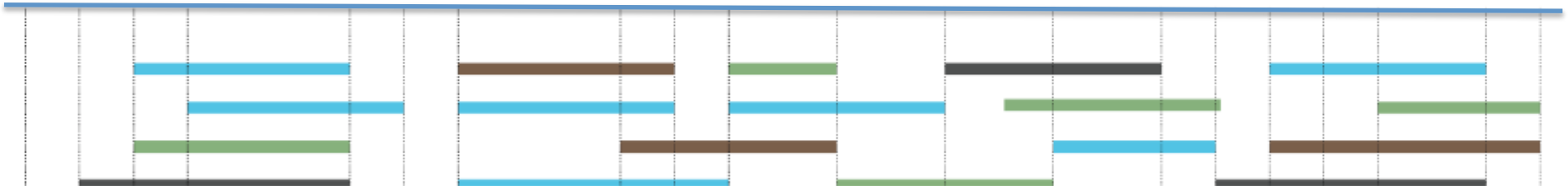
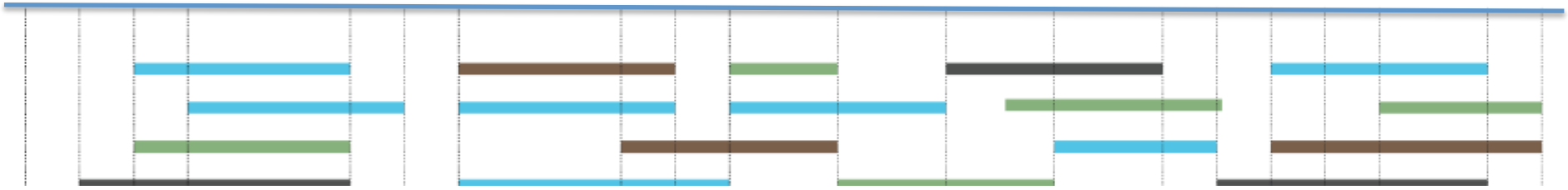
Population genetics from missing or low-depth data

Matteo Fumagalli

Challenges

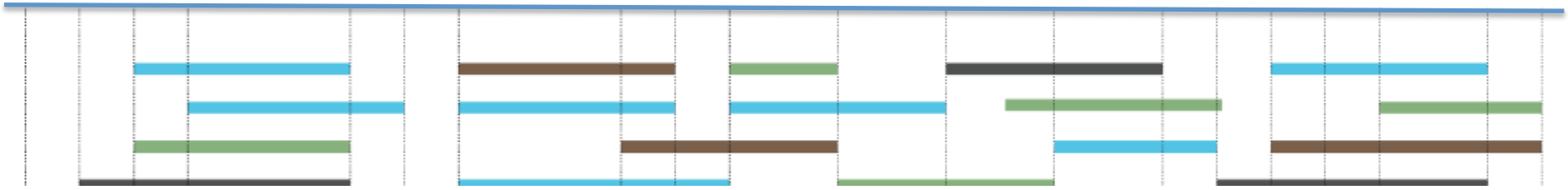


Challenges



- Variable and low depth
- High sequencing and mapping errors

Challenges

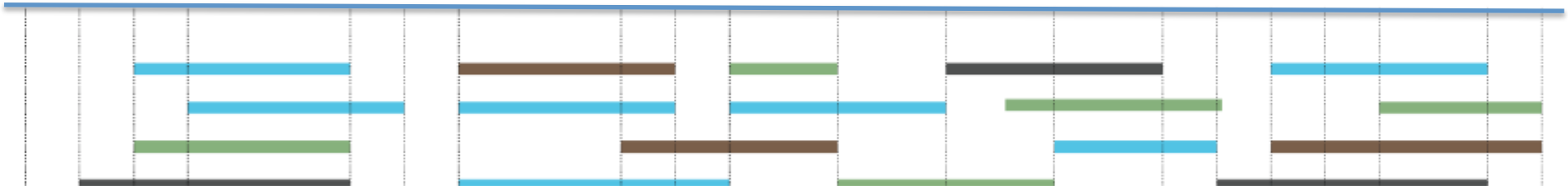


- Variable and low depth

- High sequencing and mapping errors

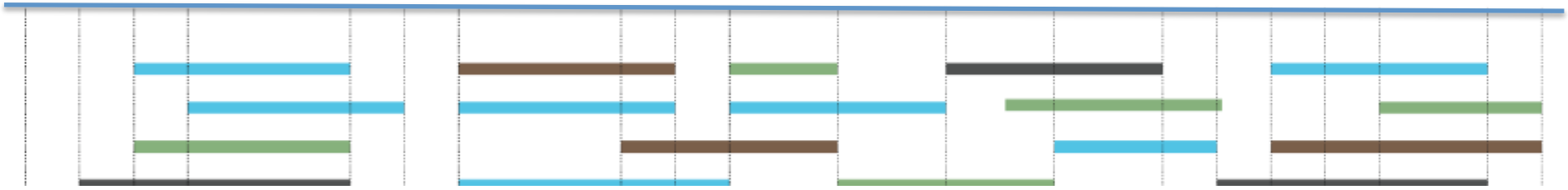
Quality control filters

Data filtering



- Variable and low depth

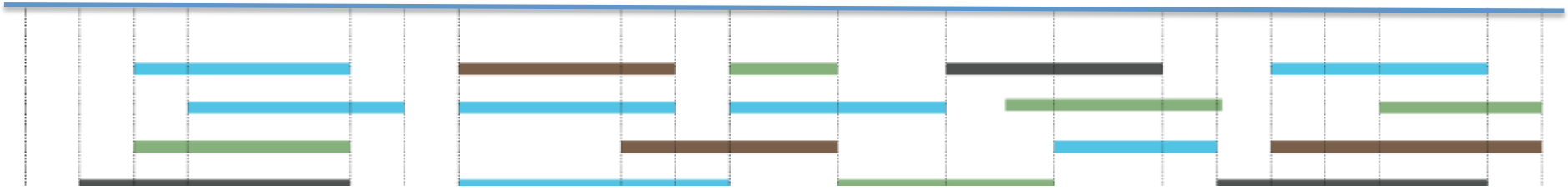
Data filtering



- Variable and low depth

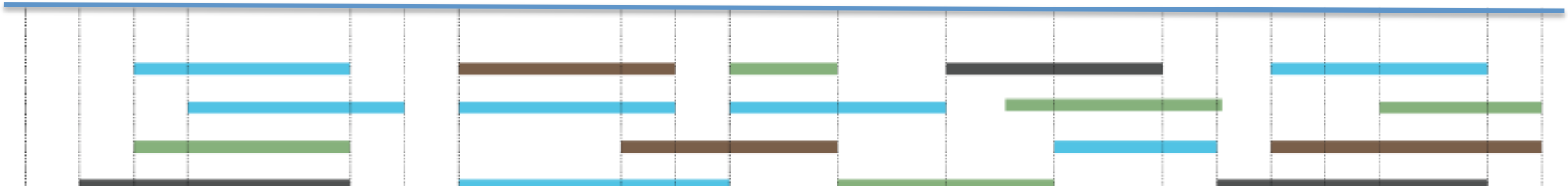
Minimum depth
Maximum depth
Even depth across samples
…

Data filtering



- Sequencing and mapping errors

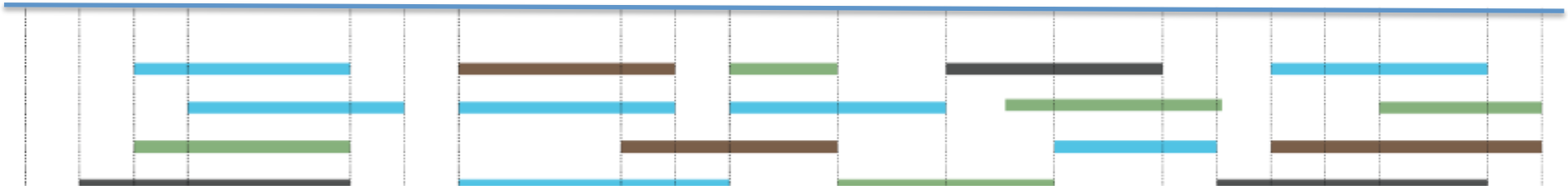
Data filtering



- Sequencing and mapping errors

Minimum base quality
Minimum mapping quality
Base quality bias
…

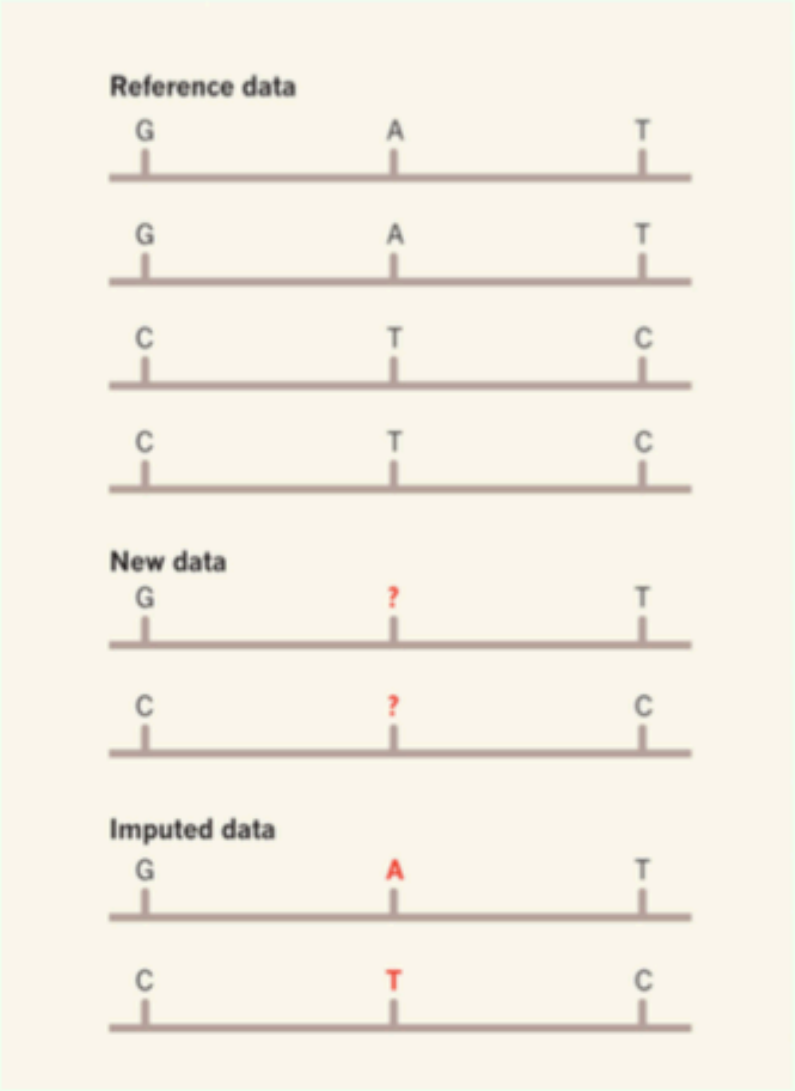
Data filtering



- Sequencing and mapping errors

Missing data!

Haplotype imputation - simplified



Reference

- 1000 Genomes
- Phased using family structures

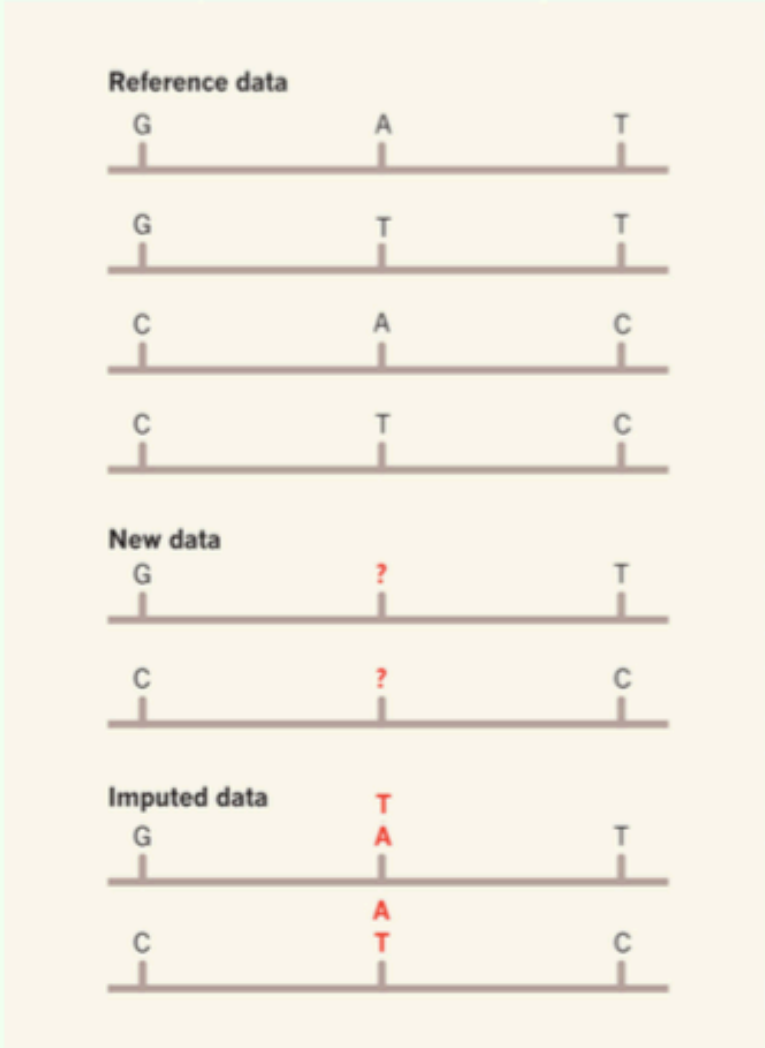
new data

- partial information

Imputed data

- Probabilistic approach
- The results retains the uncertainty of both the genotype and the haplotypes

Haplotype imputation - simplified



Reference

- 1000 Genomes
- Phased using family structures

new data

- Data with known and unknown genotypes

Imputed data

$p(? = T) =$
$p(? = A) =$

Anders Albrechtsen

Haplotype imputation - simplified

Reference

- haplotype frequencies

new data

- Data with known and unknown genotypes

first haplotype

$$p(? = T) = \frac{0.56}{0.56+0.03} = 0.95$$
$$p(? = A) = \frac{0.03}{0.56+0.03} = 0.05$$

second haplotype

$$p(? = T) = \frac{0.21}{0.21+0.2} = 0.51$$
$$p(? = A) = \frac{0.2}{0.21+0.2} = 0.49$$

