Introduction to hail

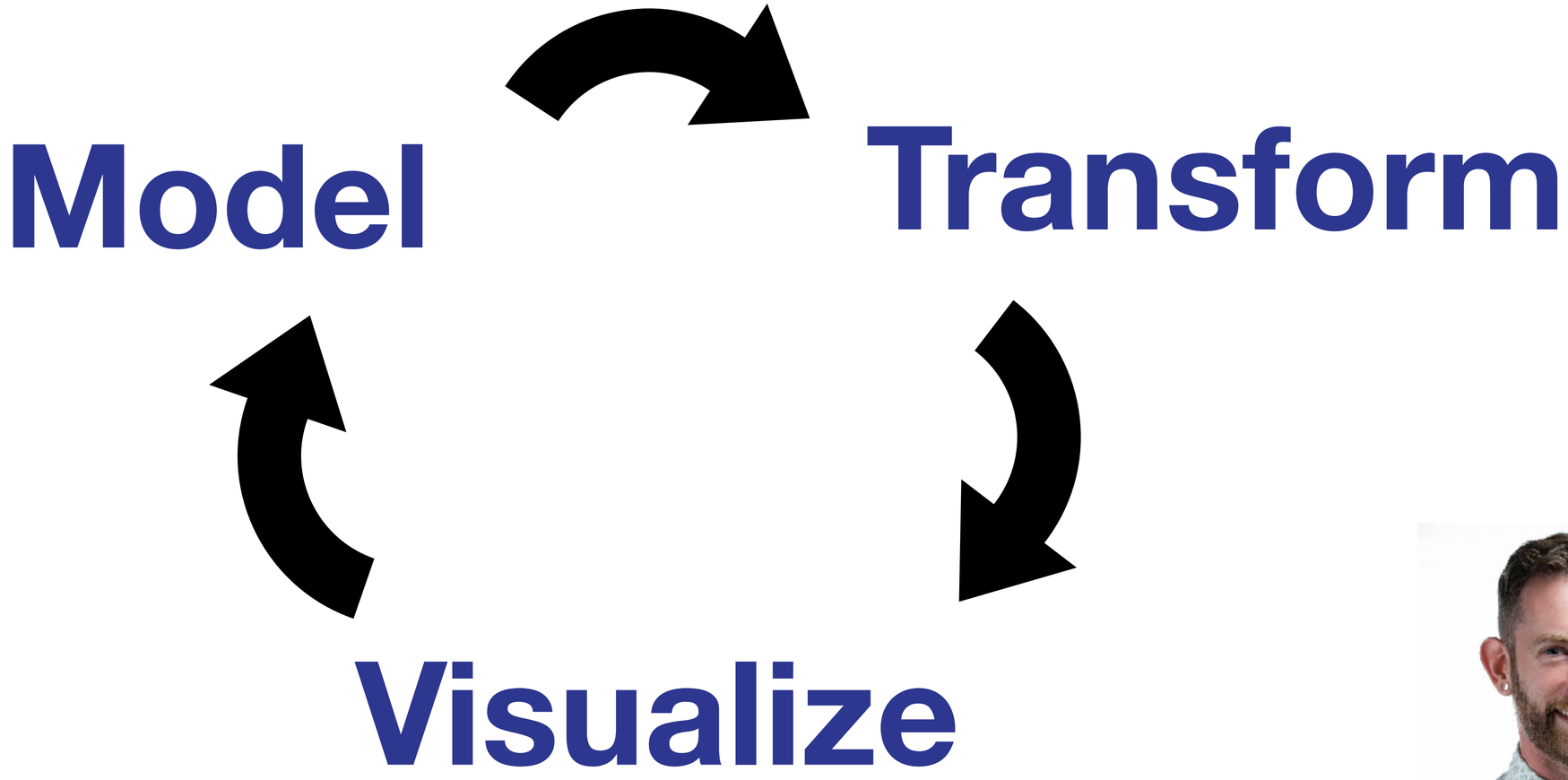
Institute for Behavioral Genetics

March 7, 2019

Outline

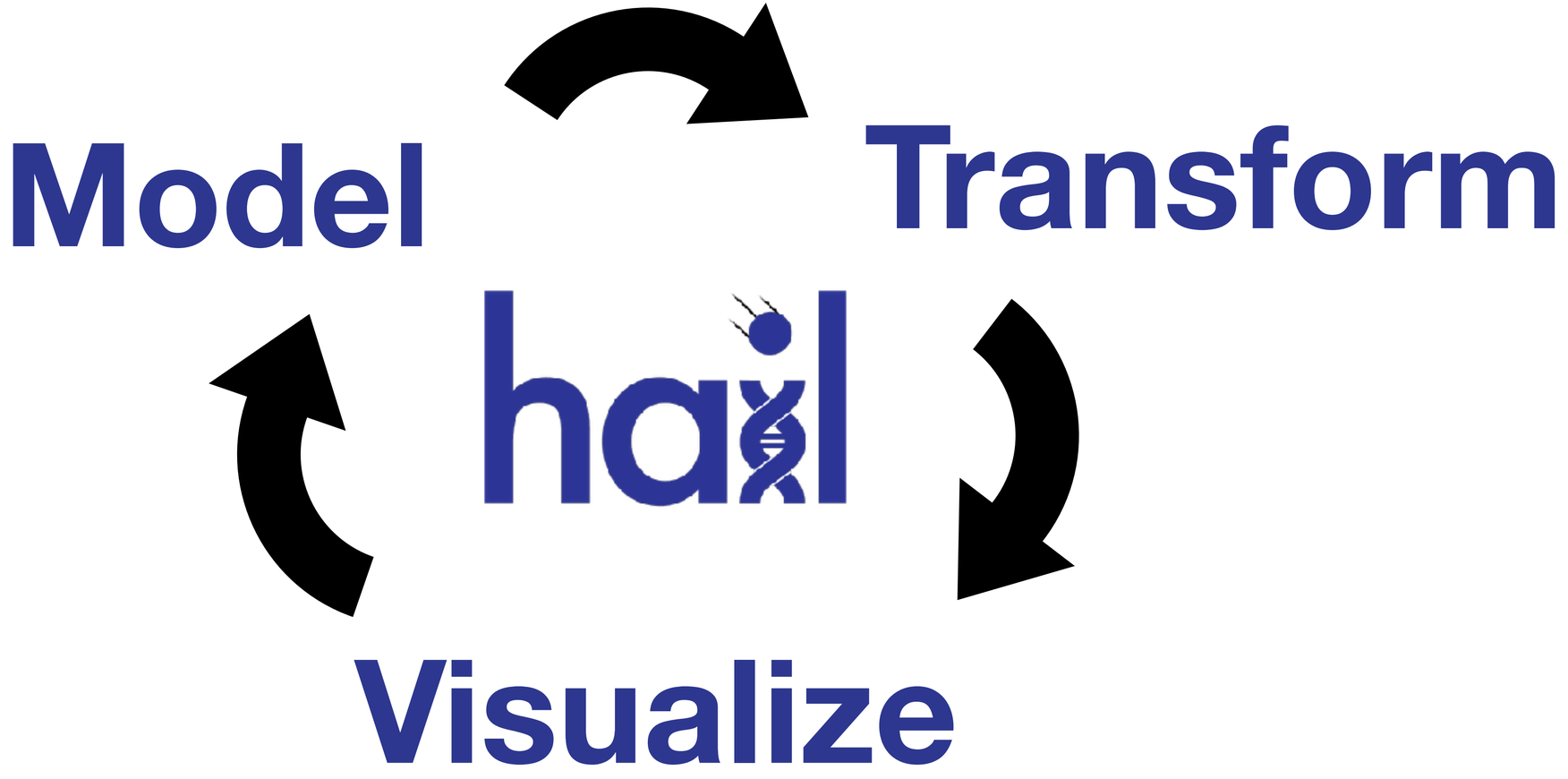
- **Introduction to Hail**
- Practical 1: QC
- Practical 2: GWAS
- Computational Landscape for Bioinformatics
- Practical 3: Inferring Ancestry
- Practical 4: Computing F_ST
- Practical 5: Gene Burden Test
- Pratical 6: De Novo Caller

Understanding data



Hadley Wickham, 2018

Understanding sequencing data



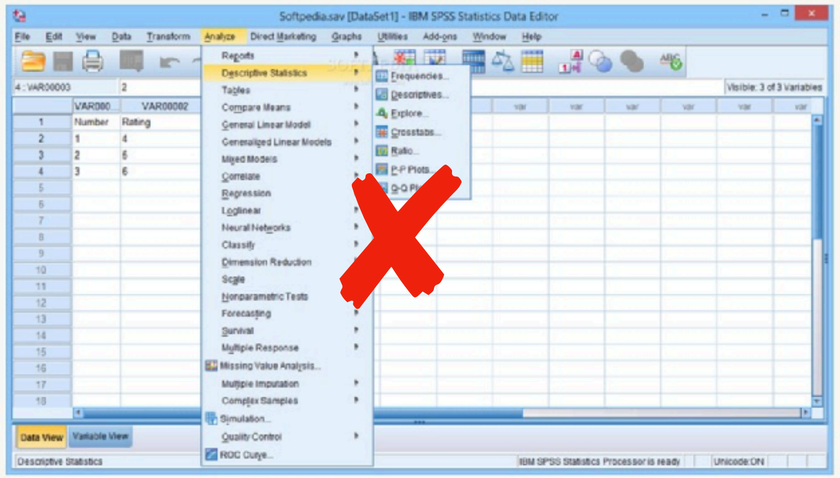
Why is programming preferable for statistics?

