Welcome! While waiting for our session to start:

- Please ensure that your microphone is muted during the presentation. **<u>But</u>** we'd love if you could unmute yourself temporarily (by pressing the *spacebar* or CMD+A):
  - To giggle or laugh (we think the presenters may be funny)
  - To comment / ask questions

- If you would like to turn on your video, great! It would be nice to see everyone. Otherwise, we respect your privacy and prerogative ☺

- Issues with the Zoom? Please use Slack or the zoom chat box. Arcturus and I will check it periodically.

twitter @mkveerapen / @hailgenetics
veerapen@broadinstitute.org
#ATGUstrong

Scalable Genomics for Common Variants

Broad E Workshop 2021

April 8th, 2021

12:00 – 3:00 PM (EST)

Zoom

Kumar Veerapen, PhD
Hail Support and Community Outreach Manager
Tim Poterba and Carolin Diaz
Software Engineer

https://hail.is
@mkveerapen / @hailgenetics
veerapen@broadinstitute.org
#scalableGenomics
#hailGenetics #ATGUstrong

Outline

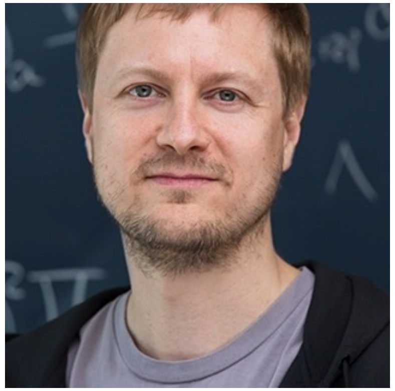
- Who are we?
- Who are you?
- What is Hail?
- Why Hail?
- How can you use Hail?

BROAD INSTITUTE

The Hail Team is a systems engineering team building tools to accelerate biological research.

BROAD INSTITUTE

Hail Team


Cotton Seed, PhD
Team Leader

Tim Poterba

Dan King

Jackie Goldstein

Daniel Goldstein

Patrick Schultz, PhD

Whitney Wade
Operations

Kumar Veerapen, PhD
Support and Outreach

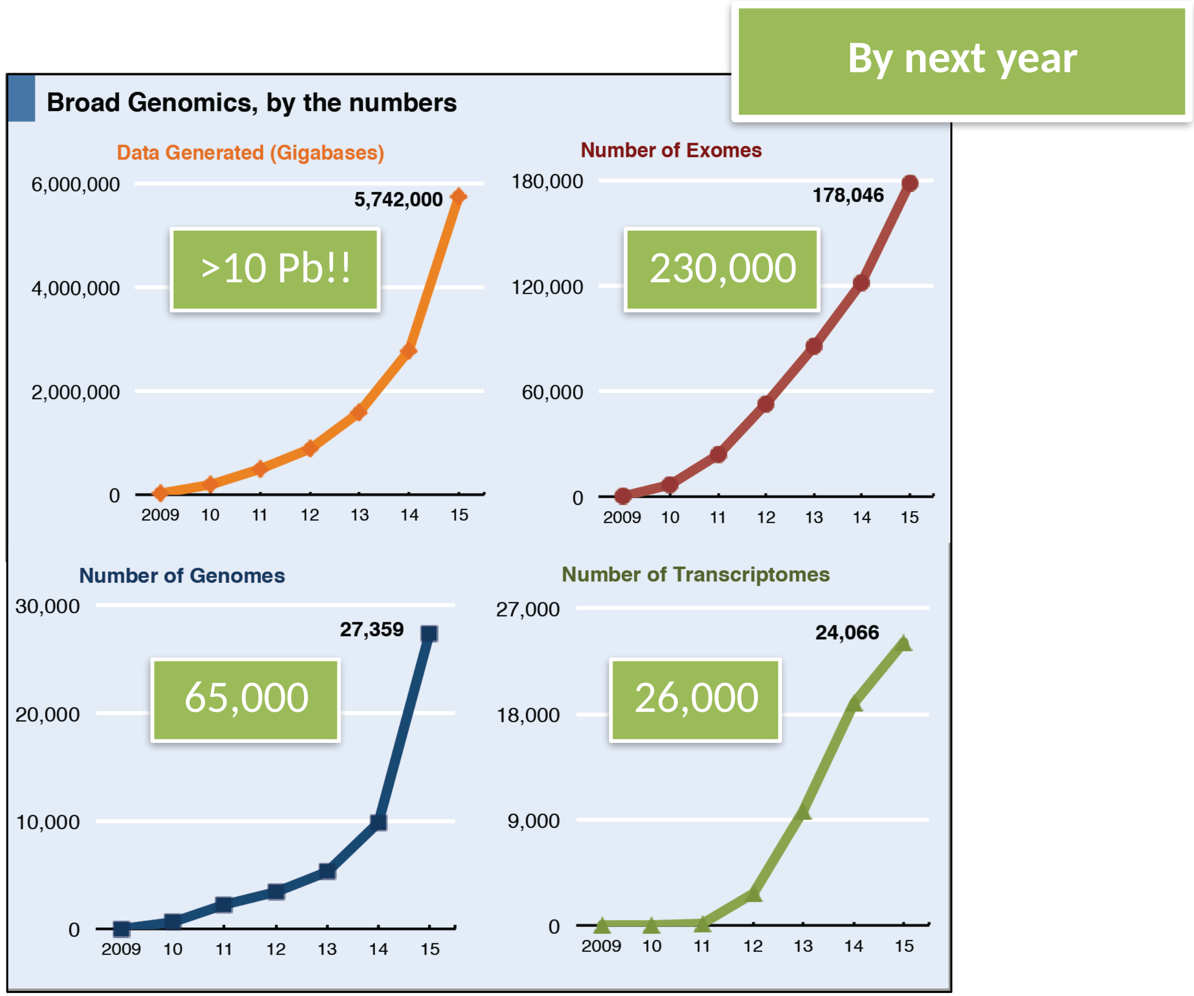
John Compitello

Carolin Diaz

Chris Vittal

Patrick Cummings

Accelerating Genomic Data e.g. Call Sets, variant files etc



By next year

Broad Genomics, by the numbers

Data Generated (Gigabases)

>10 Pb!!

