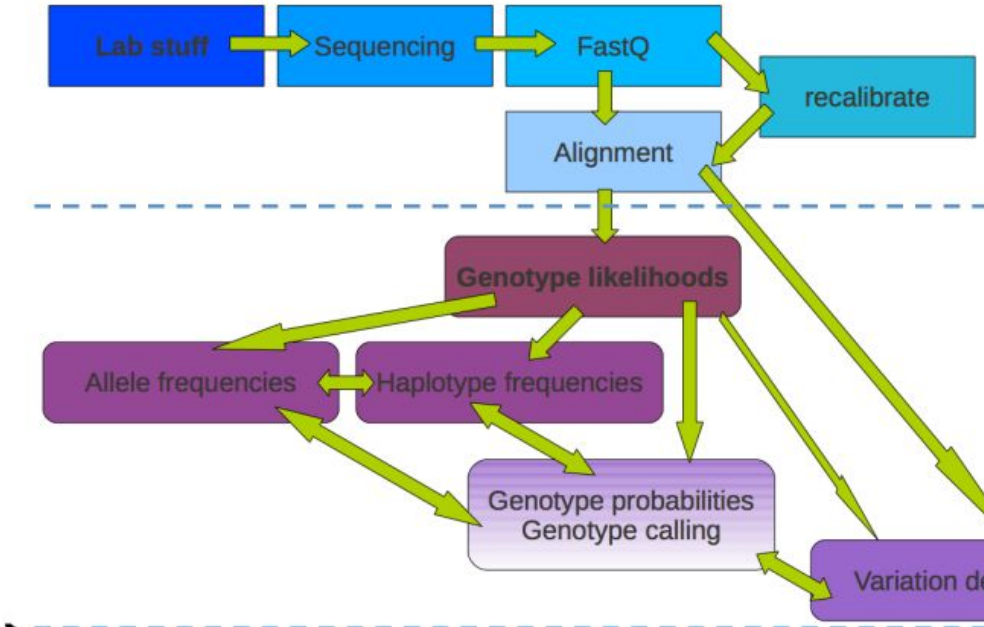
GENOME ANALYSIS &
BIOINFORMATICS

Week 5:

Genetic variation and variant calling from NGS data

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

Genome (FASTA)

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

Reads (FASTQ)

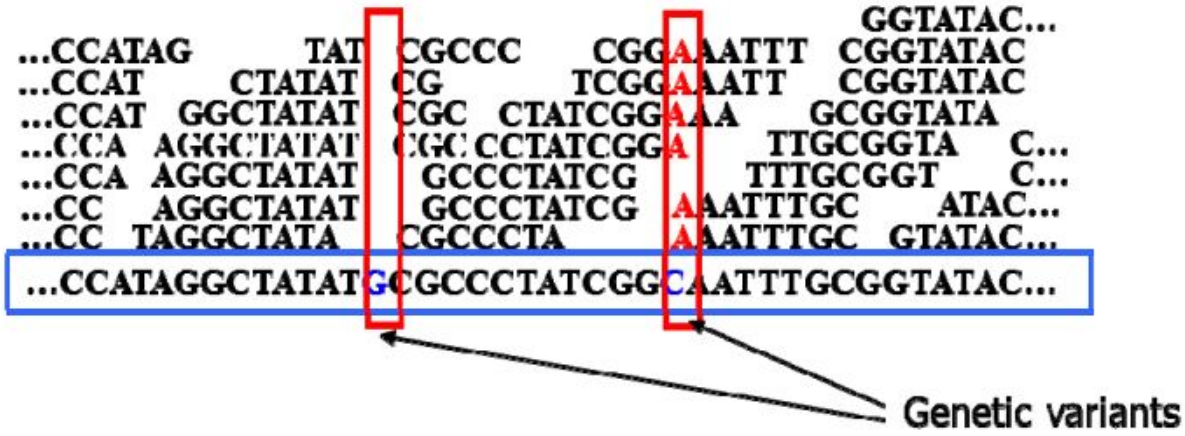
```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

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)

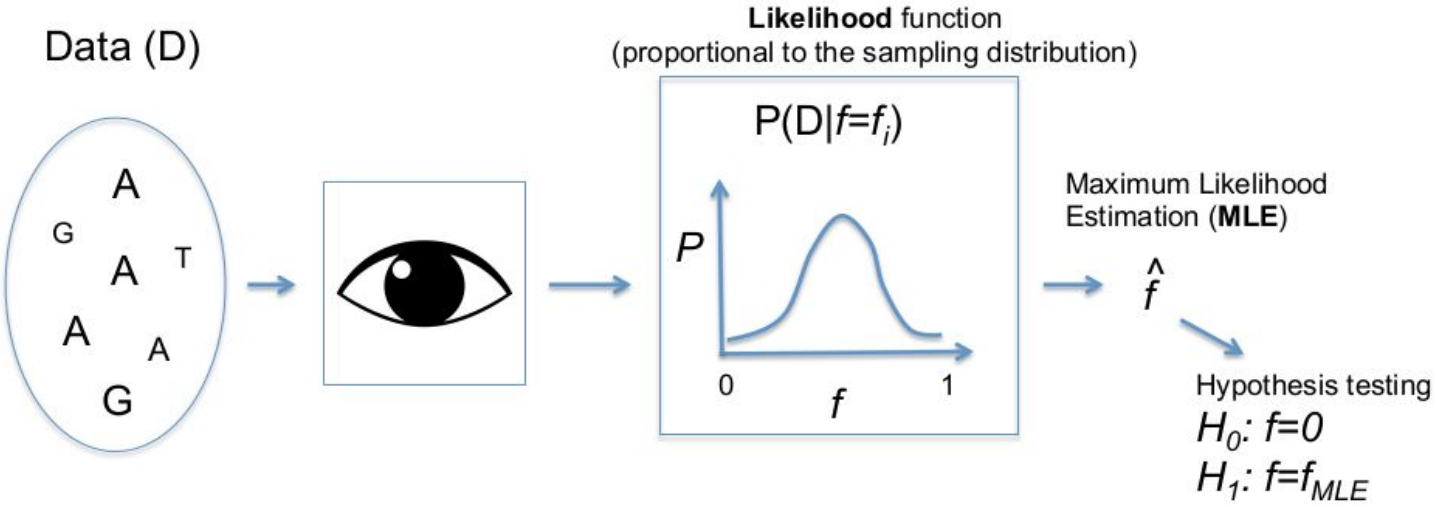
```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23andme2vcf
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  GENOTYPE
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
```



- Counting high-confident , non-reference allele (i.e. Quality >= 20)
  - Freq <20% or > 80%: homozygous genotype
  - Otherwise: heterozygous

- Works well for "deeply sequenced regions" (DSR), i.e. depth > 25x
  - But suffer from under-calling of heterozygous genotypes for low-coverage regions
  - And can't give an objective measurement for reliability