1. Code is text
2. Code is read-able
3. Code is shareable
4. Code is open

Hadley Wickham, 2018

Why is programming preferable for statistics?

1. Code is text
2. Code is read-able
3. Code is shareable
4. Code is open



Hadley Wickham, 2018

The language of data science

- Don’t memorize the ***vocabulary***, learn the ***grammar***.
  - To learn vocabulary: Hail documentation!

- Don’t worry about the function names and exact code syntax, think about **what’s happening to transform the data**!

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

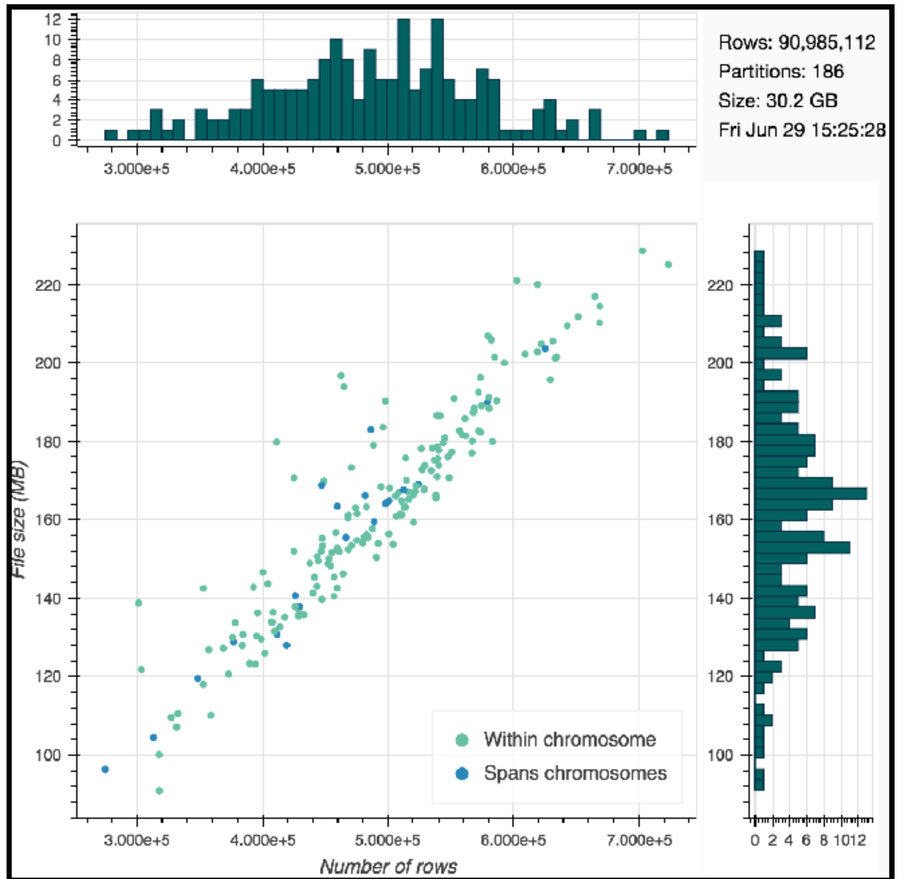
- Built-in wrappers for VEP, Nirvana
- Join with annotations by variant, locus, interval, gene
- ReferenceGenome is a first-class concept, for all our sanity

Hail as a data science library

Data slinging

Analytical toolbox

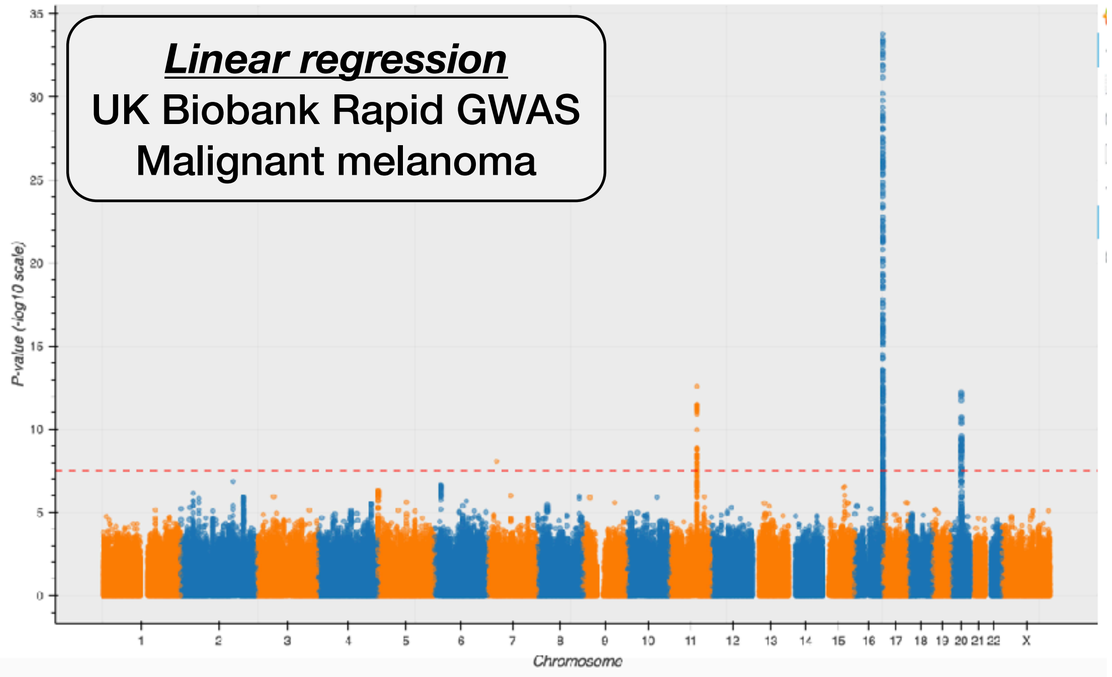
- Read and write common formats
- Filter, group, aggregate
- Annotation
- **Visualization**