Number of Exomes

230,000

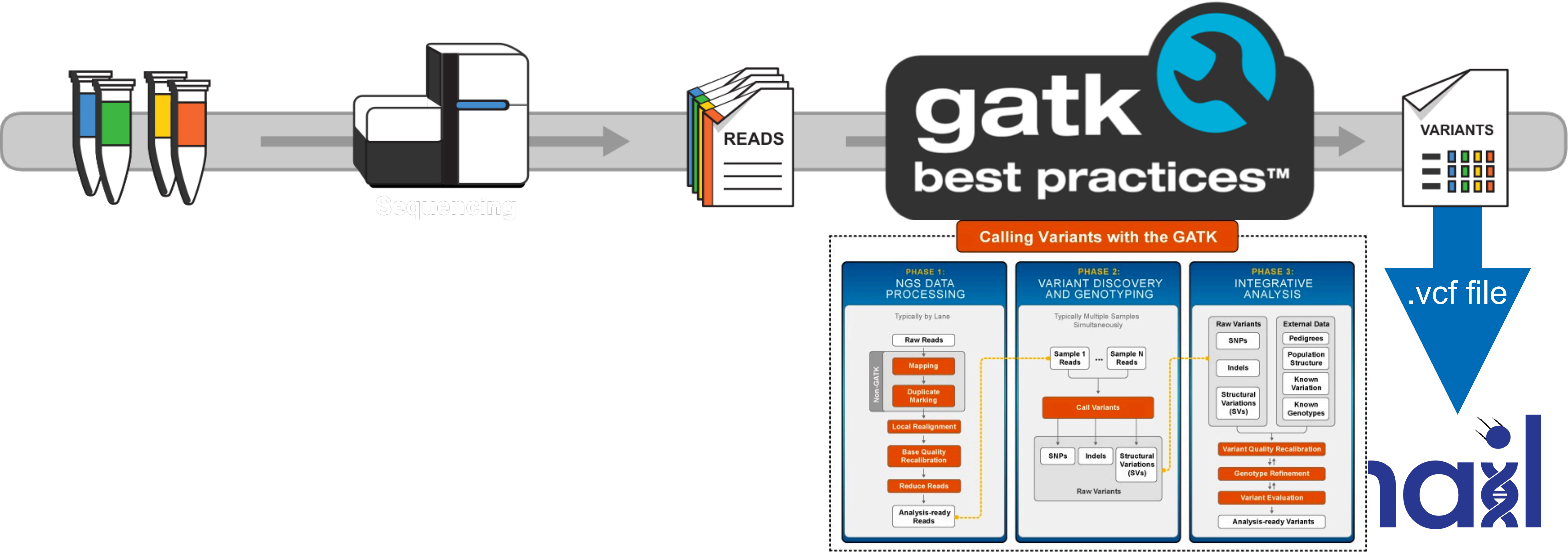
Number of Genomes

65,000

Number of Transcriptomes

26,000

What is Hail’s role in callset generation?



What is Hail?

“On a scale from zero to dplyr, the Hail 0.2 interface scores an 8/10 for general-purpose data analysis.” - Konrad K., lead analyst, gnomAD

hail

Learn more at Hail.is

***We can’t read your minds, so talk to us**
discuss.hail.is

Open-Source Library

Genomic analysis at every scale

Explore Biobank Scale Data

Interrogation of **biobank scale** genomic data

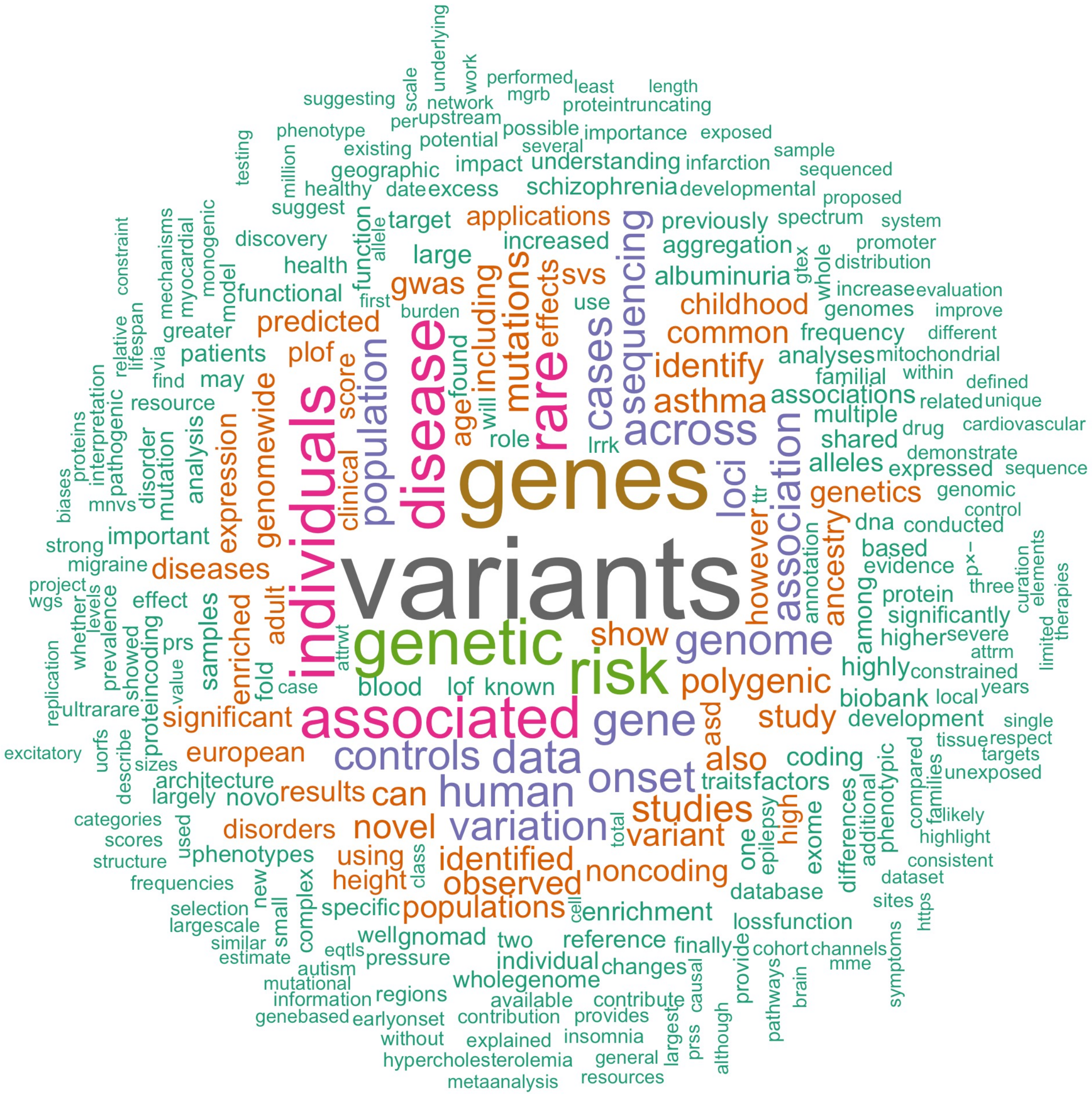
Modern Data Scaling

Efficient genomic data frame **scalability** using Hail MatrixTables.

Unified Input Platform

Tabular data frames imported as Hail MatrixTables into **unified platform**.

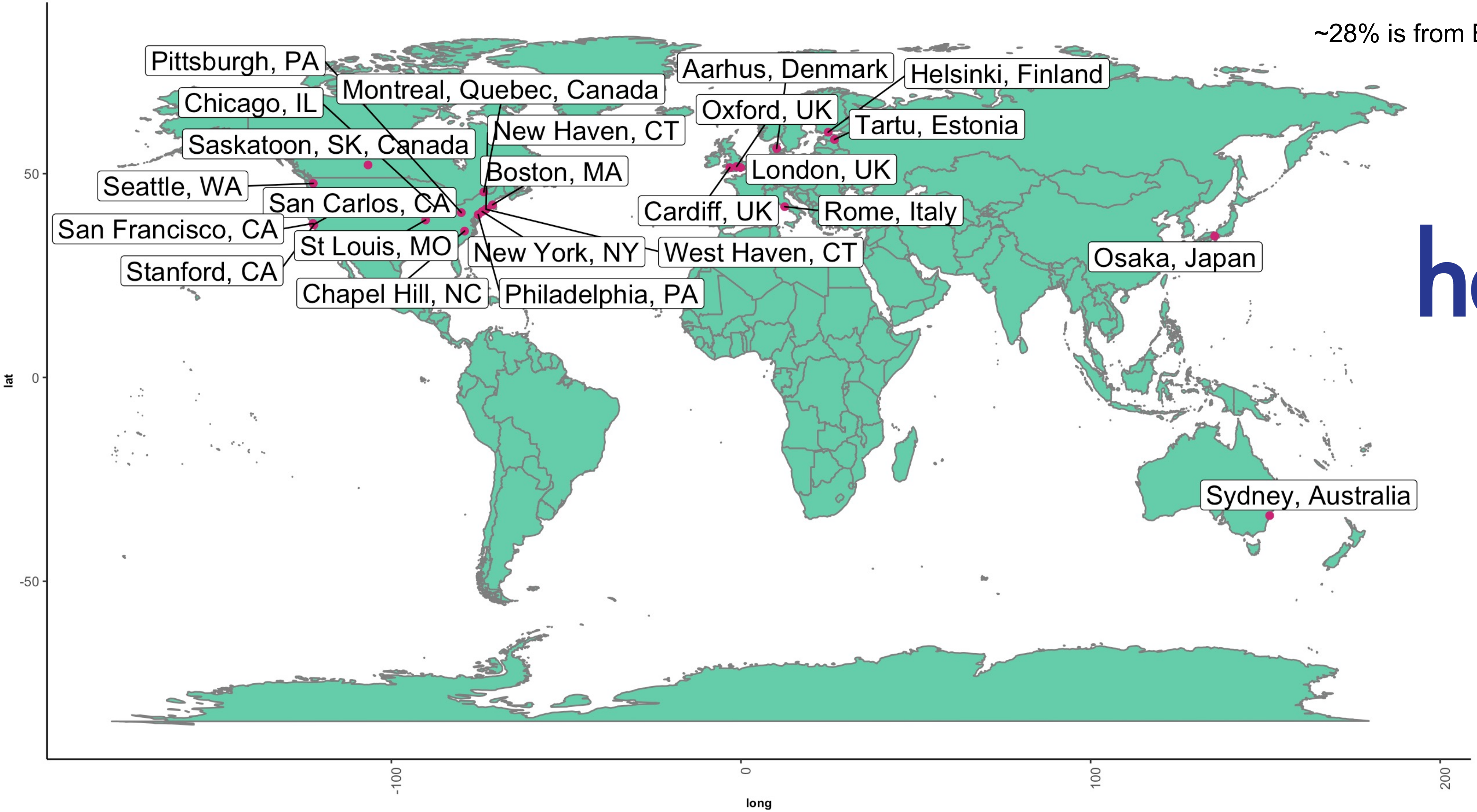
How has Hail been used? (hail.is/references.html)



