Bioinformatics Cheatsheet

Common Bioinformatics File Formats

- `.fasta`: DNA, RNA, or protein sequence format. Starts with `>`.
- `.fastq`: Sequence data with quality scores. Each entry uses 4 lines.
- `.sam` / `.bam`: Aligned sequence data (Sequence Alignment Map). `.bam` is binary.
- `.vcf`: Variant Call Format. Stores genetic variation info.
- `.gff` / `.gtf`: Gene annotation formats.
- `.bed`: Genome region coordinates (start/end).

Essential CLI Tools

- `samtools`: Manipulation of SAM/BAM files (view, sort, index, flagstat).
- `bcftools`: Manipulation of VCF files and variant calling.
- `bedtools`: Intersection, merging, and windows of genomic features.
- `seqkit`: Rapid manipulation of FASTA/Q files (stats, search, replace).
- `fastp`: Fast all-in-one preprocessing for FASTQ files (QC, filtering).
- `bwa` / `bowtie2`: DNA sequence alignment to a reference genome.
- `star`: Fast RNA-seq alignment.

Python for Bioinformatics (`Biopython`)

```python
from Bio import SeqIO

# Read a FASTA file
for record in SeqIO.parse("input.fasta", "fasta"):
    print(record.id)
    print(record.seq)
    print(len(record))
```

R for Bioinformatics (`Bioconductor`)

- `DESeq2`: Differential expression analysis for RNA-seq.
- `Seurat`: Single-cell RNA-seq analysis.
- `GenomicRanges`: Represent and manipulate genomic intervals.
- `biomaRt`: Interface to BioMart databases (e.g., Ensembl).

Databases

- `NCBI / GenBank`: Comprehensive genomic database.
- `Ensembl`: Genome browser and annotation database.
- `PDB (Protein Data Bank)`: 3D structures of proteins and nucleic acids.
- `UniProt`: Protein sequence and functional information.

Workflow Management

- `Snakemake`: Python-based workflow management system.
- `Nextflow`: Domain-specific language for data-driven pipelines.

Sequence Alignment Concepts

- `Identity`: Percentage of exact matches between sequences.
- `Similarity`: Percentage of matches considering physical-chemical properties (for proteins).
- `Gap Penalty`: Score penalty for opening or extending gaps in alignment.
- `Expect (E) Value`: The number of hits expected by chance (lower is more significant).

Pro Tips

Index Everything

Most genomics files (.bam, .vcf, .fasta) require index files (.bai, .tbi, .fai) for fast random access. Use `samtools index`, `bcftools index`, and `samtools faidx`.

Use Compression

Genomic data is huge. Use `.gz` tools like `bgzip` and work with compressed formats directly where possible.

Conda / Pixi / Flox

Isolate bioinformatics environments to avoid version conflicts between complex dependencies. Use `pixi` or `mamba` for high-speed installation.