

# Inference from NGS data

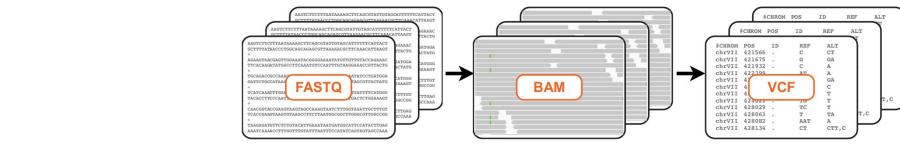
Anders Albrechtsen



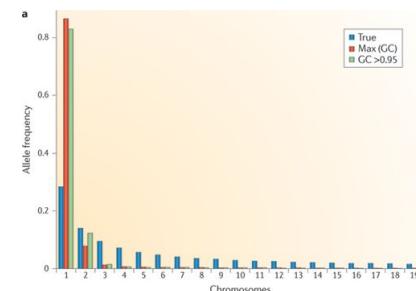
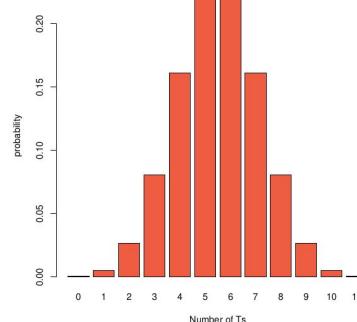
# This session

## This morning

- Discuss the issues of genotype calling from read data
- Calculate genotype likelihoods (GL)
- Use GLs as the basis for
  - Genotype calling
  - Variant calling
  - Allele frequency estimation
  - Site frequency estimation
- Exercises



$$P(X_{ij}|G) \propto \prod_{d=1}^D P(b_d|G_1, G_2) = \prod_{d=1}^D \left( \frac{1}{2} P(b_d|G_1) + \frac{1}{2} P(b_d|G_2) \right)$$



# Sequencing Depth

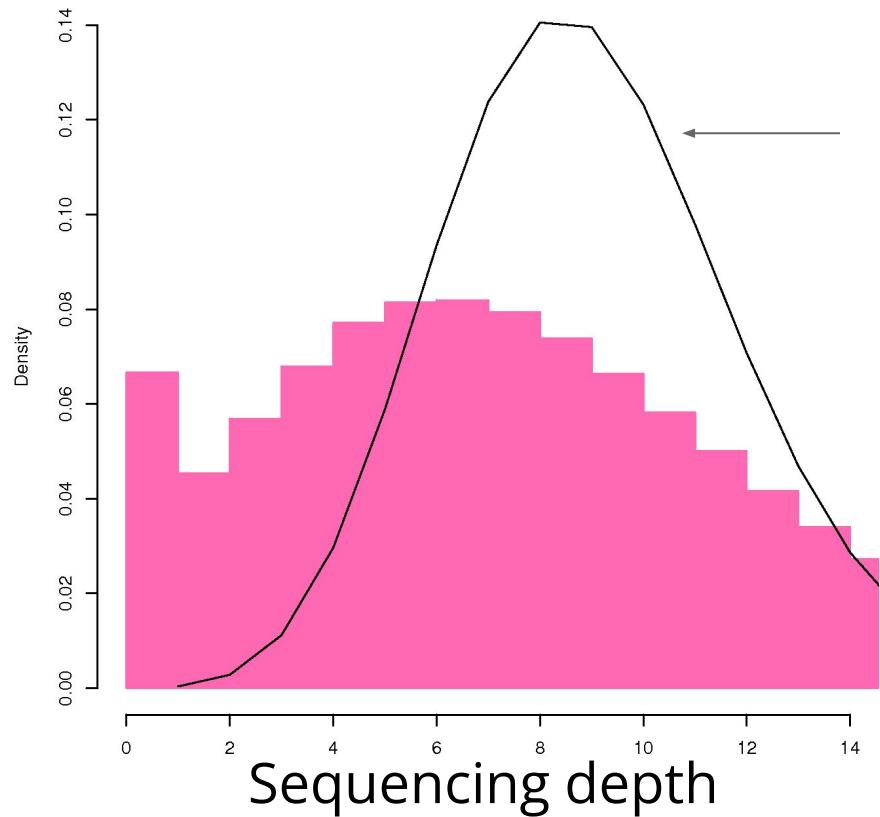
**Sequencing depth** is the number of reads covering a position

Average depth is often written as X e.g. **15X** sequencing

**Coverage** = depth or the fraction of genome with data

```
AGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAACACTG  
CAGCCACACCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
CAGCCACACCCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
TGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAC  
CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAA  
GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTAC  
TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT  
CATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT  
ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCAOGGTACCAGACAGA  
AGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC  
AGACCAGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

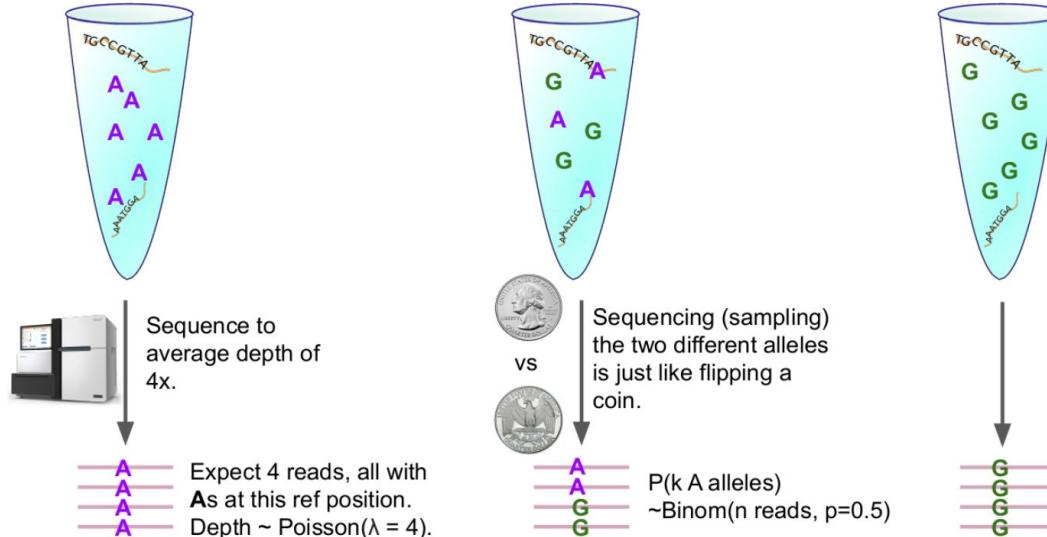
# Depth distribution



Theoretical distribution  
(Poisson) for 8X avg. depth if  
reads mapped perfectly and  
there was no bias

# Why don't we observe genotype

Each allele is sequenced separately and alleles are sampled with replacement



# Why don't we observe genotype

Question: Assuming an error rate of 1%  
Is the individual heterozygous C/T?

AGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAACACTG  
CAGCCACACCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
CAGCCACACCCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
TGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAC  
CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAA  
GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTAC  
TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT  
CATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT  
ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCAOGGTACCAGACAGA  
AGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC  
AGACCAGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA



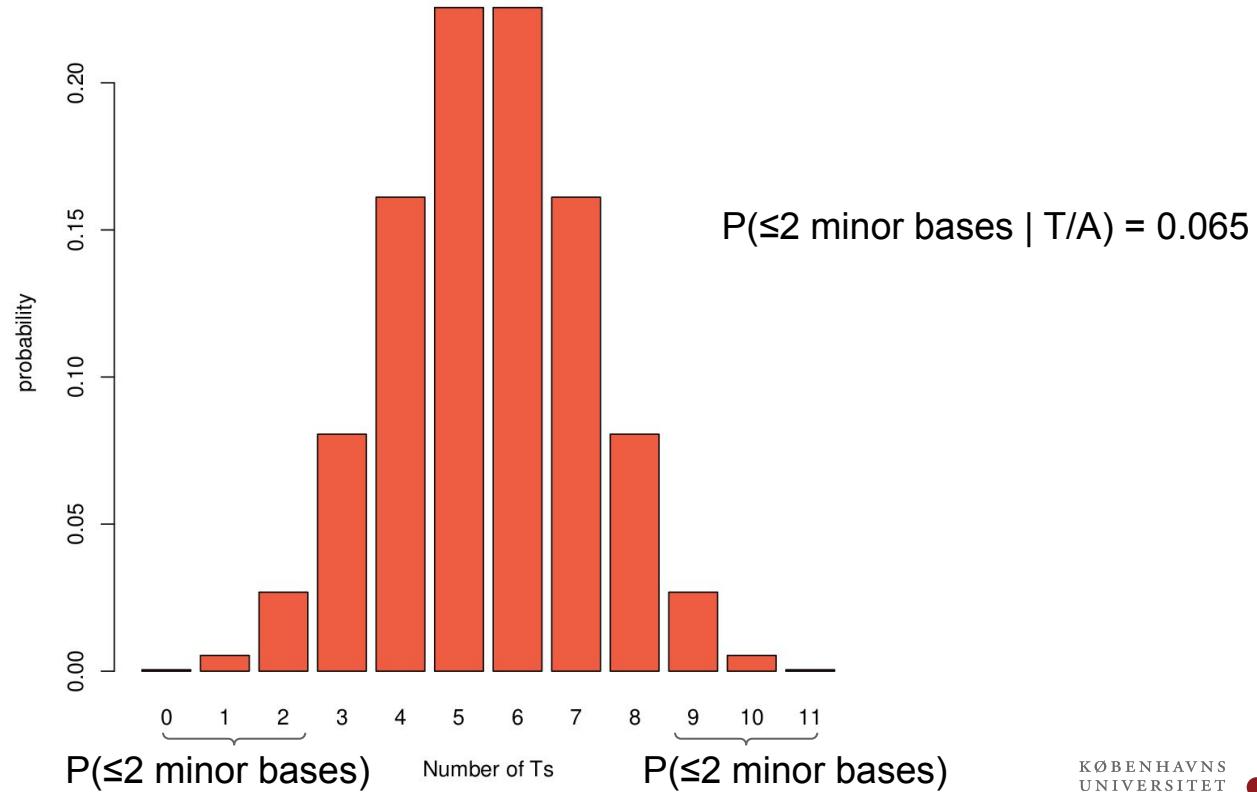
# Is the individual heterozygous or homozygous

- ⓘ Click **Present with Slido** or install our [Chrome extension](#) to activate this poll while presenting.



