Theoretical Population Biology 107 (2016) 65–76

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Eigenanalysis of SNP data with an identity by descent interpretation

Xiuwen Zheng, Bruce S. Weir *

Department of Biostatistics, University of Washington, Box 359461, Seattle, WA 98195-9461, USA

ARTICLE INFO

Article history:
Received 10 February 2015
Available online 23 October 2015

Keywords:
PCA
Relatedness
Coancestry
IBD
SNP
Admixture

ABSTRACT

Principal component analysis (PCA) is widely used in genome-wide association studies (GWAS), and the principal component axes often represent perpendicular gradients in geographic space. The explanation of PCA results is of major interest for geneticists to understand fundamental demographic parameters. Here, we provide an interpretation of PCA based on relatedness measures, which are described by the probability that sets of genes are identical-by-descent (IBD). An approximately linear transformation between ancestral proportions (AP) of individuals with multiple ancestries and their projections onto the principal components is found.

In addition, a new method of eigenanalysis "EIGMIX" is proposed to estimate individual ancestries. EIGMIX is a method of moments with computational efficiency suitable for millions of SNP data, and it is not subject to the assumption of linkage equilibrium. With the assumptions of multiple ancestries and their surrogate ancestral samples, EIGMIX is able to infer ancestral proportions (APs) of individuals. The methods were applied to the SNP data from the HapMap Phase 3 project and the Human Genome Diversity Panel. The APs of individuals inferred by EIGMIX are consistent with the findings of the program ADMIXTURE.

In conclusion, EIGMIX can be used to detect population structure and estimate genome-wide ancestral proportions with a relatively high accuracy.

© 2015 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Principal component analysis was introduced for the study of genetic data almost thirty years ago by Menozzi et al. (1978), and has since become a standard tool. Population differentiation can be inferred from multivariate statistical methods such as PCA of allele frequencies (Menozzi et al., 1978; Cavalli-Sforza and Feldman, 2003). In a new approach, Patterson et al. (2006) applied PCA to SNP genotypic data for individuals rather than populations. Their method, implemented in a software package "EIGENSTRAT", has been widely used to correct for population stratification in genome-wide association studies (GWAS) (Price et al., 2010). Although PCA is not based on a population genetics model, and may seem like a "black box" method, principal component axes often represent perpendicular gradients in geographic space (Cavalli-Sforza and Feldman, 2003; Price et al., 2006; Novembre et al., 2008). The relationship of PCA results to fundamental demographic parameters is of major interest to geneticists.

Novembre and Stephens (2008) showed that the gradient and wave patterns of principal components do not necessarily reflect migration events in history. From the perspective of coalescent theory, McVean (2009) provided a genealogical interpretation of PCA. He showed that the projection of samples onto the principal components could be obtained from the pairwise coalescence times between study individuals. Ma and Amos (2010) proposed a formulation of PCA based on the variance–covariance matrix of the sample allele frequencies.

We now provide an alternative interpretation of PCA based on relatedness measures: probabilities that sets of genes have descended from a single ancestral gene and so are identical by descent (ibd). The ibd concept is essential for genetic analyses such as linkage studies for mapping disease genes and forensic DNA profiling (Weir et al., 2006; Thompson, 2013). In population genetics, Weir and Hill (2002) extended the work of Weir and Cockerham (1984) by allowing different levels of coancestry for different populations, and by allowing non-zero coancestries between pairs of populations. Our further extension is to allow different coancestries between pairs of individuals and different inbreeding coefficients for individuals. The coancestry coefficient between two populations defined in the model of Weir and Hill is now replaced by the average kinship coefficient among pairs of study individuals from these two populations respectively, relative to a single ancestral population, so that the assumption of random

* Corresponding author.
E-mail addresses: zhengx@uw.edu (X. Zheng), bsweir@uw.edu (B.S. Weir).

http://dx.doi.org/10.1016/j.tpb.2015.09.004
0040-5809/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

rating can be relaxed. These individual-perspective measures of population structure can be used to explain the behavior of PCA.

Ancestral proportions (AP) of an individual refer to the fractions of the genome derived from specific ancestral populations (Pritchard et al., 2000; Falush et al., 2003; Tang et al., 2005; Alexander et al., 2009). The early approach for estimating AP can track back to Hanis et al. (1986), and the ancestral allele frequencies should be known to allow estimating allele admixture in this method. However, ancestral allele frequencies are usually estimated from surrogate ancestral samples in practice and later studies took into account in describing the uncertainty of estimated ancestral information.

A Bayesian approach, STRUCTURE, was developed to infer population substructure using unlinked genotypes (Pritchard et al., 2000). Later, it was extended to model linked markers (Falush et al., 2003) through admixture linkage disequilibrium (LD). STRUCTURE is computationally intensive and not likely to be suitable for large-scale studies, like GWAS, involved with thousands of individuals and hundreds of thousands of SNPs. SNP pruning has to be done before applying STRUCTURE, and this can introduce selection bias with respect to different SNP sets. A maximum-likelihood estimation method, frappe, has also been proposed to estimate AP with much less computation than STRUCTURE, but it assumes the markers are unlinked (Tang et al., 2005). The ADMIXTURE method was developed to analyze thousands of markers — it adopts the likelihood model embedded in STRUCTURE with an assumption of linkage equilibrium among the markers (Alexander et al., 2009).

Instead of estimating global ancestry via genome-wide markers, detection of local ancestry from chromosomal segments in admixed populations becomes of great interest. Recently, HAPMIX and MULTIMIX were proposed to infer local ancestry from dense SNP markers based on approximate coalescent models modeling linkage disequilibrium with two or more ancestries (Price et al., 2009; Churchhouse and Marchini, 2013). However, their methods require a fine genetic map.

The potential connection between ancestral proportions and principal components in the eigenanalysis has been investigated by the previous studies with a limited number of numerical simulations (Patterson et al., 2006; Engelhardt and Stephens, 2010). McVean (2009) indicated it is possible to identify relative admixture proportions from principal components. Ma and Amos (2012) showed how to estimate two-way admixture proportions with a proof under their framework of variance–covariance matrix. They also observed that an admixed population could divide the triangle of three parental populations in the PC plot into three small triangles with areas according to the three-way admixture proportions. However, none of these studies provided a sufficient proof for inferring admixture fractions from the principal components under their theoretical framework in the cases of more than two ancestral populations.

In our study, an approximately linear transformation between ancestral proportions (AP) of individuals with multiple ancestries and their projections onto the principal components is revealed, and a proof is given under the framework of identity by descent. This linear transformation could explain the perpendicular gradients in geographic space, and it also justifies the observation that the ratios of triangle areas correspond to admixture fractions in the study of Ma and Amos (2012). We also propose a new method of eigenanalysis "EIGMIX" to estimate individual ancestries. EIGMIX uses method of moments estimation with computational efficiency suitable for millions of SNP data, and it is not subject to the assumption of linkage equilibrium. Ancestral proportions can be estimated by making assumptions of surrogate samples for ancestral populations, but inferring ancestral allele frequencies is not necessary. The calculation uses all study individuals simultaneously without projecting the remaining individuals onto the existing axes of surrogates.

We applied various methods to the SNP data of 1198 founders from the HapMap Phase 3 project and 938 unrelated individuals from the Human Genome Diversity Project (HGDP). The ancestral proportions of individuals inferred by PCA and EIGMIX are consistent with the findings of the program ADMIXTURE. All eigenanalysis in the study are implemented in the R package "SNPRelate" (Zheng et al., 2012), allowing users to apply our method to their SNP data.

2. Methods

We develop our approach with a series of indicator variables x_{ijkl} for the kth allele, $k = 1, 2$, at the lth locus, $l = 1, 2, \ldots, L$, in the jth individual sampled from the ith population, $j = 1, 2, \ldots, n_i$; $i = 1, 2 \ldots, N$. The total sample size is $n = \sum_i n_i$. The variables take the value 1 for alleles of a specific type, e.g. the reference allele, at a locus, and the value 0 otherwise. Genotypes are indicated by $g_{ijl} = x_{ij1l} + x_{ij2l}$, and these take the values 0, 1, 2.

