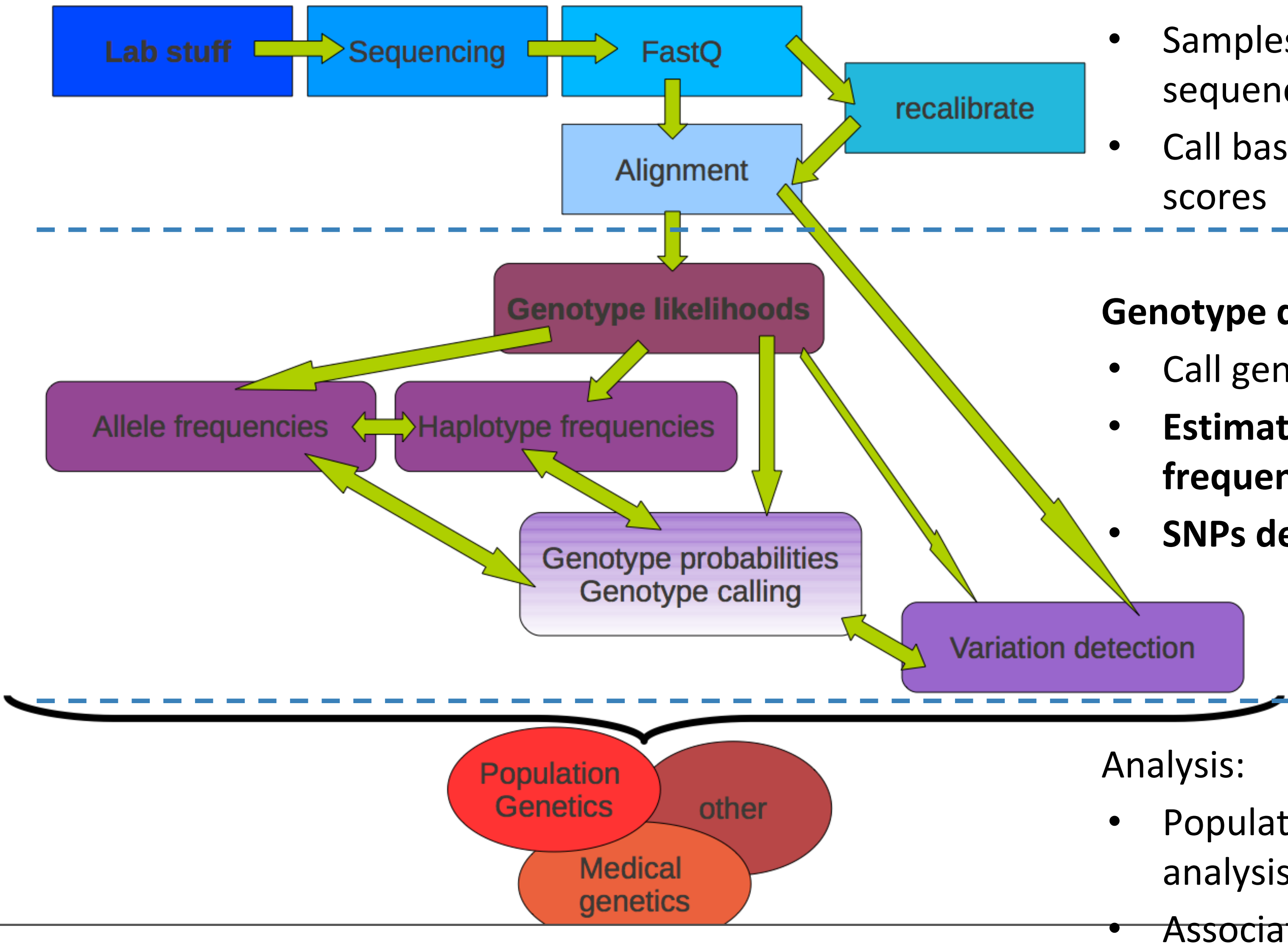
Evolutionary genomics
Data analysis module – Day 2

SNP calling and advanced methods for evolutionary inferences from NGS data

April 14th 2015

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

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 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 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)

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

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

Estimating allele frequencies

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

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

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

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

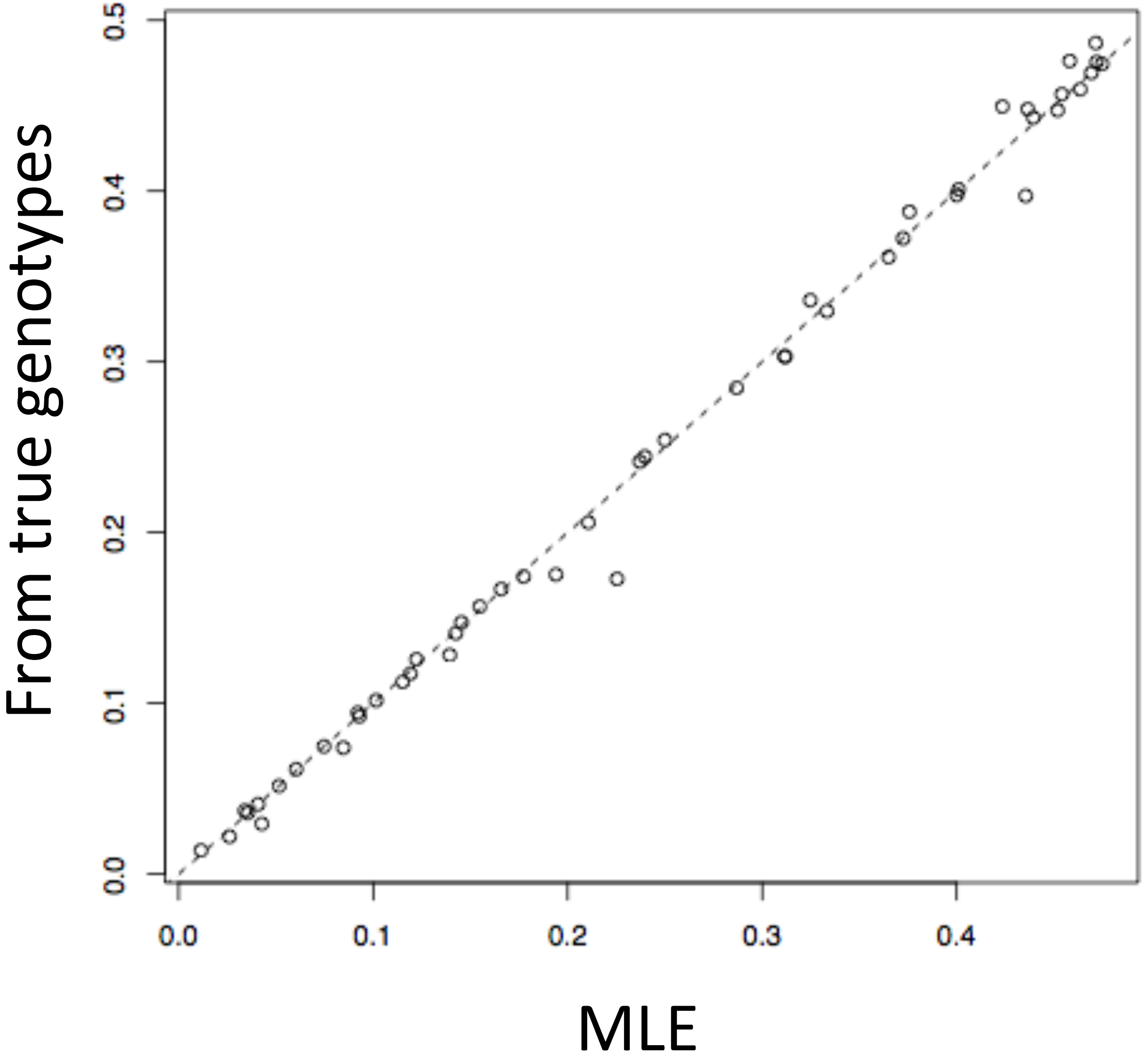
Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

