Published in final edited form as:
Nat Genet. 2016 April ; 48(4): 466–472. doi:10.1038/ng.3513.

A multiple phenotype imputation method for genetic studies

Andrew Dahl[1,8], **Valentina Iotchkova**[2,3,8], **Amelie Baud**[3], **Åsa Johansson**[4], **Ulf Gyllensten**[4], **Nicole Soranzo**[2], **Richard Mott**[1], **Andreas Kranis**[5,6], and **Jonathan Marchini**[7,1]

[1]The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

[2]Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

[3]The European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK

[4]Department of Immunology, Genetics, and Pathology, SciLifeLab Uppsala, Uppsala University, Uppsala, Sweden

[5]Aviagen Ltd, Newbridge, United Kingdom

[6]The Roslin Institute, University of Edinburgh, Midlothian, United Kingdom

[7]Department of Statistics, University of Oxford, Oxford, UK

Abstract

Genetic association studies have yielded a wealth of biologic discoveries. However, these have mostly analyzed one trait and one SNP at a time, thus failing to capture the underlying complexity of these datasets. Joint genotype-phenotype analyses of complex, high-dimensional datasets represent an important way to move beyond simple GWAS with great potential. The move to high-dimensional phenotypes will raise many new statistical problems. In this paper we address the central issue of missing phenotypes in studies with any level of relatedness between samples. We propose a multiple phenotype mixed model and use a computationally efficient variational Bayesian algorithm to fit the model. On a variety of simulated and real datasets from a range of organisms and trait types, we show that our method outperforms existing state-of-the-art methods from the statistics and machine learning literature and can boost signals of association.

Correspondence: Professor Jonathan Marchini Department of Statistics University of Oxford 1 South Parks Road Oxford OX1 3TG, UK Tel: +44 (0)1865 271125 Fax: +44 (0)1865 281333 marchini@stats.ox.ac.uk.

[8]Joint first authors of this paper

Author contributions

A.D, V.I and J.M developed the method. A.D carried out all analysis. J.M and A.D wrote the paper. A.B and R.M provided extensive advice on analysis of the rat GWAS dataset. A.J and U.G provided the NSPHS dataset. N.S provided the UKNBS dataset. A.K provided the chicken dataset and provided advice on analysis. All authors critiqued the manuscript.

Competing financial interests statement

AK is an an employee of Aviagen Ltd, a poultry breeding company that supplies broiler breeding stock world wide. AK also holds an Industry Fellowship from the Royal Society and is part-time based in The Roslin Institute.

Introduction

Genome-wide association studies (GWAS) have successfully uncovered many associated loci. Such approaches typically analyze thousands of nominally unrelated individuals and search for correlations between genetic variants and a single trait of interest. However, a complete characterization of the etiology of most traits remains elusive. This may be because the GWAS approach is quite crude, in that much of the biology between sequence and phenotype remains unmeasured. Large scale phenotyping is starting to generate invaluable data that can be harnessed by geneticists [1].

This observation motivates the analysis of multiple phenotypes, traits and sub-phenotypes, a direction that is increasingly prominent in the literature of human, plant and animal genetics [2-5] . The advantages of analyzing multiple phenotypes related to, or underlying, a phenotype of interest include boosting power to detect novel associations[6], measuring heritable covariance between traits [7] and the potential to make causal inference between traits [8].

At the same time, harnessing genetic relatedness, even amongst nominally unrelated samples, to boost power in association studies is becoming increasingly prevalent. Mixed models, re-emerging from the linkage and animal genetics literature[9-11] , are now routinely used to search for associations in the presence of relatedness or population structure and to estimate the additive genetic component of heritability. However, until recently these analyses have mostly proceeded one trait at a time.

In this paper, we consider the analysis of multiple correlated phenotypes observed on correlated samples, which arises with related individuals, cryptic relatedness, population structure or polygenicity. Crucially, the vast majority of methods for multiple phenotypes rely on all samples having fully observed phenotypes[3,6]. However, as the number of phenotypes increases the chance that at least one observation is missing increases exponentially. Removal of all samples with a missing phenotype will reduce sample size, thus attenuating the power of any statistical inference. For example, a range of real studies removed between 3%-31%[12-17] of samples. Other studies completely removed phenotypes with high levels of missing data, and imputed remaining missing data with off-the-shelf methods from mainstream statistics [18-20]. While re-phenotyping of samples is ideal, it is typically expensive or infeasible [21].

We propose a method to impute missing phenotypes in related samples, which will likely be a crucial first step for many downstream analyses. In this setting correlations will exist between phenotypes and between samples, and *both* are useful in predicting missing observations. We propose a Bayesian multiple phenotype mixed model and use a Variational Bayesian (VB) method to fit the model. We assume that the kinship between individuals in a study is known *a priori* from genetic data[22] or a pedigree. This information enables the model to decompose the correlation between traits into a genetic and a residual component. A notable feature of our method is that it can handle hundreds of traits. We call our method PHENIX.

We validate our approach with an extensive simulation study, representative of a variety of genetic studies of humans and other organisms. We compare our method to approaches that ignore either the correlations between samples or the correlations between traits, and to state-of-the-art missing data imputation techniques from mainstream statistics and machine learning. We also apply our method to five real datasets on a variety of traits from humans[2,23], yeast[24], rats[25], chickens [26] and wheat [27]. In all simulated and real datasets we show evidence that our method outperforms the competing methods in accuracy and is computationally efficient. We also apply the method to a rat GWAS of 140 phenotypes to illustrate how the method can be used to boost signals of association. Finally, we discuss the usefulness of this approach, the range of relevant datasets that the method could be applied to, and how the method might be developed further in the future.

Results

Simulations

We simulated datasets with N=300 individuals and P=15 traits varying the level of relatedness between individuals and the heritability of the traits. A standard multiple phenotype mixed model (MPMM) was used to simulate phenotypes with an underlying genetic covariance, as well as added environmental, or residual, correlation. For the genetic covariance between traits we used a model with a range of positive and negative correlations between the traits. For the residual covariance we added randomly correlated noise to the phenotypes. We varied the heritability of the traits by adjusting the relative contributions of the genetic and residual covariance terms. We used two models for relatedness between samples: Model 1 used an empirical kinship matrix derived from the Northern Sweden Population Health Study (NSPHS) [23]; Model 2 simulated 75 independent families of 4 full siblings. Missing data was added completely at random at the 5% level. The true values of missing data were kept to measure performance. We averaged results over 100 datasets simulated under each scenario. More details are given in the **Online Methods**.

We fit our method (PHENIX) to each of the simulated datasets to infer point estimates of the missing phenotypes. We assessed performance by measuring the correlation between these imputed phenotypes and their true hidden values. The results are shown in Figure 1 for both levels of relatedness. We compared our method to a range of other imputation methods from the statistical genetics, mainstream statistics and machine learning literatures (Table 1, **Online methods**). These methods model different aspects of the correlation structure in the data, in most cases ignoring genetic or phenotypic correlations; PHENIX models both aspects. Results of the methods using a mean squared error (MSE) metric and timing information are shown in Supplementary Figure 1 and Supplementary Note.

The overall pattern from Figure 1 is that PHENIX outperforms all other methods over the full range of heritability. As heritability increases the difference between PHENIX and the next best method increases. A number of other interesting patterns also emerge. Ignoring correlations between phenotypes (LMM – green line) is mostly a much worse assumption than ignoring correlations between samples (MVN – blue line), except at very high heritabilities and high levels of relatedness between samples (Model 2). In fact, ignoring correlations between samples does remarkably well, especially considering MVN is the

fastest method in our comparisons. However, the performance of MVN suffered in some real datasets with high relatedness (Figure 2) so we do not recommend it for general use. TRCMA (pink line) and SOFTIMPUTE (cyan line) seem to perform roughly equally well, and better than MICE and kNN (brown and grey lines respectively). This is likely because the former two methods partially model sample relatedness, whereas the latter two methods only model phenotypic correlations. Most methods were fast enough to be practical, although we found TRCMA to be prohibitively slow in most settings (Supplementary Note).

