Imputation for summary statistics in GWAS settings

Sina Rüeger (sina.rueger@unil.ch)
Zoltán Kutalik (zoltan.kutalik@unil.ch)

Institute of Social and Preventive Medicine, University Hospital and University of Lausanne, Lausanne, Switzerland
Swiss Institute of Bioinformatics, Lausanne, Switzerland

http://www3.unil.ch/wpmu/sgg/

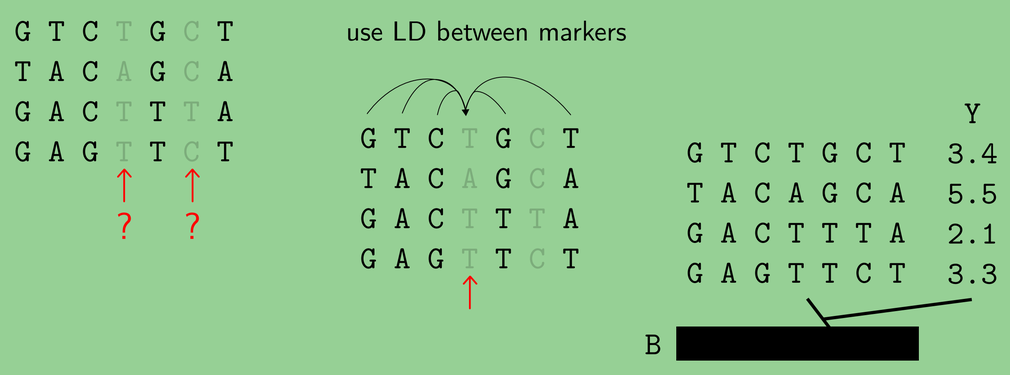
Swiss Institute of Bioinformatics

Why is it needed?

Genome-wide association studies use microarrays to measure SNPs that are often designed to tag many untyped variants, which can be imputed via the linkage disequilibrium (LD) between measured and untyped markers.

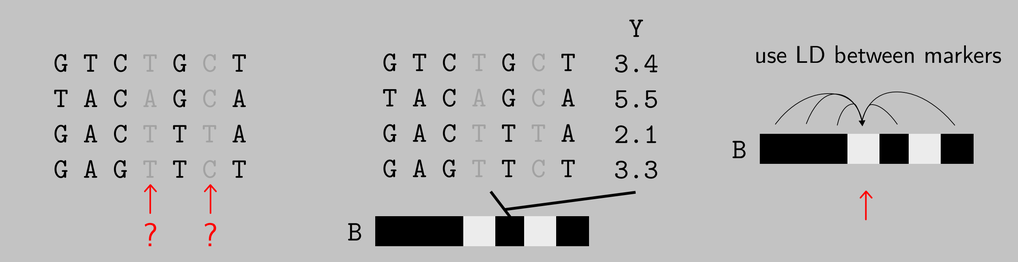
Standard imputation methods.

The imputation methods, while making most of the available data, are computationally very expensive when it comes to imputing ~30-40M variants of the 1000 Genomes panel. These imputed variants are subsequently subjected to association with various traits.



How does it work?

We propose an approach that performs imputation directly on the association summary statistics (such as t-statistics) of typed SNPs.



This allows a fast inference of the association strength of non-genotyped markers using that of the tagging SNPs. This approach bears similarities with the pioneering work of Pasaniuc et al. (2013).

$$Z_{i|t} = \Sigma_{i|t}\Sigma_{t,t}^{-1}Z_t \qquad \Sigma \overset{\text{invertible}}{=} \Sigma + \lambda I$$

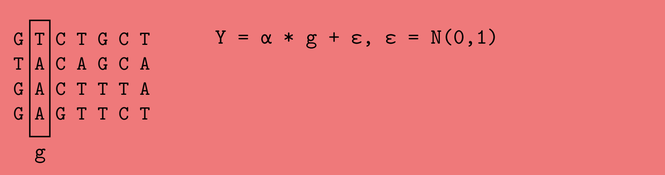
t: observed SNPs
i: **un**observed SNP
Σ: correlation matrix

Novelties (compared to Pasaniuc et al.)

