Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Improved imputation of summary statistics for mixed populations

Sina Rüeger [1,2], Aaron McDaid [1,2] and Zoltán Kutalik [1,2]*

[1] Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland and
[2] Institute of Social and Preventive Medicine, Lausanne University Hospital, Lausanne, 1010, Switzerland

*Zoltán Kutalik: zoltan.kutalik@unil.ch, tel.: +41-21 314 6750

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: *Summary statistics imputation* can be used to infer association summary statistics of an already conducted, genotype-based meta-analysis to higher genomic resolution. This is typically needed when *genotype imputation* is not feasible for some cohorts. Oftentimes, cohorts of such a meta-analysis are variable in terms of (country of) origin or ancestry. This violates the assumption of current methods that an external LD matrix and the covariance of the Z-statistics are identical.

Results: To address this issue, we present *variance matching*, an extention to the existing *summary statistics imputation* method, which manipulates the LD matrix needed for *summary statistics imputation*. Based on simulations using real data we find that accounting for ancestry admixture yields noticeable improvement only when the total reference panel size is > 1000. We show that for population specific variants this effect is more pronounced with increasing F_{ST}.

Contact: zoltan.kutalik@unil.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genotype data for genome-wide association studies (GWASs) are often collected using DNA chips, which cover only a small fraction of the variable genome. To be able to combine GWASs that measured different sets of genetic markers (due to differences in the content of commercial arrays), genetic information has to be inferred for a common set of markers. Such inference exploits the fact the neighbouring SNVs are often in linkage disequilibrium (LD), which has been well-quantified in different human populations. Statistical inference of these untyped SNVs in a study cohort, therefore, relies on an external reference panel of densely genotyped or sequenced individuals. The inference process is termed *imputation*, of which there are two main types. *Genotype imputation* (Marchini and Howie, 2010) first estimates all haplotypes both in the reference panel and the study cohort, then using a Hidden Markov Model every observed haplotype in the study cohort is assembled as a probabilistic mosaic of reference panel haplotypes. The reconstruction facilitates the computation of the probability of each genotype for every SNV of the reference panel in each individual of the study cohort. Having imputed the genotype data set, one can then run an association scan with an arbitrary trait and obtain association summary statistics. *Summary statistics imputation* Pasaniuc *et al.* (2014) on the other hand starts off with association summary statistics available for all genotyped markers and infers, combined with a reference panel, directly the association summary statistics of SNVs available in the reference panel. More specifically, estimating the local pair-wise linkage disequilibrium (LD) structure of each genetic region using the reference panel and combining it with association summary statistics allows to calculate a conditional expectation of normally distributed summary statistics. This latter approach is the central focus of our paper. Compared to *genotype imputation*, *summary statistics imputation* is much less demanding on computational resources, and requires no access to individual level genetic data.

Methods making use of summary statistics, such as calculating genetic correlation (Bulik-Sullivan *et al.*, 2015), approximate conditional analysis (Yang *et al.*, 2012) or causal inference (Burgess *et al.*, 2013), have gained interest in recent years, because they bypass the need of genotype data, but mimic it by making use of external reference panels. These methods could profit from summary statistics being available on an arbitrarily chosen panel of SNVs – provided by *summary statistics imputation*. However, it is not clear how to optimally combine different LD reference panels for summary statistics emerging from a meta-analysis of a large number of different studies (coming from different countries/regions),

