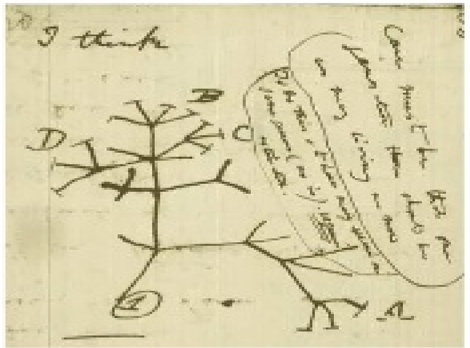
Hail: Scaling Genetic Data Analysis with Apache Spark

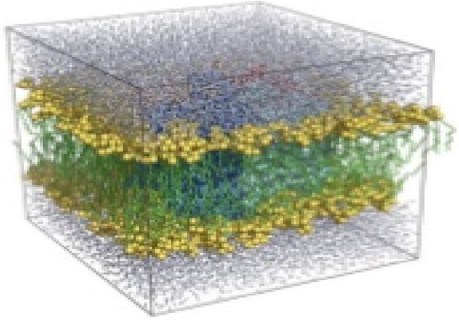
Cotton Seed, Principal Software Engineer
Tech Lead, Hail Team
Broad Institute and MGH

BROAD INSTITUTE STANLEY CENTER FOR PSYCHIATRIC RESEARCH AT BROAD INSTITUTE MASSACHUSETTS GENERAL HOSPITAL hail

Paradigms of Science

1. Empirical: describe natural phenomena
2. Theoretical: models, generalizations
3. Computational: simulate complex phenomena
4. Data Intensive: Jim Gray's 4th Paradigm
   - Automated, high-throughput data collection
   - Complex analysis pipelines
   - Experiments become computations


I think
A
B
C
D
1



Broad Institute Data

- The Broad sequences **1 genome every 10 minutes**.
- The Broad generates **17 TB** of new genomes per day.
- The Broad manages **45 PB** of scientific data.

BROAD
INSTITUTE

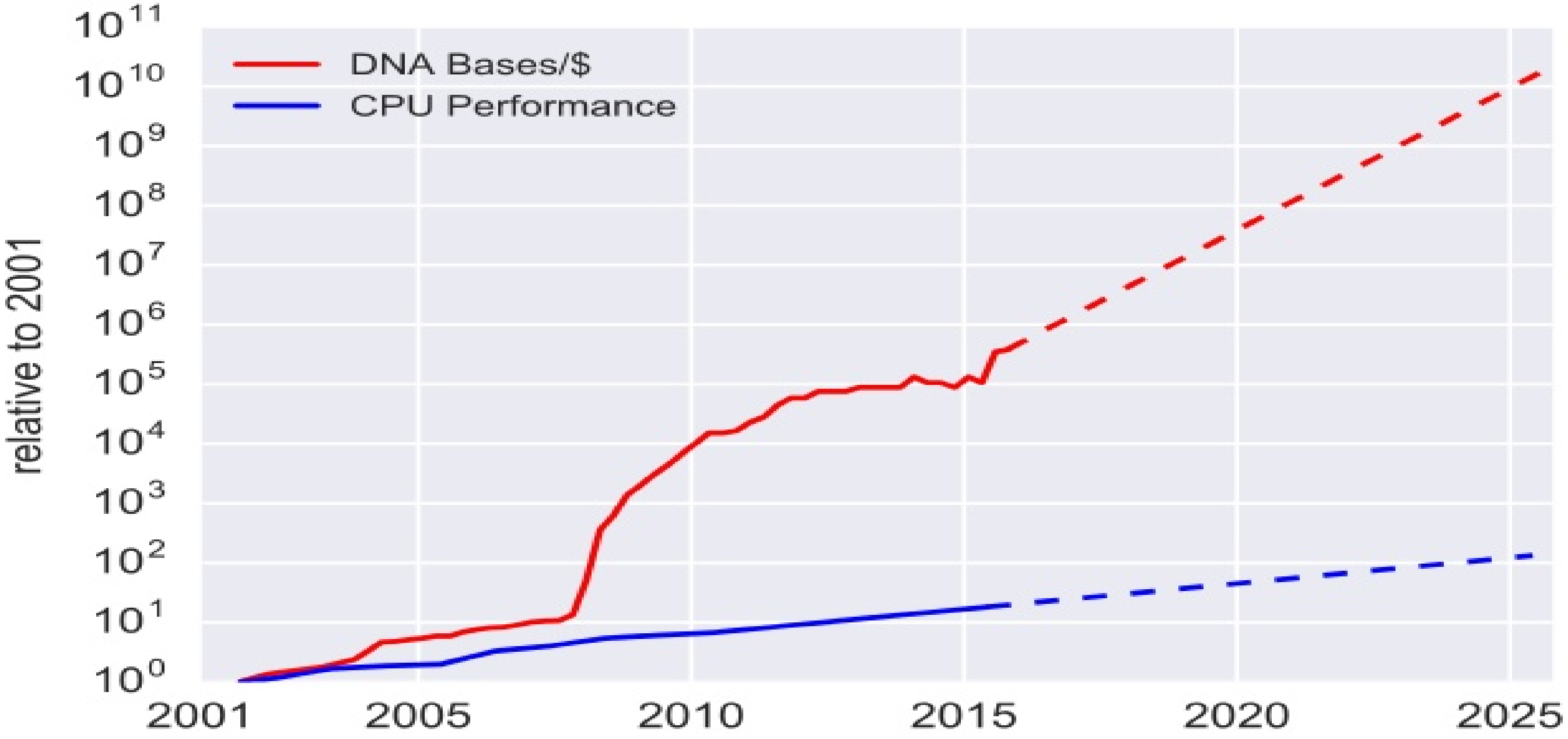
Broad Institute Data

- The Broad sequences **1 genome every 10 minutes**.
- The Broad generates **17 TB** of new genomes per day.
- The Broad manages **45 PB** of scientific data.

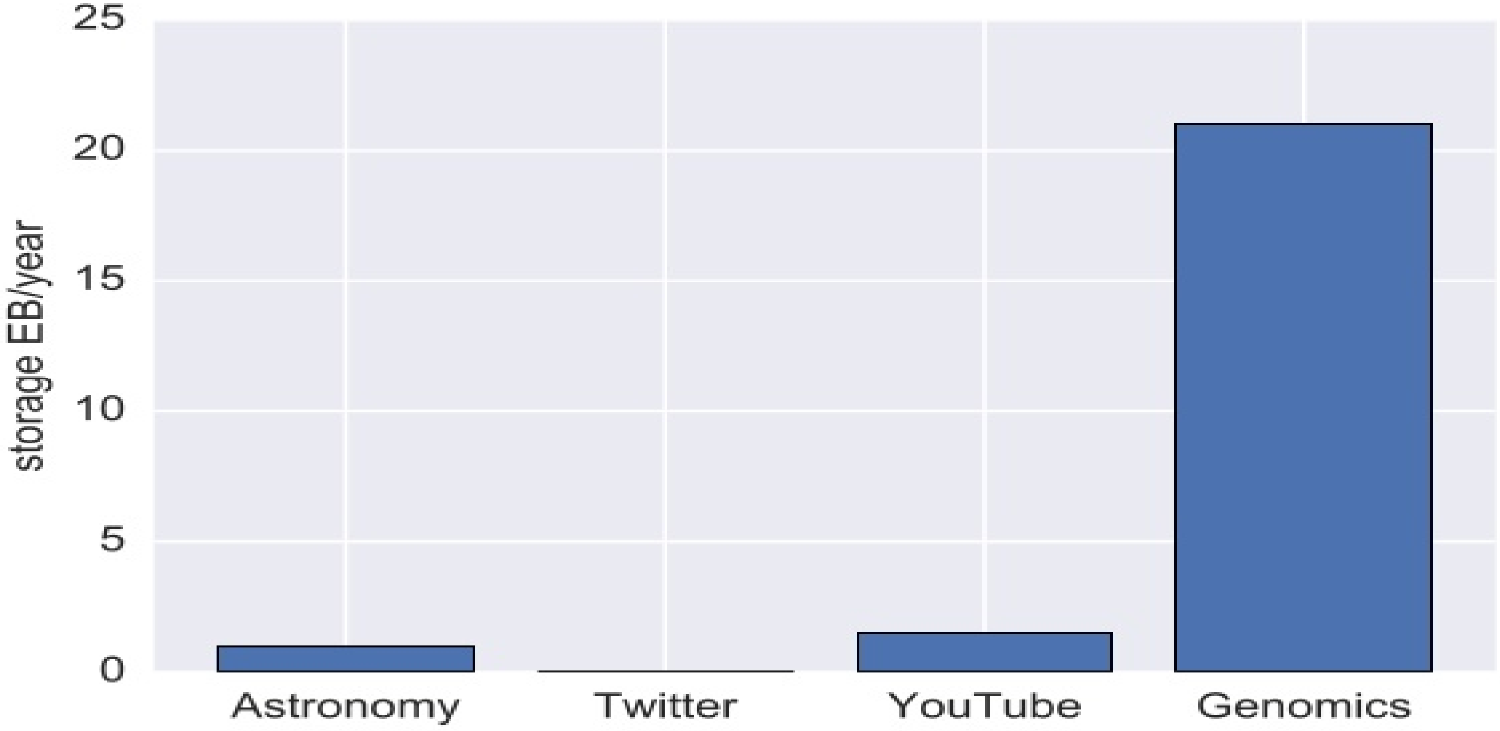
VS

- YouTube receives **24 TB** of new videos per day.
- YouTube stores about **86 PB** of video.