Increasing levels of relatedness between samples increases the accuracy of PHENIX and LMM. Both of these methods explicitly take account of the relatedness between samples via the kinship matrix. For example, when the heritability of the traits is h^2=0.3, the imputation correlation of PHENIX is 0.63 and 0.67 on Model 1 (NSPHS) and Model 2 (sibs) respectively.

As heritability increases the performance of all the best performing methods decreases, but then increases slightly again as heritability approaches 1. This occurs because the overall correlations between traits are a mixture of genetic and environmental correlations. At intermediate heritability the genetic and environmental correlations tend to cancel each other out, attenuating the performance of methods that harness phenotypic correlations. To highlight this effect we carried out simulations in which genetic and environmental covariances are the inverses of each other. At intermediate values of heritability the performance of all methods suffers (Supplementary Figure 2).

When the number of samples is increased to N=1000 and phenotypes to P=50 the performance of PHENIX improves compared to the other methods, especially for Model 1 which uses an empirical kinship matrix derived from the NSPHS study (Supplementary Figure 3). As the genetic correlation between traits increases, the residual contribution becomes less important and thus the utility from partitioning the covariance is attenuated; this means the gap between PHENIX and other methods shrinks. Conversely, when the genetic correlation shrinks, PHENIX increasingly outperforms the others (Supplementary Figure 4). Increasing the missing data rate to 10% degrades performance for all methods, especially when there are few close relationships between samples (Supplementary Figure 5). We investigated the effects of non-random missingness (Supplementary Figure 6), unmodelled, shared environmental effects (Supplementary Figure 7) and non-normally distributed phenotypes (Supplementary Figure 8), which all act to reduce performance in general. However, PHENIX remains the best performing method in all scenarios.

A likely main use of PHENIX is to impute missing phenotypes ahead of association testing of phenotypes with genome-wide SNP data. This might proceed by testing phenotypes one at a time, or by using a multi phenotype association test. As such it is important to show that our approach leads to valid statistical tests. Using simulated phenotype data and real genotype data from the NSPHS cohort (described below) we find that association testing after imputation results in well calibrated p-values under the null (Supplementary Figure 9).

There is a large literature on multi-phenotype tests [3,6,28-30] and there seems wide consensus that these tests can lead to an increase in power over single phenotype tests in many realistic

scenarios. We assessed whether imputing missing phenotypes can increase in power when testing a SNP for association. We find that imputation can lead to an increase in power when testing either one phenotype at a time, or when using a multi-phenotype test (Supplementary Figures 10 and 11). Intuitively, one of the main reasons this occurs is that imputation increases the sample size used in the test.

Real data

To further illustrate the usefulness of PHENIX we imputed missing phenotypes in several real datasets. We applied the method to a range of different organisms to illustrate that our method will be useful in a wide variety of settings and across a diverse set of phenotypes used in real genetic studies. Animal and plant studies almost always use related samples, due to study design constraints, but in some cases, like Arabidopsis, unrelated samples with considerable population structure are used.

The datasets are hematological measurements in the UK Blood Services Common Control (UKBS) collection that was studied by the HaemGen consortium [2], glycans phenotypes in the NSPHS study[4,23], phenotypes related to six disease models and measures of risk factors for common diseases in outbred rats[25], phenotypes measuring growth of yeast under different conditions [24], phenotypes relevant to a genomic selection program in a multigenerational chicken pedigree[26] and traits related to growth and yield in an inter-cross population for winter-sown wheat[27]. Table 2 details the properties of these datasets.

Each of these datasets has a different level of missing data. We created new datasets by increasing levels of missing data, keeping the true values to assess imputation performance. We applied the various imputation methods to these datasets and measured performance using the correlation between the imputed and true values. The results for each of the six datasets are presented in Figure 2, where imputation correlation (y-axis) is plotted against missing data percentage (x-axis). The true level of missing data is highlighted as a vertical, dashed black line.

As in the simulated datasets, PHENIX is the most accurate method across all six of the datasets, except at extreme levels of missingness. For realistic levels of missing data, near the actual levels in the datasets, PHENIX clearly outperforms the other methods in the yeast and chicken datasets, but the difference is smaller on the human, rat and wheat datasets. On all six datasets TRCMA, SOFTIMPUTE and MVN perform almost the same. As with the simulated data, these 3 methods tend to outperform MICE, which in turn tends to outperform kNN.

The single trait LMM method is overall the worst performing method, however it does reasonably well on the yeast and chicken datasets, where the trait heritabilities and levels of sample relatedness are high and traits are relatively uncorrelated compared to the other datasets. Appropriately, these are the datasets where PHENIX substantially outperforms TRCMA, SOFTIMPUTE and MVN.

For the human NSPHS and wheat datasets we fit a standard Multiple Phenotype Mixed Model (MPMM), with an EM algorithm[31], only to those individuals with fully observed

phenotypes, and used the estimated parameters to impute missing phenotypes in other individuals, following others[3]. MPMM will not run on the human UKNBS, yeast, chicken or rat datasets where the number of phenotypes and levels of missingness produce no samples with complete observations. When it is possible to apply this method we observed (Figure 2 **– purple lines**) that its performance drops off considerably. As the amount of missing data increases the number of samples with completely observed phenotypes will exponentially decrease, which will harm parameter estimation and subsequent imputation performance.

Application to Rat GWAS

To assess the utility of our method in the GWAS setting we re-analyzed the data from the Rat Genome Sequencing and Mapping Consortium. Specifically, we imputed all the missing phenotypes and covariates available in the deposited dataset. We then carried out GWAS for the 140 most biologically relevant phenotypes (those mapped in the original study[25]) at the 24,196 genomic locations at which HAPPY[32] descent probabilities had been calculated (see **Online Methods**). The amount of missing data in these 140 phenotypes varies from 1.5% to 87% (median=16.6%). We then compared these results to GWAS performed on the phenotypes without imputation.

In much the same way that information scores are used when carrying out downstream analyses such as GWAS on imputed genotypes[33], it is desirable to assess the accuracy of phenotype imputation. To achieve this, we added extra missing data, re-imputed the missing phenotypes and then calculated an imputation squared correlation (r^2) for each phenotype using the held out data (see **Online Methods**). This metric can be automatically calculated by the imputation functions in our `R` package, and experiments suggest that the measure is very accurately calibrated (Supplementary Figure 12). To choose a useful threshold for r^2, we used experience of filtering genotype imputation information scores, which typically filter at some value between 0.3-0.4. Ultimately, we used 82 phenotypes with $r^2 > 0.36$. The amount of missing data being imputed may also be a useful phenotype summary to consider when interpreting imputation results.

Figure 3 compares the results of the imputed and un-imputed rat GWAS for all 140 phenotypes. To report results we applied a conservative p-value threshold of $-\log_{10}(p) > 10$. We only plot p-values for genomic locations that are maximal in a 6 Mb window (+/− 3 Mb). These are plotted against the maximum $-\log_{10}(p)$ in the same 6 Mb window in the complementary (imputed or un-imputed) GWAS. Grey points are those for which $r^2 < 0.36$. The cluster of grey points with imputed $-\log_{10}(p) < 10$ and un-imputed $-\log_{10}(p) > 10$ all correspond to phenotypes with very low r^2 and high levels of missing data demonstrating that filters on r^2 and missingness can identify when imputation results should be viewed with caution.

The figure highlights that there are circumstances where phenotype imputation has a good imputation r^2 and acts to increase the signal of association (red and blue points). A cluster of associations (red points) all correspond to three related platelet phenotypes (mean platelet volume (MPV), mean platelet count (MPC) and platelet distribution width (PDW)) over an extended region of chromosome 9 between 50-80Mb. Figure 4 shows the imputed and un-imputed GWAS for these three phenotypes in this region, together with histograms of the

phenotype data, r^2 and missingness metrics. The plot highlights several peaks of association that harbor a number of genes related to platelet aggregation, adhesion and function (*Igfbp2* and *Igfbp5*[34], *Fn1*[35], *Epha4*[36], *Cps1*[37,38], *Ctla4*[39], *Hspd1*[40]).

