RESEARCH ARTICLE

The Equidistance Index of Population Structure

Yaron Granot*[1], Omri Tal[2], Saharon Rosset[3], and Karl Skorecki[1]

1 Rappaport Faculty of Medicine and Research Institute, Technion–Israel Institute of Technology, and Rambam Medical Center, Haifa, Israel **2** Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, 04103 Leipzig, Germany **3** School of Mathematical Sciences Tel Aviv University, Tel Aviv, Israel

*E-mail: yarongranot@hotmail.com

Abstract

Measures of population differentiation, such as F_{ST}, are traditionally derived from a partition of heterozygosities within and between populations. However, the emergence of population clusters from multilocus analysis is a function of genetic *structure* (departures from panmixia) rather than of diversity. If the populations are close to panmixia, slight differences between the mean pairwise distance within and between populations (low F_{ST}) can manifest as strong separation between the populations, thus population clusters are often evident even when the vast majority of diversity is partitioned within populations rather than between them. Moreover, because F_{ST} is also a function of internal diversity, it does not directly reflect the strength of separation between population clusters. For any given F_{ST} value, clusters can be tighter (more panmictic) or looser (more stratified), and in this respect higher F_{ST} does not always imply stronger differentiation. Finally, F_{ST} as a measure of structure or population distance is a 'supervised' measure, in the sense that target populations have to be predefined (samples labeled). In this study we propose a measure for the partition of structure, denoted E_{ST}, which is more consistent with results from clustering schemes. Crucially, our measure is based on a statistic of the data that is a good measure of internal structure, mimicking the information extracted by unsupervised clustering or dimensionality reduction schemes. To assess the utility of our metric, we ranked various human (HGDP) population pairs based on F_{ST} and E_{ST} and found substantial differences in ranking order. In some cases examined, most notably among isolated Amazonian tribes, E_{ST} ranking seems more consistent with demographic, phylogeographic and linguistic measures of classification compared to F_{ST}. Thus, E_{ST} may at times outperform F_{ST} in identifying evolutionarily significant differentiation.

Keywords: F_{ST}, population structure, panmixia, differentiation, standard deviation, genetic isolates

Introduction

Genetic differentiation among populations is typically derived from the ratio of within- to between-population diversity. The most commonly used metric, F_{ST}, was originally introduced as a fixation index at a single biallelic locus (Wright 1978), and subsequently adapted as a measure of population subdivision by averaging the values over many loci (Nei 1973; Weir and Cockerham 1984). F_{ST} can be expressed mathematically as $F_{ST}=1-S/T$, where S and T represent heterozygosity or some other measure of diversity in subpopulations and in the total population (Hudson et al. 1992). The validity of F_{ST} as a measure of differentiation has been brought into question, especially when gene diversity is high (e.g., in microsatellites), and various metrics, including G'_{ST} (Hedrick 2005) and Jost's *D* (Jost 2008), have been proposed to address this inadequacy (though see Whitlock 2011 for a counter-perspective).

Although these metrics vary considerably in their formulation, they all follow the same basic framework of partitioning genetic diversity into within- vs. between-group components. It has long been noted, however, that the apportionment of diversity (Lewontin 1972) does not directly reflect the strength of separation between populations, and the emergence of population clusters has been shown both empirically (Mitton 1977) and mathematically (Edwards 2003; Tal 2013) even when the vast majority of diversity is within rather than between populations. For example, humans sampled from across Europe (Nelis et al. 2009) and East Asia (Tian et al. 2008) form identifiable clusters with pairwise F_{ST} as low as 0.002, even though 99.8% of the variation is contained within populations and only 0.2% is between them. Clearly, these clusters reflect an aspect of population differentiation that is not directly captured by F_{ST}, yet there is currently no commonly used metric for partitioning structure into within- and between-population components in the same way that F_{ST} partitions diversity. Dimensionality reduction schemes such as principal component analysis (PCA) (Patterson et al. 2006) and clustering algorithms such as the widely used STRUCTURE (Rosenberg et al. 2002) are highly popular, however such programs are primarily used for visualization, and there is still value in summary statistics for quantifying complex datasets on a simple 0-1 scale.

Here we propose a novel statistic, denoted E_{ST}, based on a modified F_{ST} estimator in which the mean pairwise distance between individuals (a measure of diversity) is replaced by the standard deviation of pairwise distances (a measure of structure), thus extracting the excess structure in the total population compared to subpopulations. Conceptually, E_{ST} is formulated in three steps: 1. Population structure is defined in terms of departures from *panmixia*. 2 Panmixia is defined in terms of pairwise *equidistance* between individuals (a population is considered panmictic if all individuals are equally distant from each other). 3. Departures from equidistance are defined in terms of the *standard deviation* of pairwise distances. E_{ST} reflects the decrease in panmixia when subpopulations are pooled. The general formula is: $E_{ST}=1-SD_S/SD_T$, where SD_S and SD_T represent the standard deviations of pairwise distances in subpopulations and in the total population. While F_{ST} is weighed down by high *diversity* within populations, E_{ST} is weighed down by high *structure* within populations. Since diversity is usually greater than structure, E_{ST} is usually greater than F_{ST}.

