STANLEY CENTER FOR PSYCHIATRIC RESEARCH AT BROAD INSTITUTE

BroadE

BROAD INSTITUTE

Scalable Genomic Analysis

BroadE Workshop

March 5th, 2020

10:00 AM - 1:00 PM

75A-2001 Yellowstone

Kumar Veerapen, PhD
Patrick Schultz, PhD
John Compitello, BS

https://hail.is

twitter @hailgenetics

#scalableGenomics

#geneticsMadeEasy

#hailGenetics

Schedule and Outline

10:00 - 10:15 am Introduction to Hail

10:15 - 11:00 am Practical 1: Data Import and quality control

11:00 - 11:10 am Q & A

11:10 - 11:50 am Practical 2: GWAS and Rare Variant Analysis

11:50 - 12:00 pm Q & A

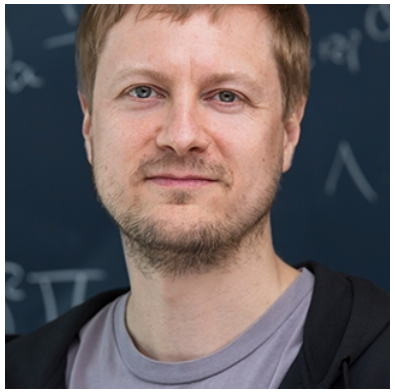
12:00 - 12:45 pm Practical 3: PCA and Ancestry