An additional association (blue point) in Figure 3 corresponds to a region associated with the CD25highCD4 phenotype (Proportion of CD4+ cells with high expression of CD25). Figure 5 shows the imputed and un-imputed GWAS for CD25highCD24 as well as two other related T cell phenotypes that also show increased levels of association (Abs_CD25CD8 (Absolute CD25+CD8+ cell count) and pctDP (Proportion of CD4−CD8− T cells)). The plots show a clear elevation of association in the region around the *Tbx21* (T-bet) gene which plays a key role in T helper cell differentiation [41].

Discussion

Missing data is a pervasive feature of the statistical analysis of genetic data. Whether it be unobserved genotypes or latent population structure in GWAS studies, partially observed genotypes in low-coverage sequencing studies, or unobserved confounding effects in GWAS and eQTL studies, accurate and efficient methods are needed to infer missing data and can often substantially enhance analysis and interpretation. In this paper, we have proposed a general method to impute missing phenotypes in samples with arbitrary levels of relatedness, population structure and missingness patterns.

While there exists a range of different methods for imputing missing data in the general statistics literature, our method focuses specifically on continuous phenotypes in genetic studies, where there is often known, or measureable, relatedness between samples. Our method leverages this relatedness to partition the phenotypic correlation structure into a genetic and a non-genetic component and to boost imputation accuracy. Using simulated and real data we have shown that our method of imputing missing phenotypes outperforms state-of-the-art methods from the statistics and machine learning literature. In the burgeoning literature of papers on mixed models applied to genetics this is the first approach we are aware of that allows for missing phenotypes.

Key features of our method are (a) boosting signals of association in GWAS when imputation quality is high, (b) not having to discard samples with partially observed phenotypes, (c) a way of assessing imputation performance via our r^2 metric, and (d) being able to handle large numbers of phenotypes in a mixed model framework. Our results of applying the method to 140 phenotypes from a rat GWAS study illustrate these key features. However, our results also suggest that imputation will not *always* boost signal, in much the same way the genotype imputation does not always increase levels of association. When imputation quality is demonstrably poor, and missingness is high, then imputation may attenuate the association signal. We recommend filtering phenotype imputation results with the same care and attention as is routine in the analysis of genotype imputation.

The method could be further developed to relax the assumption of normality to directly allow for heavy tailed distributions, or to explicitly allow for binary and categorical traits. However, our simulations have shown that PHENIX remains the currently best performing

method in some of these scenarios. In other work (unpublished data; V.I and J.M) we are extending the model to test a SNP for association with multiple phenotypes, using a spike-and-slab mixture prior on effect sizes to allow for only a subset of phenotypes to be associated. Incorporating significant SNPs into our model would likely increase imputation accuracy, especially in model organisms where loci with large effects are common; multi-trait extensions of whole-genome regression models that, intuitively, integrate SNP selection into an LMM-type model[42] could possibly improve accuracy yet further. Higher dimensional datasets, such as '3D' gene expression experiments across multiple samples, genes and tissues[43] also have missing 'phenotypes' which may be reliably imputed to boost signal in downstream analyses.

This paper addresses single imputation (SI) of phenotypes, and ignored uncertainty in these imputed values can, in theory, invalidate subsequent analyses. Multiple imputation (MI), the standard solution, propagates imputation uncertainty by performing downstream analyses on many imputed datasets, each drawn independently from their posterior. By aggregating results over these multiple datasets, MI delivers valid conclusions for any downstream analysis, regardless the imputation quality[44]. Though drawing from our approximate posterior is not a solution, as VB provably underestimates posterior covariance, it is possible to recover calibrated covariance estimates for the imputed values[45]; doing this computationally efficiently is non-trivial and left to future work. We note that our r^2 and missingness metrics dramatically attenuate this shortcoming of SI, as we only analyze phenotypes where imputation uncertainty is smallest; morever, simulations (Supplementary Figure 9) and biologically plausible results (Figures 4 and 5) suggest that SI can uncover novel true positive results in our context.

There is increasing evidence that established loci can affect multiple traits at the same time (pleiotropy)[46] and that this may explain the comorbidity of diseases[47]. It thus seems likely that studies that measure multiple phenotypes, endo-phenotypes and covariates on the same subjects will have to become more common if we are to further elucidate the causal pathways underlying human traits and diseases. Statistical methods that jointly analyze high-dimensional traits and integrate multiple 'omics' datasets will be central to this work.

Online methods

Matrix Normal Models

We develop our model using Matrix Normal (MN) distributions [55]. If an $N \times P$ random matrix X has a Matrix Normal distribution, this is denoted as

$$X \sim MN\left(M, R, C\right)$$

which implies

$$vec\left(X\right) \sim N\left(vec\left(M\right), C \otimes R\right)$$

where *vec(X)* is the column-wise vectorization of X, M is the $N \times P$ mean matrix, R is an $N \times N$ row covariance matrix, C is a $P \times P$ column covariance matrix, and $\otimes$ denotes the Kronecker product operator.

A Bayesian Multiple Phenotype Mixed Model

We let Y be an $N \times P$ matrix of P phenotypes (columns) measured on N individuals (rows). We assume that Y is partially observed and that each phenotype has been de-meaned and variance standardized. A standard Multiple Phenotype Mixed Model (MPMM) has the form

$$Y = U + \varepsilon \quad (1)$$

where U is an $N \times P$ matrix of random effects and ε is a $N \times P$ matrix of residuals and are modeled using Matrix normal distributions as follows

$$\begin{aligned} U &\sim MN(0, K, B) \\ \varepsilon &\sim MN(0, I_N, E) \end{aligned} \quad (2)$$

In this model K is the $N \times N$ kinship matrix between individuals, B is the $P \times P$ matrix of genetic covariances between phenotypes and E is the $P \times P$ matrix of residual covariances between phenotypes.

In our Bayesian MPMM (PHENIX), we fit a low-rank model for U, such that $U = S\beta$, where

$$\begin{aligned} S &\sim MN(0, K, I_P) \\ \beta &\sim MN(0, I_P, \tau^{-1} I_P) \end{aligned} \quad (3)$$

where τ is a regularization parameter. We use a Wishart prior for the residual precision matrix E^{-1}

$$E^{-1} \sim Wi\left(P+5, \frac{1}{4} I_P\right) \quad (4)$$

We fit this model using Variational Bayes (VB) [56], which is an iterative approach of approximating the posterior distribution of the model parameters. We treat missing phenotypes, which we denote as $Y^{(\text{miss})}$, as parameters in the model and infer them jointly with S, β and E. We impose that the approximate posterior factorizes over the partition $\{Y^{(miss)}, S, \beta, E\}$. The full details of the VB update equations are given in the Supplementary Methods. We let $\tau = 0$ which leads to the least low rank estimate of $U = S\beta$ under our model.

Having fit the model, for each sample with missingness the resulting approximate posterior distribution has the form of a multivariate normal distribution

$$Y_i^{(miss)} \sim N\left(\mu_i, \sigma_i^2 | Y \backslash Y^{(miss)}\right) \quad (5)$$

We use the posterior mean μ_i to impute $Y_i^{(miss)}$.

Other methods

We applied several other methods for imputing missing phenotypes from the statistical genetics, mainstream statistics and machine learning literatures. These methods are summarized briefly in Table 1. We provide brief details of each method here and more extensive details in the Supplementary Methods.

MVN—We assessed the effect of ignoring relatedness between individuals by fitting a simple multivariate normal model of covariance between traits [44]. The model is

$$Y_{i-} \sim N\left(\mu, \sigma^2\right) \quad (6)$$

where Y_{i-} denotes the ith row of the phenotype matrix Y. We use an expectation-maximization (EM) algorithm that allows for missing phenotypes to fit the model. This method was implemented in R.

LMM—To examine the effect of ignoring correlations between traits we applied a single trait linear mixed model (LMM) to each trait separately of the form

$$Y_{-p} \sim N\left(o, \sigma_{pg}^2 K + \sigma_{pe}^2 I_N\right) \quad (7)$$

where Y_{-p} denotes the pth phenotype. Missing phenotypes for each trait were predicted using the BLUP estimate of the random effect. This method was implemented in R.

