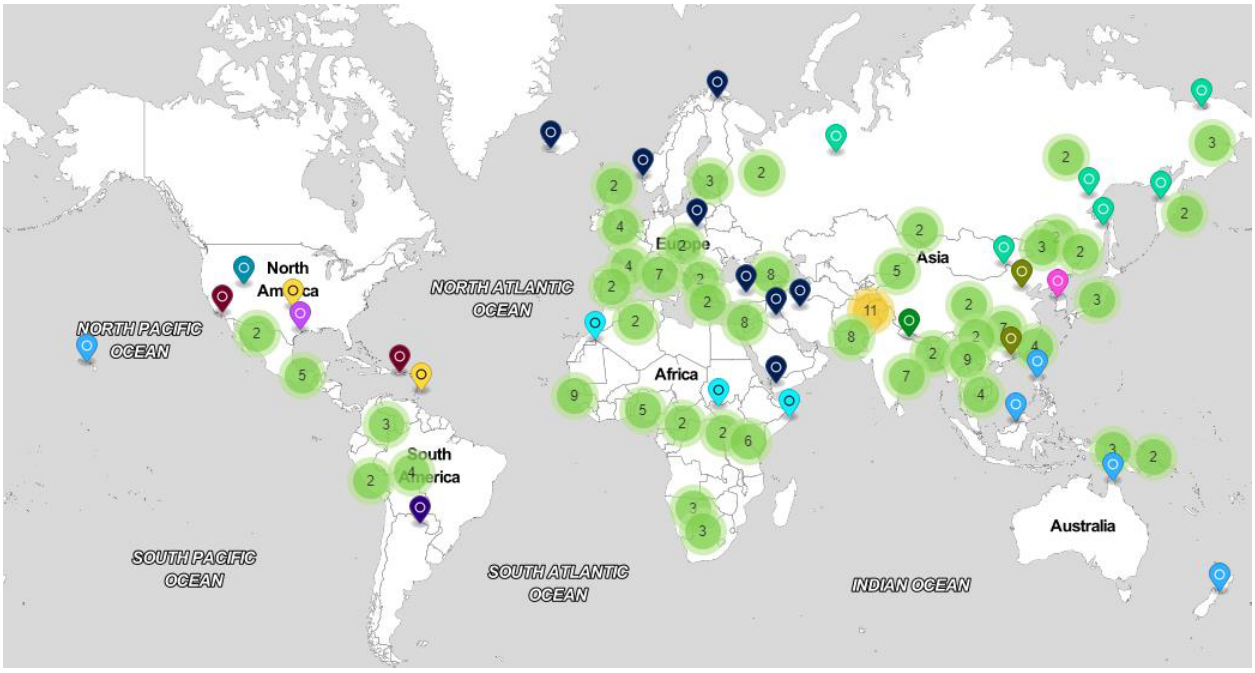


Figure 45: 1000 Genomes Project

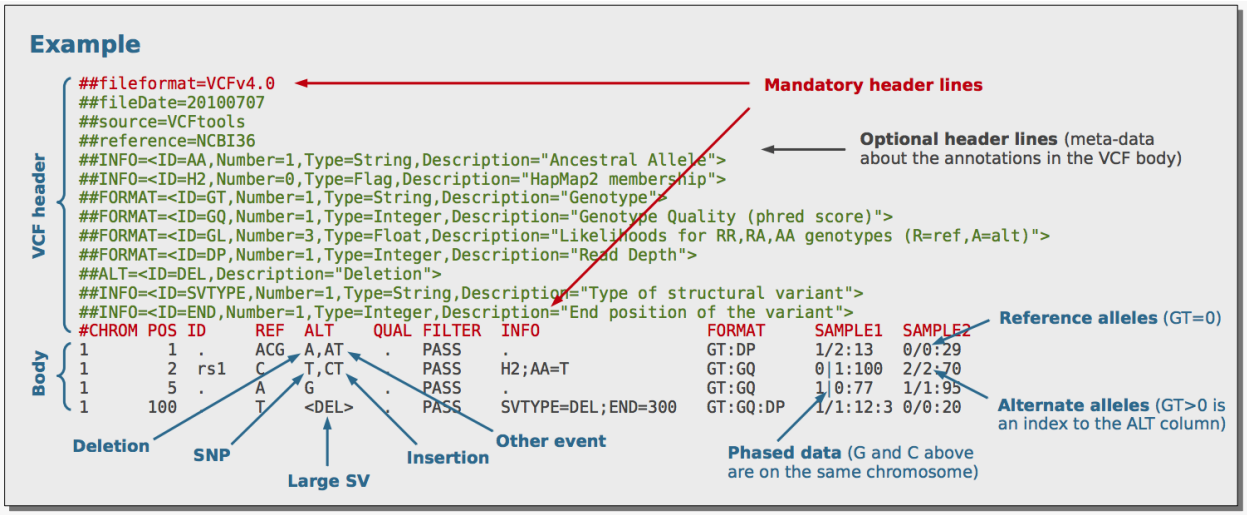


Figure 46: VCF 형식

3.3.2 유전체 통합 데이터베이스(gnomAD)

유전체 통합 데이터베이스(Genome Aggregation Database, gnomAD)는 이전에 “엑솜 통합 컨소시엄(Exome Aggregation Consortium, ExAC)”으로 알려졌으며, 엑손(단백질 코딩 영역)의 변이 데이터베이스이다. 현재 76,156명의 개인 유전체로부터 얻은 변이 데이터를 제공하고 있다.

Figure 47: gnomAD UMAP

Figure 2.22 gnomAD UMAP – 인구집단별 유전적 구조 시각화

3.3.3 dbSNP

dbSNP는 SNP를 포함한 소규모 변이의 무료 공개 데이터베이스로, 이미 1000 지놈 프로젝트, gnomAD 등의 데이터를 포함하고 있으며, 대립유전자 빈도도 제공한다. 이 데이터베이스는 미국 국립생물정보센터(NCBI)와 국립 인간 유전체 연구소(National Human Genome Research Institute, NHGRI)에서 관리한다.

3.3.4 유전적 질환 관련 데이터베이스

유전적 질환과 관련된 중요한 데이터베이스는 다음과 같다: - Online Mendelian Inheritance in Man (OMIM): 인간의 유전자, 유전적 질환, 그리고 이들 간의 관계에 대한 종합적인 온라인 데이터베이스

Figure 48: OMIM 데이터베이스

Figure 2.23 OMIM 데이터베이스

- ClinVar: NCBI에서 관리하는 유전변이와 질병 관계 데이터베이스

Figure 49: ClinVar 데이터베이스

Figure 2.24 ClinVar 데이터베이스

3.3.5 희귀 변이 및 국가별 연구

희귀 변이에 대한 연구로는 영국의 UK10K 프로젝트가 있다.

Figure 2.25 UK10K 프로젝트 – 영국 인구집단 희귀변이 연구

한국에서는 울산 10K 프로젝트가 진행되고 있다.

Figure 2.26 울산 10K 프로젝트 – 한국인 유전체 연구

3.4 생물학적 경로와 단백질 구조

3.4.1 생물학적 경로

생물학적 경로(Pathway)는 표현형 변화를 일으키는 유전자 간의 상호작용 그룹을 의미한다. 중요한 생물학적 경로로는 다음과 같은 것들이 있다: - 세포사멸(apoptosis) 경로 - DNA 복구 경로 - 암 관련 경로(예: Ras-Raf-MEK-ERK 경로)

Figure 50: UK10K 프로젝트

Figure 51: 울산 10K 프로젝트

3.4.2 유전자 집합 농축 분석

유전자 집합 농축 분석(Gene Set Enrichment Analysis, GSEA)은 쿼리 유전자를 기반으로 농축된 유전자 집합을 찾는 분석 방법이다. 관련 데이터베이스와 도구는 다음과 같다: - GSEA: https://www.gseamsigdb.org/gsea/index.jsp

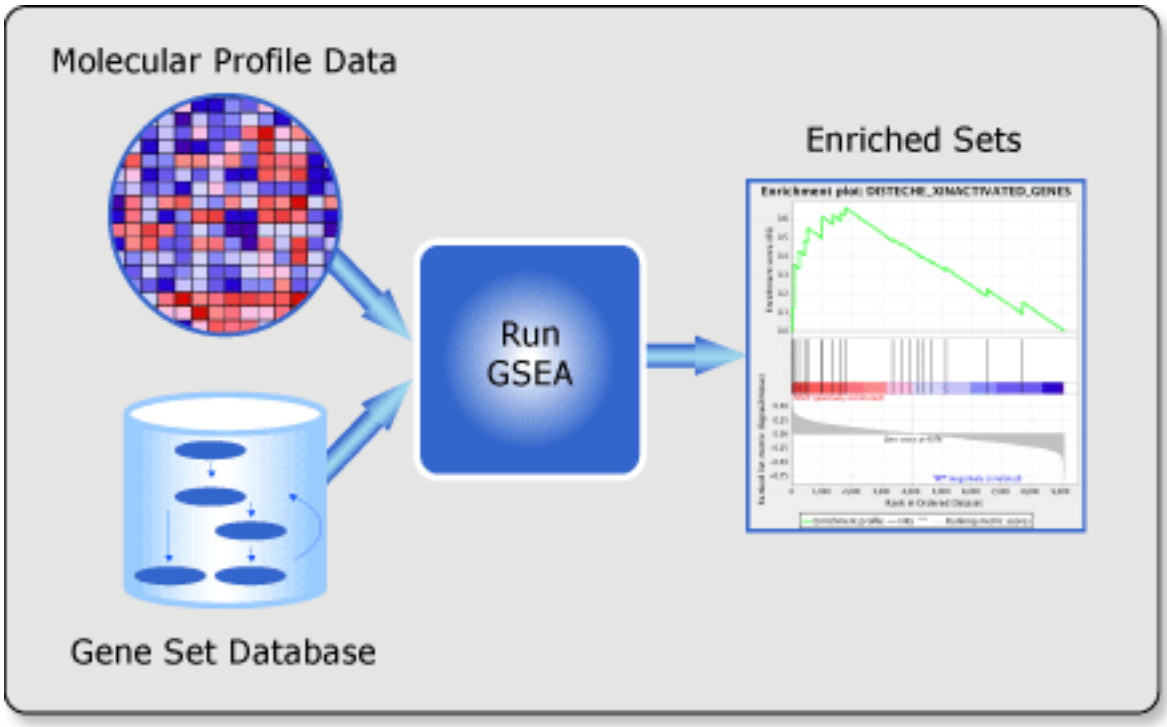


Figure 52: GSEA 분석

Figure 2.27 GSEA 분석 – 유전자 집합 농축 분석 결과

3.4.3 단백질 구조 데이터베이스

단백질 데이터 은행(Protein Data Bank, PDB)은 세계 최대의 단백질 구조 데이터베이스로, 현재 188,085개의 단백질 구조 정보를 제공하고 있다. PDB 웹사이트는 다음과 같다: - PDB: https://www.rcsb.org/

Figure 2.28 PDB 데이터베이스 – 단백질 구조 정보 저장소

3.5 생명정보학 종합 리소스: NCBI

미국 국립생물정보센터(National Center for Biotechnology Information, NCBI)는 생명과학/생명공학과 관련된 모든 정보를 제공하는 종합적인 리소스이다.

Figure 2.29 NCBI 웹사이트 – 종합 생명정보학 리소스

NCBI 웹사이트는 다음과 같다: - NCBI: https://www.ncbi.nlm.nih.gov/

NCBI에서 제공하는 중요한 서비스 중 하나로 PubMed가 있다. PubMed는 생명과학 및 의학 분야의 논문 데이터베이스로, 다음 웹사이트에서 접근할 수 있다: - NCBI PubMed: https://pubmed.ncbi.nlm.nih.gov/

Figure 2.30 PubMed – 생명과학 및 의학 논문 데이터베이스

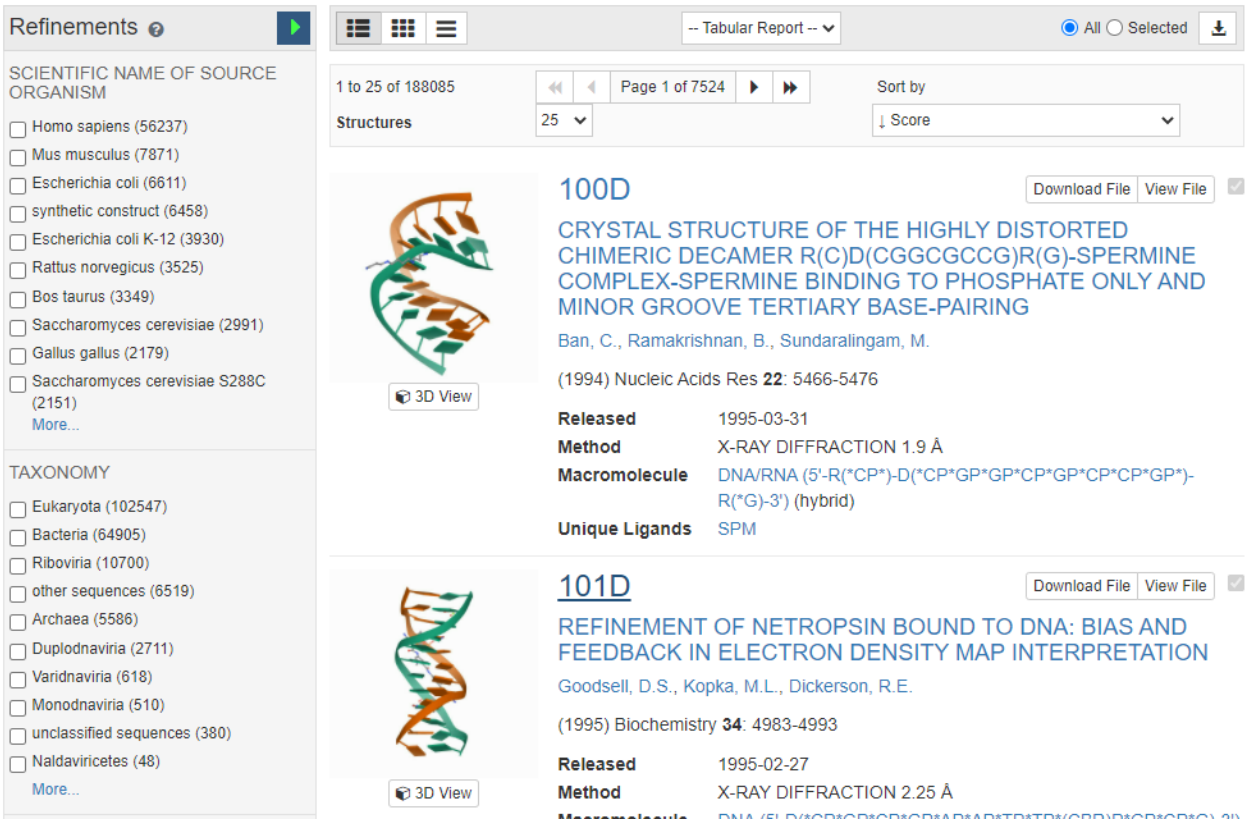


Figure 53: PDB 데이터베이스

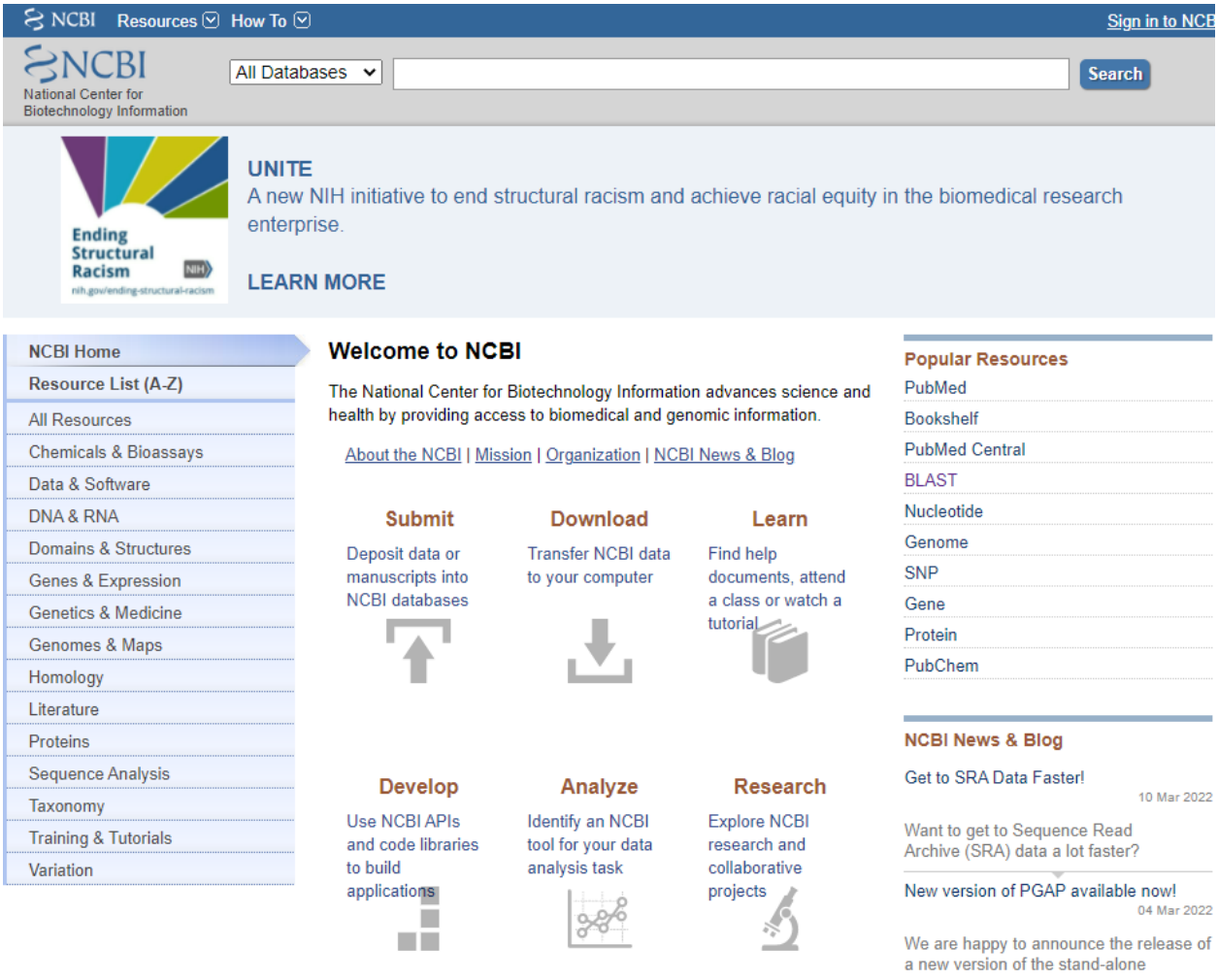


Figure 54: NCBI 웹사이트

Figure 55: PubMed

3.6 생명정보학 데이터베이스의 활용

생명정보학 연구에서는 여러 데이터베이스를 통합적으로 활용하는 것이 중요하다. 유전체 데이터, 변이 정보, 표현형 정보, 단백질 구조, 생물학적 경로 등 다양한 데이터를 종합적으로 분석함으로써 생명 현상의 복잡한 메커니즘을 이해할 수 있으며, 이는 질환 연구를 할 때에 있어 특히 유용하다.

예를 들어, 특정 질환과 관련된 연구를 수행할 때, 다음과 같은 단계로 여러 데이터베이스를 활용할 수 있다: 1. NCBI나 Ensembl에서 관심 유전자의 기본 정보 확인 2. dbSNP이나 gnomAD에서 해당 유전자의 변이 정보 조사 3. OMIM이나 ClinVar에서 관련 질환 정보 확인 4. KEGG나 Reactome에서 해당 유전자가 관여하는 생물학적 경로 분석 5. PDB에서 관련 단백질의 구조 정보 확인 6. PubMed에서 관련 연구 논문 검색

4 차세대 시퀀싱

4.1 차세대 시퀀싱의 개요

차세대 시퀀싱(Next Generation Sequencing, NGS)은 기존의 Sanger 시퀀싱 방법을 대체하는 고속 대용량 DNA 시퀀싱 기술이다. NGS 기술은 대량의 DNA 또는 RNA 단편을 병렬적으로 처리하여 단시간에 대용량의 시퀀스 정보를 생산할 수 있다. 이 기술의 등장으로 전체 유전체 시퀀싱(Whole Genome Sequencing, WGS)이 효율적으로 가능해졌으며, 이는 유전체학, 의학, 농업 및 생명과학 전반에 혁명적인 변화를 가져왔다.

4.2 주요 NGS 플랫폼의 이해

4.2.1 Illumina 시퀀싱 기술

Illumina는 현재 NGS 시장에서 가장 널리 사용되는 플랫폼으로, 높은 정확도와 광범위한 응용 프로그램을 제공한다.

Illumina 시퀀싱의 첫 단계는 Bridge PCR이라는 과정이다. 이 과정에서는:

1. DNA 단편이 어댑터가 부착된 상태로 플로우 셀(flow cell) 표면에 결합한다.
2. 각 DNA 단편이 구부러져 근처의 상보적인 어댑터와 결합한다(bridge 형성).
3. 중합효소가 상보적인 가닥을 합성한다.
4. 두 가닥이 변성되어 분리된다.
5. 이 과정이 반복되어 각 원본 DNA 단편으로부터 클러스터(cluster)가 형성된다.

Bridge PCR을 통해 동일한 DNA 단편의 수천~수백만 개의 복사본이 생성되어 시퀀싱 과정에서 신호를 증폭시키는 역할을 한다.

Illumina의 핵심 기술은 합성에 의한 시퀀싱(Sequencing by Synthesis, SBS)이다. 이 방법의 특징은:

1. 형광 표지된 가역성 종결자(reversible terminator) 뉴클레오티드를 사용한다.
2. 각 뉴클레오티드(A, T, G, C)는 서로 다른 색의 형광 물질로 표지되어 있다.
3. DNA 중합효소가 상보적인 뉴클레오티드를 하나씩 추가한다.
4. 각 주기마다 형광 신호를 이미징하여 추가된 뉴클레오티드를 식별한다.
5. 형광 표지와 종결자가 제거되어 다음 주기가 진행될 수 있도록 한다.

이 과정을 통해 클러스터 내의 모든 DNA 단편이 동시에 시퀀싱되며, 각 주기마다 한 개의 염기가 결정된다.

Figure 3.1 Illumina 시퀀싱 원리 – Bridge PCR 및 SBS 과정

Figure 3.2 Illumina 플로우 셀 – DNA 클러스터 생성을 위한 유리 슬라이드 표면

Figure 3.3 Illumina 플로우 셀 세부사항 – 레인과 클러스터 구조

Illumina 기술의 특징 중 하나는 짧은 리드(Short Read)를 생성한다는 점이다. 일반적으로 리드 길이는 75bp에서 300bp까지 다양하며, 높은 정확도(99.9% 이상)를 가진다. 그러나 Short Read는 반복 서열이나 구조적 변이를 해석하는 데 제한이 있다. 이러한 Short Read는 참조 유전체(reference genome)가 있는 경우 변이 분석에 적합하지만, 새로운 유전체를 조립하는 데는 한계가 있다.

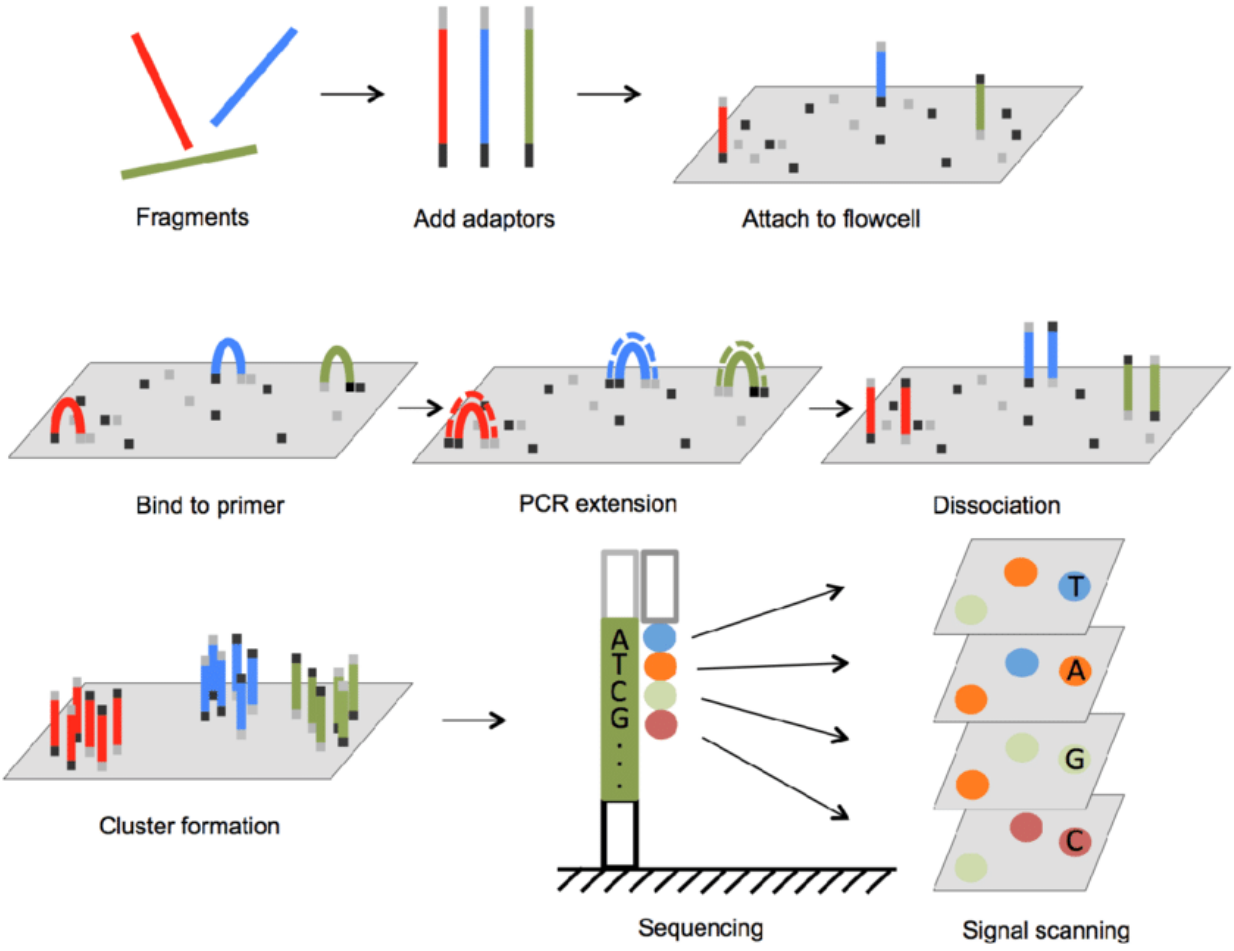


Figure 56: Illumina 시퀀싱 원리

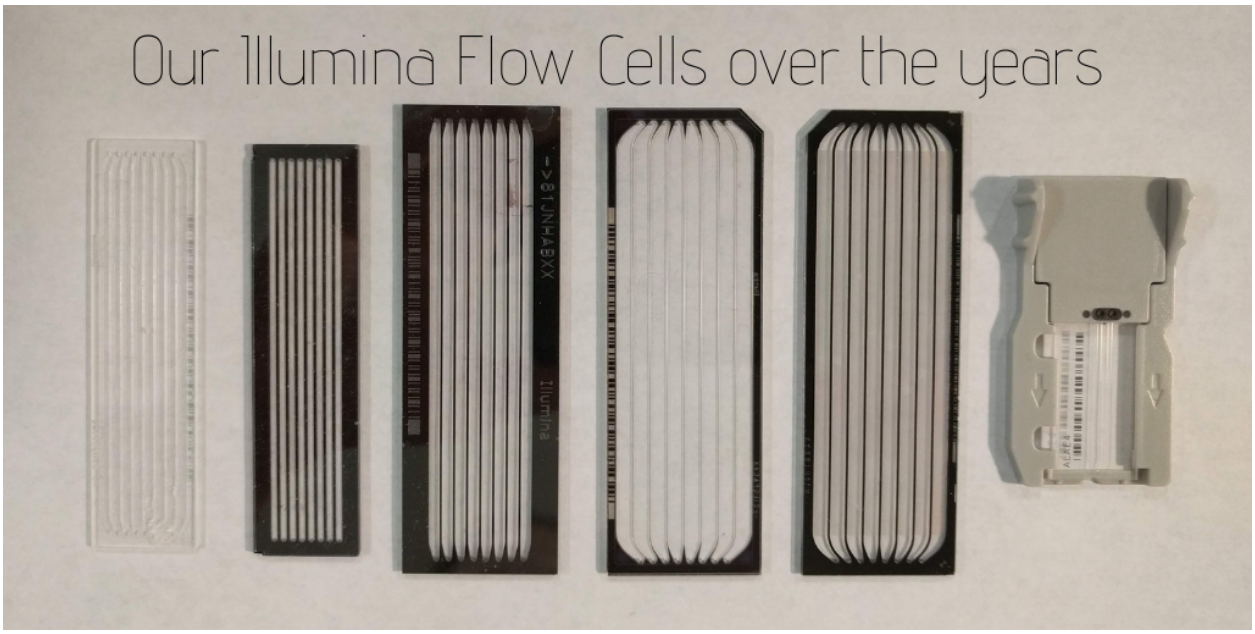


Figure 57: Illumina 플로우 셀

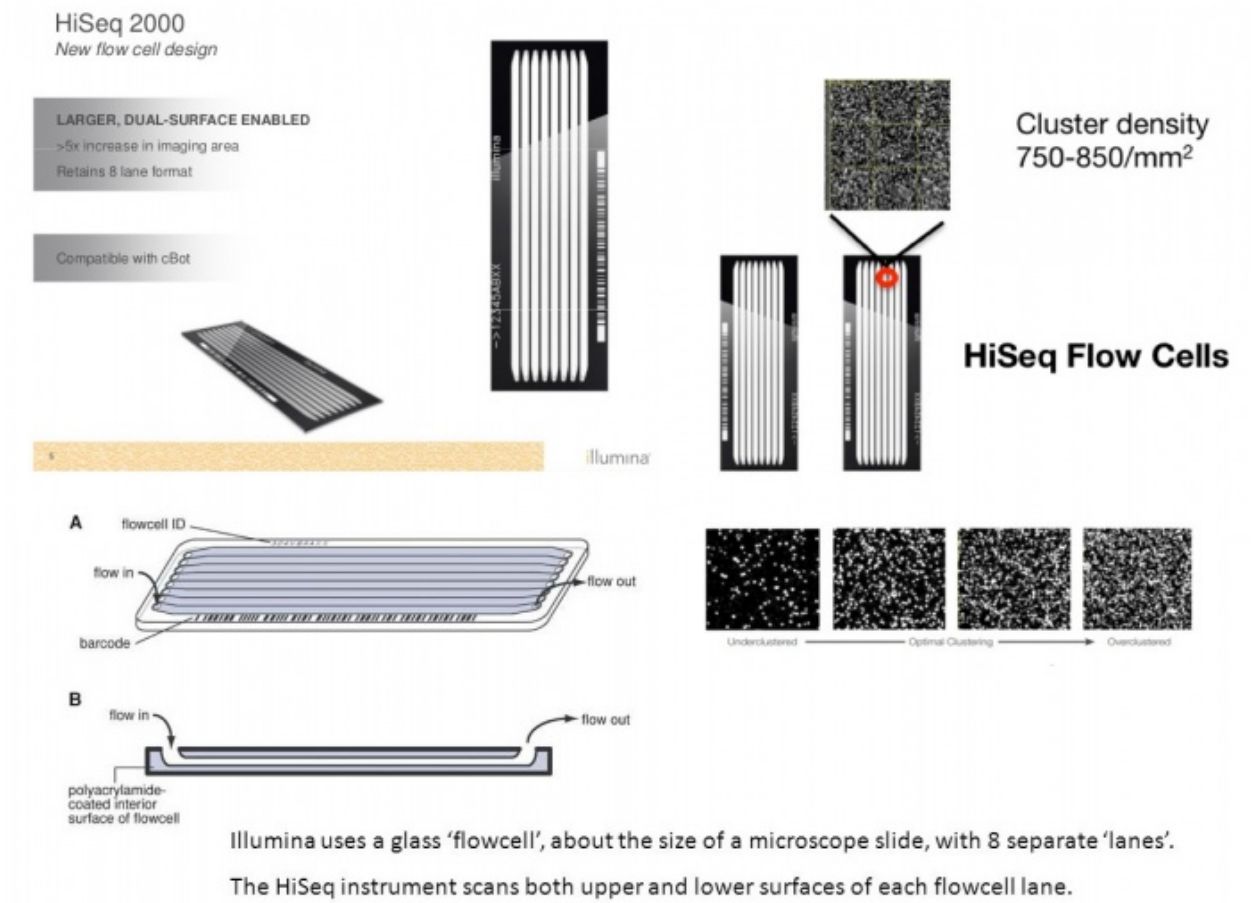


Figure 58: Illumina 플로우 셀 세부사항

4.2.2 Long Read 시퀀싱 기술: Nanopore와 PacBio

4.2.2.1 Nanopore 시퀀싱 Oxford Nanopore Technologies에서 개발한 Nanopore 시퀀싱은 전혀 다른 접근 방식을 사용한다. 나노포어 시퀀싱의 원리는 다음과 같다:

1. 단백질 나노포어(nanopore)가 전기적으로 저항성인 막에 삽입된다.
2. 전류가 나노포어를 통해 흐르게 된다.
3. DNA 또는 RNA가 나노포어를 통과할 때, 각 뉴클레오티드는 특정한 전류 차단 패턴을 생성한다.
4. 이 전류 변화를 모니터링하고 해석하여 염기 서열을 결정한다.

이 방법은 DNA 분자를 직접 읽기 때문에 PCR 증폭이나 화학적 표지가 필요하지 않다.

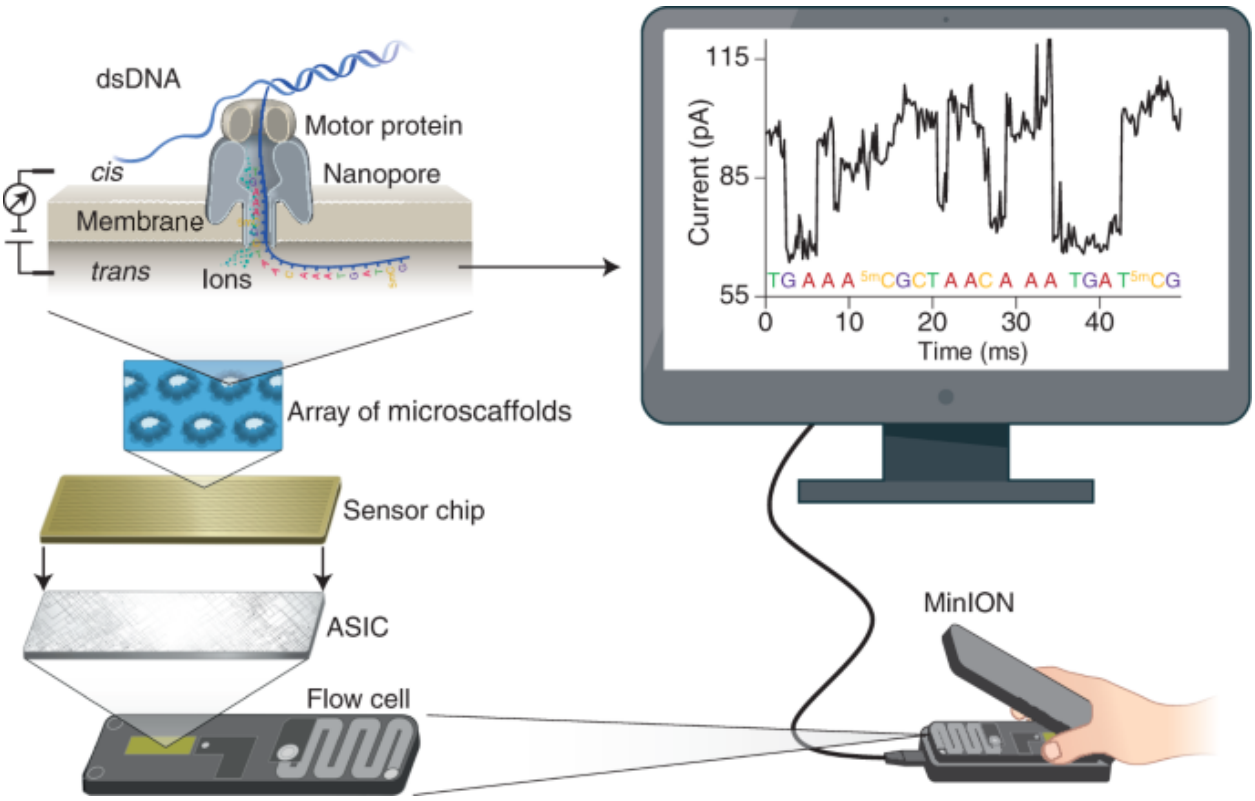


Figure 59: Nanopore 시퀀싱

Figure 3.4 Nanopore 시퀀싱 – 나노포어를 통한 전류 변화 감지

Nanopore의 가장 큰 장점은 Long Read를 생성할 수 있다는 점이다. 리드 길이에 이론적 제한이 없으며, 수백 kb 또는 그 이상의 리드가 가능하다. 실제로는 수십 kb의 평균 리드 길이가 일반적이다. 이러한 Long Read는 복잡한 유전체 영역, 구조적 변이, 반복 서열을 분석하는 데 매우 유용하다. 그러나 Nanopore 시퀀싱은 Illumina에 비해 단일 염기 정확도가 낮은 편(95–98%)이지만, 최근 기술 발전으로 정확도가 지속적으로 향상되고 있다.

4.2.2.2 PacBio 시퀀싱 Pacific Biosciences에서 개발한 PacBio 시퀀싱은 단일 분자 실시간(Single-Molecule Real-Time, SMRT) 시퀀싱 기술을 사용한다. PacBio의 HiFi Sequencing은 고리 모양 DNA를 시퀀싱하는 방식으로 작동한다:

1. DNA가 SMRTbell이라고 부르는 고리 모양 구조로 준비된다.
2. SMRTbell은 샘플 DNA의 양쪽 끝에 어댑터가 부착되어 형성된다.
3. 이 고리 모양 구조로 인해 DNA 단편이 여러 번 반복적으로 시퀀싱된다.
4. 같은 DNA 단편을 여러 번 읽음으로써 컨센서스(consensus) 서열을 생성하여 정확도를 크게 향상시킨다.

Figure 3.5 PacBio HiFi 시퀀싱 – SMRTbell 구조와 반복 시퀀싱

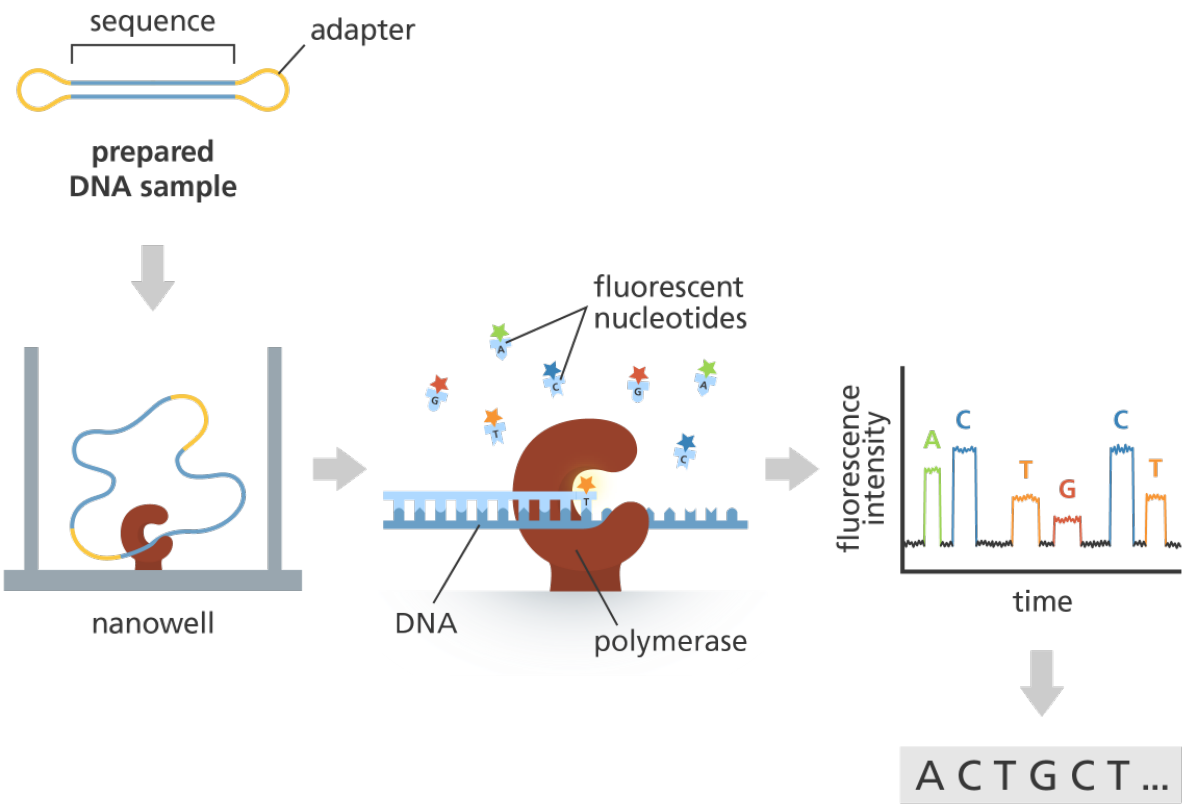


Figure 60: PacBio HiFi 시퀀싱

HiFi 시퀀싱은 Long Read 기술임에도 불구하고 Illumina에 필적하는 높은 정확도(99.9% 이상)를 제공한다. PacBio 시퀀싱은 Nanopore와 마찬가지로 Long Read를 생성하는데, HiFi 리드는 일반적으로 10–25kb 길이이다. 이러한 Long Read 길이와 높은 정확도의 조합으로 유전체 조립(genome assembly)과 구조적 변이 분석에 이상적이다. 그러나 처리량(throughput)이 Illumina보다 낮고 비용이 높은 편이다.

유전체 조립

유전체 조립은 시퀀싱으로 생성된 짧은 DNA 단편들을 원래의 긴 염색체 서열로 재구성하는 과정이다. 현재는 Short Read와 Long Read를 혼합하여 사용하는 하이브리드 조립 방법이 주로 사용되고 있다. Short Read는 높은 정확도를 제공하고, Long Read는 반복 서열과 구조적 변이 해결에 유리하여, 두 기술을 조합하면 더욱 완전하고 정확한 유전체 조립이 가능하다.

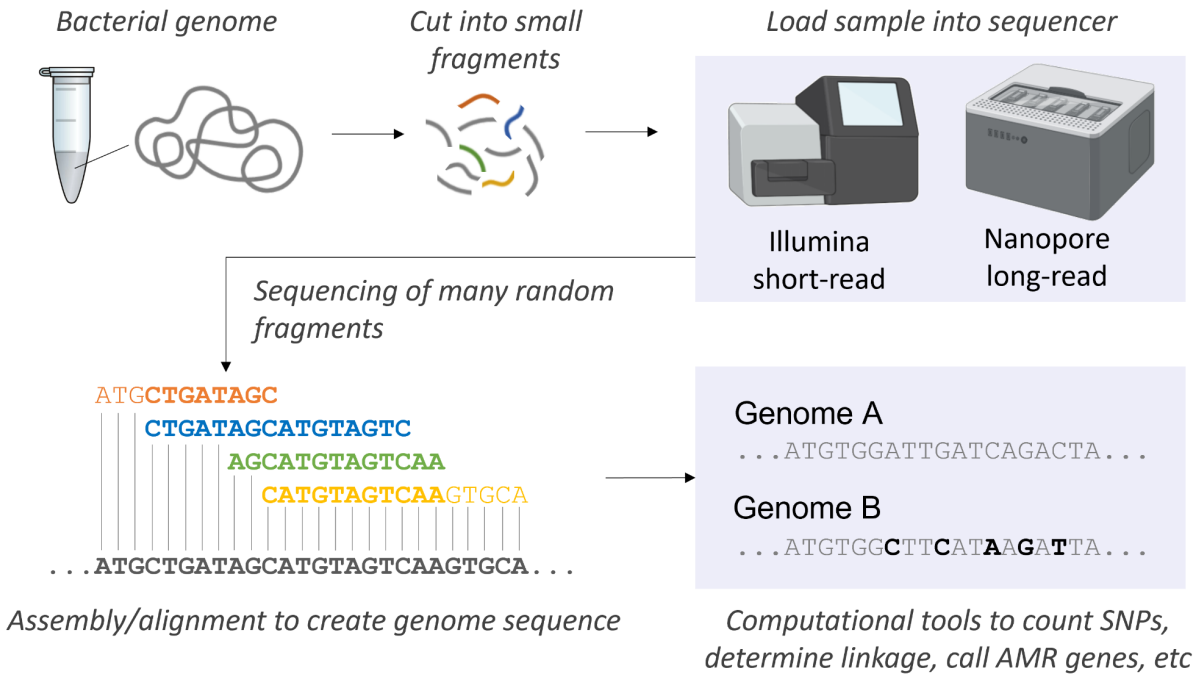


Figure 61: 유전체 조립

Figure 3.6 유전체 조립 – DNA 단편들을 원래의 염색체 서열로 재구성

Figure 3.7 Long Read와 Short Read 하이브리드 조립 – 두 기술의 장점을 결합한 조립 방법

Figure 3.8 미토콘드리아 유전체 조립 – 미토콘드리아 유전체의 원형 구조와 유전자 배치를 보여주는 원형 플롯

4.3 Illumina 시퀀싱 시스템의 구조

4.3.1 플로우 셀과 시퀀싱 칩의 이해

Illumina 시퀀싱의 핵심 구성 요소인 플로우 셀(flow cell)은 유리 슬라이드 기반의 표면으로, DNA 클러스터가 생성되는 장소이다. 표면에는 DNA 라이브러리와 결합하는 올리고뉴클레오티드가 고밀도로 부착되어 있으며, 여러 개의 레인(lane)으로 나누어져 있어 여러 샘플을 동시에 처리할 수 있다. 또한 시약이 흐르는 미세유체 채널 네트워크를 포함하고 있다.

시퀀싱 칩과 플로우 셀의 설계는 Illumina 시퀀서의 모델에 따라 다르며, 처리량과 리드 길이 간의 균형을 최적화하도록 설계되어 있다.

Figure 3.9 Illumina 시퀀싱 칩 – 플로우 셀의 물리적 구조

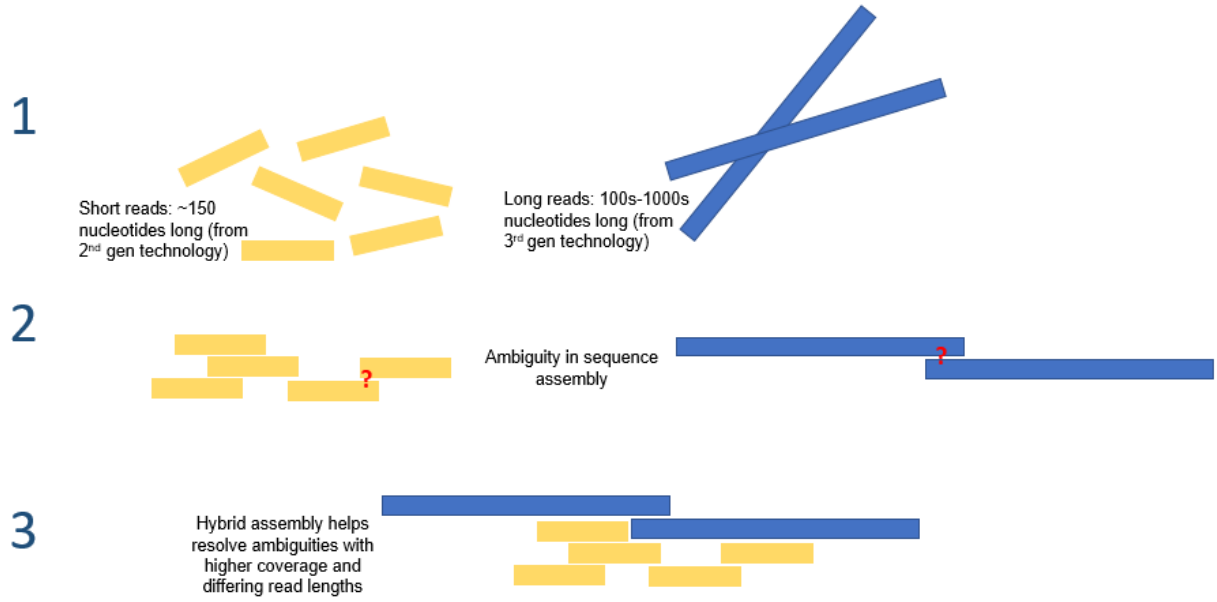


Figure 62: Long Read와 Short Read 하이브리드 조립

4.4 NGS 시퀀싱 라이브러리의 구조

NGS 분석을 위해서는 DNA 또는 RNA 샘플을 시퀀싱 라이브러리로 변환해야 한다. 시퀀싱 라이브러리는 다양한 기능적 요소를 포함하는 어댑터 시퀀스로 둘러싸인 DNA 단편들로 구성된다.

Figure 3.10 Illumina 라이브러리 구조와 시퀀싱 – 어댑터 시퀀스와 기능적 요소들

4.4.1 라이브러리 어댑터 구성요소

4.4.1.1 P5 및 P7 어댑터 시퀀스 Illumina 라이브러리에서 P5와 P7 시퀀스는 DNA 단편의 양쪽 끝에 부착되는 어댑터의 핵심 부분이다: – P5와 P7 시퀀스는 플로우 셀 표면에 결합하여 Bridge PCR을 가능하게 한다. – 이 시퀀스들은 각각 다른 올리고뉴클레오티드와 상보적으로 결합하여 DNA 라이브러리가 플로우 셀에 올바른 방향으로 부착되도록 한다. – P5는 일반적으로 Forward 리드에 연결되며, P7은 Reverse 리드에 연결된다.

4.4.1.2 인덱스 시퀀스 (바코드) 인덱스(Index, 또는 바코드라고도 함)는 다음과 같은 특성을 가진다: – 동일한 시퀀싱 런에서 여러 샘플을 구분하는 데 사용되는 고유한 시퀀스이다. – 일반적으로 6–12bp 길이의 시퀀스이다. – Index 1은 보통 P7 어댑터 가까이에, Index 2는 P5 어댑터 가까이에 위치한다. – 듀얼 인덱싱(dual indexing)은 두 개의 인덱스를 사용하여 더 많은 샘플 조합을 가능하게 한다. – 이러한 멀티플렉싱(multiplexing)을 통해 하나의 시퀀싱 런에서 여러 샘플을 동시에 처리함으로써 비용 효율성을 크게 향상시킨다.

4.4.1.3 시퀀싱 프라이머 결합 부위 시퀀싱 프라이머 결합 부위는 실제 시퀀싱 반응이 시작되는 위치를 제공한다: – Read1 시퀀싱 프라이머 결합 부위는 P5 어댑터 영역 내에 위치한다. – Read2 시퀀싱 프라이머 결합 부위는 P7 어댑터 영역 내에 위치한다. – 이 프라이머 결합 부위들은 시퀀싱 반응이 샘플 DNA 삽입물(insert)의 양 끝단에서 시퀀싱이 진행되도록 한다.

Figure 3.11 시퀀싱 프라이머 요구사항 – 시퀀싱 반응 시작을 위한 프라이머 결합 부위

4.5 시퀀싱 리드의 특성과 유형

Illumina 시퀀싱에서 Read 1과 Read 2는 각각 다른 시퀀싱 프라이머 다음부터 시작된다. Read 1은 Read1 시퀀싱 프라이머 다음에서, Read 2는 Read2 시퀀싱 프라이머 다음에서 시작되며, 두 리드는 DNA 단편의 반대 말단에서 시작된다.

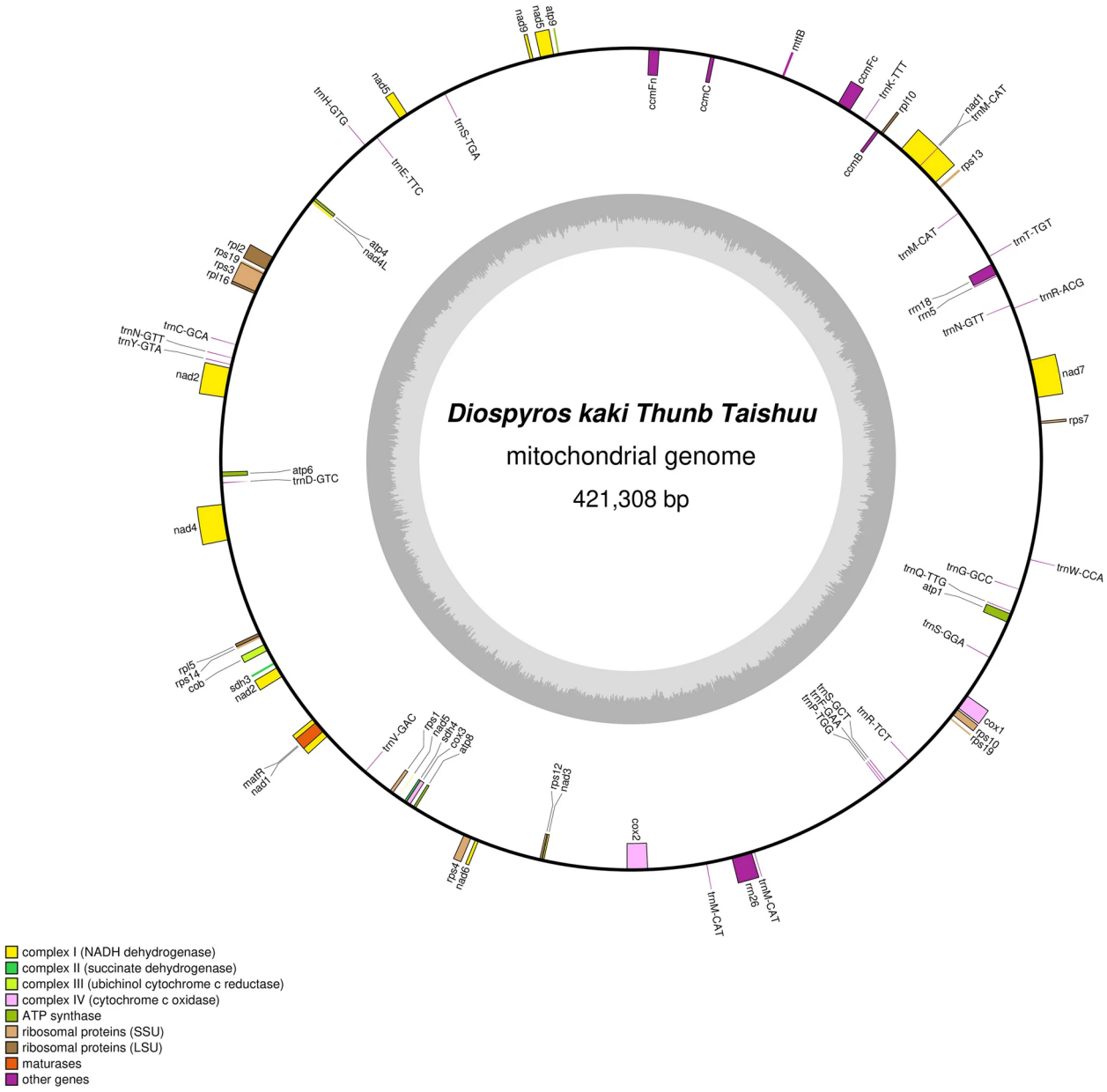


Figure 63: 미토콘드리아 유전체 조립

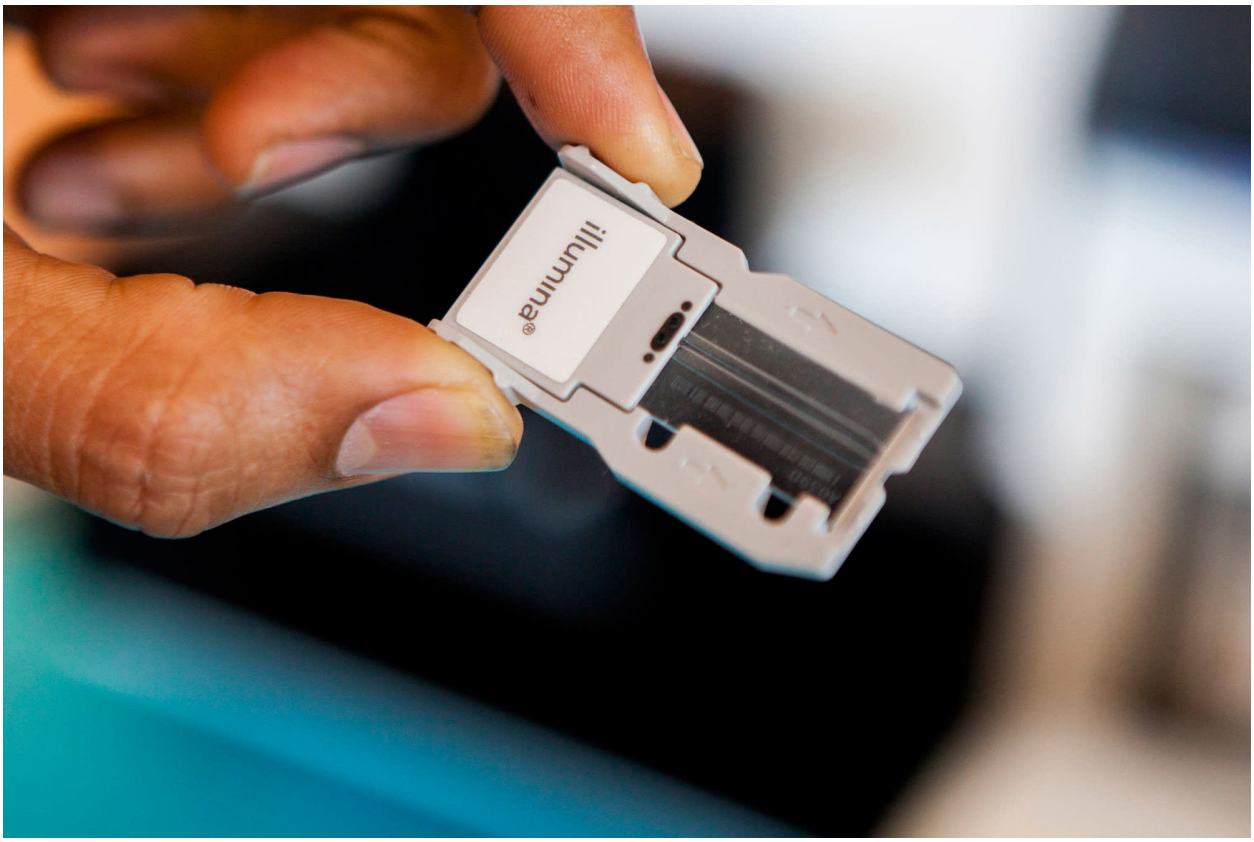


Figure 64: Illumina 시퀀싱 칩

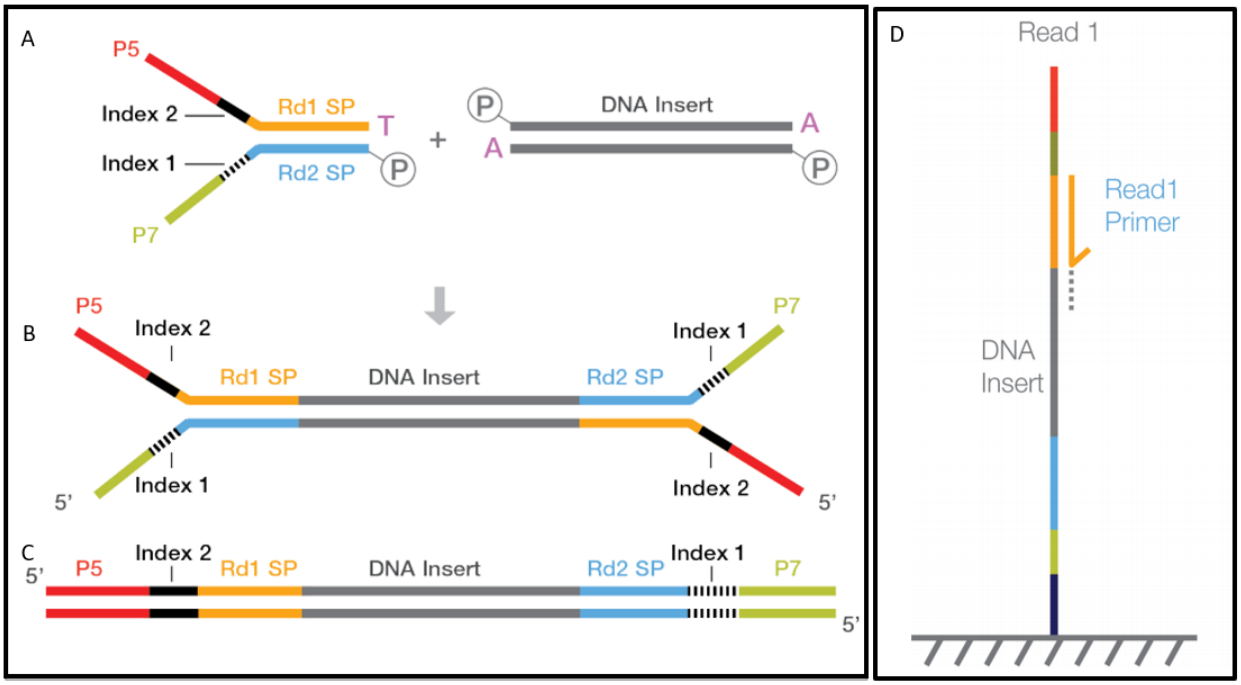


Figure 65: Illumina 라이브러리 구조와 시퀀싱

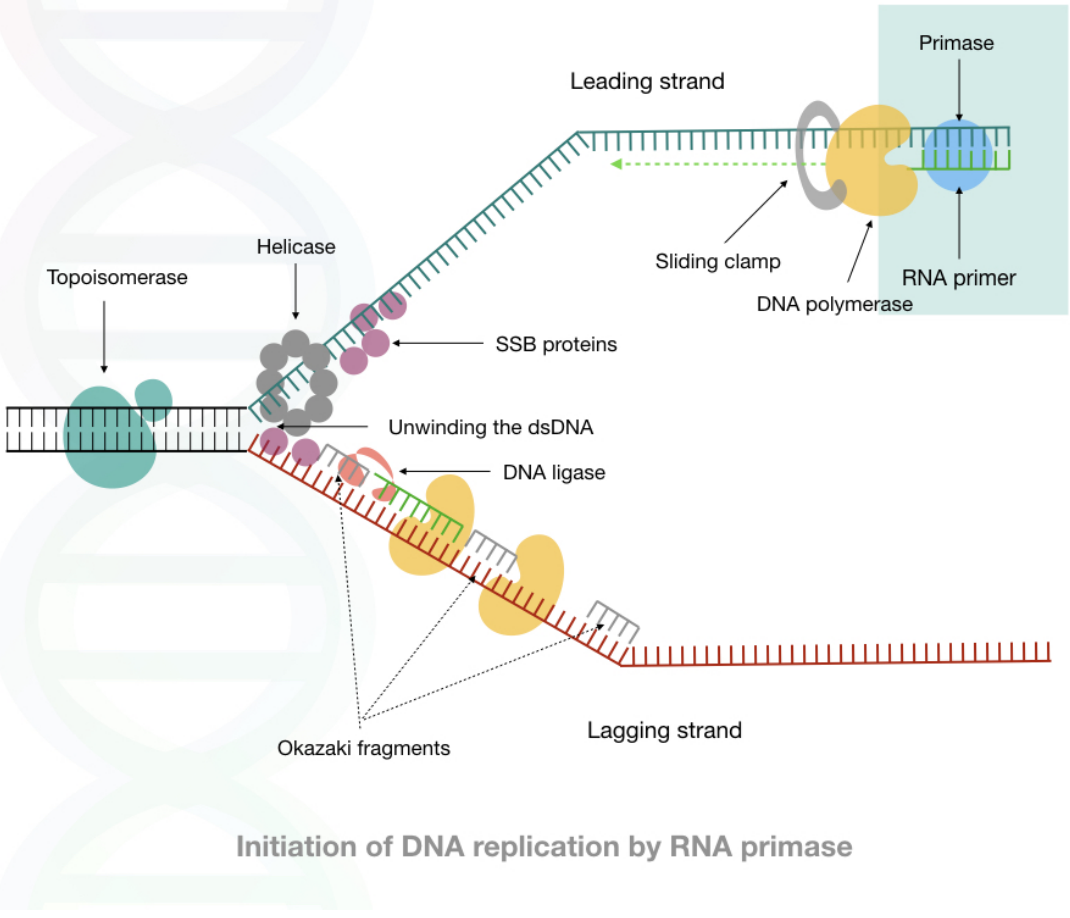


Figure 66: 시퀀싱 프라이머 요구사항

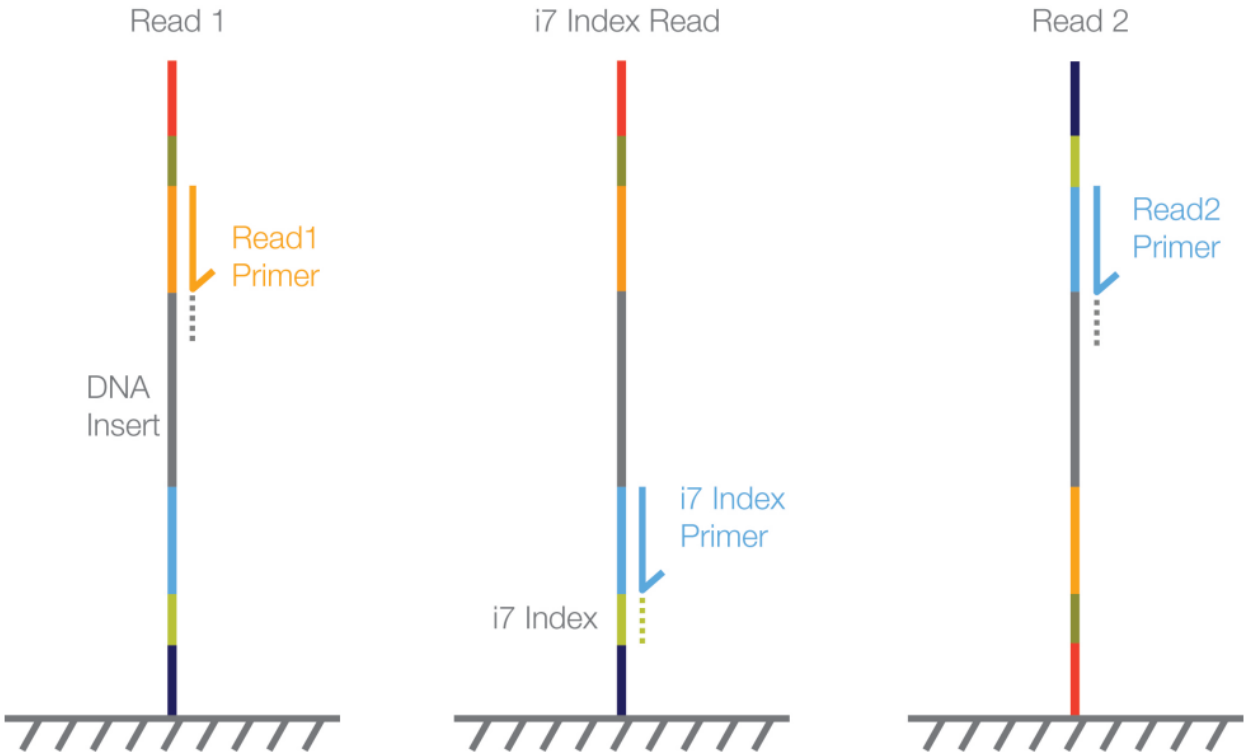


Figure 67: Read1과 Read2

Figure 3.12 Read1과 Read2 – DNA 단편 양쪽 끝에서의 시퀀싱 방향

일반적인 Illumina 시퀀싱은 각 방향에서 75bp, 100bp, 150bp 또는 250bp를 읽는다. 리드 길이는 시퀀서 모델과 키트에 따라 달라지며, 더 긴 리드는 더 많은 정보를 제공하지만 시퀀싱 비용이 증가한다.

DNA 삽입물(insert)이 리드 길이보다 짧은 경우, 리드가 반대쪽 어댑터 시퀀스를 포함할 수 있다. 이를 "리드 스루(read-through)"라고 하며, 분석 전에 어댑터 시퀀스를 트리밍(trimming)해야 한다. 이는 중복 정보를 제공하므로 시퀀싱 정확도를 향상시킬 수 있다.

4.5.1 페어드 엔드와 싱글 엔드 시퀀싱

페어드 엔드 시퀀싱(Paired-end sequencing)은 같은 DNA 단편의 양쪽 끝에서 시퀀싱하는 방식이다. Read 1과 Read 2 사이의 거리(insert size)를 알고 있기 때문에 정렬 정확도가 향상되며, 구조적 변이 감지, 유전체 조립, 반복 영역 해석에 유용하다. 다만 데이터 용량이 더 크고, 시퀀싱 시간과 비용이 더 많이 든다.

싱글 엔드 시퀀싱(Single-end sequencing)은 DNA 단편의 한쪽 끝만 시퀀싱하는 방식이다. 더 빠르고 비용이 적게 들지만, 정보량이 적다. 단순한 유전자 발현 분석이나 작은 변이 검출과 같은 응용에 적합하다.

Figure 3.13 싱글 엔드와 페어드 엔드 시퀀싱 – 두 방식의 차이점과 특징

Figure 3.14 페어드 엔드 시퀀싱 세부 – 구조적 변이 감지를 위한 리드 쌍 분석

4.6 Fastq 파일 형식과 구조

Fastq는 시퀀싱 데이터를 저장하는 표준 파일 형식으로, 시퀀스와 품질 정보를 모두 포함한다. 첫 번째 줄은 '@' 문자로 시작하고 리드 식별자를 포함한다:

```
@M04695:276:000000000-J9Y5P:1:1101:9698:1015 1:N:0:1
```

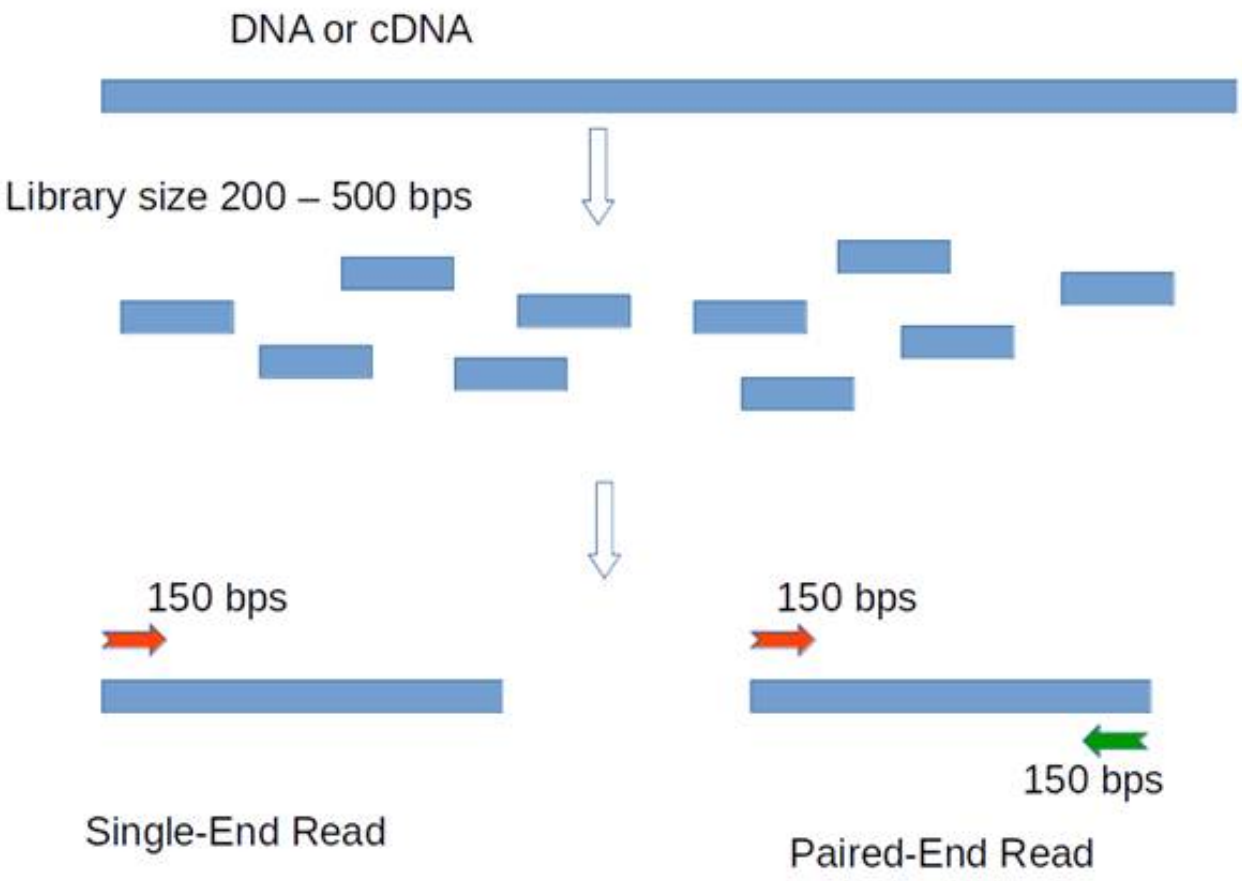


Figure 68: 싱글 엔드와 페어드 엔드

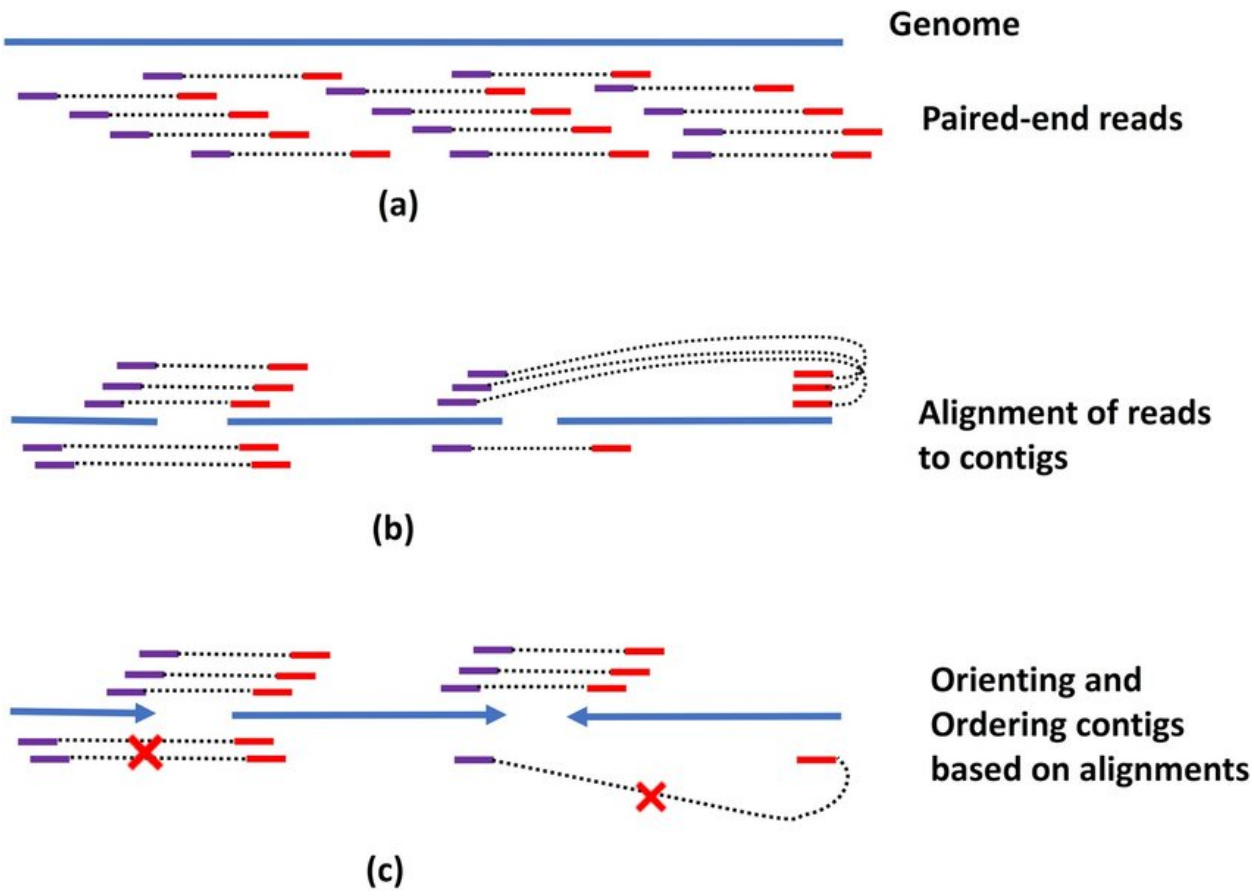


Figure 69: 페어드 엔드 시퀀싱 세부

이 식별자는 기기 ID, 런 번호, 플로우셀 ID, 레인 번호, 타일 번호, X 좌표, Y 좌표, 리드 번호, 필터 플래그, 컨트롤 번호, 인덱스 번호 등의 정보를 포함한다.

Element	Requirements	Description
@	@	Each sequence identifier line starts with @.
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID.
<run number>	Numerical	Run number on instrument.
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number.
<tile>	Numerical	Tile number.
<x_pos>	Numerical	X coordinate of cluster.
<y_pos>	Numerical	Y coordinate of cluster.
<UMI>	Restricted characters: A/T/G/C/N	Optional, appears when UMI is specified in sample sheet. UMI sequences for Read 1 and Read 2, seperated by a plus [+].
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end.
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise.
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<index>	Restricted characters: A/T/G/C/N	Index of the read.

Figure 70: Fastq Illumina 형식

Figure 3.15 Fastq Illumina 형식 – 시퀀싱 데이터의 표준 파일 형식과 구조

두 번째 줄은 시퀀싱된 염기 서열을 포함하며, 세 번째 줄은 '+' 문자로 시작한다. 네 번째 줄은 두 번째 줄의 각 염기에 대한 품질 점수를 ASCII 문자로 인코딩하여 포함한다.

Fastq 파일의 예제:

```
@M04695:276:000000000-J9Y5P:1:1101:9698:1015 1:N:0:1
NAGACGACTCTCCCCGCTATAGATN
+
#8ACCGGGGGGEF@CFGGGGFGGG#
@M04695:276:000000000-J9Y5P:1:1101:8957:1016 1:N:0:1
NTCAGCAAGAAGCCCCATCGAGATN
+
#8ACCGEGGGGGCFFGGGGGGGGG#
@M04695:276:000000000-J9Y5P:1:1101:12031:1016 1:N:0:1
NTAATCAATACGCCGCGGTTAGATN
+
#8ACCGGGGGGGGGGGGGGGFGG#
```

페어드 엔드 시퀀싱에서는 두 개의 별도 Fastq 파일이 생성되며(예: sample_R1.fastq, sample_R2.fastq), 첫 번째 파일은 Read 1 시퀀스를, 두 번째 파일은 Read 2 시퀀스를 포함한다. 두 파일의 리드 순서는 일치하여, 각 파일의 n번째 리드는 서로 대응된다. 두 Fastq 파일은 정확히 같은 수의 리드를 포함하며, 리드 식별자는 리드 번호(1 또는 2)를 제외하고 동일하다.

4.6.1 시퀀싱 품질 정보의 이해

품질 정보는 Phred 품질 점수를 사용하여 인코딩된다. Phred 점수(Q)는 염기 호출 오류 확률(P)과 로그 관계가 있다: $Q = -10 * \log_{10}(P)$. 예를 들어, Q10은 10% 오류 확률, Q20은 1% 오류 확률, Q30은 0.1% 오류 확률(일반적으로 '고품질'로 간주됨), Q40은 0.01% 오류 확률을 의미한다.

Figure 3.16 Phred 점수 – 염기 호출 품질을 나타내는 지표

ASCII 문자는 Phred 점수를 나타내는 데 사용되며, 현재 표준인 Illumina 1.8+ 형식은 ASCII 33 오프셋(! = Q0, " = Q1, ··· J = Q41, ···)을 사용한다. 이러한 품질 점수는 다운스트림 분석에서 필터링, 트리밍, 변이 호출 신뢰도 측정에 사용된다.

Figure 3.17 ASCII 코드와 Phred 점수 – 품질 점수의 ASCII 문자 인코딩

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Figure 71: Phred 점수

```
❯ ascii -d
   0 NUL    16 DLE    32      48 0    64 @    80 P    96 `   112 p
   1 SOH    17 DC1    33 !    49 1    65 A    81 Q    97 a   113 q
   2 STX    18 DC2    34 "    50 2    66 B    82 R    98 b   114 r
   3 ETX    19 DC3    35 #    51 3    67 C    83 S    99 c   115 s
   4 EOT    20 DC4    36 $    52 4    68 D    84 T   100 d   116 t
   5 ENQ    21 NAK    37 %    53 5    69 E    85 U   101 e   117 u
   6 ACK    22 SYN    38 &    54 6    70 F    86 V   102 f   118 v
   7 BEL    23 ETB    39 '    55 7    71 G    87 W   103 g   119 w
   8 BS     24 CAN    40 (    56 8    72 H    88 X   104 h   120 x
   9 HT     25 EM     41 )    57 9    73 I    89 Y   105 i   121 y
  10 LF     26 SUB    42 *    58 :    74 J    90 Z   106 j   122 z
  11 VT     27 ESC    43 +    59 ;    75 K    91 [   107 k   123 {
  12 FF     28 FS     44 ,    60 <    76 L    92 \   108 l   124 |
  13 CR     29 GS     45 -    61 =    77 M    93 ]   109 m   125 }
  14 SO     30 RS     46 .    62 >    78 N    94 ^   110 n   126 ~
  15 SI     31 US     47 /    63 ?    79 O    95 _   111 o   127 DEL
```

Figure 72: ASCII 코드와 Phred 점수

4.7 NGS 데이터의 정렬과 알고리즘

시퀀싱된 리드는 참조 유전체(reference genome)에 정렬하여 리드의 유전체 내 위치를 결정한다. 정렬기(aligner)는 각 리드의 가장 가능성 높은 유전체 위치를 찾는 소프트웨어 도구로, 여러 가능한 위치 중에서 최상의 일치 항목을 식별하고 불일치(mismatch)와 갭(gap)을 허용한다.

서열 유사성 검색에는 BLAST (Basic Local Alignment Search Tool)와 같은 도구도 중요하게 사용된다:

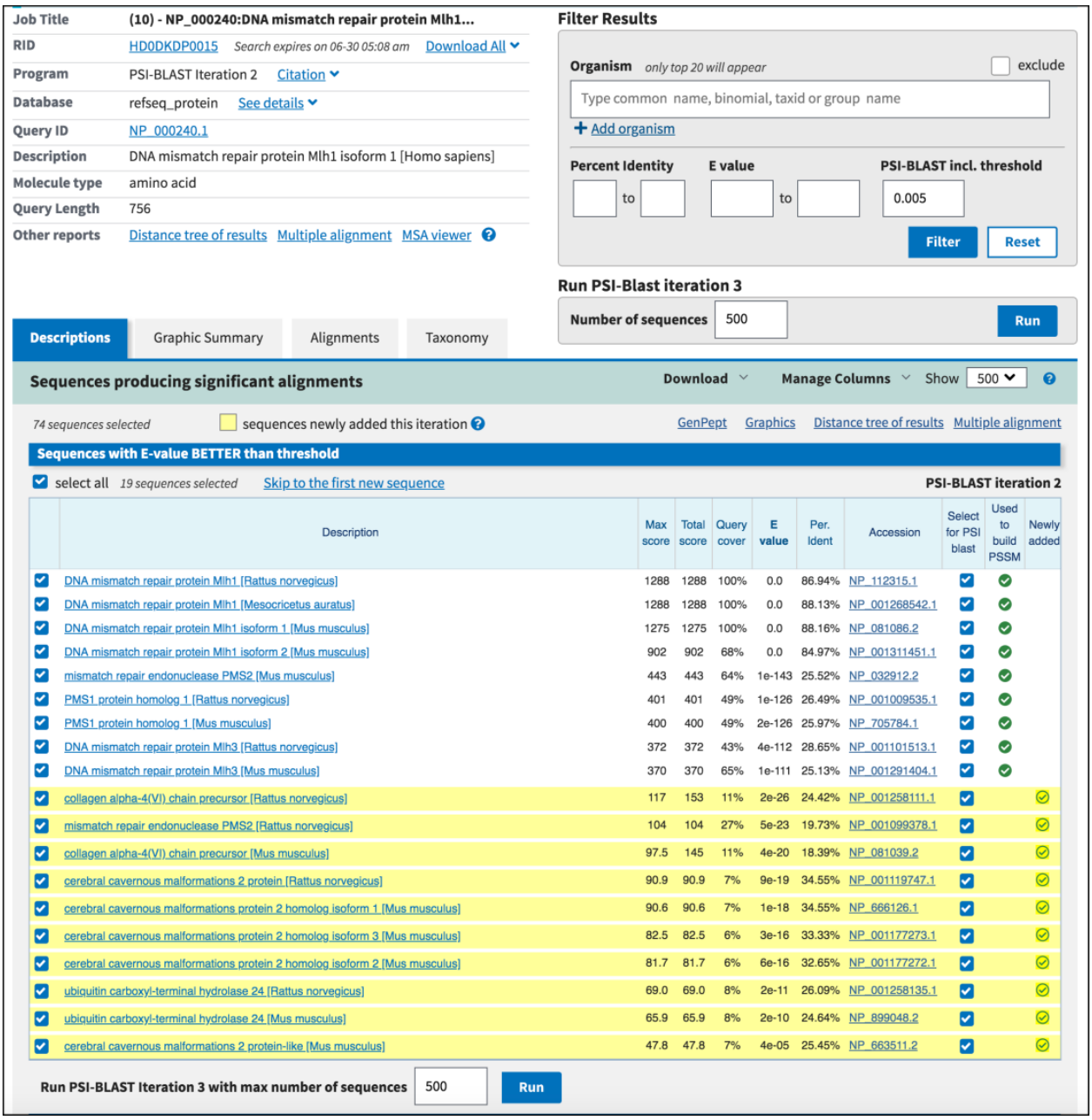


Figure 73: BLAST 알고리즘

Figure 3.18 BLAST 알고리즘 – 서열 유사성 검색을 위한 휴리스틱 접근법

다양한 정렬 도구가 사용 가능하며, 각각 특정 사용 사례에 최적화되어 있다: – BWA-MEM2: 전체 유전체 정렬에 널리 사용되는 도구로, BWA-MEM의 더 빠른 버전이다. – Diamond: 단백질 및 번역된 핵산 서열의 빠른 정렬에 특화되어 있다. – Isaac: Illumina에서 개발한 고속 정렬기로, 빠른 처리 시간이 특징이다. – Bowtie2: 짧은 리드에 최적화된

정렬기로, 전사체 분석에 자주 사용된다. - STAR: RNA-seq 데이터 정렬에 널리 사용되는 도구로, 스플라이싱을 효율적으로 처리한다.

4.7.1 주요 정렬 알고리즘의 원리

여러 알고리즘이 시퀀스 정렬에 사용되며, 각각 특정 용도에 장단점이 있다.

4.7.1.1 편집 거리와 기본 정렬 알고리즘 편집 거리(Edit Distance)는 한 문자열을 다른 문자열로 변환하는 데 필요한 최소 작업 수를 측정한다. 여기에는 삽입(insertion, 문자 추가), 삭제(deletion, 문자 제거), 대체(substitution, 문자 변경)와 같은 작업이 포함된다. 편집 거리는 두 시퀀스 간의 유사성을 정량화하는 간단한 방법으로, 편집 거리가 작을수록 시퀀스가 더 유사하다.

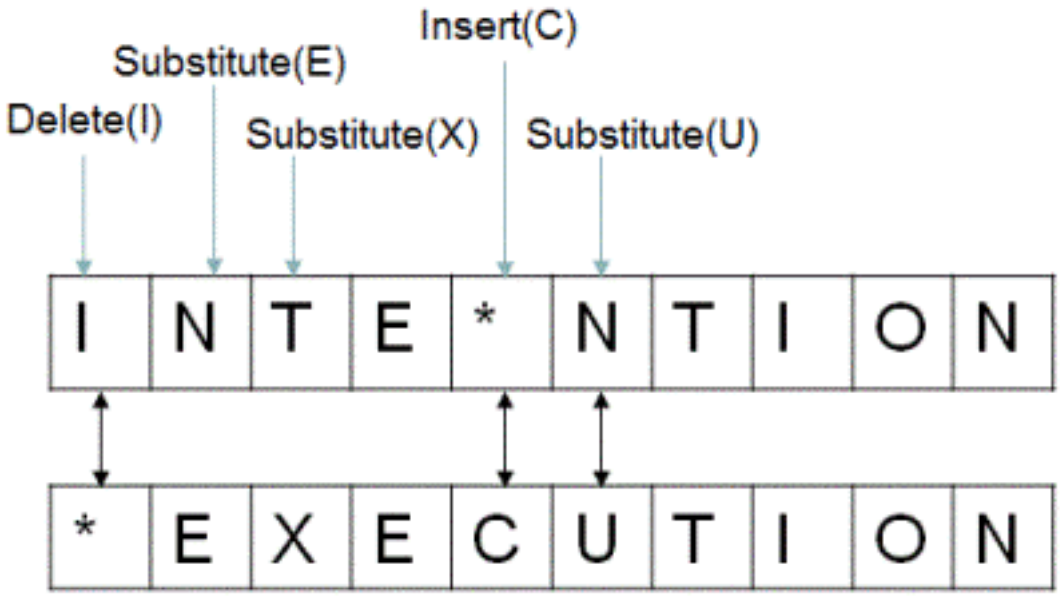


Figure 74: 해밍 거리

Figure 3.19 해밍 거리 – 두 문자열 간의 차이를 측정하는 방법

Needleman-Wunsch(NW) 알고리즘은 전역 정렬(global alignment)을 위한 동적 프로그래밍 접근 방식이다. 이 알고리즘은 두 시퀀스의 처음부터 끝까지 모두 정렬하며, 길이가 유사한 두 시퀀스를 비교할 때 적합하다.

4.7.1.1.1 3.7.1.1.1 Needleman-Wunsch 알고리즘 예시: “ATTGC”와 “ATGG” 정렬 NW 알고리즘을 사용하여 두 서열 “ATTGC”와 “ATGG”를 정렬하는 과정을 단계별로 살펴보자:

1단계: 빈 행렬 만들기

정렬하고자 하는 두 문자열의 길이를 각각 M, N이라 할 때, (0, 0)이 0인 (M+1)×(N+1) 행렬을 만든다.

		A	T	T	G	C
	0					
A						
T						
G						

	A	T	T	G	C
G					

2단계: 스코어 정의

문자가 일치하는 경우 +1, 일치하지 않으면 −1, 삽입/삭제(Indel)인 경우 −2로 스코어를 정의한다.

3단계: 행렬 채우기

다음 세 값 중 가장 큰 값으로 빈 칸을 채운다: − 두 문자가 일치할 경우 왼쪽 위 값에서 +1을, 그렇지 않을 경우 −1을 더한다 − 왼쪽 값에서 −2만큼 더한다 (insertion) − 위쪽 값에서 −2만큼 더한다 (deletion)

완성된 행렬:

		A	T	T	G	C
	0	−2	−4	−6	−8	−10
A	−2	1	−1	−3	−5	−7
T	−4	−1	2	0	−2	−4
G	−6	−3	0	1	1	−1
G	−8	−5	−2	−1	2	0

4단계: 트레이스백

맨 오른쪽 아래부터 해당 원소의 값을 계산하는데 쓰였던 원소를 찾아 트레이싱한다.

결과 정렬:

```
ATTGC
AT-GG
```

이 정렬의 최종 점수는 0이다.

Smith-Waterman(SW) 알고리즘은 지역 정렬(local alignment)을 위한 동적 프로그래밍 접근 방식이다. 이 알고리즘은 더 큰 시퀀스 내에서 유사한 영역을 찾는 데 적합하며, NGS에서는 참조 유전체 내에서 짧은 리드를 정렬하는 데 유용하다.

4.7.1.1.2 3.7.1.1.2 Smith-Waterman 알고리즘 예시: "AATGC"와 "ATGG" 정렬

SW 알고리즘을 사용하여 두 서열 "AATGC"와 "ATGG"를 정렬하는 과정을 단계별로 살펴보자:

1단계: 빈 행렬 만들기

정렬하고자 하는 두 문자열의 길이를 각각 M, N이라 할 때, 첫 열과 첫 행이 0으로 채워진 (M+1)×(N+1) 행렬을 만든다.

		A	A	T	G	C
	0	0	0	0	0	0
A	0					
T	0					
G	0					
G	0					

2단계: 스코어 정의

매치할 경우 +3, 매치하지 않을 경우 −3, 삽입/삭제(Indel)는 −2로 정의한다.

3단계: 행렬 채우기

다음 세 값 중 가장 큰 값으로 빈 칸을 채운다. 단, 음수이면 0으로 한다: – 두 문자가 일치할 경우 왼쪽 위 값에서 +3을, 그렇지 않을 경우 −3을 더한다 – 왼쪽 값에서 −2만큼 더한다 (insertion) – 위쪽 값에서 −2만큼 더한다 (deletion)

완성된 행렬:

		A	A	T	G	C
	0	0	0	0	0	0
A	0	3	3	1	0	0
T	0	1	1	6	4	2
G	0	0	0	4	9	7
G	0	0	0	2	7	6

4단계: 트레이스백

가장 큰 값(9)부터 좌측, 좌상단, 상단의 원소를 따라서 가장 큰 점수를 찾으며 올라온다. 이때 0을 발견하면 멈춘다.

결과 정렬:

```
ATG
ATG
```

이 지역 정렬의 최고 점수는 9이며, 두 서열에서 공통으로 존재하는 “ATG” 부분이 일치하는 것으로 식별되었다.

4.7.1.2 BLAST와 고급 정렬 기법 BLAST(Basic Local Alignment Search Tool)는 데이터베이스에서 유사한 시퀀스를 빠르게 검색하는 휴리스틱 알고리즘이다. 이 알고리즘은 주어진 문자열에서 “복잡한” 부분을 찾아 k-mer의 목록을 만들고, 이를 데이터베이스와 매치시킨 후 Smith-Waterman 알고리즘을 활용해 확장하여 정렬한다. BLAST는 대용량 데이터베이스에서 신속한 검색을 위해 정확도와 속도 간의 균형을 맞추기 위한 휴리스틱 접근 방식을 사용한다.

BLAST 결과에서 중요한 통계적 지표 중 하나가 E-value(Expectation value)이다. E-value는 해당 정렬 점수 이상의 결과가 우연히 나타날 것으로 예상되는 횟수를 의미한다. 예를 들어, E-value가 0.01이라면 같은 검색을 100번 수행했을 때 우연히 1번 정도는 이와 같거나 더 좋은 점수의 정렬이 나타날 것으로 예상된다는 의미이다. E-value가 낮을수록(일반적으로 10^{-5} 이하) 더 유의미한 유사성을 나타낸다. E-value는 데이터베이스 크기와 쿼리 시퀀스 길이에 따라 조정되므로, 서로 다른 검색 조건에서도 비교 가능한 신뢰도 지표를 제공한다.

4.7.1.2.1 3.7.1.2.1 k-mer 기반 검색 예시 k-mer는 DNA 서열에서 길이가 k인 모든 가능한 부분 문자열을 의미한다. 예를 들어, 시퀀스 “ATGTAC”에서 다양한 길이의 k-mer를 추출하면 다음과 같다:

k=3 (3-mer 또는 트라이머): – ATG (위치 1–3) – TGT (위치 2–4) – GTA (위치 3–5) – TAC (위치 4–6)

총 4개의 3-mer가 생성된다.

k=4 (4-mer 또는 테트라머): – ATGT (위치 1–4) – TGTA (위치 2–5) – GTAC (위치 3–6)

총 3개의 4-mer가 생성된다.

k=5 (5-mer 또는 펜타머): – ATGTA (위치 1–5) – TGTAC (위치 2–6)

총 2개의 5-mer가 생성된다.

이러한 k-mer 방식은 특히 BLAST와 같은 알고리즘에서 “시드(seed)” 역할을 하여 초기 일치점을 빠르게 찾는 데 사용된다. 일반적으로 k 값이 클수록 특이성은 높아지지만 민감도는 낮아지는 특성이 있다. NGS 분석에서는 주로 k=21~31 범위의 값이 많이 사용된다.

4.7.2 유전체 정렬 기술

전체 유전체 정렬에는 특수한 알고리즘이 필요하다. 현대 유전체 정렬기는 일반적으로 Burrows-Wheeler Transform이라는 알고리즘을 사용하여 정렬 후보 위치를 찾고, 후보 위치에서 Smith-Waterman과 유사한 알고리즘으로 정밀한 정렬을 수행한다.

4.7.2.1 Burrows-Wheeler Transform과 FM Index

문자열 "BANANA"에서 Burrows-Wheeler Transform를 수행한 후 FM Index를 구성하여 패턴을 검색하는 과정을 살펴보자.

4.7.2.1.1 Burrows-Wheeler Transform (BWT)

1단계: 원본 문자열에 종결 문자 추가

```
BANANA$
```

2단계: 모든 순환 이동(rotation) 생성

```
BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA
```

3단계: 사전순으로 정렬

```
$BANANA (1)
A$BANAN (2)
ANA$BAN (3)
ANANA$B (4)
BANANA$ (5)
NA$BANA (6)
NANA$BA (7)
```

이때 마지막 열의 문자들을 순서대로 읽으면 `A$ANNB`가 되는데, 이것이 바로 **Burrows-Wheeler Transform(BWT)**의 결과이다.

4.7.2.1.2 FM Index

BWT 결과를 바탕으로 **FM Index**를 구성한다: - 첫 열(F): 정렬된 회전의 첫 문자: `$AABNN` - 마지막 열(L): 정렬된 회전의 마지막 문자: `A$ANNB`

인덱스	F	회전된 문자열	L
1	$	$BANANA\|A\|\|2\|A\|A$BANAN	N
3	A	ANA$BAN\|N\|\|4\|A\|ANANA$B	B
5	B	BANANA$	$
6	N	NA$BANA\|A\|\|7\|N\|NANA$BA	A

4.7.2.1.3 패턴 검색: FM Index를 이용한 "ANA" 검색

FM 인덱스를 사용한 역방향 검색: 1. 마지막 문자 'A'로 시작: - L 열에서 'A'는 인덱스 1, 6, 7에 있다.

2. 다음 문자 'N'을 찾기 위해, L 열에서 'A'의 위치를 F 열에서의 'A' 위치로 매핑:
   - 'A'는 F 열에서 인덱스 2, 3, 4에 있다.
   - L 열의 'A'(인덱스 1)는 F 열의 첫 번째 'A'(인덱스 2)에 매핑된다.
   - L 열의 'A'(인덱스 6)는 F 열의 두 번째 'A'(인덱스 3)에 매핑된다.
   - L 열의 'A'(인덱스 7)는 F 열의 세 번째 'A'(인덱스 4)에 매핑된다.
3. 'A' 다음에 'N'이 나오는지 확인:
   - F 열의 인덱스 2, 3, 4에 해당하는 L 열의 문자는 각각 'N', 'N', 'B'이다.
   - 'B'는 'N'이 아니므로 인덱스 4는 제외된다.
   - 인덱스 2와 3이 유효하므로, 이들 위치에서 계속 검색한다.
4. 마지막으로 'A'를 찾기 위해, L 열에서 'N'의 위치를 F 열에서의 'N' 위치로 매핑:
   - 'N'은 F 열에서 인덱스 6, 7에 있다.
   - L 열의 'N'(인덱스 2)는 F 열의 첫 번째 'N'(인덱스 6)에 매핑된다.
   - L 열의 'N'(인덱스 3)는 F 열의 두 번째 'N'(인덱스 7)에 매핑된다.

5. ‘N’ 다음에 ’A’가 나오는지 확인:
    - F 열의 인덱스 6, 7에 해당하는 L 열의 문자는 모두 ’A’이다.
    - 두 위치 모두 유효하므로, 패턴 “ANA”는 원본 문자열에서 두 번 발견된다.

패턴 “ANA”는 “BANANA”에서 다음 위치에서 발견된다: - 인덱스 1에서 시작 (BANANA) - 인덱스 3에서 시작 (BA-NANA)

따라서 패턴 “ANA”의 존재가 확인된다. 패턴의 위치를 검색하는 알고리즘은 조금 더 복잡하며, 이는 독자의 과제로 남겨둔다.

이처럼 BWT를 활용한 FM 인덱스는 BWT를 사용하여 패턴을 효율적으로 검색할 수 있게 해준다. 이 방법은 대용량 참조 유전체에서 짧은 NGS 리드를 정렬하는 데 매우 효과적이다.

이를 활용한 도구인 BWA-MEM2 는 다양한 길이의 리드를 처리할 수 있으며, 높은 정확도와 빠른 속도로 인해 전체 유전체 시퀀싱 데이터 정렬에 널리 사용된다.

4.7.3 SAM 파일 형식과 구조

SAM(Sequence Alignment/Map) 포맷은 정렬된 시퀀스 데이터를 저장하기 위한 표준 텍스트 형식이다. SAM 파일은 헤더 섹션과 정렬 섹션으로 구성되며, 헤더 라인은 @로 시작하고 참조 시퀀스 정보와 정렬 메타데이터를 포함한다. 정렬 라인은 탭으로 구분된 필드를 포함하며, 각 리드의 정렬 정보를 나타낸다.

주요 SAM 필드는 다음과 같다: - QNAME: 쿼리 템플릿 이름으로, Fastq 파일의 리드 식별자와 일치한다. - RNAME, POS: 참조 시퀀스 이름(일반적으로 염색체 이름)과 참조 시퀀스에서 정렬의 왼쪽 끝 위치(1-기반 좌표계) - MAPQ: 매핑 품질 값으로, 정렬의 신뢰도를 나타낸다(0-255, 높을수록 더 신뢰할 수 있음). - SEQ: 리드의 염기 서열을 포함한다. - CIGAR: 리드가 참조 시퀀스에 어떻게 정렬되는지를 설명하는 문자열이다.

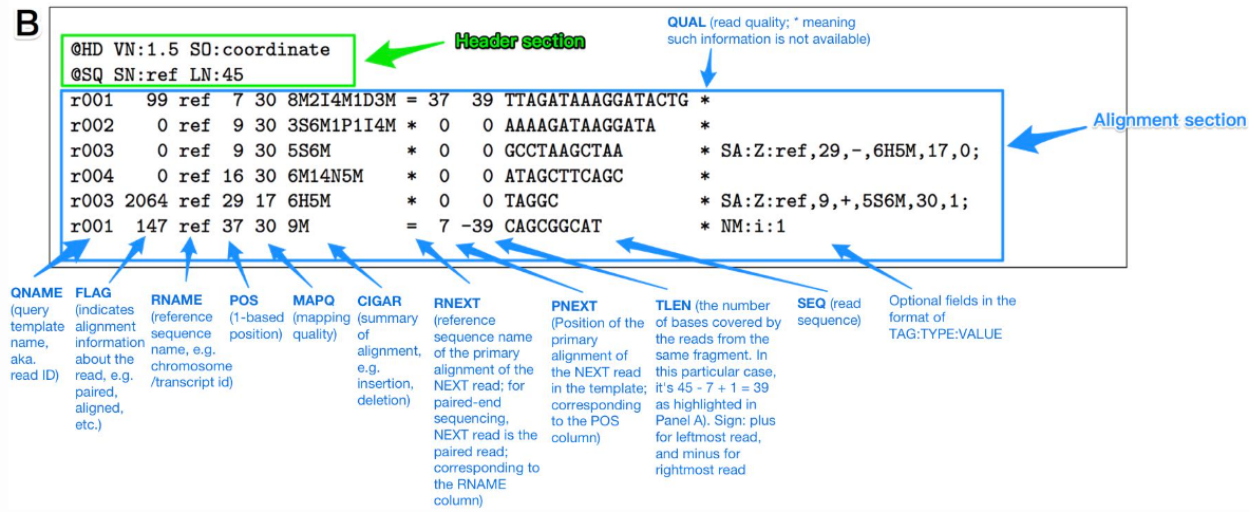


Figure 75: SAM 파일 형식

Figure 3.20 SAM 파일 형식 – 정렬된 시퀀스 데이터의 표준 텍스트 형식

CIGAR 문자열은 일련의 작업과 길이 쌍으로 구성되며, 예를 들어 “2S5M2I4M1D5M6S”는 첫 2개 염기가 소프트 클리핑되고, 다음 5개 염기가 참조 시퀀스와 정렬되며, 2개 염기가 삽입되고, 다음 4개 염기가 참조 시퀀스와 정렬되고, 참조 시퀀스에서 1개 염기가 삭제되고, 다음 5개 염기가 참조 시퀀스와 정렬되며, 마지막 6개 염기가 소프트 클리핑됨을 나타낸다.

CIGAR 문자열은 참조 시퀀스와 리드 간의 정확한 정렬 관계를 이해하는 데 중요하며, 변이 호출 및 구조적 변이 감지와 같은 다운스트림 분석에 필수적이다.

Figure 3.21 CIGAR 문자열 – 리드와 참조 시퀀스 간의 정렬 관계 표현

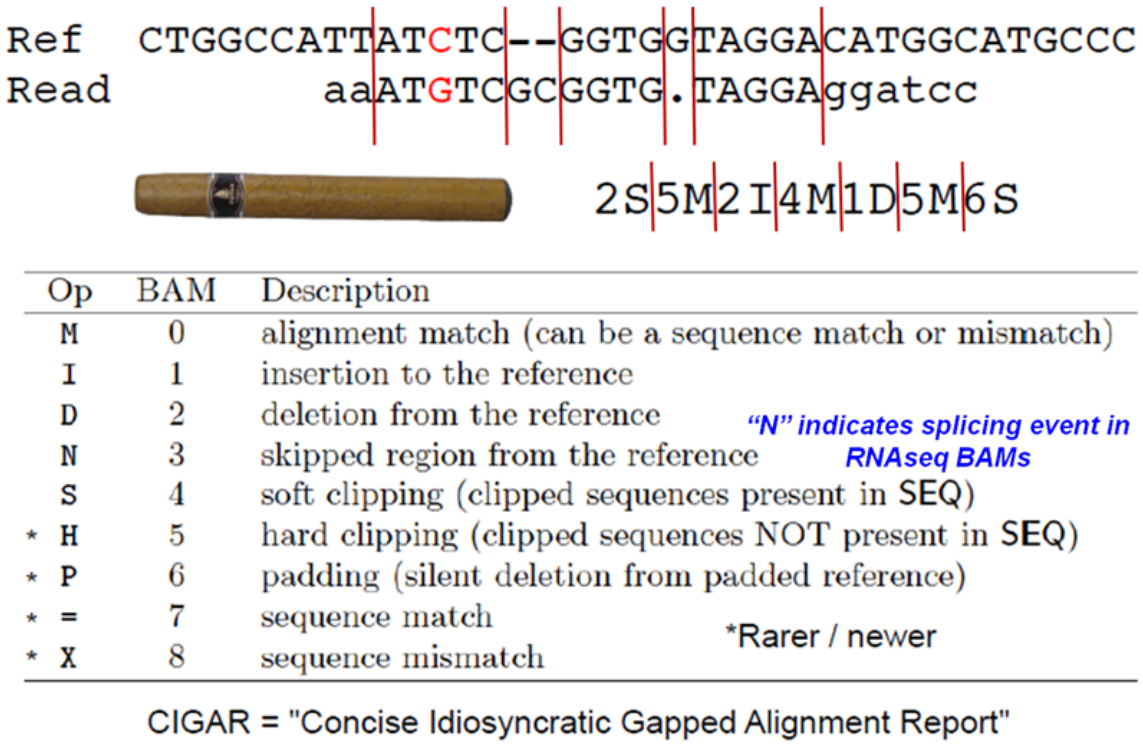


Figure 76: CIGAR 문자열

5 변이와 진화

5.1 변이 찾기 (Variant Calling)

변이 찾기는 주로 BAM 파일로부터 변이를 찾아주는 소프트웨어(variant caller)를 활용하는 과정이다. 이 과정은 개인의 유전체에서 참조 유전체과 다른 부분을 식별하는 중요한 단계이다. 현재 수십 개의 다양한 variant caller가 개발되어 있으며, 각각은 특정한 장단점과 적용 분야를 가지고 있어 연구 목적과 데이터 특성에 따라 적절한 도구를 선택해야 한다.

그 중 몇 가지 예시를 들면 다음과 같다:

- **GATK (Genome Analysis Toolkit)**: Broad Institute에서 개발한 도구로 인간 지놈 프로젝트에서 광범위하게 사용되어 왔으며 높은 정확도로 알려져 있다
- **SAMtools**: 시퀀싱 데이터 처리를 위한 포괄적인 도구 모음으로 변이 호출 기능을 포함하고 있다
- **FreeBayes**: 베이지안 통계 접근법을 사용하여 변이를 탐지하는 도구이다
- **DeepVariant**: 구글에서 개발한 딥러닝 기반 도구로 최근 높은 정확도를 보여주고 있다
- **Octopus**: 복잡한 변이 패턴도 효과적으로 탐지할 수 있는 새로운 접근 방식을 제공한다

이들 외에도 다양한 특화된 variant caller들이 계속 개발되고 있어, 특정 변이 유형이나 데이터 조건에 최적화된 도구를 선택할 수 있다.

5.2 VCF 포맷

VCF는 Variant Call Format의 약자로, 변이 정보를 저장하는 표준 파일 형식이다. 이 포맷은 유전체 위치, 참조 서열, 변이 서열 등 변이에 관한 중요 정보를 포함하고 있다.

VCF 파일의 주요 필드는 다음과 같다: - CHROM, POS: 유전체 상의 위치(Genomic Locus, 또는 복수형으로 Genomic Loci)를 나타낸다. CHROM은 염색체 번호, POS는 해당 염색체 내 위치를 의미한다. - ID: dbSNP와 같은 데이터베이스에 등록된 경우, 그 ID를 표시한다. - REF: 참조 유전체(Reference Genome) 상의 염기 서열이다. - ALT: 시퀀싱 데이터에서 얻은 변이된 염기 서열이다. - REF보다 ALT가 길면 삽입(insertion), 짧으면 결실(deletion)로, 이

둘을 합쳐 Indel이라 부른다. 길이가 같으면 단일 염기 변이(SNV, Single Nucleotide Variant)이다. – ALT가 쉼표로 구분되는 경우는 다중대립유전자 변이(Multiallelic Nucleotide Variant, MNV)이다. – FILTER: 변이 호출 과정에서 문제가 있는 경우 표시되며, 문제가 없으면 “PASS”로 표시된다.

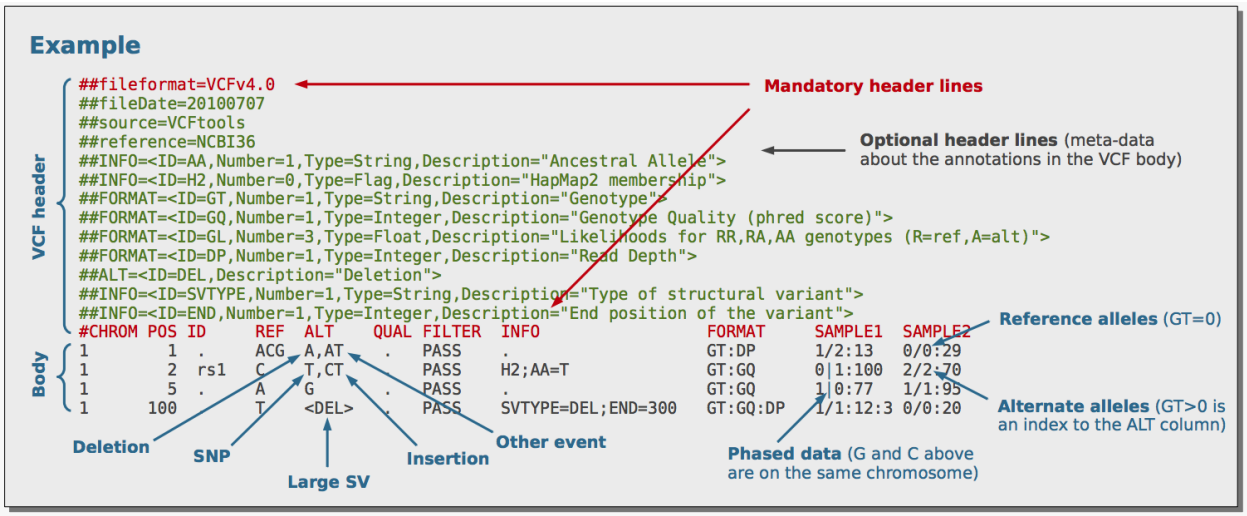


Figure 77: VCF 파일 형식

Figure 4.1 VCF 파일 형식 – 변이 정보를 저장하는 표준 파일 형식

VCF 파일은 또한 INFO 필드에 변이에 대한 추가 정보와 FORMAT 필드에 각 샘플의 유전형(genotype) 정보를 포함한다. 이러한 구조로 인해 VCF는 다양한 유전체 분석 도구와 호환되며, 변이 정보를 효율적으로 저장하고 공유할 수 있다.

5.3 Phasing

Phasing은 찾아낸 유전 변이(variant)가 포함된 대립유전자(allele)가 부계에서 왔는지 모계에서 왔는지 알아내는 과정이다. 이는 유전적 변이의 기원을 추적하고, 질병 연관성 등을 연구하는 데 중요한 정보이다.

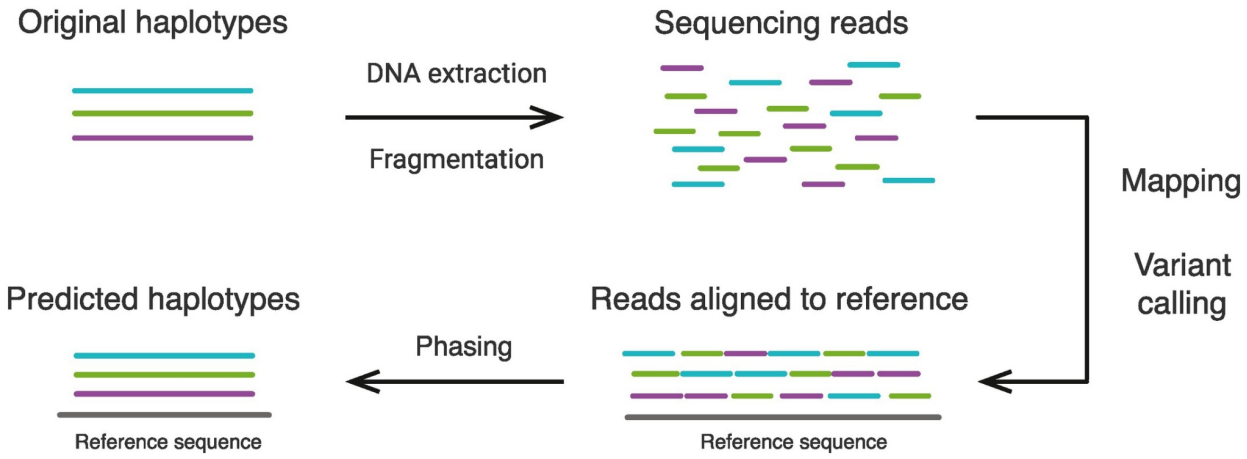


Figure 78: Phasing

Figure 4.2 Phasing – 대립유전자의 부모 기원 결정 과정

Phasing을 수행하는 두 가지 주요 방법이 있다: 1. 부모의 시퀀싱 데이터를 기반으로 부계인지 모계인지를 결정하는 방법 2. 알고리즘을 통한 통계적 추정 방법 (이 경우 부계/모계 유전인지 정확하게 파악하는 것은 어렵다)

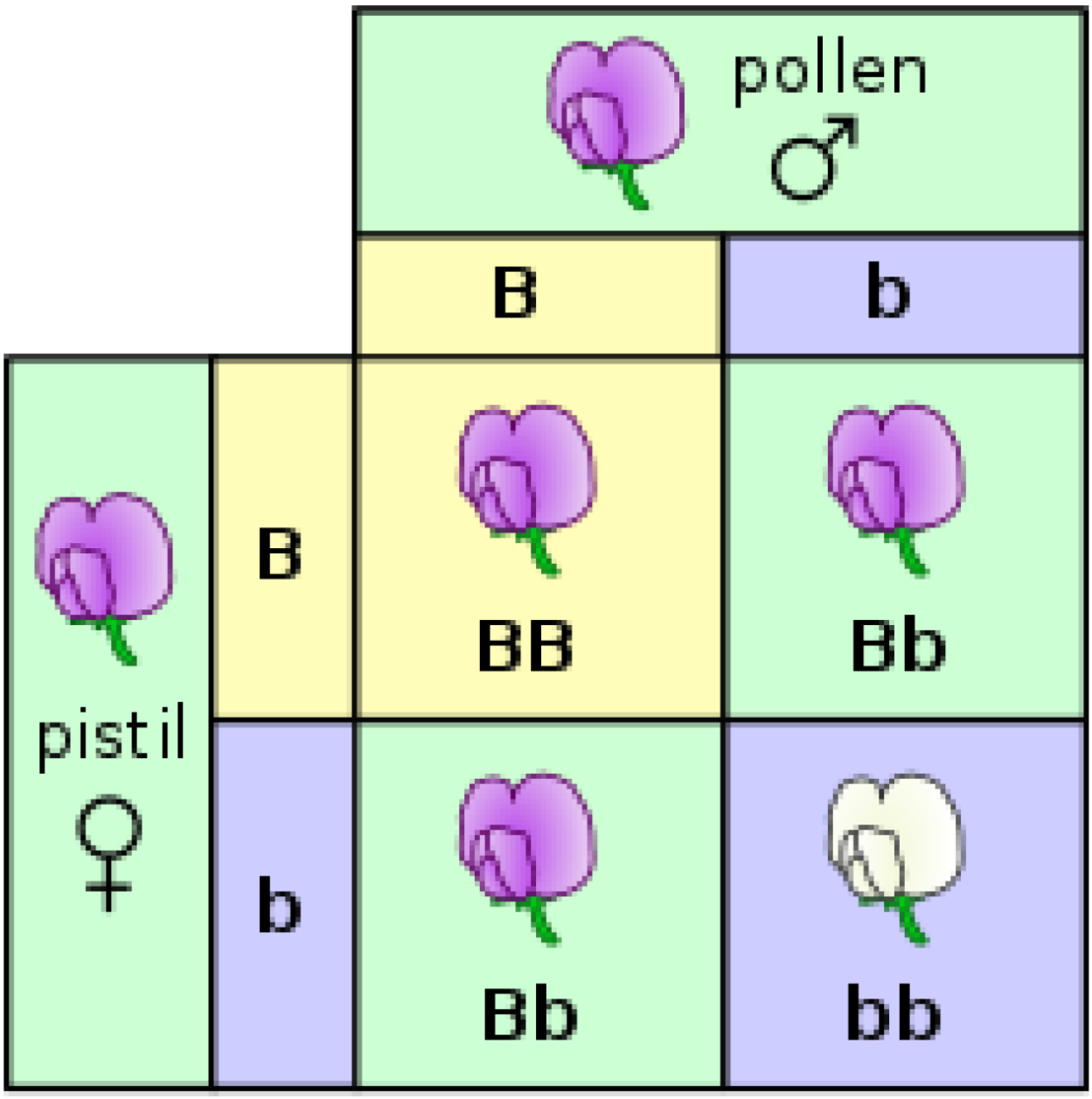


Figure 79: 대립유전자 교배

Figure 4.3 대립유전자 교배 – 꽃 색깔 유전을 보여주는 펀넷 사각형 예시

5장에서 더 자세히 다루겠지만, Phasing과 관련해 중요한 개념으로 Haplotype이 있다. Haplotype은 함께 유전되는 경향이 있는 유전자나 변이의 그룹을 의미한다. 이러한 haplotype 정보는 질병 연관성 연구, 집단 유전학, 진화 연구 등에서 중요하게 활용된다.

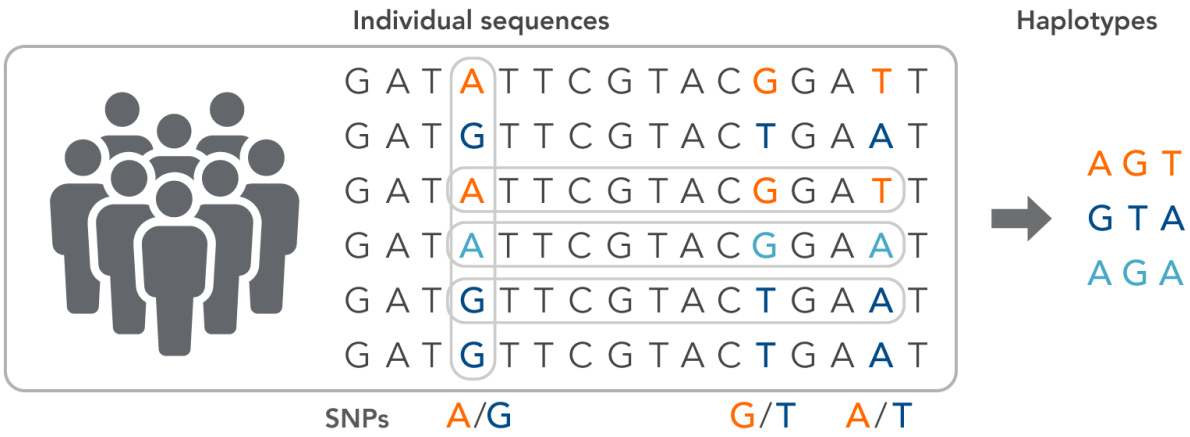


Figure 80: Haplotype

Figure 4.4 Haplotype – 함께 유전되는 유전자나 변이의 그룹

5.4 Contamination

Contamination(오염)은 variant calling 과정에서 반드시 고려해야 할 중요한 요소이다. 시퀀싱 샘플에 다른 개체의 DNA가 섞여 있거나, 실험 과정에서 오염이 발생할 수 있기 때문이다.

오염은 다양한 원인으로 발생할 수 있다: – 샘플 채취 과정에서의 오염 – 실험실 작업 중 교차 오염 – 시퀀싱 기기 자체의 오염 – 시약 오염

오염된 샘플은 잘못된 변이 호출로 이어질 수 있으며, 이는 연구 결과의 신뢰성에 심각한 영향을 미친다. 따라서 variant calling 전에 오염 여부를 확인하고, 오염된 경우 적절한 보정 방법을 적용하는 것이 중요하다.

5.5 개인 유전체 분석 과정

개인 유전체 분석은 일련의 단계를 거쳐 진행되는데, 이를 파이프라인이라고 부른다. 주요 단계는 다음과 같다:

1. Sequencing(시퀀싱): 개인의 DNA 샘플을 읽어 염기 서열 데이터를 생성한다.
2. Alignment(정렬): 생성된 시퀀스를 참조 유전체에 맞추어 정렬한다.
3. Variant Calling(변이 호출): 정렬된 데이터에서 참조 유전체과 다른 변이를 찾아낸다.
4. Gene set or Pathway analysis(유전자 세트 또는 경로 분석): 찾아낸 변이가 어떤 유전자나 생물학적 경로에 영향을 미치는지 분석한다.

이러한 파이프라인은 자동화되어 대량의 데이터를 효율적으로 처리할 수 있도록 설계된다. 각 단계마다 다양한 도구와 알고리즘이 사용되며, 연구 목적이나 대상 생물체에 따라 파이프라인을 조정할 수 있다.

Figure 4.5 전형적인 변이 호출 파이프라인 – 시퀀싱부터 변이 분석까지의 과정

5.6 Hardy–Weinberg Equilibrium

우리는 2장에서 다양한 변이 데이터베이스에 대해 배웠다. 그러나 각 변이의 빈도는 시간에 따라 일정한가? 만약 그렇지 않다면, 변이 데이터베이스는 지속적으로 갱신되어야 한다는 의미가 된다.

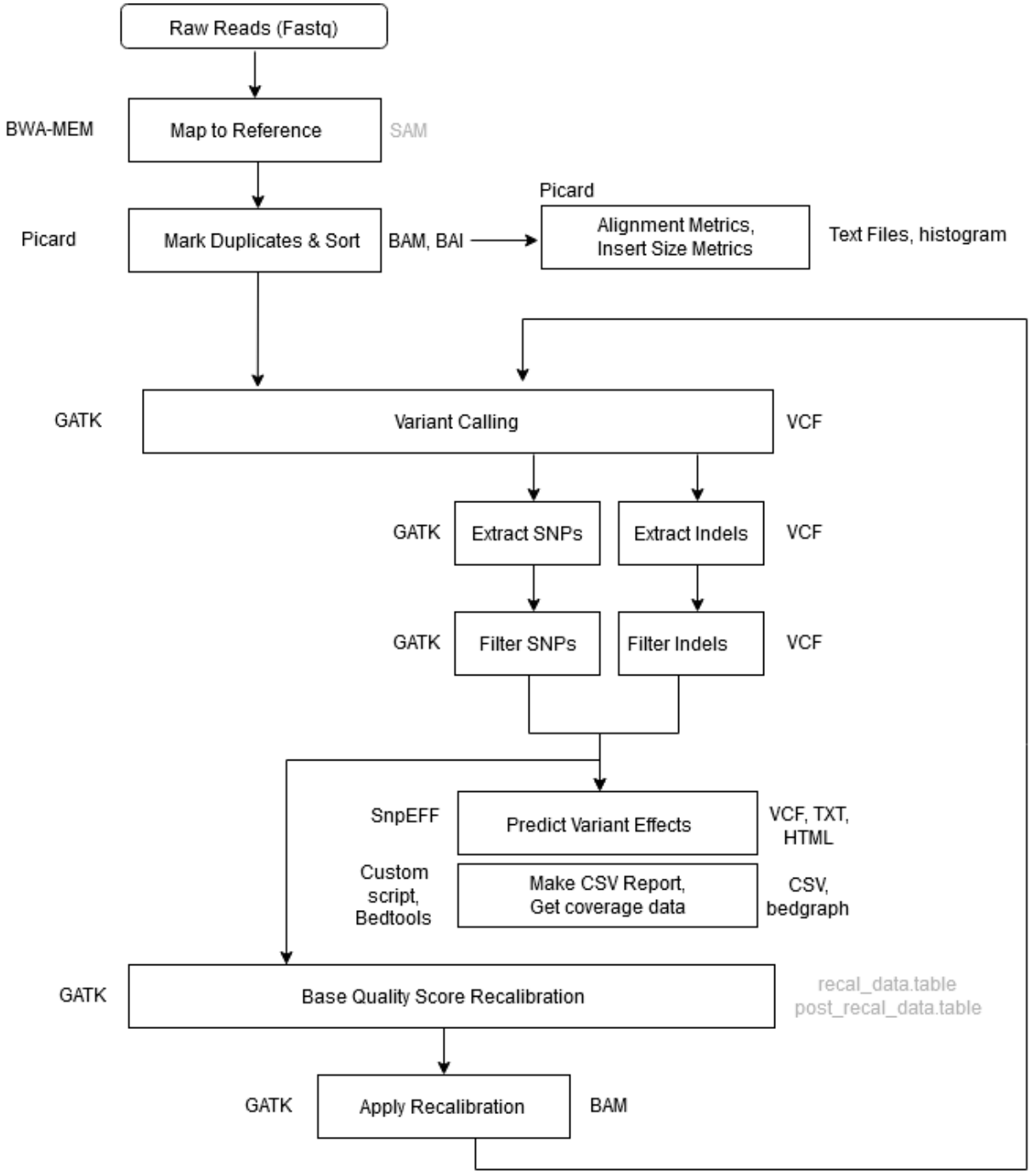


Figure 81: 전형적인 변이 호출 파이프라인

이때 등장하는 것이 Hardy-Weinberg Equilibrium(하디-바인베르크 평형) 이다. 이것은 집단 유전학의 기본 원리 중 하나로, 특정 조건 하에서 대립유전자 빈도가 세대를 거쳐도 변하지 않는다는 이론이다.

하디-바인베르크 평형이 성립하려면 다음 조건이 필요하다: 1. 무작위 교배 2. 돌연변이 없음 3. 자연선택 없음 4. 유전적 부동 없음 5. 무한한 집단 크기

당연히도 이런 조건들은 실제적으로 모두 충족되기 어렵다. 비록 상대적으로 짧은 세대에 걸쳐서는 하디-바인베르크 법칙이 어느정도 유지되는 상황들이 관찰되지만, **유전적 부동** 같은 현상으로 인해 하디-바인베르크 평형은 쉽게 깨질 수 있다.

5.7 Genetic Drift

Genetic Drift(유전적 부동)는 무작위적 과정에 의해 집단 내 대립유전자 빈도가 변화하는 현상이다. 특히 작은 집단에서 그 영향이 크게 나타난다.

라이트-피셔 모델(Wright-Fisher Model)은 유전적 부동을 설명하는 기본 모델이다: - 두 개의 대립유전자 A와 a의 빈도가 각각 p와 q인 N 크기의 집단을 고려해보자. - 총 대립유전자 수는 2N이 된다. - 모델의 단순화를 위해 다음을 가정한다: - 한 세대는 번식 후 완전히 소멸한다. - 교배는 완전히 무작위적이다. - 집단의 크기는 2N으로 고정되어 있다.

다음 세대에서 특정 대립유전자가 k개 존재할 확률은 복원 추출이므로 이항 분포를 따르게 된다.

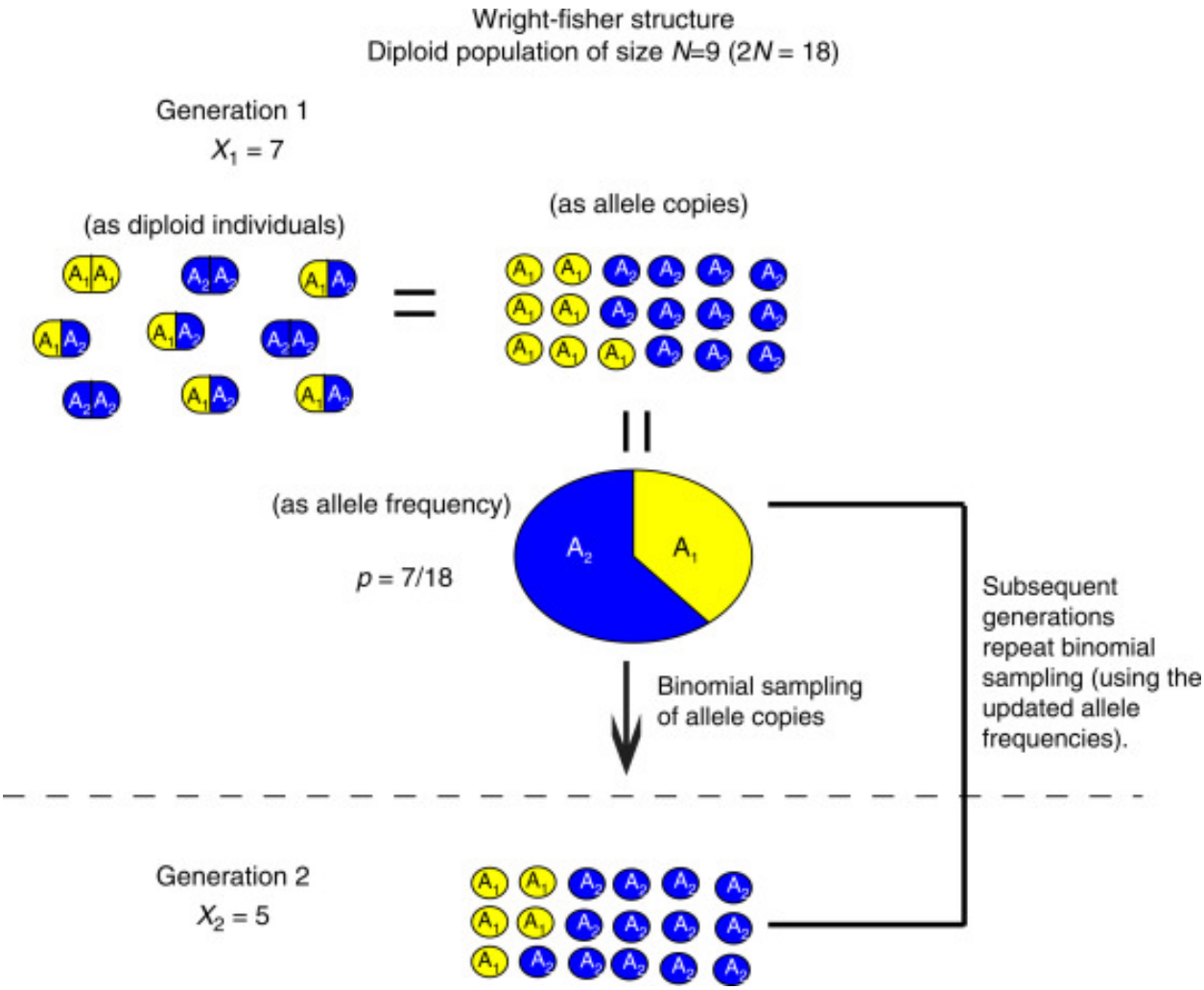


Figure 82: Wright-Fisher 모델

Figure 4.6 Wright-Fisher 모델 – 유전적 부동을 설명하는 기본 모델

$$Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Figure 83: Wright-Fisher 방정식

Figure 4.7 Wright-Fisher 방정식 – 대립유전자 빈도 변화의 수학적 모델

Figure 4.8 Wright-Fisher 모델 결과 – 시간에 따른 대립유전자 빈도 변화

이렇듯 유전은 무작위적 과정이므로, 유전적 부동을 통해 일부 대립유전자는 완전히 사라지거나 집단 내에서 우세해질 수 있다. 이는 하디-바인베르크 평형을 크게 깨뜨리는 요인이 된다.

우리는 하디-바인베르크 평형이 깨지는 현상을 "진화(Evolution)"라고 부른다. 유전적 부동은 진화를 이끄는 기본적인 힘 중 하나이다.

5.8 진화

진화(Evolution)는 생물 집단이 세대를 거치며 변화하는 과정이다. 찰스 다윈(Charles Darwin)은 이 개념을 체계화한 선구자로, HMS 비글호 항해와 다윈의 핀치 연구는 진화론 발전에 중요한 역할을 했다.

Figure 4.9 다윈의 경로 – HMS 비글호 항해와 진화론 발전

Figure 4.10 다윈의 핀치 – 자연선택과 적응 진화의 대표적 사례

또한 고대 DNA 연구는 진화 연구에 새로운 관점을 제공했다. 특히 스반테 페보(Svante Pääbo)는 네안데르탈인과 데니소바인의 유전체을 성공적으로 해독하여 현대 인류와의 교배 증거를 발견했으며, 이 공로로 2022년 노벨생리의학상을 수상했다.

Figure 4.11 스반테 페보 – 2022년 노벨생리의학상 수상자, 고유전학의 아버지

선택압(Selection Pressure)은 "진화적 압력(Evolutionary Pressure)"이라고도 불리며, 다음과 같은 요인들이 있다: – 환경(기후, 지리 등) – 다른 종과의 관계 – 교배를 촉진하거나 방해하는 기타 요인

진화의 세 가지 기본 메커니즘은 다음과 같다: 1. 무작위 돌연변이(Random Mutation) 2. 자연선택(Natural Selection) 3. 유전적 부동(Genetic Drift)

따라서 진화란 특정한 방향성을 가지지 않는 무작위적 돌연변이와 환경 변화에 기초하는 현상이며, 이러한 무작위적 과정 속에서 각 개체가 처한 환경에 맞는 선택압이 작용하는 방향으로 조금씩 적응해 가는 과정으로 이해할 수 있다.

계통학(Phylogeny)은 종과 생물 그룹이 시간에 따라 어떻게 진화했는지 연구하는 학문이다. 이는 비슷한 생물이 공통 조상을 공유한다는 개념에 기초한다. 계통학은 생물 간의 자연적 관계를 반영하는 분류체계를 만드는 데 사용된다.

Figure 4.12 생명의 나무 – 모든 생물의 진화적 관계를 나타내는 계통수

Figure 4.13 생명의 산호 – 생물 다양성의 또 다른 시각화 접근법

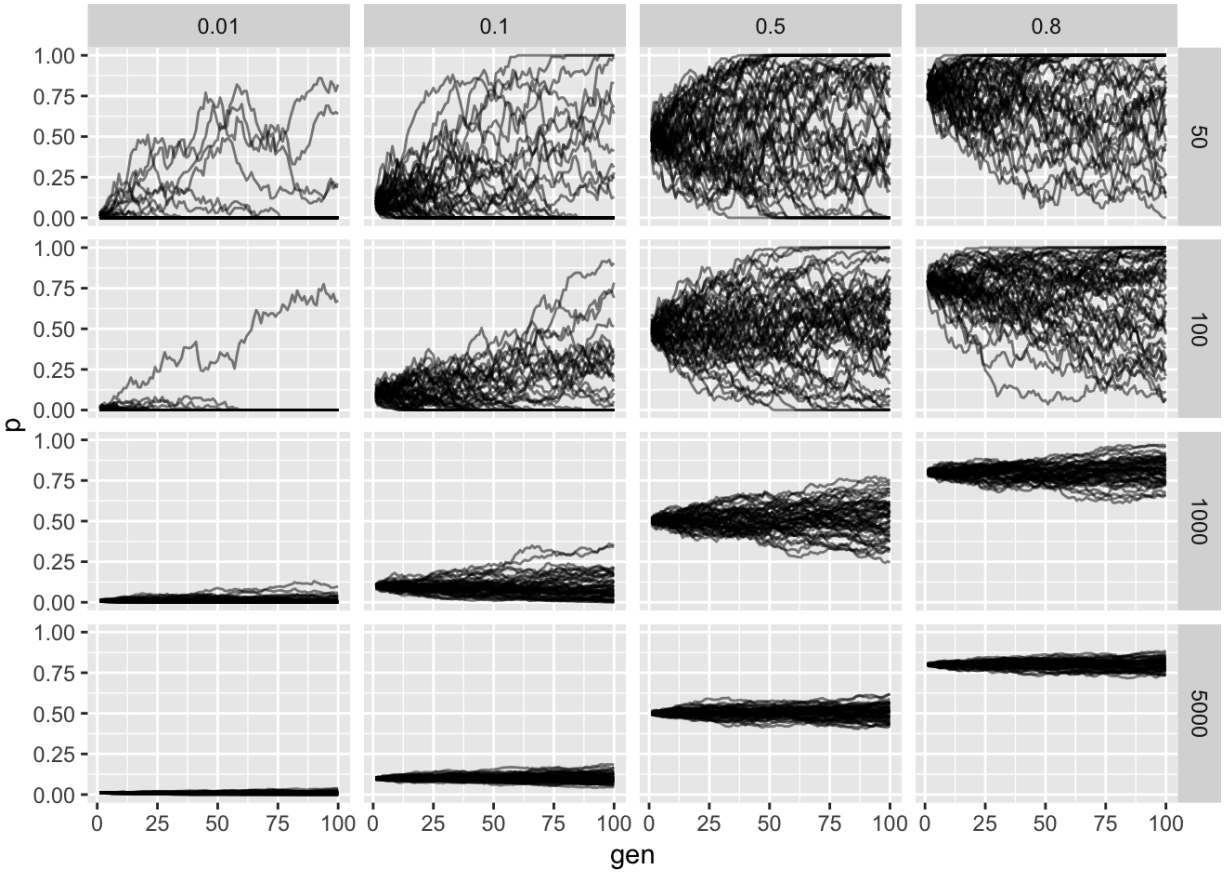


Figure 84: Wright-Fisher 모델 결과

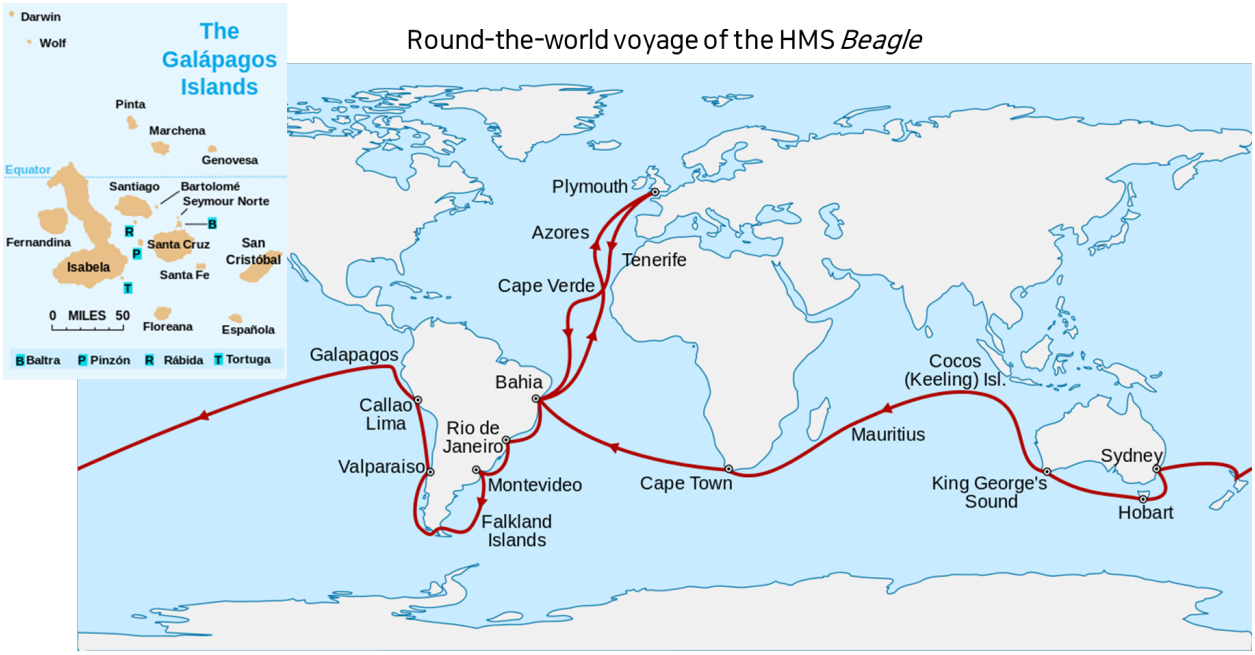


Figure 85: 다윈의 경로

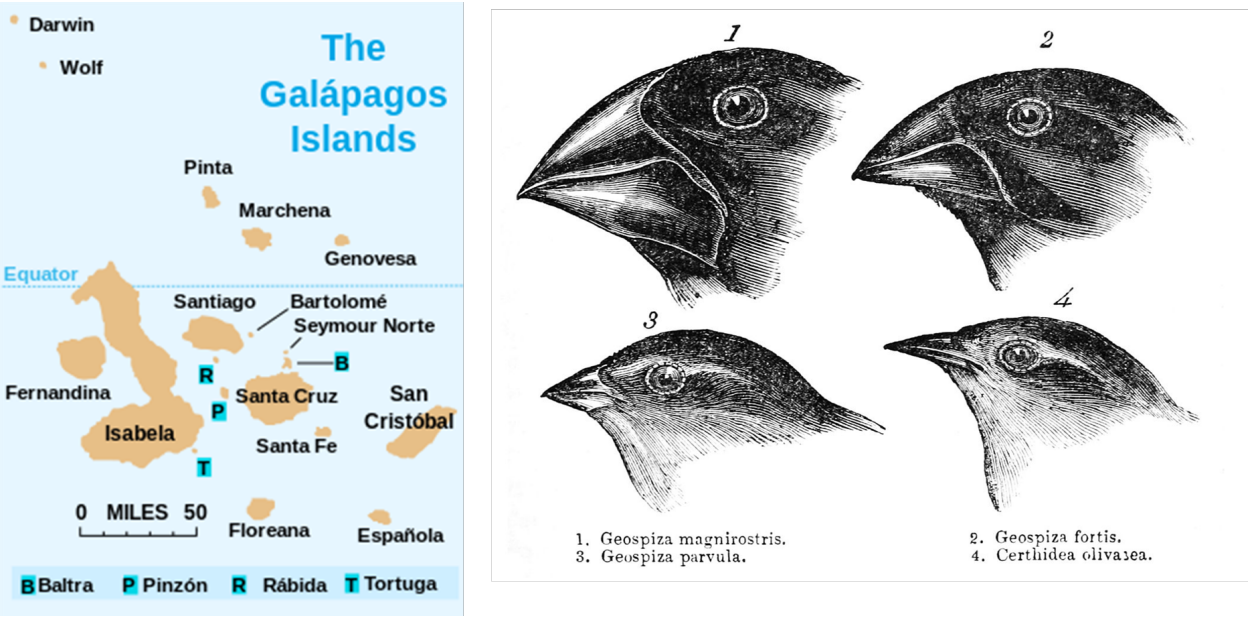


Figure 86: 다윈의 핀치

Figure 87: 스반테 페보

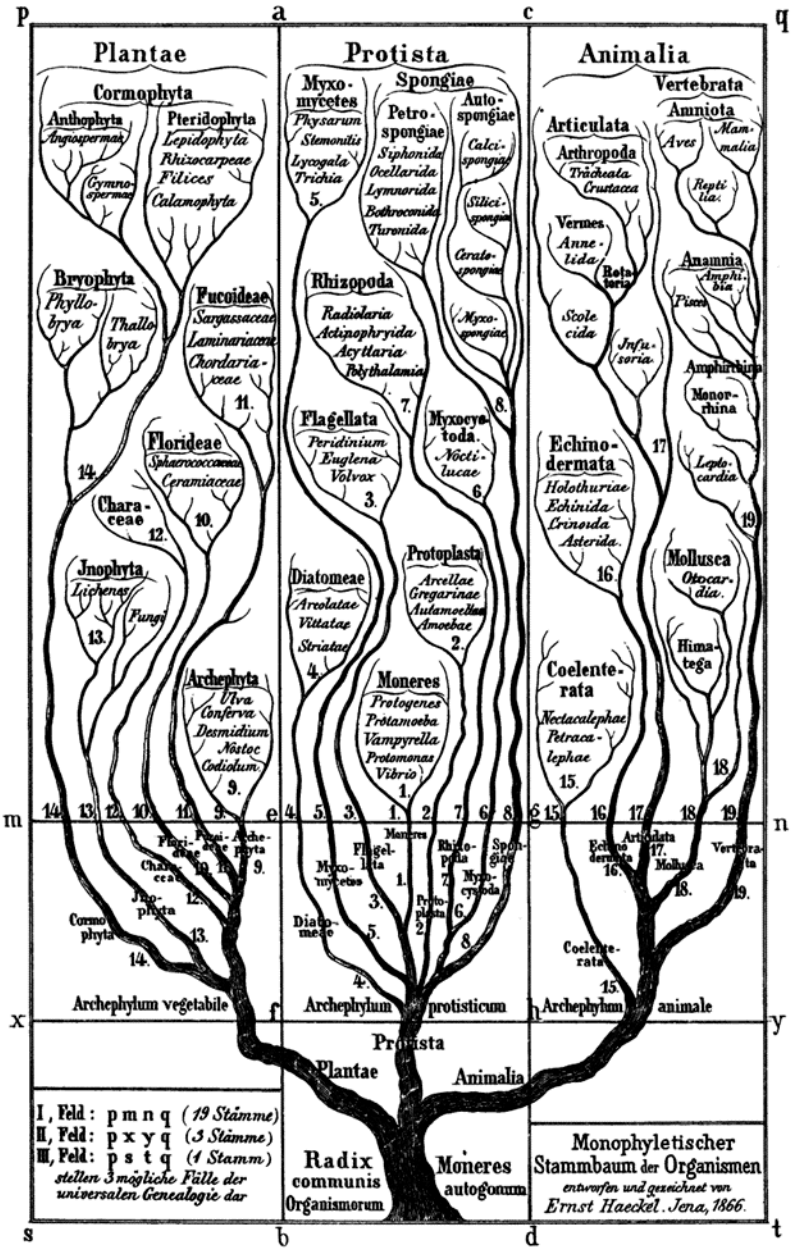


Figure 88: 생명의 나무

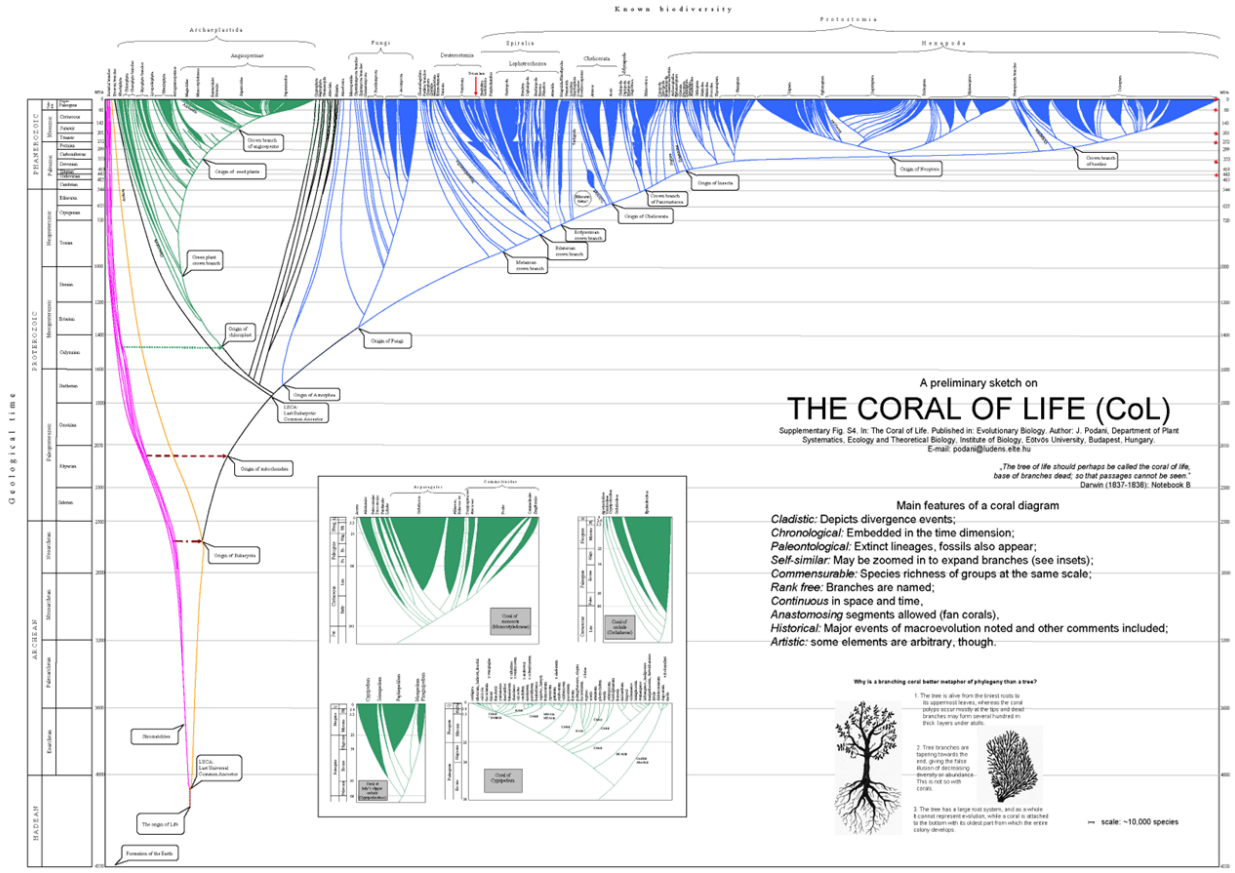


Figure 89: 생명의 산호

5.9 다중서열정렬과 분자 진화

다중서열정렬(Multiple Sequence Alignment, MSA)은 3개 이상의 DNA, RNA 또는 단백질 서열을 동시에 비교하여 진화적 관계를 파악하는 생명정보학의 기본 기법이다. MSA는 병원체의 진화 추적과 계통분류학적 분석에 활용된다.

MSA에서는 보존 영역과 가변 영역을 구분할 수 있다. 보존 영역은 진화 과정에서 변화가 적은 중요한 부위이며, 가변 영역은 종간 또는 개체간 차이가 나타나는 부위이다. 정렬 과정에서 삽입이나 결실 변이는 갭으로 표시되며, 서열 간 유사성을 통해 진화적 거리를 측정할 수 있다.

5.9.1 병원체 진화 추적

바이러스는 빠른 복제와 높은 변이율로 인해 실시간 진화 관찰이 가능한 대표적인 병원체이다. RNA 바이러스는 높은 변이율을 보이며, 숙주 면역계와 백신에 의한 선택압을 받는다. 또한 서로 다른 바이러스 간의 유전자 재조합과 유전적 부동도 바이러스 진화에 중요한 역할을 한다.

MSA 결과를 바탕으로 계통수를 구축하여 바이러스의 진화 관계를 시각화할 수 있다. 이를 통해 바이러스가 언제, 어떻게 진화했는지 추적할 수 있다.

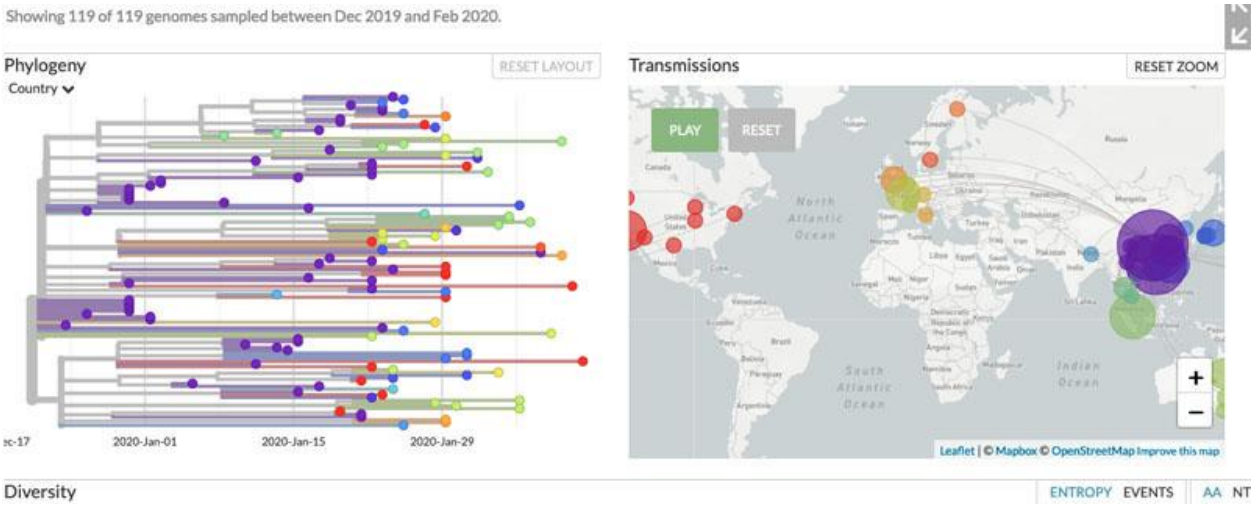


Figure 90: GISAID COVID-19 계통발생 개요

Figure 4.14 GISAID COVID-19 계통발생 개요 – 전 세계 SARS-CoV-2 변이 추적

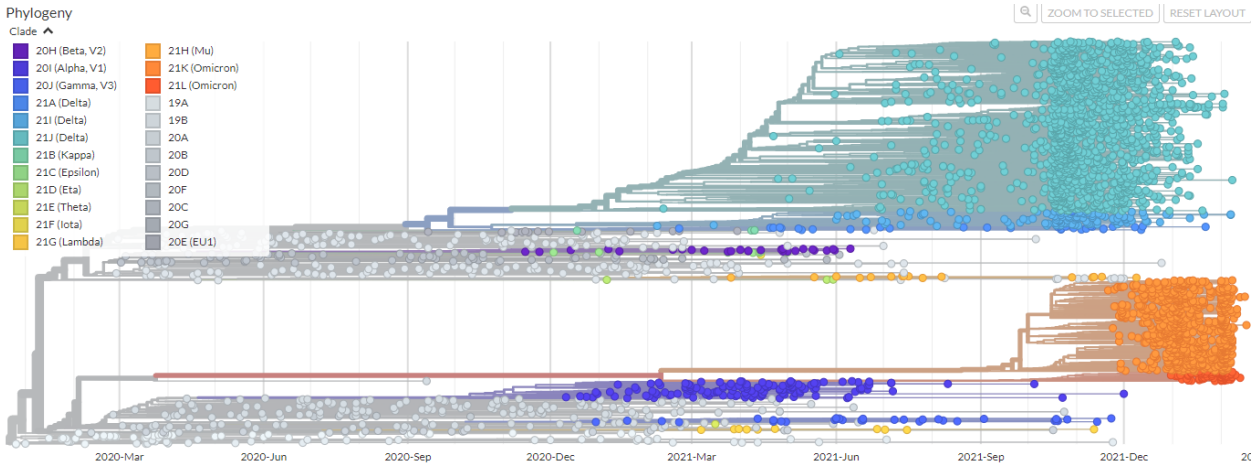


Figure 91: GISAID COVID-19 계통발생 세부

Figure 4.15 GISAID COVID-19 계통발생 세부 - 다양한 변이주의 출현과 전파

5.9.2 COVID-19 사례를 통한 감염병 전파 추적

코로나19 팬데믹은 MSA와 계통발생학적 분석의 실질적 중요성을 보여준 대표적 사례이다. SARS-CoV-2 바이러스의 유전체 분석을 통해 다양한 변이주의 출현과 전파 과정을 실시간으로 추적할 수 있었다. 이러한 성과는 전 세계 의생명정보학자들이 GISAID(Global Initiative on Sharing All Influenza Data)와 같은 플랫폼을 통해 바이러스 서열 데이터를 신속하게 공유하고 분석한 결과이다.

