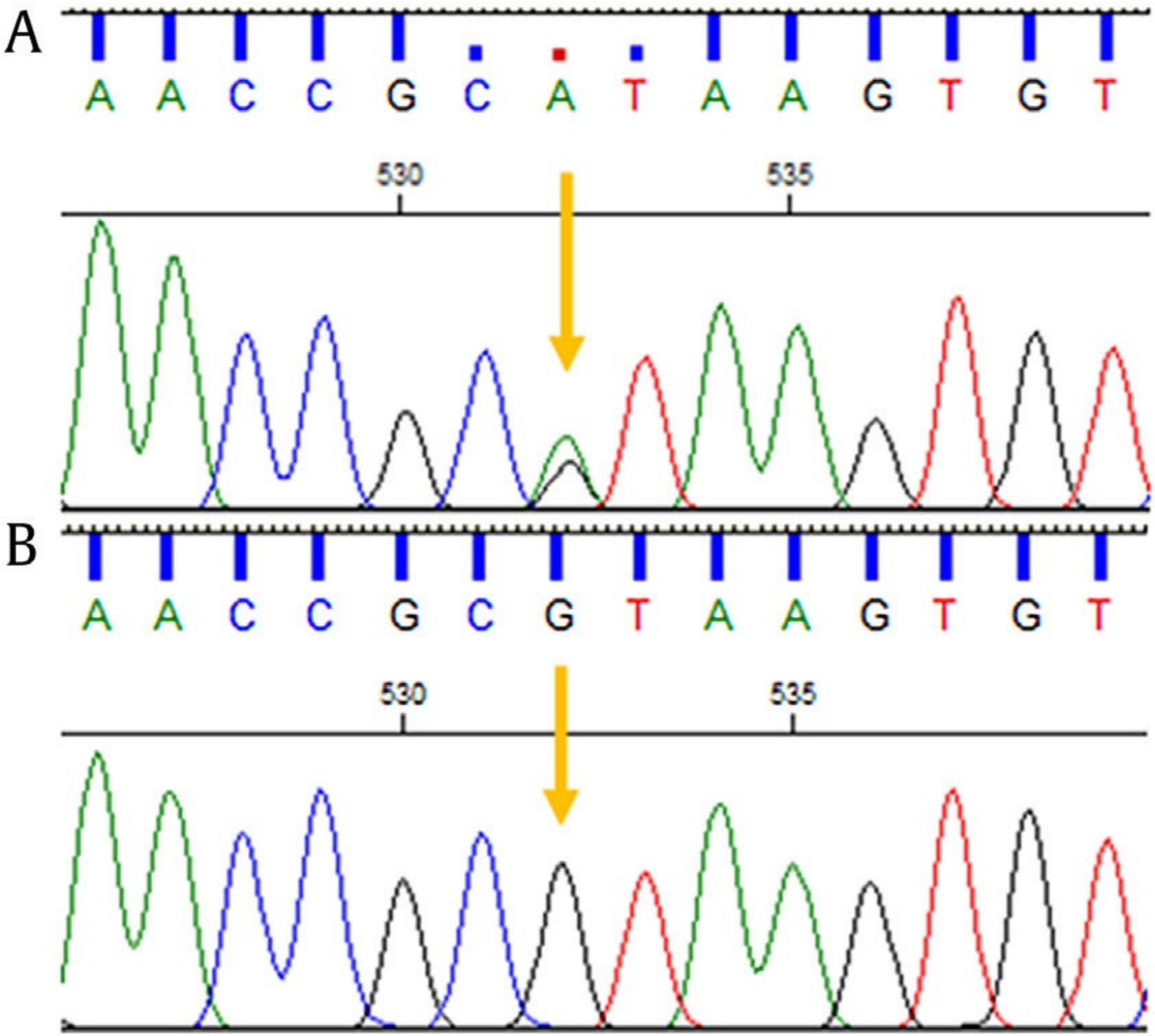
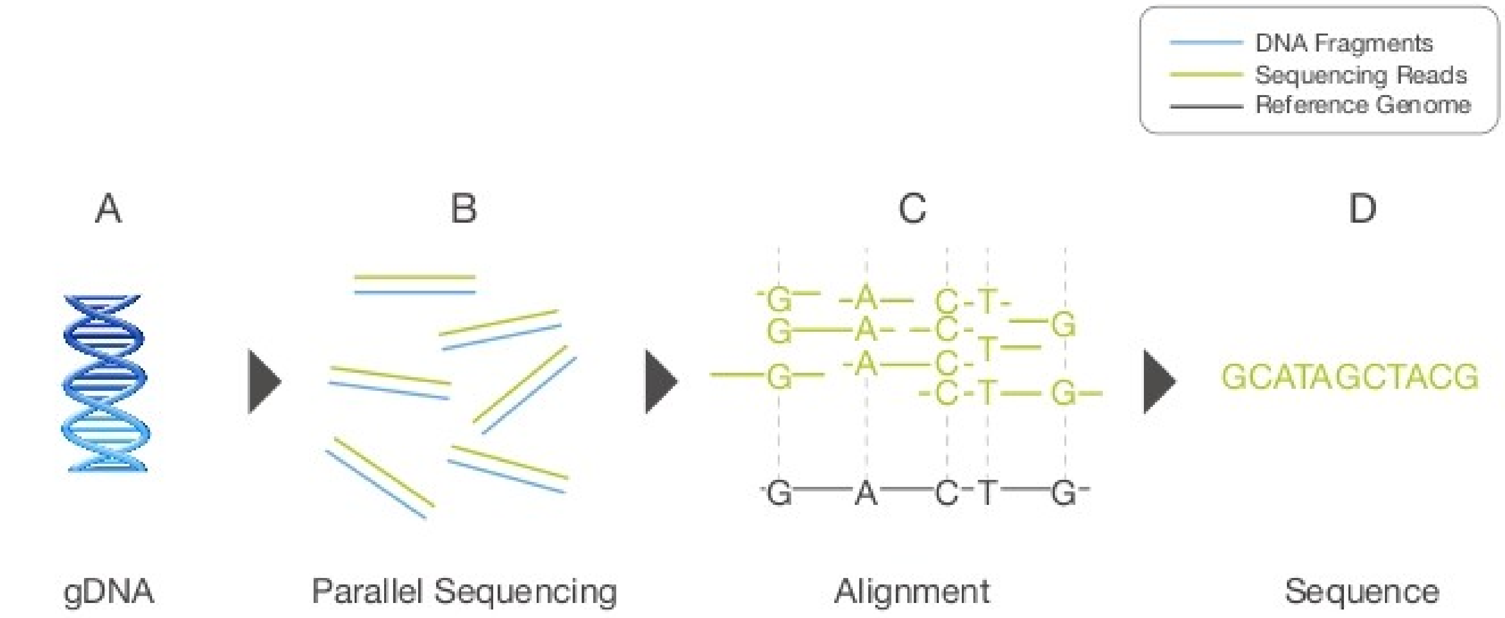
Calling variants from low-coverage NGS data

Filipe G. Vieira
Center for Ancient Environmental Genomics
GLOBE Institute
Copenhagen University
fgvieira@sund.ku.dk

Sanger Sequencing (chromatogram)



Next Generation Sequencing (NGS)



A. Extracted gDNA

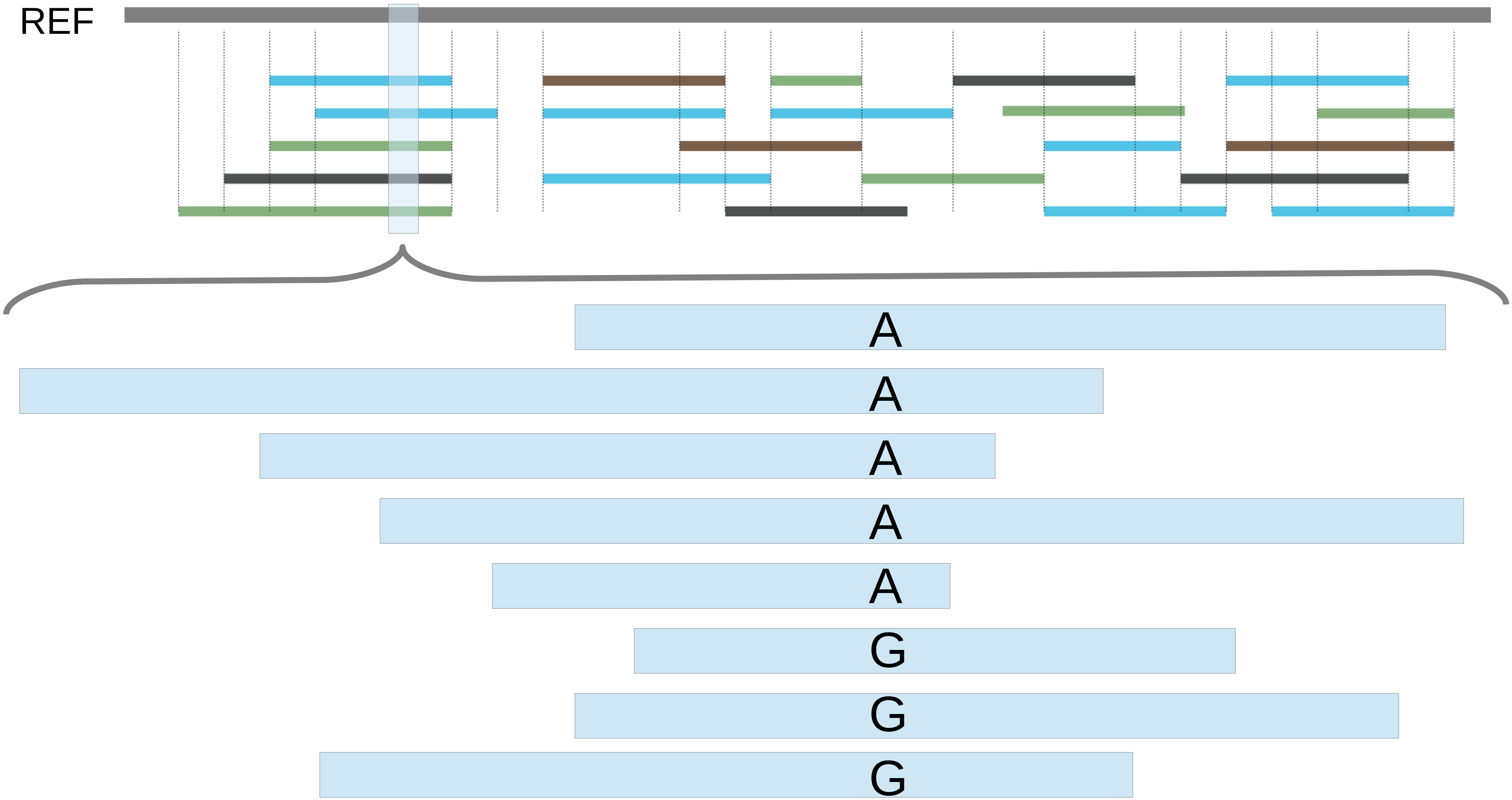
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.