Statistical inference (1)



Likelihood approach:

- All the information on the parameter is in the likelihood function (we use all the data!).
- More data leads to less bias and less variance.
- Suitable for hypothesis testing.

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

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

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

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

$$ATTT$$

$$\varepsilon = 0.01$$

Likelihood Ratio:

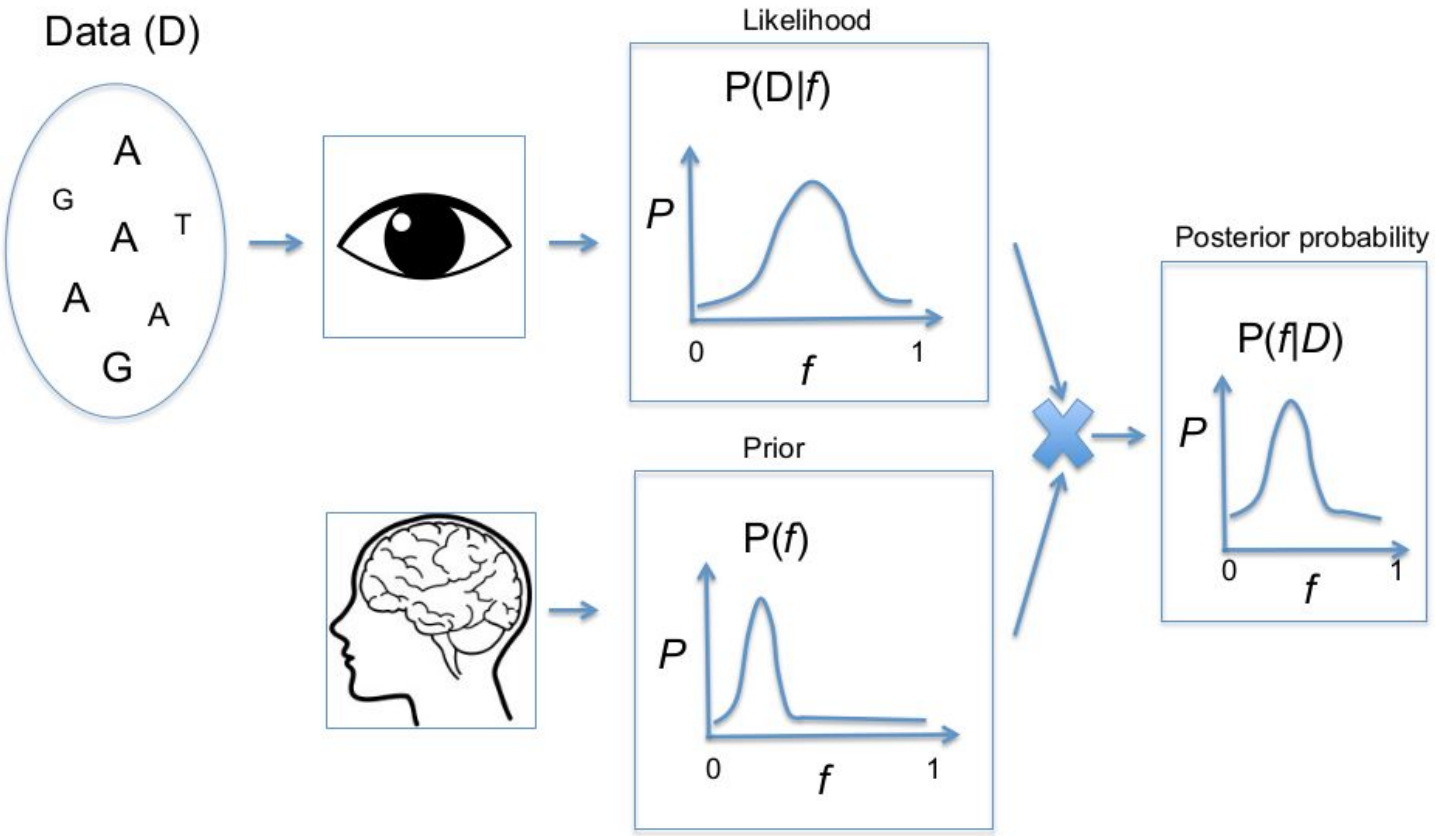
$$\log_{10} \frac{L_{G(1)}}{L_{G(2)}} > t$$

$$t = 1$$

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

(in our example t=1.27)

Statistical inference (2)



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

- Parameter is not fixed (point estimate) but rather has a probability distribution
- We update our “belief” on the parameter after performing the experiment
- As P(f|D) is a proper probability distribution, we can easily derive credible intervals

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

Genotype posterior probabilities

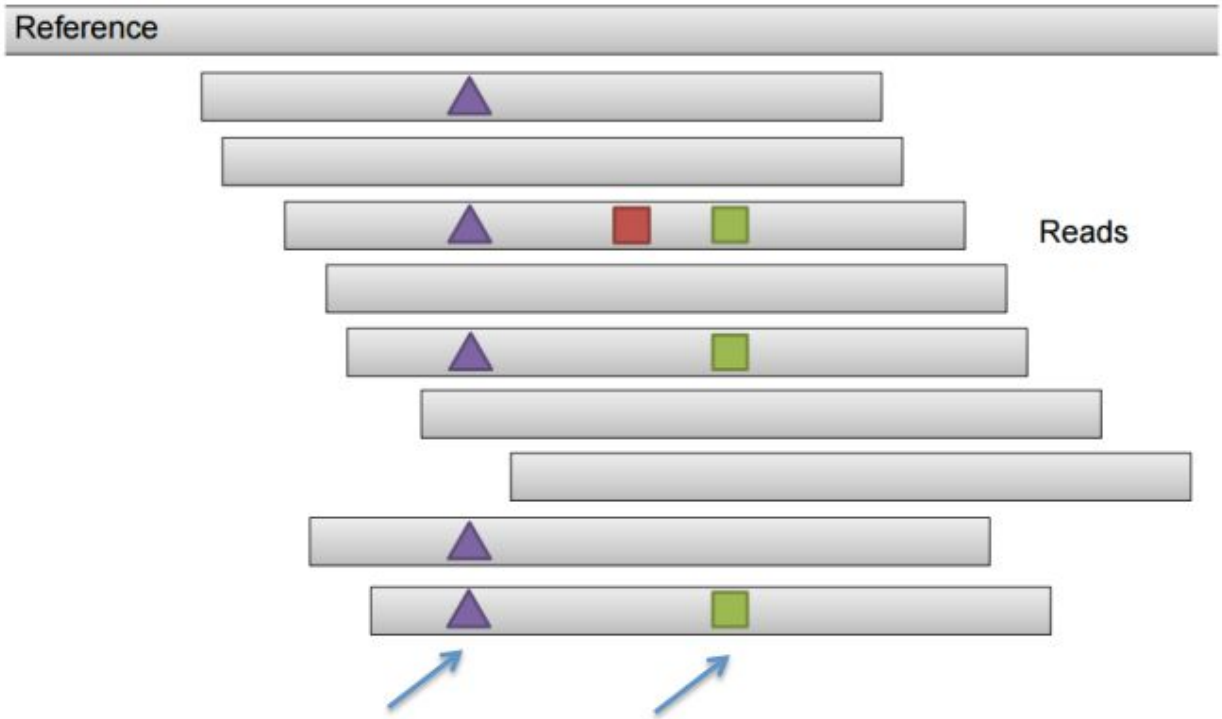
But **only** call the genotype if the largest probability is above a threshold (e.g. > 0.95)

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

Genotype posterior probabilities

P(A) = f=0.6

SNP calling procedures



We completely rely on how reads have been mapped

Haplotype-based caller

1. Define active regions
2. Determine haplotypes by assembly of active regions
3. Determine LH of haplotypes given read data
4. Assign sample genotypes
5. Determine allele frequency at all sites
6. Determine presence of SNPs

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)

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

Genotype likelihoods

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

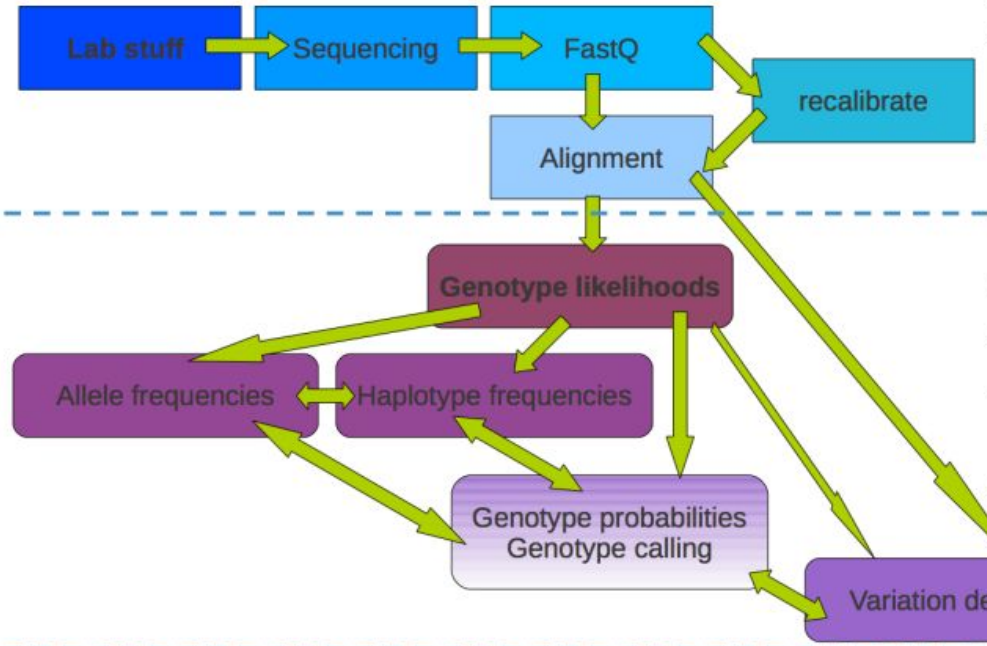
If we assume HWE:

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

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

