생물정보학 치트시트

서열 데이터 형식

- FASTA (.fa, .fasta): 서열 이름(>header)과 서열 문자열로 구성된 기본 형식.
- FASTQ (.fq, .fastq): FASTA에 각 염기의 품질 점수(Phred score)를 추가한 형식. NGS 데이터의 표준.
- SAM/BAM (.sam, .bam): 서열 정렬(alignment) 정보를 저장하는 형식. SAM은 텍스트, BAM은 이진(binary) 형식.
- VCF (.vcf): 유전 변이(variant) 정보를 저장하는 형식.
- BED (.bed): 유전체 구간(genomic intervals) 정보를 저장하는 형식.
- GFF/GTF (.gff, .gtf): 유전자 및 다른 유전체 특징(feature)의 위치와 구조를 기술하는 형식.

서열 정렬 (Sequence Alignment)

- BLAST (Basic Local Alignment Search Tool): 주어진 서열과 데이터베이스 간의 지역적 유사성을 찾는 데 사용되는 가장 기본적인 도구.
  - `blastn`: 뉴클레오티드 서열 검색
  - `blastp`: 단백질 서열 검색
  - `blastx`: 번역된 뉴클레오티드 서열을 단백질 데이터베이스에서 검색
- Bowtie2 / BWA (Burrows-Wheeler Aligner): 짧은 NGS 리드(read)를 참조 유전체(reference genome)에 빠르고 정확하게 정렬하는 데 널리 사용됨.
  - `bowtie2-build <ref.fa> <index_name>`: 참조 유전체 인덱싱
  - `bowtie2 -x <index_name> -U <reads.fq> -S <output.sam>`: 정렬 수행

SAM/BAM 파일 처리

- Samtools: SAM/BAM 파일을 조작하고 분석하기 위한 필수 도구 모음.
  - `samtools view -bS in.sam > out.bam`: SAM을 BAM으로 변환
  - `samtools sort in.bam -o sorted.bam`: BAM 파일 정렬
  - `samtools index sorted.bam`: 정렬된 BAM 파일 인덱싱 (빠른 접근을 위해 필요)
  - `samtools flagstat in.bam`: 정렬 통계 확인
  - `samtools depth`: 특정 위치의 커버리지(coverage) 계산

변이 분석 (Variant Calling)

- GATK (Genome Analysis Toolkit): 변이 분석을 위한 산업 표준 도구. Best Practices 워크플로우를 따르는 것이 중요.
  - 주요 단계: 중복 리드 제거 -> Indel 재정렬 -> 염기 품질 점수 재보정 -> 변이 호출(HaplotypeCaller) -> 변이 필터링
- BCFtools: Samtools와 함께 사용되며, 변이 호출 및 필터링을 위한 가볍고 빠른 대안.
  - `bcftools mpileup -f ref.fa sorted.bam | bcftools call -mv -o variants.vcf`

유전체 구간 분석

- BEDTools: BED, GFF, VCF 파일 등 유전체 구간 데이터를 비교, 조작, 분석하는 강력한 도구.
  - `bedtools intersect`: 두 파일 간의 겹치는 구간 찾기
  - `bedtools merge`: 겹치는 구간을 하나로 합치기
  - `bedtools closest`: 한 파일의 구간에 가장 가까운 다른 파일의 구간 찾기
  - `bedtools coverage`: 한 파일의 구간이 다른 파일의 구간에 의해 얼마나 커버되는지 계산

RNA-Seq 분석

- 전처리: `FastQC` (품질 관리), `Trimmomatic` (어댑터 및 품질 트리밍)
- 정렬: `STAR`, `HISAT2` (Splicing을 고려한 정렬)
- 정량화 (Quantification): `featureCounts`, `Salmon`, `Kallisto` (유전자 또는 전사체(transcript) 발현량 계산)
- 차등 발현 분석 (Differential Expression Analysis): `DESeq2`, `edgeR` (R 패키지) (서로 다른 조건 간에 발현량이 유의미하게 차이나는 유전자 찾기)

데이터 시각화

- IGV (Integrative Genomics Viewer): BAM, VCF, BED 등 다양한 형식의 데이터를 유전체 브라우저에서 시각적으로 탐색.
- R / Python: `ggplot2`(R), `matplotlib`/`seaborn`(Python) 등을 사용하여 Volcano plot, MA plot, Heatmap 등 다양한 그래프 생성.