Integrating Multiple Correlated Phenotypes for Genetic Association Analysis by Maximizing Heritability

Jin J. Zhou[1,2], Michael H. Cho[3,4,5], Christoph Lange[2], Sharon Lutz[6], Edwin K. Silverman[3,4,5], and Nan M. Laird[2]

March 12, 2015

[1] Division of Epidemiology and Biostatistics, College of Public Health, University of Arizona, Tucson, AZ 85724

[2]Department of Biostatistics, Harvard University

[3]Channing Division of Network Medicine, Department of Medicine, Brigham and Womens Hospital

[4]Harvard Medical School

[5]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Womens Hospital

Boston, MA 02115

[6]Department of Biostatistics and Informatics, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045

Running Title:

Integrating Correlated Phenotypes

Corresponding author:

Jin Zhou

Division of Epidemiology and Biostatistics, College of Public Health

University of Arizona, Tucson, AZ 85724

Phone: (520) 626-1393

Email: jzhou@email.arizona.edu

Abstract

Many correlated disease variables are analyzed jointly in genetic studies in the hope of increasing power to detect causal genetic variants. One approach involves assessing the relationship between each phenotype and each single nucleotide polymorphism (SNP) individually and using a Bonferroni correction for the effective number of tests conducted. Alternatively, one can apply a multivariate regression or a dimension reduction technique, such as principal component analysis (PCA), and test for the association with the principal components (PC) of the phenotypes rather than the individual phenotypes. Inspired by the previous approaches of combining phenotypes to maximize heritability at individual SNPs, in this paper, we propose to construct a maximally heritable phenotype (MaxH) by taking advantage of the estimated total heritability and co-heritability. The heritability and co-heritability only need to be estimated once, therefore our method is applicable to genome-wide scans. MaxH phenotype is a linear combination of the individual phenotypes with increased heritability and power over the phenotypes being combined. Simulations show that the heritability and power achieved agree well with the theory for large samples and two phenotypes. We compare our approach with commonly used methods and assess both the heritability and the power of the MaxH phenotype. Moreover we provide suggestions for how to choose the phenotypes for combination. An application of our approach to a COPD genome-wide association study shows the practical relevance.

Keywords:

Principal component of heritability, Co-heritability, GWAS, Multivariate analysis

1 Introduction

Complex diseases are often assessed using multiple correlated phenotypes. These phenotypes, sometimes called "endophenotypes", are heritable predicators of disease status. A standard approach to analyze multiple phenotypes is to consider each phenotype separately, but many suggestions have been made for combining the phenotypes in some way with the goal of increasing power, or elucidating disease mechanisms. A multivariate regression strategy is straightforward, but computationally intensive and the power of the approach compared to other approaches depends upon unknown effects (Korte et al., 2012; Schifano et al., 2013). Other strategies use linear combinations of the phenotypes for analysis. Principal component approach (PCA) generates linear combinations through maximizing phenotypic variances (Avery et al., 2011; Karasik et al., 2004). Multiphen (O'Reilly et al., 2012) takes the single SNP as the outcome, multiple phenotypes as the predictors and tests the association between the linear combination of phenotypes and single SNP by ordinal regression. Here we propose a linear combination of the phenotypes that maximizes the total heritability, estimated from a sample of unrelated individuals (Yang et al., 2010); as such our approach is suitable for application to a genome-wide analysis because the linear combination is selected only once, and can be applied to all SNPs on the GWAS chip. The increased heritability of the phenotype translates into improved power for association testing. In contrast, the heritability and the consequent power of the first principal component can be much lower, depending on the genetic parameters (Aschard et al., 2014).

In the linkage era, Ott and Rabinowitz (1999) introduced the approach of incorporating phenotypes into a linear combination with maximized heritability and increased power of locating genes in the context of pedigrees and the presence of pleiotropy. It also has been integrated into a family-based association test for repeated measure analysis by Lange et al. (2004). Klei et al. (2008) first applied it to association studies with independent samples. Their approach, like Lange's (Lange et al., 2004) focused on the notion of optimizing the contribution of a single genetic variant to phenotypic variance which is a fraction of the total heritability of the individual trait. Both Lange's (Lange et al., 2004) and Klei's (Klei et al., 2008) methods estimated the appropriate coefficients for each genetic variant separately. For family trios, Lange *et. al.* (Lange et al., 2004) recommended using the non-informative portion of the family data to estimate this quantity as it is independent of the remaining sample. In population studies, Klei et al. (2008) explored a method of sample splitting and cross validation to determine these coefficients from a training set and then test for association using the remainder of the sample. The method works well for individual SNPs,

but is not practical for a genome-wide association study (GWAS). Our method differs from Klei et al. (2008) and Lange et al. (2004) by globally estimating the total heritability of each single phenotype and estimating genetic covariances of pairs of phenotypes, which only need to be performed once. The combined phenotype (MaxH) is used to test all SNPs.

We compare our method with (1) single phenotype tests adjusting for multiple comparison; (2) univariate test using the first PC of PCA (Avery et al., 2011; Karasik et al., 2004) method; (3) Multiphen (O'Reilly et al., 2012), (4) multivariate regression using Mendel (Lange et al., 2013). Method (2) and (3) use the linear combination of the phenotypes and tests the association through linear regression. Mendel builds upon multivariate regression. It is a likelihood based method using both score and likelihood ratio tests (LRT) for association testing. Recent work from Aschard et al. (2014) shows that testing only the top PCs often has low power, whereas combining signals across all PCs can have greater power. We therefore compared MaxH with multivariate regression using multiple PC phenotypes. Through simulations and real examples, we find our approach proved to have higher power for testing SNPs explaining only a small fraction of the total heritability compared to other univariate association methods.

In the following sections, we first present the method of combining multiple phenotypes through maximizing total heritability and show how power can be approximated analytically for univariate regression given the phenotypic and genotypic variance matrix. In the results section, we provide simple examples illustrate how the heritability changes as a function of the number of phenotypes combined, as well as the impact of missing data. We also provide simulations to show the impact of estimating heritability on power. We use a data example and simulations to compare MaxH with the other approach described above.

2 Material and Method

2.1 Integration of Phenotypes

Let m be the unknown number of independent causal loci, indexed by k, n be the number of individuals, indexed by i, and T be the number of phenotypes, indexed by t. In the absence of any covariates or major gene effects, each phenotype is assumed to have the standard polygenic model (Falconer et al., 1981), given by

$$\begin{aligned} y_{ti} &= \mu_t + \sum_{k=1}^{m} a_{tk} x_{ki} + \epsilon_{ti} \\ &= \mu_t + g_{ti} + \epsilon_{ti}, \end{aligned} \tag{1}$$

where y_{ti} is the tth phenotypic value for the ith individual; μ_t is the mean of the phenotype; x_{ki} is the standardized minor allele count at locus k of individual i, a_{tk} is the additive allelic effect of locus k on phenotype t, $g_{ti} = \sum_{k=1}^{m} a_{tk}x_{ki}$ is the total additive genetic effect of individual i's phenotype t, and the ϵ_{ti} are the residual effects. We treat a_{tk} as random variables independent of the x_{ki}s and of each other, with zero means and common variances and covariances, so that

$$\begin{aligned} \mathrm{E}\left(g_{ti}\right) &= 0 \\ \mathrm{Var}(g_{ti}) &= \sigma^2_{at} \\ \mathrm{Cov}(g_{ti}, g_{t'i}) &= \sigma_{tt'} \\ &= \sigma_{at}\sigma_{at'}\rho_{tt'}, \end{aligned}$$

where $\sigma^2_{at} = \mathrm{Var}(\sum_{k=1}^{m} a_{tk}x_{ki})$ is the total additive genetic variance and $\sigma_{tt'}$ is the covariance between the additive effects for phenotypes t and t', average over the k causal loci. This $\sigma_{tt'}$ can be viewed as the average pleiotropy. Finally, assuming the genetic and environmental effects are independent we have

$$\boldsymbol{V}_p = \mathrm{Var}(\boldsymbol{y}_i) = \mathrm{Var}(\boldsymbol{g}_i + \boldsymbol{\epsilon}_i) = \boldsymbol{V}_g + \boldsymbol{V}_e, \tag{2}$$

where $\mathbf{y}_i$, $\mathbf{g}_i$, and $\boldsymbol{\epsilon}_i$ are the length T vectors of phenotypes, genetic and environment components for the ith individual, and

$$\begin{aligned} \boldsymbol{V}_g &= \mathrm{Var}(g_{ti}) = \begin{pmatrix} \sigma^2_{a_1} & \dots & \sigma_{a_{1T}} \\ \dots & \dots & \dots \\ \sigma_{a_{T1}} & \dots & \sigma^2_{a_T} \end{pmatrix} \\ \boldsymbol{V}_e &= \mathrm{Var}(\epsilon_{ti}) = \begin{pmatrix} \sigma^2_{e_1} & \dots & \sigma_{e_{1T}} \\ \dots & \dots & \dots \\ \sigma_{e_{T1}} & \dots & \sigma^2_{e_T} \end{pmatrix}. \end{aligned} \tag{3}$$

