Generalized Analysis of Molecular Variance

Caroline M. Nievergelt[1,3,4,5], Ondrej Libiger[1,3,4], Nicholas J. Schork[1,2,3,4,5,6*]

1 Department of Psychiatry, University of California at San Diego, La Jolla, California, United States of America, **2** Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, California, United States of America, **3** Rebecca and John Moores UCSD Cancer Center, University of California at San Diego, La Jolla, California, United States of America, **4** The Center for Human Genetics and Genomics, University of California at San Diego, La Jolla, California, United States of America, **5** The Stein Institute for Research on Aging, University of California at San Diego, La Jolla, California, United States of America, **6** Scripps Genomic Medicine and Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California, United States of America

Many studies in the fields of genetic epidemiology and applied population genetics are predicated on, or require, an assessment of the genetic background diversity of the individuals chosen for study. A number of strategies have been developed for assessing genetic background diversity. These strategies typically focus on genotype data collected on the individuals in the study, based on a panel of DNA markers. However, many of these strategies are either rooted in cluster analysis techniques, and hence suffer from problems inherent to the assignment of the biological and statistical meaning to resulting clusters, or have formulations that do not permit easy and intuitive extensions. We describe a very general approach to the problem of assessing genetic background diversity that extends the analysis of molecular variance (AMOVA) strategy introduced by Excoffier and colleagues some time ago. As in the original AMOVA strategy, the proposed approach, termed generalized AMOVA (GAMOVA), requires a genetic similarity matrix constructed from the allelic profiles of individuals under study and/or allele frequency summaries of the populations from which the individuals have been sampled. The proposed strategy can be used to either estimate the fraction of genetic variation explained by grouping factors such as country of origin, race, or ethnicity, or to quantify the strength of the relationship of the observed genetic background variation to quantitative measures collected on the subjects, such as blood pressure levels or anthropometric measures. Since the formulation of our test statistic is rooted in multivariate linear models, sets of variables can be related to genetic background in multiple regression-like contexts. GAMOVA can also be used to complement graphical representations of genetic diversity such as tree diagrams (dendrograms) or heatmaps. We examine features, advantages, and power of the proposed procedure and showcase its flexibility by using it to analyze a wide variety of published data sets, including data from the Human Genome Diversity Project, classical anthropometry data collected by Howells, and the International HapMap Project.

Citation: Nievergelt CM, Libiger O, Schork NJ (2007) Generalized analysis of molecular variance. PLoS Genet 3(4): e51. doi:10.1371/journal.pgen.0030051

Introduction

Genetic and genetic epidemiologic studies involving large numbers of individuals and/or populations are being pursued more and more often as a result of the development of high-throughput genotyping technologies and the creation of genotype data repositories such as the dbSNP (http://www.ncbi.nlm.nih.gov/SNP) and the International HapMap Project databases (http://www.hapmap.org). Many of these studies are concerned with the identification and characterization of the relationships of the populations and/or subsets of individuals in those populations on the basis of their genomic profiles or "genetic backgrounds" (i.e., whether or not these populations/individuals carry the same sets of genetic variations [1–8]). In addition, genetic epidemiologic studies are often conducted to identify relationships between specific sets of genetic variations possessed by individuals and phenotypic endpoints they might have, such as a disease. The collection of variations that an individual possesses that contribute, e.g., to his or her disease susceptibility, may vary from population to population (e.g., as defined geographically, ethnically, racially, or linguistically). This may be due to the underlying heterogeneity of disease pathogenesis, the origins of the variations both in terms of time and place, and the frequency with which those variations are transmitted across populations (e.g., via migration patterns, interpopulation matings, etc.). Thus, the genetic background of an individual—at least with respect to relevant disease-contributing variations—is as crucial in these types of investigations as it is in other types of population genetic studies. In addition, it has been shown that, due to phenomena such as varying degrees of admixture and/or cryptic relatedness in the study population, ignoring genetic background in epidemiologic studies testing associations between particular genetic variations and a phenotype can result in false positive and false negative results [9–19], which underscores the importance of genetic background analysis even in very simple genetic association studies.

Many innovative analytical methods have been developed recently to assess and accommodate genetic background heterogeneity [20–37]. The vast majority of these methods involve some form of cluster analysis, although some more recent methods do not (e.g., [29,32]). For example, hierarch-

Editor: David B. Allison, University of Alabama at Birmingham, United States of America

Received October 26, 2006; **Accepted** February 22, 2007; **Published** April 6, 2007

A previous version of this article appeared as an Early Online Release on February 22, 2007 (doi:10.1371/journal.pgen.0030051.eor).

Abbreviations: AIM, ancestry informative marker; AMOVA, analysis of molecular variance; ANOVA, analysis of variance; CEPH, Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain; FRS, nasion-bregma subtense; GAMOVA, generalized analysis of molecular variance; HGDP, CEPH Human Genome Diversity Program; IBS, identical by state; LR, Lynch-Ritland; SNPs, single nucleotide polymorphism

* To whom correspondence should be addressed. E-mail: nschork@ucsd.edu

Author Summary

Humans exhibit great genetic diversity. Understanding the factors that contribute to and sustain this diversity is an important research area. Not only can such understanding shed light on human origins, but it can also assist in the discovery of genes and genetic factors that contribute to debilitating diseases. Statistical analysis methods that can facilitate the identification of factors contributing to or associated with human genetic diversity are growing in number as new high-throughput molecular genetic assays and technologies are developed. We consider the use of an analysis method termed generalized analysis of molecular variance (GAMOVA), which builds off of previously proposed analysis methods for testing hypotheses about the factors associated with genetic background diversity. We apply the method in a wide variety of settings and show that it is both flexible and powerful. GAMOVA has great potential to assist in population-based human genetic studies, as it can be used to address questions such as: Is a sample of affected cases and unaffected controls from a homogeneous population, or is there evidence of heterogeneity that could affect the results of an association study? Is there reason to believe that the ancestry of a set of individuals influences the traits that they have?

ical clustering strategies can be used to assess genetic background clustering, and, like other cluster analysis methods, require the construction of a measure of the similarity or dissimilarity (genetic distance) between all pairs of the N individual genomes or population allele frequency profiles (e.g., between-group variation, F_{ST}) comprising a sample. The resulting $N \times N$ similarity or distance matrix is then explored statistically to identify clusters of individuals or populations that exhibit greater or lesser similarity. Problems inherent to this approach involve the choice of a similarity metric, deciding which cluster method is most appropriate (e.g., single linkage, complete linkage, etc.), the determination of the optimal number of clusters representing the data, and the biological meaning of the clusters.

With respect to the choice of a similarity metric for cluster analysis, the simplest marker-based method for the assessment of genetic similarity between two individuals is to calculate the fraction of alleles shared identical by state (IBS) by those individuals over all the loci for which the individuals have been genotyped. If N individuals have been genotyped, then all $N \times N$ pairs of individuals can be assessed in this way. In addition to providing a foundation for some cluster analysis methods, graphical displays of the similarity matrix can be produced that allow visual assessment of the potential that subgroups of individuals with similar genetic backgrounds exist in the data. This approach has been used widely, and is often referred to, when presented in graphical form as a dendrogram, or as an allele-sharing "tree of individuals" (e.g., [7,38–40]). One problem, however, with the simple IBS sharing measure of genetic background similarity is that it does not account for allele frequencies. Consider, for example, two individuals who share rare alleles. These individuals are more likely to have arisen from the same (unique) population in which those alleles arose. In this situation one may want to consider "weighting" allele sharing at each locus by the frequency of the shared (or unshared) alleles. Pairwise measures of genetic similarity that accommodate allele frequencies have been put forward and are used often in ecological and nonhuman population genetics analysis settings (e.g., [41–45]).

Cluster analysis approaches can be extended by making more explicit and rigorous assumptions about the ancestral populations from which the individuals in a sample arose. Thus, specific ancestry informative markers (AIMs), which show large frequency differences between ancestral populations, can be used to quantify the degree of admixture among individuals in a sample [18,46–49]. When an individual genotyped on such markers possesses variations that are more frequent in one of the chosen ancestral populations, then that individual's ancestral relationship to this population can be inferred. Obviously, one needs to have identified the appropriate AIMs in advance of such analyses and this requires assumptions about the ancestral populations contributing to the individual genetic backgrounds reflected in a sample.

In the following we describe a flexible alternative to cluster analysis–based methods for the statistical assessment of genetic background similarities among populations or individuals. The proposed method does not necessarily rely on AIMs, but does require genotype information on at least a few hundred (possibly less when including AIMs) genetic markers (null loci) such as microsatellites, single nucleotide polymorphisms (SNPs), and/or insertion–deletion polymorphisms. Although one can use markers that are not completely independent in the sense that they have alleles in linkage disequilibrium, this practice may require the use of a greater number of markers to make up for the lack of independence of the markers. Null loci can include genotype data available from, e.g., a previous genome-wide association or linkage studies involving the subjects or populations of interest, and could thus allow for a retrospective analysis of sample genetic background structure without additional genotyping. As in cluster analysis, the proposed method involves the construction of a genetic similarity matrix. However, it does not require cluster analysis to test hypotheses about the relationships of the individuals or populations in a sample. Rather, the method assumes that interest lies in testing the relationship between a particular grouping factor (e.g., race, country of origin, cohort, or geographical locale) or quantitative measure (such as age, cholesterol level, or weight) and variations in the genetic similarities of the individuals or populations collected. Therefore, it does not require the determination of the optimal number of clusters or, e.g., principal components, representing the data.

