Workshop on Population and Speciation Genomics, Cesky Krumlov

SNP and genotype calling (and more) – Day 3

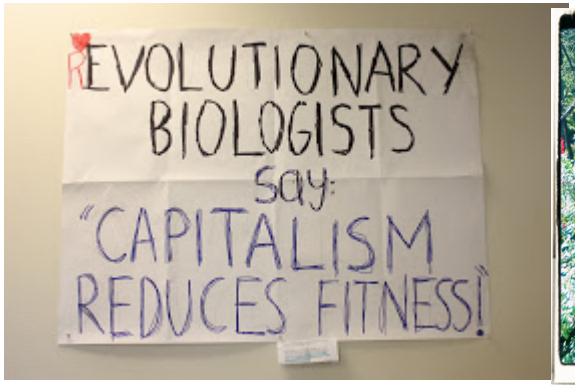
Matteo Fumagalli

January 27th 2016

Who I am

PhD in Bioengineering
Polytechnic University of Milan, Italy

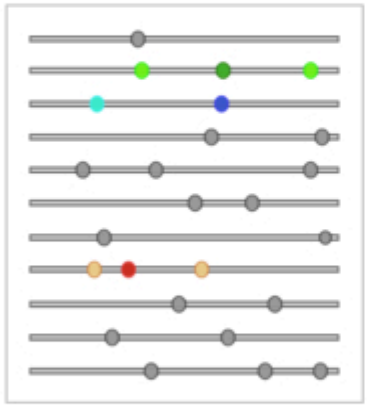
Postdoc (EMBO fellow)
Dept. Integrative Biology
Univ. of California, Berkeley, USA
R. Nielsen Group

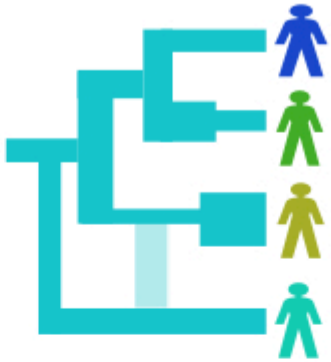


Postdoc (HFSP fellow)
UCL Genetics Institute
University College London, UK
F. Balloux Group

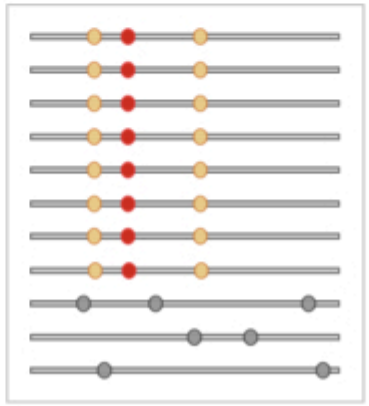
What I do

Population genetic diversity

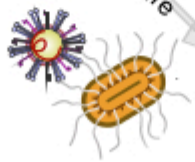




Signatures of natural selection



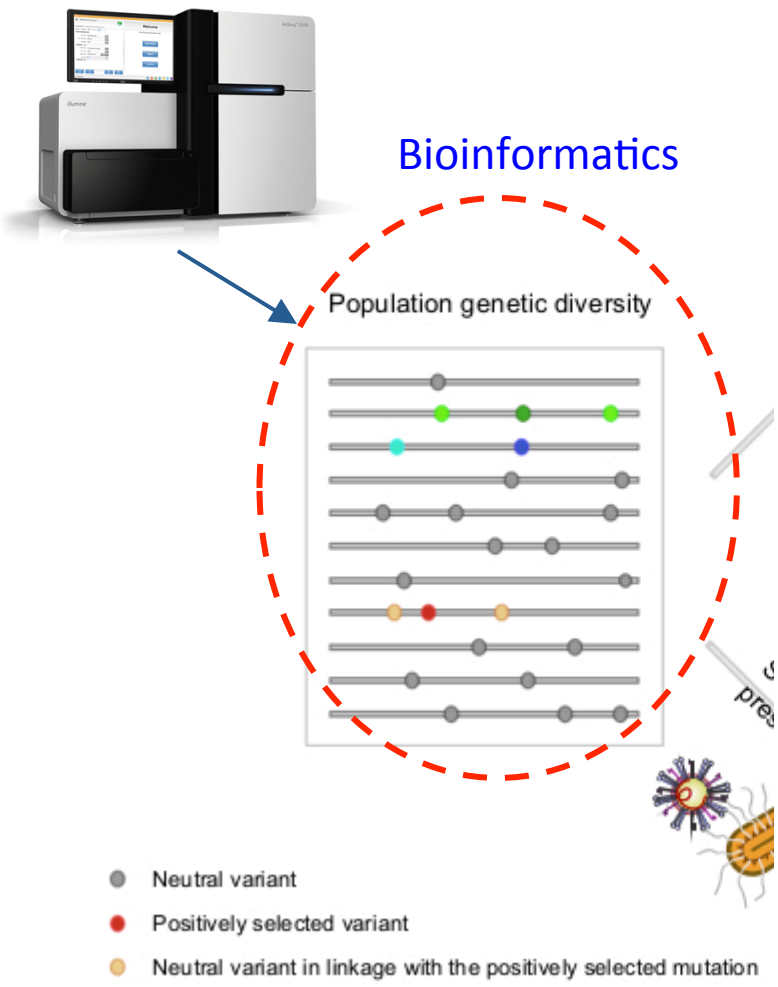
Selective pressure



- Neutral variant
- Positively selected variant
- Neutral variant in linkage with the positively selected mutation

adapted from Fumagalli and Sironi, Current Opinion in Immunology, 2014

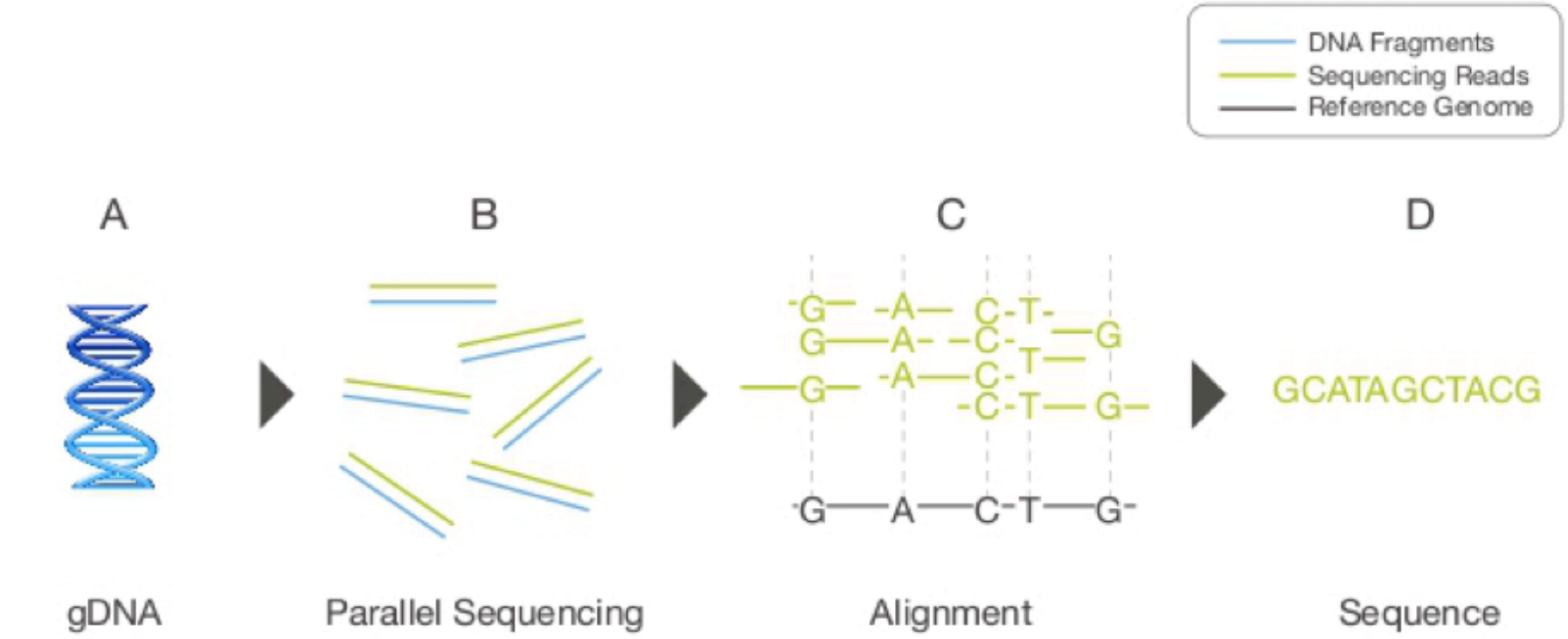
What I'll be talking about today



- Intro and basic filtering of NGS data
- Genotype calling
- SNP calling and estimation of allele frequencies
- Advanced methods for population genetic analyses for low-depth data
- Paper discussion
- Intro to practical exercises

adapted from Fumagalli and Sironi, Current Opinion in Immunology, 2014

Next-Generation Sequencing

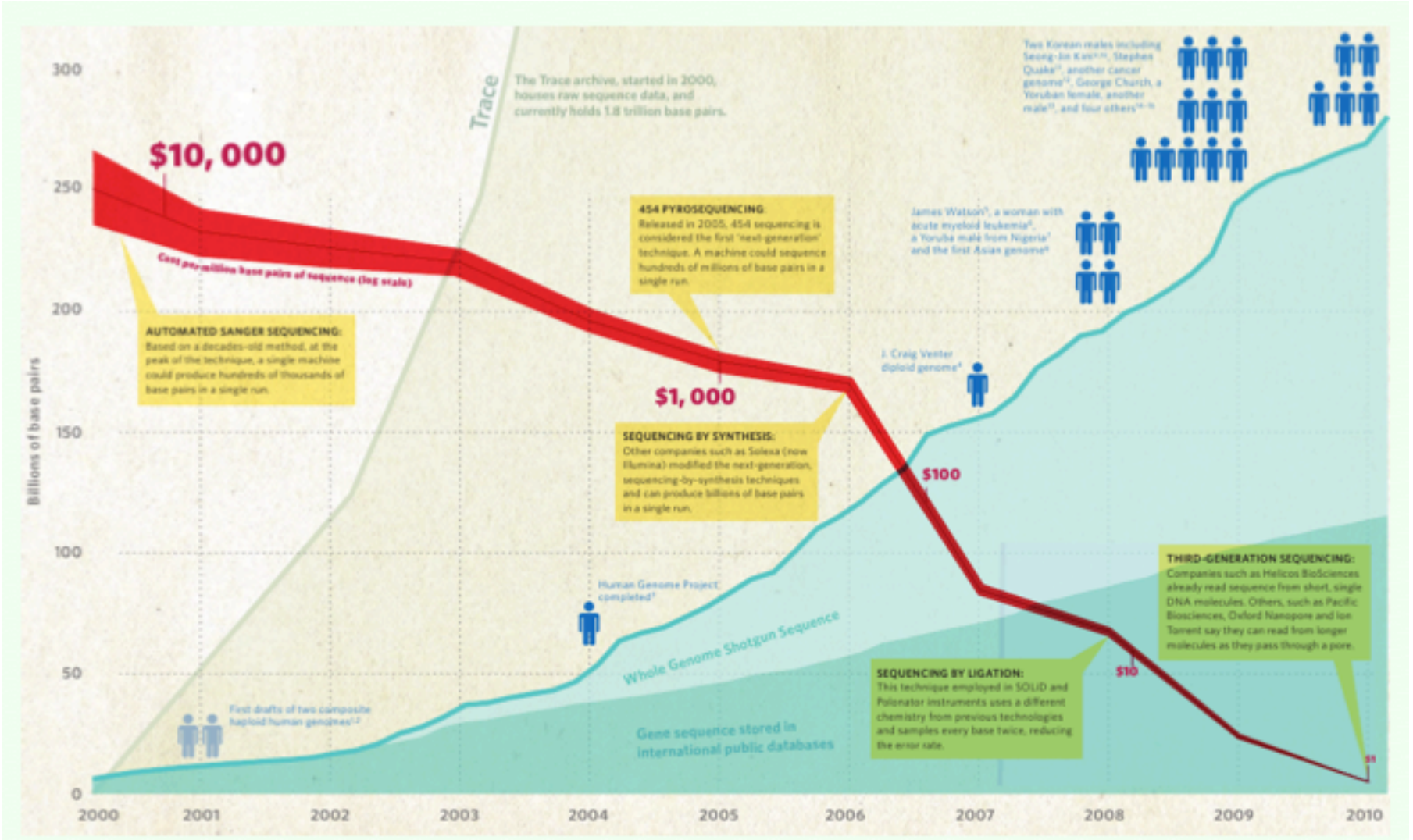


A. Extracted gDNA
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome
D. The whole-genome sequence is derived from the consensus of aligned reads.

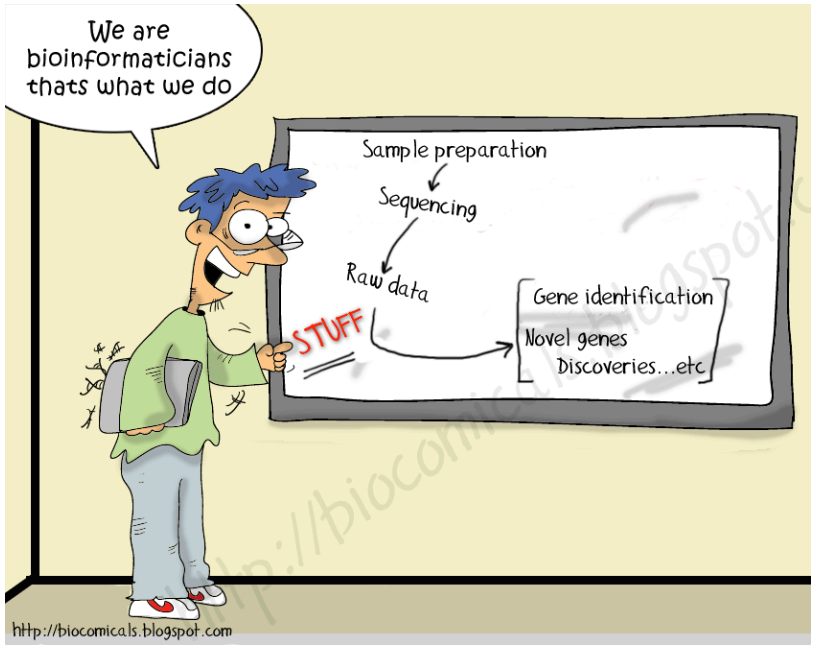
Different platforms

Technology	Read length	Gbp / day	Cost $/Mb
Sanger	**1 kb**	0.006	~ 500
454	**450 bp**	0.5	~ 20
Solexa / Illumina	2 x 100 bp	**25**	**~ 0.5**
SOLiD	2 x 50 bp	**10**	**~ 0.5**
…			
PacBio	10 kb		

Sequencing cost



New costs

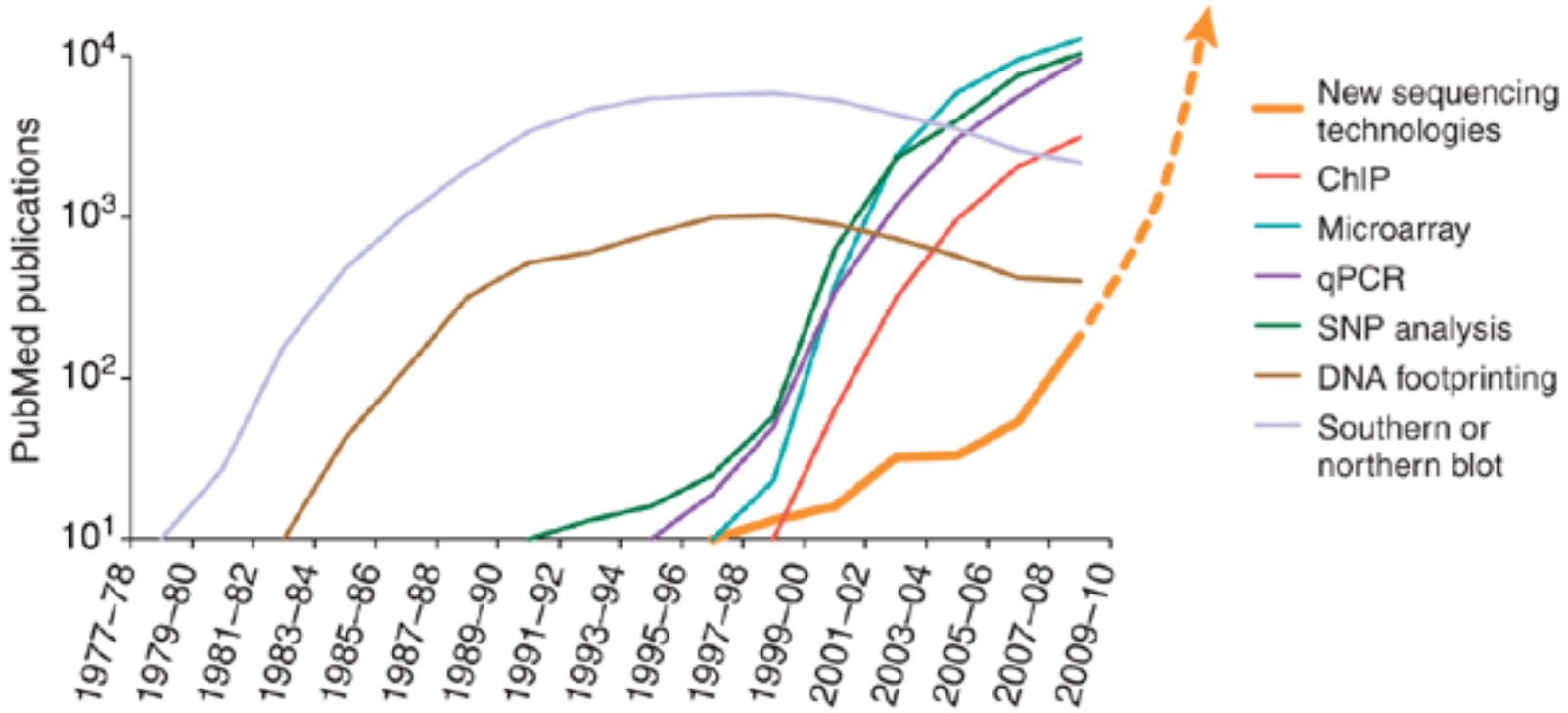

We are bioinformaticians thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc
http://biocomicals.blogspot.com

New data and new files

```
>@HWI-ST450R:198:B00R0ACXX:3:1101:1102:2231/1
NAGTTAGCAAATCGGGTGGCCTTATTTTCAACCTGGACAACCATGTACCCCGCATCGAGCACGAGAAGACATTCACATATCCCCTGGACCTGGTGCTGGAG
>@HWI-ST450R:198:B00R0ACXX:3:1101:1139:2236/1
NTTCCGATGGGGCACCGCCTTCTGGGCGGCCAACTCCCGCAGTTCGTTCAGCAGCACATGCATCTGATACTTGGAAGATCGGAAGAGCGGTTCAGCAGGAT
>@HWI-ST450R:198:B00R0ACXX:3:1101:1190:2238/1
NTCGGCAGCTGGCTTGAACACCGCTTCAACAACGTGGCTCTCGGCAGATTCCTGATGAGCTGCGTCGTTGAGGTGGAGATCAAGCAAGCGGAACAGAAGCC
>@HWI-ST450R:198:B00R0ACXX:3:1101:1421:2224/1
NTTGGATTGGATTGGATTGGGAAGTAGAGGAAGAGTCACCAAAAATAAACGGCGAAAATGTGGCCCAACTTTTTTGAGATCGGAAGAGCGGTTCAGCAGGA
>@HWI-ST450R:198:B00R0ACXX:3:1101:1257:2227/1
NATTTTTGCGGGTAATTTATTTGCGTTTTCAGAACAGAATTTGCGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTT
```

```
@FCC0WM1ACXX:8:1101:1721:2192#TAGGAATA/1
AGGATGGTGGAAGCGTGAAGCCCGCACCACCAGTTGCAGGCAGCGACGACGAGCGGGCGCAGAAGATACTCGAAGATACTACGGTGAGAGTGGGCCAACG
+
_abeceecgegggihhadgebffhihiiiifigeefhi`ghhhiiifeeccccccc_aaacccccccbcbbc^acc_bbc_bbacb`ccccS]^a^aacca
@FCC0WM1ACXX:8:1101:1922:2135#TAGGAATA/1
TGGGCAATATGCAAAAATGCCAGCGGACCCAACTCAACTCCTTCTCTTTTTTCGATTCCCTTTCCCTTTTGACCATTTCAGTCCCAAAATCCAAATCCCT
+
___ccccdgg]acfhffhhdgfffageghidaf`ffdfhhiihffffhfh_gfYP\R^baccYZ]]_U_bbab]``cbccb]bGKTR[[_]`bcccb]`b
@FCC0WM1ACXX:8:1101:1985:2180#TAGGAATA/1
CGATTTTGTGTAGAAATAAACTAAATATAAGGTCGAATTAATTAAATGAGTTTGGTGCAAAAGTGTTTTTTGTTAAAATGGTGTCGGTTTTTGGTCGATT
+
^[_eeeeega^ecgfhfhhiiiiiiiihiiiibgfhdfhhiihhhhhhhiiafgihaegbfhigbb_bfgigggeeeceebdZZ`bac^aca_caW`a_ac
@FCC0WM1ACXX:8:1101:1867:2225#TAGGAATA/1
AGAGCTAGAGCAACCAAATTTGGTATCCACACATTTACCGTGCGAGCAGCCATCAACTCCCGCGACAGCCTGGACCTAACCTTGACATAAACATTCGAAA
+
_a_eeedcggfggiihhiifhicbbaeghhhiifhiiiiicbgfhihhhfhhhhhfg_gfg]ga`_aZ^`bXXX`b^bcccbcccccbcb_bbb`bcdcc
```

Usage of NGS



Avak Kahvejian, John Quackenbush & John F Thompson
Nature Biotechnology 26, 1125 - 1133 (2008)

Applications

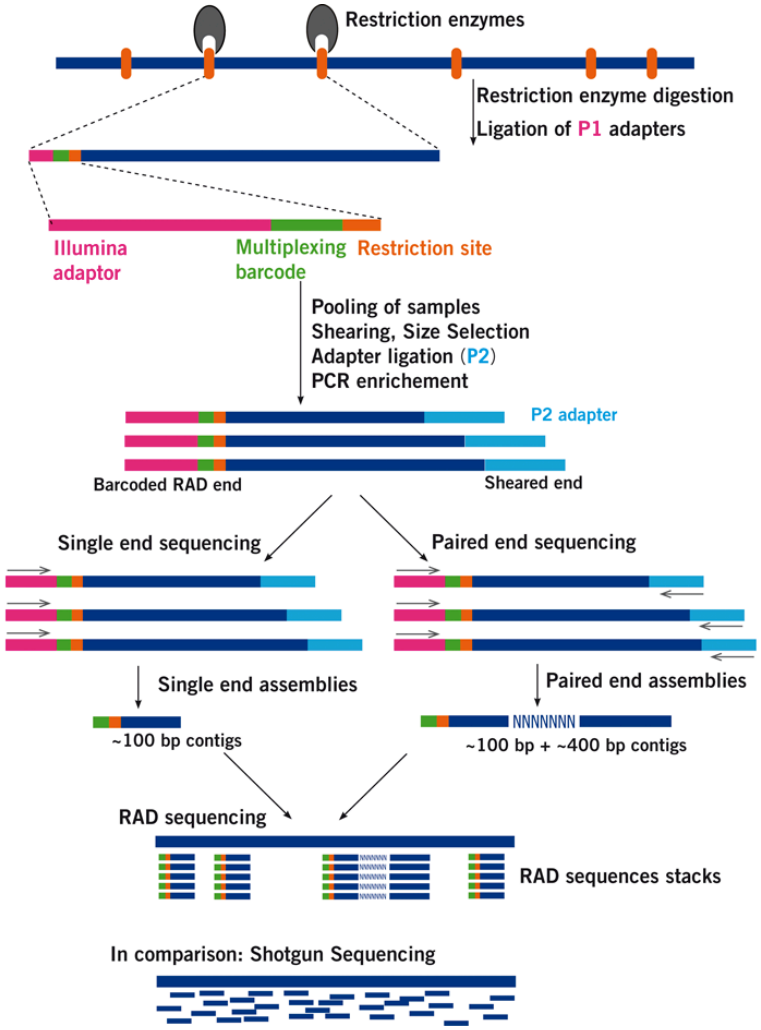
whole genome

ancient genomes

Exome capturing

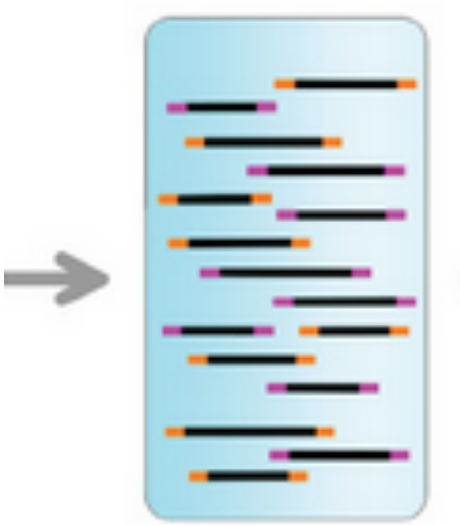
	FS5524895	FS5102038	FS5103086	FS5522194	Any 3 of 4 FS5524895 FS5102038 FS5103086 FS5522194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,766
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (MYH3)	22
...And predicted to be damaging	160	10	2	1 (MYH3)	3

