Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies

Hugues Aschard,[1,*] Bjarni J. Vilhjálmsson,[1,2] Nicolas Greliche,[3,4,5] Pierre-Emmanuel Morange,[6] David-Alexandre Trégouët,[3,4,5] and Peter Kraft[1]

Many human traits are highly correlated. This correlation can be leveraged to improve the power of genetic association tests to identify markers associated with one or more of the traits. Principal component analysis (PCA) is a useful tool that has been widely used for the multivariate analysis of correlated variables. PCA is usually applied as a dimension reduction method: the few top principal components (PCs) explaining most of total trait variance are tested for association with a predictor of interest, and the remaining components are not analyzed. In this study we review the theoretical basis of PCA and describe the behavior of PCA when testing for association between a SNP and correlated traits. We then use simulation to compare the power of various PCA-based strategies when analyzing up to 100 correlated traits. We show that contrary to widespread practice, testing only the top PCs often has low power, whereas combining signal across all PCs can have greater power. This power gain is primarily due to increased power to detect genetic variants with opposite effects on positively correlated traits and variants that are exclusively associated with a single trait. Relative to other methods, the combined-PC approach has close to optimal power in all scenarios considered while offering more flexibility and more robustness to potential confounders. Finally, we apply the proposed PCA strategy to the genome-wide association study of five correlated coagulation traits where we identify two candidate SNPs that were not found by the standard approach.

Introduction

The genetic component of common, complex diseases such as asthma or type 2 diabetes is often studied via multiple related endo-phenotypes. The identification of genetic variants that influence these correlated traits may hold the key to understanding the genetic architecture of the disease in question. Although many studies analyze each of these phenotypes separately, the joint analysis of multivariate phenotypes has recently become popular because it can increase statistical power to detect genetic loci.[1–4] However, integrating association signals at a single SNP over multiple correlated dependent variables in a single comprehensive framework is not always straightforward. Simple approaches, such as Fisher's method applied to univariate analysis of each phenotype, can inflate the type I error rate when the traits are correlated. Several advanced methods that account for the correlation between phenotypes have been proposed. Some of these methods rely on assumptions about the phenotypes or relatedness that can limit their value in practice, and some methods are computationally intensive and inapplicable to large data sets. As genotype and phenotype data sets continue to grow, both computationally efficiency and robustness will only become more important.

Currently, three different strategies are commonly used for detecting genetic associations in correlated phenotypes:[3] regression models, p value correction of univariate analysis, and data reduction methods. Regression models include mixed effects models that model the covariance structure caused by correlated phenotypes as well as population structure.[1,5] For p value correction methods, univariate association tests are first performed for each phenotype individually and then combined in a meta-analysis while accounting for the observed correlational structure between the phenotypes.[6–8] Finally, data reduction methods consist of identifying the linear combination of a set of variables that is the most highly correlated with any linear combination of a second set of variables. Two common data reduction approaches in genetic epidemiology are canonical correlation analysis[9] (which is equivalent to a one-way MANOVA when analyzing a single SNP) and principal component analysis (PCA), where principal components (PCs) are built to maximize either the phenotypic variance or heritability.[10]

In this study we review the theoretical basis for standard PCA (that maximize the phenotypic variance) and evaluate the performance of different PCA-based strategies that have been commonly applied in genetic epidemiology for linkage analysis and genome-wide association studies (GWASs).[11–18] Following the principle of dimension reduction, most studies test for associations between individual SNPs and the first few PCs that explain most of the total phenotypic variance. Downstream from the univariate analysis of the top PCs, some studies also conducted a multivariate analysis of these components.[12,13] Although

[1]Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA; [2]Medical and Population Genetics Program, Broad Institute, Cambridge, MA 02142, USA; [3]Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, 75005 Paris, France; [4]INSERM, UMR_S 1166, Genomics and Physiopathology of Cardiovascular Diseases, 75013 Paris, France; [5]Institute for Cardiometabolism and Nutrition (ICAN), 75013 Paris, France; [6]Aix-Marseille Université, INSERM UMR_S 1062, 13385 Marseille, France
*Correspondence: haschard@hsph.harvard.edu
http://dx.doi.org/10.1016/j.ajhg.2014.03.016.

previous work has demonstrated the utility of PCA for multivariate GWASs, fundamental questions remain unanswered. First, there is no clear consensus on how one chooses a "low-variance" criterion for rejection of the component from the analysis. Second, it is unclear whether and how one should combine associations across PCs and how to interpret such an association. To address these questions, we compared different PCA-based strategies when analyzing a large number of simulated correlated phenotypes.

Contrary to the current prevalent belief, our results show that principal components explaining a small amount of total phenotypic variance can be as important as those explaining large amount of variance. Interestingly, this phenomenon has been suggested in a different context when analyzing nongenetic data (e.g., Jolliffe[19]). In many realistic scenarios, these small components can capture a substantial proportion of the genetic variance. Discarding these lower-variance PCs can, therefore, severely decrease power to detect genetic variants associated with one or more of the traits. In particular, we found that combining associations across all PCs, including those explaining a small amount of variance, is a particularly powerful strategy for detecting pleiotropic genetic variants and genetic variants that are associated with single trait that is highly correlated with the other traits in the study.

Based on our analysis, we propose the combined PCA strategy as a powerful, computationally efficient, and robust method suitable for most scenarios. We compared this approach to four other methods for analyzing correlated phenotypes: MANOVA, Multiphen,[20] MTMM,[1] and TATES.[8] We found that the combined PC approach showed power close to optimal in all scenarios we considered while offering more flexibility and robustness than other methods. Finally, we confirmed the usefulness of this approach by analyzing 5 coagulation-related quantitative traits in 685 subjects from the MARTHA study,[21,22] where we identified two candidate variants that would have been missed by the standard PCA approach.

Material and Methods

Analysis of Two Phenotypes

For illustration, consider a hypothetical model with two positively correlated and normally distributed phenotypes, Y_1 and Y_2 with mean 0 and variance 1, which both depend on an unknown variable U and a scaled genotype G that are also normally distributed with mean 0 and variance 1. Let c denote the correlation between Y_1 and Y_2 due to U and v_1 and v_2 denote the proportion of variance of Y_1 and Y_2 explained by G, respectively. We assume that the effects of U and G on Y_1 and Y_2 are positive and that $(c+v_{\max})$, where $v_{\max}$ is the maximum of v_1 and v_2, can vary within [0,1], so that the two trait vectors can be expressed as:

$$\mathbf{y}_1 = \sqrt{c} * \mathbf{u} + \sqrt{v_1} * \mathbf{g} + \sqrt{(1-c-v_1)} * \boldsymbol{\varepsilon}_1 \qquad \text{(Equation 1)}$$

$$\mathbf{y}_2 = \sqrt{c} * \mathbf{u} + \sqrt{v_2} * \mathbf{g} + \sqrt{(1-c-v_2)} * \boldsymbol{\varepsilon}_2, \qquad \text{(Equation 2)}$$

where $\boldsymbol{\varepsilon_1}$ and $\boldsymbol{\varepsilon_2}$ denote independent random noises that are normally distributed with mean 0 and variance 1. For this general model, the principal components of the two traits are $\mathbf{pc}_1 = 1/\sqrt{2} * (\mathbf{y}_1 + \mathbf{y}_2)$ and $\mathbf{pc}_2 = 1/\sqrt{2} * (\mathbf{y}_1 - \mathbf{y}_2)$, which can be rewritten as:

$$\mathbf{pc}_1 = \frac{1}{\sqrt{2}} * \left((2 * \sqrt{c}) * \mathbf{u} + (\sqrt{v_1} + \sqrt{v_2}) * \mathbf{g} + \sqrt{(1-c-v_1)} * \boldsymbol{\varepsilon}_1 + \sqrt{(1-c-v_2)} * \boldsymbol{\varepsilon}_2 \right) \qquad \text{(Equation 3)}$$

$$\mathbf{pc}_2 = \frac{1}{\sqrt{2}} * \left((\sqrt{v_1} - \sqrt{v_2}) * \mathbf{g} + \sqrt{(1-c-v_1)} * \boldsymbol{\varepsilon}_1 - \sqrt{(1-c-v_2)} * \boldsymbol{\varepsilon}_2 \right). \qquad \text{(Equation 4)}$$

The total phenotypic variance explained by $\mathbf{pc_1}$ and $\mathbf{pc_2}$ and noted by s_1 and s_2, respectively, can be defined as (see Appendix A)

$$s_1 = \frac{1 + c + \sqrt{v_1 v_2}}{2} \qquad \text{(Equation 5)}$$

$$s_2 = \frac{1 - c - \sqrt{v_1 v_2}}{2}. \qquad \text{(Equation 6)}$$

From this it follows that v_{pc1} and v_{pc2}, the proportion of variance of $\mathbf{pc_1}$ and $\mathbf{pc_2}$ explained by G, respectively, can be expressed as (see Appendix A)

$$v_{pc1} = \frac{v_1 + v_2 + 2 \times \sqrt{v_1 v_2}}{2 * (c + 1 + \sqrt{v_1 v_2})} \qquad \text{(Equation 7)}$$

$$v_{pc2} = \frac{v_1 + v_2 - 2 \times \sqrt{v_1 v_2}}{2(1 - c - \sqrt{v_1 v_2})}. \qquad \text{(Equation 8)}$$

The power of the association test between G and Y_1, Y_2, PC_1, and PC_2 can then be compared for different sample size and genetic effect. The Wald test for association is equal to $N \times \hat{v}$, where N is the sample size and $\hat{v}$. is the estimated proportion of variance explained by G. Because $N \times \hat{v}$. follows a noncentral chi-square distribution with one degree of freedom (df) and a noncentral parameter equal to $\delta = N \times v$., the power is[23]

$$\text{Power} = 1 - F\left(\chi^2_{1,1-\alpha,0} \,|\, 1, \delta\right) \qquad \text{(Equation 9)}$$

Here, $F(\chi^2|d,\delta)$ is the cumulative distribution function for the noncentral chi-square distribution with d degrees of freedom and noncentrality parameter δ; $\chi^2_{d,p,\delta}$ is the inverse of F, i.e., the quantiles of the noncentral chi-square distribution; and α is the type I error rate. Because the two PCs are independents, one can define a joint test of the PCs by summing the 1 df noncentral chi-square from each PC to form a 2 df noncentral chi-square. The power for such a test is equal to

$$\text{Power} = 1 - F\left(\chi^2_{2,1-\alpha,0} \,|\, 2, \delta\right). \qquad \text{(Equation 10)}$$

For simplicity, we derived the proportion of variance explained and the power to detect a genetic association for two positively correlated traits where the genetic effects on the traits were also positive (and so in the same direction). Trivially, the same result is produced when the traits are negatively correlated and the genetic effects are in opposite direction. The extension to the situation where the genetic effect is opposite to the correlation is similarly straightforward.