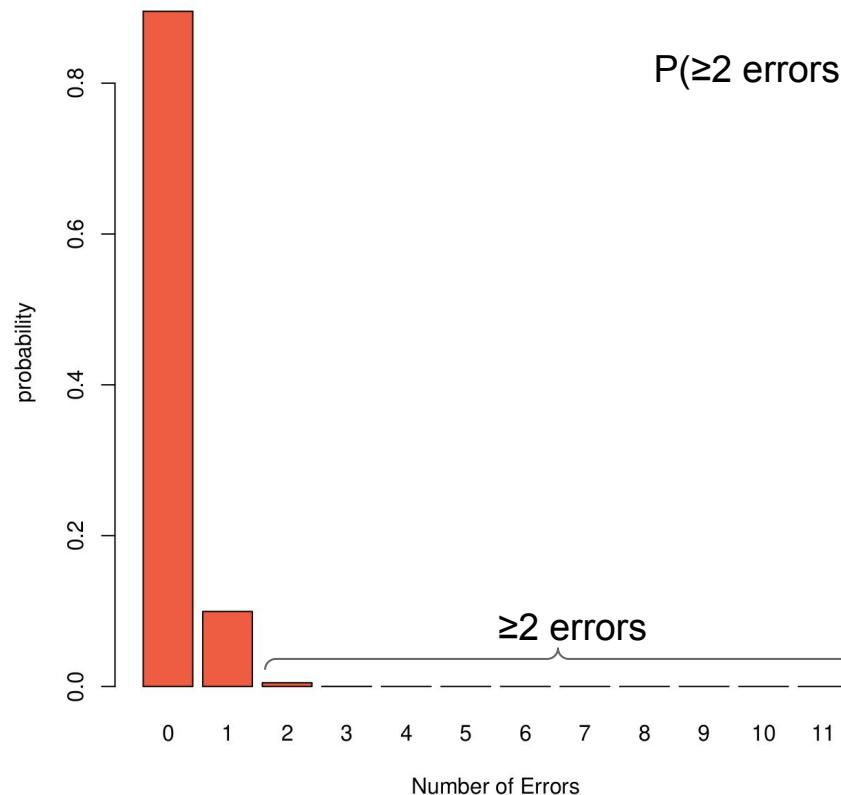
# Assuming heterozygous (C/T)

ACATTCAC  
ACACCCCG  
ACACCCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC



# Assuming homozygous (T/T)

ACATTCAC  
ACACCCCG  
ACACCCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC  
ACATTCAC



$$P(\geq 2 \text{ errors} \mid T/T) = 0.0052$$

# Why don't we observe genotype

$P(\geq 2 \text{ errors} | T/T) = 0.0052$

$P(\leq 2 \text{ minor bases} | T/C) = 0.065$

Question: Assuming an error rate of 1%  
Is the individual heterozygous C/T?

AGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAACACTG  
CAGCCACACCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
CAGCCACACCCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
TGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAC  
CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAA  
GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTAC  
TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT  
CATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT  
ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCAOGGTACCAGACAGA  
AGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGT  
AGACCAGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA

# Why don't we observe genotype

$P(\geq 2 \text{ errors} | T/T) = 0.0052$

$P(\leq 2 \text{ minor bases} | T/C) = 0.065$

Heterozygosity is 0.1%

Question: Assuming an error rate of 1%  
Is the individual heterozygous C/T?

AGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAACACTG  
CAGCCACACCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
CAGCCACACCCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT  
TGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAC  
CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAA  
GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTAC  
TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT  
CATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT  
ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA  
AGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC  
AGACCAGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA

# Why don't we observe genotype

$$P(\geq 2 \text{ errors} | T/T) = 0.0052$$

$$P(\leq 2 \text{ minor bases} | T/C) = 0.065$$

Heterozygosity is 0.1%

Question: Assuming an error rate of 1%  
Is the individual heterozygous C/T?

Approximate / rough calculations

$$\begin{aligned} P(T/T | \text{data}) &= p(\text{data}|T/T)*p(T/T)/p(\text{data}) \\ &\approx p(\text{data}|T/T)*p(T/T)/( p(\text{data}|T/T)*p(T/T) + p(\text{data}|T/C)*p(T/C)) \\ &\approx p(\text{error}\geq 2|T/T)p(\text{Hom})/( p(\geq 2\text{errors}|T/T)p(\text{Hom})+p(\text{error}\leq 2|T/C)p(\text{Het})) \\ &= 0.0052*0.999/(0.0052*0.999+0.065*0.001) \\ &= 0.987 = 98.7\% \end{aligned}$$



# Genotype likelihoods

how to quantify genotype uncertainty

# Likelihood of the data

Data ( $X_{ij}$ )		10 possible genotypes				
bases (b):		A	C	G	T	
TCCTTTTTTT	→	A	1	2	3	4
quality scores (Q):		C		5	6	7
+ ,77&&8888+		G			8	9
		T				10

The likelihood for the data for site j in indi

$$P(X_{ij}|G = \{G_1, G_2\}) = P(X_{ij}|G) \quad \text{where} \quad G \in \{A, C, G, T\}^2$$

# Simple genotype likelihoods

GATK: McKenna et al 2010

Simple calculation of genotype likelihoods :

$$P(X_{ij} | G)$$

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):

+77&&8888+

data  $X_{ij}$

*D* is the depth (number of reads at position j individual i)

*b* is the bases: e.g "TCCTTTTTTT" (i,j omitted for simplicity)



KØBENHAVNS  
UNIVERSITET

# Simple genotype likelihoods

GATK: McKenna et al 2010

Simple calculation of genotype likelihoods :

$$P(X_{ij}|G) \propto \prod_{d=1}^D P(b_d|G_1, G_2)$$

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):

+77&&8888+

data  $X_{ij}$

$D$  is the depth (number of reads at position j individual i)

$b$  is the bases: e.g "TCCTTTTTTT" (i,j omitted for simplicity)



KØBENHAVNS  
UNIVERSITET

# Simple genotype likelihoods

GATK: McKenna et al 2010

Simple calculation of genotype likelihoods :

$$P(X_{ij}|G) \propto \prod_{d=1}^D P(b_d|G_1, G_2) = \prod_{d=1}^D \left( \frac{1}{2} P(b_d|G_1) + \frac{1}{2} P(b_d|G_2) \right)$$

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):  
+,77&&8888+

data  $X_{ij}$

*D* is the depth (number of reads at position j individual i)

*b* is the bases: e.g "TCCTTTTTTT" (i,j omitted for simplicity)



KØBENHAVNS  
UNIVERSITET

# Simple genotype likelihoods

GATK: McKenna et al 2010

Simple calculation of genotype likelihoods :

$$P(X_{ij}|G) \propto \prod_{d=1}^D P(b_d|G_1, G_2) = \prod_{d=1}^D \left( \frac{1}{2} P(b_d|G_1) + \frac{1}{2} P(b_d|G_2) \right)$$

where  $P(b_d|G_l) = \begin{cases} \frac{\epsilon_d}{3} & b_d \neq G_l \\ 1 - \epsilon_d & b_d = G_l \end{cases}$ ,

where  $G = \{G_1, G_2\}$ ,  $b_d$  is the observed base and  $\epsilon_d$  is the probability of error from the quality score.

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):  
+,77&&8888+

data  $X_{ij}$

$D$  is the depth (number of reads at position j individual i)

$b$  is the bases: e.g "TCCTTTTTTT" (i,j omitted for simplicity)



KØBENHAVNS  
UNIVERSITET

# Simple genotype likelihoods

GATK: McKenna et al 2010

Simple calculation of genotype likelihoods :

$$P(X_{ij}|G) \propto \prod_{d=1}^D P(b_d|G_1, G_2) = \prod_{d=1}^D \left( \frac{1}{2} P(b_d|G_1) + \frac{1}{2} P(b_d|G_2) \right)$$

where  $P(b_d|G_l) = \begin{cases} \frac{\epsilon_d}{3} & b_d \neq G_l \\ 1 - \epsilon_d & b_d = G_l \end{cases}$ ,

where  $G = \{G_1, G_2\}$ ,  $b_d$  is the observed base and  $\epsilon_d$  is the probability of error from the quality score.

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):  
+,77&&8888+

Question: Assume you observe a read with base “T” and the ascii base quality score of “5”

- What is the  $P(b_d = "T" | G_1 = "C")$ ?

Ascii “5” is equal to a score of 20  
= 1% error rate



KØBENHAVNS  
UNIVERSITET

# Simple genotype likelihoods

GATK: McKenna et al 2010

Simple calculation of genotype likelihoods :

$$P(X_{ij}|G) \propto \prod_{d=1}^D P(b_d|G_1, G_2) = \prod_{d=1}^D \left( \frac{1}{2} P(b_d|G_1) + \frac{1}{2} P(b_d|G_2) \right)$$

where  $P(b_d|G_l) = \begin{cases} \frac{\epsilon_d}{3} & b_d \neq G_l \\ 1 - \epsilon_d & b_d = G_l \end{cases}$ ,

where  $G = \{G_1, G_2\}$ ,  $b_d$  is the observed base and  $\epsilon_d$  is the probability of error from the quality score.

