Imperial College
London

Estimation of site frequency spectra

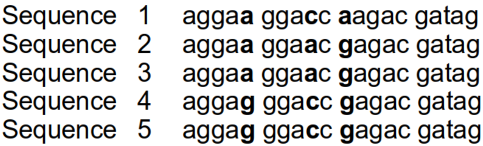
Matteo Fumagalli

Intended Learning Outcomes

By the end of this session you will be able to

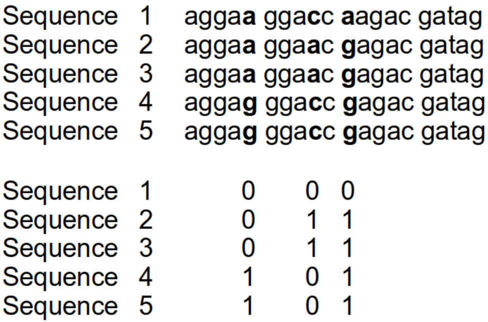
- understand the theory underlying site frequency spectrum
- appreciate how to extended such theory to low-coverage data
- acknowledge the process of inferring demography from sequencing data
- implement a pipeline in ANGSD to perform the aforementioned analyses

The Site Frequency Spectrum (SFS)



Sequence 1 agga**a** gga**c**c **a**agac gatag
Sequence 2 agga**a** gga**a**c **g**agac gatag
Sequence 3 agga**a** gga**a**c **g**agac gatag
Sequence 4 agga**g** gga**c**c **g**agac gatag
Sequence 5 agga**g** gga**c**c **g**agac gatag

The Site Frequency Spectrum (SFS)



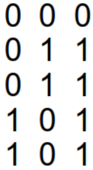
Sequence	1	agga**a** gga**c**c **a**agac gatag
Sequence	2	agga**a** gga**a**c **g**agac gatag
Sequence	3	agga**a** gga**a**c **g**agac gatag
Sequence	4	agga**g** gga**c**c **g**agac gatag
Sequence	5	agga**g** gga**c**c **g**agac gatag

Sequence	1	0	0	0
Sequence	2	0	1	1
Sequence	3	0	1	1
Sequence	4	1	0	1
Sequence	5	1	0	1

The Site Frequency Spectrum (SFS)

SFS

The SFS is obtained by tabulating the sample allele frequencies of all mutations.

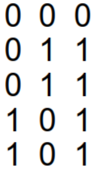


0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

The Site Frequency Spectrum (SFS)

SFS

The SFS is obtained by tabulating the sample allele frequencies of all mutations.



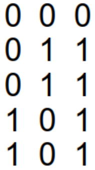
0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

The "1" alleles have frequencies 2/5, 2/5 and 4/5. The proportions of "1" alleles with a frequency of 1/5, 2/5, 3/5 and 4/5 in the sample are

The Site Frequency Spectrum (SFS)

SFS

The SFS is obtained by tabulating the sample allele frequencies of all mutations.



0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

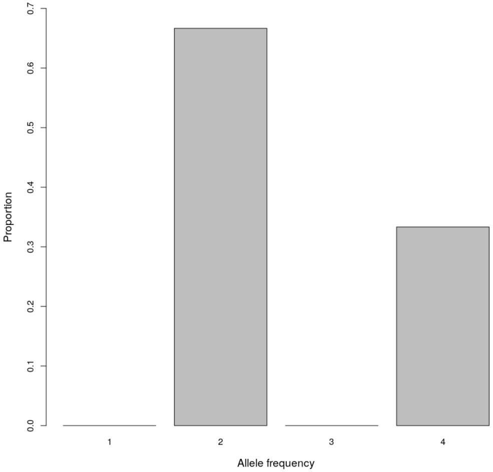
The "1" alleles have frequencies 2/5, 2/5 and 4/5. The proportions of "1" alleles with a frequency of 1/5, 2/5, 3/5 and 4/5 in the sample are $f_1 = 0$, $f_2 = 2/3$, $f_3 = 0$ and $f_4 = 1/3$.

$$\vec{f} = (f_1, f_2, ..., f_{n-1})$$

for a sample of n haploid individuals.

