Queen Mary
University of London

SNP/genotype calling and estimation of allele frequencies from NGS data

Matteo Fumagalli

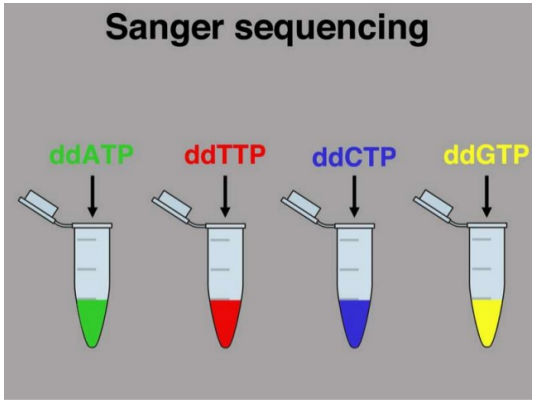
Intended Learning Outcomes

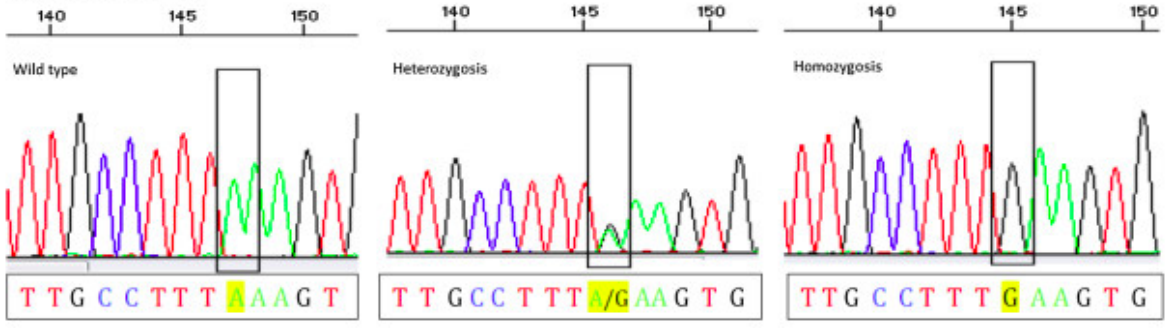
At the end of this session will be able to:

- appreciate the effect of sequencing depth and error to SNP/genotype calling
- calculate genotype and allele frequency likelihoods
- perform SNP/genotype calling from NGS data
- acknowledge the estimation of summary statistics from NGS data

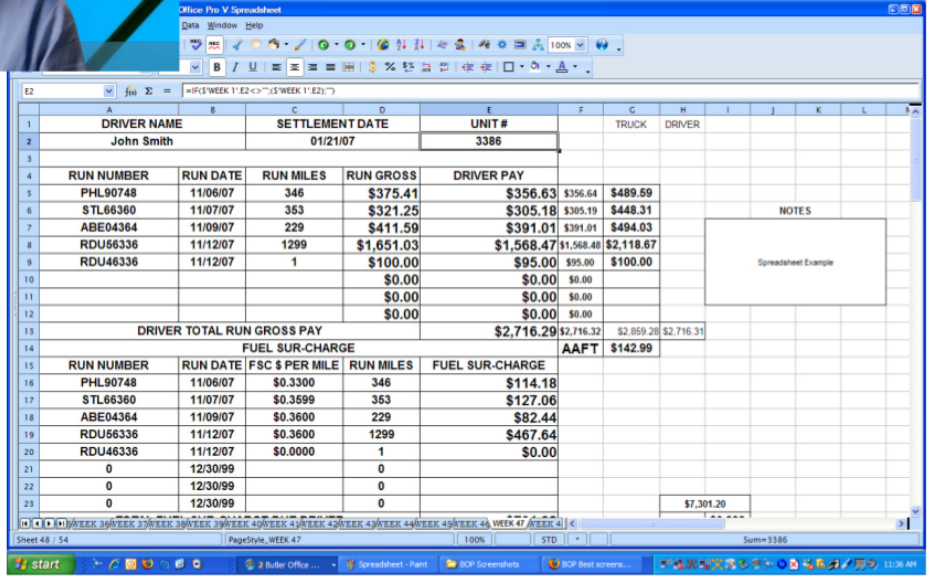
Sanger sequencing

aka first/former generation sequencing





Qualitative assignment of genotypes

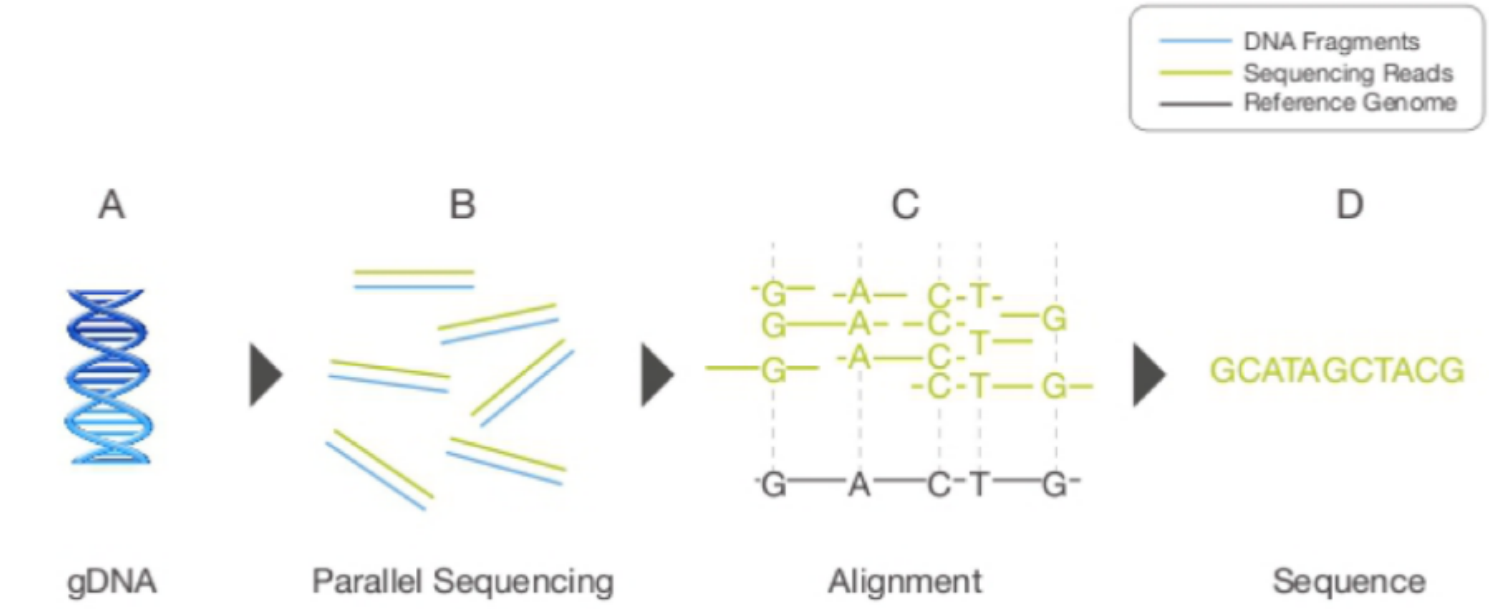


	A	B	C	D	E	F	G	H
1	DRIVER NAME		SETTLEMENT DATE		UNIT #		TRUCK	DRIVER
2	John Smith		01/21/07		3386			
3								
4	RUN NUMBER	RUN DATE	RUN MILES	RUN GROSS	DRIVER PAY			
5	PHL90748	11/06/07	346	$375.41	$356.63	$356.64	$489.59	
6	STL66360	11/07/07	353	$321.25	$305.18	$305.19	$448.31	
7	ABE04364	11/09/07	229	$411.59	$391.01	$391.01	$494.03	
8	RDU56336	11/12/07	1299	$1,651.03	$1,568.47	$1,568.48	$2,118.67	
9	RDU46336	11/12/07	1	$100.00	$95.00	$95.00	$100.00	
10				$0.00	$0.00	$0.00		
11				$0.00	$0.00	$0.00		
12				$0.00	$0.00	$0.00		
13	DRIVER TOTAL RUN GROSS PAY				$2,716.29	$2,716.32	$2,859.28	$2,716.31
14	FUEL SUR-CHARGE					AAFT	$142.99	
15	RUN NUMBER	RUN DATE	FSC $ PER MILE	RUN MILES	FUEL SUR-CHARGE			
16	PHL90748	11/06/07	$0.3300	346	$114.18			
17	STL66360	11/07/07	$0.3599	353	$127.06			
18	ABE04364	11/09/07	$0.3600	229	$82.44			
19	RDU56336	11/12/07	$0.3600	1299	$467.64			
20	RDU46336	11/12/07	$0.0000	1	$0.00			
21	0	12/30/99		0				
22	0	12/30/99		0				
23	0	12/30/99		0				$7,301.20

NOTES

Spreadsheet Example

Next Generation Sequencing



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.

