Detecting selection: from low-depth data

Matteo Fumagalli

Outline

- Brief introduction to natural selection

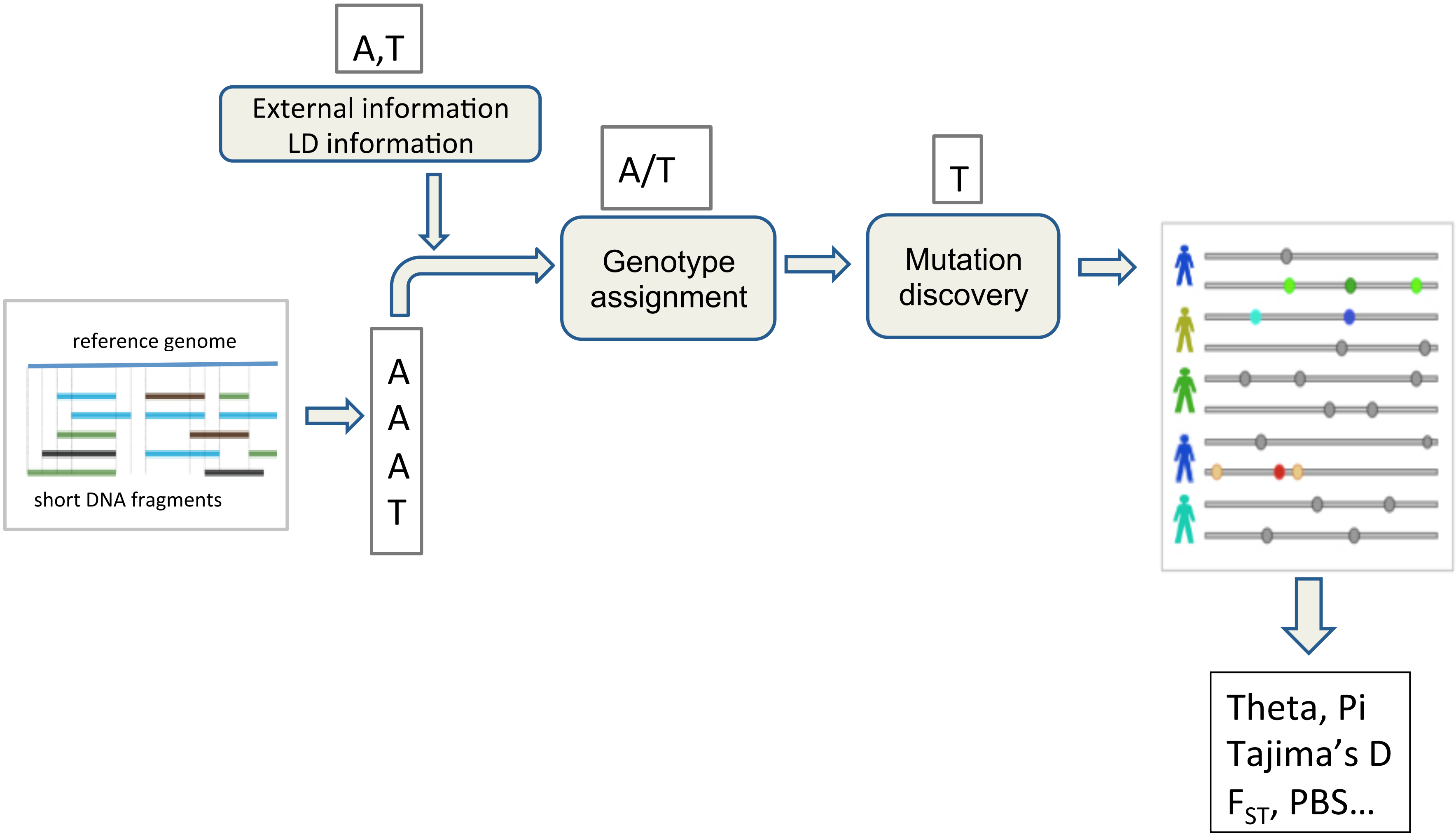
- Modes of selection

- Inferring selection at the intra-species level
  - Genetic differentiation
  - Haplotype variation
  - Model-based approaches
  - Testing for significance

- Inferring selection at the inter-species level

- Detecting selection from low-depth sequencing data
- Brief notes on optimal experimental design

