Hail Query and Sequencing Quality Control

ATGU Welcome Workshop
Tuesday 12, 2023

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

BROAD INSTITUTE

STANLEY CENTER FOR PSYCHIATRIC RESEARCH AT BROAD INSTITUTE

Schedule

15:00-15:20 Introduction to the software, Hail’s data model

15:20-16:00 Hands-on guided tutorial

16:00-16:10 Break

16:10-17:00 Resume tutorial + Questions

hail Team

Dan King
Manager

Jackie Goldstein
Technical Lead

Daniel Goldstein

Iris Rademacher

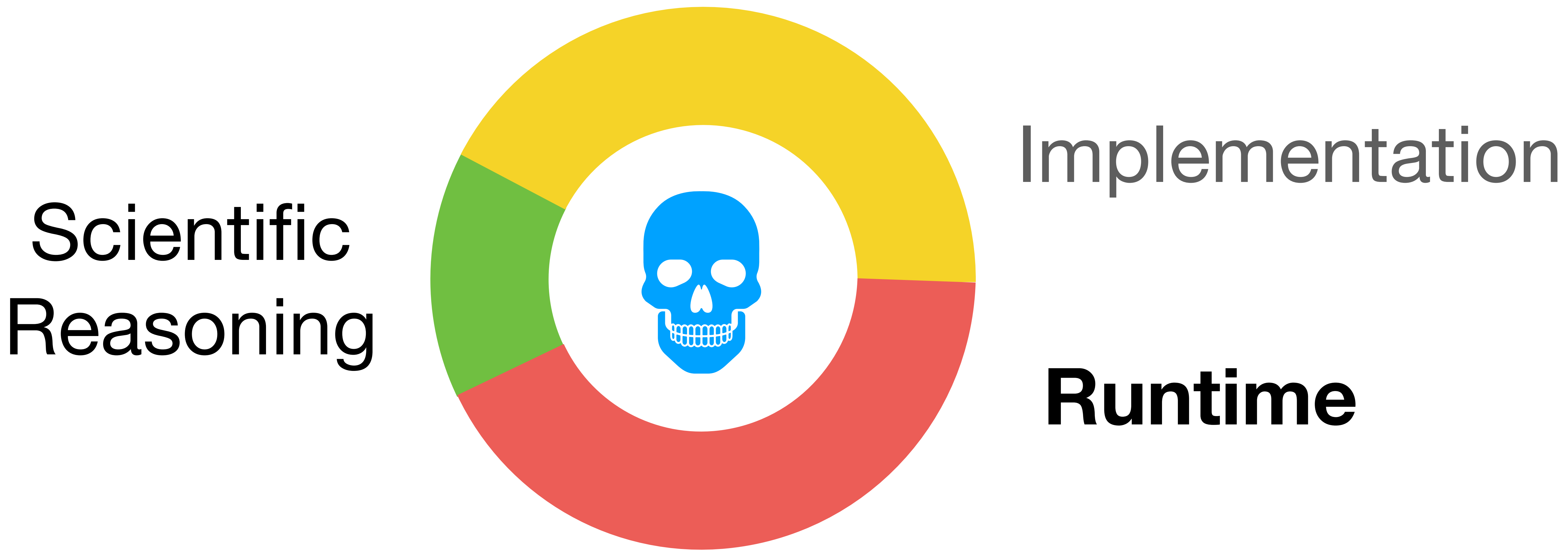
Patrick Schultz

Chris Vittal

Edmund Higham

Why hail ?

- In the last 5 years, genomic data analysis has become harder
  - ***Size*** of data =>
    time spent trying to parallelize code, waiting for results



Why hail?

`Custom Python/R scripts`

- Filter genotypes with bad allele balance
- Call *de novo* variants
- Compute transmission disequilibrium
- Dominance-encoded GWAS
- Gene count permutation tests

`PLINK`

- Detect sample duplicates or ID swaps
- Call Mendelian violations
- Relatedness
- GWAS
- …

`SNPSift`

- Genotype concordance

`bcftools`

- Split multiallelic variants
- Filter on GQ, AD, PASS

`Eigenstrat`

- PCA

`tabix`

- Subset VCFs to intervals

`vcffilterjdk`

- Filter variants

`bedtools`

- Interval annotation

Why hail ?

Custom Python/R scripts

- Filter genotypes with bad allele balance
- Call *de novo* variants
- Compute transmission disequilibrium
- Dominance-encoded GWAS
- Gene count permutation tests

Doesn’t Scale

PLINK

- Detect sample duplicates or ID swaps
- Call Mendelian violations
- Relatedness
- GWAS
- …