From genomes to variants

Genome (FASTA)

```
>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary
reference assembly
TGGAAGAGGCCTCAGCAGGCCCAGGCCACCTGGAGGGAGAGCAGACCTGCGGCTGAGGATGCAGGGCTCC
CGGGCACGGTGCTAGCCCTGCCTTGAGACACCCCGAGAGCTGTGGGAAGAGCTGTGGGATCCCCTATTGC
ATCACAAAGCGGCCCTGGAGGGCTGGTCTTTATTTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT
TAATCCGCACGCTTTAGACTCCCCGGCTGTGATTTTTGACAATGGCTCGGGGTTCTGCAAAGCGGGCCTG
TCTGGGGAGTTTGGACCCCGGCACATGGTCAGCTCCATCGTGGGGCACCTGAAATTCCAGGCTCCCTCAG
```

Reads (FASTQ)

```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

Mapped Reads (mpileup, BAM)

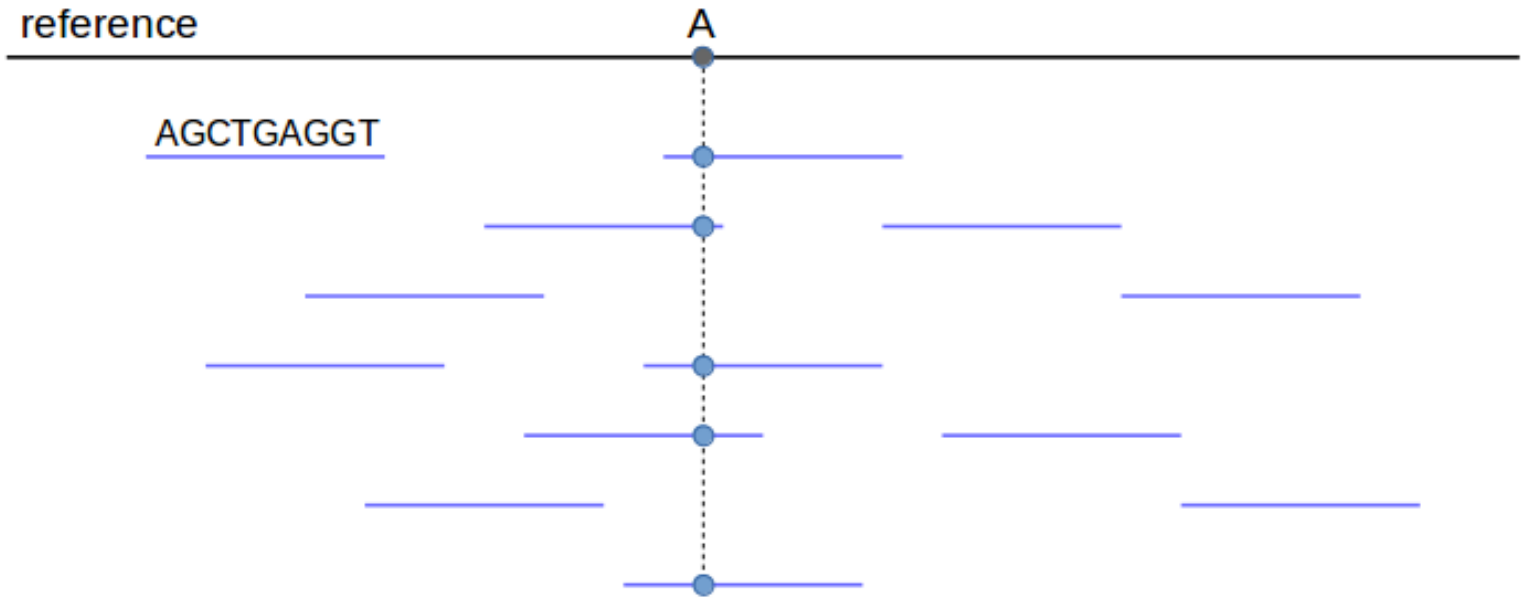
```
seq1 272 T 24  ,.$.....,,,,,,,,,,,,,,,.^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,,,,,,,,,,,,,,.A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$.....,,,,,,,,,,,,,,,,,,    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$.....,,,,,,,,,,,,,,,,,,.^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,,,,,,,,,,,,,,,,,,,  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ...,,,,,,,,,C.,,,,,,,,G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  .....,,,,,,,,,,,,,,,,,,,.^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,,,,,,,,,,,,,,,,,,,, ;75&<<<<<<<<<=<<<9<<:<<
```

Variants (VCF)

