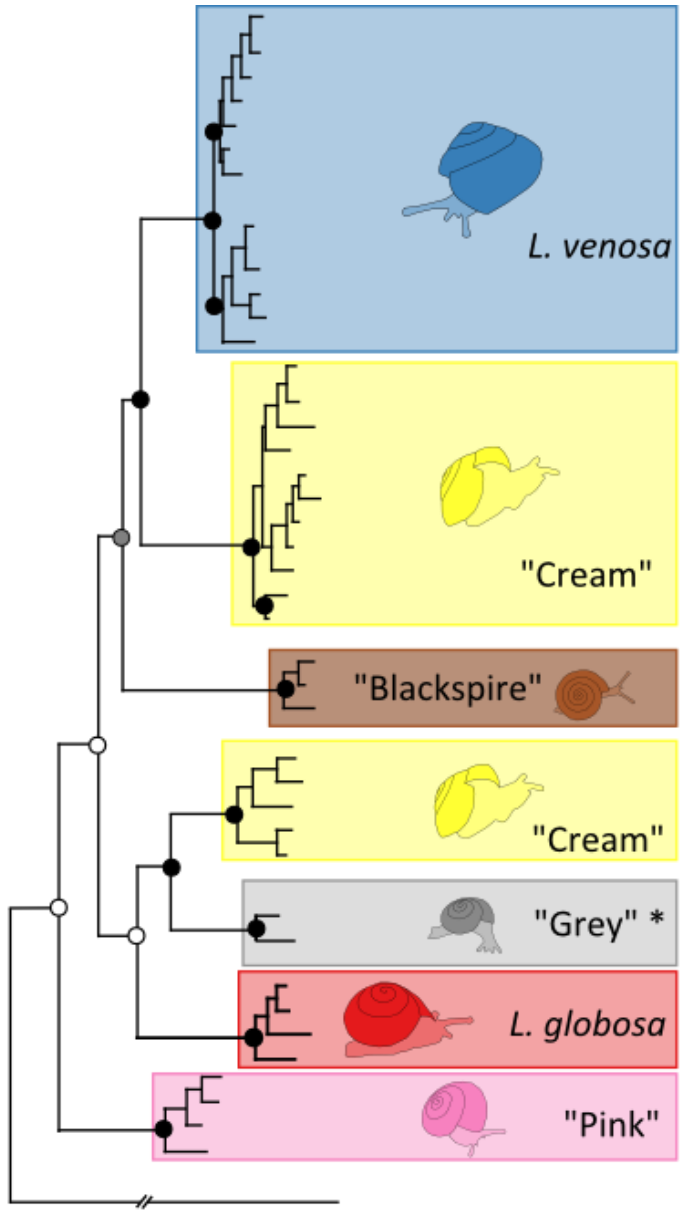
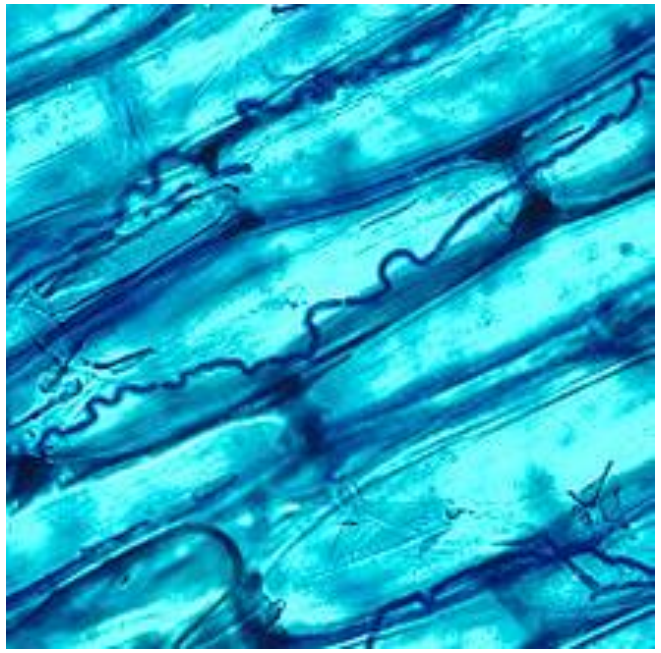
From raw reads to good SNVs

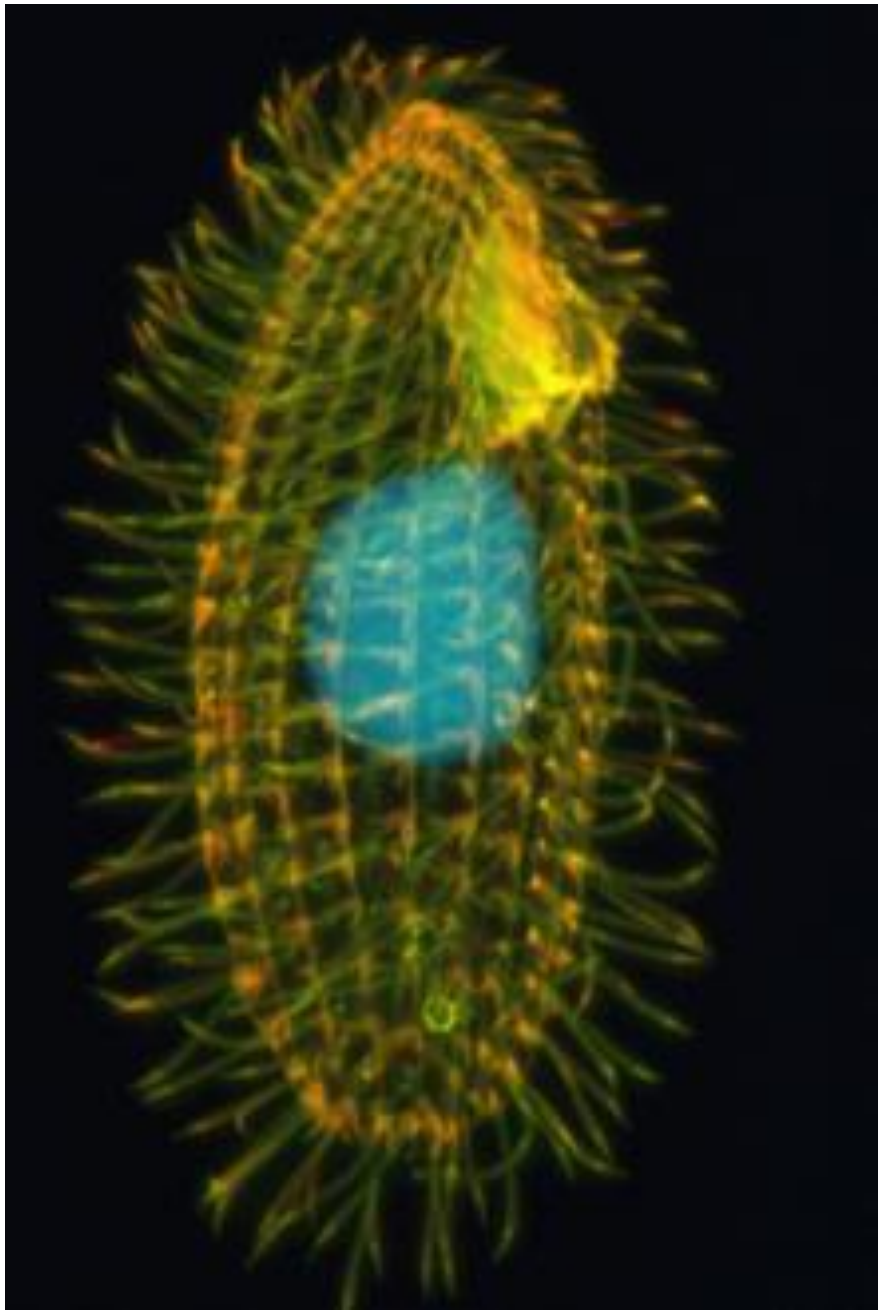
David Winter, Massey University
github.com/dwinter