12:45 - 12:55 pm Q & A

12:55 - 1:00 pm Wrap-up

BROAD INSTITUTE

Hail Team



Cotton Seed
Team Leader

Tim Poterba

Dan King

Jackie Goldstein

Alex Kotlar

Patrick Schultz

Whitney Wade
Operations

Kumar Veerapen
Support and Outreach

John Compitello

Arcturus Wang

Chris Vittal

Back

When poll is active, respond at **PollEv.com/hail2020**

Text **HAIL2020** to **22333** once to join

Visual settings

Activate

Show responses

Lock

Clear responses

What tool do you currently use to analyse your genomic dat

"fastQC"
2020-03-05T09:11:01-06:00

"GATK"
2020-03-05T11:10:03-04:00

"plink"
2020-03-05T11:09:55-04:00

"gatk"
2020-03-05T11:09:55-04:00

Total Results: 27

Logout

Back

When poll is active, respond at **PollEv.com/hail2020**

Text **HAIL2020** to **22333** once to join

Visual settings

Activate

Show responses

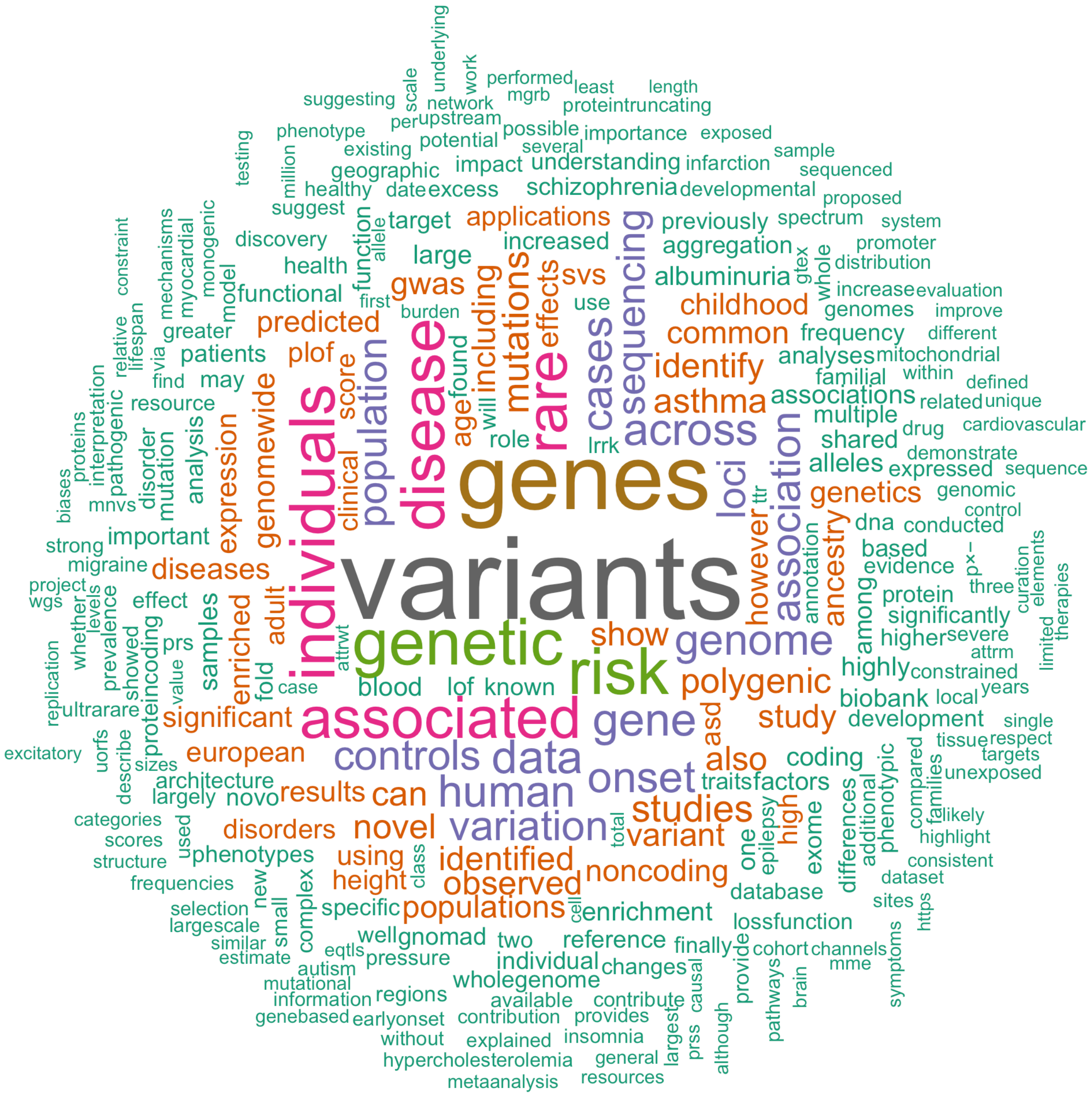
Lock

Clear responses

What word comes to mind when you think of Ha

Logout

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



Notes:

- 48 abstracts (02/03/2020)
- Word appearing > 4x

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

- **Scalable Genomics Analysis Tool**
  - Can run on a laptop, or 5000 cores on the cloud, e.g. 1 million genomes
- **Ease of Use**
  - Simple, powerful abstractions, and domain primitives
  - Most of the tools you need*, together in one place
    - Giving you the tools you need to indulge scientific curiosity
- **Reusable Infrastructure**
  - Rapid development on Apache Spark, and all open-source!
- **Learn more today and at www.hail.is**

It's works well **for big data**

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

BROAD INSTITUTE

Why would you use Hail?

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

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

Analytical toolbox

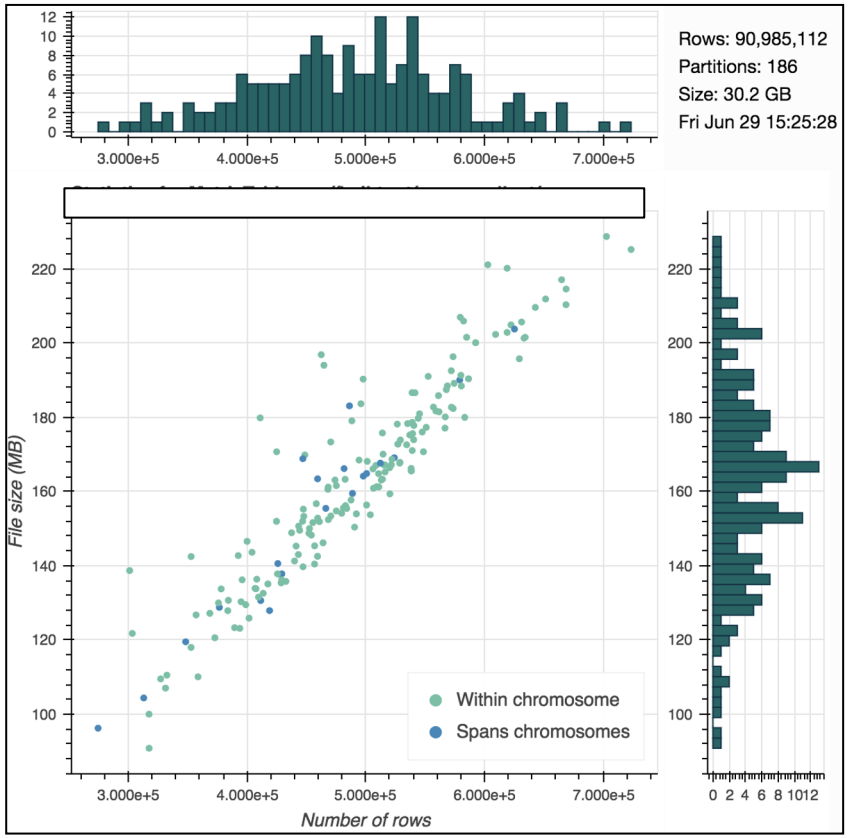
- Built-in wrappers for VEP, Nirvana
- Join with annotations by variant, locus, interval, gene
- **`ReferenceGenome`** is a first-class concept, for all our sanity
- Coming soon: annotation database