Anders Albrechtsen

Haplotype imputation - simplified

Bayes formula

$$p(H = h|f, G) = \frac{P(G|H=h)P(H=h|f)}{\sum_{h'} P(G|H=h')P(H=h'|f)}$$

$P(G|H = h)$

1 if consistent

0 otherwise

first haplotype

$$p(? = T) = \frac{0.56}{0.56+0.03} = 0.95$$

$$p(? = A) = \frac{0.03}{0.56+0.03} = 0.05$$

The likelihood for known genotypes

$G = \{G_1, G_2, \ldots, G_N\}$ with $G_i = \{G_i^1, G_i^2, \ldots, G_i^M\}$

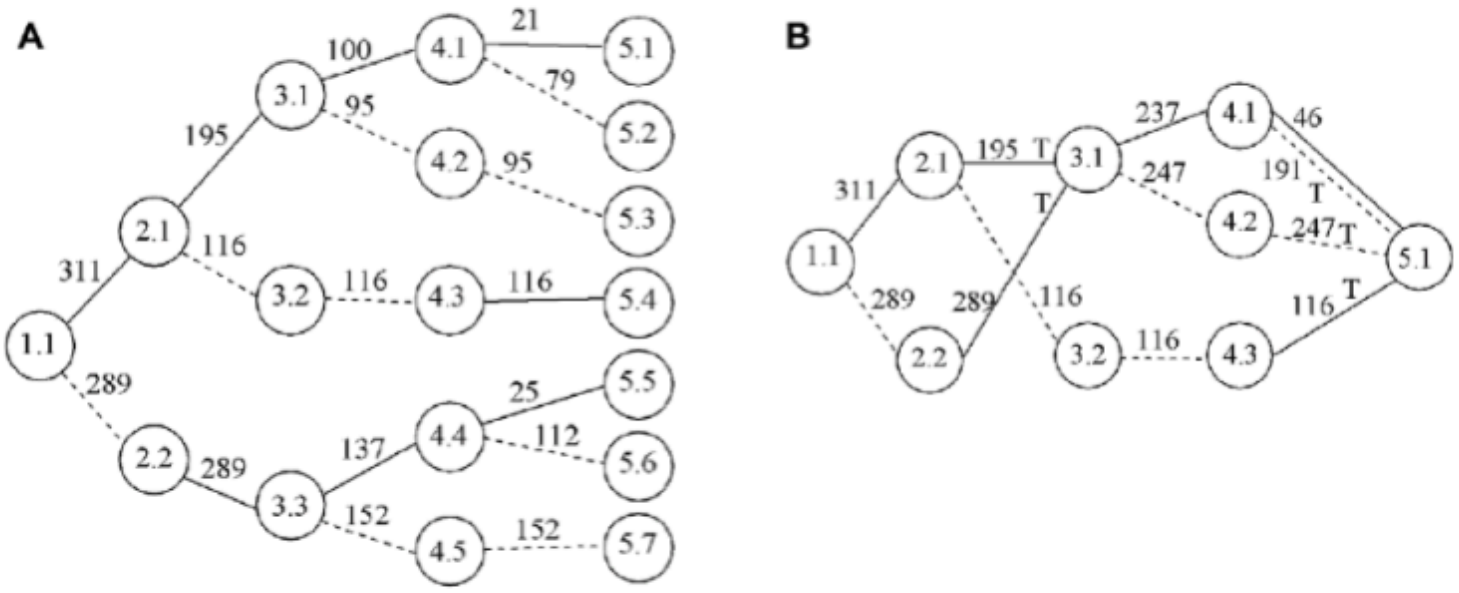
$$p(G|f) = \prod^{N} = p(G_i|f)$$

$$p(G_i|f) = \sum_{h_1} \sum_{h_2} p(G_i|H = \{h_1, h_2\}) p(H = \{h_1, h_2\}|f)$$

$$p(H = \{h_1, h_2\}|f) = f_{h_1} f_{h_2}$$

BEAGLE algorithm

Merging graphs with known haplotypes for 4 sites



Anders Albrechtsen

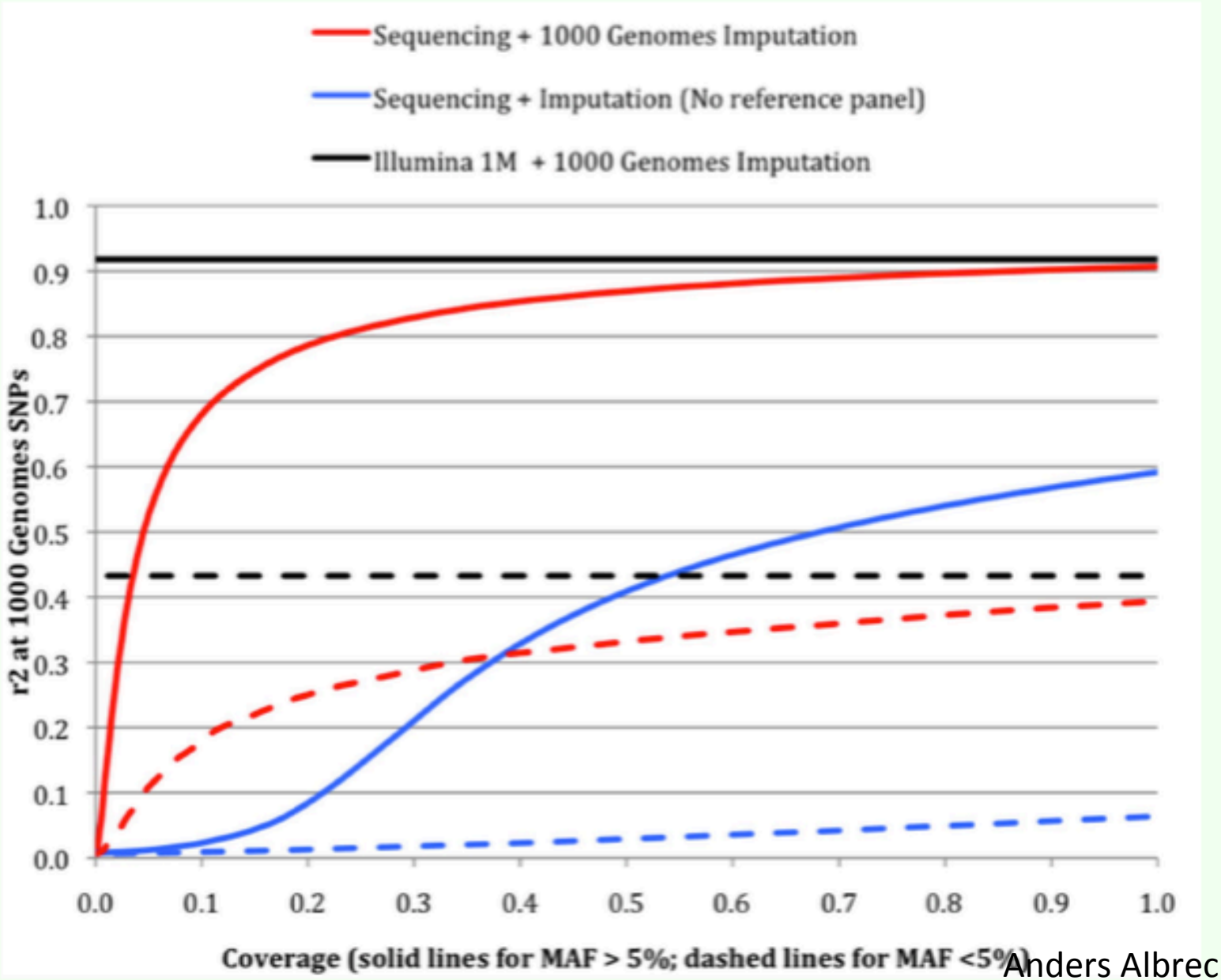
Ultra low sequencing

Medium depth vs. low depth

Ultra low sequencing

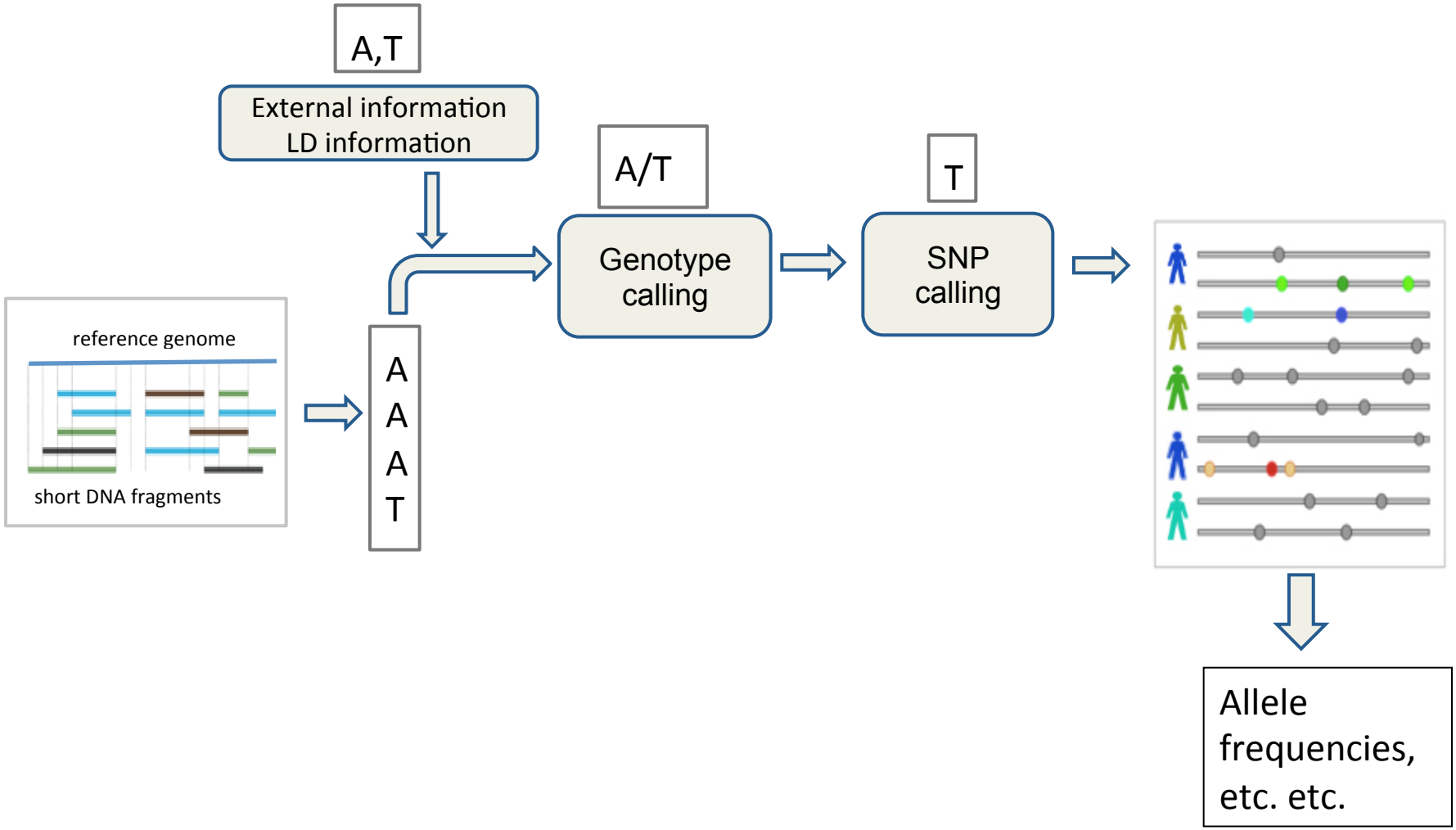
- Depth lower than 1X
- also a by product of captured data