Hail as a data science library

Data slinging

Analytical toolbox

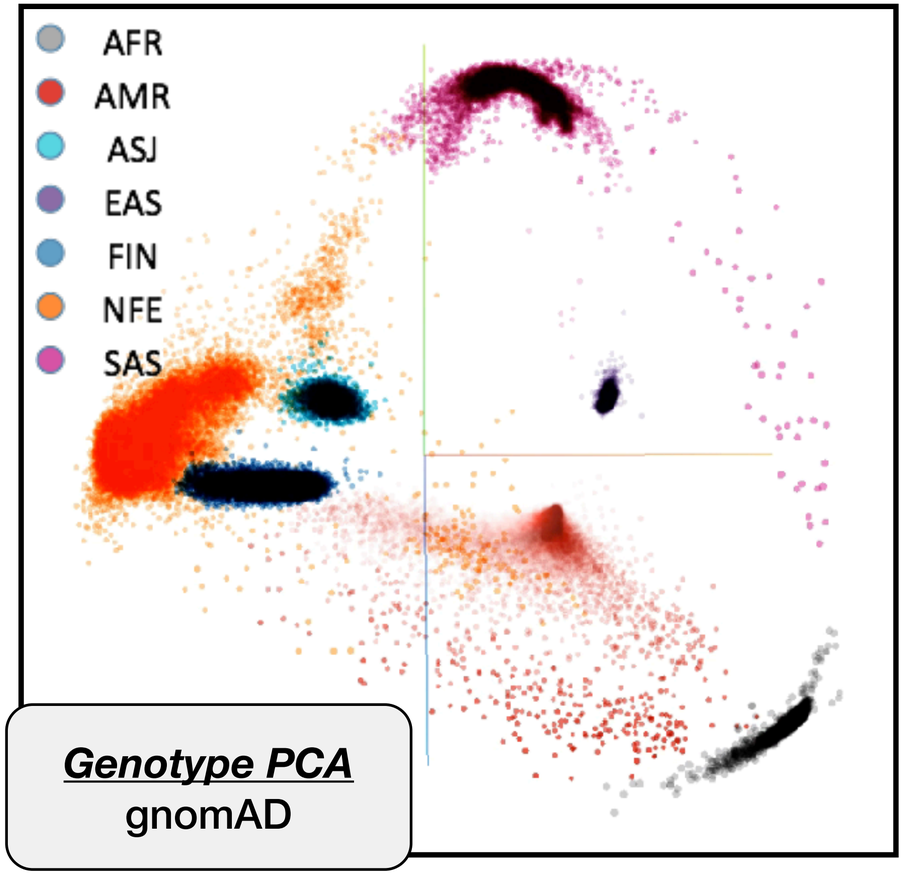


- **Statistical methods for genetics**
- Scalable linear algebra

Hail as a data science library

Data slinging

Analytical toolbox



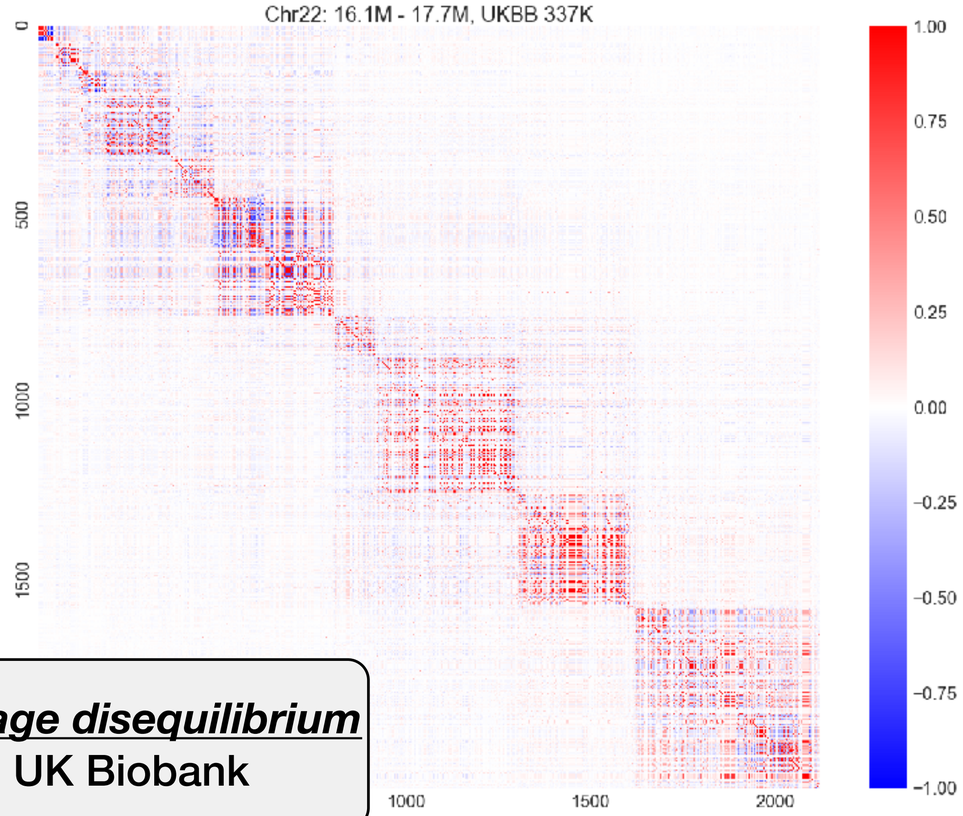
Genotype PCA
gnomAD

- **Statistical methods for genetics**
- Scalable linear algebra

Hail as a data science library

Data slinging

Analytical toolbox



Linkage disequilibrium
UK Biobank

- Statistical methods for genetics
- **Scalable linear algebra**

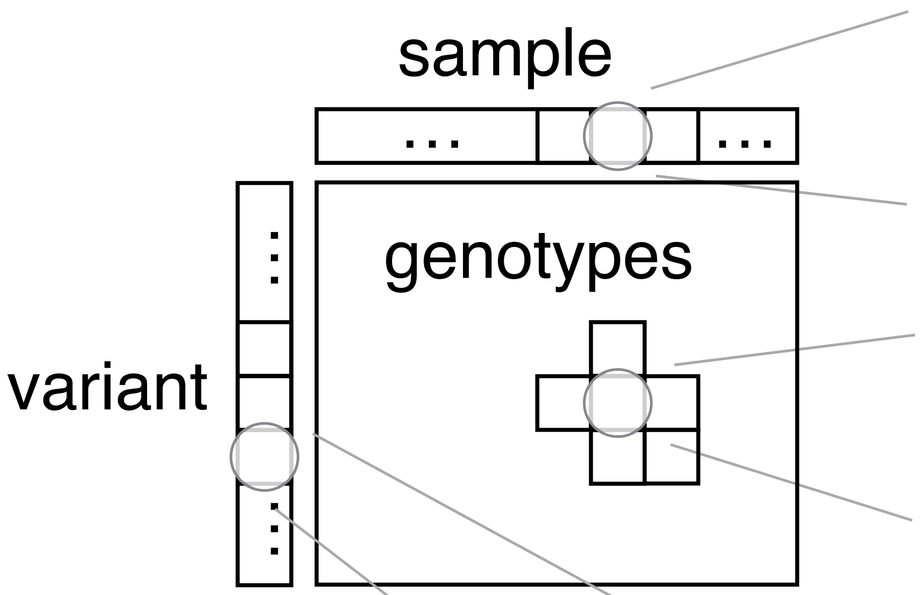
Hail as a scientific computing stack

- Most of the tools you need*, together in one place

- Worry-free scalable underlying infrastructure so you can build the rest!

***We can’t read your minds**

Variant Call Format (VCF)

