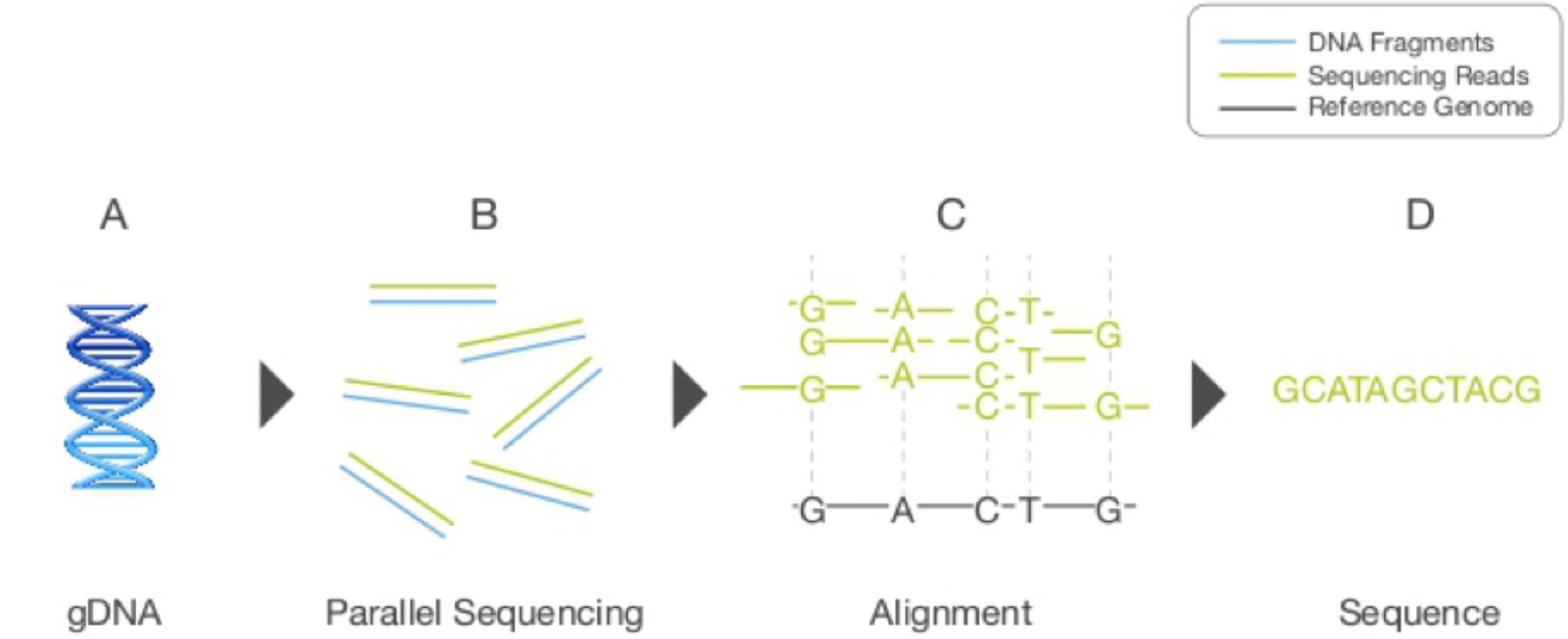
Evolutionary genomics
Data analysis module - Day 1

From raw NGS data to genotypes

April 13th 2015

Next-Generation Sequencing

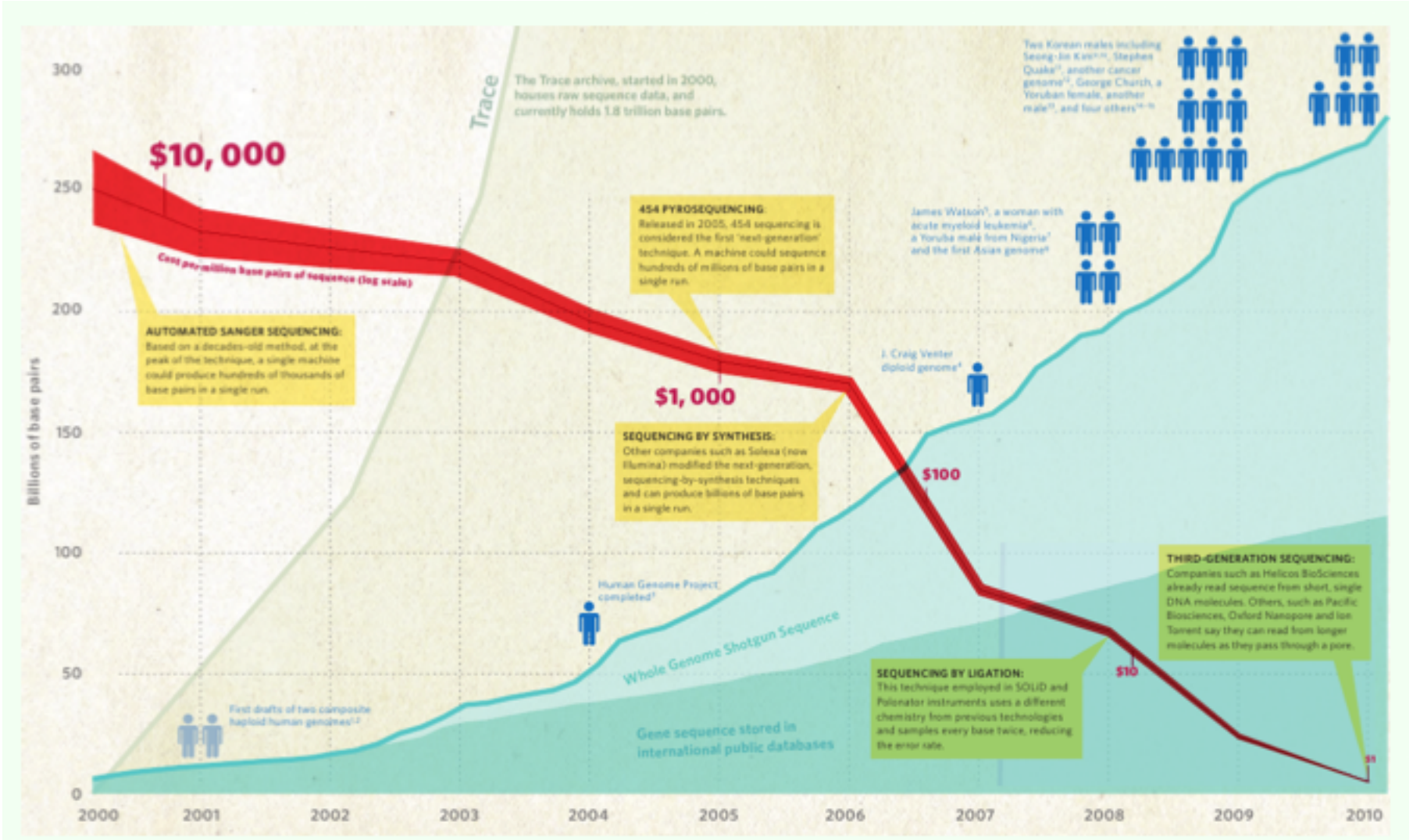


A. Extracted gDNA
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome
D. The whole-genome sequence is derived from the consensus of aligned reads.

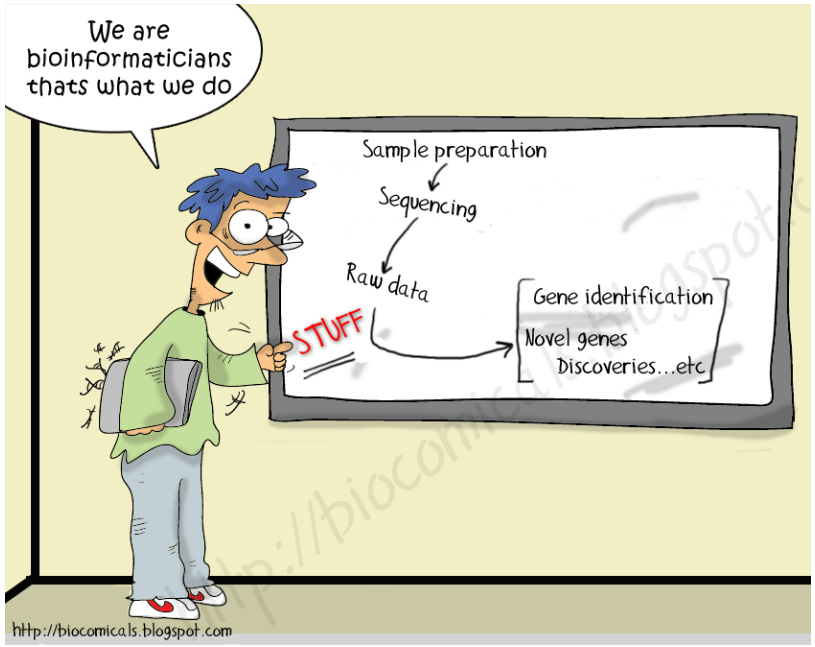
Different platforms

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

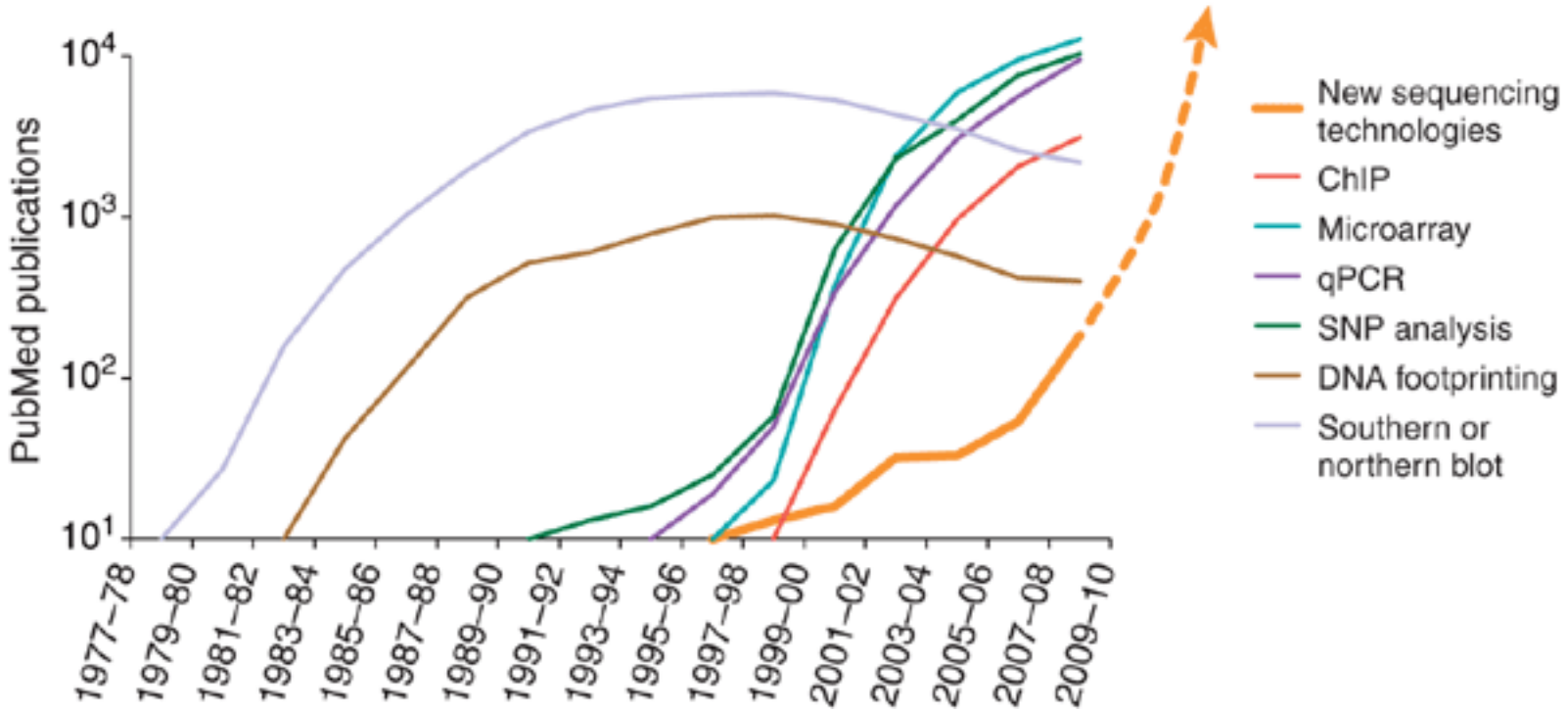

We are bioinformaticians thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc
http://biocomicals.blogspot.com

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^ecgfhfhhiiiiiiiihiiiibgfhdfhhiihhhhhhiiafgihaegbfhigbb_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

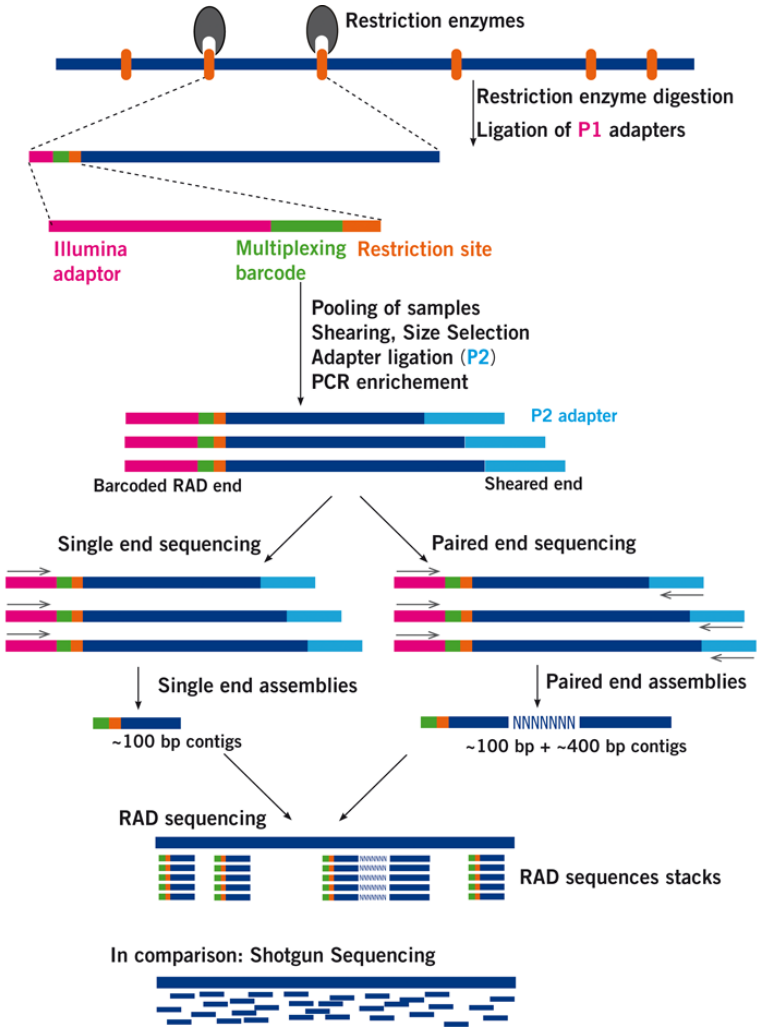
whole genome

ancient genomes

Exome capturing

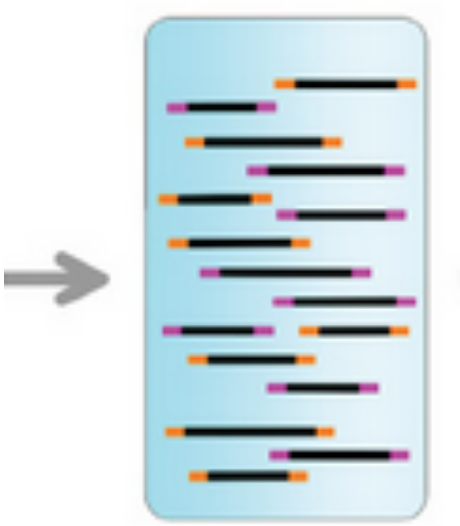
	FS5524895	FS5102038	FS5103086	FS5522194	Any 3 of 4 FS5524895 FS5102038 FS5103086 FS5522194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,766
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP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