$$\hat{f} = 0.46$$

Workflow



Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- SNPs detection

Population genetics analysis:

- **Site Frequency Spectrum**
- Summary statistics

Possible measures of genetic variation

Total raw data

the numbers of the different kinds of polymorphic sites and their distribution along the sequence/genome

Pros	**Cons**
most detailed information available	computationally intensive not practical often times impossible

SFS is the most information rich summary statistics

If the mutation rate per site is constant along the sequence, the number of segregating sites should, on average, be proportional to the length of the sequences. S should be proportional mutation rate.

S should increase with the sample size

S & Pi: reduces all information contained in the data into a single number!

SFS: an intermediate measure, between the total data and the extreme summaries.

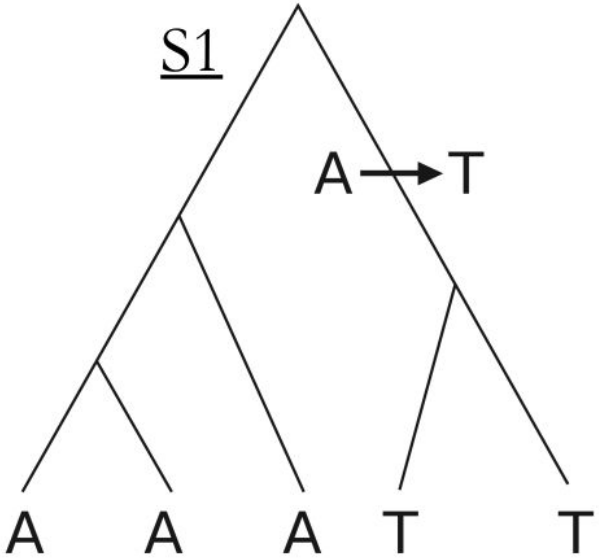
Both S & Pi can be derived from the SFS

What is the SFS?