Hail as a data science library

Data slinging

Analytical toolbox

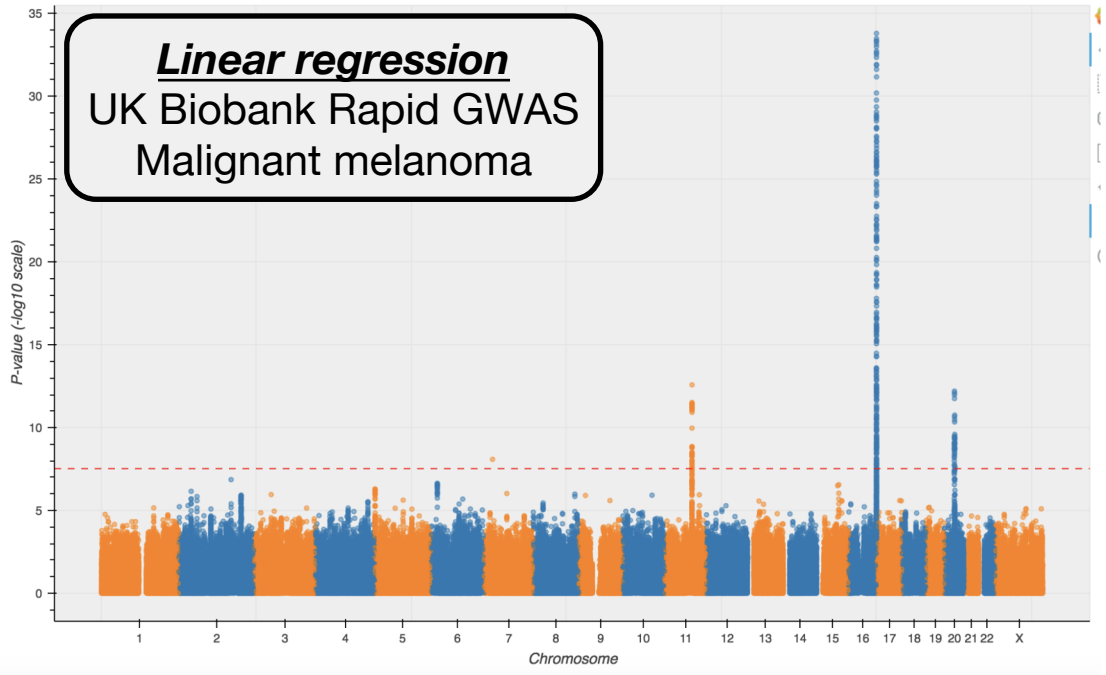
- Read and write common formats
- Filter, group, aggregate
- Annotation
- **Visualization**



