Genetic Epidemiology 32: 9–19 (2008)

Pleiotropy and Principal Components of Heritability Combine to Increase Power for Association Analysis

Lambertus Klei,[1] Diana Luca,[2] B. Devlin,[1] and Kathryn Roeder[2]*

[1]*Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania*
[2]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*

When many correlated traits are measured the potential exists to discover the coordinated control of these traits via genotyped polymorphisms. A common statistical approach to this problem involves assessing the relationship between each phenotype and each single nucleotide polymorphism (SNP) individually (PHN); and taking a Bonferroni correction for the effective number of independent tests conducted. Alternatively, one can apply a dimension reduction technique, such as estimation of principal components, and test for an association with the principal components of the phenotypes (PCP) rather than the individual phenotypes. Building on the work of Lange and colleagues we develop an alternative method based on the principal component of heritability (PCH). For each SNP the PCH approach reduces the phenotypes to a single trait that has a higher heritability than any other linear combination of the phenotypes. As a result, the association between a SNP and derived trait is often easier to detect than an association with any of the individual phenotypes or the PCP. When applied to unrelated subjects, PCH has a drawback. For each SNP it is necessary to estimate the vector of loadings that maximize the heritability over all phenotypes. We develop a method of iterated sample splitting that uses one portion of the data for training and the remainder for testing. This cross-validation approach maintains the type I error control and yet utilizes the data efficiently, resulting in a powerful test for association. *Genet. Epidemiol.* 32: 9–19, 2008.

Key words: dimension reduction; heritability; multiple phenotypes; polygenic effects; whole genome association test

*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, 15213 Pennsylvania. E-mail: roeder@stat.cmu.edu
Received 12 April 2007; Revised 23 May 2007; Accepted 23 June 2007
Published online 5 October 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20257

INTRODUCTION

Quantitative traits often underlie risk for complex disease. Risk for cardiovascular disease rises and falls with measures of blood pressure. Risk for type II diabetes increases with increasing obesity. Although less direct, risk for psychiatric disorders can follow the same pattern. For example, degrees of obsessionality, perfectionism, and rigidity predict liability to anorexia nervosa [Bulik et al., 2005]. Cardiovascular disease, obesity and anorexia are all heritable, as are the quantitative traits (QT) underlying risk. Often the QT are more heritable, and are thought to have a simpler genetic basis than the heterogeneous disease itself. For that reason, researchers hunting for disease susceptibility loci often hunt for quantitative trait loci (QTLs) underlying risk.

Suppose we wish to find a single nucleotide polymorphism (SNP) corresponding to a QTL for anorexia, say for a measure of obsessionality. It is hard to imagine that this QTL does not impact other quantitative traits, some of which could also be related to risk for anorexia, such as perfectionism and rigidity. To find this QTL, one could seek for association between individual SNPs and individual traits. We will refer to this type of analysis as phenotypic analysis (PHN). A drawback of this approach is that it can require a substantial penalty for multiple testing when dealing with many phenotypes.

Clearly, power for association analysis can be enhanced if the dimension of the problem is reduced, consequently reducing the number of tests performed. In this study we focus on decreasing the dimension by decreasing the number of phenotypes through transformations that exploit possible pleiotropic effects of the SNP. The best known method for dimension reduction involves using one or more of the principal components of the phenotypes (PCP) in place of the original phenotypes. Although this approach chooses a derived phenotype with maximum variability, it is not genetically based. Indeed, it is quite possible that the PCP has a very low heritability.

A little known alternative is a method based on the *principal component of heritability* (PCH), which derives a trait based on the measured phenotypes to enhance the heritability. This concept

was introduced in the context of pedigrees by Ott and Rabinowitz [1999]. The general notion has a history in quantitative genetics [e.g., Klei et al., 1988] and more recently for linkage analysis [Bauman et al., 2005], but it is little used for association analysis. Recently the PCH concept was applied to family-based association studies by Lange et al. [2004]. PCH is based on the notion of optimizing the phenotypic variance explained by the QTL (or genetic variant), e.g., the heritability attributable to the QTL. Although derived from the traditional meaning of the term, this notion of heritability attributable to a QTL should not be confused with the total genetic heritability of a trait [Falconer, 1985]. The latter quantity is usually calculated using family data, without reference to any specific genetic variants. In contrast, the heritability attributable to a QTL can be calculated directly from a random sample from the same population. However, heritability in this context only accounts for a small fraction of the total heritability. The remaining polygenic part of the heritability remains as part of the residual variance. Consequently the heritability attributed to a specific QTL accounts for only a fraction of the total genetic heritability of the trait itself.

This approach, which we call PCH, reduces m phenotypes to a single phenotype that has a higher heritability attributable to the particular QTL than the heritability of any of the measured phenotypes or the PCP (Fig. 1). Indeed PCH is the optimal linear combination of traits from a heritability point of view. As a result, the association between a QTL and the linear combination of phenotypes that maximizes the heritability attributable to the QTL should be easier to detect than an association with any of the individual phenotypes or the PCP. In addition the PCH only requires one test and can be based on a preset significance level α while PHN requires m tests and the significance level has to be adjusted for this multiplicity.

The main disadvantage when using PCH is that one has to estimate the appropriate loadings for each of the SNPs and each of the measured phenotypes. For family trios, Lange et al. [2004] recommend using a function of the allele frequency in the parental generation to estimate this quantity. This approach avoids double use of the data. In this paper we consider population-based studies for which this approach is not feasible. We explore an efficient method of sample splitting and cross-validation to determine these loadings from a training set and then test for association using the remainder of the sample. We investigate the power trade off with this approach compared with the PHN and PCP approaches in a variety of scenarios through analytical calculations and simulated datasets, finding that in general PCH is a more powerful approach for locating QTL.