Notes:

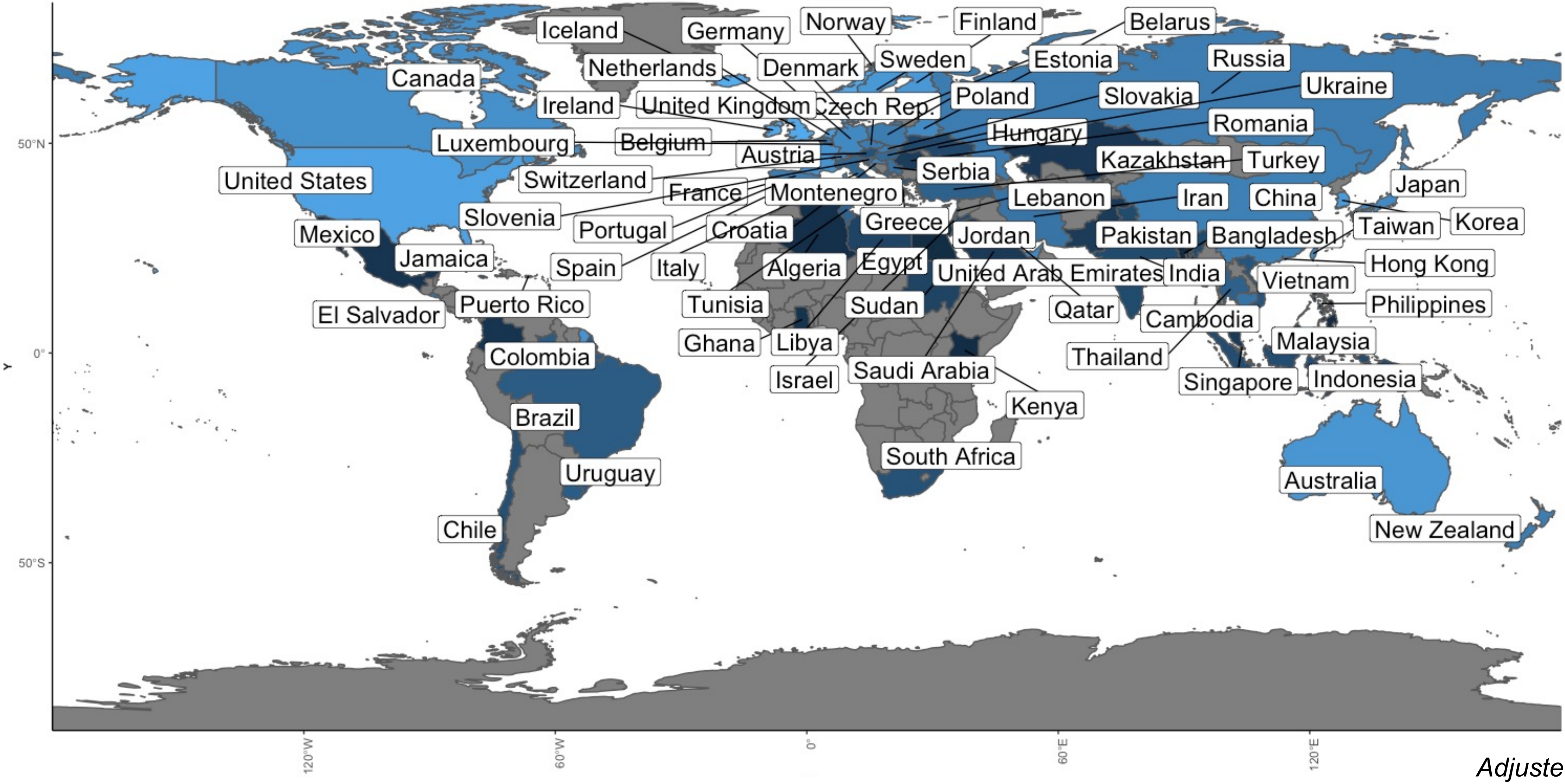
- •51 abstracts (07/20/2020)
- •Word appearing > 4x

Where has Hail been used?



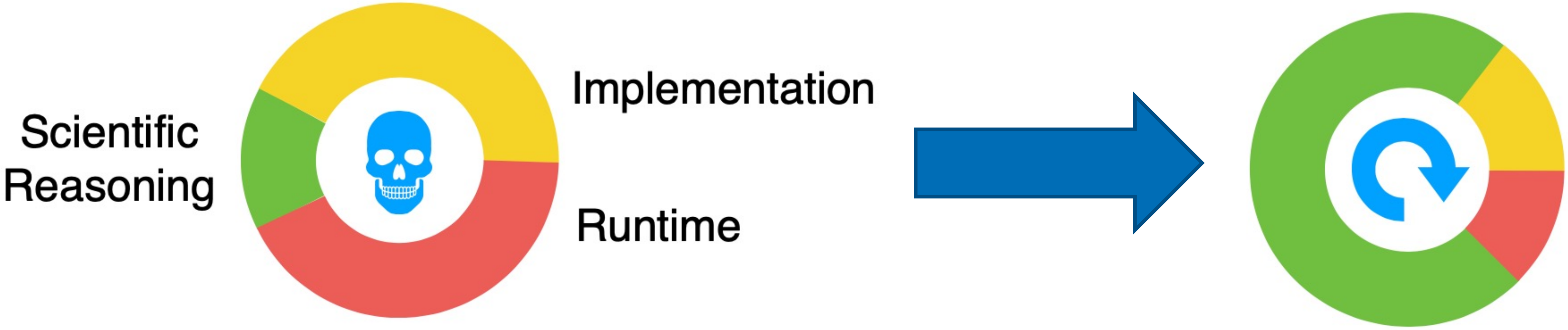
~28% is from Boston, MA

Where in the world has Hail been “pip”-ed a.k.a. downloaded?



Adjusted for total population

Why would you use Hail?