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):  
+,77&&8888+

Question: Assume you observe a read with base “C” and the ascii base quality score of “5”

- What is the  $P(b_d="C"|G_1="C")?$

Ascii “5” is equal to a score of 20  
= 1% error rate



KØBENHAVNS  
UNIVERSITET

b	$Q_{ascii}$	$Q_{score}$	$\epsilon_d$	$p(b_d T)$	$p(b_d C)$	$p(b_d G/A)$
T	+	10	0.1	0.9	0.033	0.033
C	,	11	0.079	0.026	0.92	0.026
C	7	22	0.0063	0.0021	0.99	0.0021
T	7	22	0.0063	0.99	0.0021	0.0021
T	&	5	0.32	0.68	0.11	0.11
T	&	5	0.32	0.68	0.11	0.11
T	8	23	0.005	0.99	0.0017	0.0017
T	8	23	0.005	0.99	0.0017	0.0017
T	8	23	0.005	0.99	0.0017	0.0017
T	8	23	0.005	0.99	0.0017	0.0017
T	+	10	0.1	0.9	0.033	0.033

**Data ( $X_{ij}$ )**

bases (b):

TCCTTTTTTT

quality scores (Q):  
+,77&&8888+

$$P(X|G = TC) \propto \prod_{d=1}^D P(b_d|T, C) = \prod_{d=1}^D \left( \frac{1}{2} P(b_d|T) + \frac{1}{2} P(b_d|C) \right)$$



# All 10 possible genotype likelihoods

log Likelihood  $P(D|G)$

	A	C	G	T
A	-52.84	-44.49	-52.84	-16.66
C		-43.13	-44.49	-8.31
G			-52.84	-16.66
T				-10.79

likelihood ( normal scale)

	A	C	G	T
A	0	0	0	0
C		0	0	0.00025
G			0	0
T				0.00002

Data ( $X_{ij}$ )

bases (b):

TCCTTTTTTT

quality scores (Q):  
+77&8888+



# Many genotype likelihood models

GATK haplotype caller

GATK unified genotyper caller

Samtools/bcftools

freeBayes

ATLAS ( for ancient DNA)



# Genotype calling

log Likelihood  $P(D|G)$

	A	C	G	T
A	-52.84	-44.49	-52.84	-16.66
C		-43.13	-44.49	-8.31
G			-52.84	-16.66
T				-10.79

likelihood ( normal scale)

	A	C	G	T
A	0	0	0	0
C		0	0	0.00025
G			0	0
T				0.00002

Data ( $X_{ij}$ )

bases (b):

TCCTTTTTTT

quality scores (Q):  
+77&8888+

simple genotype callers - Maximum likelihood

ML I Choose the genotype with the largest likelihood

$$\arg \max_G P(X|G)$$

ML II only call a genotype if the likelihood with much better than the second best e.g. a likelihood ratio > 2

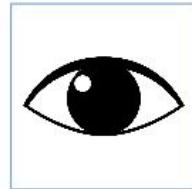


KØBENHAVNS  
UNIVERSITET

# Genotype calling: bayesian

Data  $X_{ij}$

bases (b):  
TCCTTTTTTT  
quality scores (Q):  
+,77&&8888+



Likelihood

$p(X_{ij} | G)$

	A	C	G	T
A	0	0	0	0
C	0	0	0.00025	
G	0		0	
T			0.00002	

Prior

uniform:  
 $p(A/A)=1/10$   
 $p(A/C)=1/10$   
....  
 $p(T/T)=1/10$

Bayes formula

$$p(G|X_{ij}) = \frac{p(X_{ij}|G)p(G)}{p(X_{ij})}$$

Posterior probability

	A	C	G	T
A	0	0	0	0
C	0	0	0.922	
G	0		0	
T			0.077	

Modified slide from  
matteo Fumagalli

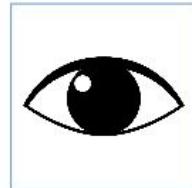


KØBENHAVNS  
UNIVERSITET

# Genotype calling: bayesian

Data  $X_{ij}$

bases (b):  
TCCTTTTTTT  
quality scores (Q):  
+,77&&8888+



Likelihood

$p(X_{ij} | G)$

	A	C	G	T
A	0	0	0	0
C	0	0	0.00025	
G	0		0	
T			0.00002	

Prior

$$\begin{aligned}f_A &= 0 \\ f_C &= 0.01 \\ f_G &= 0 \\ f_T &= 0.99\end{aligned}$$

Assume HWE

Bayes formula

$$p(G|X_{ij}, f) = \frac{p(X_{ij}|G)p(G|f)}{p(X_{ij}|f)}$$

Posterior probability

	A	C	G	T
A	0.0	0	0.0	0.0
C	0	0	0.20	
G	0		0	
T			0.80	

HWE assumption:  $p(C/C) = f_C^2, p(C/T) = 2f_C f_T, p(T/T) = f_T^2$

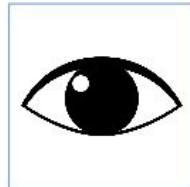


KØBENHAVNS  
UNIVERSITET

# Genotype calling: empirical bayes

Data  $X_{ij}$

bases (b):  
TCCTTTTTTT  
quality scores (Q):  
+,77&&8888+

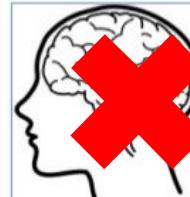


Likelihood

$p(X_{ij} | G)$

		A	C	G	T
A	0	0	0	0	
	0.00025				
C	0	0	0.00025		
			0		
G	0	0	0	0	
				0.00002	
T	0.00002				

Use all of your samples to estimate allele frequencies



Prior

$$\begin{aligned}f_A &= 0 \\f_C &= 0.05 \\f_G &= 0 \\f_T &= 0.95\end{aligned}$$

Assume HWE

Bayes formula

$$p(G|X_{ij}, f) = \frac{p(X_{ij}|G)p(G|f)}{p(X_{ij}|f)}$$

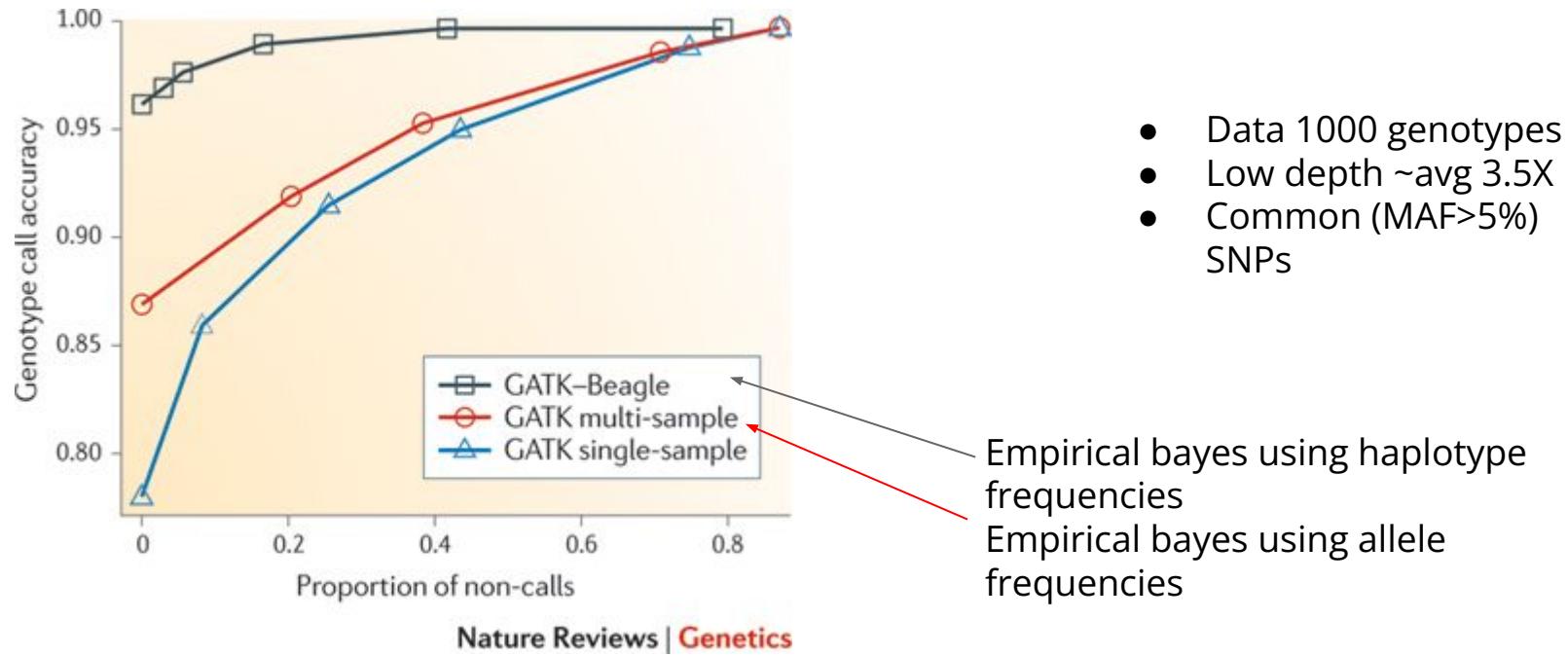
Posterior probability

		A	C	G	T
A	0.0	0.0	0.0	0.0	
	0.0	0.0	0.0	0.56	
C	0.0	0.0	0.0	0	
				0.44	
G	0.0	0.0	0.0	0	
				0.44	
T	0.0	0.0	0.0	0.44	

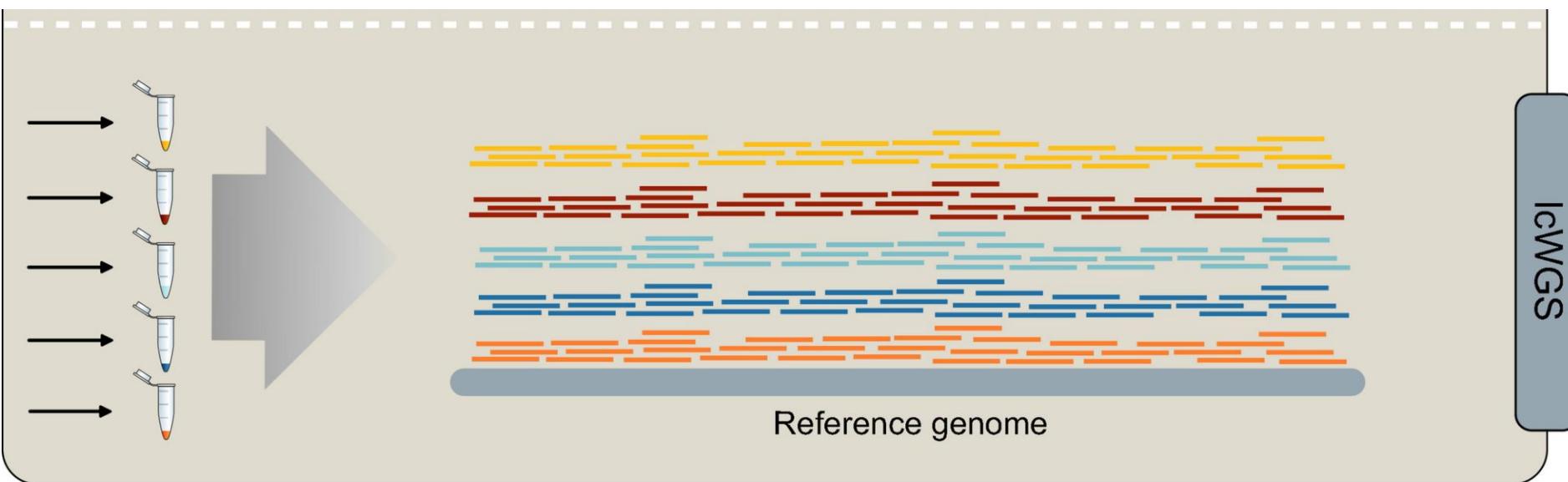
HWE assumption:  $p(C/C) = f_C^2, p(C/T) = 2f_C f_T, p(T/T) = f_T^2$



# Which genotype caller is best



# How to do inference from low depth



IcWGS



# Simple allele frequency estimator

## Simple frequency estimator

- assume only two allele types exists
- let  $n_1^i$  and  $n_2^i$  be the counts of observed alleles in individual  $i$ .