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

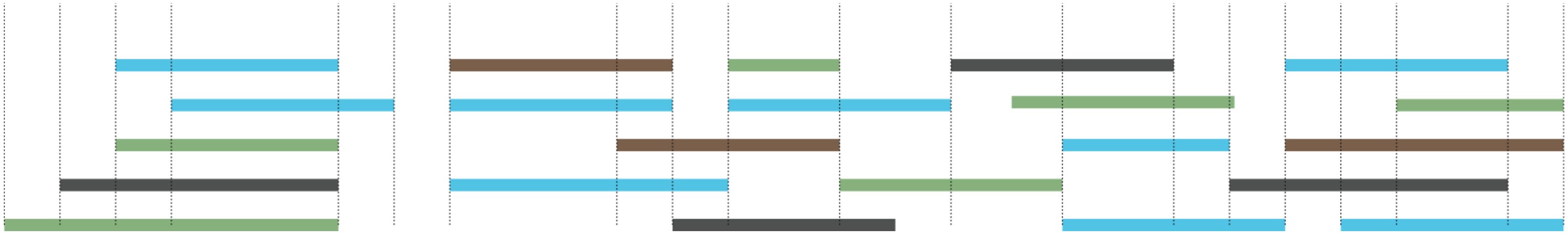
NGS data



A/G (8x)

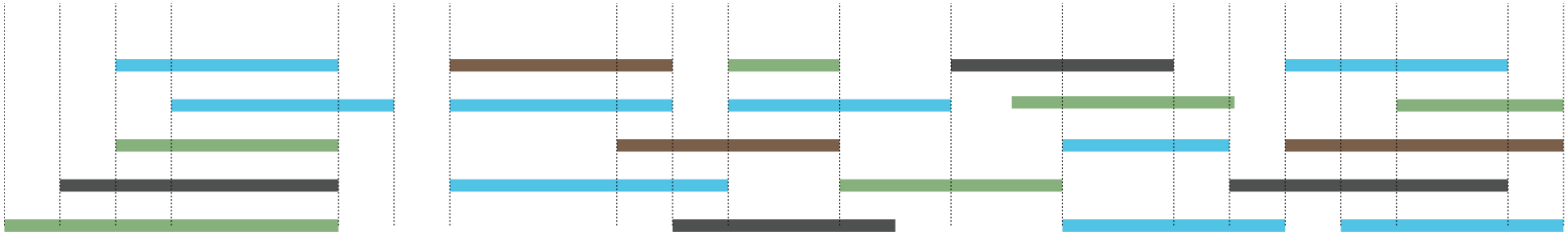
- However NGS has drawbacks:
  - High error rates
  - Heterogeneous sequencing
  - Shorter reads

Is the site variable in the sample?



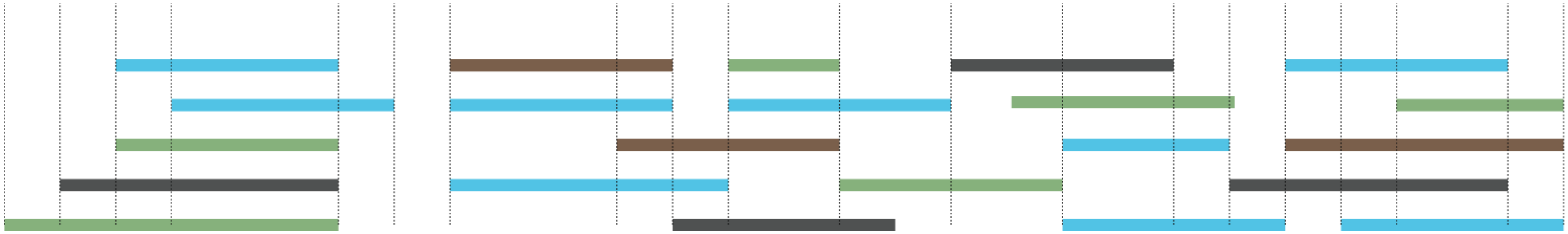
A
A
A
A
A
G
G
G

Is the site still variable in the sample?



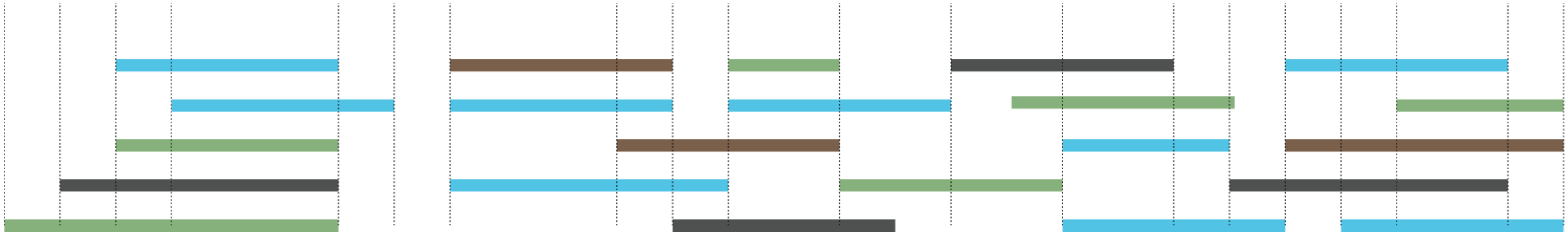
A	A
A	A
A	A
A	A
A	A
G	G
G	G
G	

Is the site still variable in the sample?



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

Is the site still variable in the sample?



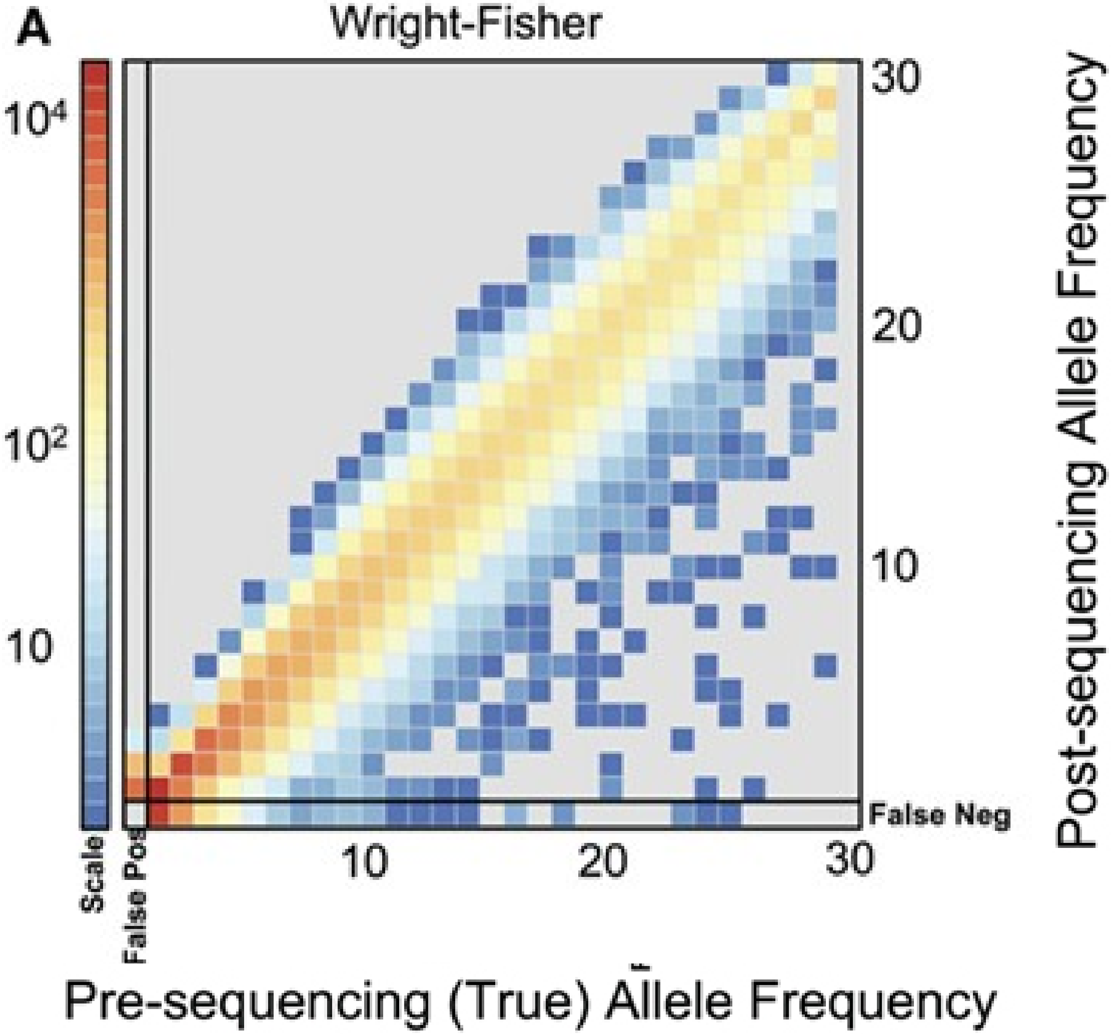
A	A	A
A	A	A
A	A	A
A	A	A
A	A	A
G	G	G
G	G	
G		

Common errors introduced here:

SNP calling: identification of variable sites.

