Genomic Imputation in ultra low coverage sequencing data of Ashkenazi Jews

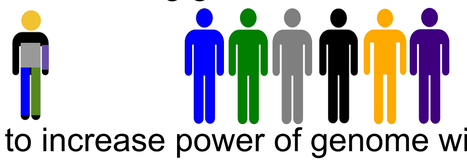
Noam Bar[1] , Shai Carmi[2]

[1]The Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel

[2]The Braun School of Public Health and Community Medicine, Hebrew University, Jerusalem, Israel

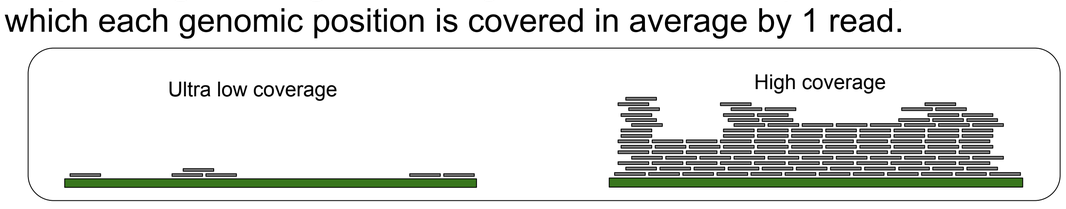
1 Introduction

Genomic imputation - Inferring genetic markers that are not directly genotyped.



Often used in order to increase power of genome wide association scans. Especially efficient in isolated populations, such as **Ashkenazi Jews**.

Sequencing coverage - Coverage of 1x refers to sequencing data, in which each genomic position is covered in average by 1 read.



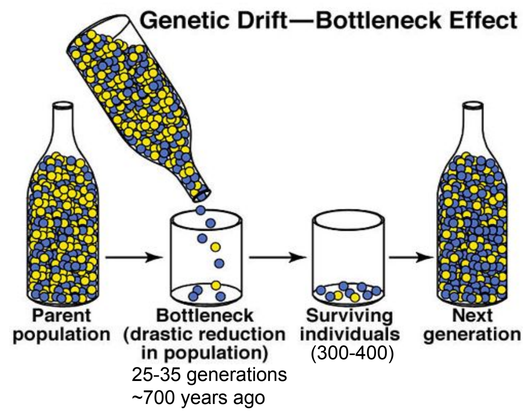
Goal - Develop a tool for genomic imputation of ultra low coverage sequencing data in Ashkenazi Jews.
We aim to successfully impute data sequenced at coverage as low as 0.1x

2 Data

128 Ashkenazi jewish individuals sequenced at high coverage, making up the **reference panel**. **>12,000,000** SNPs, covering the 22 chromosomes. Average distance between two following SNPs is ~270 bps.

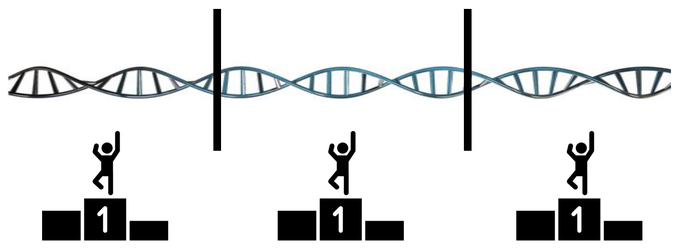
Ashkenazi Jews share long genomic segments at much higher rate than other populations.

Ashkenazi Jews



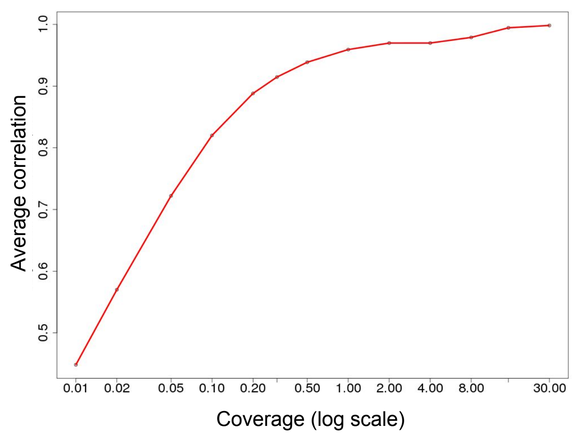
3 Find most similar individuals

1. Dividing the genome into a set of overlapping windows.
2. Ranking the individuals in the reference panel for each window using the likelihood ratio test.
3. Choosing individuals which ranked best, out of a bimodal distribution.
4. Those individuals will serve as the surrogate parents of the low coverage genome at each region.