- $f = \frac{\sum n_1^i}{\sum(n_1^i + n_2^i)}$

Heterozygous with one m allele and one M allele

## Example

i	1	2	3	4	5	6	Total
True Geno	MM	MM	Mm	Mm	mm	mm	
#M Reads	7	25	5	4	0	0	41
#m reads	0	1	3	4	2	4	14

$$f = \frac{41}{41+14} = 0.75$$



# Maximum likelihood estimator

## ML frequency estimator

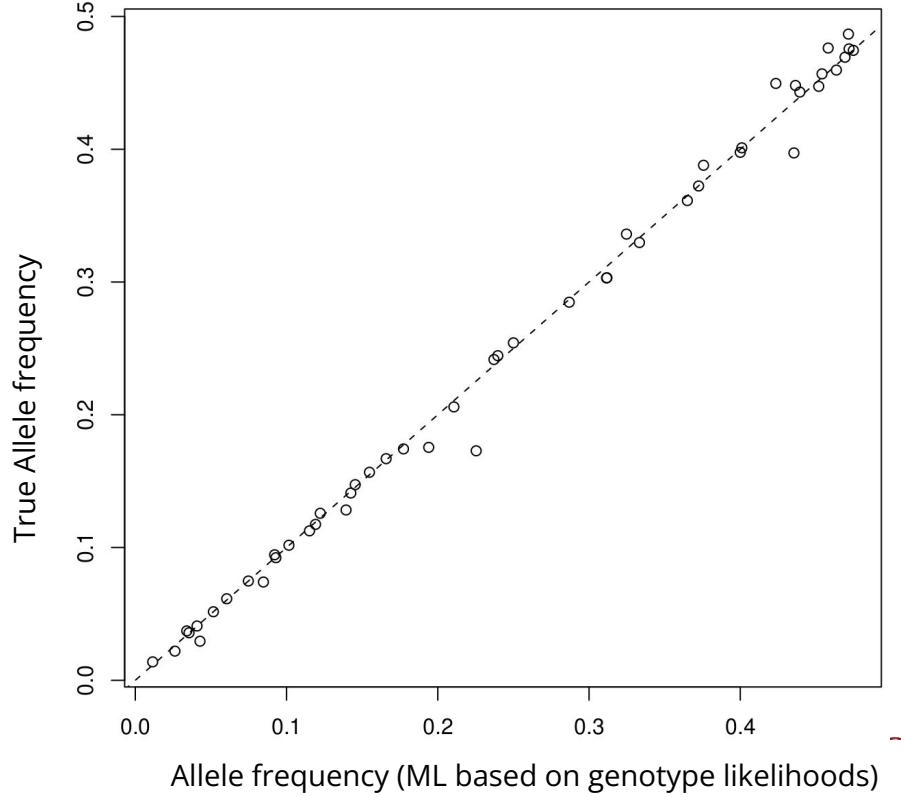
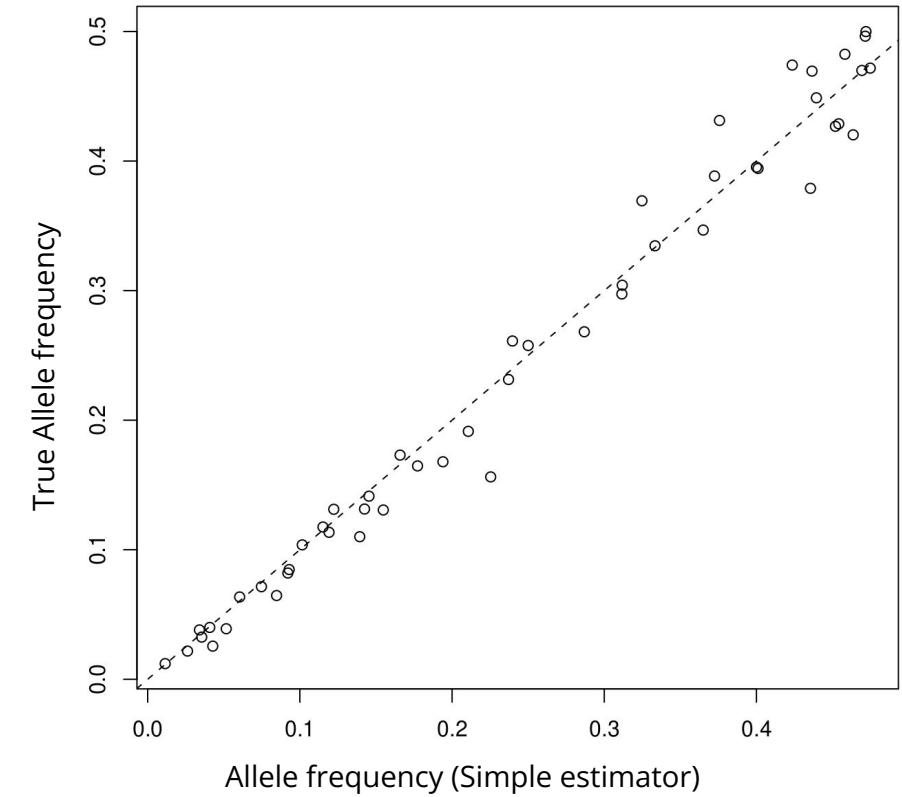
- assume only two alleles type exists
- $P(X^i|f) = \sum_{g \in \{0,1,2\}} P(X|G=g)P(G=g|f)$
- assume HWE e.g.  $P(G=A_1A_1|f) = f^2$
- $\hat{f}_{ML} = \operatorname{argmax}_p \prod_i P(X^i|f)$

### Example

i	1	2	3	4	5	6	Total
True Geno	MM	MM	Mm	Mm	mm	mm	
#M Reads	7	25	5	4	0	0	41
#m reads	0	1	3	4	2	4	14

$$\hat{f}_{ML} = 0.46$$



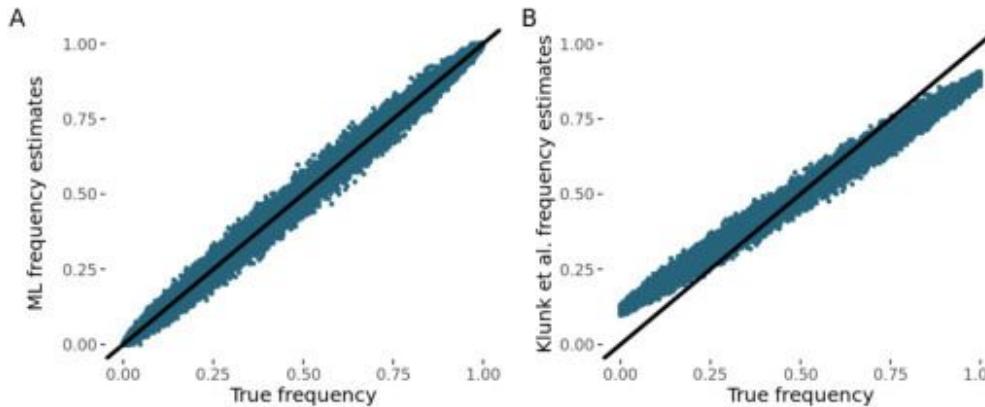


# Important to use GL correctly

[nature](#) > [articles](#) > [article](#)

Article | Published: 19 October 2022

## Evolution of immune genes is associated with the Black Death



**Insufficient evidence for natural selection associated with the Black Death**

Alison R. Barton, Cindy G. Santander, Pontus Skoglund, Ida Moltke, David Reich,  
 Iain Mathieson

doi: <https://doi.org/10.1101/2023.03.14.532615>



# Other use of frequencies from GL

## SNP calling

Null model:  $f_A=0, f_C=0, f_G=0, f_T=1$

$$L_{\text{null}} \propto p(X_j | f_A=0, f_C=0, f_G=0, f_T=1)$$

Alt model:  $f_A=0, f_C=0.05, f_G=0, f_T=0.95$

$$L_{\text{alt}} \propto p(X_j | f_A=0, f_C=0.05, f_G=0, f_T=0.95)$$

## Likelihood ratio test

$$2\log(L_{\text{alt}}/L_{\text{null}}) \sim \chi^2_1$$

## Test difference in frequency

Null model:  $f_A=f_G=0, f_C=0.05, f_T=0.95$

$$L_{\text{null}} \propto p(X_j | f_A=f_G=0, f_C=0.05, f_T=0.95)$$

Alt model:

group1:  $f_A=f_G=0, f_C=0.02, f_T=0.98$ )

group2:  $f_A=f_G=0, f_C=0.07, f_T=0.93$ )

$$L_{\text{alt}} \propto p(X_{j1} | f_A=f_G=0, f_C=0.02, f_T=0.98) * p(X_{j2} | f_A=f_G=0, f_C=0.07, f_T=0.93)$$

## Likelihood ratio test

$$2\log(L_{\text{alt}}/L_{\text{null}}) \sim \chi^2_1$$



# Software for variant & genotype calling

GATK  
samtools/bcftools  
freeBayes



- Assembly-based caller (as in GATK)  
Local re-alignment around putative variants; better resolution for INDELs detection.
