

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 Rare Variants

---

## ATGU Welcome Workshop

August 12<sup>th</sup>, 2020

2:00 – 4:00 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

- Recap: What is Hail?
- Recap: Rare Variant Analysis Lecture
- Rare Variant Analysis using Hail
- Population unmasking (if we have 30 minutes)
- What now?

# 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

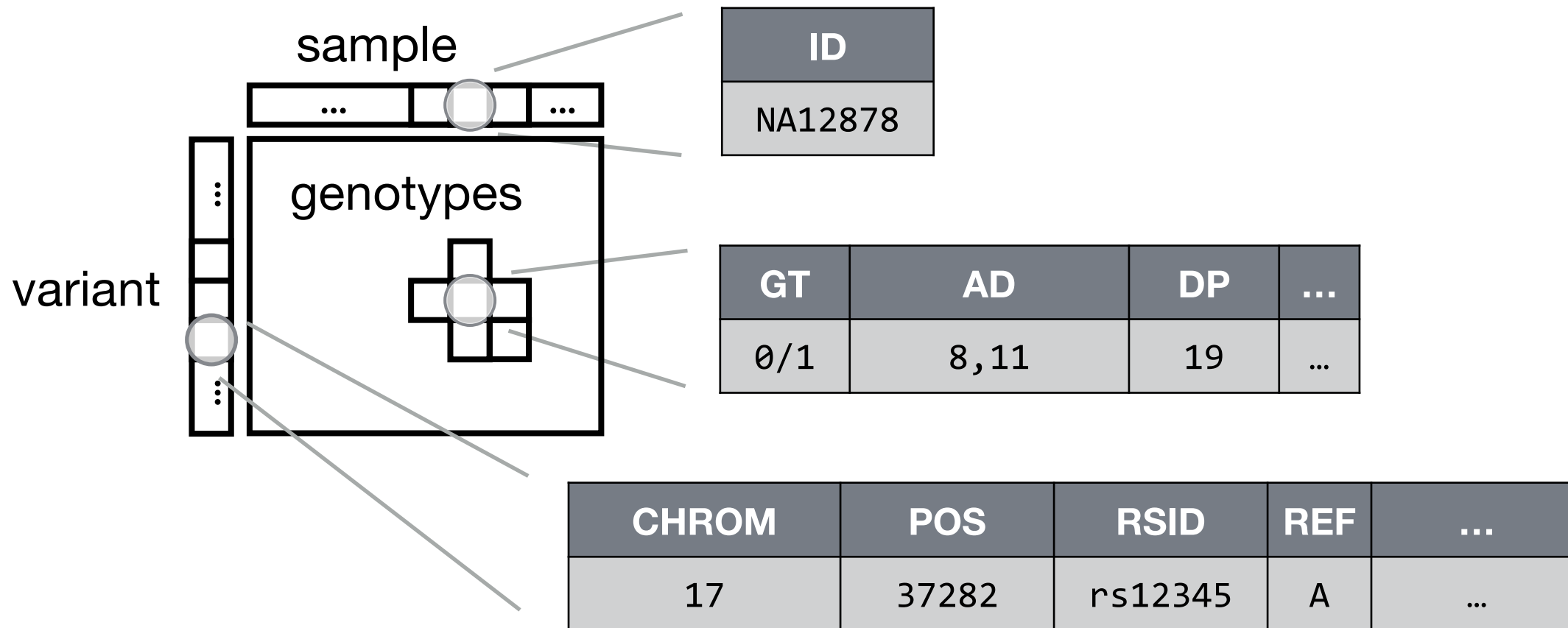
Tabular data frames  
imported as Hail  
MatrixTables into  
**unified platform.**



