Published in final edited form as:

Genet Epidemiol. 2016 September ; 40(6): 512–519. doi:10.1002/gepi.21987.

A Clustered Multiclass Likelihood-Ratio Ensemble Method for Family-Based Association Analysis Accounting for Phenotypic Heterogeneity

Yalu Wen¹ and Qing Lu^{2,*}

¹Department of Statistics, University of Auckland, Auckland, New Zealand

²Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, United States of America

Abstract

Although compelling evidence suggests that the genetic etiology of complex diseases could be heterogeneous in subphenotype groups, little attention has been paid to phenotypic heterogeneity in genetic association analysis of complex diseases. Simply ignoring phenotypic heterogeneity in association analysis could result in attenuated estimates of genetic effects and low power of association tests if subphenotypes with similar clinical manifestations have heterogeneous underlying genetic etiologies. To facilitate the family-based association analysis allowing for phenotypic heterogeneity, we propose a clustered multiclass likelihood-ratio ensemble (CMLRE) method. The proposed method provides an alternative way to model the complex relationship between disease outcomes and genetic variants. It allows for heterogeneous genetic causes of disease subphenotypes and can be applied to various pedigree structures. Through simulations, we found CMLRE outperformed the commonly adopted strategies in a variety of underlying disease scenarios. We further applied CMLRE to a family-based dataset from the International Consortium to Identify Genes and Interactions Controlling Oral Clefts (ICOC) to investigate the genetic variants and interactions predisposing to subphenotypes of oral clefts. The analysis suggested that two subphenotypes, nonsyndromic cleft lip without palate (CL) and cleft lip with palate (CLP), shared similar genetic etiologies, while cleft palate only (CP) had its own genetic mechanism. The analysis further revealed that rs10863790 (*IRF6*), rs7017252 (*8q24*), and rs7078160 (*VAX1*) were jointly associated with CL/CLP, while rs7969932 (*TBKI*), rs227731 (*17q22*), and rs2141765 (*TBKI*) jointly contributed to CP.

Keywords

nonparametric method; subphenotypes; family-based study; oral clefts

*Corresponding to: Qing Lu, 909 Fee Road, Room 601, East Lansing, MI 48824. qlu@epi.msu.edu.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

Introduction

Benefiting from the fast development of biotechnologies and the completion of the HapMap project, great progress has been made through genome-wide association studies (GWAS) in mapping genetic loci associated with complex human diseases, such as type II diabetes (T2D) [Morris et al., 2010; Zeggini et al., 2007], oral clefts, and cardiovascular diseases [Ellis et al., 2010; Qi et al., 2011]. Despite the successes made by GWAS, the effect sizes of genetic variants in most of these studies tend to be small and a significant proportion of genetic contributions to complex human diseases remains unknown [Ku et al., 2010; Manolio, 2010]. Phenotypic heterogeneity may be one of the reasons that explain small marginal effects of identified genetic variants. For example, subphenotypes of bladder cancer defined by the pathological and molecular features have their own genetic etiologies [Guey et al., 2010]. When heterogeneous subphenotypes are treated as a single entity, the effect sizes of the associated variants could be substantially reduced, leading to undetectable or small marginal effects. The use of more refined subphenotypes of diseases defined based upon disease symptoms, severity of illness, and age at onset could facilitate the identification of new genetic variants contributing to each subphenotype, which ultimately shed light on the genetic etiology of complex diseases [Morris et al., 2010]. For instance, a recent GWAS of T2D with cases categorized by body mass index, found heterogeneous genetic effects of *TCF7L2* and *FTO* between obese and nonobese forms of T2D, as well as revealed novel single nucleotide polymorphisms (SNPs) predisposing to the obese T2D [Morris et al., 2010].

Different from a typical association analysis that treats all subphenotypes as a single entity, subphenotype analysis (i.e., separately analyzing each subphenotype) investigates genetic variants predisposing to each subphenotype. However, the disadvantage of subphenotype analysis is that the sample is divided into small-size subphenotype groups, which could substantially reduce the power of association tests. Alternatively, we could use a predefined cutoff to determine the subphenotype groups (e.g., severity of illness) [Lu et al., 2009]. However, such a cutoff is chosen based on prior experience, which may not reflect the underlying genetic mechanism. It is quite possible that the resulted subphenotype groups are different from the underlying subphenotype groups of interest, resulting in either overestimating or underestimating the underlying distinct subphenotype groups.

Although it is important to consider phenotypic heterogeneity in genetic association analysis, there currently lacks the development of statistical methods for such purposes, especially for family-based data. In addition, complex diseases are likely influenced by the interplay of multiple genetic variants. Considering gene-gene interactions in the analysis could enhance our power to detect genetic variants predisposing to a specific subphenotype of a disease. Many methods have been proposed to detect gene-gene interactions that contribute to complex human diseases in family-based studies. For example, conditional logistic regression and generalized estimating equations [Liang and Zeger, 1986] are two classic methods to study gene-gene interactions in family-based studies and have been implemented in standard statistical software [Hancock et al., 2007]. In addition to these conventional methods, Martin et al. proposed a nonparametric multifactor dimensionality reduction (MDR) pedigree disequilibrium test (PDT) method, which incorporates the MDR

method into the PDT method for the identification of gene-gene interactions in family-based studies [Martin et al., 2006]. To consider the covariate adjustment, Lou et al. modified MDR-PDT and developed a pedigree-based generalized MDR method [Lou et al., 2008]. These methods have the advantages of being robust to the underlying mode of inheritance and disease models. Nonetheless, none of them is designed to consider phenotypic heterogeneity when studying gene-gene interactions.