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

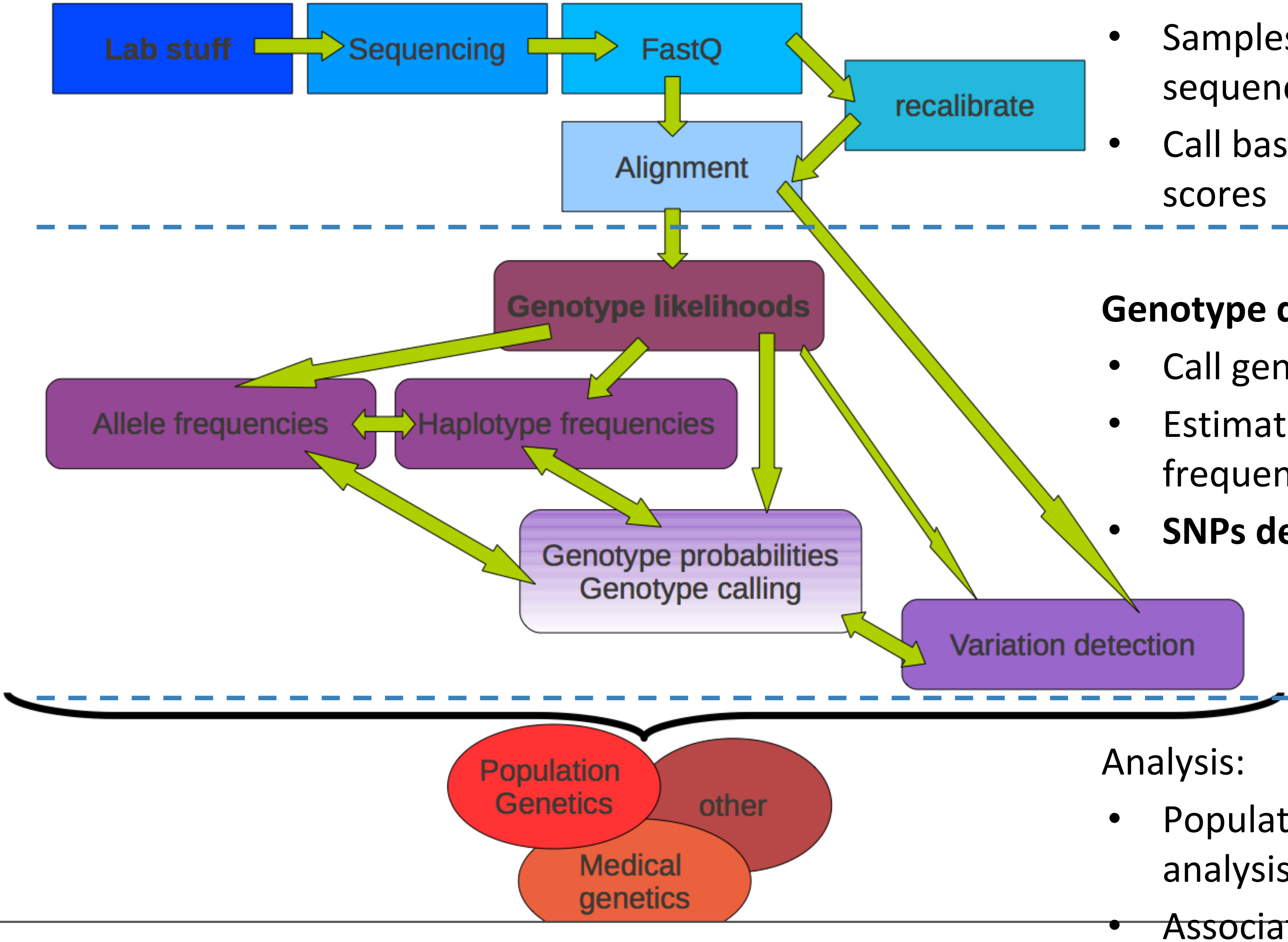
$$\hat{f} = \arg\max_p \prod_{i=1}^{N} p(D_i \mid f)$$

$$\hat{f} = 0.46$$

Allele frequency comparison



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

- 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
  - Assign as SNPs sites where the estimated allele frequency is above a certain threshold (e.g. ?)

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%

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

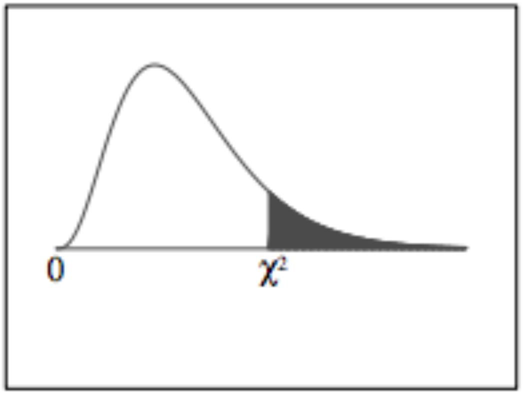
- Likelihood Ratio Test (**LRT**):

