

Welcome! While waiting for our session to start:

- Please ensure that your microphone is muted during the presentation.  
**But** 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.



STANLEY CENTER  
FOR PSYCHIATRIC RESEARCH  
AT BROAD INSTITUTE



BROAD  
INSTITUTE



# Scalable Genomics for Common Variants

---

## ATGU Welcome Workshop

July 24<sup>th</sup>, 2020

10:00 – 12:00 PM (EST)

Zoom

Kumar Veerapen, PhD  
*Hail Support and Community Outreach Manager*  
Arcturus Wang  
*Software Engineer*

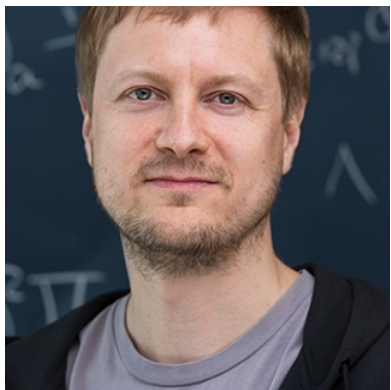


<https://hail.is>  
@mkveerapen / @hailgenetics  
veerapen@broadinstitute.org  
#scalableGenomics  
#hailGenetics #ATGUstrong

# Outline

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

# Hail Team



*Cotton Seed, PhD  
Team Leader*



*Tim Poterba*



*Dan King*



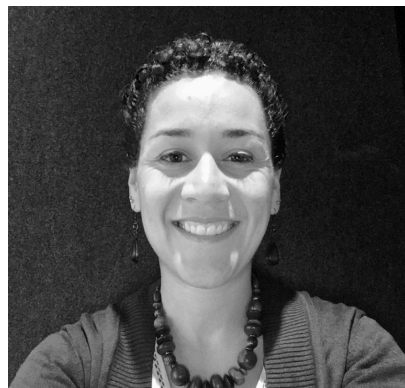
*Jackie Goldstein*



*Alex Kotlar, PhD*



*Patrick Schultz, PhD*



*Whitney Wade  
Operations*



*Kumar Veerapen, PhD  
Support and Outreach*



*John Compitello*



*Arcturus Wang*



*Chris Vittal*



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

## Where are you from?

Loading Image...

🗨 Answers to this poll are anonymous



📄 Respond at **PollEv.com/hail2020**

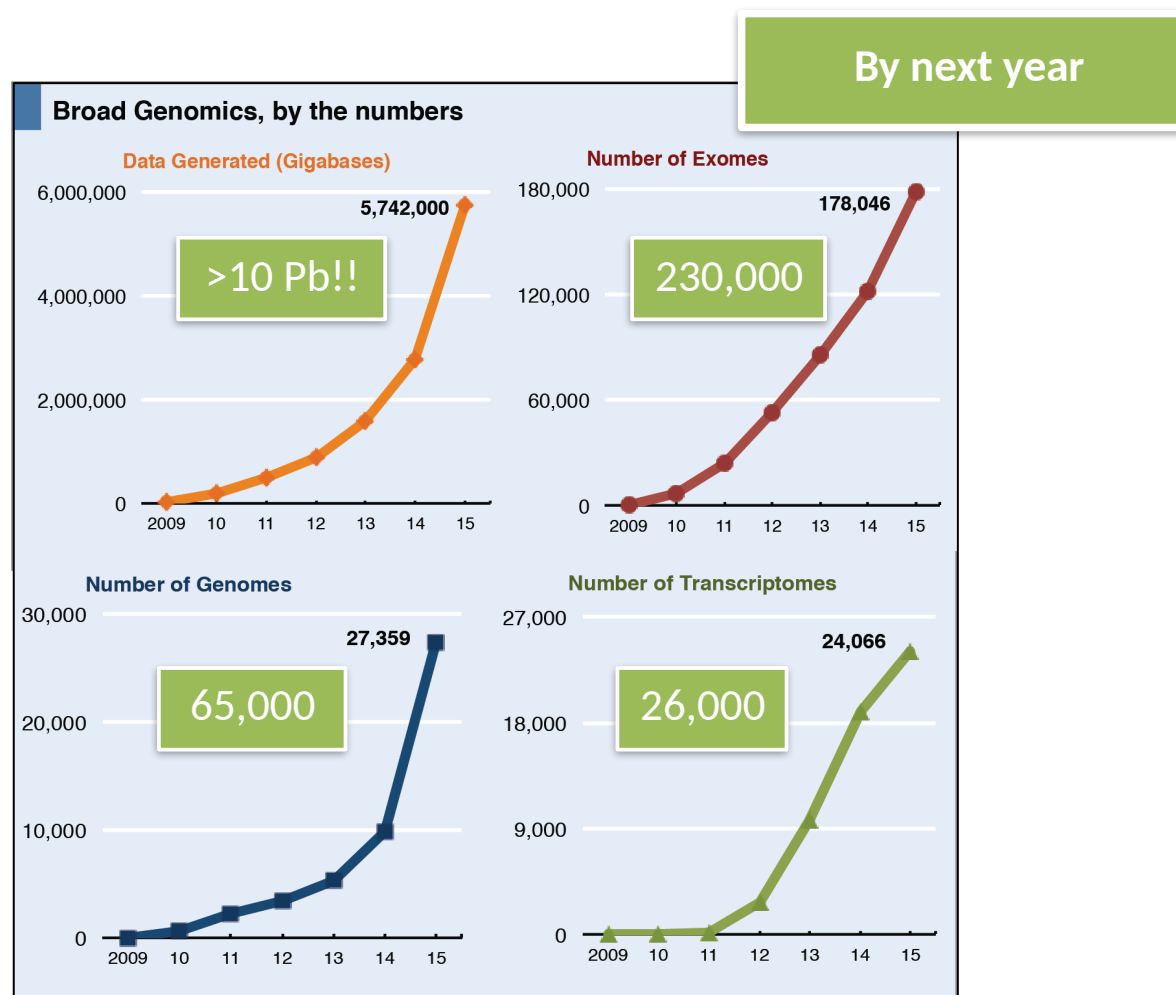
💬 Text **HAIL2020** to **37607** once to join, then text your message