To consider phenotypic heterogeneity in family-based association studies, we develop a clustered multiclass likelihood-ratio ensemble (CMLRE) method to simultaneously determine the underlying homogeneous subphenotype groups and identify gene-gene interactions predisposing to each homogeneous group. The CMLRE method is built based upon a forward U -statistic, which is capable of capturing high-order gene-gene interactions [Lu et al., 2011]. The proposed method does not require prior knowledge of disease etiology of subphenotypes, and can be applied to various pedigree structures. Therefore, it provides a powerful and flexible framework to model the complex relationship between phenotypes and genetic variants in family-based studies, allowing for heterogeneous genetic causes of each subphenotype. In the following sections, we first lay out the details of the CMLRE method, and then conduct simulation studies to evaluate the performance of the method. Finally, we apply the new method to a large-scale family-based dataset from the *ICOC* study to investigate the gene-gene interactions predisposing to subphenotypes of oral clefts.

Method

The CMLRE method searches for the optimal subphenotype groups that reflect the underlying distinct subphenotype groups, and then tests the association of genetic variants and their interactions with the determined subphenotype groups using a U -statistic.

Association Test

Let's consider a family-based association study with K initial subphenotype groups. We are interested in identifying genetic variants predisposing to these subphenotypes. Let N denote the total number of individuals from M families, and $Z = (Z_1, Z_2, \dots, Z_L)$ denote the genotypes at the L markers. We further use Y to denote phenotype status, with $Y = 0$ for controls and $Y = k, k = 1, 2, \dots, K$ for cases with k th subphenotype. Suppose there are S_0 optimal distinct subphenotype groups (the detailed selection algorithm is described in the next section). For the s th subphenotype group, I_s SNPs are selected based on a forward U -statistic [Lu et al., 2011], which form n_s multilocus genotypes with the consideration of possible gene-gene interactions. Let $G_{ij}^s = (G_{ij1}, G_{ij2}, \dots, G_{ijn_s})$ denote the n_s multilocus geno-types for the j th individual in the i th family. Let $LR^{s,a}$ and $LR^{s,u}$, respectively, denote the likelihood ratio (defined as $LR^s = P(G_{ij}^s | Y_{ij} = s) / P(G_{ij}^s | Y_{ij} = 0)$) for subjects in the s th subphenotype group and controls. The test statistic to test the association of I_s selected SNPs and their interactions (i.e., via n_s multilocus genotypes) with the s th subphenotype can be formed based on a U -statistic [Obuchowski, 1997],

$$U_{ss} = \frac{1}{N_s U} \sum_{i=1}^M \sum_{i'=1}^M \sum_{j=1}^{a_i^s} \sum_{j'=1}^{u_{i'}} \psi \left[LR^{s,a} \left(G_{ij}^s \right), LR^{s,u} \left(G_{i'j'}^s \right) \right], \quad (1)$$

where a_i^s and $u_{i'}$, respectively, denote the number of individuals with subphenotype s and the number of controls in the family i ; $N_s = \sum_{i=1}^M a_i^s$, $U = \sum_{i=1}^M u_{i'}$, and

$$\psi \left[LR^{s,a} \left(G_{ij}^s \right), LR^{s,u} \left(G_{i'j'}^s \right) \right] = \begin{cases} 1 & \text{if } LR^{s,a} \left(G_{ij}^s \right) < LR^{s,u} \left(G_{i'j'}^s \right) \\ 0.5 & \text{if } LR^{s,a} \left(G_{ij}^s \right) = LR^{s,u} \left(G_{i'j'}^s \right) \\ 0 & \text{if } LR^{s,a} \left(G_{ij}^s \right) > LR^{s,u} \left(G_{i'j'}^s \right) \end{cases}.$$

The overall test statistic for all subphenotype groups is defined as a weighted sum of the test statistic for each distinct subphenotype group,

$$U_0 = \frac{\sum_{1 \leq s \leq S_0} \omega_{ss} U_{ss}}{\sum_{1 \leq s \leq S_0} \omega_{ss}}, \quad (2)$$

where S_0 is the total number of subphenotype groups and ω_{ss} is defined as the number of the subjects in group s over the total number of subjects.

Hypothesis testing can be conducted to assess the significance of the joint association of selected SNPs with subphenotypes. Asymptotically, under the null hypothesis, $U_0 \sim N(0.5, \text{var}(U_0))$, where

$$\text{var}(U_0) = \frac{\sum_{1 \leq s \leq S} \omega_{ss}^2 \text{var}(U_{ss}) + \sum_{s, s' \in \{1, 2, \dots, S\}, s \neq s'} \text{cov}(U_{ss}, U_{s's'})}{\left(\sum_{1 \leq s \leq S} \omega_{ss} \right)^2}$$

The $\text{var}(U_{ss})$ can be estimated using a similar idea proposed by Obuchowski [1997]. The covariance between the test statistics of subphenotypes (i.e., $\text{cov}(U_{ss}, U_{s's'})$) is estimated as

$$\begin{aligned}
\text{cov}(U_{ss}, U_{s's'}) &= \frac{M/(M-1)}{N_s N_{s'}} \sum_{i=1}^M [V_{10}^s(G_{i.}^s) - a_i^s U_{ss}] \\
&\times [V_{10}^{s'}(G_{i.}^{s'}) - a_i^{s'} U_{s's'}] + \frac{M/(M-1)}{N_s U} \\
&\times \sum_{i=1}^M [V_{10}^s(G_{i.}^s) - a_i^s U_{ss}] [V_{01}^{s'}(G_{i.}^{s'}) - u_i U_{s's'}] \\
&+ \frac{M/(M-1)}{N_{s'} U} \sum_{i=1}^M [V_{10}^{s'}(G_{i.}^{s'}) - a_i^{s'} U_{s's'}] \\
&\times [V_{01}^s(G_{i.}^s) - u_i U_{ss}] + \frac{M/(M-1)}{U^2} \\
&\times \sum_{i=1}^M [V_{01}^s(G_{i.}^s) - u_i U_{ss}] [V_{01}^{s'}(G_{i.}^{s'}) - u_i U_{s's'}],
\end{aligned}$$