Information retained with ultra low sequencing

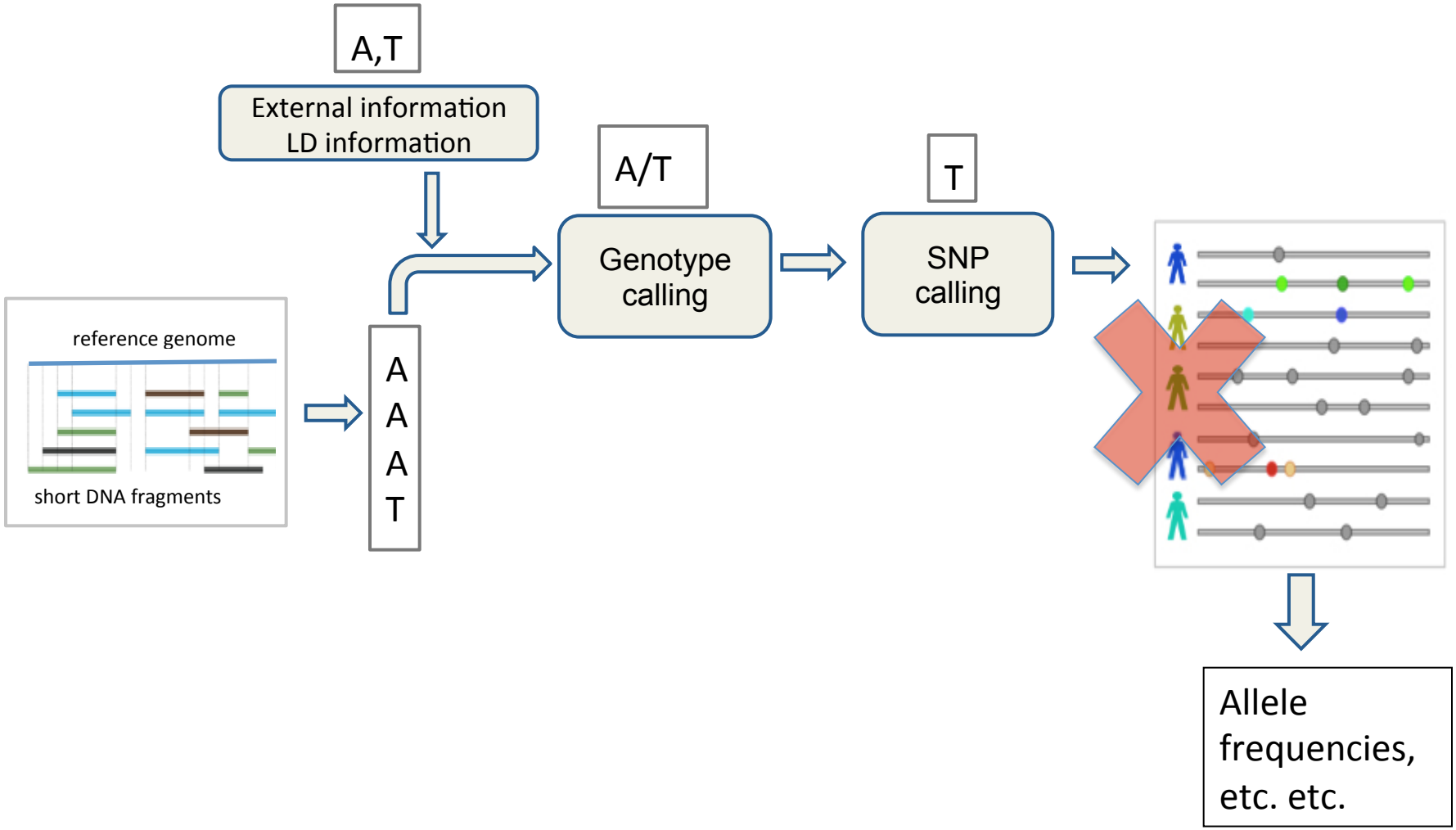


Anders Albrechtsen

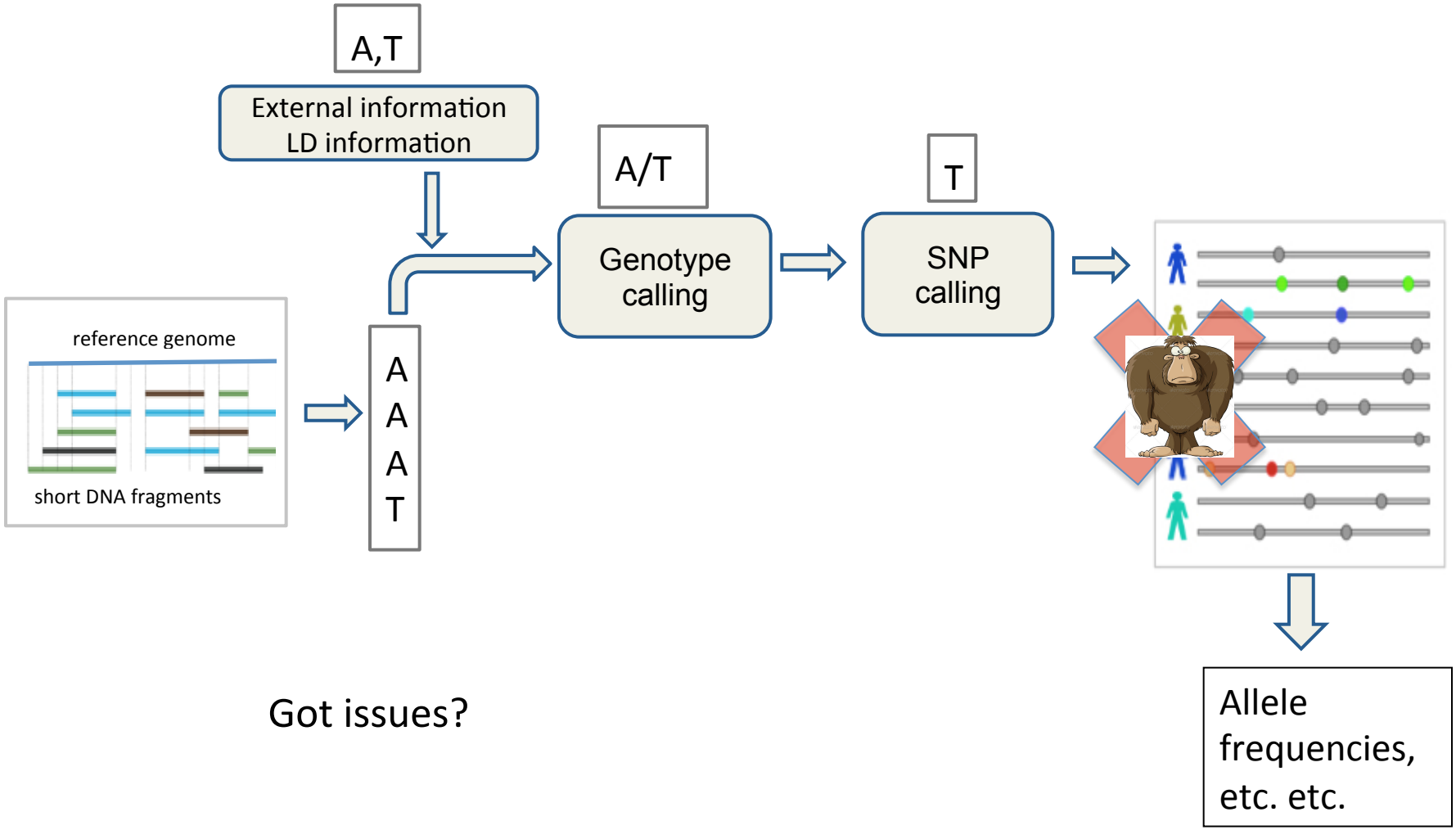
Next Generation Sequencing data processing



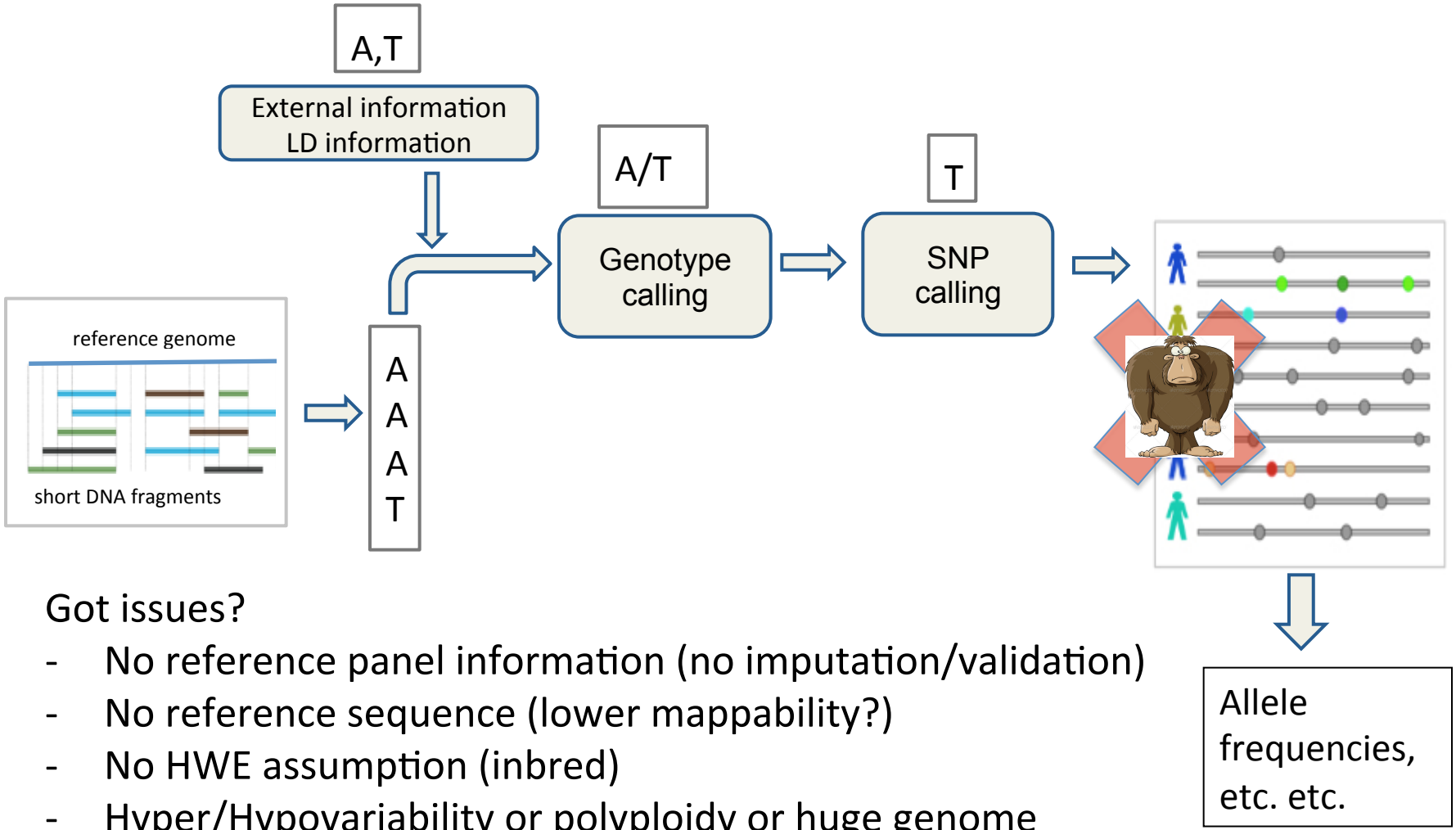
Next Generation Sequencing data processing



Next Generation Sequencing data processing in the non-model world



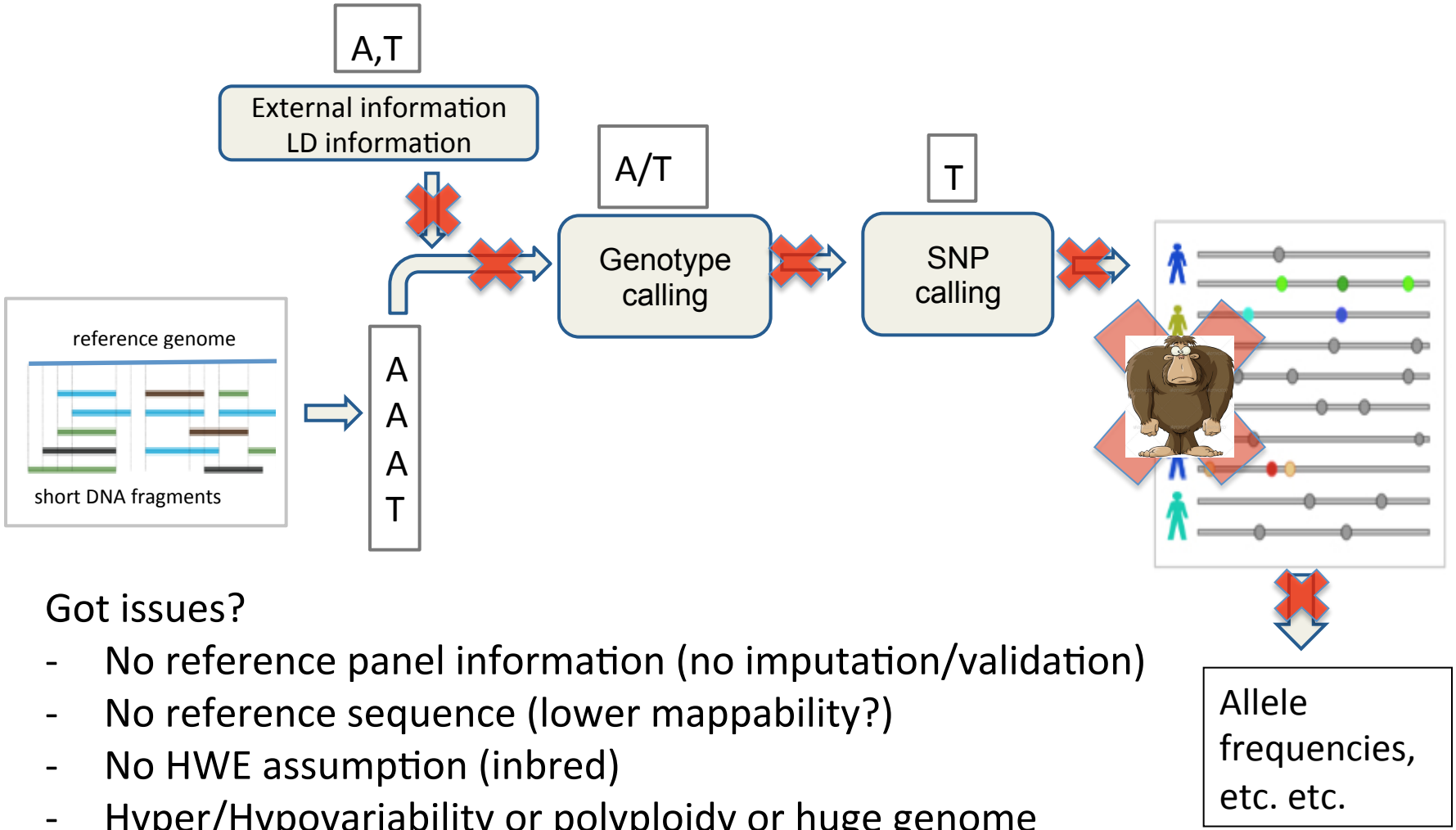
Next Generation Sequencing data processing in the non-model world



Got issues?

- No reference panel information (no imputation/validation)
- No reference sequence (lower mappability?)
- No HWE assumption (inbred)
- Hyper/Hypovariability or polyploidy or huge genome
- No money (?)
- …

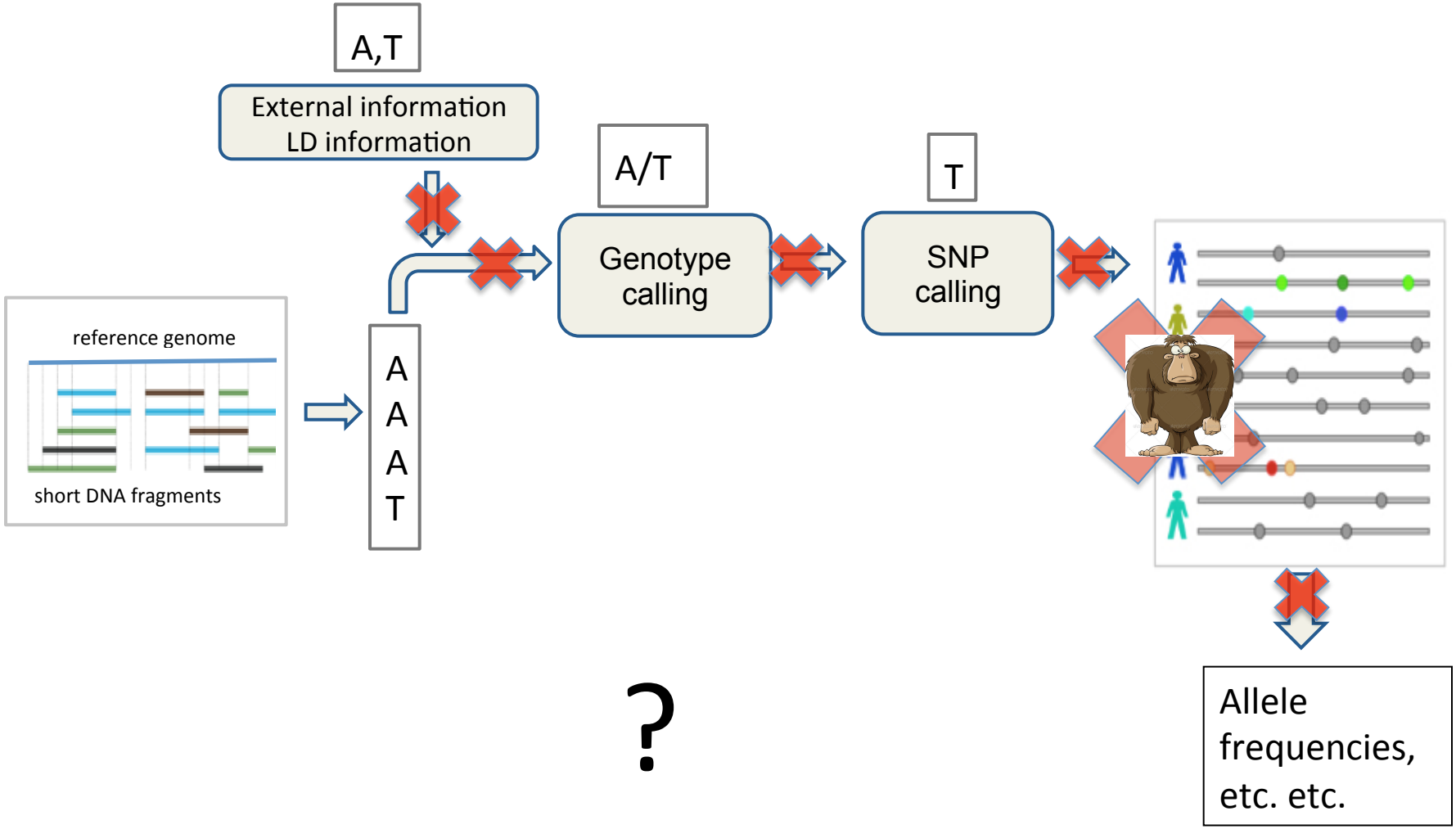
Next Generation Sequencing data processing in the non-model world



