Evolutionary genomics
Data analysis module – Day 2

Paper discussion

SNP calling and advanced methods for evolutionary inferences from NGS data

April 13th-17th 2015

Paper 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

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 24 April 2012
doi: 10.3389/fgene.2012.00066

Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data

Jacob E. Crawford and Brian P. Lazzaro*

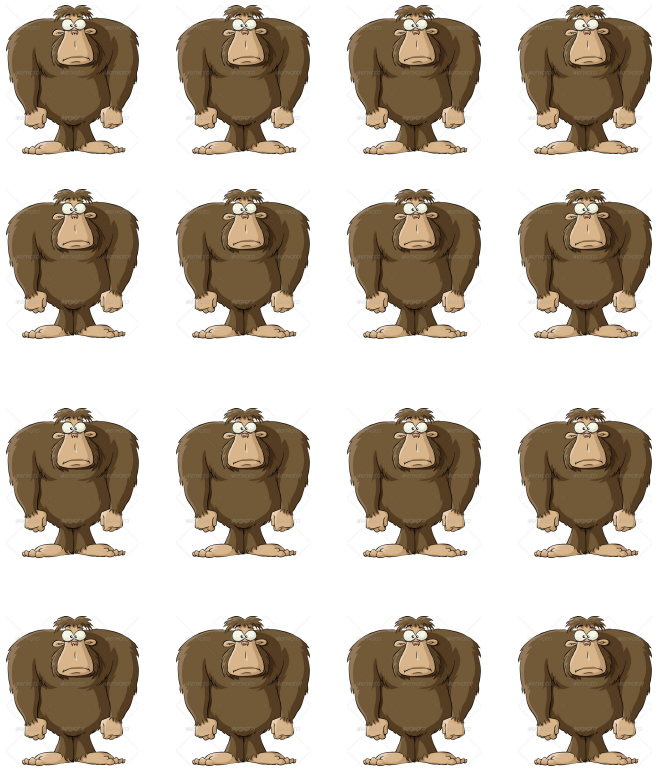
Department of Entomology, Cornell University, Ithaca, NY, USA

Experimental design

- You discovered a new species!