MPMM—We directly fit an MPMM (eqns. 1-2) to only those individuals with completely observed observations, using an EM algorithm (see Supplementary Methods) and used the resulting parameter estimates in the model to impute the missing observations. This method was implemented in R.

TRCMA—The transposable regularized covariance model (TRCM) approach[54] fits a mean restricted matrix normal model of the form

$$Y \sim MN\left(0, \mu^T 1_P + 1_N v^T, \Omega^{-1}, \Theta^{-1}\right)$$

where Ω and Θ are row and column precision matrices respectively. An EM algorithm fits maximum penalized likelihood estimates, using L_2 penalties on both Ω and Θ, and computes expected values for missing entries. TRCMA is a one-step approximation to this EM

algorithm and was proposed as a computationally tractable alternative[54]. TRCMA is much slower than all other methods we tried in this paper, especially for large N. To speed it up, we performed preliminary simulations to determine a small but useful set of regularization parameters to optimize over (5 levels for both the row and column penalties). This method was also run on fewer simulated datasets than the other methods when constructing Figure 2 due to computational reasons. We used the `R` code from the TRCMA website (see URLs) to apply this method.

SOFTIMPUTE—there is a large machine learning literature on matrix completion methods [57,58]. We picked a competitive approach [51] which estimates a low-rank approximation to the full matrix of phenotypes via a penalty on the sum of the singular values (or nuclear norm) of the approximation. If H is the set of indices of non-missing values in Y then the method seeks an estimate, X, to the full matrix, Y, that minimizes

$$\sum_{i,j \in H} (X_{ij} - Y_{ij})^2 + \lambda \|X\|_*$$

where $\|X\|_*$ is the nuclear norm of X. We used the `R` package `softImpute` to implement this method.

MICE—this approach fits regression equations to each phenotype in an iterative algorithm (MICE) and has recently been applied to a metabolite study [18]. We used the `R` package `mice` to implement this method.

kNN—We applied a nearest neighbour imputation (kNN) approach which identifies nearest neighbour observations as a basis for prediction [52]. Specifically, if Y_{ij} is a missing phenotype then the k nearest phenotypes to phenotype j are found, based on all the non-missing values. Then Y_{ij} is predicted by a weighted average of those phenotypes in the ith individual. We used the `R` package `impute` to implement this method using the default k =10.

Simulations

We simulated data from the following model

$$Y \sim MN\left(0, K, h^2 B(\rho)\right) + MN\left(0, I, \left(1 - h^2\right) E\right)$$

where K is the $N \times N$ genetic kinship matrix and h^2 is the heritability parameter which we vary between 0 and 1. For the $P \times P$ residual covariance matrix E we simulated from a Wishart distribution $Wi\left(P, \frac{1}{P} I_P\right)$, which we then scale to a correlation matrix. For the $P \times P$ genetic covariance matrix B we used an AR(1) model with $B(\rho)_{ij} = \rho^{|i-j|}$. This model produces a range of correlations between traits and is controlled by a single parameter ρ. For Figure 1 we used $\rho = 0.45$. For Supplementary Figure 3 we used $\rho = 0.275$ and $\rho = 0.675$. For the $N \times N$ genetic kinship matrix K we used two different models : Model 1 used a subset

of the empirical kinship matrix derived from the Northern Sweden Population Health Study (NSPHS) [23]; Model 2 used a kinship structure with independent sets of 4 sibs. We set N=300 and P=15 for Figure 1 and N=1000 and P=50 for Supplementary Figure 2. Missing data was added completely at random at the 5% level (Figure 1) and 10% (Supplementary Figure 4).

Genotype and phenotype data

We analyzed 6 real datasets from 5 different organisms : humans[2,23], rats[25], yeast[24], chickens[26] and wheat[27].

The human data from the UK Blood Services Common Control, collected by the Wellcome Trust Case Control Consortium, include 1,500 individuals with 6 hematological phenotypes (hemoglobin concentration, platelet, white and red blood cell counts, and platelet and red blood cell volume)[2]. DNA samples were genotyped using the Affymetrix 500K GeneChip array. Unassayed genotypes were imputed using IMPUTE2[59] and a 1000 Genomes Project Phase 1 reference panel. We calculated a genetic relatedness matrix (GRM) using code written in R. Following others[1], phenotypes were regressed on the covariates region, age and sex. Extreme outlying measurements were removed to eliminate individuals not representative of normal variation within the population.

The human data from NSPHS[23] include 1,021 individuals with 15 glycans phenotypes (desialylated glycans (DG1-DG13), antennary fucosylated glycans (FUC-A) and core fucosylated glycans (FUC-C)). DNA samples from the NSPHS individuals were genotyped using the Illumina exome chip and either Illumina Infinium HumanHap300v2 (KA06 cohort) or Illumina Omni Express (KA09 cohort) SNP bead microarrays. Unassayed genotypes were imputed using the 1000 Genomes Phase I integrated variant set as the reference panel. Genotype data were imputed with a pre-phasing approach using IMPUTE (version 2.2.2) in the two sub cohorts (KA06 and KA09) separately. We calculated a genetic relatedness matrix (GRM) using GEMMA[3]. We used only those SNPs on either of the two Illumina chips with a minor allele frequency > 1%. Following others[4], phenotypes were regressed on the covariates age and sex and residuals were then quantile normalized. Extreme outlying measurements (those more than three times the interquartile distances away from either the 75th or the 25th percentile values) were removed.

The yeast data[24] was downloaded directly from the web (see URLs) and consisted of 1,008 prototrophic haploid segregants from a cross between a laboratory strain and a wine strain of yeast. This dataset was collected via high-coverage sequencing and consists of genotypes at 30,594 SNPs across the genome. There are 46 phenotypes in this dataset and consist of measured growth in multiple conditions, including different temperatures, pHs and carbon sources, as well as addition of metal ions and small molecules[24]. Traits were mean and variance standardized and quantile normalized before analysis. We removed SNPs with MAF < 1% or missingness in > 5% of samples and calculated a GRM using code written in R.

The wheat data[27] was downloaded directly from the web (see URLs) and consists of a winter wheat population produced by the UK National Institute of Agricultural Botany (NIAB) comprising 15,877 SNPs for 720 genotypes. Seven traits were measured: yield

(YLD), flowering time (FT), height (HT), yellow rust in the glasshouse (YR.GLASS) and in the field (YR.FIELD), Fusarium (FUS), and mildew (MIL). The population was created using a multiparent advanced generation inter-cross (MAGIC) scheme. Traits were mean and variance standardized and quantile normalized before analysis. We removed SNPs with MAF < 1% or missingness in > 5% of samples and calculated a GRM using code written in R.

The chicken dataset[26] consists of 11,575 samples across 4 full generations of an animal breeding program[26] as part of a collaboration with Aviagen. We used genotypes at 52,679 SNPs. We removed samples that were missing at > 1% of SNPs and SNPs with MAF < 1% or missingness > 5% and calculated a GRM using code written in R. There are 14 traits in this dataset ((BWT) body weight, (LFI) feed intake in females, (AFI) feed intake in males, (WTG) weight gain, (AUS) ultrasound depth, (FL) condition score, (FLMORT) floor mortality , (SLMORT) slat mortality 2, (FPD) foot-pad dermatitis, (HHP) egg production, (EFERT) early fertility , (LFERT) late fertility 2, (EHOF) early hatchability, (LHOF) late hatchability). Each trait was regressed on an appropriate set of covariates, based on experience of the ongoing breeding program. Traits were mean and variance standardized and quantile normalized before analysis.