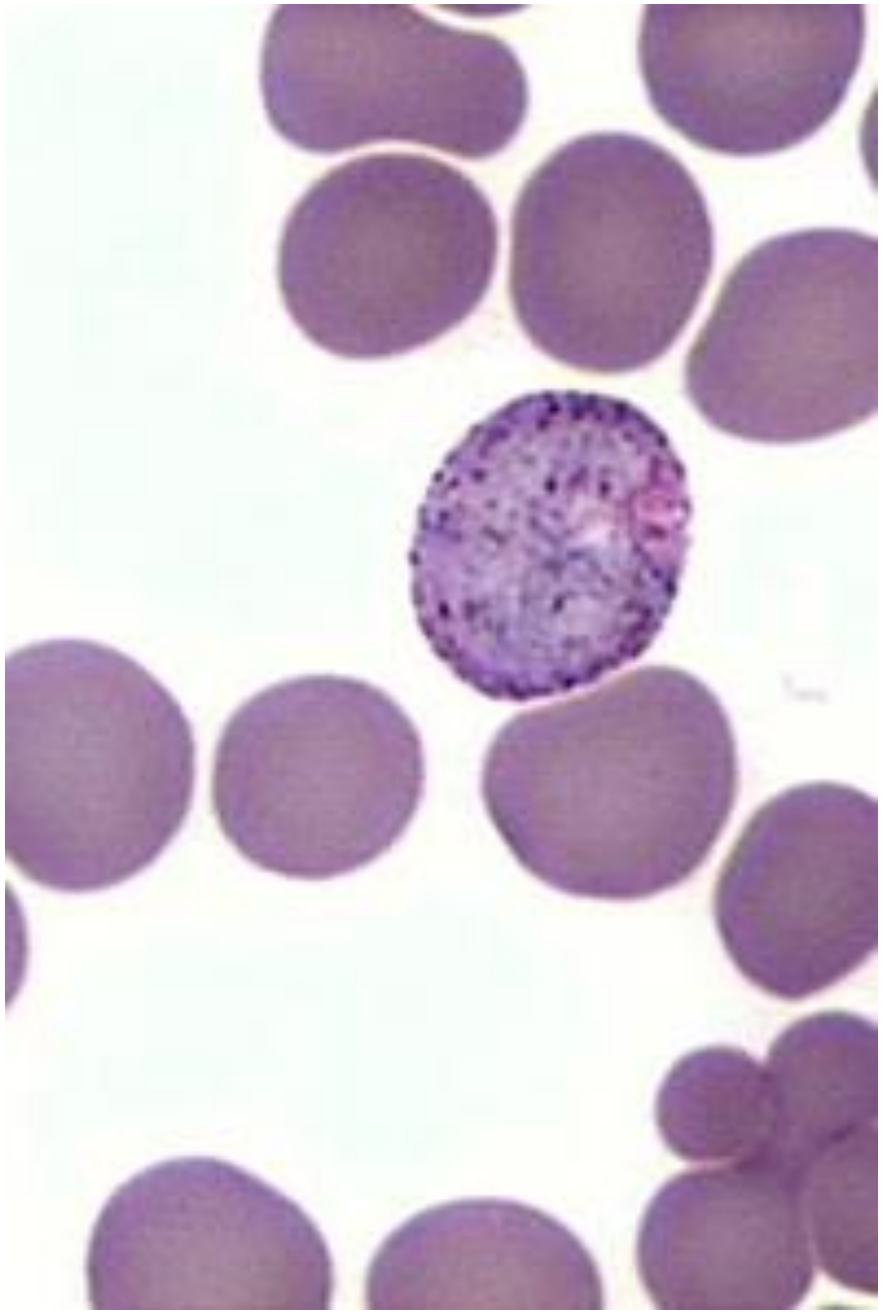


Posterior Probabilities

- ● >0.95
- ◉ 0.8 - 0.96
- ○ <0.8







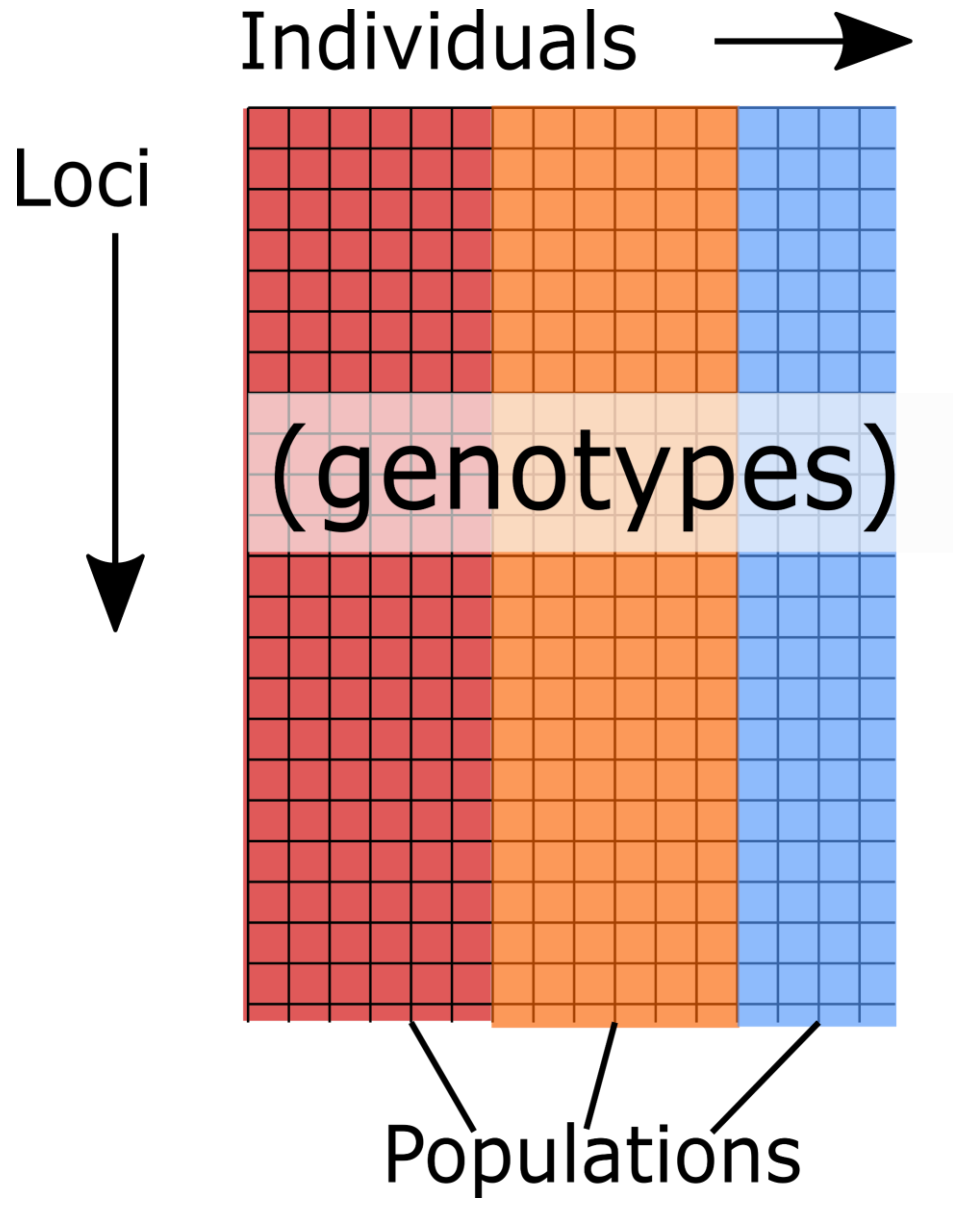
Moving from a few well-genotyped markers to thousands of probabilistically genotyped ones

