

AI and Quantum Research Center (AIQRC), Asia University, Taiwan**Activities**

Asia University and AIQRC, supported by the Ministry of Education's Taiwan Global Pathfinders Grant Program, provide Taiwan students with world-class experiential learning opportunities through programs in IBM quantum technologies, UCLA academic exchange, quantum finance, humanoid robotics, and astronaut mission training.



(A) IBM dream program



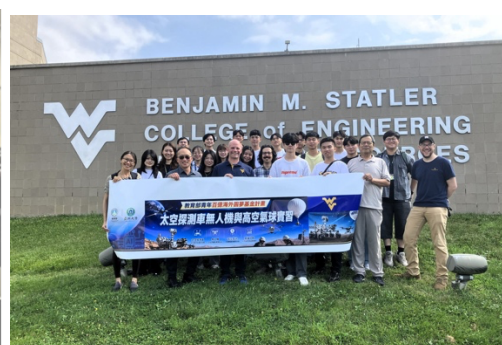
(B) UCLA elite program



(C) Quantum finance program



(D) Humanoid robot program



(E) Astronaut mission training program

Prepared and Edited by

Ka-Lok Ng ^{1,2}

Deputy Director ¹ & Distinguish Professor ²

¹ [AI and Quantum Research Center \(AIQRC\)](#),

² [Department of Bioinformatics and Medical Engineering](#), Asia University, Taiwan

Building a National-Scale Variant Catalog: Infrastructure Lessons from Combining Tens of Thousands of Genomes

Over the past few months, I have worked on a population-scale genomics initiative — a national health authority's effort to build a genetic reference resource for its citizens. Given the sensitivity of the engagement, I cannot name the country or institution involved, but the underlying challenge is one that many national genome programs

share: turning tens of thousands of individual sequencing files into a single resource that clinicians, researchers, and policymakers can actually rely on.

Why build a national variant catalog

My part of the project is the variant catalog: a consolidated map of genetic variation across the population. It exists for three reasons. First, global reference databases such as gnomAD under-represent large parts of the world's population, so variant interpretation for many citizens defaults to frequencies that do not reflect their own ancestry — a population-specific catalog corrects that. Second, it underpins national pharmacogenomic and carrier-screening programs, since knowing which variants are common locally changes which tests are worth running at scale. Third, it feeds public health epidemiology and precision-medicine planning, giving policymakers a genetic baseline for the population they serve.

Concretely, the work starts with tens of thousands of individual VCF files — one per sequenced person — that need to be combined into a single Variant Dataset (VDS) representing the whole cohort. From there we compute the statistics that make the catalog useful: allele frequencies (MAF), call rates, Hardy-Weinberg equilibrium, and variant counts across the genome, among others.

What the catalog has to be

Easy to update. New sample batches need to be folded into the existing catalog without recomputing everything from scratch — essential once the population being sequenced runs into the hundreds of thousands.

Secure. This is identifiable genetic data belonging to real citizens, so encryption, strict access control, and a full audit trail are non-negotiable design constraints, not add-ons.

Scalable and reproducible. The system has to keep working as the cohort grows toward nationwide coverage, and every statistic produced needs to be traceable back to the exact data version it came from.

Cost-efficient at scale. Compute and storage costs compound quickly across tens of thousands of genomes, so the architecture has to treat cost as a first-class requirement, not an afterthought.

How it is built

Combining tens of thousands of genome files is far beyond what a single machine can handle, so the pipeline runs on **Apache Spark**, which distributes the work across many nodes running in parallel.

On top of Spark we use **Hail**, a framework built specifically for population genomics. Spark alone is a general-purpose engine; Hail adds the pieces a project like this actually needs — a storage format designed for genomic data that avoids duplicating the vast stretches of a genome that simply match the reference, and a combiner that supports adding new samples incrementally rather than reprocessing the whole population every time. That combination is what makes the “easy to update” requirement realistic instead of aspirational.

For storage, the catalog lives on **Amazon S3** rather than on the compute cluster itself. Separating storage from compute means the Spark/Hail cluster only needs to run while a job is active and can shut down afterward, which matters directly for cost at this scale. It also lets us apply fine-grained access control and versioning to the data independently of whatever compute happens to be running — an important piece of the security requirement above.

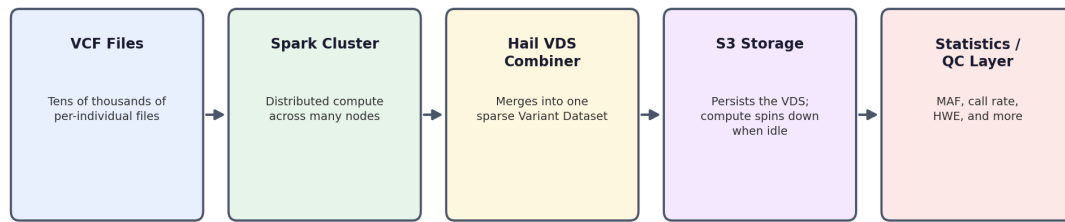


Figure 1: Variant catalog combine & QC architecture.

The broader lesson has less to do with genomics specifically and more with a pattern that shows up in any national-scale data project: the requirements that matter most — updatability, security, reproducibility, cost — are rarely satisfied by a single tool. They come from deliberately layering tools that are each good at one part of the problem.

Prepared by
Nguyen Manh Cong^{1, 2}

¹AI and Quantum Research Center (AIQRC), Asia University, Taichung, Taiwan

² DataXight
Phone: (+84) 937.123.728



AI and Quantum Research Center (AIQRC)
Room A110, Asia University, No. 500, LiuFeng Rd., WuFeng Dist., Taichung City 41354 Taiwan.
Email: gphys.qcomp@gmail.com Office: 04-23323456 ext. 6631
Web: <https://quantum.asia.edu.tw/>