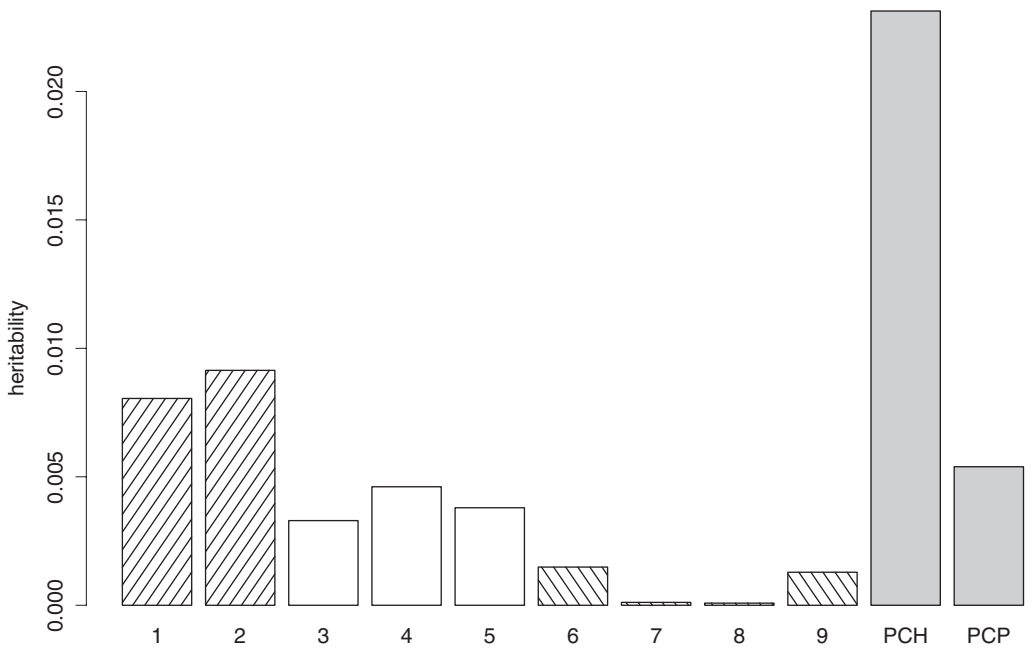

Fig. 1. Heritability of nine component traits compared with the heritability of the PCH trait and PCP trait. Data were simulated as described in the Results section with $h_j^2 = 0.08$ for phenotypes $j = 1, 2$; 0.04 for $j = 3, 4, 5$; 0.0005 for $j = 6, 7, 8, 9$ and $N = 1{,}500$. Notice that the heritability attributed to the linear combination defined by PCH is considerably larger than the heritability attributable to any of the nine traits or the linear combination defined by PCP.

METHODS

GENERAL BACKGROUND

Measure m traits $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})$ for subjects $i = 1, \ldots, N$. For each QTL $l = 1, \ldots, L$, we define the allele count as the number of copies of the minor allele. Call it x_{il} for subject i. For each phenotype j and QTL l, we assume a simple linear regression model approximates the relationship between the y_{ij} and x_{il}

$$y_{ij} = \mu_l + \beta_{jl} x_{il} + \varepsilon_{ijl}, \tag{1}$$

where μ_l is the intercept of the model and β_{jl} is the effect of the lth SNP on the jth trait. The residual error, $\mathrm{Var}[\varepsilon_{ijl}] = \sigma^2_{R_{jl}}$, includes other genetic effects, environmental effects, and measurement error. We assume that $\varepsilon = (\varepsilon_{i1l}, \ldots, \varepsilon_{iml})^t$ is approximately normally distributed with mean 0.

On the basis of this model, in the tradition of quantitative genetics [Falconer, 1985], we partition the total phenotypic covariance of $\mathbf{y}$ as

$$V_P = V_{Q_l} + V_{R_l},$$

where $V_{Ql} = \mathrm{Var}[\beta_{1l} x_l, \ldots, \beta_{ml} x_l]$ is the genetic variance due to the lth QTL, and $V_{R_l} = \mathrm{Var}[\varepsilon]$ is the residual covariance after removing the genetic effect of the lth QTL.

In the sequel, we assume that we are focusing on a particular SNP and drop the subscript l for simplicity, but note that in practice each calculation must be done for each SNP. For any linear combination of the phenotypes $\boldsymbol{w}^t\boldsymbol{y} = w_1 y_1 + \cdots + w_m y_m$ the heritability attributable to QTL x is defined as

$$h^2_w = \frac{\mathbf{w}^t V_Q \mathbf{w}}{\mathbf{w}^t V_P \mathbf{w}}. \tag{2}$$

TESTING

Three methods are investigated, two classic methods, PHN, PCP, and one more recent development, PCH [Ott and Rabinowitz, 1999; Lange et al., 2004], which we adapt for application in population-based studies. Each of these procedures involves the testing for association between one or more phenotypes ($\mathbf{w}^t\mathbf{y}$) and x. For any choice of vector $\mathbf{w}$ the linear association between $\mathbf{w}^t\mathbf{y}$ and a QTL x can be modeled by $\mathbf{w}^t\mathbf{y} = \mu + \beta x + \varepsilon$. To test the hypothesis, $\beta = 0$ vs. $\beta \neq 0$, we use a t-test $T = b/se(b)$, where b is the least-squares regression coefficient and $se(b)$ is the standard error of that quantity. From (2) and results from linear regression theory, it follows that

$$E[b] = \beta = \left(\frac{\mathbf{w}^t V_P \mathbf{w}\, h^2_w}{2p(1-p)}\right)^{1/2}$$

and

$$se(b) = \left(\frac{\mathbf{w}^t V_P \mathbf{w}\,(1-h^2_w)}{2Np(1-p)}\right)^{1/2}.$$

Consequently, $T \sim$ Normal (δ,1) with noncentrality parameter

$$\delta \equiv \left(\frac{N h^2_w}{1-h^2_w}\right)^{1/2}. \tag{3}$$

From this calculation it is clear that the power of the test for association increases if either the heritability or the sample size increases. This is the motivation behind the PCH approach.

Of the three approaches to testing we consider, PHN is the simplest, involving a univariate test of association for each of the m measured traits: $w_1 = (1, 0, \ldots, 0)$, $w_2 = (1, 0, \ldots, 0)$, and so on. This approach results in a vector of tests, $\mathbf{T} = (T_1, \ldots, T_m)$. As multiple tests are conducted, a multiple testing correction is required. The component tests are correlated due to the correlation among the measured traits, hence a Bonferroni correction for m tests is conservative.

To obtain a size α test we first determine the number of effective tests by solving for m_e such that $P(\max_j |T_j| > Z_{\alpha/2m_e}) = \alpha$ when $\delta_j = 0$, for $j = 1, \ldots, m$, $1 \le m_e \le m$. This calculation is performed assuming $\mathbf{T} \sim MVN(0, C_R)$, in which $C_R =$ correlation ($\mathbf{y}$). The power of this approach is then computed by determining $P(\max_j |T_j| > Z_{\alpha/2m_e})$. This calculation is made assuming $\mathbf{T} = MVN(\delta, C_R)$, where $\delta = (\delta_1, \ldots, \delta_m)^t$ is the vector of noncentrality parameters determined by the heritability of each of the traits.