Experimental design

Population of 1,000 individuals

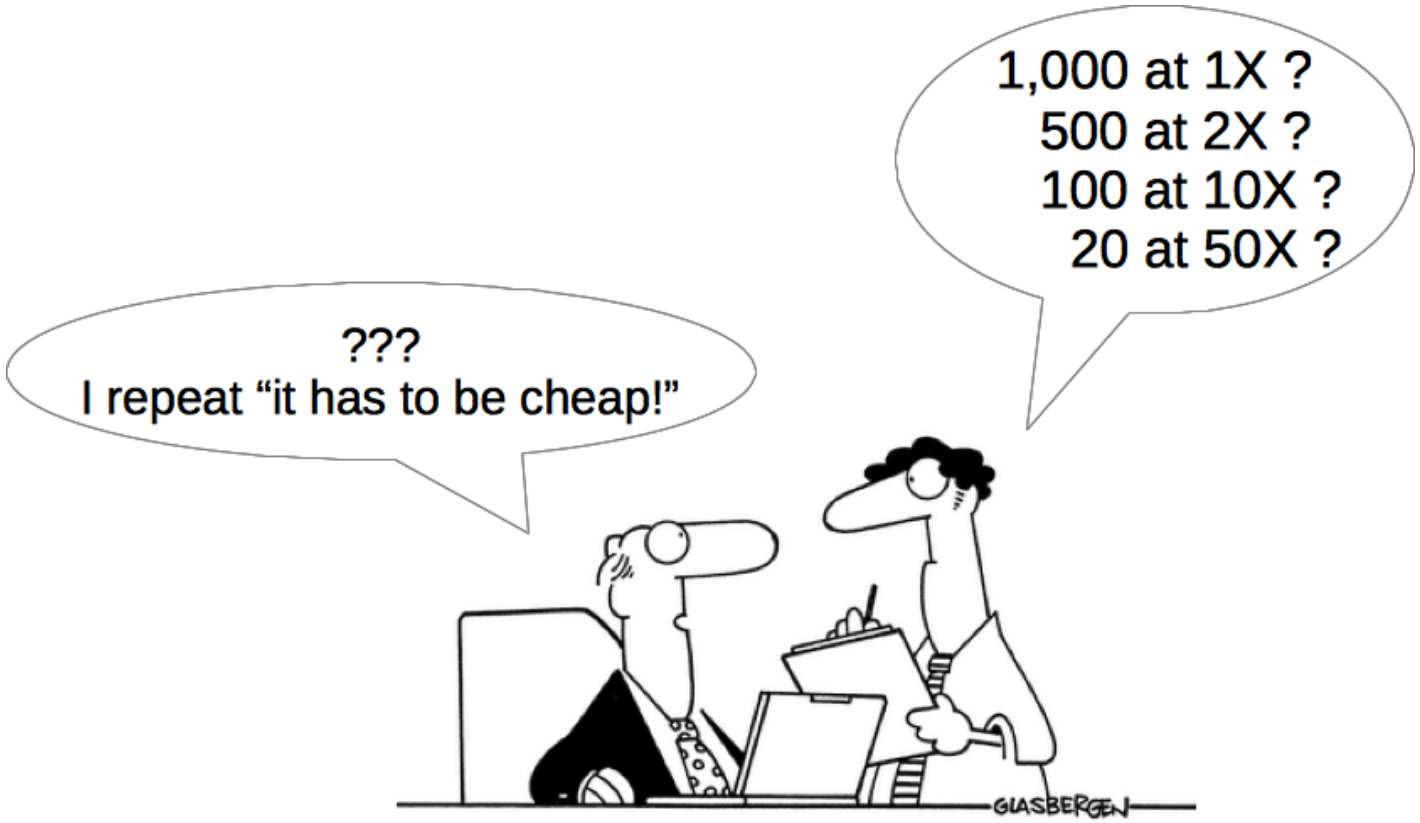


Experimental design

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

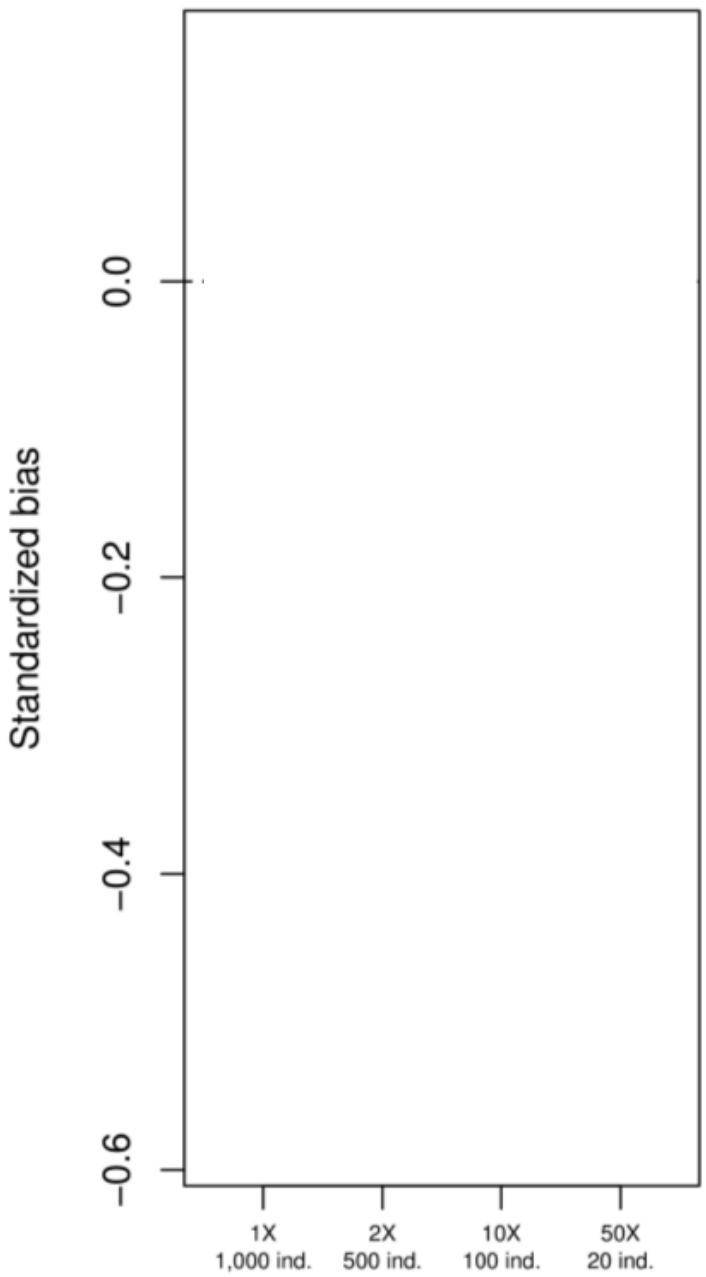
$$S = \sum_{s=1}^{L} I_s \qquad (1)$$

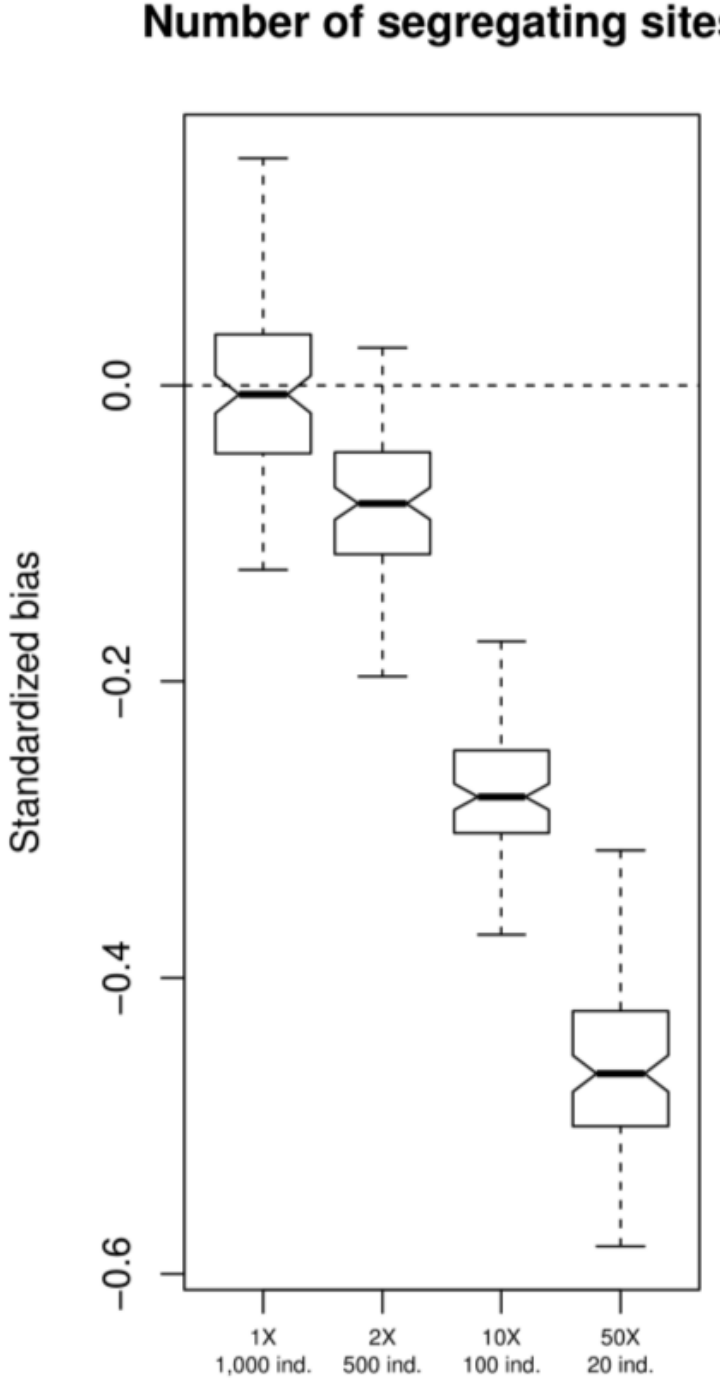
where I_s is an indicator function equal to 1 when at least one individual is heterozygous at site s, and 0 otherwise, and

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

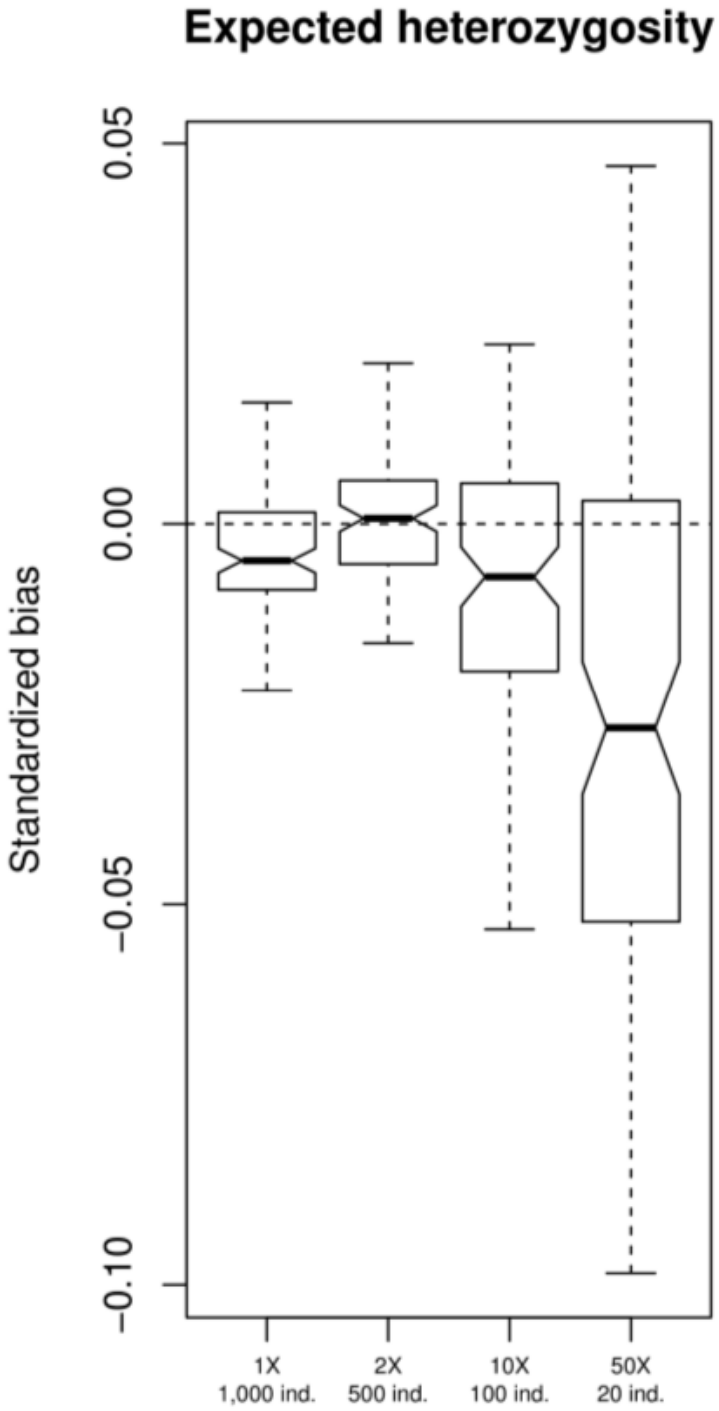
where f_s is the reference allele frequency for a site, s, in the sample.