Doesn’t Scale

SNPSift

- Genotype concordan

Doesn’t Scale

bcftools

- Split multiallelic variants
- Filter on GQ, AD, PAS

Doesn’t Scale

Eigenstrat

- PCA

Doesn’t Scale

tabix

- Subset VCFs to intervals

Doesn’t Scale

vcffilterjdk

- Filter variants

Doesn’t Scale

bedtools

- Interval annotation

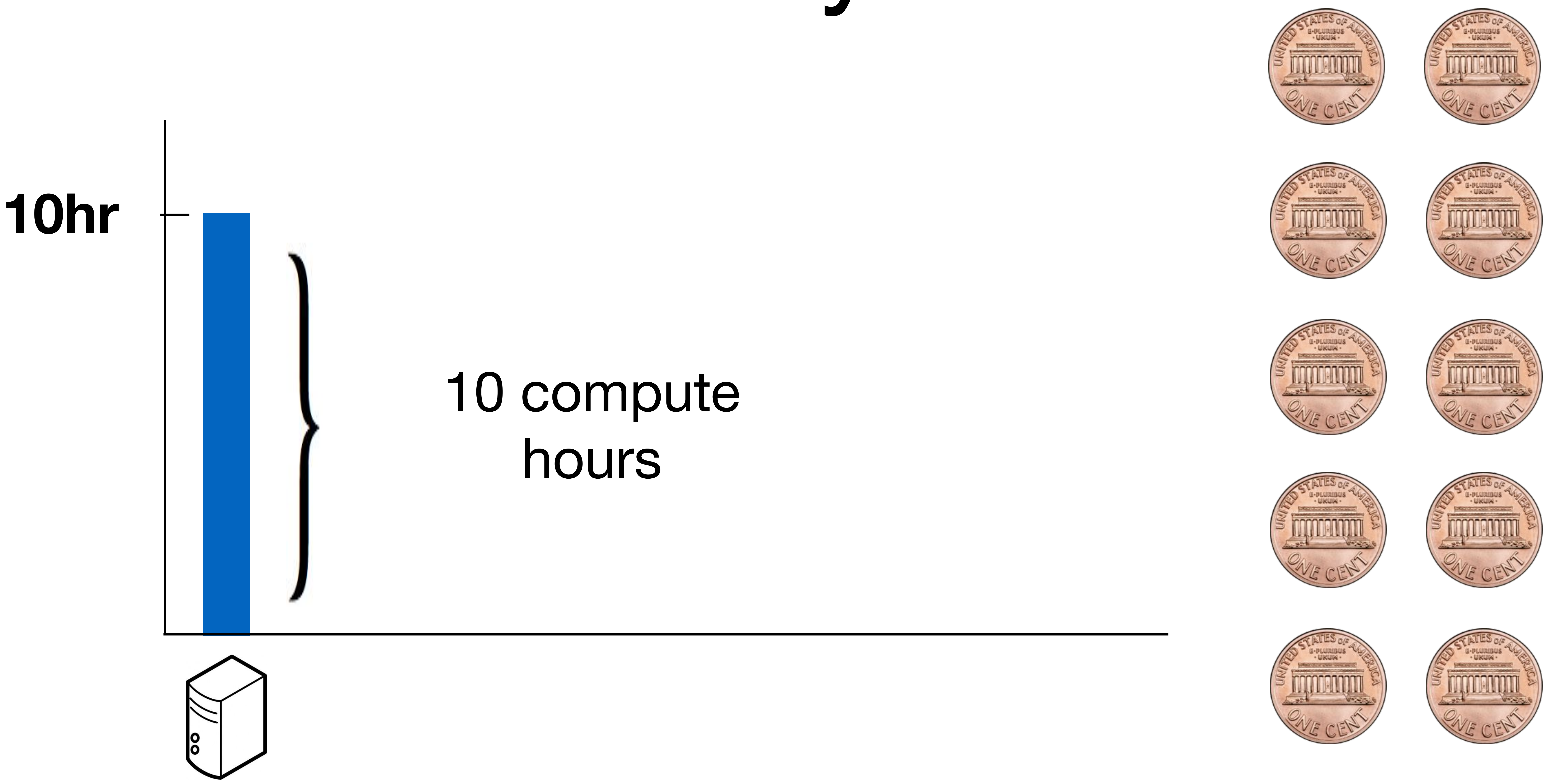
Doesn’t Scale

Hail is…

Scalable software for genomic analysis

- *scalable*: can run on a laptop, on a cluster, on the cloud

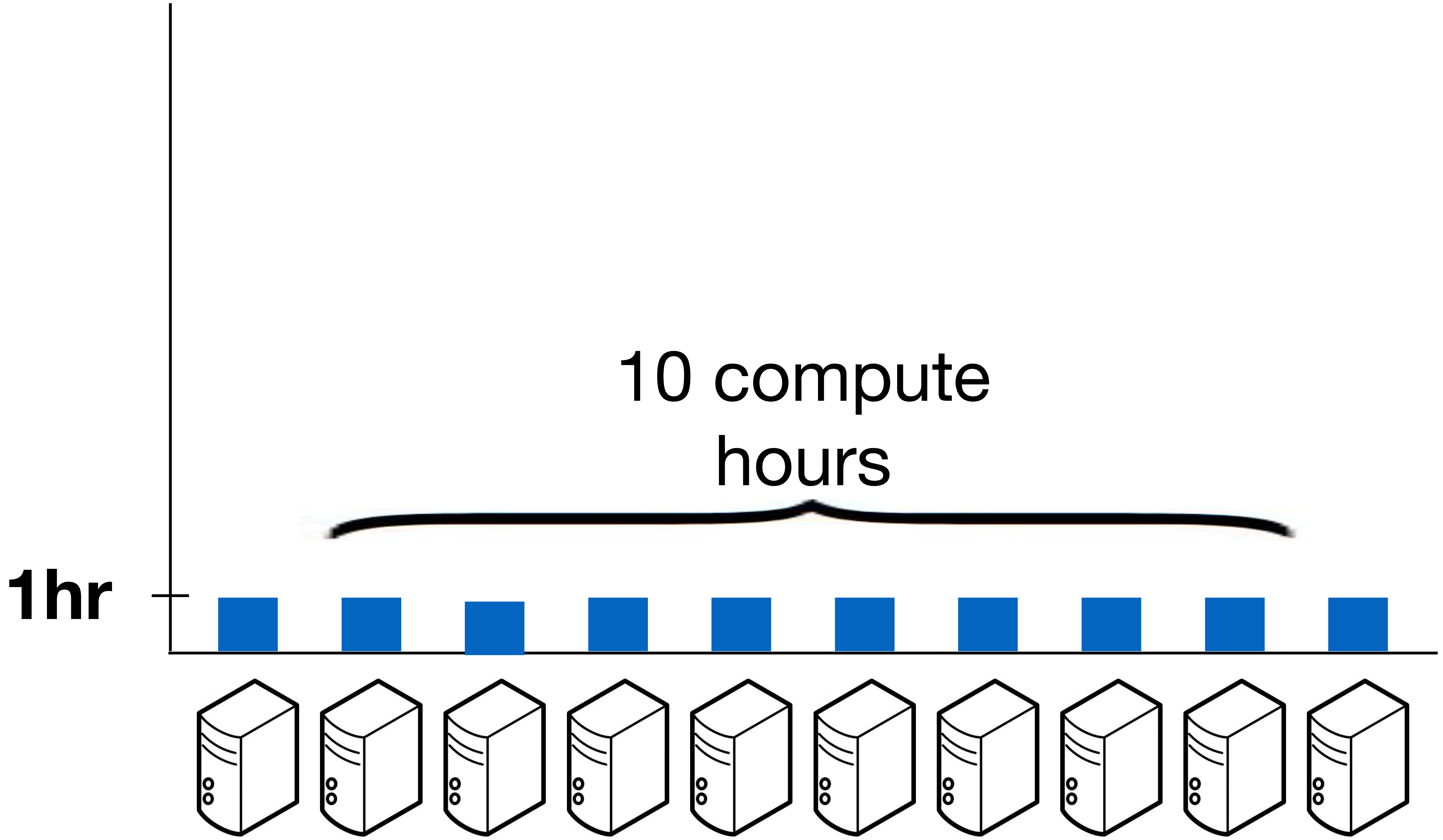
Scalability



10hr

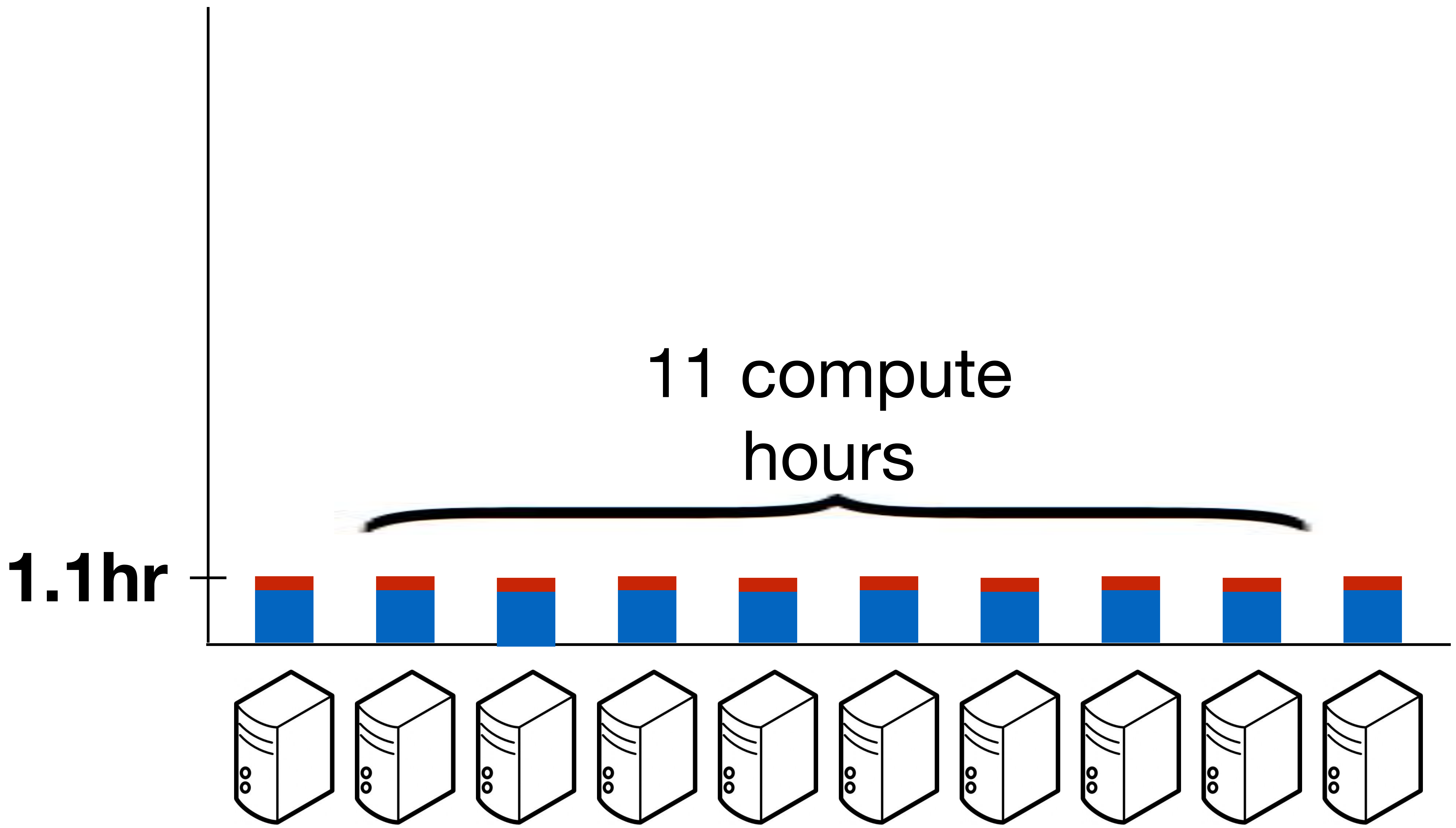
10 compute hours

Scalability



10 compute hours

Scalability

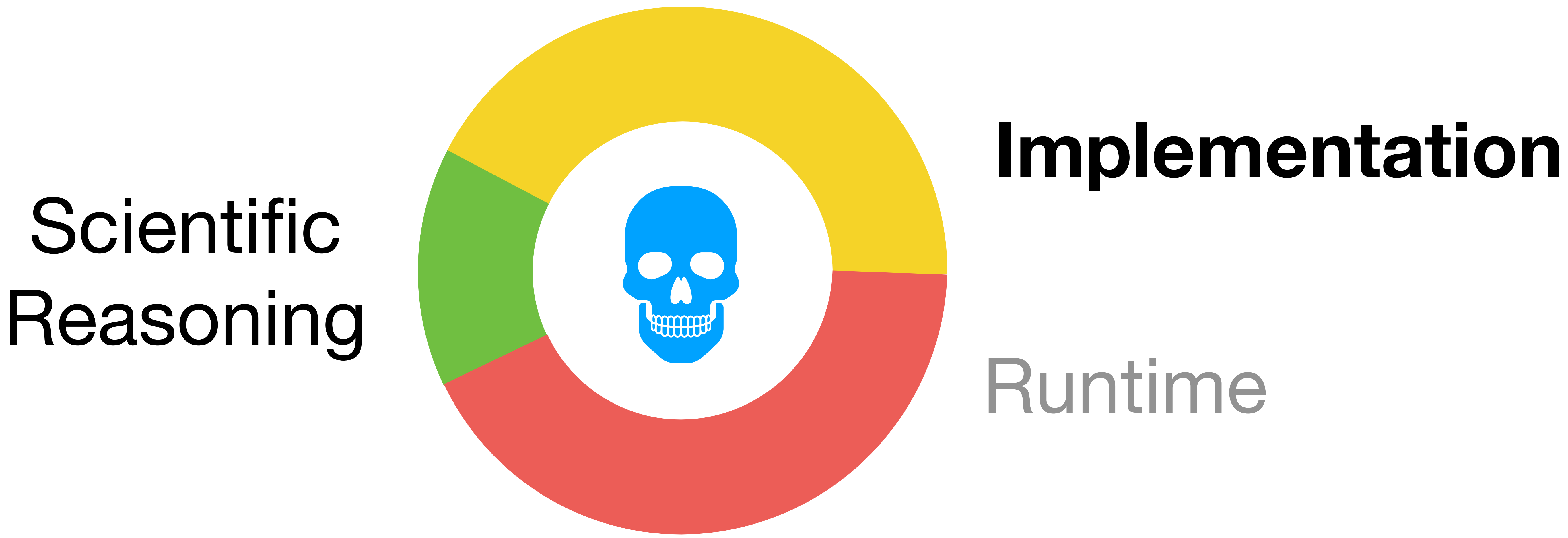


1.1hr

11 compute hours