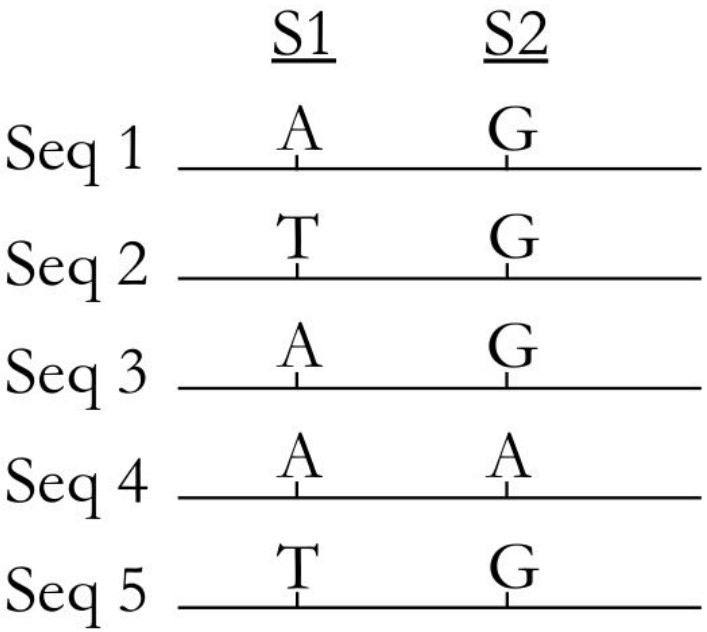
Is the number of counts of each possible kind of polymorphic site (distinguishing between sites that partition the sample in different ways), based on the correspondence between genealogical structure and observable DNA data under the infinite-sites mutation model.

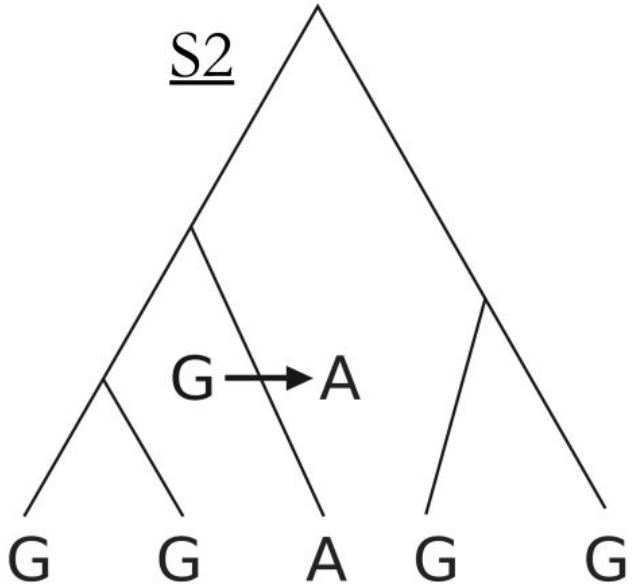
Genealogical structure & DNA data

	S1	S2
Seq 1	A	G
Seq 2	T	G
Seq 3	A	G
Seq 4	A	A
Seq 5	T	G

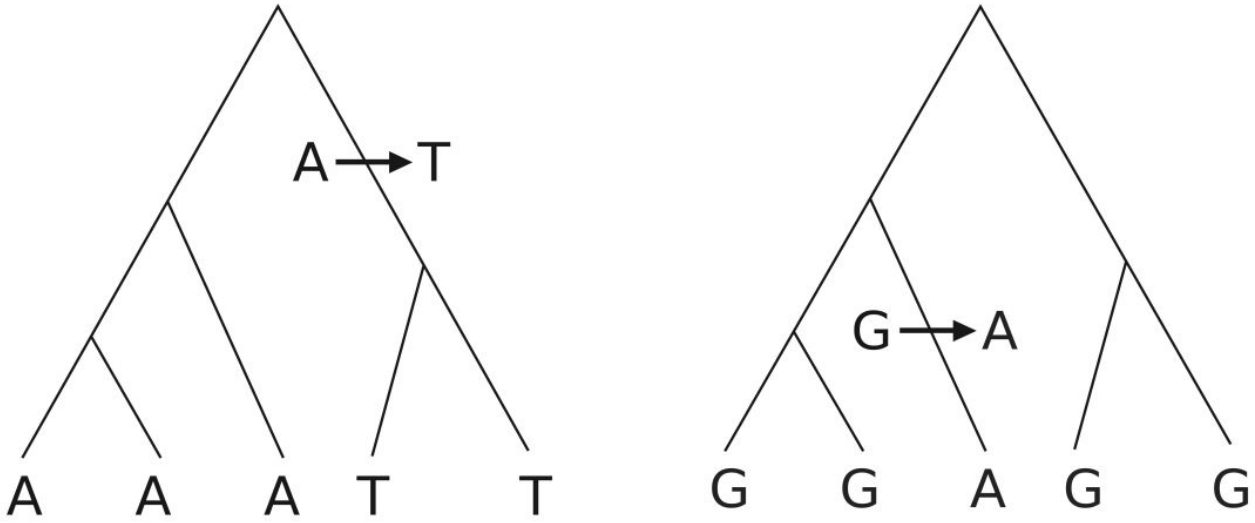


Genealogical structure & DNA data