The GWAS analysis of the rat dataset involves reconstructing the outbred rat genomes as mosaics of 8 founder haplotypes, using the program HAPPY[32]. We obtained the descent probabilities at 24,196 genomic locations based on the Rnor3.4 Rat genome assembly. For the GWAS analysis we obtained the set of pre-processed phenotypes used in the Rat Genome Sequencing and Mapping Consortium paper[25]. In total, we used 317 phenotypes to carry out phenotype imputation. The original study only carried out GWAS for 160 of these traits, deemed to be the most biological relevant traits. We re-analyzed the 140 of these 160 traits that were analyzed using mixed models in the original study. Each trait was analyzed one at a time. For this analysis we used the exact same kinship matrix used in [25]. We also assessed phenotype imputation accuracy on this dataset in Figure 2. We used exactly the 140 phenotypes and the kinship matrix from the GWAS.

When adding additional missing data to the five real datasets, we repeated this process 100 times for each level of missingness, except for the chicken dataset, which is much larger, where we used 20 simulations. The results are shown in Figure 2.

To summarize the overall levels of relatedness in each of the five datasets we calculated the following measure(Ψ), using the kinship matrix for each dataset

$$\Psi = \sqrt{\sum_{i,j} |K_{ij}|^2} / tr\,(K)$$

GWAS analysis of outbred rats

To carry out GWAS analysis of the 140 rat phenotypes we used a single-trait mixed model implemented in R. The model consisted of fixed effects that are the estimated founder descent probabilities and covariates, a single random effect with covariance as a scaled

kinship and an uncorrelated residual term. This model was fitted at each of the 24,196 genomic locations with descent probabilities. Significance was assessed using an F-test for presence or absence of the descent probabilities in the model. We carried out this analysis twice : before and after phenotype imputation.

Phenotype imputation quality metric (r^2)

We use real patterns of missing data to simulate extra missing data. We selected a rat at random and then copied its pattern of missing phenotypes to another randomly selected rat. This process continued until an extra 5% of phenotypes had been removed from the dataset. All missing phenotypes were then imputed and the squared correlation (r^2)_between the imputed values and held out values is calculated. We repeated this process 1,000 times and calculate the mean r^2.

Acknowledgments

J.M acknowledges support from the ERC (Grant no. 617306). A.D. acknowledges support from Wellcome Trust grant [099680/Z/12/Z]. This work was supported by the Wellcome Trust grant [090532/Z/09/Z]. A.K. acknowledges support from the Royal Society under the Industry Fellowship scheme.

URLs

PHENIX : https://mathgen.stats.ox.ac.uk/genetics_software/phenix/phenix.html

TRCMA : http://www.stat.rice.edu/~gallen/software.html

Yeast data : http://genomics-pubs.princeton.edu/YeastCross_BYxRM/data/cross.Rdata

Wheat data : http://www.niab.com/pages/id/402/NIAB_MAGIC_population_resources

References

1. Marx V. Human phenotyping on a population scale. Nat. Methods. 2015; 12:711–714.
2. Soranzo N, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. 2009; 41:1182–1190. [PubMed: 19820697]
3. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods. 2014; 11:407–409. [PubMed: 24531419]
4. Huffman JE, et al. Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. Hum. Mol. Genet. 2011; 20:5000–5011. [PubMed: 21908519]
5. Lauc G, et al. Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1α as a master regulator of plasma protein fucosylation. PLoS Genet. 2010; 6:e1001256. [PubMed: 21203500]
6. O'Reilly PF, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE. 2012; 7:e34861. [PubMed: 22567092]

7. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28:2540–2542. [PubMed: 22843982]

8. Schadt EE, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 2005; 37:710–717. [PubMed: 15965475]

9. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. 1998; 62:1198–1211. [PubMed: 9545414]

10. Abecasis GR, Cardon LR, Cookson WO, Sham PC, Cherny SS. Association analysis in a variance components framework. Genet. Epidemiol. 2001; 21(Suppl 1):S341–6. [PubMed: 11793695]

11. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157:1819–1829. [PubMed: 11290733]

12. Hai R, et al. Bivariate genome-wide association study suggests that the DARC gene influences lean body mass and age at menarche. Sci China Life Sci. 2012; 55:516–520. [PubMed: 22744181]

13. Piccolo SR, et al. Evaluation of genetic risk scores for lipid levels using genome-wide markers in the Framingham Heart Study. BMC Proc. 2009; 3(Suppl 7):S46. [PubMed: 20018038]

14. Choi Y-H, Chowdhury R, Swaminathan B. Prediction of hypertension based on the genetic analysis of longitudinal phenotypes: a comparison of different modeling approaches for the binary trait of hypertension. BMC Proc. 2014; 8:S78. [PubMed: 25519406]

15. Scutari M, Howell P, Balding DJ, Mackay I. Multiple quantitative trait analysis using bayesian networks. Genetics. 2014; 198:129–137. [PubMed: 25236454]

16. Park SH, Lee JY, Kim S. A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. BMC Syst Biol. 2011; 5(Suppl 2):S13. [PubMed: 22784570]

17. Cui X, Sha Q, Zhang S, Chen H-S. A combinatorial approach for detecting gene-gene interaction using multiple traits of Genetic Analysis Workshop 16 rheumatoid arthritis data. BMC Proc. 2009:S43. [PubMed: 20018035]

18. Shin S-Y, et al. An atlas of genetic influences on human blood metabolites. Nat. Genet. 2014; 46:543–550. [PubMed: 24816252]

19. Suhre K, et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature. 2011; 477:54–60. [PubMed: 21886157]

20. Meuwissen THE, Odegard J, Andersen-Ranberg I, Grindflek E. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. Genet. Sel. Evol. 2014; 46:49. [PubMed: 25158793]

21. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. Am. J. Hum. Genet. 2013; 92:744–759. [PubMed: 23643383]

22. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 2010; 42:565–569. [PubMed: 20562875]

23. Igl W, Johansson A, Gyllensten U. The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research. Rural Remote Health. 2010; 10:1363. [PubMed: 20568910]

24. Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. 2013; 494:234–237. [PubMed: 23376951]

25. Rat Genome Sequencing and Mapping Consortium. et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat. Genet. 2013; 45:767–775. [PubMed: 23708188]

26. Abdollahi-Arpanahi R, et al. Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. J. Anim. Breed. Genet. 2014; 131:183–193. [PubMed: 24460953]

27. Mackay IJ, et al. An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3 (Bethesda). 2014; 4:1603–1610. [PubMed: 25237112]

28. Ferreira MAR, Purcell SM. A multivariate test of association. Bioinformatics. 2009; 25:132–133. [PubMed: 19019849]

29. Galesloot TE, van Steen K, Kiemeney LALM, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. PLoS ONE. 2014; 9:e95923. [PubMed: 24763738]
30. Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nat. Methods. 2015; 12:755–758. [PubMed: 26076425]
31. Dahl, A.; Hore, V.; Iotchkova, V.; Marchini, J. Network inference in matrix-variate Gaussian models with non-independent noise. arXiv.org. 2013. http://arxiv.org/abs/1312.1622v1
32. Mott R, Talbot CJ, Turri MG, Collins AC, Flint J. A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. U.S.A. 2000; 97:12649–12654. [PubMed: 11050180]
33. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 2010; 11:499–511. [PubMed: 20517342]
34. Hers I. Insulin-like growth factor-1 potentiates platelet activation via the IRS/PI3Kalpha pathway. Blood. 2007; 110:4243–4252. [PubMed: 17827393]
35. Cho J, Mosher DF. Role of fibronectin assembly in platelet thrombus formation. J. Thromb. Haemost. 2006; 4:1461–1469. [PubMed: 16839338]
36. Prévost N, et al. Signaling by ephrinB1 and Eph kinases in platelets promotes Rap1 activation, platelet adhesion, and aggregation via effector pathways that do not require phosphorylation of ephrinB1. Blood. 2004; 103:1348–1355. [PubMed: 14576067]
37. Chen Y-R, et al. Y-box binding protein-1 down-regulates expression of carbamoyl phosphate synthetase-I by suppressing CCAAT enhancer-binding protein-alpha function in mice. Gastroenterology. 2009; 137:330–340. [PubMed: 19272383]
38. Shinya H, Matsuo N, Takeyama N, Tanaka T. Hyperammonemia inhibits platelet aggregation in rats. Thromb. Res. 1996; 81:195–201. [PubMed: 8822134]
39. Gilson CR, Patel SR, Zimring JC. CTLA4-Ig prevents alloantibody production and BMT rejection in response to platelet transfusions in mice. Transfusion. 2012; 52:2209–2219. [PubMed: 22321003]
40. Zufferey A, et al. Unraveling modulators of platelet reactivity in cardiovascular patients using omics strategies: Towards a network biology paradigm. Advances in Integrative Medicine. 2013; 1:25–37.
41. Szabo SJ, et al. A Novel Transcription Factor, T-bet, Directs Th1 Lineage Commitment. Cell. 2000; 100:655–669. [PubMed: 10761931]
42. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013; 9:e1003264. [PubMed: 23408905]
43. GTEx Consortium. Ardlie KG, Dermitzakis ET. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
44. Little, RJA.; Rubin, DB. Statistical analysis with missing data. John Wiley & Sons, Inc.; New York: 1987.
45. Giordano, R.; Broderick, T.; Jordan, M. Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes. arXiv.org. 2015. http://arxiv.org/abs/1506.04088v2
46. Cotsapas C, et al. Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 2011; 7:e1002254. [PubMed: 21852963]
47. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. 2013; 14:483–495. [PubMed: 23752797]
48. Listgarten J, et al. A powerful and efficient set test for genetic markers that handles confounders. Bioinformatics. 2013; 29:1526–1533. [PubMed: 23599503]
49. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 2012; 44:821–824. [PubMed: 22706312]
50. Almasy L, Dyer TD, Blangero J. Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet. Epidemiol. 1997; 14:953–958. [PubMed: 9433606]
51. Mazumder R, Hastie T, Tibshirani R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. 2010; 11:2287–2322.