Note that this model also implies

$$\begin{aligned} \mathrm{Cov}(y_{ti}, y_{ti'}) &= G_{ii'}\sigma^2_{at} \\ \mathrm{Cov}(y_{ti}, y_{t'i'}) &= G_{ii'}\sigma_{att'} \end{aligned}$$

where the $G_{ii'}$s are the genetic relationship coefficients for individuals i and i'. Elements of the $n \times n$ genetic relationship matrix, $\boldsymbol{G}$, can be determined from pedigree information (Lange, 2002) or estimated from GWAS data (Yang et al., 2010). This multivariate polygenic model is discussed in Korte et al. (2012) and Lee et al. (2012).

Narrow sense heritability of the tth phenotype is defined as the proportion of the additive genetic variance among the total phenotypic variance, i.e.,

$$h_t^2 = \frac{\sigma_{a_t}^2}{\sigma_{a_t}^2 + \sigma_{e_t}^2}.$$

To integrate multiple phenotypes, our goal is to find a vector of coefficients $\boldsymbol{l}$ such that $\boldsymbol{Yl}$ has the maximum heritability among all such linear combinations of the phenotypes, where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ is a $n \times T$ matrix of the collection of all T phenotypes. The heritability of any linear combination of phenotypes $\boldsymbol{Yl}$, can be expressed as the Rayleigh quotient (Horn and Johnson, 1985),

$$h_l^2 = \frac{l'\boldsymbol{V}_g l}{l'\boldsymbol{V}_p l}. \tag{4}$$

Henceforth we denote $\boldsymbol{Yl}$ with $\boldsymbol{l}$ chosen to maximize heritability as the set of MaxH phenotypes. The same optimization problem (4) has also been encountered in Fisher's linear discriminant analysis (LDA) for classification (Witten and Tibshirani, 2011). Detailed explanation for optimizing equation (4) can be found in Supplementary Material and the notes (Welling, 2005). Briefly, one needs to eigendecompose the matrix $\boldsymbol{V}_g^{\frac{1}{2}}\boldsymbol{V}_p^{-1}\boldsymbol{V}_g^{\frac{1}{2}}$ and the desired optimization solution is to find the biggest eigenvalue, i.e., maximized heritability h_l^2 and the corresponding eigenvector $\boldsymbol{w}$.

The above calculation assumes the parameters in $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ are known; in reality we need to estimate them. Historically $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ were estimated using data on pedigrees with known genetic relationships, i.e., $\boldsymbol{G}$. More recent work shows how to approximate $\boldsymbol{G}$ and estimate $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ from GWAS data on population based samples (Yang et al., 2010). With $\boldsymbol{G}$ treated as known, $(\boldsymbol{V}_g, \boldsymbol{V}_p)$ can be estimated using Maximum Likelihood (ML), Restricted ML (REML) or Method of Moments (MOM) approaches. When the sample size is large, the maximization is not trivial and the computation is costly. We used ML for the application example, and recommend that ML or REML be used in practice. For efficiency of computation, we used the much simpler MOM approach to estimate $\boldsymbol{V}_g$ and $\boldsymbol{V}_p$ in the simulations (Lange, 2002). We summarize the steps needed to compute the MaxH phenotype in the Supplementary Material.

2.2 Association Testing and Power Approximation

Thus far, we have focused on maximizing heritability in order to integrate multiple phenotypes. Now we consider testing and power for individual SNPs using MaxH phenotype. To test the hypothesis of no association for a single variant, we include a major gene effect and use the "mixed

model" (Korte et al., 2012)

$$y_{ti} = \mu_t + b_t x_{0i} + g_{ti} + \epsilon_{ti} \tag{5}$$

where x_{0i} is the standardized additive coding for the SNP we wish to test and $\boldsymbol{b} = (b_1, \ldots, b_T)$ is the vector of genetic effects for the T phenotypes. Letting $\boldsymbol{Y}_l = (y_{li}) = \boldsymbol{Y}\boldsymbol{l}$ denote the n-vector of MaxH phenotypes, for each element, we have

$$y_{li} = \boldsymbol{l}'\boldsymbol{y_i} = \mu_l + b_l x_{0i} + g_{li} + \epsilon_{li}$$

where $\boldsymbol{y}_i = (y_{1i}, y_{2i}, \ldots, y_{ti})'$ is individual i's T phenotypic measurments, $b_l = \boldsymbol{l}'\boldsymbol{b}$, $\mu_l = \boldsymbol{l}'(\mu_1, \ldots, \mu_T)$, $g_{li} = \boldsymbol{l}'\boldsymbol{g_i}$, and $\epsilon_{li} = \boldsymbol{l}'\boldsymbol{\epsilon}_i$. Hence

$$\begin{aligned} \mathrm{E}\,(y_{li}) &= \mu_l + b_l \\ \mathrm{Var}(y_{li}) &= \boldsymbol{l}'\boldsymbol{V}_p\boldsymbol{l}. \end{aligned}$$

To test $\mathrm{H}_0 : b_l = \boldsymbol{l}'\boldsymbol{b} = 0$, a Wald test is given by

$$W = \frac{\hat{b}_l}{\mathrm{SE}(\hat{b}_l)} \tag{6}$$

where $\hat{b}_l$ is the ordinary least squares (OLS) estimator of b_l and SE is its standard error under the regression model (Klei et al., 2008). In the calculation of $\mathrm{SE}(\hat{b})$ we have neglected the correlation of subjects' phenotypes generated by the polygenic background, since in a population based sample, the genetic relationships are small in practice. But the correlations are considered when generating MaxH phenotypes. Simulation example shows that the type I error rate is protected.

The power of any test to reject $\mathrm{H}_0 : b_l = \boldsymbol{l}'\boldsymbol{b} = 0$ depends not only on the test statistic, but also on how $\boldsymbol{b} = (b_1, \ldots, b_T)$ is chosen. The vector $\boldsymbol{b}$ can be chosen arbitrarily, but if the polygenic model is correct, in a GWAS setting with polygenic effects, it is natural to consider testing SNPs whose genetic effects are consistent with the polygenic model, i.e., $\boldsymbol{b} \sim c\mathcal{N}(0, \boldsymbol{V}_g)$, where c is a scale parameter chosen to determine the heritability of the major gene effect. When including a major gene effect, the overall genetic variance of a linear combination becomes $b_l^2 + \boldsymbol{l}'\,\mathrm{Var}(g_i)\boldsymbol{l}$. In order to maintain a fixed overall heritability (Equation (4)), we choose the major gene effect to satisfy, $b_l^2 = c^2\boldsymbol{l}'\boldsymbol{V}_g\boldsymbol{l}$, where $\boldsymbol{l}'\boldsymbol{V}_g\boldsymbol{l}$ is again the total genetic variance including the major gene effect; this implies that b_l explains a fraction c^2 of the total heritability.

