Published in final edited form as:

Stat Med. 2010 December 10; 29(28): 2932–2945. doi:10.1002/sim.4057.

Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies†

Andrew Crossett^a, Brian P. Kent^a, Lambertus Klei^b, Steven Ringquist^c, Massimo Trucco^c, Kathryn Roeder^{a,*†}, and Bernie Devlin^b

^aDepartment of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

^bDepartment of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, U.S.A.

^cDivision of Immunogenetics, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA 15224, U.S.A.

Abstract

We propose a method to analyze family-based samples together with unrelated cases and controls. The method builds on the idea of matched case–control analysis using conditional logistic regression (CLR). For each trio within the family, a case (the proband) and matched pseudo-controls are constructed, based upon the transmitted and untransmitted alleles. Unrelated controls, matched by genetic ancestry, supplement the sample of pseudo-controls; likewise unrelated cases are also paired with genetically matched controls. Within each matched stratum, the case genotype is contrasted with control pseudo-control genotypes via CLR, using a method we call *matched-CLR* (mCLR). Eigenanalysis of numerous SNP genotypes provides a tool for mapping genetic ancestry. The result of such an analysis can be thought of as a multidimensional map, or eigenmap, in which the relative genetic similarities and differences amongst individuals is encoded in the map. Once constructed, new individuals can be projected onto the ancestry map based on their genotypes. Successful differentiation of individuals of distinct ancestry depends on having a diverse, yet representative sample from which to construct the ancestry map. Once samples are well-matched, mCLR yields comparable power to competing methods while ensuring excellent control over Type I error.

Keywords

conditional logistic regression; family-based design; genome-wide association; matched case–control; population stratification

†Supporting information may be found in the online version of this article.

*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, 132 Baker Hall, Pittsburgh, PA 15213, U.S.A.. †roeder@stat.cmu.edu.

Web Resources

The URL for data presented herein is as follows: mCLR source code, <http://wpicr.wpic.pitt.edu/WPICCompGen/software.htm>.

Introduction

Collections of large samples, including case and control individuals as well as families containing affected individuals, are enhancing our ability to identify DNA variants affecting risk for disease. It has become the standard to search for genetic variants associated with complex disease using genome-wide association studies (GWAS). As of March 2010, 2403 associations for common diseases/phenotypes have been validated [1, 2]. Using combined data from three studies, genome-wide association has identified more than 30 variants for Crohn's disease [3]. Yet, even with these successes, many more variants are yet to be discovered [4]. In the pursuit of unexplained heritability more samples from many collections are being combined to increase the power to detect additional risk variants.

In addition to sample collections for specific diseases, genotype data from large samples of unrelated controls are freely available for common use [5]. Examples include the Wellcome Trust's Case Control Collaboration and the databases in the Genotypes and Phenotypes (dbGaP) archive from tens of thousands of individuals.

The two most common sampling techniques for studies of association are the case-control design and the family-based design. Owing to demographical, biological, and random forces, genetic variants differ in allele frequency in populations around the world, creating population structure or stratification reflected by ancestry. As a consequence, case-control studies are susceptible to spurious associations between genetic variants and disease status [6]. As more data are assimilated for combined analysis, the challenge of spurious associations due to population structure increases [7-9].

For the case-control design, a large panel of genetic markers can be used to estimate genetic ancestry by using principal components analysis (PCA) [10, 11] or related dimension reduction techniques [12], which are referred to as eigenmaps. These low dimensional maps encode the relative genetic similarities and differences amongst individuals. Indeed, the principal eigenvectors often reflect geographical distribution as well as hidden structures in human populations [13, 14]. Given ancestry coordinates, the effects of population stratification can be removed either by regressing out their effects [10], or by matching cases and controls of similar ancestry and performing conditional logistic regression (CLR) [12, 15-18].

Family-based designs are robust to population stratification. For simplicity, we will only consider the trio design, in which both parents are genotyped along with the affected offspring. Analysis involves designation of a case (the affected offspring) and matched 'pseudo-controls' inferred on the basis of the transmitted and untransmitted alleles of the parents [19-22]. The structure of the data is equivalent to a matched case-control sample and hence can be analyzed via CLR [23, 24]. An extension to more general types of family data can be found in the discussion.

The research problem we address here is how to use both case-control and family-based data in a test for association that is powerful yet robust to population stratification. Toward this purpose, the *combined association analysis* method was developed by Nagelkerke *et al.* [25] and refined by Epstein *et al.* [26]. Those approaches provide greater power in

association studies than using either case-control or family-based samples separately. The drawback is that they make a strong assumption about the sample with respect to population structure. If the assumption fails, the tests can lead to spurious associations. Although this assumption can be met by particularly well-planned studies, it is impossible to guarantee if data are combined across many studies.

We propose a hybrid analytical approach that is robust to differences in sampling distribution across studies, controls Type I error and yet attains good power. This method requires that sufficient genotyping is available on all samples to permit matching samples based on genetic ancestry. To test for association, the matched strata are analyzed within a CLR framework. To this end, we will refer to our method as a *matched-CLR* (mCLR) approach.

