

Inference from NGS data

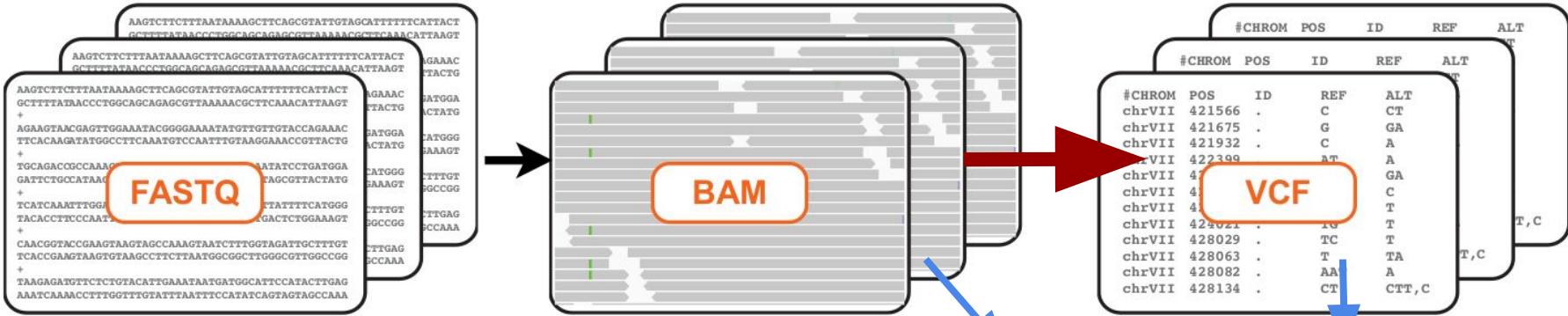
Anders Albrechtsen



Many type of sequencing



This session

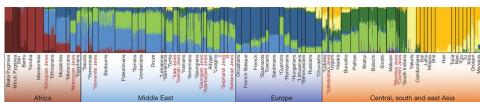


This morning

- Genotype likelihoods (GL)
- Estimate allele frequencies
- Calling variable sites
- Calling genotype

Later

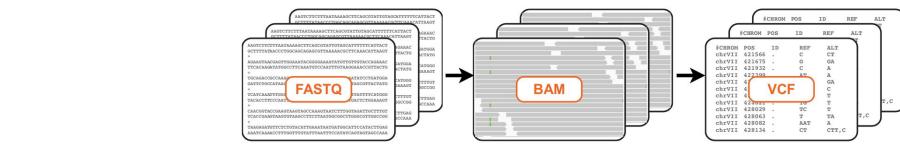
- Perform analysis from Genotypes or GLs



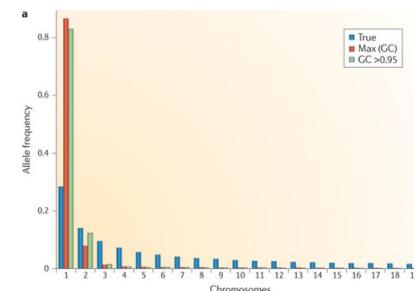
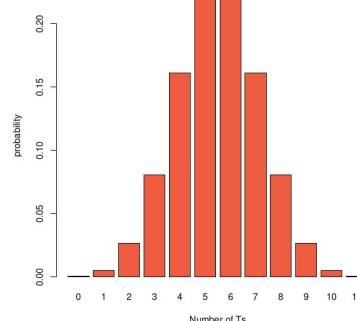
This session

This morning

- File formats and NGS data
- 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



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



Data formats

Genome (FASTA)

```
>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary  
reference assembly  
TGGAAAGAGGCCCTCAGCAGGCCACCTGGAGGGAGAGCACAGACTCGGGCTGAGGATGCAGGGCTCC  
CGGGCACGGTGCATGCCCTTGAGACACCCCAGAGCTGTGGGAAGAGCTGTGGGATCCCTATTGC  
ATCACAAAGCGGCCCTGGAGGGCTGGCTTTATTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT  
TAATCCGCACGCTTAGACTCCCCGGCTGTGATTTTGACAATGGCTCGGGGTCGAAAGCGGGCTG  
TCTGGGGAGTTGGACCCGGCACATGGTCAGCTCATGGTGGGACCTGAAATTCAAGGCTCCCTAG
```



Reads (FASTQ)

```
CCAATGATTTTTCCGTGTTTCAAATACGGTTAA  
+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  
GTTCAAAAAGAACTAAATTGTGTCAATAGAAACTC  
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

Mapped Reads (mpileup, BAM)

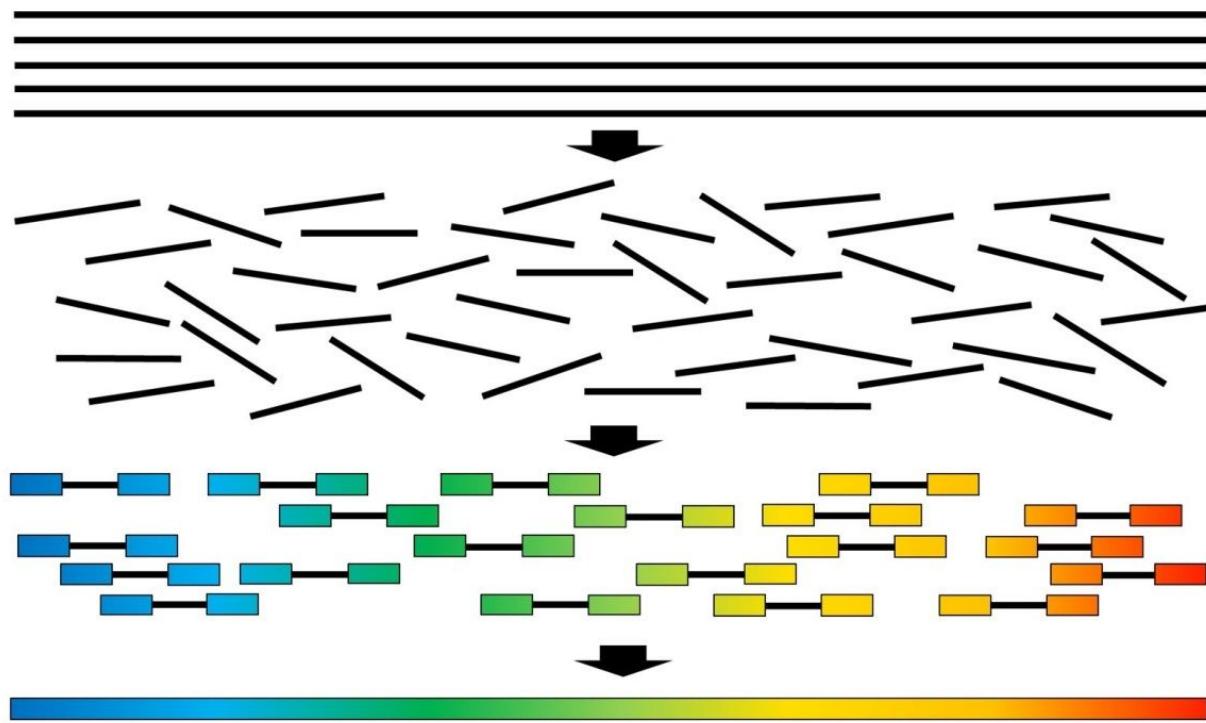
```
seq1 272 T 24 ,.S.....,.....,..^+. <<<<:;<<<<<<<<<:,<;?<&  
seq1 273 T 23 ,.....,.....,..A <<<:;<<<<<<3=<<<:;<<+  
seq1 274 T 23 ,.S.....,.....,.. 7<7:<,<<<<<<<:,<:;<<<  
seq1 275 A 23 ,S.....,.....,..^1. <<;9<<<<<<:;<:;<<<  
seq1 276 G 22 ,...T.....,.....,.. 33;<<<7<7:<&<1;<<<  
seq1 277 T 22 .....,.C.....,..G. +7<:;<<<<<&=<<;:;<<<  
seq1 278 G 23 .....,.k.....,..k. %38<:;<:7<<7<=<<:;<<<  
seq1 279 C 23 A..T.....,.....,..;758<<<<<<=<<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 GT 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 880007 rs6681049 T C . . . . GT 1  
/1
```



Next generation sequencing



Multiple versions of the same genome (DNA from many cells)

Genome fractured into small fragments

Fragments sequenced and ordered according to position on the genome



Fragment library (input DNA sample)

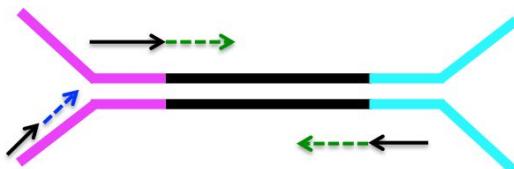
Library prep



Sequencing library

Double-stranded or Y- adaptors added

DNA sequencing



Barcode (6–12 bases) – so many samples can be run in one physical space (lane). Data is demultiplexed.

Primers

Reads (36–1000+ bases)



Single or pair of fq files.

single-end



independent reads

paired-end



two inwardly oriented
reads separated by ~200 nt

mate-paired



two outwardly oriented reads separated by ~3000 nt



fastQ (.fq.gz)