**What tool do you currently use for your genetic / genomic analyses? Or what have you learned so far?**

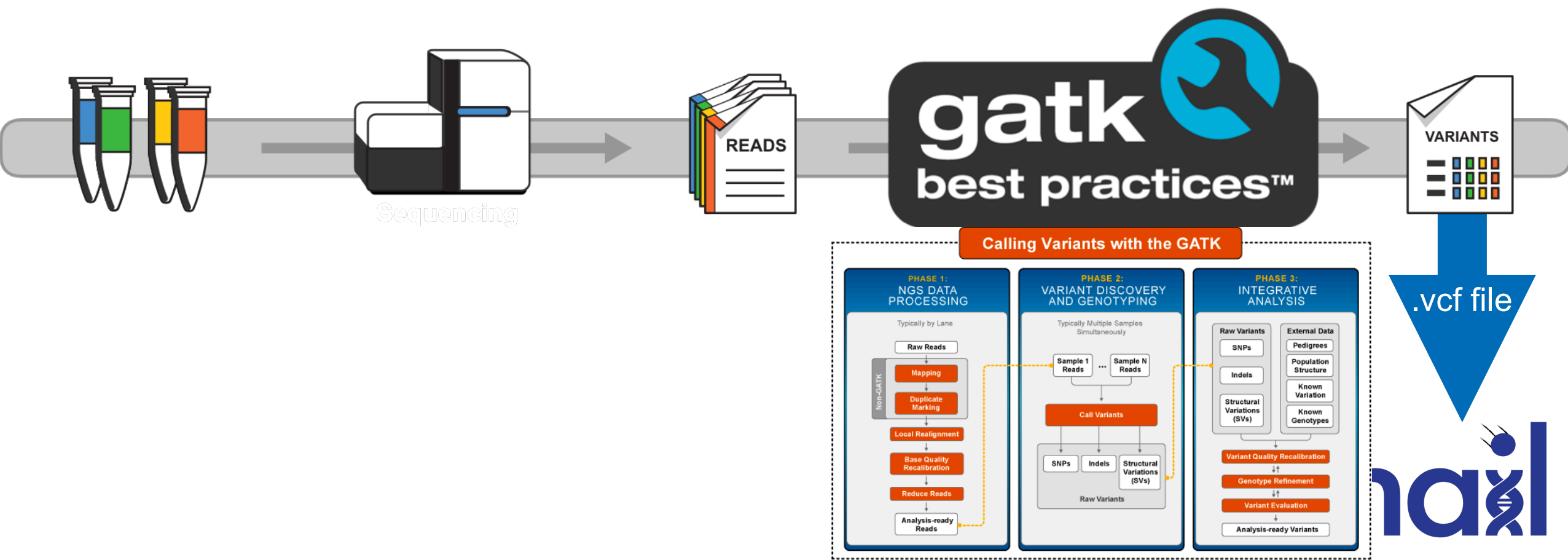
plink  
wdl/cromwell

🗳️ Answers to this poll are anonymous

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



# 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*

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

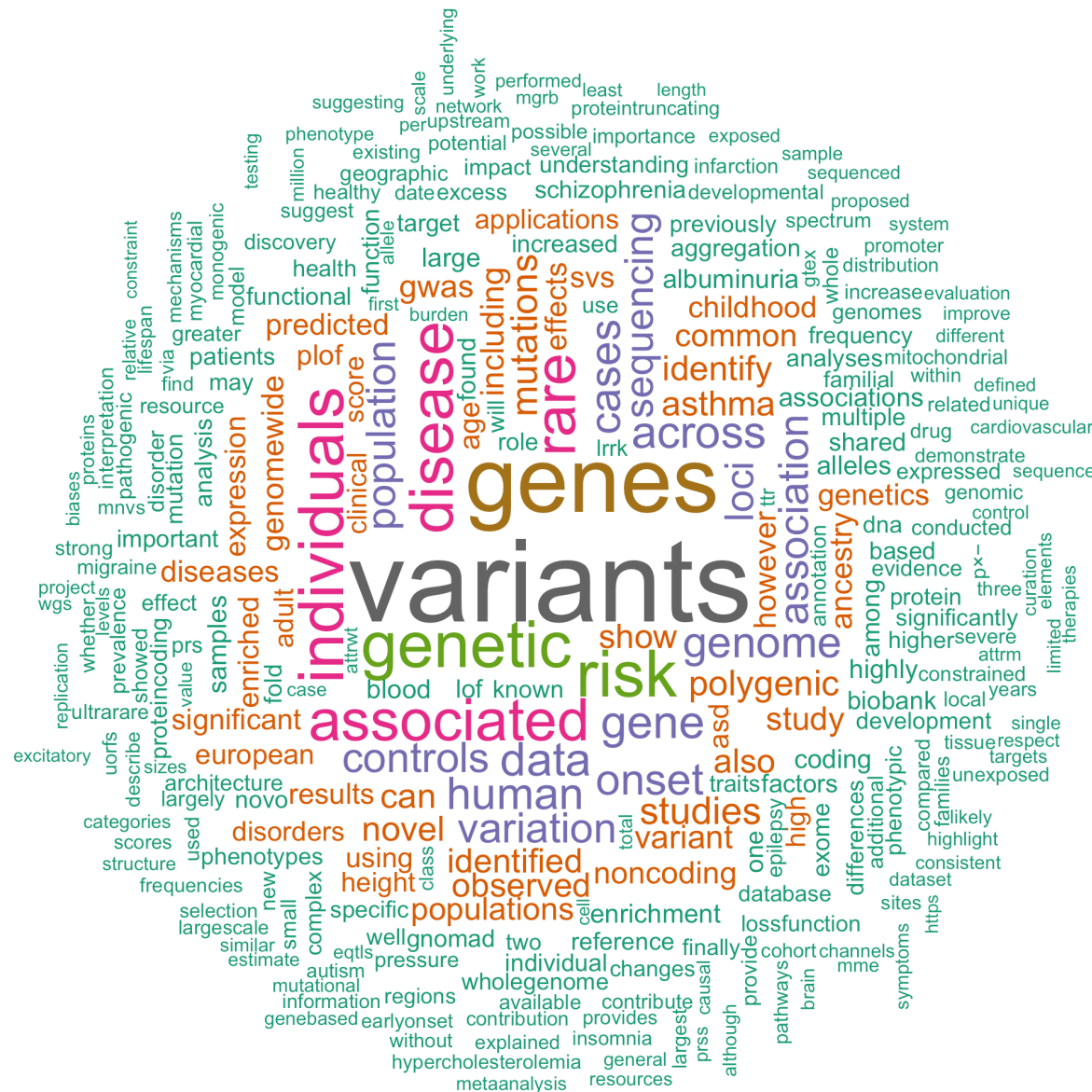
Tabular data frames  
imported as Hail  
MatrixTables into  
**unified platform.**