Next Generation Sequencing data processing



A
A
A
T

A,T

External information
LD information

A/T

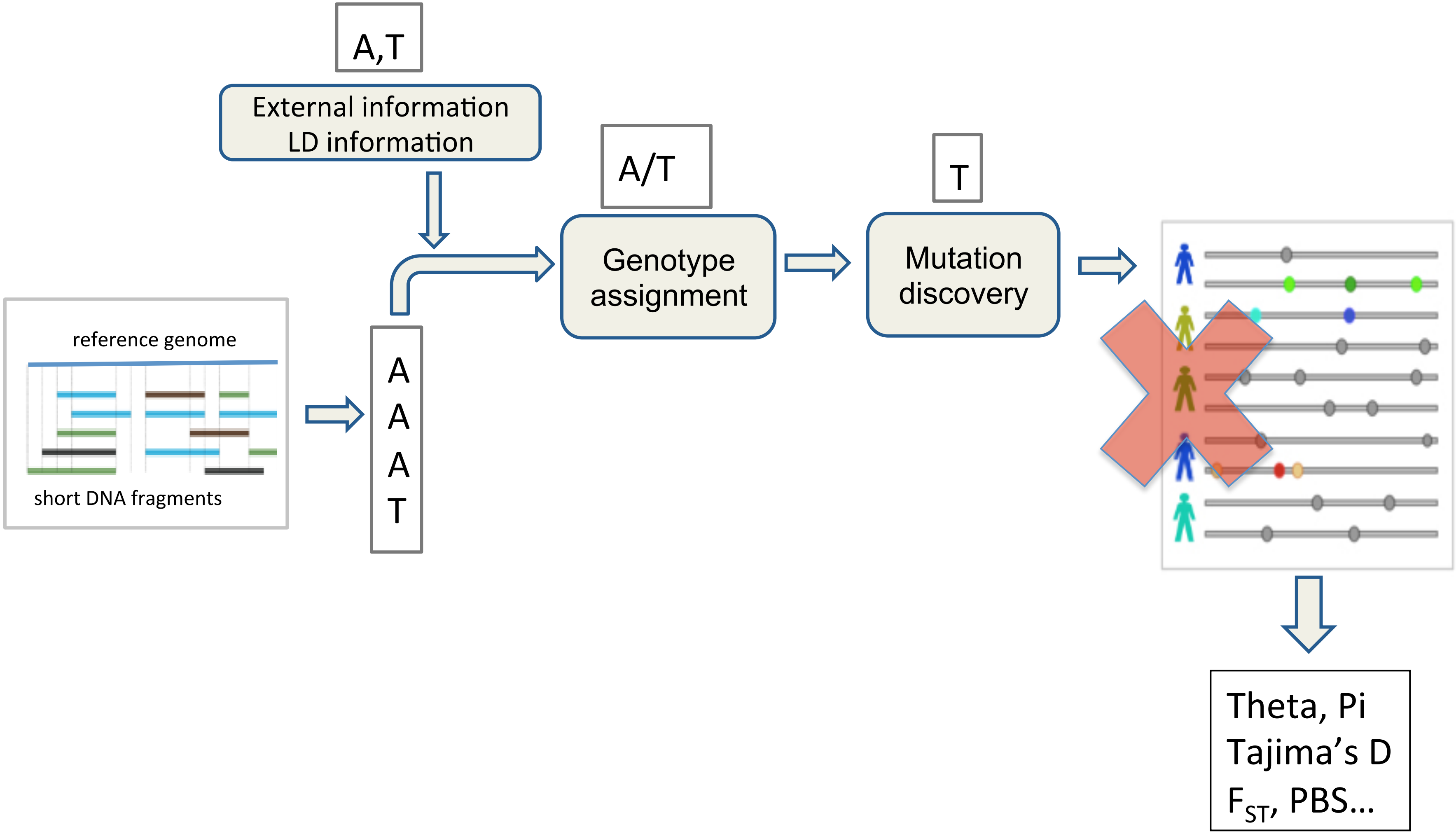
Genotype assignment

T

Mutation discovery

Theta, Pi
Tajima's D
F_{ST}, PBS…

Next Generation Sequencing data processing



A,T

External information
LD information

A/T

Genotype assignment

T

Mutation discovery

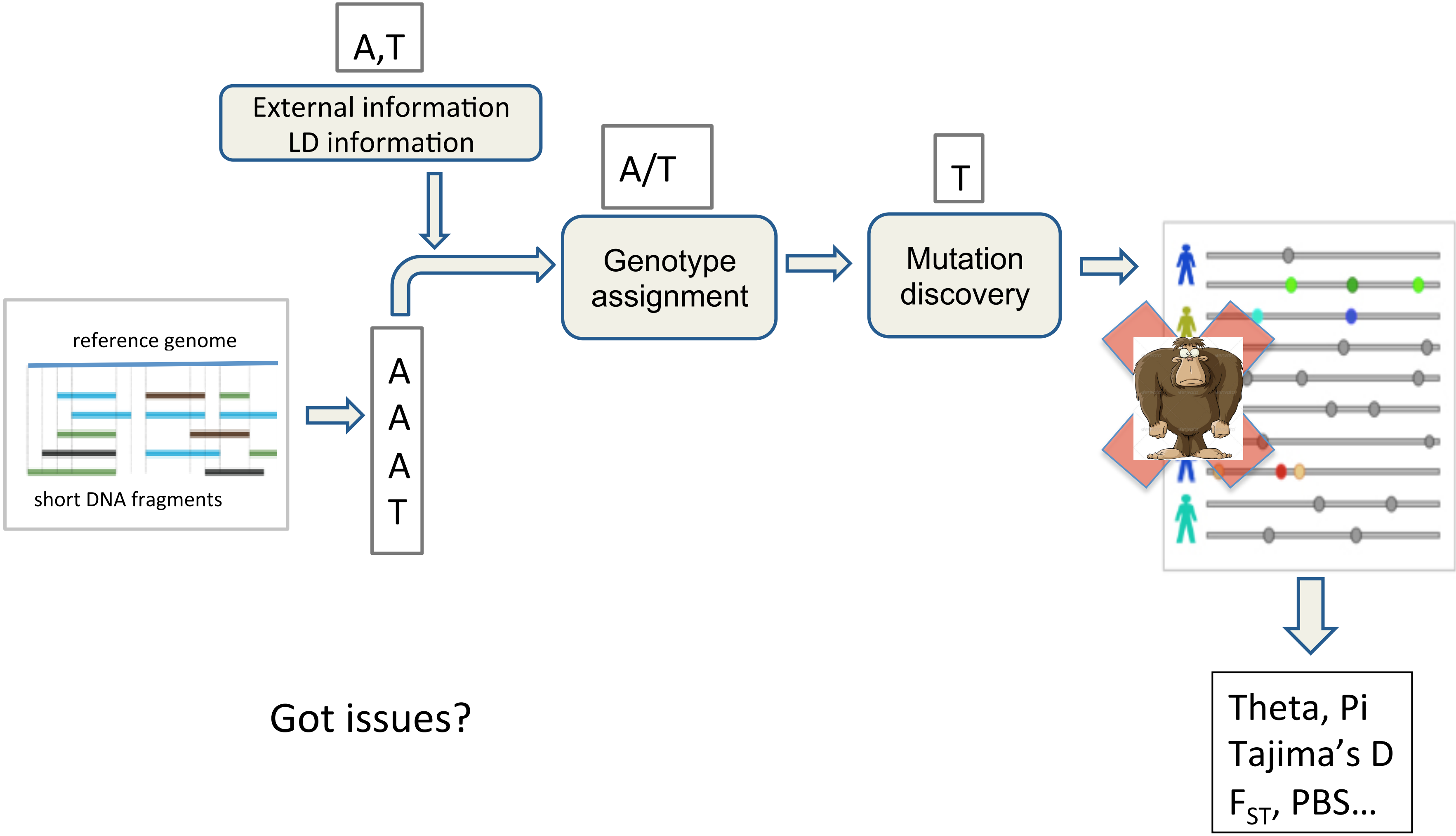