Move Over Moore's Law

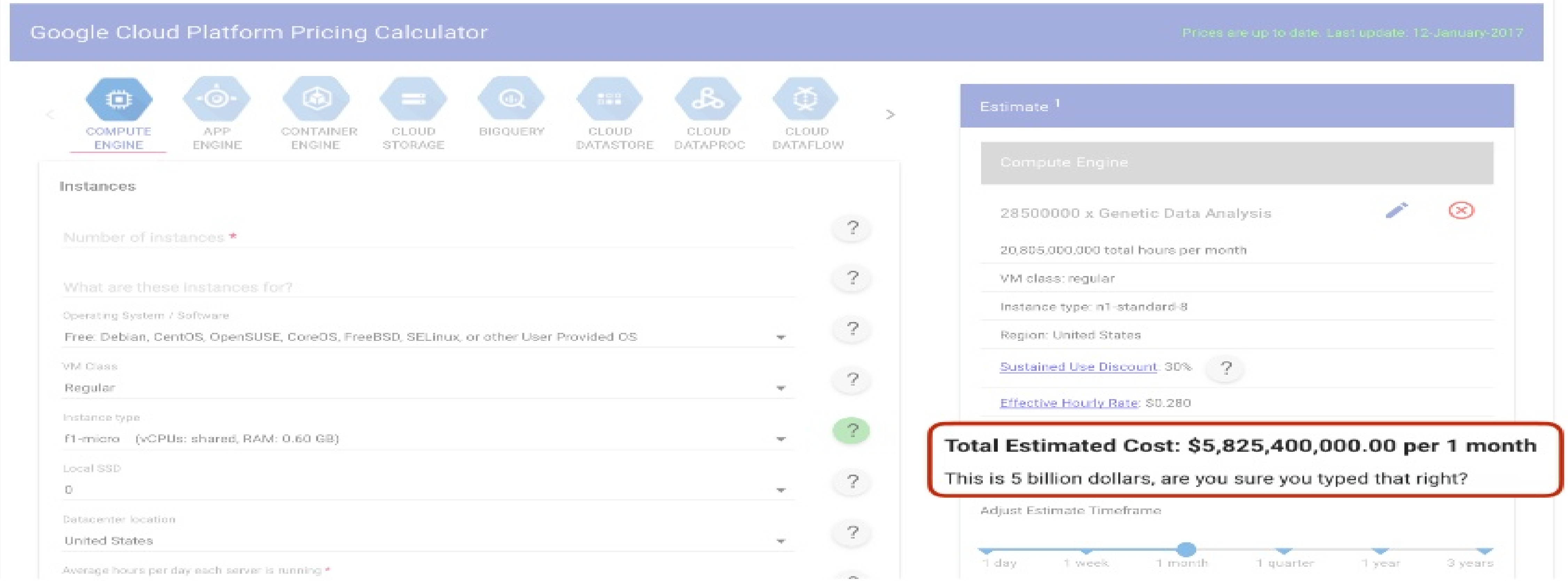


Data Acquisition in 2025



Stephens, et al., *Big Data: Astronomical or Genomical?* (2015)

2 Trillion Compute Hours



Google Cloud Platform Pricing Calculator

Prices are up to date. Last update: 12-January-2017

Instances

Number of instances *

What are these instances for?

Operating System / Software
Free: Debian, CentOS, OpenSUSE, CoreOS, FreeBSD, SELinux, or other User Provided OS

VM Class
Regular

Instance type
f1-micro (vCPUs: shared, RAM: 0.60 GB)

Local SSD
0

Datacenter location
United States

Average hours per day each server is running *

Estimate [1]

Compute Engine

28500000 x Genetic Data Analysis

20,805,000,000 total hours per month

VM class: regular

Instance type: n1-standard-8

Region: United States

Sustained Use Discount: 30%

Effective Hourly Rate: $0.280

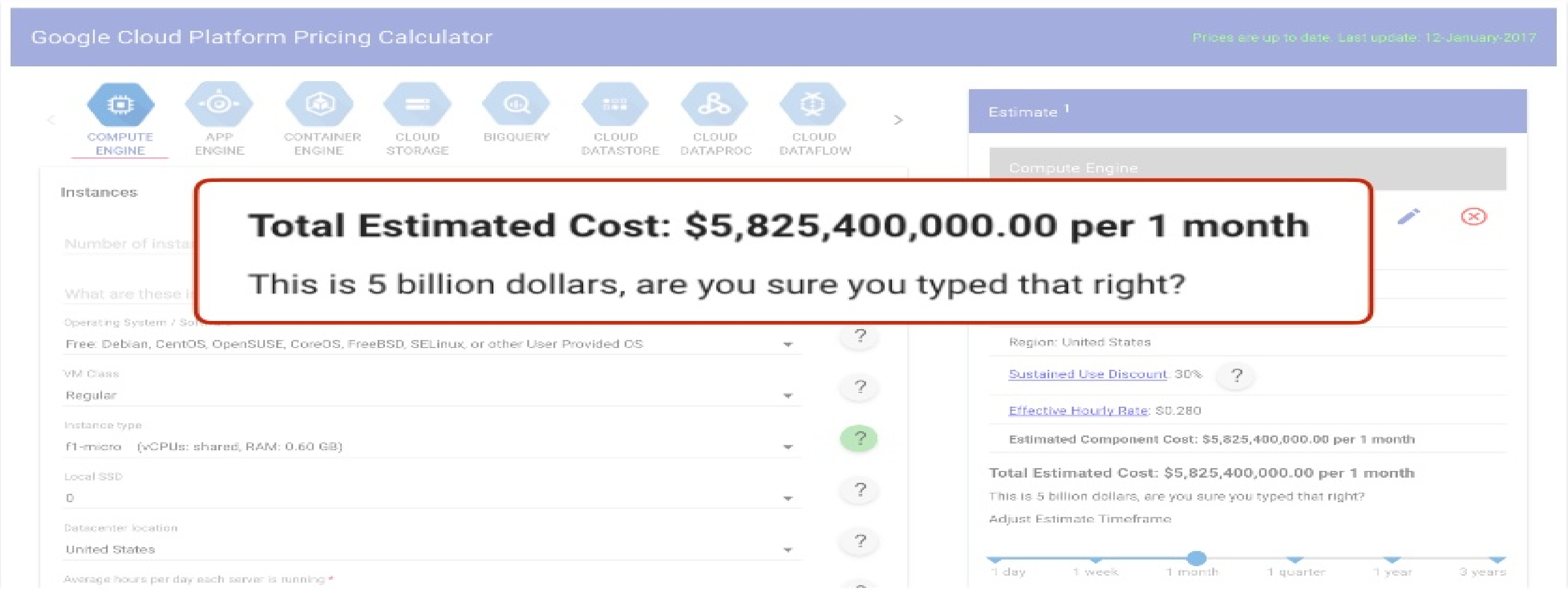
Total Estimated Cost: $5,825,400,000.00 per 1 month

This is 5 billion dollars, are you sure you typed that right?

Adjust Estimate Timeframe

1 day · 1 week · 1 month · 1 quarter · 1 year · 3 years

2 Trillion Compute Hours



Instances

Number of insta

What are these i

Operating System / Software
Free: Debian, CentOS, OpenSUSE, CoreOS, FreeBSD, SELinux, or other User Provided OS

VM Class
Regular

Instance type
f1-micro (vCPUs: shared, RAM: 0.60 GB)

Local SSD
0

Datacenter location
United States

Average hours per day each server is running *

Region: United States

Sustained Use Discount: 30%

Effective Hourly Rate: $0.280

Estimated Component Cost: $5,825,400,000.00 per 1 month

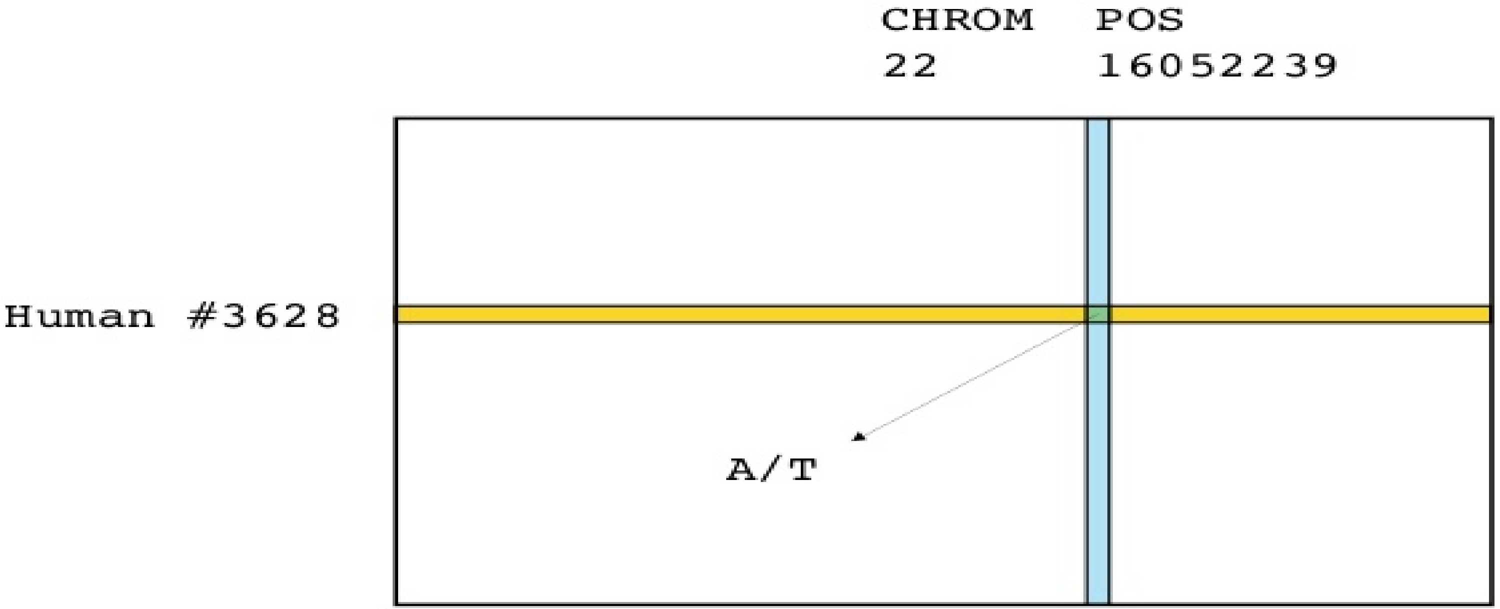
Total Estimated Cost: $5,825,400,000.00 per 1 month

This is 5 billion dollars, are you sure you typed that right?