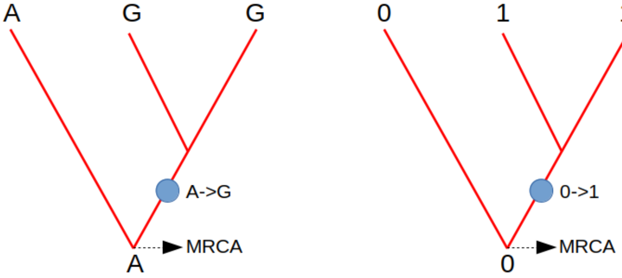
The Site Frequency Spectrum (SFS)

```
0 0 0
0 1 1
0 1 1
1 0 1
1 0 1
```



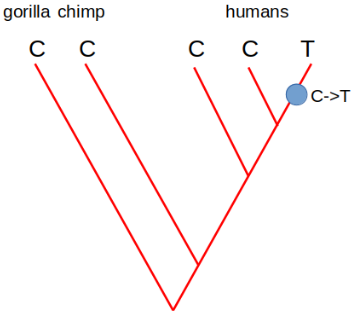
Alleles

- **ancestral** allele is the allele found in the MRCA of the sample.
- **derived** allele (or mutated) is an allele that is not ancestral.



Alleles

The ancestral allele is often inferred using **outgroups**.
e.g. if C/T polymorphism in humans and primate have C, then C is likely to be the ancestral allele.



Alleles

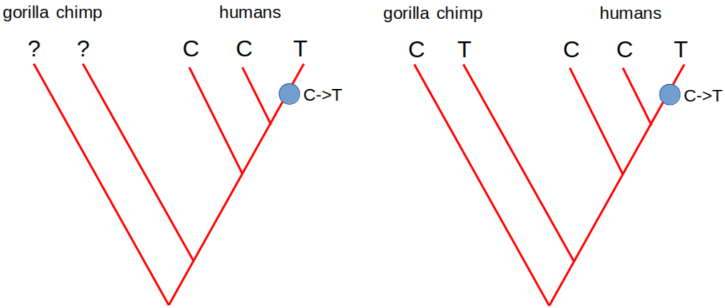


Figure 1: Uncertain ancestral allele.

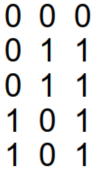
The Site Frequency Spectrum (SFS)

If no information on the ancestral allele is available, we can *fold* the frequency spectrum.

The **folded frequency spectrum** f^* is obtained by adding together the frequencies of the derived and ancestral alleles.

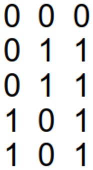
$f^* = f_i + f_{n-j}$ for $j < n/2$ and
$f^* = f_j$ for $j = n/2$
only defined for values of $f^* \leq n/2$.