The core insight here is that the asymptotic (in terms of number of SNP loci considered) standard deviation of pairwise genetic distances is a good "unsupervised" measure of internal structure, a statistic that mimics the information extracted by dimensionality reduction and clustering schemes, thus justifiable as a basis for the definition of E_{ST}. In particular, in Appendix A we prove that this asymptotic standard deviation is zero if and only if there is no internal structure (i.e., the population is panmictic).

Results and Discussion

Partitioning Diversity vs. Partitioning Structure

The difference between the partitions of diversity and structure within and between two populations from the human genome diversity project (HGDP) (Cann et al. 2002) is illustrated in Figure 1. The mean distance among mixed Russian-Chinese pairs is only marginally (~10%) higher than among Russian-Russian pairs (Figure 1*A*), reflecting the relatively low F_{ST}. However this translates to a far greater increase in total structure compared to the low structure within each population (Figure 1*B*), reflecting the much higher E_{ST}.

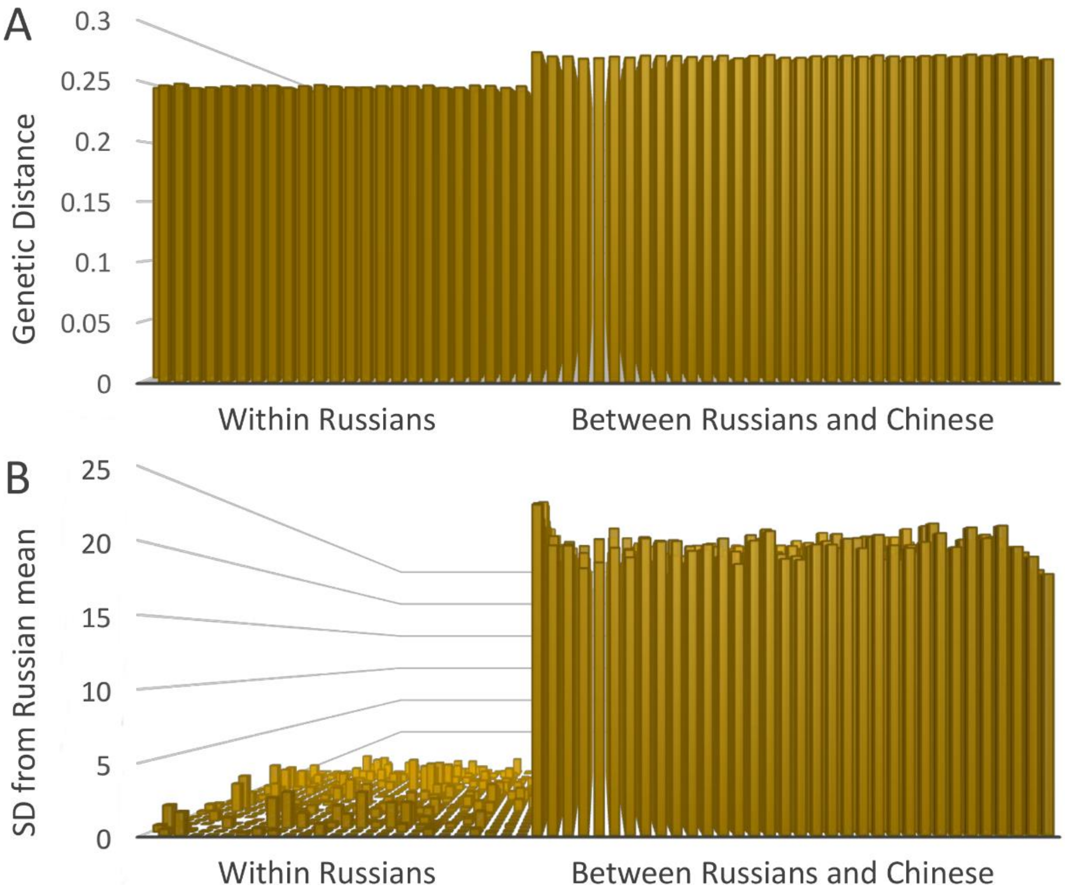

Figure 1. The mean and standard deviation of pairwise distances. Each column represents the genetic distance between a pair of individuals. (A) Partitioning diversity (F_{ST}): pairwise distances within Russians (n=25) and between Russians and Chinese (n=34). The Russians are only marginally closer to each other than to the Chinese. (B) Partitioning structure (E_{ST}): by starting the plot at the mean within-Russian distance (0.246) and defining structure in terms of SD from the Russian mean, the magnitude of population separation becomes more evident.