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

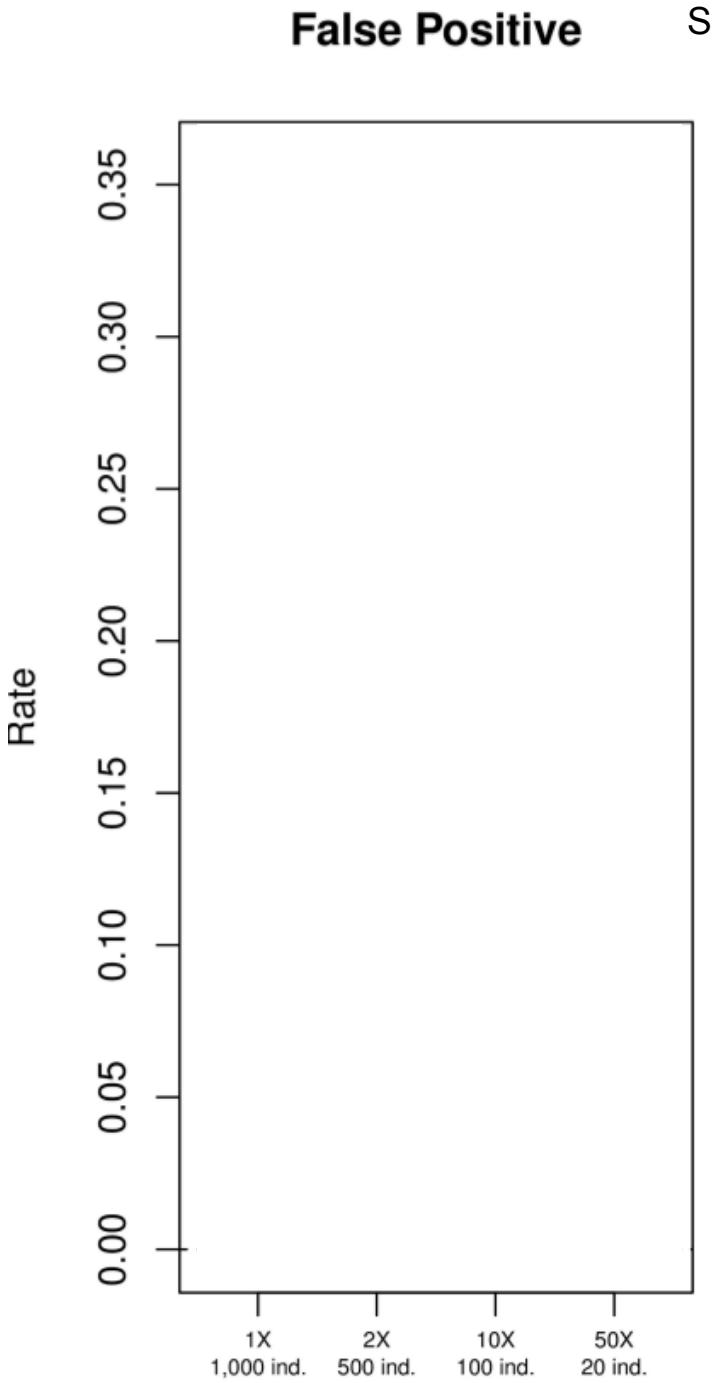

Number of segregating sites
Standardized bias
0.0
−0.2
−0.4
−0.6
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

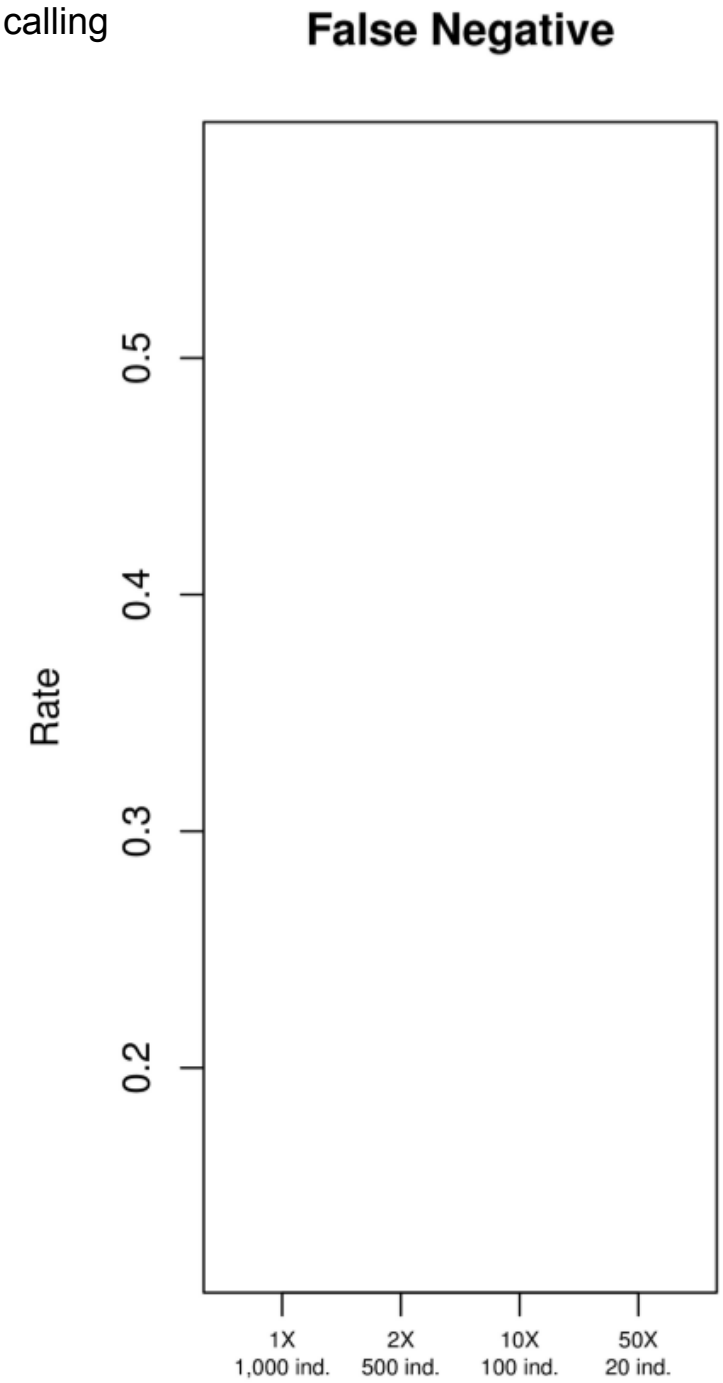

Number of segregating sites
Standardized bias
0.0
−0.2
−0.4
−0.6
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

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.

