Population genetic inferences from high-throughput (low-coverage) sequencing data

Filipe G. Vieira

Center for Genomic Medicine

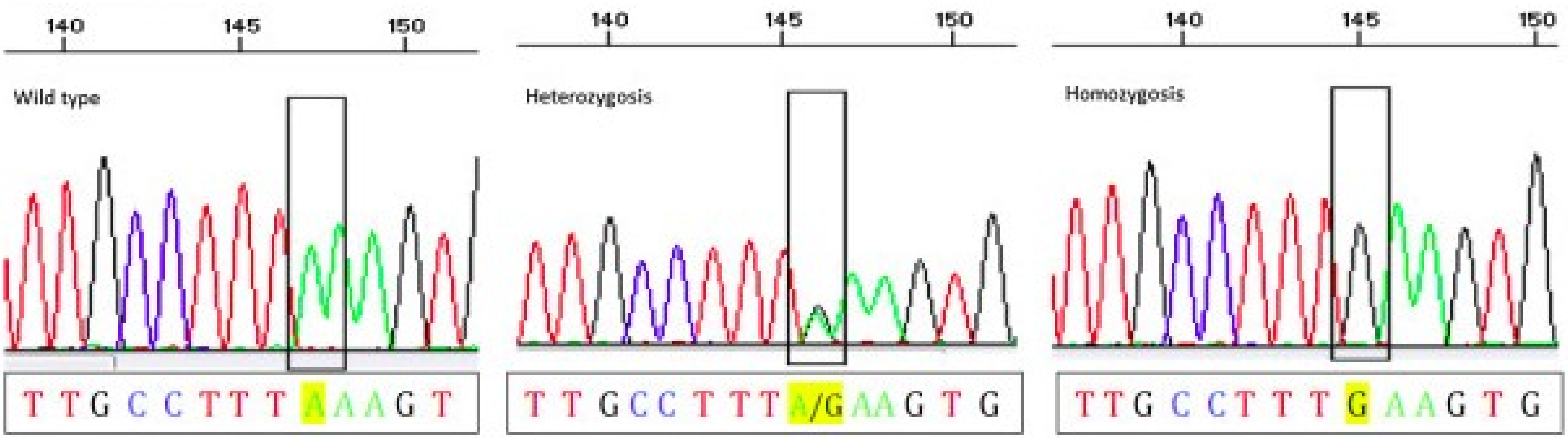
Copenhagen University Hospital, Rigshospitalet

filipe.garrett.vieira@regionh.dk

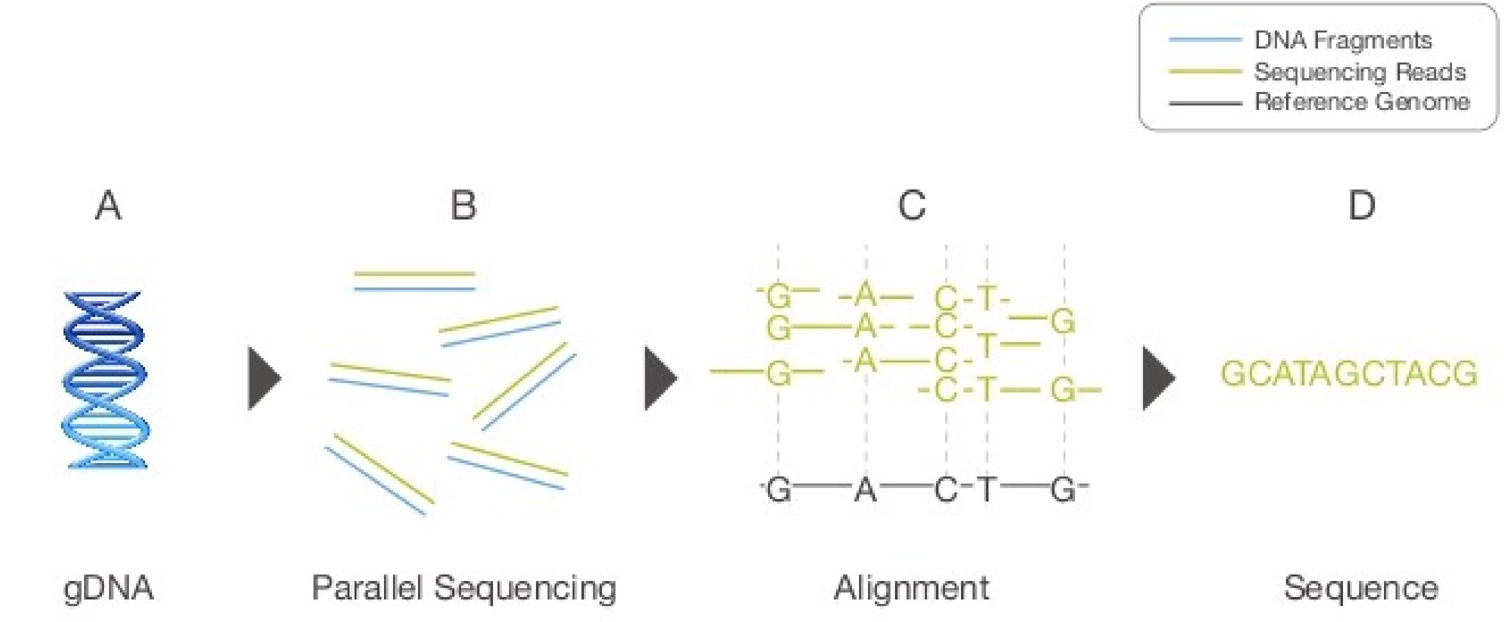
REGION Region Hovedstaden

UNIVERSITY OF COPENHAGEN

Sanger Sequencing



Next Generation Sequencing (NGS)



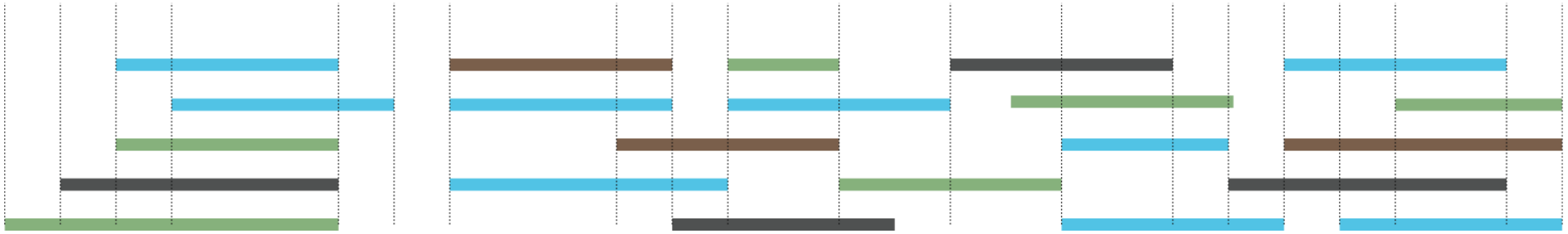
A. Extracted gDNA
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome
D. The whole-genome sequence is derived from the consensus of aligned reads.

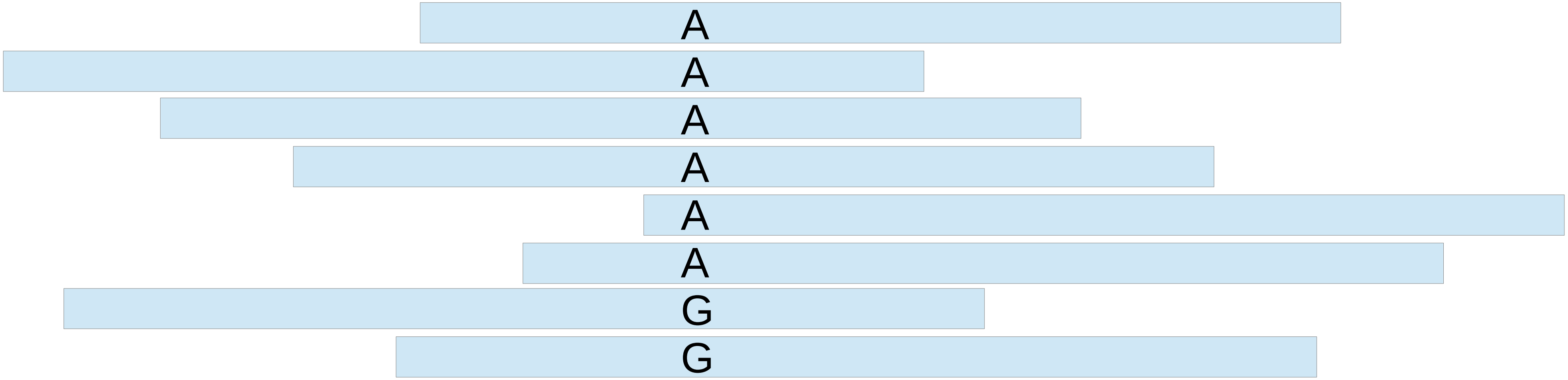
www.illumina.com

However, NGS is not perfect:

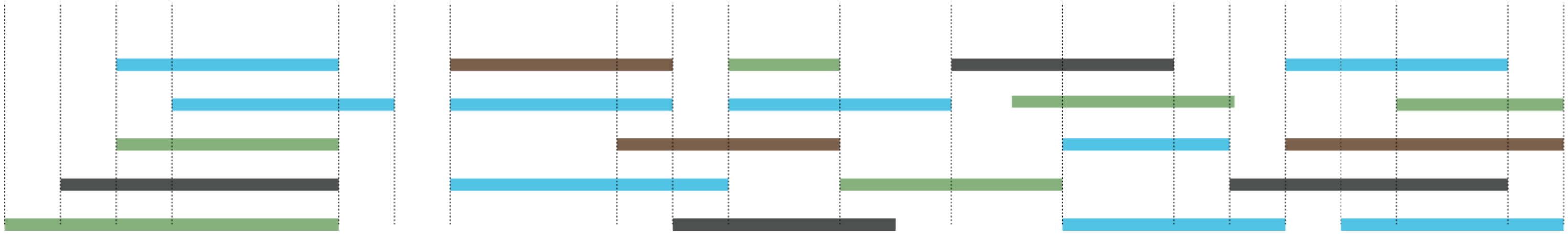
- Higher error rates
- Shorter reads

Sequencing depth / coverage



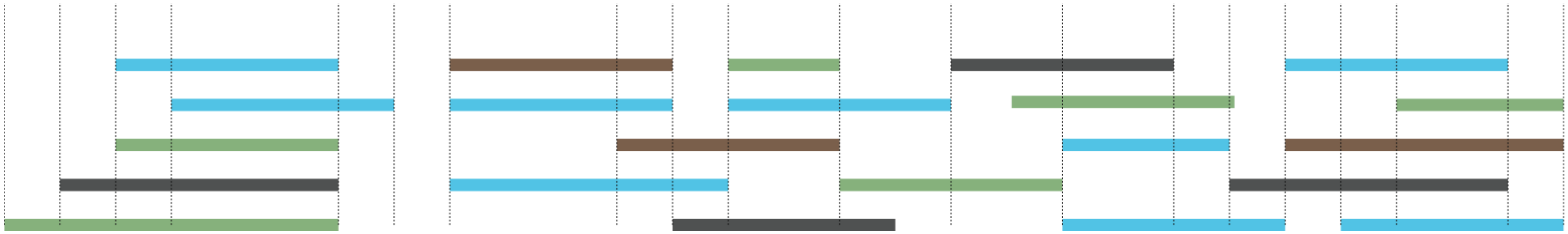


Is the site variable in the sample?



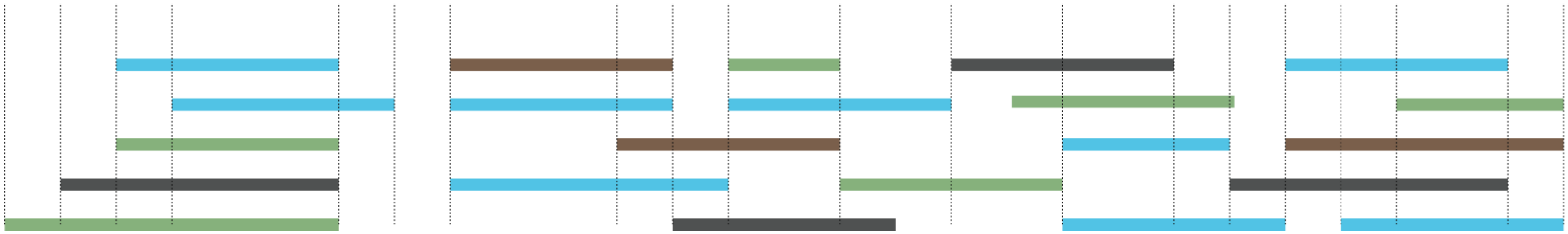
A
A
A
A
A
A
G
G

Is the site still variable in the sample?



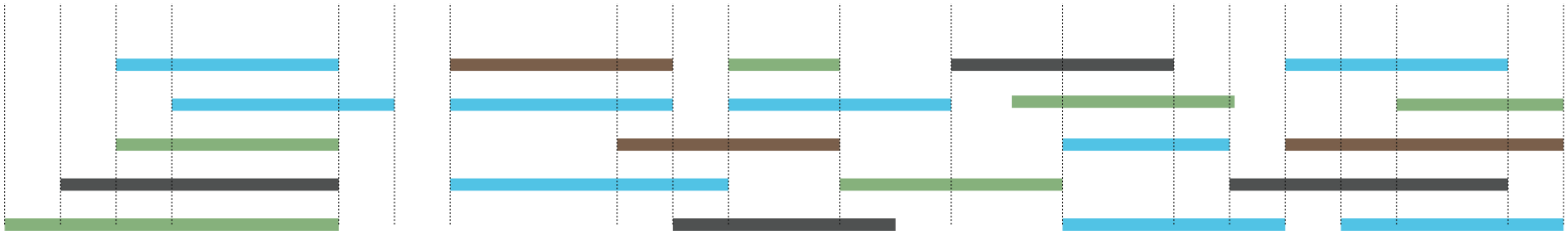
A	A
A	A
A	A
A	A
A	A
A	A
G	G
G	

Is the site still variable in the sample?



```
A   A   A
A   A   A
A   A   A
A   A   A
A   A   A
A   A   A
G   G
G
```

Is the site still variable in the sample/individual?



A	A	A
A	A	A
A	A	A
A	A	A
A	A	A
A	A	A
G	G	
G		

Common errors introduced here:

SNP calling: identification of variable sites.

Genotype calling: determination of the genotype for each site for each individual.

Possible solutions

Possible solutions

More sequencing depth?   More samples?

It depends...

- Fixed budget
  - Balance between sample size and coverage (uncertainty)
  - Depends on objective
    - Reference genome (High coverage)
    - Rare variants (Large samples at High coverage)
    - Population genetics (Large samples)
  - How low?
- How to deal with the uncertainty?
  - Stricter filtering ➔ Loss of data
  - Probabilistic framework (genotype likelihoods)
    - Improved analysis
    - Associated measure of statistical uncertainty
    - Incorporation of **prior** information

Objective

1) What are genotype likelihoods?

2) How to do SNP calling from genotype likelihoods?

3) How to do genotype calling from genotype likelihoods?

4) What is the **error** in population genetic inferences using naïve strategies for **SNP and genotype calling**?

5) What is the optimal **sequencing design** for population genetics purposes?

Objective

1) What are genotype likelihoods?

2) How to do SNP calling from genotype likelihoods?

3) How to do genotype calling from genotype likelihoods?

4) What is the error in population genetic inferences using naïve strategies for SNP and genotype calling?

5) What is the optimal sequencing design for population genetics purposes?

Genotype likelihoods

Probability of observing the read data, given particular **genotype**

$$p\left(X \middle| G = bh\right) = \frac{1}{2^r} \prod_{i=1}^{r} \left(L_b^{(i)} + L_h^{(i)}\right)$$

Likelihood of observing allele *b* at read *i*

Genotype likelihoods – an example

-A-

-A-

-C-

-T-

Where can we get the error rate from?

$$P(X|AA)=\left(\frac{L_A^{(1)}}{2}+\frac{L_A^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_A^{(2)}}{2}\right)*\left(\frac{L_A^{(3)}}{2}+\frac{L_A^{(3)}}{2}\right)*\left(\frac{L_A^{(4)}}{2}+\frac{L_A^{(4)}}{2}\right)$$

$$L_A^{(1)}=L_A^{(2)}=1-\epsilon \qquad L_A^{(3)}=L_A^{(4)}=\frac{\epsilon}{3} \qquad (1-\epsilon)+\left(\frac{\epsilon}{3}\right)+\left(\frac{\epsilon}{3}\right)+\left(\frac{\epsilon}{3}\right)=1$$

$$P(X|AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*\left(\frac{L_A^{(3)}}{2}+\frac{L_C^{(3)}}{2}\right)*\left(\frac{L_A^{(4)}}{2}+\frac{L_C^{(4)}}{2}\right)$$

$$L_A^{(1)}=L_A^{(2)}=L_C^{(3)}=1-\epsilon \qquad L_C^{(1)}=L_C^{(2)}=L_A^{(3)}=L_A^{(4)}=L_C^{(4)}=\frac{\epsilon}{3}$$

Posterior probabilities of genotypes

Prior is derived assuming **HWE** from the estimated Minor Allele Frequency.

Genotype likelihood Prior information

$$P\left(G_s^{(i)} | X_s^{(i)}\right) = \frac{P\left(X_s^{(i)} | G_s^{(i)}\right) P\left(G_s^{(i)}\right)}{\sum_{G=0}^{2} P\left(X_s^{(i)} | G_s^{(i)}\right) P\left(G_s^{(i)}\right)}$$

Nielsen et al 2012

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Priors

- Model organisms
  - Reference genome
  - SNP databases
  - Patterns of linkage disequilibrium (LD)
  - Known allele or genotype frequencies
  - …

- Non-model organisms
  - Expected genotype frequencies under some model (e.g. HWE)
    - Works for most cases
    - But not always
      - Self-polinating plants
      - Domesticated species (due to inbreeding and clonal propagation)
      - Asexual life cycles

Objective

1) What are genotype likelihoods?