Total Estimated Cost: $5,825,400,000.00 per 1 month

This is 5 billion dollars, are you sure you typed that right?

Structure of Sequence Data

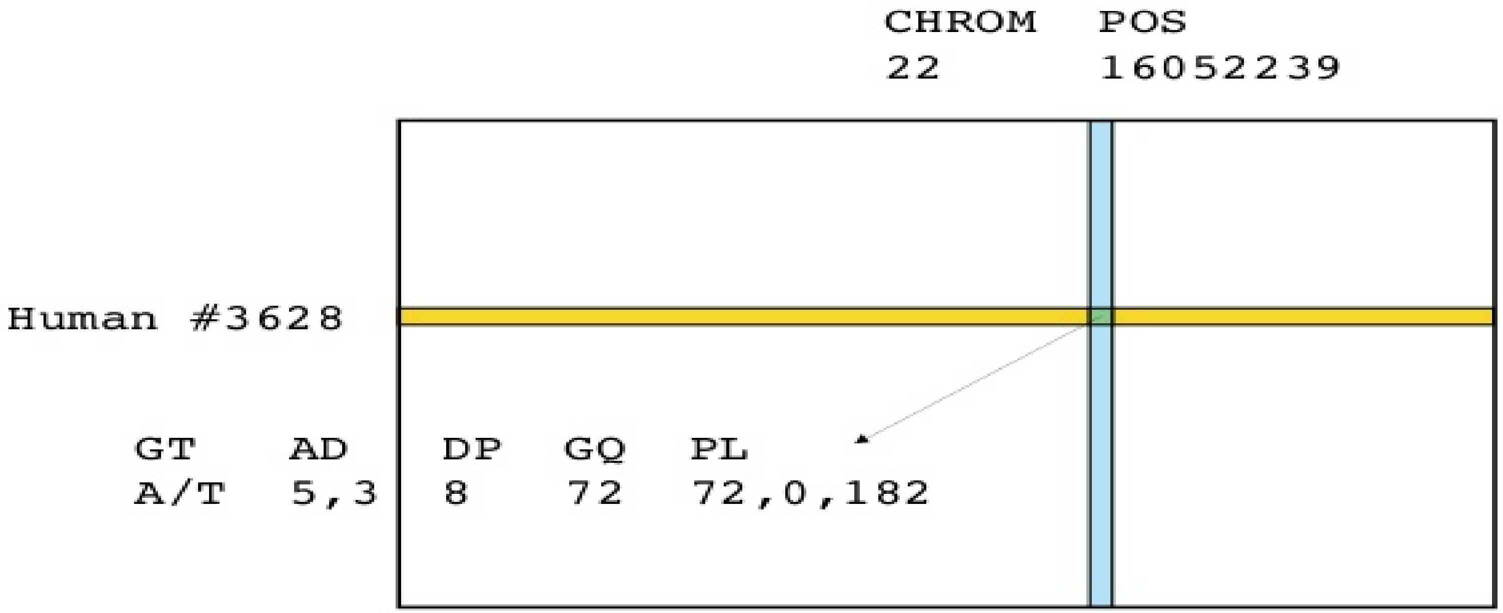


CHROM POS
22 16052239

Human #3628

A/T

Structure of Sequence Data



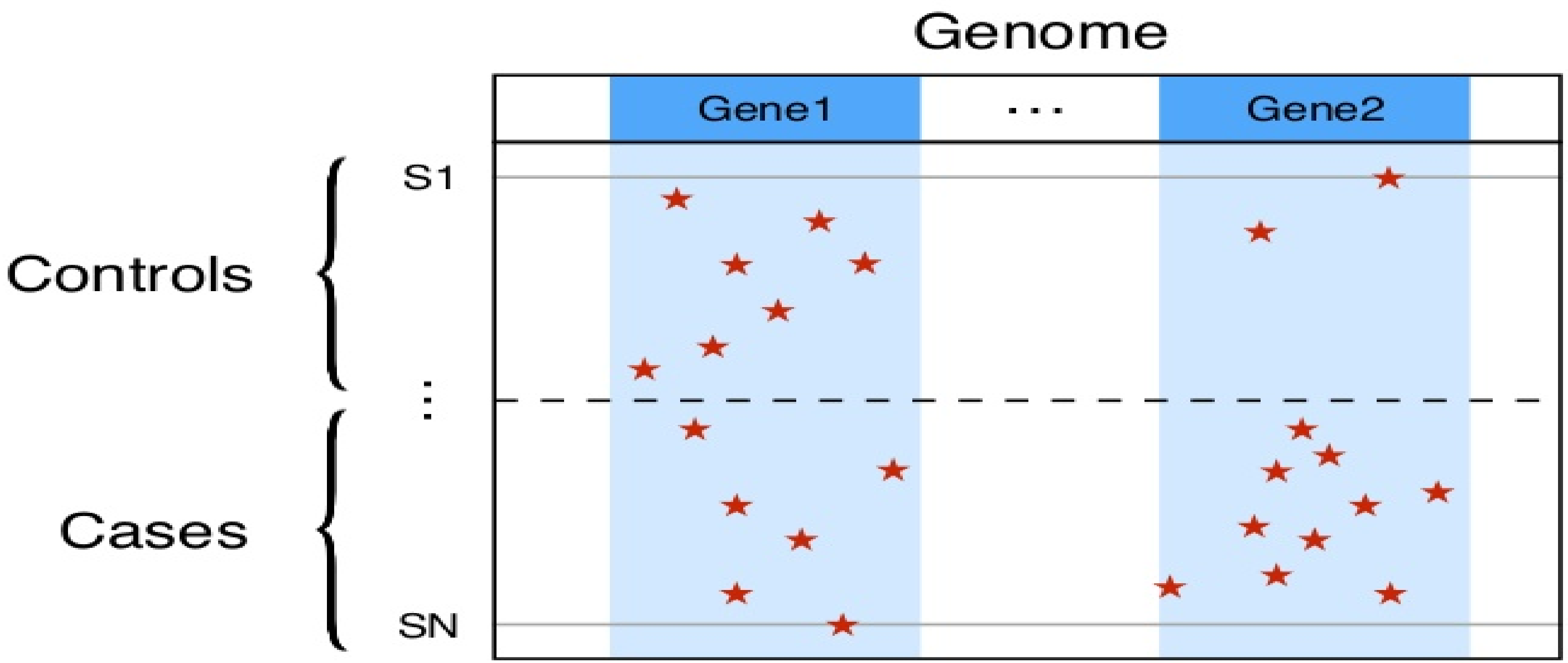
CHROM	POS
22	16052239

Human #3628

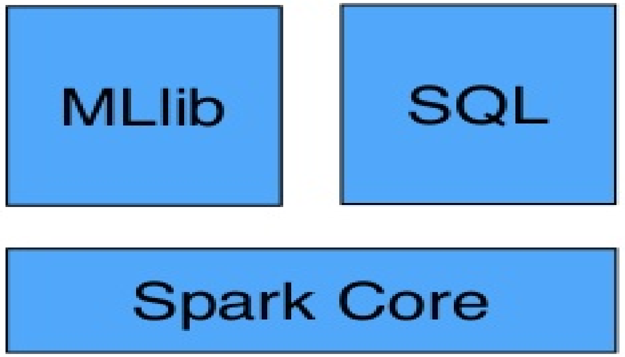
GT	AD	DP	GQ	PL
A/T	5,3	8	72	72,0,182

~100T records in current datasets

Genomic Association Analysis



- scalability
- high-level programming APIs
- linear algebra, MLlib
- Scala, python, R

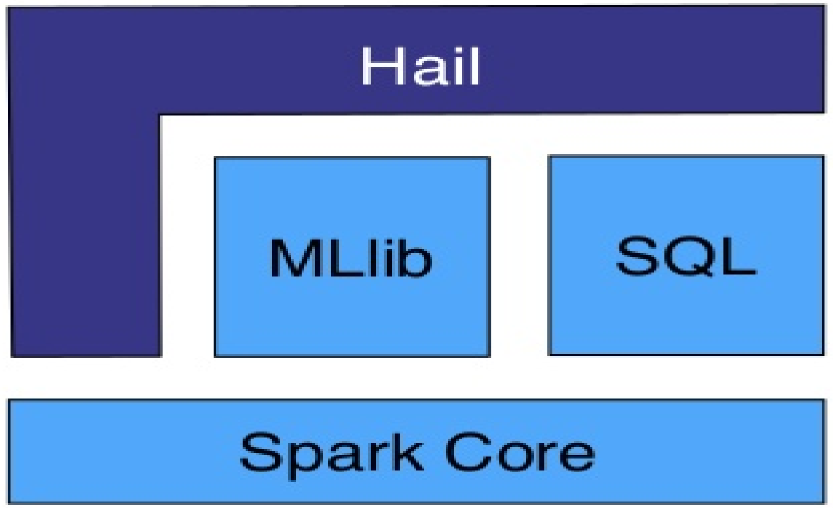


hail

- ingest genomic data
- high-level APIs for multi-dimensional data
- stats and ML methods
- Scala, python

Spark

- scalability
- high-level programming APIs
- linear algebra, MLlib
- Scala, python, R



mongoDB Solr cassandra APACHE KUDU Parquet hadoop

Ease of Use

"Hail democratizes big genetic data-analysis. You don't need to be a bioinformatician. You don’t have to know anything about parallel program execution. If you think you don't have the skills to use Hail then your only chance of actually doing any analysis with big sequence data IS Hail.”