RAD-sequencing



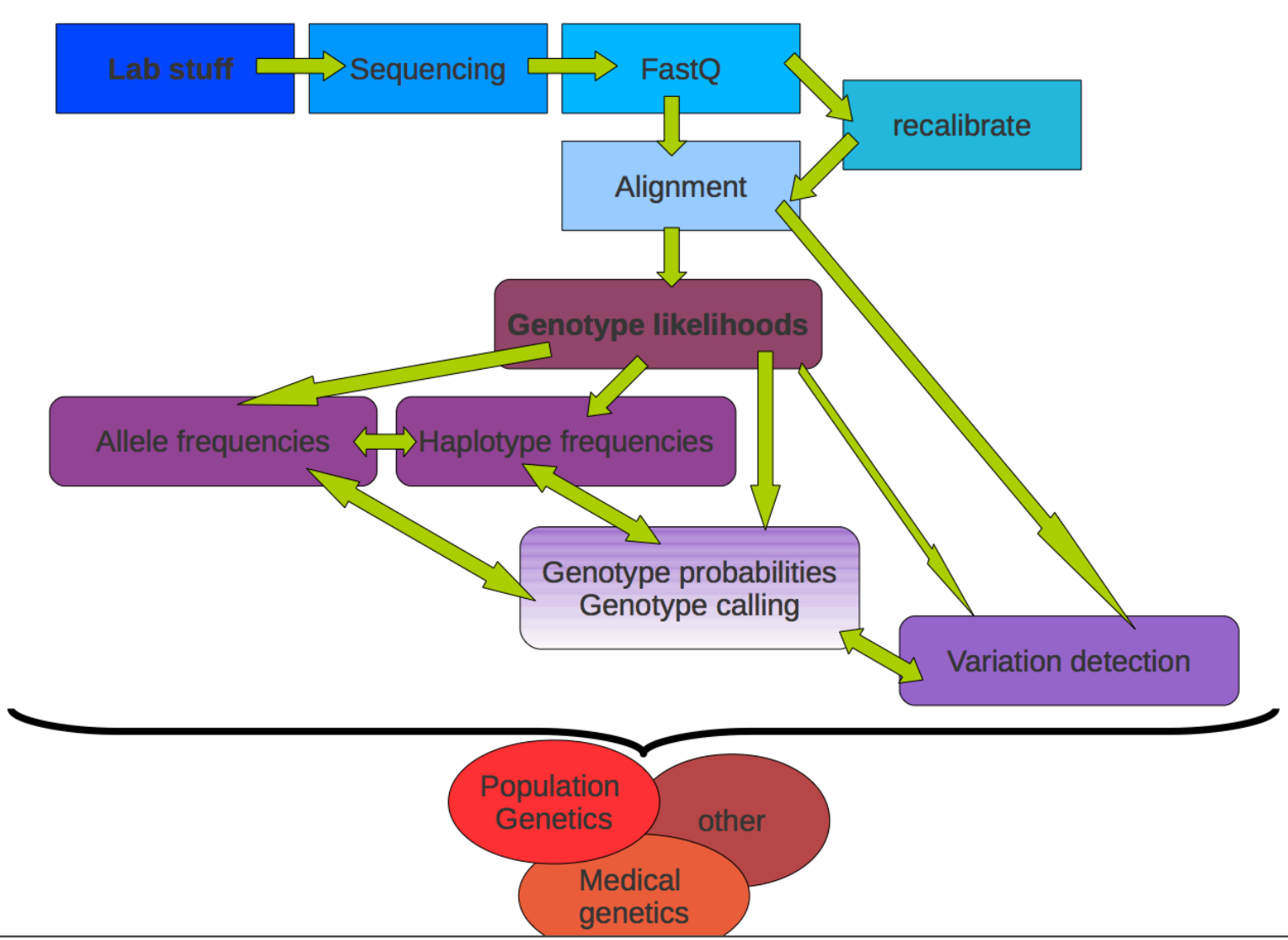
http://www.floragenex.com/

Pooled sequencing

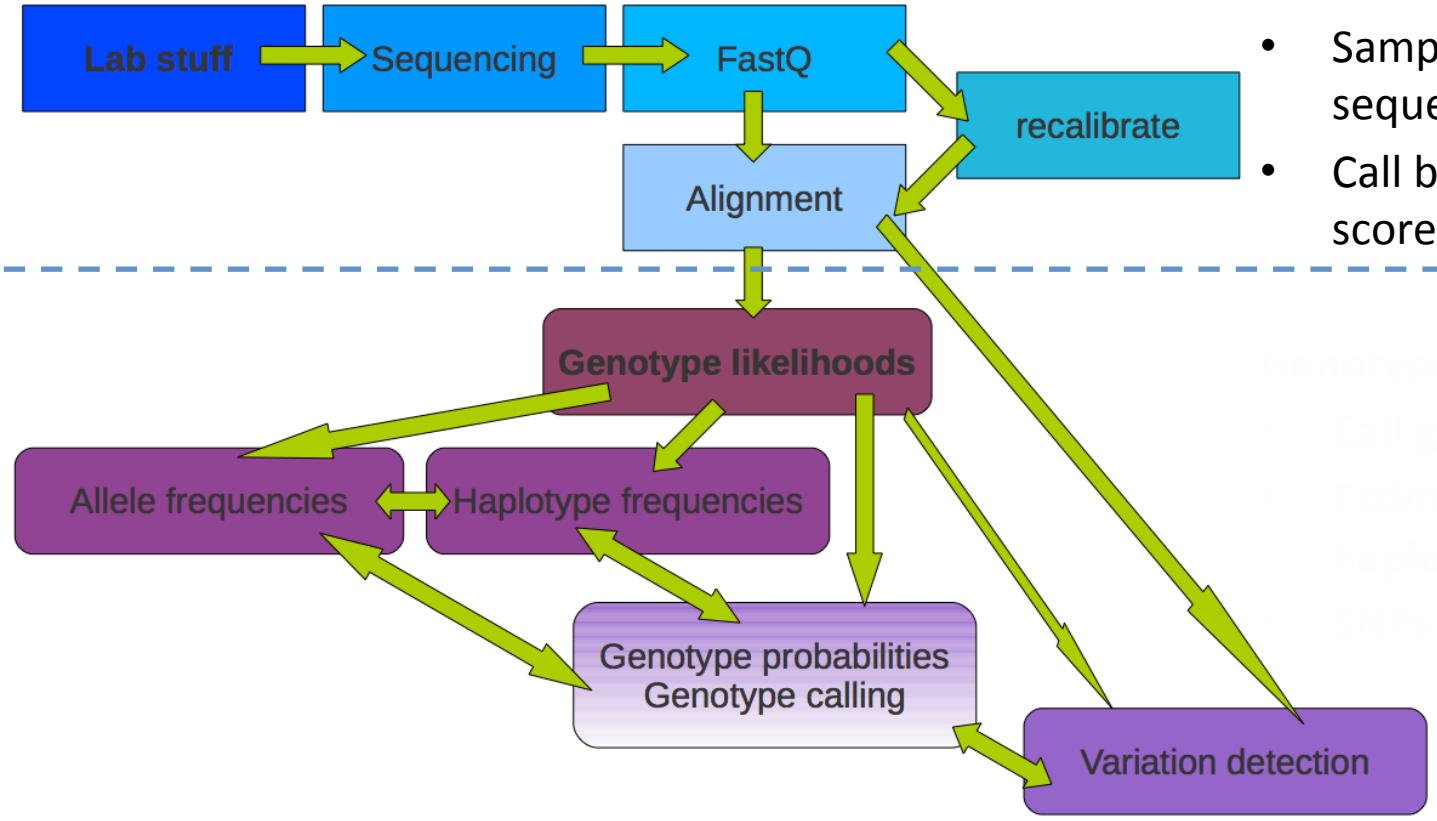


...

Workflow

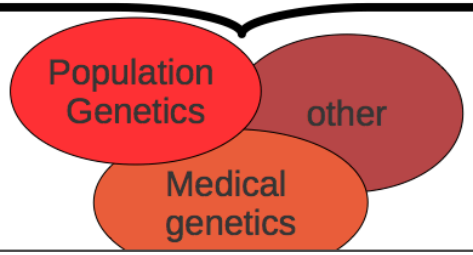

Lab stuff
Sequencing
FastQ
recalibrate
Alignment
Genotype likelihoods
Allele frequencies
Haplotype frequencies
Genotype probabilities
Genotype calling
Variation detection
Population Genetics
other
Medical genetics

Workflow

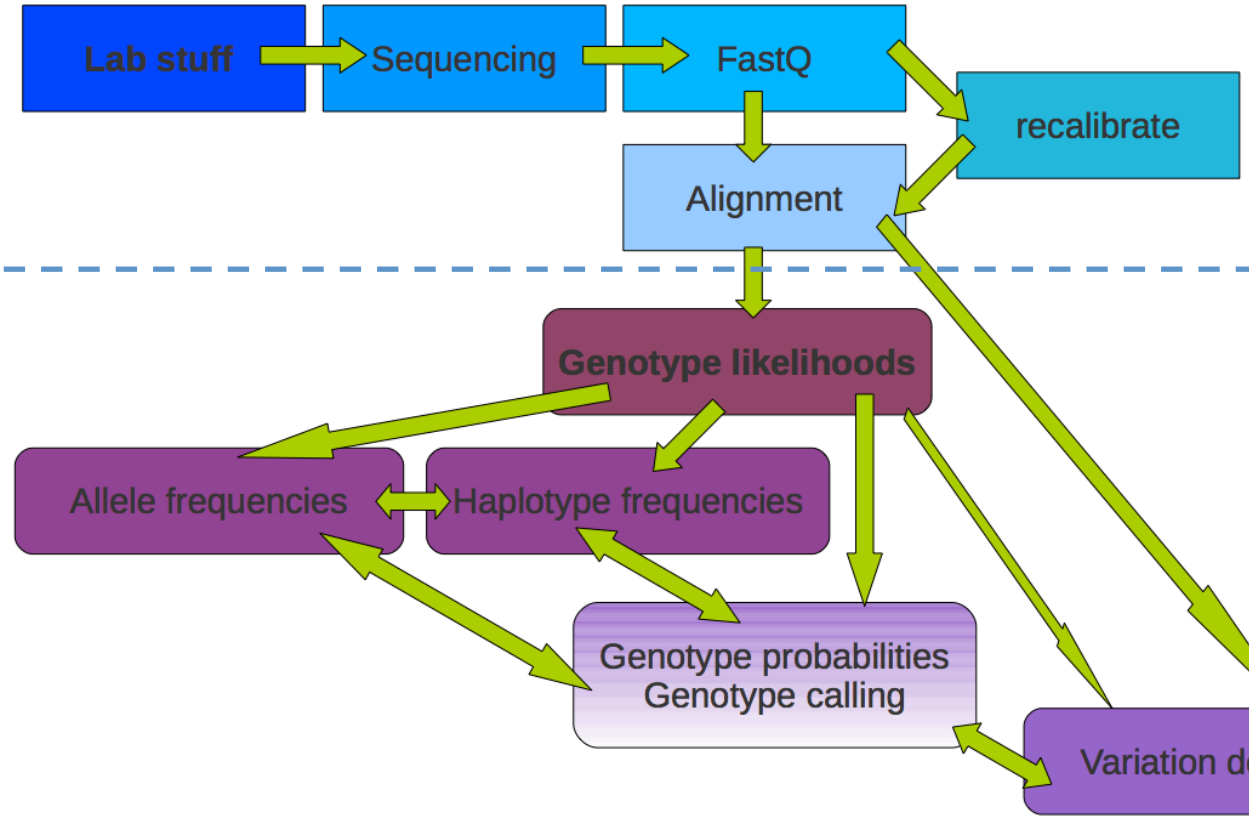


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores



Workflow

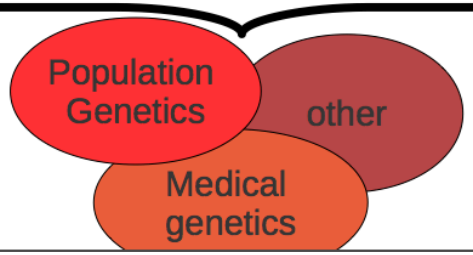


Low-level data:

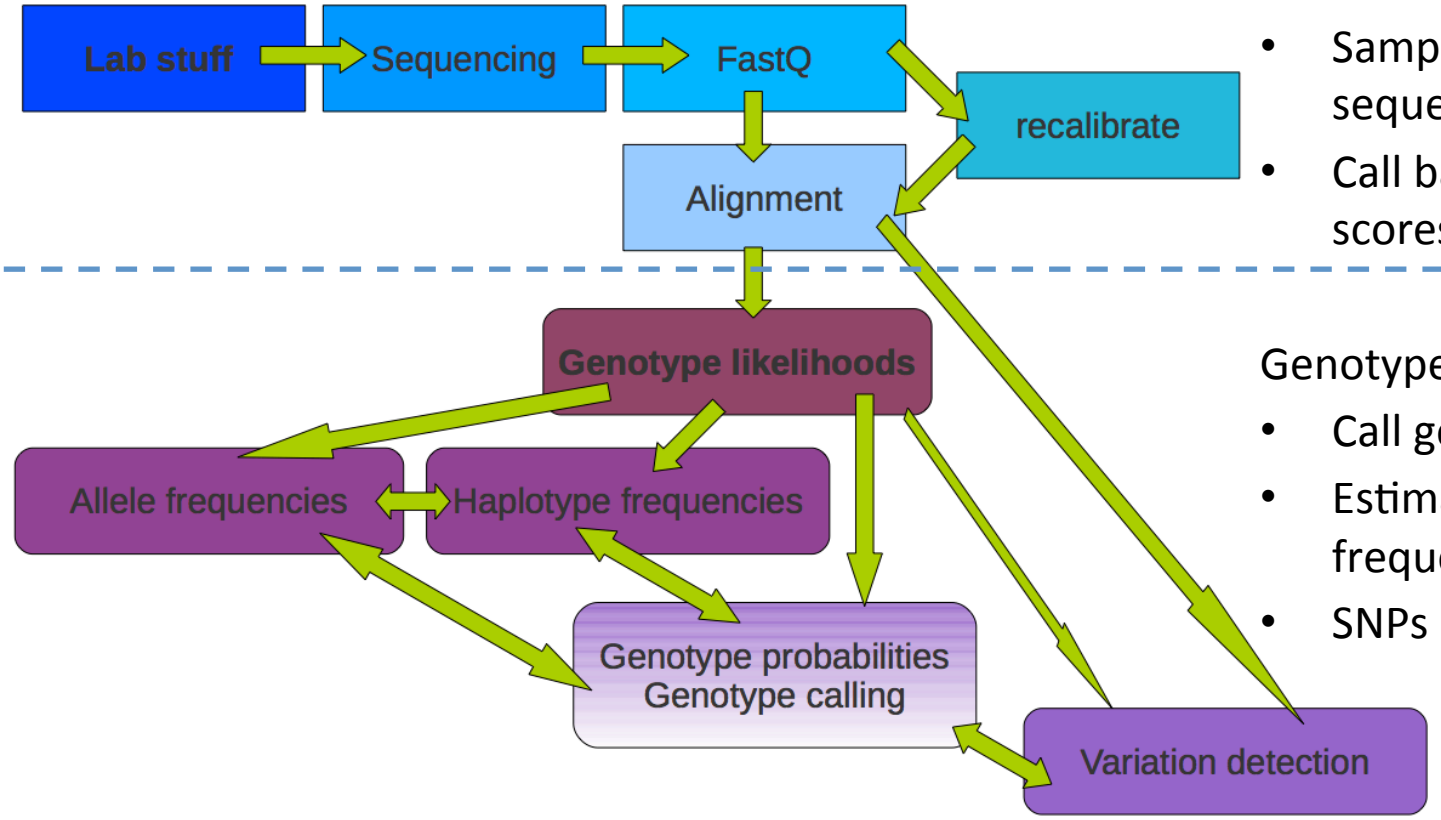
- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele/haplotype frequencies
- SNPs detection



Workflow

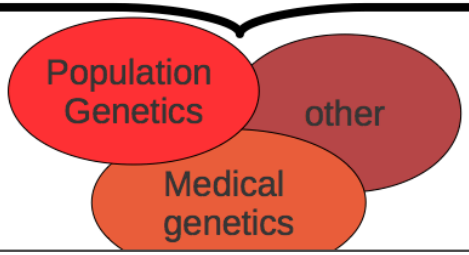


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

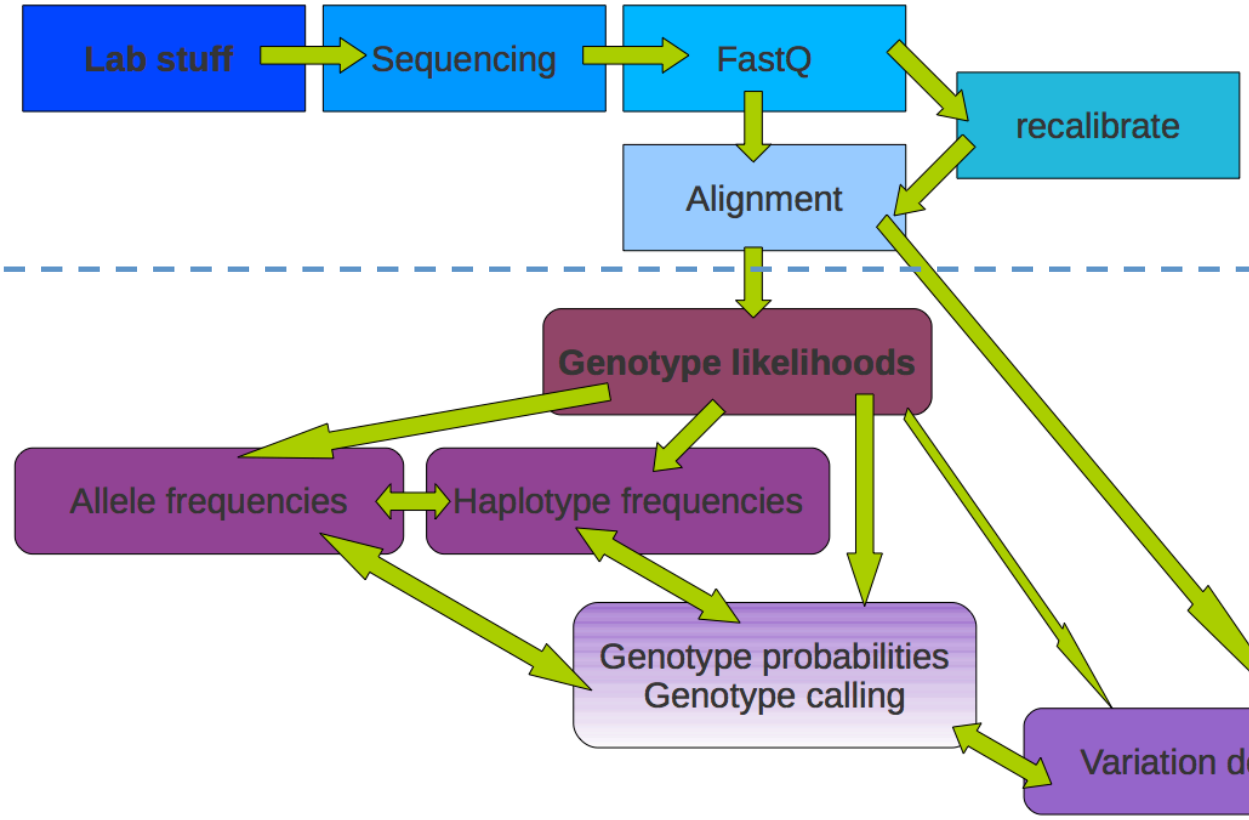
- Call genotypes
- Estimate allele/haplotype frequencies
- SNPs detection



Analysis:

- Population genetics analysis
- Association studies

Workflow

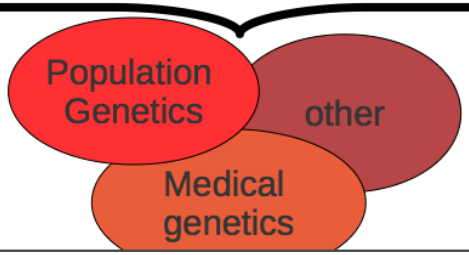


Low-level data:

- **Samples preparation + sequencing**
- **Call bases and quality scores**

Genotype data:

- Call genotypes
- Estimate allele/haplotype frequencies
- SNPs detection



Analysis:

- Population genetics analysis
- Association studies

Low-level data

FASTQ

```
a‘X_\Va\J‘KaYJHG^]b\a^BBBBBBBBBBBB  <-- quality score
@FC42BF1AAXX:6:1:5:732#0/1          <-- read ID
TGATTCTCTCGATATCCAGTCCTTAGTGNCATAGN  <-- read (bases)
+
a^_aaaa‘aa‘_aaa_aaa‘__‘‘_‘VBBBBBBBB
@FC42BF1AAXX:6:1:5:492#0/1
AACAGTGGGAGGCTGCAGCAGGAGGATTNCTGAAN
+
ababb_abbbZbabaab^‘aaTaabbaBBBBBBBB
@FC42BF1AAXX:6:1:5:480#0/1
ACCTCCTCAGAGTTCTCGAGCTCGAGAANTCTGGN
```

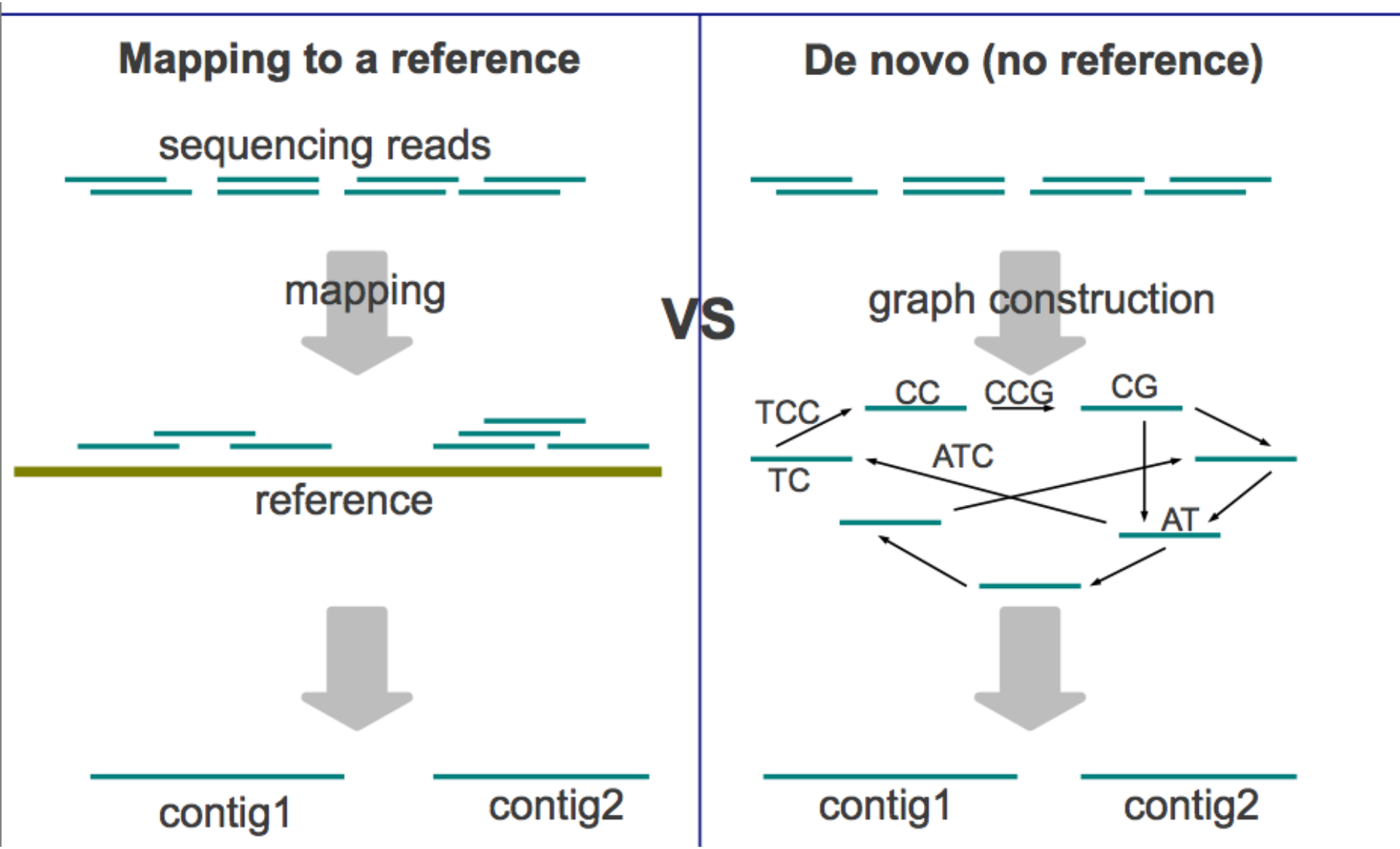
Quality scores

Qscore

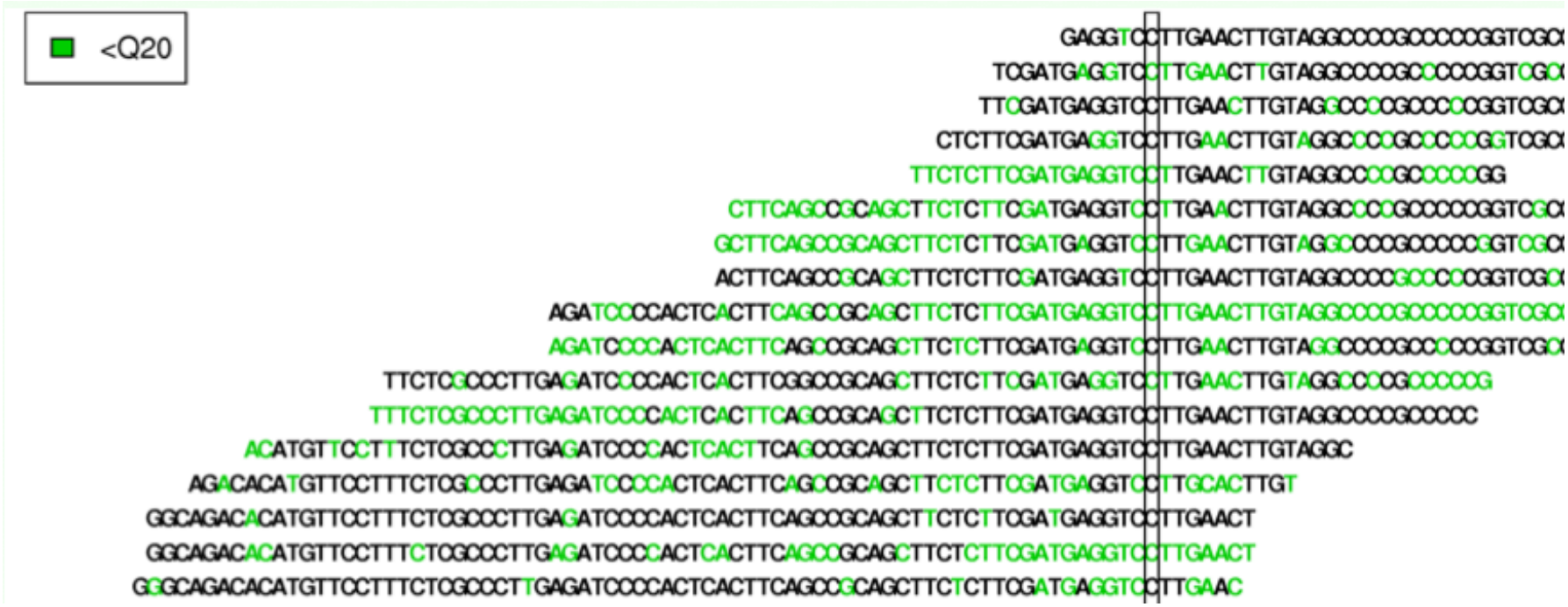
- The ASCII values can be interpreted as a probability
- A Q20 (ASCII 'T') score is probability of 1%
- The score is the probability, P, that the base is incorrect
- $$Q_{score} = -10log_{10}(P)$$
- $$P = 10^{\frac{-Q}{10}}$$