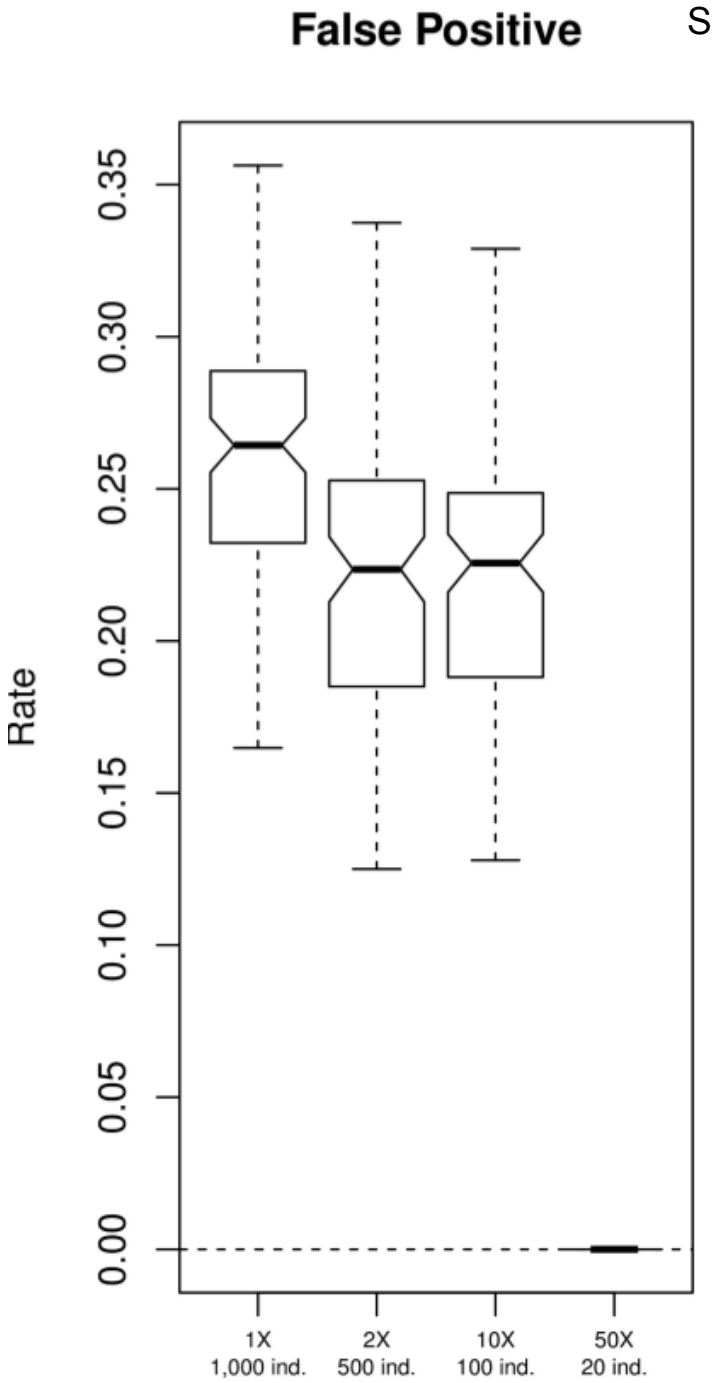

Expected heterozygosity
Standardized bias
0.05
0.00
−0.05
−0.10
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

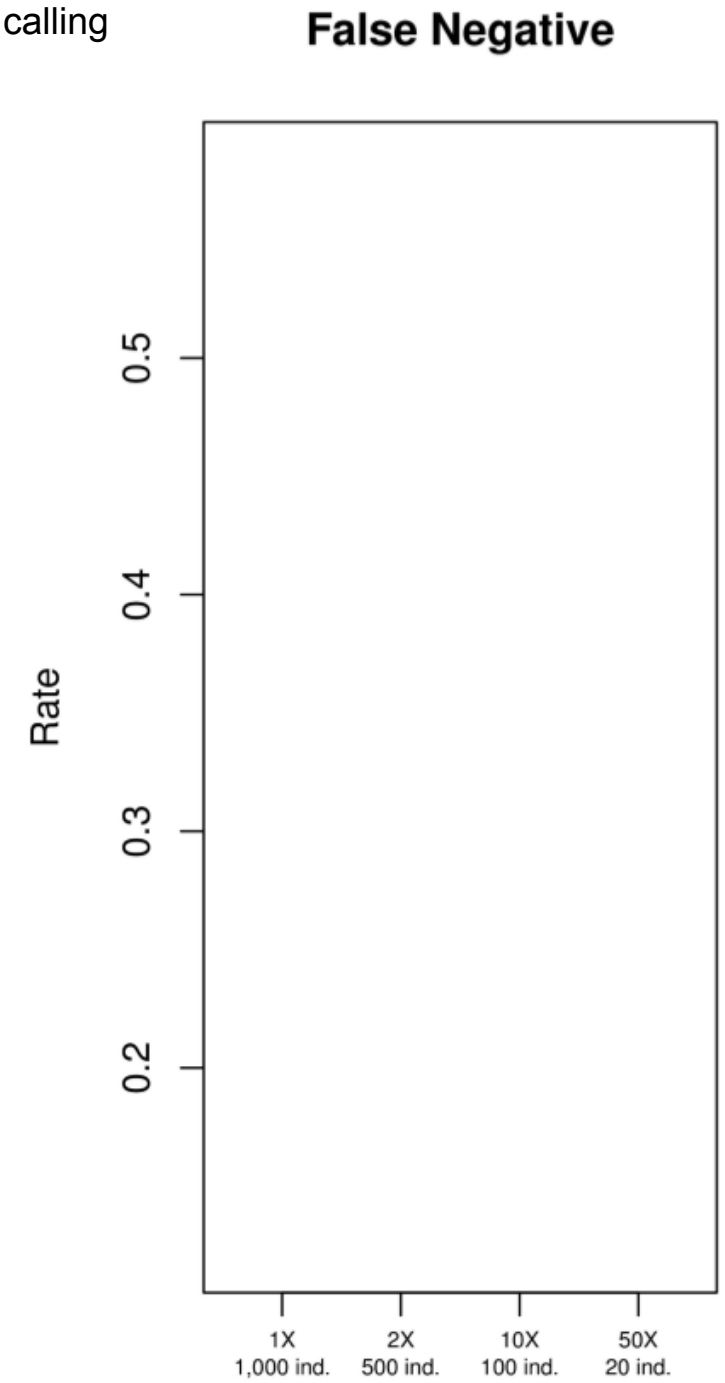
SNP calling


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.

