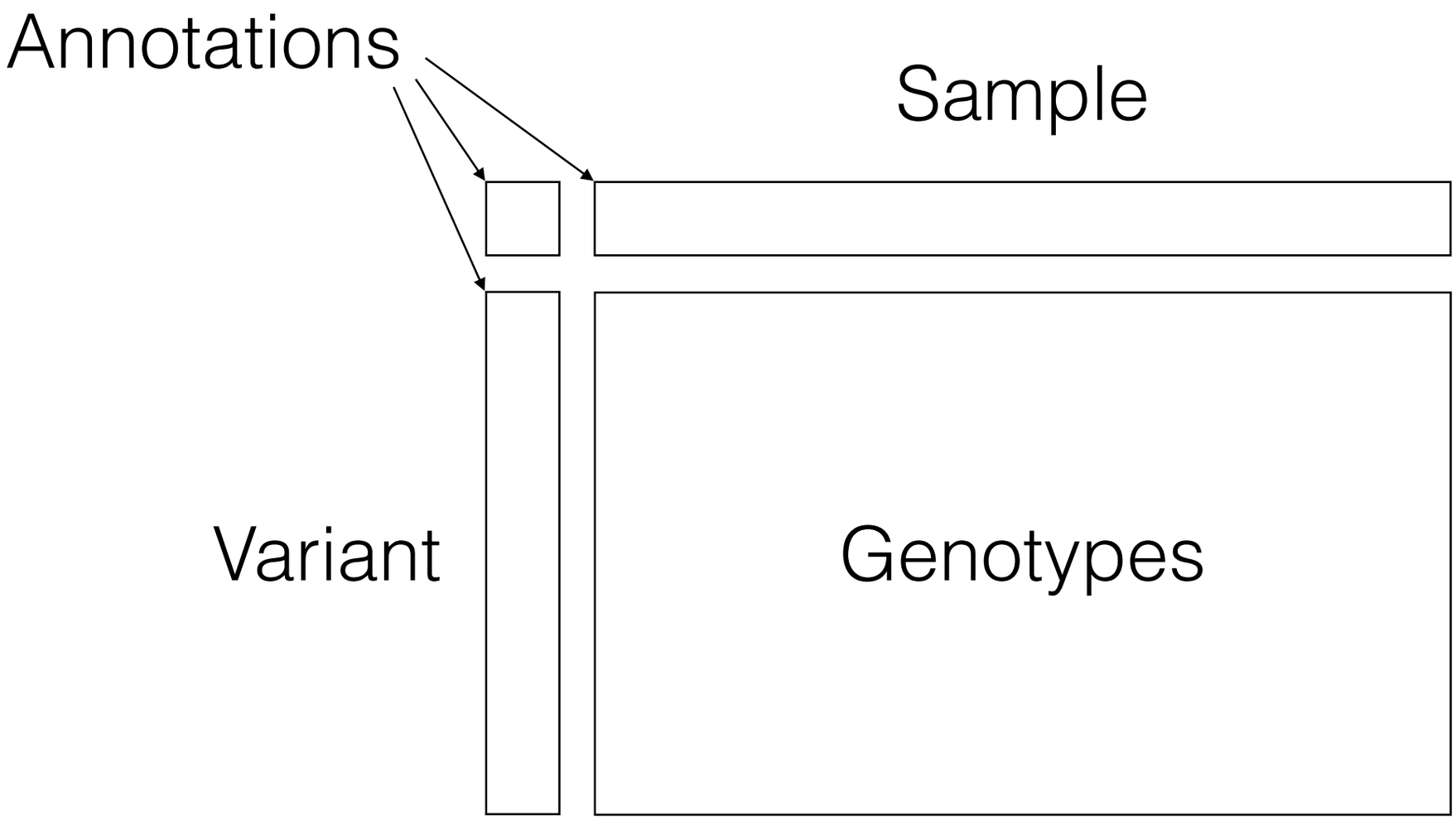
Annotations and Hail

What are annotations?