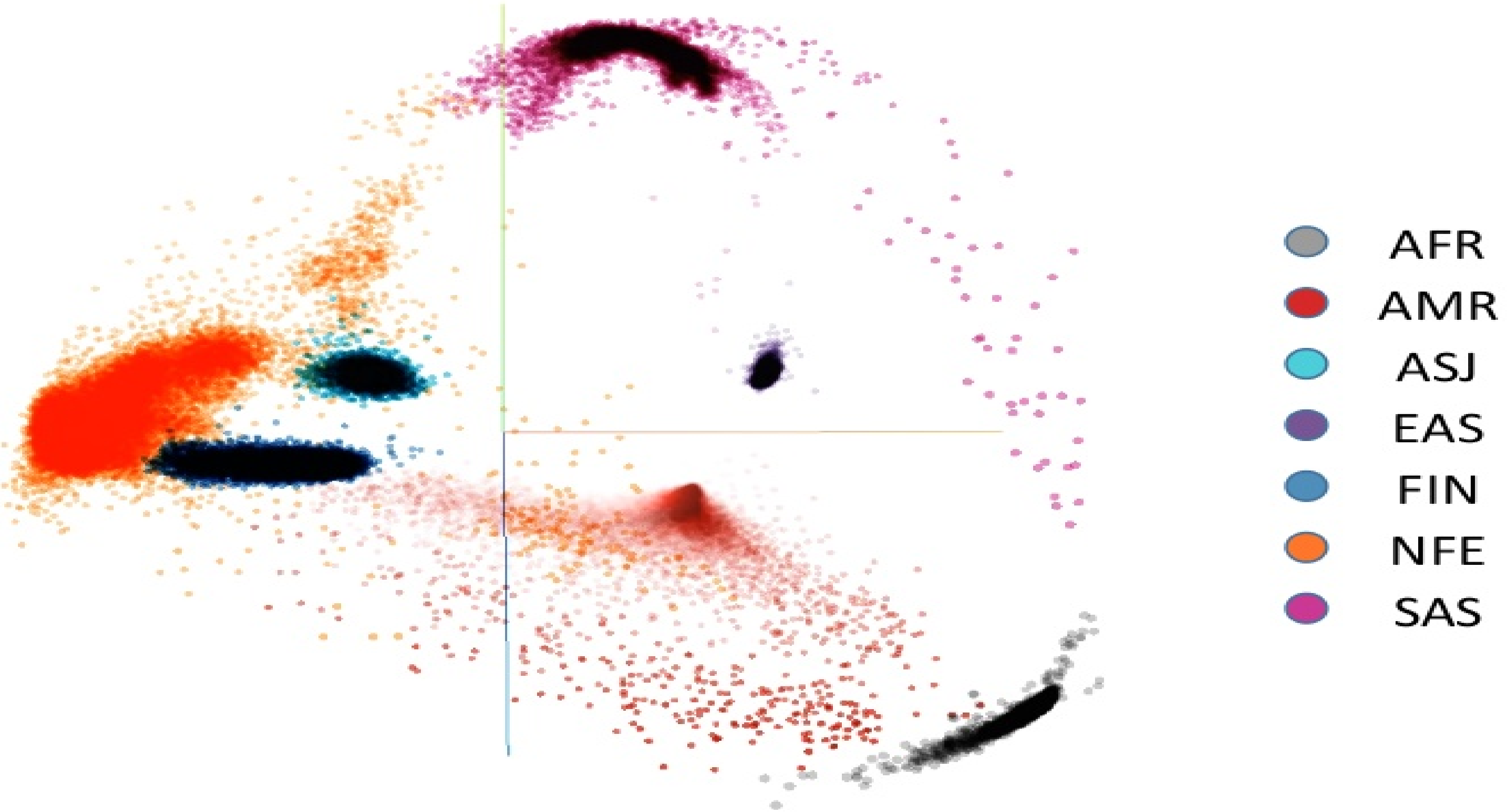
— Mitja Kurki, postdoc in statistical genetics, MGH

Hail Science

- L. Francioli, MacArthur lab, Analysis of whole-genome sequencing from 15,139 individuals
- A. Ganna et al., Ultra-rare disruptive and damaging mutations influence educational attainment in the general population, Nature Neuroscience
- A. Ganna et al., The impact of ultra-rare variants on human diseases and traits
- A. Ganna et al., The impact of rare variants on schizophrenia: whole genome sequencing of 10,000 individuals from the WGSPD consortia
- M. Kurki, Palotie Lab, Alzheimer's Disease Rare Variant Association Study in Finnish Founder Population
- M. Kurki, Palotie Lab, Genetic Architecture of Idiopathic Intellectual Disability in a Northern Finnish founder population cohort
- M. Kurki, P. Gormley, Palotie Lab, Genetic Architecture of Familial Migraine in a Family collection of 9000 Individuals in 2000 Families
- K. Karczewski, MacArthur Lab, The Human Knockout Project: analyzing loss-of-function variants across 126,216 individuals

Hail Science

- X. Li et al., Developing and optimizing a whole genome and whole exome sequencing quality control pipeline with 652 Genotype-Tissue Expression donors
- M. A. Rivas et al., Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population
- K. Satterstrom, iPSYCH-Broad Consortium, Rare variants conferring risk for autism identified by whole exome sequencing of dried bloodspots
- C. Seed et al., Neale Lab, Hail: An Open-Source Framework for Scalable Genetic Data Analysis
- G. Tiao, Pan-Cancer Analysis of Whole Genomes, Analysis of rare variation in 2,818 whole-genome germline samples from cancer patients
- S. Maryam Zekavat, P. Natarajan, Kathiresan Lab. An analysis of deep, whole-genome sequences and plasma lipids in ~16,000 multi-ethnic samples.
- S. Maryam Zekavat, Kathiresan Lab. An analysis of deep, whole-genome sequences and coronary artery disease in ~7,000 multi-ethnic samples.
- S. Maryam Zekavat, Kathiresan Lab. Analyzing the full spectrum of genomic variation with Lp(a) Cholesterol: Novel insights from deep, whole genome sequence data in 5,192 Europeans and African Americans from Estonia and from the Jackson Heart Study



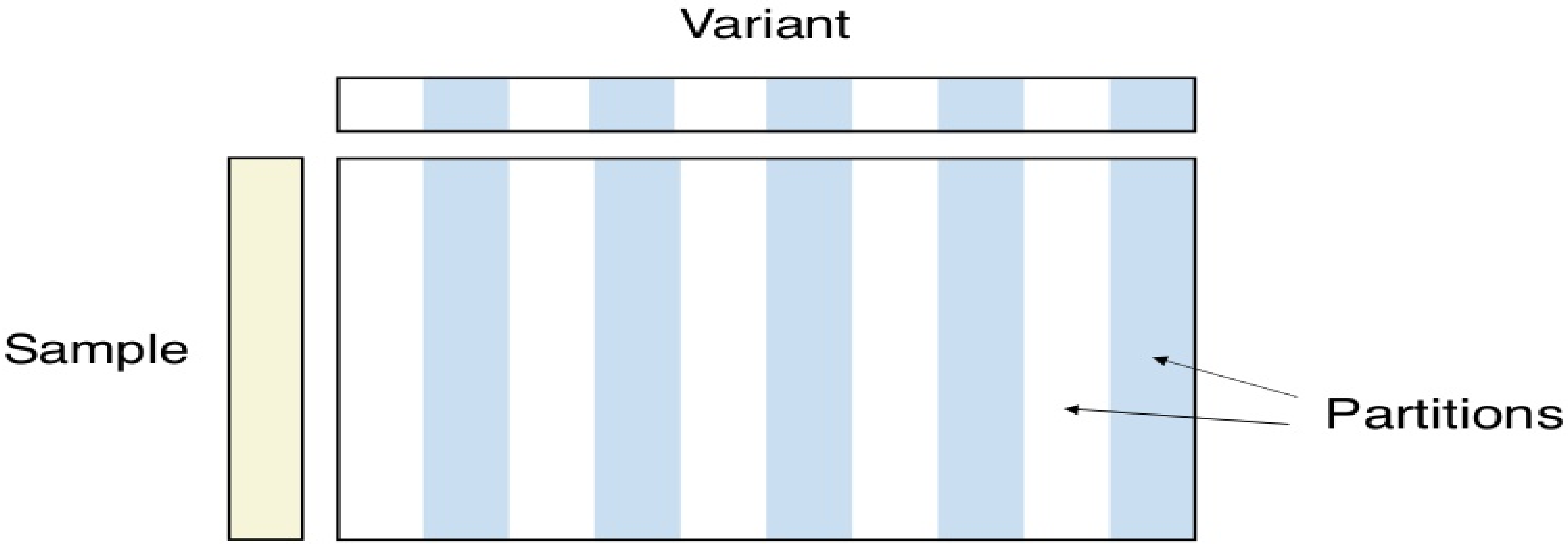
Genome Aggregation Database (gnomAD)

Genome Aggregation Database (gnomAD)