```
a`X_\Va\J`KaYJHG^]b\@a^BBBBBBBBBBBBBBB  
@FC42BF1AAXX:6:1:5:732#0/1  
TGATTCTCTCGATATCCAGTCCTAGTGNCATAGN  
+  
a^_aaaa`aa`_aaa_aaa`__```VBBBBBBBBB  
@FC42BF1AAXX:6:1:5:492#0/1  
AACAGTGGGAGGCTGCAGCAGGAGGATTNCTGAAN  
+  
ababb_abbbZbabaab^`aaTaabbaBBBBBBBBB  
@FC42BF1AAXX:6:1:5:480#0/1  
ACCTCCTCAGAGTTCTCGAGCTCGAGAANTCTGGN
```



<-- quality score

<-- read ID

<-- read (bases)



Selected issues with sequencing data

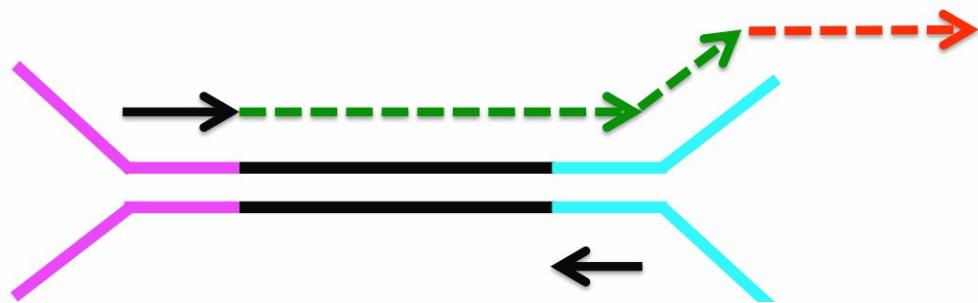
- **Adapter contamination of data**
 - If the DNA is too short we will sequence the adapters



KØBENHAVNS
UNIVERSITET

Adapter contamination

If the DNA fragment is too small you will sequencing into the adapter+junk



Solution: Identify the problem using fastQC and trim the 3' end of the read to remove the **adapter + junk (AAA...)** if needed



Selected issues with sequencing data

- Adapter contamination of data
 - If the DNA is too short we will sequence the adapters
- **Sequencing errors**
 - The reads by have errors



Sequencing error

FastQ file

```
a'X_\Va\J`KaYJHG^]b\a^BBBBBBBBBBBBBBB    <-- quality score  
@FC42BF1AAXX:6:1:5:732#0/1                  <-- read ID  
TGATTTCTCTCGATATCCAGTCCTTAGTGNCATAGN     <-- read (bases)  
TGACTCTCTCGATATCAAGTCCTTAGTGNCATAGN      <-- sequenced DNA fragment
```

Solution: Translate the quality score to error rates

Identify the scale of the problem using fastQC

Use the error rates when calling genotypes



KØBENHAVNS
UNIVERSITET

Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
\$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			



quality scores/Phred scores