where

$$V_{01}^g(G_{i.}^g) = \frac{\sum_{j=1}^{u_i} \sum_{i'=1}^M \sum_{j'=1}^{a_{i'}^g} \psi(LR^{g,a}(G_{i'j'}^g), LR^{g,u}(G_{i,j}^g))}{N_g}$$

and

$$V_{10}^g(G_{i.}^g) = \frac{\sum_{j=1}^{a_i^g} \sum_{i'=1}^M \sum_{j'=1}^{u_{i'}} \psi(LR^{g,a}(G_{ij'}^g), LR^{g,u}(G_{i',j'}^g))}{U}$$

for $g \in \{s, s'\}$. The details of the derivation of the variance is shown in supplementary Appendix S1.

Determining the Optimal Subphenotype Groups

In practice, the underlying number of subphenotype groups is usually unknown. To estimate the optimal subphenotype groups that approximate the underlying distinct subphenotype groups, we start the algorithm with all K initial subphenotypes, gradually combine the initial subphenotypes, and determine the optimal number of subphenotype groups by using a global U -statistic. The global U -statistic is defined as

$$U_g = \frac{\sum_{1 \leq k \leq k' \leq S} \omega_{kk'} U_{kk'}}{\sum_{1 \leq k \leq k' \leq S} \omega_{kk'}}, \quad (3)$$

where

$$U_{kk'} = \frac{1}{N_k N_{k'}} \sum_{i=1}^M \sum_{i'=1}^M \sum_{j=1}^{a_i^k} \sum_{j'=1}^{a_{i'}^{k'}} \psi \left[LR_b^{kk',k} \left(G_{ij}^{kk'} \right), \right. \\ \left. LR_b^{kk',k'} \left(G_{i'j'}^{kk'} \right) \right],$$

$$LR_b^{kk'} \left(G_{ij}^{kk'} \right) = \frac{P \left(G_{ij}^{kk'} | Y_k = 1 \right)}{P \left(G_{ij}^{kk'} | Y_{k'} = 1 \right)},$$

$$G_{ij}^{kk'} \in \left\{ G_{ij}^k \cup G_{ij}^{k'} \right\},$$

and $LR_b^{kk',k}$, and $LR_b^{kk',k'}$, respectively, denote the LR for individuals with subphenotype k and subphenotype k' . ω_{ss} is defined as the number of the subjects in group s over the total number of subjects.

The rationale of using the global U -statistic (i.e., U_g) as the selection criterion is that it provides a balance between overestimating (through U_{kk}) and underestimating (through $U_{kk'}$) the underlying homogenous subphenotype groups. In the case where the total number of subphenotype groups overestimates the underlying homogenous subphenotype groups, a low U_g would be expected because $U_{kk'}$ tends to be small for those subphenotype groups sharing similar genetic etiology. On the other hand, when the total number of subphenotype groups underestimates the underlying homogeneous group (i.e., some heterogeneous subphenotypes are combined together to form a subphenotype group), U_{kk} tends to be small for those heterogeneous subphenotype groups, leading to a low U_g . Therefore, the subphenotype groups achieving the maximum U_g could identify the optimal subphenotype groups that approximate the underlying homogeneous subphenotypes.

In practice, an exhaustive evaluation of U_g for all possible subphenotype combinations could be computationally intensive. To reduce the computational burden, we adopt a greedy algorithm and combine subphenotype groups one at a time. At each step, we combine two subphenotypes with the smallest mean squared error (denoted by $MSE_{k,k'}^b$) defined as

$$MSE_{k,k'}^b = \left(\Delta \tilde{U}_{k,k'} \right)^2 + var \left(\Delta \tilde{U}_{k,k'} \right) \\ = \left(\tilde{U}_{k,k'} - \tilde{U}_{k',k} \right)^2 + var \left(\tilde{U}_{k,k'} \right) + var \left(\tilde{U}_{k',k} \right) \\ - 2cov \left(\tilde{U}_{k,k'}, \tilde{U}_{k',k} \right), \quad (4)$$

where

$$\tilde{U}_{k',k} = \frac{1}{(N_k + N_{k'})U} \sum_{i=1}^M \sum_{i'=1}^M \sum_{j=1}^{a_i^{k|k'}} \sum_{j'=1}^{u_{i'}} \psi \left[LR^{k,u} \left(G_{ij}^{k'} \right), \right. \\ \left. LR^{k,u} \left(G_{i'j'}^k \right) \right]$$

and

$$\tilde{U}_{k',k} = \frac{1}{(N_k + N_{k'})U} \sum_{i=1}^M \sum_{j=1}^M \sum_{j'=1}^{a_i^{k|k'}} \sum_{j''=1}^{u_{i'}} \psi \left[LR^{k',a} \left(G_{ij}^{k'} \right), LR^{k',u} \left(G_{i'j'}^{k'} \right) \right].$$

$\tilde{U}_{k,k'} \left(\tilde{U}_{k',k} \right)$ measures the association of SNPs selected from the k th (k' th) subphenotype with the combined sample of k th and k' th subphenotypes. $a_i^{k|k'}$ denotes the number of individuals in the family i with either subphenotype k or subphenotype k' . The derivation of $\text{var}(\tilde{U}_{k,k'})$, $\text{var}(\tilde{U}_{k',k})$, and $\text{cov}(\tilde{U}_{k,k'}, \tilde{U}_{k',k})$ can be found in supplementary Appendix S2.