PCP offers a lower dimensional approach to testing multiple phenotypes. It chooses loadings $\mathbf{w}$ to create a phenotype that maximizes $\mathrm{Var}[\mathbf{w}^t\mathbf{y}]$. Here the eigenvalue decomposition of V_P is used to obtain $V_P = VDV^t$. One or more columns of V, say $\mathbf{w}_k$, are then chosen to define univariate phenotypes $\mathbf{w}^t_k\mathbf{y}$, $k = 1, \ldots, K$.

The heritability of each linear combination is described in equation (2). The power of PCP can be computed directly because the PCP vectors are orthogonal: power $= [1 - \prod_{k=1}^{K} (1 - \mathrm{power}_k)]$, where power_k is the power of detecting association between $\mathbf{w}^t_k\mathbf{y}$ and x using the kth column of $\mathbf{V}$. For each of K linear combination tested, we obtain an independent test ($T_1, \ldots, T_K$) and reject the null hypothesis if $\max_k |T_k| > Z_{\alpha/2k}$. To select K we determine which columns of $\mathbf{V}$ correspond to eigenvalues greater than 1.

PCH. The PCH method seeks loadings $\mathbf{w}$ such that the heritability attributable to the QTL is maximized (Fig. 2) for $\mathbf{w}^t\mathbf{y}$. The $\mathbf{w}$ that maximizes

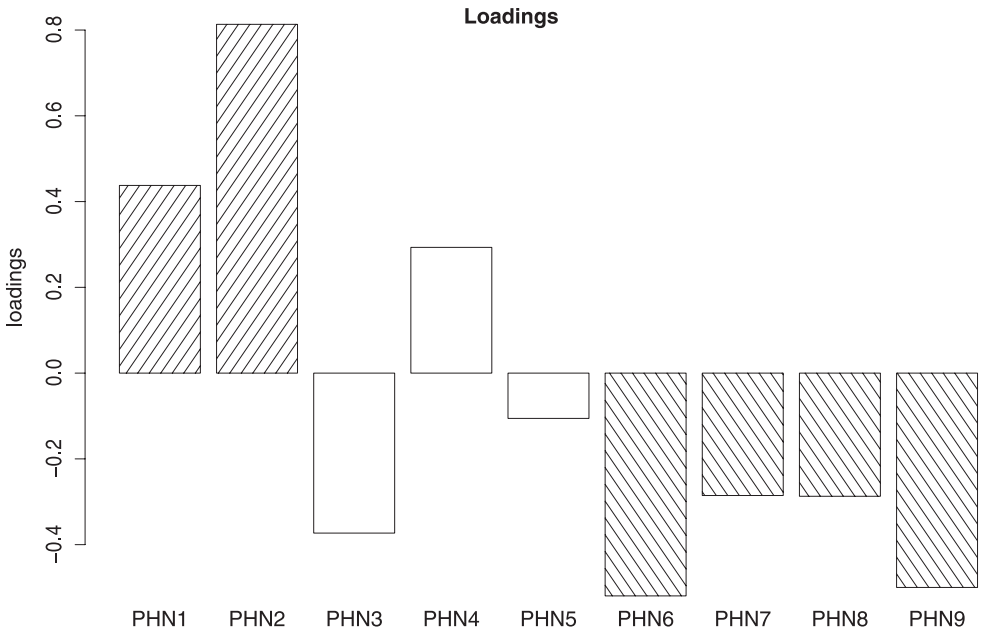

Fig. 2. The components of the loadings, *w*, that maximize the heritability of the data described in Figure 1.

this quantity is quite easily obtained by analyzing the eigenstructure of V_Q and V_R. To estimate these covariance matrices we adopt the following scheme. Define a diagonal matrix

$$\Gamma = \mathrm{diag}\left(|\beta_1|\sqrt{\mathrm{var}(x)}, \ldots, |\beta_m|\sqrt{\mathrm{var}(x)}\right) \qquad (4)$$

and vector $\mathbf{1} = (\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_m))^t$. Then the QT variance matrix is $V_Q = \Gamma \mathbf{1}\mathbf{1}^t\Gamma$. After fitting the linear model in (1) for $j = 1, \ldots, m$, replace β_j with the least-squares estimator b_j and ε_{ij} with the corresponding residual r_{ij}. Both Var(x) and the residual covariance matrix $V_R = \mathrm{Cov}(r_1, \ldots, r_m)$ can be estimated empirically: $\sigma_{jj'} = \frac{1}{N}\sum_i r_{ij}r_{ij'}$.

The principal component of heritability is based on the canonical transformation of V_P using a Cholesky decomposition to decompose $V_R = \mathbf{L}\mathbf{L}^t$, followed by an eigenvalue decomposition of $\mathbf{L}^{-1}V_Q(\mathbf{L}^{-1})^t \equiv \mathbf{P}\Lambda\mathbf{P}^t$. The canonical transformation is then defined as P^tL^{-1}. It is easy to see that this transformation of $\mathbf{y}$ reduces both the covariance matrices to diagonal form:

$$\begin{aligned}\mathrm{Var}(P^tL^{-1}\mathbf{y}) &= P^t L^{-1} V_P L^{-t} P \\ &= P^t L^{-1}(V_Q + V_R)L^{-t} P = \Lambda + \mathbf{I}.\end{aligned}$$

Because of the special structure of V_Q, which is rank 1, only one eigenvalue λ will be non-zero. Call the eigenvector that corresponds to this eigenvalue $\mathbf{v}$. For $\mathbf{w} = (L^{-1})^t\mathbf{v}$, the heritability of $\mathbf{w}^t\mathbf{y}$ is $\lambda/(\lambda+1)$. This vector maximizes the heritability as defined in equation (2); see Ott and Rabinowitz [1999] and Lange et al. [2004].

The calculations above rely on x defined as an allele count, 0, 1, or 2, which is appropriate for an additive model. For a dominant or recessive model, similar results apply with x defined as a binary variable.

For the PCH method, if $\mathbf{w}$ were known, implementation would involve a single test of association between $\mathbf{w}^t\mathbf{y}$ and x. Power would be calculated as $P(|T| > Z_{\alpha/2})$ using δ as defined in equation (3) (see Appendix for details). Implementing the PCH approach and assessing its power is complicated, however, because $\mathbf{w}$ must be estimated, as described in the following section.