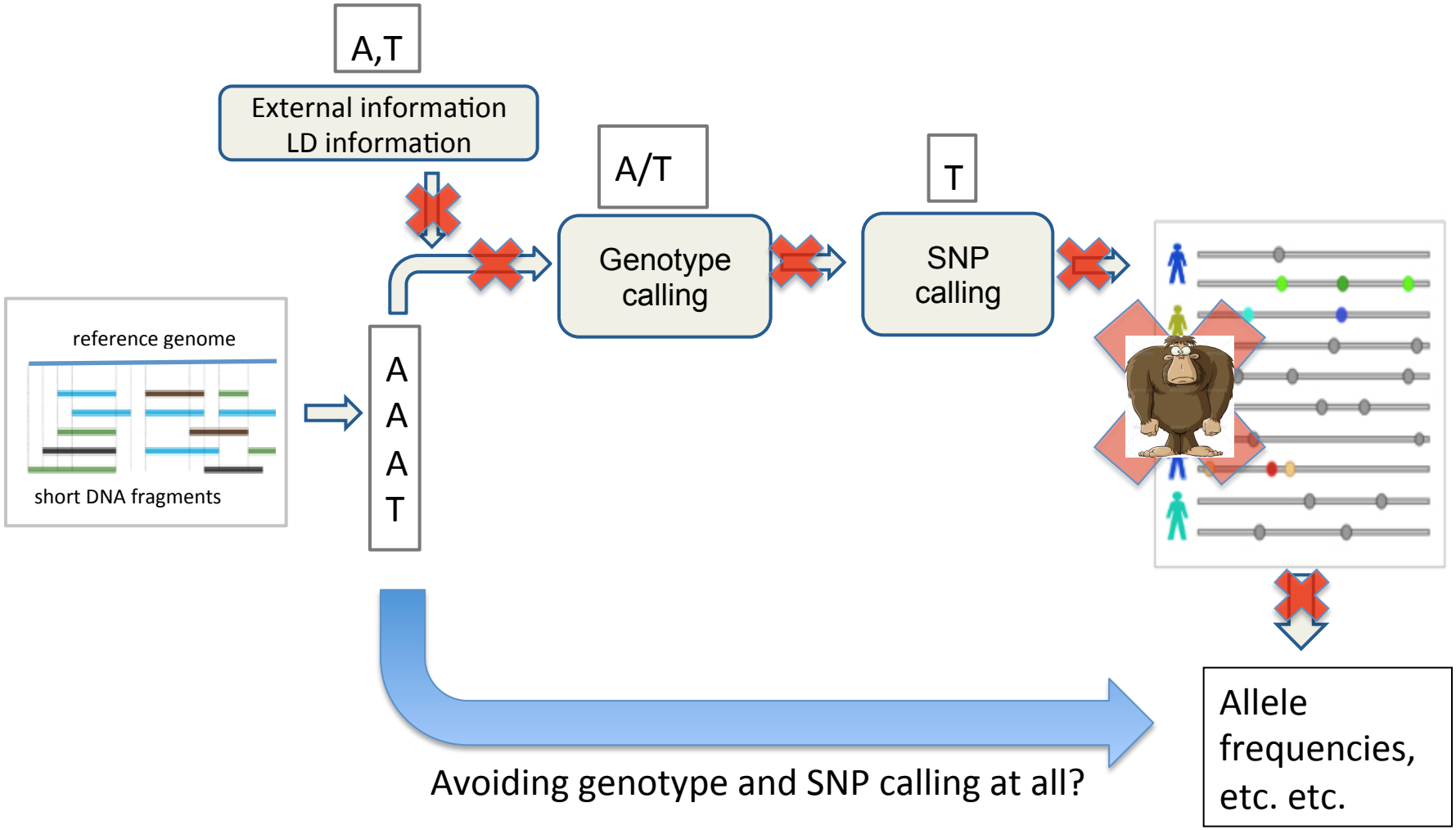
Got issues?

- No reference panel information (no imputation/validation)
- No reference sequence (lower mappability?)
- No HWE assumption (inbred)
- Hyper/Hypovariability or polyploidy or huge genome
- No money (?)
- **Your inferences will be wrong!**

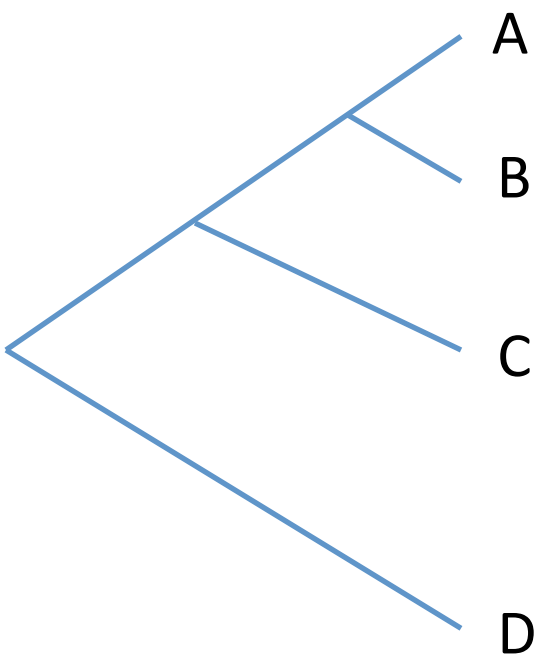
Next Generation Sequencing data processing in the non-model world



Next Generation Sequencing data processing in the non-model world

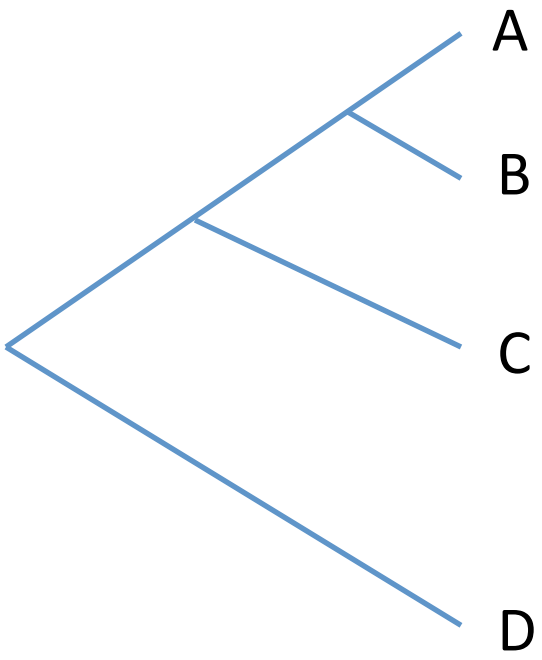


Genetic distances



Genotype 1	Genotype 2	Distance
aa	aa	0
aa	aA	1
aa	AA	2
aA	aa	1
aA	aA	0
aA	AA	2
…	…	…

Genetic distances



Genotypes are {aa, aA, AA} as {0, 1 ,2}

For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s) - g(j,s)|}{2}\right)$$

genotype of *i* at site *s*

e.g. G(i=A,s=1)=0 and G(j=B,s=1)=1 then d(i,j)=1

Genetic distances from known genotypes

Genotypes are {aa, aA, AA} as {0, 1 ,2}
For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s) - g(j,s)|}{2}\right)$$

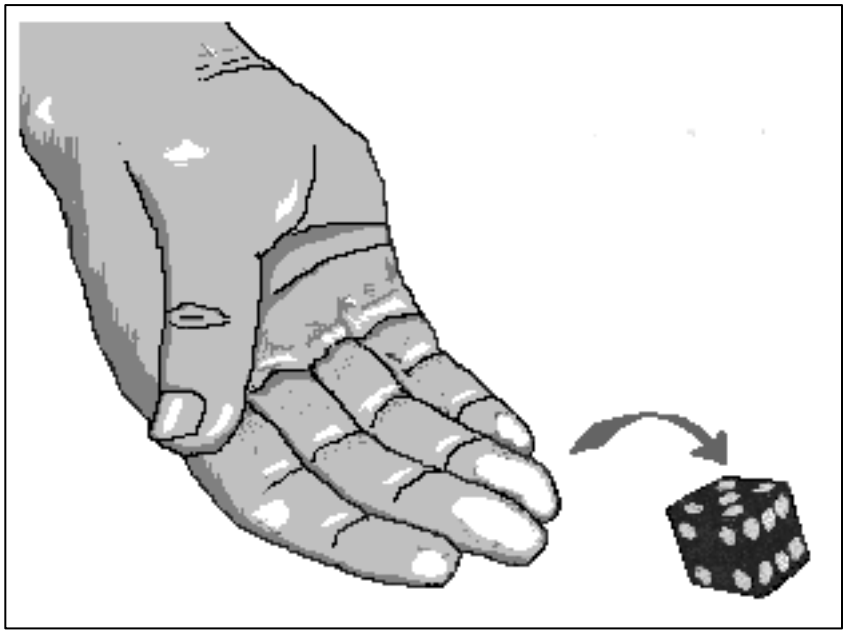
d(i,j) = 1*1.00 = 1.00/2

B

A

	0	**1**	**2**
0	0	1	0
1	0	0	0
2	0	0	0

Expected value



What are the possible outcomes?
With what probability?

from Wikipedia

Expected value

- The expected value of a discrete random variable is the probability-weighted average of all possible values

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i,$$

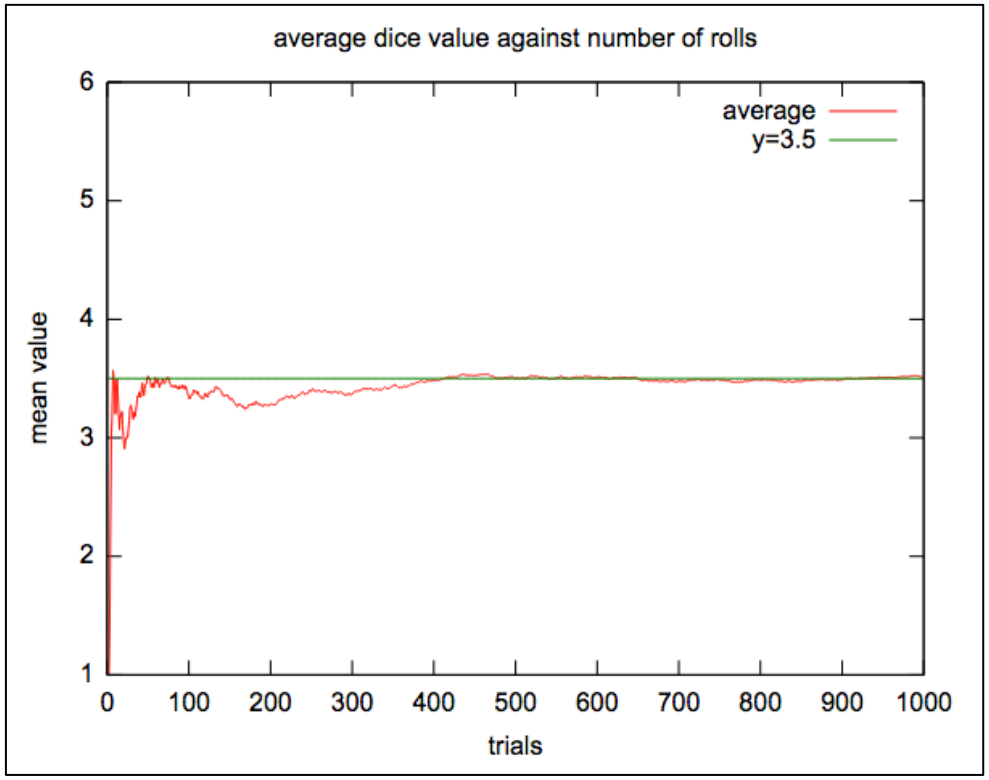
$$\mathrm{E}[X] = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_k p_k}{1} = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_k p_k}{p_1 + p_2 + \cdots + p_k}$$

from Wikipedia

Expected value

- Average value if you perform the same experiment many times

$$\mathrm{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$



from Wikipedia

Genetic distances from (un)known genotypes

Genotypes are {aa, aA, AA} as {0, 1 ,2}
For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s) - g(j,s)|}{2}\right)$$

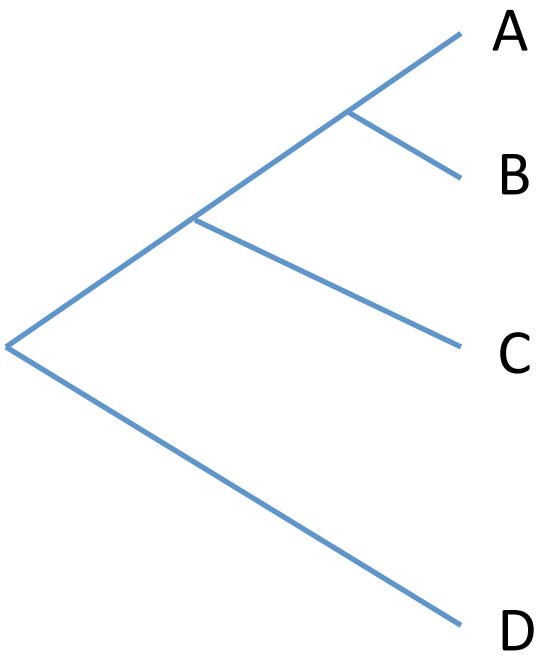
E[d(i,j)] = 0*0.30 + 1*0.50 + 2*0.10 + 1*0.10 + … = 0.80/2