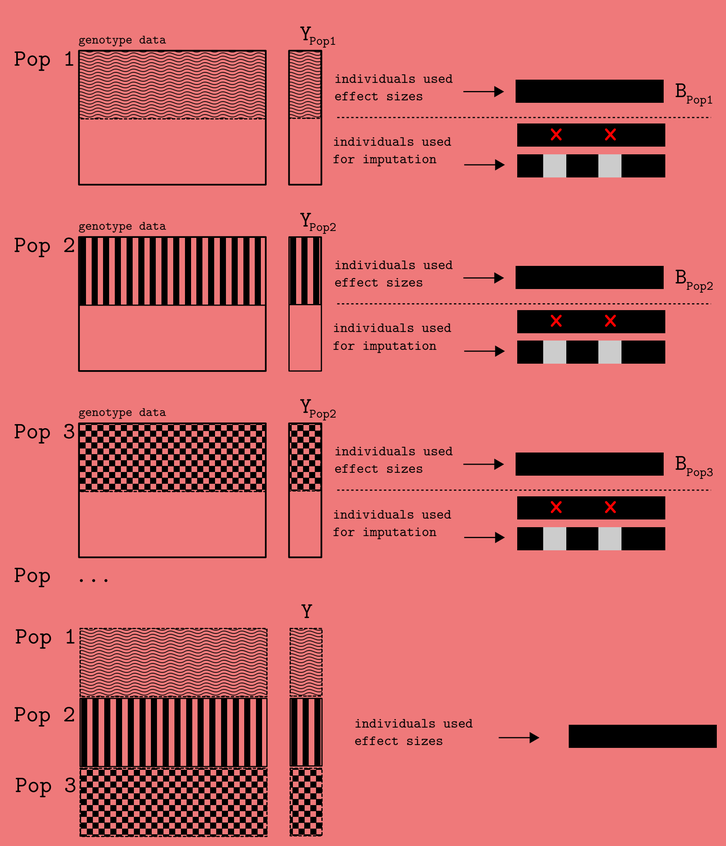
- optimal windows size for collective calculation
- optimized λ of the pair-wise marker correlation matrix
- optimized haplotype selection

Testing framework

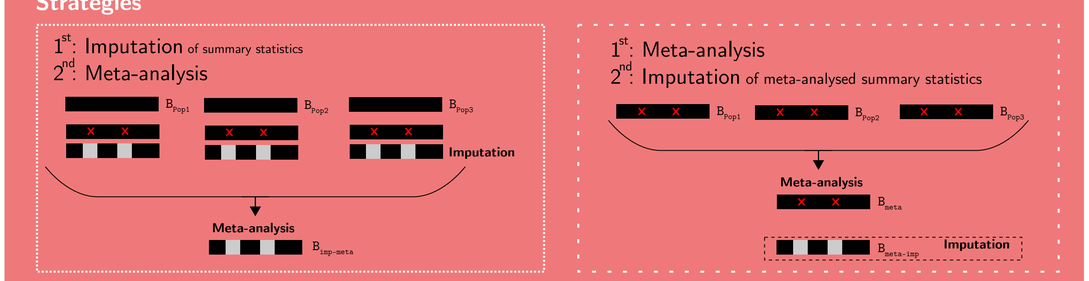
For testing we used HapMap imputed cohort data (limited to Chromosome 15) and looked particulary on regions with ancestry informative markers. We selected 6 sub-populations (CH, FR, IT, GE, SP, PT) and kept the sample size roughly equal. To generate an insilico phenotype a SNP was selected as being the "causal" one and effect sizes were generated for all SNPs (using a linear regression model).



Using the association statistics from HapMap SNPs only, we imputed the effect size of non-HapMap SNPs and compared to the "true" effect size estimates.