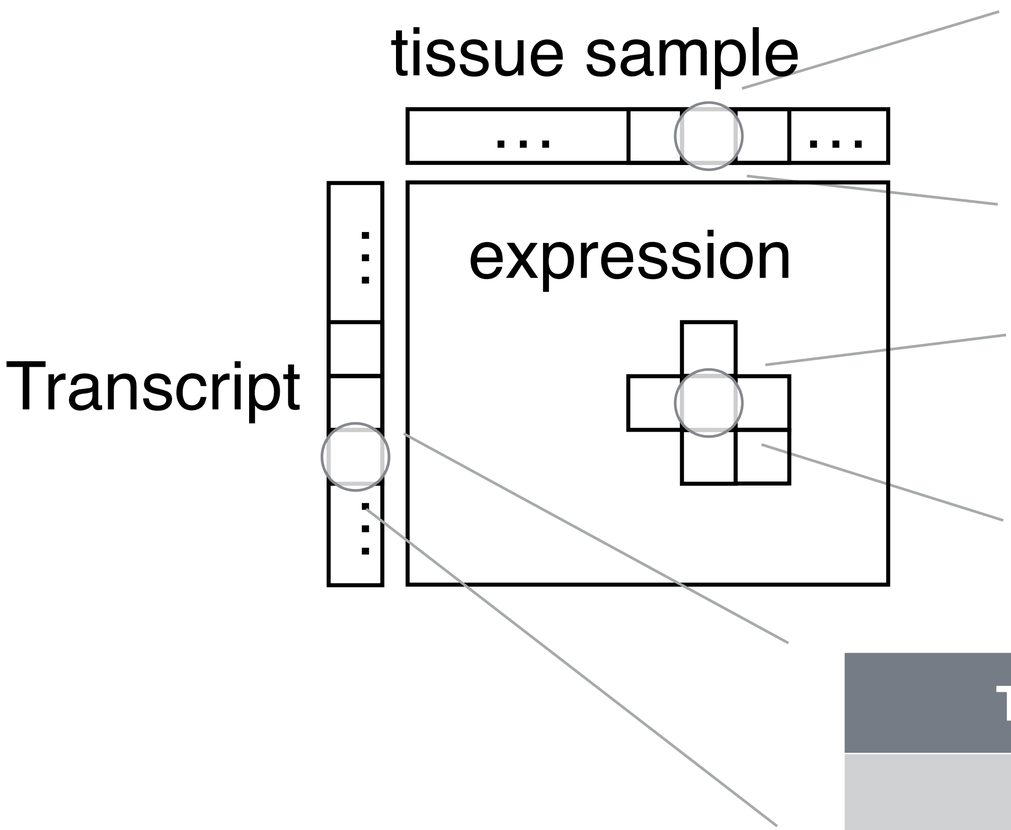


ID
NA12878

GT	AD	DP	…
0/1	8,11	19	…

CHROM	POS	RSID	REF	…
17	37282	rs12345	A	…

Transcript expression

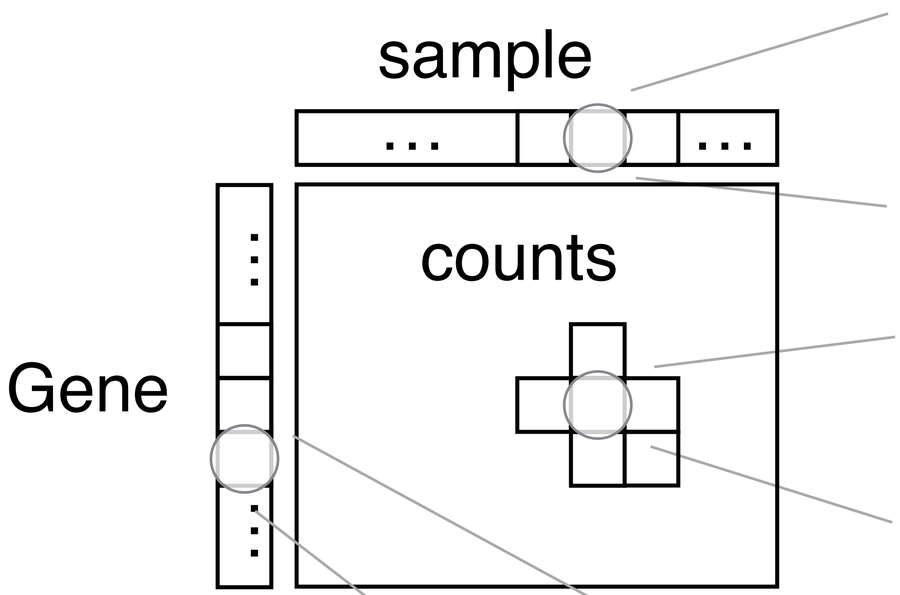


Sample	Tissue
NA12878	Liver

Raw_expression	Norm_expression
2	4.1232

Transcript	Gene
	spectrin

Rare variant aggregation

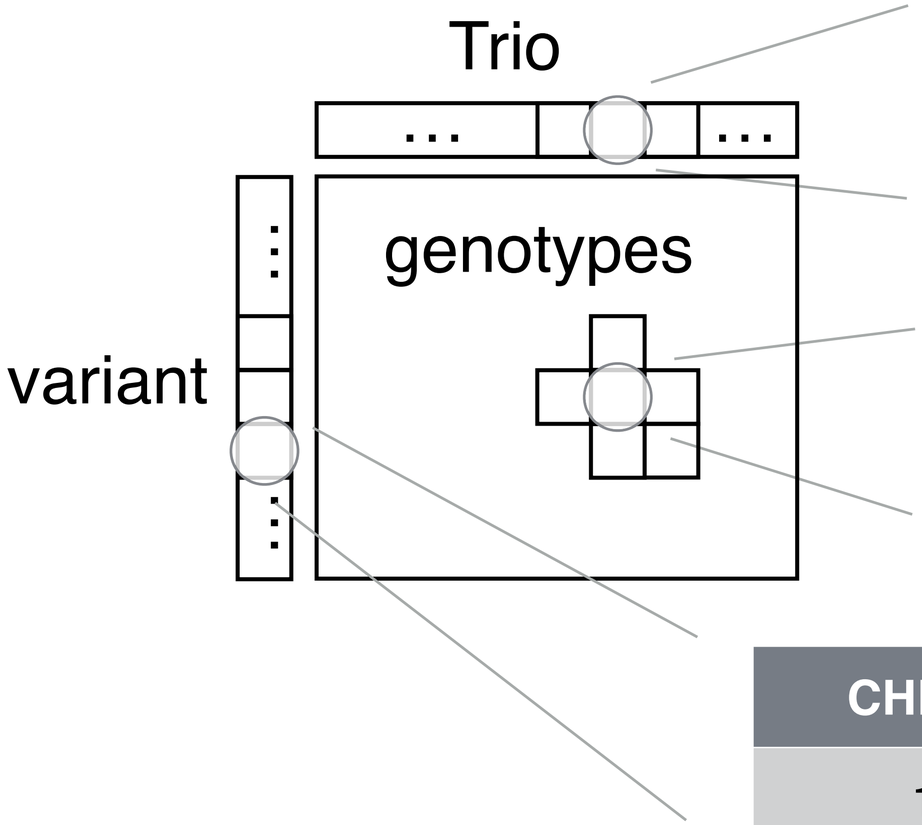


ID
NA12878

N Loss-of-function	N Missense	N Synonymous
1	9	12

Gene
spectrin

Trio data

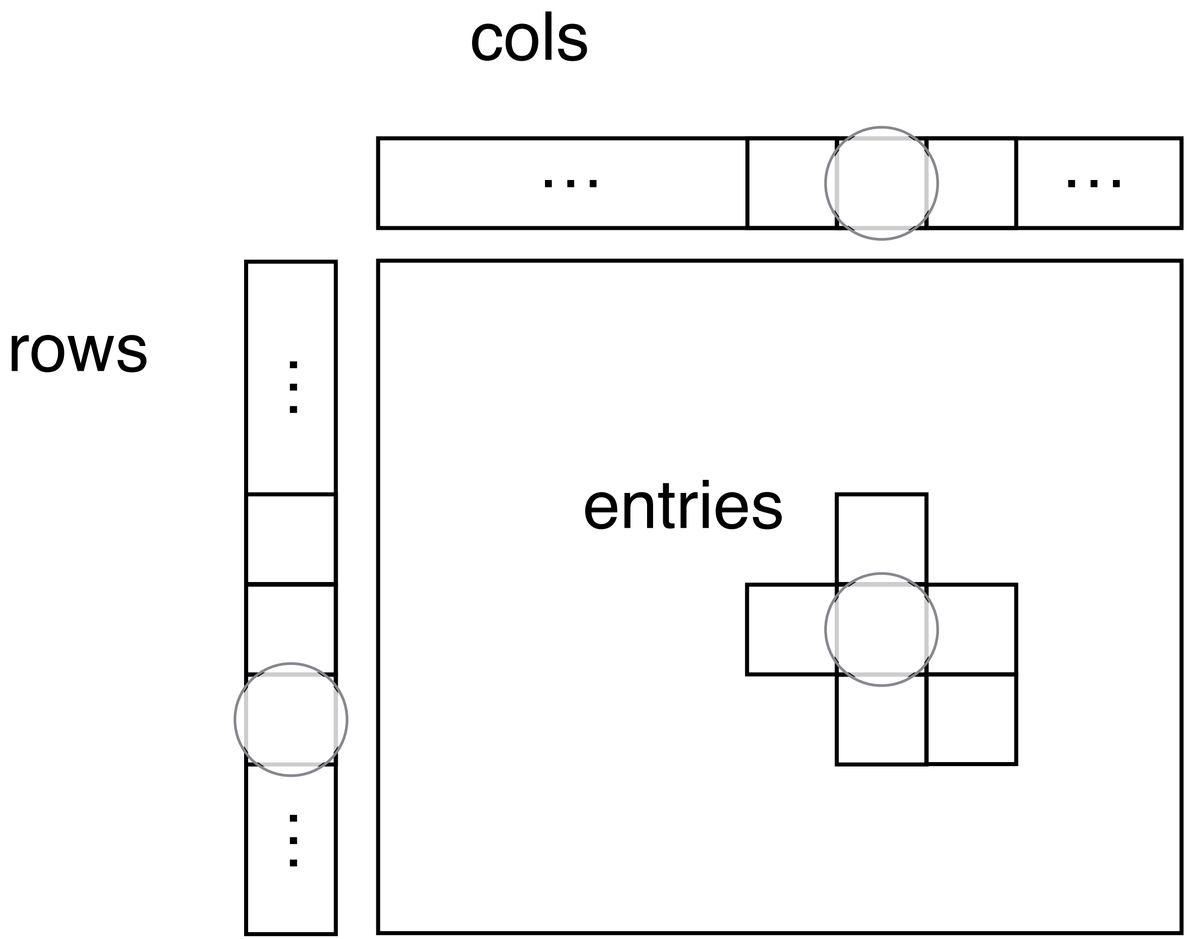


Proband	Mother	Father	…
NA12878	NA12891	NA12892	…

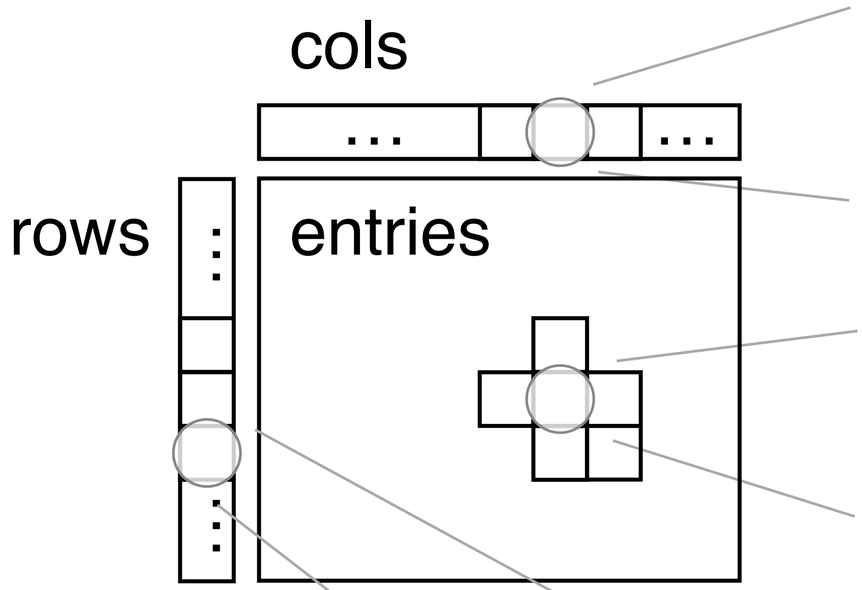
proband_GT	mother_GT	father_GT	…
0/1	0/0	0/0	…