B

A

	0	**1**	**2**
0	0.30	0.50	0.10
1	0.10	0	0
2	0	0	0

Genetic distances from unknown genotypes



Genotypes are {aa, aA, AA} as {0, 1 ,2}

For individuals i and j and N sites:

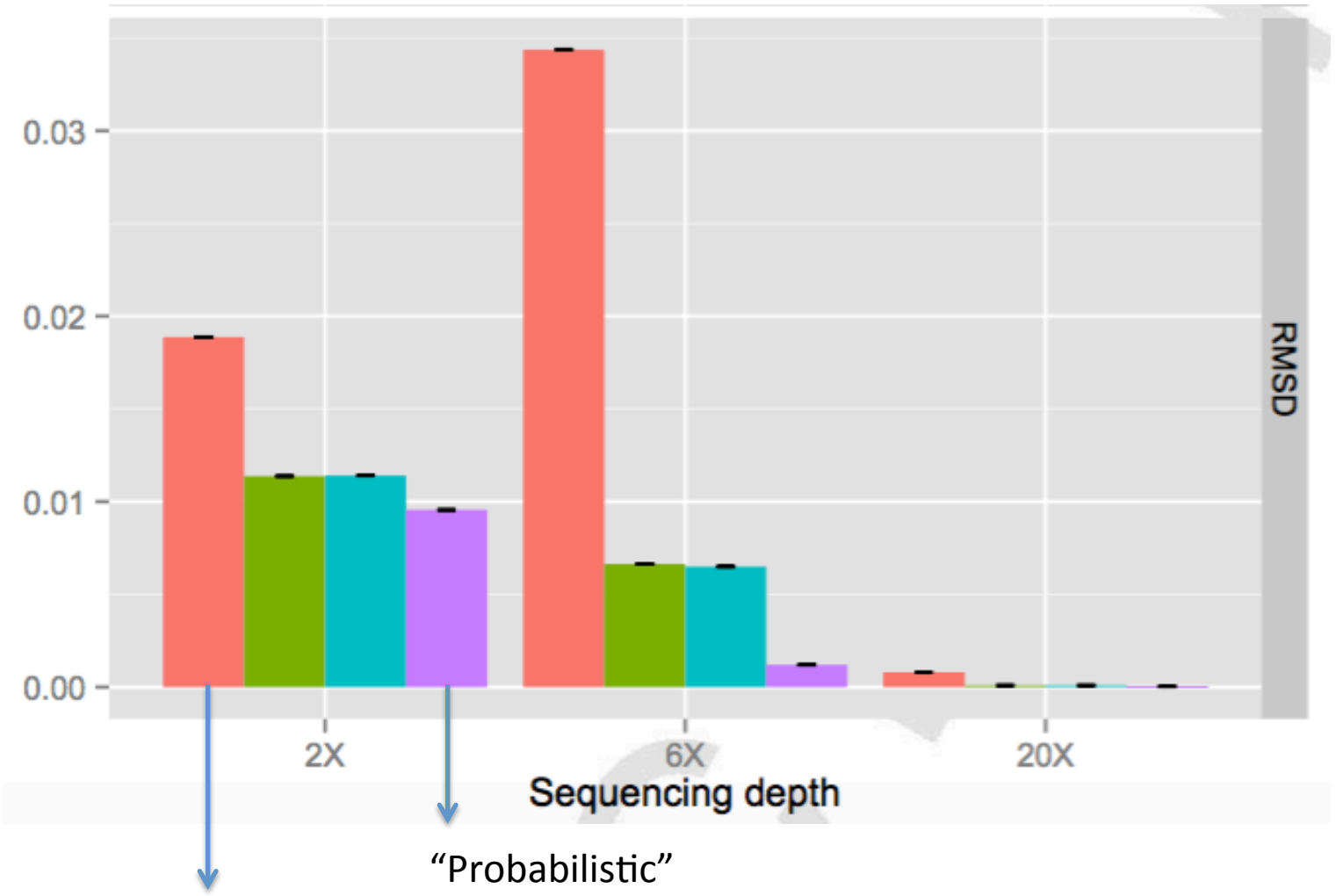
$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s) - g(j,s)|}{2}\right)$$

Iterate across all possible genotypes

Genotypes probability

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\sum_{g(i,s)=0}^{2}\sum_{g(j,s)=0}^{2}\frac{|g(i,s) - g(j,s)|}{2} * P(g(i,s), g(j,s))\right)$$

Genetic distances from unknown genotypes

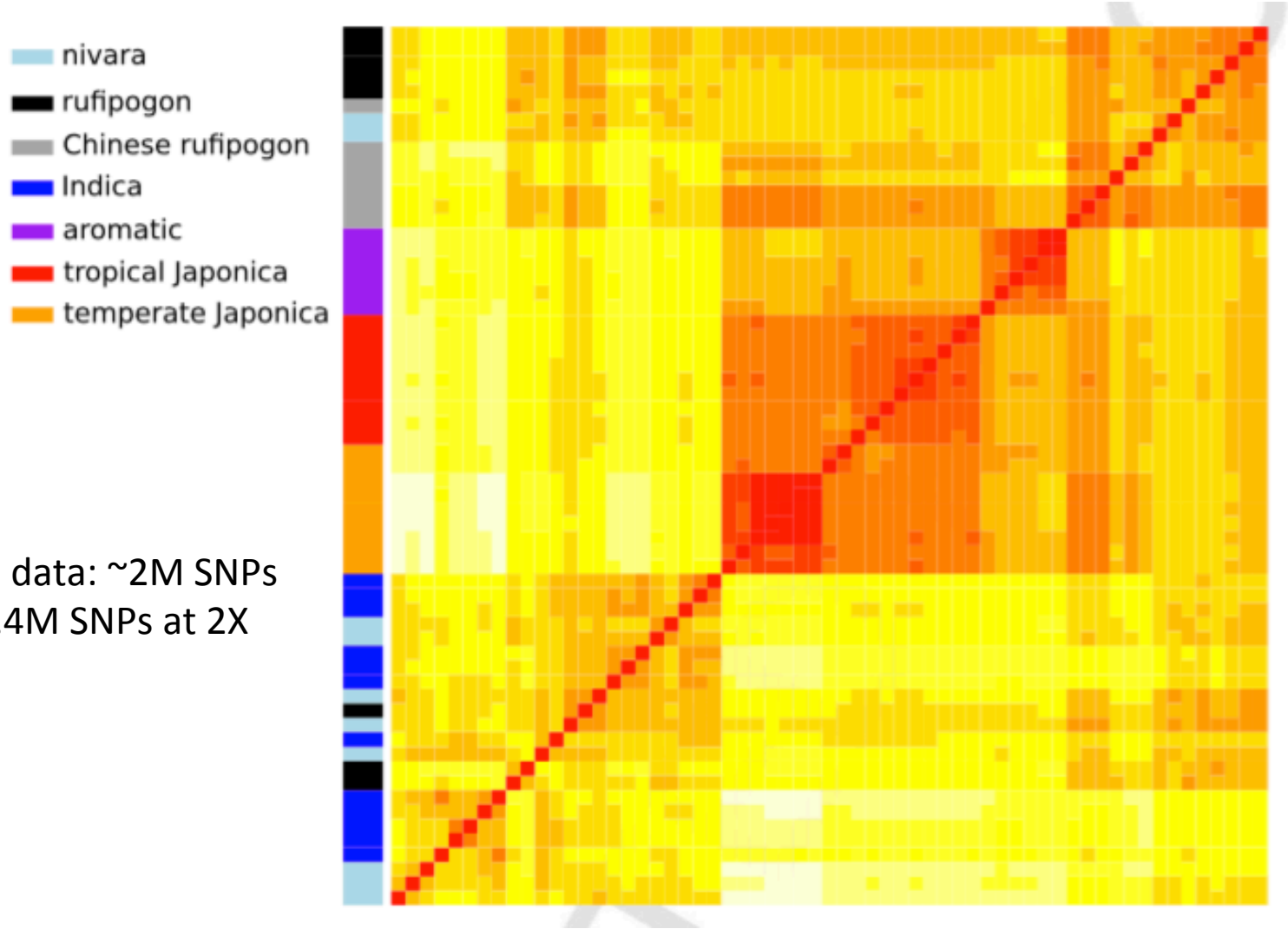


"Probabilistic"

Genotype calling (no prior)

Vieira et al. BJLS 2016

Clustering



Original data: ~2M SNPs
Here: 5.4M SNPs at 2X

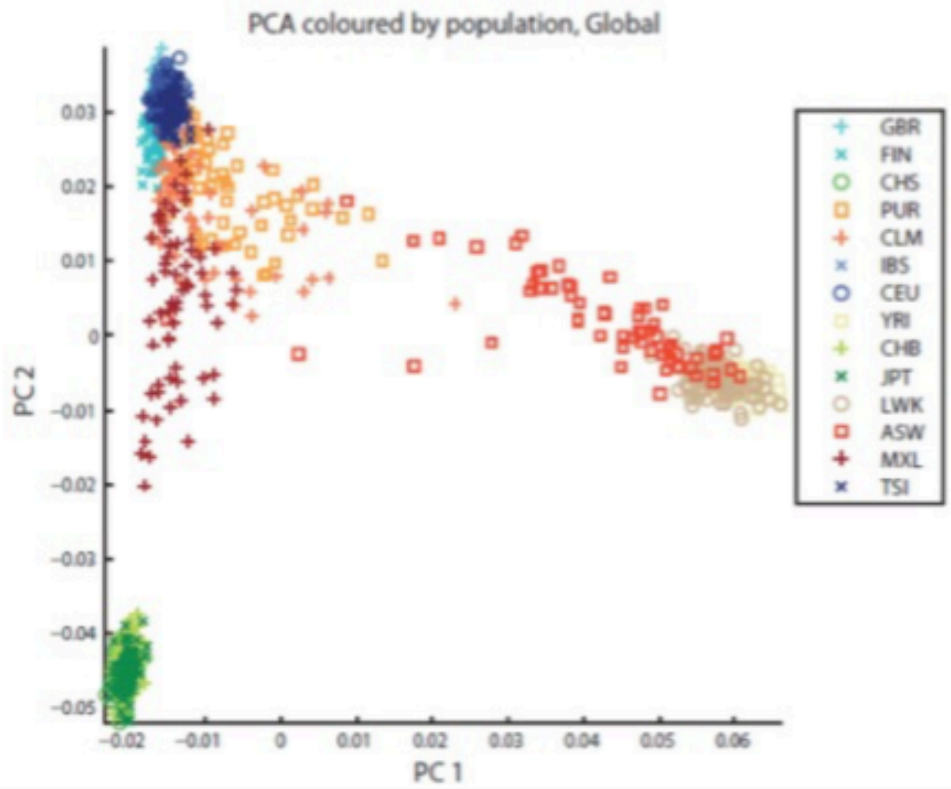
https://github.com/fgvieira/ngsDist

Vieira et al. BJLS 2016

Population structure

Principal Component Analysis (PCA) is a **data reduction** method for

- **visualization**,
- correction for population stratification,
- population history and differentiation.



1000 Genomes Consortium 2012

Covariance matrix

Genotype (0,1,2)

Allele frequency

$$cov(i,j) = \frac{1}{(m-1)} \frac{\sum_{s=1}^{m} (G_s^{(i)} - 2\hat{p}_s)(G_s^{(j)} - 2\hat{p}_s)}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

Covariance matrix

Genotype (0,1,2)

Allele frequency

$$cov(i,j) = \frac{1}{(m-1)} \frac{\sum_{s=1}^{m} (G_s^{(i)} - 2\hat{p}_s)(G_s^{(j)} - 2\hat{p}_s)}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

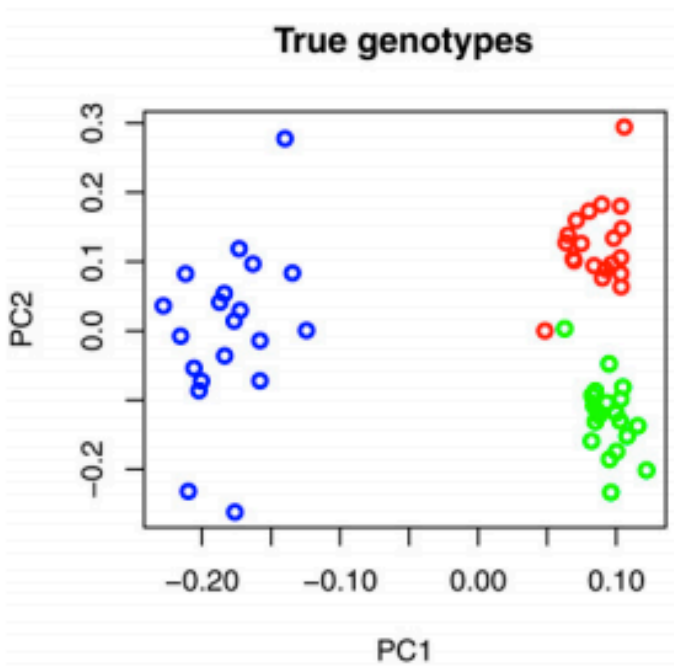
Iterate across all genotypes

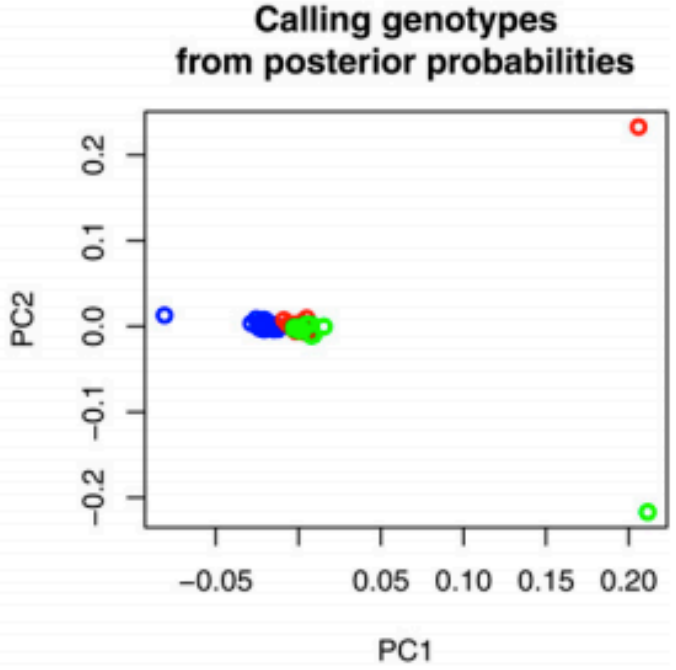
Weight by their probability

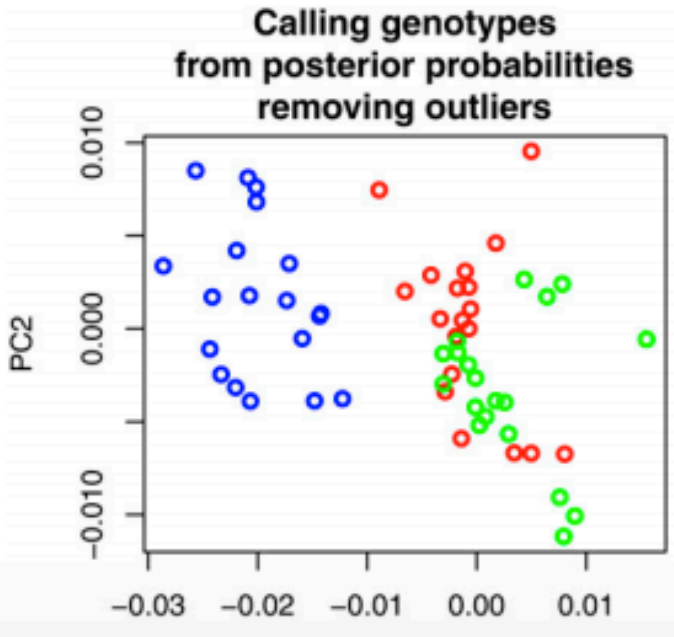
$$\hat{cov}_{(i,j)} := \frac{1}{(\sum_{s=1}^{m} P_{var,s}) - 1} \frac{\sum_{s=1}^{m} \left(\sum_{G_s^{(i)}=0}^{2} \sum_{G_s^{(j)}=0}^{2} (G_s^{(i)} - 2\hat{p}_s)(G_s^{(j)} - 2\hat{p}_s) P(G_s^{(i)} | X_s^{(i)}) P(G_s^{(j)} | X_s^{(j)}) \right) P_{var,s}}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

