Bioinformatician Technical Test

First, we would like to sincerely thank you for taking the time to complete this exercise!

Context: *In the lab, we work with many different cohorts due to our collaborations and various projects. All of these cohorts are processed through pipelines developed by the bioinformaticians in the lab. This brief exercise is intended to place you in that context, albeit on a much smaller scale.*

You are provided with two mini-cohorts (Cohort_A and Cohort_B), each containing 3 gVCF files and a metadata.tsv. The metadata file contains per-individual information with the following columns: SampleID, Age, Ancestry, and IQ.

1. Develop a workflow that takes as input a list of gVCF files and the corresponding phenotype/metadata files, filters variants to retain only heterozygous sites (0/1, 1/0) that pass QC (DP > 20 and GQ ≥ 30), and computes the number of such sites per individual. These counts should then be merged with the metadata table to produce a final structured dataset containing: SampleID, Age, Ancestry, IQ, Cohort and Het_Count.

The pipeline should automatically group samples by cohort (based on the file path, filename prefix, or metadata), and produce one output per cohort. For each cohort, generate a .parquet file containing the final merged table.

Optional: Automatically generate a .pdf report including summary visualizations of the merged table (e.g., boxplots of Age and Het_Count).

2. In addition to the implementation, please provide a short written explanation answering the following open question:

How would you optimize, benchmark, and scale this pipeline to thousands of samples? ***(max 200 words)***