We compared F_{ST}, E_{ST}, and clustering among Russian and Chinese samples, with an increasing amount of single nucleotide polymorphisms (SNPs) ranging from 10 to 660,755 (Figure 2). Using multidimensional scaling (MDS), the two population clusters gradually diverge as SNP count increases, with no corresponding increase in F_{ST}. At the same time we observe a steady increase in E_{ST} directly corresponding to the emerging clusters, indicating that the Russian and Chinese HGDP samples are close to panmixia. With few SNPs this is obfuscated by the variance of the genetic distance measure, hence E_{ST} is relatively small. The actual levels of panmixia become increasingly evident as more SNPs are added, thus revealing the population clusters (Edwards 2003). However this process does not proceed indefinitely; the finite number of pairwise differences among humans (~3 million SNPs) sets an upper limit to the number of available markers, and the amount of extractable information is further reduced by linkage disequilibrium. In our data the increase in E_{ST} as a function of marker count reaches a plateau above 100,000 SNPs (Figure S1). Although this upper bound can vary across different datasets and types of markers, it suggests that resolution may not improve substantially with further increases in marker count. Thus, these clusters can be considered close approximations of the "true" strength of separation among these populations. For this reason, E_{ST} estimates should include as many markers as possible, although fewer markers can be used and the terminal E_{ST} can be extrapolated.

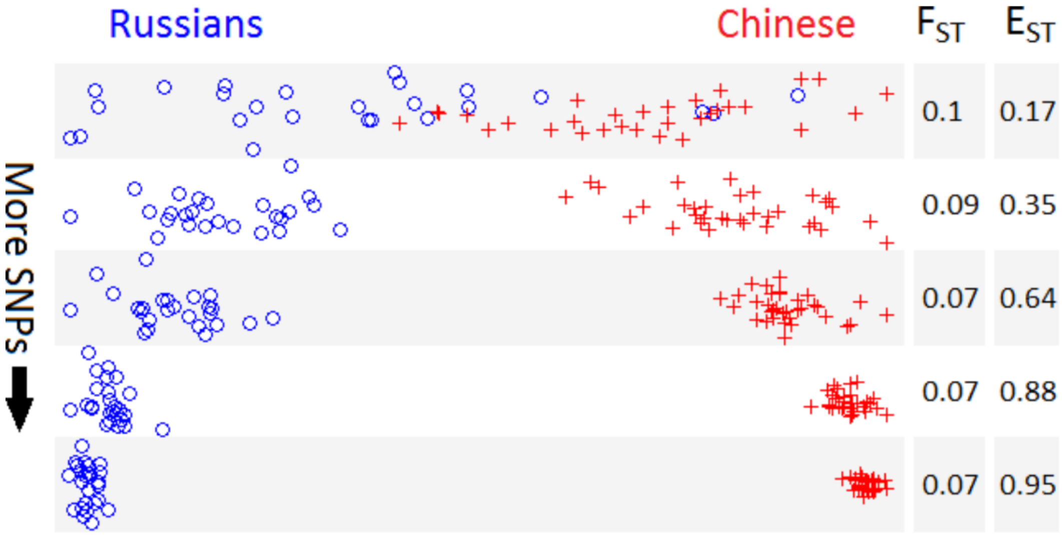

Figure 2. F_{ST} and E_{ST} vs. Clustering with increasing SNP count. Multidimensional scaling (MDS) plots with Russian (n=25) and Chinese (n=34) samples with increasing SNP count from top to bottom (10, 100, 1000, 10,000, and 660,755 SNPs). Two clusters gradually emerge as SNP count increases, along with an increase in E_{ST}, while F_{ST} remains relatively constant.

In order to determine whether or not E_{ST} adds insight to the analysis of population structure, we sought to compare the rank order of population differentiation using F_{ST} and E_{ST}. Pairwise F_{ST} and E_{ST} values from various HGDP populations are given in Table 1 (see Table S1 and Figure S2 for additional comparisons). As expected, $E_{ST} > F_{ST}$ in most population pairs. Only the Colombian-Maya pair has a slightly lower E_{ST} than F_{ST}, due to a combination of relatively low differentiation and high levels of intra-population structure. According to the HGDP browser (http://spsmart.cesga.es/search.php?dataSet=ceph_stanford, the Colombians (n=7) are the only HGDP population sample where two different tribes (Piapoco and Curripaco) were combined, which can help explain the high level of structure observed in this particular population (see Table S1, Figure S10, and Materials and Methods for further analysis of E_{ST} range).

