Disclosure Slide

Financial Disclosure for:
Andrea Haessly
Principal Software Engineer

I have nothing to disclose

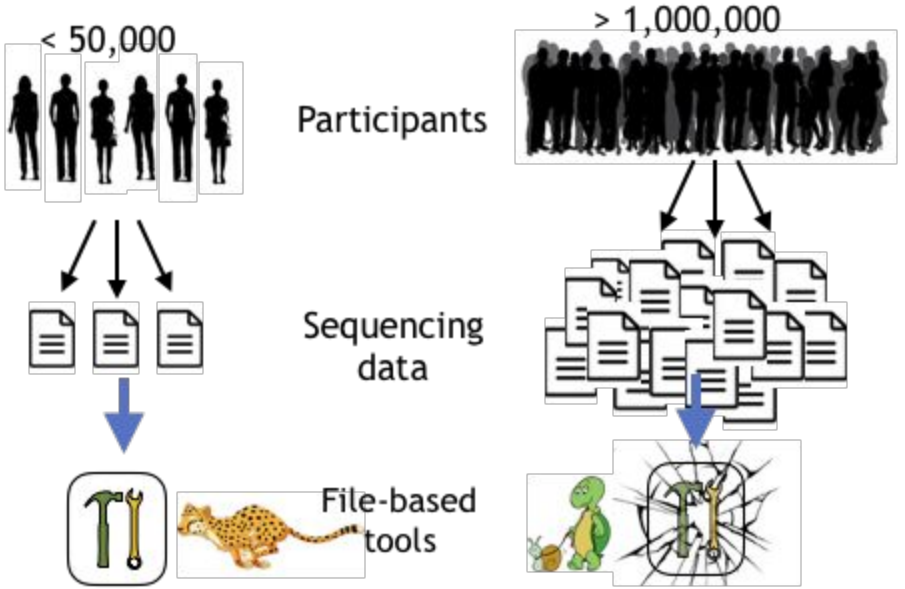
Scaling to 1M genomic samples for for the *All of Us* Research Program

A. Haessly, A. Cremer, K. Cibulskis, M. Shand, J. Soto, J. T. Smith, L. Bergelson, D. Roazen, Y. Farjoun, L. Gauthier, E. Banks

DSP, Broad Inst, Cambridge, MA, USA

Introduction

Existing tools for analyzing genomic data are largely file based and have difficulties supporting call sets with greater than 50,000 samples. The All of Us Research Program is recruiting **1 million** or more participants who will be providing health records and bio samples to be genotyped and sequenced. The genotyping data for *All of Us* has nearly **2 million probes** which can not be effectively processed with the current tools. As part of the Data and Research Center, we have developed and utilized new tools that provide efficient, scalable, and low-cost solutions for processing and managing these data.



Capabilities with a database solution

Analysis

Much of the analysis that is performed for site level quality control has, in the past, had to process all the files for the dataset in order to calculate the required summary statistics. Now, these same statistics can be quickly computed within the database.

Search

With the data accessible by sample, loci and attribute, the search capabilities are beyond what has typically been available.

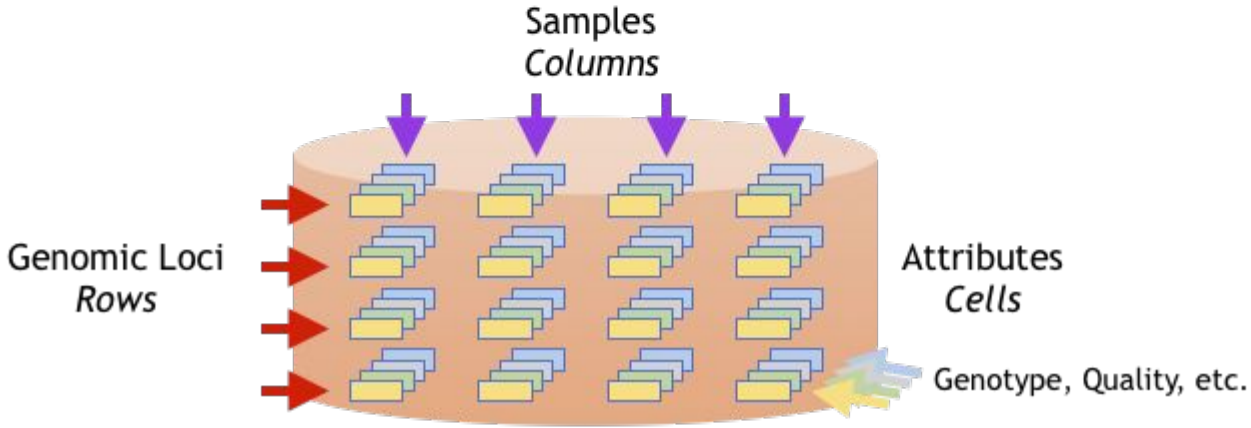
- By gene or loci
- By genotype, allele or other attribute
- Within a specific cohort of samples

Data Retrieval

The database solution allows ultimate flexibility in providing only the data of interest for additional analysis. This includes the ability to retrieve combinations of:

- Specific samples
- Specific genes or loci
- Specific attributes (e.g. genotype, phasing, quality)

And producing the required formats for downstream processing.



Google BigQuery for genomic data

Advantages

- Serverless, cloud-native solution
- Enhanced security
- Built for scalability
- Only pay for storage used
- Storage costs are separate from compute costs
- Optimized for common queries
- Run QC analytics directly on data

Cost (USD)	Per sample	1M samples
Compute for import	0.002	2,000
Storage / month	0.0017	1,700
QC Analysis	< 0.001	145
Retrieval of a 10,000 sample cohort in vcf format