The Wald test statistic W^2 in equation (6) follows a chi-square distribution with 1 degree of freedom, i.e., $\chi^2(\delta, 1)$ with non-centrality parameter (NCP)

$$\delta^2 = n\frac{c^2 h_l^2}{1 - c^2 h_l^2}. \tag{7}$$

As heritability h_l^2 increases, the NCP and the power of the test increases, as does the asymptotic power. Power gain is heavily dependent on the gain of heritability. For the MaxH phenotype, the structure of the genotypic and phenotypic variance-covariance matrix and the number of phenotypes combined determines the heritability. In practice $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ must be estimated, and sampling error may decrease power if too many phenotypes are added. This is considered later, as well as when $\boldsymbol{b}$ comes from arbitrary distributions.

3 Results

3.1 Combining Phenotypes with V_g and V_p known

First we consider combining the simple case of two phenotypes with equal heritabilities, which are standardized with mean zero and variance one. The genotypic and phenotypic variance-covariance matrices take the form,

$$\boldsymbol{V}_g = h^2 \begin{pmatrix} 1 & r_g \\ r_g & 1 \end{pmatrix} \quad \boldsymbol{V}_p = \begin{pmatrix} 1 & r_p \\ r_p & 1 \end{pmatrix}$$

where r_g and r_p are the genotypic and phenotypic correlation coefficients. Note that the phenotypic variance components are partitioned into genetic and environmental components, i.e., $\boldsymbol{V}_p = \boldsymbol{V}_g + \boldsymbol{V}_e$, thus $r_p = r_g h^2 + r_e(1 - h^2)$, where r_e is environmental correlation coefficient. Since $-1 \leq r_e \leq 1$, it follows that the genotypic and phenotypic correlation coefficients have the constraints,

$$r_p \geq h^2 - 1 + h^2 r_g \quad \text{and} \quad r_p \leq h^2 r_g + 1 - h^2. \tag{8}$$

To maximize equation (4), the eigensystem equation S(2) in the Supplementary Material has eigenvectors, $(1, 1)$ and $(1, -1)$, with eigenvalues $\frac{1+r_g}{1+r_p}h^2$ and $\frac{1-r_g}{1-r_p}h^2$. MaxH is obtained by picking the largest eigenvalue and corresponding eigenvector, subject to the constraint in (8), which also guarantees that the maximized heritability h_l^2 is bounded in $(0, 1)$.

In this simple example where the heritability of two phenotypes are the same, the eigenvectors of the PCA approach are the same as the MaxH approach and the combined phenotypes are $\boldsymbol{Y}_1 + \boldsymbol{Y}_2$ and $\boldsymbol{Y}_1 - \boldsymbol{Y}_2$, but with different eigenvalues. The eigenvalues of the two PCs are $1 \pm r_p$. When $r_g > r_p > 0$, both MaxH and the first PC phenotypes are $\boldsymbol{Y}_1 + \boldsymbol{Y}_2$. However, when $r_p > r_g > 0$, MaxH takes the combined phenotype which maximizes the heritability, i.e., $\boldsymbol{Y}_1 - \boldsymbol{Y}_2$, while PCA takes the combined phenotype which maximizes the phenotypic variance, i.e., $\boldsymbol{Y}_1 + \boldsymbol{Y}_2$. The selection of the maximal PC depends only on the sign of r_p. Thus the first PC from PCA approach is always $\boldsymbol{Y}_1 + \boldsymbol{Y}_2$ for $r_p > 0$, but it is $\boldsymbol{Y}_1 - \boldsymbol{Y}_2$ from MaxH approach when $r_p > r_g > 0$. Aschard et al. (2014)

obtained a similar result using a slightly different model. For $T = 2$, their model is equivalent to ours with the major gene x_0, except that the polygenic component g_i is omitted and the residual variance covariance matrix has positive covariance ν. The single major gene effect explains all of the heritability, and as a result $r_g = +1$ if b_1 and b_2 have the same sign or $r_g = -1$ otherwise. We consider a range of possible r_g indicating a range of pleiotropy, based on which, we integrate phenotypes that maximize heritability.

The increase of maximized heritability represents an increase in power. Figure 1 shows the maximized heritability as a function of r_p and r_g. To develop intuition for how the MaxH and its heritability behaves, we consider the two extremes of pleiotropy. When r_g equals zero, i.e., there is no correlation between the coefficients of the genetic effects at the causal loci, and no evidence for average pleiotropy. In this case, the phenotypic correlation is proportional to the residual correlation. If $r_p > 0$ the maximized heritability is $h^2/(1-r_p)$ and the MaxH phenotype is $\boldsymbol{Y}_1 - \boldsymbol{Y}_2$. Conversely, the first PC takes $\boldsymbol{Y}_1 + \boldsymbol{Y}_2$. Intuitively we see that the first PC maximizes the phenotypic (residual) covariance of the linear combination, while MaxH minimizes the residual effects. A more specific example is, when $r_g = 0$, genetic component of the first phenotype is positive (non-zero), genetic component of the second phenotype is zero, and environmental correlation is positive (i.e. $r_p > 0$), MaxH (the difference of the two single phenotypes) will not enhance the genetic signal, rather reduce the residual variances. In the absence of any information about genetic effects at a particular SNP, the phenotype with the smallest residual variance will be the best phenotype. So MaxH will do better than PC. Now consider the other extreme where $|r_g|$ approaches 1, i.e., the genetic effects for one phenotype predict perfectly the genetic effects for the second. In this case, MaxH chooses the linear combination which maximizes the variance of the combined genetic effects. The first PC continues to maximize the total phenotypic variance, and agrees with the MaxH choice when $|r_g|$ approaches 1, because most of the phenotypic covariance is in the genetic, not the residual component. Note that when $r_p = r_g$, either linear combination gives the same heritability as a single phenotype, and is also equivalent to PC. When r_p and r_g have the same magnitude, but different signs, we can expect MaxH to do much better than the case when $r_p = r_g$.

When combining more than two phenotypes, we extend the above design where pairwise correlations are the same, both phenotypic and genotypic, but heritabilities differ, i.e.,

$$\boldsymbol{V}_g = h^2 \begin{pmatrix} 1 & kr_g & \dots & k^t r_g \\ kr_g & k^2 & \dots & k^{t+1} r_g \\ \dots & \dots & \dots & \dots \\ k^{t+1} r_g & \dots & k^{2(t-1)} & k^{2t-1} r_g \\ k^t r_g & \dots & k^{2t-1} r_g & k^{2t} \end{pmatrix}, \quad \boldsymbol{V}_p = \begin{pmatrix} 1 & r_p & \dots & r_p \\ r_p & 1 & \dots & r_p \\ \dots & \dots & \dots & \dots \\ r_p & \dots & 1 & r_p \\ r_p & \dots & r_p & 1 \end{pmatrix},$$

where $0 < k \leq 1$. For simplicity, we consider the situation when genetic and phenotypic correlations are both positive. In Figure 2, we show the maximized heritability as a function of the number of phenotypes combined while varying the value of r_g, r_p, and k. For all four cases, we set $h^2 = 0.4$. In Figure 2a and Figure 2b we set $k = 1$, i.e., all combined phenotypes have the same heritability as 40%. When $r_g > r_p = 0.4$ (Figure 2a), both approaches behave the same and the heritability increases as the number of phenotypes combined increases. In Figure 2c and Figure 2d we vary $k\,(0 < k < 1)$ so that phenotypes with lower heritabilities are added in. Both figures (Figure 2c and Figure 2d) show PC loses heritability when adding phenotypes with lower heritabilities. This pattern exists even when heritabilities of phenotypes combined are fixed (Figure 2b). When $r_g < r_p$, adding more phenotypes with lower heritabilities can increase the heritability of MaxH more dramatically than combing phenotypes when $r_g > r_p$, and combining more than two phenotypes does not provide a noticeable advantage for MaxH.