CHROM	POS	RSID	REF	…
17	37282	rs12345	A	…

MatrixTable



MatrixTable

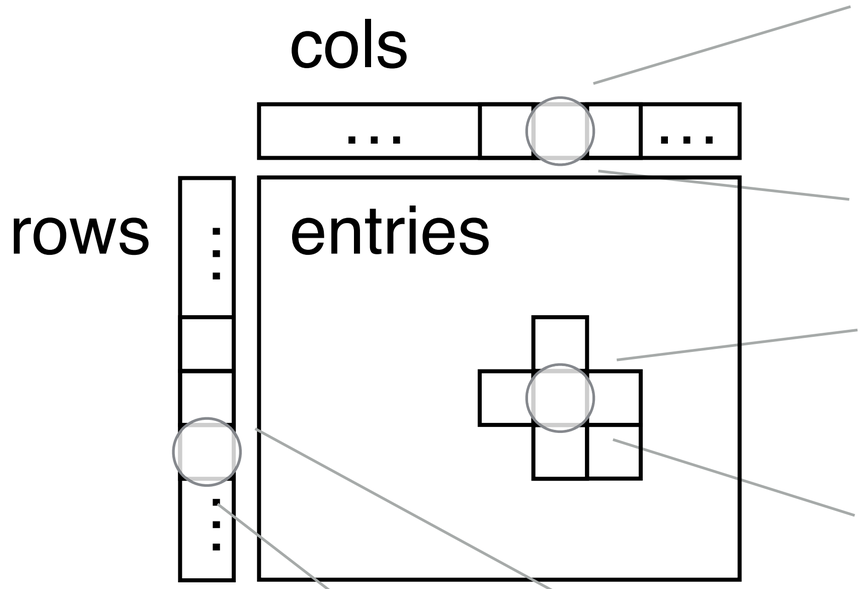


s	is_case	age	…
str	bool	int32	
NA12878	True	67	…

GT	AD	DP	…
call	array<int32>	int32	
0/1	8,11	19	…

locus	ID	…
locus<GRCh37>	str	
17:37282	rs12345	…

MatrixTable

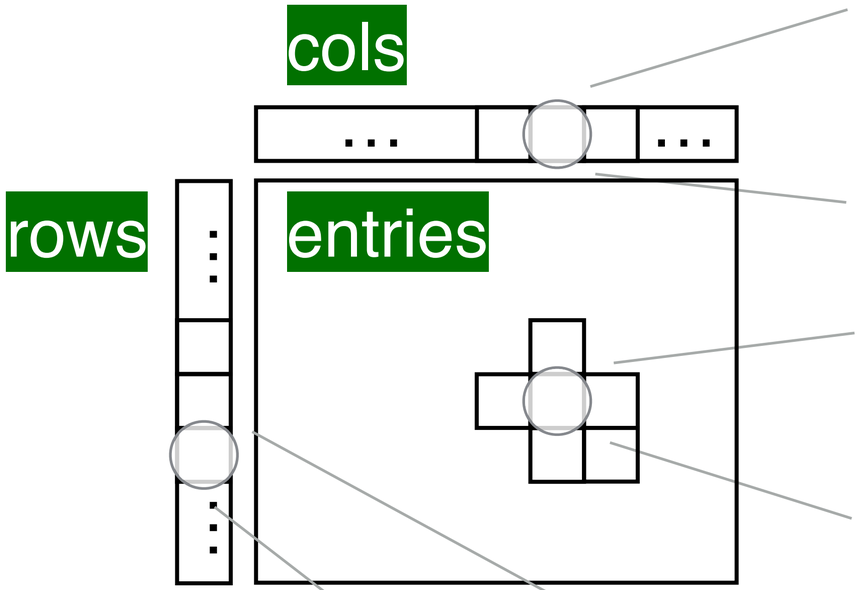


s	is_case	age	…
str	bool	int32	
NA12878	True	67	…

GT	AD	DP	…
call	array<int32>	int32	
0/1	8,11	19	…

locus	ID	…
locus<GRCh37>	str	
17:37282	rs12345	…

MatrixTable

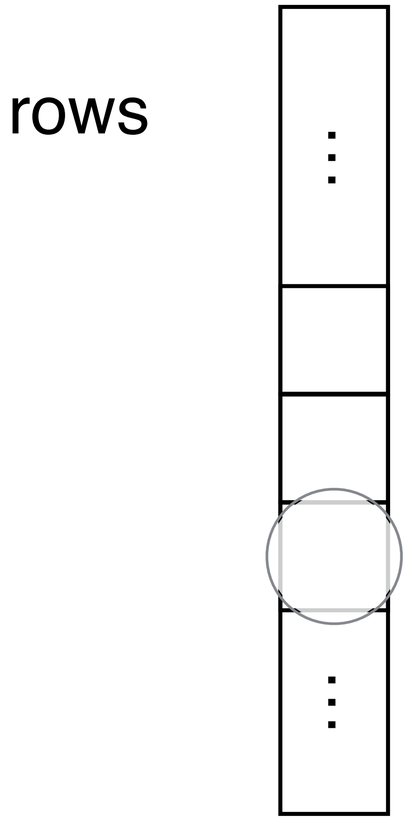


s	is_case	age	…
str	bool	int32	
NA12878	True	67	…

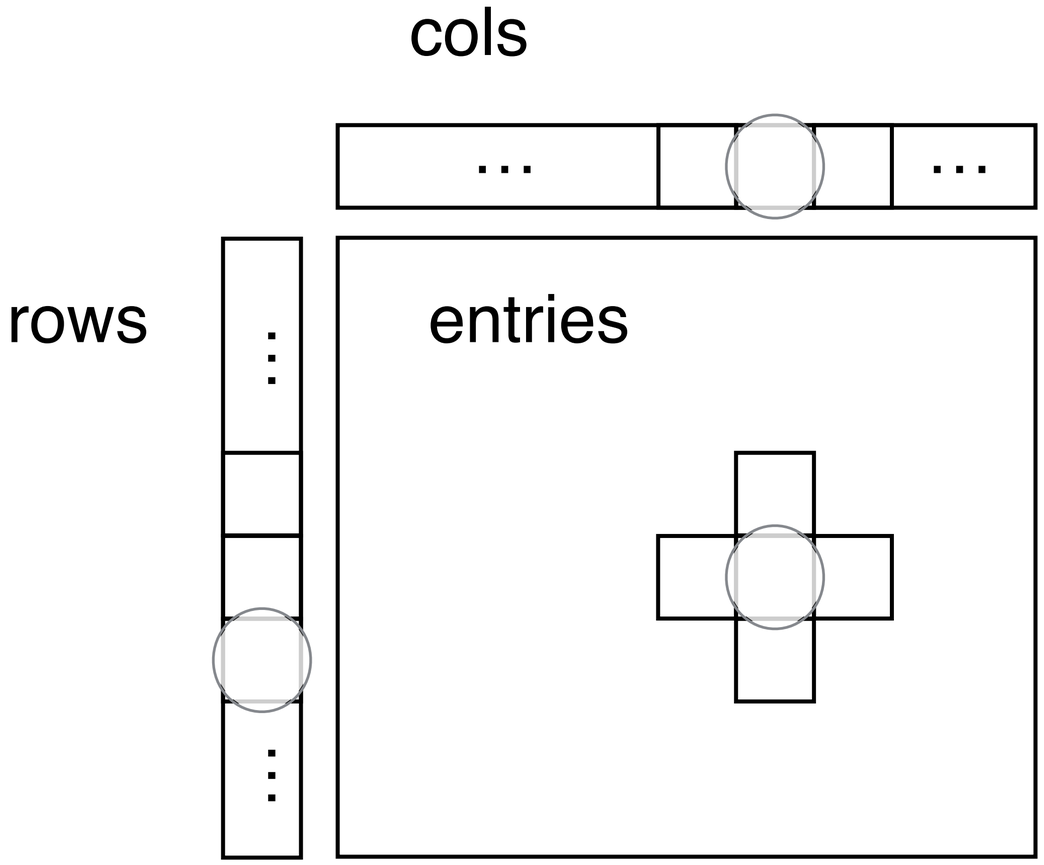
GT	AD	DP	…
call	array<int32>	int32	
0/1	8,11	19	…