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

Bias in allele frequencies



Crawford and Lazzaro 2012

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 sample sizes at high coverage)
    - Population genetics (large sample sizes)
  - How low can we go?
- How to deal with uncertainty?
  - Stricter filtering → Loss of data
  - Probabilistic framework (genotype likelihoods)
    - Increased analytical power
    - Associated measure of statistical uncertainty
    - Incorporation of **prior** information

Objective

1) What are **genotype likelihoods** (GL)?

2) How to do **SNP calling** from GL?

3) How to do **genotype calling** from GL?

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** (GL)?

2) How to do SNP calling from GL?

3) How to do genotype calling from GL?

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 a particular **genotype**

$$p\left(X \mid G=bh\right)=\frac{1}{2^{r}}\prod_{i=1}^{r}\left(L_{b}^{(i)}+L_{h}^{(i)}\right)$$

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

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

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

Nielsen et al 2012

Priors

- Model organisms
  - Reference genome
  - SNP databases
  - Patterns of LD
  - Known allele or genotype frequencies
  - ...
- Non-model organisms
  - Expected genotype frequencies under a model (e.g. HWE)
    - Works for most case, if population follows HWE
    - Exceptions:
      - Inbreeding (e.g. self-polinatign plans)
      - Asexual reproduction

Objective

1) What are genotype likelihoods (GL)?

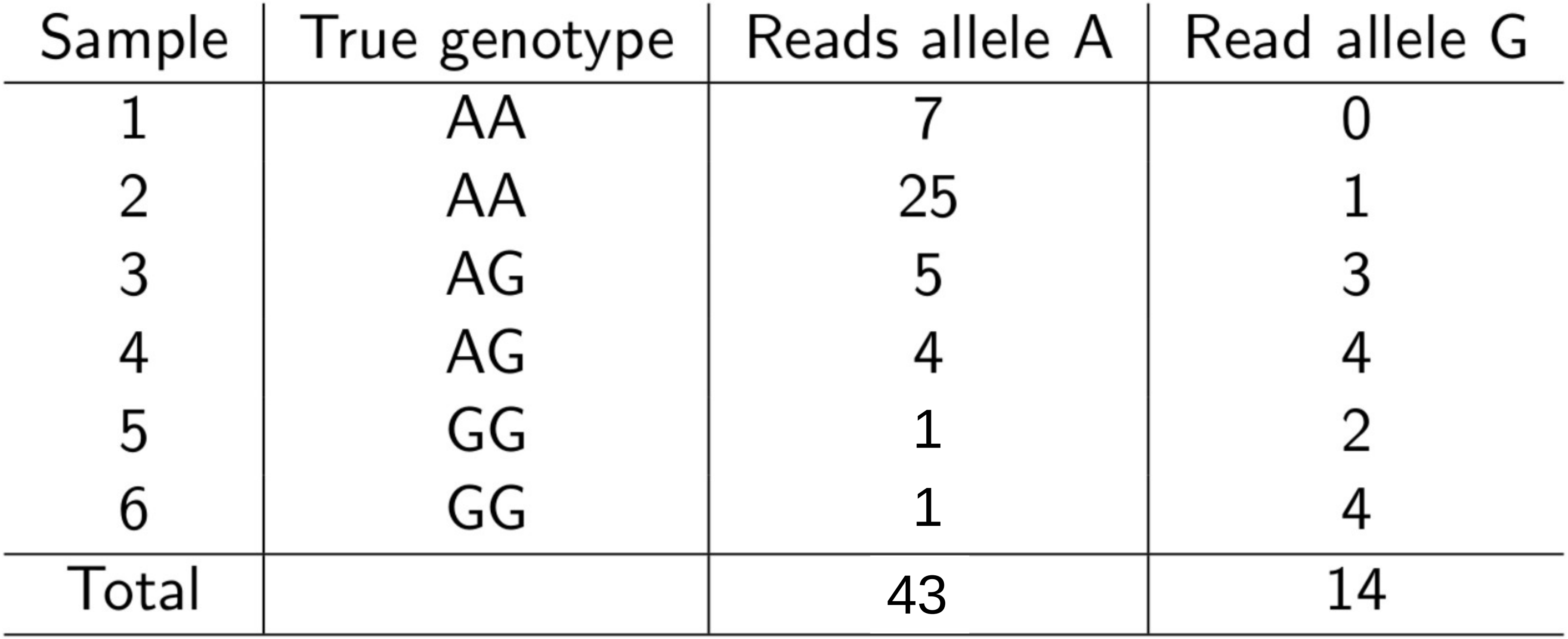
2) How to do **SNP calling** from GL?

3) How to do genotype calling from GL?

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



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	1	2
6	GG	1	4
Total		43	14

What is the true frequency?

What is the estimated frequency?

What is the problem 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
- What is?
  - P(D | G) = P (X | G)
  - P(G = g | f)
- Estimate *f*, by optimizing the likelihood function through (e.g.) EM
  - *f* = 0.46

Priors

- ANGSD uses the minor allele frequency (MAF) to call SNPs
  - Naive:
    - *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 (GL)?

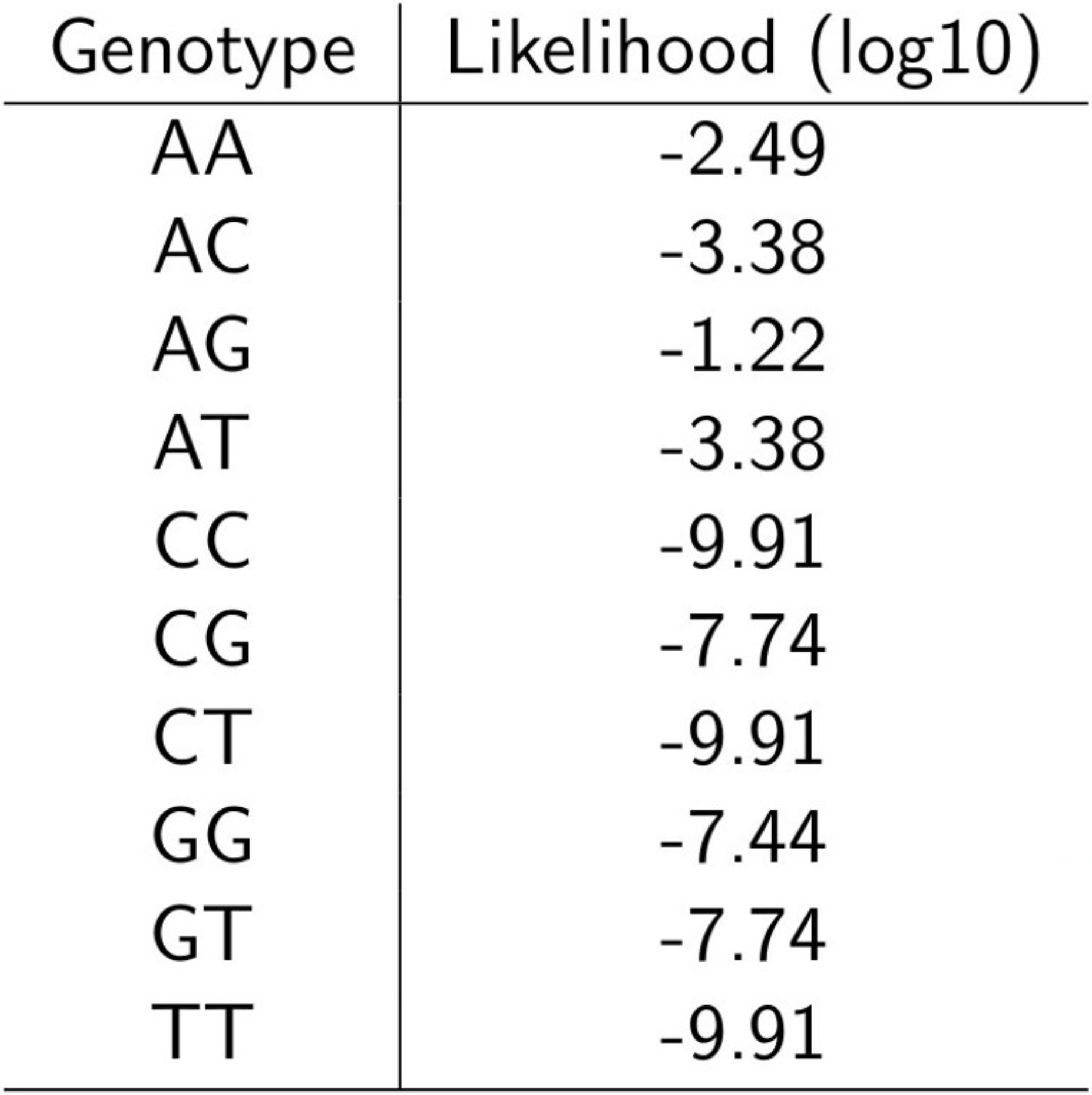
2) How to do SNP calling from GL?

3) How to do **genotype calling** from GL?

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?

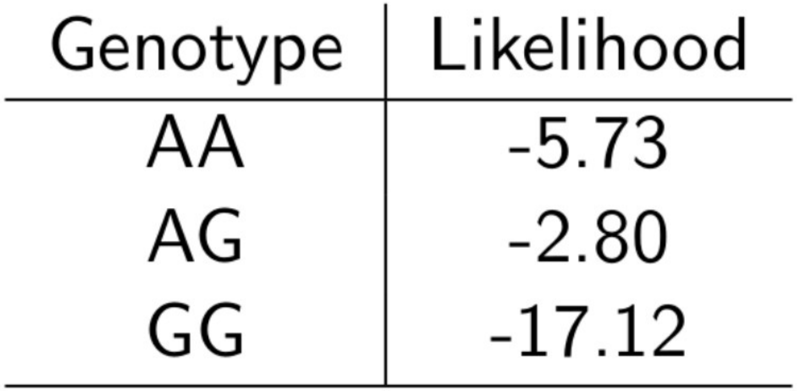
Calling genotypes – 10 GL



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 – 3 GL



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?