Scientific Reasoning

Implementation

Runtime

Hail as a data science library

Data slinging

Analytical toolbox

Hail as a data science library

Data slinging

Analytical toolbox

- **Read and write common formats**
- Filter, group, aggregate
- Annotation
- Visualization

VCF	TSV
BGEN	PLINK
JSON	GEN
BED	GTF

Hail as a data science library

Data slinging

Analytical toolbox

- Read and write common formats
- **Filter, group, aggregate**
- Annotation
- Visualization

- Compute mean depth per variant or per sample
  - Among heterozygotes
  - Grouped by ancestry labels & sex
- Count transitions & transversions called per sample

Hail as a data science library

Data slinging

Analytical toolbox

- Read and write common formats
- Filter, group, aggregate
- **Annotation**
- Visualization

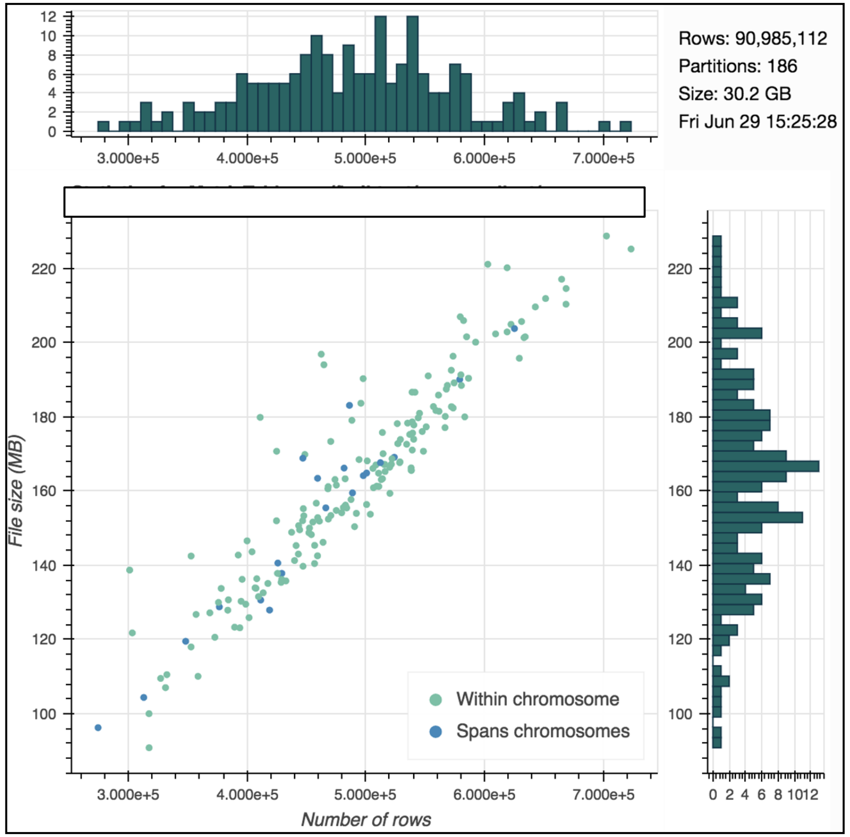
- Built-in wrappers for VEP, Nirvana
- Join with annotations by variant, locus, interval, gene
- **`ReferenceGenome`** is a first-class concept, for all our sanity
- Annotation database

Hail as a data science library

Data slinging

Analytical toolbox

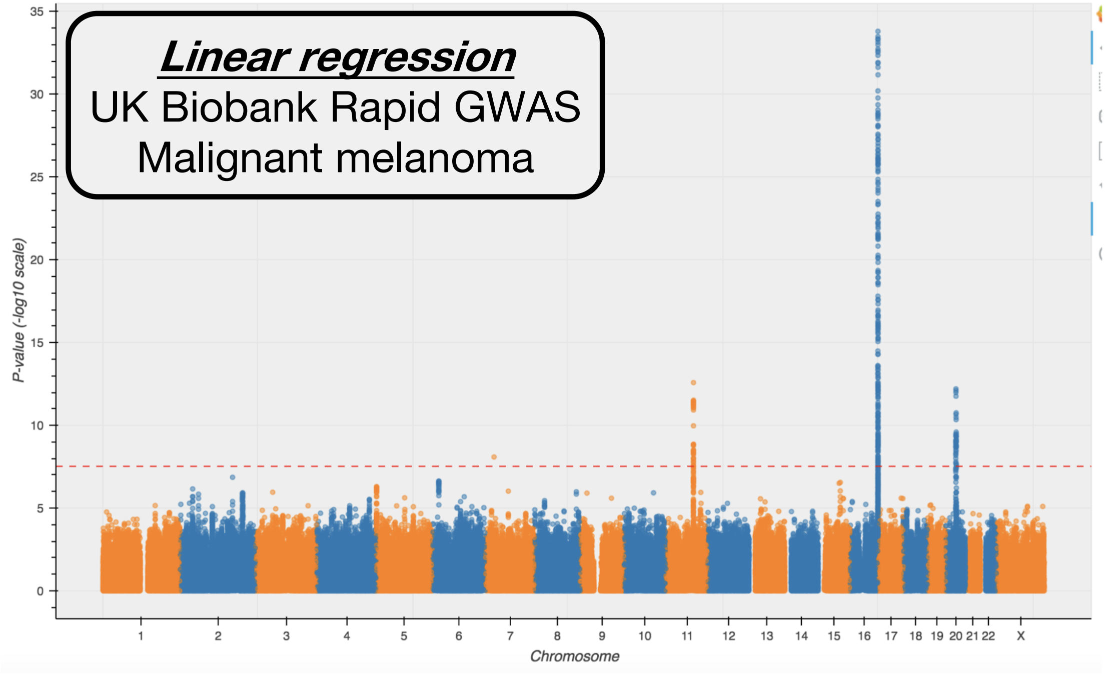
- Read and write common formats
- Filter, group, aggregate
- Annotation
- **Visualization**