!"#$%&'()*+,-./0123456789:;⇔@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Assembly


Mapping to a reference
sequencing reads
mapping
reference
contig1
contig2
VS
De novo (no reference)
graph construction
TCC
CC
CCG
CG
TC
ATC
AT
contig1
contig2

Mapped reads



- **Depth**: number of reads mapped to a position
- **Counts**: number of different alleles mapped to a position
- **Coverage**: fraction of the genome with data

Alignment file

an alignment file includes

reads	TTTGTTCTTTCTTTCTCTCTAGTCTTCTT ...
Qscore	NVFVN]^]‘^_]^^U]]‘][_VS[_^Z]_ ...
start position	chr4 53351385
multiple best hits	1
Number of mismatch	2
sequence strand	-
read quality*	V

From genome 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
```

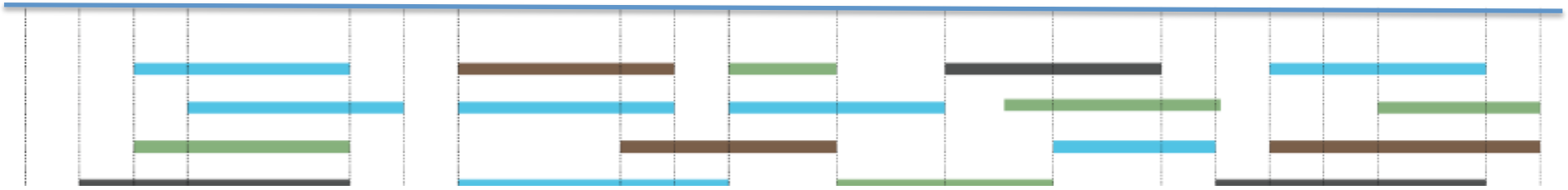
Mapped Reads (mpileup, BAM)

```
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<<:<<
```

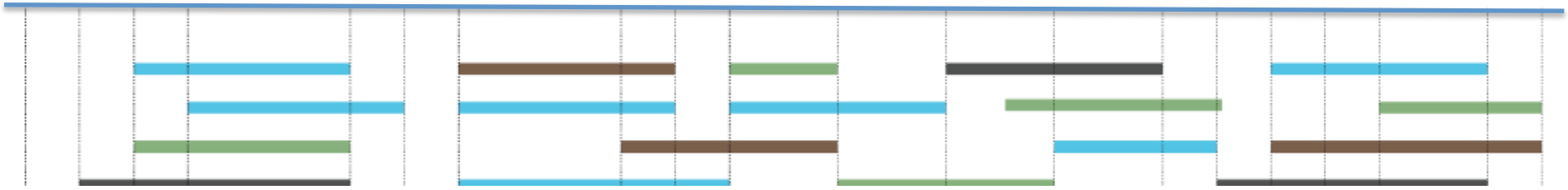
Variants (VCF)