Learn more at [Hail.is](https://hail.is)

**\*We can't read your  
minds, so talk to us**  
[discuss.hail.is](https://discuss.hail.is)

# How has Hail been used? ([hail.is/references.html](https://hail.is/references.html))

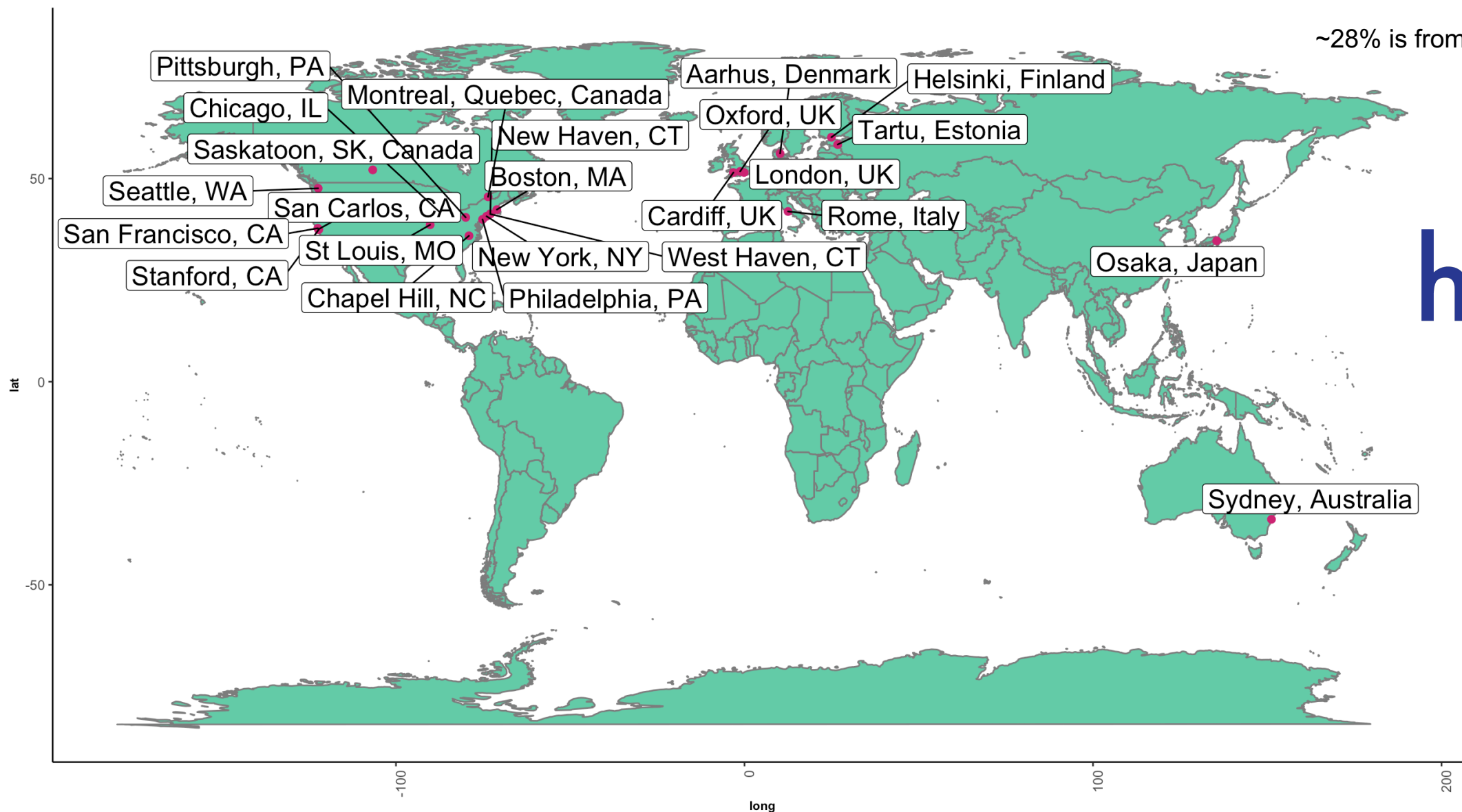


Notes:

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

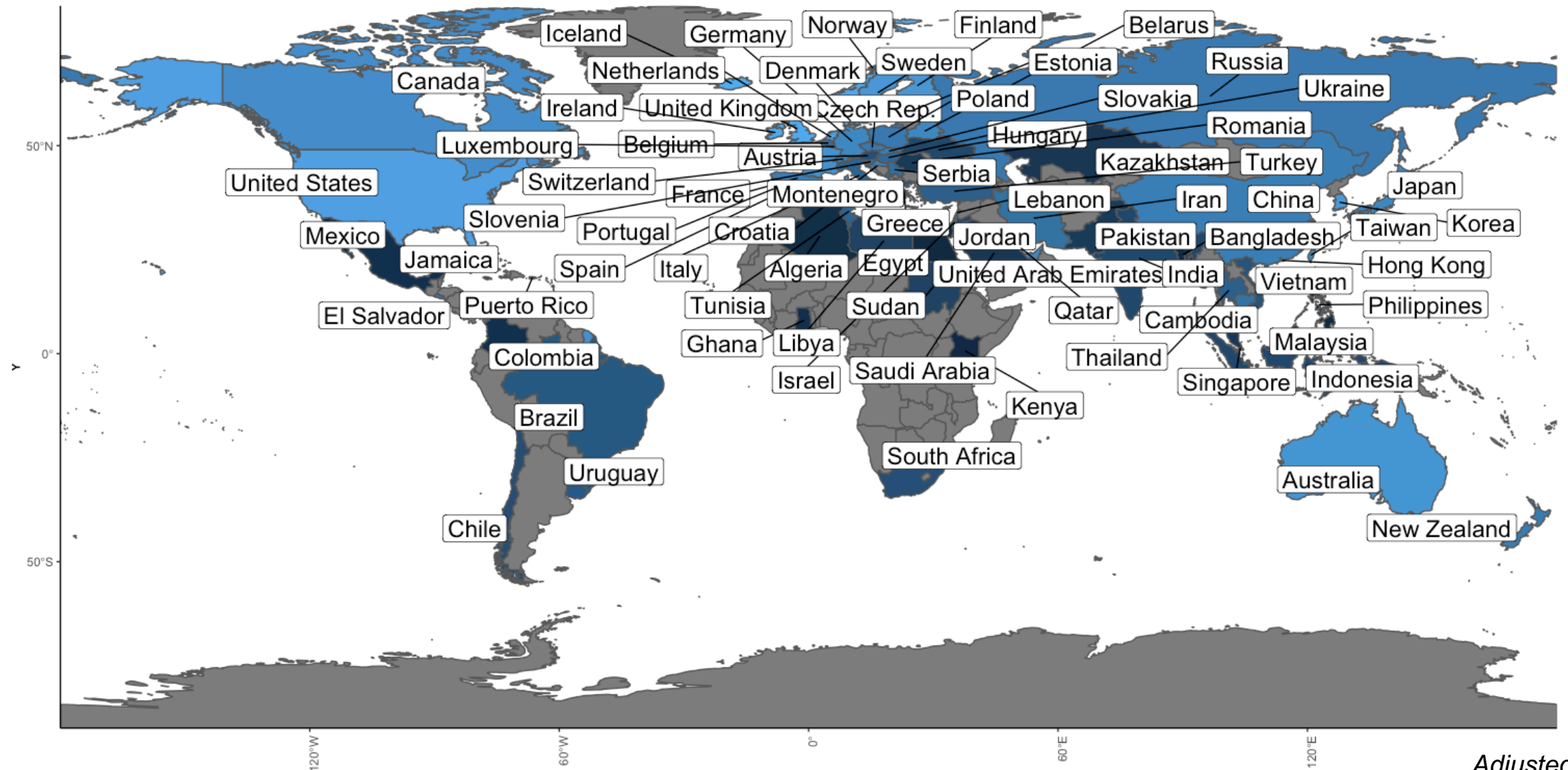
# Where has Hail been used?