- Haplotype-based caller (as in freebayes)



# VCF files

Metadata

```
##fileformat=VCFv4.1
##fileDate=20120630
##source=freeBayes version 0.9.6
##reference=W303.fasta
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observations">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT BYB1_G07-0
chrVIII 255912 . GA G 1408.89 . AO=18;... GT:GQ:DP:... 0/1:50000:...
chrVIII 263821 . G A 6257.23 . AO=201;... GT:GQ:DP:... 1/1:50000:...
chrVIII 275478 . TGCCAG TGCCAA 5885.31 . AO=185;... GT:GQ:DP:... 1/1:50000:...
chrVIII 276438 . CA C 63.5434 . AO=3;... GT:GQ:DP:... 0/1:63.5064:...
chrVIII 290238 . TA T 12.4555 . AO=5;... GT:GQ:DP:... 0/1:12.4555:...
chrVIII 298817 . CT C 482.635 . AO=13;... GT:GQ:DP:... 0/1:50000:...
chrVIII 314728 . CAT C 101.007 . AO=8;... GT:GQ:DP:... 0/1:101.007:...
chrVIII 317567 . T G,A 160.186 . AO=37;... GT:GQ:DP:... 0/1:160.126:...
chrVIII 323237 . G GA 99.7114 . AO=9;... GT:GQ:DP:... 0/1:95.3368:...
chrVIII 340061 . TTA T 360.562 . AO=5;... GT:GQ:DP:... 0/1:50000:...
chrVIII 361913 . AT A 1237.88 . AO=17;... GT:GQ:DP:... 0/1:50000:...
chrVIII 368029 . T A,G 1630.61 . AO=35,31;... GT:GQ:DP:... 1/2:50000:...
```

Body

G	A	SNP
G	GA	insertion
CT	C	deletion
TTA	T	2 bp deletion
TGCCAG	TGCCAA	complex mutation
T	A, G	multiple alternate alleles



# VCF files

## Metadata

```
##fileformat=VCFv4.1
##fileDate=20120630
##source=freeBayes version 0.9.6
##reference=W303.fasta
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observations">
##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT BYB1_G07-0
chrVIII 255912 . GA G 1408.89 .
AO=18;...
GT:GQ:DP:... 0/1:50000:...
chrVIII 263821 . G A 6257.23 .
AO=201;...
GT:GQ:DP:... 1/1:50000:...
chrVIII 275478 . TGGCCAG TGCCAA 5885.31 .
AO=185;...
GT:GQ:DP:... 1/1:50000:...
chrVIII 276438 . CA C 63.5434 .
AO=3;...
GT:GQ:DP:... 0/1:63.5064:...
chrVIII 290238 . TA T 12.4555 .
AO=5;...
GT:GQ:DP:... 0/1:12.4555:...
chrVIII 298817 . CT C 482.635 .
AO=13;...
GT:GQ:DP:... 0/1:50000:...
chrVIII 314728 . CAT C 101.007 .
AO=8;...
GT:GQ:DP:... 0/1:101.007:...
chrVIII 317567 . T G,A 160.186 .
AO=37;...
GT:GQ:DP:... 0/1:160.126:...
chrVIII 323237 . G GA 99.7114 .
AO=9;...
GT:GQ:DP:... 0/1:95.3368:...
chrVIII 340061 . TTA T 360.562 .
AO=5;...
GT:GQ:DP:... 0/1:50000:...
chrVIII 361913 . AT A 1237.88 .
AO=17;...
GT:GQ:DP:... 0/1:50000:...
chrVIII 368029 . T A,G 1630.61 .
AO=35,31;...
GT:GQ:DP:... 1/2:50000:...
```

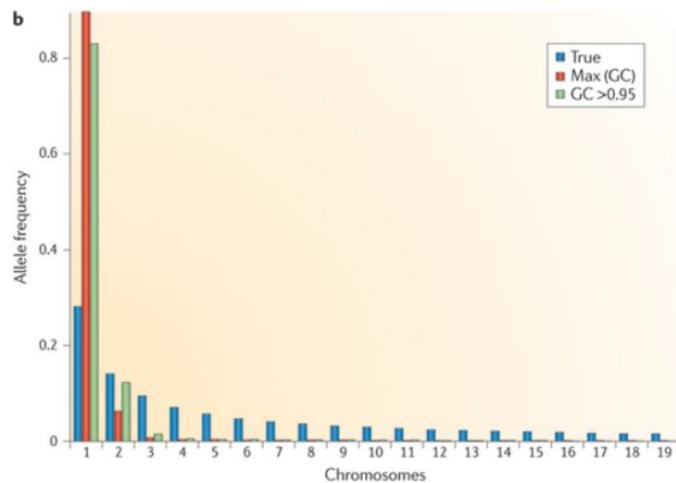
## Body

AB=0; ABP=0; AC=2; AF=1; AN=2; AO=185; CIGAR=1M1D4M1X; DP=192; DPRA=0; EPP=4.99397; EPPR=9.52472; HWE=-0; LEN=6; MEANALT=5; MQM=57.1243; MQMR=43; NS=1; NUMALT=1; ODDS=94.868; PAIRED=0.972973; PAIREDR=1; RO=3; RPP=10.3464; RPPR=9.52472; RUN=1; SAP=8.18662; SRP=9.52472; TYPE=complex; XAI=0.0102812; XAM=0.0122725; XAS=0.00199131; XRI=0; XRM=0; XRS=0; BVAR



# Should we always call genotypes

Site frequency spectrum

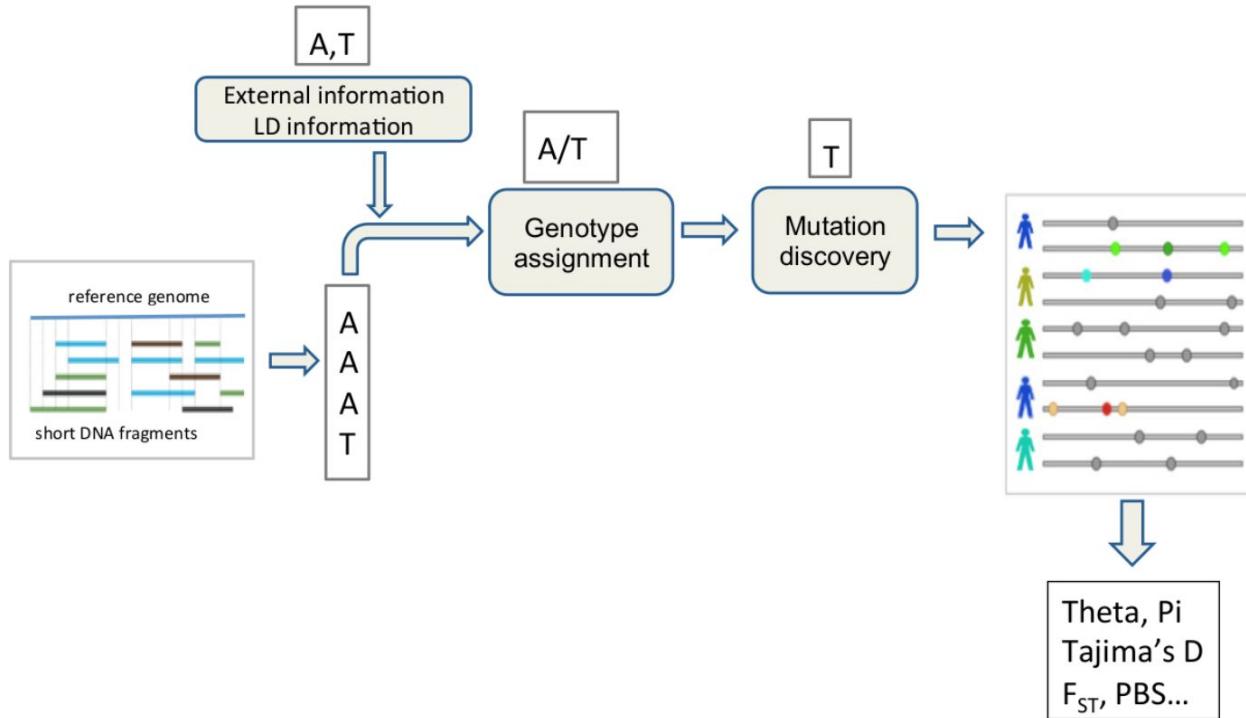


Nature Reviews | Genetics

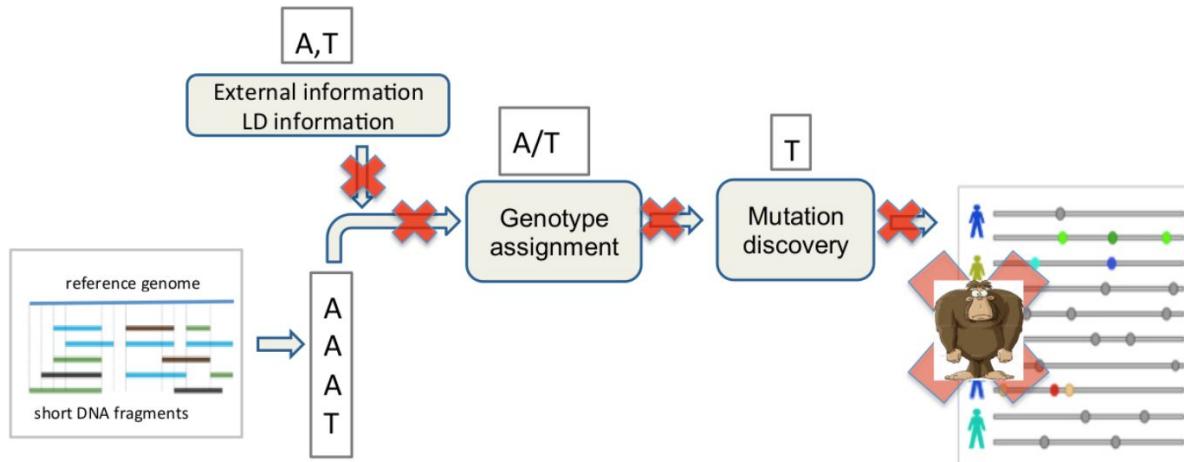
For low depth there are no filter or haplotype imputation approach that will work