```
a`x_\Va\J`KaYJHG^]b\`a^BBBBBBBBBBBBBBB    <-- quality score
```

Ascii	Dec	Qscore (Dec -33)	Error (ϵ)
+	43	10	10%
5	53	20	1%
?	63	30	0.1%
I	73	40	0.01%

Convert Qscores to sequencing error rates

$$\text{Qscore} = -10 \log_{10} (\epsilon) \Leftrightarrow \epsilon = 10^{-\text{Q}/10}$$



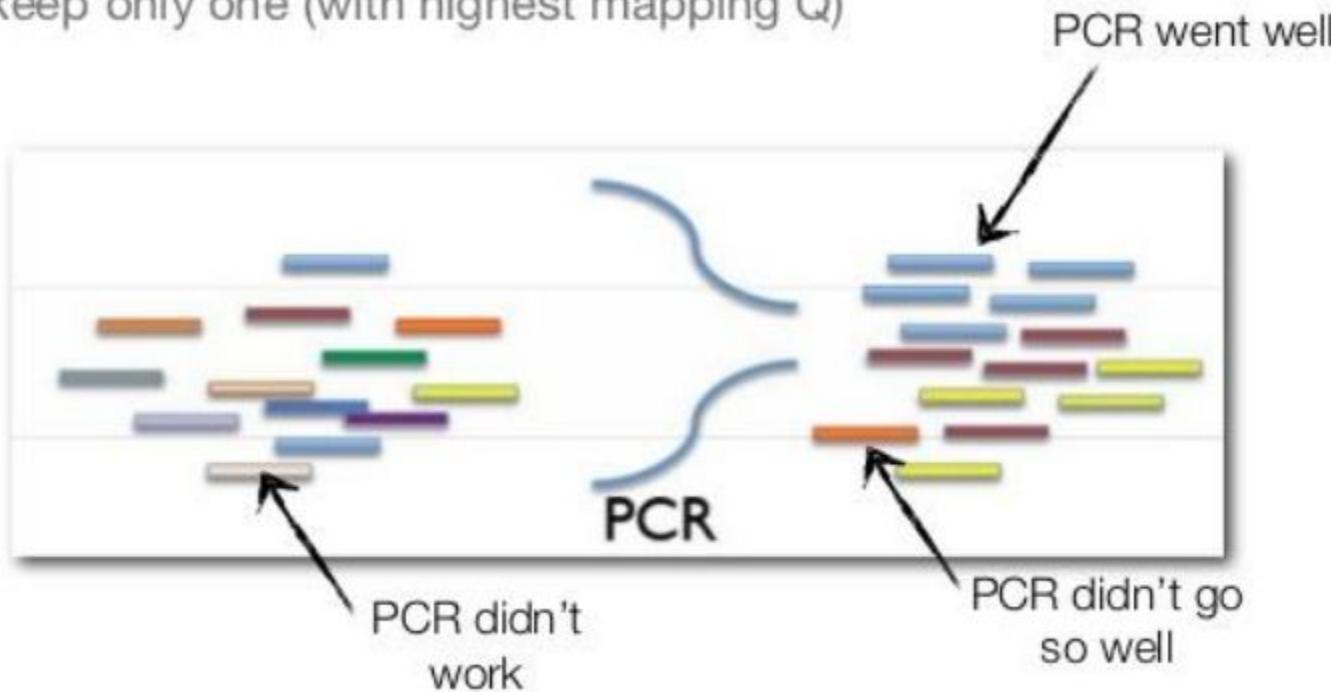
Selected issues with sequencing data

- Adapter contamination of data
 - If the DNA is too short we will sequence the adapters
- Sequencing errors
 - The reads by have errors
- **PCR or optical duplicates**
 - Reads can be duplicated ether from PCR or from the chip



Duplicated reads

> keep only one (with highest mapping Q)



Duplicated reads can cause



Solution: Identify the problem using fastQC

Identify the duplicated reads and remove or mark them



FAST QC

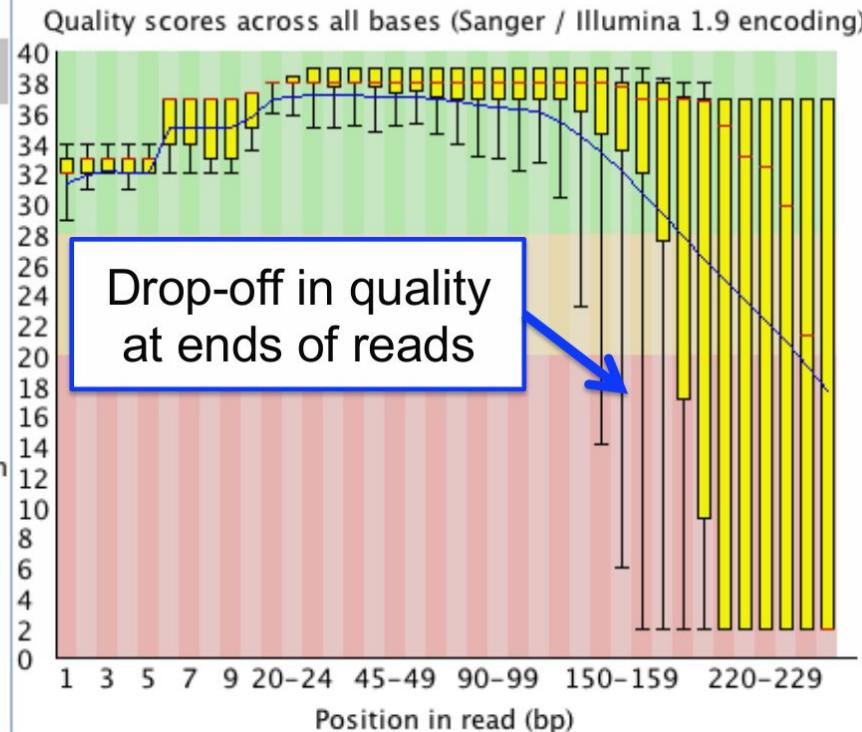
Easy to use tool for evaluating the quality of your data (fastQ or Bam files)



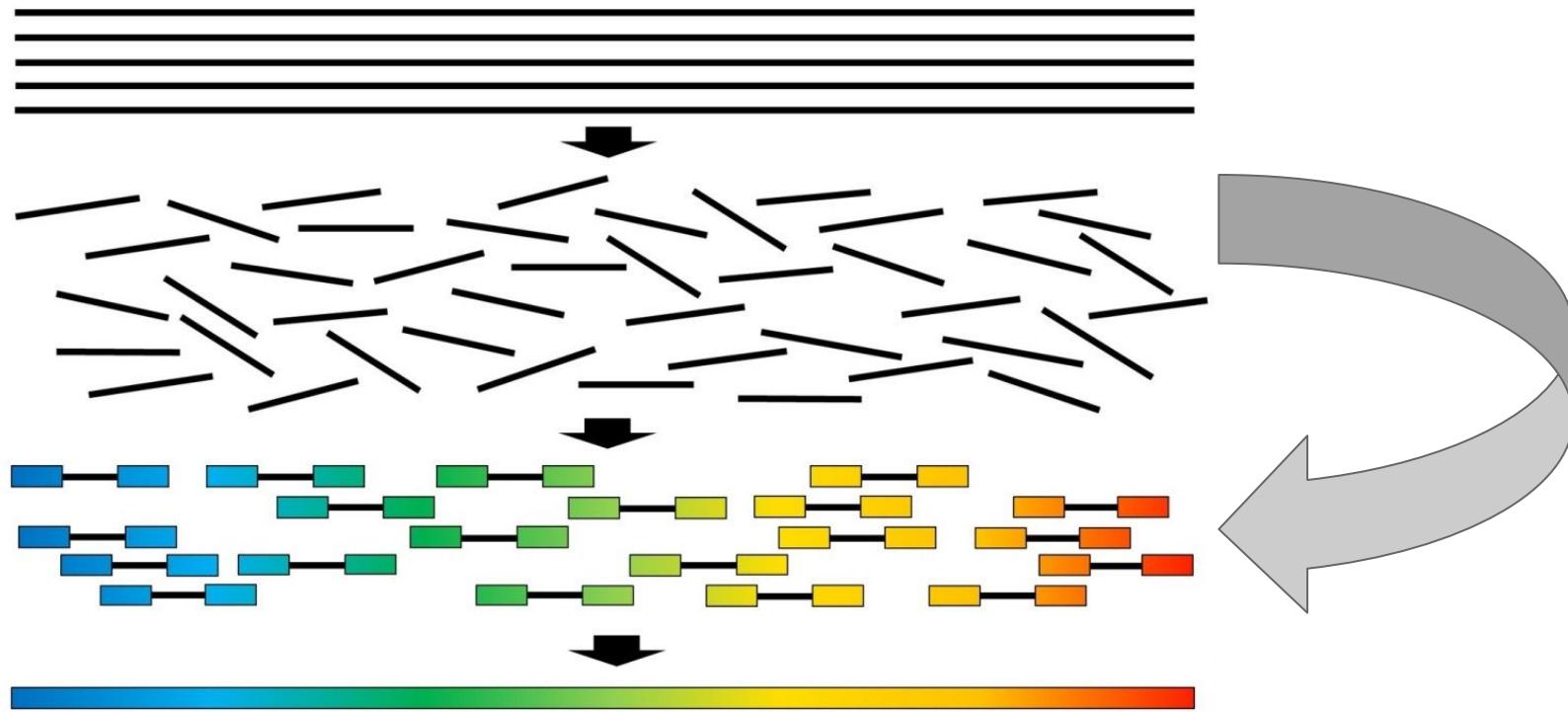
KØBENHAVNS
UNIVERSITET

Quality for each cycle

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Kmer Content

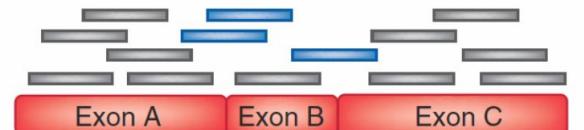
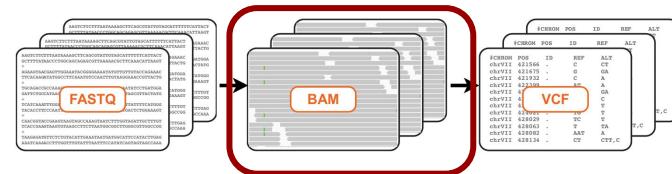


Mapping - alignment of reads



Alignment file (.bam)