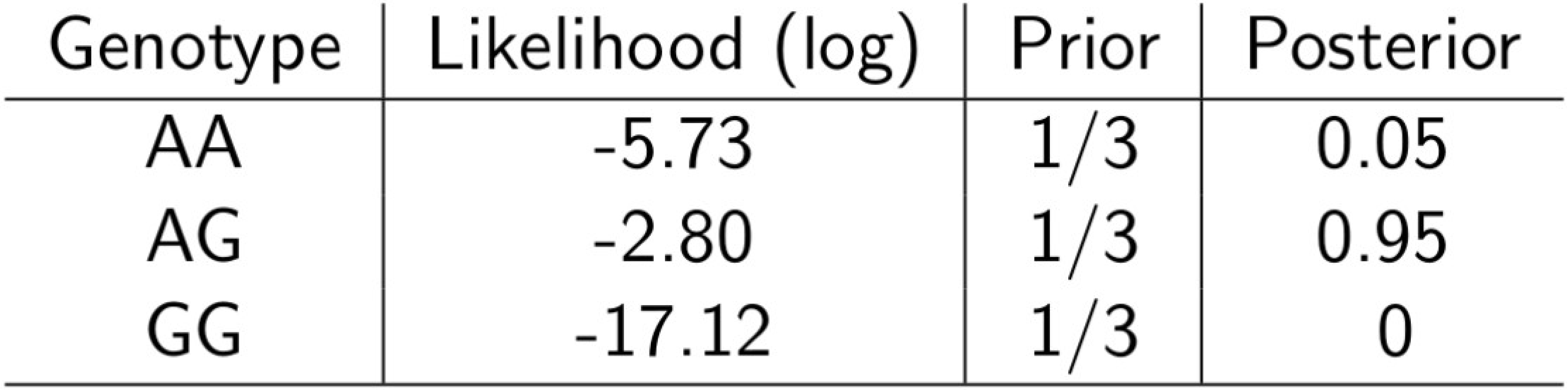
Genotype Quality?

Missing data?

Calling genotypes – Posterior Probabilities (PP)

AAAG (A,G alleles)

$\varepsilon = 0.01$



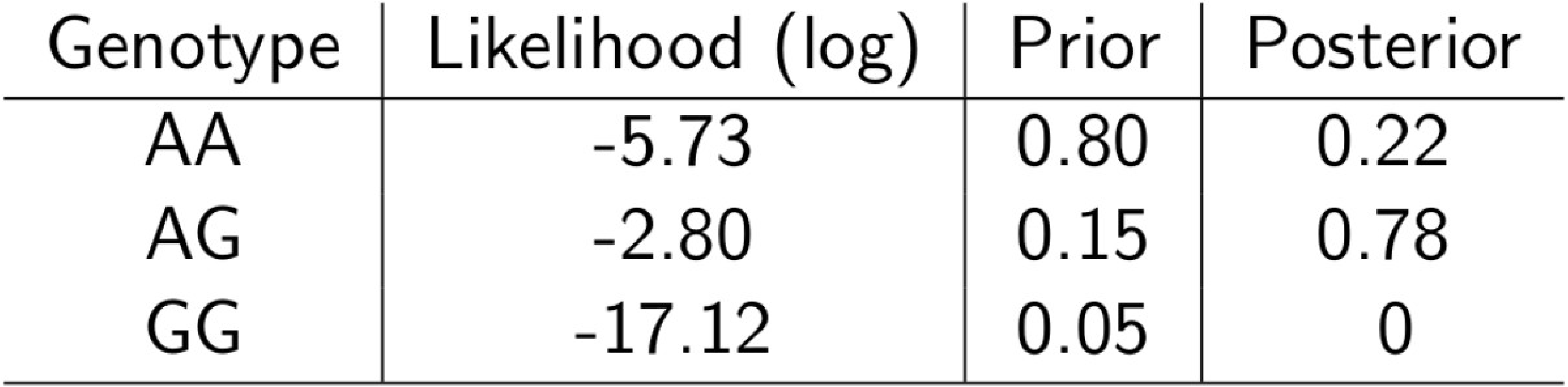
Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	1/3	0.05
AG	-2.80	1/3	0.95
GG	-17.12	1/3	0

Calling genotypes – PP (reference prior)

AAAG (A,G alleles)

ε = 0.01

A is reference → 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

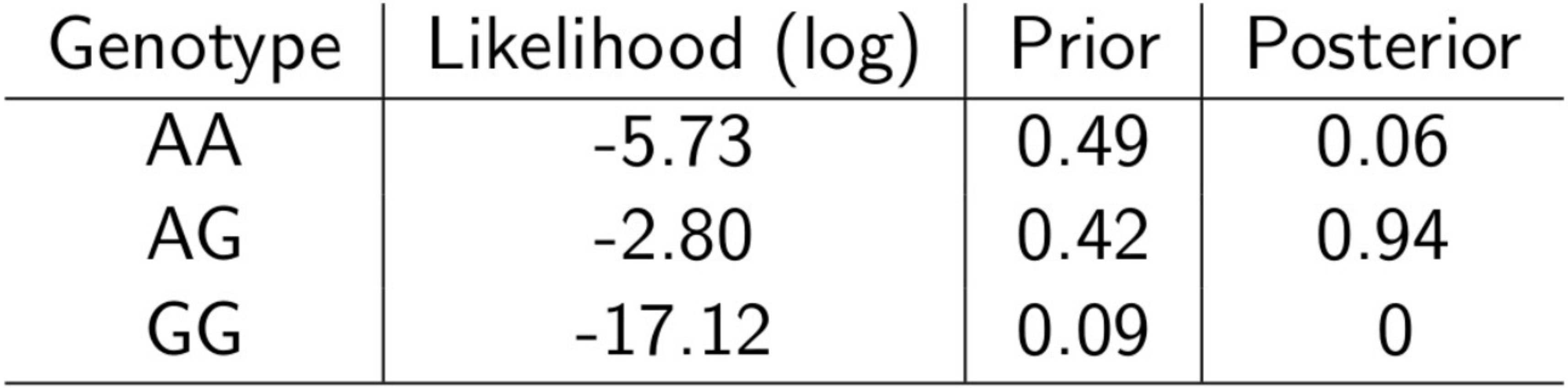
Calling genotypes – PP (HWE prior)

AAAG (A,G alleles)

ε = 0.01

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?

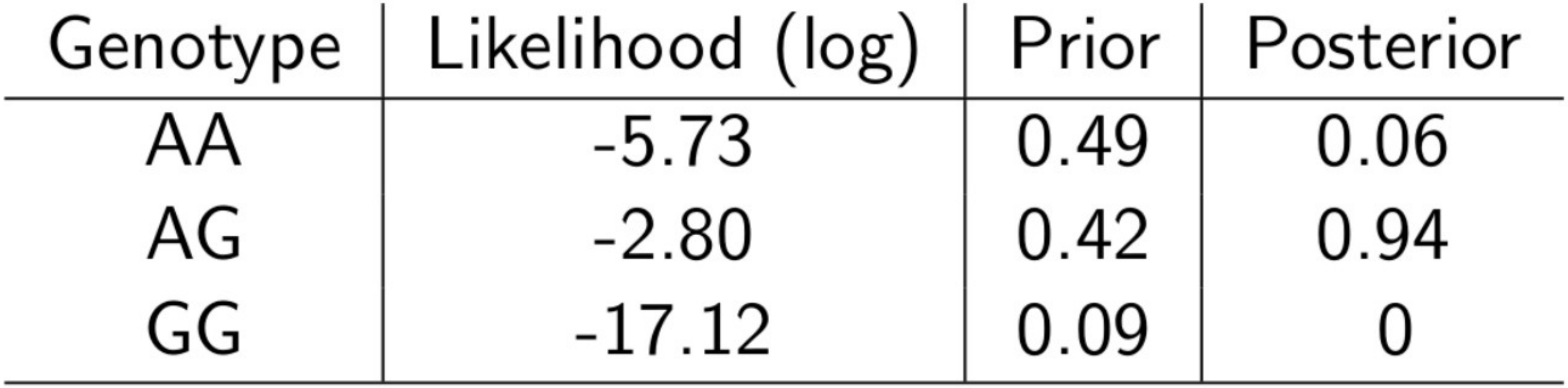
Calling genotypes – PP (HWE prior)

AAAG (A,G alleles)

$\varepsilon = 0.01$

f(a) = 0.7 (from the data itself)

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 estimate frequencies accurately?

Can we assume HWE?

Objective

1) What are genotype likelihoods (GL)?

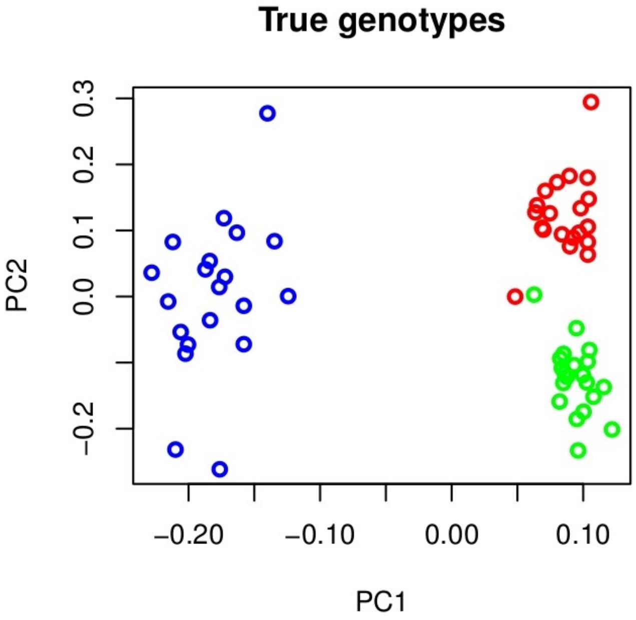
2) How to do SNP calling from GL?

3) How to do genotype calling from GL?