Sample splitting for PCH. If we estimate $\mathbf{w}$ from the data and then test for association between $\mathbf{w}^t\mathbf{y}$ and x on the same data, the type I error rate will be inflated. To implement PCH we propose a method of sample splitting similar to cross-validation (Hastie et al., 2001). The algorithm involves splitting the sample into disjoint subsets using N_0 observations for training (estimation of $\mathbf{w}$) and the remaining N_1 observations for testing, with $c = N_0/N$ typically being a small fraction, say 0.2. Conditional on the estimated quantity $\mathbf{w}$ and corresponding heritability h^2_w, the resulting test statistic $T \sim N(\delta,1)$, with $\delta = (N_1h^2_w)/(1 - h^2_w)$. Unfortunately, this approach has a substantial drawback: the resulting test statistic varies considerably depending on the choice of training and testing subsets. To circumvent this problem we propose calculating the test statistic repeatedly using random splits of the data. Ultimately, the p-value is derived from an integration of the individual test statistics.

For a given split of the sample, regress $\mathbf{w}^t\mathbf{y}$ and x and compute the resulting t-test $T_s = b/se(b)$. For each $s \in S$, the set of all possible subsets of size N_1, repeat this process. The average of these tests is

$$\overline{T} = \binom{N}{N_1}^{-1} \sum_{s \in S} T_s. \quad (5)$$

This is a generalized U-statistic, which has certain well-known properties [Lee, 1990]. For computational efficiency, $\overline{T}$ can be well approximated by taking an average of tests derived from many random subsets of size N_1. When the number of subsets, S, is very large, and c is small, $\overline{T}$ is approximately normally distributed. Under the null hypothesis of no association $E[\overline{T}] = 0$. Thus for large N, a size α test rejects H_0 if $|\overline{T}|/se(\overline{T}) > Z_{\alpha/2}$, for $se^2(\overline{T}) \equiv \mathrm{Var}[\overline{T}]$.

An analytical calculation of $se(\overline{T})$ is intractable due to correlation between pairs of tests $(T_s, T_{s'})$. An estimate of $se(\overline{T})$ can be obtained by simulating the U-statistic process under the null hypothesis by the following steps. (1) Generate data by randomly permuting the multivariate phenotypes (jointly together) independently of the individuals' genotypes. This maintains $\mathrm{Var}[\mathbf{y}] = V_P$ while breaking any relationship to the SNP genotypes. Generate g such data sets. (2) For each permutation and each SNP, calculate $\overline{T}$. For L SNPs this yields gL realizations of the statistic. (3) Compute $se^2(\overline{T})$ from the variance of the resulting gL values of $\overline{T}$.

These steps complete the outline of the process, but two other actions are required for effective analyses, namely a 'bagging sub-step' to refine the estimates of the weights $\mathbf{w}$ and a step to estimate the distribution of $\overline{T}$.

Bagging. For a given split of the data, s, we use N_0 training observations to estimate $\mathbf{w}_s$; for the remaining N_1 testing observations we compute $\mathbf{w}_s^t\mathbf{y}$; and finally we compute T_s using the testing observations. Because N_0 constitutes a small fraction of the data, $\mathbf{w}_s$ is susceptible to the influence of outliers in the training set. Bagging is a statistical technique for reducing the variability of non linear estimators of this sort [Bishop, 2006]. In its simplest form, bagging draws bootstrap samples from the training sample, applies the estimation procedure to each bootstrap sample, and then averages the resulting estimator across bootstrap samples. In our application this means, for each bootstrap sample from the particular training data set s training data, we determine V_Q and V_R and then solve for $\mathbf{w}$; call it $\mathbf{w}_{bs}$. We then take the average of these values over B bootstrap samples, to obtain $\mathbf{w}_s = \frac{1}{B}\sum_b \mathbf{w}_{bs}$. Finally, we apply these loadings and calculate T_s. We found that the use of bagging increased the power of our test, without compromising the size of the test.

It is worth noting that the canonical transformation is an orthogonal projection in which the direction (positive or negative) of the projection axes is arbitrary. General eigenvalue decomposition algorithms and software assign these directions randomly. As a result the same training set can have a large positive coordinate associated with a phenotype for one bootstrap sample, but a large negative coordinate for the next bootstrap sample. We found that requiring the coordinate for the first phenotype to be positive controlled this artifact. Also, when the minor allele frequency of an allele is less than 0.05, there may be too few individuals with the minor allele in the training sample to estimate $\mathbf{w}$. Increasing N_0 and skipping the bagging step of the algorithm can circumvent this problem.

Distribution of $\overline{\mathbf{T}}$. For a fixed partition, (N_0,N_1), the testing sample is independent of the training sample. The test statistic T_s, however, depends on $\mathbf{w}$, and hence is not independent of the training sample. This induces extra correlation between pairs of statistics $(T_s, T_{s'})$ beyond what is expected in a traditional U-statistic. This results in a statistic $\overline{T}$ with heavier tails than expected. We account for this by modeling $\overline{T}$ using a t-distribution with unspecified degrees of freedom d. We estimate d under the null hypothesis using a method of moments estimator that finds d such that $P(|\overline{T}|/se(\overline{T}) > Q_{d,0.01}) = 0.01$, where $Q_{d,a}$ is the $(1-a)$ quantile of a t-distribution with d degrees of freedom. This calculation is made using the gL realizations of $\overline{T}$ obtained in the permutation procedure. The tests rejects the hypothesis of no association when $|\overline{T}|/se(\overline{T}) > Q_{d,\alpha/2}$.

Algorithm. To obtain the P-value for a single SNP:

1. Randomly split the data into a partition called s containing (N_0, N_1) observations.
   (a) Within the N_0 observations, perform the bagging calculations to obtain $\mathbf{w}_s$.
   (b) Using $\mathbf{w}_s$ on the N_1 observations, calculate T_s.
2. Repeat the first step S times and average the tests to obtain $\overline{T}$.
3. From the calculation of the null distribution (below) obtain $se(\overline{T})$ and d.
4. Define $z_{obs} = |\overline{T}|/se(\overline{T})$. $P\text{-value} = 2P(Z > z_{obs})$ where $Z \sim t$ with d degrees of freedom.

To obtain the null distribution of $\overline{T}$:

1. Permute the data g times.
2. For each permutation and each SNP, repeat steps 1 and 2 above to obtain $\overline{T}$.

3. Using each of these realizations of the statistic, compute $se(\overline{T})$ and d.

The process of finding the standard error and degrees of freedom need only be performed once for each combination of phenotypes. These quantities can then be used for tests of any SNP potentially associated with a given set of phenotypes and SNPs.