The success of our approach depends upon the quality of the eigenmap. In practice, the map can be constructed from the full sample of individuals available or a representative *base* sample. The base sample might include individuals from a broad range of ancestry or a fairly homogeneous sample. Once constructed, new individuals can be projected onto the ancestry map based on their genotypes using the Nystrom approximation [27]. To illustrate how the map varies depending upon the choice of base sample we use two public databases that have samples of people of European ancestry and sufficient demographic data to permit classification of each person to his country of origin. In the first sample, individuals were collected for the Human Genome Diversity Project (HGDP) to reflect the genetic diversity of current human populations, thereby enhancing studies of human evolutionary history [28]. This sample emphasizes distinct populations, including isolated and geographically well-separated peoples. In contrast the Population Reference Sample (POPRES) was assembled with the goal of bringing together a set of DNA samples that would support a variety of efforts related to pharmacogenetics research [29]. It tends to represent major populations. The features of these collections will be used to examine the performance of eigenmaps constructed using a variety of base samples.

Methods

Data

The HGDP panel includes 1063 individuals from seven continental groups classified into 51 populations, eight of which are located in Europe. Individuals are genotyped at a large number of biallelic markers (single nucleotide polymorphisms or SNPs). We removed individuals with less than 95 per cent complete genotypes, SNPs with less than 99 per cent complete genotypes, or minor allele frequency less than 1 per cent. Finally, we allow for distinct subpopulation allele frequencies by adding normally distributed test statistics for Hardy Weinberg disequilibrium across tribes within subcontinents. SNPs with p -values less than 10^{-4} were removed. The POPRES database includes 2955 subjects of European ancestry. Demographic records include the individual's country of origin and that of his/her parents and grandparents. The sampling frequency varies strongly by region, including Swiss (1014), British and Irish (436), Italian (205), Portuguese and Spanish (252), French (108), north-west European (173), east-central European (75), south-eastern European (45), and other (647). Of the more than 650 000 markers typed on HGDP and 450 000 on

POPRES, focusing on the fraction that were present on both platforms, we selected tag SNPs using H-clust [30] to ensure that pairs of SNPs have correlations of 0.04 or less. Ultimately, we chose 14 650 (approximately) independent SNPs that passed our quality control in both European databases.

Matched analysis

Let G denote the minor allele count for a subject (0, 1, or 2) and D denote the disease outcome (1 affected and 0 unaffected). Define the genotype relative risk (GRR) [21] as

$$\frac{P(D=1|G=x)}{P(D=1|G=0)} = \psi_x \quad \text{for } x=1, 2.$$

For a multiplicative model, $\psi_1 = \psi$ and $\psi_2 = \psi^2$; equivalently, the log GRR is linear in G with coefficient $\beta \log(\psi)$. We wish to test the hypothesis $\psi=0$, using both case-control and family-based data.

The Euclidean distance between individuals in the eigenmap are representative of their ancestral differences. Either the fullmatch or the pairmatch algorithm [31] can be used to find genetically homogeneous strata. For the case-control design, the fullmatch algorithm minimizes distances between individuals within strata with the constraint that each stratum consists of either a single case matched to one or more controls, or a single control matched to one or more cases within. Alternatively, the pairmatch algorithm minimizes the distances between cases and controls in strata with the constraint that each case is matched with a single control. The contribution to the likelihood for the k th matched stratum, including 1 case ($i=0$) and m controls ($i=1, \dots, m$), follows the conditional logistic form,

$$e^{x_0\beta} / \sum_{i=0}^m e^{x_i\beta}.$$

A traditional approach to family-based analysis of parents and a single affected offspring (trios) is to treat the transmitted alleles as the case genotype and the remaining possible but unrealized genotypes as pseudo-controls using CLR [21-24]; X-linked loci are treated analogously for alleles. As noted by Self *et al.* [23], conceptually the family-based design is essentially equivalent to a case-control study in which the controls are sampled from hypothetical siblings. Thus for the purpose of analysis both case-control and family-based designs can lead to strata, each consisting of a case and one or more controls.

Eigenmaps

As a first step we estimate the genetic background of unrelated individuals (unrelated cases, unrelated controls, and trio probands) using a dimension reduction technique. Let x_{ij} be the minor allele count for the i th subject and the j th SNP in a matrix X . Center and scale the columns of X by subtracting the mean and dividing by the standard deviation. Assuming a sample size of N , traditional PCA decomposes XX' using eigenvalue decomposition to obtain the eigenvectors, $(\mathbf{u}_1, \dots, \mathbf{u}_N)$, and eigenvalues, $\lambda_1, \dots, \lambda_N$. Rescaled eigenvectors map the i th subject into an s -dimensional space according to

$$\left(\lambda_1^{1/2} \mathbf{u}_1(i), \dots, \lambda_s^{1/2} \mathbf{u}_s(i) \right). \quad (1)$$

Rather than using traditional PCA, we utilize a variant of this approach that arises from spectral graph theory [12]. The basic idea is to represent the population as a weighted graph, where the weights reflect the degree of similarity between pairs of subjects. As with PCA, the graph is then embedded to a lower dimensional space using the top eigenvectors of a function of the weight matrix. Lee *et al.* [12] show that the spectral graph analysis (SGA) leads to more meaningful clusters than ancestry estimated via PCA. Eigenvectors calculated based upon PCA are strongly affected by uneven sampling of populations [32]. While somewhat susceptible to this bias, the SGA is more robust to cluster size [33]. Moreover, SGA also identifies eigenvectors that successfully separate the data into homogeneous clusters that frequently correspond to demographic labels [12].