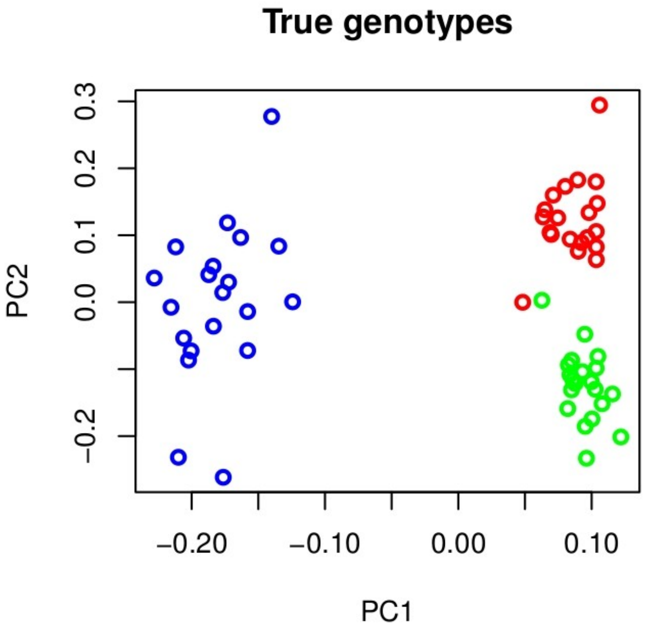
4) What is the **error** in population genetic inferences using naïve strategies for **SNP and genotype calling**?