```
reads      TTTGTTCTTCTTCTCTAGTCTTCTT ...
Qscore     NVFVN]^`^_]]^U]]'_[_VS[_^Z]_ ...
Position   chr4 53351385
Mismatch   2
strand    +
mapQ      30
Mate       mapped chr4 53351145
Alt map   chr2 15331145 with 2 mismatch
```



Processed mRNA



Mapping to genome



Mapping quality

Mapping quality – what is the probability that the read is correctly mapped to this location in the reference genome?

Read 1

ATCGGGAGATCC or ATCGGGAGATCC
||||| ||||| |||||
...TAATCGGGAGATCCGC ... TTATCGGGAGATCCGC TAGCCTAGTGTGCCGC . . .

Read 2

GCGTAGTCTGCC

|| | | | | | | |

Reference Sequence

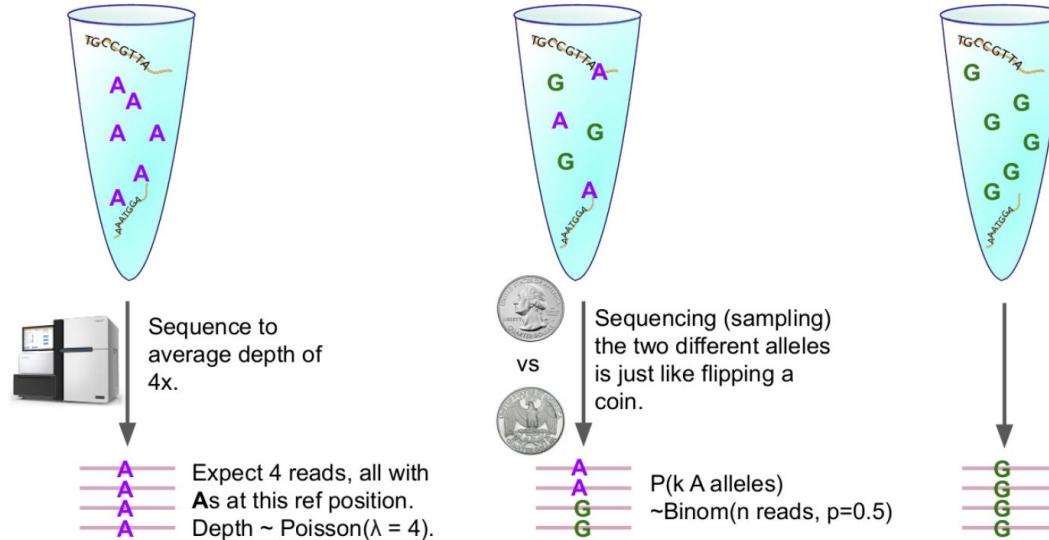
Read 1 can be mapped two places on the genome while **Read 2** only maps to one

- Which of the two reads has the highest mapping quality?



Why don't we observe genotype

Each allele is sequenced separately and alleles are sampled with replacement



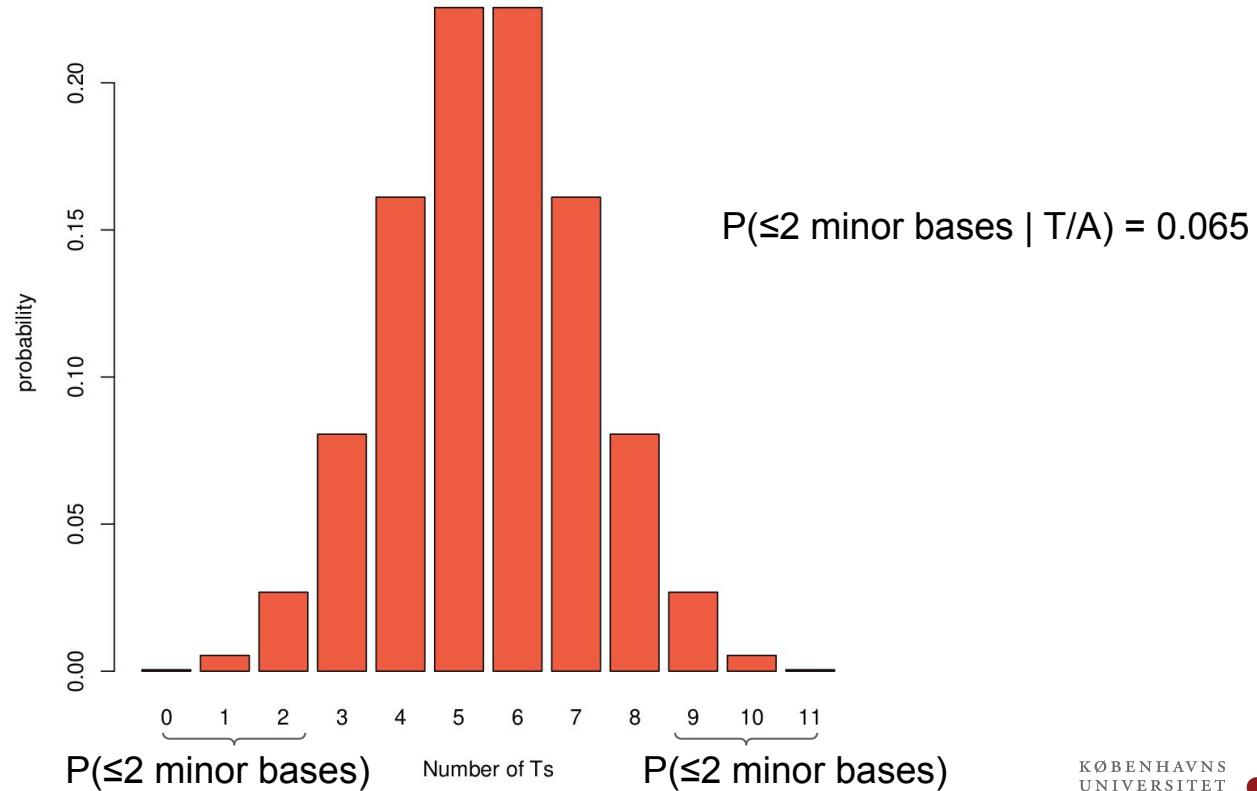
Why don't we observe genotype

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

AGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAACACTG
CAGCCACACCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT
CAGCCACACCCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT
TGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAC
CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAA
GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTAC
TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT
CATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCAOGGTACCAGACAGA
AGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA

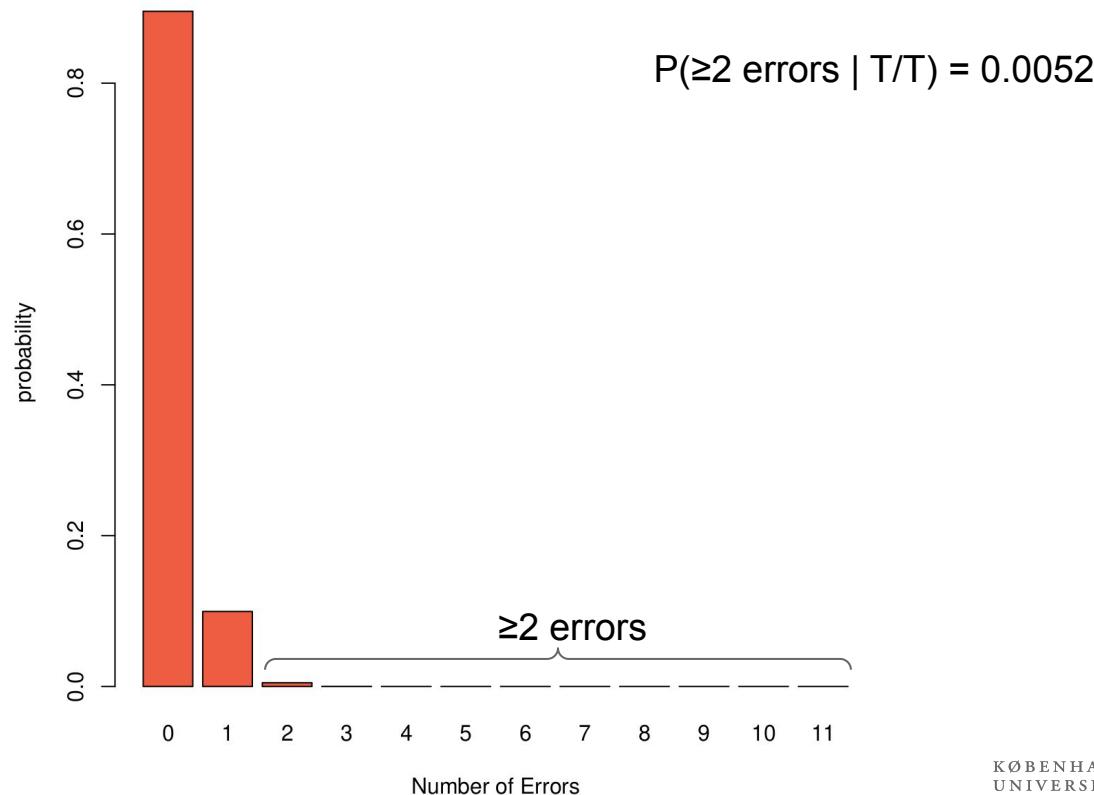
Assuming heterozygous (C/T)

ACATTCAC
ACACCCCG
ACACCCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC



Assuming homozygous (T/T)

ACATTCAC
ACACCCCG
ACACCCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC
ACATTCAC



Why don't we observe genotype

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

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

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

$P(\geq 2 \text{ errors} | T/T) = 0.052$

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

Heterozygosity is 0.1%

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

AGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAACACTG
CAGCCACACCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT
CAGCCACACACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAACACT
TGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAC
CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTACCAGACAGAAATCTCTTGCTAAA
GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCAOGGTAC
TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT
CATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCAOGGTACCAGACAGA
AGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA

Genotype likelihoods

how to quantify genotype uncertainty



KØBENHAVNS
UNIVERSITET

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 ϵ_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 ϵ_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 ϵ_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}	ϵ_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

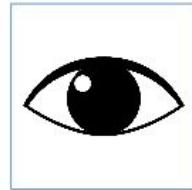


KØBENHAVNS
UNIVERSITET

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

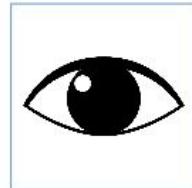
	A	C	G	T
A	0	0	0	0
C	0	0	0.922	
G	0		0	
T			0.077	

Modified slide from
matteo Fumagalli

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	
	0	0	0.00025		
C	0	0	0		
			0.00002		
G	0	0	0		
				0.00002	
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	
	0	0	0.20		
C	0	0	0		
			0		
G	0	0	0		
				0.80	
T	0.0	0	0.0		
				0.80	

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

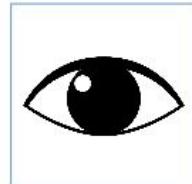


KØBENHAVNS
UNIVERSITET

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	0	0.00025		
G		0	0		
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

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

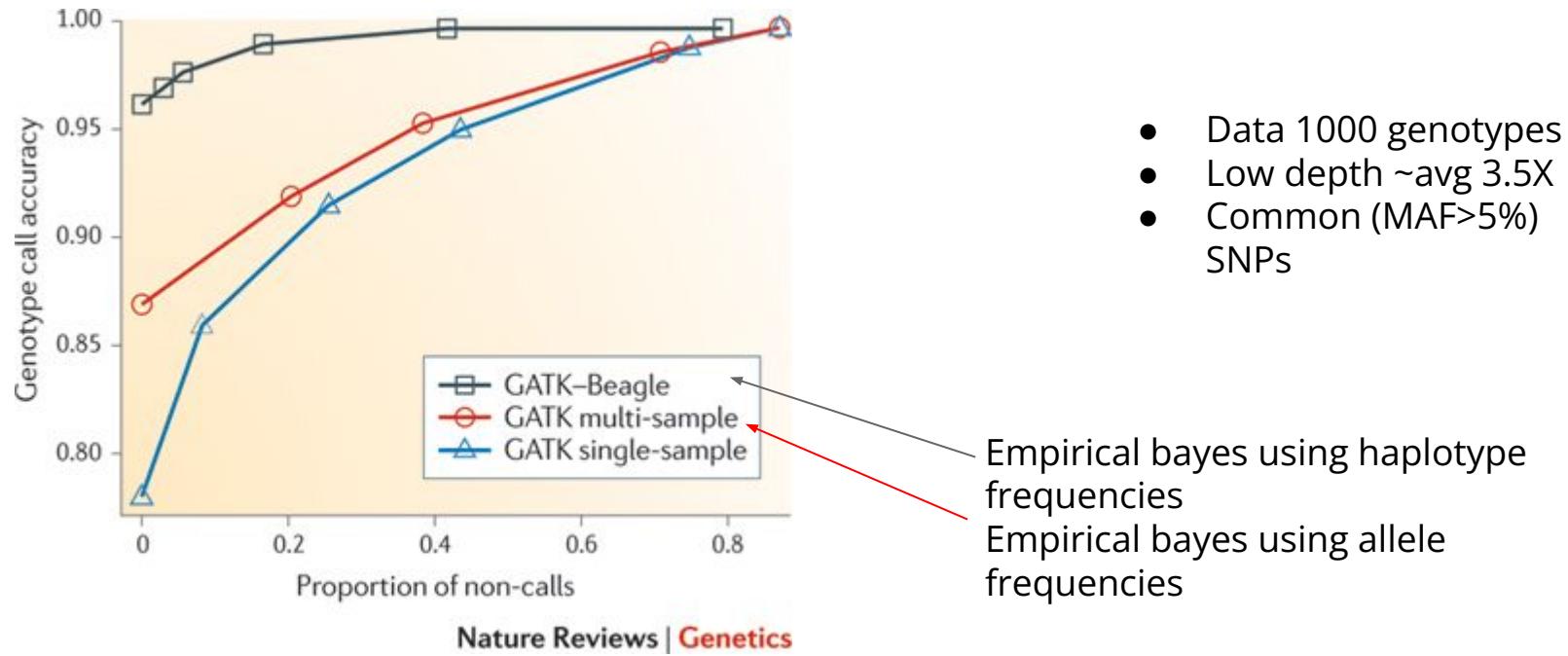
Posterior probability

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



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

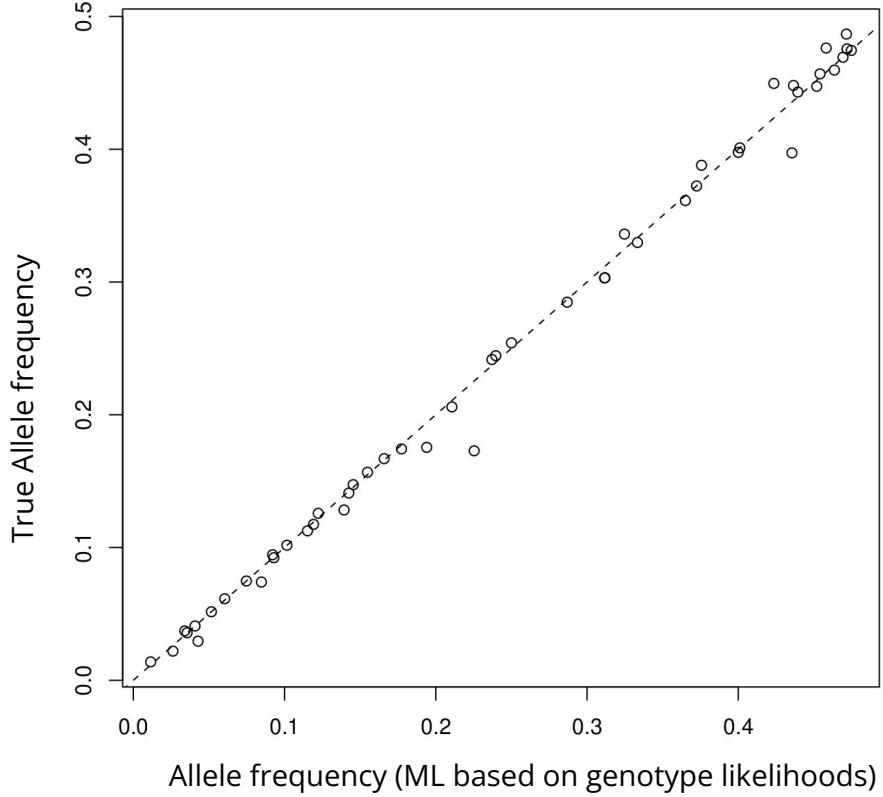
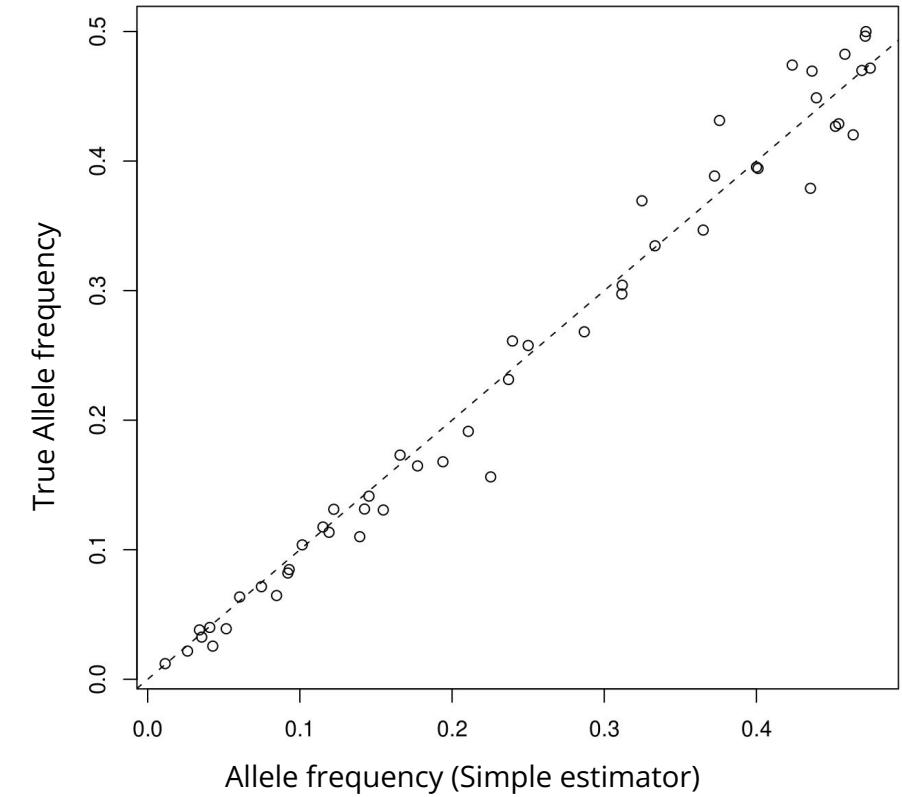
- assume only two alleles type exists
- $P(X^i|f) = \sum_{g \in \{0,1,2\}} P(X|G=g)P(G=g|f)$
- assume HWE e.g. $P(G=A_1A_1|f) = f^2$
- $\hat{f}_{ML} = \operatorname{argmax}_p \prod_i P(X^i|f)$

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

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





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)

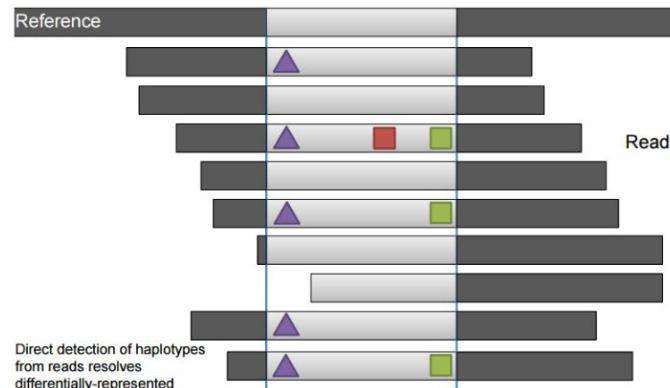


Figure from Erik Garrison

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:...
```

