Published in final edited form as:

Genet Epidemiol. 2020 January ; 44(1): 67–78. doi:10.1002/gepi.22263.

Joint analysis of multiple phenotypes using a clustering linear combination method based on hierarchical clustering

Xueling Li¹, Shuanglin Zhang¹, Qiuying Sha¹

¹Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America

Abstract

Emerging evidence suggests that a genetic variant can affect multiple phenotypes, especially in complex human diseases. Therefore, joint analysis of multiple phenotypes may offer new insights into disease etiology. Recently, many statistical methods have been developed for joint analysis of multiple phenotypes, including the clustering linear combination (CLC) method. Due to the unknown number of clusters for a given data, a simulation procedure must be used to evaluate the p-value of the final test statistic of CLC. This makes CLC method computationally demanding. In this paper, we use a stopping criterion to determine the number of clusters in the CLC method. We have named our method hierarchical clustering CLC (HCLC). HCLC has an asymptotic distribution, which is very computationally efficient and makes it applicable for genome-wide association studies (GWAS). Extensive simulations together with the COPDGene data analysis have been used to assess the type I error rates and power of our proposed method. Our simulation results demonstrate that the type I error rates of HCLC are effectively controlled in different realistic settings. HCLC either outperforms all other methods or has statistical power that is very close to the most powerful method with which it has been compared.

Keywords

Association studies; Multiple phenotypes; Joint analysis; Clustering linear combination; Hierarchical clustering

INTRODUCTION

Pleiotropy is a well-established phenomenon in which a single locus affects more than one distinct but possibly correlated phenotypic traits (Gratten & Visscher, 2016). Pleiotropy has had many important implications on physiological and medical genetics and evolutionary biology (Stearns, 2010). Substantial evidence has shown that pleiotropy is ubiquitous in complex human diseases (Sivakumaran et al., 2011). In addition, multiple correlated disease attributes (also termed as disease phenotypes) that relate to clinically meaningful outcomes,

Correspondence: Qiuying Sha, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931. qsha@mtu.edu.

Data Availability Statement: Data sharing is not applicable to this article as no new data were created or analyzed in this study. The authors have no conflict of interests to declare.

such as symptoms, exacerbations, responses to therapy, or rate of disease progression, are often collected and frequently encountered in genetic association studies (Han et al., 2010). Traditional genotype-phenotype association studies focus on the pairwise relationship between phenotypes and genotypes. However, single phenotype analyses ignore the pleiotropic effect and also suffer from multiple testing penalties, therefore these analyses may be considerably less powerful for detecting causal variants of weak effects (Sivakumaran et al., 2011). Contrasting with single phenotype analyses, joint analysis of multiple phenotypes exploits the correlations between phenotypes and aggregates multiple weak marginal effects and is therefore likely to provide important insights into the functional consequences of genetic variations.

In recent years, there has been increasing interest in jointly testing the association between a single genetic variant and multiple correlated phenotypes, with the null hypothesis that there is no association between the genetic variant of interest and any of the phenotypes while the alternative hypothesis is that the genetic variant of interest is associated with at least one of the phenotypes. The most widely used strategies for those research efforts are combining univariate analysis results, dimension reduction, and regression models. Methods involving the first strategy combine either the univariate test statistics (Kim, Bai, & Pan, 2015; O'Brien, 1984; Wei & Johnson, 1985) or p-values (Liang, Wang, Sha, & Zhang, 2016; van der Sluis, Posthuma, & Dolan, 2013; Yang, Li, Williams, & Buu, 2016); they are generally very easy to implement and can cope with a mixture of different types of phenotypes; however, the statistical power of those methods might heavily rely on the homogeneity of univariate test statistics (Zhu, Zhang, & Sha, 2015; 2018); the most popular methods in this category include O'Brien's method (O'Brien, 1984; Wei & Johnson, 1985), Trait-based Association Test that uses Extended Simes procedure (TATES) (van der Sluis, Posthuma, & Dolan, 2013), Fisher's Combination (Yang, Li, Williams, & Buu, 2016), and Adaptive Fisher's Combination (AFC) (Liang, Wang, Sha, & Zhang, 2016). For the strategy of dimension reduction, instead of testing one phenotype at a time, one first constructs a small number of latent variables, which are linear combinations of the observed phenotypes, and then tests the associations between the latent variables and the genetic variant of interest; dimension reduction methods are in general suitable only when all phenotypes are normally distributed (Yang & Wang, 2012); in addition, the newly derived latent variables are usually difficult to interpret in the real-world applications; the most popular methods in this category include principal components of the phenotypes (PCP) (Aschard et al., 2014), principal component of heritability (PCH) (Klei, Luca, Devlin, & Rowder, 2008; Wang, Sha, & Zhang, 2016; J. J. Zhou et al., 2015), and canonical correlation analysis (CCA) (Ferreira & Purcell, 2008; Tang & Ferreira, 2012). Compared with the other two strategies, the strategy involving regression models seems relatively complicated to implement but there have been a lot of R or SAS packages readily available for use and are able to handle a wide variety of single phenotype data types such as continuous, categorical, or survival, but not a mixture of them (Yang & Wang, 2012); common regression models include linear and generalized mixed effects models (Korte et al., 2012; Zhang et al., 2009; Zhou & Stephens, 2014), frailty models (Wienke, 2010; Yang & Wang, 2012), and generalized estimating equations (GEE) (Zeger & Liang, 1986; Zhang et al., 2014).