Table 1 Pairwise F_{ST} (above diagonal) and E_{ST} (below diagonal) in 5 New World and 5 Old World HGDP populations

	Surui	Karitiana	Colombian	Maya	Pima	Yakut	Mongola	Russian	Bantu	San
Surui		0.13	0.1	0.09	0.12	0.15	0.15	0.17	0.23	0.3
Karitiana	0.58		0.08	0.07	0.11	0.13	0.13	0.16	0.22	0.29
Colombian	0.51	0.57		0.03	0.06	0.09	0.09	0.12	0.18	0.25
Maya	0.52	0.63	0.02		0.04	0.07	0.06	0.08	0.15	0.21
Pima	0.57	0.63	0.37	0.43		0.1	0.09	0.12	0.19	0.25
Yakut	0.74	0.8	0.6	0.69	0.74		0.01	0.06	0.13	0.19
Mongola	0.81	0.87	0.69	0.8	0.83	0.46		0.06	0.12	0.19
Russian	0.82	0.87	0.74	0.83	0.84	0.86	0.9		0.11	0.17
Bantu	0.88	0.92	0.85	0.91	0.91	0.93	0.94	0.95		0.07
San	0.92	0.95	0.89	0.95	0.94	0.96	0.97	0.98	0.89	

Amazonians vs. Global Populations

The Surui and Karitiana have an unusually high pairwise F_{ST}. In fact, the Karitiana are as diverged from the neighboring Surui in terms of F_{ST} as they are from the Mongola on the other side of the world (Table 1, Figure 3, and Figure S6). Moreover, F_{ST} actually decreases initially with distance from the Amazon, from 0.13 between the two Amazonian tribes, to 0.08-0.1 between Amazonians and Colombians, further decreasing to 0.07-0.09 between Amazonians and the more distant Maya. Remarkably, the highest F_{ST} among all HGDP Native American populations is between the two geographically closest populations, the Surui and Karitiana. These apparent anomalies can be explained by the inflation of F_{ST} in genetic isolates. F_{ST} between pairs of isolates can be nearly twice as high as between either one of the isolates and a more cosmopolitan population, as pairwise F_{ST} reflects the *combined* isolation of both populations. Since the Surui and Karitiana are both isolated, their pairwise F_{ST} is nearly double that between any one of them and a larger, less isolated population such as the Maya. In other words, the Maya's contribution to the pairwise F_{ST} is dwarfed by that of the Amazonians.

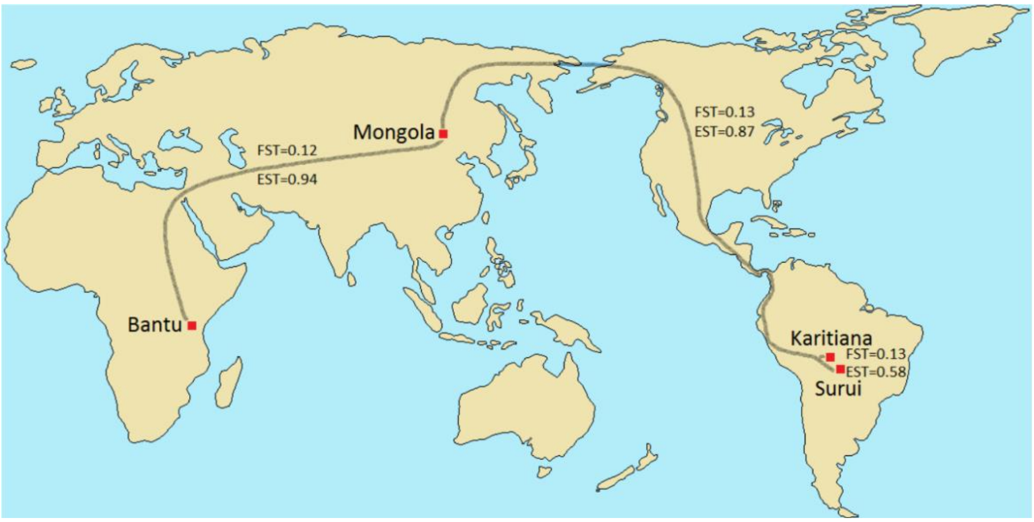

Figure 3. Geographic distance vs. F_{ST} and E_{ST} in various populations. In terms of F_{ST}, the Karitiana are roughly as diverged from the nearby Surui (F_{ST}=0.13) as they are from the Mongola on the other side of the world (F_{ST}=0.13)or as the Bantu are from the Mongola (F_{ST}=0.12). In terms of E_{ST}, differentiation is far greater among these global populations (E_{ST}≈0.9) than between the neighboring Amazonian tribes (E_{ST}≈0.6).