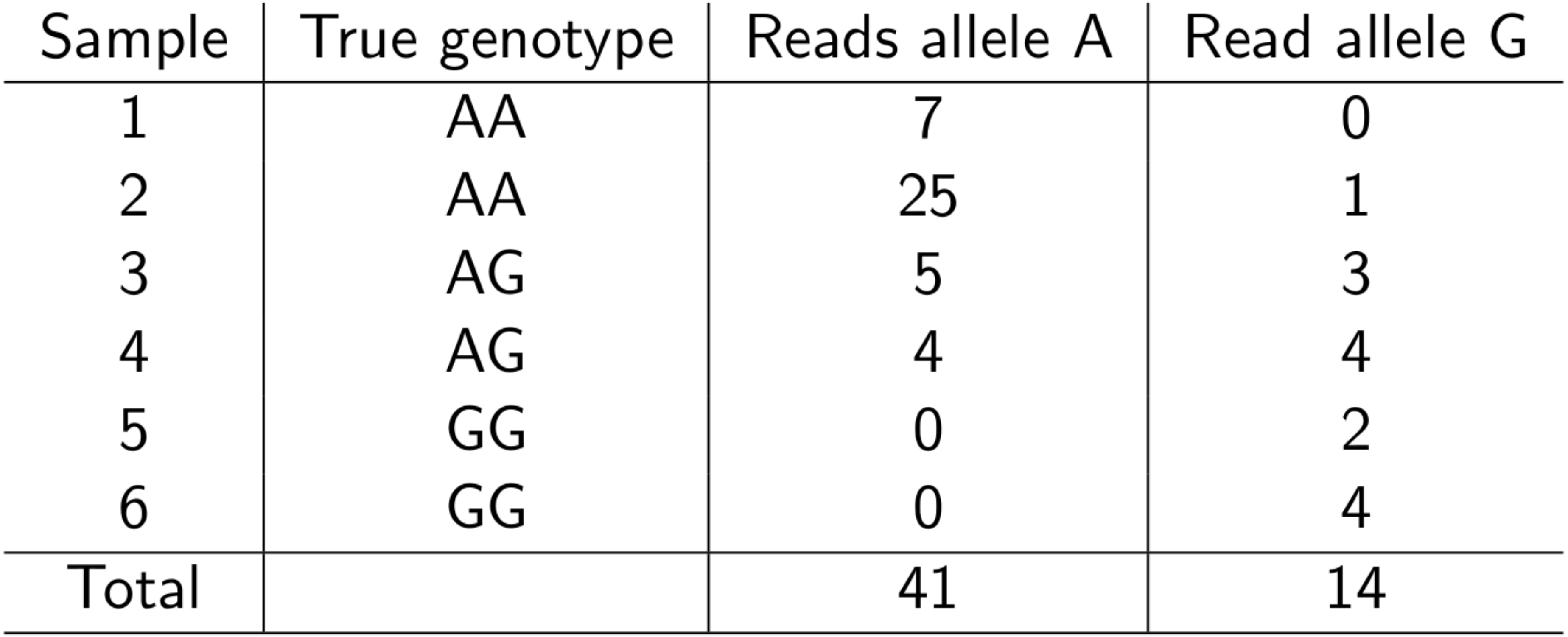
2) How to do SNP calling from genotype likelihoods?

3) How to do genotype calling from genotype likelihoods?

4) What is the error in population genetic inferences using naïve strategies for SNP and genotype calling?

5) What is the optimal sequencing design for population genetics purposes?

Estimating Allele Frequencies



Sample	True genotype	Reads allele A	Read allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Total		41	14

What is the estimated frequency?

What is the true frequency?

What is wrong with that estimate?

Estimating Allele Frequencies - ML

$$P(D|f) = \prod_{i=1}^{N} \sum_{g \in \{0,1,2\}} P(D|G=g)P(G=g|f)$$

- Likelihood function, where:
  - P(D | G)
  - P(G = g | f)
- Estimate *f*, by optimizing the likelihood function through an EM
  - *f* = 0.46

Calling SNP

- ANGSD uses the minor allele frequency (MAF) to call SNPs
  - $f > t$ (e.g., t = 1/2N)
  - Likelihood Ratio Test (LRT), comparing the goodness of fit (chi2) between:
    - null model: $f = 0$
    - alternative model: $f <> 0$

Objective

1) What are genotype likelihoods?

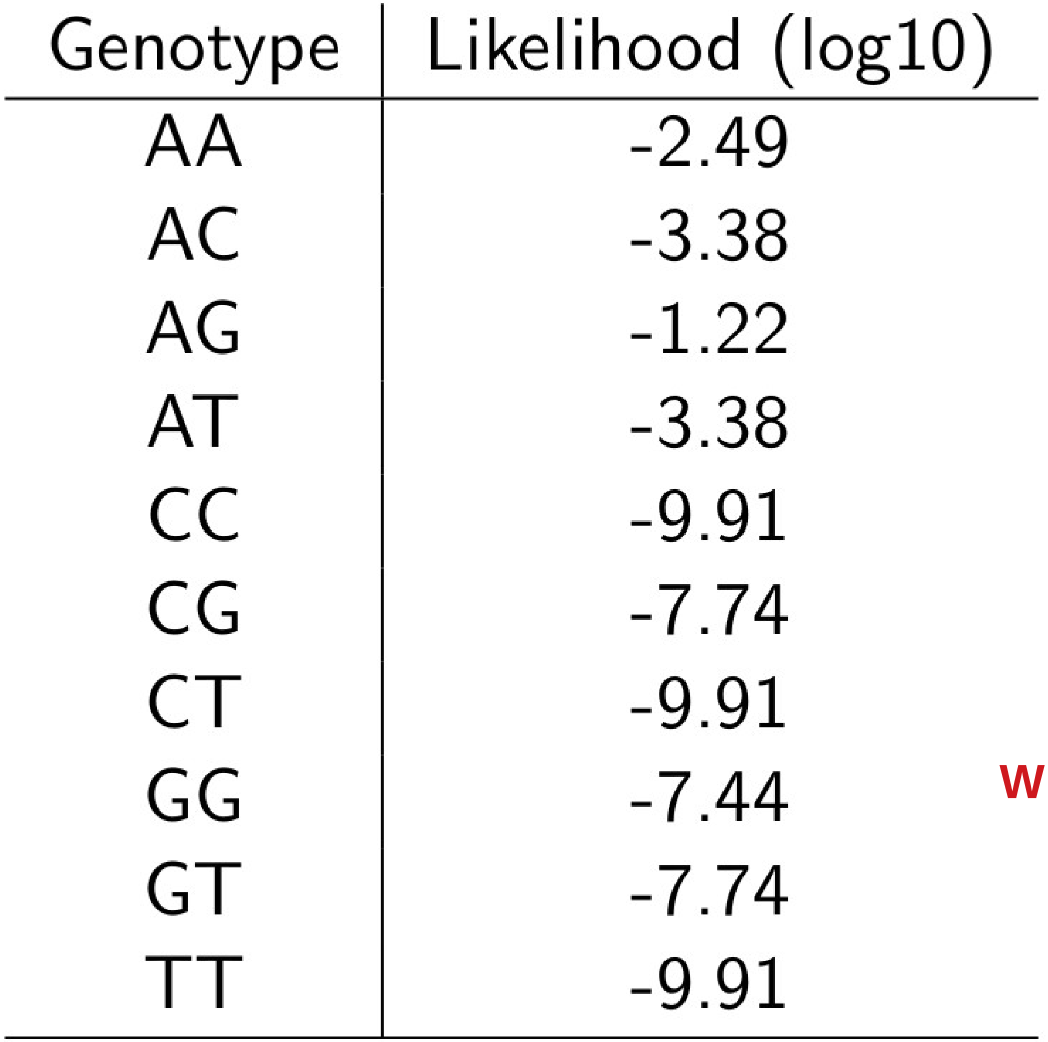
2) How to do SNP calling from genotype likelihoods?

3) How to do genotype calling from genotype likelihoods?

4) What is the error in population genetic inferences using naïve strategies for SNP and genotype calling?

5) What is the optimal sequencing design for population genetics purposes?

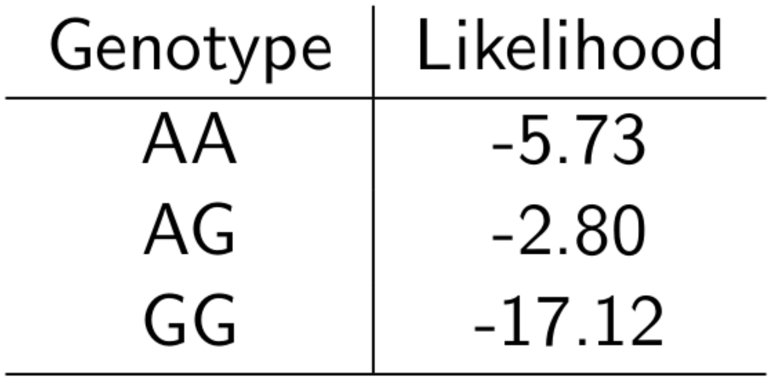
Calling genotypes – Genotype likelihoods (GL)



Genotype	Likelihood (log10)
AA	-2.49
AC	-3.38
AG	-1.22
AT	-3.38
CC	-9.91
CG	-7.74
CT	-9.91
GG	-7.44
GT	-7.74
TT	-9.91

What is the genotype?

Calling genotypes – Major / Minor alleles



Genotype	Likelihood
AA	-5.73
AG	-2.80
GG	-17.12

What is the genotype?

Calling genotypes – GL ratio

$$\log_{10} \frac{L_{G(1)}}{L_{G(2)}} > t$$

i.e. $t = 1$ meaning that the most likely genotype is 10 times more likely than the second most likely one

Pros and Cons?

Genotype Quality?