```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23andme2vcf
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  GENOTYPE
chr1    82154   rs4477212       a       .       .       .       GT      0
/0
chr1    752566  rs3094315       g       A       .       .       .       GT      1
/1
chr1    752721  rs3131972       A       G       .       .       .       GT      1
/1
chr1    798959  rs11240777      g       .       .       .       .       GT      0
/0
chr1    800007  rs6681049       T       C       .       .       .       GT      1
/1
```

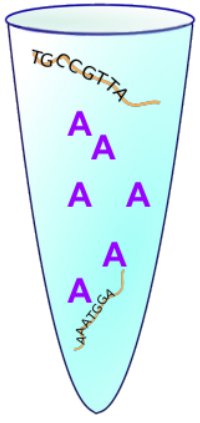
Today's starting point

Data: collection of sequenced nucleotides in **one genomic position** for **one diploid individual**

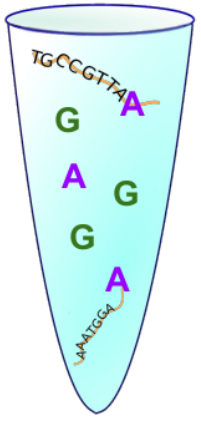


- is a **nucleotide**/base/allele with a certain **quality** score

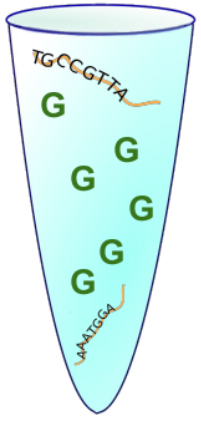
From sequenced nucleotides to alleles



The library for an individual homozygous for the **A** allele will consist only of **A**s.



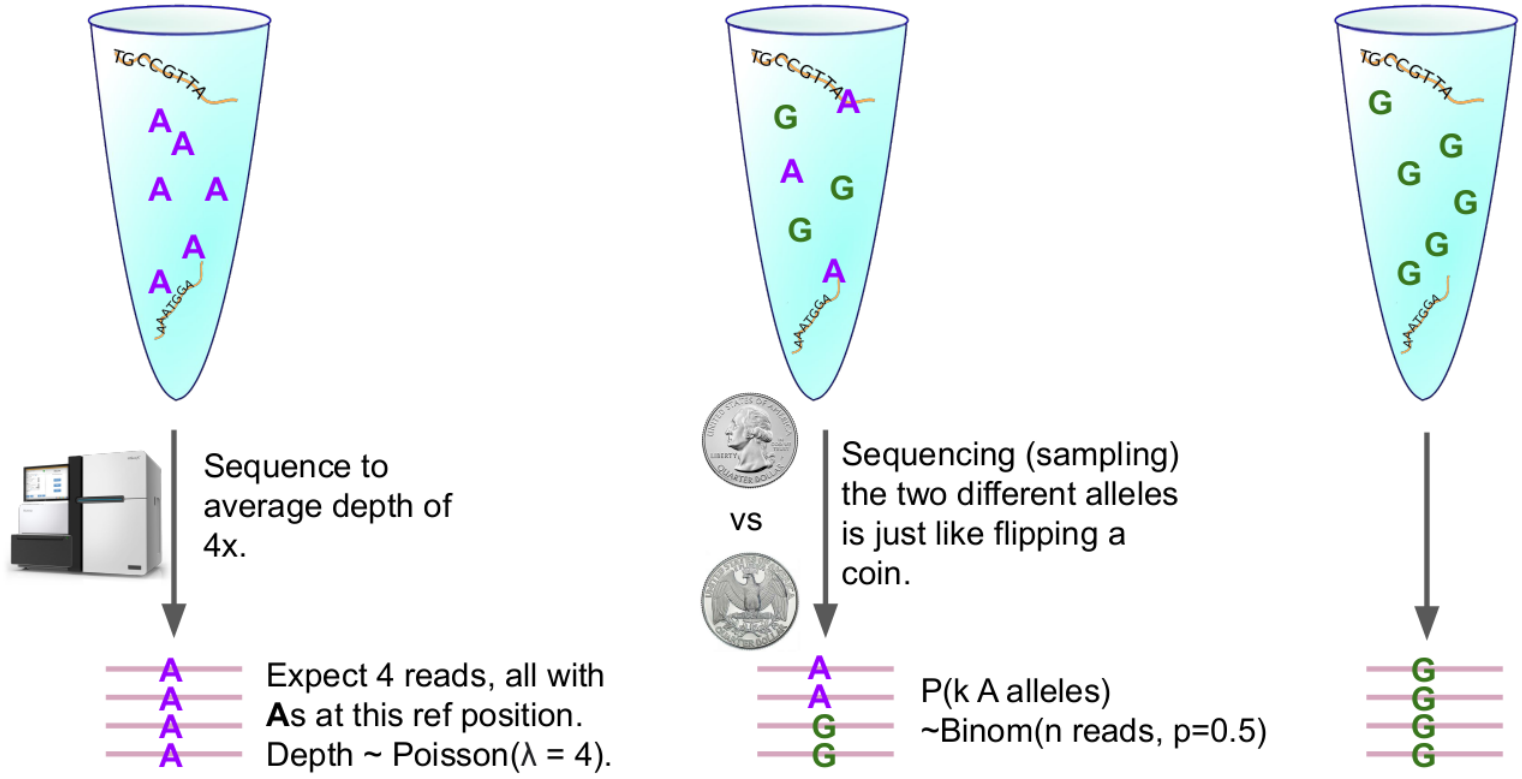
The library for a heterozygous individual at a site contains both **A**s and **G**s.



The library for an individual homozygous for the **G** allele consist only of **G**s.

(slide stolen from Tyler)

Sampling nucleotides



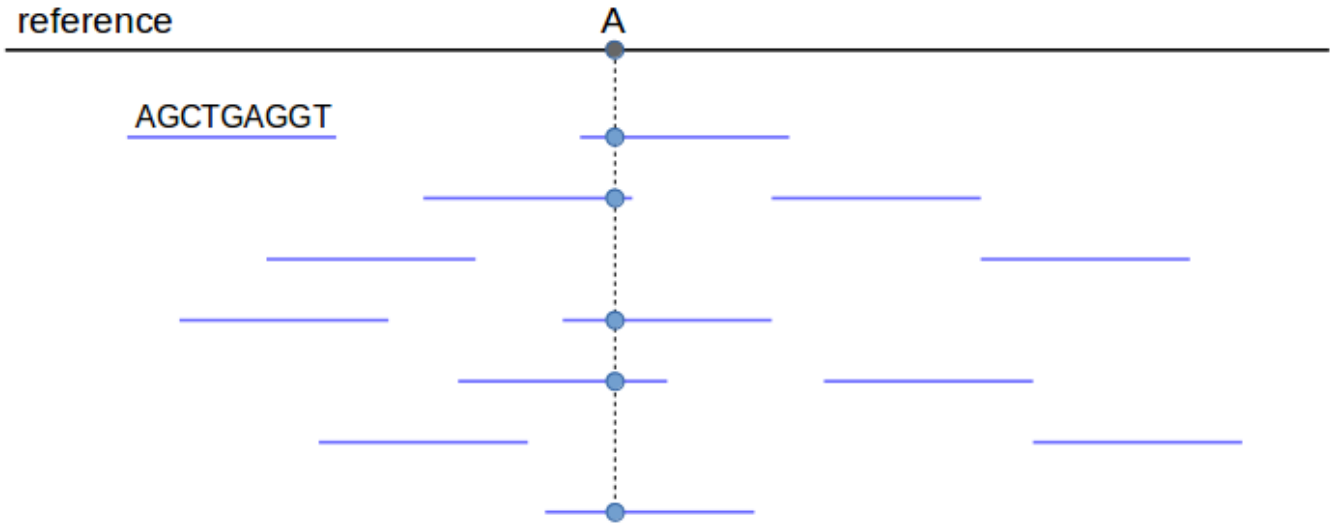
(slide stolen from Tyler)

Whiteboard + R

- What is the expected value of flipping a coin (sampling alleles)?
- What is the effect of depth?
- Any other factors affecting our sampling?

(Matteo: use whiteboard and R)

Given a possible genotype, what is the probability of observing this NGS data?



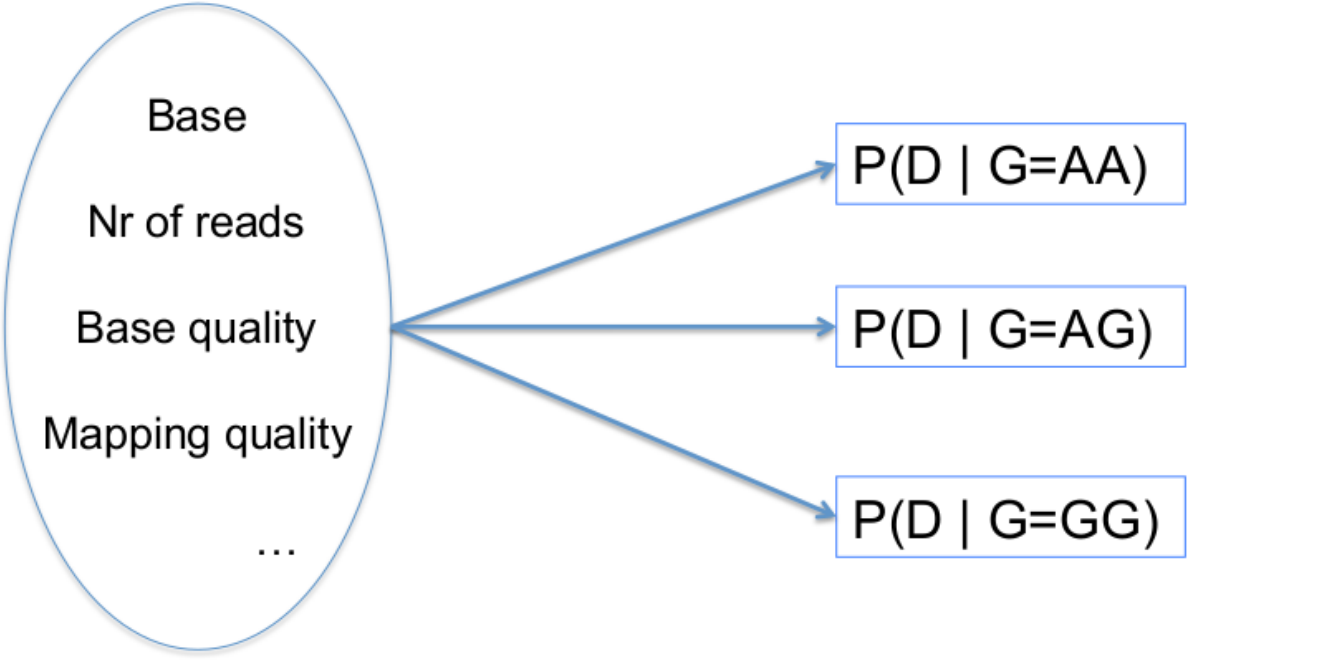
- is a **nucleotide**/base/allele with a certain **quality** score

How many genotypes likelihoods do we need to calculate for each each diploid individual at each site?

Genotype likelihoods - rationale

Chrom1 272 A 24 AAAAAGGAGAGGTAAG <<<+;<<<<<<<<<<=<;<;7<&

Base quality in Phred scale



(Matteo: whiteboard)

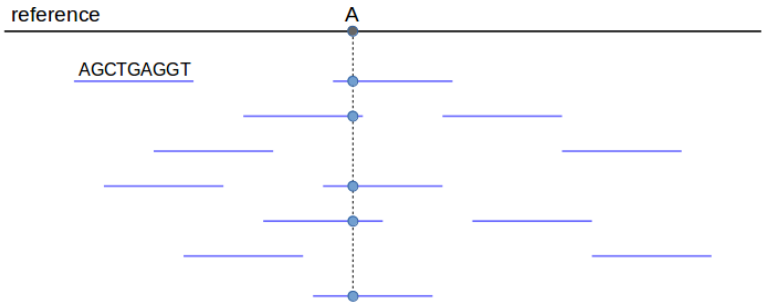
Genotype likelihoods - calculation

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

- $L_{A_j,i} = P(D|A_G = A_j)$