RESULTS

Our simulations target a portion of the parameter space in which power is not universally high. To accomplish this aim we keep heritability attributable to individual SNPs modest. In addition, SNP variation impacts only a small fraction of the traits measured for most scenarios. We investigated designs with a small, moderate and large number of measured traits ($m = 5, 10, 20$). Each trait had one of three levels of heritability attributable to a particular SNP, dubbed high (0.8%), medium (0.4%) and very low (0.005%). Heritabilties of levels (high, medium, very low) were distributed across the m traits using four patterns of counts in each category: (A) $(m, 0, 0)$; (B) $(1, 0, m-1)$; (C) $(1, 1, m-2)$; and (D) $(2, 2, m-4)$. Two levels of residual correlation between traits were explored, namely high (H) and low (L) (Table I).

As noted in the Methods, the test statistic developed here cannot be assumed to follow a normal distribution. Rather we approximate it as a t-distribution and estimate its degrees of freedom. Note that a t-distribution with a small number of degrees of freedom has notably heavier tales than a normal distribution. To evaluate this approach, we performed a simulation study. We let the fraction of the data used for training c range from 0.05 to 0.5 and $N = (400, 800, 1600)$ using model C, with correlation H and $m = 10$. (Recall that degrees of freedom d are chosen to ensure type I error control of 0.01 when $\alpha = 0.01$.) Results (Table II) indicate that when N is small or c is large, d is small indicating that the normal approximation is not adequate. For instance, when half the data are used for training, d of approximately 12 is required to control the type I error. But when $N = 1600$ and $c = 0.06$, $d = 146$ is sufficient; in this instance the distribution is approximately normal. Next, to determine if the t-distribution provides an adequate approximation to the true distribution of the statistic, we compare the achieved size of the test to the target. The results indicate that the test is well calibrated for $\alpha = 0.05$ and 0.001 (Table II). This comparison shows that choosing d to control the size of the test at $\alpha = 0.01$ successfully controls the size of the test more generally.

TABLE I. Residual covariance matrices for the high and low covariance structures

		high	med	low
H	high	0.9	0.6	0.3
	med	0.6	0.9	0.1
	low	0.3	0.1	0.1
L	high	0.5	0.3	0.1
	med	0.3	0.5	0.1
	low	0.1	0.1	0.1

TABLE II. Size of PCH test and estimated degrees of freedom (d) for population size (N), and fraction of data used for training (c)

			α	
N	c	d	0.05	0.001
400	0.10	61	0.050	0.0009
	0.20	26	0.045	0.0015
	0.25	21	0.044	0.0012
	0.33	15	0.044	0.0009
	0.50	12	0.049	0.0005
800	0.05	39	0.046	0.0010
	0.13	49	0.047	0.0005
	0.19	39	0.048	0.0017
	0.20	31	0.045	0.0012
	0.25	25	0.044	0.0019
	0.33	14	0.042	0.0009
	0.50	14	0.047	0.0007
1,600	0.06	146	0.047	0.0015
	0.09	500	0.050	0.0013
	0.20	34	0.045	0.0009
	0.25	22	0.045	0.0005
	0.33	18	0.045	0.0012
	0.50	13	0.049	0.0006

Note: Simulations were repeated 10,000 times. The standard error of the size of the test is approximately $[\alpha(1-\alpha)]^{1/2}/100$.

Because the test is valid under the null hypothesis for a broad range of choices for c, we use power to determine the best choice for this parameter (Fig. 3). From our simulations, the optimal choice varies with N. For $N = 1600$, c of about 9% is ideal, but for $N = 800$ c of about 19% is best. In both cases this involves using 150 observations for training. For $N = 400$, however, the optimal choice is less clear. We suggest using about 20% of the sample for training when the sample size is small and otherwise using about 150 observations.

Next we explored the choice of the number of bagging steps, B, and the number of subsampling steps, S. In view of our investigations it appears that both size and power of the test are essentially unchanged for B and S at least 50 (Tables III and IV). Nevertheless, we recommend using 100–200 repetitions for SNPs that yield promising results to ensure precise P-values.

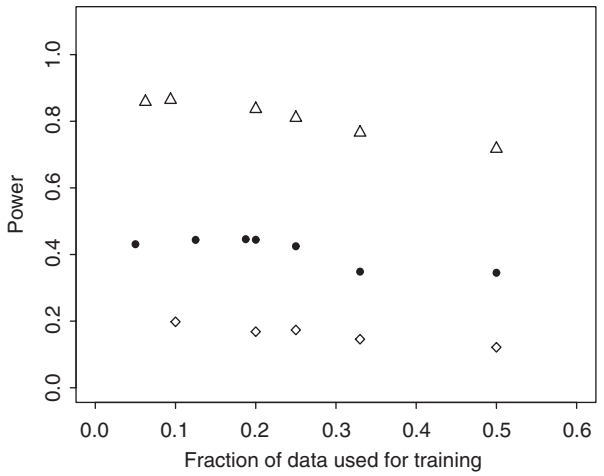

Fig. 3. Power of PCH test at $\alpha = 0.01$, for different population sizes N (◇, $N = 400$; ●, $N = 800$; △, $N = 1{,}600$) and different fractions c of data used for training. The power is maximized for $c = 0.09$ when $N = 1{,}600$ and $c = 0.19$ when $N = 800$, which is in both cases a sample of about 150 observations. For $N = 400$ the power remains almost unchanged.

TABLE III. Size of PCH, PHN and PCP tests

		α	
PCH			
S	*B*	0.05	0.001
50	50	0.047	0.0006
50	100	0.044	0.0005
50	200	0.047	0.0013
100	50	0.044	0.0007
100	100	0.046	0.0018
100	200	0.049	0.0018
200	50	0.049	0.0006
200	100	0.042	0.0007
200	200	0.048	0.0009
PHN		0.048	0.0009
PCP		0.049	0.0011

Note: For PCH test, the size was calculated for different numbers of iterations for the random samples, *B* is the number of bootstrap iterations of the bagging procedure, and *S* is the number of iterations of the subsampling procedure. Simulations were repeated 10,000 times. The size of the test is essentially unchanged for *B* and *S* of at least 50.

To compare PHN, PCP and PCH, we simulated power for each of the 24 configurations described using $N = 800$ subjects and $N_0 = 150$ (Fig. 4). For design A, which resembles a repeated-measured design, PCP is most powerful, as expected. With this type of data it makes sense to use PCP or the sample average for the phenotype, but it would be difficult to know when this model would be applicable unless the traits were direct repeated measures. Alternatively, for design B, the ''one trait'' model,

TABLE IV. Power of PCH test

		α		
S	*B*	0.05	0.01	0.001
50	50	0.68	0.44	0.21
50	100	0.66	0.40	0.15
50	200	0.69	0.45	0.21
100	50	0.69	0.44	0.18
100	100	0.69	0.44	0.19
100	200	0.70	0.46	0.22
200	50	0.70	0.46	0.21
200	100	0.68	0.43	0.18
200	200	0.70	0.46	0.22

Note: See Table III for details.