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

T is chi-squared distributed with 1 degree of freedom

SNP calling

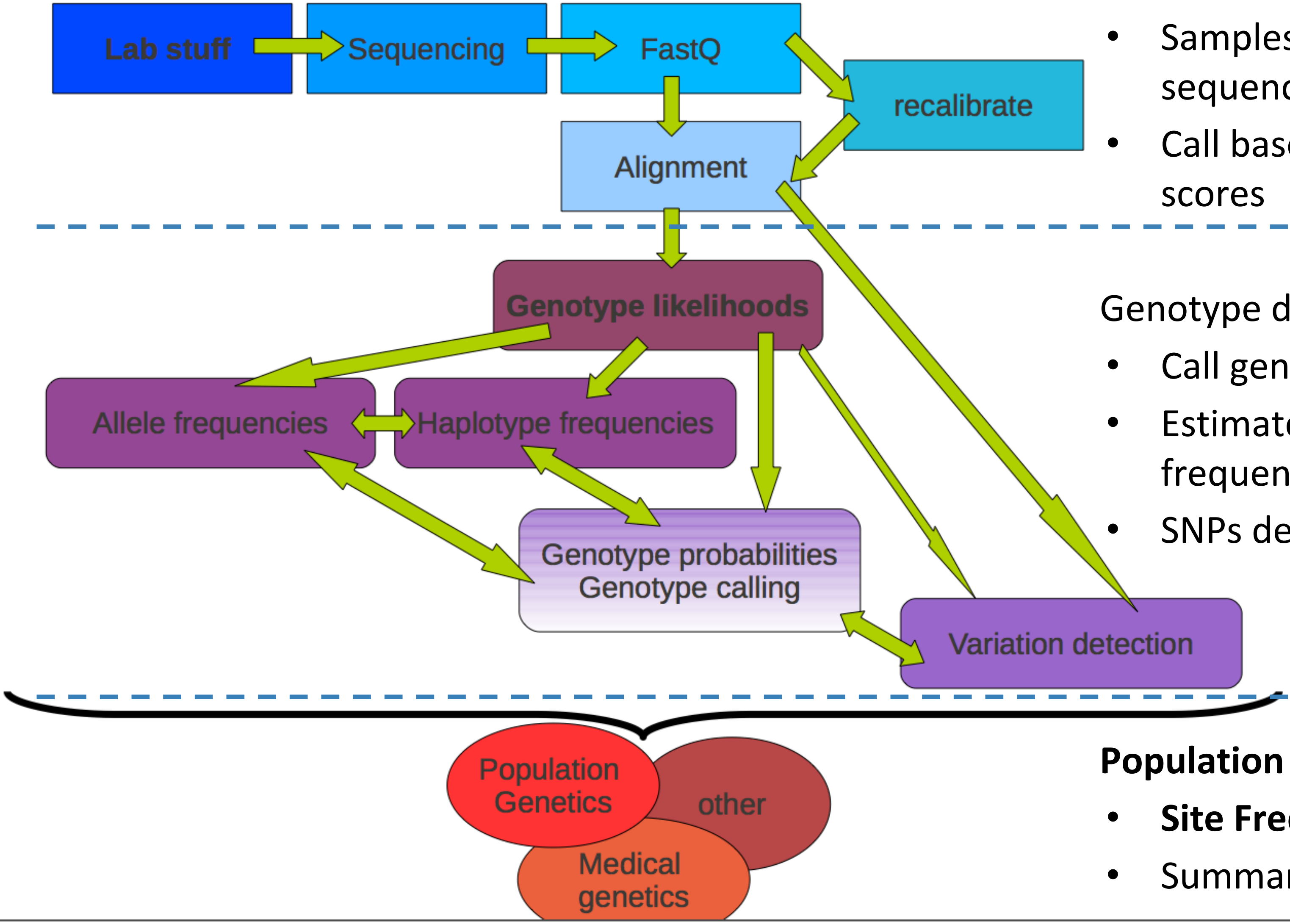
- Chi-square distribution table



http://sites.stat.psu.edu/~mga/401/tables

T	*p*-value
2.70	0.1
3.84	**0.05**
5.02	0.025
6.63	0.01
7.87	0.005

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)

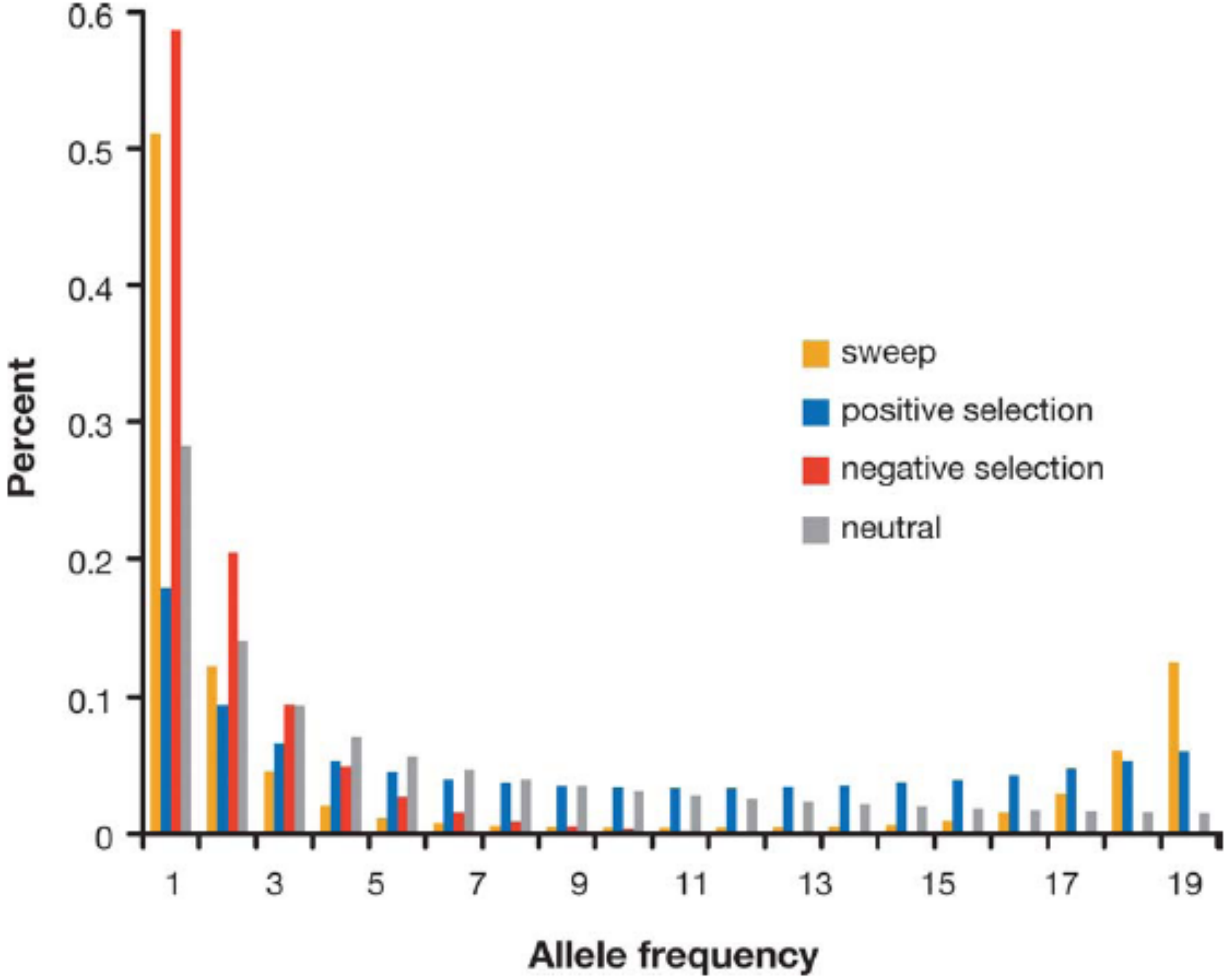
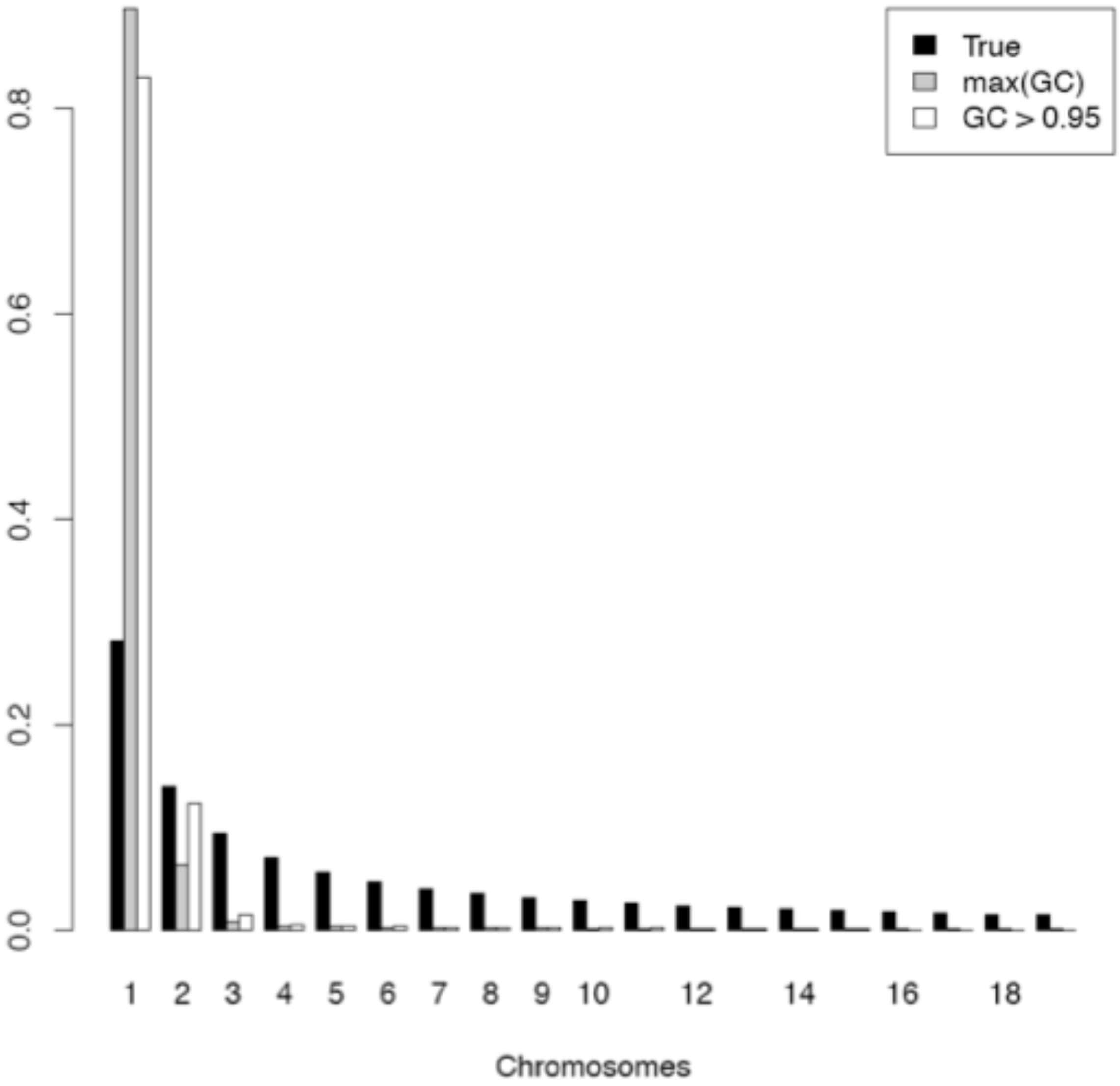


