MultiBLUP: improved SNP-based prediction for complex traits

Doug Speed and David J Balding

Genome Res. published online June 24, 2014
Access the most recent version at doi:10.1101/gr.169375.113

MultiBLUP: improved SNP-based prediction for complex traits

Doug Speed* and David J. Balding
UCL Genetics Institute, University College London, London WC1E 6BT.
* Correspondence: doug.speed@ucl.ac.uk

Abstract

BLUP (best linear unbiased prediction) is widely used to predict complex traits in plant and animal breeding, and increasingly in human genetics. The BLUP mathematical model, which consists of a single random effect term, was adequate when kinships were measured from pedigrees. However, when genome-wide SNPs are used to measure kinships, the BLUP model implicitly assumes that all SNPs have the same effect-size distribution, which is a severe and unnecessary limitation.

We propose MultiBLUP, which extends the BLUP model to include multiple random effects, allowing greatly improved prediction when the random effects correspond to classes of SNPs with distinct effect-size variances. The SNP classes can be specified in advance, for example based on SNP functional annotations, and we also provide an adaptive procedure for determining a suitable partition of SNPs.

We apply MultiBLUP to genome-wide association data from the Wellcome Trust Case Control Consortium (seven diseases), and from much larger studies of Celiac Disease and Inflammatory Bowel Disease, finding that it consistently provides better prediction than alternative methods. Moreover, MultiBLUP is computationally very efficient; for the largest dataset, which includes 12 678 individuals and 1.5 M SNPs, the total analysis can be run on a single desktop PC in under a day, and can be parallelized to run even faster. Tools to perform MultiBLUP are freely available in our software LDAK.

Introduction

BLUP (best linear unbiased prediction) is perhaps the most widely used tool for prediction of complex traits. Developed in the 1950s as a way to predict random effects in a mixed model (Henderson 1950; Henderson et al. 1959), the advent of genomic selection has further increased its role in animal and plant breeding (Meuwissen et al. 2001; Goddard and Hayes 2007; Habier et al. 2011; Scutari et al. 2013). Recently, as SNP-based heritability analyses have demonstrated the polygenic nature of many human traits (International Schizophrenia Consortium et al. 2009; Yang et al. 2010), BLUP has also gained popularity among human geneticists, where it is beginning to replace a previous emphasis on sparsity in genome-wide analyses (Yang et al. 2011a; Makowsky et al. 2011; de los Campos et al. 2013).

Central to genetic applications of BLUP is a matrix that encodes genetic similarities between pairs of individuals. It is sometimes called a genomic relatedness matrix, although we consider genomic similarity matrix (GSM) is more appropriate. The GSM is used to specify the correlation structure of a random effect term in a mixed regression model ("mixed" means that the model includes both fixed and random effects). In the past, the only available measure of genetic similarity was a kinship coefficient computed as a probability of identity by descent in a pedigree, and so a single random effect term sufficed to model genome-wide additive effects. Nowadays genetic similarity can be measured directly, and in many different ways, from genome-wide SNP data. Yet most SNP-based applications of BLUP, referred to as Genomic BLUP or GBLUP, continue to use a single random effect, which corresponds to the unrealistic assumption that all SNP effect sizes are draws from a common Gaussian distribution. Other authors have attempted to relax the Gaussian assumption, discussed further below, but we believe that the assumption of a common prior distribution is the more important limitation, and so we focus on relaxing that.

We propose MultiBLUP, which generalizes the BLUP model to accommodate multiple random effects. Different SNP classes can be allocated separate random effects, which benefits prediction when the effect-size variance differs markedly across the classes. There are many ways to define SNP classes for which different effect size distributions may be appropriate, for example: coding, intronic, flanking and inter-genic SNPs;

MHC and non-MHC SNPs; SNPs categorized according to conservation across species; and sets of eQTL SNPs for different cell types. Alternatively, we provide Adaptive MultiBLUP which automatically identifies SNP classes with different effect sizes. This adaptive approach begins by dividing the genome into many small regions, which are then merged according to rules intended to identify a small number of genomically-contiguous regions, each with effect-size variance distinct from a baseline region comprising the rest of the genome. To ease terminology, we will refer to SNP classes as if defined by genomic regions, even though there is no need for the SNPs in a class to be contiguous and classes can overlap. MultiBLUP assigns a random effect to each region, whose correlation structure is determined by a GSM calculated from the SNPs in the region. Here, we focus on GSMs encoding additive genetic effects, but it is possible to include further random effect terms corresponding to dominance or forms of epistasis.

We first apply MultiBLUP to the seven human diseases studied by The Wellcome Trust Case Control Consortium (2007) (WTCCC1). Although relatively small datasets (each comprising approximately 5 000 individuals recorded for 280 000 SNPs), these allow us to demonstrate the advantages of MultiBLUP over a range of diseases with different genetic architectures. For rheumatoid arthritis and Type 1 diabetes, we show improvements in predictive accuracy from assigning distinct random effects to SNPs within and outside the major histocompatibility complex (MHC), because for these two traits MHC SNPs tend to have larger effects. Compared to BLUP, genetic risk scores (Wray et al. 2007), stepwise regression (Purcell et al. 2007) and Bayesian Sparse Linear Mixed Models (BSLMM; Zhou et al. 2013), we find Adaptive MultiBLUP to be the overall top performing method, regardless of whether we measure the accuracy of predicted phenotypic values by correlation, mean squared error, median absolute error or area under curve (AUC). Moreover, Adaptive MultiBLUP requires only a fraction of the computational time and memory resources of BSLMM, the second best performing method.

We next tackle larger datasets, for Celiac Disease (approximately 15 000 individuals, 200 000 genotyped SNPs) and Inflammatory Bowel Disease (13 000 individuals, 1.5 M imputed SNPs). Genetic screening of patients is routinely used to guide diagnosis and hence treatment for Celiac Disease (Husby et al. 2012; Abraham et al. 2014), and so for this trait improved prediction can have immediate impact. Due to the size of these datasets, it is not feasible to run stepwise regression or BSLMM; by contrast, Adaptive MultiBLUP requires less than 4 GB of memory, and completes in about 6 hours for Celiac Disease and 24 hours for Inflammatory Bowel Disease. Again we find Adaptive MultiBLUP to be greatly superior to both standard BLUP and genetic risk scores.

Lastly, we consider 139 phenotypes from the Wellcome Trust Heterogeneous Stock mouse collection, where individuals are highly related, as is typical in plant and animal breeding. Although there is little difference between the performance of BLUP, BSLMM and Adaptive MultiBLUP, our results demonstrate that MultiBLUP can also be used when the dataset contains high levels of structure.

The tools required to perform MultiBLUP prediction are freely available in our software LDAK.

Results

The WTCCC1 data consist of two control and seven case datasets, for Bipolar Disorder (BD), Coronary Artery Disease (CAD), Crohn's Disease (CD), Hypertension (HT), Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D), and Type 2 Diabetes (T2D). Our quality control (see Methods and Supplementary Fig. 1) removed individuals inferred to be of non-Caucasian ancestry and reduced the number of genotypes to approximately 280 000. For our simulation study we use the 2 959 control individuals and the 47 546 SNPs from Chromosomes 1 and 2. For the analysis of observed phenotypes, we combine in turn the 2 959 controls with the approximately 2 000 case individuals for each of the seven traits, and used all SNPs.

For Celiac Disease, we use the data of Dubois et al. (2010). Individuals were sourced from five cohorts, labeled according to country of origin (UK1, UK2, Finland, Netherlands and Italy). After quality control, 15 283 individuals and 190 948 SNPs remained (see Methods and Supplementary Fig. 2). For Inflammatory

Bowel Disease, we combine data from WTCCC and the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK): starting with the 1 916 Crohn's Disease cases from WTCCC1, we add 8 033 individuals from WTCCC2 (5 200 population controls and 2 833 ulcerative colitis cases) and 2 788 individuals from the NIDDK (813 cases for Crohn's Disease and 947 matched controls, and 1028 ulcerative colitis cases). As genotyping was performed using multiple SNP arrays, we first imputed using IMPUTE2 against the 1000 Genome reference panel (The 1000 Genomes Project Consortium 2010; Howie et al. 2011). After quality control, 12 678 individuals and 1 487 824 SNPs remained (see Methods and Supplementary Fig. 3).

The mouse dataset consists of 1940 heterogeneous stock mice descended from 8 founder lines (Valdar et al. 2006). After quality control there were 8 516 SNPs across 19 autosomes. The equal-tailed 95% interval for the kinship coefficients is $[-0.11,0.24]$, indicating high levels of relatedness (the corresponding interval for the WTCCC1 data is $[-0.01,0.01]$). For our simulation study we use all individuals and SNPs. For the real data analysis, of the > 150 traits available, we focus on the 139 quantitative traits which have measurements available for at least 1 300 mice and coefficient of kurtosis < 6 (the kurtosis of the Gaussian distribution is 3). The chosen traits, which all had coefficient of skewness < 1.5 (the Gaussian distribution has skewness 0), span behavioral, haematological, biochemical and disease-related phenotypes. For many of the traits, the phenotypic values are strongly correlated with cage ID. Therefore, when performing cross-validation, we ensure that individuals in the same cage remain in the same fold.

Simulation study

First, we demonstrate the potential of MultiBLUP in a simple, albeit unrealistic, setting in each of two datasets (WTCCC1 and mouse). We divide the SNPs into five distinct regions and simulate quantitative phenotypes where each region contributes a specified heritability. We consider three scenarios: (1) the five regions contribute equally to heritability; (2) regions contribute to heritability in the ratios 1:2:3:4:5; and (3) only Region 5 contributes to heritability. In each region that contributes to heritability, we assign additive genetic effects to 20 random SNPs, with effect sizes drawn from a Gaussian distribution with mean zero and variance chosen to achieve the required heritability. When applying BLUP, we used a single GSM computed as average allelic correlations across all five regions. With MultiBLUP, we used five GSMs, each calculated in the same way but using SNPs from only one of the regions. For both BLUP and MultiBLUP, we divided individuals between training and test sets in the ratio 5:1, then measured prediction performance of models fitted on the training set by the correlation between simulated and predicted phenotypes in the test set.

Figure 1 shows that for both datasets, the two methods perform similarly for Scenario 1, indicating little disadvantage to assuming the more general MultiBLUP model when it is not needed. The performance of BLUP does not improve with increasing concentration of causal variation in Scenarios 2 and 3, whereas MultiBLUP does exploit the heterogeneity of effect sizes to improve prediction, dramatically so for the WTCCC1 data. Prediction performance is good in all scenarios for the mice, because close relatedness implies that almost all causal variants are tagged, but even here the improvement of MultiBLUP is noticeable. We repeated the analysis when heritability was distributed across all SNPs, rather than only a selected 20, and observed similar results (Supplementary Fig. 4). As well as prediction performance, we also measure genomic selection performance, which is the accuracy of estimation of the sum of the random effects. This is known in animal and plant genetics as the "breeding value", and represents the phenotypic value after discounting environmental noise. MultiBLUP also provides better genomic selection performance than BLUP (Supplementary Fig. 5).

WTCCC1 data

For the real phenotypes, we evaluate prediction methods using ten-fold cross-validation. When comparing methods applied to binary outcomes, it suffices to treat case/control status as a continuous variable (cases 1, controls 0), because there exists a linear relationship between prediction performance on the observed and underlying liability scales (Dempster and Lerner 1950; Yang et al. 2011a; Zhou et al. 2013). Moreover, this permits us to minimize any effects of confounding by first regressing case/control status on sex and the first

20 principal component axes.

Columns 1-4 of Table 1 report the performance of BLUP, genetic risk scores, stepwise regression and BSLMM (see Methods for details of parameter choices). Of the first three methods, there is no clear winner: stepwise regression performs best for RA and T1D, the two traits with strongest marginal associations, while BLUP and genetic risk scores fare better for the more polygenic traits BD and HT. By contrast, BSLMM, whose model allows for both sparsity and shrinkage, performs well regardless of the genetic architecture of the trait, and overall is the best of the current methods.

As a first demonstration of MultiBLUP, we consider two regions, one corresponding to the extended MHC (Chr. 6: 25-34 Mb) and one to all other SNPs (Table 1, Column 5). This relatively simple change to the BLUP model leads to greatly improved prediction for the autoimmune traits RA and T1D (correlation 0.35 and 0.56, respectively, compared to 0.21 and 0.25 for BLUP). Supplementary Figure 6 shows Manhattan plots for each trait, from which the enhanced role of the MHC for RA and T1D can be seen. Supplementary Table 1 shows heritability estimates for the MHC and non-MHC regions, and how much each contributes to prediction. Separating SNPs according to MHC improves prediction when the MHC harbors substantial heritability; conversely high heritability can be attributed to non-MHC SNPs without this contributing much to prediction, because a single SNP effect-size variance is inadequate for a large, heterogeneous region.

The SNPs in each region need not be contiguous. To illustrate this we partition the genome into two regions according to eQTL status. For this application, we classify "eQTL SNPs" as those associated ($P < 10^{-10}$) with changes in expression levels for at least one gene, according to Curtis et al. (2012). Using this threshold, approximately 5% of SNPs are classified as eQTLs. Compared to BLUP, we achieve improved prediction for RA and T1D (see Supplementary Table 2), indicating that for these traits the eQTL SNPs tend to have a larger influence than the non-eQTL ones. Similarly, MultiBLUP regions can overlap, which we illustrate for CD by constructing regions based on two pathways (IL-2 Receptor Beta Chain in T cell Activation and IL12 Pathway) and two genes (*NOD2* and *IL23R*), all of which have shown association with the trait in at least two datasets other than the WTCCC1 (Ballard et al. 2010; Wang et al. 2009; Hugot et al. 2001; Agura et al. 2001; Duerr et al. 2006). A fifth region contains all other SNPs. Prediction is slightly improved compared to BLUP (correlation 0.319 vs 0.316; see Supplementary Table 3)

Rather than rely on prior information to define SNP regions, MultiBLUP can be run adaptively, starting with many small genomic regions which are then merged as described below and in Methods. For each of the WTCCC1 traits, we begin by dividing the genome into approximately 68 000 regions of size 75 kb (on average 8 SNPs), with a 37.5 kb overlap between neighboring regions. Although our aim is to identify regions with above-average effect-size variance, because the individuals are predominantly unrelated most effect-size variances will be very small, and therefore it suffices (and is much faster) to test instead whether each effect-size variance is non-zero. Each region with $P < 10^{-6}$ is merged with any neighboring region with $P < 10^{-2}$. At the end of this process, all remaining regions are merged into a background region. For the highly polygenic traits BD, CAD, HT and T2D, this process generates 1-2 regions (including the background region); for CD, RA and T1D, we find on average 7, 5 and 8 regions, respectively (see Supplementary Fig. 7). Overall, we find Adaptive MultiBLUP (Table 1, Column 6) to be the best-performing method; it ranks first for six of the seven traits, and is only narrowly beaten by BLUP for BD. Adaptive MultiBLUP remains top if instead we measure prediction accuracy according to mean squared error, median absolute error or AUC (Supplementary Table 4).