Hail as a data science library

Data slinging

Analytical toolbox

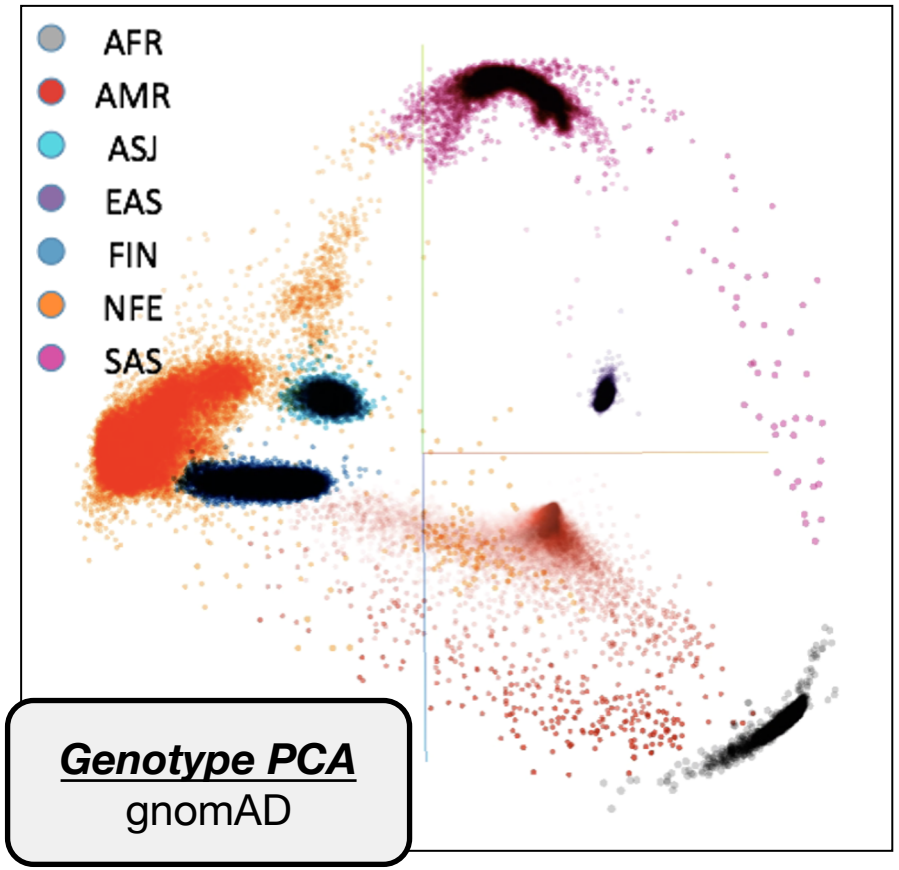


- **Statistical methods for genetics**
- Linear algebra

Hail as a data science library

Data slinging

Analytical toolbox



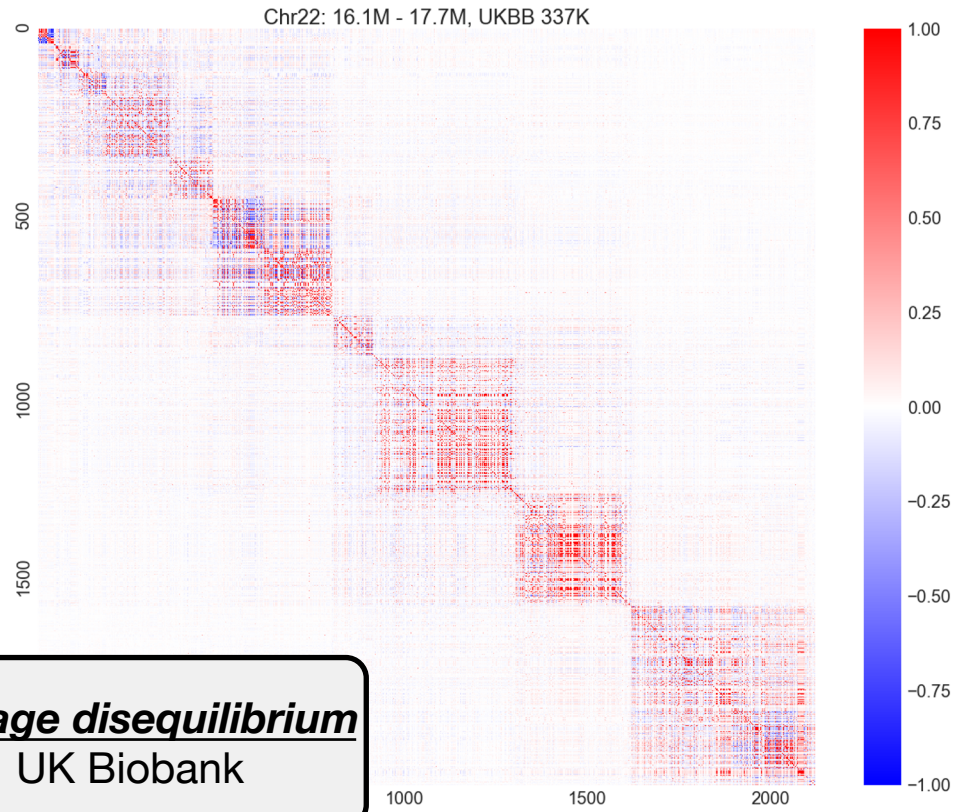
Genotype PCA
gnomAD

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

Hail as a data science library

Data slinging

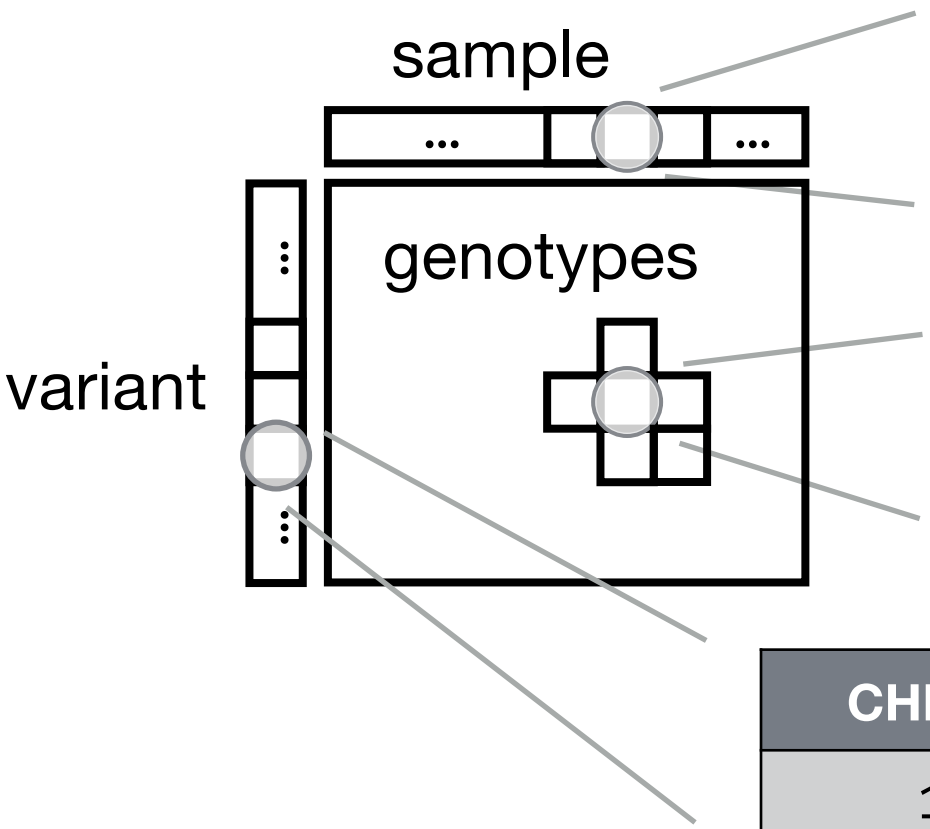
Analytical toolbox



Linkage disequilibrium
UK Biobank

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

Variant Call Format (VCF)