If two subphenotypes belong to one underlying homogenous group, we would expect the identified genetic effects on two subphenotypes to be similar, leading to a small $MSE_{k,k'}^b$.

Therefore, at each step we combine two subphenotypes with the smallest $MSE_{k,k'}^b$ into one group, and reduce the total number of subphenotype groups by one. We repeat the clustering process, and gradually combine all subphenotype groups into one group. A 10-fold cross-validation procedure is used to determine the optimum number of groups (i.e., S_o). In the cross-validation procedure, subjects in each subphenotype and the control groups are randomly divided into 10 subsets. Nine subsets are used as a training set, while the remaining one is used as the validation set. We apply CMLRE to the training set and calculate U_g based on the validation dataset. The procedure is repeated 10 times to make sure each subset serves as the validation set once. S_o is chosen as the group that maximizes the average U_g . The test statistics U_0 (Eq. (1)) is then calculated based on the optimum homogeneous groups. If the association test is conducted on the same dataset, a gene dropping based permutation test, in which genotypes of offspring are generated by randomly transmitting each allele from parent to offspring with a probability of 0.5, is used to examine the overall significance of the association. The selection procedure as described above is applied to the permutation dataset and the corresponding test statistic (i.e., U_0) is calculated. Based on the permutation distribution of the test statistic, an empirical P -value, which takes model selection into account, can be attained. In a replication study where the multilocus genotypes and the optimal subphenotype groups are predetermined from an initial study, the overall significance of the association can be obtained from the Z -test described above.

Results

Through simulations, we evaluated the performance of the proposed method, and compared it with two commonly adopted methods, analyzing all subphenotypes together (ALL) and analyzing each subphenotype separately (SEP). In the first set of simulations, we assessed the performance of the CMLRE method with a varying number of distinct underlying subphenotype groups. For the second set of simulations, we simulated various pedigree structures and investigated their impact on the proposed method. Finally, we evaluated the performance of our method by considering different initial grouping strategies.

Simulation I

In this set of simulations, we simulated four subphenotypes with varied number of underlying homogeneous subphenotype groups. We started with the scenario where all subphenotypes shared the same genetic mechanisms, and then more complex disease scenarios in which some subphenotypes shared the same disease model. Finally, we considered a disease scenario where four subphenotypes were mutually independent. For all the scenarios, we simulated a disease model with four disease-associated loci, among which two loci followed a multiplicative interaction effect model and the other two loci followed a threshold interaction effect model. The two-locus multiplicative interaction effect model and the two-locus threshold interaction effect model were independent of each other, and were set the same as those introduced by Marchini et al. [2007]. The genetic variants were simulated under the assumption of Hardy-Weinberg equilibrium (HWE). The details of the simulation setting are summarized in supplementary Table S1. To mimic a real data scenario, five noncausal loci were introduced under the assumption of HWE with minor allele frequencies drawing from a uniform distribution ranging from 0.1 and 0.5. For each scenario, 1,000 replicates were generated and in each replicate 150 trios (supplementary Figure S1.a) were simulated for each subphenotype. We analyzed each replicate by using CMLRE, SEP, and ALL. For both SEP and ALL, a U -statistic based forward selection algorithm was applied to search for gene-gene interactions predisposing to each subphenotype group [Lu et al., 2011]. A gene dropping based permutation test was then used to assess the type I error and power of all three methods, adjusting for an inflated type I error due to the model selection. The empirical null distribution was formed based on 1,000 permutations, and the empirical P -value was obtained by comparing the observed statistic with the empirical null distribution.

The type I error and power for SEP, ALL, and CMLRE are summarized in Table 1. The results showed that the type I error from all three methods were well controlled at the significance level of 0.05. We observed that the ALL method performed the best when all subphenotypes shared the same underlying disease model, but the power of the ALL method reduced significantly when heterogeneous subphenotypes were treated as a homogeneous group. In such a scenario, the strong genetic effects predisposing to specific subphenotypes became negligible in the whole population, leading to low power of association tests. In contrary to the ALL method, SEP treated each subphenotype as a unique entity, and thus was relatively robust to different settings of underlying subphenotype groups. However, when all the subphenotypes under study shared the same genetic mechanisms, the power of SEP could be substantially reduced due to the small sample size. Compared to the SEP and ALL methods, CMLRE first combined the subphenotypes with similar genetic mechanisms, and then built an overall test statistic to assess the significance of the association. It outperformed SEP and ALL methods under most of the scenarios. This indicated that CMLRE was less affected by the underlying number of homogenous subphenotype groups, and was less subjected to issues (e.g., reduced sample size) that ALL and SEP methods faced. In the case where all subphenotypes shared the same genetic mechanisms, by combining subphenotype groups, CMLRE attained higher power than the SEP method. In another extreme case where all subphenotypes were mutually independent, the CMLRE

method likely treated each subphenotype as a distinct subphenotype group. Therefore, it could achieve power similar to the SEP method but outperformed the ALL method.