# High quality data



# If you don't have high quality data

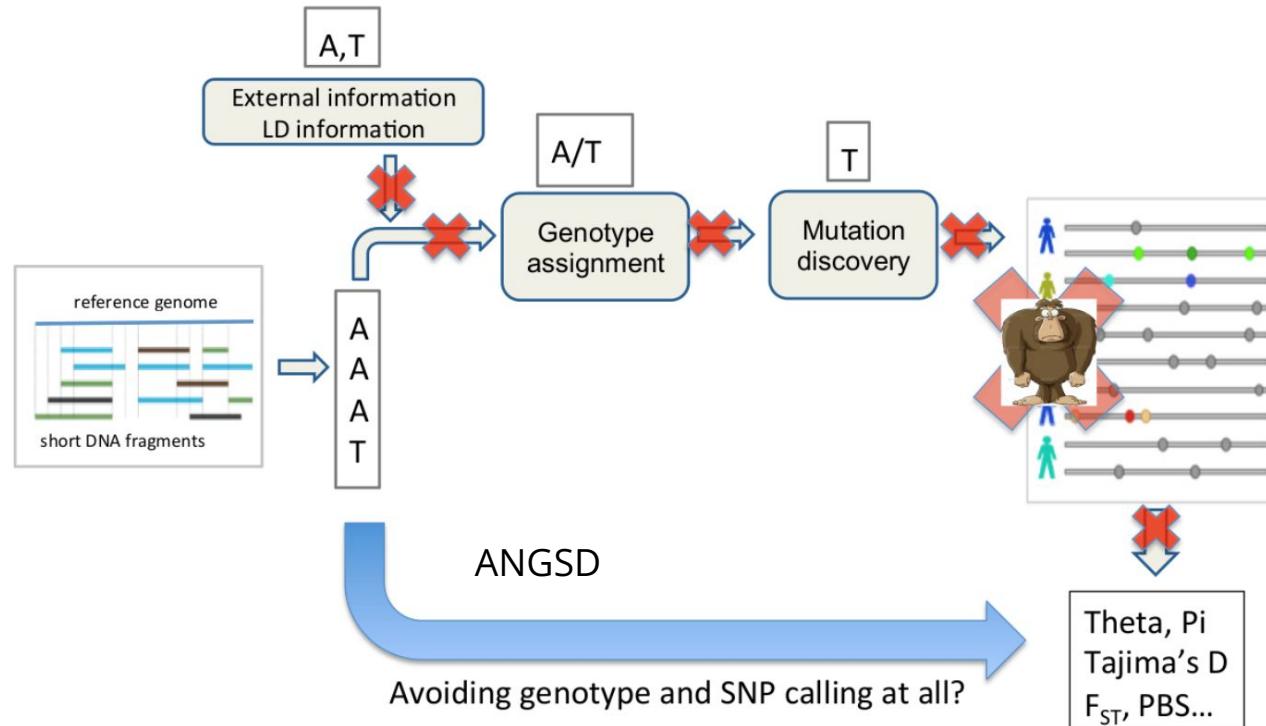


Got issues?

- 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!**



# Alternative is to not call genotypes



# Low depth vs high depth

## Genotype call based

- plink -- freq
- plink --hardy
- plink --genome
- plink --assoc
- plink --pca / smartPCA
- plink --ibs
- plink --fst
- ADMIXTURE
- ADMIXTURE project
- PCrelate
- admixtools

## GL based

- angsd -doMaf
- angsd -dohet (PCangsd -inbreeding)
- angsd -GL > ngsRelate (NGSremix)
- angsd -doAssoc
- angsd -GL > PCAngsd
- angsd -doIBS
- angsd -doSAF -> winSFS
- NGSadmix
- fastNGSadmix
- PCAngsd -kinship
- angsd -doAbbababa



Received: 1 December 2020

Revised: 30 June 2021

Accepted: 1 July 2021

DOI: 10.1111/mec.16077

SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

# A beginner's guide to low-coverage whole genome sequencing for population genomics

Runyang Nicolas Lou<sup>1</sup>  | Arne Jacobs<sup>1</sup>  | Aryn P. Wilder<sup>2</sup>  |

Nina Overgaard Therkildsen<sup>1</sup> 

<https://onlinelibrary.wiley.com/doi/10.1111/mec.16077#>

KØBENHAVNS  
UNIVERSITET



# Time for exercises

Go to

[popgen.dk/popgen24github](https://popgen.dk/popgen24github)

Download NGS intro ipynb

Go to emily server

Upload ipynb (press upload again)



# The likelihood

$$p(X | \theta) = \underbrace{\prod_{i=1}^N p(X_i | \theta)}_1 = \underbrace{\prod_{i=1}^N \sum_z p(X_i | Z_i = z)p(Z_i = z | \theta)}_{2 \quad 3}$$

- ➊ Probability of data given parameters
- ➋ Assumes individuals are independent
- ➌ Introduce latent variable using law of total probability

## Input Genotype likelihoods

i	$P(X Z=MM)$	$P(X Z=mM)$	$P(X Z=mm)$
1	0.93	0.0078	$1.0 \times 10^{-14}$
2	0.0078	$1.5 \times 10^{-8}$	$9.9 \times 10^{-51}$
3	$9.5 \times 10^{-7}$	$3.9 \times 10^{-3}$	$9.7 \times 10^{-11}$
4	$9.6 \times 10^{-9}$	$3.9 \times 10^{-3}$	$9.6 \times 10^{-9}$
5	$1.0 \times 10^{-8}$	0.25	0.98
6	$1.0 \times 10^{-8}$	0.062	0.96

## Notation

$X$  is all of the data for a single site  
 $X_i$  is the data for individual i  
 $\theta$  is the frequency of the two alleles  
 $\theta = (\theta_M, \theta_m)$   
 $N$  is the number of individuals

## Latent variable

$z$  is latent state of the genotype  
 $Z_i$  is the genotype for individual i  
 $Z_i \in \{MM, Mm, mm\}$

## Binomial ( HWE)

$$p(Z_i = MM | \theta) = \theta_M \theta_M$$

$$p(Z_i = Mm | \theta) = 2\theta_M \theta_m$$

$$p(Z_i = mm | \theta) = \theta_m \theta_m$$



# First iteration in EM algorithm

$$p(Z_i|X_i, \theta^{(0)}) = \frac{p(X_i|Z_i)p(Z_i|\theta^{(0)})}{p(X_i|\theta^{(0)})}$$

Unobserved  
Genotypes

G
MM
MM
mM
mM
m
m

Input  
Genotype likelihoods

i	$P(X Z=MM)$	$P(X Z=mM)$	$P(X Z=mm)$
1	0.93	0.0075	$4.6 \times 10^{-18}$
2	0.0026	$1.3 \times 10^{-8}$	$1.2 \times 10^{-62}$
3	$3.5 \times 10^{-8}$	$3.7 \times 10^{-3}$	$4.0 \times 10^{-13}$
4	$1.2 \times 10^{-10}$	$3.7 \times 10^{-3}$	$1.2 \times 10^{-10}$
5	$1.1 \times 10^{-5}$	0.25	0.98
6	$1.2 \times 10^{-10}$	0.061	0.96



i	$P(Z=MM X)$	$P(Z=mM X)$	$P(Z=mm X)$