RAD-sequencing



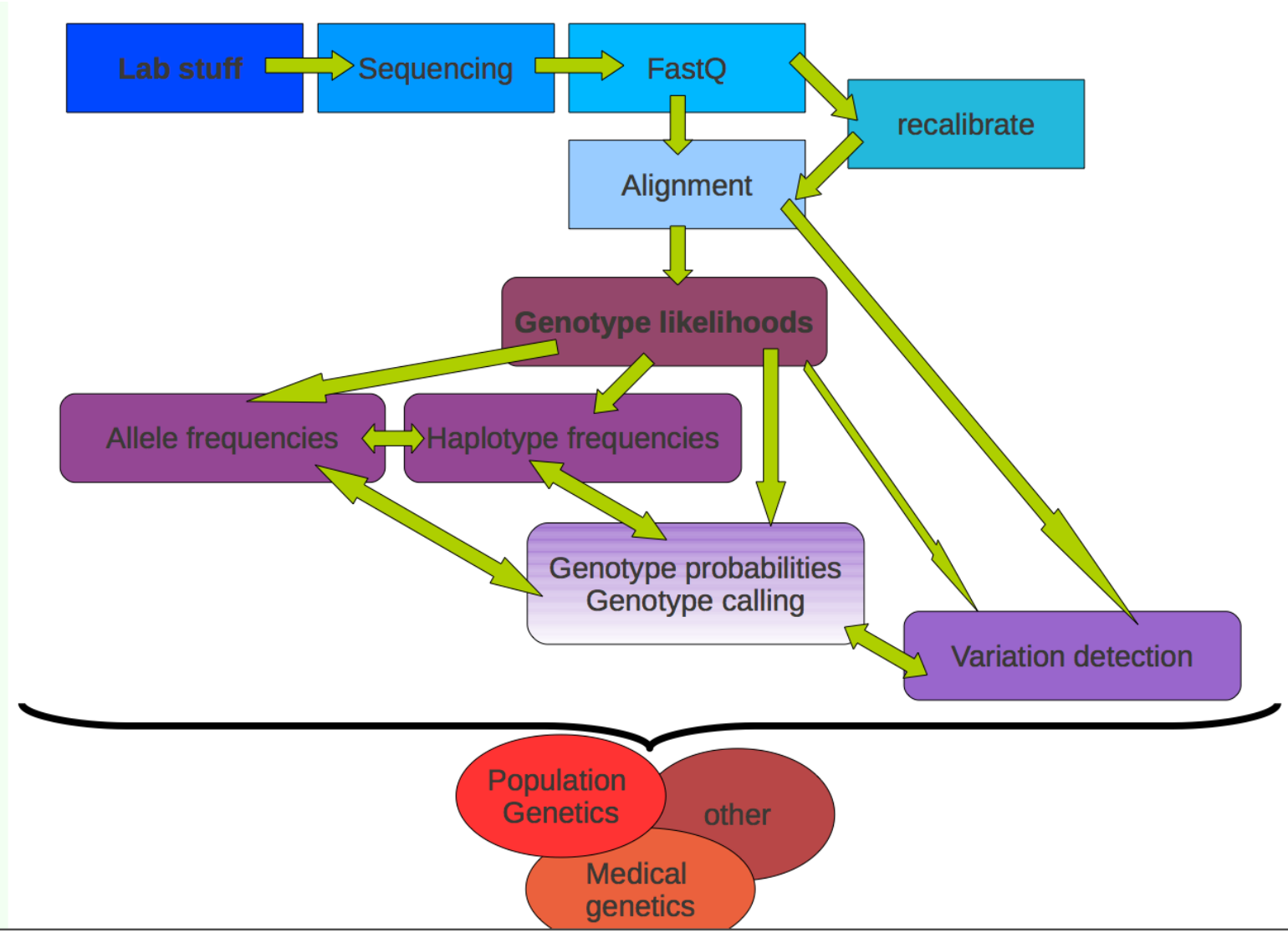
http://www.floragenex.com/

Pooled sequencing

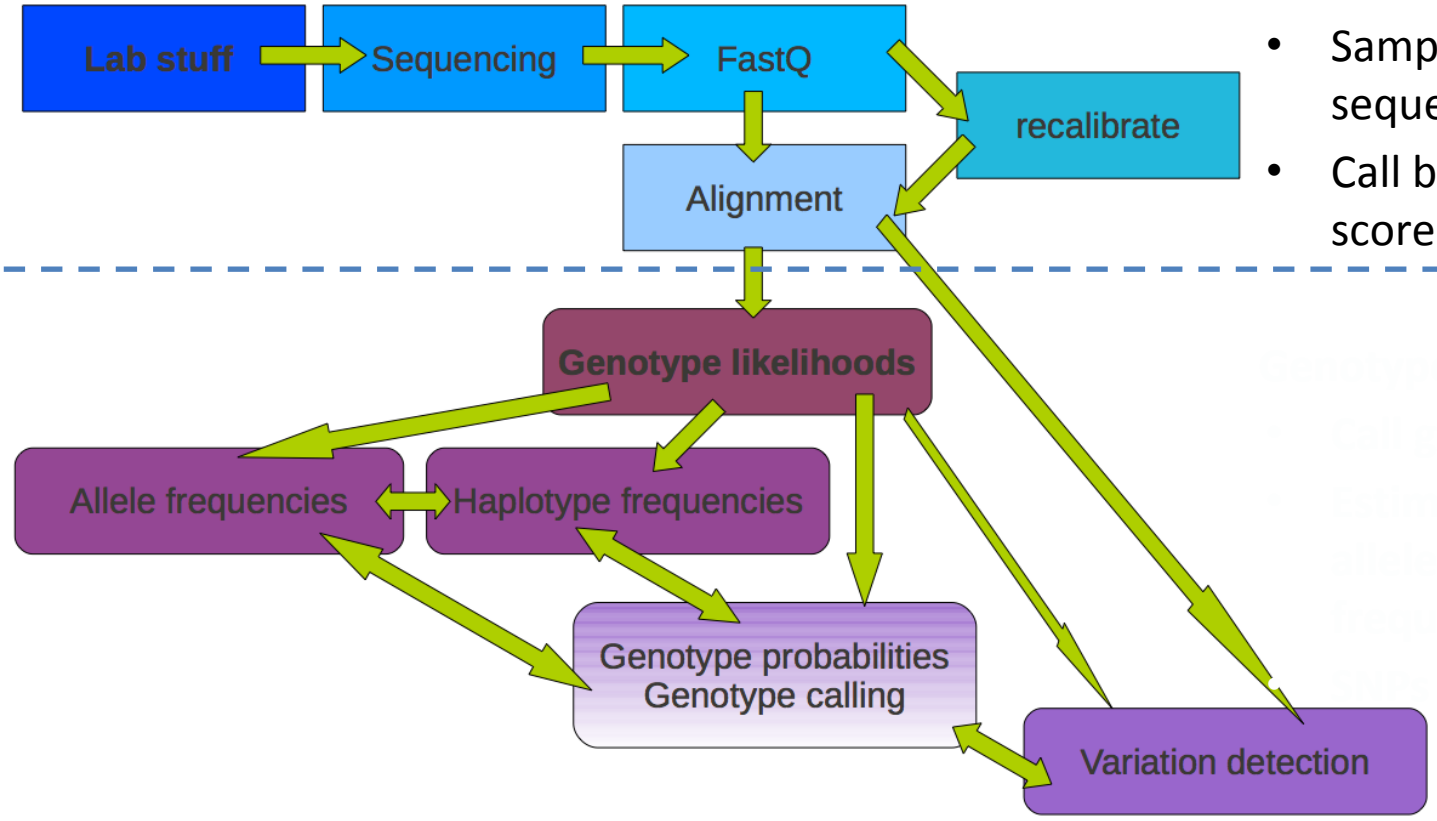


...

Workflow

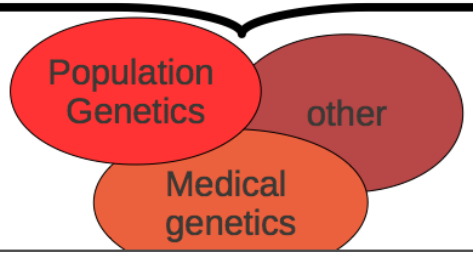

Lab stuff
Sequencing
FastQ
recalibrate
Alignment
Genotype likelihoods
Allele frequencies
Haplotype frequencies
Genotype probabilities
Genotype calling
Variation detection
Population Genetics
other
Medical genetics

Workflow

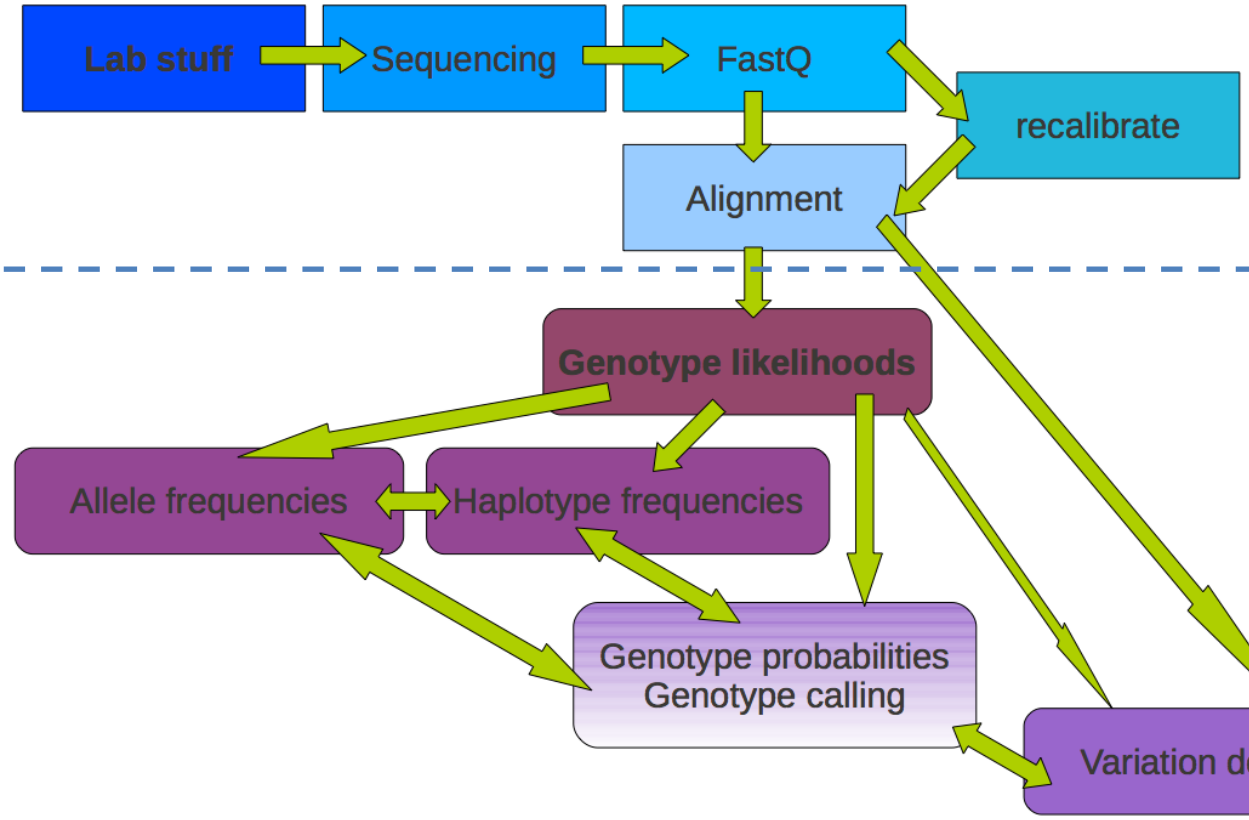


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores



Workflow

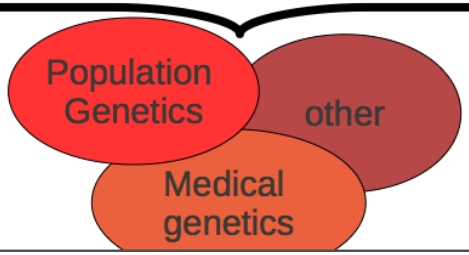


Low-level data:

