

VU LEARNING HUB

vulearninghub.xyz

HANDOUT

BIF604

Special Topics in Bioinformatics

BIF604 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

BIF604

Special Topics in Bioinformatics

Study Handout

Complete lecture-wise notes for Virtual University students

About This Handout

This handout covers BIF604 Special Topics in Bioinformatics -- the field's moving frontier: NGS, phylogenomics, structural prediction, metagenomics, and machine learning. The exam rewards pipeline-tracing, tool-choice justification, and critical evaluation. After each lesson, close the handout and trace one pipeline from memory -- bioinformatics is learned by piping, not just reading.

Table of Contents

Lesson 1	Bioinformatics: The Frontier Sampler
Lesson 2	NGS and RNA-seq Pipelines
Lesson 3	Alignment and Phylogenetics
Lesson 4	Structural Bioinformatics and Drug Design
Lesson 5	Metagenomics and Microbiomes
Lesson 6	Machine Learning for Biology
Lesson 7	Genome Assembly and Pangenomes
Lesson 8	Single-Cell Genomics
Lesson 9	Systems Biology and Networks
Lesson 10	Population Genomics and GWAS
Lesson 11	Databases and Data Stewardship
Lesson 12	Reproducible Bioinformatics
Lesson 13	Careers and the Pakistani Scene
Lesson 14	Revision: Bioinformatics Master Drill

Lesson 1: Bioinformatics: The Frontier Sampler

Learning Objectives

- Define bioinformatics
- Map data to questions
- Navigate key databases
- Plan the skill path

Bioinformatics: biology's data science -- using computation to understand biological information (DNA, RNA, proteins, systems). Why 'special topics': the field moves too fast for a fixed syllabus -- this course samples the frontier: NGS analysis, structural bioinformatics, phylogenomics, metagenomics, and machine learning for biology. The stance: the biologist who can compute, and the computer scientist who understands biology -- both win.

The data deluge: one human genome = 3 billion bases; GenBank doubles every ~18 months; the PDB, UniProt, and SRA as the public commons. The central dogma as data flow (DNA -> RNA -> protein -> function), and the bioinformatician's question at each level (what's there? what's expressed? what does it do? how does it interact?). Tools: BLAST (the search engine of biology), alignment, and the command line -- the lab bench of the computational biologist.

Skills map: sequence analysis, databases, statistics, and scripting (Python/Biopython, R/Bioconductor). The exam's orientation question: 'define bioinformatics and its scope' -- answer with data types, questions, and tools.

Key Terms & Definitions

- Central dogma as data flow — DNA->RNA->protein->function questions.
- BLAST — Sequence similarity search.
- GenBank / PDB / UniProt — The public data commons.
- Biopython / Bioconductor — Python/R bio toolkits.

Example / Illustration

- A mysterious wheat gene: BLAST it (what's it like?), align homologs (what's conserved?), model the protein (what's its shape?), check expression (when is it on?) -- the bioinformatics workflow in one gene.

Lesson Summary

- Biology's data science; the frontier sampler.
- Data deluge: compute or drown.
- BLAST, align, model, express.

Exam Pointers — Very Important

- ★ Define bioinformatics and map data types to questions -- the orientation staple.

Lesson 2: NGS and RNA-seq Pipelines

Learning Objectives

- Explain NGS technologies
- Run QC-to-variant pipelines
- Analyze RNA-seq
- Apply to Pakistani problems

NGS: millions of short reads per run (Illumina) or long reads (PacBio, Oxford Nanopore) -- the technology that made the \$1000 genome real. The analysis pipeline: quality control (FastQC -- trim the bad), alignment/mapping (BWA, Bowtie2 -- reads to reference), variant calling (GATK -- SNPs and indels), and annotation (what do the variants do? -- VEP). Each step's file formats: FASTQ, SAM/BAM, VCF -- the alphabet.