Lots of pipelines available, each with parameters to tune… but what makes a “good” set of SNVs?

Cahil & Levinton as an example

- GBS dataset from two slipper shell species from East Coast of USA
- 190 specimens from 8 collection sites
- Key question is differentiation/diversity in marginal v central pops
- Raw sequencing reads retrieved from SRA, SNVs called with UNEAK
- Output is a hapmap file

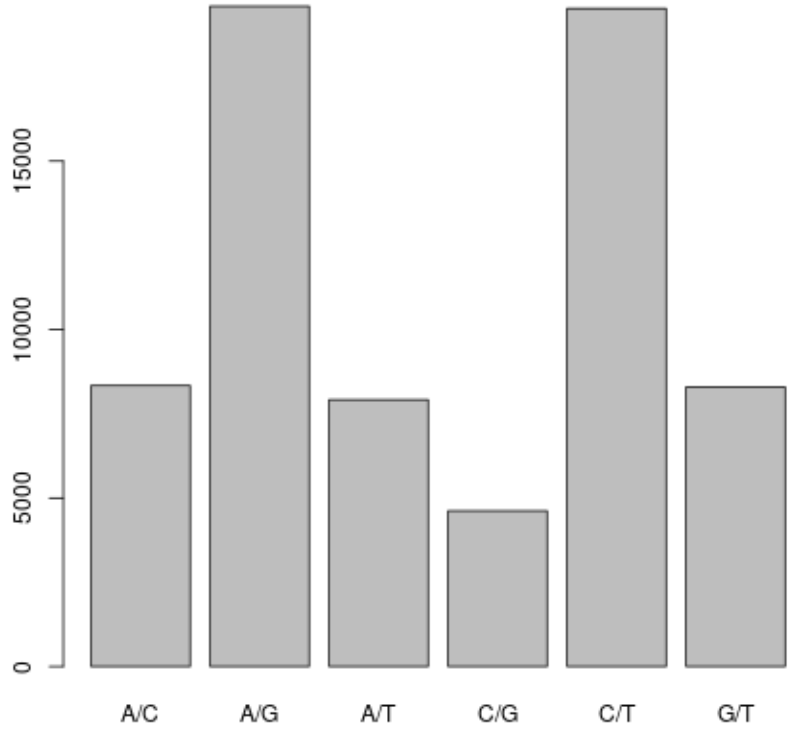
Start with a SNV matrix (.vcf, .hapmap, .tsv)


Individuals
Loci
(genotypes)
Populations

```
> SNVs <-read_hapmap("../vars/uneak_SNPS_hmp.txt.gz")
> dim(SNVs)
```

68 319 loci; 196 individuals

```
>var_spectrum <- table(SNVs$alleles)
>barplot(var_spectrum)
```



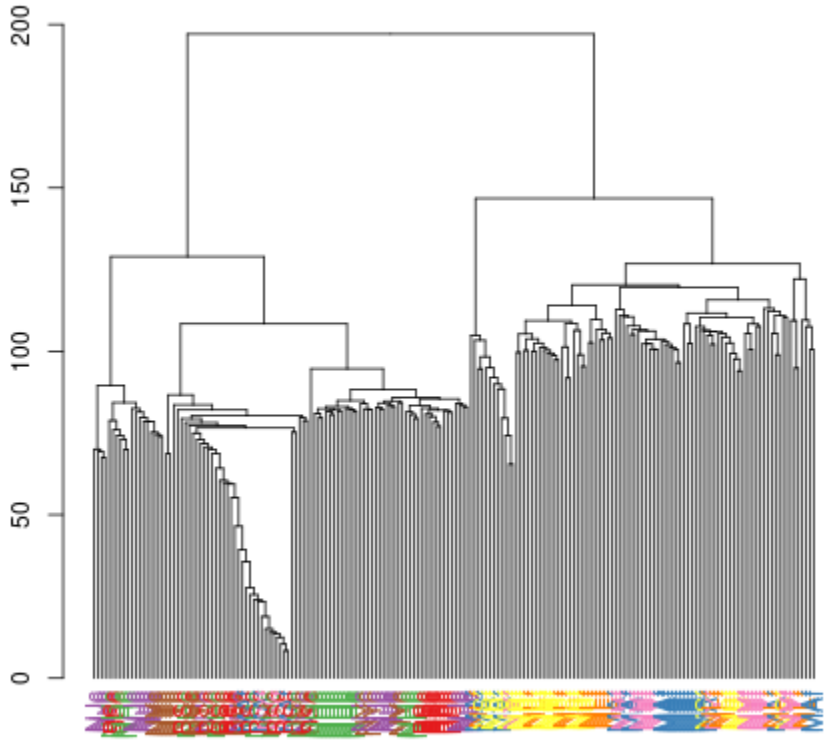
(Non-random) missingness is a feature of GBS

```
>genotypes <- SNVs[ , data_cols(SNVs)]
>missing <- genotypes == "N"
>ind_call_rate <- median(1 - colMeans(missing))
>loc_call_rate <- median(1 - rowMeans(missing))
```