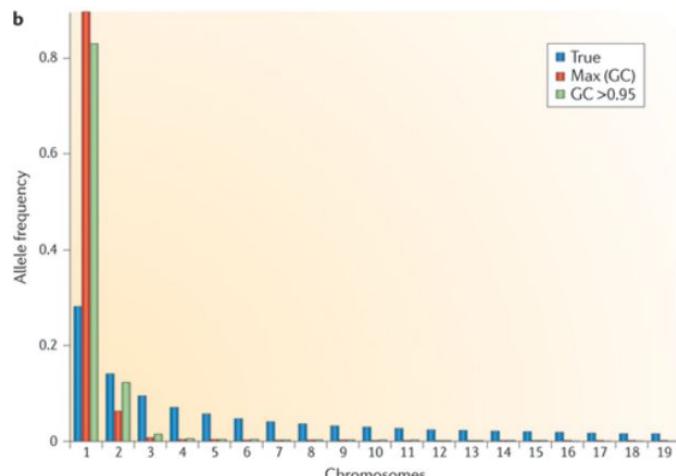
Body

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



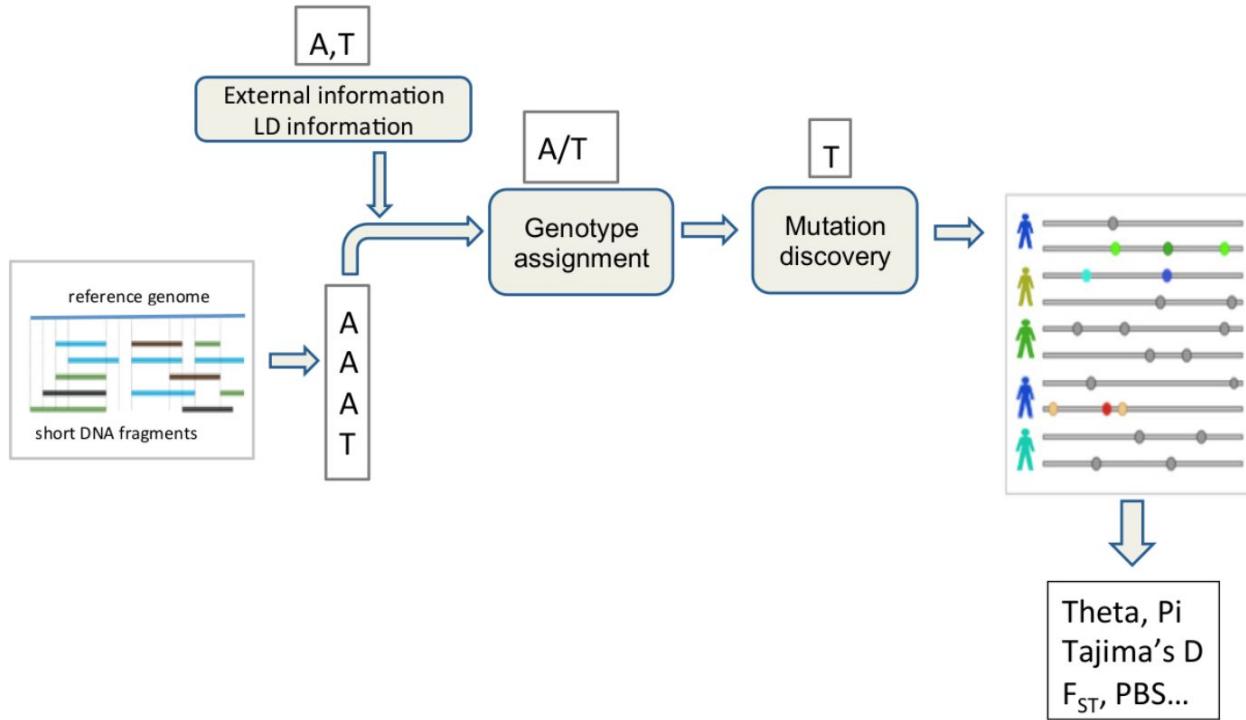
Nature Reviews | Genetics

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

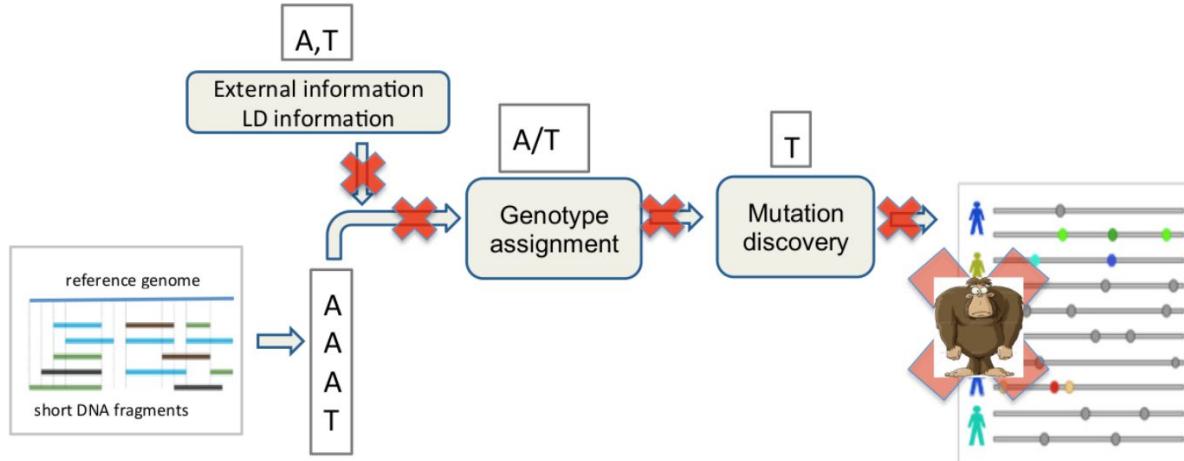


KØBENHAVNS
UNIVERSITET

High quality data



If you don't have high quality data

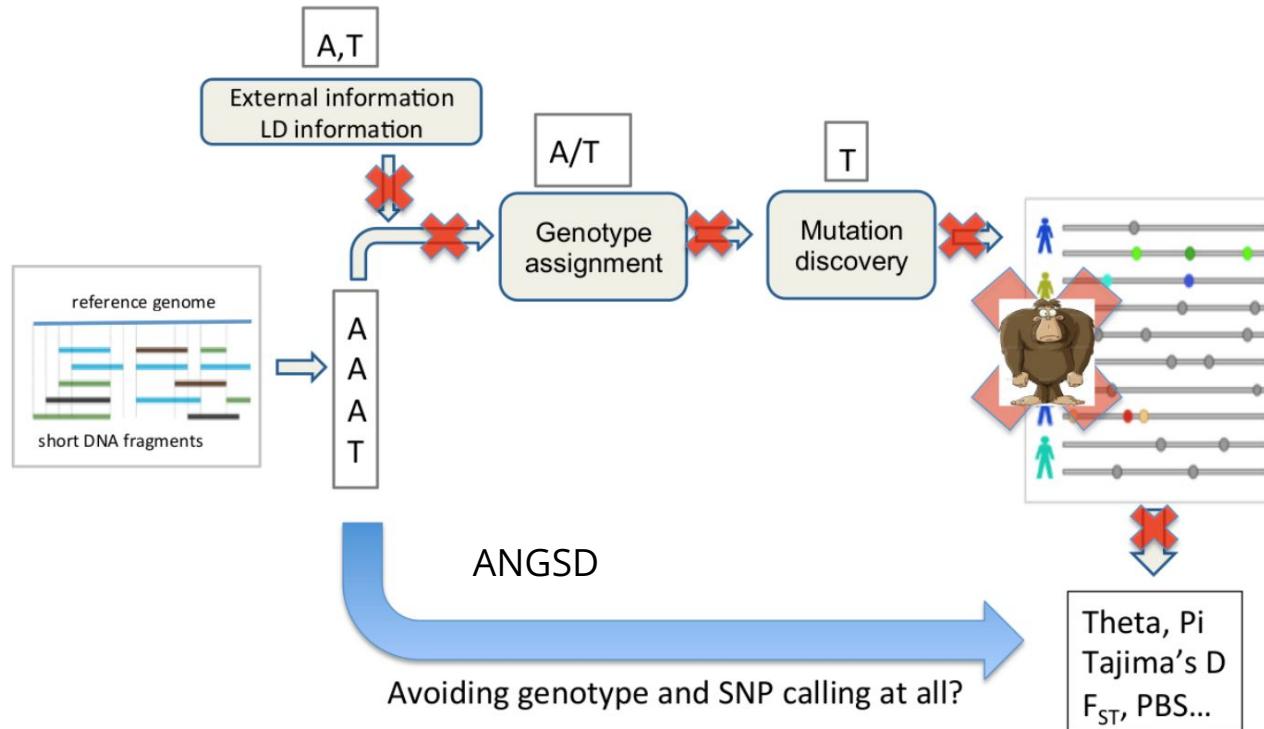


Got issues?

- 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



Time for exercises

Go to
tinyurl.com/kilifiNGS



KØBENHAVNS
UNIVERSITET

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	1.0×10^{-14}
2	0.0078	1.5×10^{-8}	9.9×10^{-51}
3	9.5×10^{-7}	3.9×10^{-3}	9.7×10^{-11}
4	9.6×10^{-9}	3.9×10^{-3}	9.6×10^{-9}
5	1.0×10^{-8}	0.25	0.98
6	1.0×10^{-8}	0.062	0.96

Notation