RNA-seq: counting transcripts -- alignment (STAR/HISAT2), quantification (featureCounts, Salmon), differential expression (DESeq2 -- the negative-binomial workhorse; fold change + adjusted p-value), and pathway analysis (GO enrichment -- what biology changed?). The exam's pipeline question: 'trace an RNA-seq analysis' -- tools and decisions at each step.

Pakistani relevance: wheat genomics (rust resistance!), cotton (leaf curl virus), the Indus dolphin's genome (conservation genomics), and COVID variant tracking (the NIH's sequencing -- public health bioinformatics). Pitfalls: batch effects, reference bias, and the 'p-value fishing' trap -- statistics with discipline.

Key Terms & Definitions

- FASTQ / BAM / VCF — Reads / alignments / variants.
- DESeq2 — Differential expression workhorse.
- GO enrichment — What biology changed.
- Batch effects — The confounder to control.

Example / Illustration

- Drought-stressed wheat: RNA-seq control vs stressed; DESeq2 finds 2,000 DE genes; GO enrichment: 'response to water deprivation' -- from reads to biology in four steps.

Lesson Summary

- QC, map, call, annotate -- the pipeline.
- DESeq2: fold change + adjusted p.
- Batch effects: control or perish.

Exam Pointers — Very Important

- ★ Trace an RNA-seq pipeline with tools -- the NGS staple.

Lesson 3: Alignment and Phylogenetics

Learning Objectives

- Align pairwise and multiple
- Choose BLAST wisely
- Build and read trees
- Apply to outbreaks

Pairwise alignment: Needleman-Wunsch (global) vs Smith-Waterman (local) -- dynamic programming's gift to biology; scoring matrices (BLOSUM62 -- the default wisdom: log-odds of substitutions) and gap penalties. BLAST: the heuristic shortcut -- seeds, extends, E-value (the expect value -- $1e-50$ means real, 10 means noise); know when BLAST fails (remote homologs -- use PSI-BLAST/HMMER).

Multiple sequence alignment (MSA): Clustal Omega, MUSCLE, MAFFT -- aligning families to find conserved domains (the active site's column!); alignment quality dictates everything downstream (garbage in, garbage tree). Phylogenetics: distance (neighbor-joining -- fast), parsimony, maximum likelihood (the model-based standard -- IQ-TREE/RAXML), Bayesian (MrBayes -- posteriors); bootstrap support (the confidence number); outgroups root trees.

Applications: SARS-CoV-2 variant tracking (the tree that guided policy!), the wheat rust's origin (phylogeography), and protein function prediction (conserved = important). The exam: 'align, tree, interpret' -- the phylogenetics practical.

Key Terms & Definitions

- BLOSUM62 — The default scoring matrix.
- E-value — BLAST's significance; $1e-50$ vs 10.
- Bootstrap — Tree confidence; >70 credible.
- Outgroup — Roots the tree.

Example / Illustration

- COVID tree: Omicron's long branch (many mutations), bootstrap 98 on its node (confident), rooted with Wuhan-Hu-1 (outgroup) -- reading the pandemic's evolution.

Lesson Summary

- Global vs local; E-value judges hits.
- Alignment quality makes or breaks trees.
- Bootstrap: the confidence number.

Exam Pointers — Very Important

- ★ Interpret a phylogenetic tree with bootstraps -- the phylogenetics practical.

Lesson 4: Structural Bioinformatics and Drug Design

Learning Objectives

- Explain the folding problem
- Use AlphaFold-era tools
- Design docking studies
- Interpret variants

Protein structure: primary to quaternary; the folding problem (Anfinsen's dogma -- sequence determines structure); PDB (the structure bank); visualization (PyMOL -- the molecular microscope). Prediction revolution: AlphaFold2 (2020-21 -- CASP14's shock; near-experimental accuracy) and RoseTTAFold -- the frontier topic of the decade; ESMFold for speed. What changed: structure for every protein, instantly -- drug design and variant interpretation transformed.