1	1.00	0.00	0.00
2	1.00	0.00	0.00
3	0.00	1.00	0.00
4	0.00	1.00	0.00
5	0.00	0.67	0.33
6	0.00	0.34	0.66
$\Sigma$	2.00	3.01	0.99

Expected genotypes

$$\theta^{(0)} = (\theta_M^{(0)} \theta_m^{(0)}) = (0.2, 0.8)$$

$$\theta_m^{(1)} = \frac{3.01 + 2 \times 0.99}{2 \times 2.00 + 2 \times 3.01 + 2 \times 0.99}$$



# EM algorithm assuming HWE

## Log likelihood

$$\log(L(\theta)) = \log(p(X|\theta)) = \sum_i \log \left( \sum_z p(X_i, Z_i = z|\theta) \right) = \sum_i \log \left( \sum_z p(X_i|Z_i = z)p(Z_i = z|\theta) \right)$$

$$\theta_m + \theta_M = 1 \quad \text{and} \quad P(X|Z, \theta) = P(X|Z) \quad \text{and} \quad P(Z_i|\theta) = B(Z_i; n = 2, \theta)$$

## E step (Q)

$$Q_i(Z_i = z) = p(Z_i = z|X_i, \theta^{(n)}) = \frac{p(X_i|Z_i = z, \theta^{(n)})p(Z_i = z|\theta^{(n)})}{\sum_{z'} p(X_i|Z_i = z', \theta^{(n)})p(Z_i = z'|\theta^{(n)})}$$

## M step

$$\theta_m^{(n+1)} = \frac{\sum_i (0 \times Q_i(Z = MM) + 1 \times Q_i(Z = mM) + 2 \times Q_i(Z = mm))}{\sum_i \sum_z Q_i(Z_i = z)}$$

$X$  is all of the data for a single site  
 $X_i$  is the data for individual i

$\theta$  is the frequency of the two alleles  
 $\theta = (\theta_M, \theta_m)$

$N$  is the number of individuals

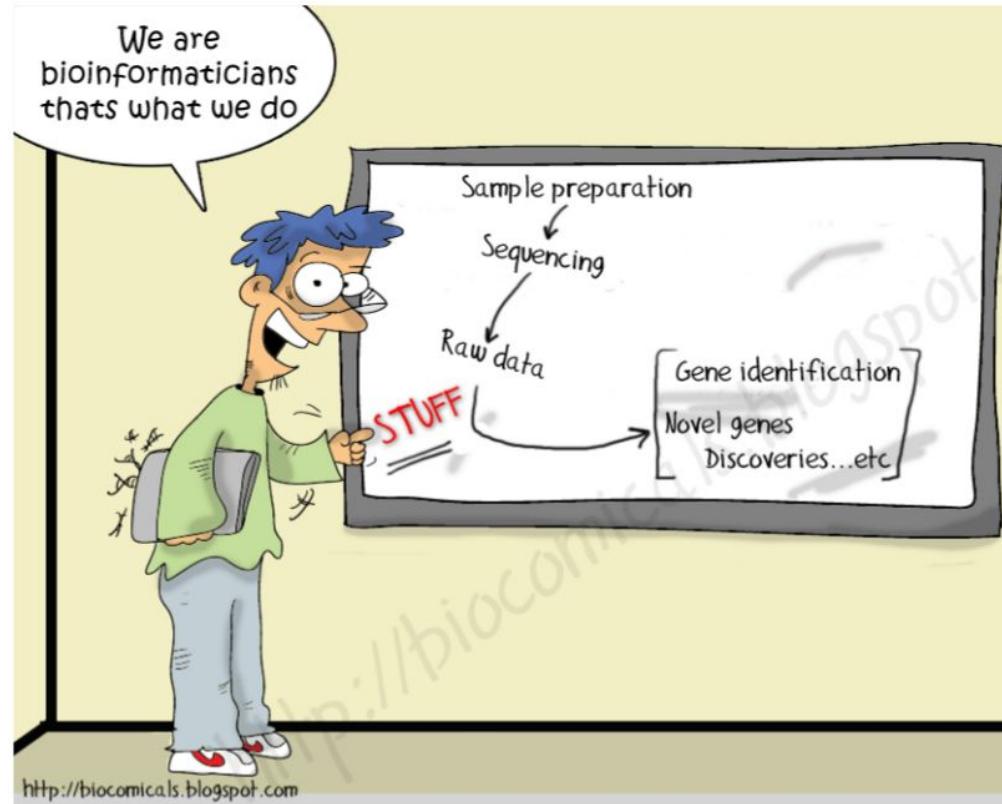
$z$  is latent state of the genotype

$Z_i$  is the genotype for individual i

$Z_i \in \{MM, Mm, mm\}$

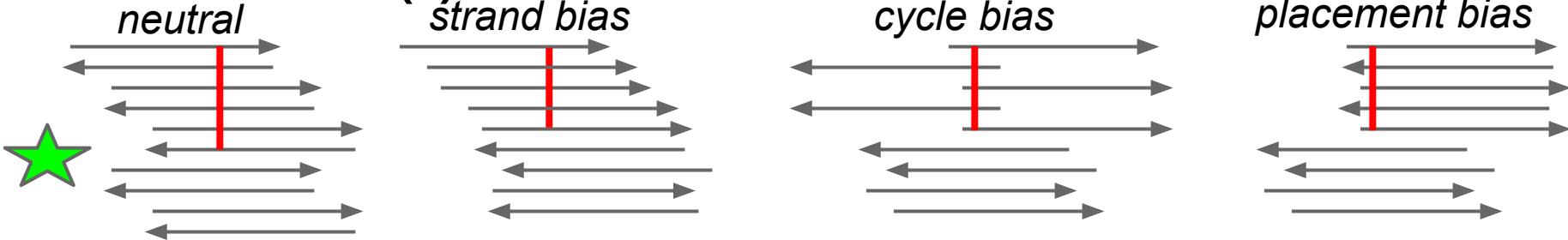


# Recap: Analysis of NGS data

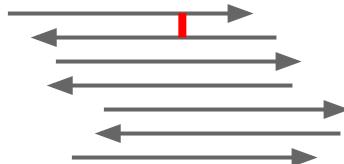


# Are our locus and alleles sequenceable?

In WGS, biases in the way we observe an allele (placement, position, strand, cycle, or balance in heterozygotes) are often correlated with error. We include this in our posterior  $P(\mathbf{G}, \mathbf{S} | \mathbf{R})$ , and to do so we need an estimator of  $P(\mathbf{S})$ .



*allele imbalance*



$$P(S) \propto$$

$$\times \prod_{\forall b \in \{B\}}$$

$$\begin{aligned} & \text{multinom}(|R \equiv b| \forall b_1, \dots, b_K); |\{R\}|, f_i, \dots, f_K) \\ & \text{binom}(|\text{forwardStrand}(\{R \equiv b\})|; |\{R \equiv b\}|, 1/2) \\ & \times \text{binom}(|\text{placedLeft}(\{R \equiv b\})|; |\{R \equiv b\}|, 1/2) \\ & \times \text{binom}(|\text{placedRight}(\{R \equiv b\})|; |\{R \equiv b\}|, 1/2) \end{aligned}$$

Mapper does not know our data e.g. it does not know whether it is haploid/diploid/pooled individual ect.

GCGGGAGTGTCCGGGAATAA.T.T.AAAA.CGATGCACACAGGGTTAGCGCGTA

ggAGGG|CCGGGAATAA.T.TAAAAA.CGATGcaca  
gAGTGTGCGGGAAATAA.TCA.AAAA.CGATGcacaaccg  
gAGTGTCCGGGAATAA.T.TAAAAA.CGATGCACACAG  
gAGTGTCCGGGAATAA.T.TAAAAA.CGATGCACACAg  
gAGTGGGCGGGAAATAA.TCA.AAAA.CGATGcacaaccg  
aGTG|GCGGGAAATAA.TCA.AAAA.CGATGCACACCgg  
aGTGTCCGGGAATAA.T.TAAAAA.CGATGCACACAgg  
aGTGTCCGGGAATAA.T.TAAAAA.CGATGCACACAgg  
gtgtGGGGAAATAA.TCA.AAAA.CGATGCACACCCggg



**TATATTAAT**GCGCGCGC**TAGGCTAGCT**

TATATTAAT--**GCGCGC**TAGGCTAGCT

TATATTAAT**GCGCGC**--TAGGCTAGCT

TATATTAAT**GCGCGC**.....

.....**GCGCGC**TAGGCTAGCT



### Read Pairs (RP)



### Read Depth (RD)



Unfortunately there is no program or pipeline that does all these things!!