```
##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
```

Challenges

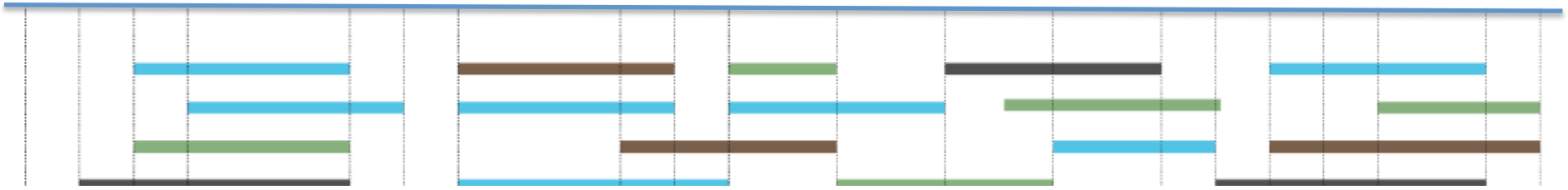


Challenges

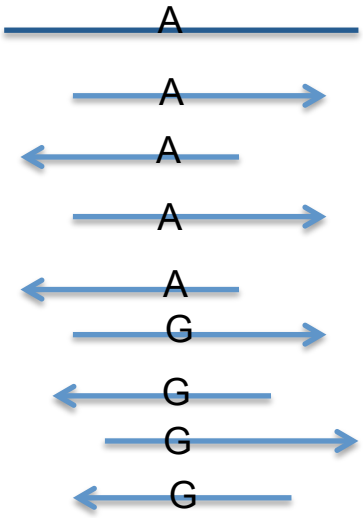


- Variable and low depth
- High sequencing and mapping errors

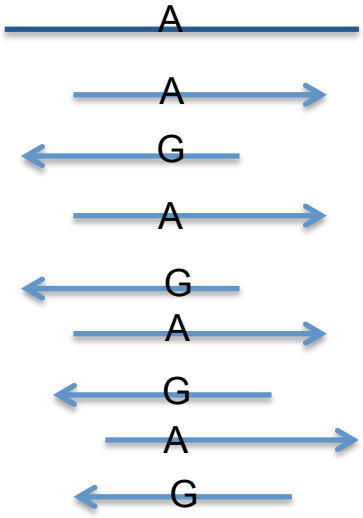
Challenges



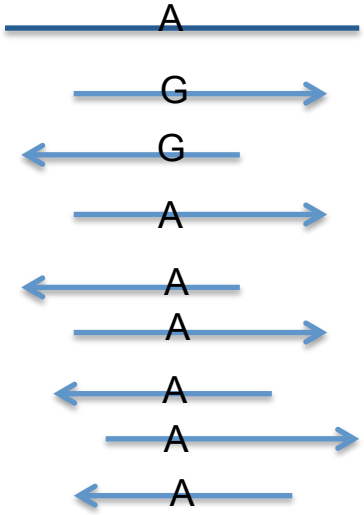
Correct (?)



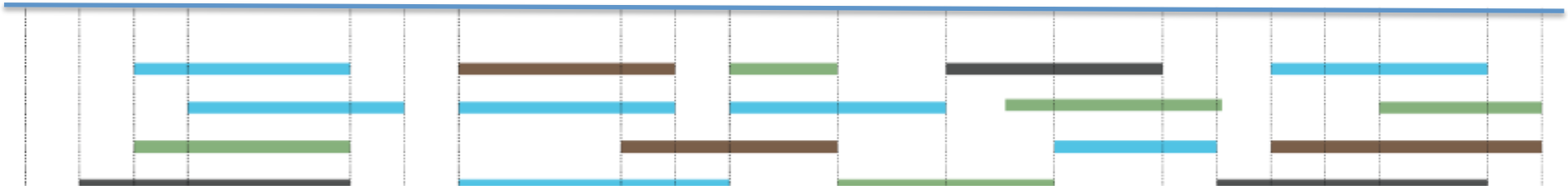
Strand bias



Allelic imbalance



Challenges

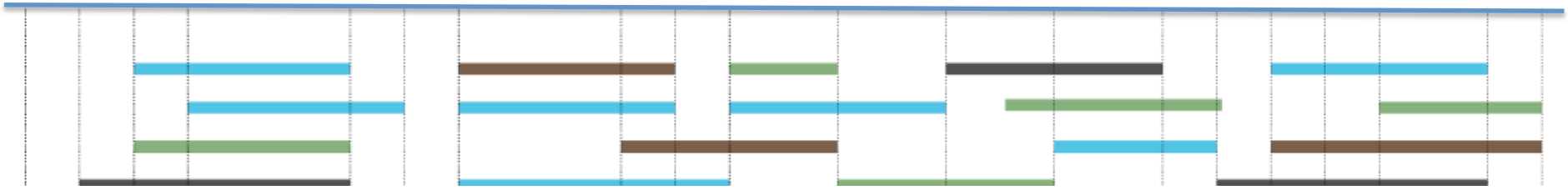


- Variable and low depth

- High sequencing and mapping errors

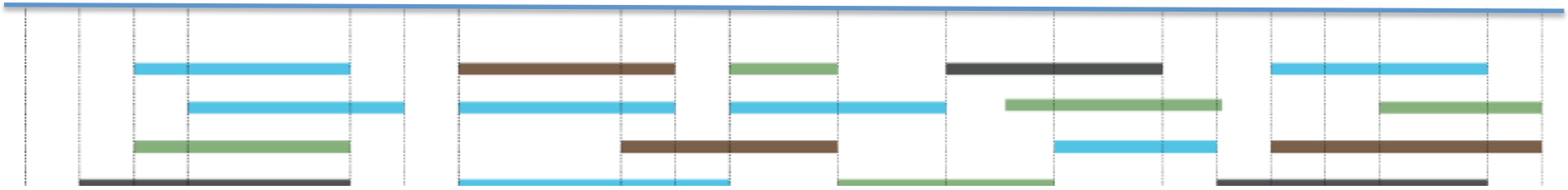
Quality control filters

Data filtering



- Variable and low depth

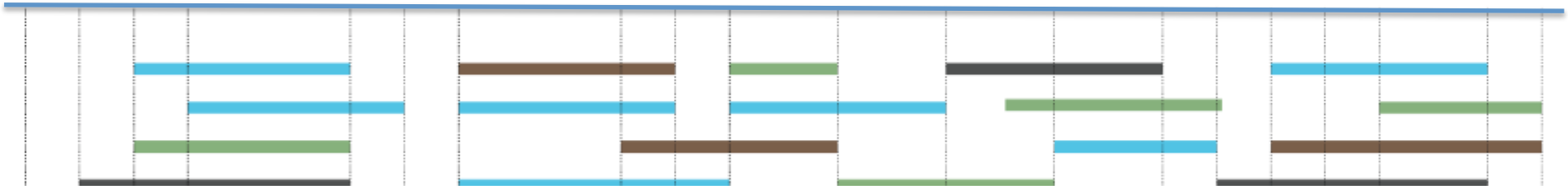
Data filtering



- Variable and low depth

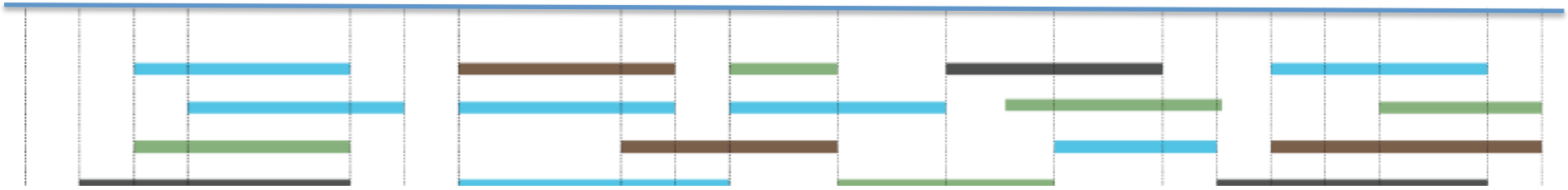
Minimum depth
Maximum depth
Even depth across samples
…

Data filtering



- Sequencing and mapping errors

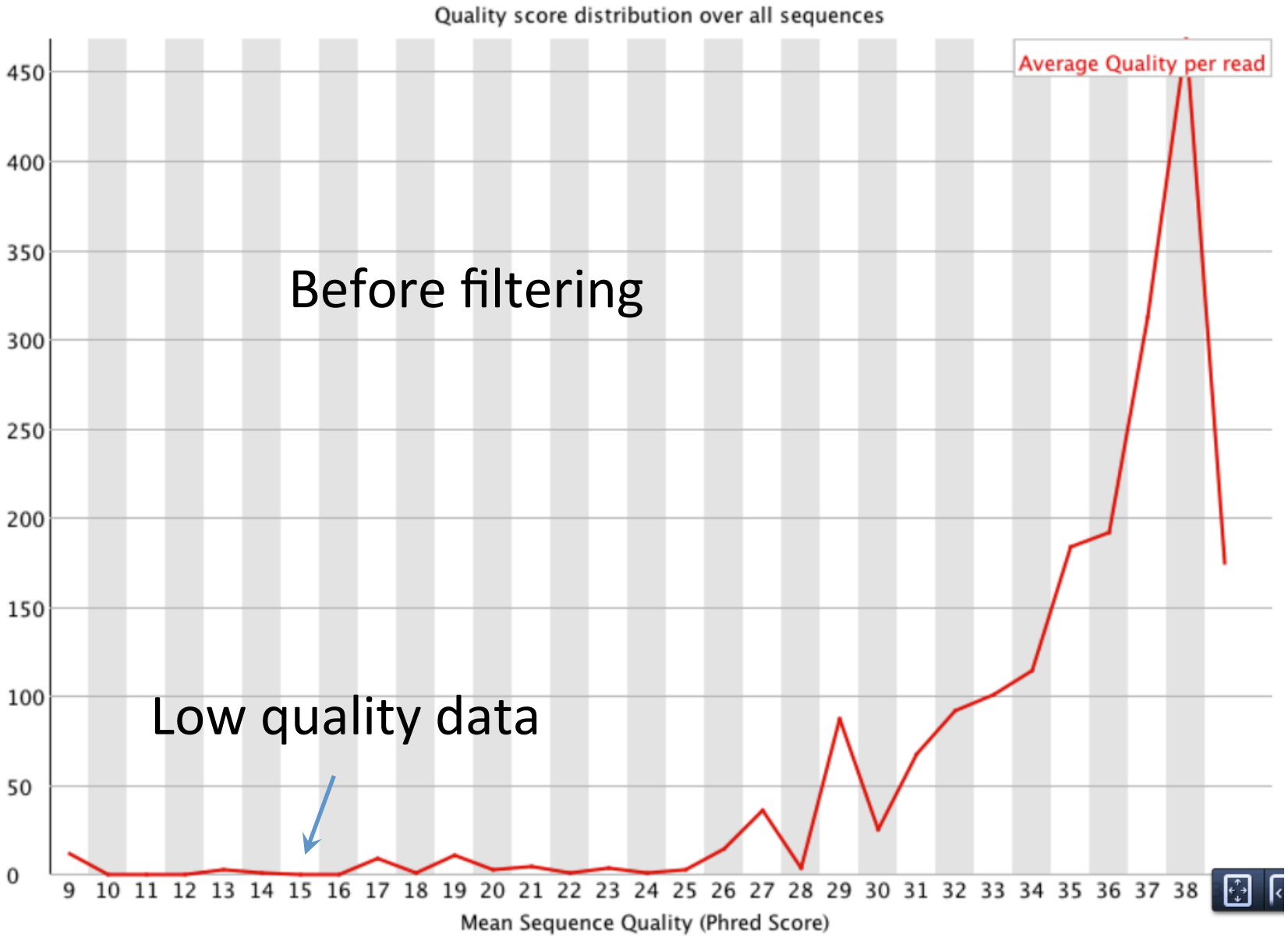
Data filtering



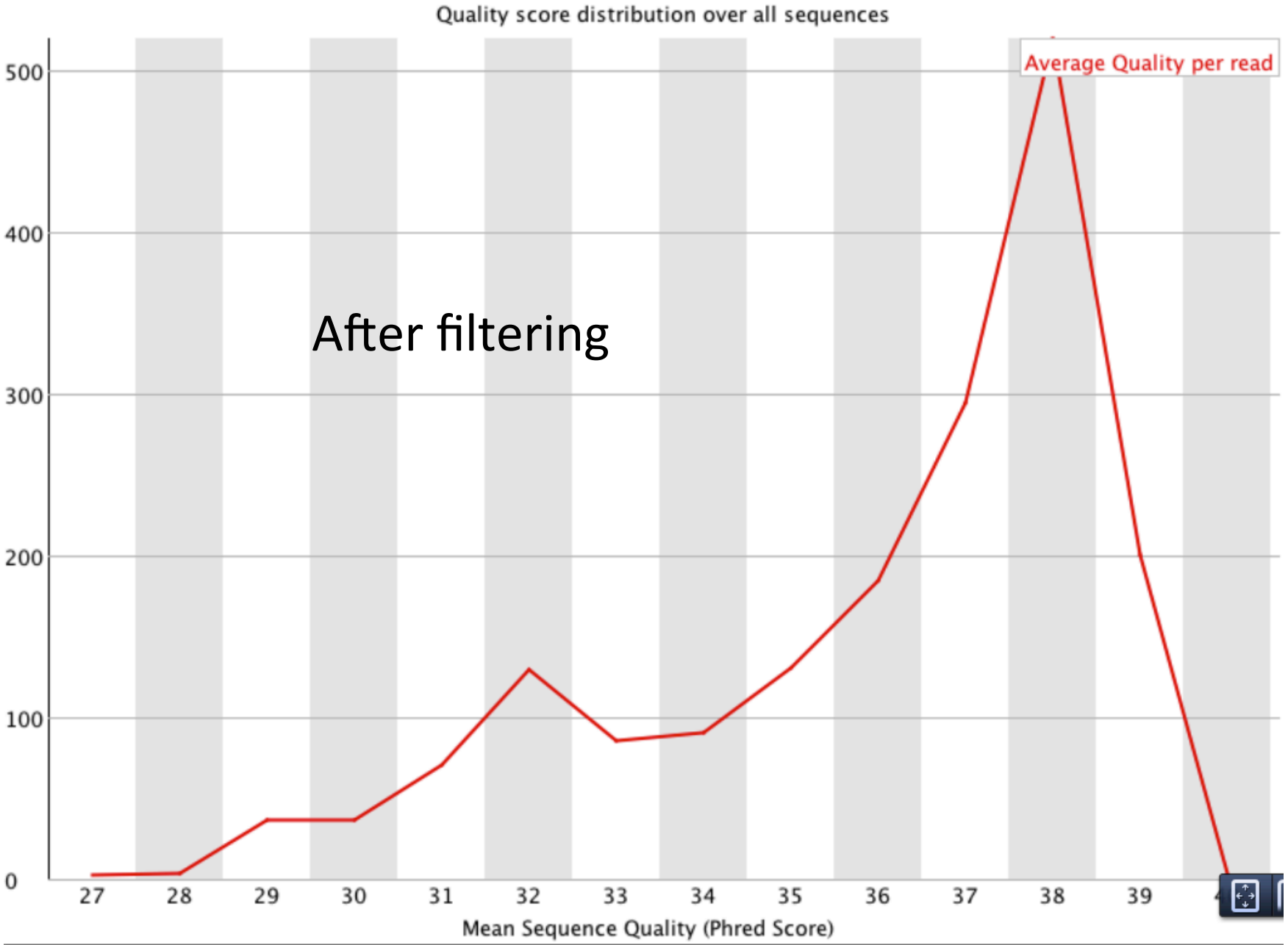
- Sequencing and mapping errors

Minimum base and mapping quality
Base quality bias
Deviation from Hardy-Weinberg Equilibrium (HWE)
…

Check your filtering



Check your filtering



Site Frequency Spectrum (SFS)

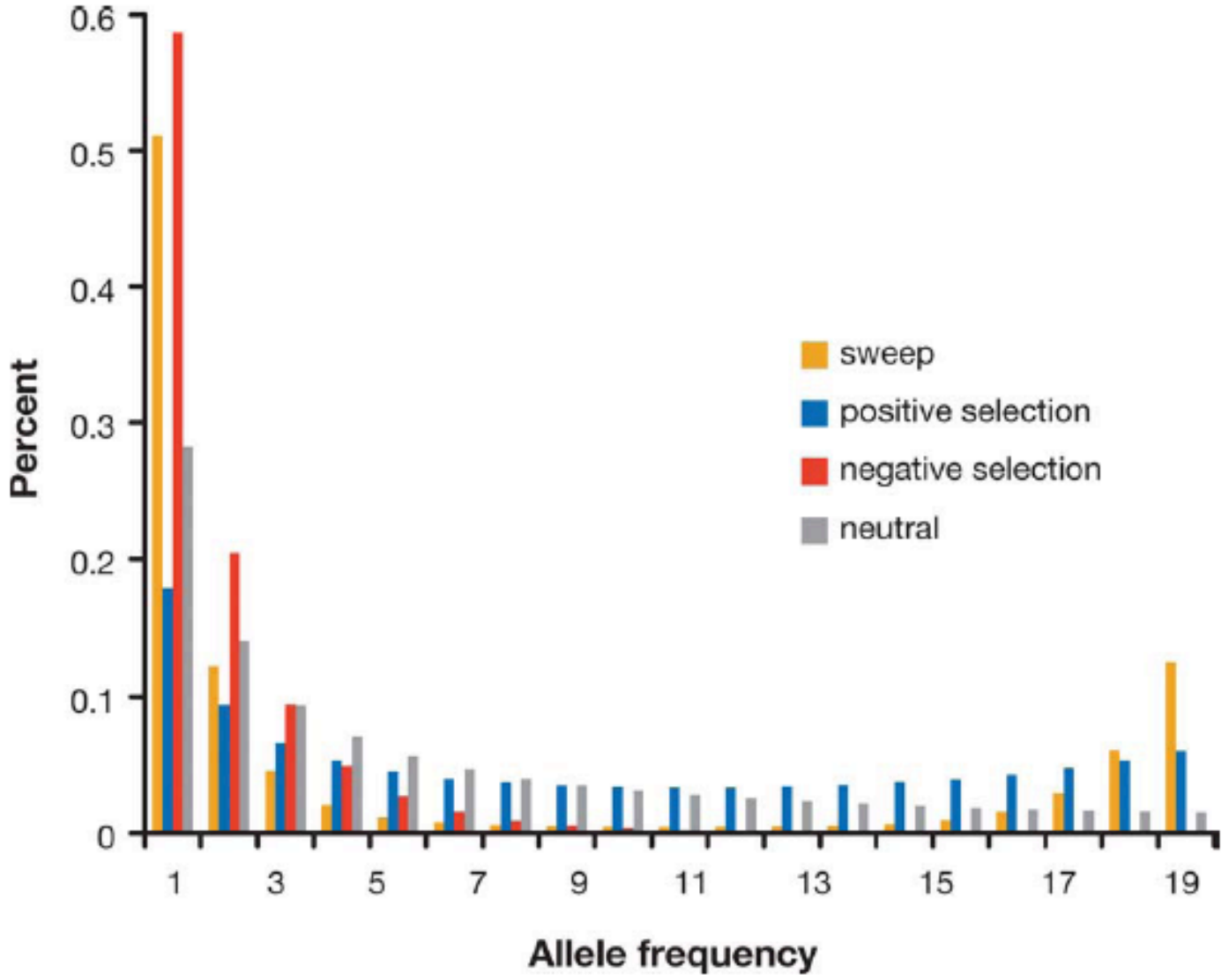
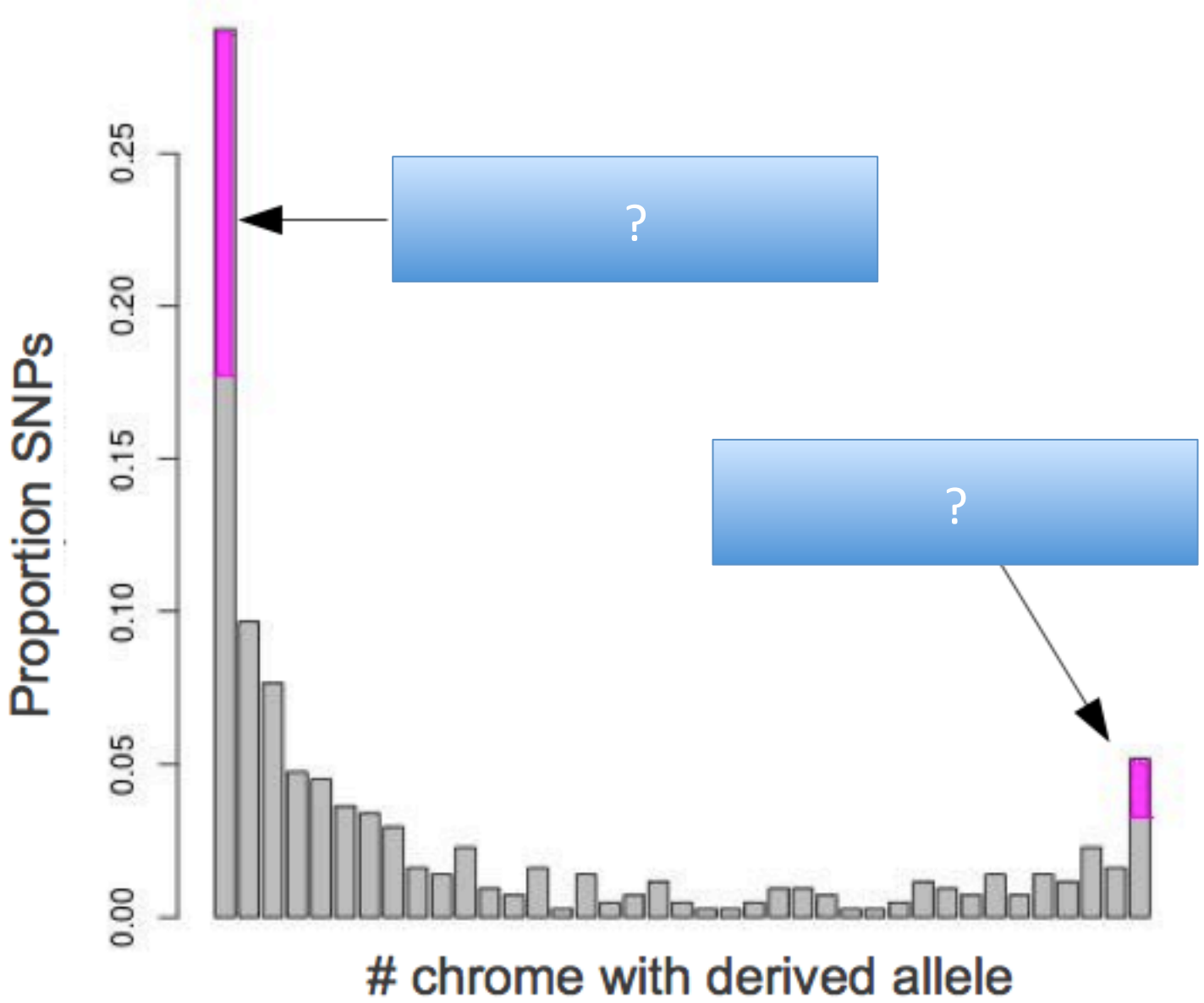
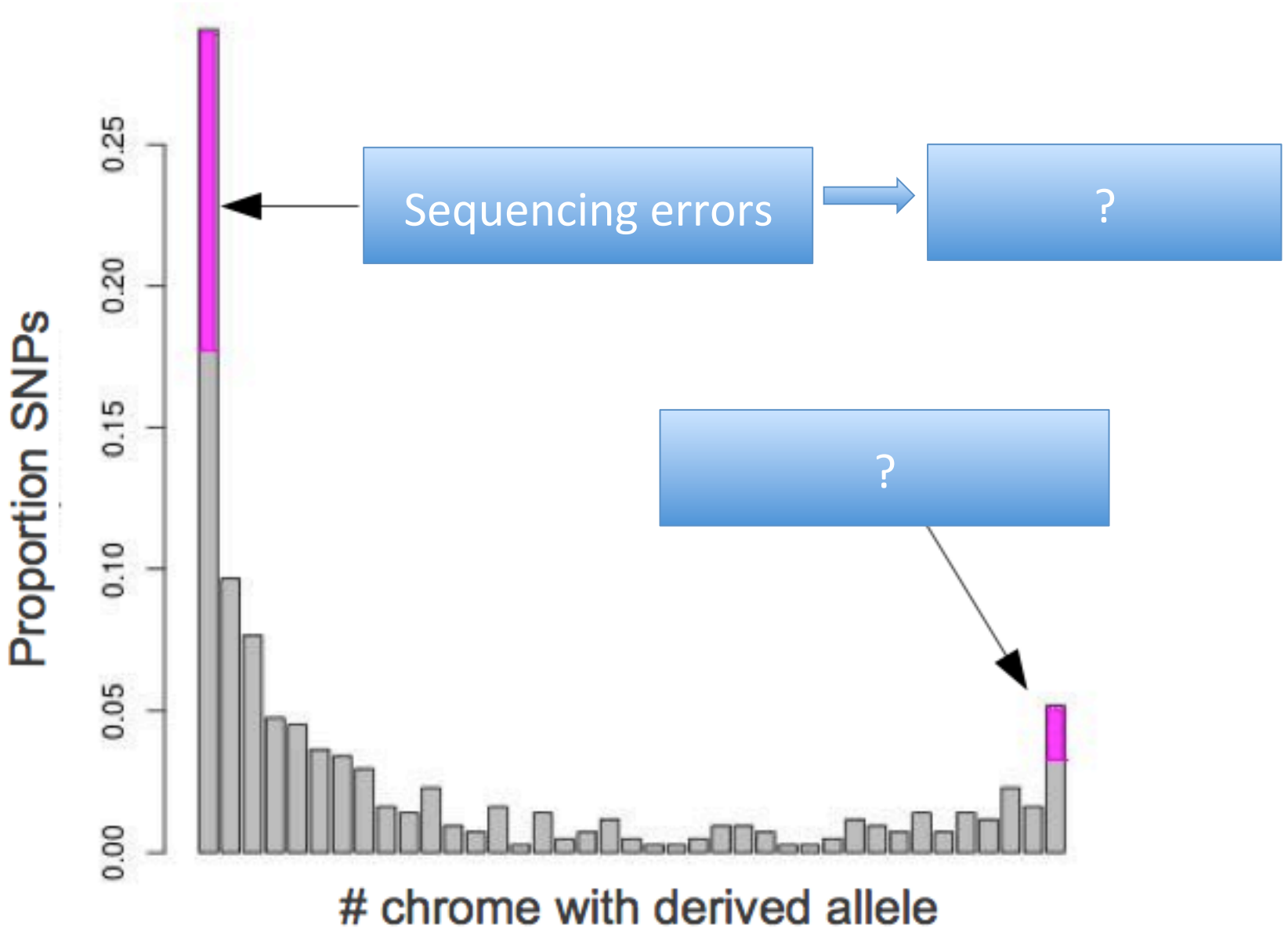


Figure 2

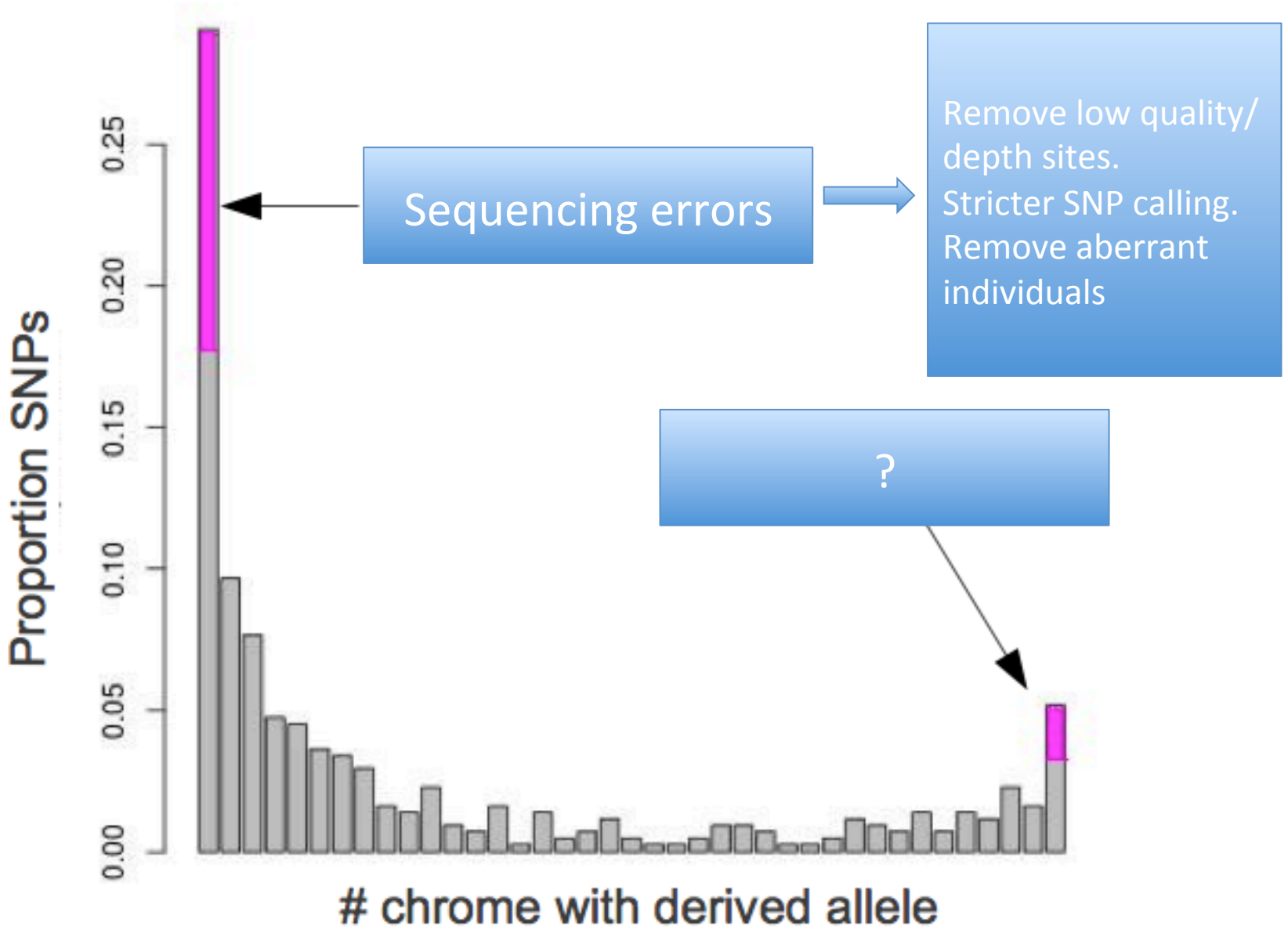
Effect of errors on the SFS



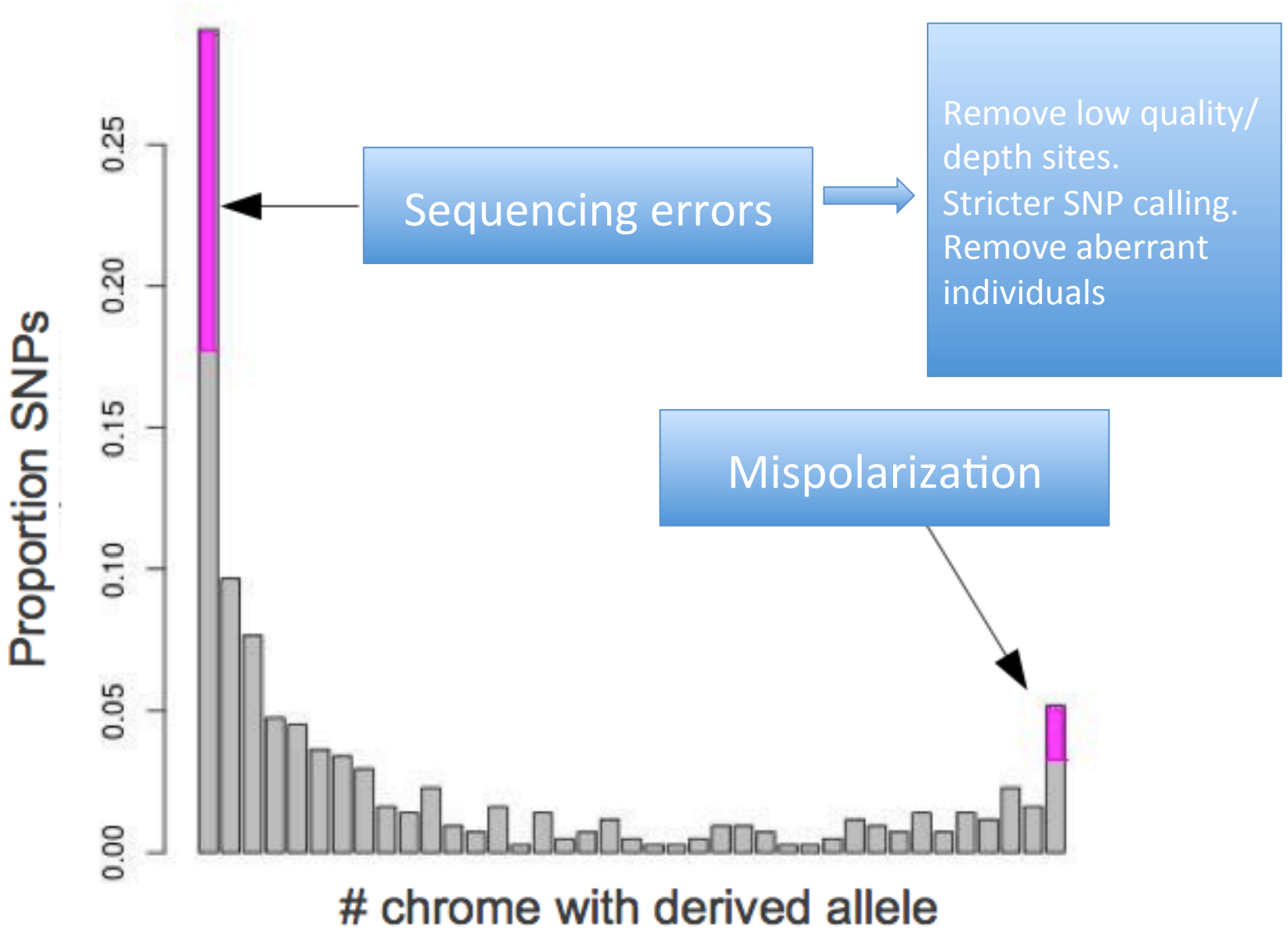
Effect of errors on the SFS



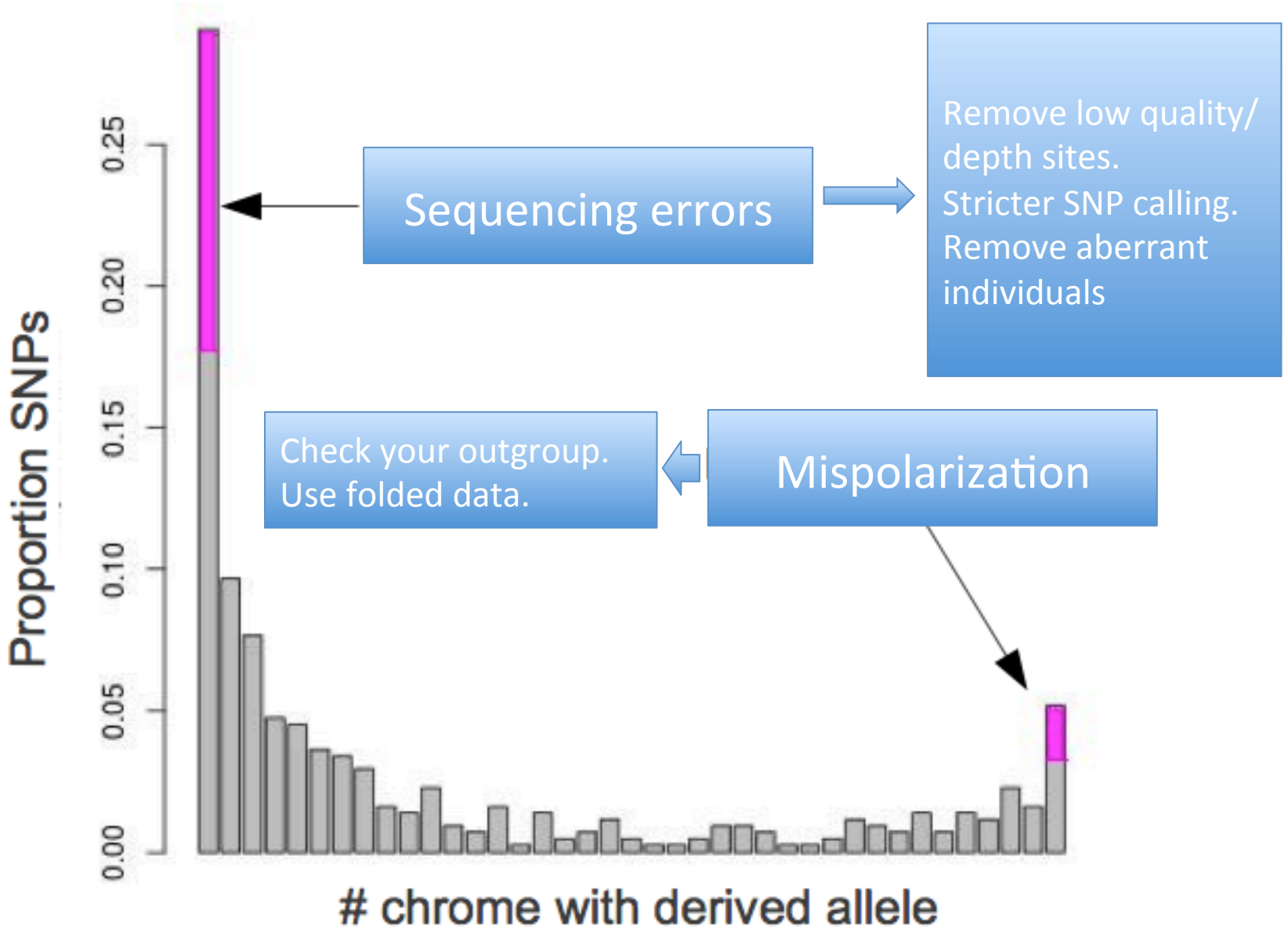
Effect of errors on the SFS



Effect of errors on the SFS



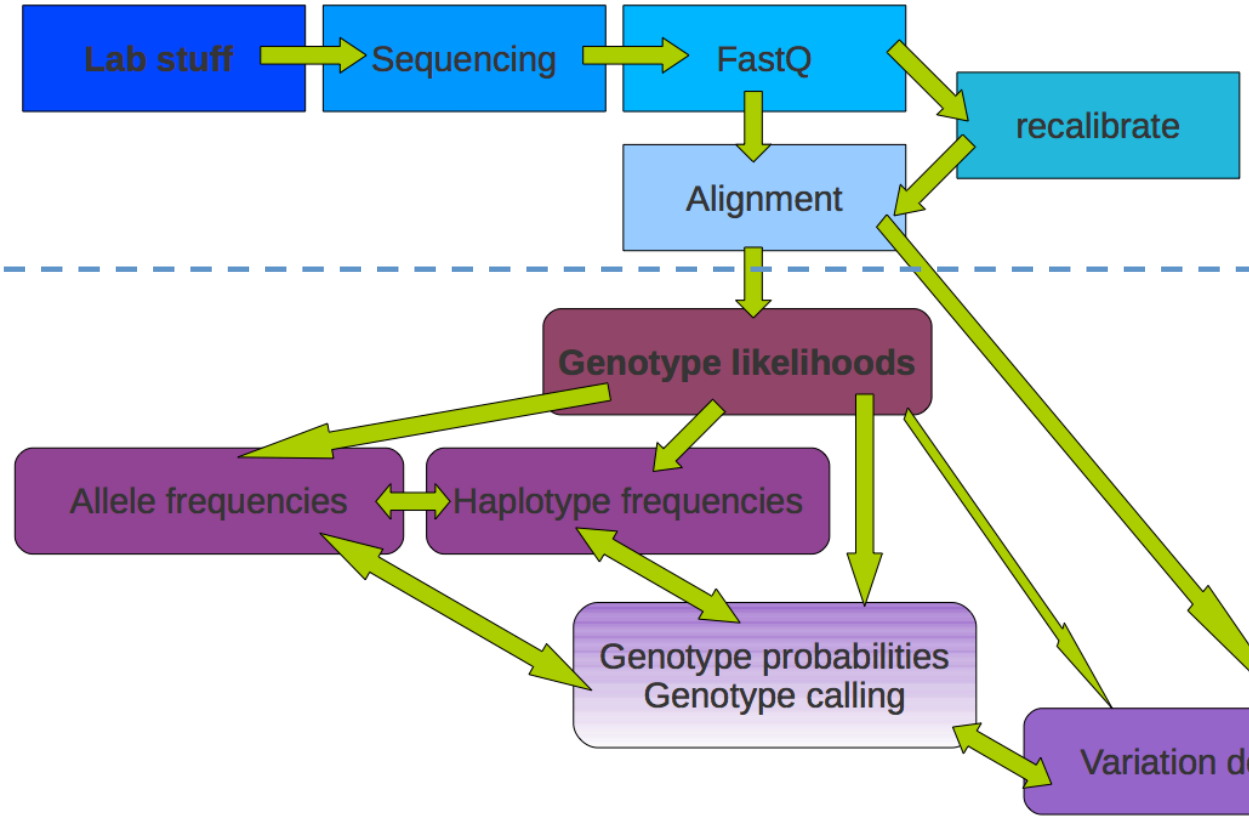
Effect of errors on the SFS



Filtering pipeline

- Dependency on your data and goals
- Check intermediate files and Site Frequency Spectrum
- Tune your parameters by iterating multiple times if necessary

Workflow

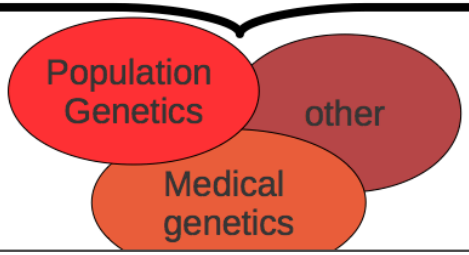


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- **Call genotypes**
- **Estimate allele/ haplotype frequencies**
- **SNPs detection**

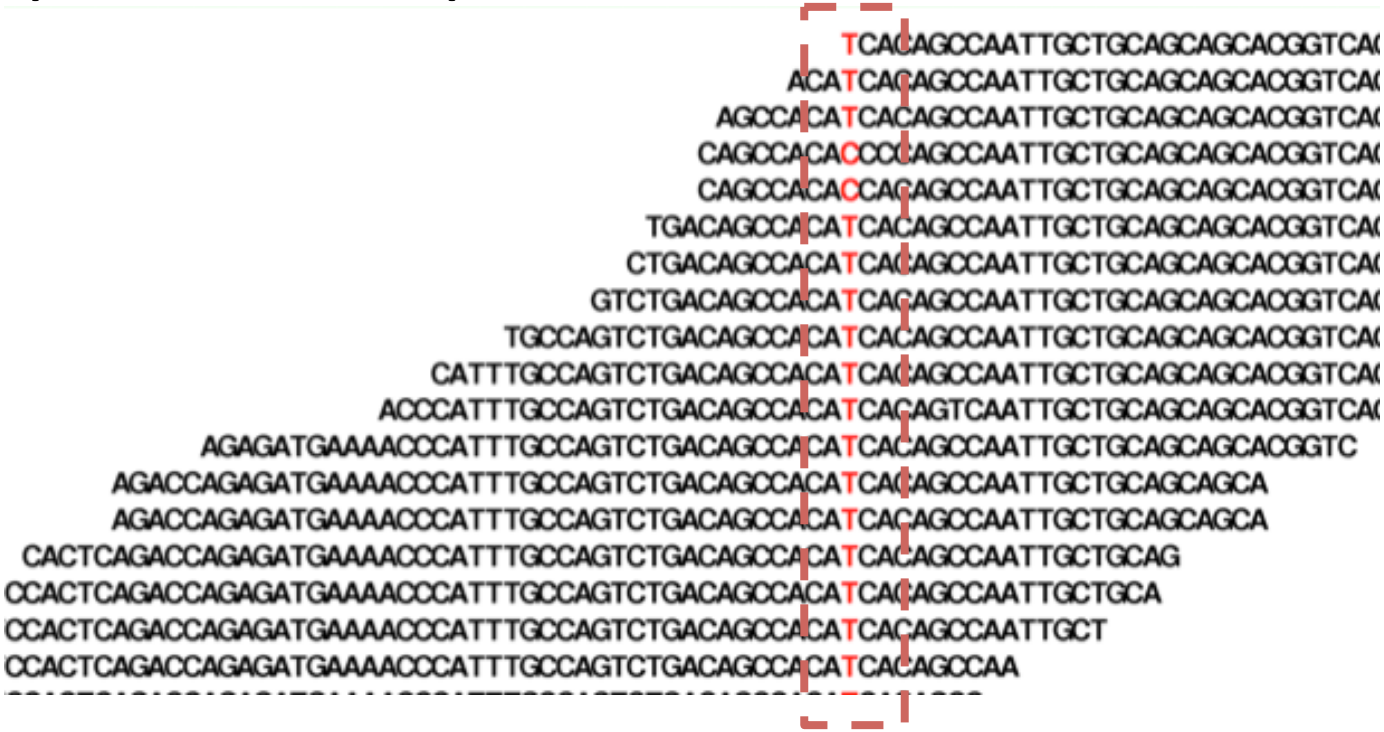


Analysis:

- Population genetics analysis
- Association studies

Genotypes calling

- **Sanger**: both alleles are amplified and sequenced at the same time
- **NGS**: each allele is sequenced separately and sampled with replacement



Likelihood

P(Data | Parameter = Value)

Maximum Likelihood Estimate (MLE):

from a set of observation, identify the value for the parameter (to be estimated) that maximize the likelihood of observing the data.

The integral of the likelihood function is not (always) 1.

Genotype likelihoods

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

How many genotype likelihoods do we have for each individual at each site?

Genotype likelihoods

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

How many genotype likelihoods do we have for each individual at each site?

3 if both alleles are known
10 if not

Genotype likelihoods

- Summarize the reads data in 10 genotype likelihoods:

bases (b):
TTTCCTTTTTTTTTTTTTT
quality scores (P):
BBGHSSBBTTTTTGHRSBB

$\longmapsto$

	A	C	G	T
A	1	2	3	4
C		5	6	7
G			8	9
T				10

Genotype likelihoods

- **SAMtools** (H Li et al., 2008): quality scores, quality dependency
- **soapSNP** (R Li et al., 2009): quality scores, quality dependency
- **GATK** (McKenna et al, 2010): quality scores
- Kim et al. (2011): type specific errors
- …

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342

A T T T	Individual 1
T T	Individual 2
A A T T	Individual 3

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

Iterate through every read for every genotypic configuration...

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

Iterate through every read for every genotypic configuration...

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

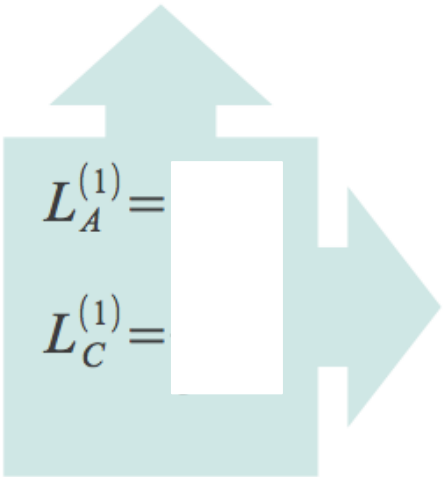
Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$



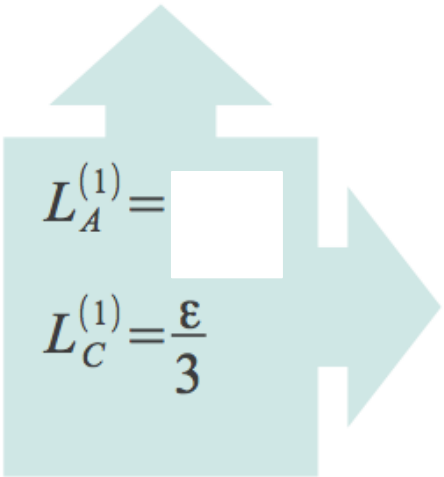
Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$



Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

$$L_A^{(1)}=1-\varepsilon$$

$$L_C^{(1)}=\frac{\varepsilon}{3}$$

$$P(X=A|G=AC)=\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*$$

$$\frac{\varepsilon}{3}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*\left(\frac{L_A^{(3)}}{2}+\frac{L_C^{(3)}}{2}\right)*\left(\frac{L_A^{(4)}}{2}+\frac{L_C^{(4)}}{2}\right)$$

$$=\left(\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6}\right)*\frac{\varepsilon}{3}*\frac{\varepsilon}{3}*\frac{\varepsilon}{3}$$

Genotype likelihoods

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	-1.22
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

$$ATTT$$

$$\varepsilon = 0.01$$

Genotype calling

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	-1.22
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

$$ATTT$$

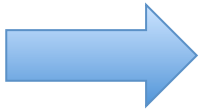
$$\varepsilon = 0.01$$

What is the genotype here?

Genotype calling

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	**-1.22**
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

Simple genotype caller:
Maximum Likelihood



AT

Choose the genotype with the largest likelihood

Genotype calling

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	-1.22
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

Simple genotype caller:
Maximum Likelihood

But **only** call the genotype if the largest likelihood is **much better** than the second best

Genotype calling

- Likelihood Ratio:

$$\log_{10} \frac{L_{G(1)}}{L_{G(2)}} > t$$

$$t = 1$$

The most likely genotype is at least **10 times** more likely than the second most likely one