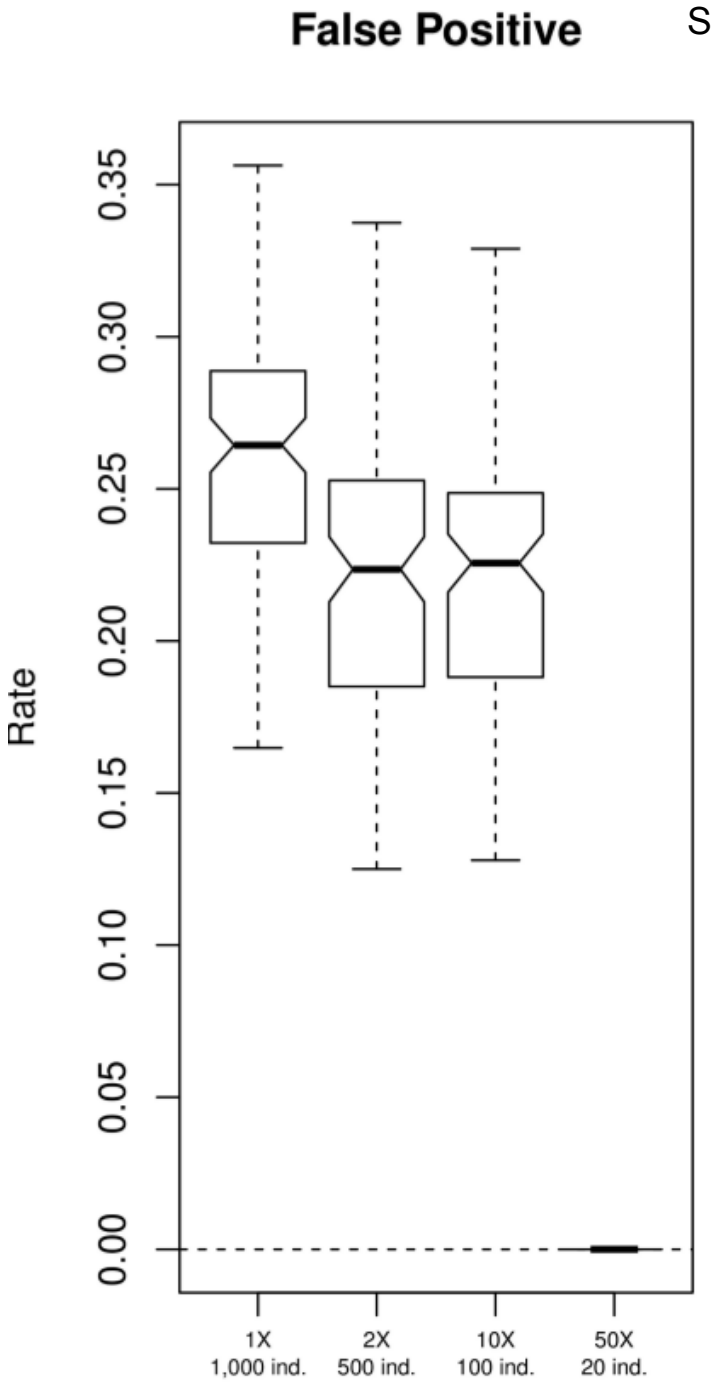

False Negative
Rate
0.2
0.3
0.4
0.5
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

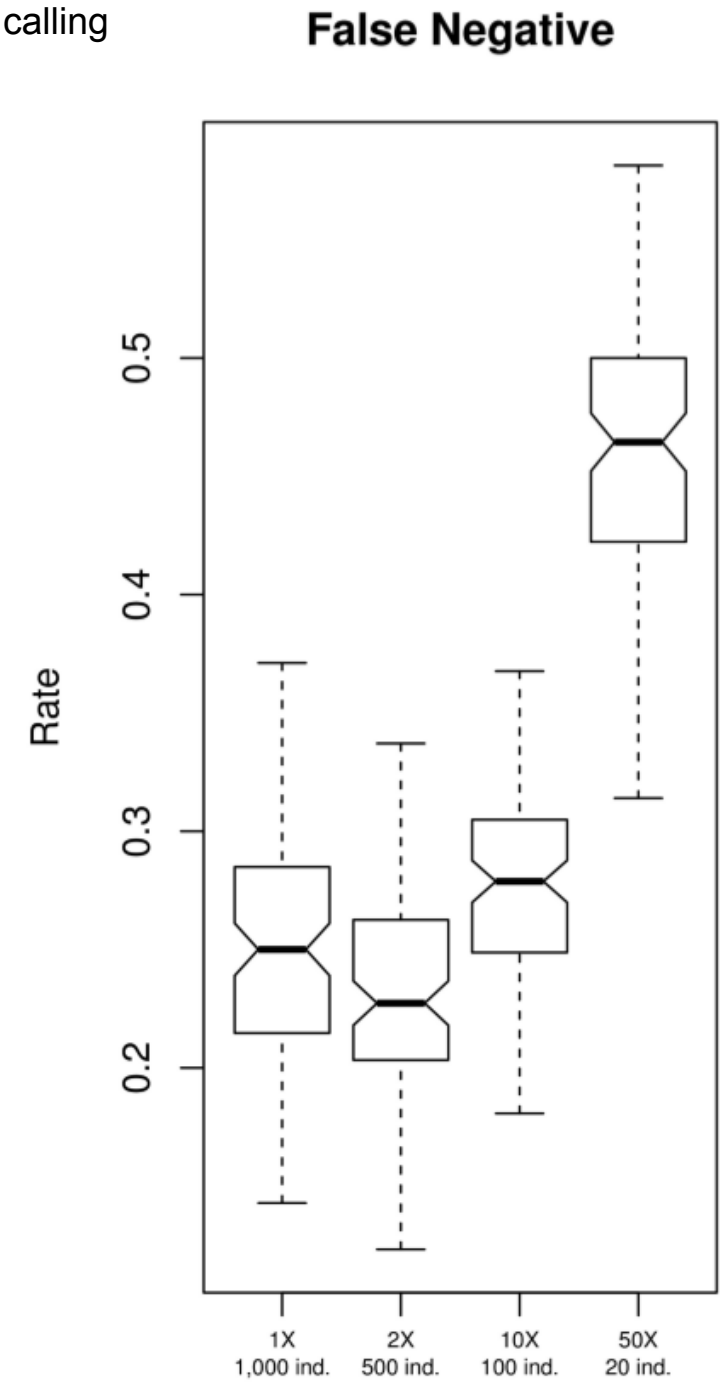
SNP calling


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.

