Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Statistical methods for detecting admixture

Pongsakorn Wangkumhang and Garrett Hellenthal

The increasing availability of large-scale autosomal genetic variation data sampled from world-wide geographic areas, coupled with advances in the statistical methodology to analyse these data, is showcasing the power of DNA as a major tool to gain insights into the demographic history of humans and other organisms. Here we review statistical techniques that shed light on a specific aspect of demography: the detection and description of admixture events where two or more genetically distinct groups intermixed at one or more times in the past. In particular we give an overview of some of the widely used methods to identify and describe admixture events using autosomal DNA from unrelated individuals, with a particular focus on analysing biallelic Single-Nucleotide-Polymorphsim (SNP) markers.

Address
University College London Genetics Institute (UGI), Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

Corresponding author: Hellenthal, Garrett (g.hellenthal@ucl.ac.uk)

Current Opinion in Genetics & Development 2018, **53**:121–127

This review comes from a themed issue on **Genetics of human origins**

Edited by **Lluis Quintana-Murci** and **Brenna Henn**

https://doi.org/10.1016/j.gde.2018.08.002

While Y-chromosome and mitochondrial (mtDNA) are extremely valuable in studying sex-biased admixture, where one of the admixing groups contributes a disproportionate number of males or females [1], current widely used approaches to infer admixture focus on analysing autosomal DNA as it contains many thousands of times more (sex-averaged) information than Y/mtDNA. As examples, we apply some of these approaches to two sets of simulated data from [2••] containing 474 491 autosomal SNPs (Figure 1). Each set consists of 20 simulated individuals, descending from a single admixture event between two sources occurring 30 generations ago. These sources are African and European in the first set, contributing ≈80% and ≈20% of the DNA, respectively, while the sources are Central-South-Asian and European in the second set, each contributing ≈50% of the DNA. Following the procedure in [3•], these simulations used DNA from 21 present-day Yoruban individuals from Nigeria, 21 present-day Brahui individuals from Pakistan and 28 present-day individuals from France as the African, Central-South-Asian and European admixing sources, respectively. In analyses below, we also include surrogate individuals representing each of these continental admixing sources, in particular using the genomes of 22 Mandenka from Africa, 21 Balochi from Central South Asia, and 23 British/Irish from Europe.