PCP has essentially no power and should be avoided. PCH is more powerful than PHN in this setting because PHN takes a multiple testing penalty. This difference is more marked for larger m. For designs C and D, which include a mix of heritabilities across traits, PCH tends to be considerably more powerful than either PHN or PCP. The difference in power between PCH and the other methods is affected by the correlation structure. As a general rule, if the traits are highly correlated, and/or numerous, then the advantage of PCH becomes more notable for all designs investigated. Finally, we note that designs C and D are just two of many possibilities that incorporate a mix of heritabilities across the traits and PCH tends to be the best method for data of this type.

We also compared theoretical power calculations (see Methods and Appendix) to empirical power achieved in the simulation experiment for all 24 scenarios. For PCH, implementation of the test requires a choice of $\mathbf{w}$ for each SNP. For a given genetic model, one can calculate the ''oracle power'' using the noncentrality parameter (2) that corresponds to the best $\mathbf{w}$, but this power cannot be achieved in practice. Power is lost due to the variability in the estimated quantity $\mathbf{w}$ versus the optimal choice. We found that, for $N = 800$ and $N_0 = 150$, the achieved power is very closely approximated by the oracle power calculated with N replaced by $0.67N$ when $m = 5$ and $m = 10$ (Fig. 5). For $m = 20$, however, the difference between the oracle and the achieved power is larger and the match between simulated and theoretical power would be better for $0.6N$.

All of our simulations thus far have had complete data. To study the effect of missing data, we simulated a complete data set consisting of 1,000 individuals, 10 traits and 4,000 SNPs. For 10 SNPs we used model D with correlation H; the rest of the SNPs were simulated under the null hypothesis. To generate phenotypes and/or genotypes missing at

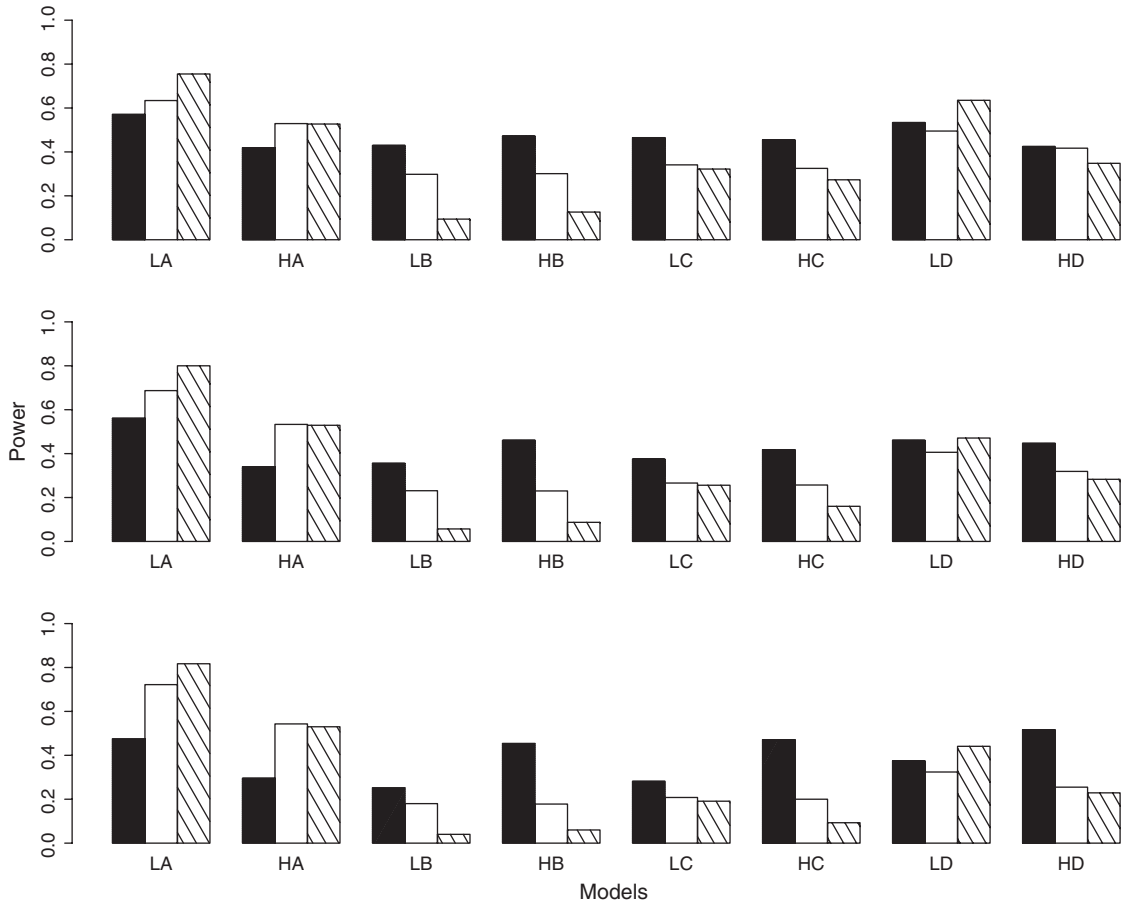

Fig. 4. Power of PCH (solid), PHN (open) and PCP (hatched) tests for different designs: the number of measured traits is $m = 5$ (upper), 10 (center), or 20 (bottom), the two types of correlation are H (L), with high (moderate) covariance between traits with similar heritabilities and moderate(low) between all remaining traits. Heritabilities of level (high, medium, very low) are distributed across the m traits using four models for each category: A $(m, 0, 0)$, B $(1, 0, m-1)$, C $(1, 1, m-2)$, D $(2, 2, m-4)$.

random we used a Poisson distribution with the expected number of missing traits per individual equal to 0.10 and expected number of missing SNPs per individual equal to 5. The procedure resulted in 79 individuals with one missing phenotype, and 7 with two missing phenotypes. When SNP genotypes were missing, we removed the subjects from the analysis; when phenotypes were missing, we either imputed the values or dropped the individuals. Then the following subsets of these data were analyzed: (i) the complete data set; (ii) missing phenotypes only, with missing traits imputed; (iii) missing SNPs only; (iv) missing phenotypes and SNPs, with missing traits imputed; and (v) missing phenotypes only, remove individuals with missing traits. Of those SNPs with a *P*-value of 0.05 or less from analysis (i), the mean *P*-value for methods (ii)–(iv) was approximately 0.025 while method (v) had a mean of 0.049. In this restricted range of values, the correlation between *P*-values obtained for approach (i) with approaches (ii)–(v) was 0.87, 0.82, 0.81, and 0.51, respectively. Only dropping individuals with missing data had a substantial effect on the analysis. This experiment suggests that the imputation of missing data successfully preserves the data structure and has a non-negligible effect on the size and power of the test. Alternatively, removing missing data appears to diminish the power of the test.