3.2 Testing a Single Locus in the Presence of Polygenic Variance

Here we estimate power for three settings when combining two phenotypes with the same heritability ($h^2 = 0.4$). First we assume $\boldsymbol{V}_g$ and $\boldsymbol{V}_p$ are known for the purpose of calculating the MaxH phenotype, then we relax that assumption. The test statistic of association and its standard error are calculated as in Section 2.2. We consider three cases, a) $r_g > r_p$ (i.e., $r_g = 0.9$, $r_p = 0.4$); b) $r_g < r_p$ (i.e., $r_g = 0.7$, $r_p = 0.8$); and c) $r_g < r_p$ (i.e., $r_g = 0.1$, $r_p = 0.5$). We simulate phenotypes based on polygenic model (1) and (3). Genotypes are taken from genome-wide SNP data of COPDGene cohort of Non-Hispanic White (NHW) population. Only SNPs ($51,428$ SNPs in total) from Chromosome 1 were used for simplicity.

Our purpose is to show that power increase is determined by the increase of the maximized heritability (Equation (7) and Figure 1), and that the magnitude of the heritability increase is a surrogate of power increase. Phenotypes were simulated based on the linear model (1). One hundred SNPs on Chromosome 1 were randomly chosen as the causal SNPs for polygenic background. Five hundred replicates, each with 3000 individuals and $T = 2$ were simulated. Our approach was then compared to the single trait association analysis and the PC approach. We consider testing only one of the 100 causal SNPs with effects chosen as described in Section 2.2 with $c = 2\%$. A different causal SNP is selected for each of the 500 replicates. Thus we compute average power for a SNP explaining 2% of the heritability. After generating the single phenotypes, MaxH phenotype and the first PC, we assess empirical type I error rate through testing all the SNPs on chromosome 2 from COPDGene NHW population which has no causal SNPs. The estimated type I error rate is

well maintained (0.048) at the significant level of 0.05.

The results are shown in Table 1. The heritability of each phenotype is 0.4, and the maximized heritabilities predicted from our theory are given in lower panel of Table 1. As predicted from our previous results in the Section 3.1, MaxH and the first PC give nearly identical results when r_g is large because they select identical linear combinations. Even a modest reduction of r_g to 0.7 with an increase of r_p to 0.8 shows substantial impact on the relative power of MaxH and PC, with MaxH doing better. For the third case, i.e., lower pleiotropy, the power of MaxH is even higher, while PC does worse than single phenotype case. The ordering of the power of MaxH in the three scenarios can be predicted from the order of the MaxH's heritability. The loss of power due to estimating r_g is negligible for case a) and c), and about 5% for case b). This is likely due to the fact that r_g and r_p can be estimated well enough to choose the correct linear combination. The estimates of all the heritabilities tend to be lower than predicted by less than 10% (Table 1). The order is preserved. As we might expect, the power loss for PC is negligible when estimating the variance components, as it does not rely on the decomposition of $\boldsymbol{V}_p$ into genetic and residual components. For other values of h^2 and T, the plots such as Figures 1 and 2 can be used as guidance for choosing which phenotypes to combine once r_g and r_p are estimated. Further studies are needed to determine loss of power for larger T and smaller n.

3.3 Empirical Power for Testing Small Effects

The 100 previously chosen SNPs from Chromosome 1 are used here as causal SNPs with each SNP explaining 1% of the total heritability. SNPs effects are generated from a bivariate normal distribution with mean zero and variance $\boldsymbol{V}_g$. Simulations are performed for a range of r_p and r_g. Five hundred pairs of the phenotypes are simulated and tested against each of the 100 SNPs. We use the same strategy to estimate type I error rate by using all the SNPs from chromosome 2 and all the MaxH phenotypes. Our empirical type I error is well maintained at the significant level of 5×10^{-4} (i.e., 4.9×10^{-4}).

We compare several methods based on the proportion of the 100 causal SNPs that have power over 80%, shown as heat maps in Figure 3 (univariate analysis) and Figure S2 (multivariate analysis). Figure 3 shows the results for MaxH and PCA. It also shows the association analysis using original single phenotypes adjusting for multiple testing. The MaxH approach generally performs the best among univariate association analysis. When $r_g = r_p$, MaxH perform poorly which is consistent with the pattern of heritability maximization (Figure 1). With certain configurations of genetic and phenotypic correlations, the MaxH method can do as well as using multivariate

phenotypes (Figure S2). Note that one could also perform a multivariate analysis using multiple phenotypes generated from our method, but it is equivalent to using original multiple phenotypes or generated from PCA method (see Discussion). We consider situation when $r_g = 0.7$ and $r_p = 0.8$ to examine the relation between effect sizes and power (Figure 4). In Figure 4, we plot the effect sizes of the 100 causal SNPs. The power for such SNPs is shown in gray scale. The pattern of black dots show that using a single phenotype ($\boldsymbol{Y}_1$ or $\boldsymbol{Y}_2$) for testing, power is the best for the loci which have the biggest effect sizes ($|b_1|$ or $|b_2|$) for the corresponding phenotypes. Using PC approach, only the loci whose effects are large on both phenotypes have good power, i.e., intersection of the black points in the bottom two plots. However, using the MaxH phenotype, the set of loci having good power is when the effect sizes follows the global genetic distribution. Especially when both $|b_1|$ and $|b_2|$ are small and have opposite sign, MaxH is the only method that reveal them with very high power. However MaxH performs poorly along the diagonal stripe, i.e., when $b_1 = b_2$, no matter the magnitude of $|b_1|$ or $|b_2|$. Using our MaxH method, 40% of the 100 loci have power over 80%. Only about 20% of the SNPs have power over 80% when using PC and single phenotypes.

Although the fixed effects b_1 and b_2 are obtained from $\mathcal{N}(0, c\boldsymbol{V}_g)$ where $c = 0.01$, they cover a broad region from -0.15 to 0.15. Our assumption about pleiotropy is that the effects of the polygenic components are drawn from a multivariate normal distribution with mean zero and variance covariance matrix $\boldsymbol{V}_g$. This does not imply equal pleiotropy for all SNPs unless $\boldsymbol{V}_g$ has rank one. This is illustrated in Figure 4 where we plot the genetic effects for a set if 100 SNPs drawn from the polygenic distribution with mean zero, variances 0.4 and correlation of 0.7. As this figure illustrates, the extent of pleiotropy differs considerably among the 100 SNPs, even though r_g is relatively high. It is natural to ask what would the power be for major SNP effects which are not drawn from this distribution, i.e., effects in the upper left and lower right corner. Intuitively we would expect that the genetic effects on the diagonal corners would be easier to detect since they are further from the origin, and this is indeed the case. Supplementary Figure S3 illustrates this point by drawing SNPs from a uniform distribution on the plane. The superiority of MaxH over PCA is clear (Figure S3).

3.4 GWAS Analysis in COPDGene NWH Population

We apply our method to COPDGene, a large case-control sample of well-characterized smokers from a genome-wide association study of respiratory disease. It includes 10,192 non-hispanic white (NHW) and African American (AA) current and former smokers with airflow obstruction ranging from none to GOLD stage 4 (very severe) COPD. The study design of COPDGene has been reported

previously (Regan et al., 2010). Briefly the subjects are included between the ages of 45 and 80 with at least a 10 pack-year smoking history. Exclusion criteria includes pregnancy, history of other lung disease except asthma, prior lobectomy or lung volume reduction surgery, active cancer undergoing treatment, or known or suspected lung cancer. We restrict our analysis to the NHW population, which includes 6678 individuals after data cleaning and exclusions. Details concerning genotyping, quality control, and imputation are posted on the COPDGene website (http: //www.copdgene.org).

We exclude SNPs that have MAF$<$ 0.01 and Hardy-Weinberg Equilibrium (HWE) p-value$<$ 10^{-8} using PLINK (Purcell et al., 2007). Only those SNPs on the autosomes are used for heritability estimation by the software package Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011a) (Table 2). Spirometry measures of lung function are performed before and after the inhalation of 180mcg (2 puffs) of albuterol. Pulmonary function measurements are collected according to the American Thoracic Society guidelines (Miller et al., 2005). Percent predicted values for FEV_1 are calculated using equations of Hankinson and colleagues (Hankinson et al., 1999). FEV_1 and FEV_1/FVC, both measurements of lung function, are used to diagnose and gauge severity of disease. Volumetric chest CT acquisitions are obtained at full inspiration (200 mAs), and at the end of normal expiration (50 mAs). Quantitative image analysis to calculate percent emphysema is performed using 3D SLICER (http://www.slicer.org/). Percent emphysema, i.e., lung destructions that can lead to decreased lung function, is estimated from using the percent below -950HU on chest CT scans.

We consider one representative example of combining three major endophenotypes of COPD: FEV_1 (post bronchodilator), FEV_1/FVC and percent of Emphysema (Table 2). From Table 2 we can see that this is not a scenario where we expect MaxH to do very well; h_l^2 is barely bigger than h^2 for FEV_1, and the $|r_p - r_g|$ are all small. Results using only FEV_1 and FEV_1/FVC are qualitatively similar (not shown). Linear regression analyses of each individual phenotype and the combined phenotypes were adjusted for age, gender, height, pack-years, and the first five genetic ancestry variables estimated by the software EIGENSTRAT (Price et al., 2006). The standardized residuals for FEV_1, FEV_1/FVC, and log-transformed emphysema are used for analysis. Univariate genome-wide association analyses are performed using PLINK (Purcell et al., 2007) and multivariate analyses are performed using the Mendel software (Lange et al., 2013). Very few SNPs reached genome-wide significant level of 5×10^{-8}. For illustration, SNPs passing the threshold 5×10^{-7} and the corresponding gene information are shown in Table 3. Detailed Manhattan plots are shown in the Supplementary Figure S4. All results are adjusted for genomic control factor (in addition to

first five genetic ancestry variables estimated using principal components).

Table 3 reports the significant results from PC and MaxH as well as multivariate regression, and Multiphen (O'Reilly et al., 2012). Full genome-wide association results for the individual phenotypes are presented in separate publications (Lutz et al and Cho et al, in preparation). SNPs in three loci, *FAM13A* (Chr 4) (Cho et al., 2010), *HHIP* (Chr 4) (Pillai et al., 2009), and *CHRNA3/CHRNA5/AGPHD1* (Chr 15) (Hardin et al., 2012; Lambrechts et al., 2010; Pillai et al., 2009) have been previously reported, and well-replicated, as associated with COPD disease status. SNPs at all of these loci are associated with MaxH, but PC, multivariate regression, and Multiphen test failed to detect the *FAM13A* region. Multiphen also fail to detect *HHIP*. All four methods confirmed the loci on Chr 15. Three other loci, *TGFB2* (Chr 1) (Soler Artigas et al., 2011), *AGER* (Chr6) (Hancock et al., 2010; Repapi et al., 2010), and *MMP12* (Chr 11) (Hunninghake et al., 2009; Korytina et al., 2008) have previously shown weaker association results in COPD GWAS. PC and MaxH found the SNP at *MMP12* significant, but the multivariate regression and Multiphen do not. All methods in Table 3 except MaxH find *AGER* significant. Only the multivariate method find *TGFB2*. The final locus *PTPRM*, found only by the multivariate method, has not previously been reported and is of uncertain validity. Although MaxH does not find the most loci (4 versus 6 for multivariate regression), it is the only approach to find all of the confirmed loci. Further we judge its performance better than PC because PC failed to find *FAM13A*.

4 Discussion

In order to discover novel genetic disease variants, multiple correlated phenotypes are frequently used in genetic association studies with the goal of improving power. One strategy uses a linear combination of the traits. The first PC derived trait is the linear combination of individual traits that accounts for the maximum phenotypic variance. In this paper, we propose an alternate dimension reduction scheme, i.e., a linear combination of the phenotypes that maximizes the heritability (MaxH) of any linear combination of the traits. In contrast to the first PC, the maximized heritability of this linear combination translates into improved power for association testing, because the coefficients are chosen to maximize the genetic variance while minimizing the residual variance. We compare several univariate and multivariate methods using both simulated and real data. We also show that a multivariate approach using all T phenotypes has better power than either univariate approach, first PC or MaxH, but depending on the parameters using a smaller subset of traits may do almost as well. Aschard et al. (2014) extends the single PC approach by including multiple PCs

of the phenotypic matrix in a multivariate regression, and shows that using all T PCs is equivalent to multivariate regression using the original T traits. It is easy to see that using all the MaxH PCs in a multivariate analysis is essentially equivalent to the multivariate analysis using the original traits because both of the PC approaches are full rank linear transformations of the $\boldsymbol{Y}$ (assuming $\boldsymbol{V}_g$ and $\boldsymbol{V}_p$ are both of full rank), and a multivariate analysis is invariant to linear transformations. However multivariate regression is usually computationally intensive and the power gain compared to other approaches depends upon unknown effects and assumptions (Korte et al., 2012; Schifano et al., 2013). In fact in a simulation study of Suo et al. (2013), multivariate analysis of analysis of variance (MANOVA) performs the worst compared to PCA and single phenotype approach.

We approximate power analytically as a simple function of the maximized heritability, given the model parameters. The improvement in maximizing heritability relative to individual trait heritability depends on the configuration of the phenotypic and genotypic correlation coefficients r_p and r_g respectively, between pairs of phenotypes. Given a data set of multiple phenotypes and SNPs from a GWAS platform, one can straightforwardly estimate the necessary parameters, $\boldsymbol{V}_g$ and $\boldsymbol{V}_p$, in order to calculate maximized heritability for any subset of the T phenotypes. When r_p and r_g are fixed and estimated for the full set of T phenotypes, by definition the maximized heritability always occurs when using the full set of T phenotypes.

Our theory assumes that the SNP effects being tested are consistent with the polygenic model. This assumption makes power calculations easy, but of course, it may not be correct. However, when $\boldsymbol{V}_g$ and $\boldsymbol{V}_p$ are estimated we assume no major gene effects, only zero mean polygenic effects. If there are major gene effects for any trait, they should make a major contribution to the estimated $\boldsymbol{V}_g$, thus enhancing the power of MaxH. This point is illustrated in Figure S3 which depicts testing polygenic effects which are not selected from the assumed polygenic distribution. Figure S3 shows that MaxH has good power when testing SNPs effects with very different pleiotropy. This is because the causal SNPs are assumed to have zero means and the the sparse areas in Figure S3 tend to be further from the origin than the many of the causal SNPs. The relationships between r_p and r_g and the individual phenotypic heritabilities can be used to infer which combined phenotypes will give larger maximized heritabilities. Our data example illustrates that even if the maximized heritability is only slightly higher than individual trait heritability, MaxH can still do well at picking up established loci. MaxH is the best way to identify SNPs associated with at least one phenotype. If a significant SNP is identified using MaxH, one should use other methods, e.g. Stephens (2013), to look for direct or indirect effects and to determine which phenotypes are directly associated.

Our method requires the estimation of the parameters once, then the combined phenotype can be used as a single trait in the standard GWAS analysis. The computational cost is relatively the same as the standard GWAS analysis. In practice, combining too many phenotypes may hurt the heritability and power, as the variance matrices that have to be estimated become too large. Large sample sizes are needed in order to accurately estimate $\boldsymbol{V}_g$ and to find the correct linear combination. In real data analysis, population substructure and environmental factors can inflate the estimation of $\boldsymbol{V}_g$ (Browning and Browning, 2011). For COPDGene data example, we employ strict QC that were suggested by Yang et al. (2011a) to minimize the potential inflation. Detailed discussions can be found in paper Zhou et al. (2013). Specifically, the proportion of estimated heritability attributed to population substructure across the whole genome is less than 1%. In controlling the effects of population substructure for association testing, we use both PCs calculated by EIGENSTRAT (Price et al., 2006) and genomic inflation factor (Yang et al., 2011b) to adjust phenotypes and test statistics. EIGENSTRAT generates PCs using only the information from genetic relationship matrix. For MaxH, we use both phenotypic and genetic relationship matrix to generate PCs and estimate MaxH phenotype. There might be more potential for bias. However, the PC's from EIGENSTRAT are PCs of genetic relationship matrix, which are different from the PC's of the heritability matrix. They therefore are still valid to be used in MaxH setting for population substructure adjustment.

Acknowledgments*

Appendix 1: NIH Grant Support and Disclaimer

The project described was supported by Award Number U01HL089897 and Award Number U01HL089856 from the National Heart, Lung, And Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, And Blood Institute or the National Institutes of Health.

Appendix 2: COPD Foundation Funding

The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens and Sunovion.

* Full list of COPD Investigators unit core and clinical centers are included in the supplementary material

References

Aschard H, Vilhjálmsson BJ, Greliche N, Morange PE, Trégouët DA, et al. (2014) Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies. Am J Hum Genet 94: 662–76.

Avery CL, He Q, North KE, Ambite JL, Boerwinkle E, et al. (2011) A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. PLoS Genet 7: e1002322.

Browning SR, Browning BL (2011) Population structure can inflate SNP-based heritability estimates. Am J Hum Genet 89: 191–3; author reply 193–5.

Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, et al. (2010) Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nat Genet 42: 200–2.

Falconer DS, et al. (1981) Introduction to quantitative genetics. Ed. 2, Longman.

Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, et al. (2010) Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet 42: 45–52.

Hankinson JL, Odencrantz JR, Fedan KB (1999) Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med 159: 179–87.

Hardin M, Zielinski J, Wan ES, Hersh CP, Castaldi PJ, et al. (2012) CHRNA3/5, IREB2, and ADCY2 are associated with severe chronic obstructive pulmonary disease in Poland. Am J Respir Cell Mol Biol 47: 203–8.

Horn RA, Johnson CR (1985) Matrix analysis. Cambridge University Press.

Hunninghake GM, Cho MH, Tesfaigzi Y, Soto-Quiros ME, Avila L, et al. (2009) MMP12, lung function, and COPD in high-risk populations. N Engl J Med 361: 2599–608.

Karasik D, Cupples LA, Hannan MT, Kiel DP (2004) Genome screen for a combined bone phenotype using principal component analysis: the Framingham study. Bone 34: 547–56.

Klei L, Luca D, Devlin B, Roeder K (2008) Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol 32: 9–19.

476 Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, et al. (2012) A mixed-model approach for
477 genome-wide association studies of correlated traits in structured populations. Nat Genet 44:
478 1066–71.

479 Korytina GF, Akhmadishina LZ, Ianbaeva DG, Viktorova TV (2008) Polymorphism in promoter
480 regions of matrix metalloproteinases (MMP1, MMP9, and MMP12) in chronic obstructive pul-
481 monary disease patients. Genetika 44: 242–9.

482 Lambrechts D, Buysschaert I, Zanen P, Coolen J, Lays N, et al. (2010) The 15q24/25 susceptibility
483 variant for lung cancer and chronic obstructive pulmonary disease is associated with emphysema.
484 Am J Respir Crit Care Med 181: 486–93.

485 Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, et al. (2004) A family-based association
486 test for repeatedly measured quantitative traits adjusting for unknown environmental and/or
487 polygenic effects. Stat Appl Genet Mol Biol 3: Article17.

488 Lange K (2002) Mathematical and Statistical Methods for Genetic Analysis. Springer.

489 Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, et al. (2013) Mendel: the Swiss army
490 knife of genetic analysis programs. Bioinformatics 29: 1568–70.

491 Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between
492 complex diseases using single-nucleotide polymorphism-derived genomic relationships and re-
493 stricted maximum likelihood. Bioinformatics 28: 2540–2.

494 Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, et al. (2005) Standardisation of
495 spirometry. Eur Respir J 26: 319–38.

496 O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, et al. (2012) MultiPhen: joint model
497 of multiple phenotypes can increase discovery in GWAS. PLoS One 7: e34861.

498 Ott J, Rabinowitz D (1999) A principal-components approach based on heritability for combining
499 phenotype information. Hum Hered 49: 106–11.

500 Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, et al. (2009) A genome-wide association study in
501 chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci.
502 PLoS Genet 5: e1000421.

503 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components
504 analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.

505 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for
506 whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–75.

507 Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, et al. (2010) Genetic epidemiology of
508 COPD (COPDGene) study design. COPD 7: 32–43.

509 Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, et al. (2010) Genome-wide association study
510 identifies five loci associated with lung function. Nat Genet 42: 36–44.

511 Schifano ED, Li L, Christiani DC, Lin X (2013) Genome-wide association analysis for multiple
512 continuous secondary phenotypes. Am J Hum Genet 92: 744–59.

513 Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, et al. (2011) Genome-wide association
514 and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet 43: 1082–90.

515 Stephens M (2013) A unified framework for association analysis with multiple related phenotypes.
516 PLoS One 8: e65245.

517 Suo C, Toulopoulou T, Bramon E, Walshe M, Picchioni M, et al. (2013) Analysis of multiple
518 phenotypes in genome-wide genetic mapping studies. BMC bioinformatics 14: 151.

519 Welling M (2005) Fisher linear discriminant analysis. Department of Computer Science, University
520 of Toronto .

521 Witten DM, Tibshirani R (2011) Penalized classification using Fisher's linear discriminant. J R
522 Stat Soc Series B Stat Methodol 73: 753–772.

523 Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a
524 large proportion of the heritability for human height. Nat Genet 42: 565–9.

525 Yang J, Lee SH, Goddard ME, Visscher PM (2011a) GCTA: a tool for genome-wide complex trait
526 analysis. Am J Hum Genet 88: 76–82.

527 Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, et al. (2011b) Genomic inflation factors under
528 polygenic inheritance. Eur J Hum Genet 19: 807–12.

529 Zhou JJ, Cho MH, Castaldi PJ, Hersh CP, Silverman EK, et al. (2013) Heritability of chronic
530 obstructive pulmonary disease and related phenotypes in smokers. Am J Respir Crit Care Med
531 188: 941–7.

Legends

Figure 1: Maximized heritability as a function of genotypic and phenotypic correlation.

Figure 2: Maximized heritability as a function of the number of phenotypes. Left two plots show the cases when $r_g > r_p = 0.4$; right two plots show the cases when $r_g < r_p = 0.8$. Upper two plots show the situation when the combine phenotypes have the same heritability ($h^2 = 0.4$ and $k = 1$) while fixing r_p and varying r_g. The lower two show the situation when heritabilities of combined phenotypes drop as a factor of k ($h^2 = 0.4$ and $k < 1$) while fixing both r_g and r_p.

Figure 3: Proportion of 100 SNPs with empirical power greater than 0.8 as a function of r_g and r_p using phenotype of first PC from MaxH and PCA method. *Association analysis was performed using both single phenotypes and used Bonferroni correction to adjusted for extra tests, i.e., 2.5×10^{-4}.

Figure 4: 100 SNPs' empirical power as a function of effects sizes of both traits, when $r_p = 0.8$ and $r_g = 0.7$. Gray scale represents the scale of power, the darker the higher power.