(in our example t=1.27)

Genotype calling

- Likelihood Ratio:

$$\log_{10}\left(\frac{L_{G(1)}}{L_{G(2)}}\right) > t$$

$$t = 1$$

The most likely genotype is at least **10 times** more likely than the second most likely one

- Higher **confidence** of called genotypes
- More **missing** data

Bayesian inference

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\sum_{\theta} P(X|\theta)P(\theta)}$$

$P(X|\theta)$ ⟵ Likelihood of θ

$P(\theta)$ ⟵ Prior probability distribution of θ

$P(\theta|X)$ ⟵ Posterior probability distribution of θ

Genotype posterior probabilities

$$p(G_s^{(i)} \mid X_s^{(i)}) \propto p(X_s^{(i)} \mid G_s^{(i)}) p(G_s^{(i)})$$

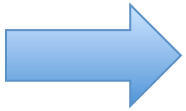
$p(X_s^{(i)} \mid G_s^{(i)})$ → Genotype likelihood

$p(G_s^{(i)})$ → Prior

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	**0.94**
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller:
Bayesian



AT

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	**0.94**
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller: **Bayesian**

But **only** call the genotype if the largest probability is above a threshold (e.g. > 0.95)

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.01	~ 0
AC	-7.74	0.01	~ 0
AG	-7.74	0.01	~ 0
AT	-1.22	0.09	**0.67**
CC	-9.91	0.01	~ 0
CG	-9.91	0.01	~ 0
CT	-3.38	0.09	0.005
GG	-9.91	0.01	~ 0
GT	-3.38	0.09	0.0005
TT	-2.49	0.81	**0.32**

Simple genotype caller: **Bayesian**

P(A) = 0.9 if A is the **reference** allele;
P(A) = 0.1 otherwise

AT (?)

Example: reference is T

$P(TT) = P(A)^2$

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49		

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

P(TT) = …
P(AT) = …
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49	0.56	

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
P(AT) = …
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22	0.38	
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49	0.56	

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
$P(AT) = 2f(1-f)$
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.06	~ 0
AC	-7.74	0	0
AG	-7.74	0	0
AT	-1.22	0.38	**0.93**
CC	-9.91	0	0
CG	-9.91	0	0
CT	-3.38	0	0
GG	-9.91	0	0
GT	-3.38	0	0
TT	-2.49	0.56	0.07

Better genotype caller: **Bayesian**

P(A) = f

Where f is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
$P(AT) = 2f(1-f)$
$P(AA) = (1-f)^2$

Assuming **f=0.75** and only **A and T** alleles

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.16	~ 0
AC	-7.74	0	0
AG	-7.74	0	0
AT	-1.22	0.48	**0.96**
CC	-9.91	0	0
CG	-9.91	0	0
CT	-3.38	0	0
GG	-9.91	0	0
GT	-3.38	0	0
TT	-2.49	0.36	0.38

Better genotype caller: **Empirical Bayesian**

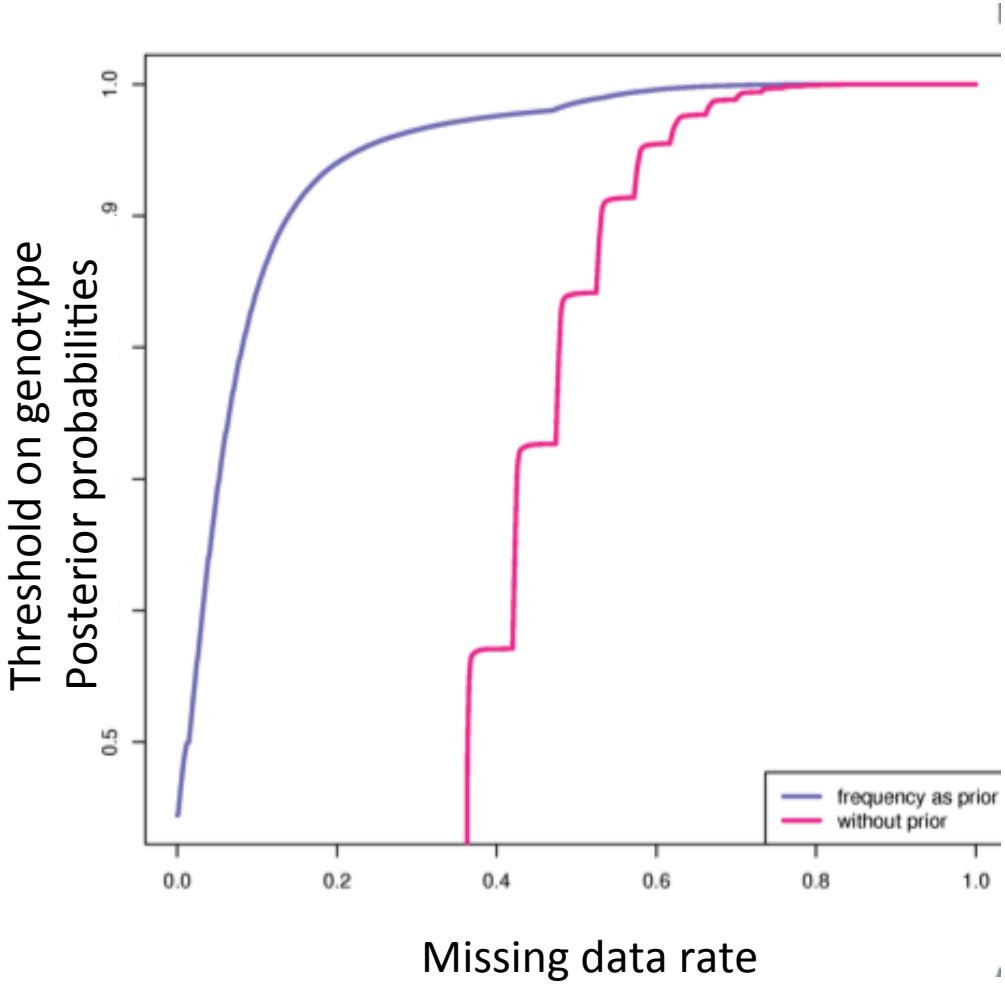
P(A) = f

Where f is the **allele frequency** estimated from the data itself

With **f=0.6**

Missing data

Mean depth 8X



Prior	Threshold	Missing data rate
No	99%	70%
No	99.9%	80%
Allele frequency	99%	50%
Allele frequency	99.9%	65%

Summary

- Data filtering should be performed keeping in mind your goals and specific features of your data.
- There is no unique perfect pipeline.
- Check your intermediate results and tune your parameters iteratively.
- Genotype calling should be performed including information from all samples.

Software

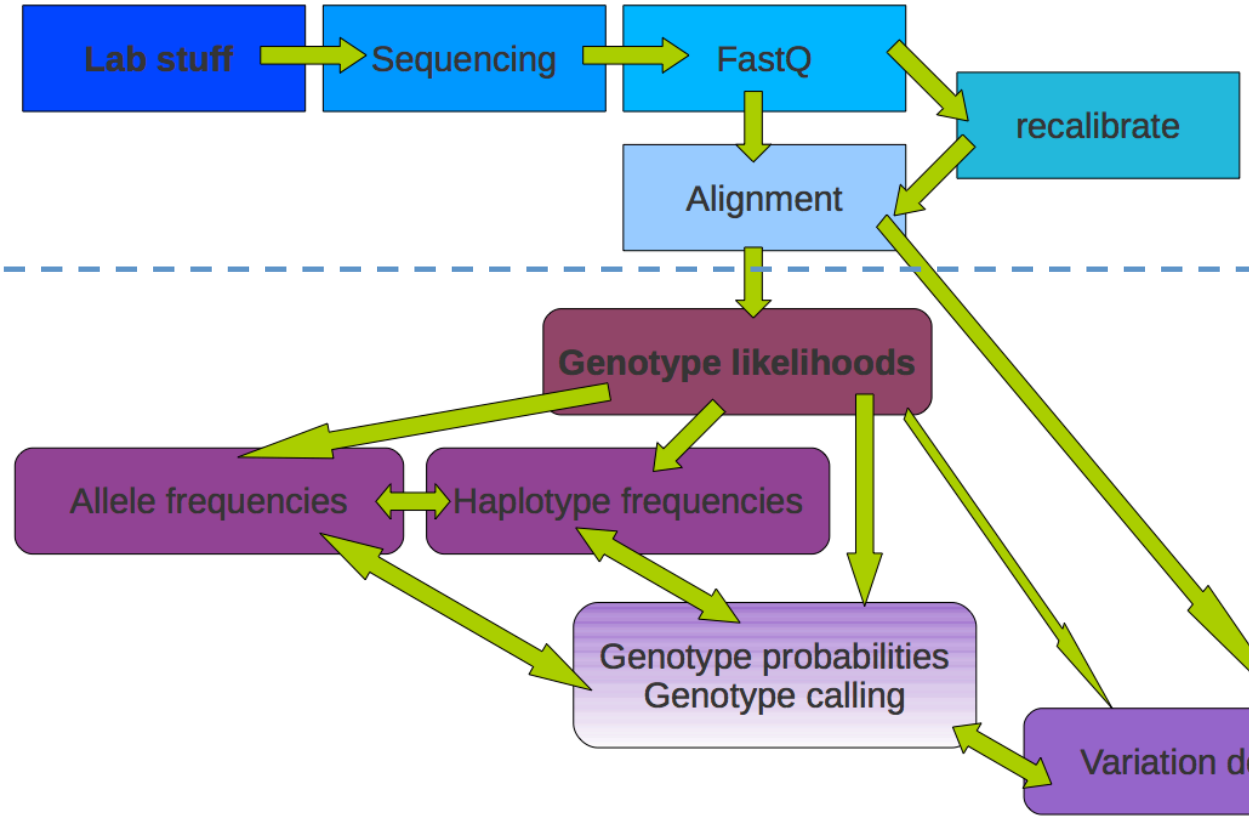
All these methods have been implemented in several software and utilities, such as:

- **SAMtools** (http://samtools.sourceforge.net)
- **GATK** (https://www.broadinstitute.org/gatk)
- **ANGSD** (http://popgen.dk/ANGSD)
- **freebayes** (https://github.com/ekg/freebayes)

Practical session using ANGSD (SAMtools)

- File formats
- Data filtering
- Assessing your filtering
- Genotype calling

Workflow

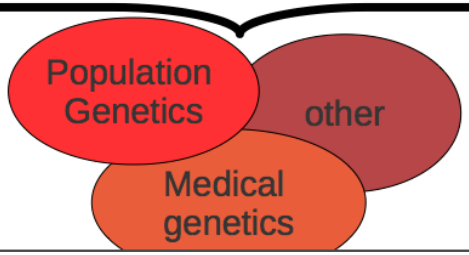


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- **Estimate allele frequencies**
- **SNPs detection**

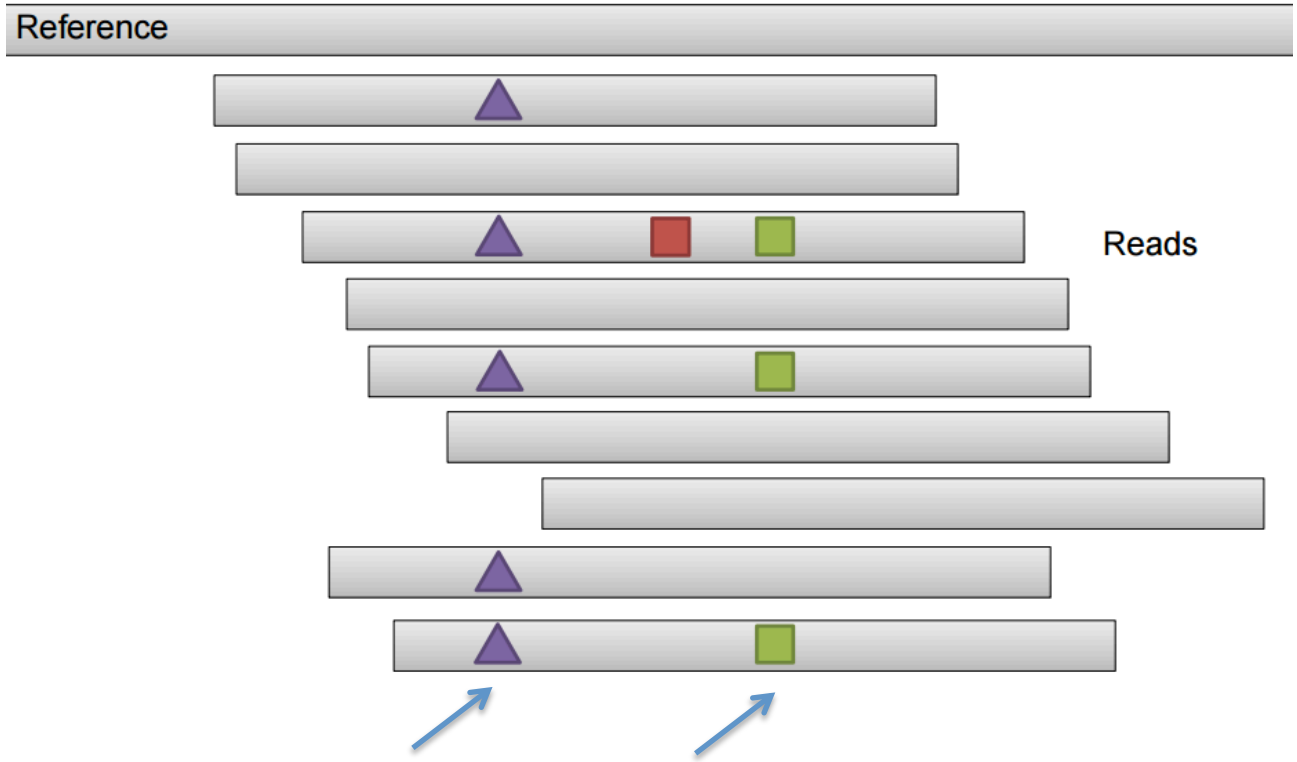


Analysis:

- Population genetics analysis
- Association studies

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-alignement around putative variants; better resolution for INDELs detection.

- Haplotype-based caller (as in freebayes)

Figure from Erik Garrison

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA		
2	AA		
3	AG		
4	AG		
5	GG		
6	GG		
Tot.			

Assume only 2 allelic types

True allele frequency is 0.50

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Assume only 2 allelic types

True allele frequency is 0.50

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

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

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})} = 0.75$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})} = 0.75$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N}(n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N}(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)}$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N}(n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N}(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)} = 0.77$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N}(n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N}(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)} = 0.77$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Simple allele frequency estimator: from **reads counts with error and weights** (Y Li et al. 2010)

$$p_i = \frac{n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)})}{(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)}$$

$$w_i = \frac{2(n_{(A,i)} + n^2_{(G,i)})}{(n_{(A,i)} + n_{(G,i)}) + 1}$$