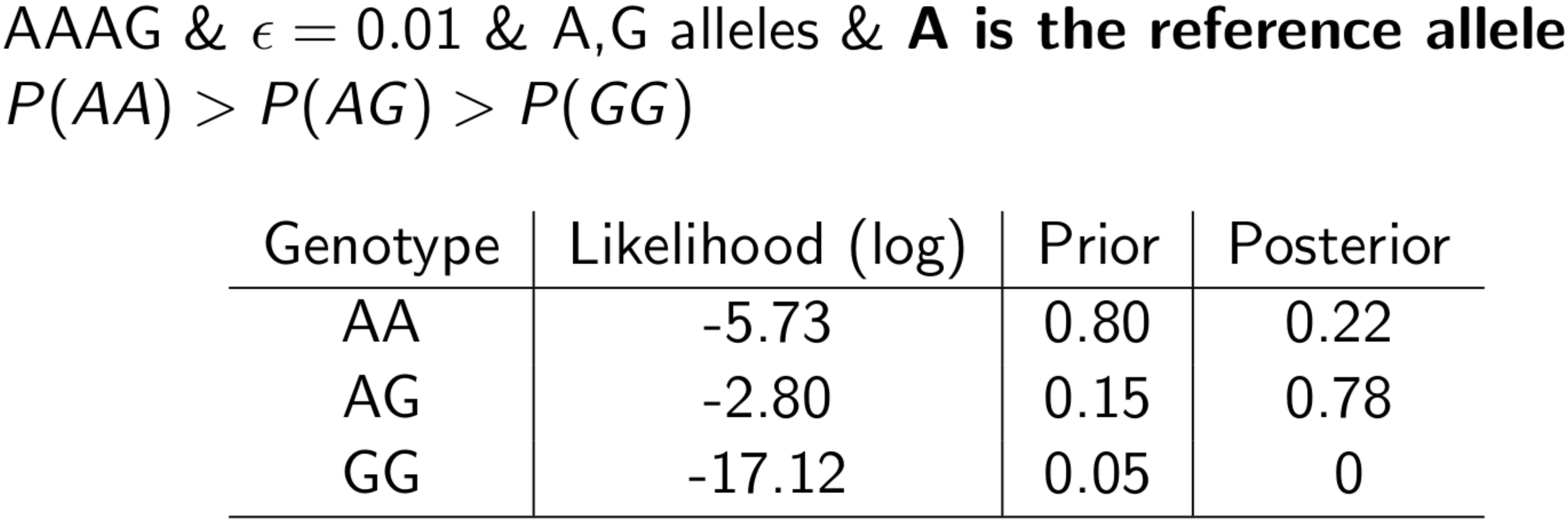
Missing data?

Calling genotypes – Posterior Probabilities (PP)

AAAG & $\epsilon = 0.01$ & A,G alleles

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	1/3	0.05
AG	-2.80	1/3	0.95
GG	-17.12	1/3	0

Calling genotypes – GPP (reference prior)



AAAG & $\epsilon = 0.01$ & A,G alleles & **A is the reference allele**

$P(AA) > P(AG) > P(GG)$

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	0.80	0.22
AG	-2.80	0.15	0.78
GG	-17.12	0.05	0

Calling genotypes – GPP (HWE prior)

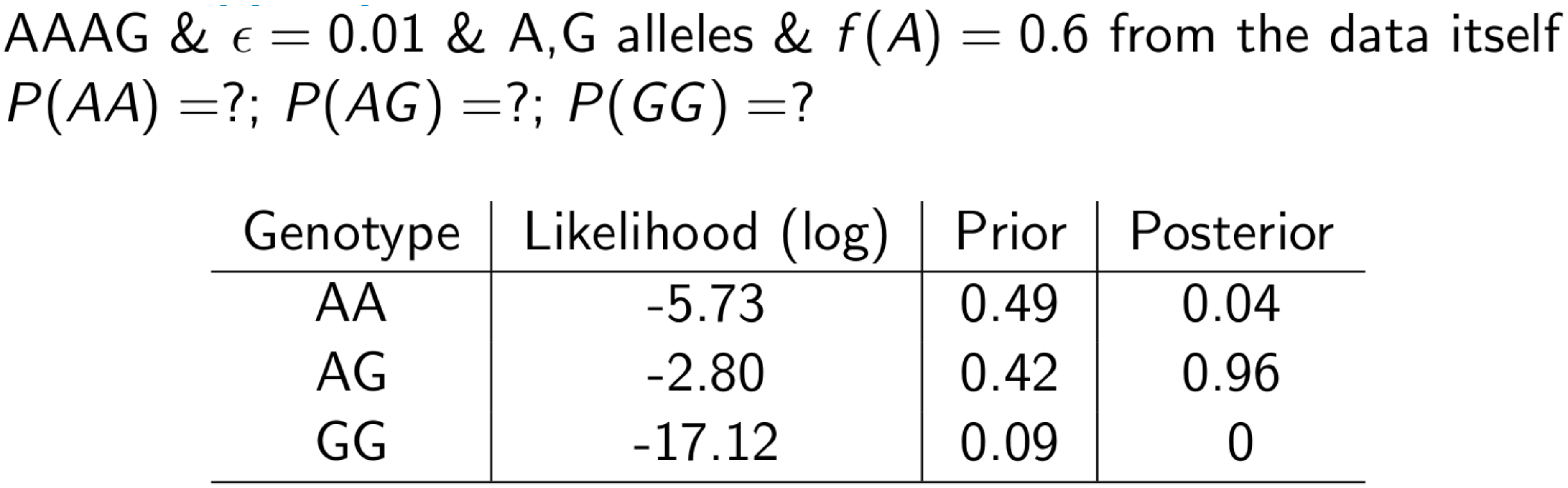
AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from a reference panel

$P(AA) = ?$; $P(AG) = ?$; $P(GG) = ?$

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	0.49	0.06
AG	-2.80	0.42	0.94
GG	-17.12	0.09	0

Can we assume HWE?

Calling genotypes – GPP (HWE prior)



AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.6$ from the data itself

$P(AA) = ?$; $P(AG) = ?$; $P(GG) = ?$

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	0.49	0.04
AG	-2.80	0.42	0.96
GG	-17.12	0.09	0

Can we assume HWE?

Can we estimate freqs accurately?

Objective

1) What are genotype likelihoods?

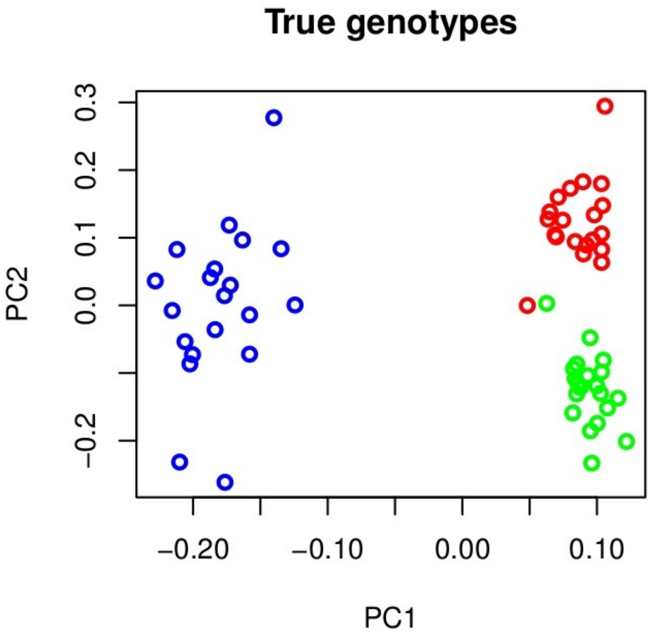
2) How to do SNP calling from genotype likelihoods?

3) How to do genotype calling from genotype likelihoods?

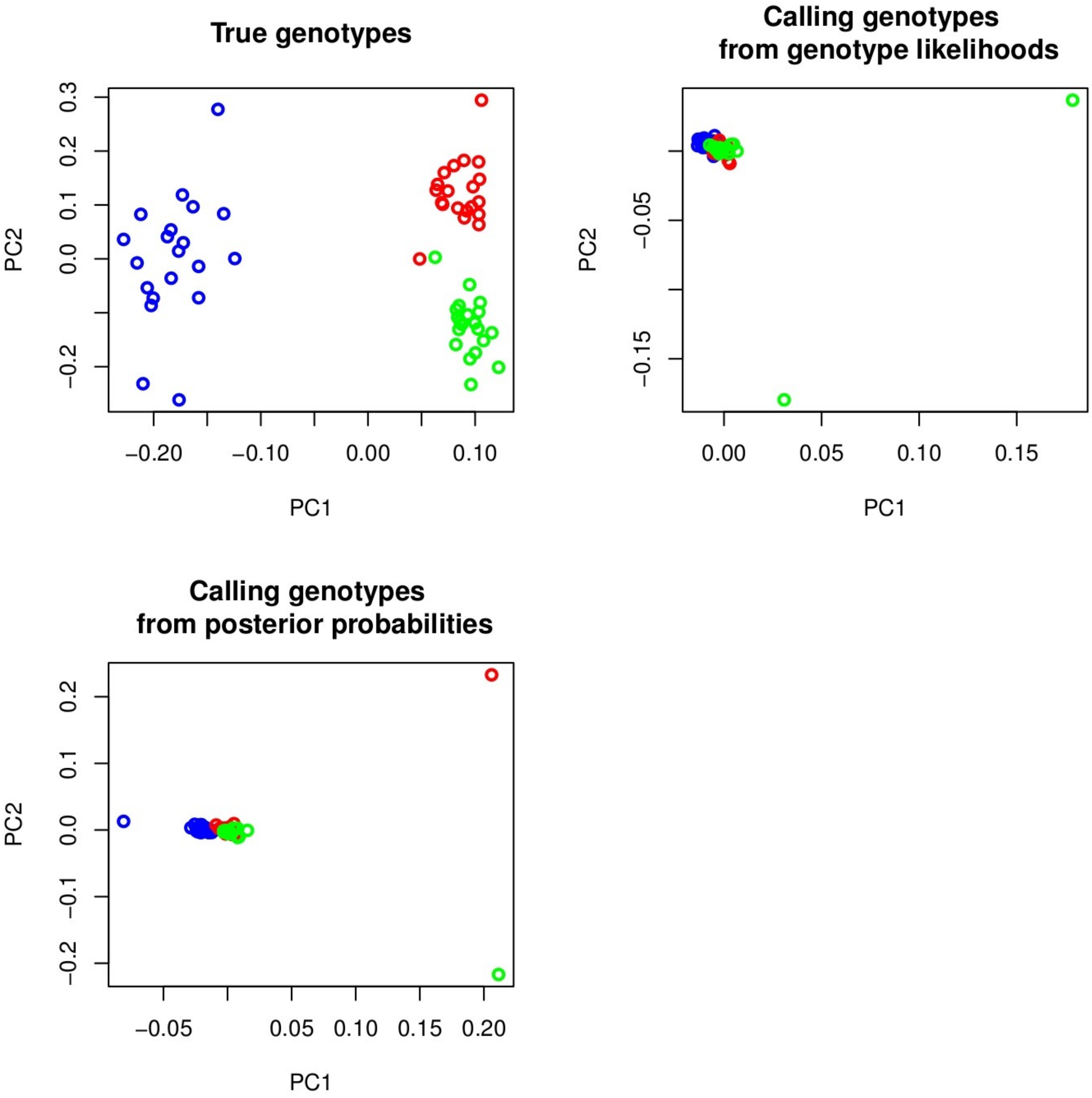
4) What is the **error** in population genetic inferences using naïve strategies for **SNP and genotype calling**?