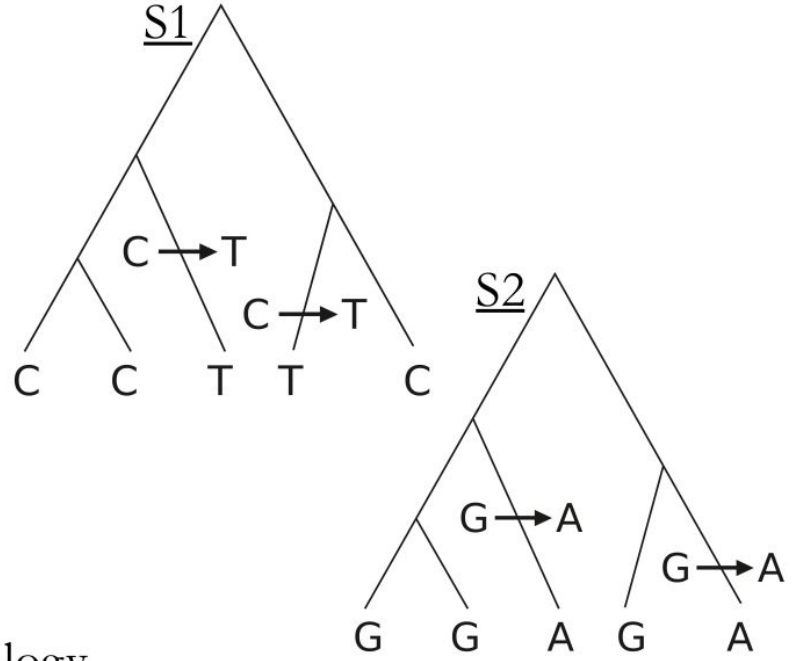
Genealogical structure & DNA data



Under the infinite sites model there is a branch in the genealogy that partitions the sample identically to the pattern at that site. If this is the case we say that the site pattern is compatible with the genealogy

Genealogical structure & DNA data

	S1	S2
Seq 1	C	G
Seq 2	C	G
Seq 3	T	A
Seq 4	T	G
Seq 5	C	A

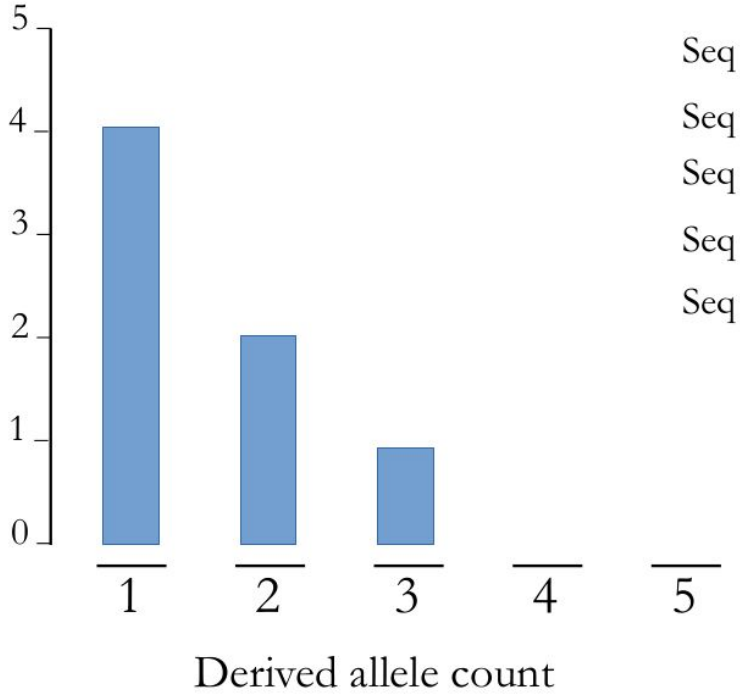


Two site patterns which are incompatible with their genealogy

Calculating the SFS

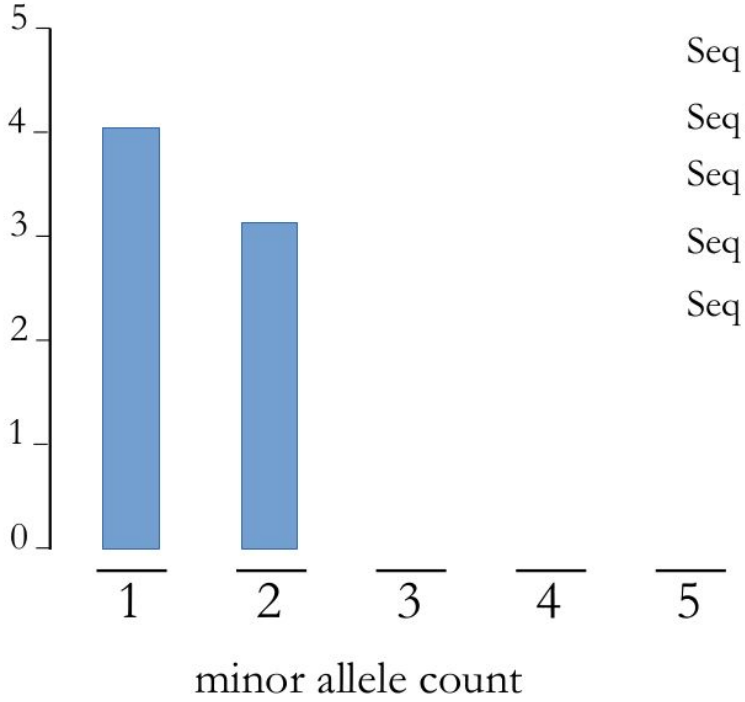