MSA를 통한 실시간 바이러스 감시는 새로운 변이주의 출현을 조기에 탐지하고, 바이러스의 전파 경로를 추적할 수 있게 한다. 특히 팬데믹 기간 동안 의생명정보학자들의 헌신적인 노력과 국제적 협력을 통해 바이러스 진화를 실시간으로 모니터링할 수 있었으며, 이러한 정보는 백신 개발, 치료제 연구, 방역 정책 수립에 중요한 근거를 제공했다.

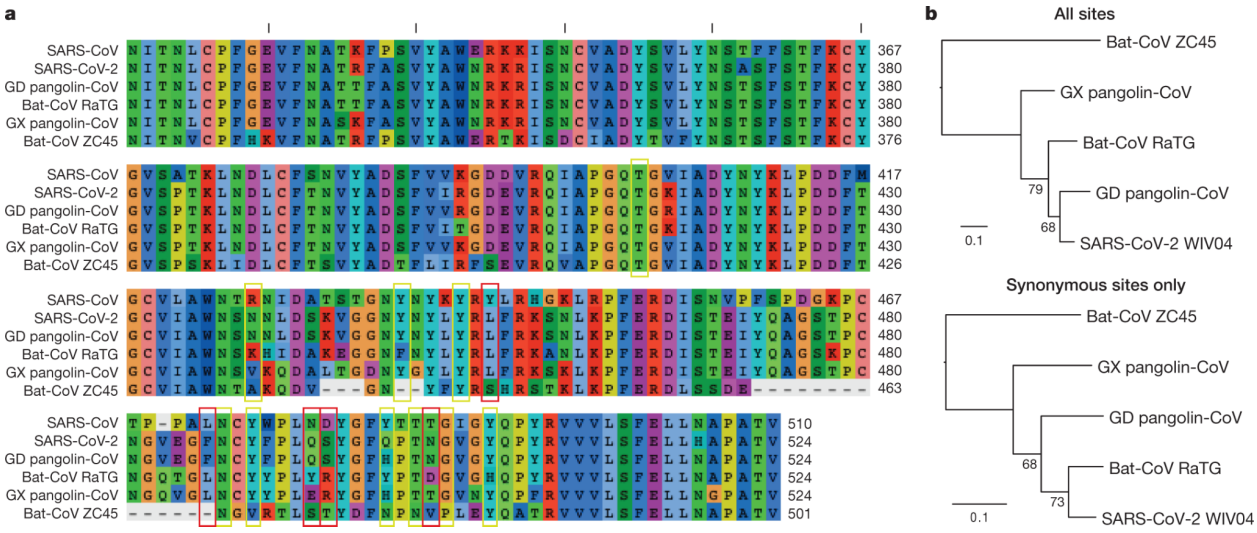


Figure 92: 다중샘플 정렬 COVID-19 계통수

Figure 4.16 다중샘플 정렬 COVID-19 계통수 - MSA를 통한 바이러스 진화 분석

5.10 고유전학

고유전학(Paleogenetics)은 고대 샘플에서 DNA를 시퀀싱하여 과거 생물의 유전 정보를 분석하는 연구 분야이다. 화석이나 고고학적 유물에서 추출한 DNA를 통해 멸종된 종이나 과거 생물집단의 유전적 특성을 연구할 수 있다.

스반테 페보(Svante Päabo)는 이 분야의 선구자로 2022년 노벨생리의학상을 수상했다. 스반테 페보와 그의 팀은 네안데르탈인의 완전한 유전체 시퀀싱과 데니소바인의 발견을 통해 인류 진화 연구에 혁신적인 기여를 했으며, 특히 현생인류와 네안데르탈인 사이의 유전자 교류를 밝혀내어 인류 진화의 복잡성을 보여주었다.

Figure 4.17 데니소바인 - 고유전학으로 발견된 새로운 인류 종

5.10.1 고유전학의 기술적 도전

고대 DNA 연구는 여러 기술적 어려움을 극복해야 한다. 시간이 지나면서 DNA는 분해되고 단편화되어 매우 짧은 조각들만 남게 된다. 또한 현대 DNA의 오염 가능성이 높아 정확한 분석을 위해 엄격한 실험 조건과 검증 절차가 필요하다. 이러한 문제들을 해결하기 위해 차세대 시퀀싱 기술과 특수한 DNA 추출 및 증폭 기법이 개발되었다.

5.10.2 고유전학의 응용 분야

고유전학은 인류학, 고고학, 진화생물학 등 다양한 분야에 응용된다. 고대 인류의 이주 경로와 집단 구조를 밝히고, 농작물과 가축의 기원과 진화 과정을 추적하며, 멸종 동물의 생태와 진화 역사를 연구하는 데 활용된다. 또한 기후 변화가 과거 생물집단에 미친 영향을 분석하여 현재의 생태계 변화를 이해하는 데에도 기여한다.

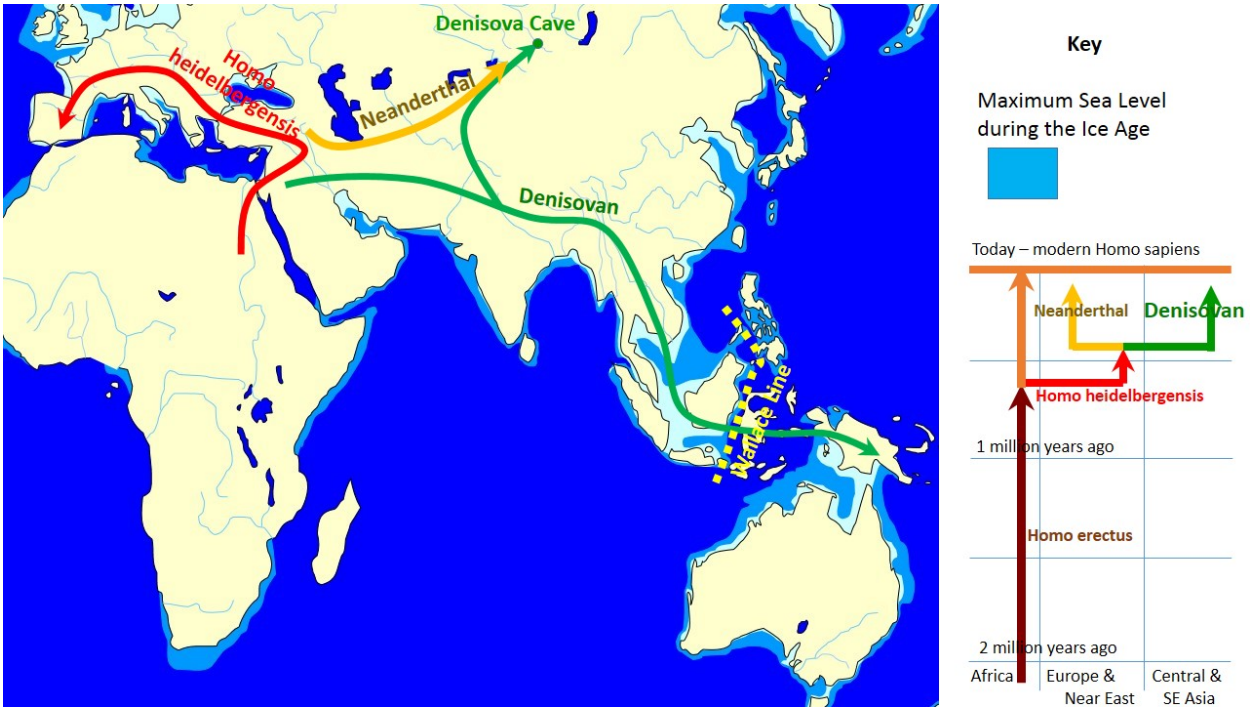


Figure 93: 데니소바인

이러한 연구들은 인류의 진화 역사와 다른 종과의 관계를 더 정확하게 이해할 수 있게 해준다. 최신 유전체 분석 기술의 발전으로 고유전학 연구는 더욱 정밀해지고 있으며, 생물의 다양성과 적응 메커니즘에 대한 새로운 발견을 계속해서 제공하고 있다.

5.11 결론

변이와 진화는 생명체의 다양성을 만들어내는 핵심 메커니즘이며, 의생명정보학은 이러한 복잡한 과정을 이해하는 데 필수적인 도구를 제공한다. 변이 찾기 기술을 통해 개인의 유전적 특성을 파악하고, VCF 형식으로 표준화된 변이 정보를 효율적으로 관리할 수 있다. Phasing과 오염 검출 같은 정교한 분석 방법은 변이 호출의 정확성을 높여준다.

Hardy-Weinberg 평형과 유전적 부동의 개념은 집단 내 유전자 빈도 변화를 설명하며, 이는 진화의 기본 원리를 이해하는 바탕이 된다. 진화의 세 가지 기본 메커니즘인 무작위 돌연변이, 자연선택, 유전적 부동은 생물의 적응과 다양성을 만들어내는 핵심 요소이다.

다중서열정렬과 계통발생학적 분석은 병원체의 진화를 실시간으로 추적할 수 있게 해준다. 특히 COVID-19 팬데믹에서 보여준 것처럼, 전 세계 의생명정보학자들의 협력을 통해 바이러스 변이를 신속하게 분석하고 공중보건 정책에 활용할 수 있었다. 고유전학 분야는 과거와 현재를 연결하여 인류와 다른 생물종의 진화 역사를 밝혀내고 있다.

이러한 의생명정보학 기술들은 단순히 과거를 이해하는 것을 넘어서 미래의 질병 예방, 치료법 개발, 생태계 보전에 중요한 기여를 하고 있다. 변이와 진화에 대한 이해는 정밀의학, 감염병 대응, 생물다양성 보전 등 현대 사회가 직면한 다양한 문제 해결의 열쇠가 되고 있다.

6 유전질환과 GWAS

6.1 유전질환

유전질환은 유전적 변이에 의해 발생하는 질병이다. 이러한 변이는 인구 집단 내에서 유전되는 경우가 많다. 유전질환은 크게 단일 돌연변이에 의한 멘델 질환과 여러 유전자의 복합적 작용에 의한 다유전자 질환으로 나눌 수 있다.

유전질환은 일반적으로 DNA 서열의 변화, 염색체 구조 이상, 또는 유전자 발현 조절 메커니즘의 문제로 인해 발생한다. 이러한 유전적 이상은 부모로부터 자녀에게 전달되거나 새롭게 발생할 수 있다. 유전질환의 심각도는 경미한 증상부터 생명을 위협하는 수준까지 다양하게 나타난다.

6.2 유전질환의 유형

6.2.1 멘델 질환(Mendelian disorders)

멘델 질환은 단일 돌연변이에 의해 발생하는 질병이다. 이러한 질환은 대개 희귀하지만, 발생 시 그 영향은 심각할 수 있다. 멘델 유전 방식을 따르기 때문에 가족 내 발생 패턴이 예측 가능하다.

멘델 질환의 몇 가지 예시는 다음과 같다:

1. 혈우병(Hemophilia): F8, F9, F11 유전자 중 하나의 돌연변이로 인해 발생하는 혈액 응고 장애 질환이다. 주로 X 연관 열성 유전을 보인다.
2. 판코니 빈혈(Fanconi Anemia): Fancm, Fancd 등의 유전자 돌연변이로 발생하는 골수 기능 부전 증후군이다. 상염색체 열성 유전을 보인다.
3. 낭포성 섬유증(Cystic Fibrosis): CFTR 유전자의 돌연변이로 인해 발생하며, 폐와 소화기계에 영향을 미치는 질환이다.
4. 헌팅턴병(Huntington's Disease): HTT 유전자의 돌연변이로 인한 신경퇴행성 질환으로, 상염색체 우성 유전을 보인다.
5. 겸상적혈구 빈혈(Sickle Cell Anemia): HBB 유전자의 돌연변이로 인해 발생하는 혈액 질환이다.

멘델 질환은 유전 방식에 따라 상염색체 우성, 상염색체 열성, X 연관 우성, X 연관 열성 등으로 분류할 수 있다. 이러한 질환은 단일 유전자의 문제로 발생하기 때문에, 특정 유전자 검사를 통해 진단이 가능하다.

6.2.2 다유전자 질환(Polygenic disorders)

다유전자 질환은 다수의 유전변이와 환경적 요인의 복합적 상호작용으로 발생한다. 이러한 질환은 멘델 질환보다 훨씬 흔하며, 현대 사회에서 가장 많이 발생하는 질병 대부분이 이 범주에 속한다.

다유전자 질환의 몇 가지 예시는 다음과 같다:

1. 치매(Dementia): 알츠하이머병 등 여러 유형의 치매는 다양한 유전적 요인과 환경적 요인의 영향을 받는다.
2. 고혈압(Hypertension): 혈압 조절에 관여하는 여러 유전자들의 변이와 생활 습관 요인이 복합적으로 작용한다.
3. 당뇨병(Diabetes): 특히 제2형 당뇨병은 여러 유전자와 생활 습관의 상호작용으로 발생한다.
4. 심혈관 질환(Cardiovascular Disease): 심장병, 뇌졸중 등은 다양한 유전적 요인과 환경적 요인에 의해 영향을 받는다.
5. 자가면역 질환(Autoimmune Diseases): 류마티스 관절염, 루푸스, 다발성 경화증 등은 여러 유전자와 환경적 요인의 복합 작용으로 발생한다.

다유전자 질환은 가족력을 보이지만, 멘델 유전처럼 명확한 패턴을 따르지 않는다. 이러한 질환의 발생 위험은 여러 유전자 변이의 조합과 환경적 요인에 의해 결정된다.

6.3 다유전자 형질(Polygenic Traits)

다유전자 형질은 여러 유전변이가 함께 작용하여 결정되는 특성을 말한다. 이는 많은 변이들이 각각 작은 효과를 발휘하여 전체적인 표현형을 만들어내며, 이러한 형질은 연속적인 분포를 보인다. 키, 체중, 피부색, 지능, 혈압 등이 대표적인 예이다.

예를 들어, 키의 경우 단순히 "키가 크다" 또는 "키가 작다"로 구분되지 않고 다양한 높이로 연속적으로 분포한다. 이는 키에 영향을 미치는 수백 개의 유전변이들이 각각 작은 효과를 발휘하기 때문이다. 또한 여기에는 환경적 요인(영양 상태, 운동 등)도 함께 작용하여 최종적인 키를 결정한다.

이때 중요한 질문이 제기된다: "각 효과 대립유전자(effect allele)의 효과 크기를 어떻게 알 수 있는가?" 즉, 특정 변이가 키를 몇 센티미터 증가시키는지, 혈압을 얼마나 상승시키는지 정량적으로 측정할 수 있을까?

이는 단순하지 않은 문제이다. 유전적 효과와 환경적 요인의 영향을 구분해야 하고, 수많은 변이들의 복합적 작용에서 개별 변이의 기여도를 분리해내야 하기 때문이다. 이러한 복잡한 문제를 해결하기 위해 대규모 집단을 대상으로 하는 정교한 통계적 기법이 필요하다. 이에 대한 답을 찾기 위해 전장유전체 연관분석(GWAS)이 개발되었다.

6.4 전장유전체 연관분석(Genome-wide Association Study, GWAS)

GWAS는 특정 형질이나 질병에 영향을 미치는 SNP(단일염기다형성)의 효과 크기를 찾아내는 강력한 통계적 도구이다. 대규모 인구 집단에서 수백만 개의 SNP를 동시에 분석하여 유전적 변이와 질병 또는 형질 간의 연관성을 체계적으로 탐색한다.

Figure 5.1 GWAS 개요 – 전장유전체 연관분석의 기본 개념과 절차

6.4.1 GWAS 설계 예시

GWAS의 기본 원리를 이해하기 위해 혈압에 관한 간단한 예시를 살펴보자. 어떤 특정 SNP가 혈압에 영향을 미치는지 알아보고 싶다고 가정해보자.

연구 설계의 주요 요소들은 다음과 같다: - **샘플**: 다양한 혈압을 가진 혼합 인구 집단 100명 - **형질**: 수축기 혈압이 표준값(120mmHg)에서 얼마나 차이나는지 - **분석 방법**: 특정 SNP에 집중한 단변량 분석

분석 과정에서는 전체 집단을 해당 SNP의 효과 대립유전자를 몇 개 가지고 있는지에 따라 세 그룹으로 나눈다: 0개(AA), 1개(Aa), 2개(aa) 보유자. 가산적 모델(additive model)을 가정하면, 효과 대립유전자를 한 개 더 가질 때마다 혈압이 일정하게 증가한다고 본다. 이를 통해 변이의 수와 혈압 변화 간의 선형 관계를 분석할 수 있다.

선형 회귀 분석을 통해 기울기(효과 크기)와 그 표준오차를 구한 후, T-통계량을 계산한다: T = 기울기 / 표준오차. 이 T-통계량은 자유도가 (n-2)인 t-분포를 따르며, 이를 바탕으로 P-값을 계산한다.

P-값은 t-분포에서 관찰된 T-값보다 극단적인 값(양쪽 꼬리 부분)이 나타날 확률의 합이다. 즉, P = P(|T| 〉 |관찰된 T-값|)으로 계산된다. 만약 계산된 T-값이 2.5라면, P-값은 t-분포에서 T 〈 -2.5 또는 T 〉 2.5인 영역의 면적을 모두 더한 값이다. P-값이 낮을수록 해당 SNP가 형질에 진짜 영향을 미칠 가능성이 높다는 의미이다.

Figure 5.2 T값을 이용한 선형 적합 – 회귀 분석에서 T-통계량 계산 과정

Figure 5.3 T분포 P값 계산 – 통계적 유의성 평가를 위한 P값 계산 과정

GWAS에서 P-값을 로그 변환하는 이유는 여러 가지가 있다. 첫째, P-값은 0과 1 사이의 값으로 매우 작은 값들(예: 10^{-8})이 자주 나타나는데, 이를 그대로 시각화하면 차이를 구별하기 어렵다. $-\log_{10}$(P-값)으로 변환하면 작은 P-값일수록 더 큰 값이 되어 시각적으로 구분하기 쉬워진다. 둘째, 맨해튼 플롯에서 유의수준 5×10^{-8}은 $-\log_{10}(5\times10^{-8}) \approx 7.3$이 되어 명확한 기준선을 제공한다. 셋째, 로그 변환을 통해 넓은 범위의 P-값을 효과적으로 표현할 수 있어 전체 유전체에서의 연관성 패턴을 한눈에 파악할 수 있다.

Figure 5.4 맨해튼 플롯 예시 – GWAS 결과를 시각화한 대표적인 맨해튼 플롯

Figure 5.5 P값 로그 변환 – GWAS에서 P값을 로그 변환하는 이유와 효과

가산적 모델 외에도 다른 유전 모델들이 사용될 수 있다. 우성 모델(dominant model)에서는 효과 대립유전자를 1개 이상 가지면 동일한 효과를 보인다고 가정하고, 열성 모델(recessive model)에서는 효과 대립유전자를 2개 가져야만 효과가 나타난다고 가정한다. 실제 GWAS에서는 주로 가산적 모델이 사용되지만, 연구 목적에 따라 다른 모델들도 고려될 수 있다.

Figure 5.6 GWAS 모델 – 가산적, 우성, 열성 모델의 비교

6.4.2 P-값의 의미

GWAS에서 P-값은 관찰된 연관성이 단순히 우연의 산물일 확률을 나타낸다. 즉, P-값이 낮을수록 우리가 발견한 연관성이 진짜일 가능성이 높다는 의미이다.

P-값의 크기를 결정하는 주요 요인들은 다음과 같다:

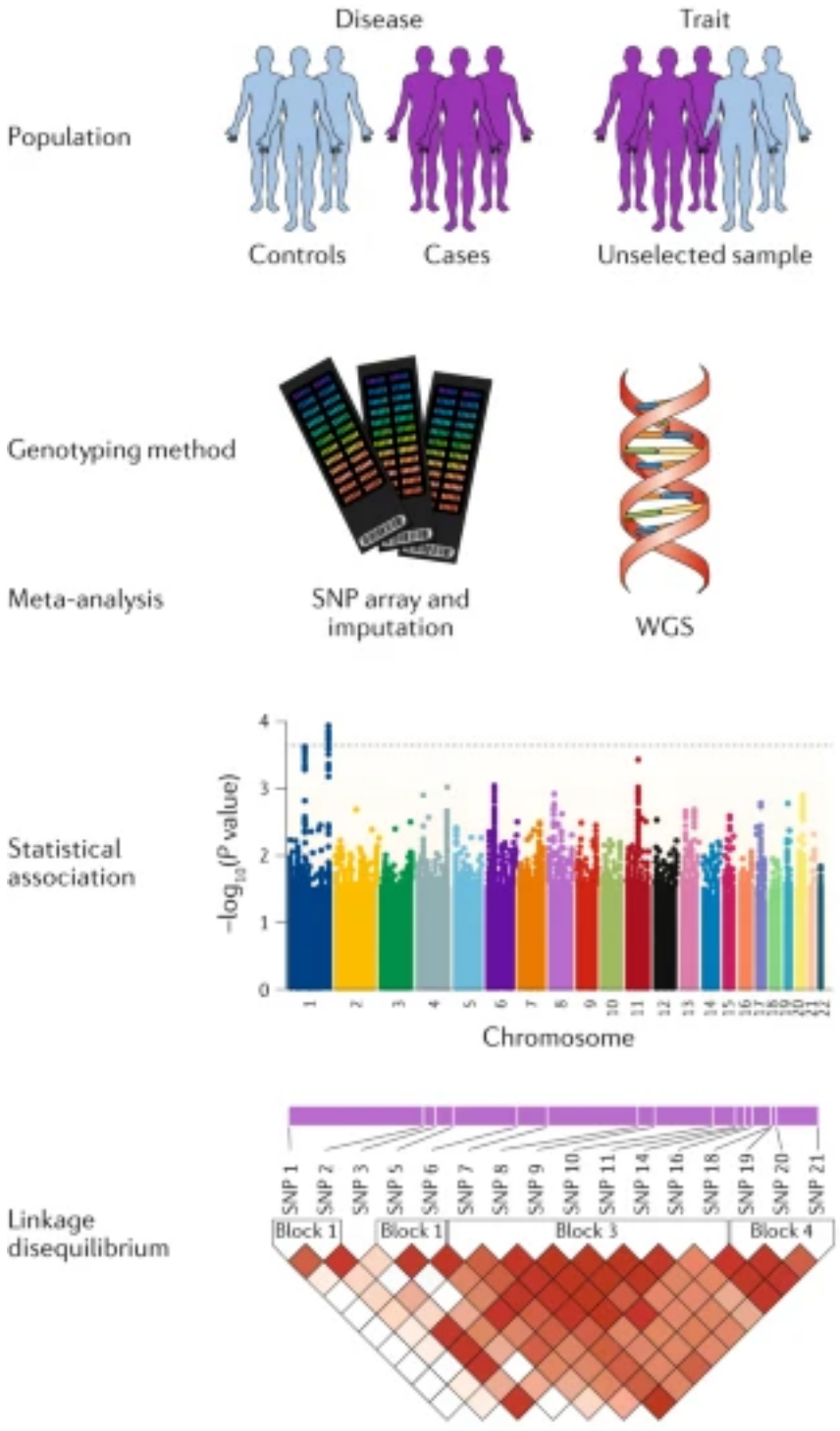


Figure 94: GWAS 개요

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\text{s. e.}(\hat{\beta})},$$

Figure 95: T값을 이용한 선형 적합

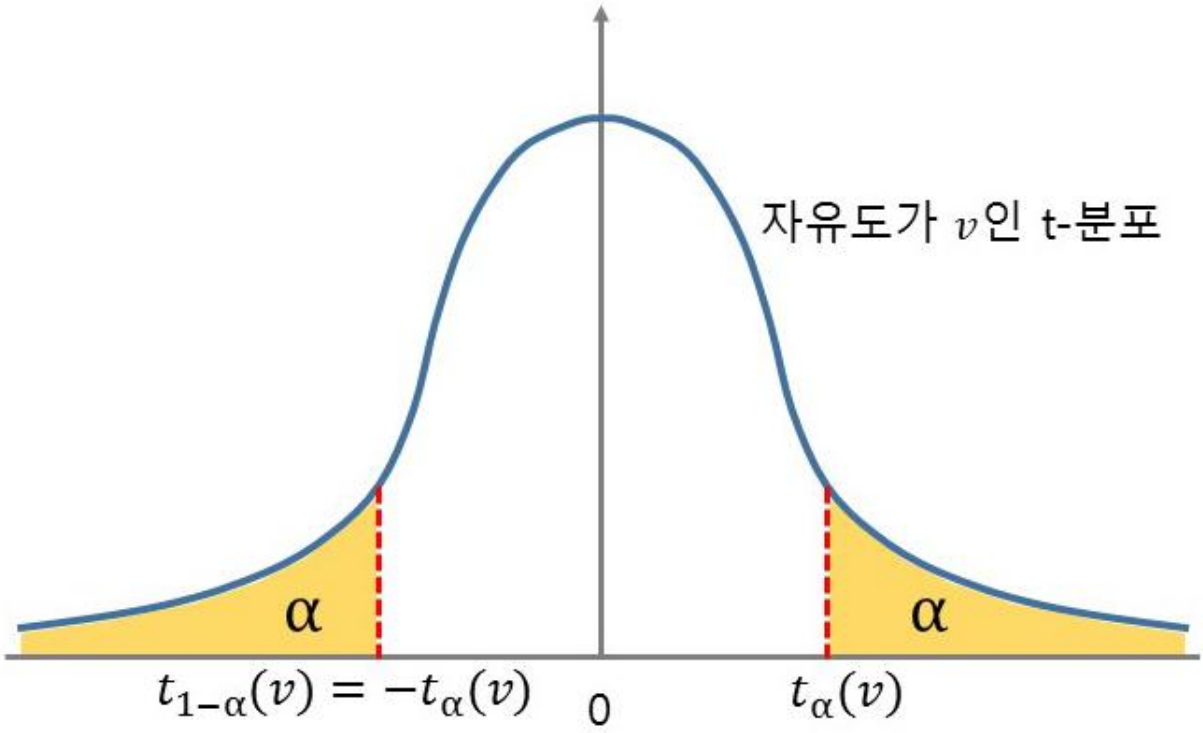


Figure 96: T분포 P값 계산

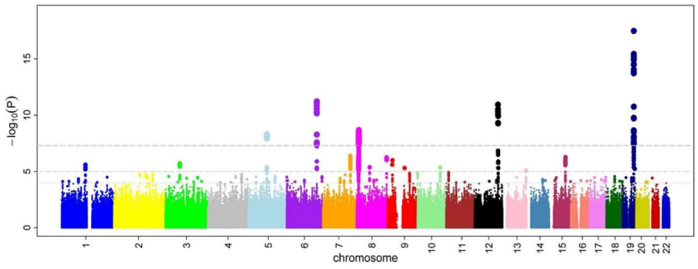


Figure 97: 맨해튼 플롯 예시

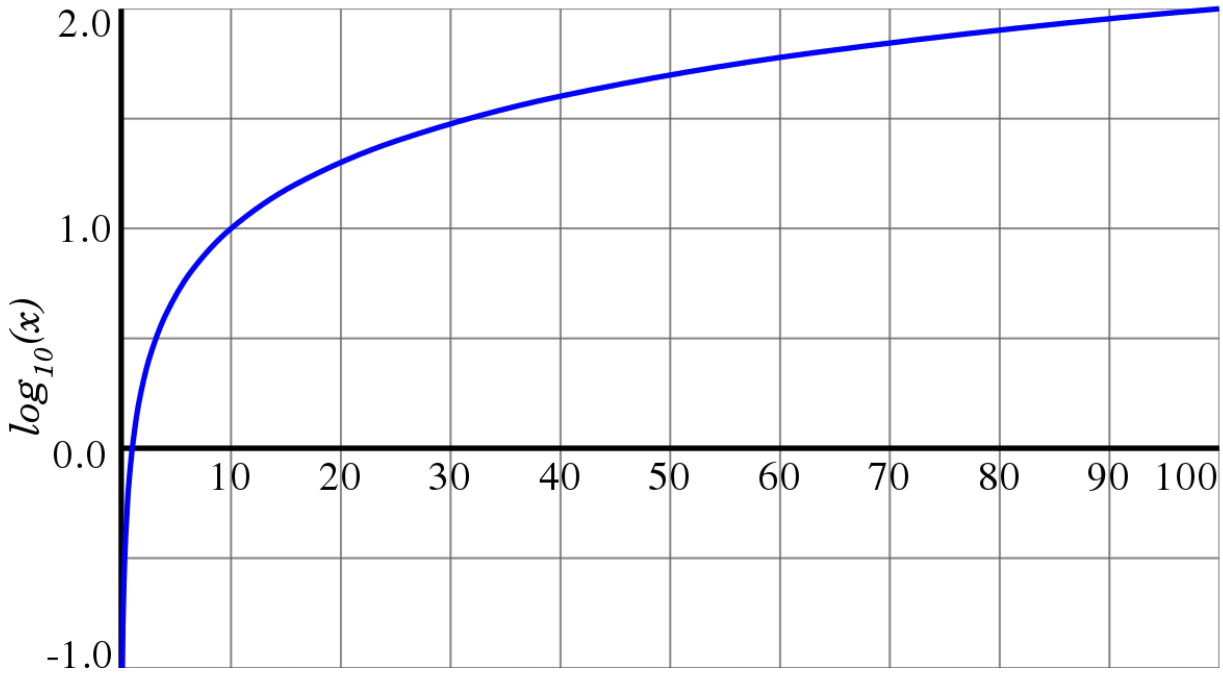


Figure 98: P값 로그 변환

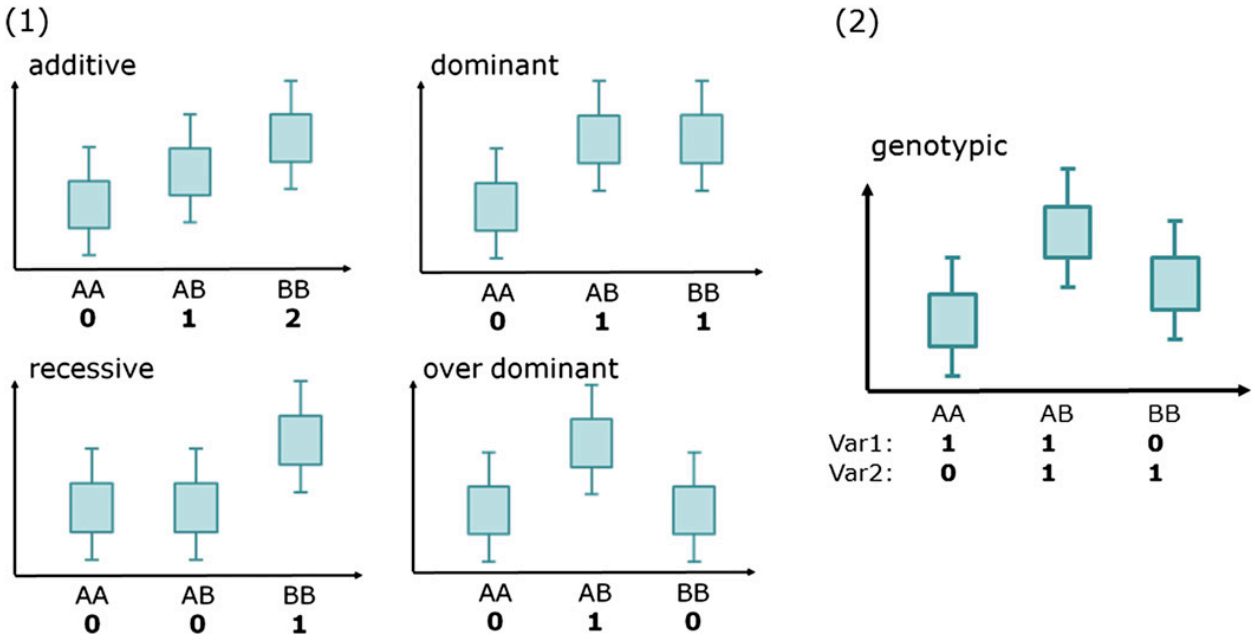


Figure 99: GWAS 모델

- **효과 크기(기울기)**: SNP가 형질에 미치는 영향이 클수록 T-통계량이 높아지고 P-값은 낮아진다
- **측정 오차**: 데이터의 변동성이 클수록 T-통계량은 낮아지고 P-값은 높아진다
- **샘플 크기**: 더 많은 샘플을 사용할수록 통계적 검정력이 높아진다

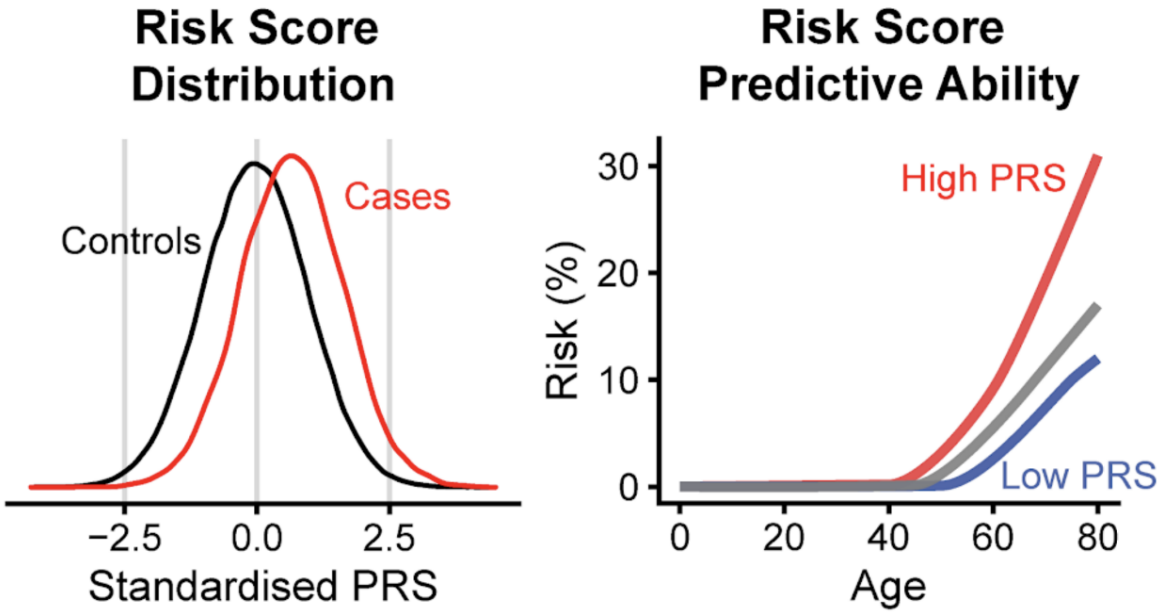


Figure 100: 다유전자 위험 점수 개념

Figure 5.7 다유전자 위험 점수 개념 – 여러 변이의 종합적 효과 평가

예를 들어, 어떤 SNP는 혈압에 큰 영향을 미치지만 데이터의 변동이 작아 낮은 P-값을 보이고, 다른 SNP는 비슷한 효과를 가지지만 데이터의 변동이 커서 상대적으로 높은 P-값을 나타낼 수 있다.

Figure 5.8 GWAS 유의한 SNV – 통계적으로 유의한 연관성을 보이는 SNP 예시

Figure 5.9 GWAS 비유의한 SNV – 통계적으로 유의하지 않은 연관성을 보이는 SNP 예시

6.4.3 맨해튼 플롯(Manhattan plot)

실제 GWAS에서는 단일 SNP가 아니라 수백만 개의 SNP를 동시에 분석한다. 각 SNP에 대해 위와 같은 통계 분석을 수행하면 수백만 개의 P-값을 얻게 된다. 이러한 방대한 데이터를 효과적으로 시각화하기 위해 맨해튼 플롯이 사용된다.

Figure 5.10 맨해튼 플롯 – 전체 유전체에서 SNP들의 연관성을 시각화한 맨해튼 플롯

맨해튼 플롯의 구성 요소들은 다음과 같다: – **X축**: 염색체 별로 정렬된 전체 유전체 상의 SNP 위치 – **Y축**: -log10(P-값)으로, 값이 클수록 더 유의미한 연관성을 의미 – **피크**: 형질과 강한 연관성을 보이는 SNP들이 높은 타워처럼 나타남

이 플롯을 '맨해튼 플롯'이라고 부르는 이유는 수많은 지점들이 만들어내는 높낮이가 마치 맨해튼의 빌딩 스카이라인을 연상시키기 때문이다. 연관성이 강한 지역은 마치 고층 빌딩처럼 높이 솟아오르고, 연관성이 약한 지역은 낮은 기준선을 이룬다.

6.5 연관 불균형(Linkage Disequilibrium, LD)

GWAS를 수행할 때 직면하는 실질적인 문제 중 하나는 분석해야 할 SNP의 개수가 너무 많다는 것이다. 인간 유전체에는 수백만 개의 SNP가 존재하며, 모든 SNP에 대해 개별적으로 통계 분석을 수행하는 것은 엄청난 계산 비용과 시간을 요구한다. 하지만 유전학자들은 유전체 연구를 통해 SNP들이 독립적으로 존재하지 않는다는 중요한 사실을 발견했다. 즉, 특정 SNP들이 서로 연관되어 함께 유전되는 패턴이 있다는 것이다. 이러한 패턴들을 LD 블록이라고 부른다.

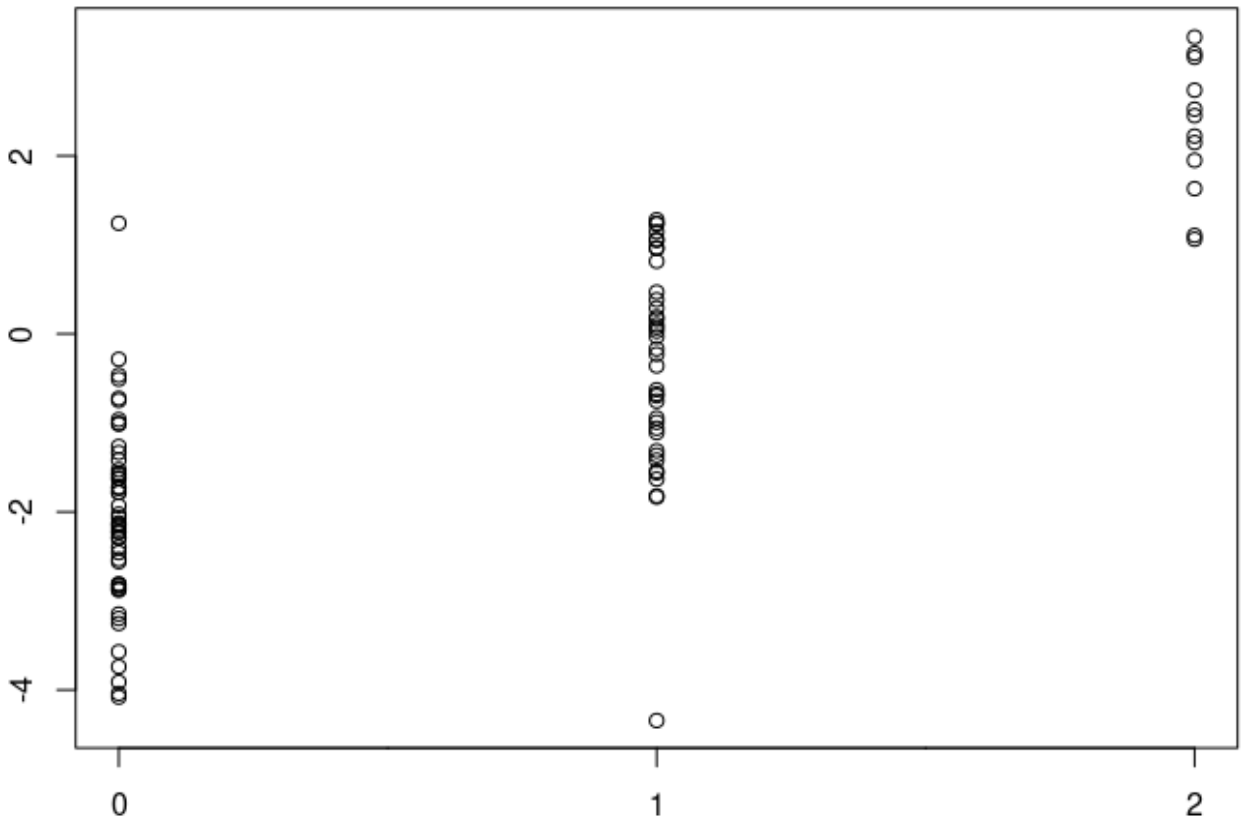


Figure 101: GWAS 유의한 SNV

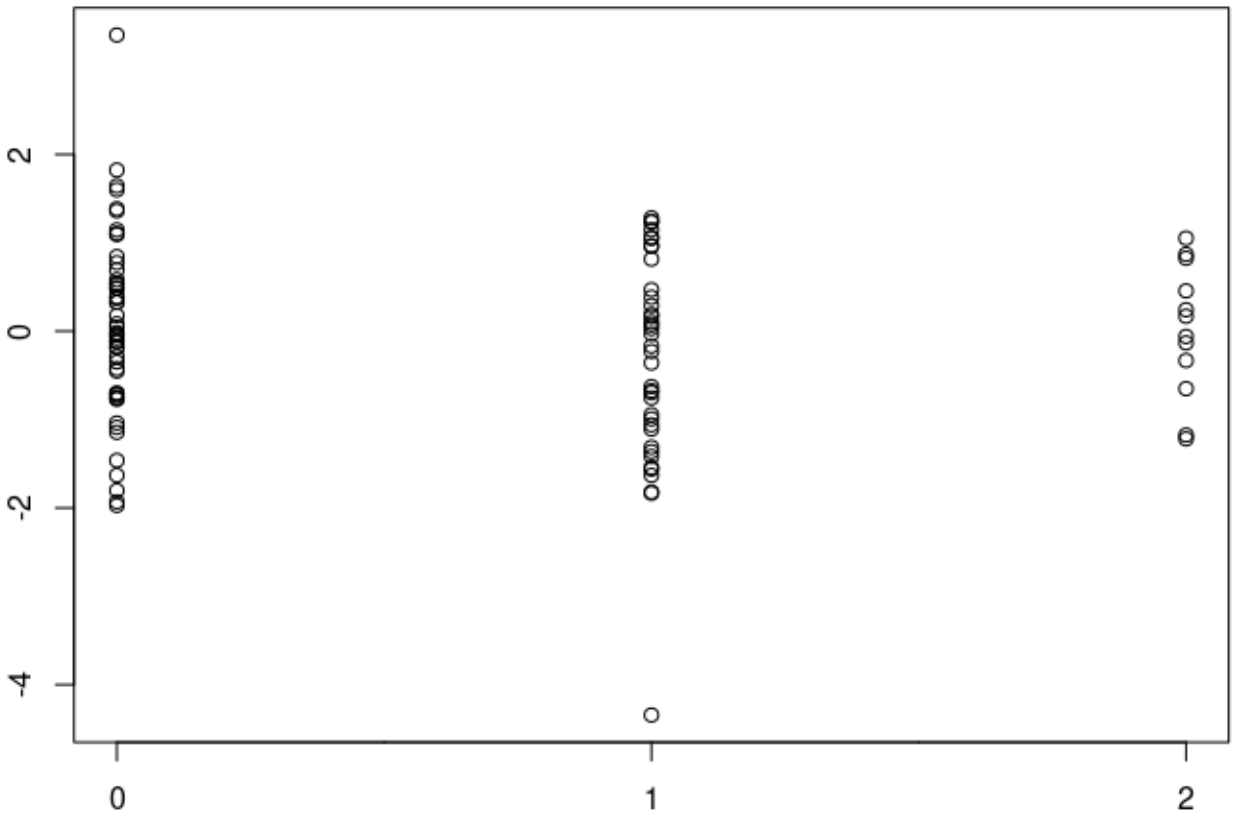


Figure 102: GWAS 비유의한 SNV

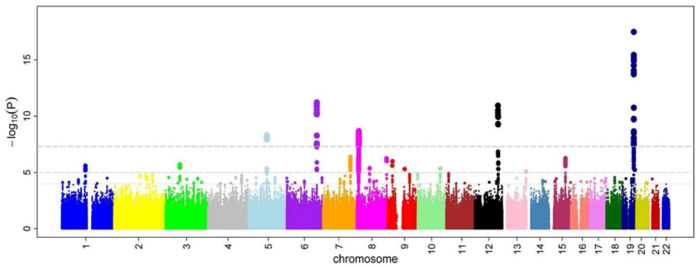


Figure 103: 맨해튼 플롯

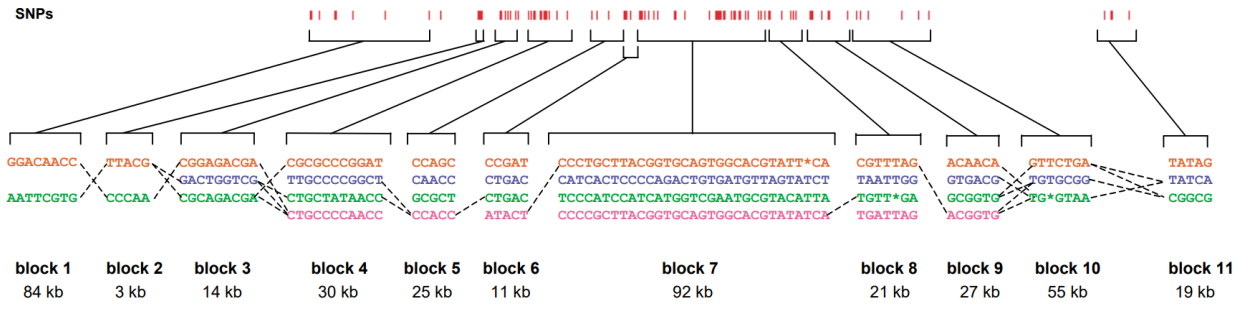


Figure 104: 연관 불균형 블록

Figure 5.11 연관 불균형 블록 – 함께 유전되는 SNP들의 블록 구조

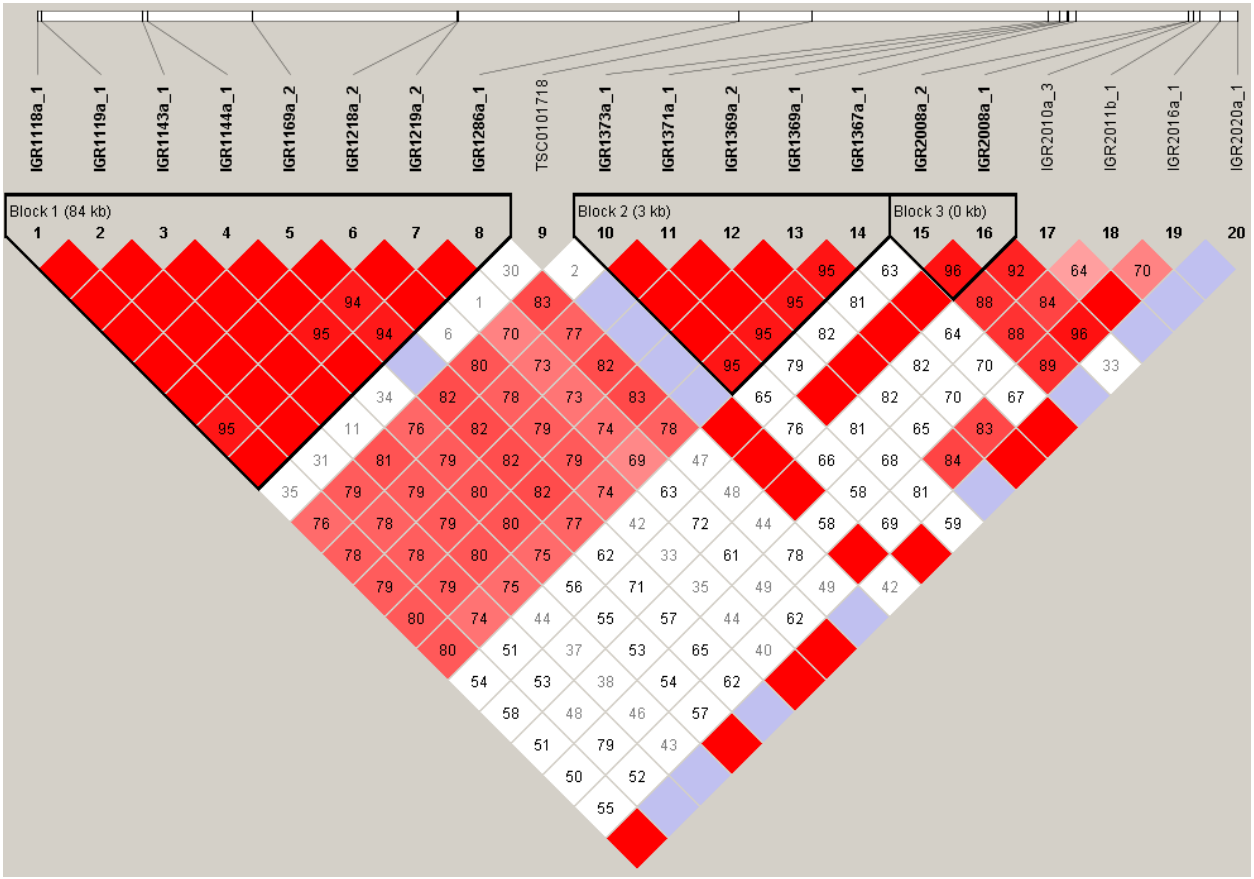


Figure 105: 연관 불균형 구조

Figure 5.12 연관 불균형 구조 – SNP 간의 연관 불균형 패턴과 상관관계

6.5.1 해플로타입(Haplotype)

유전체 상에서 서로 인접한 SNP들은 독립적으로 유전되지 않는다. 대신 특정 SNP들이 함께 묶여서 하나의 단위로 유전되는 경향을 보인다. 이러한 함께 유전되는 변이들의 그룹을 해플로타입(Haplotype)이라고 한다.

예를 들어, 염색체의 특정 구간에서 SNP A, B, C가 있다고 가정해보자. 만약 이 세 SNP가 해플로타입을 형성한다면, A에서 특정 변이를 가진 사람은 B와 C에서도 특정한 변이 패턴을 가질 가능성이 높다. 이는 진화 과정에서 재조합이 일어나지 않은 구간에서 나타나는 현상이다.

중요한 점은 해플로타입 패턴이 인구 집단마다 다르게 나타난다는 것이다. 동아시아인, 유럽인, 아프리카인 등 각 인구 집단은 서로 다른 진화 역사를 가지고 있어서 해플로타입 구조가 상이하다. 따라서 GWAS를 수행할 때는 연구 대상 인구 집단에 맞는 해플로타입 정보를 사용해야 정확한 결과를 얻을 수 있다. 이는 개인 맞춤 의학에서 인구 집단별 유전적 차이를 고려해야 하는 이유이기도 하다.

Figure 5.13 한중일 연관 불균형 구조 – 동아시아 인구 집단에서의 연관 불균형 패턴 비교

6.5.2 HapMap 프로젝트와 연관 불균형

이러한 해플로타입 패턴을 체계적으로 분석하기 위해 국제적인 HapMap 프로젝트가 수행되었다. 이 프로젝트는 다양한 인구 집단에서 SNP들 간의 연관 불균형(Linkage Disequilibrium, LD) 패턴을 지도화했다. LD란 두 개 이상의 유전적 변이가 독립적으로 분포하지 않고 함께 나타나는 경향을 의미한다.

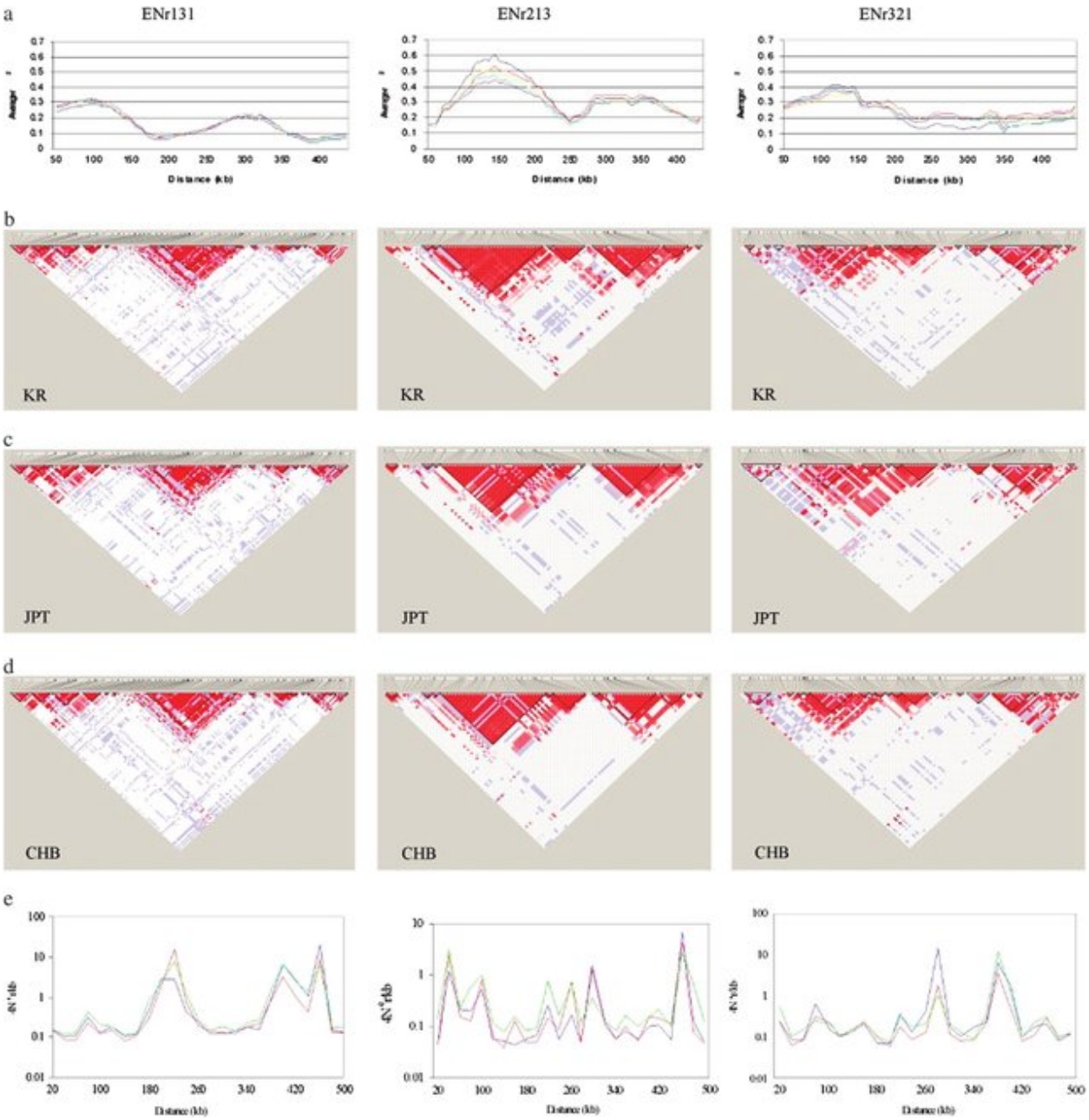


Figure 106: 한중일 연관 불균형 구조

Figure 107: HapMap 프로젝트

Figure 5.14 HapMap 프로젝트 – 인간 해플로타입 지도 작성을 위한 국제 공동 연구

HapMap 프로젝트의 성과 덕분에 우리는 어떤 SNP들이 함께 유전되는지, 즉 어떤 SNP를 분석하면 주변의 다른 SNP들의 정보도 함께 얻을 수 있는지를 알게 되었다. 이를 통해 GWAS의 효율성을 크게 향상시킬 수 있게 되었다.

6.5.3 태그 SNP를 이용한 효율적 분석

LD 정보를 활용하면 모든 SNP를 분석할 필요가 없어진다. 각 LD 블록(함께 유전되는 SNP들의 집합)에서 대표적인 "태그(tag) SNP" 또는 "마커(marker) SNP" 몇 개만 분석하면 해당 블록 전체의 정보를 얻을 수 있다. 이는 분석해야 할 SNP의 수를 수십 분의 일로 줄여주어 계산 비용과 시간을 대폭 절약할 수 있게 해준다.

하지만 이 접근법에는 한계도 있다. 태그 SNP에서 연관성이 발견되더라도, 실제로 형질에 영향을 미치는 진짜 원인 변이(causal variant)가 그 태그 SNP인지, 아니면 같은 LD 블록 내의 다른 SNP인지 구분하기 어렵다는 점이다. 따라서 GWAS 결과를 해석할 때는 이러한 한계를 염두에 두어야 한다.

6.6 다유전자 위험 점수(Polygenic Risk Score, PRS)

GWAS를 통해 질병이나 형질과 연관된 수많은 SNP들을 발견한 후, 이 정보를 어떻게 실용적으로 활용할 수 있을까? 개인의 유전적 위험을 평가하기 위한 도구로 다유전자 위험 점수(Polygenic Risk Score, PRS)가 개발되었다.

6.6.1 PRS의 기본 개념

PRS는 개인이 가지고 있는 여러 위험 변이들의 효과를 종합하여 하나의 점수로 나타낸 지표이다.

			Fred		Alice		Greg	
Genetic variant	*Effect allele*	*Effect size*	*Genotype*	*Effect*	*Genotype*	*Effect*	*Genotype*	*Effect*
rs12395	A	0.02	AA	+.02 (x2)	TT		AT	0.02
rs44346	G	-0.04	GT	-0.04	TT		TT	
rs72557	C	-0.05	CG	-0.05	CC	-.05 (x2)	GG	
rs18338	A	0.09	AT	0.09	TT		TT	
rs29849	T	0.004	TT	+.004 (x2)	CT	0.004	CT	0.004
rs43466	T	0.07	AA		TA	0.07	AA	
rs29457	G	-0.01	CC		CC		GC	-0.01
rs13458	C	0.015	AA		CA	0.0015	AA	
		Polygenic score:	0.048		-0.0245		0.014	

Figure 108: PRS 소개

Figure 5.15 PRS 소개 – 다유전자 위험 점수 계산 과정과 개인별 위험도 평가 예시

위 표는 가산적 모델(additive model)을 가정한 PRS 계산의 실제 사례를 보여준다. Fred, Alice, Greg 세 명의 개인에 대해 8개의 SNP(rs12395, rs44346, rs72557 등)의 유전형과 효과 크기를 바탕으로 PRS를 계산한 예이다. 각 SNP마다 효과 대립유전자(Effect allele)와 효과 크기(Effect size)가 정해져 있으며, 가산적 모델에 따라 효과 대립유전자의 개수에 비례하여 효과가 결정된다.

예를 들어, rs12395에서 효과 대립유전자는 A이고 효과 크기는 0.02이다. Fred는 AA 유전형을 가지므로 A 대립유전자를 2개 보유하여 0.02 × 2 = 0.04의 효과를 받는다. Alice는 TT 유전형으로 A 대립유전자가 없으므로 이 SNP로부터 0의 효과를 받는다. Greg는 AT 유전형으로 A 대립유전자를 1개 보유하여 0.02 × 1 = 0.02의 효과를 받는다.

이런 식으로 모든 SNP에 대해 계산한 후 합산하면, Fred의 PRS는 0.048, Alice의 PRS는 −0.0245, Greg의 PRS는 0.014가 된다. 이 결과에 따르면 Fred가 가장 높은 위험도를, Alice가 가장 낮은 위험도를 가지는 것으로 예측된다.

6.6.2 PRS의 활용과 한계

PRS는 개인 맞춤 의학에서 질병 예방과 조기 진단에 활용될 수 있다. 높은 PRS를 가진 사람은 해당 질병에 대한 유전적 소인이 크므로 더 주의 깊은 모니터링이나 예방 조치가 필요할 수 있다.

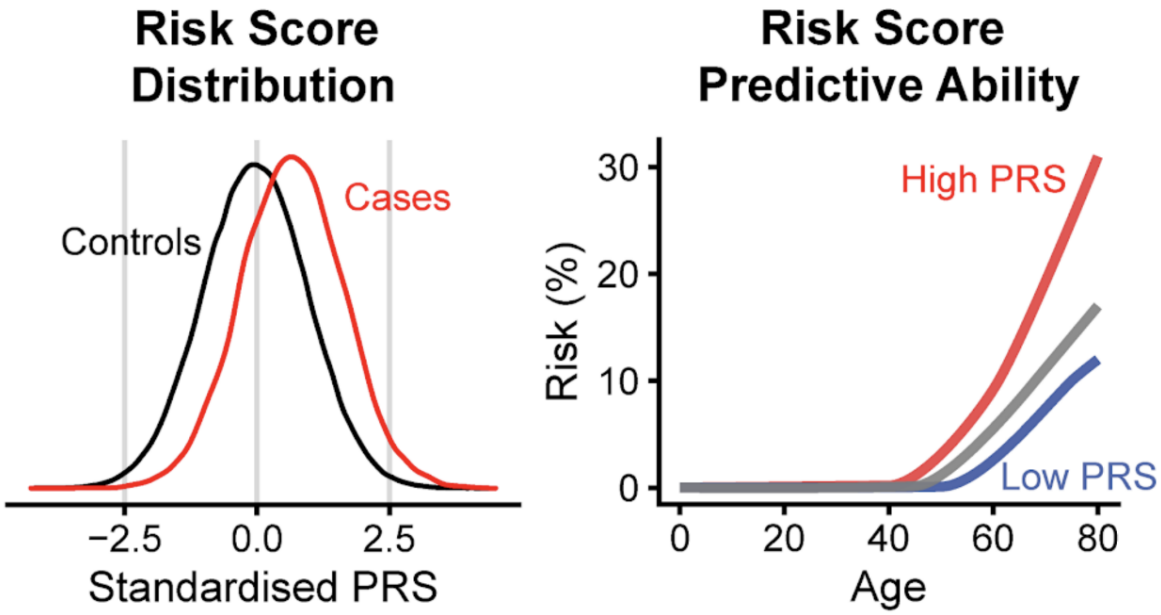


Figure 109: 다유전자 위험 점수

Figure 5.16 다유전자 위험 점수 분포 – 인구 집단에서 PRS의 분포와 질병 위험도 관계

하지만 PRS에는 중요한 한계들이 있다. 첫째, 변이–변이 상호작용이나 변이–환경 상호작용 같은 복잡한 요인들을 충분히 반영하지 못한다. 둘째, 대부분의 GWAS가 유럽계 인구를 대상으로 수행되어 다른 인구 집단에서는 정확도가 떨어질 수 있다. 셋째, 환경적 요인이나 생활 습관의 영향은 고려되지 않는다.

6.6.3 PRS 계산 도구

PRS 계산을 위해서는 전문적인 생명정보학 도구들이 사용된다. 대표적으로 Plink(https://www.cog-genomics.org/plink/)가 있다. 이러한 도구들은 GWAS 결과를 바탕으로 개인의 유전형 데이터에서 자동으로 PRS를 계산해준다.

6.7 과잉 표현 분석(Over-Representation Analysis, ORA)

GWAS 이후에는 결과를 검증해야 한다. 이를 위해 과잉 표현 분석 (Over-Representation Analysis, ORA), 경로 분석 (Pathway Analysis), 유전자 세트 농축 분석(Gene Set Enrichment Analysis, GWAS) 등의 방법과 실험을 통한 표현형 검증 등이 수행될 수 있다.

이 중 ORA는 관심 있는 유전자 세트가 특정 기능적 범주에 통계적으로 유의미하게 과잉 표현되어 있는지 평가하는 방법으로, 매우 널리 활용된다. 이는 실험에서 얻은 유전자 목록의 생물학적 의미를 해석하는 데 사용된다. 유전자 발현 데이터, 단백질체 데이터, GWAS 결과 등을 분석하는 데 활용될 수 있다.

ORA는 주로 GO(Gene Ontology) 용어(특히 생물학적 과정)와 경로 데이터베이스를 기반으로 한다.

6.7.1 유전자 온톨로지(Gene Ontology, GO)

GO는 분자 수준부터 생물체 수준까지 생물학적 시스템을 모델링하는 포괄적인 용어 목록이다. GO는 다음 세 가지 주요 영역으로 구성된다:

1. 분자 기능(Molecular Function): 단백질이나 RNA와 같은 유전자 산물이 수행하는 분자 수준의 활동
2. 세포 구성 요소(Cellular Component): 유전자 산물이 활성화되는 세포 내 위치
3. 생물학적 과정(Biological Process): 여러 분자 기능이 함께 작용하여 완성되는 더 큰 프로세스

각 GO 용어에는 해당 용어와 관련된 유전자 목록이 있다. GO는 https://geneontology.org/에서 확인할 수 있다.

Figure 5.17 유전자 온톨로지 – GO 용어의 계층 구조와 세 가지 주요 영역

Figure 5.18 GO 용어 상세 – 특정 GO 용어에 대한 상세 정보와 관련 유전자 목록

6.7.2 다양한 경로 데이터베이스

생물학적 경로 관련 주요 데이터베이스는 다음과 같다:

- KEGG Pathway Database: https://www.genome.jp/kegg/pathway.html

KEGG 경로

Figure 5.19 KEGG 경로 – KEGG 데이터베이스의 생물학적 경로 예시

- Reactome Database: https://reactome.org/

Figure 5.20 Reactome 경로 – Reactome 데이터베이스의 생물학적 반응 네트워크

Figure 5.21 Reactome 데이터베이스 – 생물학적 경로와 반응을 체계적으로 정리한 데이터베이스

- QIAGEN Ingenuity Pathway Analysis (IPA): https://digitalinsights.qiagen.com/products-overview/discovery-insightsportfolio/analysis-and-visualization/qiagen-ipa/

IPA 경로 분석

Figure 5.22 Ingenuity Pathway Analysis (IPA) – 유전자 네트워크와 경로 분석을 위한 상용 도구

- WikiPathways: https://www.wikipathways.org/

Figure 5.23 WikiPathways 플랫폼 – 커뮤니티 기반 생물학적 경로 큐레이션 플랫폼

6.7.3 초기하 분포(Hypergeometric Distribution)

ORA에서 통계적 유의성을 평가하기 위해 초기하 분포가 사용된다. 이 분포는 "비복원 추출"에서 나타나는 확률을 계산하는 데 사용되며, ORA의 핵심 통계적 기반이 된다.

6.7.3.1 초기하 분포의 개념 초기하 분포는 다음과 같은 상황에서 사용된다: – **전체 유전자 집합(N)**: 분석에 포함된 모든 유전자의 수 – **특정 경로의 유전자(M)**: 관심 있는 생물학적 경로나 기능에 속하는 유전자의 수 – **관심 유전자 목록(n)**: GWAS에서 발견된 유의미한 유전자들의 수 – **겹치는 유전자(k)**: 관심 유전자 목록과 특정 경로 모두에 속하는 유전자의 수

6.7.3.2 실제 적용 예시 구체적인 예를 들어보자. 전체 인간 유전체에 20,000개의 유전자가 있고(N=20,000), "당뇨병 관련 경로"에 100개의 유전자가 속한다고 하자(M=100). GWAS 연구에서 50개의 유의미한 유전자를 발견했는데(n=50), 이 중 10개가 당뇨병 관련 경로에 속한다면(k=10), 이것이 우연의 일치인지 통계적으로 유의미한 농축인지를 초기하 분포로 평가할 수 있다.

초기하 분포 확률 공식: $P(X = k) = C(M,k) \times C(N-M,n-k) / C(N,n)$

여기서 C(a,b)는 조합(combination)을 의미한다.

6.7.3.3 P-값 계산과 해석 ORA에서는 관찰된 겹치는 유전자 수 이상의 겹침이 일어날 확률을 계산한다: P-값 = $P(X \geq k) = \sum P(X = i)$ (i=k부터 min(n,M)까지)

P-값이 낮을수록(일반적으로 0.05 미만) 해당 생물학적 경로나 기능이 우연이 아닌 진짜 연관성을 가질 가능성이 높다고 해석한다.

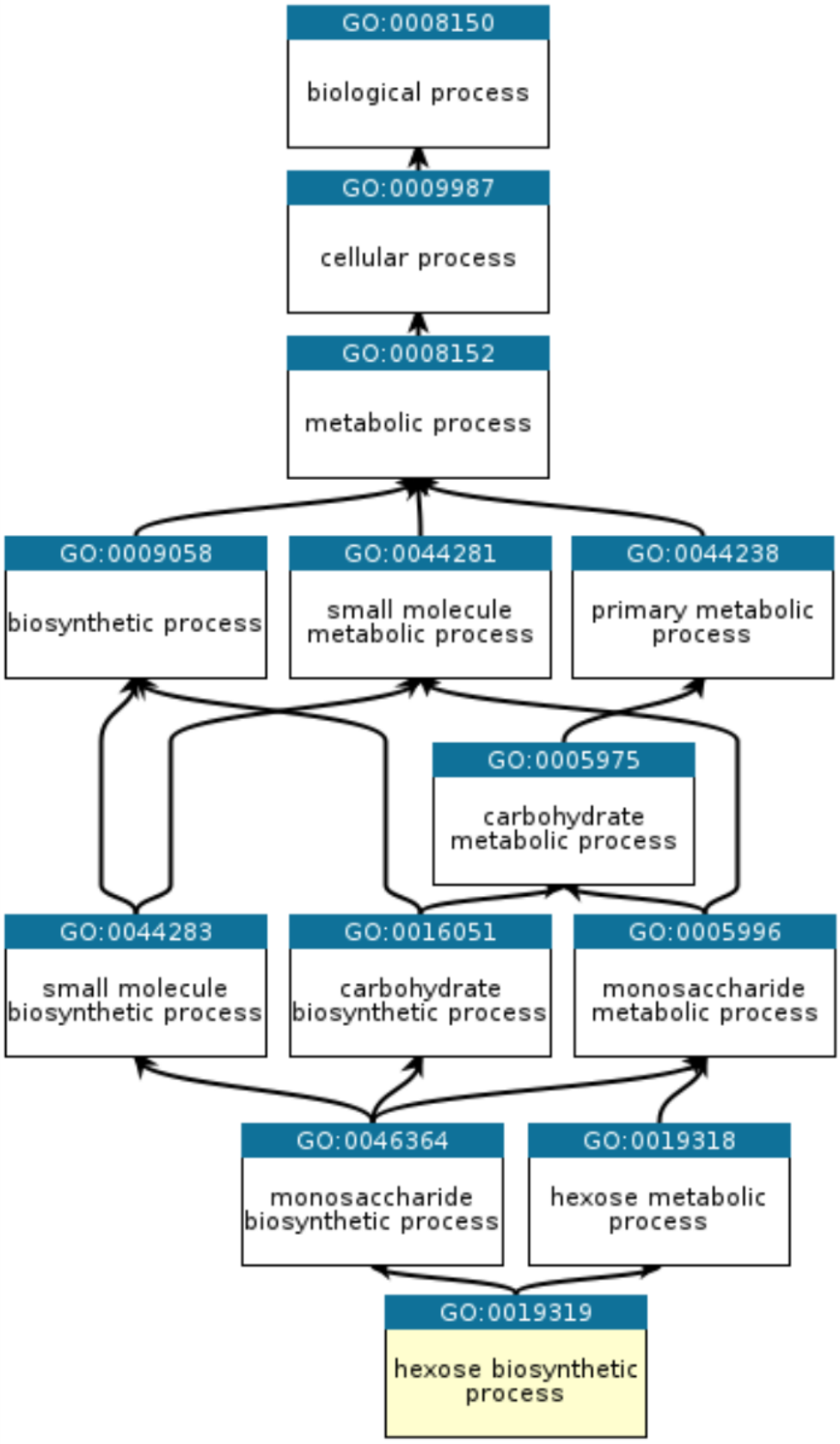


Figure 110: 유전자 온톨로지

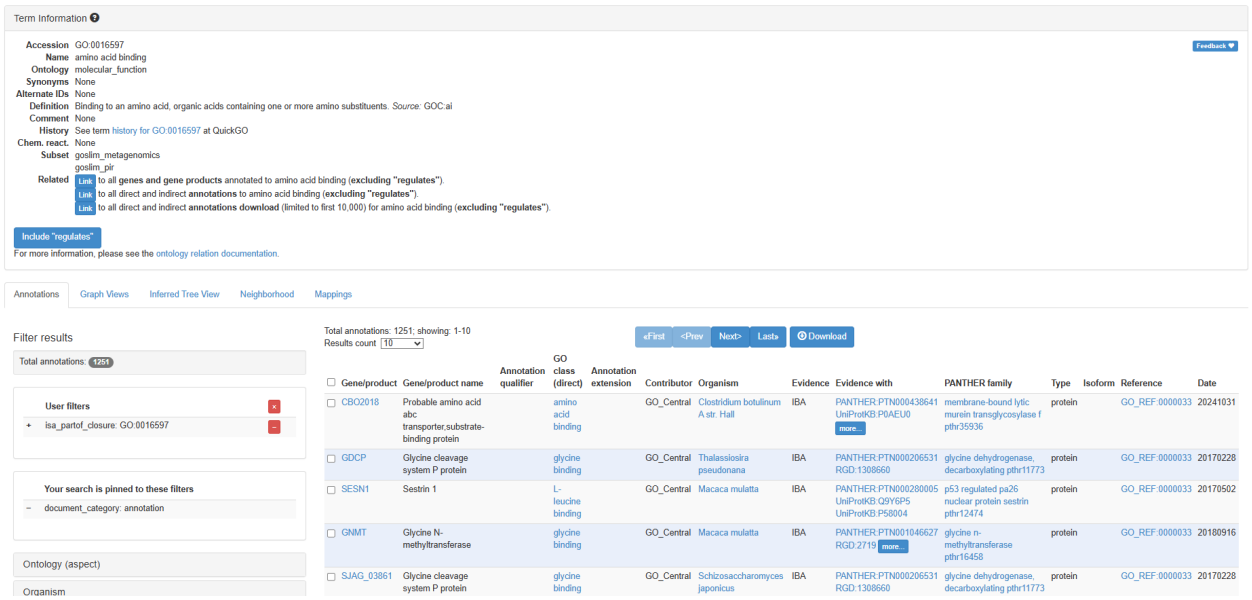


Figure 111: GO 용어 상세

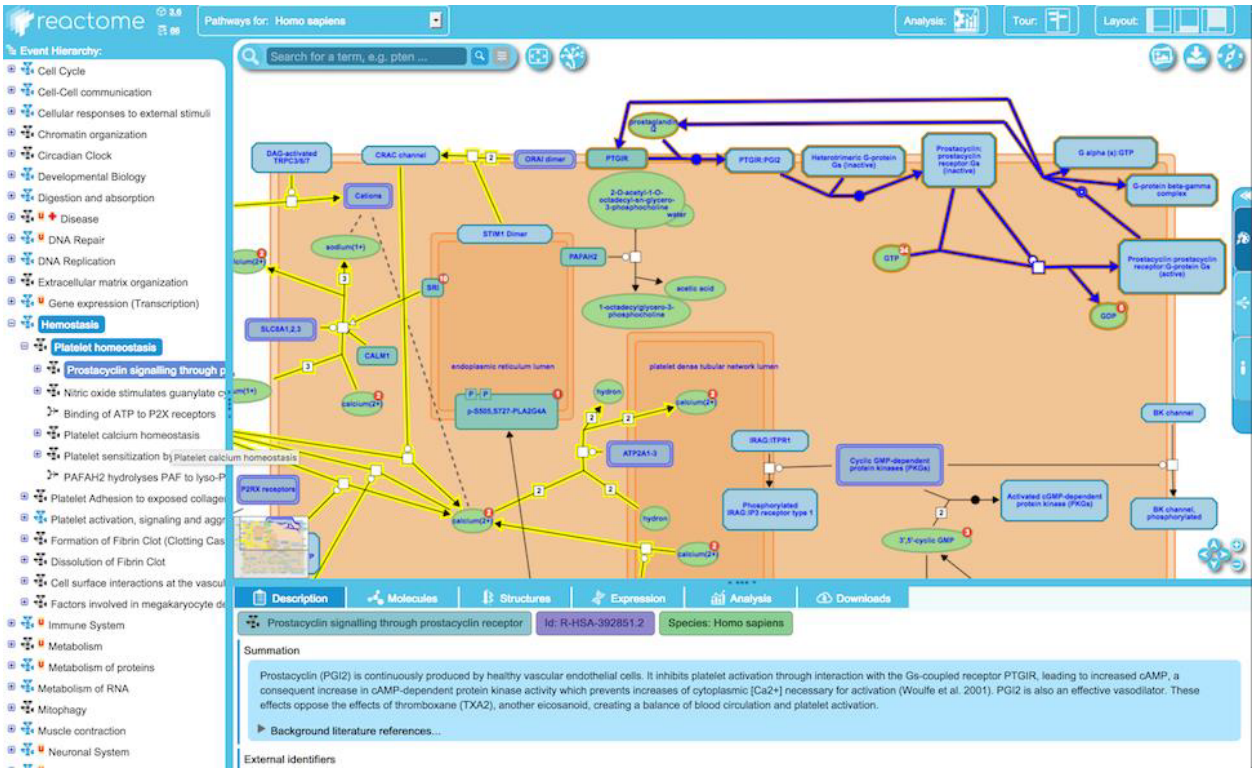


Figure 112: Reactome

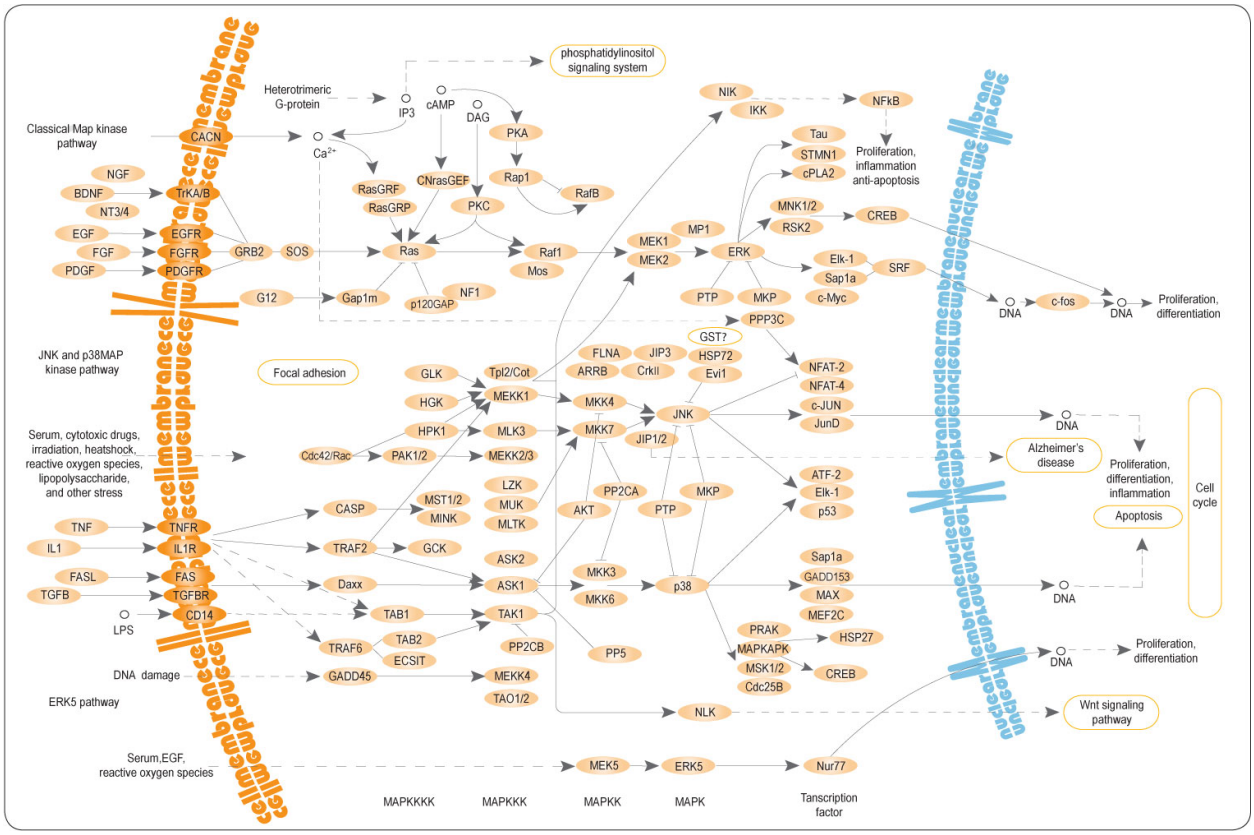


Figure 113: Reactome 데이터베이스

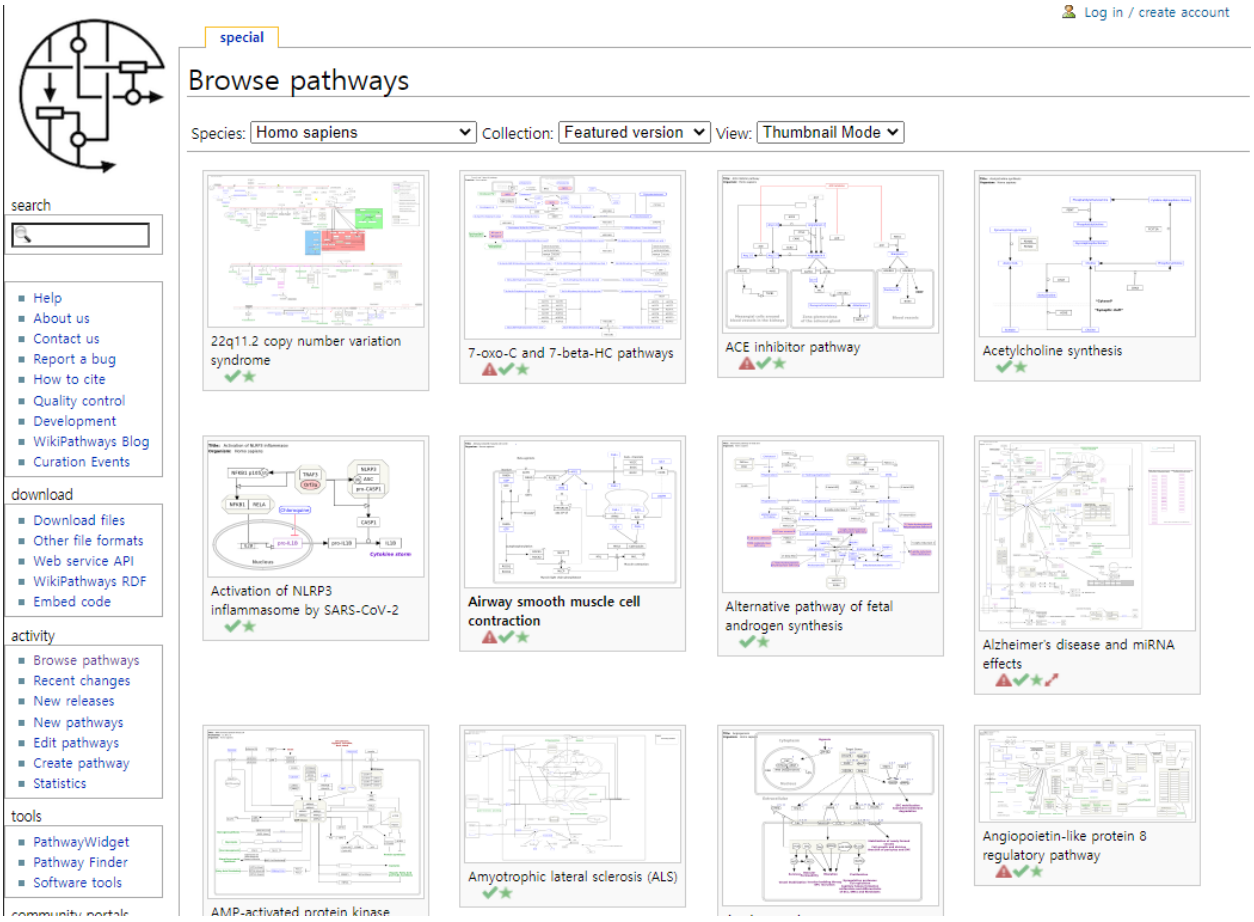


Figure 114: WikiPathways 플랫폼

6.7.3.4 다중 검정 보정 ORA에서는 수백 개의 생물학적 경로를 동시에 검정하므로 다중 검정 문제가 발생한다. 이는 많은 검정을 동시에 수행할 때 우연히 유의한 결과를 얻을 확률이 크게 증가하는 현상이다.

예를 들어, 500개의 생물학적 경로를 검정한다고 가정해보자. 각 검정에서 α = 0.05의 유의수준을 사용한다면, 실제로는 연관성이 없음에도 불구하고 우연히 유의한 결과를 얻을 확률이 각 검정마다 5%씩 존재한다. 500개 검정을 수행하면 전체적으로 약 25개의 거짓 양성 결과를 얻게 될 것이다.

이 문제를 해결하기 위해 여러 보정 방법이 사용된다. Bonferroni 보정은 가장 보수적인 방법으로 각 검정의 유의수준을 전체 검정 수로 나누어 조정한다. 예를 들어 500개 검정 시 α = 0.05/500 = 0.0001을 기준으로 사용한다. 이 방법은 가족별 오류율을 엄격히 통제하지만 너무 보수적이어서 진짜 유의한 결과도 놓칠 수 있다.

FDR(False Discovery Rate) 보정은 Benjamini–Hochberg 방법이 가장 널리 사용되며, 거짓 발견율을 일정 수준 이하로 유지하면서 검정력을 높인다. Bonferroni보다 덜 보수적이면서도 효과적으로 거짓 양성을 통제한다.

q-value는 FDR의 개념을 확장한 것으로, 각 가설에 대한 개별적인 FDR 추정치이다. q-value는 "해당 P-값 이하의 모든 검정을 유의하다고 판단할 때 예상되는 FDR"을 의미한다. 예를 들어 q-value = 0.1이면, 해당 P-값 기준으로 유의하다고 판단한 결과 중 10%가 거짓 양성일 것으로 예상됨을 뜻한다. 실무에서는 $q < 0.05$ 또는 $q < 0.1$을 기준으로 많이 사용되며, 각 개별 검정에 대한 직관적인 해석이 가능하다는 장점이 있다.

ORA 결과표에서는 일반적으로 원래 P-값, 보정된 P-값, q-value 등의 정보가 제공되며, 연구자는 연구 목적에 따라 적절한 기준을 선택하여 결과를 해석한다.

6.7.3.5 ORA 분석 도구 이러한 복잡한 계산을 수행하기 위해 다양한 도구들이 개발되었다: - **Enrichr**(https://maayanlab.cloud/Enri 웹 기반 도구로 다양한 데이터베이스 지원 - **DAVID**(https://david.ncifcrf.gov/): 포괄적인 기능 주석 도구 - **g:Profiler**(https://biit.cs.ut.ee/gprofiler/): 다중 종을 지원하는 농축 분석 도구 - **clusterProfiler**: R 기반 패키지로 프로그래밍 환경에서 사용

이러한 도구들은 사용자가 제공한 유전자 목록에 대해 자동으로 초기하 분포 기반 통계 검정을 수행하고, 결과를 시각화해준다.

Figure 5.24 Enrichr 도구 – 과잉 표현 분석을 위한 웹 기반 도구 인터페이스

6.8 결론

전장유전체 연관분석(GWAS)은 현대 의생명정보학에서 가장 중요한 연구 방법 중 하나로, 유전적 변이와 질병 또는 형질 간의 연관성을 체계적으로 탐색할 수 있게 해준다. 멘델 질환과 달리, 현대 사회의 주요 질병들은 대부분 다유전자 질환으로 여러 변이들의 복합적 작용에 의해 발생한다.

GWAS는 수백만 개의 SNP를 동시에 분석하여 각 변이의 효과 크기를 통계적으로 추정한다. 이 과정에서 P-값과 맨해튼 플롯을 통해 결과를 시각화하고, 연관 불균형과 해플로타입 개념을 활용하여 분석 효율성을 높인다. 중요한 점은 해플로타입 패턴이 인구 집단마다 다르므로, 연구 대상에 맞는 적절한 참조 데이터를 사용해야 한다는 것이다.

GWAS 결과는 다유전자 위험 점수(PRS)를 통해 개인의 유전적 위험도 평가에 활용될 수 있다. 하지만 PRS는 변이간 상호작용이나 환경 요인을 충분히 반영하지 못하는 한계가 있으며, 특히 유럽계 인구 중심의 연구 결과를 다른 인구 집단에 적용할 때는 주의가 필요하다.

GWAS 이후의 기능적 해석을 위해서는 과잉 표현 분석(ORA)과 같은 후속 분석이 필수적이다. ORA는 초기하 분포를 기반으로 하여 발견된 유전자들이 특정 생물학적 경로에 농축되어 있는지 평가한다. 이 과정에서 다중 검정 보정 등의 방법을 통해 통계적 신뢰성을 확보해야 한다.

GWAS는 정밀의학의 기반이 되는 중요한 도구이지만, 결과 해석 시에는 통계적 한계와 생물학적 복잡성을 함께 고려해야 한다. 향후 더 다양한 인구 집단을 포함한 연구와 환경 요인을 통합한 분석 방법의 발전이 필요하며, 이를 통해 개인 맞춤 의학의 실현에 한 걸음 더 다가갈 수 있을 것이다.

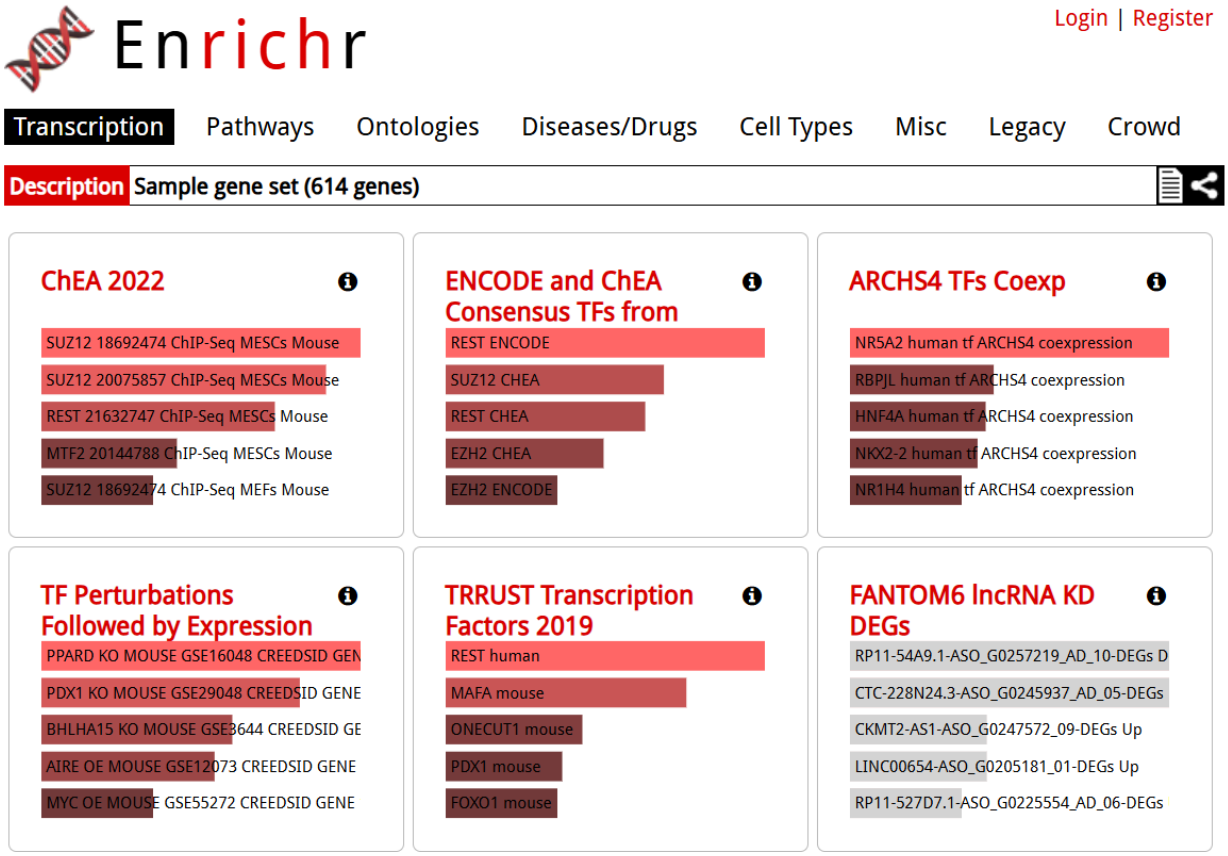


Figure 115: Enrichr 도구

7 암

7.1 암이란 무엇인가?

암은 신체의 일부 세포가 통제 불능으로 성장하여 몸의 다른 부분으로 퍼지는 질병이다(https://www.cancer.gov/about-cancer/understanding/what-is-cancer). '통제 불능'이라는 표현은 이러한 세포들이 몸의 면역 체계를 회피한다는 의미를 포함한다.

암은 인간뿐만 아니라 다른 동물에서도 발생한다. 한국인의 경우, 평생 동안 약 38%가 암에 걸리는 것으로 알려져 있다(2019년 국가암등록통계). 생활 습관 중 특히 흡연은 모든 암의 약 30%를 유발하는 주요 원인이다(국민건강보험공단). 흡연자는 비흡연자에 비해 폐암 발생 위험이 15~30배 높다(https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm).

7.2 암은 왜 발생하는가?

7.2.1 단세포 생물의 유전자 유산

단세포 생물의 생존 전략은 '증식하거나 소멸하거나(Proliferate or perish)'이다. 흥미롭게도, 우리 인간은 단세포 생물이었던 시절부터 물려받은 유전자들을 여전히 가지고 있다! 이러한 유전자들은 암을 유발할 잠재력을 가지고 있으며, 이를'**암유전자(Oncogenes)**'라고 부른다.

7.2.2 종양 억제자(Tumor suppressors)

한편, 이러한 암유전자들의 활성을 '억제'하는 유전자들도 존재한다. 이를'**항암유전자(Anti-oncogenes)**' 또는 '**종양 억제 유전자(Tumor Suppressor Genes)**'라고 한다. 이러한 유전자들은 세포 주기 조절, DNA 복구, 세포 사멸(apoptosis) 등의 과정에 관여하여 암세포의 발생을 방지한다.

7.3 암 관련 신호 전달 경로(Pathways)

이전 장에서 살펴보았듯 세포 내에서는 다양한 신호 전달 경로가 존재하는데, 이 중 일부는 암 발생과 밀접한 관련이 있다.

7.3.1 MAPK-ERK 경로

MAPK-ERK 경로는 세포 성장, 분화, 생존 등에 중요한 역할을 하는 신호 전달 체계로, 암 발생과 연관된 대표적인 경로이다. 이 경로의 비정상적인 활성화는 다양한 암 유형에서 관찰된다.

Figure 6.1 RAS 경로 - MAPK-ERK 신호 전달 경로와 암 발생에서의 역할

7.3.2 RAS 유전자의 종간 보존

RAS 유전자는 MAPK-ERK 경로에서 중요한 역할을 하는 유전자로, 다양한 생물 종에서 매우 보존되어 있다. 이는 이 유전자의 기능이 진화적으로 매우 중요하다는 것을 의미한다.

이러한 유전자 보존성을 이해하기 위해서는 유전자들 간의 진화적 관계를 알아야 한다. **상동유전자(Homologs)**는 공통 조상으로부터 유래된 유전자들을 의미한다. 이 중에서 **직교유전자(Orthologs)**는 서로 다른 종에서 유사한 기능을 하는 상동유전자들이다. 예를 들어, 인간의 RAS 유전자와 생쥐의 RAS 유전자는 직교유전자 관계에 있다. 반면 **준동유전자(Paralogs)**는 같은 종 내에서 유전자 중복을 통해 유래된 상동유전자들로, 인간에게는 KRAS, NRAS, HRAS라는 세 가지 RAS 준동유전자가 존재한다. **유사유전자(Analogs)**는 서로 다른 조상에서 유래되었지만 유사한 기능을 수행하는 유전자들을 말한다.

Figure 6.2 상동유전자 관계 - 직교유전자, 준동유전자, 유사유전자 간의 진화적 관계

RAS 유전자가 효모부터 인간까지 광범위한 생물 종에서 보존되어 있다는 사실은 세포 성장과 분화에서 이 유전자가 얼마나 필수적인 역할을 하는지를 보여준다. 동시에 이것이 RAS 유전자의 변이가 암 발생에 중요한 영향을 미치는 이유이기도 하다.

Figure 6.3 인간 RAS 경로 - 인간에서의 RAS 신호 전달 경로와 연관된 단백질들

Figure 6.4 다른 생물의 RAS 경로 - 다양한 생물 종에서 RAS 경로의 보존성

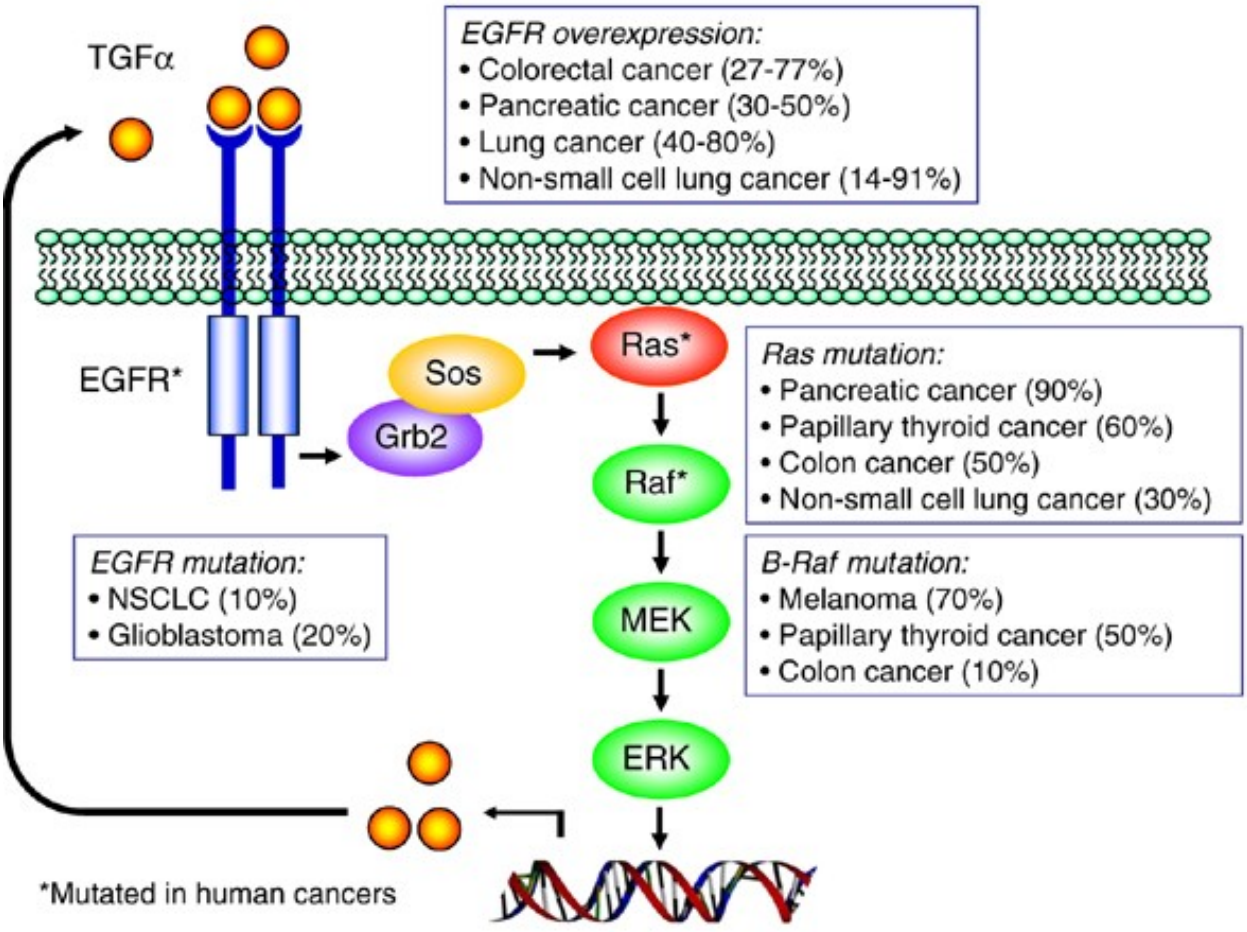


Figure 116: RAS 경로

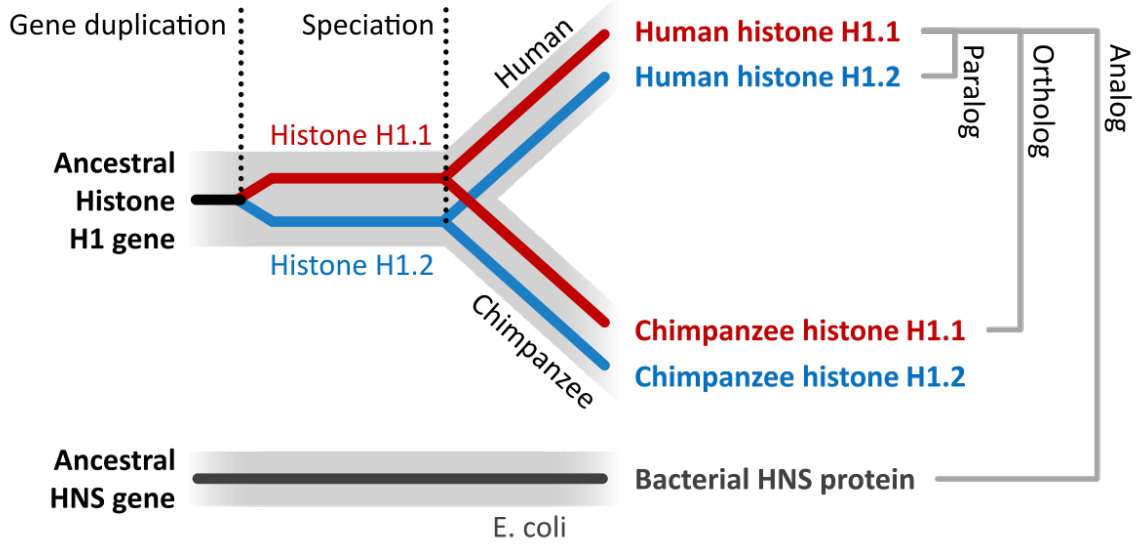


Figure 117: 상동유전자 관계

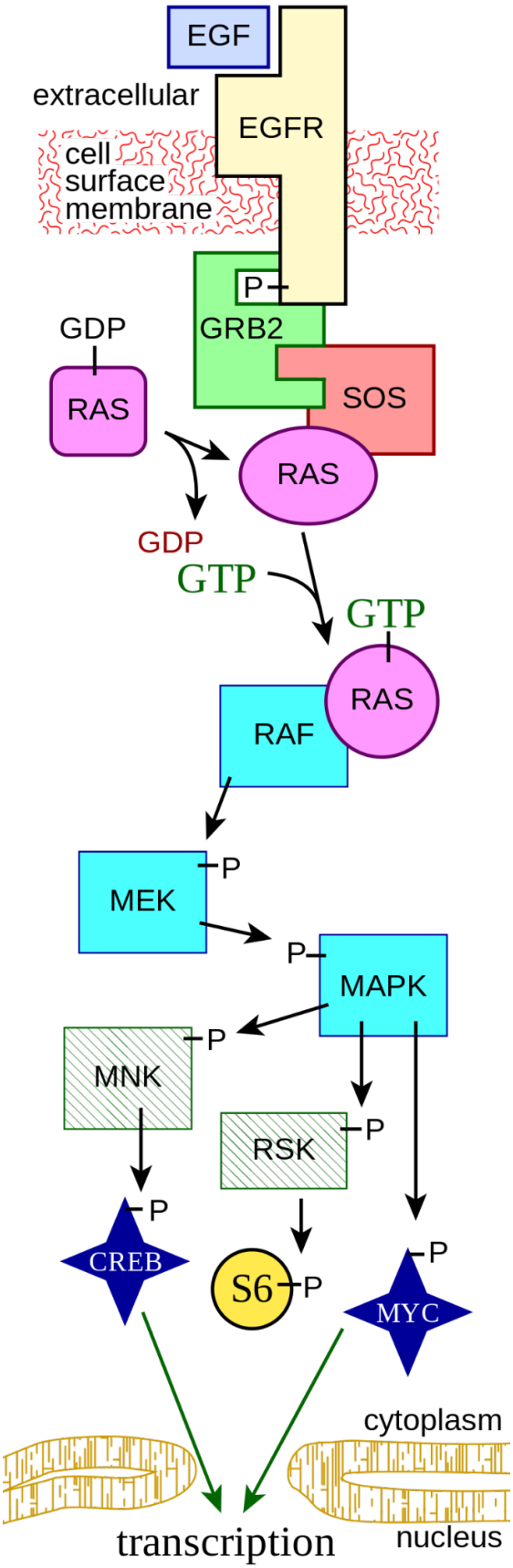


Figure 118: 인간 RAS 경로

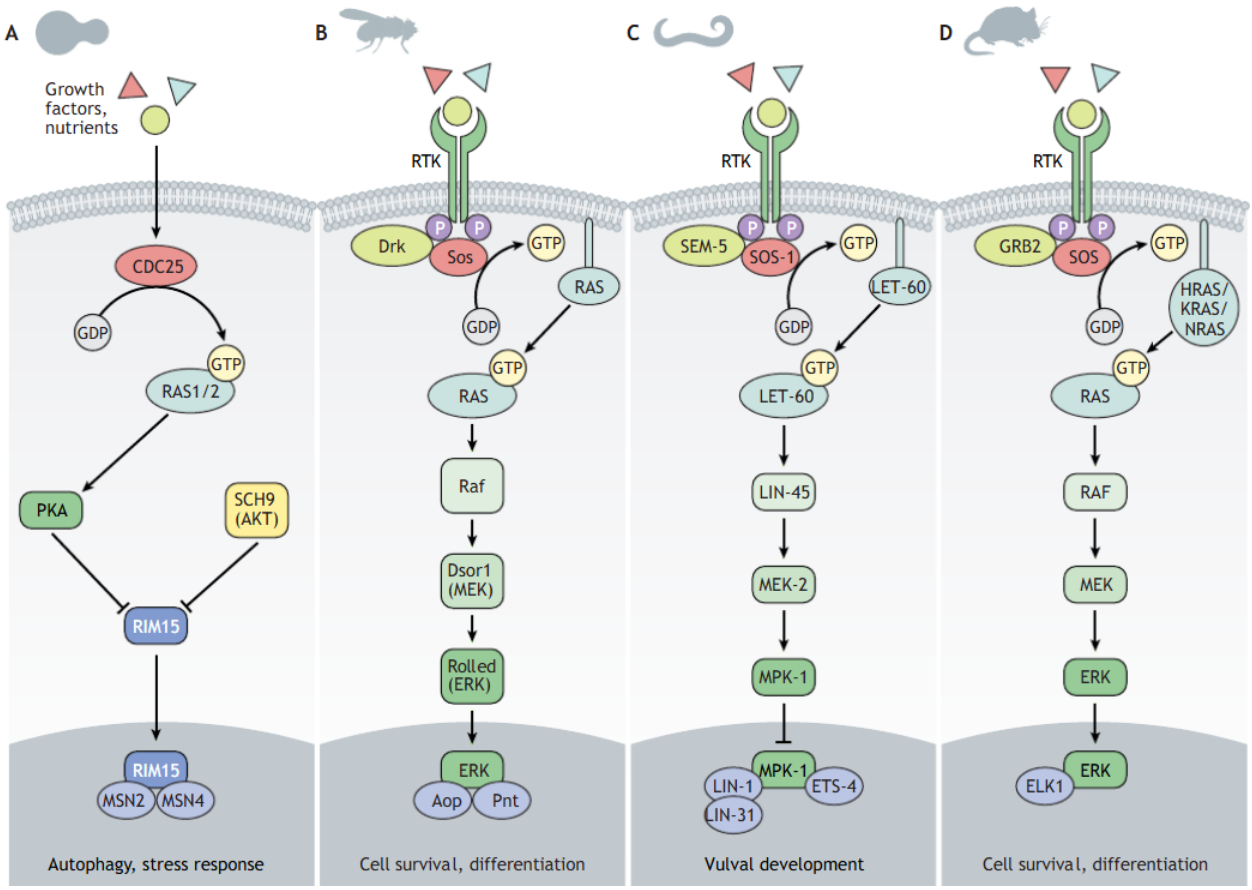


Figure 119: 다른 생물의 RAS 경로

7.4 원암유전자와 암유전자

정상 세포에서는 세포 증식, 분화, 생존 등에 관여하는 유전자를 '**원암유전자(Proto-oncogene)**'라고 한다. 이러한 원암유전자가 돌연변이를 통해 비정상적으로 활성화되면'**암유전자(Oncogene)**'가 된다. 암유전자는 세포의 비정상적인 증식을 촉진하여 암 발생에 기여한다.

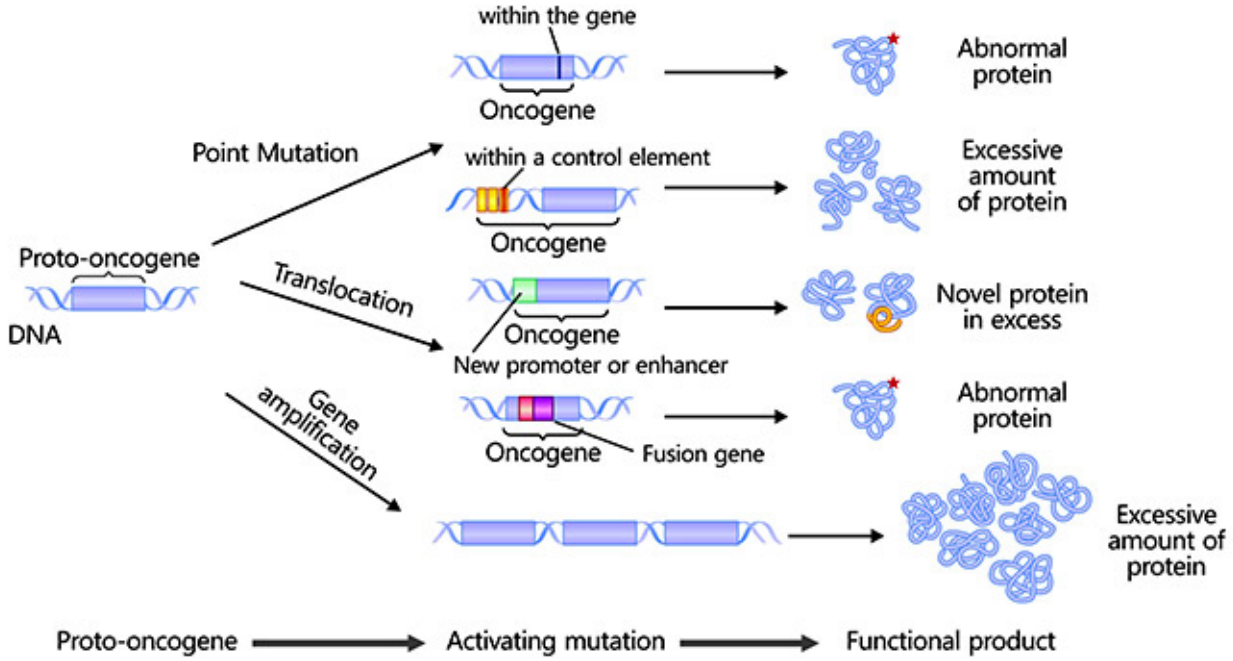


Figure 120: 원암유전자

Figure 6.5 원암유전자 – 원암유전자에서 암유전자로의 변환 과정과 주요 예시들

7.5 돌연변이

7.5.1 BCR-ABL

BCR-ABL은 9번 염색체의 ABL 유전자와 22번 염색체의 BCR 유전자가 상호 전위(translocation)되어 형성된 융합 유전자이다. 이 융합 유전자는 만성 골수성 백혈병(CML)의 주요 원인으로 알려져 있다.

Figure 6.6 전위 예시 – 염색체 간 전위에 의한 융합 유전자 형성

Figure 6.7 전위 리드 정렬 예시 – IGV에서 관찰되는 전위 관련 시퀀싱 리드 정렬 패턴

Figure 6.8 대규모 복제 – 유전자 또는 염색체 영역의 대규모 복제 현상

Figure 6.9 대규모 삽입/결실 – 대규모 DNA 삽입 및 결실 변이의 예시

Figure 6.10 BCR-ABL 융합 유전자 – 9번과 22번 염색체 전위에 의한 BCR-ABL 융합 유전자 형성

Figure 6.11 BCR-ABL과 RAS 경로 – BCR-ABL 융합 단백질이 RAS 경로에 미치는 영향

7.5.2 복제수 변이(Copy Number Variation, CNV)

CNV는 DNA 세그먼트가 정상보다 많거나 적게 반복되는 유전적 변이이다. 암에서는 종양 유전자가 증폭되거나 종양 억제 유전자가 소실되는 형태로 CNV가 관찰될 수 있다.

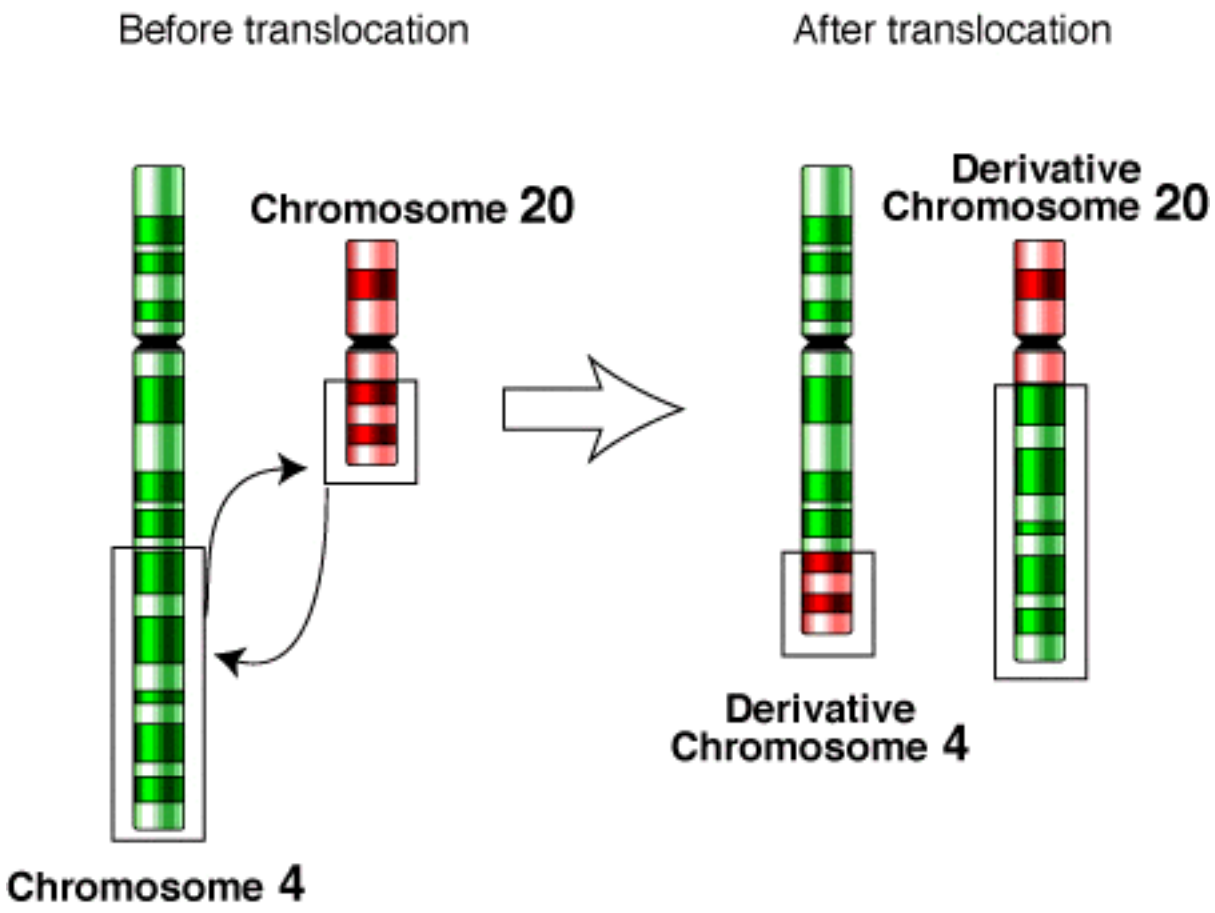


Figure 121: 전위 예시

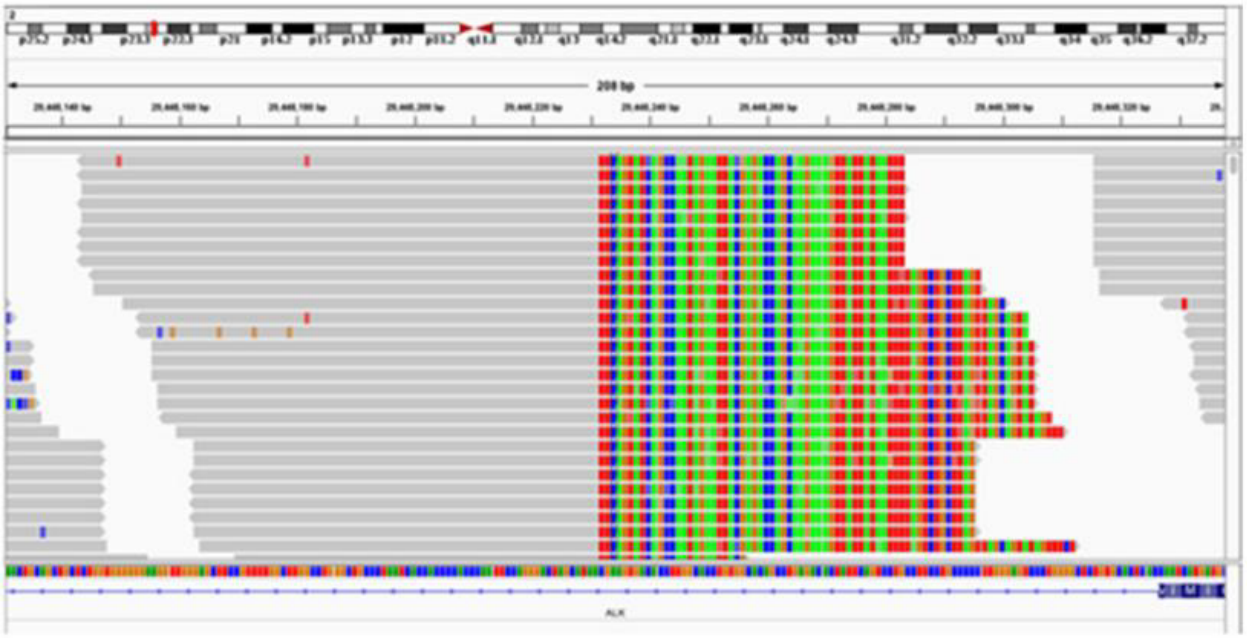


Figure 122: 전위 과정

7.6 대표적인 암유전자와 종양 억제자

7.6.1 종양 억제자

1. TP53:
   - 많은 암 유형에서 흔히 돌연변이가 발견되는 유전자이다.
   - 'genome guardian'으로 불리며, DNA 손상 시 세포 주기 정지, DNA 복구, 세포 사멸 등을 조절한다.
2. BRCA1/2:
   - 일명 "유방암 유전자"로 불린다.
   - 돌연변이가 있으면 유방암/난소암 발생 위험이 크게 증가한다.
   - DNA 이중 가닥 복구에 중요한 역할을 한다.

Figure 6.12 BRCA와 유방암 – BRCA1/2 유전자 변이와 유방암/난소암 발생 위험도

7.7 드라이버 돌연변이와 패신저 돌연변이

7.7.1 드라이버 돌연변이(Driver mutations)

암을 유발하는 돌연변이로, 주로 암유전자나 종양 억제자에서 발생한다. 이러한 돌연변이는 세포의 성장, 생존, 증식에 직접적인 영향을 미친다.

7.7.2 패신저 돌연변이(Passenger mutations)

암 발생에 직접적인 영향을 미치지 않는 양성 돌연변이이다. 암세포에서는 다음과 같은 이유로 이러한 돌연변이가 축적되는 경향이 있다: – 암은 빠르게 증식하기 때문에 돌연변이가 더 많이 발생한다. – DNA 복구 기작이 손상된 경우, 암은 더욱 공격적으로 변할 수 있다.

7.8 체세포 돌연변이(Somatic mutations)

체세포 돌연변이는 암세포에서 발견되는 유전적 변화로, 정상 세포에서는 관찰되지 않는 특징을 가진다. 이러한 돌연변이는 크기와 복잡성에 따라 여러 유형으로 분류할 수 있으며, 각각이 암의 발생과 진행에 서로 다른 방식으로 기여한다.

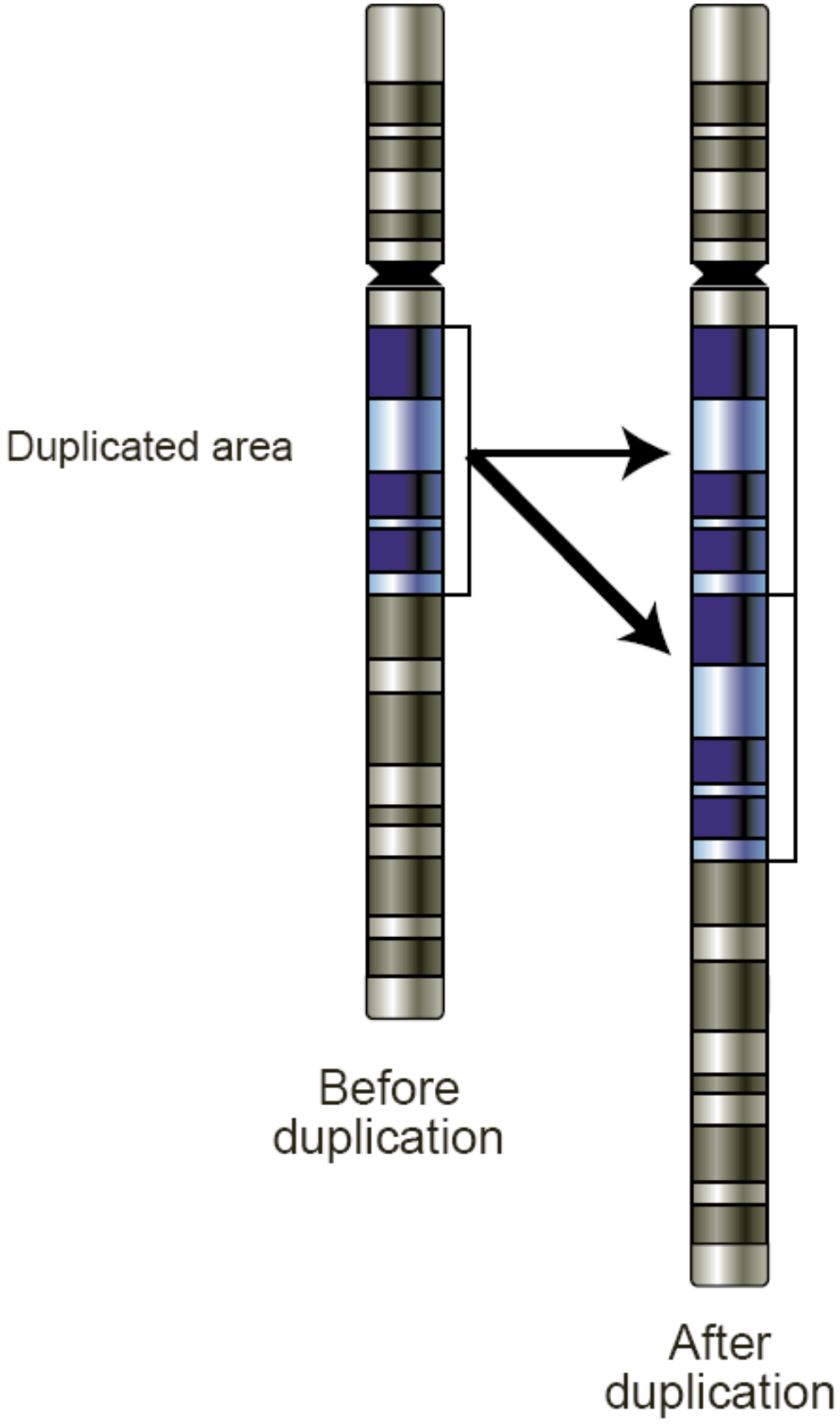


Figure 123: 대규모 복제

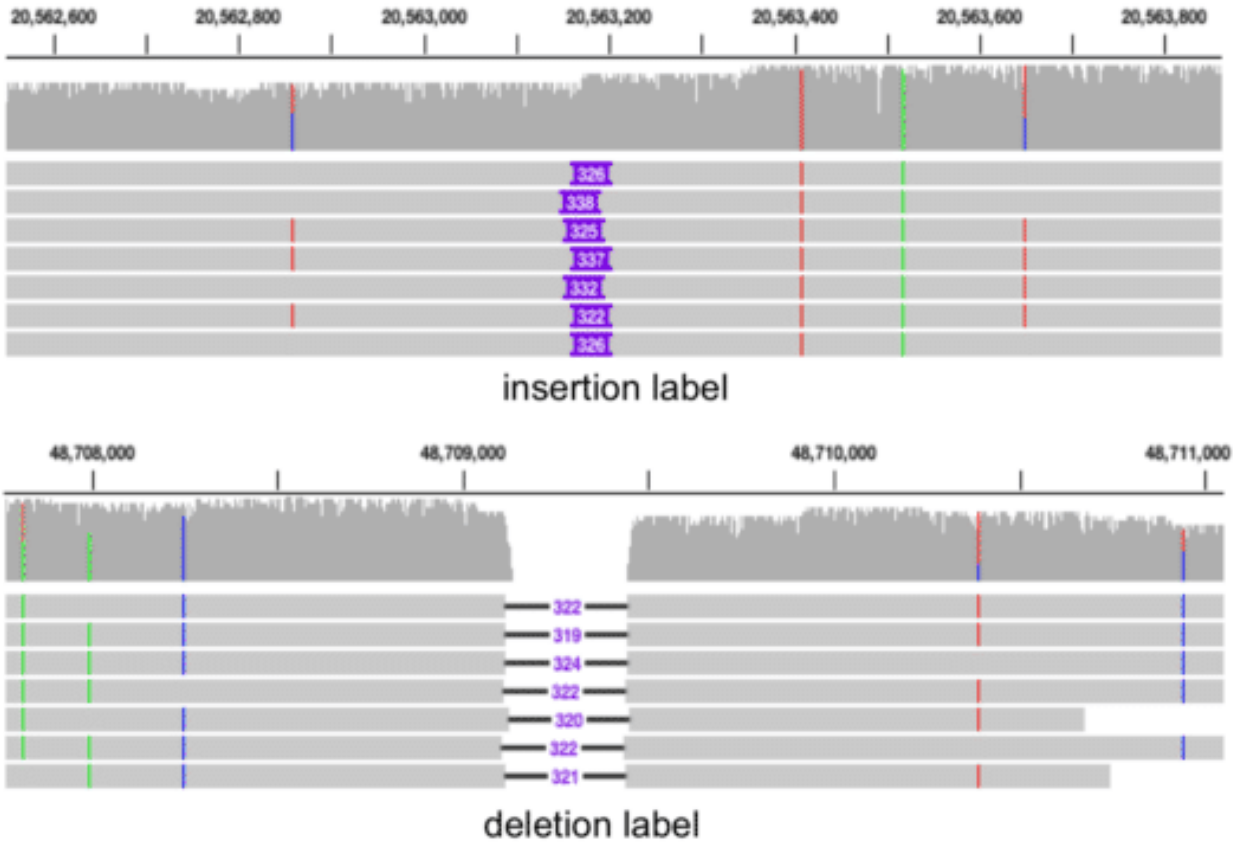


Figure 124: 대규모 삽입/결실

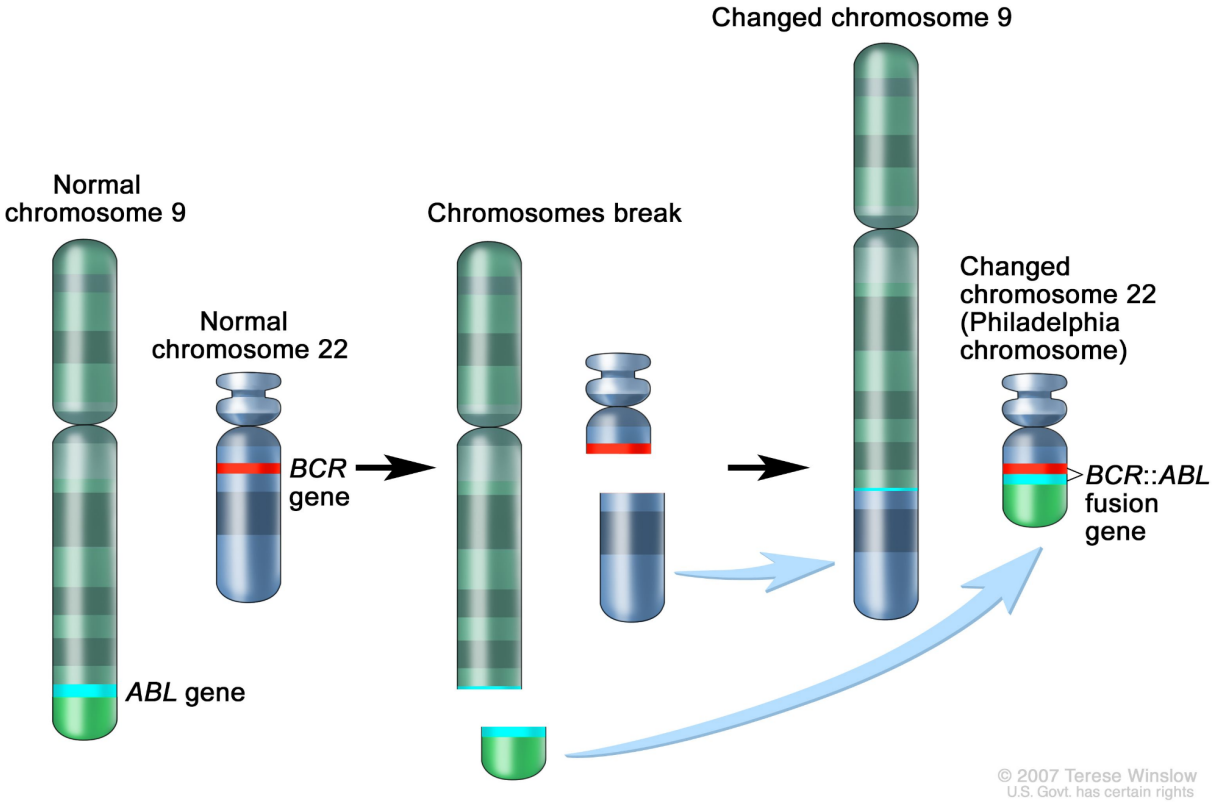


Figure 125: BCR-ABL 융합 유전자

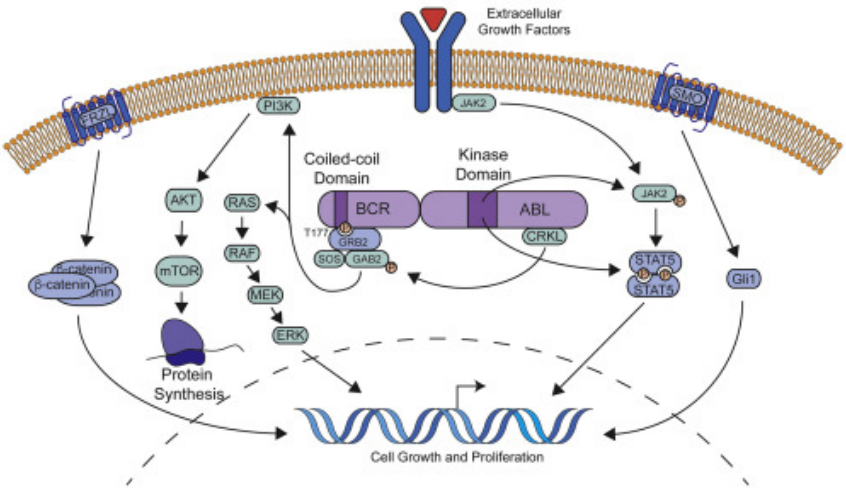
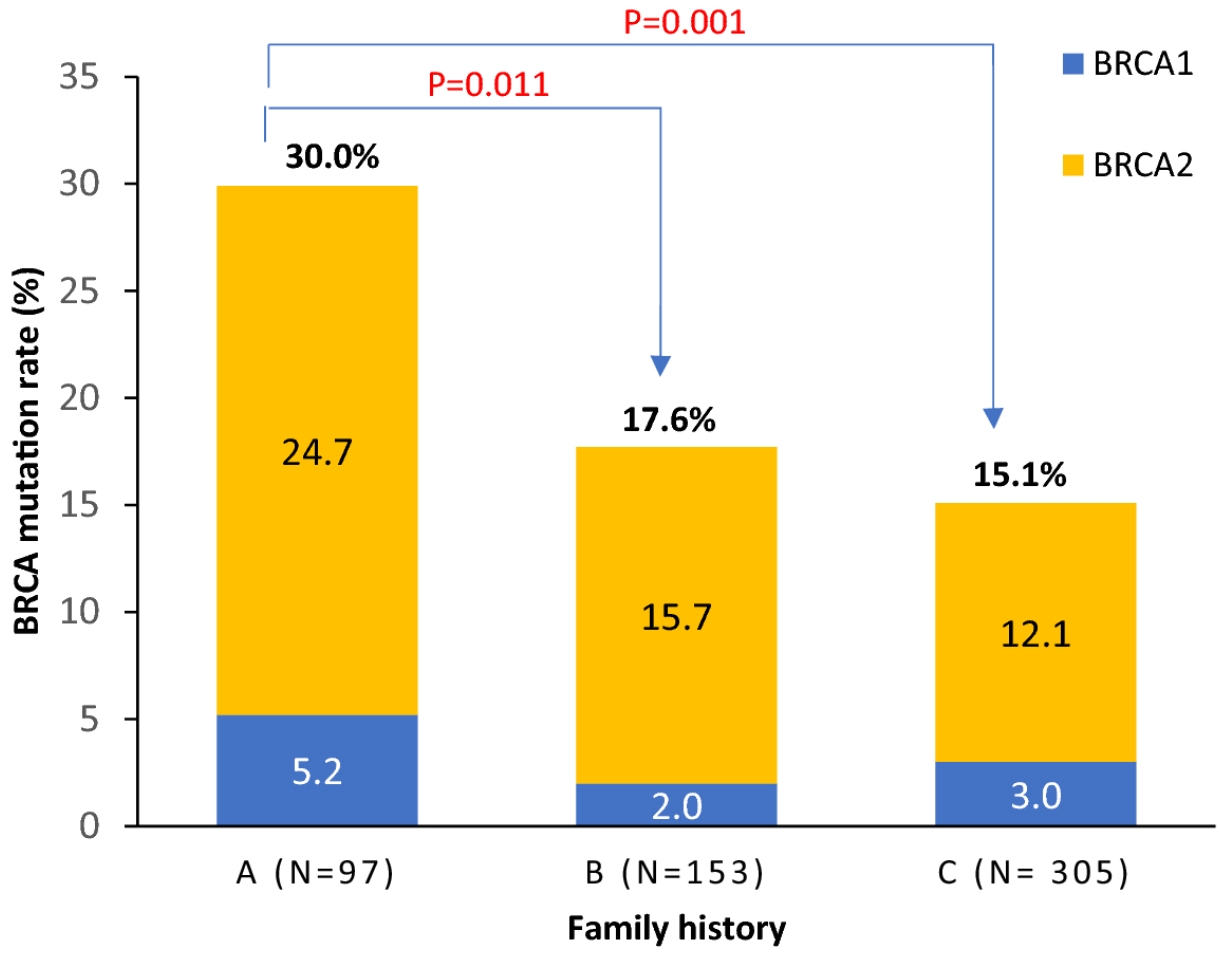


Figure 126: BCR-ABL과 RAS 경로



A: Two or more close relatives with breast cancer

B: One or more close relatives with breast cancer diagnosed at age 50 or younger

C: One or more close relatives with breast cancer at any age

Figure 127: BRCA와 유방암

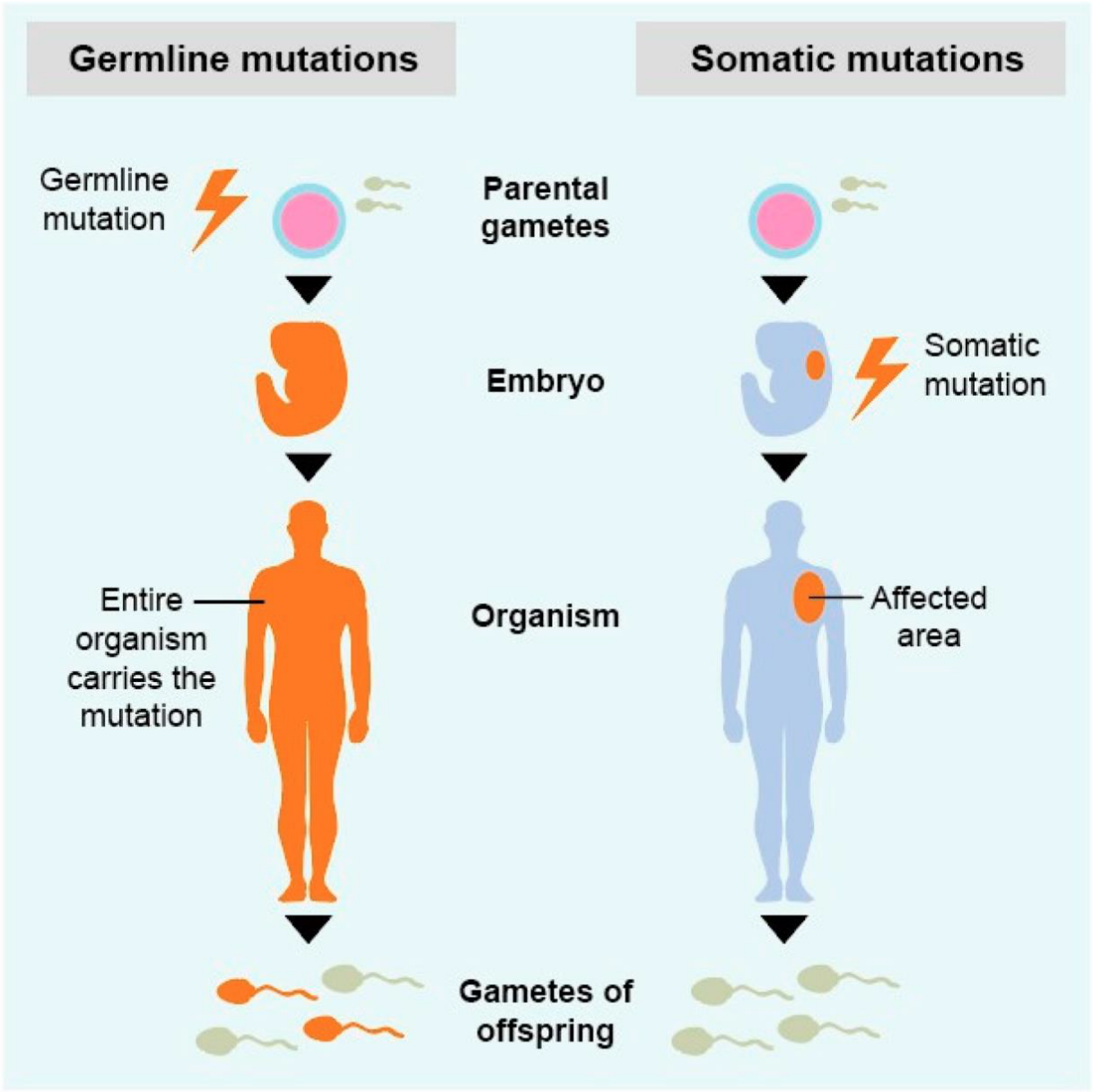


Figure 128: 체세포와 생식세포 변이

Figure 6.13 체세포와 생식세포 변이 – 체세포 변이와 생식세포 변이의 차이점과 유전 패턴

7.8.1 작은(또는 단순한) 돌연변이: SNV, Indel

가장 기본적인 형태의 체세포 돌연변이는 다음과 같다:

- **SNV(Single Nucleotide Variant)**: DNA 서열에서 하나의 염기가 다른 염기로 바뀌는 현상으로, 예를 들어 A가 T로 변하는 경우를 말한다. 이러한 변화는 단백질의 아미노산 서열을 바꾸어 단백질 기능에 영향을 미칠 수 있다.

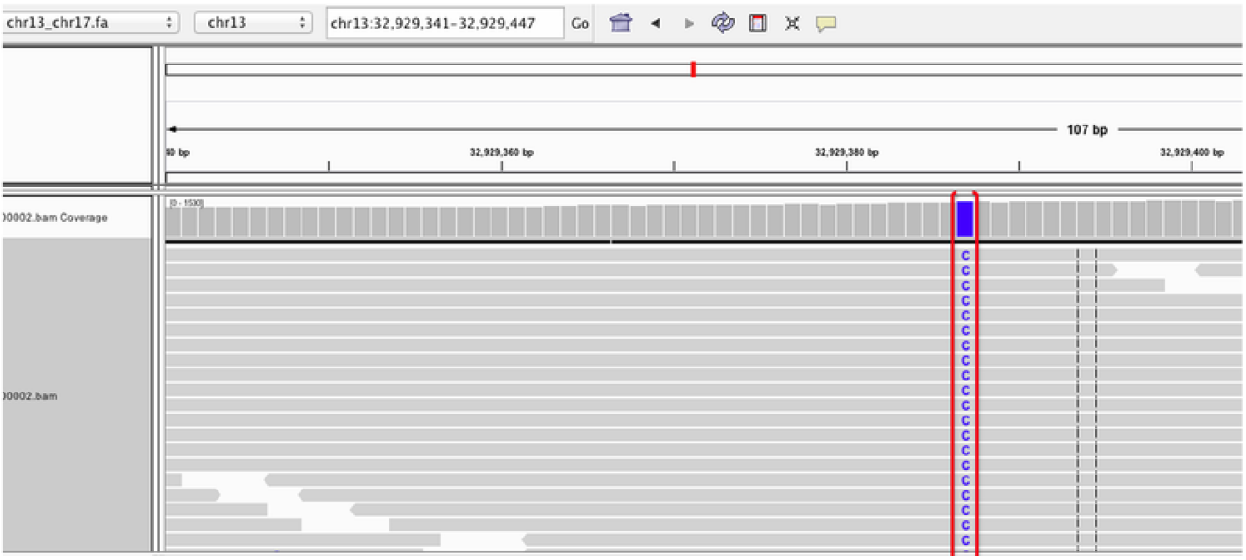


Figure 129: SNV 예시

Figure 6.14 SNV 예시 – 단일 염기 변이의 다양한 유형과 단백질에 미치는 영향

- **Indel**: 작은 규모의 삽입(insertion)이나 결실(deletion)로, 일반적으로 수십 염기 이하의 크기를 가진다. 이들은 유전자의 읽기틀을 변경하여 완전히 다른 단백질을 만들어낼 수 있다.

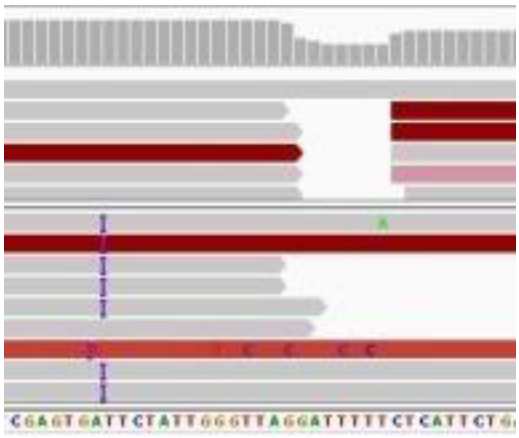


Figure 130: Indel 예시

Figure 6.15 Indel 예시 – 작은 삽입 및 결실 변이와 읽기틀 변화에 의한 영향

7.8.2 Rainfall plot

Rainfall plot은 변이들 간의 거리를 시각화하여 특정 영역에 변이가 집중되는 현상인 kataegis를 식별하는 그래프이다. Kataegis는 특정 DNA 복구 기전의 오작동으로 인해 발생하는 현상으로, 암의 발생 과정을 이해하는 중요한 단서를 제공한다.

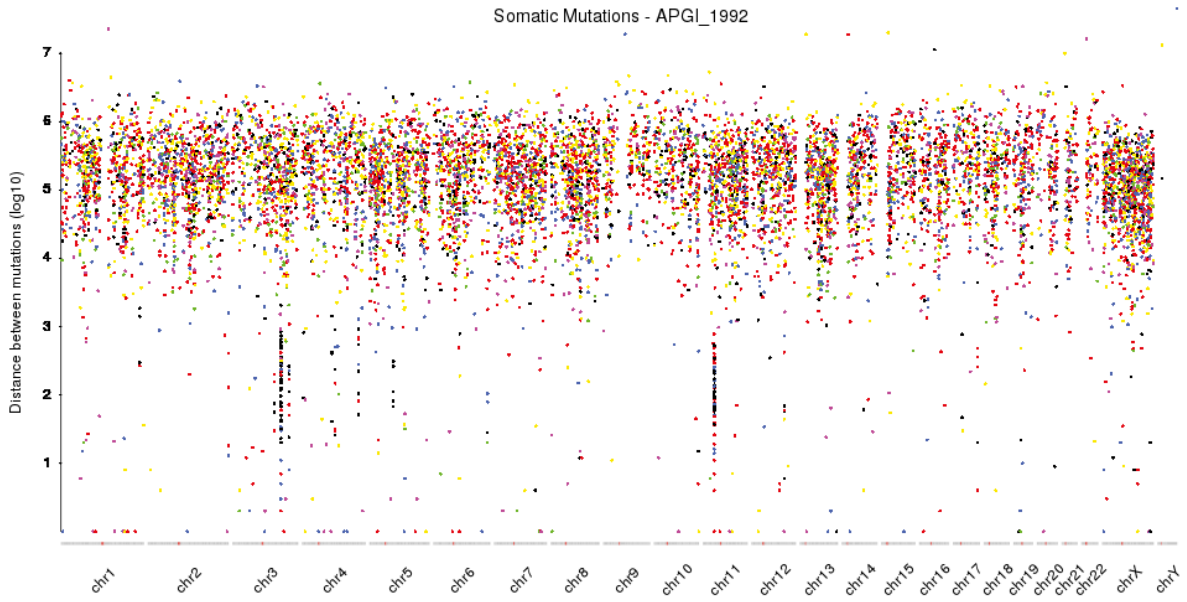


Figure 131: Rainfall plot

Figure 6.16 Rainfall plot – 변이 간 거리를 시각화하여 kataegis 현상을 식별하는 그래프

7.8.3 CNV(Copy Number Variation)

복제수 변이는 특정 DNA 영역이 정상보다 많거나 적게 존재하는 현상이다. 암에서는 종양억제유전자가 있는 영역이 결실되거나, 암유전자가 있는 영역이 증폭되는 경우가 흔히 관찰된다.

- **Focal gain/loss**: 특정 유전자나 작은 영역에 국한된 복제수 변화를 의미하며, 이는 해당 유전자의 발현량에 직접적인 영향을 미친다.
- **LOH(Loss of Heterozygosity)**: 정상적으로 두 개의 서로 다른 대립유전자를 가져야 하는 부위에서 하나의 대립유전자가 소실되는 현상이다. 이는 특히 종양억제유전자에서 중요한데, 한 개의 대립유전자만으로는 정상적인 기능을 수행하기 어려워 암 발생에 기여할 수 있다.

Figure 6.17 국소 증폭과 LOH – 국소적 복제수 증가와 이형접합자 소실의 예시

7.8.4 큰(또는 복잡한) 돌연변이: 대규모 Indel, 구조적 변이(SV)

가장 복잡한 형태의 체세포 돌연변이는 다음과 같다:

- **대규모 Indel**: 수천에서 수백만 염기에 이르는 큰 DNA 조각이 삽입되거나 결실되는 현상이다.
- **구조적 변이**: 염색체의 구조가 근본적으로 바뀌는 변화로, CNV, 역위(inversion), 전위(translocation) 등이 포함된다.

Figure 6.18 복제수 이상 – 다양한 복제수 변이와 ecDNA 형성

- **ecDNA(Extrachromosomal DNA)**: 염색체에서 분리되어 독립적으로 존재하는 원형 DNA 구조로, 주로 암유전자가 고도로 증폭된 형태로 나타난다. ecDNA는 염색체보다 더 불안정하고 빠르게 진화할 수 있어 암의 진행과 치료 저항성에 중요한 역할을 한다.

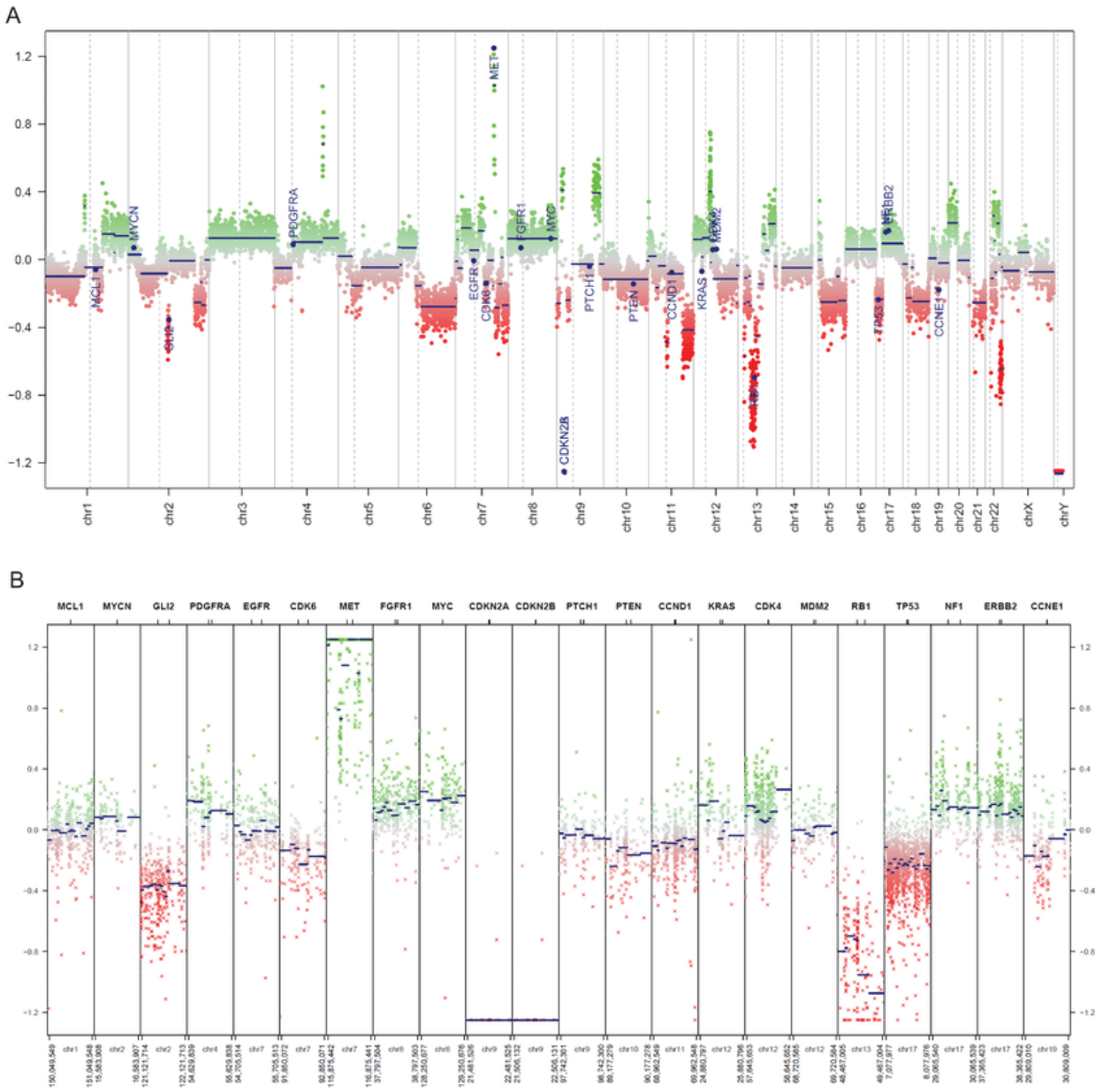


Figure 132: 국소 증폭과 LOH

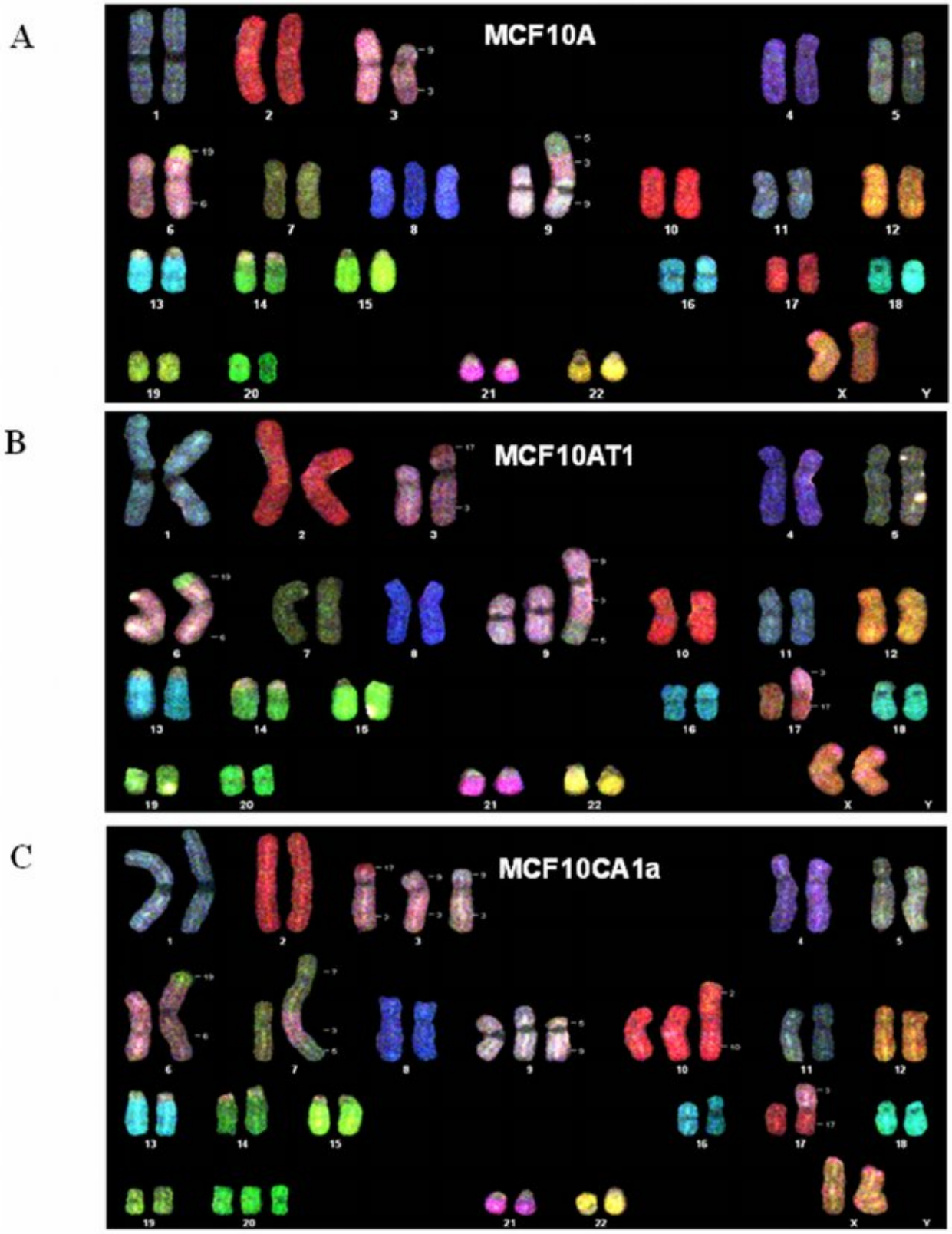


Figure 133: 복제수 이상

이러한 대규모 변이들은 여러 유전자에 동시에 영향을 미칠 수 있어 암세포의 특성을 크게 바꿀 수 있다. 또한 이들은 종종 염색체 불안정성과 연관되어 있어, 암의 진행과 전이에 중요한 역할을 한다.