Why hail ?

- In the last 5 years, genomic data analysis has become harder
  - ***Size*** of data =>
    time spent trying to parallelize code, waiting for results
  - ***Complexity*** of data / models =>
    time spent implementing scientific questions as efficient code



Hail is…

Scalable software for genomic analysis

- *scalable*: can run on a laptop, on a cluster, on the cloud

A **library** exposed through Python, with a Spark backend

- Not scalable PLINK! Like programming in Python or R
- Can recapitulate PLINK functionality, but flexibility and modularity are main goals

Hail as a scientific computing stack

Data slinging

Analytical toolbox

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

VCF

TSV

BGEN

PLINK

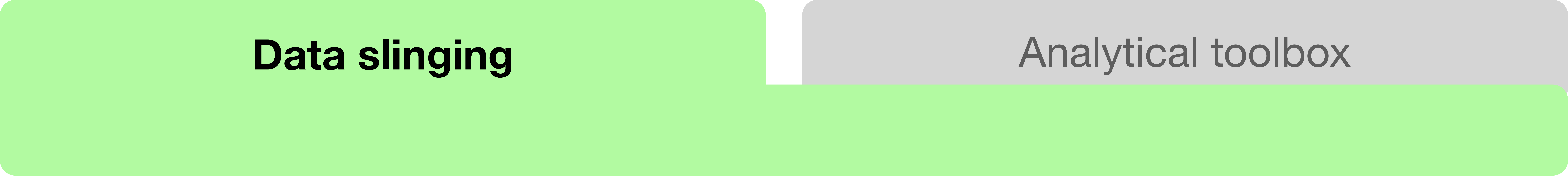
JSON

GEN

BED

GTF

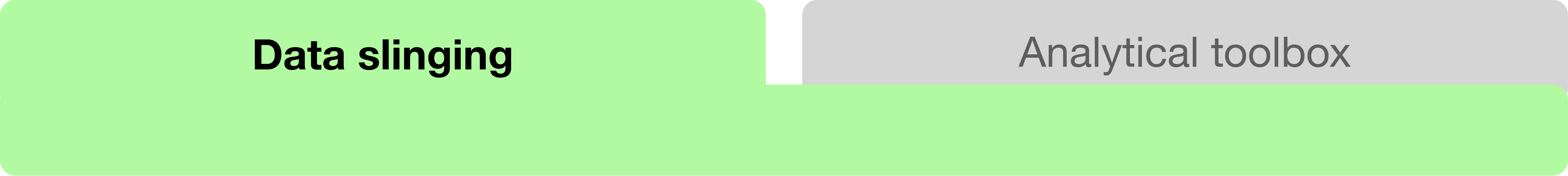
Hail as a scientific computing stack



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

- Compute AF stratified by all combinations of (sub-)population and sex
- Counting number of loss-of-function alleles per sample per gene

Hail as a scientific computing stack



Data slinging

Analytical toolbox

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

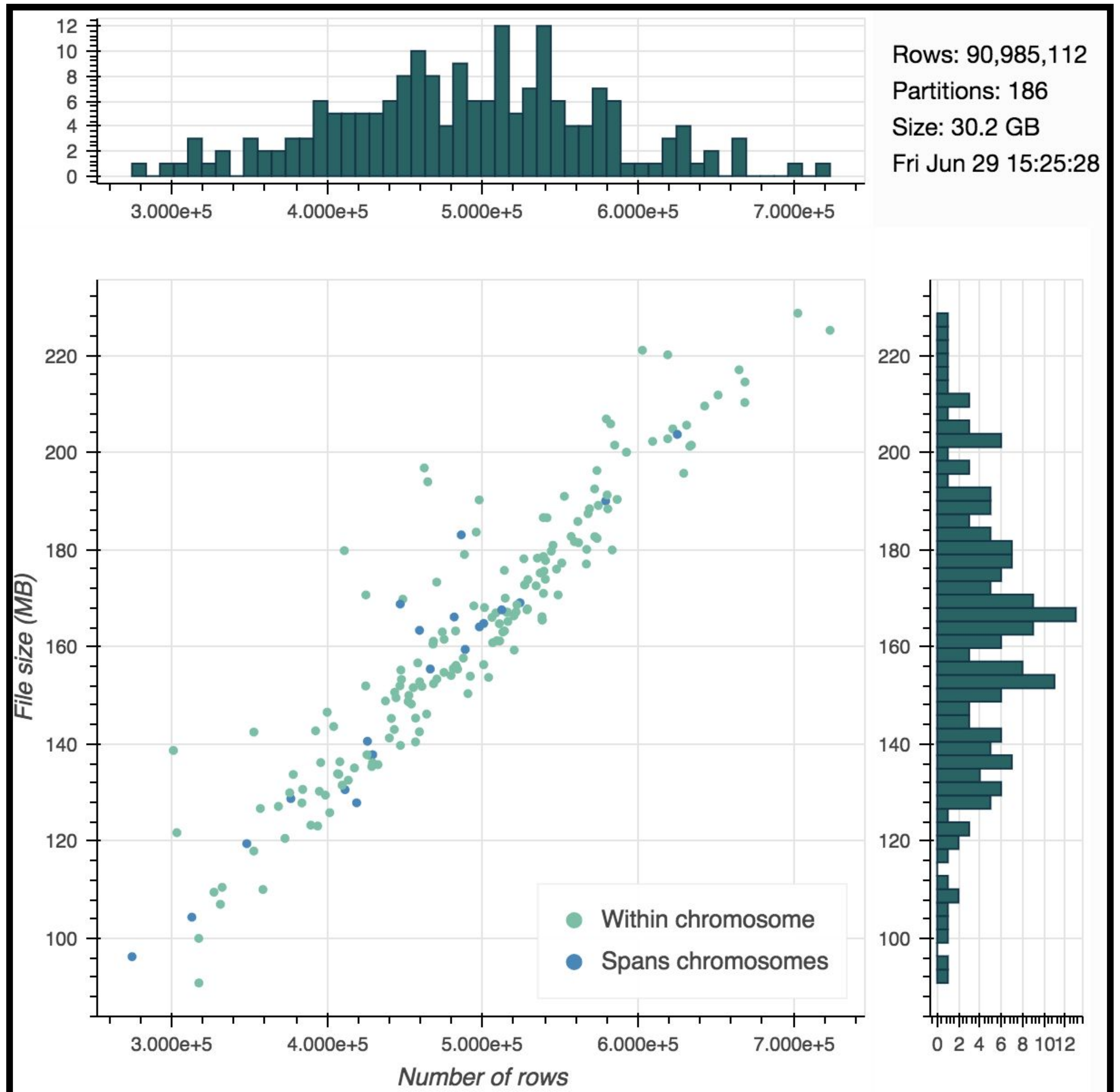
- Built-in wrappers for VEP, Nirvana
- Join with annotations by variant, locus, interval, gene
- ReferenceGenome is a first-class concept, for all our sanity