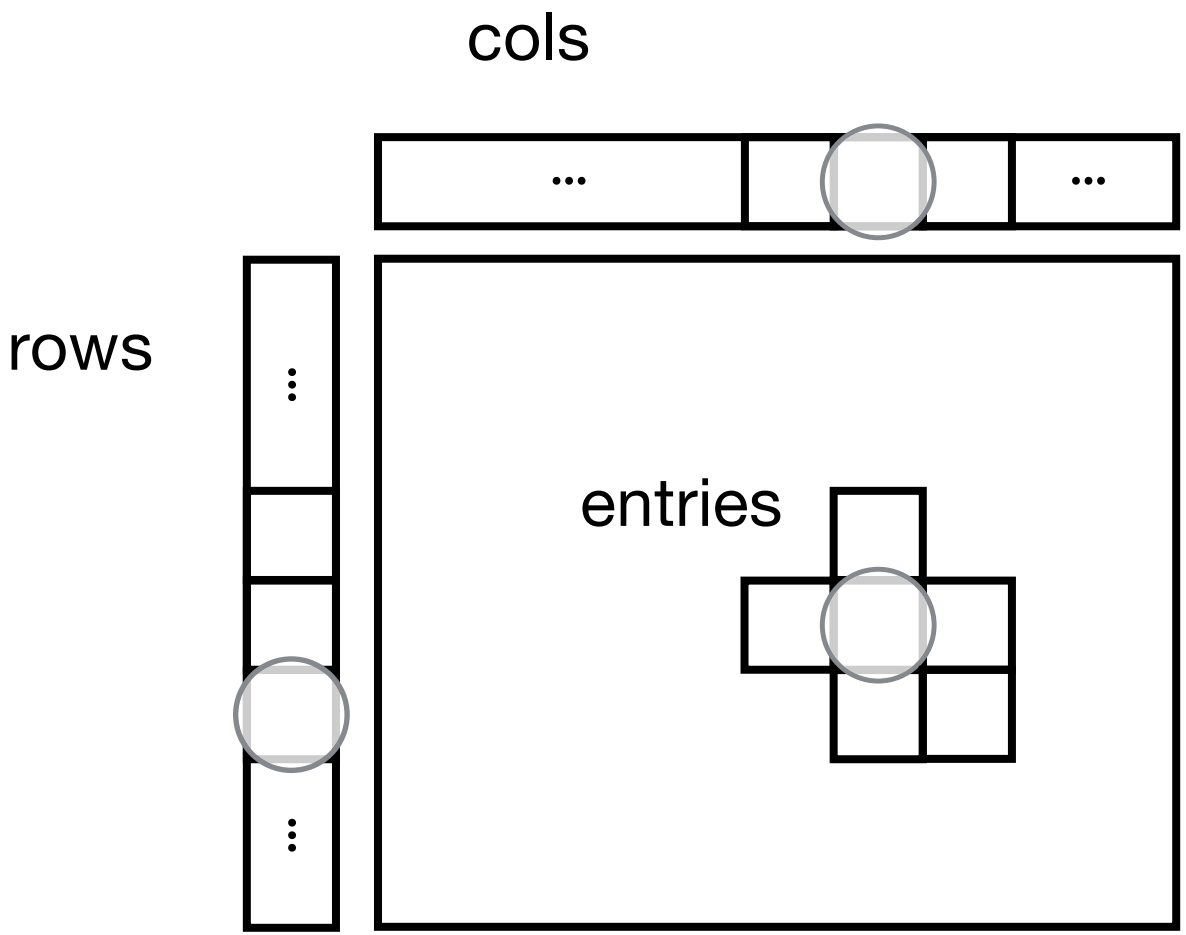


ID
NA12878

GT	AD	DP	...
0/1	8,11	19	…

CHROM	POS	RSID	REF	...
17	37282	rs12345	A	…

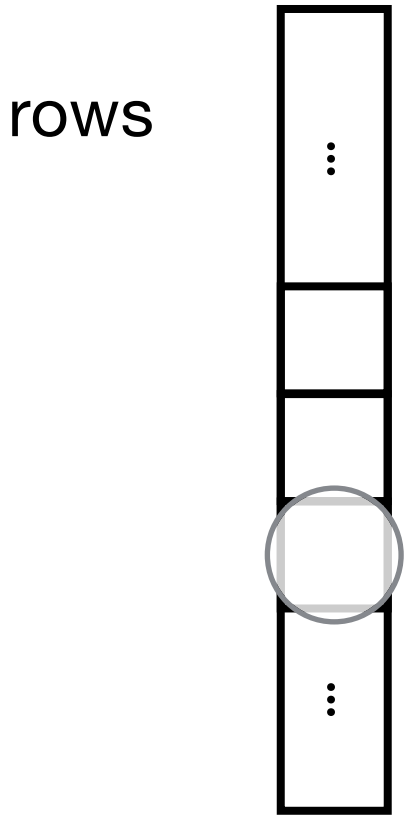
MatrixTable



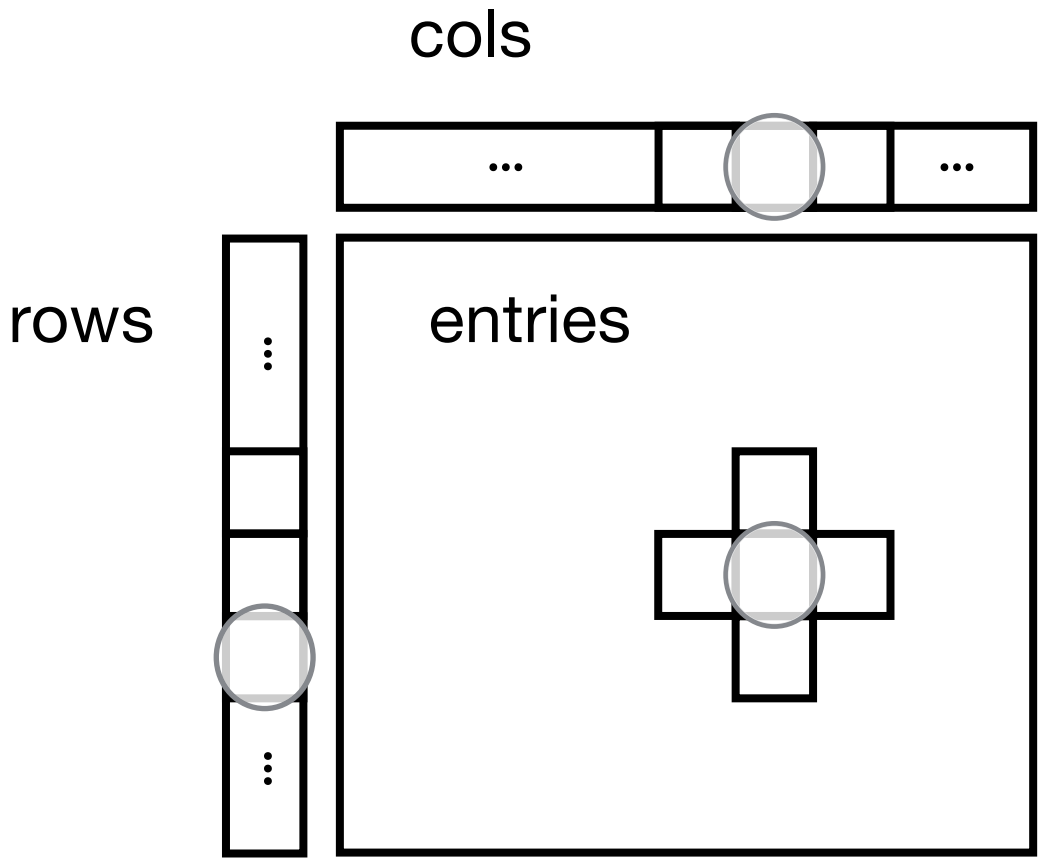

```
----------------------------------------
Global fields:
    None
----------------------------------------
Column fields:
    's': str
----------------------------------------
Row fields:
    'locus': locus<GRCh37>
    'alleles': array<str>
    'rsid': str
    'qual': float64
    'filters': set<str>
    '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

Back

Respond at **PollEv.com/hail2020**

Text **HAIL2020** to **22333** once to join, then text your message

Visual settings

Activate

Show responses

