Bioinformatics, 34, 2018, i687–i696
doi: 10.1093/bioinformatics/bty596
ECCB 2018

OXFORD

Accurate and adaptive imputation of summary statistics in mixed-ethnicity cohorts

Matteo Togninalli[1,2,*], Damian Roqueiro[1,2], COPDGene Investigators[3] and Karsten M. Borgwardt[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland and [3]COPDGene® Study

[*]To whom correspondence should be addressed.

Abstract

Motivation: Methods based on summary statistics obtained from genome-wide association studies have gained considerable interest in genetics due to the computational cost and privacy advantages they present. Imputing missing summary statistics has therefore become a key procedure in many bioinformatics pipelines, but available solutions may rely on additional knowledge about the populations used in the original study and, as a result, may not always ensure feasibility or high accuracy of the imputation procedure.

Results: We present ARDISS, a method to impute missing summary statistics in mixed-ethnicity cohorts through Gaussian Process Regression and automatic relevance determination. ARDISS is trained on an external reference panel and does not require information about allele frequencies of genotypes from the original study. Our method approximates the original GWAS population by a combination of samples from a reference panel relying exclusively on the summary statistics and without any external information. ARDISS successfully reconstructs the original composition of mixed-ethnicity cohorts and outperforms alternative solutions in terms of speed and imputation accuracy both for heterogeneous and homogeneous datasets.

Availability and implementation: The proposed method is available at https://github.com/BorgwardtLab/ARDISS.

Contact: matteo.togninalli@bsse.ethz.ch or karsten.borgwardt@bsse.ethz.ch

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

1 Introduction

The genome-wide association study (GWAS) is an invaluable tool to detect associations between a trait and genetic variants in individuals. For over a decade, GWASs have been conducted in a variety of organisms that span different plants and crops (Lin *et al.*, 2014; Meijon *et al.*, 2014; Zhao *et al.*, 2011), animal species (Kirby *et al.*, 2010; Mackay *et al.*, 2012) and humans (Freilinger *et al.*, 2012). Public databases and web services have been developed to provide easy access to the results of many of these association studies, e.g. for the model organism *Arabidopsis thaliana* (Togninalli *et al.*, 2018) and for humans (Welter *et al.*, 2014). Of the results generated by a GWAS, the *association summary statistics*, normally in the form of Z-scores, have recently gained considerable attention. They are being increasingly used for meta-analyses, conditional association methods, gene-based association tests, fine-mapping and to investigate the polygenic nature of complex traits (Pasaniuc and Price, 2017).

The growing popularity of methods that analyze summary statistics can be attributed to (i) their advantageous computational cost, especially when compared to genotype-based methods, and to (ii) the relative absence of privacy concerns when manipulating and exchanging the data. However, researchers working with summary statistics often encounter disparate datasets obtained from studies that were performed with different genotyping platforms and/or filtering criteria. This limits the types of analyses that can be conducted because of the incomplete overlap of the genetic variants—in the form of single-nucleotide polymorphisms, or SNPs. Early strategies to tackle said limitation relied on finding proxy SNP values (Meesters *et al.*, 2012), but the ubiquity of the problem justifies the development of reliable methods to impute missing summary statistics. In the last few years, several methods have been proposed as software solutions to the problem of imputing summary statistics in association studies.

Here, it is important to draw a distinction between Z-scores and other summary statistics like genotype counts and allele frequencies. Association summary statistics such as Z-scores or *P*-values identify genomic regions that have a strong association to a trait.

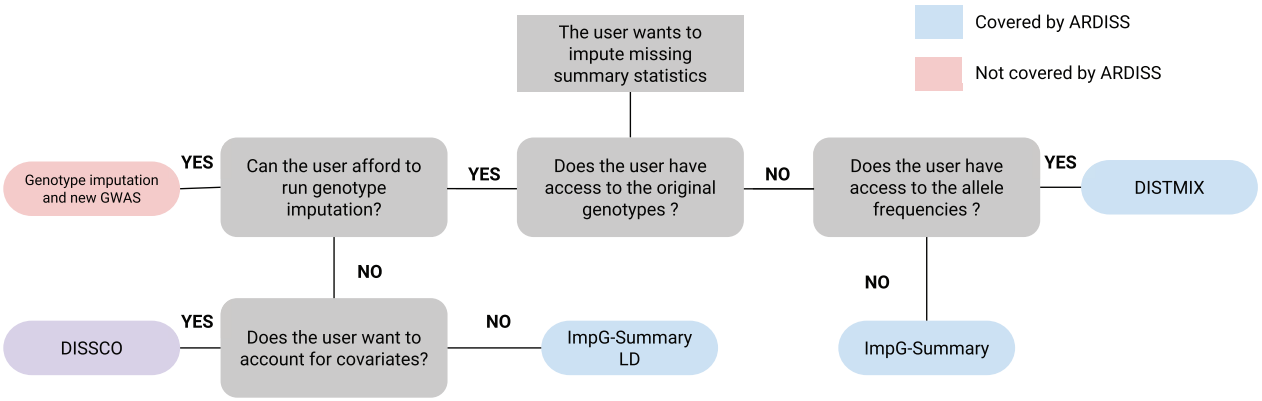

Fig. 1. Usage mapping of available methods for Summary Statistics imputation depending on data availability and computational resources. Accounting for covariates is not necessary if the covariates were taken into account during the original study (Section 4.1.2)

The imputation of Z-scores is the main focus of our work and of other state-of-the-art summary statistics imputation methods. However, the theoretical foundations of these solutions are valid for other summary statistics such as allele frequencies (Wen and Stephens, 2010) and β values.

Imputation methods based on Z-scores, depending on their data needs with respect to the samples in the original study, can be broadly categorized as (i) requiring no additional data, just the Z-scores and (ii) requiring some additional data (in addition to the Z-scores).

For the latter, the additional information can be in the form of covariates, other summary statistics like allele frequencies, or the ethnic composition of individuals in the original study. This is shown in more detail in Figure 1. The flowchart highlights the most common use cases for the imputation of summary statistics and indicates which state-of-the-art algorithm can be used for each scenario. The first distinction between the current methods is based on having access to the original genotypes. The attentive reader may wonder about the need to impute summary statistics when one has access to the original genotype data. After all, if the genotypes of each sample are available, one can rely on well-established imputation methods such as IMPUTE2 (Howie *et al.*, 2009), MaCH (Li *et al.*, 2010) and others (Browning and Browning, 2007; Servin and Stephens, 2007) to impute missing SNPs. Once the genotypes of the missing SNPs have been imputed, the association Z-scores can be computed for all SNPs (original and imputed). Nevertheless, the reason why one may pursue the imputation of summary statistics despite having access to the genotypes is due to the heavy computational cost involved in imputing genotype data. The SNPs to be imputed can be in the order of 10^6 and in a study with thousands of samples, the imputation task may require weeks of dedicated cluster computing. As shown in Figure 1, this is a plausible use case for methods like DISSCO (Xu *et al.*, 2015) or ImpG-SummaryLD (Pasaniuc *et al.*, 2014), which require access to the original genotype data.