Simulation II

The power of association tests for family-based studies could be affected by the complexity of pedigree structures. In this simulation, we evaluated the impact of pedigree structures on the performances of CMLRE, ALL, and SEP. Similar to the Simulation I, four subphenotypes were simulated with each two subphenotypes sharing the same underlying genetic mechanisms. We considered a disease model with four disease-associated loci, among which two loci followed a multiplicative interaction effect model and the other two loci followed a threshold interaction effect model. The details of the simulation setting are summarized in supplementary Table S2. Among the three pedigree structures considered, the first one used the same pedigree structure (i.e., trios) as the one used in the Simulation I. The second one had both nuclear families and sib-ships (supplementary Figs. S1.b and S1.c), and the third one was a three-generation pedigree (supplementary Fig. S1.d). For each pedigree setting, we first simulated 100,000 families and only included in the analysis the families with at least one affected member. This ended up with roughly 10,000 families per pedigree structure. Same as simulation I, 1,000 replicates, each composed of 150 families, were simulated. The gene dropping based permutation test was used to assess the type I error and the power of the three methods.

The type I error was well controlled at the significance level of 0.05 (Table 2). The power of all methods increased with the increased number of individuals per family. Compared with ALL and SEP, CMLRE attained the highest power under all simulated scenarios, regardless of the complexity of the pedigree structures. The computational time of CMLRE for second scenario (i.e., nuclear family and sib-ships) was roughly the same as the computational time for trios, and the computational time for three-generation pedigree was 1.6 times more than that of trios. The computational time increased approximately linearly with the number of individuals in families. This indicated that CMLRE had some computational advantages for analyzing large and complex pedigrees.

Simulation III

Refined subphenotypes could help to form more homogeneous subphenotype groups, and improve the power of identifying gene-gene interactions predisposing to each subphenotype. However, in practice, there is almost no gold standard to define subphenotypes. Although the CMLRE method is capable of combining subphenotypes with similar underlying genetic mechanism, it relies on the initial definition of subphenotype groups. Therefore, in this simulation, we evaluated the performance of ALL, SEP, and CMLRE methods under a variety of initial grouping strategies. We simulated five subphenotypes, among which four subphenotypes shared the same underlying disease model and the remaining one had its own disease mechanism. For each subphenotype group, we considered a disease model with four disease-associated loci, among which two loci followed a multiplicative interaction effect model and the other two loci followed a threshold interaction effect model. We started with a scenario where all the initial subphenotype groups corresponded to the underlying subphenotype groups, and then scenarios where some of the initial subphenotype groups

were from the same underlying homogeneous subphenotype groups. Finally, we considered the scenario where the initial subphenotype group composed of heterogeneous subphenotypes. The details of the simulation setting are summarized in supplementary Table S3. The power and the type I error of ALL, SEP, and CMLRE were assessed by using the gene dropping based permutation test.

As shown in Table 3, the type I error rates of three methods were well controlled. As expected, the CMLRE method performed the best when the starting subphenotypes represented the underlying subphenotype groups. Although, in general, the power of the CMLRE method decrease with the increase of the initial number of subphenotype groups, the CMLRE method could achieve relatively high power as long as the initial subphenotype groups were homogeneous. On average, about 78% of the times our method correctly combined the subphenotype groups (i.e., the combined subphenotype groups match the underlying subphenotype groups) when the initial subphenotype groups were homogeneous. Although reducing the initial number of groups may improve the performance of CMLRE, it increases the chance of having heterogeneous initial subphenotype groups. In practice, we recommend to choose an initial grouping strategy based on the purpose of the study. By focusing on the subphenotypes with the same underlying genetic mechanism, we observed that the increased number of initial subphenotype groups could have a substantial effect on the SEP method. In the simulation, the power of the SEP method could drop from 0.968 to 0.267 with the increased number of initial subphenotype groups. Compared with CMLRE and SEP, the ALL method was robust to the initial number of groups because ALL simply treated all initial subphenotype groups as one homogeneous group. Nonetheless, as shown in Table 3, the power of ALL method was consistent low when the subphenotypes had heterogeneous etiologies.

An Application to Study Phenotypic Heterogeneity of Oral Clefts

Oral clefts are generally divided into cleft lip with or without cleft palate (CL/P) and cleft palate only (CP), which involve different genetic and embryological mechanisms Dixon et al. [2011]. Recent epidemiology studies also suggest that cleft lip without cleft palate (CL) may also have different etiological features from cleft lip with cleft palate (CLP) [Mossey et al., 2009; Murray, 2002]. In order to investigate the phenotypic heterogeneity of oral clefts, we applied the proposed method to a family-based dataset from the International Consortium to Identify Genes and Interactions Controlling Oral Clefts (ICOC), investigating genetic variants and their possible interactions predisposing to three subphenotypes of oral clefts (i.e., CL, CLP, and CP). The data were composed of nearly 1,908 case-parent trios. The cases were defined as individuals with an isolated, nonsyndromic oral cleft (CL, CLP, or CP), and were recruited from a treatment center or population-based registry. DNA was collected from both the case and their parents, and was genotyped using the Illumina Human 1M DNA Analysis BeadChip. A detailed description of the study is provided in the *ICOC* study publications and also available at the dbGaP website (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1). Prior to the association analysis, we reassessed the quality of the genotype data (e.g., checking for Mendelian errors and validating the assumption of HWE). In the analyses, we excluded low-quality SNPs and individuals with either unknown phenotype or unknown cleft type. Supplementary Table S4

summarizes the number of individuals and families fell in each subphenotype category. Based on existing literature, we identified 148 SNPs that were potentially associated with oral clefts, and all these SNPs were available in the dataset.