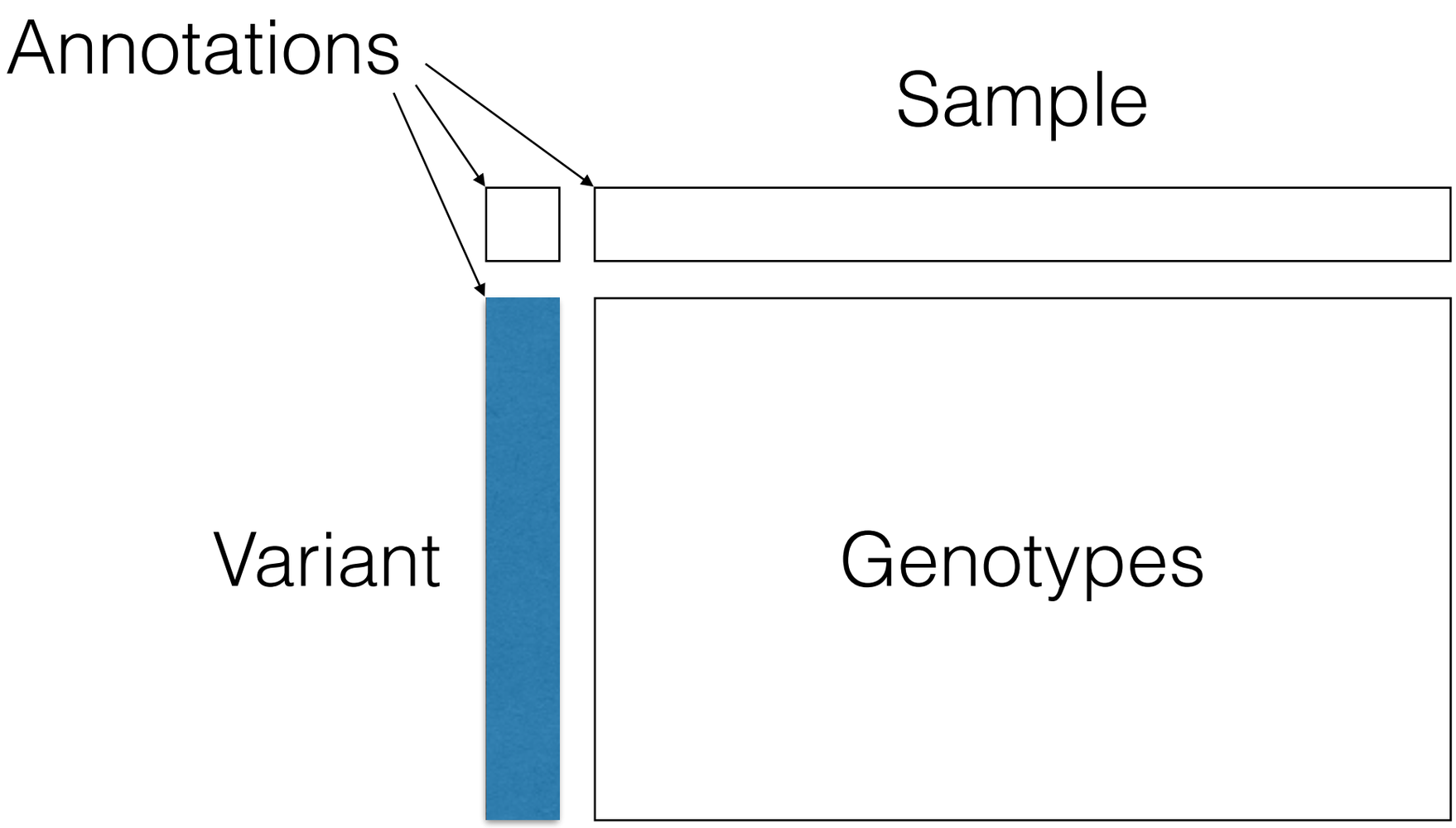
- **Variant annotations**: genome biology (functional annotation, DNAse hypersensitivity, etc)

- **Sample annotations**: phenotype and sample metadata (sex, relatives, etc)