$h^2 = 0.4,\ k = 1$

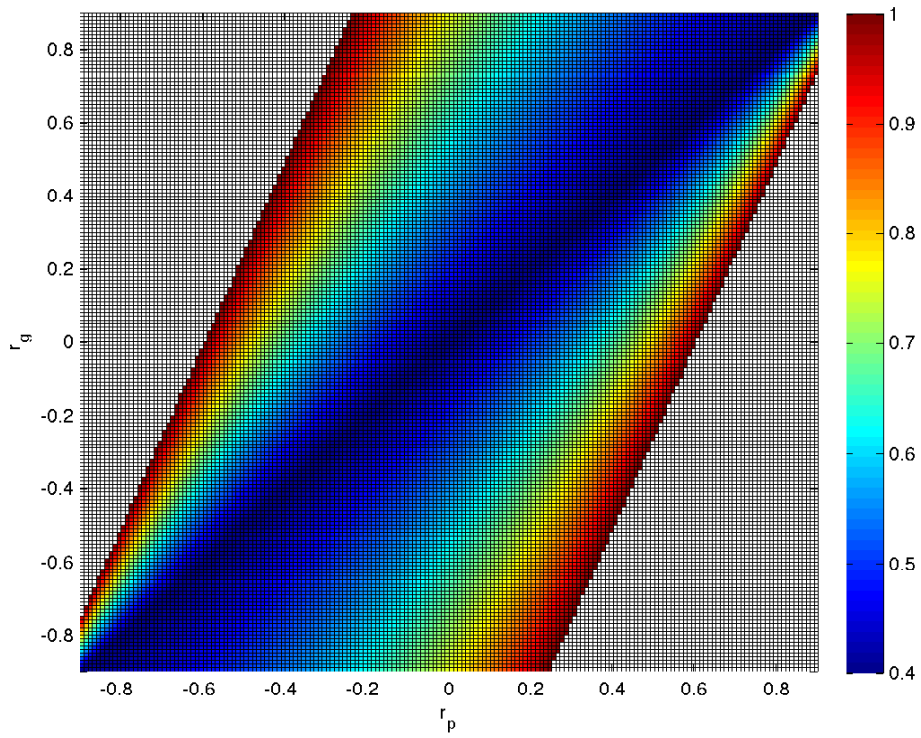


Figure 1: Maximized heritability as a function of genotypic and phenotypic correlation.

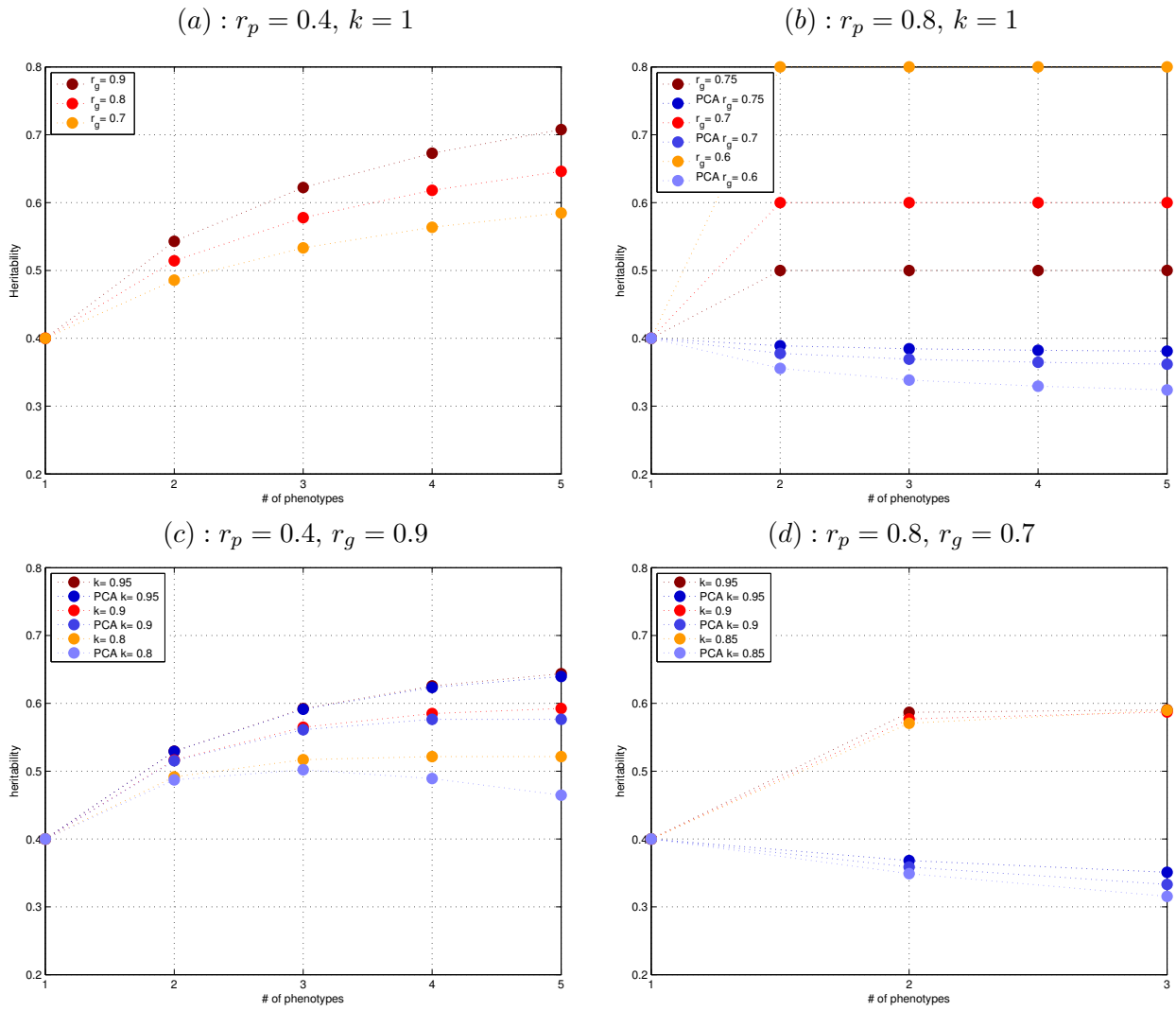
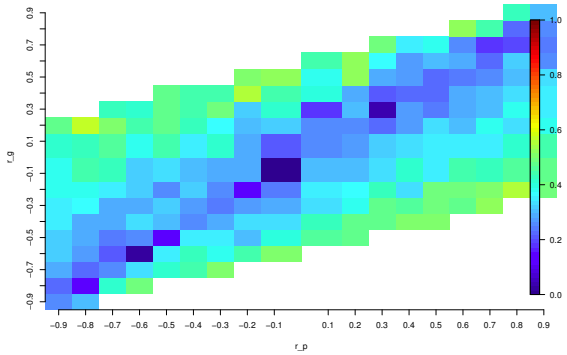
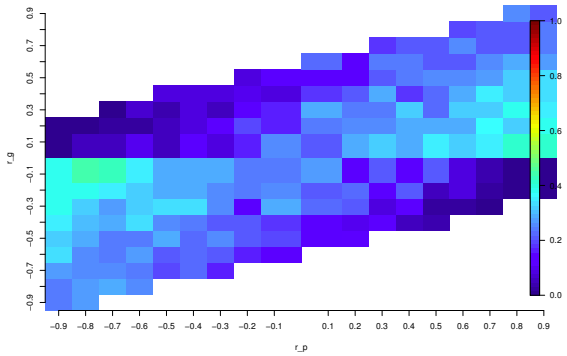


Figure 2: Maximized heritability as a function of the number of phenotypes. Left two plots show the cases when $r_g > r_p = 0.4$; right two plots show the cases when $r_g < r_p = 0.8$. Upper two plots show the situation when the combine phenotypes have the same heritability ($h^2 = 0.4$ and $k = 1$) while fixing r_p and varying r_g. The lower two show the situation when heritabilities of combined phenotypes drop as a factor of k ($h^2 = 0.4$ and $k < 1$) while fixing both r_g and r_p.