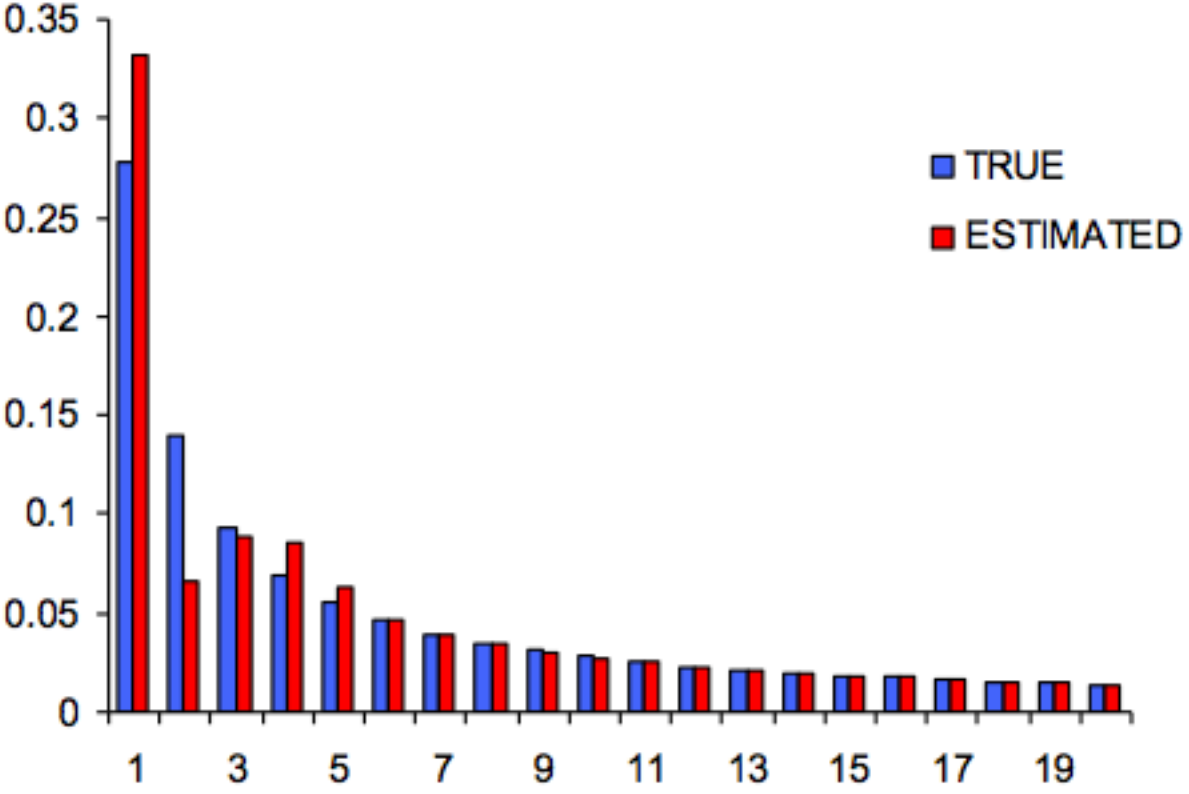
Figure 2

Effect of errors on SFS



Effect of errors on SFS

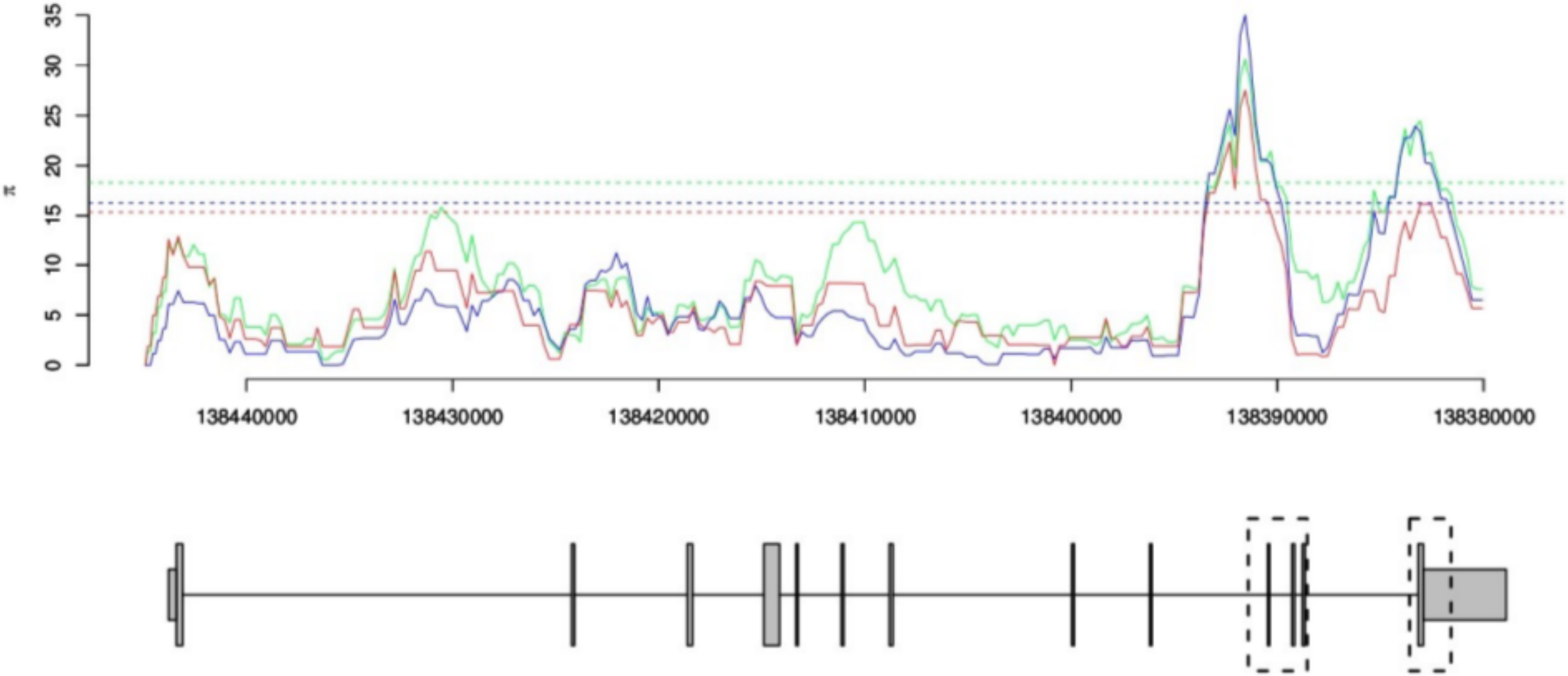
Using an ad hoc fixed cutoff for SNP calling…