reference genome

short DNA fragments

A
A
A
T

Theta, Pi
Tajima's D
F_{ST}, PBS…

Next Generation Sequencing data processing in the non-model world



A,T

External information
LD information

A/T

Genotype assignment

T

Mutation discovery

reference genome

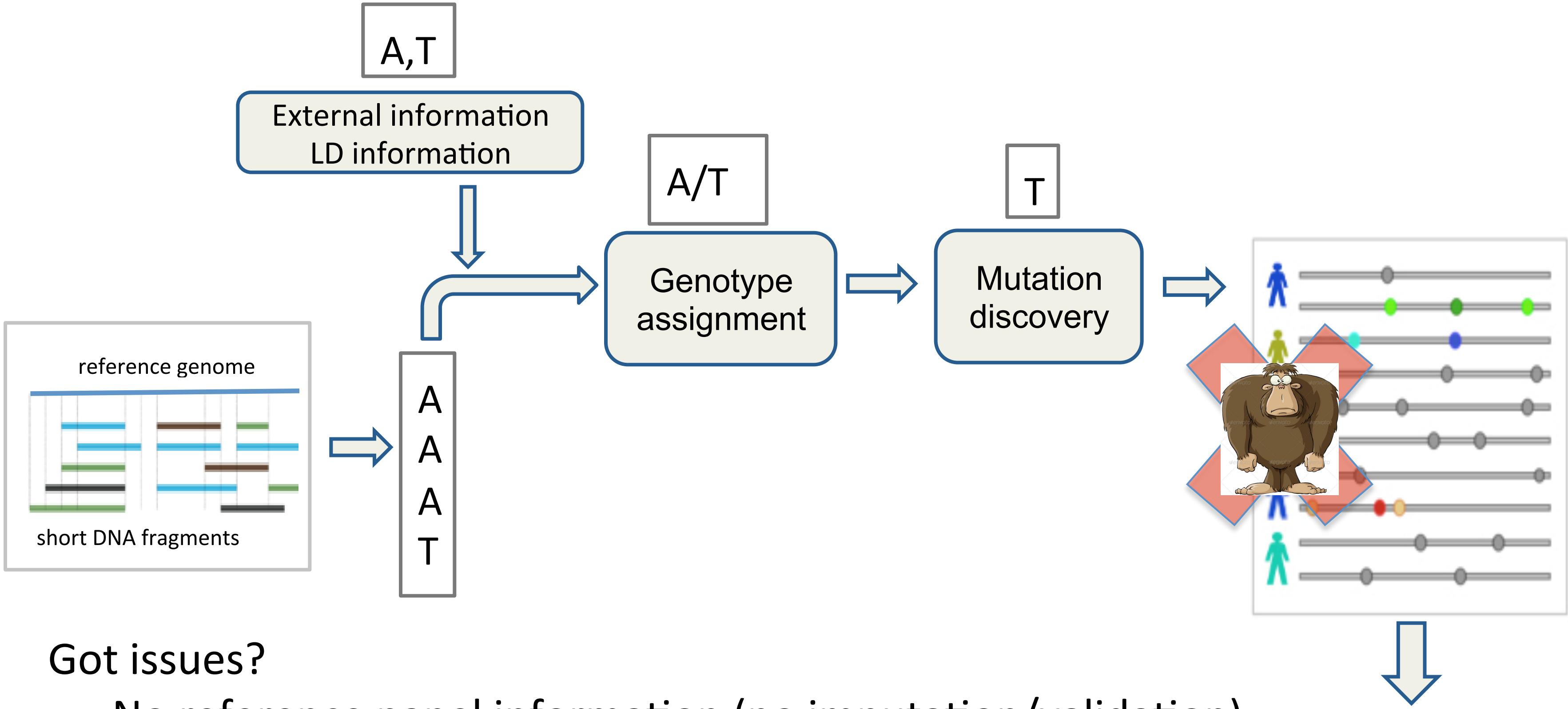
short DNA fragments

A
A
A
T

Theta, Pi
Tajima's D
F_{ST}, PBS…

Got issues?