Lock

Clear responses

If you have questions, text us!

No responses received yet. They will appear here...

Logout

Learning Objective

To be able to understand the basic genomic applications for Hail

workshop.hail.is

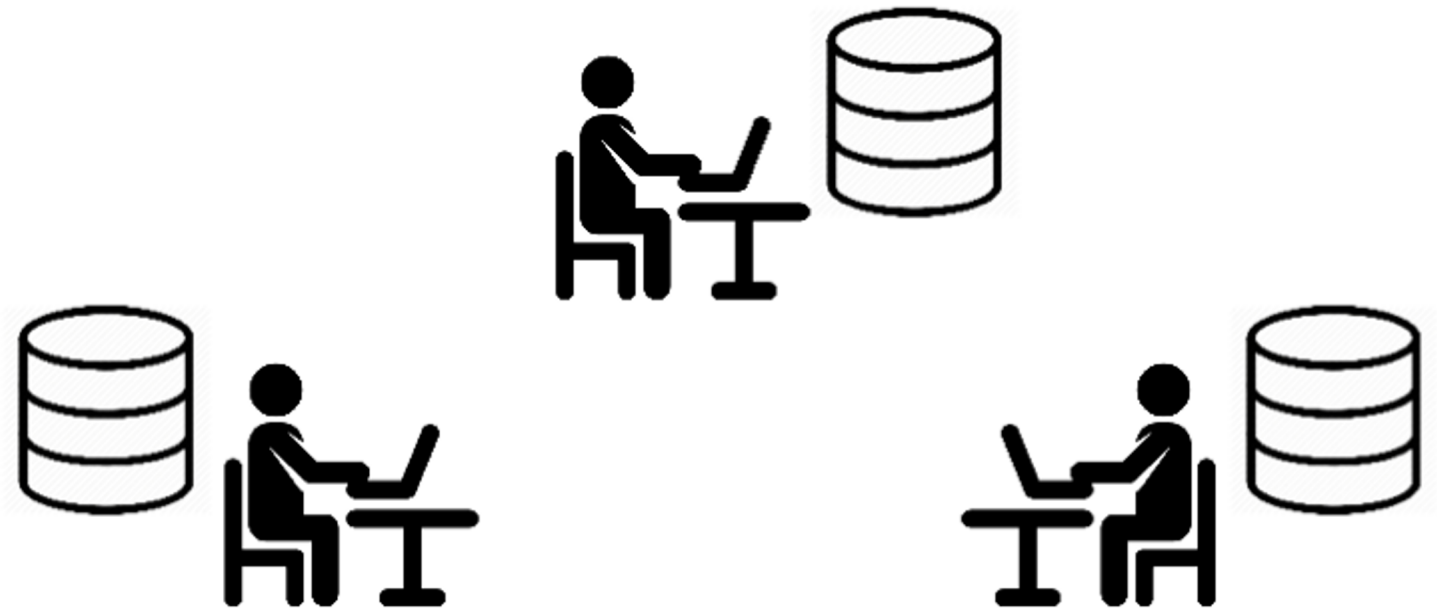
Workshop name: **`broade_march2020`**

password: **`broade`**

Large-scale datasets

- UK Biobank 500K => 5M?
  - … and many other biobanks
- gnomAD: 20K => 150K WGS
- TOPMed: >120K WGS
- All of Us: 1M
- Million Veterans Project: 1M

From Bringing Data to Researchers



To Bringing Researchers to Data

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

DOCS ▾ FORUM CHAT CODE WORKSHOPS

Hail-Powered Science

Downloaded ~140,000 times to date
https://pypistats.org/packages/hail