Variant Dataset



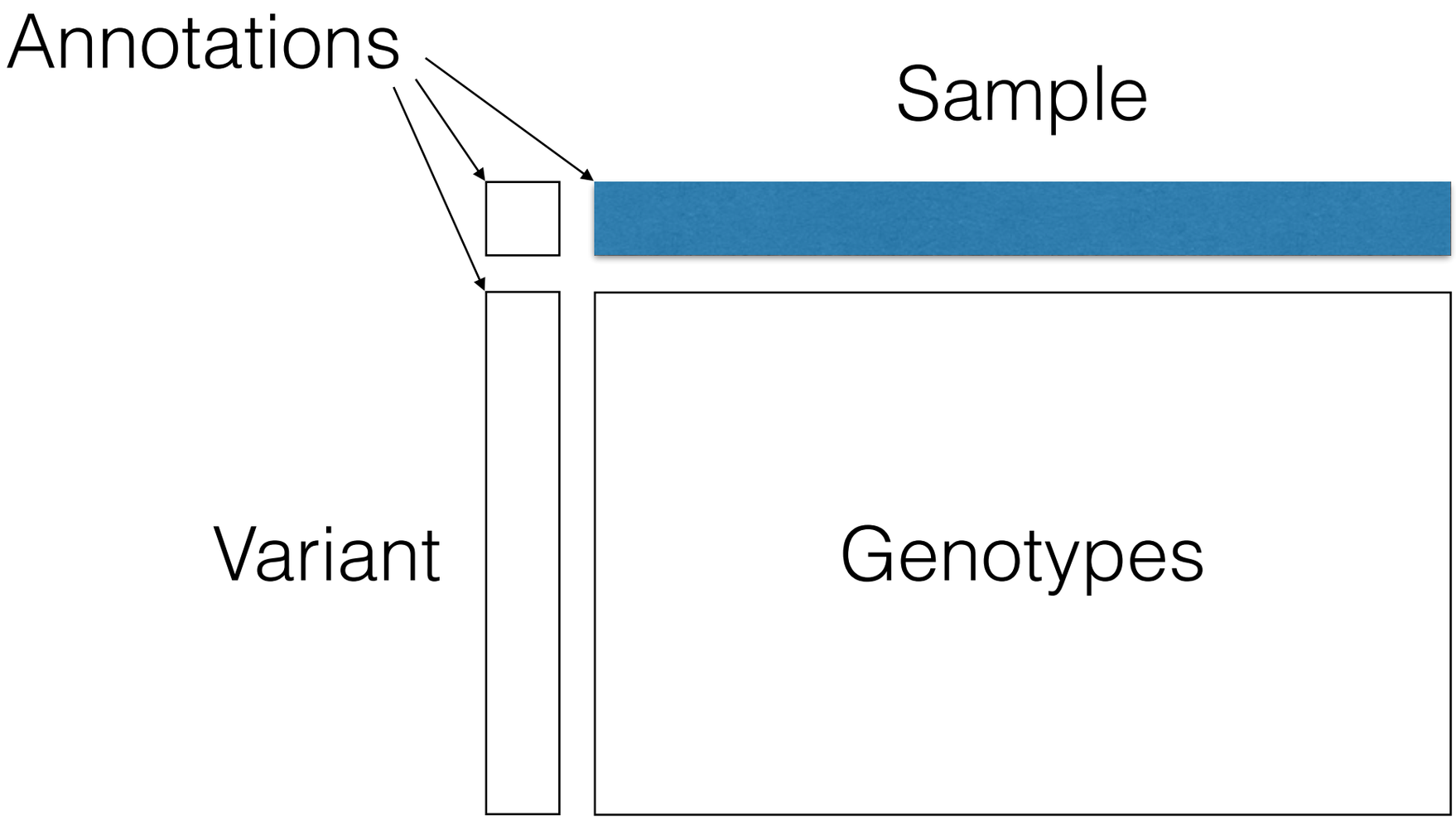
Variant Dataset



Variant annotations are genome biology

- Gene
- Allele count in reference databases like gnomAD
- rsID in dbSNP
- Functional annotations

Variant Dataset



Sample annotations are phenotype and metadata

- Reported ancestry
- Reported sex
- Binary phenotype (case / control status)
- Quantitative measures of phenotype