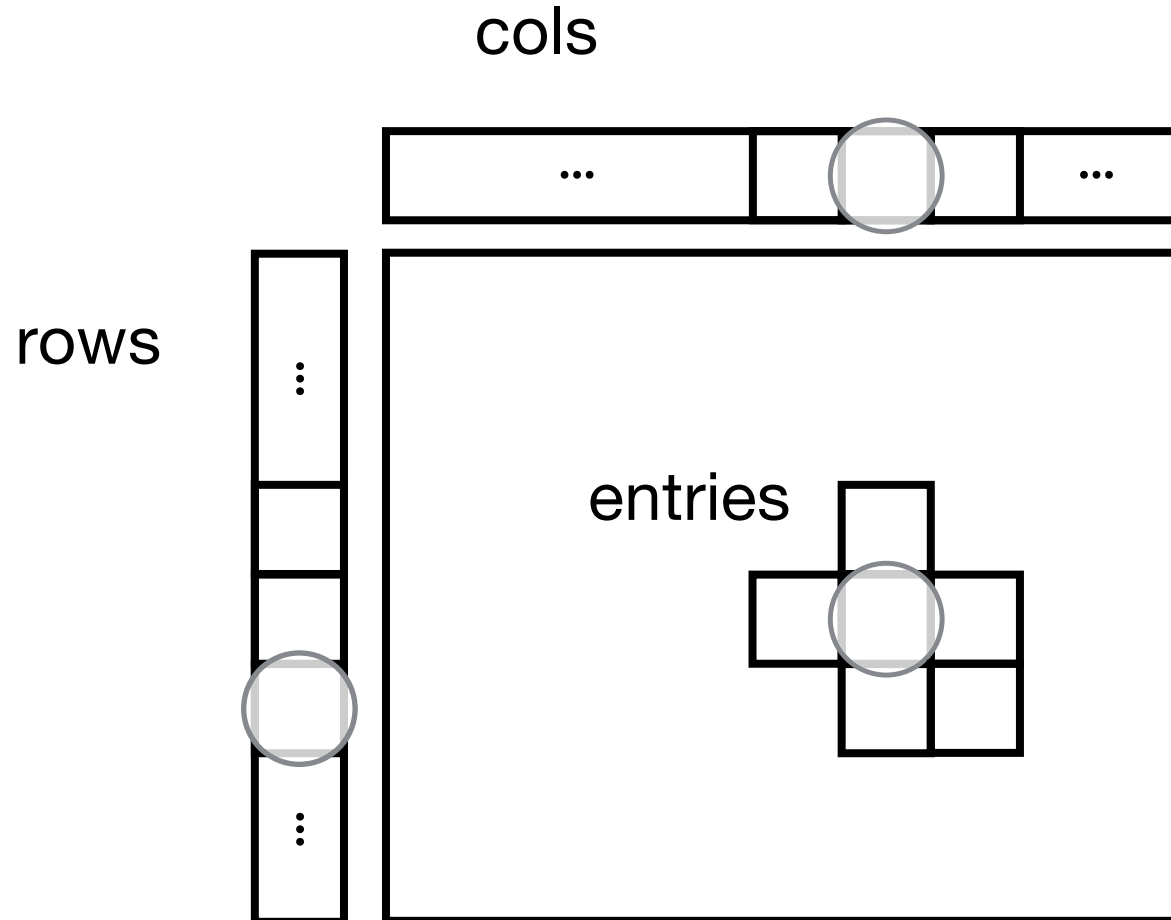
Learn more at [Hail.is](https://hail.is)