DISCUSSION

An increasing number of genetic studies have measured numerous related traits and seek to identify genetic variation having substantial impact on the population-level variance of these traits. When individually testing many traits for association with many SNPs, power can be low due to corrections for multiple testing. To improve power, one could limit the number of tests performed by

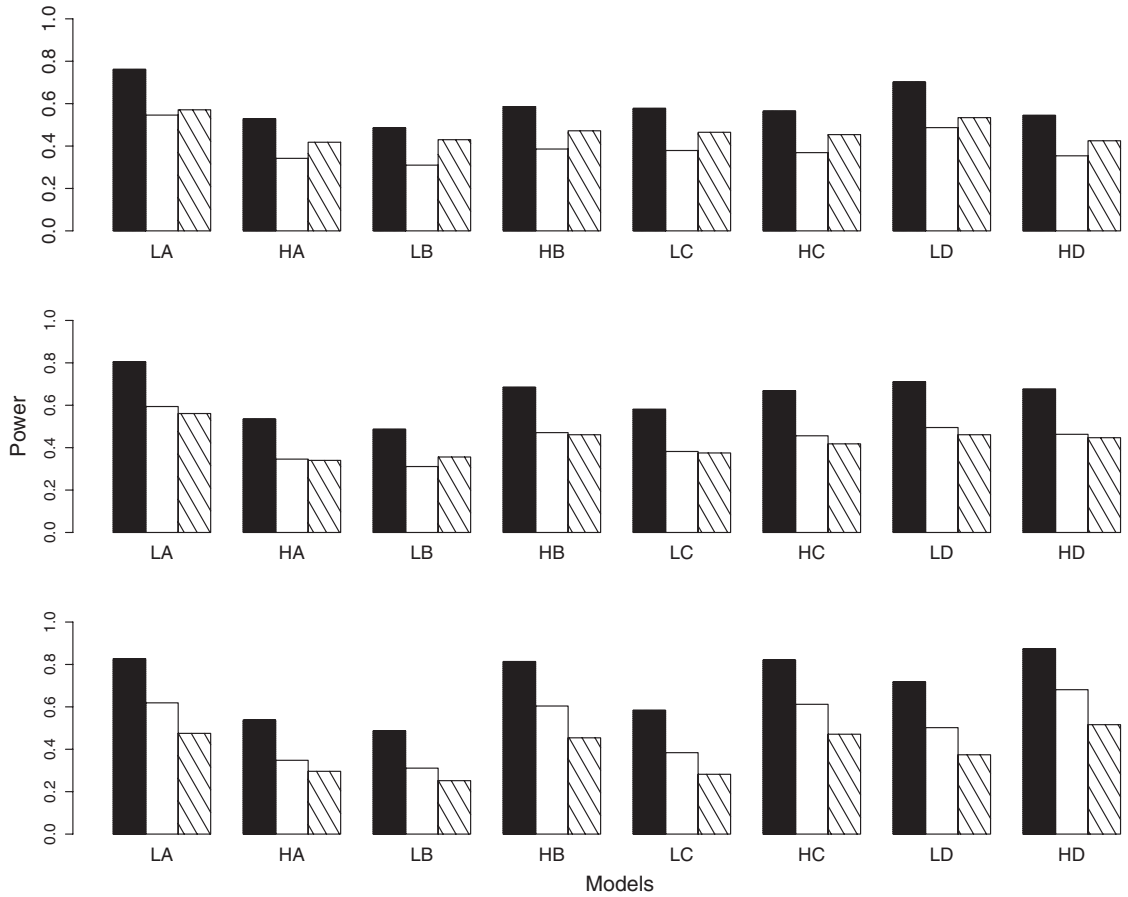

Fig. 5. Power of three tests: T(PCHm), theoretical maximum power of PCH (solid), T(PCHc), theoretical power of PCH with $c = .33$ (open), and PCH, the simulated power (hatched). Different designs: two types of correlation H (L), with high (moderate) covariance between traits with similar heritabilities and moderate (low) between all remaining traits. Heritabilities of level (high, medium, very low) are distributed across the m traits using four models for each category: A $(m, 0, 0)$, B $(1, 0, m-1)$, C $(2, 2, m-4)$, D $(1, 1, m-2)$. The number of measured traits is $m = 5$ (upper), 10 (center), and 20 (bottom).

implementing a standard dimension reduction technique. A common approach involves finding the principal component of the phenotypes. This derived trait is the linear combination of individual traits that accounts for the maximum variance. Although the principal component approach sometimes works well in practice, there is no reason to believe this linear combination of traits has particular genetic relevance. We develop an alternative method based on a different type of principal component known as the principal component of heritability. This approach finds the distinct linear combination of traits that maximizes the power to detect association with a given SNP under investigation. Because the linear combination differs for each genetic variant, we much estimate the linear combination from a set of training data. We develop a method of cross-validation that facilitates an iterative training and testing procedure for this task.

The methods proposed here build on the work of Lange et al. [2004], but their methods are tailored to family-based association designs. While our results target population-based designs, it would be interesting to see if the techniques we use here would be of value in the family-based setting. Like other analytic approaches, such as linkage analysis of complex traits [Sobel and Lange, 1993], our analysis has a stochastic component so that results can differ slightly by using different starting values. For instance, we repeated the analysis of a data set with 4,000 SNPs, 3,990 of which were not associated with the phenotypes. The second analysis differed only in the random subsets chosen. The correlation between the resulting P-values was considerably less than unity because the PCH method has this component of randomness. The variability among P-values was far greater for SNPs that were not associated with any of the phenotypes; for those SNPs with a strong signal the P-values were nearly identical for different starting values.

