Principal component analysis of genetic data

David Reich, Alkes L Price & Nick Patterson

Principal component analysis (PCA) has been a useful tool for analysis of genetic data, particularly in studies of human migration. A new study finds evidence that the observed geographic gradients, traditionally thought to represent major historical migrations, may in fact have other interpretations.

Principal component analysis (PCA) has been used for several decades to study human population migrations, resulting in remarkable inferences about history. On page 646 of this issue, John Novembre and Matthew Stephens[1] show that the geographic gradients that emerge when PCA is applied to genetic data—and that are sometimes interpreted as highly suggestive of major historical migrations—can also have other explanations. We suggest guidelines for scientists interested in using PCA in genetic analysis in light of this potential concern and highlight three applications in which PCA has continued value: detecting population substructure, correcting for stratification in disease studies and making qualified inferences about human history.

Synthetic maps in question

PCA is a statistical method for exploring and making sense of datasets with a large number of measurements (which can be thought of as dimensions) by reducing the dimensions to the few principal components (PCs) that explain the main patterns. Thus, the first PC is the mathematical combination of measurements that accounts for the largest amount of variability in the data. Luca Cavalli-Sforza and colleagues had the original insight that PCA could be applied to human genetic variation[2], and they eventually analyzed about 100 protein polymorphisms that had been measured in many human populations[3]. By superimposing the PCs on the geography of the sampled populations, they obtained "synthetic maps" that showed remarkable gradients of variation across continents suggestive of historical migrations[2]. For example, the first European PC map shows a southeast-to-northwest cline that was interpreted as reflecting the spread of Neolithic farming from the Levant throughout Europe between 9,000 and 6,000 years ago. The hypothesis of a demic diffusion of Neolithic farming has since been supported by additional genetic and archaeological data[4–6] (but see ref. 7 for a dissenting view).

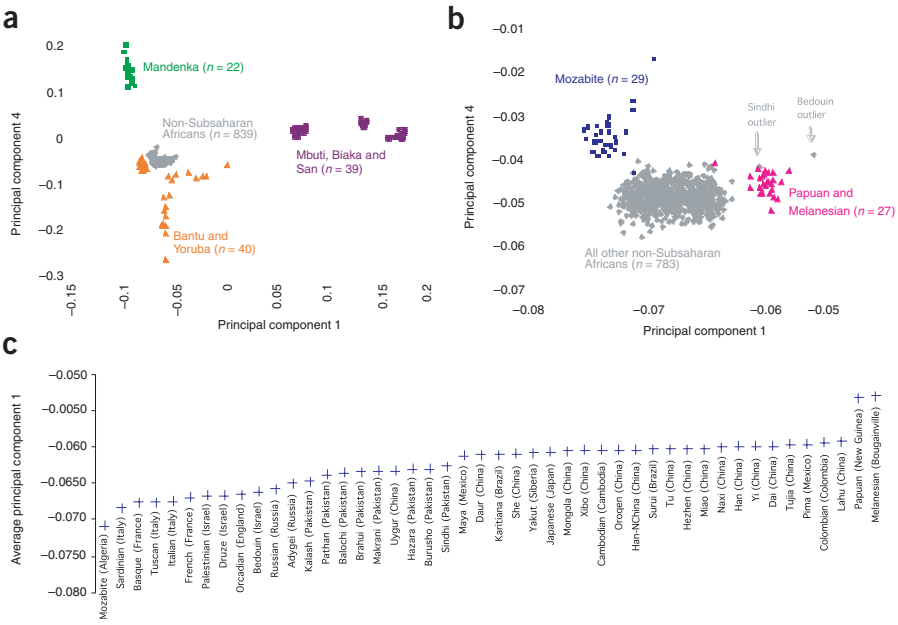

Figure 1 PCA continues to provide evidence of important migration events. (**a**) We carried out PCA on 940 individuals from the Human Genome Diversity Project that were scanned at approximately 650,000 SNPs[11] using data from 101 sub-Saharan African samples to define the PCs (Mandenka, Bantu from Kenya and South Africa, Yoruba, San, Mbuti Pygmy and Biaka Pygmy). We carried out the analysis on samples blinded to population labels (the coloring of samples was only carried out after the analysis). We plotted principal component 1 (negative values are more Bantu-related) and principal component 4 (positive values are more closely related to the Senegalese Mandenka). (**b**) Outlying populations are the Mozabite, who are more Mandenka-related, reflecting recent gene flow across the Sahara, and Papuans and Melanesians, who have inherited less Bantu-related gene flow. (**c**) To reveal the west-to-east gradient of Bantu-related ancestry across Eurasia, we averaged the first PC for each of the non-African populations and plotted the populations in rank order.

John Novembre and Matthew Stephens now show that PCs correlating with geography do not necessarily reflect major directed migrations but may instead simply reflect 'isolation by distance,' whereby there is only gene exchange among neighboring populations (thus, proximity is the determinant of how closely populations are related[1]). In computer simulations and in real data from a bird species, they show that even in the absence of major migrations, geographic gradients of PCs can emerge that look qualitatively similar to the synthetic maps. To interpret demographic history, one should consider PCs jointly, noting that some of the components correspond to real migration events, whereas others are artifacts that arise from isolation by distance.

What does this mean for interpretation of synthetic maps? Cavalli-Sforza and colleagues have emphasized the importance of combining mathematical genetics with other lines of evidence[2] before being convinced of any result. Given the strong correlation of genetic and nongenetic evidence, at least some of the migrations that they identified from the data (such as the Neolithic farming migration) are likely to be real. However, even aside from the issues raised by this new study, interpreting synthetic maps has been difficult, requiring correlation of genetic information with often incomplete data from archaeology and

David Reich and Alkes L. Price are at the Department of Genetics, Harvard Medical School, New Research Building, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. Nick Patterson is at the Broad Institute of Harvard and the Massachusetts Institute of Technology, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.
e-mail: reich@genetics.med.harvard.edu

linguistics. In light of the fact that a proportion of the PCs reflect isolation by distance, it seems even more likely that some synthetic maps have been overinterpreted.

Where does this leave PCA as a tool for analyzing genetic data? As pointed out by Novembre and Stephens[1], PCA remains useful for genetic analysis in many contexts that do not require a historical interpretation, such as in detecting the presence of population structure or in correcting for stratification in disease studies. On the other hand, if the aim is to study history and document migrations, it is important to carry out additional research to correlate the PCA results with other lines of evidence.

Population structure and stratification

PCA has a population genetics interpretation and can be used to identify differences in ancestry among populations and samples, regardless of the historical patterns underlying the structure. In particular, by assessing whether the proportion of the variance explained by the first PC is sufficiently large, it is possible to obtain a formal *P* value for the presence of population substructure and to identify the number of PCs that are statistically significant[8]. PCA is also useful as a method to address the problem of population stratification—allele frequency differences between cases and controls due to ancestry differences—that can cause spurious associations in disease association studies. We and others have described how one can correct for stratification in structured populations such as European Americans by adjusting genotypes and phenotypes by amounts attributable to ancestry along the top PCs[9,10]. Novembre and Stephens[1] emphasize that this approach is appropriate regardless of whether the PCs have arisen as a result of migrations, isolation by distance or both.

Understanding human history

Given the results of Novembre and Stephens[1], what confidence should we have in use of PCA for inferences regarding human history? To illustrate, we turned to a dataset of 940 individuals from 53 populations typed at ~650,000 SNPs as part of the Human Genome Diversity Project[11]. We used EIGENSOFT[8,9] to find the principal axes of genetic variation in the seven sub-Saharan African populations in this dataset and then projected all samples on the resulting PCs. The non-African populations fell into a rough cluster (**Fig. 1a**), which is about what would be expected if all non-African populations were founded by a single dispersal 'out of Africa'[12].

Inspecting the non-African cluster more closely, however, we found three outlier populations that have distinct relationships with sub-Saharan Africans: the Mozabite, a North African population that is well known to have received recent gene flow across the Sahara, the Papuans and the Melanesians (**Fig. 1b**). A higher-resolution analysis (**Fig. 1c**) reveals a distinct gradient of Bantu-related ancestry from west to east across Eurasia, an observation that sharply contradicts the theory that a single African migration gave rise to the entire non-African gene pool. One explanation for this is that after the initial southern-route migration out of Africa[12], there was later Bantu-related gene flow into Europe and the rest of Eurasia. Because of their geographic isolation, Papuans and Melanesians may have received a reduced contribution of this second round of gene flow, which could have arrived either via a major migration or via gradual isolation by distance[13]. This example highlights how PCA methods can provide evidence of important migration events. Interpreting the results to make reliable historical predictions, however, requires further genetic analysis and integration with other sources of information from archeology, anthropology, linguistics and geography[2].

1. Novembre, J. & Stephens, M. *Nat. Genet.* **40**, 646–649 (2008).
2. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. *Science* **201**, 786–792 (1978).
3. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes*. (Princeton University Press, Princeton, New Jersey, 1994).
4. Sokal, R.R., Oden, N.L. & Wilson, C. *Nature* **351**, 143–145 (1991).
5. Semino, O. *et al. Am. J. Hum. Genet.* **74**, 1023–1034 (2004).
6. Pinhasi, R. *et al. PLoS Biol.* **3**, e410 (2005).
7. Haak, W. *et al. Science* **310**, 1016–1018 (2005).
8. Patterson, N., Price, A. & Reich, *D. PLoS Genet.* **2**, e190 (2006).
9. Price, A.L. *et al. Nat. Genet.* **38**, 904–909 (2006).
10. Zhu, X. & Li, S. *Am. J. Hum. Genet.* **82**, 352–365 (2008).
11. Li, J.Z. *et al. Science* **319**, 1100–1104 (2008).
12. Mellars, P. *Science* **313**, 796–800 (2006).
13. Handley, L.J.L., Manica, A., Goudet, J. & Balloux, F. *Trends Genet.* **23**, 432–439 (2007).

From gene expression to disease risk

Emmanouil T Dermitzakis

Gene expression can be an indicator of cellular state, and studies characterizing variation in gene expression have been useful on the cellular level. Two new studies now provide the first direct demonstration of the successful use of the multidimensionality of gene expression to dissect the genetic architecture of complex diseases.

The genetics of variation in gene expression, or genetical genomics, has attracted significant interest in the last decade, with many studies characterizing its genetic architecture[1–5]. In addition, several studies have demonstrated the potential causal impact of differential gene expression on complex disease risk[6,7]. Two new studies[8,9] now take the field a step further toward understanding the correlation between gene expression and specific disease phenotypes by combining gene expression and clinical information or disease traits in large human population samples and segregating mouse populations, respectively.

Emmanouil T. Dermitzakis is at the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, CB10 1SA Cambridge, UK.
e-mail: md4@sanger.ac.uk

Genetics of gene expression

Gene expression can be considered as a quantitative trait that is highly heritable. Genetic variation that explains variance of gene expression is usually found in the proximal genomic region of the gene whose expression is being measured. This proximity can be mainly explained by variants in *cis* regulatory elements (promoters, enhancers, etc.), many of which segregate in natural populations at high frequencies. Some of the variance can also be explained by variability in upstream regulatory pathways and networks, and these signals are traditionally called *trans* effects.

Heritability of gene expression variation is generally high, as a result of the small number of molecular interactions between the