52. Troyanskaya O, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17:520–525. [PubMed: 11395428]

53. Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011; 45:1–67.

54. Allen GI, Tibshirani R. Transposable regularized covariance models with an application to missing data imputation. Ann Appl Stat. 2010; 4:764–790. [PubMed: 26877823]

55. Dawid AP. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. Biometrika. 1981; 68:265–274.

56. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. Machine Learning. 1999; 37:183–233.

57. Liu, D.; Zhou, T.; Qian, H.; Xu, C.; Zhang, Z. Machine Learning and Knowledge Discovery in Databases 8189. Springer; Berlin Heidelberg: 2013. p. 210-225.

58. Wang Z, et al. Rank-One Matrix Pursuit for Matrix Completion. ICML. 2014:91–99.

59. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 2012; 44:955–959. [PubMed: 22820512]

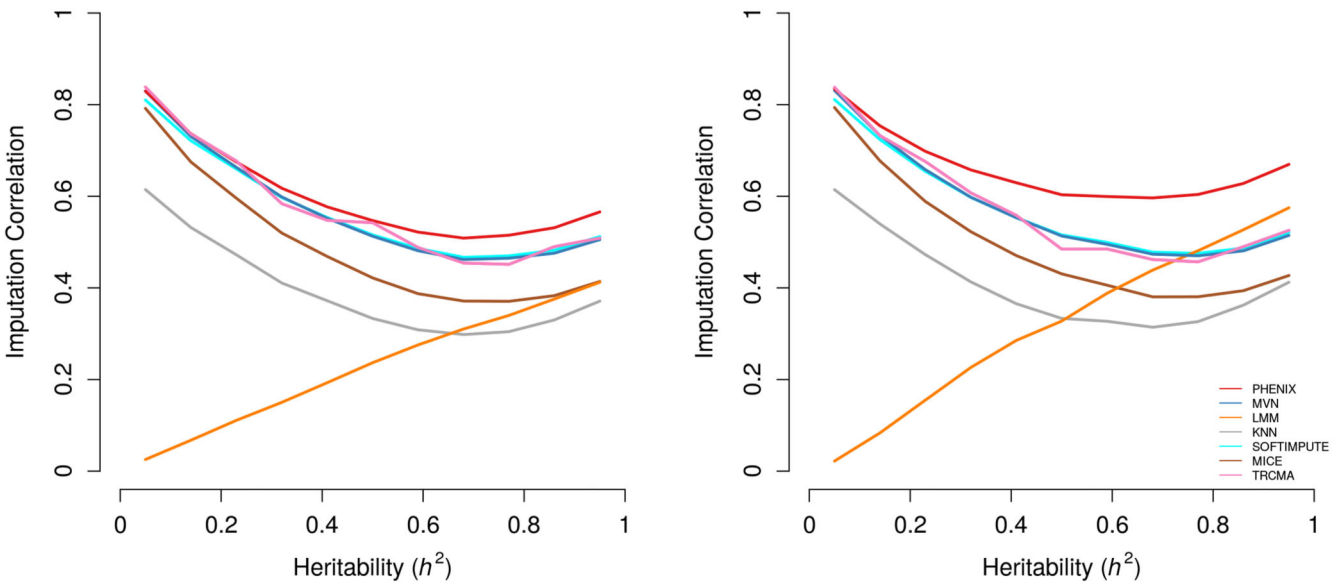

Figure 1. Simulation results

Model 1 – scenario simulated using an empirical kinship matrix derived from the human NSPHS study[23]. **Model 2** – scenario simulated using 75 families of 4 sibs. Datasets were simulated at various levels of heritability (x-axis) for the traits. 300 individuals at 15 traits were simulated. 5% of phenotype values were set as missing before imputation. 7 different methods (legend) were applied to impute the missing values. The correlation of the imputed values with the true values is plotted on the y-axis for each method. The lines for TRCMA, MVN and SOFTIMPUTE lie almost exactly on top of each other.