Hail as a data science library

Data slinging

Analytical toolbox



Linear regression

UK Biobank Rapid GWAS

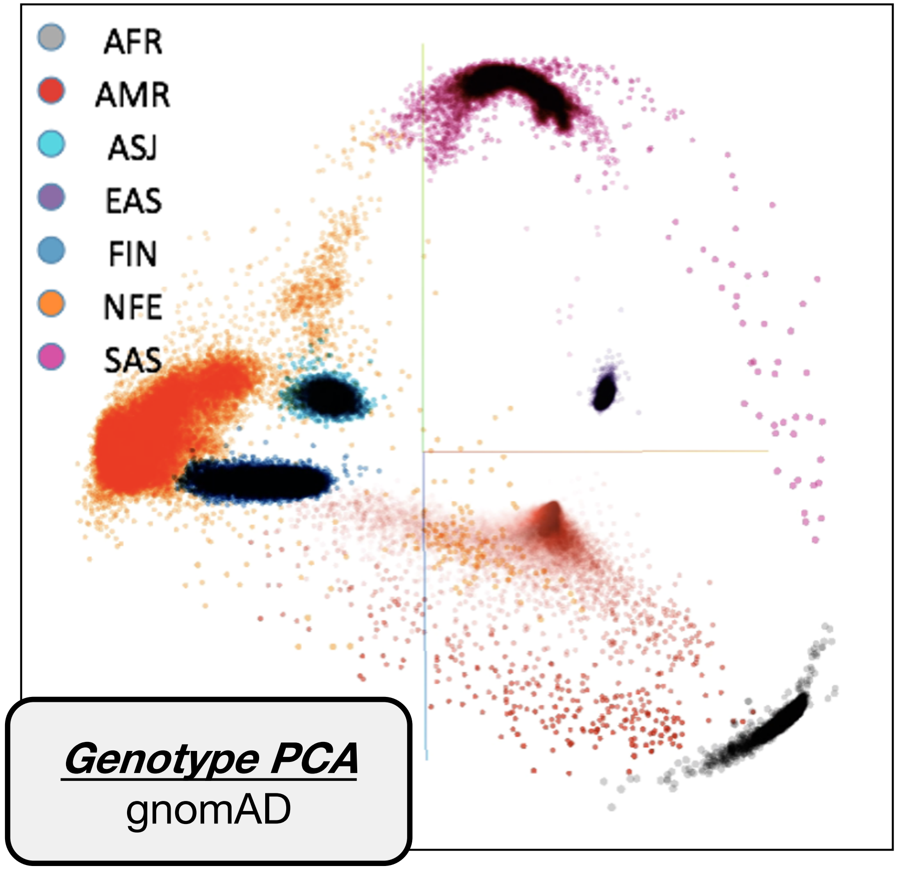
Malignant melanoma

- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

Analytical toolbox



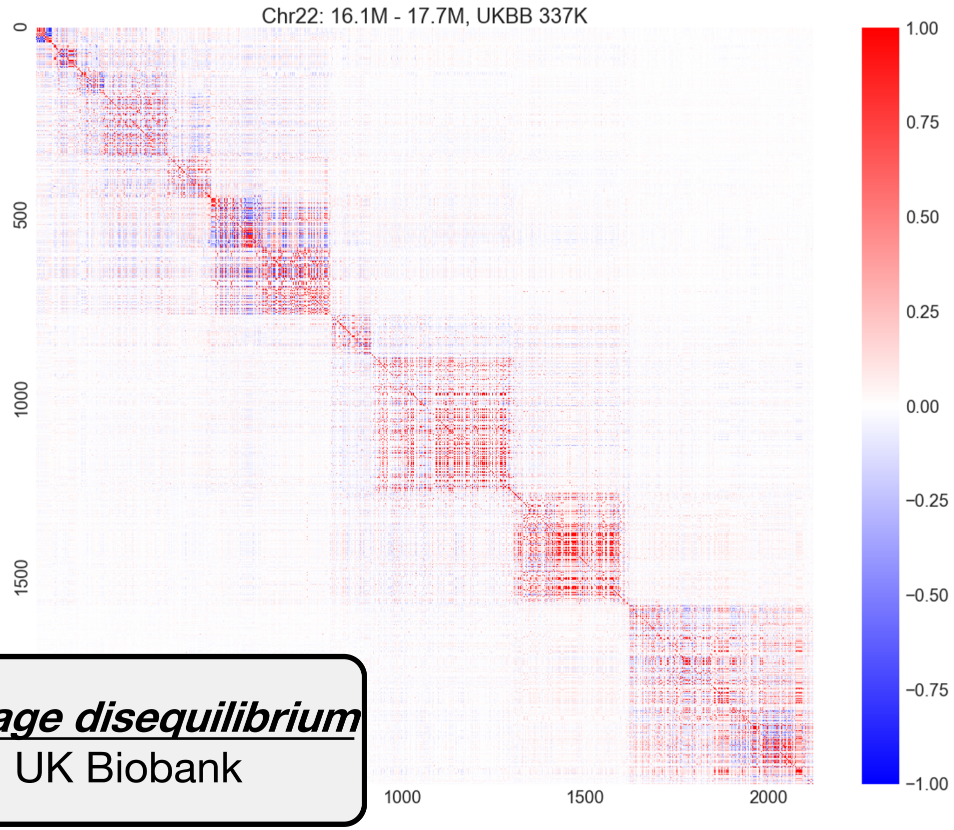
Genotype PCA
gnomAD

- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

Analytical toolbox

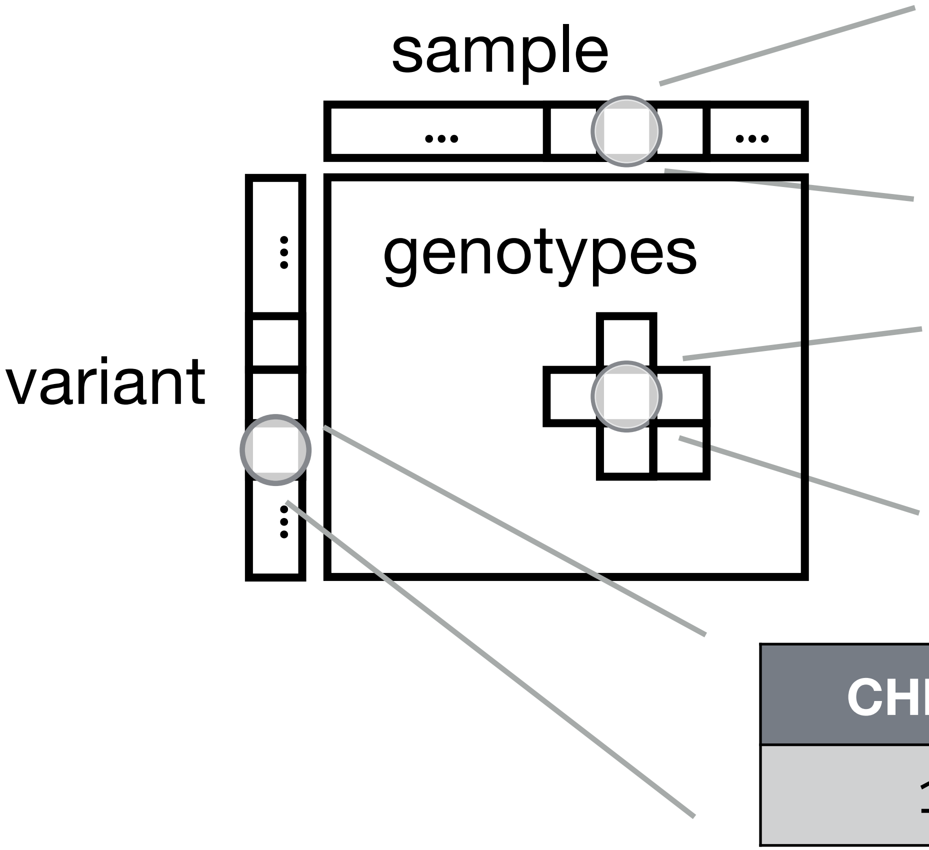


Linkage disequilibrium

UK Biobank

- Statistical methods for genetics
- **Linear algebra (early stages)**

Variant Call Format (VCF)