The following is an incomplete list of scientific work enabled by Hail. We welcome you to add additional examples by editing this page directly, after which we will review the pull request to confirm the addition is valid. Please adhere to the existing formatting conventions.

In addition to software development, the Hail team engages in theoretical, algorithmic, and empirical research inspired by scientific collaboration. Examples include Loss landscapes of regularized linear autoencoders, Secure multi-party linear regression at plaintext speed, and A synthetic-diploid benchmark for accurate variant-calling evaluation.

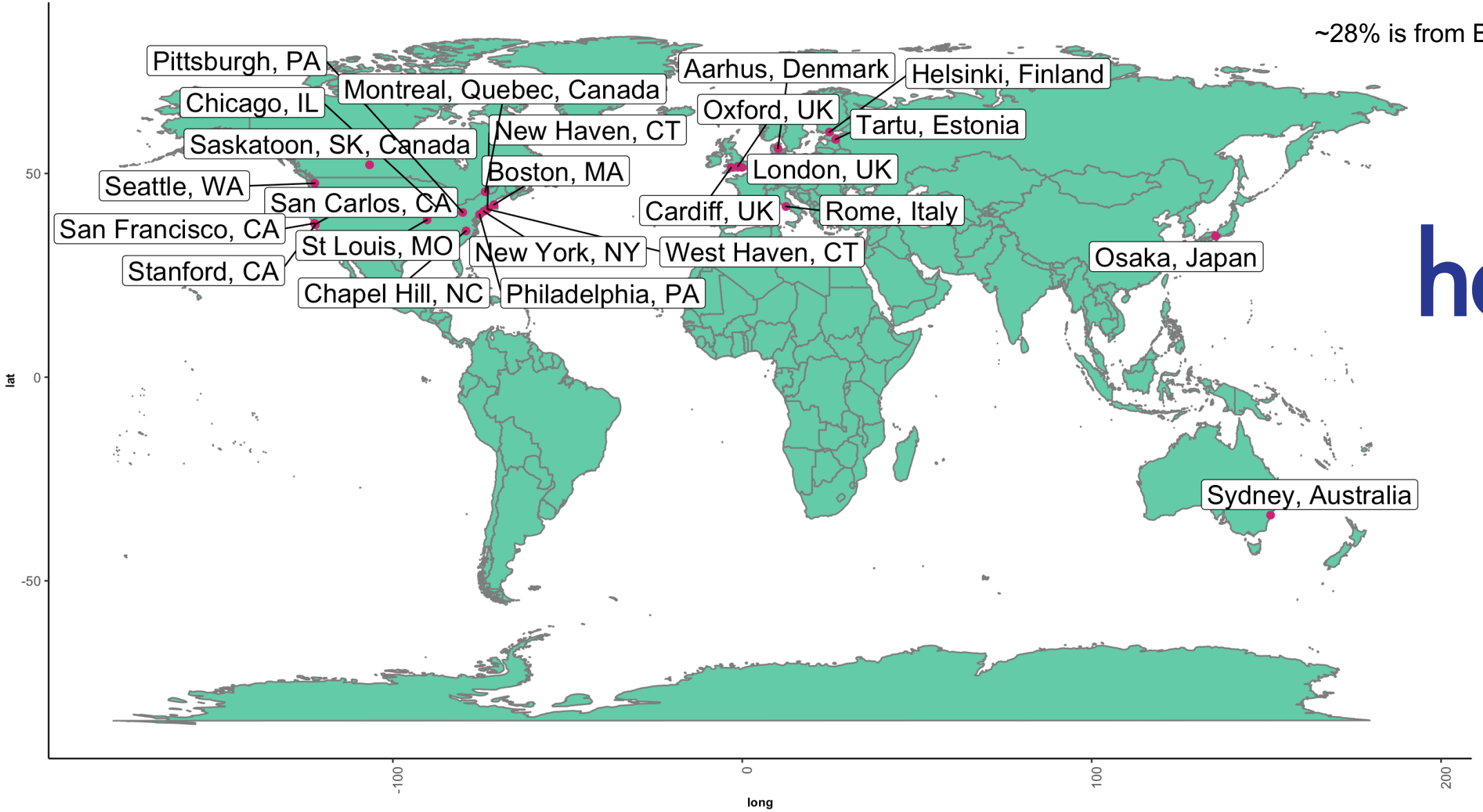
Last updated on February 3rd, 2020 at 11:00 AM EST