~28% is from Boston, MA



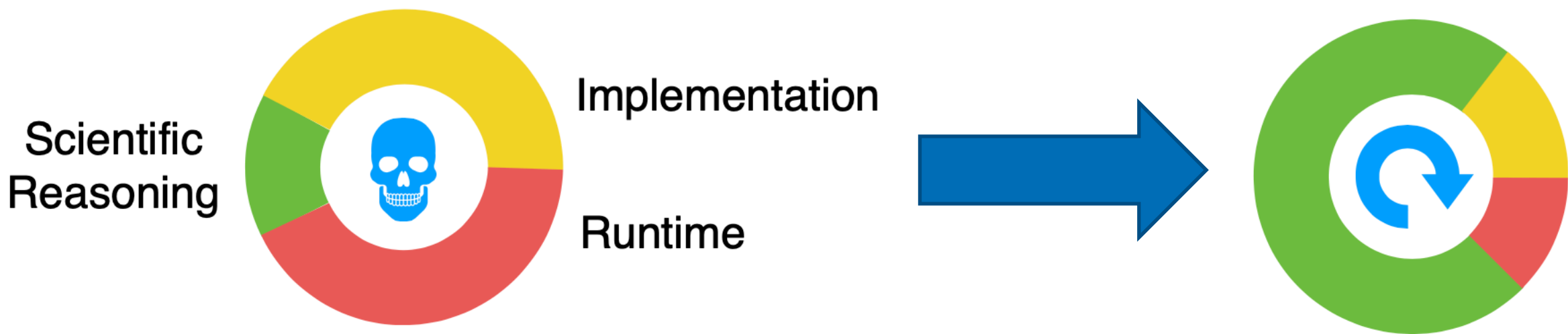
hail

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



*Adjusted for total population*

# 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

## Analytical toolbox

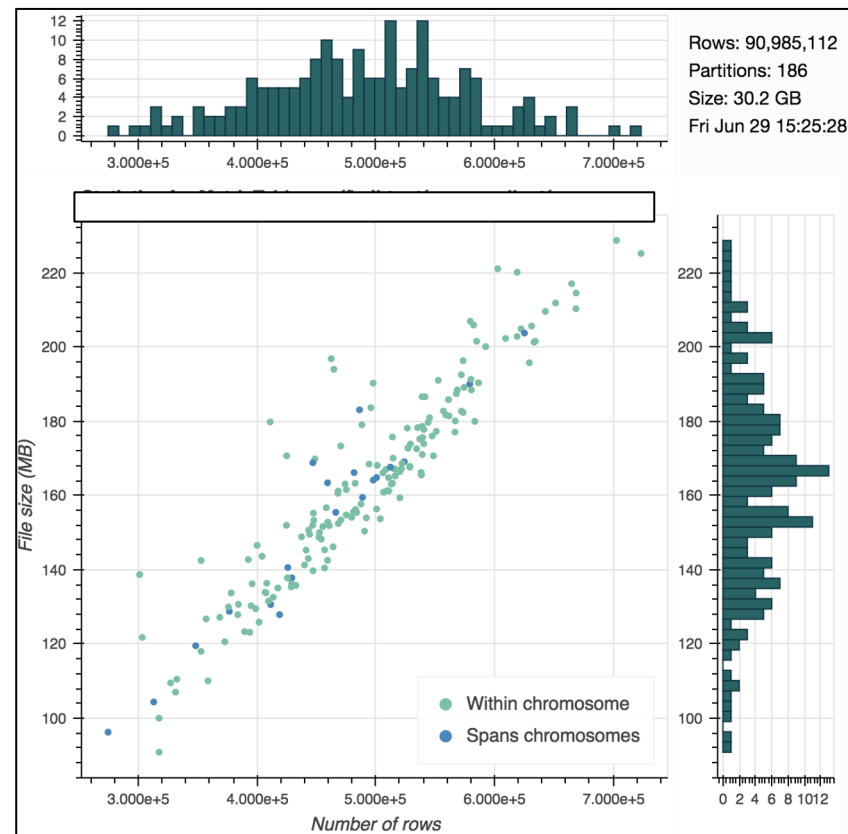
- Read and write common formats
- Filter, group, aggregate
- **Annotation**
- Visualization
- Built-in wrappers for VEP, Nirvana
- Join with annotations by variant, locus, interval, gene
- **ReferenceGenome** is a first-class concept, for all our sanity
- Annotation database