We initiated the association analysis by treating CL, CLP, and CP as distinct subphenotypes. CMLRE combined CL and CLP into one group (i.e., CL/P), and treated CP as another distinct group. Three loci, including rs 10863790 (*IRF6*), rs7017252 (*8q24*), rs7078160 (10q25.3), were jointly associated with CL/P, while SNPs rs7969932 (*TBKI*), rs227731 (*17q22*), rs2141765 (*TBKI*) jointly predisposed to CP (Table 4). The two SNPs, rs7969932 and rs2141765, located on gene *TBKI* are in linkage disequilibrium ($r^2 = 0.311$). The association reached a significance level that was less than 0.001 based on the gene dropping based permutation test. To further investigate the effects of identified genetic variants, additional analyses were conducted using the generalized estimation equations (GEE) method, where up to three-way interactions were considered. Based on the result from GEE, there was no evidence of an interaction between the identified genes with both CL/P and CP, while the main effects of the selected loci were significant at the level of 0.05. The adjusted odds ratio and the corresponding 95% confidence interval for each selected locus are summarized in Table 5.

We also applied the SEP and ALL methods to the same dataset to search for gene-gene interactions associated with oral clefts. We found the genetic variants selected by SEP were similar for CL and CLP with one exception (i.e., rs126280), while the genetic variants identified for CP were completely different from CL/P, indicating that the genetic etiologies of CL/P and CP were different. Compared with the CMLRE method, SEP may lack of power to identify risk loci with small to moderate effect size for CL/P due to the reduced sample size. Regarding the ALL method, only two genetic variants were selected and both of them were risk factors for CL/P. The genetic variants predisposing to CP were completely missed by the ALL method. This could be partially explained by the fact that the strong genetic effects predisposing to a specific subphenotype in a small homogenous population became negligible in the whole population when all the heterogeneous subphenotypes were treated as a single entity.

Discussion

Converging evidences suggest that subphenotypes of a disease with similar clinical manifestations could have heterogeneous pathophysiological and etiological mechanisms [Guey et al., 2010; Lu et al., 2009; Morris et al., 2010]. In the presence of phenotypic heterogeneity, methods that treat distinct subphenotypes as the same disease are subject to low power, as the strong genotypic effects on a small homogeneous subpopulation would become small or even negligible in the whole population. On the other hand, subphenotype analyses, where subphenotype groups are formed based upon prior knowledge (e.g., disease symptom, severity, and age of onset) may overestimate or underestimate the underlying subphenotype groups, resulting in low power of association analyses.

In this paper, we introduced a CMLRE method for family-based association studies considering phenotypic heterogeneity. The CMLRE method gradually combines

subphenotypes into homogeneous groups, and thus does not require prior knowledge of underlying subphenotype groups. It is developed based on a forward U -statistic method, which allows for the identification of genetic variants with the consideration of high-order gene-gene interactions. Moreover, it makes no assumption of mode of inheritance, and is thus robust to various underlying disease models [Lu et al., 2011]. CMLRE takes the correlation between family members by using a clustered receiver operating characteristic (ROC) method [Obuchowski, 1997] that has no restriction on pedigree structures, making it applicable to various pedigree settings.

Through simulations, we showed that when the empirical criteria overestimated the underlying homogeneous groups, the subphenotype analysis had low power due to the reduced sample size. On the other hand, when heterogeneity existed in some subphenotypes, the ALL method was subject to low power. In contrast to both subphenotype analysis and the analysis in which all subphenotypes were treated as a single entity, the CMLRE method in general attained powerful and robust performance by simultaneously searching for the optimal subphenotype groups and identifying gene-gene interactions predisposing to each subphenotype. Though excess number of initial subphenotype groups reduced the power of CMLRE, it significantly outperformed the other two methods. Therefore, the CMLRE provides a powerful and flexible framework to model genetic variants with complex human diseases by allowing for heterogeneous genetic causes among subphenotypes of a disease.

We further applied the CMLRE method to study the genetic variants associated with CLP, CL, and CP. Consistent with previous findings, our method indicated that CLP and CL shared similar genetic variants, while the genetic etiology of CP was different from CL/P [Mitchell et al., 2002]. In the association analysis, we have identified a three-locus joint model for CL/P that included rs10863790 (*IRF6*), rs7017252 (*8q24*), and rs7078160 (*VAX1*). All of the identified loci have been previously reported to be associated with CL/P [Beaty et al., 2010; Birnbaum et al., 2009; Blanton et al., 2010; Nikopensius et al., 2010]. Further investigation with the GEE model indicated no interaction effect among three identified genetic variants. Our result is consistent with the previous finding by Birnbaum et al., who claimed no evidence of an interaction effect between *IRF6* and *8q24* [Birnbaum et al., 2009]. However, Blanton et al. [2010] have identified a significant interaction effect on CL/P between SNPs located on *IRF6* and *8q24.21* in a non-Hispanic white population. Therefore, follow-up studies are needed to thoroughly investigate the role of *IRF6* and *8q24.21* on CL/P etiology. We also identified a three-locus joint model for CP, including rs7969932 (*TBKI*), rs227731 (*17q22*), and rs2141765 (*TBKI*). The rs7969932 (*TBKI*) on chromosome 12q14.2 has been reported to be associated with CP in the presence of maternal smoking by a recent GWAS, while rs2141765 (*TBKI*) was significantly associated with CP regardless of maternal smoking status [Beaty et al., 2011]. Though rs7969932 and rs2141765 were in linkage disequilibrium ($r^2 = 0.311$) in the dataset, the associations between CP and each of the locus remained significant when the effects of the other loci were adjusted. rs227731 (*17q22*) was first reported to be associated with CP/L by Mangold et al. in 2010, and to our knowledge the association between rs227731 and CP has never been reported. Therefore, follow-up studies will be needed to further replicate and study the effect of *17q22* on the etiology of CP.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Project was supported by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and the National Natural Science Foundation of China (Award No. 81502887), the National Institute of Dental & Craniofacial Research under Award Number R03DE022379, and the National Institute on Drug Abuse under Award Number K01DA033346. The datasets used for the analyses was obtained from dbGaP through dbGaP accession number [phs000094.v1.p1].