5) What is the optimal sequencing design for population genetics purposes?

Population structure


True genotypes

Population structure

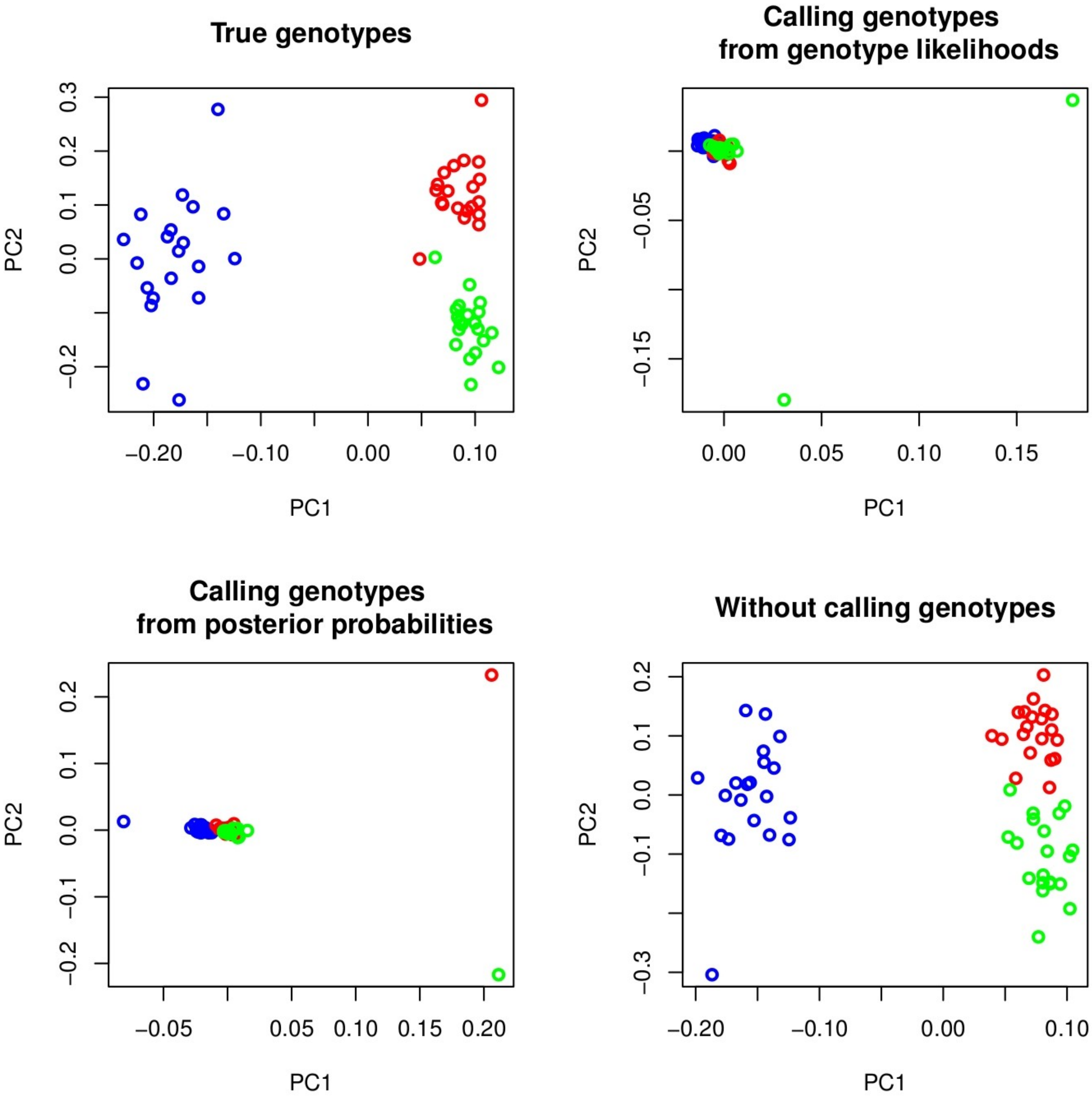


True genotypes

Calling genotypes from genotype likelihoods

Calling genotypes from posterior probabilities

Population structure



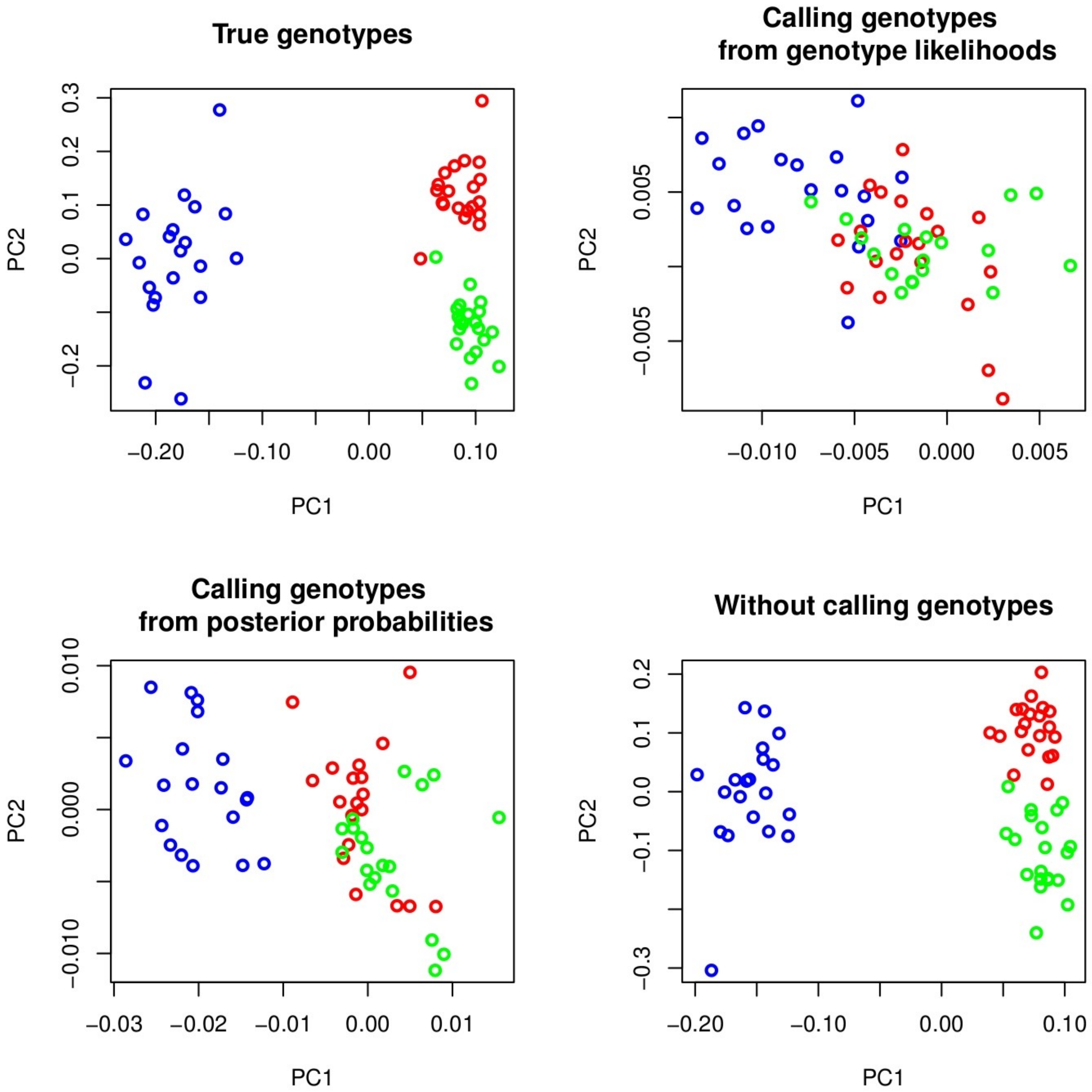
True genotypes

Calling genotypes from genotype likelihoods

Calling genotypes from posterior probabilities

Without calling genotypes

Population structure



Objective

1) What are genotype likelihoods?

2) How to do SNP calling from genotype likelihoods?

3) How to do genotype calling from genotype likelihoods?