To perform spectral graph analysis (SGA), we start with the PCA kernel, $X X^t$ and create a weight matrix W for spectral analysis:

$$w_{ij} = \begin{cases} \sqrt{\mathbf{x}_i^t \mathbf{x}_j} & \text{if } \mathbf{x}_i^t \mathbf{x}_j \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are row vectors of X . These w_{ij} are the edge weights of the graph. Let $d_i = \sum_{j=1}^n w_{ij}$ be the degree of vertex i , and let $D = \text{diag}(d_1, \dots, d_n)$ be a diagonal matrix. The normalized Laplacian matrix for W is defined as $I - L$ where $L = D^{-1/2} W D^{-1/2}$. Let ν_i and $\mathbf{u}_i$ be the eigenvalues and eigenvectors of $I - L$ and let $\lambda_i = \max \{0, 1 - \nu_i\}$. Map the i th subject into the S -dimensional eigenmap using (1). The dimension of the eigenmap, S , is determined using the eigengap heuristic to test for the number of significant eigenvalues in L (not including the trivial dimension). Given the S -dimensional representation, we use Ward's algorithm to partition the data into large homogeneous clusters [12, 17]. A cluster is considered homogeneous provided the eigenvalues are not significant based on the eigengap heuristic [12]. SGA is available as an R library, SpectralGEM (www.r-project.org).

The base sample, consisting of subject $i = 1, \dots, n$ corresponding to the centered and scaled allele count vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$, defines X and determines the eigenmap. To project a new individual with scaled allele count vector $\mathbf{z}$ onto this basis we use the Nystrom approximation. The weight associated with an edge between $\mathbf{z}$ and $\mathbf{x}$ is

$$w(\mathbf{z}, \mathbf{x}) = \begin{cases} \sqrt{\mathbf{z}^t \mathbf{x}} & \text{if } \mathbf{z}^t \mathbf{x} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The degree associated with $\mathbf{z}$ is $d(\mathbf{z}) = \sum_{i=1}^n w(\mathbf{z}, \mathbf{x}_i) + w(\mathbf{z}, \mathbf{z})$. Finally,

$$L(\mathbf{z}, \mathbf{x}_i) = [d(\mathbf{z}) d_i]^{-1/2} w(\mathbf{z}, \mathbf{x}_i).$$

The eigenvector coordinates for dimensions $k=1, \dots, S$ for $\mathbf{z}$ are

$$\nabla \mathbf{u}_k(\mathbf{z}) = \lambda_k^{-1} \sum_{i=1}^n L(\mathbf{z}, \mathbf{x}_i) \mathbf{u}_k(\mathbf{x}_i).$$

Using these eigenvector coordinates we can map new individuals into an existing eigenmap using (1).

Combining trios, cases, and controls

As a first step we estimate the genetic background of unrelated individuals (cases, controls, and trio probands) using a dimension reduction technique. Here as an illustration we consider genotypic data from the International HapMap Project (30 CEU trios) and from the POPRES database [29]. Trio probands are matched to one or more controls that are genetically similar based on the eigenmap (Figure 1) [17]. The Euclidean distance between individuals in this eigenspace are representative of their genetic differences. When data consist of family-based samples as trios of parents and their affected offspring, as well as additional controls, we will prefer one case:many control matching.

For trios, pseudo-controls are automatically matched by ancestry with the corresponding proband, and will be contrasted to the case genotype. Additional information can be garnered by clustering trio probands with unrelated controls. In this way we identify additional controls who are matched by ancestry to the probands (Figure 1). The structure of the data is equivalent to a matched case-control sample and hence can be analyzed via CLR.

Some adjustments to the fullmatch algorithm are necessary in practice. There is a computational advantage to limiting each stratum to include only one case. Specifically, for computational reasons, the conditional logit algorithm works best for $1:m$ or $m:1$ matching. In the general case of $n:m$ matching the contribution of the k -th stratum to the conditional likelihood is

$$l_k(\beta) = \frac{\prod_{i=1}^m e^{x_i \beta}}{\sum_{j=1}^{c_k} \prod_{i_j=1}^m e^{x_{ji} \beta}},$$

where $c_k = (n+m)!/m!n!$. One can see that by adding multiple cases to a stratum we are increasing the number of terms in the denominator. For instance, if $m=20$ and $n=1$, 20 terms are in the denominator, whereas at $n=2$ it is 190 terms, and at $n=3$ it is 1140 terms. By design there are multiple pseudo-controls per case. Therefore, we maintain the $1:m$ structure by matching additional controls to the case, if the ancestral match is suitable. Moreover, to be useful in the association analysis, every unrelated case must be matched to an unrelated control. Thus we first match unrelated cases to one or more unrelated controls. These individuals are then set aside as matched strata. Next we use fullmatch to cluster trio probands with the remaining unrelated controls. If fullmatch defines a cluster that includes multiple trio probands, one proband is selected at random to remain in the stratum. The extra

probands are each moved to their own strata together with their ancestrally identical pseudo-controls. We now have K strata consisting of 1 case and n_k controls in stratum k .

Some unrelated controls will not be similar enough to any probands to merit inclusion in the study. For example, the HapMap trios can only be successfully matched to a subset of the full European sample in POPRES (Figure 1). Likewise some unrelated cases might not be well matched by any unrelated controls in the study. SpectralGEM provides features that facilitate the removal of individuals who cannot be successfully matched because their genetic ancestry is too remote, relative to the others in the sample [12, 17]. These individuals should be removed from further consideration in the association study. Once the strata are established, a natural next step is to compare the differences in genotypes between the case and controls by using CLR.