- $A_i \in \{A, C, G, T\}$
- R is the depth (nr. of reads)
- N is the ploidy level (nr. of chromosomal copies)

Genotype likelihoods - example

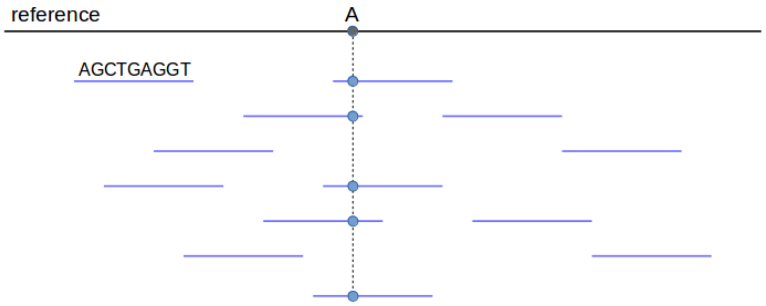


- is a **nucleotide**/base/allele with a certain **quality** score

A
A
A
G

with all with quality scores equal to 20 (in phred score)

Genotype likelihoods - example



- is a **nucleotide**/base/allele with a certain **quality** score

A
A
A
G

with all with quality scores equal to 20 (in phred score)

What is $P(D|G = AC) =?$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20

$$P(D|G = \{A, C\}) = ...$$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20
$N = 2; i = 1; A_1 = A; A_2 = C$

$$P(D|G = \{A, C\}) = (\frac{L_{A,1}}{2} + \frac{L_{C,1}}{2}) \times ...$$

What are $L_{A,1}$ and $L_{C,1}$?

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,1} =$$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,1} = \frac{\epsilon}{3}$$

$$L_{A,1} =$$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,1} = \frac{\epsilon}{3}$$

$$L_{A,1} = 1 - \epsilon$$

$$P(D|G = \{A, C\}) = (\frac{1-\epsilon}{2} + \frac{\epsilon}{6}) \times ...$$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,4} =$$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,4} = \frac{\epsilon}{3}$$

$$L_{A,4} =$$

Genotype likelihoods - example

Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,4} = \frac{\epsilon}{3}$$

$$L_{A,4} = \frac{\epsilon}{3}$$

$$P(D|G = \{A, C\}) = (\frac{1-\epsilon}{2} + \frac{\epsilon}{6})^3 \times \frac{\epsilon}{3}$$

What is ϵ?

Genotype likelihoods - example

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

A

A

A

G

$\epsilon = 0.01$

Genotype calling

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

AAAG & $\epsilon = 0.01$

What is the genotype here?

Genotype calling

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

AAAG & $\epsilon = 0.01$ What is the genotype? AG.

Maximum Likelihood

The simplest genotype caller: choose the genotype with the highest likelihood.

Major and minor alleles

Likelihood function

$$\log P(D|G = A) = \sum_{i=1}^{R} \log L_{A_j,i}$$