locus	ID	…
locus<GRCh37>	str	
17:37282	rs12345	…

Table

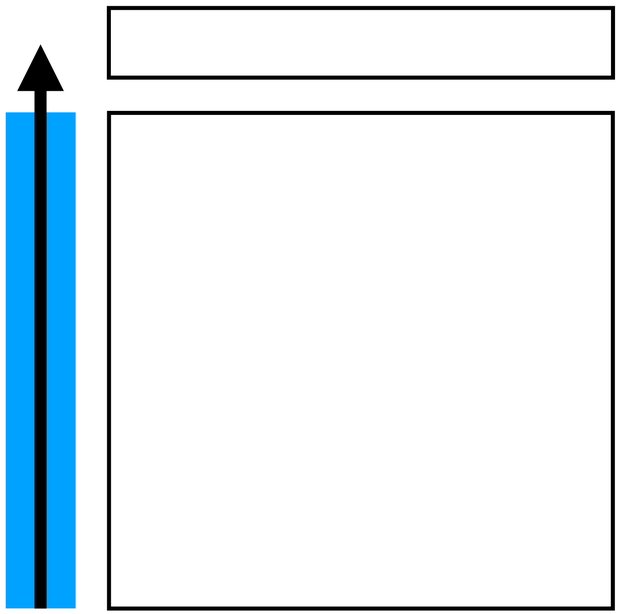


MatrixTable

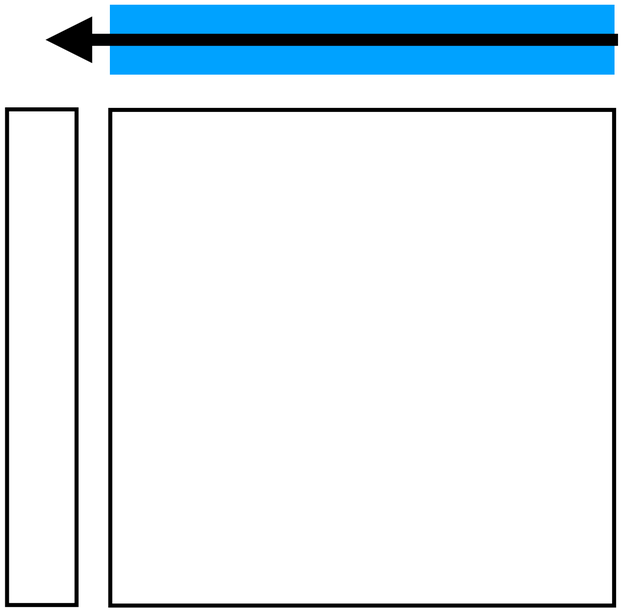


aggregate

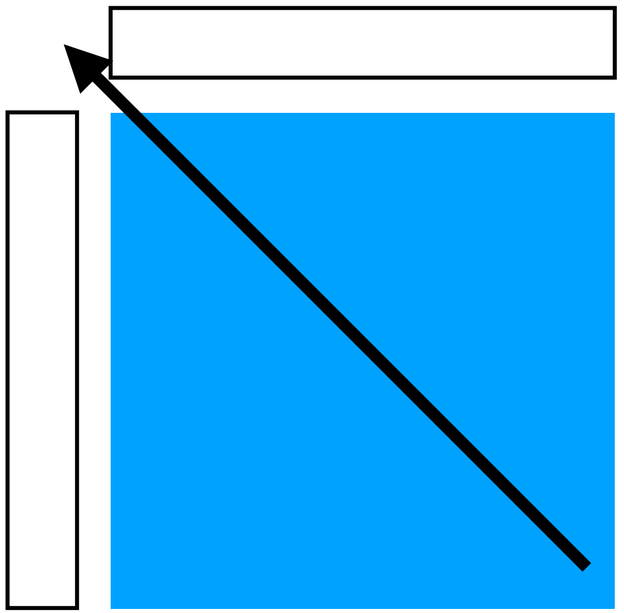
aggregate_rows



aggregate_cols

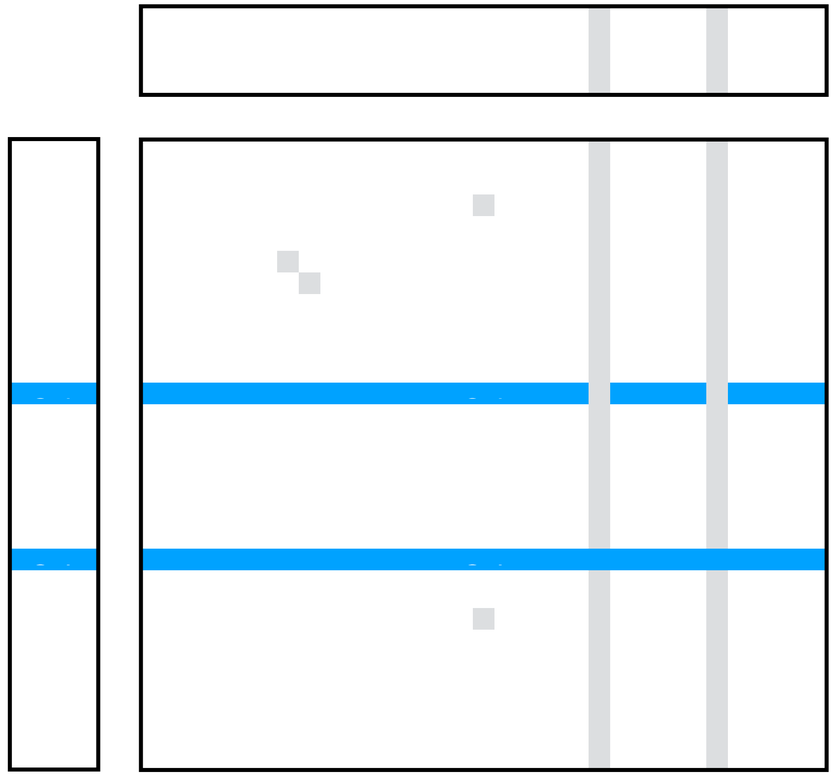


aggregate_entries



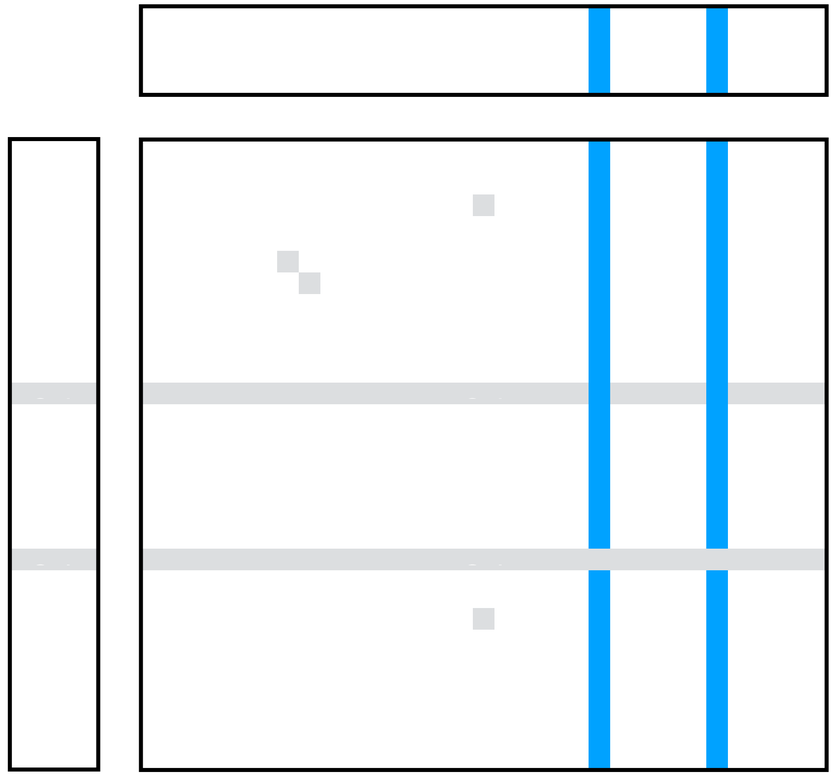
filter

- filter_rows
- filter_cols
- filter_entries



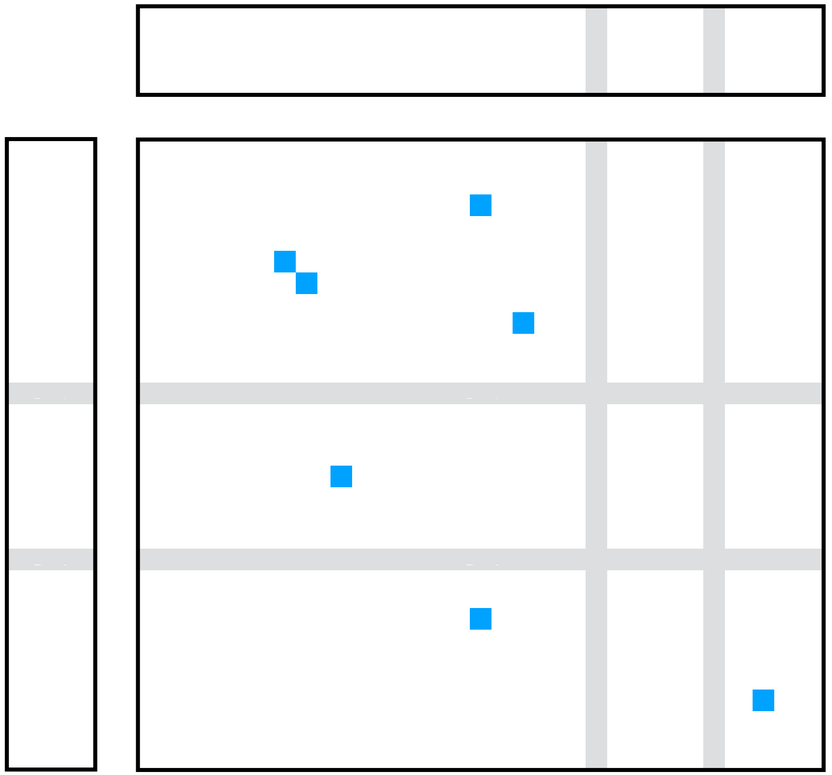
filter

- filter_rows
- filter_cols
- filter_entries



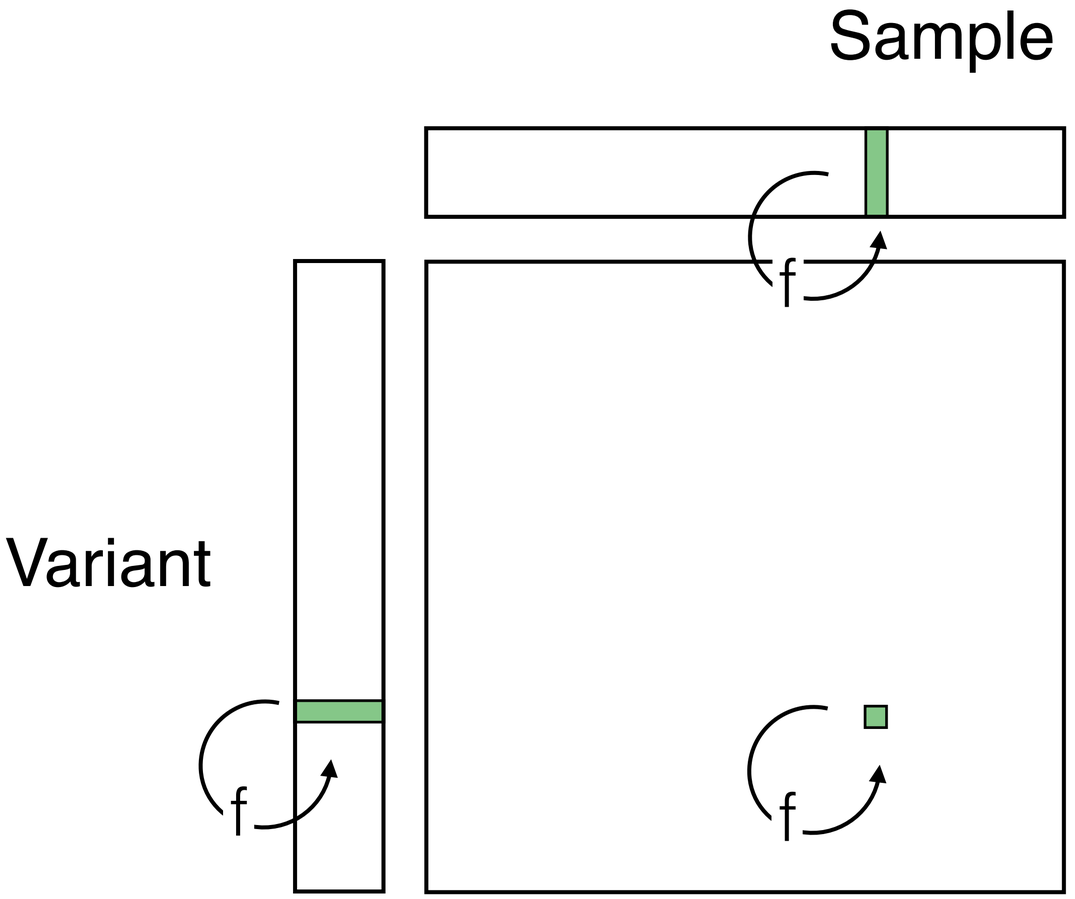
filter

- filter_rows
- filter_cols
- filter_entries



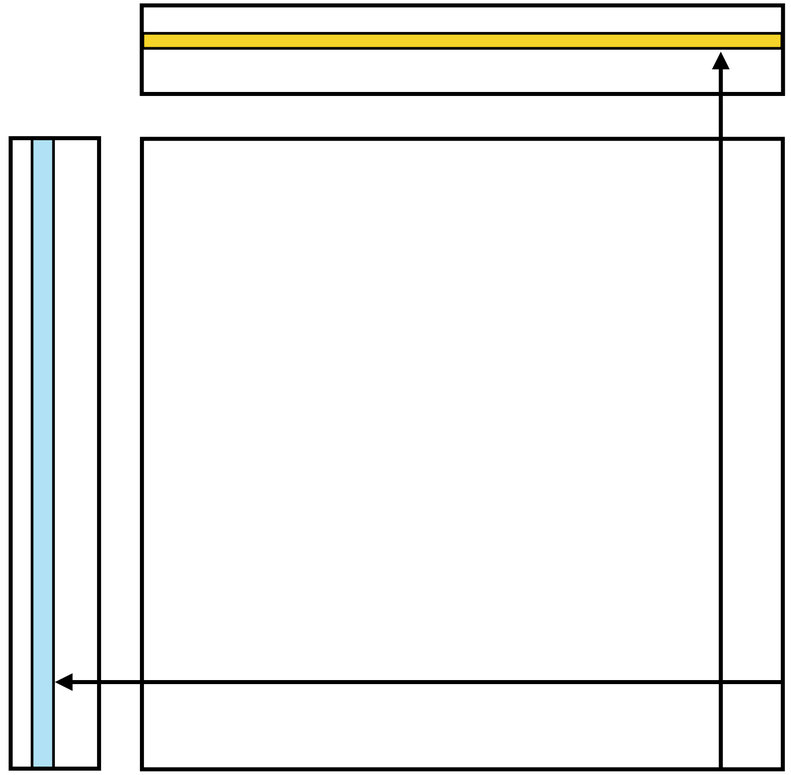
`annotate:` compute new fields

- `annotate_rows`
- `annotate_cols`
- `annotate_entries`



`annotate:` compute new fields

- `annotate_rows`
- `annotate_cols`



Mastering Hail takes practice