Results

Ensuring a robust eigenmap for matching by ancestry

Based on our analysis of eigenmaps estimated from data for each continent in the HGDP sample, we can see that individuals cluster with remarkable consistency by population and geographic region (Figure 2, Supplementary Figures 1–2). For instance, the six African populations formed six clusters with near perfect concordance; the eight European populations formed five clusters, distinguishing the Adygei, French Basque, Russian and Sardinian and grouping the French, Northern Italian, Tuscan and Orcadian populations.

To represent the genetic diversity of Europe in the POPRES sample we selected a stratified random sample from the database, including 50 British, 50 French, 100 Italian, 100 Portuguese or Spanish, 50 Swiss, 50 East-Central Europeans, and 45 South-Eastern Europeans. These 495 individuals, combined with the 156 Europeans in HGDP, were split into a base set ($n = 330$) for construction of the eigenmap and a projection set ($n = 321$). The latter samples were projected into the eigenmap via the Nystrom approximation (Supplementary Figure 3). The projected individuals mimic the distribution of the base sample over the space. This shows that the eigenvectors separate the observations based on underlying features in the data and these same features are present in the projected data.

To see how the eigenspace varies depending on the choice of base sample, we estimate the eigenvectors using various base samples: (a) European HGDP data; (b) European POPRES; (c) HGDP and half of POPRES; and (d) the 330 individuals from HGDP and POPRES described above. The remaining data were projected onto the eigenspace to illustrate the estimated ancestry distribution (Figures 3(a)–(d)). Regardless of the choice of base, four of the HGDP populations (Adygei, French Basque, Sardinian and Russian) are distinct from other HGDP populations in the eigenspace. The other four populations (French, Northern Italian, Orcadian, and Tuscan), which are more similar to the POPRES sample, separate best in the eigenspace if at least some of the POPRES sample is included in the base (b, c, d). Overall the HGDP sample is more heterogeneous than the POPRES sample (a, c); however, this distinction is muted when the HGDP sample is not included in the base calculations (b). In essence, the eigenspace aims to separate clusters like those included in the base. As a result, when using HGDP as a base, the axes do not highlight the differences in the POPRES

sample causing them to clump together in the center of the eigenspace (a). Likewise, when using a POPRES base, the axes do not capture the strong differences in the HGDP data (b). Using data from both repositories produces an eigenspace that better reflects the full range of variability in the data (c, d). Using a balanced sample from the available data improves the separation between these populations (d).

Using the balanced sample we compare the HGDP populations with particular subsets of the POPRES data (Figures 4(a)–(d)). The HGDP-French correspond well with the bulk of the POPRES-French sample (a). Likewise the core of the POPRES-British & Irish sample corresponds well with HGDP-Orcadian population (b). The POPRES-Italian sample is more heterogeneous, spanning the range of the HGDP-Northern Italians and HGDP-Tuscans (c). The HGDP-French-Basque map midway between the POPRES-French and POPRES-Spanish/Portuguese samples on the first dimension, but separate in the second dimension (d). The POPRES sample is not well represented by individuals from eastern Europe, hence there is no natural comparison group for the HGDP-Russian and Adygei samples. Similarly none of the populations sampled in POPRES overlaps with the HGDP-Sardinian population. In total the similarity of the populations of like ancestry suggests that the eigenmap based on ancestrally balanced samples is providing a meaningful representation of ancestry that can be used to match cases and controls of unknown ancestry in practice.

Assuming a public reference sample is available to serve as a control, the objective is to select a set of controls with ancestry similar to the cases without the aid of detailed demographic records of ancestry. To this end we conduct an experiment to see how well we can match individuals in the projected sample to those in the base sample by pair matching to minimize the total pairwise distance in the eigenmap [31]; and by matching at random within each of the seven strata in POPRES and eight strata in HGDP. Distances observed for the two different matching criterion are similar (Supplementary Figure 4), which suggests that the eigenvectors are mapping populations in correspondence with subtle demographic labels.

We conjecture that eigenmaps are most successful when the base sample is a diverse but representative sample. If our conjecture is correct we predict that analysis of worldwide samples will highlight continental differentiation, but obscure fine-scale ancestral differentiation. To examine this prediction we construct an eigenmap using the full sample of 51 populations from HGDP. Using this base, we identified 12 significant dimensions of ancestry. In clustering individuals based on this 12 dimensional space, we successfully clustered individuals by continent, but lost the ability to identify many of the individual populations within continents that were apparent at the continental level (Supplementary Figures 1 and 5). For example, the six formerly distinct African populations now cluster together; the six regionally distinct clusters of East Asians now cluster into a southern and northern component; and all Europeans cluster together.

Simulations to demonstrate effectiveness of control over stratification in mixed population and family-based samples

To simulate the marker information for the unrelated case–control data we use the Balding–Nichols model [34]. We generate samples from C subpopulations with a fixed F_{st} to model

the difference in allele frequencies between populations. Trios are also generated from each of the C populations.