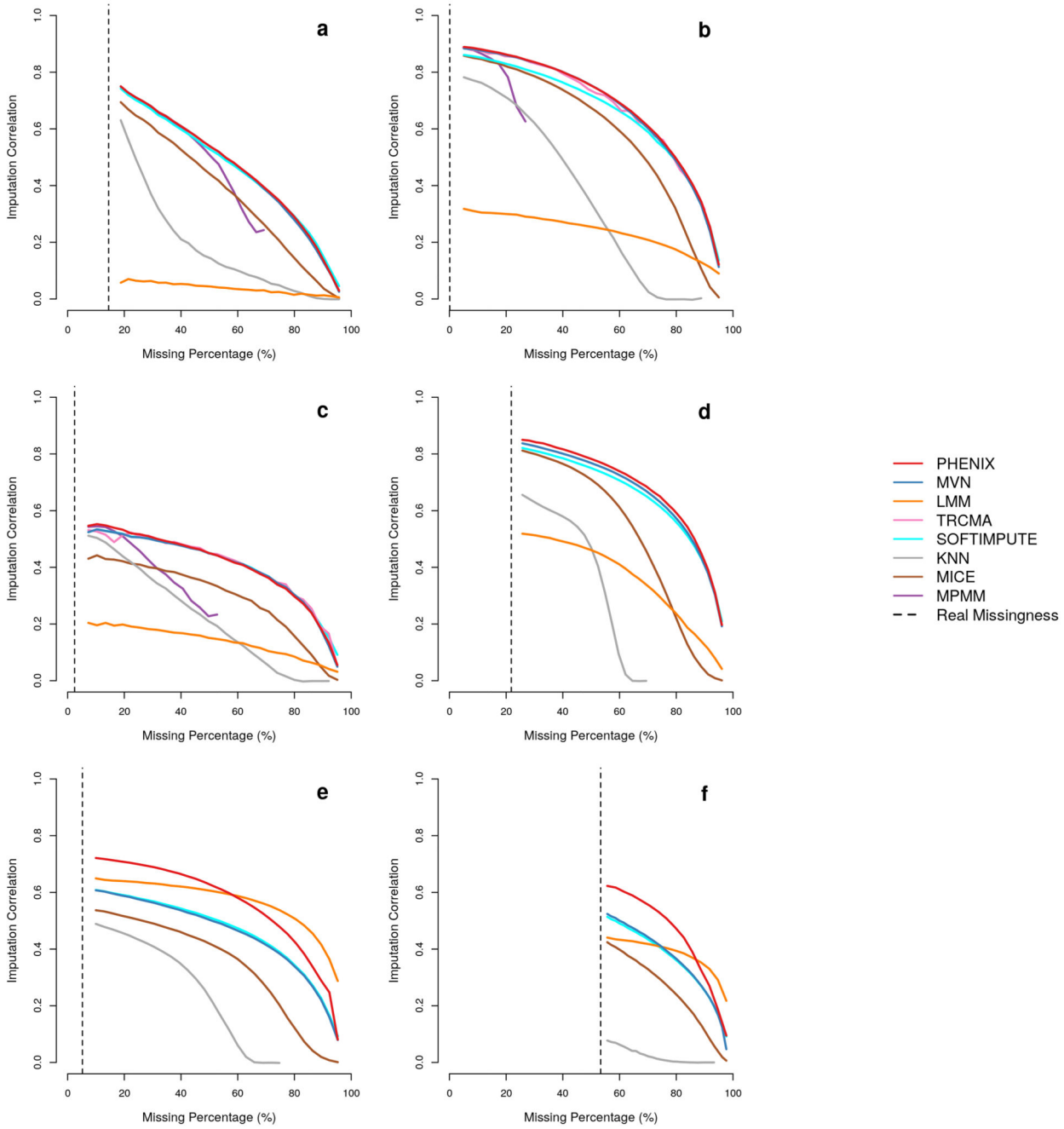

Figure 2. Imputation performance in real datasets

There is one plot for each of the six real datasets. The vertical dotted black line shows the true level of missingness in the dataset. Extra missingness was added to each dataset , and the x-axis shows the amount of missing data in these reduced datasets. The y-axis shows imputation correlation between the imputed missing data and the held out data. The legend denotes the different methods that were applied to the datasets. Not all methods were run on all datasets. TRCMA and MPMM were only run on the human NSPHS and wheat datasets for computational reasons.

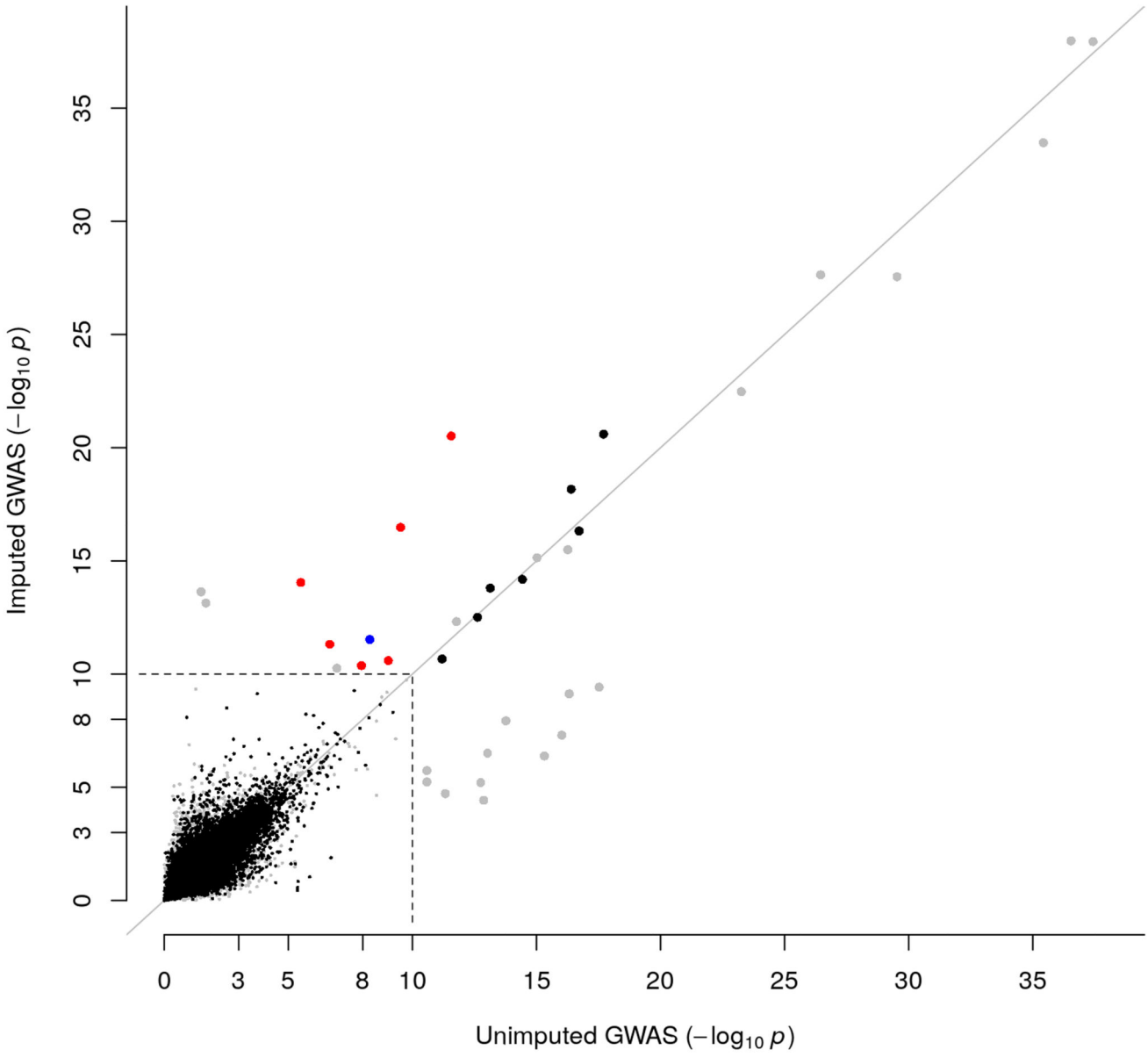

Figure 3. Missing phenotype imputation in 140 rat GWAS
The x-axis and y-axis show the −log10(p) for the GWAS on the un-imputed and imputed phenotypes respectively. Each point corresponds to a region in both scans. The dashed black lines denote a conservative threshold of −log10(p)>10 that was applied to highlight associated regions (large points). Points in grey have imputation r^2<0.36. Associations with platelet phenotypes on chr 9 and T cell phenotypes on chr 10 are highlighted with red and blue points respectively.