Analysis of Five Correlated Phenotypes

To explore the performance of different methods, we simulated traits under four different models of correlation pattern. For each of these, we simulated 10,000 sets of correlated traits measured on 5,000 subjects. For each subject we generated a SNP G with minor allele frequency of 0.3 and five phenotypes (Y_1 to Y_5). The phenotypes were generated as defined in Equation 1 where c was different for each phenotype so that the average correlation pattern between the five phenotypes matches the following correlation matrices:

$$model1 = \begin{pmatrix} 1 & 0.50 & 0.31 & 0.15 & 0.07 \\ 0.50 & 1 & 0.31 & 0.15 & 0.15 \\ 0.31 & 0.31 & 1 & 0.09 & 0.04 \\ 0.15 & 0.15 & 0.09 & 1 & 0.02 \\ 0.07 & 0.07 & 0.04 & 0.02 & 1 \end{pmatrix},$$

$$model2 = \begin{pmatrix} 1 & 0.80 & 0.63 & 0.32 & 0.09 \\ 0.80 & 1 & 0.63 & 0.32 & 0.09 \\ 0.63 & 0.63 & 1 & 0.09 & 0.07 \\ 0.32 & 0.32 & 0.09 & 1 & 0.03 \\ 0.09 & 0.09 & 0.07 & 0.03 & 1 \end{pmatrix}$$

$$model3 = \begin{pmatrix} 1 & 0.30 & 0.30 & 0.30 & 0.30 \\ 0.30 & 1 & 0.30 & 0.30 & 0.30 \\ 0.30 & 0.30 & 1 & 0.30 & 0.30 \\ 0.30 & 0.30 & 0.30 & 1 & 0.30 \\ 0.30 & 0.30 & 0.30 & 0.30 & 1 \end{pmatrix},$$

$$model4 = \begin{pmatrix} 1 & 0.70 & 0.70 & 0.70 & 0.70 \\ 0.70 & 1 & 0.70 & 0.70 & 0.70 \\ 0.70 & 0.70 & 1 & 0.70 & 0.70 \\ 0.70 & 0.70 & 0.70 & 1 & 0.70 \\ 0.70 & 0.70 & 0.70 & 0.70 & 1 \end{pmatrix}$$

The proportion of variance explained by the simulated genetic variant G was drawn from a uniform distribution with minimum 0.001 and maximum 0.005, independently of the phenotypic correlation. For the pleiotropic effects, $K = 2,..,5$ phenotypes affected by the SNP were randomly chosen (with equal probability). We also simulated situations where the pleiotropic effect of the SNP reflects the phenotypic correlation pattern; that is, assuming that the most highly correlated traits are more likely to be associated with the same genetic variants. To do so, we set the probability that the i^{th} phenotype was associated with G to be proportional to its correlation with other traits. For example, in the presence of a pleiotropic effect on two phenotypes under model 2, the two traits Y_1 and Y_2, correlated at 0.8, had much higher chances to be selected. Note that for model 3 and model 4, this does not matter because the phenotypes are equally correlated.

Combined Analysis of Venous Thrombosis-Related Phenotypes

To demonstrate the applicability of the proposed method, we analyzed five quantitative intermediate phenotypes for venous thrombosis (MIM 188050) risk measured in the MARTHA study. The study sample consists of unrelated European individuals with venous thromboembolism, consecutively recruited at the Thrombophilia center of La Timone Hospital (Marseille, France) between January 1994 and October 2005. All individuals were genotyped with the Illumina Human610-Quad and Human660W-Quad Beadchips. Five coagulation-related phenotypes were studied: the activated partial thromboplastin time (aPTT), the standardized Anticoagulant response to Agkistrodon contortrix venom (ACVn), and plasma levels of three coagulation factors, fibrinogen (FIB), factor VIII (FVIII), and von Willebrand factor (vWF). GWASs have already been conducted for these phenotypes with either raw genotypes or imputed SNPs from HapMap 2. A detailed description of the cohort and the phenotypes can be found in Oudet-Mellakh et al.,[21] Antoni et al.,[22] and Tang et al.[24] In this study, we imputed the genotyped by using the 1000 Genomes[25] 2012-02 release with the minimac software. Only SNPs with imputation quality Rsq > 0.3 and minor allele frequency (MAF) > 0.01 (n = 8,862,493) were used in this study. The imputed SNPs were tested for association with all of the phenotypes individually and the derived PCs via a linear regression where the allele dosages for the imputed SNPs were used. To achieve this, we employed the mach2qtl software.[26] These association analyses were conducted in a sample of 685 individuals with no missing phenotype information and they were adjusted for age, sex, anticoagulant therapy, smoking, and the first four principal components derived from the genome-wide genotypes.

Results

Comparison of Power for the Analysis of Two Phenotypes

Let us first consider the analysis of two correlated traits, Y_1 and Y_2 (Equations 1 and 2), where the tested SNP affects only the first trait, Y_1, and has no effect on the second, Y_2. If the genetic effect is small, such that the contribution of the SNP on the correlation between the two traits is negligible, then the two PCs (Equations 3 and 4) can be approximated by

$$\mathbf{pc}_1 \approx \frac{1}{\sqrt{2}} * \left(\left(2 * \sqrt{c}\right) * \mathbf{u} + \left(\sqrt{v_1}\right) * \mathbf{g} + \sqrt{2 * (1-c)} * \boldsymbol{\varepsilon}_1' \right) \quad \text{(Equation 11)}$$

$$\mathbf{pc}_2 \approx \frac{1}{\sqrt{2}} * \left(\left(\sqrt{v_1}\right) * \mathbf{g} + \sqrt{2 * (1-c)} * \boldsymbol{\varepsilon}_2' \right), \quad \text{(Equation 12)}$$

where $\boldsymbol{\varepsilon}'$ follow a normal distribution with mean 0 and variance 1. Under this assumption, the proportion of the total variance explained by PC_1 and PC_2 (Equations 5 and 6) become a linear function of the correlation c, i.e., $s_1 = (1+c)/2$ and $s_2 = (1-c)/2$. Hence, s_1 is approximately equal to s_2 when c is small but is much larger than s_2 when c increases.

From Equations 11 and 12 we see that PC_1 primarily depends on the effect of U, the unmeasured variable representing the shared effect. Conversely, the effect of U is, by construction in this example, not captured by PC_2, which depends only on the effect of the SNP on Y_1 plus some residual noise. This noise term scales with the phenotypic correlation, decreasing dramatically with an increasing value of c. Therefore, a test of association with the smaller principal component, PC_2, has greater power than a test with PC_1 to detect a genetic variant that affects a single phenotype when c is different from 0. Interestingly, it can also have greater power than testing for association with Y_1, even though the genetic variant does

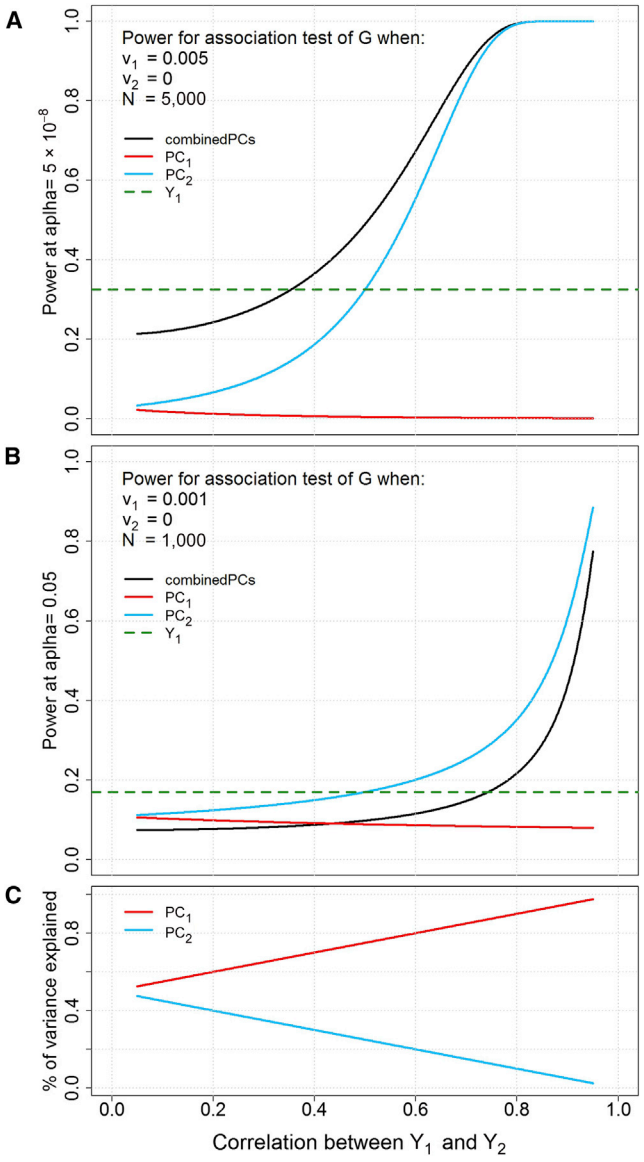

Figure 1. Power to Detect a SNP Associated with a Single Trait in a Bivariate Analysis

Power to detect the SNP associated with Y_1 based on the tests of PC_1, PC_2, the combined PCs, and Y_1 for different sample size and genetic effects (A and B), and proportion of phenotypic variance explained by PC_1 and PC_2 (C). The power of each of the four tests is presented as a function of c the correlation between Y_1 and Y_2, the sample size N, and v_1 and v_2, the proportion of the variance of Y_1 and Y_2 explained by the SNP, respectively.

not affect the second trait. This is because the ratio of genetic effect over total variance can become greater in PC_2 as compared to Y_1, thus increasing the signal-to-noise ratio. Indeed, the difference between Y_1 and Y_2 can be viewed as a form of adjustment of Y_1 for Y_2 (and conversely). When the correlation between the two traits is large, a substantial amount of noise (the effect of U) is removed from Y_1 by subtracting Y_2, while at the same time the genetic effect remains constant because the SNP is associated with Y_1 only. Figure 1 presents a comparison of the power for the test of association between G and respectively PC_1, PC_2, and the combined PCs approach, against the test of association between G and Y_1 when the SNP is associated with Y_1 but not with Y_2. In this case the combined PC test and the test of PC_2 have greater power than the test of Y_1 when the correlation is large but lower power when the correlation is low. This gain in power tends to be larger for the combined PCs as compared to PC_2 when the effect of G on the trait is large or when the sample size increases (see Figure 1A). Conversely, the test of PC_2 becomes optimal when sample size or genetic effect is small and correlation is larger than 0.5 (see Figure 1B).