7.9 암 변이 데이터베이스

암 연구에서는 대규모 유전체 데이터를 체계적으로 수집하고 분석하기 위해 다양한 데이터베이스와 도구가 개발되었다. 이러한 자원들은 암의 분자적 특성을 이해하고 치료 전략을 개발하는 데 필수적인 역할을 한다.

7.9.1 주요 암 데이터베이스

- **COSMIC(Catalogue of Somatic Mutations in Cancer)**: 전 세계에서 수집된 암의 체세포 돌연변이에 관한 가장 포괄적인 정보를 제공하는 데이터베이스이다. 수백만 개의 변이 정보와 함께 변이 시그니처, 약물 저항성 정보 등을 포함한다.
- **ICGC(International Cancer Genome Consortium)**: 다양한 암 유형의 유전적, 후성유전적, 전사체적 특성을 연구하는 국제 협력 프로젝트로, 50개 이상의 암 유형에 대한 포괄적인 분자 정보를 제공한다.

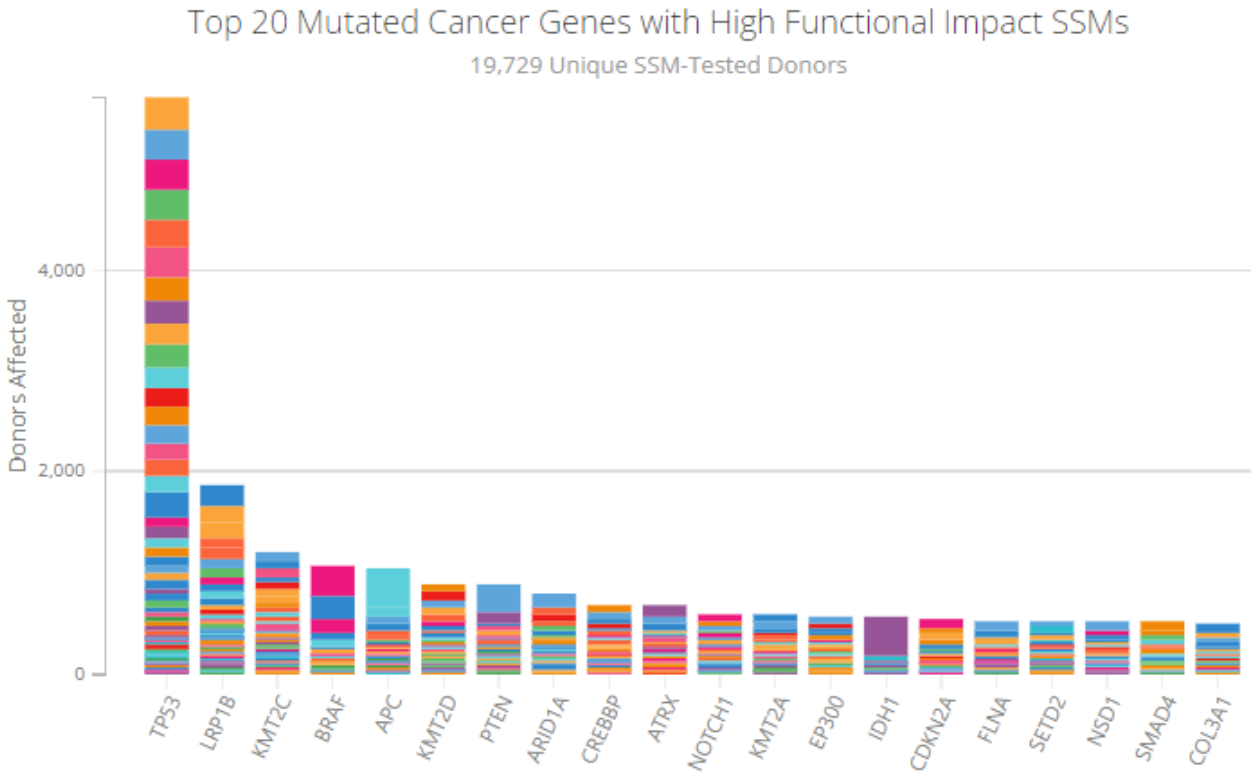


Figure 134: ICGC 주요 결과

Figure 6.19 ICGC 주요 결과 – 국제 암 유전체 컨소시엄의 주요 연구 성과

- **TCGA(The Cancer Genome Atlas)**: 33개 암 유형에 대해 11,000명 이상의 환자 샘플을 분석한 대규모 연구 프로젝트로, 유전체, 전사체, 후성유전체, 단백체 데이터를 통합적으로 제공한다.

7.9.2 데이터 시각화 도구

cBioPortal(https://www.cbioportal.org/)은 암 유전체 데이터를 시각화하고 분석할 수 있는 웹 기반 플랫폼이다. 사용자는 특정 유전자나 암 유형에 대한 변이 패턴을 쉽게 검색하고 분석할 수 있다.

7.9.3 Oncoprint

Oncoprint는 여러 환자 샘플에서 다양한 유전적 변이를 시각적으로 표현하는 히트맵 형태의 그래프이다. 각 행은 유전자를, 각 열은 환자 샘플을 나타내며, 색상과 패턴을 통해 변이 유형을 구분한다. 이를 통해 특정 암 유형에서 어떤 유전자가 빈번하게 변이되는지, 그리고 여러 유전자의 변이가 어떻게 함께 나타나는지를 쉽게 파악할 수 있다.

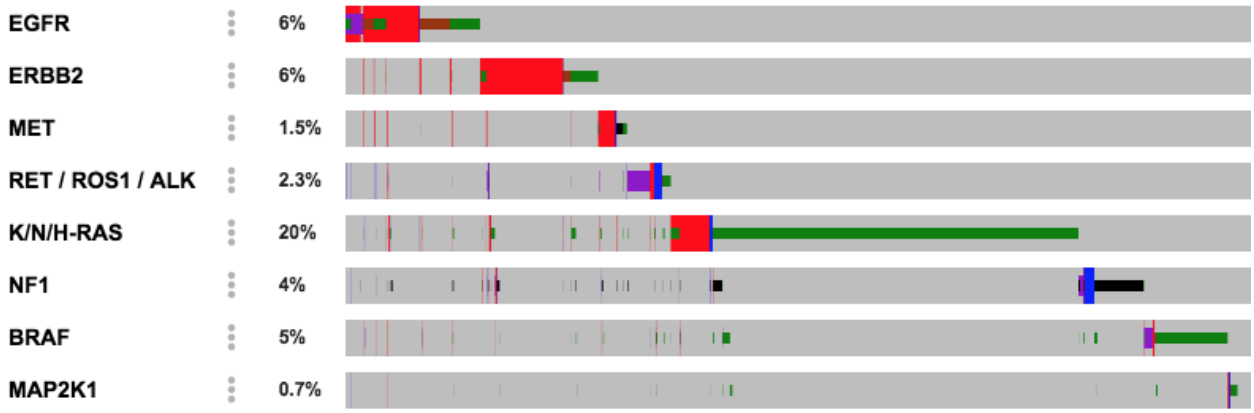


Figure 135: Oncoprint

Figure 6.20 Oncoprint – 여러 환자 샘플에서 다양한 유전적 변이를 시각화한 히트맵

7.10 변이 주석(Variant Annotation)

변이 호출(variant calling) 후, dbSNP, Clinvar, 1000 Genomes, gnomAD 등의 데이터베이스를 참조하여 변이에 주석을 달 수 있다. ANNOVAR는 이러한 변이 주석 작업을 수행하는 도구 중 하나이다.

7.11 종양 순도(Tumor Purity)

종양 순도는 샘플 내 종양 세포의 비율을 의미한다. 체세포 변이 호출 시 종양 순도를 고려하는 것이 중요하다. 순도가 낮을수록 변이 검출이 어려워지므로, 분석 결과 해석 시 이를 고려해야 한다.

Figure 6.21 종양 순도 – 종양 샘플 내 종양 세포의 비율과 변이 검출에 미치는 영향

7.12 암 치료를 위한 정밀 의학

정밀 의학 접근법을 통한 암 치료 과정은 다음과 같다:

1. 대조군(정상 조직)과 종양 조직의 시퀀싱
2. 참조 유전체에 대한 정렬
3. 두 샘플의 변이 호출
4. 변이 제거: (종양 변이) – (대조군 변이) – (dbSNP의 일반적인 변이) = (체세포 돌연변이)
5. 돌연변이 유전자 세트 찾기(특히 rainfall plot에서)
6. 암유전자 조사 및 경로 분석 수행

Figure 6.22 전형적인 암 유전체학 파이프라인 – 정밀 의학을 위한 암 유전체 분석 과정

7.13 암 유전체의 특성

7.13.1 종양 이질성(Tumor Heterogeneity)

종양 이질성은 암 조직 내에서 세포들이 서로 다른 유전적, 후성유전적 특성을 보이는 현상이다. 이는 암 치료의 효과를 제한하는 주요 요인 중 하나이다. 종양 이질성은 크게 두 가지 유형으로 나뉜다.

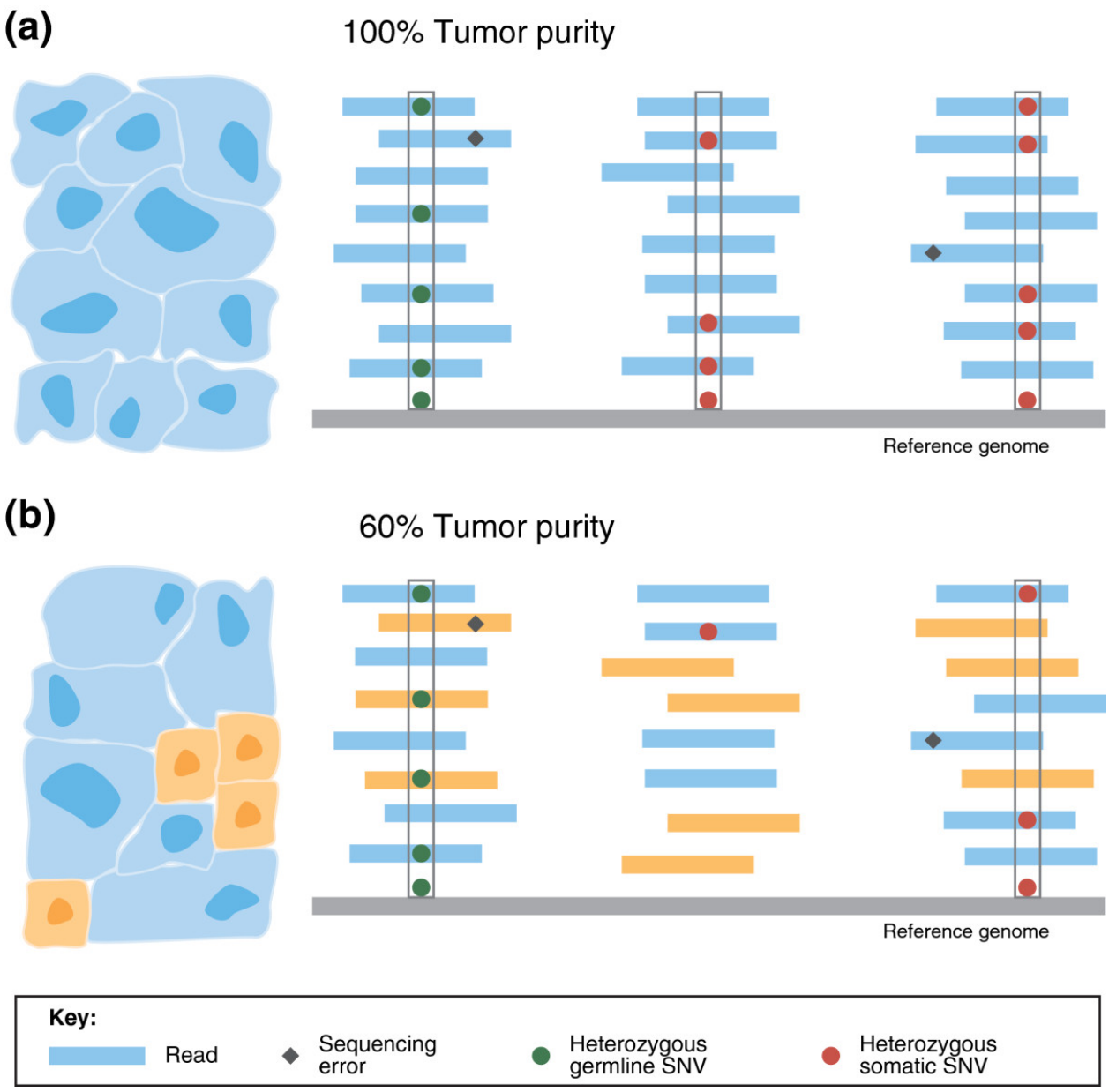


Figure 136: 종양 순도

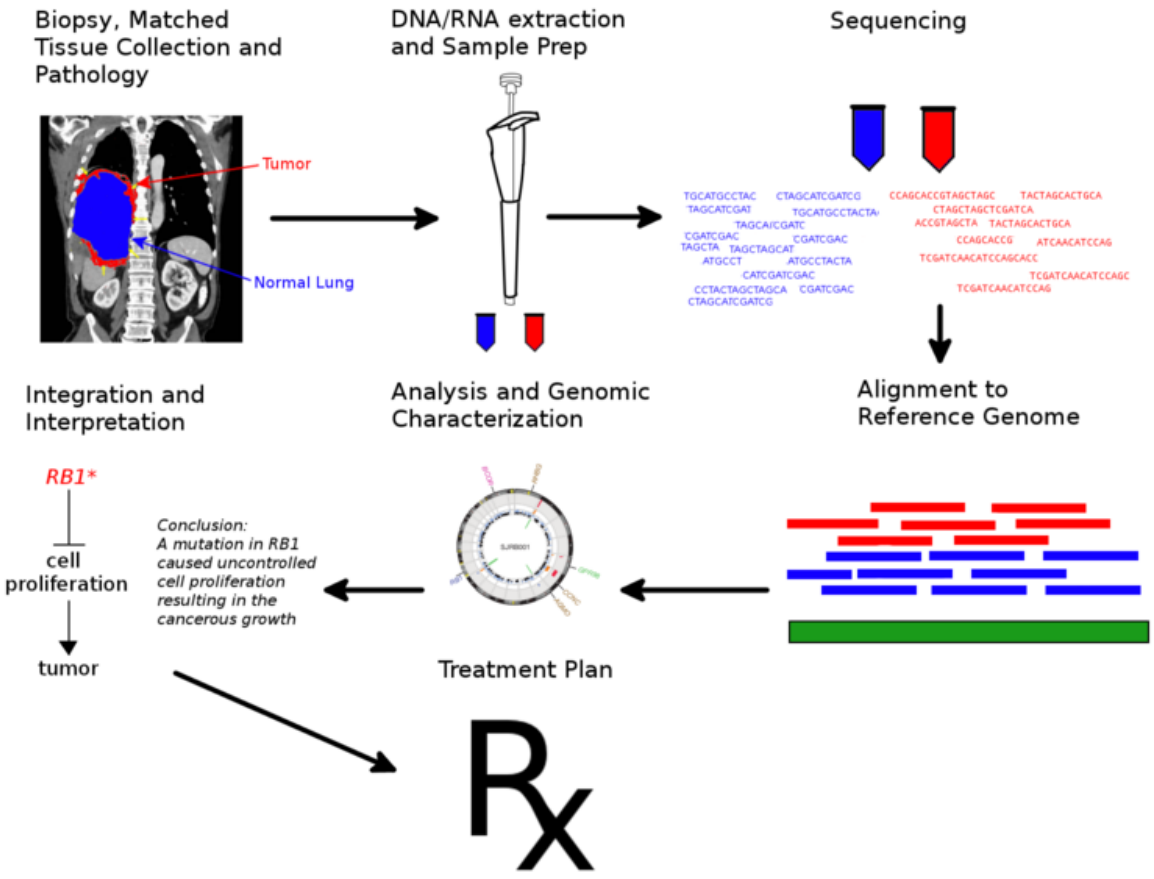


Figure 137: 전형적인 암 유전체학 파이프라인

종양간 이질성(Inter-tumor heterogeneity)은 같은 암 유형이라도 환자마다 다른 유전적 변이 패턴을 보이는 현상이다. 예를 들어, 폐암 환자 A와 B는 모두 폐암이지만 서로 다른 드라이버 돌연변이를 가질 수 있다. 종양내 이질성(Intra-tumor heterogeneity)은 같은 종양 내에서도 부위별로 다른 세포 집단이 존재하는 현상이다. 이로 인해 조직검사로 얻은 일부 샘플만으로는 전체 종양의 특성을 완전히 파악하기 어렵다.

7.13.2 클론 진화(Clonal Evolution)

클론 진화는 암이 발생하고 진행되는 과정에서 암세포들이 계속해서 새로운 변이를 획득하며 진화하는 현상이다. 이는 다윈의 자연선택 이론과 유사한 메커니즘으로 작동한다.

정상 세포에서 시작하여 첫 번째 드라이버 돌연변이가 발생하면 전암 병변이 형성된다. 추가적인 돌연변이가 축적되면서 암세포로 변환되고, 치료 과정에서 저항성을 가진 클론이 선택되어 살아남는다. 이 과정은 암의 재발과 전이의 주요 원인이 되며, 정밀의학에서 지속적인 모니터링이 필요한 이유이기도 하다.

7.13.3 변이 시그니처(Mutational Signatures)

변이 시그니처은 특정한 변이 유발 요인이 DNA에 남기는 고유한 변이 패턴을 의미한다. 각각의 변이 유발 요인은 특정한 염기 변화 패턴을 만들어내며, 이를 통해 암의 원인을 추정할 수 있다.

COSMIC(Catalogue of Somatic Mutations in Cancer) 데이터베이스는 암에서 발견되는 체세포 돌연변이들을 체계적으로 분류하고 분석한 결과를 제공한다. 돌연변이 시그니처는 특정한 돌연변이 패턴을 나타내며, 각각은 특정한 돌연변이 유발 과정과 연관되어 있다.

대표적인 변이 시그니처으로는 흡연에 의한 G>T 변이 증가, 자외선에 의한 C>T 변이 증가, APOBEC 효소에 의한 C>G 변이 증가 등이 있다. COSMIC 데이터베이스에서는 현재까지 약 60여 개의 서로 다른 변이 시그니처이 정의되어 있으며, 이는 암의 병인을 이해하고 치료 전략을 수립하는 데 중요한 정보를 제공한다.

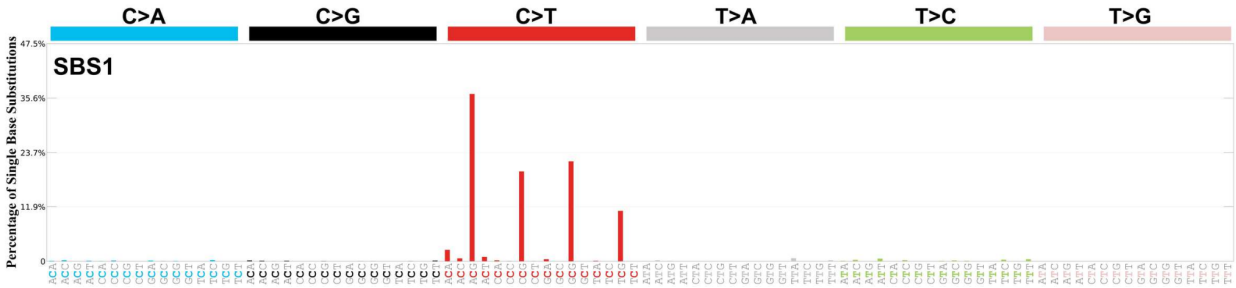


Figure 138: 변이 시그니처

Figure 6.23 변이 시그니처 – 특정 변이 유발 요인이 남기는 고유한 변이 패턴

돌연변이 시그니처 분석은 단순히 돌연변이 패턴을 파악하는 것을 넘어 임상적으로 매우 중요한 정보를 제공한다. 특정 화학물질이나 DNA 수선 기전의 결함과 연관된 고유한 시그니처를 통해 암의 근본 원인을 추정할 수 있다. 이러한 분석 결과는 치료에도 활용된다. DNA 수선 결함의 종류에 따라 특정 표적치료제의 효과가 다를 수 있어, 시그니처 분석을 통해 개인 맞춤형 치료 전략을 수립할 수 있다. 또한 환자의 돌연변이 부담 정도와 시그니처 패턴을 분석하여 치료 반응성이나 예후를 예측할 수 있다. 환경적 요인들도 특징적인 시그니처를 남기므로, 과거의 환경 노출 이력을 추정하고 예방 전략을 세우는 데 활용할 수 있다.

7.13.4 마이크로위성체 불안정성(Microsatellite Instability, MSI)

마이크로위성체 불안정성은 DNA 복구 시스템의 결함으로 인해 마이크로위성체 서열에서 반복적인 삽입/결실이 발생하는 현상이다. MSI는 주로 mismatch repair (MMR) 유전자의 기능 상실로 인해 발생한다.

MSI-High (MSI-H) 종양은 면역치료에 대한 반응성이 높은 것으로 알려져 있다. 이는 MMR 결함으로 인해 비정상적인 단백질이 많이 생성되어 면역계가 이를 외부 항원으로 인식하기 때문이다. 대장암의 약 15%, 자궁내막암의 약 30%에서 MSI-H가 관찰되며, 이러한 환자들은 PD-1/PD-L1 억제제와 같은 면역치료제에 좋은 반응을 보인다.

Type of microsatellite	Example sequence	Shorthand nomenclature
Uninterrupted	CCCCCC CTCTCTCTCTCT CTGCTGCTGCTGCTGCTG	$(C)_6$ $(CT)_6$ or $(TC)_5$ $(CTG)_6$
Interrupted	CCACCC CTCTAACTCTCT CTGCTGCTGAAACTGCTG	$(C)_2A(C)_3$ $(CT)_2AA(CT)_3$ $(CTG)_3AAA(CTG)_2$
Compound	CCCAAA CTCTCTCACACA CTGCTGCTGCAGCAGCAG	$(C)_3(A)_3$ $(CT)_3(CA)_3$ $(CTG)_3(CAG)_3$

Figure 139: 마이크로위성체

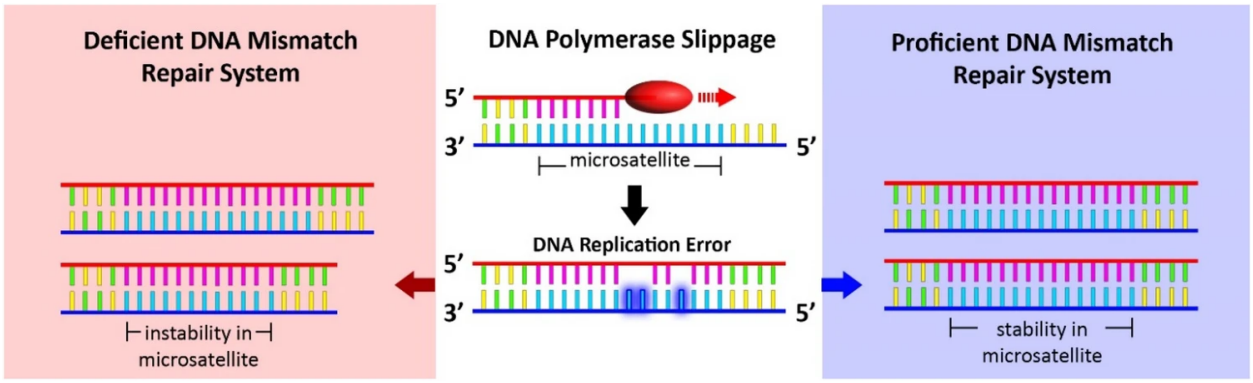


Figure 140: 마이크로위성체 불안정성

Figure 6.24 마이크로위성체 – 마이크로위성체 서열의 구조와 반복 단위

Figure 6.25 마이크로위성체 불안정성 – MSI 현상과 DNA 복구 결함에 의한 영향

7.13.5 종양 변이 부담(Tumor Mutational Burden, TMB)

종양 변이 부담은 암 유전체에서 관찰되는 체세포 변이의 총 개수를 의미한다. 일반적으로 백만 염기쌍당 변이 개수로 표현된다. TMB는 면역치료 효과를 예측하는 중요한 바이오마커로 여겨진다.

TMB가 높은 종양일수록 더 많은 신항원(neoantigen)을 생성할 가능성이 높아 면역계의 인식과 공격을 받기 쉽다. 흑색종, 폐암, 방광암 등은 일반적으로 높은 TMB를 보이며, 이러한 암종에서 면역치료의 효과가 좋은 것으로 알려져 있다. 반면 소아암이나 일부 혈액암은 상대적으로 낮은 TMB를 보인다.

7.14 액체생검과 순환종양DNA

7.14.1 액체생검(Liquid Biopsy)의 개념

액체생검은 혈액, 소변, 타액 등의 체액에서 순환하는 종양 유래 물질을 분석하여 암을 진단하고 모니터링하는 기술이다. 기존의 조직생검이 침습적이고 제한적인 정보만 제공하는 반면, 액체생검은 비침습적이며 종양 전체의 정보를 반영할 수 있다.

액체생검의 주요 분석 대상에는 순환종양DNA(circulating tumor DNA, ctDNA), 순환종양세포(circulating tumor cells, CTCs), 엑소좀(exosomes) 등이 있다. 이 중에서도 ctDNA가 가장 활발히 연구되고 있으며 임상 적용이 확대되고 있다.

7.14.2 순환종양DNA(Circulating Tumor DNA, ctDNA)

ctDNA는 종양세포가 죽으면서 혈액으로 방출되는 짧은 DNA 조각들이다. 이는 종양의 유전적 특성을 반영하므로 암의 진단, 치료 모니터링, 재발 감시 등에 활용될 수 있다.

ctDNA의 양은 일반적으로 전체 순환 DNA의 0.1–10% 정도로 매우 적어 검출에 고감도 기술이 필요하다. 종양의 크기, 위치, 혈관성, 전이 여부 등에 따라 ctDNA의 농도가 달라지며, 일반적으로 전이성 암에서 더 높은 농도로 검출된다.

7.14.3 액체생검의 임상 적용

액체생검은 여러 영역에서 임상적으로 활용되고 있다. 조기 진단 영역에서는 무증상 고위험군에서 암을 조기에 발견하기 위한 스크리닝 도구로 연구되고 있다. 치료 선택에서는 표적치료제의 적응증을 판단하기 위한 동반진단으로 사용된다.

치료 모니터링에서는 치료 중 ctDNA 농도 변화를 통해 치료 효과를 실시간으로 평가할 수 있다. 재발 감시에서는 수술 후 잔존 질병이나 재발을 조기에 탐지하는 데 활용된다. 또한 치료 저항성 획득 과정을 모니터링하여 치료 전략을 조정하는 데도 도움이 된다.

액체생검은 비침습적이고 반복적인 검사가 가능하여 종양의 시간적 변화를 추적할 수 있다는 장점이 있다. 그러나 아직 기술적 한계와 표준화 문제가 있어 지속적인 연구와 개발이 필요한 상황이다.

7.15 유전체학 기반 정밀의학

전통적인 암 치료는 주로 세포독성 항암제에 의존했다. 플라틴 계열 항암제(카보플라틴, 시스플라틴)는 DNA에 손상을 가해 빠르게 분열하는 암세포를 공격하지만, 동시에 정상 세포도 손상시켜 심각한 부작용을 야기했다. 그러나 암 유전체학의 발전은 이러한 “모든 암에 동일한 치료” 접근법에서 벗어나 “암의 분자적 특성에 맞춘 정밀치료”로의 혁명적 전환을 가능하게 했다.

Figure 6.26 플라티늄 화학요법 – 전통적인 세포독성 항암제의 작용 메커니즘

Figure 6.27 정밀 의학 – 변이 필터링을 통한 임상적으로 유의한 변이 발굴 워크플로우

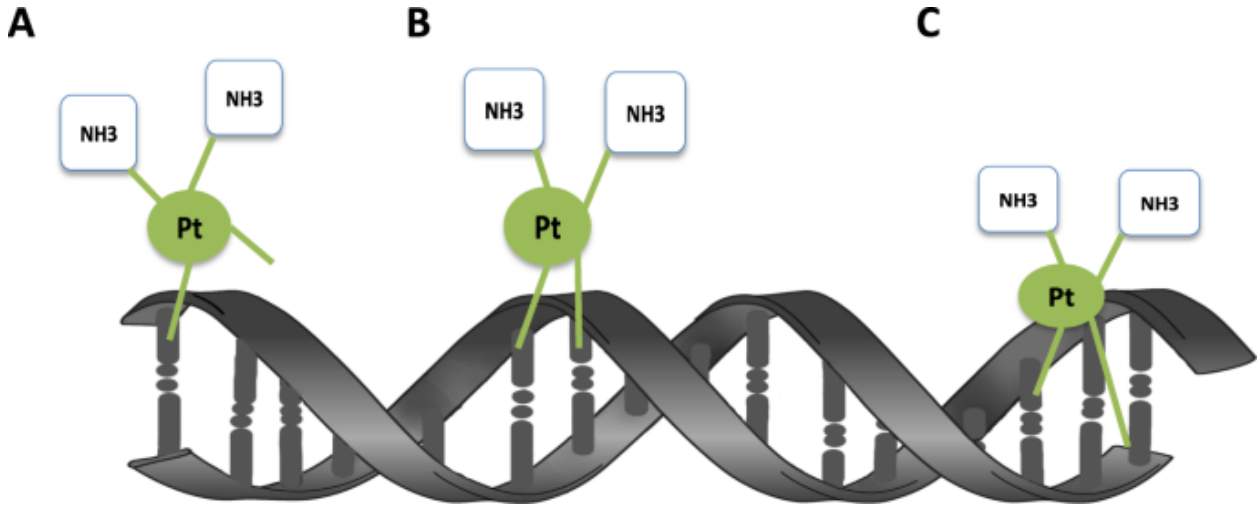


Figure 141: 플라티늄 화학요법

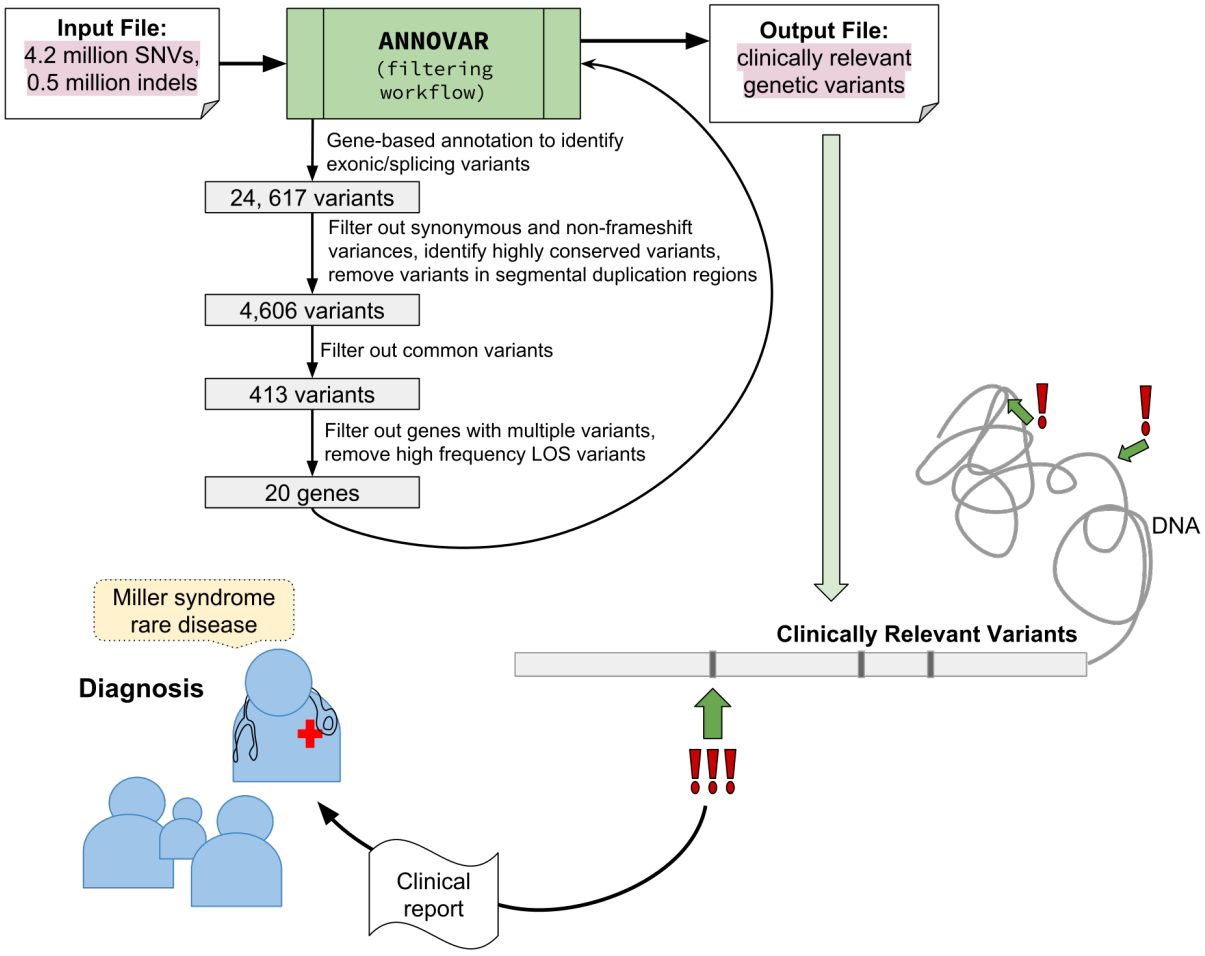


Figure 142: 정밀 의학

7.15.1 표적치료

암 유전체 시퀀싱을 통해 각 암의 고유한 유전적 변화가 밝혀지면서, 이러한 변화를 특이적으로 표적하는 치료법이 개발되었다. 이마티닙의 성공은 이러한 접근법의 가능성을 극명하게 보여준다. BCR-ABL 융합 단백질만을 선택적으로 차단하는 이 약물은 만성 골수성 백혈병 환자의 5년 생존율을 30%에서 90% 이상으로 끌어올렸으며, 동시에 기존 화학요법의 심각한 부작용은 크게 줄였다.

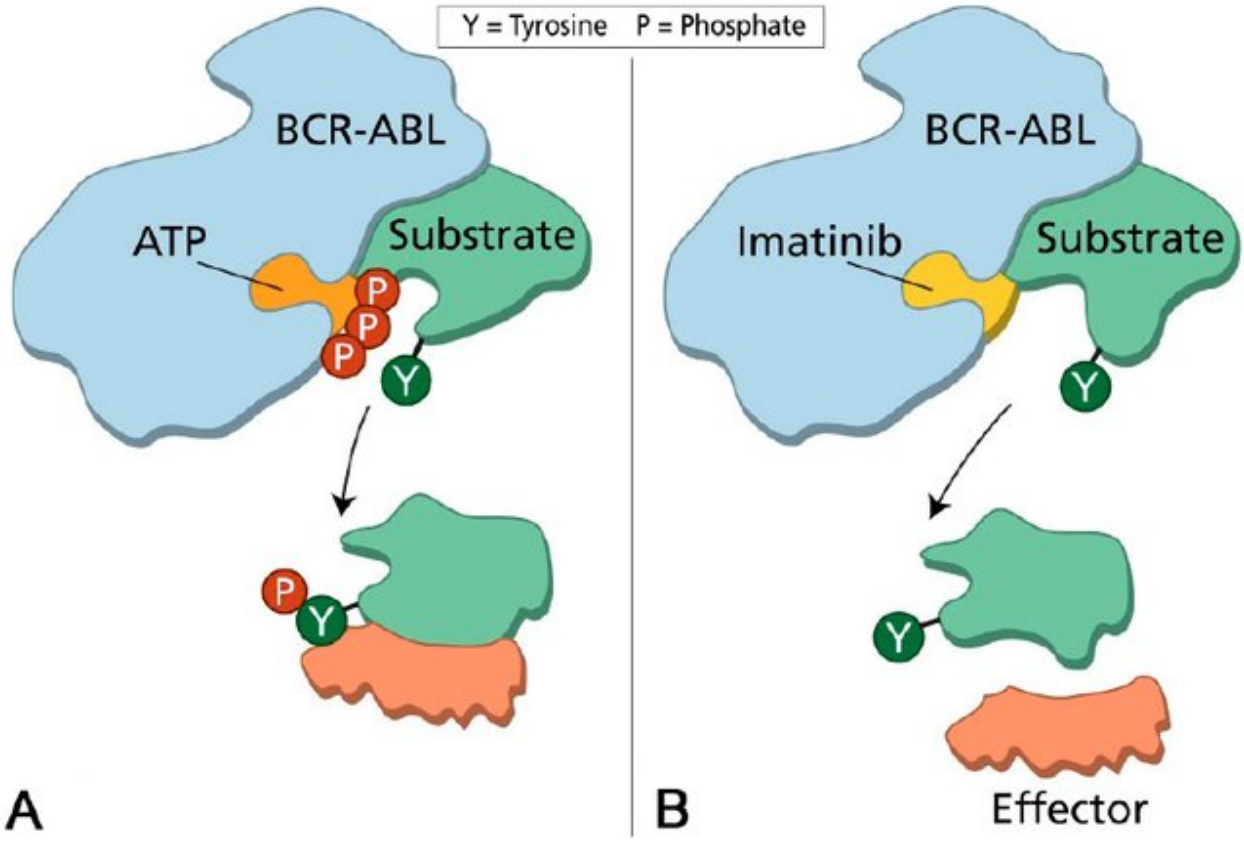


Figure 143: BCR-ABL과 이마티닙

Figure 6.28 BCR-ABL과 이마티닙 – 표적치료제 이마티닙의 BCR-ABL 단백질 선택적 차단

이러한 성공에 힘입어 EGFR 변이를 표적하는 게피티닙, HER2 증폭을 겨냥하는 트라스투주맙, BRAF 변이를 차단하는 베무라페닙 등 다양한 표적치료제가 개발되었다. 각각은 특정 유전적 변화를 가진 환자에게만 효과적이지만, 해당 환자들에게는 기존 치료법보다 훨씬 우수한 결과를 보여준다.

7.15.2 면역치료

암 유전체학의 또 다른 혁명적 발견은 면역치료의 토대를 마련했다. DNA 복구 결함으로 인해 높은 돌연변이 부담(TMB)을 가지거나 마이크로위성체 불안정성(MSI-H)을 보이는 종양은 수많은 신항원을 생성한다. 이러한 신항원은 면역계가 암세포를 "외부 침입자"로 인식할 수 있는 표지가 되며, 이는 면역관문 억제제의 효과를 극대화한다.

펨브롤리주맙과 같은 PD-1 억제제는 특히 MSI-H 대장암이나 높은 TMB를 가진 흑색종에서 놀라운 효과를 보인다. 일부 환자에서는 완전관해가 수년간 지속되는 결과를 보여주어, "암 완치"라는 희망을 현실로 만들고 있다.

Figure 6.29 PD-L1 기적 – 면역관문 억제제의 혈색종 치료 성과

이와 함께 CAR-T 세포 치료법과 같은 세포 치료법도 혈액암 치료에서 혁신적인 결과를 보이고 있다. CRISPR 기술을 이용한 유전자 편집 치료법도 암 치료의 새로운 가능성을 열어주고 있다.

Figure 6.30 CAR-T 세포 치료 – 유전자 조작된 T세포를 이용한 맞춤형 면역치료

카터 전 대통령, 면역항암제로 흑색종 치료

안경진 기자 | 입력 2015.12.08 13:19 | 수정 2015.12.09 09:02 | 댓글 0

MSD키트루다 '카터효과'로 또한번 주목

▲ 지미카터 전미 대통령(출처: 페이스북)

지미 카터(Jimmy Carter) 전 미국 대통령의 '흑색종 완치' 소식이 화제다.

지난 8월 흑색종이 간에 이어 뇌까지 전이됐다고 밝혔던 카터 전 대통령이 6일(현지시각) "최근 시행한 뇌 MRI 결과, 암이 사라졌다(cancer free). 펨브롤리주맙을 이용한 3주간의 정규 면역치료를 지속할 계획"이라는 내용의 성명을 발표한 것.

물론 의학적으로는 완치보다는 항암치료 후 시행한 검사상 암세포가 관찰되지 않는 상태란 의미의 '완전관해(complete remission)'가 더 정확한 표현이겠다. 덕분에 그가 투여받았다는 면역치료제 키트루다(성분명 펨브롤리주맙)가 이른바 '카터효과'로 홍보효과를 톡톡히 누리고 있다.

Figure 144: PD-L1 기적

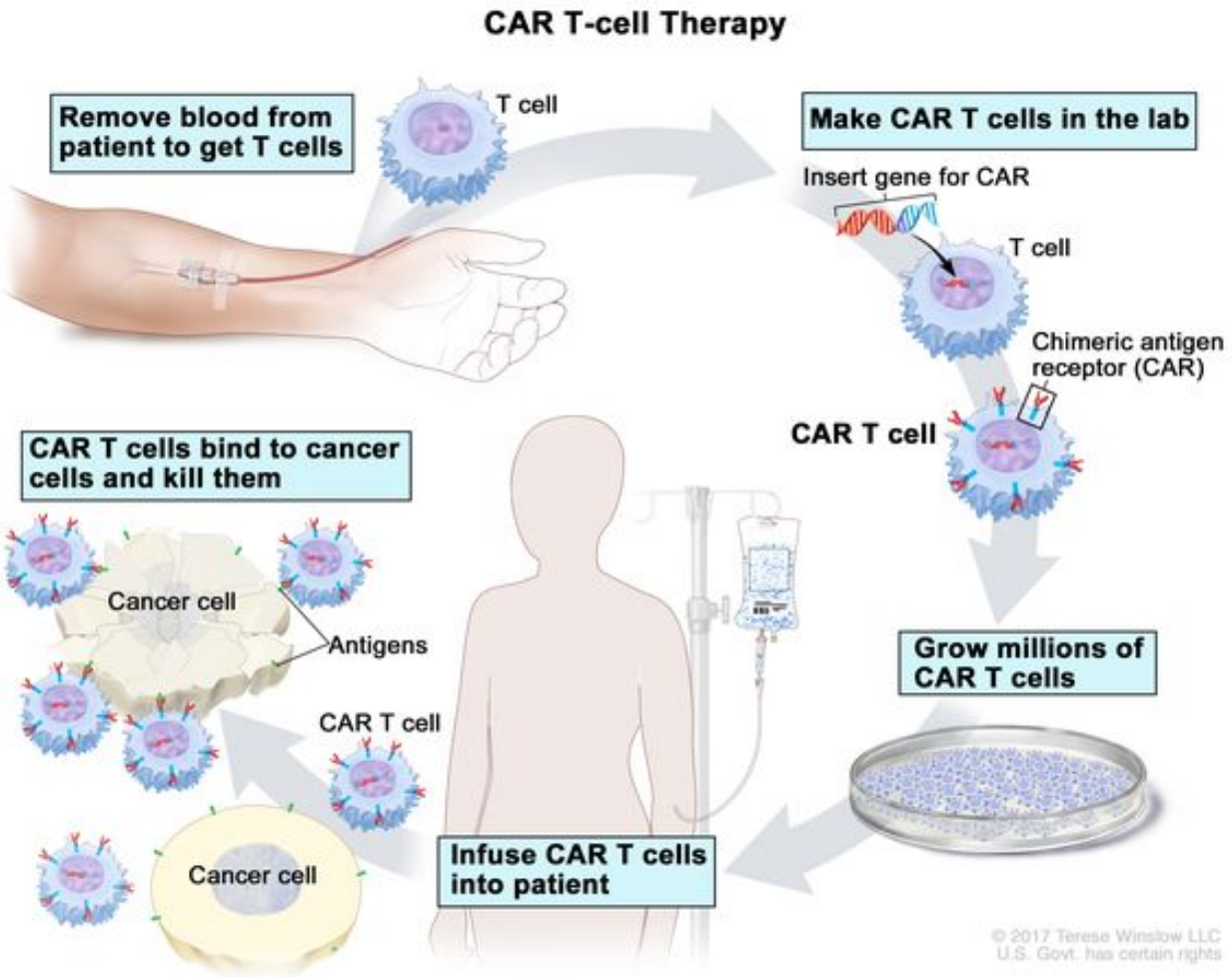


Figure 145: CAR-T 세포 치료

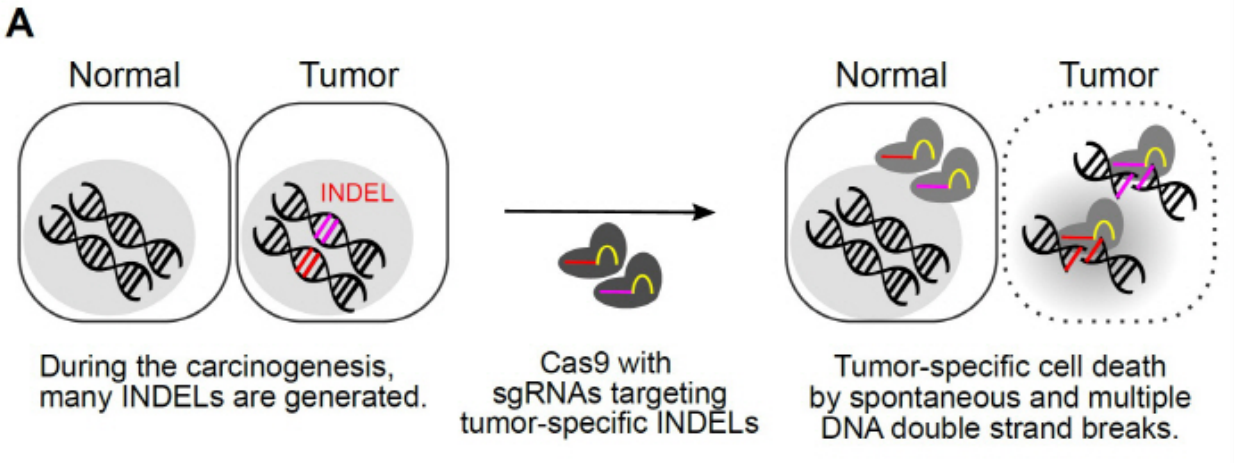


Figure 146: CRISPR 항암 치료

Figure 6.31 CRISPR 항암 치료 – CRISPR 기술을 이용한 유전자 편집 치료법

7.15.3 차세대 치료 기술

암 유전체학의 발전은 전통적인 약물 개발의 한계를 뛰어넘는 혁신적 기술들을 탄생시켰다. 항체-약물 접합체(ADC)는 암세포 표면의 특정 항원을 인식하는 항체에 강력한 세포독성 약물을 연결한 "스마트 미사일"이다. 이는 정상 세포에는 영향을 주지 않으면서 암세포만을 선택적으로 공격할 수 있게 한다.

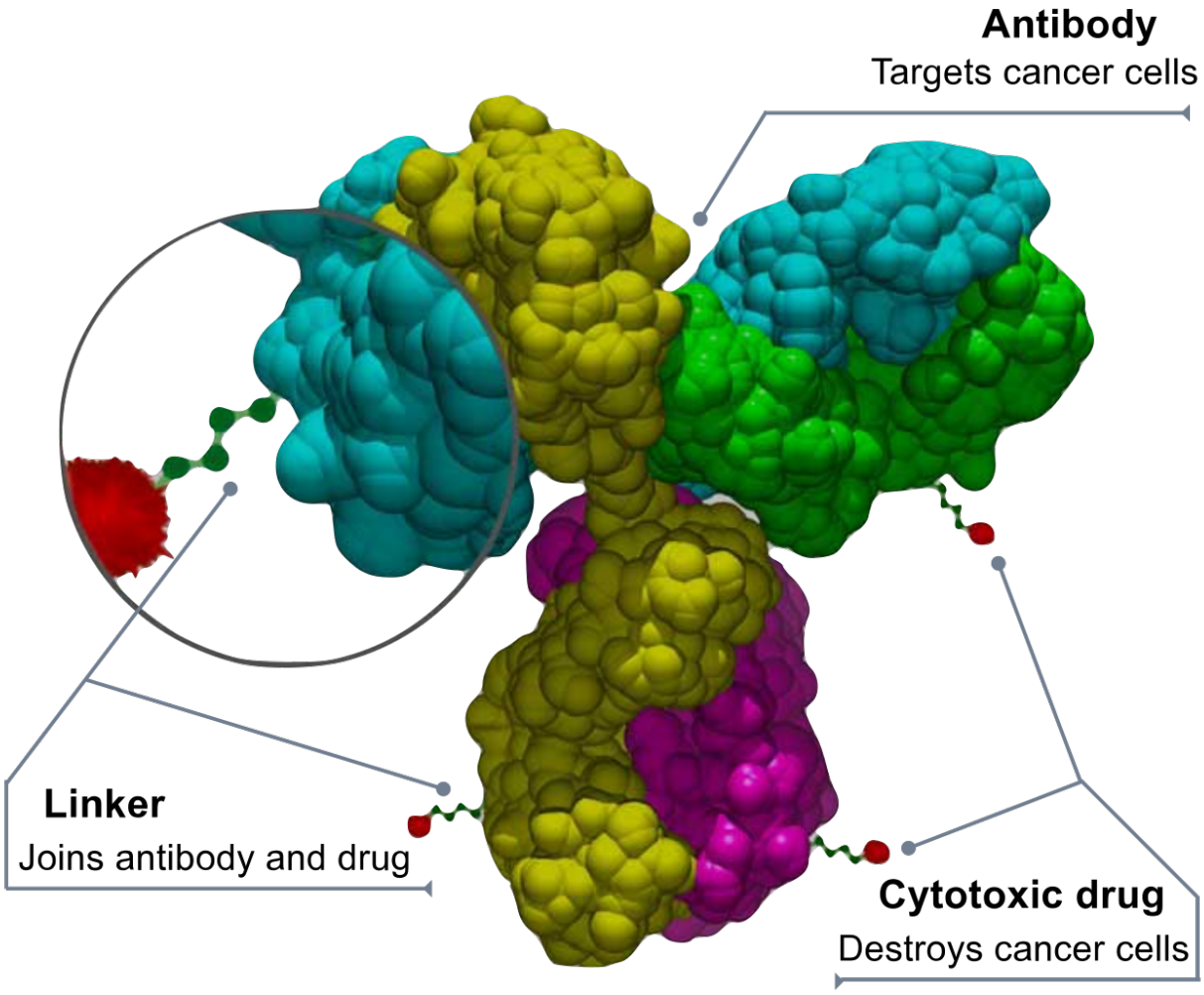


Figure 147: ADC 구조

Figure 6.32 ADC 구조 – 항체-약물 접합체의 구조와 작용 메커니즘

환자별 종양 엑솜 시퀀싱을 통해 개인화된 신항원을 예측하고 맞춤형 암 백신을 제작하는 기술도 주목할 만한 발전이다. 각 환자의 고유한 변이 패턴을 분석하여 해당 환자만의 신항원 펩타이드를 합성하고, 이를 백신으로 활용하여 특이적인 T세포 반응을 유도한다. 이는 암 유전체학이 개인 맞춤형 치료의 구체적인 현실로 구현되는 과정을 보여주며, 임상시험에서 유망한 결과를 보이고 있다.

Figure 6.33 신항원 백신 – 개인 맞춤형 신항원 백신의 제작과 작용 원리

단백질 분해 표적 키메라(PROTAC) 기술은 또 다른 혁신적 접근법을 제시한다. 기존의 억제제가 문제 단백질의 활성만을 차단했다면, PROTAC은 해당 단백질을 세포 내에서 완전히 제거한다. 이는 기존에 "타겟하기 어려운(undruggable)" 단백질들도 치료 표적으로 만들 수 있는 가능성을 열어주었다. 특히 전사인자나 구조 단백질처럼 전통적인 약물로는 접근하기 어려웠던 표적들이 이제 치료 가능한 영역으로 들어오고 있다.

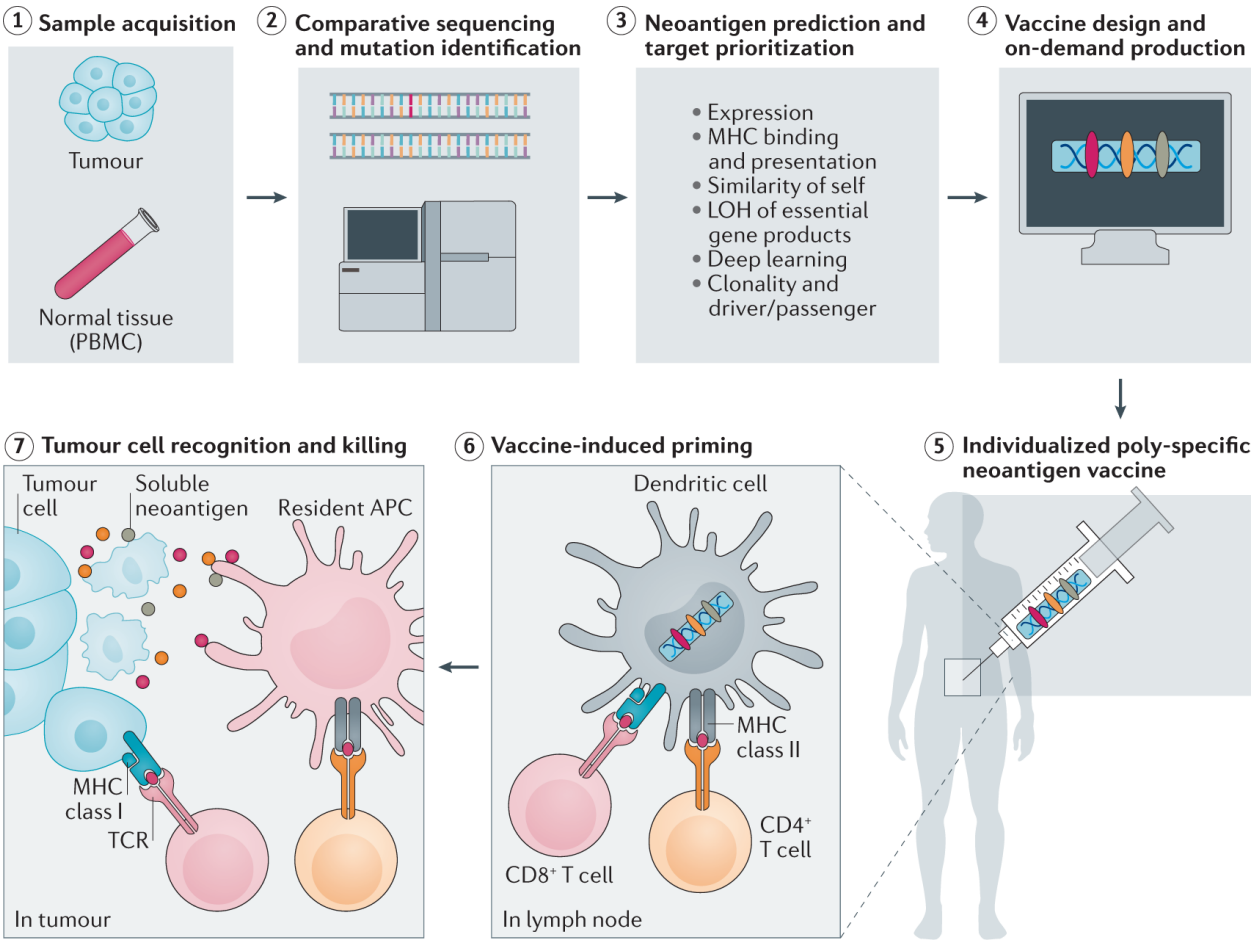


Figure 148: 신항원 백신

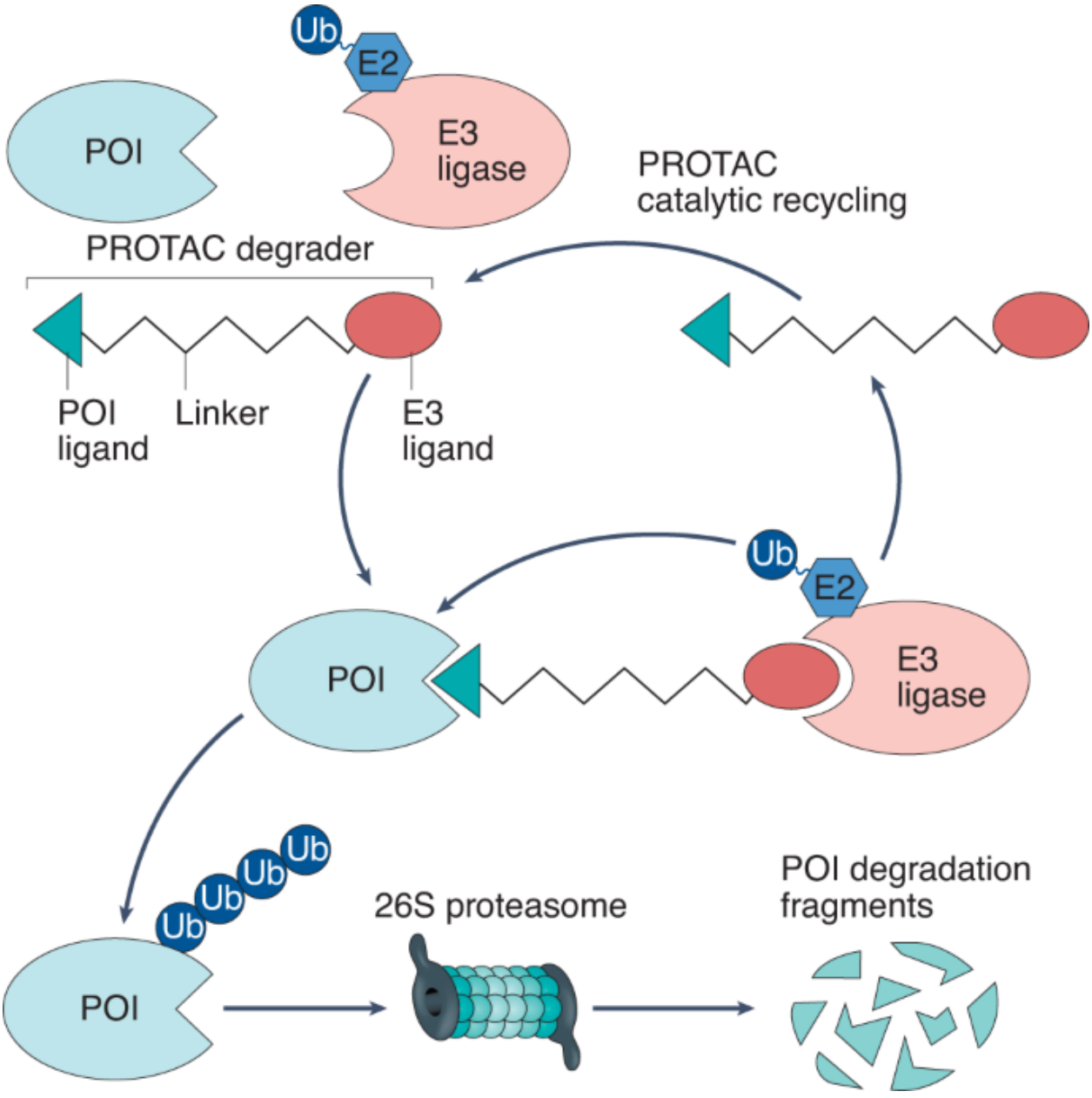


Figure 149: PROTAC 기술

Figure 6.34 PROTAC 기술 – 단백질 분해 표적 키메라 기술의 원리

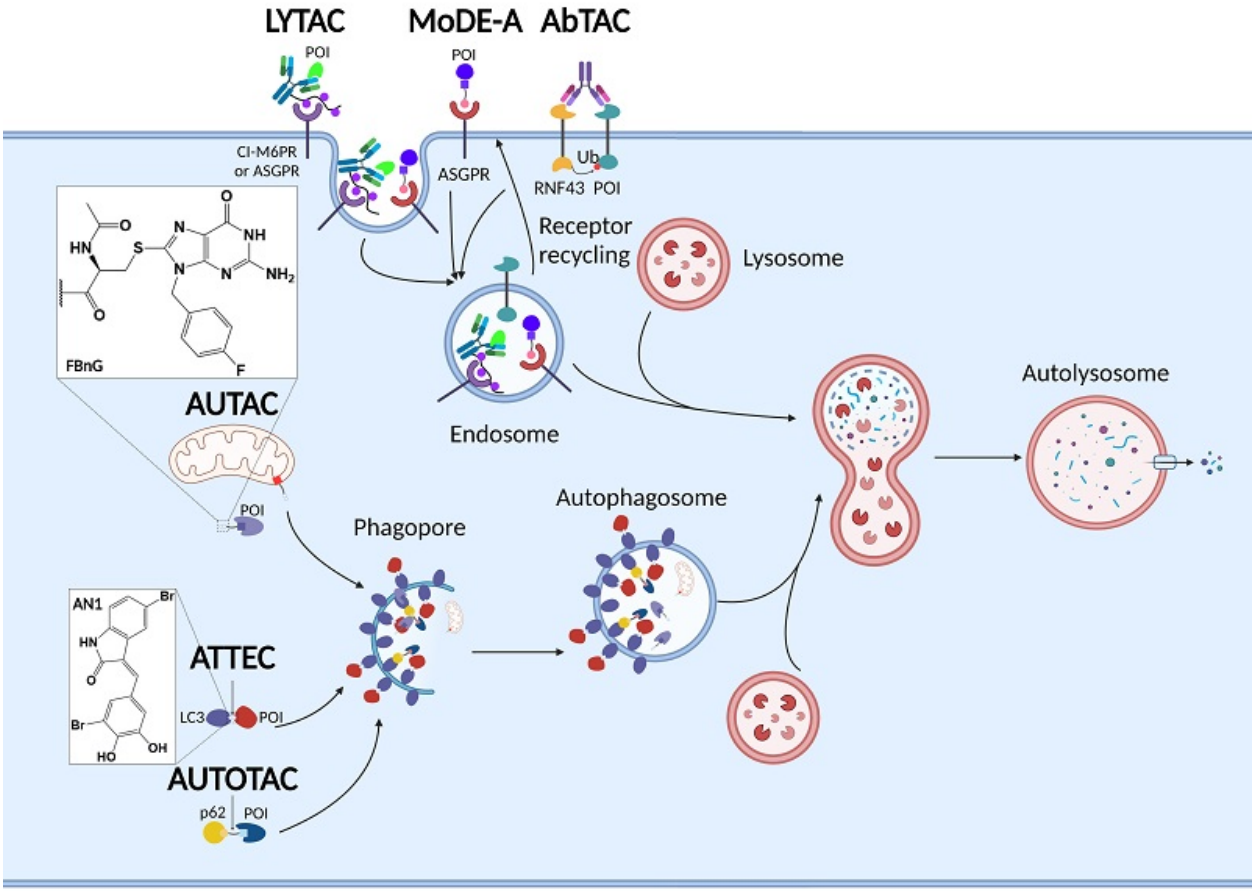


Figure 150: 다양한 TAC 기술

Figure 6.35 다양한 TAC 기술 – PROTAC 외에 개발된 다양한 표적 단백질 분해 기술들

7.16 암과 노화

암 연구와 노화 연구는 놀라운 유사성을 보인다. 두 분야 모두 DNA 손상과 복구 메커니즘, 세포 주기 조절, 후성유전학적 변화, 텔로미어 역학 등의 공통된 분자 기전을 탐구한다.

Figure 6.36 노화와 암 – 노화와 암 연구의 많은 공통점과 분자적 메커니즘

세포노화는 암과 많은 분자 기전을 공유한다. 두 현상 모두 DNA 손상, p53 활성화, 텔로미어 단축 등의 동일한 경로를 거친다.

Figure 6.37 세포 노화 – 세포 노화를 유발하는 다양한 분자적 메커니즘과 후성유전학적 변화

7.17 결론

암은 복잡하고 다면적인 질병으로, 단일 유전자의 변이가 아닌 다수의 유전적 변이가 축적되어 발생하는 질환이다. 암유전자와 종양억제유전자의 균형이 깨지면서 세포의 정상적인 성장 조절 메커니즘이 무너지고, 이는 결국 악성 종양으로 발전한다.

현대 암 연구는 차세대 시퀀싱 기술의 발전과 함께 급속도로 발전하고 있다. 종양 이질성, 클론 진화, 변이 시그니처 등의

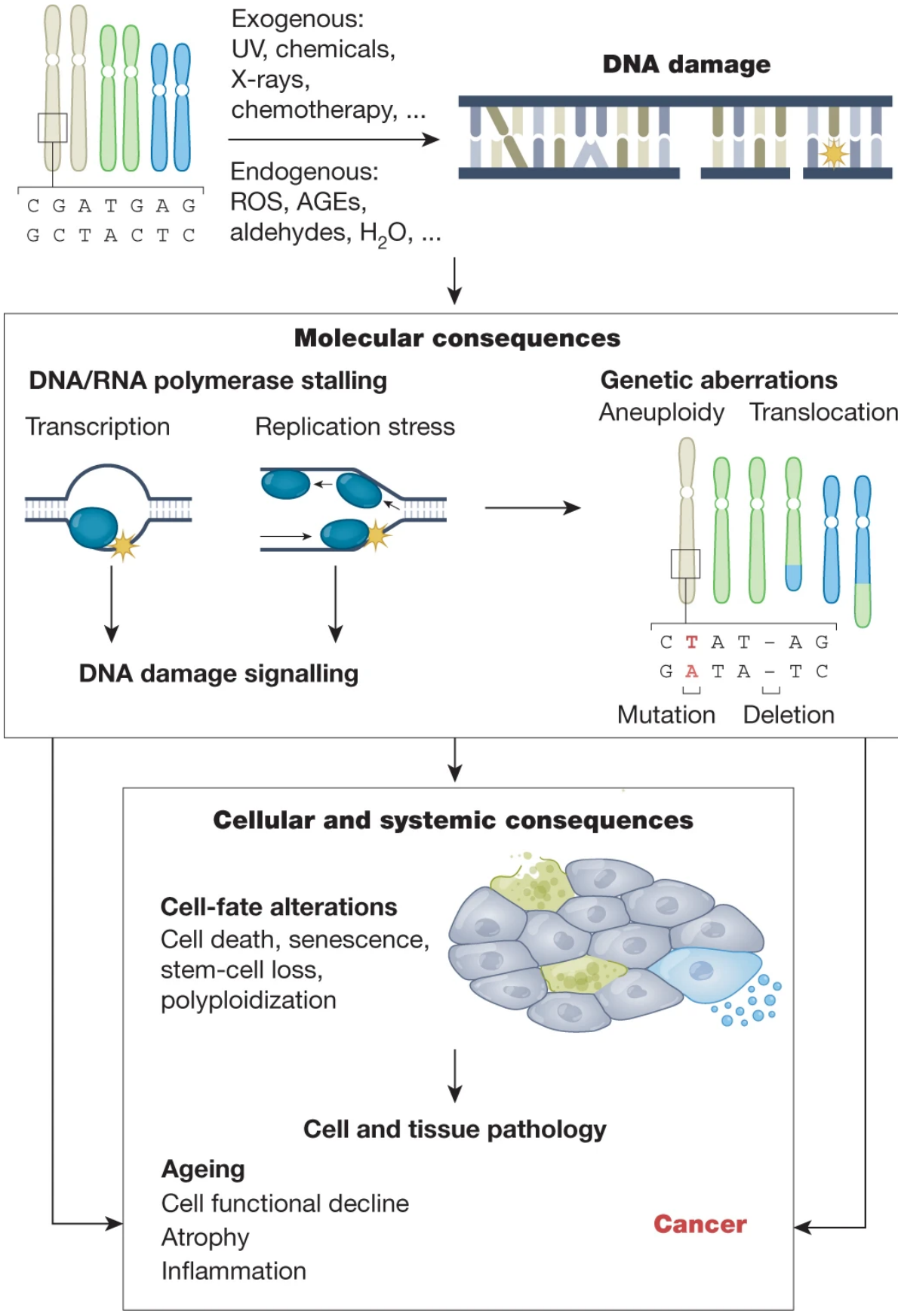


Figure 151: 노화와 암

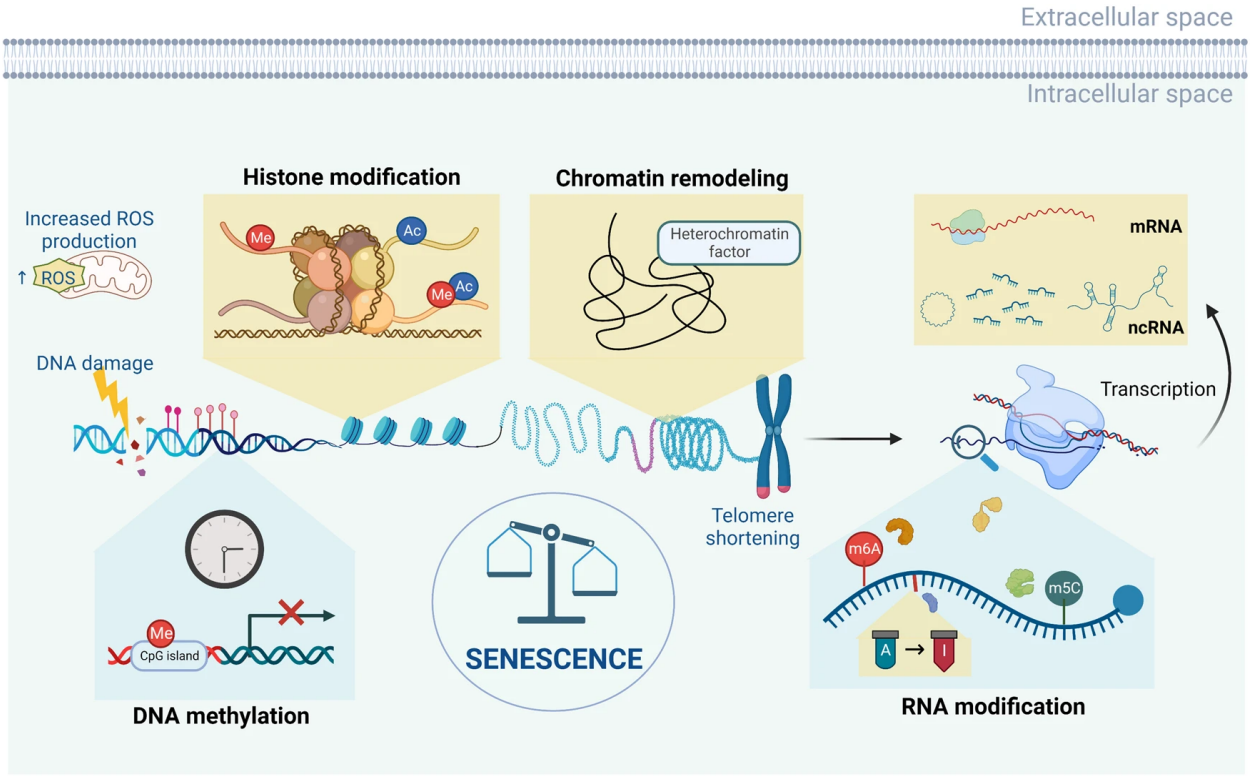


Figure 152: 세포 노화

개념을 통해 암의 발생과 진행 과정을 보다 정확하게 이해할 수 있게 되었다. MSI와 TMB 같은 바이오마커는 면역치료의 효과를 예측하는 중요한 지표로 활용되고 있다.

치료 측면에서는 전통적인 세포독성 항암제에서 정밀의학 기반 치료로 패러다임이 변화하고 있다. 환자의 종양을 유전체 분석하여 특정 변이나 바이오마커를 확인하고, 이에 맞는 표적치료제나 면역치료제를 선택하는 것이 표준 진료가 되었다. 표적치료제, 면역치료제, ADC, PROTAC 등 다양한 치료법이 개발되어 암을 "불치병"에서 "관리 가능한 만성질환" 또는 "완치 가능한 질병"으로 인식을 전환시키고 있다. 또한, 액체생검과 ctDNA 분석 기술은 비침습적으로 암을 진단하고 모니터링할 수 있는 새로운 가능성을 제시하고 있다. 이는 특히 치료 효과 모니터링과 재발 감시에서 중요한 역할을 할 것으로 기대된다.

의생명정보학은 이러한 모든 과정에서 핵심적인 역할을 담당하고 있다. 대용량 유전체 데이터의 분석, 변이의 기능적 해석, 치료 표적의 발굴, 바이오마커의 개발 등 모든 영역에서 의생명정보학적 접근이 필수적이다. 앞으로도 의생명정보학의 발전은 암 정복을 위한 중요한 열쇠가 될 것이다.

8 비코딩 유전체와 발현 조절

8.1 인간 유전체의 구성

인간 유전체는 단백질을 코딩하는 영역이 전체의 극히 일부에 불과하다는 특징을 가진다. 단백질 코딩 유전자는 전체 유전체의 1.5%에 불과하며, 나머지는 인트론(26%), LINE(Long Interspersed Nuclear Elements, 20%), SINE(Short Interspersed Nuclear Elements, 13%), LTR 레트로트랜스포존(8%), 단순 반복 서열(3%), 분절 중복(5%) 등의 비코딩 영역으로 구성되어 있다.

Figure 7.1 인간 유전체 구성 – 단백질 코딩 영역과 비코딩 영역의 비율

Figure 7.2 트랜스포제이블 요소 – 인간 유전체에 존재하는 다양한 트랜스포제이블 요소들

Figure 7.3 LINE 반복 서열 – LINE 요소와 질병에 관한 연구 논문

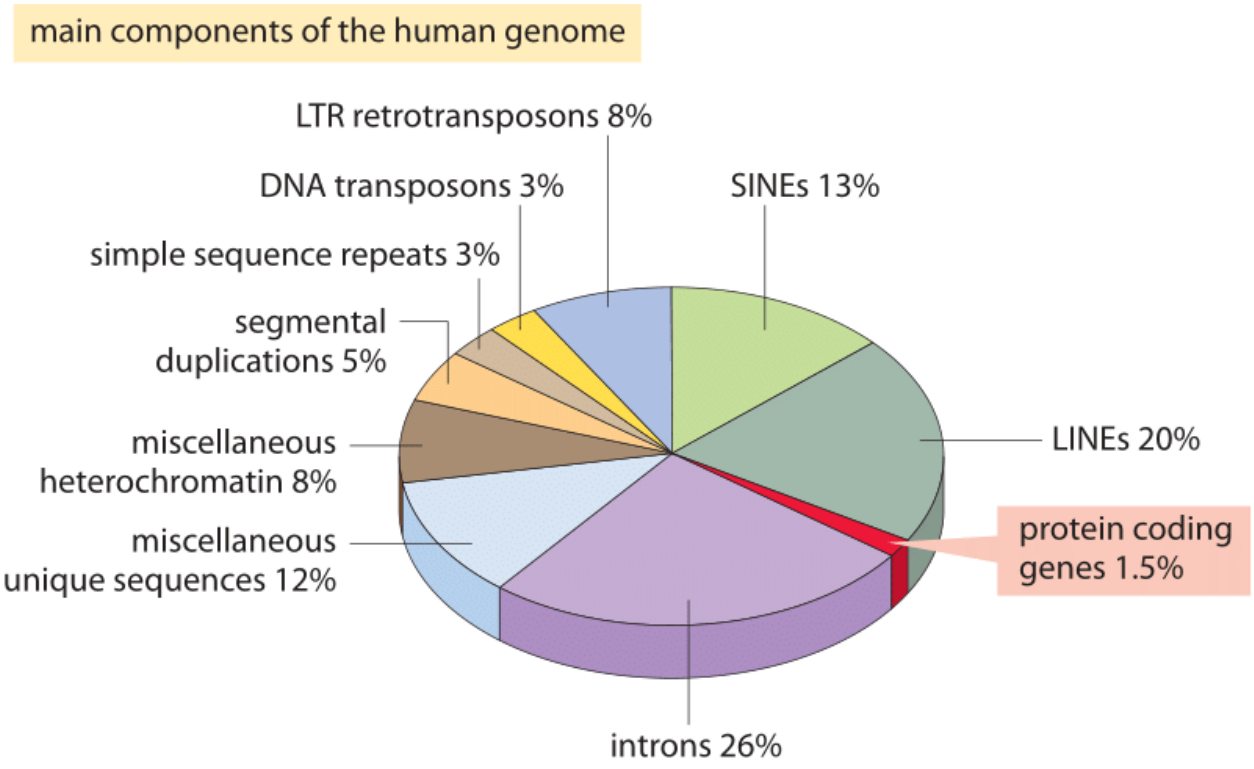


Figure 153: 인간 유전체 구성

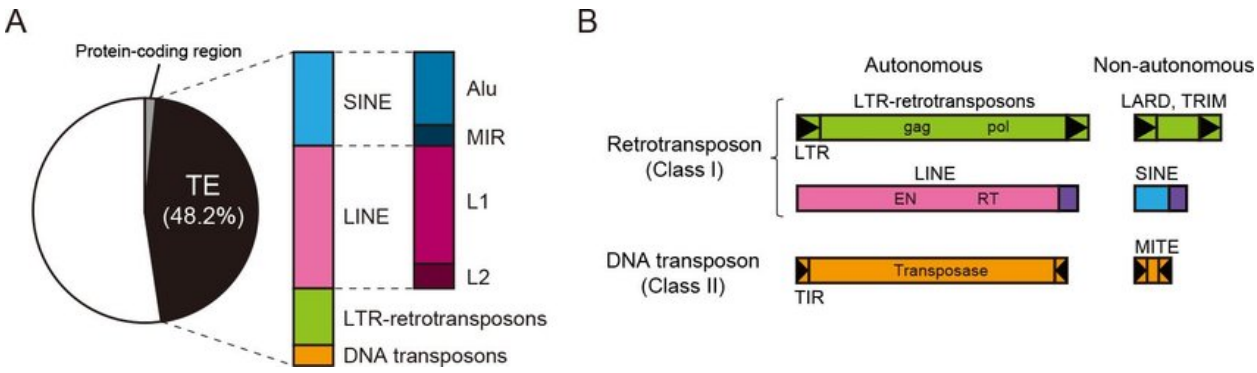
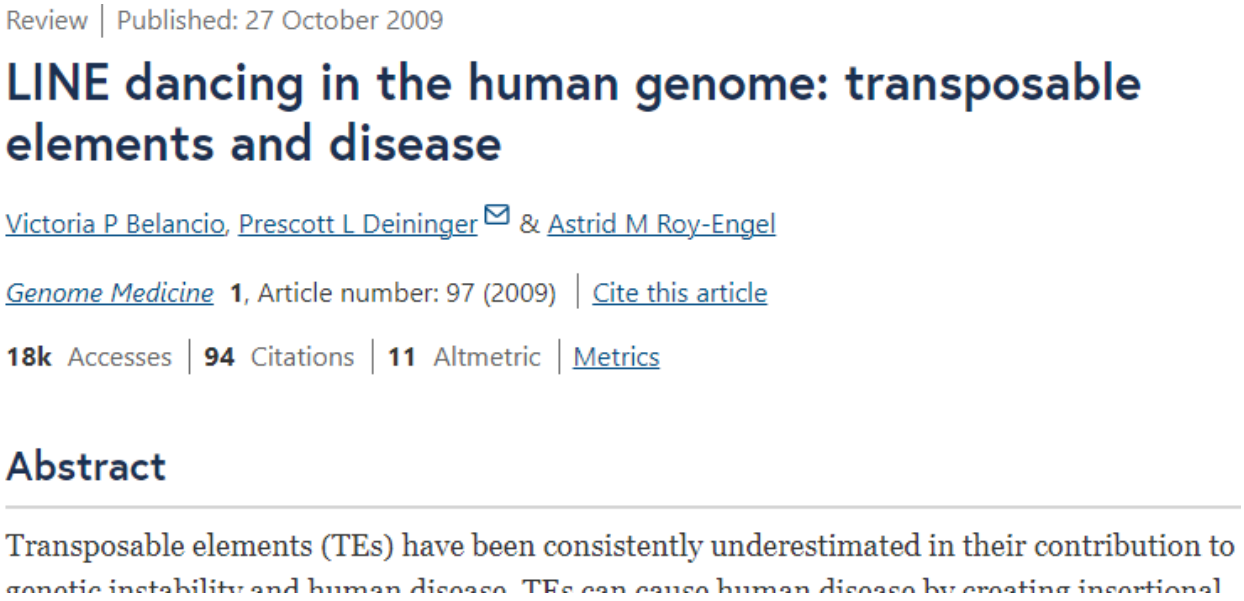


Figure 154: 트랜스포제이블 요소



Review | Published: 27 October 2009

LINE dancing in the human genome: transposable elements and disease

Victoria P Belancio, Prescott L Deininger & Astrid M Roy-Engel

Genome Medicine **1**, Article number: 97 (2009) | Cite this article

18k Accesses | **94** Citations | **11** Altmetric | Metrics

Abstract

Transposable elements (TEs) have been consistently underestimated in their contribution to genetic instability and human disease. TEs can cause human disease by creating insertional

Figure 155: LINE 반복 서열

이러한 구성을 통해 알 수 있듯이, 인간 유전체의 대부분은 비코딩 영역으로 이루어져 있다. 과거에는 이러한 비코딩 영역을 '정크 DNA(junk DNA)'라고 불렀지만, 현재는 이들이 유전자 발현 조절과 유전체 구조 유지에 중요한 역할을 한다는 것이 밝혀지고 있다.

8.2 GWAS 변이의 분포와 비코딩 영역의 중요성

전장유전체연관분석(Genome-Wide Association Studies, GWAS)을 통해 발견된 질병 관련 변이들의 분포를 분석해보면 흥미로운 패턴을 발견할 수 있다. GWAS 변이의 대부분은 비코딩 영역(인터제닉 및 인트라제닉)에 집중되어 있으며, 프로모터 영역에 일부가, 그리고 코딩 영역에는 극히 일부만이 분포하고 있다.

8.2.1 질병별 GWAS 변이 분포

질병별 분석 결과를 보면, 대부분의 질병에서 인핸서 영역에 많은 연관 SNP가 집중되어 있다. 일부 질병은 인핸서 및 프로모터 영역에 고르게 분포하는 패턴을 보이기도 한다.

Figure 7.4 GWAS 변이 분포 – 질병 관련 GWAS 변이들의 유전체 영역별 분포 패턴

Figure 7.5 전사인자와 인핸서 – 전사인자, TBP, RNA 중합효소의 인핸서-프로모터 상호작용 메커니즘

이러한 분포 패턴은 질병 감수성이 단백질 서열의 변화보다는 유전자 발현 조절의 변화와 더 밀접한 관련이 있음을 시사한다. 따라서 후성유전체학적 접근이 질병의 분자적 기전을 이해하는 데 필수적이다.

8.3 발현 수량적 형질 유전자좌(eQTL)

8.3.1 QTL과 eQTL의 개념

QTL(Quantitative Trait Loci)은 다유전자 형질에 영향을 미치는 유전변이들의 위치를 의미한다. 이 중에서 **eQTL(Expression Quantitative Trait Loci)**은 특히 유전자의 발현에 영향을 미치는 유전변이들의 위치를 가리킨다.

8.3.2 eQTL의 분류

eQTL은 작용 방식에 따라 크게 두 가지로 분류된다. Cis-eQTL은 동일한 염색체상에서 비교적 가까운 거리(일반적으로 표적 유전자로부터 1Mb 이내)에 있는 유전자의 발현에 직접적인 조절 작용을 통해 영향을 미치는 변이이다. 반면 Trans-

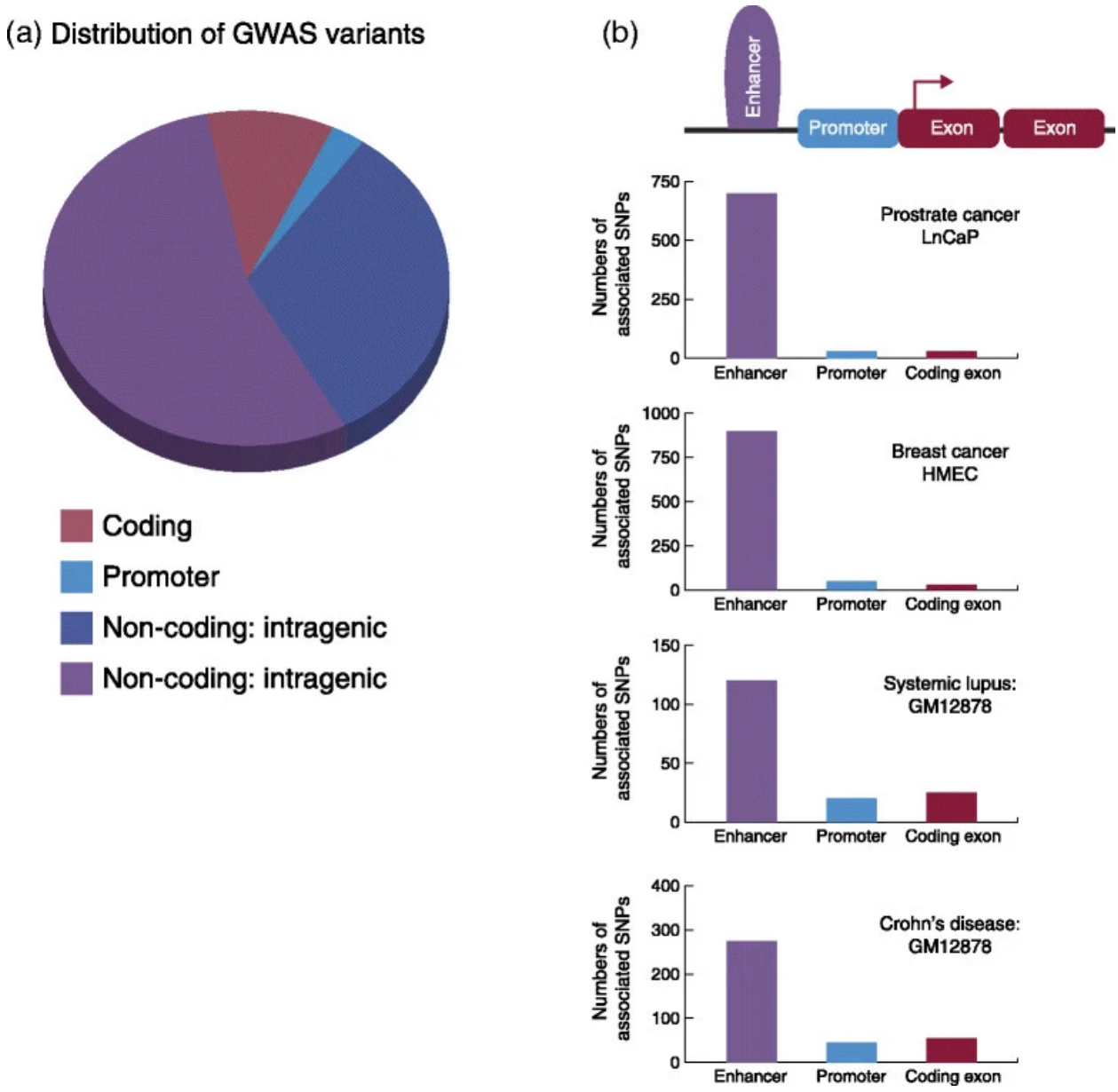


Figure 156: GWAS 변이 분포

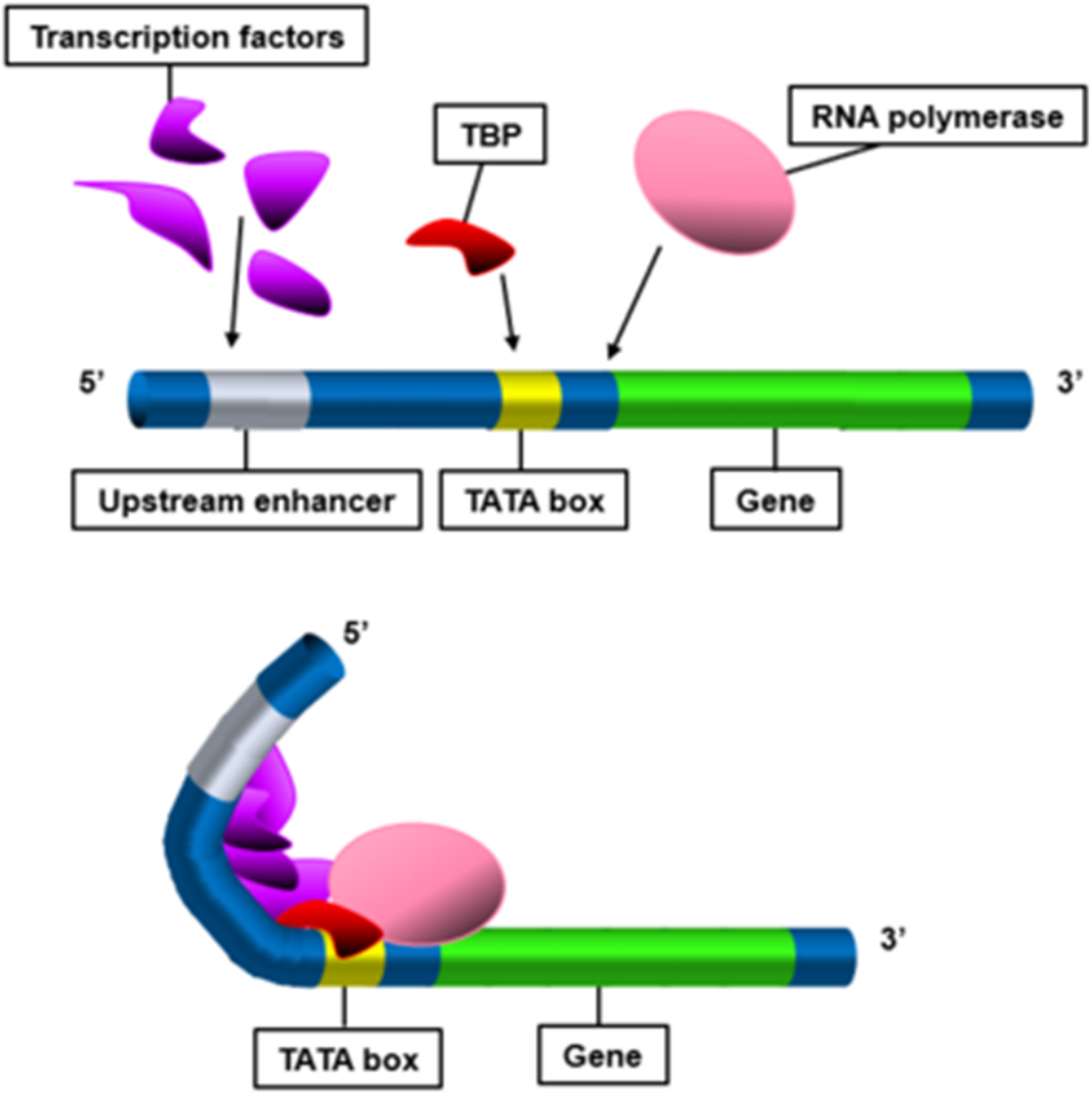


Figure 157: 전사인자와 인핸서

eQTL은 다른 염색체나 같은 염색체라도 멀리 떨어진 유전자의 발현에 영향을 미치는 변이로, 전사인자나 신호전달 경로를 통한 간접적인 조절 작용을 하며 네트워크 효과를 통해 다수 유전자의 발현에 동시에 영향을 미칠 수 있다.

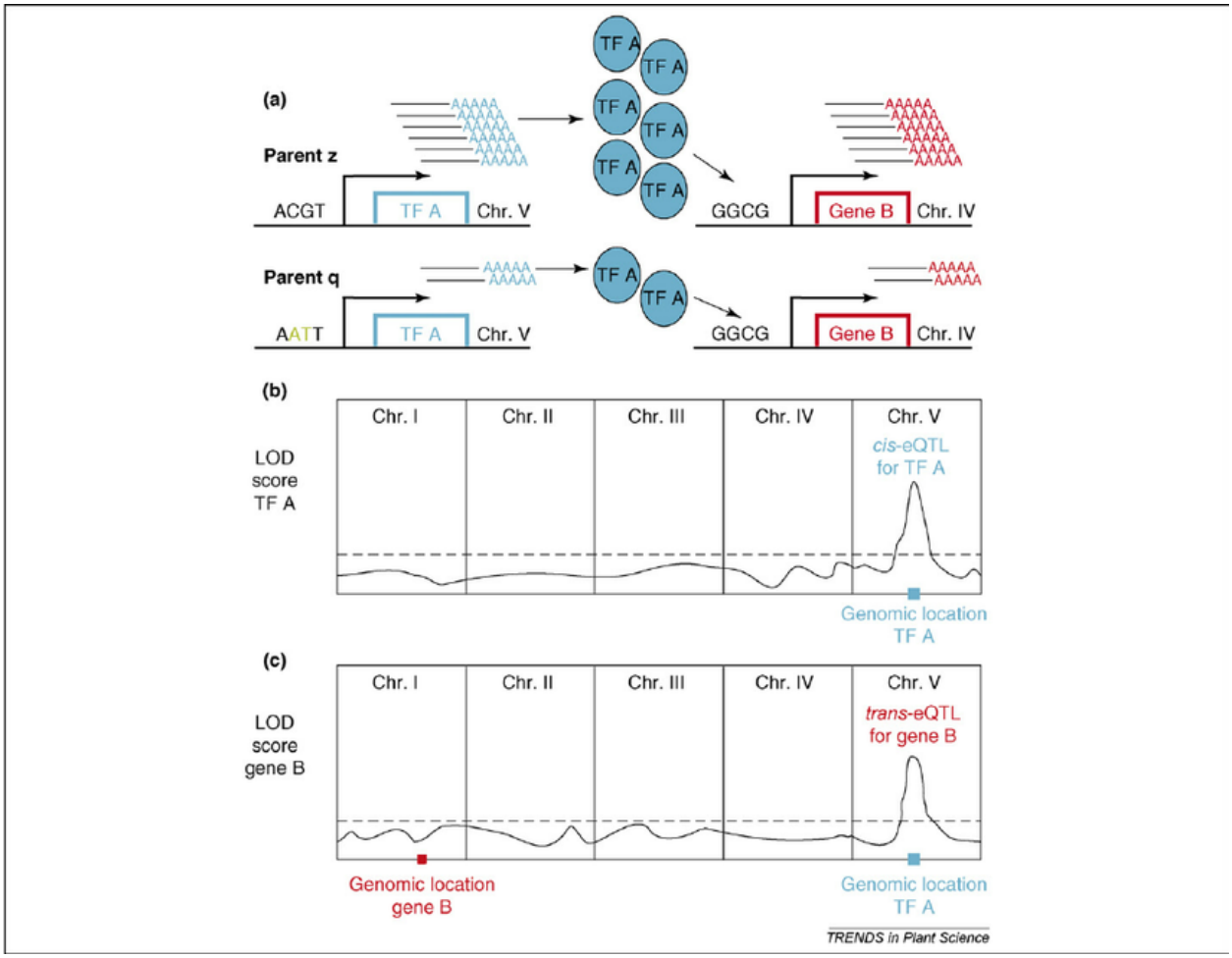


Figure 158: eQTL 개념

Figure 7.6 eQTL 개념 – Cis-eQTL과 Trans-eQTL의 차이점과 작용 메커니즘

8.3.3 eQTL 분석의 의의

GWAS 연구에서 질병과 연관된 변이를 발견하는 것은 첫 번째 단계일 뿐이다. 이러한 변이가 실제로 어떤 유전자의 발현에 영향을 미치는지 파악해야 질병의 분자적 메커니즘을 이해할 수 있는데, 바로 여기서 eQTL 분석이 중요한 역할을 한다.

eQTL 분석을 통해 우리는 여러 중요한 정보를 얻을 수 있다. 먼저 특정 유전적 변이가 실제로 어떤 유전자의 발현을 조절하는지 명확히 할 수 있으며, 그 조절 효과가 얼마나 강한지, 그리고 발현을 증가시키는지 감소시키는지도 정량적으로 측정할 수 있다. 특히 같은 변이라도 조직에 따라 다른 영향을 미칠 수 있기 때문에, 조직별 발현 조절 패턴을 파악하는 것은 질병의 조직 특이성을 이해하는 데 필수적이다. 나아가 여러 유전자들이 어떻게 네트워크를 이루어 질병 발생에 기여하는지도 eQTL 분석을 통해 규명할 수 있다.

8.3.4 GTEx 프로젝트

GTEx(Genotype-Tissue Expression) 프로젝트는 인간의 다양한 조직에서 유전자 발현 패턴과 유전적 변이의 연관성을 체계적으로 분석하는 대규모 연구 프로젝트다. 이 프로젝트는 49개의 서로 다른 조직에서 약 900명의 기증자로부터 수집된 데이터를 바탕으로 조직별 eQTL 지도를 구축했다.

GTEx 데이터베이스는 eQTL 연구의 표준 참조자료로 활용되고 있다. 이 데이터베이스를 통해 연구자들은 특정 유전적 변이가 어떤 조직에서 어떤 유전자의 발현에 영향을 미치는지 확인할 수 있으며, 조직별 발현 조절의 차이를 비교분석할 수 있다. GTEx는 또한 질병 관련 GWAS 변이들이 실제로 어떤 조직에서 어떤 유전자의 발현을 조절하는지 해석하는 데 핵심적인 역할을 한다.

8.4 트랜스포존과 유전체 진화

레트로트랜스포존(Retrotransposon)은 RNA를 중간체로 하여 역전사 과정을 통해 유전체 내에서 자신을 복사하고 이동하는 유전 요소다. 이들은 유전체 내에 존재하는 레트로바이러스의 흔적으로 여겨진다.

내인성 레트로바이러스(Endogenous Retroviruses, ERVs)는 과거에 생식세포계열에 감염되어 유전체에 통합된 레트로바이러스의 남은 서열이다. 바이러스 RNA와 단백질이 리보핵단백질 복합체를 형성하고, 역전사 과정을 통해 RNA가 DNA로 변환된다. 형성된 DNA는 숙주 세포의 유전체에 통합되어 새로운 바이러스 입자를 생산한다.

트랜스포존은 유전체에 다양한 영향을 미친다. 이들은 프로모터나 인핸서 역할을 통해 유전자 발현을 조절하고, 상동 재조합을 통해 큰 규모의 구조적 변이를 일으켜 유전체를 재배열한다. 또한 다른 방식의 스플라이싱을 통해 새로운 엑손을 제공하여 유전자 다양성을 증가시키고, 메틸화 패턴의 변화를 통한 후성유전학적 조절 기능을 수행한다.

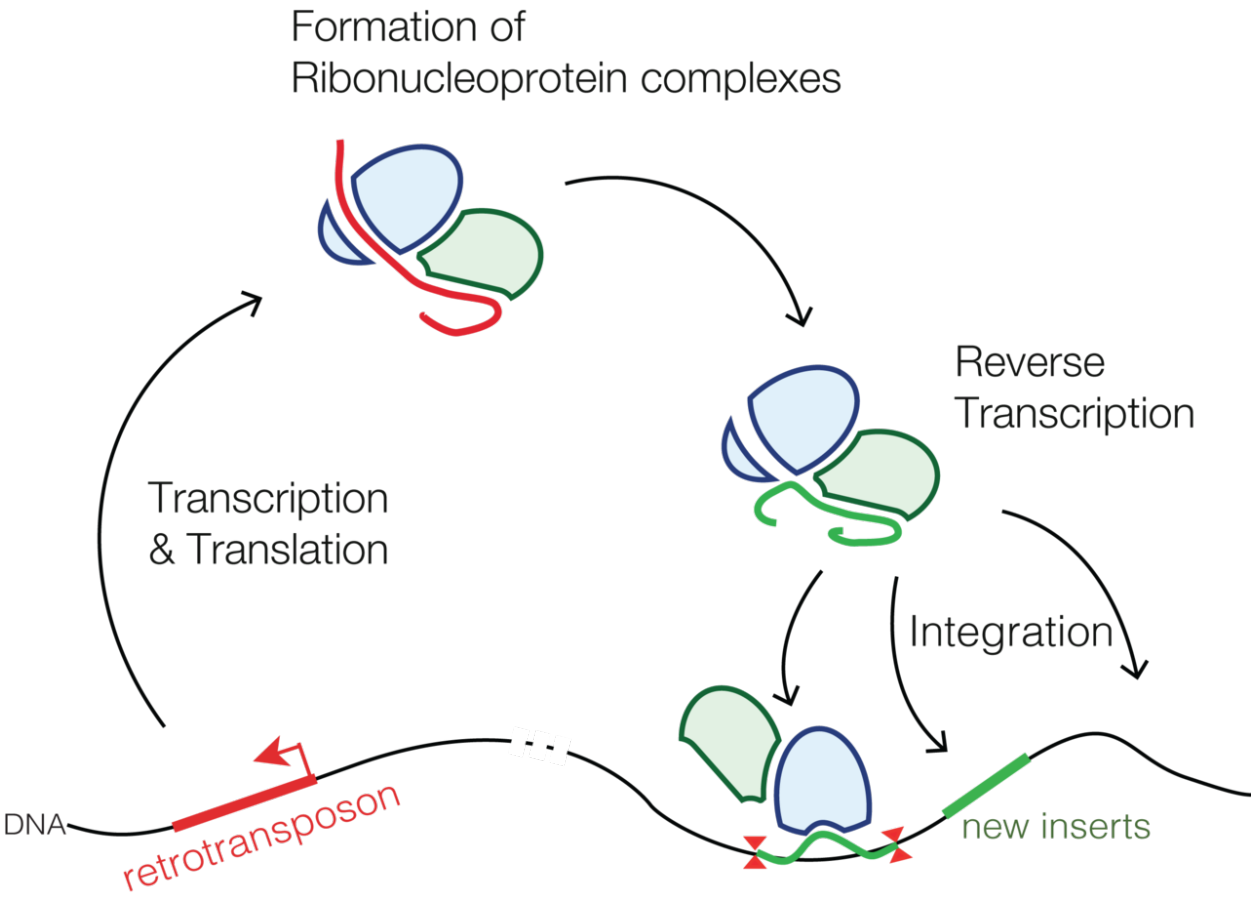


Figure 159: 레트로트랜스포존

Figure 7.7 레트로트랜스포존 – 레트로트랜스포존의 구조와 역전사 과정을 통한 복제 메커니즘

8.5 비코딩 RNA의 종류와 기능

비코딩 RNA(non-coding RNA, ncRNA)는 단백질로 번역되지 않지만 다양한 생물학적 기능을 수행하는 RNA 분자들이다. 인간 유전체에서 전사되는 RNA의 대부분이 비코딩 RNA이며, 유전자 발현 조절에 중요한 역할을 한다.

주요 비코딩 RNA로는 마이크로RNA (miRNA), 긴 비코딩 RNA (lncRNA), 작은 간섭 RNA (siRNA), 환형 RNA (circRNA) 등이 있다. 마이크로RNA는 약 20개 뉴클레오타이드 길이의 짧은 RNA로, 표적 mRNA에 결합하여 유전자 발현을 억제한다. 암에서는 종양억제 miRNA의 감소와 종양촉진 miRNA의 증가가 관찰되며, 이를 바탕으로 miRNA 기반 치료제가 개발되고 있다.

긴 비코딩 RNA는 200개 뉴클레오타이드 이상의 긴 RNA로, 크로마틴 구조 변화, 전사 조절, miRNA 스펀지 역할을 한다. 일부 긴 비코딩 RNA는 암 발생과 진행에 관여하는 것으로 알려져 있다. 작은 간섭 RNA는 약 21–23개 뉴클레오타이드 길이의 이중가닥 RNA로, 표적 mRNA의 특이적 분해를 통해 유전자 발현을 억제한다. 현재 여러 siRNA 치료제가 임상에서 승인되어 사용되고 있다.

8.6 비코딩 RNA의 의생명정보학적 분석

비코딩 RNA 연구에는 다양한 의생명정보학적 방법과 도구가 활용된다. RNA 시퀀싱을 통해 비코딩 RNA의 발현을 분석하고, 다양한 예측 도구를 사용하여 miRNA의 표적 유전자를 예측한다. 또한 유전자 온톨로지와 경로 분석을 통해 비코딩 RNA의 기능을 분석한다.

주요 데이터베이스로는 miRNA, 긴 비코딩 RNA, 암 관련 변이, 조직별 발현 패턴을 다루는 다양한 전문 데이터베이스들이 있다. 이들은 각각 서열과 주석 정보, 질병 관련 변이, 조직별 발현 패턴 등을 제공한다.

생물정보학적 도구로는 miRNA 표적 예측과 RNA 구조 예측을 위한 예측 소프트웨어, 비코딩 RNA 발현 분석을 위한 작업 과정, 그리고 네트워크 분석과 발현 패턴 시각화 도구 등이 있다. 이러한 도구들은 비코딩 RNA의 기능과 질병과의 연관성을 이해하는 데 필수적이다.

8.7 치료적 응용과 전망

비코딩 RNA는 현재 다양한 치료법의 표적이 되고 있다. siRNA 치료제는 유전성 질환 치료에서 승인을 받아 임상에서 사용되고 있으며, 안티센스 올리고뉴클레오타이드는 여러 질환 치료에 활용되고 있다. miRNA 기반 치료법도 임상시험이 진행되고 있어 향후 새로운 치료 옵션이 될 것으로 기대된다.

미래에는 환자별 비코딩 RNA 발현 양상을 바탕으로 한 개인 맞춤형 치료가 가능할 것으로 전망된다. 비코딩 RNA는 질병의 진단, 예후 예측, 치료 반응 모니터링을 위한 바이오마커로도 활용될 것이다. 또한 나노기술을 이용한 표적 전달 시스템의 발전으로 비코딩 RNA 기반 치료제의 효율성과 안전성이 더욱 향상될 것으로 예상된다.

8.8 결론

비코딩 유전체는 과거 '정크 DNA'로 여겨졌으나, 현재는 유전자 발현 조절에 핵심적인 역할을 하는 것으로 밝혀졌다. GWAS 변이의 대부분이 비코딩 영역에 분포한다는 사실은 질병이 단백질 서열 변화보다는 발현 조절 변화와 더 밀접한 관련이 있음을 보여준다.

비코딩 RNA는 miRNA, lncRNA, siRNA 등 다양한 종류가 있으며, 각각 고유한 방식으로 유전자 발현을 조절한다. 이들의 이상은 암, 신경질환, 심혈관질환 등과 연관되어 있어 새로운 진단 바이오마커와 치료 표적으로 주목받고 있다.

의생명정보학적 측면에서는 RNA-seq 기술과 다양한 분석 도구들을 통해 비코딩 RNA의 기능을 예측하고 분석할 수 있다. 특히 miRNA 표적 예측, 발현 패턴 분석, 네트워크 분석 등이 중요한 연구 방법론이다.

현재 siRNA 치료제가 임상에서 성공을 거두고 있으며, 향후 비코딩 RNA를 활용한 개인 맞춤형 치료법 개발이 기대된다. 이러한 비코딩 영역의 이해는 정밀의학 시대의 핵심 요소가 될 것이다.

9 후성유전체학

9.1 크로마틴 구조와 후성유전학적 조절

9.1.1 크로마틴의 기본 구조

DNA는 히스톤 단백질을 감싸 **뉴클레오솜(nucleosome)**을 형성하고, 이들이 연결되어 크로마틴 구조를 이룬다. 크로마틴의 응축 정도에 따라 유전자 접근성이 달라지며, 이는 유전자 발현 조절의 핵심 메커니즘이다.

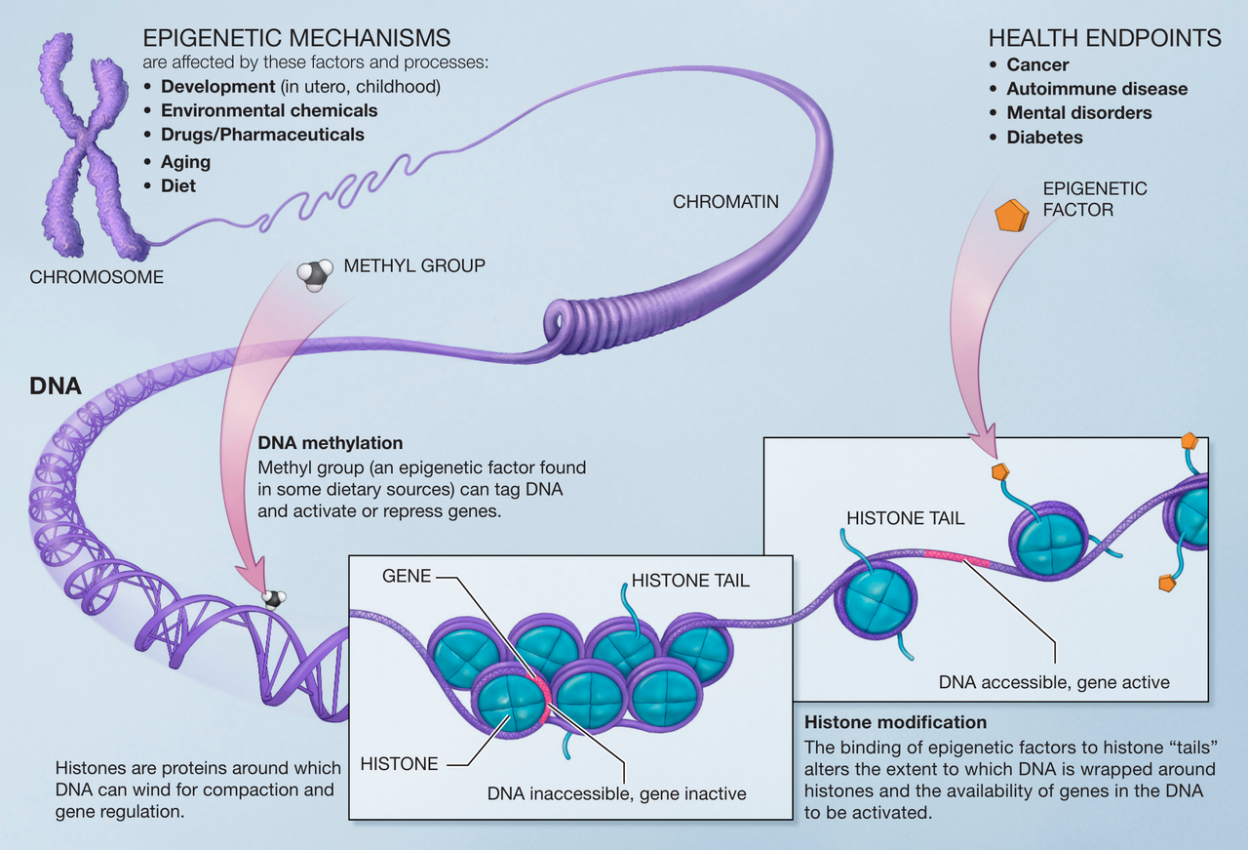


Figure 160: 후성유전체

Figure 8.1 후성유전체 – DNA 메틸화, 히스톤 수정, 크로마틴 구조와 유전자 발현 조절

9.1.2 후성유전학적 조절 요인

후성유전학적 조절은 다양한 내외부적 요인의 영향을 받는다. 내부적 요인으로는 세포 분화와 발달 과정에서의 프로그래밍과 관련된 발달 단계와 세포 분열과 관련된 크로마틴 변화를 수반하는 세포 주기가 있다. 외부적 요인으로는 내분비 교란물질이나 중금속 같은 환경 화학물질, 치료제나 독성물질 노출과 관련된 약물, 시간 경과에 따른 점진적 변화로 나타나는 노화, 그리고 영양소와 대사물질의 영향을 포함하는 식이 요인 등이 있다.

9.1.3 질병과의 연관성

후성유전학적 변화는 다양한 질병과 밀접한 관련이 있다. 암에서는 종양 억제 유전자의 과메틸화와 암유전자의 저메틸화가 관찰되며, 이러한 변화가 암 발생과 진행에 중요한 역할을 한다. 자가면역질환에서는 면역 관련 유전자의 비정상적 발현 조절이 나타나 면역 반응의 균형이 깨지게 된다. 정신질환의 경우 신경발달과 신경가소성 관련 유전자의 조절 이상이 주요 특징이며, 당뇨병에서는 대사 관련 유전자의 발현 조절 장애가 질병 발생의 핵심 메커니즘으로 작용한다.

9.2 DNA 메틸화의 분자적 기전

9.2.1 DNA 메틸화의 기본 개념

DNA 메틸화는 사이토신 염기에 메틸기(–CH_3)를 부착하는 후성유전학적 변형으로, 주로 CpG 디뉴클레오타이드에서 발생한다. 이러한 메틸화 과정은 유전자 발현을 조절하는 중요한 메커니즘이다.

9.2.2 메틸화와 유전자 발현의 관계

메틸화와 유전자 발현 간의 관계는 두 가지 주요 상태로 나타난다. 정상적인 저메틸화 상태에서는 크로마틴이 개방되어 전사인자의 접근이 용이하며, RNA 중합효소 복합체가 효율적으로 전사를 수행한다. 반면 과메틸화 상태에서는 메틸화된 CpG가 메틸 CpG 결합 단백질을 유인하여 크로마틴이 응축되고, 히스톤 탈아세틸화효소 복합체의 동원으로 인해 전사가 억제된다.

Figure 8.2 DNA 메틸화 – 메틸화와 유전자 발현 억제 메커니즘

9.2.3 메틸화의 생화학적 특성

DNA에서는 여러 형태의 메틸화가 존재할 수 있다. 5-메틸사이토신(5mC)은 진핵생물과 원핵생물 모두에서 가장 잘 알려진 형태이다. 6-메틸아데닌(6mA) 및 4-메틸사이토신(4mC)은 주로 원핵생물에서 발견된다.

Figure 8.3 다양한 메틸화 형태 – 5mC, 6mA, 4mC 등 다양한 DNA 메틸화 형태들의 화학 구조

9.3 전장 바이설파이트 시퀀싱(WGBS)

9.3.1 WGBS의 원리

전장 바이설파이트 시퀀싱(Whole Genome Bisulfite Sequencing, WGBS)은 유전체 전체에서 5-메틸사이토신(5mC)의 위치를 단일 염기 해상도로 검출하는 기술이다. 이 기술의 핵심은 바이설파이트 화학적 처리를 통해 비메틸화된 사이토신을 우라실로 변환시키는 것인데, 이때 우라실은 DNA에서 일반적으로 사용되지 않는 염기이다. 일반적으로 DNA에서는 티민(T)이 사용되고 RNA에서만 우라실(U)이 사용되지만, 바이설파이트 처리 과정에서는 이러한 변환을 이용하여 메틸화 상태를 구별할 수 있게 된다.

9.3.2 바이설파이트 변환의 화학적 원리

바이설파이트 처리의 핵심 원리는 다음과 같다. 먼저 비메틸화된 사이토신은 바이설파이트 처리 시 우라실로 변환되는 반면, 메틸화된 사이토신은 이 처리에 저항하여 사이토신으로 유지된다. 이후 PCR 증폭 과정에서 우라실은 티민으로, 사이토신은 사이토신으로 유지되며, 최종적으로 시퀀싱을 통해 이러한 변환 패턴을 분석하여 원래의 메틸화 상태를 추론할 수 있다.

Figure 8.4 WGBS 바이설파이트 처리 – 바이설파이트 화학적 처리를 통한 메틸화 검출 원리

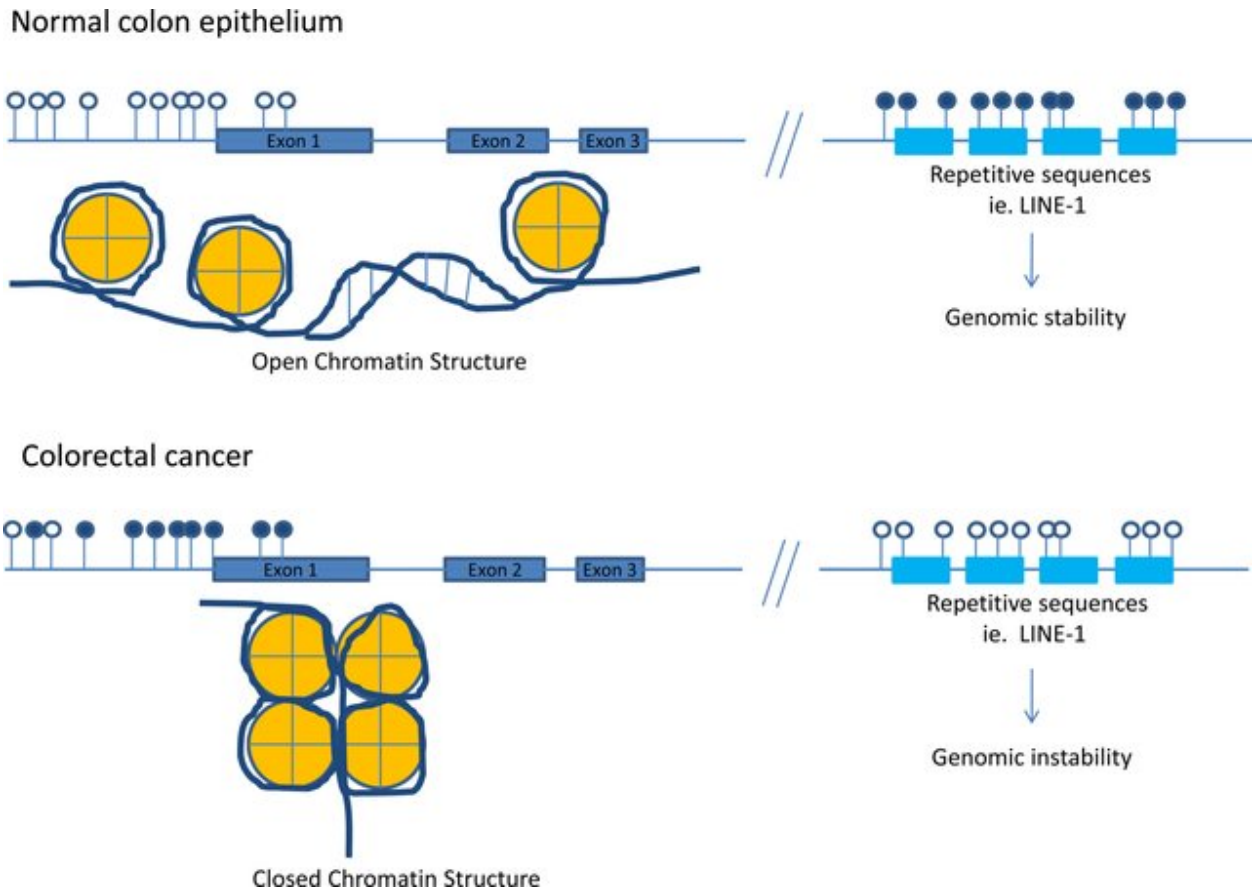


Figure 161: DNA 메틸화

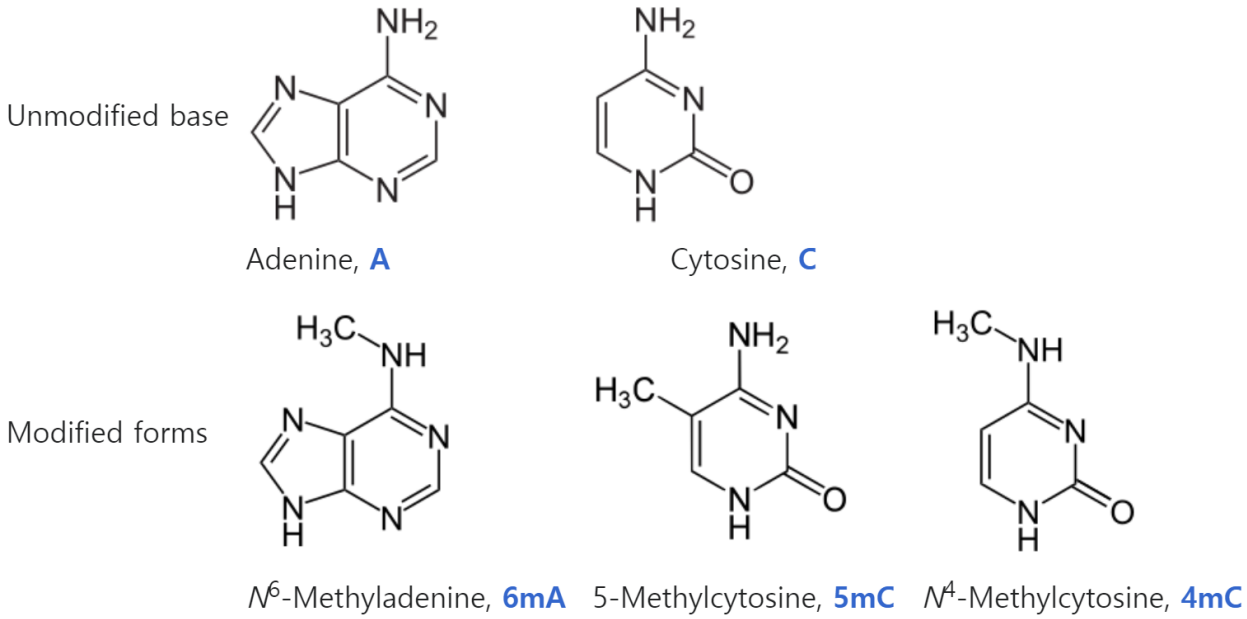


Figure 162: 다양한 메틸화 형태

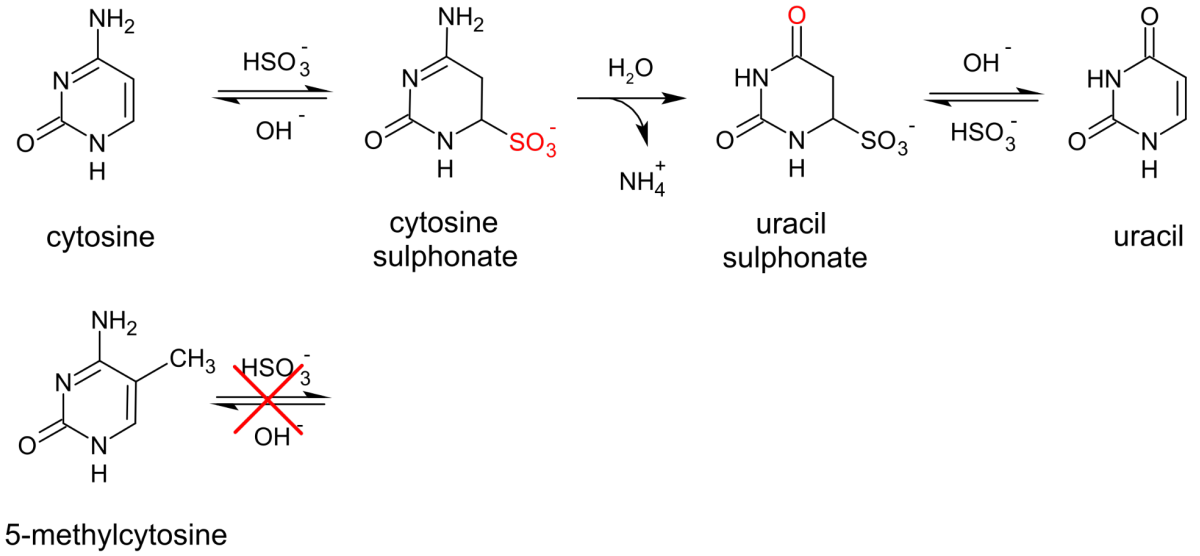


Figure 163: WGBS 바이설파이트 처리

9.3.3 WGBS 실험 과정

WGBS 파이프라인은 여러 단계로 구성된다. 먼저 고품질 유전체 DNA를 추출하여 샘플을 준비하고, 화학적 처리를 통해 사이토신을 바이설파이트로 변환한다. 이후 시퀀싱용 DNA 라이브러리를 구축하고 대용량 병렬 시퀀싱을 수행한다. 마지막으로 생물정보학 분석 단계에서는 BSMAP, Bismark 등의 전용 도구를 사용하여 정렬을 수행하고, 각 CpG 위치별 메틸화 수준을 계산하며, 변환 효율성 및 커버리지에 대한 품질 관리를 수행한다.

Figure 8.5 WGBS 개요 – 전장 바이설파이트 시퀀싱의 전체적인 실험 과정

9.3.4 WGBS 데이터의 해석

WGBS 데이터 분석에서는 여러 주요 요소들을 고려해야 한다. 변환 효율성은 비메틸화 사이토신의 변환율로 95% 이상을 권장하며, 정확한 메틸화 수준 측정을 위해서는 충분한 시퀀싱 깊이와 커버리지가 필요하다. 또한 유전체 영역별 CpG 밀도의 차이와 세포나 조직 간 메틸화 패턴의 차이인 조직 특이성도 중요한 고려사항이다.

9.4 후성유전체 연관 분석(EWAS)

9.4.1 EWAS의 정의와 목적

후성유전체 연관 분석(Epigenome-wide Association Study, EWAS)은 유전체 전반의 메틸화 패턴과 특정 표현형(질병, 형질) 간의 연관성을 통계적으로 검정하는 연구 방법이다. EWAS는 6장에서 다룬 유전체 전장 연관 분석(GWAS)과 본질적으로 유사한 접근법을 사용하지만, DNA 서열 변이 대신 후성유전학적 변화(주로 DNA 메틸화)를 분석 대상으로 한다는 점에서 차이가 있다.

9.4.2 EWAS와 GWAS의 차이점

EWAS의 연구 설계는 GWAS와 기본적으로 동일하지만, 별다른 측정 대상과 고려사항들로 인해 여러 중요한 차이점들이 있다. 가장 기본적인 차이는 측정 대상에서 나타난다. GWAS가 DNA 서열 변이(예: SNP, CNV)를 대상으로 하는 반면, EWAS는 CpG 지점별 DNA 메틸화 수준을 분석한다.

기술적 고려사항에서도 중요한 차이가 있다. 현실에서 블액 내 세포 유형별 메틸화 패턴의 차이로 인해 세포 구성 효과를 반드시 보정해야 하며, 메틸화는 환경 요인에 의해 더 큰 영향을 받는 특성이 있다. 또한 메틸화 패턴이 시료 보관 조건에

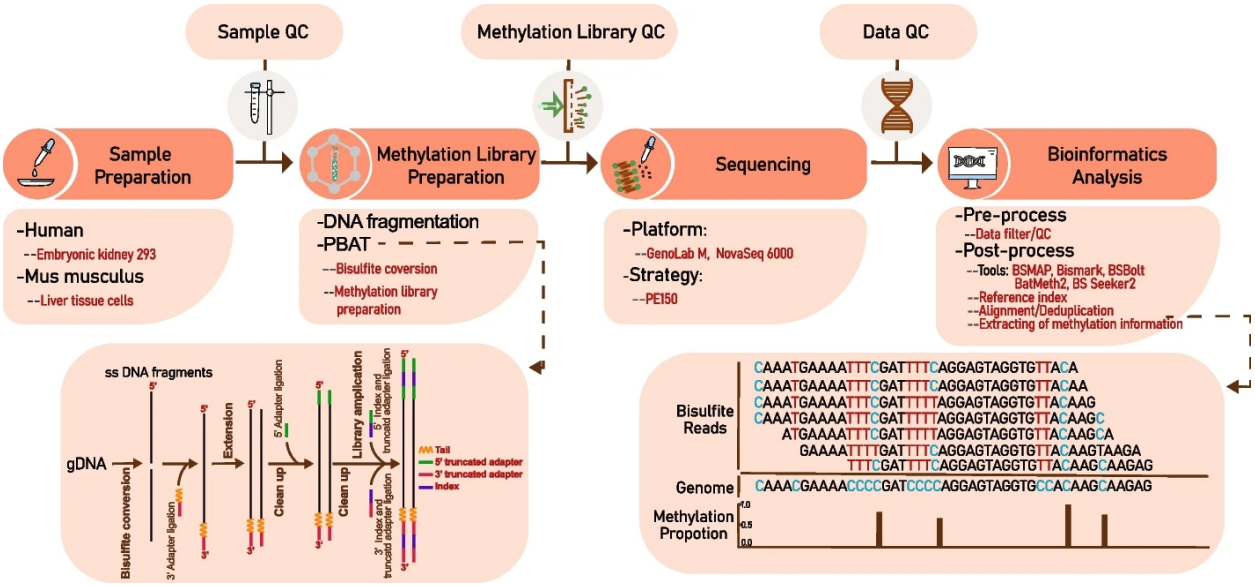


Figure 164: WGBS 개요

더 민감하기 때문에 시료 안정성에도 주의를 기울여야 한다. 통계적 측면에서는 EWAS에서 CpG 지점 수가 약 2700만 개에 달하므로 GWAS보다 더 엄격한 유의수준을 적용해야 한다.

9.4.3 EWAS의 주요 발견들

EWAS 연구를 통해 다양한 환경 요인과 질병에서의 메틸화 변화가 밝혀졌다. 환경 요인 측면에서는 흡연이 특정 CpG 지점에서 일관된 메틸화 변화를 일으키고, 대기오염은 염증 관련 유전자의 메틸화 변화와 연관되어 있다. 또한 스트레스는 시상하부-뇌하수체-부신 축 관련 유전자의 조절 변화를 유발한다.

질병 관련 연구에서는 여러 중요한 발견들이 있었다. 당뇨병에서는 포도당 대사 관련 유전자의 메틸화 이상이 발견되었고, 심혈관질환에서는 지질 대사와 염증 관련 유전자에서 의미 있는 변화가 관찰되었다. 정신질환의 경우에도 신경발달 및 신경전달물질 관련 유전자의 조절 이상이 확인되었다.

9.5 크로마틴 면역침강 시퀀싱(ChIP-seq)

9.5.1 ChIP-seq의 기본 원리

크로마틴 면역침강 시퀀싱(Chromatin Immunoprecipitation sequencing, ChIP-seq)은 특정 단백질(전사인자, 히스톤 변형 등)이 결합한 DNA 서열을 유전체 전체에서 식별하는 기술이다.

9.5.2 ChIP-seq 실험 과정

9.5.2.1 주요 실험 단계 ChIP-seq 실험 과정은 5개의 주요 단계로 구성된다. 첫 번째 단계인 교차결합(Crosslinking)에서는 포름알데하이드를 사용하여 살아있는 세포 내에서 일어나는 단백질-DNA 상호작용을 고정하여 in vivo 상태를 보존한다. 두 번째 단계는 크로마틴 전단(Shearing) 과정으로, 초음파 처리를 통한 물리적 방법 또는 MNase 등을 이용한 효소적 방법으로 크로마틴을 200-500bp 크기로 절단한다.

Figure 8.6 ChIP-seq 과정 - 크로마틴 면역침강 시퀀싱의 실험 과정

세 번째 단계인 면역침강(Immunoprecipitation)에서는 표적 단백질에 특이적인 항체를 사용하여 Protein A/G 비드를 통해 항체-항원 복합체를 포획하고, 충분한 세척을 통해 비특이적 결합을 제거한다. 네 번째 단계에서는 역교차결합 및 DNA 정제 과정을 통해 고온 처리로 단백질-DNA 교차결합을 해제하고, 프로테아제 K 처리로 단백질을 제거하여 고순도 DNA를 회수한다. 마지막 단계에서는 라이브러리 제작 및 시퀀싱을 위해 시퀀싱 어댑터를 부착하고 PCR 증폭을 거쳐 고처리량 시퀀싱을 수행한다.

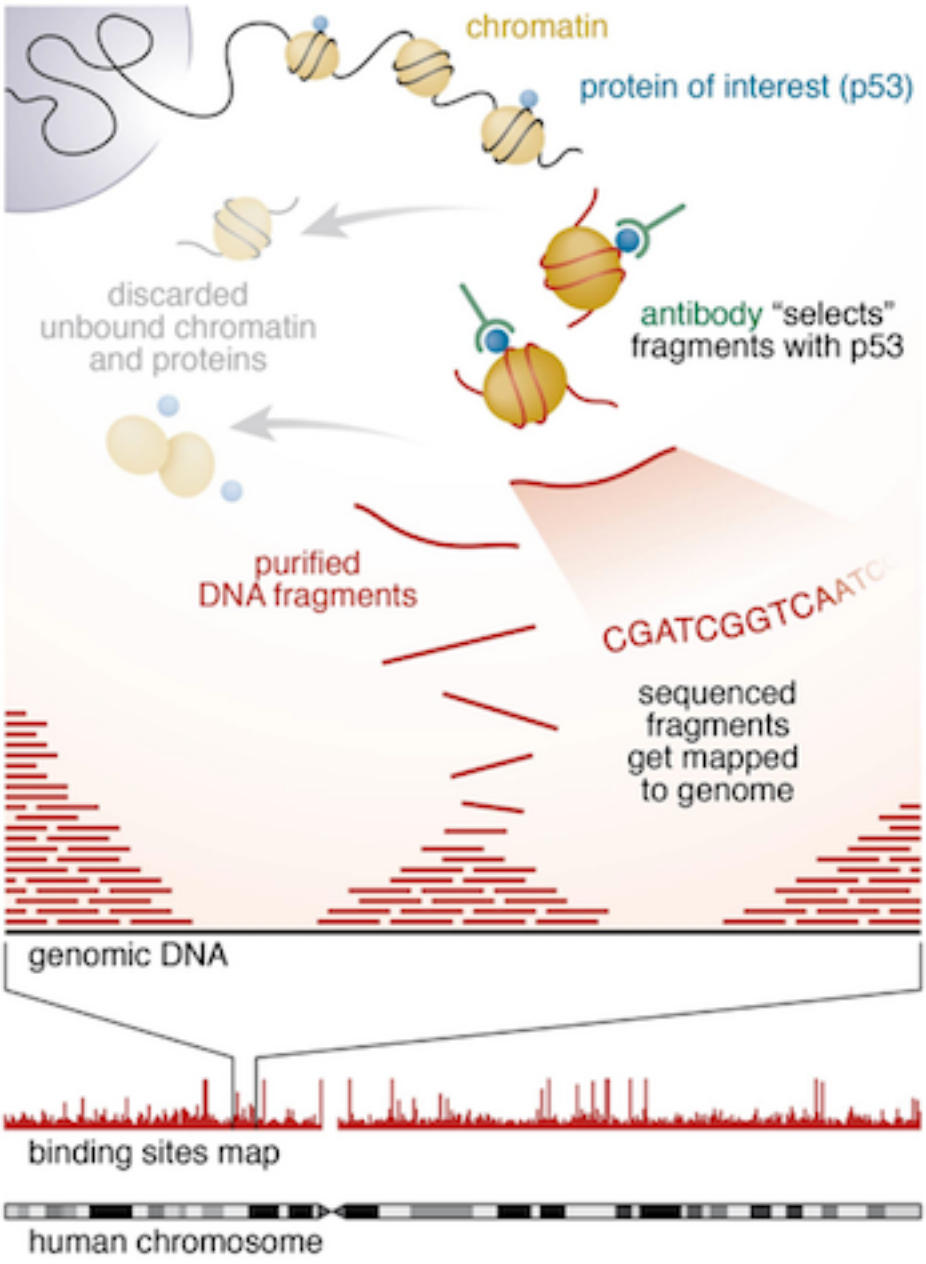


Figure 165: ChIP-seq 과정

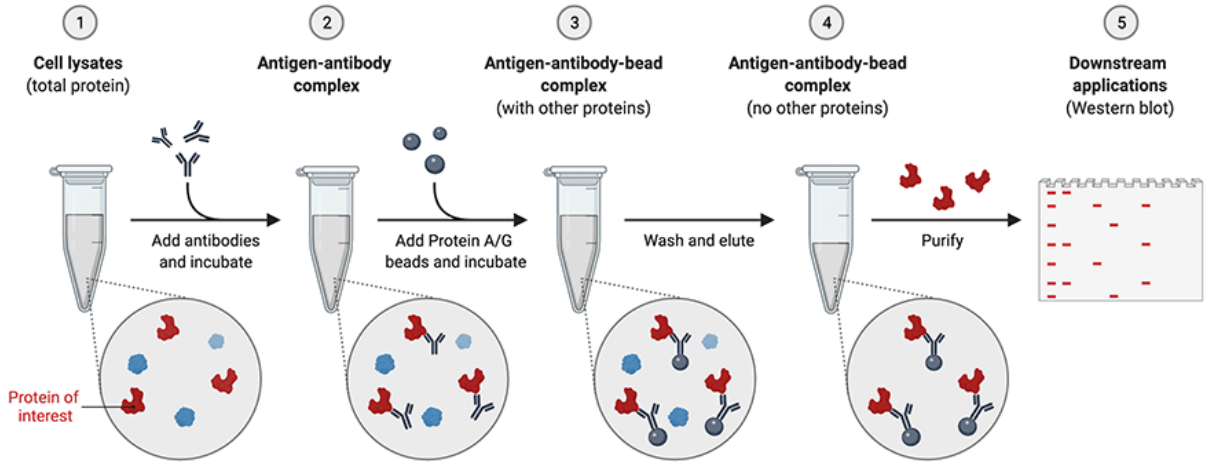


Figure 166: 면역침강 과정

Figure 8.7 면역침강 과정 – 항체–항원 결합과 자성 비드를 이용한 분리 과정

ChIP–seq 실험에서 대조 실험은 결과의 신뢰성을 확보하는 데 필수적이다. Input DNA는 면역침강 없이 동일하게 처리한 전체 크로마틴으로 배경 신호를 평가하는 데 사용되며, IgG 대조는 비특이적 항체를 사용한 음성 대조로 비특이적 결합을 평가한다. 또한 실험 재현성 확보를 위해 독립적인 생물학적 복제 실험을 수행해야 한다.

9.5.3 면역침강의 생화학적 원리

면역침강의 생화학적 원리는 5개의 순차적 단계로 이루어진다. 먼저 표적 단백질에 대한 특이적 항체–항원 결합이 일어난 후, Protein A/G가 항체의 Fc 부위에 결합하여 비드 복합체를 형성한다. 이후 자성 비드를 이용한 자기 분리를 통해 복합체를 회수하고, 체계적인 세척 과정을 통해 비특이적 결합물질을 제거한다. 마지막으로 용출 과정을 통해 표적 복합체만을 선택적으로 회수한다.

9.5.4 ChIP–seq 데이터 분석

9.5.4.1 전처리 및 품질 관리 ChIP–seq 데이터 분석은 전처리 및 품질 관리 단계에서 시작된다. 이 단계에서는 FastQC 등을 이용한 시퀀싱 품질 평가와 BWA, Bowtie2 등을 이용한 레퍼런스 유전체 정렬을 수행하며, PCR 중복으로 인한 편향을 제거하기 위해 중복 서열을 제거한다.

9.5.4.2 피크 호출(Peak Calling) 피크 호출은 ChIP–seq 데이터에서 단백질이 결합한 유전체 영역을 식별하는 핵심 단계이다. MACS2, HOMER 등의 알고리즘을 사용하여 대조군 대비 통계적으로 유의한 신호 강화 영역을 식별한다. 피크의 형태는 표적 단백질의 특성에 따라 구분되는데, 전사인자의 경우 특정 DNA 서열에 직접 결합하여 좁은 피크(Sharp peaks)를 형성하는 반면, 히스톤 변형의 경우 넓은 크로마틴 도메인에 걸쳐 분포하여 넓은 피크(Broad peaks)를 나타낸다.

9.5.4.3 기능적 주석화 피크 호출을 통해 식별된 결합 부위의 생물학적 의미를 파악하는 기능적 주석화 과정이 이어진다. 프로모터 분석에서는 전사개시점(TSS) 주변의 결합 패턴을 조사하여 유전자 전사에 미치는 직접적 영향을 평가하며, 인핸서 식별을 통해 원위 조절 요소를 발견하여 장거리 유전자 조절 메커니즘을 규명한다. 유전자 온톨로지 분석을 통해 결합 표적 유전자들의 기능적 특성을 분석하고, 모티프 분석을 통해 단백질이 인식하는 특정 DNA 서열 패턴의 특성을 규명한다.

9.5.5 ChIP-seq의 응용 분야

9.5.5.1 전사 조절 연구 ChIP-seq는 전사 조절 메커니즘을 규명하는 핵심 도구로 활용된다. 전사인자 결합 지도 작성을 통해 조직 및 세포 특이적 결합 패턴을 밝혀내고, 여러 전사인자의 협동적 결합을 통한 조합적 조절 메커니즘을 이해할 수 있다. 이러한 결과를 통합하여 유전자 발현을 조절하는 복잡한 조절 네트워크의 전체적인 구조를 규명한다.

9.5.5.2 후성유전학적 연구 후성유전학적 연구 분야에서 ChIP-seq는 크로마틴 구조와 기능을 이해하는 데 필수적인 기술이다. H3K4me3, H3K27me3 등 다양한 히스톤 변형의 유전체 전반에 걸친 분포 지도를 작성하여 크로마틴의 기능적 상태를 파악하고, 활성 및 억제 크로마틴 도메인을 식별한다. 특히 발달 과정이나 질병 상태에서 일어나는 후성유전학적 스위치 현상을 추적하여 크로마틴 상태의 동적 변화를 규명할 수 있다.

9.6 표적 시퀀싱과 캡처 기술

9.6.1 표적 시퀀싱의 필요성

전장 유전체 시퀀싱이나 전장 후성유전체 분석은 비용과 시간이 많이 소요되는 반면, 연구 목적에 따라서는 특정 영역에만 집중하는 것이 더 효율적일 수 있다. 표적 시퀀싱은 이러한 요구를 충족하는 비용 효율적인 대안이다.

9.6.2 엑솜 시퀀싱과 표적 후성유전체 분석

표적 시퀀싱은 특정 관심 영역에 대한 프로브를 사용하여 해당 영역의 DNA만 선택적으로 포획하고 시퀀싱하는 기술이다.

9.6.2.1 전체 엑솜 시퀀싱(Whole Exome Sequencing) 전체 엑솜 시퀀싱은 10만 개 이상의 엑솜 영역을 대상으로 하여 전체 유전체의 약 1-2%에 해당하는 코딩 영역을 포괄적으로 분석하는 기술이다. 이 방법은 단백질 기능에 직접적인 영향을 미치는 변이를 집중적으로 탐지할 수 있어 유전질환 진단 및 암 유전학 연구에서 널리 활용되고 있다.

Figure 8.8 표적 시퀀싱 - 특정 유전체 영역에 대한 선택적 시퀀싱 과정

9.6.2.2 표적 바이설파이트 시퀀싱 표적 바이설파이트 시퀀싱은 프로모터, 인핸서, CpG island 등 핵심 조절 영역에 집중하여 DNA 메틸화 패턴을 분석하는 기술이다. 이 접근법은 전장유전체 바이설파이트 시퀀싱(WGBS)에 비해 훨씬 경제적이면서도 특정 질병이나 생물학적 경로에 집중된 후성유전학 연구에서 높은 해상도의 결과를 제공한다.

9.7 ATAC-seq: 접근 가능한 크로마틴 분석

9.7.1 ATAC-seq의 기본 원리

ATAC-seq(Assay for Transposase-Accessible Chromatin using sequencing)은 Tn5 전이효소의 태그멘테이션(tagmentation) 반응을 이용하여 접근 가능한 크로마틴 영역을 식별하는 기술이다.

Figure 8.9 ATAC-seq 원리 - Tn5 전이효소를 이용한 접근 가능한 크로마틴 분석

9.7.2 Tn5 태그멘테이션의 분자적 기전

9.7.2.1 태그멘테이션 과정

1. **Tn5 결합**: 전이효소가 접근 가능한 DNA에 결합
2. **동시 절단 및 삽입**: DNA 절단과 어댑터 삽입을 한 번에 수행
3. **이중가닥 절단**: 약 300bp 간격으로 양쪽 가닥 모두 절단
4. **어댑터 부착**: 시퀀싱용 어댑터가 절단 부위에 직접 삽입

Figure 8.10 태그멘테이션 과정 - Tn5 전이효소의 동시 절단 및 어댑터 삽입

9.7.2.2 크로마틴 접근성과 단편 길이 크로마틴의 구조적 특성에 따라 Tn5 전이효소의 접근 정도와 생성되는 DNA 단편의 길이가 달라진다. 열린 크로마틴 영역에서는 전이효소가 쉽게 접근하여 짧은 단편을 생성하고 높은 시퀀싱 신호를 나타내는 반면, 뉴클레오솜이 결합된 영역에서는 약 200bp 주기의 특징적인 길이 분포를 보인다. 응축된 크로마틴에서는 전이효소의 접근이 제한되어 신호가 현저하게 감소한다.

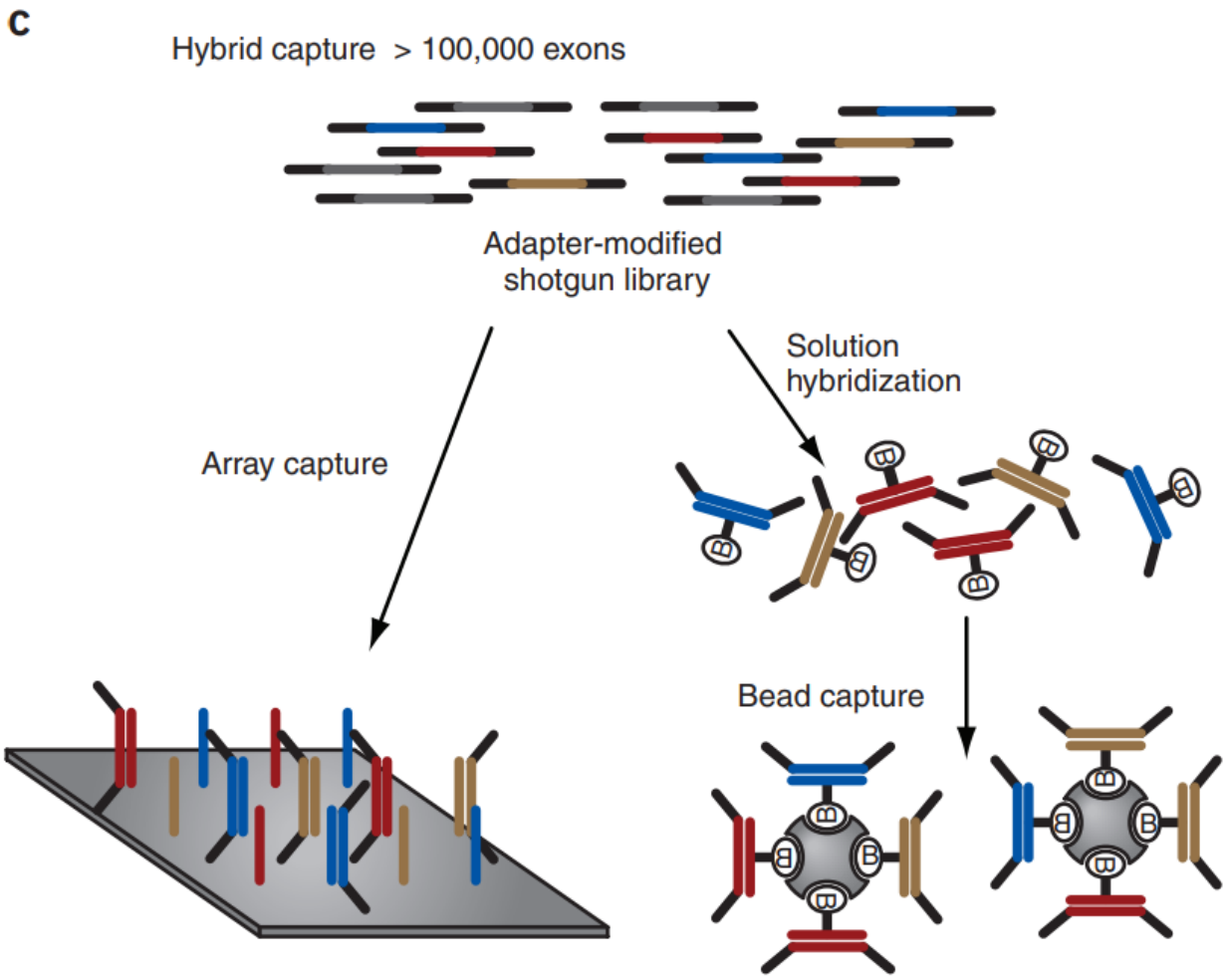


Figure 167: 표적 시퀀싱

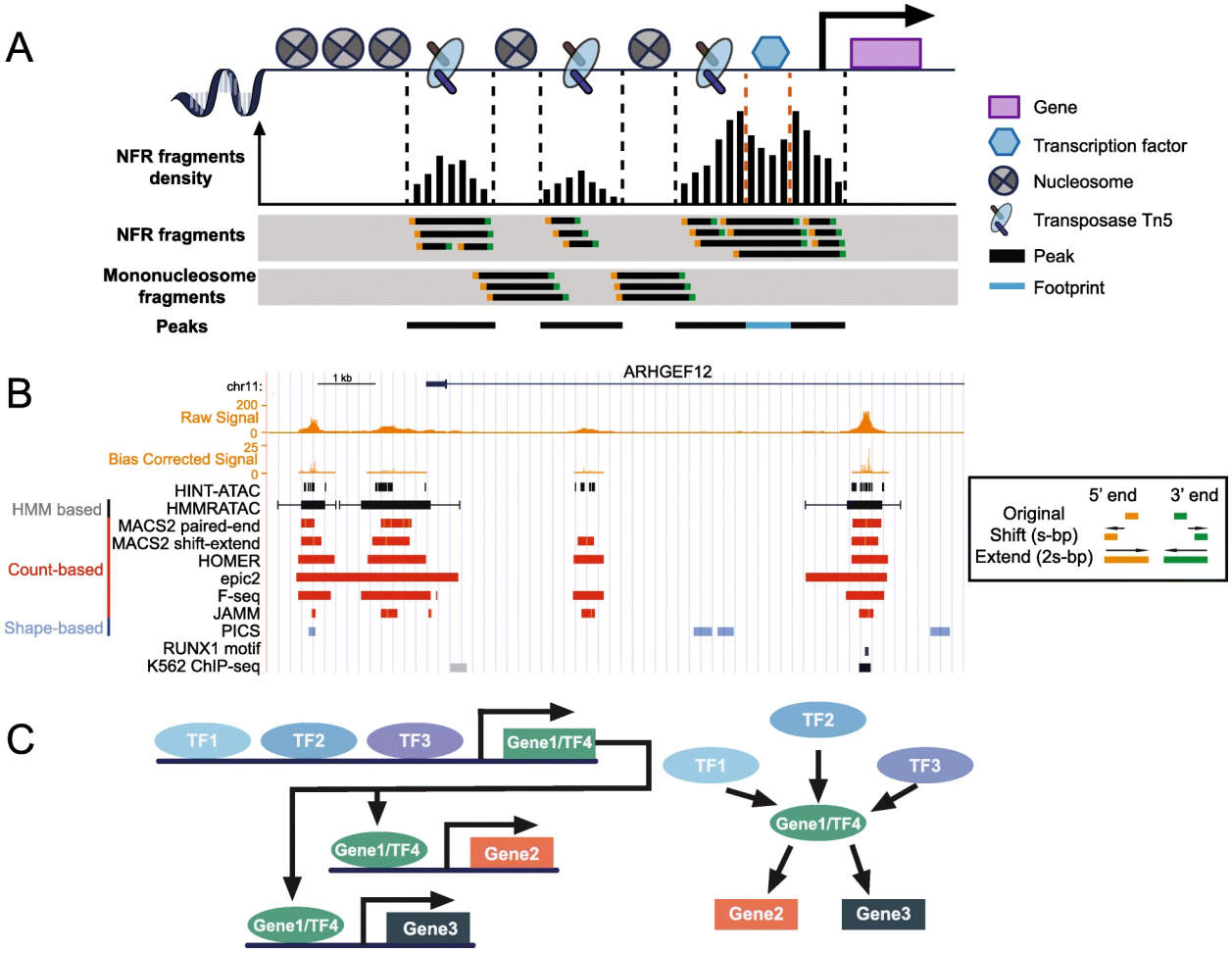


Figure 168: ATAC-seq 원리

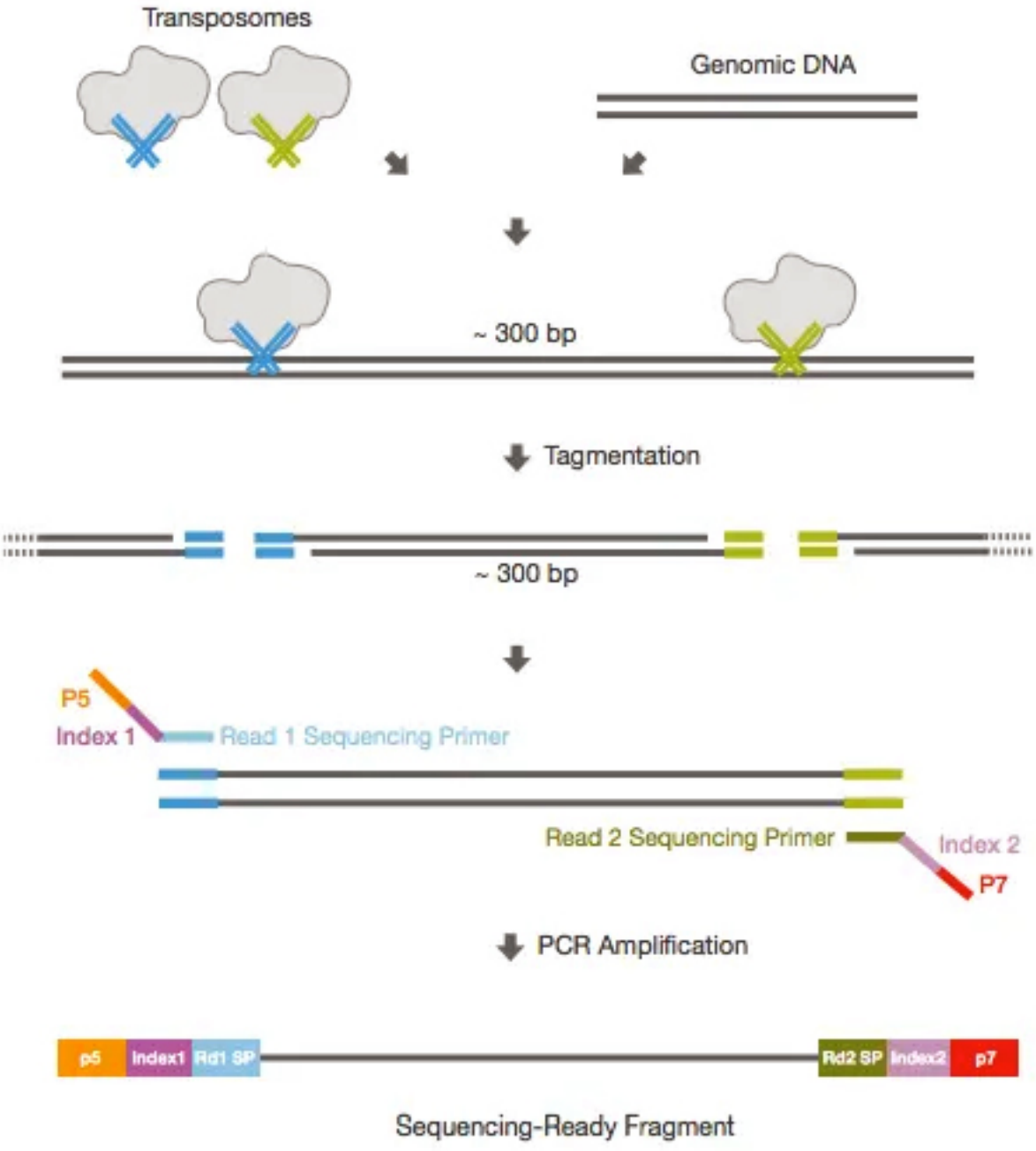


Figure 169: 태그멘테이션 과정

9.7.3 ATAC-seq 실험 과정

9.7.3.1 샘플 준비 ATAC-seq 실험의 성공을 위해서는 고품질 샘플 준비가 필수적이다. 신선한 세포를 분리하거나 핵만을 선택적으로 추출하여 사용하며, 일반적으로 50,000에서 100,000개의 세포를 사용한다. 전 과정에서 적절한 이온 강도와 pH를 유지하는 버퍼 조건을 유지하여 크로마틴 구조를 보존한다.