can never produce unbiased estimates.

Effects of low-depth data

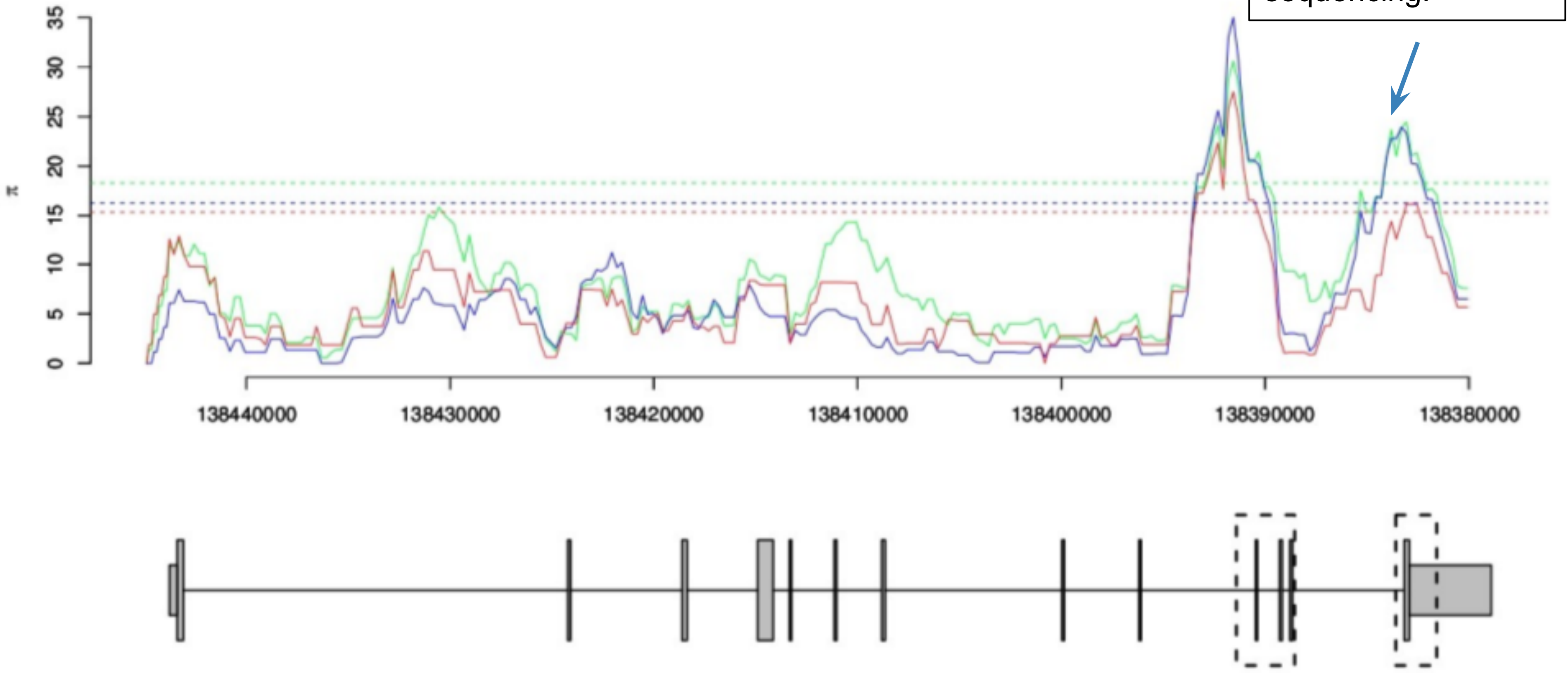
Nucleotide diversity scan using 1000 Genomes Project data (low-depth)



Cagliani et al. 2012

Effects of low-depth data

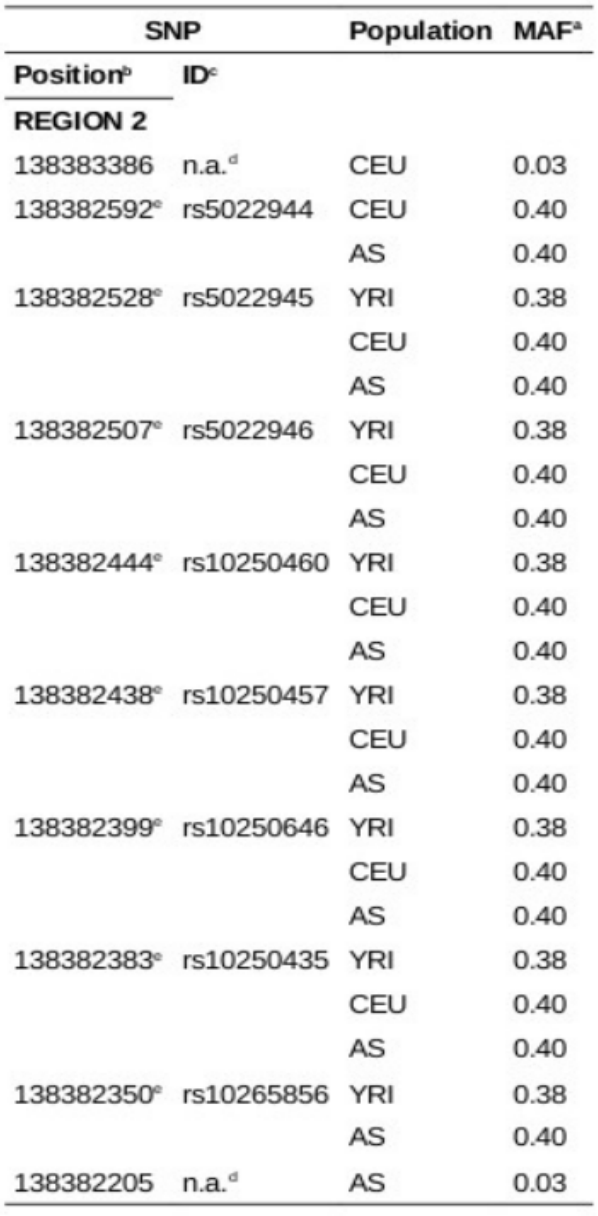
Nucleotide diversity scan using 1000 Genomes Project data (low-depth)



Highest peak based on Sanger sequencing!