Celiac Disease and Inflammatory Bowel Disease

For these two traits, the sizes of the datasets make it infeasible to run stepwise regression or BSLMM, so we restrict comparison to BLUP, genetic risk scores and Adaptive MultiBLUP (starting as before with overlapping 75 kb regions). Increasing the sample size improves the resolution of Adaptive MultiBLUP; for example, for Inflammatory Bowel Disease, the method identifies on average 27 distinct local regions (Supplementary Figure 8). We again find Adaptive MultiBLUP to be the best performing method (Table 2 and Supplementary Table 5). For Celiac Disease, we additionally consider a linear prediction model constructed from 77

susceptibility SNPs: 6 SNPs tagging four human leukocyte antigen (HLA) haplotypes (Monsuur et al. 2008) and 71 SNPs based on Romanos et al. (2014). For this model, the average correlation is 0.40 and the average AUC is 0.78 (see Supplementary Table 6), demonstrating here it is better to incorporate genome-wide SNP data than use only top associated SNPs. Unlike Celiac Disease, genetic testing is not yet routinely used for Inflammatory Bowel Disease, but its potential for distinguishing subtypes has been discussed. For example, while the low prevalence of Crohn's disease makes prediction at the population level difficult, genetic data could aid in the diagnosis of patients presenting with abdominal pain, diarrhea and weight loss (Jostins and Barrett 2011), and the case for this is strengthened by the improved predictive accuracy of MultiBLUP.

Mouse data

Supplementary Figure 9 shows the performance of BLUP, genetic risk scores, BSLMM and Adaptive MultiBLUP across the 139 mouse phenotypes. Again, we start Adaptive MultiBLUP with overlapping 75 kb regions, but owing to the smaller size of the mouse genome, relax the initial significance threshold to $P < 5 \times 10^{-6}$. We find that genetic risk scores is by far the worst performing method (average correlation 0.27), because the basic single-SNP association test it uses copes poorly with the structure present in the dataset. Overall, the performances of BLUP, BSLMM and Adaptive MultiBLUP are very similar (average correlations 0.335, 0.336 and 0.336, respectively), with different methods performing best for different phenotypes. As explained below, the advantage of BSLMM and Adaptive MultiBLUP relative to BLUP comes from being able to identify individual causal loci with relatively strong influence on the phenotype; however, the high levels of relatedness and low SNP density present in the mouse data will in general make this difficult.

In this application, Adaptive MultiBLUP is slowed down due to the high levels of relatedness; despite there being less than 2000 individuals, it now takes about three hours to analyze each phenotype, approximately as long as BSLMM. This is because when deciding how to divide SNPs into regions, the shortcut used for the human data (testing whether each initial region has effect-size variance greater than zero) is no longer valid. However, this step is parallelizable and we anticipate that it can be made orders of magnitude faster by implementing algorithmic speed-ups similar to those proposed by Listgarten et al. (2012).

Comparison between Adaptive MultiBLUP and BSLMM

The prediction models used by Adaptive MultiBLUP and BSLMM have much in common. For both methods, a relatively small number of SNPs are used to capture the contributions of distinct causal loci, while the majority of SNPs influence the prediction model only through a polygenic term (in Adaptive MultiBLUP, this corresponds to the background region). The major difference is that in BSLMM each causal locus is typically represented by only one or two SNPs, whereas a local region in Adaptive MultiBLUP will generally include multiple SNPs. The former approach might be expected to perform better when a reasonably strong causal variant is well tagged by a single SNP, but even then, prediction is unlikely to suffer much by including some extra SNPs. By contrast, when the causal variant is difficult to detect through single-SNP analysis, either because it is not well tagged or has effect size too weak, or when a local region contains two or more causal variants, using multiple SNPs can provide improved prediction.

This would suggest that the accuracy of Adaptive MultiBLUP, and the potential to outperform BSLMM, will tend to increase with SNP density. Adaptive MultiBLUP was noticeably better than BSLMM for the WTCCC1 datasets, which have on average 8 SNPs per 75 kb, and we predict that had it been feasible to apply BSLMM to the Inflammatory Bowel Disease dataset (44 SNPs per 75 kb), the gap between the two methods would have been even larger. Conversely, for the mouse data the genotyping is much more sparse (most 75 kb regions contain only a single SNP), and Adaptive MultiBLUP no longer has an advantage. The difference between prediction models also explains the disparity in computational demands. The BSLMM model has one effect size for each SNP, plus additional parameters, whose values are estimated using Markov Chain Monte Carlo (MCMC). By contrast, Adaptive MultiBLUP has many fewer parameters (one variance component for each region, plus one for the environmental noise term); this allows the prediction model to be fitted deterministically, which is much faster than using MCMC and avoids issues of parameter convergence.

Similarly, the memory demands of Adaptive MultiBLUP are much lower. Both methods must store a (genome-wide) GSM, but whereas BSLMM must also read in the entire dataset, Adaptive MultiBLUP requires only those SNPs included in local regions (which is typically a small fraction of the total number of SNPs). For this reason, while it was not possible to apply BSLMM to the Celiac or Inflammatory Bowel Disease datasets, Adaptive MultiBLUP could realistically be run on even larger datasets: up to around 50 000 individuals with full genome sequence data.

Discussion

We have presented MultiBLUP, a powerful and efficient method for prediction of complex traits from genome-wide SNP data. The statistical model underlying BLUP was developed for use with kinship coefficients derived from pedigrees, but SNP data allows additional flexibility that has not previously been exploited. Specifically, the BLUP model assumes that SNP effect sizes have the same distribution for all SNPs. MultiBLUP generalizes this model by introducing multiple random effects, allowing different effect-size variances for different classes of SNPs. The SNP classes used in MultiBLUP can be identified using prior information, for example about genes and pathways relevant to the trait or other functional annotation of SNPs. Alternatively, Adaptive MultiBLUP can automatically identify SNP regions with different effect-size variances. In fact, there is no need for the correlation structure of MultiBLUP's random effects to be defined by SNPs, its prediction model can integrate multiple sources of data including copy number variants, measures of gene expression or methylation, and pedigree information.

Previous attempts to generalize the BLUP model have mainly focused on weakening the Gaussian assumption for SNP effect sizes, which has been rightly criticized because of the "thin tails" property of the Gaussian distribution. The t, double-exponential and normal-exponential-gamma distributions have been suggested as alternatives, as well as mixture distributions that allow many SNPs to have zero or negligible effect; see Zhou et al. (2013) for a review. It is not practical to compare MultiBLUP with all rival methods, so in addition to BLUP, genetic risk scores and stepwise regression, we chose BSLMM, an approach which seeks to incorporate ideas from many of the BLUP generalizations, and which has been shown to outperform a number of alternative methods (Zhou et al. 2013). The advantage of BSLMM over BLUP, genetic risk scores and stepwise regression was apparent in our analyses of the WTCCC1 data, but we found it to be inferior to Adaptive MultiBLUP for all traits, with Adaptive MultiBLUP requiring only 10% of the computation time and 5% as much memory as BSLMM. For the much larger Celiac Disease and Inflammatory Bowel Disease datasets, we showed that MultiBLUP continued to outperform the computationally-feasible alternatives, while the mouse data demonstrated that MultiBLUP can also perform well for structured datasets.

Consortia now exist for a wide variety of traits, combining data across tens of thousands of patients, while initiatives such as the 100K Genome Project (www.genomicsengland.co.uk) are set to recruit individuals in even larger numbers. At the same time, with next generation sequencing becoming more widely available, and with our ability to interrogate other sources of information (for example, transcriptomic and epigenomic) constantly improving, the number of predictors available will continue to increase. With much of the algorithm parallelizable, there is essentially no limit to the number of predictors that MultiBLUP can analyze. Instead, the runtime of MultiBLUP depends primarily on the number of individuals, as this affects how long it takes to estimate the variance components. The current implementation of Adaptive MultiBLUP can analyze 50 000 individuals, and we expect algorithmic advances to lead to increases in this number. We also envisage a meta-analysis version of MultiBLUP, where prediction models are constructed locally and then combined, allowing MultiBLUP to be used by meta-analysis consortia where data can not be shared centrally.

SNP-based prediction of phenotype is central to genomic selection, which is revolutionizing animal and plant breeding. For humans, prediction is more challenging because we are largely outbreeding which leads to low levels of relatedness in most populations. Moreover, the binary nature and low prevalence of many disease phenotypes imply that prediction of disease onset is typically not useful in a general population. However, prediction of disease state from genotype already has clinical utility in individuals selected to be of high risk on the basis of non-genetic risk factors. Moreover, where decisions about treatment options are

already based on risk factor scores, genomic information can contribute substantially to improved decision making at the population level, even when individual predictions are imprecise. As the costs of genome-wide genotyping continue to fall, additional clinical uses of genomic information for prediction of traits in humans will be found, for example to generate more realistic, individual-specific baselines from which to assess environmental impacts in population health studies. These will allow additional benefits to be obtained from the superior predictions of MultiBLUP.

Methods

The usual BLUP model assumes that $\boldsymbol{Y}$, the vector of phenotypic values for the n individuals, is influenced by random effects $\boldsymbol{g}$ (genetic) and $\boldsymbol{e}$ (environmental) via

$$\boldsymbol{Y} = \boldsymbol{g} + \boldsymbol{e} \quad \text{with} \quad \boldsymbol{g} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{K}\sigma^2) \quad \text{and} \quad \boldsymbol{e} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2), \tag{1}$$

where $\boldsymbol{K}$ is a GSM specifying the correlation structure of $\boldsymbol{g}$, $\boldsymbol{I}$ is an $n \times n$ identity matrix, while σ^2 and σ_e^2 are variances (for simplicity, fixed effects have been ignored). A common SNP-based GSM is allelic correlations averaged over SNPs (Astle and Balding 2009):

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}'/p, \tag{2}$$

where $\boldsymbol{X}$ is a matrix of (normalized) SNP genotypes, $\boldsymbol{X}'$ is its transpose and p is the number of SNPs. If (2) holds, then (1) can be expressed as a linear regression with random coefficients (Hayes et al. 2009):

$$\boldsymbol{Y} = \sum_j \boldsymbol{X}_j \beta_j + \boldsymbol{e} \quad \text{with} \quad \beta_j \sim \mathbb{N}(0, \sigma^2/p), \tag{3}$$

where $\boldsymbol{X}_j$ denotes the jth column of $\boldsymbol{X}$, and so β_j is a measure of effect size for the jth SNP.

MultiBLUP extends (1) to include random effects $\boldsymbol{g}^1, \ldots, \boldsymbol{g}^M$, with correlation structures specified by $\boldsymbol{K}^1, \ldots, \boldsymbol{K}^M$, and the corresponding variances $\sigma_1^2, \ldots, \sigma_M^2$:

$$\boldsymbol{Y} = \sum_{m=1}^{M} \boldsymbol{g}^m + \boldsymbol{e} \quad \text{with} \quad \boldsymbol{g}^m \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{K}^m\sigma_m^2) \quad \text{and} \quad \boldsymbol{e} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2). \tag{4}$$

When each $\boldsymbol{K}^m$ is of the form (2) for a matrix $\boldsymbol{X}^m$ with columns corresponding to a set of SNPs R_m of size p_m, the corresponding random regression model is

$$\boldsymbol{Y} = \sum_{m=1}^{M} \sum_{j \in R_m} \boldsymbol{X}_j^m \beta_j^m + \boldsymbol{e} \quad \text{with} \quad \beta_j^m \sim \mathbb{N}(0, \sigma_m^2/p_m). \tag{5}$$

As with BLUP, the key computational step of MultiBLUP is the estimation of the variance parameters $\sigma_1^2, \ldots, \sigma_M^2$ and σ_e^2. This can be achieved using (a generalized version of) REML (Corbeil and Searle 1976), which maximizes the log likelihood

$$-\frac{n}{2}\log(2\pi) - \frac{1}{2}\boldsymbol{Y}'\boldsymbol{V}^{-1}\boldsymbol{Y} - \frac{1}{2}\log|\boldsymbol{V}| \quad \text{where} \quad \boldsymbol{V} = \sigma_e^2\boldsymbol{I} + \sigma_1^2\boldsymbol{K}^1 + \ldots + \sigma_M^2\boldsymbol{K}^M. \tag{6}$$

MultiBLUP computes $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2$ and $\hat{\sigma_e^2}$, estimates of the variance components, using average information REML (Gilmour et al. 1995; Lee and van der Werf 2006). If the total proportion of variance explained by a kinship matrix is below 0.01% for two consecutive iterations, its contribution is set to zero; there is no limit on how many variance terms can be set to zero, allowing MultiBLUP to be run with very many regions.