Differentiation based on E_{ST} (Surui-Karitiana=0.58, Karitiana-Mongola=0.87, and Mongola-Bantu=0.94) seems more consistent with the geographic distances among these populations (Figure 3). It should be noted that the Surui-Karitiana E_{ST} might be somewhat underestimated due to cryptic sampling of close relatives (Rosenberg 2006), however the wide range of heterozygosity values (which are less sensitive to the sampling of close relatives) and the elevated structure across all Native American HGDP populations (Figures S3-S5) suggest that this is not merely a sampling artifact. In some cases E_{ST} also decreases with distance from the Amazon (Table 1), however this decrease is more moderate than the decrease in F_{ST} (Figure S6).

Neighbor-joining trees of individual similarities (Jorde and Wooding 2004) are a convenient tool for representing multidimensional genetic data on a two-dimensional plane, while simultaneously displaying distances within and between populations. Two pairs of such trees, for Surui-Karitiana and Yoruba-Russians, are given in Figure 4, and we can see that in both cases distances are greater between individuals (black branches) than between populations (red branches) (Figure 4*A*).

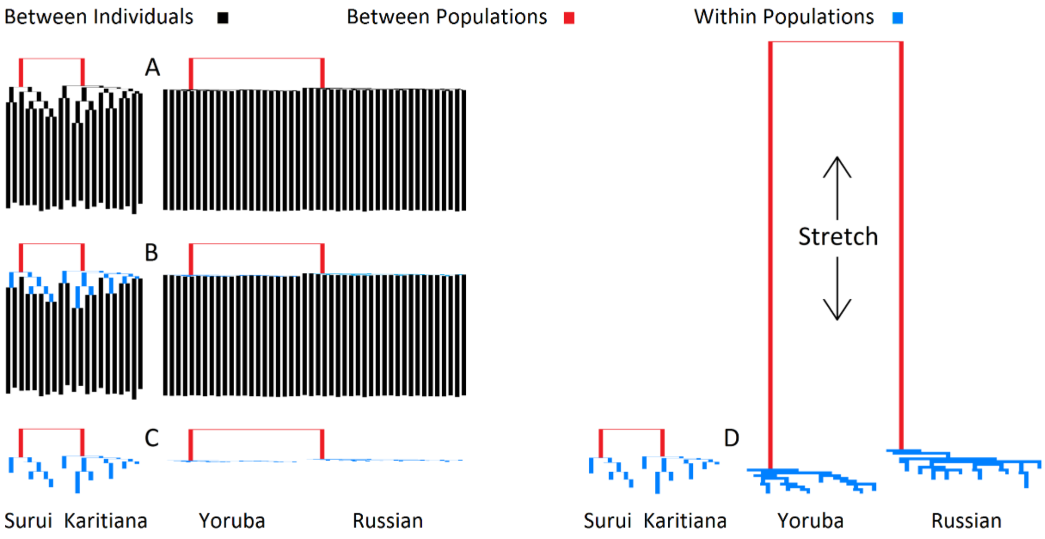

Figure 4. Surui-Karitiana vs. Yoruba-Russian NJ trees of individual similarities. (A) Diversity is apportioned into individual (black) and population (red) components. (B) A third component, structure within populations (blue), is added. (C) The individual component is removed. (D) The Yoruba-Russian tree is stretched to roughly match the level of structure within the Surui-Karitiana tree.

The ratio of within- to between-population distance is roughly equivalent in the two population pairs, however the Yoruba-Russian tree is significantly *flatter*, indicating greater panmixia within these two populations (Figures S7-S8). Adding a third dimension of intra-population structure (blue branches) highlights this discrepancy (Figure 4*B*), which is further accentuated by removing the inter-individual component (Figure 4*C*) and stretching the Yoruba-Russian tree to match the level of structure observed in the Surui-Karitiana tree (Figure 4*D*). At first glance the Amazonian tribes, with their long population branches, appear to be as differentiated as the Yoruba are from the Russians. Upon closer inspection, however, the Yoruba and Russians appear more strongly diverged. The Amazonian tribes are highly structured not only between them, but also within them, resulting in distant, but loosely separated clusters. This aspect of population structure is not captured by F_{ST}, which is actually slightly higher between the Surui and Karitiana (0.13) than between Yoruba and Russians (0.12), but is revealed by the higher E_{ST} between Yoruba and Russians (0.97) compared to the Surui and Karitiana (0.58).