The proposed method is similar to the analysis of molecular variance (AMOVA) method introduced by Excoffier and colleagues, but is more flexible and provides a much more intuitive and generalizable derivation of relevant test statistics [50]. The description of the AMOVA procedure provided by Excoffier et al.[50] includes relevant sum-of-squares calculations to formulate analysis of variance (ANOVA)-like hypothesis-oriented test statistics that consider differences between groups of individuals or populations with respect to genetic background. As described in the Methods section, the proposed approach builds off an analysis method we have termed multivariate distance matrix regression analysis and can be used to test hypotheses about not only categorical or grouping factors and genetic background, but quantitative traits as well [51,52]. In addition, the formulation of the proposed test statistics can be adapted for

use in multiple regression-like test settings, so that the relationships of multiple factors to genetic background can be explored. As a result of the connections between the proposed approach and the AMOVA approach of Excoffier et al. [50], we have labeled the proposed approach generalized molecular analysis of variance (GAMOVA).

In addition to the AMOVA procedure, the proposed GAMOVA procedure also has some similarities to the Mantel-based test statistic approach reviewed and extended by Smouse et al. [53]. The Mantel test is used to test the relationship between the entries or cells in two (or more) distance/similarity matrices. Thus, one could have a genetic background similarity matrix computed from different populations whose relationship to, e.g., a geographic distance matrix computed for the populations is of interest. The proposed GAMOVA procedure considers the relationship between the $N \times N$ entries (or cells; where N is the number of individuals or populations being studied) in a genetic background distance matrix and information, represented as N-dimensional vectors, on the N individuals or populations whose genetic background distances are reflected in the matrix.

Below we apply the GAMOVA procedure to three data sets available in the public domain to address some prevailing questions: (1) an analysis of the Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain (CEPH)–Human Genome Diversity Project (HGDP) Cell Line Panel, (2) an analysis of the morphological data made available by Howells [54,55] on human craniometric characters, and (3) an analysis of the International HapMap Project data addressing questions about the similarity of the individual chromosomes possessed by the subjects genotyped as part of the project. In addition, we also consider aspects of the power of the GAMOVA procedure via simulation studies.

Results

Analysis of the CEPH-HGDP Cell Line Panel Dataset

We considered the use of the proposed GAMOVA analysis to analyze the CEPH-HGDP Cell Line Panel data [56] in a number of ways. We constructed several distance matrices over 1,040 subjects collected from 51 worldwide populations based on: (1) individual IBS allele sharing and (2) Lynch-Ritland (LR) frequency weighted allele-sharing distance, and (3) the standard between-population genetic distance measure F_{ST} (see Methods). We then considered the relationship between additional information collected on those individuals (and/or populations) and variation in the similarity among the individuals and populations using the proposed GAMOVA procedure. The additional information included, for each individual, which of the 51 populations or ethnic groups they were from, the geographic location of that population (i.e., one of the five or seven global world regions associated with populations), and its distance from Addis Ababa in Africa [8]. In addition to the analyses based on individuals, geographic location and distance from Addis Ababa were also considered in analyses involving the 51 populations as a whole. By considering the distance of each population from Addis Ababa we could address hypotheses about global historical migration patterns and the impact these migration patterns have on genomic diversity, as has been recently pursued through the use of different statistical methods [6,57].

To visually assess the potential for genetic background clustering we first constructed neighbor-joining trees based on the IBS distance matrix of the CEPH-HGDP individuals. We color-coded each branch (representing an individual) based on: (1) which of 5 major geographic regions (Figure 1A, left panel) and (2) which of 51 populations an individual was from (Figure 1B, right panel). Figure 1 shows a fairly dramatic clustering of the individuals that is roughly consistent with the population of origin for each individual. Note that, as observed by Rosenberg et al. [1], the Mozabite (the population labeled with a "6"), a Berber ethnic group living in the Sahara in Northern Africa, clusters with Middle Eastern populations (assigned labels "4," "5," and "7").

We then considered two analyses designed to assess how much of the genetic background variation exhibited by the CEPH-HGDP individuals and ethnic groups could be explained by the world regions each individual or population was associated with, as well as the distance of that world region from Addis Ababa, using the GAMOVA procedure. We created simple 0–1 indicator variables that reflected which world region an individual or population was associated with and used these indicator variables as independent or predictor variables in the GAMOVA regression procedure (see the Methods section for details) along with distance from Addis Ababa as a continuous variable. Table 1 provides the results assuming either a seven-world region breakdown (Table 1; East Asia, Africa, Oceania, Central and South Asia, America, the Middle East, and Europe) or a five-world region breakdown (Table 1; Eurasia, East Asia, Oceania, America, and Africa) as defined previously [1]. We also compare GAMOVA regression models that did not consider (Table 1) distance from Addis Ababa as a predictor to contrast the results with the findings of models that included it (Table 1).

The top half of Table 1 reflects the analysis of the IBS allele sharing among individuals and suggests that approximately 9%–11% of the variation in the similarity of individual genetic backgrounds can be explained by world region either in conjunction with the distance of that world region from Addis Ababa or not. Approximately 68%–72% of the variation in genetic background similarity of the populations as a whole, assuming the F_{ST} measure of genetic distance, could be explained by world region and distance of those world regions from Addis Ababa (Table 1; bottom half). This clearly reflects the greater diversity among individual genomes within a population than allele frequency differences between populations as a whole.

It is also interesting to note that, as found by others [6,57], the distance from Addis Ababa is the strongest predictor of genetic background similarity among the individuals and populations, but the world regions explain variation in genetic background similarity over and above this measure, suggesting that diversity among individuals within populations situated within the same world region is not completely captured by their distance from Addis Ababa. Also of note is the strength of the contributions of the various world regions to variation in genetic background similarity, which reflect factors such as the populations' individual demographic histories and selective environmental pressures. For example, Africa is the strongest contributor to individual genetic background similarity after accounting for each world