To simulate genotypes for case individuals drawn from subpopulation $c = 1, \dots, C$, allele counts 0, 1, and 2 are assigned with probabilities

$$\frac{(1 - p_c)^2}{t}, \frac{2\psi p_c (1 - p_c)}{t} \quad \text{and} \quad \frac{\psi^2 p_c^2}{t}$$

respectively, where p_c is the minor allele frequency in population c for controls,

$t = (1 - p_c)^2 + 2\psi p_c (1 - p_c) + \psi^2 p_c^2$ and ψ is the GRR. For the trio data, there are 10 family types. For every locus or marker, l , and c , a family type was generated using the probabilities found in Table I. These probabilities are based on the assumption of the Hardy–Weinberg equilibrium in the parental generation.

To compare the *mCLR* method with the *combined association* approach, we simulated a simple scenario including population stratification by sampling the trio data in proportion, q and $1 - q$, from $C = 2$ subpopulations. The unrelated controls were sampled with equal probability from the two subpopulations. For this sampling scenario, the two samples were combinable without concern for population heterogeneity only when $q \approx 0.5$. To examine the false positive rate when the sampling proportions in subpopulations are not the same, we varied q between 0.1 and 0.5, and set the GRR at $\psi = 1$ (under a multiplicative model with no risk). Each simulation included 500 controls and 500 trios. Three levels of stratification were simulated: $F_{st} = 0.001, 0.01, 0.05$. As expected the *mCLR* did a better job than the combined association analysis in controlling for spurious associations in the presence of population stratification (Figure 5). When $F_{st} = 0.05$, the combined association analysis produced unacceptably high Type I errors at every level of q . Even when the two populations are quite similar genetically ($F_{st} = 0.001$), the combined association analysis produced false positives at a rate above the threshold value of $\alpha = 0.05$ when sampling ratios are substantially unbalanced. Epstein *et al.* [26] recommends testing whether the data should be combined prior to performing the analysis. If that test were successful, it would prevent inflated Type I errors, but would also fail to capture the information in the unrelated controls samples.

We next compared the power of the *mCLR* design with the combined association analysis using a model that favors the combined association analysis. Data are generated with no population structure ($C = 1$) so that matching is unnecessary. In this scenario it is well known that matching leads to a modest loss in power [35]. Using 500 controls and 500 trios, we calculated the power for ψ ranging from 1 to 2. From Figure 6(a), we can see that even in the worst case scenario, *mCLR* exhibits only a modest loss of power. The greatest loss of power occurred in the interval $1.1 < \psi < 1.4$, with a maximum reduction of 6 per cent occurring at $\psi 1.2$.

Thus far power calculations were based on simulations restricted to 1:1 matching of probands to unrelated controls. Next we examined what would happen to the power if we increased the ratio of controls matched to the trio proband within each stratum. For instance

in Figure 1 each HapMap trio could be matched to many POPRES controls. Therefore, we decided to vary the ratio of unrelated controls to each trio proband for three levels of relative risk ($\psi = 1.3, 1.4, 1.5$). To make the simulations more realistic we used features of the POPRES database [29]. In the European sample of POPRES we identified $C = 6$ large, genetically homogeneous clusters [12]. Within each of these six clusters we computed the allele frequencies for 10 000 randomly selected SNPs, each with minor allele frequency greater than 0.05. Using these allele frequencies we generated 50 trios from each of the six subpopulations. Next, we generated unrelated controls from these six subpopulations to obtain, on average, a matching ratio of R . We vary R from 0 to 20. From Figure 6(b) we can see that the power increases as we add in controls to every case but it appears to plateau as R approaches 10. Adding many more controls does not add any new information to the model.

Application to type 1 diabetes

In previous studies Type 1 diabetes (T1D) has demonstrated a strong association with the HLA region of chromosome 6 [36]. To illustrate our method we consider joint analysis of 19 T1D trios with just over 2000 independent controls. All family and control samples are of European ancestry; for details about the data see Luca *et al.* [17]. First, we estimated the ancestry of the controls and plotted them against their two most significant axes of genetic variation using SpectralGEM. We then projected the 19 trio probands onto the control's eigenmap using the Nystrom approximation (Supplementary Figure 6). The fullmatch algorithm identified 19 distinct strata, each including exactly one trio proband, and between 19 and 359 controls. We call these unbalanced strata 'all controls', to indicate that we matched the full sample of controls. For our analysis we also chose the closest m controls to each case, where $m = 5$ and 10. For SNPs in the HLA region, we evaluated mCLRs success at detecting association with T1D. From our results it is apparent that as m increases our power to detect certain SNPs increases (Figure 7). The best p -value is over an order of magnitude better for $m = 10$ than $m = 0$ and well over 2 orders of magnitude better when using all of the controls. The strongest signals occur at SNPs rs241427 and rs9273363 located near the confirmed T1D susceptibility locus HLA-DQB1 within the HLA class II region [36].

Discussion

In a genetic association study, as the sample size grows, the effect of population substructure becomes more serious. If not modeled correctly, even subtle correlations between individuals of common ancestry begin to affect the distribution of tests of association causing a greater number of spurious associations [7-9]. Thus, for sound inferences from GWAs, especially those using samples of diverse ancestry, it is important to control for ancestry differentiation. Family-based samples and association analyses, such as trios of parent and affected offspring and analyzed by FBAT [37], are robust to population structure.