9.7.3.2 태그멘테이션 반응 태그멘테이션 반응은 ATAC-seq의 핵심 단계로, 37°C에서 30분에서 60분간 수행된다. 사용하는 세포 수에 따라 Tn5 전이효소의 농도를 신중하게 조절하여 최적의 태그멘테이션 효율을 달성하고, 반응 완료 후에는 EDTA 처리를 통해 효소 활성을 중단시켜 반응을 종료한다.

9.7.3.3 라이브러리 증폭 태그멘테이션 반응으로 생성된 DNA 단편에 삽입된 어댑터를 이용하여 PCR 증폭을 수행한다. 이 과정에서는 과도한 증폭으로 인한 편향을 방지하기 위해 적절한 PCR 사이클 수를 신중하게 결정해야 하며, 마지막으로 크기 선별 과정을 통해 원하는 크기 범위의 DNA 단편만을 선택적으로 회수한다.

9.7.4 ATAC-seq 데이터 분석

9.7.4.1 품질 평가 지표

- **TSS 농축도**: 전사 시작 지점 주변의 신호 강화 정도
- **뉴클레오솜 패턴**: 단편 길이 분포에서 200bp 주기성 확인
- **미토콘드리아 DNA 비율**: 핵 DNA의 순도 평가

9.7.4.2 피크 호출 및 분석 ATAC-seq 데이터로부터 의미 있는 생물학적 정보를 추출하기 위해 피크 호출 및 분석 과정을 수행한다. 접근성 피크는 통계적으로 유의한 신호 강화를 보이는 영역으로 활성 조절 요소를 나타내며, 전사인자 풋프린팅 분석을 통해 피크 내에서 보호된 영역을 식별하여 결합 인자를 예측할 수 있다. 또한 조직이나 세포 간 차별적 접근성 패턴을 비교하여 세포형 특이적 인핸서를 식별한다.

9.8 크로마틴 형태 포획(3C 및 Hi-C)

9.8.1 크로마틴 형태 포획의 개념

크로마틴 형태 포획(Chromatin Conformation Capture, 3C)은 물리적으로 근접한 크로마틴 영역 간의 상호작용을 탐지하여 3차원 유전체 구조를 분석하는 기술이다.

Figure 8.11 크로마틴 형태 포획 - 3C 기술을 이용한 크로마틴 상호작용 분석

9.8.2 3C 기술의 원리

9.8.2.1 기본 과정 3C 기술은 5단계의 순차적 과정을 통해 크로마틴 상호작용을 분석한다. 첫 번째 단계에서는 포름알데하이드를 사용하여 살아있는 세포 내에서 일어나는 크로마틴 상호작용을 고정하고, 두 번째 단계에서는 제한효소를 처리하여 크로마틴을 특정 부위에서 절단한다. 세 번째 단계에서는 희석된 조건에서 DNA 연결효소를 이용하여 공간적으로 근접한 DNA 단편들을 분자간 연결시키며, 네 번째 단계에서는 역교차결합을 통해 단백질을 제거하고 DNA를 회수한다. 마지막 단계에서는 PCR 또는 시퀀싱을 통해 상호작용하는 DNA 영역들을 검출하고 정량화한다.

9.8.2.2 생물학적 의의 근접한 크로마틴 영역들이 서로 상호작용하여 그 양을 측정함으로써 3차원 공간에서의 유전체 구조를 추론할 수 있다.

Figure 8.12 3차원 유전체 구조 - 크로마틴 상호작용을 통한 공간적 구조

9.8.3 Hi-C: 유전체 전체 크로마틴 상호작용

9.8.3.1 Hi-C의 특징 Hi-C 기술은 3C 기술의 발전된 형태로 여러 독특한 특징을 가지고 있다. 전유전체 접근법을 통해 유전체 내 모든 가능한 크로마틴 상호작용을 동시에 분석할 수 있으며, 수 킬로베이스 단위의 고해상도로 상호작용 지도를 작성할 수 있다. 또한 상호작용 빈도를 정량적으로 측정하여 크로마틴 상호작용의 강도를 정확하게 평가할 수 있다.

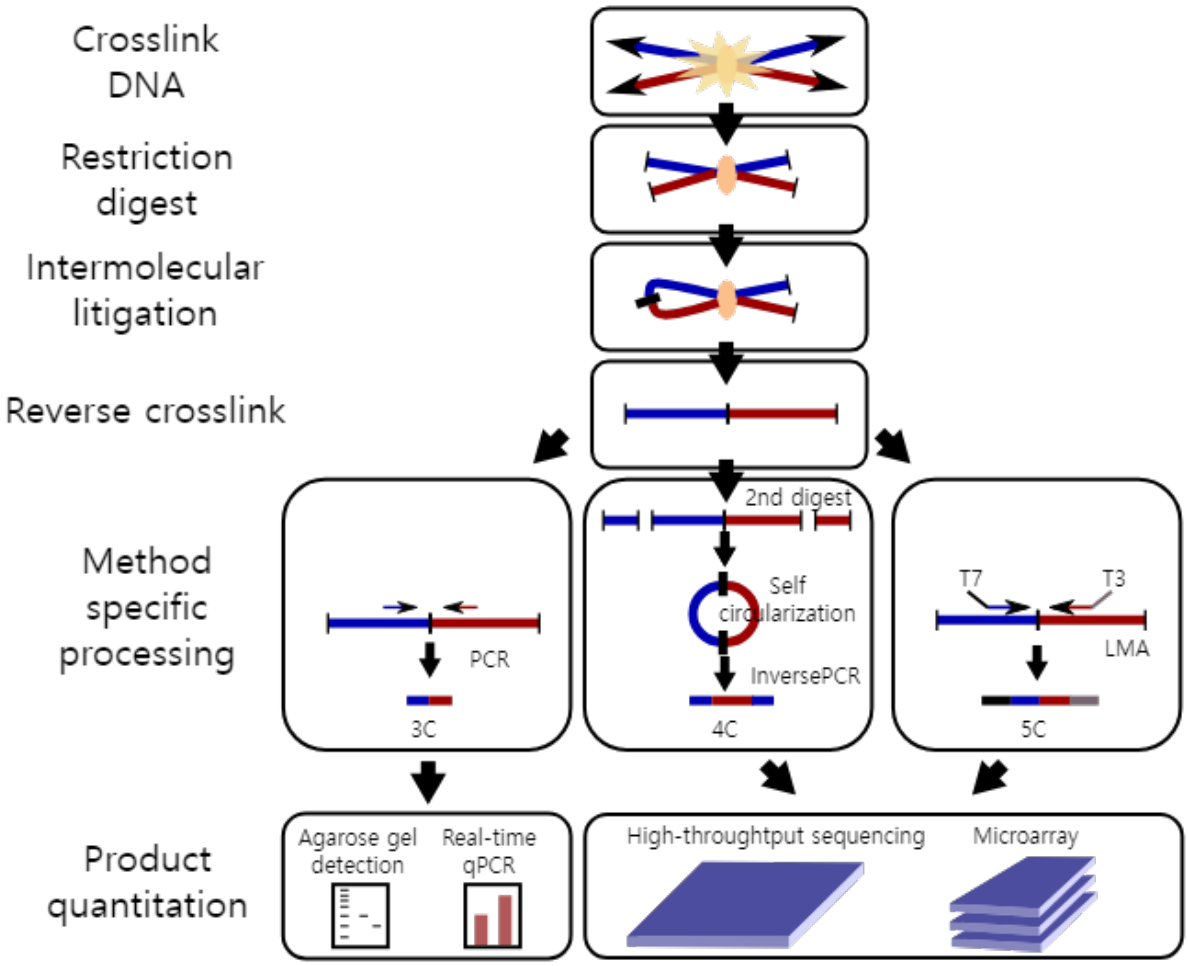


Figure 170: 크로마틴 형태 포획

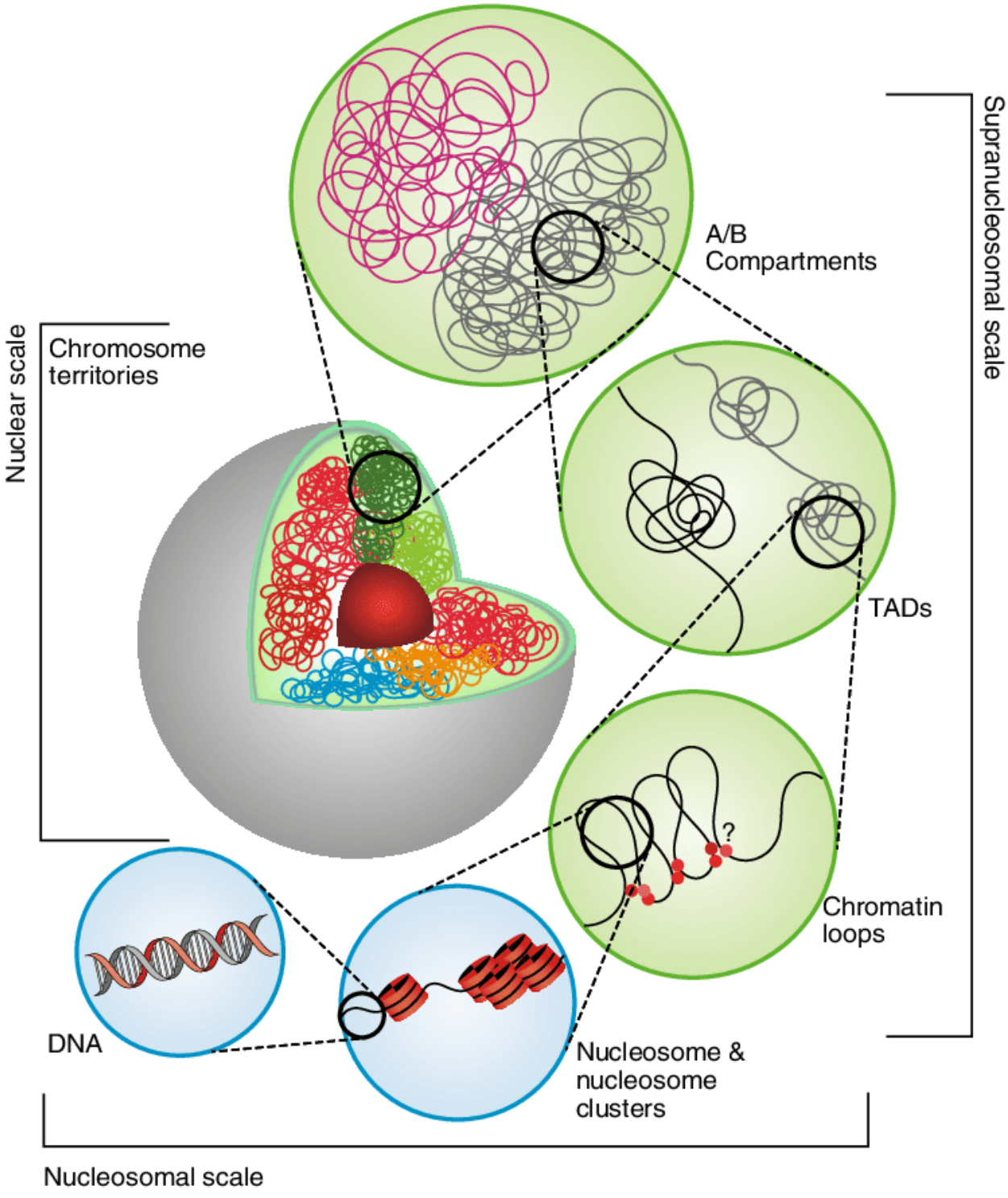


Figure 171: 3차원 유전체 구조

9.8.3.2 Hi-C 데이터의 활용 Hi-C 데이터는 3차원 유전체 구조를 이해하는 데 다양하게 활용된다. 프로모터와 인핸서 간의 장거리 상호작용 등 유전자 발현 조절에 중요한 3차원 구조를 규명한다.

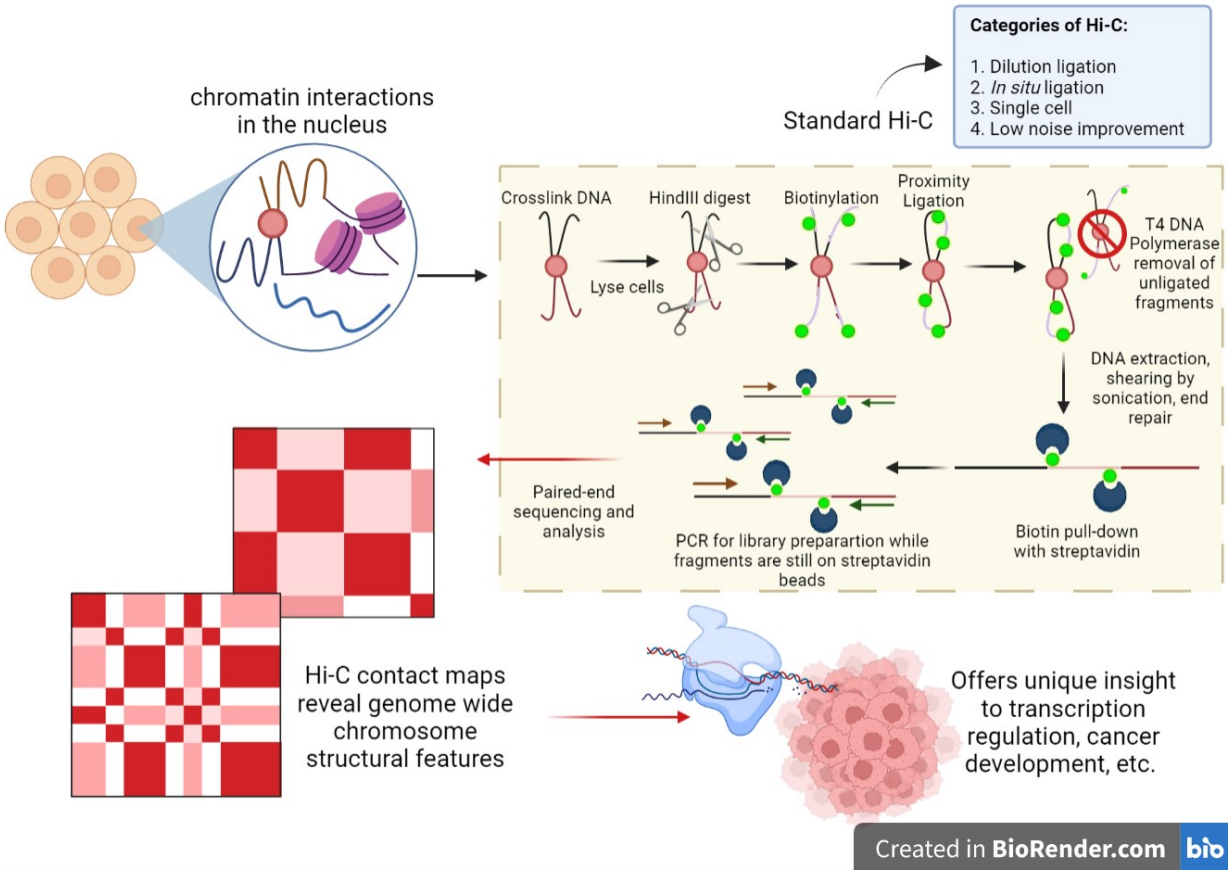


Figure 172: Hi-C 기술

Figure 8.13 Hi-C 기술 – 전유전체 크로마틴 상호작용 분석 과정

Figure 8.14 Hi-C 결과 – 크로마틴 상호작용 빈도 지도와 구조적 도메인

9.9 피크 호출 방법론

9.9.1 피크 호출의 기본 원리

피크 호출은 ChIP-seq, ATAC-seq 등 후성유전체 데이터에서 통계적으로 유의한 신호 영역을 식별하는 핵심 과정이다. 기본적으로 배경 신호 대비 현저히 높은 시퀀싱 읽기 수를 보이는 연속된 유전체 영역을 찾아내며, 이때 거짓 양성을 최소화하면서 진정한 결합 부위나 활성 영역을 정확히 검출하는 것이 목표이다.

9.9.2 피크 형태와 특성

9.9.2.1 피크 유형별 특징 후성유전체 데이터에서 관찰되는 피크는 표적 단백질과 실험 조건에 따라 서로 다른 형태를 보인다. 전사인자 결합 부위는 일반적으로 좁고 뾰족한 피크(Sharp peaks)를 형성하며 수백 염기쌍 범위에서 강한 신호를 나타낸다. 반면 히스톤 변형은 넓고 완만한 피크(Broad peaks)를 형성하여 수 킬로염기쌍에 걸쳐 분포하며, 크로마틴 접근성 신호는 중간 정도의 폭을 가진 피크를 보인다.

9.9.2.2 신호 분포 패턴 각 피크 유형은 고유한 신호 분포 패턴을 가진다. 전사인자 피크는 중심에서 최고점을 보이고 양쪽으로 급격히 감소하는 패턴을 보이며, 히스톤 변형 피크는 평탄한 고원 형태나 여러 개의 작은 봉우리가 연결된 형태를

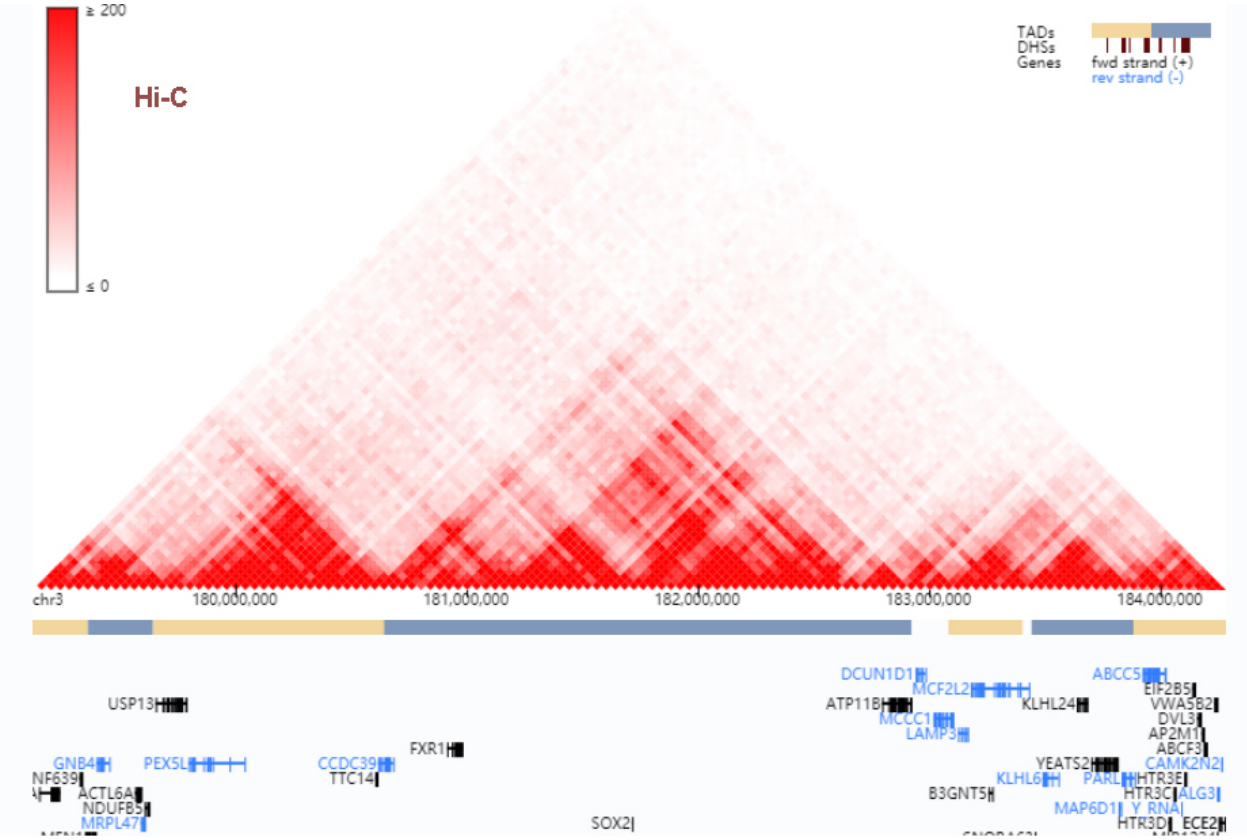


Figure 173: Hi-C 결과

나타낸다. 이러한 패턴의 차이는 피크 호출 알고리즘 선택에 중요한 기준이 된다.

9.9.3 피크 호출 방법론

9.9.3.1 통계적 접근법 MACS2, HOMER 등의 전통적인 피크 호출 도구는 포아송 분포나 음이항 분포를 기반으로 한 통계적 모델을 사용한다. 이들은 국소적 배경 신호를 추정하고 이를 기준으로 통계적 유의성을 계산하여 피크를 식별한다. 은닉 마르코프 모델(HMM)을 활용한 방법은 인접한 위치 간의 공간적 연속성을 고려하여 더 정확한 피크 경계를 결정할 수 있다.

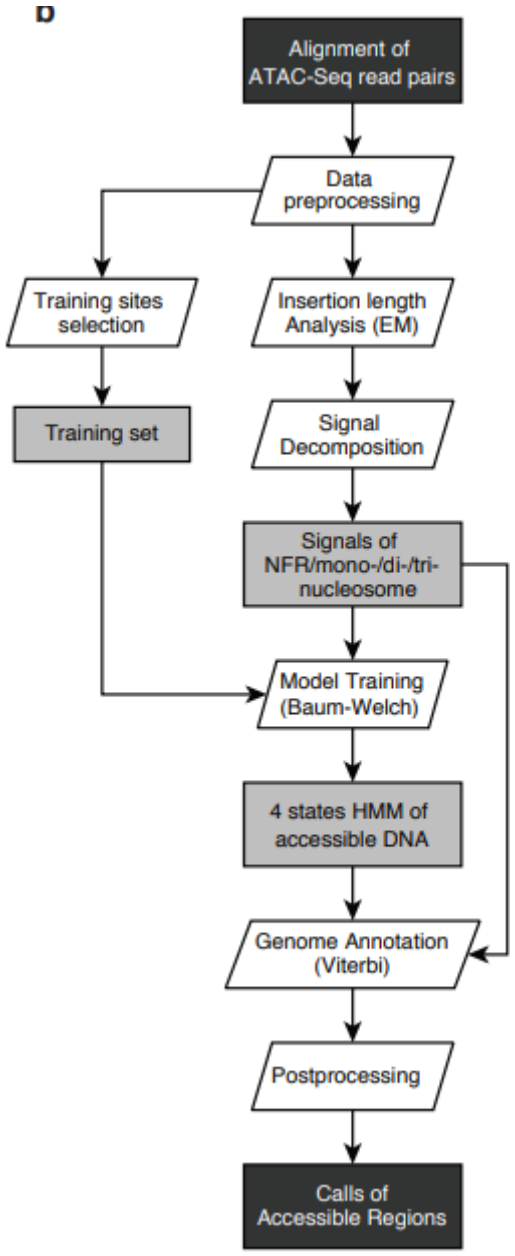


Figure 174: ATAC-seq 피크 호출

Figure 8.15 ATAC-seq 피크 호출 – HMM 기반 피크 호출 예제와 결과

9.9.3.2 기계학습 기반 접근법 최근에는 딥러닝을 포함한 기계학습 기법이 피크 호출에 적용되고 있다. 합성곱 신경망(CNN)은 피크의 형태적 특징을 학습하여 복잡한 신호 패턴을 인식할 수 있으며, 순환 신경망(RNN)은 시계열 신호의 장거리 의존성을 모델링할 수 있다. 이러한 방법들은 대용량 훈련 데이터를 통해 다양한 실험 조건에서 robust한 성능을 보이는 장점이 있다.

9.9.3.3 통합적 접근법 현대적인 피크 호출 파이프라인은 여러 방법을 조합하여 사용한다. 초기 후보 영역을 통계적 방법으로 식별한 후 기계학습 모델로 정제하거나, 서로 다른 알고리즘의 결과를 ensemble하여 신뢰도를 높이는 접근법이 널리 사용되고 있다. 또한 멀티모달 데이터를 활용하여 여러 후성유전체 신호를 동시에 고려하는 통합적 피크 호출 방법도 개발되고 있다.

9.10 결론

후성유전체학은 DNA 서열 변화 없이도 유전자 발현이 조절되는 메커니즘을 연구하는 분야로, 현대 의생명정보학에서 핵심적인 역할을 담당하고 있다. 유전체 서열 정보만으로는 설명할 수 없는 생물학적 현상들이 점점 더 많이 발견되면서 후성유전체학의 중요성이 더욱 부각되고 있다.

동일한 유전체 서열을 가진 일란성 쌍둥이의 표현형 차이, 환경 요인에 의한 유전자 발현 변화, GWAS 변이의 기능적 해석 등은 모두 후성유전학적 정보 없이는 완전히 이해할 수 없는 현상들이다. 이에 따라 현대 생의학 연구는 "무엇이 있는가"를 보여주는 유전체 정보에서 "언제, 어디서, 어떻게 조절되는가"를 설명하는 후성유전체 정보를 함께 분석하는 통합적 접근법을 채택하고 있다.

WGBS, ChIP-seq, ATAC-seq, Hi-C 등 다양한 후성유전체 기술들을 통합적으로 활용함으로써 암, 신경질환, 대사질환 등의 분야에서 새로운 이해를 얻고 있으며, 단일세포 분석과 기계학습 기법의 발전과 함께 개인 맞춤 의학 실현에 핵심적인 기여를 할 것으로 기대된다.

10 전사체학 기초

10.1 전사체학 개요

전사체학(Transcriptomics)은 특정 조건에서 세포 또는 조직이 발현하는 모든 RNA 분자를 종합적으로 분석하는 학문 분야이다. 유전체학이 생명체가 가진 유전적 잠재력을 보여준다면, 전사체학은 특정 시점과 조건에서 실제로 활성화되고 있는 유전자들의 발현 패턴을 제시한다. 이는 생명 현상을 이해하는 데 있어 정적인 설계도에서 동적인 실행 과정으로의 관점 전환을 의미한다.

전사체학은 다양한 생물학적 질문에 답할 수 있다. 예를 들어, 특정 질병 상태에서 어떤 유전자들이 비정상적으로 발현되는지, 약물 치료에 대한 세포의 반응은 어떤지, 발생 과정에서 세포가 어떻게 분화하는지 등을 분자 수준에서 규명할 수 있다. 또한 환경 변화나 스트레스에 대한 생물체의 적응 반응을 전사체 차원에서 해석할 수 있다.

Figure 9.1 유전자 발현 히트맵 – 다양한 조건에서의 유전자 발현 패턴 비교

10.2 마이크로어레이 기술

10.2.1 마이크로어레이의 기본 원리

마이크로어레이(microarray)는 RNA-seq 이전에 유전자 발현을 대규모로 분석하는 주요 기술이다. 이 기술은 고체 기판(주로 유리 슬라이드) 위에 수천에서 수만 개의 DNA 프로브를 배열하여 특정 유전자의 발현량을 측정한다. 각 프로브는 특정 유전자 서열에 상보적인 단일가닥 DNA 올리고뉴클레오타이드로 구성되며, 표적 RNA와 혼성화(hybridization)를 통해 결합한다.

Figure 9.2 마이크로어레이 – Affymetrix GeneChip 마이크로어레이 칩의 실제 모습

실험 과정에서는 먼저 샘플에서 추출한 RNA를 역전사하여 cDNA로 변환한 후, 형광 염료로 표지한다. 일반적으로 대조군과 실험군을 서로 다른 색깔의 형광 염료(Cy3, Cy5)로 표지하여 동일한 칩에서 동시에 분석할 수 있다. 표지된 cDNA가 칩 위의 해당 프로브와 혼성화하면 형광 신호가 발생하며, 이 신호의 강도가 해당 유전자의 발현량에 비례한다.

Figure 9.3 마이크로어레이 유전자 발현 – 형광 신호를 통한 발현량 측정

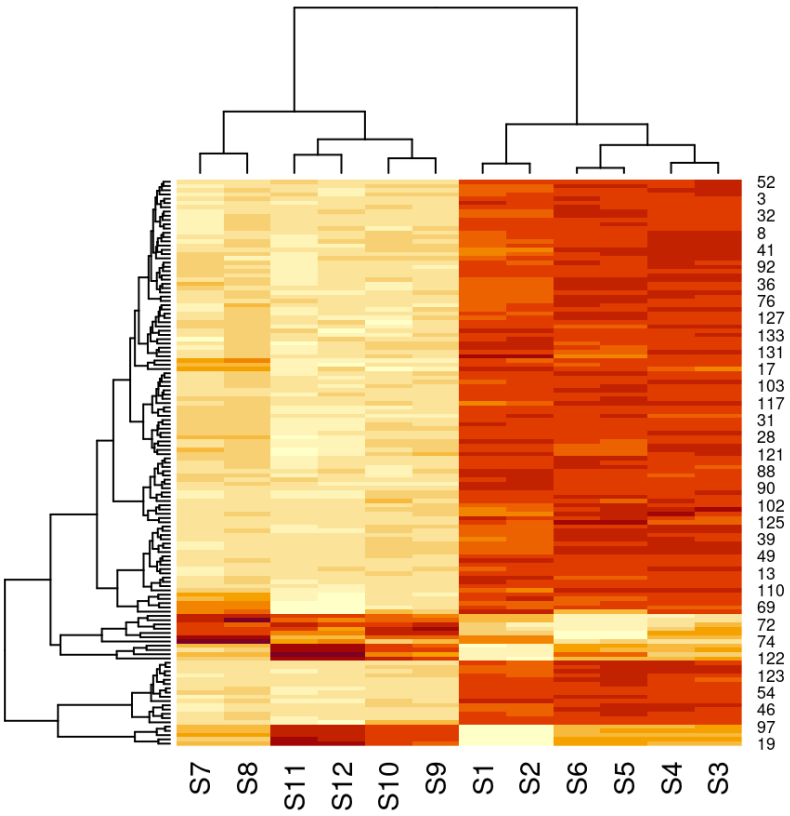


Figure 175: 유전자 발현 히트맵

Figure 176: 마이크로어레이

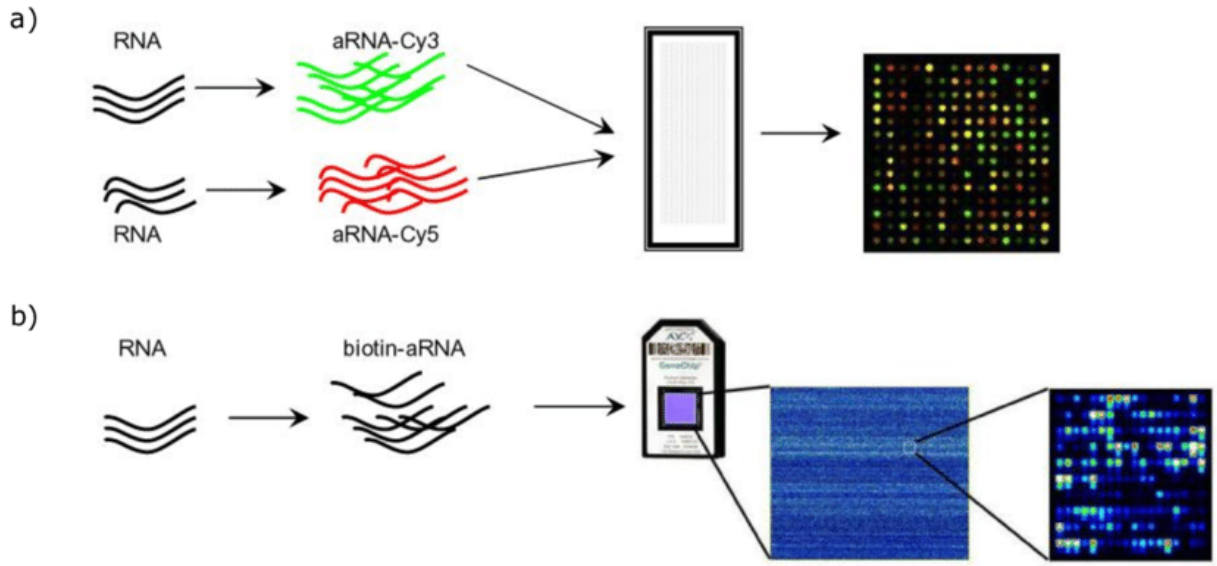


Figure 177: 마이크로어레이 유전자 발현

10.2.2 마이크로어레이의 장단점

장점으로는 상대적으로 저렴한 비용과 표준화된 프로토콜로 인한 높은 재현성을 들 수 있다. 또한 잘 알려진 유전자들에 대해서는 매우 정확한 정량이 가능하며, 대용량 샘플 처리에 적합하다. 데이터 분석 파이프라인이 성숙되어 있어 결과 해석이 용이하다.

단점으로는 미리 설계된 프로브에 해당하는 유전자만 측정할 수 있어 새로운 전사체나 이소폼을 발견할 수 없다는 제한이 있다. 또한 발현량이 매우 높거나 낮은 유전자의 정확한 측정이 어려우며, 특이성이 낮은 문제도 한계로 지적된다.

10.2.3 마이크로어레이 데이터 분석

마이크로어레이 데이터 분석은 여러 전처리 단계를 거친다. **배경 보정(background correction)**에서는 비특이적 혼성화로 인한 배경 신호를 제거하고, **정규화(normalization)**를 통해 칩 간의 기술적 변이를 보정한다. 대표적인 정규화 방법으로는 분위수 정규화(quantile normalization)와 RMA(Robust Multi-array Average) 방법이 있다.

차등 발현 분석에서는 t-검정, ANOVA, 또는 더 정교한 통계 모델을 사용하여 조건 간 발현 차이를 검정한다. 다중 검정 보정을 통해 위양성률을 조절하며, fold change와 통계적 유의성을 동시에 고려하여 생물학적으로 의미 있는 유전자를 선별한다.

10.3 정량적 중합효소연쇄반응(qPCR)

10.3.1 qPCR의 기본 원리

정량적 중합효소연쇄반응(quantitative PCR, qPCR 또는 real-time PCR)은 특정 유전자의 발현량을 정확하고 신속하게 측정할 수 있는 기술이다. 전사체학 연구에서는 주로 RT-qPCR(Reverse Transcription qPCR) 형태로 사용되며, 먼저 RNA를 cDNA로 역전사한 후 특정 유전자 서열을 실시간으로 증폭하면서 그 양을 정량화한다.

qPCR의 핵심은 PCR 증폭 과정에서 실시간으로 형성되는 PCR 산물의 양을 모니터링하는 것이다. 매 사이클마다 측정되는 형광 신호의 강도는 초기 주형 DNA(또는 cDNA)의 양에 비례하므로, 이를 통해 원래 샘플에 존재했던 RNA의 양을 역추적할 수 있다.

10.3.2 qPCR의 장단점과 응용

장점으로는 높은 정확도와 정밀도, 빠른 분석 속도(2-3시간), 넓은 발현 범위(7-8 log), 높은 특이성을 들 수 있다. 또한 상대적으로 저렴한 비용으로 소수의 유전자를 정확하게 분석할 수 있어 마이크로어레이나 RNA-seq 결과의 검증에 자주 사용된다.

단점으로는 한 번에 분석할 수 있는 유전자 수가 제한적이며(일반적으로 1-100개), 새로운 전사체나 스플라이싱 변이를 발견할 수 없다. 또한 참조 유전자의 선택이 결과에 큰 영향을 미칠 수 있다.

10.4 Bulk RNA 시퀀싱

RNA 시퀀싱(RNA-seq)은 차세대 시퀀싱(NGS) 기술을 기반으로 한 전사체 분석 기법이다. 마이크로어레이나 qPCR과 달리 RNA-seq은 미리 설계된 프로브나 프라이머 없이도 샘플에 존재하는 모든 RNA 분자를 정량화할 수 있다. 이는 알려지지 않은 전사체나 새로운 스플라이싱 변이체의 발견을 가능하게 하며, 전장 유전체 수준에서 전사체의 전체적인 그림을 제공한다.

NGS 기술의 발전과 비용 감소로 인해 RNA-seq은 현재 전사체학 연구의 표준 방법으로 자리잡았다. 한 번의 실험으로 수만 개의 유전자를 동시에 분석할 수 있으며, 매우 넓은 발현 범위와 높은 해상도를 제공한다. 또한 정량적 분석뿐만 아니라 새로운 전사체 발견, 대립유전자별 발현 분석, 융합 유전자 검출 등 다양한 분석이 가능하다.

10.4.1 실험 과정

Bulk RNA 시퀀싱은 많은 수의 세포에서 추출한 RNA를 집합적으로 분석하는 방법이다. 실험 과정은 여러 단계로 구성되며, 각 단계는 정확한 전사체 정보를 얻기 위해 필수적이다.

전통적인 RNA-seq 방법에서는 Polyadenylation을 이용한 mRNA 선택이 중요한 첫 단계이다. Polyadenylation은 전사 후 과정으로, mRNA의 3' 끝에 약 200개의 아데닌(A) 뉴클레오타이드가 연속으로 붙는 현상이다. 이 polyA 꼬리는 mRNA의 안정성을 증가시키고, 세포질로의 이동과 번역을 조절하는 역할을 한다. RNA-seq 실험에서는 이 polyA 꼬리에 올리고-dT 프라이머를 결합시켜 mRNA를 선택적으로 포획한다.

Figure 9.4 폴리아데닐화 - mRNA 3' 끝의 polyA 꼬리 형성과 선택적 포획

포획 또는 선택된 RNA는 이후 적절한 크기로 단편화되고, 역전사효소를 이용하여 cDNA로 합성되며, 그 이후는 일반적인 NGS 시퀀싱 라이브러리 준비 과정과 동일하다.

Figure 9.5 RNA-seq 라이브러리 구축 - RNA에서 시퀀싱 라이브러리까지의 과정

Poly-A 기반 RNA 시퀀싱의 한계

하지만 이러한 polyA 선택 방법에는 중요한 한계가 있다. 첫째, 많은 비암호화 RNA(lncRNA, circRNA 등)들이 polyA 꼬리를 갖지 않아 분석에서 제외된다. 둘째, 히스톤 mRNA처럼 polyA 꼬리가 없는 일부 코딩 RNA도 놓치게 된다. 셋째, RNA 분해 과정에 있거나 부분적으로 손상된 RNA는 polyA 꼬리가 짧아져 포획되지 않을 수 있다. 이러한 한계를 극복하기 위해 Total RNA-seq이 개발되기도 하였다.

10.4.2 데이터 정량화와 정규화

RNA-seq 데이터에서 유전자 발현을 정량화할 때는 단순히 매핑된 리드 수를 세는 것만으로는 충분하지 않다. 유전자의 길이에 따라 더 많은 리드가 매핑될 수 있으며, 이를 보정해야 정확한 발현 수준을 측정할 수 있다. 길이가 긴 유전자는 자연스럽게 더 많은 리드를 생성하므로, 실제 발현량을 비교하기 위해서는 길이 보정이 필수적이다.

이러한 문제를 해결하기 위해 FPKM(Fragments Per Kilobase of transcript per Million mapped reads)과 TPM(Transcripts Per Million) 같은 정규화 방법이 사용된다.

Figure 9.6 RNA-seq 정규화 필요성 - 유전자 길이와 라이브러리 크기 보정의 중요성

FPKM (Fragments Per Kilobase of transcript per Million mapped reads):

```
FPKM = (Raw Counts × 10^9) / (Gene Length in Kbp × Total Mapped Reads)
```

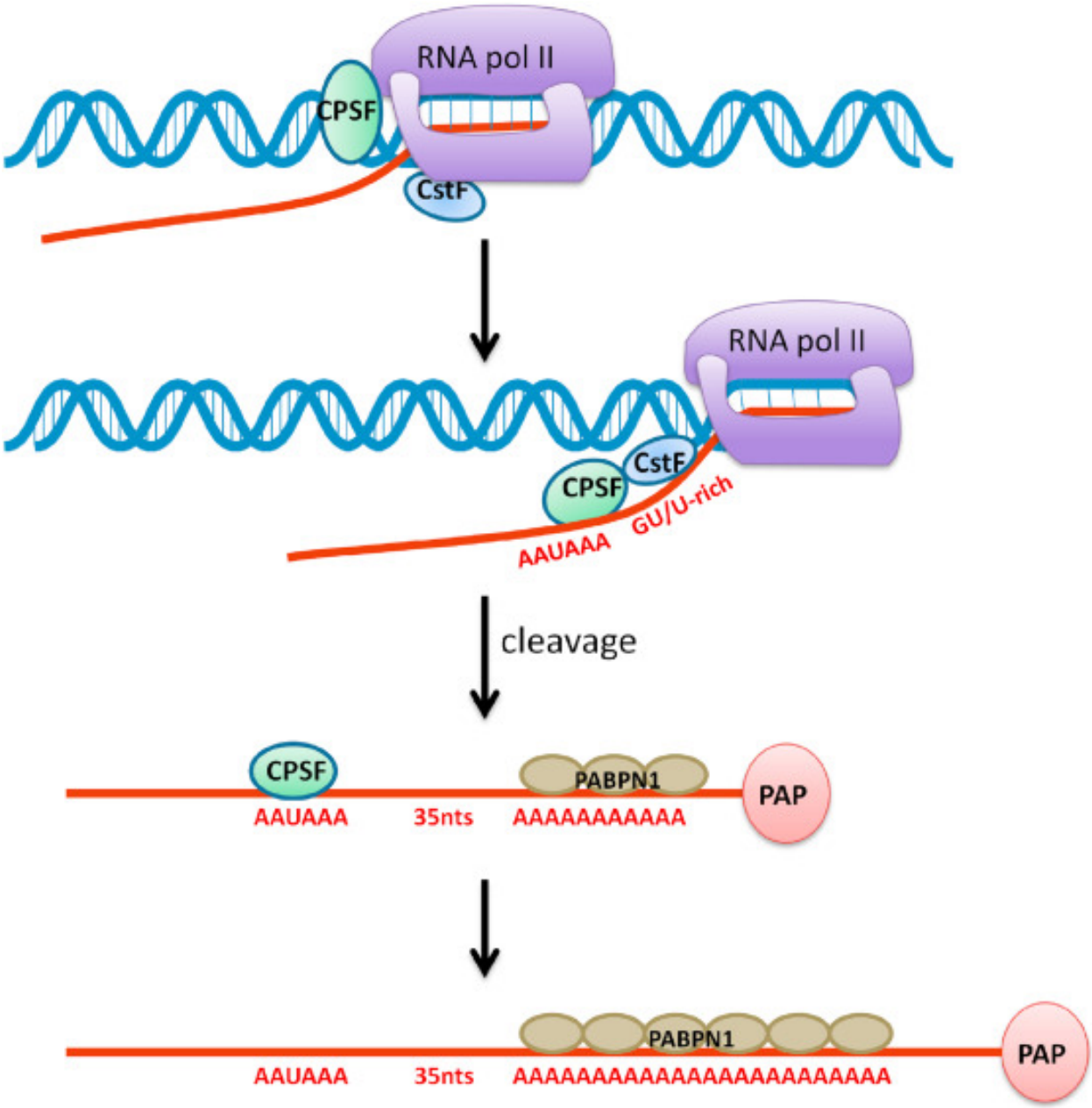


Figure 178: 폴리아데닐화

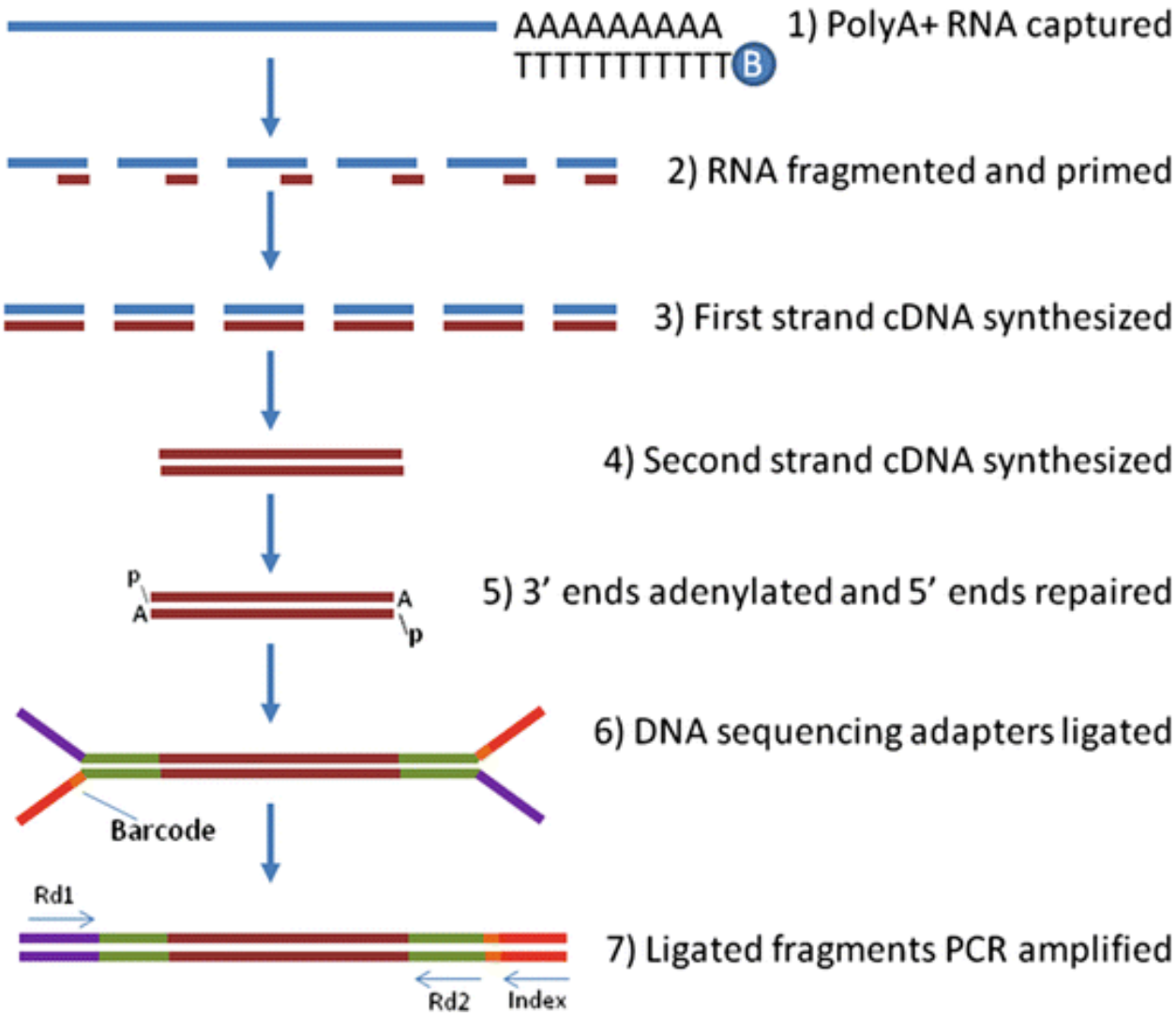


Figure 179: RNA-seq 라이브러리 구축

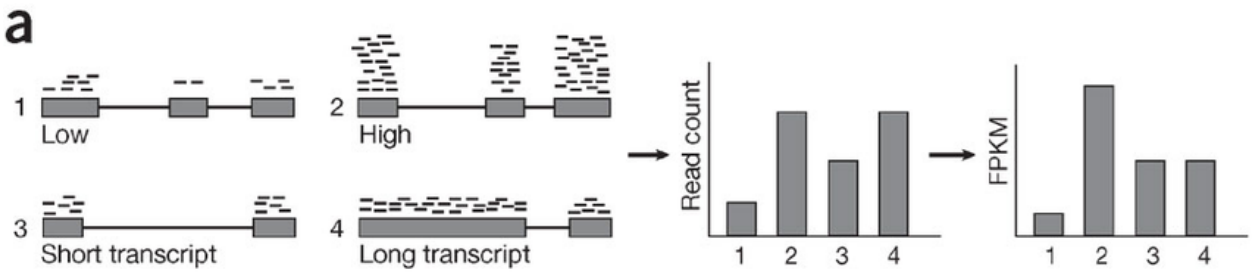


Figure 180: RNA-seq 정규화 필요성

FPKM은 원시 카운트를 유전자 길이(염기쌍 단위)와 총 매핑된 리드 수로 나누어 계산한다. 계산 과정에서 10^9을 곱하는 이유는 킬로베이스와 백만 단위로 정규화하기 위함이다.

TPM (Transcripts Per Million):

```
1. RPK = Raw Counts / Gene Length in Kbp
2. TPM = RPK / (Sum of all RPK) × 1,000,000
```

TPM은 먼저 각 유전자의 원시 카운트를 길이로 나누어 rate를 계산한 후, 모든 유전자의 rate 합으로 나누고 백만을 곱하여 정규화한다. 흥미롭게도 TPM은 RPK 대신에 FPKM 값을 사용하여 계산해도 동일한 결과를 얻을 수 있다. 즉, `TPM = FPKM / (Sum of all FPKMs) × 1,000,000`로도 계산이 가능하다.

정규화의 필요성을 이해하기 위해 다음과 같은 예시를 살펴보자. 5개 유전자에 대한 원시 데이터가 있다고 가정하면: - 유전자 A: 길이 1Kbp, 리드 수 200개 - 유전자 B: 길이 1.2Kbp, 리드 수 300개 - 유전자 C: 길이 0.8Kbp, 리드 수 150개 - 유전자 D: 길이 1.5Kbp, 리드 수 400개 - 유전자 E: 길이 0.9Kbp, 리드 수 250개

단순히 리드 수만 비교하면 D 〉 B 〉 E 〉 A 〉 C 순서가 되지만, 실제로는 유전자 길이를 고려해야 한다. 위 예시에서 총 매핑된 리드가 1천만 개 (10^7 개) 라고 가정하면, FPKM 공식을 사용하여 다음과 같이 계산할 수 있다:

```
FPKM = (Raw Counts × 10^9) / (Gene Length in Kbp × Total Mapped Reads)
     = (Raw Counts / Gene Length in Kbp) × (10^9 / Total Mapped Reads)
```

- 유전자 A: (200/1) / 10 = 20 FPKM
- 유전자 B: (300/1.2) / 10 = 25 FPKM
- 유전자 C: (150/0.8) / 10 = 18.75 FPKM
- 유전자 D: (400/1.5) / 10 = 26.67 FPKM
- 유전자 E: (250/0.9) / 10 = 27.78 FPKM

이제 발현 순서가 E 〉 D 〉 B 〉 A 〉 C로 바뀌어, 유전자 길이를 보정한 실제 발현 수준을 반영한다.

동일한 예시에서 TPM을 계산해보면, 모든 유전자의 RPK의 합이 20,000이라고 가정할 때: - 유전자 A: 200/1000 = 0.2 RPK - 유전자 B: 300/1200 = 0.25 RPK - 유전자 C: 150/800 = 0.1875 RPK - 유전자 D: 400/1500 = 0.267 RPK - 유전자 E: 250/900 = 0.278 RPK

Sum of RPK = 20,000

각 유전자의 TPM: - 유전자 A: (0.2/20,000) × 1,000,000 = 10 TPM - 유전자 B: (0.25/20,000) × 1,000,000 = 12.5 TPM - 유전자 C: (0.1875/20,000) × 1,000,000 = 9.375 TPM - 유전자 D: (0.267/20,000) × 1,000,000 = 13.33 TPM - 유전자 E: (0.278/20,000) × 1,000,000 = 13.89 TPM

TPM의 총합은 정확히 1,000,000이 되며, 발현 순서는 FPKM과 동일하게 E 〉 D 〉 B 〉 A 〉 C이다.

두 방법의 주요 차이점은 정규화 순서에 있다. FPKM은 먼저 전체 라이브러리 크기로 정규화한 후 유전자 길이로 보정하는 반면, TPM은 먼저 유전자 길이로 정규화한 후 라이브러리 크기를 보정한다. 이로 인해 TPM은 각 샘플 내에서 모든 유전자의 TPM 값의 합이 정확히 1,000,000이 되는 특성을 가지며, 이는 샘플 간 비교에서 더 일관성 있는 결과를 제공한다. 최근에는 DESeq2나 edgeR 같은 도구들이 사용하는 정규화된 카운트 방법들이 차등 발현 분석에서 더 선호되고 있다.

10.4.3 유전자 발현의 확률적 모델

유전자 발현은 본질적으로 확률적(stochastic) 과정으로 이해할 수 있다. 즉, RNA Pol II가 DNA 상의 여러 유전자들 중 무작위로 특정 유전자에 결합하는 것으로 모델링할 수 있다. 이러한 무작위적 확률 과정을 “랜덤 프로세스” 라 부른다.

사고실험을 통해 랜덤 프로세스로부터 유전자 발현을 모델링하면, 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 나타내는 포아송 분포(Poisson distribution)가 자연스럽게 도출된다. 포아송 분포는 다음과 같은 확률 질량 함수를 가진다:

$f(x; \theta) = (\theta^x e^{(-\theta)}) / x!$

여기서 θ = (평균) = (분산)으로, 포아송 분포의 핵심 특징은 평균과 분산이 같다는 점이다. 그러나 유전자 발현을 포함하여 자연에서 발견되는 실제 데이터는 많은 경우 평균이 높아질 때 그보다 분산이 높아지는 과대산포(Overdispersion) 현상을 보인다.

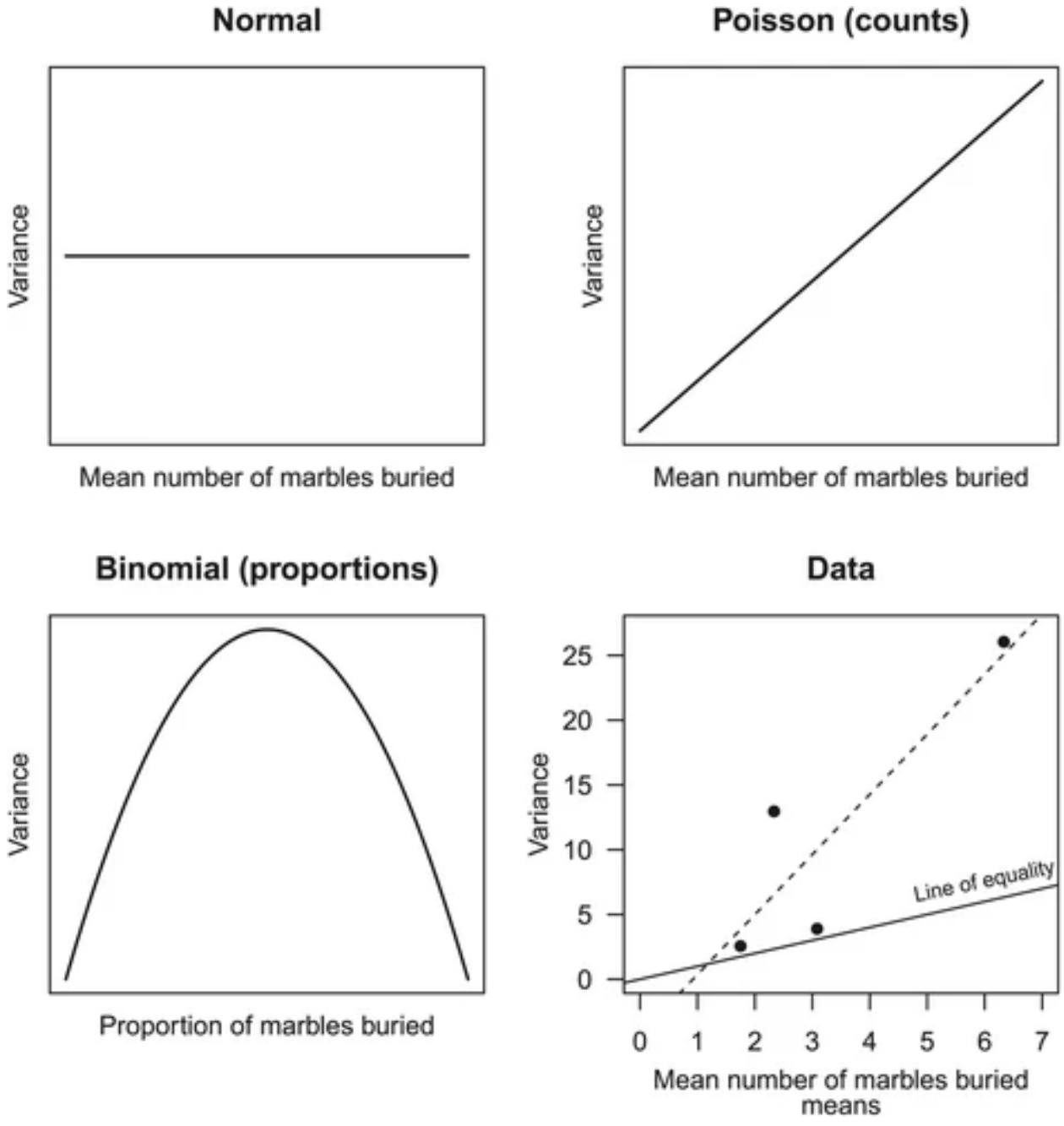


Figure 181: 평균-분산 관계와 과대산포

Figure 9.7 평균-분산 관계와 과대산포 - 포아송 분포 대비 실제 데이터의 과대산포 현상

실제 유전자 발현 데이터를 분석해보면 포아송 분포의 기본 가정과 달리 과대산포 현상이 광범위하게 관찰된다. 이는 유전자 발현이 완전한 무작위 과정이 아님을 나타낸다. 예를 들어, Bulk RNA-seq 데이터의 mean-variance plot에서는 대부분의 유전자에서 분산이 평균보다 훨씬 큰 값을 보이는 것이 확인되며, 따라서 이를 단순한 포아송 분포로는 적절히 모델링할 수 없다. 이러한 과대산포 문제를 해결하기 위해 음이항분포 모델이 도입되었다.

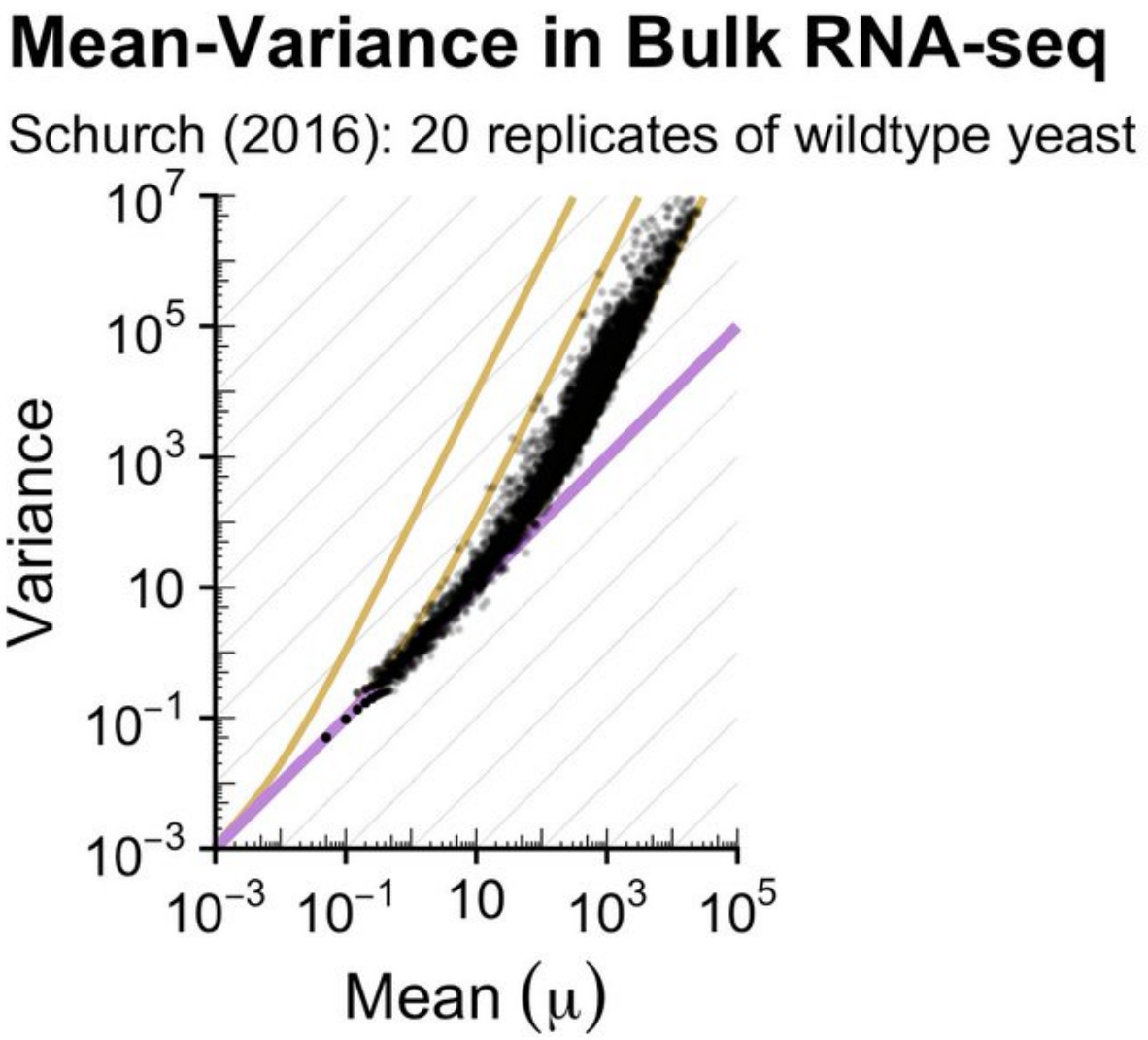


Figure 182: RNA-seq 평균-분산 관계

Figure 9.8 RNA-seq 평균-분산 관계 - Bulk RNA-seq 데이터에서 관찰되는 과대산포

음이항분포는 분산이 평균보다 큰 데이터를 효과적으로 모델링할 수 있으며, 다양한 모수 조합(n=20 p=0.25, n=20 p=0.5, n=20 p=0.75)에 따라 서로 다른 분포 형태를 보인다.

Figure 9.9 음이항분포 - 다양한 모수 조합에 따른 음이항분포의 형태

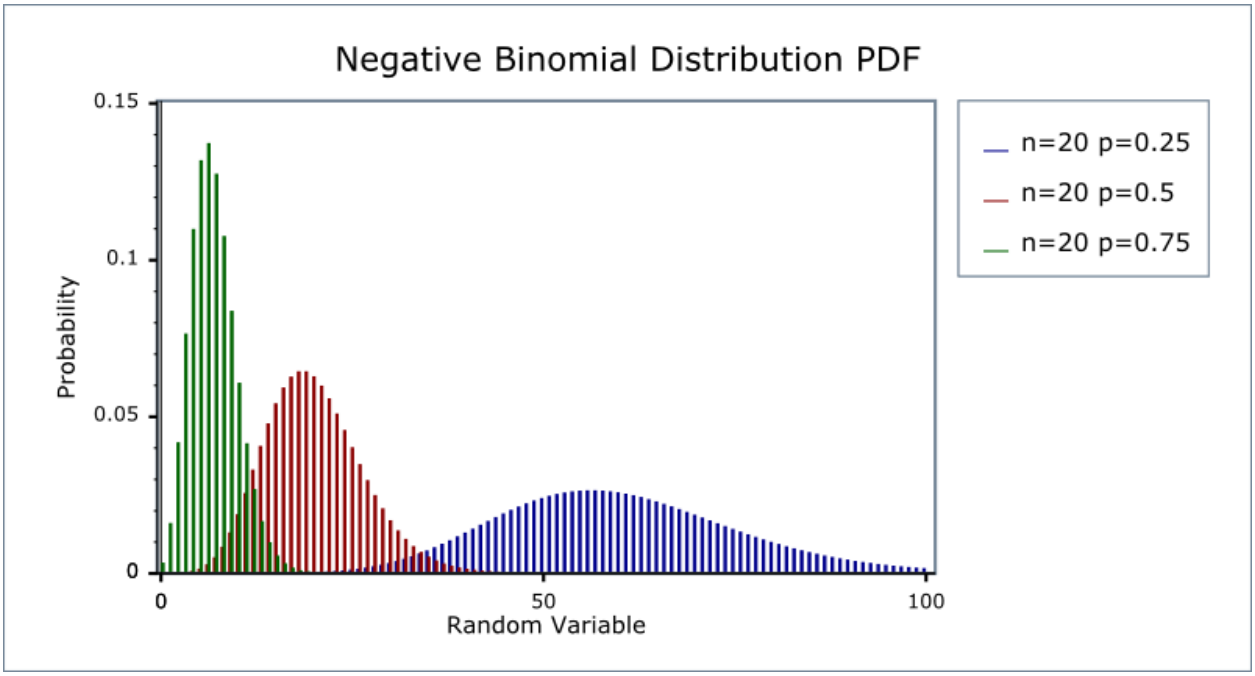


Figure 183: 음이항분포

실제 유전자 발현 데이터의 히스토그램들을 관찰하면, 모든 실험 조건과 복제에서 일관되게 음이항분포의 특징적인 형태를 보이는 것을 확인할 수 있다. 이는 음이항분포가 RNA-seq 데이터의 본질적 특성을 잘 반영하는 통계 모델임을 보여주며, 현재 DESeq2, edgeR 같은 주요 차등 발현 분석 도구들이 모두 음이항분포를 기반으로 설계되어 있는 이유이기도 하다.

Figure 9.10 RNA-seq 분포 – 실제 유전자 발현 데이터의 음이항분포 특성

10.4.4 데이터 변환과 정규화

RNA-seq 데이터는 카운트 데이터의 특성상 이분산성(heteroscedasticity)을 보인다. 이분산성이 있는 데이터의 분산을 평평하게 만들어주는 변환이 필요하며, 이를 위해 여러 가지 변환 방법이 사용된다.

분산 안정화 변환(Variance-stabilizing transform, VST)은 이분산성이 있는 데이터들의 분산을 평평하게 만들어주는 변환이다. sqrt(x), pearson residual, log(x+1) 등 여러 가지가 있다.

이 중에서 **로그 변환(log transform)**은 RNA-seq 데이터 분석에 일반적으로 가장 널리 활용되는 방법으로, 가장 기본적인 방법으로 log(x+1)을 취하는 것이다. 로그 변환의 목적은 분포를 정규분포(normal distribution)와 비슷하게 만들어주고, 평균에 따른 분산을 일정하게 만들어주는 것이다.

Figure 9.11 RNA-seq 로그 변환 분포 – 로그 변환 후 정규분포에 근접한 형태

10.5 배치 효과와 보정

10.5.1 배치 효과의 원인

배치 효과(batch effect)는 RNA-seq 데이터에서 생물학적 변이가 아닌 기술적 변이로 인해 발생하는 체계적 편향을 의미한다. 이는 실험 설계와 수행 과정에서 다양한 요인에 의해 발생할 수 있다. 주요 원인으로는 실험 날짜, 시퀀싱 장비, 시약 배치(lot), 실험자, 실험실 환경(온도, 습도) 같은 기술적 요인이 있다. 또한 샘플의 유전적 배경, 연령, 성별, 질병 상태, 조직 채취 조건 같은 생물학적 요인도 배치 효과를 유발할 수 있다. 이러한 배치 효과를 적절히 제거하지 않으면 실제 생물학적 신호가 가려져 잘못된 결론에 도달할 수 있다.

Figure 9.12 배치 보정 – 배치 효과 전후의 데이터 또는 평균 비교

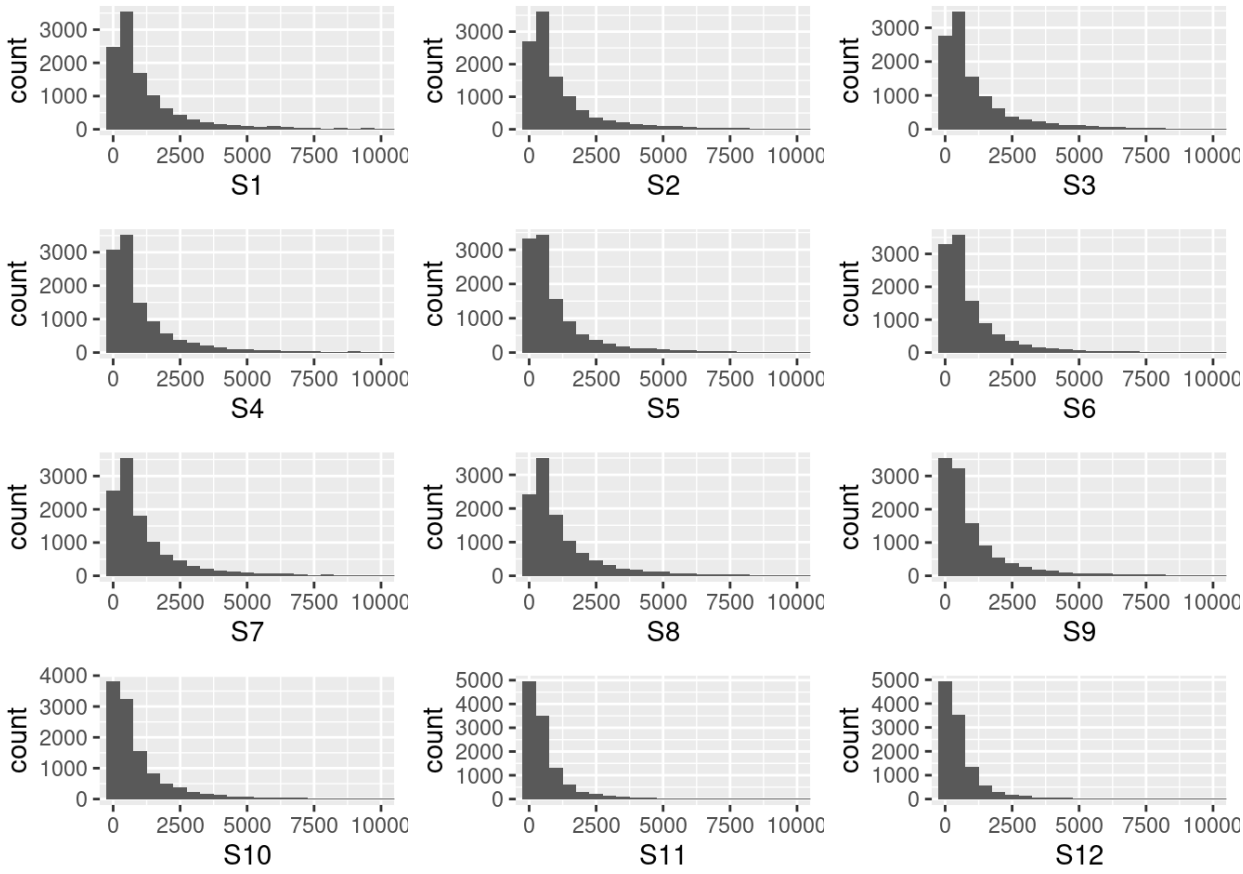


Figure 184: RNA-seq 분포

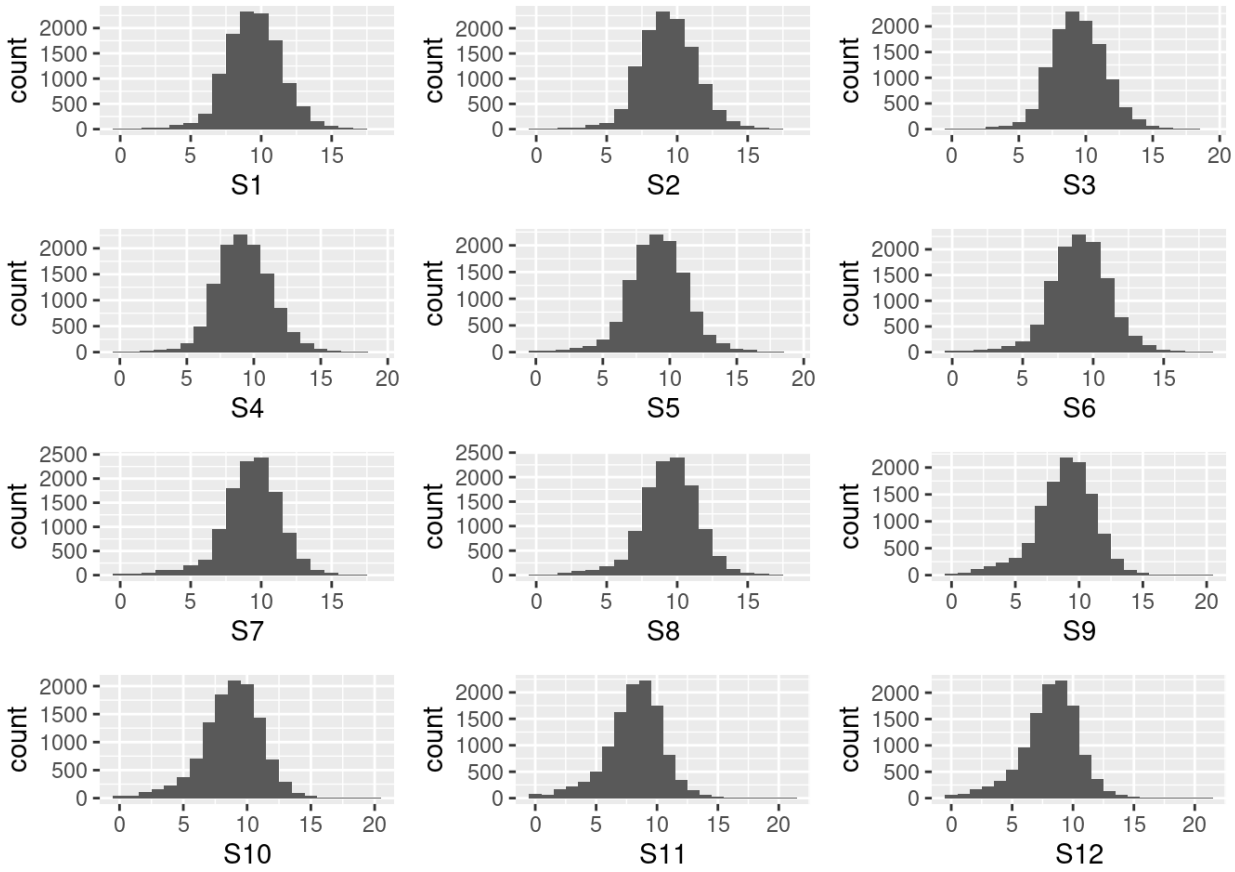


Figure 185: RNA-seq 로그 변환 분포

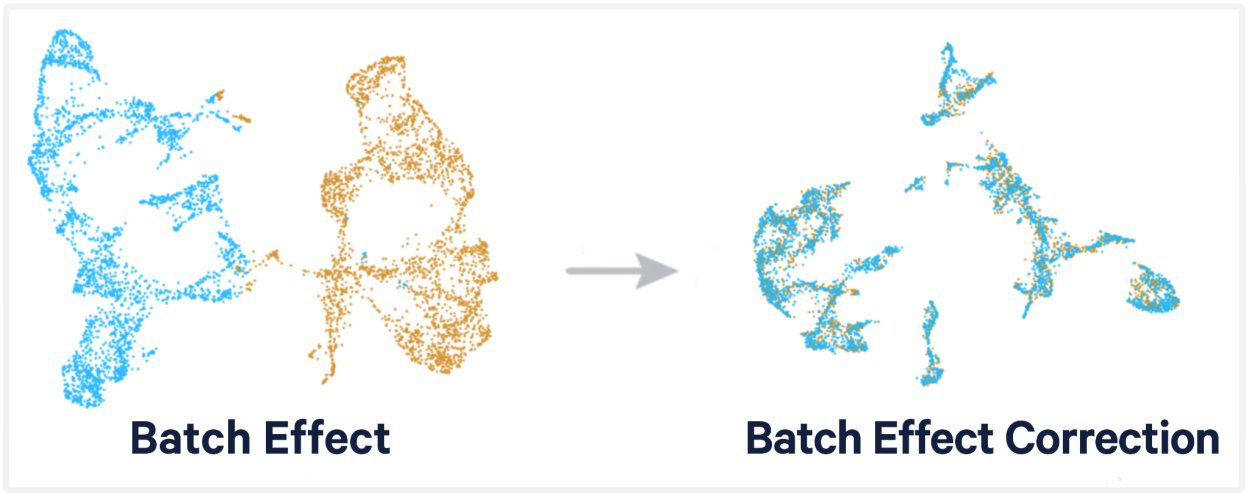


Figure 186: 배치 비교

10.5.2 정규화 기법

배치 효과를 보정하고 샘플 간 비교 가능성을 높이기 위해 다양한 정규화 기법이 사용된다.

분위수 정규화(Quantile normalization)는 모든 샘플이 동일한 분포를 갖도록 조정하는 방법이다. 각 샘플의 유전자 발현값을 순위별로 정렬한 후, 동일한 순위의 값들을 평균으로 대체하여 모든 샘플이 같은 분포를 갖도록 만든다.

상위 분위수 정규화(Upper quartile normalization)는 상위 75% 분위수를 기준으로 정규화하는 방법으로, 낮은 발현값에서 발생하는 노이즈의 영향을 줄이고 높은 발현값에 초점을 맞춘다.

중앙값 정규화(Median normalization)는 각 샘플의 중앙값을 기준으로 정규화하는 방법으로, 극값의 영향을 최소화하여 견고한(robust) 결과를 얻을 수 있다.

DESeq 정규화(DESeq normalization)는 기하평균(geometric mean)을 이용해 크기 인자(size factor)를 계산하고, 제로 카운트가 많은 RNA-seq 데이터의 특성을 고려하여 보정하는 방법이다. 각 유전자의 모든 샘플에 대한 기하평균을 계산한 후, 이를 기준으로 샘플별 크기 인자를 구하여 정규화를 수행한다.

10.6 차등 발현 분석

10.6.1 개념

차등 발현 분석(differential expression analysis)은 두 개 이상의 조건 사이에서 발현량이 유의하게 다른 유전자를 찾아내는 과정이다. 실무에서는 DEA(Differential Expression Analysis) 또는 DEG(Differentially Expressed Gene) 분석이라는 표현으로 자주 사용된다. 이는 RNA-seq 분석의 핵심 단계로, 질병과 정상 상태, 치료 전후, 시간 경과에 따른 변화 등을 비교하여 생물학적으로 의미 있는 유전자를 식별하는 것이다.

10.6.2 주요 분석 도구

DESeq2는 음이항 분포 모델(negative binomial model)과 수축 추정(shrinkage estimation)을 활용하여 작은 샘플 크기에서도 안정적으로 동작하는 도구이다. 베이지안 접근법을 사용하여 log2 fold change를 보정하고, 분산을 정확히 추정한다.

edgeR은 일반화 선형 모델(generalized linear model)과 정확 검정(exact test)을 제공하며, 복잡한 실험 설계도 처리할 수 있다. 음이항 분포를 기반으로 하며, 태그별 분산(tagwise dispersion)을 추정하여 각 유전자의 특성을 고려한다.

limma-voom은 전통적인 선형 모델을 RNA-seq 데이터에 적용할 수 있도록 분산을 조정하는 방법이다. voom 변환을 통해 평균-분산 관계를 모델링하고, limma의 경험적 베이즈 방법을 사용한다.

10.6.3 핵심 통계 지표

log2 fold change (log2FC)는 조건 간 발현 비율을 로그 값으로 표현한 것이다. $|\log2FC| \geq 1$이면 두 배 이상의 발현 차이를 의미하며, 양수는 증가, 음수는 감소를 나타낸다.

p-value는 관찰된 발현 차이가 우연히 발생할 확률을 나타낸다. 일반적으로 $p < 0.05$를 유의수준으로 사용한다.

adjusted p-value (FDR)는 다중 검정 보정을 통해 위양성률을 조절한 값이다. Benjamini-Hochberg 방법이 가장 널리 사용된다.

10.6.4 화산 플롯(Volcano plot)

화산 플롯은 차등 발현 분석 결과를 시각화하는 대표적인 방법이다. 이름은 그래프의 형태가 화산 모양과 유사하기 때문에 붙여졌다. X축에 log2FC, Y축에 `-log10(p-value)`를 두고 시각화하면 오른쪽 상단은 유의하게 증가한 유전자, 왼쪽 상단은 유의하게 감소한 유전자를 나타낸다. 일반적으로 `|log2FC|  1`, `p-value  0.05`를 컷오프로 사용하여 통계적으로 유의하고 생물학적으로 의미 있는 유전자를 식별한다.

Figure 9.13 화산 플롯 - 차등 발현 분석 결과의 시각화

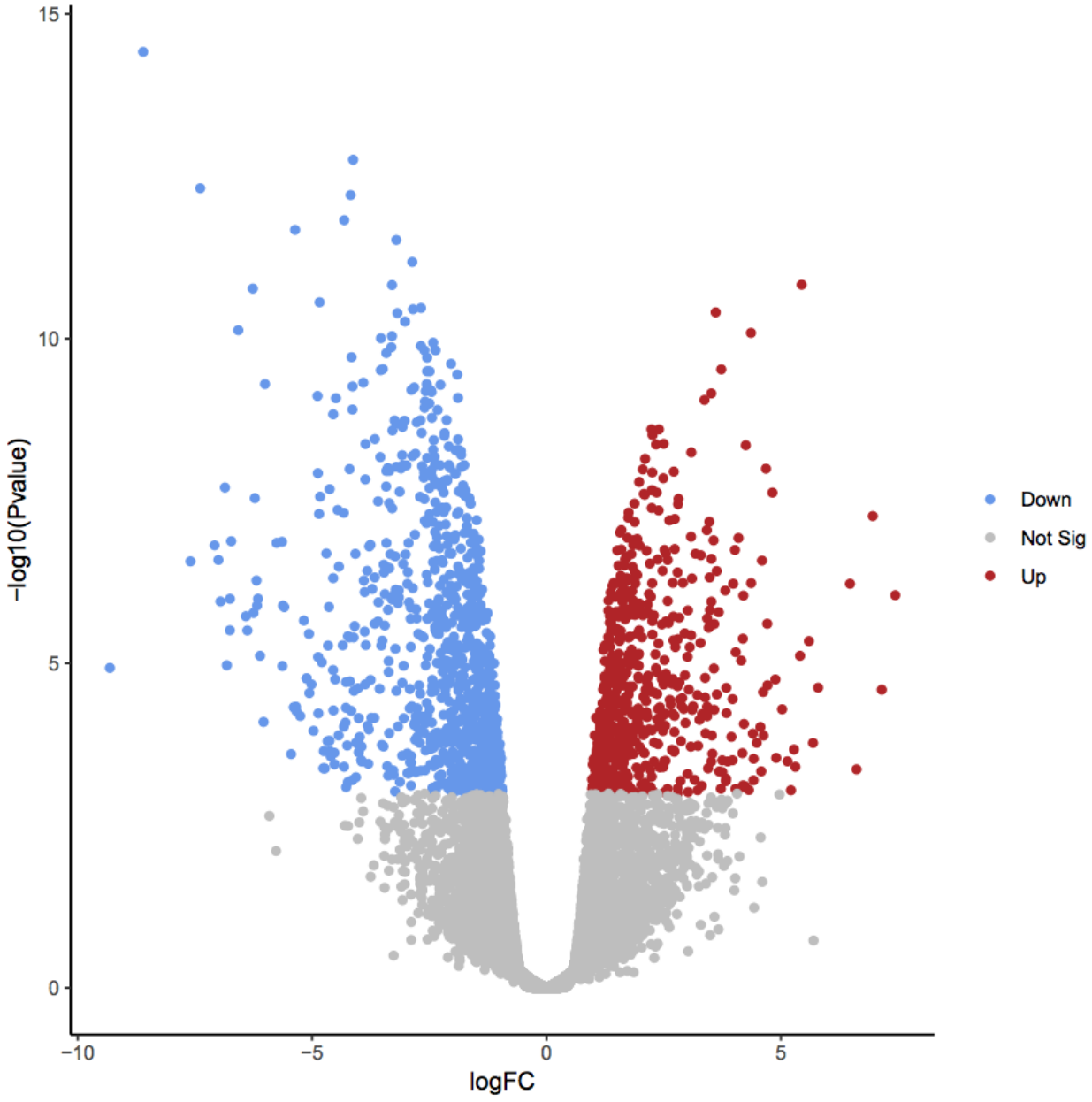


Figure 187: 화산 플롯

10.7 RNA-seq 데이터 분석 파이프라인

RNA-seq 데이터 분석은 원시 시퀀싱 데이터에서 생물학적으로 의미 있는 정보를 추출하기 위한 체계적인 과정이다. 이 파이프라인은 데이터의 특성과 연구 목적에 따라 단계별로 수행되며, 각 단계는 다음 단계의 분석 품질에 직접적인 영향을 미친다.

10.7.1 데이터 전처리

품질 평가와 필터링: 시퀀싱에서 생성된 원시 FASTQ 파일은 먼저 품질 평가를 거쳐야 한다. FastQC는 리드의 품질 점수 분포, GC 함량, 서열 중복도, 어댑터 오염 등을 종합적으로 평가한다. 품질이 낮은 리드나 어댑터 서열이 포함된 부분은 후속 분석의 정확도를 떨어뜨리므로 제거해야 한다. 이 과정에서 리드의 양 끝에서 품질이 낮은 염기를 잘라내고(trimming), 전체적으로 품질이 낮은 리드는 완전히 제거한다(filtering).

참조 유전체 정렬: 전처리된 리드들은 참조 유전체에 매핑되어야 한다. RNA-seq 데이터의 특수성은 인트론 스플라이싱으로 인해 리드가 여러 엑손에 걸쳐 있을 수 있다는 점이다. 따라서 STAR나 HISAT2 같은 스플라이스 인식 정렬기(splice-aware aligner)를 사용한다. 이들 도구는 알려진 스플라이스 사이트 정보를 활용하고, 새로운 스플라이스 접합부도 발견할 수 있다. 정렬 결과는 SAM/BAM 형식으로 저장되며, 각 리드의 유전체 상 위치와 매핑 품질 정보를 포함한다.

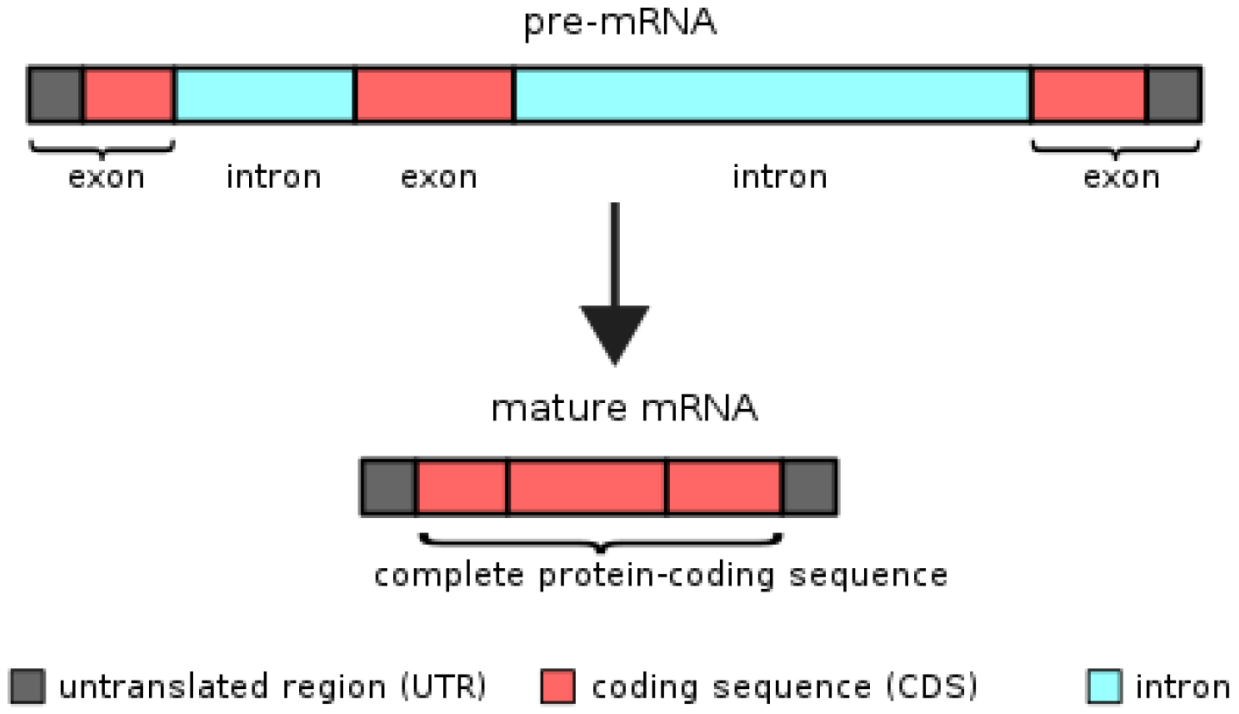


Figure 188: RNA 스플라이싱

Figure 9.14 RNA 스플라이싱 – 엑손 연결과 인트론 제거 과정

발현량 정량화: 정렬된 리드를 바탕으로 각 유전자의 발현량을 계산한다. HTSeq나 featureCounts는 유전자 수준에서 리드 카운트를 계산하는 도구로, 각 유전자의 엑손 영역에 매핑된 리드 수를 집계한다. 반면 kallisto나 salmon은 전사체 수준에서 발현량을 추정하는 도구로, 각 전사체 이소폼별로 발현량을 정량화할 수 있다. 이들은 k-mer 기반 알고리즘을 사용하여 빠른 속도로 정량화를 수행한다.

10.7.2 다운스트림 데이터 분석

주성분 분석(PCA): PCA는 고차원 데이터의 주요 변이를 저차원으로 투영하여 샘플 간 유사성을 시각화한다. 이를 통해 실험 조건에 따른 샘플 그룹화, 배치 효과의 존재, 이상값(outlier) 샘플을 확인할 수 있다. 생물학적 복제들이 서로 가깝게

클러스터되고, 실험 조건이 다른 그룹들이 명확히 분리되는 것이 이상적이다.

클러스터링 분석: 클러스터링 분석은 샘플 간의 유사성을 바탕으로 그룹을 형성하여 데이터 구조를 파악한다. 샘플 간 거리 행렬을 계산하고 이를 시각적으로 표현하여 샘플들의 관계를 이해할 수 있다. 이는 PCA와 상호 보완적인 정보를 제공하며, 샘플 간의 상대적 유사성을 정량적으로 평가할 수 있다. 유클리드 거리나 상관계수 기반 거리를 사용하여 샘플 간 유사도를 계산한다.

품질 검증: 생물학적 복제 간 상관 분석을 통해 실험의 재현성을 평가한다. 높은 상관계수($r > 0.9$)는 실험이 일관되게 수행되었음을 의미하며, 낮은 상관계수는 기술적 문제나 샘플 품질 이슈를 시사한다.

10.7.3 생물학적 해석

차등 발현 분석을 통해 개별 유전자의 변화를 확인한 후, 기능적으로 연관된 유전자 집합 수준에서의 해석이 필요하다. 이를 위해 세 가지 주요 접근법이 사용된다.

과대표 분석(Over-Representation Analysis, ORA): 5장에서 소개한 ORA는 차등 발현 유전자 목록에서 특정 생물학적 경로나 기능 범주의 유전자들이 통계적으로 유의하게 많이 포함되어 있는지 검정한다. Fisher의 정확 검정이나 하이퍼기하분포를 사용하여 우연히 발생할 확률을 계산한다.

유전자 집합 풍부도 분석(Gene Set Enrichment Analysis, GSEA): GSEA는 미리 정의된 임계값 없이 전체 유전자 목록을 순위화하여 분석한다. 특정 생물학적 경로나 기능 범주에 속하는 유전자들이 전체적으로 상향 또는 하향 조절되는지 통계적으로 검정한다. 이는 개별 유전자 수준에서는 유의하지 않지만 경로 수준에서는 의미 있는 변화를 포착할 수 있어, 생물학적 메커니즘을 이해하는 데 중요하다.

네트워크 기반 분석: 유전자 간의 상호작용 정보를 활용하여 생물학적 네트워크 상에서 변화를 해석한다. IPA(Ingenuity Pathway Analysis) 같은 도구는 차등 발현 유전자들을 알려진 상호작용 네트워크에 매핑하여 핵심 조절 분자와 활성화된 경로를 예측한다.

10.8 결론

RNA 시퀀싱은 유전자 발현을 전사체 수준에서 정량적으로 분석할 수 있는 강력한 기술이다. 벌크 RNA-seq을 통해 조직이나 세포 집단의 평균적인 발현 패턴을 파악할 수 있으며, 차등 발현 분석을 통해 다양한 조건 간의 유전자 발현 변화를 체계적으로 규명할 수 있다.

RNA-seq 데이터 분석에서는 음이항분포 기반의 통계 모델을 사용하여 과대산포 현상을 적절히 처리하고, 다양한 정규화 기법을 통해 샘플 간 비교 가능성을 확보한다. 특히 DESeq2, edgeR, limma-voom 같은 전용 도구들이 개발되어 신뢰성 있는 차등 발현 분석을 수행할 수 있다.

분석 파이프라인은 원시 데이터 전처리부터 생물학적 해석까지 체계적으로 구성되며, 각 단계에서 적절한 품질 관리와 검증이 필요하다. 최종적으로는 ORA, GSEA, 네트워크 기반 분석을 통해 개별 유전자 수준을 넘어선 생물학적 의미를 도출할 수 있다.

RNA-seq 기술은 질병 바이오마커 발굴, 약물 반응성 예측, 정밀의학 구현 등 임상 분야에서도 광범위하게 활용되고 있다. 앞으로는 장거리 시퀀싱, 실시간 분석, 다중 오믹스 통합 등의 기술 발전과 함께 딥러닝 기반 자동 해석이 RNA-seq의 가치를 더욱 높일 것으로 전망된다.

11 단일 세포 전사체학

11.1 단일세포 RNA 시퀀싱의 혁신

전통적인 벌크 RNA 시퀀싱(bulk RNA-seq)은 수백만 개의 세포에서 추출한 RNA를 혼합하여 분석하므로, 전체 세포 집단의 평균적인 발현 패턴만을 제시한다. 이러한 접근법은 조직 전체의 일반적인 특성을 파악하는 데는 유용하지만, 개별 세포 간의 차이나 희귀 세포 유형의 존재를 감지하기 어렵다는 한계가 있다.

단일세포 RNA 시퀀싱(single-cell RNA sequencing, scRNA-seq)은 이러한 한계를 극복하고 세포 이질성(cellular heterogeneity)을 직접 측정할 수 있는 혁신적인 기술이다. 이 기술을 통해 연구자들은 겉보기에 동일해 보이는 세포 집단 내에서도 유전자 발현 패턴이 현저히 다른 세포 아집단들을 발견할 수 있었다. 예를 들어, 종양 조직 내에서 암세포의

다양한 아형들이 서로 다른 전사체 프로파일을 보이며, 이러한 차이가 치료 반응성과 전이 능력에 영향을 미칠 수 있다는 것이 밝혀졌다.

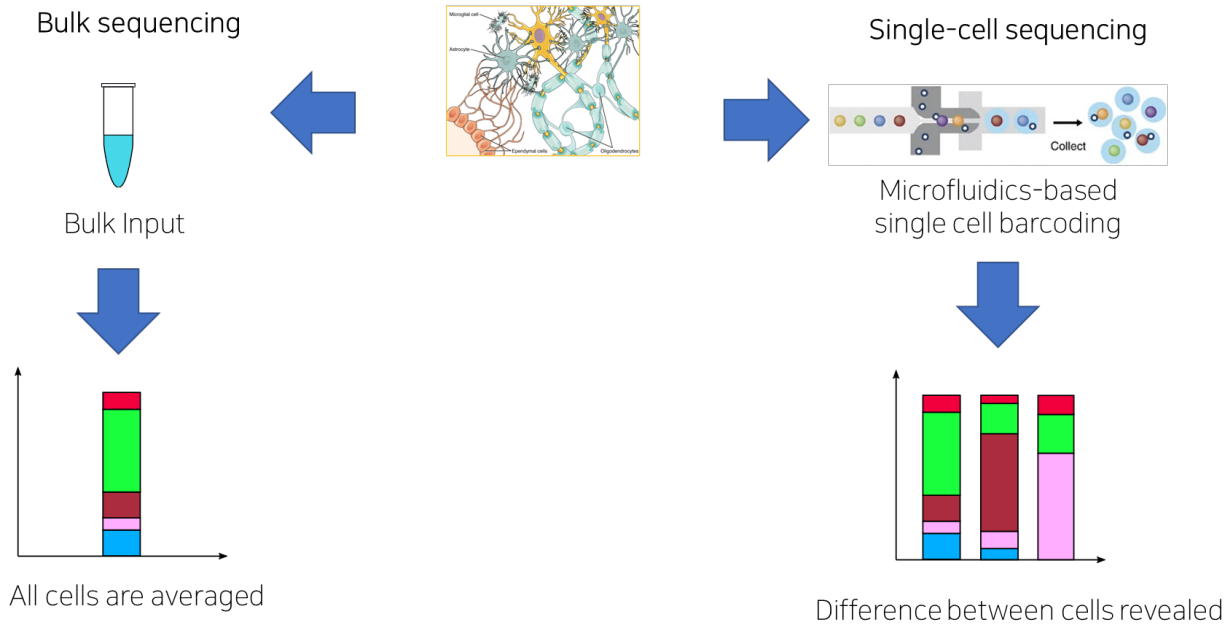


Figure 189: 벌크와 단일세포 RNA-seq 비교

Figure 10.1 벌크와 단일세포 RNA-seq 비교 – 벌크 분석의 평균화된 신호와 단일세포 분석의 개별 세포 정보

11.1.1 세포 분화와 발생학적 이해

scRNA-seq 기술은 세포 분화 과정을 시간적 순서에 따라 재구성할 수 있는 강력한 도구를 제공한다. 의사시간(pseudotime) 분석을 통해 발생 과정에서 줄기세포가 특화된 세포로 분화하는 경로를 추적할 수 있다. 이는 실제 시간 경과를 따라 실험을 수행하지 않고도, 서로 다른 분화 단계에 있는 세포들의 전사체 유사성을 바탕으로 발생학적 궤적을 재구성하는 계산적 방법이다.

Figure 10.2 단일세포 RNA-seq 응용 – 세포 분화, 발생학적 궤적, 세포 유형 발견의 다양한 응용 분야

RNA velocity 분석은 이러한 분석을 한 단계 더 발전시킨 기법으로, 각 세포에서 미성숙 mRNA(unspliced)와 성숙 mRNA(spliced)의 비율을 분석하여 유전자 발현의 동적 변화 방향을 예측한다. 이를 통해 세포가 어떤 방향으로 분화하고 있는지, 어떤 분기점에서 세포 운명이 결정되는지를 실시간으로 파악할 수 있다.

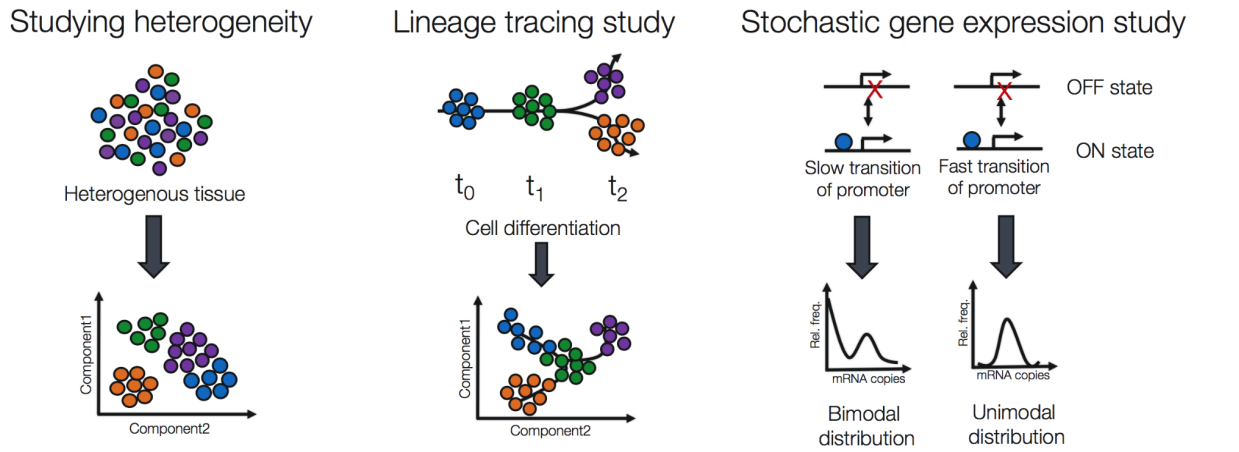
11.1.2 유전자 발현의 확률적 특성

단일세포 수준에서 관찰되는 유전자 발현은 벌크 분석에서 예상되는 것과는 매우 다른 특성을 보인다. 개별 세포에서 유전자 발현은 연속적이고 균일한 과정이 아니라, 간헐적이고 폭발적인 버스트(burst) 형태로 일어난다. 이는 유전자의 프로모터 영역이 활성 상태와 비활성 상태를 번갈아 가며, 활성 상태일 때 다량의 mRNA가 한꺼번에 생산되기 때문이다.

이러한 확률적 발현 패턴은 단일세포 데이터에서 이중모드(bimodal) 또는 다중모드 분포로 나타난다. 일부 세포에서는 특정 유전자가 높은 수준으로 발현되는 반면, 다른 세포에서는 거의 발현되지 않는 ON/OFF 스위치와 같은 패턴을 보인다. 이러한 발현의 불균질성은 세포가 환경 변화에 대응하거나 분화 신호에 반응할 때 표현형 다양성을 생성하는 중요한 메커니즘으로 작용한다.

11.1.3 단일세포 데이터의 클러스터링

단일세포 데이터에서 세포 유형을 분류하기 위해서는 고차원 데이터에 적합한 클러스터링 알고리즘이 필요하다. 주요 클러스터링 방법과 차원 축소 기법의 상세한 내용은 11장 “차원 축소와 데이터 분석”에서 다룬다.



Liu S and Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges, F1000 Research 2016 (doi: 10.12688/f1000research.7223.1)
Junker and van Oudenaarden; Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology, Cell 2014 (doi: 10.1016/j.cell.2014.02.010)

Figure 190: 단일세포 RNA-seq 응용

11.2 플로우 사이토메트리

11.2.1 기술의 물리적 원리

플로우 사이토메트리(flow cytometry)는 유체역학적 집속(hydrodynamic focusing) 원리를 이용하여 세포를 한 개씩 분리하고 분석하는 기술이다. 이 시스템에서 세포가 포함된 시료는 중앙의 세포 스트림을 형성하고, 양쪽에서 순환하는 전해질 용액(sheath fluid)에 의해 둘러싸여 층류(laminar flow) 상태를 만든다. 이러한 층류 조건에서 세포들은 단일 파일로 정렬되어 레이저 빔을 통과하게 된다.

세포가 레이저와 만나는 순간, 빛의 산란과 형광 신호가 발생한다. 전방 산란(Forward Scatter, FSC)은 세포의 상대적 크기를 반영하며, 측방 산란(Side Scatter, SSC)은 세포 내부의 복잡도나 과립의 정도를 나타낸다. 이 두 매개변수만으로도 혈액 세포의 주요 유형들(림프구, 단핵구, 과립구)을 구별할 수 있다.

Figure 10.3 플로우 사이토메터 – 실제 플로우 사이토메터 장비와 레이저 빔 경로

11.2.2 형광 활성화 세포 분류(FACS)

형광 활성화 세포 분류(Fluorescence-Activated Cell Sorting, FACS)는 플로우 사이토메트리 기술을 확장하여 관심 있는 세포를 물리적으로 분리할 수 있는 시스템이다. 이 과정에서 세포가 포함된 스트림은 진동에 의해 균일한 크기의 미세한 액적(droplet)으로 분할된다. 각 액적에는 통계적으로 0개 또는 1개의 세포가 포함되며, 세포의 형광 특성에 따라 특정 전하로 대전된다.

대전된 액적들은 전기장을 통과하면서 편향되어 서로 다른 수집 용기로 분류된다. 이를 통해 특정 마커를 발현하는 세포만을 선택적으로 회수할 수 있으며, 분류된 세포들은 후속 실험에 바로 사용할 수 있을 정도로 높은 순도와 생존율을 유지한다. FACS는 특히 희귀 세포 집단의 농축이나 줄기세포 분리에 널리 활용된다.

Figure 10.4 FACS 분류 – 형광 활성화 세포 분류를 통한 세포 집단 분리 과정

11.2.3 다중 매개변수 분석

현대의 플로우 사이토메터는 10개 이상의 서로 다른 형광 채널을 동시에 측정할 수 있어, 복잡한 세포 집단의 다차원적 분석이 가능하다. 그러나 여러 형광 염료들의 방출 스펙트럼이 부분적으로 겹치는 경우가 많아, 이를 보정하기 위한 스펙트럼 보상(spectral compensation) 과정이 필요하다.

보상 과정에서는 각 형광 염료로 단독 염색된 대조군 시료들을 이용해 보상 매트릭스를 계산한다. 이 매트릭스는 각 검출기에서 측정된 신호 중 다른 형광 염료로부터 누출된 신호의 비율을 정량화하여, 실제 신호에서 이를 수학적으로 제거한다.

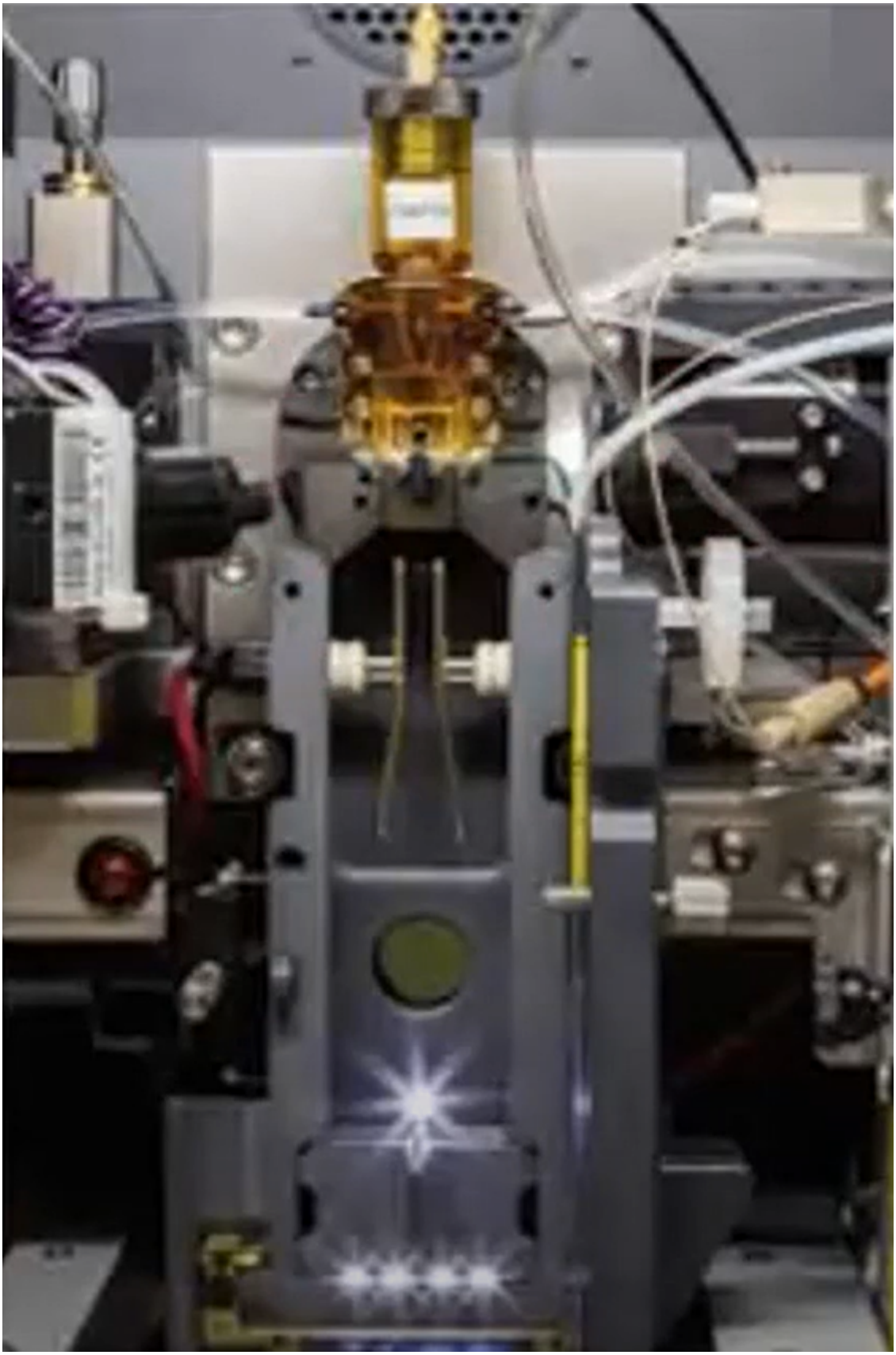


Figure 191: 플로우 사이토메터

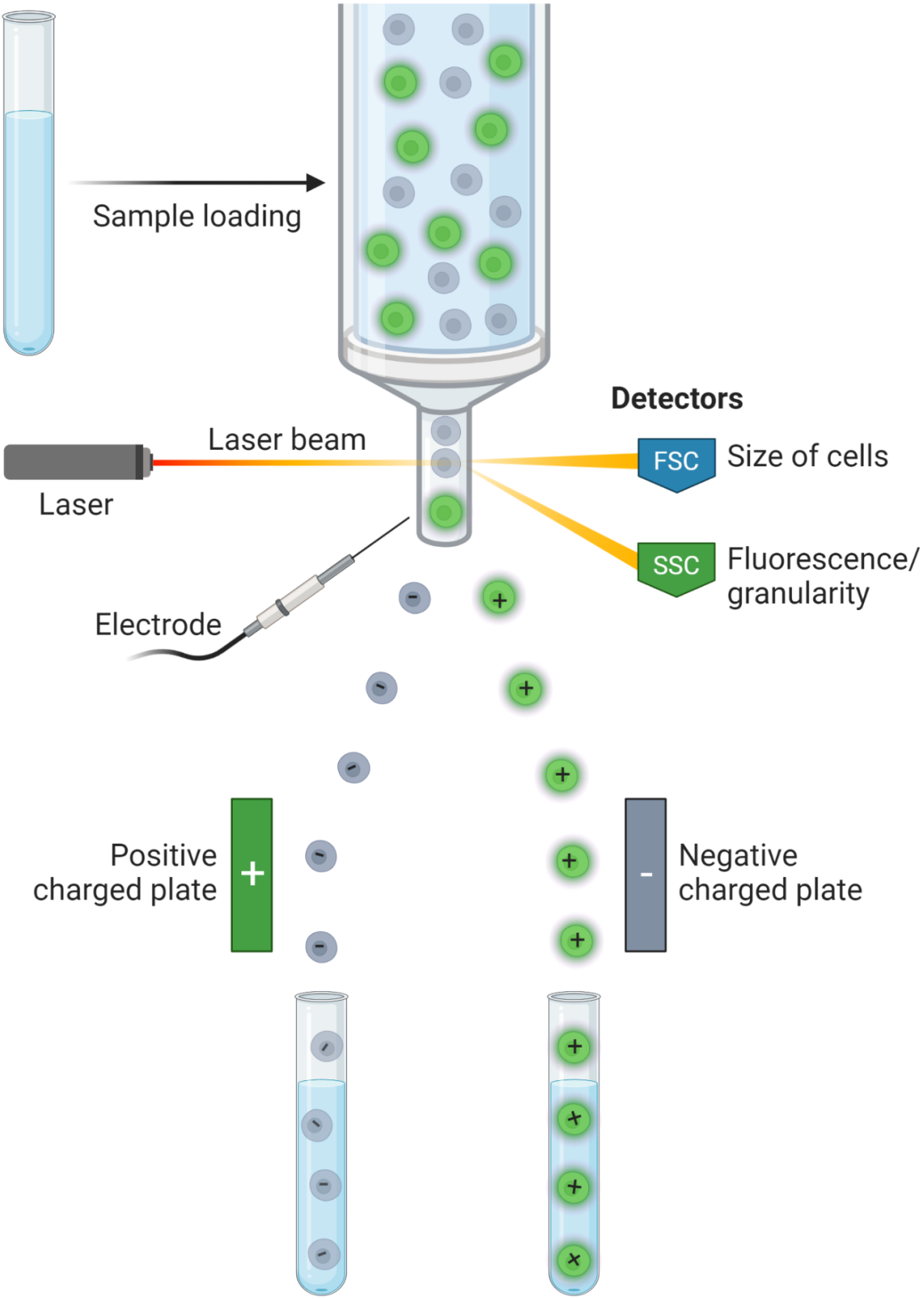


Figure 192: FACS 분류

데이터 분석에서는 단계적 게이팅(sequential gating) 전략을 사용한다. 먼저 FSC/SSC 분포를 이용해 세포 파편과 응집체(doublet)를 제거하고, 생존/사멸 염료를 이용해 생존 세포만을 선별한다. 이후 특정 표면 마커나 세포내 단백질의 발현 패턴에 따라 순차적으로 세포 아집단을 정의한다. 이러한 체계적 접근을 통해 복잡한 조직에서도 정확한 세포 유형 분류와 정량 분석이 가능하다.

11.3 미세유체학과 단일세포 포획 기술

11.3.1 미세유체학의 물리적 특성

미세유체학(microfluidics)은 마이크로미터 규모의 채널에서 유체의 흐름을 제어하는 기술이다. 플로우 사이토메트리가 1960년대부터 발전해온 것과 달리, 미세유체학은 2000년대 후반부터 단일세포 분석 분야에 본격적으로 도입되었다.

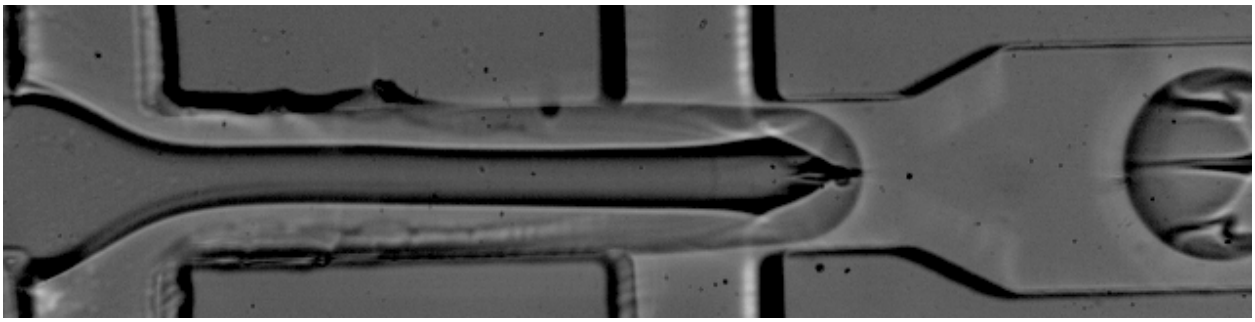


Figure 193: 액적 단일세포 포획

Figure 10.5 액적 단일세포 포획 – 미세유체학적 액적 생성을 통한 단일세포 포획 과정

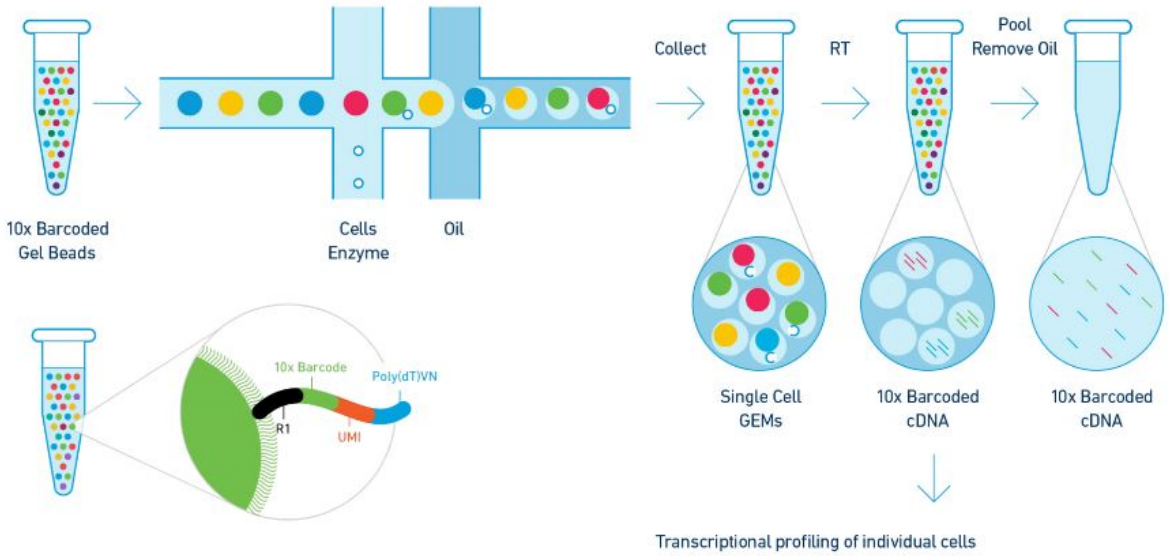


Figure 194: 10x 시퀀싱

Figure 10.6 10x 시퀀싱 – 10x Genomics 플랫폼의 단일세포 분석 워크플로우

마이크로미터 규모에서는 유체역학적 특성이 거시적 환경과 크게 달라지는데, 예를 들어 점성력이 상대적으로 더 커지게 되어 층류(laminar flow)가 지배적이게 된다. 이러한 특성 때문에 서로 다른 유체 흐름이 만나더라도 난류 혼합 없이 안정적인 계면을 형성하게 된다.

이러한 물리적 특성을 활용하여 T자형 또는 flow-focusing 구조의 채널에서 균일한 크기의 액적(droplet)을 생성할 수 있다. 이러한 기술의 발전은 이를 활용한 단일세포 전사체 분석 방법들이 등장하는 계기가 되었다.

11.3.2 Drop-seq 기술의 원리

Drop-seq는 미세유체학적 액적 생성 기술을 단일세포 전사체 분석에 적용한 방법이다. 이 시스템의 핵심은 각 액적에 세포와 바코드된 비드(barcoded bead)를 함께 포함시키는 것이다. 비드 표면에는 poly-T 서열, 고유한 세포 바코드(Cell Barcode, BC), 그리고 분자 식별자(Unique Molecular Identifier, UMI)가 순서대로 연결된 올리고뉴클레오타이드가 부착되어 있다.

세포와 비드의 액적 내 분포는 포아송 분포를 따르므로, 희석된 농도에서는 대부분의 액적에 0개 또는 1개의 세포가 포함된다. 세포 바코드는 일반적으로 12-16개 염기로 구성되어 수만 개의 서로 다른 세포를 구별할 수 있으며, UMI는 증폭 과정에서 발생할 수 있는 편향을 보정하기 위해 각 mRNA 분자를 고유하게 표지한다.

액적이 형성된 후, 계면활성제를 이용해 세포막을 용해시키고 mRNA를 방출시킨다. 비드 표면의 poly-T 서열이 mRNA의 poly-A tail에 결합하여 역전사 반응이 일어나며, 이 과정에서 각 cDNA 분자에는 해당 세포의 고유한 바코드와 UMI가 함께 부착된다.

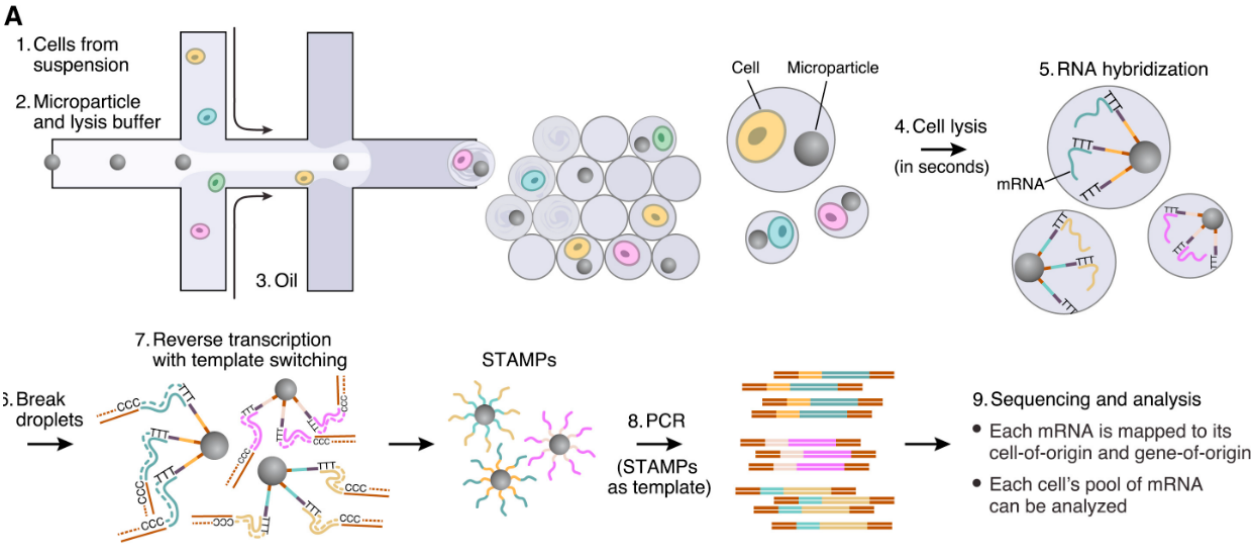


Figure 195: Drop-seq 기술

Figure 10.7 Drop-seq 기술 – 바코드된 비드와 단일세포를 이용한 전사체 분석 과정

11.3.3 라이브러리 제작과 상용화 플랫폼

역전사 반응이 완료된 후에는 계면활성제를 이용해 액적을 파괴하고 바코드된 cDNA를 회수한다. 회수된 cDNA는 PCR 증폭을 통해 시퀀싱에 충분한 양으로 늘리며, 이 과정에서 샘플 식별을 위한 추가 바코드와 Illumina 시퀀싱에 필요한 어댑터 서열이 부착된다.

현재 여러 상용화된 플랫폼이 이러한 원리를 구현하고 있다. 10x Genomics의 Chromium 시리즈는 가장 널리 사용되는 플랫폼으로, 한 번의 실험에서 수천에서 수만 개의 세포를 분석할 수 있다. BD Biosciences의 Rhapsody 시스템은 다중 시료 처리에 최적화되어 있으며, Bio-Rad의 ddSEQ 시스템은 비용 효율적인 대안을 제공한다.

Figure 10.8 단일세포 라이브러리 구조 – 10x Genomics 라이브러리의 바코드와 UMI 구조

상용 플랫폼들은 표준화된 프로토콜과 자동화된 처리 과정을 제공하여 실험 간 재현성을 향상시켰다. 세포 포획 효율, 유전자 검출 감도, 처리량 등이 개선되면서 단일세포 전사체학의 활용 범위가 확대되고 있다.

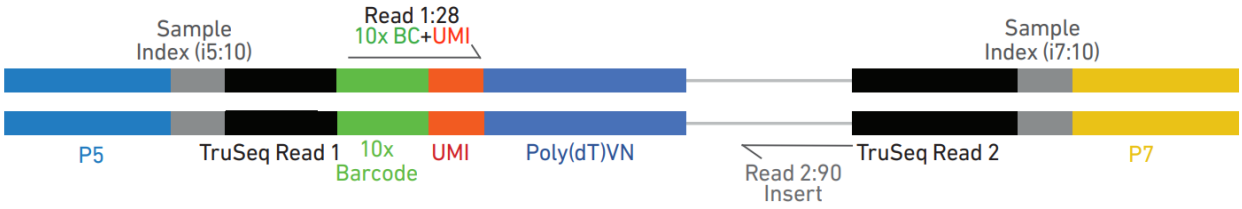


Figure 196: 단일세포 라이브러리 구조

11.4 전사체 데이터 분석 방법론

11.4.1 통계적 검정과 다중 비교 보정

전사체 데이터, 특히 단일세포 RNA-seq 데이터는 일반적인 생물학적 데이터와는 매우 다른 통계적 특성을 보인다. 가장 두드러진 특징은 "영점 과다(zero inflation)" 현상으로, 많은 유전자가 특정 세포에서 전혀 발현되지 않아 측정값이 0이 되는 경우가 빈번하다. 이는 기술적 한계(낮은 검출 감도, 증폭 편향)와 생물학적 특성(확률적 발현, 세포 특이적 발현) 모두에 기인한다.

이러한 특성 때문에 전통적인 t-검정이나 분산분석과 같은 모수적 검정보다는 비모수적 검정이 선호된다. Wilcoxon rank-sum 검정(또는 Mann-Whitney U test 라고도 불림)은 분포에 대한 가정 없이 두 그룹 간의 차이를 검정할 수 있어 널리 사용된다.

전사체 분석에서는 수천에서 수만 개의 유전자를 동시에 검정하므로 다중 비교 문제가 심각하다. 이를 해결하기 위해 거짓 발견률(False Discovery Rate, FDR)을 제어하는 Benjamini-Hochberg 방법이나 더 보수적인 본페로니 보정이 적용된다.

11.4.2 차원 축소 기법

단일세포 전사체 데이터의 시각화와 분석을 위해서는 차원 축소가 필수적이다. PCA, t-SNE, UMAP 등의 주요 차원 축소 기법과 그 특성에 대한 상세한 설명은 11장 "차원 축소와 데이터 분석"에서 다룬다.

11.4.3 클러스터링과 품질 평가

차원 축소된 데이터에서 세포들을 유사한 그룹으로 분류하는 클러스터링은 세포 유형 발견의 핵심 과정이다. 다양한 클러스터링 알고리즘과 품질 평가 방법에 대한 상세한 내용은 11장 "차원 축소와 데이터 분석"에서 다룬다.

11.5 단일세포 데이터 분석 파이프라인의 개요와 중요성

단일세포 RNA 시퀀싱(single-cell RNA sequencing, scRNA-seq) 기술의 발전은 의생명정보학 분야에 혁신적인 변화를 가져왔다. 전통적인 벌크 RNA 시퀀싱이 수백만 개 세포의 평균적인 발현 정보를 제공하는 것과 달리, 단일세포 분석은 개별 세포의 유전자 발현 패턴을 측정할 수 있게 해준다. 이러한 기술적 혁신을 통해 연구자들은 조직 내 세포 이질성을 직접 관찰하고, 희귀 세포 유형을 발견하며, 세포 분화 과정을 상세히 추적할 수 있게 되었다.

Figure 10.9 단일세포 RNA-seq 개요 – 단일세포 분석의 전체 워크플로우와 주요 단계들

단일세포 데이터 분석 파이프라인은 원시 시퀀싱 데이터에서 생물학적으로 의미 있는 정보를 추출하기 위한 체계적인 절차를 의미한다. 이 파이프라인은 데이터의 특수한 특성을 고려하여 설계되어야 한다. 단일세포 데이터는 일반적으로 다음과 같은 특징을 가진다:

희소성(Sparsity): 단일세포 RNA-seq 데이터에서는 전체 측정값의 60-90%가 0 또는 매우 낮은 값을 나타낸다. 이는 기술적 한계(검출 감도, 포획 효율)와 생물학적 특성(확률적 유전자 발현) 모두에 기인한다.

고차원성(High-dimensionality): 수만 개의 유전자를 수천에서 수십만 개의 개별 세포에서 측정하므로, 차원의 저주 문제가 심각하게 나타날 수 있다.

노이즈(Noise): 세포 포획, 역전사, PCR 증폭 등 여러 실험 단계에서 발생하는 기술적 변이와 세포 주기, 환경 반응 등으로 인한 생물학적 변이가 복합적으로 작용한다.

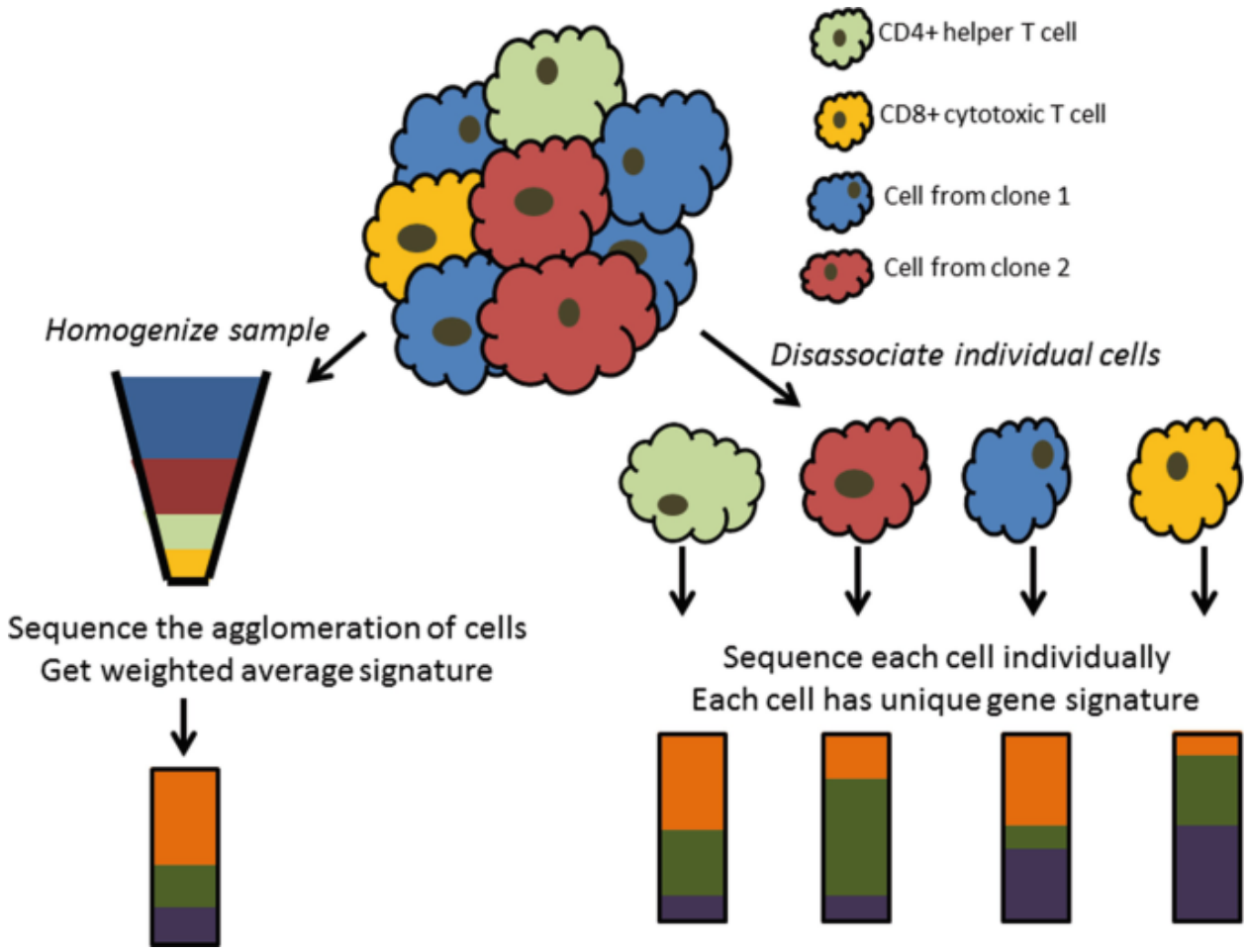


Figure 197: 단일세포 RNA-seq 개요

이러한 특성으로 인해 단일세포 데이터 분석은 전통적인 벌크 분석과는 완전히 다른 접근법을 요구하며, 전용 분석 도구와 방법론의 개발이 필요하다.

11.6 단일세포 데이터 분석의 핵심 단계

단일세포 RNA-seq 데이터 분석은 일련의 체계적인 과정을 통해 수행된다. 각 단계는 데이터의 품질과 최종 분석 결과에 직접적인 영향을 미치므로, 각 과정의 원리와 목적을 정확히 이해하는 것이 중요하다.

11.6.1 데이터 정량화와 품질 관리

정량화 과정의 이해: 단일세포 분석의 첫 번째 단계는 각 세포에서 각 유전자의 발현량을 정확히 측정하는 것이다. 이 과정에서 생성되는 것이 세포×유전자 행렬(cell × gene matrix)로, 이는 모든 후속 분석의 기반이 된다. 벌크 RNA-seq과 달리 단일세포 분석에서는 각 개별 세포가 하나의 샘플이 되므로, 수천에서 수만 개의 개별 샘플을 동시에 처리하는 것과 같다.

정량화 지표로는 다음과 같은 두 가지 방법이 주로 사용된다. - **Raw count**: 각 유전자에 대해 매핑된 리드의 수를 직접 계산한 값 - **UMI count**: Unique Molecular Identifier를 이용하여 PCR 증폭 편향을 보정한 값

품질 관리 지표: 단일세포 데이터의 품질을 평가하기 위해 여러 지표를 사용한다. 각 세포에서 검출된 총 유전자 수는 세포의 건강 상태를 나타내는 중요한 지표이다. 너무 적은 유전자가 검출되면 세포가 손상되었거나 죽은 상태일 가능성이 높다. 미토콘드리아 유전자의 발현 비율도 중요한 품질 지표로, 세포막이 손상되면 세포질 mRNA는 손실되지만 미토콘드리아 mRNA는 상대적으로 보존되어 이 비율이 높아진다.

10x Genomics 플랫폼의 기술적 특징: 현재 가장 널리 사용되는 단일세포 RNA 시퀀싱 플랫폼 중 하나인 10x Genomics는 독특한 바코딩 시스템을 사용한다. 각 세포는 16nt의 Cell Barcode로 구분되며, 각 RNA 분자에는 12nt의 UMI(Unique Molecular Identifier)가 부착된다. Poly(dT) 프라이머를 사용하여 mRNA만을 선택적으로 포획한다. 실험 과정에서 세포를 마이크로드롭에 가두고 용해한 뒤 역전사를 진행하면서 바코드와 UMI를 부착하고, 이어서 cDNA를 증폭해 라이브러리를 제작한다.

UMI의 중요성: UMI는 같은 RNA 분자에서 나온 모든 리드를 하나로 묶어 PCR 편향을 제거하는 핵심 역할을 한다. 이를 통해 리드 수 대신 실제 분자 수를 정확하게 계산할 수 있어, 정량화의 정확도가 크게 향상된다.

Figure 10.10 단일세포 RNA-seq 정량화 - 10x Genomics 플랫폼에서의 세포 바코드와 UMI 기반 정량화 과정

Figure 10.11 PCR 편향 - PCR 증폭 과정에서 발생하는 편향과 UMI의 역할

Figure 10.12 UMI 원리 - Unique Molecular Identifier를 이용한 분자 수 정확한 정량화

드롭아웃 현상의 이해: 단일세포 데이터의 가장 특징적인 현상 중 하나는 드롭아웃(dropout)으로, 전체 값의 95% 가까이가 0으로 나타나기도 한다. 이는 낮은 포획 효율, 역전사 실패, 시퀀싱 깊이 부족 같은 기술적 요인과 실제 발현 부재, 세포 주기, 세포 특이적 발현 같은 생물학적 요인이 복합적으로 작용한 결과이다.

Figure 10.13 단일세포 FASTQ 파일 미리보기 Read1 - 단일세포 RNA-seq Read1의 FASTQ 파일 구조와 바코드 정보

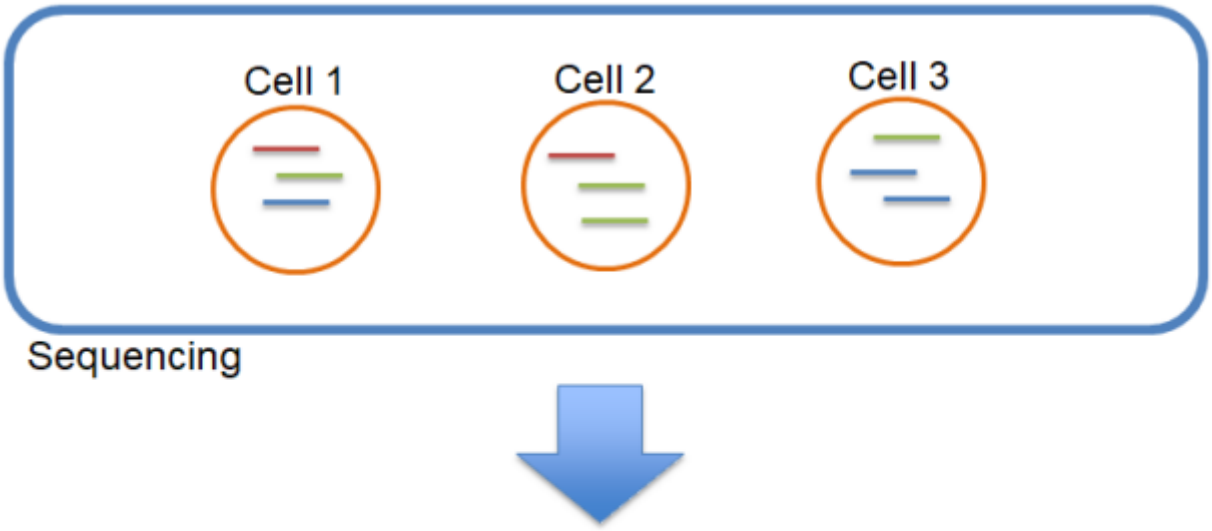
Figure 10.14 단일세포 FASTQ 파일 미리보기 Read2 - 단일세포 RNA-seq Read2의 FASTQ 파일 구조와 cDNA 서열 정보

Figure 10.15 단일세포 데이터의 드롭아웃과 과대산포 - 단일세포 데이터의 특징적인 통계적 성질

11.6.2 배치 효과 교정과 데이터 정규화

배치 효과의 이해: 단일세포 실험에서 배치 효과(batch effect)는 생물학적 변이가 아닌 기술적 요인으로 인해 발생하는 체계적 편향을 의미한다. 실험 수행 일자, 실험자, 시약 로트, 장비 상태, 시료 보관 조건 등 다양한 요인이 배치 효과를 유발할 수 있다. 이러한 효과는 실제 생물학적 신호를 가릴 수 있어 반드시 보정되어야 한다.

배치 교정 방법들: ComBat은 경험적 베이즈 방법을 사용하여 유전자별로 배치 효과를 모델링하고 제거한다. Harmony는 차원 축소된 공간에서 배치 효과를 교정하는 방법으로, 계산 효율성이 높고 대용량 데이터에 적합하다. Seurat Integration은 서로 다른 데이터셋 간의 공통 구조를 찾아 통합하는 방법으로, 상호 최근접 이웃(mutual nearest neighbors)을 이용한다. 추가적으로 BBKNN과 같은 도구들도 단일세포 데이터에서 배치 간 차이를 조정하는 데 사용된다.



	Cell 1	Cell 2	Cell3
Gene 1	1	1	0
Gene 2	1	2	1
Gene 3	1	0	2

Figure 198: 단일세포 RNA-seq 정량화

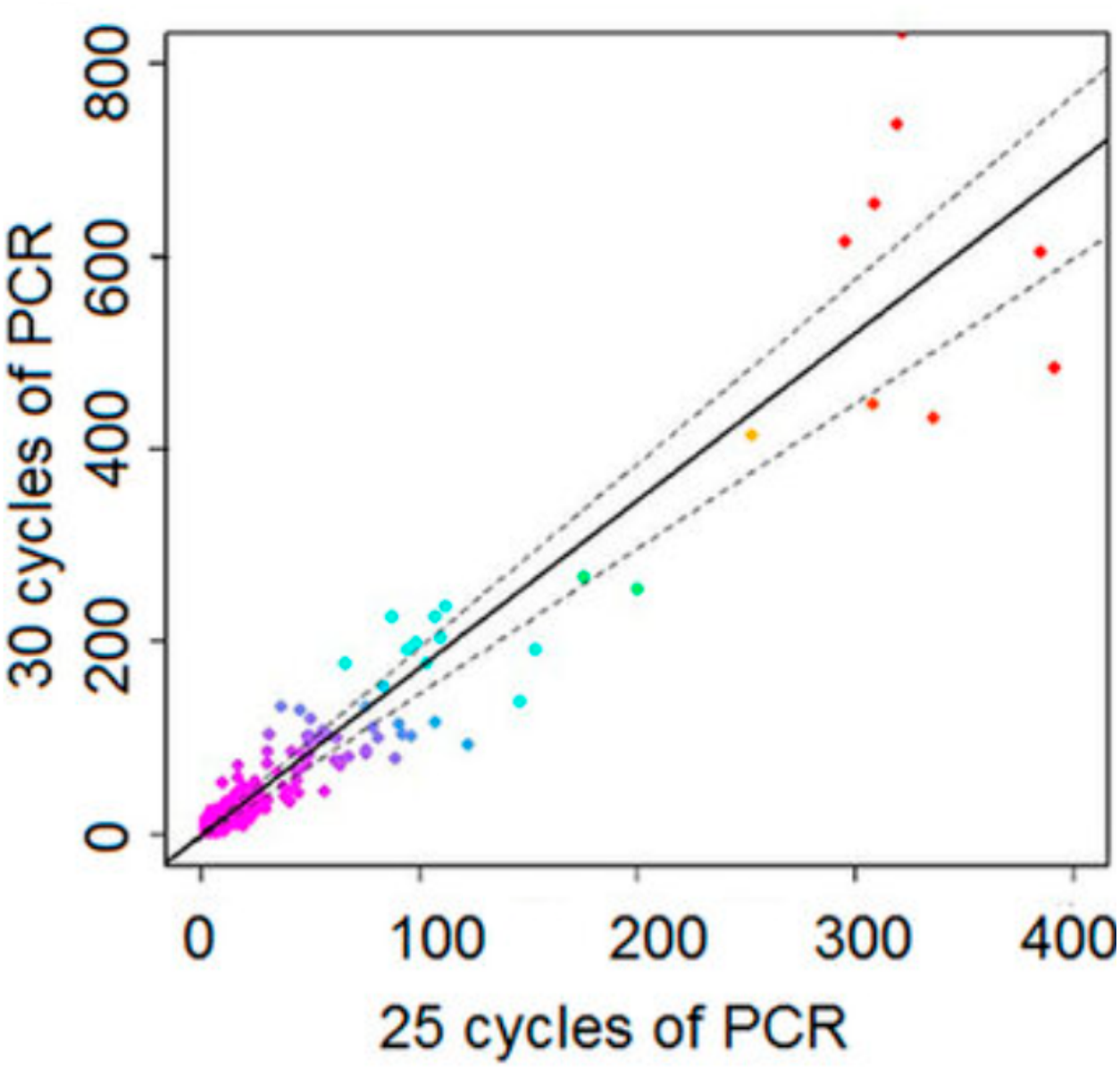


Figure 199: PCR 편향

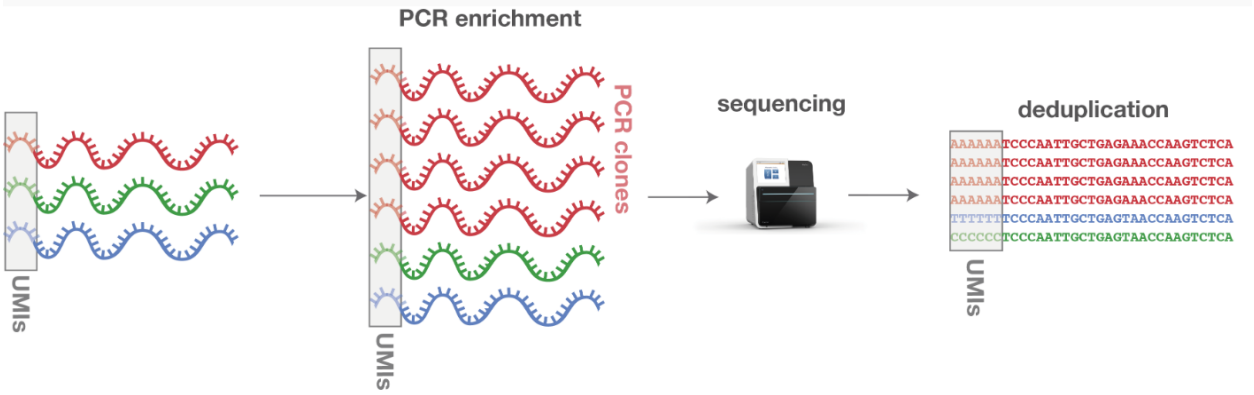


Figure 200: UMI 원리

```
❯ zcat pbmc_1k_v3_S1_L001_R1_001.fastq.gz | head
@A00228:279:HFWFVDMXX:1:1101:8486:1000 1:N:0:NCATTACT
NGTGATTAGCTGTACTCGTATGTAAGGT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 1:N:0:NCATTACT
NTCATGAAGTTTGGCTAGTTATGTTCAT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:12626:1000 1:N:0:NCATTACT
NACCAAAAGGACATCGCGGTAGAGGCTT
```

Figure 201: 단일세포 FASTQ 파일 미리보기 Read1

```
❯ zcat pbmc_1k_v3_S1_L001_R2_001.fastq.gz | head
@A00228:279:HFWFVDMXX:1:1101:8486:1000 2:N:0:NCATTACT
NACAAAGTCCCCCCCATAATACAGGGGGAGCCACTTGGGCAGGAGGCAGGGAGGGGTCCATTCCCCCTGGTGGGGCTGGTGGGGAGCTGTA
+
#FFFFFFFFFFFFFF:FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:10782:1000 2:N:0:NCATTACT
NTTGCAGCTGAACTGGTAAACTTGTCCCTAAAGAGACATAAGAATGGTCAACTGGAATGTGGATTCATCTGTAACATTACTCAGTGGGCCT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00228:279:HFWFVDMXX:1:1101:12626:1000 2:N:0:NCATTACT
GCTCCGTAGCCCTGACTGCCTTGGTGACTGTATCCTCAGAAGAATTTGAAGACAAGTGGTTCAGAAAGATCAAAGACCATTTCTGTCCATT
```

Figure 202: 단일세포 FASTQ 파일 미리보기 Read2

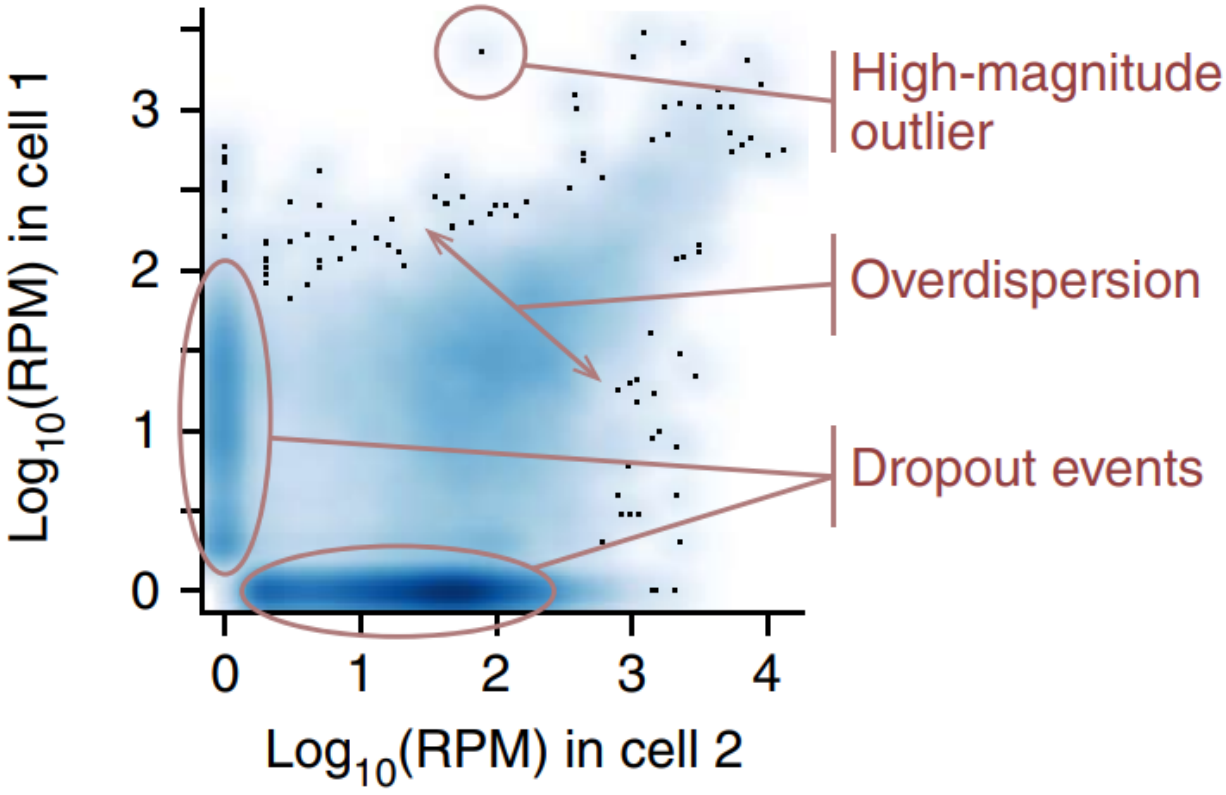


Figure 203: 단일세포 데이터의 드롭아웃과 과대산포

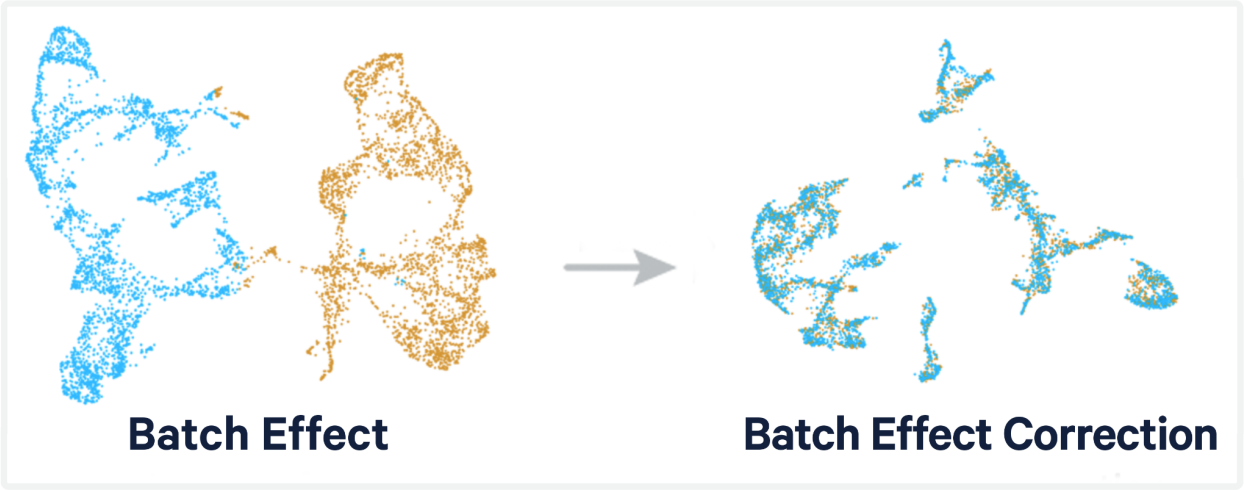


Figure 204: 배치 보정

Figure 10.16 배치 보정 – 서로 다른 배치 사이의 기술적 변이 제거 과정

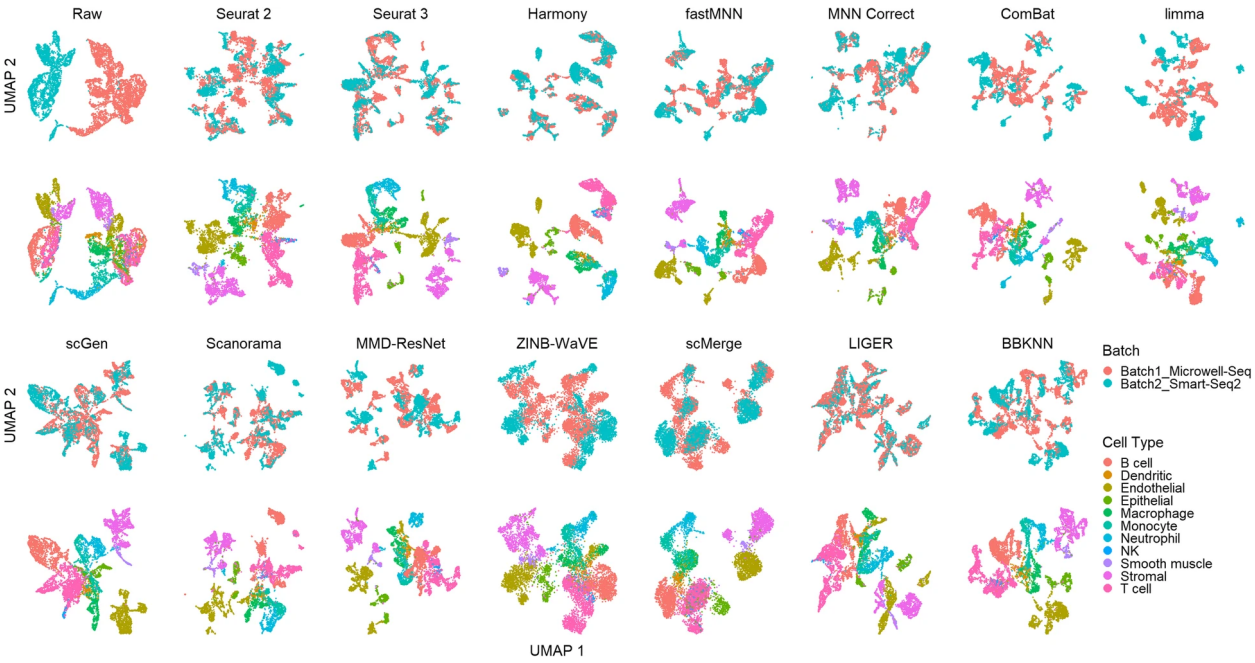


Figure 205: 다양한 단일세포 배치 보정 알고리즘

Figure 10.17 다양한 단일세포 배치 보정 알고리즘 – ComBat, Harmony, Seurat Integration 등 주요 배치 보정 방법들

정규화의 필요성: 세포마다 포획된 mRNA의 총량이 다르고, 유전자마다 길이가 달라 정규화가 필수적이다. 라이브러리 크기 정규화는 세포 간 시퀀싱 깊이 차이를 보정하고, 유전자 길이 정규화는 긴 유전자가 더 많은 리드를 생성하는 편향을 교정한다.

11.6.3 차원 축소와 그래프 기반 클러스터링

분석 파이프라인: 단일세포 데이터 분석에서는 차원 축소와 클러스터링이 핵심적인 단계이다. PCA를 통한 초기 차원 축소, k-최근접 이웃 그래프 구성, Leiden 클러스터링, 그리고 t-SNE/UMAP을 이용한 시각화 등의 구체적인 방법론과 매개변수 설정에 대한 상세한 내용은 11장 “차원 축소와 데이터 분석”에서 다룬다.

11.7 세포 유형 식별과 마커 유전자 분석

단일세포 분석에서 클러스터링이 완료되면, 각 클러스터가 어떤 세포 유형을 나타내는지 결정하는 것이 다음 중요한 단계이다. 이 과정은 생물학적 해석의 핵심이며, 정확한 세포 유형 주석(cell type annotation)이 후속 분석의 신뢰성을 좌우한다.

11.7.1 마커 유전자 발견

차별적 발현 분석: 각 클러스터에서 특이적으로 발현되는 유전자를 찾기 위해 차별적 발현 분석을 수행한다. 이는 특정 클러스터의 세포들과 다른 모든 세포들 간의 유전자 발현 차이를 통계적으로 검정하는 과정이다.일반적으로 Wilcoxon rank-sum test가 주로 사용되나, 최근의 추세는 “슈도벌크 접근법” 을 사용하는 것이다.

슈도벌크 접근법: 단일세포 데이터에서 차등 발현 분석의 또 다른 중요한 방법은 슈도벌크(pseudobulk) 접근법이다. 이 방법은 동일한 조건과 세포 유형의 세포들을 묶어 “슈도벌크” 샘플을 만든 뒤, DESeq2나 edgeR 같은 전통적인 벌크 분석 도구를 적용한다. 이 접근법의 장점은 생물학적 복제를 고려하면서 신뢰도 높은 차등 발현 분석을 수행할 수 있다는 점이다.

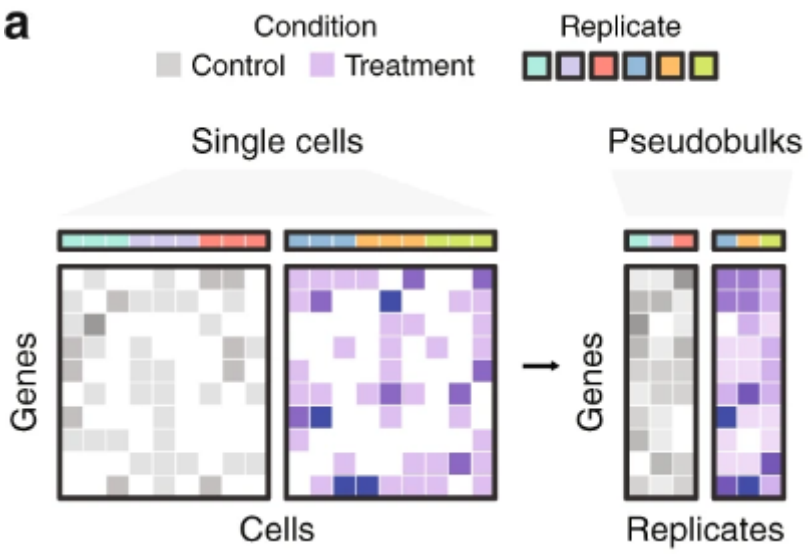


Figure 206: 슈도벌크 분석

Figure 10.18 슈도벌크 분석 – 단일세포 데이터에서 세포 집단을 묶어 벌크 분석 수행

마커 유전자의 특성: 좋은 마커 유전자는 다음과 같은 특성을 가져야 한다. 첫째, 특정 세포 유형에서 높은 발현량을 보여야 하고(높은 평균 발현), 둘째, 해당 세포 유형의 많은 세포에서 일관되게 발현되어야 하며(높은 발현 빈도), 셋째, 다른 세포 유형과 명확한 발현 차이를 보여야 한다(높은 특이성).