AAAG & $\epsilon = 0.01$

Allele	Likelihood
A	**-2.49**
C	-3.38
G	**-1.22**
T	-3.38

We can reduce the genotype space to 3 entries (from 10, for diploids).

Genotype calling

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

Genotype	Likelihood
AA	-5.73
AG	-2.80
GG	-17.12

At what extent is the data affecting the called genotype and its **confidence**? (Matteo: examples using ngsJulia in jupyter-notebook

Genotype likelihood 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?

- Yes:

Genotype likelihood 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?

- Yes: genotype are called with higher **confidence**
- No:

Genotype likelihood 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?

- Yes: genotype are called with higher **confidence**
- No: more **missing** data

Practical: genotype likelihoods

`https://github.com/mfumagalli/Copenhagen`

The monster dilemma

Figure 1: Nessie, the Loch Ness Monster. Truth or hoax?

"Eyes" thinking

What's "wrong"?
Our inference on N, our parameter, is driven solely by our observations, given by our likelihood function (data).

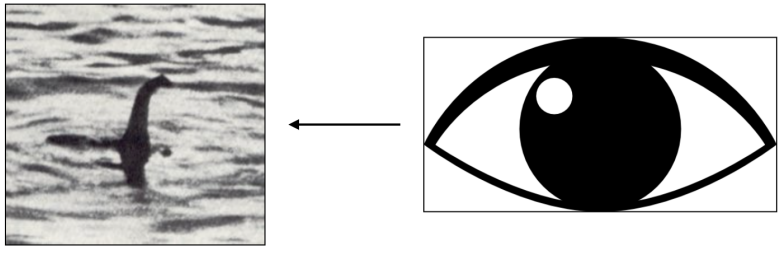


Figure 2: The eye: a "likelihood" organ.

"Blind Brain" thinking

In real life we take many decisions based not only on what we observe but also on some "blind" believes of ours*.

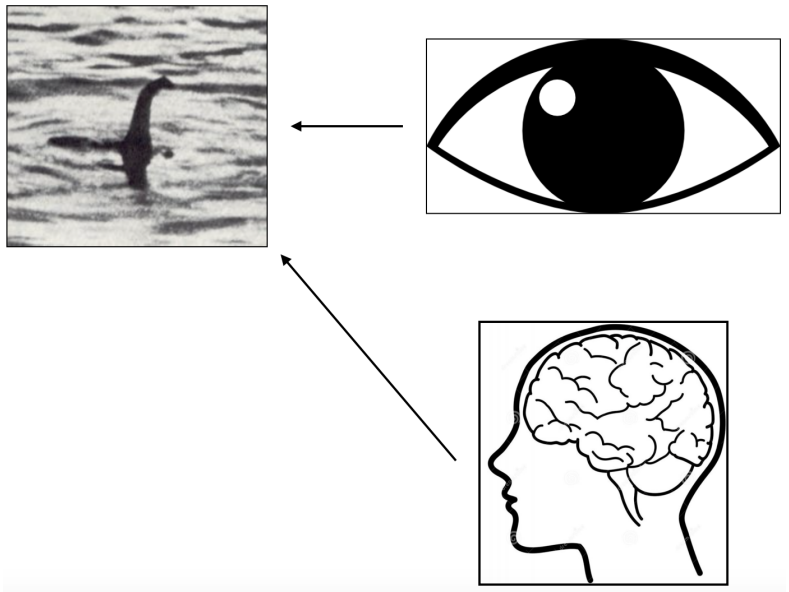


Figure 3: The brain: a "non-likelihood" organ.

* unfortunately in many cases

"Eyes + Blind Brain" thinking

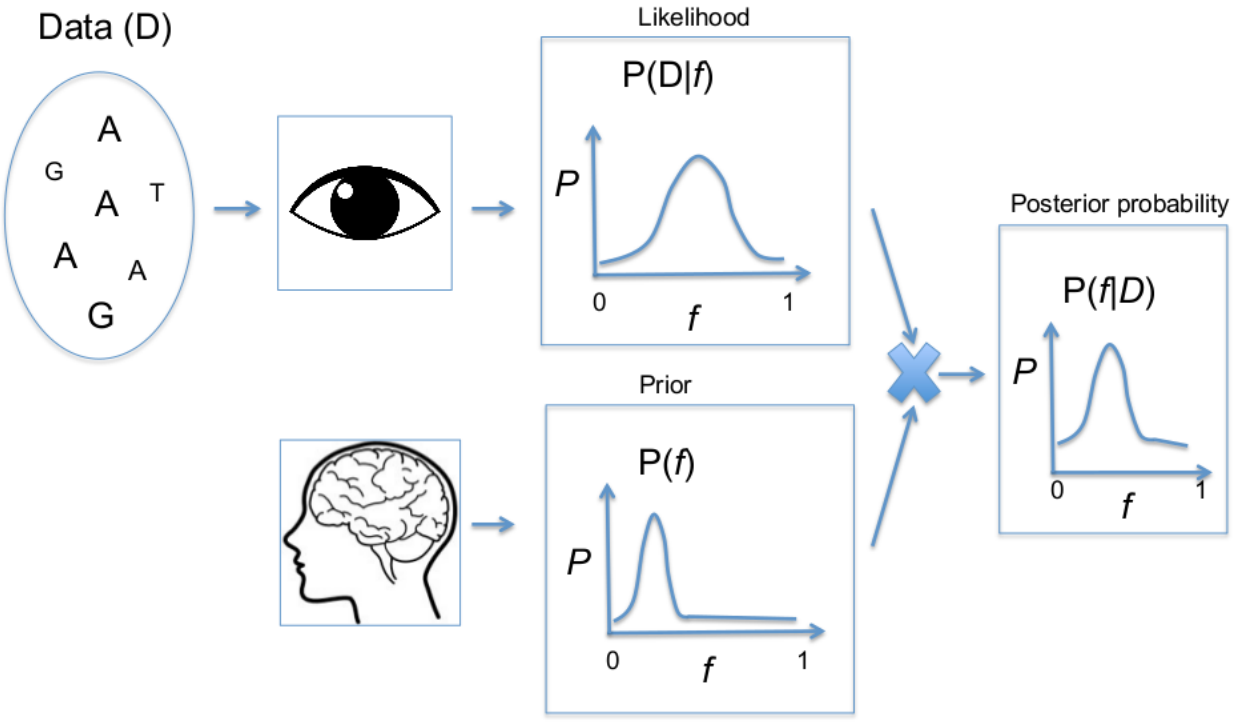
The "belief" function is called **prior probability** and the joint product of the likelihood and the prior is proportional to the **posterior probability**.

The use of posterior probabilities for inferences is called Bayesian statistics.

Bayesian *vs.* Likelihoodist

- we obtain legitimate probability distributions of our parameters rather than point estimates
- a probability is assigned to a hypothesis rather than a hypothesis is tested
- we can "accept" the null hypothesis rather than "fail to reject" it

Bayesian inference



Bayes' Theorem

$$p(G|D) = \frac{f(D|G)\pi(G)}{\int f(D|G)\pi(G)dG}$$