$$\hat{f} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} p_i w_i = 0.57$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads 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
Tot.		41	14

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Genotype likelihoods

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g)\, p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Genotype likelihoods

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

If we assume HWE:

$$p(G = AA \mid f) = f^2$$

$$p(G = AG \mid f) = 2f(1-f)$$

$$p(G = GG \mid f) = (1-f)^2$$

Estimating allele frequencies

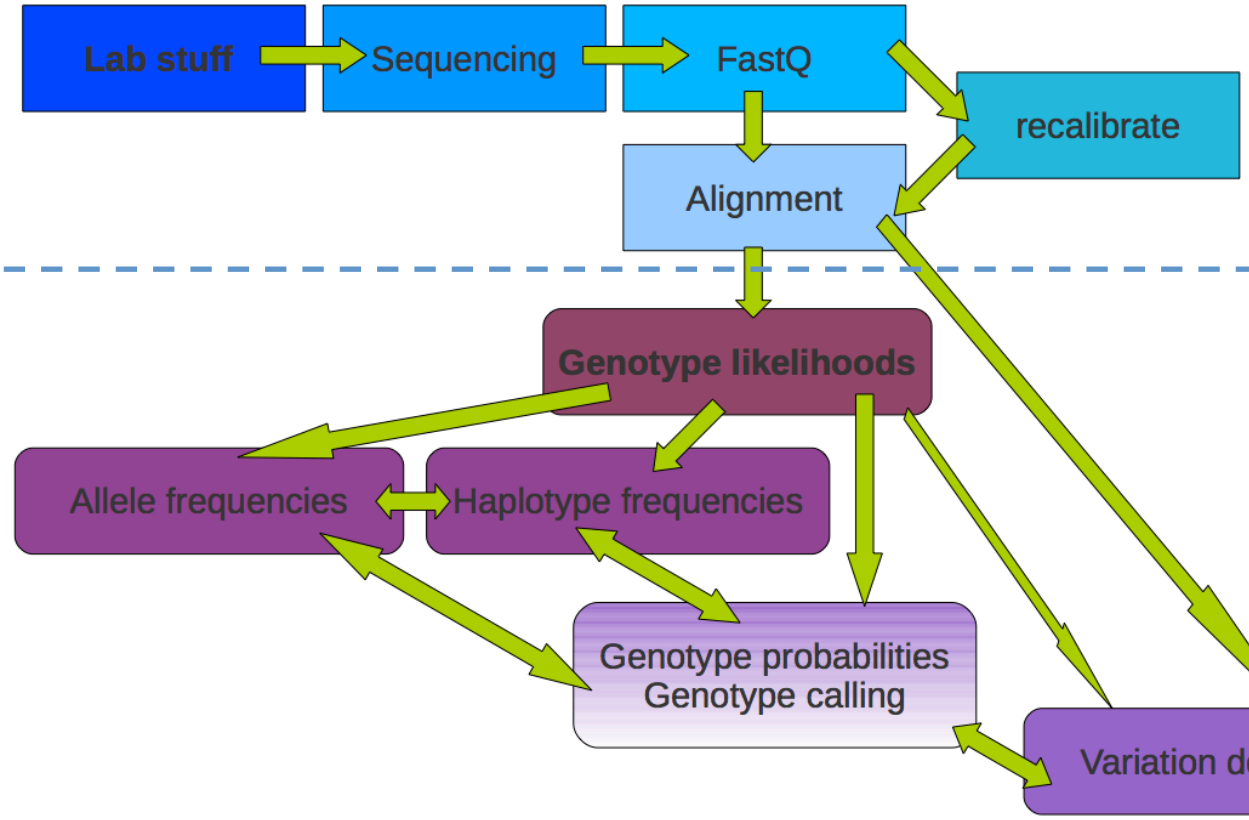
Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$\hat{f} = \arg\max_p \prod_{i=1}^{N} p(D_i \mid f)$$

$$\hat{f} = 0.46$$

Workflow

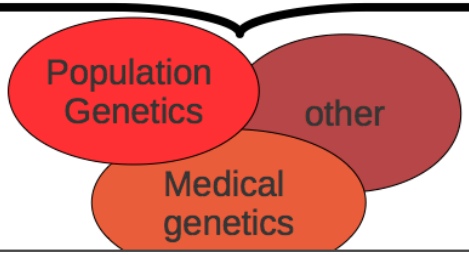


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- **SNPs detection**



Analysis:

- Population genetics analysis
- Association studies

SNP calling

- A lot of missing data if calling genotypes at low depth (heterozygotes can be lost!)
- Rare variants are hard to detect
- Trade-off between False Positives and False Negatives

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate?

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate? >99%

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate? >99%

Heavy filtering of data (error rate of 0.001):

False positive rate?

SNP calling – effect of errors

Calling SNPs if 2 alternate alleles are observed (5X and 100 samples and error rate of 0.01):

False positive rate? >99%

Heavy filtering of data (error rate of 0.001):

False positive rate? 60%

Numbers from R. Nielsen's slide

SNP calling

- What is the most straightforward method to for SNP calling?

SNP calling

- What is the most straightforward method to for SNP calling?
  - Assign as SNPs sites where at least one heterozygote has been called
  - …

SNP calling

- What is the most straightforward method to for SNP calling?
  - Assign as SNPs sites where at least one heterozygote has been called
  - Assign as SNPs sites where the estimated allele frequency is above a certain threshold (e.g. ?)

SNP calling

- MLE of allele frequency at each site:

Call a SNP if

$$\hat{f}_{MLE} > t$$

Where *t* can be defined as the minimum sample allele frequency detectable (e.g. with 10 samples *t* can be set to 0.05)

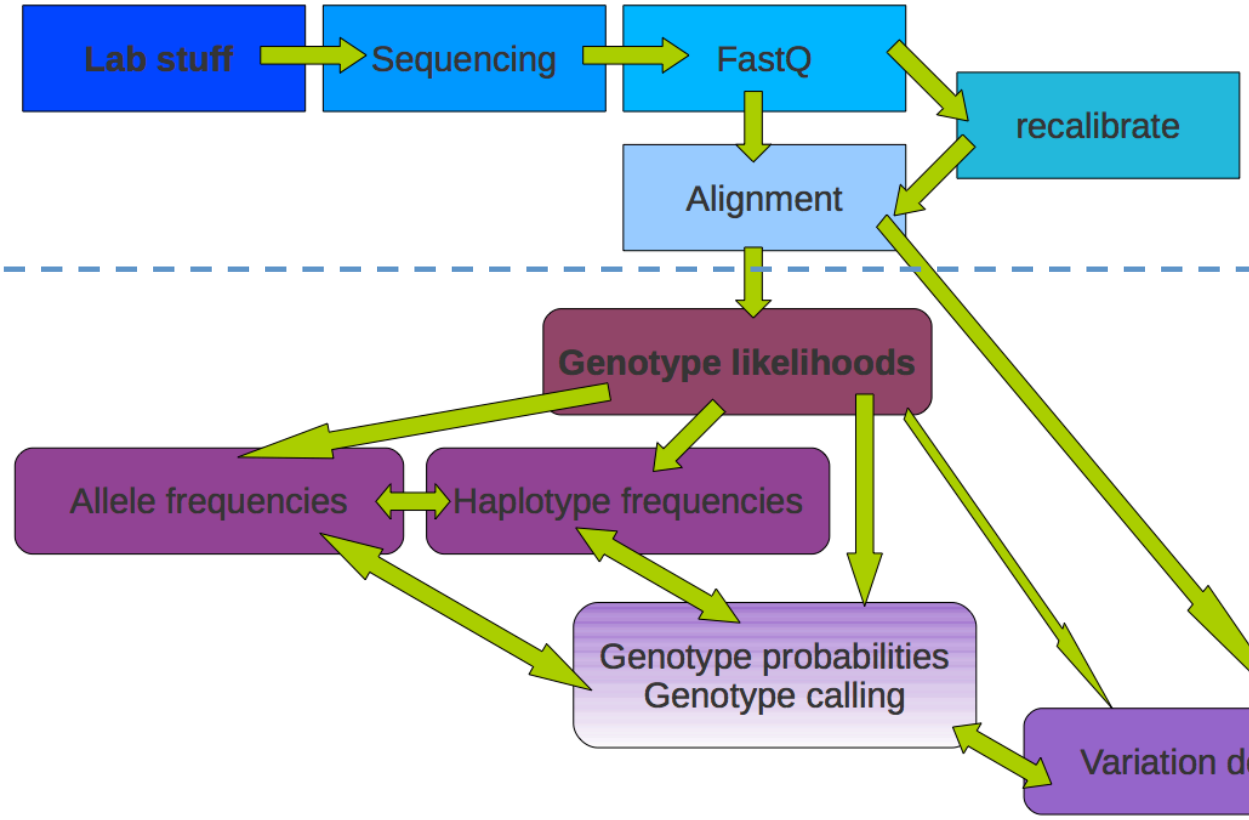
SNP calling

- Likelihood Ratio Test (**LRT**): test statistical hypotheses based on comparing the maximum likelihood under 2 different models.

$$T = -2 \ln \left(\frac{L(f = 0)}{L(f \neq 0)} \right)$$

T is chi-squared distributed with 1 degree of freedom -> assign a *p*-value

Workflow

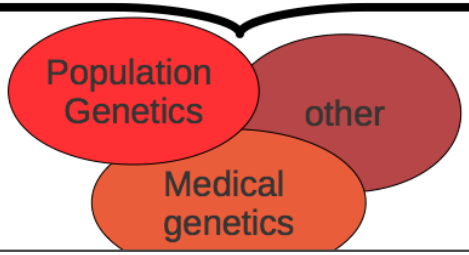


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- SNPs detection



Population genetics analysis:

- **Site Frequency Spectrum**
- Summary statistics

Site Frequency Spectrum (SFS)

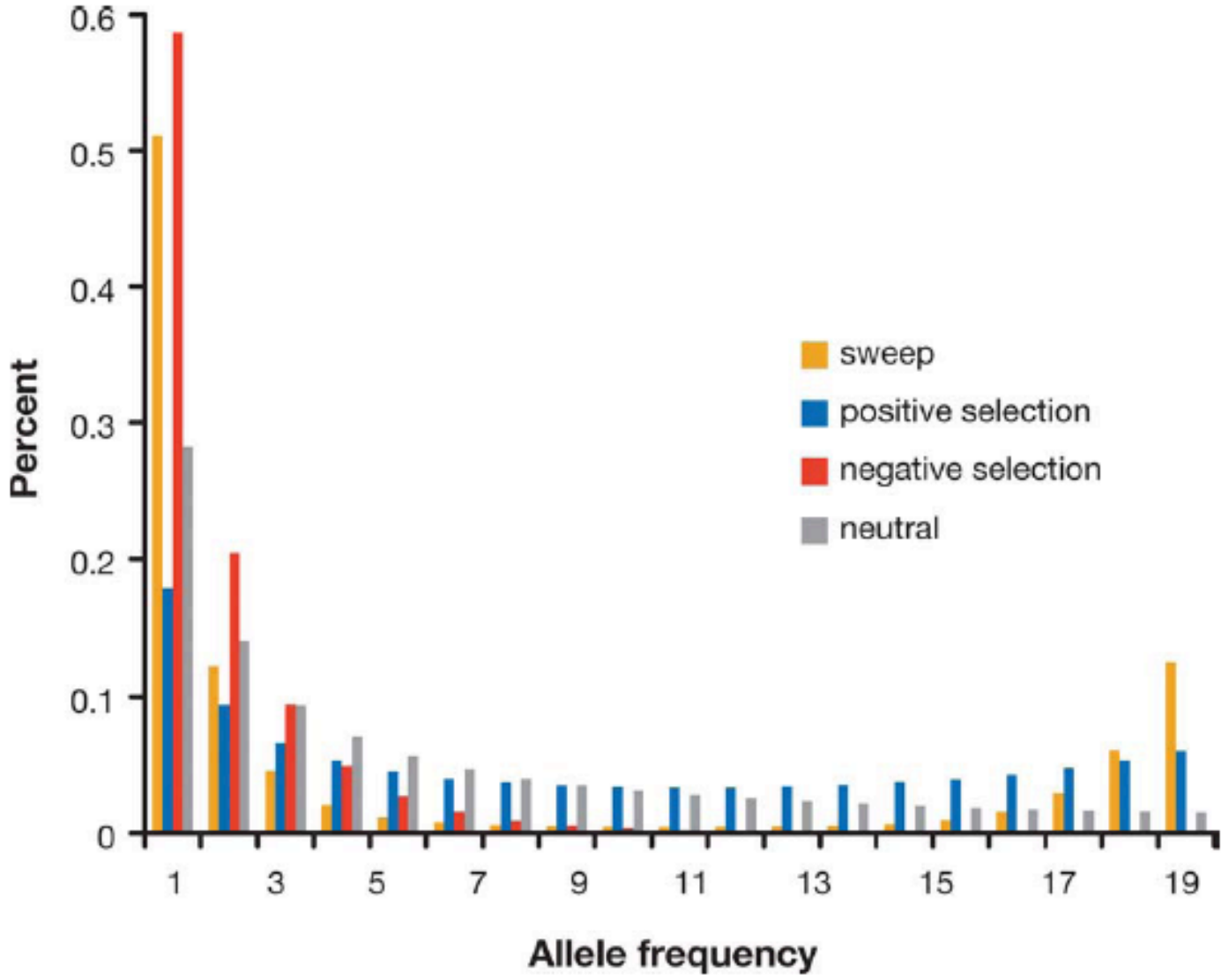
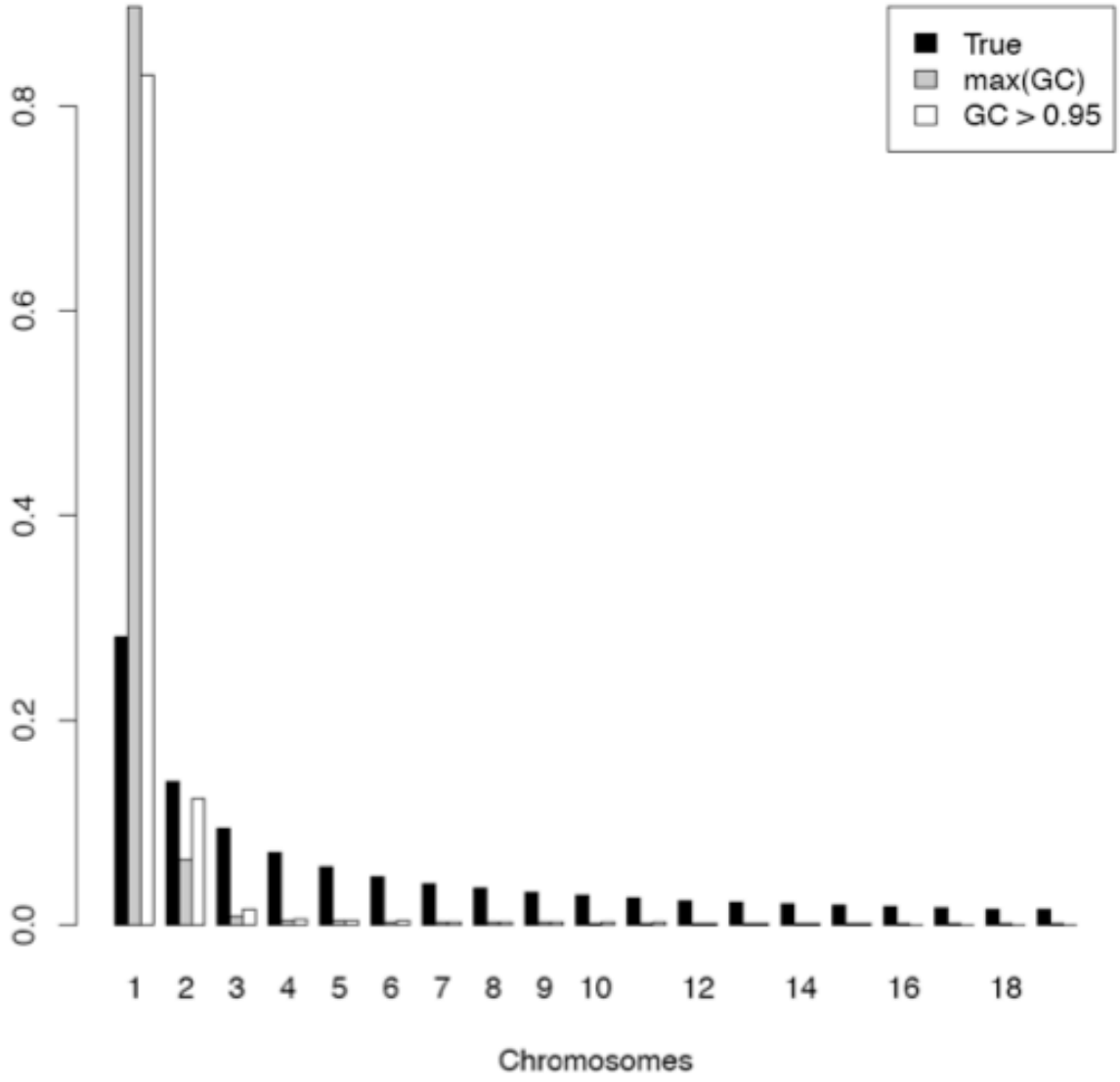


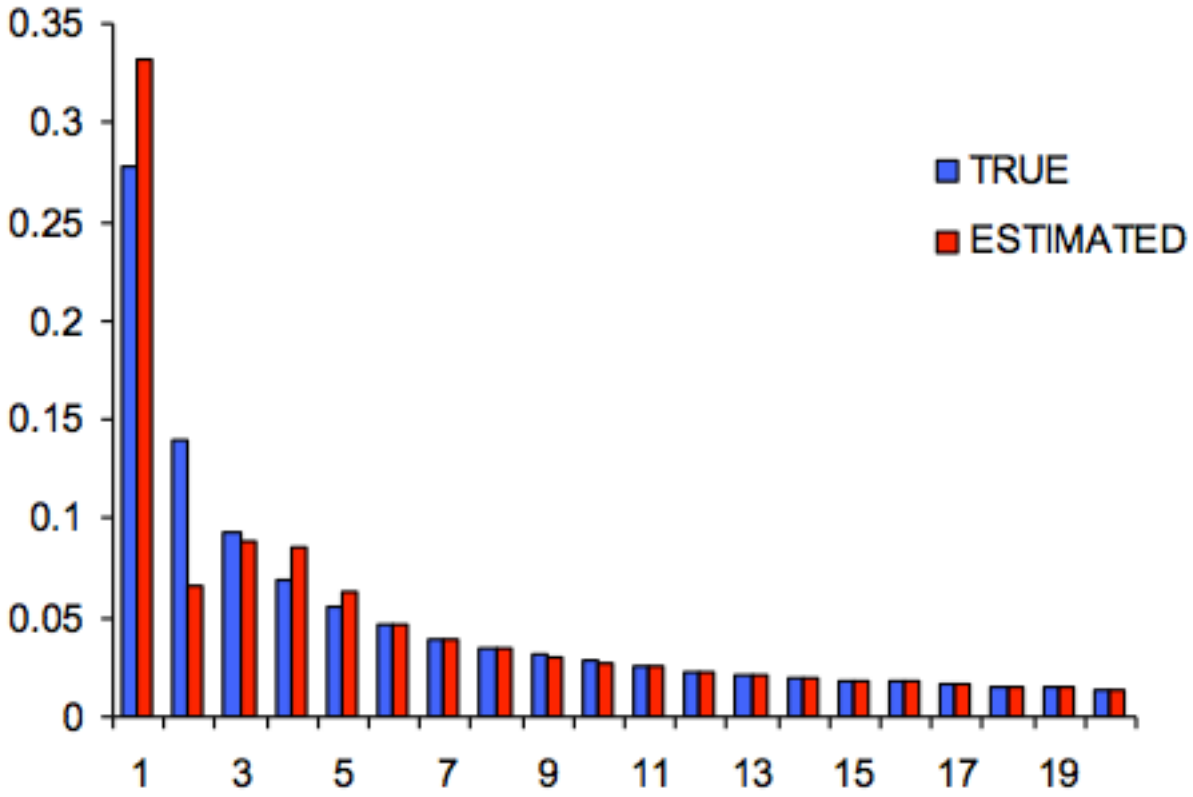
Figure 2

Effect of errors on SFS



Effect of errors on SFS

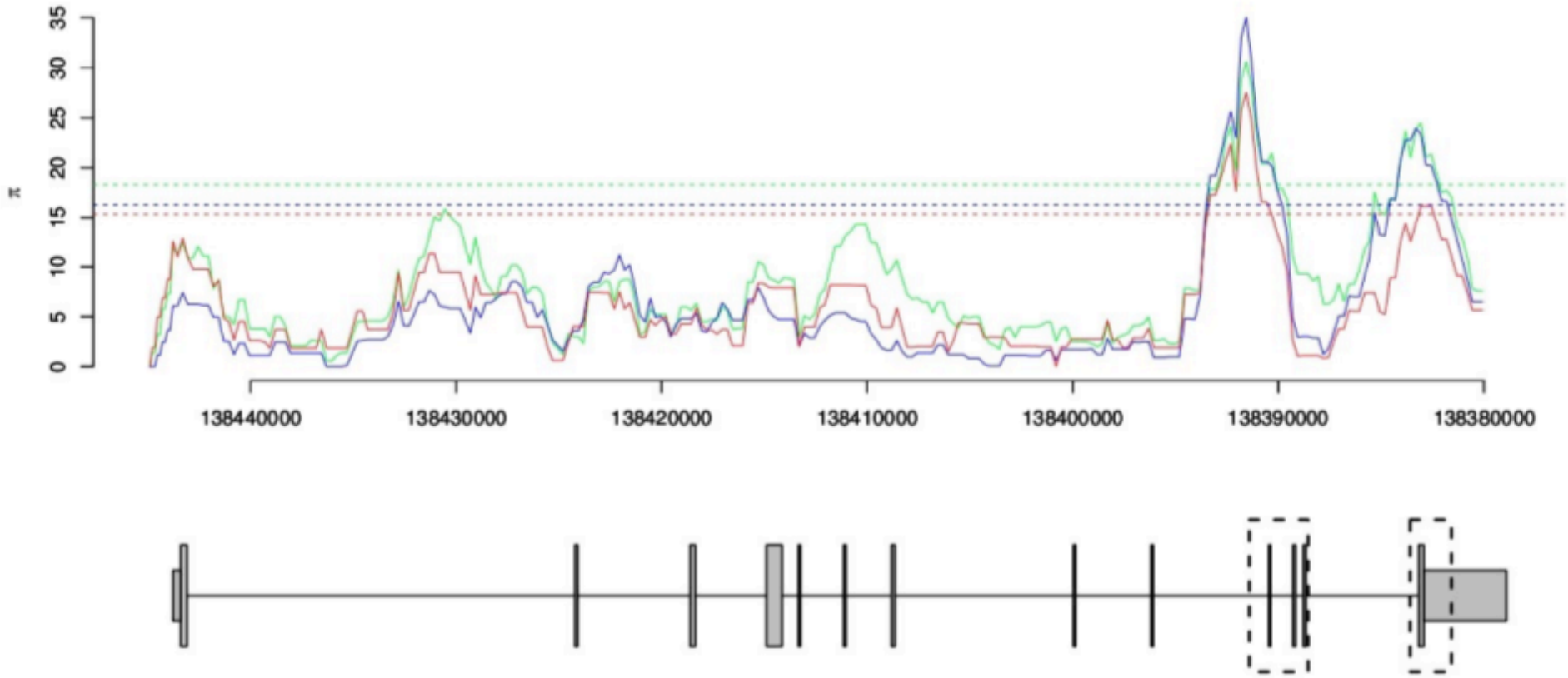
Using an ad hoc fixed cutoff for SNP calling…



can never produce unbiased estimates.

Effects of low-depth data

Nucleotide diversity scan using 1000 Genomes Project data (low-depth)

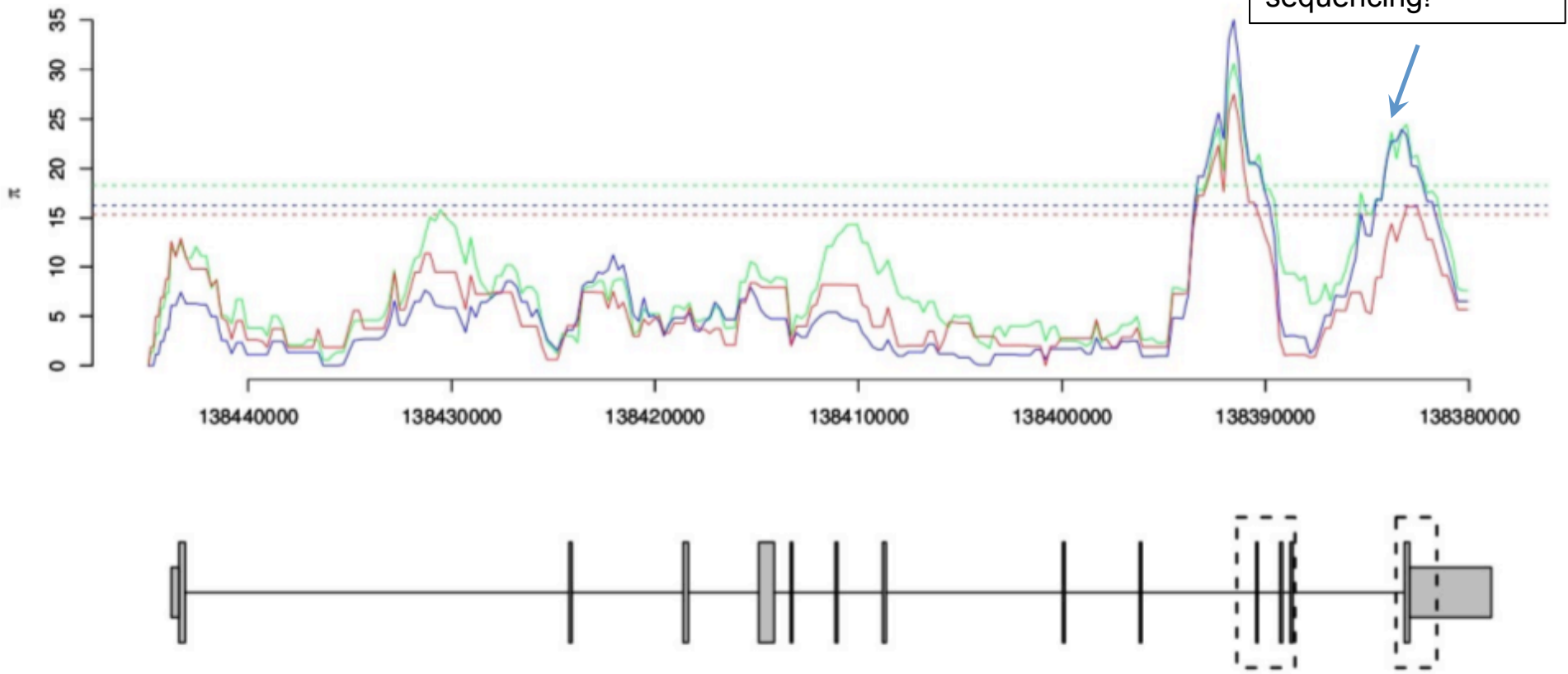


Cagliani et al. MBE. 2012

Effects of low-depth data

Nucleotide diversity scan using 1000 Genomes Project data (low-depth)

Highest peak based on Sanger sequencing!



Cagliani et al. MBE 2012

Effects of low-depth data

SNP		Population	MAF[a]
Position[b]	ID[c]		
REGION 2			
138383386	n.a.[d]	CEU	0.03
138382592[e]	rs5022944	CEU	0.40
		AS	0.40
138382528[e]	rs5022945	YRI	0.38
		CEU	0.40
		AS	0.40
138382507[e]	rs5022946	YRI	0.38
		CEU	0.40
		AS	0.40
138382444[e]	rs10250460	YRI	0.38
		CEU	0.40
		AS	0.40
138382438[e]	rs10250457	YRI	0.38
		CEU	0.40
		AS	0.40
138382399[e]	rs10250646	YRI	0.38
		CEU	0.40
		AS	0.40
138382383[e]	rs10250435	YRI	0.38
		CEU	0.40
		AS	0.40
138382350[e]	rs10265856	YRI	0.38
		AS	0.40
138382205	n.a.[d]	AS	0.03

- Sanger: detected a total of 24 variants
- NGS: only 13

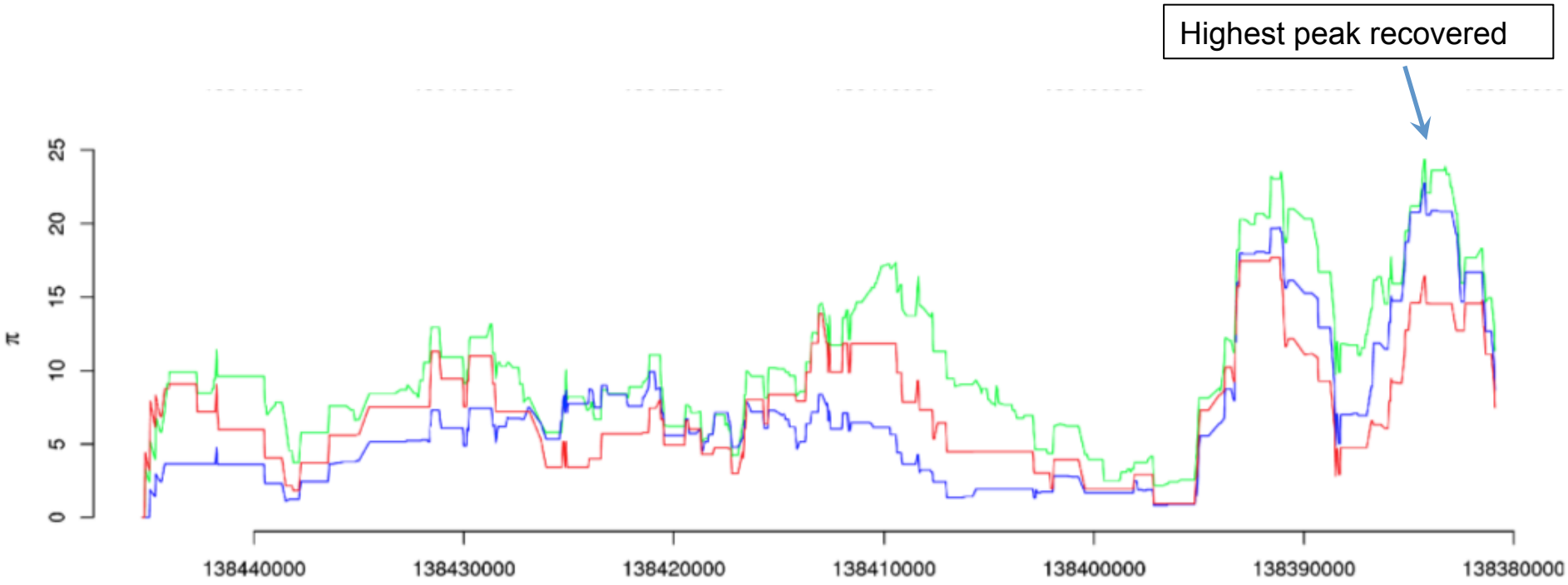
Most of them (n=8) have intermediate frequency in all populations.

They are located within an AluSx element in the 3'UTR.

A large portion of “inaccessible Sites” in the low-depth1000 Genomes data maps to repetitive sequences.

Cagliani et al. MBE 2012

Masked data



- Missing data
- Unpredictable effects

Cagliani et al. MBE 2012

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

- Parameterize the SFS, with *k* individuals

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

If unfolded, *?* entries

If folded, *?* entries

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

- Parameterize the SFS, with *k* individuals

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

If unfolded, *2k+1* entries

p_0	p_1	p_2	p_3	…	p_{2k}

If folded, *2k* entries

p_0	p_1	p_2	…	p_k

ML estimation of the SFS

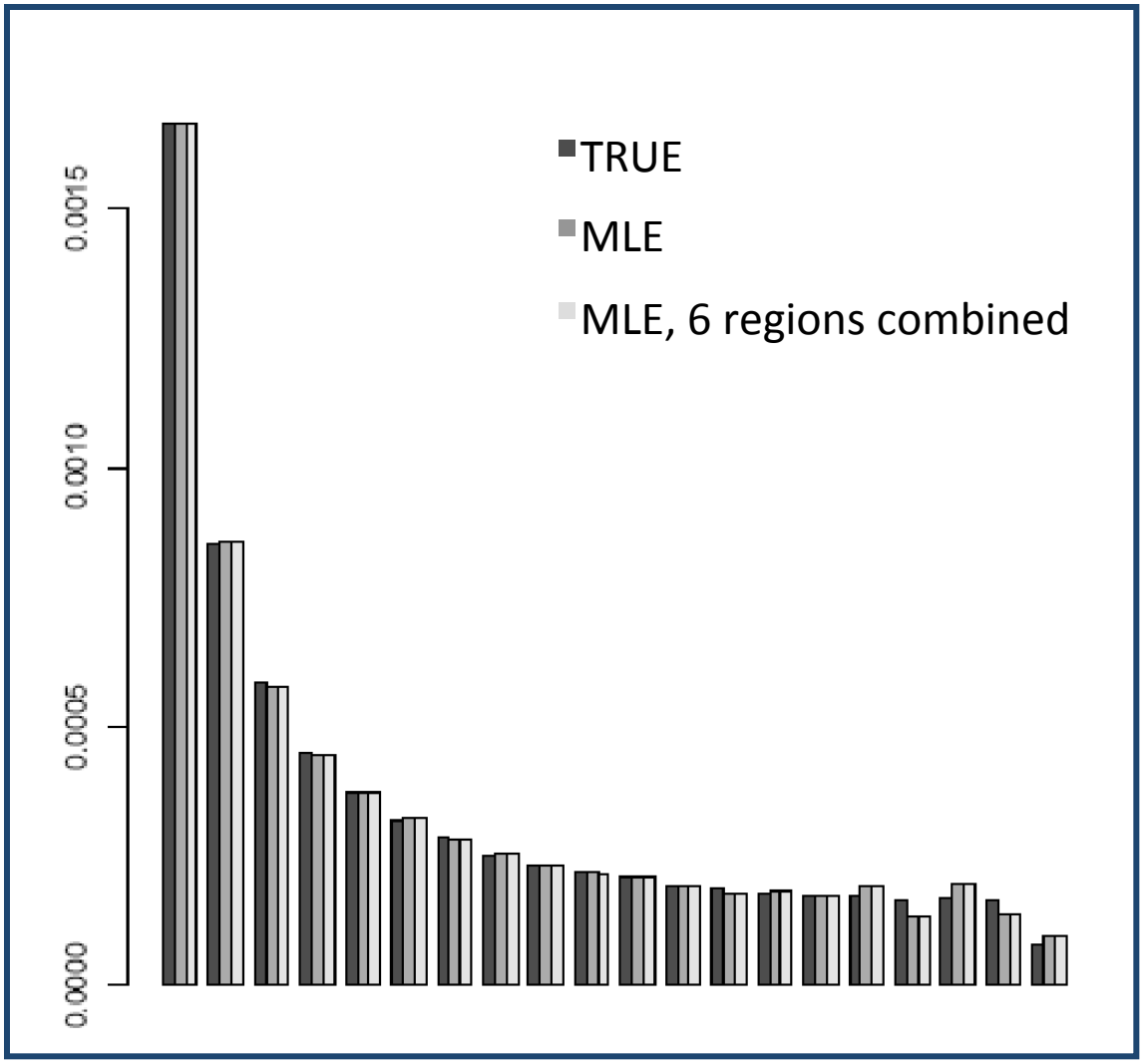
Summing across all unknown genotypes and multiplying the likelihood across sites.

- Likelihood function:

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

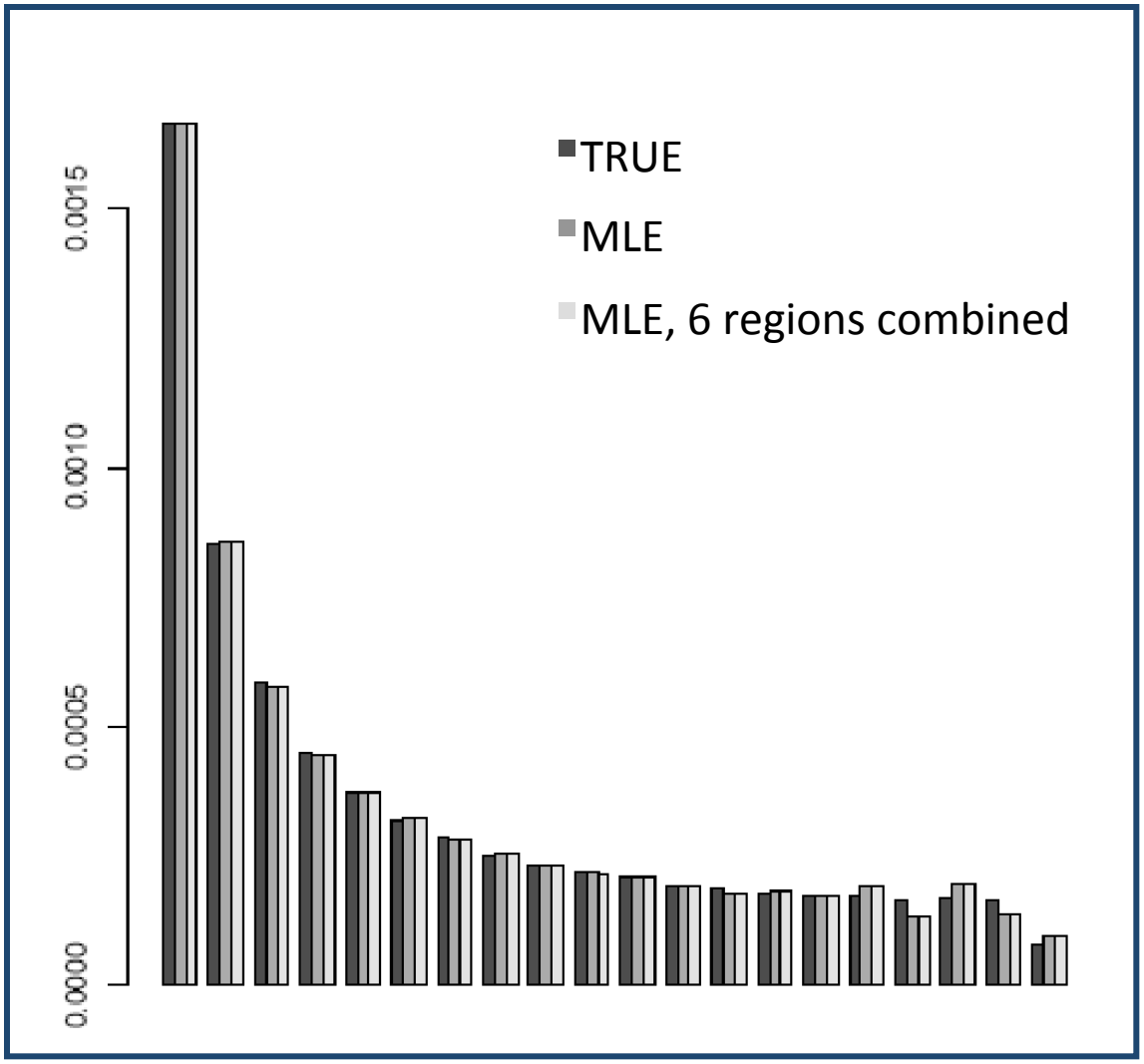
Nielsen et al. 2012 PLoS One

ML estimation of the SFS



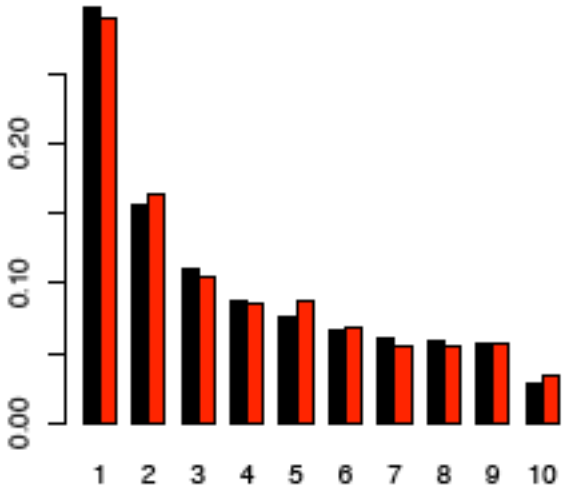
Simulated 30Mb
Error rate of 0.3%
Mean depth of 5X

ML estimation of the SFS



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

Mean depth of 1X:



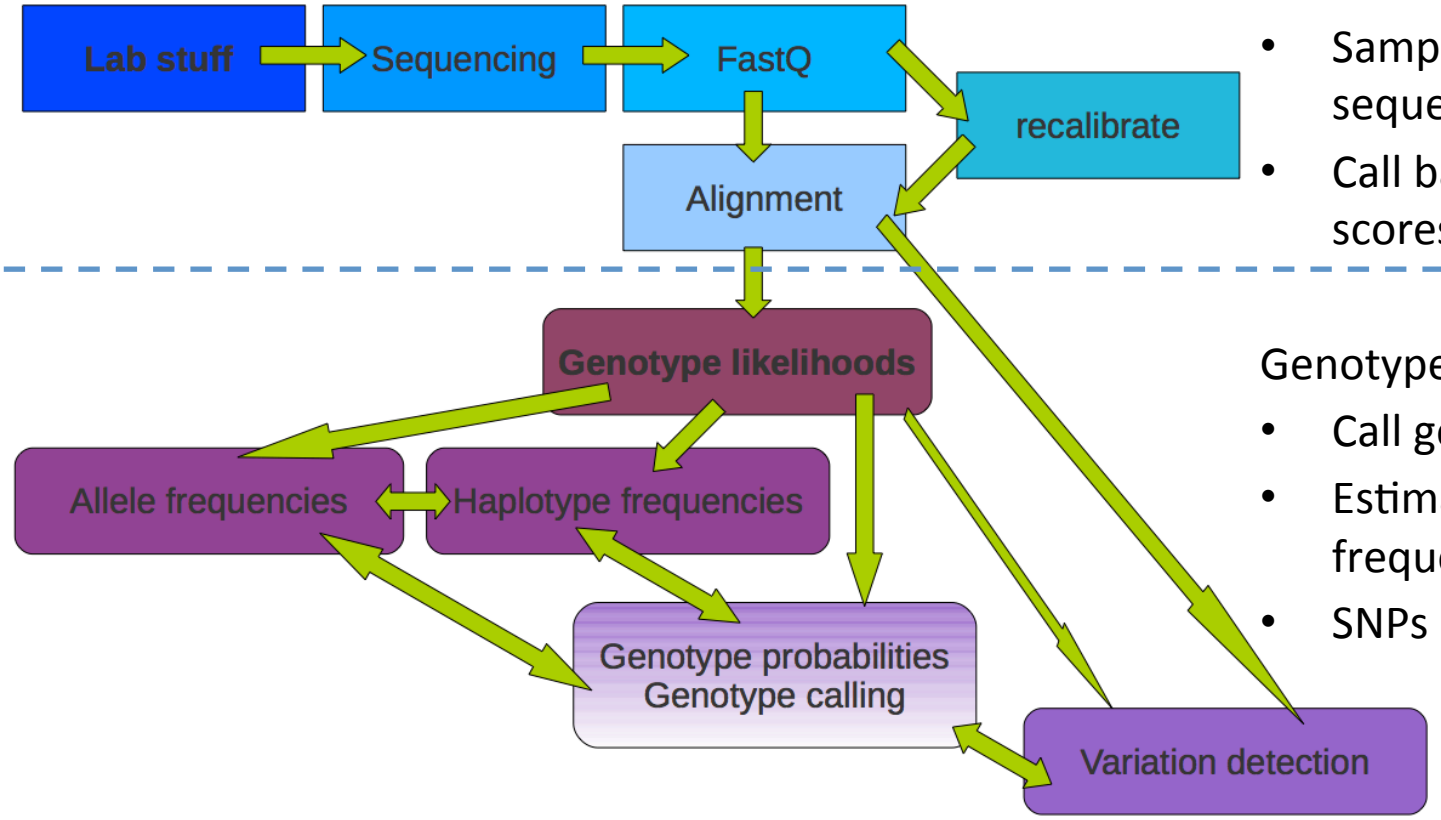
R. Nielsen's slide

ML estimation of the SFS

Can be used for:

- SNP calling
- Genotype calling