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

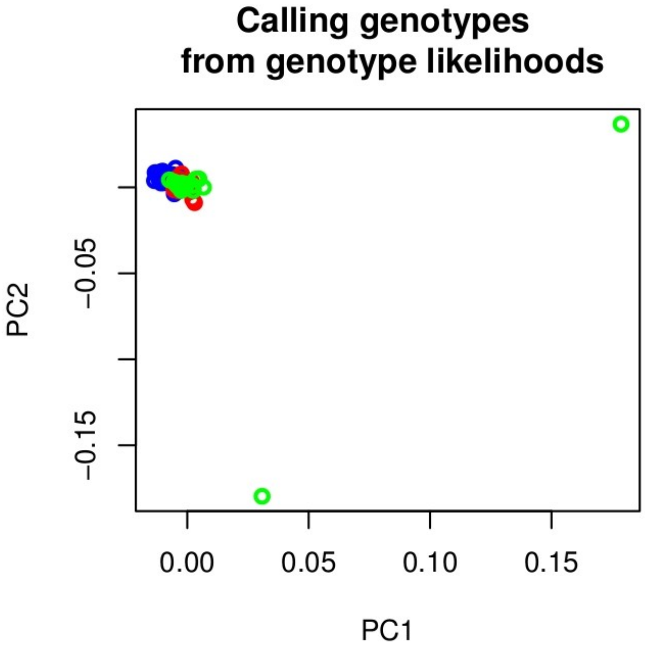
Population structure - PCA



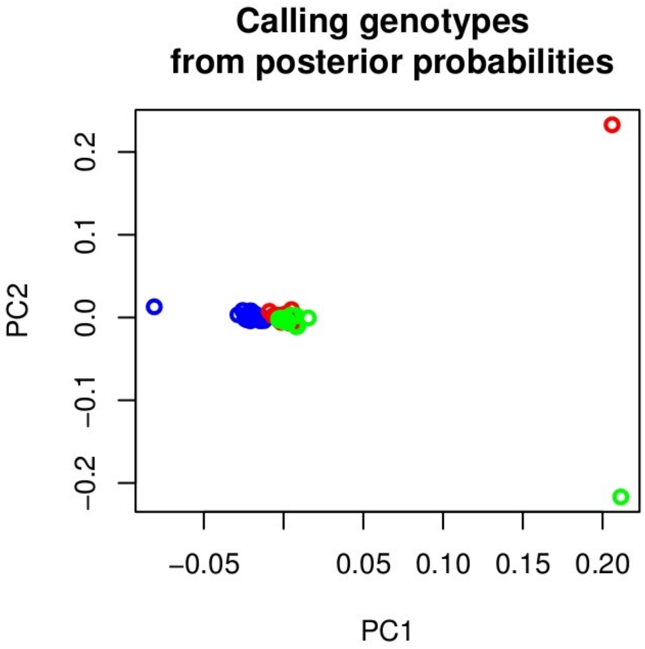
Population structure - PCA



True genotypes

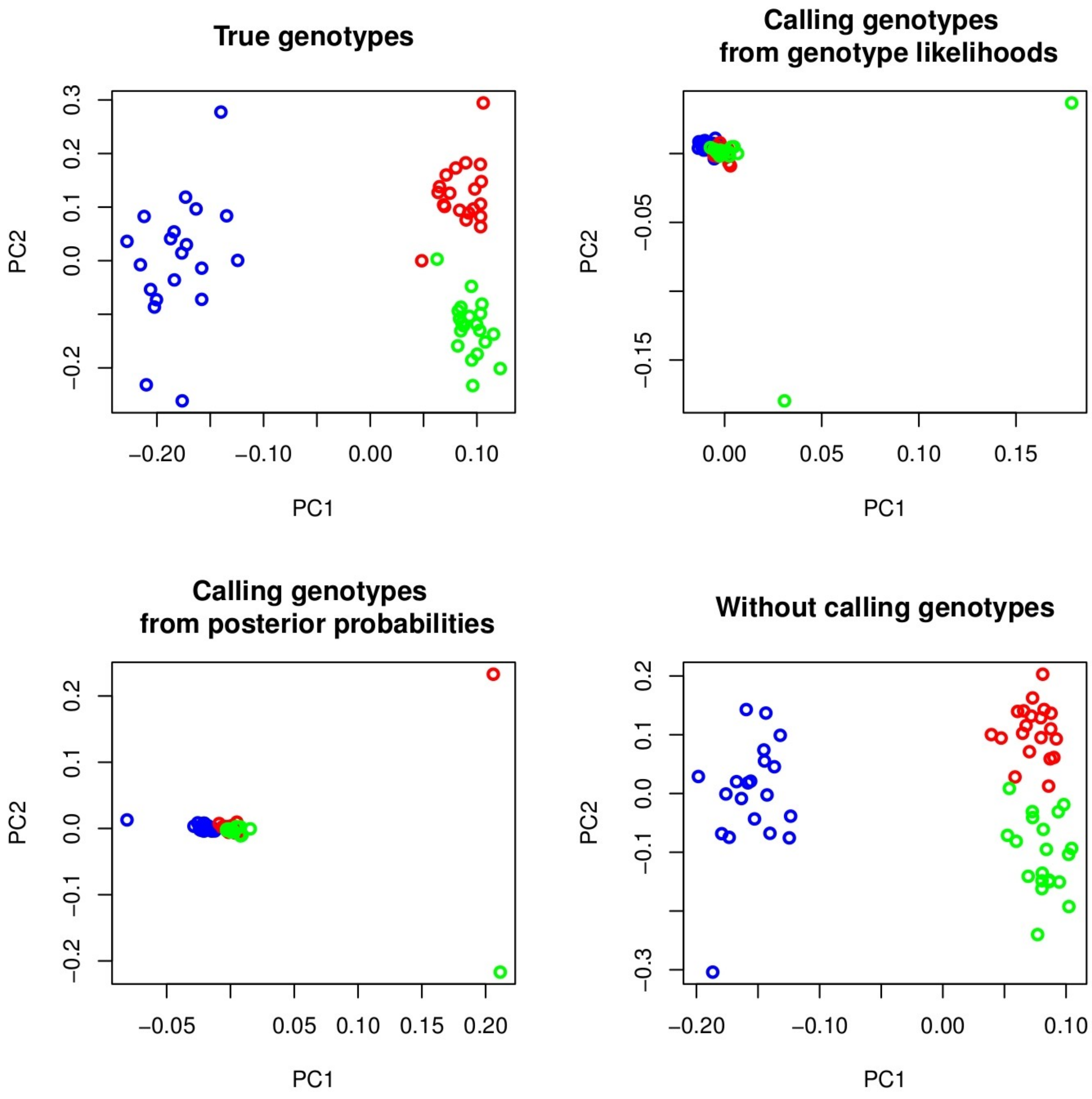


Calling genotypes
from genotype likelihoods



Calling genotypes
from posterior probabilities

Population structure - PCA



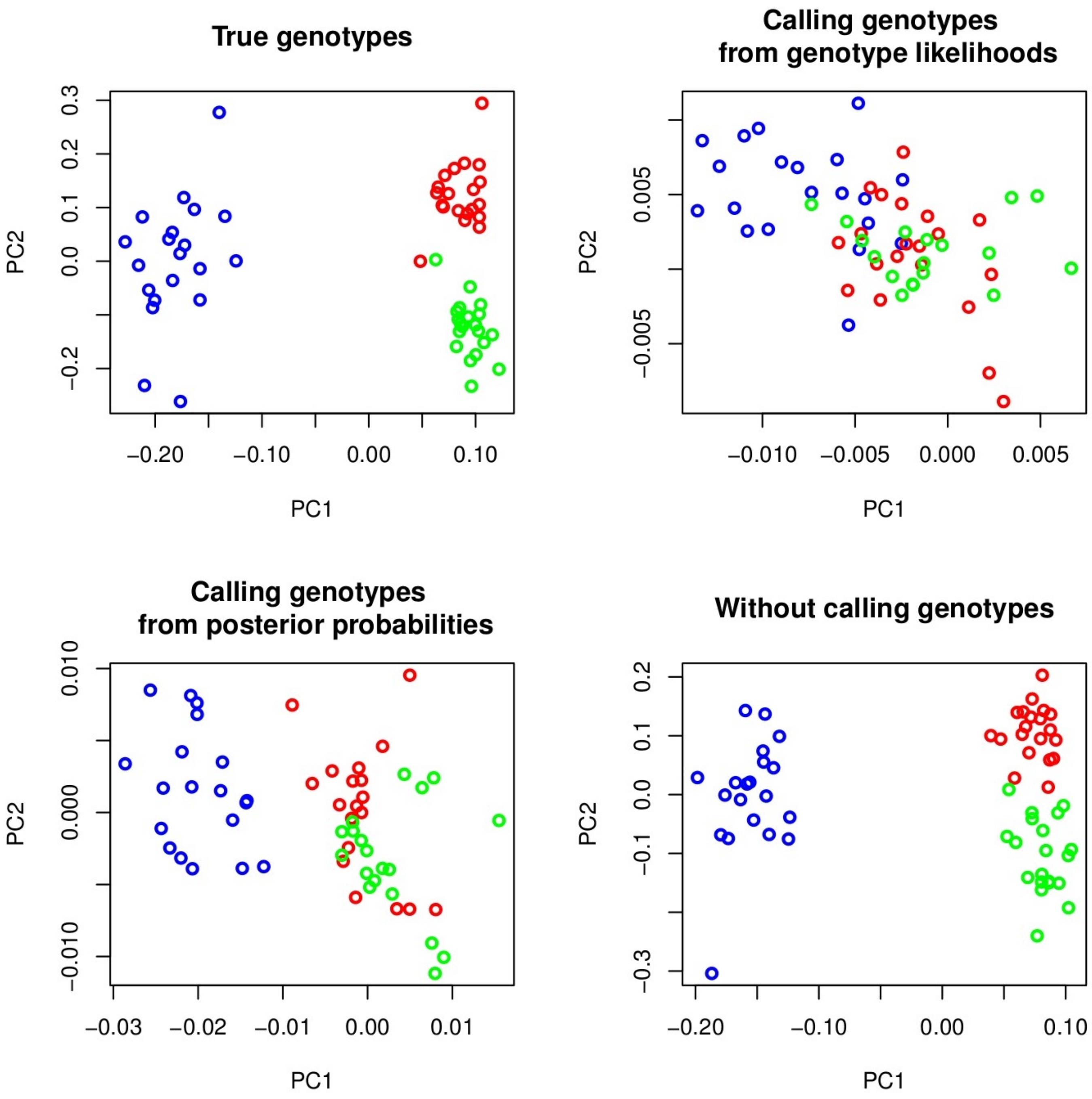
True genotypes

Calling genotypes from genotype likelihoods

Calling genotypes from posterior probabilities

Without calling genotypes

Population structure – PCA (no outliers)



Objective

1) What are genotype likelihoods (GL)?

2) How to do SNP calling from GL?

3) How to do genotype calling from GL?

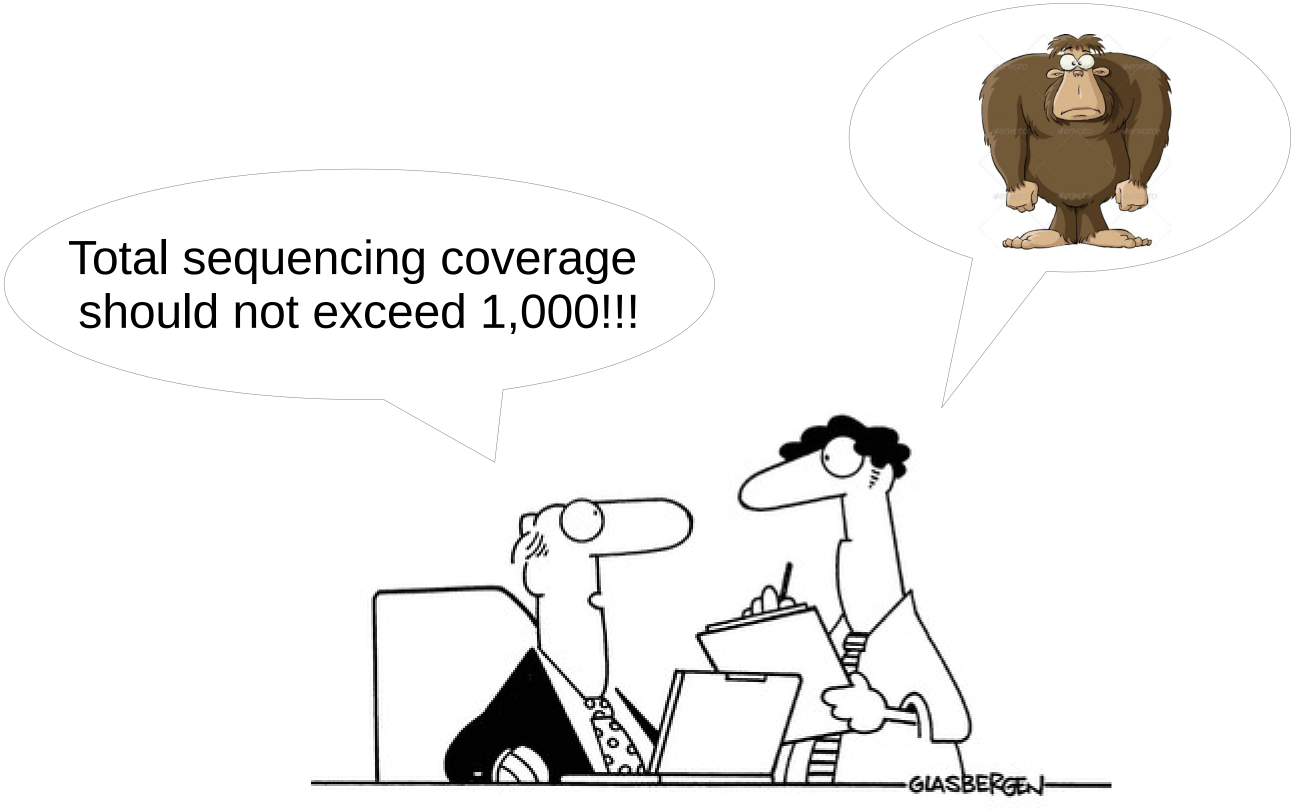
4) What is the error in population genetic inferences using naïve strategies for SNP and genotype calling?

