**AIV Whole‑Genome Resequencing Pipeline — Methods & Rationale**

*A conceptual, step‑by‑step guide explaining what we did, why we did it, and how to interpret each stage. This document is species‑agnostic; apply it in each <Species> folder.*

**0) Project layout & data provenance**

**Why**: Keep inputs/outputs standardized across species and make the workflow reproducible.

**Layout** (per species under $PROJ\_ROOT):

<Species>/

raw/ # raw FASTQs (after species sorting)

fastq\_merged/ # lane-merged FASTQs + samples.tsv

fastqc\_pre/ # FastQC + MultiQC (pre-trim)

trim/ # fastp‑trimmed FASTQs + reports

fastqc\_post/ # FastQC + MultiQC (post-trim)

align/ # BAMs (.sorted, .markdup or .rg.md) + .bai

gvcf/ # per‑sample .g.vcf.gz

vcf/ # combined.g.vcf.gz, raw.vcf.gz, filtered SNPs

pca/ # PLINK bed/bim/fam, eigen\* files, PCA plots

tree/ # distance matrices, trees (PDF/NWK)

logs/ tmp/ # LSF logs & temp

<Species>.fa # reference genome + index files

**Species sorting**:

* Raw filenames: ln<batch><Code><id>\_S#\_L###\_R1/2\_001.fastq.gz.
* Species code → species name: A→*Amaranthus cruentus*, B→*Brassica carinata*, N→*Solanum nigrum*, C→*Cleome hirta*, J→*Corchorus olitorius*, V→*Vigna unguiculata*, S→*Crotalaria brevidens*.
* We distributed files by the single letter after ln<digits> and lane‑merged reads per sample.

**Sample sheet (samples.tsv)**: 3 columns → sample\tR1\tR2 (used by array jobs).

**1) Quality control (pre‑trim)**

**What**: Run FastQC on raw reads; aggregate with MultiQC.

**Why**: Identify adapter contamination, low quality tails, GC issues, overrepresented sequences.

**Key outputs**: fastqc\_pre/\*fastqc.zip/html, fastqc\_pre/multiqc\_report.html.

**Interpretation** (high level):

* **Per base quality** should be high; if 3’ end quality drops, trimming will help.
* **Adapter content** or **overrepresented sequences** → fastp will remove.
* **Sequence duplication** high values may reflect PCR duplicates or biological complexity; deduplicate post‑alignment.

**2) Trimming with fastp**

**What**: Adapter/quality trimming, minimal length filtering.

**Why**: Improve mapping and reduce artifacts.

**Key params** used:

* --detect\_adapter\_for\_pe (auto‑detect adapters)
* --qualified\_quality\_phred 15 (lenient base quality floor)
* --length\_required 50 (drop very short reads)

**Outputs**: trim/${sample}\_R1.fastq.gz, trim/${sample}\_R2.fastq.gz, JSON/HTML reports in trim/reports/.

**QC**: After trimming, run FastQC + MultiQC again (fastqc\_post/). Adapter/quality warnings should reduce.

**3) Alignment & duplicate marking**

**What**: Align to species reference with **BWA‑MEM2** (or **minimap2** for distant reference), then **MarkDuplicates**.

**Why**: BWA‑MEM2 is robust for short reads; duplicate marking needed for accurate variant calling.

**Key details**:

* **Read groups** (@RG) include ID, SM, PL; required by GATK tools.
* We sort BAMs and mark duplicates with Picard (.markdup.bam). If RG tags were missing, we add them and re‑mark (.rg.md.bam).

**Outputs**: align/${sample}.markdup.bam (+ .bai) or align/${sample}.rg.md.bam (+ index).

**QC**: Check mapping rate, insert size, duplication rate (Picard metrics), and BAM indexes.

**Cross‑species references**:

* If a species lacks a reference (e.g., *Crotalaria brevidens*), we used a close species (e.g., *Crotalaria pallida*).
* Expect lower mapping rate & higher missingness; consider minimap2 -ax sr as alternative aligner.