The folded SFS



0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

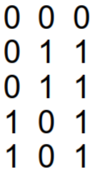
The folded SFS



$$\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{array}$$

$$\vec{f}^{*} =$$

The folded SFS



0	0	0
0	1	1
0	1	1
1	0	1
1	0	1

$$\vec{f^*} = (f_1^* = 1/3, f_2^* = 2/3)$$

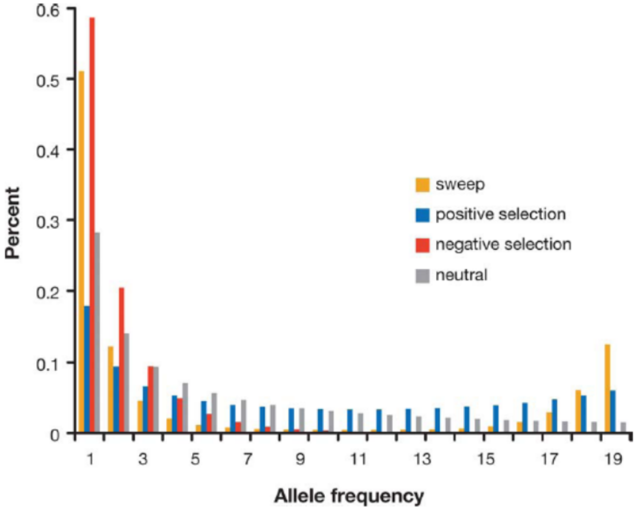
The Site Frequency Spectrum

- S and π can be calculated directly from $\vec{f}$ but the opposite is not true.
- Alleles segregating at frequency of $1/n$ are called **singletons**.
- The expected SFS under the standard coalescence model with infinite sites mutations is

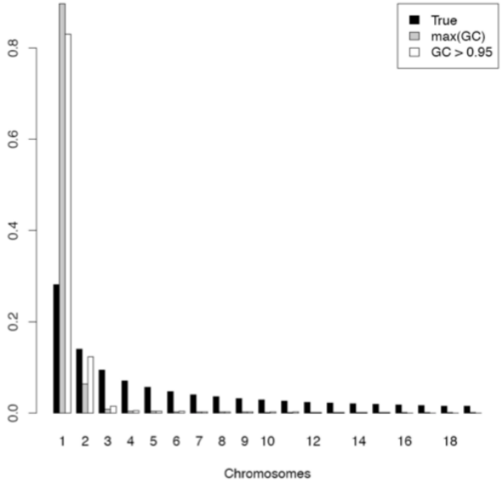
  $$E[f_i] = \frac{1/j}{\sum_{k=1}^{n-1} \frac{1}{k}} \qquad (1)$$

  with $j = 1, 2, ..., n-1$

Fundamental statistics to infer demography of your population of interest.

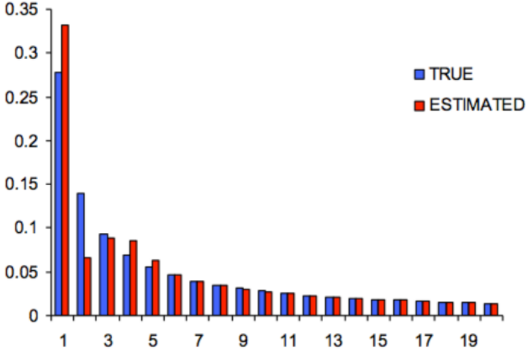

0.6
0.5
0.4
0.3
0.2
0.1
0
Percent
sweep
positive selection
negative selection
neutral
1
3
5
7
9
11
13
15
17
19
Allele frequency

Effect of errors on SFS



Effect of errors on SFS

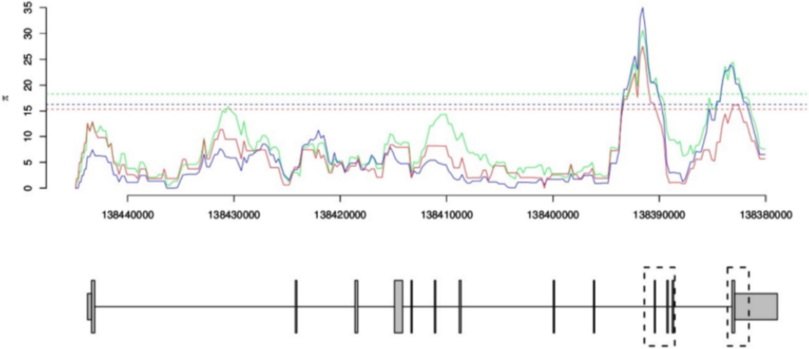
Using an ad hoc fixed cutoff for SNP calling…



can never produce unbiased estimates.