If the genotypes are not available, one has to assess if the Z-scores to be imputed arise from a study with a cohort of mixed-ethnicities. Depending on the organism on which the GWAS was conducted, this may or may not be an issue. As an example, in humans, GWASs tend to be conducted on homogeneous cohorts–or at least as homogeneous as the designers of the study envisioned it–to avoid spurious associations due to population stratification. Nevertheless, ethnicity in these studies is self-reported and individuals may not be aware of their true genetic background when recruited for the study. This often creates a mixed-ethnicity cohort and the Z-scores of association reflect that. This problem is exacerbated in other organisms such as plants. In these cases, it is almost guaranteed that the Z-scores were derived from a non-homogeneous cohort and the imputation of missing Z-scores should take this into consideration to avoid reporting false positives. DISTMIX (Lee *et al.*, 2015) addresses the issue of mixed-ethnicities, but in order to perform the proper adjustment of imputed Z-scores, it requires the allele frequencies in the original study to estimate the different compositions of ethnic groups in that study. We cannot take lightly the fact that DISTMIX requires the allele frequencies to perform an accurate imputation, especially because these additional data may be simply unavailable or hard to obtain.

Therefore, there is no unique and adaptive method that ideally fits all scenarios. Hence, we here propose a new imputation method named `ARDISS` that is able to approximate the ethnic composition of the samples in an association study by simply relying on (i) the Z-scores obtained from such study and (ii) a reference panel. `ARDISS` does not require neither additional summary statistics nor any information of the samples in the study, thus preserving the anonymity of the original samples. Our method relies on Automatic Relevance Determination (ARD), a common strategy used in the Gaussian Process Regression literature for feature selection in high-dimensional spaces (MacKay, 1994). `ARDISS` uses automatic relevance determination to weigh the contribution of single samples to the observed Z-scores. Moreover, our method is highly parallelizable and this results in considerable speed improvements with respect to current solutions. The remainder of the document is organized as follows: Section 2 presents background information about the imputation of summary statistics and introduces our new method `ARDISS`. Section 3 describes the datasets and the experimental setup we used. Section 4 compares our method with state-of-the-art tools and provides an interpretation of the results. Section 5 discusses the applicability scenarios of the state-of-the-art methods and how these relate to `ARDISS`. Section 6 concludes the paper.

2 Materials and methods

2.1 Existing methods

Several methods have been proposed for the task of imputing summary statistics in association studies. In particular, DIST (Lee *et al.*, 2013), ImpG-Summary (Pasaniuc *et al.*, 2014), DISSCO (Xu *et al.*, 2015) and DISTMIX (Lee *et al.*, 2015) are considered state-of-the-art for this task. Despite their differences in terms of additional data requirements, all methods share a commonality: in order to impute

missing Z-scores for specific SNPs, they rely on a reference panel of genotyped individuals. For example, these reference panels are normally obtained from the 1000 Genomes Project (Abecasis *et al.*, 2012), for humans, or from the 1001 Genomes Project (Alonso-Blanco *et al.*, 2016) for the plant *A. thaliana*. These methods impute missing Z-scores by approximating them with a multi-variate Gaussian distribution over neighboring SNPs' values. They differentiate between typed (Z_t) and untyped (Z_u) values and use the underlying linkage disequilibrium (LD) structure to approximate the distribution of the Z-scores Z according to variations of the following formula:

$$Z_{u|t} = \Sigma_{ut}\Sigma_{tt}^{-1}Z_t \tag{1}$$

Where Σ_{ut} is the matrix of correlations between untyped and typed SNPs and Σ_{tt} is the matrix of correlations between typed SNPs. The correlations are obtained by looking at the SNP genotype values of samples from a reference panel (e.g. the 1000 Genomes samples) and measuring their Pearson's correlation coefficient. This approach can be translated in a naïve Gaussian Process Regression (Rasmussen and Williams, 2006) that uses a simple linear kernel k and 0 mean:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}')) \tag{2}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{d} x_i x_i' \tag{3}$$

Where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the standardized feature vectors of two SNPs (i.e. the standardized genotype values for the d individuals in the reference panel). The Gaussian Process then generates predicted means and variances of the missing points according to the following formulas:

$$\begin{gathered} \boldsymbol{f_u}|X_i, X_t, \boldsymbol{f_t} \sim \mathcal{N}(\mu_K, \sigma_K) \\ \mu_K = K(X_u, X_t)K(X_t, X_t)^{-1}\boldsymbol{f_t} \\ \sigma_K = K(X_u, X_u) - K(X_u, X_t)K(X_t, X_t)^{-1}K(X_t, X_u) \end{gathered} \tag{4}$$

Where $\boldsymbol{f_u}$ and $\boldsymbol{f_t}$ are equivalent to Z_u and Z_t, respectively; X_u and X_t are the matrices of features for the untyped and typed SNPs and $K(X, X')$ is the $n \times n'$ matrix of the covariance values evaluated at all pairs of points using Equation (3). To simplify notation, we will refer to $K(X_t, X_t)$, $K(X_u, X_u)$, $K(X_u, X_t)$ and $K(X_t, X_u)$ as K_{tt}, K_{uu}, K_{ut} and $K_{ut}^\top$, respectively.

To account for the noise observed in the typed data, it is common to add a noise component in the covariance between typed data as follow:

$$K_y = K_{tt} + \sigma_{\text{noise}}^2 I \tag{5}$$

and replace K_{tt} by K_y. Notice that this step is often related as $\Sigma_{tt}^{\text{adj}} = \Sigma_{tt} + \lambda I$ in the summary statistics imputation literature.

Other variations of the formulas presented here have been used to account for mixed-ethnicity cohorts. While ImpG-Summary does not account at all for mixed populations during imputation, other methods that do so (DISSCO and DISTMIX) require extra information about the study population: the user should either report the original study genotypes, the allele frequencies of the study genotypes or an estimate of the population structure under the form of fractions that sum up to one. This information is then used to compute adjusted partial correlations between SNPs. Such requirements are not ideal in a realistic setting: when access to the original genotypes is possible, genotype imputation should be preferred (Pasaniuc *et al.*, 2014) and allele frequencies are not often shared due to privacy concerns (Erlich and Narayanan, 2014; Homer *et al.*, 2008), see Section 5 for a more detailed discussion.