X is all of the data for a single site
 X_i is the data for individual i
 θ 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	4.6×10^{-18}
2	0.0026	1.3×10^{-8}	1.2×10^{-62}
3	3.5×10^{-8}	3.7×10^{-3}	4.0×10^{-13}
4	1.2×10^{-10}	3.7×10^{-3}	1.2×10^{-10}
5	1.1×10^{-5}	0.25	0.98
6	1.2×10^{-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
Σ	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

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

N is the number of individuals

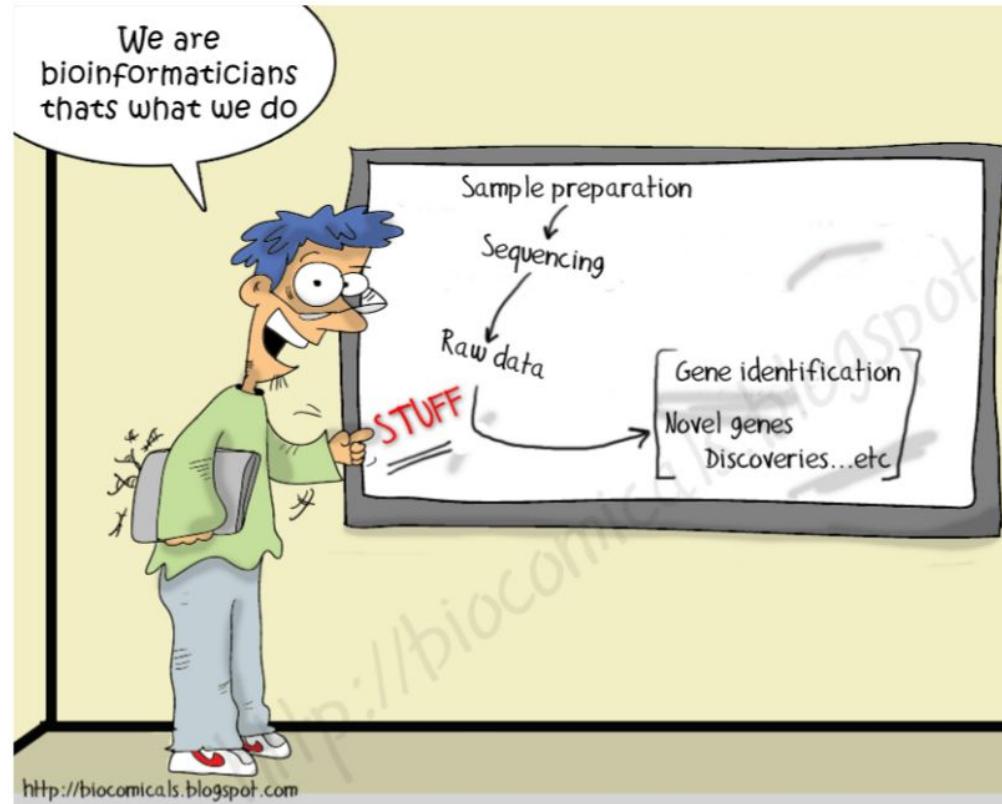
z is latent state of the genotype

Z_i is the genotype for individual i

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

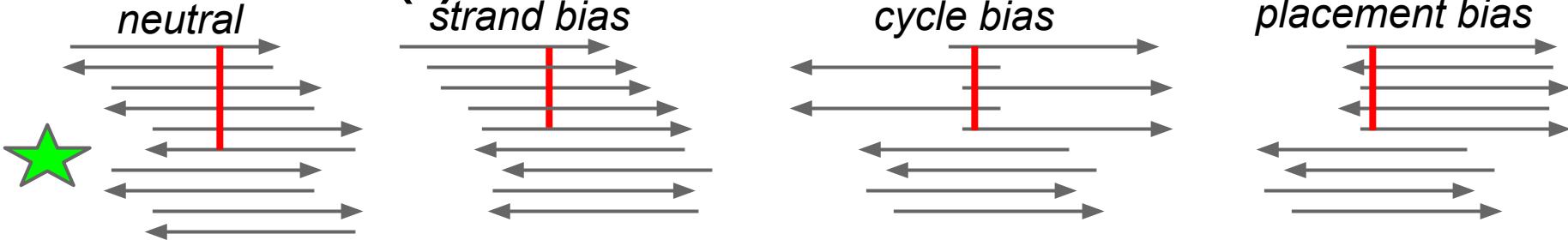


Recap: Analysis of NGS data

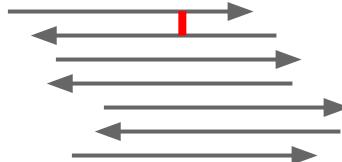


Are our locus and alleles sequenceable?