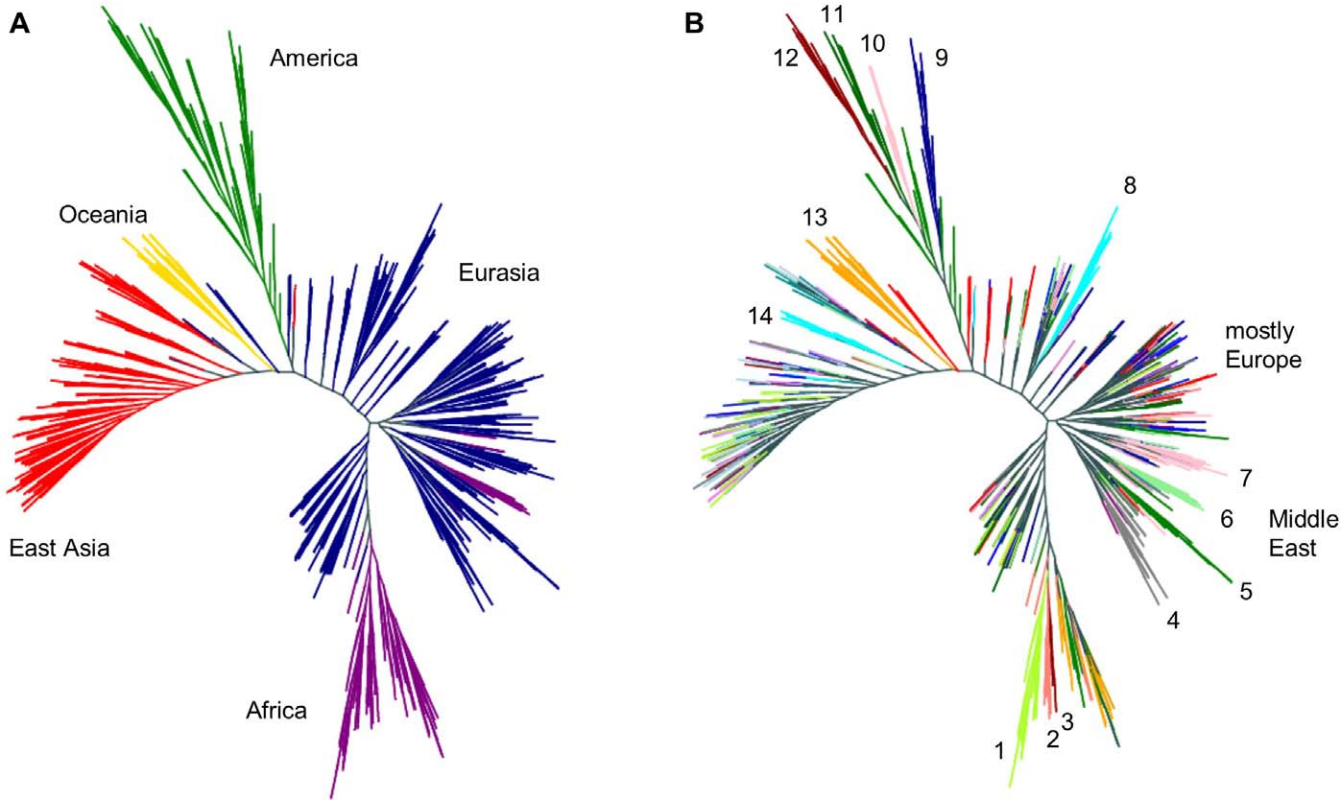

Figure 1. Neighbor-Joining Trees Depicting the Genetic Relationships of 1,040 Individuals from 51 World Populations Collected by the CEPH-HGDP
(A) Individuals are color coded according to which of five major geographic regions of the globe they are collected from.
(B) Individuals are color coded according to which of the 51 populations they are associated with (1: Biaka Pygmy, 2: San, 3: Mbuti Pygmy, 4: Druze; 5: Bedouin, 6: Mozabite, 7: Palestinian, 8: Kalash, 9: Pima, 10: Columbian, 11: Karitiana, 12: Surui, 13: New Guinea, 14: Yakut).
doi:10.1371/journal.pgen.0030051.g001

region's distance from Addis Ababa (Distance considered/IBS Matrix portion of Table 1), which is consistent with the deep genetic structure of this continent [58]. On the other hand, the strongest contributor to pairwise population distances (F_{ST}) after accounting for geographic distance from Addis Ababa (Distance considered/F_{ST} portion of Table 1) was found to be America, consistent with findings by Ramachandran et al. [58].

We also considered analyses that took into consideration all the populations studied, assuming both the IBS allele-sharing measure of genetic background similarity and the LR allele frequency-weighted measure (Table 2). Overall, the individual populations that the study subjects were from could explain approximately 16%–19% of the variation in genetic background similarity exhibited by the individuals in the CEPH-HGDP database. Interestingly, the analyses using the IBS and LR measures did not agree perfectly—although they are similar—suggesting that allele frequency weighting can make a difference in assessing individual genetic background similarity. In addition, our GAMOVA analysis suggests that individuals from three populations in the Americas (the Surui, the Karitiana, and the Pima) have the most divergent genomes from the other individuals' genomes, which has been observed by others as well (e.g., [1,58]).

Analysis of the Craniometric Data Collected by Howells

We also considered analyses involving morphological data made available by Howells [54,55] on human craniometric characters collected on individuals from ten worldwide populations. We computed the median of each of 43 craniometric measures for males and females separately from each of these populations. We combined the data with the genetic data on the CEPH-HGDP subjects by geographically matching the countries and regions represented in the CEPH-HGDP with those for which we had craniometric data in a fashion identical to the one outlined by Roseman [59]. The median values for each of the 43 craniometric measures were then considered as a regressor or covariate in a GAMOVA analysis of the genetic distance matrix computed for the ten corresponding CEPH-HGDP populations. The goal was to test associations between craniometric features of the people within the populations and genetic background similarities those people might have with people in other populations. We want to emphasize that many of the craniometric measures are correlated so that associations between any one of these measures and genetic background suggest that other measures may also be associated with genetic background, just not necessarily independently of the others.

Table 3 describes the results of the analyses for males and females. The cranial feature most strongly associated with genetic background similarity is the nasion-bregma subtense (FRS), which "explains" ~54% and ~49% of the variation in genetic background similarity for males and females, respectively. Other measures, such as glabella projection, minimum cranial breadth and basion-prosthion length for males, and

Table 1. GAMOVA Analysis Estimates of the Poportion of Variation in Genetic Background Similarity Explained by Seven or Five World Regions, Respectively, Including and Excluding Geographic Dstances (between Each Population and Addis Ababa, See Text)

Distance from Addis Ababa as a Predictor	Matrix	Seven-World Region Breakdown						Five-World Region Breakdown					
		Seven Regions	SS	Pseudo-*F*	*p*	Prop	Cum Prop	Five Regions	SS	Pseudo-*F*	*p*	Prop	Cum Prop
Distance considered	IBS	Geographic distances	8.018	39.235	0.0001	0.0365	0.0365	Geographic distances	8.018	39.235	0.0001	0.0365	0.0365
	IBS	Africa	5.205	26.088	0.0001	0.0237	0.0601	Africa	5.205	26.088	0.0001	0.0237	0.0601
	IBS	East Asia	4.804	24.625	0.0001	0.0218	0.0820	East Asia	4.804	24.625	0.0001	0.0218	0.0820
	IBS	Oceania	2.433	12.613	0.0001	0.0111	0.0930	Oceania	2.433	12.613	0.0001	0.0111	0.0930
	IBS	Central and South Asia	1.128	5.873	0.0001	0.0051	0.0982	America	1.064	5.539	0.0001	0.0048	0.0979
	IBS	America	1.117	5.847	0.0001	0.0051	0.1032	Eurasia	0.474	2.470	0.0001	0.0022	0.1000
	IBS	Middle East	0.847	4.447	0.0001	0.0039	0.1071						
	IBS	Europe	0.010	0.050	1	0	0.1071						
Distance not considered	IBS	America	6.786	33.016	0.0001	0.0309	0.0309	America	6.786	33.016	0.0001	0.0309	0.0309
	IBS	East Asia	6.179	30.927	0.0001	0.0281	0.0589	East Asia	6.179	30.927	0.0001	0.0281	0.0589
	IBS	Africa	5.245	26.911	0.0001	0.0238	0.0828	Africa	5.245	26.911	0.0001	0.0238	0.0828
	IBS	Oceania	2.310	11.978	0.0001	0.0105	0.0933	Eurasia	2.310	11.978	0.0001	0.0105	0.0933
	IBS	Central and South Asia	1.243	6.479	0.0001	0.0057	0.0990	Oceania	0.000	0.000	1	0	0.0933
	IBS	Europe	0.785	4.104	0.0001	0.0036	0.1025						
	IBS	Middle East	0.000	0.000	1	0	0.1025						
Distance considered	F_{ST}	Geographic distances	0.048	28.379	0.0001	0.3575	0.3575	Geographic distances	0.048	28.379	0.0001	0.3575	0.3575
	F_{ST}	America	0.018	13.450	0.0001	0.1362	0.4937	America	0.018	13.450	0.0001	0.1362	0.4937
	F_{ST}	Africa	0.012	10.356	0.0001	0.0883	0.5820	Africa	0.012	10.356	0.0001	0.0883	0.5820
	F_{ST}	East Asia	0.010	9.762	0.0001	0.0706	0.6527	East Asia	0.010	9.762	0.0001	0.0706	0.6527
	F_{ST}	Oceania	0.007	8.034	0.001	0.0507	0.7034	Eurasia	0.007	8.034	0.001	0.0507	0.7034
	F_{ST}	Middle East	0.002	2.606	0.0916	0.0159	0.7193	Oceania	0.000	0.002	0.9472	0.0000	0.7034
	F_{ST}	Europe	0.001	1.126	0.3287	0.0069	0.7261						
	F_{ST}	Central and South Asia	0.000	0.000	1	0	0.7261						
Distance not considered	F_{ST}	America	0.046	26.757	0.0001	0.3441	0.3441	America	0.046	26.757	0.0001	0.3441	0.3441
	F_{ST}	Africa	0.023	17.180	0.0001	0.1677	0.5118	Africa	0.023	17.180	0.0001	0.1677	0.5118
	F_{ST}	East Asia	0.014	13.633	0.0001	0.1063	0.6181	Eurasia	0.015	14.999	0.0001	0.1144	0.6262
	F_{ST}	Oceania	0.008	9.032	0.0171	0.0605	0.6786	East Asia	0.007	7.814	0.0297	0.0523	0.6786
	F_{ST}	Central and South Asia	0.001	1.110	0.3321	0.0074	0.6860	Oceania	0.000	0.000	1	0	0.6786
	F_{ST}	Europe	0.000	0.305	0.7015	0.0021	0.6881						
	F_{ST}	Middle East	0.000	0.000	1	0	0.6881						

Note: Estimates are based on IBS allele-sharing information across 1,040 individuals (top half) genotyped on 783 markers, and pairwise genetic distances (F_{ST}) between 51 populations (bottom half).