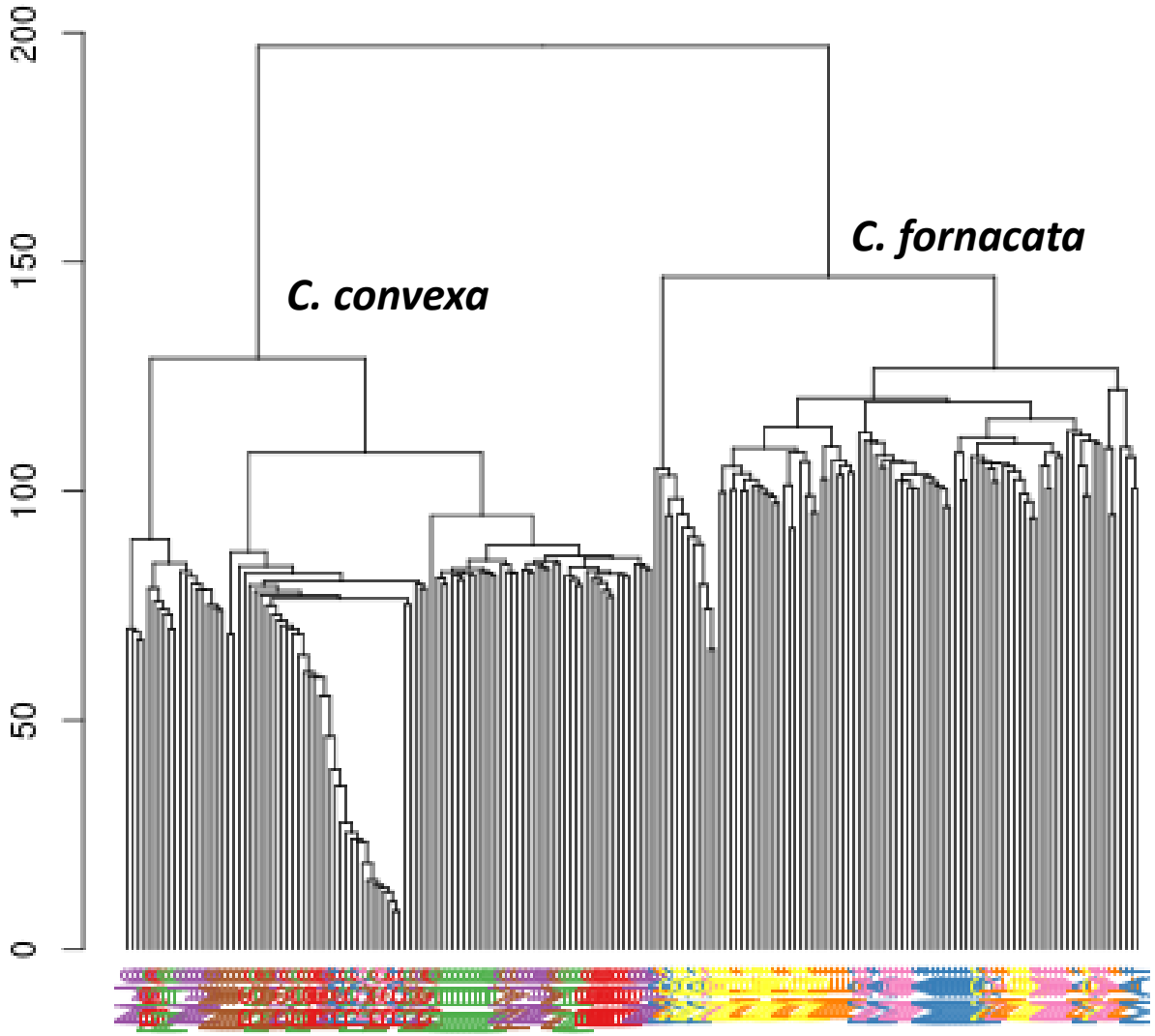
Median ind. call-rate = 0.17

Median locus call-rate = 0.11

```
>D <- dist(t(missing))
>tr <- as.dendrogram(hclust(D))
>plot(tr)
```



```
>D <- dist(t(missing))
>tr <- as.dendrogram(hclust(D))
>plot(tr)
```

