Mini Project: End-to-End Genomic Data Analysis of Fungal Isolates (Galaxy)

Background
Whole-genome resequencing (WGS) enables variant discovery and comparative genomics across fungal isolates. In this project, students will process short-read FASTQ files from **5 known fungal species** plus **several unknown isolates**. They will perform quality control, read trimming, reference-guided alignment, variant calling, and annotation in **Galaxy**.

Objectives
- Navigate public databases to find references, annotations, and reads (NCBI/ENA/SRA, Ensembl Fungi).
- Execute a complete WGS pipeline in **Galaxy**: QC → trimming → alignment → post-processing → joint variant calling → filtering → annotation → core SNP matrix.

Dataset Provided
A project folder containing:
- **FASTQ** (paired-end) reads for:
  1. 5 known fungal species (≥2 isolates each if available).
  2. Several **unknown** isolates to classify.

- A metadata sheet (samples.tsv) with columns: sample_id, status (Known/Unknown), species (for known), SRA_accession (if applicable), library_layout, read_group, notes.

If you prefer, provide **SRA accessions** only; students will fetch reads inside Galaxy using "NCBI SRA Tools".

Tasks
1) Database Navigation (NCBI/ENA/SRA & Ensembl Fungi)
Goal: identify and download all inputs reproducibly.
Steps (students document each):
1. **Reference genome**:
   - Find a **chromosome-level** or best-available assembly for each **known species** on **NCBI Assembly** or **Ensembl Fungi**.
   - Download: reference.fasta and annotation.gff3 (or GTF).
   - Record assembly accession (e.g., GCA_XXXXXXXXX.X) and version.
2. **Reads** (if not pre-provided):
   - Use **SRA Run Selector** to list runs, confirm **paired-end**, Illumina platform, and similar read length.
   - Note the SRA accessions for each sample and add to samples.tsv.
3. **Document** database pages/screenshots + accessions in a short "Data Provenance" note.

2) Load Data into Galaxy
Goal: organise the project as reproducible Galaxy histories & collections.
Steps:
- Create a Galaxy **History** named Fungal_Genomics_Project_<YourName>.

- Upload or fetch:
    - FASTQs (or use **Get Data → NCBI SRA Tools: Fasterq-dump**).
    - References (reference.fasta) + annotation.gff3 for each species.
- Build **Dataset Collections** for paired reads (R1/R2).
- Rename datasets with clear labels: SPC1_iso1_R1 / SPC1_iso1_R2, etc.

3) Quality Control (Galaxy)
Tools & sequence (typical choices in parentheses):

1. **FastQC** on all raw reads.
2. **MultiQC** to summarize FastQC results.
3. **Adapter/quality trimming** (e.g., **Trim Galore!** or **fastp**):
    - Typical params: adapter auto-detect; quality cutoff 20; min length 50–70.
4. **FastQC** (post-trim) → **MultiQC** (compare improvement).
   Output: MultiQC HTML reports for **raw** and **trimmed** reads.

4) Alignment & Post-Processing (Galaxy)
Reference choice:

- Option A (simplest): use a **single reference** from the species you expect most unknowns to belong to.
- Option B (rigorous): map each isolate to its **species-specific reference**, then combine variants in a species-aware manner. (Pick A for first run; B as bonus.)

Tools & steps (BWA-MEM2 pipeline example):

1. **BWA-MEM/MEM2**: index reference.fasta; map paired reads → SAM.
2. **Samtools sort** → BAM; **Samtools index**.
3. **Picard MarkDuplicates** (or GATK MarkDuplicates).
4. **Alignment metrics**:
   Samtools flagstat and **idxstats**;
   Optional: **Qualimap BamQC**;
   bedtools genomecov for coverage summaries.

Output: deduplicated, indexed BAMs; metrics tables; coverage summaries.

5) Variant Calling & Joint Genotyping (Galaxy)
Two solid routes (pick one):
Route 1: bcftools mpileup/call

1. **bcftools mpileup** (per sample) with -Ou -f reference.fasta.
2. **bcftools call** (per sample) with -mv (variants only) → per-sample VCF.
3. **bcftools merge** (multi-sample) to form a **joint VCF**.

Route 2: FreeBayes (population calling)

- **FreeBayes** on a **collection** of BAMs to emit one **multi-sample VCF**.

Filtering (either route):

- **bcftools filter**: depth (e.g., DP≥8–10), quality (QUAL≥30), genotype quality (GQ≥20), missingness (retain sites with ≥80% genotyped).

- Optionally **vcftools**: --max-missing 0.8, --maf 0.01.

Output: a **filtered multi-sample VCF**.

6) Variant Annotation (Galaxy)

- Build or select a database for your species:
  SnpEff: create/select the fungal genome database (Ensembl Fungi GFF3).
  Alternatively, **VEP** (if available) for functional consequence.

- Run **SnpEff** on the filtered VCF → annotated VCF (*.snpeff.vcf) + summary HTML.

Output: functionally annotated VCF + summary.

7) Core SNP Matrix (Galaxy)

Goal: derive a core SNP alignment.

1. **Extract biallelic SNPs only** (e.g., **bcftools view -v snps -m2 -M2**).

2. **Create alignment**:
   vcf2phylip (Galaxy wrapper) or **SNP-sites** to convert VCF → PHYLIP/FASTA SNP alignment.

Analysis

- **Data quality effects**: Do low-coverage samples behave erratically? Reference bias?

- **Variant filters sensitivity**: Show how stricter/looser DP/QUAL thresholds affect clustering.

- **Functional signals** (optional): Are there species-specific HIGH-impact variants?

Expected Output

1. **Galaxy artifacts**
   Two **MultiQC** reports (pre- and post-trim).
   Deduplicated, indexed **BAM** files + alignment/coverage metrics.
   Filtered multi-sample VCF and **SnpEff-annotated VCF** + HTML summary.
   Core SNP alignment (FASTA/PHYLIP).

2. **Presentation (10–12 slides)**
   Figures: MultiQC screenshots, pipeline schematic.
   Discussion of agreement/disagreement; sensitivity to filters; limitations.

3. **Reproducibility bundle**
   Galaxy **Workflow export (.ga)**, **History export (.tar)**, and a README with tool versions/parameters.

Practical Tips & Parameter Hints

- **Trim Galore!** : Q=20, min len=50–70; auto-adapters.
- **Alignment**: BWA-MEM2 default; ensure read groups if needed.
- **Filtering**: start with DP≥10, GQ≥20, QUAL≥30, max missing ≤20%; then explore sensitivity.
- **SnpEff**: confirm genome build matches reference; document database/version.