Figure 10.19 마커 유전자 – 세포 유형별 특이적 발현 패턴을 보이는 마커 유전자들

11.7.2 세포 유형 주석 과정

문헌 기반 주석: 발견된 마커 유전자들을 기존 문헌과 비교하여 세포 유형을 결정한다. 예를 들어, CD3E, CD3D, CD3G가 높게 발현되는 클러스터는 T세포로, CD19, CD20(MS4A1), PAX5가 발현되는 클러스터는 B세포로 주석할 수 있다. 이 과정에서는 단일 마커보다는 여러 마커의 조합을 고려하는 것이 중요하다.

데이터베이스 활용: CellMarker, PanglaoDB, Human Cell Atlas 같은 공개 데이터베이스를 활용하여 조직 특이적이고 세포 유형 특이적인 마커 정보를 얻을 수 있다. 이러한 데이터베이스는 수많은 연구에서 검증된 마커 정보를 제공하여 주석 과정의 객관성과 정확성을 높인다.

시각화를 통한 검증: 마커 유전자의 발현 패턴을 다양한 방법으로 시각화하여 세포 유형 주석의 타당성을 검증한다. 점 그래프(dot plot)는 각 클러스터에서 마커 유전자의 평균 발현량과 발현 세포 비율을 동시에 보여준다. Feature plot은 UMAP이나 t-SNE 좌표계에서 특정 유전자의 발현 분포를 색상으로 표현하여 공간적 패턴을 확인할 수 있게 한다.

Figure 10.20 점 그래프 – 세포 유형별 마커 유전자의 발현량과 발현 비율 시각화

11.7.3 자동화된 세포 유형 주석

참조 기반 방법: 기존에 잘 주석된 참조 데이터셋을 이용하여 새로운 데이터의 세포 유형을 자동으로 예측하는 방법이다. 이는 주로 참조 데이터셋의 각 세포 유형별 발현 프로파일과 새로운 데이터의 세포들을 비교하여 가장 유사한 세포 유형을 할당하는 방식으로 동작한다.

기계학습 기반 분류: 지도학습 알고리즘을 이용하여 세포 유형을 분류하는 방법도 널리 사용된다. 이는 미리 라벨링된 훈련 데이터셋을 이용하여 분류 모델을 학습시키고, 이를 새로운 데이터에 적용하는 방식이다. 최근에는 딥러닝 기반의 방법들도 개발되어 복잡한 세포 유형 분류에서 높은 성능을 보이고 있다.

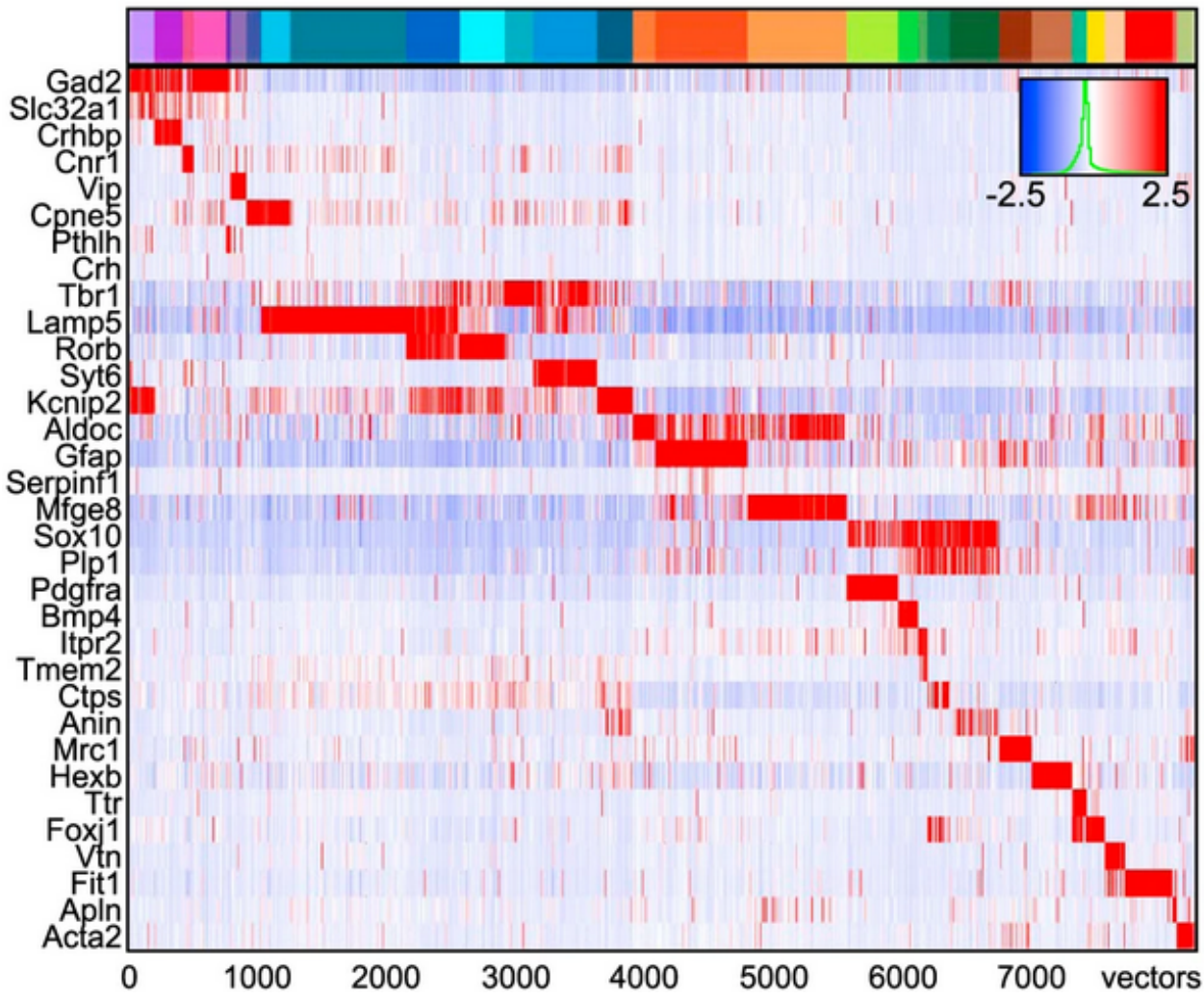


Figure 207: 마커 유전자

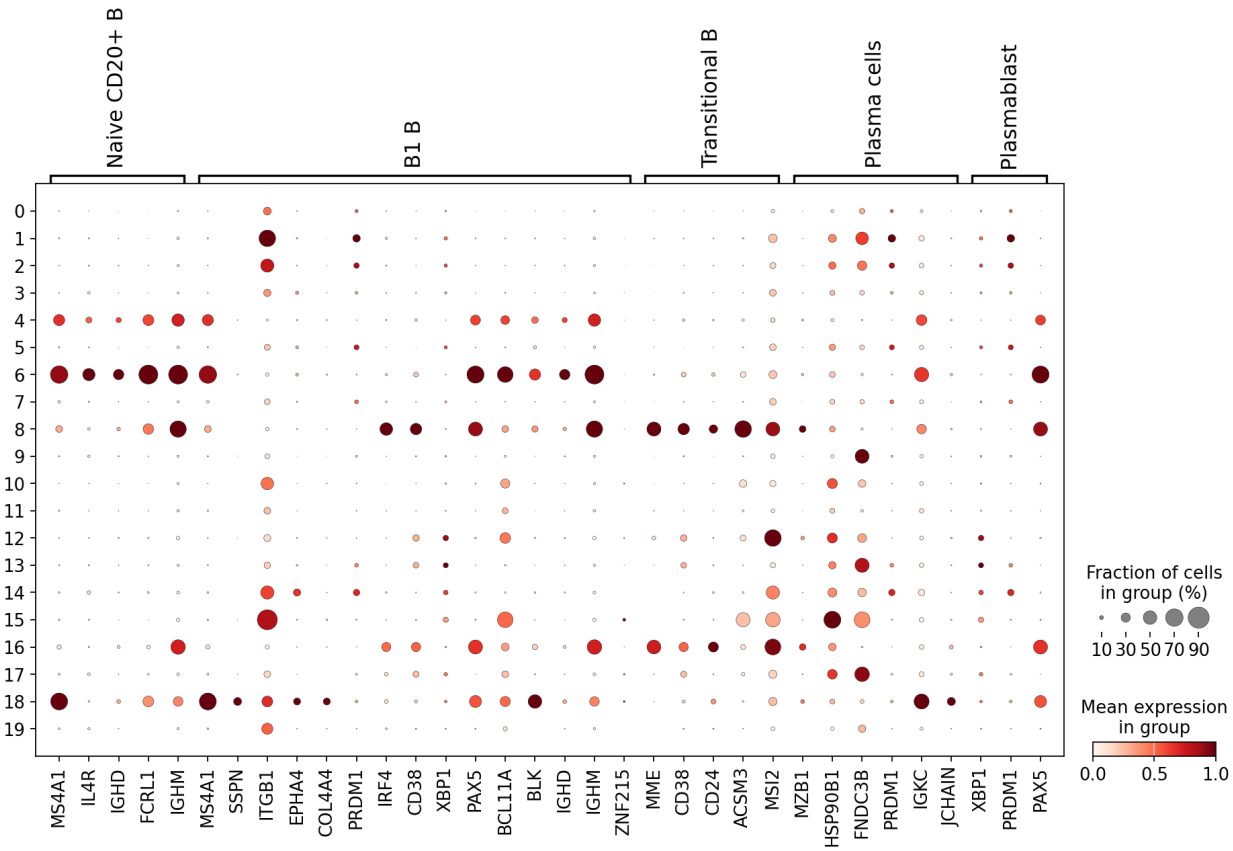


Figure 208: 점 그래프

11.8 대규모 단일세포 연구와 세포 아틀라스 프로젝트

단일세포 기술의 발전과 함께 인간과 모델 생물의 모든 세포 유형을 망라하는 포괄적인 세포 아틀라스 구축 프로젝트들이 활발히 진행되고 있다. 이러한 대규모 연구들은 생명과학 연구의 패러다임을 바꾸고 있으며, 질병 연구와 치료법 개발에 새로운 기회를 제공하고 있다.

11.8.1 Human Cell Atlas (HCA) 프로젝트

프로젝트의 비전과 목표: Human Cell Atlas는 인간 몸의 모든 세포 유형을 단일세포 수준에서 포괄적으로 매핑하는 국제적 연구 컨소시엄이다. 2016년 Aviv Regev(MIT/Broad Institute)와 Sarah Teichmann(Wellcome Sanger Institute)이 공동으로 설립한 이 프로젝트는 인간 발생, 성인 조직, 질병 상태에서의 모든 세포를 분자 수준에서 특성화하는 것을 목표로 한다.

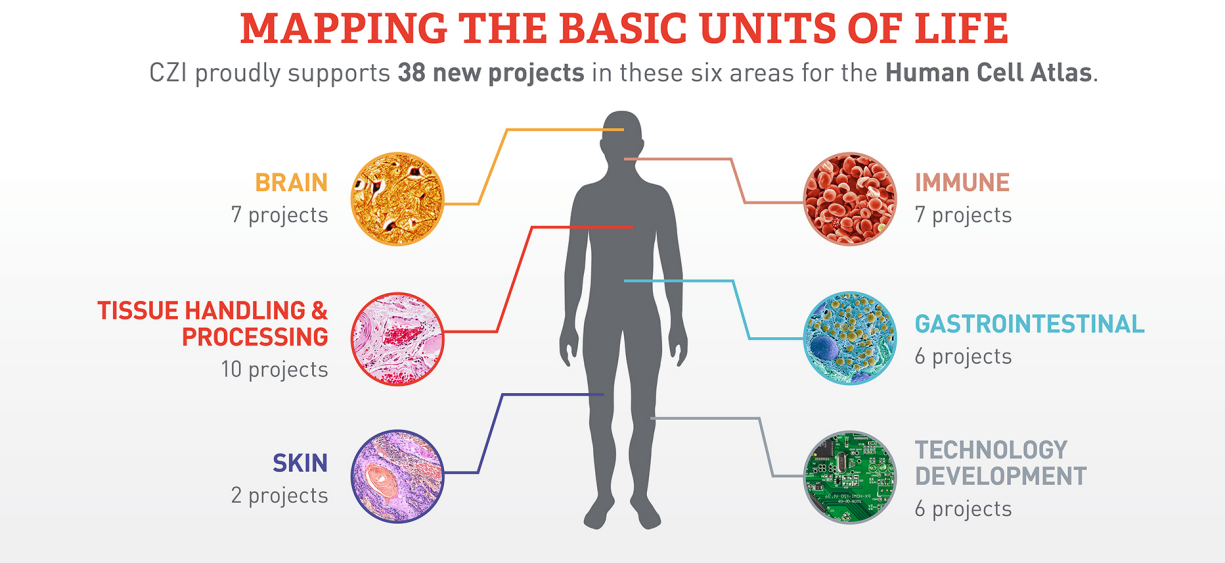


Figure 209: 휴먼 셀 아틀라스와 챈 주커버그 이니셔티브

Figure 10.21 휴먼 셀 아틀라스와 챈 주커버그 이니셔티브 – HCA 프로젝트와 주요 후원 기관들

이 프로젝트는 단순히 데이터를 수집하는 것을 넘어서, 표준화된 실험 프로토콜, 데이터 분석 방법론, 그리고 데이터 공유 체계를 구축하여 전 세계 연구자들이 일관된 방식으로 협력할 수 있는 기반을 마련하고 있다. 현재까지 수백만 개의 세포 데이터가 생성되었으며, 이는 인간 생물학에 대한 우리의 이해를 근본적으로 변화시키고 있다.

Figure 10.22 휴먼 참조 아틀라스 – HCA에서 개발된 인간 세포 유형의 포괄적 참조 데이터셋

Figure 10.23 휴먼 셀 아틀라스 – 인간 모든 세포 유형의 포괄적 매핑 프로젝트

주요 성과와 영향: HCA를 통해 이전에 알려지지 않았던 새로운 세포 유형들이 발견되었고, 기존에 알려진 세포 유형들의 아형(subtype)과 상태(state) 변화가 상세히 밝혀졌다. 특히 면역계, 신경계, 발생 과정에서의 세포 다양성과 상호작용에 대한 새로운 통찰이 제공되었다.

11.8.2 모델 생물 아틀라스 프로젝트

Tabula Muris (2018): 이 프로젝트는 마우스의 20개 장기와 조직에서 약 10만 개의 세포를 분석하여 포괄적인 마우스 세포 아틀라스를 구축했다. 이는 인간 연구의 중요한 참조점이 되며, 질병 모델 연구에서 세포 수준의 비교 분석을 가능하게 한다. 특히 aging, 성별 차이, 조직 특이적 세포 분포 등에 대한 중요한 기초 자료를 제공한다.

Figure 10.24 Tabula Muris – 마우스 20개 장기의 포괄적 단일세포 아틀라스

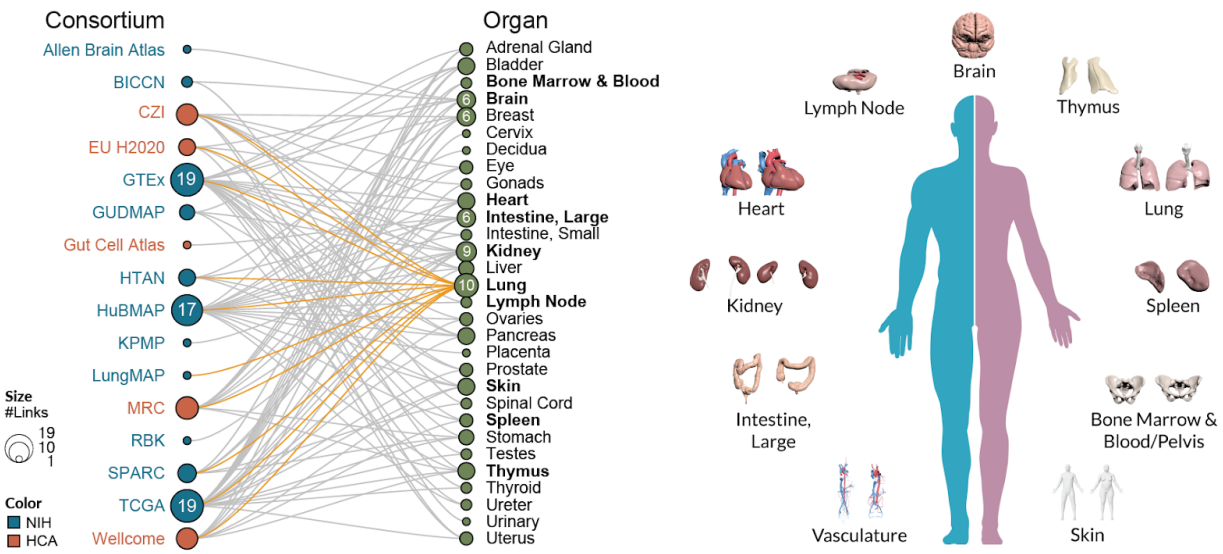


Figure 210: 휴먼 참조 아틀라스

Figure 211: 휴먼 셀 아틀라스

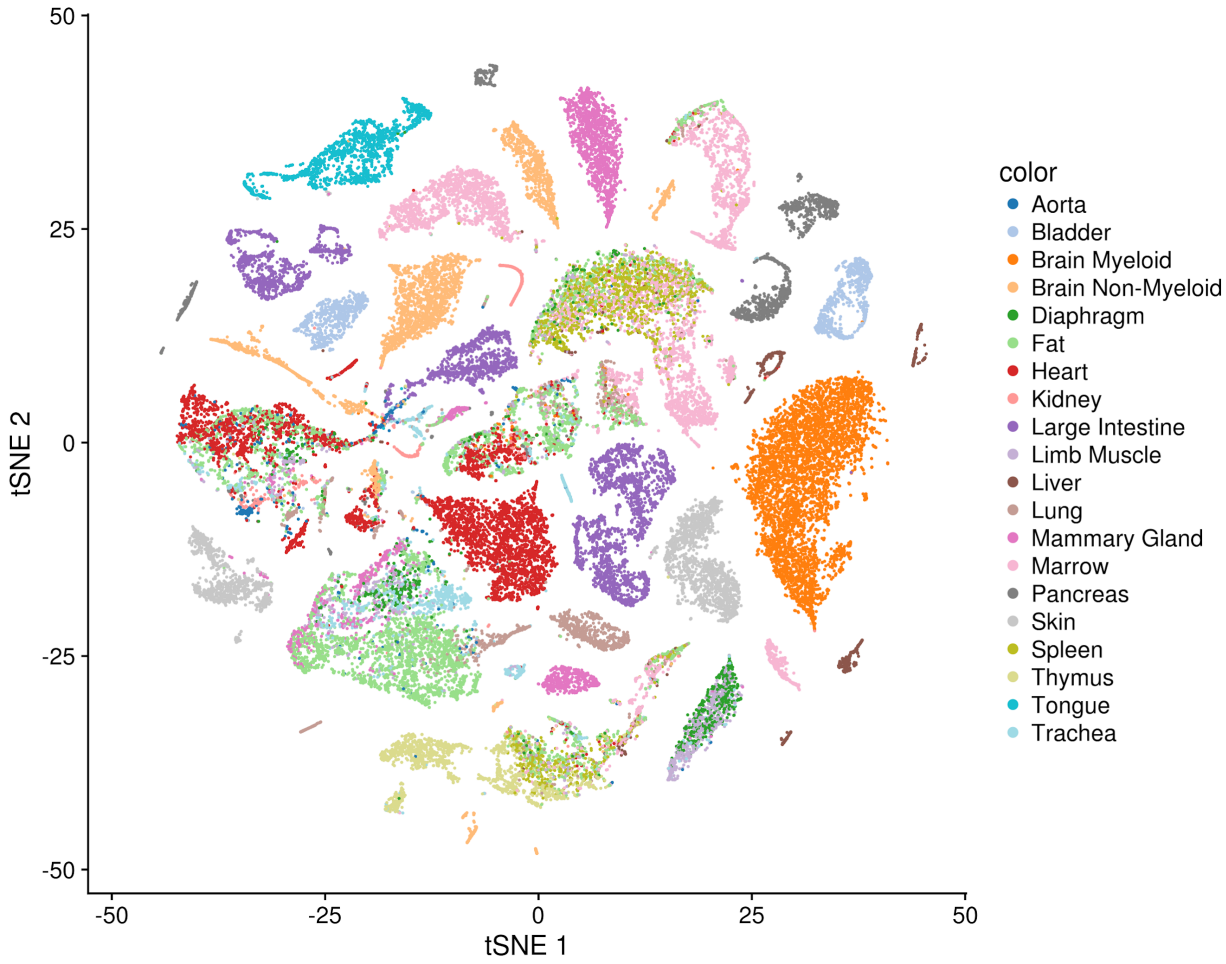


Figure 212: Tabula Muris

Tabula Sapiens (2022): 성인 인간의 24개 장기와 조직을 대상으로 한 대규모 단일세포 연구로, 약 50만 개의 세포 데이터를 포함한다. 이 연구는 인간 성인 조직의 세포 구성과 기능을 이해하는 데 중요한 기준점을 제공하며, 질병 연구와 치료법 개발에 직접적으로 활용되고 있다.

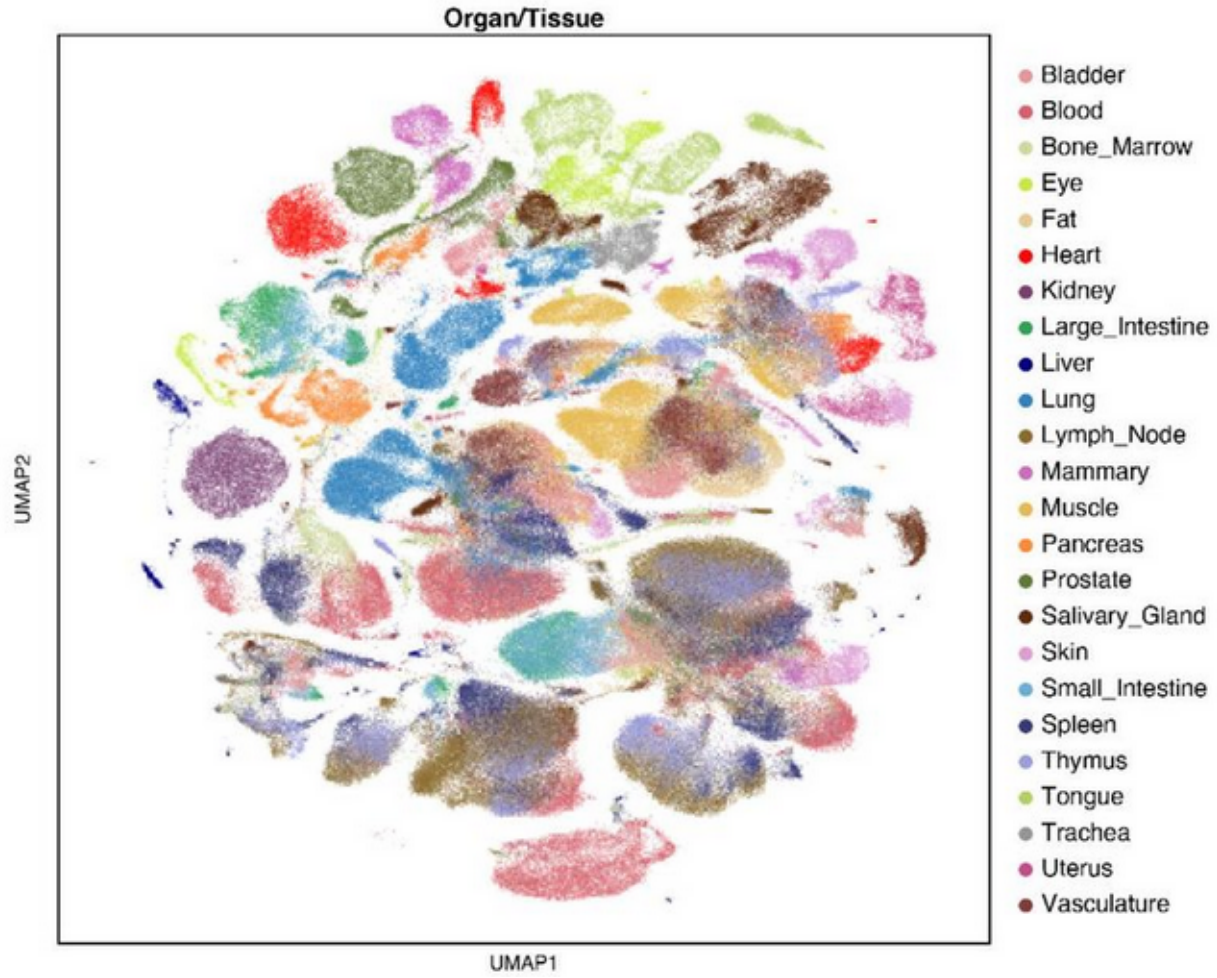


Figure 213: Tabula Sapiens

Figure 10.25 Tabula Sapiens – 성인 인간 24개 장기의 대규모 단일세포 연구

11.8.3 참조 매핑과 전이 학습

참조 기반 세포 유형 예측: 잘 구축된 세포 아틀라스는 새로운 실험 데이터에서 세포 유형을 자동으로 식별하는 참조 데이터셋으로 활용된다. 이는 연구자들이 매번 처음부터 세포 유형을 주석할 필요 없이, 기존의 고품질 참조 데이터와 비교하여 빠르고 정확하게 세포 유형을 결정할 수 있게 해준다.

파운데이션 모델과 제로샷 학습: 최근에는 수백만 개의 세포 데이터로 사전 훈련된 대규모 기계학습 모델들이 개발되고 있다. scGPT, nicheformer와 같은 모델들은 새로운 조직이나 질병 샘플에서도 사전 훈련 없이 세포 유형을 예측할 수 있는 제로샷(zero-shot) 전이 학습 능력을 보여준다. 이러한 접근법은 희귀 질병이나 새로운 조직에서의 세포 분석을 크게 용이하게 한다.

11.9 세포 디콘볼루션: 벌크 데이터에서 세포 구성 추정

벌크 RNA 시퀀싱 데이터는 조직 전체의 평균적인 유전자 발현을 측정하지만, 실제로는 다양한 세포 유형이 혼재되어 있다. 세포 디콘볼루션(cellular deconvolution)은 이러한 혼합된 신호에서 각 세포 유형의 비율과 특성을 추정하는

계산적 방법이다. 이는 특히 임상 샘플이나 대규모 코호트 연구에서 세포 구성 변화와 질병의 관련성을 이해하는 데 중요하다.

11.9.1 참조 기반 디콘볼루션 방법

기본 원리와 수학적 모델링: 참조 기반 디콘볼루션은 벌크 RNA-seq 데이터가 개별 세포 유형의 발현 프로파일의 선형 조합으로 표현될 수 있다는 가정에 기반한다. 수학적으로는 다음과 같이 표현된다:

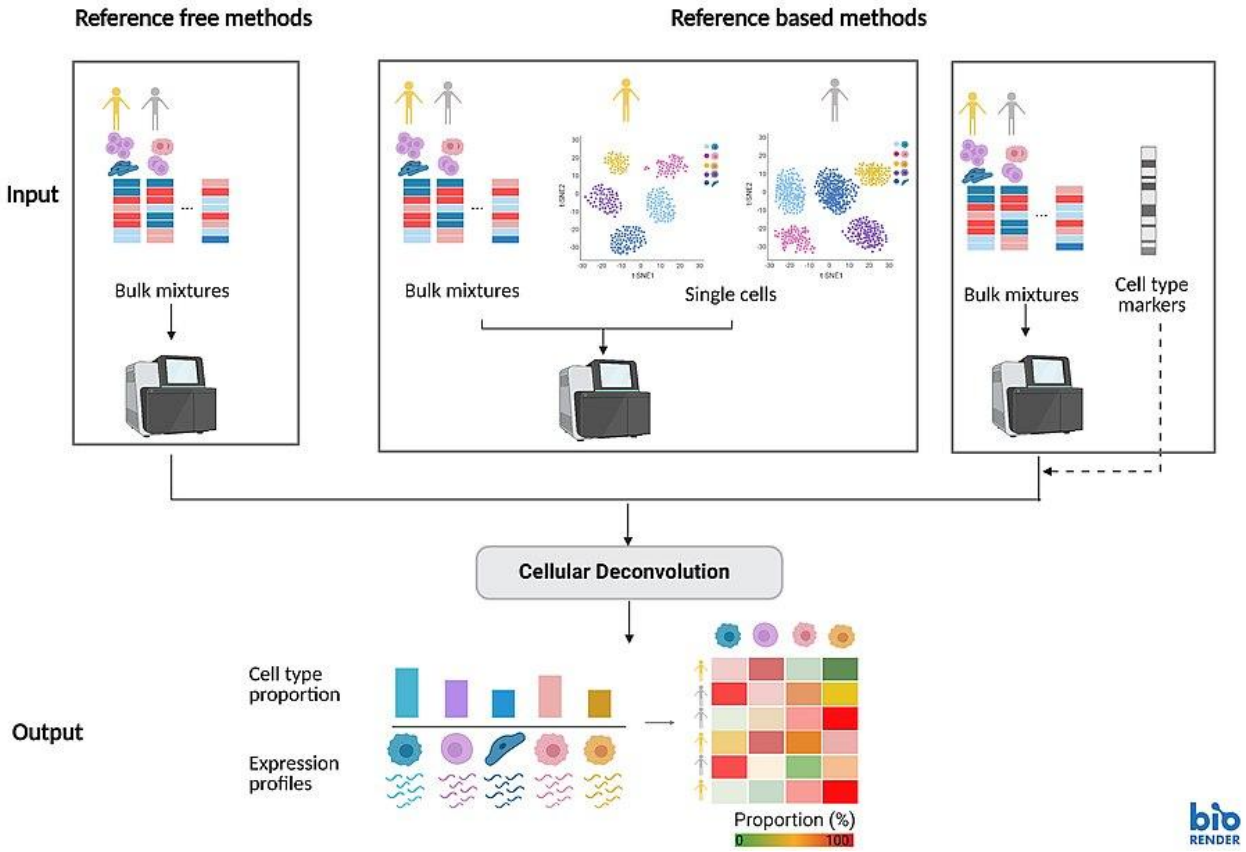


Figure 214: 세포 디콘볼루션 원리

Figure 10.26 세포 디콘볼루션 원리 – 벌크 데이터에서 세포 유형별 비율 추정의 수학적 모델

$$Y = X \times \beta + \varepsilon$$

여기서 Y는 벌크 샘플의 유전자 발현 벡터, X는 각 세포 유형의 참조 발현 프로파일 행렬, β는 각 세포 유형의 비율 벡터, ε는 노이즈 항이다. 이 방정식을 풀어 β를 추정하는 것이 디콘볼루션의 핵심이다. 여기에는 대표적으로 CIBERSORT와 같은 방법론이 있다.

참조 데이터셋의 중요성: 정확한 디콘볼루션을 위해서는 고품질의 세포 유형별 참조 발현 프로파일이 필요하다. 이러한 참조 데이터는 단일세포 RNA-seq, 정제된 세포 유형별 RNA-seq, 또는 면역형광으로 분류된 세포들의 발현 데이터에서 얻을 수 있다. 참조 데이터의 품질과 포괄성이 디콘볼루션 결과의 정확성을 크게 좌우한다.

11.9.2 참조 프리 디콘볼루션 방법

비음수 행렬 분해(Non-negative Matrix Factorization, NMF): 참조 데이터가 없는 상황에서는 NMF와 같은 비지도 학습 방법을 사용할 수 있다. NMF는 벌크 발현 행렬 V를 두 개의 비음수 행렬 W(세포 유형별 시그니처)와 H(세포 유형 비율)의 곱으로 분해한다:

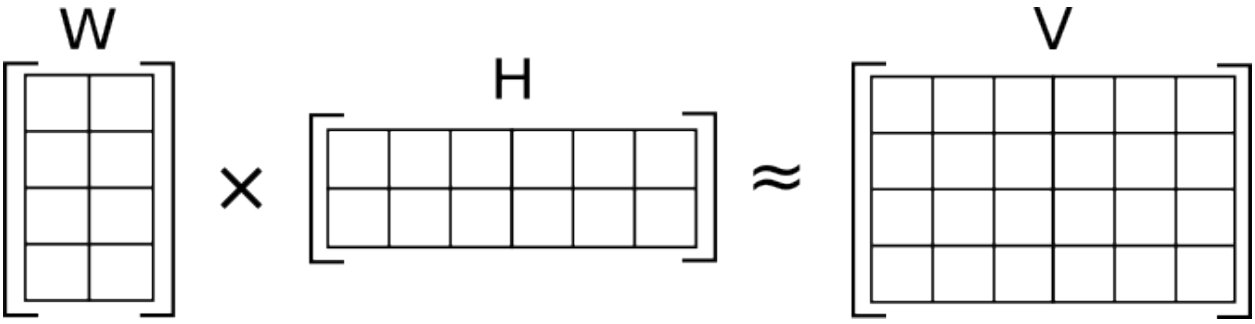


Figure 215: 비음수 행렬 분해

Figure 10.27 비음수 행렬 분해 – NMF를 이용한 참조 프리 디콘볼루션 방법

$$\mathbf{H}^{n+1}_{[i,j]} \leftarrow \mathbf{H}^{n}_{[i,j]} \frac{\left((\mathbf{W}^n)^T \mathbf{V}\right)_{[i,j]}}{\left((\mathbf{W}^n)^T \mathbf{W}^n \mathbf{H}^n\right)_{[i,j]}}$$

and

$$\mathbf{W}^{n+1}_{[i,j]} \leftarrow \mathbf{W}^{n}_{[i,j]} \frac{\left(\mathbf{V}(\mathbf{H}^{n+1})^T\right)_{[i,j]}}{\left(\mathbf{W}^n \mathbf{H}^{n+1}(\mathbf{H}^{n+1})^T\right)_{[i,j]}}$$

Figure 216: NMF 알고리즘

Figure 10.28 NMF 알고리즘 – 비음수 행렬 분해의 반복적 최적화 알고리즘과 수렴 과정

```
V = W × H
```

이 방법의 장점은 사전 정의된 세포 유형에 의존하지 않고도 데이터 자체에서 세포 유형 시그니처를 발견할 수 있다는 것이다. 하지만 발견된 시그니처의 생물학적 해석이 어려울 수 있다는 한계가 있다.

통계적 최적화: NMF의 목적 함수는 일반적으로 다음과 같이 정의된다:

```
min ||V - WH||²_F + R(W,H)
```

여기서 첫 번째 항은 재구성 오차를, 두 번째 항은 정규화 항을 나타낸다. 정규화 항은 과적합을 방지하고 생물학적으로 의미 있는 해를 찾는 데 도움이 된다.

11.9.3 디콘볼루션 결과의 검증과 해석

검증 방법들: 디콘볼루션 결과의 정확성을 검증하기 위해 여러 방법이 사용된다. 모의 데이터(simulated data)를 이용한 검증에서는 알려진 세포 비율로 구성된 가상의 벌크 데이터를 생성하고 추정 결과와 비교한다. 실험적 검증으로는 면역조직화학염색이나 플로우 사이토메트리로 측정한 세포 비율과 비교할 수 있다.

임상적 응용: 디콘볼루션은 암 연구에서 종양 미세환경의 면역세포 침윤 정도를 평가하거나, 자가면역질환에서 조직별 면역세포 분포 변화를 분석하는 데 활용된다. 또한 치료 반응 예측이나 예후 평가에서도 중요한 바이오마커로 사용되고 있다.

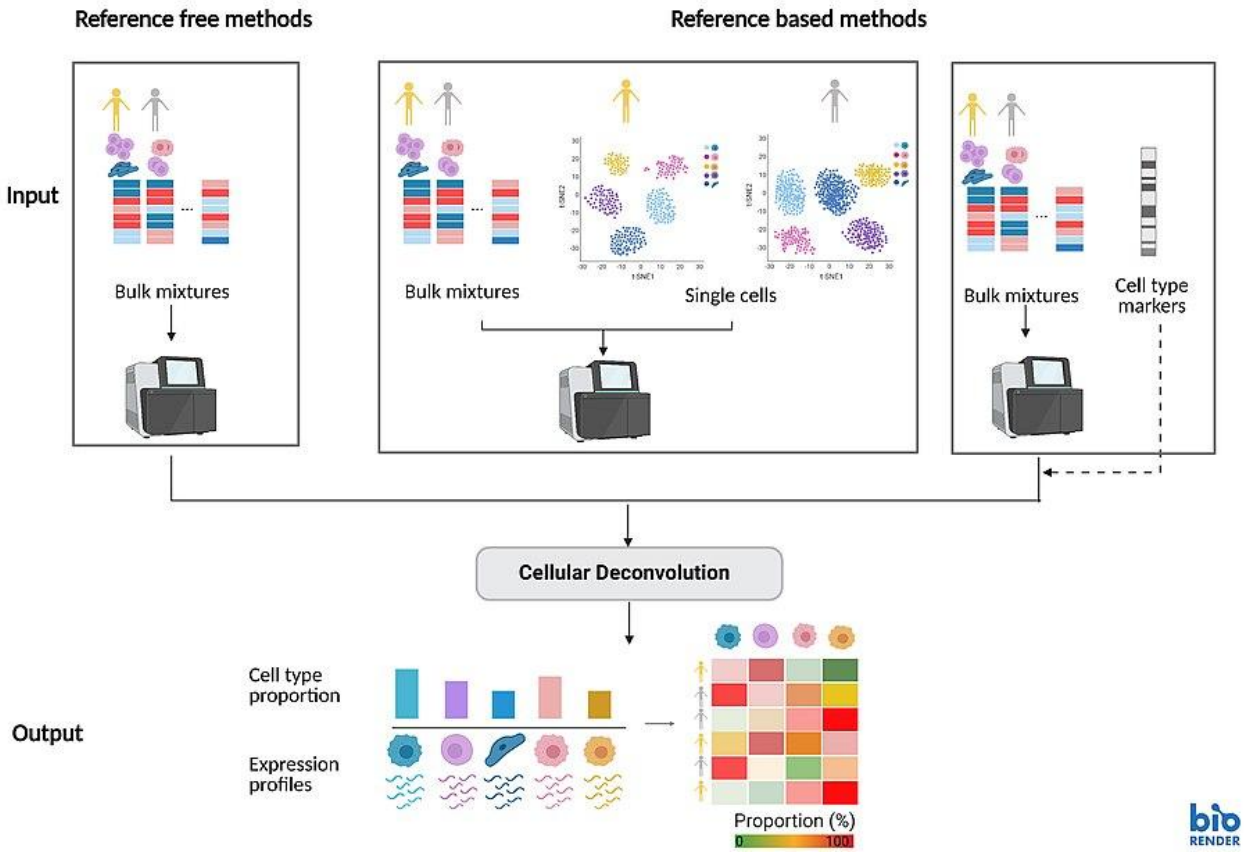


Figure 217: 세포 디콘볼루션 검증

Figure 10.29 세포 디콘볼루션 검증 – 디콘볼루션 결과의 검증 방법과 임상적 응용 사례

11.10 임상 응용과 미래 전망

11.10.1 정밀의학에서의 역할

단일세포 전사체학은 개별 환자의 세포 구성과 기능 상태를 정밀하게 분석하여 맞춤형 치료법을 선택하는 정밀의학의 핵심 도구로 자리잡고 있다. 특히 면역치료 효과 예측, 치료 저항성 메커니즘 규명, 그리고 새로운 치료 표적 발견에서 중요한 역할을 담당하고 있다.

11.10.2 기술 발전 방향

향후 단일세포 기술은 더 높은 처리량, 더 낮은 비용, 그리고 실시간 분석 능력을 갖춘 방향으로 발전할 것으로 예상된다. 또한 다중 오믹스 통합 분석과 인공지능 기반 자동화 분석 파이프라인의 개발이 가속화될 것이다.

11.10.3 도전 과제와 해결 방안

표준화 부족, 배치 효과, 대용량 데이터 처리 등의 기술적 도전과 함께, 개인정보 보호와 윤리적 문제들도 해결해야 할 중요한 과제이다. 이를 위해 국제적 협력과 표준화된 프로토콜 개발이 지속되고 있다.

단일세포 전사체학은 생명과학 연구와 의학 발전에 지속적으로 혁신적인 변화를 가져올 것이며, 이를 통해 질병의 이해와 치료에서 새로운 돌파구를 제공할 것으로 기대된다.

12 차원 축소와 데이터 분석

12.1 고차원 데이터의 문제점과 차원 축소의 필요성

의생명정보학에서 다루는 데이터는 종종 고차원의 특성을 가진다. 예를 들어, 유전자 발현 데이터에서는 수만 개의 유전자가 변수로 작용하고, 단일세포 RNA 시퀀싱 데이터에서는 각 세포가 2만여 개의 유전자 발현값을 가질 수 있다. 이러한 고차원 데이터는 분석과 해석에 있어 여러 가지 근본적인 문제를 야기한다.

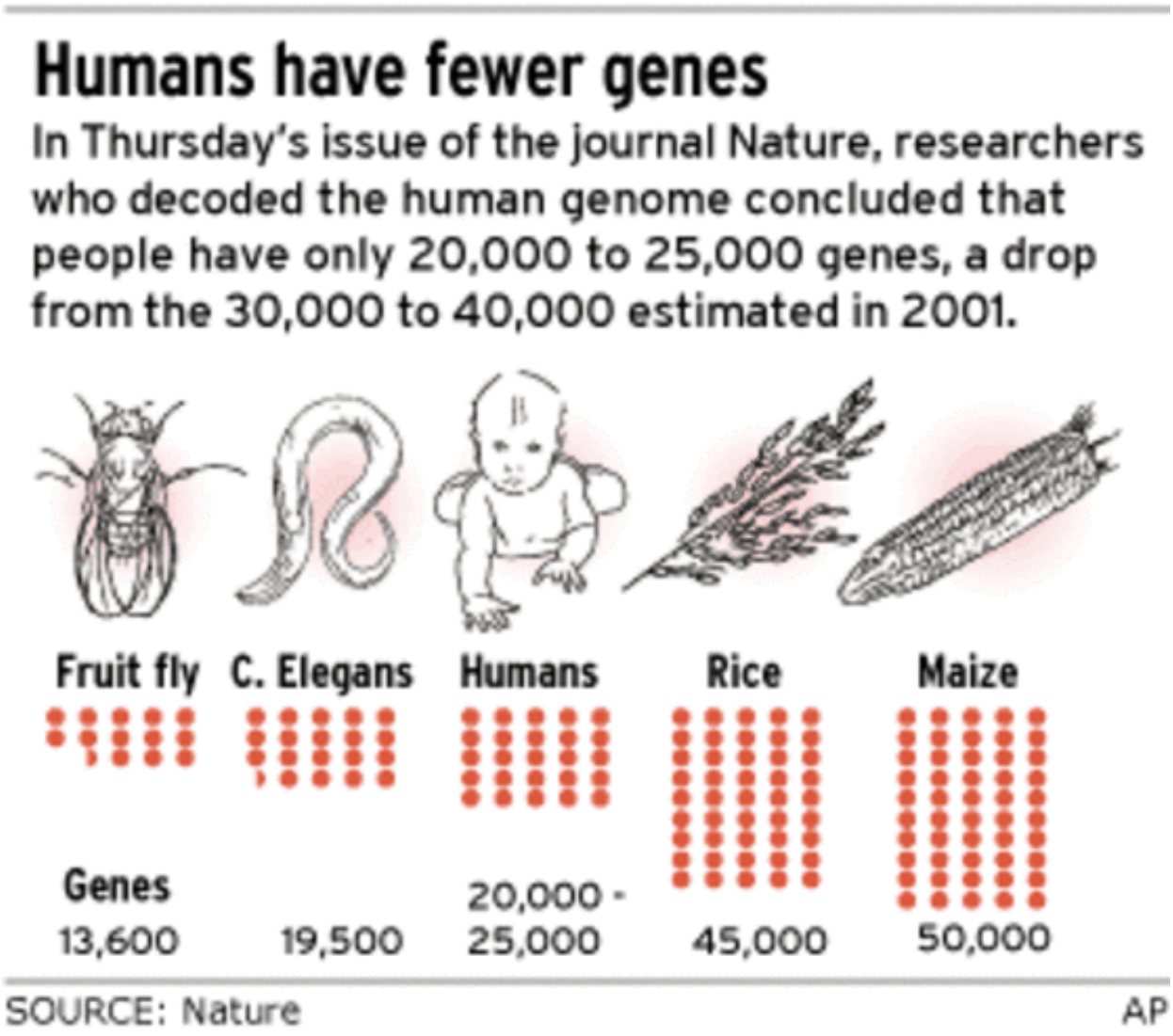


Figure 218: 고차원 데이터의 유전자 수

Figure 11.1 고차원 데이터의 유전자 수 – 의생명정보학 데이터에서 다루어야 할 방대한 수의 유전자

12.1.1 차원의 저주(Curse of Dimensionality)

차원의 저주는 고차원 공간에서 데이터 분석이 직면하는 핵심적인 문제이다. 이 현상을 수학적으로 이해하기 위해 유클리드 거리 공식을 살펴보자:

$$d(p,q) = \sqrt{[(p-q)^2 + (p-q)^2 + \ldots + (p-q)^2]}$$

차원 수 n이 증가할수록 거리 계산에 포함되는 항의 수가 늘어나며, 이는 데이터 포인트들 간의 거리가 점점 비슷해지는 결과를 낳는다. 고차원 공간에서는 모든 점들이 서로 "멀리" 떨어져 있게 되어, 거리 기반 알고리즘의 분별력이 급격히 저하된다. 이는 최근접 이웃 탐색, 클러스터링, 분류 등 많은 기계학습 알고리즘의 성능에 심각한 영향을 미친다.

12.1.2 희소성(Sparsity) 문제

의생명정보학 데이터, 특히 단일세포 RNA 시퀀싱 데이터는 극심한 희소성을 보인다. 전체 측정값의 70~90%가 0이거나 거의 0에 가까운 값을 가지는 것이 일반적이다. 이러한 희소성은 여러 요인에 기인한다:

기술적 요인: 개별 세포에서 추출할 수 있는 RNA의 양이 제한적이고, 현재 시퀀싱 기술의 검출 한계로 인해 낮은 발현량을 가진 유전자들이 측정되지 않을 수 있다. 또한 PCR 증폭 과정에서 발생하는 편향이나 시퀀싱 깊이의 한계도 희소성을 증가시킨다.

생물학적 요인: 많은 유전자들이 특정 조건이나 시점에서만 발현되며, 전사 과정 자체가 확률적(stochastic) 특성을 가지고 있다. 특히 단일세포 수준에서는 유전자 발현이 버스트(burst) 형태로 일어나기 때문에 측정 시점에 따라 0 또는 높은 값을 보일 수 있다.

12.1.3 밀집(Dense)과 희소(Sparse) 데이터의 처리 전략

데이터의 희소성 정도에 따라 적절한 분석 전략을 선택해야 한다. **밀집 데이터**는 대부분의 변수에서 의미 있는 값이 관측되는 경우로, 전통적인 통계 기법과 기계학습 알고리즘이 안정적으로 작동한다. 반면 **희소 데이터**는 관측되지 않은 값(missing value)이나 0값이 많아 모델의 불안정성을 야기할 수 있다.

희소 데이터를 다룰 때는 다음과 같은 특별한 고려사항이 필요하다: 1) 0값이 진정한 0인지 아니면 기술적 한계로 인한 결측값인지 구별해야 하고, 2) 희소성을 고려한 정규화 기법을 적용해야 하며, 3) 희소 행렬에 최적화된 알고리즘을 선택해야 한다. 이러한 문제들을 해결하기 위해 차원 축소는 데이터의 본질적인 구조를 보존하면서도 계산적으로 다루기 쉬운 저차원 표현으로 변환하는 핵심적인 역할을 담당한다.

12.2 주성분 분석(Principal Component Analysis, PCA)

주성분 분석은 차원 축소의 가장 기본이 되는 선형 기법으로, 1901년 Karl Pearson에 의해 처음 소개되었다. PCA의 핵심 아이디어는 고차원 데이터의 분산을 최대한 보존하면서 더 적은 수의 변수(주성분)로 데이터를 표현하는 것이다.

단일 세포 데이터 분석의 주 목적 중 하나는 세포별로 그 세포 유형을 구분해내는 것이다. 그런데 동일한 생물학적 경로(pathway)에 속한 유전자들은 함께 발현되는 경향이 있어 서로 높은 상관성을 나타내므로, 유전자 간 상관관계(공분산)을 기반으로 한 차원 축소 기법인 PCA가 특히 유용하다.

Figure 11.2 유전자 간 상관관계 – 동일 생물학적 경로에 속한 유전자들의 높은 상관성

12.2.1 PCA의 수학적 원리

PCA의 수학적 과정은 다음과 같은 단계로 구성된다:

1단계: 데이터 중심화(Centering) 먼저 각 변수(유전자)의 평균을 0으로 맞춘다. 이는 모든 데이터 포인트에서 해당 변수의 평균값을 빼는 과정이다.

2단계: 공분산 행렬 계산 중심화된 데이터로부터 공분산 행렬을 계산한다. 이 행렬은 변수들 간의 선형 관계를 나타낸다.

3단계: 고유값 분해(Eigenvalue Decomposition) 공분산 행렬을 고유값과 고유벡터로 분해한다. 고유벡터는 주성분의 방향을 나타내고, 고유값은 해당 방향에서의 분산의 크기를 나타낸다.

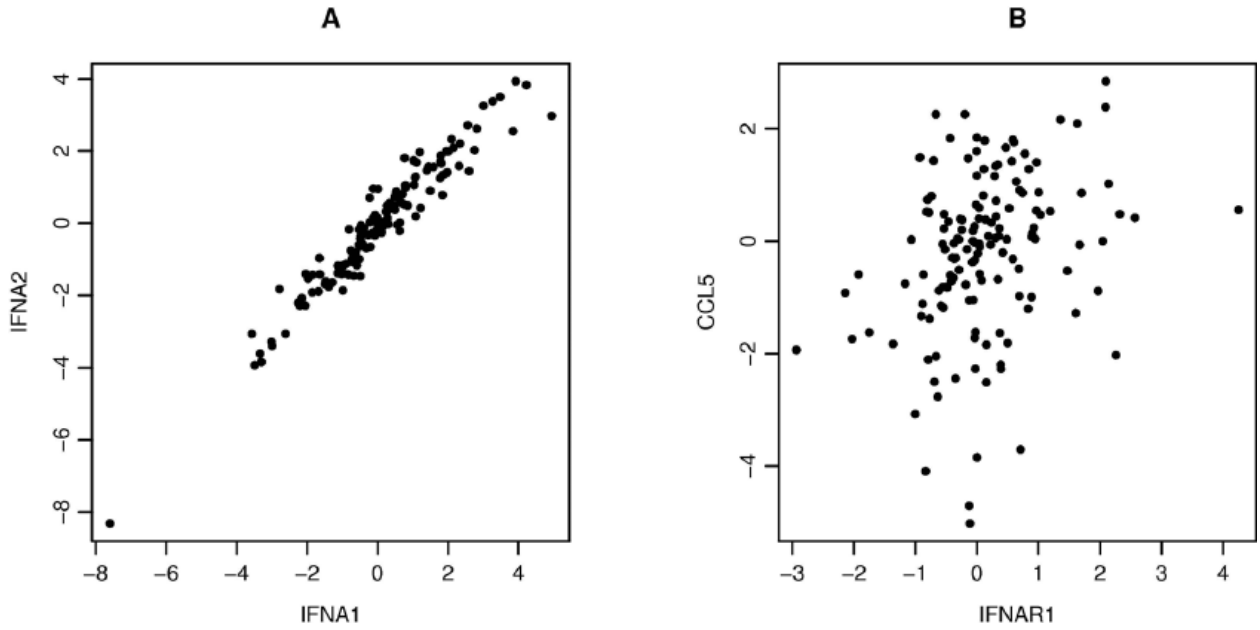


Figure 219: 유전자 간 상관관계

공분산 행렬은 에르미트 행렬(Hermitian matrix)이므로 다음과 같은 중요한 성질을 가진다: - 모든 고유값이 실수이다 - 고유벡터들이 서로 직교한다(orthogonal) - 고유값의 크기에 따라 고유벡터를 정렬할 수 있으며, 가장 큰 고유값에 해당하는 고유벡터가 데이터의 최대 분산 방향을 나타낸다

Figure 11.3 PCA 개요 - 주성분 분석의 기본 원리와 고유벡터 방향

12.2.2 최적 주성분 개수 결정

적절한 주성분의 개수를 선택하는 것은 PCA 분석에서 중요한 결정이다. 이를 위해 여러 방법이 사용된다:

엘보우 플롯(Elbow Plot): 주성분의 개수에 따른 설명된 분산의 비율을 그래프로 나타낸 것이다. 그래프에서 기울기가 급격히 완만해지는 지점(엘보우 포인트)을 선택한다.

누적 분산 기여율: 일반적으로 전체 분산의 80-95%를 설명하는 주성분까지를 선택한다. 단일 세포 데이터에서는 보통 상위 50-100개의 주성분으로도 충분한 정보를 보존할 수 있다.

Figure 11.4 엘보우 플롯 - 최적 주성분 개수 결정을 위한 설명된 분산비율 그래프

12.2.3 의생명정보학에서의 PCA 응용 사례

암 아형 분류: 다양한 암 유형의 유전자 발현 데이터에 PCA를 적용하면, 각 암 유형이 주성분 공간에서 서로 다른 클러스터를 형성함을 시각화하여 볼 수 있다. 이는 각 암 유형이 고유한 전사체 시그니처를 가지고 있음을 보여준다.

세포 유형 식별: 단일세포 RNA 시퀀싱 데이터에서 PCA를 수행하면 상피세포(epithelial cells), 간질세포(stromal cells), T세포, 골수계 세포(myeloid cells) 등 서로 다른 세포 유형들이 주성분 공간에서 뚜렷이 구별된다. 이를 통해 복잡한 조직 내의 세포 이질성을 이해할 수 있다.

Figure 11.5 PCA 암 세포 유형 - PCA를 이용한 다양한 암 세포 유형의 분류와 시각화

Figure 11.6 PCA 예시 - 단일세포 데이터에서 PCA를 이용한 세포 유형 식별

12.3 비선형 차원 축소

선형 기법으로는 궤적 구조나 매니폴드 같은 복잡한 패턴을 온전히 표현하기 어렵다. 이를 보완하기 위해 비선형 방법이 등장했다.

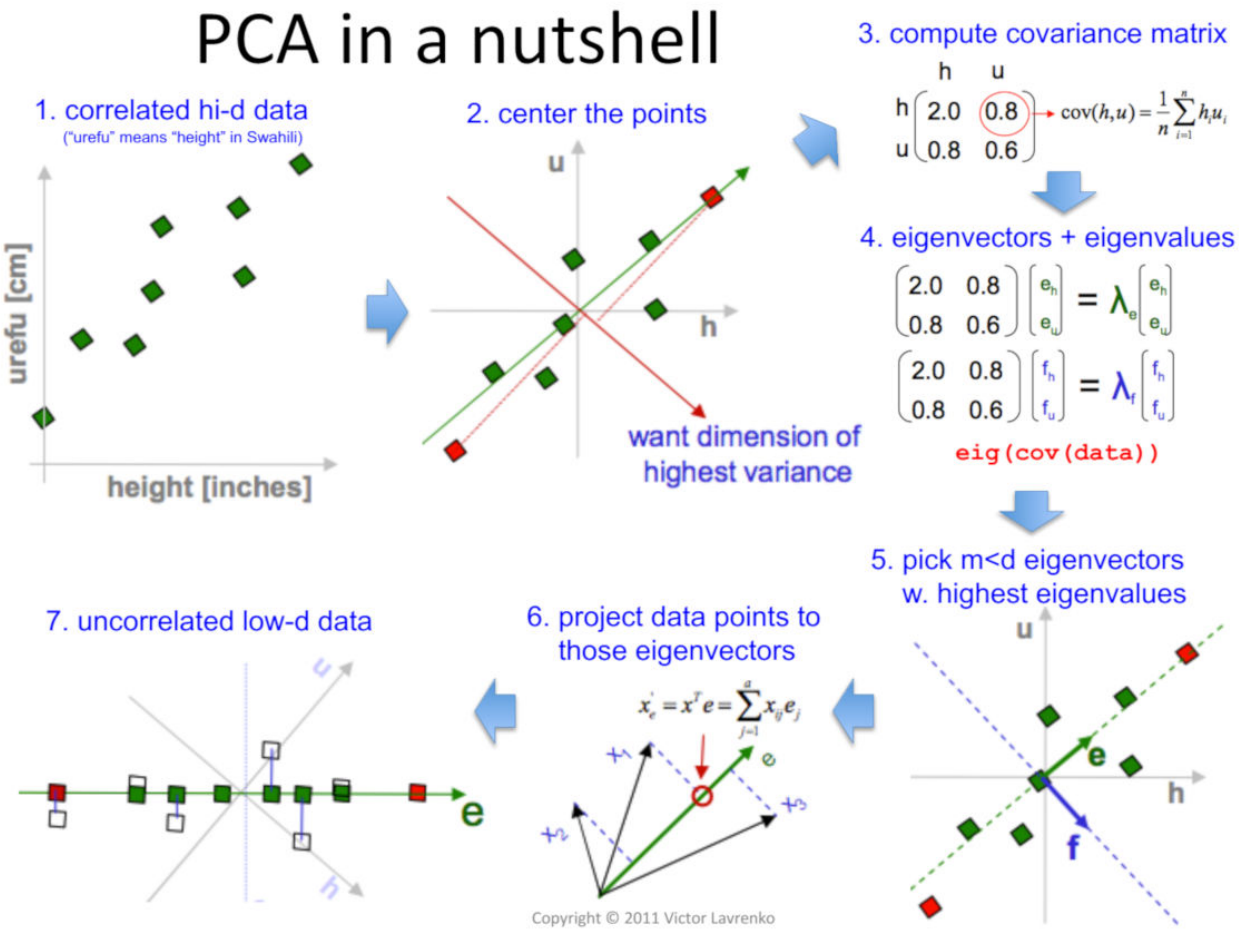


Figure 220: PCA 개요

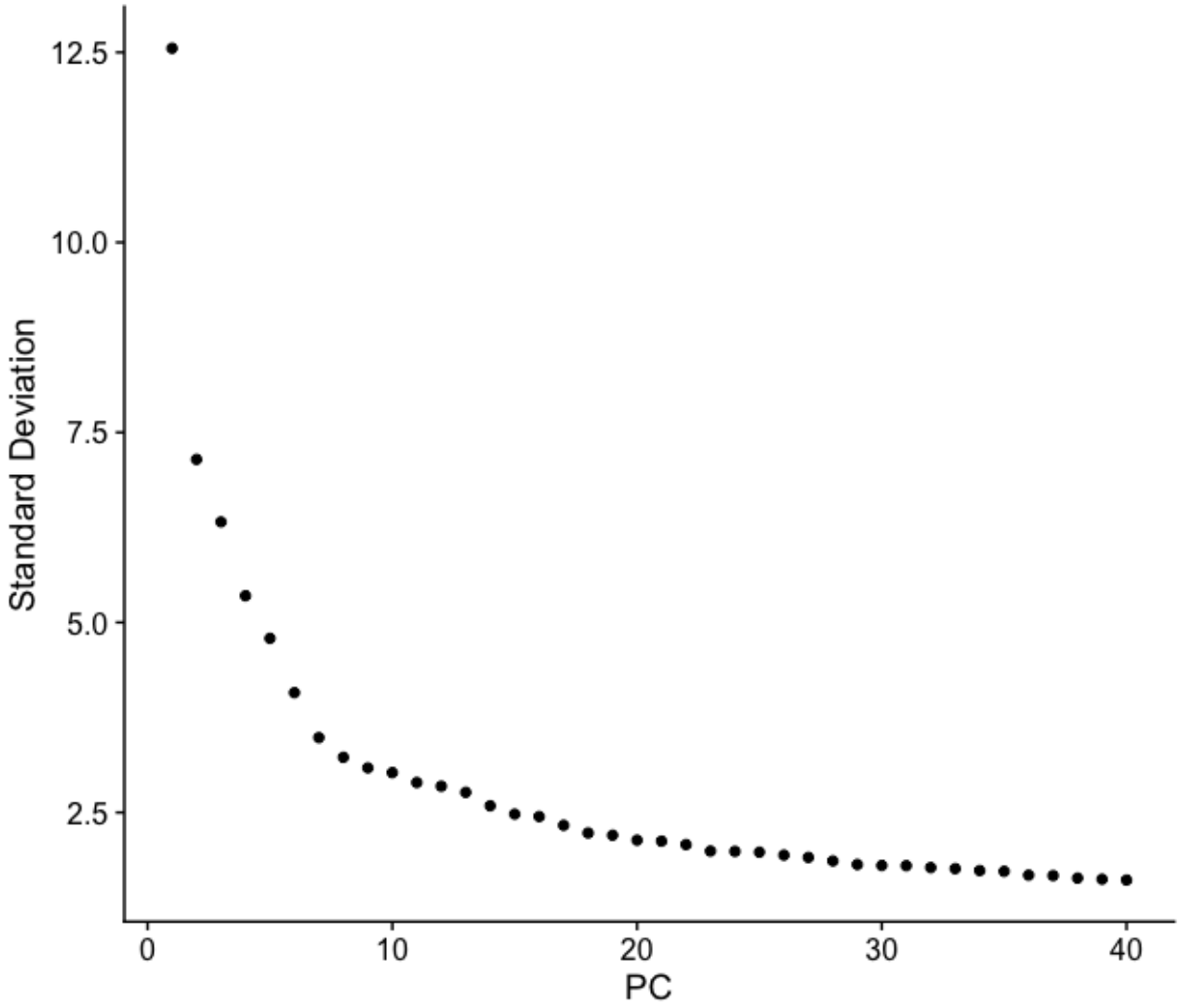


Figure 221: 엘보우 플롯

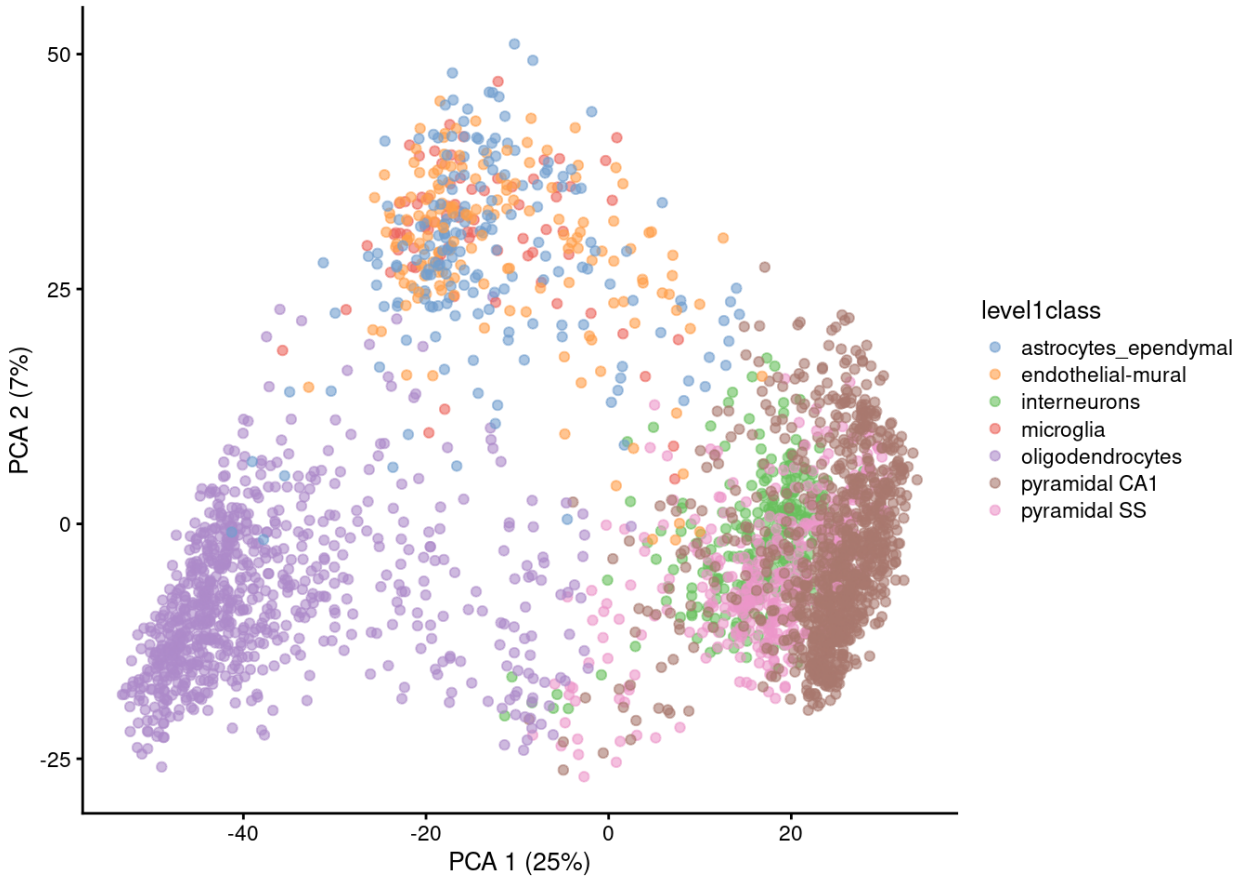


Figure 222: PCA 암 세포 유형

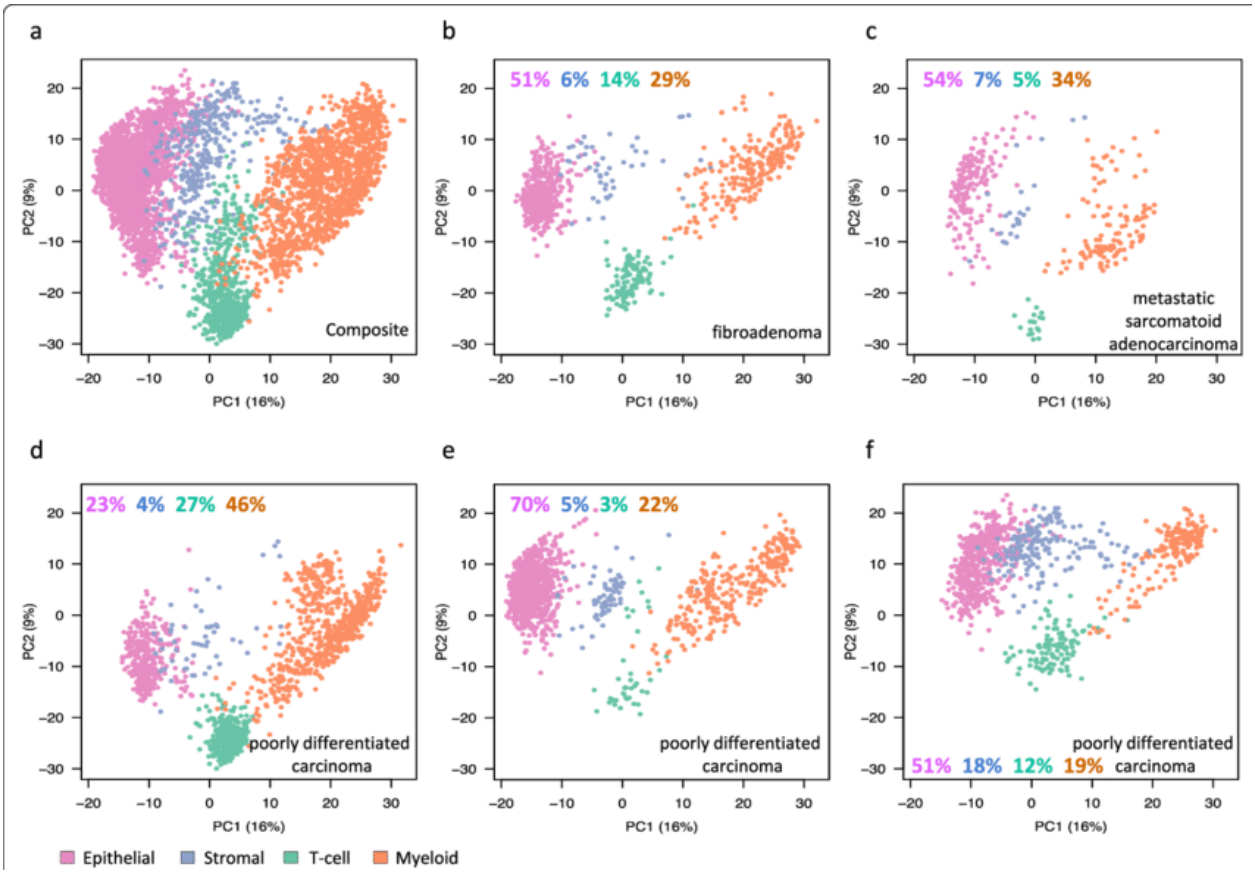


Figure 223: PCA 예시

12.3.1 확산 지도(Diffusion Maps)

가우시안 커널 k(x, y) = exp(-||x-y||^2 /)로 포인트 간 유사성을 계산하고, 그래프 라플라시안 L^( ) = D^{- } L D^{- }을 이용해 확산 과정을 모델링한다. 확산 좌표는 데이터의 연속적 변화를 잘 표현해 생물학적 궤적 분석에 자주 쓰이며, 여기서 파생된 확산 의사시간(Diffusion Pseudotime)은 세포 분화 순서를 추정하는 데 활용된다.

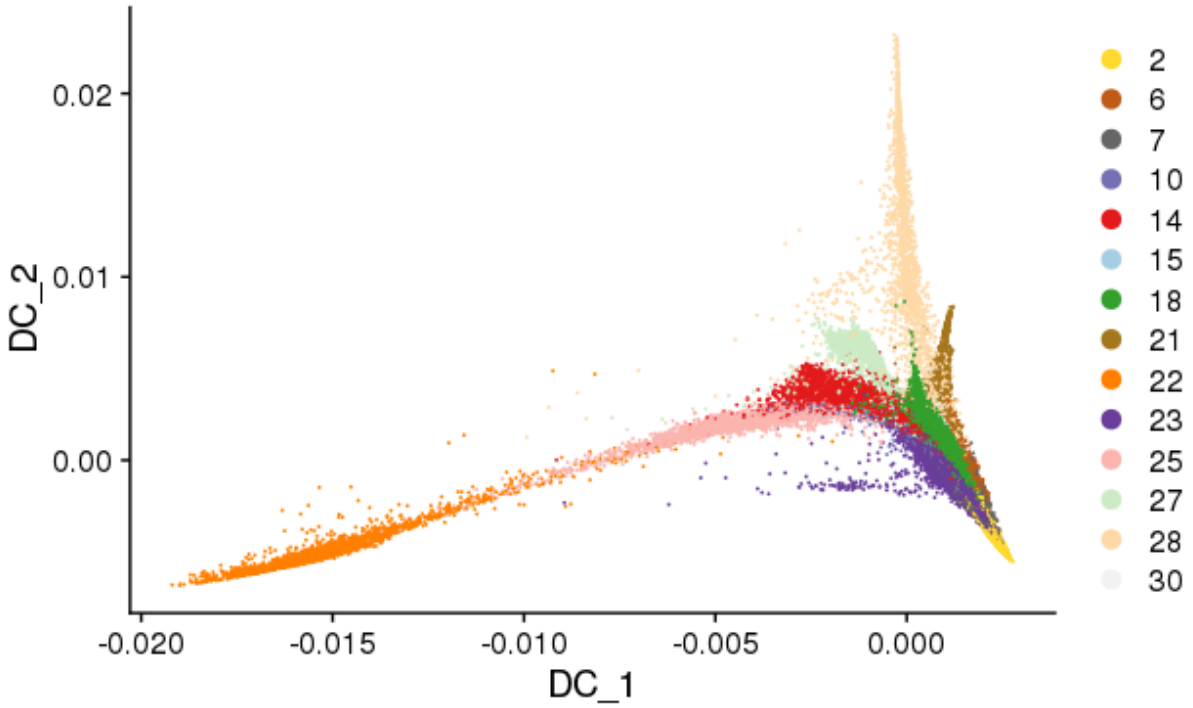


Figure 224: 확산 지도

Figure 11.7 확산 지도 – 확산 과정을 모델링하여 생물학적 궤적을 표현하는 비선형 차원 축소

12.3.2 t-SNE

t-SNE는 고차원에서 가까운 점이 저차원에서도 가깝도록 확률 분포를 맞춘다. 고차원에서는 가우시안, 저차원에서는 자유도가 낮은 t-분포를 사용해 KL 발산(Kullback-Leibler divergence)을 최소화한다.

KL 발산은 두 확률 분포 간의 차이를 측정하는 방법으로, 다음과 같이 정의된다:

KL(P||Q) = Σ P(i) log(P(i)/Q(i))

여기서 P는 고차원 공간의 확률 분포, Q는 저차원 공간의 확률 분포이다. t-SNE는 이 값을 최소화하여 두 공간에서의 확률 분포를 가능한 한 유사하게 만든다.

t-SNE는 지역 구조 시각화에는 뛰어나지만, 클러스터 간 거리나 크기는 해석이 어렵다.

Figure 11.8 t-SNE 알고리즘 – t-SNE의 고차원과 저차원 확률 분포 매칭 과정

12.3.3 UMAP

UMAP은 k-최근접 이웃(KNN)으로 퍼지 연결 그래프를 만든 뒤, 교차 엔트로피를 최소화하는 방식으로 저차원 임베딩을 학습한다.

C((A,), (A,)) = Σ[(a) log((a)/ (a)) + (1- (a)) log((1- (a))/(1- (a)))]

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for t=1 **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end

Figure 225: t-SNE 알고리즘

UMAP은 지역 구조와 전역 구조를 균형 있게 보존하고, 계산 속도와 재현성에서 장점이 있다.

Figure 11.9 UMAP 퍼지 그래프 – UMAP의 k-최근접 이웃 기반 퍼지 연결 그래프 구성

12.4 기법 비교

12.4.1 Swiss Roll 예제

3차원 나선형 구조인 Swiss Roll 데이터에 기법을 적용하면 차이가 분명하다. PCA는 나선을 펼치지 못하고, Diffusion Map은 연속적인 나선 구조를 보존한다. t-SNE는 클러스터는 잘 보이지만 전체 모양이 왜곡되고, UMAP은 지역과 전역 구조를 모두 고려해 균형 잡힌 결과를 제공한다.

Figure 11.10 Swiss Roll 데이터 – 3차원 나선형 구조를 가진 벤치마크 데이터셋

Figure 11.11 Swiss Roll PCA와 확산 지도 – 선형 및 확산 기반 차원 축소 결과 비교

Figure 11.12 Swiss Roll t-SNE와 UMAP – 비선형 차원 축소 기법들의 성능 비교

12.4.2 MNIST 손글씨

숫자 이미지를 차원 축소하면 각 숫자가 얼마나 잘 분리되는지 평가할 수 있다. t-SNE와 UMAP은 서로 다른 숫자를 잘 구분하지만, UMAP이 전역 구조까지 더 잘 유지하는 편이다.

Figure 11.13 MNIST 데이터 – 손글씨 숫자 이미지 데이터셋의 예시

Figure 11.14 MNIST PCA와 확산 지도 – 선형 및 확산 기반 차원 축소로 숫자 분류

Figure 11.15 MNIST t-SNE와 UMAP – 비선형 차원 축소로 숫자 클러스터링

12.5 클러스터링 기법

차원 축소 후에는 의미 있는 그룹을 찾기 위해 클러스터링을 적용한다.

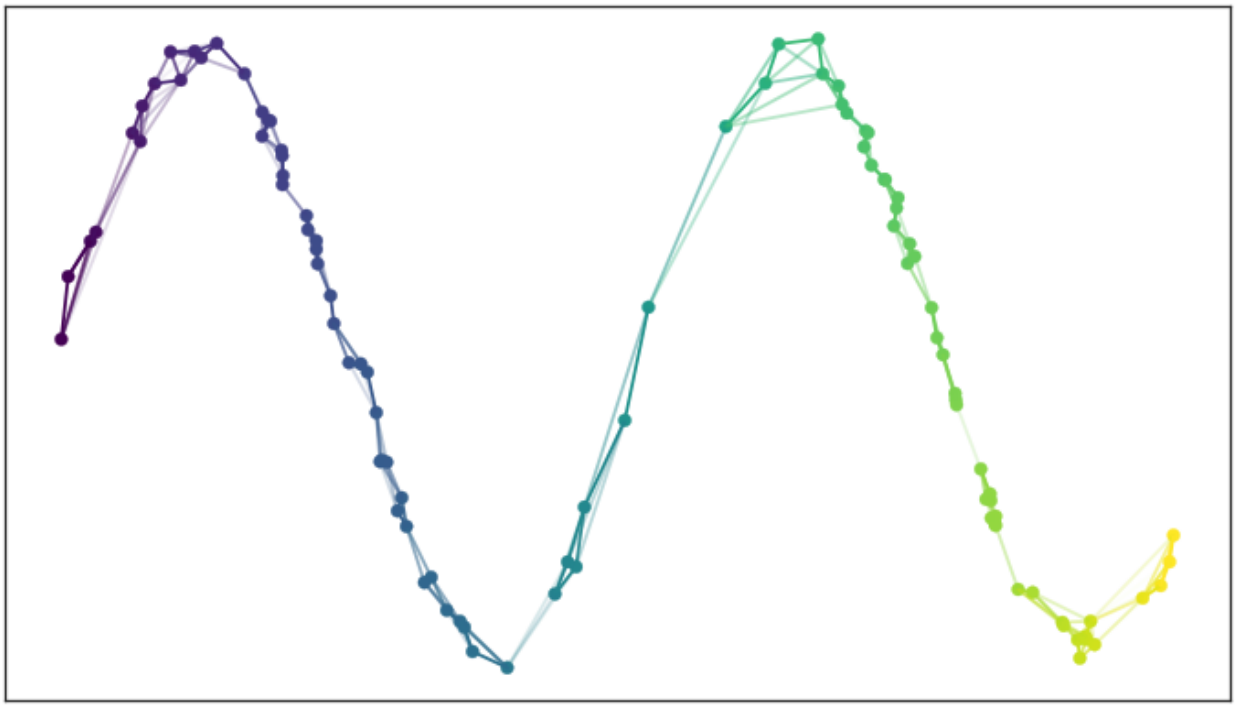


Figure 226: UMAP 퍼지 그래프

12.5.1 K-means

1. k개의 중심을 초기화한다.
2. 각 점을 가장 가까운 중심에 할당한다.
3. 각 클러스터의 평균으로 중심을 업데이트한다.
4. 변화가 없을 때까지 반복한다.

애니메이션으로 보면 중심이 움직이며 클러스터가 형성되는 과정을 직관적으로 이해할 수 있다.

Figure 11.16 K-means 애니메이션 – K-means 클러스터링 알고리즘의 반복적 최적화 과정

12.5.2 네트워크 기반 클러스터링

단일세포 데이터에서는 KNN 그래프를 만든 뒤 커뮤니티 탐지 알고리즘을 적용한다. Louvain은 모듈성을 최대화하는 방법이며, Leiden은 Louvain을 보완해 더 안정적이고 재현성 높은 결과를 제공하여 단일세포 분석에 널리 활용된다.

Figure 11.17 KNN 그래프 – k-최근접 이웃 그래프 구성과 네트워크 기반 클러스터링

Figure 11.18 Louvain 커뮤니티 탐지 – Louvain 알고리즘을 이용한 네트워크 커뮤니티 식별

Figure 11.19 Leiden 커뮤니티 탐지 – Leiden 알고리즘을 이용한 더 안정적인 커뮤니티 발견

Figure 11.20 Leiden 다양한 해상도 – 해상도 매개변수에 따른 Leiden 클러스터링 결과 변화

12.6 주요 도구

- **Monocle3**: 단일세포 궤적과 pseudotime 분석을 위한 R 패키지. (https://cole-trapnell-lab.github.io/monocle3/)
- **Destiny**: 확산 지도 분석을 지원하는 R 패키지. (https://academic.oup.com/bioinformatics/article/32/8/1241/1744143)

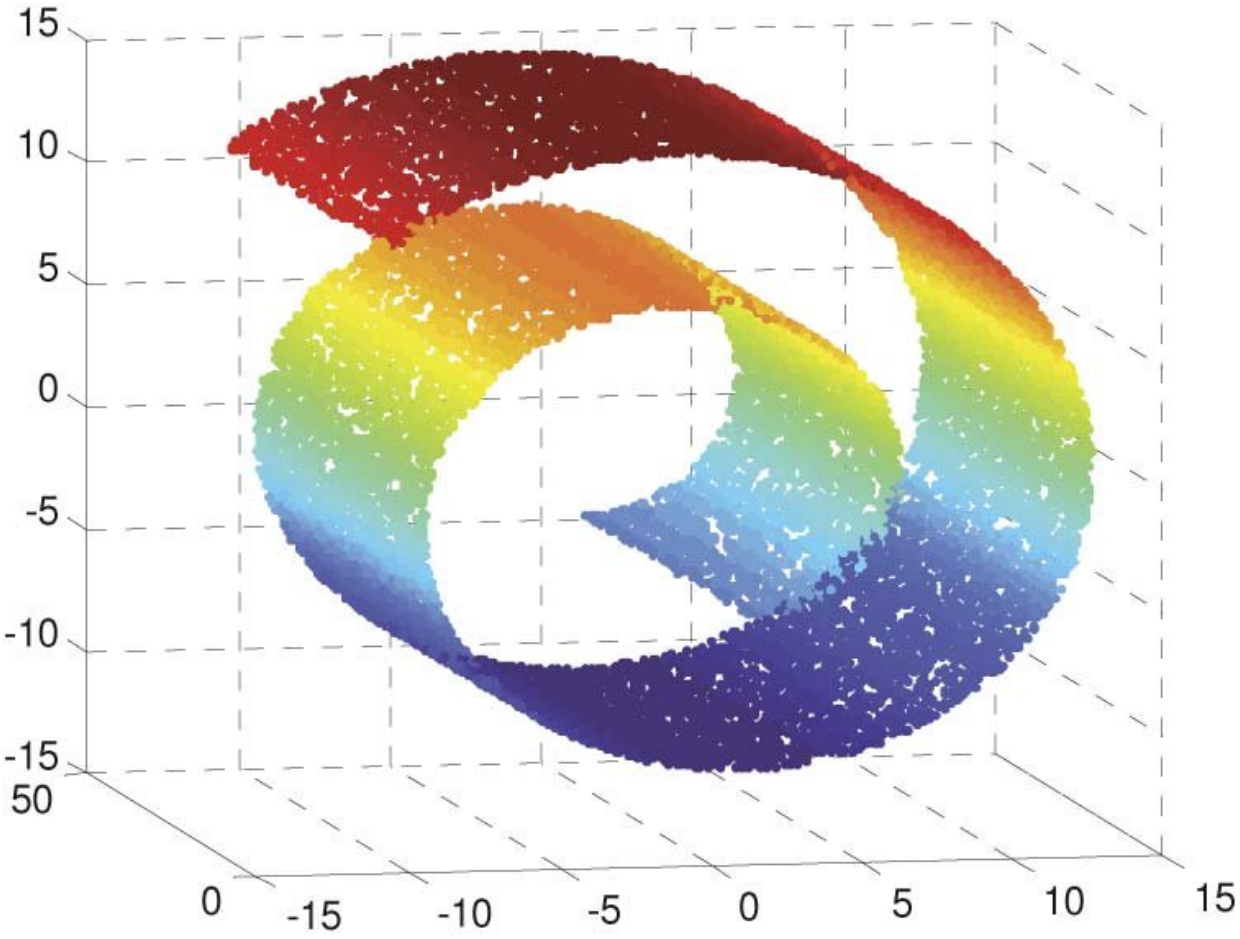


Figure 227: Swiss Roll 데이터

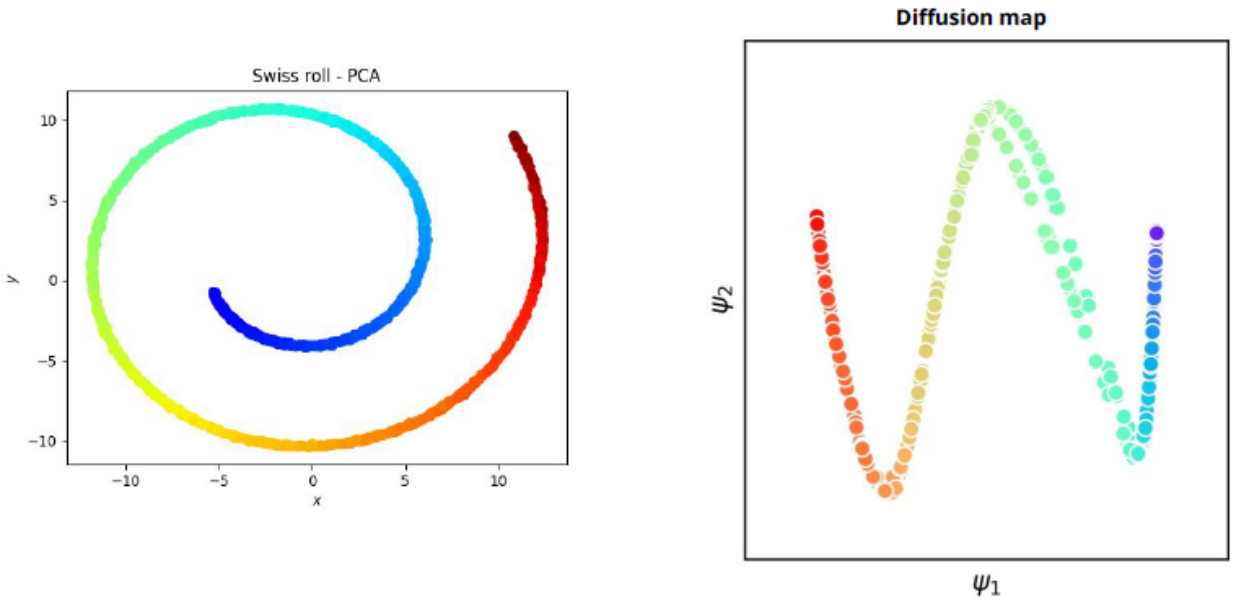


Figure 228: Swiss Roll PCA와 확산 지도

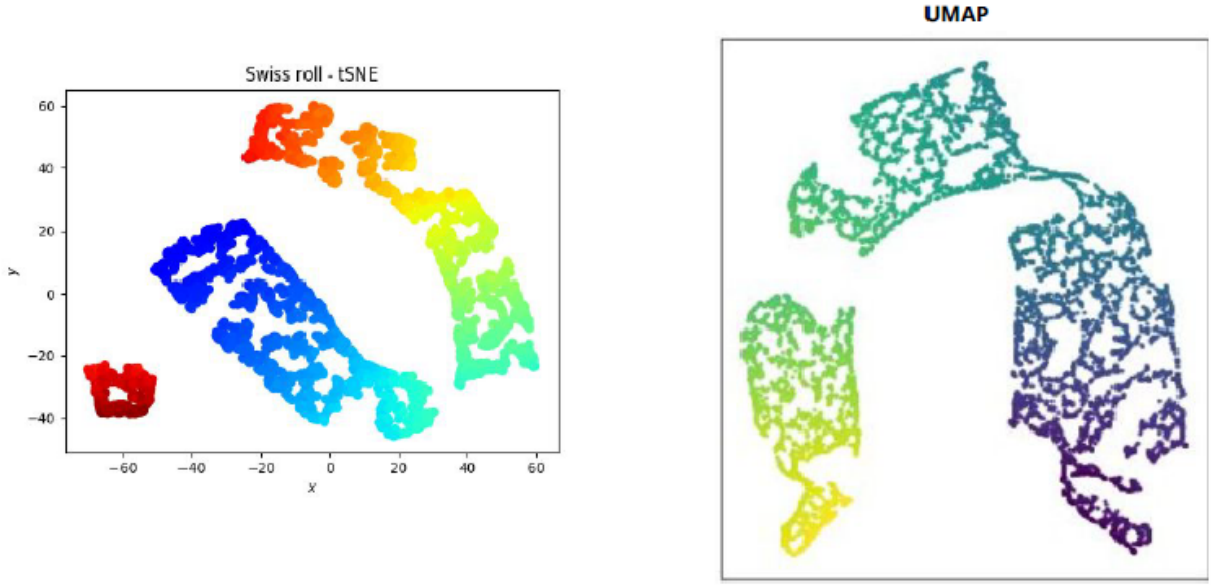


Figure 229: Swiss Roll t-SNE와 UMAP

Figure 230: MNIST 데이터

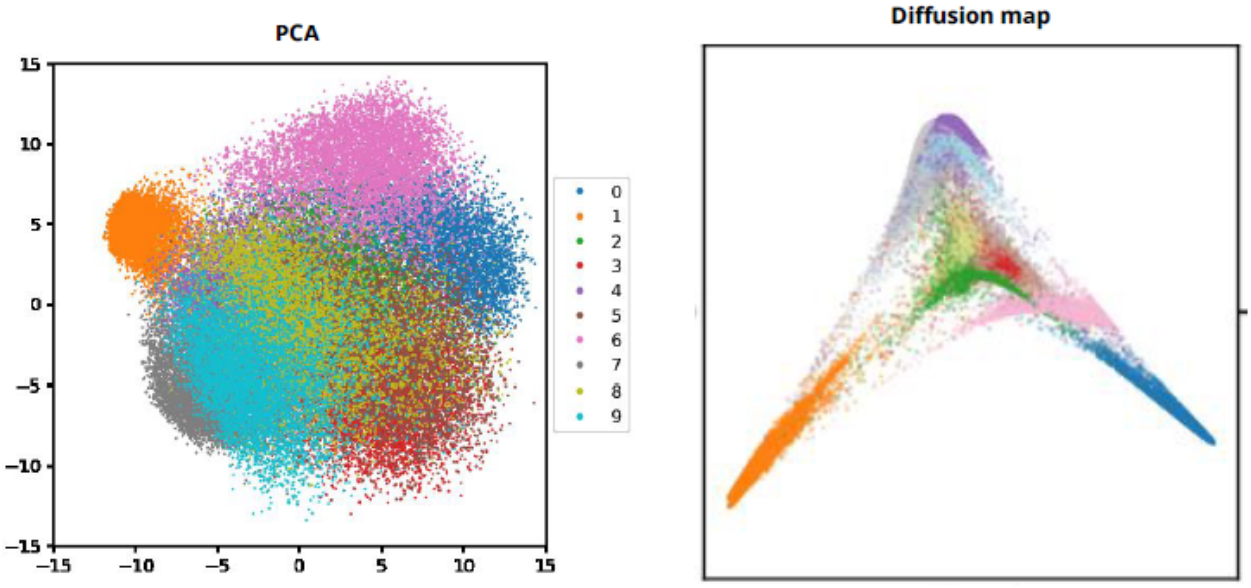


Figure 231: MNIST PCA와 확산 지도

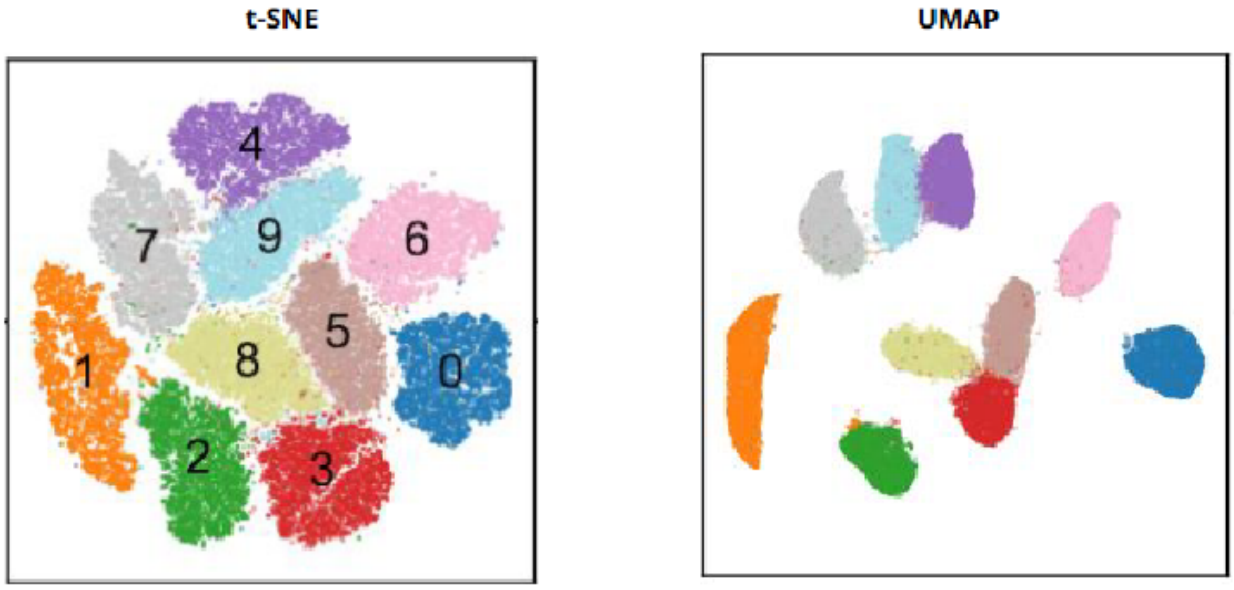


Figure 232: MNIST t-SNE와 UMAP

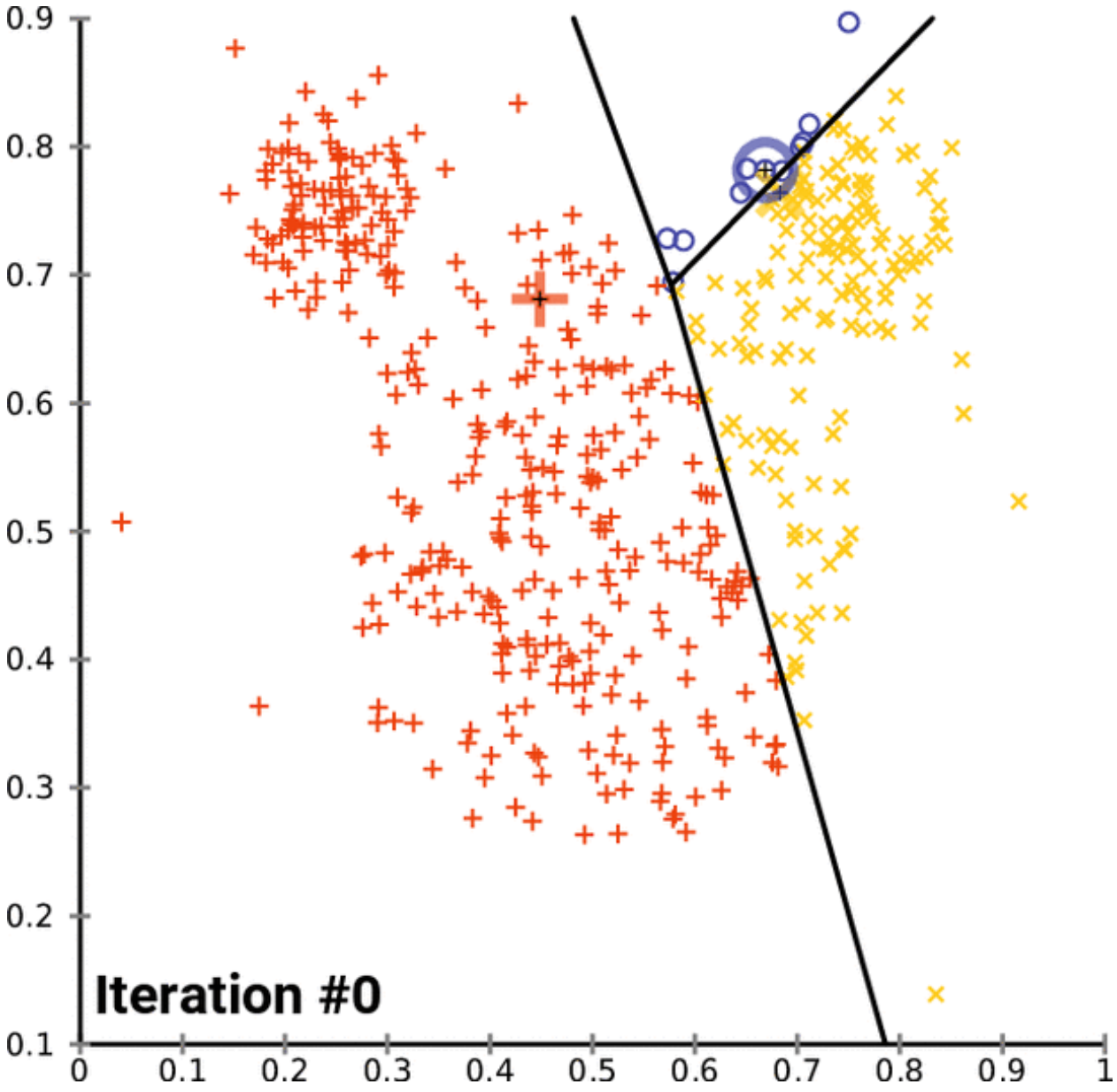


Figure 233: K-means 애니메이션

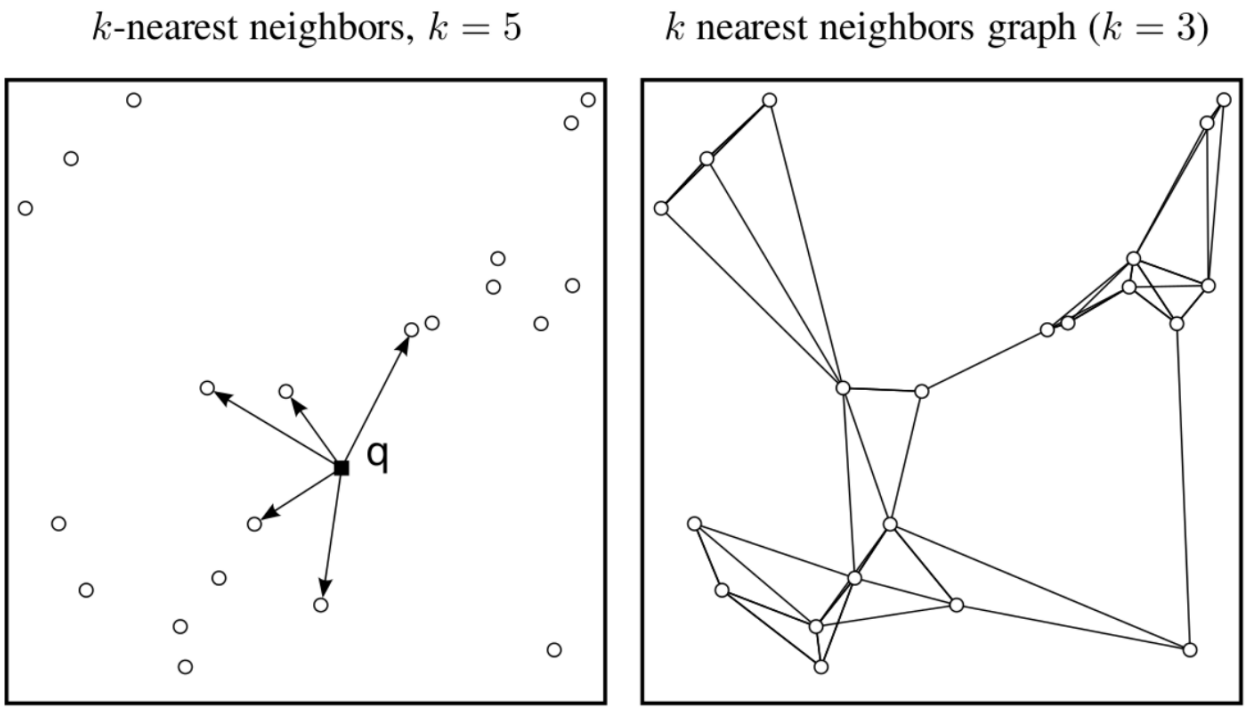


Figure 234: KNN 그래프

12.7 해석 시 주의점

12.7.1 t-SNE

- 클러스터 간 거리는 의미가 없을 수 있다.
- 클러스터 크기는 실제 비율을 반영하지 않는다.
- Perplexity 등 하이퍼파라미터에 민감하다.

특히 Perplexity 파라미터를 잘 조절하는 것이 중요하다. Distill의 "How to Use t-SNE Effectively"(https://distill.pub/2016/misread tsne/)를 참고할 만하다.

12.7.2 UMAP 매개변수

n_neighbors, min_dist 같은 매개변수에 따라 결과가 달라질 수 있다. 데이터 특성과 분석 목표에 맞게 값을 조정하고, 여러 설정을 비교해 보는 것이 좋다. 원 논문은 https://arxiv.org/pdf/1802.03426.pdf 에서 확인할 수 있다.

12.7.3 2차원 시각화의 한계와 왜곡의 필연성

고차원 데이터를 2차원으로 축소하는 과정에서는 정보 손실과 왜곡이 필연적으로 발생한다. 이는 수학적으로 고차원 공간의 복잡한 구조를 저차원 공간에 완벽하게 보존하는 것이 불가능하기 때문이다. 특히 수천 개의 유전자 발현 데이터를 x, y 좌표만으로 표현할 때 고차원에서 멀리 떨어진 두 점이 2차원에서는 가깝게 보이거나, 반대로 가까운 점들이 멀어져 보일 수 있다.

따라서 **t-SNE나 UMAP의 2차원 시각화 결과만으로는 정량적 분석을 수행해서는 안 된다.** 이러한 시각화는 데이터의 전반적인 패턴을 탐색하고 가설을 세우는 용도로만 사용되어야 하며, 실제 분석과 결론 도출은 원본 고차원 데이터나 적절한 차원(예: PCA 상위 50개 성분)에서 수행되어야 한다. 시각화는 "보여주기 위한" 도구이지 "분석하기 위한" 도구가 아님을 명심해야 한다.

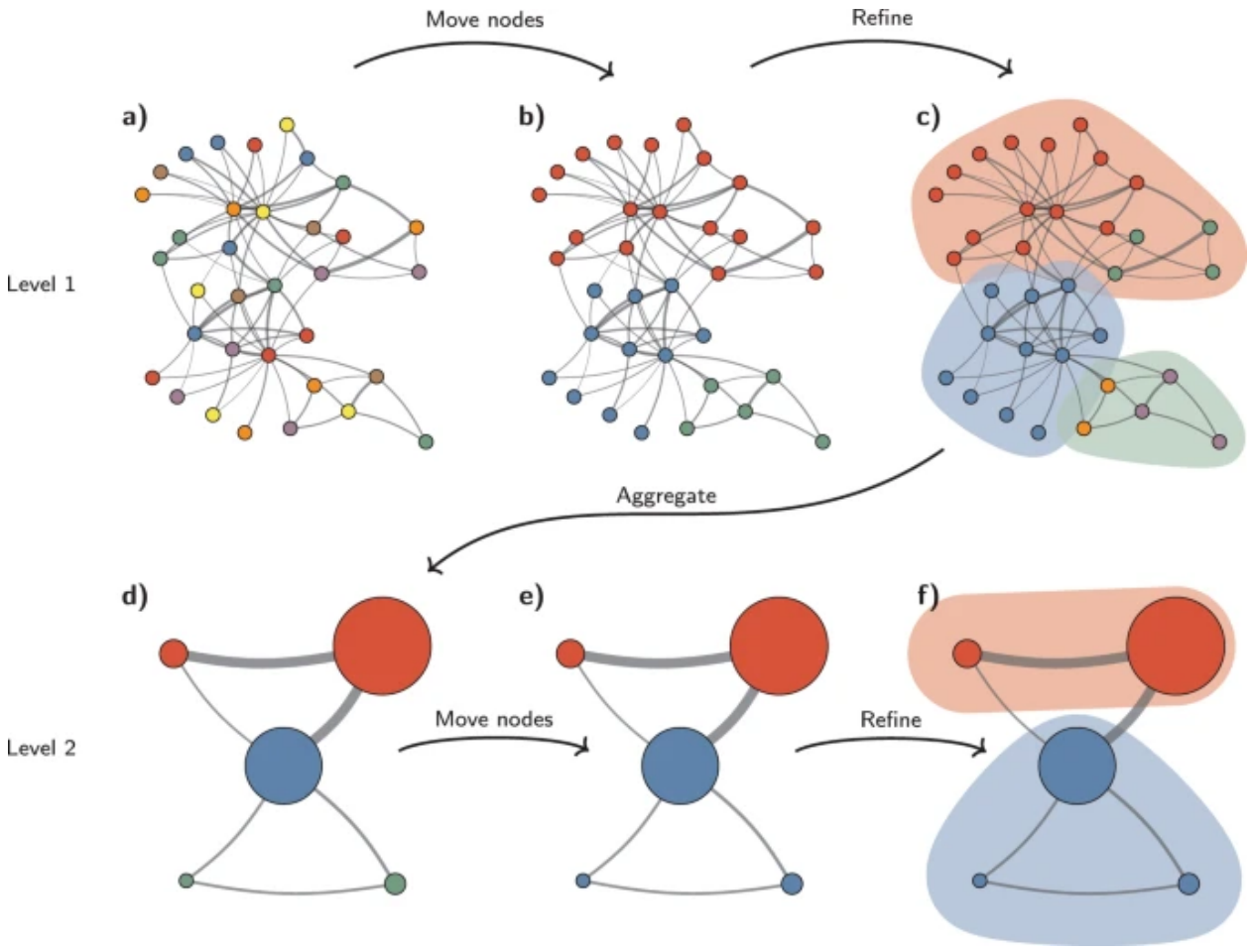


Figure 235: Louvain 커뮤니티 탐지

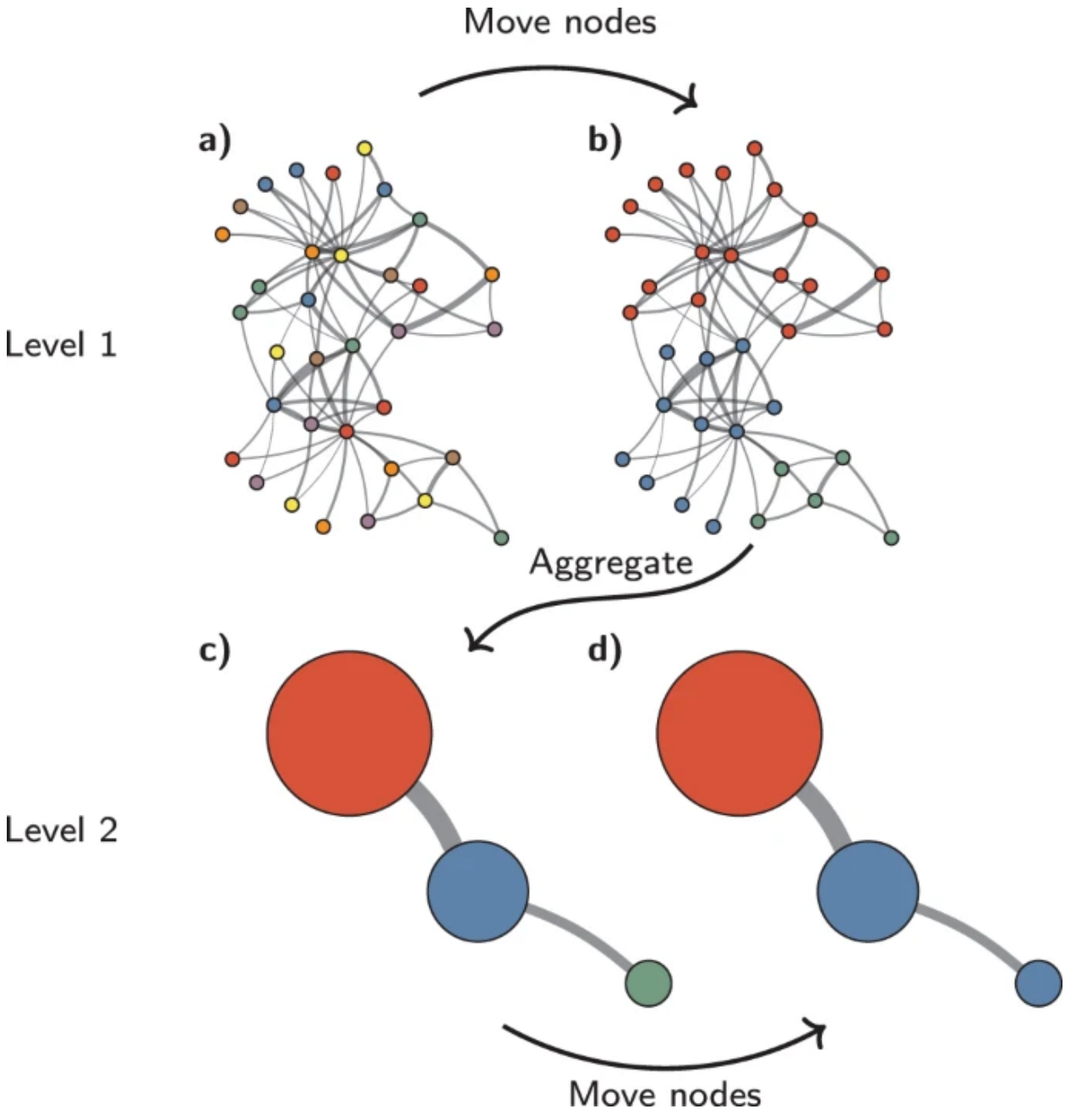


Figure 236: Leiden 커뮤니티 탐지

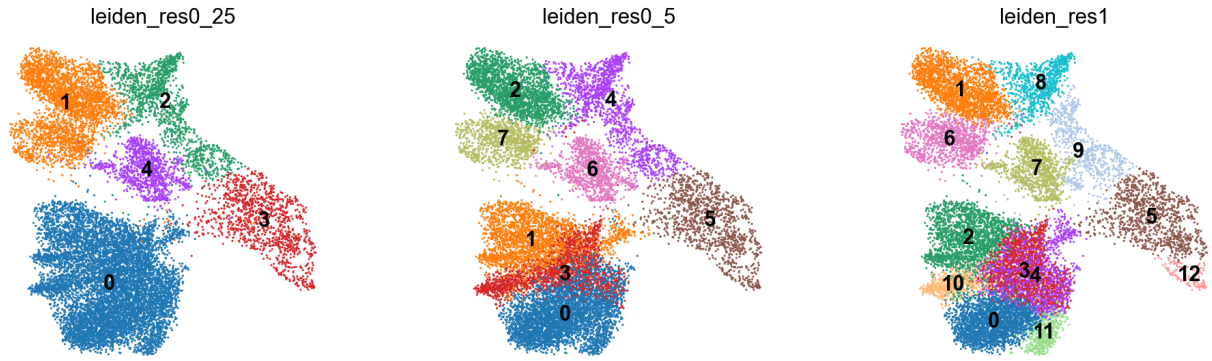


Figure 237: Leiden 다양한 해상도

12.8 결론

차원 축소는 현대 의생명정보학에서 고차원 생물학 데이터를 이해하고 해석하는 데 없어서는 안 될 핵심 도구로, PCA와 같은 선형 기법은 계산이 빠르고 해석이 용이하며 유전자 발현 데이터의 공분산 구조를 효과적으로 활용하는 반면, t-SNE, UMAP, 확산 지도와 같은 비선형 기법은 복잡한 매니폴드 구조를 발견하고 국소적 패턴을 보존하는 데 뛰어나다. 특히 단일세포 RNA 시퀀싱 데이터 분석에서는 세포 유형을 정의하는 소수의 중요 유전자들과 동일 경로 유전자들 간의 높은 상관성을 고려할 때 적절한 차원 축소가 생물학적 신호를 노이즈로부터 분리하는 데 필수적이며, 실제 분석에서는 PCA로 초기 노이즈 제거와 계산 효율성을 확보한 후 t-SNE나 UMAP으로 세밀한 구조를 탐색하는 통합적 접근이 권장된다.

적절한 기법과 매개변수를 신중히 선택하고 결과를 생물학적 맥락에서 올바르게 해석할 때 차원 축소는 정밀의학과 개인맞춤 치료를 향한 깊이 있는 통찰을 제공하여 복잡한 생명 현상을 이해하고 질병 메커니즘을 밝히는 데 중요한 기여를 할 것으로 기대된다.

13 조직 병리학과 공간체학

13.1 공간 전사체학과 조직 병리학의 통합

최근 10년간 공간체학(spatial omics) 기술의 급속한 발전은 전통적인 조직 병리학의 패러다임을 근본적으로 변화시키고 있다. 기존의 형태학적 관찰에 국한되었던 병리학적 진단이 이제는 조직 내 각 세포의 위치별 분자적 특성을 정밀하게 분석할 수 있게 되었다. 공간 전사체학, 공간 단백체학, 공간 대사체학 등의 기술들이 통합되면서 조직의 3차원적 분자 지도 작성이 가능해졌고, 이는 질병 메커니즘 이해와 정밀 진단에 새로운 가능성을 열어주고 있다. 본 장에서는 이러한 기술 혁신이 조직 병리학에 미치는 영향과 최신 공간 분석 기술들을 다룬다.

13.1.1 조직학적 염색법의 원리

전통적인 조직병리학에서는 조직의 형태학적 특성을 관찰하기 위해 다양한 염색 기법을 사용한다. 가장 기본적인 헤마톡실린-에오신(Hematoxylin & Eosin, H&E) 염색에서 헤마톡실린은 양전하를 띠어 음전하를 가진 핵산(DNA, RNA)에 결합하여 세포핵을 파란색-보라색으로 염색한다. 반면 에오신은 음전하를 띠어 양전하를 가진 세포질 단백질에 결합하여 분홍색-빨간색으로 염색한다.

세포질의 염색 양상은 세포의 기능 상태를 반영한다. 거친 소포체(rough endoplasmic reticulum)가 발달한 세포는 많은 양의 리보솜 RNA 때문에 호염기성(basophilic)을 보이며, 단백질 합성이 활발한 세포는 호산성(eosinophilic) 염색 양상을 보인다. 분화가 덜 된 세포나 소기관이 적은 세포는 무색소성(amphophilic) 또는 양친성을 나타낸다.

Figure 12.1 H&E 염색 – 헤마톡실린-에오신 염색의 기본 원리와 세포 구조 시각화

Figure 12.2 전형적인 H&E 염색 – 조직 내 다양한 세포 유형의 H&E 염색 양상

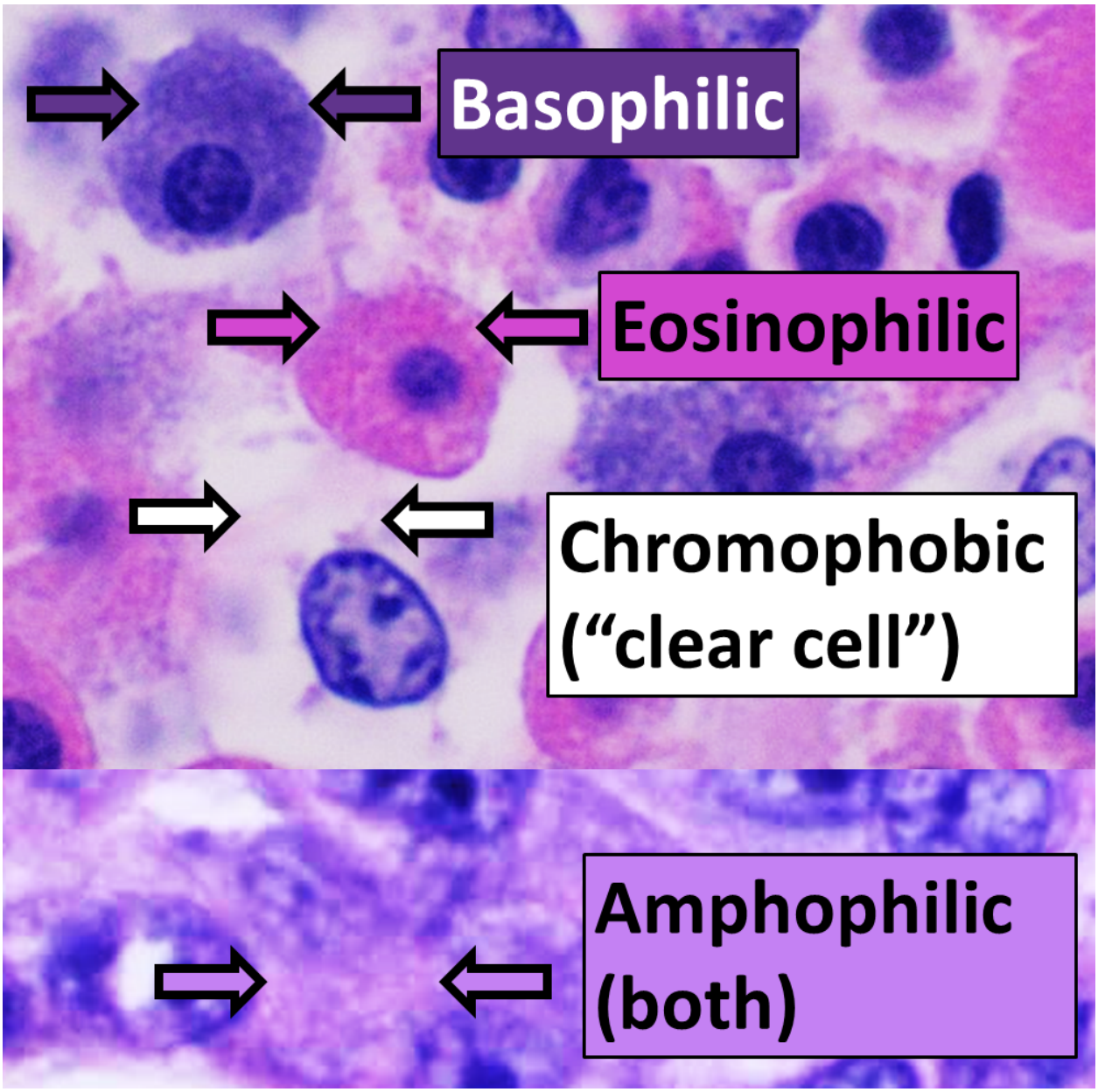


Figure 238: H&E 염색

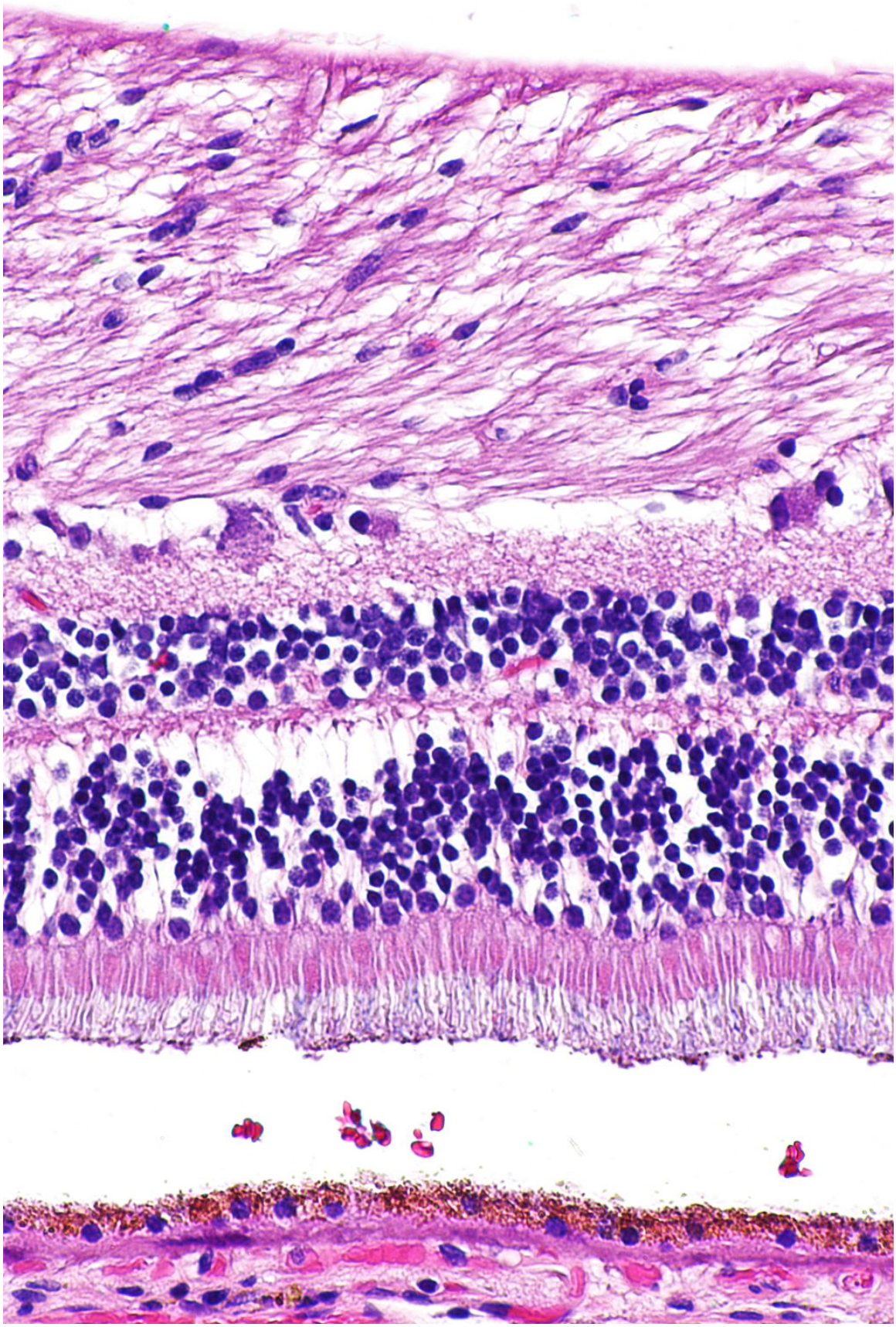


Figure 239: 전형적인 H&E 염색

13.1.2 형광 이미징과 다중 표지

1장에서 살펴본 이미징 방법들 중 형광 현미경법은 특정 분자를 표적으로 하는 정밀한 분석을 가능하게 한다.

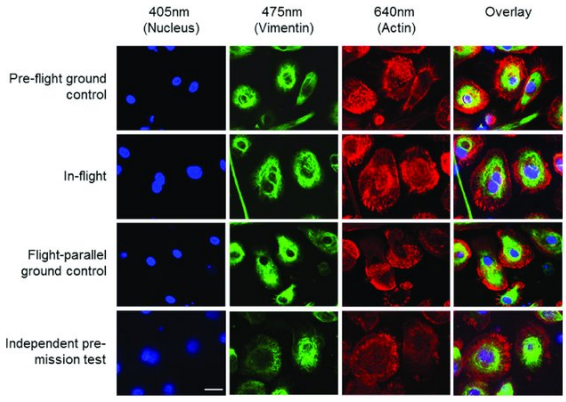


Figure 240: 형광 이미징

Figure 12.3 형광 이미징 – 다중 형광 표지를 이용한 조직 내 다양한 분자 시각화

DAPI(4',6-diamidino-2-phenylindole)는 DNA의 AT-rich 부위에 결합하여 자외선을 받으면 파란색 형광을 방출한다. 이를 통해 세포핵의 위치와 형태를 정확히 파악할 수 있으며, 세포 분열 단계나 세포 사멸 여부도 판단할 수 있다.

Figure 12.4 DAPI 염색 – DNA 결합 형광 염료를 이용한 세포핵 시각화

Figure 12.5 형광 이미지에서의 DAPI – 파란색 DAPI 신호로 표시된 세포핵의 분포

다중 형광 표지(multiplexed fluorescence labeling)를 통해 여러 단백질이나 RNA를 동시에 시각화할 수 있다. 각기 다른 자극/방출 파장을 가진 형광 염료들을 조합하면, 한 조직 절편에서 다양한 세포 유형이나 분자적 상태를 구별할 수 있다. 공초점 현미경이나 초해상도 현미경을 사용하면 세포내 소기관 수준까지 정밀한 위치 정보를 얻을 수 있다.

13.1.3 공간 전사체학의 임상적 의의

공간 전사체학(spatial transcriptomics)은 조직 내에서 각 세포의 위치 정보를 보존하면서 전사체 분석을 수행하는 기술이다. 이 기술을 통해 종양의 이질성, 종양-기질 상호작용, 면역 미세환경의 공간적 분포 등을 정밀하게 분석할 수 있다.

예를 들어, 종양 조직에서는 종양 중심부, 침윤 경계, 주변 기질 등 위치에 따라 서로 다른 유전자 발현 패턴을 보인다. 이러한 공간적 이질성은 종양의 진행, 전이, 치료 반응성과 밀접한 관련이 있다. 또한 면역세포의 공간적 분포와 활성화 상태를 분석하면 면역치료의 효과를 예측하거나 저항성 메커니즘을 이해할 수 있다.

형태학적 변화와 분자적 변화를 통합적으로 분석하면 기존의 조직병리학적 진단을 한 단계 발전시킬 수 있다. 예를 들어, 형태학적으로는 유사해 보이는 종양이라도 분자적 프로파일이 다르면 예후나 치료 반응성이 달라질 수 있다. 이러한 정밀 진단은 개인맞춤 치료의 기초가 된다.

Figure 12.6 공간 전사체학 종합 이미지 – 다양한 공간 전사체학 기술들의 통합적 개념도

13.2 공간전사체학 기술의 발전과 응용

13.2.1 공간전사체학의 기본 원리

공간 바코딩 시스템: 공간전사체학의 핵심은 조직의 각 위치를 고유하게 식별할 수 있는 공간 바코드(spatial barcode) 시스템이다. 초기 방법에서는 유리 슬라이드 위에 poly-T 올리고뉴클레오타이드 프로브를 격자 형태로 배열하고, 각 위치마다 고유한 바코드 서열을 부착했다. 조직 절편을 이 슬라이드 위에 올리면 mRNA가 해당 위치의 프로브에 포획되고, 이후 시퀀싱을 통해 각 위치에서의 유전자 발현을 측정할 수 있다.

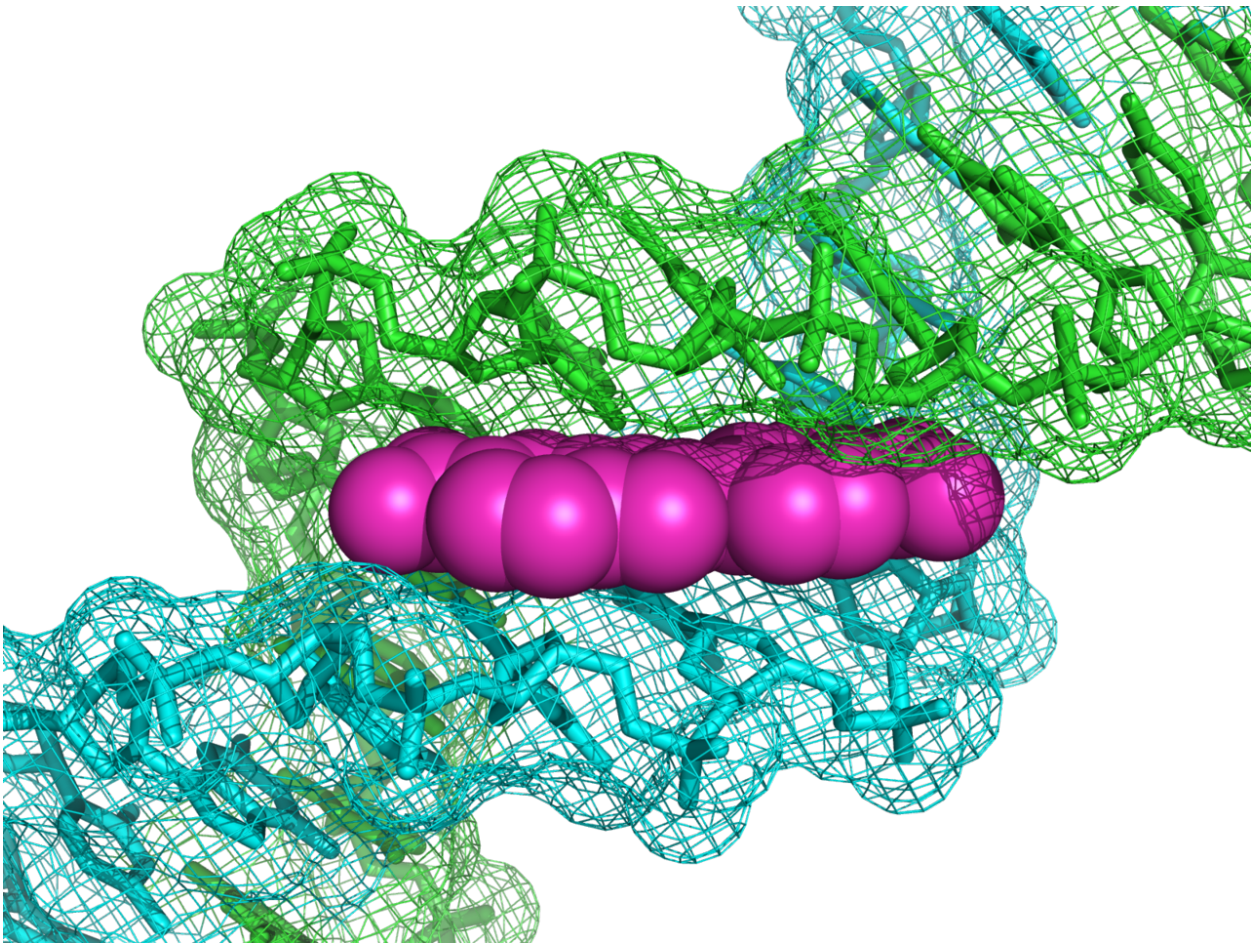


Figure 241: DAPI 염색

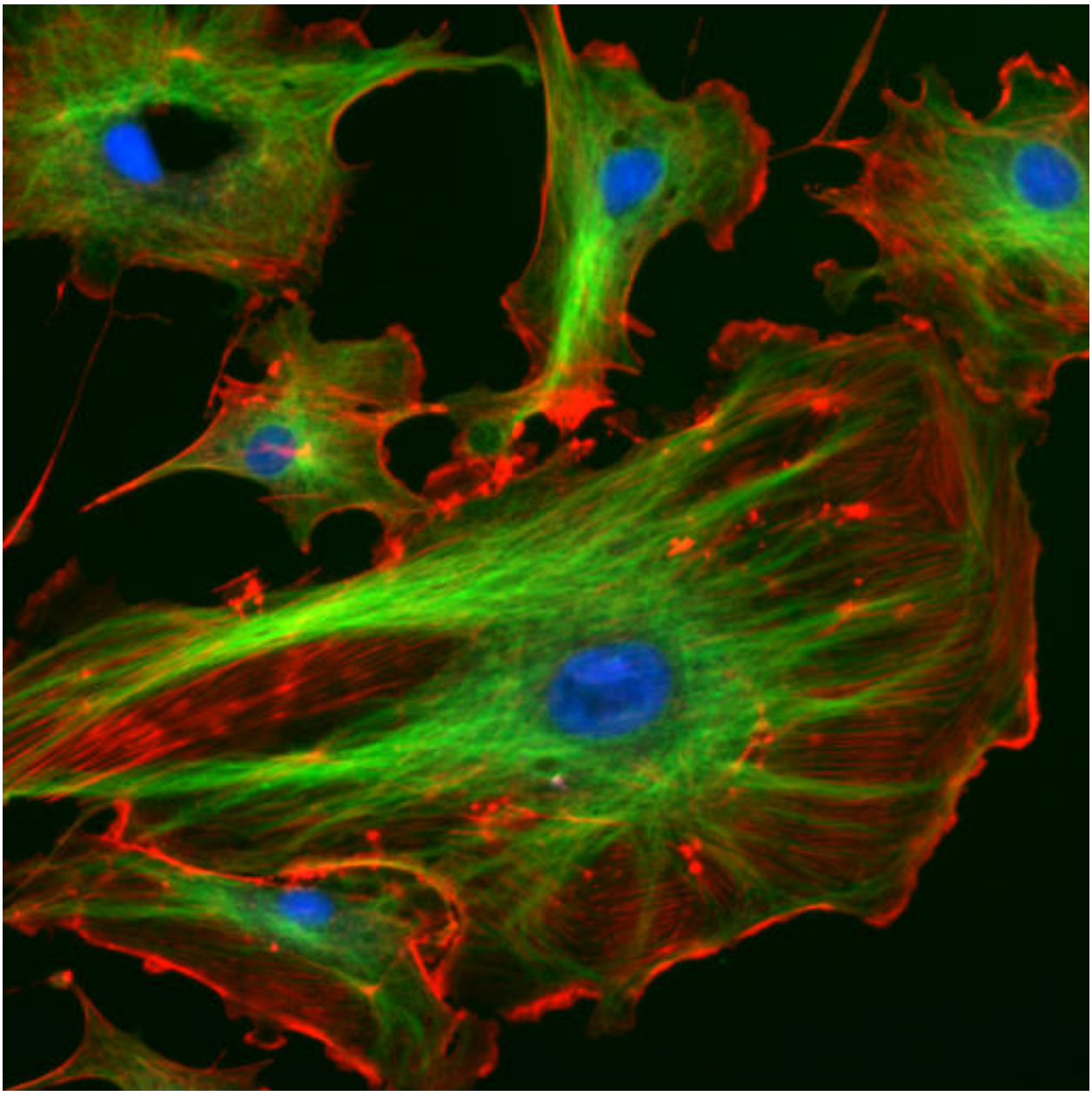


Figure 242: 형광 이미지에서의 DAPI

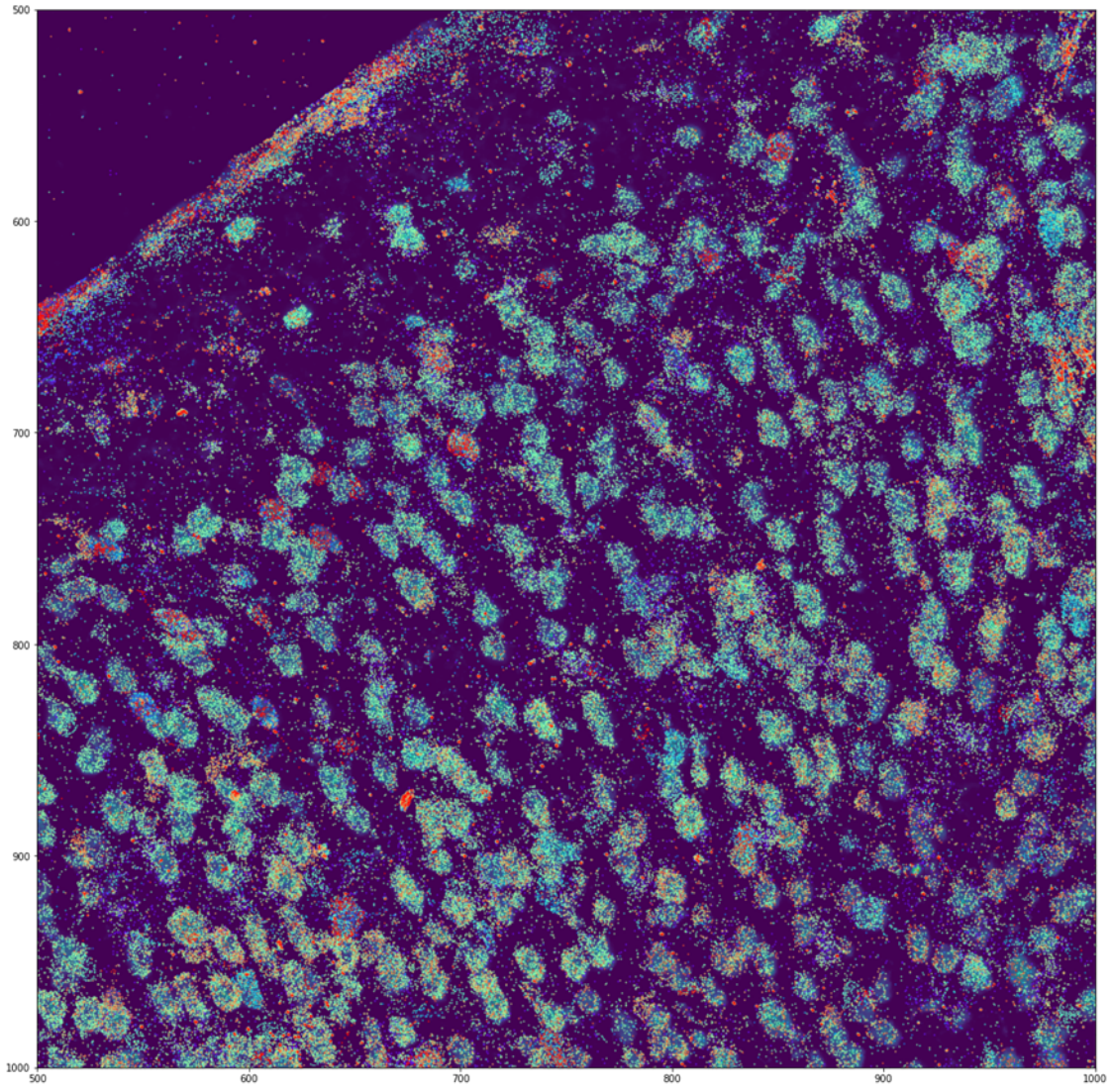


Figure 243: 공간 전사체학 종합 이미지

이 과정에서 각 프로브에는 공간 바코드와 함께 UMI(Unique Molecular Identifier)도 포함되어 있어, 개별 mRNA 분자를 정확히 계수할 수 있다. 이를 통해 조직의 형태학적 구조와 분자적 정보를 동시에 얻을 수 있어, 기존의 조직학적 분석과 분자생물학적 분석을 통합할 수 있다.

13.2.2 Visium 기술과 발전 방향

Visium 플랫폼의 특성: 10x Genomics에서 개발한 Visium은 현재 가장 널리 사용되는 상업적 공간전사체학 플랫폼이다. Visium은 6.5mm × 6.5mm 영역에 약 5,000개의 스팟을 배열하며, 각 스팟의 직경은 55μm이다. 이는 일반적인 세포 크기(10–20μm)보다 크므로, 하나의 스팟에는 여러 개의 세포가 포함될 수 있다.

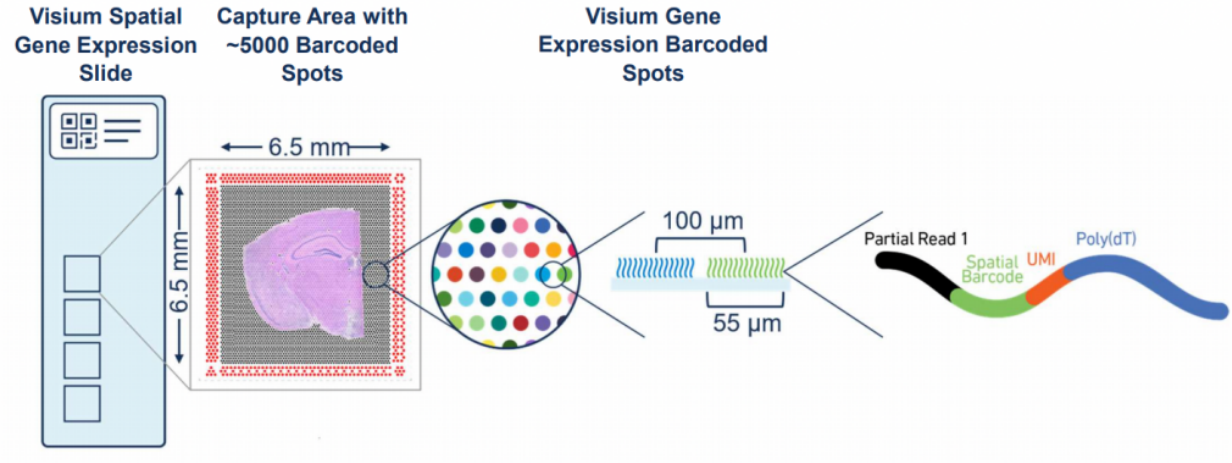


Figure 244: Visium 기술

Figure 12.7 Visium 기술 – 10x Genomics Visium 플랫폼의 공간 바코딩 시스템

이러한 해상도 제한은 스팟 디콘볼루션(spot deconvolution)이라는 추가적인 분석 과정을 필요로 한다. 스팟 디콘볼루션은 각 스팟에 포함된 서로 다른 세포 유형의 비율을 추정하는 과정으로, 단일세포 참조 데이터와 computational 방법을 결합하여 수행된다.

Figure 12.8 Visium 디콘볼루션 필요성 – 스훙 크기와 세포 크기 차이로 인한 디콘볼루션의 필요성

고해상도 기술 발전: Visium HD는 기존 Visium의 해상도를 크게 개선한 차세대 기술로, 더 작은 스팟 크기를 통해 단일세포에 가까운 해상도를 제공한다. 이를 통해 세포 경계를 더 정확히 구분할 수 있고, 세포 간 상호작용을 더 정밀하게 분석할 수 있다.

Figure 12.9 Visium HD – 고해상도 공간 전사체학을 위한 차세대 Visium 기술

Figure 12.10 Visium과 Visium HD 비교 – 기존 Visium과 고해상도 Visium HD 기술의 성능 차이

STOmics(Stereo-seq) 기술은 DNA 나노볼(DNB) 패턴 칩을 이용하여 서브 마이크로미터 해상도까지 달성할 수 있는 기술이다. 이는 세포보다 작은 픽셀 단위로 유전자 발현을 측정할 수 있어, 세포 내부의 공간적 구조까지 분석할 수 있는 가능성을 제시한다.

Figure 12.11 STOmics Stereo-seq – 서브 마이크로미터 해상도를 제공하는 STOmics 공간 전사체학 기술

13.2.3 형광 기반 공간 분석 기술

형광의 물리적 원리: 형광 현상은 분자가 특정 파장의 빛을 흡수하여 전자가 들뜬 상태로 이동한 후, 바닥 상태로 돌아가면서 더 긴 파장의 빛(형광)을 방출하는 현상이다. 흡수 파장과 방출 파장 간의 차이를 스토크스 이동(Stokes shift)이라고 하며, 이 원리를 이용하여 서로 다른 분자들을 구별할 수 있다.

Figure 12.12 형광 원리 – 분자가 빛을 흡수하고 형광을 방출하는 물리적 과정

Figure 12.13 스토크스 이동 – 흡수 파장과 방출 파장의 차이를 보여주는 스토크스 이동 현상

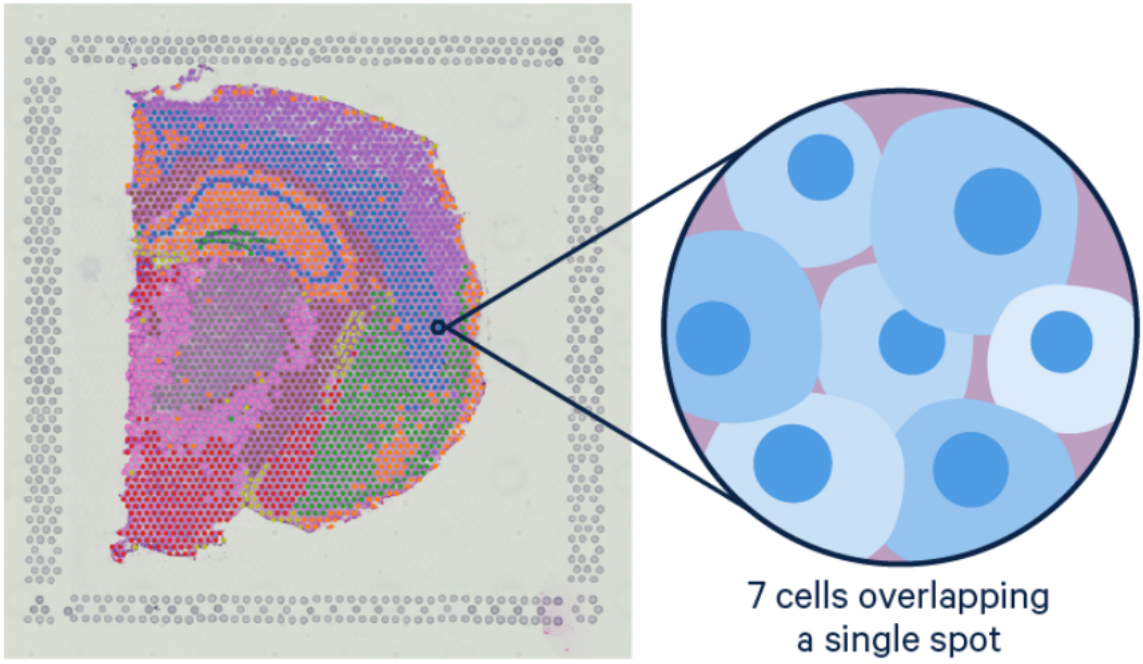


Figure 245: Visium 디콘볼루션 필요성

단일분자 형광 제자리 혼성화(smFISH): smFISH는 특정 mRNA 서열에 상보적인 단일가닥 DNA 올리고뉴클레오타이드 프로브를 형광 염료로 표지하여 조직 내에서 개별 mRNA 분자를 시각화하는 기술이다. 일반적으로 하나의 mRNA에 대해 20–50개의 짧은 프로브를 설계하여 신호를 증폭시킨다.

Figure 12.14 FISH 개요 – 형광 제자리 혼성화 기술의 전체적인 실험 과정

Figure 12.15 FISH 이미징 – 형광 제자리 혼성화를 이용한 세포 내 RNA 지역화 시각화

Figure 12.16 회절 한계 – 공간 이미징에서 발생하는 광학적 회절 한계와 해상도 제약

Figure 12.17 이미징 기술들 – 다양한 공간 이미징 및 바이오이미징 기술의 비교

실험 과정에서는 조직을 포름알데하이드나 메탄올로 고정한 후, 특이적 프로브를 이용한 혼성화 반응을 수행한다. 핵은 DAPI로 염색하여 세포의 위치를 확인할 수 있다. 이 기술의 장점은 단일 분자 해상도를 제공하며, 정량적 분석이 가능하다는 것이다.

또한 정확한 RNA의 위치 파악을 위해 신호 증폭 기술이 개발되었으며, 그 대표적인 예시 중 하나는 RNAscope를 들 수 있다. RNAscope은 분지형 DNA 증폭 시스템을 사용하여 단일 mRNA 분자도 검출할 수 있을 정도로 민감도를 높였다.

Figure 12.18 RNAscope – 분지형 DNA 증폭을 이용한 고감도 RNA 검출 기술

고도 다중화 형광 기술: MERFISH(Multiplexed Error-Robust Fluorescence In Situ Hybridization)와 SeqFISH는 여러 라운드의 이미징과 바코딩을 결합하여 한 번의 실험에서 수백에서 수천 개의 유전자를 동시에 분석할 수 있다.

Figure 12.19 MERFISH 증폭 – MERFISH 기술에서 사용되는 신호 증폭 시스템

Figure 12.20 MERFISH 조합 바코딩 – 다중 유전자 분석을 위한 조합적 바코딩 시스템

Figure 12.21 SeqFISH – 순차적 형광 제자리 혼성화를 통한 다중 유전자 분석 기술

Figure 12.22 MERSCOPE – MERFISH 기술을 상업화한 Vizgen의 MERSCOPE 플랫폼

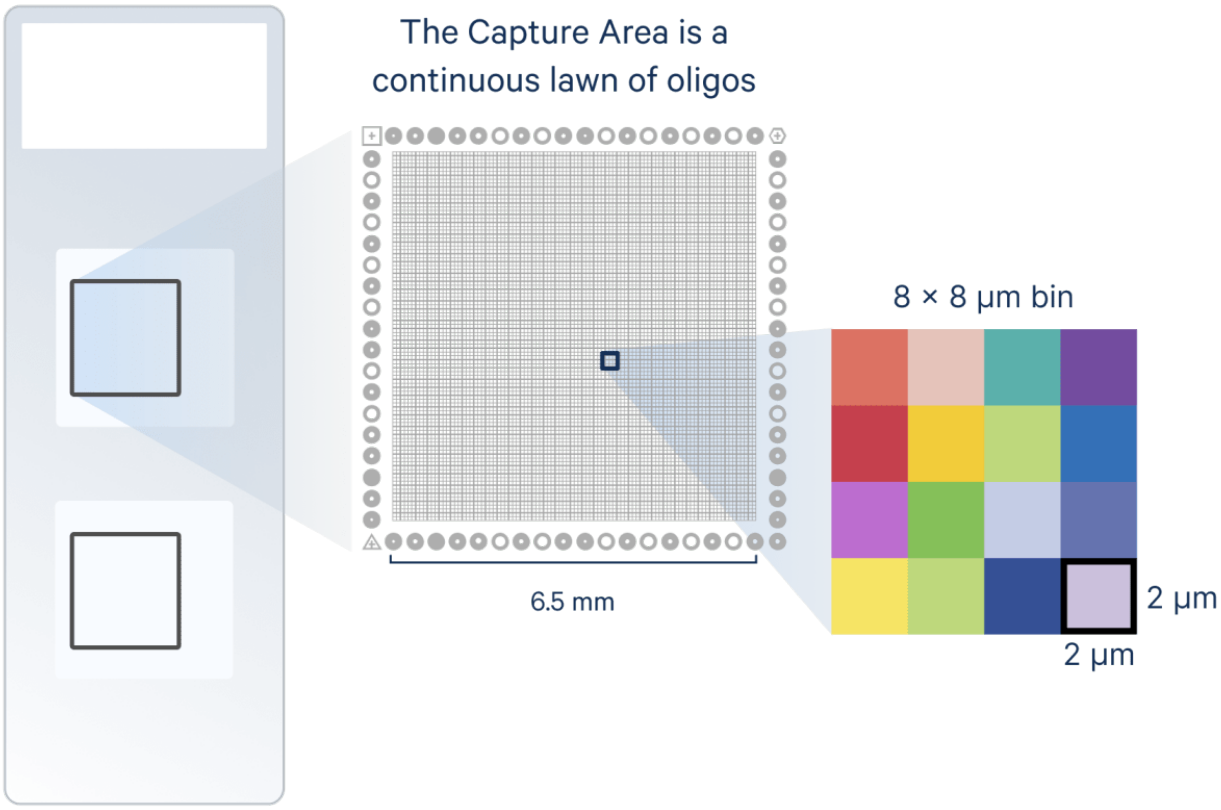


Figure 246: Visium HD

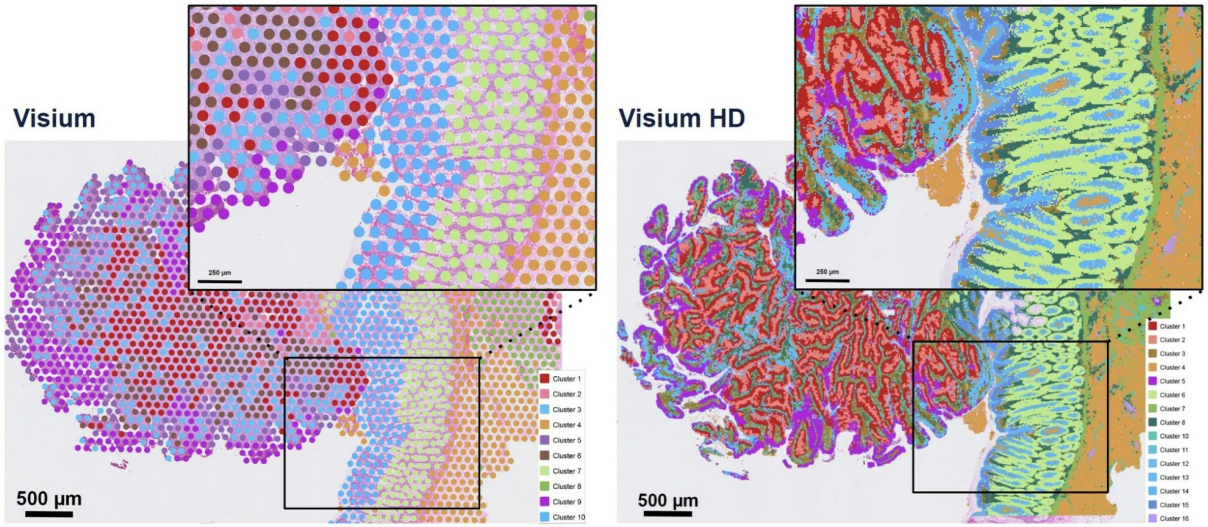


Figure 247: Visium과 Visium HD 비교

13.2.4 제자리 시퀀싱 기술

ISS와 FISSEQ의 원리: 제자리 시퀀싱(In Situ Sequencing, ISS)은 조직에서 RNA를 추출하지 않고 그 자리에서 직접 시퀀싱하는 기술이다. FISSEQ(Fluorescent In Situ Sequencing)는 이 분야의 선구적 기술로, 복잡한 프라이머 설계와 증폭 과정을 통해 다양한 RNA 서열을 in situ 에서 읽어낼 수 있다.

Xenium 플랫폼: 10x Genomics의 Xenium은 ISS와 FISSEQ 기술을 기반으로 한 상업적 플랫폼으로, 수백 개의 유전자를 세포 해상도로 동시에 분석할 수 있다. 이 플랫폼은 사용자 친화적인 워크플로우와 자동화된 분석 파이프라인을 제공하여, 연구자들이 복잡한 기술적 세부사항에 얽매이지 않고 생물학적 질문에 집중할 수 있게 해준다.

Figure 12.23 Xenium 제자리 시퀀싱 – 10x Genomics Xenium 플랫폼의 제자리 시퀀싱 과정

실험 과정에서는 조직 절편에 특수하게 설계된 프라이머를 이용하여 역전사 반응을 수행하고, 여러 라운드의 시퀀싱 반응을 통해 각 mRNA 분자의 서열을 결정한다. 이때 각 시퀀싱 라운드에서는 서로 다른 형광 염료를 사용하여 뉴클레오타이드를 구별하며, 고해상도 현미경을 통해 신호를 검출한다.

Figure 12.24 제자리 시퀀싱 – 조직에서 RNA를 추출하지 않고 직접 시퀀싱하는 기술 개념도

Figure 12.25 FISSEQ – 형광 제자리 시퀀싱 기술의 선구적 방법

13.2.5 공간전사체학 데이터 분석의 특수성

공간 정보의 활용: 공간전사체학 데이터 분석에서는 기존의 단일세포 분석에 공간 정보가 추가된다. 이를 통해 세포 간 물리적 거리, 조직 구조와의 관계, 국소적 미세환경의 영향 등을 분석할 수 있다. 공간적 자기상관(spatial autocorrelation) 분석을 통해 특정 유전자가 공간적으로 클러스터링되는 패턴을 확인할 수 있으며, 이는 세포 간 신호전달이나 조직 기능과 관련된 중요한 정보를 제공한다.

세포 간 상호작용 추론: 공간 정보를 활용하여 인접한 세포들 간의 리간드-수용체 상호작용을 추론할 수 있다. 이를 통해 특정 신호전달 경로가 조직의 어느 영역에서 활성화되는지 등을 매핑할 수 있다.