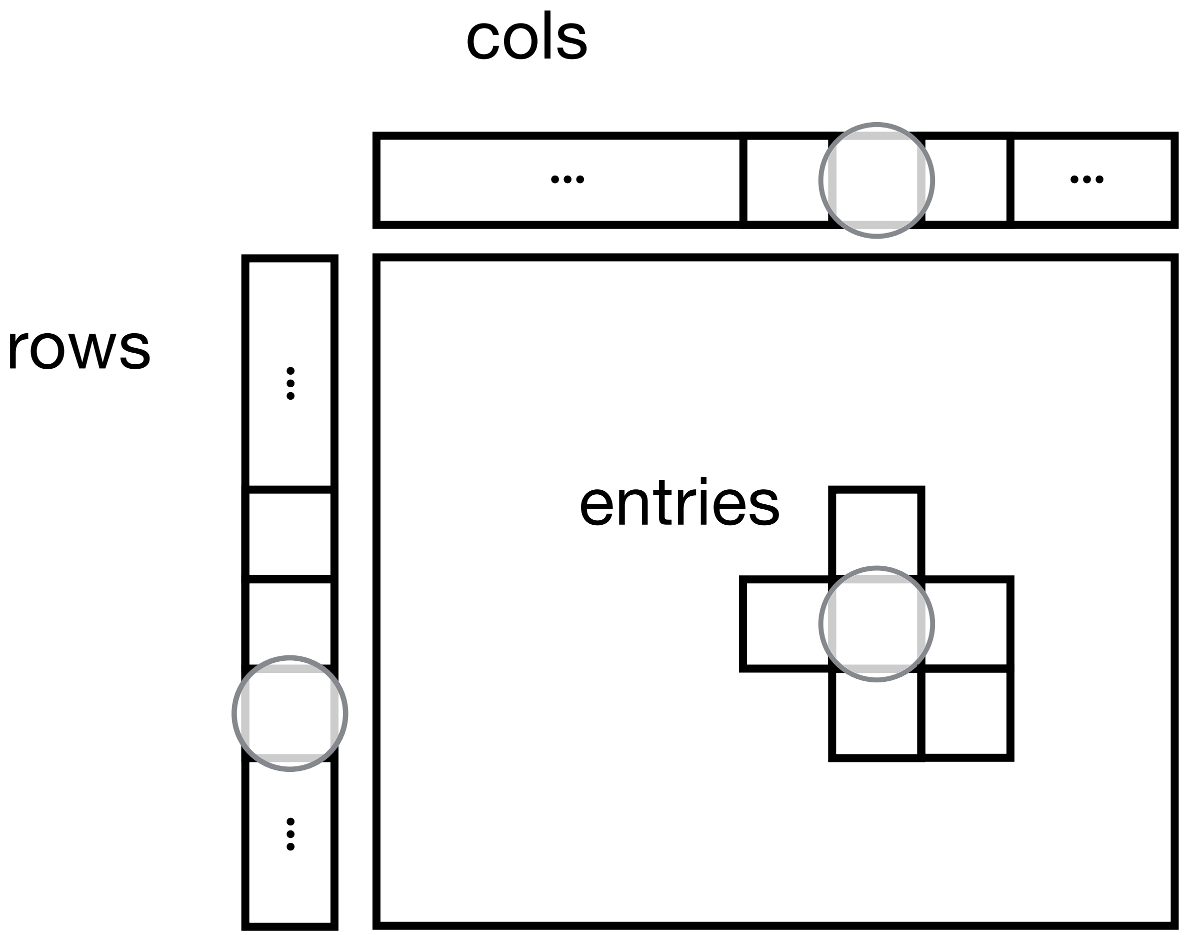


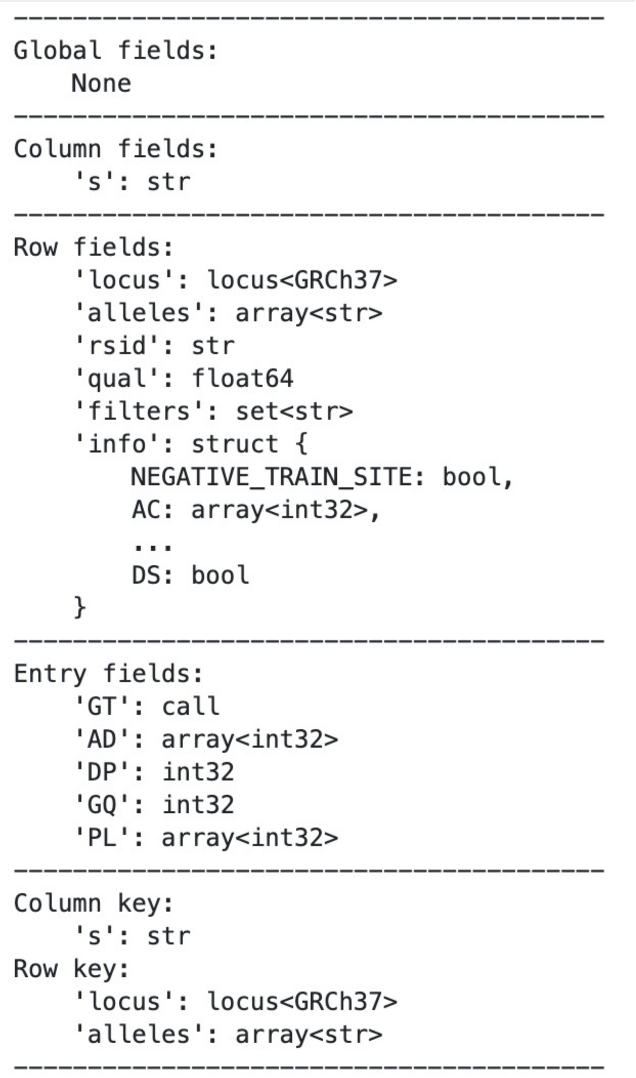
ID
NA12878

GT	AD	DP	…
0/1	8,11	19	…

CHROM	POS	RSID	REF	…
17	37282	rs12345	A	…

MatrixTable

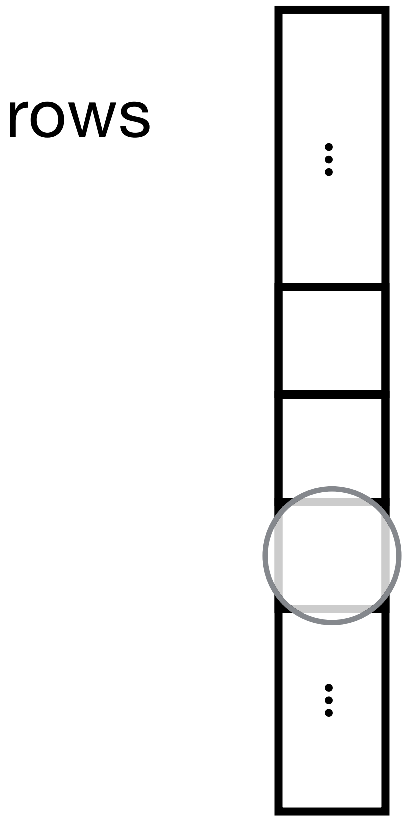




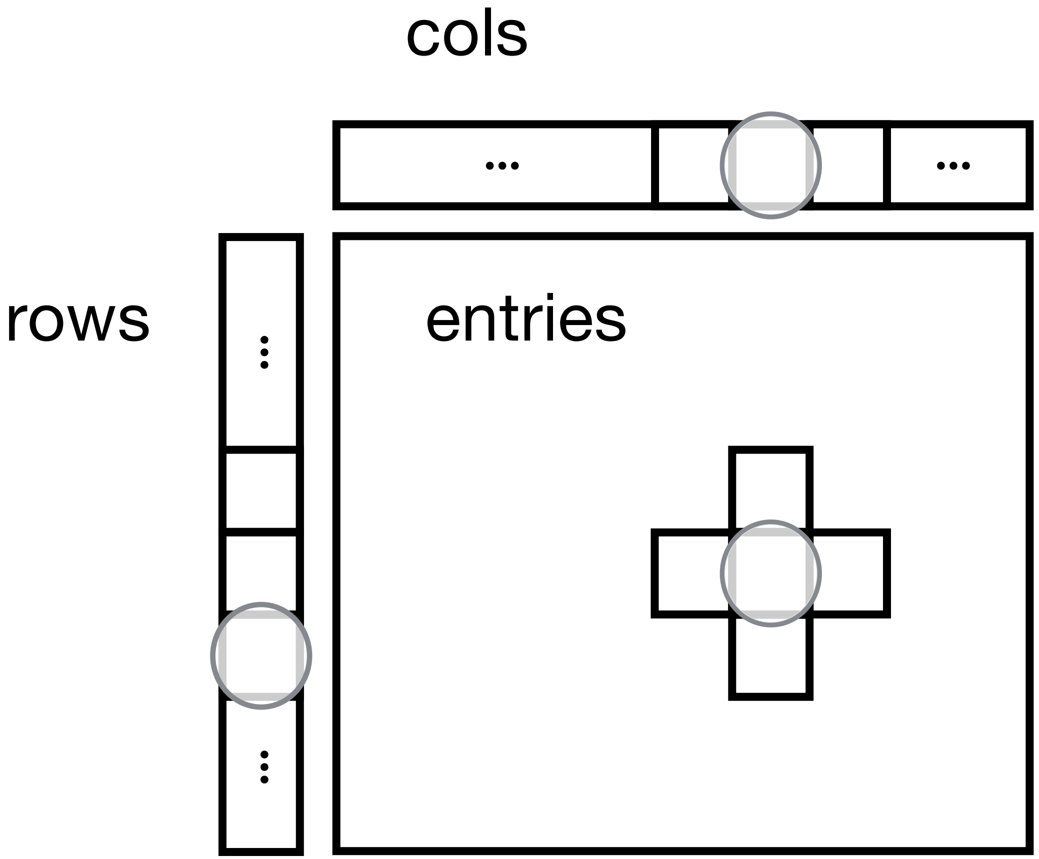
```
----------------------------------------
Global fields:
    None
----------------------------------------
Column fields:
    's': str
----------------------------------------
Row fields:
    'locus': locus<GRCh37>
    'alleles': array<str>
    'rsid': str
    'qual': float64
    'filters': set<str>
    'info': struct {
        NEGATIVE_TRAIN_SITE: bool,
        AC: array<int32>,
        ...
        DS: bool
    }
----------------------------------------
Entry fields:
    'GT': call
    'AD': array<int32>
    'DP': int32
    'GQ': int32
    'PL': array<int32>
----------------------------------------
Column key:
    's': str
Row key:
    'locus': locus<GRCh37>
    'alleles': array<str>
----------------------------------------
```