SS, sum of squares from the analysis; Pseudo-F,statistic measuring the influence of the population on individual genetic background similarities; *p, p*-value associated with the pseudo-F based on 1000 data permutations; Prop, proportion of variation in the genetic background similarity matrix explained by the population; Cum Prop,: cumulative proportion of variation explained by the population.

doi:10.1371/journal.pgen.0030051.t001

brasion-prosthion length and dacryon subtense for females, were found to be also associated with variation in population genetic profile similarity over and above the FRS. The multicollinearity among the 43 measures precluded fitting a GAMOVA model with all 43 measures as predictors, so only those measures that had associations with genetic background similarity that were independent of the others were considered in Table 3 (i.e., as in standard multiple regression contexts). A strong association between the frontal bone curvature FRS with genetic background has also been reported by Roseman and Weaver using a principal components analysis [60]. As found by others (e.g., [61]), our analysis suggests that certain morphological features, namely craniometric features, segregate with genetic background across different global populations much like, e.g., skin color [62].

Analysis of the HapMap Dataset

We next considered the application of the GAMOVA procedure to the analysis of the large-scale genotyping effort associated with the International HapMap Project (http://www.hapmap.org; [63]). The data consist of genotypes at over a million SNP loci on 209 individuals associated with four different population groups (Northern European, West African, Japanese, and Han Chinese). Computational methods were used to "phase" individuals based on the genotype data (i.e., probabilistically assign unique chromosome pairs to each individual based on linkage disequilibrium patterns) by

Table 2. GAMOVA Analysis Estimates of the Proportion of Variation in CEPH-HGDP Individual Genetic Background Similarity Explained by the 51 Populations from Which the Subjects Were Collected on the Basis of IBS Allele-Sharing Information (Left Half) and an Allele Frequency Weighted Measure LR (Right Half)

IBS Allele-Sharing Matrix							Lynch-Ritland Unbiased Estimates of Pairwise Relatedness						
Region	**Population**	**SS**	**Pseudo-*F***	***p***	**Prop**	**Cum Prop**	**Region**	**Population**	**SS**	**Pseudo-*F***	***p***	**Prop**	**Cum Prop**
America	Surui	3.3381	14.9873	0.001	0.0142	0.0142	America	Surui	7.4226	15.002	0.001	0.0142	0.0142
America	Karitiana	3.0619	13.9181	0.001	0.0131	0.0273	America	Karitiana	7.0345	14.401	0.001	0.0135	0.0277
America	Pima	2.5814	11.8569	0.001	0.011	0.0383	America	Pima	5.5845	11.5489	0.001	0.0107	0.0385
Africa	Biaka Pygmies	2.288	10.6067	0.001	0.0098	0.0481	Africa	Biaka Pygmies	3.6481	7.5925	0.001	0.007	0.0455
Oceania	Melanesian	1.5306	7.1376	0.001	0.0065	0.0546	Oceania	Melanesian	2.9607	6.1927	0.001	0.0057	0.0512
Africa	Mbuti Pygmies	1.5059	7.0634	0.001	0.0064	0.061	America	Columbian	2.9101	6.1169	0.001	0.0056	0.0567
Africa	Yoruba	1.5005	7.0796	0.001	0.0064	0.0674	America	Maya	2.9224	6.1737	0.001	0.0056	0.0623
Africa	Mandenka	1.5108	7.1708	0.001	0.0064	0.0738	Central and South Asia	Kalash	2.7421	5.8198	0.001	0.0053	0.0676
America	Maya	1.4739	7.0365	0.001	0.0063	0.0801	Africa	Yoruba	2.3232	4.9495	0.001	0.0045	0.0721
America	Columbian	1.4346	6.8881	0.001	0.0061	0.0862	Oceania	New Guinea	2.3014	4.9218	0.001	0.0044	0.0765
Oceania	New Guinea	1.2945	6.2471	0.001	0.0055	0.0918	Africa	Mbuti Pygmies	2.2503	4.8306	0.001	0.0043	0.0808
East Asia	Han	1.2763	6.1904	0.001	0.0054	0.0972	Africa	Mandenka	2.258	4.8653	0.001	0.0043	0.0851
East Asia	Japanese	1.1492	5.5992	0.001	0.0049	0.1021	East Asia	Han	2.1541	4.6579	0.001	0.0041	0.0893
Central and South Asia	Kalash	1.133	5.5443	0.001	0.0048	0.1069	East Asia	Yakut	2.0283	4.4005	0.001	0.0039	0.0932
East Asia	Yakut	1.0295	5.0579	0.001	0.0044	0.1113	East Asia	Japanese	1.9907	4.3329	0.001	0.0038	0.097
Africa	San	0.9006	4.4397	0.001	0.0038	0.1152	Middle East	Druze	1.7833	3.8924	0.001	0.0034	0.1004
Middle East	Druze	0.8567	4.2363	0.001	0.0037	0.1188	Middle East	Bedouin	1.6137	3.531	0.001	0.0031	0.1035
Africa	Mozabite	0.8117	4.0259	0.001	0.0035	0.1223	Africa	Mozabite	1.6457	3.6102	0.001	0.0032	0.1067
Africa	Bantu	0.809	4.0243	0.001	0.0034	0.1257	Middle East	Palestinian	1.5315	3.3675	0.001	0.0029	0.1096
Middle East	Bedouin	0.8165	4.0738	0.001	0.0035	0.1292	Africa	San	1.5192	3.348	0.001	0.0029	0.1125
Middle East	Palestinian	0.8084	4.0456	0.001	0.0034	0.1327	Europe	Sardinian	1.3811	3.0499	0.001	0.0027	0.1152
Europe	Sardinian	0.6942	3.4823	0.001	0.003	0.1356	Africa	Bantu	1.3605	3.0102	0.001	0.0026	0.1178
Europe	Basque	0.605	3.041	0.001	0.0026	0.1382	East Asia	Lahu	1.337	2.9641	0.001	0.0026	0.1204
East Asia	Lahu	0.5881	2.9618	0.001	0.0025	0.1407	Europe	Basque	1.2466	2.7684	0.001	0.0024	0.1227
Europe	French	0.5602	2.8264	0.001	0.0024	0.1431	Europe	French	1.0849	2.4128	0.001	0.0021	0.1248
Europe	Russian	0.527	2.6634	0.001	0.0022	0.1453	Europe	Russian	1.1159	2.4853	0.001	0.0021	0.127
Europe	Orcadian	0.545	2.7588	0.001	0.0023	0.1477	Europe	Orcadian	1.1049	2.4642	0.001	0.0021	0.1291
Central and South Asia	Brahui	0.5225	2.6495	0.001	0.0022	0.1499	Central and South Asia	Brahui	1.1198	2.5012	0.001	0.0021	0.1312
Central and South Asia	Balochi	0.4868	2.4718	0.001	0.0021	0.152	Central and South Asia	Burusho	0.9984	2.2328	0.001	0.0019	0.1332
Central and South Asia	Makrani	0.5269	2.68	0.001	0.0022	0.1542	Central and South Asia	Balochi	1.044	2.3379	0.001	0.002	0.1352
Central and South Asia	Sindhi	0.5417	2.76	0.001	0.0023	0.1565	Central and South Asia	Sindhi	1.0895	2.4433	0.001	0.0021	0.1373
Central and South Asia	Pathan	0.5677	2.8982	0.001	0.0024	0.1589	Central and South Asia	Makrani	1.1999	2.6953	0.001	0.0023	0.1396
Central and South Asia	Burusho	0.6277	3.2116	0.001	0.0027	0.1616	Central and South Asia	Pathan	1.206	2.7136	0.001	0.0023	0.1419
Europe	Adygei	0.6619	3.3947	0.001	0.0028	0.1644	Europe	Adygei	1.2703	2.8637	0.001	0.0024	0.1443
Europe	Bergamo	0.6752	3.471	0.001	0.0029	0.1673	Europe	Bergamo	1.241	2.8026	0.001	0.0024	0.1467
Europe	Tuscan	0.5209	2.6826	0.001	0.0022	0.1695	Europe	Tuscan	0.9923	2.2438	0.001	0.0019	0.1486
Central and South Asia	Hazara	0.4889	2.5215	0.001	0.0021	0.1716	Central and South Asia	Hazara	1.0283	2.3282	0.001	0.002	0.1506

Table 2. Continued.