Effects of low-depth data

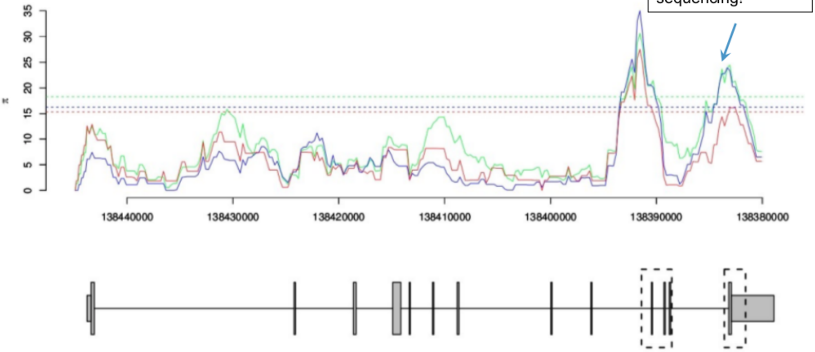
Nucleotide diversity scan using 1000 Genomes Project data (low-depth)



Cagliani et al. MBE. 2012

Effects of low-depth data

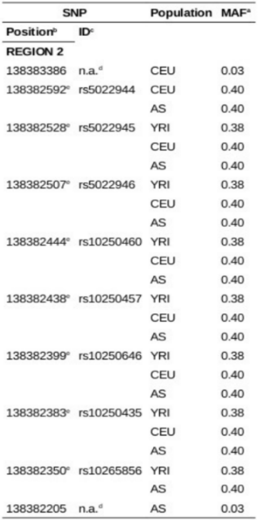
Nucleotide diversity scan using 1000 Genomes Project data (low-depth)



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

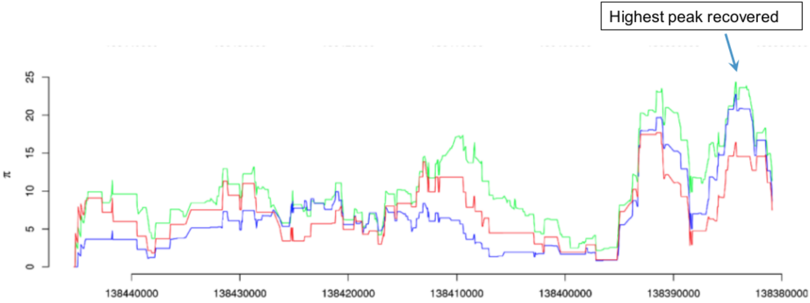
Most of them (n=8) have intermediate frequency in all populations.

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



Highest peak recovered

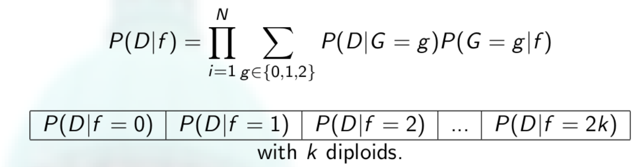
- ○ Missing data
- ○ Unpredictable effects

Cagliani et al. MBE 2012

How can we estimate the site frequency spectrum from low-coverage sequencing data?

How can we estimate the site frequency spectrum from low-coverage sequencing data?
Can we count alleles over genotypes?
Can we "round up" estimated allele frequencies?
Can we estimate the SFS directly from genotype likelihoods?

Sample allele frequency (saf) likelihoods



$$P(D|f) = \prod_{i=1}^{N} \sum_{g \in \{0,1,2\}} P(D|G=g)P(G=g|f)$$

$P(D\|f=0)$	$P(D\|f=1)$	$P(D\|f=2)$	...	$P(D\|f=2k)$

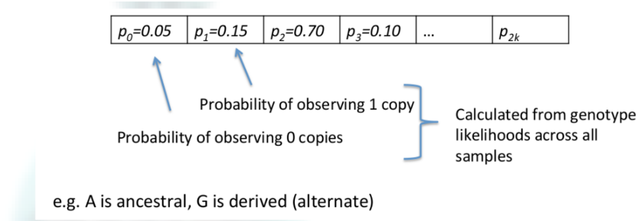
with k diploids.