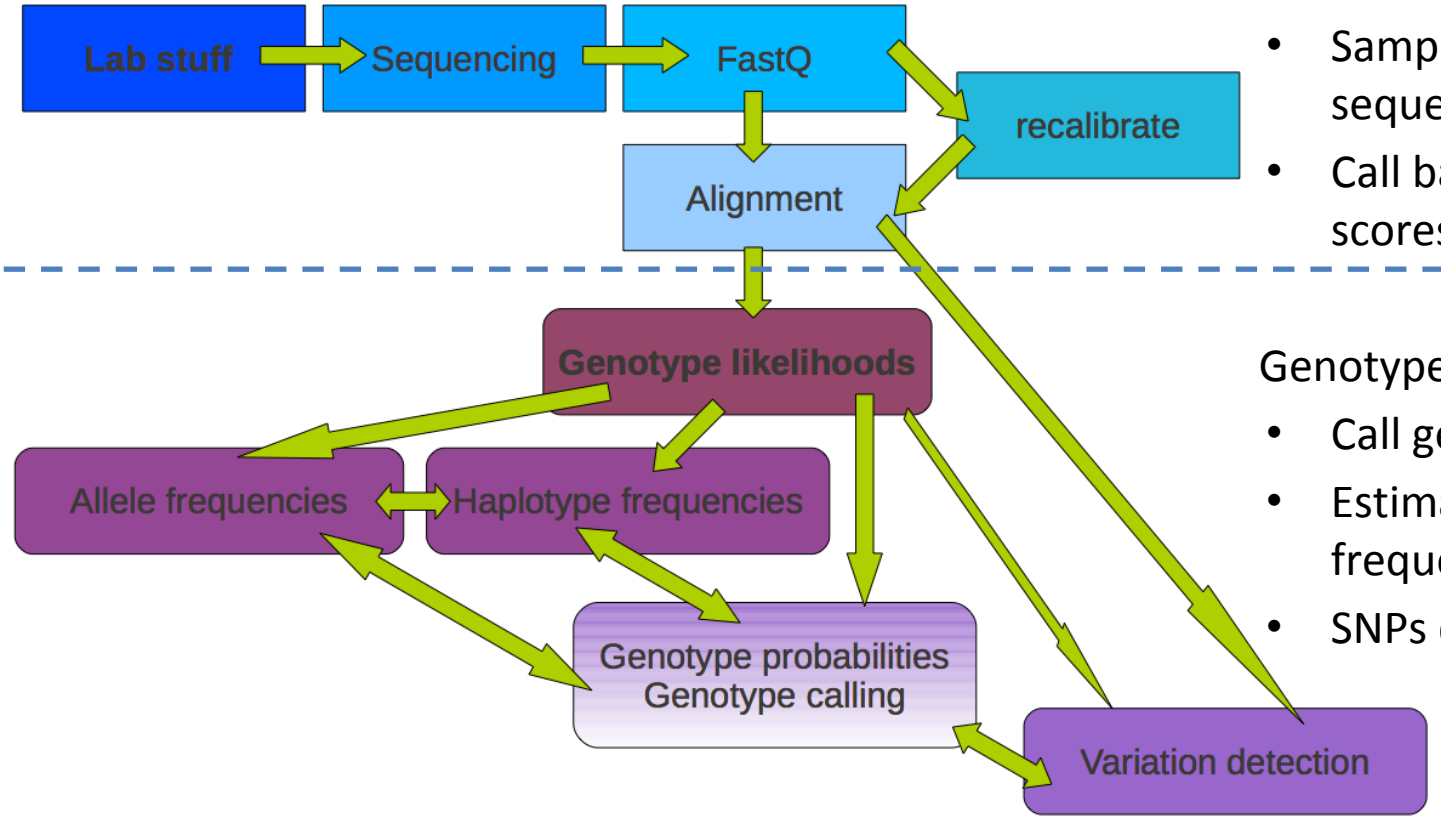
- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

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



Workflow

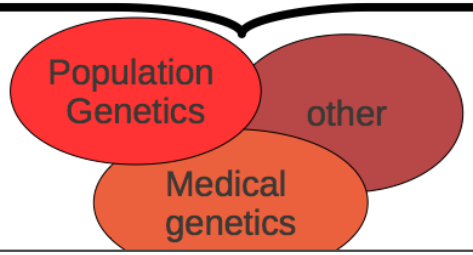


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

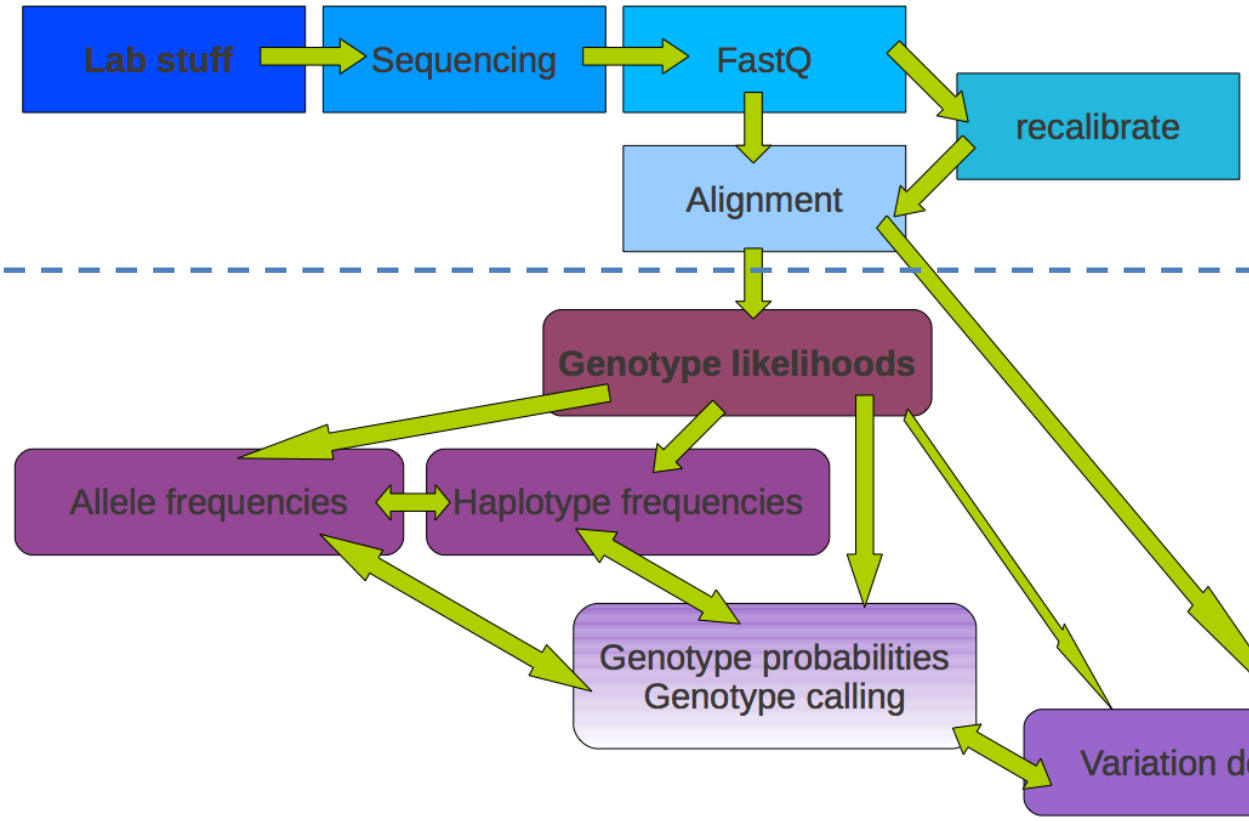
- Call genotypes
- Estimate allele/haplotype frequencies
- SNPs detection



Analysis:

- Population genetics analysis
- Association studies

Workflow

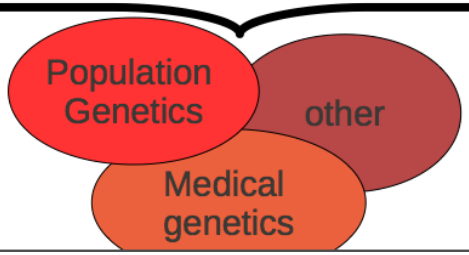


Low-level data:

- **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
```

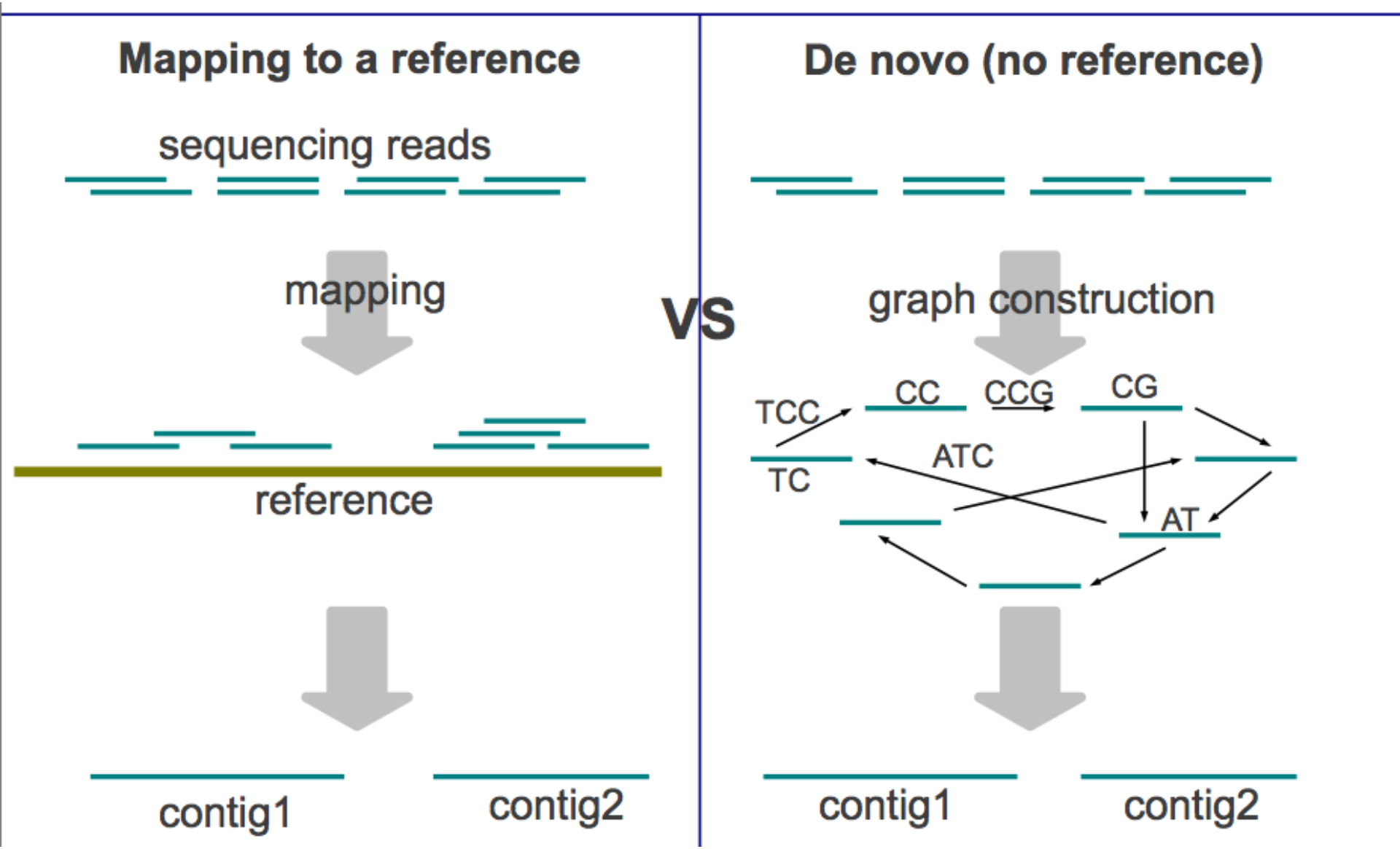
Quality scores

Qscore

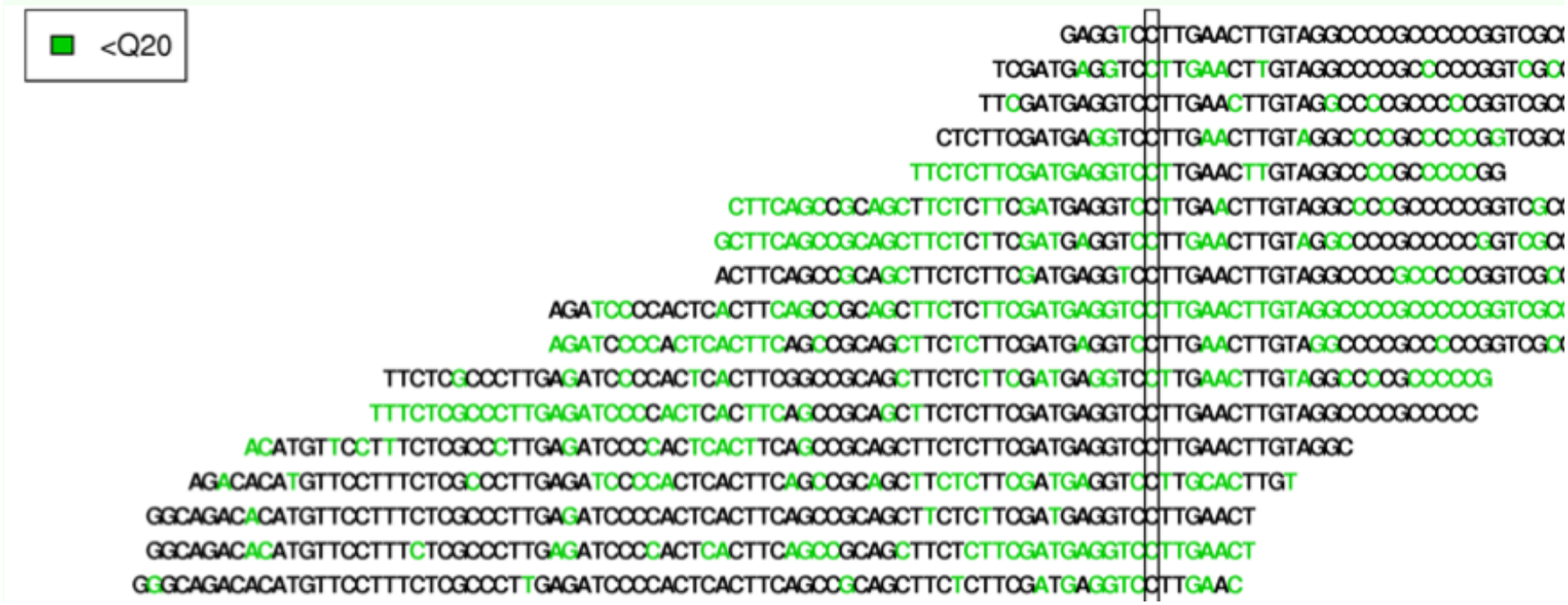
- The ASCII values can be interpreted as a probability
- A Q20 (ASCII 'T') score is probability of 1%
- The score is the probability, P, that the base is incorrect
- $$Q_{score} = -10log_{10}(P)$$
- $$P = 10^{\frac{-Q}{10}}$$

!"#$%&'()*+,-./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

BAM/SAM file

HS1:109:C01CCABXX:1:2101:18857:56640: 99 Contig287516 168 150 87M2H = 168 88
GGCATTTTCACCTTGGGCATCTTCAGGTGCCAGTCTGGGCCATGAACATCCACATCTGGGGCACTGATGTCTA
CTTTAGGGCCTTTG
CCCFFFFFHHHHHJJJJJJJJJJJJJJHHIIJJHHJJJJJJJJJJJJJIJJHIIHIJJJJJJJJIJIIJJHHHFHHHHFEFFFDDEED
PG:Z:novoalign AS:i:31 UQ:i:31 NM:i:1 MD:Z:59A27 PQ:i:61 SM:i:150 AM:i:150

HS1:109:C01CCABXX:1:2101:18857:56640: 147 Contig287516 168 150 89M = 168 -88
GGCATTTTCACCTTGGGCATCTTCAGGTGCCAGTCTGGGCCATGAACATCCACATCTGGGGCACTGATGTCTA
CTTTAGGGCCTTTGAG
DDDDDEECA;DFFFFHHHGHHJIJJJJJJJJIJIIJJJJJJJJIJIIGGIIHJJIJJJJJJGJJJJJJJJJJJJJJHHHHHFFFFFCCC
PG:Z:novoalign AS:i:30 UQ:i:30 NM:i:1 MD:Z:59A29 PQ:i:61 SM:i:150 AM:i:150

HS1:109:C01CCABXX:1:1104:16690:184446:73 Contig287516 170 150 99M1H = 170 0