References

- Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet.* 2010; 42(6):525–529. others. [PubMed: 20436469]
- Beaty TH, Ruczinski I, Murray JC, Marazita ML, Munger RG, Hetmanski JB, Murray T, Redett RJ, Fallin MD, Liang KY. Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genet Epidemiol.* 2011; 35(6):469–478. others. [PubMed: 21618603]
- Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, Almeida de Assis N, Alblas MA. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet.* 2009; 41(4):473–477. others. [PubMed: 19270707]
- Blanton SH, Burt A, Stal S, Mulliken JB, Garcia E, Hecht JT. Family-based study shows heterogeneity of a susceptibility locus on chromosome 8q24 for nonsyndromic cleft lip and palate. *Birth Defects Res A Clin Mol Teratol.* 2010; 88(4):256–259. [PubMed: 20196142]
- Dixon MJ, Marazita ML, Beaty TH, Murray JC. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet.* 2011; 12(3):167–178. [PubMed: 21331089]
- Ellis KL, Pilbrow AP, Frampton CM, Doughty RN, Whalley GA, Ellis CJ, Palmer BR, Skelton L, Yandle TG, Palmer SC. A common variant at chromosome 9P21.3 is associated with age of onset of coronary disease but not subsequent mortality. *Circ Cardiovasc Genet.* 2010; 3(3):286–293. others. [PubMed: 20400779]
- Guey LT, Garcia-Closas M, Murta-Nascimento C, Lloreta J, Palencia L, Kogevinas M, Rothman N, Vellalta G, Calle ML, Marenne G. Genetic susceptibility to distinct bladder cancer subphenotypes. *Eur Urol.* 2010; 57(2):283–292. others. [PubMed: 19692168]
- Hancock DB, Martin ER, Li YJ, Scott WK. Methods for interaction analyses using family-based case-control data: conditional logistic regression versus generalized estimating equations. *Genet Epidemiol.* 2007; 31(8):883–893. [PubMed: 17565751]
- Ku CS, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: where are we now? *J Hum Genet.* 2010; 55(4):195–206. [PubMed: 20300123]
- Liang KY, Zeger SL. Longitudinal data-analysis using generalized linear-models. *Biometrika.* 1986; 73(1):13–22.
- Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD. A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am J Hum Genet.* 2008; 83(4):457–467. [PubMed: 18834969]
- Lu Q, Song Y, Wang X, Won S, Cui Y, Elston RC. The effect of multiple genetic variants in predicting the risk of type 2 diabetes. *BMC Proc.* 2009; 3(Suppl 7):S49. [PubMed: 20018041]
- Lu, Q., Wei, C., Ye, C., Li, M., Elston, R. A likelihood ratio based Mann-Whitney approach finds novel replicable joint gene action for type 2 diabetes. 2011. Submitted to *Genet Epidemiol*
- Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet.* 2010; 42(1):24–26. others. [PubMed: 20023658]

- Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*. 2010; 363(2):166–176. [PubMed: 20647212]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007; 39(7):906–913. [PubMed: 17572673]
- Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genet Epidemiol*. 2006; 30(2):111–123. [PubMed: 16374833]
- Mitchell LE, Beaty TH, Lidral AC, Munger RG, Murray JC, Saal HM, Wyszynski DF. Guidelines for the design and analysis of studies on nonsyndromic cleft lip and cleft palate in humans: summary report from a Workshop of the International Consortium for Oral Clefts Genetics. *Cleft Palate Craniofac J*. 2002; 39(1):93–100. [PubMed: 11772175]
- Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, McCarthy MI. A powerful approach to subphenotype analysis in population-based genetic association studies. *Genet Epidemiol*. 2010; 34(4):335–343. [PubMed: 20039379]
- Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. *Lancet*. 2009; 374(9703):1773–1785. [PubMed: 19747722]
- Murray JC. Gene/environment causes of cleft lip and/or palate. *Clin Genet*. 2002; 61(4):248–256. [PubMed: 12030886]
- Nikopensius T, Birnbaum S, Ludwig KU, Jagomagi T, Saag M, Herms S, Knapp M, Hoffmann P, Nothen MM, Metspalu A. Susceptibility locus for non-syndromic cleft lip with or without cleft palate on chromosome 10q25 confers risk in Estonian patients. *Eur J Oral Sci*. 2010; 118(3):317–319. others. [PubMed: 20572868]
- Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997; 53(2):567–578. [PubMed: 9192452]
- Qi L, Ma J, Qi Q, Hartiala J, Allayee H, Campos H. Genetic risk score and risk of myocardial infarction in Hispanics. *Circulation*. 2011; 123(4):374–380. [PubMed: 21242481]
- Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007; 316(5829):1336–1341. others. [PubMed: 17463249]

Table 1

Power comparison of ALL, SEP, and CMLRE under a varied number of underlying homogeneous subphenotype groups

Underlying subphenotypes groups ^a		CMLRE	ALL	SEP1 ^b	SEP2 ^c	SEP3 ^d	SEP4 ^e
(1,2,3,4)	Power	0.609	0.916	0.254	0.242	0.234	0.184
	Type I error	0.051	0.051	0.064	0.061	0.055	0.044
(1,2,3);4	Power	0.866	0.561	0.385	0.334	0.410	0.648
	Type I error	0.048	0.052	0.058	0.061	0.048	0.056
(1,2);(3,4)	Power	0.890	0.235	0.512	0.506	0.460	0.425
	Type I error	0.049	0.039	0.057	0.058	0.049	0.043
1;2;3;4	Power	0.729	0.052	0.114	0.142	0.133	0.812
	Type I error	0.050	0.037	0.048	0.058	0.041	0.051

^aSubphenotypes in parentheses share the same disease mechanisms.