Next Generation Sequencing data processing in the non-model world



A,T

External information
LD information

reference genome

short DNA fragments

A
A
A
T

A/T

Genotype assignment

T

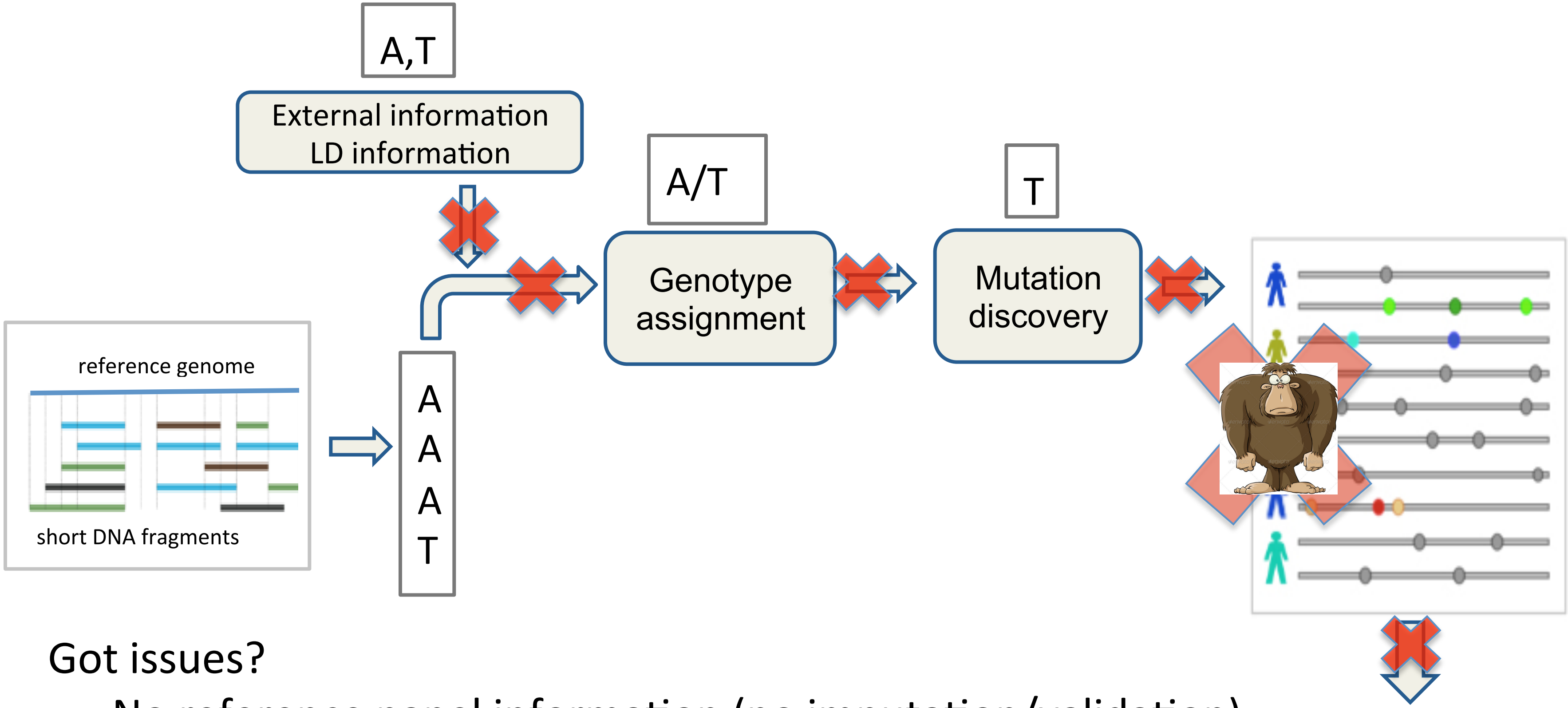
Mutation discovery

Theta, Pi
Tajima's D
F_{ST}, PBS…

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



A,T

External information
LD information

reference genome

short DNA fragments

A
A
A
T

A/T

Genotype assignment

T

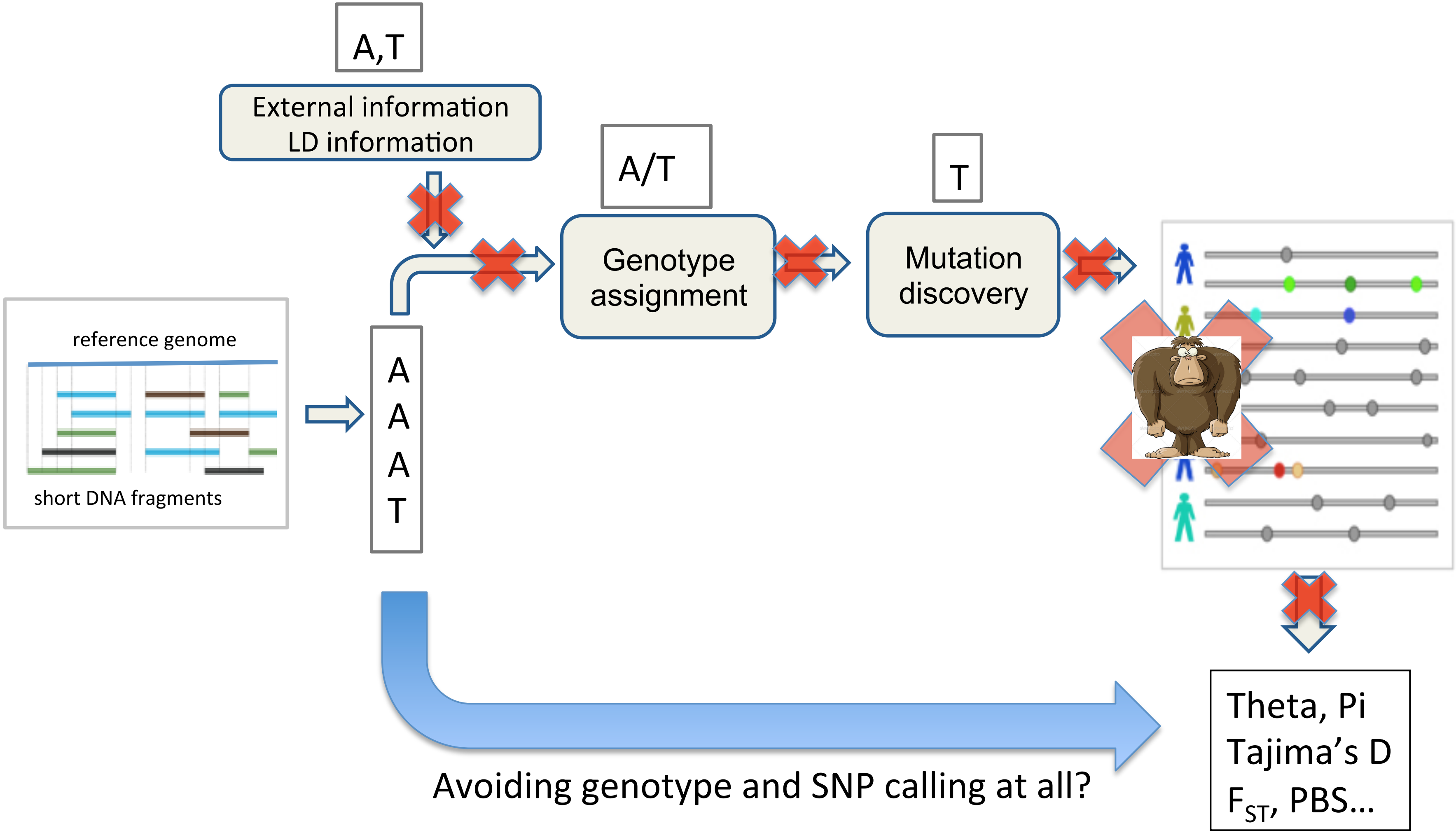
Mutation discovery

Theta, Pi
Tajima's D
F_{ST}, PBS…

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



A,T

External information
LD information

A/T

Genotype assignment

T

Mutation discovery

reference genome

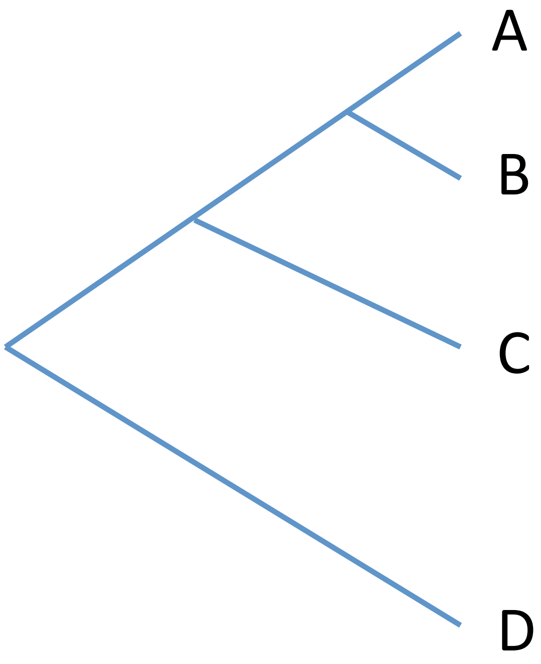
short DNA fragments

A
A
A
T

Theta, Pi
Tajima's D
F_{ST}, PBS…

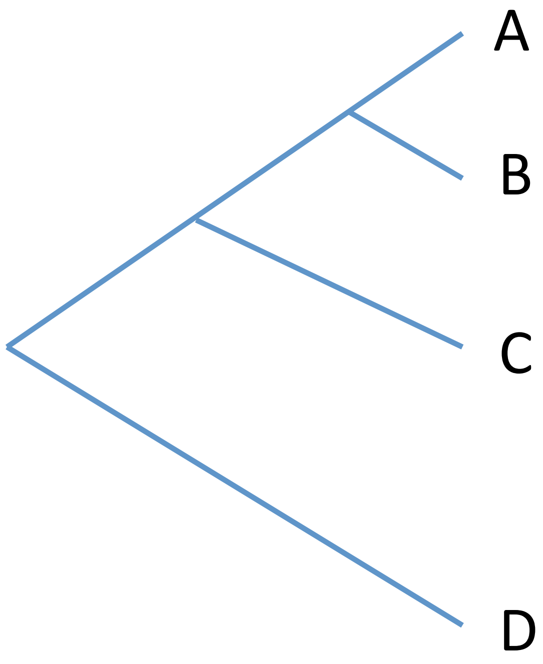
Avoiding genotype and SNP calling at all?