### Split Reads (SR)



### Assembly (AS)



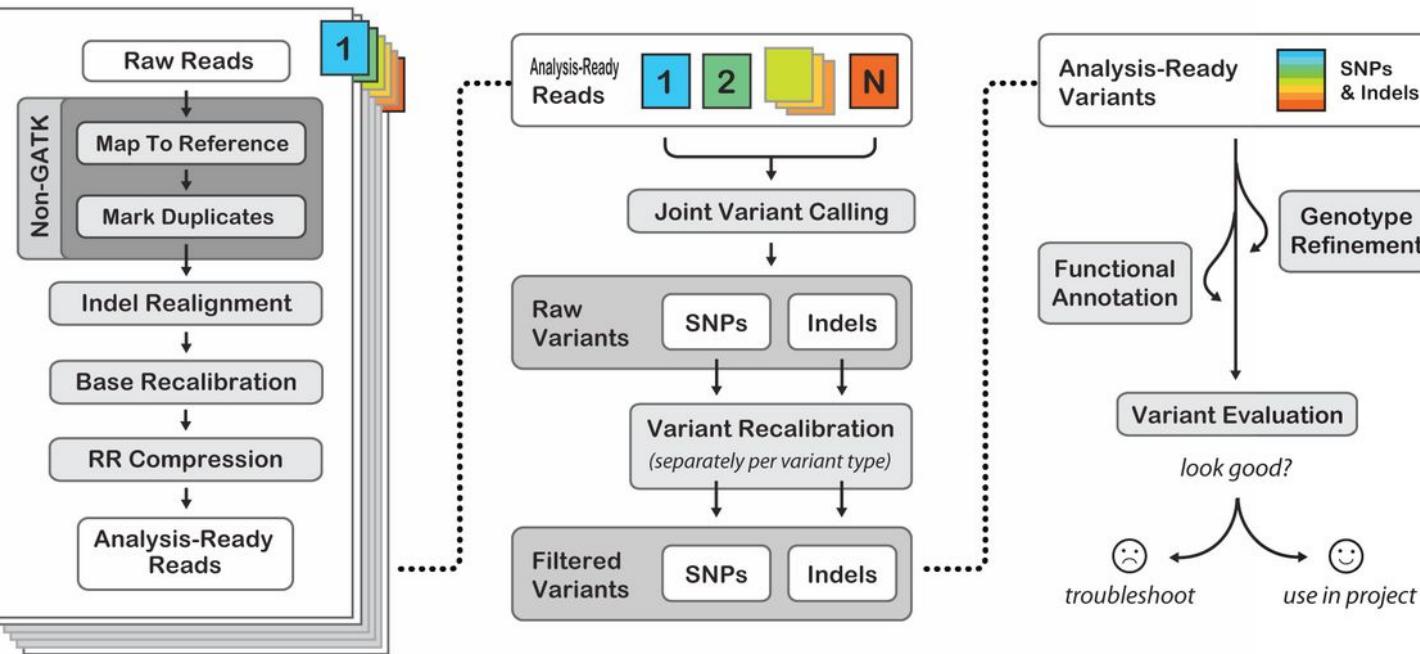
## Data Pre-processing

&gt;&gt;

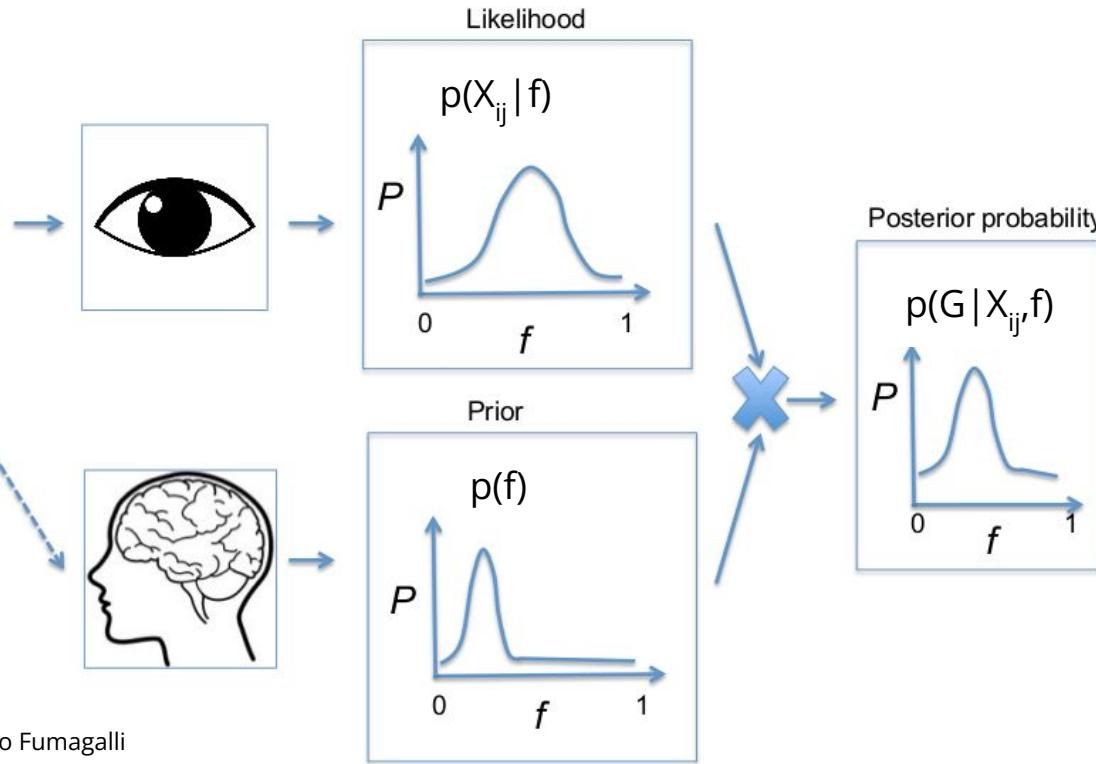
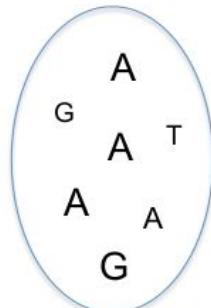
## Variant Discovery

&gt;&gt;

## Preliminary Analyses



**Data  $X_{ij}$**



Slide from matteo Fumagalli



## The likelihood for allele frequencies

### The likelihood

$$\underbrace{p(X | \theta)}_1$$

- ➊ Probability of data given parameters

### notation

$X$  is all of the data for a single site

$X_i$  is the data for individual  $i$

$\theta$  is the frequency of the two alleles

$\theta = (\theta_M, \theta_m)$

$N$  is the number of individuals



## The likelihood

$$\underbrace{p(X | \theta)}_1 = \prod_{i=1}^N \underbrace{p(X_i | \theta)}_2$$

- ① Probability of data given parameters
- ② Assumes individuals are independent

## notation

$X$  is all of the data for a single site

$X_i$  is the data for individual  $i$

$\theta$  is the frequency of the two alleles

$\theta = (\theta_M, \theta_m)$

$N$  is the number of individuals



## The likelihood

$$\underbrace{p(X | \theta)}_1 = \underbrace{\prod_{i=1}^N p(X_i | \theta)}_2 = \underbrace{\prod_{i=1}^N \sum_z p(X_i | Z_i = z) p(Z_i = z | \theta)}_3$$

- ① Probability of data given parameters
- ② Assumes individuals are independent
- ③ Introduce latent variable using law of total probability

## notation

$z$  is latent state of the genotype

$Z_i$  is the genotype for individual  $i$

$Z_i \in \{MM, Mm, mm\}$

## notation

$X$  is all of the data for a single site

$X_i$  is the data for individual  $i$

$\theta$  is the frequency of the two alleles

$\theta = (\theta_1, \dots, \theta_n)$



## The likelihood

$$\underbrace{p(X | \theta)}_1 = \underbrace{\prod_{i=1}^N p(X_i | \theta)}_2 = \underbrace{\prod_{i=1}^N \sum_z p(X_i | Z_i = z) p(Z_i = z | \theta)}_3$$

- ① Probability of data given parameters
- ② Assumes individuals are independent
- ③ Introduce latent variable using law of total probability

$$p(X_i | \theta, Z_i = z) = p(X_i | Z_i = z)$$

$p(X_i | Z_i = z)$  is the genotype likelihood



## The likelihood

$$\underbrace{p(X | \theta)}_1 = \underbrace{\prod_{i=1}^N p(X_i | \theta)}_2 = \underbrace{\prod_{i=1}^N \sum_z p(X_i | Z_i = z) p(Z_i = z | \theta)}_3$$

- ① Probability of data given parameters
- ② Assumes individuals are independent
- ③ Introduce latent variable using law of total probability

$$p(Z_i = z | \theta)$$

Binomial with parameters

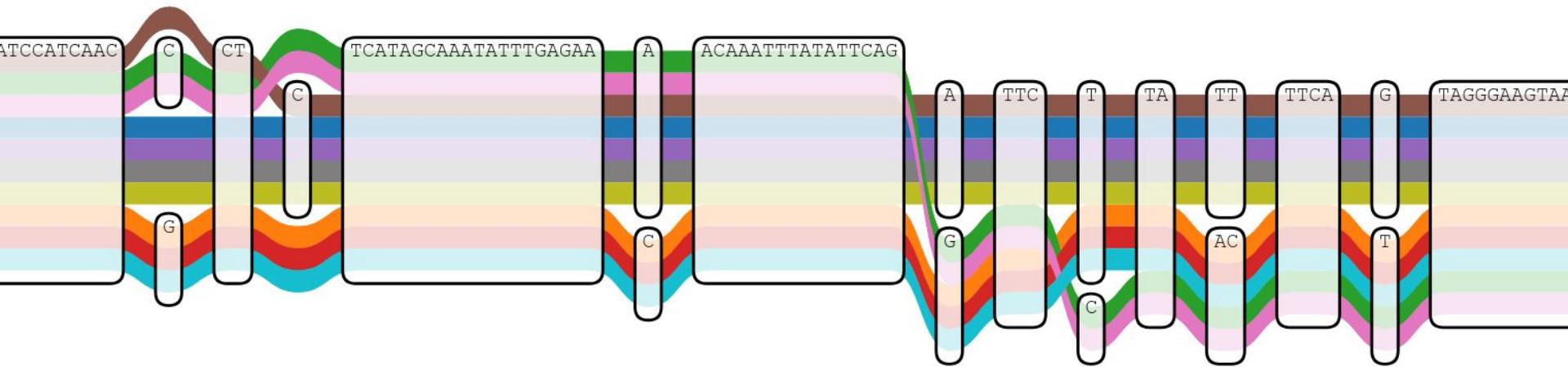
$$n = 2, p = \theta_m, k = \#m$$

$$p(X_i | \theta, Z_i = z) = p(X_i | Z_i = z)$$

$p(X_i | Z_i = z)$  is the genotype likelihood

$$p(X_i | \theta_z) = B(k : n, p)$$





## Variation graph

<https://vgteam.github.io/sequenceTubeMap/>



# Indel mapping is not consistent

TATATTAAT**GCGCGCGC**TAGGCTAGCT  
TATATTAAT--**GCGCGC**TAGGCTAGCT  
TATATTAAT**GCGCGC**--TAGGCTAGCT  
TATATTAAT**GCGCGC**.....  
.....**GCGCGC**TAGGCTAGCT