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

Discovery of a “new” species/population

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

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

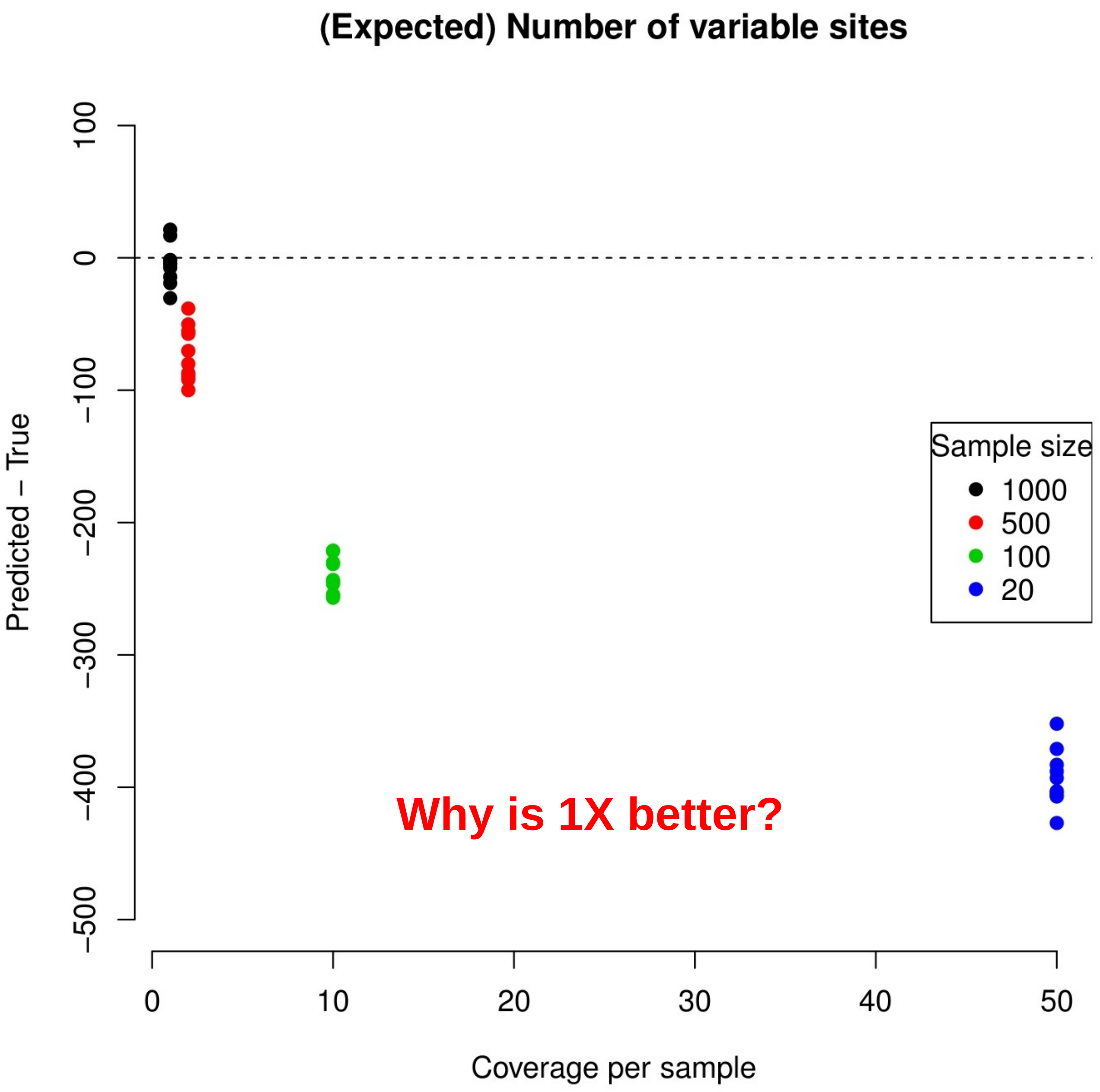


Planning the experiment

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

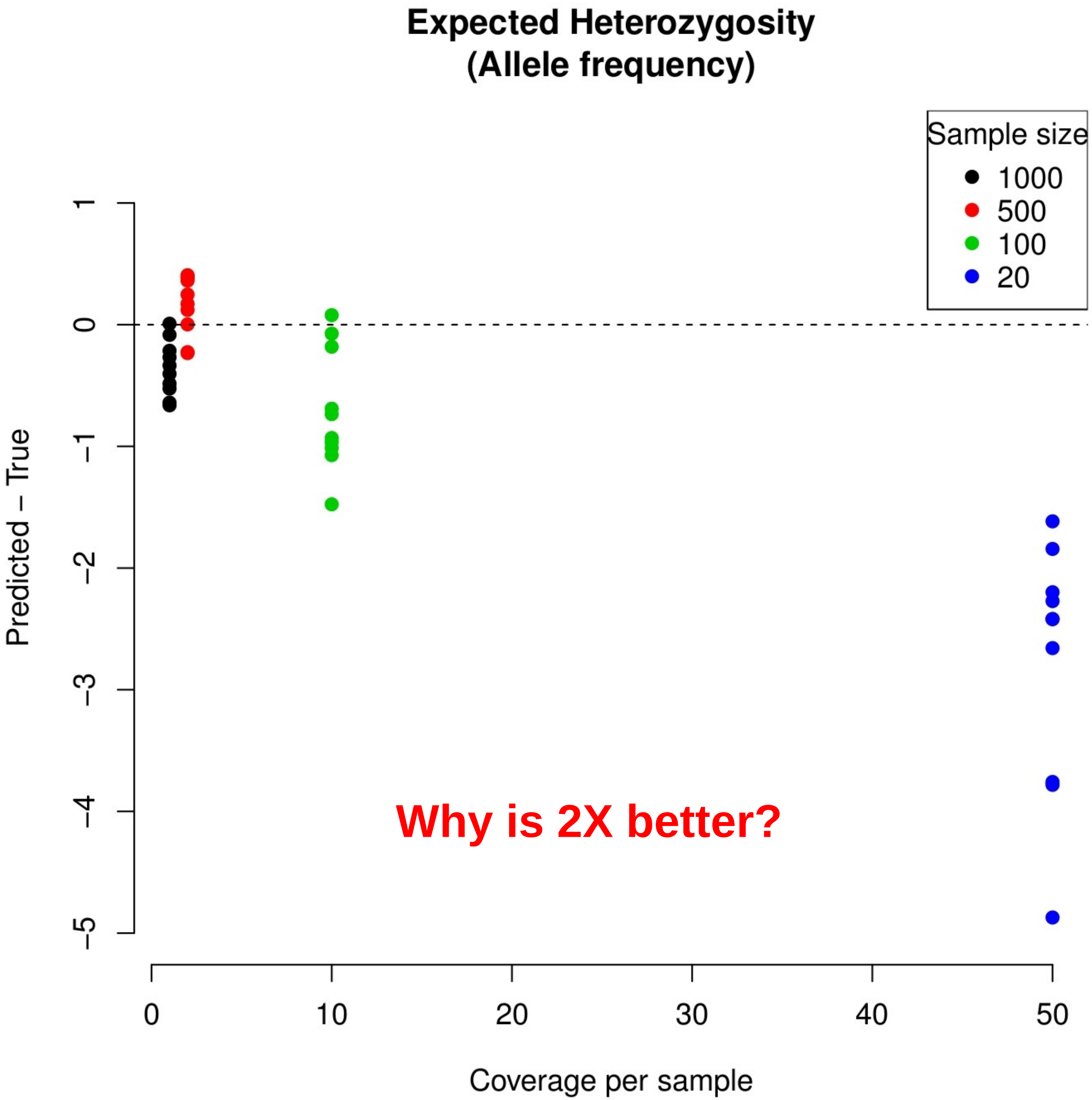
Genome is **100,000 bp** long.

How many polymorphic sites?



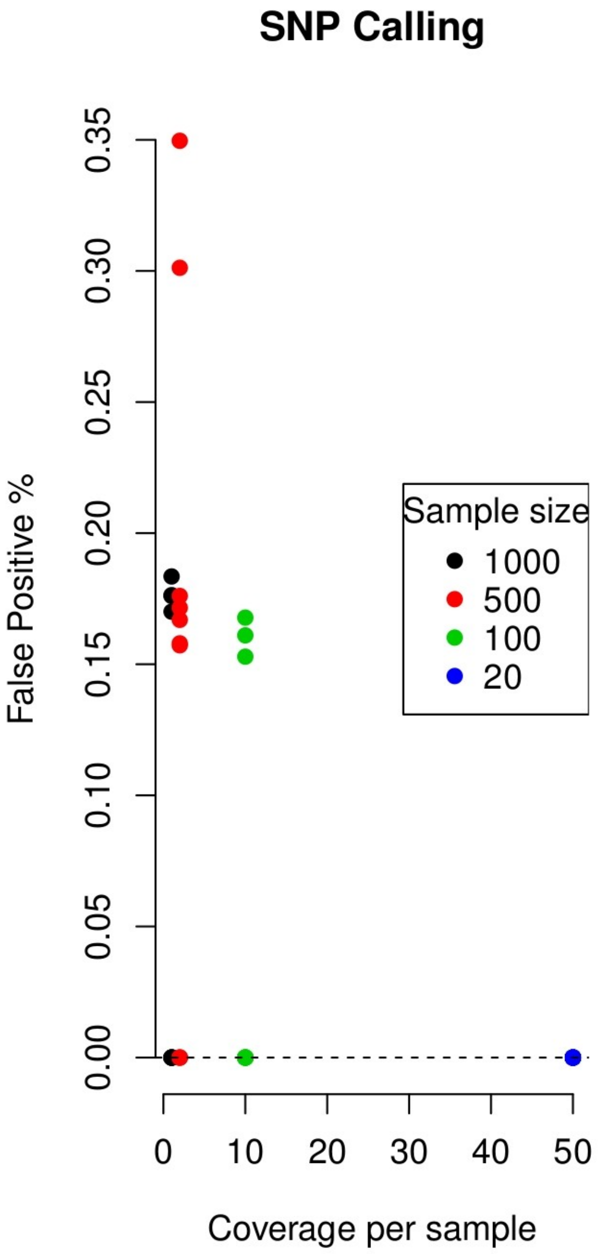
Why is 1X better?

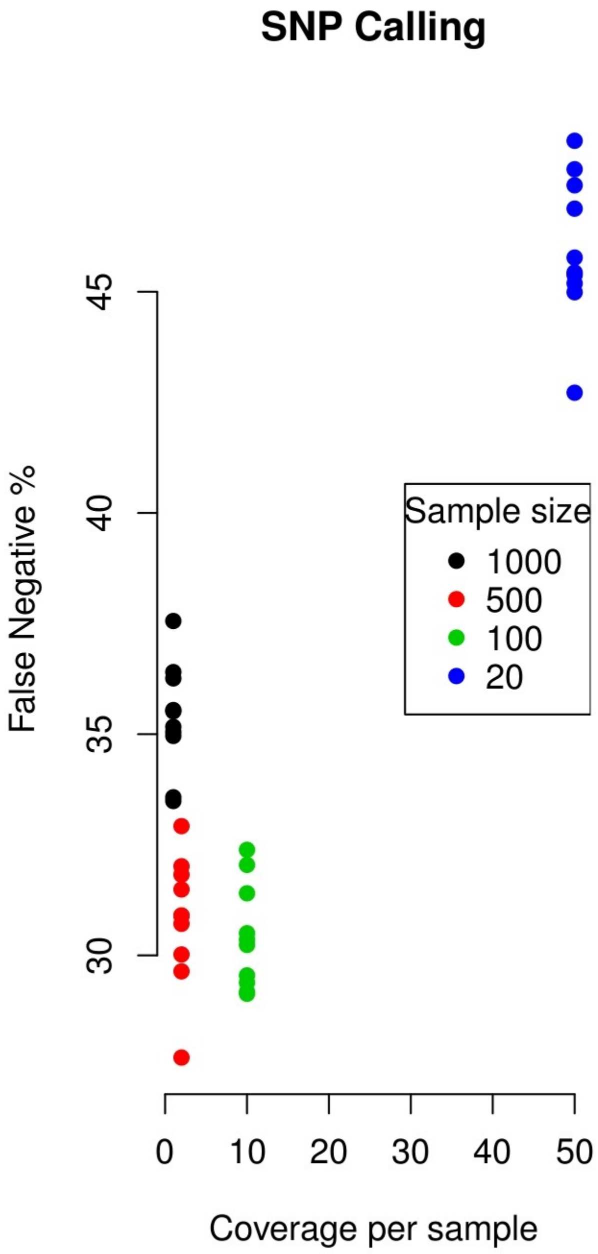
How about the allele frequencies?



Why is 2X better?

Do you get the right SNPs?