2.2 Automatic relevance determination

In order to account for mixed ethnicity cohorts without relying on other sources of information, we introduce `ARDISS` (ARD for Imputation of Summary Statistics), a new summary statistics imputation method that only relies on the typed statistics and a reference panel of genotypes [e.g. for humans, the panel from the 1000 Genomes Project (Abecasis *et al.*, 2012)]. In order to do so, we borrow elements from the Gaussian Process field, with a focus on ARD.

A Gaussian Process is characterized by its mean function $m(x)$, kernel k and hyperparameters θ. Moreover, if the mean is zero, one can compute the marginal likelihood (or evidence) to evaluate how the parameters fit the observed data according to:

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^\top K_y^{-1}\boldsymbol{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log(2\pi) \tag{6}$$

To have an optimal fit of the Gaussian Process, we want to maximize the likelihood. Hence, we can set the optimal hyperparameters by computing the partial derivatives of the marginal likelihood with respect to the hyperparameters [See Chapter 5 of (Rasmussen and Williams, 2006)]:

$$\frac{\partial}{\partial\theta_j}\log p(\boldsymbol{y}|X) = \frac{1}{2}\ \text{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - K_y^{-1}\right)\frac{\partial K_y}{\partial\theta_j}\right) \quad \text{where} \quad \boldsymbol{\alpha} = K_y^{-1}\boldsymbol{y} \tag{7}$$

and using the partial derivatives in a gradient based optimizer.

The idea of ARD is to add weights to every feature used to construct the kernel and to fit those weights so as to optimally represent the observed values with a Gaussian Process. In our case, this would mean to weigh the contribution of the individual genotypes in the reference panel so as to ideally match the reported Z-scores. This can be seen as a proxy to represent the original population of the GWAS as closely as possible. The new linear kernel function simply becomes (MacKay, 1994):

$$k_{\text{ARD}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{d} \sigma_i^2 x_i x_i' \tag{8}$$

And, we fit the σ_i values by using a gradient descent optimizer on the negative of the likelihood in Equation (6). The partial derivative of K_y with respect to σ_j is a simple outer product of the genotype values for sample j across the SNPs of interest multiplied by $2\sigma_j$.

2.3 `ARDISS`: implementation

`ARDISS` combines ARD with a moving-window imputation of missing GWAS Z-scores. The algorithm consists of two steps. We first iterate over the available Z-scores across one chromosome to obtain the consensus weight for every sample in the reference panel. Then, we apply the obtained weights to the genotype values and run the imputation with a moving window across the chromosome.

We rely on an external library *GPflow* (Matthews *et al.*, 2017) for the optimization of the weights. Since deriving the weights on a single large window enclosing all the SNPs for which Z-scores are available is computationally very expensive due to the many matrix inversions required ($O(n^3)$), the optimization procedure is carried out on subsets of SNPs in a window-based approach, as shown in lines 1–7 of Algorithm 1. This implementation can benefit from parallel computing on graphics processing units (GPUs) and allows very fast runtimes. Any optimizer object can be used for the optimization

of the weights (line 15) and we observed that the RMSProp optimizer implementation of GPflow with a learning rate of 0.1 and momentum of 0.001 yields satisfactory results. Once the ARD weights have been optimized, we average them across the chromosome (line 17) and we use the following formula to compute the untyped values:

$$Z_{i|t} = K_{it}^{\text{ARD}}[K_{tt}^{\text{ARD}} + \sigma_{\text{noise}}^2 I]^{-1} Z_t \quad (9)$$

Where the entries of K_{tt}^{ARD} are given by Equation (8). To accelerate execution, all the genotypes are multiplied element-wise with the ARD weights and standardized as seen in line 4 of Algorithm 2. The imputation procedure is then run using a moving-window with a window size given by the number of neighboring SNPs (a window size of 100 SNPs guarantees excellent results under various scenarios) rather than the commonly used approach of splitting the data in fixed-size blocks. This enables faster matrix operations which further speed up the execution. In particular, we can update the inverse of the correlation matrix in $O(n^2)$ by using the Sherman–Morrison formula. Furthermore, the imputation procedure always centers the window around the SNPs of interest, ensuring to find the strongest LD structures. A total of N loops are performed (where N is the number of typed SNPs). The covariance matrix K_{tt}, its inverse and the typed Z-score vector Z_t are all initialized before iterating through all typed SNPs (Lines 5–9). Boundary SNPs are treated slightly differently and are imputed with all the window-size SNPs at the boundary. Every iteration then updates the necessary entries (Lines 11–14), quickly retrieves X_u for the missing SNPs located between the two central typed SNPs (e.g. between the 50th and the 51st typed SNPs for a window size of 100) using specific Python data structures (line 15) and imputes their Z-score values (lines 16–17).

The overall complexity of `ARDISS` is $O(Nkw \cdot \max(w, d))$ for the weight learning step and $O(N(w^2 + n_u wd))$ for the imputation step, where N is the number of typed SNPs, k is the maximum number of iterations of the optimizer, w is the window size, d is the number of samples in the reference panel and n_u is the number of untyped SNPs in a single window. We can furthermore assume that, on average, $n_u = \frac{N_u}{N}$, where N_u is the overall number of untyped SNPs and have a final runtime complexity of $O(Nw^2 + N_u wd)$ for the imputation step. Due to the simple inner products needed to compute the covariance matrix, the method scales linearly for the number of samples in the reference panel—with fixed number of SNPs and a fixed window size. Please, refer to Supplementary Figure S1 for an empirical validation.

Algorithm 1 `ARDISS_get_weights`

Input: Standardized genotypes for typed SNPs $X_t \in \mathbb{R}^{N \times d}$, typed Z-scores $Z_t \in \mathbb{R}^N$, window size w, optimizer Opt
Output: Average ARD weights across chromosome
1: $W \leftarrow \lfloor \text{length of } X/w \rfloor, weights \leftarrow \emptyset$
2: **for** k in $\{1 \ldots W\}$ **do**
3 ▷ Slice the array to get batch samples
4: $X_{\text{batch}} \leftarrow X_{t,i\bullet} \quad i \in \{(k-1) \cdot w \ldots k \cdot w\}$
5: $Z_{\text{batch}} \leftarrow Z_{t,i} \quad i \in \{(k-1) \cdot w \ldots k \cdot w\}$
6: ▷ Initialize the ARD weights to a vector of d ones
7: $\sigma_{\text{ARD}}^2 \leftarrow 1_{1 \times d}$
8: **for** i in $\{1 \ldots Opt.maxiter\}$ **do:**
9 ▷ Compute the kernel matrix
10: $K_y \leftarrow X_{\text{batch}}\ \text{diag}(\sigma_{\text{ARD}}^2) X_{\text{batch}}^\top + \sigma_{\text{noise}} I$
11: $\alpha \leftarrow K_y^{-1} Z_{\text{batch}}$
12 ▷ Compute the σ_{ARD} gradients with Equation (7)
13: $\text{grads} \leftarrow \frac{1}{2}\ \text{tr}((\alpha\alpha^\top - K_y^{-1}) \frac{\partial K_y}{\partial \theta_i})_{j=\{1 \ldots d\}}$
14: ▷ Update the ARD weights with the optimizer of choice
15: $\sigma_{\text{ARD}} \leftarrow Opt.update(grads)$
16: Append σ_{ARD} to $weights$
17: Return average of $weights$ along second axis