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

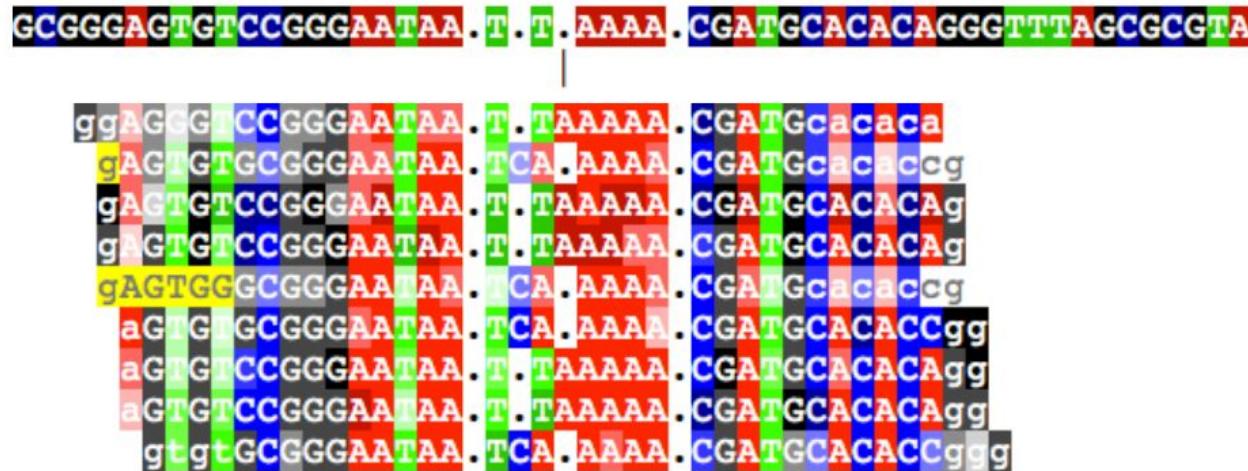


$$P(S) \propto$$

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

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

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



TATATTAATGCGCGCGC**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)



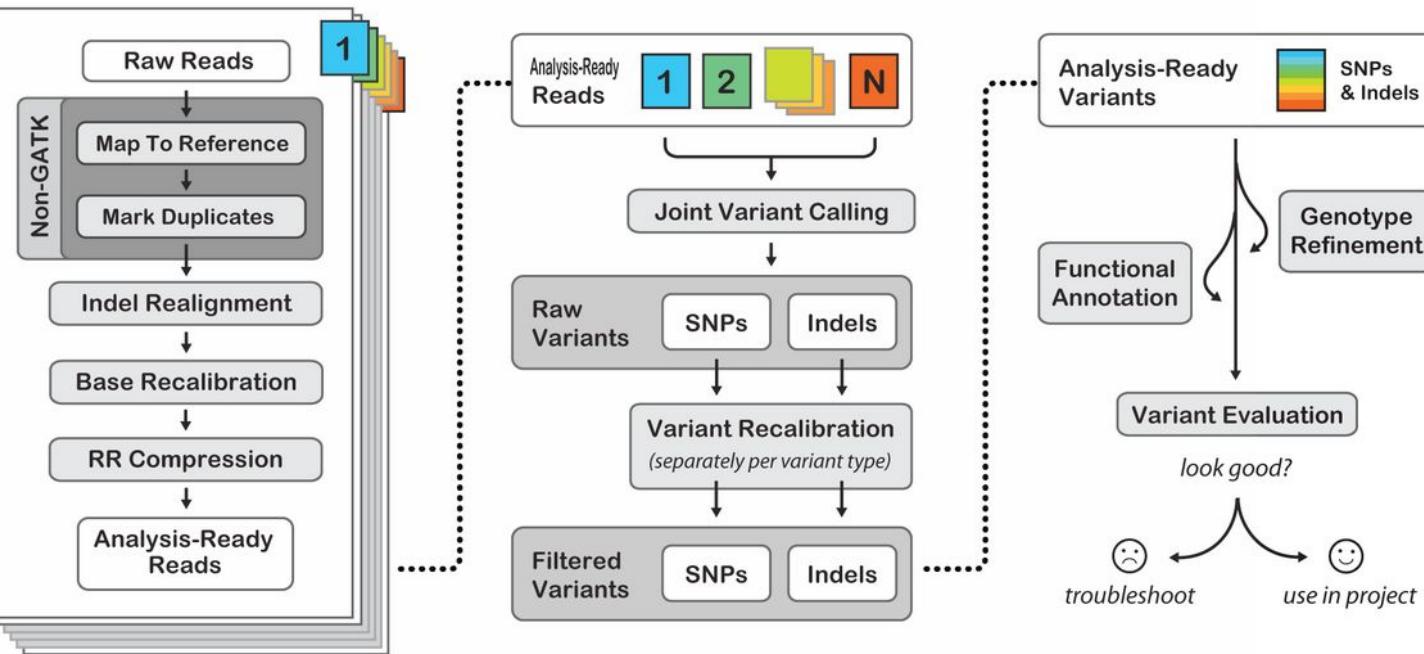
Data Pre-processing

>>

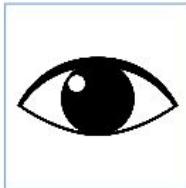
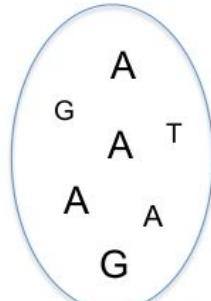
Variant Discovery

>>

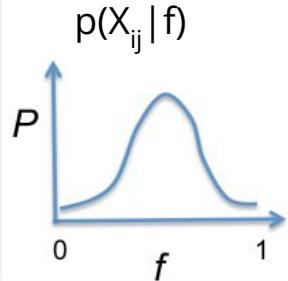
Preliminary Analyses



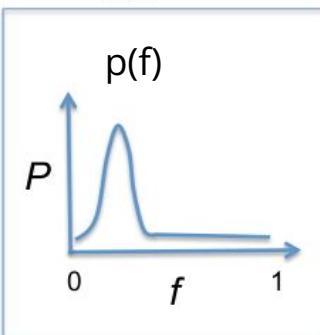
Data X_{ij}



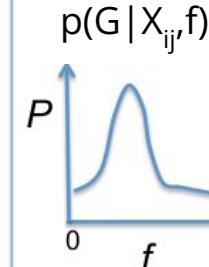
Likelihood



Prior



Posterior probability



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

θ 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

θ 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

θ is the frequency of the two alleles

$\theta = (\theta_1, \theta_2)$



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

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

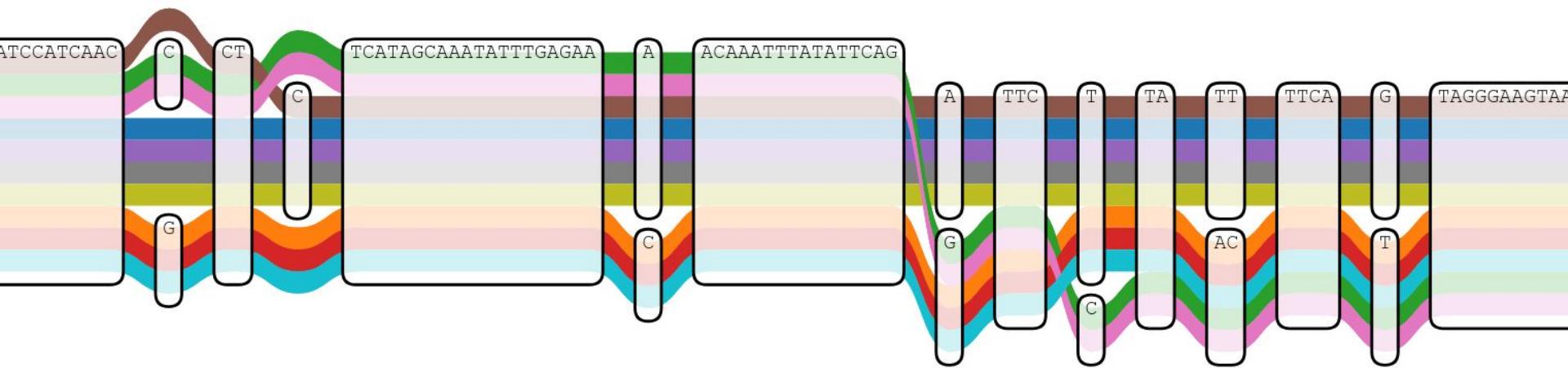
$$p(X_i | \theta, Z_i = z) = p(X_i | Z_i = z)$$

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

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

Hardy weinberg
NGS inference





Variation graph

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



Indel mapping is not consistent

TATATTAAAT**GCGCGCGC**TAGGCTAGCT
TATATTAAAT--**GCGCGC**TAGGCTAGCT
TATATTAAAT**GCGCGC**--TAGGCTAGCT
TATATTAAAT**GCGCGC**.....
.....**GCGCGC**TAGGCTAGCT