Can be extended to rare variant aggregation, trio, transcript expression

Table



MatrixTable



We have cheatsheets for this too!
https://hail.is/docs/0.2/cheatsheets.html

Hands on using

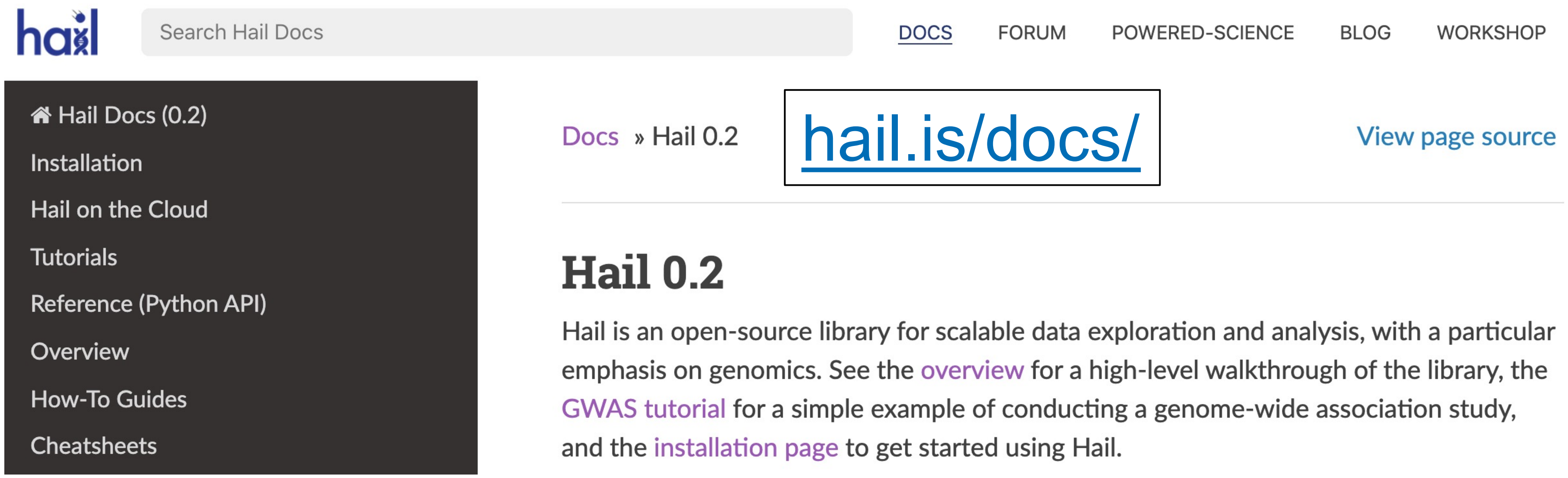
workshop.hail.is

workshop name: `broade_april2021`

password: `broade`

Your next steps

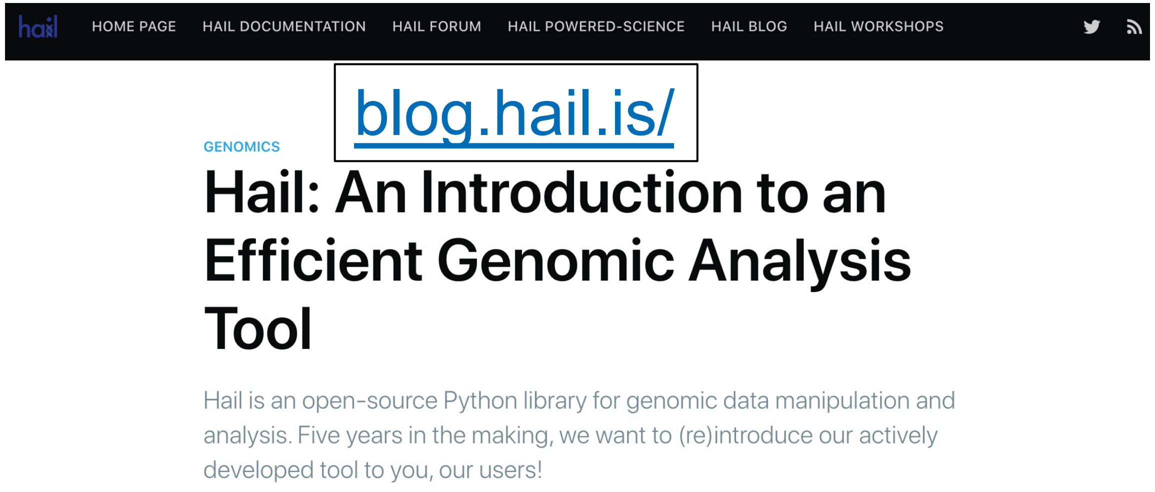
```
pip install hail
```



hail.is/docs/

Hail 0.2

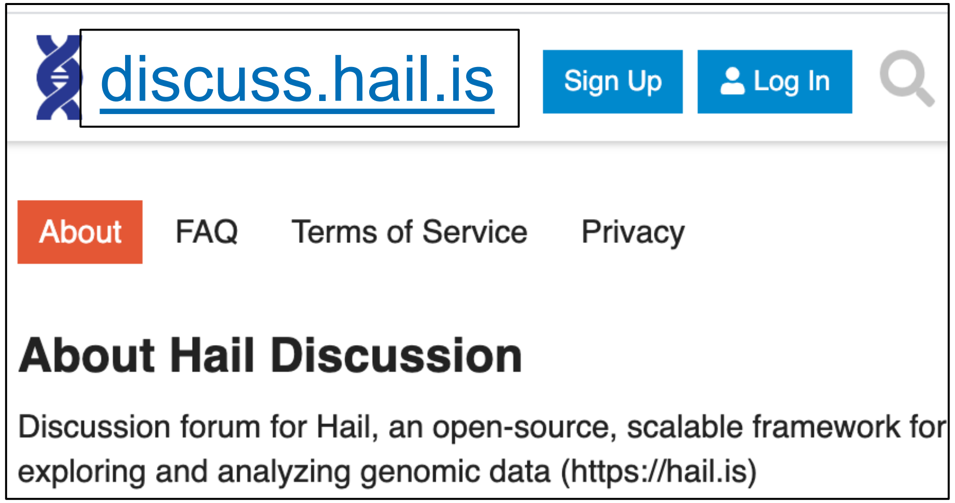
Hail is an open-source library for scalable data exploration and analysis, with a particular emphasis on genomics. See the overview for a high-level walkthrough of the library, the GWAS tutorial for a simple example of conducting a genome-wide association study, and the installation page to get started using Hail.