IBS Allele-Sharing Matrix							Lynch-Ritland Unbiased Estimates of Pairwise Relatedness					
Region	**Population**	**SS**	**Pseudo-*F***	***p***	**Prop**	**Cum Prop**	**Region**	**Population**	**SS**	**Pseudo-*F***	***p***	**Prop**	**Cum Prop**
Central and South Asia	Uygur	0.3096	1.5975	0.001	0.0013	0.1729	East Asia	Cambodian	0.6929	1.5697	0.001	0.0013	0.1519
East Asia	Cambodian	0.2941	1.5187	0.001	0.0013	0.1742	Central and South Asia	Uygur	0.6881	1.5597	0.001	0.0013	0.1532
East Asia	Naxi	0.2864	1.4795	0.001	0.0012	0.1754	East Asia	Naxi	0.658	1.4922	0.001	0.0013	0.1545
East Asia	She	0.2865	1.4805	0.001	0.0012	0.1766	East Asia	She	0.6466	1.4669	0.001	0.0012	0.1557
East Asia	Dai	0.274	1.4167	0.001	0.0012	0.1778	East Asia	Oroqen	0.6333	1.4375	0.001	0.0012	0.1569
East Asia	Oroqen	0.2427	1.2553	0.001	0.001	0.1788	East Asia	Dai	0.6048	1.3734	0.001	0.0012	0.1581
East Asia	Hezhen	0.2262	1.1702	0.017	0.001	0.1798	East Asia	Hezhen	0.5488	1.2465	0.001	0.0011	0.1592
East Asia	Daur	0.217	1.1227	0.053	0.0009	0.1807	East Asia	Daur	0.5346	1.2144	0.001	0.001	0.1602
East Asia	Tu	0.2166	1.1205	0.057	0.0009	0.1817	East Asia	Tu	0.5189	1.1791	0.006	0.001	0.1612
East Asia	Xibo	0.2115	1.0941	0.12	0.0009	0.1826	East Asia	Yizu	0.519	1.1794	0.003	0.001	0.1622
East Asia	Yizu	0.2135	1.1048	0.102	0.0009	0.1835	East Asia	Mongola	0.5306	1.2061	0.002	0.001	0.1632
East Asia	Mongola	0.2167	1.1217	0.058	0.0009	0.1844	East Asia	Xibo	0.5418	1.2319	0.001	0.001	0.1642
East Asia	Miaozu	0.1733	0.8966	0.908	0.0007	0.1851	East Asia	Miaozu	0.4199	0.9547	0.749	0.0008	0.1650
East Asia	Tujia	0	0	1	0	0.1851	East Asia	Tujia	0	0	1	0	0.1650

Note: Calculations are based on the analysis of 1,040 individuals.

SS, sum of squares from the analysis; Pseudo-*F*, statistic measuring the influence of the population on individual genetic background similarities; *p*, *p*-value associated with the pseudo-*F* based on 1,000 data permutations; Prop, proportion of variation in the genetic background similarity matrix explained by the population; Cum Prop, cumulative proportion of variation explained by the population.

doi:10.1371/journal.pgen.0030051.t002

the HapMap investigators; see the description of the phasing procedures at the International HapMap Project Web site. We undertook an analysis investigating the fraction of genetic similarity explained by the four population subgroups on a per-chromosome basis using a simple IBS measure of genotype similarity for the 209 individuals, as well as individual chromosomal similarity based on the $209 \times 2 = 418$ chromosomes obtained from the phase-resolved data. The goal of this analysis was to determine how much of the similarity or distances between the multilocus diploid genomes, on a per-chromosome basis, could be explained by the population groups associated with the HapMap individuals. We also wanted to determine how much of the similarity or distances between the individual chromosomes (with each person contributing two to the total pool of $209 \times 2 = 418$) could be explained by the population groups associated with the Hapmap individuals.

Table 4 describes the results and suggests that roughly 20%–22% of the individual chromosomal similarity can be explained by the populations associated with each chromosome (i.e., assigned haplotypes) (left half of Table 4) and roughly 28%–30% of the individual chromosomal similarity based on each individual's diploid genotype can be explained by the population origins of the subjects. We note that the percentages are consistent across the chromosomes, as one might expect. In addition, the Yoruban population has the most divergent chromosomes, followed by the Northern Europeans. The distinction between the Han Chinese and Japanese chromosomes, although significant, is much weaker, as expected, since the residual variation after accounting for African and European background effects is very small. In addition, whereas the effect of Chinese origin was more significant on individual chromosome similarity, the effect of Japanese origin was more significant on genotyping similarity.

Power Estimation

We also considered the power of the proposed GAMOVA procedure to detect varying degrees of differentiation between two populations using simulated data. We chose simulation settings that were consistent with those recently described by Patterson et al. [64]. We simulated four different settings/datasets with two populations each, whose pairwise genetic distances ranged from $F_{ST} = 0$ to $F_{ST} = 0.01$ (see Methods). We performed a GAMOVA analysis on these data with known group membership taken as a predictor variable. These analyses were repeated for a total of 1,000 simulations in each setting. Results were binned in groups having different F_{ST} statistics calculated for each data set (i.e., knowing the assumed F_{ST} used to generate the data may differ from the F_{ST} calculated from the simulated sample). Figure 2 shows the relationship of F_{ST} between the two populations to power of GAMOVA to detect that level of differentiation at a type-I error rate of 0.05. In general, GAMOVA shows excellent power at very low F_{ST} values around 0.0002, which is in the range of the least differentiated human populations described in literature (e.g., for different geographic regions of Iceland, a homogenous genetic isolate [14]). As noted by Patterson et al. [64], we found that at a fixed data size ($D =$ number of markers $\times$ number of subjects), genetic differentiation is easier to detect for larger sample sizes, even though a smaller number of markers is used, than for smaller sample sizes using a larger number of markers.

Table 3. GAMOVA Analysis Investigating the Relationship between Craniometric Measures Collected by Howells and Genetic Background

Gender	Measure	SS	Pseudo-*F*	*p*	Prop	Cum Prop	Comment
Male	FRS	0.0129	9.298	0.002	0.5375	0.54	nasion-bregma subtense
	GLS	0.0044	4.611	0.030	0.1837	0.72	glabella projection
	WCB	0.0032	5.364	0.004	0.1316	0.85	minimum cranial breadth
	BPL	0.0013	2.813	0.039	0.0530	0.91	basion-prosthion length
	NAS	0.0010	2.950	0.081	0.0400	0.95	nasio-frontal subtense
	FRC	0.0008	4.521	0.093	0.0326	0.98	nasion-bregma chord
	OCC	0.0006	−16.485	0.974	0.0246	1.00	lambda-opisthion chord
	OBH	0.0000	−0.313	0.673	0.0014	1.00	orbit height
Female	FRS	0.0114	6.676	0.009	0.4881	0.49	nasion-bregma subtense
	BPL	0.0058	5.760	0.016	0.2507	0.74	basion-prosthion length
	SOS	0.0022	2.917	0.068	0.0962	0.84	supraorbital projection
	MDH	0.0017	3.149	0.103	0.0727	0.91	mastoid height
	DKS	0.0018	17.017	0.021	0.0785	0.99	dacryon subtense
	OCC	0.0003	55.968	0.008	0.0134	1.00	lambda-opisthion chord
	FMB	0.0002	−1.059	0.744	0.0086	1.01	bifrontal breadth

SS, sum of squares from the analysis; Pseudo-*F*, statistic measuring the influence of the population on individual genetic background similarities; *p*, *p*-value associated with the pseudo-*F* based on 10,000 data permutations; Prop, proportion of variation in the genetic background similarity matrix explained by the population; Cum Prop, cumulative proportion of variation explained by the population.

doi:10.1371/journal.pgen.0030051.t003

Discussion

As DNA sequencing and genotyping costs decrease, a greater number of population scientists, geneticists, clinical researchers, and epidemiologists will seek to identify and characterize genetic variations that underlie phenotypic variations as well as the biological relationships among individuals. Flexible analysis tools that can be used to test appropriate hypotheses will thus be needed for these investigations. We have proposed an analysis procedure, GAMOVA, that not only extends an analysis of variance approach that is used widely [50] for assessing relationships between genetic variations, phenotypic variations, and the population origins of individuals, but also complements widely used cluster analysis approaches for these purposes [20,23]. Specifically, the proposed GAMOVA approach can be used to test hypotheses about the relationship between variables collected on individuals or populations (such as particular phenotypes or population-level migration patterns) and variation in the genetic similarity or distance of

Table 4. Percentage of the Variation in the Dissimilarity of Individual Chromosomes or Diploid Genotypes Explained as a Function of the Population Designations of the 209 Subjects Genotyped as Part of the HapMap Project