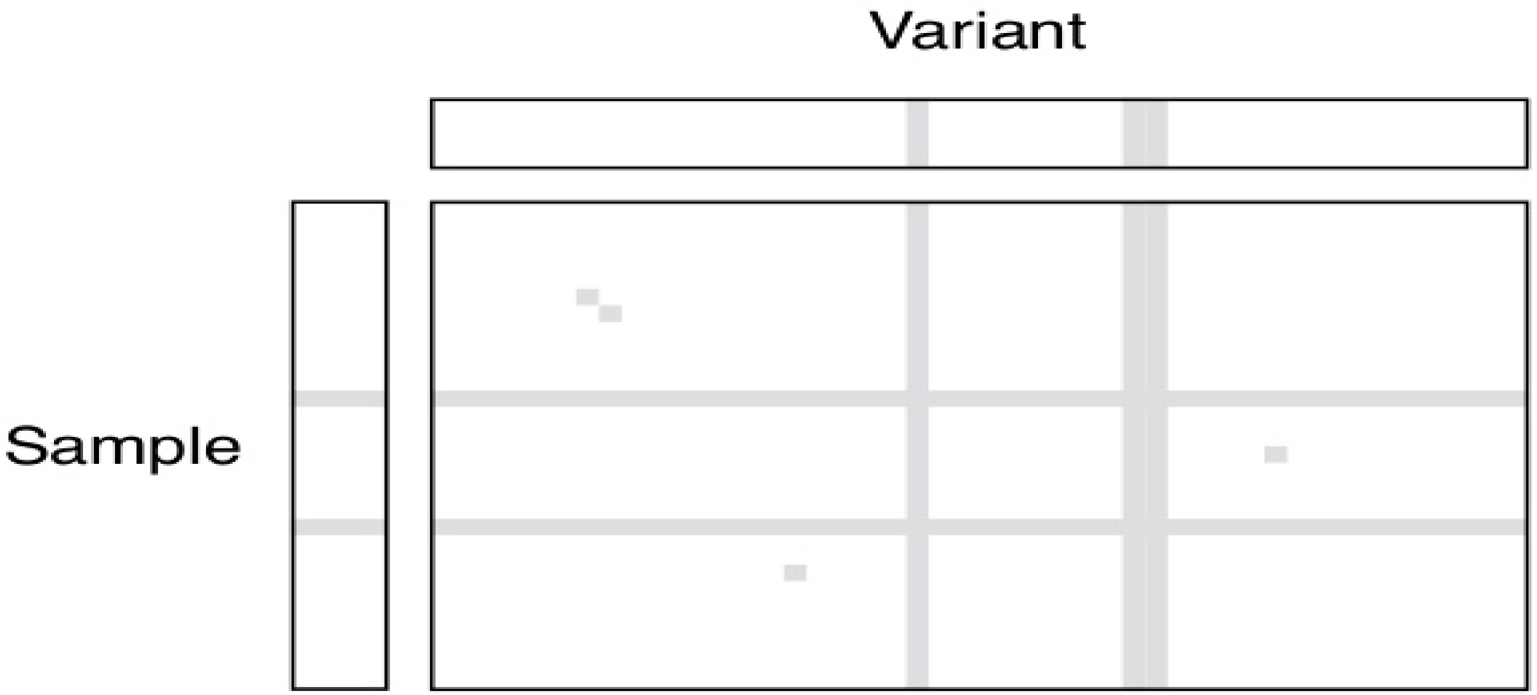
- Successor to ExAC
- Public resource: http://gnomad.broadinstitute.org
- ~6M hits in last year
- >140K people, ~280TB VCF
- Flexibility and speed enable rapid iteration on analysis
- Raw data to initial release in 1 week with Hail