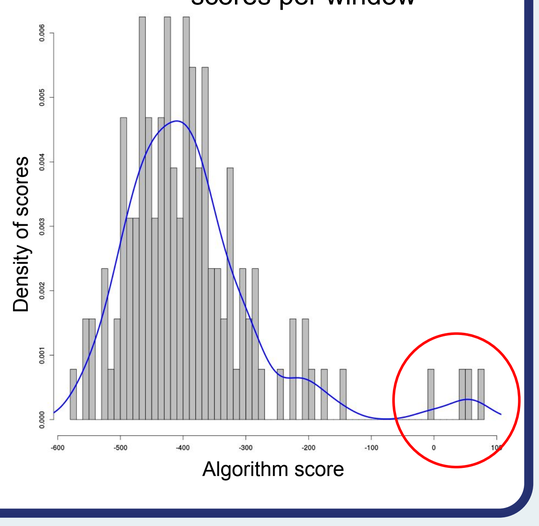


score: $\log\left(\frac{P(X|related)}{P(X|not\ related)}\right)$

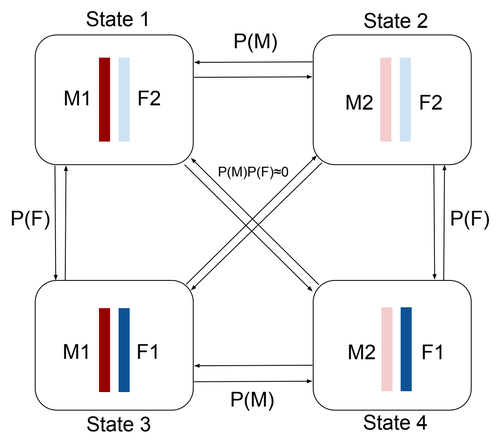
Average correlation between low coverage scores and high coverage score



Bimodal distribution of scores per window

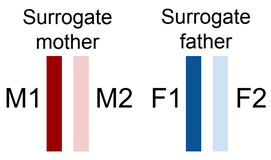


4 Using HMM algorithms



Transitions
P(M) - Probability of phase switch at surrogate mother
P(F) - Probability of phase switch at surrogate father
d = Distance between the two variants
270 bps - Average distance between variants
P(F)=P(M) =1-0.99^(d/270)

Reference's Haplotypes

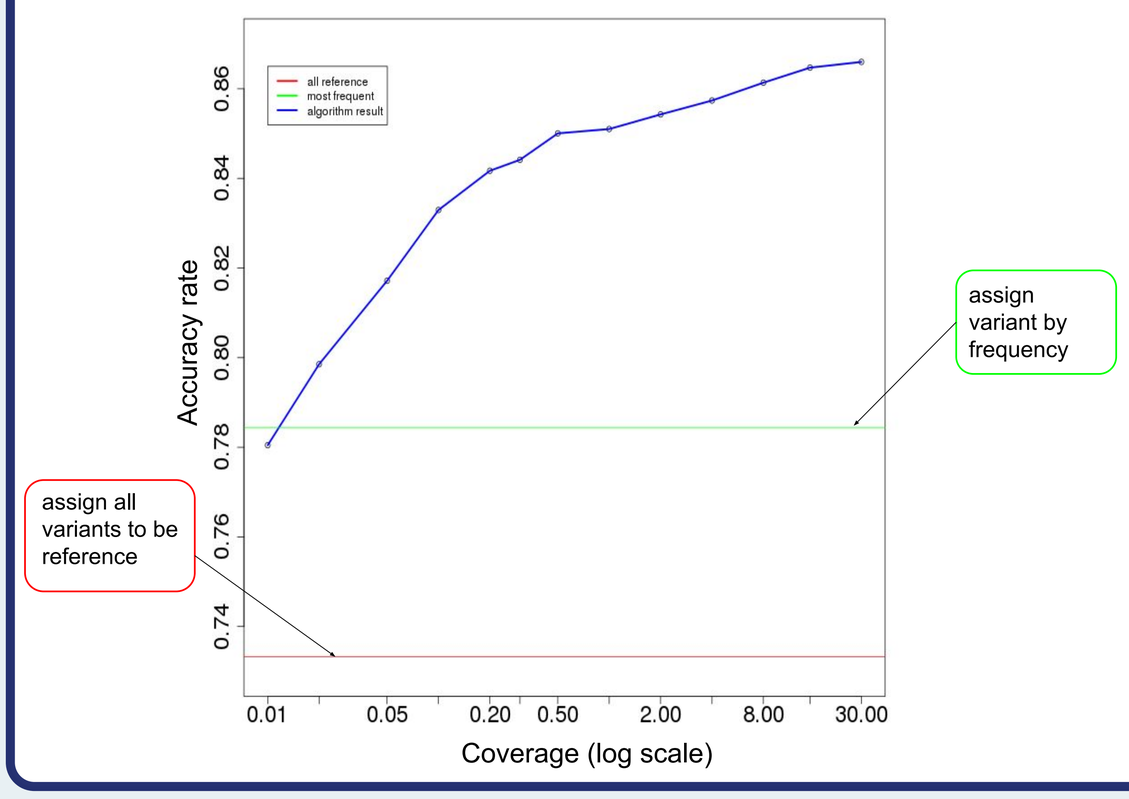


Emissions
The number of reads at the site: n
Sequencing error probability: ε
n_A reads of allele A, n_B reads of allele B

Induced genotypes at shared haplotypes	Probability of observation
AA	$\binom{n_A+n_B}{n_A}(1-\epsilon)^{n_A}\epsilon^{n_B}$
AB	$\binom{n_A+n_B}{n_A}2^{-(n_A+n_B)}$
BB	$\binom{n_A+n_B}{n_A}(1-\epsilon)^{n_B}\epsilon^{n_A}$

5 Imputation results

Accuracy rate of imputing algorithm using vs coverage

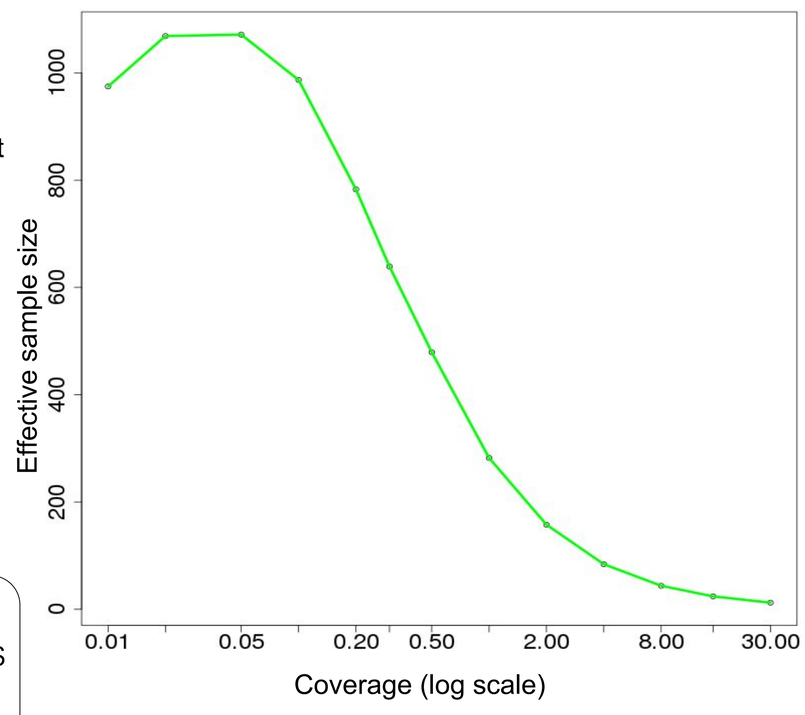


6 Effective sample size for a fixed budget

Effective sample size vs coverage for a fix budget of 100,000$

The effective sample size (ESS) is calculated as follows:
N - number of individuals sequenced at low coverage.
r - The correlation between the genotype as returned by the imputing algorithm, and the actual genotype.

ESS = N*r²



Cost assumptions:
Sample preparation - 30$
Sequencing 1x - 133$

7 Summary

1. Genomic imputation in Ashkenazi jews using a reference panel of 128 individuals.
2. The algorithm used to infer haplotypes is HMM based.
3. Maximum in effective sample size achieved at sequencing coverage of 0.05-0.1x.

References

S. Carmi, K.Y. Hui, E. Kochav, X. Liu, J. Xue, et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun 5, 4835 (2014).

B. Pasaniuc, N. Rohland, P.J. McLaren, K. Garimella, N. Zaitlen, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44, 631 (2012).