# Hail as a data science library

## Data slinging

## Analytical toolbox

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

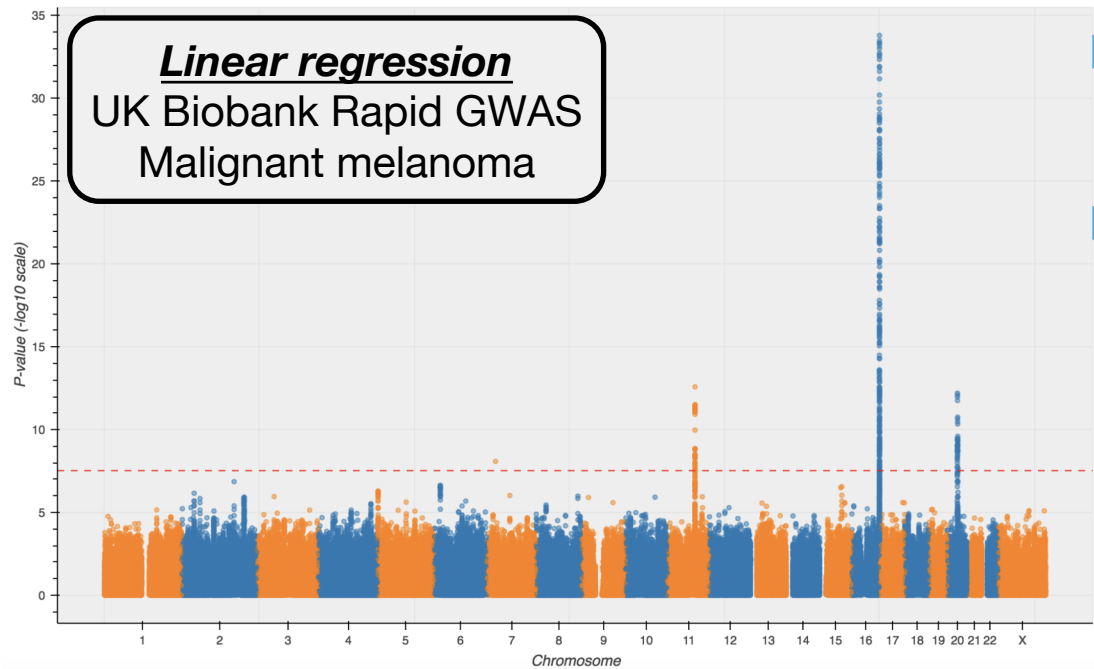




# Hail as a data science library

Data slinging

Analytical toolbox

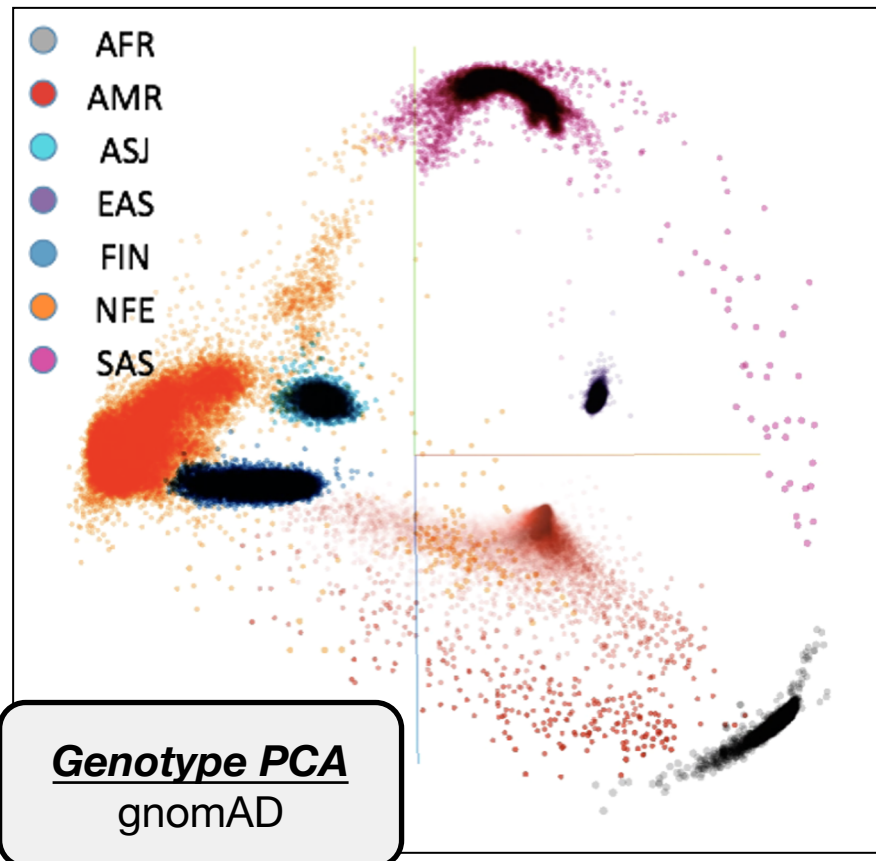


- **Statistical methods for genetics**
- Linear algebra

# Hail as a data science library

Data slinging

Analytical toolbox

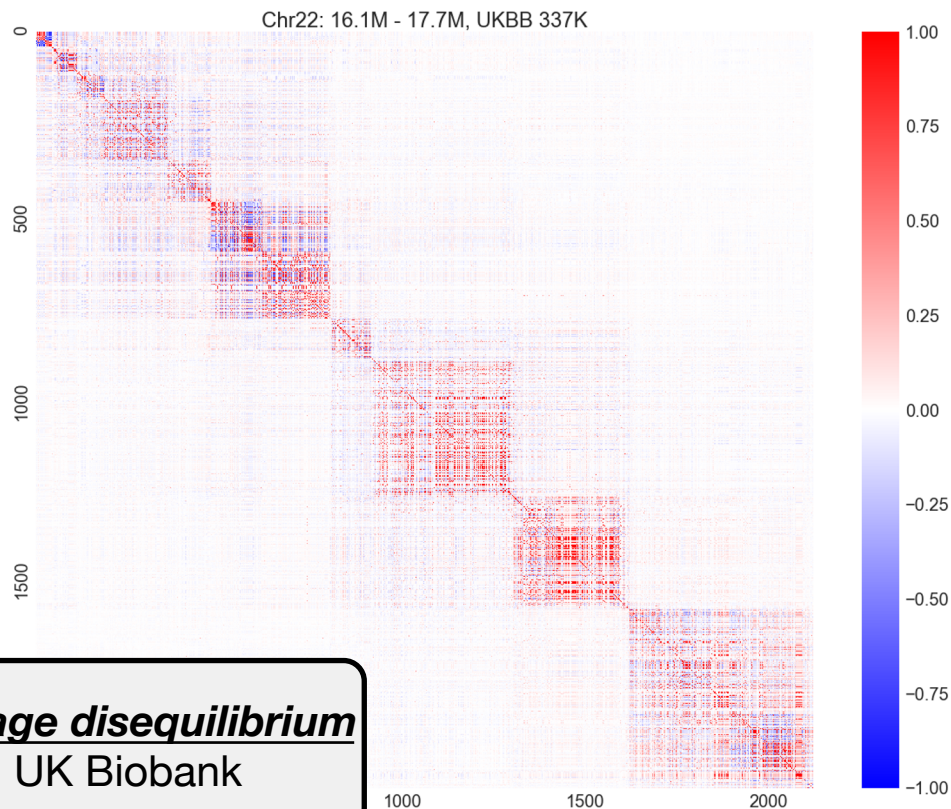