Hierarchical Clustering (HC) is a cluster analysis approach which builds a hierarchy of clusters (Ding & He, 2002; Johnson, 1967; Karypis, Han, & Kumar, 1999; Rokach & Maimon, 2005). There are two main types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). An agglomerative method starts with each object as a single cluster and merges the two clusters that have the smallest distance at each step of the clustering iteration until a given stopping criterion is met or there is only one cluster left. A divisive method starts with all objects belonging to the same cluster and repeatedly partitions a cluster into two such that the distance between the two new clusters are maximized in each step of the iteration until a given stopping criterion is met or each object is in its own singleton cluster. Traditional hierarchical clustering algorithms often merge or split one cluster at a time. In addition, there are many different ways of defining distance between clusters; these ways are often referred to as link functions. The most popular distances used for hierarchical clustering are single-link distance, complete-link distance, and average-link distance (Ding & He, 2002). The arrangement of the clusters generated by hierarchical clustering can be easily visualized in a dendrogram, a tree-based hierarchical taxonomy that records the sequences of merges or partitions. In the world of data mining and statistics, we usually refer the smallest distance in each step of the iteration as the height of the merged cluster in the dendrogram. For both types of hierarchical clustering approaches, we do not need to specify the number of clusters in advance, however, we need to specify a termination condition of the clustering process. In practice, the clustering iteration stops at the step that provides maximum cluster separation.

The Clustering Linear Combination (CLC) method is a recently developed approach which combines individual test statistics for joint analysis of multiple phenotypes in association analyses (Sha, Wang, Zhang, & Zhang, 2018). The CLC method works particularly well with phenotypes that have natural groupings. In the CLC method, one first clusters individual statistics from the association tests for each phenotype into positively correlated clusters, then the CLC test statistic was developed to combine the individual statistics linearly within each cluster and combine the between-cluster terms in a quadratic form. It was theoretically proved that if the individual statistics can be clustered correctly, the CLC test statistic is the most powerful test among all tests with certain quadratic forms. Due to the unknown number of clusters for a given data, the final test statistic of CLC method is the minimum p-value among all p-values of the CLC test statistics obtained from each possible number of clusters. Therefore, a simulation procedure must be used to evaluate the p-value of the final test statistic. This makes CLC method computationally demanding.

In this paper, we use a stopping criterion to determine the number of clusters in the clustering step. Using the clusters identified by the stopping criterion, the test statistic of CLC has an asymptotic distribution, which is very computationally efficient and makes this approach applicable for genome-wide association studies (GWAS). We have named our method hierarchical clustering CLC (HCLC). Extensive simulation studies have been conducted to evaluate the performance of HCLC. Five competitive methods for joint analysis of multiple phenotypes, i.e., multivariate analysis of variance (MANOVA) (Cole, Maxwell, Arvey, & Salas, 1994), joint model of multiple phenotypes (MultiPhen) (O'Reilly et al., 2012), TATES (van der Sluis, Posthuma, & Dolan, 2013), AFC (Liang, Wang, Sha, & Zhang, 2016), and CLC (Sha, Wang, Zhang, & Zhang, 2018), have been applied to compare

with our proposed method. Our results indicate that HCLC can control Type I error rates very well in all simulation scenarios and is either the most powerful method or has statistical power that is very similar to the most powerful method among the five existing methods we have compared it with. We also validate our proposed method by applying it to a COPDGene real dataset.

METHODS

Consider a sample of n unrelated individuals, where each individual has been genotyped at a genetic variant and has K potentially correlated phenotypes. Let $\mathbf{x} = (x_1, \dots, x_n)^T$ denote the genotypic score of the n individuals at the genetic variant of interest, where $x_i \in \{0, 1, 2\}$ is the number of minor alleles that the i^{th} individual carries at the genetic variant. Let $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_K)$ denote the $n \times K$ phenotype matrix, where $\mathbf{Y}_k = (y_{1k}, \dots, y_{nk})^T$ is the k^{th} phenotype of the n individuals.

First, we give a brief introduction of the CLC method previously developed by our group (Sha, Wang, Zhang, & Zhang, 2018). In the CLC method, we first apply a HC method (Liang, Wang, Sha, & Zhang, 2018) to partition the original K phenotypes into M disjoint clusters ($M = 1, 2, \dots, K$); the phenotypes within each cluster highly correlate with each other, while the phenotypes between clusters are much less likely to be correlated. More specifically, in the HC step, an agglomerative hierarchical clustering is directly applied on a phenotypic distance matrix D , whose $(i, j)^{\text{th}}$ entry d_{ij} is the distance between the i^{th} phenotype and the j^{th} phenotype. We use correlation between the two phenotypes to define the distance, $d_{ij} = 1 - \text{corr}(\mathbf{Y}_i, \mathbf{Y}_j)$, where corr denotes the correlation. We estimate the phenotypic distance matrix D through the sample correlation matrix of the K phenotypes. That is, $\hat{D} = J_K - C^s(\mathbf{Y})$, where J_K is a $K \times K$ matrix with each entry equals to 1, and $C^s(\mathbf{Y})$ is the sample correlation matrix of the phenotypes.

We define the distance between any two clusters to be the average linkage of the two clusters. For example, the distance between cluster C_m and cluster C_ℓ is calculated by the following equation:

$$D(C_m, C_\ell) = \frac{1}{|C_m| \cdot |C_\ell|} \sum_{i \in C_m, j \in C_\ell} d_{ij}, \quad (1)$$

where $|C_m|$ denotes the number of phenotypes in cluster C_m and $|C_\ell|$ denotes the number of phenotypes in cluster C_ℓ . The agglomerative hierarchical clustering starts with each phenotype as a singleton cluster and then successively merges pairs of clusters that have the smallest distance until all clusters have been merged into a single cluster that contains all phenotypes.