Annotations are much more

- **Variant annotations**: any piece of information indexed by variant

- **Sample annotations**: any piece of information indexed by sample

Variant annotations are ~~genome biology~~ **anything**

- *p*-values and betas from association
- transmission rate among trios
- QC statistics: call rate, Hardy-Weinberg equilibrium
- Concordance with orthogonal datasets

Sample annotations are ~~phenotype and metadata~~ **anything**

- Principal components from PCA
- Imputed sex (compare to reported sex!)
- QC statistics: call rate, number of singletons
- Computed polygenic risk score

Annotations are everywhere

Produce annotations

- **annotate_alleles** (1)
- **annotate_global** (4)
- **annotate_variants** (7)
- **annotate_samples** (5)
- **concordance** (concordance matrix)
- **impute_sex** (inbreeding coeff., etc)
- **linreg** (association stats)
- **logreg** (association stats)
- **lmmreg** (association stats)
- **pca** (sample PCs, variant loadings, eigenvalues)
- **sample_qc** (mixed bag of useful stats)
- **variant_qc** (mixed bag of useful stats)
- **tdt** (transmission count and t-statistic)
- **vep** (just you wait)
- **split_multi** (original index, and whether a variant was split)

Consume annotations

- **linreg** (covariates)
- **logreg** (covariates)
- **lmmreg** (covariates)
- **export_vcf** (add to INFO field)
- **export_plink** (.fam file columns)
- **impute_sex** (population frequency prior)
- **Anything expr!**
  - **Filter**
  - **Annotate**
  - **Query**

Outline of Hail Practicals

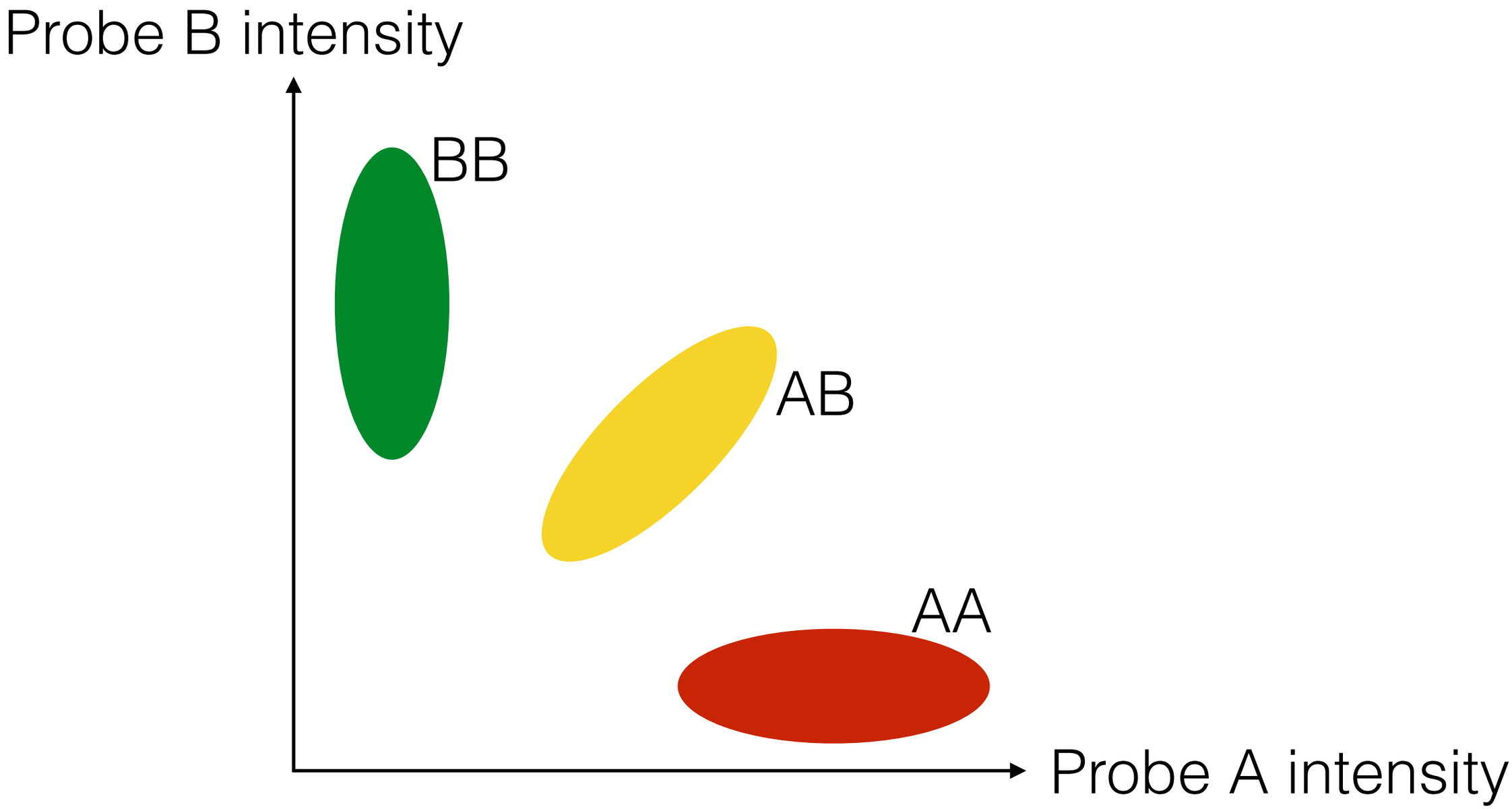
1. Importing, schemas, simulated data
2. The Hail expression language
3. **Annotation, query and plotting**
4. Aggregables: working with massive data
5. Understanding GQ and DP in sequence data
6. Unmasking ancestry
7. Basic association analysis