^bSubphenotype analysis for subphenotype 1.

^cSubphenotype analysis for subphenotype 2.

^dSubphenotype analysis for subphenotype 3.

^eSubphenotype analysis for subphenotype 4.

Table 2

Power comparison of ALL, SEP, and CMLRE under different pedigree structures

Pedigree		CMLRE	ALL	SEP1 ^a	SEP2 ^b	SEP3 ^c	SEP4 ^d
Trios	Power	0.881	0.307	0.478	0.476	0.392	0.454
	Type I error	0.028	0.069	0.033	0.064	0.043	0.036
Mixed	Power	0.902	0.414	0.614	0.558	0.459	0.408
	Type I error	0.046	0.053	0.058	0.047	0.046	0.049
Third generation	Power	0.967	0.812	0.78	0.831	0.638	0.592
	Type I error	0.048	0.054	0.046	0.066	0.052	0.051

^aSubphenotype analysis for subphenotype 1.^bSubphenotype analysis for subphenotype 2.^cSubphenotype analysis for subphenotype 3.^dSubphenotype analysis for subphenotype 4.

Table 3

The impact of the initial number of subphenotype groups on the performances of three methods

Starting groupings ^a		CMLRE	ALL	SEP1 ^d	SEP2 ^e	SEP3 ^f	SEP4 ^g	SEP5 ^h
1 ^b ; (2 + 3 + 4 + 5)	Power	0.981	0.406	0.661	0.968	N/A	N/A	N/A
	Type I error	0.050	0.049	0.049	0.050	N/A	N/A	N/A
1 ^b ; (2 + 3); (4 + 5)	Power	0.927	0.476	0.704	0.595	0.625	N/A	N/A
	Type I error	0.056	0.047	0.075	0.056	0.052	N/A	N/A
1 ^b ; (2 + 3 + 4); 5	Power	0.913	0.590	0.683	0.888	0.216	N/A	N/A
	Type I error	0.034	0.061	0.045	0.038	0.054	N/A	N/A
1 ^b ; 2; 3; (4 + 5)	Power	0.902	0.462	0.726	0.200	0.236	0.680	N/A
	Type I error	0.055	0.048	0.056	0.052	0.060	0.068	N/A
1 ^b ; 2; 3; 4; 5	Power	0.830	0.476	0.716	0.225	0.267	0.254	0.234
	Type I error	0.055	0.024	0.064	0.053	0.057	0.059	0.061
(1 ^b +2) ^c ; (3 + 4 + 5)	Power	0.805	0.491	0.184	0.880	N/A	N/A	N/A
	Type I error	0.049	0.060	0.044	0.040	N/A	N/A	N/A
(1 ^b +2+3) ^c ; (4 + 5)	Power	0.585	0.499	0.131	0.592	N/A	N/A	N/A
	Type I error	0.034	0.044	0.027	0.052	N/A	N/A	N/A

^aSubphenotypes in parentheses are treated as a homogeneous group initially.^bSubphenotype 1 has its own mechanism, while the other four subphenotypes share the same mechanism.^cHeterogeneous subphenotype group.^dSubphenotype analysis for subphenotype group 1.^eSubphenotype analysis for subphenotype group 2.^fSubphenotype analysis for subphenotype group 3.^gSubphenotype analysis for subphenotype group 4.^hSubphenotype analysis for subphenotype group 5.

Table 4

Identified genetic variants predisposing to cleft lip with/without palate (CL/P) and cleft palate only (CP) by ALL, SEP, and CMLRE

Models	CMLRE		SEP			ALL
	CL/P	CP	CL	CLP	CP	CL, CLP, CP
SNPs	rs10863790	rs7969932	rs10863790		rs7969932	
	(<i>IRF6</i>)	(<i>TBK1</i>)	(<i>IRF6</i>)	rs10863790	(<i>TBK1</i>)	rs10863790
	rs7017252	rs227731	rs7017252	(<i>IRF6</i>)	rs227731	(<i>IRF6</i>)
	(<i>8q24</i>)	(<i>17q22</i>)	(<i>8q24</i>)	rs7017252	(<i>17q22</i>)	rs7017252
	rs7078160	rs2141765	rs126280	(<i>8q24</i>)	rs2141765	(<i>8q24</i>)
	(<i>VAX1</i>)	(<i>TBK1</i>)			(<i>TBK1</i>)	
<i>P</i> -value ^a	<0.001			<0.001		<0.001

^aEstimated based on 1,000 permutations.

Table 5

GEE analyses for cleft lip with/without palate (CL/P) and cleft palate only (CP)

	CP				CL/P	
SNPs	rs7969932 (CT/CC)	rs227731 (AC/CC)	rs2141765 (AC/CC)	rs10863790 (AC/CC)	rs7017252 (CT/TT)	rs7078160 (AG/AA)
OR ^a	1.69	0.63	1.50	0.59	1.60	1.33
CI ^b	(1.37,2.08)	(0.52,0.75)	(1.22,1.85)	(0.52,0.68)	(1.41,1.8)	(1.19,1.5)

^aOdds ratio.^b95% confidence interval.