When the tested SNP affects both traits, one can similarly derive the power for each of the three previous tests against the test of Y_1 and Y_2. Figure 2 presents such results for phenotypes with correlation c equal to 0.1 (Figure 2A), 0.5 (Figures 2A and 2B), and 0.9 (Figure 2C), while assuming G explains 0.5% of the variance of Y_1 and between 0% and 1% of the variance of Y_2. Furthermore, we considered the directionality of the SNP effects for the second trait: the same direction as on the first trait (positive pleiotropy) and effects in opposite directions on the two traits (negative pleiotropy). Similarly to the previous scenario, the proportion of the total variance explained by PC_1 increases with c, but conversely PC_1 displays most of the genetic effect if G has similar effects on both of the traits (i.e., is in same direction). This can be explained by considering v_{pc1} and v_{pc2}, the proportion of variance explained by G on the PCs (Equations 7 and 8), for example, when $v_1 = v_2 = v'$, v_{pc2} is null and v_{pc1} can be approximated by $v' * 2/(c+1)$. Because c is by definition smaller than 1, the test of PC_1 in this specific case will always outperform the test of each trait independently and the test of PC_2. When the genetic effects are in opposite directions, large gains in power can be achieved by testing for PC_2, whereas testing for PC_1 has almost no power. This is consistent with the results of Korte et al.,[5] where they also observed increased power to detect genetic variants with phenotype-specific effects and negative pleiotropic genetic effects when accounting for the correlation between the traits. In the latter case, the effect of G on PC_2 can be approximated by $v' * 2/(1-c)$, so that the effect of G increases exponentially as the phenotype correlation increases. Although the expected power of the combined PC approach in all these situations is usually lower than the power of single PCs when assuming a specific direction of genetic effect on the traits, it offers a good compromise, allowing for reasonable power without assuming a specific hypothesis about the genetic effect. A summary of how the different tests based on PC_1, PC_2, and the combined test of the two PCs behave for moderate sample size is shown in Table 1.

Comparison of Power for the Analysis of Five Phenotypes

When more than two phenotypes are considered, deriving the power of the various analysis strategies quickly becomes too complex to be comprehensively expressed. To

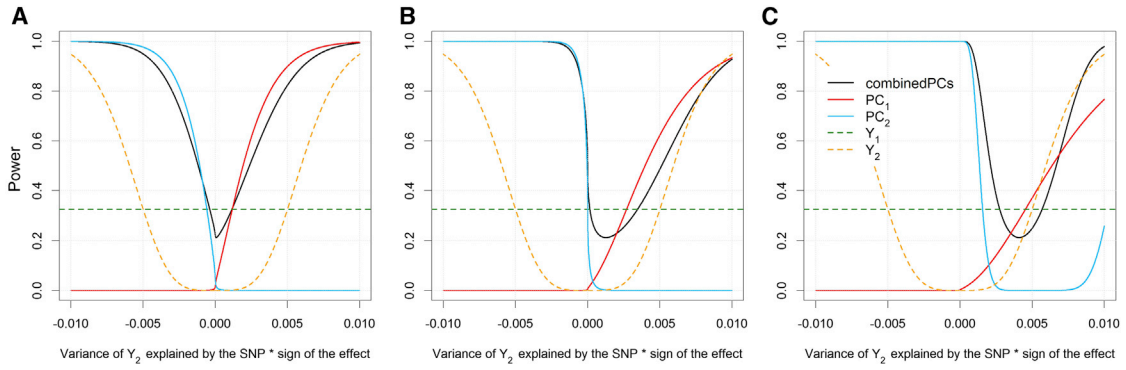

Figure 2. Power to Detect a SNP Associated with Two Traits in a Bivariate Analysis
Power at 5×10^{-8} significance level to detect the SNP associated with the *Y*. using the independent tests of Y_1, Y_2, PC_1, and PC_2 and a combined PCs test when analyzing 5,000 individuals. The genetic variant has a fixed effect on the trait Y_1. The power of each of the four tests is presented as a function of the effect on the second trait Y_2 for three levels of correlation between Y_1 and Y_2: 0.1 (A), 0.5 (B), and 0.9 (C).

compare the power of univariate and multivariate tests in such a setting, we simulated five phenotypes, Y_1 to Y_5, under four different correlation scenarios: (1) a gradient of moderate to low correlations, (2) a gradient of strong to low correlations, (3) uniform moderate correlations, and (4) uniform strong correlations (see Material and Methods). For each scenario we compared the power of four approaches: the univariate test of the original phenotypes, the univariate test of each PC, the test of the PC displaying the largest genetic effect after correcting for multiple testing, and the combined test of all PCs. We simulated and studied cases where the genetic variant is associated with a single trait (see Figure S1 available online) and cases where the genetic variant has a pleiotropic effect (see Figure 3).

No single PC analysis approach was found to have optimal power for all of the scenarios considered. The association pattern between the SNP and the five components varied greatly across the set of parameters we used. Sometimes the strongest associations were with the PCs explaining most of the variance (e.g., when the genetic variant is associated with the 5 traits and the correlation between the traits is smaller than 0.5); sometimes the strongest associations were with the PCs explaining the smallest amount of the total variance (e.g., when the genetic variant is associated with several, but not all, traits and there is correlation >0.5 among all traits). Hence, in most situations, the strategy of testing all PCs and picking the one with the strongest signal was more efficient than focusing only on the PCs explaining most of the variance. However, because most PCs showed association signals, the combined approach was on average the most powerful except when SNPs had an effect on a single trait that had correlation lower than 0.5 with the other traits analyzed, or in the presence of a fully pleiotropic effect (i.e., the SNP is associated with all 5 traits) where the correlation between traits is homogeneous.

In these simulations we considered a variety of scenarios, where we varied the correlation structure of the phenotypes while simulating independently positive pleiotropic pattern of the tested SNP. When the pleiotropic effect reflects the phenotypic correlation (see Material and Methods), the association patterns (see Figure S2) were similar to those observed without accounting for the correlation among traits (see Figure 3). Conversely, when we simulated negative pleiotropy (pleiotropy against the phenotype correlation), we observed dramatic increases in power for all PC-based approaches (see Figure S3).

When a Very Large Number of Phenotypes Are Analyzed

The simulations in the previous section show that when a relatively small number of phenotypes are analyzed jointly, the genetic association signal is usually spread across most or all of the PCs, making the analysis of a single PC or a subgroup of PCs less powerful as compared to the combined analysis of all PCs. However, when the number of phenotypes becomes very large this will not be always true, because the large increase in the degrees of freedom can outweigh the benefit in combining many small associations with individual PCs.

Consider a scenario with a very large number of correlated traits, for example, circulating levels of 100 metabolites. Under such a scenario, when analyzing 2,000 individuals and assuming that the genetic effects are of the same order of magnitude as in the previous simulations (i.e., proportion of variance explained between 0.1% and 0.5%), the power of the univariate test of the raw phenotypes at genome-wide significance level (and before any correction for the 100 tests conducted) is below 1%. Depending on the correlation pattern and the level of pleiotropy, focusing on a subgroup of PCs may increase power. This is demonstrated in Figure S4, where the power at 5×10^{-8} significance level is shown for different tests when analyzing 100 phenotypes simulated via a gradient of correlation from 0 to 0.9 (extended model 2 from the previous section, but with the more complex simulation

Table 1. Rationale for Testing Genetic Association with PCs in a Bivariate Analysis

Genetic Model	PC_1	PC_2	Combined PCs
$\beta_{Y1} \neq 0$ and $\beta_{Y2} = 0$	almost no power, converging to 0 with increase correlation	most powerful for small sample size and lowβ_{Y1}; power increases with correlation	most powerful for large sample size and large β_{Y1}; power increases with correlation
β_{Y1} in the same direction as β_{Y2}	most powerful when correlation and $\beta_{Y.}$ are moderate	very low power; power increase slightly with correlation	most powerful when correlation and $\beta_{Y.}$ are high
β_{Y2} opposite to β_{Y1}	almost no power; minor variation with increase correlation	very powerful; power increase with correlation	very powerful; power increase with correlation

The two positively traits are denoted Y_1 and Y_2 and and genetic effect of G on Y_1 (β_{Y1}) and Y_2 (β_{Y2}).

scheme SC1 described in Appendix B and Figure S5, and simulating genetic pleiotropic effects independently of the phenotypic correlation). We constructed two series of 100 tests by combining either the smallest n PCs that explain the least amount of the total phenotypic variance (i.e., the smallest eigenvalues) or the largest n PCs that explain the largest amount of the total phenotypic variance (i.e., the largest eigenvalues), with n varying from 1 to 100. When the genetic variant affected 5 traits, we observed a substantial gain in power when combining the signal from the 10 PCs corresponding to the 10 smallest eigenvalues as compared to the combined analysis of all PCs (power was 0.59 and 0.23, respectively). However, with the same simulation scheme, that same test (based on the last 10 PCs) was severely underpowered when the SNPs were associated with 20 traits (power was 0.63 and 0.99, respectively).

As our simulations demonstrate, the optimal strategy strongly depends on the underlying model. However, a naive approach that consists of analyzing the PCs jointly in a few subgroups based on their eigenvalues can significantly improve power. For example, in Figure S4 we show that combining the joint signal from the PCs with the largest eigenvalues with the joint signal from the remaining PCs (with the smallest eigenvalues) can increase power in the two scenarios described above. To achieve this, we propose Fisher's method. We defined a global multistep combined PC (mCPC) score as

$$T_K = -2 * \left[log\left(1 - F\left(\sum_{i=1}^{K} \chi_i^2 \,|\, K\right)\right) + log\left(1 - F\left(\sum_{i=K+1}^{N} \chi_i^2 \,|\, N-K\right)\right)\right]. \quad \text{(Equation 13)}$$

Here, K is the number of top PCs included in the first group and N is the total number of PCs. The χ_i^2 is the chi-square association statistic between the SNP and the i^{th} PC, and $F(\chi^2|d)$ is the cumulative distribution function of the central chi-square distribution with d degrees of freedom. Because all PCs are independent, T_K follows a chi-square distribution with 4 degree of freedom under the null hypothesis of no association with any of the PCs. Note that this test can easily be extended to more than two groups.

Comparison to Other Methods

We compared the power of the combined analysis of all PCs (CPC) approach against four existing methods: multivariate analysis of variance (MANOVA), ordinal regression of the genotypes against all phenotypes as implemented in Multiphen,[20] multiple trait mixed models (MTMM, see Appendix C),[2,5] and the p value correction method TATES,[8] which combines the p values from standard univariate regression of the phenotype. We did not consider principal component of heritability, because previous work found that such methods can be less powerful than MANOVA.[3] Under the scenario described in Figure S2, Multiphen, MANOVA, MTMM, and CPC perform similarly, with high gain in power as compared to the univariate test of each phenotype and TATES (see Figure 4). We also compared these methods by using 100 simulated phenotypes under three different generating models: an extension of the model from Figure S2 (scenario SC1) and two other models more similar to those used in the TATES paper.[8] Here, the SNPs tested were indirectly associated with the phenotypes through unmeasured latent variables driving the phenotypic correlation. We considered both a small (30) and a very large (1,000) number of latent variables (scenarios SC2 and SC3, respectively). The details of the three scenarios are described in Figure S5 and Appendix B, and Figure S6 illustrates the correlation structure and pleiotropic effects induced in each of the three simulation schemes. Because we observed inflated false positives for Multiphen in these settings (see Figure S7), we decided to exclude it from the 100 phenotype simulations. The MTMM test was included in these simulations, but we used the true phenotypic covariance structure instead of estimating it from the data (see Appendix C). In practice, estimating the phenotypic covariance structure from the data, as done in MTMM and other software,[27,28] for a large number of phenotypes (e.g., >10) is computationally intractable.[1]

In all of the simulated scenarios, MANOVA was more powerful than the other methods. However, consistent with results described above, we observed that the multistep combined PC (mCPC) strategy (Equation 13) can substantially improve power (Figure 5) and outperform MANOVA in some cases. The optimal number of groups and number of PCs per group (K in Equation 13) that will maximize the gain in power depends on the

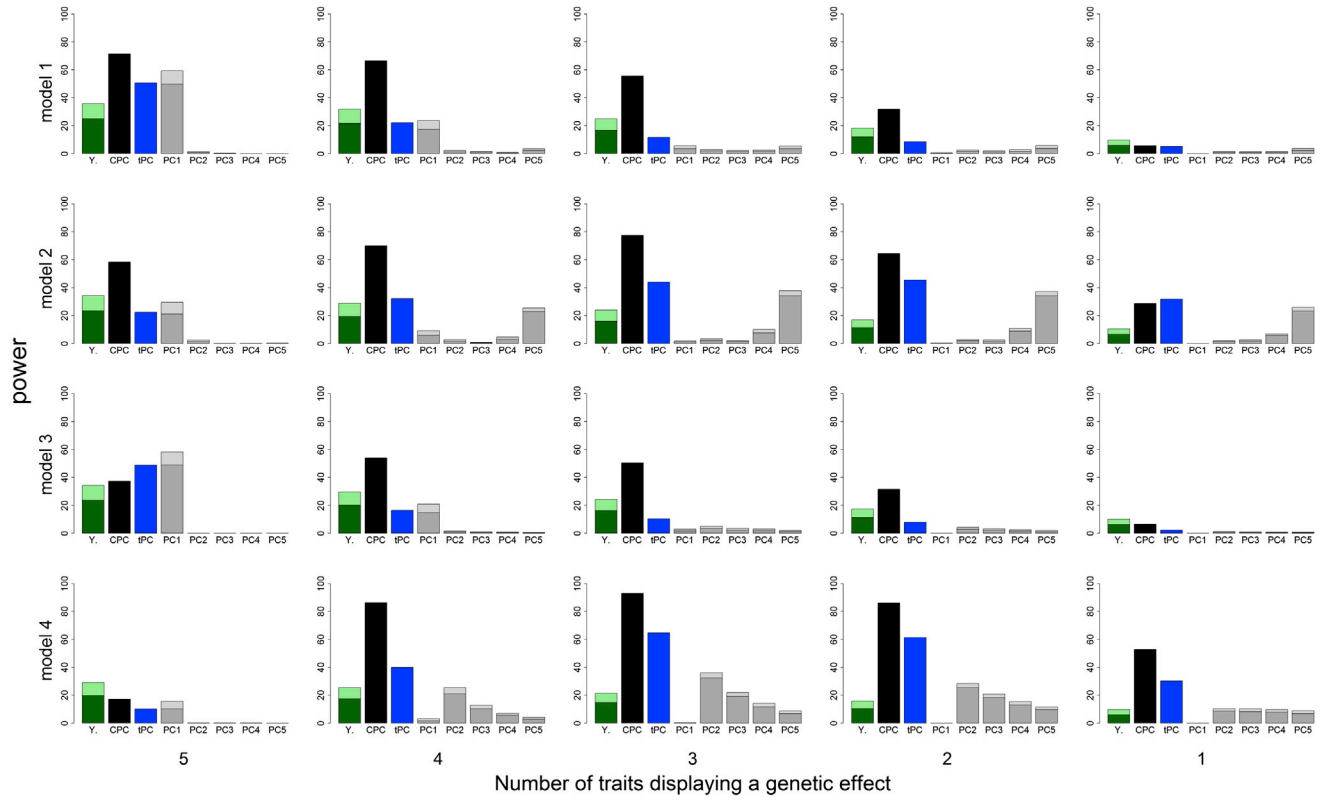

Figure 3. Power Comparison for the Multivariate Analysis of Five Traits in the Presence of Pleiotropic Effect
Power at 5×10^{-8} significance level for the detection of a genetic variant when analyzing five phenotypes. Between one and five phenotypes are simulated as a function of the genetic variant, where its proportion of variance explained was randomly chosen between 0.1% and 0.5%. All genetic effects were positive and the associated phenotypes were randomly selected with equal probability. The bars represent the power of eight different tests. The univariate tests for each PC are shown in light gray and dark gray after correcting for the multiple testing, and the univariate test for the most significant PC (tPC) is in blue. The combined test of all five PCs is shown in black, and the most significant univariate test of all *Y.* in light green (dark green after correcting for the multiple testing). The power is shown for four different correlation models and 10,000 simulation replicates with 5,000 individuals.

underlying architecture of the traits. For this purpose, the level of pleiotropy and the phenotypic structure, i.e., the phenotypic correlation matrix, are probably key components. In the case of two groups, if there is marked phenotypic structure, then the mCPC approach is unlikely to improve power whatever the value of *K* (see Figure 5C and Figure S6, right panels). If there is no marked structure, then the choice of *K* would mainly depend on the level of pleiotropy. For low pleiotropy among the correlated traits, using a high value for *K* (i.e., combining the top 80% PCs and the last 20% PCs) might be more powerful (see Figure 5A and Figure S6, left panels). Conversely, if most correlated traits shared genetic effects, then using a low value for *K* can be more powerful (Figure 5B and Figure S6, middle panels). Although defining an optimal strategy would require a fine study exploring various parameters, these simulations highlight that the multistep combined PC method can incorporate investigators' hypotheses regarding the specific genetic architecture of the correlated traits, with limited loss of power when the assumption does not hold.

We then evaluated the robustness and flexibility of the four methods for handling several issues that commonly arise in GWASs. We observed that MANOVA, TATES, and CPC are sensitive to the presence of outliers and can make the test invalid, whereas Multiphen is not. However, our simulations showed that the type I error is inflated only in the presence of a large number of outliers with values that are an order of magnitude larger than expected (data not shown). Moreover, deviation from the null hypothesis affects mostly genetic variants with very low frequencies (e.g., MAF $< 1\%$ in 1,000 individuals) but have a small impact on common genetic variants. Contrariwise, Multiphen suffers from an inflated type I error when the ratio of number of phenotypes over number of individual is relatively large (>0.01), e.g., for 50 phenotypes in 1,000 individuals, the lambda value is 1.3 (see Figure S7), whereas CPC, TATES, and MANOVA do not suffer from this problem.

We have focused on analyses of unrelated individuals. To our knowledge, there is no simple extension of MANOVA or Multiphen to family data. Conversely, for TATES and CPC, one can easily apply well-established methods such as mixed models. For illustrative purposes we simulated 10,000 replicates of 200 nuclear families of two parents and one to five children for a total of 1,000 subjects and

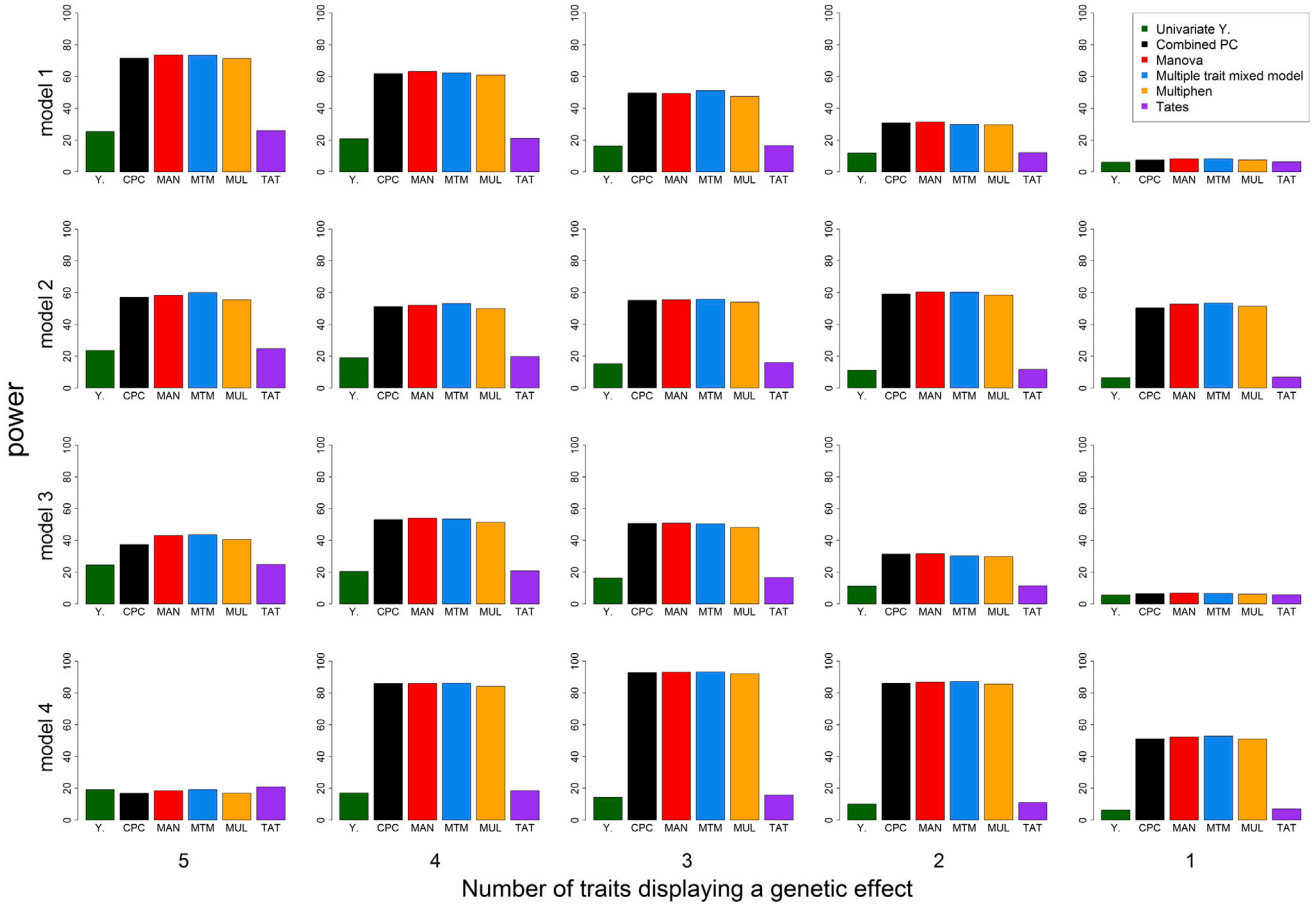

Figure 4. Power of Alternative Methods for the Multivariate Analysis of Five Traits
A comparison of the power at 5×10^{-8} significance level for detecting a genetic variant by five different multiple trait analysis: the combined PCs (CPC, in black); MANOVA (MAN, in red); multitrait mixed model (MTM, in blue); Multiphen (MUL, in orange); and TATES (TAT, in purple). These tests were applied to five traits where the number of traits with causal genetic effect was varied between one and five. All genetic effects were positive and associated phenotypes were selected randomly with probability proportional to their level of correlation with other phenotype. The proportion of variance explained by the causal variant was randomly chosen between 0.1% and 0.5%. The power is shown for four different correlation models and 10,000 simulation replicates with 5,000 individuals.

analyzed 10 phenotypes under the null hypothesis of no association with the tested genotypes. Although applying any of the tests described above for unrelated individuals to family data results in an inflated type I error rate, except for MTMM, applying a mixed model (as implemented in the software EMMAX[29]) to the univariate phenotypes and to the univariate PCs for TATES and CPC, respectively, solves this issue (Figure 6A). When generating data under the alternative (Figure 6B), we observed that CPC applied in conjunction with a mixed model and MTMM had the highest power (with MTMM being slightly more powerful) as compared to applying genomic control correction to the test displaying inflated type I error rate. Overall, the combined PC approach and TATES offer more flexibility than do integrated approaches such as MANOVA and Multiphen. In particular, it is straightforward to apply alternative tests for association that capture nonlinear effects, such as the DC test we recently developed[30] or tests of homogeneity of variance by genotypic classes.[31,32] As showed in Figure S8, when applied to the test of variances, CPC has a well-calibrated type I error rate under the null whereas combining signal across univariate phenotypes does not.

GWAS of Coagulation-Related Phenotypes

To illustrate the importance of including principal components that explain a small proportion of total phenotypic variance in the analysis, we conducted a genome-wide scan of 5 coagulation-related phenotypes in 685 individuals from the MARTHA study, namely fibrinogen (FIB), factor VIII (FVIII), von Willebrand factor (vWF), the activated partial thromboplastin time (aPTT), and the standardized anticoagulant response to Agkistrod on contortrix venom (ACV). All these phenotypes reflect global coagulation activity and display moderate to strong correlation. The correlation matrix between these traits (Table S1) was very similar to that simulated in model 2, with a gradient of absolute value correlation varying between 0.75 (FVIII and vWF) to 0.013 (FIB and aPTT). The five principal components extracted from the standardized phenotypes explained 46.22, 18.86, 18.32, 12.48, and 4.11 percent of

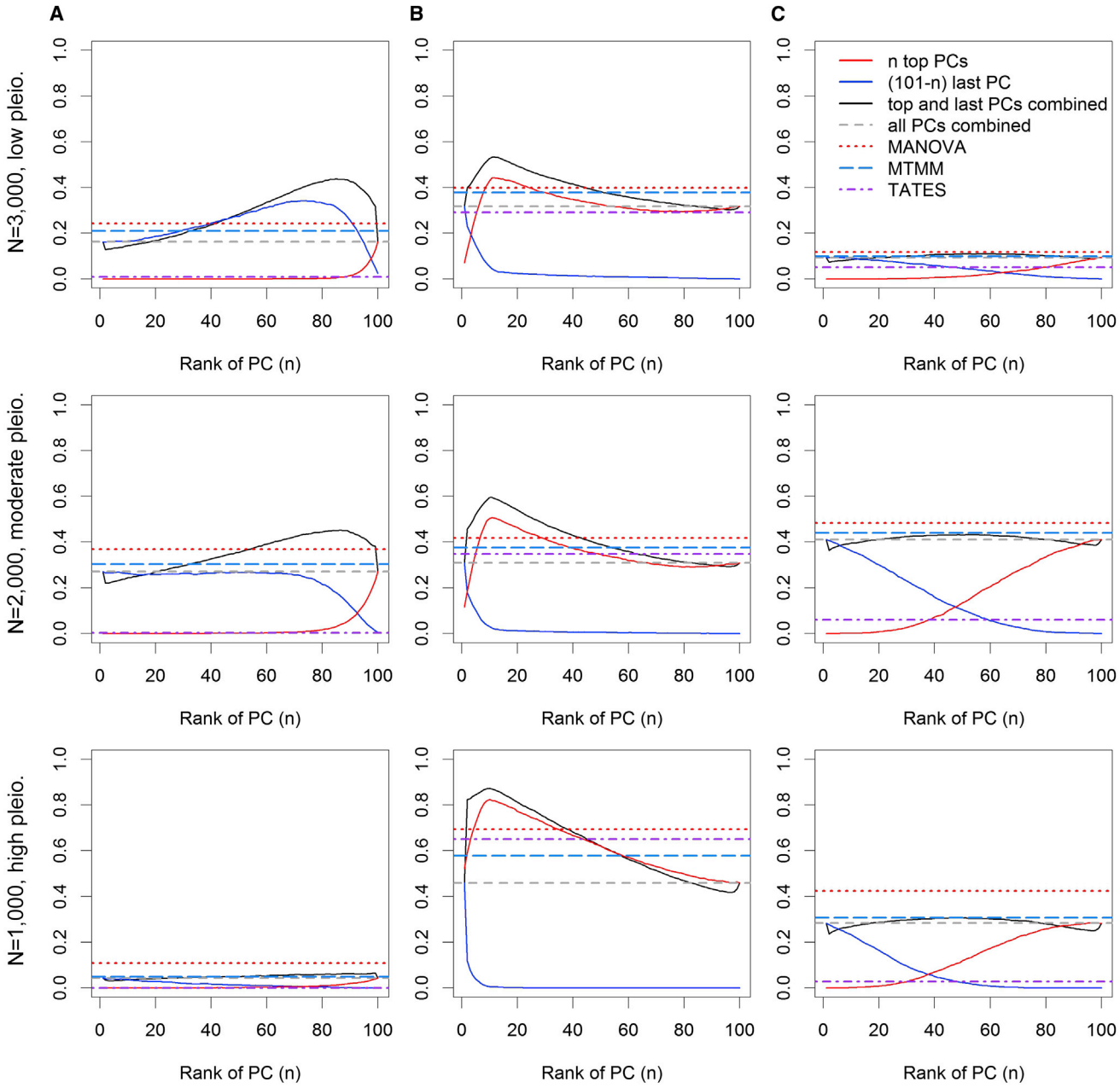

Figure 5. Power Comparison for the Multivariate Analysis of 100 Traits
Power at 5×10^{-8} significance level for seven different tests when analyzing 100 phenotypes across 10,000 replicates. Plots were simulated under schemes SC1 (A), SC2 (B), and SC3 (C) (described in Appendix B and Figure S5). The top, middle, and bottom rows show the power for low, moderate, and high level of pleiotropy with sample sizes of 3,000, 2,000, and 1,000, respectively. The red curve corresponds to the test combining signals from the *n* PCs associated with the largest eigenvalues, the dark blue curve corresponds to the test combining signals on the 101-*n* PCs associated with the smallest eigenvalues, and the black curve corresponds to the combined test of latter two tests by the Fisher's method, with *n* varying from 1 to 100. The dashed lines correspond to the test of all PCs combined (gray), MANOVA (red), multitrait mixed model (MTMM, blue), and TATES (purple).

the total phenotypic variance, respectively. The individual trait loadings for each of the five PCs are presented in Table S2.

All tests we applied showed correct distribution of p values (Table S3). The results from the univariate analysis of each original trait and the univariate analysis of each PC (adjusted for multiple hypothesis testing by Bonferroni correction) compared to the multivariate analysis of the PCs were consistent with the conclusions drawn from the simulation study. All of the five loci that were found to be genome-wide significant ($p < 5 \times 10^{-8}$) in a single-trait analysis were also significant in the combined PC analysis (Table 2). Conversely, focusing for example on the top two PCs explaining more than 55% of the variance as done in Avery et al.[13] would have identified only one of these five SNPs. The combined PCs analysis furthermore

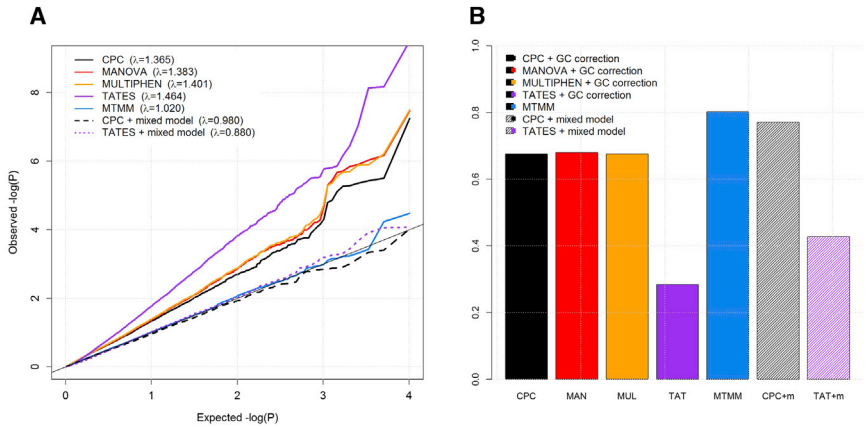

Figure 6. Multivariate Analysis of Family Data

Comparison of five multivariate tests for the analysis of 10 phenotypes in 200 nuclear families including two parents and one to five children for a total of 1,000 subjects. Pairwise phenotypic correlations follow a gradient from 0 to 0.8 (extended model 2 from Figure 3).

(A) QQplots and lambda values under the null hypothesis of no association between the tested SNP and any of the ten phenotypes.

(B) Power under the alternative, when the SNP is associated with three phenotypes chosen randomly, and proportion of phenotypic variance explained by the SNP varying in [0, 0.025]. We compared MANOVA (red), MultiPhen (orange), TATES (purple), CPC (combined PCs analysis, black), and MTMM (multitrait mixed model, blue). Under the null, all tests, except MTMM, show inflated type I error rate when the family structure is not accounted for. Applying a mixed model to the univariate phenotype or univariate PC analysis for TATES and CPC, respectively (dashed lines), solve this issue. Under the alternative, we applied a genomic control (GC) correction to all tests showing type I error inflation. Power was derived at a significance level of 5×10^{-3}.

identified two variants, one on chromosome 18q21.2 located between two genes, *DCC* (MIM 120470) and the *RPS8P3* pseudo-gene ($p = 1.7 \times 10^{-8}$). This SNP was suggestive genome significant based on the univariate analysis of FVIII ($p = 1.7 \times 10^{-7}$ and beta coefficient of the coded allele [allele T against C] of 1.8) and slightly associated with vWF ($p = 0.035$ and beta coefficient of 0.89). The second locus found in the combined PCs analysis was just below the genome-wide significant level ($p = 5.8 \times 10^{-8}$). It is located on chr10p11.22 lying between the interesting genes *ITGB1* (MIM 135630) and *NRP1* (MIM 602069). This SNP had a marginal effect on FVIII and vWF (the beta coefficients of the coded allele [allele A against G] were 0.502 and 0.297, respectively) that was not suggestive of genome-wide association (p value of 6.6×10^{-7} and 0.0073, respectively).

Several patterns of association observed in the simulation study were also observed in the empirical data. First, the genetic signal at the most associated loci was spread out across the PCs with association pattern changing across loci; thus, focusing only on the univariate signal from the top PCs was suboptimal. Applying the combined PCs approach on the top two or three PCs was underpowered for the same reason (data not shown). When the SNP was associated to a single trait that was moderately correlated to the others, there was no gain in using PCA-based approaches (e.g., SNPs rs6025, rs710446, and rs191945075 in Table 2); however, when the affected trait had correlation with another trait above 0.5, the combined PCs had the highest power (e.g., SNPs rs576123, rs183013917, and rs76854392 in Table 2). Indeed, we noted that association signals at the additional loci identified by the combined PCs analysis were driven by signal from the last PC, which explained 4% of the total phenotypic variance. These signals involve nonpleiotropic effects, or at least unbalance genetic effect on FVIII and vWF, the most correlated traits. This confirms first that PCA of multiple traits can improve detection of both genetic variants harboring pleiotropic effects and those affecting a single trait; and second that PCs explaining a low amount of variance can be as important as those explaining a large amount.

Discussion

Principal component analysis is a common tool that has been widely used for the combined analysis of correlated phenotypes in genetic linkage and association studies. In this study we show that PC-based analyses that focus on the few components explaining most of the phenotypic variance, as done in many studies, is generally suboptimal. By deriving the power for PC analysis in a simple case of two phenotypes and by conducting simulations for more complex situations, we show that a genetic association signal may in practice be spread across many or all of the principal components. Under many realistic scenarios, important genetic signals, e.g., trait-specific or negative pleiotropic genetic effects (e.g., positive correlation and opposite genetic effects), are captured by the PCs explaining the least amount of the total phenotypic variance. We demonstrated that combining the signal from all PCs can, therefore, be a more efficient strategy than the standard PCA approach. Although focusing on only a few PCs is unlikely to improve statistical power when the total number of phenotypes analyzed jointly is small (e.g., <10), this is not necessarily the case when analyzing a very large number of phenotypes. When considering a more diverse range of underlying models, involving multiple latent variables and various patterns of genetic effects, we found that increase in power can be achieved by applying a naive multistep approach, where signal on PCs are merged into subgroups based on their eigenvalues and the association signal across all groups tested jointly by the Fisher's

Table 2. Genome-wide Significant SNPs from the Three Approaches: Univariate Phenotypes, Univariate PCs, and Combined PCs Approach

SNP ID	Chr	Gene[a]	MAF	Rsq	p Value									
					ACVn	FIB	FVIII	aPTT	vWF	PC1	PC2	PC3	PC4	PC5	combPC
rs6025	1	F5	0.90	0.93	5.3×10^{-23}*	–	–	–	–	0.21	0.0091	1.1×10^{-17}	–	–	1.8×10^{-18}
rs710446	3	KNG1	0.54	0.99	–	–	–	5.2×10^{-11}*	–	0.064	0.0083	–	8.2×10^{-8}	–	3.3×10^{-10}
rs1801020	5	F12	0.78	0.86	0.29	0.19	–	2.2×10^{-12}	–	–	1.0×10^{-5}	0.0057	5.8×10^{-9}	–	7.6×10^{-15}*
rs576123	9	ABO	0.55	0.86	–	–	2.7×10^{-9}	9.7×10^{-5}	1.1×10^{-13}	3.6×10^{-10}	0.00011	–	0.0062	–	1.2×10^{-14}*
rs76854392	10	NRP1	0.90	0.74	–	–	6.6×10^{-7}	–	0.0073	0.00075	–	0.029	0.28	0.00034	5.8×10^{-8}*
rs191945075	11	LRP4	0.96	0.60	2.8×10^{-25}*	0.095	–	–	–	0.017	0.00072	1.5×10^{-17}	–	–	1.3×10^{-19}
rs183013917	18	DCC	0.99	0.82	–	–	1.7×10^{-7}	–	0.035	0.0017	–	–	0.12	7.0×10^{-6}	1.8×10^{-8}*

Abbreviations are as follows: Chr, chromosome; MAF, minor allele frequency; Rsq, quality control imputation criterion; FIB, Fibrinogen; FVIII, factor VIII; vWF, von Willebrand factor; aPTT, the activated partial thromboplastin time; ACVn, the standardized Anticoagulant response to Agkistrodon contortrix venom; combPC, combined PCs approach.
The p values for univariate raw phenotype analysis and univariate PC analysis were adjusted for multiple tests via the Bonferroni correction. p values greater than 0.3 were replaced by a dash "–." The most significant test is indicated with an asterisk (*).
[a]Physically closest gene.

method. This strategy may be worth exploring more deeply in the future.

We compared the power of the combined PC (CPC) approach against four existing methods: MANOVA; an ordinal logistic regression with the genetic variant as the outcome and the phenotypes as predictors (Multiphen); a multitrait linear mixed model (MTMM); and the p value correction method TATES. Overall, CPC showed power close to the optimal in all of the scenarios considered while offering more flexibility and robustness than other methods. In particular, it can account for population and family structure when applied together with a mixed linear model, which enables applications to family data whatever the number of traits in a reasonable computation time, whereas MTMM, for example, would be computationally intractable for more than 5–10 traits.[1] It can also be applied in conjunction with other univariate tests, e.g., nonparametric tests, or tests of heterogeneity of variance by phenotypic classes, an approach now commonly applied in genomic data.[32] Comparatively, MANOVA was the most powerful approach when a large number of phenotypes was analyzed (e.g., >50). The performance of Multiphen was similar to CPC and MANOVA when a small number of phenotypes was analyzed, but it had an inflated type I error when the total number of phenotypes analyzed was large as compared to the number of subjects. However, to our knowledge, no statistical method has been developed to account for complex population or family structure in MANOVA and Multiphen; these methods are therefore not applicable to family data. The MTMM approach offers a more global framework than CPC to study multiple correlated phenotypes, accounting for population structure and providing additional estimates including the genetic and environmental variance of each phenotype. However, this increased complexity comes at a dramatic cost in computation time so that the approach can be applied to only a limited number of traits. Although TATES offers substantial flexibility—it can both handle structured data and be applied to various statistical tests—this method was dramatically underpowered as compared to the other approaches in most scenarios we considered.

In summary, we believe that the combined PC approach is an attractive approach because it retains relatively good power across a wide range of alternatives while preserving computational efficiency. Moreover, as shown in Figure 5, we can improve the power of CPC by incorporating prior knowledge about the underlying architecture of the multivariate phenotypes, but not for the other approaches. Finally, we note that performing a meta-analysis across multiple studies of a multiphenotypes test, including CPC, is more complex than conducting a meta-analysis of a univariate test. Simple approaches such as combining p values across studies by the Fisher's method are possible, but more efficient strategies might be developed. We are actively pursuing this goal.

We simulated a wide range of trait correlation and pleiotropic patterns, but no simulation study can be exhaustive.

It is possible that analyses based on specific mechanistic hypotheses may be more powerful when the mechanistic hypotheses hold. But such methods often lose power when the hypothesized mechanism does not hold. For example, the recently proposed TATES statistic was shown to outperform MANOVA in a range of scenarios when the genetic effect of a SNP was constant across multiple traits (e.g., $v_1 = v_2$ in Equations 1 and 2).[8] When we compared the performance of tests in situations where the genetic effects varied across traits, TATES had notably less power than other approaches, including the CPC strategy (Figures 4 and 5). Although there exist situations where some tests may have more power, we believe the combined PC approach is an attractive approach because it retains relatively good power across a wide range of alternatives. This makes the combined PC approach particularly appealing when the underlying mechanism is unknown.

The genetic variants identified in the MARTHA study at genome-wide significance level are mostly known variants. The association between *ABO* (MIM 110300) and FVIII and vWF has been known for decades;[33,34] the two variants associated to aPTT (rs710446 in *KNG1* [MIM 612358] and rs1801020 in *F12* [MIM 610619]) have been reported previously;[24] and the two loci associated with ACVn, *LRP4* (MIM 604270) and *F5* (MIM 612309), have already been described by Oudot et al.[21] Two additional variants were identified by the combined PCs approach at genome-wide or nearly genome-wide significance level: the SNP rs183013917 (which is mainly associated with FVIII) and the SNP rs76854392 near *NRP1* (which is associated with both FVIII and vWF). Although these are potential candidates, especially *NRP1* because of previous studies showing association with angiogenesis (e.g., Vander Kooi et al.[35]), they deserve further replication before being confirmed.

A number of studies have used principal component analysis for the multivariate analysis of correlated phenotypes. Most of them followed the standard strategy that consists of reducing the dimension of the outcome data by focusing on the few components that explain the most variability in the outcomes and removing those explaining a low amount of total variance. In this work, we show that contrary to this widespread practice, testing the top PCs only can be dramatically underpowered because PCs explaining a low amount of the total phenotypic variance can harbor a substantial part of the total genetic association. We also demonstrate that PCA-based strategies achieve a moderate gain in power only in the presence of positive pleiotropy, but have great potential to detect negative pleiotropy or genetic variants that are associated with a single trait highly correlated to others.

Appendix A

The proportion of the total variance explained by PC_1 and PC_2, respectively s_1 and s_2, is defined as

$$
\begin{aligned}
s_1 &= \frac{var(PC_1)}{var(Y_1) + var(Y_2)} \\
&= \frac{var(PC_1)}{2} \\
&= \left(\frac{4 * c}{2} + \frac{(\sqrt{v_1} + \sqrt{v_2})^2}{2} + \frac{(1 - c - v_1)}{2} + \frac{(1 - c - v_2)}{2}\right) \Big/ 2 \\
&= \left(\frac{4 * c + v_1 + v_2 + 2 * \sqrt{v_1 v_2} + 1 - c - v_1 + 1 - c - v_2}{2}\right) \Big/ 2 \\
&= \left(\frac{2 + \ 2 * c + 2 * \sqrt{v_1 v_2}}{2}\right) \Big/ 2 \\
&= (1 + \ c + \sqrt{v_1 v_2})/2
\end{aligned}
$$

$$
\begin{aligned}
s_2 &= 1 - s_1 \\
&= (1 - c - \sqrt{v_1 v_2})/2.
\end{aligned}
$$

The proportion of variance of PC_1 and PC_2 explained by G, respectively v_{pc1} and v_{pc2}, can be expressed as the ratio of the genetic effect of G on the variance of PC_1 and PC_2, that we denoted τ_1 and τ_2, respectively, divided by the total variance of PC_1 and PC_2, which are equal to s_1 and s_2, respectively

$$
\begin{aligned}
v_{pc1} &= \frac{\tau_1}{var(PC_1)} \\
&= \frac{\left(\frac{\sqrt{v_1} + \sqrt{v_2}}{\sqrt{2}}\right)^2}{(1 + \ c + \sqrt{v_1 v_2})} \\
&= \frac{v_1 + v_2 + 2 * \sqrt{v_1 v_2}}{2 * (1 + c + \sqrt{v_1 v_2})}
\end{aligned}
$$

$$
\begin{aligned}
v_{pc2} &= \frac{\tau_2}{var(PC_2)} \\
&= \frac{\left(\frac{\sqrt{v_1} - \sqrt{v_2}}{\sqrt{2}}\right)^2}{(1 - c - \sqrt{v_1 v_2})} \\
&= \frac{v_1 + v_2 - 2 * \sqrt{v_1 v_2}}{2(1 - c - \sqrt{v_1 v_2})}.
\end{aligned}
$$

Appendix B

When simulating 100 phenotypes, we considered three different simulation schemes (SC1, SC2, and SC3), which are illustrated in Figure S5. In model SC1 the correlation between phenotypes is a result of a limited number (30) of independent latent variables, each affecting 40 phenotypes on average and explaining altogether 30% of the total phenotypic variance. The magnitude of the genetic effects on the phenotypes was generated independently from the latent variables, but the associated phenotypes were selected while accounting for the pairwise phenotypic correlation. In model SC2, the phenotypic correlation was generated in similar fashion as in model SC1, i.e., using the same latent variables, but the genetic variant

was associated with some of these latent variables but not directly to the phenotypes. In model SC3, we considered a more complex model involving a thousand latent variables together explaining 90% of the total phenotypic variance, subgroups of these latent variables affecting cluster of phenotypes. As for model SC2, the genetic variant was associated with some of the latent variables.

More specifically, in model SC1, the n_{phe} phenotypes were generated as follows: $\mathbf{y} = \sqrt{\mathbf{c}} \times (\boldsymbol{\beta}^{\boldsymbol{t}} \times \mathbf{u}) + \sqrt{\gamma} \times \mathrm{g} + \sqrt{(1 - \gamma - \mathbf{c})} \times \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of phenotypic values for a given subject, g is a SNP, γ is the vector of proportion of variance explained by that SNP on the n_{phe} phenotypes, $\mathbf{u}$ is a vector of realization of n_{u} independent latent variables $U = (U_1, \ldots U_{n_u})$ normally distributed with mean 0 and variance 1, $\boldsymbol{\beta}$ is a matrix of (positive) weights with n_{u} rows and n_{phe} columns that defines the contribution of each variable U_{i} to the phenotypes (these weights are defined so that for each phenotype j, $\sum_{i=1}^{n_u} \beta_{ij}^2 = 1$), $\mathbf{c}$ is a vector of n_{phe} weights that defined the proportion of variance explained by each linear combination $\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{t}} \times u_i$, and $\boldsymbol{\varepsilon}$, the residual variance, is a vector of n_{phe} independent variables, normally distributed with mean 0 and variance 1. In models SC2 and SC3, the n_{phe} phenotypes were generated as follows: $\mathbf{y} = \sqrt{\mathbf{c}} \times (\boldsymbol{\beta}^{\boldsymbol{t}} \times \mathbf{u}) + \sqrt{(1 - \mathbf{c})} \times \boldsymbol{\varepsilon}$. A subsample of the u_{i} were generated as a function of a SNP G such that $u_{\text{i}} = \sqrt{\delta_{\text{i}}} \times \mathrm{g} + \sqrt{(1 - \delta_{\text{i}})} \times \mathrm{e}_i$, where δ_{i} is the effect of SNP on the latent variable U_{i}, and e_{i}, the residual, is normally distributed with mean 0 and variance 1.

The parameters $\mathbf{c}$ and $\boldsymbol{\beta}$ and γ of the three models were defined empirically to obtain a similar distribution of pairwise phenotypic correlation and a similar distribution of proportion of variance explained by the PCs. Because the three models were very different, it was difficult to generate a similar range of genetic effects on the phenotypes. However, this was of secondary importance because the aim of this experiment was not to compare the models but to compare the power of different methods for analyzing correlated traits under various scenarios. Regardless, with these constraints, the genetic effects simulated (either directly on the phenotypes or on the latent variables) were explaining a very small proportion of the total phenotypic variance (<0.05%). For example, in model SC1, the SNP was associated with 10 traits and the variance that it explained for each phenotype was drawn from a uniform distribution with minimum 0.1% and maximum 0.5%, so that the total phenotypic variance explained was on average 0.025%. In model SC2, the SNP had a similar effect size on 3 latent variables, so that the average contribution of the SNP was 0.007%.