Strategies



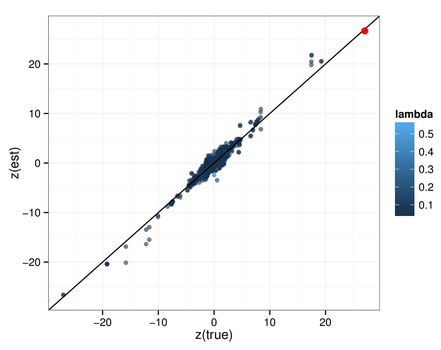
Reference panels for imputation

Reference data sets should represent the population used to calculate the effectsizes (usually a mixture of different populations).

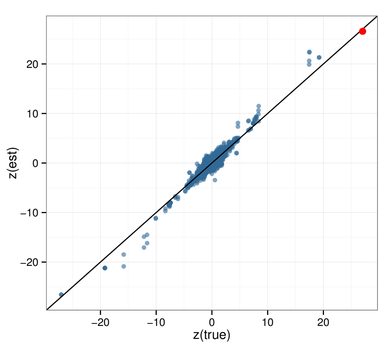
How can we improve the imputation performance?

For the insilico phenotype we generated, the results suggest that our test statistics agree closer (mean square error = 0.024, optimized λ) with the true values than the estimates provided by previous methods (mean square error = 0.028, λ = 0.1). The optimized λ makes only little difference in this setting. However, as the λ gets smaller, e.g. λ = 1e-07, the median square error increases to 0.047. λ = 0.1 is optimal for small reference data sets.

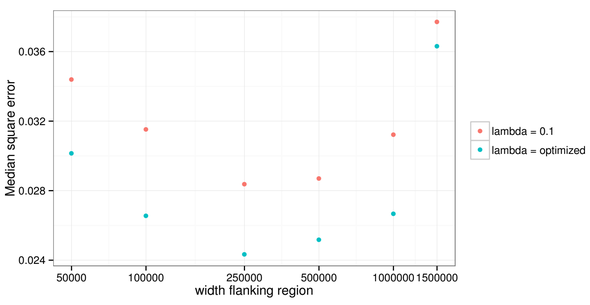
Optimized lambda



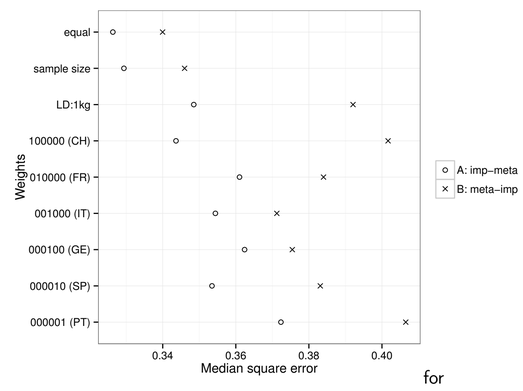
Lambda = 0.1



Finding the optimal window size is important. Therefore we varied the flanking region and ran the same data set for both, optimized λ and λ = 0.1. The optimal **flanking region** width in our setting is 250e3.

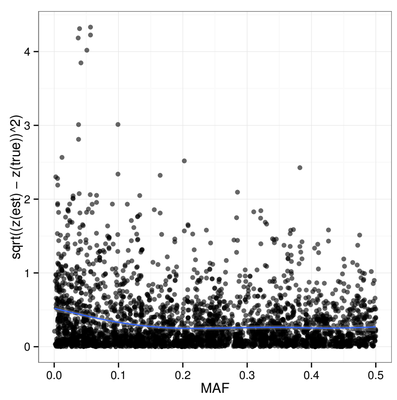


Additionally we looked at **haplotype mixtures** (for λ = 0.1), where population mixtures according to sample size was performing the best (in this case sample size was almost equal).
The order of using the methods - imputation is done before doing meta-analysis - plays an important role. Our results show a consistent better performance.



for

Error is decreasing as the minor allele frequency increases.



Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, Price AL (2014) Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics.