Relationships among evolutionary distance, the variance– covariance matrix, multidimensional scaling, and principal component analysis

Kazuharu Misawa^{1, 2*}

¹ Department of Human Genetics, Yokohama City University Graduate School of Medicine, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan.

² RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

* Correspondence: kazu_misawa@hotmail.com; Tel.: +81-45-787-2606, Fax: +81-45-786-5219.

Abstract

Principal component analyses (PCAs) are often used to visualize patterns of genetic variation in human populations. Previous studies showed a close correspondence between genetic and geographic distances. In such PCAs, the principal components are eigenvectors of the data's variance-covariance matrix, which is obtained by a genetic relationship matrix (GRM). However, it is difficult to apply GRM to multiallelic sites. In this paper, I showed that a PCA from GRM is equivalent to multidimensional scaling (MDS) from nucleotide differences. Therefore, a PCA can be conducted using nucleotide differences. The new method provided in this study provides a straightforward method to predict the effects of different demographic processes on genetic diversity.

Keywords: principal component analysis; evolutionary distance; genetic variation; variance–covariance matrix; multidimensional scaling; kernel trick

1. Introduction

Understanding the structure of human populations is important for human genetic studies. Principal components analysis (PCA) was first applied to genetic data by Cavalli-Sforza and colleagues¹. PCAs were widely used to reveal the present patterns of human genetic geography and revealed several hidden patterns of genetic variation in human populations¹⁻⁷. Previous studies, such as^{4,7,8}, showed that geographical maps naturally reflect a two-dimensional summary of genetic variation. For example, Novembre and his colleagues⁴ showed a close correspondence between genetic and geographic distances in Europe. Moreover, McVean⁹ showed that the projection of samples onto principal components can be directly inferred by considering the average coalescent times between pairs of haploid genomes.

The principal components are eigenvectors of the data's variance–covariance matrix, which is obtained by a genetic relationship matrix (GRM). However, it is difficult to apply a GRM to multiallelic sites. The aim of this study was to develop a method to facilitate PCA using nucleotide differences. Therefore, I investigated the relationship between PCA and nucleotide differences.

2. Materials and Methods

I first briefly outline the theoretical formulation of PCA based on McVean's notation⁹. I assumed that N haploid individuals were sequenced. Let us define a new matrix $\mathbf{X}$ with row sums equal to 0.

2.2. Variance Covariance Matrix and Genetic Relationship Matrix

The principal components are eigenvectors of the data's covariance matrix. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. Let us define a matrix $\mathbf{M}_s$ at locus s . The ij -th element of $\mathbf{M}_s$ is defined using the following equation:

$$\mathbf{M}_{sij} = X_{si}X_{sj}, \quad (1)$$

The second step of PCA is to obtain eigenvectors of the matrix $\mathbf{M}$. Let us denote the i th eigenvalue and eigenvector of the matrix $\mathbf{M}$ by λ_i and u_i .

$$\lambda_i u_i = \mathbf{M} u_i \quad (2)$$

2.2. Genetic Relationship Matrix

In the following section, I assumed that all polymorphic sites were biallelic; that is, they have two alleles that are the result of a single mutation. Let Z_{si} be the allelic state for individual i at locus k , coding the ancestral allele as 0 and the derived allele as 1. However, the following also applies for any coding, for example the minor allele coded as 1, as mentioned by McVean⁹. $\mathbf{Z}$ consists of an $L \times N$ binary matrix where L is the number of SNPs.

The vector $\mathbf{z}_s$ and $\mathbf{1}$ are defined by:

$$\mathbf{z}_s = (Z_{s1} \ Z_{s2} \ \cdots \ Z_{sN})^T,$$

$$\mathbf{1} = (1 \ 1 \ \cdots \ 1)^T$$

The GRM is obtained as $\mathbf{z}_s \mathbf{z}_s^T$ using the genotype valued decomposition¹⁰. The GRM

is used in the software GCTA¹¹ and in PCA³.

The variance–covariance matrix is also obtained by $\mathbf{1}$ and $\mathbf{z}_s$.

$$m_s = \mathbf{1}^T \mathbf{z}_s = \mathbf{z}_s^T \mathbf{1}, \quad (3)$$

where

$$X_{si} = Z_{si} - m_s/N. \quad (4)$$

$\mathbf{M}_s$ can be obtained by the following equation:

$$\mathbf{M}_s = \mathbf{z}_s \mathbf{z}_s^T - \frac{m_s}{N} \mathbf{1} \mathbf{z}_s^T - \frac{m_s}{N} \mathbf{z}_s \mathbf{1}^T + \frac{m_s^2}{N^2} \mathbf{1} \mathbf{1}^T. \quad (5)$$

The variance–covariance matrix of the whole genome is obtained as follows:

$$\mathbf{M} = \sum_{s=1}^L \mathbf{M}_s, \quad (6)$$

These equations are essentially the same as equations (8)–(10) proposed by McVean⁹.

2.3. Evolutionary Distance

I defined a new variable, d_{sij} , as 1 when the allelic states of i and j at locus s differed; otherwise, d_{sij} was defined as 0. Because Z_{si} and Z_{sj} are either 0 or 1, when they both equal 1, then $Z_{si}Z_{sj} = 1$; otherwise, $Z_{si}Z_{sj} = 0$.

D is the number of different sites between the haploid samples i and j . Let us

assume the additivity of evolutionary distances. I obtained the evolutionary distance as follows:

$$\mathbf{D} = \sum_{s=1}^L \mathbf{D}_s, \quad (7)$$

Note the equivalent formula for this defined variable:

$$d_{sij} = -2Z_{si}Z_{sj} + Z_{si} + Z_{sj}. \quad (8)$$

$$\mathbf{D}_s = -2\mathbf{z}_s\mathbf{z}_s^T + \mathbf{1}\mathbf{z}_s^T + \mathbf{z}_s\mathbf{1}^T. \quad (9)$$

2.3. Number of Multiallelic Sites

Number of multiallelic sites were counted from the genetic variation data of 2,504 people determined by 1000G which were downloaded from the International 1000 Genomes Project ¹².

3. Results

3.1 Number of Multiallelic Sites

In this paper, I used the double centering of multidimensional scaling (MDS) ¹³, because the sum of each column is required to be 0 to perform an effective PCA. The following equation was used to perform the MDS ¹⁴:

$$\mathbf{D}_s' = \mathbf{D}_s - \frac{1}{N} \mathbf{11}^T \mathbf{D}_s - \frac{1}{N} \mathbf{D}_s \mathbf{11}^T + \frac{1}{N^2} \mathbf{11}^T \mathbf{D}_s \mathbf{11}^T \quad (10)$$

Using equation (5), $\mathbf{D}'$ is obtained as:

$$\mathbf{D}_s' = -2\mathbf{z}_s \mathbf{z}_s^T + 2 \frac{m_s}{N} \mathbf{1z}_s^T + 2 \frac{m_s}{N} \mathbf{z}_s \mathbf{1}^T - 2 \frac{m_s^2}{N^2} \mathbf{11}^T, \quad (11)$$

(for details, see Appendix A).

$$\mathbf{D}' = \sum_{s=1}^L \mathbf{D}_s, \quad (12)$$

From equations (5), (6), (11), and (12), we obtain

$$\mathbf{D}' = -2\mathbf{M}. \quad (13)$$

Equation (13) shows a PCA from GRM is equivalent to MDS from nucleotide differences.

3.2. Number of Multiallelic Sites

Among genomic sequences of 2,504 people determined by 1000G, 81 million sites were biallelic and 274,620 sites were multiallelic (Table 1).

4. Discussion

I showed that principal components can be directly obtained from the evolutionary distances of haploid genomes.

When GRM among individuals is used for PCA, the GRM is obtained by

$$\mathbf{M} = \sum_{s=1}^L w_s \mathbf{M}_s. \quad (14)$$

where w_s is weight of locus s . Weights are usually included to normalize the data in each column to have the same variance^{2,3}. McVean showed that normalization has little effect upon SNP data⁹.

A close correspondence between genetic and geographic distances indicates that the geographical distance between two persons is a linear function of the time to their common ancestor. The relationships among migration, geographical isolation, and admixture have also been studied on the basis of a modest number of discrete demes^{7,15,16}. McVean⁹ showed that the projection of samples onto the principal components can be obtained directly by considering the average coalescent times between pairs of haploid genomes. His result provided a direct route to understanding the influence that various demographic scenarios can have on the relationships between samples identified from PCA.

Moreover, Charlesworth et al¹⁷ reviewed some of the theory and provided expected coalescence times, which provided new insight into continuous human movement.

Continuous isolation-by-distance models have been studied to obtain the probability of identity-by-descent¹⁸⁻²⁰ and to study the relationship between gene frequencies and geographic distances^{21,22}. Previous theoretical studies showed that coalescent times vary with deme location^{23,24}. Further studies must be necessary. However, it is worth noting that the evolutionary distance between two randomly chosen genes increases almost linearly with coalescent time, and coalescent time refers to the time to most recent common ancestor.

I assumed polymorphic sites were biallelic. However, in the datasets with large sample sizes, many variant sites were observed to be multiallelic. It is difficult to apply GRM to multiallelic sites, but evolutionary distances can easily be calculated using multiallelic sites. PCA using evolutionary distances will be widely applicable for studying genome variation.

In this study, I assumed additivity of evolutionary distance. However, there are many causes of nonlinear evolutionary distances²⁵. When evolutionary distance is not linear, PCA needs to be conducted using a nonlinear dimension reduction approach²⁶.

Author Contributions: K. M. confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Acknowledgments: I thank Prof. Fujio Tajima and Prof. Naomichi Matsumoto for their advice and encouragement. I thank Mallory Eckstut, PhD, from Edanz (<https://jp.edanz.com/ac>) for editing a draft of this manuscript. This work was supported by JSPS KAKENHI Grant Numbers JP17K08682, JP19K22647, JP20K07316. This work was a part of the branding program as a world-leading research university on intractable immune and allergic diseases supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflicts of Interest: The author declares no competing interests.

Appendix A

From equation (5), it is easy to obtain equations (A1), (A2), and (A3):

$$-\frac{1}{N} \mathbf{11}^T \mathbf{D}_s = \frac{2m_s}{N} \mathbf{1z}_s^T - \mathbf{1z}_s^T - \frac{m_s}{N} \mathbf{11}^T, \quad (\text{A1})$$

$$-\frac{1}{N} \mathbf{D}_s \mathbf{11}^T = \frac{2m_s}{N} \mathbf{z}_s \mathbf{1}^T - \frac{m_s}{N} \mathbf{11}^T - \mathbf{z}_s \mathbf{1}^T, \quad (\text{A2})$$

$$\text{and } \frac{1}{N^2} \mathbf{11}^T \mathbf{D} \mathbf{11}^T = -\frac{2m_s^2}{N^2} \mathbf{11}^T + \frac{m_s}{N} \mathbf{11}^T + \frac{m_s}{N} \mathbf{11}^T. \quad (\text{A3})$$

By adding equations (9), (A1), (A2), and (A3), I obtained equation (11).

Table 1. Number of SNPs in the 1000G dataset.

Type	Number of SNPs
Biallelic	81,106,364
Triallelic	272,987
Tetraallelic	1,633
Multiallelic sum	274,620

References

1. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786-92 (1978).
2. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
3. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
4. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).
5. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* **40**, 646-9 (2008).
6. Yamaguchi-Kabata, Y. *et al.* Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet* **83**, 445-56 (2008).
7. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet* **21**, 1277-85 (2013).

8. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309-14 (2015).
9. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet* **5**, e1000686 (2009).
10. Misawa, K. Genotype Value Decomposition: Simple Methods for the Computation of Kernel Statistics. *Adv Genet* (**Submitted**).
11. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
12. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
13. Tzeng, J., Lu, H.H. & Li, W.H. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics* **9**, 179 (2008).
14. Bishop, C.M. Pattern Recognition and Machine Learning. in *Pattern Recognition and Machine Learning* (2007).
15. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44**, 243-6 (2012).
16. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* **48**, 94-100 (2016).
17. Brian Charlesworth, Deborah Charlesworth, a. & Barton, N.H. The Effects of Genetic and Geographic Structure on Neutral Variation. *Annual Review of Ecology, Evolution, and Systematics* **34**, 99-125 (2003).
18. Barton, N.H., Depaulis, F. & Etheridge, A.M. Neutral evolution in spatially continuous populations. *Theor Popul Biol* **61**, 31-48 (2002).
19. Novembre, J. & Slatkin, M. Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution* **63**, 2914-25 (2009).
20. Joseph, T.A., Hickerson, M.J. & Alvarado-Serrano, D.F. Demographic inference under a spatially continuous coalescent model. *Heredity (Edinb)* **117**, 94-9 (2016).
21. Hardy, O.J. & Vekemans, X. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity (Edinb)* **83** (Pt 2), 145-54 (1999).
22. Furstenau, T.N. & Cartwright, R.A. The effect of the dispersal kernel on isolation-by-distance in a continuous population. *PeerJ* **4**, e1848 (2016).
23. Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597-601 (1989).
24. Wilkins, J.F. & Wakeley, J. The coalescent in a continuous, finite, linear

- population. *Genetics* **161**, 873-88 (2002).
25. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*, (Oxford University Press, New York, 2000).
 26. Zhang, M., Liu, Y., Zhou, H., Watkins, J. & Zhou, J. A novel nonlinear dimension reduction approach to infer population structure for low-coverage sequencing data. *BMC Bioinformatics* **22**, 348 (2021).