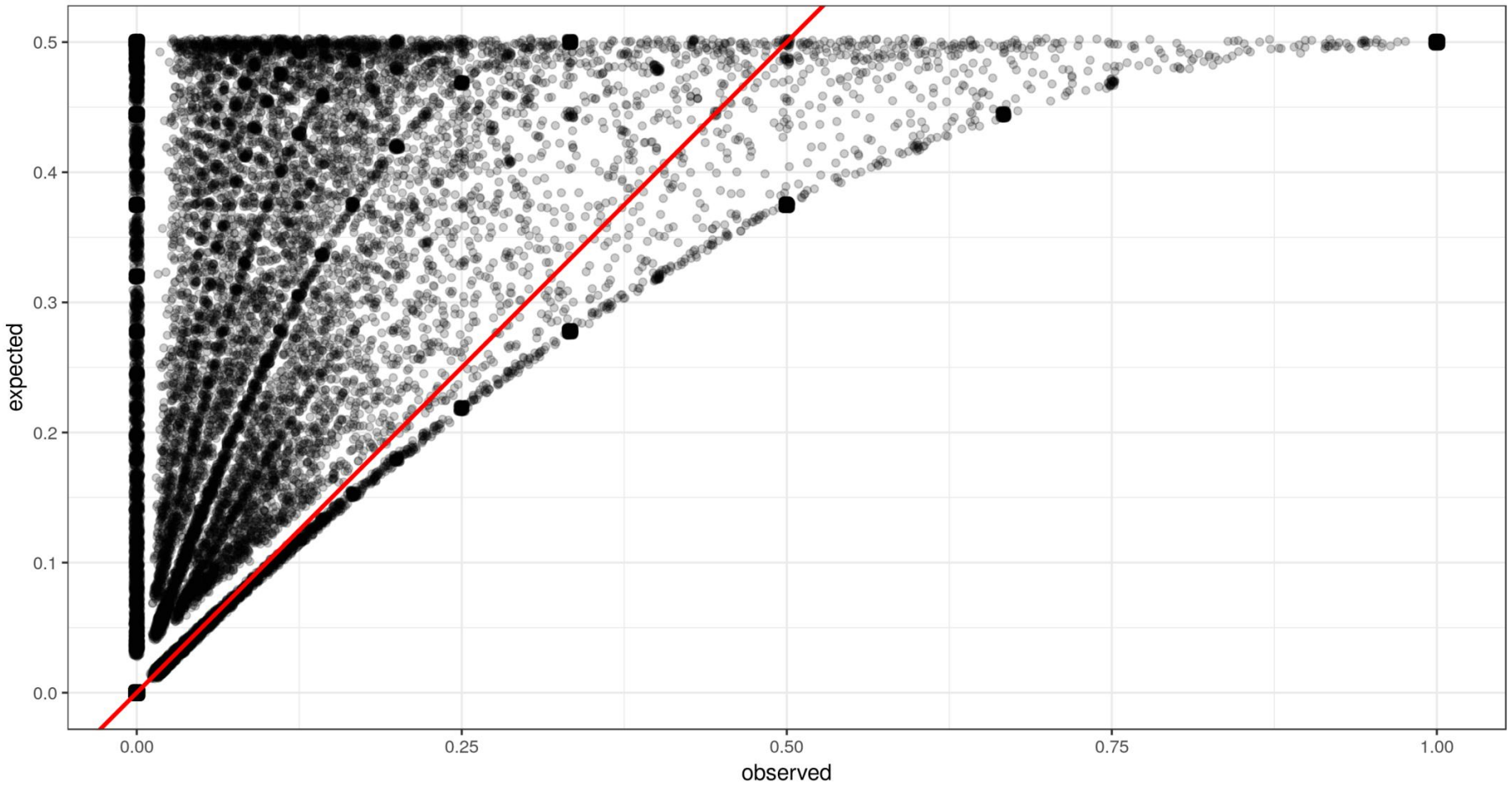


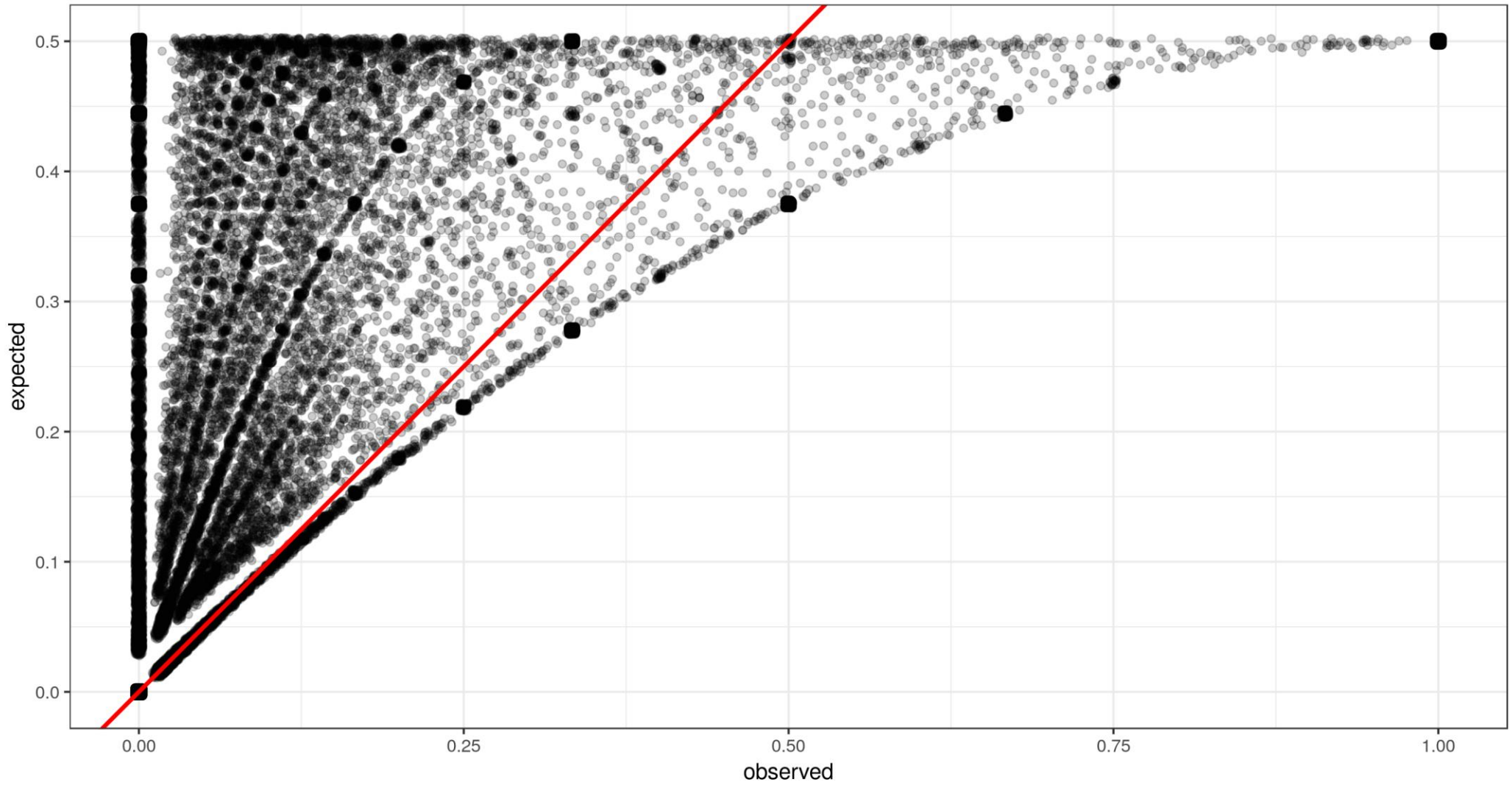
```
>conv <- genotypes[,get_spp(genotypes) == "c"]
#remove monomorphic loci
>conv <- conv[!is.nan(maf(AF(conv))),]
>dim(conv)
```

33982 loci, 95 individuals

Excess heterozygosity suggests sequencing artefacts

```
> conv_afreq  <- AF(conv)
> conv_He  <- He(conv_afreq)
> conv_Ho  <- Ho(conv)
```




0.5
0.4
0.3
0.2
0.1
0.0
expected
0.00
0.25
0.50
0.75
1.00
observed

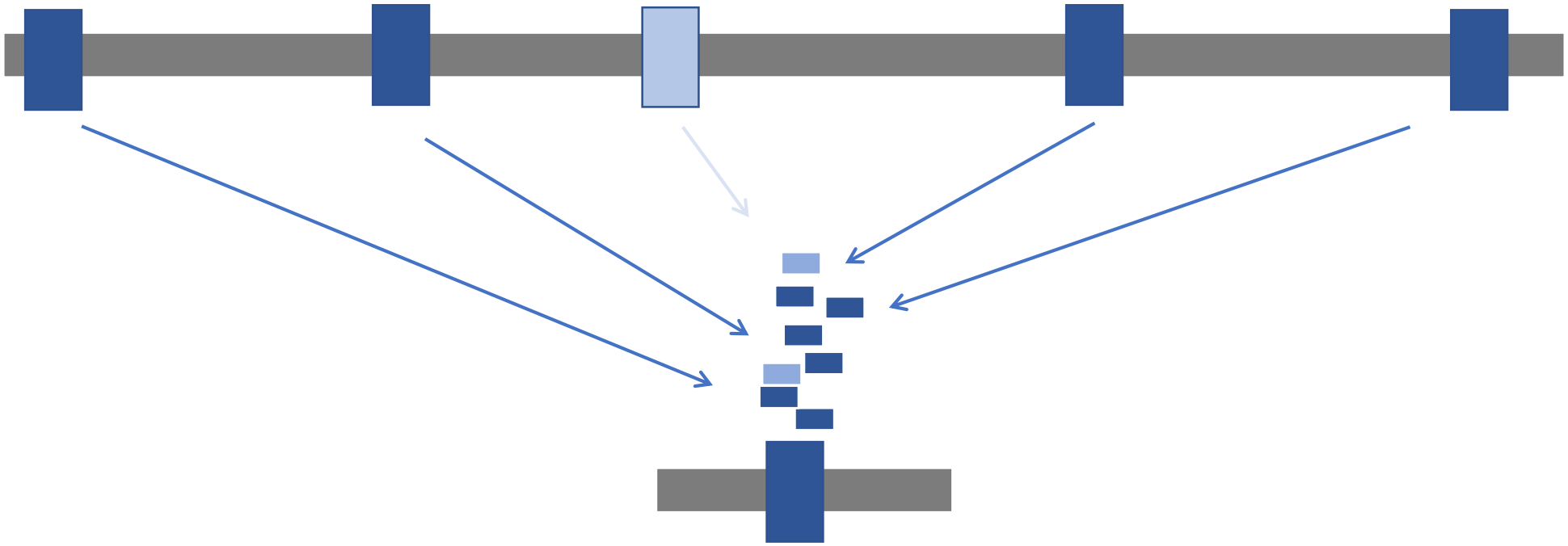
Sequencing depth and heterozygosity

Low-coverage GBS (or any other) sequencing will under-call heterozygotes.