After applying the HC method described above to partition the original K phenotypes into L disjoint clusters ($L = 1, 2, \dots, K$), we use $T_{CLC}^L = (WT)^T (W \Sigma W^T)^{-1} (WT)$ to test the association between the K phenotypes and the genetic variant with L clusters; where $T = (T_1, \dots, T_K)^T$ and T_k is the score test statistic to test the association between the genetic variant and the k^{th} phenotype ($k = 1, \dots, K$) under the generalized linear model (Nelder and

Wedderburn, 1972), $g(E(y_{ik} | x_i)) = \beta_{0k} + \beta_{1k}x_i$; $W = B^T^{-1}$, where B is a $K \times L$ matrix with the $(k, l)^{th}$ entry denoted by b_{kl} with $b_{kl} = 1$ if the k^{th} phenotype belongs to the l^{th} cluster or otherwise $b_{kl} = 0$, and Σ is the variance-covariance matrix of T . Sha, Wang, Zhang, & Zhang (2018) showed that Σ can be estimated by the sample correlation matrix of the K phenotypes. Under the null hypothesis that none of phenotypes are associated with the genetic variant, T_{CLC}^L follows a chi-square distribution with degrees of freedom equal to L . Since for a given data set, the number of clusters of the phenotypes is unknown, in the last step of the CLC method, we use $T_{CLC} = \min_{1 \leq L \leq K} p_L$ as the final test statistic, where p_L denote the p-value of T_{CLC}^L for $L = 1, \dots, K$. Since T_{CLC} does not have an asymptotic distribution, we need to use a simulation procedure to evaluate the p-value of T_{CLC} .

Although the CLC method described above is a powerful test and can analyze different types of phenotypes, due to the unknown number of clusters for a given data, a simulation procedure must be used to evaluate the p-value of the final test statistic T_{CLC} . This makes the CLC method computationally demanding. In this paper, instead of considering all possible number of clusters in the HC step, we use a stopping criterion to determine the number of clusters. We determine the number of clusters in the HC step by maximizing the cluster separation (Bühlmann, Rütimann, van de Geer, & Zhang, 2013; Liang, Wang, sha, & Zhang, 2018). Let d_b denote the smallest distance between any two clusters in the b^{th} step of iteration ($b - 1$) in the HC step. If we let

$$\hat{b} = \underset{b \geq 1}{\operatorname{argmax}} (d_{b+1} - d_b), \quad (2)$$

then the total number of clusters yielded at the step $\hat{b}$ is our desired number of clusters.

Suppose that the number of clusters using the stopping criterion in the HC step is M . Then we can use $T_{CLC}^M = (WT)^T (W \Sigma W^T)^{-1} (WT)$ to test the association between multiple phenotypes and the genetic variant. We refer to this test statistic as HCLC. Under the null hypothesis that none of the K phenotypes are associated with the genetic variant of interest, T_{HCLC} follows a chi-square distribution with degrees of freedom equal to M . Comparing with CLC, HCLC does not need to use a simulation procedure to evaluate the p-value, so it is computationally more efficient than CLC.

COMPARISON OF METHODS

We compare the performance of our HCLC method with those of the other five existing methods: MANOVA (Cole, Maxwell, Arvey, & Salas, 1994), MultiPhen (O'Reilly et al., 2012), TATES (van der Sluis, Posthuma, & Dolan, 2013), AFC (Liang, Wang, Sha, & Zhang, 2016), and CLC (Sha, Wang, Zhang, & Zhang, 2018b). Since we previewed the CLC approach in the method section, here, we briefly introduce the other four methods as they apply to the current study.

MANOVA (Cole, Maxwell, Arvey, & Salas, 1994):

Consider a multivariate simple linear regression model: $Y = x\beta^T + e$, where β is a K -dimensional vector of coefficients with the K elements corresponding to the K phenotypes, while e is an $n \times K$ residual matrix, with each row following an independent identically distributed (i.i.d.) multivariate normal distribution with mean $\mathbf{0}$ and a constant variance-covariance matrix. To test $H_0: \beta = \mathbf{0}$, the Wilk's Lambda test statistic is commonly used, which is equivalent to the test statistic of the likelihood ratio test. Under the null hypothesis, the MANOVA test statistic has an asymptotic χ_K^2 distribution.

MultiPhen (O'Reilly et al., 2012):

MultiPhen uses the ordinal regression (also known as proportional odds logistic regression) and inverts the general linear regression model of a single phenotype on multiple genotypes. That is, MultiPhen treats the genetic variant of interest as an ordinal response variable and the multiple correlated phenotypes as regressors. A likelihood ratio test is then used to test association between the genetic variant and the phenotypes. The resulting test statistic asymptotically follows the χ_K^2 distribution

TATES (van der Sluis, Posthuma, & Dolan, 2013):

TATES combines phenotype-specific p-values obtained from standard univariate GWAS while considering the correlations between components. Denote p_k the p-value of the test statistic to test the association between the k^{th} phenotype and the genetic variant, $p_{(k)}$ the k^{th} smallest p-value among all p_k 's, where $k = 1, 2, \dots, K$. Then, the p-value of TATES is given by $\min_{1 \leq k \leq K} \left(\frac{m_e p(k)}{m_{e(k)}} \right)$, where m_e represents the effective number of independent p-values among all K p-values, and $m_{e(k)}$ represents the effective number of independent p-values among the first k smallest p-values.