Current data repositories include samples of families large enough to generate intriguing results, but typically not large enough to yield genome wide significance for variants with small to moderate effect size. We propose a hybrid design we call mCLR that simultaneously utilizes the information from unrelated case-control samples, trio data, and

freely available controls obtained from a generic database. The method builds on the principal of matching by ancestry to remove the potential confounding effects of population stratification. Thus trio probands are matched to unrelated controls based on ancestry, and pseudo-controls based on genetic transmission. Unrelated cases are matched to unrelated controls based on ancestry. Both family-based and case-control study designs produce genetically matched strata consisting of a single case and one or more controls. These data can be analyzed using the conditional logistic model. Simulations show that the resulting method is both powerful and robust to population stratification. Thus through careful matching, the mCLR approach has the advantages of family-based studies, but the enhanced power of a case-control study.

A cautionary note about combining case-control and family-based samples is worthwhile. While mCLR controls for ancestry, it cannot control for hidden biases inherent in the designs. For example, family-based studies require relatively intact families [37], which could impose conditions quite different from those inherent in a case-control collection. Combining the data by mCLR has advantages for a genetic study only when case-control and family-based samples are not strongly differentiated for risk factors correlated with the genetics of risk.

Up to this point we considered only families consisting of trios. Our method extends to more general family-based designs. Larger pedigrees can be split into trios. When one or more parent is not genotyped, transmissions can be inferred, provided a sufficient number of relatives have been sampled [38]. When families include multiple affected siblings, the contributions of multiple transmissions are independent if there are no disease loci in the region under examination. Nonindependence due to linkage is usually handled using a robust Huber-White variance estimation [39, 40]. This method makes an empirical adjustment to the variance/covariance matrix of the parameter estimate to account for the correlation among siblings [41-43].

Other methods have been proposed for the joint analysis of family-based and unrelated samples. Zhu *et al.* [44] suggest a model that utilizes PCA to estimate the genetic ancestry of sampled individuals. The effect of ancestry is regressed out of both genotypes and phenotypes prior to testing for association. Rather than modeling transmissions, the approach treats families as correlated clusters of observations. This is in contrast to our method, which preserves the family structure inherent in the trio design. Finally, these authors assume that parents and offspring are phenotyped, which is often not the case in practice. Another more general approach is known as ROADTRIPS [45]. This procedure uses a covariance matrix estimated from genome-screen data to correct for unknown population and pedigree structure, as well as accounting for known pedigree information. While this method has the advantage of flexibility, it does not model transmissions within families. Both of these methods work best if the cases and controls are sampled from a common population. When the controls are obtained as a sample of convenience, approaches that regress out the effect of ancestry are not fully robust to confounding [17].

Methods such as mCLR, and in fact any related methods controlling for heterogeneity statistically, require good eigenmaps. We show such a map, one that successfully identifies

clusters of genetically distinct individuals, requires a sufficiently diverse, yet representative base sample. It is not sufficient to use only the most genetically similar and diverse populations available for the base sample. Genetic isolates alone are not ideal for creating an eigenmap meant to differentiate typical individuals in modern populations. A large sample of convenience is also not optimal. A smaller number of individuals chosen to represent the full range of ancestry in the sample of interest will produce a better eigenmap. In the near future, when cases and controls will be matched prior to genomewide sequencing, sound eigenmaps are likely to be even more important.

Genetic matching can be achieved via PCA [10, 11, 17], the SGA [12], or based on measures of identity by state [18]. Various software programs are available for estimating ancestry; for example, Eigenstrat [11], PLINK [46], and SpectralGEM [12]. Given pairwise distances or similarities, strata can be formed using the fullmatch algorithm, implemented in *R* (cran.r-project.org) via the optmatch library [31]. Finally, provided the pseudo-controls are delineated, and the matched strata defined, analysis can be performed using any standard software for conditional logistic analysis. For example, the clogit function, part of the survival library is available in *R*. We provide a suite of *R* programs to implement all of the algorithms necessary to perform the full set of analyses described herein from our website (see mCLR source code).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Institute of Mental Health grant MH057881 and Autism Speaks grant for the Autism Genome Project (awarded to BD and KR), Department of Defense grant W81XWH-07-1-0619 and W81XWH-08-1-0428 (awarded to MT), and Office of Naval Research grant N0014-08-1-0673.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*. 2009; 106(23):9362–9367.
2. Manolio TA, Brooks LD, Collins FS. A hapmap harvest of insights into the genetics of common diseases. *Journal of Clinical Investigation*. 2008; 118(5):1590–1605. [PubMed: 18451988]
3. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverber MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhardt AH, Targan SR, Xavier RJ, the NIDDK IBD Genetics Consortium; Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot J-P, de Vos Severine Vermeire M, Louis E, the Belgian-French IBD Consortium; the Wellcome Trust Case Control Consortium. Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghorri J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly M. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nature Genetics*. 2008; 40(8):955–962. [PubMed: 18587394]
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL,