- Hail is harder to learn than command-line tools
  - It's not about memorizing command-line calls!
  - It's about building a foundational understanding of how to explore any kind of data
- Prior experience with a data frame library* will help
  - *`R, dplyr, pandas`, etc
- Hail is about giving you the tools you need to indulge scientific curiosity on biological data, and that's not always easy.
- Feedback is **very** welcome!

notebook.hail.is

class key: **IBG2019**

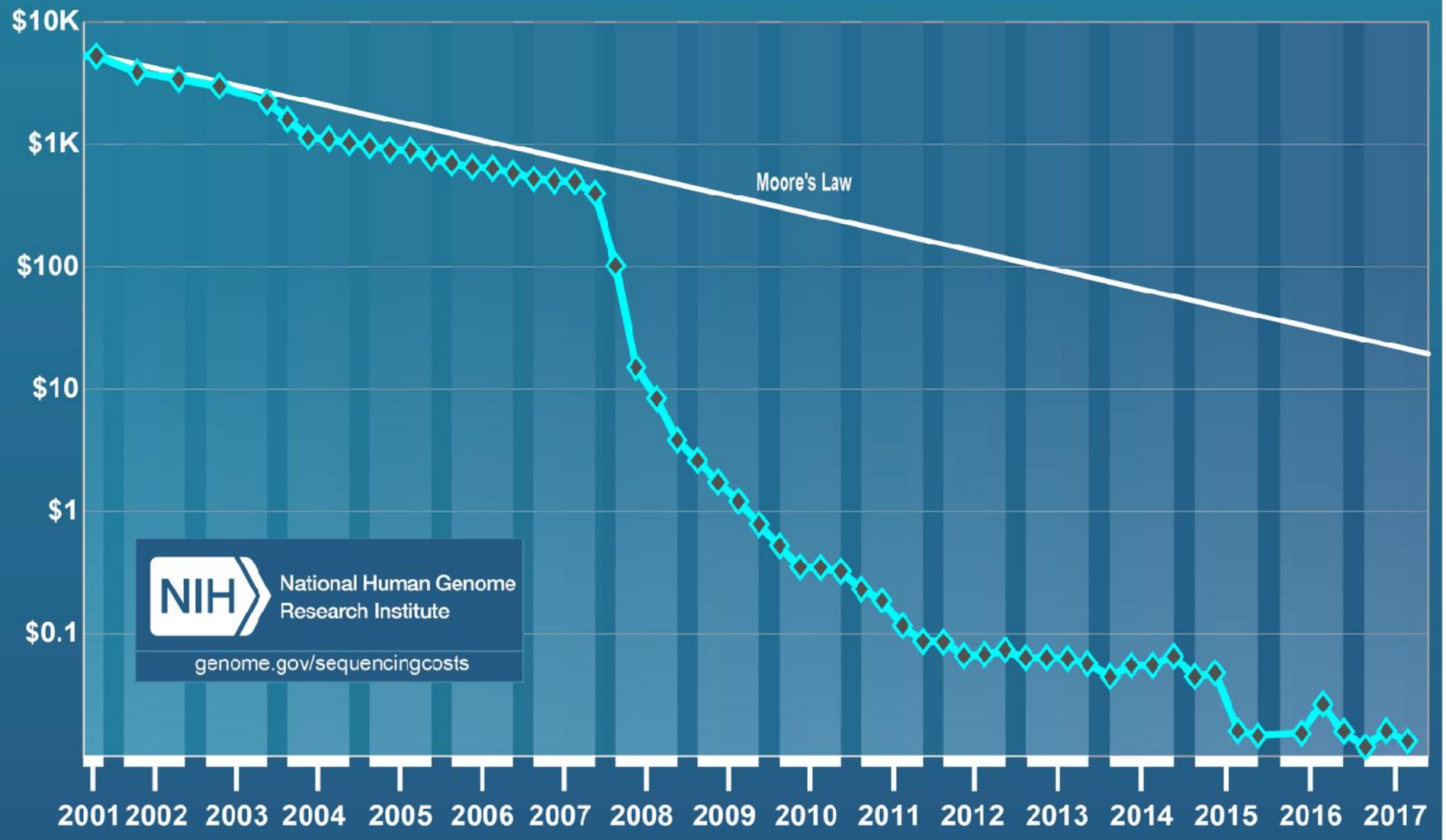
Outline

- Introduction to Hail
- Practical 1: QC
- Practical 2: GWAS
- **Computational Landscape for Bioinformatics**
- Practical 3: Inferring Ancestry
- Practical 4: Computing F_ST
- Practical 5: Gene Burden Test
- Pratical 6: De Novo Caller

Let’s run our pipeline on **all** of 1000 Genomes!