**4) Reference preparation**

**What**: Index the reference once per species.

**Why**: Required by aligners and variant callers.

**Indices**: BWA (\*.bwt.2bit.64, .0123, .amb, .ann, .pac), FASTA index (.fai), Picard dict (.dict).

**Idempotent**: Scripts skip building if files exist.

**5) Variant calling (per‑sample gVCFs)**

**What**: GATK HaplotypeCaller per sample with -ERC GVCF to produce .g.vcf.gz.

**Why**: gVCF mode captures genotype likelihoods across sites for joint calling; scales well to cohorts.

**Inputs**: RG+dedup BAMs; **Reference**: <Species>.fa.

**Outputs**: gvcf/${sample}.g.vcf.gz (+ .tbi).

**6) Joint genotyping**

**What**: Combine per‑sample gVCFs, then joint genotype.

**Why**: Ensures consistent cohort‑level genotypes and variant sites.

**Tools**: gatk CombineGVCFs → gatk GenotypeGVCFs.

**Outputs**: vcf/${Species}.combined.g.vcf.gz, vcf/${Species}.raw.vcf.gz (+ .tbi).

**7) Variant filtration (hard filters)**

**What**: Split SNPs/INDELs, apply GATK hard filters to SNPs, keep PASS, then thin to biallelic and basic MAF/missingness.

**Why**: Remove likely artifacts without a known truth set (VQSR not practical for non‑model species/cohort sizes).

**Thresholds used** (SNPs):

* QD < 2.0, FS > 60.0, MQ < 40.0, SOR > 3.0, MQRankSum < -12.5, ReadPosRankSum < -8.0 → **filter out**
* Keep **PASS** only, **biallelic SNPs**, then apply **MAF ≥ 0.01** and **F\_MISSING ≤ 0.2** (tunable; see “Notes”).

**Outputs**: vcf/${Species}.snps.filt.vcf.gz → snps.pass.vcf.gz → snps.clean.vcf.gz (+ .tbi).

**Notes**:

* For small cohorts, consider easing missingness (e.g., ≤0.3) if too few SNPs remain.
* Cross‑species alignment often increases missingness; balance stringency vs. dataset size.

**8) PCA preparation (PLINK2)**

**What**: Convert clean VCF to PLINK bed/bim/fam with unique variant IDs; LD prune (or thin if n<50); PCA.

**Why**: Reduce correlated markers; PCA reveals population structure and outliers.

**Details**:

* Unique IDs: --set-all-var-ids @:#:$r:$a to avoid duplicate rsIDs.
* If **samples < 50**: LD estimates unstable → use MAF/geno filters + --thin as a fallback.
* Standard: --indep-pairwise 50 5 0.2 to prune, then --pca 20.

**Outputs**: pca/${Species}.clean.\*, ${Species}.pca.\*, ${Species}.eigenvec/.eigenval.

**Plotting**:

* We generate labeled PC1–PC2 scatterplots with ggrepel to avoid overlaps.
* Alternative: SNPRelate (GDS + LD pruning inside R) for small/edge cases.

**9) Phylogenetic inference**

**Approach A (IBS + NJ)**:

* **What**: PLINK pairwise IBS distance (--distance square ibs), then **Neighbor‑Joining** via **ape**.
* **Why**: Fast, reasonable for within‑species diversity; easy visualization.
* **Outputs**: tree/${Species}.mdist.gz → ${Species}\_NJ\_tree.pdf & .nwk.

**Approach B (SNPhylo)**:

* **What**: End‑to‑end SNP filtering + ML tree (RAxML) with bootstrap.
* **Why**: Alternative per project plan; provides support values.
* **Note**: Requires additional dependencies; we include an automated script.

**10) Job scheduling (LSF) & resources**

**Queues** used\*\*:\*\* short (≤~1 h) for quick steps; long for alignment, gVCF, joint genotyping.

**Typical requests**: -n 8, -R 'span[hosts=1] rusage[mem=4..8]' depending on step.

**Array jobs**: Steps 02/03/03fix/04 run as arrays over samples.tsv.

**Check pending reasons**:

bjobs -l <JOBID> | sed -n '/PENDING REASONS/,+30p'

**Tuning**:

* If stuck pending, try bmod -q short <JOBID>, reduce -n, or lower rusage[mem=…].
* Use job name prefixes including species (we do via ${PWD##\*/}) for clarity.

**11) Cross‑species mapping considerations**

* **Reference choice** matters: closer reference ⇒ higher mapping rate, better genotyping.
* For *Crotalaria brevidens* we used *C. pallida*; expect higher missingness and possible reference bias.
* Consider **minimap2** for distant references and compare mapping metrics.
* Adjust SNP filtering thresholds (especially F\_MISSING) to retain informative sites.

**12) Reproducibility & documentation**

* All step scripts live in $PROJ\_ROOT/lsf\_templates/ and are versioned in GitHub.
* The **GitHub Pages** site (docs/) hosts one‑liners with Copy buttons.
* Each species folder is self‑contained with consistent outputs.
* Logs in logs/ capture parameters, stdout/stderr, and tool versions.