Ni = The number of sites "**ξi**" at which the mutant base is present in i copies and the ancestral base is in n – i copies

Calculating the SFS



Unfolded SFS

Calculating the SFS



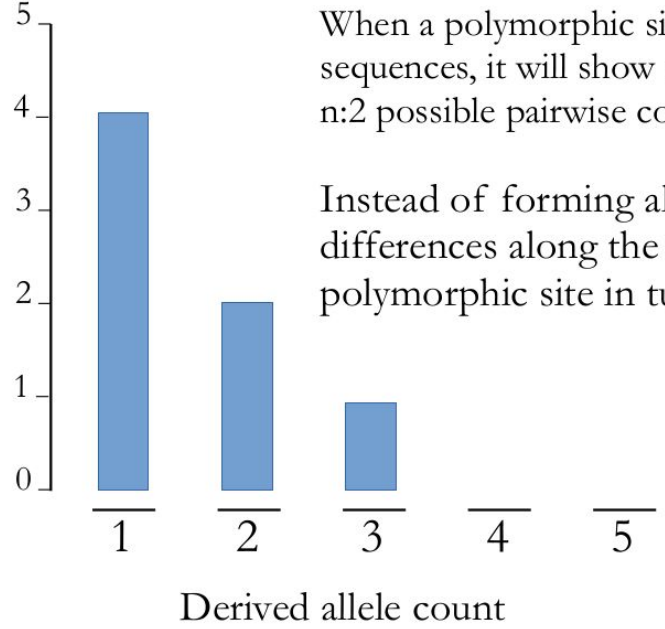
Folded SFS

Deriving summary stats from the SFS

$$S = \sum_{i=1}^{[n/2]} \eta_i$$

For S both simply count the number of poly sites

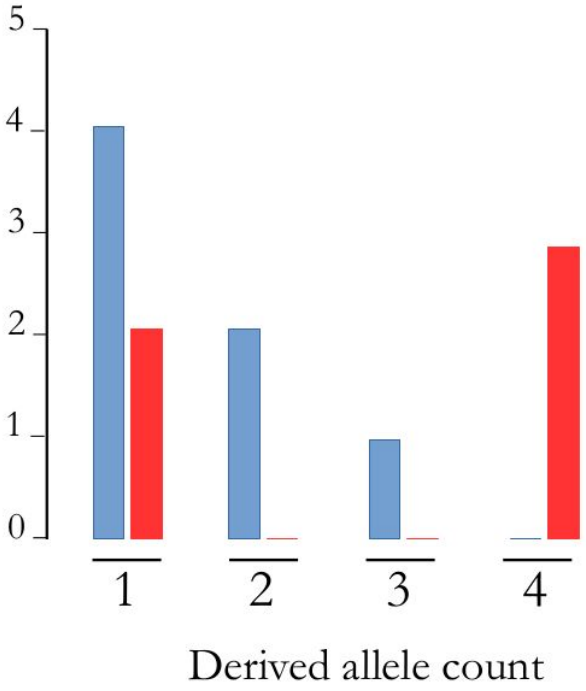
$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{[n/2]} i(n-i)\eta_i.$$

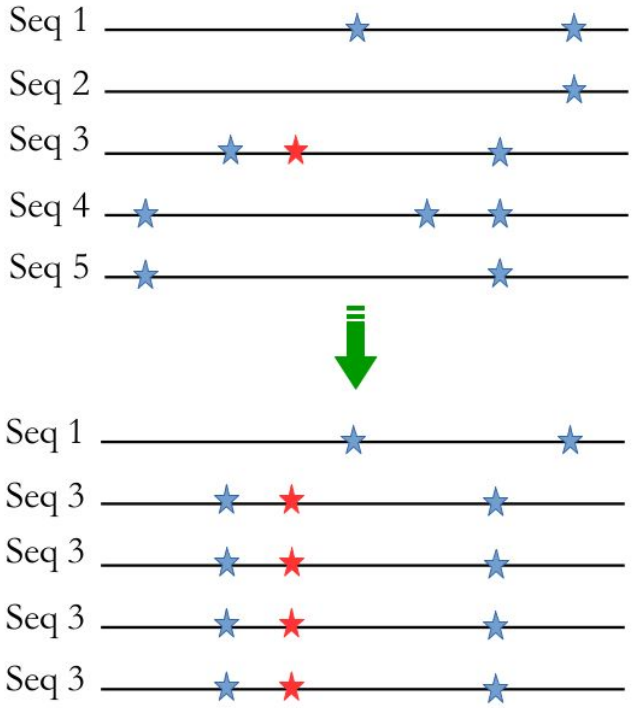