- G is not a fixed parameter but a random quantity with prior distribution $\pi(G)$
- $p(G|D)$ is the posterior probability distribution of G
- $\int p(G|D)dG = 1$

Genotype posterior probability

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

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73		

Genotype posterior probability

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

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	1/3	

Genotype posterior probability

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

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

Genotype posterior probability

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		

Genotype posterior probability

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

What is the called genotype? What's its confidence?

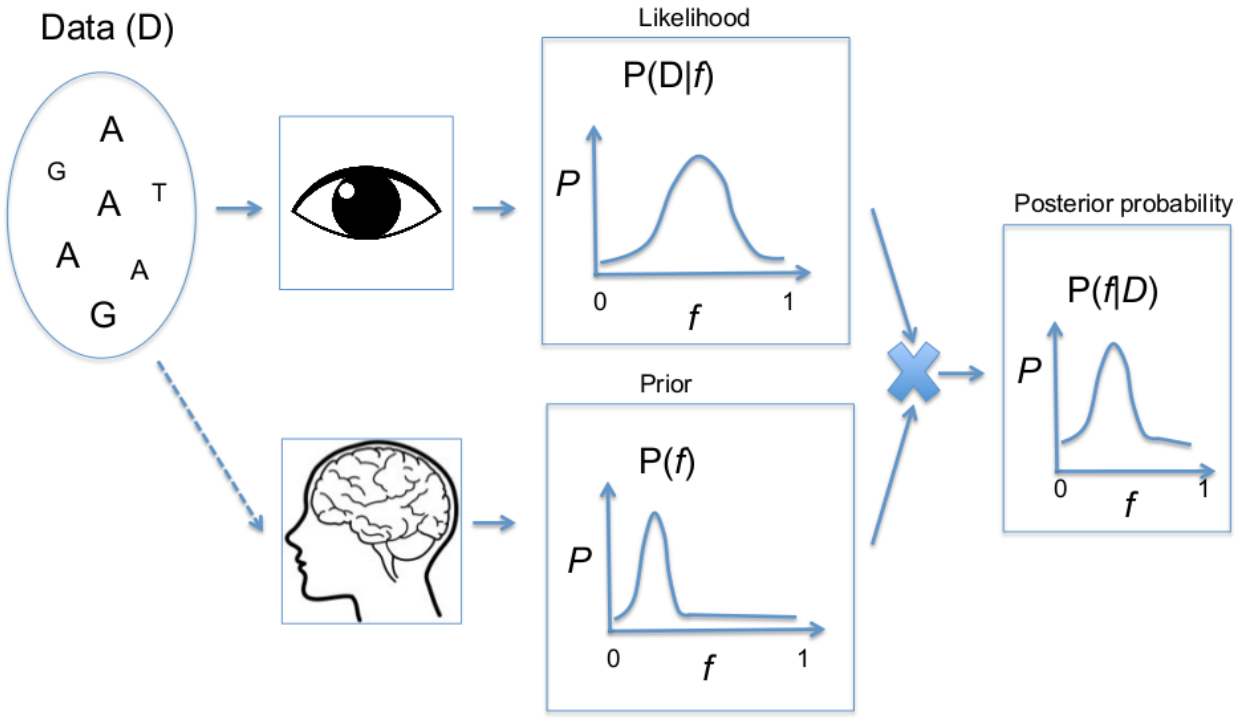
Genotype posterior probability

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

What is the called genotype? What's its confidence?
Only call genotypes if the largest probability is above a certain threshold (e.g. 0.95).

"Eyes + non-Blind Brain" inference



Empirical Bayesian

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from the data itself

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73		

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from the data itself

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

- if the assumption of HWE(+-F) can be met (no population structure)
- if enough samples to have a robust estimate of the allele frequencies

Practical: genotype calling

`https://github.com/mfumagalli/Copenhagen`

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from the data itself

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

How can we estimate allele frequencies from NGS data?

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

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

What is the simplest estimator of allele frequencies?

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

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

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

