frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 25 October 2013
doi: 10.3389/fgene.2013.00215

Semi-supervised spectral clustering with application to detect population stratification

Binghui Liu[1,2], Xiaotong Shen[2] and Wei Pan[1]*

[1] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA
[2] School of Statistics, University of Minnesota, Minneapolis, MN, USA

Edited by:
Shuang Wang, Columbia University, USA

Reviewed by:
Weihua Guan, University of Minnesota, USA
Ren-Hua Chung, National Health Research Institutes, Taiwan

***Correspondence:**
Wei Pan, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455, USA
e-mail: weip@biostat.umn.edu

In genetic association studies, unaccounted population stratification can cause spurious associations in a discovery process of identifying disease-associated genetic markers. In such a situation, prior information is often available for some subjects' population identities. To leverage the additional information, we propose a semi-supervised clustering approach for detecting population stratification. This approach maintains the advantages of spectral clustering, while is integrated with the additional identity information, leading to sharper clustering performance. To demonstrate utility of our approach, we analyze a whole-genome sequencing dataset from the 1000 Genomes Project, consisting of the genotypes of 607 individuals sampled from three continental groups involving 10 subpopulations. This is compared against a semi-supervised spectral clustering method, in addition to a spectral clustering method, with the known subpopulation information by the Rand index and an adjusted Rand (ARand) index. The numerical results suggest that the proposed method outperforms its competitors in detecting population stratification.

Keywords: clustering, genome-wide association studies (GWAS), population stratification, semi-supervised spectral clustering, single nucleotide variant (SNV)

1. INTRODUCTION

With the rapid advance of high-throughput technologies, genome-wide association studies (GWAS) and whole-exome or whole-genome sequencing studies have become popular (International HapMap Consortium, 2003). However, in a population-based association study, presence of undetected population stratification, also referred to as the population structure, becomes a potential issue leading to false discovery (Marchini et al., 2004). Population stratification occurs in presence of a systematic difference in allele frequencies between cases and controls due to different ancestries. One direct consequence of ignoring population stratification is inflated false positives and false negatives (Lander and Schork, 1994; Hirschhorn and Daly, 2005; Thomas et al., 2005).

Clustering has been an effective means to detect and describe known or cryptic population stratification (Paschou et al., 2010). For detecting or adjusting for population stratification, three major methods have been proposed, including genomic control (Devlin and Roeder, 1999; Devlin et al., 2004), structured association mapping and other clustering methods (Pritchard et al., 2000; Satten et al., 2001), and principal component analysis [PCA, (Patterson et al., 2006; Zhang et al., 2012)] and spectral methods (Lee et al., 2009; Zhang et al., 2009). As argued in Lee et al. (2009), different methods may be applicable in different situations, for instance, a combination of PCA and a clustering method may be preferable when the method is applied to preprocess in association studies. Despite progress, issues remain. One important issue is how to utilize additional prior information to enhance clustering performance to adjust for population stratification. In a situation where some subjects' population identities are known *priori*, a semi-supervised approach is more suitable. Towards this end, we propose two methods to detect population stratification, that is, semi-supervised clustering methods that are integrated with PCA and another clustering method, respectively. These methods are developed to (1) integrate the prior information for clustering, (2) to avoid that dense clusters are collapsed into a single group, whereas their sparser counterparts are divided into more multiple clusters, and (3) utilize the prior information to separate highly overlapped subpopulations.

For (1), we incorporate prior information through constraints, as in Grira et al. (2004). The constraints are expressed in terms of pairwise *must-links* and *cannot-links* imposed over a subset of the subjects with known population identities, where a must-link connects two subjects from the same subpopulation, whereas a cannot-link deals with different subpoluations.

For (2), we develop our semi-supervised clustering method based on a local scale spectral clustering method (Zelnik-manor and Perona, 2004). In some situations, subpopulations may not be in a same scale, then we consider a spectral clustering method involving a local scaling parameter to guard against potential disruptive influence caused by the different densities of the different subpoluations.

For (3), we introduce a continuous parameter to adjust similarities between subjects with cannot-links and those without any cannot-link, in addition to adjusting similarities between must-link pairs and cannot-link pairs. As indicated in the numerical results in Section 3.1, many pairs of subjects with cannot-links in two different subpoluations were assigned into one cluster by an existing semi-supervised method, which is in contrast to the proposed semi-supervised spectral clustering method.

The paper is organized as follows. Section 2 gives a motivating data example, and introduces the proposed methods. Section 3 presents our analysis of a low-coverage whole-genome sequencing data, published on the 1000 Genomes Project website. This is followed by a discussion in Section 4.

2. MATERIALS AND METHODS

2.1. DATA

In this study, we used a low-coverage whole-genome sequencing dataset to evaluate the performance of our semi-supervised spectral clustering algorithm. The processed data were downloaded from the 1000 Genomes Project (1000 Genomes Project Consortium 2010) web site http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-08.html. The phased data contain the DNA sequences of $n = 607$ individuals of three continental groups: Africans (AFR), Europeans (EUR) and Asians (ASN); there are 3, 4, and 3 subgroups in the three continental groups respectively (**Table 1**) after we removed three subgroups (2 PUR and 1 MXL) from the downloaded data due to their small sample sizes.

We used all the $p = 7, 459, 664$ SNVs appearing in all the three continental groups on chromosomes 1 to 22. In the 7,456,664 SNVs, there are 343,782 rare variants (RVs, with minor allele frequencies, MAFs $< 1\%$), 1,189,061 low frequency variants (LFVs, $1\% \leq$ MAFs $< 5\%$) and 5,926,821 common variants (CVs, MAFs $\geq 5\%$). There are 132,742, 525,440 and 1,107,080 monomorphic variants in each of the three continental groups: AFR, EUR and ASN, respectively, and there are 18,559 variants that are monomorphic in all the three continental groups. Furthermore, there are 101,279 variants that are monomorphic in AFR but polymorphic in EUR, and 67,661 variants that are monomorphic in AFR but polymorphic in ASN; there are 493,977 variants that are monomorphic in EUR but polymorphic in AFR, and 133,388 variants that are monomorphic in EUR but polymorphic in ASN; there are 1,041,999 variants that are monomorphic in ASN but polymorphic in AFR, and 715,028 variants that are monomorphic in ASN but polymorphic in EUR.

Denote the data by an $n \times p$ matrix Z, with rows indexing n individuals, and columns indexing p SNVs. For each SNV, we chose the minor allele as the reference allele. Let $Z_{ij} \in \{0, 1, 2\}$ be the number of minor alleles for SNV j of individual i. We centered each column (SNV) to have mean 0; denote the centered data matrix $Z_c = AZ$, where $A = I - \frac{1}{n}\mathbf{1}\mathbf{1}^t$ is an $n \times n$ centering matrix, I denotes the $n \times n$ identity matrix and $\mathbf{1}$ denotes the length-n vector with each entry equal to 1. Then, we used PCA for dimension reduction (Menozzi et al., 1978; Cavalli-Sforza et al., 1994): we computed the $n \times n$ sample covariance matrix $H = Z_c Z_c^t$, and then used the re-scaled eigenvectors of H as coordinates for subject i, $x_i = (\sqrt{\lambda_1}u_1(i), \ldots, \sqrt{\lambda_J}u_J(i))$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J \geq 0$ are the M largest eigenvalues of H and $u_j = (u_j(1), \ldots, u_j(n))^t$, $j \in \{1, \ldots, J\}$, are the corresponding eigenvectors. Typically, eigenvectors that correspond to large eigenvalues reveal important ancestry axes.

Table 1 | 10 subgroups of 607 individuals.

AFR:	YRI	LWK	ASW	
# Samples	78	67	24	
label	1	2	3	
EUR:	**GBR**	**FIN**	**CEU**	**TSI**
# Samples	43	36	90	92
label	4	5	6	7
ASN:	**CHS**	**CHB**	**JPT**	
# Samples	25	68	84	
label	8	9	a	

2.2. SEMI-SUPERVISED SPECTRAL CLUSTERING

Existing semi-supervised clustering methods can be categorized into two: search-based and similarity-based. The former is a modified clustering method in that the prior constraints are used to yield appropriate partitions (Demiriz et al., 1999; Wagstaff et al., 2001; Basu et al., 2002). The latter is a clustering method based on a modified similarity metric (Bilenko and Mooney, 2003; Xing et al., 2003; Yang et al., 2008). We think that the latter may be more efficient, since it embeds prior constraints only by simply modifying the similarity metric, while the former may use prior constraints to yield appropriate partitions in each iteration.

With this in mind, in this paper we developed a semi-supervised spectral clustering method to infer population structure. Before proposing this method in detail, we first review some spectral clustering algorithms, which were developed from the studies of weighted graph partitioning problems (Shi and Malik, 2000; Meila and Shi, 2001; Ng et al., 2001; Kannan et al., 2004). The spectral clustering algorithms are similarity-based. A popular choice for defining the similarity between a pair of subjects $(x_i, x_j)(i \neq j)$ is letting $W_{ij} = \exp(-||x_i - x_j||^2/\sigma_{ij}^2)$, where the scale parameter σ_{ij} controls the size of local neighborhoods in the weighted graph. Although a global scale $\sigma_{ij} = \sigma$ is often used, as mentioned in Zelnik-manor and Perona (2004), using a local scale parameter, $\sigma_{ij} = (\sigma_i\sigma_j)^{1/2}$ with $\sigma_i, \sigma_j > 0$, for each pair (i, j) may obtain better performance, especially when the clusters of the data have different volumes. Below we review the local scale spectral clustering algorithm proposed by Zelnik-manor and Perona (2004).

Given a set of n points $\{x_1, \ldots, x_n\}$ in the J-dimensional Euclidean space $\mathbf{R}^J$ and the neighborhood parameter T, cluster them into K clusters as follows:

1. Compute alocal scale $\sigma_i = d(x_i, x_{i_T})$ for each point x_i, where $d(., .)$ is the Euclidean distance metric and x_{i_T} is the T-th nearest neighbor of point x_i.
2. Form a weight matrix W with its ij-th element $W_{ij} = \exp(-\frac{d^2(x_i, x_j)}{\sigma_i\sigma_j})$ for each i and $j \in \{1, \ldots, n\}$ with $i \neq j$ and $W_{ii} = 0$ for each $i \in \{1, \ldots, n\}$.
3. Define D to be a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$ and construct the normalized Laplacian matrix $\mathcal{L} = I - D^{-1/2}WD^{-1/2}$.
4. Find $u_1, \ldots, u_K$, the smallest K eigenvectors of $\mathcal{L}$, and let U be the matrix containing the vectors $u_1, \ldots, u_K$ as columns.

5. For $i = 1, \dots, n$ and $j = 1, \dots, K$, let $U'_{ij} = U_{ij}/(\sum_{j'=1}^{K} U_{ij'}^2)^{1/2}$, and then U' is a matrix with elements U'_{ij}.
6. For $i = 1, \dots, n$, let $y_i = (y_i^{(1)}, \dots, y_i^{(K)})^t \in \mathbf{R}^K$ be the vector corresponding to the ith row of U', and then cluster the points $\{y_1, \dots, y_n\}$ with the K-means algorithm into K clusters $\{S_1, \dots, S_K\}$.
7. Assign the original point x_i to cluster S_j if and only if y_i is assigned to cluster S_j.

Let $\mathcal{M}$ denote the must-link matrix and $\mathcal{C}$ denote the cannot-link matrix for clustering n points $\{x_1, \dots, x_n\}$, where $\mathcal{M}_{ii} = \mathcal{C}_{ii} = 0$ $(i = 1, \dots, n)$ and for $i \neq j$, $\mathcal{M}_{ij} = 1$ (or $\mathcal{C}_{ij} = 1$) means that x_i, x_j are already known to be in the same (or different) cluster(s); $\mathcal{M}_{ij} = 0$ $(\mathcal{C}_{ij} = 0)$ means that we do not know whether x_i, x_j are in the same cluster. Given the must-link matrix $\mathcal{M}$ and the cannot-link matrix $\mathcal{C}$, Yang et al. (2008) proposed a semi-supervised algorithm by modifying the second step of the local scale spectral clustering algorithm above as follows: for each pair $i \neq j$, if $\mathcal{M}_{ij} = 1$, then let $W_{ij} \equiv 1$; if $\mathcal{C}_{ij} = 1$, then let $W_{ij} \equiv 0$. By letting $W_{ij} \equiv 1$ for a must-link pair (i, j), the algorithm forces the pair (i, j) to be clustered into the same cluster. However, in general, letting $W_{ij} \equiv 0$ for a cannot-link pair (i, j) may not force the pair (i, j) to be clustered into two different clusters. $W_{ij} \equiv 0$ only means that observations i and j are far away; if there exists an observation k such that W_{ik} and W_{kj} are large enough, i and j may still be clustered into one cluster. Thus, for embedding cannot-link information into a spectral clustering algorithm, only letting the weights of all cannot-link pairs to be zero is not enough. To avoid clustering a cannot-link pair (i, j) into one cluster, we adjust W_{ik} and W_{kj} for each k without any cannot-link information, based on which we propose a new semi-supervised spectral clustering algorithm.

Before introducing the algorithm, to make best use of the semi-supervised information, we may first adjust the must-link matrix and cannot-link matrix as follows: (1) Adjust the must-link matrix $\mathcal{M}$ such that: for each pair (i, j) $(i \neq j)$, $\mathcal{M}_{ij} = 1$ whenever there exists a $k \neq i, j$ such that $\mathcal{M}_{ik} = 1$ and $\mathcal{M}_{kj} = 1$; (2) Adjust the cannot-link matrix $\mathcal{C}$ such that: for each pair (i, j) $(i \neq j)$, $\mathcal{C}_{ij} = 1$ whenever there exists a $k \neq i, j$ such that $\mathcal{M}_{ik} = 1$ and $\mathcal{C}_{kj} = 1$. After the adjustment, if there exists any contradictory pair (i, j) $(i \neq j)$ with $\mathcal{C}_{ij} = 1$ and $\mathcal{M}_{ik} = 1$, to avoid being misled we will let $\mathcal{C}_{ij} = 0$ and $\mathcal{M}_{ik} = 0$.

In fact, though there have been much reported success with using pairwise constraints for clustering, there are two limitations (Davidson and Ravi, 2005; Davidson et al., 2006). First, if the constraints are poorly specified and then using cannot-link constraints may make the feasibility problem intractable (Davidson and Ravi, 2005); second, some constraints may have adverse effects to semi-supervised clustering (Davidson et al., 2006). There were some discussions about how to deal with the limitations, and accordingly some methods were specifically designed to overcome such limitations. Because the concern of the limitations is not the focus of this paper, we will not introduce these methods in detail.

Let $V_{\mathcal{M}} = \bigcup_{\mathcal{M}_{ij}=1}\{i, j\}$ and $V_{\mathcal{C}} = \bigcup_{\mathcal{C}_{ij}=1}\{i, j\}$. Now we are ready to show our semi-supervised spectral clustering (SSSC) algorithm.

Algorithm SSSC Given a set of n points $\mathscr{D} = \{x_1, \dots, x_n\}$ in $\mathbf{R}^J$, a must-link matrix $\mathcal{M}$, a cannot-link matrix $\mathcal{C}$, and the parameters $\alpha \in \{0, 1\}$, $\beta \geq 1$ and the neighborhood parameter T, cluster the points into K clusters as follows:

1. Compute the local scale $\sigma_i = d(x_i, x_{i_T})$ for each point x_i, where x_{i_T} is the T-th neighbor of point x_i.
2. Form the weight matrix W:
   a. Initially let $W_{ij} = \exp(-\frac{d^2(x_i, x_j)}{\sigma_i \sigma_j})$ for $i \neq j$ and $W_{ii} = 0$.
   b. For each i and j $(i \neq j)$, if $\mathcal{M}_{ij} = 1$, let $W_{ij} = 1$; and if $\mathcal{C}_{ij} = 1$, let $W_{ij} = 0$.
   c. For each $k \in V \backslash V_{\mathcal{C}}$, let $c_k = \arg\max_{c \in V_{\mathcal{C}}}(W_{kc})$. Then for each $i \in V_{\mathcal{C}}$ with $\mathcal{C}_{c_k i} = 1$, let W_{ik} and W_{ki} be replaced by W_{ik}/β.
   d. For each $k \in V \backslash V_{\mathcal{M}}$, let $m_k = \arg\max_{m \in V_{\mathcal{M}}}(W_{km})$. If $\alpha = 1$, then for each $j \in V_{\mathcal{M}}$ with $\mathcal{M}_{m_k j} = 1$, let W_{jk} and W_{kj} be replaced by $W_{m_k k}$.
3. Define D to be a diagonal matrix with $D_{ii} = \sum_j W_{ij}$ and construct the normalized Laplacian matrix $\mathcal{L} = I - D^{-1/2} W D^{-1/2}$.
4. Find $u_1, \dots, u_K$, the first K eigenvectors of $\mathcal{L}$, and let U be the matrix containing the vectors $u_1, \dots, u_K$ as columns.
5. For $i = 1, \dots, n$ and $j = 1, ..., K$, let $U'_{ij} = U_{ij}/(\sum_{j'=1}^{K} U_{ij'}^2)^{1/2}$, and then U' is a matrix with elements U'_{ij}.
6. For $i = 1, \dots, n$, let $y_i = (y_i^{(1)}, \dots y_i^{(K)})^t \in \mathbf{R}^K$ be the vector corresponding to the ith row of U', and then cluster the points $\{y_1, \dots, y_n\}$ with the K-means algorithm into K clusters $\{S_1, \dots, S_K\}$.
7. Assign the original point x_i to cluster S_j if and only if y_i was assigned to cluster S_j.

Note that in the Step 2.c of our new algorithm above, we believe that for each $k \in V \backslash V_{\mathcal{C}}$ and each cannot-link pair (i, j), if x_k is nearer to x_i, then it should be much farther away from x_j, because the distance between x_i and x_j has already been set to the maximum. Thus, we penalize the similarity between x_k and x_j by letting $W_{jk} = W_{kj} = W_{jk}/\beta$ $(\beta > 1)$. On the other hand, we set a parameter α to determine whether we force the similarities between a sample $k \in V \backslash V_{\mathcal{M}}$ and a must-link pair (i, j) to be the same. In fact, if $\alpha = 0$ and $\beta = 1$, then our algorithm reduces to that of Yang et al. (2008).

2.3. CHOOSING THE PARAMETERS

We develop a cross-validation procedure to choose the parameters for the **Algorithm SSSC**, modified from a criterion used in Tibshirani and Walther (2005) for the K-mean clustering. In addition, we borrow the idea of cluster reproducibility index (RI) (Shen et al., 2009) to define a new prediction strength. We summarize the procedure as follows.

Given a data set $\mathscr{D}$ and a candidate set of parameters $\Theta = \mathscr{K} \times \mathscr{A} \times \mathscr{B} \times \mathscr{T}$, where $\mathscr{K}$ and $\mathscr{T}$ are sets of positive integers, $\mathscr{A} = \{0, 1\}$, and $\mathscr{B}$ is a set of real numbers equal to or larger than 1. Randomly permute the sample index set $V = [N]$ of $\mathscr{D}$, and then partition the permuted sample index set into two roughly equal parts. Select one part as the test index subset V^{te} for the test data $\mathscr{D}^{te} = \{X_n\colon n \in V^{te}\}$ and take the remaining part as the training index subset V^{tr} for the training data $\mathscr{D}^{tr} = \{X_n\colon n \in V^{tr}\}$. Let $\mathcal{M}^{tr} = \mathcal{M}_{V^{tr}V^{tr}}$, $\mathcal{M}^{te} = \mathcal{M}_{V^{te}V^{te}}$, $\mathcal{C}^{tr} = \mathcal{C}_{V^{tr}V^{tr}}$ and $\mathcal{C}^{te} = \mathcal{C}_{V^{te}V^{te}}$. For each $\theta = (K, \alpha, \beta, T) \in \Theta$, apply **Algorithm SSSC** to divide $\mathscr{D}^{tr}$ into K clusters with parameters α, β, T and the must-link matrix $\mathcal{M}^{tr}$, the cannot-link matrix $\mathcal{C}^{tr}$; apply **Algorithm SSSC** to divide $\mathscr{D}^{te}$ into K clusters with parameters α, β, T and the must-link matrix $\mathcal{M}^{te}$, the cannot-link matrix $\mathcal{C}^{te}$. Let l_{tr} and l_{te} denote the corresponding clustering assignments. Divide the test data $\mathscr{D}^{te}$ into K clusters under the guidance of l_{tr}, that is, assign each sample in $\mathscr{D}^{te}$ into the closest cluster of $\mathscr{D}^{tr}$ characterized by l_{tr} in the sense of the Euclidean distance, and then let $l_{te|tr}$ denote the corresponding clustering assignment. Note that here the distance between a sample and a cluster is defined as the minimum distance between this sample and each sample in the cluster. Next, compute the adjusted Rand index (Hubert and Arabie, 1985) between $l_{te|tr}$ and l_{te} as the prediction strength. Repeat the above steps for a number of times with different randomly selected permuted samples, and finally choose $\hat{\theta} = (\hat{K}, \hat{\alpha}, \hat{\beta}, \hat{T}) \in \Theta$ with the highest average prediction strength.

Note that while using PCA for dimension reduction in Section 2.1, we did not mention how to choose an appropriate number of PCs. There are many studies about this problem for traditional PCA, such as Jackson (1991), Jolliffe (2002) and Pedro et al. (2005). Because in this paper we only focus on the performance of a clustering algorithm, we propose using a special procedure that is related to the clustering performance. In fact, we view the number of PCs as a parameter and then decide it in the above cross-validation procedure. Especially, we first choose $\hat{\theta}_J \in \Theta$ using the above cross-validation procedure for each J in a set of candidate numbers of PCs $\mathcal{J}$, and then choose $\hat{J} \in \mathcal{J}$ with the highest average prediction strength among $\mathcal{J}$ as the best fitted number of PCs.

3. RESULTS

3.1. MAIN RESULTS

We used all the SNVs appeared in all the three continental groups in chromosomes 1-22 to extract the top t principle components (PCs). As shown in the left panel of **Figure 1**, the top 2 PCs could completely separate the three continental groups. However, some subgroups could not be completely separated. We used the local scale spectral clustering algorithm introduced in Section Methods to cluster the 607 t-dimension vectors into 10 clusters. As shown in **Table 2** and **Figure 1**, subgroup GBR ('4') cannot be completely distinguished from CEU ('6'); CHS ('8') cannot be completely distinguished from CHB ('9').

The spectral clustering algorithm used above is an unsupervised clustering algorithm without using any additional clustering information. However, in many cases, partial knowledge is available concerning pairwise (must-link or cannot-link) constraints among a subset of subjects. Thus, we propose a semi-supervised local scale spectral clustering algorithm to make use of the pre-known constraints. We show the performance of our algorithm by varying the number of available must-link or cannot-link constraints. We let SSR denote the semi-supervised ratio, and randomly selected a fraction SSR of individuals from each subgroup. Then we obtained a must-link matrix and a cannot-link matrix according to the selected individuals and their subgroup identities, which were input to our semi-supervised algorithm. We used the algorithm in Yang et al. (2008) and our new proposed algorithm to cluster the 607 individuals into 10 clusters with the top 10, 20, and 30 PCs respectively, and then compared the Rand index (Rand, 1971) and an adjusted Rand (ARand) index (Hubert and Arabie, 1985) between the true subgroups and the clustering results. We repeated this process for 100 times and at each time we randomly selected some individuals for getting the pre-known must-link matrix and cannot-link matrix by setting a different seed in **R** software. Then we indicated the average results of these 100 simulations in **Figure 2**. From **Figure 2**, we can see that when using the top 10, 20 and 30 PCs for clustering, our algorithm performed much better than the existing one with ($\alpha = 0$, $\beta = 1$) (Yang et al., 2008) in terms of the Rand index and adjusted Rand index for almost all the values of SSR. It is clear that the blue vertical lines for our new algorithm appeared with smaller SSR values, indicating that our algorithm made use of the given semi-supervised information more efficiently. In fact, while using other numbers of the top PCs, we also obtained similar results (not shown). Additionally, here in our new algorithm we used $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$.

Tables 3, **4** present the numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using the existing SSSC algorithm ($\alpha = 0$, $\beta = 1$) (Yang et al., 2008) and our **Algorithm**

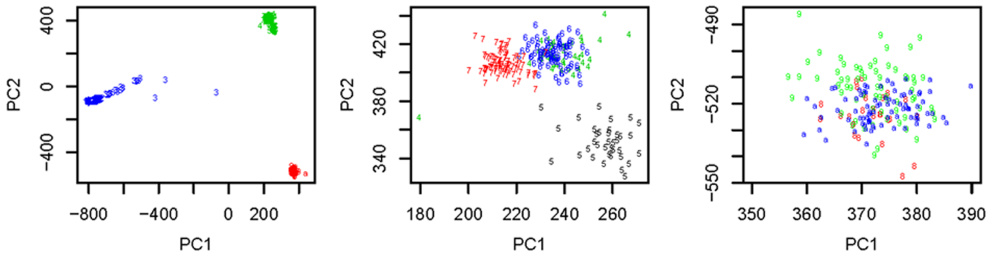

FIGURE 1 | Left: Top: 2 PCs after PCA dimension reduction for the SNV data from chromosomes 1-22, where subjects are self-identified as EUR (green), AFR (blue) and ASN (red). **Middle:** Top 2 PCs for EUR group: GBR (green), FIN (black), CEU (blue), TSI (red); **Right:** Top 2 PCs for ASN group: CHS (red), CHB (green), JPT (blue).

SSSC with SSR = 0.5. We can see that our new algorithm performed much better than the existing one.

To further illustrate the difference among the unsupervised local scale spectral clustering algorithm, the existing SSSC algorithm ($\alpha = 0$, $\beta = 1$) (Yang et al., 2008) and our new SSSC algorithm ($\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$), we plotted the first two co-ordinates (of y_i's in Step 6) for each of the three algorithms (see **Figure 3**). To better observe the separation between the two subgroups CHS ('8') and CHB ('9'), we particularly plotted for the two subgroups, where the colors of the subjects in CHB were still kept red, however, those in CHS were changed to black (see the right three sub-figures of **Figure 3**). The top two sub-figures are for the unsupervised local scale spectral clustering algorithm, the middle two are for the existing SSSC algorithm ($\alpha = 0$, $\beta = 1$) and the bottom two are for our SSSC algorithm ($\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$). From the first two sub-figures of **Figure 3** and **Table 2**, we see that the two pairs of subgroups, GBR-CEU and CHS-CHB were inseparable, respectively. Then for the middle two sub-figures of **Figure 3** and **Table 3**, by adjusting the similarities between must-link pairs to be 1 and those between cannot-link pairs to be 0, the subjects in GBR and CEU were a little more separable, however the subjects in CHS and CHB were still inseparable. Finally for the last two

Table 2 | The numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using the unsupervised local scale spectral clustering algorithm.

Groups	Subgroups	S1	S2	S3	S4	S5	S6	S7	S8	S9	Sa	All
AFR	YRI	78	0	0	0	0	0	0	0	0	0	78
	LWK	0	60	7	0	0	0	0	0	0	0	67
	ASW	0	0	1	23	0	0	0	0	0	0	24
EUR	GBR	0	0	4	0	39	0	0	0	0	0	43
	FIN	0	0	0	0	0	36	0	0	0	0	36
	CEU	0	0	0	0	90	0	0	0	0	0	90
	TSI	0	0	2	0	0	0	90	0	0	0	92
ASN	CHS	0	0	0	0	0	0	0	25	0	0	25
	CHB	0	0	0	0	0	0	0	68	0	0	68
	JPT	0	0	0	0	0	0	0	0	17	67	84

(Rand = 0.961, ARand = 0.821).

Table 3 | The numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using the existing SSSC algorithm ($\alpha = 0$, $\beta = 1$) with SSR = 0.5.

Groups	Subgroups	S1	S2	S3	S4	S5	S6	S7	S8	S9	Sa	All
AFR	YRI	78	0	0	0	0	0	0	0	0	0	78
	LWK	0	62	5	0	0	0	0	0	0	0	67
	ASW	0	0	1	23	0	0	0	0	0	0	24
EUR	GBR	0	0	2	0	39	2	0	0	0	0	43
	FIN	0	0	0	0	0	0	36	0	0	0	36
	CEU	0	0	0	0	17	73	0	0	0	0	90
	TSI	0	0	1	0	0	0	0	91	0	0	92
ASN	CHS	0	0	0	0	0	0	0	0	25	0	25
	CHB	0	0	0	0	0	0	0	0	68	0	68
	JPT	0	0	0	0	0	0	0	0	1	83	84

(Rand = 0.975, ARand = 0.881).

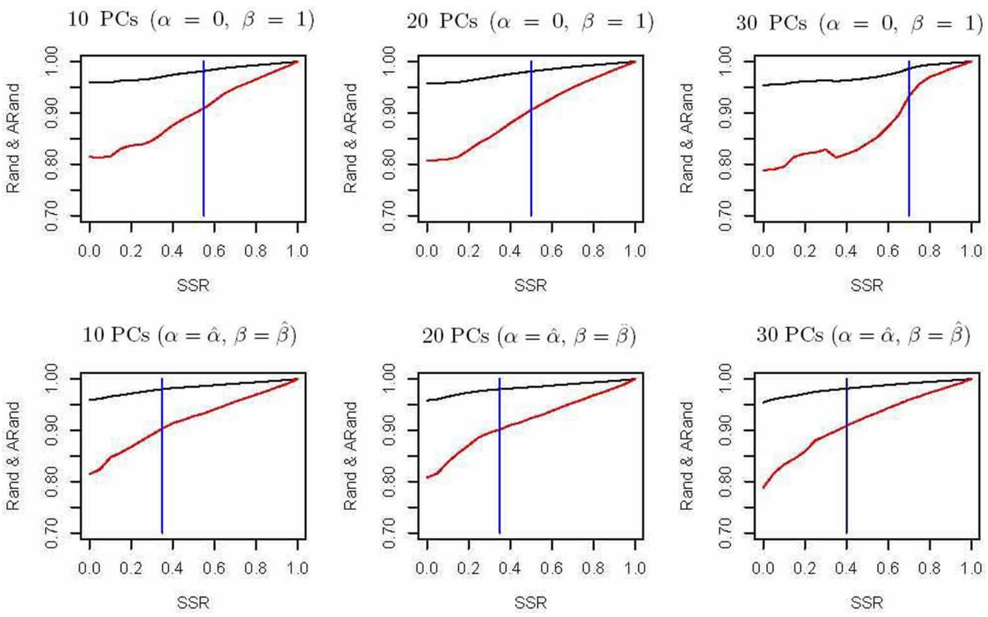

FIGURE 2 | Rand indices (black) and adjusted Rand indices (red) between the true subgroups and the clustering results with various numbers of the top PCs. A vertical blue line indicates the point of *SSR*, at which the Rand index and adjusted Rand index are both larger than 0.90 for the first time.

Table 4 | The numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using our new SSSC algorithm ($\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$) with SSR = 0.5.

Groups	Subgroups	S1	S2	S3	S4	S5	S6	S7	S8	S9	Sa	All
AFR	YRI	78	0	0	0	0	0	0	0	0	0	78
	LWK	0	67	0	0	0	0	0	0	0	0	67
	ASW	0	0	24	0	0	0	0	0	0	0	24
EUR	GBR	0	0	0	7	34	2	0	0	0	0	43
	FIN	0	0	0	0	0	0	36	0	0	0	36
	CEU	0	0	0	82	8	0	0	0	0	0	90
	TSI	0	0	0	0	0	92	0	0	0	0	92
ASN	CHS	0	0	0	0	0	0	0	20	5	0	25
	CHB	0	0	0	0	0	0	0	4	64	0	68
	JPT	0	0	0	0	0	0	0	0	1	83	84

(Rand = 0.984, ARand = 0.923).

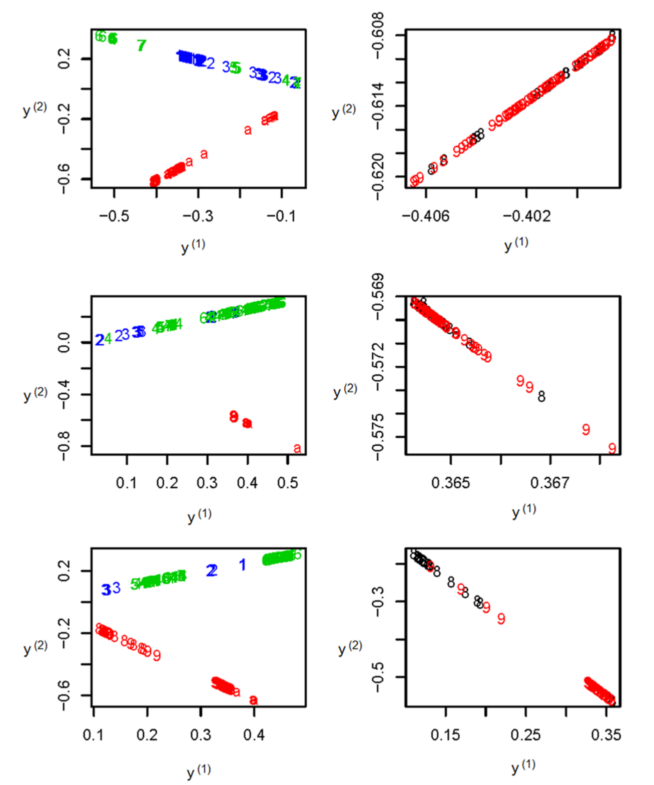

FIGURE 3 | The top two sub-figures are for the unsupervised local scale spectral clustering algorithm, the middle two are for SSSC algorithm ($\alpha = 0$, $\beta = 1$) and the bottom two are for our SSSC algorithm ($\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$). Note that in the right three sub-figures, the colors of the subjects in CHS ('8') were changed from red to black. Note that the $y^{(1)}$ and $y^{(2)}$ axes are just the first two co-ordinates of the data points $\{y_1, \ldots, y_n\}$ in Step 6 of **Algorithm SSSC**.

sub-figures of **Figure 3** and **Table 4**, we can see that the subjects in all the subgroups were more separable, and in particular, in the bottom right sub-figure the subjects in CHS and CHB were more separable.

3.2. SEMI-SUPERVISED CLUSTERING VERSUS CLASSIFICATION

We also did some numerical experiments to compare classification (supervised learning) with our semi-supervised clustering. For illustration and to have a easier problem for classification, we only took the individuals in the EUR continental group and used common variants (CVs) with minor allele frequencies (MAFs) greater than 5% on chromosome 1. Furthermore, we used PLINK (Purcell et al., 2007) to prune out correlated SNVs with a sliding window of size 50 (shifted by 5) and a threshold of $r^2 \leq 0.05$, after which we had 11,840 CVs.

First, we randomly chose a fraction SSR of individuals from each of the above four subgroups as semi-supervised information for our clustering algorithm and as the training data for a classification algorithm. We used penalized multinomial logistic regression with the Lasso or the Ridge penalty for classification; the penalization parameter was chosen by 5-fold cross-validation. We used the trained classifier to predict the subgroup labels for the remaining data, and combined the known labels in the training data and the predicted labels in the test data together to compare with the true labels in terms of the Rand indices and adjusted Rand indices. The top two sub-figures in **Figure 4** summarize the corresponding results based on 100 simulations; it is demonstrated that in our experiments, our semi-supervised clustering algorithm performed much better than both Lasso- and Ridge-penalized regression, especially for cases with low SSRs.

On the other hand, in some cases the given semi-supervised information may not involve all the four subgroups. For example, we only had information about a subset of subjects from the CEU and GBR subgroups, but not any from the other subgroups. As before, we randomly selected a fraction (*SSR*) of individuals from the CEU and GBR subgroups respectively; they were used as semi-supervised information for our algorithm and as training data for a classification algorithm. The bottom three sub-figures in **Figure 4** show the corresponding comparisons, indicating that our semi-supervised clustering algorithm performed overwhelmingly better than Lasso- and Ridge-penalized regression, because the classification algorithm predicted all the individuals of the unknown TSI or FIN subgroup as of either CEU or GBR subgroup. This illustrates an obvious advantage of a semi-supervised clustering approach for discovery of novel classes.

4. DISCUSSION

4.1. DIMENSION REDUCTION

In Section 3, we have demonstrated good performance of our algorithm with a few top PCs. In addition, we also obtained similar results (not shown) with other methods for dimension reduction, such as the spectral graph approach used in `SpectralR` and `SpectralGEM` (Lee et al., 2010). We used the spectral method in (Lee et al., 2010) for dimension reduction, then clustered the data into 10 clusters; we took the same procedure to compare the Rand index and adjusted Rand index values. **Figure 2**.

4.2. ALL OR A SUBSET OF SNVs?

In the previous section, we used all 7,459,664 SNVs appearing in all the three continental groups on chromosomes 1

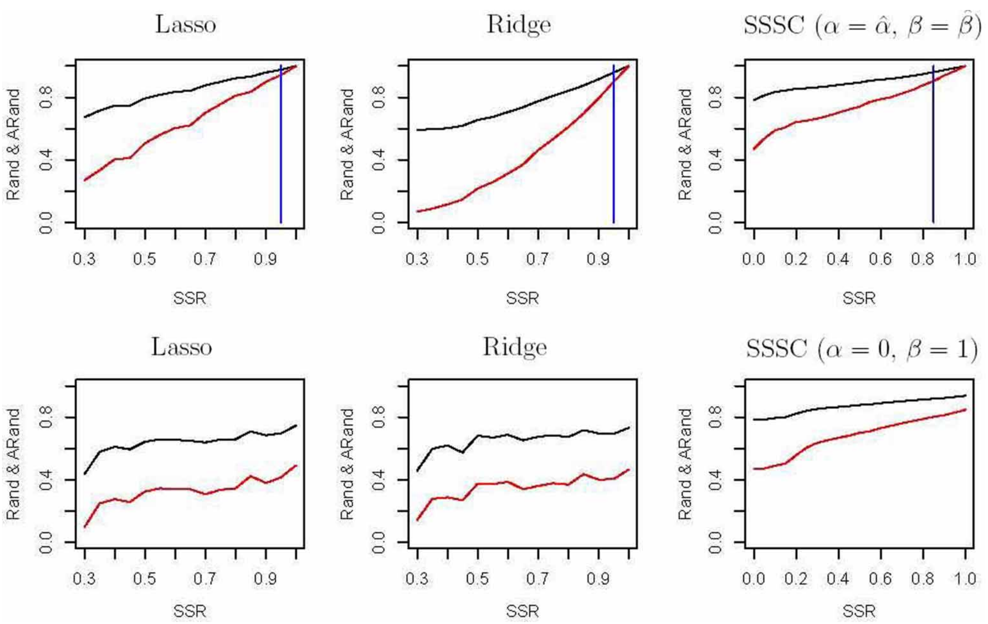

FIGURE 4 | Rand indices (black) and adjusted Rand indices (red) between the true subgroups and the SSSC or classification results. A vertical blue line indicates the point of SSR, at which the Rand and adjusted Rand are both bigger than 0.90 for the first time. The top three sub-figures demonstrate comparison of Lasso, Ridge and SSSC for the cases that semi-supervised information involves all the four subgroups, while the bottom three for the cases that semi-supervised information involves two subgroups.

Table 5 | The numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using the unsupervised local scale spectral clustering algorithm.

Groups	Subgroups	S1	S2	S3	S4	S5	S6	S7	S8	S9	Sa	All
AFR	YRI	59	19	0	0	0	0	0	0	0	0	78
	LWK	0	67	0	0	0	0	0	0	0	0	67
	ASW	24	0	0	0	0	0	0	0	0	0	24
EUR	GBR	0	0	0	43	0	0	0	0	0	0	43
	FIN	0	0	0	0	36	0	0	0	0	0	36
	CEU	0	0	0	90	0	0	0	0	0	0	90
	TSI	0	0	0	0	0	92	0	0	0	0	92
ASN	CHS	0	0	0	0	0	0	19	6	0	0	25
	CHB	0	0	0	0	0	0	21	7	40	0	68
	JPT	0	0	0	0	0	0	0	16	0	68	84

(Rand = 0.948, ARand = 0.758).

Table 6 | The numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using the existing SSSC algorithm ($\alpha = 0$, $\beta = 1$) with SSR = 0.5.

Groups	Subgroups	S1	S2	S3	S4	S5	S6	S7	S8	S9	Sa	All
AFR	YRI	78	0	0	0	0	0	0	0	0	0	78
	LWK	0	67	0	0	0	0	0	0	0	0	67
	ASW	0	0	24	0	0	0	0	0	0	0	24
EUR	GBR	0	0	0	43	0	0	0	0	0	0	43
	FIN	0	0	0	0	36	0	0	0	0	0	36
	CEU	0	0	0	12	0	78	0	0	0	0	90
	TSI	0	0	0	0	0	0	92	0	0	0	92
ASN	CHS	0	0	0	0	0	0	0	25	0	0	25
	CHB	0	0	0	0	0	0	0	10	58	0	68
	JPT	0	0	0	0	0	0	0	0	0	84	84

(Rand = 0.988, ARand = 0.941).

to 22 without pruning out SVNs in linkage disequilibrium. We used common variants (CVs) with minor allele frequencies (MAFs) greater than 5% on chromosomes 1 to 22, and used PLINK (Purcell et al., 2007) to prune out correlated SNVs with a sliding window of size 50 (shifted by 5) and a threshold of $r^2 \leq 0.5$, after which we had 1,022,090 CVs. Next, we used PCA for dimension reduction and then use the three algorithms to analyze the resulting data after dimension reduction.

Tables 5–7 present the numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using the unsupervised spectral clustering algorithm, the existing semi-supervised spectral clustering algorithm and our new algorithm with $SSR = 0.5$. From these results and those indicated by **Tables 2–4**, we see that using all the SNVs was better than using the pruned data in terms of the performance of the unsupervised spectral clustering. For the two semi-supervised spectral clustering algorithms, we find that while using the pruned data, the new

Table 7 | The numbers of subjects assigned to each of the 10 clusters based on the top 10 PCs using our new SSSC algorithm ($\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$) with SSR = 0.5.

Groups	Subgroups	S1	S2	S3	S4	S5	S6	S7	S8	S9	Sa	All
AFR	YRI	78	0	0	0	0	0	0	0	0	0	78
	LWK	0	67	0	0	0	0	0	0	0	0	67
	ASW	0	0	24	0	0	0	0	0	0	0	24
EUR	GBR	0	0	0	43	0	0	0	0	0	0	43
	FIN	0	0	0	0	36	0	0	0	0	0	36
	CEU	0	0	0	6	0	84	0	0	0	0	90
	TSI	0	0	0	0	0	0	92	0	0	0	92
ASN	CHS	0	0	0	0	0	0	0	25	0	0	25
	CHB	0	0	0	0	0	0	0	5	62	1	68
	JPT	0	0	0	0	0	0	0	0	0	84	84

(Rand = 0.991, ARand = 0.957).

semi-supervised spectral clustering algorithm still performed better than the existing one (Yang et al., 2008) as in Section 3.1 with all the SNVs.

4.3. LOCAL SCALE SPECTRAL CLUSTERING

Our semi-supervised spectral clustering algorithm is based on the local scale spectral clustering (Zelnik-manor and Perona, 2004), because we believe that local scales work better than choosing a single global scale for all pairs of subjects. In some situations the subgroups might not have the same scale; from our experience, given a fixed number of clusters, the subjects in a sparser group are more likely to be divided into more clusters, and the individuals in a denser group are more likely to be merged together. In these cases, it will be difficult to choose a suitable single global scale. In contrast, using local scales automatically adjusts for the heterogeneous scales in the subgroups. We did some experiments to compare the spectral clustering algorithms with a global scale and with local scales. We used several candidate values for a global scale, and found that even the best clustering result (in terms of the Rand indices and adjusted Rand indices) was almost the same as that obtained by using local scales. Because it is not the main point of this study, we do not show the detailed comparisons here.

5. CONCLUSIONS

We have proposed a new semi-supervised spectral clustering algorithm based on a more efficient use of the cannot link constraints in prior data. A whole-genome sequencing dataset from the 1000 Genomes Project was analyzed to compare the performance of our and other algorithms. In our experiments, unsupervised clustering algorithms could not completely separate some subgroups, such as the CEU-GBR and CHB-CHS subgroups; our semi-supervised spectral clustering algorithm, along with a subset of individuals with known subgroup identities, distinguished these subgroups much better. Our proposed method may be potentially useful in genetic association studies. Its extensions to other clustering (Thalamuthu et al., 2006) and dimension reduction approaches are to be studied.

ACKNOWLEDGMENTS

Funding: This work was supported by NIH grants R01GM081535, R01-HL65462 and R01-HL105397, by NSF grants DMS-0906616 and DMS-1207771, and by NSFC grant 11101182.

REFERENCES

Basu, S., Banerjee, A., and Mooney, R. J. (2002). "Semisupervised clustering by seeding," in *Proceedings of 19th International Conference on Machine Learning*, (San Francisco, CA), 27–34.

Bilenko, M., and Mooney, R. J. (2003). "Adaptive duplicate detection using learnable string similarity measures," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Washington, DC), 39–48. doi: 10.1145/956750.956759

Cavalli-Sforza, L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*, Princeton, NJ: Princeton Univ. Press.

Davidson, I., Basu, S., and Wagstaff, K. (2006). "Measuring constraint-set utility for partitional clustering algorithms," in *Proceedings of the Tenth European Principles and Practice of KDD (PKDD)*, (Berlin).

Davidson, I., and Ravi, S. (2005). "Clustering with constraints: feasibility issues and the k-means algorithm," in *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM-05)*, (Newport Beach, CA).

Demiriz, A., Bennett, K. P., and Embrechts, M. J. (1999). Semi-supervised clustering using genetic algorithms. *Artif. Neural Netw. Eng.* 809–814.

Devlin, B., Bacanu, S. A., and Roeder, K. (2004). Genomic control to the extreme. *Nat. Genet.* 36, 1129–1130. doi: 10.1038/ng1104-1129.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Grira, N., Crucianu, M., and Boujemaa, N. (2004). "Unsupervised and Semi-supervised Clustering: a brief survey," in *A Review of Machine Learning Techniques for Processing Multimedia Content*, Report of the MUSCLE European Network of Excellence (FP6).

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108. doi: 10.1038/nrg1521

Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 1993–1218. doi: 10.1007/BF01908075

International HapMap Consortium. (2003). The international HapMap project. *Nature* 426, 789–796. doi: 10.1038/nature02168

Jackson, D. A. (1991). *A User's Guide to Principal Components*. New York, NY: Wiley.

Jolliffe, I. T. (2002). *Principal Component Analysis*, (2nd Edn). New York, NY: Springer.

Kannan, R., Vempala, S., and Vetta, V. (2004). On spectral clustering-good, bad and spectral. *J. ACM* 51, 497–515.

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* 265, 2037–2048. doi: 10.1126/science.8091226

Lee, C., Abdool, A., and Huang, C. H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics* 10(Suppl. 1):S73. doi: 10.1186/1471-2105-10-S1-S73

Lee, A. B., Luca, D., and Roeder, K. (2010). A spectral graph approach to discovering genetic ancestry. *Ann. Appl. Stat.* 4, 179–202. doi: 10.1214/09-AOAS281

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat. Genet.* 36, 512–517. doi: 10.1186/1471-2156-6-S1-S109

Meila, M., and Shi, J. (2001). Learning segmentation by random walks, *Adv. Neural Inf. Process. Syst.* 470–477.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science* 201, 786–792. doi: 10.1126/science.356262

Ng, A. Y., Jordon, M. I., and Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* 14, 849–856.

Paschou, P., Lewis, J., Javed, A., and Drineas, P. (2010). Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J. Med. Genet.* 47, 835–847. doi: 10.1136/jmg.2010.078212

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigen analysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Pedro, R. P.-N., Donald, A. J., and Keith, M. S. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* 49, 974–997. doi: 10.1016/j.csda.2004.06.015

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations, *Am. J. Hum. Genet.* 67, 170–181. doi: 10.1086/302959

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.3410/f.1162373.622875

Purcell, S., and Sham, P. (2004). Properties of structured association approaches to detecting population stratification. *Hum. Hered.* 58, 93–107. doi: 10.1159/000083030

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Associat.* 66, 846–850. doi: 10.2307/2284239

Satten, G. A., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* 68, 466–477. doi: 10.1086/318195

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 22, 2906–2912. doi: 10.1093/bioinformatics/btp543

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 888–905. doi: 10.1109/34.868688

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288.

Tibshirani, R., and Walther, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Stat.* 14, 511–528. doi: 10.1198/106186005X59243

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* 22, 2405–2412. doi: 10.1093/bioinformatics/btl406

Thomas, D. C., Haile, R. W., and Duggan, D. (2005). Recent developments in genomewide association scans: a workshop summary and review. *Am. J. Hum. Genet.* 77, 337–345. doi: 10.1086/432962

Wagstaff, K., Cardie, C., Rogers, S., and Schroedl, S. (2001). "Constrained K-Means clustering with background knowledge." in *Proceedings of 18th International Conference on Machine Learning*, 577–584.

Xing, E. P., Ng, A. Y., Jordan M. I., and Russel, S. (2003). Distance metric learning, with application to clustering with side-information. *Adv. Neural Inf. Process. Sys.* 15, 505–512.

Yang, C., Zhang, X., Jiao, L., and Wang, G. (2008). Self-tuning semi-supervised spectral clustering. *Comput. Intell. Secur.* 1, 1–5. doi: 10.1109/CIS.2008.141

Zelnik-manor, L., and Perona, P. (2004). Self-tuning spectral clustering. *Adv. Neural Inf. Process. Syst.* 17, 1601–1608.

Zhang, Y., Guan, W., and Pan, W. (2012). Adjustment for population stratification via principal components in association association analysis of rare variants. *Genet. Epidemiol.* 37, 99–109. doi: 10.1002/gepi.21691

Zhang, J., Niyogi, P., and McPeek, M. S. (2009). Laplacian eigenfunctions learn population structure. *PLoS ONE* 4:e7928. doi: 10.1371/journal.pone.0007928

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 August 2013; paper pending published: 29 August 2013; accepted: 04 October 2013; published online: 25 October 2013.

Citation: Liu B, Shen X and Pan W (2013) Semi-supervised spectral clustering with application to detect population stratification. Front. Genet. ***4****:215. doi: 10.3389/fgene.2013.00215*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.