AFC (Liang, Wang, Sha, & Zhang, 2016):

Using the same notations in TATES, let $p_{(k)}$ denote the k^{th} smallest p-value among all p_k 's, where $k = 1, 2, \dots, K$. Define $T_k = -\sum_{i=1}^k (\log p_{(i)})$, where $k = 1, 2, \dots, K$, and let p_{T_k} denote the p-value of T_k . The statistic of AFC is given by $T_{AFC} = \min_{1 \leq k \leq K} p_{T_k}$. A permutation procedure is then used to evaluate the p-values of T_{AFC} .

SIMULATION STUDIES**Simulation settings**

In order to evaluate the Type I error rates and statistical power of our proposed method, we simulate genotype and phenotype data for n unrelated individuals. The genotype of each individual at a variant of interest is generated based on the minor allele frequency (MAF) assuming Hardy Weinberg equilibrium; the phenotypes of each individual are generated according to the following factor model used by van der Sluis, Posthuma, & Dolan (2013), Yan et al. (2013), Wang, Sha, Zhang (2016), and Zhu, Zhang, & Sha (2018)

$$\mathbf{y} = \lambda x + c\gamma\mathbf{f} + \sqrt{1 - c^2} \times \boldsymbol{\varepsilon} \quad (3)$$

where $\mathbf{y} = (y_1, \dots, y_K)^T$ is a K -dimensional phenotype vector; x is the genotypic score at a genetic variant; $\lambda = (\lambda_1, \dots, \lambda_K)^T$ is the effect sizes of the genetic variant on the K phenotypes; c is a constant; γ is a $K \times R$ block diagonal matrix used for setting up various simulation scenarios; $\mathbf{f}$ is a vector of factors and $\mathbf{f} = (f_1, \dots, f_R)^T \sim MVN_R(\mathbf{0}, \Sigma)$, where R is the number of factors, $\Sigma = (1 - \rho)I + \rho J$, I is an identity matrix, J is a matrix with elements of all 1's, and ρ is the correlation between factors; $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_K)^T$ is a vector of residuals, where $\varepsilon_1, \dots, \varepsilon_K$ are pairwise independent and each $\varepsilon_k \sim N(0, 1)$ for $k = 1, \dots, K$. Based on equation (3), we consider the following four models for which the within-factor correlation equals c^2 , the between-factor correlation equals ρc^2 .

Model 1: There is only one factor and the genetic variant impacts all of the K phenotypes but with different effect sizes. That is, $R = 1$, $\lambda = \beta(1, 2, \dots, K)^T$, and $\gamma = (1, \dots, 1)^T$.

Model 2: There are two factors and the genetic variant impacts one of the factors. That is, R

$= 2$, $\lambda = \left(0, \dots, 0, \underbrace{\beta, \dots, \beta}_{K/2}\right)^T$, and $\gamma = \text{bdiag}(D_1, D_2)$, where “bdiag” indicates the block diagonal matrix, and $D_i = \left(\underbrace{1, \dots, 1}_{K/2}\right)^T$ for $i = 1, 2$.

Model 3: There are five factors and the genetic variant impacts two of the factors. That is, $R = 5$, $\lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T$, and $\gamma =$

$\text{bdiag}(D_1, D_2, D_3, D_4, D_5)$, where $D_i = \left(\underbrace{1, \dots, 1}_{K/5}\right)^T$ for $i = 1, \dots, 5$, $k = \frac{K}{5}$,

$\beta_{11} = \dots = \beta_{1k} = \beta_{21} = \dots = \beta_{2k} = \beta_{31} = \dots = \beta_{3k} = 0$, $\beta_{41} = \dots = \beta_{4k} = -\beta$, and

$(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1} * (1, \dots, k)$.

Model 4: There are five factors and the genetic variant impacts four of the factors. That is, $R = 5$, $\lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T$, and $\gamma =$

$\text{bdiag}(D_1, D_2, D_3, D_4, D_5)$, where $D_i = \left(\underbrace{1, \dots, 1}_{K/5}\right)^T$ for $i = 1, \dots, 5$, $k = \frac{K}{5}$, $\beta_{11} = \dots = \beta_{1k} = 0$,

$\beta_{21} = \dots = \beta_{2k} = \beta$, $\beta_{31} = \dots = \beta_{3k} = -\beta$, $\beta_{41} = \dots = \beta_{4k} = -\frac{2\beta}{k+1} * (1, \dots, k)$, and

$(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1} * (1, \dots, k)$.

To evaluate type I error rates, we let $\beta = 0$, MAF = 0.3, and vary significance level α , total number of phenotypes K , and sample size n . To evaluate statistical power, we set sample size n to be 5,000, MAF = 0.3, and vary the values of β , the total number of phenotypes K , within-factor correlation c^2 , and between-factor correlation ρc^2 .

Simulation results

In each simulation scenario, the p-values of the test statistics of MANOVA, MultiPhen, TATES, and HCLC are estimated based on their asymptotic distributions. The p-values of AFC and CLC are estimated using 10,000 permutations.

We use 10,000 replicated samples to evaluate the type I error. For 10,000 replicated samples, the 95% confidence intervals (CIs) for type I error rates at the nominal levels 0.05, 0.01, and 0.001 are (0.0457, 0.0543), (0.0080, 0.0120), and (0.0004, 0.0016), respectively. Table 1 summarizes the estimated type I error rates of HCLC. The results indicate that all of the estimated type I error rates are within the 95% CIs, which confirms that HCLC is a valid method.