CATTTTCACCTTGGGCATCTTCAGGTGCCAGTCTGGGCCATGAACATCCACATCTGGAGCATCAATGTCCACT
TTAGGGCCTTTGATATCAACATCAGG
BCCFFFFFHHHHHJJJIJJJJJJJJHJJJIJIJJJJJJJJJJJIJIJJJJJJJJIJJGJJJJJIIJJJJJJJJJGEHHHCFFFCEEDEFDED
DDDDDDDPG:Z:novoalign AS:i:240 UQ:i:240 NM:i:8 MD:Z:61C0T0G5T16G0G2T2T5

HS1:109:C01CCABXX:1:1201:4604:153988: 99 Contig287516 184 150 90M = 184 89
GCATCTTCAGGTGCCAGTCTGGGCCATGAACATCCACATCTGGGGCACTGATGTCTACTTTAGGGCCTTTGAG
GTCTACTTCAGGGCCTT
CCCFFFFFHHHFHIIHJJHIJJJIJJJJJJJJJJIJIJJJJJJJJJIJJJJIJJJIJJJJIIIJIHHIIHDHH??DFFDFEEEE@BBBC
PG:Z:novoalign AS:i:30 UQ:i:30 NM:i:1 MD:Z:43A46 PQ:i:61 SM:i:150 AM:i:150

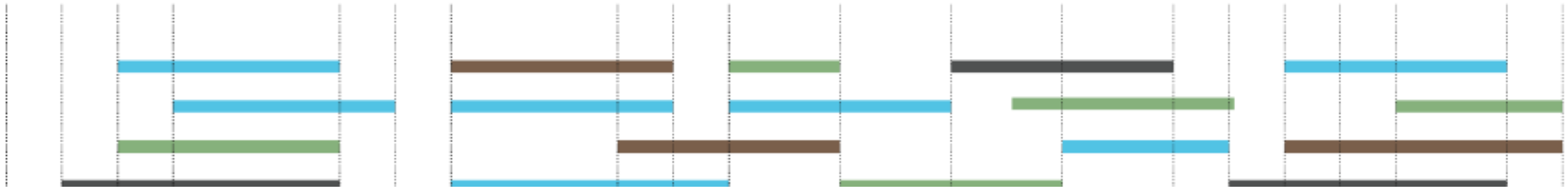
HS1:109:C01CCABXX:1:1201:4604:153988: 147 Contig287516 184 150 90M = 184 -89
GCATCTTCAGGTGCCAGTCTGGGCCATGAACATCCACATCTGGGGCACTGATGTCTACTTTAGGGCCTTTGAG
GTCTACTTCAGGGCCTT
DDDDDDEDEDEEFFB@HGHHJIHBJIJIHEJJJJIHJJJIIJJJIIIJJIJJJJJGHJJIJJJJJJJJJJJJJJIIHHHGHHFFFFD@C
@ PG:Z:novoalign AS:i:31 UQ:i:31 NM:i:1 MD:Z:43A46 PQ:i:61 SM:i:150 AM:i:150

PILEUP

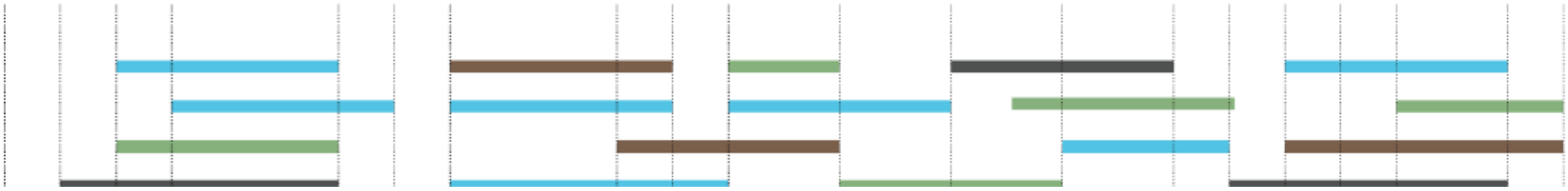
one individual:

Contig86016 522 A 36 ,,,.....,.,..,.,,,.,,....,..,.....,.
FFHJIH>JJG?FDJ-JBIEJHIEJJJIJFDCJJCEF
Contig86016 523 T 36 ,,,.....,.,..,.,,,.,,....,..,.....,.
FFHJIH@JJF>FDJ;JEIAJHJGJJJIIJDHJJFDF