What did we learn?

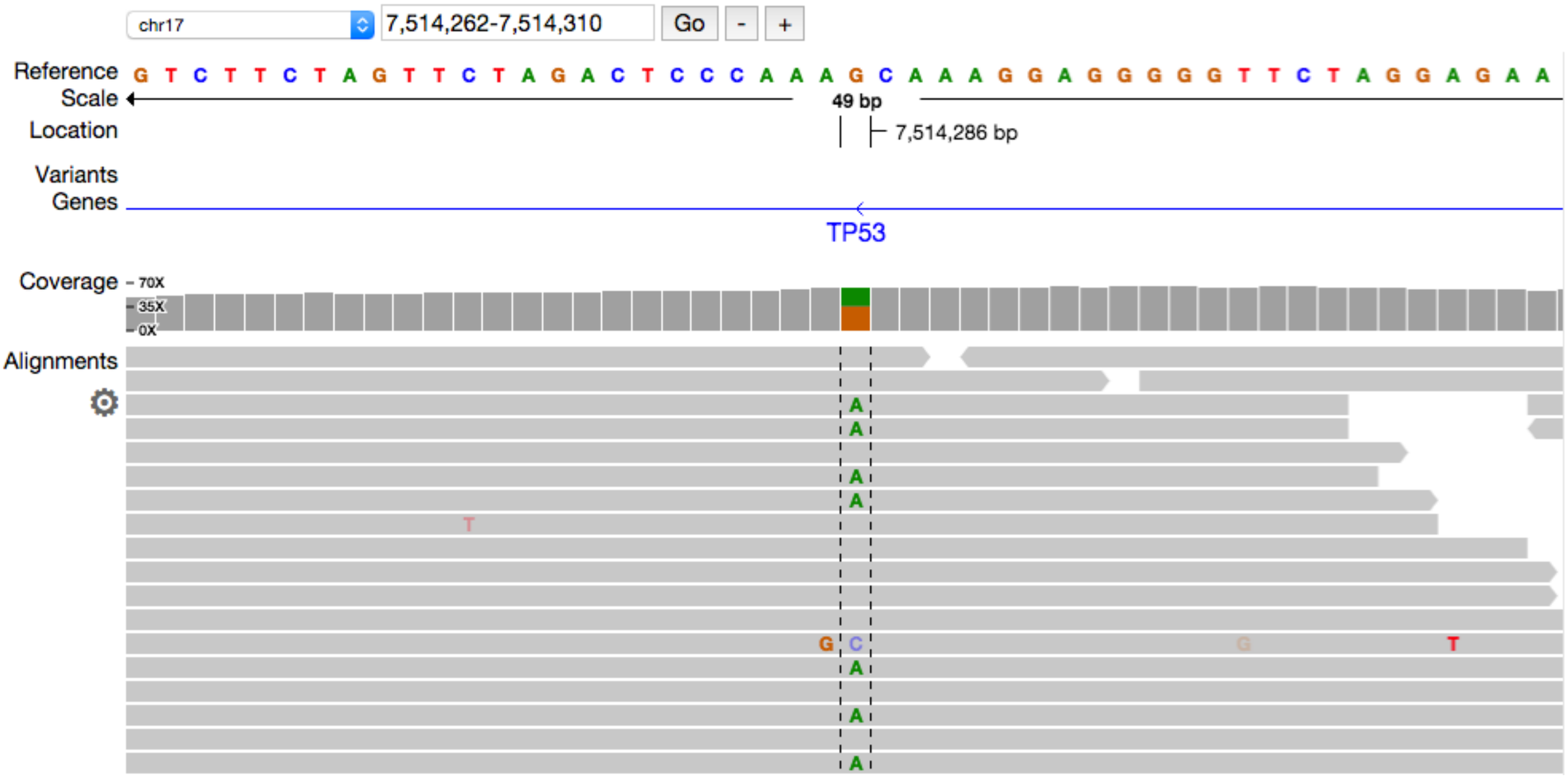
- Annotations are a way to organize analyses
- Everything is together: inputs, results, summary statistics
- Query methods provide an interface to understand properties of your data

Next-Gen Sequencing Data Science

Diploid genotyping



Shotgun sequencing



NGS means metadata

```
Genotype(GT=1,
         AD=[4, 3],
         DP=7,
         GQ=85,
         PL=[85, 0, 109])
```

NGS means metadata

```
Genotype(GT=1,
         AD=[4, 3],
         DP=7,
         GQ=85,
         PL=[85, 0, 109])
```

GT: Genotype call

- For biallelics, is the same as PLINK
- For multiallelics, is more complicated…
- Best-guess genotype call

NGS means metadata

```
Genotype(GT=1,
AD=[4, 3],
DP=7,
GQ=85,
PL=[85, 0, 109])
```

AD: Allele Depth

- List with one element per allele (2 for biallelic)
- Number of informative reads for each allele

NGS means metadata

```
Genotype(GT=1,
         AD=[4, 3],
         DP=7,
         GQ=85,
         PL=[85, 0, 109])
```

DP: Depth

- Total depth
- Usually is the sum of AD, but sometimes is larger

NGS means metadata

```
Genotype(GT=1,
AD=[4, 3],
DP=7,
GQ=85,
PL=[85, 0, 109])
```

GQ: Genotype Quality

- 0 to 99
- Phred scaled: -10 log10
- GQ 10 = 90% confidence
- GQ 20 = 99% confidence
- GQ 99 = very confident
- Often not quite calibrated

NGS means metadata

```
Genotype(GT=1,
         AD=[4, 3],
         DP=7,
         GQ=85,
         PL=[85, 0, 109])
```

PL: Phred-scaled Likelihoods

- One per possible genotype*
- Phred scaled: -10 log10
- GQ 10 = 90% confidence
- GQ 20 = 99% confidence
- GQ 99 = very confident (often not well calibrated)

*3 for biallelic, 6 for triallelic

NGS means metadata

```
Genotype(GT=1,
         AD=[4, 3],
         DP=7,
         GQ=85,
         PL=[85, 0, 109])
```

x 10 trillion (gnomAD)

Science starts with familiarity

- Making good decisions in QC and analysis requires more than push-button solutions
- The most effective sequence data analysts are comfortable with:
  - dozens of summary statistics
  - their distributions
  - their correlations!

Outline of Hail Practicals

1. Importing, schemas, simulated data
2. The Hail expression language
3. Annotation, query and plotting
4. Aggregables: working with massive data
5. **Understanding GQ and DP in sequence data**
6. Unmasking ancestry
7. Basic association analysis

What did we learn?

- Sequence data has strange properties that can be explained with data science

- Same process can be applied to QC and analysis!