- Modeling uncertainty in population genetics analyses

Workflow

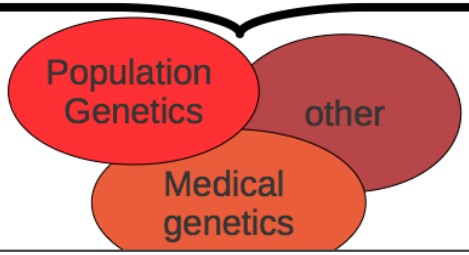


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- SNPs detection



Population genetics analysis:

- Site Frequency Spectrum
- **Summary Statistics**

Sample allele frequency posterior probabilities

S_m: sample allele frequency at site *m*

Likelihood Prior

$$p(S_m = j \mid X) \propto p(X \mid S_m = j)\, p(S_m = j)$$

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

Sample allele frequency posterior probabilities

S_m: sample allele frequency at site *m*

Likelihood

Prior

$$p(S_m = j \mid X) \propto p(X \mid S_m = j)\, p(S_m = j)$$

Estimate of the overall SFS

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

Sample allele frequency posterior probabilities

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

- Estimating allele frequency

Expected value

- The expected value of a discrete random variable is the probability-weighted average of all possible values
- Average value if you perform the same experiment many times
- It is the value that one could expect on average

Sample allele frequency posterior probabilities

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

- Estimating allele frequency

$$\hat{f} =$$

Used as prior for genotype calling

Sample allele frequency posterior probabilities

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

- Estimating allele frequency

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

Used as prior for genotype calling

Sample allele frequency posterior probabilities

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

- SNP calling

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

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

with *t* being 0.05, 0.01., 0.001 and so on.

Sample allele frequency posterior probabilities

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

- SNP calling

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

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

with *t* being 0.05, 0.01., 0.001 and so on.

Nr of segregating sites

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

Nr of segregating sites

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

Nr of segregating sites

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

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

Nucleotide diversity

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

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

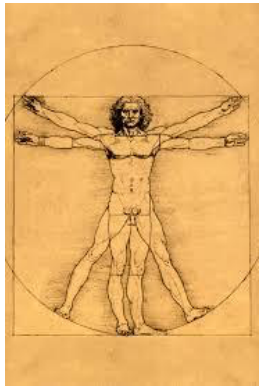
$$E[D] =$$

Nucleotide diversity

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

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

Applications



- Model and non-model species
- Plants
- Vertebrates and invertebrates
- Ancient DNA

…

Software

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

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

which we will explore during the practical session.

Summary

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

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

References

- Nielsen *et al.* Nat Rev Genet 2011 (21587300)
- Li H. Bioinformatics 2011 (21903627)
- Kim *et al.* BMC Bioinformatics 2011 (21663684)
- Fumagalli M. PLoS One 2013 (24260275)

* PubMed ID: http://www.ncbi.nlm.nih.gov/pubmed/*

Practical exercises

- Estimating allele frequencies
- SNP calling
- Estimating the Site Frequency Spectrum
- Estimating summary statistics

Paper(s) discussion

MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3028–3035 doi: 10.1111/mec.12105

Population genomics based on low coverage sequencing: how low should we go?

C. ALEX BUERKLE* and ZACHARIAH GOMPERT†

*Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY, USA, †Department of Biology, Texas State University, San Marcos, TX, USA

OPEN ACCESS Freely available online

PLOS | ONE

Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences

Matteo Fumagalli*

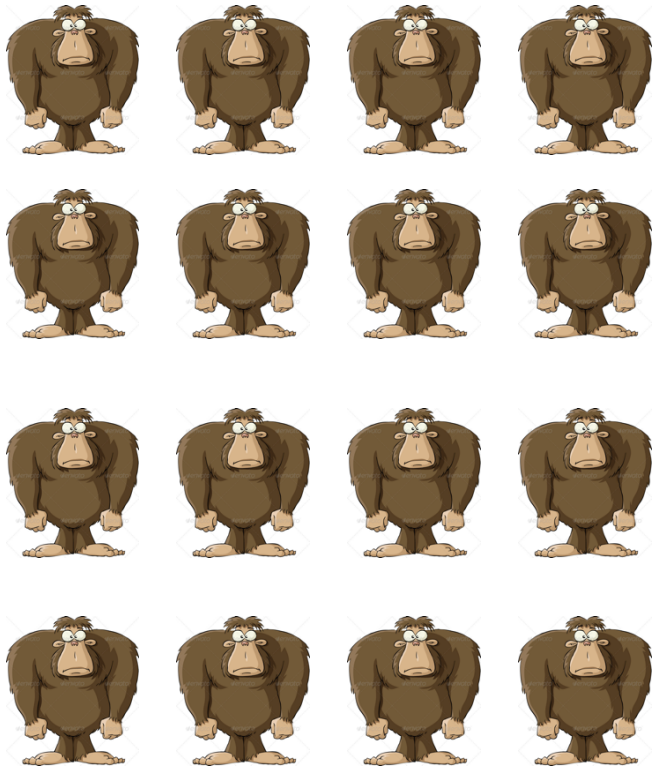
Department of Integrative Biology, University of California, Berkeley, California, United States of America

Experimental design

- You discovered a new species!

Experimental design

Population of 1,000 individuals

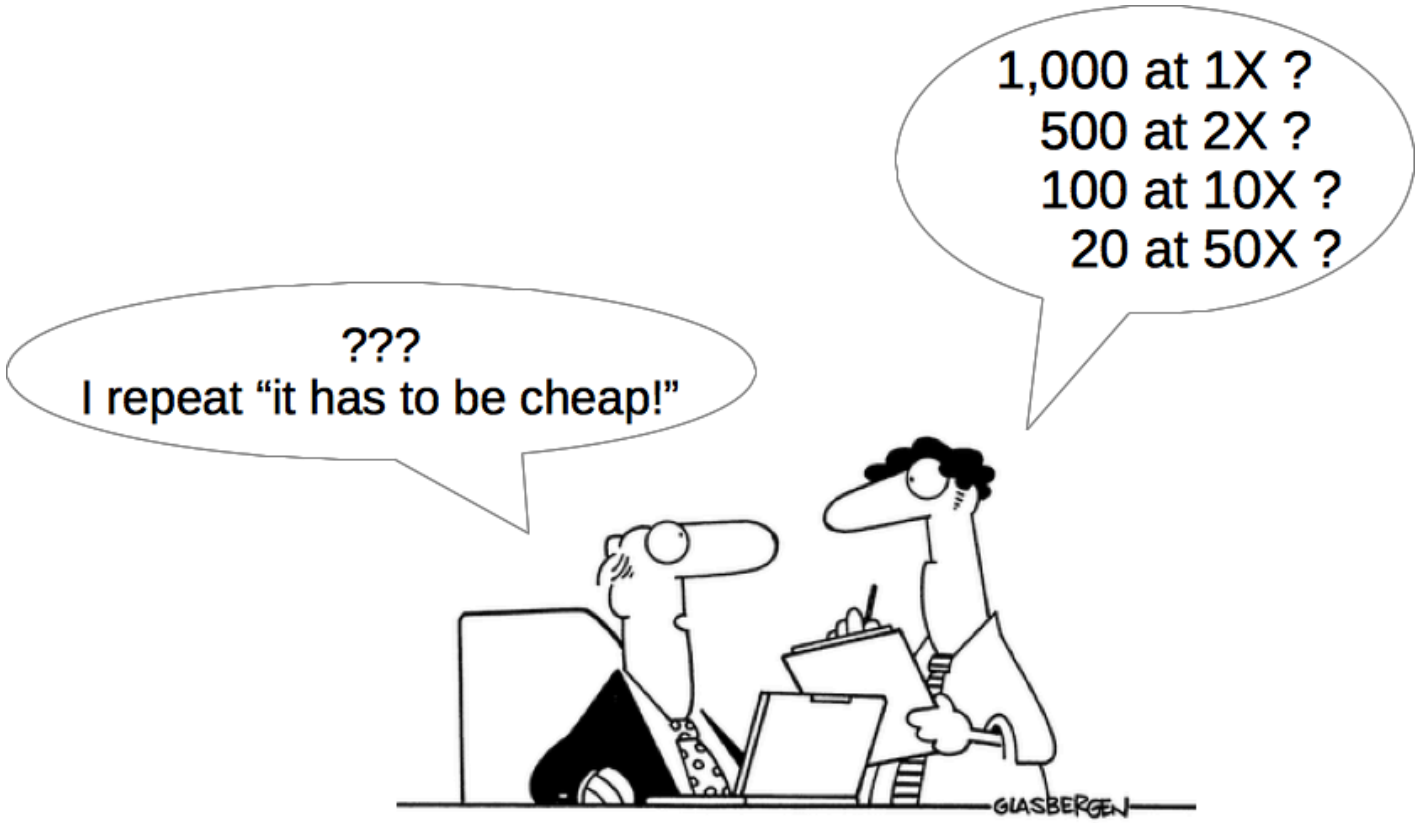


Experimental design

Experimental design

Experimental design

Experimental design



Experimental design

At a fixed budget: 

- sequencing **more samples** will lower the per-sample sequencing depth, and, as a consequence, increase the genotype uncertainty.

- **higher sequencing coverage** will decrease genotyping uncertainty, but will also restrict the analysis to a smaller sample of individuals, which may be a poor representation of the genomic variation of the entire population

Experimental design

At a fixed budget:

- sequencing **more samples** will lower the per-sample sequencing depth, and, as a consequence, increase the genotype uncertainty.
- **higher sequencing coverage** will decrease genotyping uncertainty, but will also restrict the analysis to a smaller sample of individuals, which may be a poor representation of the genomic variation of the entire population

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

Experimental design

At a fixed budget:

- sequencing **more samples** will lower the per-sample sequencing depth, and, as a consequence, increase the genotype uncertainty.
- **higher sequencing coverage** will decrease genotyping uncertainty, but will also restrict the analysis to a smaller sample of individuals, which may be a poor representation of the genomic variation of the entire population

ARTICLE

Deep Whole-Genome Sequencing of 100 Southeast Asian Malays