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

# 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>

---

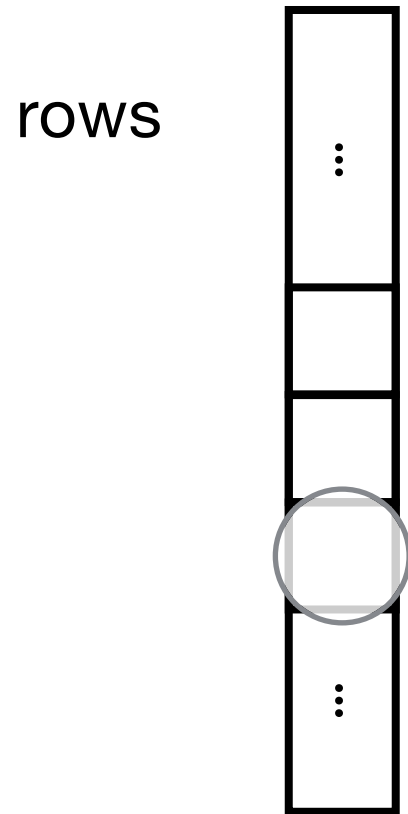
Column key:  
's': str

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

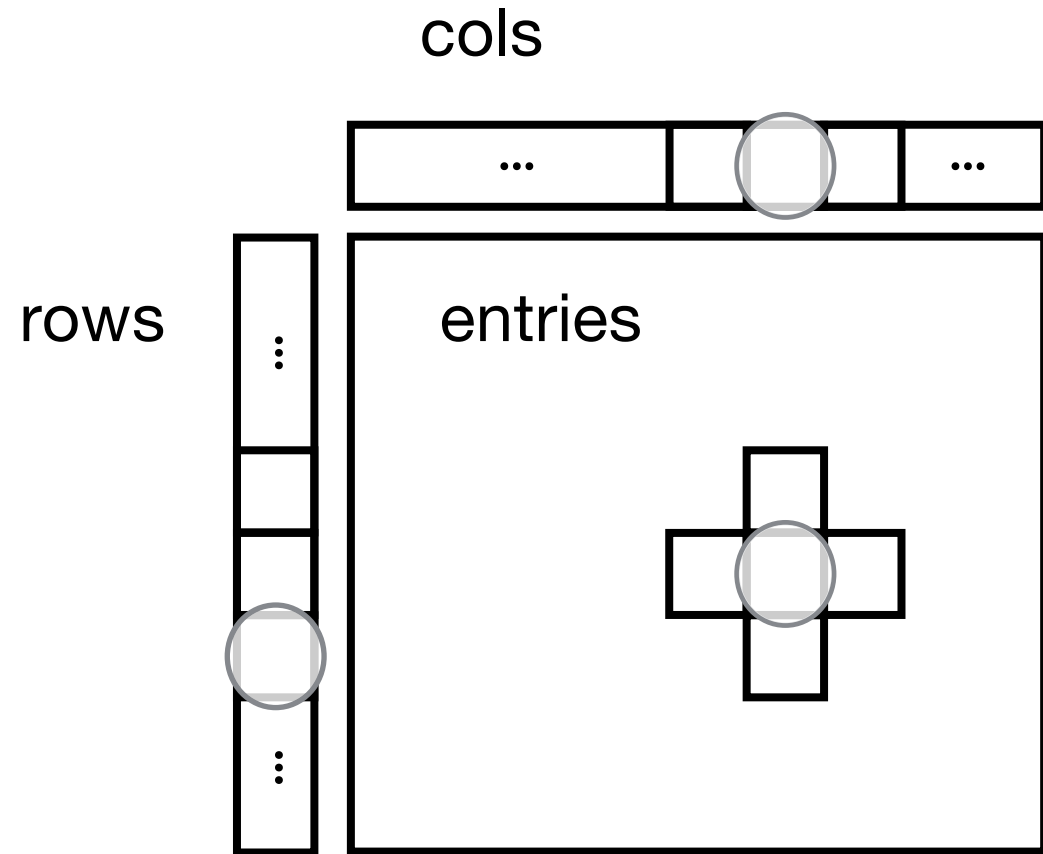
---

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

# Table



# MatrixTable



# Common versus rare

Binary trait – e.g., disease (schizophrenia) or tail of distribution (LDL > 200 md/dl)

## Common variants

Pick SNP



Compare in cases vs controls,  
Calculate effect size

## Rare variants

Pick a gene

Which mutations to **aggregate**?

- coding region
- non-synonymous

Which mutations to **filter**?