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:

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

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

B

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

Expected value

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

- Average value if you perform the same experiment many times

Genetic distances from (un)known genotypes

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

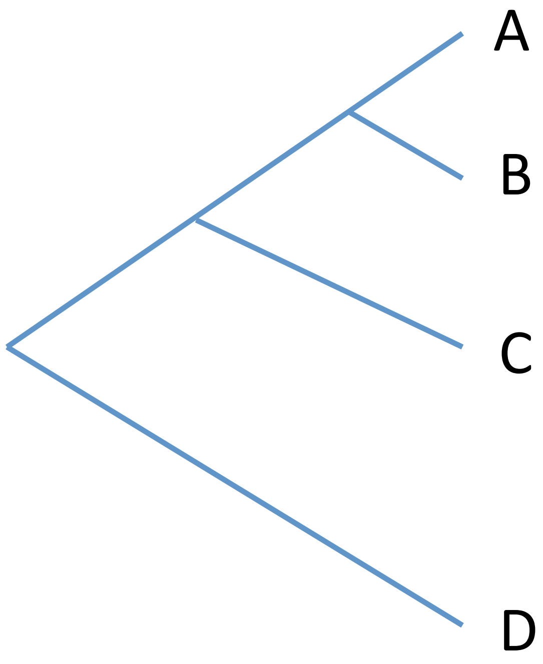
$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N} \frac{|g(i,s) - g(j,s)|}{2}\right)$$

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

B

		0	1	2
A	**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:

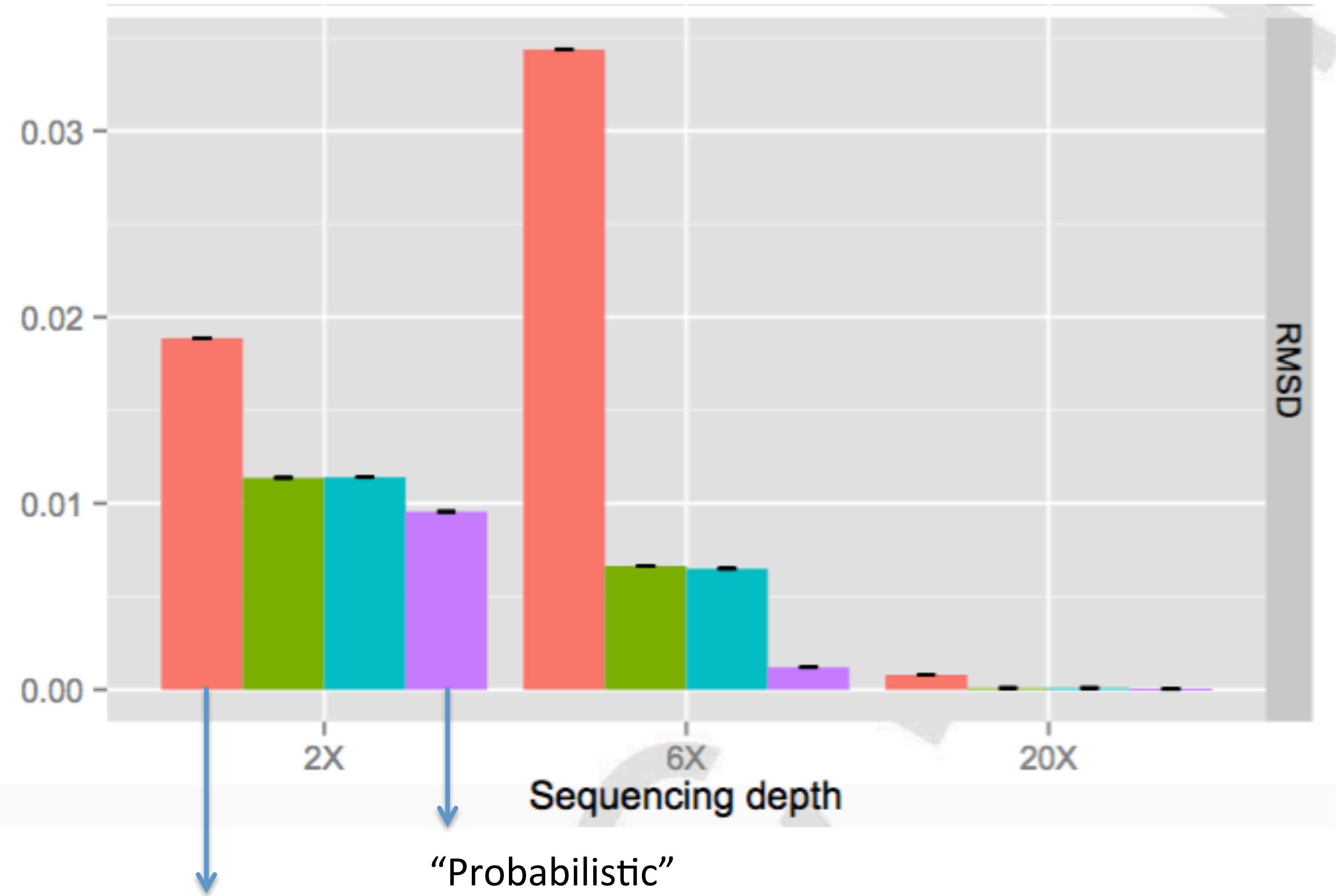
$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s)-g(j,s)|}{2}\right)$$

Iterate across all possible genotypes

Genotypes probability

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

Genetic distances from unknown genotypes



“Probabilistic”

Genotype calling (no prior)

