RAISS: Robust and Accurate imputation from Summary Statistics

Hanna Julienne[1,*] Huwenbo Shi[2], Bogdan Pasaniuc[2], Hugues Aschard[1]

[1]Groupe de Génétique Statistique, Département de Génomes and Génétique, C3BI, Institut Pasteur, Paris, France.
[2] Departments of Pathology and Lab Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed.

Abstract
Motivation: Multi-trait analyses using public summary statistics from genome-wide association studies (GWAS) are becoming increasingly popular. A constraint of multi-trait methods is that they require complete summary data for all traits. While methods for the imputation of summary statistics exist, they lack precision for genetic variants with small effect size. This is benign for univariate analyses where only variants with large effect size are selected a posteriori. However, it can lead to strong p-value inflation in multi-trait testing. Here we present a new approach that improve the existing imputation methods and reach a precision suitable for multi-trait analyses.
Results: We fine-tuned parameters to obtain a very high accuracy imputation from summary statistics. We demonstrate this accuracy for variants of all effect sizes on real data of 28 GWAS. We implemented the resulting methodology in a python package specially designed to efficiently impute multiple GWAS in parallel.
Availability: The python package is available at: https://gitlab.pasteur.fr/statistical-genetics/raiss, its accompanying documentation is accessible here http://statistical-genetics.pages.pasteur.fr/raiss/.
Contact: hanna.julienne@pasteur.fr

1 Introduction

By solving practical and ethical challenges, public summary statistics has become a gold entry point for the study of complex traits (B. Pasaniuc & Price, 2017). In the past years, multi-trait methods using summary statistics have attracted much scientific attention and yields many applications including *e.g.* multi-trait testing (Liu & Lin, 2018; Turley et al., 2018) or correction for pleiotropy in mendelian randomization (Verbanck, Chen, Neale, & Do, 2018). Most multi-trait methods are only applicable to single nucleotide polymorphisms (SNPs) with complete data for the traits of interest, and imputation of missing statistics is mandatory in many real data analyses. However, in the multi-traits context, imputation must reach a very high level of accuracy, even SNPs with moderate effect sizes, to avoid false association signal. Existing solutions do not achieve this level of accuracy (see Supplementary Fig.1). The imputation must also be time efficient so computation does not become a bottleneck in pre-processing many traits.

We improved an existing imputation solution (Bogdan Pasaniuc et al., 2014) on two key points to make it suitable for multi-trait applications: 1) we optimize a hyper-parameters through a systematic space search . 2) We designed the python package RAISS (Robust and Accurate imputation from Summary Statistics) so multiple traits can be imputed in parallel.

2 Methods

2.1 Statistical model

The statistical model used in RAISS is similar to the one described in (Lee, Bigdeli, Riley, Fanous, & Bacanu, 2013). Summary statistics are given as Z-scores. The model assumes that the Z-scores are under the null hypothesis but is robust to realistic violations of this assumption (see Supplementary data). The idea behind summary statistics imputation is to leverage linkage disequilibrium (LD) to compute Z-scores of missing SNPs from neighboring observed SNPs:

$$\mathbb{E}(\boldsymbol{z_i}|\boldsymbol{z_t}) = \boldsymbol{\Sigma_{i,t}}\,\boldsymbol{\Sigma_{t,t}^{-1}}\boldsymbol{z_t}$$

Where z_i is the vector of missing SNPs, z_t is the vector of observed SNPs and Σ is linkage disequilibrium matrix between SNPs. The conditional variance of z_i is estimated as:

$$\boldsymbol{Var}(\boldsymbol{z_i}|\boldsymbol{z_t}) = \boldsymbol{\Sigma_{i,i}} - \boldsymbol{\Sigma_{i,t}}\boldsymbol{\Sigma_{t,t}^{-1}}\boldsymbol{\Sigma_{t,i}}$$

It follows that the variance of a missing SNPs j explained by observed SNPs equals ${R^2}_j = 1 - Var(z_j|\boldsymbol{z_t})$. We use the standardized conditional expectation of z_j as its estimator:

$$\hat{z}_j = E(z_j|\boldsymbol{z_t}) \Big/ \sqrt{{R^2}_j}$$

Details on the derivation are provided in supplementary data along simulation results showing the robustness of this estimator in the presence of causal SNPs (Supplementary Figures 3 and 4).

2.2 Ensuring correct inversion of $\boldsymbol{\Sigma_{t,t}}$

Neighboring SNPs are highly correlated variables which makes the inversion of $\boldsymbol{\Sigma_{t,t}}$ prone to numerical instabilities. We invert $\boldsymbol{\Sigma_{t,t}}$ with the Moore-Penrose pseudo inverse. To ensure numerical stability, we applied a very stringent pruning of small eigen-values –*i.e.* eigen values below a given threshold are set to zero in the computation of $\boldsymbol{\Sigma_{t,t}}$ pseudo inverse We will denote this threshold *rcond* in this paper (as its corresponding parameter in the scipy.linalg.pinv function).

2.3 RAISS pipeline and computation time optimization

2.3.1 Precomputation of linkage disequilibrium

We derived LD using individuals of European ancestry from the 1000 genome panel (Genomes Project et al., 2012) (see RAISS documentation). To avoid repeated estimation of LD when imputing statistics for multiple GWAS, RAISS precomputes pairwise linkage disequilibrium between SNPs present in the reference panel (see Supplementary data and Supplementary Figures 5 and 6). The execution times of 483 various imputation tasks are reported in Supplementary table 3. On chromosome 1, the imputation took on average 2 hours and never exceeded 5 hours 30 minutes.