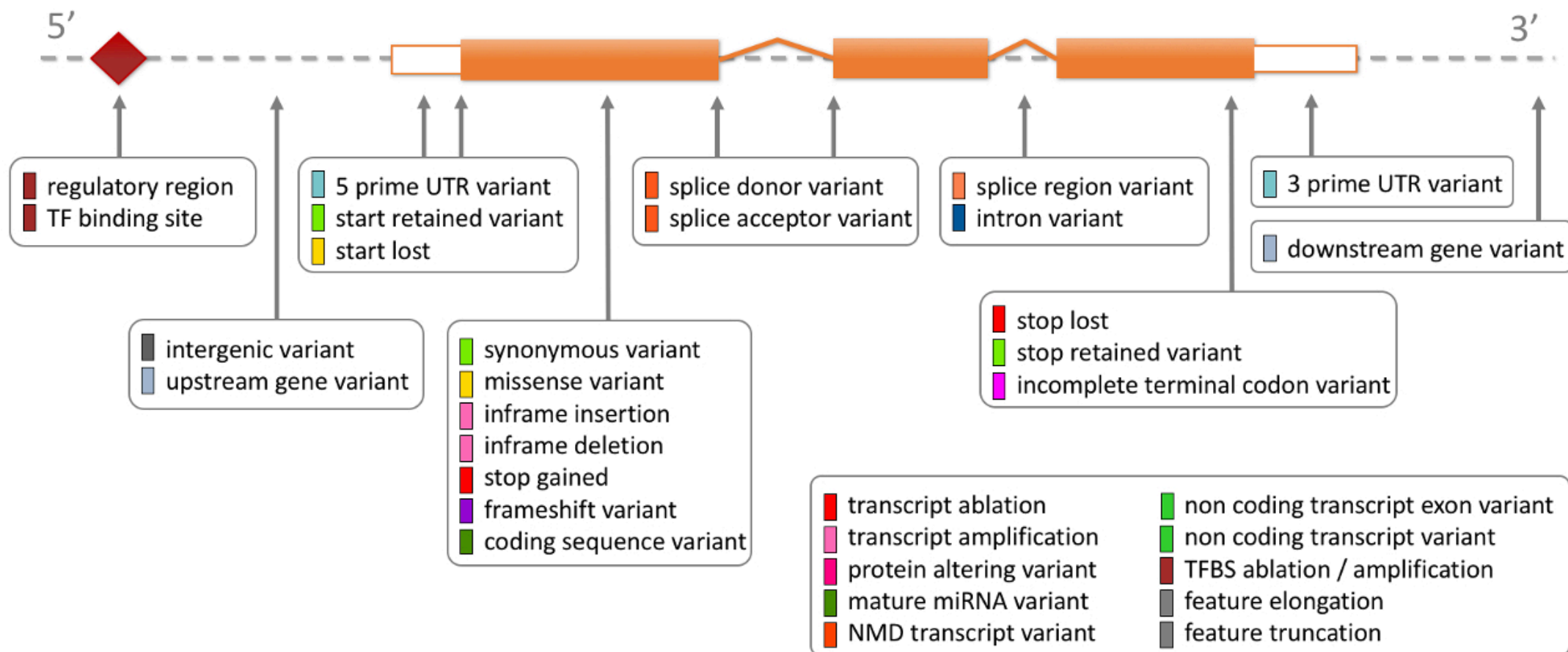
- By type (missense?)
- By frequency
- By predicted function