Probability of the site being variable
(to avoid SNP calling)

Depth=2X

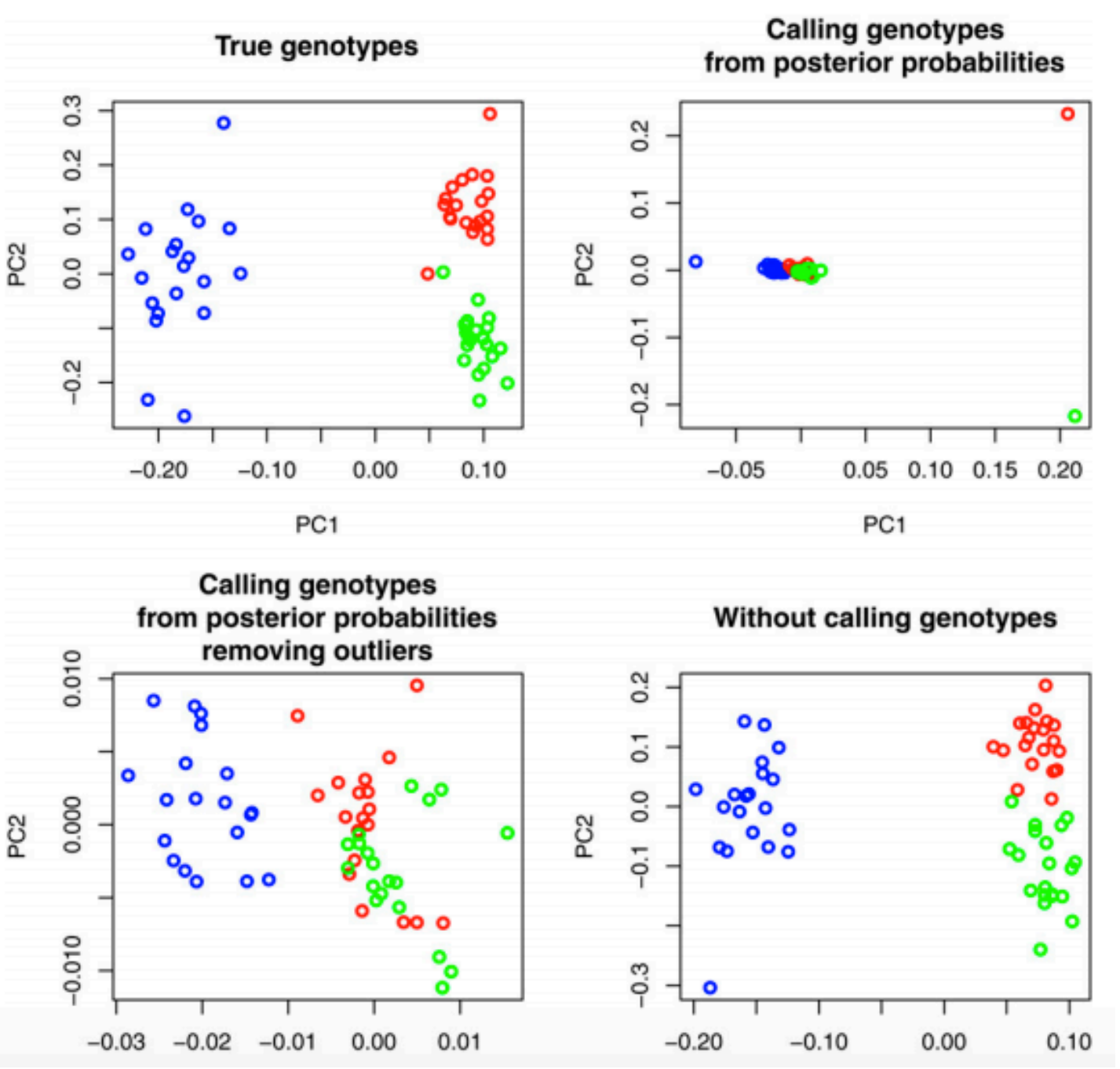

True genotypes
PC2
PC1
0.3
0.2
0.1
0.0
−0.2
−0.20
−0.10
0.00
0.10


Calling genotypes
from posterior probabilities
PC2
PC1
0.2
0.1
0.0
−0.1
−0.2
−0.05
0.05
0.10
0.15
0.20


Calling genotypes
from posterior probabilities
removing outliers
PC2
0.010
0.000
−0.010
−0.03
−0.02
−0.01
0.00
0.01

Without calling genotypes

Depth=2X


True genotypes
PC2
PC1
Calling genotypes
from posterior probabilities
PC2
PC1
Calling genotypes
from posterior probabilities
removing outliers
PC2
Without calling genotypes
PC2

Site Frequency Spectrum (SFS)

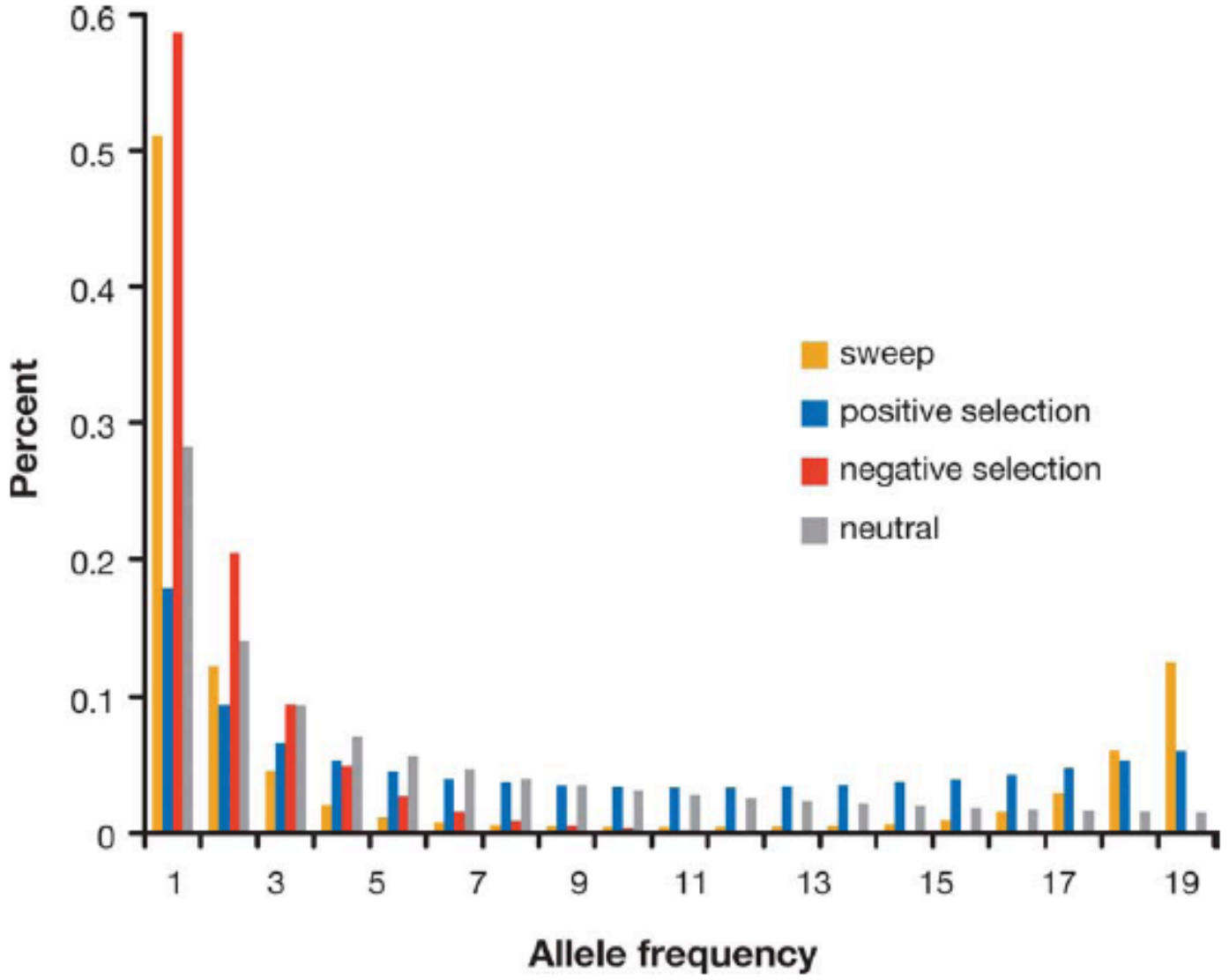
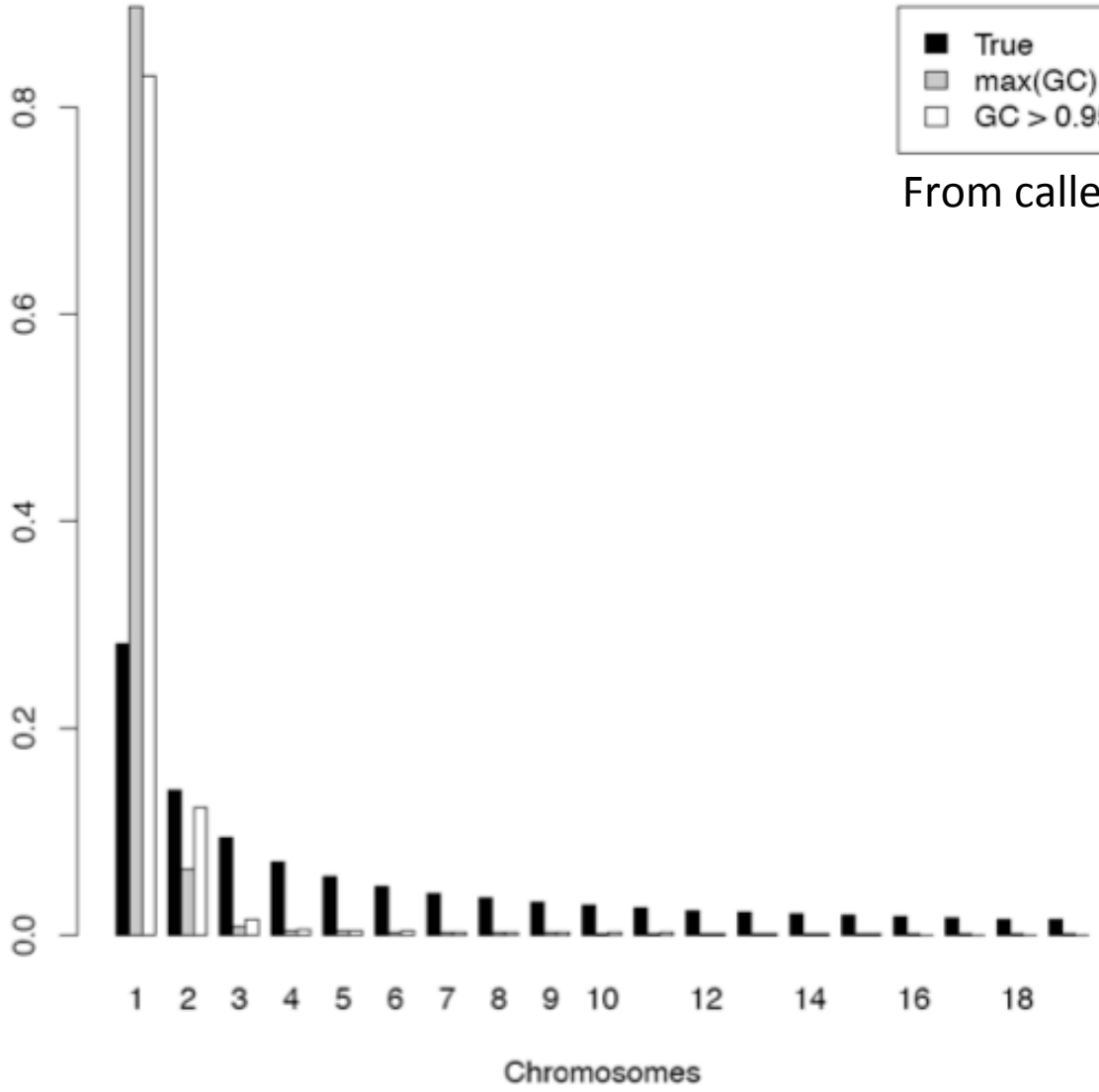