"Without Hail, we would have been totally screwed."
— Daniel MacArthur

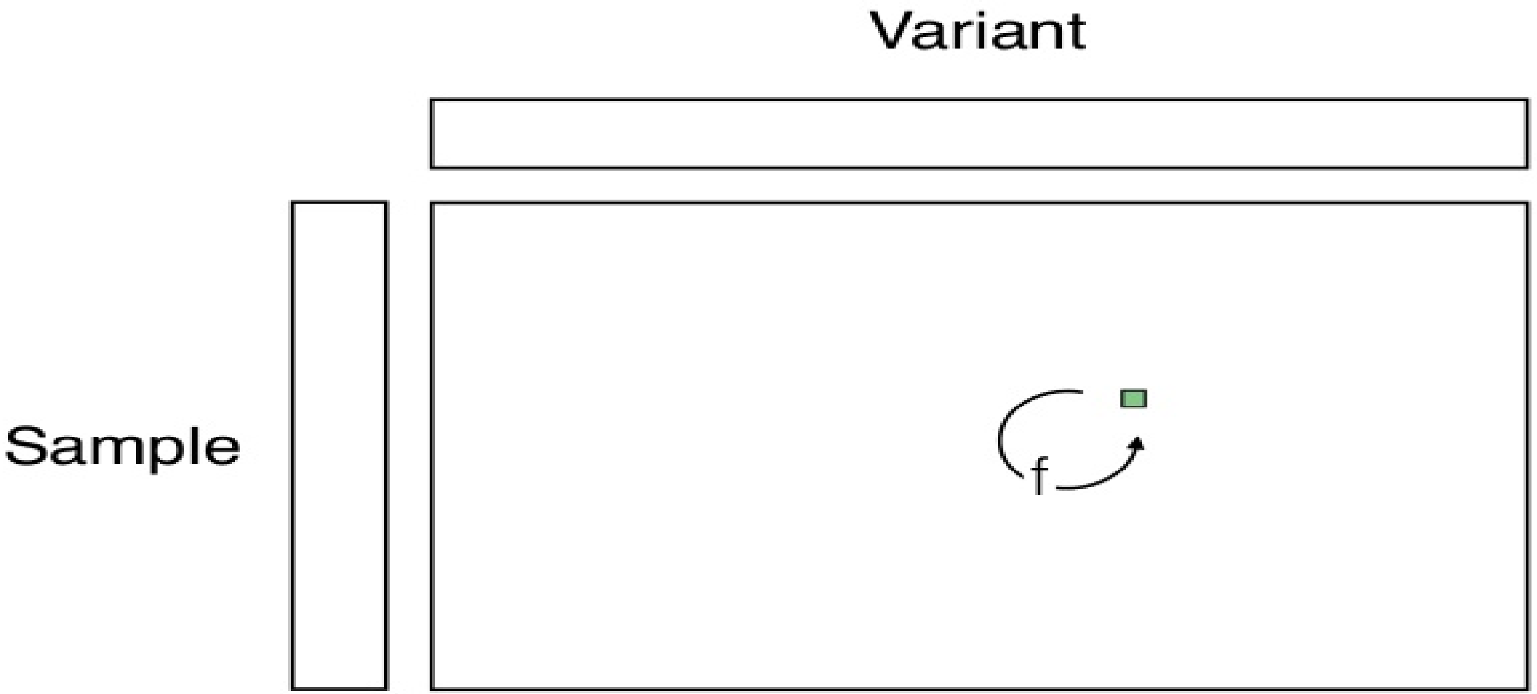
Variant-Sample Matrix



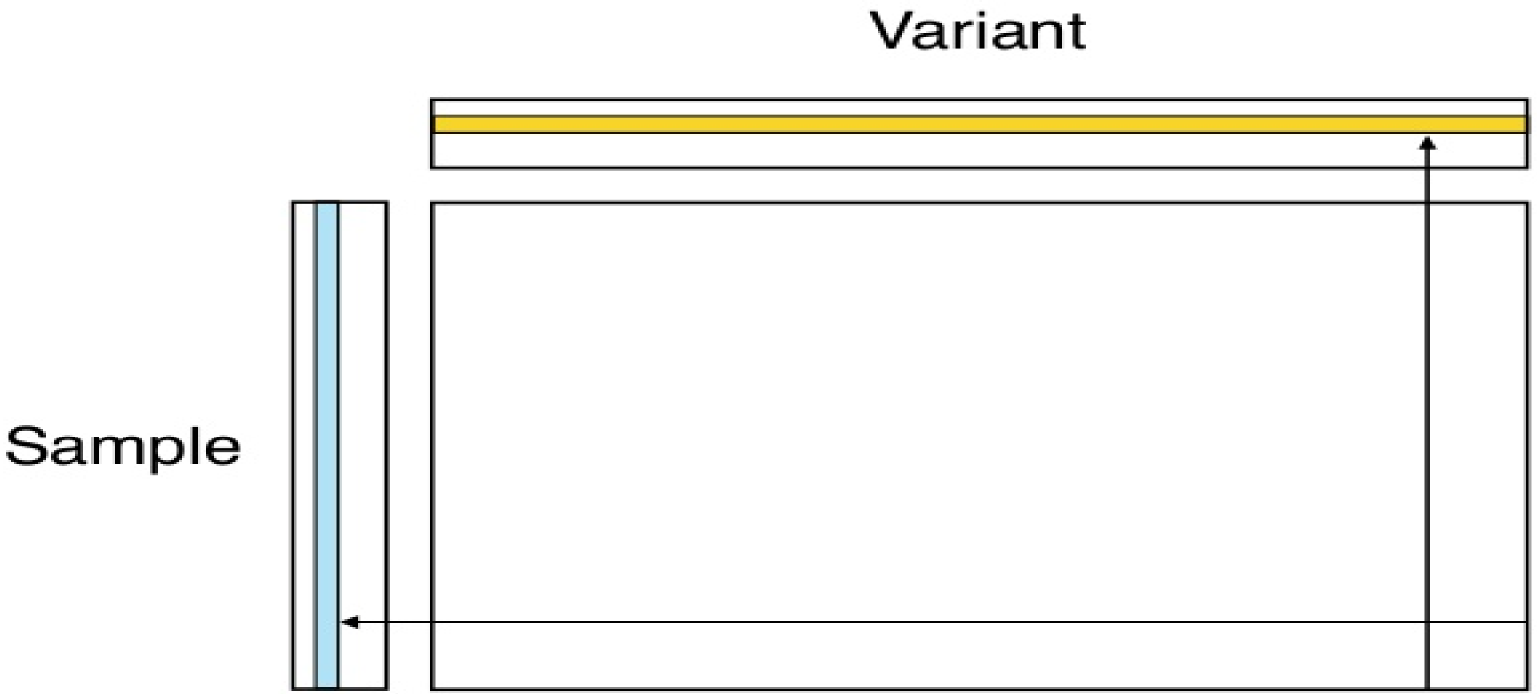
`filter`



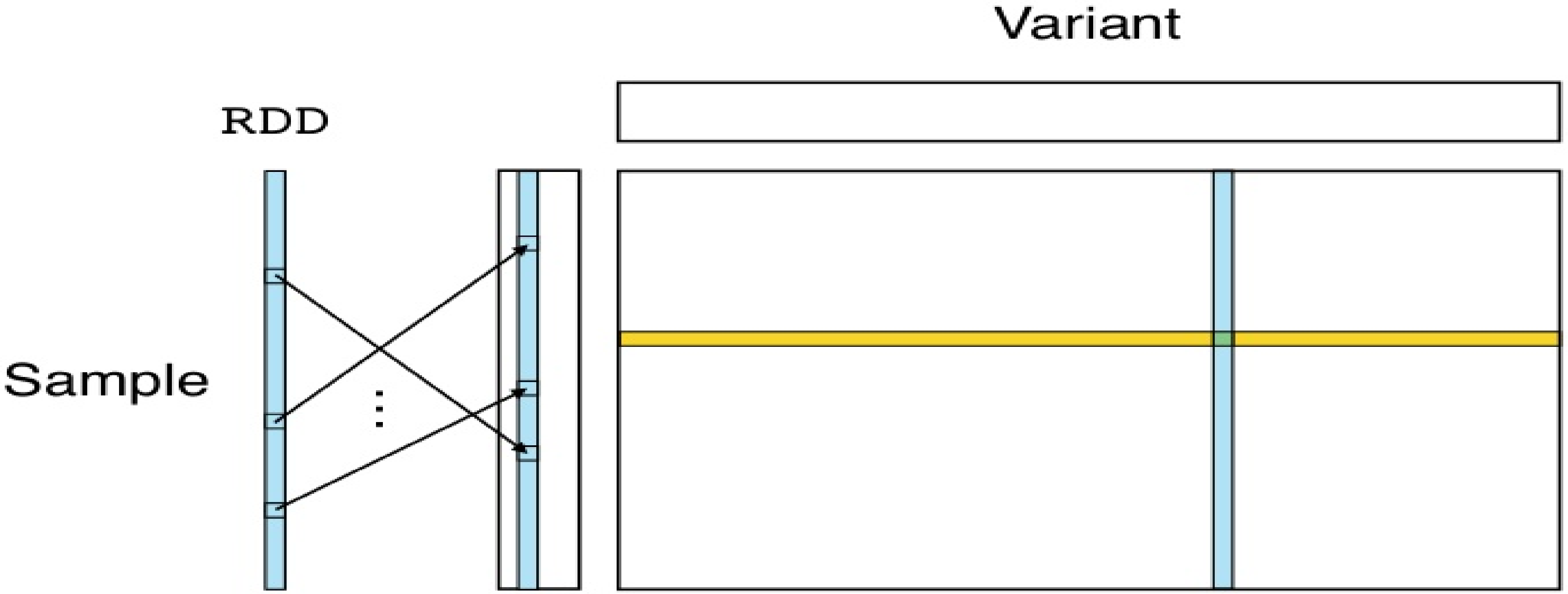
map



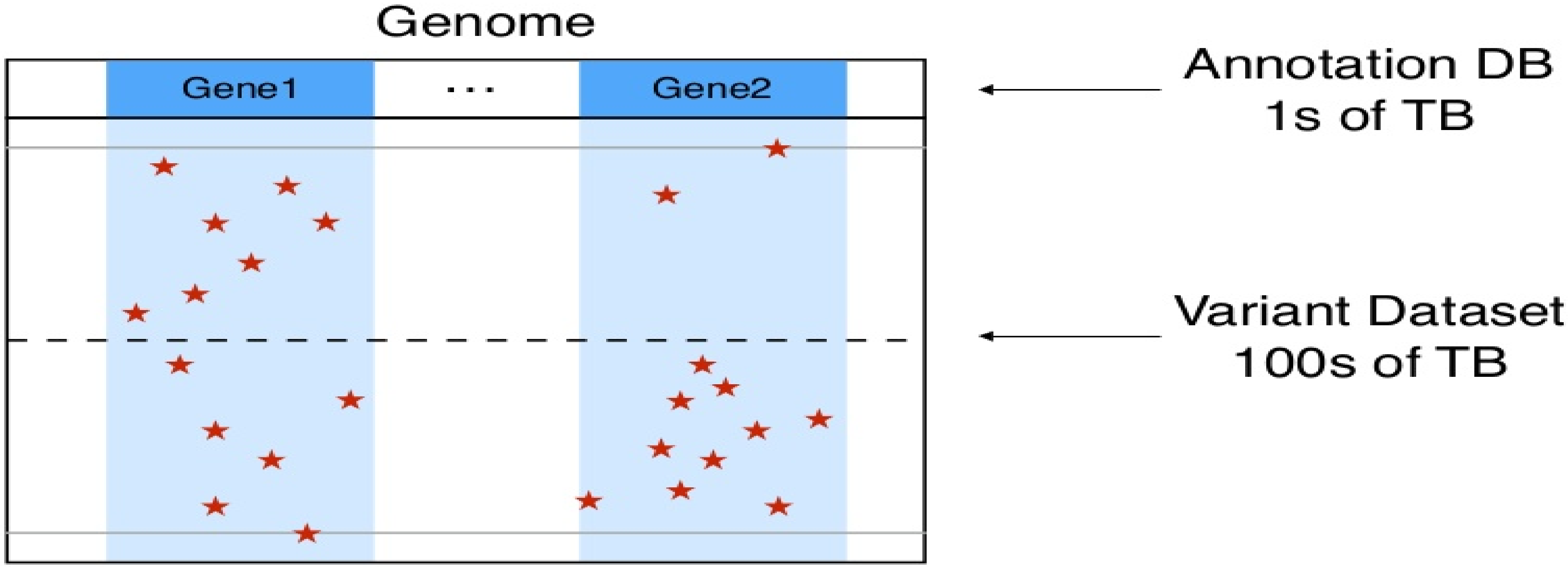
reduce



join



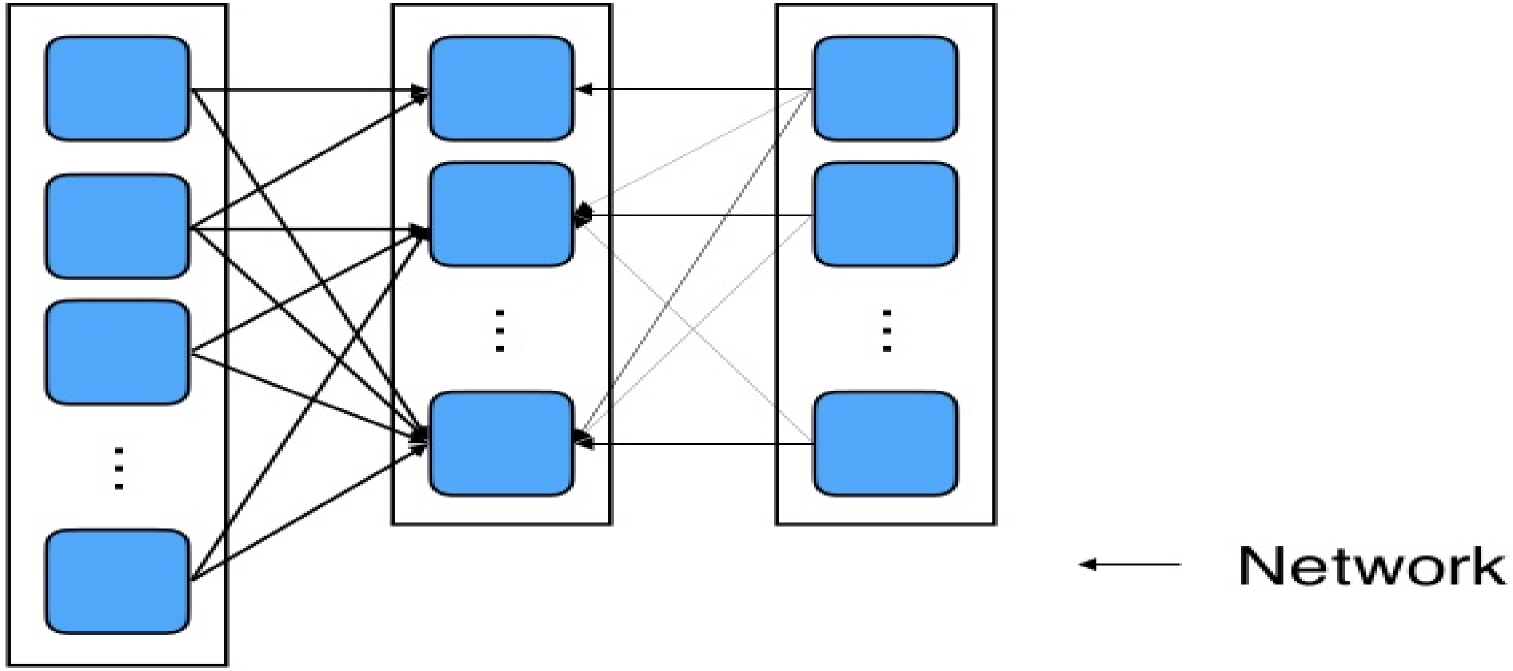
OrderedRDD motivation



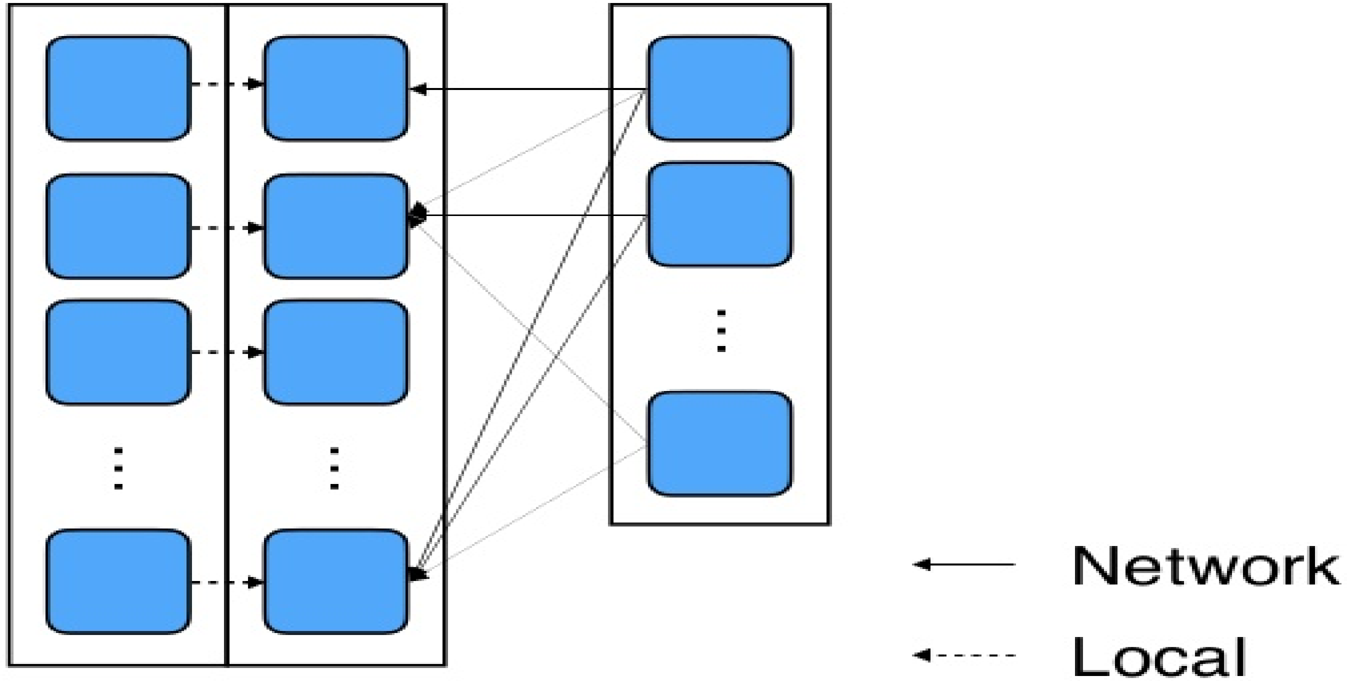
OrderedRDD

- Generalizes Spark's `RangePartitioner`.
- Partitioning preserved through `write`/`read`.
- Support range join.
- Push predicates through partitioning.

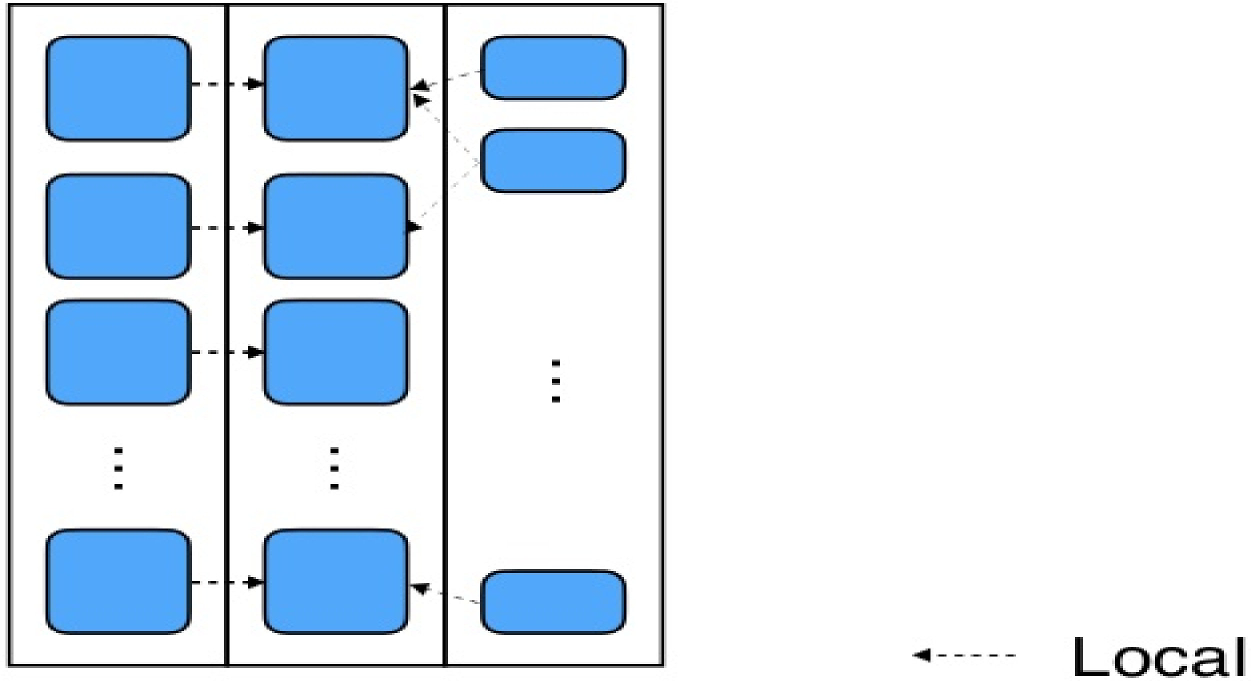
Join