Contig86016 524 C 36 ,,,...A.,.,.$.,.,,,.,,....,..,.....,.
CFFJFH(IJIDCDJ5JHJCJHJHJJJJJJFDJJHEF
Contig62808 313 C 27 .,.,T.,,.,.....,,.,.,.....
BBDI9HIHJJIJJJJJJJIH=JHHHHF
Contig222765 857 G 23 ,.$.a,...A,..,..,,.,,,,.
FAI.CDFJ65DB<HHFHEDDDDJ
Contig409192 281 A 35 .,.,,T.,,,,,.,,,..,,,..,,,.,.,,.,.. JHGJH-
(<IJFJHIJ>@IJCHGEHGFJIJIGFJ@C

Challenges

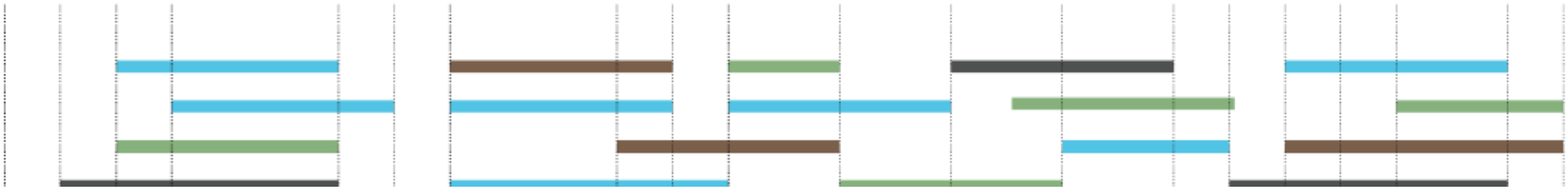


Challenges



- Variable and low depth
- High sequencing and mapping errors

Challenges

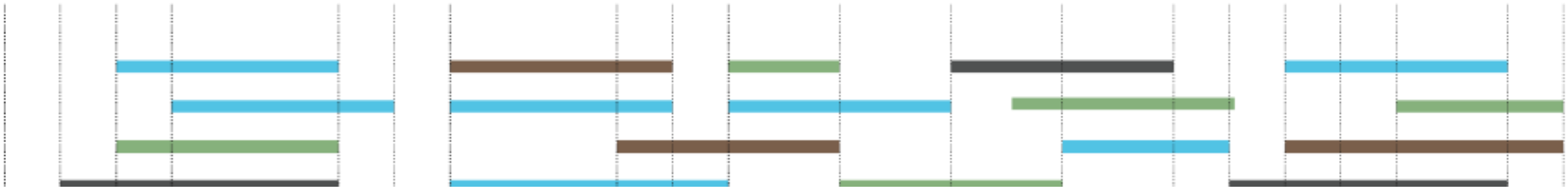


- Variable and low depth

- High sequencing and mapping errors

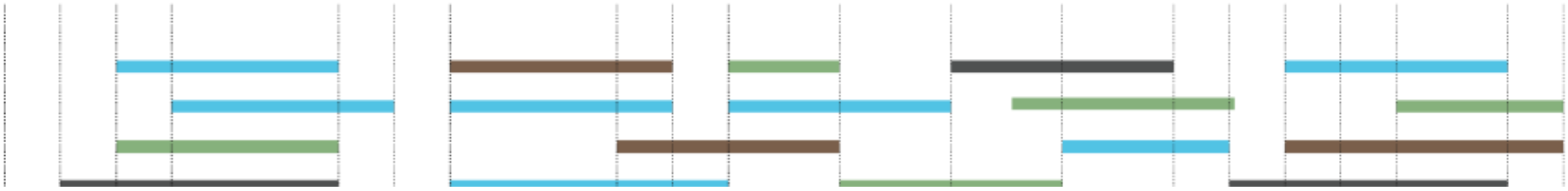
Quality control filters

Data filtering



- Variable and low depth

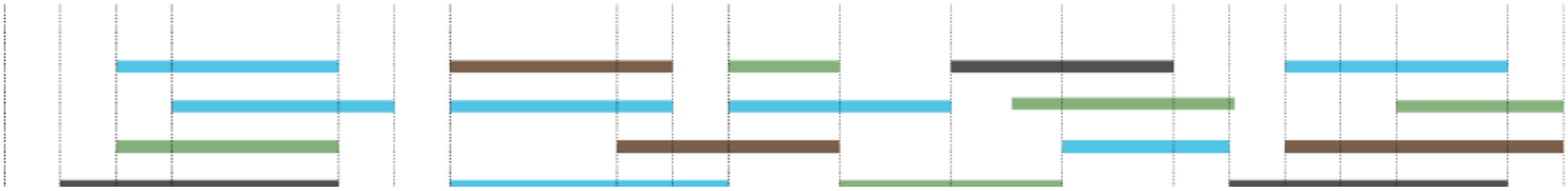
Data filtering



- Variable and low depth

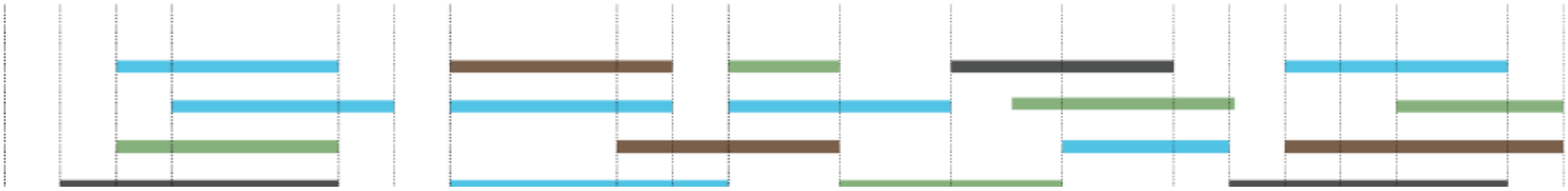
Minimum depth
Maximum depth
Even depth across samples
…

Data filtering



- Sequencing and mapping errors

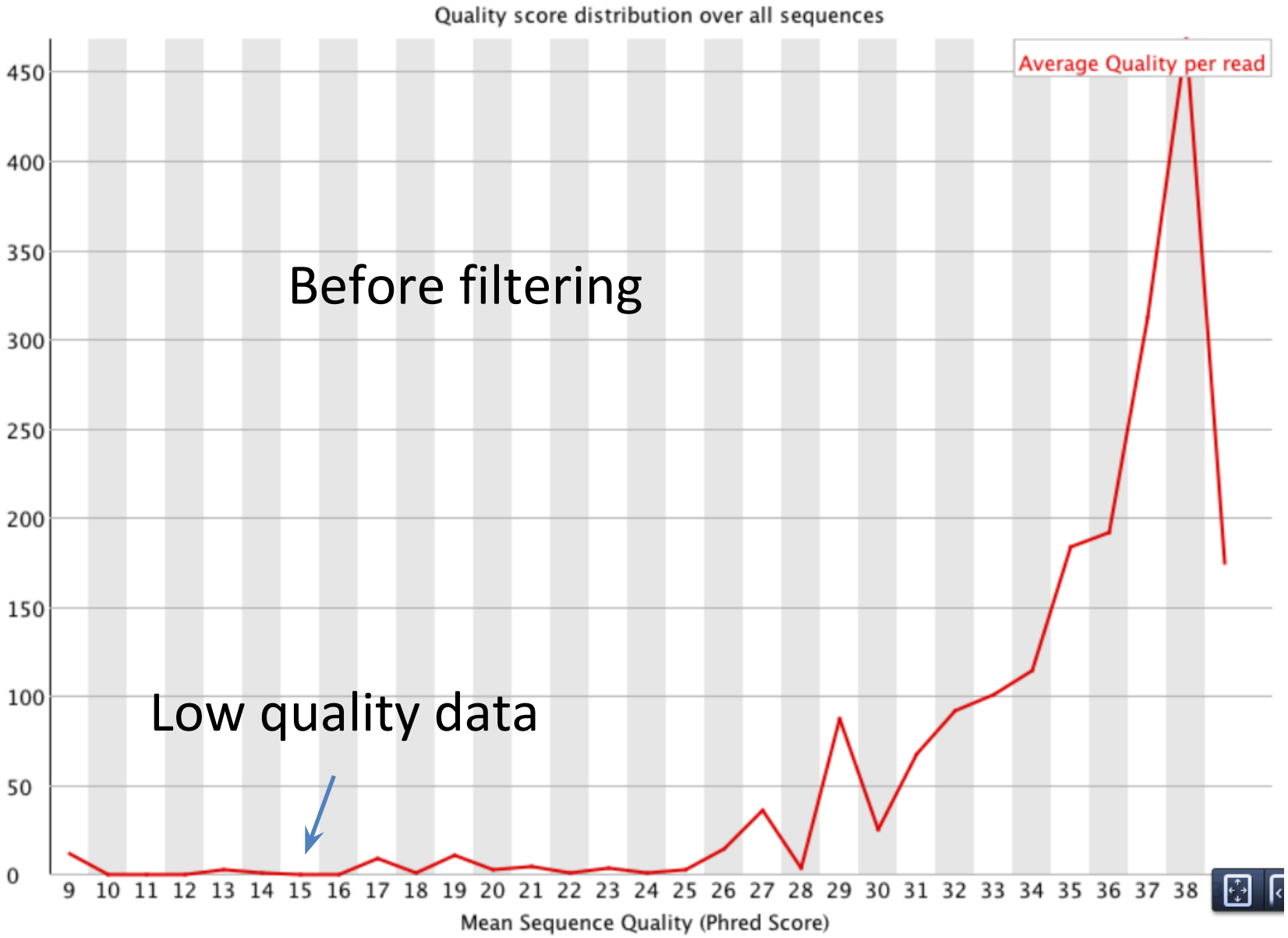
Data filtering



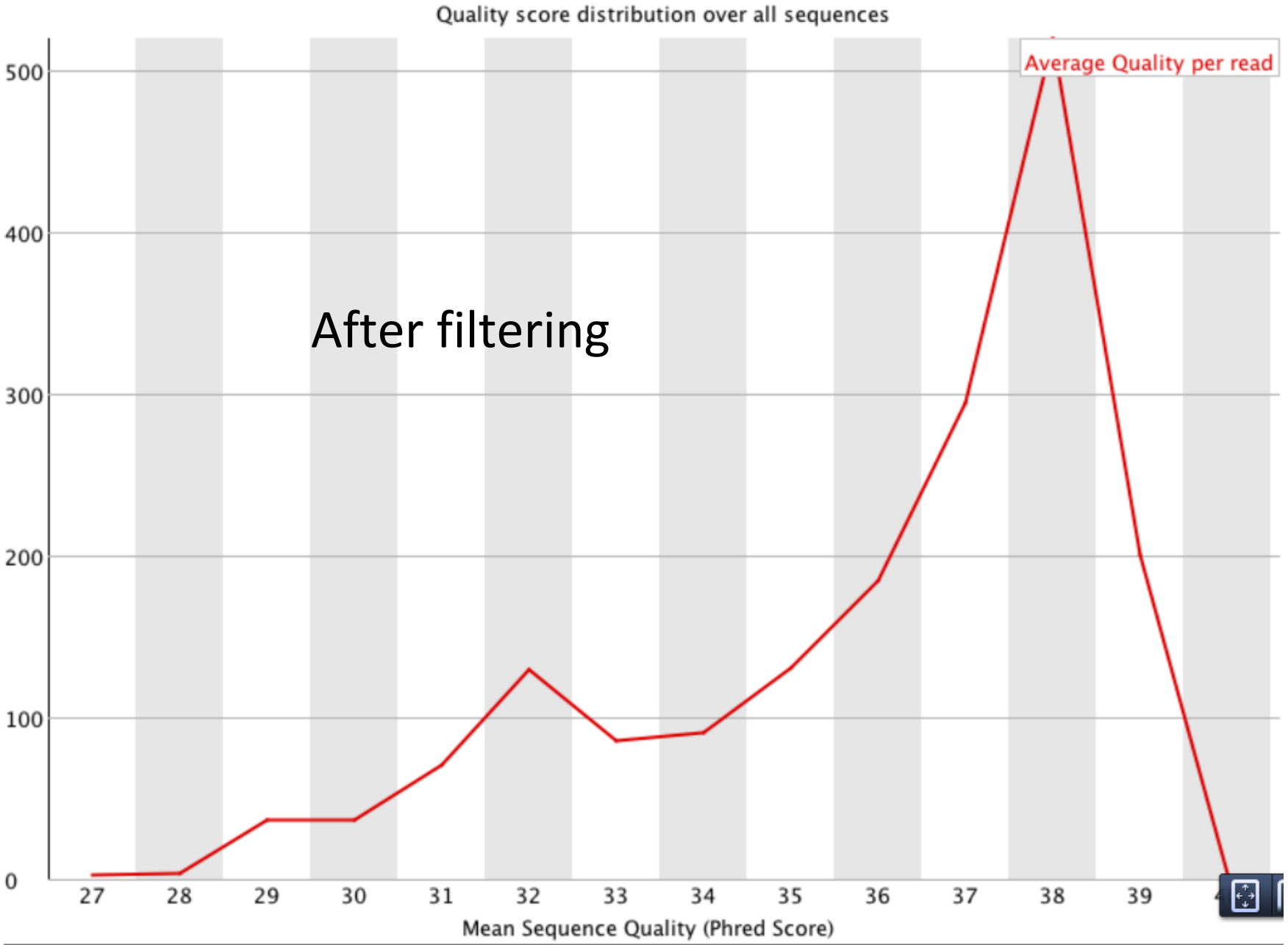
- Sequencing and mapping errors

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

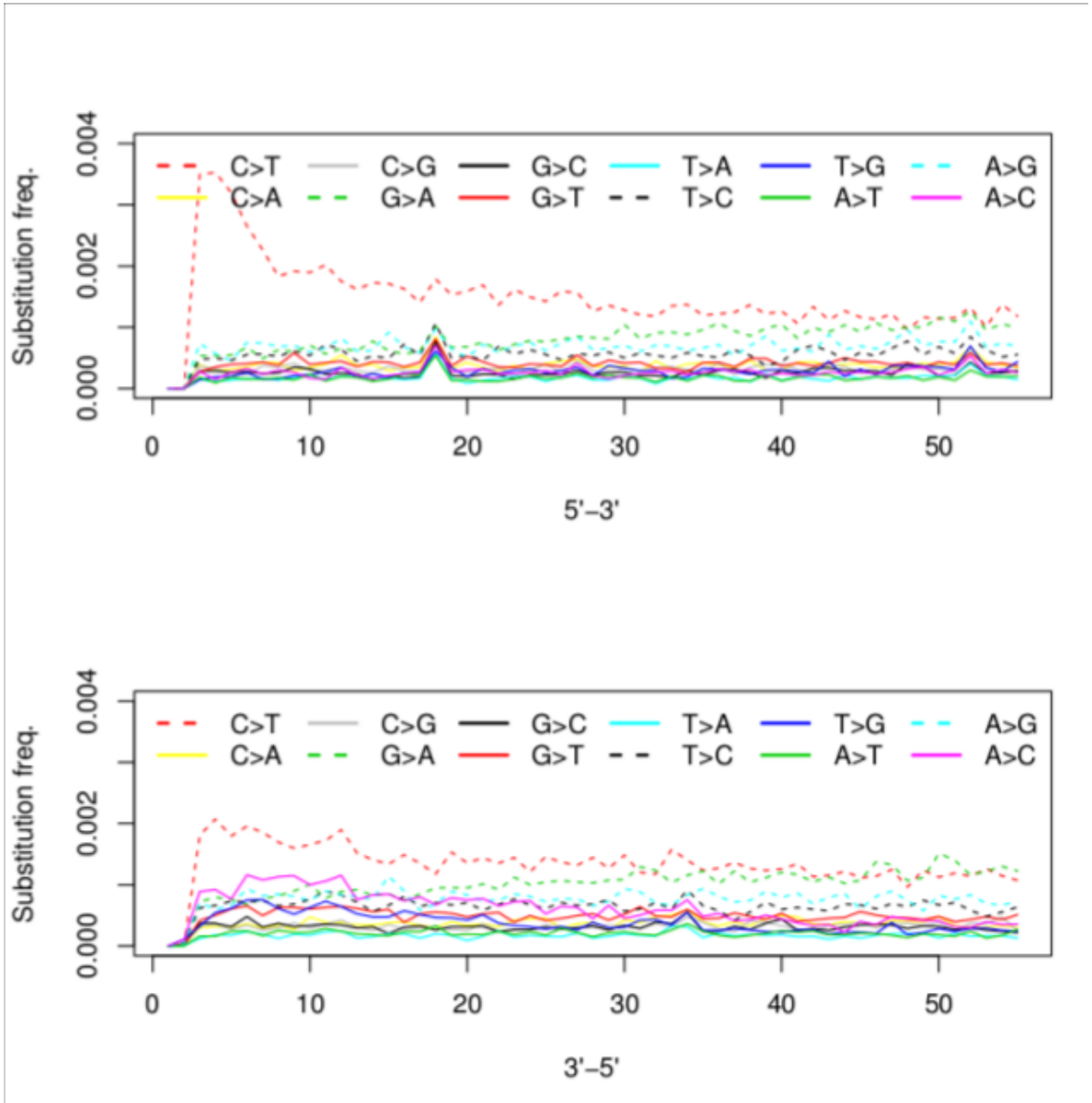
Check your filtering



Check your filtering



Mutation frequency bias



Site Frequency Spectrum (SFS)

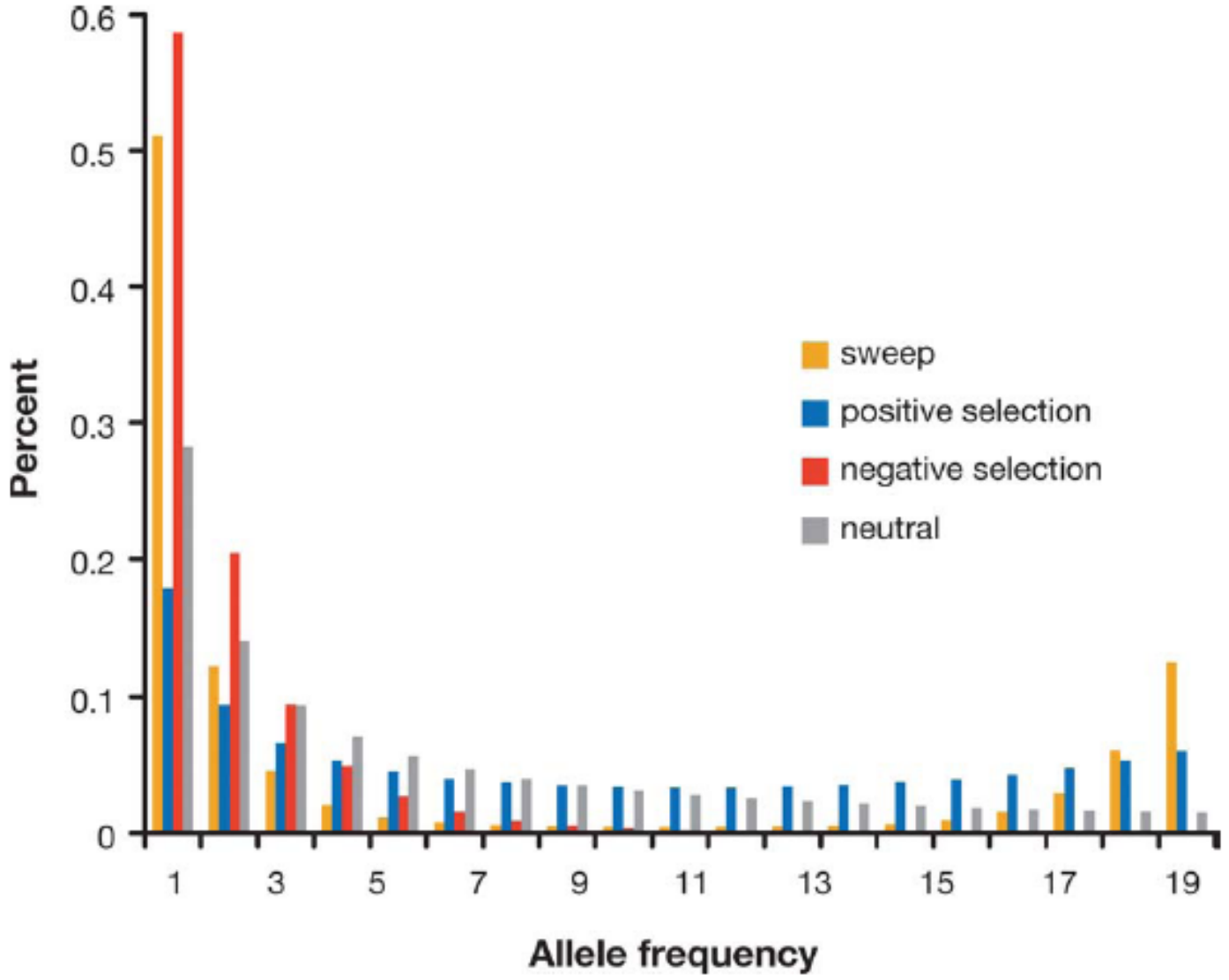
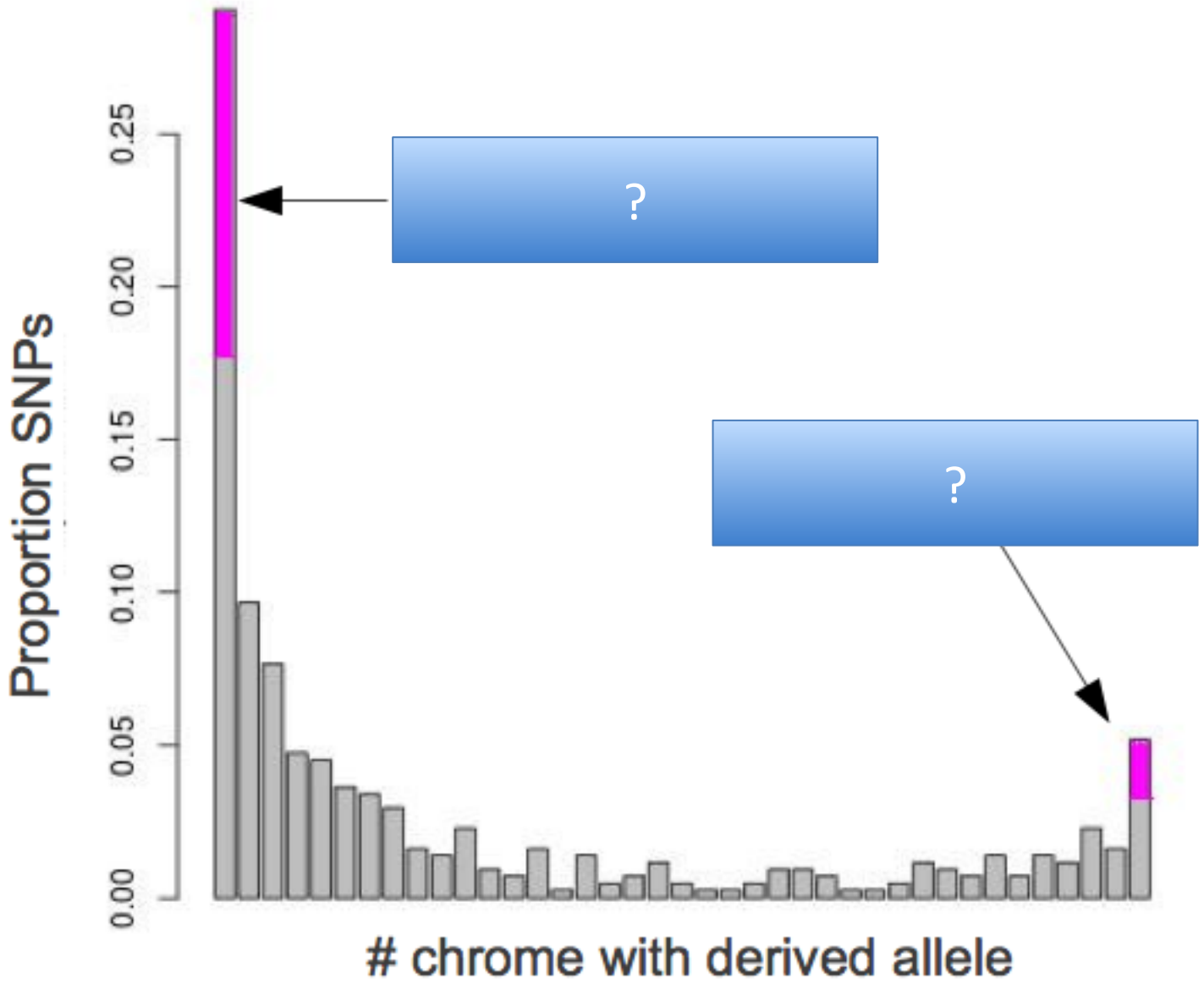
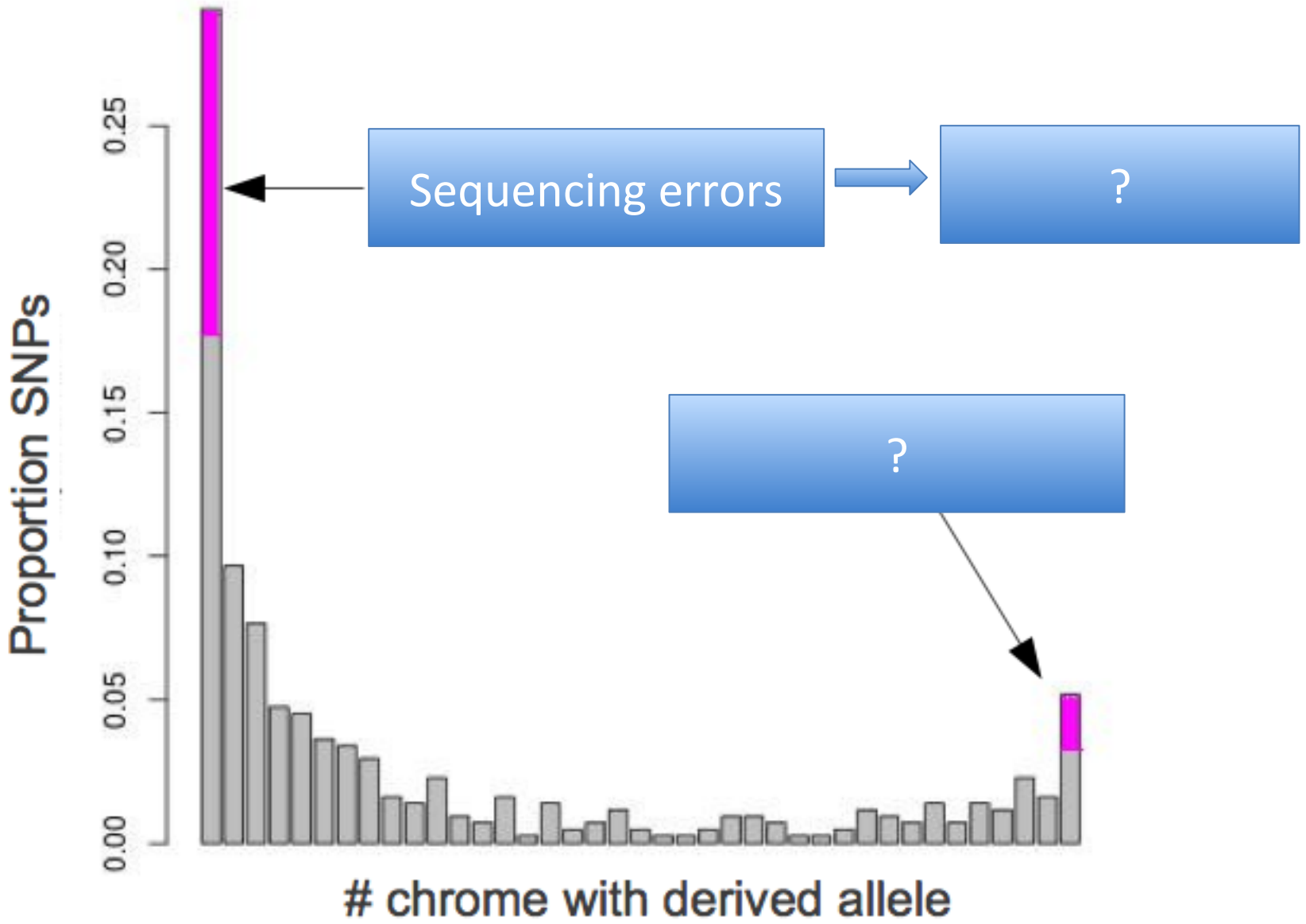


Figure 2

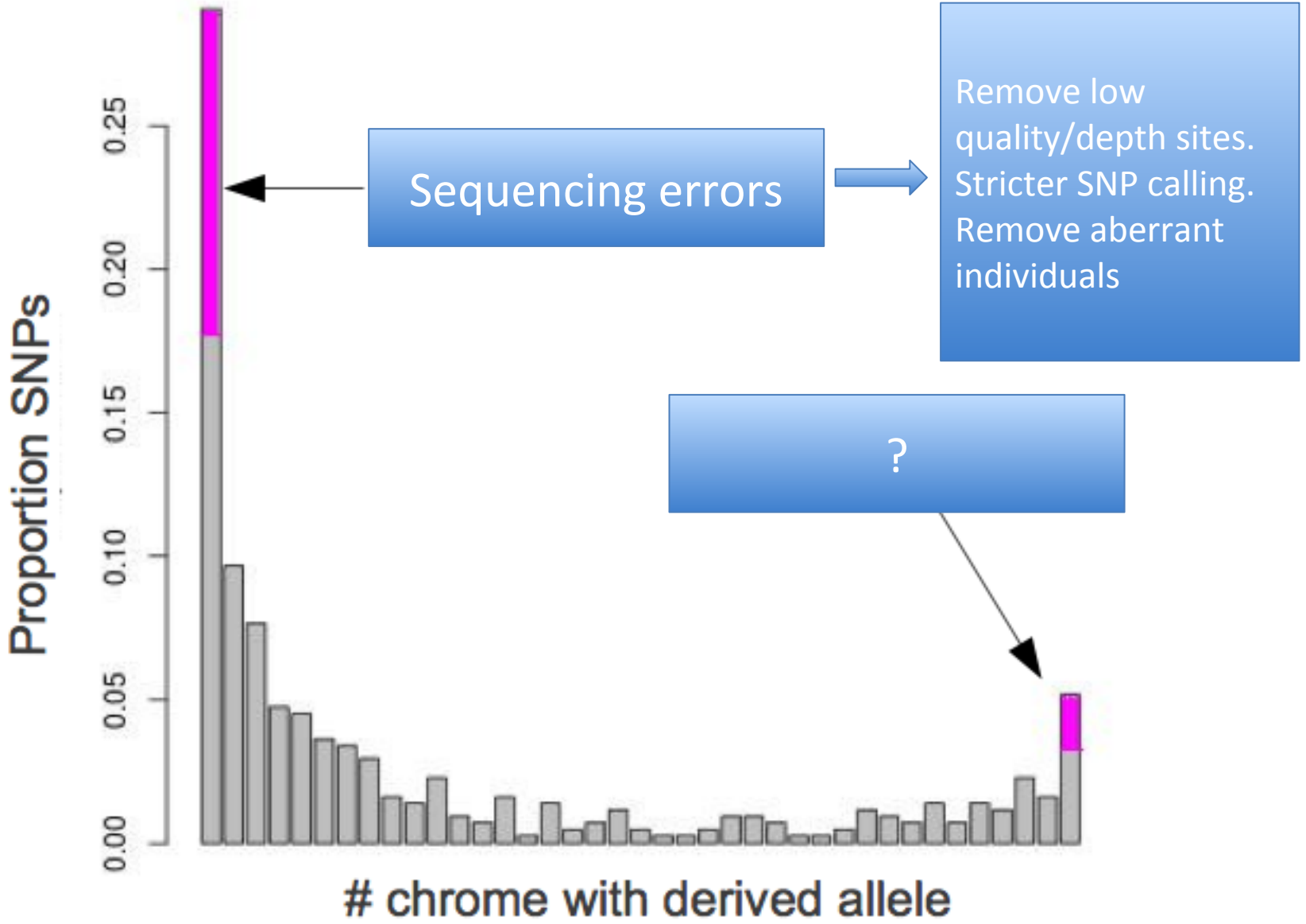
Effect of errors on the SFS



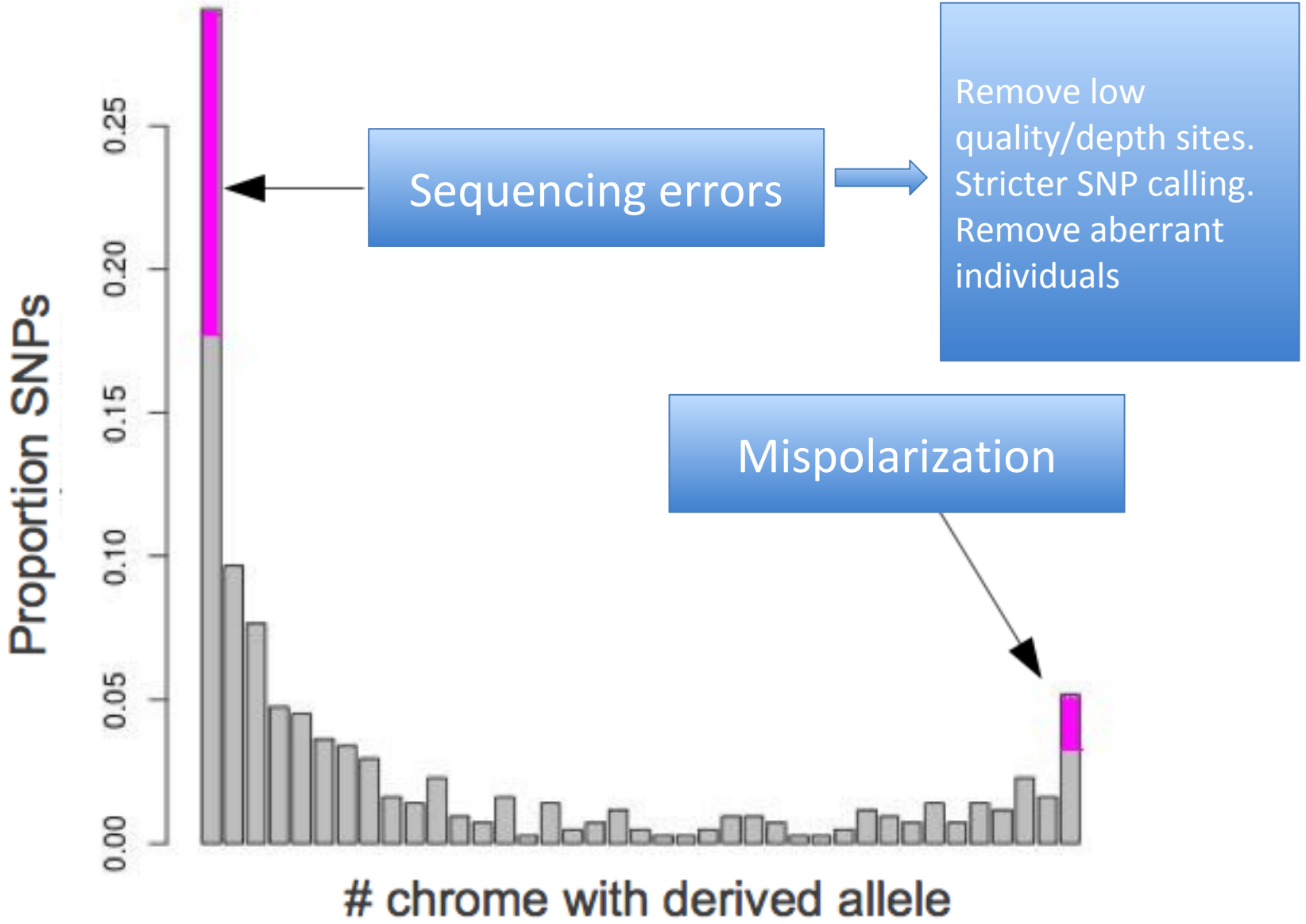
Effect of errors on the SFS



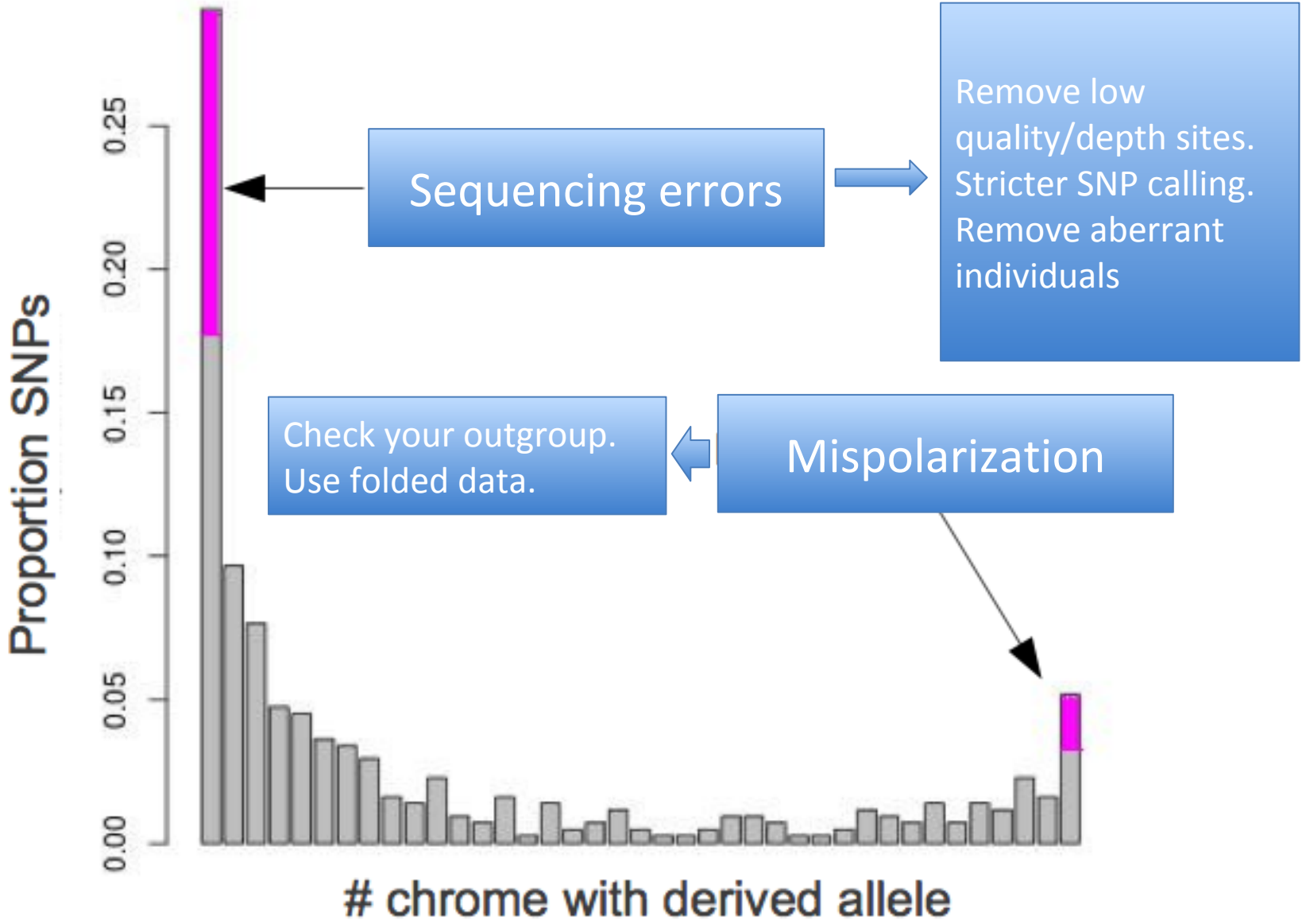
Effect of errors on the SFS



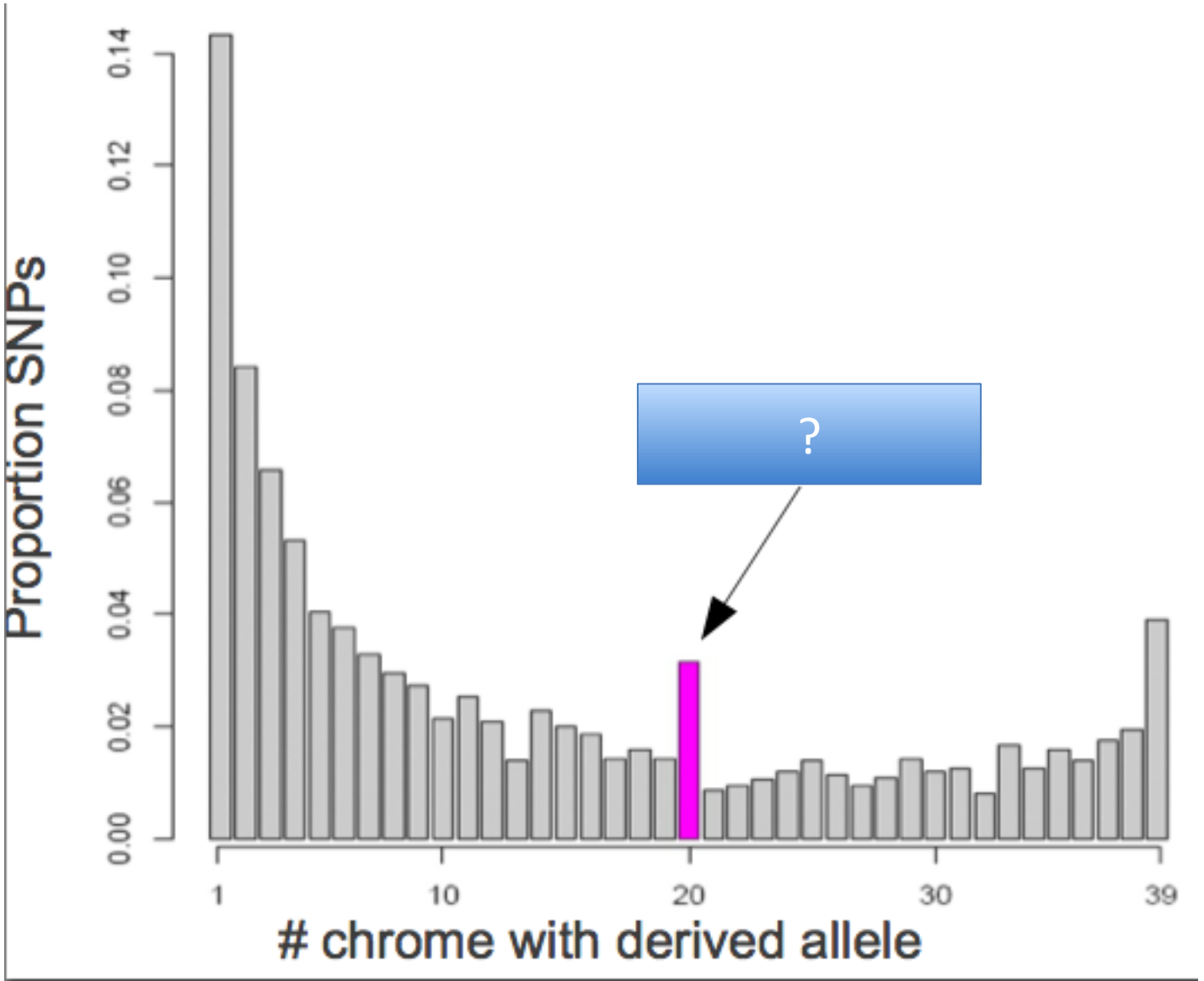
Effect of errors on the SFS



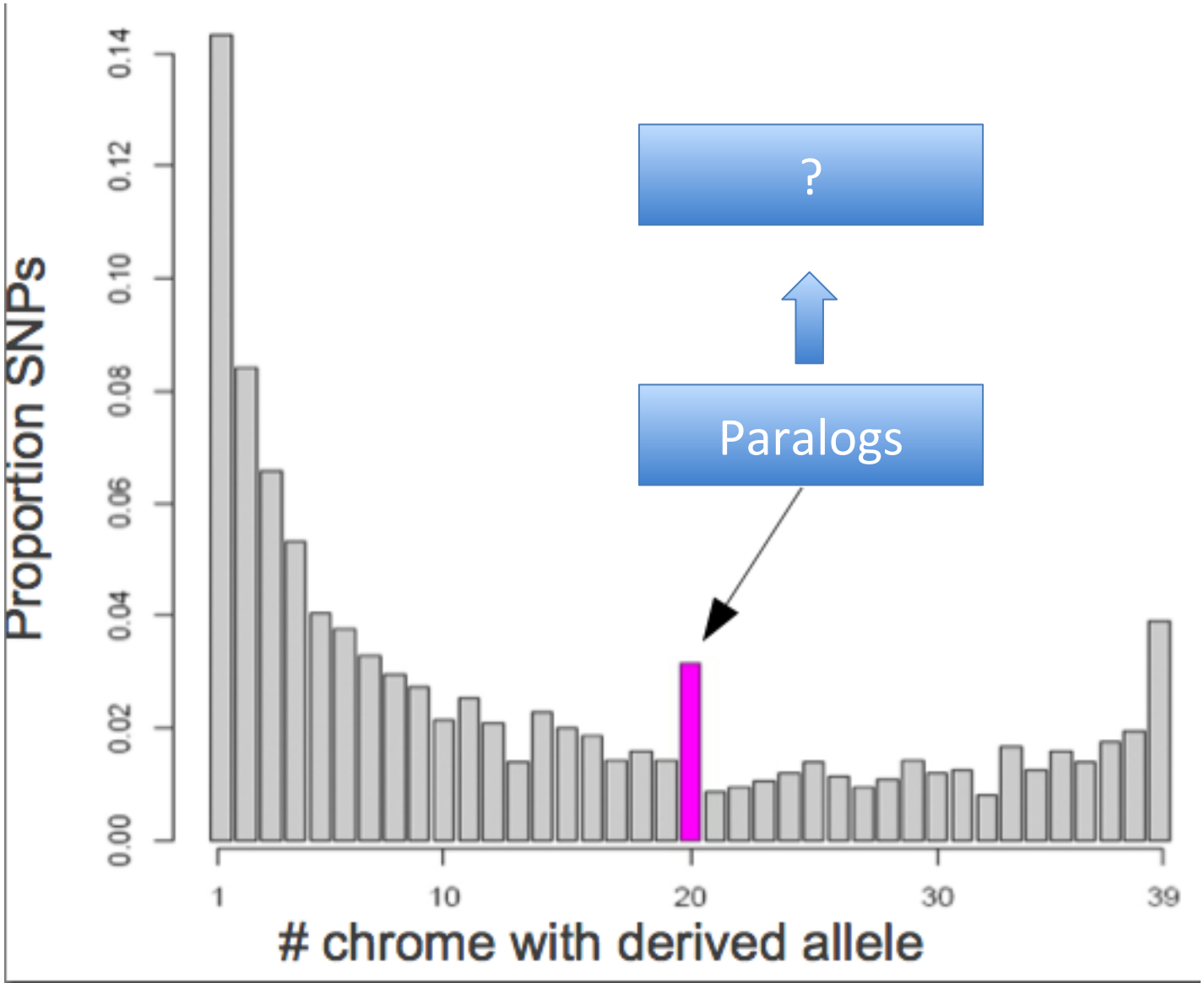
Effect of errors on the SFS



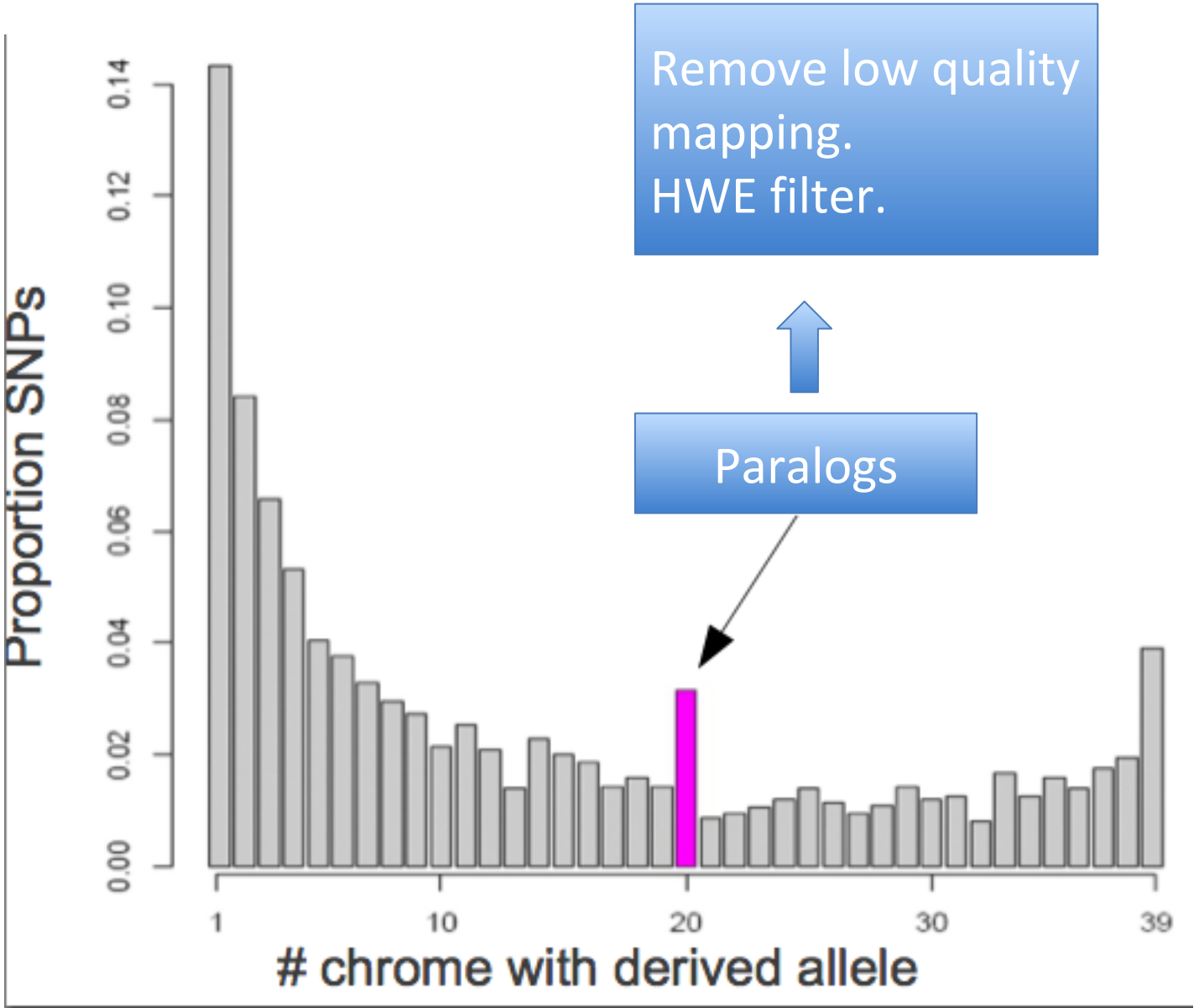
Effect of errors on the SFS



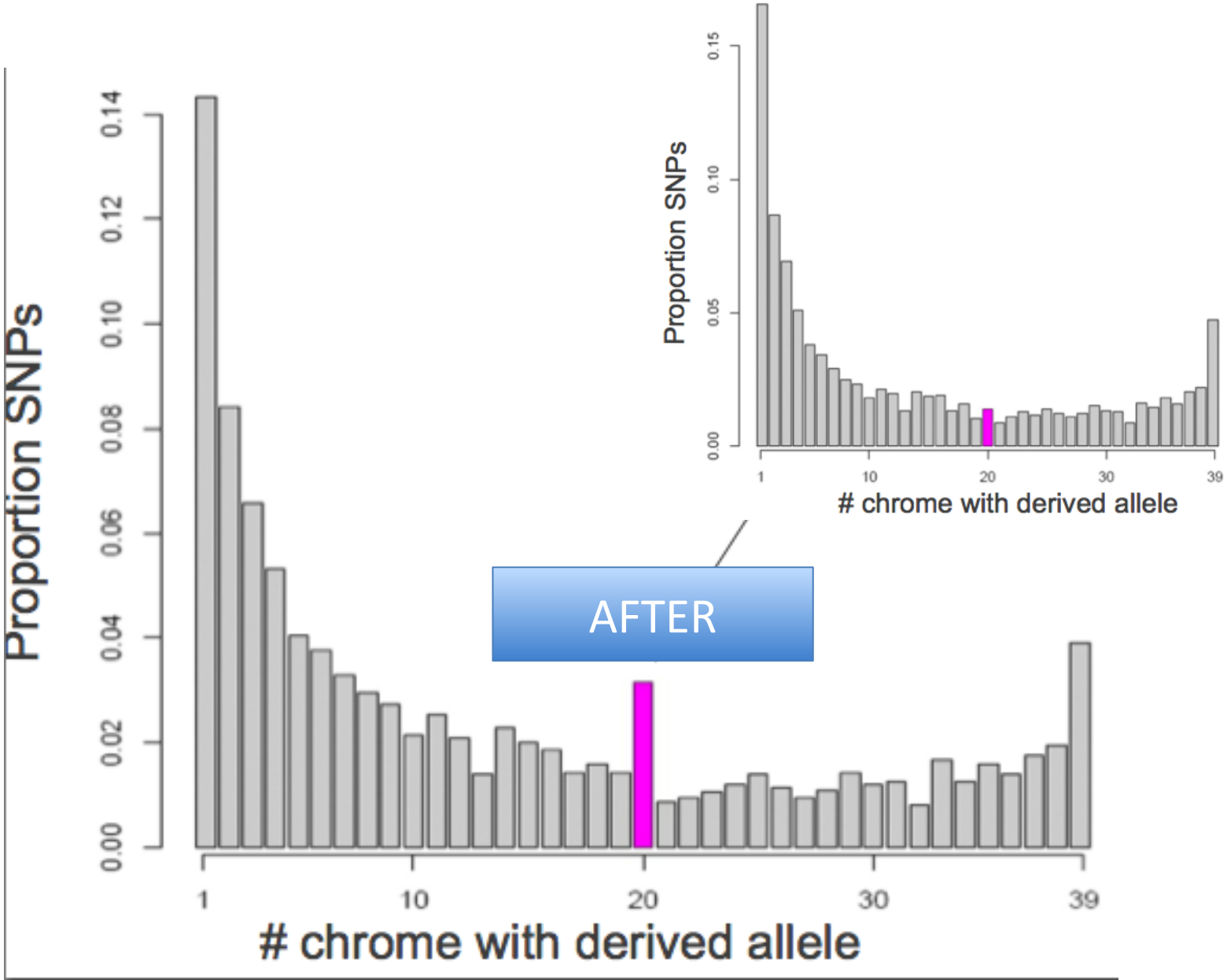
Effect of errors on the SFS



Effect of errors on the SFS



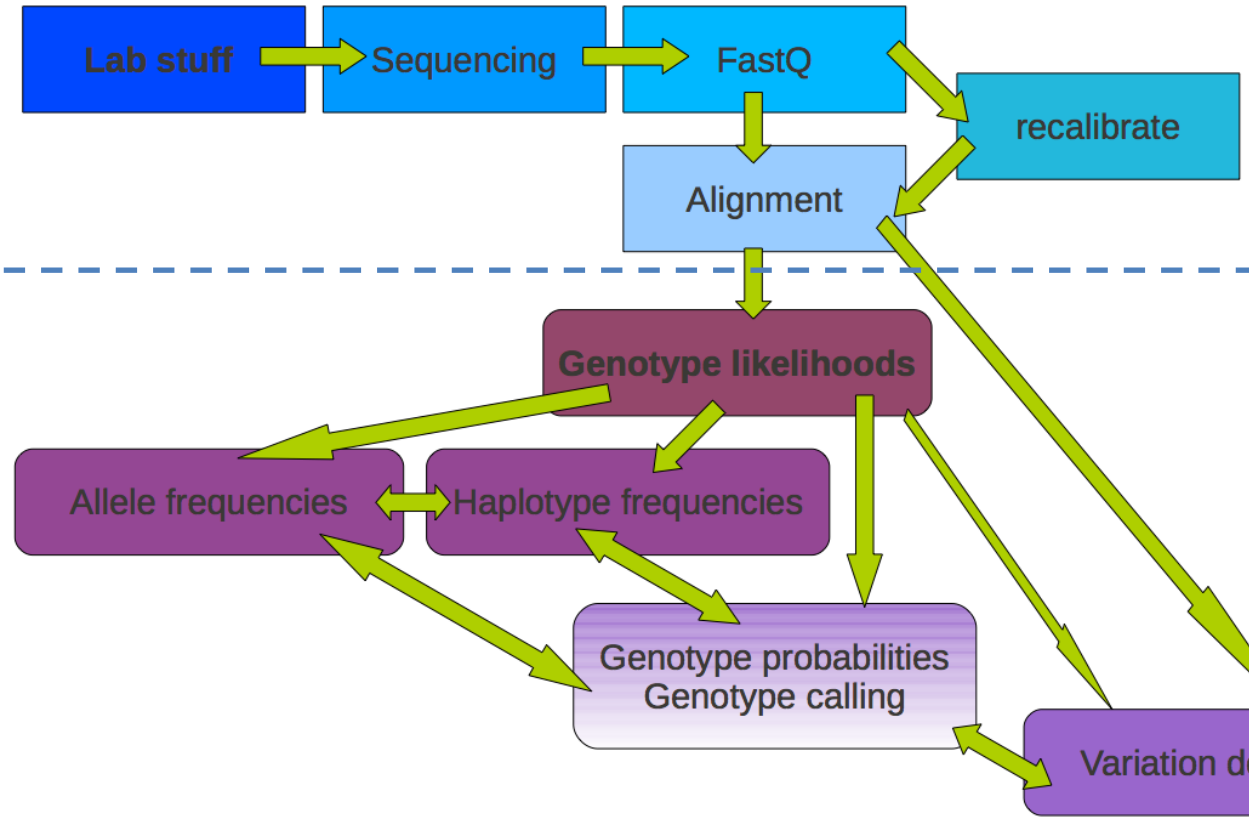
Effect of errors on the SFS



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

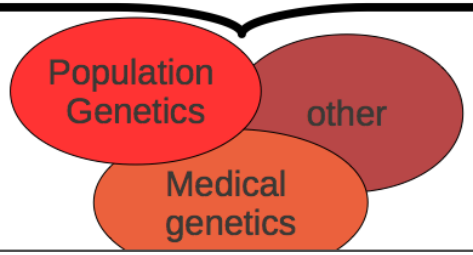


Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

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

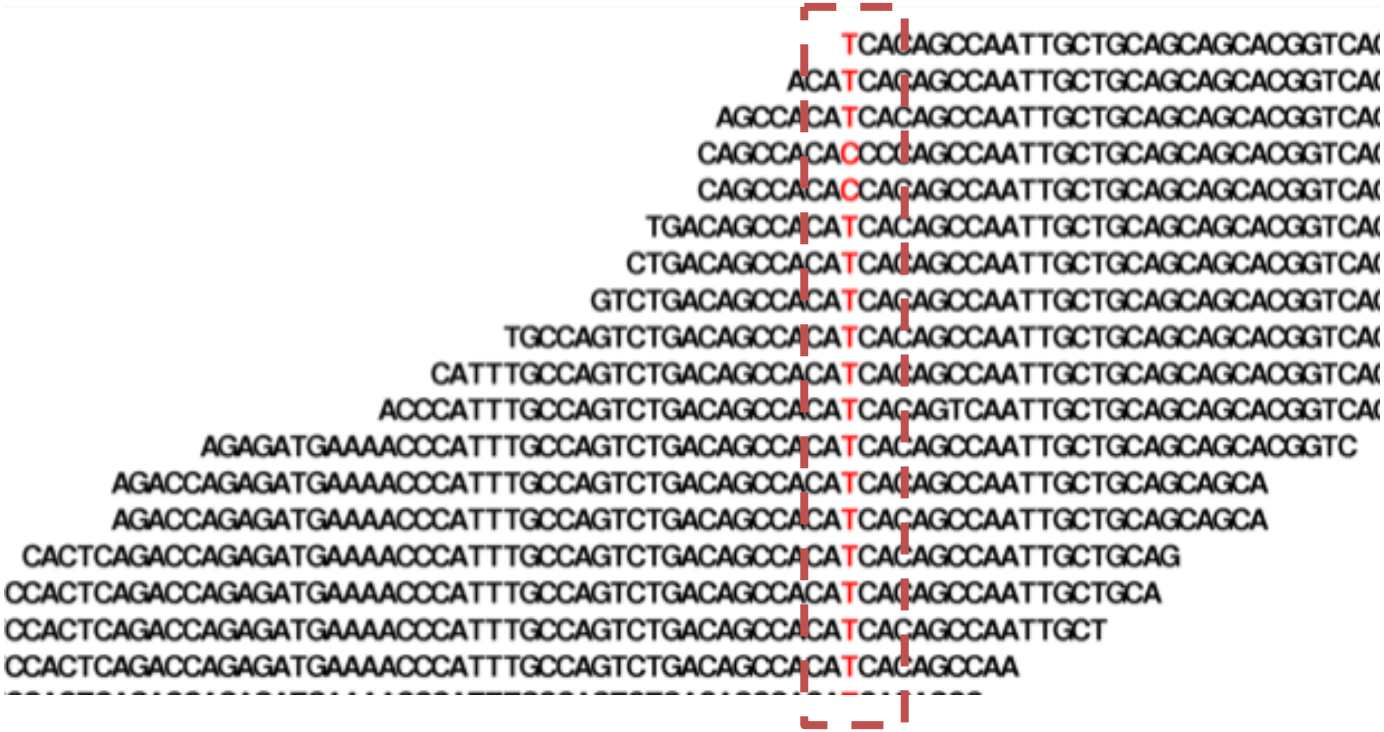


Analysis:

- 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



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

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

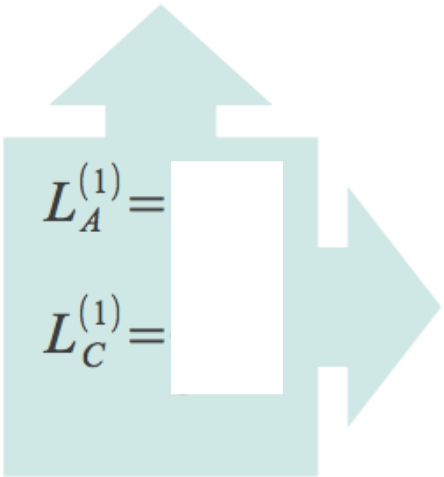