Figure 2

Effect of errors on SFS



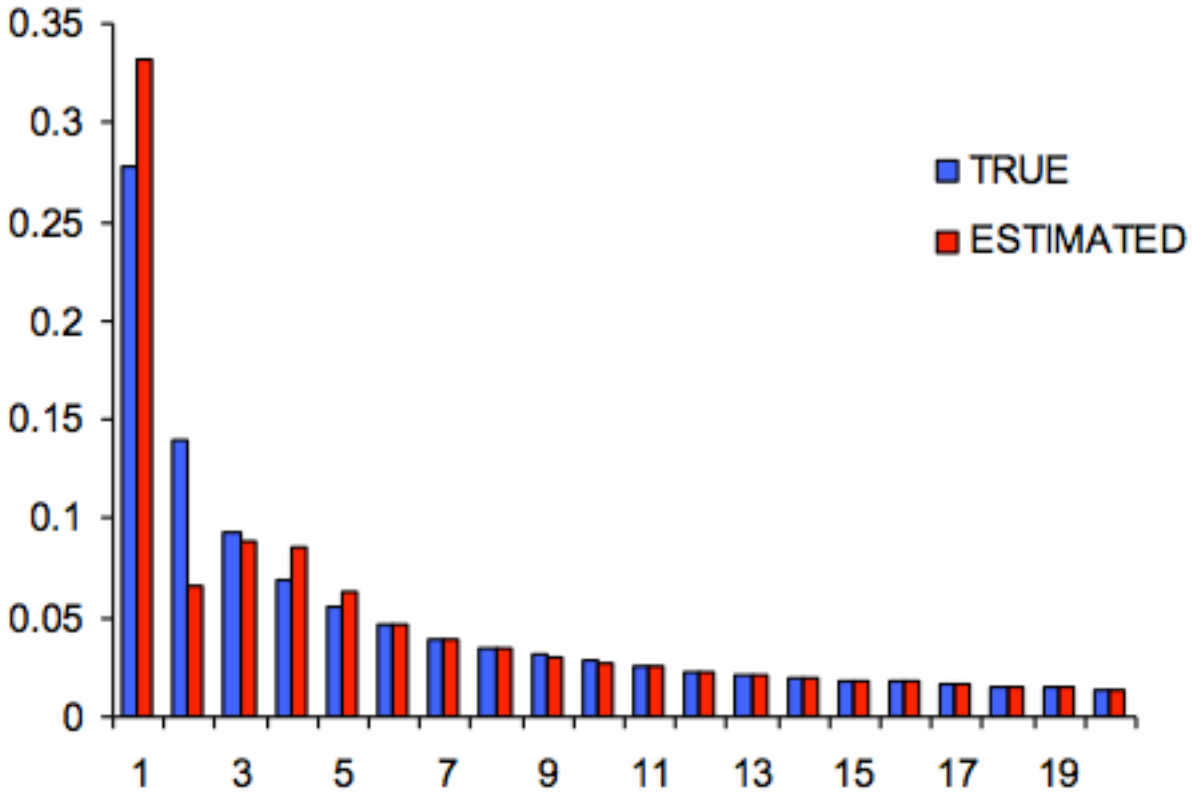
5X

From called genotypes

Slide by R. Nielsen

Effect of errors on SFS

Using an ad hoc fixed cutoff for SNP calling…



can never produce unbiased estimates.

Slide by R. Nielsen

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded?

If folded?

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries

p_0	p_1	p_2	p_3	…	p_{2k}

If folded, *k+1* entries

p_0	p_1	p_2	…	p_k

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries

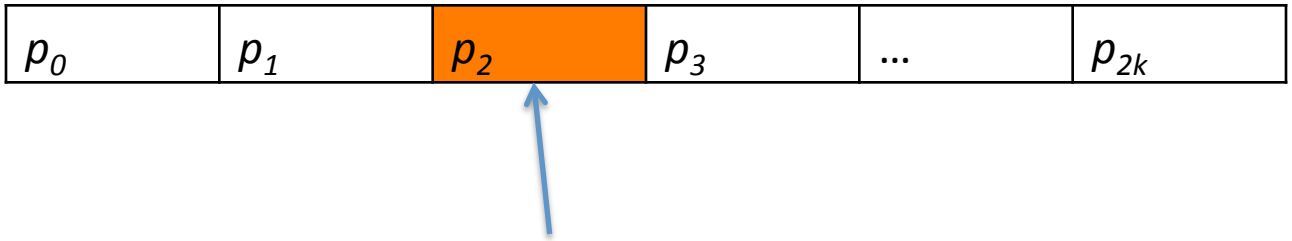
p_0	p_1	p_2	p_3	…	p_{2k}

e.g. A is ancestral, G is derived (alternate)
AA AA AG AA AG AA AA AA AA

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries



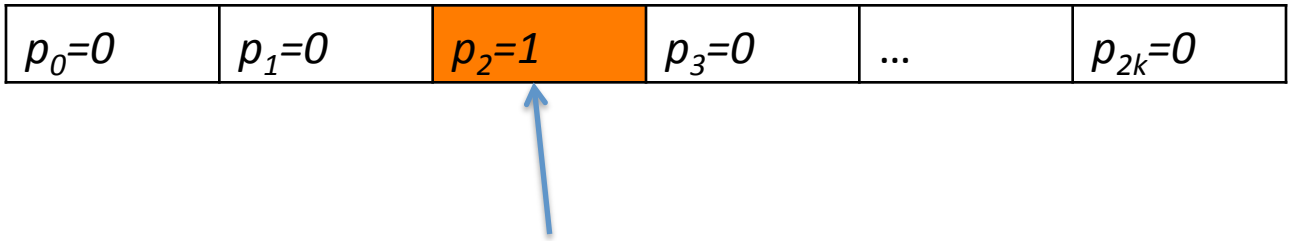
e.g. A is ancestral, G is derived (alternate)

AA AA AG AA AG AA AA AA AA

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries



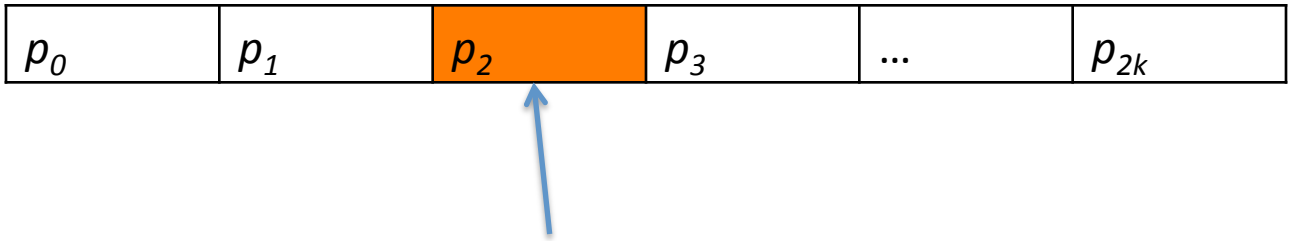
e.g. A is ancestral, G is derived (alternate)

AA AA AG AA AG AA AA AA AA

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries

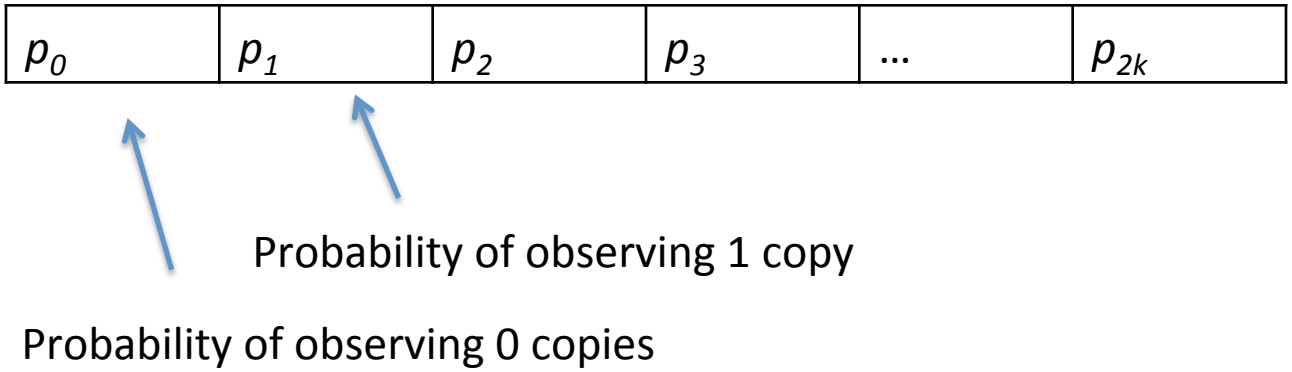


e.g. A is ancestral, G is derived (alternate)
If genotypes are unknown? Counting is not possible?

Sample allele frequency

- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries

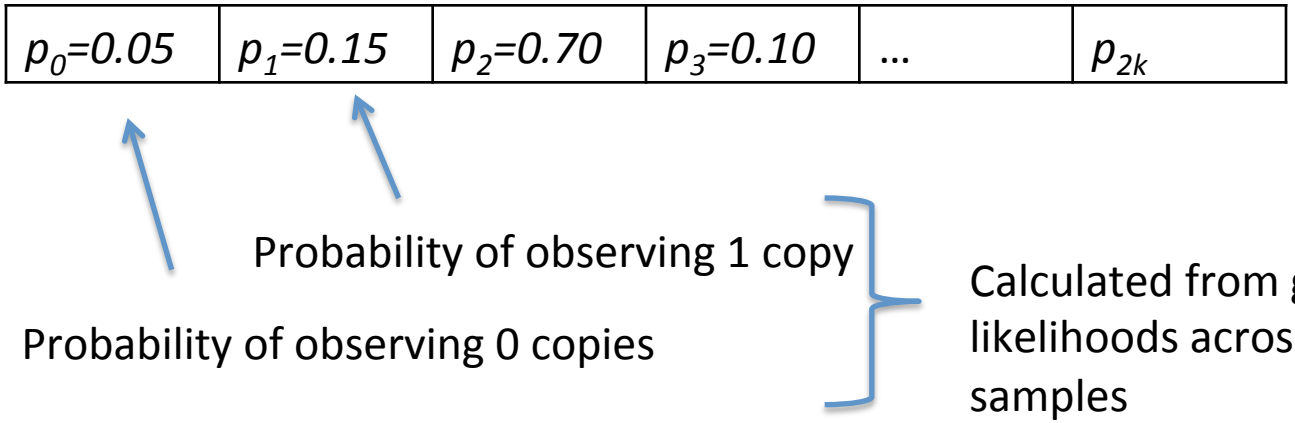


e.g. A is ancestral, G is derived (alternate)

Sample allele frequency

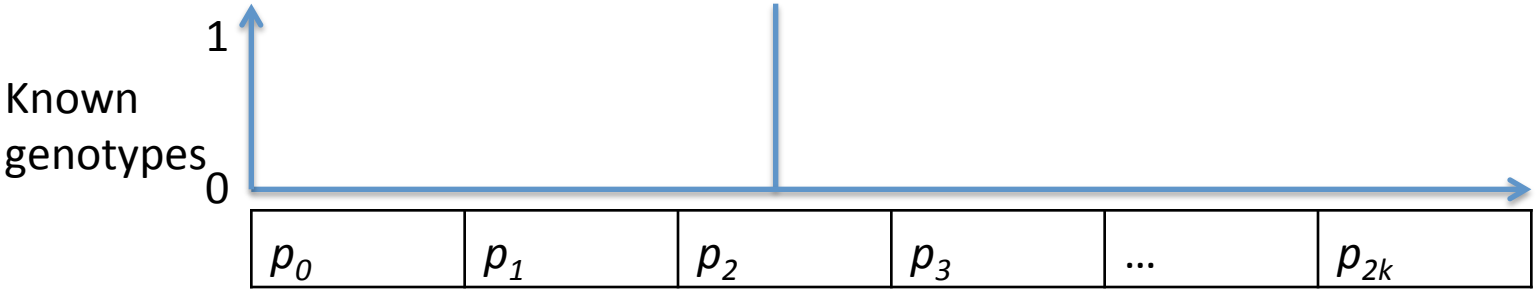
- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries

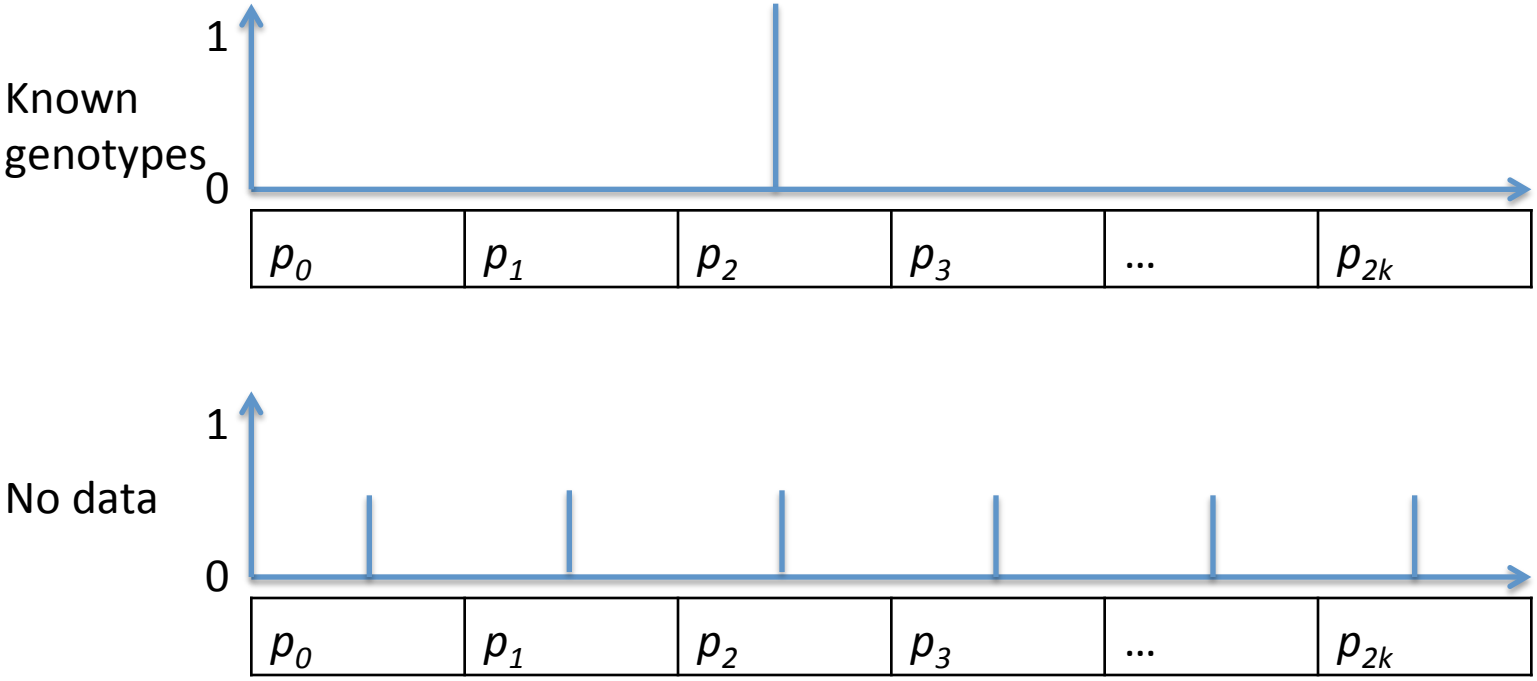


e.g. A is ancestral, G is derived (alternate)