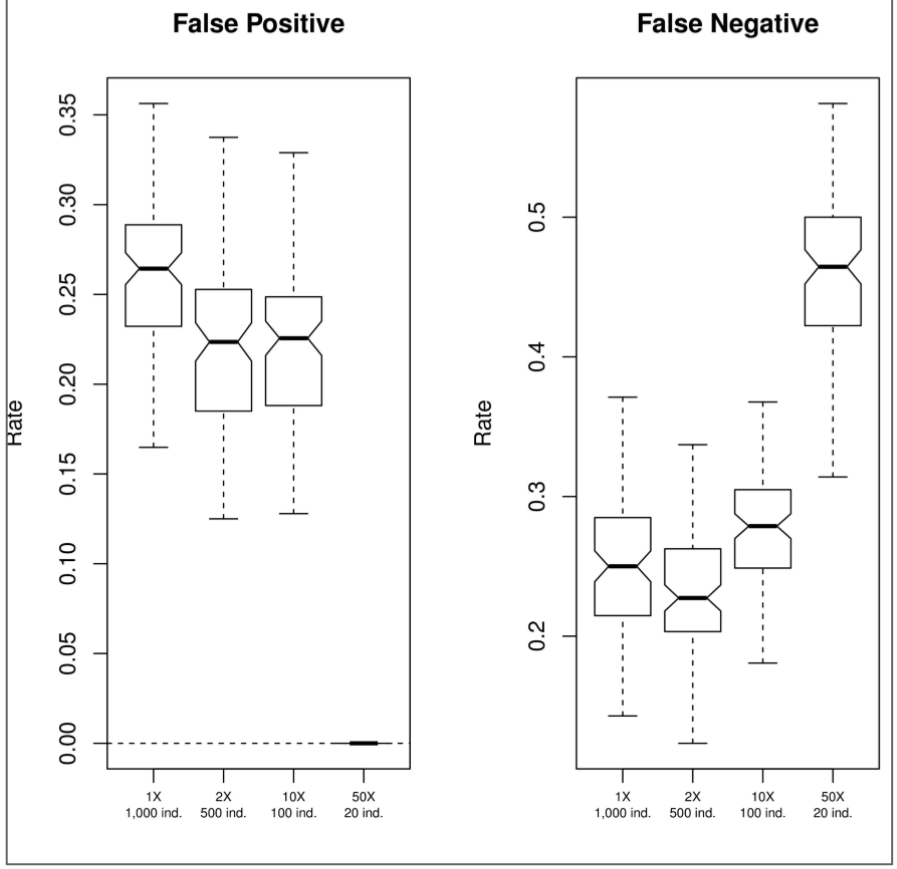
SNP calling


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.

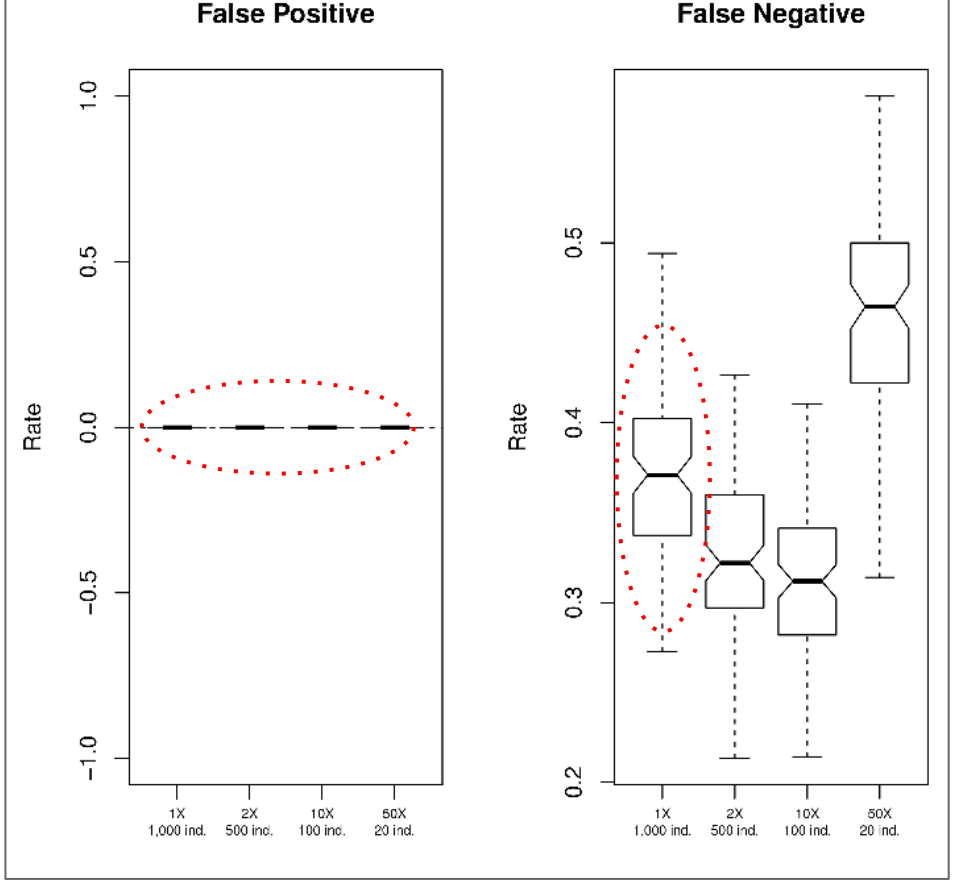

False Negative
Rate
0.2
0.3
0.4
0.5
1X
1,000 ind.
2X
500 ind.
10X
100 ind.
50X
20 ind.

