RNA-seq Data Processing

Raw RNA sequencing data were obtained from the NCBI SRA database, and metadata, including the tissue of origin and experimental conditions for each SRA entry, were compiled. Data were extracted using v0.6.7 (https://github.com/rvalieris/parallel-fastq-dump). Quality control and adapter trimming were performed using Fastp v 0.23.41 to remove low-quality sequences. Clean reads were aligned to reference genome using HISAT2 v2.2.1[2], with genome indexing performed using hisat2-build. BAM files were generated and sorted using Samtools v1.19.2[3]. Samples with mapping rates below 60% were excluded from downstream analysis. Read counts were assigned to genes based on the reference annotation using featureCounts v2.0.6[4], producing gene expression matrices for both paired-end and single-end datasets. Gene expression was normalized using FPKM and TPM for paired-end data and TPM for single-end data, with all calculations performed using custom scripts (https://github.com/MIEFIGH/OrphancerealsOmicsDatabase).

SNP Calling Processing

Raw sequencing data in SRA format were obtained from public databases and converted to FASTQ files using parallel-fastq-dump v0.6.7 (https://github.com/rvalieris/parallel-fastq-dump). Quality control and adapter trimming were performed using Fastp v 0.23.4[1] to remove low-quality sequences. Reads were aligned to the reference genome using BWA-MEM2 v0.7.18[5]. The resulting BAM files were sorted using Samtools v1.19.2[3] to ensure proper genome coordinate ordering. These sorted and quality-controlled BAM files were used as input for variant calling. SNP calling was performed using the GATK v3.7[6]**.** The analysis was conducted separately for each chromosome, with temporary directories created for efficient processing. The UnifiedGenotyper was executed with multi-threading enabled and a variant confidence threshold set to ensure high-quality SNP calls. These chromosome-level VCFs were then merged using BCFtools v1.17[3], producing a final compressed VCF file.

References

1. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
2. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**,

907–915 (2019).

3. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
4. Liao, Y., Smyth, G. K. & Shi, W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).
5. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
6. Auwera, G. A. V. der *et al.* From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11.10.1-11.10.33 (2013).