- **Statistical methods for genetics**
- Linear algebra

# Hail as a data science library

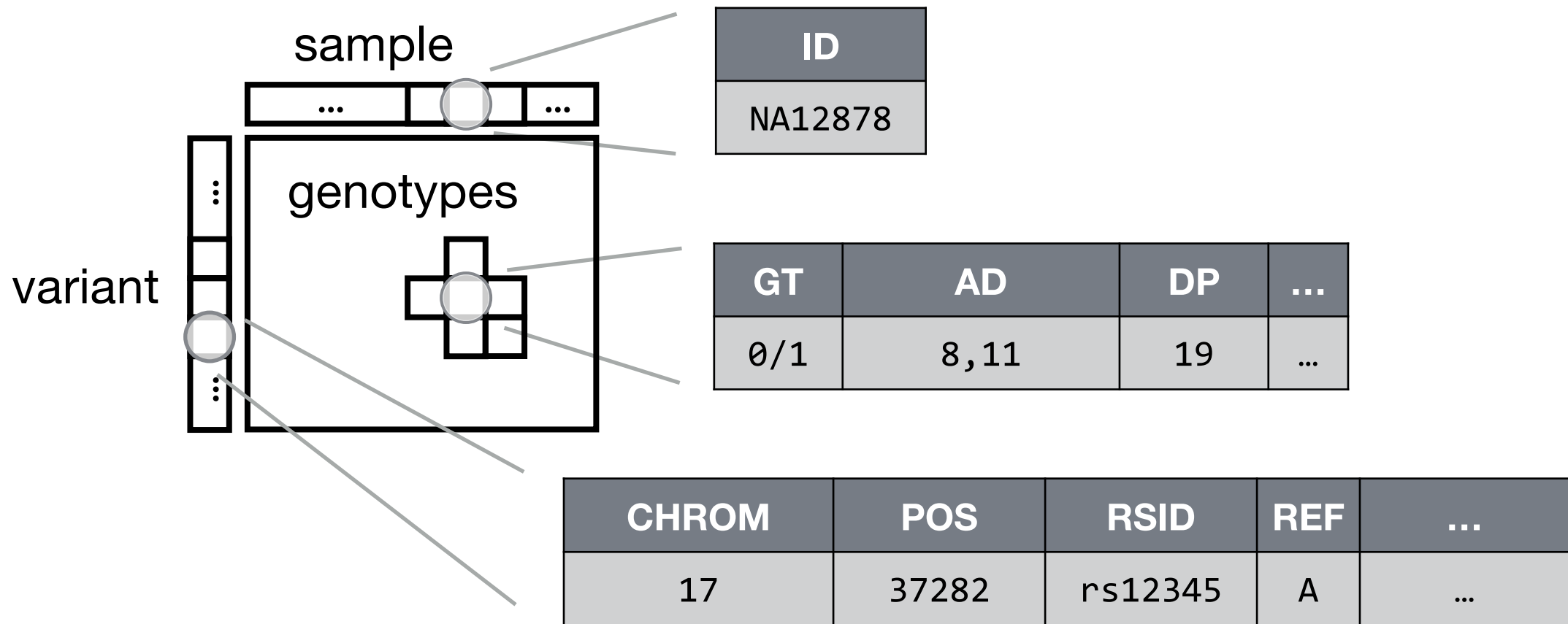
Data slinging

Analytical toolbox

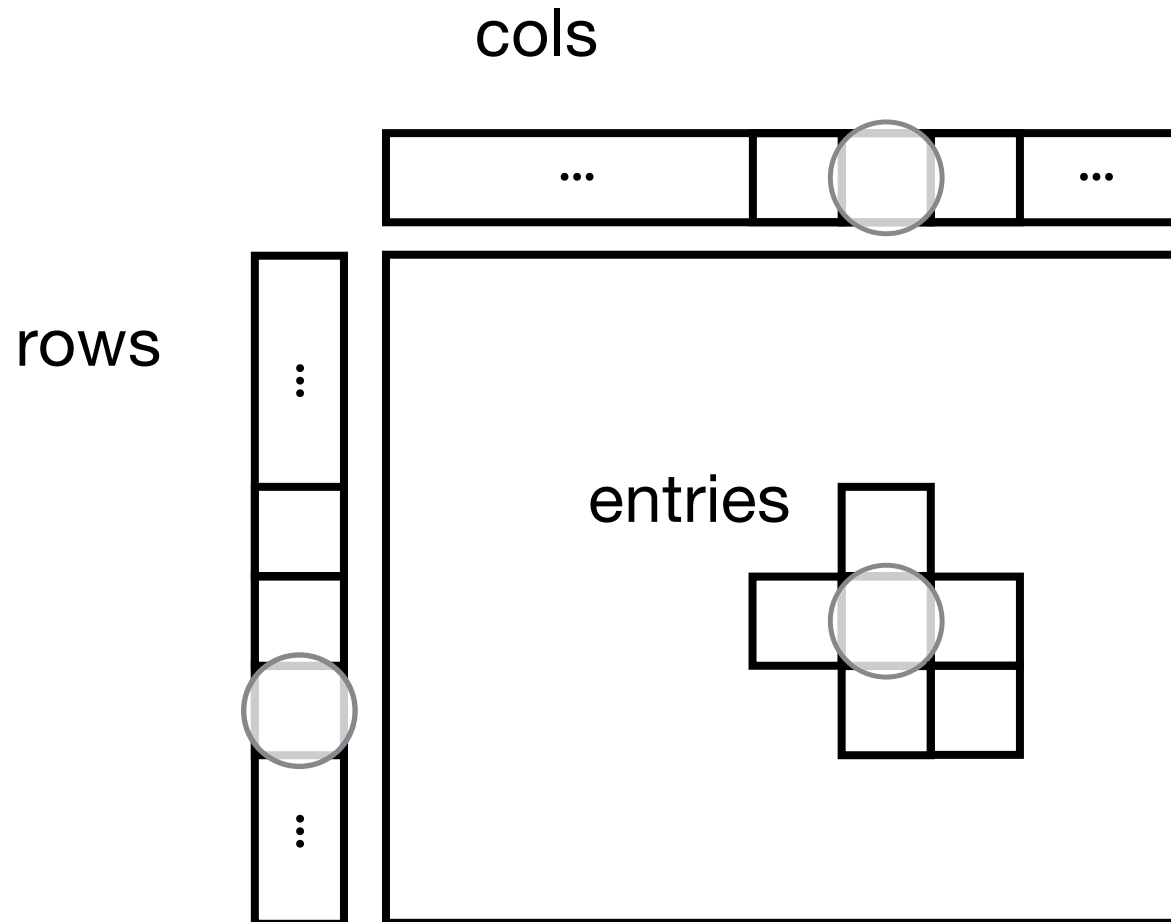


- Statistical methods for genetics
- **Linear algebra (early stages)**

# Variant Call Format (VCF)



# 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

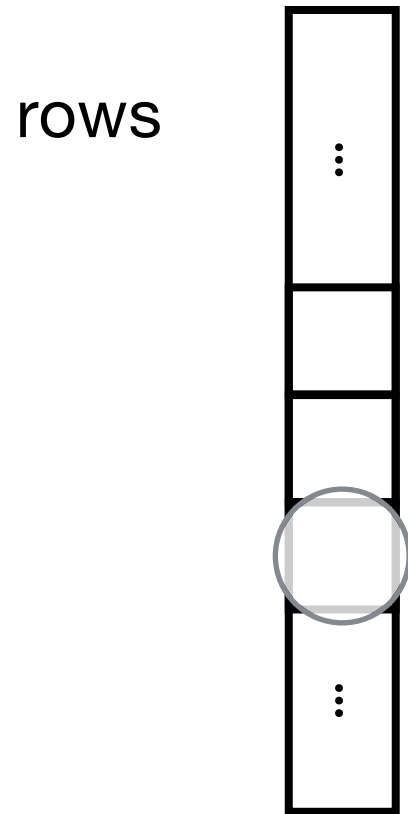
Row key:  
'locus': locus<GRCh37>  
'alleles': array<str>