Hail as a scientific computing stack

Data slinging

Analytical toolbox

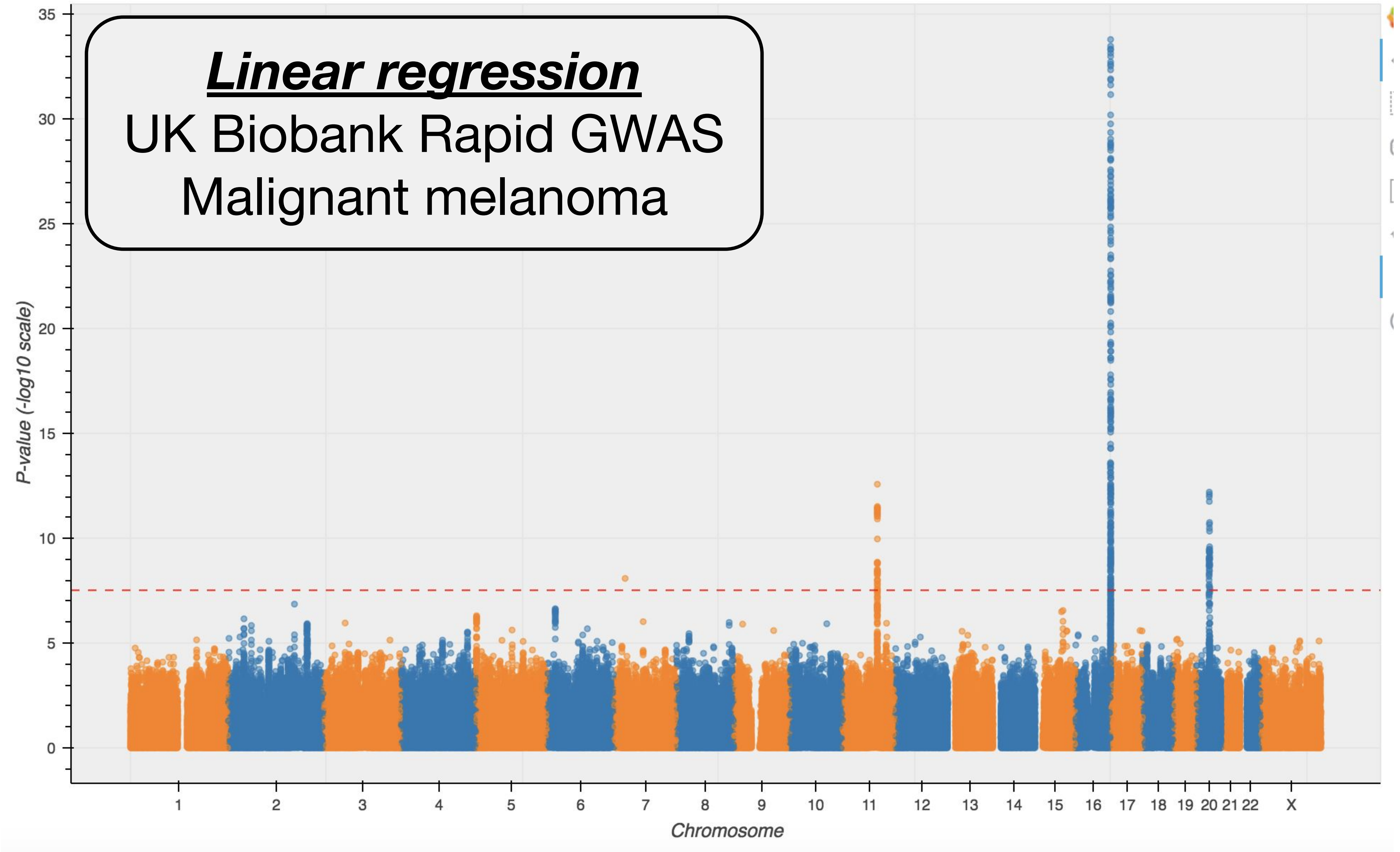
- Read and write common formats
- Filter, group, aggregate
- Annotation
- **Visualization**