MultiBLUP includes two computational optimizations to reduce memory usage and time requirements. Firstly, when a region contains fewer SNPs than the number of individuals, the corresponding GSM, $\boldsymbol{K}^m = \boldsymbol{X}^m(\boldsymbol{X}^m)'/p_m$, is computed on-the-fly, meaning that only $\boldsymbol{X}^m$, rather than $\boldsymbol{K}^m$ need be stored. Secondly, in

each iteration, the most time-consuming step is inverting $\boldsymbol{V}$; however, this process can be sped up whenever at most one GSM has full rank and the total number of SNPs contributing to the remaining GSMs is less than the number of individuals (which is generally the case for Adaptive MultiBLUP). Suppose that $\boldsymbol{K}^1$ is the GSM with full rank, then it can be decomposed as $\boldsymbol{K}^1 = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}'$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{E}$ diagonal, and therefore $\sigma_e^2\boldsymbol{I} + \sigma_1^2\boldsymbol{K}^1 = \boldsymbol{U}(\sigma_e^2\boldsymbol{I} + \sigma_1^2\boldsymbol{E})\boldsymbol{U}'$. The Woodbury Matrix Identity states that

$$(\boldsymbol{A} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}')^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{Z}(\boldsymbol{D}^{-1} + \boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{A}^{-1}. \tag{7}$$

Let $\boldsymbol{A} = \sigma_e^2\boldsymbol{I} + \sigma_1^2\boldsymbol{K}^1$, concatenate the remaining regional SNP matrices into $\boldsymbol{Z} = [\boldsymbol{X}^2\ \boldsymbol{X}^3 \ldots\ \boldsymbol{X}^M]$, and construct the diagonal matrix $\boldsymbol{D}$ with diagonal elements consisting of σ_m^2/p_m repeated p_m times, for $m = 1, \ldots, M$. Then $\boldsymbol{V}$ is in the form required to apply (7). Because $\boldsymbol{A}^{-1} = \boldsymbol{U}(\sigma_e^2\boldsymbol{I} + \sigma_1^2\boldsymbol{E})^{-1}\boldsymbol{U}'$ and $\boldsymbol{D}$ is diagonal, the only inversion required is of the lower-dimensioned matrix $(\boldsymbol{D}^{-1} + \boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z})^{-1}$. Moreover, by keeping $\boldsymbol{V}^{-1}$ in the form $\boldsymbol{U}\boldsymbol{W}\boldsymbol{U}'$, it is possible to carry out the REML iterations without computing $\boldsymbol{V}^{-1}$ explicitly, avoiding the need to multiply matrices of size $n \times n$. Additionally, this implementation avoids problems caused by local region kinship matrices being low-rank and therefore not invertible.

Predicting phenotypes. Suppose that phenotypes are recorded for individuals indexed by the set S, and we wish to predict those for individuals in the set T. In addition to estimating the variance parameters, REML also obtains $\hat{\boldsymbol{g}}_S^1, \ldots, \hat{\boldsymbol{g}}_S^M$, estimates of the genetic random effects for individuals in S. To predict phenotypes for individuals in T, we estimate $\boldsymbol{g}_T^1, \ldots, \boldsymbol{g}_T^M$ by their expected values given $\hat{\boldsymbol{g}}_S^1, \ldots, \hat{\boldsymbol{g}}_S^m$:

$$\hat{\boldsymbol{g}}_T^m = \mathbb{E}(\boldsymbol{g}_T^m | \hat{\boldsymbol{g}}_S^M) = \boldsymbol{K}_{TS}^m(\boldsymbol{K}_{SS}^m)^{-1}\hat{\boldsymbol{g}}_S^m, \tag{8}$$

where $\boldsymbol{K}_{TS}^m$ and $\boldsymbol{K}_{SS}^m$ are submatrices of $\boldsymbol{K}^m$ defined by the subscripts. We can then predict phenotypes for individuals in T via $\hat{Y}_T = \hat{\boldsymbol{g}}_T^1 + \ldots + \hat{\boldsymbol{g}}_T^M$. When the GSM takes the form (2), we have

$$\hat{\boldsymbol{g}}_T^m = \boldsymbol{X}_T^m\hat{\boldsymbol{\beta}}^m \quad \text{where} \quad \hat{\boldsymbol{\beta}}^m = \boldsymbol{X}_S^{m\,T}(\boldsymbol{X}_S^m(\boldsymbol{X}_S^m)')^{-1}\hat{\boldsymbol{g}}_S^m, \tag{9}$$

so that $\hat{\boldsymbol{\beta}}^m$ is the vector of effect sizes for SNPs in $\boldsymbol{X}^m$. When $\boldsymbol{K}^m$ corresponds to a local region, it is typically not invertible, so instead we use

$$\hat{\boldsymbol{g}}_T^m = \boldsymbol{X}_T^m\hat{\boldsymbol{\beta}}^m \quad \text{where} \quad \hat{\boldsymbol{\beta}}^m = ((\boldsymbol{X}_S^m)'\boldsymbol{X}_S^m)^{-1}(\boldsymbol{X}_S^m)'\hat{\boldsymbol{g}}_S^m. \tag{10}$$

Use of (9) and (10) is often more convenient than (8), as then phenotypes for test individuals can be predicted without needing to refer to data for training individuals.

Adaptive MultiBLUP. If regions m and m', of sizes p_m and $p_{m'}$, have equal effect-size variances (i.e., $\sigma_m^2/p_m = \sigma_{m'}^2/p_{m'}$), then Model (5) is unaffected by merging the two regions or, equivalently, replacing $\boldsymbol{g}^m$ and $\boldsymbol{g}^{m'}$ in (4) with a single random effect with correlation structure $(p\boldsymbol{K}_m + p'\boldsymbol{K}_{m'})/(p + p')$. Therefore, our adaptive strategy starts by dividing the genome into genomically-local SNP regions, then testing for each region whether its effect-size variance is significantly greater than that for all other regions combined. The formal test for Region m is performed by calculating l_0, the maximum value of (6) using a single GSM $\boldsymbol{K} = \sum_{m'} p_{m'}\boldsymbol{K}^m / \sum_{m'} p_{m'}$, and l_1, the maximum value using two GSMs: $\boldsymbol{K}^m$ and its complement $\boldsymbol{K}^{-m} = \sum_{m' \neq m} p_{m'}\boldsymbol{K}^{m'} / \sum_{m' \neq m} p_{m'}$. A p-value is obtained by comparing the test statistic $2(l_1 - l_0)$ to a $\chi^2(1)$ distribution. When levels of relatedness are low, it suffices to instead test whether the contribution to heritability from each region is significantly different from zero, using highly efficient computations similar to those outlined by Listgarten et al. (2013). The starting region size (we chose 75 kb for the human data) is intended to be small enough to separate distinct causal loci, but in case a causal locus spans multiple 75 kb regions, we then merge adjacent significant regions as described above. To test sensitivity for the WTCCC1 data, we additionally ran Adaptive MultiBLUP starting with 37.5 and 150 kb regions, or using significance thresholds $P < 10^{-5}$ and $P < 10^{-7}$ (instead of the Bonferroni-derived $P < 10^{-6}$); in all cases we observed little difference in prediction performance, and Adaptive MultiBLUP remained the best performing method (see Supplementary Table 7).

Model (4) is the same as that used by genome partitioning, a method for estimating the variance explained by subsets of SNPs (Yang et al. 2011b). The advantage of MultiBLUP over BLUP arises when

relatively large fractions of phenotypic variance can be assigned to relatively small SNP classes. However, it is important to bear in mind that the focus of MultiBLUP is to obtain the prediction model $\boldsymbol{Y} = \sum_{m=1}^{M} \boldsymbol{g}^m$ (or equivalently $\boldsymbol{Y} = \sum_{m=1}^{M} \sum_{j \in R_m} \boldsymbol{X}_j^m \beta_j^m$); the estimates $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2$ will only accurately reflect variance explained when individuals are distantly related and when the decision of how to construct each $\boldsymbol{K}^m$ was made *a priori*, rather than adaptively based on the data.