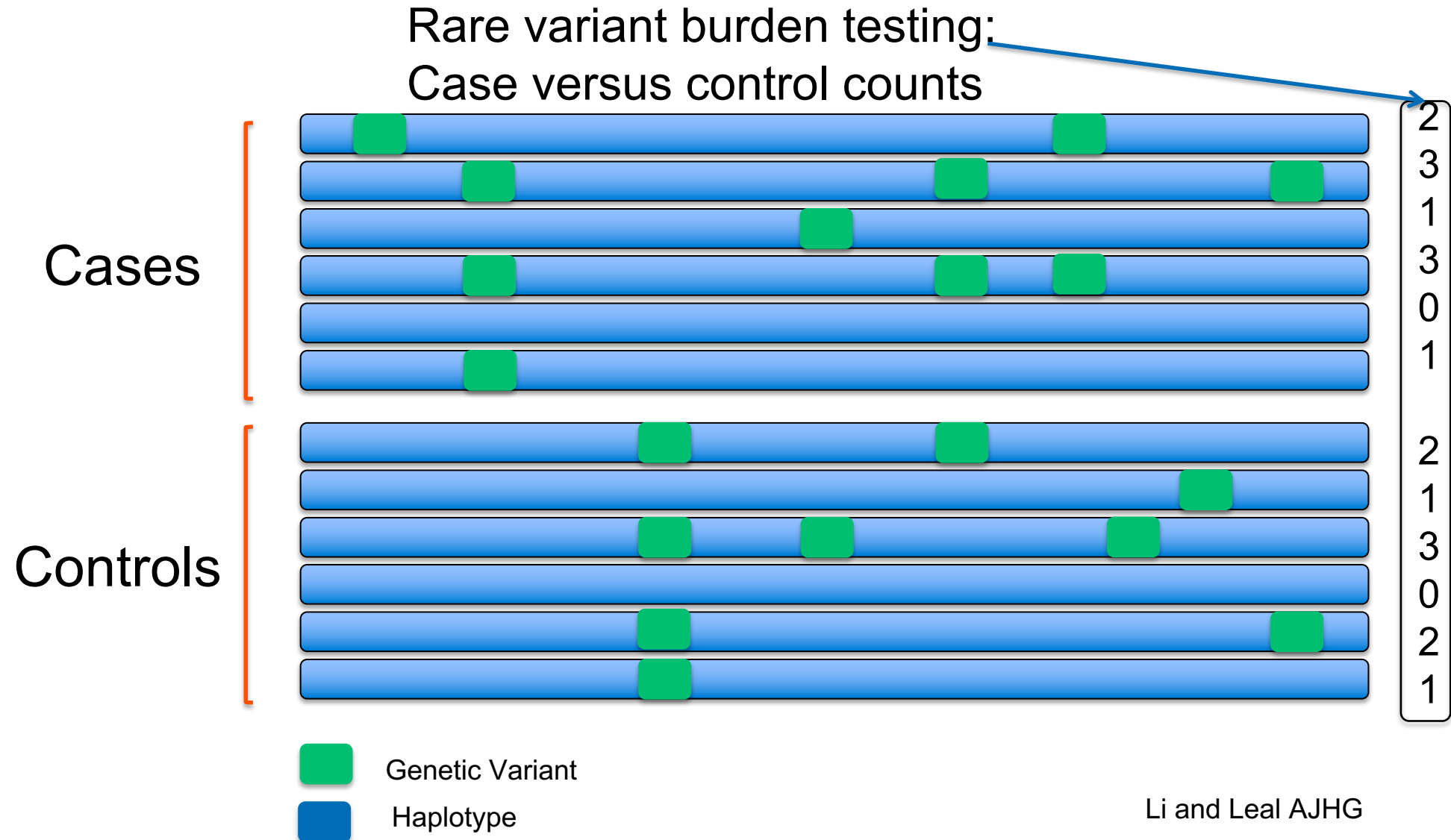
Compare in cases vs controls,  
Infer effect size



# Exonic variant annotation



# Visual representation of sequence data testing



# Rare Variant Analysis with Hail

- Group variants with similar annotations
- Run a burden with Hail
- Output:
  - summary statistics (beta, p-value, etc) for each group (e.g. gene + annotation) for each phenotype

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>)

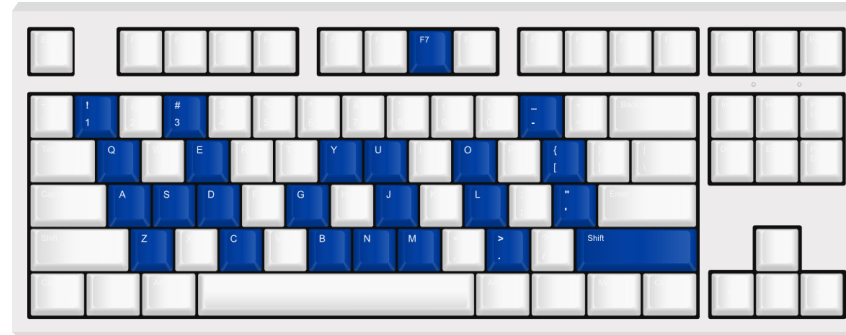




STANLEY CENTER  
FOR PSYCHIATRIC RESEARCH  
AT BROAD INSTITUTE



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