blog.hail.is/

Hail: An Introduction to an Efficient Genomic Analysis Tool

Hail is an open-source Python library for genomic data manipulation and analysis. Five years in the making, we want to (re)introduce our actively developed tool to you, our users!



discuss.hail.is

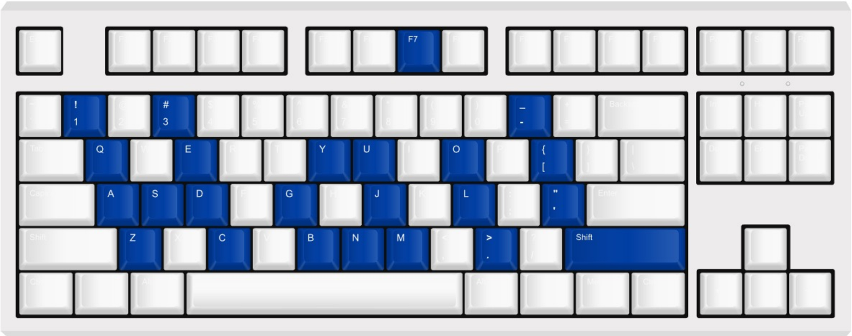
About Hail Discussion

Discussion forum for Hail, an open-source, scalable framework for exploring and analyzing genomic data (https://hail.is)

Coming soon….

Broad E workshop: Hail Batch

Fall 2021



Thank you!

Broad E Workshop 2021

Have questions? We may have answers!

Kumar Veerapen, PhD
Hail Support and Community Outreach Manager
Tim Poterba and Carolin Diaz
Software Engineer

https://hail.is
twitter @mkveerapen / @hailgenetics
✉ veerapen@broadinstitute.org
#scalableGenomics
#hailGenetics #ATGUstrong

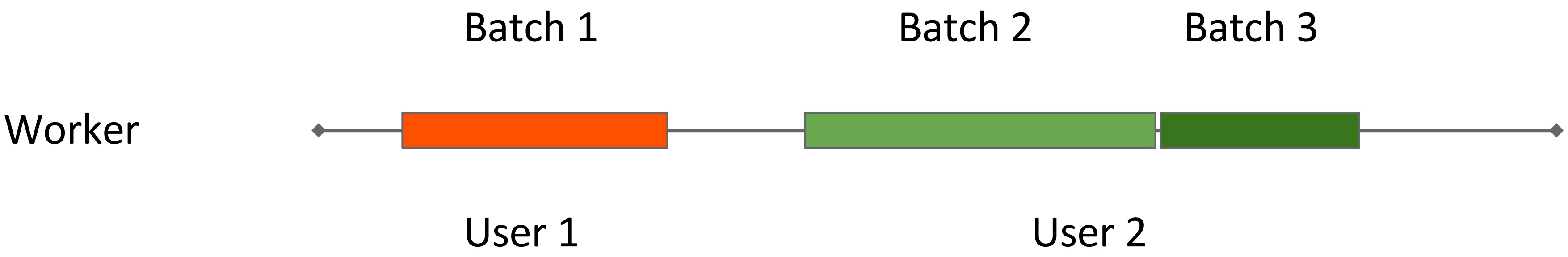
Hail Batch is a serverless, autoscaling, multitenant HPC service.

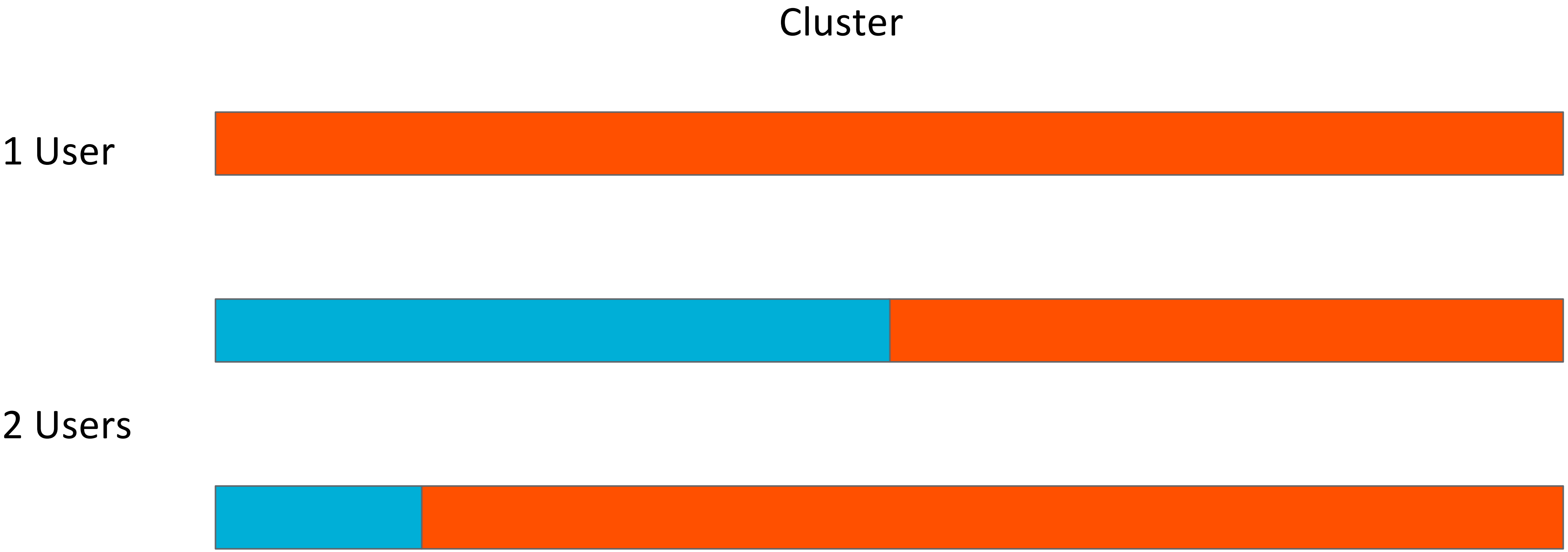
- A Python library that allows you to easily build computational workflows including managing file copying and job dependencies automatically
- A shared compute cluster in Google Cloud that the Hail team is managing (think UGER in the cloud)

- Different than the current Hail Python library a lot of people are using to analyze data with Dataproc
- Rebranding Hail => Hail Query
- Hail Batch will be the execution engine of the Hail Query Service (HaaS)

- REST API and Python client library
- Schedules static graphs of docker containers
- Handles file localization
  - gcsfuse also supported
  - Aside: copy tool
- Web UI for monitoring batches, viewing logs, etc.
  - https://batch.hail.is
- Local backend

- Jobs (containers) are scheduled on workers in our GCP project
- Schedule jobs on pools of worker instances
  - 3 pools: standard, highmem, highcpu
  - Pool workers share local SSD for very fast disk performance
  - Support for non-preemptibles, custom instance types
- We track usage
- Only pay for what you use
- Workers multiplex jobs from multiple users, batches
- Have seen >30x cost reduction batches of small jobs


Batch 1
Batch 2
Batch 3
Worker
User 1
User 2



Cluster

1 User

2 Users

- Bill by the millisecond
- Spending limits
- Roughly cost of underlying compute plus $0.01/core/hr service fee (same as Google Dataproc model)
- Details:
  - https://hail.is/docs/batch/service.html#billing

- Batch client part of the **hail** PyPI package
- Hail as a Service Sign-up: https://auth.hail.is/signup
  - $10 credit
- Batch docs: https://hail.is/docs/batch/
- Live support: https://hail.zulipchat.com/
  - “Batch support” stream
- To set up billing account, contact Whitney Wade <wwade@broadinstitute.org>