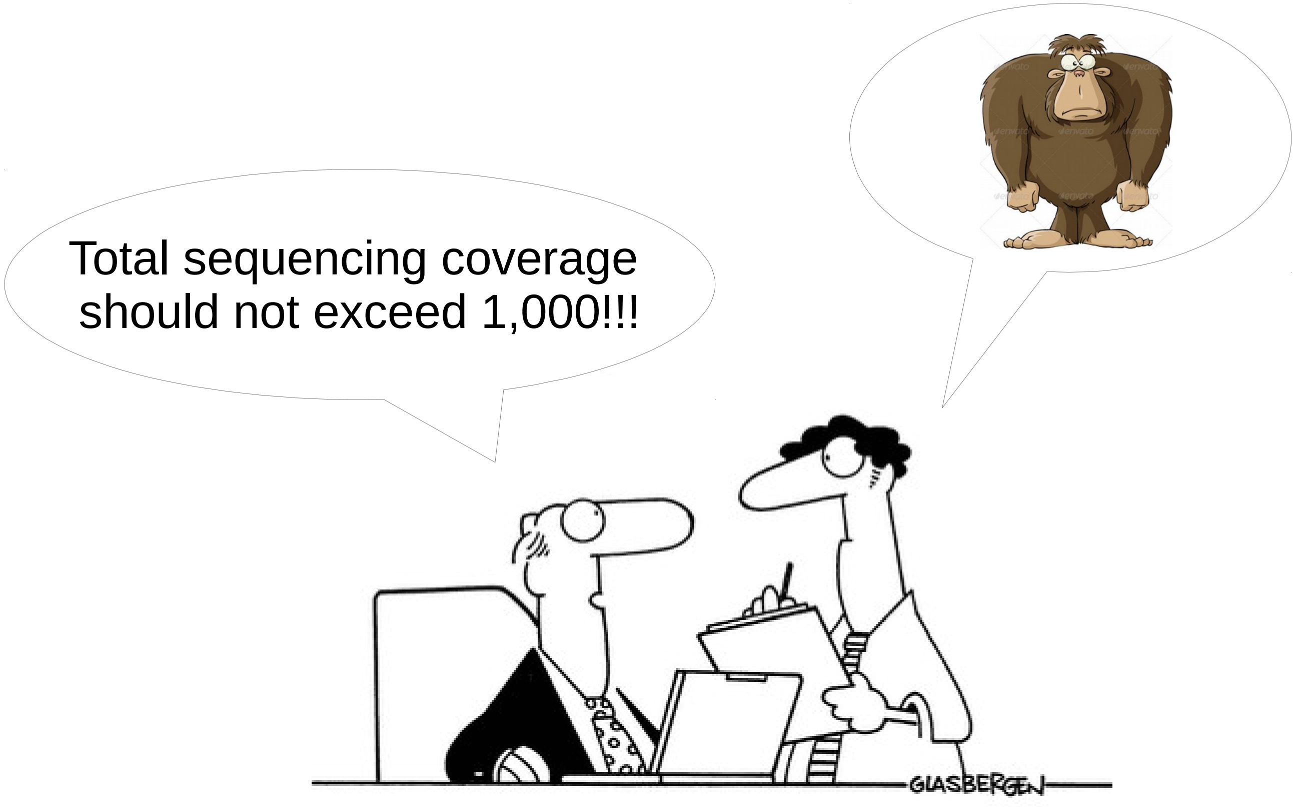
4) What is the error in population genetic inferences using naïve strategies for SNP and genotype calling?

5) What is the optimal **sequencing design** for population genetics purposes?

Discovery of a “new” species

Population is comprised of **1,000 individuals**.

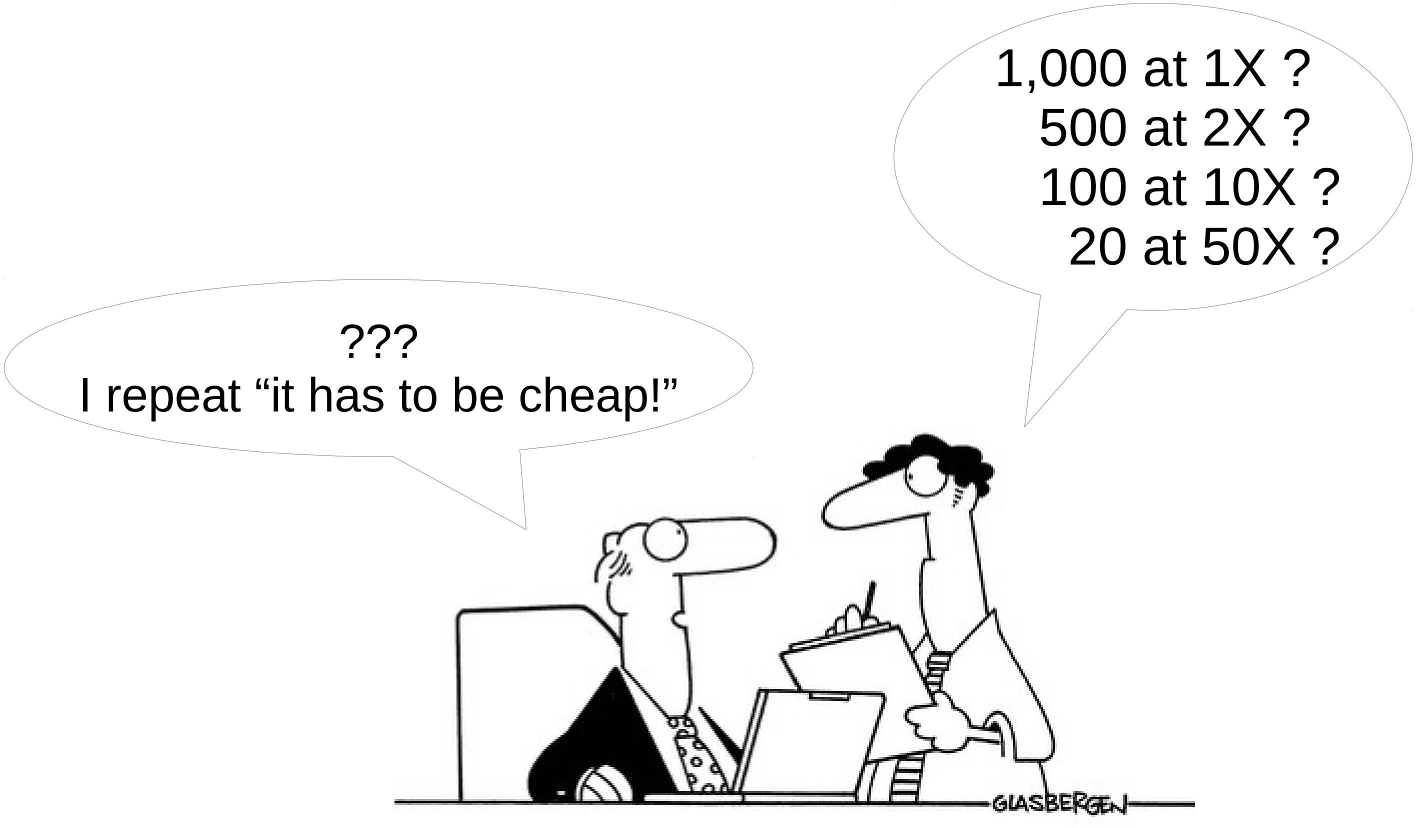
Genome is **100,000 bp** long.



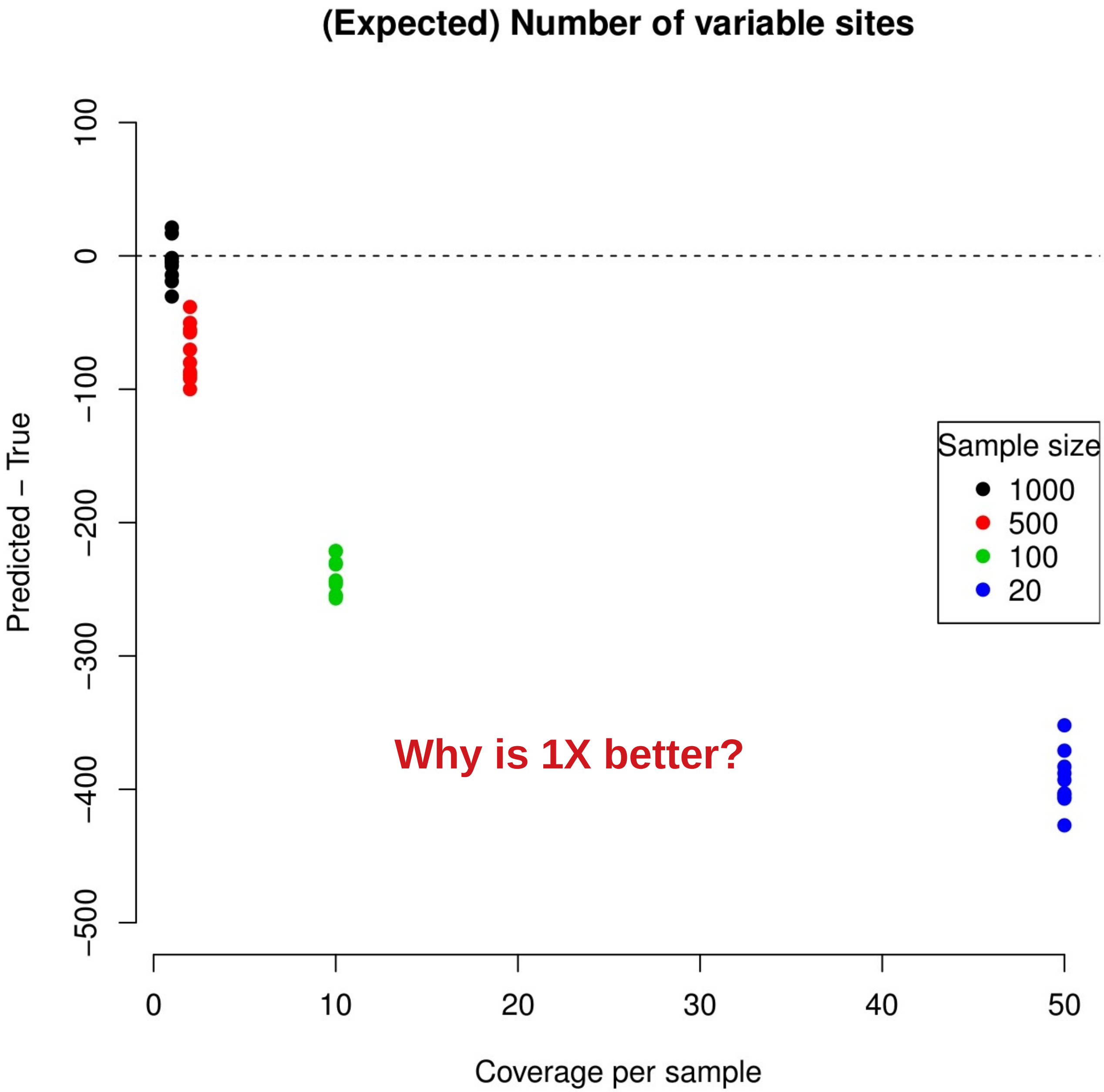
Planning the experiment

Population is comprised of **1,000 individuals**.