E_{ST} and the Dissimilarity Fraction

The dissimilarity fraction, ω, is defined (Witherspoon et al. 2007) as the probability that individuals are genetically more similar to members of a different population than to members of their own population. For pairs of populations, this probability should have a 0-0.5 range, with ω=0 indicating that individuals are always closer to members of their own population and ω=0.5 indicating that individuals are just as likely to be closer to members of the other population as to members of their own population. Witherspoon et al. reported that that when many thousands of loci are analyzed, individuals from "geographically separated populations" are never closer to each other than to members of their own populations. The definition of "geographically separated" is, of course, open to interpretation. We found no overlap (ω=0) between the Adygei and Uygur HGDP samples, but some overlap (ω > 0) between Mayans and Surui, despite a 4x higher F_{ST} (Figure 5). Thus, F_{ST} and the dissimilarity fraction (ω) are not necessarily congruent. The E_{ST} values for these two population pairs are more consistent with ω, showing strong separation between the Adygei and Uygur (0.79) and more moderate separation between Colombians and Maya (0.52) (see Figure S9 for a more detailed plot).

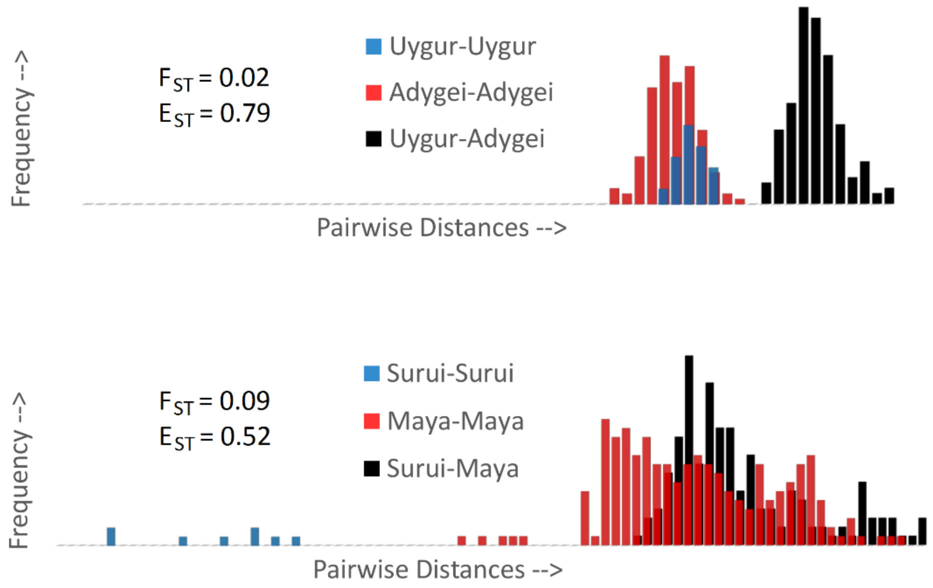

Figure 5. F_{ST} vs. genetic similarity in various population pairs. Pairwise distances are colored red or blue within populations and black between populations. (A) Even at a relatively low F_{ST} of 0.02 all within-population pairs among the Uygur and Adygei samples are genetically more similar than all the between-population pairs. (B) Separation is more ambiguous among Native Americans. Despite a relatively high F_{ST} of 0.09, there is substantial overlap between Maya-Maya (red) and Maya-Surui (black) samples. E_{ST} values are more consistent with the within- vs.-between population overlap and the dissimilarity fraction (ω).

Summary and Conclusions

The main distinction between F_{ST} and E_{ST} is that F_{ST} partitions diversity, whereas E_{ST} partitions structure within and between populations. F_{ST} is more sensitive to effective population size, while E_{ST} is more sensitive to outliers, though this is largely mitigated by using E_{ST}median rather than E_{ST}mean (see Materials and Methods). F_{ST} is often weighed down by high levels of intrapopulation diversity and can be close to zero even when population clusters are completely separated. This is not a flaw in F_{ST}, but it does demonstrate a conceptual disconnect between F_{ST} and clustering. Sewall Wright proposed a series of arbitrary F_{ST} thresholds ranging from 0.05 to 0.25, denoting little to very great differentiation (Wright 1978), however these are only broad guidelines, and the highest ranking of "very great differentiation" leaves most of the range (0.25-1) undefined.

Given its wider empirical range and more direct correlation with clustering and classification (Figure 2), phylogeography (Figure 3), and the dissimilarity fraction (Figure 5), such arbitrary thresholds may not be necessary for E_{ST}. E_{ST}>0.5 simply indicates that most of the structure is between populations rather than within them, corresponding to moderately separated populations such as Russians and Adygei (E_{ST}=0.5), Bantu from South Africa and Kenya (E_{ST}=0.48), or French and Sardinians (E_{ST}=0.48) (Table S1). E_{ST}<<0.5 indicates weak differentiation and E_{ST}>>0.5 indicates strong differentiation. E_{BT} is similar in many ways to E_{ST}, though its HGDP ranking order is often intermediate between F_{ST} and E_{ST} (Table S1). Interestingly, some East Asians populations have relatively low E_{BT}, such as Cambodians vs. Mongola (E_{BT}=0.13) and Japanese vs. Chinese (E_{BT}=0.16).