Algorithm 2 `ARDISS`

Input: Standardized genotypes from reference sample $X \in \mathbb{R}^{M \times d}$, typed Z-scores $Z_t \in \mathbb{R}^N$, window size w, optimizer Opt
Output: Imputed Z-scores
1: Split genotypes in typed and untyped $X_t, X_u \leftarrow X$
2: $\sigma_{\text{ARD}} \leftarrow$ `ARDISS_get_weights`(X_t, Z_t, w, Opt)
3: ▷ Element-wise multiplication followed by standardization is a way to further speed up the computations afterwards
4: $X_{i\bullet} \leftarrow \text{Standardize } X_{i\bullet} \odot \sigma_{\text{ARD}} \quad i \in \{1 \ldots M\}$
5: $X_{t,\text{window}} \leftarrow X_{t,i\bullet} \quad i \in \{0 \ldots w\}$
6: $Z_{t,\text{window}} \leftarrow Z_{t,i} \quad i \in \{0 \ldots w\}, Z_u \leftarrow \emptyset$
7: $N \leftarrow \text{length of } X_t$
8: $K_{tt} \leftarrow X_{t,\text{window}} X_{t,\text{window}}^\top + \sigma_{\text{noise}}^2 I$
9: Compute K_{tt}^{-1}
10: ▷ Boundary conditions are treated differently
11: **for** i in $\{\frac{w}{2} + 1 \ldots N - \frac{w}{2}\}$ **do**
12: Update $X_{t,\text{window}}$, $Z_{t,\text{window}}$ and K_{tt}
13: ▷ Use Sherman-Morrison formulas
14: $K_{tt}^{-1} \leftarrow$ `update_inverse`(K_{tt}^{-1}, K_{tt})
15: $X_u \leftarrow$ `get_untyped_snps_for_window`$()$
16: $K_{ut} \leftarrow X_u X_{t,\text{window}}^\top$
17: $Z_{u,\text{window}} \leftarrow K_{ut} K_{tt}^{-1} Z_{t,\text{window}}$
18: Append $Z_{u,\text{window}}$ to Z_u
19: Return Z_u

3 Experiments

In this section we describe the experiments we conducted to evaluate the performance of `ARDISS` in different use cases. All experiments were conducted on real datasets in which we highlight the strengths of the method and compare its performance to that of comparison partners.

3.1 COPDGene

We obtained genotype data from participants in the COPDGene study (Regan *et al.*, 2011). The goal of this study is to identify genetic risk factors associated to chronic obstructive pulmonary disease (COPD). The study was conducted on two different ethnic groups: African Americans (AA) and non-Hispanic whites (NHW). We combined the samples of the two populations and kept 615 906 SNPs that overlapped in both datasets. Of these SNPs, those that did not fulfill the following criteria were removed from the study:

i. Minor allele frequency <0.01.
ii. Hardy–Weinberg equilibrium <1.0e–6.

Additionally, due to genotyping errors, some combinations of samples and SNPs had missing genotypes. In these cases, the missing genotypes were imputed as described in (Cho *et al.*, 2014). Of the 7993 samples in the combined dataset, 3633 are individuals diagnosed with COPD (cases) and 4360 are controls. Table 1 provides

Table 1. Details of the samples in the original populations of COPDGene

Population	Disease status			Gender	
	Case	Control	Total	Male	Female
African–Americans	821	1826	2647	1498	1149
Non-Hispanic whites	2812	2534	5346	2816	2530
Total	3633	4360	7993	4314	3679

Note: The column 'Case' refers to individuals who were diagnosed with COPD. The number of SNPs in the intersection of both populations is 615 906 and this is the starting point of our analysis.

additional details of the sample sizes. This combined dataset was then subsampled to create cohorts of mixed ethnicities as described below.

3.1.1 Randomized cohorts of mixed-ethnicity

In order to simulate cohorts of mixed ethnicities, we created 11 randomized partitions of the combined COPDGene dataset. On the two ends of the spectrum, we have homogeneous populations of 100% AA and 100% NHW samples, respectively. In between, we created cohorts with mixed ethnicities by increments of 10%, i.e. 90% AA with 10% NHW; 80% AA with 20% NHW, all the way to 10% AA with 90% NHW. Additionally, we randomly selected samples from the two populations in a stratified manner to guarantee the same ratio of cases/controls per population. We set all randomized partitions to contain the same number of samples: 2313.

3.1.2 Association analysis

For each of the randomized partitions previously described, we conducted a GWAS using a linear mixed model to account and to correct for population structure in the mixed cohort (Price *et al.*, 2010). In particular, the analyses were performed with the tool FaST-LMM (Lippert *et al.*, 2011) and for each of the 615 906 SNPs we obtained a Z-score of association. It is important to note that, when imputing summary statistics in a real setting, the genotypes of the individuals in the study will, most likely, not be available. Having access to the original genotypes in COPDGene allowed us to create randomized cohorts of varying ethnic composition and to perform the corresponding association tests. The Z-scores were the starting point to the execution of ARDISS and of the other comparison partners.

3.1.3 Randomized SNPs (typed and untyped)

To contrast the performance of ARDISS versus that of its comparison partners, we assumed that for certain SNPs the Z-scores of association were missing. Of the 615 950 SNPs across the whole genome we randomly chose 10% and flagged them as missing (these are the ones we want to impute and we refer to them as *untyped SNPs*). The remaining 90% are the *typed SNPs*, i.e. the ones for which we know the Z-score and that are used to infer the untyped ones. This randomization was repeated 10 times in order to get a good genome-wide coverage. The genomic locations of the SNPs are based on the hg19 version of the human genome. All methods were asked to impute SNPs not present in the typed set, for a total of 11 671 761 imputed SNPs.