When a polymorphic site divides the sample into i and n– i sequences, it will show a pairwise difference in exactly i(n–i) of the n:2 possible pairwise comparisons.

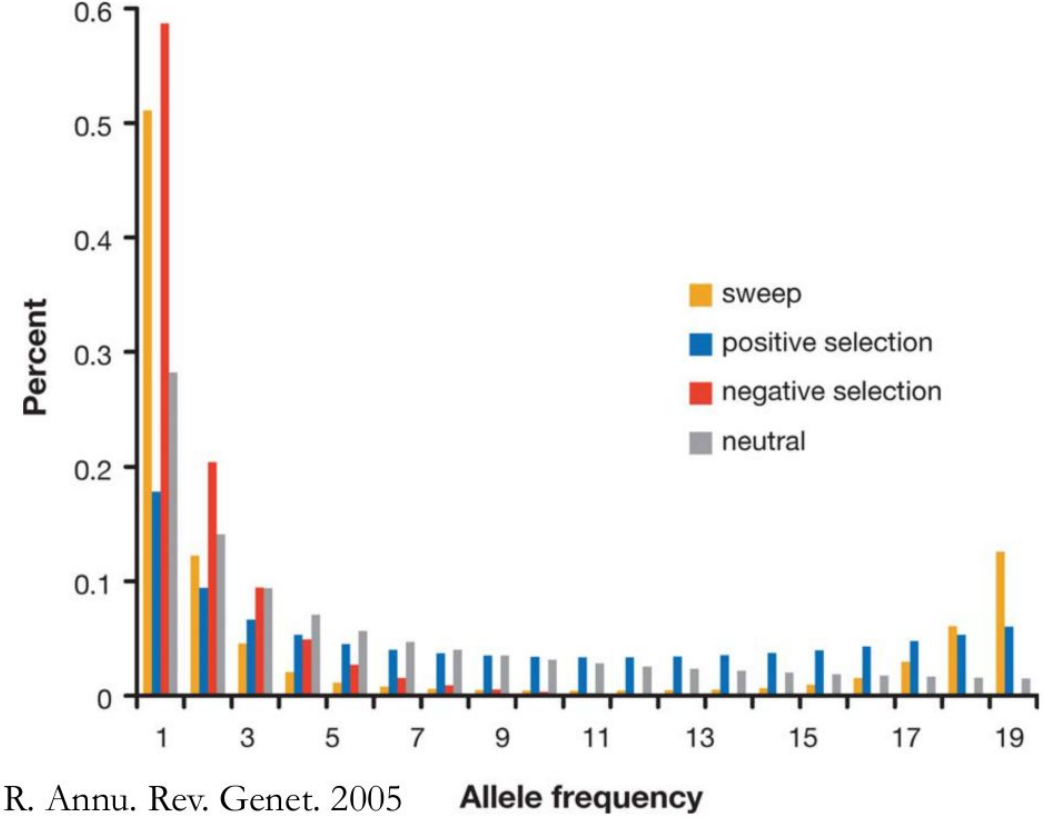
Instead of forming all pairs and counting total numbers of differences along the sequence, we can consider each polymorphic site in turn and calculate its contribution to π

Positive selection & SFS





Influence of evolutionary forces on the SFS



Nielsen R. Annu. Rev. Genet. 2005

Effect of errors on the SFS