Vieira et al. BJLS 2016

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

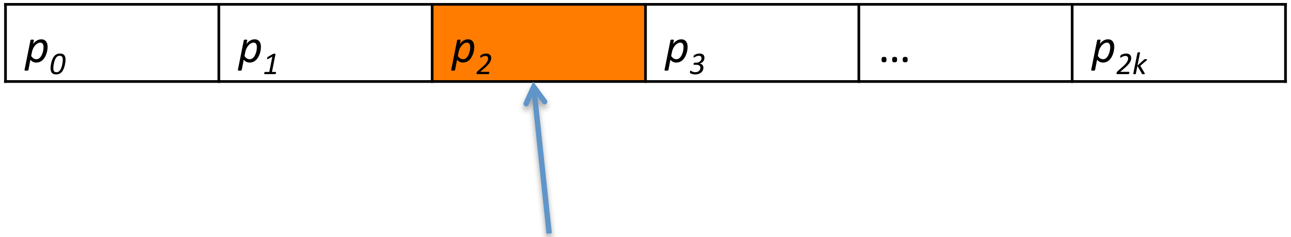
p_0	p_1	p_2	p_3	…	p_{2k}

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



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

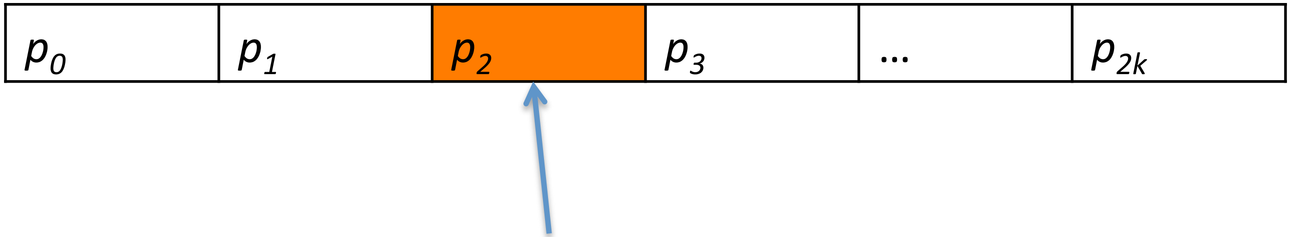
$p_0=0$	$p_1=0$	$p_2=1$	$p_3=0$	…	$p_{2k}=0$