Hail as a scientific computing stack

Data slinging

Analytical toolbox



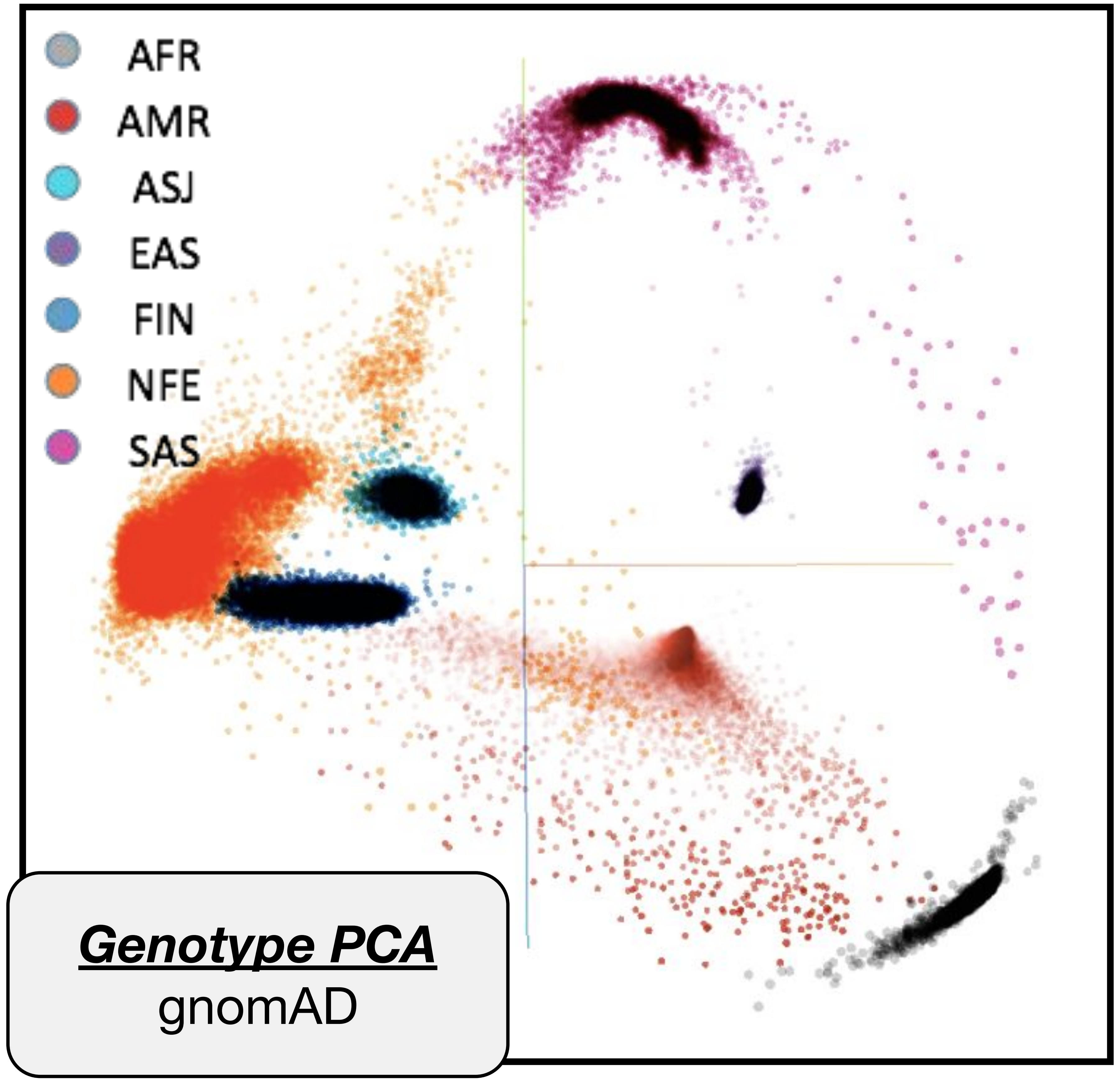
Linear regression
UK Biobank Rapid GWAS
Malignant melanoma