- Actually 2,504 genomes, 36M variants, 14M filtered variants
- `cloudtools` simplifies Hail cluster management on Google’s cloud
- 125x 8-core preemptible workers => 1000 cores

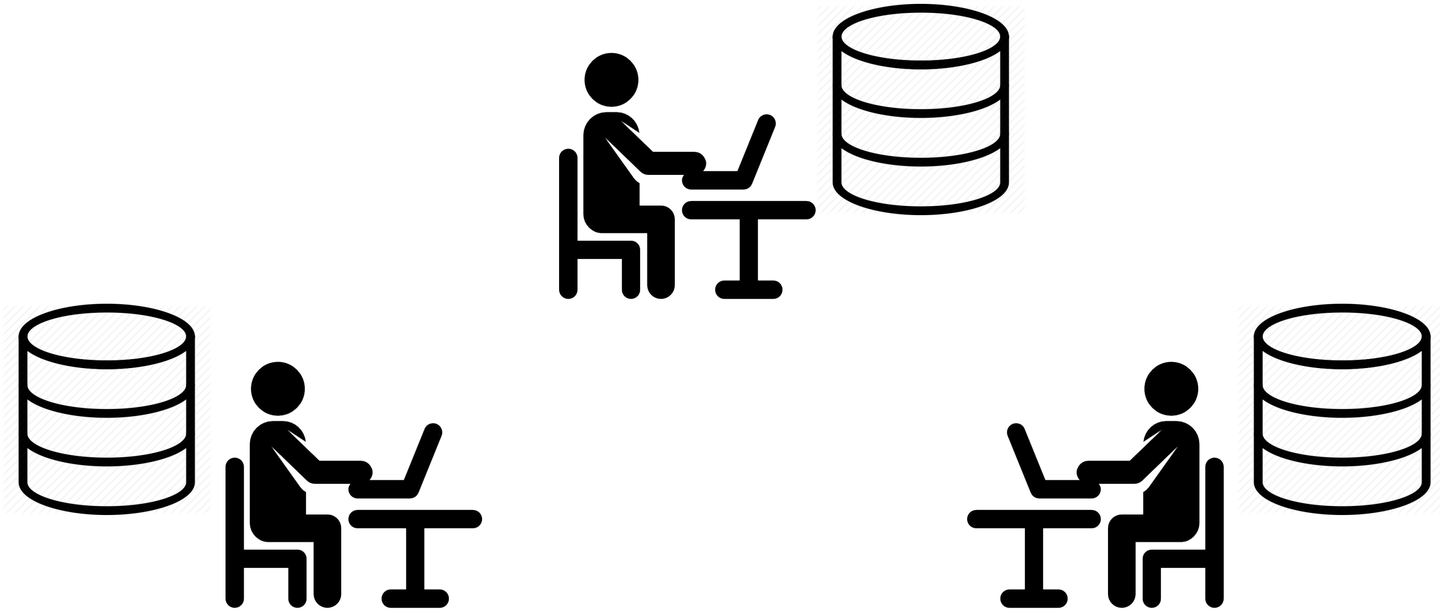
Cost per Raw Megabase of DNA Sequence



Large-scale datasets

- UKB 500K => 5M?
  - … and many other biobanks
- gnomAD: 20K => 120K WGS, 200K WES => 1M? WES
- TOPMed: >100K WGS
- All of Us 1M
- MVP 1M

From Bringing Data to Researchers



To Bringing Researchers to Data

Computational Landscape

- Laptop/Desktop
- Server
- HPC cluster
- Cloud

Computational Landscape

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

Computational Landscape

- Laptop/Desktop
  pip install hail
- Server/HPC cluster single node
  pip install hail
- HPC cluster
  On-prem Spark cluster
  Hail does not support HPC schedulers like SLURM, Grid Engine and LSF
- Cloud
  GCP: pip install cloudtools
  AWS, see:
  - https://github.com/hms-dbmi/hail-on-AWS-spot-instances
  - https://discuss.hail.is/t/spin-up-aws-emr-clusters-with-hail/818

Let's run our pipeline on **all** of 1000 Genomes!

- Actually 2,504 genomes, 36M variants, 14M filtered variants
- `cloudtools` simplifies Hail cluster management on Google's cloud
- 125x 8-core preemptible workers => 1000 cores

```
cluster start ibg -p 125

cluster connect ibg notebook

cluster stop ibg
```

hail-cseed/ | 1kg-gwas-demo

localhost:8123/notebooks/hail-cseed/1kg-gwas-demo.ipynb

jupyter 1kg-gwas-demo Last Checkpoint: 11 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Hail

Run Code Show running Spark jobs

In [2]:
```
import hail as hl
from hail.plot import show
hl.plot.output_notebook()
```

Loading BokehJS ...

In [3]:
```
%%time

annotations_path = 'gs://hail-tutorial/1kg_annotations.txt'
mt_path = 'gs://hail-1kg/1kg-all.mt'
purcell_5k = 'gs://hail-tutorial/purcell5k.loci'

ht = hl.import_table(annotations_path, impute=True).key_by('Sample')

mt = hl.read_matrix_table(mt_path)

mt = mt.annotate_cols(pheno = ht[mt.s])
mt = hl.sample_qc(mt)
mt = mt.filter_cols((mt.sample_qc.dp_stats.mean >= 4) & (mt.sample_qc.call_rate >= 0.97))
ab = mt.AD[1] / hl.sum(mt.AD)
filter_condition_ab = ((mt.GT.is_hom_ref() & (ab <= 0.1)) |
                       (mt.GT.is_het() & (ab >= 0.25) & (ab <= 0.75)) |
                       (mt.GT.is_hom_var() & (ab >= 0.9)))
mt = mt.filter_entries(filter_condition_ab)
mt = hl.variant_qc(mt)

pruned = hl.import_table(purcell_5k, no_header=True, min_partitions=20)
pruned = pruned.key_by(locus = hl.parse_locus(pruned.f0))
pruned_mt = mt.filter_rows(hl.is_defined(pruned[mt.locus]))
pruned_mt = pruned_mt.select_rows().select_cols().repartition(10)
```

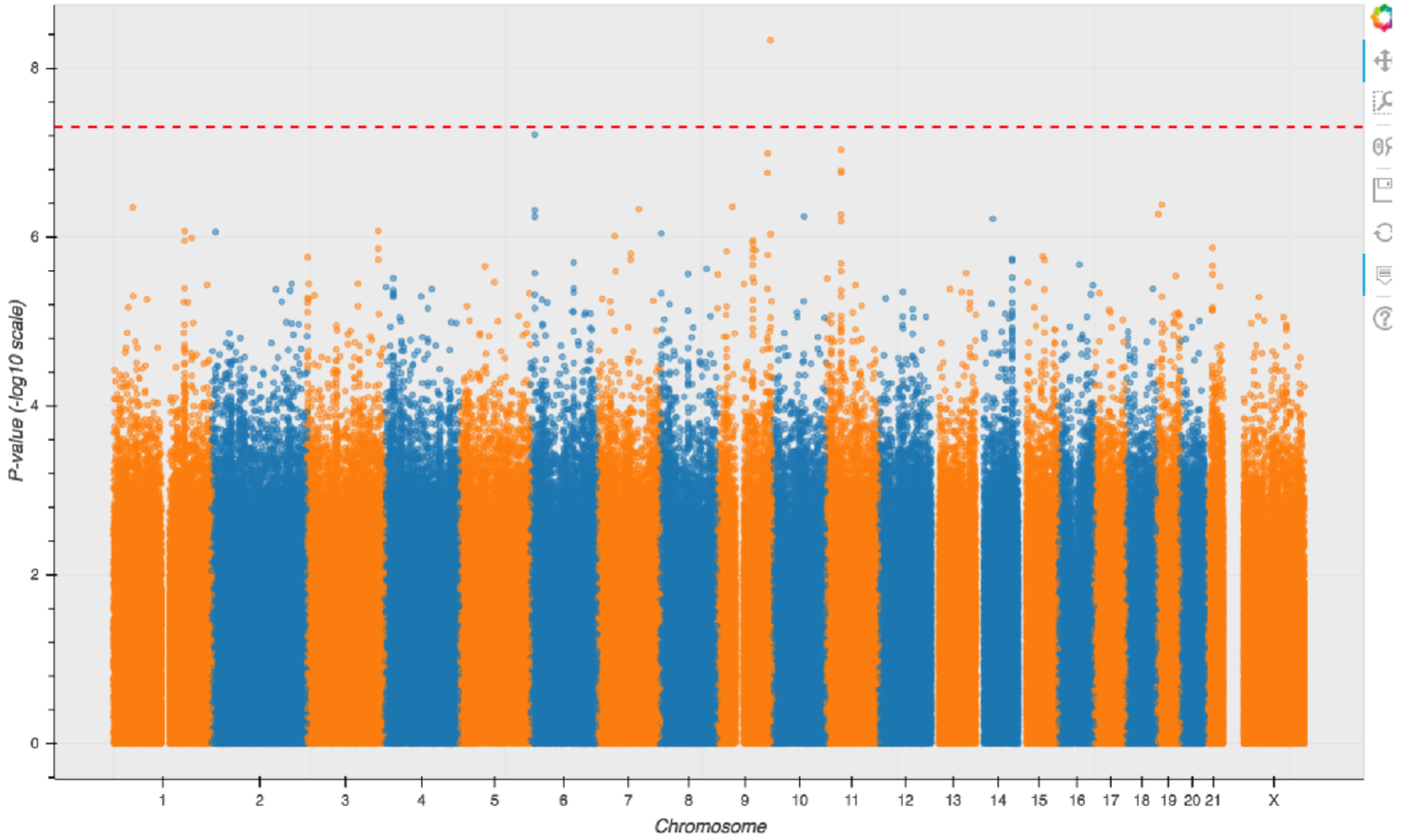
hail-cseed/ | 1kg-gwas-demo

localhost:8123/notebooks/hail-cseed/1kg-gwas-demo.ipynb

jupyter 1kg-gwas-demo Last Checkpoint: 11 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Hail

Run | Code | Show running Spark jobs




```
CPU times: user 8.18 s, sys: 244 ms, total: 8.43 s
Wall time: 8min 44s
```

Your next steps

```
pip install hail
pip install cloudtools
```

Docs, tutorials, chat, forum, code	**hail.is**
Hail cloudtools for Google cloud	**github.com/Nealelab/cloudtools**
Genome aggregation database	**gnomad.broadinstitute.org**
Medical & Pop. Genetics primers	**broadinstitute.org/mpg**
Models, Inference & Algorithms	**broadinstitute.org/mia**