Fast Imputation of Summary Statistics Based on Local LD Structure

Matteo Togninalli, Damian S. Roquerio, Karsten M. Borgwardt
ETH Zürich

Meta-analyses of genome-wide association studies (GWAS) have increasingly been used to improve statistical power of association tests and discover new genetic risk variants associated with specific traits (Evangelou and Ioannidis 2013). In order to achieve this, GWAS results – under the form of summary statistics – are pooled and combined via different statistical methods. Since meta-analysis is carried independently for each single nucleotide polymorphism (SNP), conducting one comes at a price: for maximum power, all GWAS needs to include results for the same SNPs. Therefore, genotype imputation is often used to ensure positional overlap of all involved GWAS. However, genotype imputation suffers from two major drawbacks: (i) it can be computationally burdensome for large datasets and (ii) it requires the availability of the genotype data, whose access remains difficult for privacy and organizational reasons.

The need for reliable summary statistics imputation methods is therefore evident for the scientific community, both to ease large consortiums' efforts and to democratize meta-analysis to other researchers. This is an acknowledged problem and, in the last few years, a few methods have been proposed as software solutions: DIST (Lee et al. 2013), ImpG-Summary (Pasaniuc et al. 2014), DISSCO (Xu et al. 2015) and DISTMIX (Lee et al. 2015). They all assume that multivariate Gaussian distribution accurately model the local distribution of summary statistics values (Wen and Stephens 2010) and they use this observation to impute missing summary statistics using linkage disequilibrium (LD) structure information gathered from an external reference panel such as the 1000 genomes panel (Auton et al. 2015).

Correlation matrices used in the imputation step are approximated by the ones obtained from the external panel. In order to avoid computing a very large correlation matrix traversing an entire chromosome, the presented methods use smaller windows that still ensure accurate imputation while decreasing computational burden. However, the proposed methods rely on fixed window sizes (either in number of SNPs or number of base pairs), which hardly reflects the structure of LD across the genome.

In this study, we propose a summary statistics imputation method that uses a moving window which allows for matrix computation speed-ups while better considering local LD structures and ultimately results in faster imputation. Furthermore, our method does not ask the user to enter window-size parameter as we identify the best window-size based on the provided SNPs, we therefore avoid the tedious trial-and-error process required to fine-tune parameters prior to imputation.

Our preliminary observations show that much smaller LD structures are often enough to reach accurate imputation and result in faster runtimes. Moreover, the flexibility of the method is guaranteed over various datasets and available SNPs densities. Our method will also be benchmarked against state-of-the-art tools used in the imputation of summary statistics.

References:

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. doi:10.1038/nature15393.

Evangelou, Evangelos, and John P. A. Ioannidis. 2013. “Meta-Analysis Methods for Genome-Wide Association Studies and beyond.” *Nature Reviews Genetics* 14 (6). Nature Publishing Group: 379–89. doi:10.1038/nrg3472.

Lee, Donghyung, T. Bernard Bigdeli, Brien P. Riley, Ayman H. Fanous, and Silviu Alin Bacanu. 2013. “DIST: Direct Imputation of Summary Statistics for Unmeasured SNPs.” *Bioinformatics* 29 (22): 2925–27. doi:10.1093/bioinformatics/btt500.

Lee, Donghyung, T. Bernard Bigdeli, Vernell S. Williamson, Vladimir I. Vladimirov, Brien P. Riley, Ayman H. Fanous, and Silviu Alin Bacanu. 2015. “DISTMIX: Direct Imputation of Summary Statistics for Unmeasured SNPs from Mixed Ethnicity Cohorts.” *Bioinformatics* 31 (19): 3099–3104. doi:10.1093/bioinformatics/btv348.

Pasaniuc, Bogdan, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P. Strachan, Nick Patterson, and Alkes L. Price. 2014. “Fast and Accurate Imputation of Summary Statistics Enhances Evidence of Functional Enrichment.” *Bioinformatics (Oxford, England)* 30 (20): 2906–14. doi:10.1093/bioinformatics/btu416.

Wen, Xiaoquan, and Matthew Stephens. 2010. “Using Linear Predictors to Impute Allele Frequencies from Summary or Pooled Genotype Data.” *Annals of Applied Statistics* 4 (3): 1158–82. doi:10.1214/10-AOAS338.

Xu, Zheng, Qing Duan, Song Yan, Wei Chen, Mingyao Li, Ethan Lange, and Yun Li. 2015. “DISSCO: Direct Imputation of Summary Statistics Allowing Covariates.” *Bioinformatics* 31 (15): 2434–42. doi:10.1093/bioinformatics/btv168.