Cagliani et al. 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

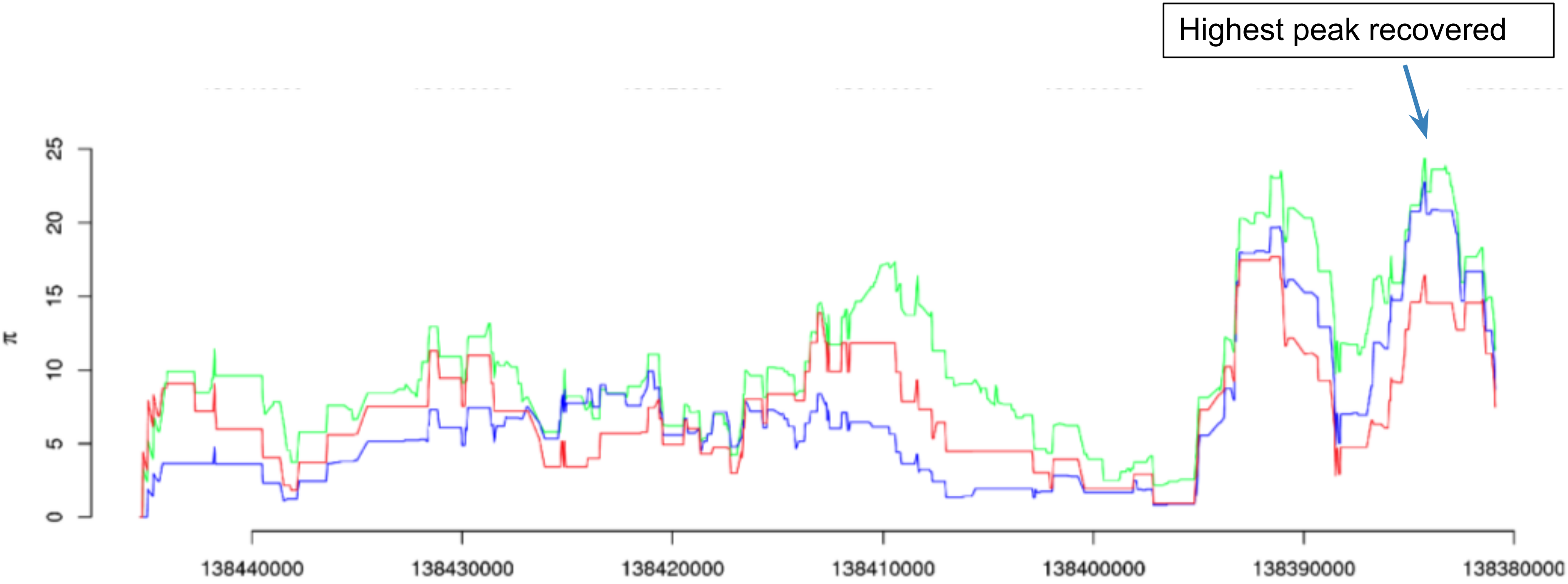
- Sanger: detected a total of 24 variants
- NGS: only 13

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

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

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

Masked data



- ○ Missing data
- ○ Unpredictable effects

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

- Parameterize the SFS, with *k* individuals

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

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:

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

Nielsen et al. 2012 PLoS One

ML estimation of the SFS

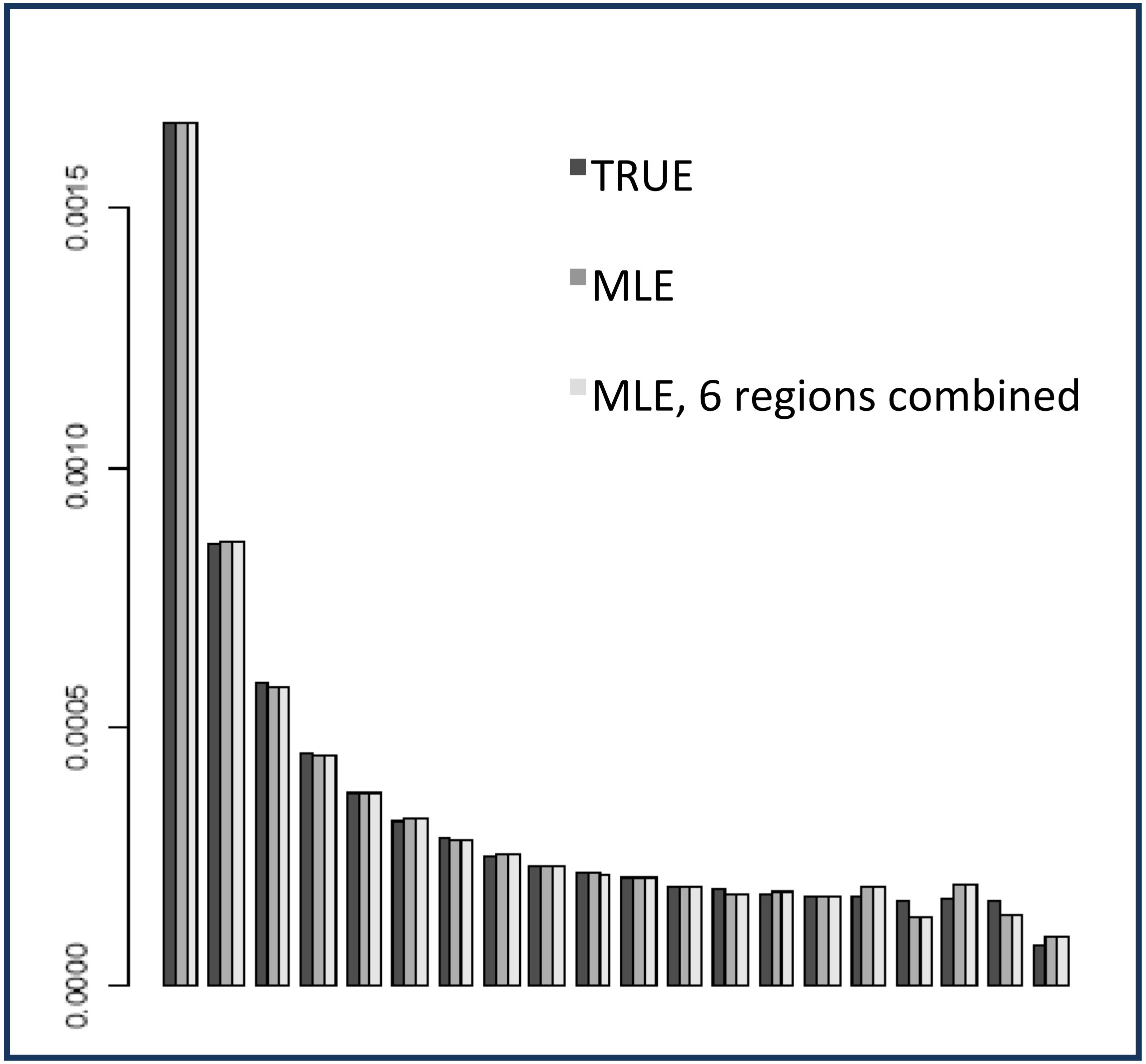
Summing across all unknown genotypes and multiplying the likelihood across sites.