For power comparisons, we use a 5% significance level. The power of each method is estimated under four different models using 1,000 replicated samples. Figure 1 and Figure 2 provide the power comparisons of the six methods (MANOVA, MultiPhen, TATES, AFC, CLC, and HCLC) as a function of genetic effect size β with 20 phenotypes and 40 phenotypes, respectively. From these two figures, we can see that HCLC is the most powerful test among all six methods we compared with under the four models except for model 4 when $\beta = 0.03$. CLC is the second most powerful method among the six methods under models 3 and 4. MANOVA and MultiPhen have similar statistical power under all four models and they are the second most powerful method under models 1 and 2. TATES and AFC have similar power under models 1 to 3. Under model 4, AFC performs better than TATES. TATES and AFC are the least powerful methods under model 1 and perform better than MANOVA and MultiPhen under model 2. As anticipated, when the effect size β increases, the statistical powers of all six methods increase.

Figure 3 and Figure 4 provide the power comparisons of the six methods as a function of the within-factor correlation c^2 with 20 and 40 phenotypes, respectively. The pattern of the powers of the six models are similar to these observed in Figures 1 and 2 except that under model 1, when the within-factor correlation c^2 is small (<0.3), AFC and CLC are more powerful than the other four methods. In general, the powers of all six methods decreases as the within-factor correlation c^2 increases.

We also provide the power comparisons of the six methods as a function of the between-factor correlation ρc^2 with 20 and 40 phenotypes, respectively (Figures S1 and S2). The pattern of the powers of the six models are similar to these observed in Figures 1 and 2. When the between-factor correlation ρc^2 increases, the statistical powers of the six methods almost all stay at the same levels, indicating that these methods are not sensitive to the changes in the between-factor correlation.

REAL DATA ANALYSIS

COPD data set

Chronic Obstructive Pulmonary Disease (COPD) is a chronic inflammatory lung disease that obstructs airflow (Chu et al., 2014). Possible signs and symptoms of COPD include breathing difficulty, cough, mucus (sputum) production, and wheezing (Chung & Pavord,

2008). People with COPD have a higher risk of developing heart disease, lung cancer, and numerous other afflictions. Though cigarette smoking is commonly recognized as a trigger for COPD, genetic risk factors also seem to play an important role in the development of the disease (Mannino & Buist, 2007; Pillai et al., 2009; Regan et al., 2011; Sandford, Weir, & Pare, 1997; Schellenberg et al., 1998; Silverman et al., 1998; Silverman & Weiss, 2004). The COPDGene is one of the largest studies to uncover the underlying genetic factors of COPD and other smoking-related diseases and to use this important information to develop new therapeutic approaches to cure these diseases. There was a total of 10,192 smokers who were potentially affected by COPD, of which 6,784 were non-Hispanic white and 3,408 were African-American, recruited for the COPDGene study (Chu et al., 2014).

In this paper, we apply six methods, MANOVA, MultiPhen, TATES, AFC, CLC, and HCLC, to the non-Hispanic white cohort of the COPDGene study to discover genetic variants associated with COPD-related phenotypes. Following a similar study (Liang, Wang, Sha, & Zhang, 2016) and the literature on which it was based, we select seven quantitative COPD-related phenotypes and four covariates. The seven phenotypes are % predicted FEV1 (FEV1), Emphysema (Emph), Emphysema Distribution (EmphDist), Gas Trapping (GasTrap), Airway Wall Area (Pi10), Exacerbation Frequency (ExacerFreq), Six-minute Walk Distance (6MWD). With reference to a previous study (Chu et al., 2014), we perform a log transformation on the phenotype EmphDist. The correlation plot of the seven phenotypes are given in Figure S3. We also change the signs of phenotypes FEV1 and 6MWD because their correlations with the other five phenotypes are negative. After this modification, the pair-wise correlations between phenotypes are all positive. The four covariates considered in this study are BMI, Age, Pack-Years (PackYear), and Sex. We eliminate participants with missing SNP and missing values in any of the 11 variables. After the data preprocessing steps, there remains a total of 5,430 subjects and 630,860 SNPs. For each of the seven phenotypes, we adjust the phenotype values for the four covariates through a linear regression (Sha, Wang, Zhang, & Zhang, 2018). All the subsequent analyses are based on the adjusted phenotypes. To identify SNPs associated with the seven COPD-related phenotypes, we use the standard genome-wide significance p-value threshold of 5×10^{-8} to account for multiple testing.

Real data analysis results

In the COPDGene real data analysis, HC divides the seven phenotypes into five clusters, that is, $M=5$. The dendrogram of HC on the seven phenotypes is given in Figure S4. One of the clusters contains phenotypes FEV1, Emph, and GasTrap, while the other four clusters each contain only a single phenotype. This finding is aligned with a previous study, which showed that GasTrap is a “hub” in the phenotypic network; the pairings of GasTrap with FEV1 and GasTrap with Emph are both highly correlated in the race-specific networks (Chu et al., 2014). Table 2 summarizes the 14 significant SNPs that have been identified by at least one of the six methods. All of these 14 SNPs were previously reported to be associated with COPD (Brehm et al., 2011; Cho et al., 2010; Cho et al., 2014; Cui, Ge, & Ma, 2014; Du, Xue, & Xiao, 2016; Hancock et al., 2010; Li et al., 2011; Liang, Sha, Rho, & Zhang, 2018; Lutz et al., 2015; Pillai et al., 2009; Sha, Wang, Zhang, & zhang, 2018; Wilk et al., 2009; Wilk et al., 2012; Young et al., 2010; Zhang, Summah, Zhu, & Qu, 2011; Zhu et al., 2014).

As seen in Table 2, MultiPhen identifies 14 SNPs; HCLC, CLC, and MANOVA each identifies 13 SNPs; AFC identifies 12 SNPs; and TATES identifies nine SNPs. The reason that TATES only identifies nine SNPs may be due to the fact that this method depends heavily on the smallest p-value of the seven univariate analysis results. The results of the real data analysis are consistent with the findings from our simulations, corroborating that the HCLC has similar or superior performance to that of the other five methods.

DISCUSSION

In this paper, we develop a novel method, HCLC, for joint analysis of multiple phenotypes in genetic association studies. HCLC has several important advantages over existing methods. First of all, HCLC takes advantage of the natural grouping information of the phenotypes, which can be easily obtained from hierarchical clustering. In addition, HCLC is easy to implement and computationally efficient for GWAS. HCLC avoids the computational burden of AFC and CLC, which use permutation to evaluate the p-values of their test statistics; instead, HCLC has an asymptotic distribution. Furthermore, HCLC does not require access to individual phenotypes themselves; it only requires a distance matrix of phenotypes. When individual phenotype data is not available, this distance matrix of phenotypes can be estimated from the summary statistics of univariate GWAS (Zhu et al., 2015).

Our simulation results demonstrate that HCLC has controlled type I error rates effectively and almost always exceeds the other five competing methods in terms of statistical power within various simulation scenarios. Additionally, the real data analysis results indicate that HCLC has great potential for GWAS.

In this study, we use the bottom-up hierarchical clustering method to cluster phenotypes in the HC step. For future studies, we can explore other clustering approaches and incorporate the corresponding clustering information into the clustering step.

In addition, we use correlation to define phenotypic distances in the study. We can also use the Euclidean distance, maximum distance, and absolute distance, among others to define phenotypic distances. We choose the average linkage as the distance between two clusters in this study. In fact, we can also choose other linkages; for example, single linkage and complete linkage. However, the outcome of using other distances and applying other linkages within our proposed method still needs further investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R15HG008209. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This research used data generated by the COPDGene study (phs000179/HMB and phs000179/DS-CS-RD), which was supported by National Institutes of Health (NIH) grants U01HL089856 and U01HL089897. 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. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board comprised of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion.

Superior, a high-performance computing infrastructure at Michigan Technological University, was used in obtaining results presented in this publication.

REFERENCES

- Aschard H, Vilhjalmsdottir BJ, Greliche N, Morange PE, Tregouet DA, & Kraft P (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet*, 94(5), 662–676. doi:10.1016/j.ajhg.2014.03.016 [PubMed: 24746957]
- Brehm JM, Hagiwara K, Tesfaigzi Y, Bruse S, Mariani TJ, Bhattacharya S, ..., & Avila L (2011). Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease. *Thorax*, 66(12), 1085–1090. [PubMed: 21921092]
- Bühlmann P, Rütimann P, van de Geer S, & Zhang CH (2013). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11), 1835–1858.
- Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, ..., & Sparrow D (2010). Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature genetics*, 42(3), 200. [PubMed: 20173748]
- Cho MH, McDonald M-LN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, ..., & Laird NM (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The lancet Respiratory medicine*, 2(3), 214–225. [PubMed: 24621683]
- Chu JH, Hersh CP, Castaldi PJ, Cho MH, Raby BA, Laird N, ..., & Silverman EK (2014). Analyzing networks of phenotypes in complex diseases: methodology and applications in COPD. *BMC Systems Biology*, 8(1), 78. [PubMed: 24964944]
- Chung KF, & Pavord ID (2008). Prevalence, pathogenesis, and causes of chronic cough. *The Lancet*, 371(9621), 1364–1374.
- Cole DA, Maxwell SE, Arvey R, & Salas E (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin*, 115(3), 465.
- Cui K, Ge X, & Ma H (2014). Four SNPs in the CHRNA3/5 alpha-neuronal nicotinic acetylcholine receptor subunit locus are associated with COPD risk based on meta-analyses. *PLoS ONE*, 9(7), e102324. [PubMed: 25051068]
- Ding C, & He X (2002). Cluster merging and splitting in hierarchical clustering algorithms. Paper presented at the Data Mining, 2002. ICDM 2003. Proceedings 2002 IEEE International Conference on.
- Du Y, Xue Y, & Xiao W (2016). Association of IREB2 gene rs2568494 polymorphism with risk of chronic obstructive pulmonary disease: a meta-analysis. *Medical science monitor: international medical journal of experimental and clinical research*, 22, 177. [PubMed: 26775557]
- Ferreira MA, & Purcell SM (2008). A multivariate test of association. *Bioinformatics*, 25(1), 132–133. [PubMed: 19019849]
- Gratten J, & Visscher PM (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine*, 8, 78. doi:10.1186/s13073-016-0332-x. [PubMed: 27435222]
- Han MK, Agustí A, Calverley PM, Celli BR, Criner G, Curtis JL, ..., & MacNee W (2010). Chronic obstructive pulmonary disease phenotypes: the future of COPD. *American journal of respiratory and critical care medicine*, 182(5), 598–604. [PubMed: 20522794]
- Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marcianti KD, ..., & Barr RG (2010). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics*, 42(1), 45. [PubMed: 20010835]
- Johnson SC (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254. [PubMed: 5234703]