MaxH



PCA



Single Phenotype*

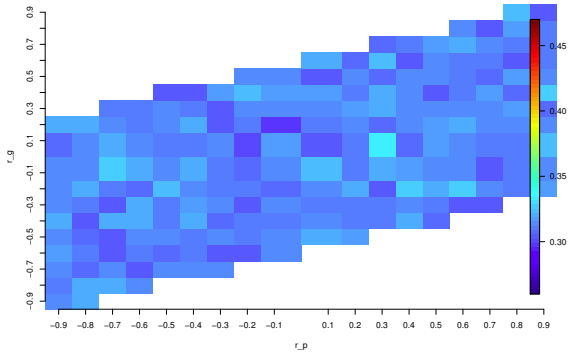


Figure 3: Proportion of 100 SNPs with empirical power greater than 0.8 as a function of r_g and r_p using phenotype of first PC from MaxH and PCA method. *Association analysis was performed using both single phenotypes and used Bonferroni correction to adjusted for extra tests, i.e., 2.5×10^{-4}.

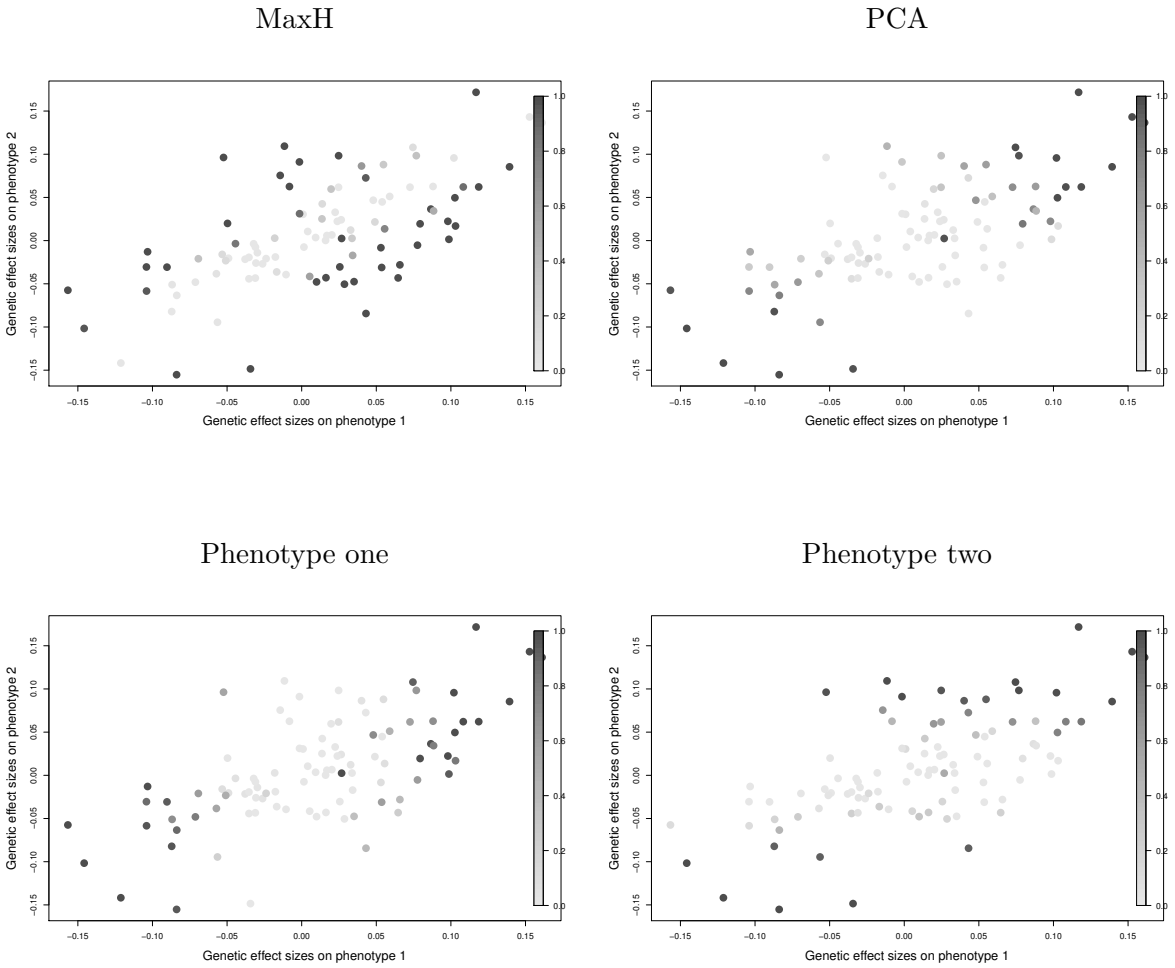


Figure 4: 100 SNPs' empirical power as a function of effects sizes of both traits, when $r_p = 0.8$ and $r_g = 0.7$. Gray scale represents the scale of power, the darker the higher power.

Table 1: Empirical power for a single major locus in the presence of polygenic variance are shown when using MaxH, PC phenotypes, and two single phenotypes (upper panel). Estimated and predicated MaxH phenotype's heritabilities are shown in the lower panel. Both empirical power and estimated heritabilities are assessed when r_g and r_p are known and when r_g and r_p are unknown.

	a		b		c	
	$r_g = 0.9,\ r_p = 0.4$		$r_g = 0.7,\ r_p = 0.8$		$r_g = 0.1,\ r_p = 0.5$	
	Power					
r_p and r_g Known	MaxH	0.716	MaxH	0.780	MaxH	0.796
	PCA	0.716	PCA	0.692	PCA	0.652
r_p and r_g Estimated	MaxH	0.706	MaxH	0.748	MaxH	0.792
	PCA	0.706	PCA	0.704	PCA	0.644
Single Trait	Trait 1	0.630	Trait 1	0.664	Trait 1	0.664
	Trait 2	0.638	Trait 2	0.672	Trait 2	0.668
	MaxH Heritablity					
Predicted	0.54		0.60		0.72	
r_p and r_g Known	0.506(0.049)		0.566(0.113)		0.682(0.035)	
r_p and r_g Estimated	0.509(0.049)		0.573(0.110)		0.676(0.035)	

Table 2: Heritability estimates are listed on the diagonal. Phenotypic r_p (upper diagonal) and genotypic r_g (lower diagonal) correlations are listed on the off-diagonal. (MaxH = -0.892FEV_1-$0.349\text{FEV}_1/\text{FVC}+0.283\log(\text{pctEmph})$; PCA=$-0.583\text{FEV}_1$-$0.631\text{FEV}_1/\text{FVC}$ $+0.511\log(\text{pctEmph})$)

	FEV_1	FEV_1/FVC	$\log(\text{pctEmph})$	MaxH	PCA
FEV_1	0.383	0.837	-0.440	-	-
FEV_1/FVC	0.882	0.372	-0.637	-	-
$\log(\text{pctEmph})$	-0.623	-0.814	0.283	-	-
MaxH	-	-		0.395	-
PCA	-	-		-	0.390

Table 3: The number of SNPs in COPDGene sample passing genome-wide significant level 5×10^{-7} by different methods and the minimum and maximum $-log_{10}$(p-value) when using FEV_1, FEV_1/FVC, and log(pctEmph).

Chr	Nearest Gene		MaxH		PCA		Multivariate		MultiPhen	
			Min	Max	Min	Max	Min	Max	Min	Max
1	**TGFB2*	Num					2			
		$-\log_{10}$(P)					6.5	6.7		
4	#*FAM13A*	Num	2							
		$-\log_{10}$(P)	6.3	6.4						
4	#*HHIP*	Num	6		7		5			
		$-\log_{10}$(P)	6.4	7.8	6.4	8.3	7.9	8.3		
6	**AGER*	Num			1		1		1	
		$-\log_{10}$(P)			7.6	7.6	6.8	6.8	6.8	6.8
11	**MMP12*	Num	1		1					
		$-\log_{10}$(P)	6.4	6.4	6.3	6.3				
15	#*CHRNA3-5* *AGPHD1* *IREB2*	Num	13		15		13		9	
		$-\log_{10}$(P)	6.3	10.9	6.5	9.0	6.5	11.3	6.3	8.3
18	*PTPRM*	Num					1			
		$-\log_{10}$(P)					6.6	6.6		

confirmed in prior COPD GWAS
* supportive evidence from other studies
See text.