Partitioned Join



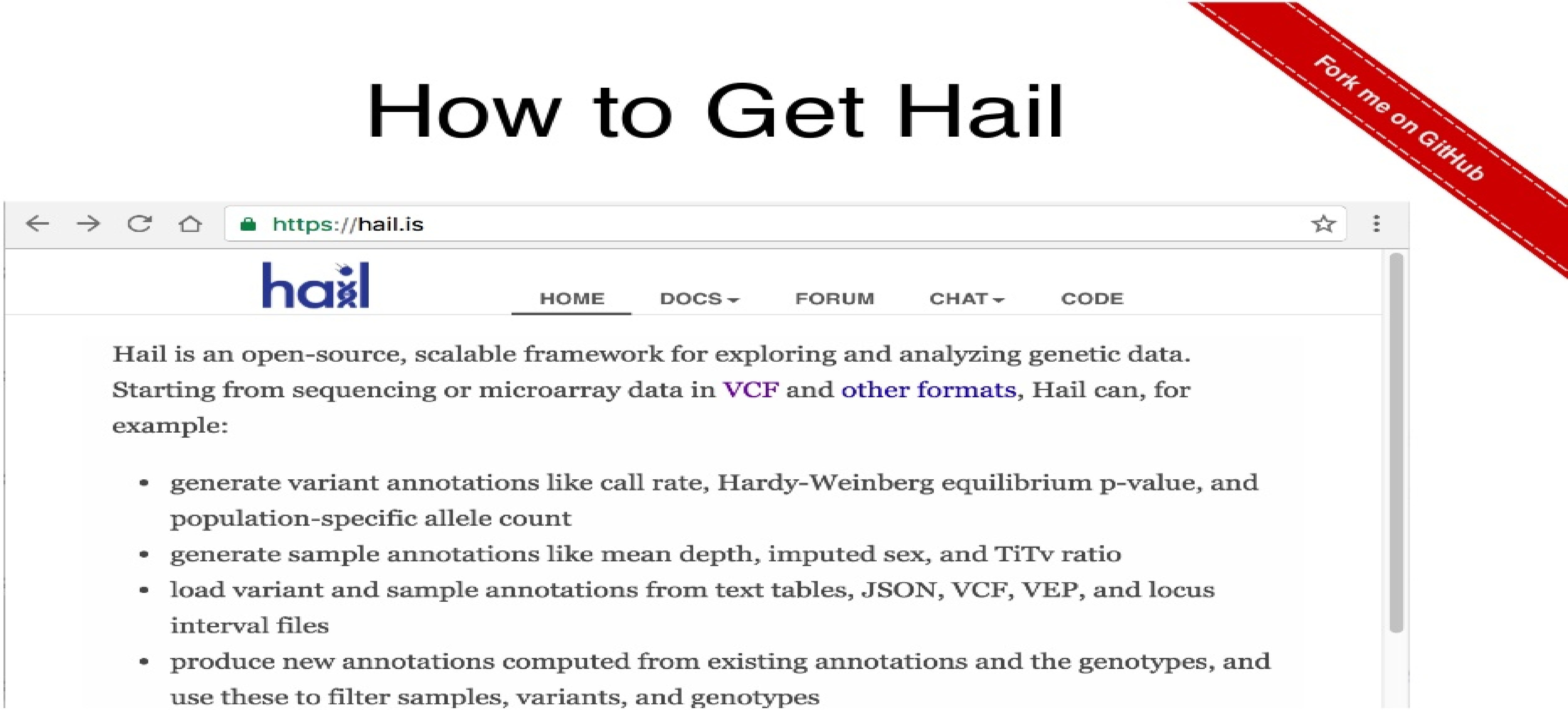
Range Join



Future Directions

- Ontogeny recapitulates phylogeny: `DataFrame`/`Dataset`-like APIs to take advantage of Spark 2 performance improvements.
- Need partitioned `DataSources`.
- Separate general-purpose abstractions from genetic-specific code to make available to wider Spark community.
- Versioned release.
- Domain-specific genetics functionality, of course…

How to Get Hail



Hail is an open-source, scalable framework for exploring and analyzing genetic data. Starting from sequencing or microarray data in VCF and other formats, Hail can, for example:

- generate variant annotations like call rate, Hardy-Weinberg equilibrium p-value, and population-specific allele count
- generate sample annotations like mean depth, imputed sex, and TiTv ratio
- load variant and sample annotations from text tables, JSON, VCF, VEP, and locus interval files
- produce new annotations computed from existing annotations and the genotypes, and use these to filter samples, variants, and genotypes

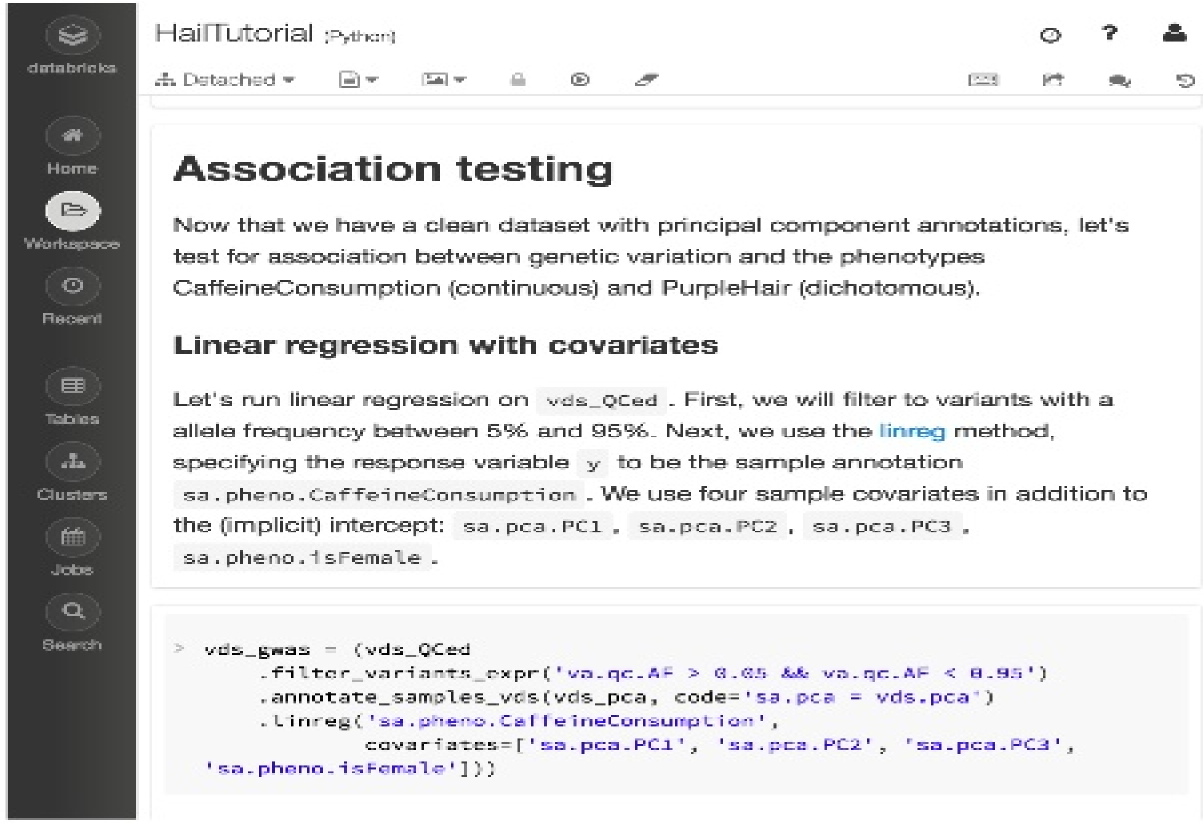
databricks

hail

Try Hail on Databricks!

Sign up for your Databricks free trial at:
https://accounts.cloud.databricks.com/registration.html#signup

Import the Hail tutorial notebook here:
https://docs.databricks.com/spark/latest/training/1000-genomes.html



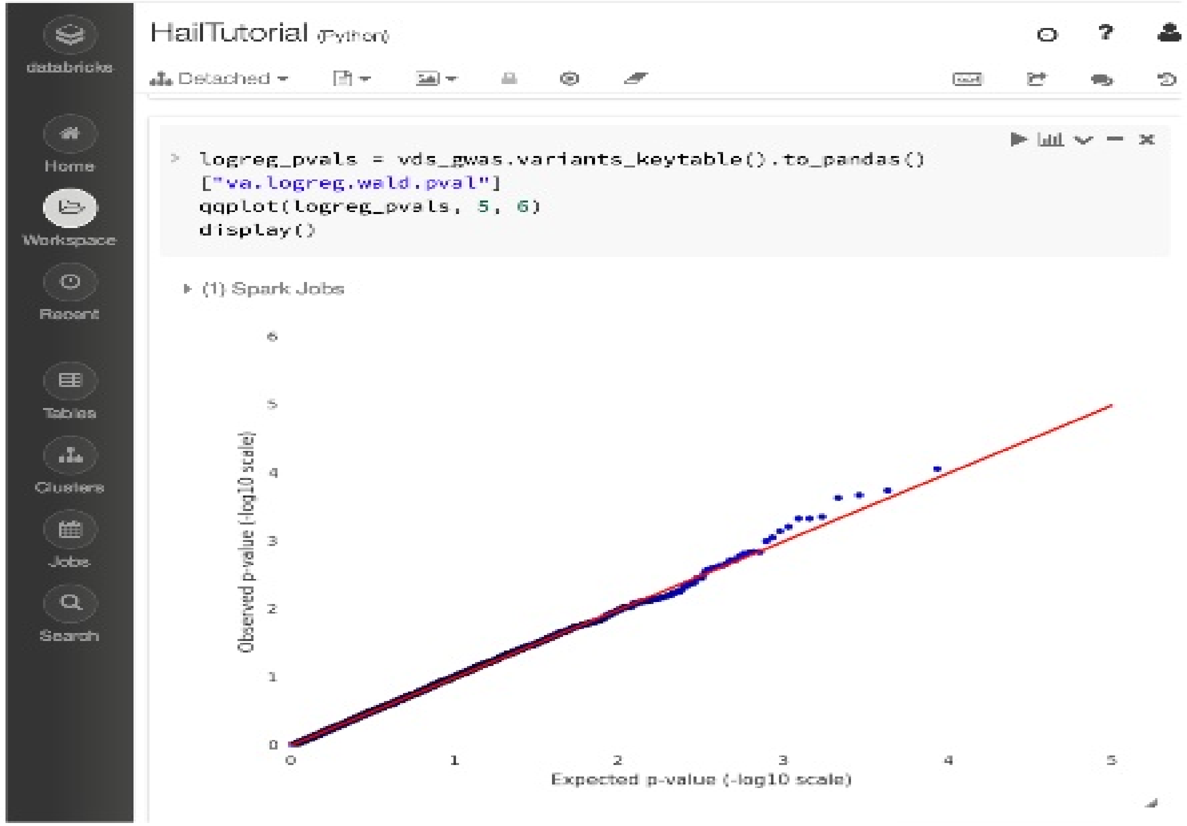
Association testing

Now that we have a clean dataset with principal component annotations, let's test for association between genetic variation and the phenotypes CaffeineConsumption (continuous) and PurpleHair (dichotomous).

Linear regression with covariates

Let's run linear regression on `vds_QCed`. First, we will filter to variants with a allele frequency between 5% and 95%. Next, we use the linreg method, specifying the response variable `y` to be the sample annotation `sa.pheno.CaffeineConsumption`. We use four sample covariates in addition to the (implicit) intercept: `sa.pca.PC1`, `sa.pca.PC2`, `sa.pca.PC3`, `sa.pheno.isFemale`.

```
> vds_gwas = (vds_QCed
      .filter_variants_expr('va.qc.AF > 0.05 && va.qc.AF < 0.95')
      .annotate_samples_vds(vds_pca, code='sa.pca = vds.pca')
      .linreg('sa.pheno.CaffeineConsumption',
              covariates=['sa.pca.PC1', 'sa.pca.PC2', 'sa.pca.PC3',
  'sa.pheno.isFemale']))
```

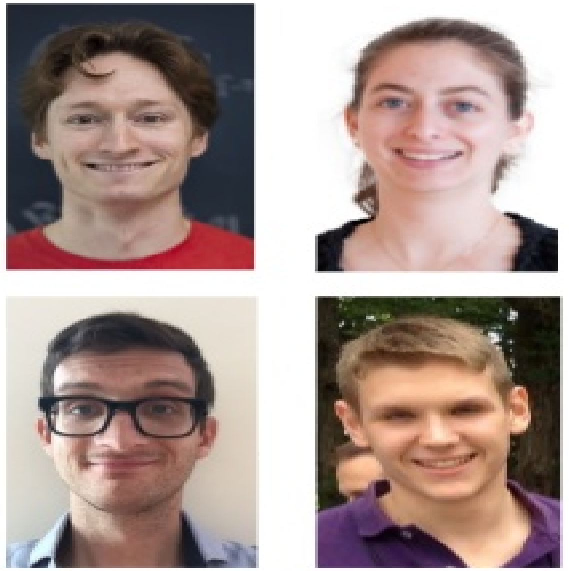


```
> logreg_pvals = vds_gwas.variants_keytable().to_pandas()
  ["va.logreg.wald.pval"]
  qqplot(logreg_pvals, 5, 6)
  display()
```

▸ (1) Spark Jobs

6
5
4
3
2
1
0
Observed p-value (-log10 scale)
0
1
2
3
4
5
Expected p-value (-log10 scale)

Thank You



Hail Team

Jon Bloom
Jackie Goldstein
Daniel King
Tim Potherb

Contributors

Szabolcs Berecz
Alex Bloemendal
John Compitello
Laurent Francioli
Konrad Karczewski
Jack Kosmicki
Mitja Kurki
Mark Pinese
Ben Weisburd
Tom White
Alex Zamoshchin
@shusson

And **Prof. Ben Neale** and scientific collaborators.