Data quality control. All analyses used only autosomal SNPs. For the WTCCC1 data (www.wtccc.org.uk), we filtered to remove population outliers identified through principal component analysis (Supplementary Fig. 1), after which 2 959 controls remained, while each of the case/control studies were left with between 4 859 and 4 928 individuals. Then we removed SNPs with either minor allele fraction (MAF) < 0.01 or call-rate (CR) < 0.995, or $P < 0.05$ from either a test for Hardy-Weinberg equilibrium (HWE) or differential missingness between cases and controls, after which studies contained between 270 319 and 284 913 SNPs. Even subtle differences in population between cases and controls can lead to artificial gains in prediction. To guard against this, we first regressed disease status on sex plus the top 20 principal component axes, then used the (continuous-valued) residuals for subsequent analyses. A potential drawback of this approach is that any true causal signal contained within the top axes is discarded and so unable to contribute towards prediction; however, as population stratification is likely to benefit most methods whose prediction models contain very many SNPs, we thought it better to err on the conservative side. By way of comparison, we instead regressed disease status on sex and two ancestry axes derived from the HapMap reference panel (The International HapMap Consortium 2003) observing slightly higher prediction performance for the WTCCC1 traits (see Supplementary Table 7), suggesting that the true prediction potential lies somewhere in between these two sets of values.

For the Celiac Disease data, the initial quality control steps are described in Dubois et al. (2010). Principal component analysis indicated that the dataset was sufficiently homogeneous (Supplementary Fig. 2), so we retained all 15 283 individuals, but removed SNPs with MAF < 0.01, CR < 0.995 or HWE $P < 0.05$, after which 190,948 remained. For Inflammatory Bowel Disease, the data came from five cohorts: 1 916 Crohn's Disease cases from WTCCC1, 5 200 controls and 2 833 ulcerative colitis cases from WTCCC2, 813 Crohn's Disease cases and 947 matched controls, and 1 028 ulcerative colitis cases from NIDDK (www.niddk.nih.gov). Separately for each cohort, we first removed outlying samples based on principal component analysis (Supplementary Fig. 3), and SNPs with MAF < 0.01, CR < 0.95 or HWE $P < 10^{-6}$, then imputed against the 1000 Genome reference panel using IMPUTE2 (The 1000 Genomes Project Consortium 2010; Howie et al. 2011). Then we combined samples, filtering out SNPs with (expected) MAF < 0.01, (expected) CR < 0.995, or IMPUTE2 Info Score < 0.98, and finally excluding 213 individuals who appeared to be duplicates (estimated kinship > 0.7 with another individual in the dataset). 12 678 individuals and 1 487 824 SNPs remained.

For the mouse data (downloaded from www.mus.well.ox.ac.uk), no individuals were excluded, but SNPs were removed if they had MAF < 0.01, HWE $P < 10^{-4}$ or call-rate < 0.99. Each of the supplied phenotypes had been pre-adjusted for marginally significant covariates, such as age, sex and body weight, and when performing ten-fold cross-validation, we ensured that mice in the same cage were kept in the same fold.

Genetic risk scores. We constructed a linear predictor $\sum_j \beta_j \boldsymbol{X}_j$ using all SNPs achieving a p-values from marginal association analysis below a specified threshold, with effect sizes estimated from the same analysis. We considered five threshold values (1 to 5 on the $-\log_{10}$ scale), expecting higher p-value thresholds to provide better prediction for more highly polygenic traits, and vice versa. Prediction from genetic risk scores can be impaired due to high levels of linkage disequilibrium, so for the human data we repeated the analysis having first pruned to obtain a subset of SNPs in approximate linkage equilibrium; results were noticeably different only for the Inflammatory Bowel Disease dataset, which uses imputed genotypes, so for this trait we instead report results from the pruned analysis. Because genetic risk scores estimates SNP effect sizes independently, the method is not expected to perform well when judged according to mean squared error or median absolute error, and this proved to be the case for all traits.

Stepwise regression. We performed multiple runs of single-SNP association analysis, each time conditioning on the SNPs already selected and adding the most strongly associated SNP to the model. We stopped

when no SNP was (conditionally) significant at $P < 10^{-6}$, then estimated coefficients for the selected SNPs using least squares. For BD, CAD, HT and T2D, the average model size was between 0 and 2 SNPs, while for CD, RA, and T1D, the average model size was 7, 6 and 17 SNPs, respectively.

BSLMM. We ran BSLMM with the parameters at their default values, meaning that the first 100 000 MCMC iterations were discarded, then posterior estimates were obtained from the next 1 000 000. For both the human and mouse data, we used a standardized kinship matrix (option `-gk 2`), matching the way GSMs were computed for BLUP and MultiBLUP.

Computing resources. The first step in MultiBLUP is to compute one or more GSMs. When these represent allelic correlations, the time required scales approximately linearly in the total number of SNPs and quadratically in the number of individuals; for example, with optimized code (Gray et al. 2012), this step took about 15 hours for the Inflammatory Bowel Disease data (c. 13 000 individuals, 1.5 M SNPs), and is readily parallelized. For Adaptive MultiBLUP, each initial region must be tested, which when individuals are predominantly unrelated is very fast (about 1 hour for the Inflammatory Bowel Disease data), but slower when individuals are highly related (about 3 hours for the mouse data), however, this step can also be parallelized. The time to estimate the variance terms scales approximately quadratically in the number of individuals, taking about 5 hours for the Inflammatory Bowel Disease data. The memory required by MultiBLUP scales quadratically with the number of individuals and, if each GSM has full rank, linearly with the number of random effects (e.g., with two full-rank GSMs, MultiBLUP requires about twice the memory of BLUP); however, GSMs corresponding to small subsets of SNPs can be computed on-the-fly, meaning that Adaptive MultiBLUP typically requires only slightly more memory than BLUP (about 4 Gb for the Inflammatory Bowel Disease data).

Software Availability

The tools required to apply MultiBLUP are freely available in our software LDAK (www.ldak.org).

Acknowledgments

We thank David van Heel of Queen Mary University of London for providing the Celiac Disease data and Sang Lee of the Queensland Institute of Medical Research for helpful advice regarding average information REML. Analyses were performed with the use of the UCL Legion High Performance Computing Facility (Legion@UCL), and with the help of the associated support services. Access to Wellcome Trust Case Control Consortium data was authorized as work related to the project "Genome wide association study of susceptibility and clinical phenotypes in epilepsy," while access to data from the National Institute of Diabetes and Digestive and Kidney Disease was granted under Project 5938, "Using genome-wide SNP data to predict disease behavior for Crohn's disease." This work is funded by the UK Medical Research Council under grant G0901388, with support from the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

Disclosure

The authors declare no conflicts of interest.