When measuring multiple phenotypes, the genetic underpinnings vary between two extremes. On one

extreme each phenotype is measuring essentially the same underlying trait. In this scenario each measurement has a similar heritability. Traits will tend to have a high correlation also, due to the common influence of other genes and environmental factors. With this type of data, often called repeated measures, a nearly ideal method of analysis assesses the correlation between the SNPs and the average measured phenotype. PCP generates a single trait that is highly correlated to the sample average, so this approach is tailored for this type of data. With PHN, each repeated measurement of the trait is compared with the SNP genotypes, sequentially. Because the traits are highly correlated, the effective number of tests will be near one; nevertheless, some power is lost due to multiple testing and extra residual variability. Finally, PCH will estimate a linear combination similar to the one used by PCP; however, power will be lost due to the penalty for estimation that is not incurred by PCP.

On the opposite extreme, consider a study for which only one phenotype is genetically linked to the variant under investigation. The remaining phenotypes are linked with the heritable trait only through environmental factors and perhaps other genes. In this scenario, the analysis should assess the correlation between the genetic variant and the appropriate phenotype. PHN analysis will make this comparison, but it does not achieve it without incurring a multiple-testing penalty. This penalty can be substantial if the correlation between traits is low. PCP will create a linear combination that includes many phenotypes that are unrelated to the variant of interest, generating a noisy phenotype. Finally, PCH comes the closest to estimating the optimal trait. The PCH-defined trait features the desired phenotype, and it can also enhance the power by taking a linear combination of phenotypes that reduce the variability due to measurement error [see Lange et al., 2004]. In spite of the penalty for estimation, this approach yields the best power in this scenario.

On the basis of our simulations we find that models between these two extremes also favor PCH in terms of power. The method performs well with a broad range of phenotypes, ranging from 5 to 20 in our simulations (Fig. 4). Remarkably, the approach has good power even when 19 of the 20 traits have negligible heritability. Thus, we believe the PCH is a good choice for analysis provided the traits are not primarily repeated measures of a single trait. If several of the traits are repeated measures, we suggest replacing these repeated measures with a simple average of the measures. This will reduce the dimension of the problem, which facilitates the estimation of the trait loadings in the analysis. Also, for any investigation of many traits, one might perform an additional exploratory analysis using PHN, to identify SNP/phenotype combinations that might merit further investigation.

An alternative approach to test for an association between a suite of traits and a genotype is traditional multivariate analysis of variance. In an informal simulation study this approach showed no advantages (results not shown). Overall, the power is not greater, and it has the disadvantage of relying on many assumptions.

Three other features of our analyses should be noted. (1) We did not explore the performance of the dimension reduction approach when the number of traits was greater than 20. If a very large number of traits have been measured, we suggest a preliminary analysis step in which related measures are averaged, or subsets of traits are analyzed separately. (2) As in any population-based test for association population, structure can lead to spurious associations. To remove this effect while using the PCH method, one could apply the eigenstrat principle [Price et al., 2006] as follows: compute the principal component decomposition following Price et al.; obtain the residuals for each phenotype after regressing out several principal components; and finally, perform the association study using the residuals rather than the original phenotypes. (3) When PCH is applied SNP by SNP, as we explore it does not assimilate information across a gene to discover the PCH for a gene. Such an approach could be beneficial in some circumstances and we plan to extend the method to handle more diverse types of data.

An interesting contrast can be drawn between the family-based and population-based implementations of PCH. With the former, heritability can be increased by using a linear combination of traits even if many of the traits do not have a genetic effect. This advantage is less likely to occur for the latter implementation. In a family-based design the loadings are estimated using ''supplementary information'', specifically the expected allele transmission, given the parents genotypes. Thus no information is lost due to estimation of the loadings. Power can be gained by using additional correlated traits that reduce the overall variance. The same would be true in population-based studies if the loadings were estimated from a pilot study, published data, or an earlier stage of a multistage design (e.g., a two-stage genome-wide association study).

ELECTRONIC DATABASE INFORMATION

PCHAT http://wpicr.wpic.pitt.edu/WPICCompGen/ (software for Principal Components of Heritability Association Tests)

ACKNOWLEDGMENTS

This work was supported by National Institute of Mental Health grant MH057881.

REFERENCES

Bauman LE, Almasy L, Blangero J, Duggirala R, Sinsheimer JS, Lange K. 2005. Fishing for pleiotropic QTLs in a polygenic sea. Ann Hum Genet 69:590–611.

Bishop CM. 2006. Pattern recognition and machine learning. New York: Springer.

Bulik CM, Bacanu SA, Klump KL, Fichter MM, Halmi KA, Keel P, Kaplan AS, Mitchell JE, Rotondo A, Strober M, Treasure J, Woodside DB, Sonpar VA, Xie W, Bergen AW, Berrettini WH, Kaye WH, Devlin B. 2005. Selection of eating-disorder phenotypes for linkage analysis. Am J Med Genet B Neuropsychiatr Genet 139:81–87.

Falconer DS. 1985. Introduction to quantative genetics, 2nd ed. New York: Longman.

Hastie T, Tibshirani R, Friedman J. 2001. The elements of statitical learning, data mining, inference, and prediction. New York: Springer.

Klei L, Pollak EJ, Quaas RL. 1988. Genetic and environmental parameters associated with linearized type appraisal scores. J Dairy Sci 71:2744–2752.

Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B. Murphy A, Silverman EK, MacGregor A, Weiss ST, Laird NM. 2004. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat Appl Genet Mol Biol 3:(Article17).

Lee AJ. 1990. U-statistics: theory and practice. New York: Marcel Dekker.

Ott J, Rabinowitz D. 1999. A principal-components approach based on heritability for combining phenotype information. Hum Heredity 49:106–111.

Sobel E, Lange K. 1993. Metropolis sampling in pedigree analysis. Stat Methods Med Res 2:263–282.

APPENDIX

The following describes an algorithm for calculating the power of PCH, given a particular genetic model.

1. Choose the heritability of each measured trait ($h_1^2, \ldots, h_m^2$), allele frequency p, and residual variance matrix V_R. For convenience we assume the variables are scaled so that the latter is a correlation matrix.
2. From equation (2) the heritability of a single phenotype is

$$h_j^2 = \frac{2p(1-p)\beta_j^2}{2p(1-p)\beta_j^2 + \sigma_{R_j}^2}.$$

   Solving for β_j we find

$$\beta_j^2 = \frac{h_j^2}{(1-h_j^2)2p(1-p)}.$$

3. Solve for Γ and then for V_Q in terms of $(\beta_1,\ldots,\beta_m)$ using equation (4).
4. Solve for $\mathbf{w}$ in terms of V_Q and V_R using the canonical decomposition.
5. Solve for nonconcentrality parameter δ using equation (3), with sample size $N^* = 0.67\,N$, for $m \approx 10$.
6. Approximate power is $P(|T| \geq Z_{\alpha/2})$, using δ calculated above.