3.2 Insomnia complaints

We obtained summary statistics from the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP; Leslie *et al.*, 2014). The full GRASP catalog currently contains publicly-available association scores of more than 2000 GWASs. Of these, we analyzed the results of a study aimed at identifying the genetic risk factors associated with insomnia complaints (Hammerschlag *et al.*, 2017). The original study was conducted on 113 006 individuals of self-reported European descent. The samples were obtained from the May 2015 release of the UK Biobank (Sudlow *et al.*, 2015).

We downloaded from GRASP the results file #2 with full summary statistics on males and females. Among the many columns with summary statistics, we used BETA: the β of the logistic regression and SE: the standard error of the logistic regression β. The Z-score for each SNP was computed as $\frac{\texttt{BETA}}{\texttt{SE}}$.

Our analysis was limited to one chromosome. Of the original 430 235 Z-scores in chromosome 12, we randomly masked 10% and imputed them. Similarly to COPDGene, this process of random masking was performed 10 times.

3.3 Evaluation of imputation performance

To assess the performance of each method, we compared the imputed Z-scores with the original Z-scores, only on the untyped SNPs. The evaluation metrics commonly found in the literature are the Pearson's correlation coefficient, the R^2 score and the root-mean-square-error (RMSE) between the imputed and original values. We computed all these metrics for the three methods but only report the correlations. Supplementary Table S1 has additional details on R^2 scores and RMSE.

3.4 Reference panel

As mentioned in Section 1, all methods used as comparison partners rely on a reference panel to perform the imputation. For our analyses we used the reference panel from the 1000 Genomes Project, Ref. 1, release 3 (Abecasis *et al.*, 2012). This panel is based on the hg19 version of the human genome and contains 14 populations grouped in 4 superpopulations (Table 2).

3.5 Implementation and speed measurements

All our runtime analyses were performed on a dedicated server running Ubuntu 14.04.5 LTS, with 2 CPUs (Intel® Xeon® E5-2620 v4 @ 2.10 GHz), 8 GPUs (NVIDIA® GeForce® GTX 1080) and 128 GB of RAM. The code was implemented in Python 3.

In order to assess the runtime required by each of the methods, we ran them independently on our server, with no other concurrent processes. We measured the imputation runtime for every chromosome separately as these processes are highly parallelizable, depending on the computing capabilities available to the users. Speed measurements were taken for the imputation of all the missing SNPs described in Section 3.1.3 when using ARDISS and ImpG-Summary. Due to the considerably slower nature of DISTMIX, we could only run it on five chromosomes (chromosomes 18 through 22). Note that this only refers to the comparison of runtimes. We used the same setup to run the speed tests for varying reference panel sizes and window sizes.

4 Results

This section presents the results on the COPDGene dataset and on the insomnia complaints study. As described in Section 3.1.1, the two COPDGene populations were used to create cohorts of mixed ethnicities. These cohorts, in turn, allowed us to do a thorough analysis of the weights computed by ARDISS and to assess the accuracy of the imputation methods under different conditions. The insomnia study provided a very realistic scenario for the imputation of

Table 2. Details of the samples in the 1000 Genomes Project that were used in our analyses as reference panel

Super population	Population	Name	Samples
AFR	ASW	African–American SW	61
	LWK	Luhya	97
	YRI	Yoruba	88
AMR	CLM	Colombian	60
	MXL	Mexican–American	66
	PUR	Puerto Rican	55
EAS	CHB	Han Chinese	97
	CHS	Southern Han Chinese	100
	JPT	Japanese	89
EUR	CEU	CEPH (Utah residents)	85
	FIN	Finnish	93
	GBR	British	89
	IBS	Spanish	14
	TSI	Tuscan	98

The four superpopulations are: AFR (African), AMR (ad-mixed American), EAS (East Asian), EUR (European).

summary statistics: one in which the data are publicly available and the imputation task has limited knowledge of the samples in the original study.

4.1 COPDGene

We applied ARDISS, ImpG-Summary and DISTMIX to the Z-scores obtained from the GWASs performed on the 11 mixed-ethnicity cohorts detailed in Section 3.1.1. Although we had access to the original genotypes and covariates, we did not run DISSCO because we wanted a more realistic imputation scenario in which the comparison partners do not require knowledge of the covariates from the original samples.

4.1.1 Weights optimization

Prior to performing imputation, ARDISS optimizes and outputs the specific weights derived for every sample in the reference panel. The individual weights were pooled according to the ethnic background of the sample with which they were associated, and used to reconstruct the population composition as depicted in Supplementary Figure S2.

As expected, ARD also detects some residual signal from the other populations and their contribution is kept in the imputation procedure. This can be seen in Figures 2 and 3. The nature of the window-based optimization causes this weak noise contamination: since full chromosome optimization of the weights is computationally unrealistic, smaller windows (100×100 SNPs) are used for optimization and the final weights are obtained by averaging over the chromosome.

The weights obtained by ARD were also pooled by individual populations and compared to the weights derived by DISTMIX from the allele frequencies of the original study samples. The overall correlation obtained by the weights is 0.839. The correlation between the populations of interest, however, is 0.936, as seen in Figure 4. Therefore, ARDISS correctly reconstructs the study population weights by using only the typed Z-scores and without the need for allele frequencies.

4.1.2 Imputation performance

Once the weights are optimized, ARDISS proceeds with the imputation of the missing Z-scores. We compared our method with

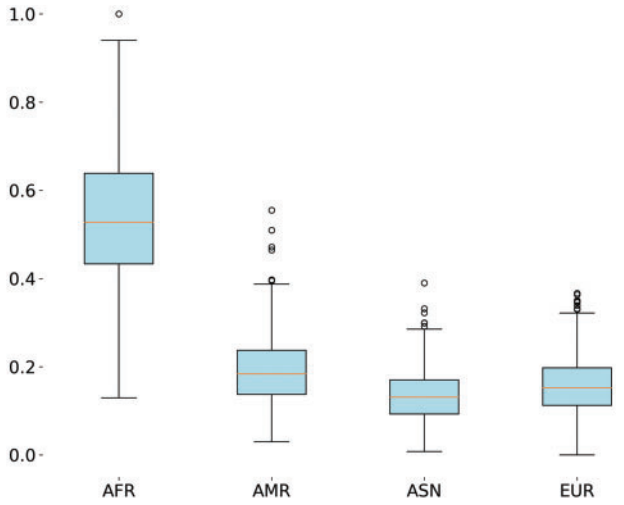

Fig. 2. Scaled contribution of weights from individual samples detected by ARD for the 100% AA | 0% NHW mix of population. Some residual weights for non-African populations are picked up. Super-populations codes are reported in Table 2. The boxplots are generated by taking the weights output by ARDISS, i.e. one per sample in the reference panel and grouping them by their super-population code