Signatures of admixture

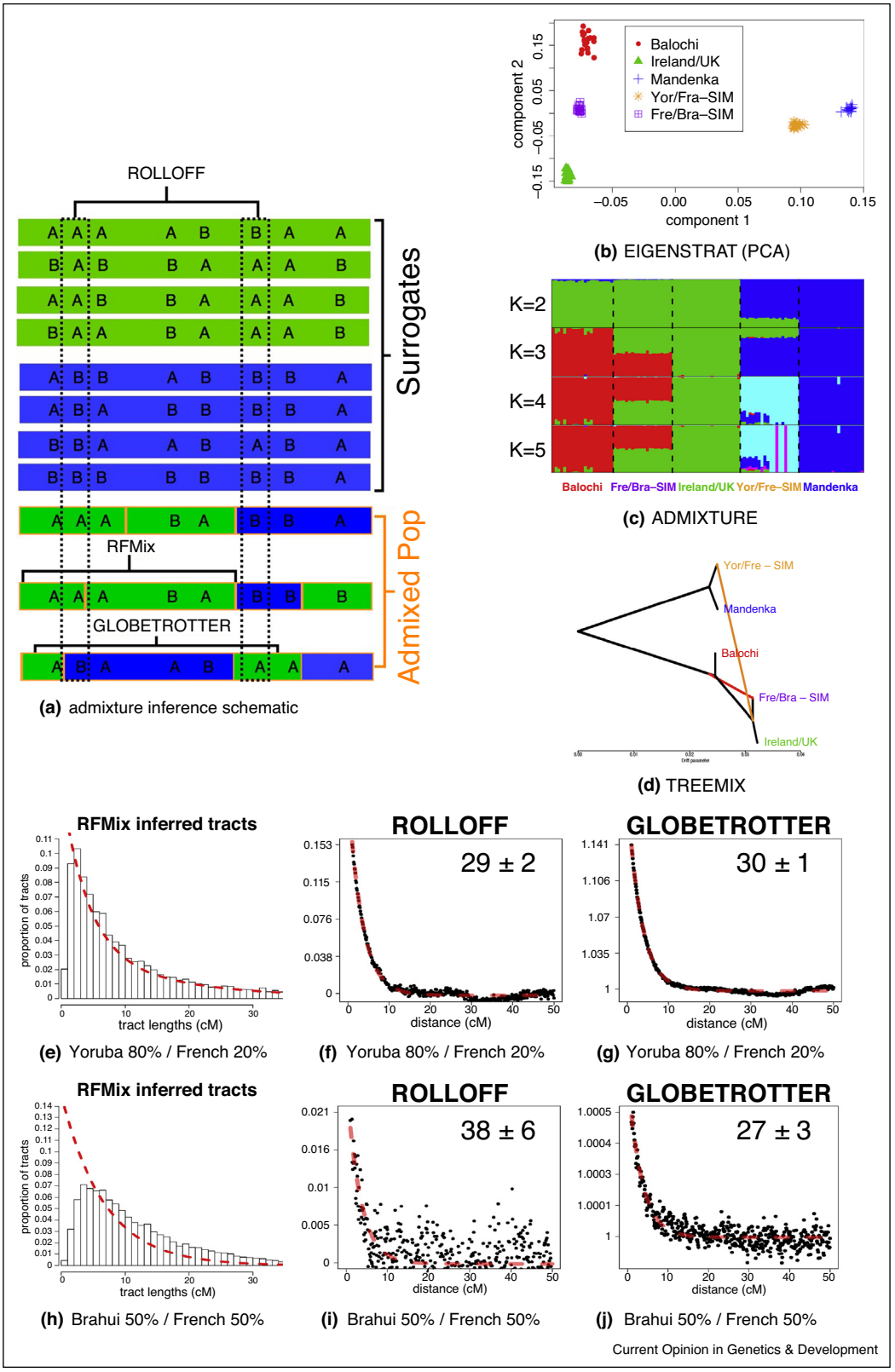
Two of the most widely used techniques for describing population structure and demography are principal components analysis (PCA) and clustering algorithms. While the genetic patterns captured by each approach may be attributable to factors other than admixture, they can each highlight individuals who descend from multiple intermixing source groups.

PCA and spatial techniques

PCA is an algebraic technique that applies eigenvector decomposition to reduce high-dimensional data to a small number of independent variables, termed principal components, that explain a large proportion of variation in the complete data. Typically PCA is applied to the genome-wide genotype data of multiple individuals, and the top few principal components are plotted to visualize genetic distance among the individuals [4,5••,6–9]. For example, Figure 1b plots the first two principal components from the PCA program EIGENSTRAT [[4,5••,6–9]] applied to our simulated data and the surrogate individuals. Note that surrogate individuals from the three continental groups fall into different corners of the plot, while each set of admixed individuals falls between the two continental groups comprising their ancestry in a manner roughly consistent with admixture proportions. However, caution is strongly warranted against concluding admixture wherever such a pattern is observed, as PCA projections can depend on sample size, SNP ascertainment, and features of demography besides admixture [9,10].

To better interpret these patterns, recent approaches have utilized isolation-by-distance models that assume genetic similarity decreases with geographic distance between samples. For example, SpaceMix [11•] highlights populations whose allele frequencies are more correlated than expected when assuming genetic similarity decays exponentially with distance, while EEMS [12•] relates genetic similarity to the distance among individuals' geographic locations (e.g. birthplace or sampling information) assuming migration occurs between

Figure 1



Identifying admixture in simulated data from [2••]. **(a)** Populations represented by blue/green surrogate groups mix *30* generations in the past, generating admixed haplotypes at bottom. A and B denote allele types at each biallelic SNP; orange bars separate segments in admixed