If unfolded, *2k+1* entries

$p_0=0$	$p_1=0$	$p_2=1$	$p_3=0$	…	$p_{2k}=0$

e.g. A is ancestral, G is derived (alternate)

AA AA AG AA AG AA AA AA AA



$p_0=0.05$	$p_1=0.15$	$p_2=0.70$	$p_3=0.10$	...	p_{2k}

Probability of observing 1 copy

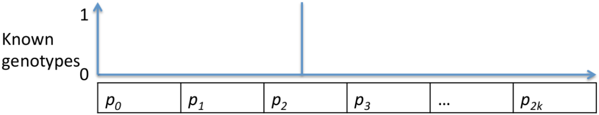
Probability of observing 0 copies

Calculated from genotype likelihoods across all samples

e.g. A is ancestral, G is derived (alternate)

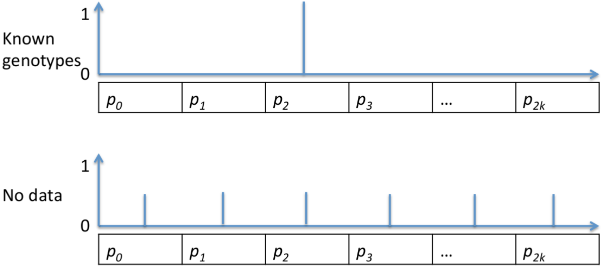
If genotypes are unknown and counting is not possible.

Sample allele frequency (saf) likelihoods

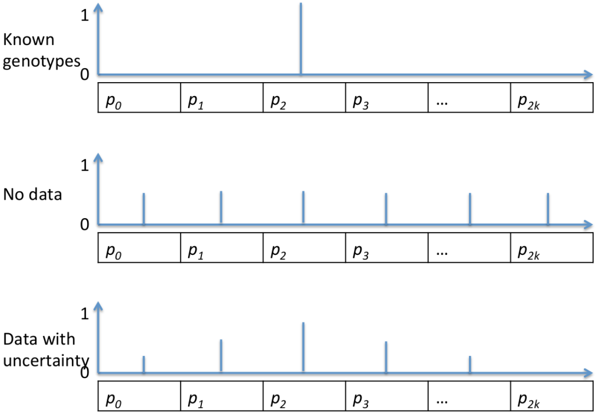


Imperial College London

Sample allele frequency (saf) likelihoods



Sample allele frequency (saf) likelihoods



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

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

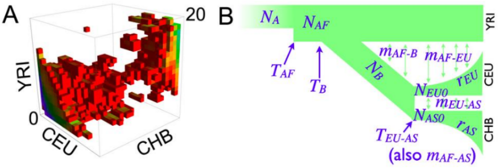
- Likelihood function:

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

Nielsen et al. 2012 PLoS One

Can we go beyond the statistical estimation of unfolded SFS for one single population from low-coverage sequencing data? What are the issues if we have more populations?

Multi-dimensional site frequency spectrum (multi-SFS)



Gutenkunst al. 2009

example on whiteboard?

Estimation of multi-SFS

For *N* populations and *theta* being the SFS and *D* the data and *X* the allele frequency for site *s*:

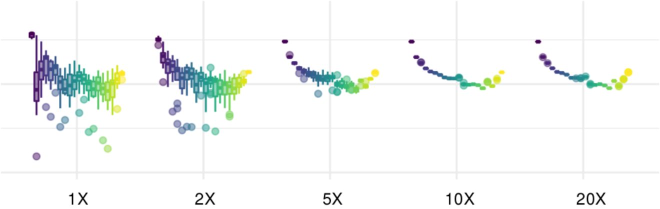
$$L_s(D|\theta) = \underbrace{\prod_{n_1=1}^{n^1} \prod_{n_2=1}^{n^2} \cdots \prod_{n_N=1}^{n^N}}_{N} p(D^1|X = n_1) p(D^2|X = n_2) \cdots p(D^N|X = n_N)$$

with $\Sigma(2n^i+1)$ parameters; optimized using an accelerated EM.