Genome is **100,000 bp** long.

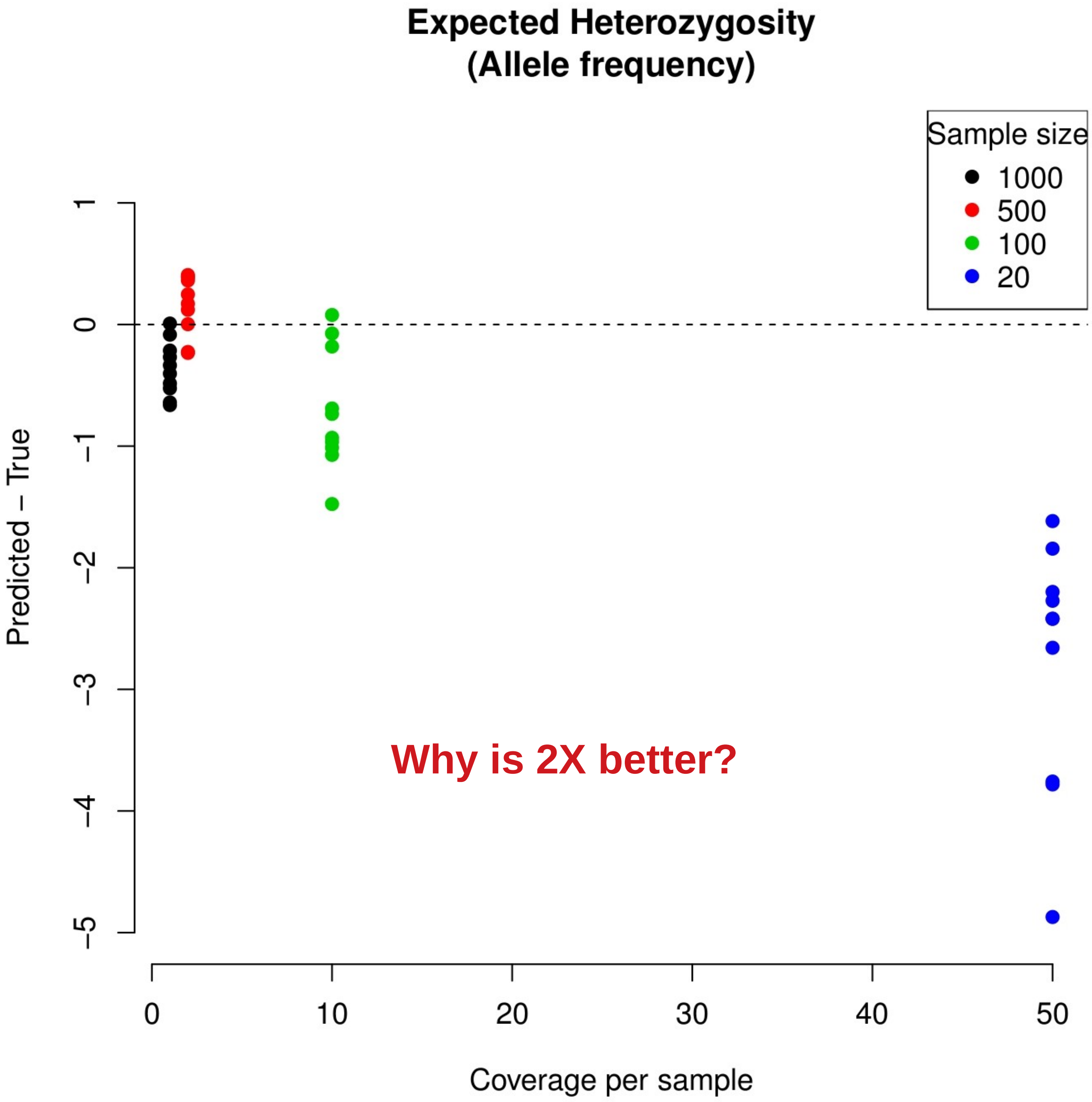


How many polymorphic sites?



Why is 1X better?

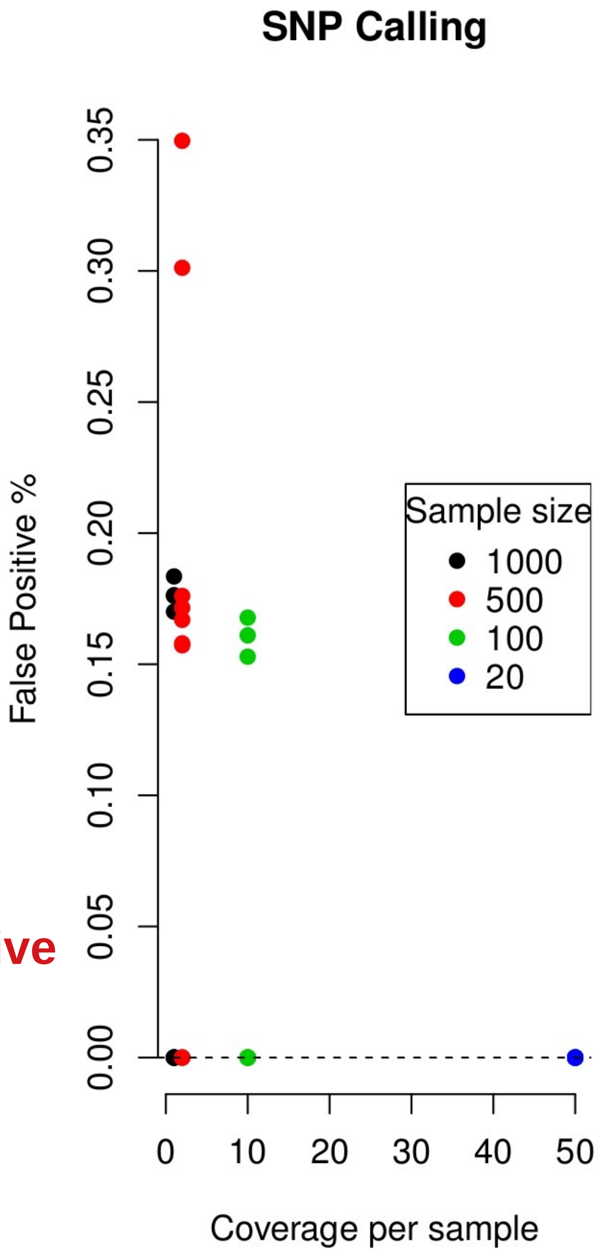
How about the allele frequencies?

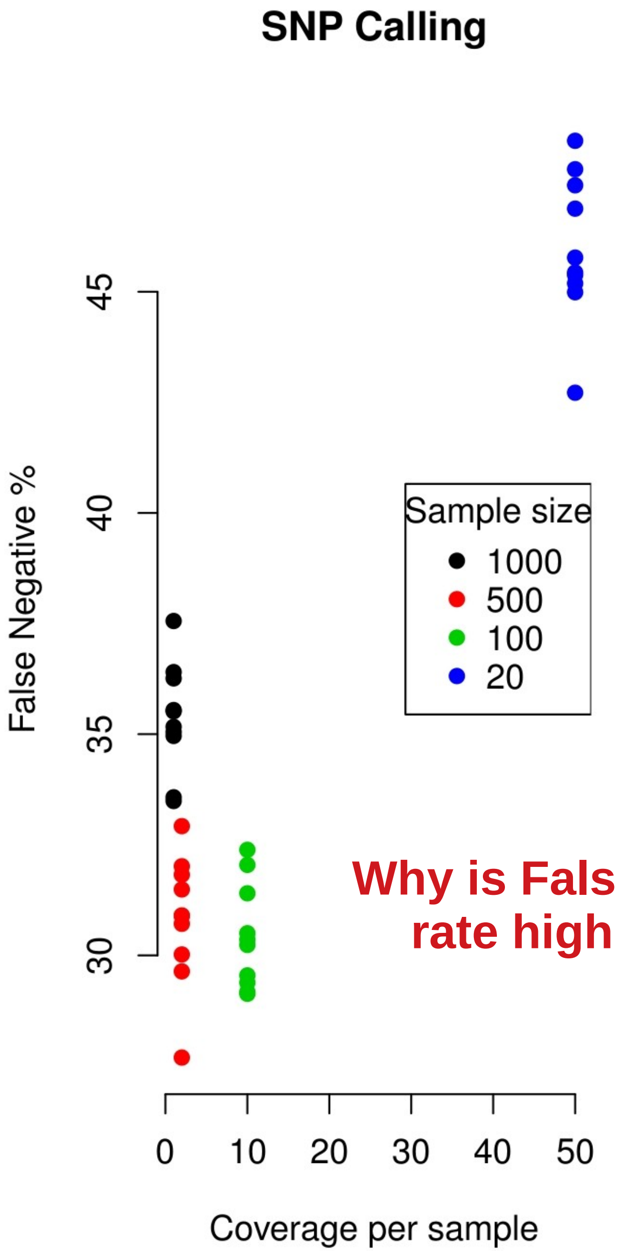


Expected Heterozygosity (Allele frequency)

Why is 2X better?