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

Hail as a scientific computing stack

Data slinging

Analytical toolbox



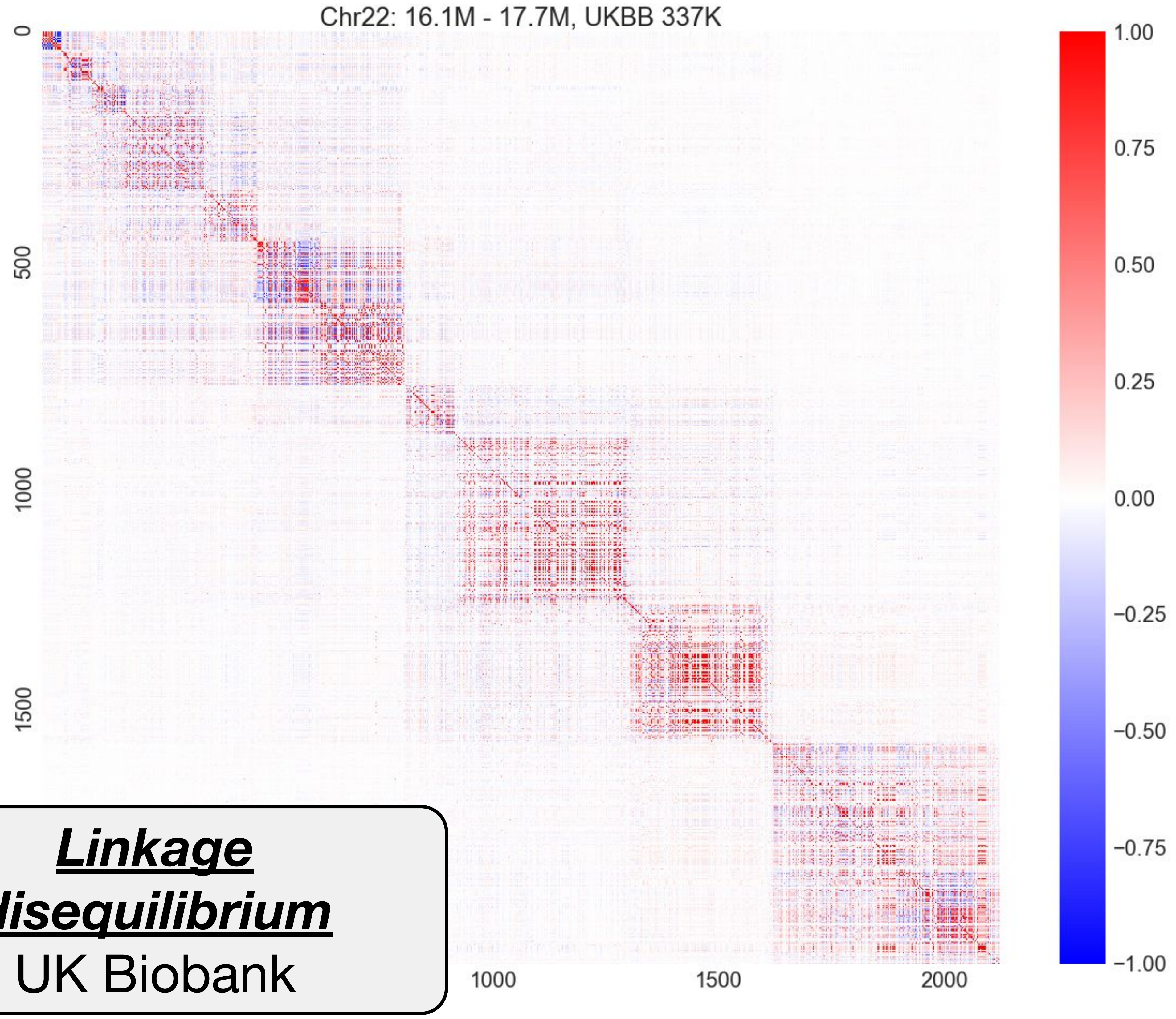
Genotype PCA
gnomAD

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

Hail as a scientific computing stack

Data slinging

Analytical toolbox



Linkage disequilibrium
UK Biobank

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

Scalable tools + elastic clouds

UK Biobank Rapid GWAS: 361K samples, 4200 phenotypes

- 200,000 CPU hours to run 115B regressions
- Research compute cluster: 6 months
- Cloud (20,000 cores): 10 hours
- Cost: $4,000

gnomAD QC: 20K genomes, 120K exomes

- 20,000 CPU hours *for one iteration*
- Research compute cluster: 17 days
- Cloud (4000 cores): 5 hours
- Cost: $400

Scalable tools + elastic clouds

UK Biobank Rapid GWAS: 361K samples, 4200 phenotypes

- 200,000 CPU hours to run 115B regressions
- Research compute cluster: **6 months**
- Cloud (20,000 cores): 10 hours
- Cost: $4,000

Research compute cluster
50 cores per user

gnomAD QC: 20K genomes, 120K exomes

- 20,000 CPU hours *for one iteration*
- Research compute cluster: **17 days**
- Cloud (4000 cores): 5 hours
- Cost: $400

Scalable tools + elastic clouds

UK Biobank Rapid GWAS: 361K samples, 4200 phenotypes

- 200,000 CPU hours to run 115B regressions
- Research compute cluster: **6 months**
- Cloud (20,000 cores): **10 hours**

🐖 **Cost: $4,000 ($1 per phenotype!)**

gnomAD QC: 20K genomes, 120K exomes

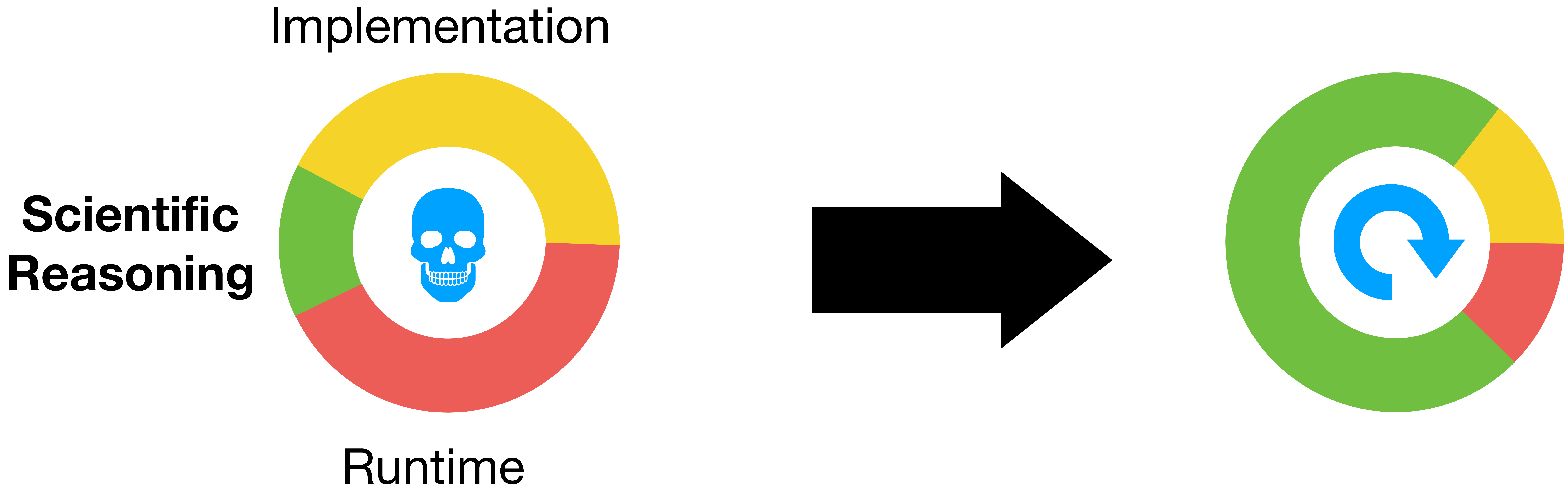
- 20,000 CPU hours *for one iteration*
- Research compute cluster: **17 days**
- Cloud (4000 cores): **5 hours**

🐖 **Cost: $400 (per iteration)**

Research compute cluster
50 cores per user

Public clouds
Pay per CPU-hour for what you want

Why hail ?



Implementation

Scientific Reasoning

Runtime

Who can Hail help?