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

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{A,i}}{\sum_{i=1}^{N} (n_{A,i} + n_{G,i})}$$

Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

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

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{A,i}}{\sum_{i=1}^{N} (n_{A,i} + n_{G,i})}$$

$\hat{f} = 0.75$

What is wrong with this estimator? What improvements can we suggest?

Estimating allele frequencies

Maximum Likelihood estimator

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

with N samples.

What are $P(D|G = g)$ and $P(G = g|f)$?

Estimating allele frequencies

Maximum Likelihood estimator

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

$P(D|G=g)$ is the genotype likelihood and $P(G=g|f)$ is given by HWE (for instance).

In our previous example, $\hat{f} = 0.46$ which is much closer to the true value than previous estimators.

SNP calling (for low-coverage NGS data)

Challenges

SNP calling (for low-coverage NGS data)

Challenges

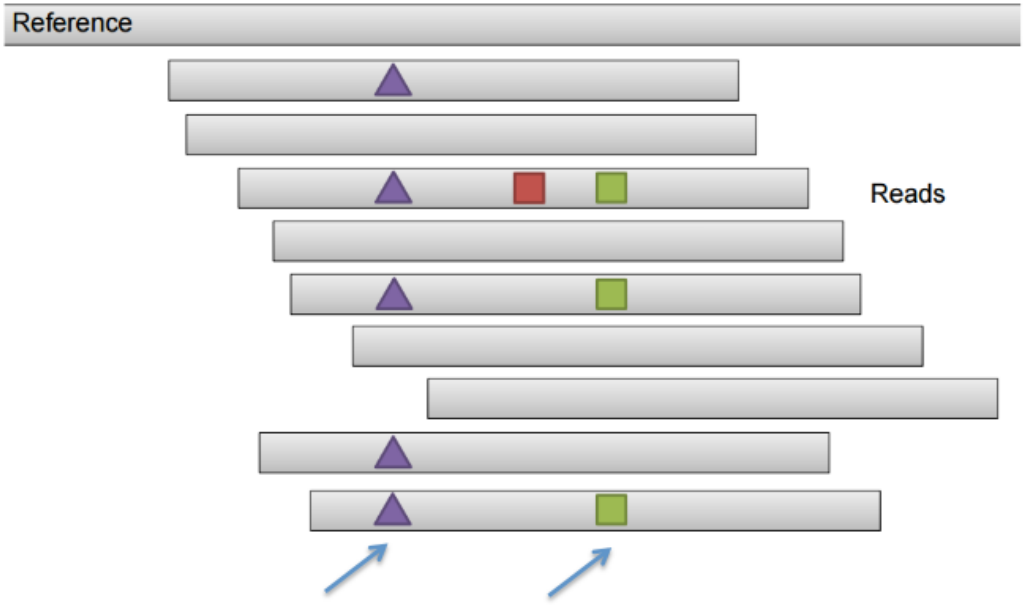
- If high levels of missing data, then genotypes can be lost.
- Rare variants are hard to detect.
- Trade off between false positive and false negative rates.

How to call SNPs (traditionally)?

- If at least one heterozygous genotype has been called.
- If the estimated allele frequency is above a certain threshold.

SNP calling procedures

- Alignment-based caller



We completely rely on how reads have been mapped

Figure from Erik Garrison

SNP calling procedures

- Assembly-based caller (as in GATK)

Local re-alignment around putative variants; better resolution for INDELs detection.

- Haplotype-based caller (as in freebayes)

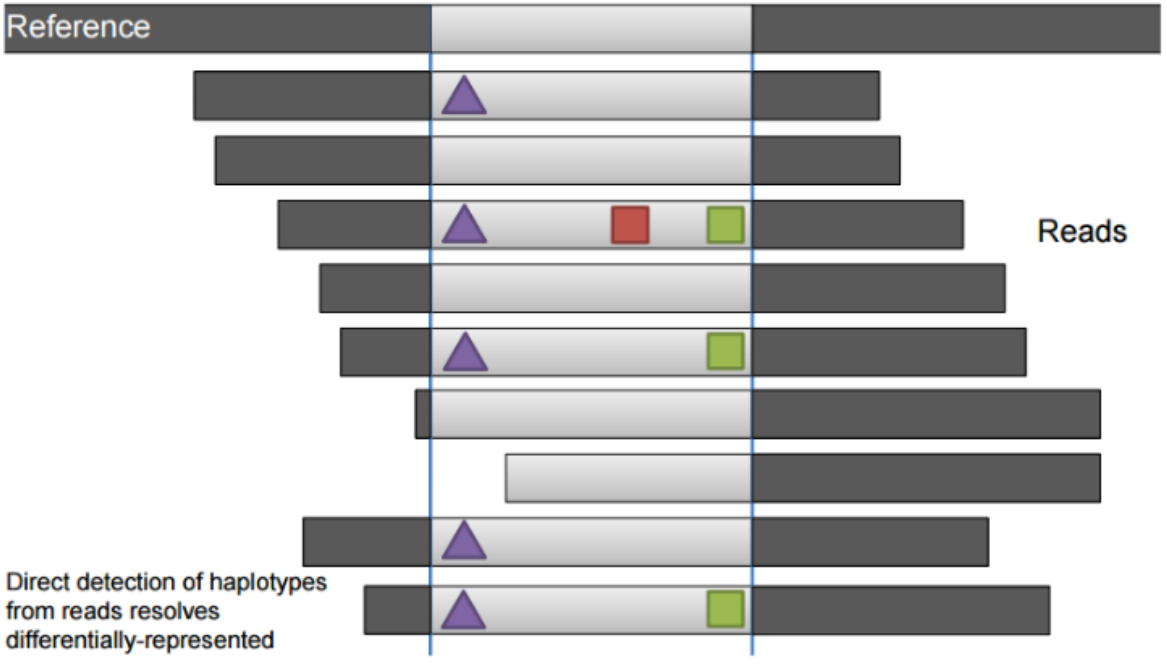


Figure from Erik Garrison

SNP calling

Call a SNP if

$$\hat{f} \geq t$$

where t can be the minimum sample allele frequency detectable (e.g. $t = 1/2N$ with N diploids).

Likelihood Ratio Test

A Likelihood Ratio Test (LRT) compares the goodness of fit between the null and the alternative model:

- Null model: $f = 0$
- Alternative model: $f \neq 0$