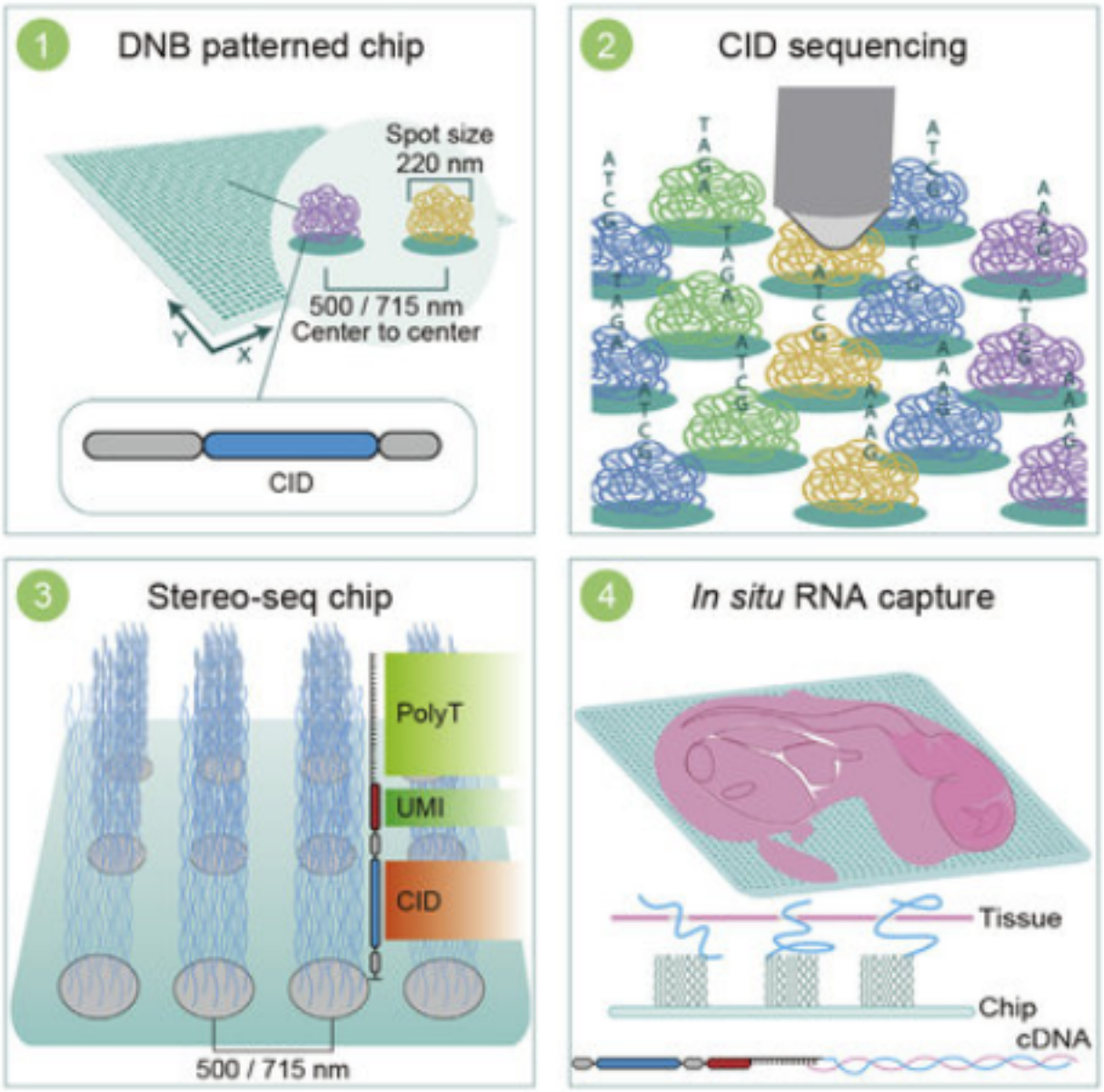


Figure 248: STOmics Stereo-seq

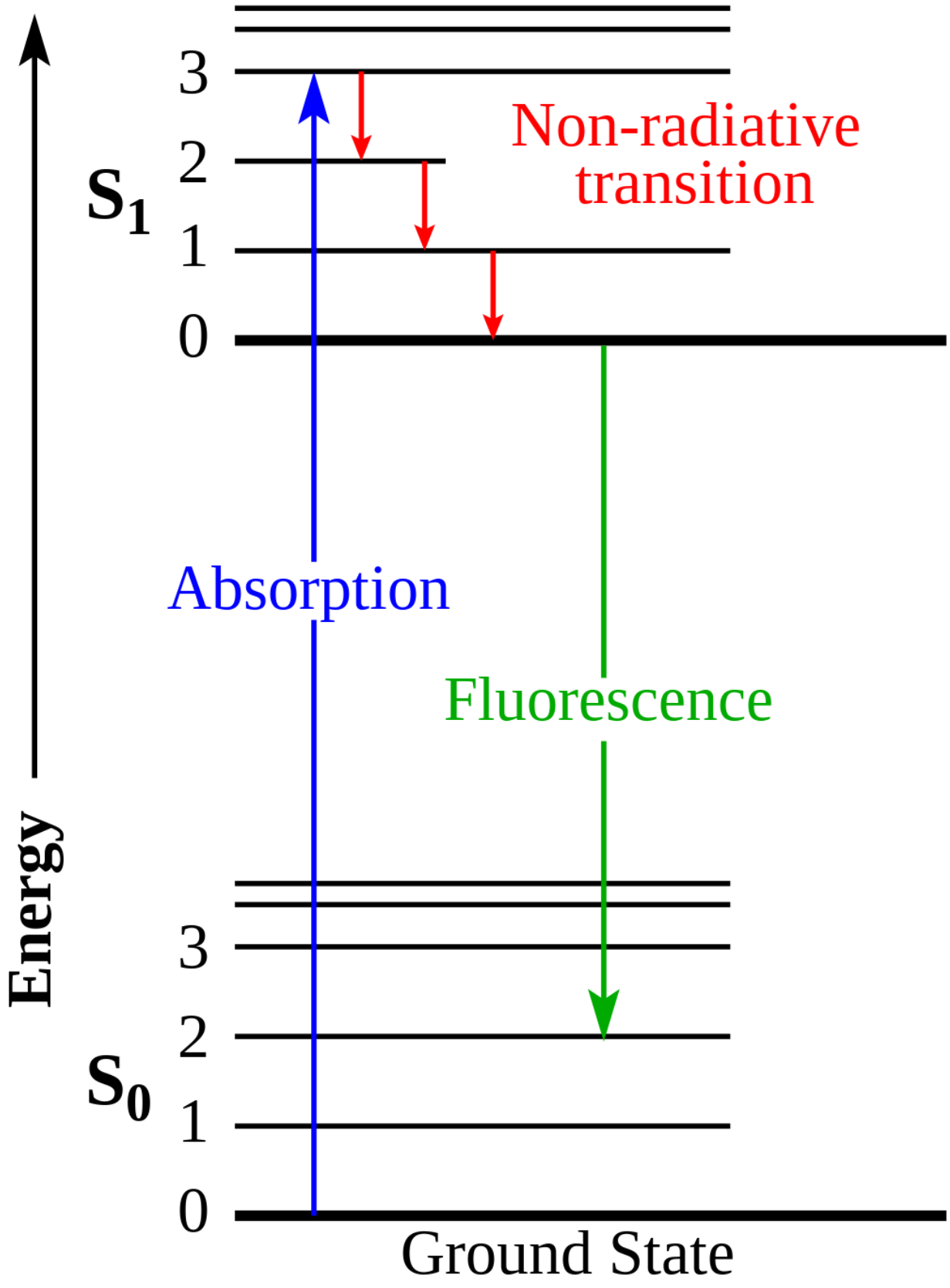


Figure 249: 형광 원리

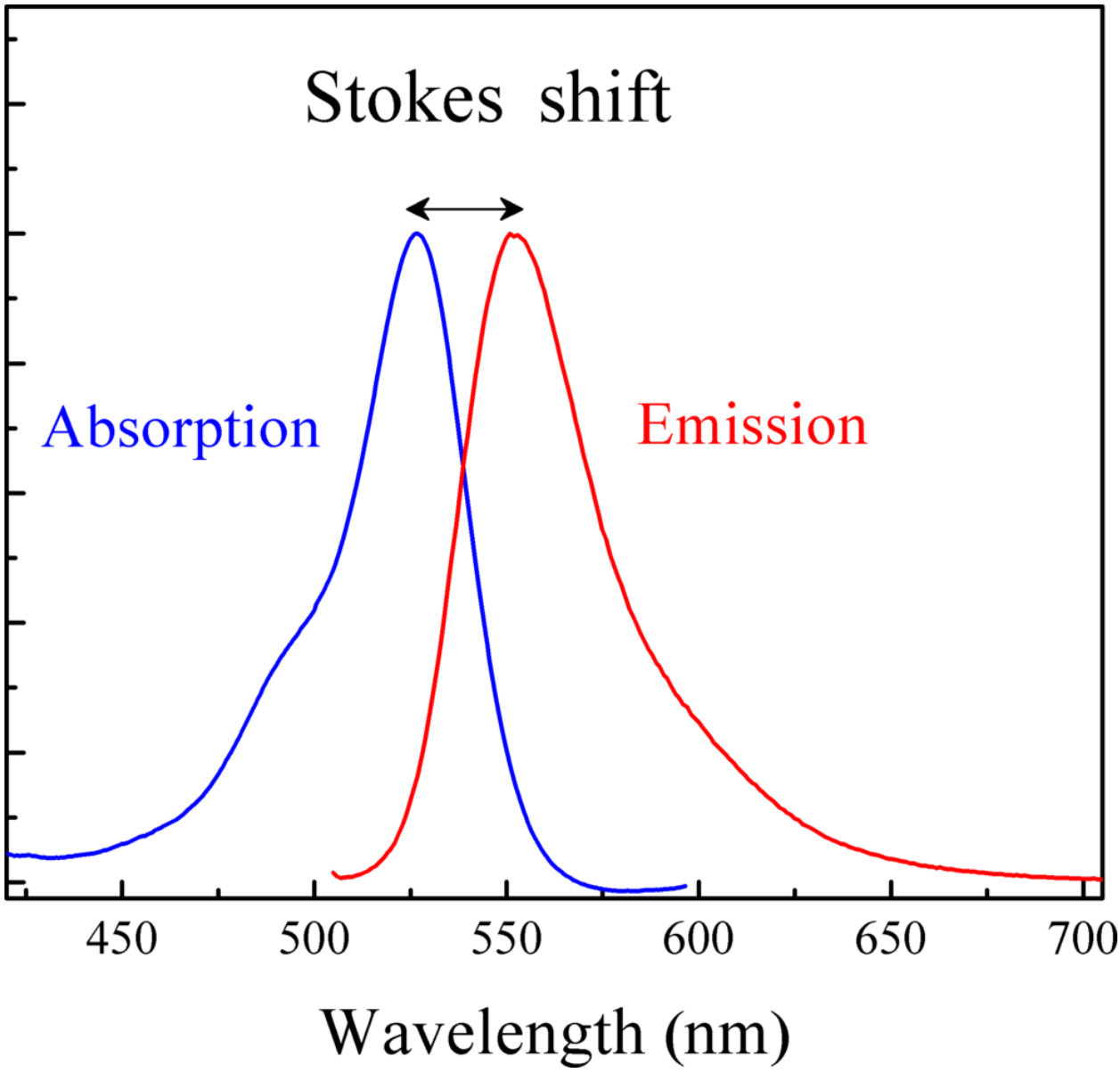


Figure 250: 스토크스 이동

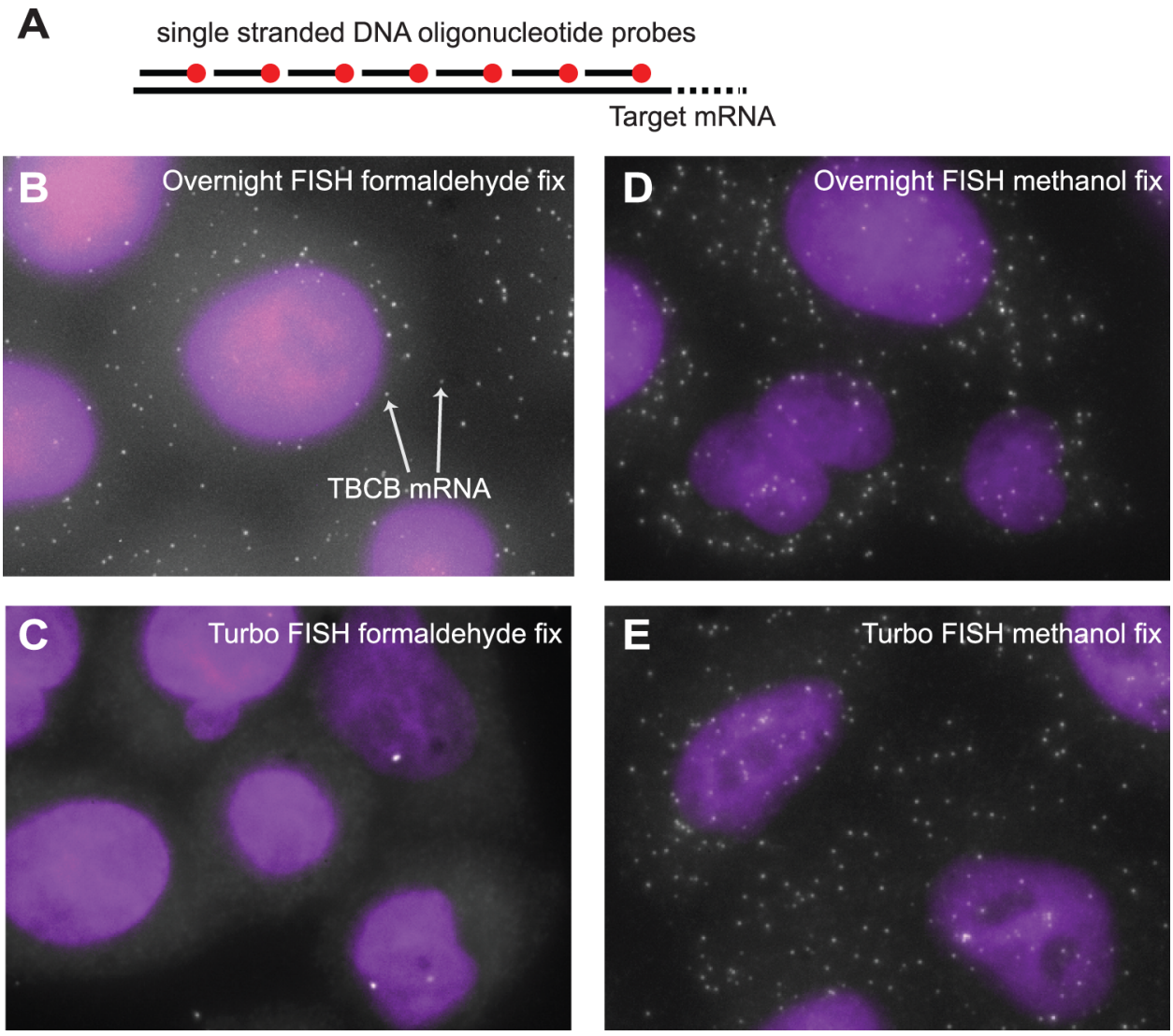


Figure 251: FISH 개요

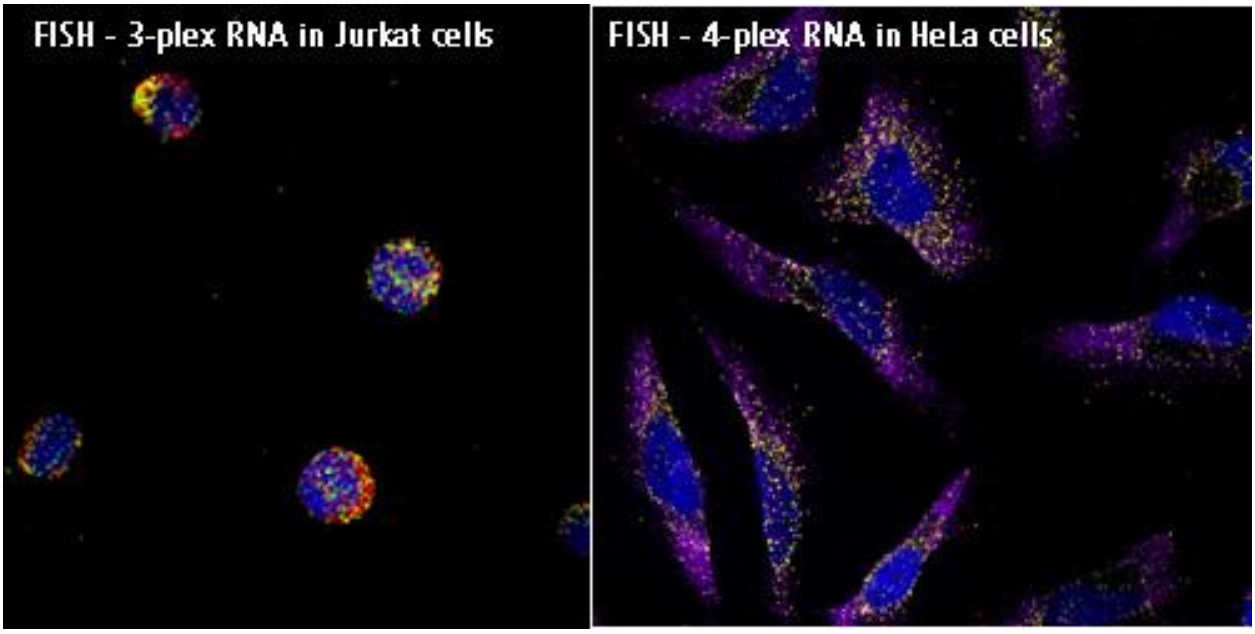


Figure 252: FISH 이미징

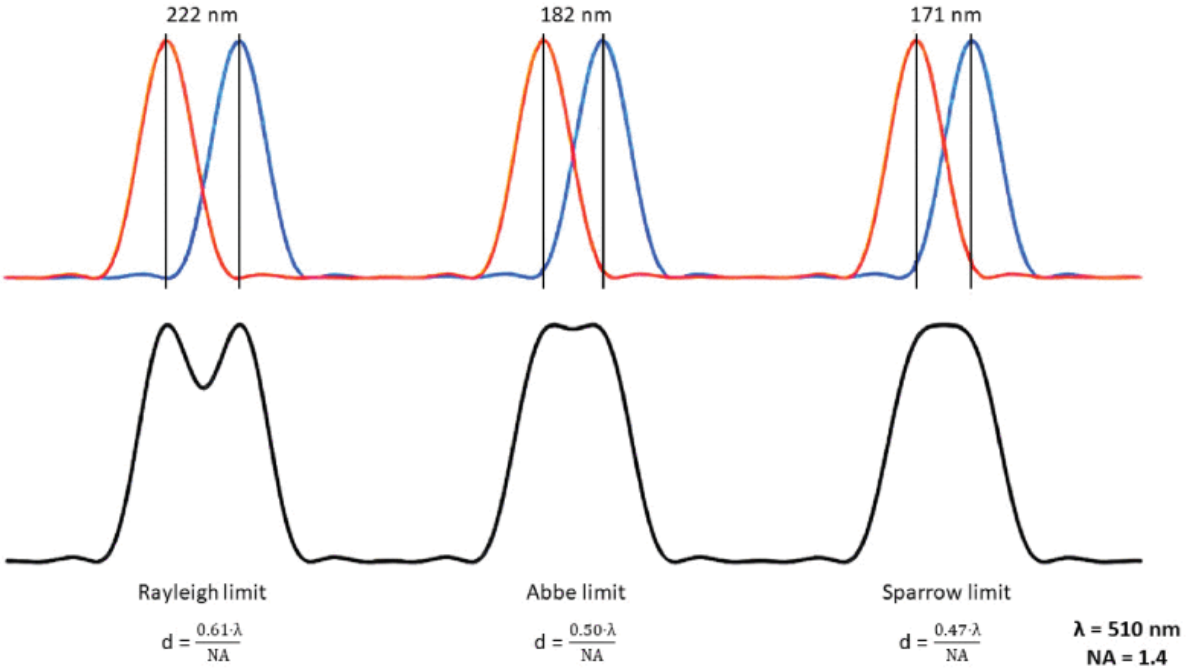


Figure 253: 회절 한계

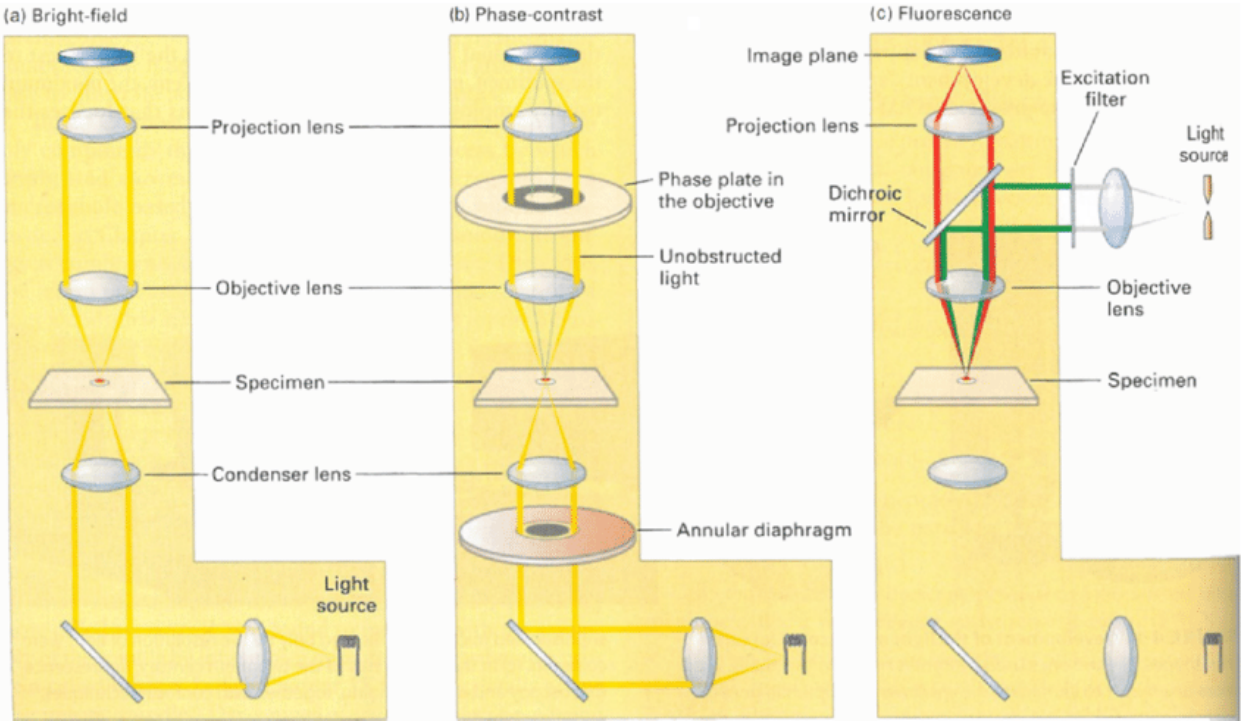


Figure 254: 이미징 기술들

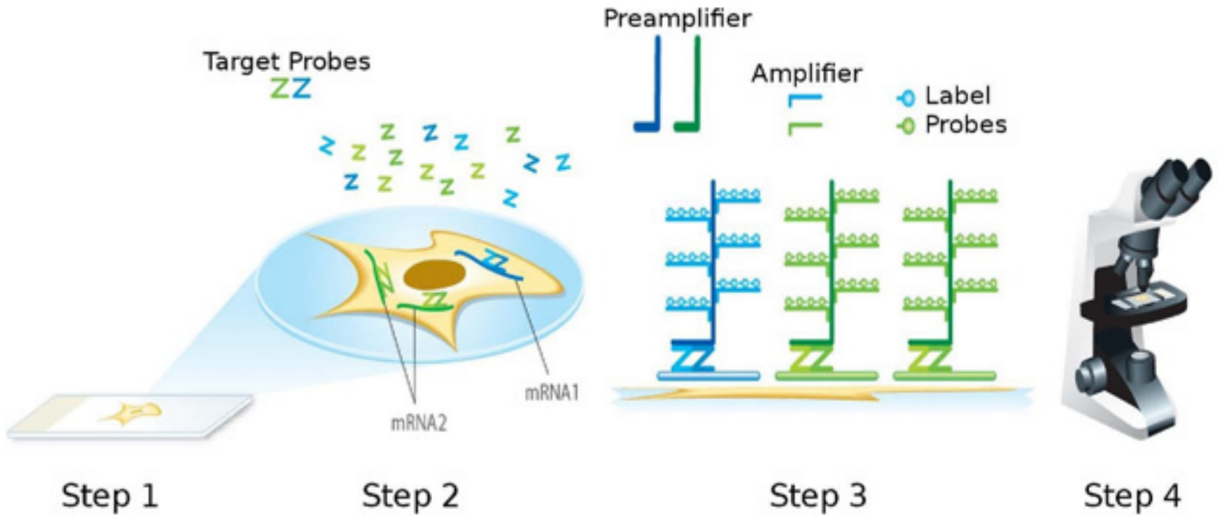


Figure 255: RNAscope

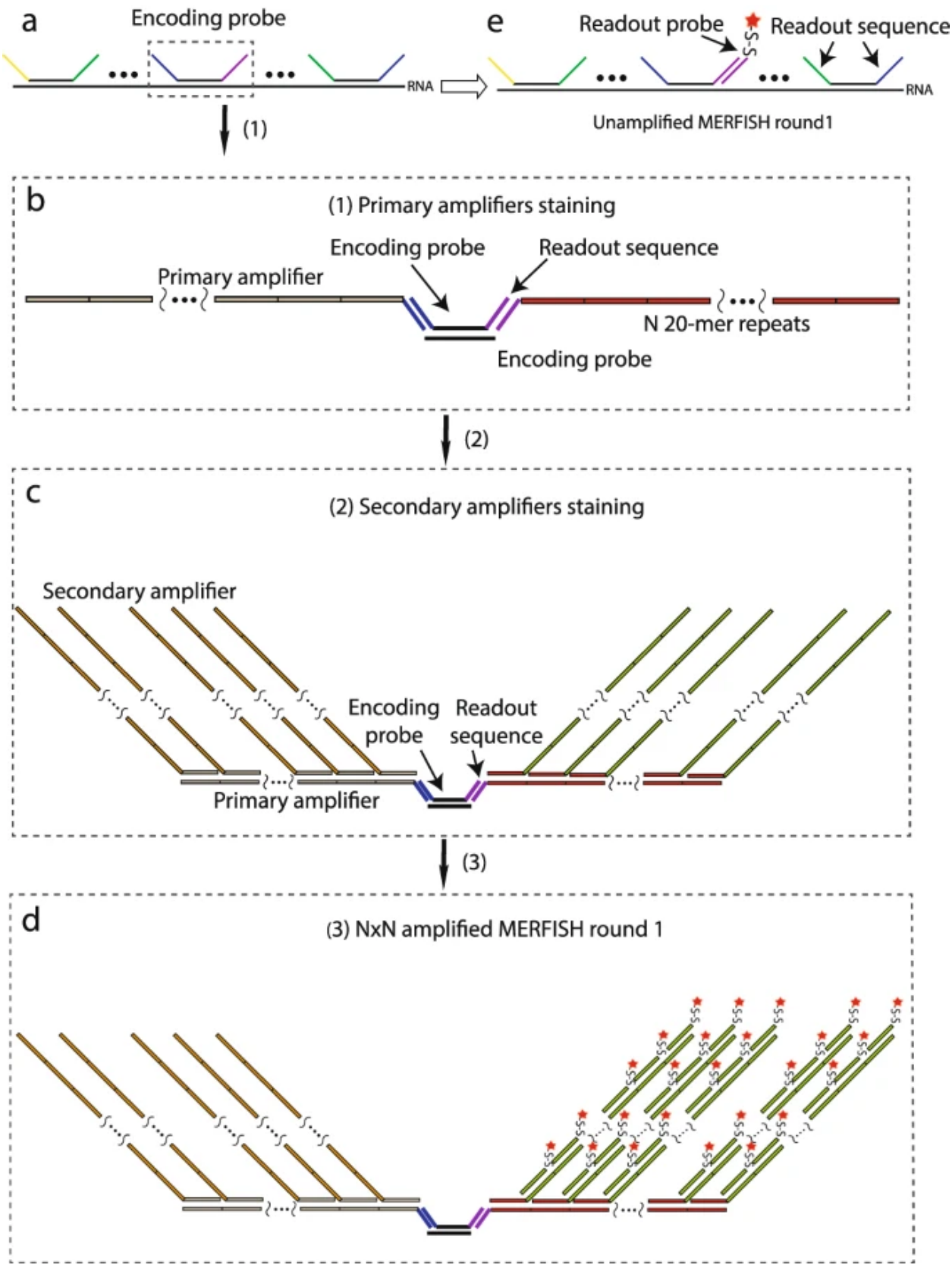


Figure 256: MERFISH 증폭

RNA	Barcode
A	11…0
B	01…1
C	10…1
D	10…0
E	01…0
F	00…1
G	11…1

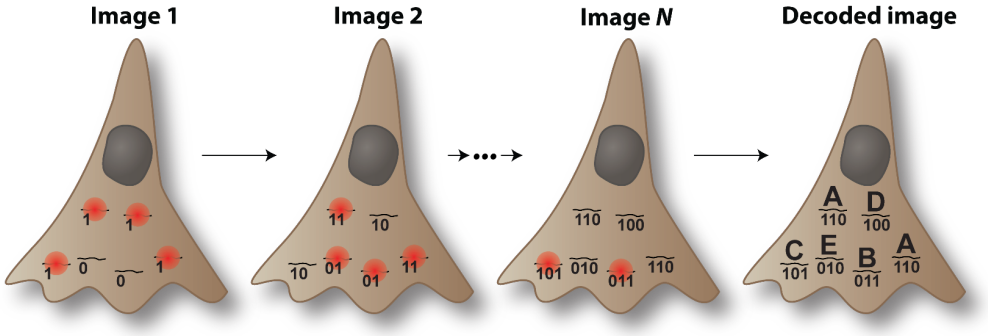


Figure 257: MERFISH 조합 바코딩

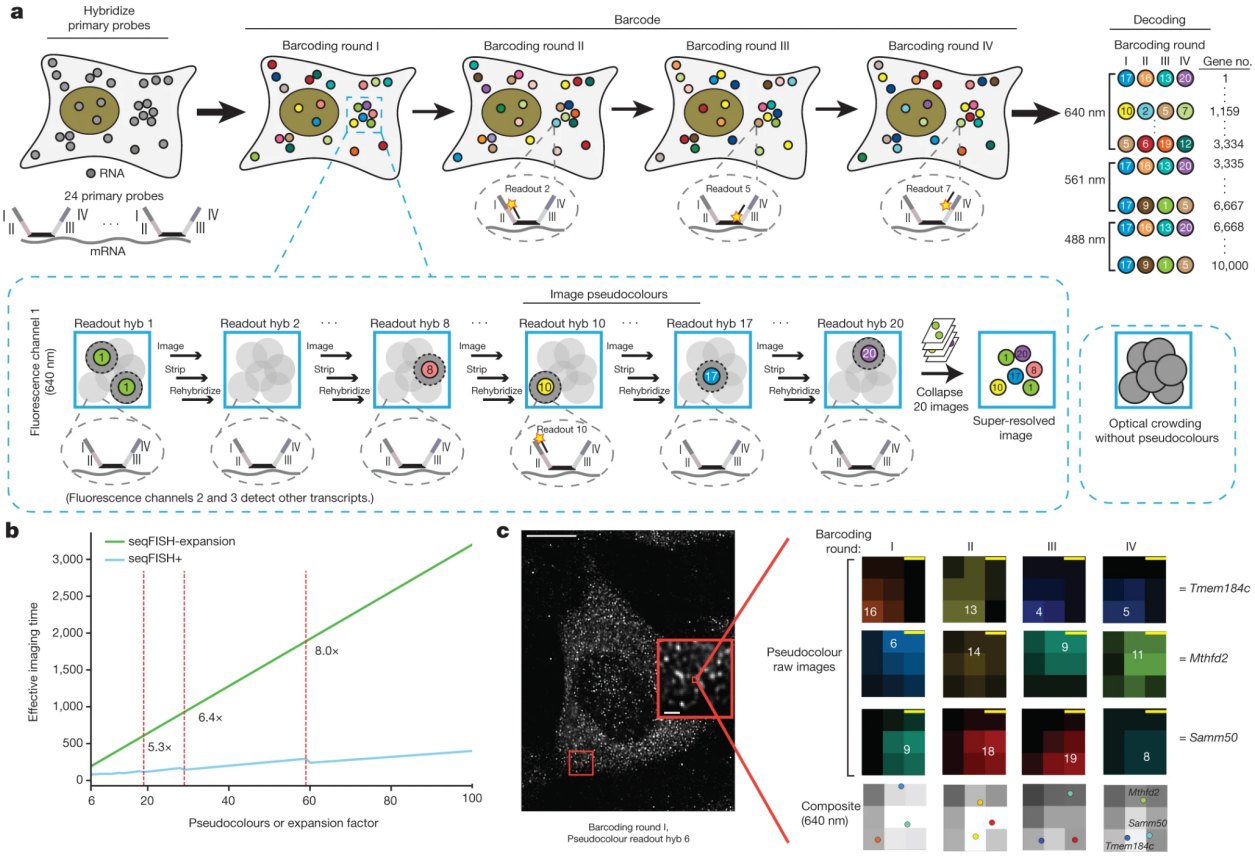


Figure 258: SeqFISH

Figure 259: MERSCOPE

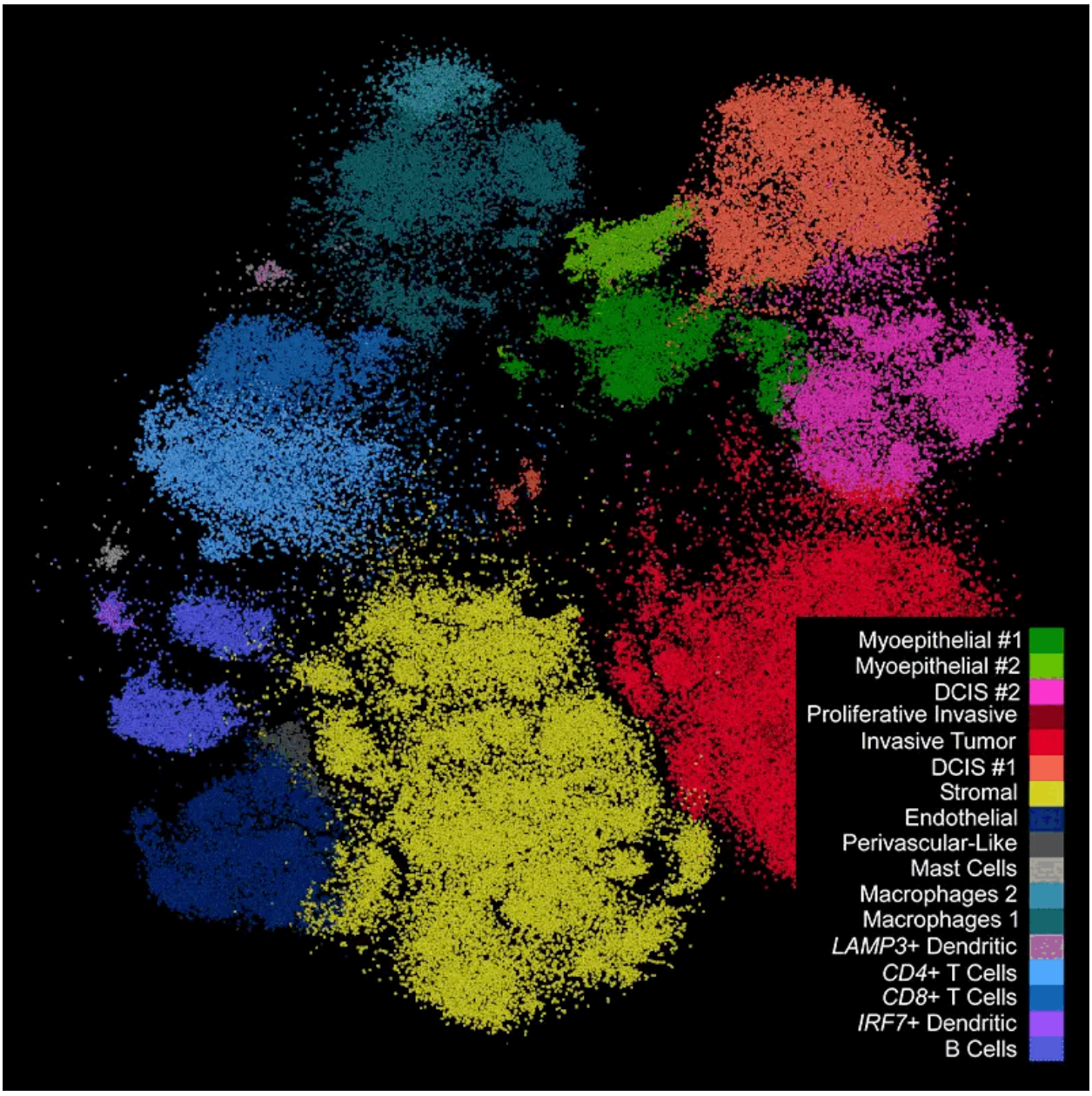
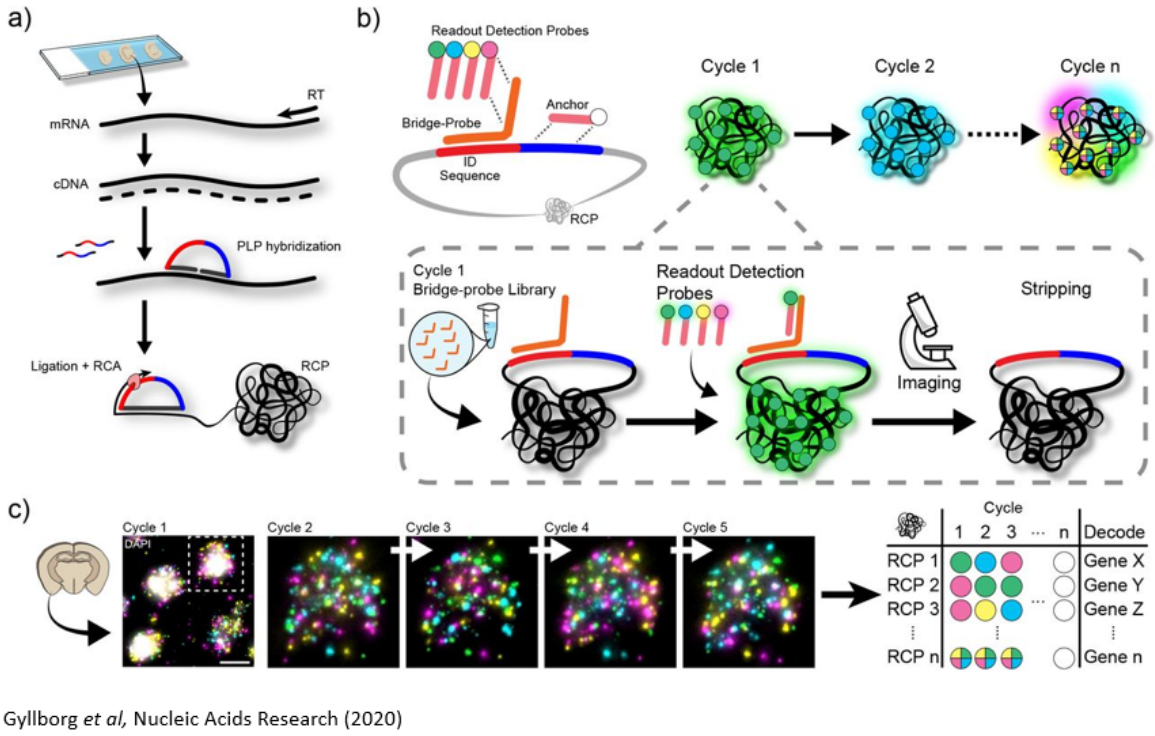


Figure 260: Xenium 제자리 시퀀싱



Gyllborg *et al,* Nucleic Acids Research (2020)

Figure 261: 제자리 시퀀싱

13.2.6 단일세포 수준에서의 공간 전사체학

세포 분할: 공간 전사체학을 단일세포 수준에서 분석하기 위해서는 먼저 조직 이미지에서 개별 세포를 정확히 구분하고 분할하는 과정, 즉 세포 분할(cell segmentation)이 필요하다.

Figure 12.26 세포 분할 – 공간 이미징 데이터에서 개별 세포를 분할하고 식별하는 전산 방법

세포 유형 식별: 세포 분할이 성공적으로 수행되면, 각 개별 세포에 대한 유전자 발현 프로파일을 얻을 수 있게 된다. 이때는 기존의 단일세포 RNA-seq 분석과 동일한 방법론을 적용하여 세포 유형을 식별할 수 있다. 구체적으로는 차원 축소 기법(PCA, t-SNE, UMAP), 클러스터링 알고리즘(Leiden, Louvain), 그리고 마커 유전자 기반 주석화 등의 표준적인 단일세포 분석 파이프라인을 그대로 사용할 수 있다.

이 접근법의 가장 큰 장점은 세포 유형 정보와 함께 정확한 공간적 위치 정보를 동시에 얻을 수 있다는 것이다. 즉, "어떤 세포 유형이 조직의 어느 위치에 분포하는가"라는 질문에 대해 단일세포 해상도로 답할 수 있게 된다. 이를 통해 세포 간 공간적 상호작용, 조직 구조와 세포 유형의 관계, 질병 진행에 따른 세포 분포 변화 등을 정밀하게 분석할 수 있다.

Figure 12.27 공간 세포 유형 식별 – 공간 전사체학 데이터를 이용한 조직 내 세포 유형 매핑

그러나 세포 분할은 기술적으로 쉽지 않은 편이다. 주요 도전 과제들은 다음과 같다. 첫째, 조직 내에서 세포들이 매우 밀집하게 배치되어 있어 경계를 명확히 구분하기 어렵다. 둘째, 세포마다 크기와 모양이 다른데, 특히 비정형적 모양의 세포들은 전통적인 알고리즘으로 처리하기 극히 어렵다. 셋째, DAPI나 세포질 마커 등의 염색 품질이 균일하지 않아 일부 세포가 과소 또는 과대 분할될 위험이 있다. 따라서 최근에는 세포 분할 과정을 생략하고도 단일세포 수준의 공간 전사체 분석을 수행할 수 있는 segmentation-free 방법들이 개발되고 있다. 이러한 방법들은 유전자 발현 패턴의 공간적 변화를 직접 분석한다.

대표적인 예로 SSAM 같은 방법이 있다. 이 방법은 공간적 좌표와 유전자 발현 정보를 동시에 고려하여 비슷한 전사체 프로파일을 가진 영역을 클러스터링하고, 이를 통해 세포 유형을 추론한다. 이러한 접근법은 세포 경계를 명시적으로 정의하지 않고도 세포 수준의 생물학적 정보를 추출할 수 있어, 공간 전사체학의 또 다른 가능성을 열어주고 있다.

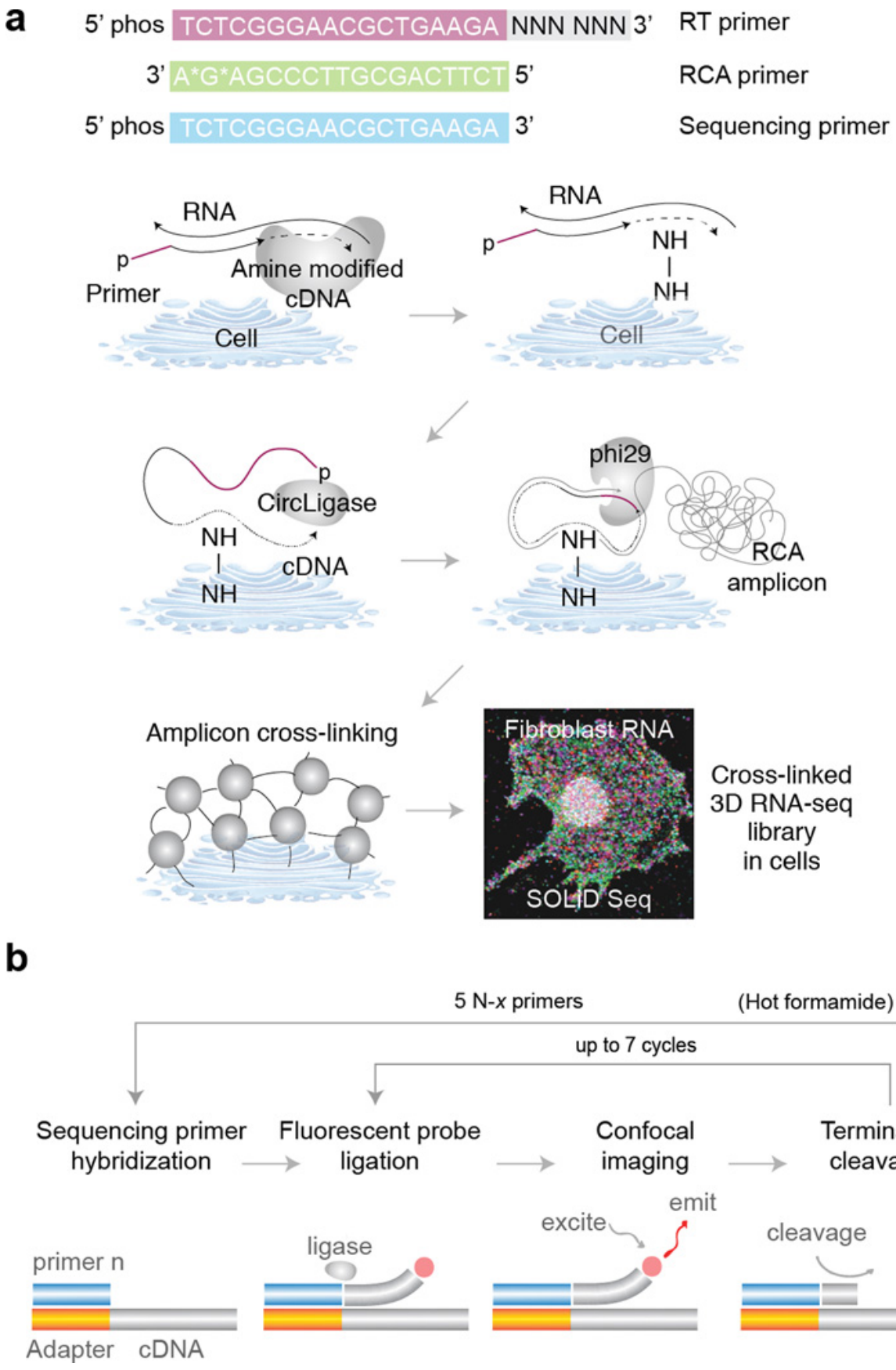


Figure 262: FISSEQ

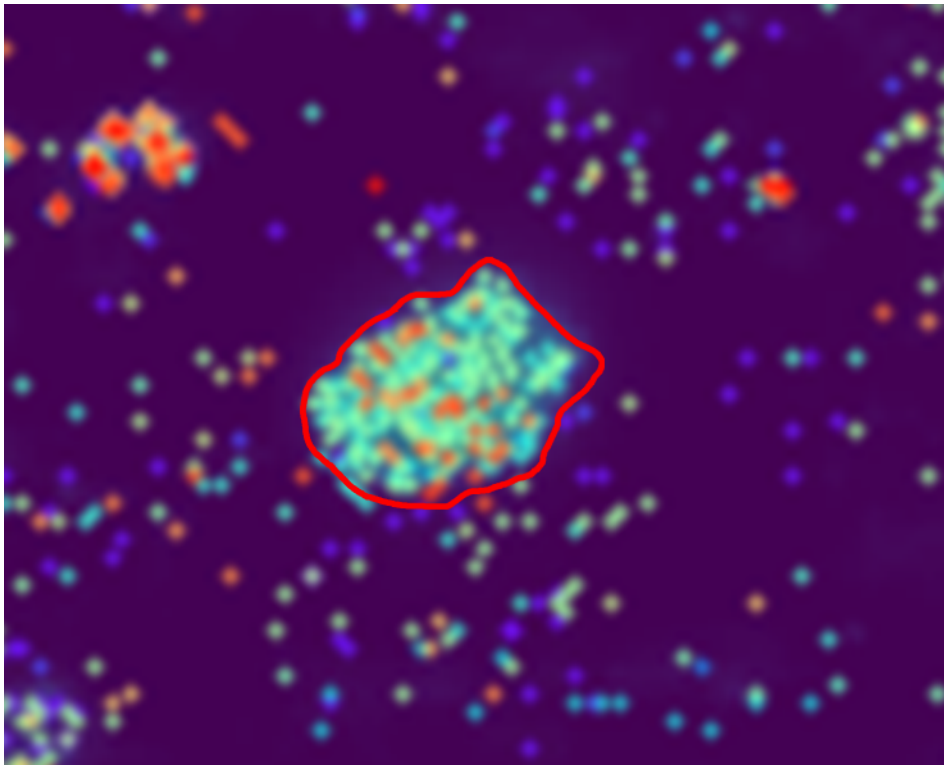


Figure 263: 세포 분할

Figure 12.28 SSAM – 공간 전사체학 데이터의 통계 분석과 하위 세포 매핑 방법론

13.3 통합 분석과 다중 모달 접근법

13.3.1 단일세포와 공간전사체학 데이터의 통합

참조 기반 스팟 디콘볼루션: 단일세포 RNA-seq 데이터를 참조로 활용하여 공간전사체학의 각 스팟에 포함된 세포 유형 구성을 추정하는 방법이다. 이는 공간전사체학의 상대적으로 낮은 해상도를 보완하여 세포 수준의 공간 정보를 얻을 수 있게 해준다.

가상 공간 재구성: 단일세포 데이터의 세포들을 실제 조직 구조에 매핑하여 가상의 공간 전사체 지도를 생성하는 방법도 개발되고 있다. 이는 공간전사체학 실험을 수행하기 어려운 상황에서 공간 정보를 추론하는 데 활용된다.

13.3.2 다중 오믹스 통합

동시 측정 기술: 최근에는 공간상에서 전사체와 후성유전체(epigenome), 단백질체(proteome) 정보를 동시에 측정하는 기술들이 개발되고 있다. Spatial CITE-seq은 항체를 이용하여 단백질 발현과 RNA 발현을 공간상에서 동시에 측정하며, ATAC-seq과 결합한 방법들은 전사체와 염색질 접근성을 함께 분석할 수 있다.

계산적 통합 방법: 이러한 공간 다중 오믹스 기술들은 측정될 수 있는 유전자 수, 정확도 등의 한계가 있다. 따라서 단일세포 실험을 통해 얻은 다중 오믹스 데이터를 공간 오믹스 데이터와 통합하는 계산적 방법들도 활발히 개발되고 있다.

13.4 기술의 한계와 도전 과제

13.4.1 기술적 한계

감도와 특이성의 균형: 현재의 공간전사체학 기술들은 모두 감도와 특이성 사이의 트레이드오프를 가지고 있다. 높은 공간 해상도를 얻으려면 측정할 수 있는 유전자 수가 제한되고, 많은 유전자를 측정하려면 공간 해상도가 떨어진다. 이러한

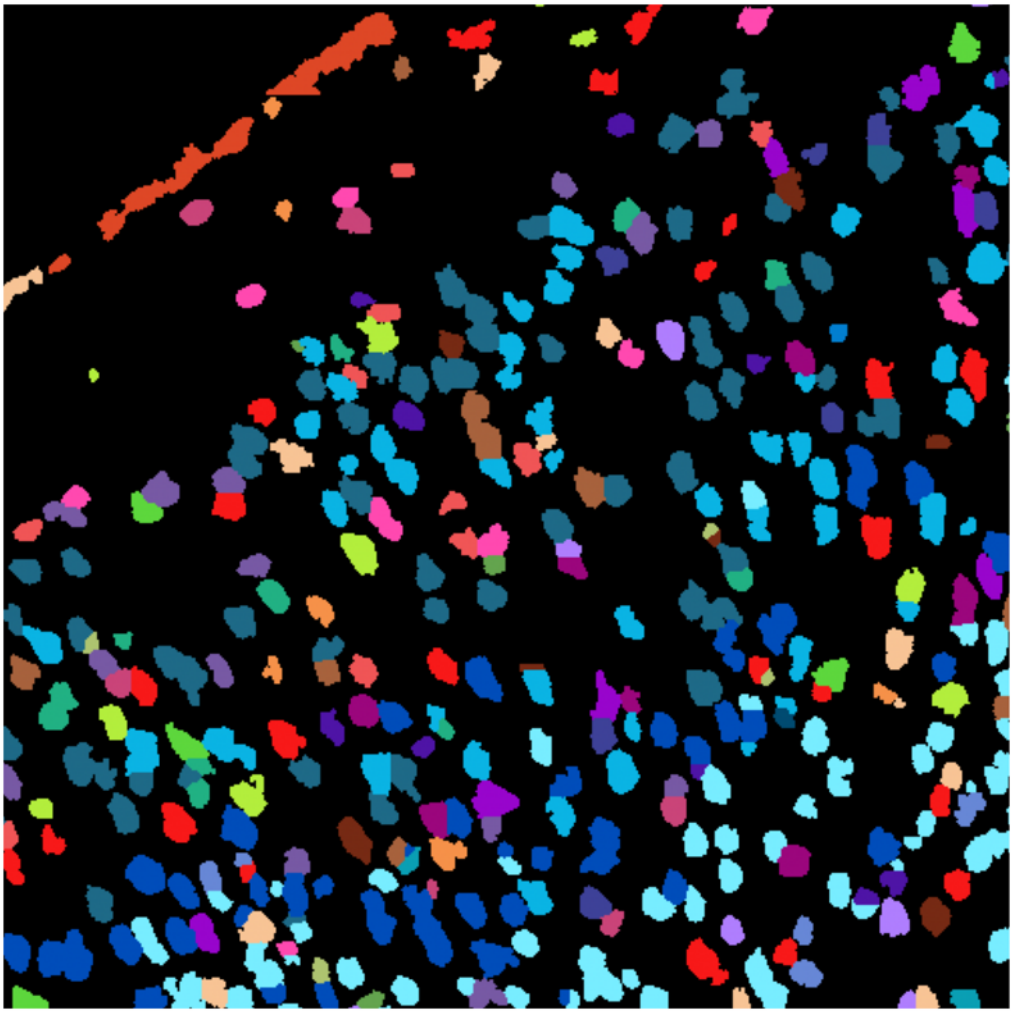


Figure 264: 공간 세포 유형 식별

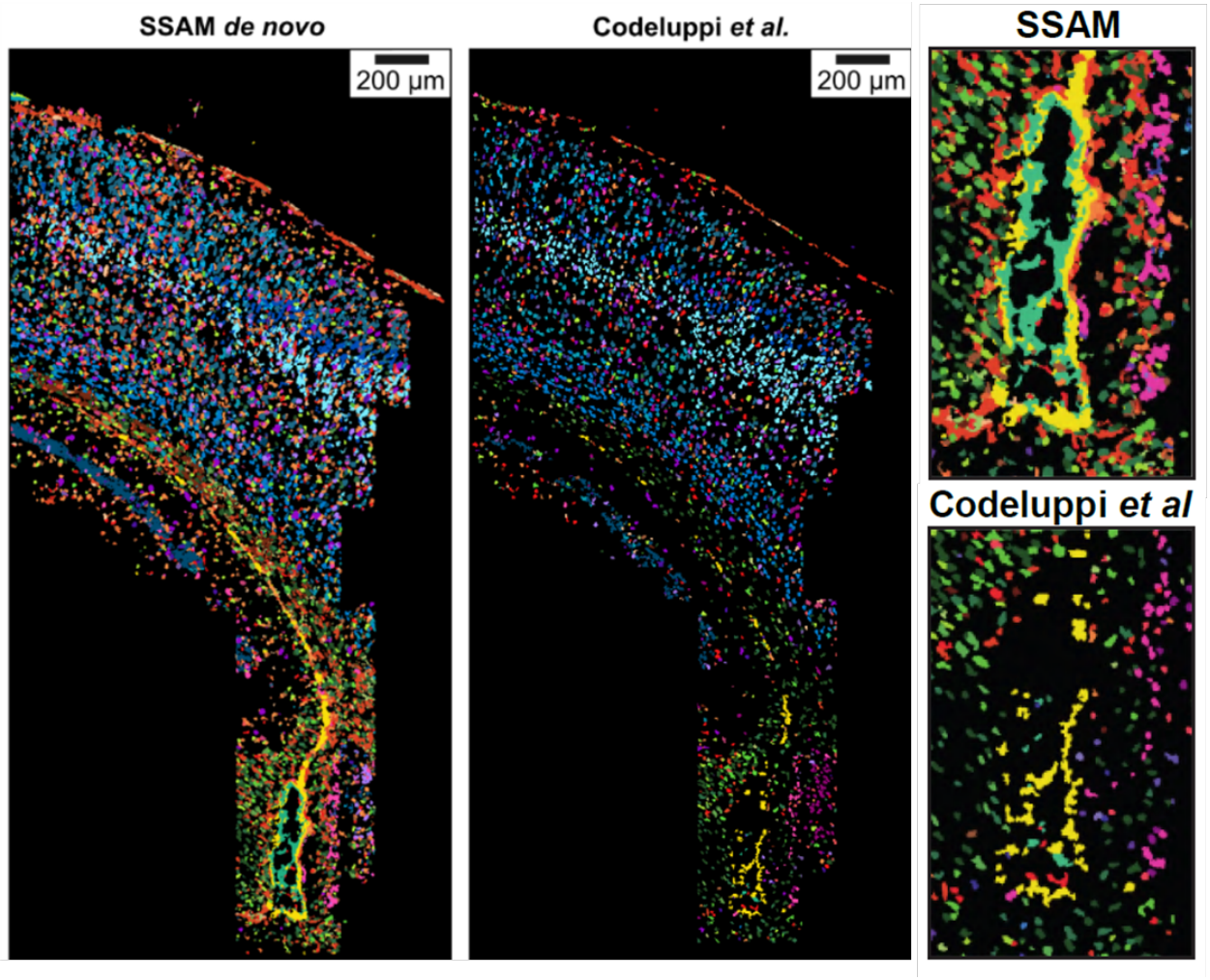


Figure 265: SSAM

기술적 제약은 실험 설계 시 연구 목적에 따른 신중한 선택을 요구한다.

조직 처리로 인한 아티팩트: 조직의 고정, 동결, 절편 과정에서 RNA 품질이 저하되거나 공간 구조가 변형될 수 있다. 특히 RNA의 분해는 측정 결과에 직접적인 영향을 미치므로, 샘플 준비 과정의 표준화가 중요하다.

13.4.2 분석상의 도전 과제

배치 효과와 표준화: 서로 다른 플랫폼, 실험실, 또는 실험 조건에서 생성된 데이터를 통합할 때 배치 효과가 심각한 문제가 될 수 있다. 특히 공간전사체학에서는 조직의 고정 방법, 절편 두께, 보관 조건 등이 모두 배치 효과의 원인이 될 수 있어, 이를 보정하는 방법론의 개발이 중요하다.

통계적 검정력: 공간전사체학 데이터는 일반적으로 측정되는 유전자 수가 전체 전사체에 비해 제한적이므로, 통계적 검정력이 상대적으로 낮을 수 있다. 이는 특히 희귀한 세포 유형이나 미묘한 발현 변화를 탐지하는 데 제약이 될 수 있다.

13.5 임상 응용과 미래 전망

13.5.1 질병 연구에서의 응용

암 연구: 공간전사체학은 종양 미세환경의 복잡한 구조를 이해하는 중요한 도구이다. 암세포와 면역세포, 혈관내피세포, 섬유아세포 등 다양한 세포 유형들 간의 공간적 관계와 상호작용을 직접 관찰할 수 있어, 암의 진행, 전이, 치료 저항성 메커니즘을 이해하는 데 중요한 기여를 하고 있다.

신경과학 연구: 뇌 조직의 복잡한 구조와 신경세포 네트워크를 이해하기 위해 공간전사체학이 적극 활용되고 있다. 특히 알츠하이머병, 파킨슨병 등 신경퇴행성 질환에서 질병 진행에 따른 공간적 변화 패턴을 추적하는 연구가 활발히 진행되고 있다.

13.5.2 정밀의학으로의 발전

환자 맞춤 치료: 개별 환자의 조직에서 세포 구성과 공간적 배치를 분석하여 최적의 치료법을 선택하는 정밀의학 접근법이 현실화되고 있다. 특히 면역치료의 효과를 예측하기 위해 종양 내 면역세포의 분포와 활성화 상태를 분석하는 연구가 주목받고 있다.

치료 반응 모니터링: 치료 전후의 조직을 비교 분석하여 치료 효과를 모니터링하고, 치료 저항성이 발생하는 부위와 메커니즘을 식별하는 데 공간전사체학이 활용되고 있다.

13.5.3 기술 발전 방향

해상도와 처리량의 개선: 향후 기술 발전은 더 높은 공간 해상도와 더 많은 유전자를 동시에 측정할 수 있는 처리량 개선에 초점을 맞출 것으로 예상된다. 단일세포 수준의 공간 해상도에서 전체 전사체를 측정하는 것이 궁극적인 목표이다.

실시간 분석: 현재는 고정된 조직에서만 분석이 가능하지만, 향후에는 살아있는 조직에서 실시간으로 유전자 발현 변화를 추적할 수 있는 기술 개발이 기대된다.

다중 분자 측정: 전사체뿐만 아니라 단백질, 대사물질, 후성유전체 정보를 동시에 공간적으로 측정할 수 있는 통합적 다중오믹스 기술의 개발이 진행되고 있다.

13.6 결론

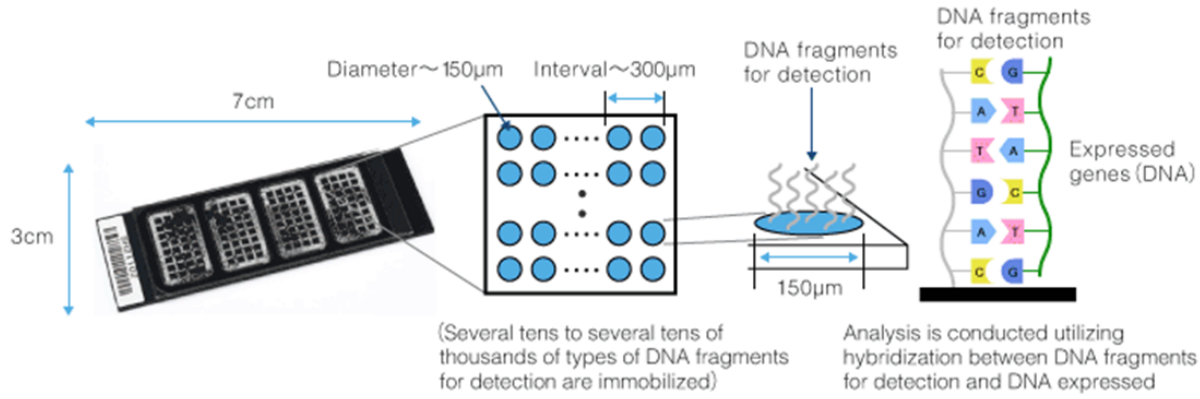
공간 전사체학과 단백체학은 조직 내 분자적 정보와 공간적 정보를 통합하여 생물학적 현상을 이해하는 새로운 접근법을 제공한다. 이러한 기술들의 통합과 발전을 통해 질병의 메커니즘 규명, 치료법 개발, 그리고 정밀의학 구현에서 중요한 돌파구를 제공할 것으로 기대된다.

특히 암 연구, 신경과학, 발생생물학 등 다양한 분야에서 조직의 공간적 구조와 기능적 특성을 동시에 분석할 수 있게 됨으로써, 생명 현상에 대한 더욱 깊이 있는 이해가 가능해질 것이다. 앞으로의 기술 발전을 통해 더 높은 해상도, 더 많은 분자 종류의 동시 측정, 그리고 실시간 분석이 가능해질 것으로 전망된다.

14 단백체학

14.1 단백체학 개요

단백체학(proteomics)은 생물체나 세포에서 발현되는 모든 단백질의 집합인 단백체(proteome)를 연구하는 분야이다. 유전체학이 생명체의 유전적 청사진을 다룬다면, 단백체학은 실제로 세포에서 기능하는 분자들을 직접적으로 연구한다. 단백질은 유전자의 최종 산물로서 세포의 구조를 형성하고, 효소 반응을 촉매하며, 신호전달을 담당하는 등 생명 현상의 핵심적인 역할을 수행한다.



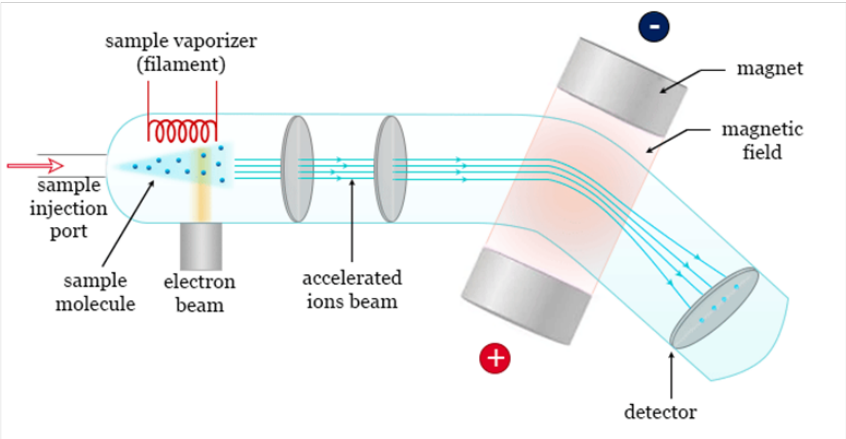


Figure 266: 마이크로어레이와 질량분석법

Figure 13.1 마이크로어레이와 질량분석법 – 전통적인 마이크로어레이와 현대적인 질량분석법 기술의 비교

단백체는 유전체와 달리 시간과 조건에 따라 동적으로 변화한다. 같은 유전체를 가진 세포라도 발달 단계, 환경 조건, 질병 상태에 따라 서로 다른 단백질 프로파일을 보인다. 이러한 특성 때문에 단백체학은 질병 진단, 치료 표적 발굴, 바이오마커 개발 등에서 중요한 역할을 한다. 또한 약물의 작용 메커니즘을 이해하고 부작용을 예측하는 데도 필수적인 정보를 제공한다.

최근에는 단일세포 수준에서 단백질을 분석하는 기술과 공간적 위치 정보를 보존하면서 단백질 분포를 관찰하는 기술들이 급속히 발전하고 있다. 이러한 기술적 진보는 조직 내 세포 간 상호작용, 질병의 진행 과정, 치료에 대한 반응 등을 이해하는 새로운 관점을 제공하고 있다.

14.2 단백체학 기법

14.2.1 질량분석법 기반 단백체학

매트릭스 보조 레이저 탈착/이온화 비행시간 질량분석법(Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry, MALDI-TOF MS)은 생체분자의 질량을 측정하는 분석 기술이다. 이 방법에서는 분석하고자 하는 시료를 특별한 매트릭스 물질과 혼합하여 공결정화시킨 후, 레이저 펄스를 조사하여 이온화시킨다.

매트릭스는 레이저 에너지를 흡수하여 시료 분자를 부드럽게 이온화시키는 역할을 한다. 이 과정에서 단백질이나 펩타이드와 같은 큰 생체분자도 분해되지 않고 그대로 이온화될 수 있다. 이온화된 분자들은 전기장에 의해 가속되어 비행관을 통과하며, 질량이 다른 이온들은 서로 다른 속도로 검출기에 도달한다. 비행시간을 측정함으로써 각 이온의 질량 대 전하비(m/z)를 정확하게 계산할 수 있다.

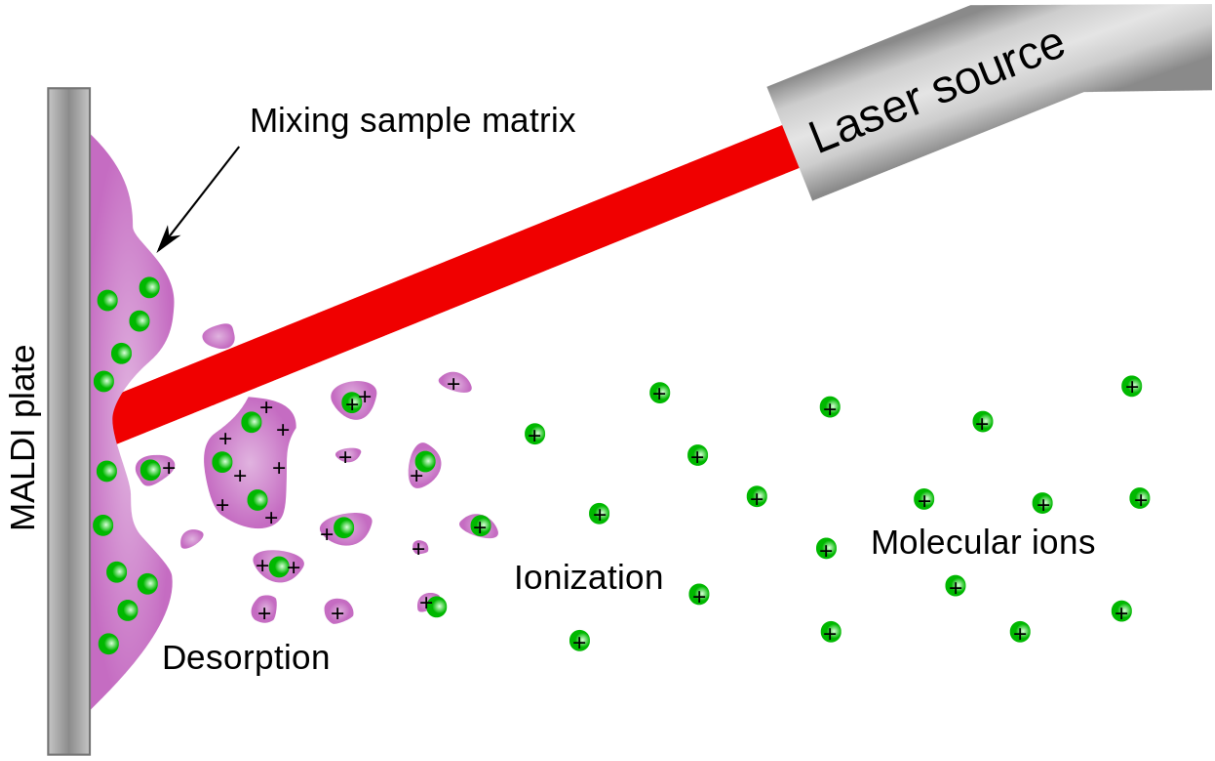


Figure 267: MALDI 원리

Figure 13.2 MALDI 원리 – 매트릭스 보조 레이저 탈착/이온화 과정의 기본 원리

Figure 268: TOF 질량분석기

Figure 13.3 TOF 질량분석기 – 비행시간 측정을 통한 질량 분석 원리

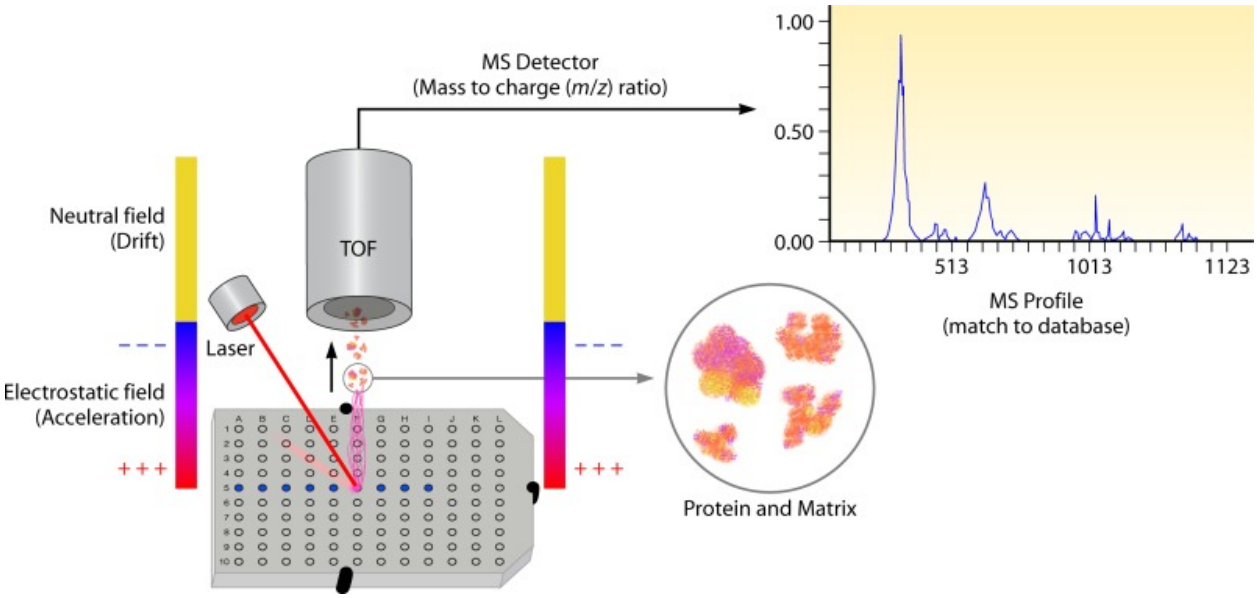


Figure 269: MALDI-TOF 시스템

Figure 13.4 MALDI-TOF 시스템 – MALDI와 TOF를 결합한 질량분석 시스템의 전체 구성

Figure 270: 다나카 고이치

Figure 13.5 다나카 고이치 – 생체고분자의 질량분석법 개발로 2002년 노벨 화학상을 수상한 다나카 고이치

액체 크로마토그래피-질량분석법(Liquid Chromatography–Mass Spectrometry, LC–MS)은 복잡한 생체 시료에서 화합물을 분리하고 동정하는 분석 도구이다. 먼저 액체 크로마토그래피에서 시료 성분들이 화학적 성질에 따라 시간차를 두고 분리되고, 이어서 질량분석기에서 각 성분의 정확한 질량과 구조 정보를 얻는다.

Figure 13.6 HPLC 시스템 – 고성능 액체 크로마토그래피를 이용한 단백질 분리 시스템

Figure 13.7 질량분석기 – 생체분자 분석을 위한 질량분석기의 기본 구성

Figure 13.8 질량분석법 과정 – 시료의 이온화부터 검출까지 전체적인 질량분석 과정

정량적 단백체학에서는 서로 다른 조건에서 단백질 발현량의 변화를 측정한다. 여기에는 라벨링 기반 방법과 라벨프리

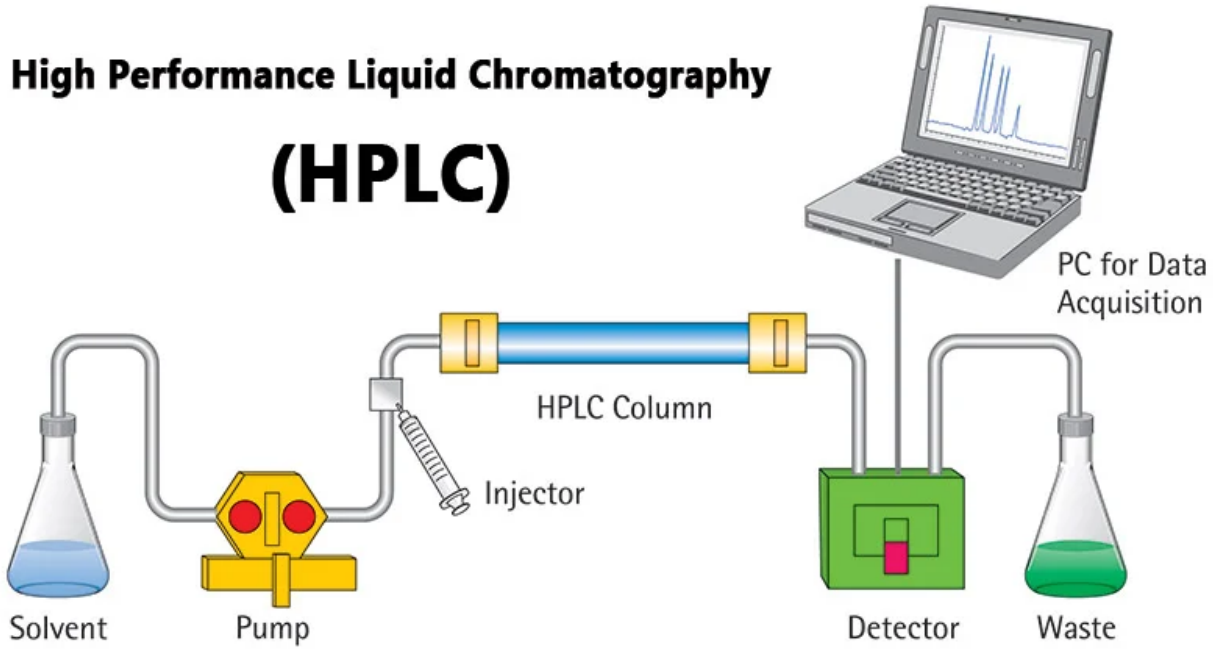


Figure 271: HPLC 시스템

방법이 주로 사용된다. 라벨링 방법은 동위원소 표지를 통해 정확한 정량이 가능하지만 비용이 높고 실험 설계가 복잡하다. 라벨프리 방법은 상대적으로 간단하지만 실험 간 변이를 보정하는 것이 중요하다.

14.2.2 항체 기반 단백체학

항체 기반 단백체학은 특정 단백질에 대한 항체의 특이적 결합을 이용하여 단백질을 검출하고 정량하는 방법들을 포함한다. 면역조직화학(immunohistochemistry)은 조직 절편에서 특정 단백질의 위치와 발현 정도를 시각화하는 기본적인 방법이다.

웨스턴 블롯(Western blot)은 전기영동으로 분리된 단백질을 막으로 전사한 후 특정 항체로 검출하는 방법으로, 단백질의 크기와 발현량을 정확히 분석할 수 있다. 단백질 어레이(protein array)는 다수의 항체나 단백질을 고밀도로 배열하여 단백질 간 상호작용이나 다중 단백질 발현을 동시에 분석할 수 있는 고처리량 기술이다.

14.3 공간 단백체학

14.3.1 질량분석 기반 공간 이미징

질량분석 이미징(Mass Spectrometry Imaging, MSI)은 조직 절편의 공간적 분자 분포를 시각화하는 기술이다. 조직 절편의 각 픽셀에서 질량 스펙트럼을 획득하여 특정 분자들의 공간적 분포 지도를 생성한다. 이를 통해 종양 조직에서 항암제의 분포나 대사물질의 변화를 직접 관찰할 수 있으며, 약물의 조직 침투성이나 효과를 평가할 수 있다.

Figure 13.9 MALDI 이미징 – 조직 절편에서 분자의 공간적 분포를 시각화하는 MALDI 이미징 기술

이미징 질량 세포분석법(Imaging Mass Cytometry, IMC)은 조직 절편에서 금속 태그가 결합된 항체를 이용하여 다중 단백질 분석을 수행하는 기술이다. 이 방법은 레이저 절제를 통해 조직의 특정 영역을 제거하면서 동시에 질량분석을 수행하여 1마이크로미터 해상도로 40개 이상의 단백질 마커를 동시에 검출할 수 있다.

Figure 13.10 질량 세포분석법 – CyTOF를 이용한 단일세포 수준 다중 단백질 분석 원리

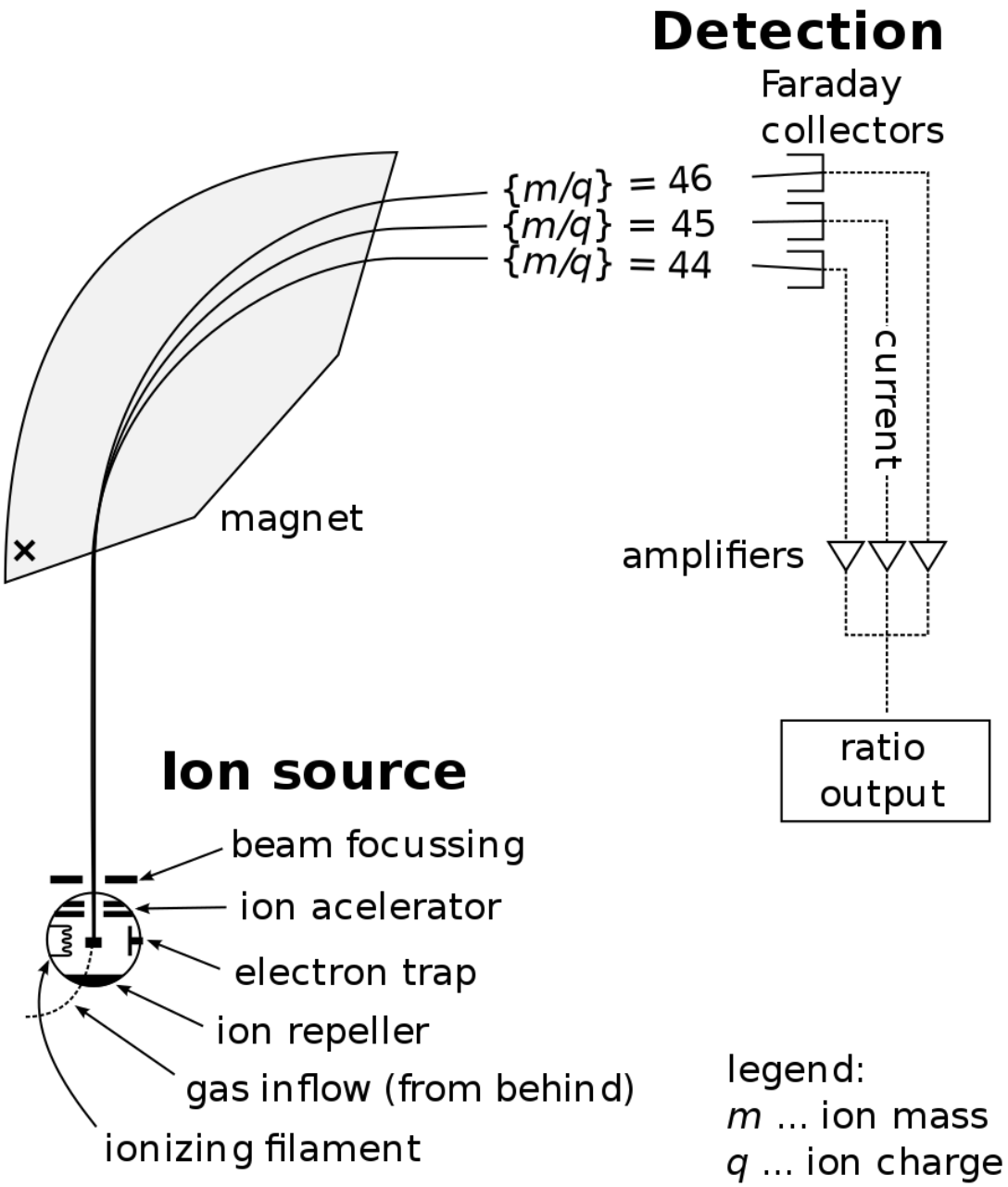


Figure 272: 질량분석기

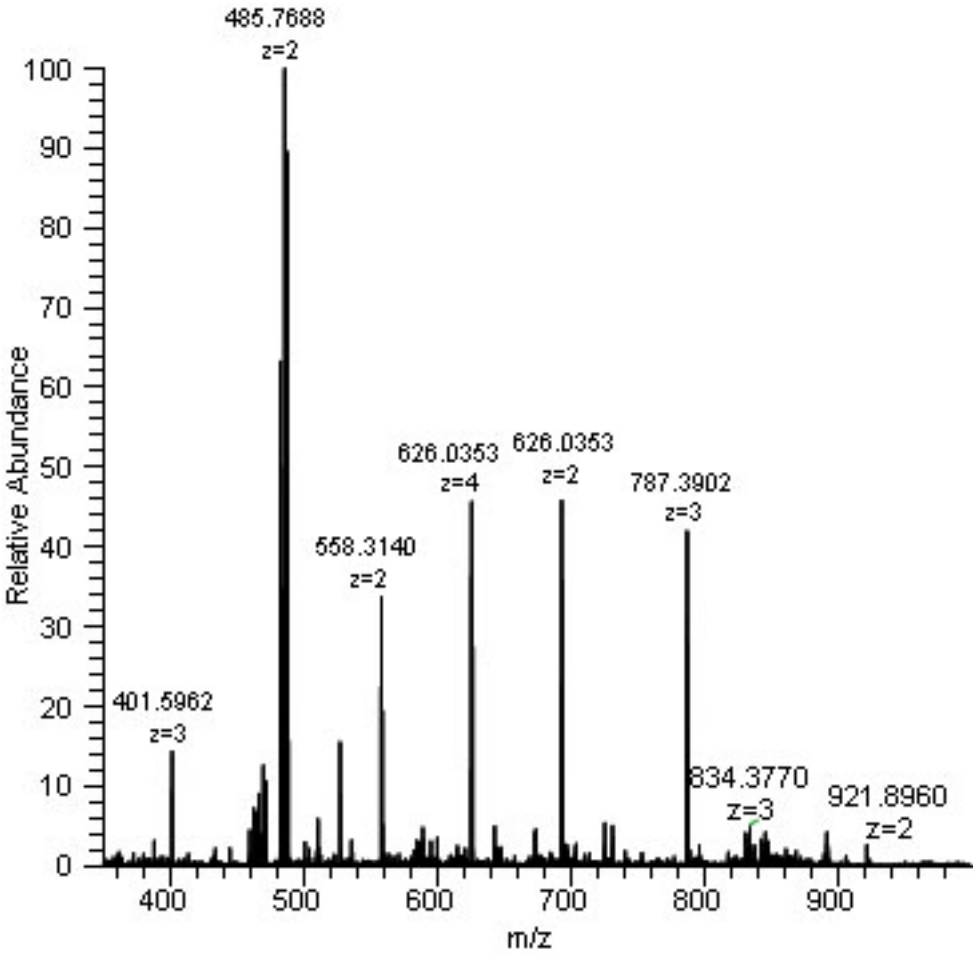


Figure 273: 질량분석법 과정

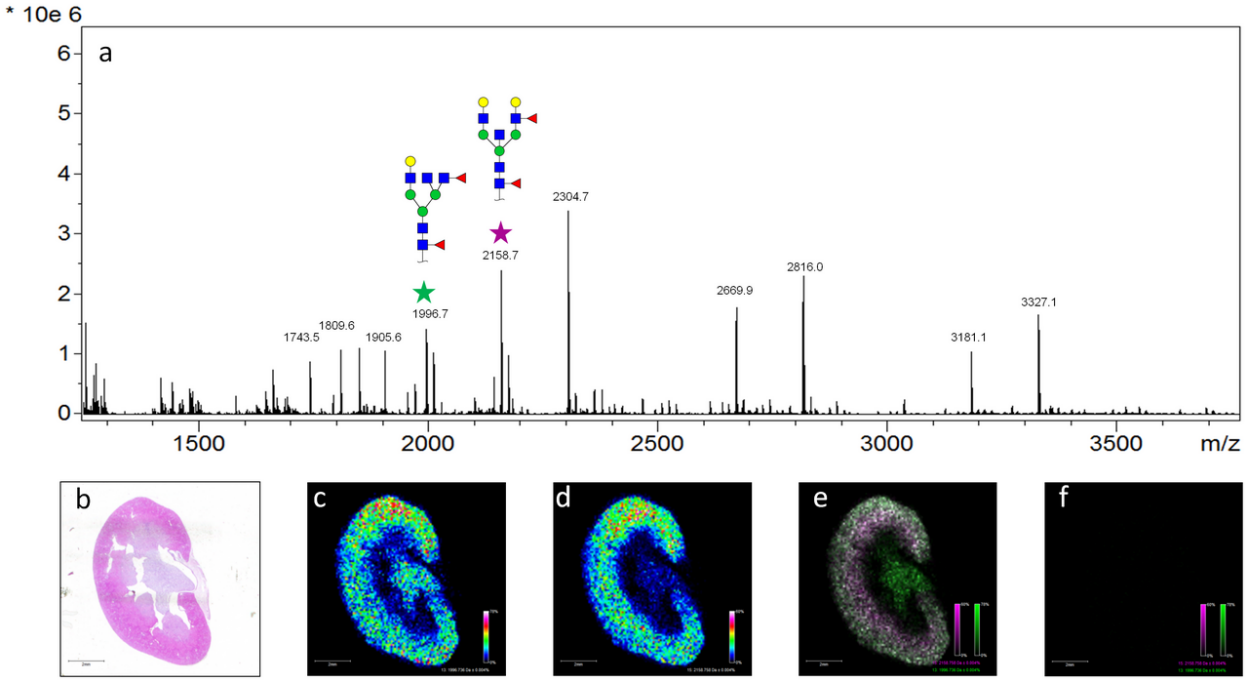


Figure 274: MALDI 이미징

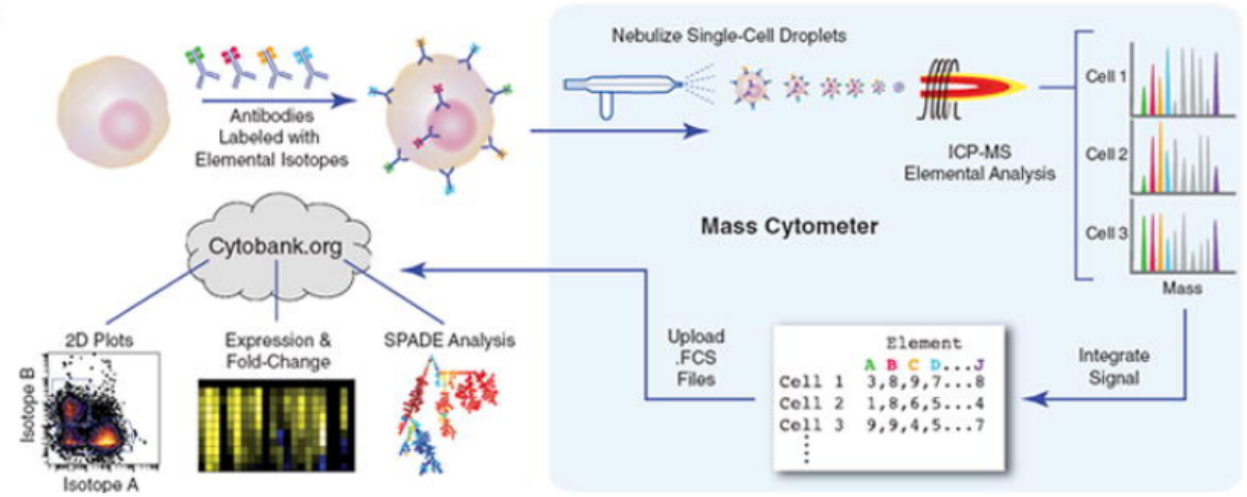


Figure 275: 질량 세포분석법

14.3.2 형광 기반 공간 이미징

형광 기반 공간 단백체학 방법으로는 CODEX(CO-Detection by indexing)가 대표적이다. CODEX는 동일한 조직 절편에서 형광 항체를 순차적으로 염색하고 탈색하는 과정을 반복하여 수십 개의 단백질을 분석하는 기술이다. 각 염색 라운드에서 3-4개의 형광 채널을 사용하여 다수의 마커를 순차적으로 검출하며, 최종적으로 모든 데이터를 통합하여 다중 단백질 발현 지도를 생성한다.

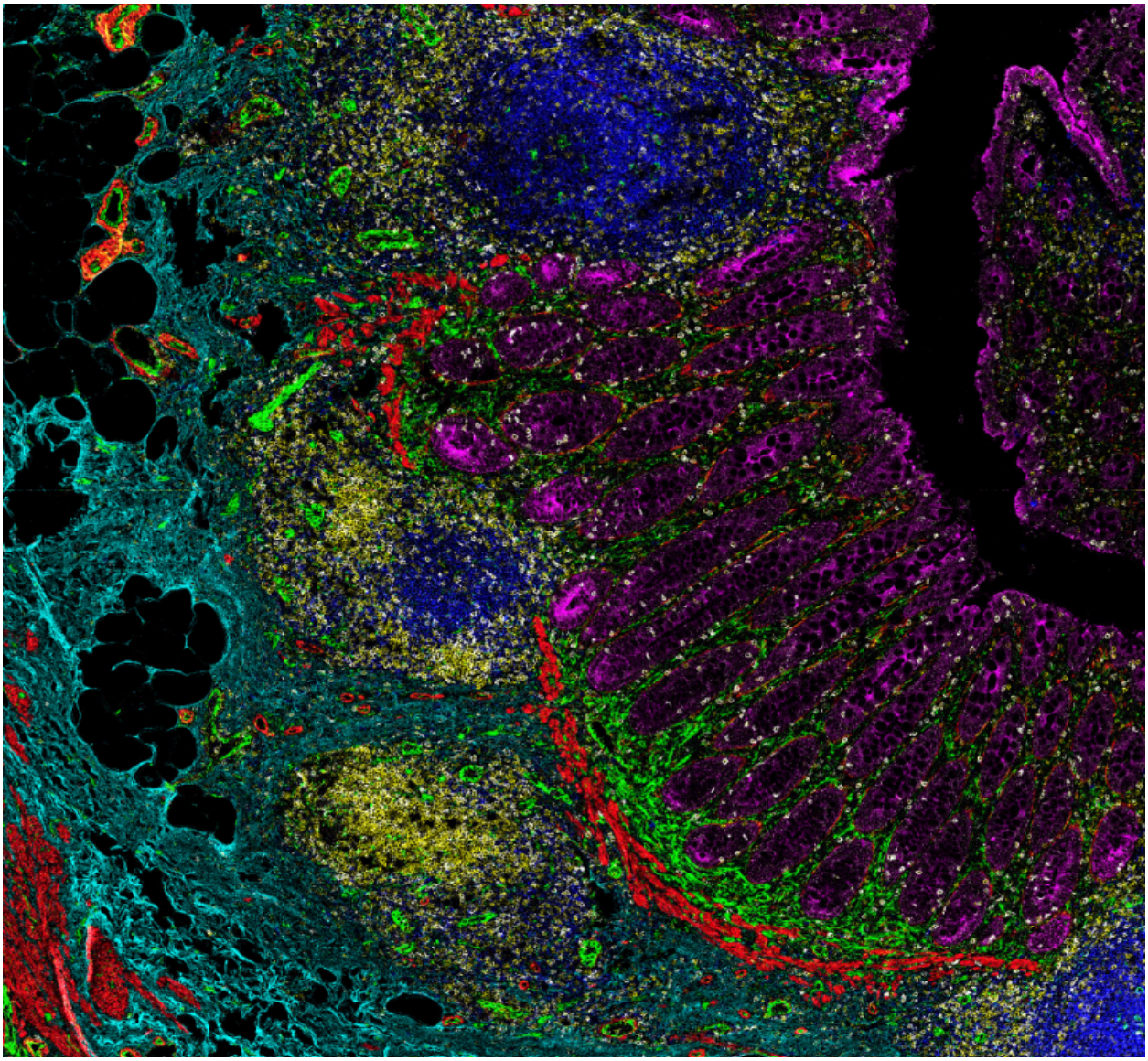


Figure 276: CODEX 결과

Figure 13.11 CODEX 결과 - CODEX 기술을 이용한 다중 단백질 공간 이미징 결과

다중 면역조직화학(Multiplex IHC) 기술들도 발전하여 한 번의 염색으로 여러 단백질을 동시에 관찰할 수 있게 되었다. 이러한 기술들은 면역학, 종양학, 신경과학 등 다양한 분야에서 조직의 분자적 구조를 이해하는 데 활용되고 있으며, 공간 전사체학(spatial transcriptomics) 기술과의 통합을 통해 더욱 포괄적인 분자 지도를 제공하고 있다.

14.4 단일세포 단백체학

14.4.1 질량 세포분석법 (CyTOF)

질량 세포분석법(Cytometry by Time-of-Flight, CyTOF)은 전통적인 플로우 사이토메트리의 한계를 극복한 단일세포 분석 기술이다. 이 기술의 핵심은 형광 염료 대신 희토류 금속 동위원소를 항체에 결합시켜 세포 표면이나 세포내 단백질을 표지하는 것이다. 희토류 금속들은 자연 상태에서 생체 내에 거의 존재하지 않으므로 배경 신호가 극히 낮고, 각각의 질량이 뚜렷이 구별되어 스펙트럼 겹침 현상이 거의 없다.

측정 과정에서 금속 표지된 세포들은 고온의 아르곤 플라즈마를 통과하면서 완전히 이온화된다. 이때 금속 태그들도 이온화되어 질량분석기로 전달되며, 비행시간 측정을 통해 각 금속의 질량을 정확히 구별한다. 이러한 원리를 통해 현재 40개 이상의 매개변수를 동시에 측정할 수 있어, 세포의 복잡한 분자적 특성을 매우 정밀하게 분석할 수 있다.

CyTOF는 단일세포 수준에서 표현형과 기능을 동시에 평가할 수 있는 대표적인 기술이다. 세포 표면 마커를 통한 세포 유형 분류와 동시에 세포내 신호 전달 단백질, 전사 인자, 세포 주기 마커 등을 측정하여 각 세포의 기능 상태를 정밀하게 파악할 수 있다. 이는 특히 면역계의 복잡한 상호작용이나 종양 미세환경의 이질성을 이해하는 데 매우 유용하다.

14.4.2 다중 오믹스 단일세포 분석

CITE-seq(Cellular Indexing of Transcriptomes and Epitopes by sequencing)는 단일세포에서 RNA와 단백질을 동시에 측정하는 혁신적인 기술이다. 이 방법은 DNA 바코드가 결합된 항체를 사용하여 단백질을 표지하고, 10X Genomics와 같은 단일세포 RNA 시퀀싱 플랫폼에서 전사체와 단백질 정보를 함께 획득한다. 이를 통해 같은 세포에서 유전자 발현과 단백질 발현의 관계를 직접적으로 분석할 수 있으며, 전사 후 조절 메커니즘을 이해하는 데 중요한 정보를 제공한다.

REAP-seq(RNA Expression and Protein sequencing)와 같은 유사한 기술들도 개발되어 단일세포에서 다중 오믹스 정보를 획득할 수 있게 되었다. 이러한 전사체-단백체 통합 분석을 통해 유전자 발현과 단백질 발현 간의 불일치를 관찰하고, 번역 후 조절이나 단백질 안정성 등에 대한 정보를 얻을 수 있다.

14.5 단백체학 데이터 분석 및 해석

14.5.1 질량분석 데이터 처리

단백체학에서 생성되는 질량분석 데이터는 고차원적이고 복잡한 특성을 가진다. 각 단백질은 고유한 질량을 가지며, 번역 후 수정(post-translational modification)에 따라 다양한 형태로 존재할 수 있다. 또한 단백질의 농도 범위가 매우 넓어서 고농도 단백질에 의해 저농도 단백질의 신호가 가려지는 동적 범위(dynamic range) 문제가 존재한다.

데이터 전처리 과정에서는 배경 신호 제거, 노이즈 감소, 피크 검출, 질량 보정 등의 단계가 필요하다. 특히 서로 다른 실험 조건이나 시점에서 얻은 데이터를 비교할 때는 정규화(normalization)와 배치 효과 보정이 중요하다.

14.5.2 단일세포 단백체학 데이터 분석

단일세포 단백체학 데이터 분석 과정은 10장과 11장에서 설명한 단일세포 전사체 데이터의 차원 축소, 클러스터링 등과 기본적으로 동일하다. 그러나 단백질 마커의 수가 전사체에 비해 매우 적기 때문에(일반적으로 40-100개), 데이터의 특성에 맞는 정규화 방법이 필요하다.

특히 단일세포 수준의 단백체학 데이터에서는 CLR(Centered Log-Ratio) 정규화와 같은 변형된 형태의 로그 변환을 사용한다. CLR 정규화는 각 세포에서 모든 마커의 기하평균을 구한 후, 각 마커의 값을 이 기하평균으로 나누고 로그 변환하는 방법이다. 이는 낮은 정량화 효율로 인한 기술적 변동을 효과적으로 제거할 수 있다.

정규화된 데이터는 차원 축소 기법(t-SNE, UMAP)과 클러스터링 알고리즘을 통해 세포 유형을 분류하고 발달 궤적을 추적할 수 있다. 이를 통해 세포 분화 과정에서 단백질 발현의 동적 변화나 질병 진행 과정에서 세포 상태 전환을 관찰할 수 있다.

14.5.3 정량적 분석 및 통계

통계적 분석에서는 단백질 발현의 차이를 검정하기 위해 t-test, ANOVA, 또는 더 정교한 선형 모델을 사용한다. 다중 검정 보정을 통해 위양성을 제어하고, 효과 크기(effect size)를 고려하여 생물학적으로 의미 있는 변화를 식별한다.

단일세포 데이터의 경우 세포 간 이질성과 기술적 변이를 고려한 특별한 통계적 방법이 필요하다. 계층적 모델링이나 베이지안 접근법이 이러한 복잡성을 다루는 데 유용하다.

14.5.4 기능적 주석 및 경로 분석

식별된 단백질들의 생물학적 의미를 해석하기 위해 기능적 주석(functional annotation)과 경로 분석(pathway analysis)을 수행한다. 이는 5장에서 설명한 것과 같이 Gene Ontology(GO) 분석을 통해 분자 기능, 생물학적 과정, 세포 구성 요소 관점에서 단백질들을 분류하고, KEGG, Reactome, WikiPathways, Ingenuity Pathway Analysis (IPA) 등의 데이터베이스를 활용하여 관련된 생물학적 경로를 식별한다.

14.6 임상 단백체학의 응용

14.6.1 질병 진단을 위한 단백체학 활용

단백체학 기술들이 실제 병원에서 어떻게 활용되고 있는지 살펴보자. MALDI-TOF, LC-MS, CyTOF, IHC 등의 기술들이 질병 진단과 치료에 중요한 역할을 하고 있다.

혈액 검사를 통한 바이오마커 발굴에서는 질량분석법이 핵심적으로 사용된다. 예를 들어, 전립선암 진단에 사용되는 PSA(전립선 특이 항원) 검사나 난소암 진단의 CA125 검사는 모두 특정 단백질의 양을 정확히 측정하는 기술에 기반한다. 이러한 검사에서 생성되는 데이터는 연속형 수치 데이터로, 정상 범위와 비교하여 질병 여부를 판단한다.

면역조직화학(IHC)은 암 진단에서 필수적인 검사이다. 유방암에서 에스트로겐 수용체(ER), 프로게스테론 수용체(PR), HER2 단백질을 확인하는 것이 대표적인 예시이다. 이 검사 결과는 0점부터 3점까지의 점수로 표현되며, 환자의 치료 방향을 결정하는 중요한 정보가 된다.

단일세포 단백체학 기술도 임상에서 점차 활용되고 있다. CyTOF를 이용하여 환자의 면역세포를 자세히 분석하거나, CITE-seq를 통해 암 조직의 세포 구성을 파악하는 연구가 진행되고 있다.

14.6.2 공간 단백체학의 임상 적용

공간 단백체학 기술들이 임상에서 어떻게 활용되는지 알아보자. 특히 이미징 질량 세포분석법(IMC)과 CODEX와 같은 기술들이 암 진단과 치료에 새로운 가능성을 제시하고 있다.

암 조직을 분석할 때 단순히 암세포만 보는 것이 아니라, 주변 면역세포들과 혈관, 섬유조직 등이 어떻게 배치되어 있는지 파악하는 것이 중요하다. IMC 기술을 사용하면 조직에서 40개 이상의 단백질을 동시에 확인하면서 각각의 위치까지 정확히 알 수 있다. 이를 통해 암세포 주변에 면역세포가 얼마나 많이 있는지, 면역세포들이 활성화되어 있는지 확인할 수 있다.

CODEX 기술도 마찬가지로 여러 단백질을 순차적으로 염색하여 조직의 복잡한 구조를 파악한다. 이러한 정보는 환자가 면역항암제 치료에 잘 반응할지 예측하는 데 도움이 된다.

공간 단백체학 데이터는 일반적인 단백질 데이터와 달리 위치 정보가 추가로 포함되어 있어 분석이 더 복잡하다. 단백질 발현량뿐만 아니라 세포들 간의 거리와 배치 패턴을 함께 고려해야 한다.

14.6.3 단백체학 데이터 분석의 도전과 해결방안

단백체학 데이터를 임상에 적용할 때 여러 어려움이 있다. 특히 질병을 진단하거나 예측하는 데 유용한 단백질(바이오마커)을 찾는 일은 쉽지 않다.

가장 큰 문제는 측정할 수 있는 단백질의 종류는 매우 많은데(수천 개), 실제 환자 수는 상대적으로 적다는 점이다. 예를 들어, LC-MS를 이용해 3000개의 단백질을 측정했지만 환자는 100명뿐이라면, 우연히 관련이 있어 보이는 단백질을 잘못 선택할 위험이 크다.

이 문제를 해결하기 위해서는 통계적 방법들이 중요하다. 정말 중요한 단백질만 선별하거나, 데이터의 차원을 줄이는 방법을 사용해야 한다. 또한 여러 단백질을 동시에 검사할 때는 우연의 일치로 인한 잘못된 결론을 피하기 위해 통계적 보정이 필요하다.

14.6.4 통합적 단백체학 접근법의 미래

단백질 정보만으로는 질병의 전체적인 모습을 이해하기 어렵다. 따라서 유전자, RNA, 단백질, 대사물질 등 여러 종류의 생체 정보를 함께 분석하는 다중 오믹스 접근법이 중요해지고 있다.

예를 들어, 환자의 유전자 변이 정보(GWAS 결과)와 함께 단백질 발현 정보를 분석하면 더 정확한 질병 예측이 가능하다. 또한 단일세포 RNA 데이터와 CITE-seq 데이터를 함께 분석하면 같은 세포에서 유전자 발현과 단백질 발현의 관계를 이해할 수 있다.

서로 다른 종류의 데이터를 함께 분석할 때는 각 데이터의 특성이 다르기 때문에 주의해야 한다. 유전자 데이터는 수만 개의 변수가 있지만 단백질 데이터는 수천 개일 수 있다. 이런 차이를 고려하여 기능적 주석과 경로 분석 방법을 활용한다.

단백질 상호작용 네트워크 정보를 활용하면 더 정확한 분석이 가능하다. 최근에는 인공지능과 기계학습 방법을 이용하여 복잡한 다중 오믹스 데이터의 패턴을 찾는 연구가 활발하게 진행되고 있다.

14.7 결론

단백체학은 생명현상의 실제 실행자인 단백질을 연구하는 분야로서, 질량분석법, 항체 기반 기술, 공간 단백체학, 단일세포 단백체학 등 다양한 기술을 통해 질병의 이해와 치료에 기여하고 있다. 이러한 기술들에서 생성되는 복잡한 데이터는 적절한 정보학적 분석을 통해 의미 있는 생물학적 지식으로 변환되며, 임상에서는 질병 진단과 치료법 선택에 실질적으로 활용되고 있다. 미래의 단백체학은 다중 오믹스 통합 분석과 인공지능 기술의 발전을 통해 개인맞춤 의학 구현에 중요한 역할을 할 것으로 기대된다.

15 고급 유전체학 기법들

유전체학 분야는 개별 생물체의 유전체 해독과 분석에서 시작되어 현재는 복합 미생물 군집 분석, 작물의 유전적 개량, 그리고 유전체의 정밀 편집에 이르기까지 광범위한 영역으로 발전하였다. 본 장에서는 이러한 발전 과정을 반영하여 군유전체학(metagenomics), 식물 유전체학, 그리고 CRISPR/Cas9 유전체 편집 기술이라는 세 가지 분야를 다룬다.

군유전체학은 환경 및 임상 샘플에서 미생물 군집 전체의 유전 정보를 분석하여 배양이 어려운 미생물을 포함한 생태계를 이해하는 기법이다. 식물 유전체학에서는 배수성(polyploidy)이 식물의 진화와 농업적 개량에 미치는 영향을 살펴보며, 특히 밀, 딸기, 감자 등 주요 작물이 어떻게 배수화를 통해 수확량과 품질이 향상되었는지를 다룬다. 마지막으로 유전체 편집 기술은 단순히 유전체를 읽고 분석하는 차원을 넘어서 유전체를 정밀하게 수정할 수 있는 혁신적 도구로, CRISPR/Cas9 시스템을 중심으로 그 원리와 응용을 다룬다.

이러한 기술들은 인간 마이크로바이옴 연구, 작물 개량, 그리고 유전 질환 치료에서 실제 응용되고 있으며, 현대 생명과학과 의학 발전의 핵심 도구로 자리잡고 있다.

15.1 군유전체학(Metagenomics)

15.1.1 군유전체학의 정의와 의의

전통적인 미생물학 연구는 단일 균주를 순수 배양하여 분석하는 방법에 의존해왔다. 그러나 자연 환경에 존재하는 미생물의 99% 이상이 현재까지 실험실 조건에서 배양되지 않는다는 사실이 밝혀지면서, 배양에 의존하지 않는 새로운 접근법이 필요하게 되었다.

군유전체학(Metagenomics)은 환경이나 임상 샘플에서 추출한 모든 미생물의 유전 물질을 배양 과정 없이 직접 분석하는 학문 분야이다. 이 접근법은 샘플 내에 존재하는 전체 미생물 군집의 집합적 유전체를 의미하는 메타유전체(metagenome)를 대상으로 한다. 군유전체학을 통해 배양 불가능한 미생물을 포함한 전체 미생물 생태계의 구조와 기능을 파악할 수 있으며, 개별 미생물이 아닌 군집 수준에서의 대사 경로와 생태학적 상호작용을 이해할 수 있다.

Figure 14.1 군유전체학 – 환경 샘플에서 미생물 군집 전체의 유전 정보를 분석하는 군유전체학 개념도

이 분야의 발전은 차세대 시퀀싱(NGS) 기술의 급속한 발전과 생물정보학 분석 도구의 개발에 크게 의존한다. 고처리량 시퀀싱을 통해 대용량의 메타유전체 데이터를 생산할 수 있게 되었으며, 복잡한 미생물 군집 데이터를 분석할 수 있는 알고리즘과 소프트웨어가 지속적으로 개발되고 있다. 또한 미생물 유전자와 기능에 대한 포괄적인 참조 데이터베이스의 확장과 대용량 데이터 처리를 위한 컴퓨팅 인프라의 발달도 군유전체학 연구의 활성화에 기여하고 있다.

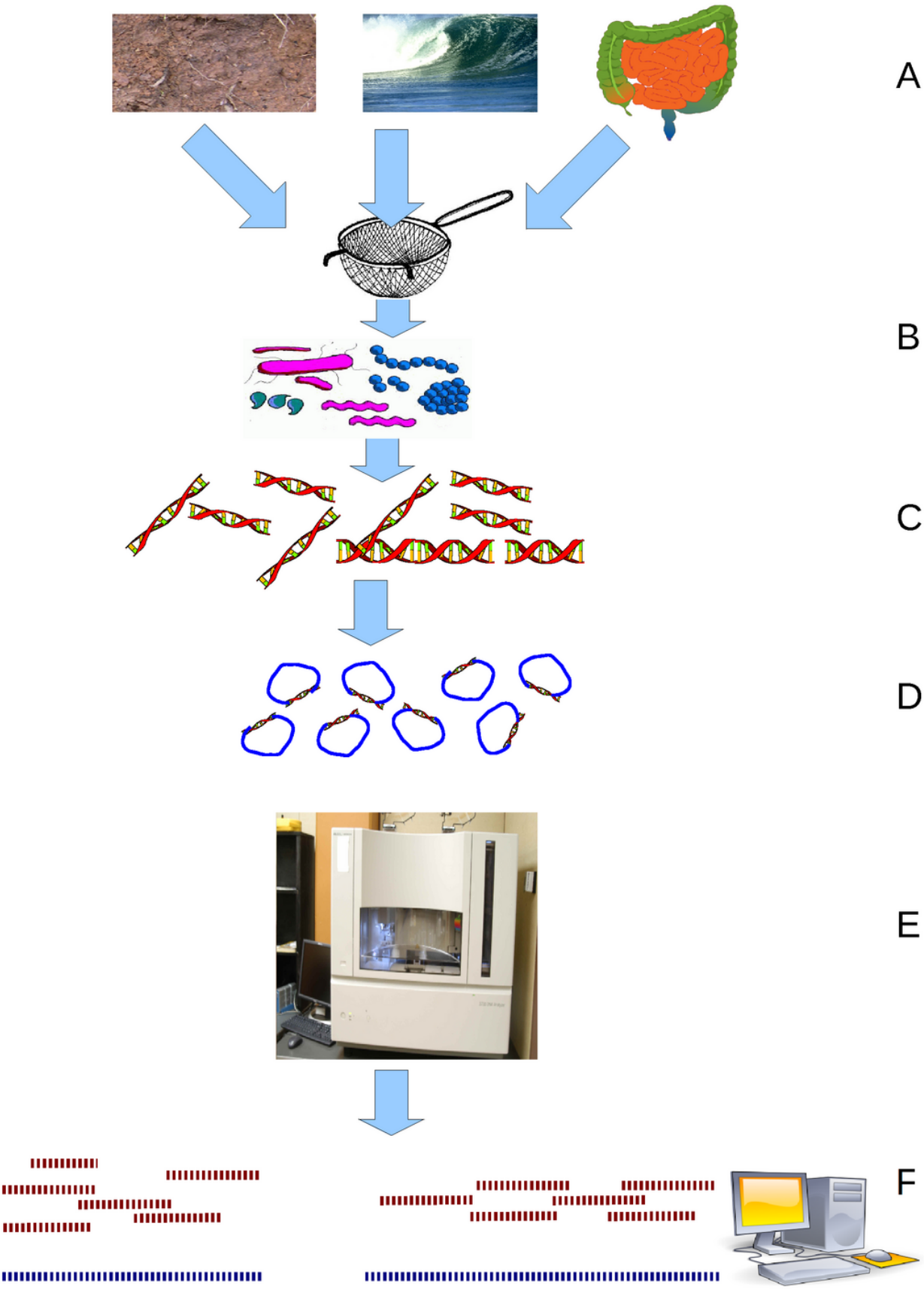


Figure 277: 군유전체학

15.1.2 군유전체학 연구 접근법

군유전체학 연구는 크게 두 가지 주요 접근법으로 구분된다. 첫 번째는 16S rRNA 유전자에 기반한 분석 방법이고, 두 번째는 전체 메타유전체를 대상으로 하는 샷건 메타유전체학이다.

15.1.2.1 16S rRNA 유전자 기반 분석 16S rRNA 유전자 기반 분석은 분류학적 마커를 이용하여 미생물 군집의 구성을 파악하는 방법이다. 이 접근법은 비교적 비용이 적게 들고 분석이 간단하여 대규모 코호트 연구에 적합하다. 16S rRNA 유전자는 모든 세균과 고세균에 존재하면서도 종 특이적 차이를 보이는 영역을 포함하고 있어 미생물 군집 분석의 표준 마커로 사용된다.

이 방법을 통해 미생물 군집의 분류학적 구성, 다양성 지수, 그리고 군집 간 유사성을 비교할 수 있다. 그러나 기능적 정보는 제한적이며, 유전자 복사본 수의 차이로 인한 정량적 편향이 존재할 수 있다는 한계가 있다.

15.1.2.2 샷건 메타유전체학 샷건 메타유전체학은 샘플 내 모든 미생물의 전체 DNA를 무작위로 시퀀싱하는 방법이다. 이 접근법은 분류학적 정보뿐만 아니라 기능적 유전자의 존재와 풍부도를 동시에 파악할 수 있어 미생물 군집의 대사적 잠재력을 이해하는 데 유용하다.

샷건 메타유전체학을 통해 항생제 저항성 유전자, 독성 인자, 대사 경로 등을 직접 분석할 수 있으며, 메타유전체 조립 유전체(Metagenome-Assembled Genomes, MAGs)를 구성하여 배양되지 않은 미생물의 유전체 정보를 얻을 수도 있다. 과거에는 높은 비용이 단점이었으나, 최근 시퀀싱 비용의 급격한 감소로 이 방법의 활용이 크게 증가하고 있다.

15.1.3 16S rRNA 유전자 분석

16S ribosomal RNA 유전자는 16S rRNA 유전자 기반 분석 방법의 핵심 도구로, 미생물 군집의 분류학적 구성을 파악하는 데 가장 널리 사용되는 마커이다. 이 유전자는 세균과 고세균의 작은 리보솜 소단위체(30S)에 위치하며, 단백질 합성에서 중요한 기능을 담당한다.

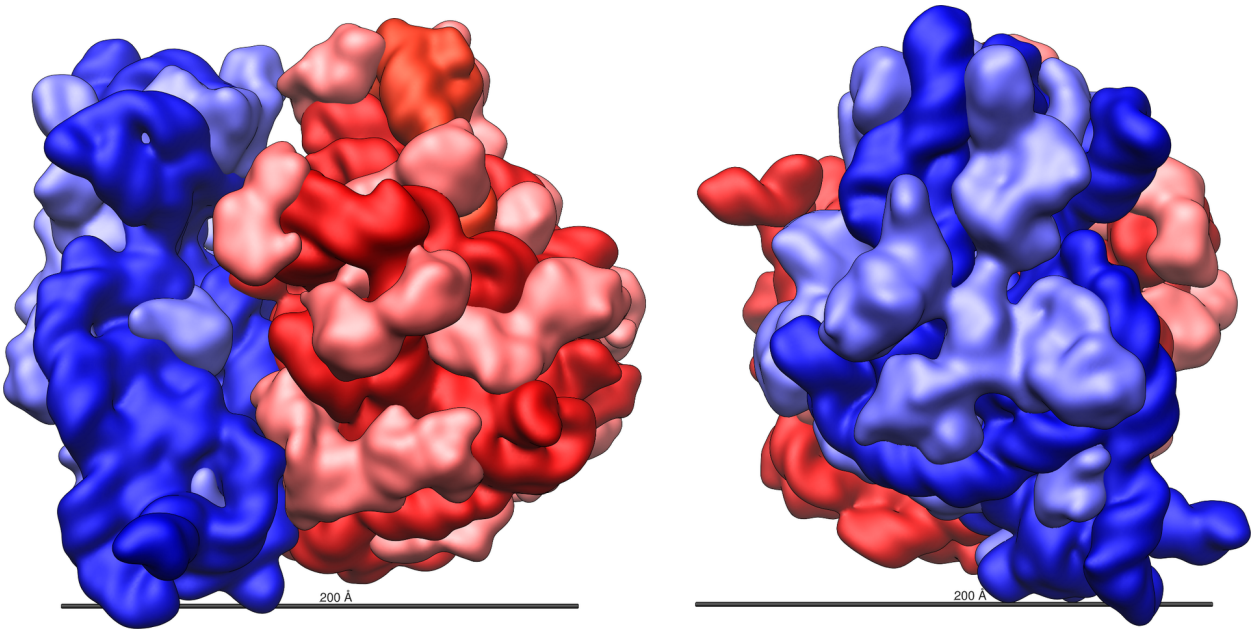


Figure 278: 리보솜 구조

Figure 14.2 리보솜 구조 – 16S rRNA가 위치한 리보솜의 3차원 구조와 단백질 합성 과정

16S rRNA 유전자가 미생물 분류의 표준 마커로 사용되는 이유는 몇 가지 독특한 특성 때문이다. 먼저 이 유전자는 모든 세균과 고세균에 보편적으로 존재하면서도, 리보솜 기능의 중요성으로 인해 진화 과정에서 비교적 보존되어 있다. 동시에 약 1,500 염기쌍의 적절한 길이를 가지고 있어 PCR 증폭과 시퀀싱 분석에 적합하다.

특히 중요한 것은 16S rRNA 유전자 내에 보존 영역과 가변 영역이 교대로 나타나는 구조이다. 9개의 가변 영역(V1–V9)은 종 특이적 차이를 보이면서도, 그 사이의 보존 영역은 프라이머 설계의 기준점을 제공한다. 현재 대부분의 연구에서는 V3–V4 또는 V4 영역을 표적으로 하는 시퀀싱이 주로 사용되며, 이를 통해 속 수준에서 종 수준까지의 분류학적 구분이 가능하다.

분석 과정에서는 먼저 환경 샘플에서 추출한 DNA를 대상으로 16S rRNA 유전자 특이적 프라이머를 이용한 PCR 증폭을 수행한다. 증폭된 산물은 차세대 시퀀싱을 통해 분석되며, 얻어진 서열 데이터는 참조 데이터베이스와 비교하여 분류군을 할당한다.

RNA World 가설

DNA와 단백질이 등장하기 이전의 원시 생명체는 RNA만으로 자기복제와 대사를 수행했을 것으로 추정된다. 리보솜 RNA는 현재도 펩타이드 결합 형성에 직접 관여하는 리보자임(ribozyme) 활성을 가지고 있어 이 가설을 뒷받침한다. 특히 16S rRNA는 진화적으로 고도로 보존된 영역과 가변 영역을 모두 포함하고 있어, 이러한 진화적 역사를 간직한 분자로서 계통분류학적 연구에 핵심적인 역할을 한다.

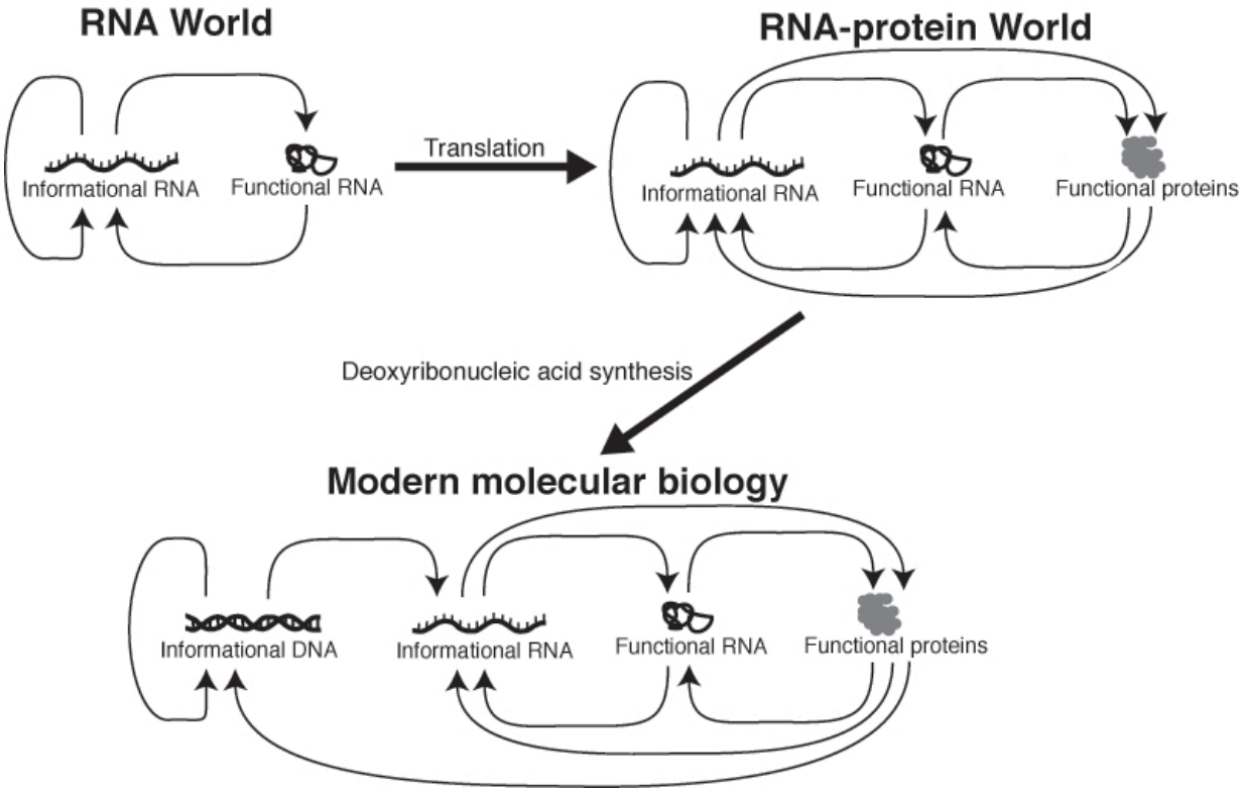


Figure 279: RNA World 가설

Figure 14.3 RNA World 가설 – DNA와 단백질 이전의 원시 생명체에서 RNA의 중심적 역할

15.1.4 미생물 다양성 지표

미생물 군집의 다양성은 생태학적 관점에서 세 가지 계층적 수준으로 평가된다. 이러한 다양성 개념은 1960년대 생태학자 Robert Whittaker에 의해 제안된 것으로, 현재 군유전체학 연구에서 미생물 군집의 특성을 이해하는 핵심 도구로 활용된다.

15.1.4.1 알파 다양성 (α-diversity) 알파 다양성은 단일 샘플 또는 특정 서식지 내에서 관찰되는 미생물의 다양성을 의미한다. 이는 해당 환경에 서식하는 미생물 종의 수와 그들의 상대적 풍부도를 종합적으로 반영하는 지표이다. 알파 다양성이 높다는 것은 그 환경에 많은 종류의 미생물이 비교적 균등하게 분포하고 있음을 의미하며, 일반적으로 생태계의 안정성과 건강성을 나타내는 지표로 해석된다.

15.1.4.2 베타 다양성 (β-diversity) 베타 다양성은 서로 다른 샘플이나 서식지 간의 미생물 군집 구성의 차이를 나타내는 지표이다. 이는 공간적 또는 시간적으로 분리된 환경들 사이에서 미생물 군집이 얼마나 다른지를 정량화한다. 베타 다양성이 높다는 것은 비교하는 샘플들 간에 미생물 군집 구성이 크게 다르다는 것을 의미하며, 환경 변화나 지리적 거리에 따른 미생물 군집의 분화 정도를 이해하는 데 중요하다.

15.1.4.3 감마 다양성 (γ-diversity) 감마 다양성은 연구 대상 전체 지역이나 생태계에서 관찰되는 총 다양성을 의미한다. 이는 개별 서식지의 알파 다양성과 서식지 간 베타 다양성을 통합한 개념으로, 전체 연구 영역에서 발견되는 모든 미생물 종의 총합을 나타낸다. 감마 다양성은 대규모 생태계나 지리적 영역에서의 미생물 다양성 패턴을 이해하고, 보전 전략을 수립하는 데 중요한 정보를 제공한다.

15.1.5 군유전체학의 응용 사례

군유전체학은 다양한 환경과 임상 분야에서 활용되고 있다. 대표적인 예로 인간 마이크로바이옴 연구에서는 장내 미생물 군집 분석을 통해 소화기 질환과의 연관성을 밝히고 있으며, 구강 미생물 분석으로는 치주염과 심혈관 질환의 관계를 발견하였다. 특히 폐에서는 과거 무균 장기로 여겨졌던 것과 달리 Prevotella, Streptococcus 등의 미생물이 발견되어 호흡기 질환 연구에 새로운 관점을 제시하고 있다.

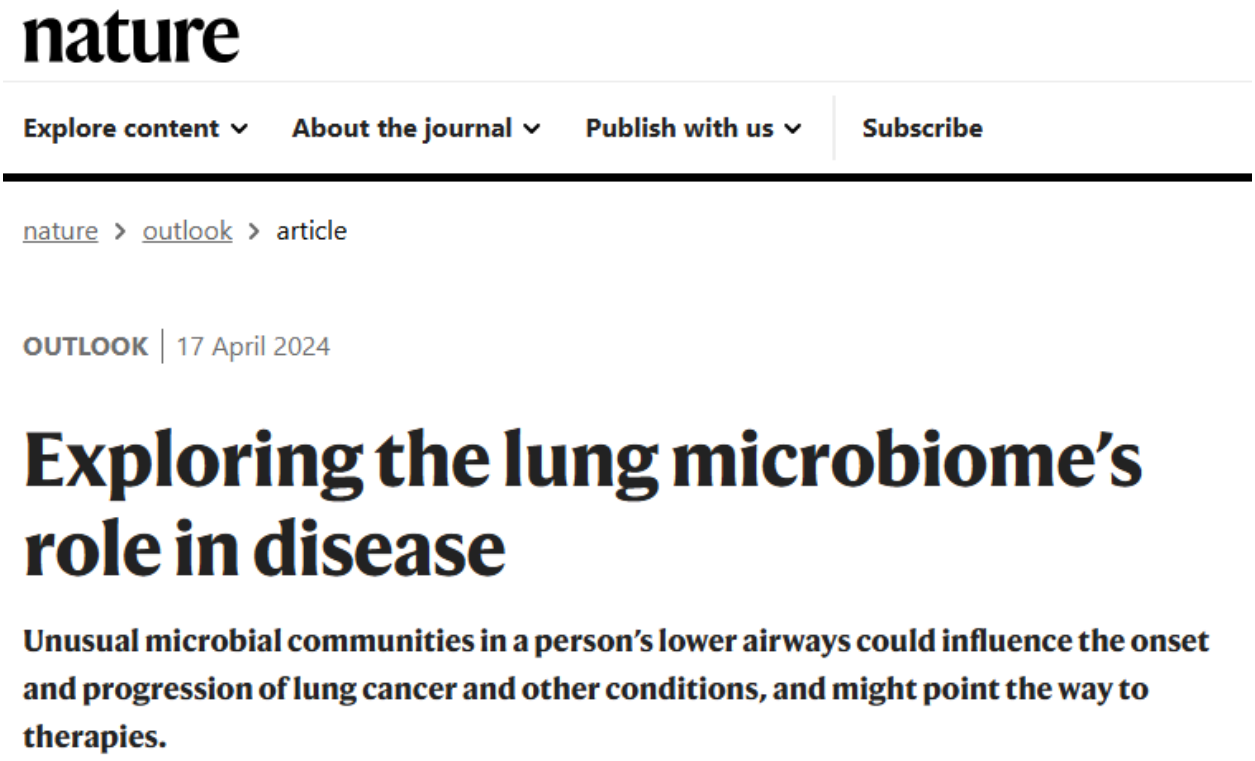


Figure 280: 폐 마이크로바이옴

Figure 14.4 폐 마이크로바이옴 – 과거 무균 장기로 여겨졌던 폐에서 발견된 미생물 군집

15.1.5.1 장-뇌 축 (Gut-Brain Axis) 장-뇌 축은 장내 미생물군집과 중추신경계 간의 복잡한 양방향 소통 네트워크를 의미한다. 이 축은 신경계, 내분비계, 면역계를 통해 작동하며, 장내 미생물이 뇌 기능과 행동에 직접적인 영향을 미칠 수 있음을 보여준다. 장내 미생물은 세로토닌, 도파민, GABA 등의 신경전달물질을 생산하여 기분과 인지 기능을 조절한다. 특히 장내 세로토닌의 90% 이상이 장크롬친화세포에서 생산되며, 이는 미생물 대사산물의 영향을 받는다.

미생물이 생산하는 단쇄지방산, 특히 부티레이트는 혈뇌장벽을 통과하여 미세아교세포의 활성을 조절하고 신경염증을 억제한다. 또한 미주신경을 통한 직접적인 신호전달과 시상하부-뇌하수체-부신피질 축(HPA axis)의 조절을 통해

스트레스 반응과 감정 상태에 영향을 미친다. 이러한 메커니즘을 통해 장내 미생물 불균형은 우울증, 불안장애, 자폐스펙트럼장애, 심지어 알츠하이머병과 파킨슨병 등의 신경퇴행성 질환과도 연관성이 보고되고 있다.

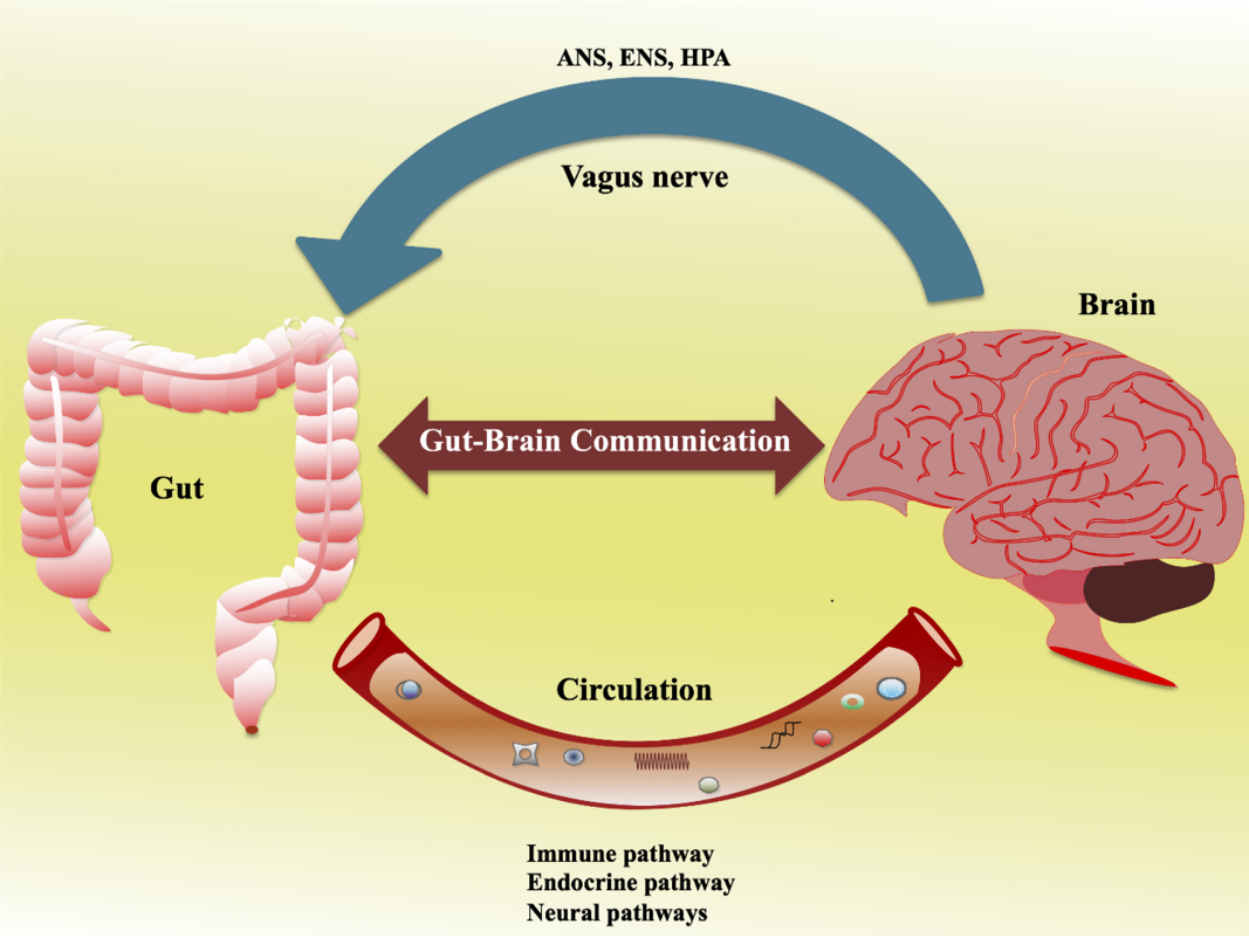


Figure 281: 장-뇌 축

Figure 14.5 장-뇌 축 - 장내 미생물과 중추신경계 간의 양방향 소통 네트워크

Figure 14.6 장-뇌 축 상세 - 미생물 대사산물이 뇌 기능에 미치는 영향의 분자적 메커니즘

환경 분야에서는 해양의 극한 환경에서 새로운 항생물질 생산 미생물을 발견하거나, 토양 미생물 분석을 통해 탄소 순환과 기후 변화의 관계를 이해하고 있다. 또한 식품 발효 과정의 미생물 군집 변화를 추적하여 발효 품질을 개선하고, 농업에서는 식물 근권 미생물이 식물 생장에 미치는 긍정적 효과를 규명하여 친환경 농법 개발에 기여하고 있다.

15.1.5.2 분변 미생물 이식술 (Fecal Microbiota Transplantation, FMT) 분변 미생물 이식술은 건강한 공여자의 분변에서 추출한 미생물 군집을 환자에게 이식하여 장내 미생물 생태계를 복원하는 치료법이다. 이는 항생제 사용으로 인한 장내 미생물 불균형이나 병원성 미생물의 과증식을 치료하는 데 활용된다.

Figure 14.7 분변 미생물 이식술 개요 - 건강한 공여자의 장내 미생물을 환자에게 이식하는 FMT 과정

최근 연구에서는 FMT가 암 치료 분야에서도 주목받고 있다. 특히 면역관문억제제(immune checkpoint inhibitor) 치료에 반응하지 않는 암 환자들에게서 FMT를 통한 장내 미생물 개선이 항암 면역반응을 증강시킬 수 있다는 연구 결과들이 보고되고 있다. 예를 들어 특정 미생물 종들이 PD-1/PD-L1 억제제의 효능을 향상시킬 수 있는 가능성이 보고되고 있다.

Figure 14.8 분변 이식술 응용 - FMT를 통한 장내 미생물 생태계 복원에 따른 PD-1 억제제 반응성 향상 가능성 리포트

그러나 공여자 선별의 까다로움, 전파 감염의 위험성, 그리고 아직 완전히 규명되지 않은 장기적 안전성 등의 한계가

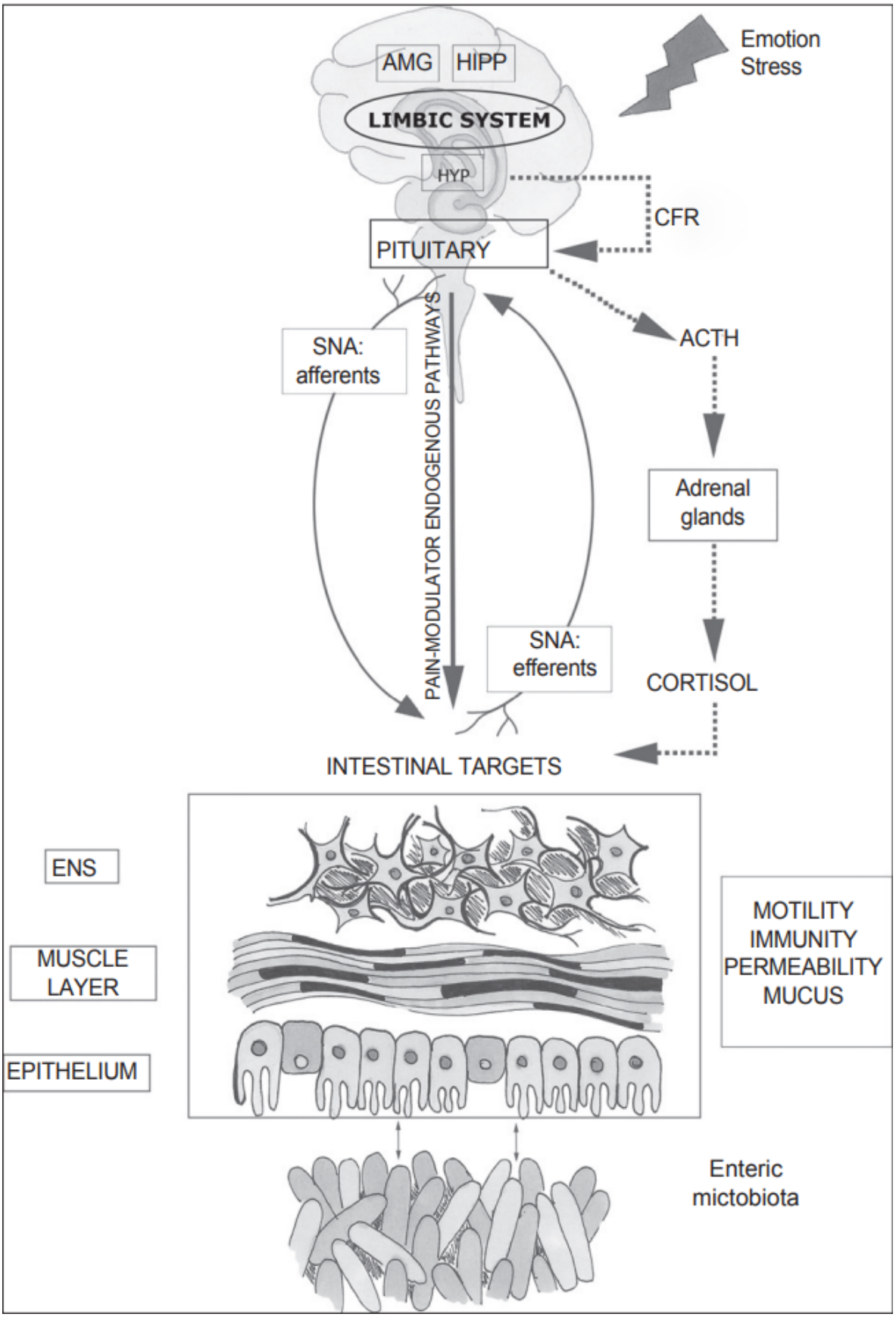


Figure 282: 장-뇌 축 상세

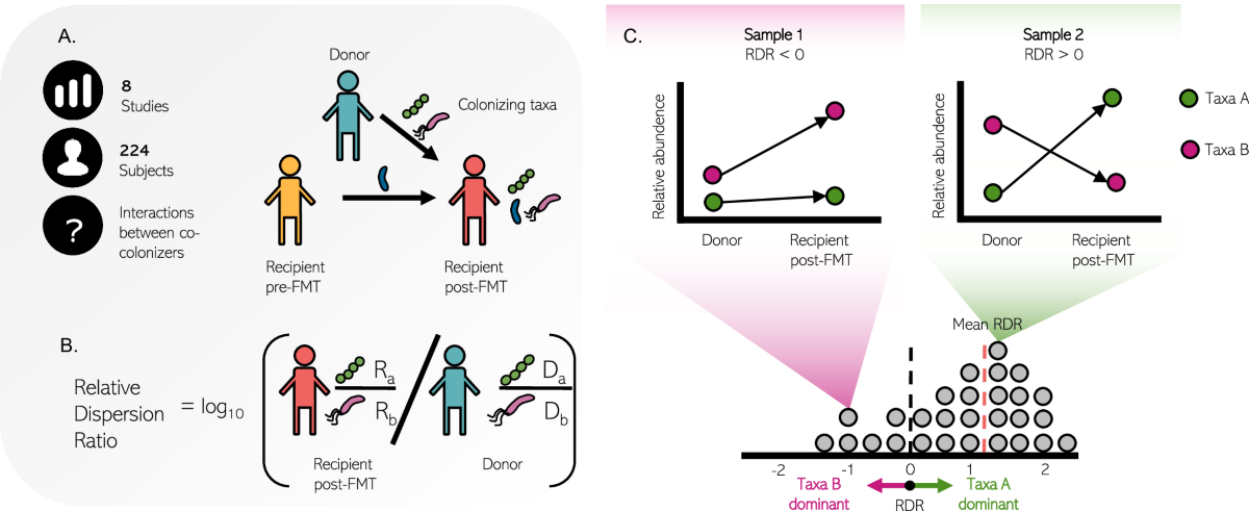


Figure 283: 분변 미생물 이식술 개요

Signal Transduction
and Targeted Therapy

Research Highlight | Open access | Published: 08 May 2021

Fecal microbiota transplantation: can it circumvent resistance to PD-1 blockade in melanoma?

Lisa Derosa ✉ & Laurence Zitvogel ✉

Signal Transduction and Targeted Therapy **6**, Article number: 178 (2021) | Cite this article

4895 Accesses | **2** Altmetric | Metrics

Figure 284: 분변 이식술 응용

존재한다. 특히 2019년에는 분변 이식술을 시술받은 환자의 사망 사례가 발생하면서 더 조심스럽게 접근해야 할 필요성이 조명되고 있다.

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › nature briefing › article

NATURE BRIEFING | 14 June 2019

Daily briefing: Fecal transplant linked to patient's death

FDA halts fecal transplant clinical trials, how to break science's language barrier and Neolithic farmers built artificial islands.

By Flora Graham

Sign up for Nature Briefing

Figure 285: 분변 이식술 위험성

Figure 14.9 분변 이식술 사망 사례 – FMT 치료 과정에서 발생할 수 있는 안전성 문제

15.2 식물 유전체학

식물 유전체학은 식물의 유전체 구조와 기능을 연구하는 분야로, 동물과는 다른 독특한 특성들을 다룬다. 식물은 동물에 비해 훨씬 높은 빈도로 배수성을 나타내며, 이는 식물의 진화와 농업적 개량에 중요한 역할을 한다. 또한 식물은 고정된 위치에서 다양한 환경 변화에 적응해야 하므로, 유전체 수준에서 유연성과 다양성을 유지하는 독특한 전략을 발달시켰다. 본 절에서는 식물 유전체학의 핵심 개념인 배수성을 중심으로, 이것이 식물의 진화와 농업에 미치는 영향을 살펴본다.

15.2.1 배수성

배수성(Polyploidy)은 식물에서 매우 흔하게 관찰되는 현상으로, 정상적인 2배체보다 많은 염색체 세트를 가지는 상태를 의미한다. 식물에서 배수성은 진화, 육종, 그리고 농업적 개선에 핵심적인 역할을 한다.

15.2.1.1 주요 작물의 배수성 사례 **곡물류**: – **밀 (Triticum aestivum)**: 6배체 (2n = 6x = 42) – **귀리 (Avena sativa)**: 6배체 (2n = 6x = 42) – **기장 (Setaria italica)**: 4배체 (2n = 4x = 36)

과일류: – **딸기 (Fragaria × ananassa)**: 8배체 (2n = 8x = 56) – **바나나 (Musa spp.)**: 3배체 (2n = 3x = 33) – **키위 (Actinidia deliciosa)**: 6배체 (2n = 6x = 174)

채소류: - **감자 (Solanum tuberosum)**: 4배체 (2n = 4x = 48) - **고구마 (Ipomoea batatas)**: 6배체 (2n = 6x = 90)

기호품: - **커피 (Coffea arabica)**: 4배체 (2n = 4x = 44) - **담배 (Nicotiana tabacum)**: 4배체 (2n = 4x = 48)

15.2.2 배수성의 진화적 의의

배수성은 식물이 환경 변화에 적응하고 진화하는 데 중요한 역할을 한다. 동일한 유전자의 여러 복사본을 보유함으로써 식물은 유전적 유연성을 확보하게 된다. 예를 들어 하나의 유전자에 해로운 돌연변이가 발생하더라도 다른 복사본들이 정상 기능을 유지할 수 있어, 치명적인 변이의 영향을 완화하는 완충 역할을 한다.

또한 배수성은 식물의 환경 스트레스 내성을 향상시킨다. 가뭄, 염분, 극한 온도와 같은 무생물학적 스트레스뿐만 아니라 병원체와 해충 같은 생물학적 스트레스에 대한 저항성도 증가한다. 이는 다양한 대립유전자 조합이 환경 변화에 대한 다양한 반응을 가능하게 하기 때문이다.

배수성은 종분화에도 중요한 기여를 한다. 특히 홀수 배수체의 경우 감수분열 과정에서 염색체 분리에 문제가 생겨 생식적 격리가 일어나며, 이는 새로운 종의 형성으로 이어질 수 있다. 서로 다른 종 간의 교배로 생긴 잡종이 배수화를 통해 생식 능력을 회복하는 경우도 있어, 식물 진화에서 배수성은 핵심적인 메커니즘으로 작용한다.

15.2.3 재배종과 야생종의 비교: 딸기의 사례

딸기는 배수성이 농작물 개량에 어떤 혁신적 변화를 가져올 수 있는지를 보여주는 훌륭한 사례이다. 우리가 마트에서 구입하는 큰 딸기와 산에서 자라는 작은 야생 딸기의 차이는 단순히 품종의 차이가 아니라, 염색체 수의 근본적인 차이에서 비롯된다.

야생 딸기는 이배체(2n = 2x = 14)로, 인간과 같이 각 염색체가 2개씩 존재하는 일반적인 형태로, 이들의 열매는 지름이 5–10mm 정도로 매우 작은 편이다. 야생 딸기는 종자로도 번식하고 러너(runner)라는 줄기를 통해서도 번식할 수 있어, 자연 환경에서의 생존과 확산에 유리한 특성을 가지고 있다.

현재 우리가 먹는 재배 딸기는 8배체(2n = 8x = 56)로, 야생 딸기보다 4배나 많은 염색체를 가지고 있다. 이러한 배수화의 결과로 딸기의 크기는 직경 20–40mm로 야생 딸기의 3–4배가 되었다. 더 놀라운 것은 개체당 수확량이 10–20배나 증가했다는 점이다. 또한 과실의 품질이 균일해져서 상업적 재배와 유통에 적합하게 되었으며, 단단한 과육으로 인해 운송과 저장이 용이해졌다.

Figure 14.10 재배종과 야생종 딸기 – 8배체 재배 딸기와 2배체 야생 딸기의 크기와 특성 비교

Figure 14.11 딸기 재배 역사 – 딸기의 배수화 과정과 농업적 개량을 통한 품종 발전 역사

딸기의 8배체화는 단순히 크기만 키운 것이 아니다. 염색체 수가 증가하면서 유전자의 복사본이 많아져, 환경 변화에 대한 적응력이 향상되었다. 이로 인해 다양한 기후 조건에서도 안정적으로 재배할 수 있게 되었고, 병해충에 대한 저항성도 증가했다. 또한 잼이나 주스 등 가공품 제조에도 적합한 특성을 갖게 되어, 딸기 산업 전체의 발전을 이끌었다. 이는 배수성이 어떻게 하나의 작물을 완전히 변화시켜 현대 농업의 핵심 작물로 만들 수 있는지를 보여주는 대표적인 예라고 할 수 있다.

15.2.4 배수성과 현대 농업

배수성은 자연에서 자발적으로 일어나기도 하지만, 현대에는 인위적으로 유도하여 새로운 품종을 개발하는 데 활용되고 있다. 자연적으로 발생한 배수성 식물들은 오랜 진화 과정을 거쳐 안정화되어 있어 특정 환경에 잘 적응된 특성을 보인다. 반면 인위적으로 유도된 배수성은 빠른 시간 내에 원하는 특성을 얻을 수 있다는 장점이 있다.

인위적 배수성 유도의 대표적인 방법은 콜히친(colchicine)이라는 화학물질을 사용하는 것이다. 콜히친은 세포 분열 과정에서 방추사 형성을 억제하여 염색체가 두 배로 복제되도록 유도한다. 이 방법은 1930년대부터 사용되어 왔으며, 현재도 많은 배수성 품종 개발에 활용되고 있다. 또한 조직 배양 기술을 통해서도 배수성 식물을 얻을 수 있는데, 이는 콜히친 처리보다 안전하고 효율적인 방법으로 평가받고 있다.

배수성의 농업적 활용은 매우 다양하다. 가장 잘 알려진 것은 씨 없는 과일 생산이다. 홀수 배수체는 감수분열 과정에서 정상적인 생식 세포를 만들 수 없어 씨가 생기지 않는데, 이를 이용해 씨 없는 수박, 포도 등을 만들 수 있다. 또한 배수성 식물은 일반적으로 환경 스트레스에 강하고 수확량이 많아, 기후 변화에 대응하는 새로운 품종 개발에도 중요한 역할을 하고 있다.

Figure 14.12 재배종과 야생종 감자 – 4배체 재배 감자와 2배체 야생 감자의 크기 차이

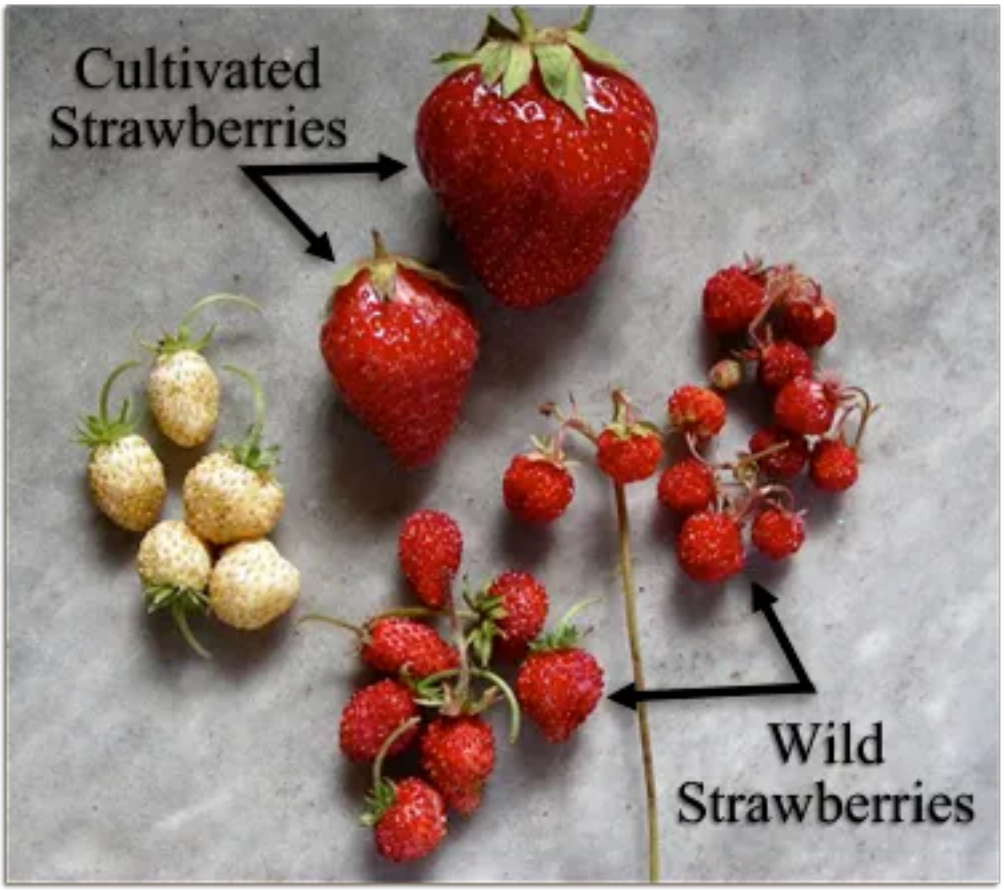


Figure 286: 재배종과 야생종 딸기

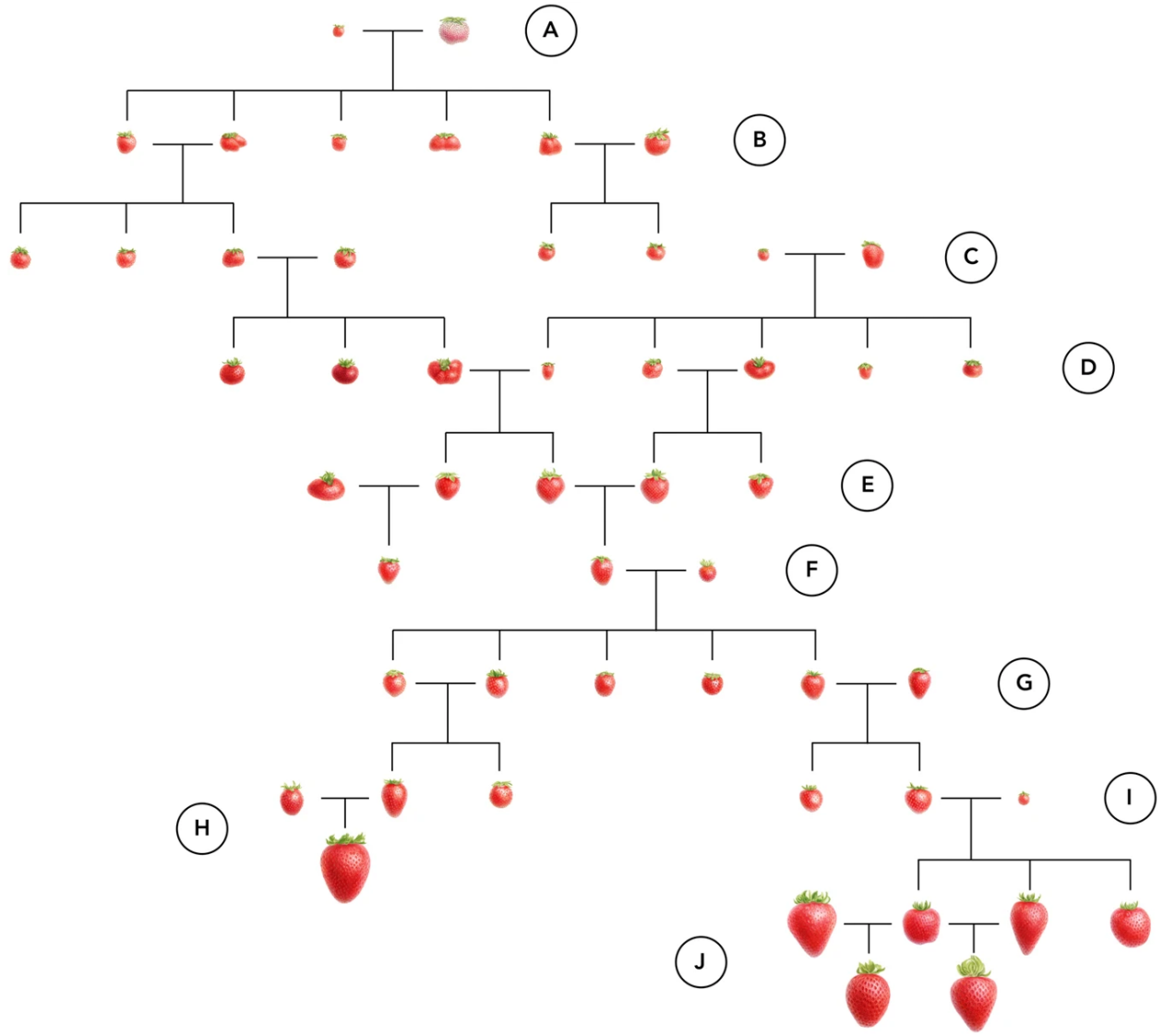


Figure 287: 딸기 재배 역사

Figure 288: 재배종과 야생종 감자

Figure 289: 야생 바나나

Figure 14.13 야생 바나나 – 2배체 야생 바나나

15.3 유전체 편집 기술

유전체 편집 기술은 살아있는 세포에서 DNA를 정밀하게 수정할 수 있는 분자 도구들을 의미한다. 이 기술은 유전 질환 치료와 작물 개량 분야에서 혁신적인 변화를 가져오고 있다. 임상 분야에서는 겸상적혈구병, 베타지중해빈혈과 같은 단일 유전자 질환의 치료에 성공적으로 적용되고 있으며, 유전성 실명, 면역결핍증 등의 치료에도 활발히 연구되고 있다. 농업 분야에서는 병해충 저항성, 가뭄 내성, 영양 성분 강화 등의 특성을 가진 새로운 품종 개발에 활용되어, 기후 변화와 인구 증가에 대응하는 지속 가능한 농업 기술로 주목받고 있다.