Do you get the right SNPs?

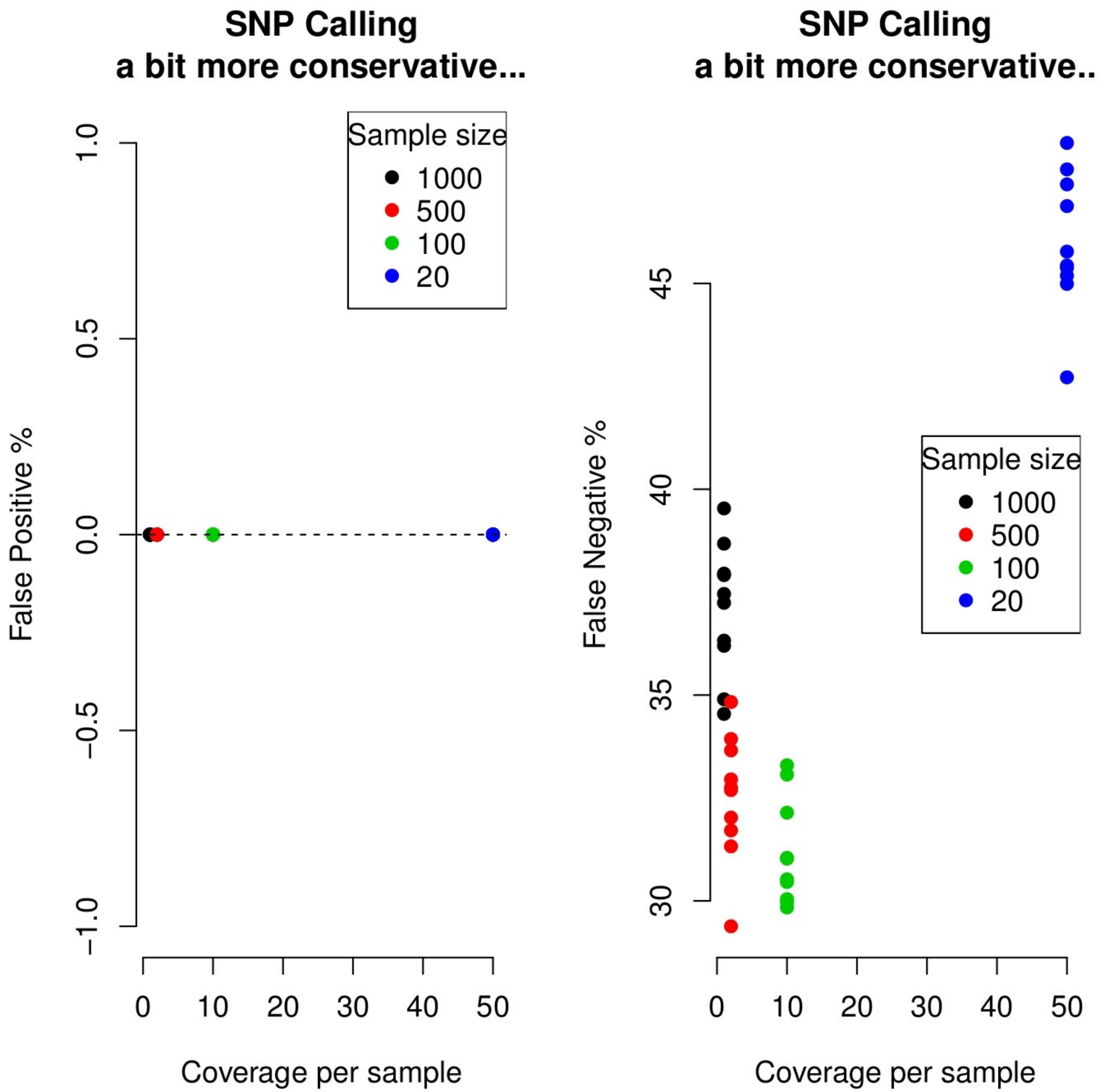




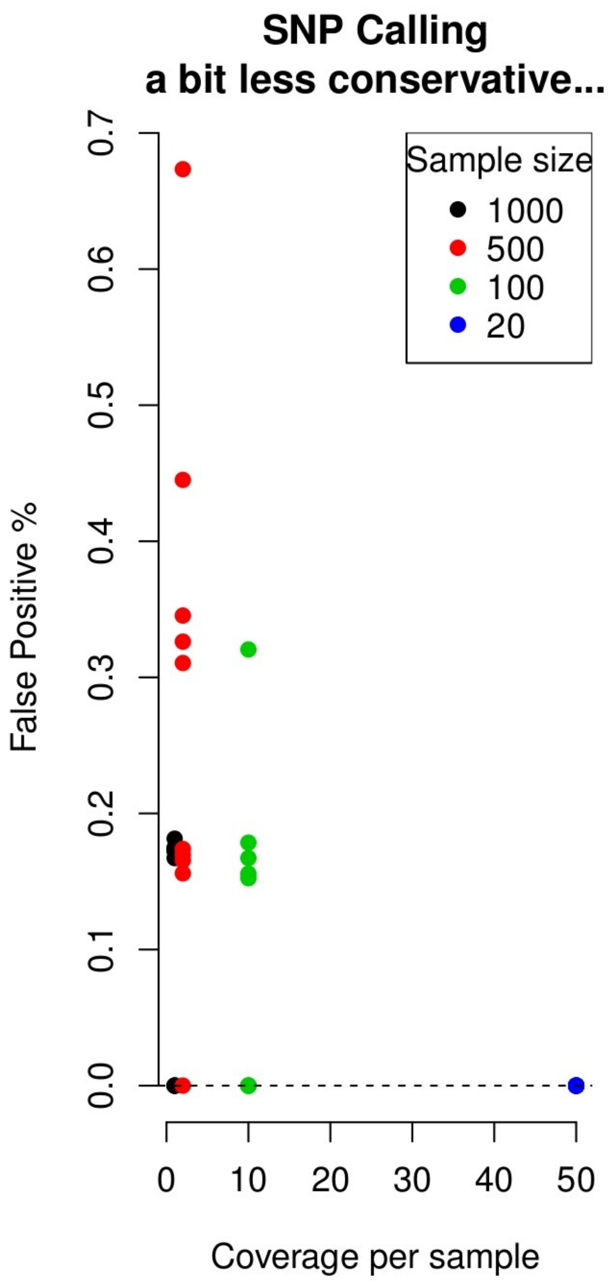
Why is False Positive rate high at low coverage?

Why is False Negative rate high at 50X?

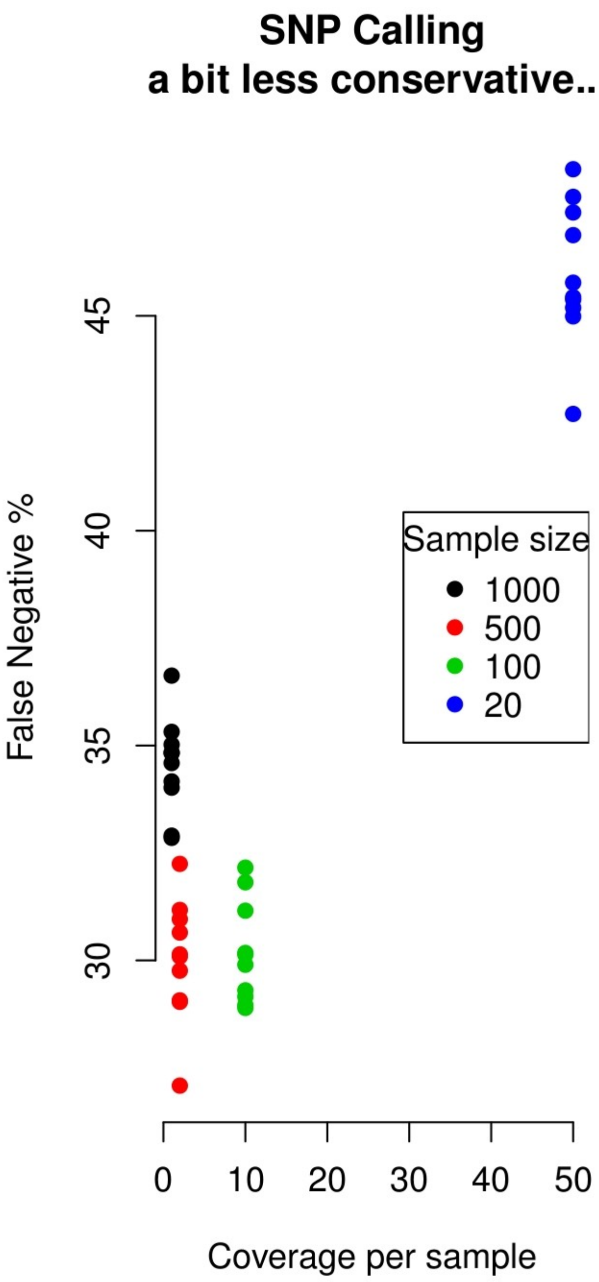
Do you get the right SNPs (second try)?



Do you get the right SNPs (last try)?



SNP Calling
a bit less conservative...



SNP Calling
a bit less conservative..

Conclusions

It is important to take **statistical uncertainty** into account, specially for low coverage samples.

The methods presented provide **tools** for investigating population genetic variation for multiple populations on a large scale.

The great improvement in accuracy for low coverage data can be explained by the fact that we **do not call SNPs or genotypes**.

Acknowledgments

Rasmus Nielsen

UNIVERSITY OF COPENHAGEN

Thorfinn Korneliussen

Anders Albrechtsen

Imperial College London

Matteo Fumagalli

You for the attention!

Software available at:
http://www.popgen.dk/angsd/
https://github.com/fgvieira
https://github.com/mfumagalli/ngsTools