Lai-Ping Wong,[1,14] Rick Twee-Hee Ong,[1,14] Wan-Ting Poh,[1,14] Xuanyao Liu,[1,2,14] Peng Chen,[1] Ruoying Li,[1] Kevin Koi-Yau Lam,[1] Nisha Esakimuthu Pillai,[3] Kar-Seng Sim,[4] Haiyan Xu,[1] Ngak-Leng Sim,[4] Shu-Mei Teo,[1,2] Jia-Nee Foo,[4] Linda Wei-Lin Tan,[1] Yenly Lim,[1] Seok-Hwee Koo,[5] Linda Seo-Hwee Gan,[6] Ching-Yu Cheng,[1,10,11] Sharon Wee,[1] Eric Peng-Huat Yap,[6] Pauline Crystal Ng,[4] Wei-Yen Lim,[1] Richie Soong,[7] Markus Rene Wenk,[8,9] Tin Aung,[10,11] Tien-Yin Wong,[10,11] Chiea-Chuen Khor,[1,4,10,12] Peter Little,[3] Kee-Seng Chia,[1] and Yik-Ying Teo[1,2,3,4,13,*]

Whole-genome sequencing across multiple samples in a population provides an unprecedented opportunity for comprehensively characterizing the polymorphic variants in the population. Although the 1000 Genomes Project (1KGP) has offered brief insights into the value of population-level sequencing, the low coverage has compromised the ability to confidently detect rare and low-frequency variants. In addition, the composition of populations in the 1KGP is not complete, despite the fact that the study design has been extended to more than 2,500 samples from more than 20 population groups. The Malays are one of the Austronesian groups predominantly present in Southeast Asia and Oceania, and the Singapore Sequencing Malay Project (SSMP) aims to perform deep whole-genome sequencing of 100 healthy Malays. By sequencing at a minimum of 30× coverage, we have illustrated the higher sensitivity at detecting low-frequency and rare variants and the ability to investigate the presence of hotspots of functional mutations. Compared to the low-pass sequencing in the 1KGP, the deeper coverage allows more functional variants to be identified for each person. A comparison of the fidelity of genotype imputation of Malays indicated that a population-specific reference panel, such as the SSMP, outperforms a cosmopolitan panel with larger number of individuals for common SNPs. For lower-frequency (<5%) markers, a larger number of individuals might have to be whole-genome sequenced so that the accuracy currently afforded by the 1KGP can be achieved. The SSMP data are expected to be the benchmark for evaluating the value of deep population-level sequencing versus low-pass sequencing, especially in populations that are poorly represented in population-genetics studies.

Simulations design

The sequencing strategy can easily be modelled in terms of the number of sequenced samples and the per-sample sequencing depth.

Sample size	Per-sample depth
1,000	1X
500	2X
100	10X
20	50X

total depth is 1,000X

Simulations design

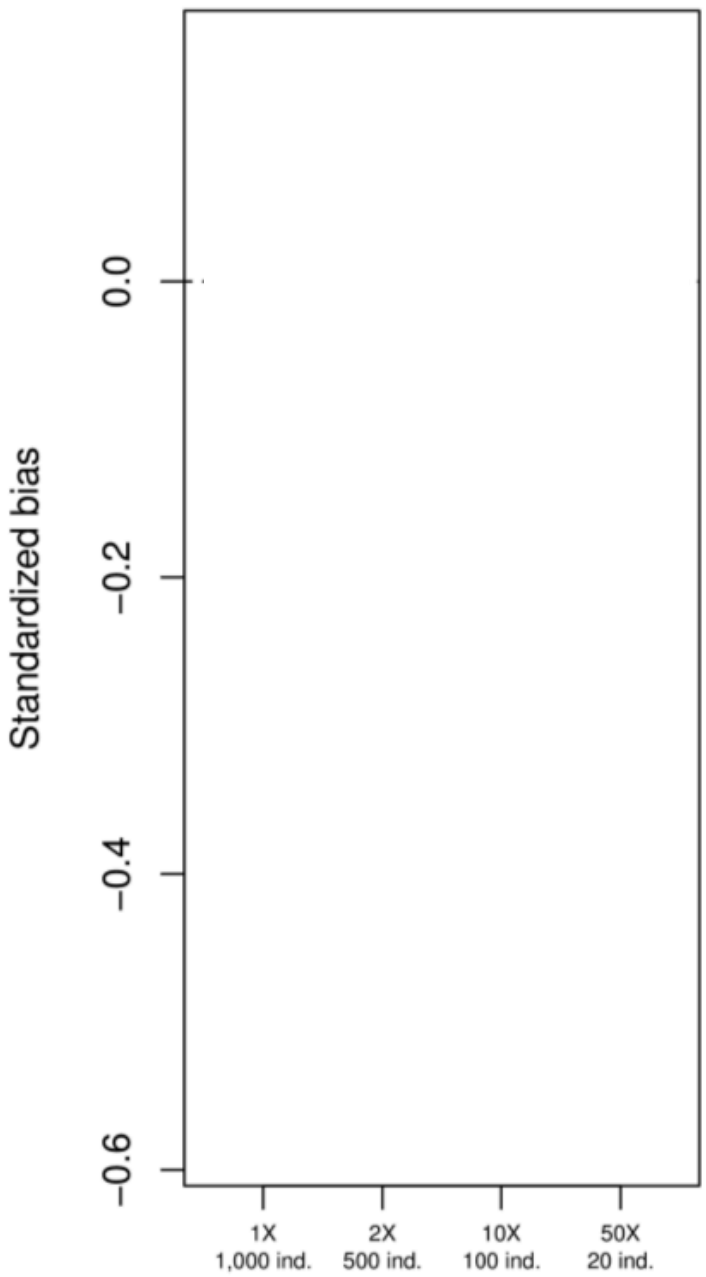
$$S = \sum_{s=1}^{L} I_s \qquad (1)$$

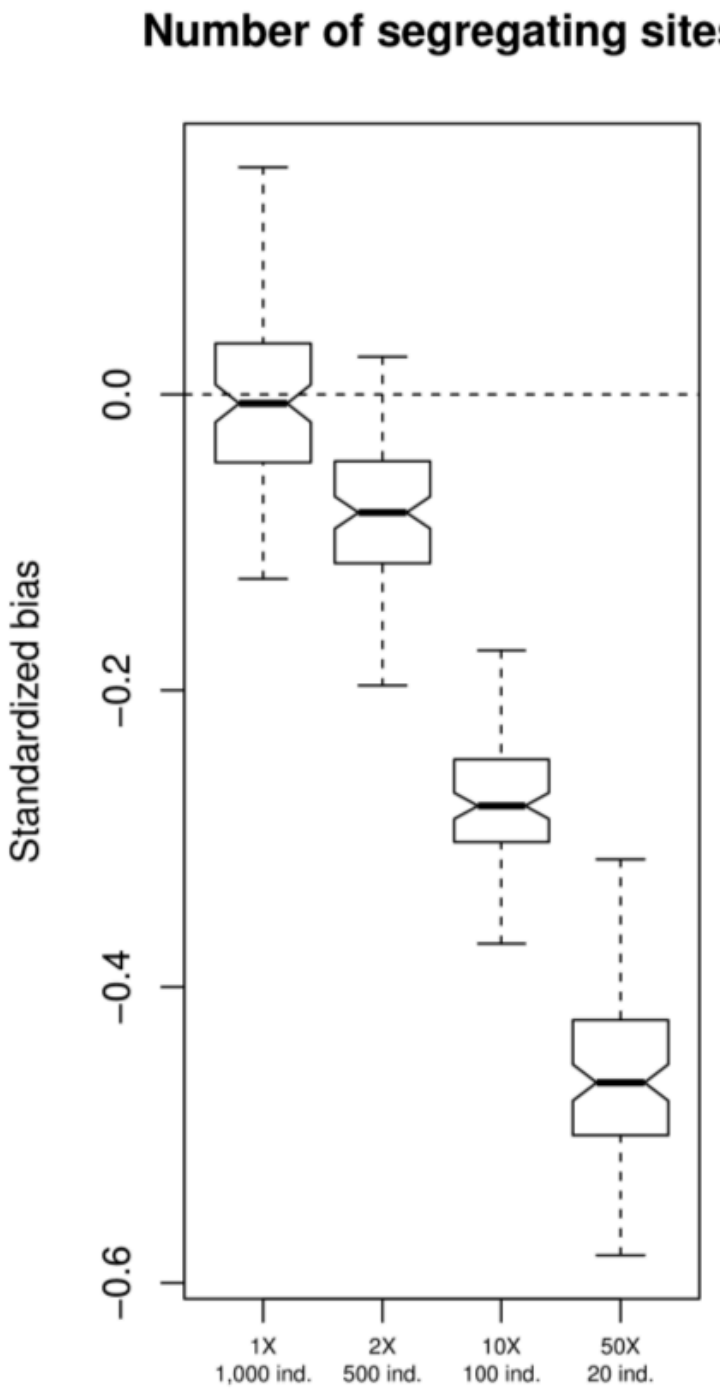
where I_s is an indicator function equal to 1 when at least one individual is heterozygous at site s, and 0 otherwise, and

$$H = \sum_{s=1}^{L} 2f_s(1 - f_s); \qquad (2)$$

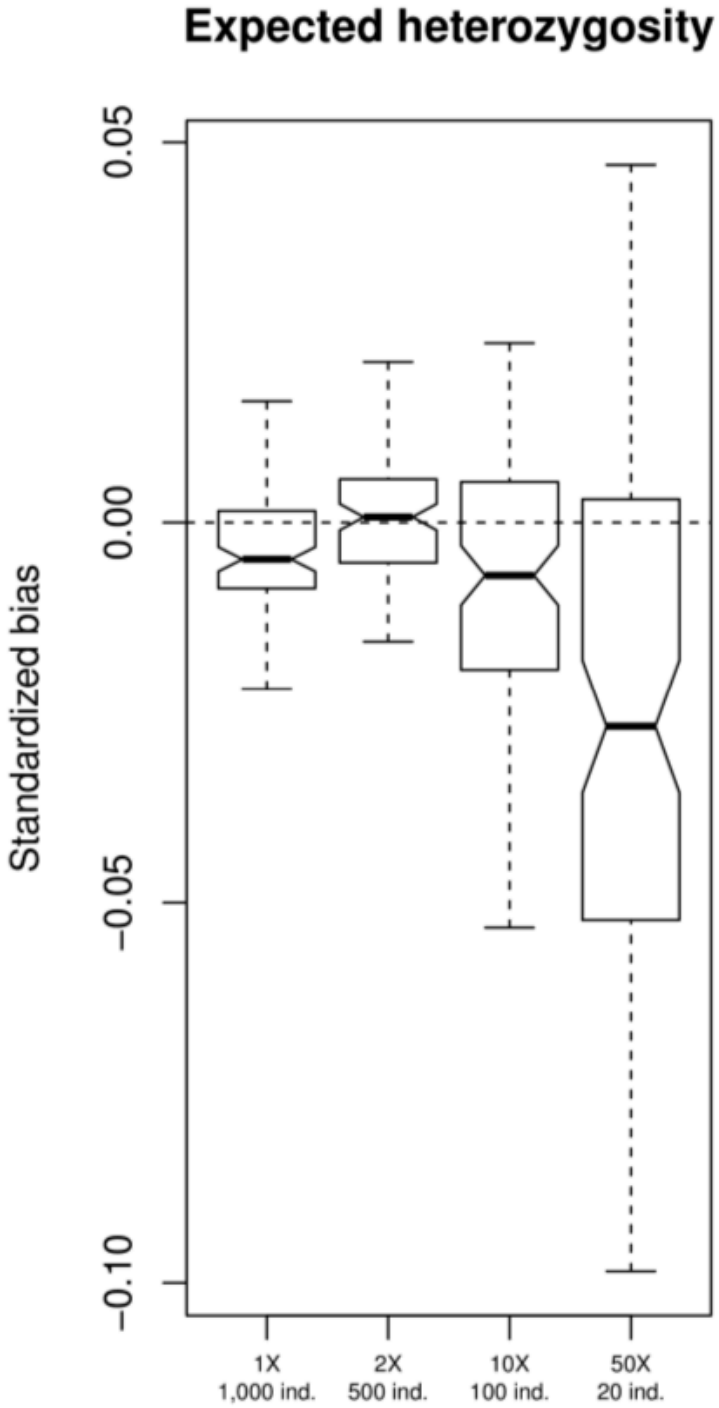
where f_s is the reference allele frequency for a site, s, in the sample.

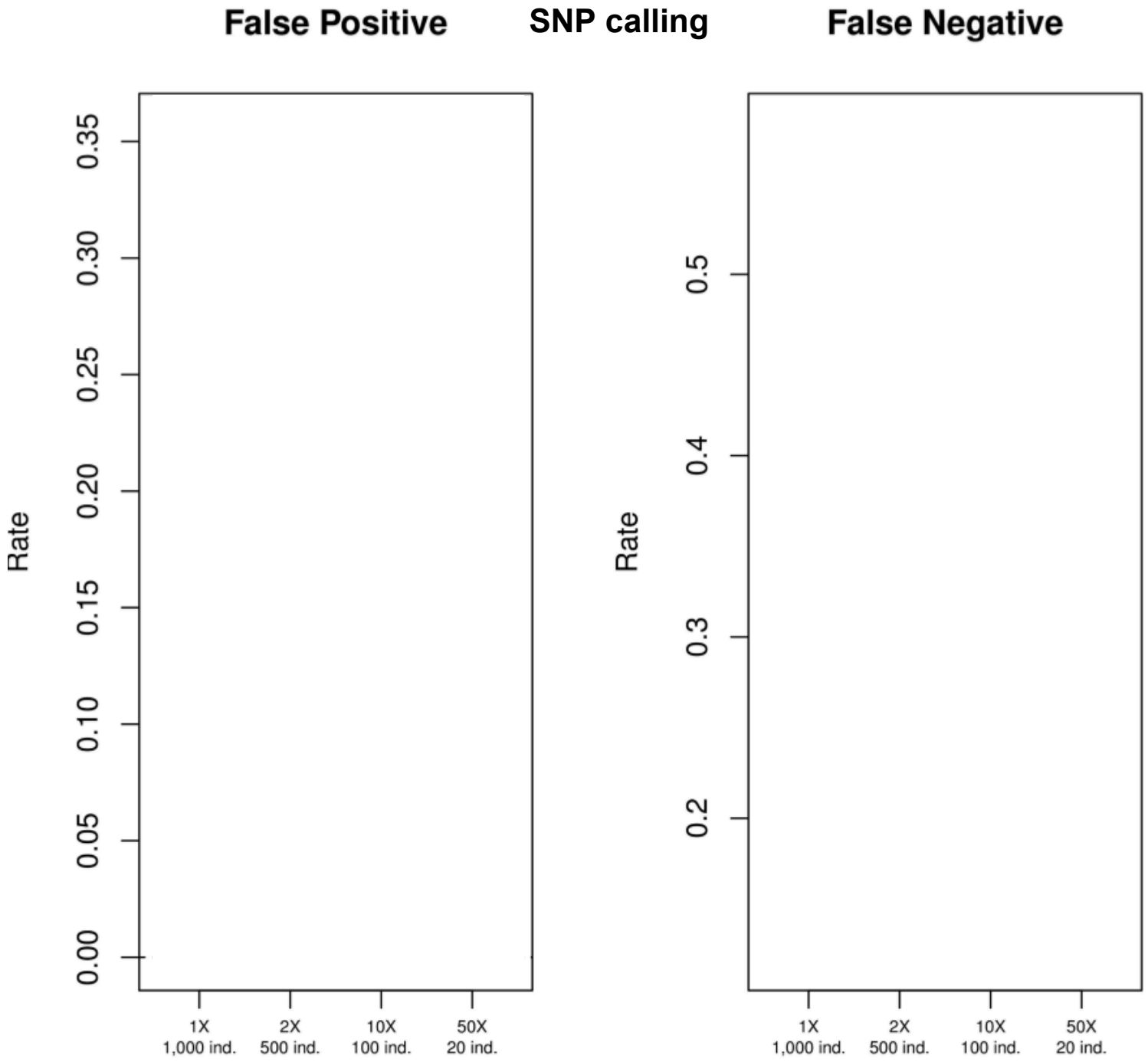
Assess the bias in the estimation → $Bias(S) = \frac{\hat{S} - S}{S}$

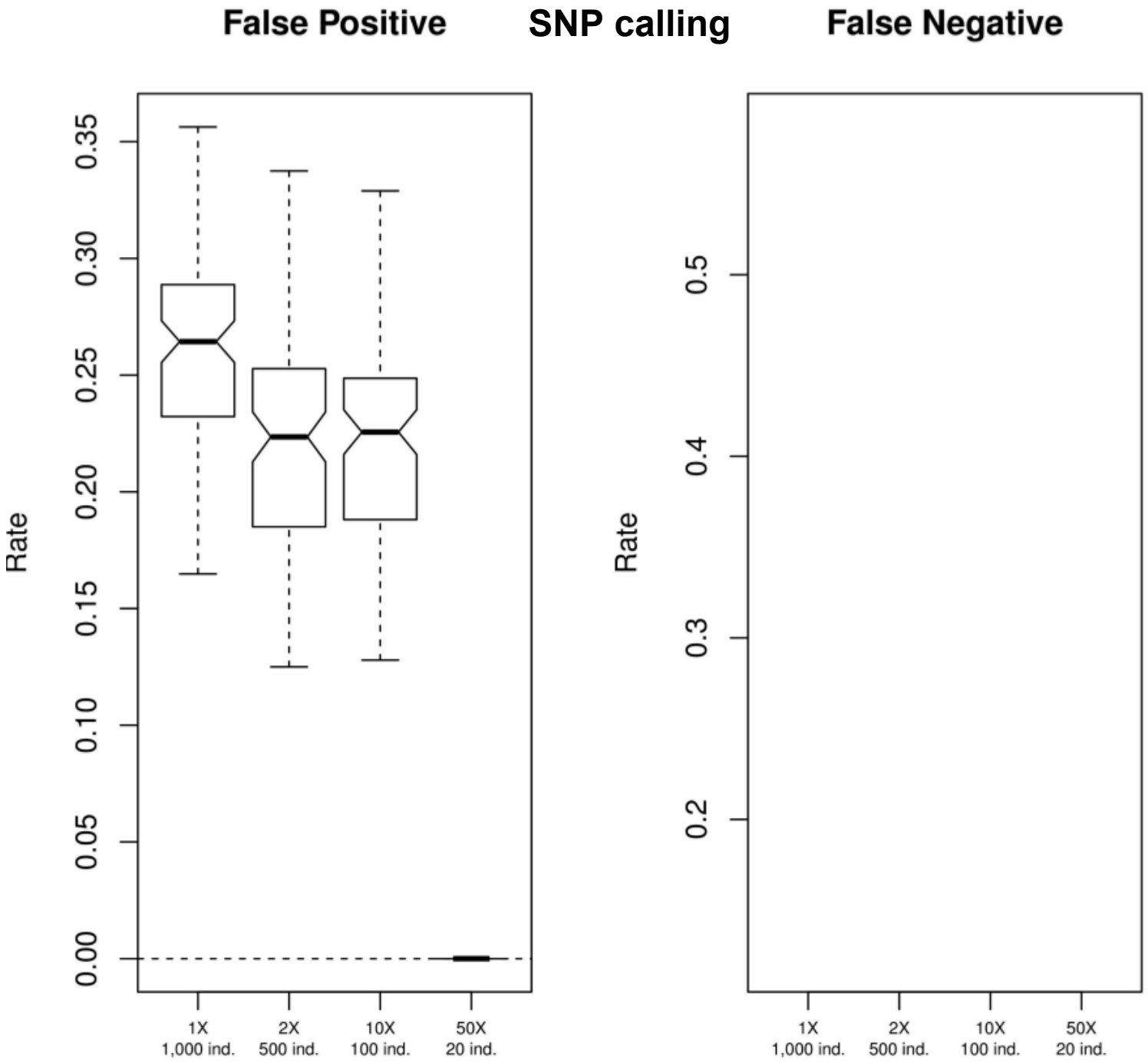

Number of segregating sites
Standardized bias
0.0
−0.2
−0.4
−0.6
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

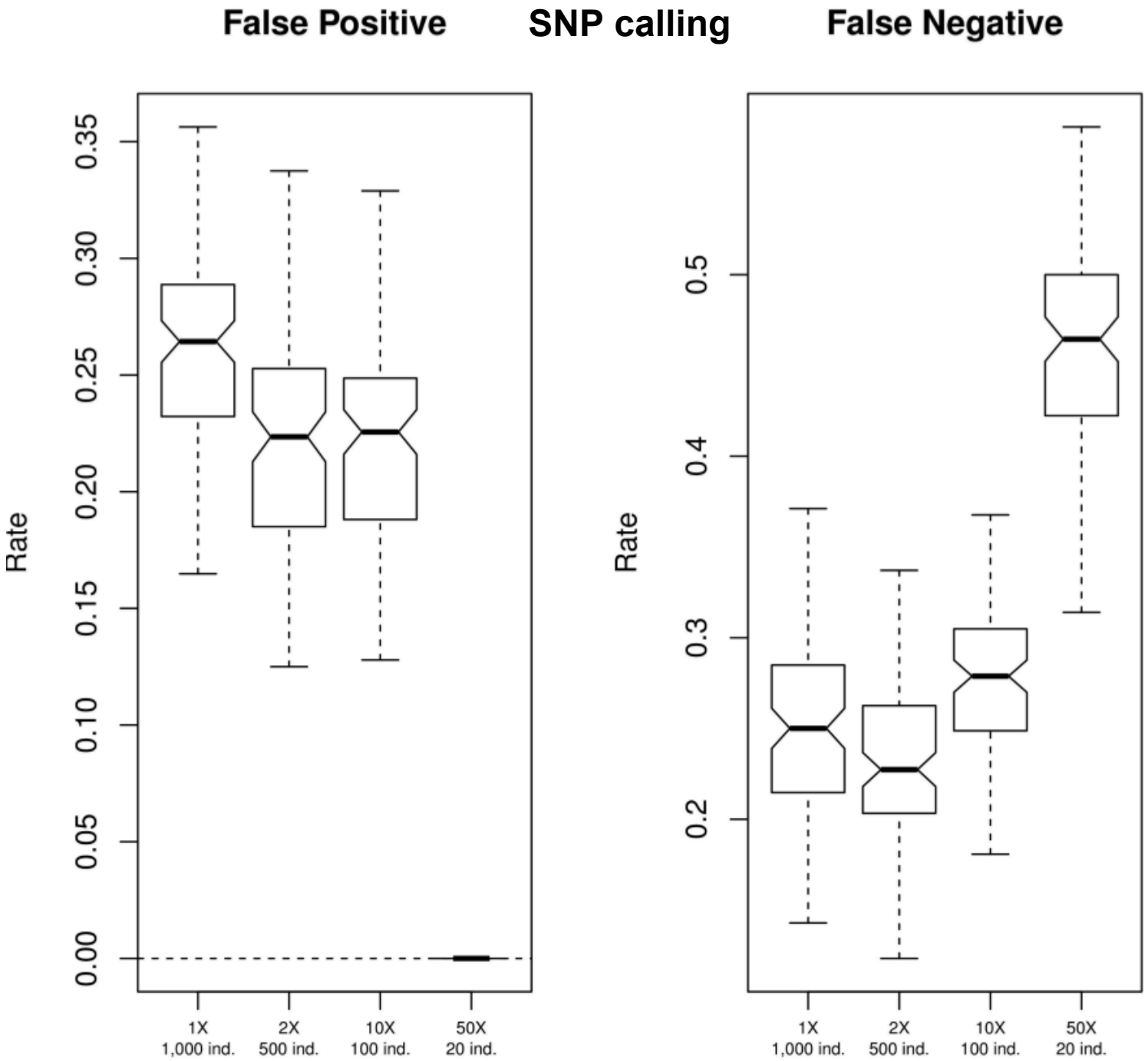

Number of segregating sites
Standardized bias
0.0
−0.2
−0.4
−0.6
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

Expected heterozygosity
Standardized bias
0.05
0.00
−0.05
−0.10
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.


Expected heterozygosity
Standardized bias
0.05
0.00
−0.05
−0.10
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

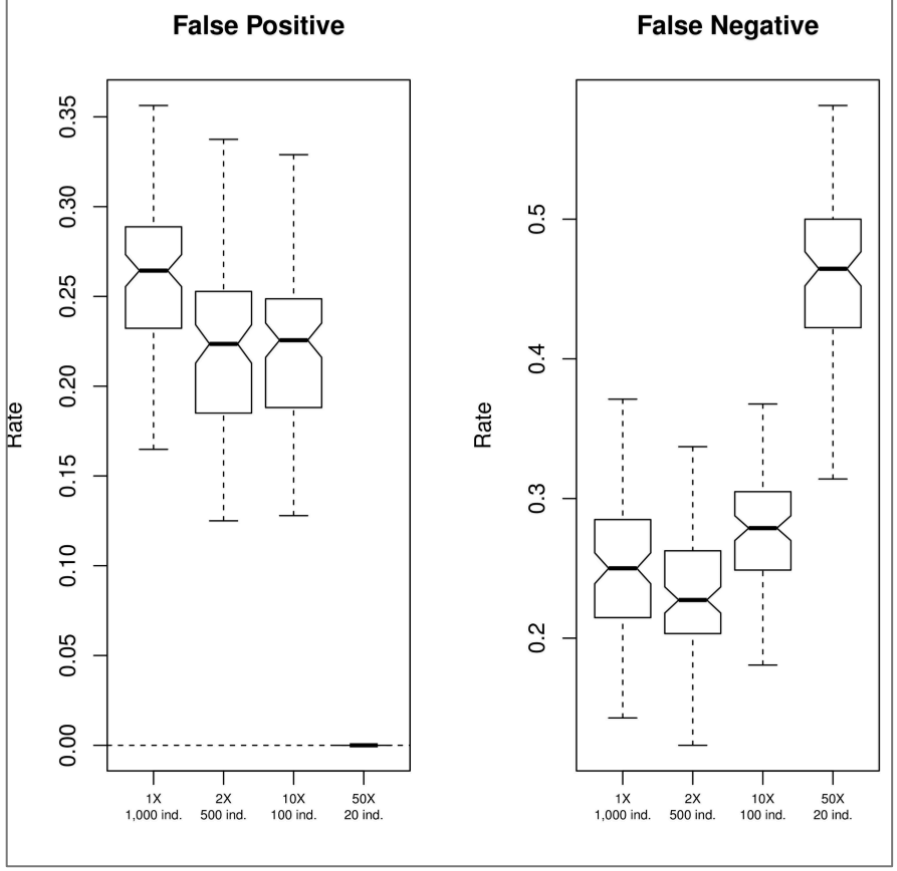

SNP calling
False Positive
Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.
False Negative
Rate
0.2
0.3
0.4
0.5
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.


False Positive
SNP calling
False Negative
Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.
Rate
0.2
0.3
0.4
0.5
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

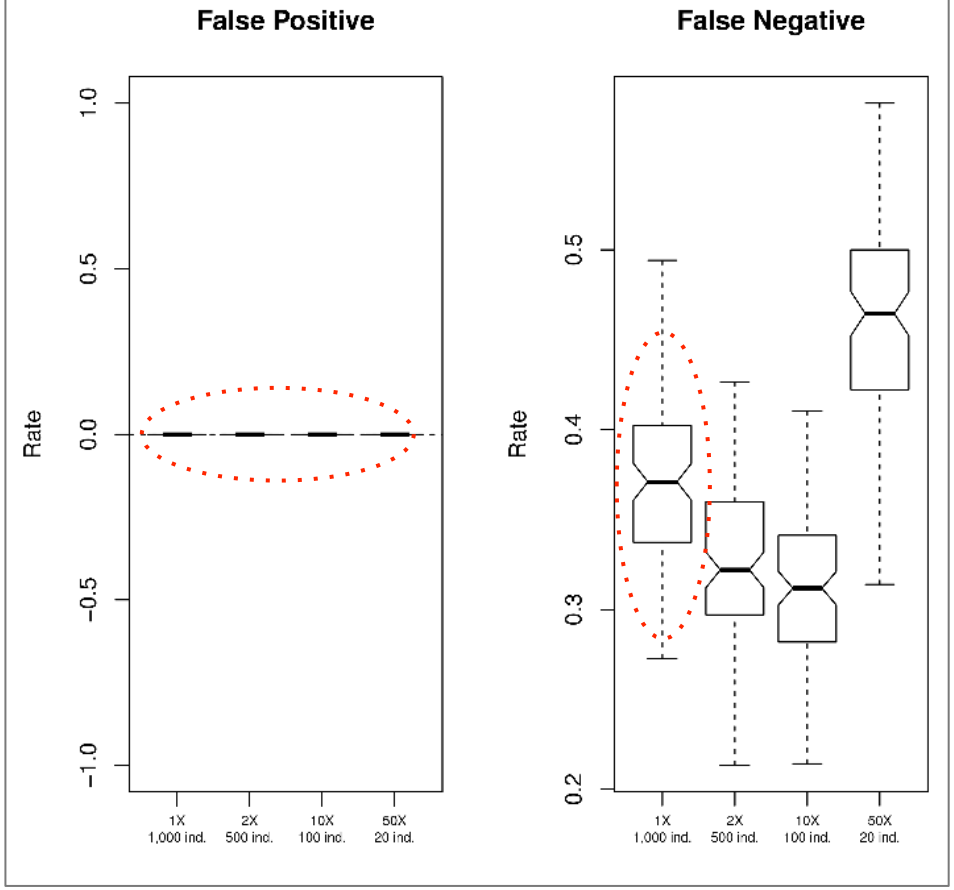

False Positive
SNP calling
False Negative
Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.
Rate
0.2
0.3
0.4
0.5
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

Question for discussion - 1

SNP is assigned if allele frequency is > 1/(2N)

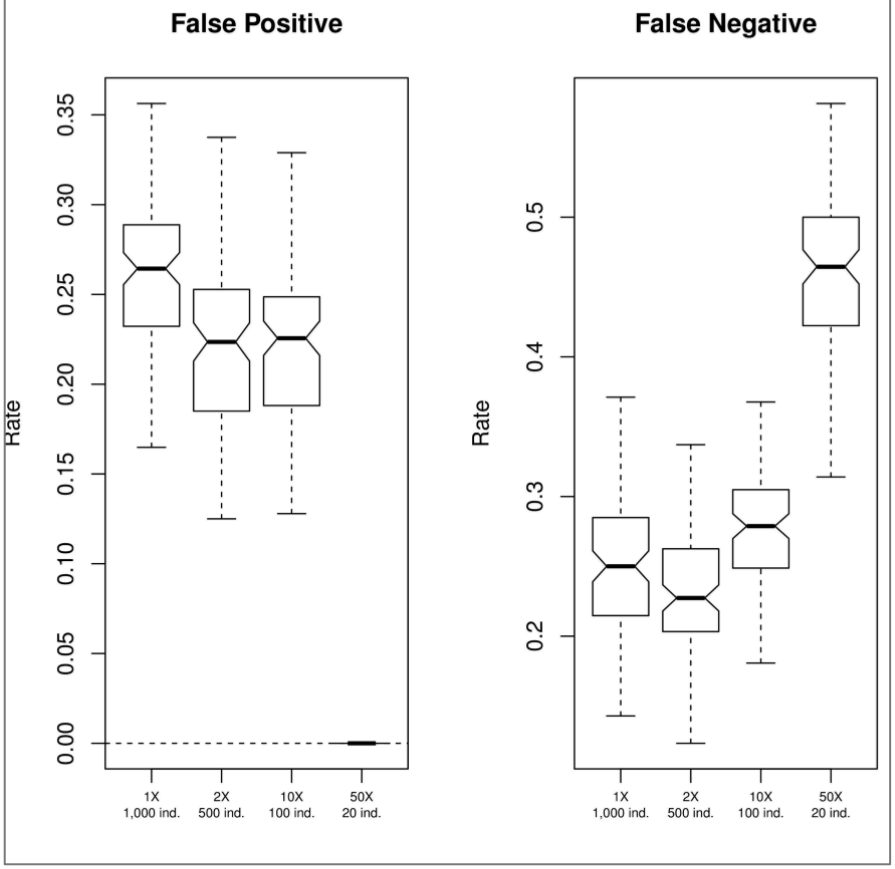


SNP is assigned if ?

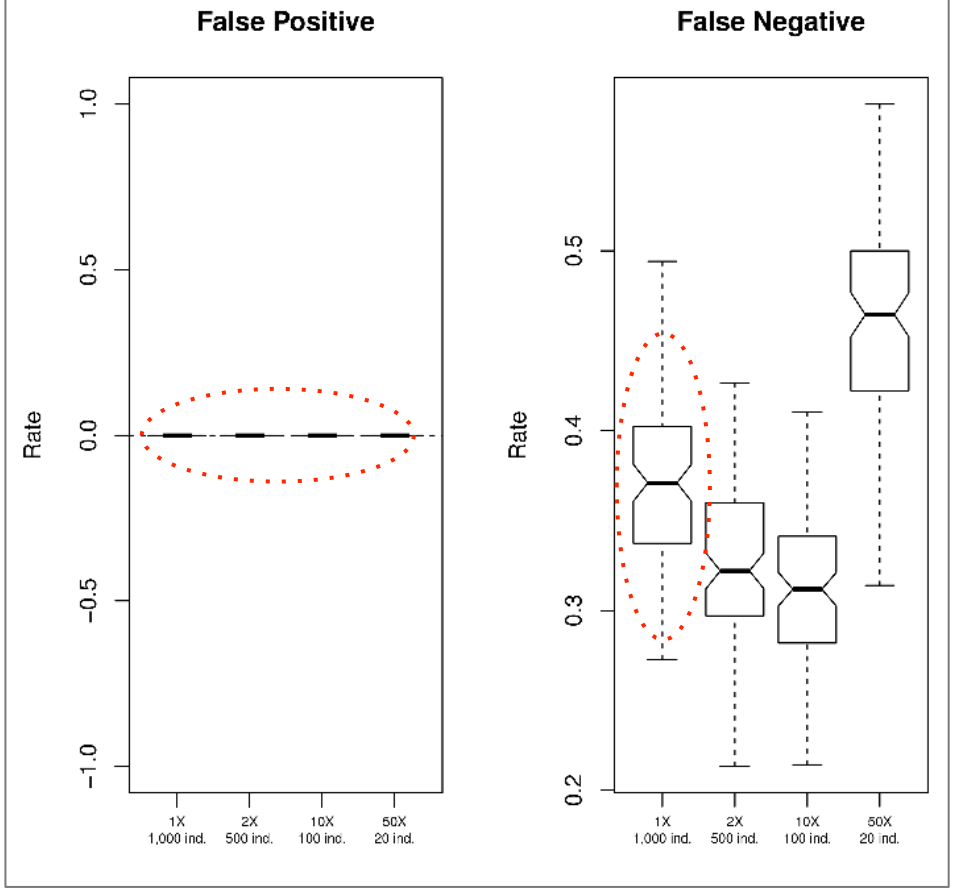


Question for discussion - 1

SNP is assigned if allele frequency is > 1/(2N)



SNP is assigned if the probability of being variable is > 0.95

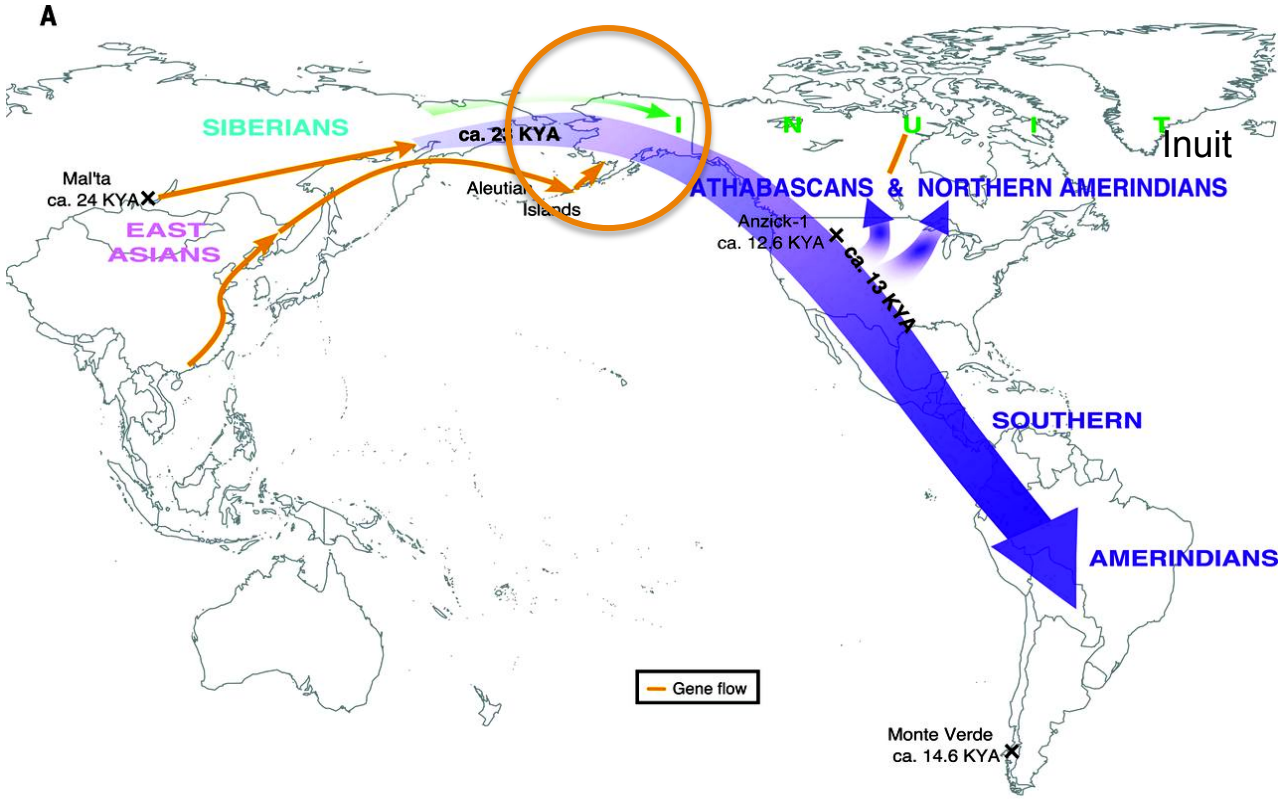


Conclusions

- The results suggest that at a fixed sequencing budget, it is desirable to sequence **a large number of individuals**, at the cost of reducing the per-sample sequencing depth.
- To estimate allele frequencies and identify polymorphic sites, sequencing the largest possible sample size with at least a per-sample sequencing depth of 2X is recommended.
- **State-of-the-art statistical methods** to estimate genetic variation from NGS data should be adopted in all population genetics studies using low-medium coverage sequencing data.

Practical

- Basic filtering
- Estimation of allele frequencies and SNP calling
- Genotype calling
- Advanced methods to estimate SFS
- Exercise: identification of allele frequency differentation between, with admixture assessment and quantification, from low-depth data: the case of FADS genetic variation in Native Americans



Raghavan et al. 2015 Science