동아일보 | IT/의학

‘유전자 가위’로 선천성 실명질환 치료 성공했다

김민수 동아사이언스 기자

입력 2020-10-23 03:00 | 업데이트 2020-10-23 03:00

가

기존 기술보다 정확도 뛰어난 ‘염기 교정 가위’로 시력 되돌려

희귀 유전질환 치료 길 열어

올해 노벨 화학상은 생명공학에서 가장 혁신적인 기술로 각광받고 있는 크리스퍼(CRISPR) 유전자가위를 개발한 2명의 여성 화학자에게 돌아갔다. 크리스퍼 유전자가위는 생명체의 정보를 담고 있는 DNA에서 유전질환을 유발하는 특정한 부위의 염기를 잘라내 교정하는 기술이다. 최근에는 유전자가위 기술의 정확도와 효율성이 높아져 선천성 유전질환 치료 가능성을 높이고 있다.

한국 연구자들이 주축으로 구성된 크시슈토프 팔체프스키 미국 어바인 캘리포니아대 가빈허버트안과연구소 교수 연구팀은 DNA의 단일 염기를 교정하는 ‘염기 교정 유전자가위’ 기술을 이용해 선천성 유전질환인 ‘레버 선천성 흑암시’를 유발하는 유전자 교정의 효율성을 대폭 높이는 데 성공했다고 국제학술지 ‘네이처 바이오메디컬 엔지니어링’ 19일자(현지 시간) 온라인판에 발표했다.

Figure 290: CRISPR 희귀질환 치료

Figure 14.14 CRISPR 희귀질환 치료 – CRISPR 기술을 이용한 선천성 실명질환 치료 사례

Figure 14.15 CRISPR 항암 치료 – CRISPR 기술을 활용한 암 치료법 개발

15.3.1 유전체 편집 기술의 발전사

첫 번째 세대인 아연 핑거 Nuclease(ZFN, Zinc Finger Nuclease)는 1990년대에 개발되었다. ZFN은 특정 DNA 서열을 인식하는 아연 핑거 단백질과 DNA를 절단하는 FokI Nuclease를 결합한 형태이다. 그러나 원하는 DNA 서열을 표적으로 하는 아연 핑거 단백질을 설계하고 제작하는 과정이 복잡하고 비용이 많이 들어 널리 사용되지 못했다.

두 번째 세대인 TALEN(Transcription Activator-Like Effector Nuclease)은 2010년경에 등장했다. TALEN은 식물 병원균에서 발견된 TAL 이펙터 단백질을 이용하여 DNA를 인식하며, ZFN보다 설계가 쉽고 정확도가 높았다. 하지만 여전히 각 표적마다 새로운 단백질을 만들어야 한다는 한계가 있었다.

세 번째 세대인 CRISPR/Cas 시스템은 2012-2013년에 혁신적인 변화를 가져왔다. 이 시스템의 특징은 DNA 결합 단백질 대신 작은 가이드 RNA가 표적 특이성을 결정한다는 점이다.

Figure 14.16 CRISPR 유전체 편집 – CRISPR/Cas9 시스템의 DNA 절단 메커니즘

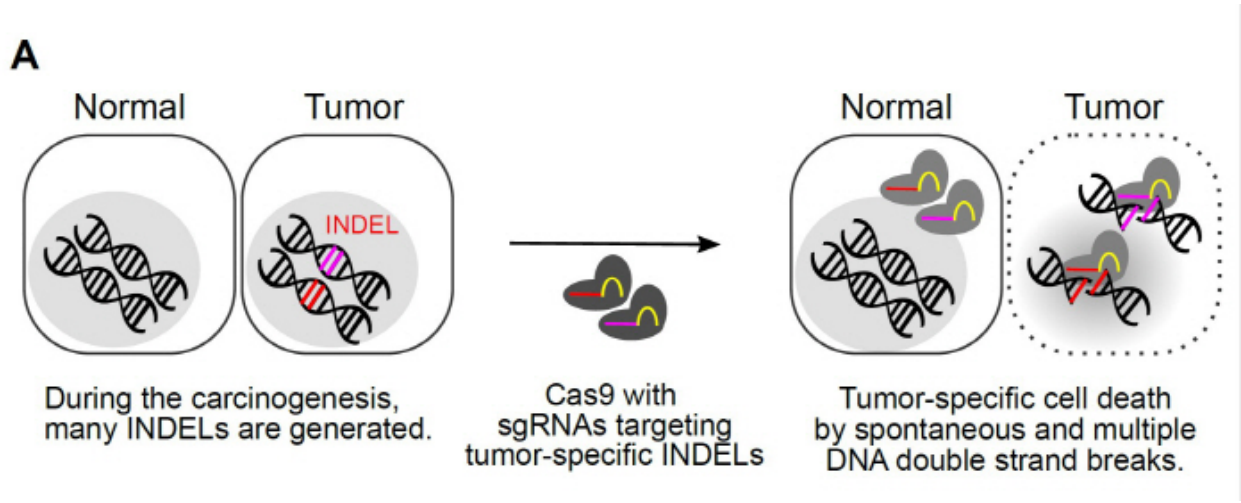


Figure 291: CRISPR 항암 치료

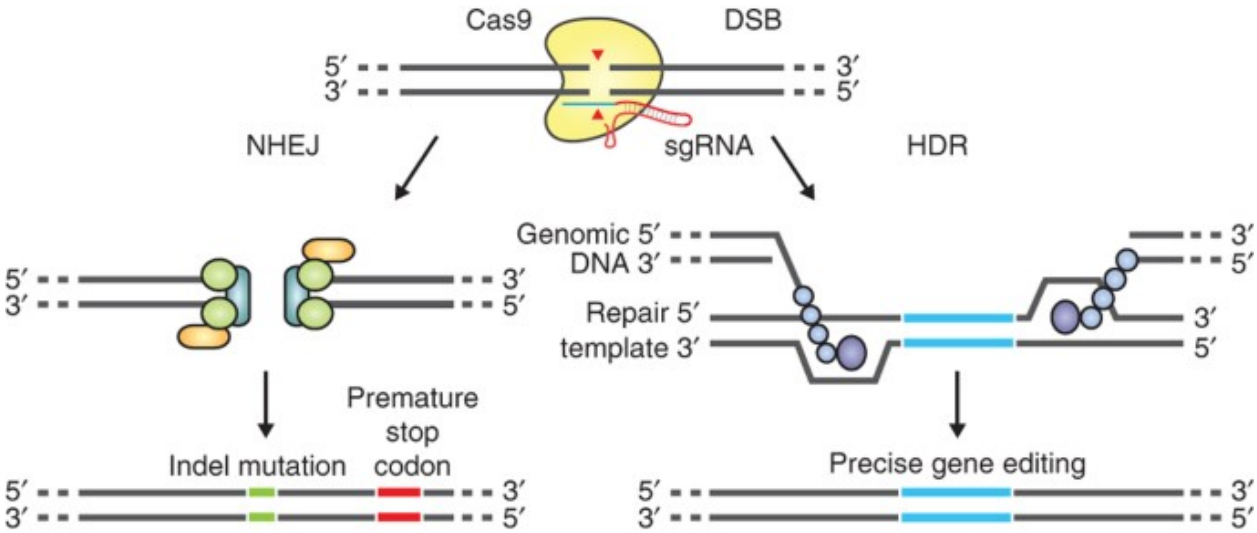


Figure 292: CRISPR 유전체 편집

15.3.2 CRISPR/Cas 시스템의 구조와 원리

CRISPR/Cas 시스템은 CRISPR 연관 단백질 9(Cas9), CRISPR RNA(crRNA), 그리고 트랜스 활성화 crRNA(tracrRNA)로 구성된다. 또는 crRNA와 tracrRNA를 하나로 합친 단일 가이드 RNA(sgRNA)와 Cas9 단백질로 구성될 수도 있다. 가이드 RNA의 가이드 서열은 프로토스페이서라고 불리는 20염기쌍의 표적 DNA 서열과 상보적이며, 이 표적 서열은 프로토스페이서 인접 모티프(PAM, Protospacer Adjacent Motif) 바로 옆에 위치해야 한다.

CRISPR 시스템에서 사용되는 다양한 Cas 단백질들은 각각 고유한 PAM 서열을 인식한다:

- SpCas9 (Streptococcus pyogenes): 5'-NGG-3'
- SaCas9 (Staphylococcus aureus): 5'-NNGRRT-3'
- NmCas9 (Neisseria meningitidis): 5'-NNNNGMTT-3'
- Cas12a (AsCpf1/LbCpf1): 5'-TTTV-3'
- Cas12f1: 5'-TTTR-3'
- CasX: 5'-TTCN-3'

Cas9 단백질은 HNH와 RuvC라는 두 개의 Nuclease 도메인을 가지고 있어 DNA 이중나선을 모두 절단한다. 그러나 이를 변형하여 특별한 용도로 사용할 수도 있다. Cas9 Nickase는 두 Nuclease 도메인 중 하나를 불활성화시켜 DNA의 한 가닥만 절단하는 변형 단백질이다. 이는 오프타겟 효과를 줄이고 더 정밀한 편집을 가능하게 한다. 또한 dead Cas9(dCas9)은 두 Nuclease 도메인을 모두 불활성화시켜 DNA를 절단하지 않고 특정 부위에만 결합하는 형태로, 전사 조절이나 후성유전학적 변형에 활용된다.

이러한 다양한 Cas 단백질들은 자연에서 발견되거나 단백질 공학을 통해 개발되었으며, 서로 다른 PAM 서열을 인식하여 유전체 편집의 표적 범위를 대폭 확장시켰다. 이를 통해 연구자들이 원하는 유전체 부위를 보다 유연하게 편집할 수 있게 되었다.

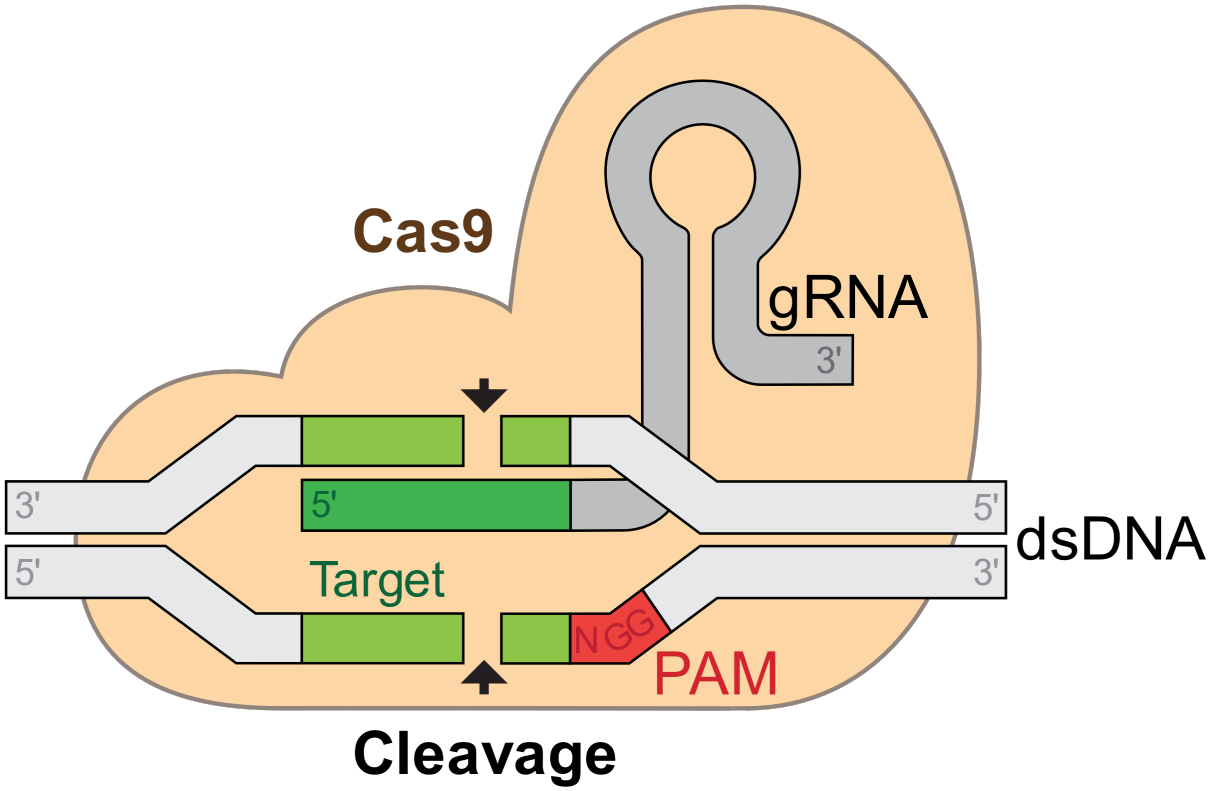


Figure 293: CRISPR PAM과 sgRNA

Figure 14.17 CRISPR PAM과 sgRNA – PAM 서열 인식과 단일 가이드 RNA의 표적 특이성 결정 및 절단 메커니즘

15.3.3 염기 편집 기술

CRISPR/Cas9 기술의 발전과 함께 보다 정밀하고 안전한 유전체 편집을 위한 차세대 기술들이 개발되고 있다. 그 중에서도 염기 편집 기술은 DNA 이중나선을 절단하지 않고도 단일 염기를 정밀하게 치환할 수 있는 혁신적인 도구이다.

염기 편집기(Base Editor)는 Cas9 Nickase와 염기 탈아미노화 효소를 결합한 구조로, 표적 부위에서 한 가닥의 DNA만을 절단하여 편집 창을 형성한다. 사이토신 염기 편집기(CBE)는 사이토신 탈아미노화 효소를 이용해 C를 U로 변환한 후 세포의 DNA 복구 과정을 통해 최종적으로 C를 T로 바꾼다. 아데닌 염기 편집기(ABE)는 아데닌 탈아미노화 효소를 사용하여 A를 이노신으로 변환하고, 이것이 복제 과정에서 G로 인식되어 결과적으로 A를 G로 치환한다. 이 기술은 이중나선 절단으로 인한 원치 않는 삽입이나 결실(indel) 생성을 피할 수 있어 더욱 안전한 편집이 가능하다.

15.3.4 프라임 편집 기술

프라임 편집기(Prime Editor)는 더욱 진보된 기술로, 최대 80염기쌍까지의 정밀한 삽입, 결실, 치환을 동시에 수행할 수 있다. 이 시스템은 Cas9 Nickase와 역전사효소를 결합한 형태로, 프라임 편집 가이드 RNA(pegRNA)가 원하는 편집 정보를 담고 있어 매우 정확한 편집이 가능하다. 프라임 편집기는 기존 CRISPR로는 어려웠던 복잡한 편집을 가능하게 하여 유전자 치료의 새로운 지평을 열고 있다.

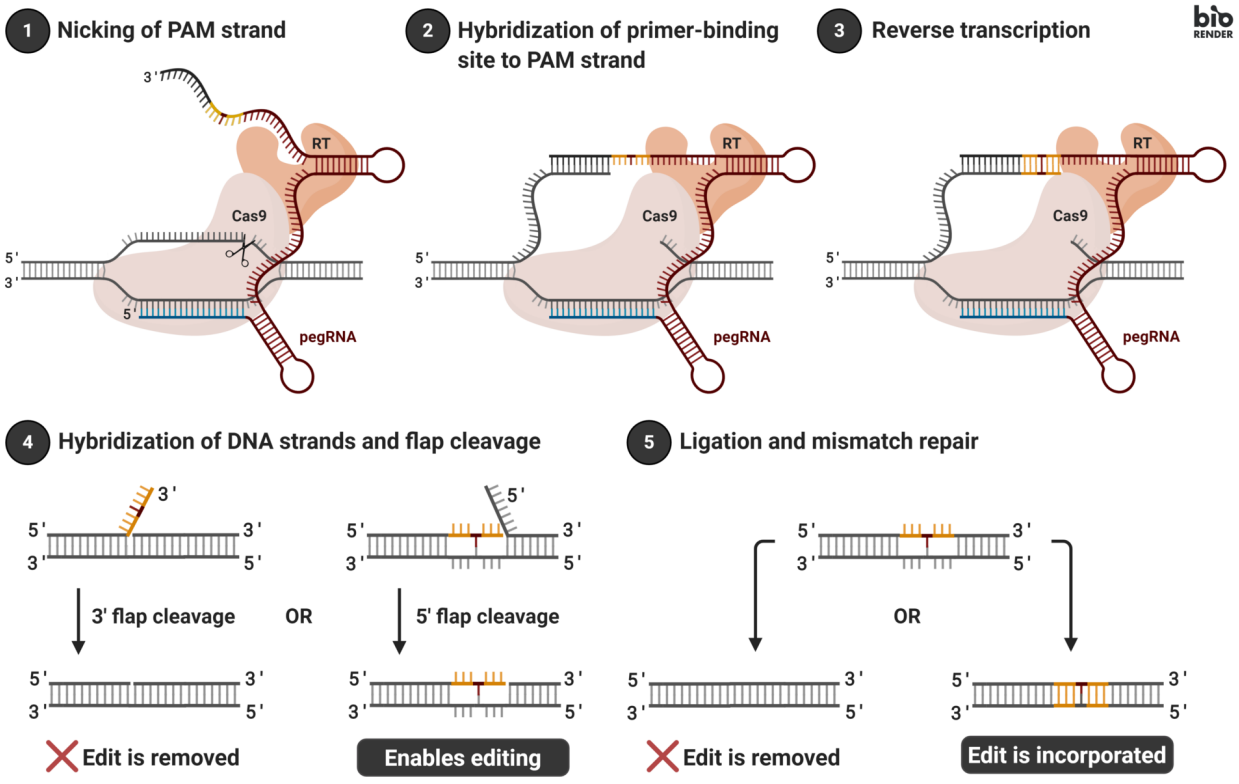


Figure 294: 프라임 편집기

Figure 14.18 프라임 편집기 – 정밀한 삽입, 결실, 치환을 가능하게 하는 차세대 유전체 편집 기술

15.3.5 CRISPR 기반 진단 기술

CRISPR 기술은 편집뿐만 아니라 진단 분야에서도 혁신을 가져왔다. SHERLOCK(Specific High-sensitivity Enzymatic Reporter unLOCKing)은 Cas13 단백질을 이용한 핵산 검출 시스템으로, 바이러스나 세균의 유전 물질을 매우

민감하게 탐지할 수 있다. 이 기술은 COVID-19와 같은 감염병 진단에서 PCR과 유사한 민감도를 보이면서도 더 빠르고 간편한 검사를 가능하게 한다. 또한 DETECTR(DNA Endonuclease-Targeted CRISPR Trans Reporter)과 같은 유사한 시스템들도 개발되어 현장 진단(point-of-care) 도구로 활용되고 있다.

15.3.6 오프타겟 효과와 예측 도구

CRISPR/Cas 시스템의 가장 큰 도전 과제 중 하나는 오프타겟(off-target) 효과이다. 이는 가이드 RNA가 의도한 표적 부위가 아닌 다른 DNA 서열을 인식하여 원치 않는 편집이 일어나는 현상을 말한다. 오프타겟 효과는 임상 응용에서 심각한 부작용을 초래할 수 있기 때문에, 이를 예측하고 최소화하는 것이 매우 중요하다.

오프타겟이 발생하는 주요 원인은 가이드 RNA와 표적 DNA 간의 불완전한 상보적 결합이다. 가이드 RNA의 20염기쌍 서열 중 일부가 다른 유전체 부위와 유사할 경우, Cas 단백질이 그 부위를 잘못 인식하여 절단할 수 있다. 특히 PAM 서열 근처의 시드(seed) 영역에서의 미스매치는 오프타겟 가능성을 크게 높인다.

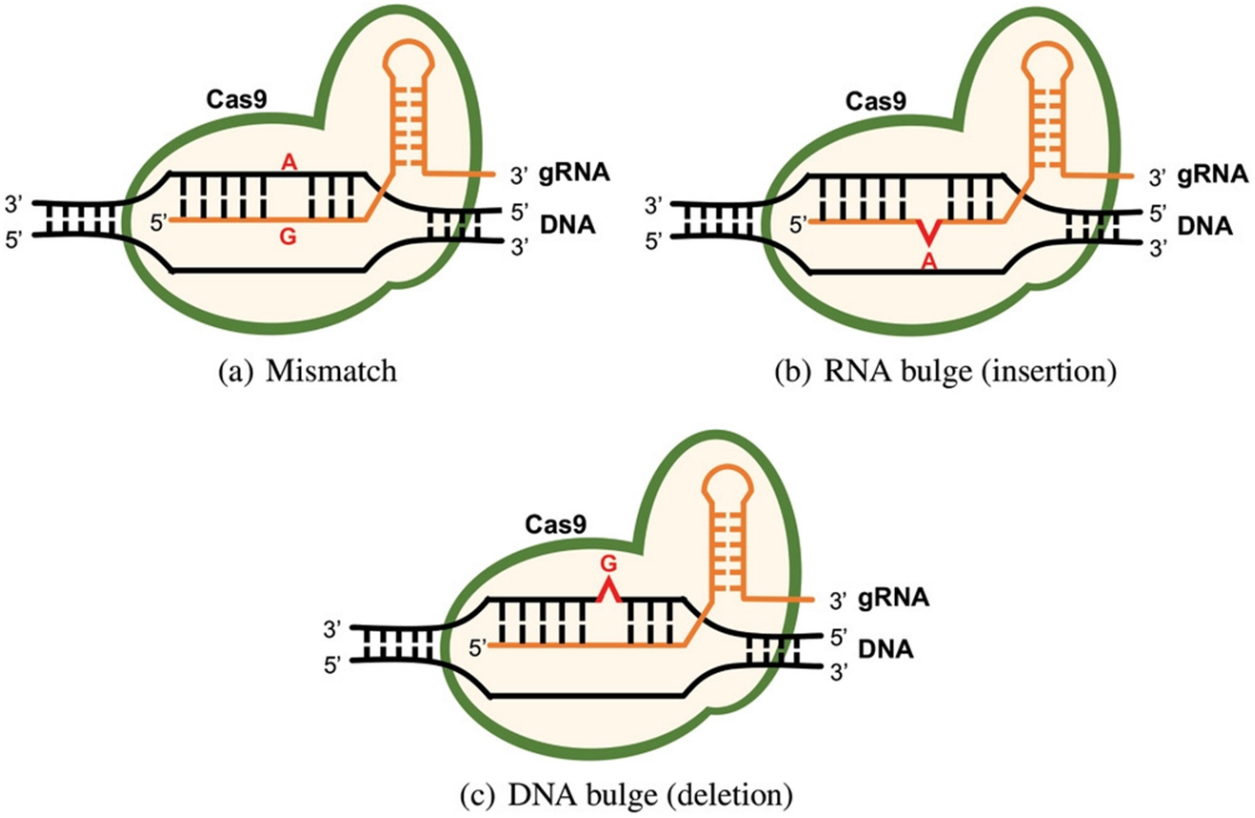


Figure 295: CRISPR 오프타겟

Figure 14.19 CRISPR 오프타겟 – CRISPR 시스템에서 발생할 수 있는 오프타겟 효과

이러한 문제를 해결하기 위해 다양한 계산 도구들이 개발되었다. 현재 가장 널리 활용되는 도구 중 하나는 Cas-OFFinder로, 이는 GPU 가속을 이용하여 전체 유전체에서 잠재적 오프타겟 부위를 빠르게 검색하는 도구이다. 이 프로그램은 사용자가 지정한 미스매치 허용 범위 내에서 모든 가능한 오프타겟 부위를 찾아 점수를 매겨 위험도를 평가한다.

Figure 14.20 Cas-OFFinder 알고리즘 – GPU 가속을 이용한 빠른 오프타겟 검색 알고리즘의 동작 원리

Cas-OFFinder를 포함한 CRISPR RGEN Tools (http://rgenome.net)는 국내에서 개발된 통합 설계 플랫폼으로, 가이드 RNA 설계부터 오프타겟 예측까지 전 과정을 지원한다. 이 도구는 다양한 Cas 단백질(Cas9, Cas12a 등)에 대응하며, 실험적으로 검증된 오프타겟 데이터를 바탕으로 예측 정확도를 높이고 정확한 유전체 편집을 가능하게 한다.

Figure 14.21 CRISPR RGEN Tools – 가이드 RNA 설계부터 오프타겟 예측까지 지원하는 통합 CRISPR 설계 플랫폼

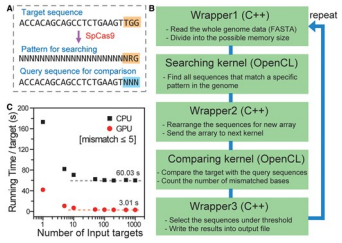


Figure 296: Cas-OFFinder 알고리즘

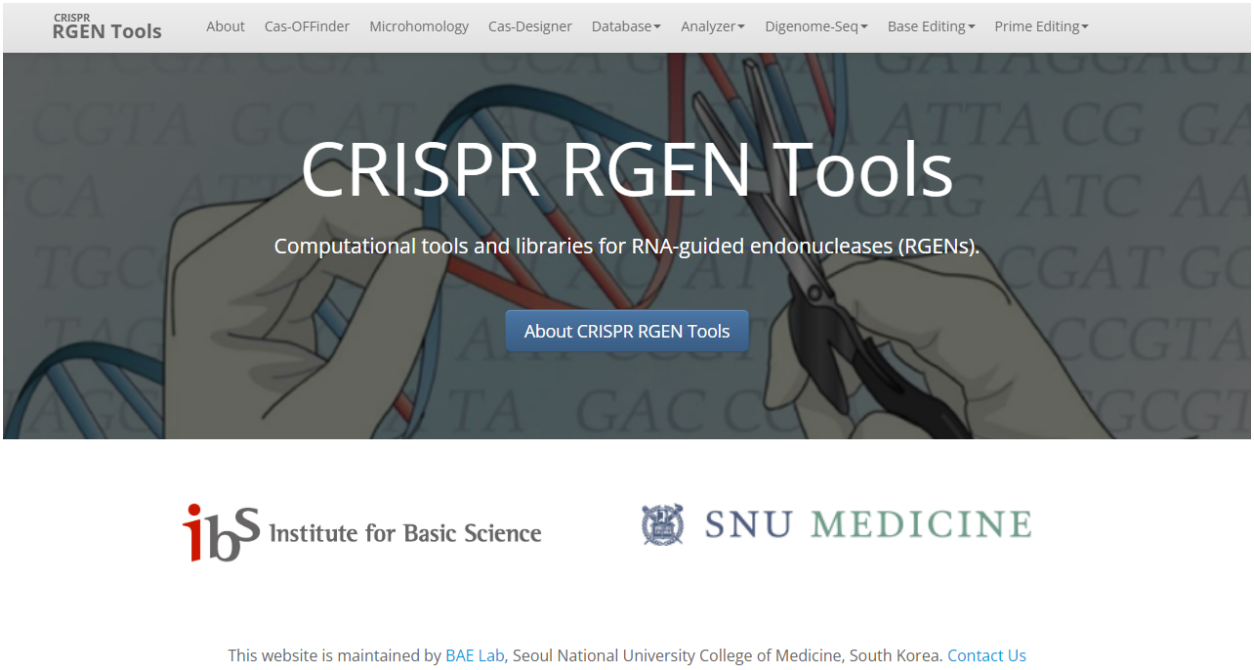


Figure 297: CRISPR RGEN Tools

최근에는 머신러닝과 딥러닝 기술을 활용한 오프타겟 예측 모델들도 등장하고 있다. 이들은 대량의 실험 데이터를 학습하여 기존의 규칙 기반 방법보다 정확한 예측을 가능하게 한다.

15.3.7 세포 내 소기관 유전체 편집

유전체 편집 기술이 발전하면서, 세포 내 소기관인 미토콘드리아와 엽록체의 유전체 편집에도 관심이 높아지고 있다. 그러나 미토콘드리아와 엽록체는 세포 내에서 이중막으로 둘러싸여 있어 편집 도구의 전달이 어려운 편이며, 따라서 이들에 대한 유전체 편집은 기술적 도전에 직면해 왔다.

최근에 개발된 미토콘드리아 유전체 편집은 미토콘드리아 질환 치료에 새로운 가능성을 제시한다. 근래 등장한 미토콘드리아 특이적 염기 편집 기술은 미토콘드리아 DNA에 직접 작용하여 점 돌연변이를 교정할 수 있다. 이는 미토콘드리아 유전 질환의 근본적 치료법 개발에 중요한 진전을 의미한다.

또한 식물의 엽록체 유전체 편집은 농업 분야에서 새로운 돌파구를 마련하고 있다. 엽록체는 광합성의 핵심 장소로, 이곳의 유전자를 편집하면 광합성 효율 향상, 제초제 저항성 부여, 그리고 고부가가치 물질 생산이 가능하다. 특히 엽록체 유전체는 모계 유전되는 특성이 있어 유전자 확산에 대한 우려를 줄일 수 있다는 장점이 있다.

15.3.8 CRISPR를 이용한 이종장기이식

이종장기이식(xenotransplantation)은 다른 종의 동물에서 인간에게 장기를 이식하는 기술로, 만성적인 장기 부족 문제를 해결할 수 있는 혁신적인 대안으로 주목받고 있다. 특히 돼지는 인간과 장기 크기와 생리학적 기능이 유사하여 이종장기이식의 가장 유망한 공여 동물로 여겨지고 있다.

그러나 이종장기이식은 몇 가지 중대한 도전 과제를 가지고 있다. 가장 큰 문제는 면역 거부 반응이다. 인간의 면역 시스템은 다른 종의 장기를 외부 침입자로 인식하여 강력한 면역 공격을 가하게 된다. 또한 이종 감염의 위험도 존재한다. 돼지의 내인성 레트로바이러스(PERV) 같은 동물 특이적 병원체가 인간에게 전파될 가능성이 있어 심각한 안전성 문제를 야기할 수 있다.

CRISPR 기술의 등장은 이러한 문제들을 해결할 수 있는 도구를 제공하였다. 첫째, 면역 거부 반응을 줄이기 위해 돼지의 면역 관련 유전자들을 인간형으로 교체할 수 있다. 대표적으로 주조직적합성복합체(MHC) 유전자들을 인간의 HLA 유전자로 교체하여 면역 인식성을 향상시킬 수 있다. 둘째, 이종 감염 위험을 제거하기 위해 돼지 유전체에 내재된 모든 PERV 서열을 비활성화할 수 있다.

Figure 14.22 CRISPR와 이종장기이식 – CRISPR 기술을 이용한 이종장기이식용 돼지 개발 관련 뉴스 및 규제로 인한 국내의 현황

2022년 미국 메리랜드 대학교 의료진은 CRISPR로 편집된 돼지의 신장을 뇌사한 환자에게 이식하는 세계 최초의 이종장기이식 수술을 성공적으로 수행하였다. 이 수술에 사용된 돼지는 10개의 유전자가 CRISPR로 편집되었다: 3개의 유전자가 끄기고(노크아웃), 6개의 인간 유전자가 삽입되었으며, 1개의 돼지 유전자가 수정되었다. 비록 환자는 2개월 후 사망하였지만, 이는 이식된 신장의 문제가 아니라 기존의 심장 질환 때문이었다. 이식된 돼지 신장은 수술 후 정상적으로 기능하였으며, 이는 CRISPR 기반 이종장기이식의 가능성을 입증하는 중요한 이정표가 되었다.

현재 여러 연구기관에서는 더욱 안전하고 효과적인 이종장기이식용 돼지 개발을 위한 연구가 진행되고 있다. 여기에는 더 많은 면역 관련 유전자의 편집, PERV 완전 제거, 그리고 인간 대사에 최적화된 장기 기능 개선 등이 포함된다. 이러한 연구들이 성공한다면, CRISPR 기반 이종장기이식은 전 세계적인 장기 부족 문제를 해결하는 혁신적인 치료법이 될 것으로 기대된다.

15.4 결론

첨단 유전체학 분야는 군유전체학을 통한 미생물 생태계의 이해, 식물 유전체학에서의 배수성 활용, CRISPR/Cas 시스템을 이용한 정밀 유전체 편집, 그리고 유전체 역학을 통한 질병의 분자적 이해 등을 통해 의학과 농업 분야에 혁신적 변화를 가져오고 있다.

분자 역학의 발전은 감염병의 전파 과정을 분자 수준에서 추적하고 예측할 수 있게 하였으며, 다층 오믹스 데이터의 통합은 개인 맞춤형 정밀 의학의 실현을 가능하게 하고 있다. RNA-seq 기술의 발전은 전사체 연구를 혁신적으로 변화시켰으며, GWAS와 로지스틱 회귀를 통한 유전적 위험 요인 분석은 질병 예방과 치료 전략 수립에 핵심적인 역할을 하고 있다.

이러한 첨단 기술들은 각각 독립적으로 발전하면서도 서로 융합하여 더욱 강력한 도구로 발전하고 있으며, 앞으로도 인류의 건강과 복지 향상, 그리고 지속가능한 미래 사회 구축에 중요한 역할을 할 것으로 기대된다.

사람에 돼지 장기 이식 잇따라…국내 핵심기업은 존속 위기

2024.03.31 08:00

가 가

미국에서 돼지 신장을 살아 있는 사람한테 이식한 세계 최초의 사례가 나왔다. 의료진이 이식 수술에 쓰인 돼지 신장을 꺼내고 있다. 매사추세츠 종합병원 제공

미국과 중국에서 최근 잇따라 유전자를 교정한 돼지 장기를 실제 사람에게 이식하는 데 성공했다는 소식이 전해졌다. 아직 사람 대상의 임상은 초기 수준이지만 간이나 심장, 신장 등 장기 이식이 반드시 필요한 환자들을 위한 이식용 장기 부족 문제 해결에 대한 기대감도 덩달아 커지고 있다. 특히 이종 이식 기술과 노하우를 국내에서도 확보해야 하는 것 아니냐는 목소리도 나온다.

30일 관련업계에 따르면 이처럼 세계적으로 이종 이식을 향한 기대감이 커지고 있지만 국내는 정부 예산을 받아 국가 사업을 이끌던 이종 이식 핵심 기업 중 하나가 상장 폐지를 앞두고 있는 것으로 나타났다.

Figure 298: CRISPR와 이종장기이식

16 유전체 데이터 보안

16.1 유전체 데이터 보안의 필요성과 윤리적 고려사항

개인정보 중에서도 유전체 정보는 특별한 보호가 필요한 영역이다. 과연 유전체 데이터는 일반적인 개인정보와 무엇이 다를까? 왜 우리는 유전체 데이터에 대해 더 신중한 접근이 필요할까? 본 장에서는 이러한 질문들에 답하기 위해 유전체 정보가 갖는 고유한 특성들을 살펴보고, 이것이 개인과 사회에 미치는 영향에 대해 알아본다.

16.1.0.1 15.1.1 SNP 기반 개인 식별 유전체 데이터의 가장 중요한 특성 중 하나는 개인을 고유하게 식별할 수 있다는 점이다. DNA에 존재하는 단일염기다형성(SNP, Single Nucleotide Polymorphism)은 각 개인마다 다른 패턴을 나타내며, 이는 익명화가 매우 어려운 이유가 된다.

SNP는 개인 간에 DNA 서열에서 단일 염기(A, T, G, C)가 다른 위치를 말한다. 예를 들어, 어떤 사람은 특정 위치에서 'A'를 가지고 있지만, 다른 사람은 같은 위치에 'G'를 가질 수 있다. 인간 유전체에는 약 300만 개의 SNP가 존재하며, 이들의 조합은 각 개인마다 고유한 유전적 지문을 형성한다.

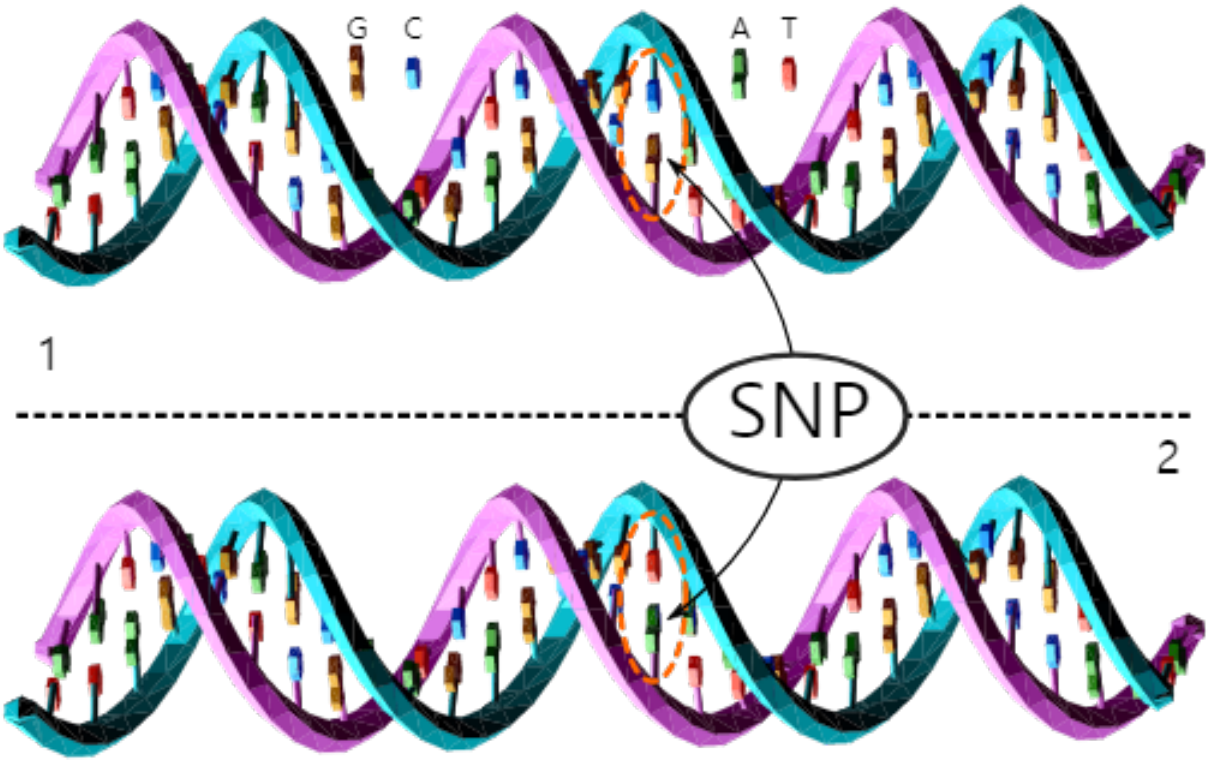


Figure 299: 단일염기다형성(SNP)의 개념도

Figure 15.1 단일염기다형성(SNP)의 개념. 개인 간 유전체에서 단일 염기가 다른 위치를 보여주며, 이러한 변이들의 조합이 각 개인만의 고유한 유전적 지문을 형성한다.

수만 개의 SNP 데이터를 조합하면 개인을 특정할 수 있는 고유한 패턴이 나타난다. 이는 다른 개인정보와 달리 유전체 데이터가 본질적으로 익명화에 한계가 있음을 의미한다. 심지어 여러 사람의 DNA가 혼재된 데이터베이스에서도 개별 기여자를 구별해낼 수 있으며, 소량의 데이터만으로도 개인 재식별이 가능하다는 보안상 우려가 존재한다.

Figure 15.2 SNP 데이터를 활용한 개인 식별 과정. 수만 개의 SNP 조합을 통해 개인을 고유하게 구별할 수 있는 패턴이 형성되며, 이는 유전체 데이터의 익명화가 어려운 이유를 보여준다.

이러한 재식별 위험성은 2008년 발표된 연구에서 실증적으로 확인되었다: – "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays"

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, David W. Craig

Published: August 29, 2008 • https://doi.org/10.1371/journal.pgen.1000167

Abstract

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

Figure 300: SNP 데이터를 활용한 개인 식별 과정

(2008)

이 연구는 고밀도 SNP 분석을 통해 복잡한 DNA 혼합물에서도 개별 기여자를 구별해낼 수 있음을 보여주었다.

16.1.0.2 15.1.2 STR 분석과 가족 관계 추론 단일염기반복서열(STR, Short Tandem Repeat)은 또 다른 형태의 개인 식별 가능한 유전체 정보이다. STR은 DNA 내에서 짧은 서열이 연속적으로 반복되는 영역으로, 개인마다 반복 횟수가 다르다. 이러한 반복 횟수의 개인차는 체세포 분열 과정에서도 안정적으로 유지되며, 집단 내에서 높은 다형성을 보인다. 이런 특징을 기반으로 한 STR 분석은 법의학 분야에서도 개인 식별 목적으로 널리 활용되고 있다.

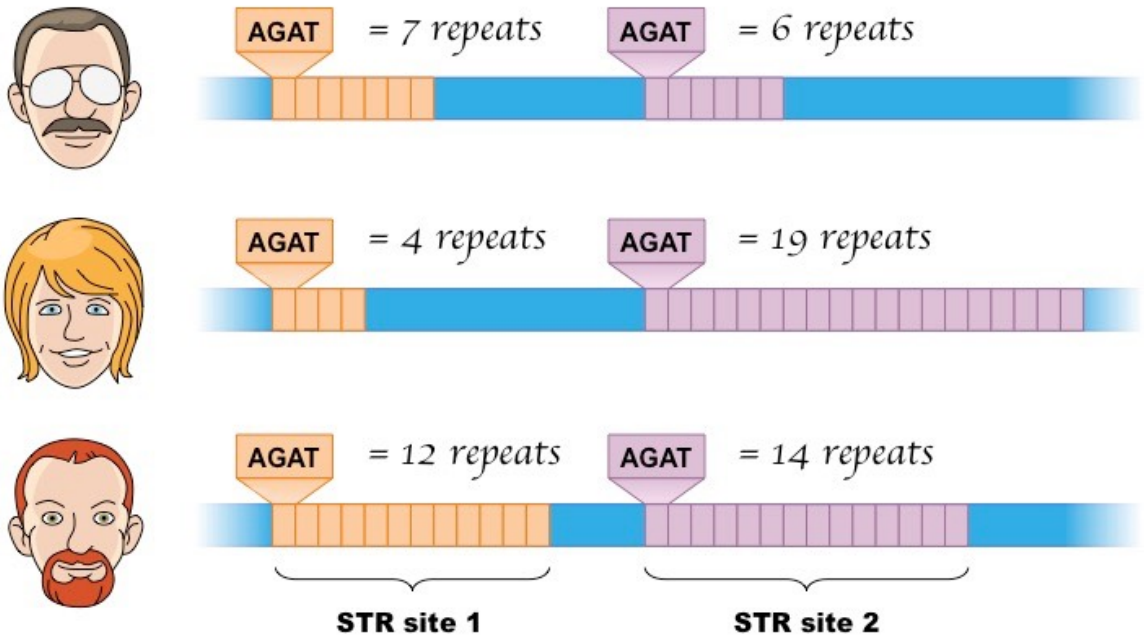


Figure 301: 단일염기반복서열(STR)의 구조

Figure 15.3 단일염기반복서열(STR)의 구조. DNA 내에서 짧은 서열이 연속적으로 반복되는 영역으로, 개인마다 반복 횟수가 달라 개인 식별에 활용된다.

Figure 15.4 법의학 분야에서의 STR 분석 활용. STR 마커들의 조합을 통해 개인 식별과 친자 확인이 가능하며, 높은 정확도로 개인을 구별할 수 있다.

최근 연구에서는 Y염색체 STR 분석을 통한 실제 개인 식별 사례가 보고되었다. Gymrek 등은 GenBank에 공개된 개인의 Y염색체 서열에서 STR을 추출하고, 이를 공개 계보 웹사이트의 정보와 비교하여 가장 가능성 높은 성씨를 찾아낸 후, 나이와 거주 주 정보를 결합하여 특정 개인을 식별해냈다. 911명의 개인을 대상으로 한 분석에서는 미국 백인 남성 인구 내에서 12%의 성공률을 보였으며, 공개 저장소에 유전체 정보가 있는 개인들의 가족까지 높은 확률로 식별 가능함을 보여주었다 (Gymrek et al., Science, 2013).

Figure 15.5 Y염색체 STR을 이용한 성씨 추론 과정. 공개된 계보 데이터베이스와 Y-STR 정보를 연계하여 개인의 성씨를 추론하고, 추가 정보와 결합하여 특정 개인을 식별하는 과정을 보여준다.

즉, STR 역시 개인 식별에 활용될 수 있다는 점에서 유전체 데이터 보안상 중요한 고려사항이 된다.

16.1.1 개인 유전체 공개로 인한 위험성

특히 유전체 데이터의 경우 질병 위험 예측이 가능하다는 특성으로 인해 위험이 존재한다. 가령, TP53 변이는 높은 암 발생 위험을, BRCA1/2 변이는 유방암 및 난소암에 대한 높은 감수성을 나타낸다. CHRNA5/A3/B4 유전자군 변이는 니코틴 중독 및 흡연 관련 질환 위험과 연관되며, 다유전자 위험 점수(PRS)를 통해서는 여러 유전적 변이를 종합한 질병

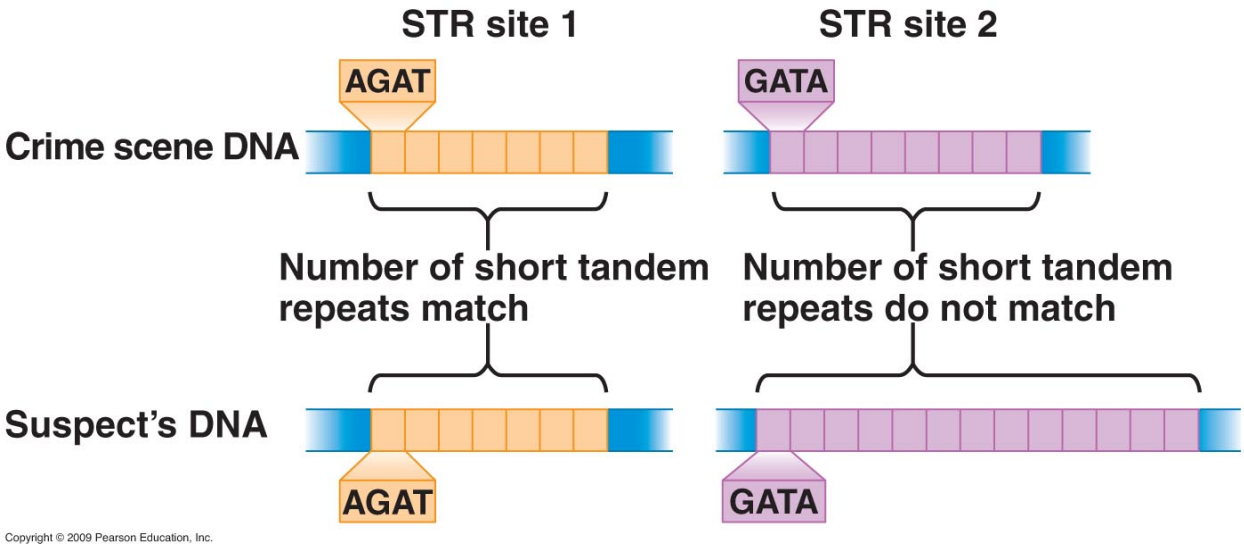


Figure 302: 법의학 분야에서의 STR 분석

위험 예측이 가능하다. 이러한 예측 정보는 개인의 미래 건강 상태에 대한 정보를 제공하지만, 부적절하게 사용될 경우 개인에게 불이익을 초래할 수 있다.

이러한 개인 유전체 데이터가 부적절하게 노출될 경우 개인이 직면할 수 있는 위험들이 존재한다:

- 보험 가입 거절: 유전적 질병 위험이 높은 개인에 대한 보험사의 차별적 대우
- 고용 차별: 특정 질병에 대한 감수성이 높은 지원자에 대한 고용 배제
- 사회적 낙인: 정신질환이나 유전질환과 관련된 사회적 편견
- 범죄 악용: 개인정보를 이용한 금융 사기나 신원 도용, 협박 등

16.2 유전체 데이터 익명화 기술

16.2.1 메타데이터 개요

의생명정보학 데이터 분석에서 유전체 데이터는 종종 환자의 다양한 메타데이터와 함께 취급된다. 1장에서 학습한 의료정보학적 데이터의 범주를 기반으로, 유전체 연구에서 활용되는 메타데이터는 다음과 같이 분류할 수 있다.

전자의무기록(EMR) 기반 메타데이터: 전자의무기록에서 추출되는 정보는 유전체 연구의 핵심 메타데이터를 구성한다: – 진찰 기록: 임상 증상, 신체 검사 소견, 의사의 평가 및 진단 – 환자 이력: 가족력, 과거 병력, 처방력, 알레르기 정보, 생활습관(술/담배 이력 등) – 정량적 데이터: 키, 몸무게, 혈압, 맥박 등의 생체 지표 – 검사 결과: 혈액 검사, 소변 검사, 생화학 검사 수치

Figure 15.6 전자의무기록(EMR) 시스템. 환자의 진찰 기록, 검사 결과, 처방 정보 등이 디지털 형태로 저장되며, 유전체 연구에서 중요한 메타데이터를 제공한다.

표준화된 진단 코드: 국제질병분류(ICD) 시스템을 통해 표준화된 진단 정보가 메타데이터로 활용된다: – 세계보건기구(WHO)에서 관리하는 ICD-11을 기반으로 한 진단 코드 – 한국에서는 한국표준질병사인분류(KCD)를 사용 – 이러한 표준화된 코드는 국제적인 유전체 연구 협력과 데이터 통합을 가능하게 함

Figure 15.7 국제질병분류(ICD) 시스템. WHO에서 관리하는 표준화된 진단 코드 체계로, 유전체 연구에서 표준화된 진단 정보를 제공하여 국제적인 연구 협력을 가능하게 한다.

DICOM 기반 의료영상 메타데이터: 의료영상 및 통신 표준(DICOM, Digital Imaging and Communications in Medicine)은 유전체 연구에서 중요한 영상 메타데이터를 제공한다:

DICOM 표준의 특징: – 거의 모든 현대 의료영상(X-ray, CT, MRI, PET, 초음파 등)이 DICOM 포맷으로 저장됨 – 병원 내 PACS(Picture Archive and Communication System)를 통해 관리 – 영상 데이터뿐만 아니라 촬영 조건, 환자 정보,

REPORT

Identifying Personal Genomes by Surname Inference

MELISSA GYMREK, AMY L. MCGUIRE, DAVID GOLAN, ERAN HALPERIN, AND , YANIV ERLICH Authors Info & Affiliations

SCIENCE • 18 Jan 2013 • Vol 339, Issue 6117 • pp. 321-324 • DOI: 10.1126/science.1229566

691 642 GET ACCESS

Anonymity Compromised

The balance between maintaining individual privacy and sharing genomic information for research purposes has been a topic of considerable controversy. **Gymrek *et al.*** (p. 321; see the Policy Forum by **Rodriguez *et al.***) demonstrate that the anonymity of participants (and their families) can be compromised by analyzing Y-chromosome sequences from public genetic genealogy Web sites that contain (sometimes distant) relatives with the same surname. Short tandem repeats (STRs) on the Y chromosome of a target individual (whose sequence was freely available and identified in GenBank) were compared with information in public genealogy Web sites to determine the shortest time to the most recent common ancestor and find the most likely surname, which, when combined with age and state of residency identified the individual. When STRs from 911 individuals were used as the starting points, the analysis projected a success rate of 12% within the U.S. male population with Caucasian ancestry. Further analysis of detailed pedigrees from one collection revealed that families of individuals whose genomes are in public repositories could be identified with high probability.

Figure 303: Y염색체 STR을 이용한 성씨 추론

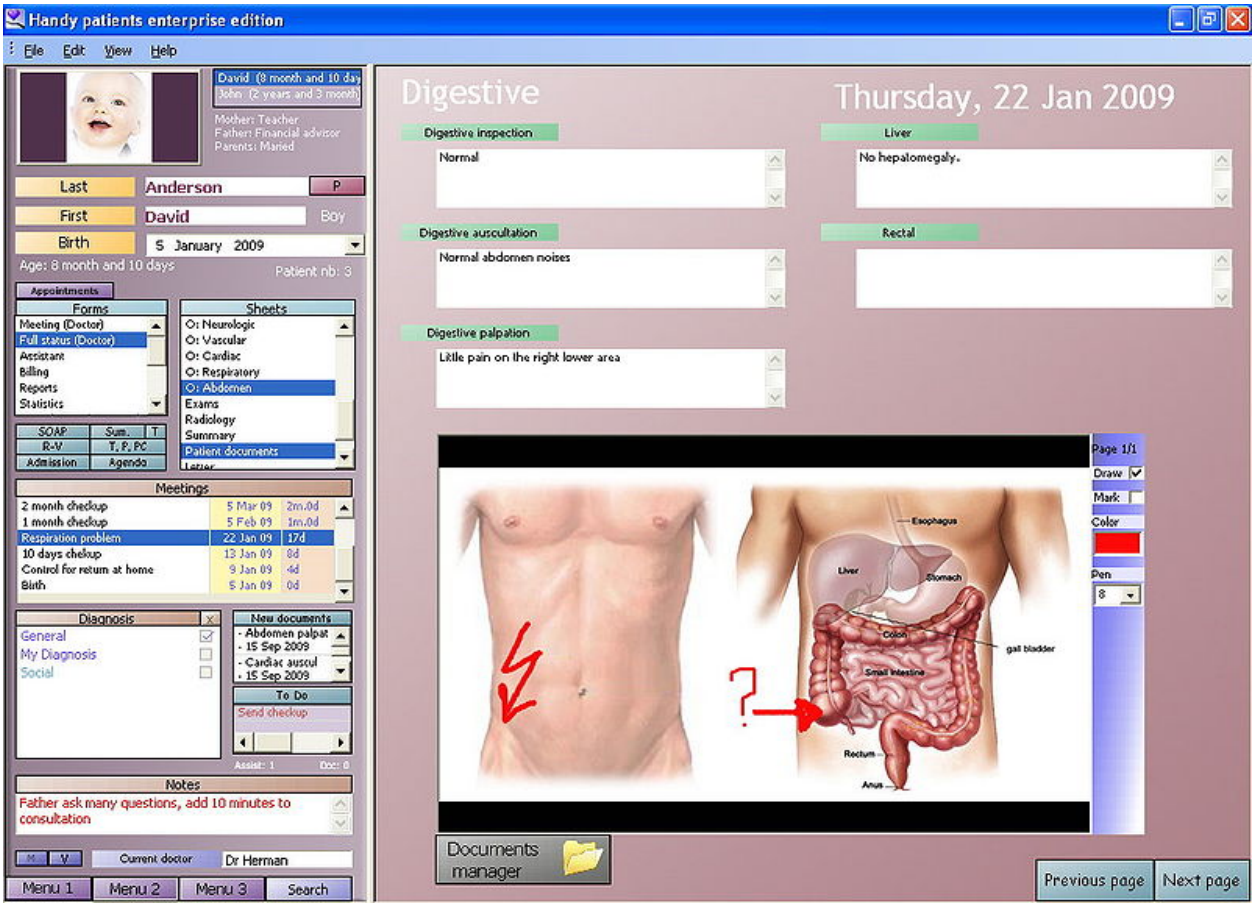


Figure 304: 전자의무기록(EMR) 시스템

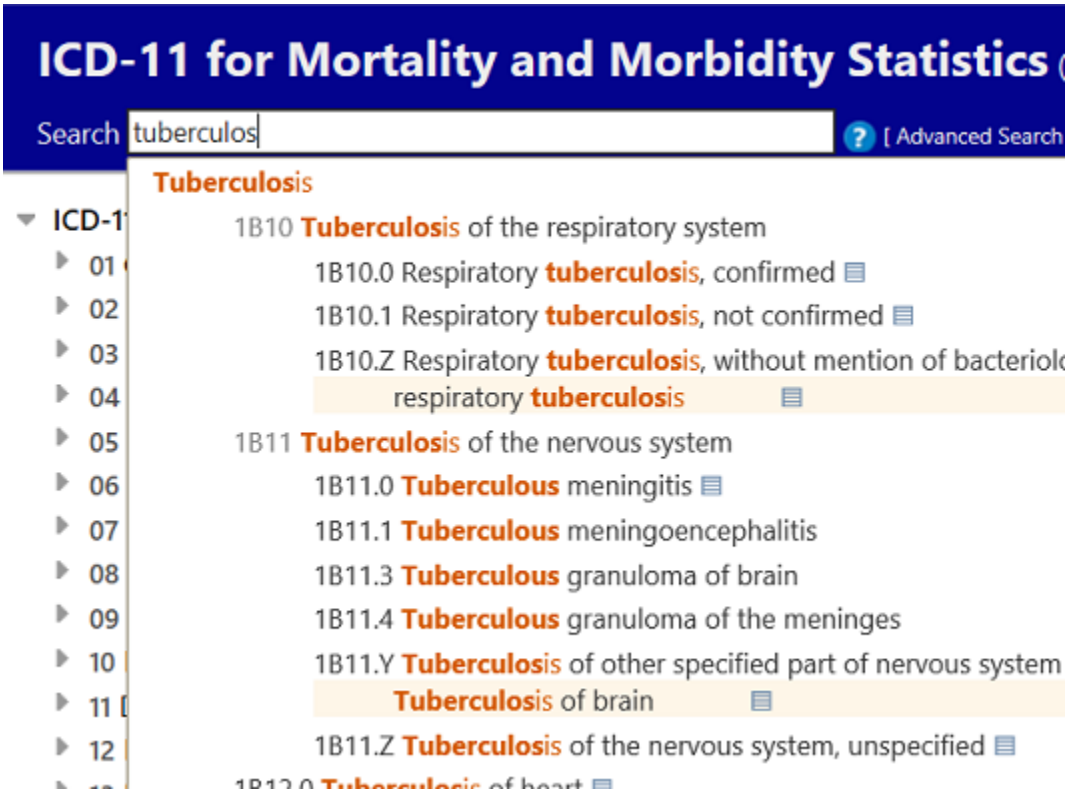


Figure 305: 국제질병분류(ICD) 시스템

장비 정보 등의 메타데이터도 포함

DICOM 메타데이터의 구성 요소: - Patient Information: 환자 ID, 나이, 성별, 체중 등 - Study Information: 검사 날짜, 검사 목적, 임상 정보 - Series Information: 촬영 프로토콜, 영상 파라미터 - Equipment Information: 장비 제조사, 모델명, 소프트웨어 버전 - Image Information: 픽셀 데이터, 영상 크기, 해상도

Figure 15.8 DICOM 표준 기반 의료영상 데이터. 의료영상과 함께 촬영 조건, 환자 정보, 장비 정보 등의 메타데이터가 표준화된 형태로 저장되어 유전체 연구에 활용된다.

비디지털 데이터의 처리: 1장에서 언급했듯이, 모든 의료 데이터가 디지털 포맷으로 저장되어 있지는 않다. 수기 기록이나 필름 데이터가 여전히 존재하며, 이를 유전체 연구에 활용하기 위해서는: - 영상처리 및 필기인식 기술을 통한 디지털 변환 - 자연어 처리 기술을 활용한 텍스트 데이터 추출 및 구조화 - 이러한 과정에서 개인정보 보호를 위한 추가적인 익명화 처리

이러한 다양한 메타데이터는 유전체 데이터와 결합되어 개인의 건강 상태와 질병 위험에 대한 포괄적인 정보를 제공하지만, 동시에 개인 식별 위험을 증가시키는 요소이기도 하다.

16.2.2 메타데이터 보안 처리

유전체 데이터의 메타데이터 처리는 데이터의 과학적 가치를 유지하면서도 개인 식별 위험을 최소화하는 방향으로 이루어진다. 샘플 식별자는 원본 정보와 완전히 분리된 랜덤 코드로 대체되며, 이때 비가역적 해시 함수를 사용하여 원본 정보로의 역추적을 방지한다.

유전적 변이 정보의 경우 개별 변이보다는 집단 수준의 통계적 정보로 변환하여 제공하는 경우가 많다. 희귀 변이나 개인 고유의 변이 패턴은 별도의 보안 처리를 거쳐 연구 목적에 필요한 범위 내에서만 제한적으로 공개된다.

지리적 정보나 인구학적 정보는 광범위한 지역 단위나 인구 집단 단위로 일반화된다. 시간 정보는 정확한 날짜 대신 연도나 분기 단위로 제공되며, 기술적 세부사항은 연구 재현성에 필요한 최소한의 정보만 포함한다.

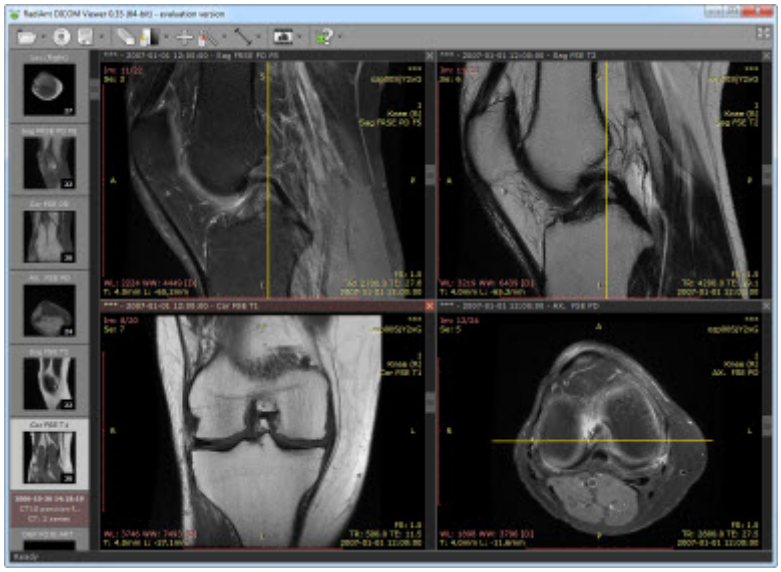


Figure 306: DICOM 표준 기반 의료영상

16.2.3 메타데이터 기반 재식별 위험

유전체 데이터에서 메타데이터는 재식별 위험을 증가시킬 수 있는 요소로 작용한다. 유전체 정보가 개인 고유의 특성을 가지고 있어, 메타데이터와 결합될 경우 개인 식별 가능성이 존재한다.

희귀 유전적 변이와 관련된 메타데이터의 경우 특별한 주의가 필요하다. 낮은 빈도로 나타나는 변이를 가진 개인의 경우, 해당 변이 정보와 인구학적 메타데이터의 조합을 통해 개인 식별이 이루어질 수 있다. 가족력이나 유전적 계보 정보가 메타데이터에 포함된 경우에는 공개된 계보 데이터베이스와의 연계 분석을 통한 개인 특정 가능성이 있다.

따라서 유전체 메타데이터 보안은 개인정보 제거뿐만 아니라 유전체 데이터의 고유한 특성을 고려한 조심스러운 접근이 요구된다.

16.3 통제 접근 저장 시스템

유전체 데이터의 익명화 기술만으로는 완전한 개인정보 보호를 보장하기 어렵다는 한계가 인식되면서, 데이터 접근 자체를 통제하는 방식의 보안 시스템이 개발되었다. 통제 접근 저장 시스템은 데이터를 공개하지 않고 승인된 연구자에게만 제한적으로 접근을 허용함으로써 보안을 강화하면서도, 연구의 효율성을 유지할 수 있도록 설계된 시스템이다.

16.3.1 dbGaP (Database of Genotypes and Phenotypes)

dbGaP는 미국 국립생물정보센터(NCBI)가 운영하는 대표적인 통제 접근형 유전체 데이터베이스로, 연구자들이 대규모 유전체 연구 데이터에 접근할 수 있도록 하면서도 개인정보 보호를 위한 엄격한 통제 시스템을 적용하고 있다.

dbGaP의 접근 통제 과정은 다단계로 이루어진다:

- 연구자 자격 검증: 소속 기관 확인 및 연구 경력 심사
- 연구 계획 심사: 데이터 사용 목적과 방법에 대한 상세한 검토
- 제한적 사용 권한: 승인된 연구 목적에 한정된 데이터 접근
- 사용 추적: 모든 데이터 접근 및 사용 내역의 로깅과 모니터링

이러한 체계적인 과정을 통해 데이터의 오남용을 방지하면서도 연구자들이 필요한 데이터에 접근할 수 있도록 한다.

데이터 보안 측면에서 dbGaP는 연구자의 소속 기관이 데이터 보안 정책 준수를 확약하도록 요구한다. 또한 암호화된 데이터 전송 및 저장을 통한 기술적 보안과 승인된 장소에서만 데이터 접근을 허용하는 물리적 보안 조치를 동시에 적용한다. 이러한 다층적 보안 체계를 통해 연구 데이터의 활용성과 개인정보 보호를 동시에 달성하고 있다.

16.3.2 EGA (European Genome-phenome Archive)

EGA는 유럽연합의 엄격한 개인정보보호 규정(GDPR)을 준수하는 유전체 데이터 저장소로, 유럽의 법적 프레임워크와 문화적 특성을 반영한 독특한 데이터 거버넌스 체계를 갖추고 있다.

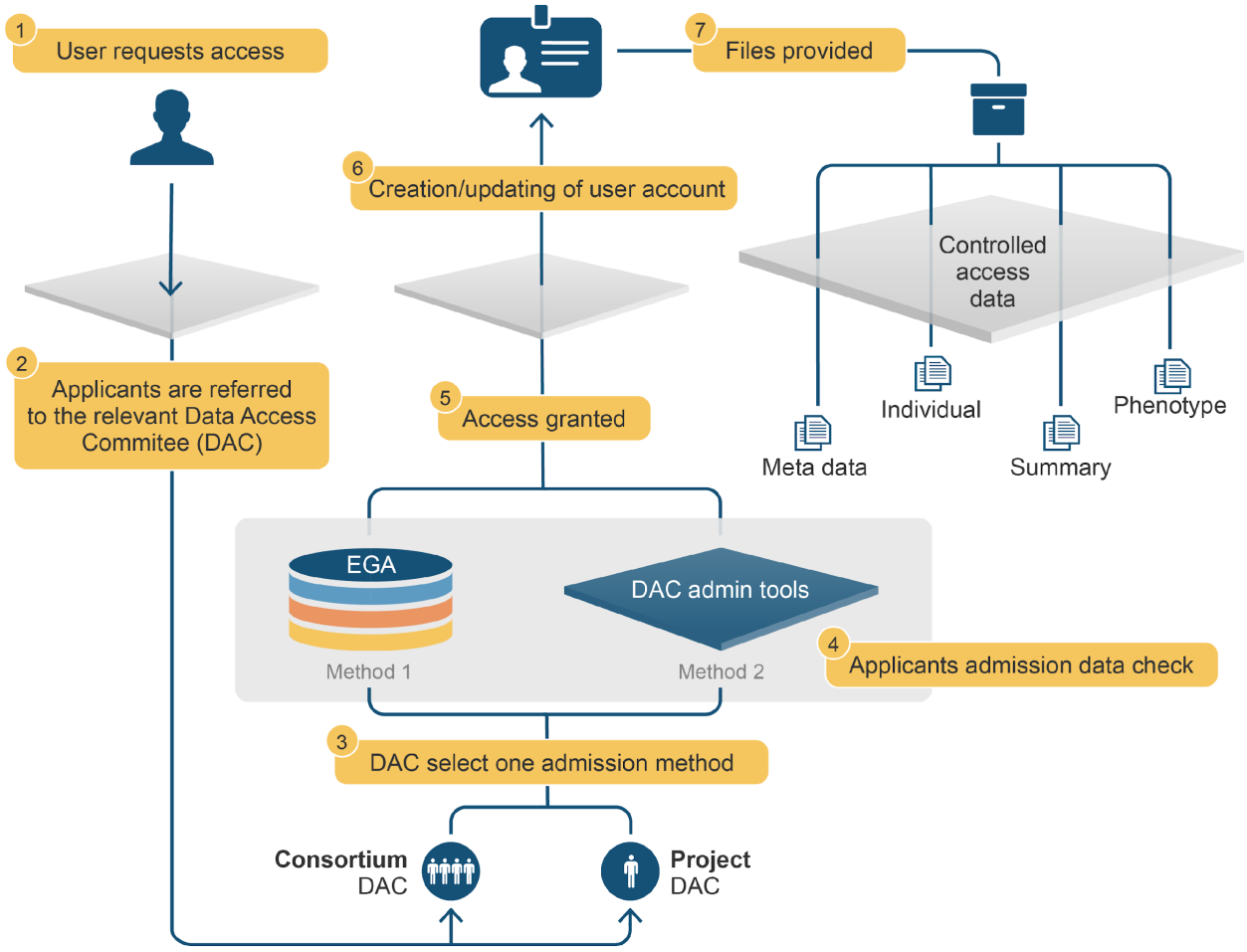


Figure 307: 유럽 유전체-표현형 아카이브(EGA)

Figure 15.9 유럽 유전체-표현형 아카이브(EGA). 유럽의 GDPR 규정을 준수하는 통제 접근형 유전체 데이터베이스로, 독립적인 데이터 접근 위원회를 통한 엄격한 보안 체계를 갖추고 있다.

EGA의 데이터 거버넌스는 유럽식 접근법을 기반으로 한다:

- GDPR 준수: 유럽연합의 일반개인정보보호규정에 따른 엄격한 보호
- 데이터 접근 위원회: 독립적인 전문가 위원회를 통한 접근 승인
- 명확한 승인 범위: 접근 가능한 데이터 범위와 사용 기간의 명시적 규정
- 표준화된 메타데이터: 국제적 호환성을 위한 표준화된 데이터 기술

이러한 체계는 개인정보 보호를 최우선으로 하면서도 연구의 투명성과 재현성을 보장한다.

또한 EGA는 국제 공동연구를 적극 지원한다. 다른 국제 데이터베이스와의 상호 운용성을 확보하고, 국경을 넘나드는 대규모 연구 프로젝트를 지원하며, 각국의 문화적 차이와 법적 요구사항을 고려한 유연한 접근 방식을 채택하고 있다.

16.4 기타 보안 강화 기술

유전체 데이터 보안 분야에서는 암호화와 접근 제어 외에도 다양한 기술들이 개발되고 있다. 데이터의 상호 운용성을 확보하면서도 보안 기술을 융합하여 보다 포괄적인 보안 체계를 구축하는 접근법들이 활용되고 있다.

16.4.1 표준화 기반 보안 강화 기술

OMOP-CDM (Observational Medical Outcomes Partnership - Common Data Model): OMOP-CDM은 서로 다른 의료 기관과 연구소 간의 데이터 통합을 위한 표준 데이터 모델이다. 이 표준은 유전체 데이터 보안에 다음과 같은 이점을 제공한다:

- **표준화된 데이터 구조**: 다양한 소스의 의료 데이터를 일관된 형태로 변환하여 통합 분석 가능
- **개인정보 보호 강화**: 표준화 과정에서 자동적인 식별자 제거 및 데이터 정규화
- **연구 협력 촉진**: 기관 간 데이터 공유 시 개인 식별 위험 최소화
- **품질 관리**: 표준화된 검증 절차를 통한 데이터 품질 보장

폐쇄망 환경에서의 안전한 데이터 접근: OMOP-CDM의 가장 중요한 보안 장점 중 하나는 폐쇄망(closed network) 환경에서의 안전한 데이터 접근을 가능하게 한다는 점이다. CDM 기반 시스템에서는 다음과 같은 방식으로 보안이 강화된다:

- **네트워크 격리**: 의료 데이터가 외부 인터넷과 완전히 분리된 폐쇄망 내에서만 처리
- **표준화된 쿼리 인터페이스**: 연구자들이 직접 원본 데이터에 접근하지 않고도 표준화된 쿼리를 통해 필요한 분석 수행
- **중앙집중식 접근 제어**: 모든 데이터 접근이 중앙 시스템을 통해 로깅되고 모니터링됨
- **분산 분석 환경**: 각 기관이 자체 폐쇄망에서 동일한 CDM 구조를 사용하여 분석 결과만 공유

이러한 폐쇄망 기반 접근 방식은 특히 민감한 유전체 데이터의 경우 다음과 같은 추가적인 보안 이점을 제공한다:

- **물리적 보안**: 데이터가 외부 네트워크에 노출될 위험 원천 차단
- **감사 추적**: 모든 데이터 접근과 분석 활동의 완전한 기록 유지
- **연구 재현성**: 표준화된 환경에서 동일한 분석을 다른 기관에서도 재현 가능
- **규제 준수**: 각국의 의료 데이터 보호 규정을 준수하면서도 국제 협력 연구 가능

16.4.2 블록체인 기반 보안 기술

Nebula Genomics의 블록체인 기반 데이터 주권: Nebula Genomics는 블록체인 기술을 활용하여 개인이 자신의 유전체 데이터를 직접 통제하고 연구자에게 판매할 수 있는 플랫폼을 구현한 대표적인 사례이다. 이 시스템의 핵심 원리는 다음과 같다:

Figure 15.10 Nebula Genomics의 블록체인 기반 데이터 주권 시스템. 개인이 블록체인과 스마트 계약을 통해 자신의 유전체 데이터를 직접 통제하고 연구자에게 판매할 수 있는 플랫폼의 구조를 보여준다.

기술적 구조: - **암호화 키 저장**: 개인의 유전체 데이터 암호화 키가 블록체인에 안전하게 저장됨 - **스마트 계약 기반 자동화**: 미리 정의된 조건이 충족되면 스마트 계약을 통해 자동으로 암호화 키가 발급됨 - **분산 원장**: 모든 데이터 접근과 거래 기록이 변조 불가능한 분산 원장에 기록됨 - **개인 통제권**: 개인이 직접 자신의 유전체 데이터에 대한 접근 권한과 사용 조건을 설정

개인-연구자 직접 거래 모델: 이 시스템은 개인이 연구자에게 직접 유전체 데이터를 판매할 수 있는 플랫폼을 제공한다: - 개인이 데이터 사용 목적, 기간, 범위를 직접 설정 - 연구자가 조건에 동의하면 스마트 계약이 자동 실행 - 데이터 사용료가 블록체인을 통해 투명하게 지급 - 모든 거래 내역이 공개적으로 검증 가능 - 개인이 자신의 유전체 데이터에 대한 완전한 통제권 보유 - 데이터 사용 내역의 투명한 기록과 활용 수익의 공정한 분배

16.5 윤리적 고려사항과 미래 전망

16.5.1 현대의 유전체 데이터 보안 기술

유전체 데이터의 고유한 특성을 고려한 보안 기술들이 지속적으로 개발되고 있다. 이러한 기술들은 유전체 데이터의 민감성과 복잡성을 다루면서도 연구의 과학적 가치를 유지하는 것을 목표로 한다.

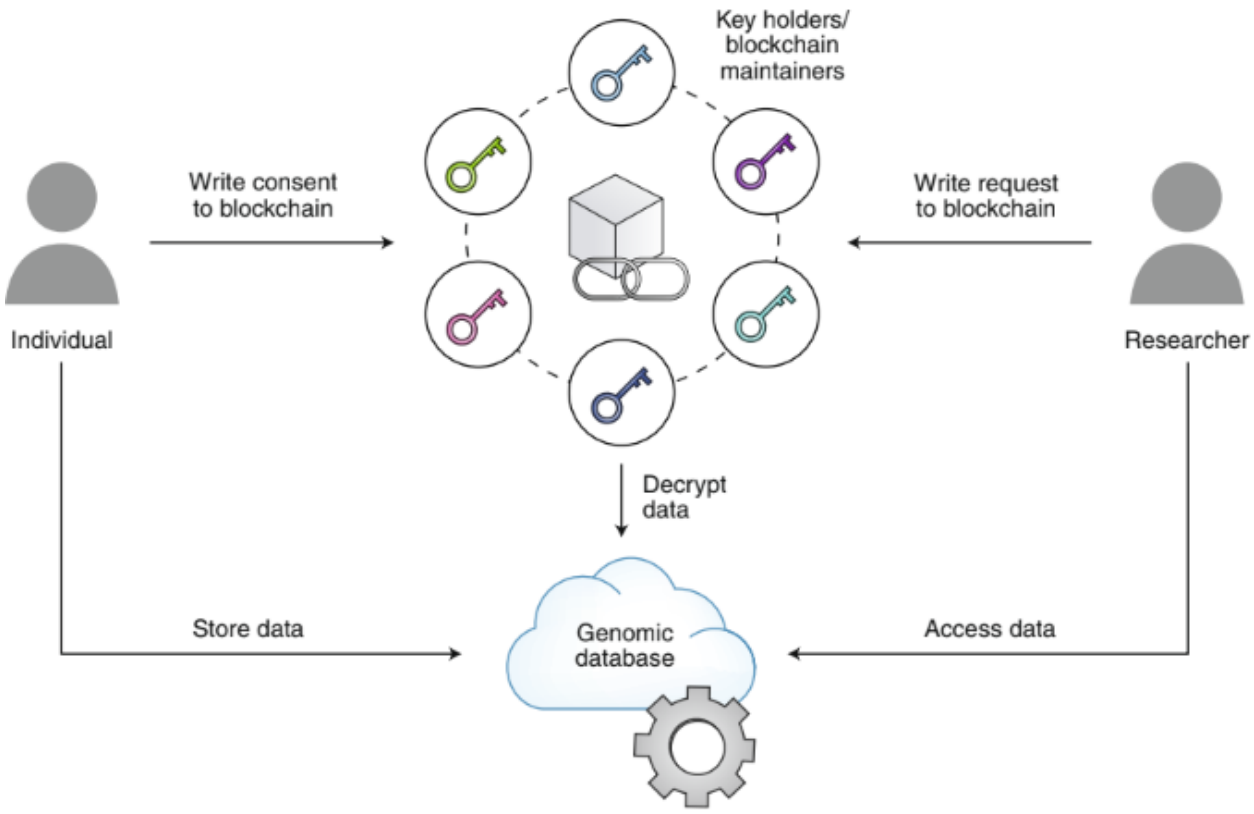


Figure 308: Nebula Genomics의 블록체인 기반 시스템

동형암호화(Homomorphic Encryption)는 데이터를 암호화된 상태로 유지하면서도 연산을 가능하게 하는 기술이다. 유전체 연구 분야에서는 암호화된 서열 데이터에서 직접 변이 검출과 유전자 발현 분석을 수행할 수 있다. GWAS 연구에서는 개인 유전체 정보를 노출하지 않고 집단 수준의 통계 분석이 가능하며, 여러 기관이 각자의 데이터를 공개하지 않으면서도 공동 연구를 수행할 수 있게 한다. 또한 클라우드 환경에서 민감한 유전체 데이터를 안전하게 처리할 수 있어, 연구자들이 실제 유전체 서열에 직접 접근하지 않고도 필요한 분석을 수행할 수 있다.

딥러닝 기반 메타데이터 익명화 기술은 유전체 데이터와 함께 수집되는 복잡한 메타데이터의 패턴을 학습하는 방법이다. 이 기술은 임상 정보, 인구학적 정보, 생활 양식 정보 등의 복합적 패턴을 분석하여 개인 식별 위험이 높은 메타데이터 조합을 자동으로 탐지한다. 기존의 규칙 기반 익명화 방법과 달리, 딥러닝 모델은 연구 목적에 필요한 정보는 유지하면서 재식별 위험을 최소화하는 더욱 정교한 개인정보 보호를 제공할 수 있다.

16.6 결론

유전체 데이터 보안은 현대 의생명정보학의 핵심 영역으로, 과학적 발전과 개인정보 보호 사이의 균형을 요구한다. 유전체 데이터는 개인의 생물학적 지문과 같은 고유성을 가지며, 한 번 공개되면 되돌릴 수 없는 특성을 가진다. 따라서 전통적인 익명화 기법만으로는 충분하지 않으며, 동형암호화, 블록체인, 딥러닝 기반 익명화 등의 혁신적 기술을 통한 보안 강화가 필수적이다.

기술적 진보가 사회적 혜택으로 이어지기 위해서는 윤리적 고려사항이 기술 개발과 함께 발전해야 한다. 개인의 자율성 존중, 사회적 이익과 개인 권리의 균형, 그리고 다학제적 협력을 통한 지속적인 정책 개선이 필요하다. 유전체 데이터 보안 분야의 발전은 단순히 기술적 성취를 넘어서, 인류가 직면한 보건 의료 문제의 해결과 더 나은 미래 사회 구축에 기여할 수 있는 잠재력을 가지고 있다.

17 인공지능과 의생명정보학의 미래

현대 의생명정보학은 인공지능과 기계학습의 발전과 함께 새로운 전환점을 맞고 있다. 지금까지 우리는 유전체학, NGS, GWAS, 암 연구, 신약 개발, 후성유전체학 등 다양한 분야를 살펴보았다. 이 모든 영역에서 생성되는 대용량 데이터를 효과적으로 분석하고 의미 있는 정보를 추출하는 것이 현대 의생명정보학의 핵심 과제가 되었다.

인공지능은 단순히 데이터를 처리하는 도구를 넘어서, 복잡한 생물학적 패턴을 발견하고 새로운 가설을 제시하며, 심지어 치료법을 제안하는 역할까지 담당하고 있다. 본 장에서는 의생명정보학에서 활용되는 주요 인공지능 기술들을 살펴보고, 이들이 어떻게 질병 진단, 치료, 신약 개발 등에 혁신을 가져오고 있는지 알아본다. 또한 앞으로 의생명정보학이 나아가야 할 방향과 인공지능 기술의 발전이 인류의 건강에 미칠 영향에 대해서도 논의한다.

Figure 16.1 기계학습과 딥러닝의 개요. 인공지능의 하위 분야로서 기계학습과 딥러닝의 관계를 보여주며, 의생명정보학에서 활용되는 다양한 AI 기술들의 전체적인 구조를 나타낸다.

17.1 인공지능과 기계학습의 의생명정보학 응용

17.1.1 딥러닝의 기본 원리와 의생명정보학에서의 역할

인공지능에서 학습은 최적화의 과정으로 이해할 수 있다. 모델은 훈련 과정에서 중요하지 않은 정보는 망각하고, 문제 해결에 핵심적인 패턴만을 학습하여 일반화 능력을 획득한다. 이러한 원리는 복잡하고 고차원적인 생명과학 데이터에서 의미 있는 신호를 추출하는 데 활용된다.

Figure 16.2 딥러닝의 기본 구조. 다층 신경망을 통해 복잡한 패턴을 학습하는 딥러닝의 원리를 보여주며, 입력층에서 출력층까지의 정보 흐름과 각 층에서의 특징 추출 과정을 나타낸다.

현대 의생명정보학에서 인공지능의 주요 역할 중 하나는 이질적인 데이터 유형들을 통합하여 분석하는 것이다. DNA 서열, 단백질 구조, 유전자 발현 프로파일, 의료 영상, 그래프 네트워크, 시계열 데이터, 자연어 형태의 임상 기록, 전자의무기록 등 서로 다른 형태와 크기를 가진 데이터들을 “잠재 공간(latent space)” 라고 불리는 공통된 수학적 표현으로 변환한다. 이러한 통합된 표현을 통해 분류, 클러스터링, 시각화, 생성 모델링 등 다양한 분석 작업을 수행할 수 있으며, 단일 데이터 유형만으로는 얻을 수 없는 정보를 제공한다.

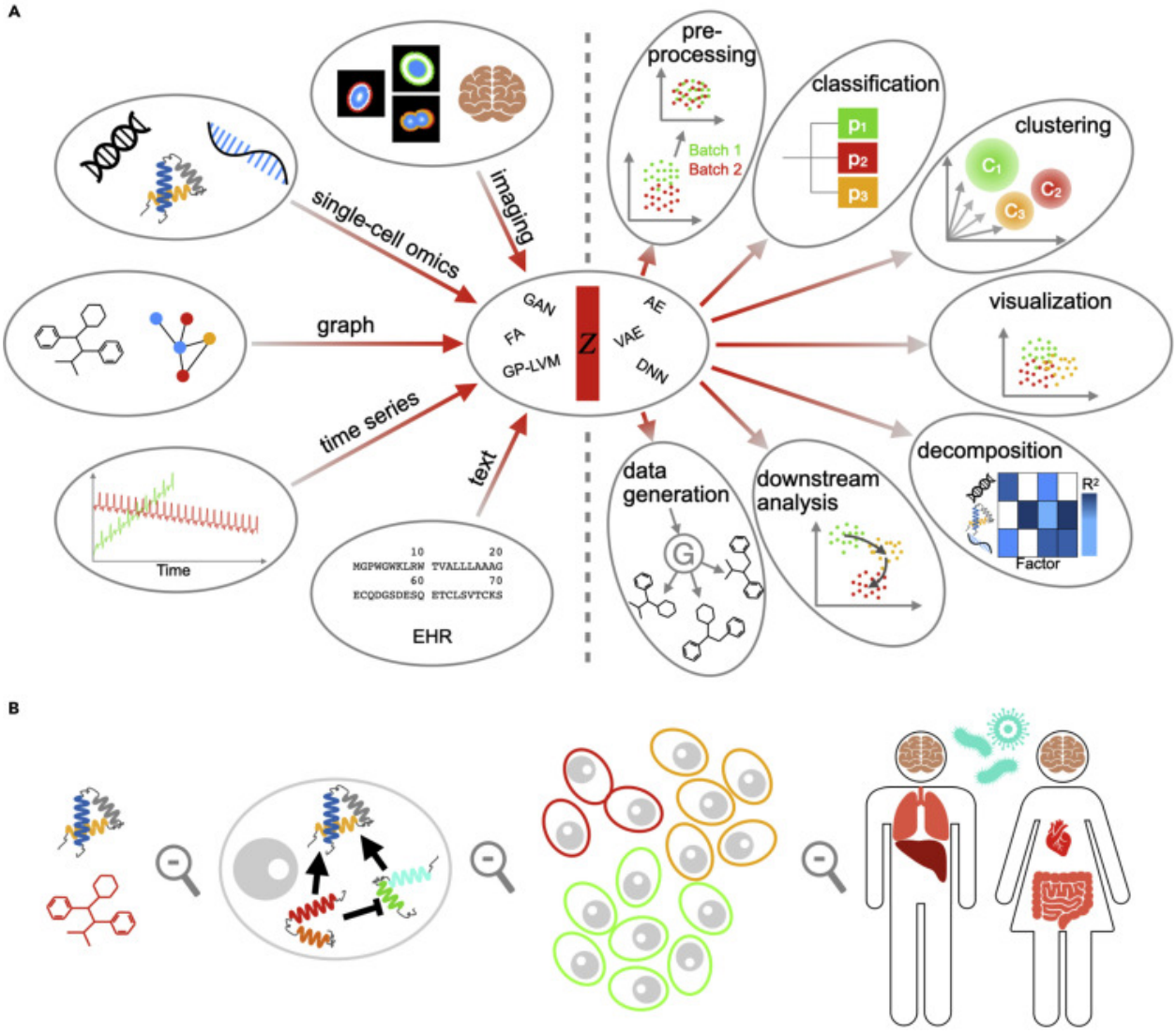


Figure 309: 기계학습과 딥러닝의 개요

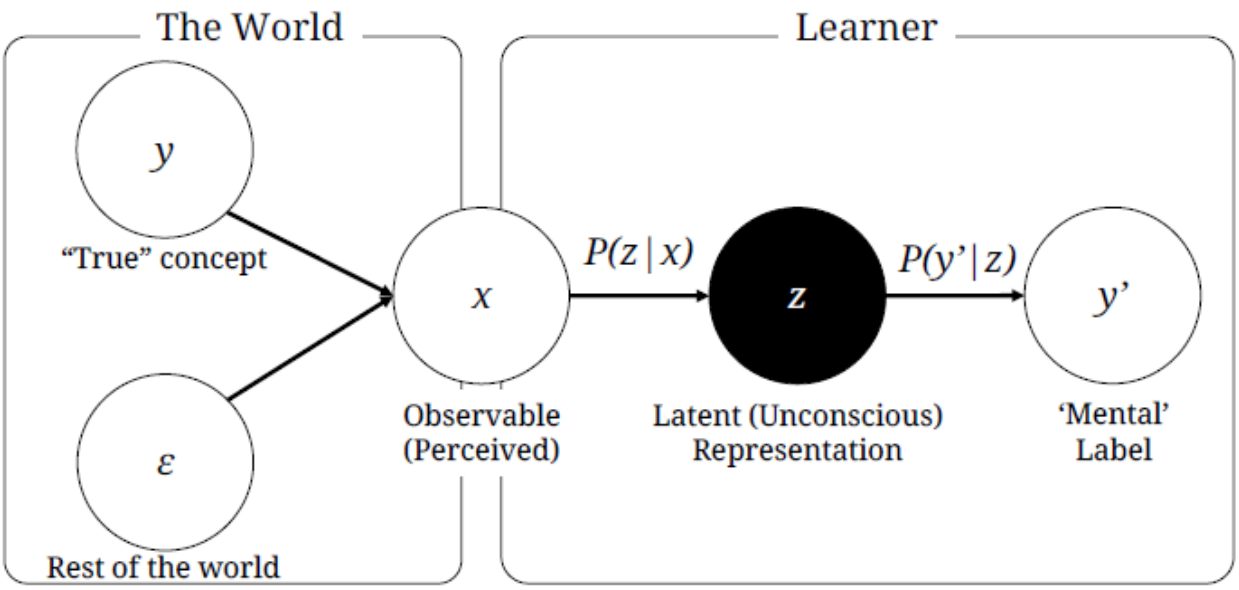


Figure 310: 딥러닝의 기본 구조

17.1.2 차원 축소와 잠재 공간 모델링

생명과학 데이터는 일반적으로 고차원성을 가지고 있다. 단일세포 RNA-seq 데이터는 수만 개의 유전자 차원을 가지며, 의료 영상은 수백만 개의 픽셀로 구성된다. 이러한 고차원 데이터를 직접 분석하는 것은 계산적으로 비효율적일 뿐만 아니라, 차원의 저주 문제로 인해 통계적 분석의 정확성도 떨어뜨린다.

잠재 공간(latent space)은 고차원 데이터의 핵심적인 특징을 저차원으로 압축하여 표현하는 추상적 공간이다. 이 공간에서는 유사한 특성을 가진 데이터들이 가까이 위치하게 되며, 데이터 간의 관계가 더욱 명확하게 드러난다. 예를 들어, 수만 개의 유전자 발현 값을 가진 세포들을 2차원이나 3차원의 잠재 공간으로 매핑하면, 세포 유형이나 발달 단계에 따른 클러스터를 시각적으로 관찰할 수 있다.

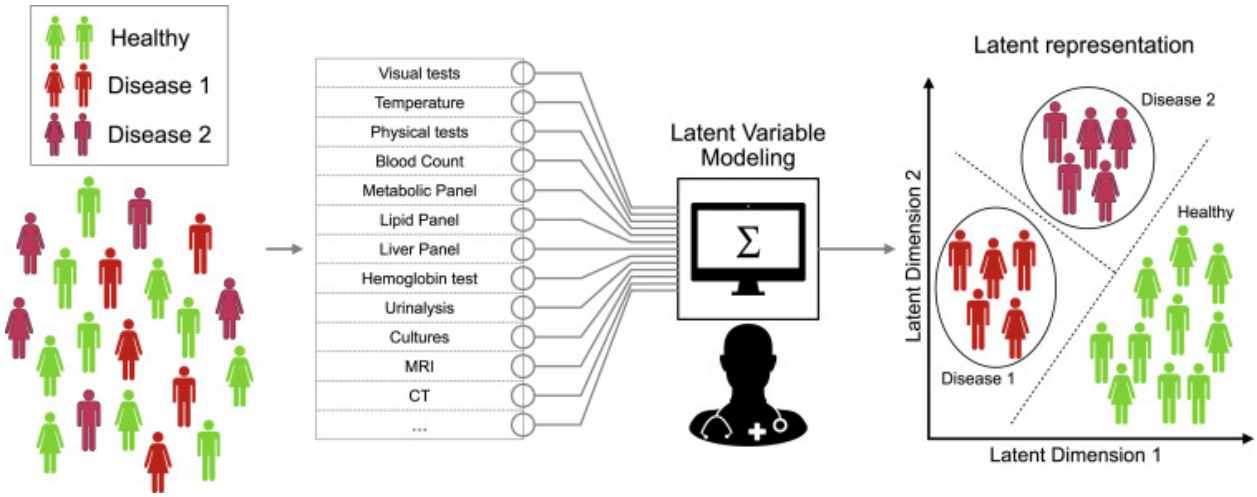


Figure 311: 잠재 공간의 개념

Figure 16.3 잠재 공간의 개념. 고차원 데이터를 저차원의 잠재 공간으로 매핑하는 과정을 시각화하여, 유사한 특성을 가진 데이터들이 잠재 공간에서 가까이 위치하는 원리를 보여준다.

11장에서 다룬 전통적인 차원 축소 방법들과 최신 딥러닝 기법들은 잠재 공간을 구성하는 방식에서 중요한 차이를

보인다. 주성분 분석(PCA), 비음수 행렬 분해(NMF) 등의 전통적인 방법들은 선형 변환을 기반으로 한다. PCA는 분산을 최대화하는 선형 조합을 찾고, NMF는 비음수 제약을 통해 부분-전체 관계를 모델링한다.

반면 딥러닝 기반의 차원 축소 기법들은 비선형 변환을 통해 더욱 복잡하고 유연한 잠재 공간을 구성할 수 있다. 오토인코더, 변분 오토인코더(VAE), 적대적 생성 신경망(GAN) 등은 다층 신경망을 통해 비선형 매핑을 학습한다. 이를 통해 전통적 방법으로는 포착하기 어려운 복잡한 데이터 구조와 패턴을 잠재 공간에 효과적으로 표현할 수 있다. 특히 생명과학 데이터의 경우, 유전자 간의 상호작용이나 세포 상태의 연속적 변화와 같은 비선형적 관계를 더 잘 모델링할 수 있다는 장점이 있다.

17.1.3 베이지안 접근법과 딥러닝의 확률적 원리

딥러닝의 핵심 원리는 베이지안 확률론에 기반한다. 베이지안 확률은 확률을 "믿음의 정도"로 해석하며, 새로운 데이터가 들어올 때마다 기존의 믿음을 업데이트하는 과정이다. 이는 딥러닝 모델이 학습하는 방식과 본질적으로 동일하다.

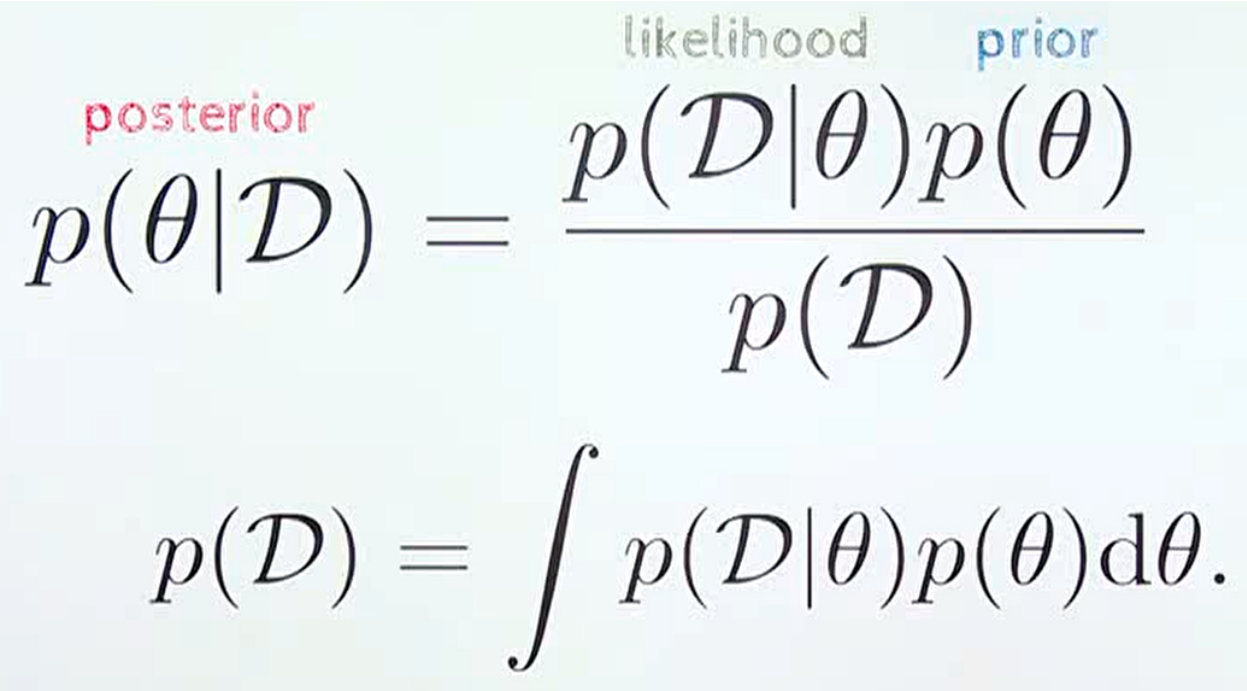


Figure 312: 베이지안 추론의 원리

Figure 16.4 베이지안 추론의 원리. 사전 확률에서 시작하여 새로운 증거를 통해 사후 확률을 업데이트하는 베이지안 학습 과정을 도식화하여, 딥러닝의 확률적 기초를 설명한다.

베이지안 정리는 다음과 같이 표현된다:

P(H|E) = P(E|H) × P(H) / P(E)

여기서 P(H|E)는 증거 E가 주어졌을 때 가설 H의 사후 확률, P(E|H)는 가설 H가 참일 때 증거 E가 관찰될 우도, P(H)는 사전 확률, P(E)는 증거의 주변 확률이다.

딥러닝 모델은 본질적으로 확률 분포를 학습한다. 신경망의 가중치는 확률 분포로 표현될 수 있으며, 새로운 훈련 데이터가 입력될 때마다 이 분포가 조금씩 업데이트된다. 각 층의 뉴런은 입력에 대한 확률적 변환을 수행하고, 전체 네트워크는 복잡한 확률 분포를 근사하게 된다. 따라서, 딥 러닝 모델은 많은 경우 확률분포로 나타내어진다. 이러한 확률적 접근은 모델이 불확실성을 표현하고, 예측의 신뢰도를 제공할 수 있게 한다.

17.1.4 딥러닝 모델의 학습 원리와 손실 함수

딥러닝 모델의 학습, 즉 확률 분포를 업데이트하기 위해서는 공통적으로 손실 함수(loss function)를 최소화하는 방향으로 학습을 진행한다. 손실 함수는 모델의 예측과 실제 정답 간의 차이를 수치화한 지표로, 모델이 얼마나 잘못 예측하고

있는지를 나타낸다. 학습 과정에서는 역전파(backpropagation) 알고리즘을 통해 손실 함수의 기울기를 계산하고, 경사하강법(gradient descent)을 사용하여 손실을 줄이는 방향으로 모델의 가중치를 업데이트한다.

대표적인 손실 함수들은 다음과 같다. 분류 문제에서는 교차 엔트로피 손실(cross-entropy loss)이 주로 사용되며, 회귀 문제에서는 평균 제곱 오차(mean squared error)나 평균 절댓값 오차(mean absolute error)가 활용된다. 특히 생물학적 서열 분석이나 단백질 구조 예측과 같은 복잡한 문제에서는 문제의 특성에 맞는 맞춤형 손실 함수가 설계되기도 한다.

손실 함수: $L(\theta) = f(y_pred, y_true)$ 경사하강법: $\theta \leftarrow \theta - \alpha \nabla L(\theta)$

여기서 θ는 모델의 매개변수, α는 학습률, $\nabla L(\theta)$는 손실 함수의 기울기를 의미한다.

17.2 딥러닝 기반 의생명 데이터 분석

17.2.1 오토인코더

오토인코더는 인코더와 디코더로 구성된 신경망 구조로, 입력 데이터를 저차원의 잠재 공간으로 압축한 후 다시 원본과 유사한 출력으로 복원하는 과정을 학습한다. 이 과정에서 잠재 공간은 입력 데이터의 핵심적인 특징을 압축된 형태로 표현하게 된다. 인코더는 입력 데이터를 잠재 표현으로 변환하고, 디코더는 잠재 표현을 다시 원본 공간으로 복원한다.

$E(x) \rightarrow z$ $D(z) \rightarrow x'$ 훈련 목표: $||x - x'||$ 최소화

여기서 E는 인코더 함수, D는 디코더 함수, x는 입력 데이터, z는 잠재 표현, x'는 복원된 출력을 의미한다.

기본 오토인코더의 확장으로 변분 오토인코더(VAE)와 적대적 오토인코더(AAE) 등의 생성 모델들이 개발되기도 하였으며, 이들은 의생명 데이터 분석에 널리 활용되고 있다.

Figure 16.5 적대적 오토인코더의 구조. 인코더-디코더 구조에 적대적 네트워크를 결합하여 더 나은 잠재 표현을 학습하는 과정을 보여주며, 생성 모델로서의 활용 가능성을 나타낸다.

17.2.2 합성곱 신경망(CNN)

합성곱 신경망(CNN)은 의료 영상 분석에서 가장 성공적인 딥러닝 아키텍처 중 하나이다. CNN은 이미지의 지역적 특징을 효과적으로 추출하는 합성곱 연산과 공간적 차원을 축소하는 풀링 연산을 결합하여 영상의 계층적 특징을 학습한다.

이 중 핵심이 되는 합성곱 연산은 작은 크기의 필터(커널)를 입력 이미지 위에서 슬라이딩하면서 각 위치에서 필터와 이미지 픽셀 값들의 내적을 계산하는 과정을 말한다.

Figure 16.6 합성곱 연산의 원리. 필터(커널)가 입력 이미지 위에서 슬라이딩하면서 지역적 특징을 추출하는 합성곱 연산의 과정을 단계별로 보여준다.

이를 포함한 일반적인 CNN의 핵심 구성 요소는 다음과 같다:

- 합성곱 층(Convolution Layer): 필터(커널)를 이용해 이미지의 지역적 특징을 추출
- 활성화 함수(Activation Function): 비선형성을 도입하여 복잡한 패턴 학습 가능
- 풀링 층(Pooling Layer): 특징 맵의 크기를 줄이고 중요한 정보만 선별
- 완전 연결 층(Fully Connected Layer): 추출된 특징을 최종 분류 결과로 변환

의료 영상에서 CNN은 계층적 특징 추출 과정을 통해 작동한다. 초기 층에서는 가장자리나 텍스처 같은 저수준 특징을 추출하고, 중간 층에서는 해부학적 구조의 부분적 패턴을 학습하며, 최종 층에서는 질병의 특이적 패턴이나 병변의 전체적 형태를 인식한다.

Figure 16.7 합성곱 신경망의 전체 구조. 합성곱층, 활성화 함수, 풀링층, 완전연결층으로 구성된 CNN의 전체 아키텍처와 의료 영상 분석에서의 계층적 특징 추출 과정을 나타낸다.

CNN은 다양한 의료 영상 분야에 활용되고 있다. X-ray 영상에서는 폐렴, 폐결핵, 골절 등을 진단하고, CT 스캔에서는 종양 탐지, 뇌출혈 진단, 폐렴 분석 등을 수행할 수 있다. MRI에서는 뇌종양, 치매, 다발성 경화증 등의 신경계 질환을 진단하며, 병리 슬라이드에서는 암세포 분류와 조직학적 등급 결정 등에 사용된다.

한국의 의료 AI 분야는 세계적 수준의 기술력을 보유하고 있다. 뷰노(VUNO), 루닛(Lunit), 딥노이드(Deepnoid), 제이엘케이(JLK), 삼성전자 등의 기업들이 의료 영상 분석 분야에서 혁신적인 솔루션을 개발하고 있으며, 이들 기업들은 다양한 의료 영상 모달리티와 질병 진단 영역에서 경쟁력 있는 기술을 선보이고 있다.

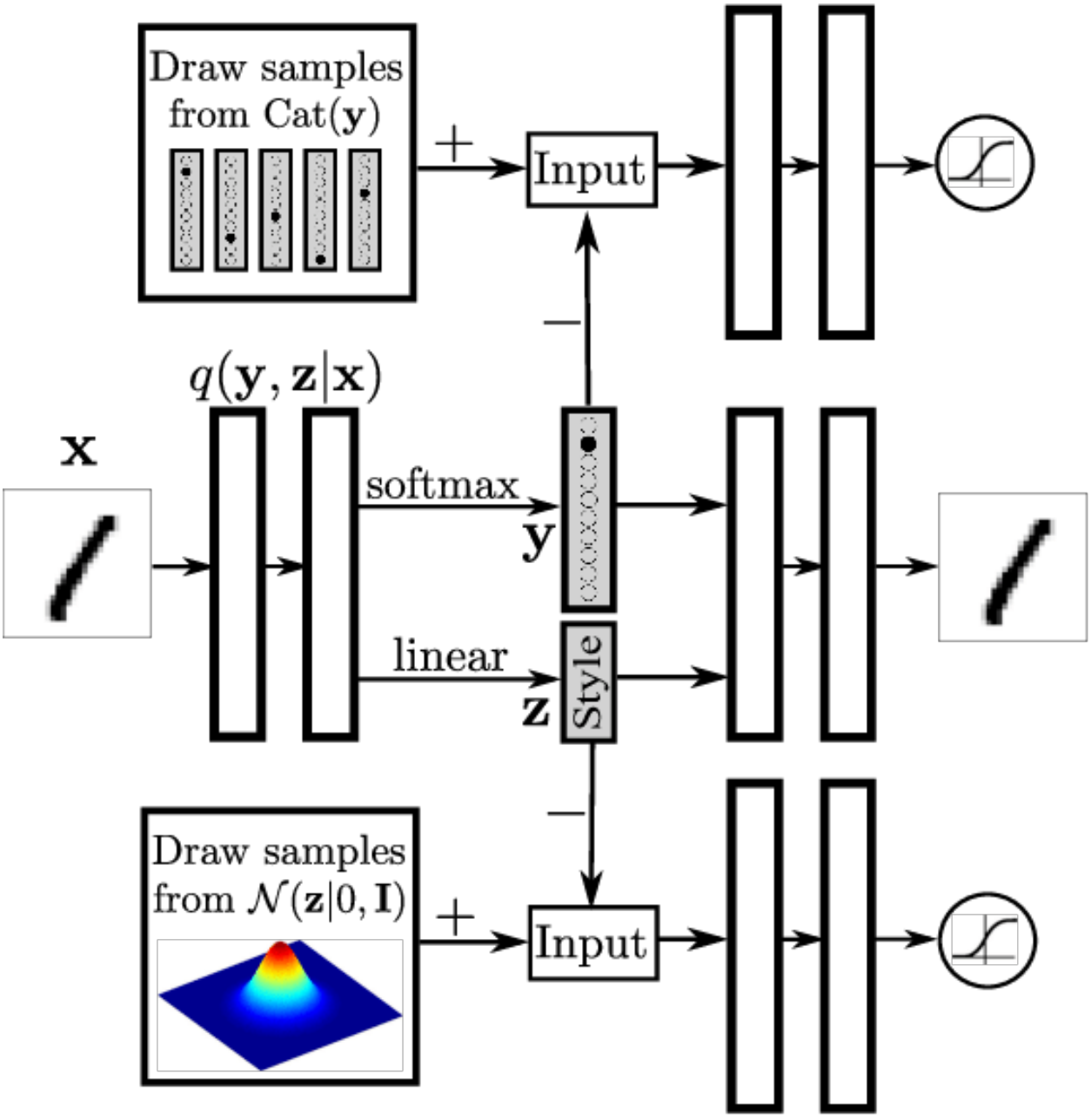


Figure 313: 적대적 오토인코더의 구조

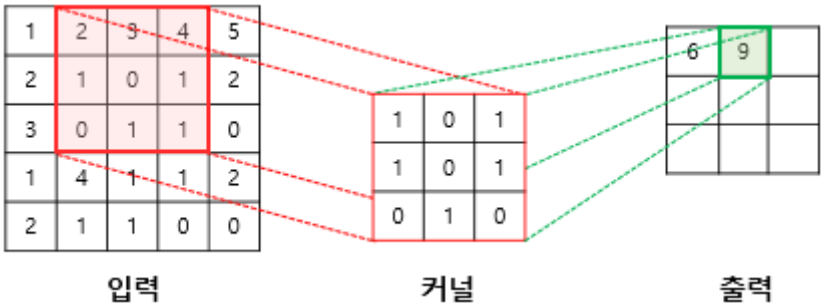


Figure 314: 합성곱 연산의 원리

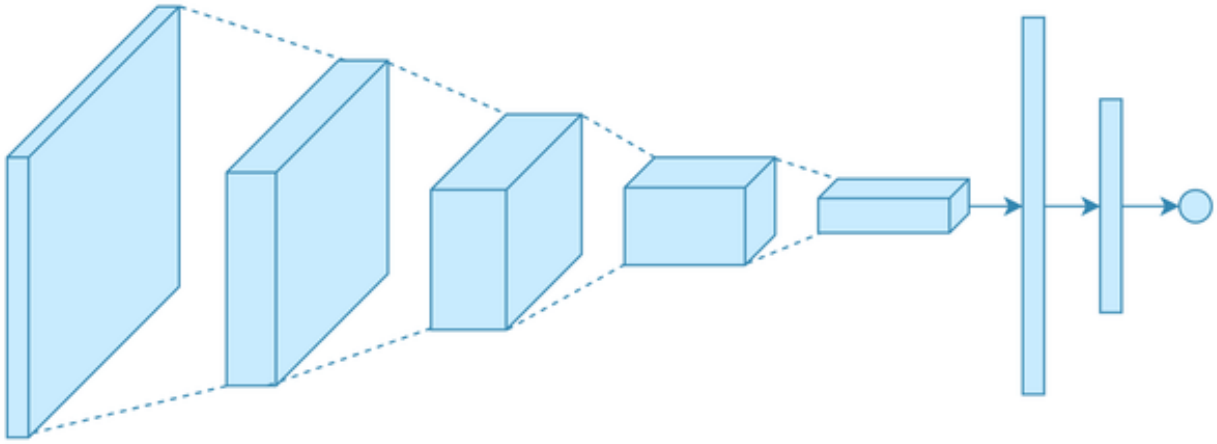
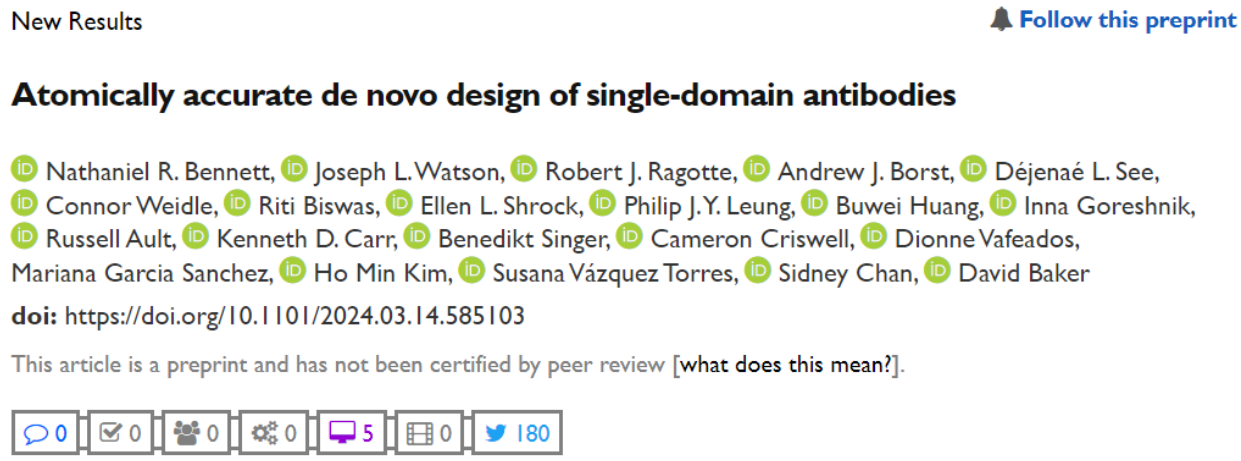


Figure 315: 합성곱 신경망의 전체 구조

17.2.3 확산 모델(Diffusion Models)

확산 모델은 최근 생성형 AI 분야에서 주목받고 있는 새로운 패러다임이다. 이 모델은 노이즈를 점진적으로 추가하는 순방향 과정과 노이즈를 제거하는 역방향 과정을 학습하여 고품질의 데이터를 생성한다. 의생명정보학 분야에서는 다음과 같은 응용이 가능하다.

분자 설계 분야에서 확산 모델은 새로운 약물 후보 분자를 생성하는 데 활용된다. 기존의 알려진 분자 구조에서 시작하여 점진적으로 변형을 가해 새로운 특성을 가진 분자를 설계할 수 있다. 또한 단백질 구조 예측에서도 아미노산 서열로부터 3차원 단백질 구조를 생성하는 데 사용되고 있다.


New Results

Follow this preprint

Atomically accurate de novo design of single-domain antibodies

Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, Andrew J. Borst, Déjenaé L. See, Connor Weidle, Riti Biswas, Ellen L. Shrock, Philip J. Y. Leung, Buwei Huang, Inna Goreshnik, Russell Ault, Kenneth D. Carr, Benedikt Singer, Cameron Criswell, Dionne Vafeados, Mariana Garcia Sanchez, Ho Min Kim, Susana Vázquez Torres, Sidney Chan, David Baker

doi: https://doi.org/10.1101/2024.03.14.585103

This article is a preprint and has not been certified by peer review [what does this mean?].

0 0 0 0 5 0 180

Figure 316: RFdiffusion 단백질 설계

Figure 16.8 RFdiffusion을 이용한 단백질 설계. 확산 모델을 활용하여 노이즈에서 시작해 점진적으로 새로운 단백질 구조를 생성하는 과정을 보여주며, 합성 생물학에서의 응용 가능성을 나타낸다.

의료 영상 분야에서는 부족한 의료 데이터를 보완하기 위한 합성 데이터 생성에 활용된다. 예를 들어, 희귀 질병의 영상 데이터가 부족할 때 확산 모델을 통해 현실적인 합성 영상을 생성하여 진단 모델의 성능을 향상시킬 수 있다. 이는 의료 AI 모델의 훈련 데이터 부족 문제를 해결하는 중요한 접근법으로 주목받고 있다.

17.2.4 순환 신경망(RNN)

순환 신경망(RNN)은 시간적 의존성을 가진 데이터를 처리하는 데 특화된 신경망이다. 일반적인 신경망과 달리 RNN은 이전 시점의 정보를 기억하여 현재 시점의 예측에 활용할 수 있다. 특히 LSTM(Long Short-Term Memory)은 장기 의존성 문제를 해결하여 긴 시퀀스의 패턴을 효과적으로 학습할 수 있다.

Figure 16.9 순환 신경망의 기본 구조. 시간적 의존성을 가진 데이터를 처리하기 위해 이전 시점의 정보를 현재 시점으로 전달하는 RNN의 기본 원리를 보여준다.

Figure 16.10 LSTM의 구조. 장기 의존성 문제를 해결하기 위한 LSTM의 게이트 메커니즘(forget gate, input gate, output gate)과 셀 상태의 흐름을 나타낸다.

Figure 16.11 LSTM의 상세 동작 원리. 각 게이트(forget, input, output)의 상세한 작동 메커니즘과 셀 상태와 숨겨진 상태의 업데이트 과정을 단계별로 보여준다.

의료 분야에서 RNN은 다양한 시계열 데이터 분석에 활용되고 있다. 생체 신호 모니터링에서는 심전도(ECG), 뇌전도(EEG), 혈압 변화 등의 실시간 분석을 통해 이상 신호를 감지한다. 임상 기록 분석에서는 환자의 진료 기록, 투약 이력, 검사 결과의 시간적 패턴을 분석하여 의미 있는 정보를 추출한다. 또한 만성 질환의 경과 예측이나 합병증 발생 위험도 평가, 약물 치료나 수술 후 회복 과정의 추적 등에도 사용된다.

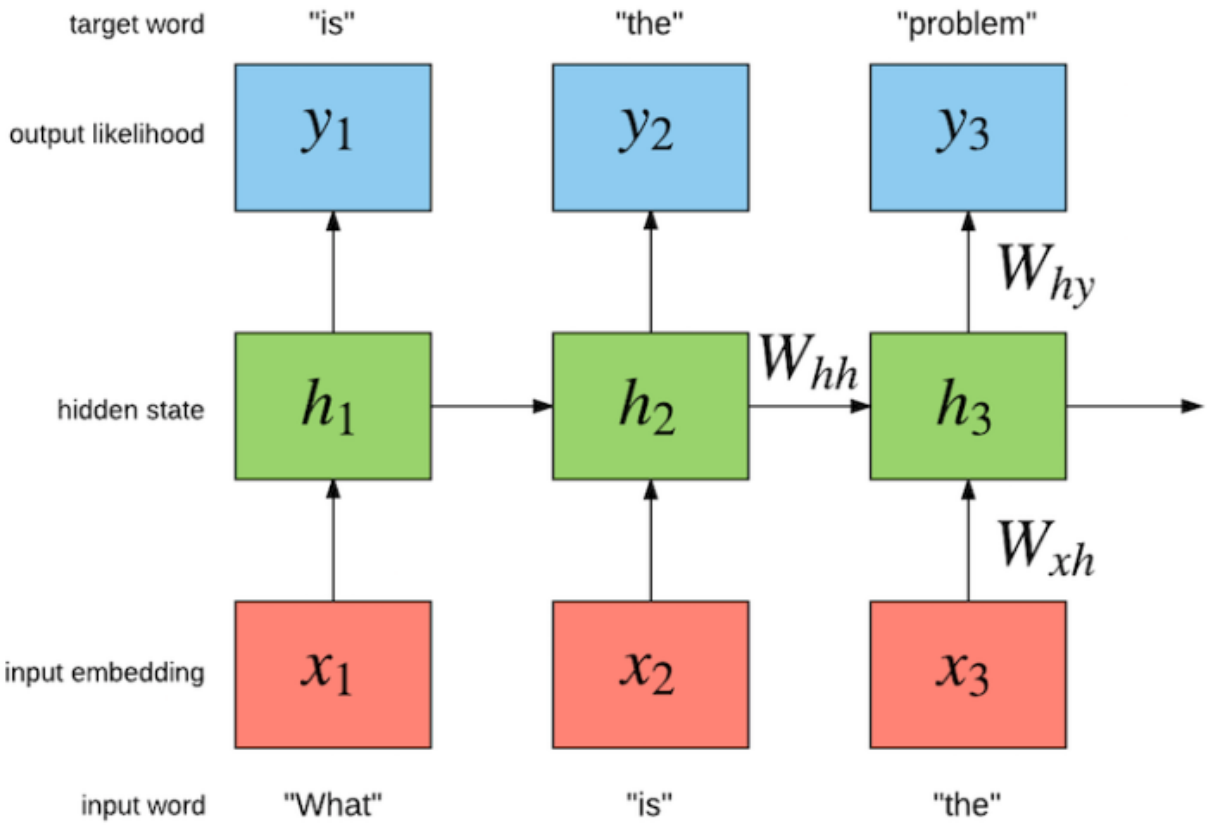


Figure 317: 순환 신경망의 기본 구조

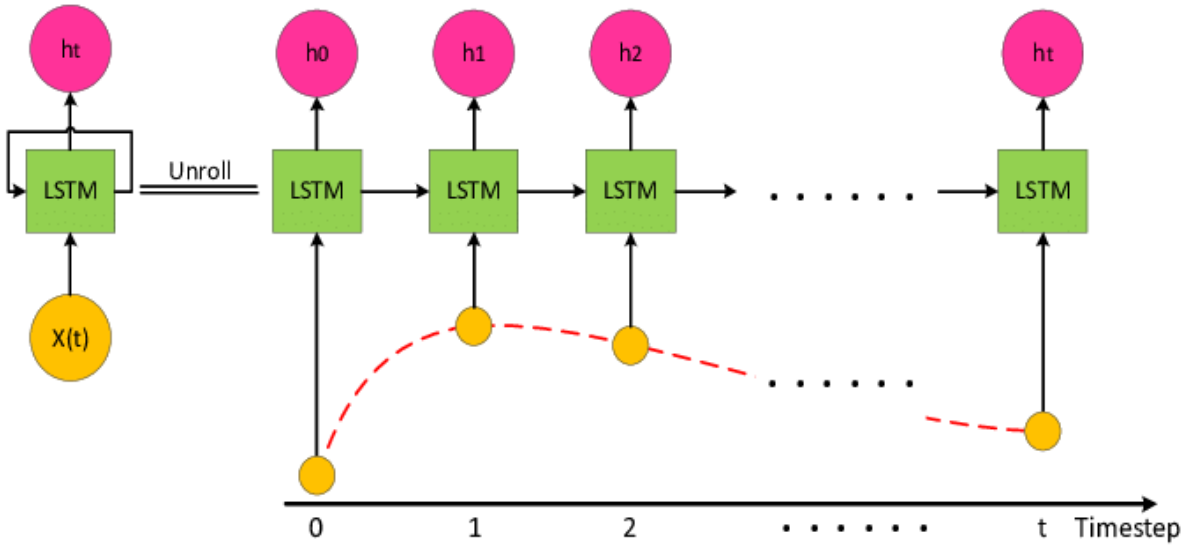


Figure 318: LSTM의 구조

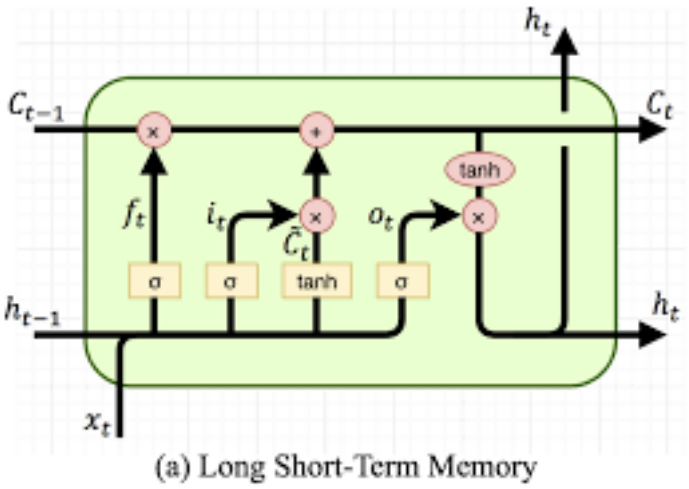


Figure 319: LSTM의 상세 동작 원리

17.2.5 트랜스포머와 대규모 언어 모델

지금까지 우리는 CNN을 통한 의료 영상 분석, 확산 모델을 통한 분자 설계, RNN을 통한 시계열 데이터 분석을 살펴보았다. 이제 현재 AI 분야에서 가장 주목받고 있는 트랜스포머 아키텍처와 그 의생명정보학 응용에 대해 알아보자.

앞서 다룬 RNN과 LSTM이 순차적 데이터 처리에 강점을 보였지만, 긴 시퀀스에서는 여전히 한계가 있었다. 이를 해결하기 위해 등장한 것이 seq2seq(sequence-to-sequence) 모델이다. seq2seq는 인코더-디코더 구조로 구성되며, 인코더에서 입력 시퀀스를 압축된 표현으로 변환하고, 디코더에서 이를 출력 시퀀스로 변환한다. 초기 seq2seq 모델은 LSTM을 기본 블록으로 사용했지만, 긴 시퀀스에서 정보 손실 문제가 발생했다.

이 문제를 해결하기 위해 seq2seq에 어텐션 메커니즘이 도입되었다. 어텐션은 간단히 말해 "주의를 기울이는" 기능으로, 모델이 출력을 생성할 때 입력 데이터의 어느 부분에 집중해야 하는지를 결정한다. 예를 들어, 긴 DNA 서열을 분석할 때 전체 서열 중에서 현재 분석하고 있는 부분과 관련된 중요한 위치들에 더 많은 주의를 기울이는 것이다.

트랜스포머(Transformer)는 이러한 발전의 정점으로, LSTM 구조를 완전히 제거하고 어텐션 메커니즘만으로 구성된 아키텍처를 제시했다. 트랜스포머는 LSTM의 순차적 처리 방식과 달리 병렬 처리가 가능하여 학습 속도가 빠르고, 더 긴 시퀀스를 처리할 수 있어 자연어 처리 분야를 혁명적으로 변화시켰다.

Figure 16.12 트랜스포머 아키텍처. 어텐션 메커니즘을 기반으로 한 인코더-디코더 구조와 멀티-헤드 어텐션, 피드포워드 네트워크로 구성된 트랜스포머의 전체 구조를 나타낸다.

GPT(Generative Pre-trained Transformer)는 트랜스포머 아키텍처를 기반으로 한 생성형 언어 모델의 일반적인 명칭이다. GPT는 "생성형(Generative)" "사전 훈련된(Pre-trained)" "트랜스포머(Transformer)"의 줄임말로, 대량의 텍스트 데이터로 사전 훈련되어 텍스트를 생성할 수 있는 모델을 의미한다. 우리가 잘 아는 ChatGPT는 OpenAI에서 개발한 GPT 시리즈 중 하나이며, GPT라는 용어 자체는 ChatGPT만을 지칭하는 것이 아니라 이러한 구조를 가진 모든 모델을 포괄하는 용어이다. GPT 모델들은 사전 훈련 후 특정 작업에 맞게 미세 조정되어 인간과 자연스러운 대화가 가능한 수준까지 발전했다.

트랜스포머 기술은 생명과학 분야에서도 다양하게 활용되고 있다. DNA 서열을 분석하는 DNABERT, 단백질 기능을 예측하는 ProteinBERT, 화학 분자의 특성을 예측하는 ChemBERTa, 단일세포 데이터를 분석하는 scBERT와 scGPT, 공간 전사체학 분석을 위한 Nicheformer 등이 개발되어 각각의 전문 분야에서 성과를 보이고 있다. 이들 모델은 모두 자연어를 처리하는 것처럼 생물학적 서열이나 데이터를 "언어"로 취급하여 분석한다.

Figure 16.13 DNA 언어 모델의 개념. DNA 서열을 자연어처럼 처리하는 대규모 언어 모델의 원리를 보여주며, 생물학적 서열 데이터를 토큰화하여 분석하는 과정을 나타낸다.

Figure 16.14 scGPT 단일세포 분석 모델. 단일세포 RNA-seq 데이터를 GPT 아키텍처로 분석하는 scGPT 모델의 구조와 세포 유형 분류, 유전자 발현 예측 등의 응용 분야를 보여준다.

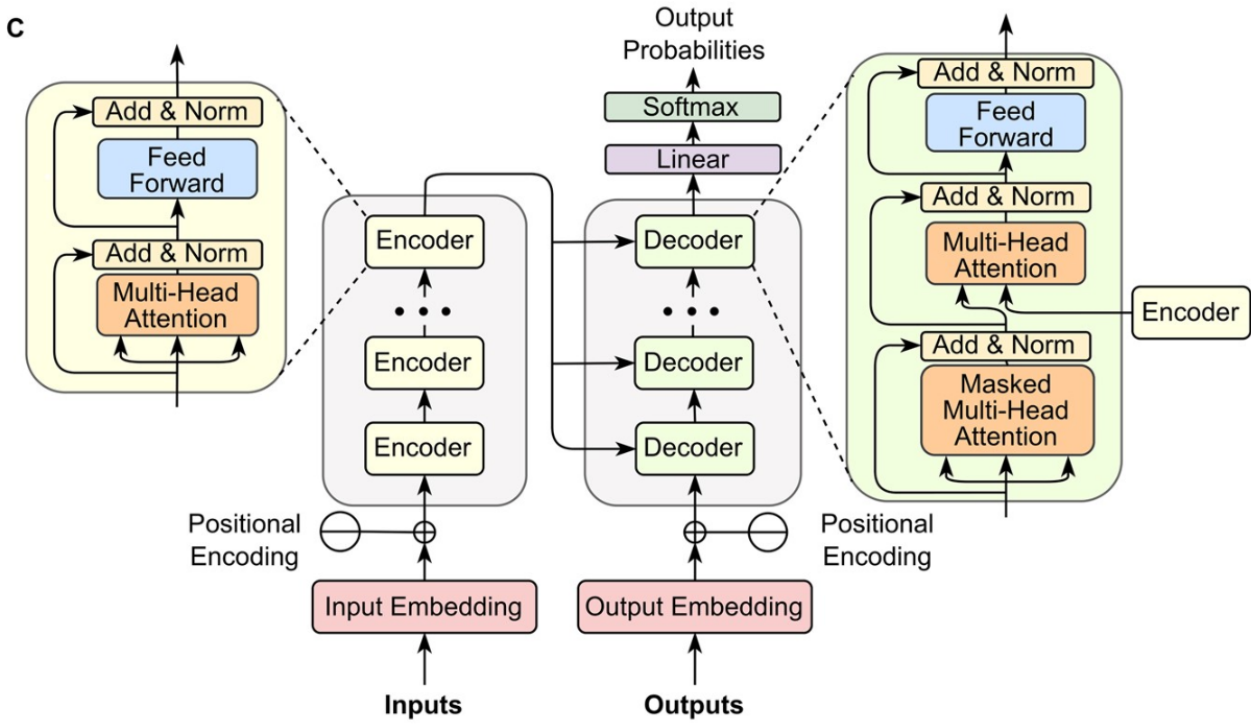


Figure 320: 트랜스포머 아키텍처

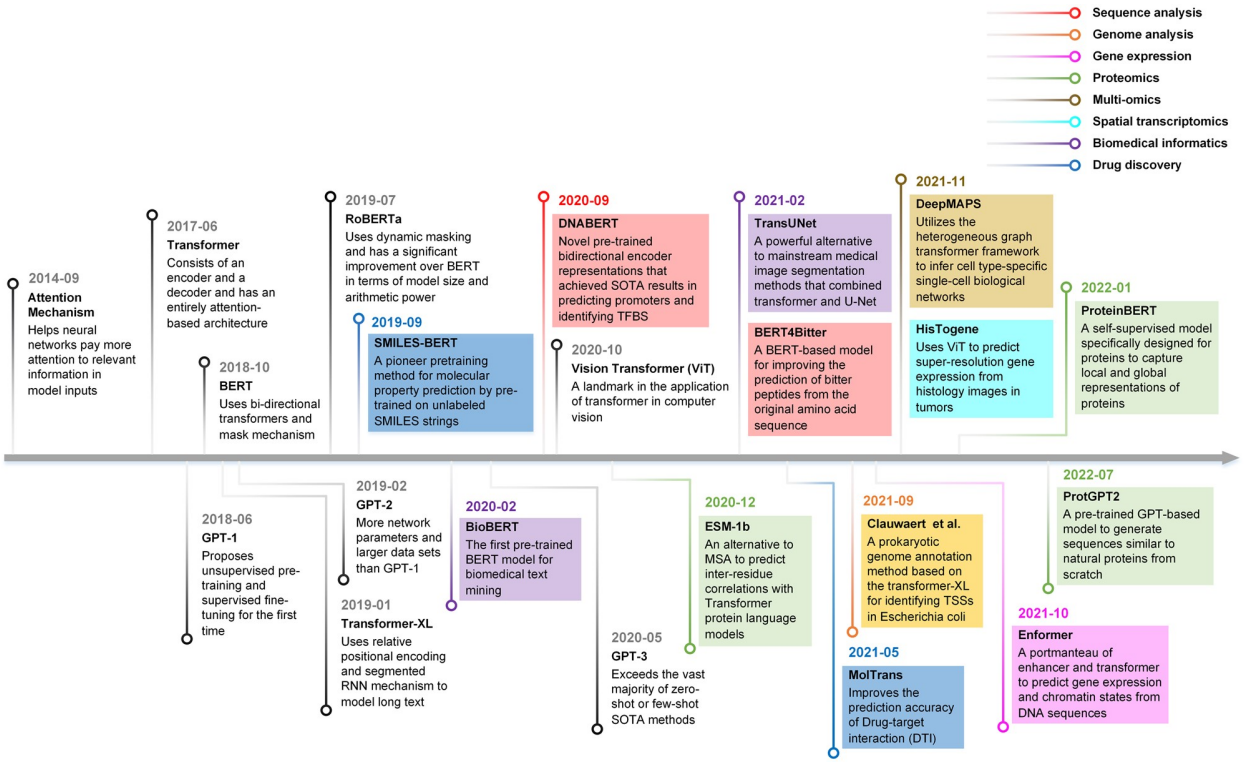


Figure 321: DNA 언어 모델의 개념

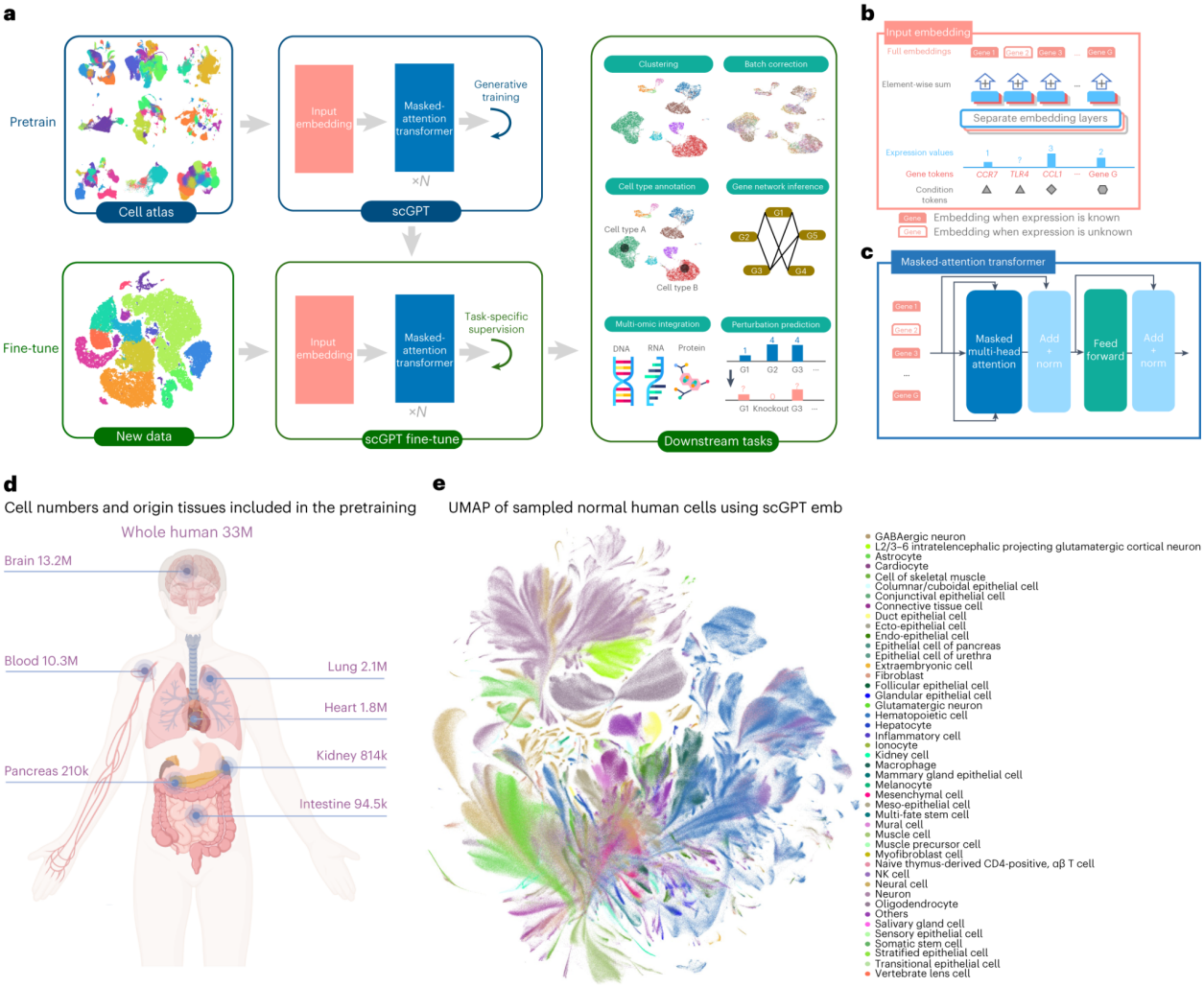


Figure 322: scGPT 단일세포 분석 모델

이러한 기술들은 실제 연구와 임상에서 다양하게 응용되고 있다. 유전체 연구에서는 DNA의 기능이 알려지지 않은 부분을 분석하고, 새로운 단백질을 설계하며, 신약 개발에서는 약물과 질병 표적 간의 상호작용을 예측한다. 또한 서로 다른 종류의 오믹스 데이터를 통합하여 분석하거나, 조직 내에서 유전자가 어떻게 발현되는지 공간적 패턴을 분석하는 데도 활용된다.

우리가 일상에서 사용하는 ChatGPT와 같은 생성형 AI도 의료 분야에서 활용되고 있다. 의료진이 복잡한 의학 문헌을 빠르게 요약하거나, 진단을 보조하거나, 환자와의 상담을 도와주는 역할을 한다. 실제로 부산대학교병원에서는 의사가 작성하는 진료 기록의 오타를 자동으로 수정하고, 의학 용어를 표준화하며, 진단 코드를 자동으로 매핑하는 AI 시스템을 도입하여 의료진의 업무 효율성을 높이고 있다.

17.2.6 트랜스포머를 넘어선 새로운 아키텍처

최근에는 트랜스포머 아키텍처의 한계를 극복하기 위한 새로운 모델들이 등장하고 있다. 특히 생물학적 서열 분석에서는 기존 트랜스포머보다 효율적이고 강력한 성능을 보이는 차세대 아키텍처들이 주목받고 있다.

Evo2는 이러한 차세대 모델의 대표적인 예시로, Arc Institute에서 개발한 대규모 DNA 기반 모델이다. Evo2는 기존 트랜스포머의 계산 복잡도 문제를 해결하기 위해 새로운 아키텍처를 채택했으며, 최대 131만 개의 토큰까지 처리할 수 있는 긴 컨텍스트 처리 능력을 갖추고 있다. 이는 전체 박테리아 유전체나 인간 염색체의 큰 부분까지 한 번에 분석할 수 있는 수준이다. Evo2는 생성 모델로서 새로운 DNA 서열을 설계할 수 있는 능력도 갖추고 있어, 합성 생물학 분야에서의 활용 가능성도 크다.

17.3 인공지능을 이용한 단백질 구조 예측

17.3.1 단백질 구조 예측의 중요성과 역사

단백질의 3차원 구조는 그 기능을 결정하는 핵심 요소이다. 단백질은 아미노산이 일차원적으로 연결된 서열로 시작하지만, 이들이 접혀서 만드는 3차원 구조가 실제 생물학적 기능을 수행한다. 따라서 단백질의 구조를 정확히 예측할 수 있다면 신약 개발, 질병 이해, 효소 설계 등 다양한 분야에 혁신적인 변화를 가져올 수 있다.

Figure 16.15 COVID-19 스파이크 단백질 구조. SARS-CoV-2 바이러스의 스파이크 단백질 3차원 구조를 보여주며, 단백질 구조 예측의 중요성과 신약 개발에서의 활용 가능성을 나타낸다.

그동안 단백질 구조를 실험적으로 결정하는 방법으로는 X선 결정학, 핵자기공명분광법(NMR), 전자현미경 등이 사용되어 왔지만, 이들 방법은 많은 시간과 비용이 소요되며 모든 단백질에 적용하기 어렵다는 한계가 있었다. 이러한 배경에서 컴퓨터를 이용한 단백질 구조 예측 연구가 수십 년간 진행되어 왔으며, 2024년 노벨 화학상은 바로 이 분야의 혁신적 성과를 인정받은 것이다.

2024년 노벨 화학상은 데이비드 베이커(David Baker)와 구글 딥마인드의 데미스 하사비스(Demis Hassabis), 존 점퍼(John Jumper)에게 수여되었다. 베이커는 단백질 설계 분야의 선구자로서 Rosetta 소프트웨어를 개발하여 새로운 단백질을 설계하는 기반을 마련했고, 하사비스와 점퍼는 AlphaFold를 개발하여 단백질 구조 예측 문제를 혁신적으로 해결했다.

단백질 데이터 뱅크(PDB)의 역할: 현대 단백질 구조 예측의 혁신은 단백질 데이터 뱅크(Protein Data Bank, PDB)라는 중요한 기반 데이터베이스가 있었기 때문에 가능했다. PDB는 1971년에 설립된 세계 최초의 생체분자 구조 데이터베이스로, 실험적으로 결정된 단백질, 핵산, 복합체의 3차원 구조 정보를 수집하고 공유하는 글로벌 저장소이다.

Figure 16.16 단백질 데이터 뱅크(PDB). 전 세계 단백질 구조 정보를 수집하고 공유하는 PDB 데이터베이스의 인터페이스를 보여주며, AI 단백질 구조 예측 모델의 훈련 데이터로서의 역할을 나타낸다.

PDB는 X선 결정학, 핵자기공명분광법(NMR), 전자현미경 등을 통해 결정된 수십만 개의 단백질 구조를 보유하고 있으며, 이러한 방대한 실험 데이터가 AlphaFold와 같은 AI 모델의 훈련 데이터로 활용되었다. 특히 AlphaFold는 PDB에 축적된 구조 정보와 진화적 관련성을 가진 단백질들의 서열 정보를 결합하여 학습함으로써 혁신적인 성과를 달성할 수 있었다. 즉, 지난 50여 년간 과학자들이 실험을 통해 축적한 PDB의 구조 데이터가 AI 기반 단백질 구조 예측 혁명의 토대가 된 것이다.

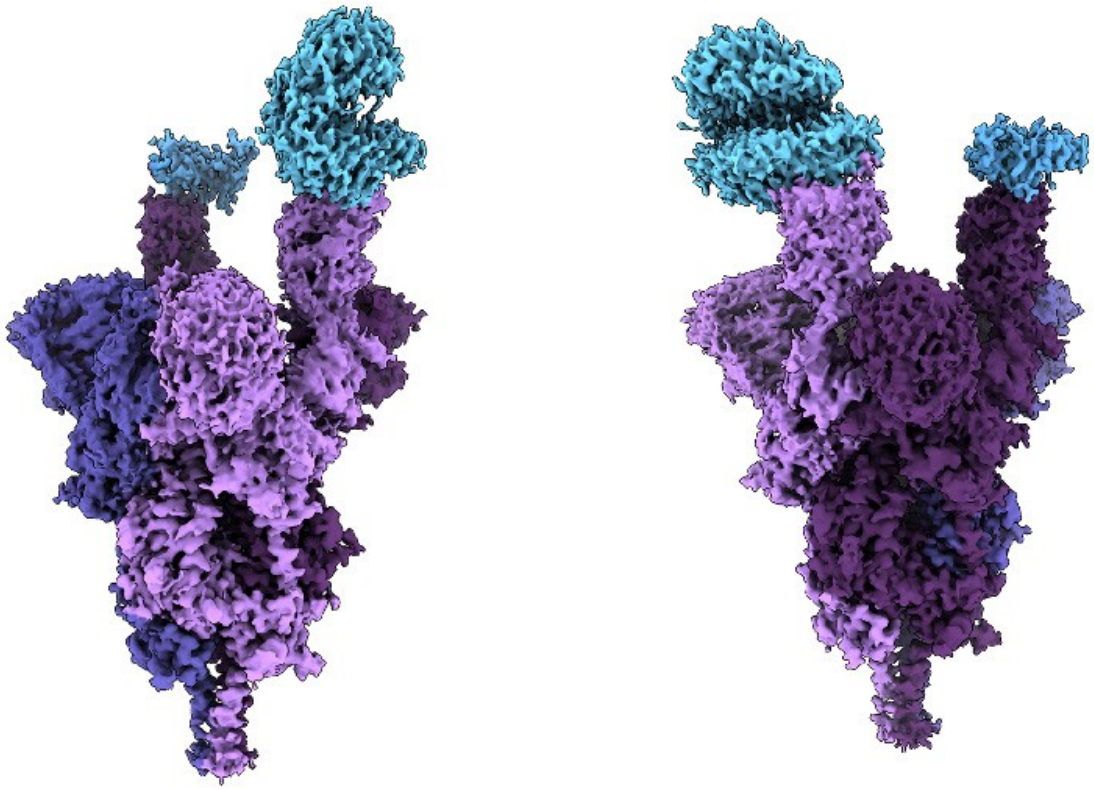


Figure 323: COVID-19 스파이크 단백질 구조

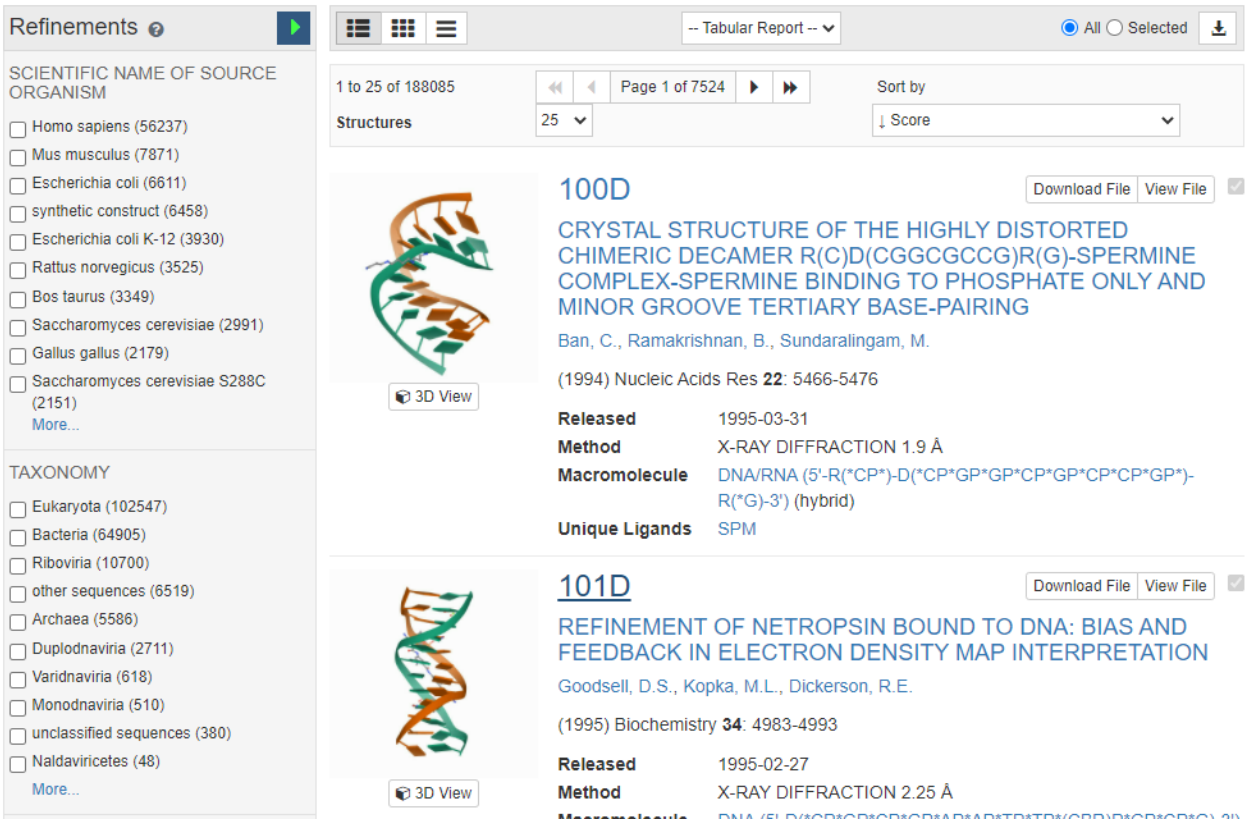


Figure 324: 단백질 데이터 뱅크(PDB)

17.3.2 AlphaFold의 혁신

AlphaFold는 구글 딥마인드에서 개발한 딥러닝 기반 단백질 구조 예측 시스템으로, 2020년 CASP14(Critical Assessment of protein Structure Prediction) 경진대회에서 기존 방법들을 압도적으로 앞서는 성과를 보여주었다. AlphaFold는 아미노산 서열만을 입력으로 받아 단백질의 3차원 구조를 예측하며, 실험적으로 결정된 구조와 거의 동일한 수준의 정확도를 달성했다.

Figure 325: AlphaFold 시스템

Figure 16.17 AlphaFold 시스템. 구글 딥마인드에서 개발한 AlphaFold의 주요 기능과 인터페이스를 보여주며, 아미노산 서열로부터 3차원 단백질 구조를 예측하는 혁신적인 AI 시스템을 나타낸다.

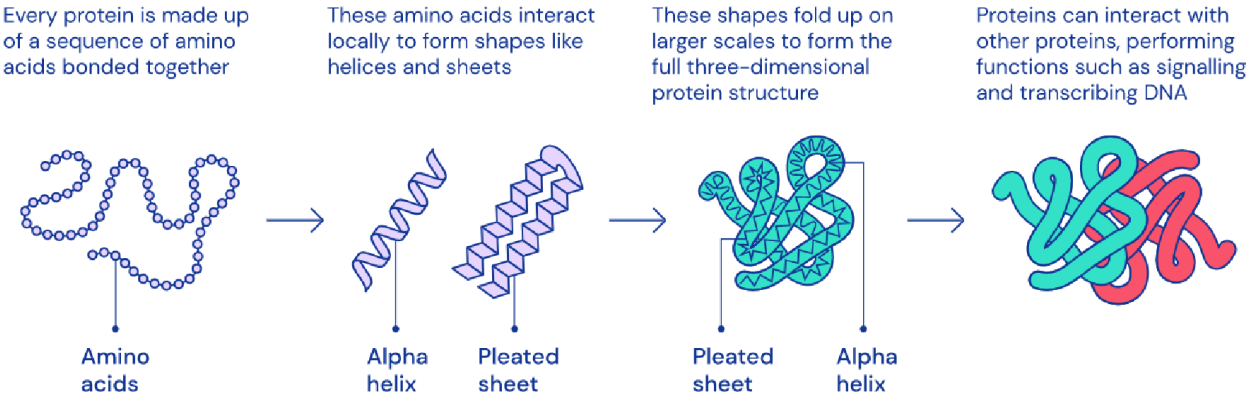


Figure 326: AlphaFold의 작동 원리

Figure 16.18 AlphaFold의 작동 원리. 어텐션 메커니즘을 활용하여 아미노산 간의 관계를 분석하고 3차원 공간에서의 근접성을 예측하는 AlphaFold의 핵심 기술적 원리를 단계별로 보여준다.

Figure 16.19 CASP에서의 AlphaFold 성과. 단백질 구조 예측 경진대회인 CASP14에서 AlphaFold가 달성한 혁신적인 성과를 보여주며, 기존 방법들을 압도적으로 앞지르는 정확도를 나타낸다.

Figure 16.20 CASP 다른 참가들의 성과. CASP14에서 AlphaFold 외에 참가한 다른 단백질 구조 예측 모델들의 성능을 비교하여, AlphaFold의 압도적인 우수성을 명확하게 보여준다.

AlphaFold의 핵심 기술은 어텐션 메커니즘을 활용한 트랜스포머 아키텍처이다. 이 모델은 아미노산 간의 관계를 분석하여 어떤 아미노산들이 3차원 공간에서 가까이 위치할지를 예측한다. 또한 진화 정보를 활용하여 유사한 서열을

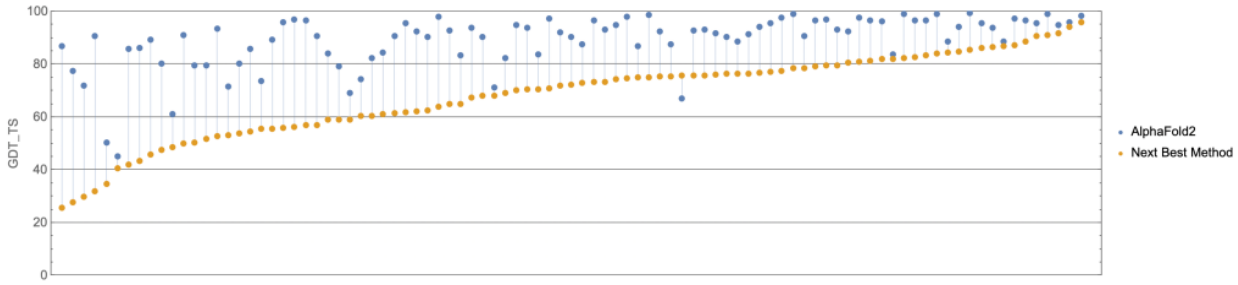


Figure 327: CASP에서의 AlphaFold 성과

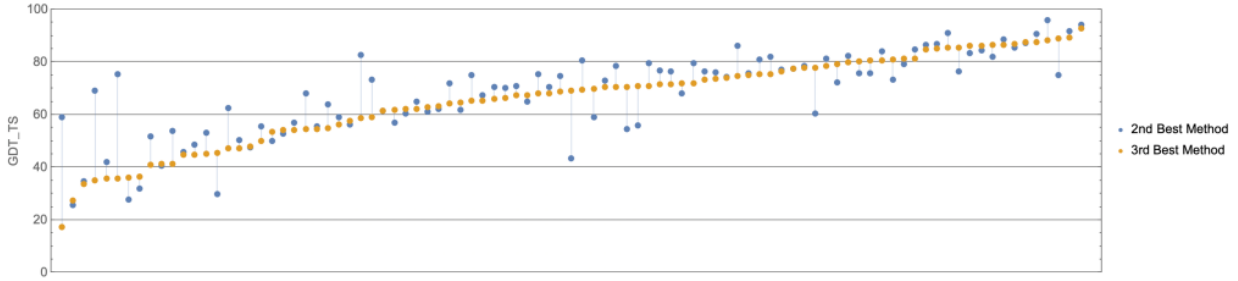


Figure 328: CASP 다른 참가들의 성과

가진 단백질들의 정보를 통합적으로 분석한다. AlphaFold는 현재 2억 개 이상의 단백질 구조를 예측하여 무료로 공개하고 있으며, 이는 전 세계 연구자들에게 혁신적인 자원을 제공하고 있다.

17.3.3 최신 단백질 구조 예측 모델들

AlphaFold 외에도 다양한 최신 단백질 구조 예측 모델들이 개발되고 있다. RoseTTAfold는 데이비드 베이커 연구팀에서 개발한 모델로, AlphaFold와 유사한 성능을 보이면서도 더 빠른 속도로 구조를 예측할 수 있다. 이 모델은 3개의 서로 다른 네트워크가 협력하여 서열, 거리, 구조 정보를 동시에 처리하는 독특한 아키텍처를 가지고 있다.

RFdiffusion은 확산 모델을 단백질 설계에 적용한 혁신적인 접근법이다. 기존의 구조 예측과 달리, RFdiffusion은 원하는 기능을 가진 전혀 새로운 단백질을 설계할 수 있다. 이 모델은 노이즈에서 시작하여 점진적으로 단백질 구조를 생성하는 확산 과정을 학습하여, 자연에 존재하지 않는 새로운 단백질을 창조할 수 있다.

ESMfold는 메타(Meta, 구 페이스북)에서 개발한 모델로, 대규모 언어 모델의 원리를 단백질 서열에 적용했다. ESM(Evolutionary Scale Modeling) 언어 모델을 기반으로 하여 단백질 서열의 패턴을 학습하고, 이를 통해 구조를 예측한다. 특히 ESMfold는 다른 모델들에 비해 계산 속도가 빠르면서도 높은 정확도를 유지한다는 장점이 있다.

17.3.4 단백질 구조 예측의 응용과 미래

이러한 AI 기반 단백질 구조 예측 기술들은 다양한 분야에서 활용되고 있다. 신약 개발에서는 질병 관련 단백질의 구조를 예측하여 약물 표적을 발견하고, 기존 약물의 작용 메커니즘을 이해하는 데 활용된다. 또한 새로운 치료용 단백질이나 백신을 설계하는 데도 중요한 역할을 하고 있다.

환경 문제 해결에도 기여하고 있다. 플라스틱을 분해하는 효소나 이산화탄소를 포집하는 단백질 등을 설계하여 지속가능한 기술 개발에 활용되고 있다. 농업 분야에서도 작물의 수확량을 높이거나 병해충에 저항성을 가진 새로운 단백질을 설계하는 연구가 진행되고 있다.

앞으로는 단일 단백질 구조 예측을 넘어서 단백질 복합체, 단백질-DNA 상호작용, 단백질의 동적 변화 등을 예측하는 기술로 발전할 것으로 예상된다. 또한 예측된 구조의 신뢰도를 정확히 평가하고, 실험 데이터와 통합하여 더욱 정확한 예측을 하는 방향으로 연구가 진행될 것이다.