- Karypis G, Han EH, & Kumar V (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8), 68–75.
- Kim J, Bai Y, & Pan W (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic epidemiology*, 39(8), 651–663. [PubMed: 26493956]
- Klei L, Luca D, Devlin B, & Roeder K (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(1), 9–19.
- Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, & Nordborg M (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9), 1066. [PubMed: 22902788]
- Li X, Howard TD, Moore WC, Ampleford EJ, Li H, Busse WW, ..., & Erzurum SC (2011). Importance of hedgehog interacting protein and other lung function genes in asthma. *Journal of Allergy and Clinical Immunology*, 127(6), 1457–1465. [PubMed: 21397937]
- Liang X, Sha Q, Rho Y, & Zhang S (2018). A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genetic epidemiology*, 42(4), 344–353. [PubMed: 29682782]
- Liang X, Wang Z, Sha Q, & Zhang S (2016). An adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies. *Scientific reports*, 6, 34323. [PubMed: 27694844]
- Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald M-L, ..., & Parker M (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC genetics*, 16(1), 138. [PubMed: 26634245]
- Mannino DM, & Buist AS (2007). Global burden of COPD: risk factors, prevalence, and future trends. *The Lancet*, 370(9589), 765–773.
- Nelder JA & Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*;135(3):370–384.
- O'Brien PC (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4), 1079–1087. doi:10.2307/2531158 [PubMed: 6534410]
- O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin M-R, & Coin LJM (2012). MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*, 7(5), e34861. doi:10.1371/journal.pone.0034861 [PubMed: 22567092]
- Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, ..., & Gulsvik A (2009). A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*, 5(3), e1000421. [PubMed: 19300482]
- Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, ..., & Crapo JD (2011). Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1), 32–43.
- Rokach L, & Maimon O (2005). Clustering methods In *Data mining and knowledge discovery handbook* (pp. 321–352): Springer.
- Sandford A, Weir T, & Pare PD (1997). Genetic risk factors for chronic obstructive pulmonary disease. *European Respiratory Journal*, 10(6), 1380–1391. [PubMed: 9192947]
- Schellenberg D, Pare PD, Weir TD, Spinelli JJ, Walker BA, & Sandford AJ (1998). Vitamin D binding protein variants and the risk of COPD. *American journal of respiratory and critical care medicine*, 157(3), 957–961. [PubMed: 9517617]
- Sha Q, Wang Z, Zhang X, & Zhang S (2018). A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *Bioinformatics*, 35(8): 1373–1379. 10.1093/bioinformatics/bty810
- Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, ..., & Mentzer S (1998). Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease: risk to relatives for airflow obstruction and chronic bronchitis. *American journal of respiratory and critical care medicine*, 157(6), 1770–1778. [PubMed: 9620904]
- Silverman EK, & Weiss ST (2004). Risk factors for the development of COPD In *Long-Term Intervention in Chronic Obstructive Pulmonary Disease* (pp. 81–98): CRC Press.

- Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, ..., & Campbell H (2011). Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet*, 89(5), 607–618. doi:10.1016/j.ajhg.2011.10.004 [PubMed: 22077970]
- Stearns FW (2010). One hundred years of pleiotropy: a retrospective. *Genetics*, 186(3), 767–773. doi:10.1534/genetics.110.122549 [PubMed: 21062962]
- Tang CS, & Ferreira MA (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics*, 28(6), 845–850. doi:10.1093/bioinformatics/bts051 [PubMed: 22296789]
- van der Sluis S, Posthuma D, & Dolan CV (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, 9(1), e1003235. doi:10.1371/journal.pgen.1003235 [PubMed: 23359524]
- Wang Z, Sha Q, & Zhang S (2016). Joint analysis of multiple traits using “optimal” maximum heritability test. *PLoS ONE*, 11(3), e0150975. [PubMed: 26950849]
- Wei L, & Johnson WE (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika*, 72(2), 359–364.
- Wienke A (2010). *Frailty models in survival analysis*: CRC Press.
- Wilk JB, Chen T. h., Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, ..., & Weiss ST (2009). A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet*, 5(3), e1000429. [PubMed: 19300500]
- Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, ..., & Tang W (2012). Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *American journal of respiratory and critical care medicine*, 186(7), 622–632. [PubMed: 22837378]
- Yan T, Li Q, Li Y, Li Z, & Zheng G (2013). Genetic association with multiple traits in the presence of population stratification. *Genetic epidemiology*, 37(6):571–80. doi: 10.1002/gepi.21738. [PubMed: 23740720]
- Yang JJ, Li J, Williams LK, & Buu A (2016). An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC bioinformatics*, 17(1), 19. [PubMed: 26729364]
- Yang Q, & Wang Y (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics*, 2012, 13. doi:10.1155/2012/652569.
- Yang X, Zhang S, & Sha Q (2019). Joint analysis of multiple phenotypes in association studies based on cross-validation prediction error. *Scientific Reports*, 9, 1073. [PubMed: 30705317]
- Young R, Whittington C, Hopkins R, Hay B, Epton M, Black P, & Gamble G (2010). Chromosome 4q31 locus in COPD is also associated with lung cancer. *European Respiratory Journal*, 36(6), 1375–1382. [PubMed: 21119205]
- Zeger SL, & Liang KY (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121–130. [PubMed: 3719049]
- Zhang J, Summah H, Zhu YG, & Qu JM (2011). Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis. *Respiratory research*, 12(1), 158. [PubMed: 22176972]
- Zhang Y, Xu Z, Shen X, Pan W, & the Alzheimer’s Disease Neuroimaging Initiative (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96, 309–325. [PubMed: 24704269]
- Zhang Z, Buckler ES, Casstevens TM, & Bradbury PJ (2009). Software engineering the mixed model for genome-wide association studies on large samples. *Briefings in Bioinformatics*, 10(6), 664–675. doi:10.1093/bib/bbp050 [PubMed: 19933212]
- Zhou JJ, Cho MH, Lange C, Lutz S, Silverman EK, & Laird NM (2015). Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Human heredity*, 79(2), 93–104. [PubMed: 26111731]
- Zhou X, & Stephens M (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4), 407. [PubMed: 24531419]
- Zhu AZ, Zhou Q, Cox LS, David SP, Ahluwalia JS, Benowitz NL, & Tyndale RF (2014). Association of CHRNA5-A3-B4 SNP rs2036527 With Smoking Cessation Therapy Response in African-American Smokers. *Clinical Pharmacology & Therapeutics*, 96(2), 256–265. [PubMed: 24733007]

- Zhu H, Zhang S, & Sha Q (2015). Power comparisons of methods for joint association analysis of multiple phenotypes. *Human heredity*, 80(3), 144–152. [PubMed: 27344597]
- Zhu H, Zhang S, & Sha Q (2018). A novel method to test associations between a weighted combination of phenotypes and genetic variants. *PLoS ONE*, 13(1), e0190788. [PubMed: 29329304]
- Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, ..., & Edwards TL (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics*, 96(1), 21–36. [PubMed: 25500260]

Figure 1.

The power comparisons of the six methods (MANOVA, MultiPhen, TATES, AFC, CLC and HCLC) for 20 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the effect size β , where MAF is 0.3, the sample size is 5,000, and the number of replications is 1,000, the within-factor correlation is 0.5 ($c^2 = 0.5$), and the between-factor correlation is 0.1 ($\rho c^2 = 0.1$).

Figure 2.

The power comparisons of the six methods (MANOVA, MultiPhen, TATES, AFC, CLC and HCLC) for 40 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the effect size β , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, the within-factor correlation is 0.5 ($c^2 = 0.5$), and the between-factor correlation is 0.1 ($\rho c^2 = 0.1$).

Figure 3.

The power comparisons of the six methods (MANOVA, MultiPhen, TATES, AFC, CLC and HCLC) for 20 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the within-factor correlation c^2 , where MAF is 0.3, the sample size is 5,000, and the number of replications is 1,000, and the between-factor correlation is 0.1 ($\rho c^2 = 0.1$).

Figure 4.

The power comparisons of the six methods (MANOVA, MultiPhen, TATES, AFC, CLC and HCLC) for 40 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the within-factor correlation c^2 , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, the between-factor correlation is 0.1 ($\rho c^2 = 0.1$).

Table 1.

The estimated type I error rates of HCLC, where MAF is 0.3, the number of replications is 10,000. K is the number of phenotypes, and α is the significance level. All estimated type I error rates are within the corresponding 95% CIs.

K	α	Sample Size	Model			
			1	2	3	4
20	0.050	2000	0.0473	0.0528	0.0516	0.0477
		5000	0.0505	0.0482	0.0520	0.0493
	0.010	2000	0.0099	0.0114	0.0104	0.0100
		5000	0.0111	0.0099	0.0104	0.0106
	0.001	2000	0.0008	0.0007	0.0011	0.0011
		5000	0.0014	0.0011	0.0009	0.0010
40	0.050	2000	0.0497	0.0481	0.0489	0.0498
		5000	0.0484	0.0495	0.0520	0.0470
	0.010	2000	0.0096	0.0083	0.0098	0.0113
		5000	0.0090	0.0114	0.0110	0.0099
	0.001	2000	0.0004	0.0010	0.0005	0.0011
		5000	0.0008	0.0012	0.0004	0.0009

Table 2.

Significant SNPs and the corresponding p-values in the COPDGene real data analysis. The p-values of MANOVA, MultiPhen, TATES, and HCLC are estimated based on their asymptotic distributions and the p-values of AFC and CLC are estimated using 10,000 permutations. The graying out p-values indicate the p-values greater than 5×10^{-8} .

Chr	Position	Variant Identifier	MANOVA	MultiPhen	TATES	AFC	CLC	HCLC
4	145431497	rs1512282	1.69×10^{-9}	1.03×10^{-9}	5.77×10^{-9}	1.10×10^{-8}	0	5.28×10^{-10}
4	145434744	rs1032297	6.52×10^{-14}	7.69×10^{-14}	6.22×10^{-13}	0	0	1.17×10^{-13}
4	145474473	rs1489759	1.11×10^{-16}	1.22×10^{-16}	2.52×10^{-16}	0	0	0
4	145485738	rs1980057	6.68×10^{-17}	8.14×10^{-17}	9.35×10^{-17}	0	0	0
4	145485915	rs7655625	7.12×10^{-17}	9.13×10^{-17}	1.64×10^{-16}	0	0	0
15	78882925	rs16969968	1.32×10^{-11}	7.84×10^{-12}	2.98×10^{-8}	0	0	6.18×10^{-11}
15	78894339	rs1051730	1.41×10^{-11}	8.16×10^{-12}	2.63×10^{-8}	0	0	3.18×10^{-11}
15	78898723	rs12914385	1.76×10^{-12}	1.48×10^{-12}	5.14×10^{-10}	0	0	1.09×10^{-12}
15	78911181	rs8040868	2.74×10^{-12}	2.59×10^{-12}	2.40×10^{-9}	0	0	2.96×10^{-12}
15	78878541	rs951266	1.77×10^{-11}	1.02×10^{-11}	5.17×10^{-8}	0	0	6.36×10^{-11}
15	78806023	rs8034191	2.14×10^{-10}	7.74×10^{-11}	1.02×10^{-7}	1.40×10^{-8}	0	8.08×10^{-10}
15	78851615	rs2036527	3.99×10^{-10}	1.77×10^{-10}	1.56×10^{-7}	2.90×10^{-8}	8.33×10^{-10}	1.21×10^{-9}
15	78826180	rs931794	2.35×10^{-10}	9.09×10^{-11}	1.18×10^{-7}	6.30×10^{-8}	0	3.63×10^{-9}
15	78740964	rs2568494	1.05×10^{-7}	4.23×10^{-8}	2.88×10^{-5}	5.00×10^{-6}	3.98×10^{-7}	1.23×10^{-6}