Question for discussion - 1

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

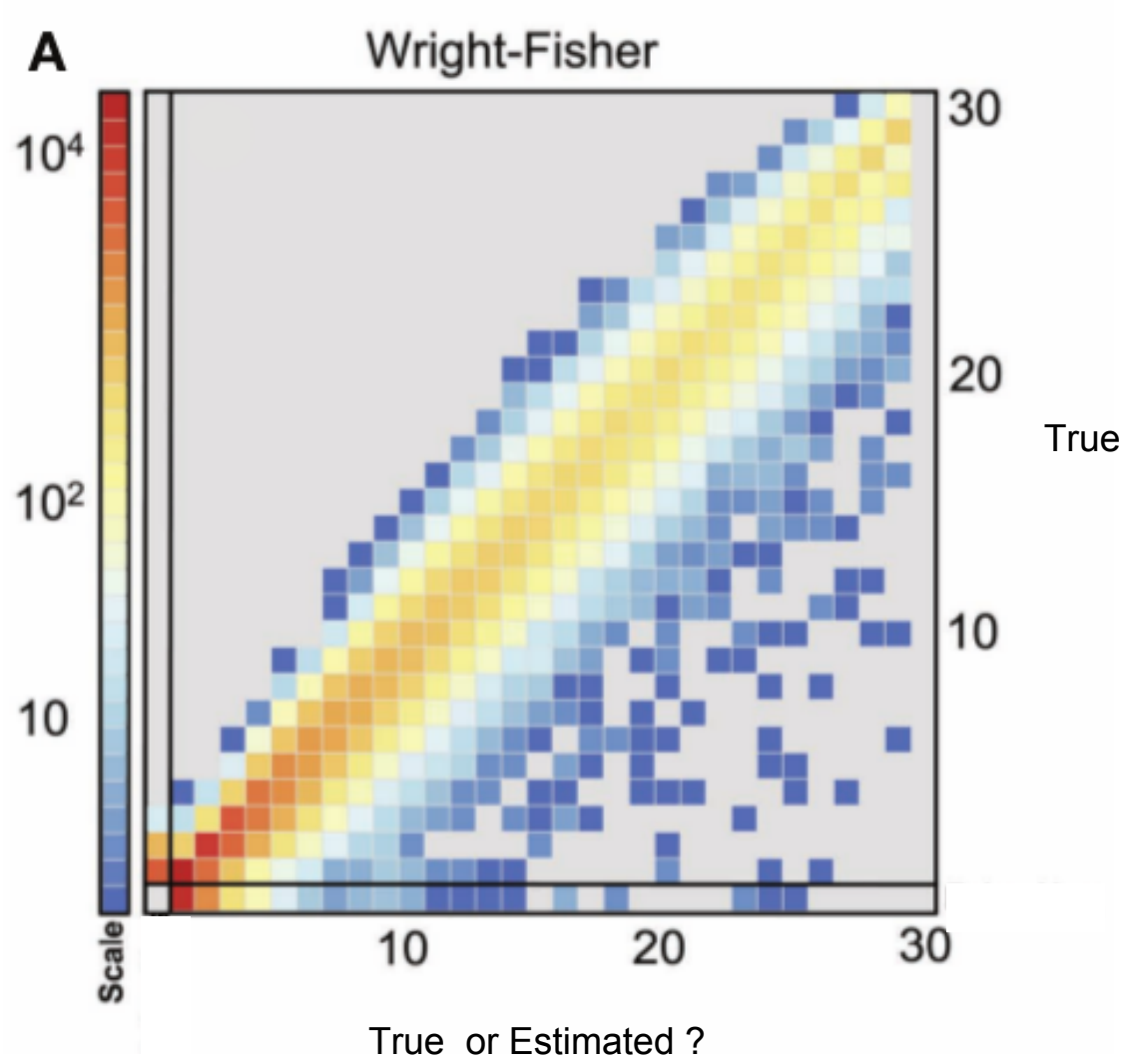


SNP is assigned if ?



Question for discussion - 2

Estimation of allele frequencies



True or Estimated ?

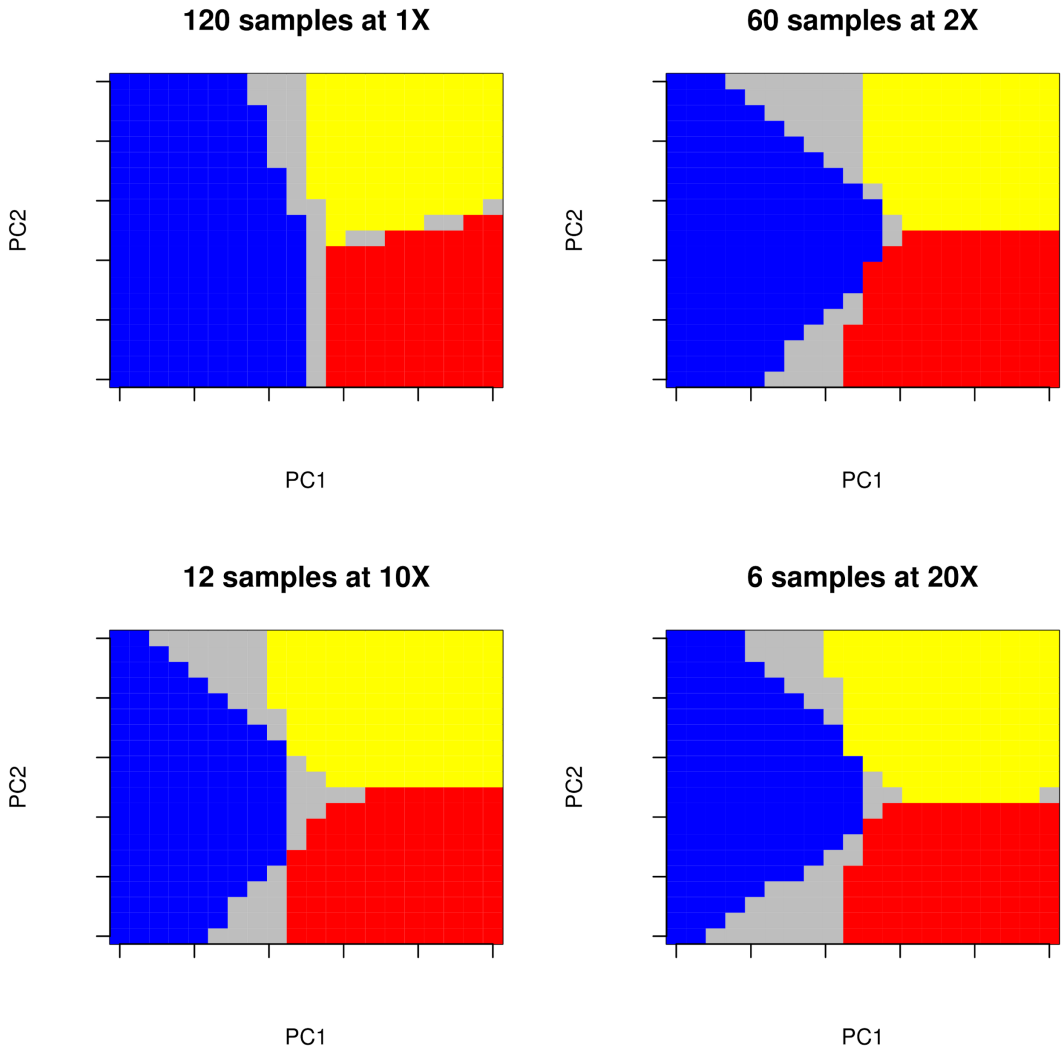
Conclusions

- 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.

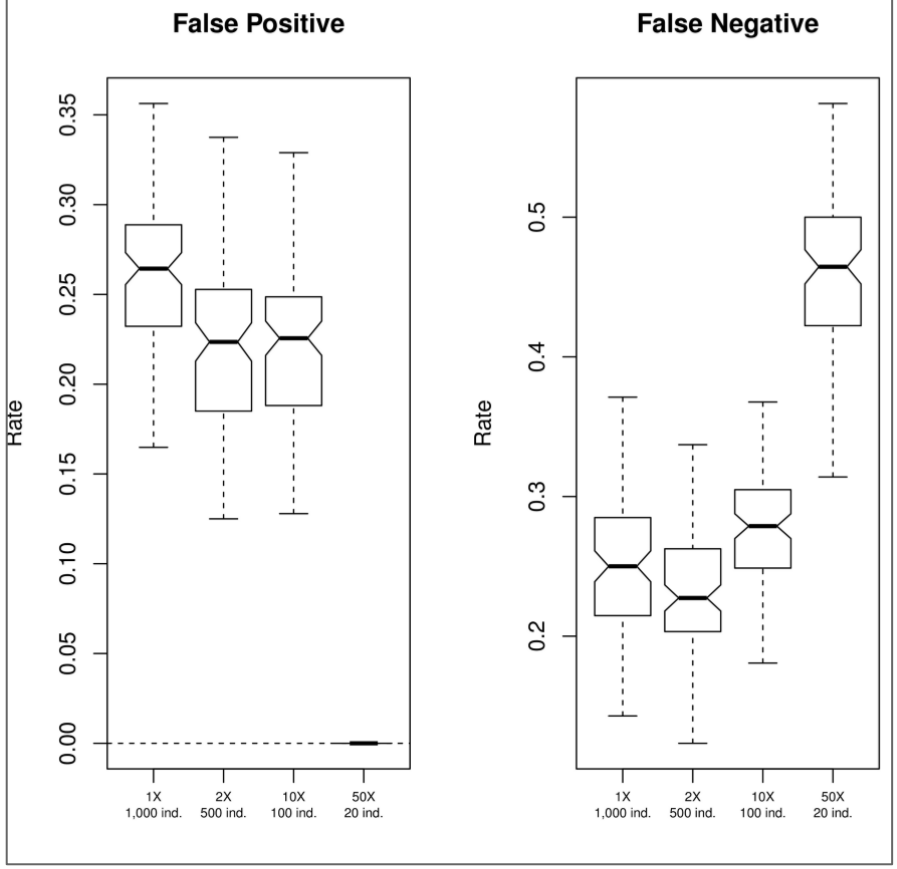
- State-of-the-art statistical methods to estimate genetic variation from NGS data should be adopted in all population genetics studies using low-medium coverage sequencing data.