Chromosome	Chromosomes				Diploid Genotypes			
	Total	Yoruban	CEPH	Chinese	Total	Yoruban	CEPH	Japanese
1	22.15	14.15	7.61	0.39	29.20	18.37	10.06	0.59
2	23.43	14.78	8.24	0.41	30.39	19.05	10.74	0.61
3	22.29	14.53	7.40	0.37	29.12	18.77	9.77	0.57
4	21.90	14.79	6.73	0.38	28.71	19.22	8.91	0.59
5	21.31	13.44	7.46	0.41	28.10	17.57	9.90	0.63
6	21.21	13.37	7.46	0.37	28.03	17.58	9.85	0.59
7	21.14	13.35	7.39	0.40	27.85	17.41	9.82	0.62
8	21.76	14.27	7.09	0.40	28.57	18.65	9.31	0.60
9	20.89	12.67	7.79	0.43	27.50	16.59	10.24	0.67
10	21.83	13.45	7.96	0.42	28.66	17.51	10.52	0.63
11	20.79	13.36	7.01	0.42	27.42	17.51	9.26	0.65
12	21.85	13.62	7.84	0.38	28.80	17.83	10.37	0.59
13	20.79	12.91	7.47	0.41	27.48	17.01	9.85	0.62
14	21.68	13.58	7.68	0.42	28.54	17.74	10.17	0.63
15	23.61	14.16	9.02	0.43	30.67	18.41	11.60	0.66
16	21.81	13.92	7.49	0.40	28.68	18.19	9.89	0.60
17	23.25	15.34	7.54	0.37	30.17	19.79	9.81	0.57
18	19.64	13.02	6.21	0.41	26.06	17.13	8.28	0.64
19	20.96	13.50	7.08	0.38	27.54	17.59	9.33	0.61
20	21.67	14.10	7.14	0.43	28.42	18.34	9.43	0.65
21	20.16	12.81	6.93	0.41	26.59	16.77	9.21	0.60
22	21.49	13.68	7.38	0.43	28.34	17.89	9.76	0.69

doi:10.1371/journal.pgen.0030051.t004

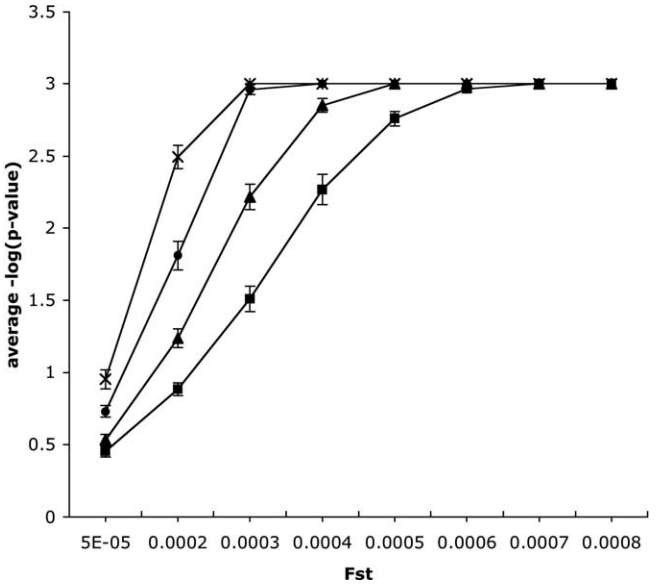

Figure 2. Relationship between the Genetic Differentiation among Two Populations as Measured by Wright's F_{ST} and the Average (±S.E.M.) Power of the GAMOVA Procedure to Detect that Differentiation

Results are based on 1,000 simulation studies involving four sets of two equally sized populations, each generated according to varying genetic differentiation. Known group membership was used as predictor in the GAMOVA analysis. For a constant data size (number of markers × number of subjects), genetic differentiation can be detected at lower F_{ST} values in larger populations with fewer markers compared to smaller populations with more markers (squares: 32 individuals, 32768 markers; triangles: 64 individuals, 16384 markers; circles: 128 individuals, 8192 markers; stars: 256 individuals, 4096 markers).

doi:10.1371/journal.pgen.0030051.g002

those individuals or populations as characterized by individual genotype or allele frequency data.

Our applications of the proposed GAMOVA procedure suggest that it can be used to address a number of population genetic questions concerning the relationships of individuals at the DNA sequence level; e.g., it can be used to directly quantify the degree to which certain factors, such as race, self-reported ethnicity, admixture, migration patterns, and anecdotally derived connections between individuals and populations, are associated with the genetic similarity of individuals and populations. The exploration of such relationships has been the hallmark of applied population genetics research for decades [65–68]. However, one particularly important area of application for the proposed procedure is in the area of genetic epidemiology, and genetic association studies in particular, for at least two reasons. First, it is well known that the polygenic and/or multifactorial nature of many traits and diseases can influence the identification of the individual loci contributing to the expression of those traits and diseases if not accounted for appropriately [69,70]. Second, it is also well known that population stratification or genetically distinct subdivisions within a population sampled for an association study can lead to both false positive and false negative results if ignored [12–16,18,69]. In these two contexts, the proposed GAMOVA approach can be used to test hypotheses about the relationship between a phenotype of interest and genetic background similarity among the subjects to be used in an association study (provided that they have been genotyped on an appropriate set of markers to assess genetic background [28,71,72]). If an association is found, then steps can be taken to accommodate the influence of genetic background on the trait or disease in question as described by many researchers. The steps that can be taken to control for genetic background heterogeneity within the context of the GAMOVA analysis could involve identifying the leading eigenvalues of the distance/similarity matrix and using the corresponding eigenvectors as regressor variables or covariates in an appropriate linear model relating the specific genetic variation in question to the phenotype of interest [29,64].

Properties of the GAMOVA procedure, i.e., its robustness, power, level accuracy, etc., have been studied in some very general contexts, such as those involving genetic association analyses, gene expression analyses, and DNA sequence–based association studies [51,52], as well as in the simulation studies presented here. For population genetic analyses, our simulations suggest that the GAMOVA procedure is sensitive enough to detect very low levels of population structure in epidemiological samples. In addition, the use of permutation tests provides a very robust method for testing hypotheses demonstrating that the procedure is powerful in many different settings. In addition, virtually all of these studies document the flexibility of the method.

In addition to the applications showcased here, as well as those outlined by Wessel and Schork [51] and Zapala and Schork [52], we have routinely made use of the GAMOVA analysis to test for, e.g., differences across studies due to laboratory effects or genotyping artifacts, genotyping quality shifts over time, and genetic background differences between subjects from an original and replication sample [73]. Finally, the GAMOVA procedure is also applicable to the identification of informative markers for specific cohorts or communities under study, since one can use the procedure to test the effect of each SNP on variation in a genetic background similarity matrix for informativeness without requiring knowledge about the ancestral history of the subjects under study.

There are, however, a few limitations inherent in the proposed GAMOVA approach that may provide fertile ground for further research. For example, the choice of a similarity or distance measure is crucial. Although the IBS and LR measures for individual genetic similarity and the F_{ST} and related measures for population-level genetic similarity (e.g., [6]) are the standards, it is unclear which of these measures are the most powerful to use in the GAMOVA procedure (or even other methods relying on distance measures besides GAMOVA). In this context the power of the proposed GAMOVA approach in different population analysis settings and locus effect scenarios deserves detailed attention. However, since the procedure is rooted in the derivation of traditional ANOVA, regression, and general linear models, many of the same intuitions and findings related to the power of these modeling procedures apply. For example, the proposed procedure assesses the question of how much of the variation in the similarity/dissimilarity exhibited by a group of individuals can be explained by another factor, which is analogous to questions concerning how much of the variation in a quantitative particular trait is explained by a certain factor in regression and ANOVA contexts.

A final concern with the proposed approach, which is an issue with all analysis methodologies that involve high-

dimensional data types, involves missing genotype data. One can handle missing genotype data in a number of ways. First, one could restrict the construction of the similarity measure to only those individuals with complete data—which may result in a substantially reduced sample—or simply construct the measure with the data that are available on each pair of subjects. This latter approach will be problematic if a number of individuals are missing genotype data at the most heavily weighted (e.g., functional or informative) loci. Another approach to handling missing data would involve imputing or assigning individuals genotype data based on linkage disequilibrium information. This approach would only be as useful as the strength of the linkage disequilibrium between alleles at the loci with missing data and those without. The approach we took to handling missing data was to use whatever genotype information was available on the subjects for the similarity calculations.

Finally, we note that a web-based GAMOVA tool is available from the authors at http://polymorphism.scripps.edu/~cabney/cgi-bin/mmr.cgi.

Materials and Methods

Computing a similarity matrix. As noted, the proposed procedure requires the computation of a "distance" matrix that reflects the dissimilarity of the genetic backgrounds of the individuals or populations being analyzed. There are many possible measures that could be used to construct such a matrix, and we considered two methods for computing the similarity of individuals' genetic backgrounds based on genotype data collected on them. The resulting similarity measure can be translated into a distance or dissimilarity measure as described later. The first similarity measure is widely used and is based on simple IBS allele sharing [38] and can be calculated as the fraction of alleles shared identical by state for each pair of individuals in a sample over all the loci for which the individuals have been genotyped:

$$\frac{1}{2L}\sum_{l=1}^{L}\hat{r}_{IBS},l \tag{1}$$

where $\hat{r}_{IBS}$ is the individual, locus-specific allele-sharing value and $L=$ number of loci considered in the calculations.

The second similarity measure essentially considers weighting loci in the computation of IBS-based allele sharing by allele frequency and was introduced by Lynch and Ritland [44]. The LR regression-based method-of-moments estimator has been shown to have some desirable properties relative to other methods, especially in the case of populations consisting of individuals with a low degree of relatedness [45,74], and has been widely discussed in the population genetics and behavioral ecology literature (e.g., [75–77]). The LR estimator uses a regression approach to infer relationships (i.e., one individual of a pair serves as a "reference" individual and the probabilities of the locus-specific genotypes of the second individual are then conditioned on those of the reference individual). The LR coefficient of relatedness is:

$$\hat{r}_{xy}=\frac{p_a(S_{bc}+S_{bd})+p_b(S_{ac}+S_{ad})-4p_ap_b}{(1+S_{ab})(p_a+p_b)-4p_ap_b} \tag{2}$$

where p_a and p_b equal the frequencies of alleles a and b in the population. The reference individual is assumed to have alleles a and b (such that if this individual is homozygous, S_{ab}=1, if heterozygous, S_{ab} = 0), and the proband has alleles c and d. Multilocus estimates of genetic background similarity can be obtained by summing the single estimates, weighted by the inverse of their sampling variance:

$$\hat{r}_{xy}=\frac{1}{W_{r,x}}\sum_{l=1}^{L}W_{r,x}(l)\hat{r}_{xy}(l) \tag{3}$$

where

$$W_{r,x}(l)=\frac{1}{Var[\hat{r}_{xy}(l)]}\frac{(1+S_{ab})(p_a+p_b)-4p_ap_b}{2p_ap_b}, \tag{4}$$

which is computed under the assumption that the two individuals in question are unrelated (i.e., have 0.0 relatedness).

The similarity matrices were transformed into a dissimilarity or "distance" matrix by subtracting the components of the matrix from 1.0 if the IBS measure is used, or subtracting them from 1.0 after each component in the matrix is divided by the theoretical or empirical maximum of the similarity measure to scale the entries to lie between 0 and 1.

Multivariate distance matrix regression analysis. Once one has computed a distance matrix it can be subjected to a regression analysis testing hypotheses regarding, e.g., whether or not variation in the level of similarity/dissimilarity exhibited by pairs of individuals reflected in that matrix can be explained by other features those individuals posses (e.g., whether they are from a particular ethnic group or a specific country). To describe the regression model, we assume that each of N individuals or study subjects has been genotyped at L unlinked polymorphic loci (bi- or multiallelic) and that M grouping or phenotypic variables have been collected on the N subjects. These grouping or phenotypic variables could include information about the country of origin (coded using dummy variables, such as a 1 assigned to individuals from a particular country, and 0 assigned to individuals from a different country), the continental origin of that country and its distance from Addis Ababa, and craniometric diversity data, as we have considered.

We note that the proposed regression procedure, which is an extension of the procedure described by McArdle and Anderson [78] and a general reformulation of the AMOVA procedure discussed by Excoffier et al. [50], does not require that the distance matrix used have metric properties. Let this distance matrix and its elements be denoted by $D=d_{ij}$ $(i,j=1,\ldots,N)$, for the N subjects. The possibility that $N \ll L$ will not pose problems in the proposed regression analysis setting. Let X be an $N \times M$ matrix harboring information on the M grouping or phenotypic variables, which will be modeled as predictor or regressor variables whose relationships to the values in the genomic similarity matrix are of interest. Compute the standard projection matrix, $H = X(X'X)^{-1}X'$, typically used to estimate coefficients relating the predictor variables to outcome variables in multiple regression contexts. Next, compute the matrix $A=(a_{ij})=(-[1/2]d_{ij}^2)$ and center this matrix using the transformation discussed by Gower [79] and denote this matrix G:

$$G=\left(I-\frac{1}{n}11'\right)A\left(I-\frac{1}{n}11'\right) \tag{5}$$

An F-statistic can be constructed to test the hypothesis that the M regressor variables have no relationship to variation in the genomic distance or dissimilarity of the N subjects reflected in the $N \times N$ distance/dissimilarity matrix as [78]:

$$F=\frac{tr(HGH)}{tr[(I-H)G(I-H)]} \tag{6}$$

If the Euclidean distance is used to construct the distance matrix on a single quantitative variable (i.e., as in a univariate analysis of that variable) and appropriate numerator and denominator degrees of freedom are accommodated in the test statistics, the F-statistic above is equivalent to the standard ANOVA F-statistic [78]. The distributional properties of the F-statistic are complicated for alternative distance measures computed for more than one variable, especially if those variables are discrete, as in genotype data. However, permutation tests can then be used to assess statistical significance of the pseudo F-statistic [80,81]. The M regressor variables can be tested individually or in a step-wise manner. All matrix-based regression analyses we have performed in this paper used 10,000 permutations to calculate p-values, except for the analysis of the CEPH-HGDP data in Table 2, for which we used 1,000 permutations. In addition, one can calculate the percentage of variation in similarity/distances within the distance matrix explained by the regressor variables, r^2, through the formula:

$$r^2=\frac{tr(HGH)}{tr(G)} \tag{7}$$

Graphical display of similarity matrices. Similarity matrices of the type we have described can be represented graphically in a number of ways (e.g., heatmaps and trees) that can facilitate interpretation. We considered trees that are constructed such that individuals with greater genomic similarity are placed next to each other (i.e., they are represented as adjacent branches of the tree) and less similar individuals are represented as branches some distance away from each other, using the module neighbor of the program PHYLIP v.

3.64 (http://evolution.genetics.washington.edu/phylip.html) to construct a neighbor-joining tree. By color coding the individual branches based on the phenotype values possessed by the individuals they represent, one can see if there are patches of a certain color on neighboring branches, which would indicate that phenotype values cluster along with genetic similarity (e.g., using HyperTree v.1.0.0, http://www.kinase.com/tools/HyperTree.html).

The CEPH-HGDP Cell Line Panel. We used genotype data from the publicly available CEPH-HGDP Cell Line Panel [56], which have been investigated recently in numerous studies (e.g., reviewed in [4]). The datasets used here include 377 and 783 autosomal microsatellites typed on 1,040 people from 51 populations distributed worldwide (China and United States Han subjects were pooled). We included the same 1,040 subjects as originally described in Rosenberg et al. [1], with the exception of 16 duplicated or mislabeled samples [5]. In addition, we also used geographic data (i.e., the distance from Addis Ababa to each of the 51 CEPH-HGDP populations), kindly provided by Dr. François Balloux [8], and pairwise F_{ST} values [82] between 51 populations based on 783 microsatellites kindly provided by Dr. Noah Rosenberg [83].

The anthropometric data of Howells. We used craniometric diversity data (the median across the subjects of each of 45 features for each gender) gathered on 489 males and 459 females from ten populations (nine populations for females) made available through the work by Howells [54,55]. The craniometric data was paired according to geographic regions with genetic data from 415 subjects from 19 populations from the CEPH-HGDP panel genotyped on 783 markers as described in Table 1 of Roseman [59]. Pairwise F_{ST} between the ten populations (merged from an original 19 CEPH-HGDP populations to represent the locations sampled from, for the craniometric data) was calculated according to standard formulae [82] for diploid data using genotypes at 786 microsatellite loci from the CEPH-HGDP. The pairwise F_{ST} analysis produced a 10×10 genetic distance matrix that we used in the proposed GAMOVA procedure to determine if relationships exist between the 45 cranial measurements and genetic background similarity.

The HapMap data set. We downloaded the ~700,000 SNP markers from the phase I data available on the 209 individuals genotyped as part of the International HapMap Project (http://www.hapmap.org; [60]). These 209 individuals included 60 individuals of Northern European descent (i.e., the "CEPH-HGDP" derived individuals), 44 individuals of Japanese descent, 45 individuals of Han Chinese descent, and 60 individuals of West African descent (i.e., the "Yoruban" population-derived individuals). Since these 209 individuals had been phased (i.e., assigned haplotypes), we considered the data as providing both 209 multilocus genotypes on each of the 22 autosomes, as well as providing 418 individual chromosomes, from each of the four populations, and analyzed it in this light.

Power estimations. A Python computer program was used to generate four sets of two populations, each with M markers and N subjects with the same constant data size ($D = N \times M = 2^{20}$) as discussed by Patterson et al. [64]. Allele frequencies of all biallelic loci for the first population were generated by assuming they followed a beta-distribution with parameters 0.75 and 0.75. For the second population, for each locus, the allele frequencies of the first population were modified by adding random numbers so that the two populations would exhibit certain genetic distances based on Wright's F_{ST} measure of population differentiation ([82], Formula 5.12). For each of the four sets, 1,000 populations were simulated with F_{ST} values that ultimately were randomly distributed between 0 and 0.01. We assigned hypothetical individuals in the simulated samples alleles at each of the M loci based on the allele frequencies. A GAMOVA analysis was then performed on an IBS distance matrix constructed from the allelic profiles of the simulated individuals as described above with known population membership taken as a predictor variable. Permutations (1,000) of the data were performed to determine the significance of each pseudo-F statistic from the GAMOVA analysis.

Acknowledgments

The authors would like to thank Drs. Noah Rosenberg and François Balloux for providing insight into, and data associated with, the CEPH-HGDP Cell Line Panel. The authors would also like to thank Dr. Marti Anderson for advice and encouragement regarding the multivariate distance matrix regression method that forms the basis of the proposed GAMOVA procedure.

Author contributions. CMN and NJS conceived and designed the experiments, performed the experiments, analyzed the data, and wrote the paper. CMN, OL, and NJS contributed reagents/materials/analysis tools.