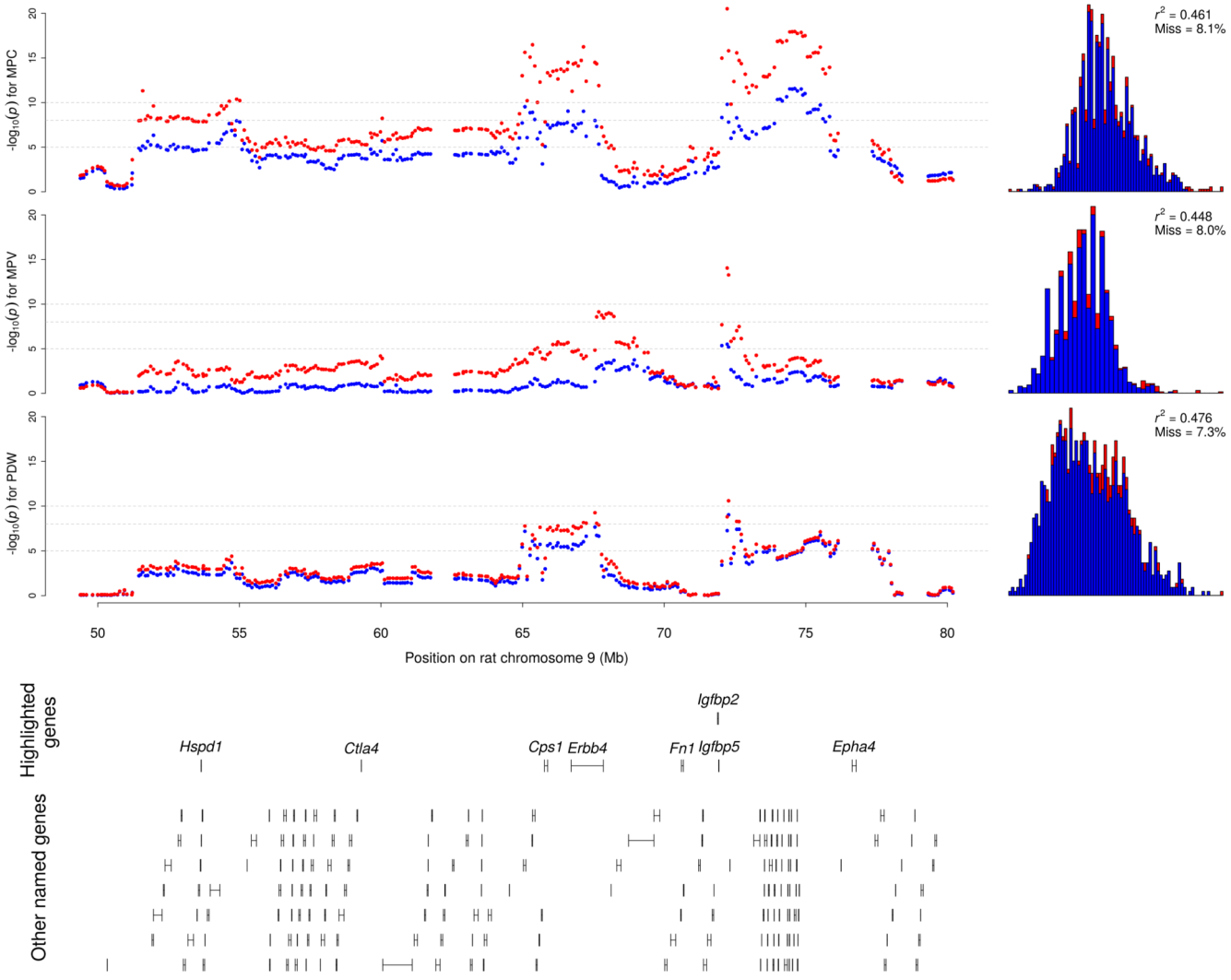

Figure 4. Platelet phenotype associations

GWAS results for un-imputed (blue points) and imputed phenotypes (red points) for three platelet phenotypes (MPC, MPV, PDW) measured in rats, on rat chromosome 9 (50-80Mb). Genes are shown below the plots, with some (named) genes with relevant annotation to platelet function, adhesion and aggregation highlighted in a separate track. Histograms on the right show the distribution of observed (cyan) and imputed (purple) phenotypes, together with missingness and r^2 metrics.

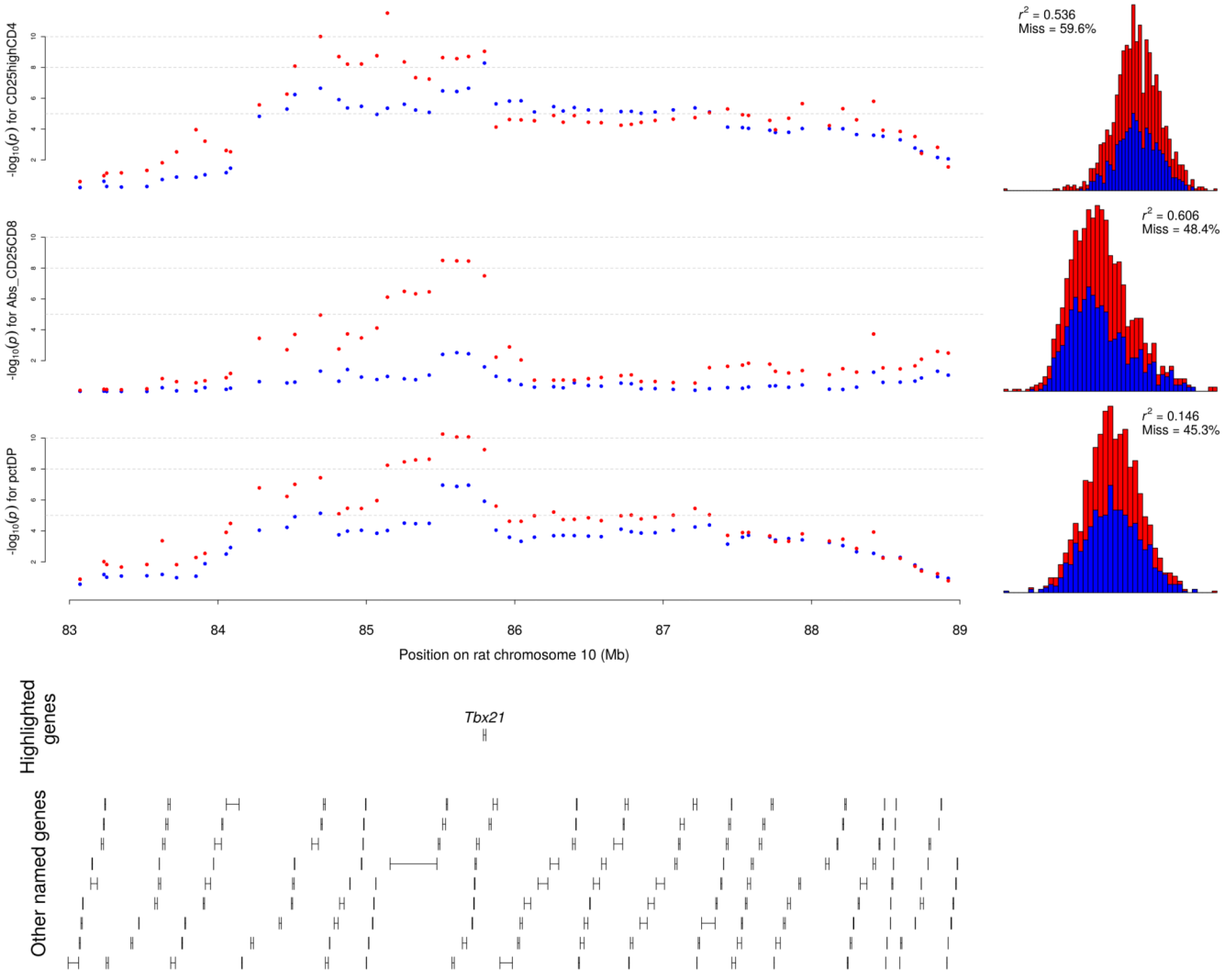

Figure 5. T cell phenotype associations

GWAS results for un-imputed (blue points) and imputed phenotypes (red points) for three T cell phenotypes (CD25highCD4, Abs_CD25CD8, pctDP) measured in rats, on rat chromosome 10 (83-89Mb). Genes are shown below the plots, with some (named) genes with relevant annotation to T cell phenotypes highlighted in a separate track. Histograms on the right show the distribution of observed (cyan) and imputed (purple) phenotypes, together with missingness and r^2 metrics.

Table 1
Brief summary of methods applied to simulated and real datasets

Method	Description and Properties	References
PHENIX	Bayesian multivariate mixed model fitted via Variational Bayes	This paper
MVN	Multivariate normal model of covariance between traits, fit using an EM algorithm. Ignores genetic covariance between samples.	44
LMM	Single trait linear mixed model, with estimated BLUP used to impute missing values. Ignores covariance between phenotypes.	48,49
MPMM	Multiple Phenotype Mixed Model, fit using EM algorithm to only samples without missing data.	3,50
SOFT-IMPUTE	Low-rank approximation to phenotype matrix via nuclear norm penalty function	51
kNN	Nearest neighbour imputation	52
MICE	Multivariate Imputation by Chained Equations	53
TRCMA	Fits a single matrix normal model to the data by estimating penalized row and column covariances	54

Table 2

Summary of real datasets analyzed

The relatedness measure (Ψ) is defined in the Online Methods.

Dataset	Number of samples	Number of phenotypes	Missing data (%)	Relatedness Measure	Reference
Rats	1,407	205	15.8	0.12	[25]
Yeast	1,008	46	5.2	0.10	[24]
Wheat	720	7	2.4	0.09	[27]
Chickens	11,575	12	57.1	0.06	[26]
NSPHS	1,021	15	0.1	0.05	[23]
UKBS	1,500	6	14.5	0.03	[2]