- Mackay TFC, McCarroll SA, Visscher PM. Finding missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. [PubMed: 19812666]
5. Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K. Genome-wide association database developed in the Japanese integrated database project. *Journal of Human Genetics*. 2009; 54(9):543–546. [PubMed: 19629137]
 6. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994; 265(5181):2037–2048. [PubMed: 8091226]
 7. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55(4):997–1004. [PubMed: 11315092]
 8. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*. 2001; 60(3):155–166. [PubMed: 11855950]
 9. Devlin B, Bacanu SA, Roeder K. Genomic control to the extreme. *Nature Genetics*. 2004; 36(11):1129–1130. [PubMed: 15514657]
 10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006; 38:904–909. [PubMed: 16862161]
 11. Patterson NJ, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genetics*. 2006; 2(12):e190. DOI: 10.1371/journal.pgen.0020190. [PubMed: 17194218]
 12. Lee AB, Luca D, Klei L, Devlin B, Roeder K. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*. 2010; 34:51–59. [PubMed: 19455578]
 13. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D, Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvanen AC, Meneton P, Hercberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Genin E, Cardon LR, Lathrop M. Investigation of the fine structure of european populations with applications to disease association studies. *European Journal of Human Genetics*. 2008; 16:1413–1429. [PubMed: 19020537]
 14. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. Genes mirror geography within Europe. *Nature*. 2008; 456:98–101. [PubMed: 18758442]
 15. Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *American Journal of Human Genetics*. 2007; 80(5):921–930. [PubMed: 17436246]
 16. Lee WC. Case-control association studies with matching and genomic controlling. *Genetic Epidemiology*. 2004; 27(1):1–13. [PubMed: 15185398]
 17. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *American Journal of Human Genetics*. 2008; 82(2):453–463. [PubMed: 18252225]
 18. Guan W, Liang L, Boehnke M, Abecasis GR. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genetic Epidemiology*. 2009; 33(6):508–517. [PubMed: 19170134]
 19. Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*. 1987; 57:455–464.
 20. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American Journal of Human Genetics*. 1993; 52:506–516. [PubMed: 8447318]
 21. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics*. 1993; 53:1114–1126. [PubMed: 8213835]
 22. Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and parents. *American Journal of Human Genetics*. 1994; 55:402–409. [PubMed: 8037216]
 23. Self SG, Longton G, Kopecky KJ, Liang KY. On estimating hla/disease association with application to a study of aplastic anemia. *Biometrics*. 1991; 47:53–61. [PubMed: 2049513]

24. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms with a gene using case/control or family data: application to hla in type I diabetes. *American Journal of Human Genetics*. 2002; 70:124–141. [PubMed: 11719900]
25. Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *European Journal of Human Genetics*. 2004; 12:964–970. [PubMed: 15340361]
26. Epstein MP, Veal CD, Trembath R, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. *American Journal of Human Genetics*. 2005; 76:592–608. [PubMed: 15712104]
27. Bengio Y, Delalleau O, Le Roux N, Paiement JF, Vincent P, Ouimet M. Learning eigenfunctions links spectral embedding and kernel pca. *Neural Computation*. 2004; 16(10):2197–2219. [PubMed: 15333211]
28. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 2008; 319(5866):1100–1104. [PubMed: 18292342]
29. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH. The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *American Journal of Human Genetics*. 2008; 83(3):347–358. [PubMed: 18760391]
30. Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. Characterization of multilocus linkage disequilibrium. *Genetic Epidemiology*. 2005; 28(3):193–206. [PubMed: 15637716]
31. Hansen BB. Full matching in an observational study of coaching for the (sat). *Journal of the American Statistical Association*. 2004; 99:609–618.
32. McVean G. A genealogical interpretation of principal components analysis. *PLoS Genetics*. 2009; 5(10):e1000, 686.
33. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*. 2002; 14:585–591.
34. Balding D, Nichols R. A method for quantifying differentiation between populations at multi-allelic locus and its implications for investigating identify and paternity. *Genetica*. 1995; 3:3–12. [PubMed: 7607457]
35. Breslow N. Design and analysis of case-control studies. *Annual Review of Public Health*. 1982; 3:29–54.
36. Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SCL, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farral M, Barnett AH, Bain SC, Todd JA. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature*. 1994; 371:130–136. [PubMed: 8072542]
37. Lange C, Laird NM. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genetic Epidemiology*. 2002; 23(2):165–180. [PubMed: 12214309]
38. Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test. *American Journal of Human Genetics*. 1999; 64(3):861–870. [PubMed: 10053021]
39. Huber P. The behaviour of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*. 1967; 1:221–233.
40. White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50:1–25.
41. Schaid DJ. Likelihoods and tdt for the case-parent design. *Genetic Epidemiology*. 1999; 16:250–260. [PubMed: 10096688]
42. Clayton D. Tdt for uncertain haplotypes. *American Journal of Human Genetics*. 1999; 65:1170–1177. [PubMed: 10486336]

43. Cordell HJ. Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. *Genetic Epidemiology*. 2004; 26(3):186–205. [PubMed: 15022206]
44. Zhu X, Li S, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. *American Journal of Human Genetics*. 2008; 82:352–365. [PubMed: 18252216]
45. Thornton T, McPeck MS. Roadtrips: case–control association testing with partially or completely unknown population and pedigree structure. *American Journal of Human Genetics*. 2010; 86:172–184. [PubMed: 20137780]
46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 2007; 81(3):559–575. [PubMed: 17701901]

Figure 1.

HapMap trios matched by ancestry to POPRES controls. The 30 offsprings from the HapMap, CEU sample, trios serve as cases and the 2184 individuals of European ancestry from the POPRES data serve as controls. (a) The plot displays the top two principal components of ancestry for cases (red) and controls (black) obtained using SGA. Based on the distribution of points in the eigenmap, many available controls would not be good matches to the HapMap trios. Only those delineated in blue are considered further. Each case is matched to one or more controls that are genetically similar based on the eigenvectors. (b) Distance between controls and closest case when matching in a random subset drawn from the full sample of controls versus (c) the distances when the controls consist of the restricted sample delineated in blue.

Population	Cluster					
	A	B	C	D	E	F
Biaka Pygmies	22	0	0	0	1	0
Mandenka	0	23	0	0	0	0
Yoruba	0	0	21	0	0	0
Mbuti Pygmies	0	0	0	11	0	0
Bantu	0	0	0	0	14	5
(a) San	0	0	0	0	0	4

Population	Cluster					
	A	B	C	D	E	F
Japanese	28	1	0	0	0	0
Yakut	0	25	0	0	0	0
Yizu	0	0	9	0	1	0
Tu	0	0	9	0	0	1
Naxi	0	0	7	0	0	0
Daur	0	0	0	0	0	9
Hezhen	0	0	0	0	0	8
Mongola	0	3	0	0	0	7
Xibo	0	1	0	0	0	8
Oroqen	0	1	0	0	0	8
Lahu	0	0	0	6	0	0
Cambodians	0	0	0	10	0	0
Dai	0	0	0	10	0	0
Han	0	0	1	11	30	1
Tujia	0	0	0	0	10	0
Miaozu	0	0	0	3	7	0
(b) She	0	0	0	0	9	0

Population	Cluster				
	A	B	C	D	E
Sardinian	28	0	0	0	0
French Basque	0	24	0	0	0
Russian	0	0	25	0	0
Adygei	0	0	0	17	0
French	0	0	0	0	28
North Italian	0	0	0	0	13
Orcadian	0	0	0	0	13
(c) Tuscan	0	0	0	0	8

Figure 2.

(a) African, (b) East Asian, and (c) European clusters identified by SGA. The 51 population samples within HGDP were analyzed to identify homogeneous clusters using SGA applied to continental samples. Analysis was performed separately for each continent using SpectralGEM. Population labels were ignored in the analysis. The display is organized to emphasize when a population or group of populations falls into a common cluster. Groups of populations that fall into a common cluster are often from a common region; see Supplementary Figures 1 and 2.

Figure 3.

HGDP and POPRES eigenmap representations plotted for various ancestry bases. In each panel, the eigenvectors (labeled PC) are calculated using a portion of the data, called the base. The remaining samples are projected using the Nystrom approximation. For each eigenmap we show only the top two principal components, POPRES (turquoise) and HGDP (black). (a) Base = HGDP, projected = POPRES; (b) Base = POPRES, projected = HGDP; (c) Base = HGDP + half of POPRES, projected = half of POPRES; and (d) Base = half of the balanced subset of countries including HGDP, projected = remaining half of the balanced subset.

Figure 4.

Comparing ancestry of selected groups in HGDP versus POPRES for the top two principal components. SGA was performed using the balanced sample (Figure 3(d)). Individuals selected for comparison from POPRES and HGDP are highlighted using colors other than turquoise. (a) HGDP-French (black) versus POPRES-French (fuchsia); (b) HGDP-Orcadian (black) versus POPRES-British & Irish (fuchsia); (c) HGDP-Tuscan (black) and HGDP-N. Italian (blue) versus POPRES-Italian (fuchsia); and (d) HGDP-French Basque (black) versus POPRES-French (fuchsia), POPRES-Spanish & Portuguese (blue).

Figure 5.

Type I error analysis at $\alpha = 0.05$. Solid line represents Type I error for *mCLR* method and dashed line represents Type I error for *combined association analysis* with $F_{st} = 0.05$ (a), $F_{st} = 0.01$ (b), and $F_{st} = 0.001$ (c). Results are based on 5000 replications of 500 unrelated controls and 500 trios.

Figure 6.

Power analysis at $\alpha = 0.05$. (a) *mCLR* method (solid line) versus *combined association analysis* (dashed line). Results are based on 5000 replications of 500 unrelated controls and 500 trios. (b) Power of *mCLR* method plotted against the theoretical ratio of controls to case (R). Results are based on 10 000 replications under the assumption that $\psi = 1.3, 1.4, 1.5$.

Figure 7. Association between HLA markers and Type 1 diabetes. $-\log_{10}(p\text{-values})$ are plotted versus individual SNPs in the HLA region of chromosome 6. (a) All controls matched; (b) 1:10 matching; (c) 1:5 matching; and (d) Trios only. The strongest association occurs for rs241427 (diamond) and next strongest for rs9273363 (triangle).

Table I

Family type probabilities for trios. $t=(1-p_c)^2+2\psi p_c(1-p_c)+\psi^2 p_c^2$.

Family type		
Parental genotypes	Proband genotype	f_k
$AA \times AA$	AA	$p_c^2 \psi^2 / t$
$AA \times AB$	AA	$2 p_c^3 (1-p_c) \psi^2 / t$
$AA \times AB$	AB	$2 p_c^3 (1-p_c) \psi / t$
$AA \times BB$	AB	$2 p_c^2 q_c^2 \psi / t$
$AB \times AB$	AA	$p_c^2 (1-p_c)^2 \psi^2 / t$
$AB \times AB$	AB	$2 p_c^2 (1-p_c)^2 \psi / t$
$AB \times AB$	BB	$p_c^2 (1-p_c)^2 / t$
$AB \times BB$	AB	$2 p_c (1-p_c)^3 \psi / t$
$AB \times BB$	BB	$2 p_c (1-p_c)^3 / t$
$BB \times BB$	BB	$(1-p_c)^4 / t$