Likelihood Ratio Test

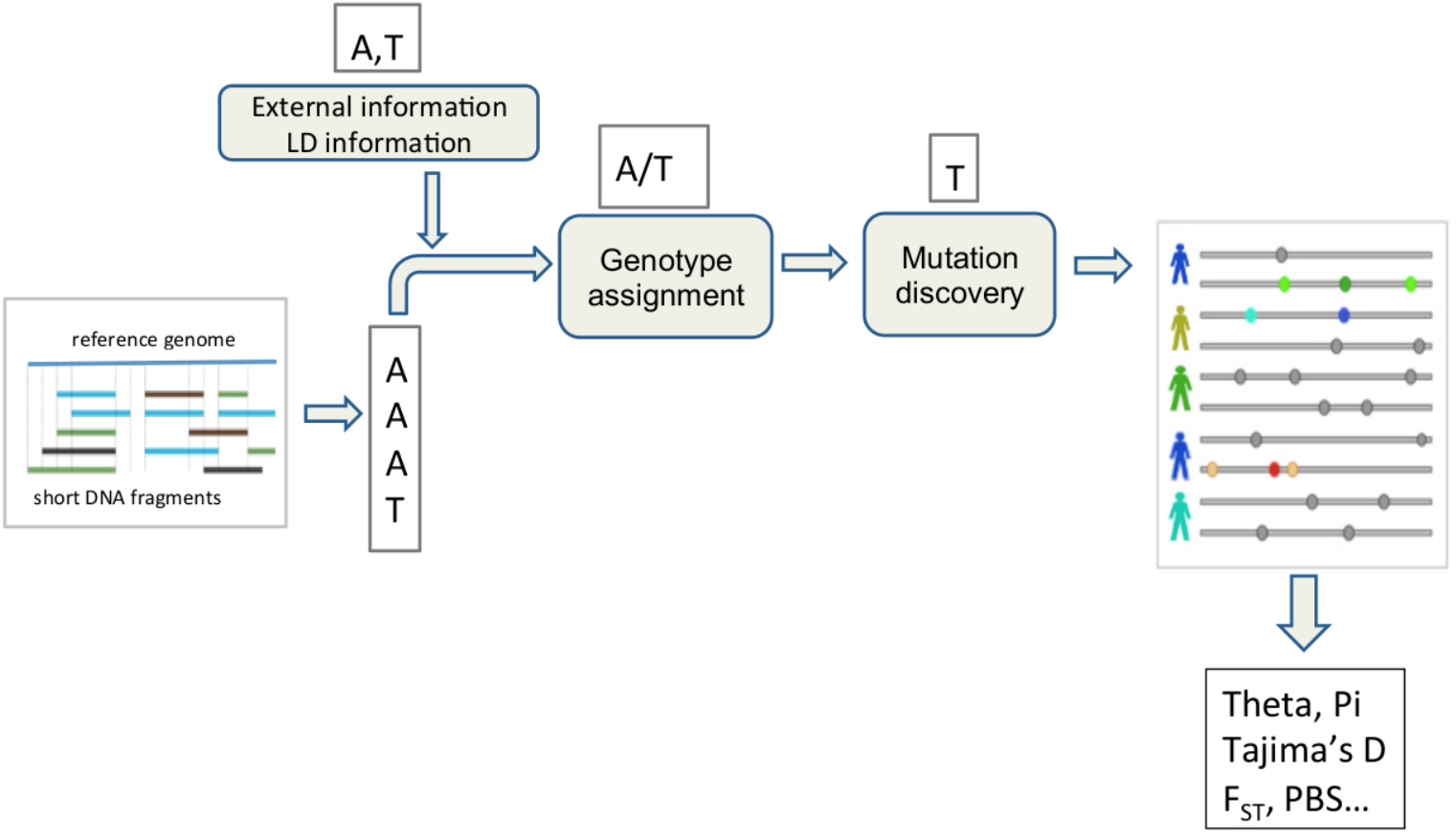
A Likelihood Ratio Test (LRT) compares the goodness of fit between the null and the alternative model:

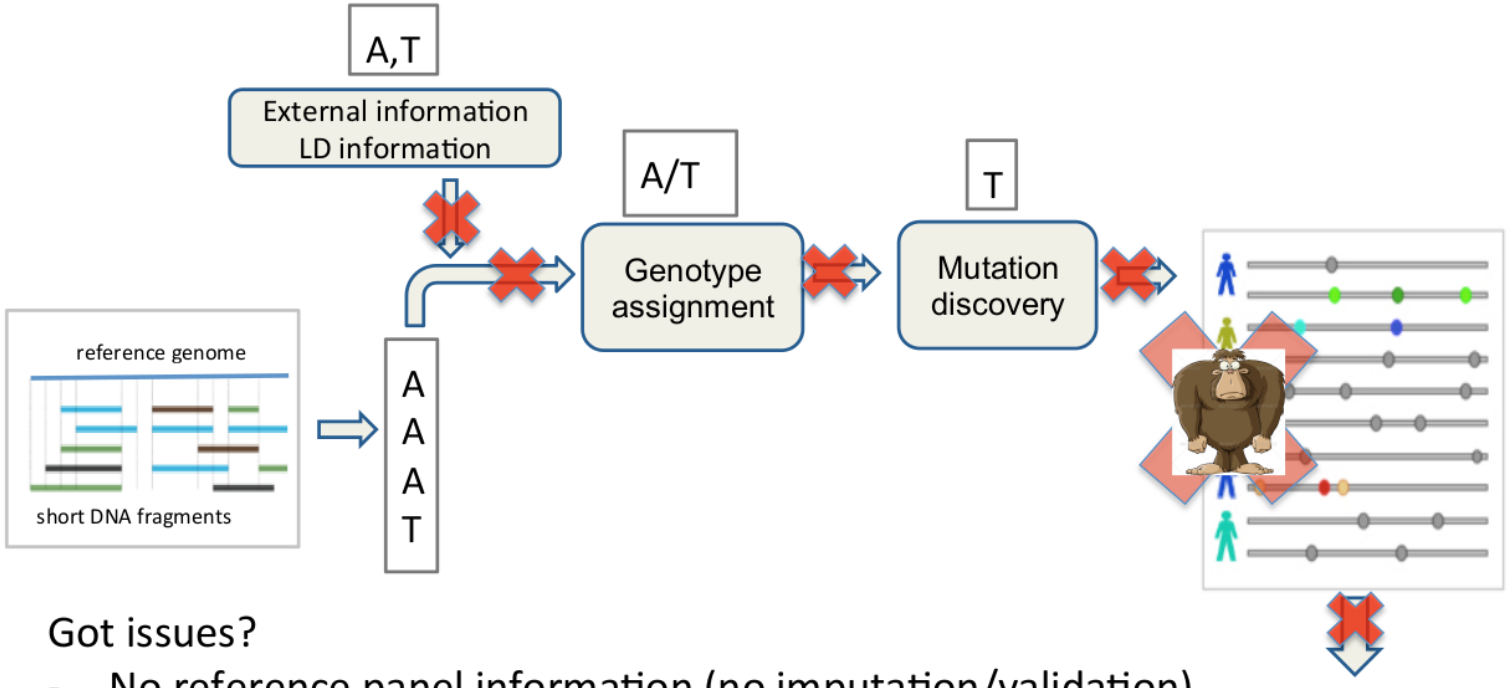
- Null model: $f = 0$
- Alternative model: $f \neq 0$

$$T = -2 \log \frac{L(f = 0)}{L(f = \hat{f}_{MLE})}$$

where T is χ^2 distributed with 1 degree of freedom.

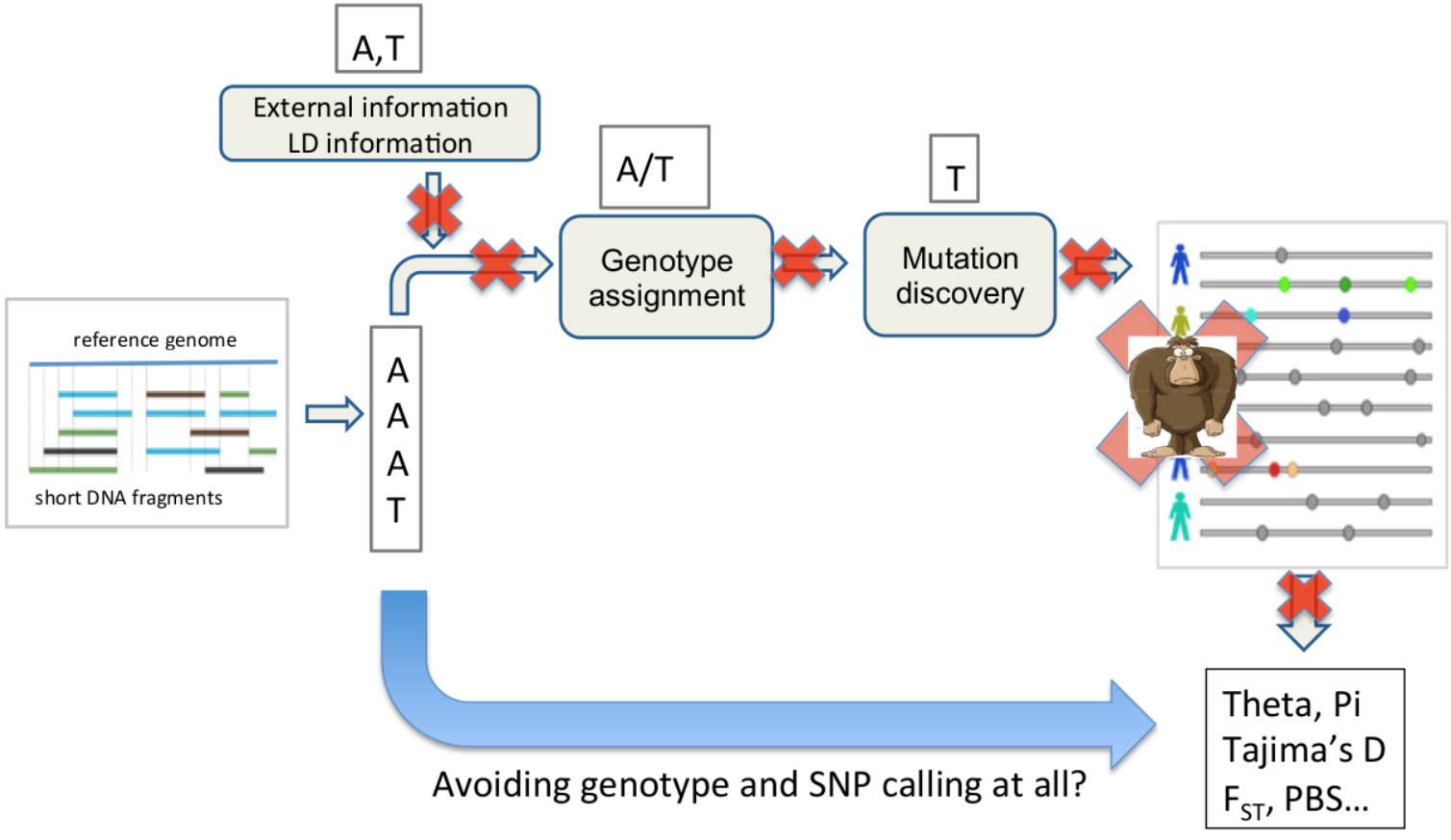
Practical: allele frequencies and SNP calling
`https://github.com/mfumagalli/Copenhagen`


A,T
External information
LD information
A/T
T
reference genome
short DNA fragments
A
A
A
T
Genotype assignment
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!**


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?

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}

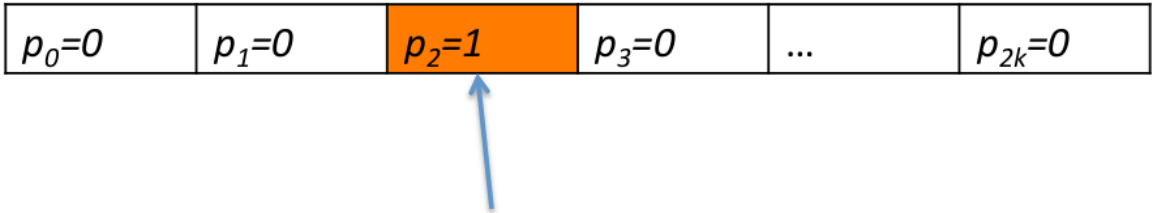
If folded, *k+1* entries

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

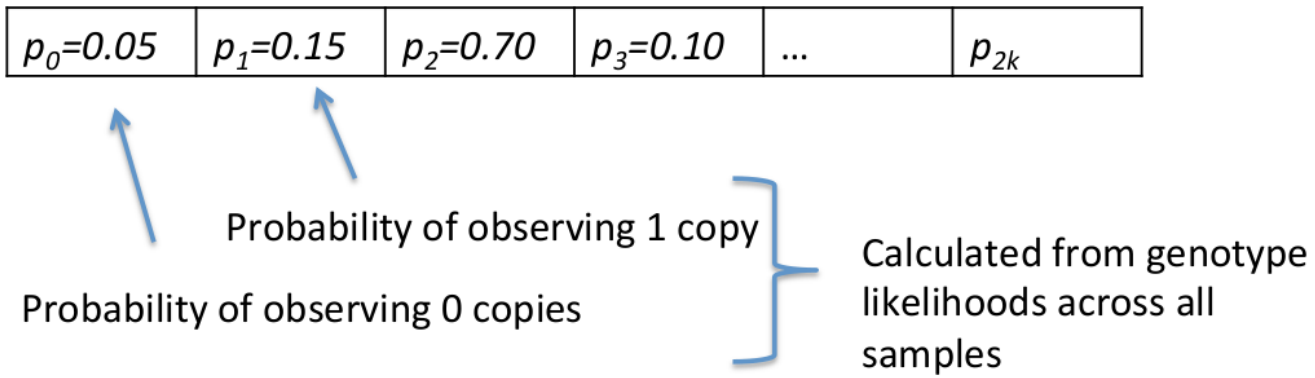


e.g. A is ancestral, G is derived (alternate)
AA AA AG AA AG AA AA AA AA

Sample allele frequency

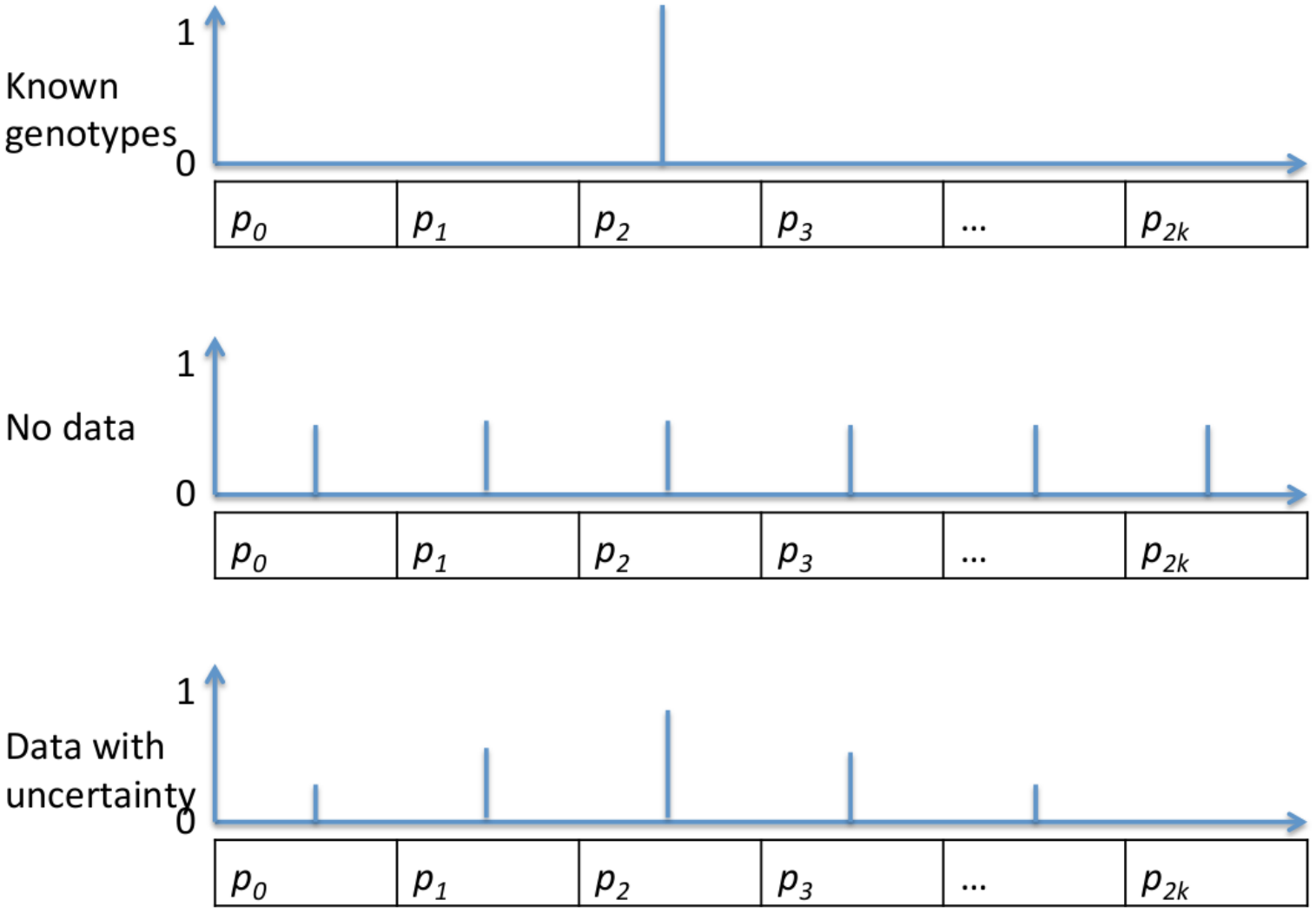
- *With k* diploid individuals, how many possible sample allele frequencies can I observe?

If unfolded, *2k+1* entries



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

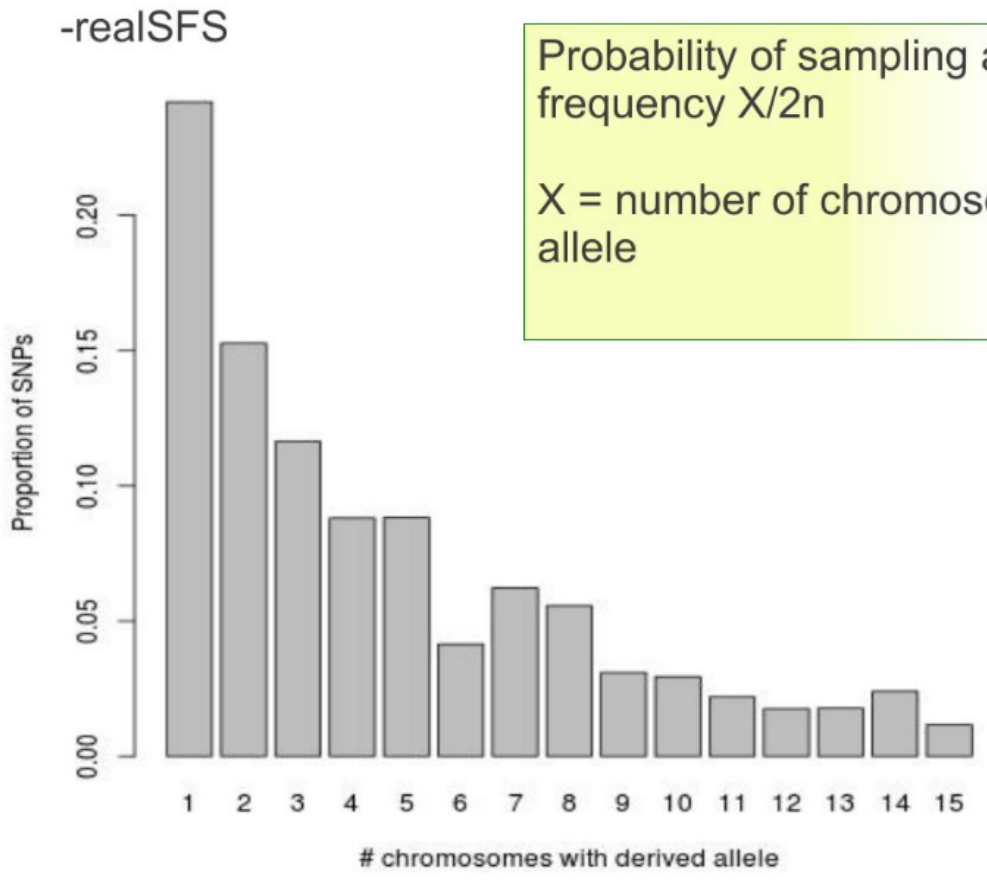
Sample allele frequency



Allele frequency likelihoods

	0	1	2	3	4	. . .	2n
Site1	0.00	-2.24	-4.53	-6.99	-9.63		-232.69
Site2	0.00	-2.24	-4.53	-6.99	-9.63		-232.69
Site3	-76.63	-37.87	-10.42	0.00	-9.59		-467.13
Site4	0.00	-2.24	-5.53	-6.99	-9.63		-237.55
.							
.							
.							
.							
.							
.							
Sitek	0.00	-8.62	-19.22	-30.67	-43.27		-626.78

Allele frequency likelihoods



Probability of sampling a site with allele frequency X/2n

X = number of chromosomes with alternative allele

Intended Learning Outcomes

At the end of this session you are now able to:

- appreciate the effect of sequencing depth and error to SNP/genotype calling
- calculate genotype and allele frequency likelihoods
- perform SNP/genotype calling from NGS data
- acknowledge the estimation of summary statistics from NGS data