2.3.2 Command line tool for chromosome imputation

The simplest access to the imputation function in RAISS is the shell command *raiss* (accessible in a terminal after installing the package). This command impute the summary statistics for one trait and one chromosome and filter the results according to the R^2 value (see Supplementary Fig.5 and RAISS documentation).

3 Results

We tested RAISS performances using the following procedure. For a chromosome and a trait: 1) Remove randomly 5000 SNPs in the Z-score file, 2) Impute these 5000 SNPs, 3) Set imputation hyper-parameters, 4) Compute the correlation between the real Z-scores and the imputed Z-scores.

3.1 Effect of hyper-parameters

We ran the above validation procedure on chromosome 22 for a height GWAS (Wood et al., 2014). We varied the stringency in pruning small eigen-value from 10e-15 (default value in scipy) to 0.1, and the imputation R^2 filtering threshold from 0.1 to 0.9.

The pruning threshold for small eigen-value turns out to be the most important hyper-parameter to ensure a good correspondence between observed and imputed Z-scores. The imputation quality concomitantly increases with the threshold to reach a high correlation of 0.96 (see Figure 1). A more complete discussion about the setting of the *rcond* parameter is available in Supplementary data (see

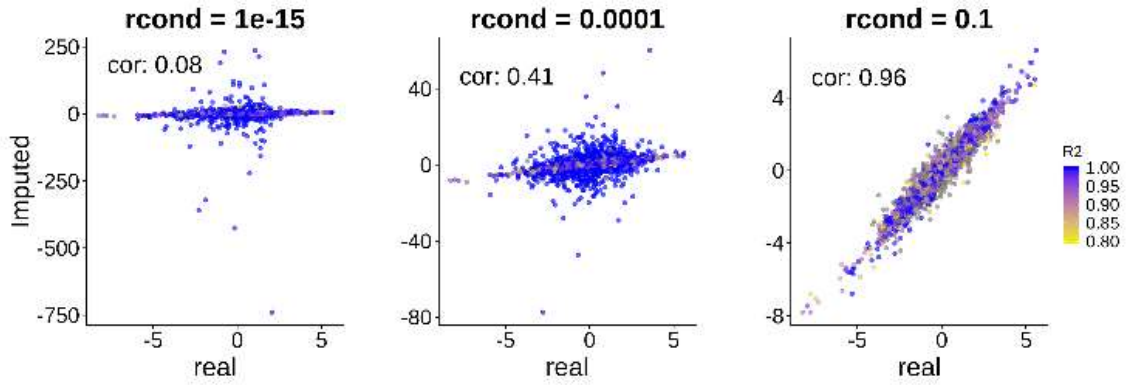

Figure 1 Imputation performance measured on real data (height GWAS). The x-axis shows the measured SNP zscores and the y-axis shows the imputed zscores for the same SNPs (the imputation was performed with the plotted SNPs masked). The color scale shows the variance explained by the imputation model (R^2) for each SNPs. Only SNPs with a R^2 above 0.8 are represented. Each panel corresponds to a different *rcond* parameter which determines the eigen vector used to perform the LD matrix inversion (see section 2.2).

Supplementary Fig.2). Filtering imputed SNPs by their R^2 improves only slightly the imputation accuracy. Moreover, if the R^2 threshold is set too high most of the imputed SNP would be filtered (see Supplementary Fig.7 and Supplementary data).

3.2 Performance on a large panel of traits

To further assess the relevance of the hyper-parameters defined using the height GWAS data (*rcond* = 0.1 and R^2 >0.6), we applied the final procedure for the analysis of 28 GWAS (see Supplementary table 1). The correlation between real and imputed Z-scores varied from 0.9 to 0.97 (see Supplementary table 2) dramatically increasing performances as compared to existing approach. We used imputed summary statistics from RAISS as input for a multivariate test method currently available at http://jass.pasteur.fr/index.html and we did not observe any p-value inflation as measured by the genomic control coefficient (see Supplementary Fig. 8).

4 Conclusion

We implemented an efficient tool allowing for the imputation of multiple summary statistics in parallel. We demonstrate a greatly improved accuracy for small size-effect variants in the real data analysis of 28 GWAS. Thus, the RAISS package has an appropriate level of confidence that makes it suited for the imputation of summary statistic for various multi-trait analyses.

Funding

This work has been supported by the NIH grant R03DE025665.

Conflict of Interest: none declared.

References

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56-65. doi:10.1038/nature11632

Lee, D., Bigdeli, T. B., Riley, B. P., Fanous, A. H., & Bacanu, S. A. (2013). DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics*, *29*(22), 2925-2927. doi:10.1093/bioinformatics/btt500

Liu, Z., & Lin, X. (2018). Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics*, *74*, 165-175. doi:10.1111/biom.12735

Pasaniuc, B., & Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet, 18*(2), 117-127. doi:10.1038/nrg.2016.142

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., . . . Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics (Oxford, England), 30*(20), 2906-2914. doi:10.1093/bioinformatics/btu416

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., . . . Social Science Genetic Association, C. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet, 50*(2), 229-237. doi:10.1038/s41588-017-0009-4

Verbanck, M., Chen, C. Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet, 50*(5), 693-698. doi:10.1038/s41588-018-0099-7

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., . . . Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics, 46*(11), 1173-1186. doi:10.1038/ng.3097