Appendix C

For N_{phe} phenotypes and N_{ind} individuals, a common modeling for multiple traits mixed model is as follows: $\mathrm{Y} = \beta \mathbf{X}^T + \mathbf{G} + \mathbf{E}$, where $\mathbf{G} = \mathbf{V}_{\mathrm{G}} \otimes \mathbf{K}$ and $\mathbf{E} = \mathbf{V}_{\mathrm{E}} \otimes \mathbf{I}$, Y is a column vector of N_{phe} phenotypes, $\mathbf{X}$ is a $N_{\text{ind}} \times N_{\text{phe}}$ by N_{phe} matrix defines as $\mathbf{I}_{\boldsymbol{N_{phe}}} \otimes \boldsymbol{g}$ where g is a vector of genotypes for a given genetic variant and β is a N_{phe} vector of its effect sizes; $\mathbf{G}$ is a $N_{\text{ind}} \times N_{\text{phe}}$ by $N_{\text{ind}} \times N_{\text{phe}}$ matrix of genetic effects and $\mathbf{E}$ is a $N_{\text{ind}} \times N_{\text{phe}}$ by $N_{\text{ind}} \times N_{\text{phe}}$ matrix of residual errors; $\mathbf{K}$ is a N_{ind} by N_{ind} relatedness matrix and $\mathbf{I}$ is the N_{ind} by N_{ind} identity matrix; and $\mathbf{V}_{\mathrm{G}}$ is a N_{phe} by N_{phe} matrix of genetic variance and $\mathbf{V}_{\mathrm{E}}$ is a N_{phe} by N_{phe} matrix of environmental variance.

Although the standard MTMM test includes estimating all of the variance components for $\mathbf{G}$ and $\mathbf{E}$ from the data, we decided to use the true parameters because estimating variance components for a large set of correlated traits is computationally intensive. The variance component parameters used exclude the effect of the tested SNP, which leads to a better calibrated statistic and improved power over the approximate LRT approach used by MTMM.[1] We derived generalized least square estimate for β for each simulated SNP X as follows:

$$\widehat{\beta} = \frac{\boldsymbol{X}^t \boldsymbol{V}^{-1} Y}{\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X}}, \text{ where } \mathbf{V} = \mathbf{G} + \mathbf{E}.$$

We obtained the p value by using an F-test, where the residual sum of squares for the null model, $\mathrm{RSS}_0 = \mathrm{Y}^t \mathbf{V}^{-1} \mathrm{Y}$, and the alternative model, $\mathrm{RSS}_1 = (\mathrm{Y} - \mathbf{X}\widehat{\beta})^t \mathbf{V}^{-1} (\mathrm{Y} - \mathbf{X}\widehat{\beta})$, assuming the genotypes and phenotypes have mean zero and variance one.

Although multivariate mixed models is feasible for a small number of traits (e.g., $N_{\text{phe}} < 10$), it becomes computationally intensive for more traits, because it requires estimating a large number of dependent covariance parameters.[1,5]

For our simulations with unrelated individuals, we assumed that pairs of different individuals were zero (i.e., $\mathbf{K} = \mathbf{I}$) and we do not separate genetic and environmental variance. Under this assumption, deriving $\mathbf{V}^{-1}$ can be simplified to $\mathbf{V}^{-1} = (\mathbf{V}_{\mathrm{G+E}} \otimes \mathbf{I})^{-1} = \mathbf{V}_{\mathrm{G+E}}^{-1} \otimes \mathbf{I}$. Because $\mathbf{V}_{\mathrm{G+E}}$ is a N_{phe} by N_{phe} matrix, it can very easily be inverted, even when analyzing hundreds of traits. Furthermore, the simple form of $\mathbf{V}^{-1}$ allows the derivation of $\widehat{\beta}$, RSS0, and RSS1 through linear algebra, which avoids saving $\mathbf{V}^{-1}$, a practical and necessary advantage for 100 traits and 1,000 individuals, because such matrix would require more than 500 Gb of memory.

Supplemental Data

Supplemental Data include eight figures and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2014.03.016.

Acknowledgments

We would like to thank Alkes L. Price, Alexander Gusev, Amit Joshi, Po-Ru Loh, Nan Laird, and Matthew Stephens for valuable discussions and helpful comments. H.A. was supported by grant R03HG006720. P.K. and H.A. were supported by grant R21CA165920. The MARTHA project was supported by a grant from the Program Hospitalier de la Recherche Clinique. Statistical

analyses of the MARTHA data sets were performed with the C2BIG computing cluster, funded by the Région Ile de France, Pierre and Marie Curie University, and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05).

Received: January 15, 2014
Accepted: March 24, 2014
Published: April 17, 2014

Web Resources

The URLs for data presented herein are as follows:

CRAN – Package MultiPhen, http://cran.us.r-project.org/web/packages/MultiPhen/index.html
CTG Lab software (including TATES), http://ctglab.nl/software/?/software
Hugues Aschard software, http://www.hsph.harvard.edu/hugues-aschard/software/
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
Mach2qtl, https://helix.nih.gov/Applications/mach2qtl.html
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
UCSC Human Genome Browser, http://genome.ucsc.edu/cgi-bin/hgGateway

References

1. Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods. Published online February 16, 2013. http://dx.doi.org/10.1038/nmeth.2848.
2. Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. PLoS ONE *8*, e65245.
3. Yang, Q., and Wang, Y. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. J. Probab. Stat. *2012*, 13.
4. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495.
5. Korte, A., Vilhjálmsson, B.J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. *44*, 1066–1071.
6. O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. Biometrics *40*, 1079–1087.
7. Yang, Q., Wu, H., Guo, C.Y., and Fox, C.S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet. Epidemiol. *34*, 444–454.
8. van der Sluis, S., Posthuma, D., and Dolan, C.V. (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. *9*, e1003235.
9. Ferreira, M.A., and Purcell, S.M. (2009). A multivariate test of association. Bioinformatics *25*, 132–133.
10. Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet. Epidemiol. *32*, 9–19.
11. Elston, R.C., Buxbaum, S., Jacobs, K.B., and Olson, J.M. (2000). Haseman and Elston revisited. Genet. Epidemiol. *19*, 1–17.
12. Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M.M., Hysi, P.G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M.A., et al. (2012). A genome-wide association study identifies five loci influencing facial morphology in Europeans. PLoS Genet. *8*, e1002932.
13. Avery, C.L., He, Q., North, K.E., Ambite, J.L., Boerwinkle, E., Fornage, M., Hindorff, L.A., Kooperberg, C., Meigs, J.B., Pankow, J.S., et al. (2011). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. PLoS Genet. *7*, e1002322.
14. He, L.N., Liu, Y.J., Xiao, P., Zhang, L., Guo, Y., Yang, T.L., Zhao, L.J., Drees, B., Hamilton, J., Deng, H.Y., et al. (2008). Genome-wide linkage scan for combined obesity phenotypes using principal component analysis. Ann. Hum. Genet. *72*, 319–326.
15. Karasik, D., Cheung, C.L., Zhou, Y., Cupples, L.A., Kiel, D.P., and Demissie, S. (2012). Genome-wide association of an integrated osteoporosis-related phenotype: is there evidence for pleiotropic genes? J. Bone Miner. Res. *27*, 319–330.
16. Karasik, D., Cupples, L.A., Hannan, M.T., and Kiel, D.P. (2004). Genome screen for a combined bone phenotype using principal component analysis: the Framingham study. Bone *34*, 547–556.
17. Rainwater, D.L., Mahaney, M.C., VandeBerg, J.L., Brush, G., Almasy, L., Blangero, J., Dyke, B., Hixson, J.E., Cole, S.A., and MacCluer, J.W. (2004). A quantitative trait locus influences coordinated variation in measures of ApoB-containing lipoproteins. Atherosclerosis *176*, 379–386.
18. Zhang, F., Guo, X., Wu, S., Han, J., Liu, Y., Shen, H., and Deng, H.W. (2012). Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. PLoS ONE *7*, e53320.
19. Jolliffe, I.T. (1982). A note on the use of principal components in regression. J. R. Stat. Soc. Ser. C Appl. Stat. *31*, 300–303.
20. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE *7*, e34861.
21. Oudot-Mellakh, T., Cohen, W., Germain, M., Saut, N., Kallel, C., Zelenika, D., Lathrop, M., Trégouët, D.A., and Morange, P.E. (2012). Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. Br. J. Haematol. *157*, 230–239.
22. Antoni, G., Oudot-Mellakh, T., Dimitromanolakis, A., Germain, M., Cohen, W., Wells, P., Lathrop, M., Gagnon, F., Morange, P.E., and Tregouet, D.A. (2011). Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. BMC Med. Genet. *12*, 102.
23. Xu, S. (2013). Major gene detection. In Principles of Statistical Genomics (New York: Springer), pp. 61–78.
24. Tang, W., Schwienbacher, C., Lopez, L.M., Ben-Shlomo, Y., Oudot-Mellakh, T., Johnson, A.D., Samani, N.J., Basu, S., Gögele, M., Davies, G., et al. (2012). Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. Am. J. Hum. Genet. *91*, 152–162.
25. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
26. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate

haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
27. Gilmour, A., Thompson, R., and Cullis, B.R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics *51*, 1440–1450.
28. Meyer, K. (2007). WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). J. Zhejiang Univ. Sci. B *8*, 815–821.
29. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–1723.
30. Aschard, H., Zaitlen, N., Tamimi, R.M., Lindström, S., and Kraft, P. (2013). A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. Genet. Epidemiol. *37*, 323–333.
31. Paré, G., Cook, N.R., Ridker, P.M., and Chasman, D.I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet. *6*, e1000981.
32. Yang, J., Loos, R.J., Powell, J.E., Medland, S.E., Speliotes, E.K., Chasman, D.I., Rose, L.M., Thorleifsson, G., Steinthorsdottir, V., Mägi, R., et al. (2012). FTO genotype is associated with phenotypic variability of body mass index. Nature *490*, 267–272.
33. Orstavik, K.H., Magnus, P., Reisner, H., Berg, K., Graham, J.B., and Nance, W. (1985). Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. Am. J. Hum. Genet. *37*, 89–101.
34. Gill, J.C., Endres-Brooks, J., Bauer, P.J., Marks, W.J., Jr., and Montgomery, R.R. (1987). The effect of ABO blood group on the diagnosis of von Willebrand disease. Blood *69*, 1691–1695.
35. Vander Kooi, C.W., Jusino, M.A., Perman, B., Neau, D.B., Bellamy, H.D., and Leahy, D.J. (2007). Structural basis for ligand and heparin binding to neuropilin B domains. Proc. Natl. Acad. Sci. USA *104*, 6152–6157.