Calculating genotype likelihoods

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

Example:

Chrom6 342 A T T T

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



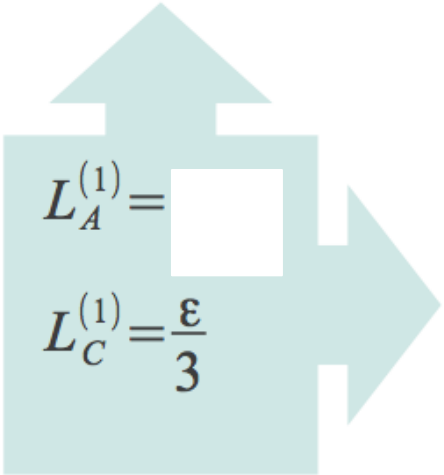
Calculating genotype likelihoods

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

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

$$ATTT$$

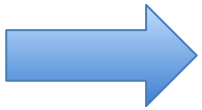
$$\varepsilon = 0.01$$

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

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

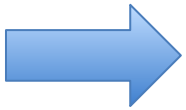
$p(X_s^{(i)} \mid G_s^{(i)})$ → Genotype likelihood

$p(G_s^{(i)})$ → 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	0.01	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

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

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**

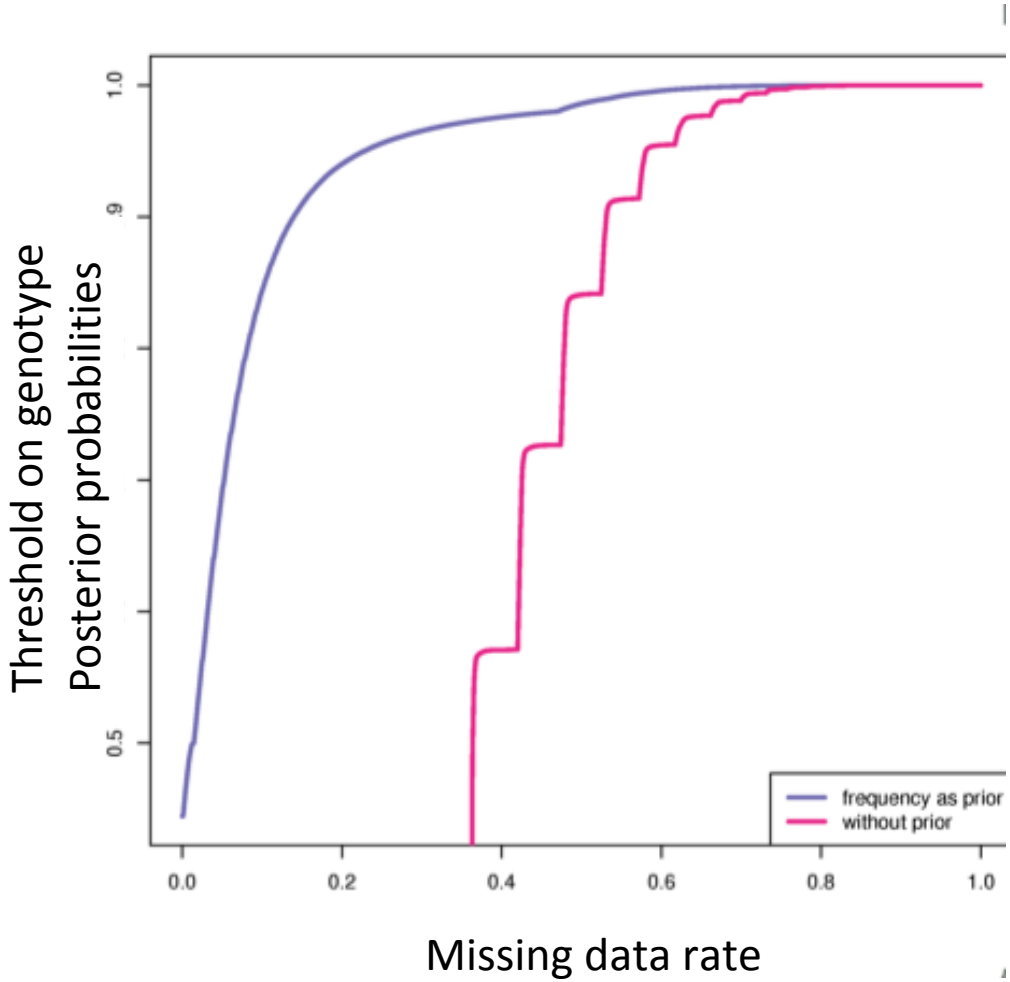
P(A) = f

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)
- **ANGSD** (http://popgen.dk/ANGSD)
- **GATK** (https://www.broadinstitute.org/gatk)

which we will explore during the practical session.

Practical session

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

Paper discussion

- ?