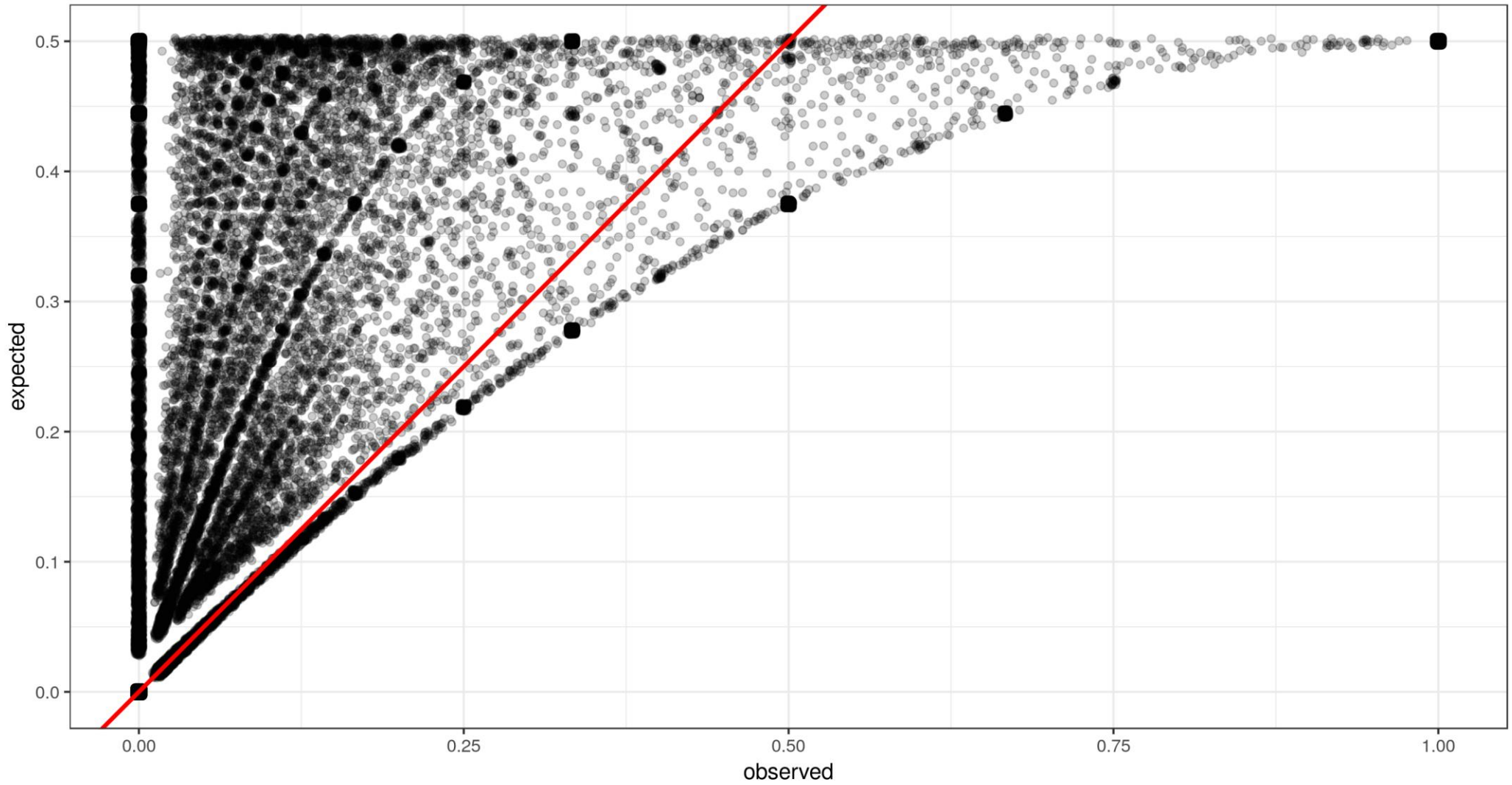
Importance of this fact depends on down-stream analysis