Docking and drug design: molecular docking (AutoDock -- ligand into binding pocket; scoring functions rank poses), virtual screening (millions of compounds, computationally), and the Pakistani angle (medicinal plants -- docking their compounds against targets; the traditional-medicine-to-lead pipeline). Limitations: docking scores aren't binding energies; dynamics need MD simulations (GROMACS -- the movie of the molecule).

Protein function: domains (Pfam), motifs, and the structure-function link (the active site's geometry = chemistry). Variant interpretation: missense SNP in a cancer gene -- benign or pathogenic? (SIFT/PolyPhen + AlphaFold context). The exam: 'design a docking study' -- target, library, protocol, validation.

Key Terms & Definitions

- AlphaFold2 — The prediction revolution (2021).
- Docking — Ligand into pocket; AutoDock.
- MD simulations — GROMACS; the molecular movie.
- Pfam — Domain database.

Example / Illustration

- Neem compounds vs dengue NS2B-NS3 protease: dock 50 limonoids (AutoDock), rank by score, inspect top poses (H-bonds to catalytic triad?), validate with known inhibitors -- the in-silico screen.

Lesson Summary

- AlphaFold: structure for every protein.
- Dock, score, inspect, validate.
- Dynamics need MD; scores aren't energies.

Exam Pointers — Very Important

- ★ Design a docking-based drug screen -- the structural capstone.

Lesson 5: Metagenomics and Microbiomes

Learning Objectives

- Explain metagenomics
- Run community pipelines
- Apply to health/agriculture
- Handle compositional stats

Metagenomics: sequencing whole communities (soil, gut, ocean) -- who is there (16S/ITS amplicons; shotgun for function), bypassing culture (99% of microbes won't grow on plates!). Pipeline: QC, assembly (metaSPAdes/MEGAHIT -- the jigsaw), binning (MAGs -- metagenome-assembled genomes), taxonomic profiling (Kraken2), functional annotation (KEGG pathways). The exam's pipeline: 'trace a gut-microbiome study.'

Applications: human gut (diet-disease links; the Pakistani gut -- under-studied!), soil (wheat rhizosphere -- the yield microbiome), and wastewater surveillance (polio/COVID tracking -- public-health metagenomics; Pakistan's polio program uses it!). eDNA (ZOO620's link!): biodiversity from water samples -- the same sequencing, ecological questions.

Challenges: contamination (the kitome!), database bias (reference genomes skew Western), and the correlation-causation trap (dysbiosis claims need experiments). Statistics: compositional data (relative abundances lie -- use CLR transforms), diversity metrics, and PERMANOVA. The frontier: long-read metagenomics (complete MAGs) and the microbiome-therapy race.

Key Terms & Definitions

- MAGs — Metagenome-assembled genomes.
- 16S vs shotgun — Who's there vs what they do.
- Kitome — Reagent contamination trap.
- Compositional data — Relative abundances lie; CLR.

Example / Illustration

- Polio wastewater: Karachi sewage sequenced monthly; poliovirus reads detected -> vaccination campaign targeted -- metagenomics guiding public health in real time.

Lesson Summary

- Sequence communities, not isolates.
- Binning builds genomes from mixtures.
- Contamination and composition: the traps.

Exam Pointers — Very Important

- ★ Trace a metagenomics pipeline -- the community-sequencing staple.

Lesson 6: Machine Learning for Biology

Learning Objectives

- Explain ML paradigms
- Build bio-ML pipelines
- Evaluate critically
- Apply to variants/drugs

Machine learning for biology: the paradigm -- learn patterns from data (features -> labels), not rules from experts. Algorithms: random forests (the robust default), SVMs, neural networks (deep learning -- CNNs for images, transformers for sequences), and the biology-specific (AlphaFold's Evoformer!). Tasks: classification (pathogenic vs benign variants), regression (binding affinity), clustering (cell types in scRNA-seq), and generation (protein design -- the frontier!).