Definitely:

- People working with sequencing call sets of any size
- People working with big genotyping data

Maybe:

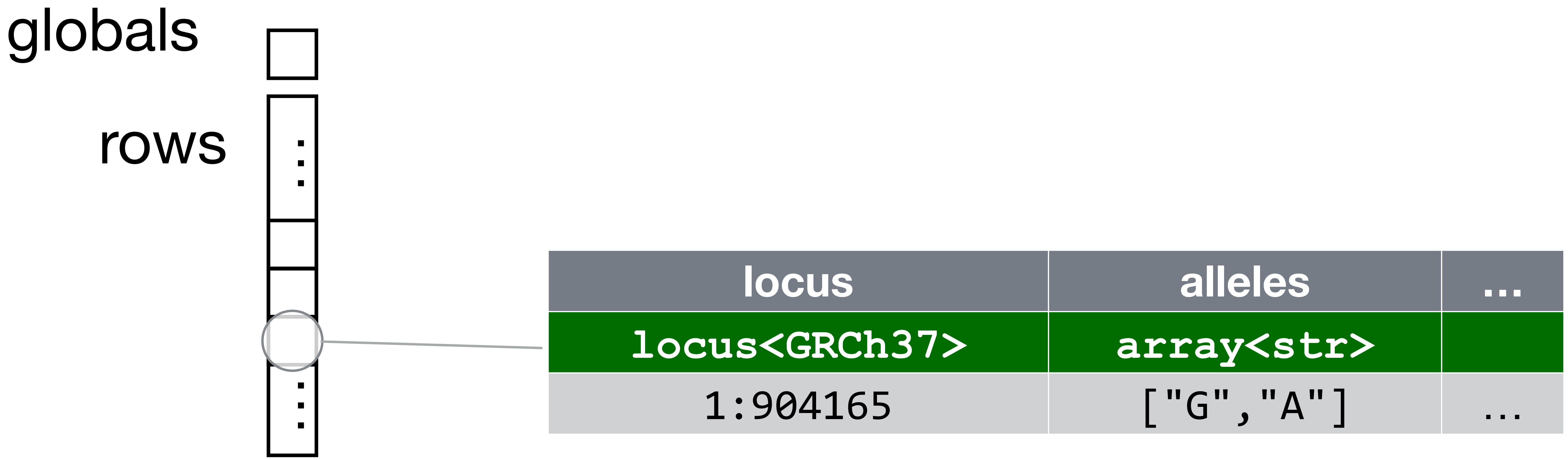
- People working with big RNA-Seq data

Not yet:

- Clinical geneticists (sequence one genome, deliver one report)
    - **But** Hail was critical for building gnomAD, Seqr, …

Review of Hail Data Structures

Table

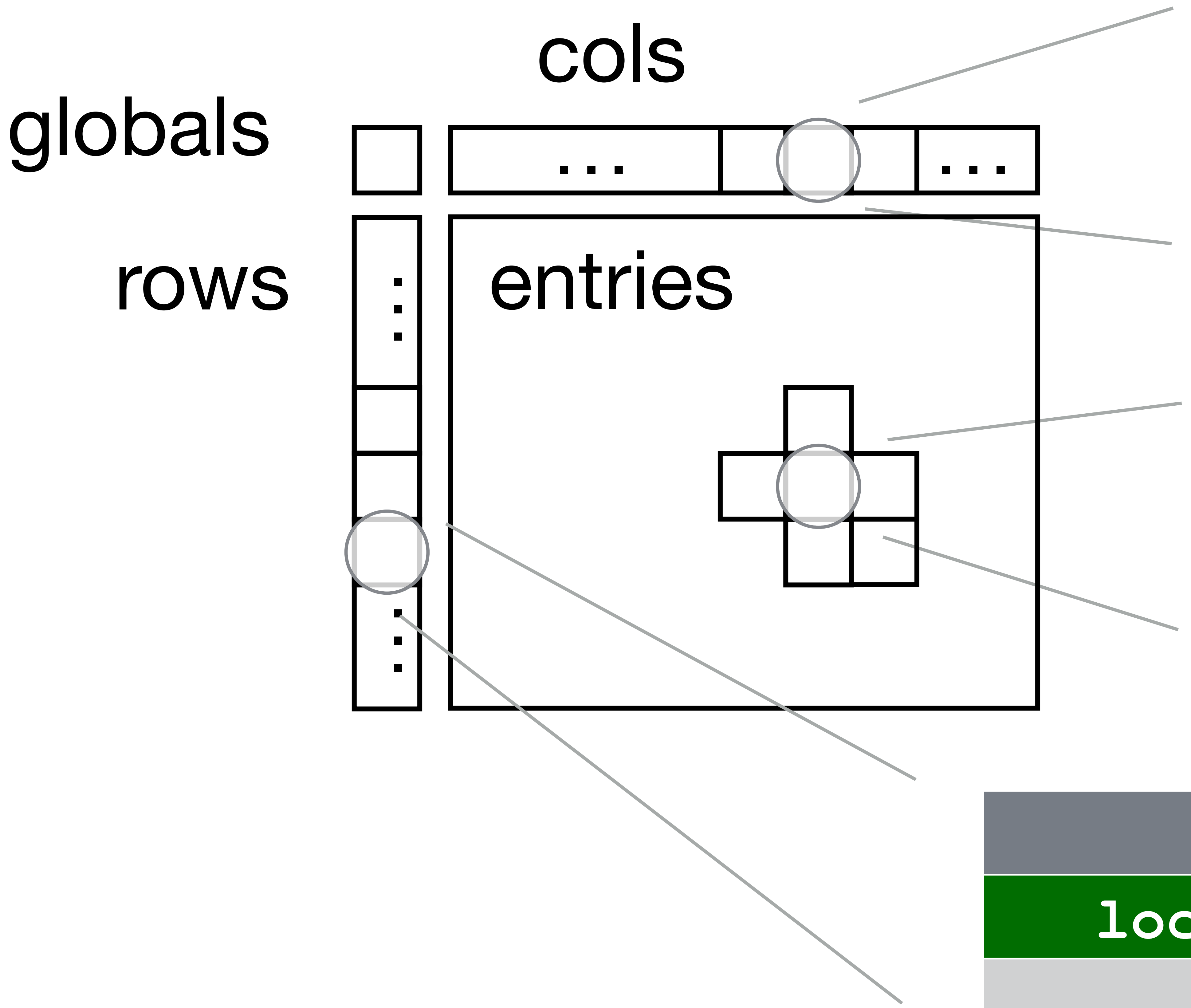


globals

rows

locus	alleles	...
locus<GRCh37>	array<str>	
1:904165	["G","A"]	...

MatrixTable



ID	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	...

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* or SQL 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!

Get the tutorial materials:

```
git clone git@github.com:hail-is/ATGU-Hail_Workshop2023.git
```

Thank you!