Filtering for only high-depth loci may inflate heterozygosity and increase noise

Real genome



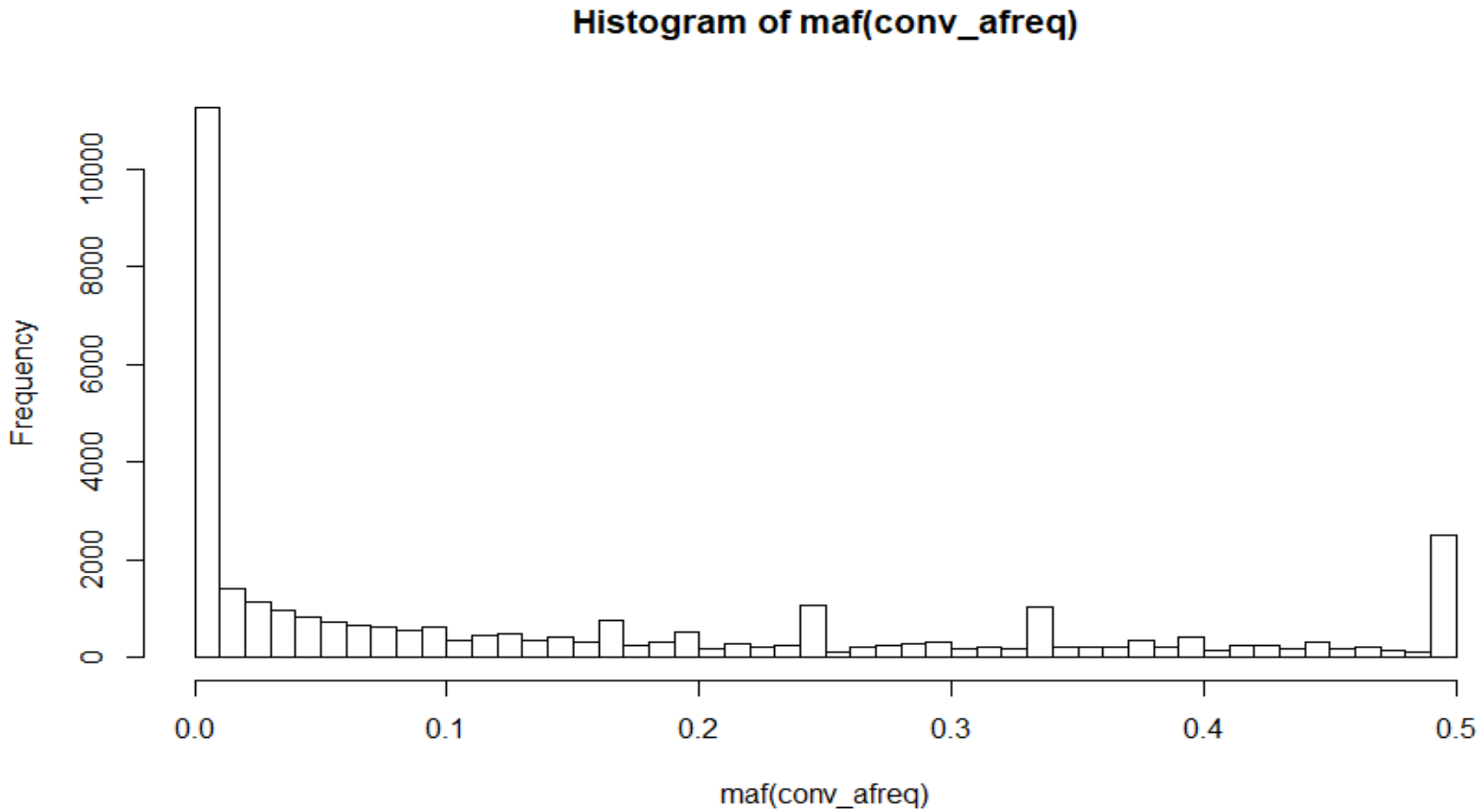
Rad locus


0.5
0.4
0.3
0.2
0.1
0.0
expected
0.00
0.25
0.50
0.75
1.00
observed

Sequencing depth and heterozygosity

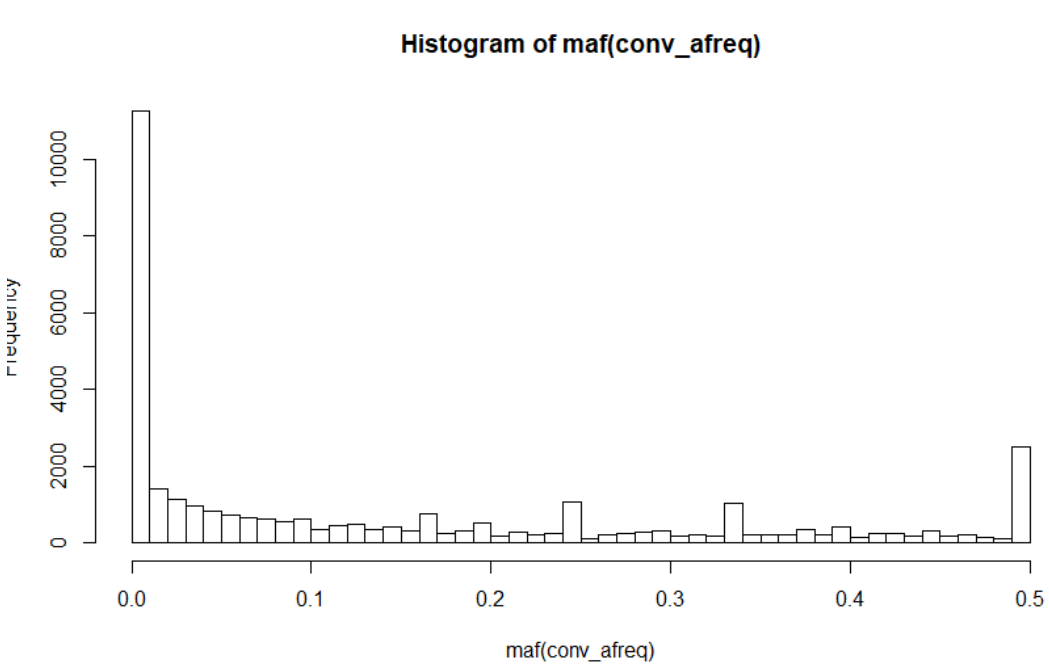
- Check for “minor allele frequency” constantly different than 0.5
- Filter out reads with v. high relative coverage
- If low-coverage loci are included
  - Use statistics that are robust to mis-estimated heterozygosity
  - Use genotype likelihood methods for F_{IS} relatedness etc

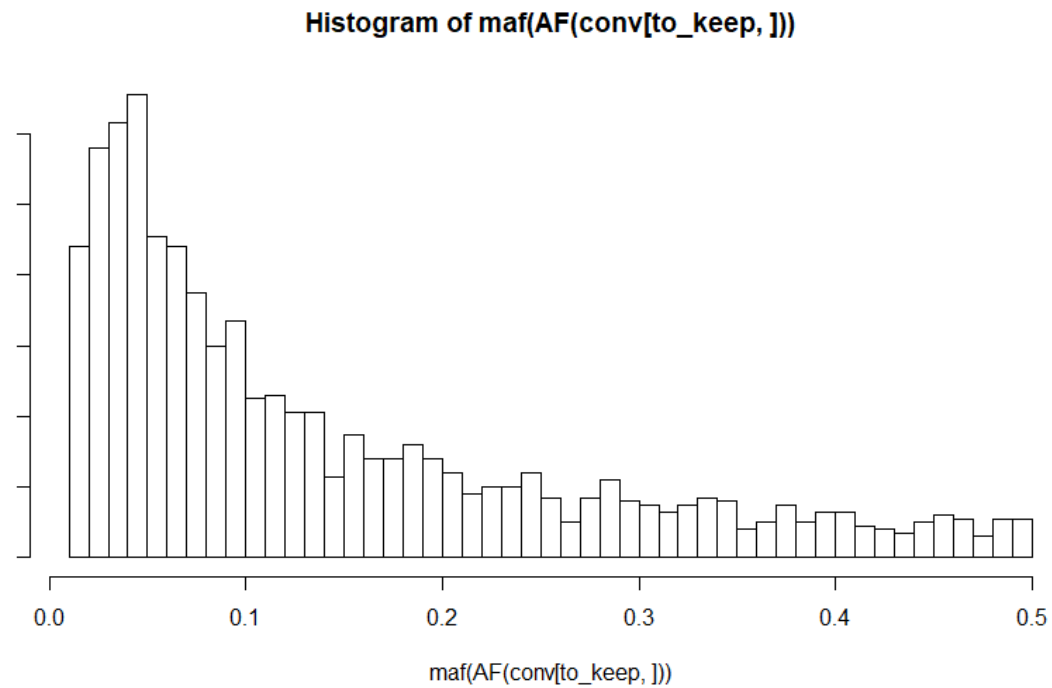
> hist(maf(conv_afreq), breaks=50)



```
> to_keep <-  (rowMeans(conv == "N") < 0.4) & (conv_Ho <= conv_He)
> sum(to_keep)
```