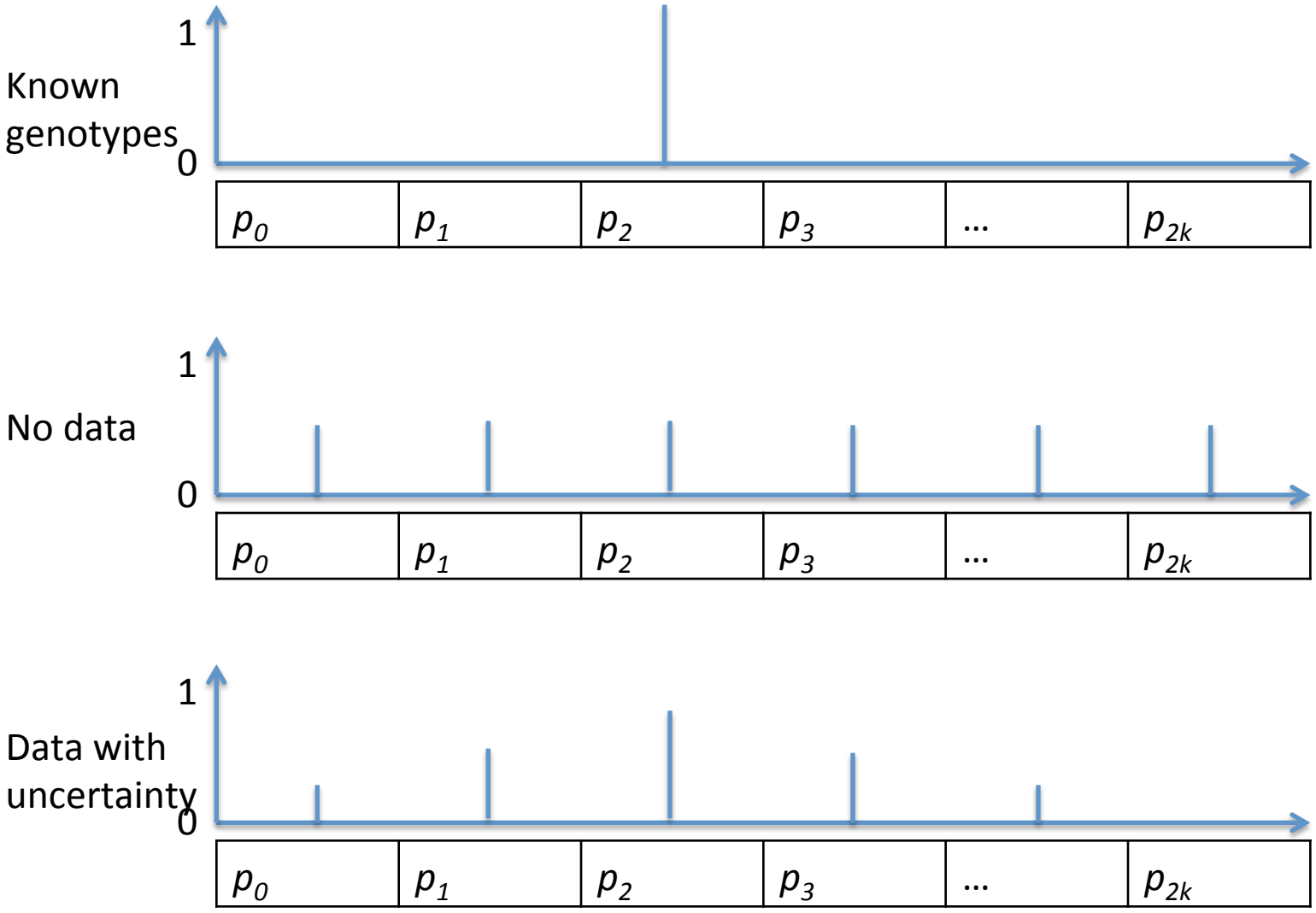
Sample allele frequency probabilities



Sample allele frequency probabilities



Sample allele frequency probabilities



Sample allele frequency posterior probabilities

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

- Estimating allele frequency

$$\hat{f} =$$

Sample allele frequency posterior probabilities

$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

- Estimating allele frequency

$$\hat{f} = \sum_{i=0}^{2k} \left(\frac{i}{2k} \right) p(S = i)$$

Sample allele frequency posterior probabilities

With 6 chromosomes (3 diploids)

p_0=0.10	p_1=0.15	p_2=0.50	p_3=0.15	p_4=0.05	p_5=0.05	p_6=0.00

- SNP calling

$$p_{\text{var}} = \ ?$$

$$p_{\text{var}} > t$$

with *t* being 0.95, 0.99, 0.999 and so on.

Sample allele frequency posterior probabilities

p_0=0.10	p_1=0.15	p_2=0.50	p_3=0.15	p_4=0.05	p_5=0.05	p_6=0.00

- SNP calling

$$p_{\text{var}} = 1 - p(S=0) - p(S=2k) \quad = 0.90$$

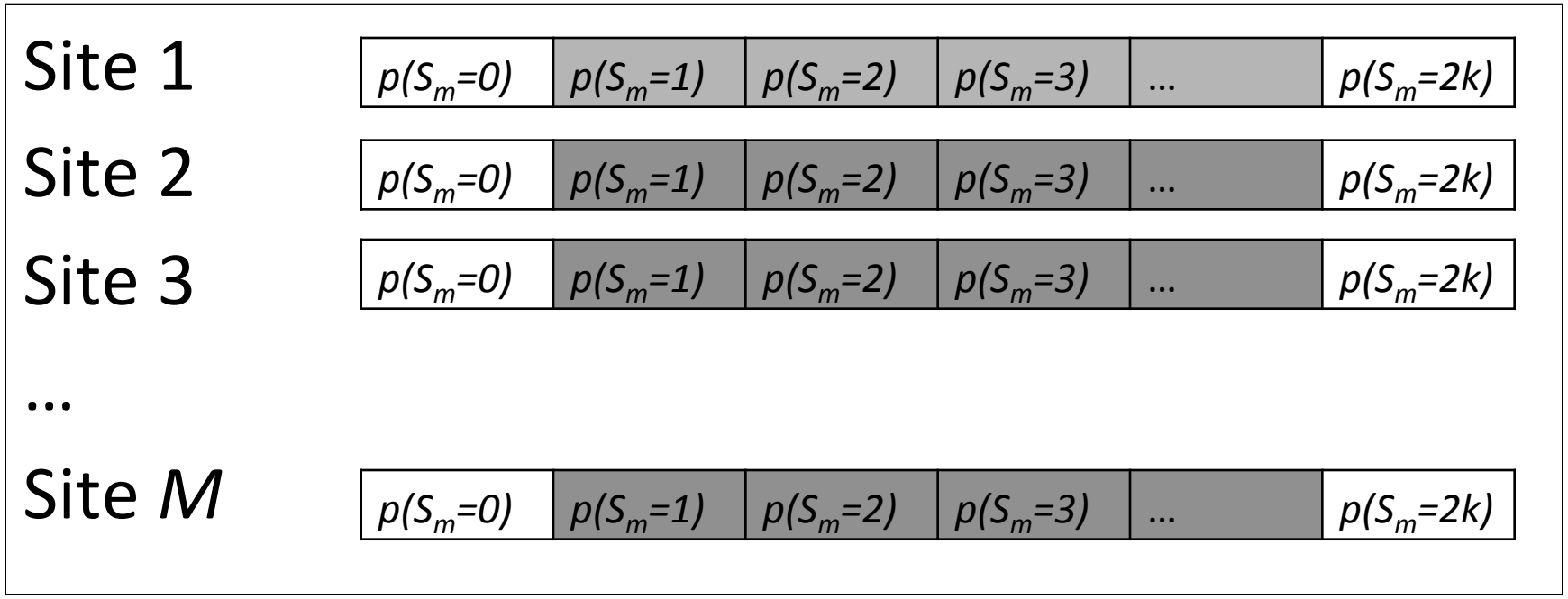
$$p_{\text{var}} > t$$

with *t* being 0.95, 0.99, 0.999 and so on.

Nr of segregating sites

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

Nr of segregating sites



Nr of segregating sites

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$E[S] = \sum_{m=1}^{M} p_{\text{var}}^{(m)} = \sum_{m=1}^{M} (1 - p(S_m = 0) - p(S_m = 2k))$$

Nucleotide diversity

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$D = 2f(1-f)$$

$$E[D] =$$

Nucleotide diversity

Site 1	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 2	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
Site 3	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$
…						
Site *M*	$p(S_m=0)$	$p(S_m=1)$	$p(S_m=2)$	$p(S_m=3)$	…	$p(S_m=2k)$

$$E[D] = \sum_{m=1}^{M} \sum_{j=0}^{2k} 2 \left(\frac{i}{2k} \right) \left(\frac{2k-i}{2k} \right) p(S_m = i)$$

Maximum Likelihood Estimation (MLE) of the **Site Frequency Spectrum**

- Parameterize the SFS, with *k* individuals

$$\overline{P} = (p_0, p_1, \cdots p_{2k})$$

If unfolded, *2k+1* entries

p_0	p_1	p_2	p_3	…	p_{2k}

If folded, *2k* entries

p_0	p_1	p_2	…	p_k

ML estimation of the SFS

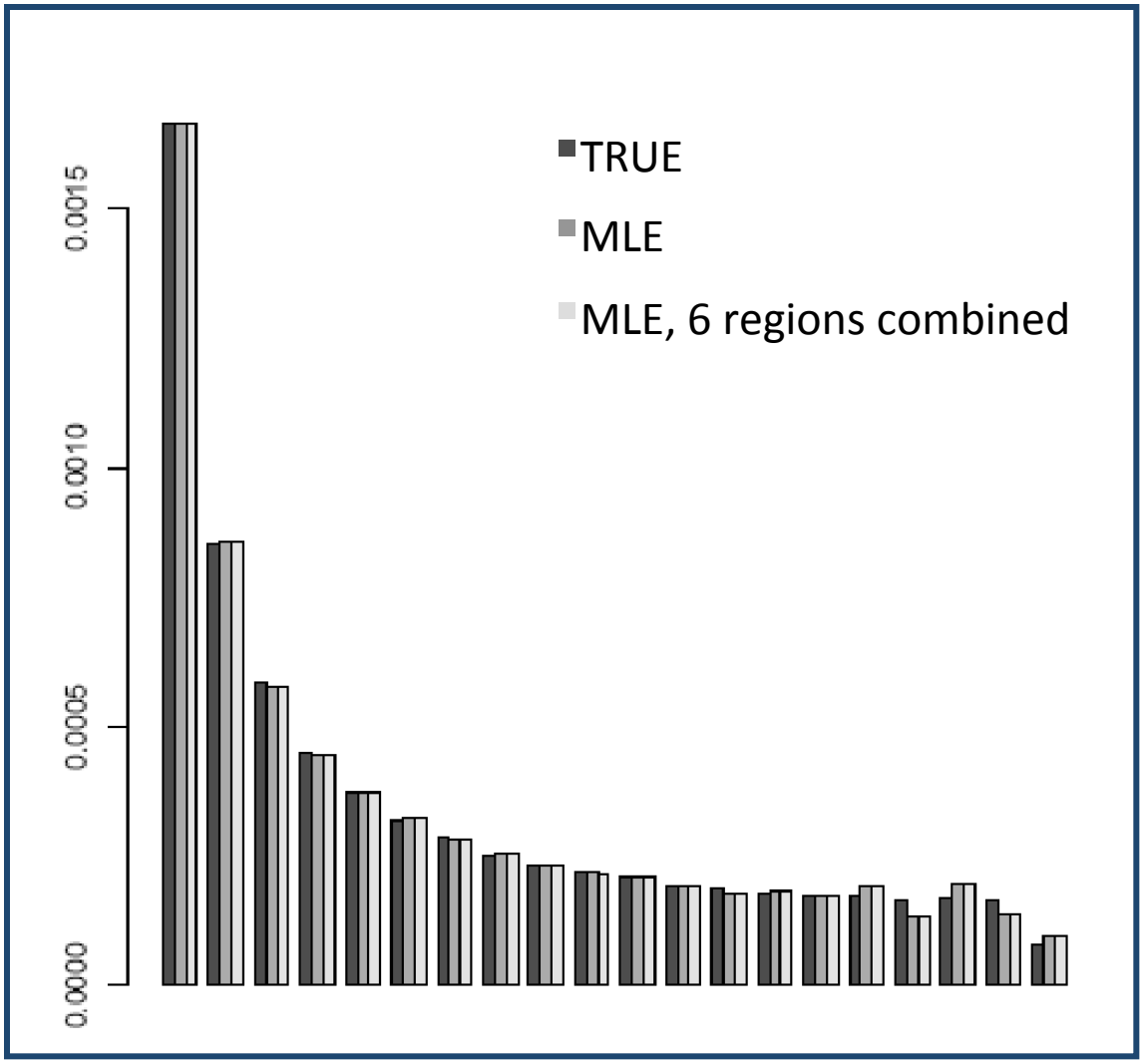
Summing across all unknown genotypes and multiplying the likelihood across sites.

- Likelihood function:

$$L(\bar{P}) = \prod_{v} \left(\sum_{j=0}^{2k} p_j \left[\sum_{G_1^{(v)}} \cdots \sum_{G_k^{(v)}} c(j, G^{(v)}) \prod_{d=0}^{k} p(X_d^{(v)} \mid G_k^{(v)}) \right] \right)$$

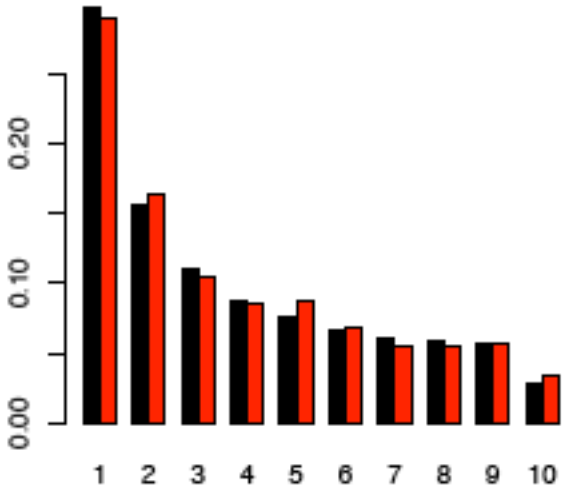
Nielsen et al. 2012 PLoS One

ML estimation of the SFS



Simulated 30Mb
Error rate of 0.3%
Mean depth of 5X

Mean depth of 1X:



R. Nielsen

Applications

...

- Model and non-model species
- Plants
- Ancient genomes
- …