- Likelihood function:

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

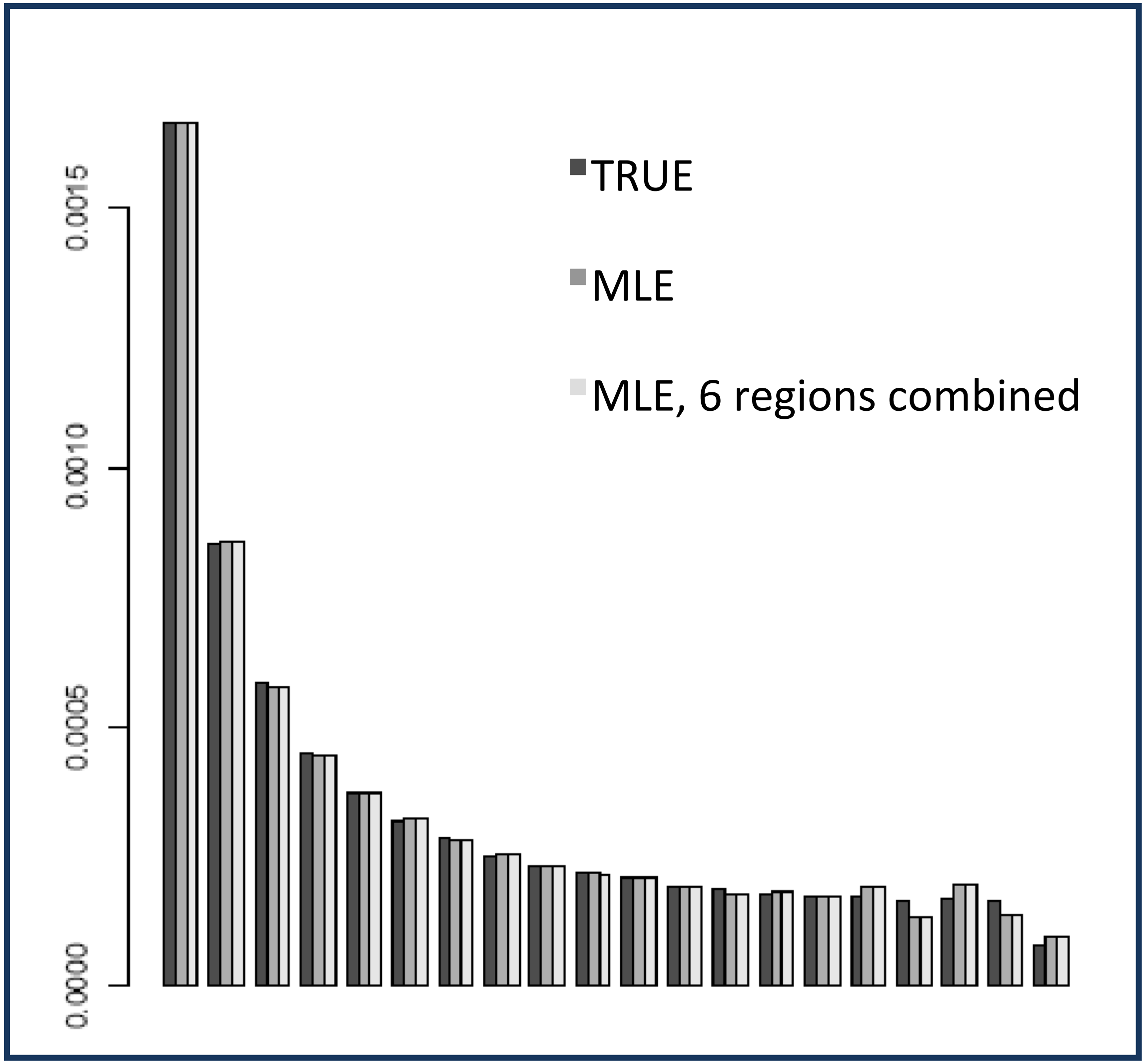
Nielsen et al. 2012 PLoS One

ML estimation of the SFS



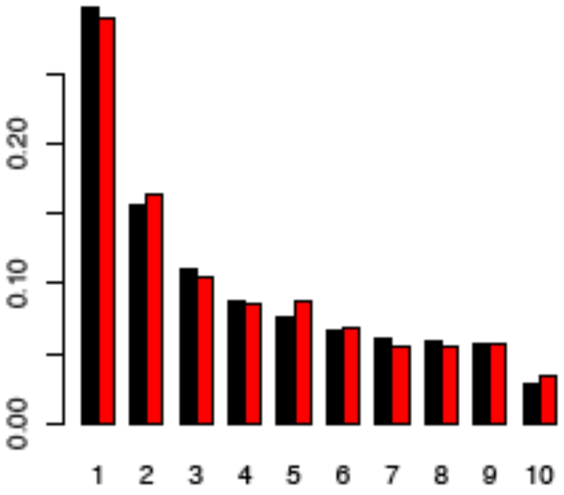
Simulated 30Mb
Error rate of 0.3%
Mean depth of 5X

ML estimation of the SFS



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

Mean depth of 1X:

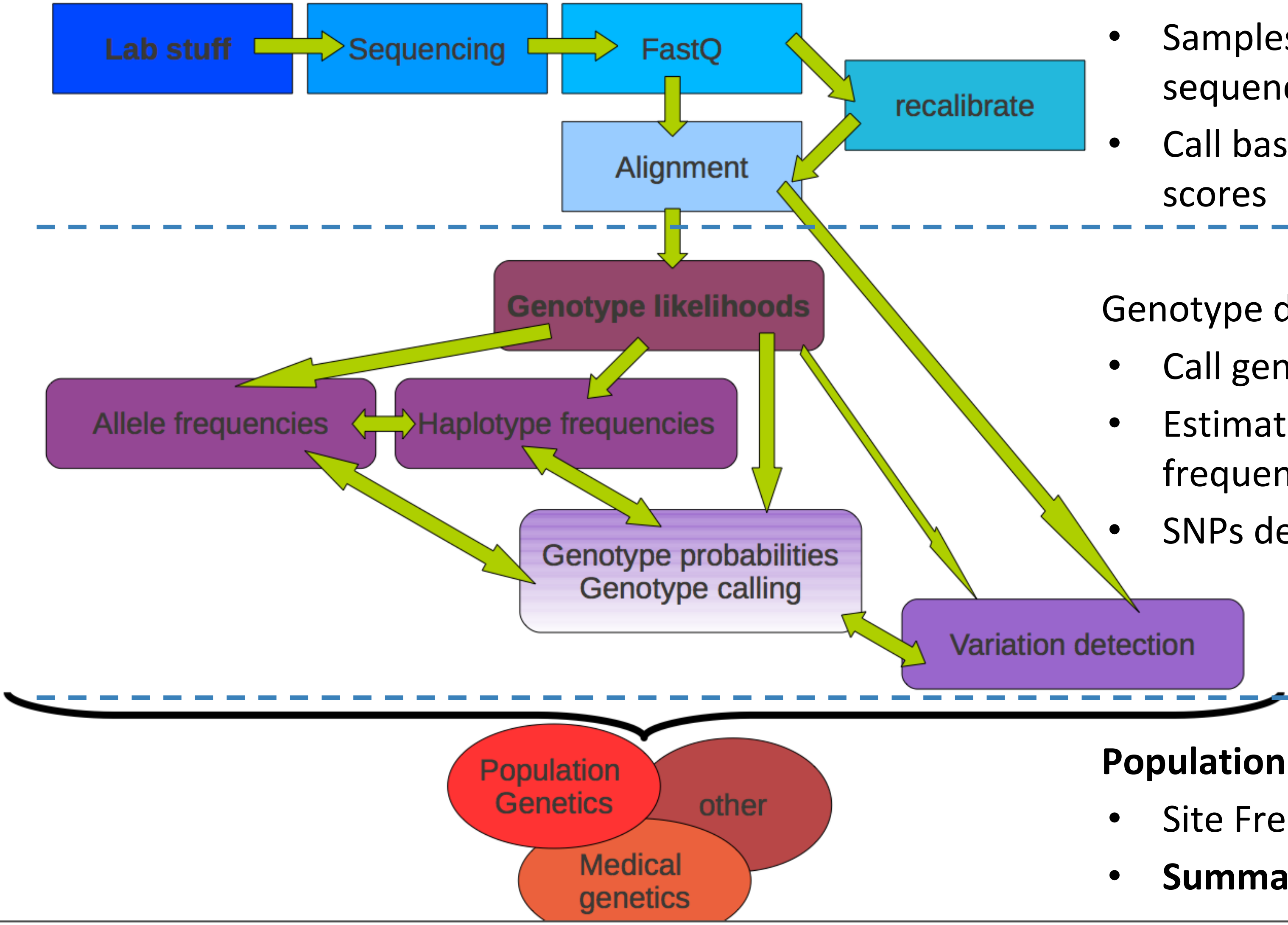


ML estimation of the SFS

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*

$$p(S_m = j \mid X) \propto \overset{\text{Likelihood}}{p(X \mid S_m = j)}\,\overset{\text{Prior}}{p(S_m = j)}$$

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*

$$p(S_m = j \mid X) \propto \overset{\text{Likelihood}}{p(X \mid S_m = j)}\,\boxed{\overset{\text{Prior}}{\underset{\textbf{Estimate of the overall SFS}}{p(S_m = j)}}}$$

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

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

- Estimating allele frequency

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

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

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

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

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

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

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

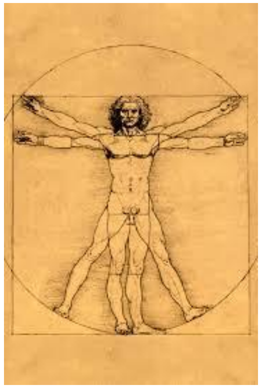
$$E[D] =$$

Nucleotide diversity

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

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

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

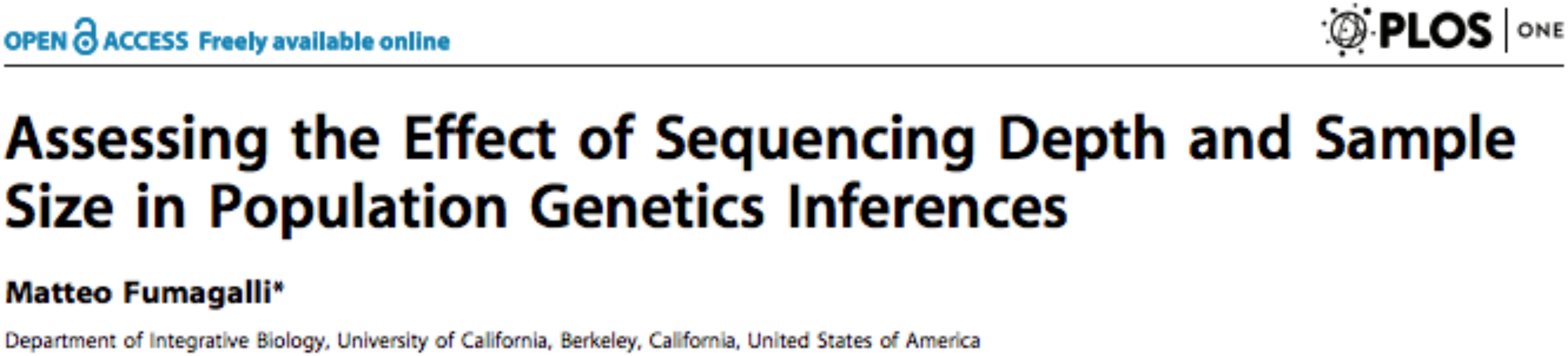
- Nielsen *et al.* Nat Rev Genet 2011 (21587300)
- Li H. Bioinformatics 2011 (21903627)
- Kim *et al.* BMC Bioinformatics 2011 (21663684)
- Fumagalli M. PLoS One 2013 (24260275)

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

Practical exercises

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

Study discussion



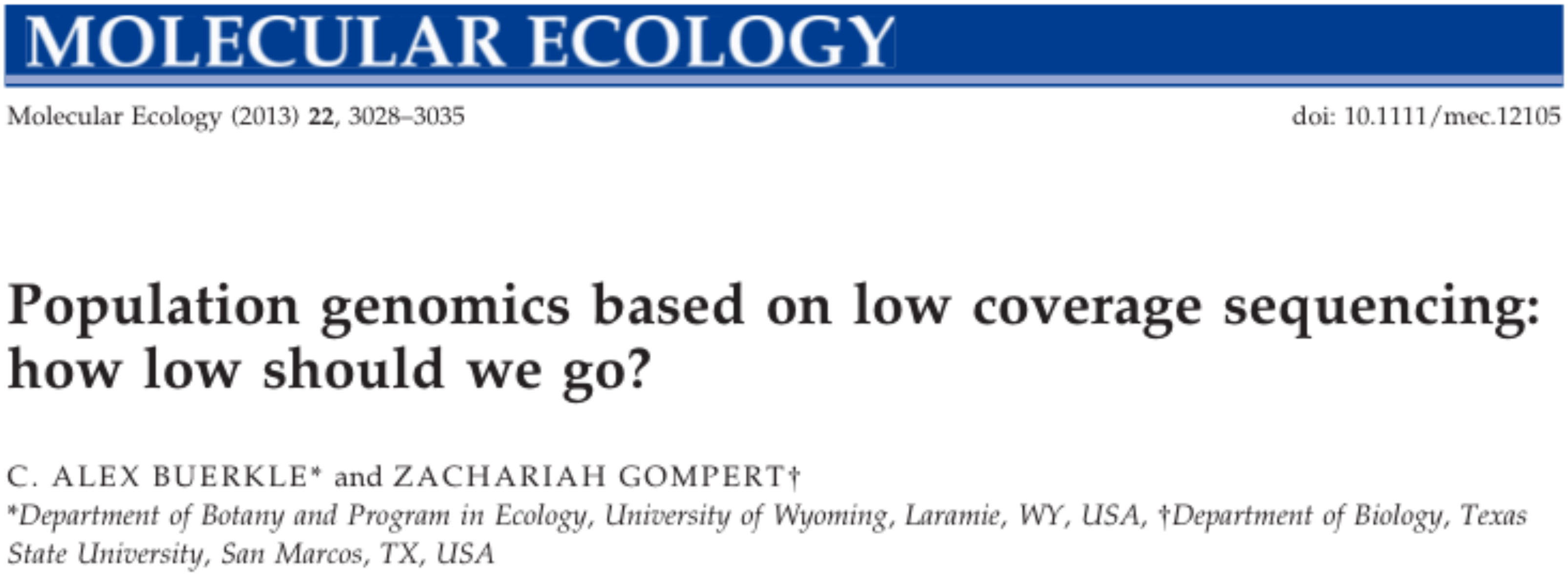
OPEN ACCESS Freely available online

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



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*