Figure Legend

Tables

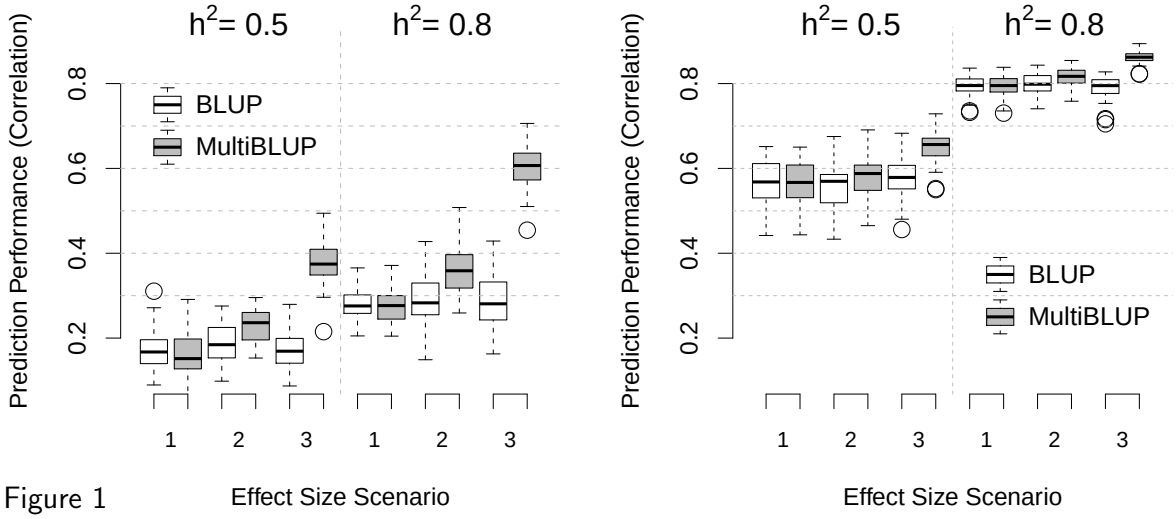


Figure 1

Figure 1. Prediction performance of BLUP and MultiBLUP on simulated quantitative traits. The two plots correspond to unrelated humans (left) and related mice (right). They show across 50 repetitions the correlation between predicted and observed phenotypes in the test set, for BLUP (white boxes) and MultiBLUP (shaded boxes). The x-axis indexes the simulation scenarios, with increasing heterogeneity of effect sizes across the five regions. Here MultiBLUP uses five GSMs, one for each region. Within each plot, the true (simulated) heritability is 0.5 (left half) or 0.8 (right half).

Table 1. Prediction of case/control status for WTCCC1 human traits

	Current methods				MultiBLUP	
Trait	BLUP	Risk Score ($-\log_{10}(P)$)	Stepwise Regression	BSLMM	Two-region MHC/non-MHC	Adaptive
Bipolar Disorder	**0.27**	0.25 (1)	0.02	0.27	0.27	0.27
Coronary Artery Disease	0.13	0.12 (1)	0.08	0.15	0.13	**0.16**
Crohn's Disease	0.32	0.28 (1)	0.18	0.34	0.29	**0.36**
Hypertension	0.15	0.14 (1)	0.00	0.14	0.14	**0.17**
Rheumatoid Arthritis	0.21	0.28 (3)	0.32	0.33	0.35	**0.37**
Type 1 Diabetes	0.25	0.34 (5)	0.54	0.57	0.56	**0.59**
Type 2 Diabetes	0.16	0.14 (1)	0.10	0.17	0.16	**0.18**
Average across 7 traits	0.21	0.22	0.18	0.28	0.27	**0.30**

Current methods BLUP, genetic risk scores, stepwise regression and BSLMM (Bayesian Sparse Linear Mixed Models) are compared with MultiBLUP (regions defined according to MHC/non-MHC) and Adaptive MultiBLUP (starting with 75 kb regions). Values report correlation between observed and predicted phenotypes based on ten-fold cross-validation. For the genetic risk scores, we consider five p-value thresholds (1 to 5 on the $-\log_{10}$ scale), and report the best prediction across these (and the corresponding threshold in brackets). The largest correlation observed for each trait is marked in **bold**.

Table 2. Prediction of case/control status for Celiac Disease and Inflammatory Bowel Disease

	BLUP		Risk Score ($-\log_{10}(P)$)		Adaptive MultiBLUP	
Trait (Number of Samples)	r	AUC	r	AUC	r	AUC
Celiac Disease, All Samples (15283)	0.46	0.79	0.45 (1)	0.79 (1)	**0.57**	**0.86**
Celiac Disease, UK Cohorts Only (10118)	0.42	0.78	0.44 (1)	0.79 (1)	**0.55**	**0.86**
Celiac Disease, UK2 $\rightarrow$ UK1 (6785 $\rightarrow$ 3333)	0.41	0.78	0.44 (1)	0.80 (1)	**0.53**	**0.86**
Inflammatory Bowel Disease (12678)	0.15	0.58	0.21 (1)	0.61 (1)	**0.33**	**0.68**
Crohn's Disease (8826)	0.16	0.60	0.25 (4)	0.67 (4)	**0.29**	**0.68**
Ulcerative Colitis (9978)	0.15	0.59	0.17 (3)	0.61 (3)	**0.27**	**0.66**

Average correlation (r) and area under curve (AUC) between observed and predicted phenotypes for BLUP, genetic risk scores, and Adaptive MultiBLUP (starting with 75 kb regions). For Celiac Disease, we consider all samples and separately UK samples only; for Inflammatory Bowel Disease, we consider all samples, only Crohn's Disease cases and only ulcerative colitis cases. The results are based on ten-fold cross-validation, except that for Celiac Disease, we also perform out-of-sample prediction from one UK cohort into the other. The best performing method for each measure is marked in **bold**.

References

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**:1061–1073.

Abraham, G., Tye-Din, J., Bhalala, O., Kowalczyk, A., Zobel, J., and Inouye, M., 2014. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet*, **10**:e1004137.

Agura, Y., Bonen, D., Inohara, N., Nicolae, D., Chen, F., Ramos, R., Britton, H., Moran, T., Karaliuskasn, R., Duerr, R., *et al.*, 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**:603–606.

Astle, W. and Balding, D., 2009. Population structure and cryptic relatedness in genetic association studies. *Statist Sci*, **24**:451–471.

Ballard, D., Abraham, C., Cho, J., and Zhao, H., 2010. Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Med Genom*, **3**:25.

Corbeil, R. and Searle, S., 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**:31–38.

Curtis, C., Shah, S., Chin, S., Turashvili, G., Rueda, O., Dunning, M., Speed, D., Lynch, A., Samarajiwa, S., Yuan, H., *et al.*, 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**:346–352.

de los Campos, G., Hickey, J., Pong-Wong, R., and Daetwyler, H., 2013. Whole genome regression and prediction methods applied to plan and animal breeding. *Genetics*, **193**:327–345.

Dempster, E. and Lerner, I., 1950. Heritability of threshold characters. *Genetics*, **35**:212–236.

Dubois, P., Trynka, G., Franke, L., Hunt, K., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G., Ádány, R., and Aromaa, A., *et al.*, 2010. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet*, **42**:295–302.

