

VU Learning HUB

BIF502 Quiz File

Chapter-wise Solved MCQs

vulearninghub.xyz

Unit 1: Genomes and Data

Q1. What percentage of the human genome consists of protein-coding genes?

- A) ~50%
- B) ~20%
- C) ~1.5%
- D) ~10%

Q2. Which sequencing generation produces long reads (10 kb to 1 Mb+) capable of resolving complex genomic repeats?

- A) First generation (Sanger sequencing)
- B) Second generation (Illumina NGS)
- C) Third generation (PacBio and Oxford Nanopore)
- D) Microarray-based sequencing

Q3. Which sequence file format includes per-base Phred quality scores alongside sequence data?

- A) FASTA
- B) FASTQ
- C) VCF
- D) BAM

Q4. What is the primary advantage of using a pangenome graph over a single traditional reference genome?

- A) It eliminates the need for sequence alignment algorithms.
- B) It represents variation across diverse populations and reduces reference bias.
- C) It replaces FASTQ files with VCF files.
- D) It converts protein sequences directly back into raw DNA reads.

Answer Key

Q1: C — Protein-coding genes make up only ~1.5% of the human genome, while non-coding and repetitive elements dominate the remaining space

Q2: C — Third-generation sequencing technologies (PacBio and Oxford Nanopore) generate long reads (10 kb-1 Mb+) that resolve repetitive regions and enable complete assemblies

Q3: B — Unlike standard FASTA files, FASTQ files store per-base Phred quality scores to quantify base-calling reliability

Q4: B — Pangenomes use graph structures to incorporate genomic variation from multiple individuals, overcoming single-reference ancestral bias

Unit 2: Sequence Alignment

Q5. How does the Smith-Waterman local alignment algorithm differ from the Needleman-Wunsch global alignment algorithm in its dynamic programming matrix calculation?

- A) Smith-Waterman floors cell scores at 0 and tracebacks from the maximum score in the matrix.
- B) Smith-Waterman does not allow gap penalties.
- C) Smith-Waterman runs in linear $O(n)$ time complexity.
- D) Smith-Waterman only aligns protein sequences using BLOSUM matrices.

Q6. What does a low BLAST Expect value (such as $E = 0.001$) indicate about an alignment result?

- A) The alignment is likely due to random chance.
- B) The match is statistically significant and unlikely to occur by chance.
- C) The query sequence contains sequencing errors.
- D) The database contains fewer than 1,000 sequences.

Q7. Which indexing strategy is used by read-mapping tools like BWA and Bowtie2 for rapid exact seed lookups?

- A) Hidden Markov Models (HMM)
- B) Burrows-Wheeler Transform (BWT) and FM-index
- C) Neighbor-Joining trees
- D) Dynamic programming matrix traceback

Q8. What is a major limitation of progressive multiple sequence alignment algorithms like Clustal?

- A) They cannot align more than three sequences simultaneously.
- B) Errors made in early pairwise alignments propagate throughout the alignment process.
- C) They require complete whole-genome assemblies as input.
- D) They ignore conserved amino acid positions.

Answer Key

Q5: A — Smith-Waterman floors negative scoring matrix values at 0 and tracebacks from the maximum score anywhere in the matrix to find optimal local alignments

Q6: B — The E-value measures expected chance matches, so a very small E-value (e.g., 0.001) signifies a statistically reliable homology result

Q7: B — Short-read mappers compress and index the reference genome using the Burrows-Wheeler transform and FM-index for ultra-fast seed matching

Q8: B — Progressive alignments follow a fixed guide tree, meaning mistakes made in early pairwise steps cannot be fixed later ("once a gap, always a gap")

Unit 3: Assembly and Annotation

Q9. What does the N50 metric represent in genome assembly evaluation?

- A) The percentage of reads that mapped to the reference genome.
- B) The contig length such that half of the total assembly length is contained in contigs of this size or larger.
- C) The number of single-base variants detected per kilobase.
- D) The average error rate across third-generation sequencing reads.

Q10. How do ab initio gene prediction tools like Genscan and Augustus identify gene structures in unannotated DNA?

- A) By mapping short reads using the Burrows-Wheeler transform.
- B) By using Hidden Markov Models (HMMs) to detect statistical sequence signals such as start/stop codons and splice sites.
- C) By comparing whole genomes using dot plots.
- D) By sequencing RNA directly using Nanopore devices.

Q11. Which type of genetic variation represents single-base changes between a sample sequence and a reference genome?

- A) Single Nucleotide Polymorphism (SNP)
- B) Copy-Number Variant (CNV)
- C) Transposon
- D) Structural Variant (SV)

Q12. Why is read count normalization (such as TPM or DESeq2 size factors) essential in RNA-seq expression analysis?

- A) To convert RNA sequences into protein alignments.
- B) To correct for differences in sequencing depth and gene length across samples.
- C) To remove introns from raw genomic DNA sequences.
- D) To build de Bruijn graphs from transcriptomic reads.

Answer Key

Q9: B — N50 is a contiguity metric where 50% of the assembled genome sequence is contained in contigs equal to or longer than the N50 value

Q10: B — Ab initio predictors rely on statistical models (HMMs) trained on intrinsic sequence signals like GT-AG splice junctions and codon bias

Q11: A — A Single Nucleotide Polymorphism (SNP) is a single-base substitution identified during variant calling against a reference genome

Q12: B — Normalization accounts for variations in total library sequencing depth and gene feature lengths so expression levels can be accurately compared

Unit 4: Comparative Genomics and Applications

Q13. In comparative genomics, what does the term "synteny" refer to?

- A) The substitution error rate in short reads.
- B) The conserved order of genes along chromosomes across different species.
- C) The percentage of GC content in non-coding regions.
- D) The assembly metric measuring gene completeness.

Q14. Which statistical technique is widely used in phylogenetic tree building to measure branch support and confidence?

- A) Dynamic programming traceback
- B) Bootstrapping
- C) Burrows-Wheeler transform
- D) Variant Quality Score Recalibration

Q15. What is the key operational difference between 16S rRNA sequencing and shotgun metagenomics in microbiome studies?

- A) 16S rRNA targets a specific marker gene for taxonomic identification, while shotgun metagenomics sequences all community DNA to assess functional capacity.
- B) 16S rRNA is used only for eukaryotic organisms, while shotgun metagenomics is used for viruses.
- C) 16S rRNA produces long reads, while shotgun metagenomics produces only FASTA headers.
- D) 16S rRNA requires cell culturing, while shotgun metagenomics does not.

Q16. Why is traditional anonymisation or de-identification ineffective for protecting human genomic privacy?

- A) Genomic sequence files are too large to encrypt safely.
- B) Genomic data is inherently identifying and can be re-linked to individuals or relatives through genealogical databases.
- C) Reference genomes automatically strip out all non-coding sequence data.
- D) VCF files automatically attach individual names and clinical records.

Answer Key

Q13: B — Synteny describes the conservation of gene order blocks across species, helping track evolutionary rearrangements and duplications

Q14: B — Bootstrapping resamples alignment columns with replacement to evaluate how strongly the sequence data supports each evolutionary clade

Q15: A — 16S rRNA profiling identifies "who is there" via taxonomic marker genes, whereas shotgun metagenomics profiles total DNA to reveal "what they can do"

Q16: B — Human genomes are uniquely identifying; re-identification techniques and forensic genealogy demonstrate that true genomic anonymisation is impossible

End of Quiz

BIF502 — 16 solved MCQs
Prepared by VU Learning HUB
vulearninghub.xyz