Spatial structure

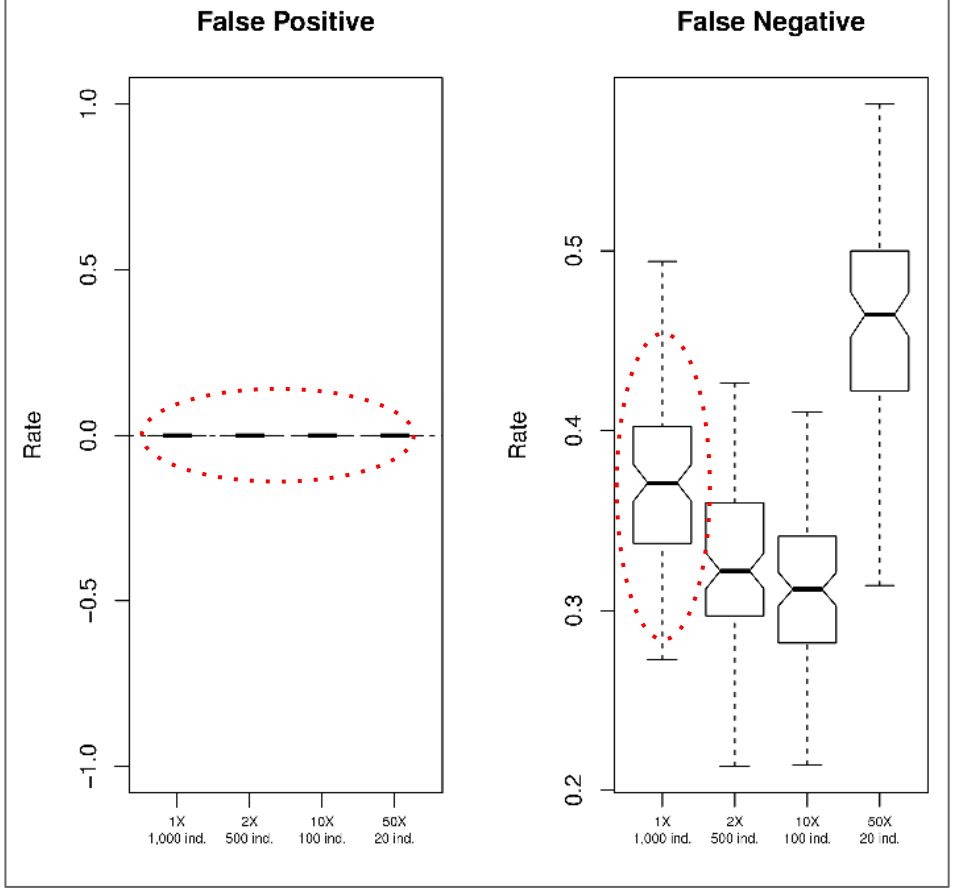


Question for discussion - 1

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

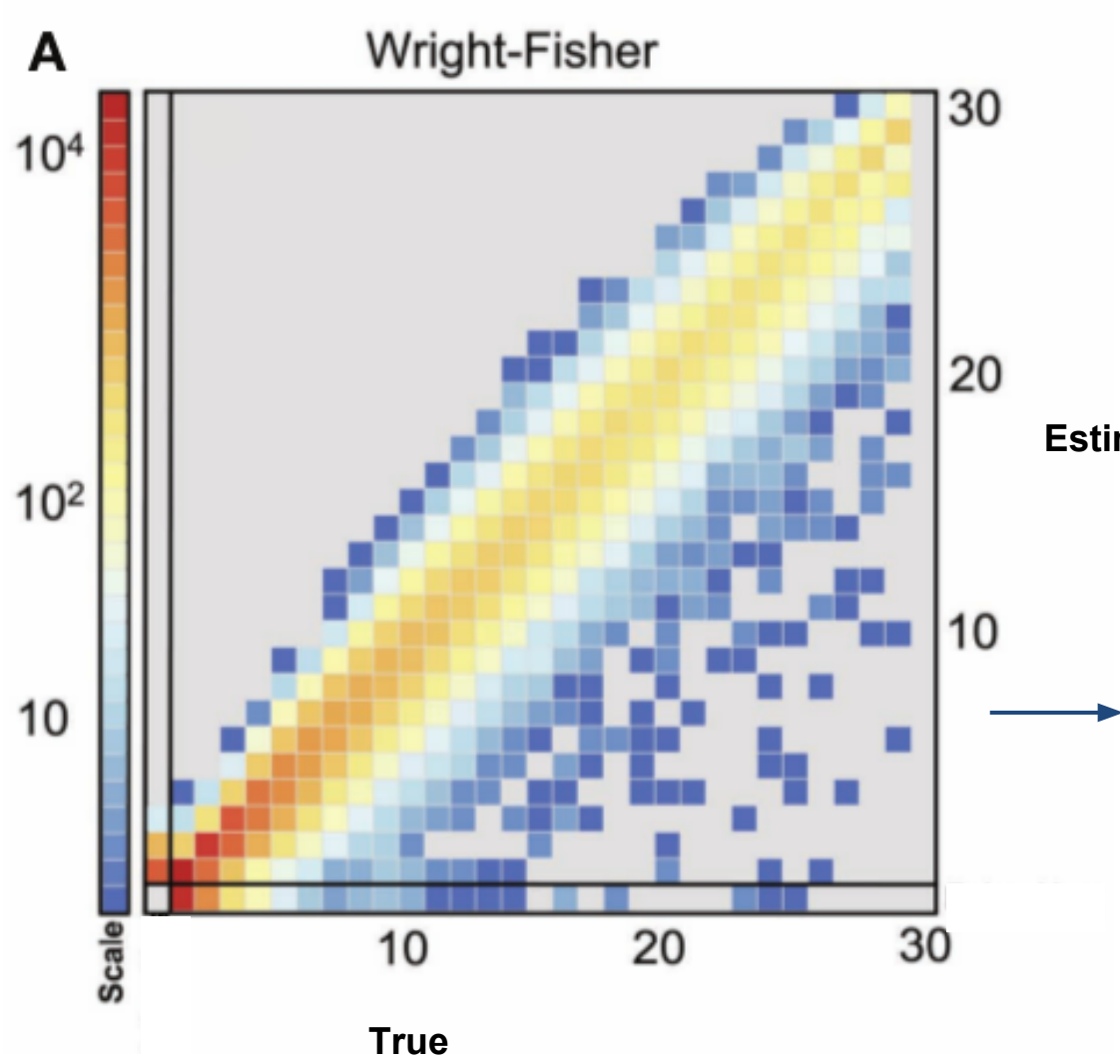


SNP is assigned if the probability of being variable is > 0.95



Question for discussion - 2

Estimation of allele frequencies



values tend to be under-estimated