Duerr, R., Taylor, K., Brant, S., Rioux, J., Silverberg, M., Daly, M., Steinhart, A., Abraham, C., Regueiro, M., Griffiths, A., *et al.*, 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**:1461–1463.

Gilmour, A., Thompson, R., and Cullis, B., 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**:1440–1450.

Goddard, M. and Hayes, B., 2007. Estimation of genetic parameters. *J. Anim Breed Genet*, **124**:323–330.

Gray, A., Stewart, I., and Tenesa, A., 2012. Advanced complex trait analysis. *Bioinformatics*, **28**:3134–3136.

Habier, D., Fernando, R., Kizilkaya, K., and Garrick, D., 2011. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **186**:186–197.

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M., 2009. Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci*, **92**:433–443–.

Henderson, C., 1950. Estimation of genetic parameters. *Ann Math Stat*, **21**:309–310.

Henderson, C., Kempthorne, O., Searle, S., and von Krosigk, C., 1959. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**:192–218.

Howie, B., Marchini, J., and Stephens, M., 2011. Genotype imputation with thousands of genomes. *G3*, **1**:457–470.

Hugot, J., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J., Belaiche, J., Almer, S., Tysk, C., O'Morain, C., and Gassull, M., *et al.*, 2001. Asssociation of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**:599–603.

Husby, S., Koletzko, S., Korponay-Szabó, I., Mearing, M., Phillips, A., Shamir, R., Troncone, R., Giersiepen, K., Branski, D., Catassi, C., *et al.*, 2012. European society for pediatric gastroenterology, hepatology, and nutrition guidelines for the diagnosis of coeliac disease. *Pediatr Gastr Nurtr*, **54**:136–160.

The International HapMap Consortium, 2003. The International HapMap Project. *Nature*, **426**:789–796.

International Schizophrenia Consortium, Purcell, S., Wray, N., Stone, J., Visscher, P., O'Donovan, M., Sullivan, P., and Sklar, P., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**:748–752.

Jostins, L. and Barrett, J., 2011. Genetic risk prediction in complex disease. *Hum Mol Genet*, **20**:R182–R188.

Lee, S. and van der Werf, J., 2006. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet Sel Evol*, **38**:25–43.

Listgarten, J., Lippert, C., Kadie, C., Davidson, R., Eskin, E., and Heckerman, D., 2012. Improved linear mixed models for genome-wide association studies. *Nat. Methods*, **9**:525–526.

Listgarten, J., Lippert, C., Kang, E., Xiang, J., Kadie, C., and Heckerman, D., 2013. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, **29**:1526–1533.

Makowsky, R., Pajewski, N., Klimentidis, Y., Vazquez, A., Duarte, C., Allison, D., and de losz Campos, G., 2011. Beyond missing heritability: prediction of complex traits. *PLoS Genet*, **7**:e1002051.

Meuwissen, T., Hayes, B., and Goddard, M., 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**:1819–1829.

Monsuur, A., de Bakker, P., Zhernakova, A., Pinto, D., Verduijn, W., Romanos, J., Auricchio, R., Lopez, A., van Heel, D., Crusius, J., *et al.*, 2008. Effective detection of human leukocyte antigen risk alleles in coeliac disease using tag single nucleotide polymorphisms. *PLoS ONE*, **3**:e2270.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., *et al.*, 2007. Plink: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, **81**:559–575.

Romanos, J., Rosén, A., Kumar, V., Trynka, G., Franke, L., Szperl, A., Gutierrez-Achury, J., van Diemen, C., Kanninga, R., Jankipersadsing, S., *et al.*, 2014. Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. *Gut*, **63**:415–422.

Scutari, M., Mackay, I., and Balding, D., 2013. Improving the efficiency of genomic selection. *Stat Appl Genet Mol*, **12**:517–527.

Valdar, W., Solberg, L., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W., Taylor, M., Rawlins, J., Mott, R., and Flint, J., *et al.*, 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet*, **38**:879–887.

Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J., Russell, R., Sleiman, P., Imielinski, M., Glessner, J., Hou, C., *et al.*, 2009. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. *Am J Hum Genet*, **84**:399–405.

The Wellcome Trust Case Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**:661–678.

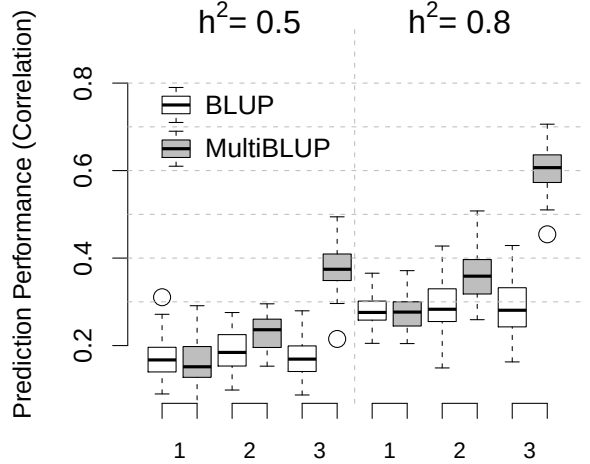
Wray, N., Goddard, M., and Visscher, P., 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*, **17**:1520–1528.

Yang, J., Benjamin, B., McEvoy, B., Gordon, S., Henders, A., Nyholt, D., Madden, P., Heath, A., Martin, N., Montgomery, G., *et al.*, 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, **42**:565–569.

Yang, J., Lee, S., Goddard, M., and Visscher, P., 2011a. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, **88**:76–82.

Yang, J., Manolio, T., Pasquale, L., Boerwinkle, E., Caporaso, N., Cunningham, J., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M., *et al.*, 2011b. Genomic partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*, **43**:519–525.

Zhou, X., Carbonetto, P., and Stephens, M., 2013. Polygeneic modeling with Bayesian sparse linear mixed models. *PLoS Genet*, **9**:e1003264.


h2= 0.5
h2= 0.8
BLUP
MultiBLUP
Prediction Performance (Correlation)
0.2
0.4
0.6
0.8
1
2
3
1
2
3
Effect Size Scenario

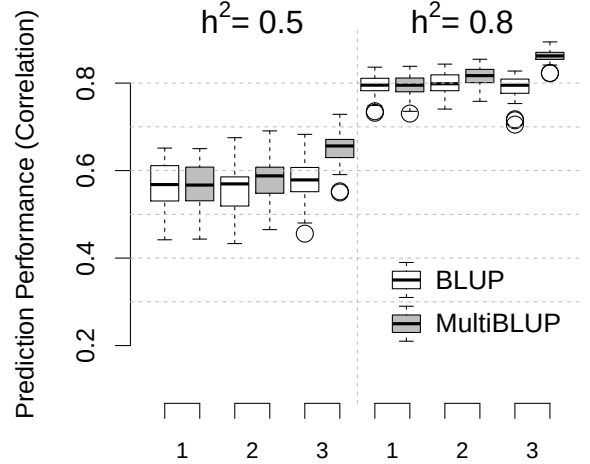

h2= 0.5
h2= 0.8
BLUP
MultiBLUP
Prediction Performance (Correlation)
0.2
0.4
0.6
0.8
1
2
3
1
2
3
Effect Size Scenario

Figure 1