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



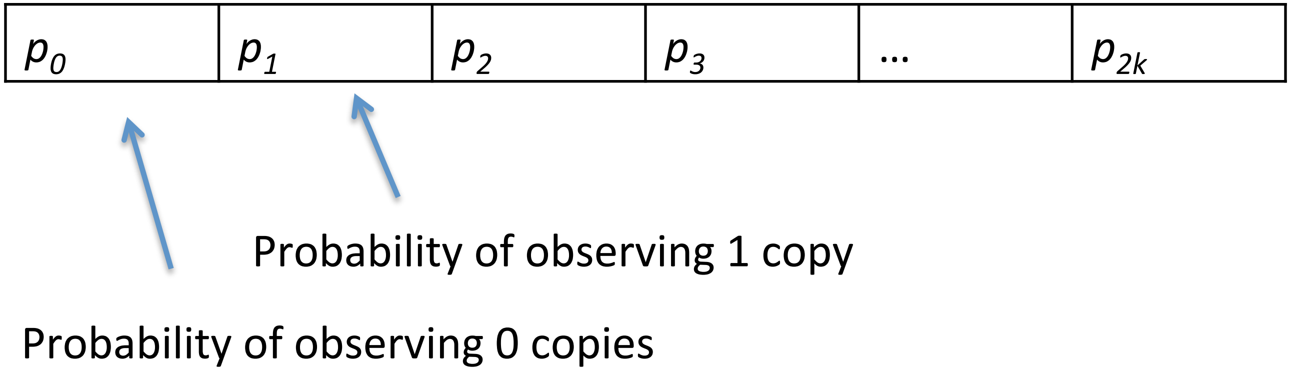
p_0	p_1	p_2	p_3	…	p_{2k}

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



p_0	p_1	p_2	p_3	…	p_{2k}

Probability of observing 1 copy

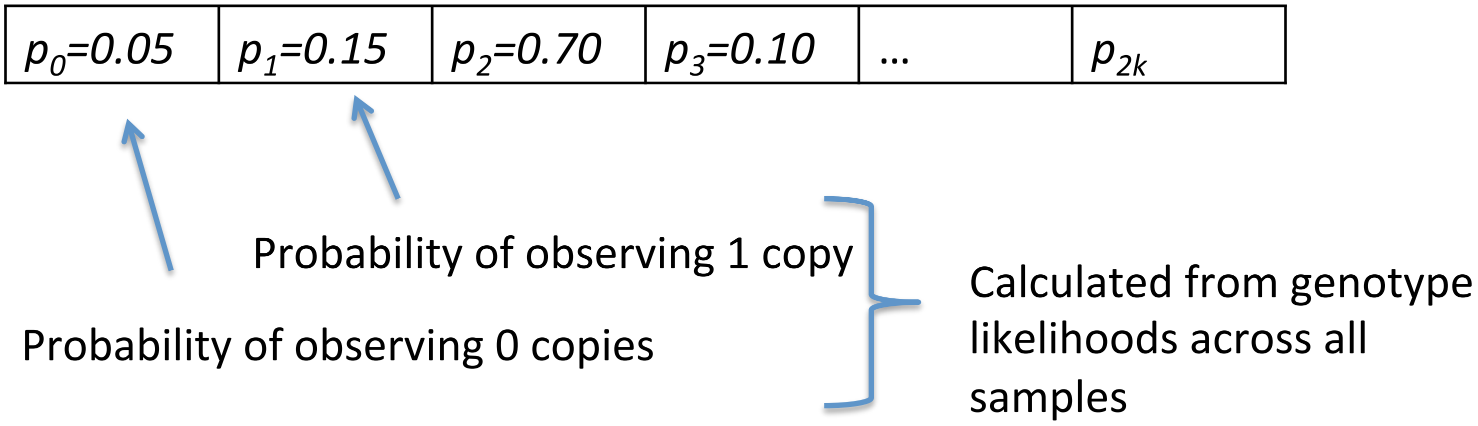
Probability of observing 0 copies

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



$p_0=0.05$	$p_1=0.15$	$p_2=0.70$	$p_3=0.10$	…	p_{2k}

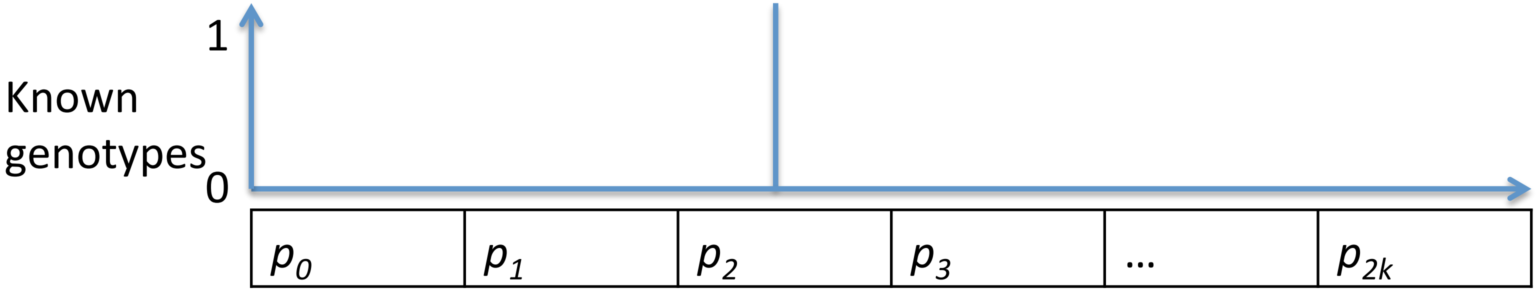
Probability of observing 1 copy

Probability of observing 0 copies

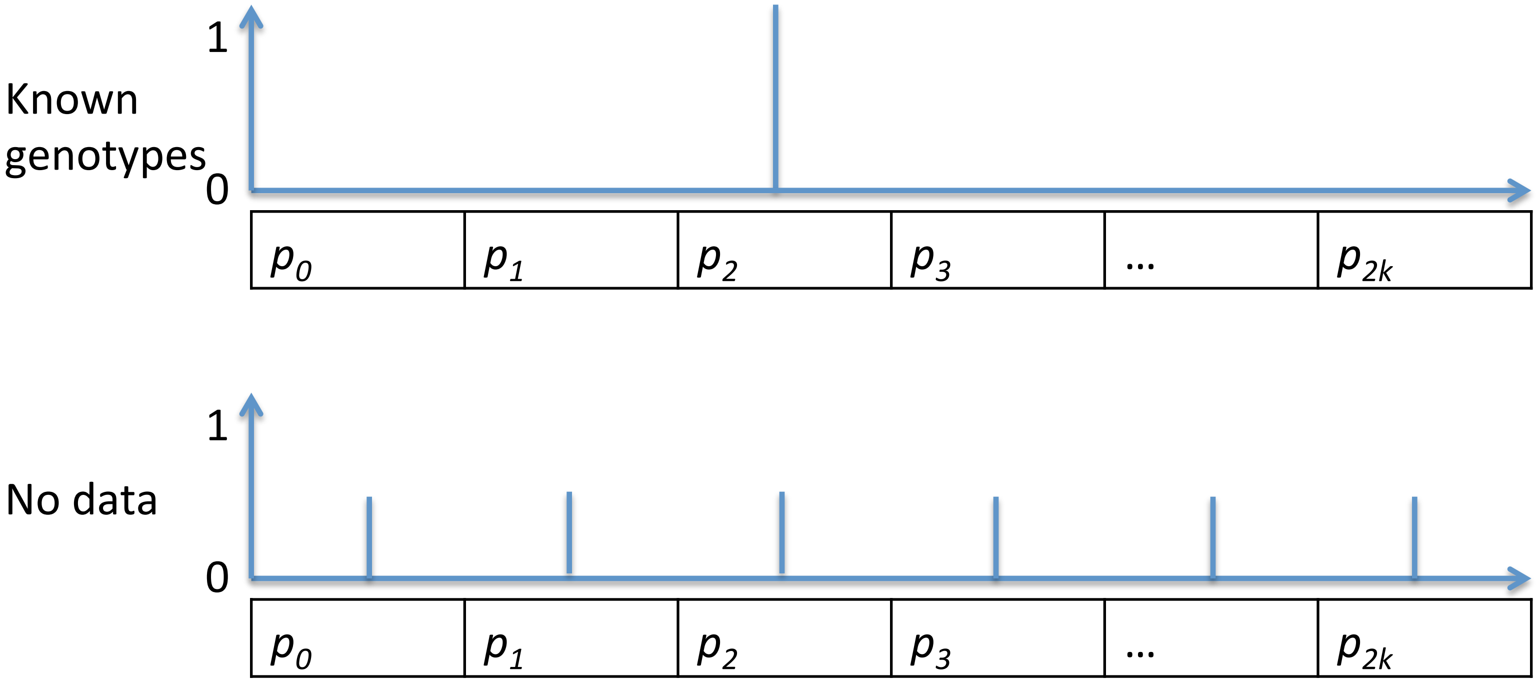
Calculated from genotype likelihoods across all samples