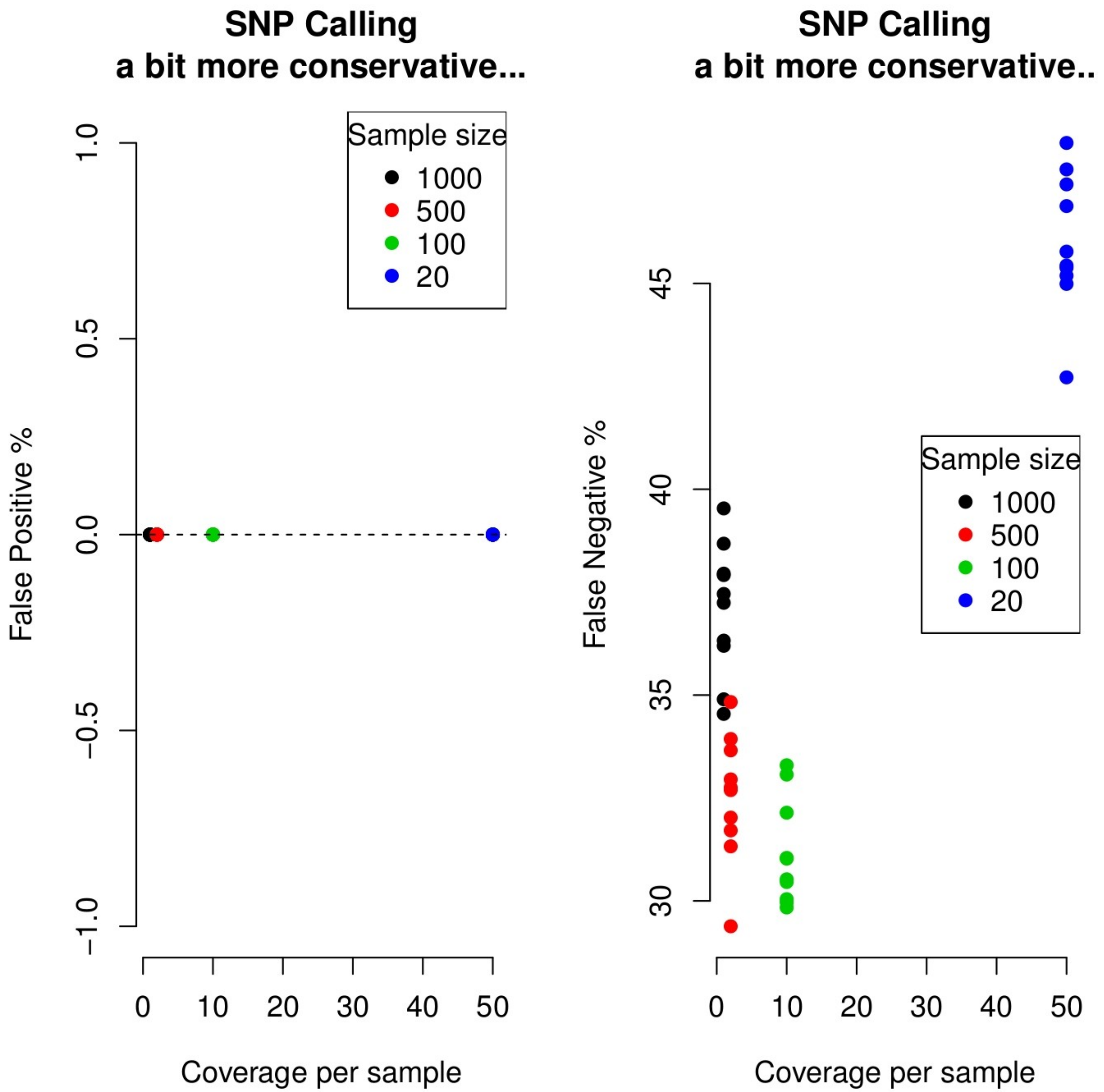




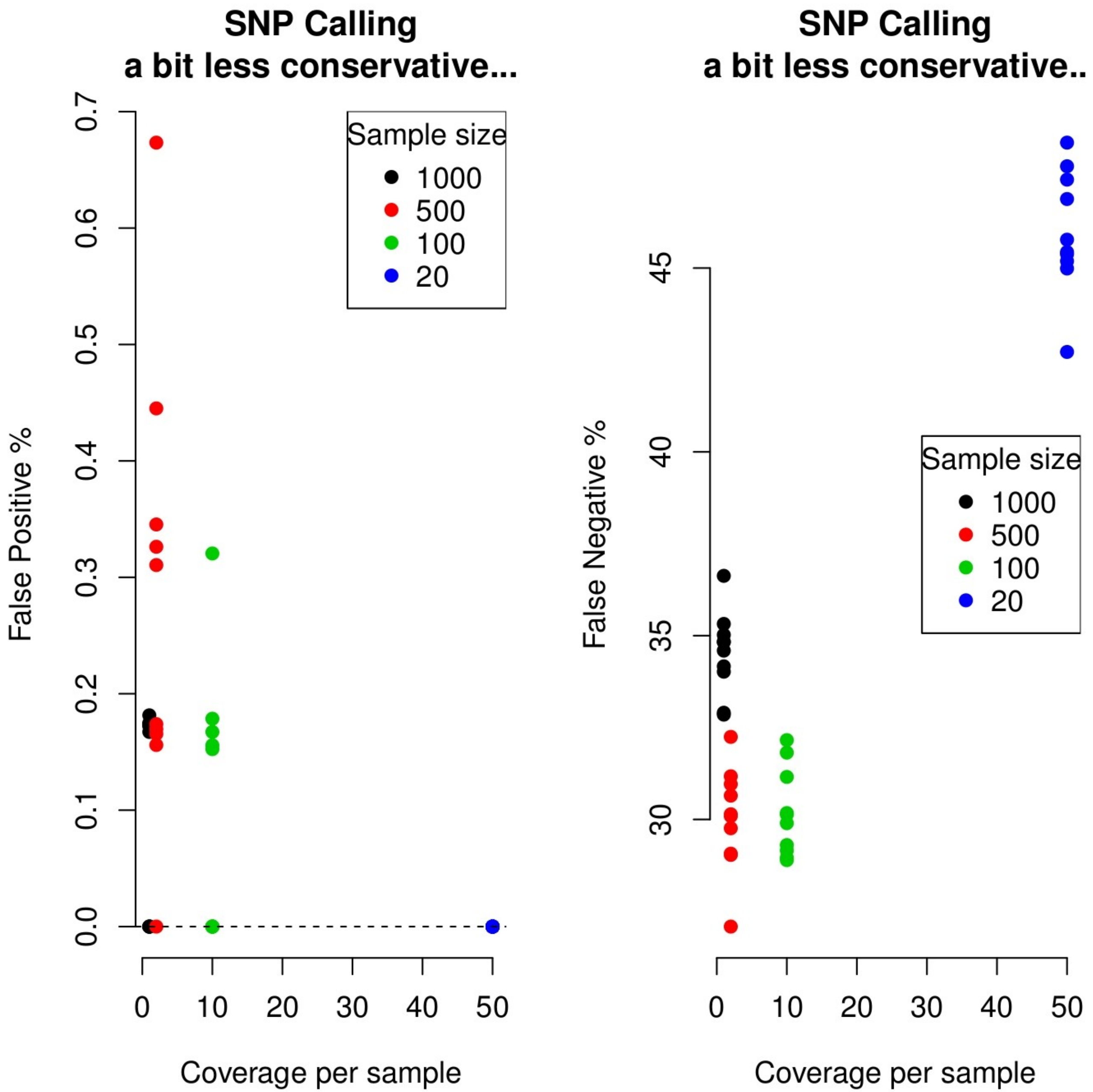
Why is FN rate high at high coverage?

Why is FP rate high at low coverage?

Do you get the right SNPs (more strict)?



Do you get the right SNPs (less strict)?



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

Queen Mary University of London

Matteo Fumagalli

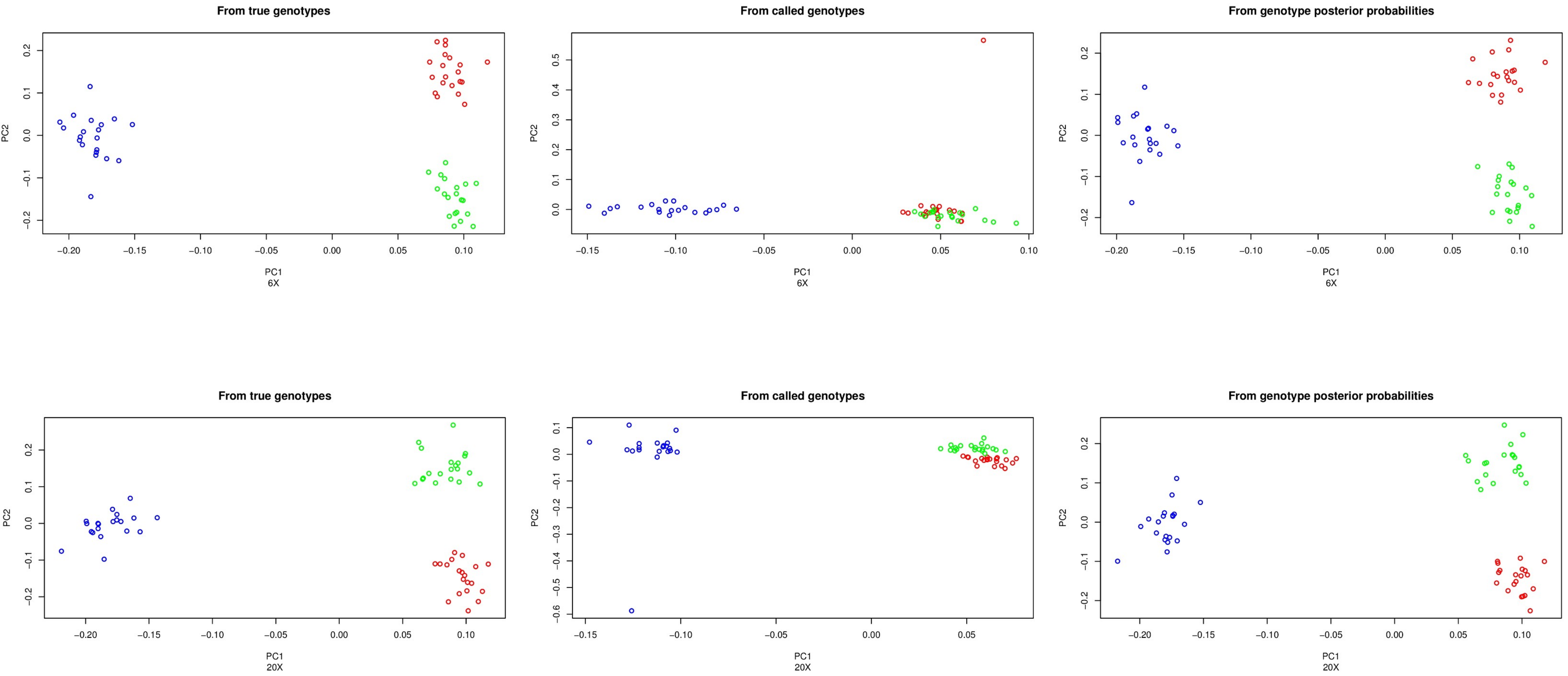
Software available at:

http://popgen.dk/software/angsd.html

https://github.com/fgvieira

https://github.com/mfumagalli/ngsTools

Performance of PCA


From true genotypes
From called genotypes
From genotype posterior probabilities
PC1
PC2
6X
From true genotypes
From called genotypes
From genotype posterior probabilities
PC1
PC2
20X