**Provenance tips**:

* Save multiqc\_report.html (pre/post) for methods/QC sections.
* Archive final snps.clean.vcf.gz, PCA plots, and tree PDFs per species.
* Record reference genome source/versions (NCBI accessions) in a REFERENCE.md per species.

**13) Parameters cheat‑sheet**

**fastp**: Q≥15; min length 50; auto adapters; paired mode.

**BWA‑MEM2**: -R "@RG\tID:sample\tSM:sample\tPL:ILLUMINA".

**GATK SNP filters**: QD<2; FS>60; MQ<40; SOR>3; MQRankSum<-12.5; ReadPosRankSum<-8 ⇒ filtered; keep **PASS** only.

**bcftools**: biallelic SNPs; MAF≥0.01; F\_MISSING≤0.2 (tune by species/cohort).

**PLINK2 LD prune**: 50 5 0.2 (or fallback thin if n<50).

**SNPRelate LD prune**: ld.threshold=0.2, maf=0.05, missing.rate=0.2.

**14) Common pitfalls & fixes**

* **"Sam file header missing Read Group"**: add RG via Picard, then re‑MarkDuplicates (we provide 03\_fix\_rg\_and\_md\_picard.lsf).
* **PLINK2 error about duplicate IDs / contigs**: use --set-all-var-ids @:#:$r:$a and --allow-extra-chr.
* **Too few samples for LD**: Use the fallback path (thin/no‑miss pipeline) or SNPRelate.
* **Picard dict already exists**: benign; our scripts skip re‑creating indexes.
* **FastQC truncated file warning**: re‑check upstream FASTQs; ensure lane merge finished and files not corrupted.

**15) Deliverables per species (what to archive)**

* fastqc\_pre/ & fastqc\_post/ MultiQC HTMLs (QC evidence)
* align/\*.markdup.bam + .bai (or .rg.md.bam) (optional long‑term)
* vcf/${Species}.raw.vcf.gz (+ .tbi) and ${Species}.snps.clean.vcf.gz (+ .tbi)
* pca/\*\_PCA\_\*\_Labeled.png (final PCA figure with labels)
* tree/\*\_NJ\_tree.pdf and .nwk (phylogeny)
* A short METADATA.md noting sample counts, reference, filters used, date, tool versions

**16) Where to find the runnable scripts**

All .lsf files live in: **$PROJ\_ROOT/lsf\_templates/**. The GitHub Pages site shows one‑liners that submit them with sensible resources and job names.

* 00\_prep\_ref.lsf — reference indexing
* 01\_fastqc\_pre.lsf / 01b\_fastqc\_post.lsf — QC
* 02\_fastp.lsf — trimming
* 03\_align\_markdup.lsf — alignment + duplicates
* 03\_fix\_rg\_and\_md\_picard.lsf — ensure RG + MD
* 04\_hc\_gvcf.lsf — per‑sample gVCF
* 05\_joint\_geno.lsf — joint genotyping
* 06\_filter\_vcfs.lsf — hard filters → clean SNPs
* 07\_plink2\_pca\_all.lsf — PCA build (PLINK2)
* 07c\_pca\_nomiss.lsf — PCA fallback (drop empty)
* 08\_snprelate\_pca.lsf — PCA via SNPRelate
* 10\_plot\_pca\_ggrepel.lsf — labeled PCA plot
* 11\_tree\_distance.lsf — IBS distance
* 12\_tree\_nj\_pretty.lsf — NJ tree with labels
* 12b\_snphylo.lsf — SNPhylo pipeline (alt.)

**17) Notes for the manuscript Methods section**

* **Data generation**: library type, read length, platform (fill in from lab records).
* **Read processing**: fastp settings; pre/post QC with FastQC/MultiQC.
* **Alignment**: reference source (NCBI accession), aligner (BWA‑MEM2), duplicate marking, RG tags.
* **Variant discovery**: GATK HaplotypeCaller (gVCF) + joint genotyping.
* **Filtering**: hard thresholds (list) with motivation; biallelic SNPs; MAF/missingness.
* **Population genetics**: PCA (PLINK2 and/or SNPRelate), IBS distances, NJ tree; any bootstrap support if SNPhylo used.
* **Cross‑species mapping**: justify reference substitution (e.g., *C. pallida* for *C. brevidens*) and note expected biases.
* **Computational env**: NCSU HPC (LSF); conda environment & key tool versions.

*End of document.*