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

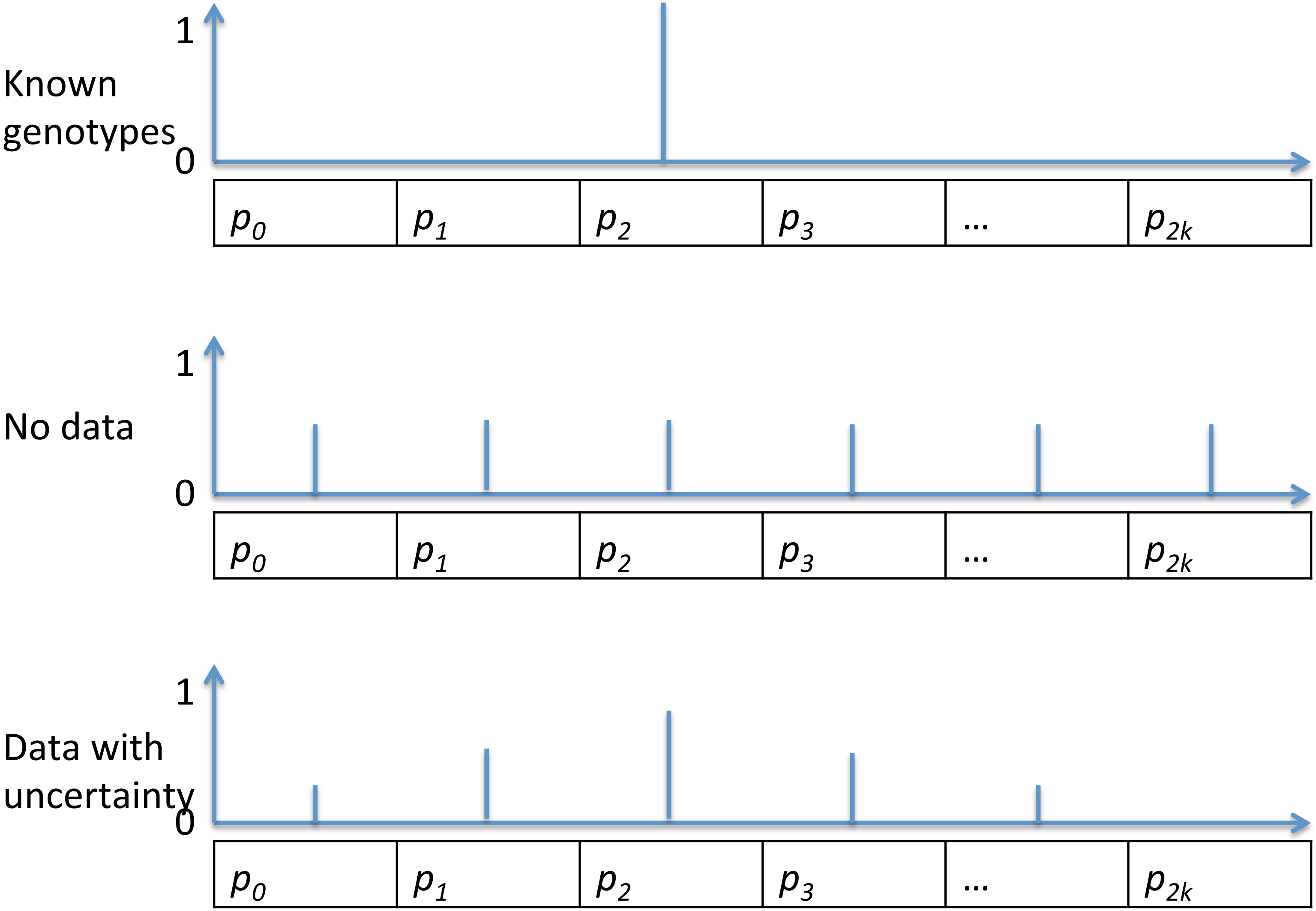
Sample allele frequency probabilities



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

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

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

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

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

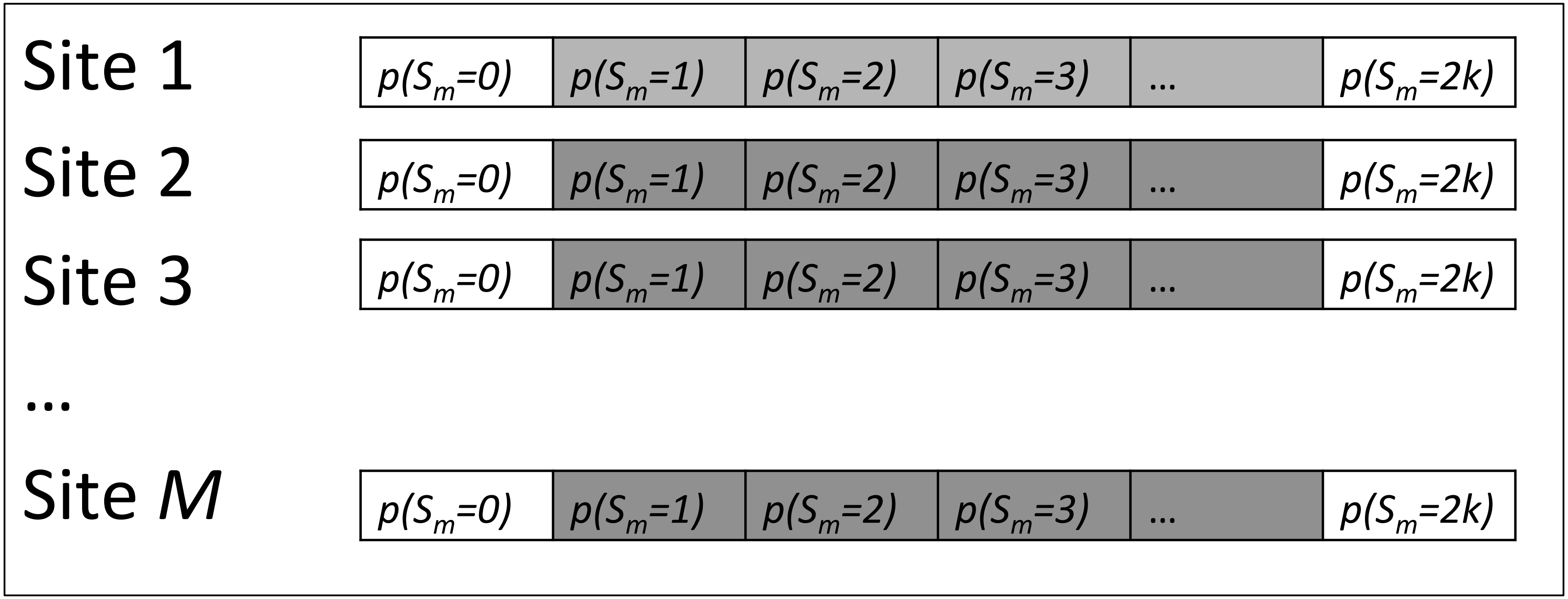
$$p_{\mathrm{var}} > t$$

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



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

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)

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

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

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

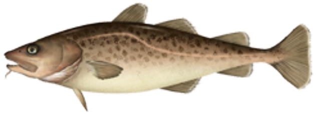
$$E[D] =$$

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

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

Applications



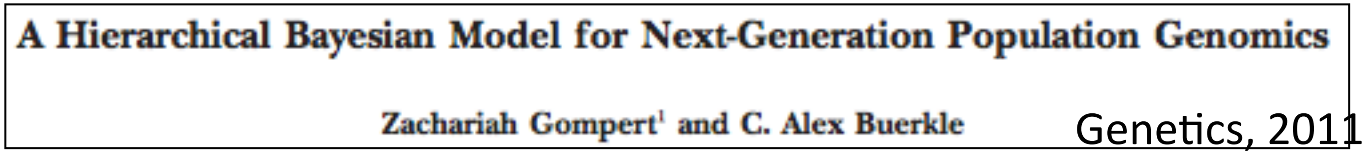
…

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

Software

Such advanced methods have been implemented in several software and utilities, such as:

- **ANGSD** (http://popgen.dk/ANGSD)
- **ngsTools** (https://github.com/mfumagalli/ngsTools)
- http://jnpopgen.org/software/



A Hierarchical Bayesian Model for Next-Generation Population Genomics

Zachariah Gompert[1] and C. Alex Buerkle

Genetics, 2011

which we will explore during the practical session.

Summary

- SNP calling should be performed including information from all samples (and inbreeding coefficient estimates, if relevant)

- Probabilistic methods for estimation of allele frequencies and statistics should be preferred (especially for mean sequencing depth < 20X)