Differentiation metrics are judged by their ability to quantify meaningful evolutionary divergence, and can be indispensable in identifying *Evolutionarily Significant Units* (ESU) and *Distinct Population Segments* (DPS) for conservation (Waples 1991). For example given several subpopulations within a species, it is reasonable to prioritize the most highly differentiated subpopulation for conservation in order to maximize biodiversity. However, higher F_{ST} does not necessarily reflect stronger separation and lower misclassification, as with the Uygur and Adygei, whose clusters are better defined than those of the Surui and Maya despite a fourfold lower F_{ST} (Figure 5). In this context humans can be a useful model species simply because we know so much about human populations due to our “long habit of observing ourselves” (Darwin 1871). This allows us to make educated inferences about human populations that might otherwise be overlooked, e.g., we can be skeptical of the high Surui-Karitiana F_{ST}, and realize that this is most likely due to the relatively recent isolation of two small tribes. This is a luxury that we do not usually have with other species, in which case high F_{ST} can be misinterpreted as a deep phylogenetic divide, potentially leading to misguided conservation strategies. Our hope is that by combining information from both *fixation* (F_{ST}) and *equidistance* (E_{ST}) indices, researchers could make more informed decisions.

Unlike F_{ST}, which can be estimated from a handful of markers, E_{ST} requires large datasets with thousands of markers, which were unavailable to previous generations of population geneticists. With the latest SNP chips containing well over 100,000 markers, accurate estimates of departures from panmixia are finally within reach, and there is no longer a need for the simplifying assumption that subpopulations are effectively panmictic. By deriving an F_{ST}–type statistic for apportioning structure within and between populations, namely E_{ST}, we hope to add a new useful metric to the 21st century population genetics toolkit.

Materials and Methods

The HGDP data used in our analysis are available at: http://www.hagsc.org/hgdp/files.html. After removing the 163 mitochondrial SNPs and 105 samples previously inferred to be close relatives (Rosenberg 2006), the final file included 660,755 SNPs from 938 samples in 53 populations. Strings of SNPs were treated as sequences, with mismatches summed and divided by the sequence length. Pairwise distances, based on Allele Sharing Distance (ASD) (Gao and Martin 2009), were calculated as one minus half the average number of shared alleles per locus.

We used Hudson’s F_{ST} estimator (Hudson et al. 1992):

$$F_{ST}=1-S/T \qquad (1)$$

Where S and T are the mean pairwise distances within subpopulations and in the total pooled population.

The general equation for E_{ST} is:

$$E_{ST}=1-SD_S/SD_T \qquad (2)$$

Where SD_S and SD_T are the standard deviations (SD) of pairwise distances within subpopulations and in the total population. This E_{ST} estimator is referred to as E_{ST}mean. We used three additional E_{ST} estimators: E_{ST}min, E_{ST}median, and E_{ST}max (Figure S10). All four estimators use the same basic formula, with only the type of SD_S differing among estimators. In E_{ST}min, E_{ST}median, and E_{ST}max, SD_S is respectively replaced with the smallest, median, and largest *individual SD*, where the individual SD is the standard deviation of pairwise distances between a single sample and all other samples in the population. E_{ST}min uses the smallest individual SD_S from each population, i.e., the SD of the most panmictic sample, E_{ST}median uses the median individual SD_S, and E_{ST}max uses the highest individual SD_S. Each of these metrics has different sensitivities to various sampling biases. Due to E_{ST}mean's sensitivity to the sampling of close relatives, we used E_{ST}median (which is unaffected by the inclusion of relatives as long as at least 50% of the samples are unrelated) as the primary measure of E_{ST} in this study. In the rare event that >50% of the samples are closely related, E_{ST}max may be preferable, as long as at last one individual has no close relatives among the samples. E_{ST} values, especially E_{ST}min and E_{ST}mean, can be negative if structure is high and differentiation is low (Figure S10). Small sample sizes were often sufficient for estimating heterozygosity (Figure S11) and F_{ST} and E_{ST} (Figure S12) using all the SNPs in the HGDP dataset.