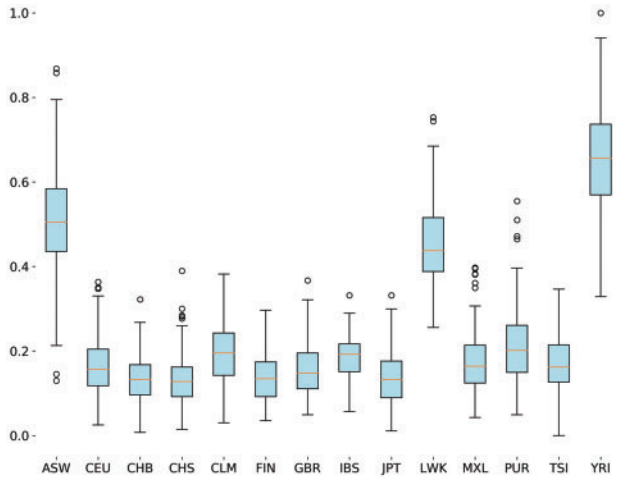

Fig. 3. Population-level individual samples' weights scaled contribution detected by ARD for the 100% AA | 0% NHW ethnicity mixture. Some residual weights for non-African populations are picked up. Population codes are reported in Table 2

ImpG-Summary and DISTMIX. When running ImpG-Summary, we used all the available samples in the reference panel to mimic a realistic setting where little is known about the original population. Using only a subset (AFR and EUR samples) did not yield better results for ImpG-Summary, as shown in Supplementary Figure S3.

ARDISS reports better performance than ImpG-Summary and DISTMIX for all mixtures of population. All the methods perform better on the homogeneous cohort of 100% non-Hispanic whites than on the cohort of 100% African American samples. There are two main possible reasons for this: (i) the imputation works better on samples of European descent because there are more European genotypes in the reference panel (379 EUR versus 246 AFR), and (ii) populations of African-descent have higher genetic diversity and less LD (Campbell and Tishkoff, 2008), making it more challenging to

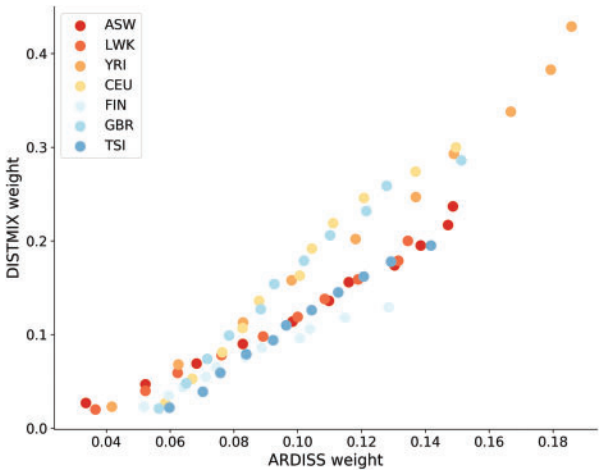

Fig. 4. Weights obtained by ARDISS (*x* axis) and by DISTMIX using the allele frequencies of the original study (*y* axis) for a selection of populations. The color code indicates the population to which the weight belongs and the different points were obtained from the different mixture of ethnicity sets. Population codes are reported in Table 2

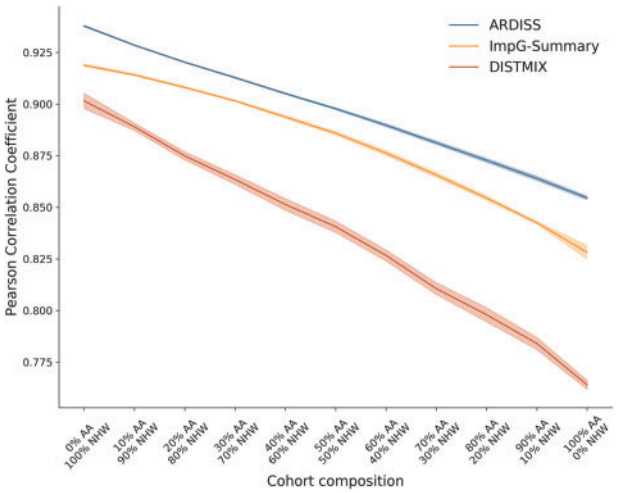

Fig. 6. Pearson's correlation coefficients obtained during full genome imputation imputation across different mixtures of ethnicity sets using ARDISS and available methods. ImpG-Summary was run using all the samples in the reference panel and DISTMIX computed the optimal weights from the allele frequencies. Replications of 10 were performed. The shaded areas represent the SD interval

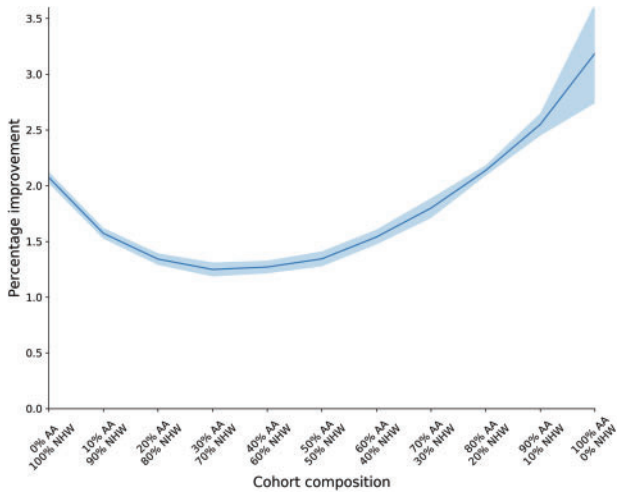

Fig. 5. Relative improvement of ARDISS over ImpG-Summary for different randomized mixed-ethnicity cohorts. ARDISS outperforms ImpG-Summary in all mixture scenarios, with both methods being equally accurate in cases of very heterogeneous cohorts (with practically 50% of AA and NHW). The shaded area represents the SD interval

cover all the haplotype diversity with few reference panels. Another surprising aspect is that DISTMIX, after optimizing the weights with allele frequencies of the original study, still performs worse than ImpG-Summary. While no proper comparison of the two methods was reported before, this could imply that the cohort genotype correlation matrix derived by DISTMIX is not taking advantage of the weights.