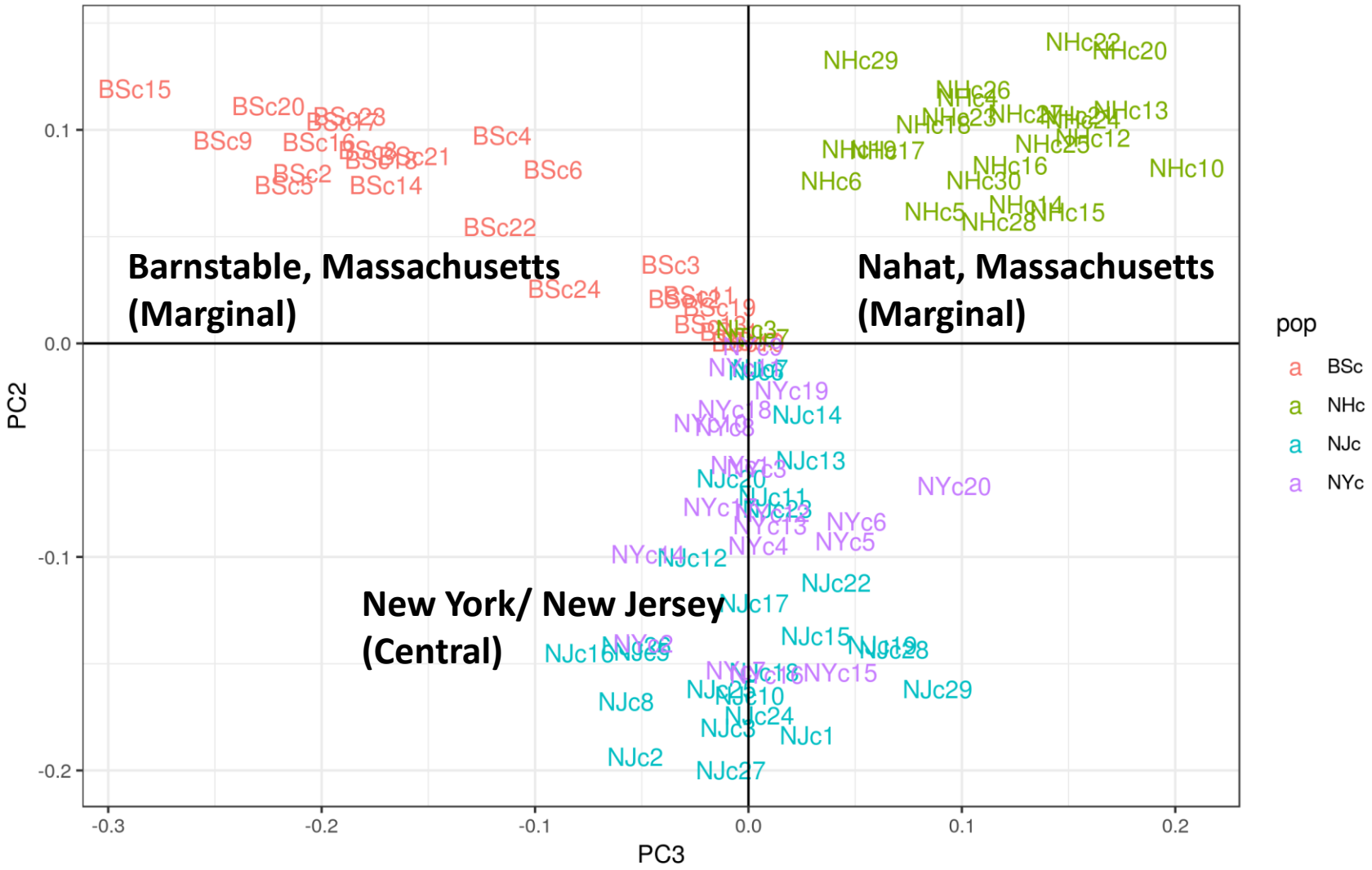
1,6014 loci (still 94 individuals)

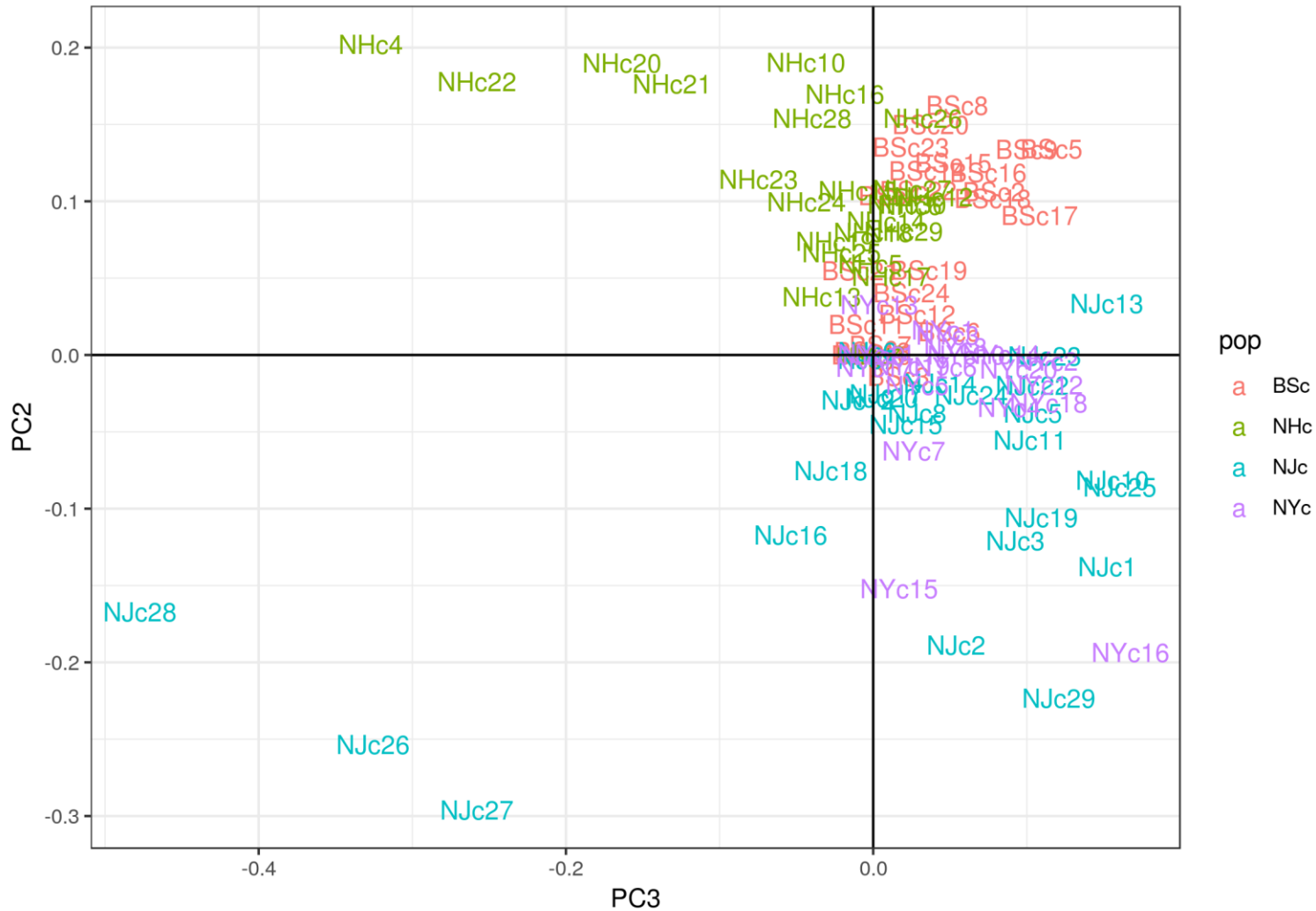

Histogram of maf(conv_afreq)
Frequency
0
2000
4000
6000
8000
10000
0.0
0.1
0.2
0.3
0.4
0.5
maf(conv_afreq)


Histogram of maf(AF(conv[to_keep,]))
0.0
0.1
0.2
0.3
0.4
0.5
maf(AF(conv[to_keep,]))

Perform your “real” analyses on different SNV sets

Biological results should be robust to pipeline used. If they are not, what about one pipleline changes downstream results?


Barnstable, Massachusetts
(Marginal)
Nahat, Massachusetts
(Marginal)
New York/ New Jersey
(Central)
PC2
PC3
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2
pop
a BSc
a NHc
a NJc
a NYc


PC2
PC3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
NHc4
NHc22
NHc20
NHc21
NHc10
NHc16
NHc28
NHc23
NHc24
NHc13
NJc13
NJc18
NJc16
NJc28
NJc26
NJc27
NJc19
NJc3
NJc1
NJc2
NJc29
NYc7
NYc15
NYc16
NJc11
NJc5
NJc25
BSc8
BSc23
BSc17
BSc24
pop
a BSc
a NHc
a NJc
a NYc

Take home messages

- Perform exploratory analyses to check for common issues
- Compare pipelines / paramaters based on potential artefacts
- Filter data to exclude misleading SNVs
- Consider qualities of the data when choosing analyses
- Compare “final” stats between call-sets derived from different pipelines