2.1. Population coancestry framework of Weir and Hill (2002)

Under the framework of Weir and Hill (2002), the expectations for first and second moments of the x's are

$$\begin{aligned}
\mathcal{E}[x_{ijkl}] &= p_l \\
\mathcal{E}[x^2_{ijkl}] &= p_l \\
\mathcal{E}[x_{ijkl}\, x_{ijk'l}] &= p_l^2 + p_l(1-p_l)F_{ij}, \quad k \neq k', \text{ the same individual} \\
\mathcal{E}[x_{ijkl}\, x_{ij'k'l}] &= p_l^2 + p_l(1-p_l)\theta_i, \quad j \neq j', \text{ the same population} \\
\mathcal{E}[x_{ijkl}\, x_{i'j'k'l}] &= p_l^2 + p_l(1-p_l)\theta_{ii'}, \quad i \neq i', \text{ different populations.}
\end{aligned}$$

Here expectation is over both repeated samples from the population and over evolutionary replicates of the populations. These expressions introduce the total inbreeding coefficient F_{ij}, the within-population coancestries θ_i, and the between-population-pair coancestries $\theta_{ii'}$. The quantities p_l are the overall, or ancestral, frequencies of the reference alleles if all study individuals can be traced back to a single reference population. This reference population could be common ancestors at a point in time of the past. The equal values for $\mathcal{E}[x_{ij1l}\, x_{ij2l}]$ and $\mathcal{E}[x_{ijkl}\, x_{ij'k'l}]$ require an assumption of random mating.

The coancestry coefficient θ_i refers to the ibd probability for a random pair of alleles in population i, and the pair of alleles can come from the same individual. The coancestry coefficient $\theta_{ii'}$ refers to the ibd probability for a random pair of alleles, one from population i and the other from population i'. Note that we implicitly assume θ_i and $\theta_{ii'}$ are the same at each locus, and in practice θ_i and $\theta_{ii'}$ are actually the average inbreeding and coancestry coefficients over all L loci.

Now consider an individual perspective measures of population structure, i.e., a special case of Weir and Hill's model where each population i has only one sampled individual ($n_i = 1$) so $j = 1$ for each population. The assumption of random mating is relaxed, and the sample size n is also the number of populations r. Therefore,

$$\begin{aligned}
&\bar{p}_l = \frac{1}{n}\sum_{i=1}^{n}\bar{p}_{il} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\sum_{j=1}^{1}(x_{ij1l} + x_{ij2l})\right] \\
&\mathcal{E}[\bar{p}_l] = p_l \\
&\mathrm{Var}[\bar{p}_{il}] = \frac{1}{2}p_l(1-p_l)(1+\theta_i) \\
&\mathrm{Cov}[\bar{p}_{il}, \bar{p}_{i'l}] = p_l(1-p_l)\theta_{ii'} \\
&\mathrm{Var}[\bar{p}_l] = \frac{n-1}{n}p_l(1-p_l)\theta_T + \frac{1}{2n}p_l(1-p_l)(1+\theta_l) \\
&\mathcal{E}[\bar{p}_l(1-\bar{p}_l)] = \frac{n-1}{n}p_l(1-p_l)(1-\theta_T) + \frac{1}{2n}p_l(1-p_l)(1-\theta_l)
\end{aligned} \tag{1}$$

where $\theta_l = \sum_{i=1}^{n}\theta_i/n$, the average inbreeding coefficient among all study individuals, and $\theta_T = \sum_{i,i'=1,\ i\neq i'}^{n}\theta_{ii'}/[n(n-1)]$,

Download English Version:

https://daneshyari.com/en/article/6372320

Download Persian Version:

https://daneshyari.com/article/6372320

Daneshyari.com