When comparing with ImpG-Summary, the performance improvement is more noticeable with non-mixed cohorts. The 0% AA | 100% NHW, 90% AA | 10% NHW and 100% AA | 0% NHW are the mixtures for which ARDISS is better able to spot the composition with 2.08%, 2.55% and 3.20% improvements over ImpG-Summary, respectively (Fig. 5). Once the mixture approaches 50%|50%, the gain in weighting individual contributions decreases (1.34% improvement over ImpG-Summary), as the weight distribution gets closer to having all equivalent weights (this is the case for ImpG-Summary with all samples in the reference panel). The improvement on the African–American population is considerably higher. This can be explained by the ability ARDISS has to draw information from other individual samples that might not be in the same super population group. Similarly, ARDISS does considerably better than DISTMIX, with an improvement ranging from 4.03% for 0% AA | 100% NHW to 11.85% for 100% AA | 0% NHW. Supplementary Figure S4 provides more details on this. We also evaluated the performance of ImpG-Summary using (i) only EUR samples, (ii) only AFR samples and (iii) a combination of both on chromosome 12. Additionally, we computed the performance of DISTMIX when manually fed with 'best-guess' weights, an approach that is somewhat realistic in a setting for which no information about the original population is known. For each ethnicity mixture, we attributed the effective percentage of weights to the ASW (Americans of African Ancestry in Southwest USA) and to the CEU [Utah Residents (CEPH) with Northern and Western European Ancestry]. None of these approaches yielded better results than the results reported in Figure 6 and the 'best-guess' weights resulted in the worst performance (data now shown). For a thorough examination of alternative scenarios, please refer to Supplementary Figure S3.

Additionally, we considered the effect of the window size on the imputation accuracy. Performance initially increases with increasing window size, but starts deteriorating for larger windows. This is due to the non-overlapping nature of the ARD step: with larger window sizes, that are potentially overlapping multiple LD regions, the obtained weights are more evenly distributed and no sample is clearly selected. Moreover, during the imputation step, long-distance, low-LD SNPs covered by large windows add noise to the imputed values, decreasing the quality of the imputation. An overview of the performance of ARDISS for various window sizes can be found in Supplementary Figures S5 and S6. We observed optimal performance with a window size of 100 and while this value might depend on the GWAS in use and its available typed Z-scores, the general behavior is similar in other studies.

Table 3. Example percentage of recovered top 100 SNPs after imputation on chromosome 12 for the 10% AA | 90% NHW cohort

	Typed	ARDISS	ImpG-summary
With covariates	100	70	55
Without covariates	100	64	61

Table 4. Imputation performance on the publicly available summary statistics for Insomnia GWAS reported in correlation between imputed and typed SNPs

Method	Insomnia	
	Correlation	RMSE
ARDISS	**0.956 ± 0.001**	**0.093 ± 0.002**
ImpG-summary	0.889 ± 0.003	0.218 ± 0.005

Note: Bold characters indicate the best performances.

In order to evaluate whether ARDISS is adversely affected from not accounting for covariates, we decided to analyze the percentage of recovered top hits. As important as the correlation between imputed and original Z-scores is, it may overlook the ranks of Z-scores. By this we mean it is also very relevant that, if the original Z-score of a SNP ranks high when compared to the rest, then the imputed Z-score should also rank high. To validate this, we selected the top 100 SNPs from the *untyped* SNPs, i.e. the highest absolute value of the Z-scores marked as missing and compared them with the top 100 imputed Z-scores. Table 3 shows how ARDISS recovers a comparable number of top hits when the original association test is conducted with or without covariates. In the case of COPD, the covariates used as confounders where (i) age and (ii) pack-years of smoking.

4.2 Insomnia complaints

For the insomnia study, we only compared ARDISS to ImpG-Summary due to its wider adoption and ease of use. As mentioned in Section 3.2, the GWAS on insomnia complaints conducted by Hammerschlag et al. used samples from the UK Biobank, a large dataset of self-reported traits and genotypes. The participants' ethnicities were also self-reported, making them subject to uncertainty and, thus, a perfect use case for our method. Table 4 reports the results obtained by the two methods. ARDISS clearly outperforms ImpG-Summary on the imputation task, suggesting that it successfully evaluates the study population's structure and ideally imputes values for it. This result highlights the advantage of using an adaptive method such as ARDISS in a setting where the ethnic background of the participants in a study is not clearly defined.

As mentioned in Section 1, a method like ARDISS can be easily extended to perform imputation of β values. In this study, the correlation between the imputed and masked values is 0.804 ± 0.008 for β values on chromosome 12. The imputation accuracy is lower than for Z-scores because the Z-score—defined as the ratio of β over its standard error—contains more information about the association between the SNP and the phenotype.

4.3 Speed performance

ARDISS leverages state-of-the art scientific computing libraries and can be deployed on GPU architectures to speed up the ARD computation. When compared with available solutions, ARDISS showed

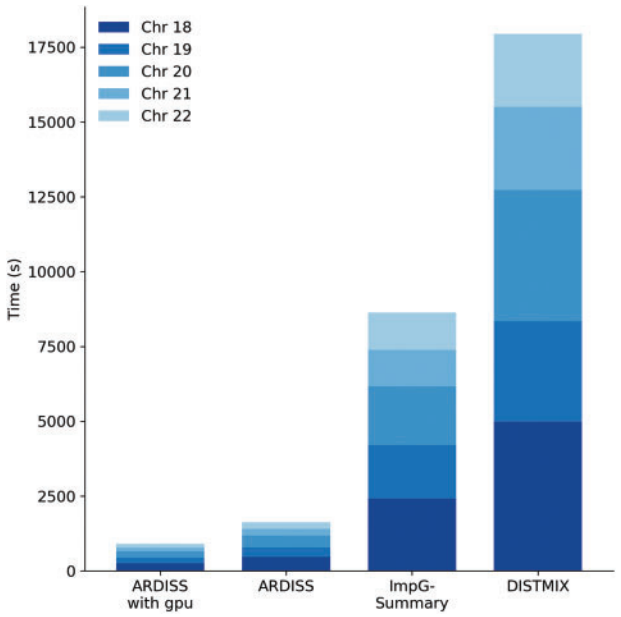

Fig. 7. Breakdown of the run times for sequential imputation of summary statistics across chromosomes 18 to 22

large improvements in runtime performance. The total runtime required to impute the missing SNPs described in Section 3.1.3 using ImpG-Summary was of ∼22 h (79 205.53 s) compared to ∼4h15 min (15 287.61 s) for ARDISS and ∼2h20 (8530.12 s) when using ARDISS on a GPU. Alternatively, users with large time constraints also have the option to omit ARD and get even faster imputation: ∼35 min (2118.95 s) for the whole genome, at a cost of slightly less accurate imputation. On the other end of the spectrum, DISTMIX was too slow for full sequential imputation and could only be measured for a subset of chromosomes. In order to impute 1 131 674 SNPs on chromosomes 18 to 22, DISTMIX took ∼5 h (17 947.63 s), compared to the ∼15 min (907.50 s) of ARDISS on a GPU. Figure 7 shows the sequential run times of ARDISS, ImpG-Summary and DISTMIX on a subset of chromosomes. Since these methods are usually ran in parallel to impute multiple chromosomes separately, we computed the average ratio of run times of ARDISS and comparison partners across chromosomes: when using GPUs the user can expect, on average, a method that is 9.38 times faster than ImpG-Summary and 19.88 times faster than DISTMIX. When dropping the ARD step, the fold change increases to 38.35 and 83.10, respectively.