Bio-ML pipeline: data (curated! -- garbage in...), features (k-mers, physicochemical properties, embeddings -- ESM protein language models!), train/validate/test splits (no leakage!), metrics (accuracy lies on imbalanced data -- use AUROC, F1, MCC), and interpretation (SHAP -- which features mattered?). The exam's critique: 'evaluate this ML paper' -- hunt leakage, tiny test sets, and missing baselines.

Applications: variant effect prediction (EVE, AlphaMissense -- the 2023 leap), drug-target interaction, medical imaging (the diabetic-retinopathy model), and protein design (RFdiffusion -- designing proteins from scratch, 2023+). Pakistani angle: the talent

pipeline -- bio-ML as the career frontier; Kaggle + Rosalind as training grounds. Ethics: bias in training data (whose genomes?), and the hype-reality gap.

Key Terms & Definitions

- No leakage — Train/test separation; homologs too.
- AUROC / MCC — Imbalanced-data metrics.
- Embeddings — ESM; protein language models.
- AlphaMissense — 2023 variant-effect leap.

Example / Illustration

- '99% accuracy' variant classifier: test set 90% benign (accuracy lies!) -- demand AUROC/MCC, check leakage (homologs in train AND test?), require a simple baseline (does BLAST beat it?).

Lesson Summary

- Features, splits, metrics, interpretation.
- Accuracy lies; AUROC tells truth.
- Leakage and baselines: the critique kit.

Exam Pointers — Very Important

- ★ Critique a bio-ML paper's evaluation -- the ML-skepticism practical.

Lesson 7: Genome Assembly and Pangenomes

Learning Objectives

- Explain assembly
- Evaluate with N50/BUSCO
- Annotate genes
- Use pangenomes

De novo assembly: building genomes without a reference -- the jigsaw problem (short reads + repeats = the hard case); assemblers (SPAdes for small, Flye/Canu for long reads); metrics (N50 -- the contiguity number; BUSCO -- the completeness check against conserved genes). Long reads (PacBio HiFi, Nanopore): spanning repeats -- the telomere-to-telomere (T2T) human genome (2022) as the triumph; the \$1000 T2T era coming.

Annotation: finding genes -- ab initio predictors (AUGUSTUS), homology (lift-over), and RNA-seq evidence (the three-legged stool); functional annotation (InterPro, GO terms); organelles and repeats (RepeatMasker). The wheat genome's hexaploid nightmare (16 Gb, three subgenomes!) -- why assembly matters for Pakistan's crop. Quality: BUSCO >95% or redo.

Pangenomes: one reference lies -- the species' gene pool (core vs accessory genes); graph genomes (the reference as a network, not a line); human pangenome (2023 -- 47 diverse genomes); crop pangenomes (the resistance genes hiding in wild relatives!). The exam: 'evaluate an assembly' -- N50, BUSCO, then biology.

Key Terms & Definitions

- N50 / BUSCO — Contiguity / completeness metrics.
- T2T — Telomere-to-telomere assembly.
- Pangenome — Core + accessory genes.
- Graph genomes — Reference as network.

Example / Illustration

- New assembly: N50 2 Mb (good contiguity), BUSCO 97% (complete) -- but the resistance-gene cluster is fragmented (repeats!); add HiFi reads, re-scaffold with Hi-C -- the improvement plan.

Lesson Summary

- Long reads solved the repeat problem.
- N50 contiguity; BUSCO completeness.
- One reference per species is obsolete.

Exam Pointers — Very Important

- ★ Evaluate a genome assembly's quality -- the assembly practical.

Lesson 8: Single-Cell Genomics

Learning Objectives

- Explain scRNA-seq
- Run the analysis pipeline
- Interpret clusters
- Handle traps

Single-cell RNA-seq: every cell counted separately (10x Genomics droplets) -- tissues unmixed; the Human Cell Atlas ambition. Pipeline: QC (dead cells out -- mitochondrial %!), normalization (SCTransform), dimensionality reduction (PCA then UMAP -- the beautiful blob plot), clustering (Louvain -- cell types emerge), and marker genes (what defines each cluster?). The exam's pipeline: 'trace scRNA-seq.'

Applications: cancer (tumor heterogeneity -- the resistant clone hiding), immunity (the COVID immune atlas), development (lineage trajectories -- pseudotime!), and plants (root cell types). Spatial transcriptomics: genes mapped on tissue slices (Visium) -- the geography of expression. Pakistani angle: the cost barrier (collaborate!); the analysis skill as the portable asset.

Traps: doublets (two cells, one barcode), batch effects (integrate with Harmony), and over-clustering (is it biology or noise? -- validate with markers). Statistics: the dropout problem (zeros everywhere -- model it, don't fear it). The frontier: single-cell multi-omics (RNA + ATAC + protein -- the cell's full state).

Key Terms & Definitions

- UMAP — The blob plot; clusters = cell types.
- Doublets — Two cells, one barcode; filter.
- Pseudotime — Developmental trajectories.
- Spatial transcriptomics — Genes on tissue geography.

Example / Illustration

- Tumor UMAP: 8 clusters; cluster 5 expresses stem markers (the resistant suspects!); trajectory analysis shows cluster 5 -> cluster 2 (differentiation?) -- from blobs to hypotheses.

Lesson Summary

- Unmix tissues cell by cell.
- QC ruthlessly; integrate batches.
- Validate clusters with markers.

Exam Pointers — Very Important

- ★ Interpret a UMAP with clusters -- the single-cell practical.

Lesson 9: Systems Biology and Networks

Learning Objectives

- Analyze biological networks
- Model systems
- Integrate omics
- Prioritize targets

Networks: biology as graphs -- protein-protein interaction (STRING database), gene regulatory (transcription-factor targets), metabolic (KEGG pathways). Analysis: hubs (the essential proteins -- lethality-centrality rule!), modules (MCODE clusters = complexes/pathways), and enrichment (your gene list's pathways). Cytoscape: the network workbench. The exam: 'analyze a PPI network.'

Systems modeling: from parts to behavior -- ODE models of pathways (the toggle switch!), flux balance analysis (FBA -- genome-scale metabolic models; predict growth knockouts), and Boolean networks (logic of regulation). Applications: drug targets (hub proteins in pathogen networks), metabolic engineering (the biofuel yeast), and the cancer-network rewiring.

Multi-omics integration: genome + transcriptome + proteome + metabolome (the vertical stack); MOFA (factor analysis across omics); the Pakistani crop angle (drought response across omics layers). Philosophy: reductionism's limits -- the network as the unit; but models are maps, not territory (validate experimentally!).

Key Terms & Definitions

- Hubs — Essential network nodes.
- FBA — Flux balance analysis; metabolic prediction.
- Cytoscape — Network visualization/analysis.
- Multi-omics — Vertical data integration.

Example / Illustration

- Malaria PPI: hub proteins essential (lethality-centrality); filter for no human homolog (safety!); check druggability (pocket?) -- network to target list in three filters.

Lesson Summary

- Hubs are essential; modules are pathways.
- FBA predicts metabolic behavior.
- Networks guide; experiments decide.

Exam Pointers — Very Important

- ★ Prioritize drug targets from a pathogen PPI network -- the network practical.

Lesson 10: Population Genomics and GWAS

Learning Objectives

- Analyze population structure
- Run GWAS logic
- Scan for selection
- Handle ethics

Population genomics: variation within species -- SNPs across populations; the data (1000 Genomes; the Pakistani genome projects -- the under-represented frontier!). Statistics: F_{st} (differentiation), Tajima's D (selection's signature), PCA (population structure -- the scatter plot of ancestry), ADMIXTURE (ancestry proportions). The exam: 'interpret a PCA of populations.'

GWAS: genotype-phenotype association -- millions of SNPs, Manhattan plots (the skyline of significance), the genome-wide threshold ($5e-8$), and the replication requirement. Hits to biology: fine-mapping, eQTLs (the SNP that tunes expression). Pakistani relevance: consanguinity studies (recessive disease mapping -- the population's structure as a research asset!), diabetes/cardiovascular GWAS.

Selection scans: iHS, XP-EHH (the long-haplotype trick); classic hits (lactase, sickle cell, altitude adaptation -- Tibetans' EPAS1). Ethics: group-stigmatization risks (report carefully!), data sovereignty (whose genomes, whose benefit? -- the Nagoya Protocol's spirit).

Key Terms & Definitions

- F_{st} / Tajima's D — Differentiation / selection stats.
- Manhattan plot — GWAS skyline; $5e-8$ line.
- eQTL — Expression-tuning variants.
- Consanguinity mapping — Recessive-disease asset.

Example / Illustration

- The lactase peak: chromosome 2's skyscraper ($p=1e-40$); fine-map to the enhancer SNP; eQTL shows it tunes LCT -- from association to mechanism, the full arc.

Lesson Summary

- F_{st} measures divergence; D detects selection.
- GWAS: associate, replicate, fine-map.
- Structure is an asset; report responsibly.

Exam Pointers — Very Important

- ★ Interpret a Manhattan plot -- the GWAS practical.

Lesson 11: Databases and Data Stewardship

Learning Objectives

- Use core databases
- Deposit data
- Apply FAIR
- Query programmatically

The commons: GenBank/ENA/DDBJ (sequences), SRA (raw reads), PDB, UniProt -- the submission duty (publish = deposit!); accession numbers as the currency. FAIR data (ZOO620's link!): findable, accessible, interoperable, reusable -- the exam's acronym again, bioinformatics edition. Metadata: the MIAME/MINSEQE standards (experiments described enough to reuse).

Databases as tools: Ensembl (genome browser -- the locus page), NCBI Taxonomy, KEGG (pathways), Reactome, and the specialist (FlyBase, TAIR for Arabidopsis -- the model-organism portals). Query craft: Entrez, BioMart, and APIs (programmatic access -- the scale skill); the UCSC Table Browser.

Curation crisis: databases rot without curators (the funding gap); the biocurator profession; text mining's promise (and limits). Pakistani contribution: deposit local data (the wheat landrace sequences!); the visibility argument (deposited data gets cited). The exam: 'plan data management for a sequencing project.'

Key Terms & Definitions

- Accession numbers — The data currency.
- MIAME/MINSEQE — Experiment standards.
- BioMart / APIs — Programmatic queries.
- Biocuration — The maintenance profession.

Example / Illustration

- Wheat landrace project: raw reads -> SRA, assembly -> GenBank, variants -> EVA, phenotypes -> MIAPPE-compliant spreadsheet; accessions in the paper -- the complete deposition.

Lesson Summary

- Publish = deposit; accessions = currency.
- Metadata makes data reusable.
- Deposit local data; get cited.

Exam Pointers — Very Important

- ★ Plan FAIR data management for a project -- the stewardship practical.

Lesson 12: Reproducible Bioinformatics

Learning Objectives

- Ensure reproducibility
- Use workflows and containers
- Version control
- Budget compute

Reproducibility crisis: the pipeline that ran once (on the student's laptop, in 2019) -- bioinformatics' shame. Solutions: workflow managers (Snakemake, Nextflow -- the pipeline as code; nf-core's curated pipelines!), environments (conda/bioconda -- the package fix; containers -- Docker/Singularity: ship the whole OS!), and version control (Git/GitHub -- the lab notebook).

The reproducible project: README, environment file, workflow script, config, and test data (the five artifacts); CI testing (GitHub Actions); literate analysis (Jupyter/R Markdown -- code + narrative). The exam: 'make this analysis reproducible' -- list the five artifacts.

HPC and cloud: SLURM (the cluster's queue), the cloud bill (spot instances for cheap!), and the Pakistani reality (HEC's smart classrooms to cloud credits -- the access ladder). Cost craft: the \$1000 genome's \$100 analysis -- budget compute. The frontier: workflow languages converging; the well-documented pipeline as the publication's companion.

Key Terms & Definitions

- Snakemake / Nextflow — Workflow managers.
- Conda / containers — Environments; ship the OS.
- nf-core — Curated pipelines.
- Five artifacts — README, env, workflow, config, test.

Example / Illustration

- The 2019 script: hard-coded paths, unknown package versions -- rescue: conda env export, Snakemake rules, config.yaml, GitHub push, test dataset -- the five-artifact fix.

Lesson Summary

- Pipeline as code; environment as file.
- Five artifacts per project.
- Reproducible or it didn't happen.

Exam Pointers — Very Important

- ★ Convert a laptop analysis into a reproducible workflow -- the reproducibility practical.

Lesson 13: Careers and the Pakistani Scene

Learning Objectives

- Map Pakistani opportunities
- Build the skill portfolio
- Plan careers
- Watch the frontier

The Pakistani scene: universities (the bioinformatics programs -- Karachi, Lahore, Islamabad hubs), the NIH/HEC sequencing capacity, agri-biotech (wheat, cotton, rice genomics -- the food-security mission), and the diaspora bridge (collaborations as the accelerator). Career paths: academia, biotech/pharma (the docking/NGS analyst), clinical genomics (the hospital's variant interpreter -- the growing niche!), and the global remote market (the bioinformatician's laptop travels).

Skill portfolio: Linux + Python/R (non-negotiable), one pipeline deeply (RNA-seq or variants), statistics (the differentiator!), communication (the paper/thesis), and the GitHub profile (the portfolio). Learning ladder: Rosalind (problems), Biostar (Q&A),

Coursera specializations, then the project (the thesis as the portfolio piece).

Frontier watch: protein design (RFDiffusion), single-cell multi-omics, pangenome graphs, AI agents for bioinformatics (the copilots), and the climate-genomics nexus. The course's closing: special topics never settle -- the skill is learning the next topic fast. The exam's essay: 'bioinformatics and Pakistan's future' -- food, health, and the talent bet.

Key Terms & Definitions

- Non-negotiables — Linux + Python/R + statistics.
- Portfolio — GitHub as the CV.
- Clinical genomics — The growing hospital niche.

Example / Illustration

- The portfolio: GitHub with an RNA-seq Snakemake pipeline + a docking screen + a blog post explaining AlphaFold -- the job application that writes itself.

Lesson Summary

- Agri + health + remote = careers.
- Pipeline deeply; statistics differentiates.
- Learn to learn the next topic.

Exam Pointers — Very Important

- ★ Argue bioinformatics' role in Pakistan's future -- the closing essay.

Lesson 14: Revision: Bioinformatics Master Drill

Master drill (two per lesson): (1) data-to-question map; (2) FASTQ/BAM/VCF; (3) DESeq2; (4) BLOSUM62; (5) E-value; (6) bootstrap; (7) AlphaFold; (8) docking protocol; (9) MAGs; (10) 16S vs shotgun; (11) no leakage; (12) AUROC; (13) N50/BUSCO; (14) pangenome; (15) UMAP; (16) hubs; (17) FBA; (18) Manhattan plot; (19) Fst; (20) FAIR; (21) five artifacts. Pipeline drills: RNA-seq, variant calling, metagenomics, scRNA-seq, ML evaluation -- trace each from memory with tools. Essay frames: 'bioinformatics and Pakistan', 'the AlphaFold revolution', 'ML: promise and pitfalls'. Final: the next topic is coming -- learn to learn it.