```
import hail as hl
print(hl.citation())
```

```
Hail Team. Hail 0.2.13-81ab564db2b4. https://github.com/hail-is/hail/releases/tag/0.2.13.
```

Where has Hail been used?

~28% is from Boston, MA



Computational Landscape

- Laptop/Desktop
- Server
- High Performance Computing (HPC) cluster
- Cloud

Computational Landscape

- Laptop/Desktop
  development, small data (10s of WGS, 100s of WES)
- Server
  medium data (1Ks WGS, 10Ks of WES)
- High Performance Computing (HPC) cluster
  large (1M WGS, 10M WES)
- Cloud
  large (1M WGS, 10M WES)

hail

Computational Landscape

- Laptop/Desktop

  `pip install hail`

- Server/High Performance Computing (HPC) cluster single node

  `pip install hail`

- High Performance Computing (HPC) cluster

  On-prem Spark cluster

  Hail *does not support* HPC schedulers like SLURM, UGER, and LSF

- Cloud

  Google Cloud Platform (GCP):

  `pip install hail`

  `hailctl dataproc start CLUSTER`

  Amazon Web Services (AWS): some support

  - https://github.com/hms-dbmi/hail-on-AWS-spot-instances
  - https://discuss.hail.is/t/spin-up-aws-emr-clusters-with-hail/818

Your next steps

```
pip install hail
```

hail.zulipchat.com