5 Discussion

The growing interest for genomics methods based on GWAS summary statistics has brought a panoply of tools for imputing missing values. However, as we highlighted in our analysis, some of the available methods suffer from usability issues. DISTMIX relies entirely on allele frequencies for accurate imputation results. While these data might be accessible in certain cases, their exchange has been severely reduced after they were proven to be an effective mean to identify participants in a GWAS (Homer *et al.*, 2008). Moreover, as privacy concerns constantly grow, access to sensitive information will not become easier in the future. On the contrary, with more informed study participants, research groups will likely tighten their access policies for external collaborators. Furthermore, large public repositories of SNP-trait association data, such as the GWAS

Catalog (Welter *et al.*, 2014), provide summary statistics from which Z-scores can be derived, i.e. SNP effect and its standard error, but seldom report allele frequencies and never provide covariates from the original samples. On the one hand, a method like DISTMIX is severely handicapped in the case of missing allele frequencies. On the other hand, a method that requires covariates on the original data, like DISSCO, cannot simply be executed when these are not present (this is a documented limitation of the tool). In fact, due to its requirements, it can be argued that the usability of DISSCO is circumscribed to a small niche of users. These users have access to the original genotype data in a study (and its covariates) but prefer to impute summary statistics on the missing SNPs rather than perform the more accurate (yet more computationally intensive) task of imputing the missing genotypes with IMPUTE2, MaCH or others.

Because of its own merits, ImpG-Summary offers excellent performance on certain well-defined datasets, but may lack the flexibility necessary to impute missing values in slightly more complex studies. In our experiments, the GWAS study on insomnia shows that an easily adaptable method such as ARDISS has the potential to yield much better imputation performance, even for a self-reported homogeneous cohort. Our motivation to develop ARDISS was to simplify the task of imputing summary statistics by providing a unique and robust solution that encompasses all scenarios described in Figure 1, while at the same time, providing superior imputation accuracy and better runtimes. In fact, another aspect that was central in the development of ARDISS was to improve the runtime efficiency by exploiting parallel computing methods and highly-efficient open-source libraries.

Finally, the ever increasing body of publicly available results from association studies in plants, humans and other model organisms, enables researchers that use GWAS results to ask questions that go beyond the SNP-trait association. The integration of Z-scores from different studies makes the imputation of missing values a necessity which, coupled with the limited time a researcher has to gather additional sample information from a study publication, creates opportunities for software tools that minimize the need for additional data. Therefore, a method like the one we propose here, which accurately imputes Z-scores for ethnically-mixed populations without requirements for additional input from the user, is bound to be successful and widely adopted by the community.

6 Conclusion

We presented ARDISS, a fast, accurate and adaptable method to impute missing Z-scores while inferring the underlying population composition without the need for any extra information such as allele frequencies or covariates of the original study population. Our method matches typical use-case scenarios betten than other available solutions. It outperforms other methods, not only in imputation performance but also in speed as it is highly parallelizable on platforms with available GPUs. It entirely relies on open-source libraries and the code is publicly available online.

Funding

The COPDGene project was supported by award number R01HL089897 and award number R01HL089856 from the National Heart, Lung, And Blood Institute. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board composed of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens, GlaxoSmithKline and Sunovion.

Conflict of Interest: none declared.

References

Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1, 092 human genomes. *Nature*, **491**, 56–65.

Alonso-Blanco,C. *et al.* (2016) 1, 135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, **166**, 481–491.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Campbell,M.C. and Tishkoff,S.A. (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.*, **9**, 403–433.

Cho,M.H. *et al.* (2014) Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Resp. Med.*, **2**, 214–225.

Erlich,Y. and Narayanan,A. (2014) Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.*, **15**, 409–421.

Freilinger,T. *et al.* (2012) Genome-wide association analysis identifies susceptibility loci for migraine without aura. *Nat. Genet.*, **44**, 777–782.

Hammerschlag,A.R. *et al.* (2017) Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.*, **49**, 1584–1592.

Homer,N. *et al.* (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet.*, **4**, e1000167.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

Kirby,A. *et al.* (2010) Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics*, **185**, 1081–1095.

Lee,D. *et al.* (2013) DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics*, **29**, 2925–2927.

Lee,D. *et al.* (2015) DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. *Bioinformatics*, **31**, 3099–3104.

Leslie,R. *et al.* (2014) Grasp: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, **30**, i185–i194.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Lin,T. *et al.* (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.*, **46**, 1220–1226.

Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

MacKay,D.J.C. (1994) Bayesian non-linear modelling for the energy prediction competition. *ASHRAE Trans.*, **100**, 1053–1062.

Mackay,T.F. *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, **482**, 173–178.

Matthews,A.G.D.G. *et al.* (2017) GPflow: a Gaussian process library using TensorFlow. *J. Mach. Learn. Res.*, **18**, 1–6.

Meesters,C. *et al.* (2012) Quick,'imputation-free' meta-analysis with proxy-SNPs. *BMC Bioinformatics*, **13**, 231.

Meijon,M. *et al.* (2014) Genome-wide association study using cellular traits identifies a new regulator of root development in Arabidopsis. *Nat. Genet.*, **46**, 77–81.

Pasaniuc,B. and Price,A.L. (2017) Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.*, **18**, 117–127.

Pasaniuc,B. *et al.* (2014) Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics (Oxford, England)*, **30**, 2906–2914.

Price,A.L. *et al.* (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.

Rasmussen,C.E. and Williams,C.K.I. (2006) *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA, USA.

Regan,E.A. *et al.* (2011) Genetic epidemiology of COPD (COPDGene) study design. *COPD: J. Chronic Obstructive Pulmonary Dis.*, **7**, 32–43.

Servin,B. and Stephens,M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.*, **3**, e114.

Sudlow,C. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.

Togninalli,M. *et al.* (2018) The AraGWAS Catalog: a curated and standardized *Arabidopsis thaliana* GWAS catalog. *Nucleic Acids Res.*, **46**, D1150–D1156.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Wen,X. and Stephens,M. (2010) Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.*, **4**, 1158.

Xu,Z. *et al.* (2015) DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics*, **31**, 2434–2442.

Zhao,K. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.*, **2**, 467.