---

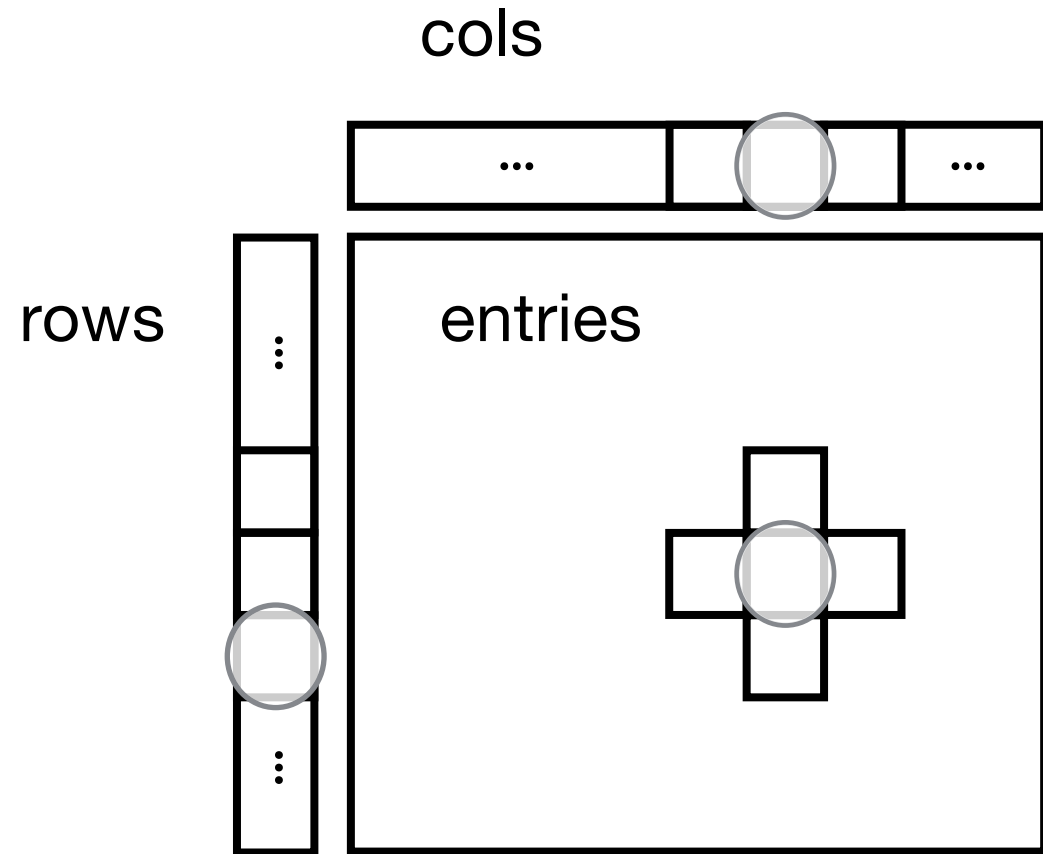
*Can be extended to rare variant aggregation, trio, transcript expression*



# Table



# MatrixTable



Hands on using  
[workshop.hail.is](https://workshop.hail.is)

workshop name: atgu\_workshop2020

password: atgu

#ATGUstrong

# Your next steps

```
pip install hail
```

[DOCS](#)[FORUM](#)[POWERED-SCIENCE](#)[BLOG](#)[WORKSHOP](#)[Hail Docs \(0.2\)](#)[Installation](#)[Hail on the Cloud](#)[Tutorials](#)[Reference \(Python API\)](#)[Overview](#)[How-To Guides](#)[Cheatsheets](#)[Docs](#) » [Hail 0.2](#)[hail.is/docs/](https://hail.is/docs/)[View page source](#)

## 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.

[HOME PAGE](#)[HAIL DOCUMENTATION](#)[HAIL FORUM](#)[HAIL POWERED-SCIENCE](#)[HAIL BLOG](#)[HAIL WORKSHOPS](#)[blog.hail.is/](https://blog.hail.is/)[GENOMICS](#)

## 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](https://discuss.hail.is)[Sign Up](#)[Log In](#)[About](#)[FAQ](#)[Terms of Service](#)[Privacy](#)

### About Hail Discussion

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

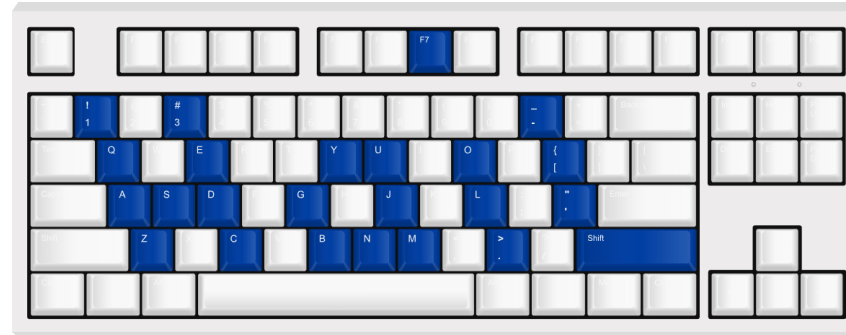




STANLEY CENTER  
FOR PSYCHIATRIC RESEARCH  
AT BROAD INSTITUTE



BROAD  
INSTITUTE



# Thank you!

## ATGU Welcome Workshop

*Have questions? We may have answers!*

Kumar Veerapen, PhD  
*Hail Support and Community Outreach Manager*  
Arcturus Wang  
*Software Engineer*



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