extended from Nielsen et al. 2012 PLoS ONE

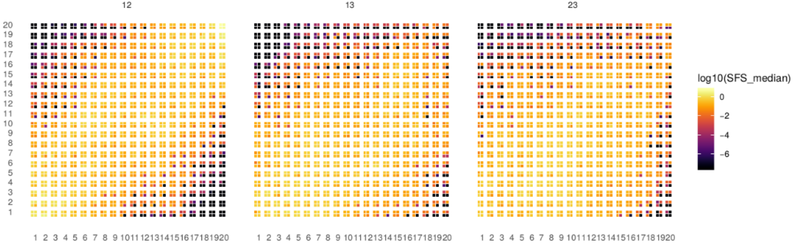
Thorfinn Korneliussen
Anders Albrechtsen

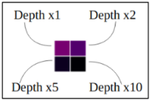
Estimation of 1D SFS



Alex Mas Sandoval

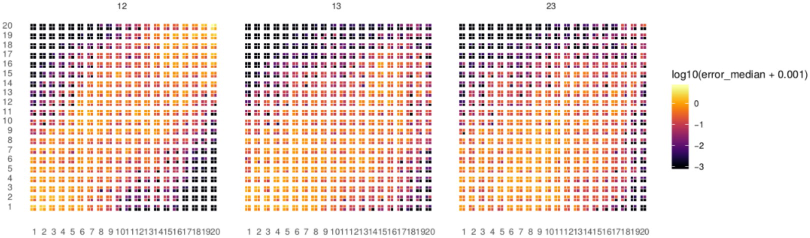
Estimation of 3D SFS

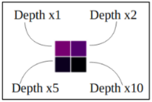




Alex Mas Sandoval

Estimation of 3D SFS (error)

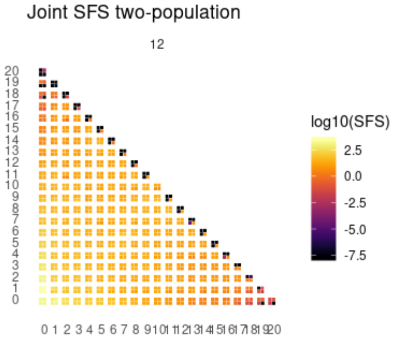




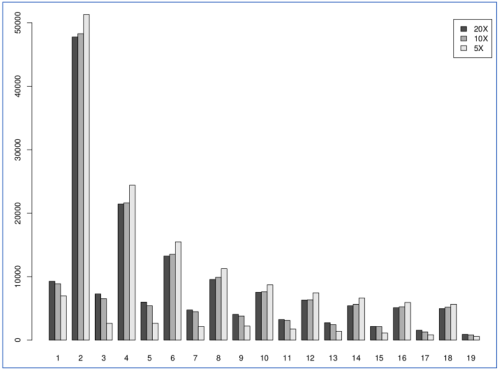
Alex Mas Sandoval

Imperial College London

Folded SFS

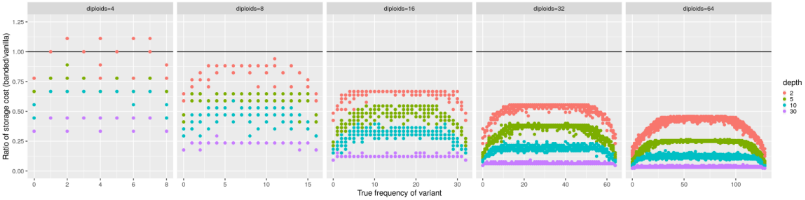


Estimation of SFS for inbred species



extended from Vieira *et al*. 2013 Genome Res

Fast and efficient estimation and data storage



'score-limited' algorithm (Han *et al*. 2015 Bioinformatics):
"*to compute the SAF likelihood: all non-negligible values of the SAF likelihood are concentrated on a few cells around the best-guess allele counts.*"

Nate Pope

Imperial College London

Intended Learning Outcomes

At the end of this session you are now able to

- understand the theory underlying site frequency spectrum
- appreciate how to extended such theory to low-coverage data
- acknowledge the process of inferring demography from sequencing data
- implement a pipeline in ANGSD to perform the aforementioned analyses