Funding. NJS and his laboratory are supported in part by the following research grants: the National Heart, Lung, and Blood Institute Family Blood Pressure Program (FBPP; U01 HL064777–06); the National Institute on Aging Longevity Consortium (U19 AG023122–01); the National Institute of Mental Health Consortium on the Genetics of Schizophrenia (COGS; 5 R01 HLMH065571–02); National Institutes of Health R01s HL074730–02 and HL070137–01; Scripps Genomic Medicine; and the Donald W. Reynolds Foundation (Helen Hobbs, Principal Investigator).

Competing interests. The authors have declared that no competing interests exist.

References

1. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
2. Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5: 598–609.
3. Serre D, Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. Genome Res 14: 1679–1685.
4. Cavalli-Sforza LL (2005) The Human Genome Diversity Project: Past, present and future. Nat Rev Genet 6: 333–340.
5. Mountain JL, Ramakrishnan U (2005) Impact of human population history on distributions of individual-level genetic distance. Hum Genomics 2: 4–19.
6. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102: 15942–15947.
7. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, et al. (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2: 81–89.
8. Liu H, Prugnolle F, Manica A, Balloux F (2006) A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79: 230–237.
9. Kittles RA, Chen W, Panguluri RK, Ahaghotu C, Jackson A, et al. (2002) CYP3A4-V and prostate cancer in African Americans: Causal or confounding association because of population stratification? Hum Genet 110: 553–560.
10. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598–604.
11. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, et al. (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36: 388–393.
12. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512–517.
13. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, et al. (2005) Demonstrating stratification in a European American population. Nat Genet 37: 868–872.
14. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37: 90–95.
15. Reiner AP, Ziv E, Lind DL, Nievergelt CM, Schork NJ, et al. (2005) Population structure, admixture, and aging-related phenotypes in African American adults: The cardiovascular health study. Am J Hum Genet 76: 463–477.
16. Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case-control association studies. PLoS Genet 1: e32. doi:10.1371/journal.pgen.0010032
17. Berger M, Stassen HH, Kohler K, Krane V, Monks D, et al. (2006) Hidden population substructures in an apparently homogeneous population bias association studies. Eur J Hum Genet 14: 236–244.
18. Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, et al. (2006) European population substructure: Clustering of northern and southern populations. PLoS Genet 2: e143. doi:10.1371/journal.pgen.0020143
19. Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of populations. Genetics 174: 875–891.
20. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
21. Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet Res 78: 59–77.
22. Satten GA, Flanders WD, 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.

23. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
24. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, et al. (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72: 1492–1504.
25. Stassen HH, Hoffman K, Scharfetter C (2003) Similarity by state/descent and genetic vector spaces: Analysis of a longitudinal family study. BMC Genet 4 Suppl 1: S59.
26. Kohler K, Bickeboller H (2005) Case-control association tests correcting for population stratification. Ann Hum Genet 69: 98–115.
27. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical and study design considerations. Genet Epidemiol 28: 289–301.
28. Tsai HJ, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard EG, et al. (2005) Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. Hum Genet 118: 424–433.
29. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
30. Redden DT, Divers J, Vaughan LK, Tiwari HK, Beasley TM, et al. (2006) Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. PLoS Genet 2: e137. doi:10.1371/journal.pgen.0020137
31. Rosenberg NA, Nordborg M (2006) A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed, or spatially distributed populations. Genetics 173: 1665–1678.
32. Setakis E, Stirnadel H, Balding DJ (2006) Logistic regression protects against population structure in genetic association studies. Genome Res 16: 290–296.
33. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
34. McKeigue PM (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 63: 241–251.
35. Wu B, Liu N, Zhao H (2006) PSMIX: An R package for population structure inference via maximum likelihood method. BMC Bioinformatics 7: 317.
36. Zhang S, Zhu X, Zhao H (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 24: 44–56.
37. Chen HS, Zhu X, Zhao H, Zhang S (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 67: 250–264.
38. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368: 455–457.
39. Mountain JL, Cavalli-Sforza LL (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 61: 705–718.
40. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1: 274–286.
41. Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. Hum Hered 43: 45–52.
42. Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. Evolution 43: 258–275.
43. Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res: 175–185.
44. Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152: 1753–1766.
45. Wang J (2002) An estimator for pairwise relatedness using molecular markers. Genetics 160: 1203–1215.
46. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402–1422.
47. Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, et al. (2004) Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet 114: 263–271.
48. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, et al. (2004) A high-density admixture map for disease gene discovery in African Americans. Am J Hum Genet 74: 1001–1013.
49. Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, et al. (2006) A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet 79: 640–649.
50. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131: 479–491.
51. Wessel J, Schork NJ (2006) Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 79: 792–806.
52. Zapala MA, Schork NJ (2006) Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci U S A 103: 19430–19435.
53. Smouse PE, Long JC (1992) Matrix correlation analysis in anthropology and genetics. American Journal of Physical Anthropology 35: 187–213.
54. Howells WW (1973) Cranial variation in man: A study of multivariate analysis of patterns of difference among recent human populations. Cambridge (Massachusetts): Harvard University Press, Peabody Museum of Archaeology and Ethnology. 259 p.
55. Howells WW (1989) Skull shapes and the map: craniometric analyses in the dispersion of modern Homo. Cambridge (Massachusetts): Harvard University Press, Peabody Museum of Archaeology and Ethnology. 189 p.
56. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.
57. Manica A, Prugnolle F, Balloux F (2005) Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet 118: 366–371.
58. Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72: 1171–1186.
59. Roseman CC (2004) Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. Proc Natl Acad Sci U S A 101: 12824–12829.
60. Roseman CC, Weaver TD (2004) Multivariate apportionment of global human craniometric diversity. Am J Phys Anthropol 125: 257–263.
61. Martinez-Abadias N, Gonzalez-Jose R, Gonzalez-Martin A, Van der Molen S, Talavera A, et al. (2006) Phenotypic evolution of human craniofacial morphology after admixture: a geometric morphometrics approach. Am J Phys Anthropol 129: 387–398.
62. Parra EJ, Kittles RA, Shriver MD (2004) Implications of correlations between skin color and genetic ancestry for biomedical research. Nat Genet 36: S54–S60.
63. Consortium TIH (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
64. Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190
65. Lewontin R (1972) The apportionment of human diversity. Evolutionary Biology 6: 381–398.
66. Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33 Suppl: 266–275.
67. Barbujani G, Goldstein DB (2004) Africans and Asians abroad: Genetic diversity in Europe. Annu Rev Genomics Hum Genet 5: 119–150.
68. Mountain JL, Risch N (2004) Assessing genetic contributions to phenotypic differences among "racial" and "ethnic" groups. Nat Genet 36: S48–S53.
69. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
70. Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for "race" and medicine. Nat Genet 36: S21–S27.
71. Turakulov R, Easteal S (2003) Number of SNPS loci needed to detect population structure. Hum Hered 55: 37–45.
72. Yang BZ, Zhao H, Kranzler HR, Gelernter J (2005) Characterization of a likelihood based method and effects of markers informativeness in evaluation of admixture and population group assignment. BMC Genet 6: 50.
73. Nievergelt C, Kelsoe J, Shimizu C, Burns J, Schork N. (2006) Multivariate distance based methods for testing and accommodating population substructure [abstract 2203]. Annual meeting of The American Society of Human Genetics, 12 October 2006, New Orleans, Louisiana. Available: http://www.ashg.org/genetics/ashg06s/index.shtml. Accessed 5 March 2007.
74. Van de Casteele T, Galbusera P, Matthysen E (2001) A comparison of microsatellite-based pairwise relatedness estimators. Mol Ecol 10: 1539–1549.
75. Widdig A, Nurnberg P, Krawczak M, Streich WJ, Bercovitch FB (2001) Paternal relatedness and age proximity regulate social relationships among adult female rhesus macaques. Proc Natl Acad Sci U S A 98: 13769–13773.
76. Parsons K, Durban J, Claridge D, Balcomb K, Noble L, et al. (2003) Kinship as a basis for alliance formation between male bottlenose dolphins, *Tursiops truncatus*, in the Bahamas. Anim Behav 66: 185–194.
77. Lobo JA, Quesada M, Stoner KE (2005) Effects of pollination by bats on the mating system of *Ceiba pentandra* (Bombacaceae) populations in two tropical life zones in Costa Rica. Am J Bot 92: 370–376.
78. McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: A comment on distance-based redundancy analysis. Ecology 82: 290–297.
79. Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–338.
80. Good P (1994) Permutation tests. New York: Springer-Verlag. 228 p.
81. Edington ES (1995) Randomization tests. New York: Marcel Dekker. 409 p.
82. Weir BS (1996) Genetic data analysis II: Methods for discrete population genetic data. Sunderland (Massachusetts): Sinauer Associates. 445 p.
83. Rosenberg NA (2007) Human genetic variation: Answers to frequently asked questions. In: Bamshad M, editor. Human genetic diversity: Implications for race, ancestry, and health. Oxford: Oxford University Press.