17.4 AI 에이전트의 등장과 의생명정보학 데이터 분석

최근 의생명정보학 분야에서는 복잡한 데이터 분석 과정을 자동화하고 지능화하기 위한 AI 에이전트들이 등장하고 있다. 특히 단일세포 분석과 공간전사체학 분석을 위한 차원 축소와 데이터 분석 기법들의 복잡성을 극복하기 위해, AI 에이전트는 연구자가 복잡한 분석 파이프라인을 직접 구성하지 않아도 자동으로 적절한 분석 방법을 선택하고 실행할 수 있도록 도와주는 역할을 한다.

단일세포 RNA 시퀀싱과 공간전사체학 데이터 분석에서는 세포 유형 분류, 발달 궤적 추적, 세포 간 상호작용 분석, 공간적 패턴 인식 등 다양한 분석 단계가 필요하다. 각 단계마다 적절한 알고리즘과 매개변수를 선택해야 하며, 분석 결과를 해석하고 다음 단계로 연결하는 과정에서 전문적인 지식이 요구된다. AI 에이전트는 이러한 복잡한 의사결정 과정을 자동화하여, 연구자가 생물학적 질문에 더 집중할 수 있도록 지원한다.

17.4.1 AI 에이전트의 구조와 작동 원리

AI 에이전트는 단순한 AI 모델과 달리 자율적으로 작업을 수행하고 환경과 상호작용할 수 있는 지능형 시스템이다. 이러한 에이전트들은 크게 세 가지 핵심 구성 요소로 이루어져 있다.

첫째, 중앙 처리 장치 역할을 하는 대화형 AI 모델이 있다. 이는 일반적으로 GPT나 Claude와 같은 대규모 언어 모델을 기반으로 하며, 사용자의 요청을 이해하고 적절한 응답을 생성한다. 둘째, 다양한 외부 도구와 연결할 수 있는 인터페이스가 있다. 이를 통해 에이전트는 데이터베이스 검색, 계산 수행, 파일 처리, 웹 검색 등의 작업을 실행할 수 있다. 셋째, 과거의 대화 내용과 작업 결과를 저장하고 참조할 수 있는 메모리 시스템이 있다.

최근에는 MCP(Model Context Protocol)라는 표준화된 프로토콜이 개발되어, 서로 다른 AI 에이전트들이 일관된 방식으로 외부 도구와 상호작용할 수 있게 되었다. 이는 마치 컴퓨터에서 서로 다른 소프트웨어들이 공통된 운영체제를 통해 하드웨어에 접근하는 것과 유사하다. MCP를 통해 AI 에이전트는 데이터 분석 도구, 시각화 프로그램, 데이터베이스 등을 표준화된 방식으로 활용할 수 있다.

또한 A2A(Agent-to-Agent) 통신이라는 개념도 등장하고 있다. 이는 여러 개의 전문화된 AI 에이전트들이 서로 협력하여 복잡한 작업을 수행하는 방식이다. 예를 들어, 하나의 에이전트가 데이터 전처리를 담당하고, 다른 에이전트가 통계 분석을 수행하며, 또 다른 에이전트가 결과를 시각화하는 식으로 분업할 수 있다. 이러한 협력적 접근법은 각각의 에이전트가 특정 영역에 특화되어 더 높은 성능을 발휘할 수 있게 해준다.

17.4.2 의생명정보학 분야의 AI 에이전트 활용 사례

SpatialAgent는 이러한 AI 에이전트의 대표적인 예시로, 공간전사체학 데이터 분석을 위해 개발된 지능형 시스템이다. 이 시스템은 사용자가 제공한 데이터와 연구 질문을 바탕으로 적절한 분석 방법을 자동으로 선택하고, 결과를 해석하여 생물학적 통찰을 제공한다. 또한 다양한 데이터베이스와 연동하여 추가적인 정보를 자동으로 수집하고, 분석 결과를 시각화하여 연구자가 쉽게 이해할 수 있도록 돕는다.

Figure 16.21 SpatialAgent AI 에이전트. 공간전사체학 데이터 분석을 위한 지능형 AI 에이전트의 구조와 기능을 보여주며, 사용자의 요청을 자동으로 분석하고 적절한 분석 방법을 선택하는 과정을 나타낸다.

이러한 AI 에이전트들은 자동 요약 및 보고서 생성, 생명과학 지식의 체계적 관리, 단계별 추론 과정 수행, 결과의 자기 검증 등의 기능을 통해 연구자를 지원한다. 현재 Google Colab과 같은 클라우드 플랫폼을 통해 접근할 수 있으며, 연구의 민주화를 이끌어 고도의 계산 자원이나 전문 지식이 부족한 연구자들도 최신 기술을 활용할 수 있게 해주고 있다.

17.5 인공지능과 의생명정보학의 미래

17.5.1 기술 발전의 방향

인공지능 기술은 의생명정보학 분야에서 점점 더 정교하고 실용적인 형태로 발전하고 있다. 앞서 살펴본 딥러닝 모델들과 AI 에이전트 시스템은 단순히 데이터를 분석하는 도구를 넘어서, 생물학적 현상을 이해하고 예측할 수 있는 지능형 연구 동반자로 진화하고 있다. 이러한 발전은 개별 기술의 성능 향상뿐만 아니라, 서로 다른 기술들 간의 융합을 통해 이루어지고 있다.

특히 멀티모달 AI의 발전은 주목할 만하다. 텍스트, 이미지, 유전체 데이터, 공간 정보 등 다양한 형태의 생명과학 데이터를 동시에 처리하고 통합할 수 있는 AI 시스템들이 개발되고 있다. 이는 기존에 분리되어 있던 여러 분석 과정을 하나의 통합된 시스템으로 연결하여, 더욱 포괄적인 생물학적 이해를 가능하게 한다.

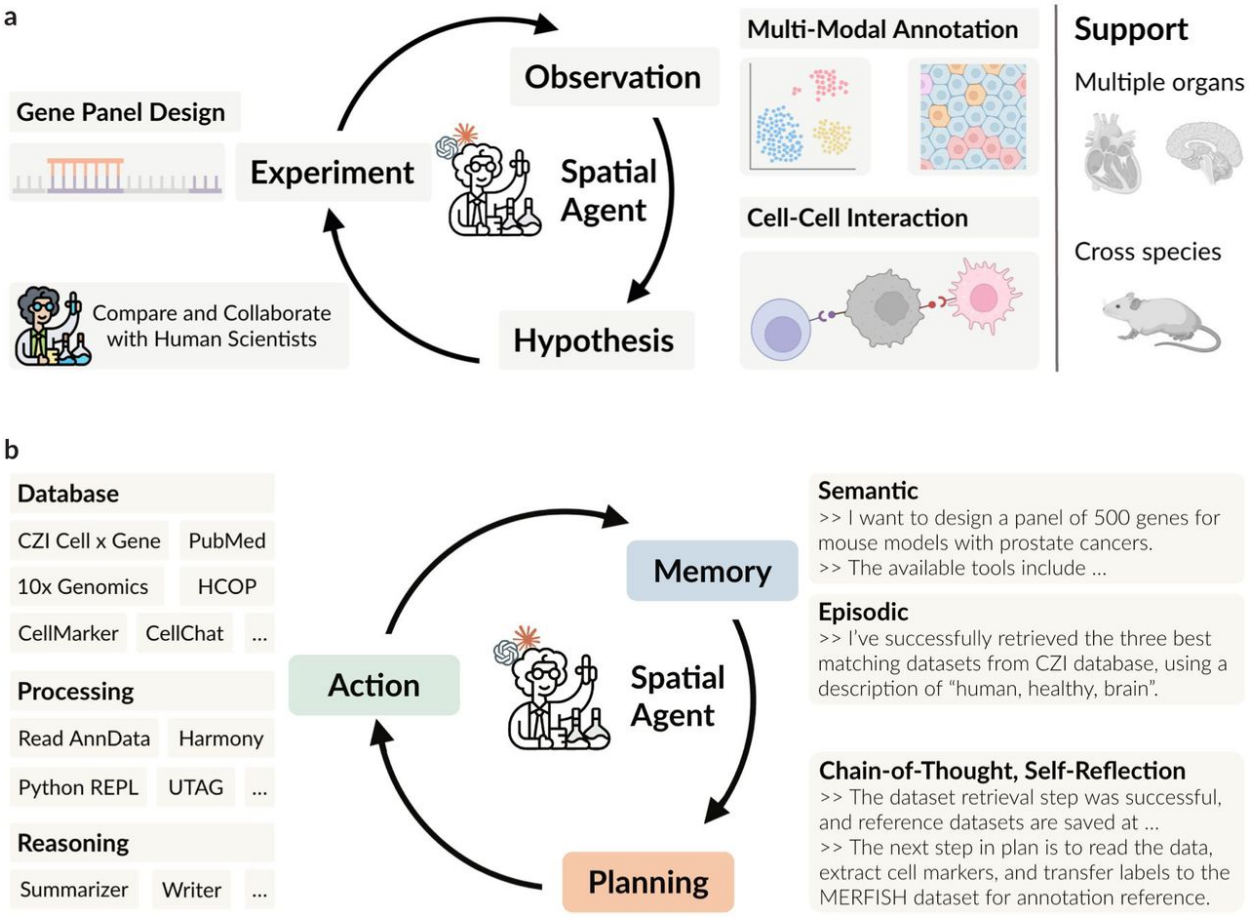


Figure 329: SpatialAgent AI 에이전트

또한 설명 가능한 AI(Explainable AI) 기술의 발전도 의생명정보학 분야에서 중요한 의미를 가진다. AI가 도출한 결과에 대한 근거와 추론 과정을 명확히 제시할 수 있게 되면서, 연구자들이 AI의 분석 결과를 더욱 신뢰하고 활용할 수 있게 되었다. 이는 특히 의료 분야에서 AI 기술의 임상 적용을 위한 필수 요건이기도 하다.

17.5.2 실용화와 접근성

AI 기술의 실용화 측면에서도 괄목할 만한 진전이 이루어지고 있다. 과거에는 고도의 전문 지식과 막대한 계산 자원이 필요했던 복잡한 분석들이 이제는 일반 연구자들도 쉽게 접근할 수 있는 도구로 변화하고 있다. 클라우드 기반의 플랫폼들과 사용자 친화적인 인터페이스의 개발로 인해, 프로그래밍 지식이 부족한 연구자들도 최신 AI 기술을 활용할 수 있게 되었다.

이러한 접근성의 향상은 연구의 민주화를 이끌고 있다. 대형 연구기관이나 자원이 풍부한 조직에서만 가능했던 고급 분석 기법들이 이제는 소규모 연구팀이나 개별 연구자들도 활용할 수 있게 되면서, 더 다양하고 창의적인 연구들이 진행될 수 있는 환경이 조성되고 있다.

17.5.3 도전 과제와 해결 방향

하지만 이러한 기술적 진보와 함께 해결해야 할 과제들도 존재한다. 가장 중요한 것은 데이터의 품질과 표준화 문제이다. AI 모델의 성능은 입력 데이터의 품질에 크게 의존하므로, 정확하고 일관성 있는 데이터의 확보가 필수적이다. 이를 위해서는 국제적 수준의 데이터 표준화 노력과 품질 관리 체계의 구축이 필요하다.

개인정보 보호와 데이터 보안 역시 중요한 고려사항이다. 의생명정보학 데이터는 개인의 민감한 유전 정보와 의료 기록을 포함하고 있어, 이를 안전하게 보호하면서도 연구 목적으로 활용할 수 있는 기술적, 법적 체계가 필요하다. 연합 학습이나 차등 프라이버시와 같은 기술들이 이러한 문제 해결에 기여할 수 있을 것으로 기대된다.

또한 AI 기술의 임상 적용을 위해서는 엄격한 검증 과정과 규제 체계가 필요하다. 연구 환경에서 개발된 AI 모델이 실제 의료 현장에서 안전하고 효과적으로 작동할 수 있도록 하는 검증 프로토콜의 정립이 중요하다.

17.6 결론

인공지능은 의생명정보학의 패러다임을 근본적으로 변화시키고 있다. 딥러닝 모델들은 유전체 분석부터 의료 영상 진단, 신약 개발에 이르기까지 다양한 분야에서 혁신적인 성과를 보여주고 있으며, 멀티모달 AI의 발전으로 서로 다른 오믹스 데이터를 통합 분석하여 생물학적 시스템의 전체적 이해가 가능해졌다. 특히 AI 에이전트 시스템의 등장은 연구 과정 자체를 자동화하고 지능화하여, 연구자들이 더 높은 수준의 생물학적 질문에 집중할 수 있게 해주고 있다. 이러한 기술들은 연구의 접근성을 대폭 향상시켜 의생명정보학 연구의 민주화를 이끌고 있으며, 앞으로 개인맞춤 의학과 정밀 의료의 실현에 핵심적인 역할을 할 것으로 기대된다.

18 컴퓨터 기반 신약 개발과 임상시험

18.1 서론

지금까지 학습한 생명정보학 지식들은 궁극적으로 인간의 질병을 이해하고 치료하는 데 활용된다. 유전체학을 통해 질병의 유전적 원인을 파악하고, 차세대 염기서열 분석(NGS)으로 개인의 유전적 변이를 확인하며, 유전체 전장 연관성 분석(GWAS)을 통해 질병과 연관된 유전자를 발견하고, 암 유전체학을 통해 종양의 분자적 특성을 규명하는 모든 과정은 결국 효과적인 치료법 개발로 이어진다. 이러한 생명정보학적 발견들을 실제 환자의 치료에 적용하기 위해서는 결국 새로운 의약품의 개발이 필수적이다.

신약 개발은 현대 의학과 생명과학에서 가장 중요하면서도 도전적인 분야 중 하나이다. 전통적인 신약 개발 과정은 10–17년의 긴 시간과 약 13억 달러의 막대한 비용이 소요되며, 성공률은 10% 미만에 그치는 매우 복잡하고 어려운 과정이다. 그러나 앞서 학습한 생명정보학 기술들의 발전과 함께 컴퓨터 기반 신약 설계(Computer-Aided Drug Design, CADD)와 인공지능 기술이 신약 개발 과정에 도입되어 혁신적인 변화를 가져오고 있다.

생명정보학에서 축적된 방대한 유전체 데이터, 단백질 구조 정보, 그리고 질병-유전자 연관성 데이터는 신약 표적 발견의 핵심 자원이 되고 있다. 예를 들어, GWAS를 통해 발견된 질병 연관 유전자는 새로운 약물 표적이 될 수 있으며, NGS를 통해 밝혀진 암의 돌연변이 패턴은 정밀 의료용 표적 치료제 개발의 기초가 된다. 컴퓨터 기반 신약 설계는 이러한 생명정보학적 지식과 계산 화학, 분자 모델링 기술을 융합하여 표적 발견(target discovery), 리드 화합물 발굴(lead

identification), 리드 최적화(lead optimization), 그리고 ADMET(Absorption, Distribution, Metabolism, Excretion, Toxicity) 예측 등의 단계에서 중요한 역할을 수행한다.

또한 컴퓨터 기반으로 개발된 신약 후보물질이 실제 환자에게 사용되기까지는 엄격한 임상시험 과정을 거쳐야 한다. 임상시험은 새로운 치료법의 안전성과 효능을 체계적으로 평가하는 과정으로, 전임상 연구부터 4상 임상시험까지 단계적으로 진행된다. 이 과정에서는 무작위 대조 임상시험, 맹검법, 통계적 분석 등 엄격한 과학적 방법론을 적용하여 치료법의 인과관계를 확립하고 근거 기반 의학의 토대를 마련한다. 특히 상관관계와 인과관계를 구분하는 것은 임상시험의 핵심이며, 이를 통해 환자에게 안전하고 효과적인 치료법을 제공할 수 있게 된다.

18.2 신약 개발 과정의 개요

18.2.1 전통적인 신약 개발 파이프라인

신약 개발 과정은 크게 전임상 단계와 임상시험 단계로 구분되며, 각 단계마다 엄격한 평가 기준을 통과해야만 다음 단계로 진행할 수 있는 체계적이고 순차적인 과정이다. 전체 과정은 다음과 같은 단계들로 구성된다:

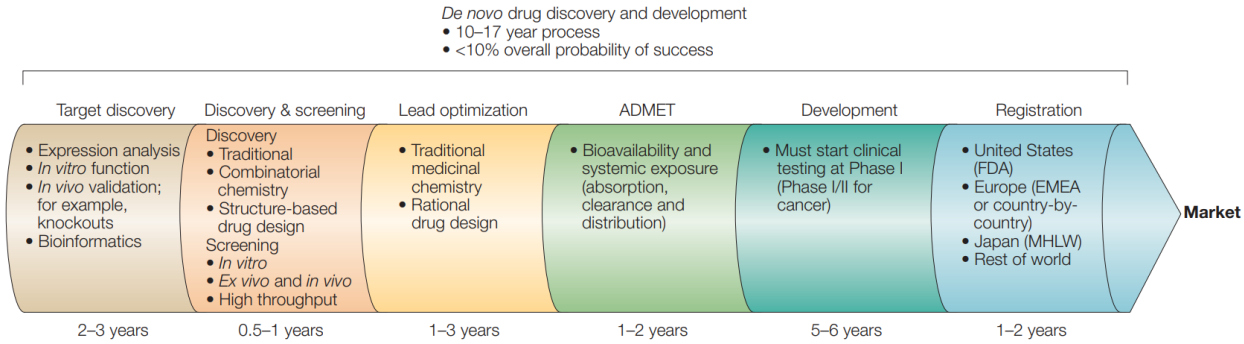


Figure 330: 신약 개발 과정

Figure 17.1 신약 개발 과정 – 표적 발견부터 시장 출시까지의 파이프라인

표적 발견 단계 (2–3년): 이 단계는 질병과 관련된 분자적 표적을 식별하고 그 타당성을 검증하는 것이 핵심이며, 다양한 오믹스 기술이 활용된다: – 발현 분석(expression analysis) – 시험관 내 기능 연구(in vitro function) – 생체 내 검증(in vivo validation) – 유전자 제거 실험(knockouts) – 생명정보학적 분석

발견 및 스크리닝 단계 (0.5–1년): 확인된 표적에 대해 활성을 나타내는 화합물을 찾아내는 것이 목표이며, 수만에서 수백만 개의 화합물을 체계적으로 검사하는 과정이 포함된다: – 전통적 방법(traditional methods) – 조합화학(combinatorial chemistry) – 구조 기반 신약 설계(structure–based drug design) – 시험관 내 스크리닝(in vitro screening) – 생체 내 및 동물 실험(ex vivo and in vivo studies) – 고처리량 스크리닝(high throughput screening)

리드 최적화 단계 (1–3년): 초기에 발견된 활성 화합물의 효능을 향상시키고 부작용을 최소화하는 과정으로, 의약화학자들이 화합물의 구조를 체계적으로 변형하여 더 나은 특성을 가진 후보물질을 개발한다: – 전통적 의약화학(traditional medicinal chemistry) – 합리적 신약 설계(rational drug design)

ADMET 단계 (1–2년): 개발된 화합물이 실제 생체 내에서 어떻게 작용하는지를 평가하며, 독성 여부를 확인하는 단계로 후보물질이 인체에 투여되기 전에 안전성을 사전에 평가하는 중요한 과정이다: – 생체 이용률 및 전신 노출(bioavailability and systemic exposure) – 흡수, 청소율 및 분포(absorption, clearance and distribution)

개발 단계 (5–6년): 전임상 연구와 임상시험을 포함하는 가장 긴 단계로, 동물 실험을 통한 안전성과 효능 확인 후 인간을 대상으로 한 임상시험이 1상부터 3상까지 순차적으로 진행된다: – 1상 임상시험에서 시작하는 임상시험 (암 대상)

등록 단계 (1–2년): 각국의 규제 기관에 신약 승인을 신청하고 심사를 받는 과정을 거치며, 승인을 받은 후에야 시장에서 환자들에게 처방될 수 있다: – 미국(FDA), 유럽(EMEA 또는 국가별), 일본(MHLW), 기타 지역에서의 승인

18.3 컴퓨터 기반 신약 설계 (CADD)

18.3.1 표적 발견 기술

표적 발견은 신약 개발의 첫 번째 단계로서, 질병의 원인이 되는 분자적 표적을 식별하고 그 타당성을 검증하는 과정이다. 현대 신약 개발에서는 전통적인 생화학적 접근법에서 벗어나 다양한 오믹스 기술을 활용한 체계적이고 포괄적인 접근이 이루어지고 있다.

오믹스 기반 신약 발견: 현대 신약 개발에서는 오믹스(omics) 기술들이 통합적으로 활용되어 포괄적이고 체계적인 신약 발견을 가능하게 한다. 이러한 다중 오믹스 접근법은 유전체학(genomics), 전사체학(transcriptomics), 단백체학(proteomics), 대사체학(metabolomics) 등을 통합하여 질병의 분자적 메커니즘을 종합적으로 이해하고 새로운 치료 표적을 발견하는 데 활용된다.

OMCS for Drug Discovery

Figure 17.2 신약 발견을 위한 오믹스 기술

18.3.2 스크리닝 방법

스크리닝은 확인된 표적에 대해 활성을 나타내는 화합물을 발굴하는 과정으로, 전통적인 실험적 방법과 컴퓨터 기반 방법이 상호 보완적으로 활용되고 있다.

고처리량 스크리닝 (High-Throughput Screening, HTS): 고처리량 스크리닝은 로봇 자동화 시스템을 이용하여 수십만에서 수백만 개의 화합물을 빠르게 테스트하는 기술로, 현대 신약 개발에서 필수적인 도구가 되었다. 이 방법의 주요 목표는: - “히트(hits)” 화합물의 발견 - “오프타겟(off-targets)” 효과의 최소화

HTS는 현재 완전 자동화된 로봇 시스템을 통해 수행되며, 이를 통해 동시에 많은 수의 화합물을 테스트할 수 있다. 이러한 자동화 시스템은 인간의 개입을 최소화하여 실험의 일관성을 보장하고, 대용량 스크리닝을 가능하게 한다.

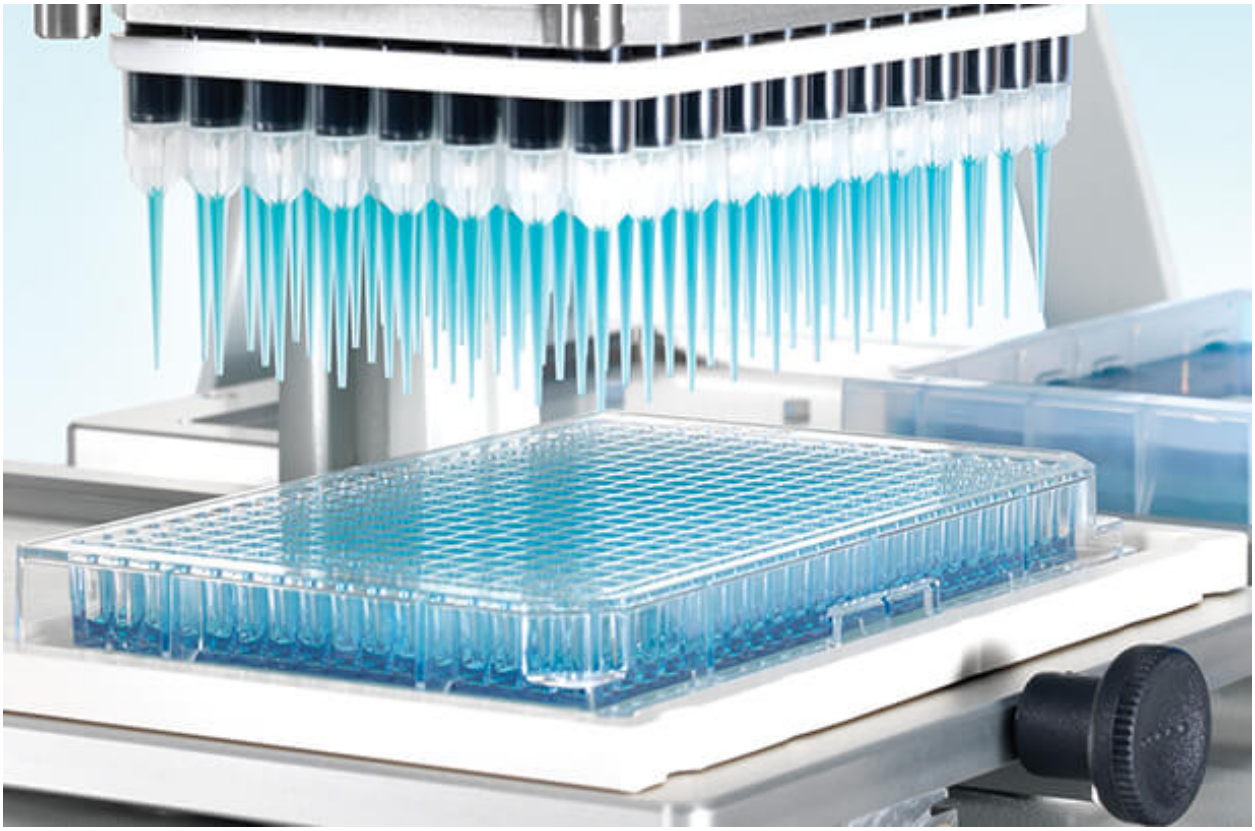


Figure 331: 고처리량 스크리닝 시스템

Figure 17.3 고처리량 스크리닝 시스템

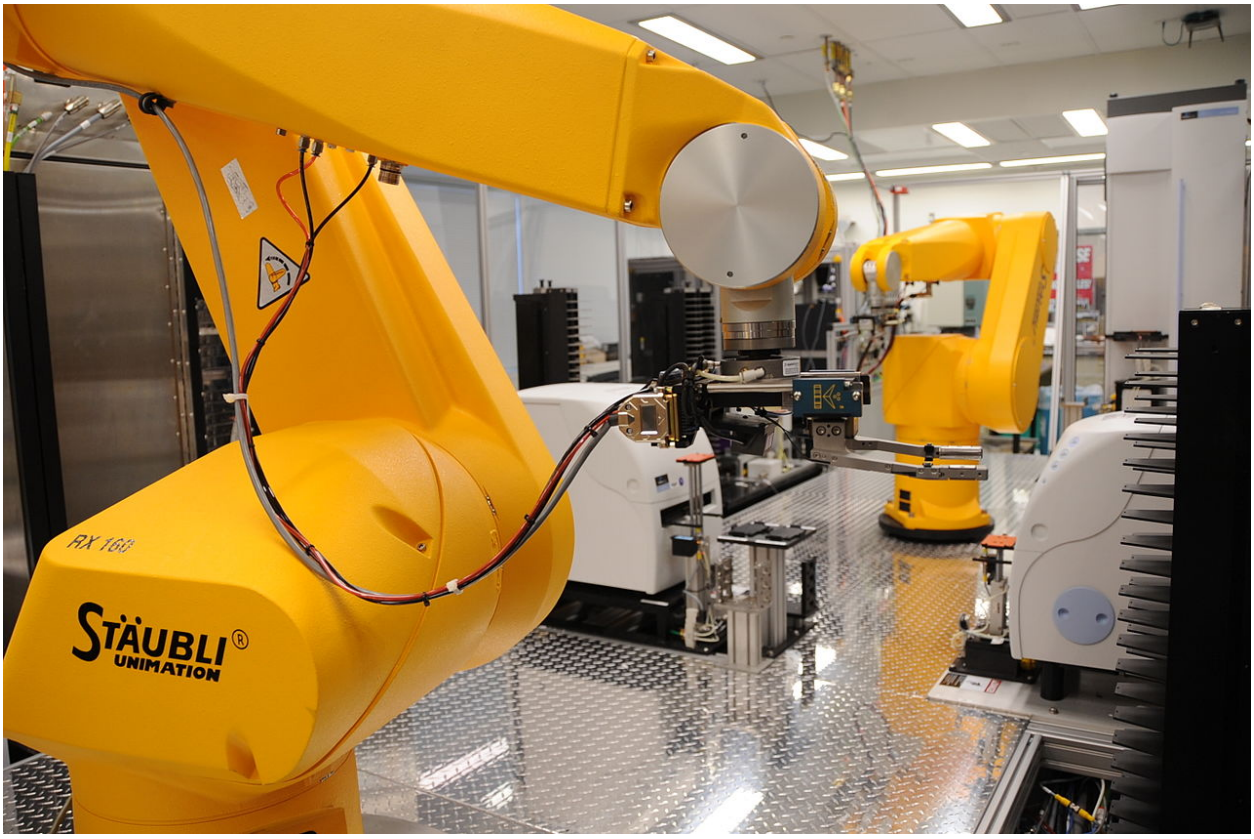


Figure 332: HTS 로봇 자동화 시스템

Figure 17.4 HTS 로봇 자동화 시스템

가상 스크리닝 (Virtual Screening, VS): 가상 스크리닝은 컴퓨터 기반 계산 방법을 통해 히트 화합물을 찾는 기술로, 실험적 스크리닝에 앞서 후보 화합물을 사전에 선별하여 효율성을 크게 향상시킨다. 이 방법은 다음을 기반으로 한다: - 표적 구조 (예: PDB 데이터베이스) - 리간드 구조 (예: ChEMBL 데이터베이스)

가상 스크리닝의 핵심은 방대한 화학 데이터베이스를 활용하는 것이다. ChEMBL은 약물 유사 특성을 가진 생물활성 분자들의 수동 큐레이션된 데이터베이스로, 화학적 생물활성과 유전체 데이터를 결합하여 유전체 정보를 효과적인 새로운 약물로 변환하는 데 도움을 준다. ChEMBL 데이터베이스는 수백만 개의 화합물과 활성도 정보를 포함하고 있어, 이를 기반으로 가상으로 새로운 약물 후보를 체계적으로 탐색할 수 있다.

Figure 17.5 ChEMBL 약물 데이터베이스

18.3.3 분자 도킹 (Molecular Docking)

분자 도킹은 컴퓨터 기반 신약 설계의 핵심 기술 중 하나로, 표적 단백질과 리간드 분자 간의 최적 결합 방식을 예측하는 계산 기법이다. 이 기술은 물리화학적 원리에 기반하여 분자 간 상호작용을 모델링하고, 가장 안정한 결합 구조를 예측한다.

Figure 17.6 분자 도킹

18.3.4 리드 최적화

리드 최적화는 신약 개발 과정에서 가장 중요하고 도전적인 단계 중 하나로, 초기 스크리닝에서 발견된 히트 화합물을 실제 치료제로 발전시키기 위한 체계적인 분자 설계 과정이다.

Figure 17.7 리드 최적화

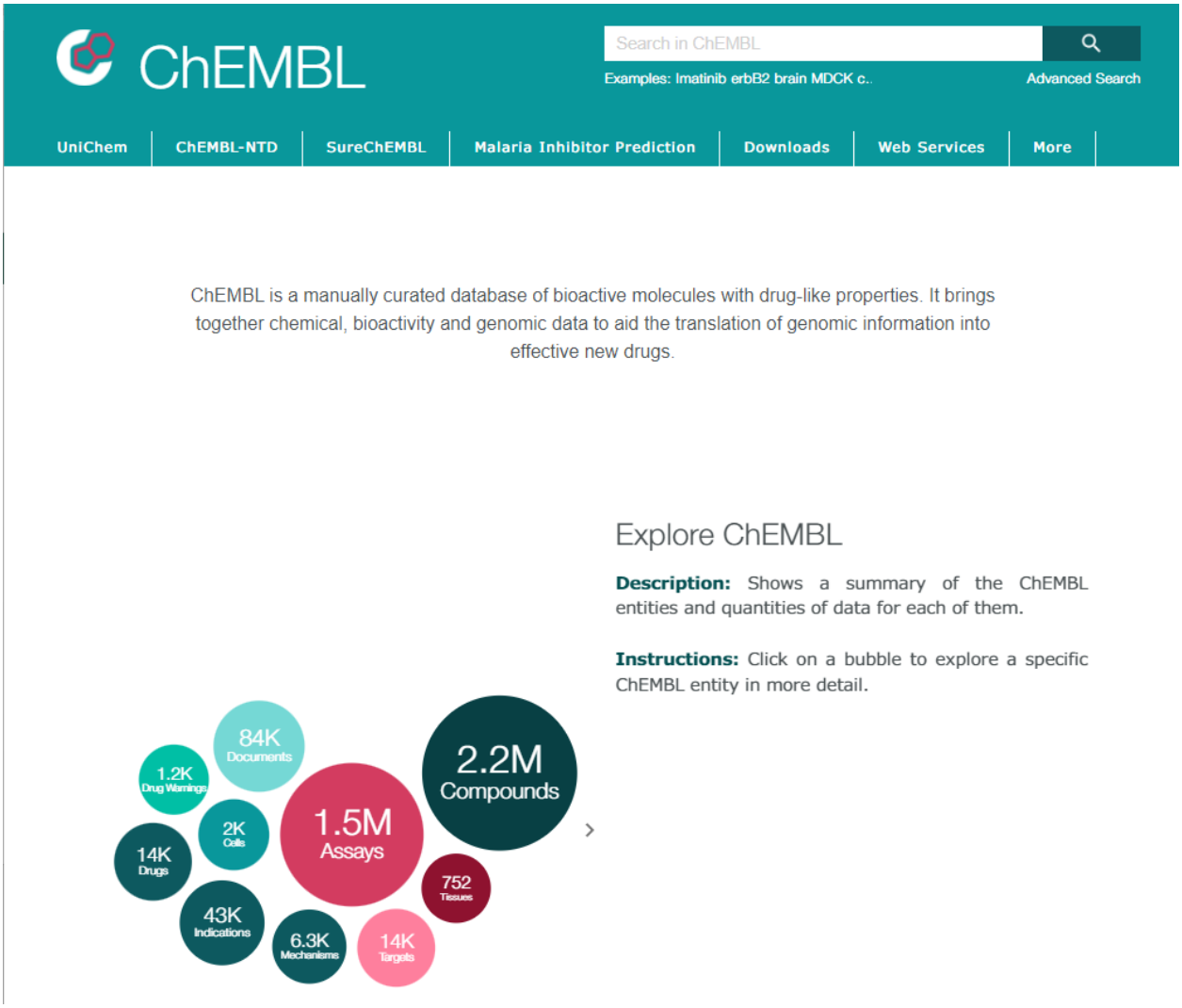


Figure 333: ChEMBL 약물 데이터베이스

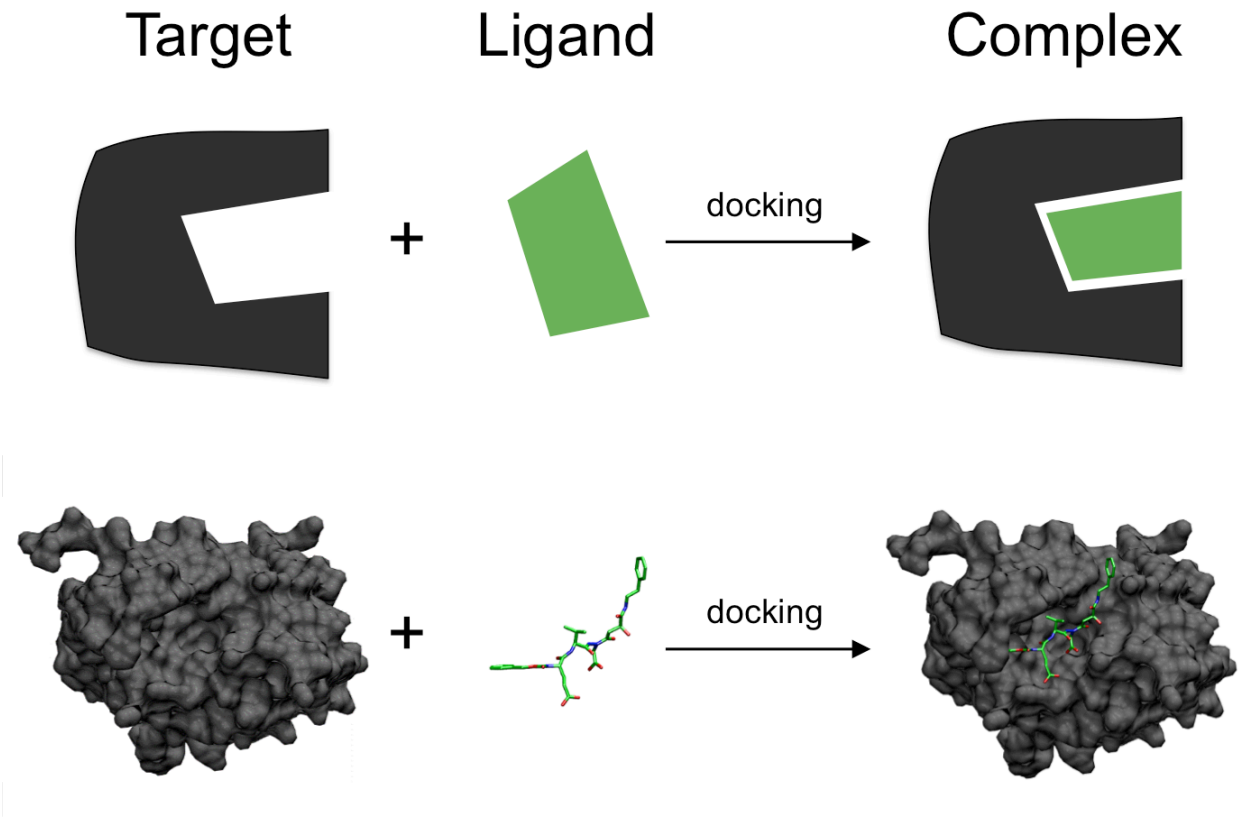


Figure 334: 분자 도킹 과정

Figure 335: 리드 최적화 과정

히트에서 리드로의 전환: 초기 스크리닝에서 발견된 히트 화합물을 약물로 개발하기 위해서는 다방면의 특성을 동시에 개선해야 하는 체계적인 최적화 과정이 필요하다. 이 과정에서는 구조적 단순화와 리드 최적화를 통해 복잡한 구조를 더 간단하고 합성하기 쉬운 형태로 변환하면서도, 생물학적 활성은 유지하거나 향상시켜야 한다. 의약화학자들은 구조–활성 관계(Structure–Activity Relationship, SAR) 연구를 통해 분자의 어떤 부분이 활성에 필수적인지, 어떤 부분을 변형할 수 있는지를 체계적으로 파악한다.

리드 최적화 과정에서는 효능(potency), 선택성(selectivity), 약동학적 특성(pharmacokinetics), 독성(toxicity), 합성 용이성(synthetic accessibility) 등 다양한 요소를 고려하여 화합물의 구조를 개선하는 과정이다.

이러한 최적화 과정은 수백 개에서 수천 개의 유사체 화합물을 합성하고 평가하는 반복적인 과정을 통해 이루어진다.

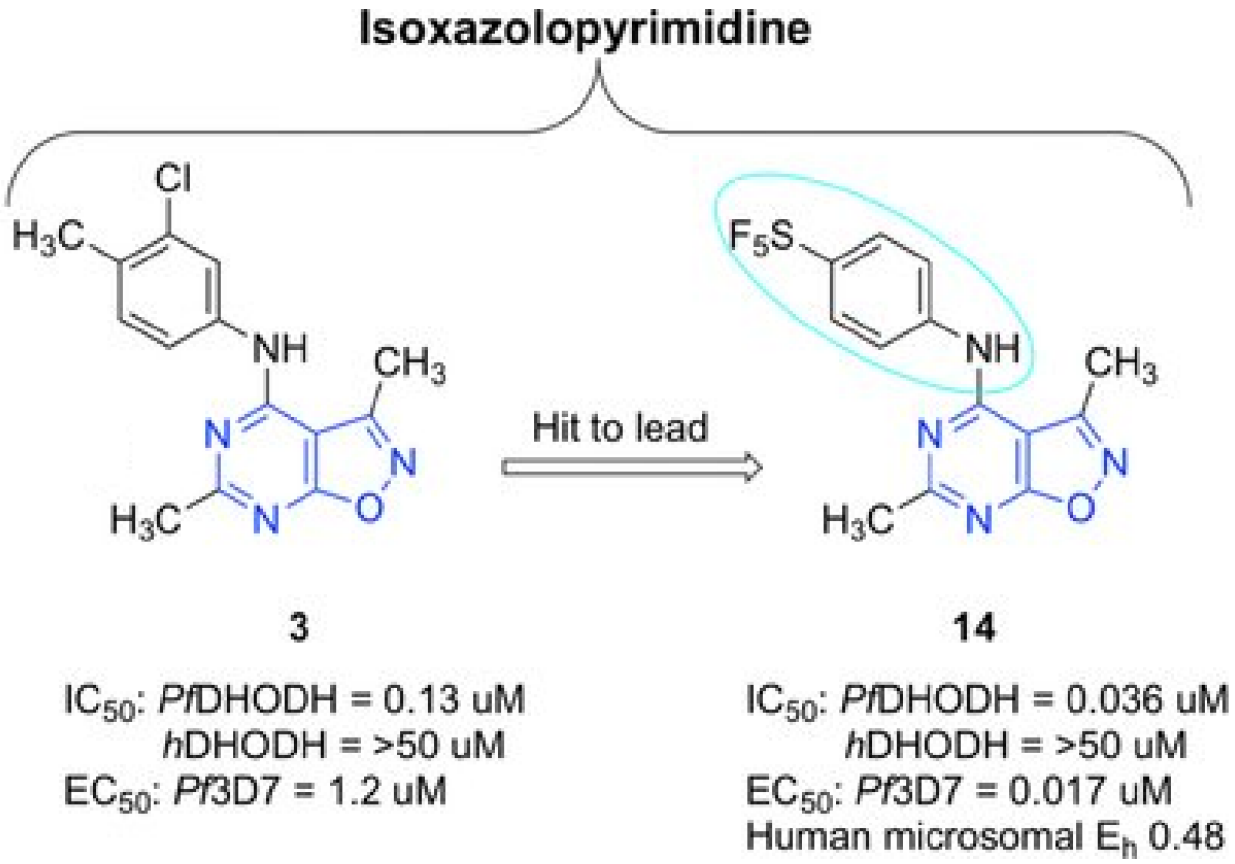


Figure 336: 히트에서 리드로의 전환

Figure 17.8 히트에서 리드로의 전환 – 복잡한 구조를 간단하고 합성하기 쉬운 형태로 변환

18.3.5 ADMET 예측

ADMET는 약물의 흡수(Absorption), 분포(Distribution), 대사(Metabolism), 배설(Excretion), 독성(Toxicity)을 의미하며, 이러한 특성들은 약물의 효능과 안전성을 결정하는 핵심 요소들이다. 전통적으로 ADMET 특성은 개발 후반부에 평가되었지만, 현재는 신약 개발 초기 단계에서부터 예측하고 최적화하는 것이 일반적이다.

ADME 특성: 각 ADME 특성은 약물이 체내에서 겪는 일련의 과정을 나타낸다: – **흡수**: 위장관에서의 약물 흡수 과정으로, 생체이용률과 직접적으로 연관된다. 흡수율은 약물의 용해도, 투과성, 안정성에 의해 결정된다. – **분포**: 혈류를 통한 조직 분포 및 혈장 단백질 결합으로, 약물이 표적 조직에 도달하는 정도를 결정한다. 분포 용적과 단백질 결합률이 주요 지표이다. – **대사**: 간에서의 효소에 의한 대사 과정으로, 약물의 반감기와 활성 지속 시간에 영향을 미친다. 주로 간의 사이토크롬 P450 효소군이 관여한다. – **배설**: 신장을 통한 배설 및 장벽 통과로, 약물의 체내 축적을 방지하고 청소율을 결정한다.

Figure 17.9 ADME 특성

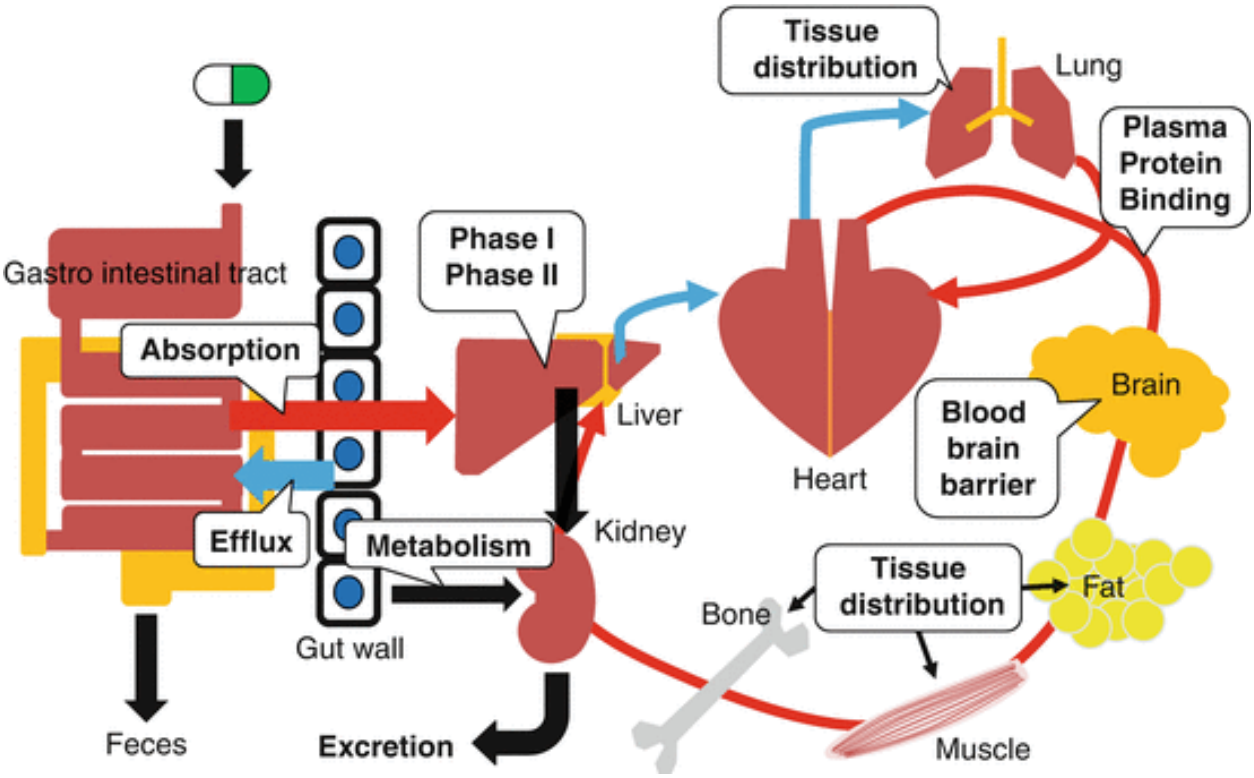


Figure 337: ADME 특성 평가

독성 예측: ADMET 과정에서 독성 예측은 특히 중요한 부분으로, 약물 개발 실패의 주요 원인 중 하나가 예상치 못한 독성 때문이다. 최근 연구에 따르면, 기계학습 소프트웨어가 화학물질의 독성 예측에서 동물 실험보다 우수한 성능을 보이는 것으로 나타났다. 이는 동물 실험의 한계를 극복하고 더 효율적인 안전성 평가를 가능하게 한다. 현대의 독성 예측 모델은 다양한 독성 종류(간독성, 심독성, 신독성 등)를 개별적으로 예측할 수 있으며, 구조-독성 관계(Structure-Toxicity Relationship)를 기반으로 한다.

Figure 17.10 동물실험을 능가하는 성능을 보인 기계학습

18.4 현대적 치료제 및 신약 개발 기술

18.4.1 항체-약물 접합체(ADC)

항체-약물 접합체(Antibody-Drug Conjugate, ADC)는 표적 단백질에 특이적으로 결합하는 단클론 항체와 강력한 세포독성 약물을 연결한 혁신적인 표적 치료제이다. ADC는 항체의 특이성과 세포독성 약물의 효능을 결합하여, 정상 세포에는 영향을 최소화하면서 암세포만을 선택적으로 제거할 수 있다.

Figure 17.11 항체-약물 접합체(ADC) 기술 – 단클론 항체와 세포독성 약물을 연결한 선택적 암 치료제

ADC 기술의 발전 역사는 지난 20여 년간 지속적인 혁신을 통해 이루어져 왔다. 초기 ADC들은 안정성과 효능 면에서 한계가 있었지만, 링커 기술의 발전과 새로운 세포독성 약물의 개발로 많은 개선이 이루어졌다.

Figure 17.12 ADC 발전 역사 – 지난 20여 년간 초기 ADC에서 현대적 고효능 ADC로의 발전

18.4.2 CAR-T 세포 치료

CAR-T(Chimeric Antigen Receptor T-cell) 세포 치료는 환자의 T 세포를 추출하여 유전적으로 개조한 후 다시 환자에게 주입하는 개인맞춤형 면역 치료법이다. 이 기술은 특정 암 세포를 인식하고 공격할 수 있도록 설계된 인공

NEWS | 11 July 2018

Software beats animal tests at predicting toxicity of chemicals

Machine learning on mountain of safety data improves automated assessments.

Richard Van Noorden

Computer programs can, in some cases, predict chemical toxicity as well as tests done on rats and other animals. Credit: Coneyl Jay/SPL

Machine-learning software trained on masses of chemical-safety data is so good at predicting some kinds of toxicity that it now rivals – and sometimes outperforms – expensive animal studies, researchers report.

Computer models could replace some standard safety studies conducted on millions of animals each year, such as dropping compounds into rabbits' eyes to check if they are irritants, or feeding chemicals to rats to work out lethal doses, says Thomas Hartung, a

Figure 338: 기계학습이 동물실험을 능가하는 성능

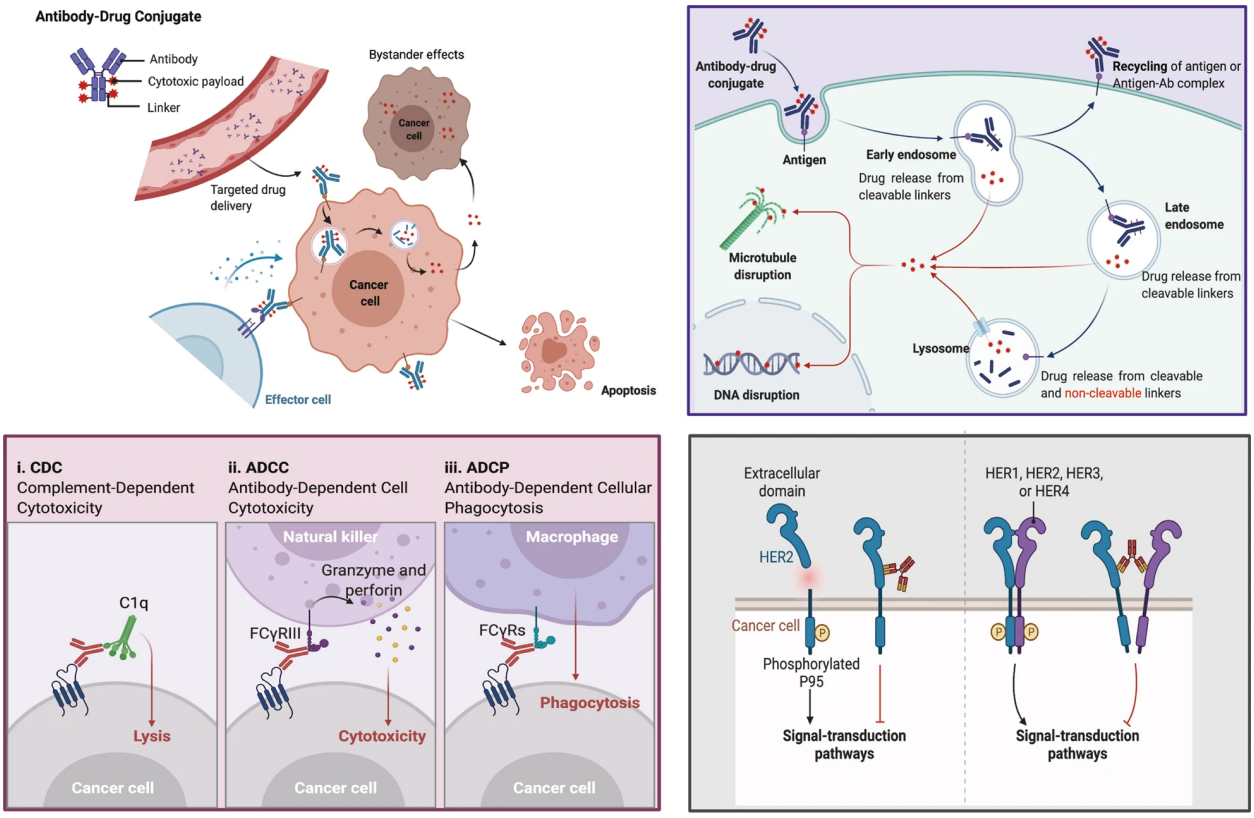


Figure 339: ADC 기술

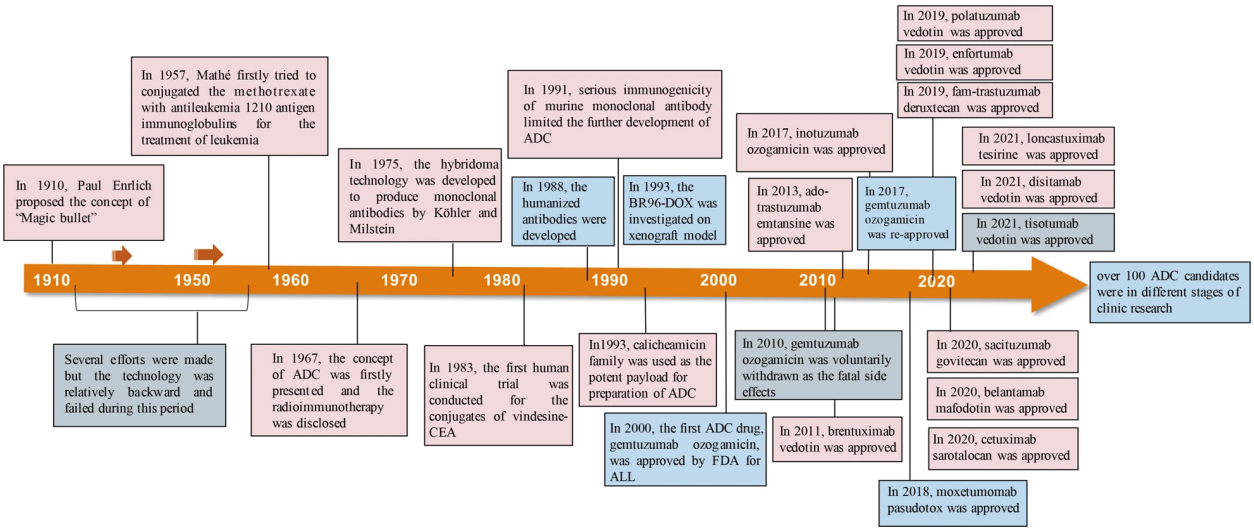


Figure 340: ADC 발전 역사

수용체를 T 세포에 도입하여, 자체 면역 시스템이 암을 보다 효과적으로 인식하고 제거할 수 있도록 한다. CAR-T 세포 치료의 효능과 안전성을 향상시키기 위해서 유전자 편집 기술이 적극적으로 활용되고 있다.

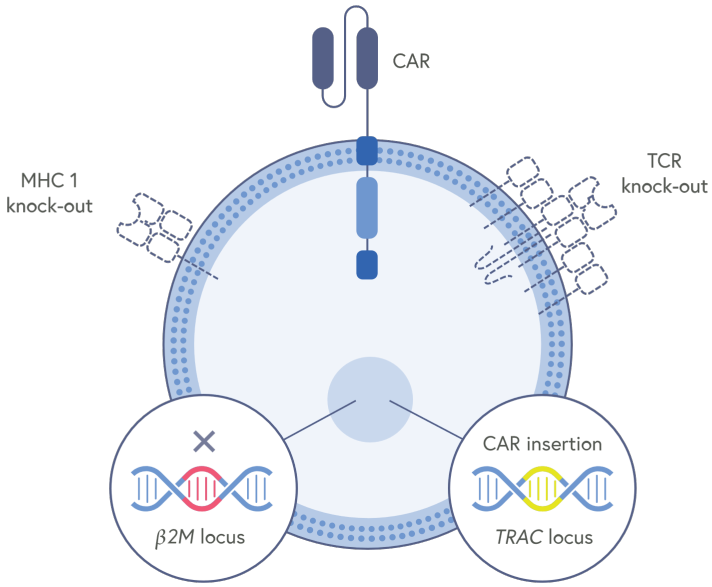


Figure 341: CAR-T 세포 구조

Figure 17.13 CAR-T 세포 구조

18.4.3 PROTAC 기술

PROTAC(PROteolysis TArgeting Chimera)는 대상 단백질을 단백질 분해 시스템으로 유도하여 선택적으로 분해시키는 혁신적인 약물 개발 기술이다. 이 기술은 전통적인 단백질 억제와는 다른 접근 방식으로, 대상 단백질을 아예 제거하여 보다 지속적이고 강력한 치료 효과를 달성할 수 있다.

Figure 17.14 PROTAC 발전 역사

18.4.4 면역 체크포인트 억제제

면역 체크포인트 억제제는 암 세포가 면역 시스템의 공격을 피하기 위해 사용하는 체크포인트 단백질을 차단하여, 면역 세포가 암 세포를 다시 인식하고 공격할 수 있도록 하는 혁신적인 치료법이다. PD-1/PD-L1 경로는 이러한 체크포인트 메커니즘의 대표적인 예시이다.

Figure 17.15 PD-L1 체크포인트 억제제

18.5 인공지능 기반 신약 발견

인공지능과 기계학습 기술의 급속한 발전은 신약 발견 분야에 혁명적인 변화를 가져오고 있다. 전통적인 신약 개발이 직면한 높은 비용, 긴 개발 기간, 낮은 성공률 등의 문제를 해결하기 위해 AI 기술이 신약 개발의 모든 단계에 도입되고 있으며, 이는 패러다임의 근본적인 전환을 의미한다.

18.5.1 AI 기반 신약 발견 워크플로우

현대 신약 발견에서 인공지능은 전통적인 순차적 접근법을 대체하여 통합되고 순환적인 워크플로우를 제공한다. AI 시스템은 대용량 데이터를 동시에 처리하고 다중 목표를 최적화할 수 있어, 기존 방법론보다 훨씬 효율적인 신약 개발이 가능하다.

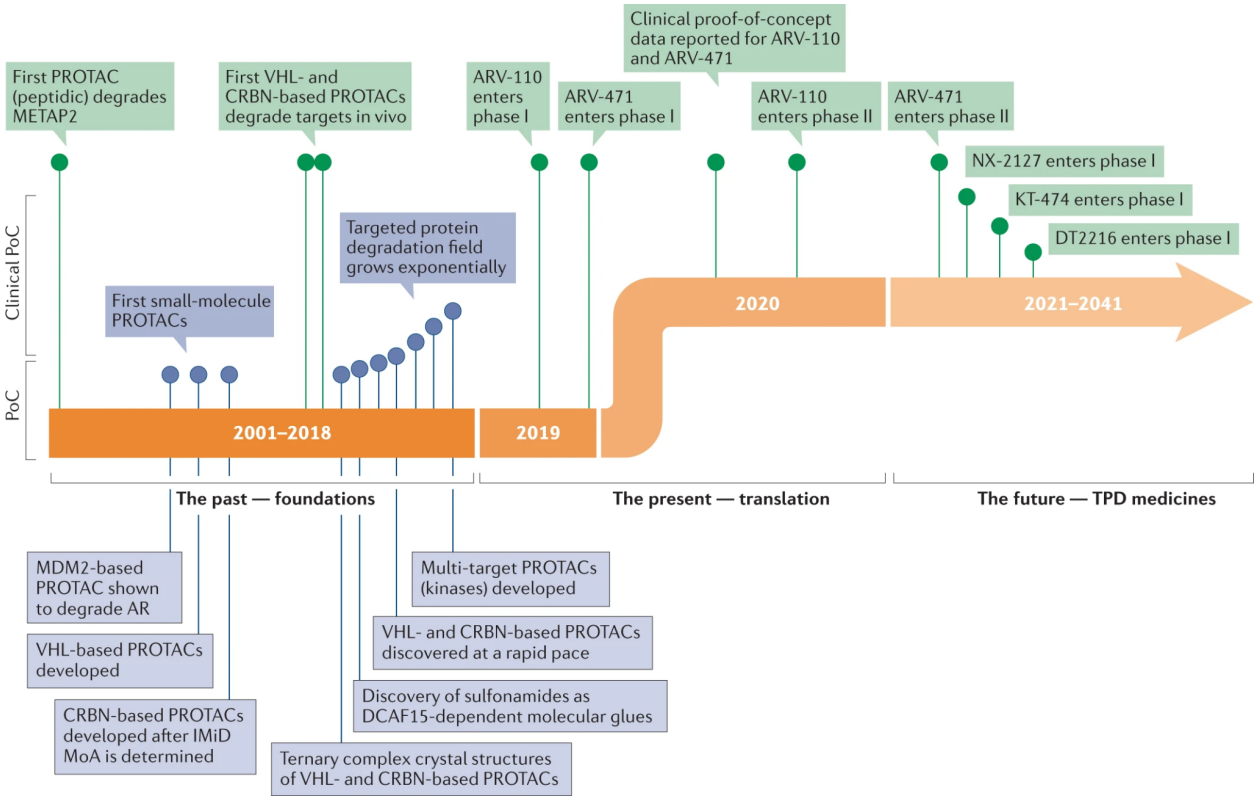


Figure 342: PROTAC 발전 역사

실제 사례를 보면, 인공지능 기반 접근법은 150만 개 이상의 화합물 라이브러리에서 시작하여 가상 스크리닝과 고처리량 스크리닝을 거쳐 각각 1,478개, 58,371개의 화합물 후보군으로부터 1,901개의 화합물을 선별하고, 최적화를 거쳐 최종적으로 10개의 화합물을 도출하는 효율적인 과정을 보여준다. 이는 전통적인 방법 대비 수백 배의 효율성 향상을 의미한다.

Figure 17.16 AI 기반 신약 발견 워크플로우 예시

18.5.2 최신 AI 기법의 활용

신약 발견 분야에서는 다양한 최신 AI 기법들이 혁신적인 결과를 창출하고 있으며, 각각의 기술은 신약 개발의 특정 병목 지점을 해결하는 데 특화되어 있다.

기계학습 기반 접근법: 분자 도킹과 결합된 기계학습 접근법에서는 랜덤 포레스트(Random Forests) 알고리즘을 활용한 사례가 있다: - 알려진 리간드 데이터를 이용한 모델 훈련 - 여러 리간드 형태(conformations)에 대한 분자 도킹 수행 - 미지의 화합물 라이브러리에 대한 가상 스크리닝 실시

Figure 17.17 랜덤 포레스트 알고리즘 예시 - 알려진 리간드 데이터로 모델 훈련 후 미지 화합물 가상 스크리닝

딥러닝과 신경망 기법: 최근에는 딥러닝 기법이 신약 발견의 핵심 도구로 부상하고 있다. 특히 다음과 같은 영역에서 혁신적인 성과를 보이고 있다: - **분자 생성 모델**: VAE(Variational Autoencoder), GAN(Generative Adversarial Networks), Transformer 기반 모델들이 새로운 약물 분자를 생성한다. - **분자 특성 예측**: 그래프 신경망(Graph Neural Networks)이 분자의 3차원 구조 정보를 활용하여 생물학적 활성을 예측한다. - **약물-표적 상호작용 예측**: 멀티모달 딥러닝 모델이 단백질과 화합물의 상호작용을 정확하게 예측한다.

단백질 구조 예측과 설계: AlphaFold의 성공 이후 단백질 구조 예측 분야에서 AI의 활용이 폭발적으로 증가했다. RFdiffusion과 같은 최신 AI 모델들은 단백질 구조 생성과 설계에 혁신적인 발전을 가져왔으며, 다음과 같은 응용이 가능하다: - 새로운 표적 단백질의 구조 예측 - 기존 구조의 개선 및 최적화 - 완전히 새로운 기능을 가진 단백질 설계 - 단백질-단백질 상호작용 설계

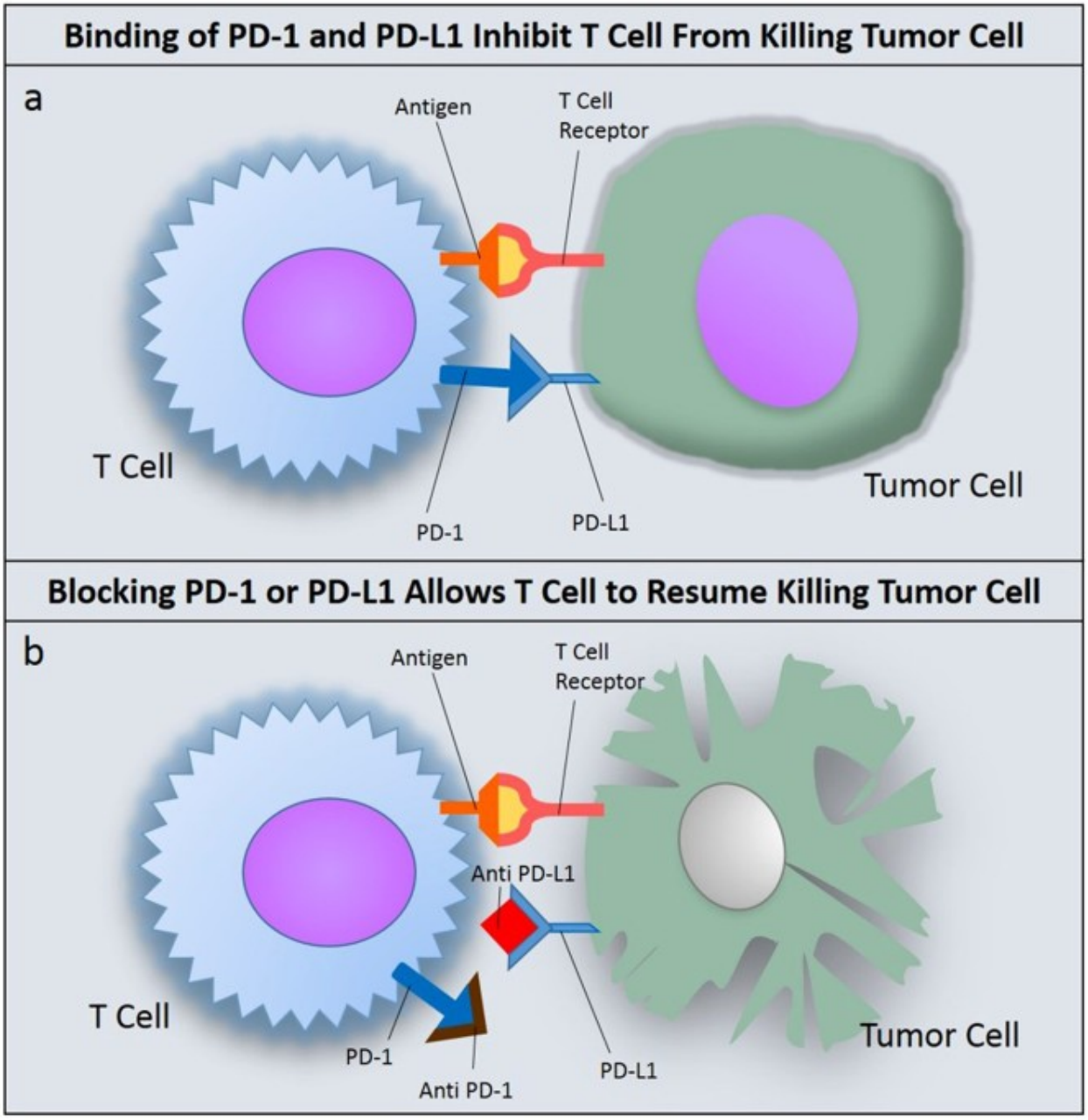


Figure 343: PD-L1 체크포인트

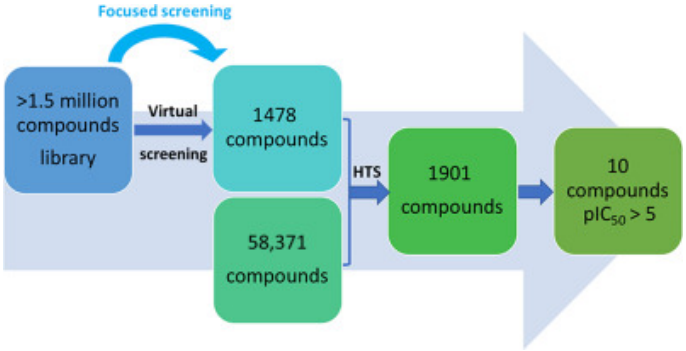


Figure 344: AI 기반 신약 발견 워크플로우

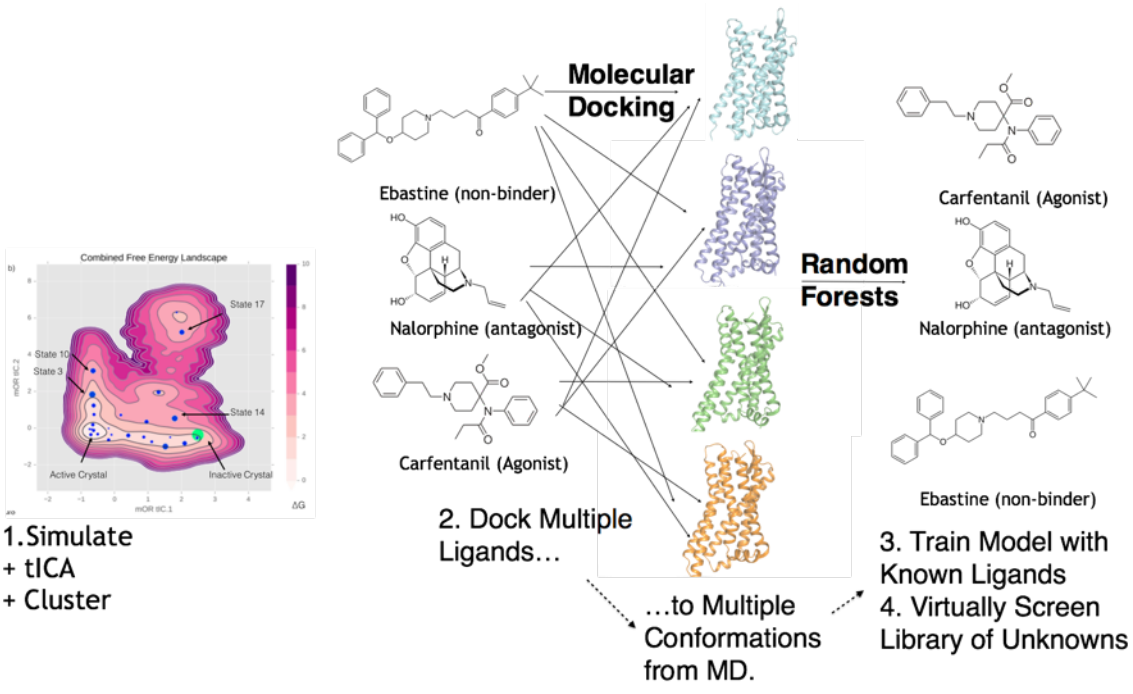


Figure 345: 랜덤 포레스트 알고리즘 예시

이러한 기술들은 기존에 약물 디자인이 불가능하다고 여겨졌던 (undruggable) 표적들에 대한 새로운 접근 방법을 제시하고 있다.

18.5.3 AI 기반 신약 발견의 성공 사례

최근 몇 년간 AI 기반 신약 발견 분야에서 주목할 만한 성공 사례들이 등장하고 있어, 이 기술의 실용성과 잠재력을 입증하고 있다.

COVID-19 치료제 개발: 팬데믹 상황에서 AI 기술은 신속한 치료제 발굴에 중요한 역할을 했다. 기존 약물의 재사용(drug repurposing)을 위한 AI 모델들이 개발되어, 수개월 내에 유력한 후보 물질들을 식별했다. 예를 들어, 렘데시비르와 같은 기존 약물들이 AI 기반 스크리닝을 통해 COVID-19 치료제로 재용도화되었다 (https://www.sciencedirect.com/science/article/pii/S

알츠하이머병 치료제 개발: 복잡한 신경퇴행성 질환인 알츠하이머병에 대해서도 AI 기반 접근법이 새로운 가능성을 제시하고 있다. 멀티오믹스 데이터를 활용한 AI 모델들이 새로운 표적을 발굴하고, 기존에 실패했던 접근법들의 한계를 극복하는 새로운 전략을 제시하고 있다.

항생제 내성 문제 해결: AI 기술은 항생제 내성 문제 해결에도 기여하고 있다. 새로운 작용 기전을 가진 항생제 후보 물질들이 AI 기반 스크리닝을 통해 발굴되고 있으며, 기존 항생제의 효과를 증진시키는 조합 요법도 AI를 통해 개발되고 있다.

18.6 임상시험의 설계와 방법론

18.6.1 임상시험의 개요

임상시험은 새로운 약물의 안전성과 효능을 평가하기 위해 인간을 대상으로 수행되는 체계적인 연구이다. 임상시험은 엄격한 과학적 방법론과 윤리적 기준에 따라 설계되고 수행되어야 한다. 이러한 연구는 환자의 안전을 최우선으로 하면서도, 과학적으로 신뢰할 수 있는 결과를 도출하여 **근거 기반 의학**의 토대를 제공한다.

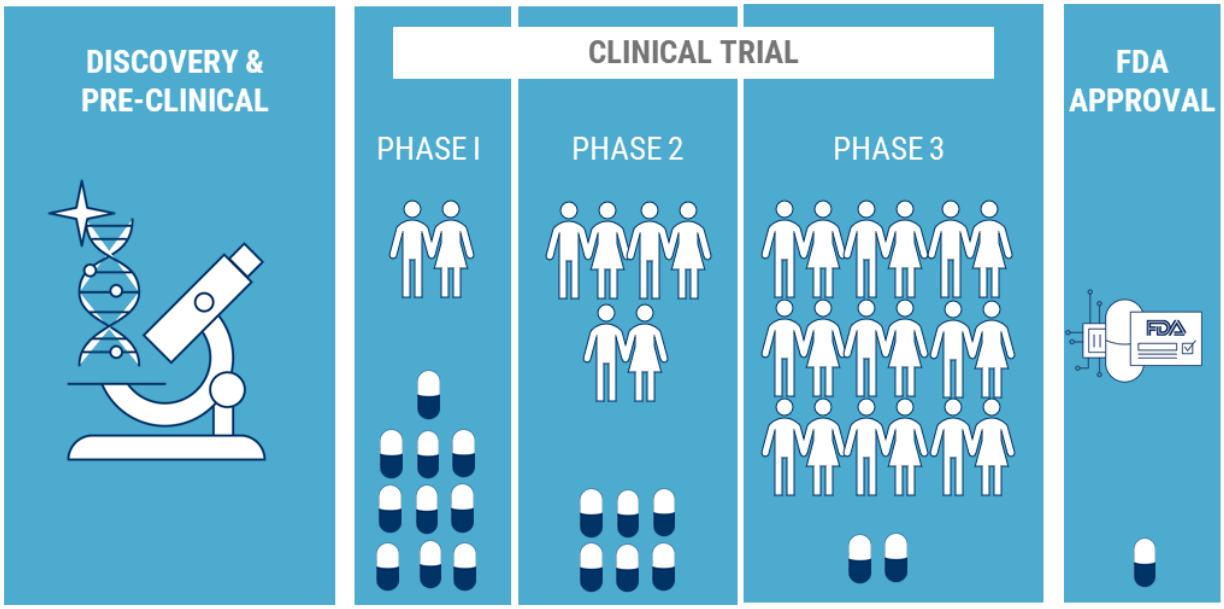


Figure 346: 임상시험의 개요

Figure 17.18 임상시험 과정

근거 기반 의학과 근거의 위계: 근거 기반 의학(Evidence-Based Medicine)은 현재 이용 가능한 최선의 과학적 근거를 통합하여 의학적 결정을 내리는 접근법이다. 모든 과학적 근거가 동등한 가치를 갖는 것이 아니라, 연구 설계의 엄격성과

방법론에 따라 질적 차이가 있다. 의학 연구에서 근거의 질은 피라미드 구조로 표현되며, 위에서부터 아래로 갈수록 근거의 질이 낮아진다:

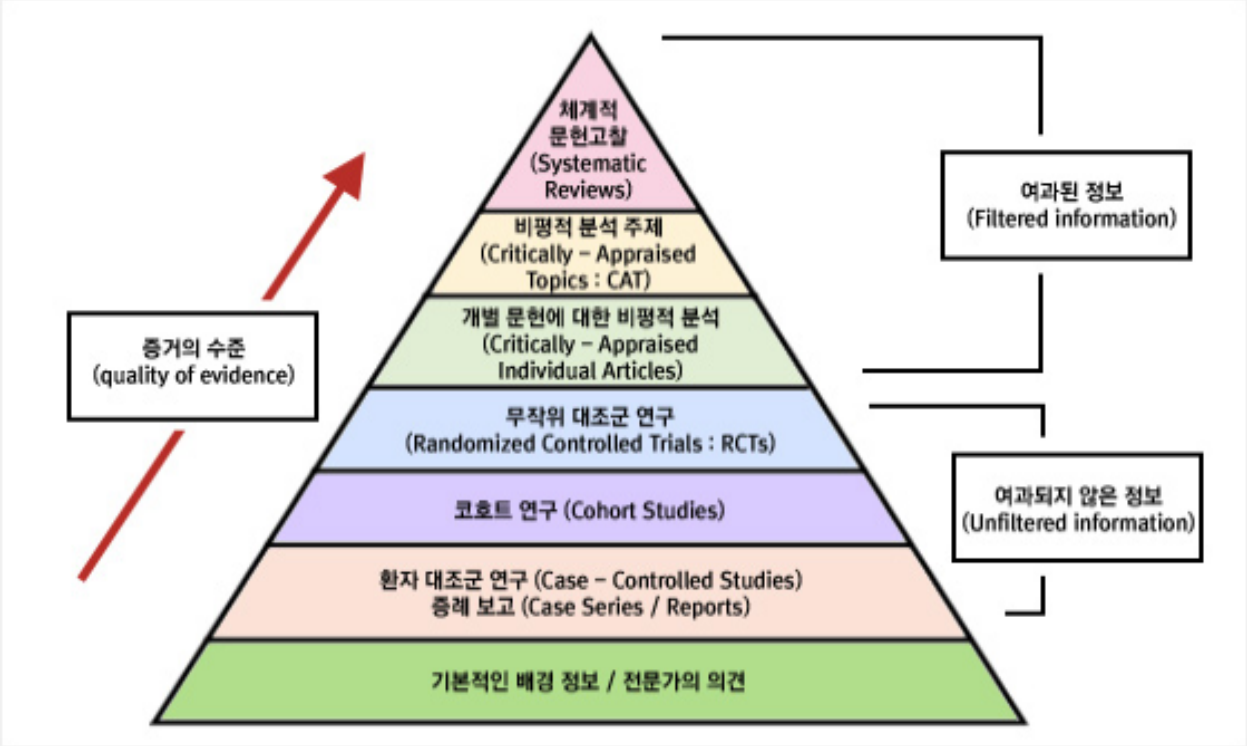


Figure 347: 근거의 피라미드

Figure 17.19 근거의 피라미드

1. **체계적 고찰 및 메타분석 (Systematic Reviews & Meta-analyses)**: 최고 수준의 근거
2. **무작위 대조 임상시험 (Randomized Controlled Trials, RCT)**: 높은 수준의 근거
3. **코호트 연구 (Cohort Studies)**: 중간 수준의 근거
4. **사례-대조 연구 (Case-Control Studies)**: 낮은 수준의 근거
5. **사례 보고서 (Case Reports)**: 더 낮은 수준의 근거
6. **전문가 의견 (Expert Opinion)**: 가장 낮은 수준의 근거

사례 보고서와 전문가 의견은 근거 위계에서 가장 낮은 단계에 위치하지만, 새로운 의학 지식의 출발점이 되거나 다른 근거가 부족한 상황에서 의학 발전의 초기 단계에서 중요한 역할을 한다.

그러나 전문가 의견은 개인의 경험과 주관적 판단에 의존하는 특성상 체계적 편견이나 오류의 가능성이 높다. 역사적으로 많은 의학적 관행들이 전문가의 권위에 의존하여 오랫동안 유지되었지만, 후에 엄격한 과학적 연구를 통해 효과가 없거나 오히려 해로운 것으로 밝혀진 사례들이 있다. 따라서 근거 기반 의학에서는 전문가의 판단을 맹신하기보다는 객관적이고 재현 가능한 과학적 증거를 우선시하며, 전문가 의견은 다른 근거가 부족한 상황에서만 보조적으로 활용해야 한다.

Figure 17.20 권위에 대한 맹신의 위험성

상관관계와 인과관계의 구분: 임상시험의 관점에서 가장 중요한 개념 중 하나는 상관관계와 인과관계를 정확히 구분하는 것이다. 상관관계는 두 변수 간의 연관성을 의미하지만, 반드시 인과관계를 의미하지는 않는다. 많은 경우 강한 상관관계가 있어도 실제로는 제3의 변수(혼동 요인, confounding factor)나 우연의 일치에 의한 것일 수 있다.

진정한 인과관계를 확립하기 위해서는 다음과 같은 조건들이 충족되어야 한다: - 시간적 순서의 확립 (원인이 결과보다 선행해야 함) - 혼동 요인의 제거 (제3의 변수를 통제) - 생물학적 개연성 (생물학적으로 합리적인 메커니즘) - 용량-반응 관계 (노출량이 늘수록 효과 증가)

이러한 이유로 임상시험에서는 무작위 배정과 대조군 설정을 통해 인과관계를 확립하려고 한다. 무작위 배정은 알려진 및

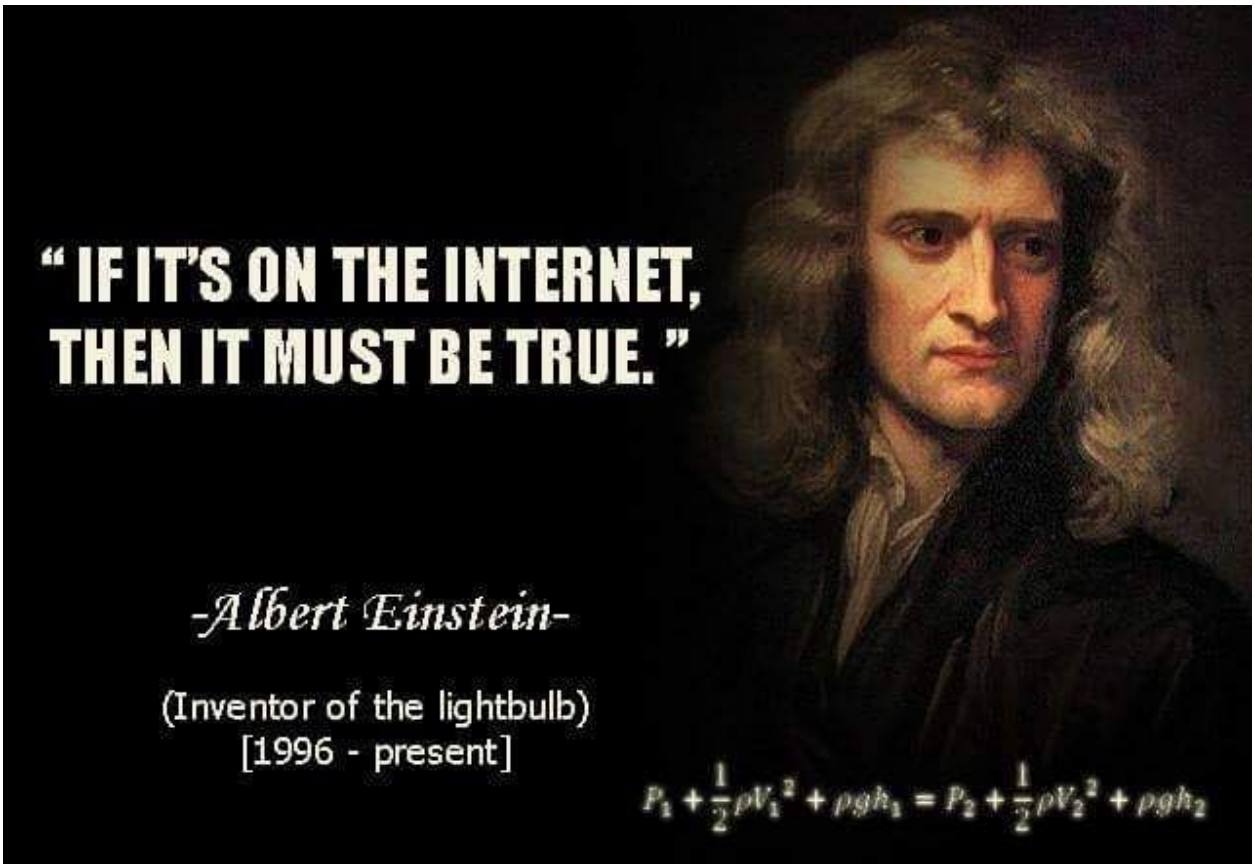


Figure 348: 권위에 대한 맹신의 위험성

알려지지 않은 모든 혼동 요인을 실험군과 대조군에 균등하게 분포시켜, 치료법의 순수한 효과만을 분리하여 평가할 수 있게 한다.

18.6.2 혼동 요인과 임상시험에서의 통제

혼동 요인(confounding factor)은 노출 변수(치료법)와 결과 변수(치료 결과) 모두에 영향을 미치는 제3의 변수로, 진정한 인과관계를 왜곡할 수 있는 요소를 말한다.

대표적인 예로 초콜릿 소비량과 노벨상 수상자 수 간의 상관관계를 들 수 있다. 한 연구에서는 국가별 1인당 초콜릿 소비량과 인구 1천만 명당 노벨상 수상자 수 사이에 강한 양의 상관관계(r=0.791, p<0.0001)가 발견되었다. 그러나 이는 명백한 가성 상관관계의 사례이다. 초콜릿은 일종의 사치품으로 볼 수 있고, 부유한 국가일수록 초콜릿 소비가 높은 편이다. 따라서 실제로는 국내총생산(GDP)이라는 혼동 요인이 존재한다고 볼 수 있으며, GDP가 초콜릿 소비량과 노벨상 수상자 수 모두에 인과적 영향을 미치는 진짜 원인일 수 있는 것이다.

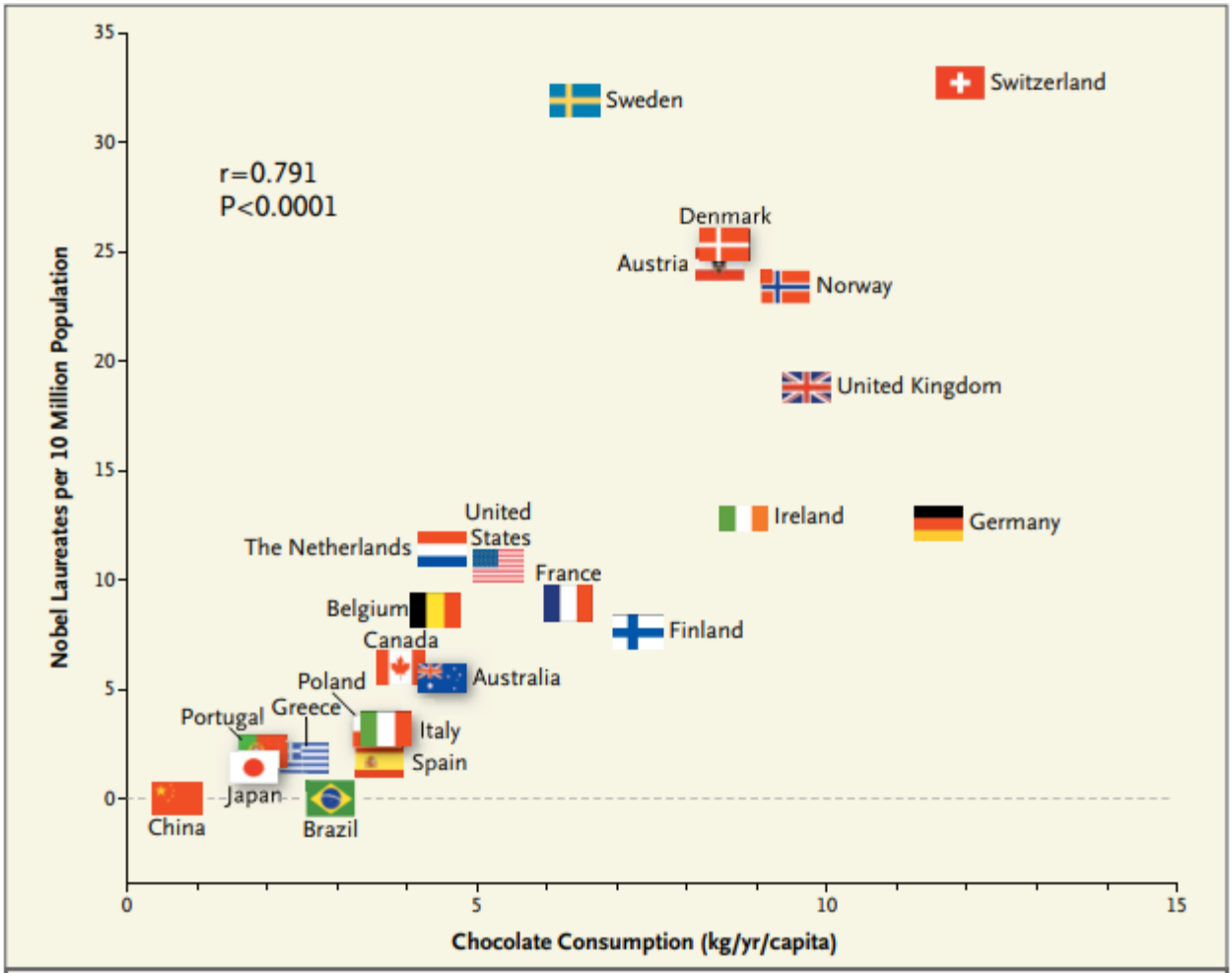


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure 349: 초콜릿과 노벨상의 상관관계

Figure 17.22 초콜릿과 노벨상의 상관관계

Figure 17.23 혼동 요인

혼동 요인을 인식하는 것은 주장의 타당성에 대한 사전 믿음에 따라 달라질 수 있다. 상식적으로 받아들이기 어려운

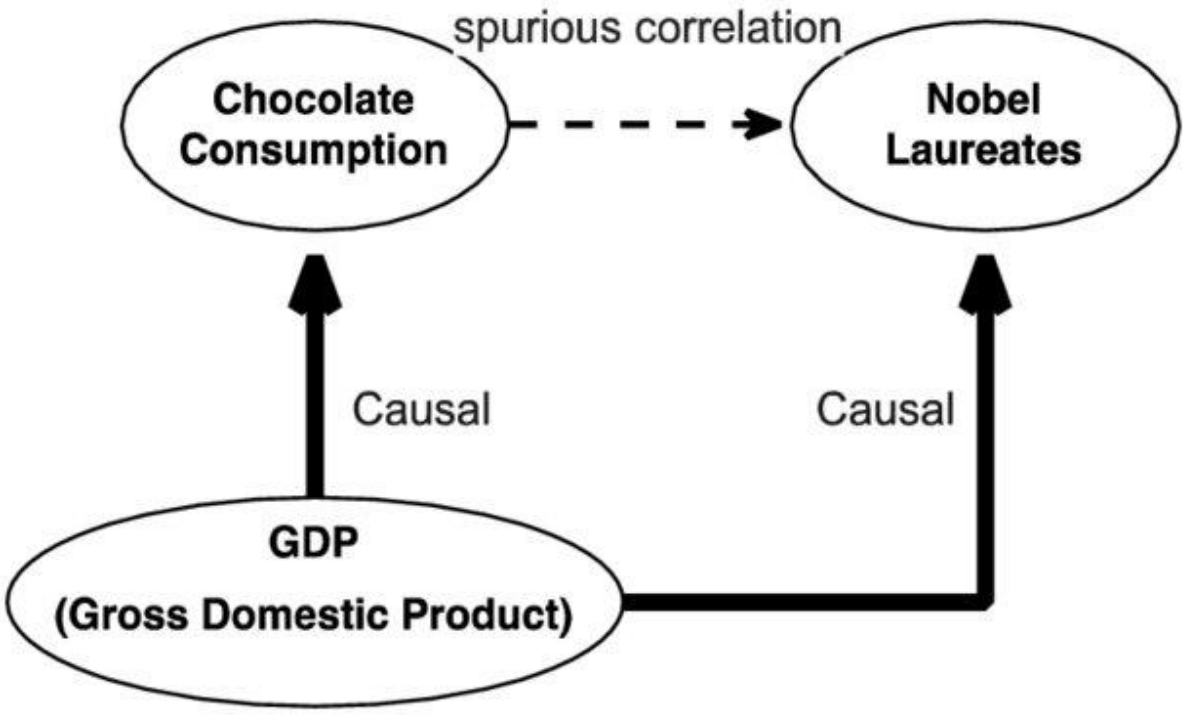


Figure 350: 혼동 요인의 개념

주장("초콜릿을 먹기만 해도 뇌 기능이 좋아진다")에 대해서는 혼동 요인을 쉽게 의심하지만, 믿고 싶거나 그럴듯해 보이는 주장일수록 혼동 요인의 존재를 간과하기 쉽다. 예를 들어, 유럽연합 집행위원회가 발표한 "백신 접종률이 높을수록 사망률이 낮다"는 데이터는 많은 사람들이 사실로 받아들이기 쉽다. 그러나 이 경우에도 **국가별 소득 수준**이라는 중요한 혼동 요인이 존재한다. 경제적으로 발달한 국가일수록 백신 접종과 무관하게 우수한 의료 시스템으로 인해 사망률이 낮을 수 있는 것이다.

Figure 17.21 혼동 요인의 예시 – 백신 접종률과 사망률의 상관관계에서 국가별 경제 수준이 혼동 요인으로 작용함

임상시험에서의 혼동 요인 통제: 임상시험에서 흔히 접하는 혼동 요인들은 다음과 같다: – **인구학적 요인**: 나이, 성별, 인종, 사회경제적 상태 – **질병 관련 요인**: 질병의 중증도, 유병 기간, 동반 질환 – **생활습관 요인**: 흡연, 음주, 운동, 식습관 – **치료 관련 요인**: 이전 치료 경험, 동반 약물 사용, 치료 순응도

무작위 배정의 중요성: 임상시험에서 혼동 요인을 통제하는 가장 일반적인 방법은 **무작위 배정(randomization)**이다. 무작위 배정은 알려진 혼동 요인뿐만 아니라 알려지지 않은 혼동 요인까지도 치료군과 대조군에 균등하게 분배한다. 이를 통해 두 군 간의 차이가 순수하게 치료법에 의한 것임을 보장할 수 있다.

무작위 배정의 핵심 장점은 연구자가 미처 고려하지 못한 혼동 요인들까지도 자동으로 통제된다는 점이다. 이러한 이유로 무작위 대조 임상시험은 근거 기반 의학에서 높은 수준의 근거로 평가받으며, 새로운 치료법의 효과를 입증하는 표준이 되고 있다.

18.6.3 임상시험의 단계

임상시험은 안전성과 효능을 체계적으로 평가하기 위해 단계적으로 진행된다. 각 단계는 특정한 목적과 방법론을 가지고 있으며, 이전 단계에서 얻은 결과를 바탕으로 다음 단계로 진행할지 결정한다.

전임상 단계 (Preclinical): 전임상 단계는 인간에게 적용하기 전에 실험실과 동물 모델에서 기본적인 안전성과 효능을 평가하는 단계이다. 이 단계에서는 다음과 같은 연구가 수행된다: – 시험관 내 실험 (in vitro) – 생체 내 실험 (in vivo) – 기본적인 안전성 및 효능 평가

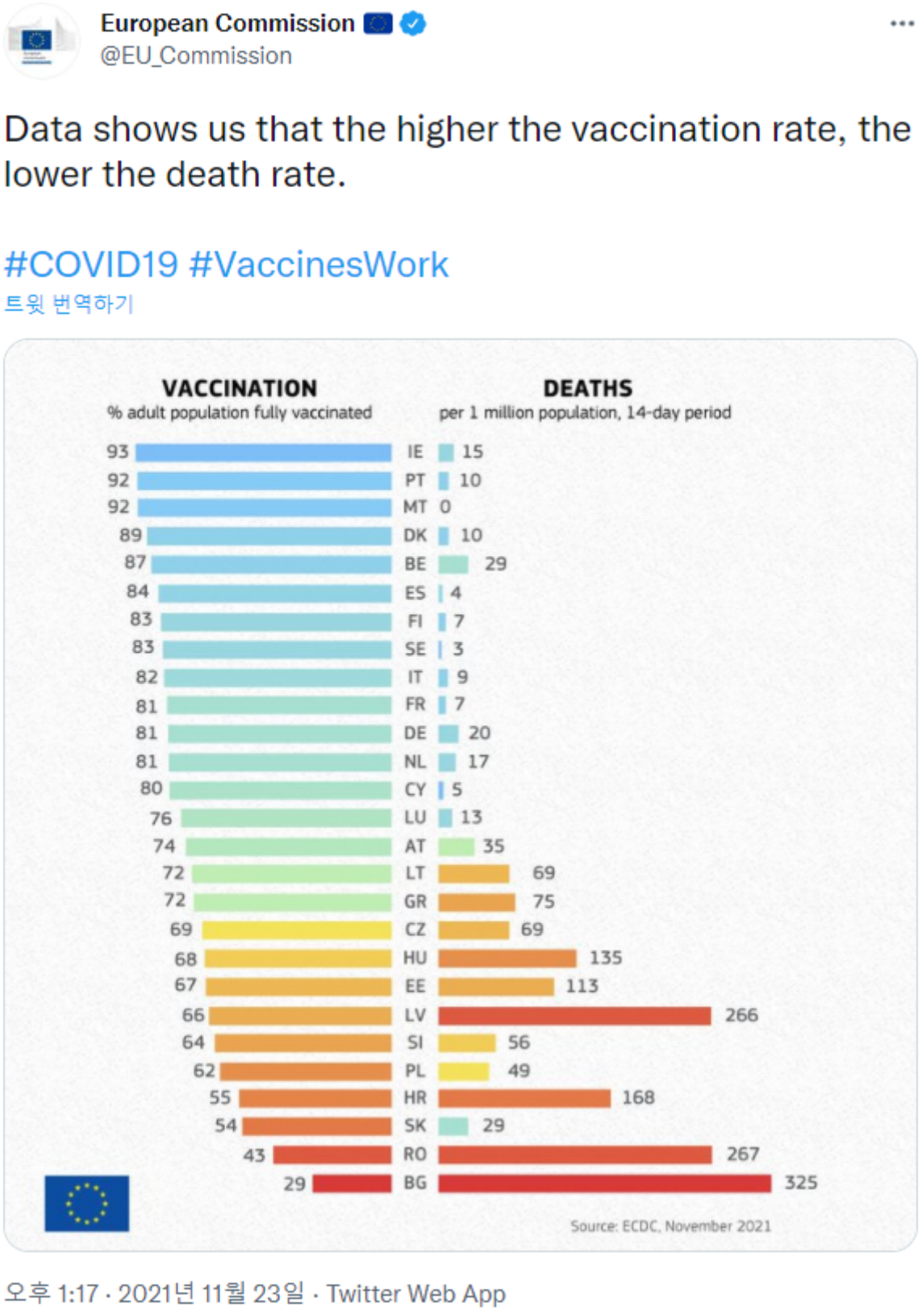


Figure 351: 유럽연합 집행위원회의 백신 접종률과 사망률 상관관계 데이터

1상 임상시험 (Phase I): 1상 임상시험은 새로운 치료법의 첫 번째 인간 적용 단계로, 주로 안전성 평가에 중점을 둔다. 이 단계에서는 소수의 참여자를 대상으로 조심스럽게 용량을 증가시키면서 안전한 투여 범위를 찾는다: - 참여자: 10-100명의 지원자 - 기간: 수개월 - 목적: 주로 안전성 평가, 적정 용량 결정 - 건강한 지원자 또는 환자를 대상으로 실시

2상 임상시험 (Phase II): 2상 임상시험은 1상에서 확인된 안전한 용량 범위 내에서 치료법의 효능을 평가하는 단계이다. 이 시점에서 연구자들은 치료법이 실제로 질병에 대해 치료 효과를 나타내는지 확인하고, 최적의 용량을 결정하기 위한 보다 정교한 분석을 수행한다: - 참여자: 100-1,000명의 지원자 - 기간: 수개월에서 수년 - 목적: 안전성, 부작용, 그리고 효능 평가 - 특정 질환을 가진 환자를 대상으로 실시

3상 임상시험 (Phase III): 3상 임상시험은 임상개발의 마지막 단계로, 새로운 치료법이 기존 표준 치료법보다 우수하거나 최소한 동등한 효과를 보이는지 확인하는 결정적인 단계이다. 이 단계의 결과는 규제 당국의 승인 결정에 직접적인 영향을 미치며, 따라서 가장 엄격한 과학적 기준을 적용한다. 3상 임상시험에서 안전성과 효능이 성공적으로 입증되면, 해당 약물은 규제 당국(FDA, EMA 등)의 최종 승인을 받아 시장에 출시될 수 있다: - 참여자: 100-10,000명의 지원자 - 기간: 2-10년 - 목적: 장기간 안전성과 확실한 효능 입증 - 대규모 다기관 연구로 진행

4상 임상시험 (Phase IV): 4상 임상시험은 시판 후 감시(post-marketing surveillance) 단계로, 치료법이 실제 임상 환경에서 사용될 때의 장기적인 안전성과 효능을 지속적으로 모니터링한다. 이 단계는 임상시험 환경에서는 발견하기 어려운 드문 부작용이나 특정 인구 집단에서의 특별한 반응을 탐지하는 데 중요한 역할을 한다: - 시장 승인 후 무제한 지속적인 데이터 수집 - 대중에게 공개된 후 장기간 안전성 모니터링 - 드물게 나타나는 부작용 탐지

Figure 17.24 4상 임상시험 과정

Figure 17.25 4상 임상시험 결과 예시

18.6.4 임상시험 설계 원칙

임상시험의 가장 중요한 목적은 단순한 상관관계를 넘어서 진정한 인과관계를 확립하는 것이다. 앞서 논의한 바와 같이, 두 변수 간의 연관성만으로는 한 변수가 다른 변수를 직접적으로 유발한다고 결론내릴 수 없다. 임상시험의 설계 원칙들은 모두 이러한 인과관계 확립을 위해 혼동 요인을 통제하고 편견을 제거하는 데 초점을 맞추고 있다.

무작위 배정 (Randomization) 무작위 배정은 상관관계와 인과관계를 구분하는 가장 강력한 도구이다. 치료군과 대조군을 무작위로 배정함으로써, 치료법 이외의 모든 요인들이 두 군에 균등하게 분포되도록 한다. 이를 통해 관찰되는 결과의 차이가 순수하게 치료법에 의한 것임을 보장할 수 있다: - **혼동 요인의 균등한 분포**: 알려진 요인(나이, 성별, 질병 중증도 등)과 알려지지 않은 요인 모두를 두 군에 고르게 분배 - **선택 편견의 제거**: 연구자나 환자의 주관적 판단이 군 배정에 영향을 미치지 않도록 방지 - **제3의 변수 효과 최소화**: 치료법과 결과 사이에 개입할 수 있는 다른 변수들의 영향을 통제

맹검법 (Blinding) 맹검법은 인과관계 확립에서 중요한 심리적 혼동 요인들을 통제하는 방법이다. 환자나 의료진이 치료법을 알고 있을 경우, 이러한 지식 자체가 결과에 영향을 미칠 수 있어 진정한 치료 효과와 심리적 효과를 구분하기 어려워진다. 맹검법을 통해 다음과 같은 편견을 제거할 수 있다: - **위약 효과 통제**: 환자의 기대감이 실제 치료 효과와 분리되어 측정됨 - **관찰자 편견 제거**: 의료진의 선입견이 결과 평가에 미치는 영향 차단 - **행동 변화 방지**: 치료법 인지로 인한 환자나 의료진의 행동 변화 최소화

맹검법의 유형: - **단일맹검 (Single-blind)**: 환자만 치료법을 모르는 경우 - **이중맹검 (Double-blind)**: 환자와 의료진 모두 치료법을 모르는 경우 - **삼중맹검 (Triple-blind)**: 환자, 의료진, 분석자 모두 치료법을 모르는 경우

18.6.5 통계적 분석 방법

임상시험 데이터의 적절한 통계적 분석은 연구 결과의 과학적 신뢰성을 확보하는 데 필수적이다. 다양한 연구 설계에 따라 적절한 통계적 지표를 선택하고 해석하는 것이 중요하며, 이를 통해 치료의 효과를 정량적으로 평가할 수 있다.

위험비 (Risk Ratio, RR): 위험비는 코호트 연구나 무작위 대조 임상시험에서 주로 사용되는 지표로, 노출군과 비노출군 간의 사건 발생 위험을 비교한다. RR이 1보다 클 경우 위험 증가를, 1보다 작을 경우 위험 감소를 의미한다:

RR = (실험군에서의 발생률) / (대조군에서의 발생률)

오즈비 (Odds Ratio, OR): 오즈비는 사례-대조 연구에서 주로 사용되는 지표로, 특정 노출이 있는 경우와 없는 경우의 오즈비를 비교한다. OR이 1보다 클 경우 양의 연관성을, 1보다 작을 경우 음의 연관성을 의미한다. 이 지표는 교차표 분석을 통해 계산된다:

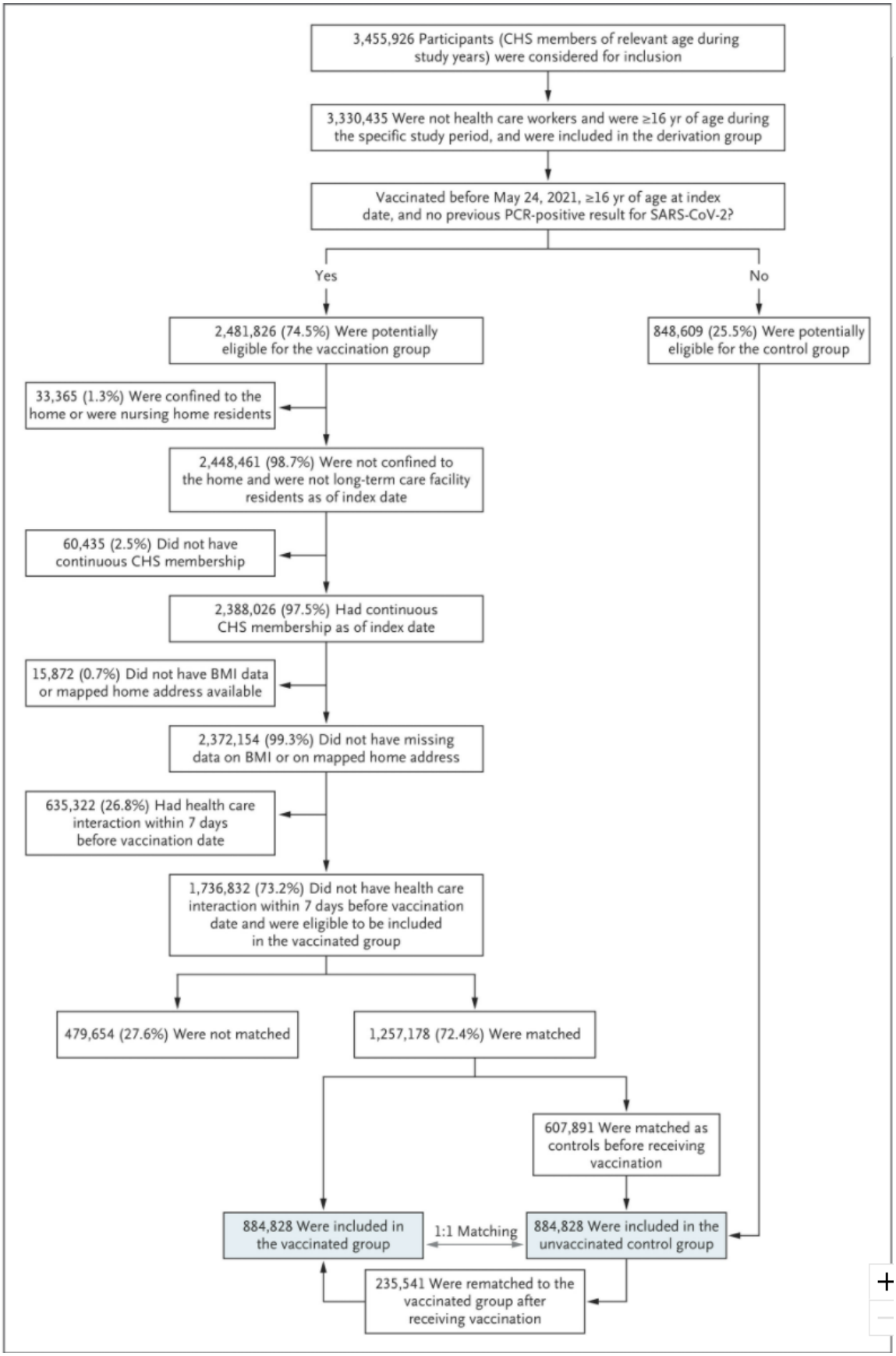


Figure 352: 4상 임상시험 개요

Table 2. Adverse Events Associated with SARS-CoV-2 Vaccination.*

Event	Adverse-Event Cohort in Each Group	Vaccinated Group	Control Group	Risk Ratio (95% CI)	Risk Difference (95% CI)
	no. of persons	*no. of events*			*no. of events/100,000 persons*
Acute kidney injury	912,019	20	45	0.44 (0.23 to 0.73)	−4.6 (−7.8 to −1.8)
Anemia	709,267	298	378	0.79 (0.67 to 0.93)	−18.7 (−32.1 to −6.1)
Appendicitis	900,289	95	66	1.40 (1.02 to 2.01)	5.0 (0.3 to 9.9)
Arrhythmia	856,152	254	284	0.89 (0.74 to 1.04)	−6.1 (−14.7 to 1.8)
Arthritis or arthropathy	731,340	64	70	0.95 (0.65 to 1.34)	−0.8 (−6.3 to 4.2)
Bell's palsy	923,692	81	59	1.32 (0.92 to 1.86)	3.5 (−1.1 to 7.8)
Cerebrovascular accident	917,598	45	55	0.84 (0.54 to 1.27)	−1.6 (−5.3 to 2.0)
Deep-vein thrombosis	925,380	39	47	0.87 (0.55 to 1.40)	−1.1 (−4.5 to 2.7)
Herpes simplex infection	876,328	219	205	1.13 (0.95 to 1.38)	4.8 (−1.9 to 12.4)
Herpes zoster infection	888,647	283	204	1.43 (1.20 to 1.73)	15.8 (8.2 to 24.2)
Intracranial hemorrhage	933,130	13	30	0.48 (0.20 to 0.89)	−2.9 (−5.6 to −0.5)
Lymphadenopathy	823,006	660	279	2.43 (2.05 to 2.78)	78.4 (64.1 to 89.3)
Lymphopenia	938,939	2	7	0.26 (0.00 to 1.03)	−0.9 (−2.0 to <0.1)
Myocardial infarction	892,785	59	60	1.07 (0.74 to 1.60)	0.8 (−3.3 to 5.2)
Myocarditis	938,812	21	6	3.24 (1.55 to 12.44)	2.7 (1.0 to 4.6)
Neutropenia	919,291	20	22	0.87 (0.46 to 1.66)	−0.5 (−2.8 to 1.8)
Other thrombosis†	932,469	12	22	0.46 (0.19 to 0.91)	−2.2 (−4.6 to −0.3)
Paresthesia	827,478	552	496	1.12 (0.98 to 1.24)	10.8 (−1.8 to 21.4)
Pericarditis	936,197	27	18	1.27 (0.68 to 2.31)	1.0 (−1.6 to 3.4)
Pulmonary embolism	937,116	10	17	0.56 (0.21 to 1.15)	−1.5 (−3.6 to 0.4)
Seizure	913,091	36	35	0.99 (0.62 to 1.64)	−0.4 (−3.0 to 3.1)
Syncope	858,068	326	267	1.12 (0.94 to 1.34)	6.2 (−3.2 to 15.4)
Thrombocytopenia	923,123	56	60	0.94 (0.63 to 1.27)	−0.6 (−4.6 to 2.3)
Uveitis	933,217	26	20	1.27 (0.68 to 2.67)	1.0 (−1.5 to 3.8)
Vertigo	773,263	433	395	1.12 (0.97 to 1.28)	9.3 (−2.5 to 20.0)

* Estimates were calculated with the use of the Kaplan–Meier estimator 42 days after vaccination or SARS-CoV-2 infection. Confidence intervals (CIs) were estimated with the use of the percentile bootstrap method with 500 repetitions.
† The "other thrombosis" category is a composite diagnosis that includes arterial embolism and thrombosis, venous embolism and thrombosis, vascular insufficiency of the intestine, portal-vein thrombosis, or cranial venous sinus thrombosis.

Figure 353: 4상 임상시험 결과

OR = (a × d) / (b × c)

여기서 a, b, c, d는 2×2 교차표의 각 셀 값이다.

백신 효능 (Vaccine Efficacy): 백신 효능은 백신이 질병을 예방하는 능력을 나타내는 지표로, 전체 인구에서 백신 접종으로 인해 예방될 수 있는 질병 사례의 비율을 나타낸다. 100%에 가까울수록 높은 예방 효과를 의미한다:

VE = [1 − (백신 접종자에서의 사례 수 / 미접종자에서의 사례 수)] × 100%

18.6.6 COVID-19 백신 임상시험 사례

COVID-19 팬데믹 상황에서 신속하게 개발된 백신들의 임상시험은 현대적인 임상시험 방법론의 우수성과 효율성을 입증하는 대표적인 사례이다. 이러한 임상시험들은 전 세계적인 위기 상황에서도 과학적 엄격성을 유지하면서 신뢰할 수 있는 결과를 도출한 성공적인 사례로 평가된다.

Pfizer/BNT162b2 백신 임상시험: Pfizer/BNT162b2 백신의 3상 임상시험은 대규모 무작위 대조 임상시험의 모범 사례로 여겨진다. 이 연구는 엄격한 과학적 기준과 계획된 국제적 협력을 통해 단기간에 신뢰할 수 있는 결과를 도출했다: – 총 참여자: 44,320명 – 다국가 연구: 아르헨티나, 브라질, 독일, 남아프리카, 터키, 미국 – 인구학적 다양성 확보 – 백신 효능: 94.6% (2차 접종 후) – 포괄적인 안전성 모니터링 – 다양한 평가 지표에서 백신군과 대조군 비교

Figure 17.29 COVID-19 3상 임상시험 포함 기준 예시 – 다양한 인구학적 배경을 포함하여 연구 결과의 대표성 확보

Figure 17.30 COVID-19 3상 임상시험 과정 예시

Figure 17.31 COVID-19 3상 임상시험 결과 예시

이 연구는 대규모 무작위 대조 임상시험의 전형적인 예시로, 엄격한 방법론과 다양한 인구 집단을 포함한 포괄적인 평가를 보여준다.

18.6.7 임상시험 연구 설계의 중요성

연구 설계는 연구 결과의 신뢰성과 유용성을 결정하는 가장 중요한 요소이다. 잘 설계된 연구는 편견을 최소화하고 신뢰할 수 있는 결과를 도출하는 반면, 부실한 연구 설계는 아무리 정교한 통계 분석을 수행해도 의미 있는 결론을 도출하기 어렵다.

Table 1. Demographic Characteristics of the Participants in the Main Safety Population.*

Characteristic	BNT162b2 (N=18,860)	Placebo (N=18,846)	Total (N=37,706)
Sex — no. (%)			
Male	9,639 (51.1)	9,436 (50.1)	19,075 (50.6)
Female	9,221 (48.9)	9,410 (49.9)	18,631 (49.4)
Race or ethnic group — no. (%)†			
White	15,636 (82.9)	15,630 (82.9)	31,266 (82.9)
Black or African American	1,729 (9.2)	1,763 (9.4)	3,492 (9.3)
Asian	801 (4.2)	807 (4.3)	1,608 (4.3)
Native American or Alaska Native	102 (0.5)	99 (0.5)	201 (0.5)
Native Hawaiian or other Pacific Islander	50 (0.3)	26 (0.1)	76 (0.2)
Multiracial	449 (2.4)	406 (2.2)	855 (2.3)
Not reported	93 (0.5)	115 (0.6)	208 (0.6)
Hispanic or Latinx	5,266 (27.9)	5,277 (28.0)	10,543 (28.0)
Country — no. (%)			
Argentina	2,883 (15.3)	2,881 (15.3)	5,764 (15.3)
Brazil	1,145 (6.1)	1,139 (6.0)	2,284 (6.1)
South Africa	372 (2.0)	372 (2.0)	744 (2.0)
United States	14,460 (76.7)	14,454 (76.7)	28,914 (76.7)
Age group — no. (%)			
16–55 yr	10,889 (57.7)	10,896 (57.8)	21,785 (57.8)
>55 yr	7,971 (42.3)	7,950 (42.2)	15,921 (42.2)
Age at vaccination — yr			
Median	52.0	52.0	52.0
Range	16–89	16–91	16–91
Body-mass index‡			
≥30.0: obese	6,556 (34.8)	6,662 (35.3)	13,218 (35.1)

* Percentages may not total 100 because of rounding.
† Race or ethnic group was reported by the participants.
‡ The body-mass index is the weight in kilograms divided by the square of the height in meters.

Figure 354: COVID-19 3상 임상시험 포함 기준

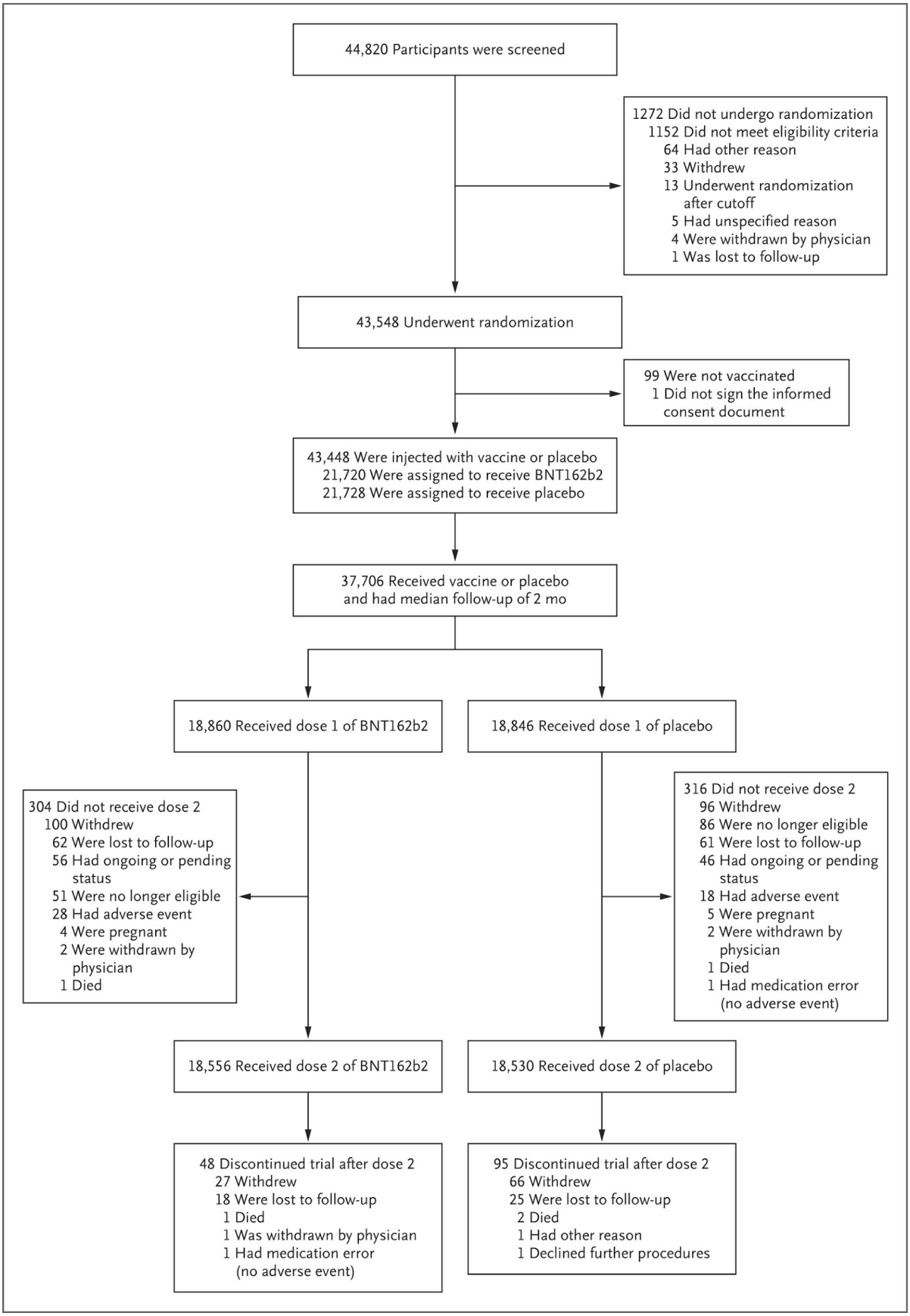


Figure 355: COVID-19 3상 임상시험 개요

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.*

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval)‡	Posterior Probability (Vaccine Efficacy >30%)§
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
		(N=19,965)		(N=20,172)		
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.
† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.
‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.
§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

Figure 356: COVID-19 3상 임상시험 결과

대표성 있는 연구 집단의 중요성을 위해서는 연구 대상이 전체 인구를 적절히 대표해야 한다. 만약 연구 대상이 특정 인구 집단에만 편중되어 있다면, 그 결과를 다른 인구 집단에 적용하기 어려울 수 있다. 편향된 연구 집단은 잘못된 결론으로 이어질 수 있다.

낙하산 임상시험 (Parachute Study)는 연구 설계의 중요성을 강조하기 위해 제시된 예시이다. 이 연구는 낙하산의 효능을 무작위 대조 임상시험으로 검증하자고 제안하면서, 자명한 사실에 대해서도 잘못된 연구 설계가 잘못된 과학적 결론으로 이어질 수 있음을 풍자적으로 보여주었다. 이 예시는 부적절한 연구 설계가 어떻게 잘못된 결론으로 이어질 수 있는지를 보여주는 예시이다. 이를 통해 대표성 있는 연구 집단의 중요성과 적절한 연구 방법론의 필요성을 보여준다.

Figure 17.32 낙하산의 치명률 RCT 연구 개요

Figure 17.33 낙하산의 치명률 RCT 연구 포함 기준

Figure 17.34 낙하산의 치명률 RCT 연구 결과

Figure 17.35 낙하산의 치명률 RCT 연구 각주 – 저자들의 재치를 볼 수 있음

18.6.8 멘델리안 무작위배정

멘델리안 무작위배정(Mendelian Randomization, MR)은 전통적인 관찰 연구에서 피할 수 없는 혼동 요인과 역인과관계의 문제를 해결하기 위해 개발된 혁신적인 연구 방법이다. 이 방법은 유전의 기본 원리인 멘델의 법칙을 활용하여, 유전적 변이가 자연적인 무작위 배정 도구 역할을 할 수 있다는 점에 착안하였다. 특히 GWAS(유전체 전장 연관성 분석) 데이터의 축적과 함께 최근 10년간 급속히 발전하고 있는 분야이다.

기본 개념과 원리: 멘델리안 무작위배정은 유전적 변이를 도구변수(instrumental variable)로 사용하여 인과관계를 추론하는 방법이다. 이 방법의 핵심은 부모로부터 자녀에게 유전자가 전달되는 과정이 본질적으로 무작위적이라는 멘델의 독립 분리 법칙에 기반한다. 즉, 개인이 특정 유전적 변이를 가지게 되는 것은 태생적으로 결정되며, 이후의 환경적 요인이나 생활 습관과는 독립적이다. 이러한 특성으로 인해 유전적 변이는 수정불가능하고 무작위로 배정되는 특성을 가지고 있어, 전통적인 무작위 대조 임상시험에서 달성하고자 하는 무작위 배정의 효과를 자연스럽게 만들어낸다.

MR 연구가 유효하기 위해서는 세 가지 핵심 가정이 충족되어야 한다: (1) 관련성 가정(relevance assumption) – 유전적 변이가 노출 변수와 강하게 연관되어야 함, (2) 독립성 가정(independence assumption) – 유전적 변이가 알려진 또는 알려지지 않은 혼동 요인과 연관되지 않아야 함, (3) 배제 제약 가정(exclusion restriction assumption) – 유전적 변이가 노출 변수를 통해서만 결과 변수에 영향을 미쳐야 함.

적용 분야와 활용 사례: MR은 다양한 의학 연구 분야에서 활용되고 있다. 대표적인 예로는 혈중 콜레스테롤 수치와 심혈관 질환의 인과관계 규명, 비만과 당뇨병의 관계 분석, 알코올 섭취와 간질환의 연관성 평가 등이 있다. 특히 흡연과 폐암의 관계처럼 윤리적으로 무작위 대조 임상시험을 수행하기 어려운 경우나, 비타민 D 수치와 다양한 질병의 관계처럼 장기간의 노출 효과를 평가해야 하는 경우에 매우 유용하다.

Figure 17.36 멘델리안 무작위배정 – 유전적 변이를 도구변수로 사용하여 흡연과 폐암 사이의 인과관계 추론

최근에는 약물 표적 발굴에도 활용되고 있는데, 특정 단백질의 기능을 모방하는 약물의 효과를 예측하기 위해 해당 단백질을 코딩하는 유전자의 변이를 분석하는 방식으로 사용되고 있다. 이를 통해 임상시험 전에 약물의 효능과 부작용을 미리 예측할

Research » Christmas 2018: Look Before You Leap

Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial

BMJ 2018 ; 363 doi: https://doi.org/10.1136/bmj.k5094 (Published 13 December 2018)

Abstract

Objective To determine if using a parachute prevents death or major traumatic injury when jumping from an aircraft.

Design Randomized controlled trial.

Setting Private or commercial aircraft between September 2017 and August 2018.

Participants 92 aircraft passengers aged 18 and over were screened for participation. 23 agreed to be enrolled and were randomized.

Intervention Jumping from an aircraft (airplane or helicopter) with a parachute versus an empty backpack (unblinded).

Main outcome measures Composite of death or major traumatic injury (defined by an Injury Severity Score over 15) upon impact with the ground measured immediately after landing.

Results Parachute use did not significantly reduce death or major injury (0% for parachute *v* 0% for control; P>0.9). This finding was consistent across multiple subgroups. Compared with individuals screened but not enrolled, participants included in the study were on aircraft at significantly lower altitude (mean of 0.6 m for participants *v* mean of 9146 m for non-participants; P<0.001) and lower velocity (mean of 0 km/h *v* mean of 800 km/h; P<0.001).

Conclusions Parachute use did not reduce death or major traumatic injury when jumping from aircraft in the first randomized evaluation of this intervention. However, the trial was only able to enroll participants on small stationary aircraft on the ground, suggesting cautious extrapolation to high altitude jumps. When beliefs regarding the effectiveness of an intervention exist in the community, randomized trials might selectively enroll individuals with a lower perceived likelihood of benefit, thus diminishing the applicability of the results to clinical practice.

Figure 357: 낙하산 연구 개요

Characteristics	Parachute	Control
Total	11 (100)	12 (100)
Demographics		
Median (SD) age (years)	38.1 (8.7)	38.6 (11.0)
Women	4 (36)	6 (50)
Men	7 (64)	6 (50)
Ethnic group:		
American Indian or Alaska Native	0 (0)	0 (0)
East Asian or South Asian	4 (36)	4 (33)
Black or African American	0 (0)	0 (0)
More than one race	0 (0)	0 (0)
White	7 (64)	8 (67)
Mean (SD) height (cm)	171.8 (9.1)	171.7 (8.4)
Mean (SD) weight (kg)	75.9 (24.4)	74.6 (13.0)
Medical history		
Broken bones	4 (36)	5 (42)
Acrophobia	3 (27)	6 (50)
Parachute use	3 (27)	0 (0)
Family history of parachute use	2 (18)	0 (0)
Frequent flier (average >4 flights per month)	0 (0)	4 (33)
Flight		
International v domestic:		
International	0 (0)	0 (0)
Domestic	11 (100)	12 (100)
Aircraft type:		
Jetliner	0 (0)	0 (0)
Biplane	5 (46)	6 (50)
Helicopter	6 (55)	6 (50)
Mean (SD) velocity (km/h)	0 (0)	0 (0)
Mean (SD) altitude (m)	0.6 (0.1)	0.6 (0.1)

Figure 358: 낙하산 연구 포함 기준

Endpoint	Parachute	Control	Mean difference (95% CI)	P value
On impact				
Death or major traumatic injury	0 (0)	0 (0)	0	>0.9
Mean (SD) Injury Severity Score	0 (0)	0 (0)	0	>0.9
30 days after impact				
Death or major traumatic injury	0 (0)	0 (0)	0	>0.9
Mean (SD) Injury Severity Score	0 (0)	0 (0)	0	>0.9
Health status				
Mean (SD) Short Form Health Survey score	43.9 (1.8)	44.0 (2.4)	0.1 (−2.0 to 2.2)	0.9
Mean (SD) physical health subscore	19.6 (0.7)	19.7 (0.5)	0.04 (−0.5 to 0.6)	0.9
Mean (SD) mental health subscore	24.3 (1.3)	24.3 (2.1)	0.08 (−1.6 to 1.8)	0.9

Figure 359: 낙하산 연구 결과

Footnotes

Contributors: RWY had the original idea but was reluctant to say it out loud for years. In a moment of weakness, he shared it with MWY and BKN, both of whom immediately recognized this as the best idea RWY will ever have. RWY and LRV wrote the first draft. CS, DBK, JBS, EAS, and JLH provided critical review. RMD provided subject matter expertise. DSK took this work to another satirical level. All authors suffered substantial abdominal discomfort from laughter. RWY worried that BKN would not keep his mouth shut until the Christmas issue was published. All authors had full access to the data in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. RWY is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Figure 360: 낙하산 연구 각주

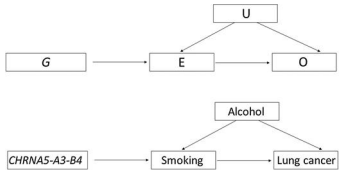


Figure 361: 멘델리안 무작위배정

수 있어 신약 개발의 효율성을 높이는 데 기여하고 있다.

이러한 방법론적 발전과 함께 MR은 근거 기반 의학에서 중요한 도구로 자리잡고 있으며, 특히 관찰 연구의 한계를 보완하고 임상시험 설계에 중요한 정보를 제공하는 역할을 하고 있다.

18.7 결론

컴퓨터 기반 신약 설계와 임상시험은 현대 의학 발전의 핵심 동력이다. 인공지능과 기계학습 기술의 발전은 전통적인 신약 개발 과정을 혁신하고 있으며, 보다 효율적이고 비용 효과적인 신약 발견을 가능하게 하고 있다. 가상 스크리닝, 분자 도킹, ADMET 예측 등의 계산 기법들은 실험적 접근법과 상호 보완적으로 작용하여 신약 개발의 성공률을 높이고 개발 기간을 단축시키고 있다. 임상시험 분야에서도 엄격한 과학적 방법론과 통계적 분석을 통해 새로운 치료법의 안전성과 효능을 입증하는 체계가 지속적으로 발전하고 있다. 무작위 대조 임상시험을 비롯한 다양한 연구 설계 방법론들은 근거 기반 의학의 토대를 제공하며, 환자에게 최선의 치료를 제공하기 위한 과학적 근거를 마련한다. 앞으로 AI 기술과 디지털 헬스 기술의 발전, 실세계 증거의 활용, 개인 맞춤형 의학의 구현 등이 신약 개발과 임상시험 분야에 더욱 혁신적인 변화를 가져올 것으로 기대된다. 이러한 기술적 진보와 함께 국제적 협력과 데이터 공유를 통해 인류의 건강 증진에 기여하는 보다 효과적이고 안전한 의약품 개발이 가능할 것이다.

19 리눅스

19.1 리눅스의 중요성

대부분의 생명정보학 도구는 리눅스용으로 출시된다. 따라서 의생명정보학 연구를 위해서는 리눅스 환경에서의 작업이 필수적이다. 이는 리눅스 환경이 고성능 컴퓨팅과 대용량 데이터 처리에 최적화되어 있기 때문이며, 주로 다음과 같은 이유로 생명정보학 분야에서 널리 사용된다:

1. 다양한 오픈소스 도구의 가용성
2. 명령줄 인터페이스를 통한 효율적인 데이터 처리
3. 분산 컴퓨팅 환경과의 호환성
4. 배치 처리를 통한 대규모 분석 자동화 가능성

생명정보학을 공부하고 실무에 적용하기 위해서는 리눅스를 공부하는 것이 매우 중요하다.

19.2 리눅스의 역사와 철학

19.2.1 자유 소프트웨어 재단 (Free Software Foundation, FSF)

자유 소프트웨어 재단은 1985년 리처드 스톨만(Richard Stallman, 일명 “RMS”)에 의해 창설되었다. 이 재단의 핵심 철학은 “자유(Free) 소프트웨어”의 개념이다. 여기서 말하는 “자유”란 누구나 자유롭게 소프트웨어를 공부하고, 변경하고, 배포할 수 있도록 그 소스 코드가 공개된 소프트웨어를 의미한다.

이는 소프트웨어를 무료로 제공해야 한다는 의미가 아니다. 리처드 스톨만의 유명한 말에 따르면 “자유 소프트웨어는 언론의 자유와 같은 자유를 의미하는 것이지, 무료 맥주를 의미하는 것이 아니다(Free in the sense of free speech, not free beer).”

자유 소프트웨어 재단은 일반 공중 사용 허가서(General Public License, GPL)를 통해 자유 소프트웨어의 권리를 보호한다. GPL은 자유 소프트웨어를 사용하여 개발한 소프트웨어도 반드시 자유 소프트웨어가 되어야 한다는 조건을 포함한다.

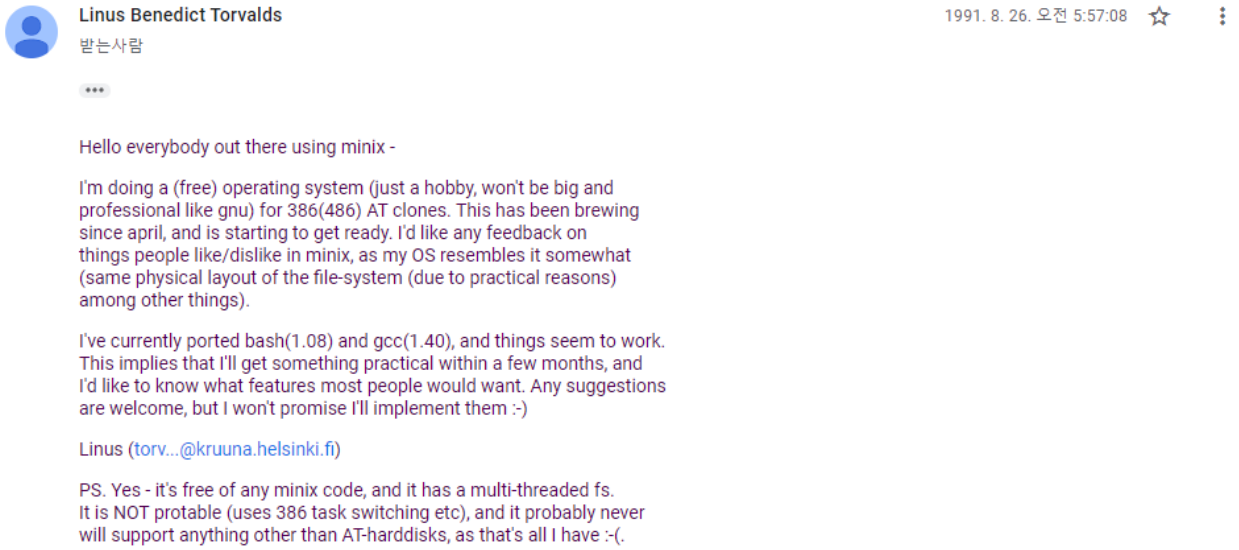
19.2.2 GNU 프로젝트

GNU(GNU is not UNIX!)는 자유 소프트웨어 재단의 주요 프로젝트 중 하나이다. 이 프로젝트의 목표는 당시 가장 인기 있었던 운영체제인 유닉스(UNIX)의 “클론”을 제작하는 것이었다. 유닉스는 커널(Kernel)과 다양한 유틸리티로 구성되어 있었다.

리처드 스톨만은 이를 위해 수많은 유틸리티를 개발했다: - Emacs (텍스트 에디터) - GCC (GNU Compiler Collection, C 컴파일러) - GDB (디버거) - GNU make (빌드 자동화 도구) 등 다양한 도구들이 개발되었다.

19.2.3 리눅스 커널의 탄생

한편, 지구 반대편 핀란드에서는 리누스 토발즈(Linus Torvalds)가 “재미로(just for fun)” 운영체제 커널을 개발하고 있었다. 이 커널은 후에 GNU 프로젝트의 일부로 편입되게 된다.



Linus Benedict Torvalds
받는사람
1991. 8. 26. 오전 5:57:08

Hello everybody out there using minix -

I'm doing a (free) operating system (just a hobby, won't be big and professional like gnu) for 386(486) AT clones. This has been brewing since april, and is starting to get ready. I'd like any feedback on things people like/dislike in minix, as my OS resembles it somewhat (same physical layout of the file-system (due to practical reasons) among other things).

I've currently ported bash(1.08) and gcc(1.40), and things seem to work. This implies that I'll get something practical within a few months, and I'd like to know what features most people would want. Any suggestions are welcome, but I won't promise I'll implement them :-)

Linus (torv...@kruuna.helsinki.fi)

PS. Yes - it's free of any minix code, and it has a multi-threaded fs. It is NOT protable (uses 386 task switching etc), and it probably never will support anything other than AT-harddisks, as that's all I have :-(.

Figure 362: 리눅스 발표

Figure 18.1 리누스 토발즈가 리눅스를 처음 발표한 역사적 순간 (1991년)

19.2.4 GNU/Linux 시스템

GNU 프로젝트의 유틸리티와 리눅스 커널의 결합으로 GNU/Linux 시스템이 탄생했다. 이 시스템은 유닉스와 완전히 호환되는 POSIX 준수 시스템으로, 현재 99% 이상의 서버 시스템이 리눅스를 채택하고 있다.

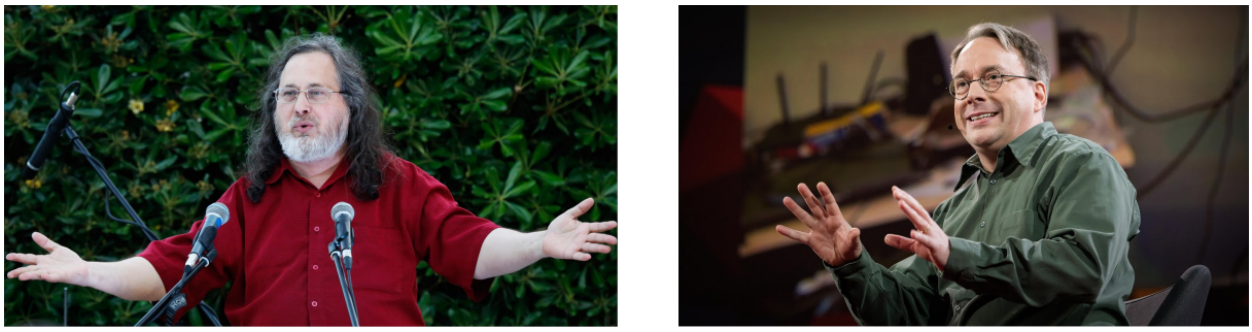


Figure 363: 스톨만과 토발즈

Figure 18.2 리처드 스톨만과 리누스 토발즈

현재는 이 시스템을 간단히 “리눅스”라고 부르는 것이 일반적이다. 이는 시스템에 포함된 모든 소프트웨어가 GPL을 따르는 것은 아니기 때문이다. 물론 리처드 스톨만은 리눅스라는 명칭보다는 GNU/Linux라는 명칭을 선호한다.

19.3 리눅스 배포판

19.3.1 배포판의 개념

리눅스 배포판은 각자의 취향과 목적에 맞는 소프트웨어를 기본으로 포함하는 리눅스 운영체제이다. 다양한 배포판이 존재하며, 이들은 웹사이트 https://distrowatch.com/ 에서 확인할 수 있다.

19.3.2 주요 배포판

가장 널리 사용되는 두 카테고리의 배포판은 다음과 같다: - 데비안 계열: 우분투 리눅스(Ubuntu Linux)가 대표적이다. - 레드햇 계열: CentOS가 대표적이다.

19.3.3 패키지 매니저

배포판의 주요 차이점 중 하나는 패키지 매니저이다: - 데비안 계열: APT (Advanced Package Tool) - 레드햇 계열: YUM (Yellowdog Updater Modified), 최근에는 DNF (Dandified YUM)

이들은 비슷한 활용법을 가지고 있다: - `apt install` (데비안 계열) - `yum install` 또는 `dnf install` (레드햇 계열)

패키지 매니저는 “루트(root)” 권한이 있는 사용자만 활용할 수 있다.

19.4 리눅스 기본 개념

19.4.1 루트 (Root)

리눅스에서 “루트”는 두 가지 의미를 가진다: 1. 시스템 관리자 (슈퍼유저) 2. 디렉토리 구조에서 가장 첫 번째 노드 (/)

19.4.2 디렉토리

디렉토리는 파일 및 다른 디렉토리를 포함할 수 있는 논리적 공간을 의미한다. 윈도우의 “폴더”와 같은 개념이다.

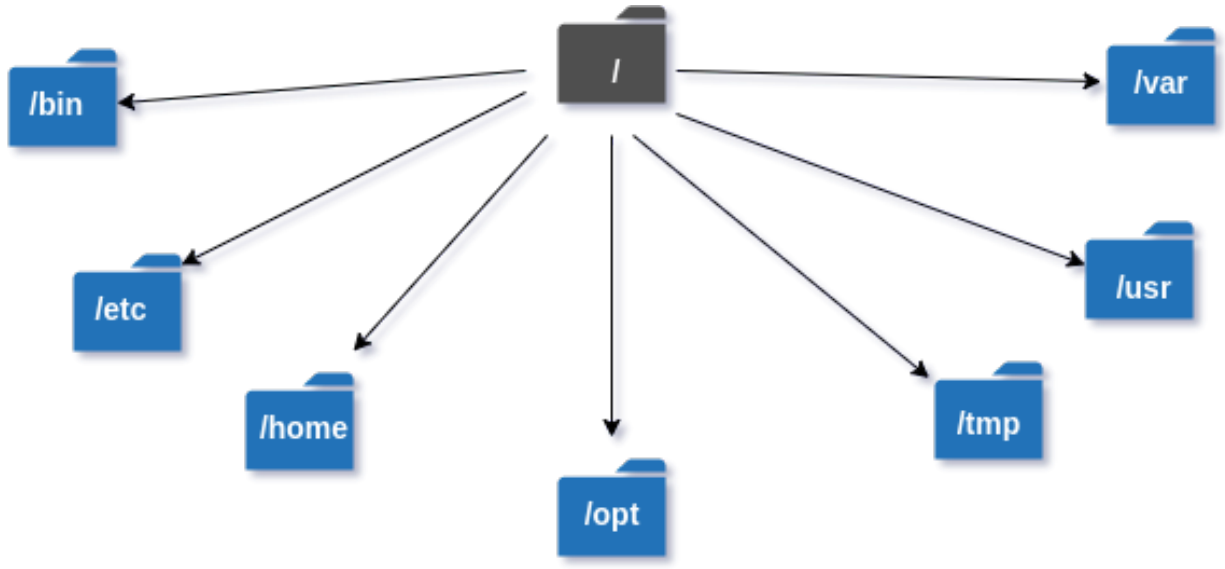


Figure 364: 파일 시스템 경로

Figure 18.3 리눅스 파일 시스템의 경로 구조와 절대경로/상대경로

19.4.3 쉘 (Shell)

쉘은 커널과 사용자 간의 인터페이스 역할을 한다. 주로 CUI(Command-Line User Interface) 기반으로, 키보드를 통해 명령어를 입력하고 화면에 그 결과를 확인할 수 있다.

주요 쉘의 종류: - sh: Bourne shell (UNIX 표준 쉘) - bash: GNU Bourne-Again Shell (GNU 프로젝트 표준 쉘) - csh, zsh, fish 등: 편의기능이 추가된 쉘

19.5 가상 환경 구축: Conda

Conda는 프로젝트별로 분리된 가상 환경을 구축할 수 있게 해주는 도구이다. 이는 서로 다른 프로젝트에서 서로 다른 버전의 소프트웨어를 사용해야 할 때 특히 유용하다.

예를 들어, `project1`에서는 Python 3.4가 필요하고, `project2`에서는 Python 3.10이 필요한 경우, Conda를 사용하면 두 프로젝트의 환경을 분리하여 서로 다른 버전의 파이썬을 설치할 수 있다.

19.5.1 Conda 설치

Conda는 처리 속도가 느린 편이므로, 더 빠른 속도로 동작하는 Conda의 클론인 Micromamba를 설치하는 것을 추천한다. Micromamba는 https://mamba.readthedocs.io/en/latest/user_guide/micromamba.html 에서 설치할 수 있다.

19.5.2 가상환경 생성 예제

다음은 두 가지 가상환경을 생성하고, 서로 다른 버전의 Python을 설치하는 예제이다:

```
# project1     (Python 3.9)
$ micromamba create -n project1 python=3.9
$ micromamba activate project1
(project1) $ python --version
Python 3.9.x

#
(project1) $ micromamba install -c conda-forge numpy pandas matplotlib
(project1) $ python -c "import numpy; print(numpy.__version__)"
1.xx.x

#
(project1) $ micromamba deactivate

# project2     (Python 3.13)
$ micromamba create -n project2 python=3.13
$ micromamba activate project2
(project2) $ python --version
Python 3.13.x

#        (    ,    )
(project2) $ micromamba install -c conda-forge numpy pandas matplotlib
(project2) $ python -c "import numpy; print(numpy.__version__)"
2.xx.x

#
(project2) $ micromamba deactivate

#
$ micromamba env list
```

이 예제에서는 micromamba를 사용하여 두 개의 독립된 가상환경(project1과 project2)을 생성하고, 각각 Python 3.9와 Python 3.13을 설치했다. 각 환경에서는 동일한 패키지(numpy, pandas, matplotlib)를 설치했지만, Python 버전이 다르기 때문에 패키지 버전도 다를 수 있다. 각 환경은 완전히 독립적이므로, 한 환경에서의 변경이 다른 환경에 영향을 주지 않는다. 이는 서로 다른 버전의 패키지가 필요한 여러 프로젝트를 동시에 진행할 때 매우 유용하다.

19.6 생명정보학 문제 해결: Rosalind

Rosalind는 생명정보학 문제풀이 웹사이트(https://rosalind.info/)로, 생명정보학을 코딩테스트처럼 공부할 수 있게 해준다. 생명정보학 개념을 실습을 통해 배우고 싶다면 Rosalind의 문제를 풀어보는 것이 좋은 방법이다.

19.7 리눅스 필수 명령어

19.7.1 파일 시스템 탐색

- **pwd** (Print Working Directory): 현재 디렉토리 경로 확인
  - 경로가 /로 시작하는 경우 "절대경로", 그렇지 않은 경우 "상대경로"

Figure 365: pwd 명령어

Figure 18.4 pwd 명령어를 통한 현재 작업 디렉토리 확인

- **cd** (Change Directory): 디렉토리 변경
  - 탭키를 활용하여 긴 경로를 쉽게 입력할 수 있다.
- **ls** (List): 디렉토리 내용 표시

19.7.2 파일 및 디렉토리 관리

- **chmod** (Change Mode): 파일/디렉토리 권한 변경
- **groups**: 현재 내가 소속된 그룹 확인
- **cp** (Copy): 파일/디렉토리 복사
  - 디렉토리 복사 시에는 -r 옵션이 필요하다.
  - `cp { }`
  - `cp -r`
- **mv** (Move): 파일/디렉토리 이동 또는 이름 변경
  - `mv { }`
- **mkdir** (Make Directory): 디렉토리 생성
  - `mkdir`
- **rmdir** (Remove Directory): 디렉토리 삭제 (비어있는 디렉토리만 삭제 가능)
  - `rmdir`
- **rm** (Remove): 파일/디렉토리 삭제
  - 디렉토리 삭제를 위해서는 -r 옵션을 사용해야 한다.
  - `rm`
  - `rm -r`
  - 한번 삭제된 파일은 복구할 수 없으므로 신중히 사용해야 한다.

19.7.3 시스템 모니터링

- **htop**: 현재 실행 중인 프로세스 확인

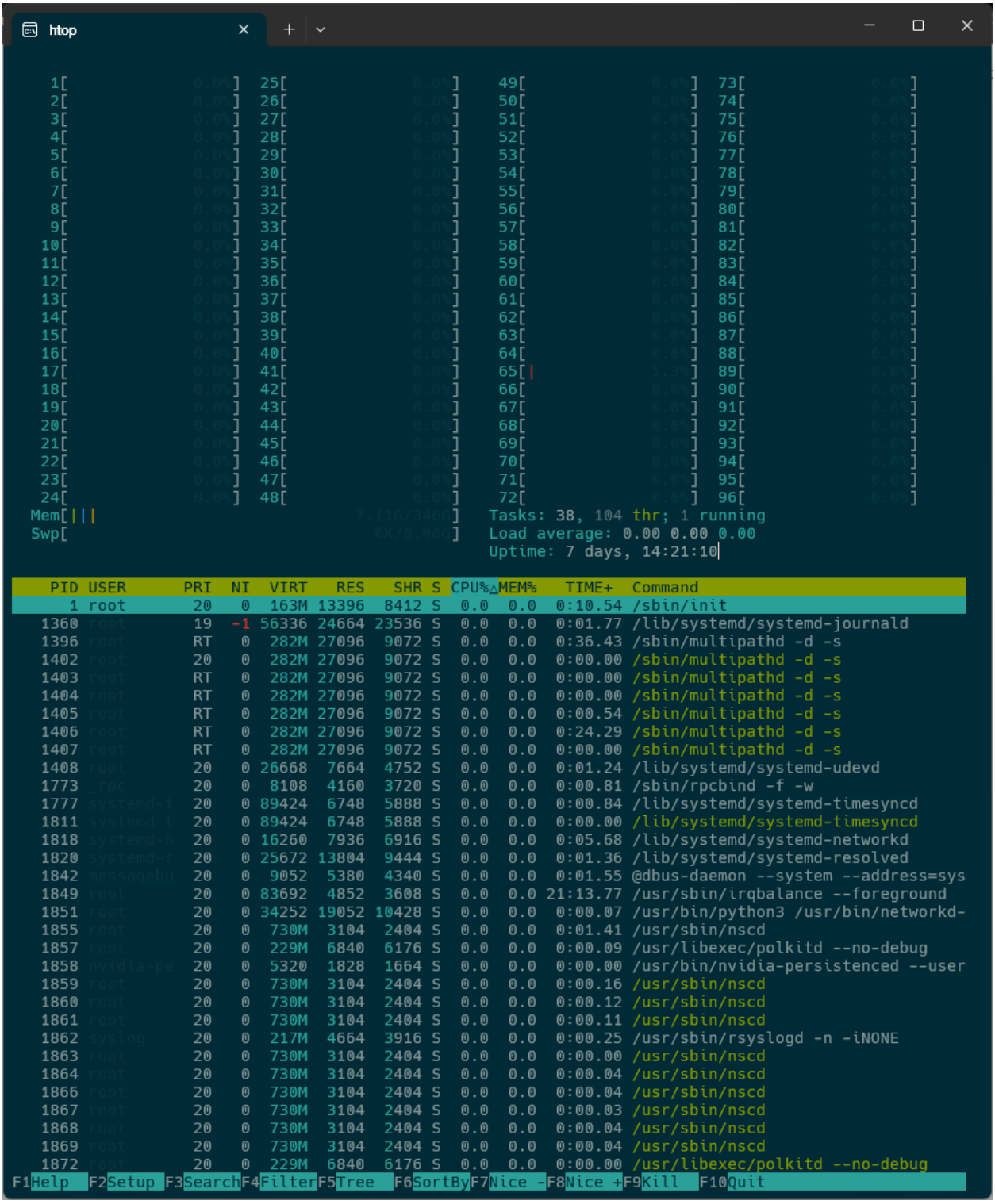


Figure 366: htop 화면

Figure 18.5 htop을 이용한 시스템 리소스 모니터링

19.7.4 파일 조작

- **cat / less**: 파일 내용 확인
- **zcat / zless**: gz 압축된 파일 내용 확인
- **nano**: 간단한 텍스트 편집기

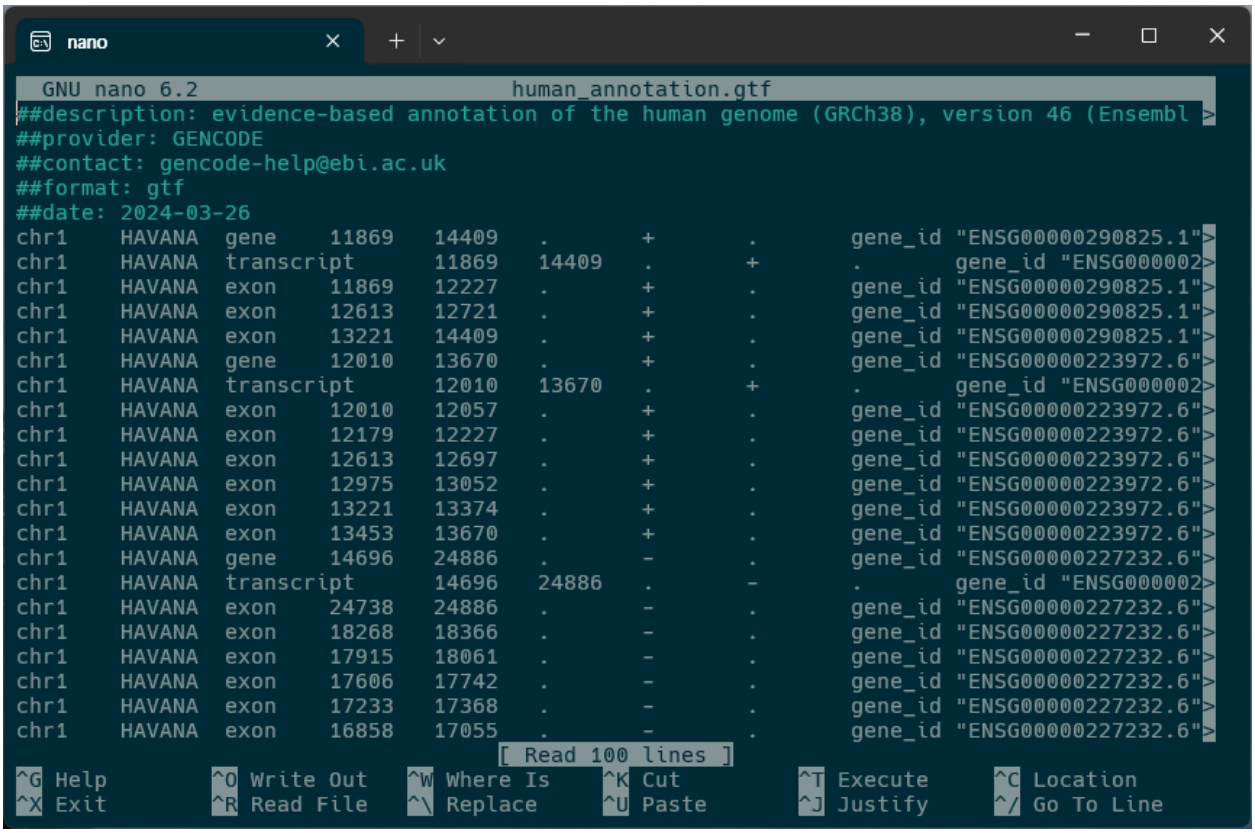


Figure 367: nano 에디터

Figure 18.6 nano 텍스트 에디터의 인터페이스와 기본 사용법

- **wget**: 인터넷에서 파일 다운로드

19.8 생명정보학에서의 리눅스 활용

리눅스는 생명정보학에서 필수적인 도구이다. 대부분의 생명정보학 소프트웨어와 데이터베이스는 리눅스 환경에서 개발되고 운영된다. 또한, 대용량 데이터 처리, 병렬 컴퓨팅, 자동화된 워크플로우 구축 등 생명정보학 연구에 필요한 다양한 작업을 리눅스 환경에서 효율적으로 수행할 수 있다.

리눅스의 명령줄 인터페이스(CLI)는 반복적인 작업을 자동화하고, 복잡한 데이터 처리 파이프라인을 구축하는 데 유용하다. 또한, 서버 환경에서의 원격 작업 및 배치 작업 실행에도 리눅스의 기능이 필수적이다.