We derived an additional equidistance index, denoted E_{BT}, which is less sensitive to intra-population structure and the inclusion of relatives. Recall that E_{ST} reflects equidistance (E) within subpopulations (S) compared to the total (T) population. Similarly, E_{BT} reflects equidistance (E) between subpopulations (B) compared to the total (T) population:

$$E_{BT}=1-SD_B/SD_T \qquad (3)$$

Where SD_B and SD_T are the standard deviations of pairwise distances between individuals from different subpopulations, and in the total pooled population. In most cases $SD_T \geq SD_B$, because SD_T includes pairs of individuals from the same population as well as pairs from different populations, whereas SD_B only includes pairs of individuals from different populations. Pairs of individuals from the same population are likely to have a higher SD due to relatives in the samples, which disrupt the panmixia (see Naxi population in Figures S3-S5). Panmictic populations are not just equidistant among themselves, they are also equidistant towards each other. Such populations should have similar SD_S and SD_B, and thus similar E_{ST} and E_{BT}. All F_{ST}, E_{ST} and E_{BT} estimates in this study are based on pairwise comparisons between two populations or population groups. Each of the two paired populations was given equal weight, as were the within- and between-population pairs. Thus, 25% of the total weight was given to each population, and 50% to between-population pairs.

We developed a custom MATLAB code for extracting genetic distances from SNP data and estimating heterozygosity, pairwise distances, F_{ST}, E_{ST}, and E_{BT}. The code corrects for missing data and small sample sizes, and identifies outliers, but includes no further assumptions or corrections. Phylogenetic trees and MDS plots were also generated with MATLAB. Equal angle and square neighbor-joining trees of individual similarities were generated from matrices of pairwise distances with the *seqneighjoin* command. An alternative script, based on the internal MATLAB *seqpdist* command for sequence distance, yielded similar results.

Acknowledgments

We thank Alan Templeton for helpful advice and Sagi Abelson for help with the MATLAB script.

Appendix A – *The standard deviation of pairwise distances as* a measure of *population structure*

Our goal in this appendix is to substantiate the *asymptotic* (in terms of number of SNP loci) *standard deviation* of pairwise genetic distances as a good “unsupervised” measure of internal structure, thus justifiable as a basis for the definition of E_{ST}. In particular, we prove that this asymptotic standard deviation is zero if and only if there is no internal structure (i.e., the population is panmictic).

A model of pairwise genetic distances for genotypes from two diploid populations

Let p_i denote the frequency at locus *i* of allele ‘*A*’ in population 1, and let and q_i denote the frequency of the same allele in population 2 and assume that both populations are effectively very large and have the same contribution to the total population. The commonly-used *allele sharing distance* (ASD) measures the dissimilarity of two individual genotypes. For *diploid* genotypes, it is commonly defined as 2 minus the number of shared alleles at each locus, averaged across loci (Nakamura et al., 2005; Gao and Martin, 2009). For multiple loci genotypes we use a normalized (by the number of considered loci) version of ASD to simplify the analysis of means and variances of the ASD distribution, as in Tal (2013). Under the assumption of Hardy-Weinberg Equilibrium, allele frequencies fully determine per-locus genotype frequencies.

Let a categorical random variable X_i represent the ASD at diploid locus *i*, and let D_n represent the normalized ASD across *n* loci for pairs of genotypes sampled from the *total* population,

$$D_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We are interested in arriving at an expression for the variance (and ultimately the asymptotic standard variation) of D_n. Under the standard assumption of *linkage equilibrium* (LE) within each of our two populations, the X_i for the *total-population* pairs are *not* statistically independent, and therefore the formulation for the variance of D_n requires a partition into conditional expectations. From basic principles,

$$Var[D_n] = E[D_n^2] - E[D_n]^2$$

Now to evaluate $E[D_n^2]$ we need to condition it upon classification of pairs of genotypes as within- or between-population,

$$\begin{aligned}
E[D_n^2] &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n^2} E\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i\neq j}^{n}(X_i \cdot X_j)\right] = \frac{1}{n^2}\left(\sum_{i=1}^{n} E[X_i^2] + 2\sum_{i\neq j}^{n} E[X_i \cdot X_j]\right) \\
&= \frac{1}{4}\cdot\frac{1}{n^2}\left(\sum_{i=1}^{n} E[X_i^2] + 2\sum_{i\neq j}^{n} E[X_i]\cdot E[X_j]\right)\Bigg|\, both\ genotypes\ \ from\ pop\, 1 \\
&+ \frac{1}{4}\cdot\frac{1}{n^2}\left(\sum_{i=1}^{n} E[X_i^2] + 2\sum_{i\neq j}^{n} E[X_i]\cdot E[X_j]\right)\Bigg|\, both\ genotypes\ \ from\ pop\ 2 \qquad (1) \\
&+ \frac{1}{2}\cdot\frac{1}{n^2}\left(\sum_{i=1}^{n} E[X_i^2] + 2\sum_{i\neq j}^{n} E[X_i]\cdot E[X_j]\right)\Bigg|\, one\ genotype\ \ from\ pop\, 1\, and\ one\ \ from\ pop\ 2
\end{aligned}$$

where $E[X_i X_j] = E[X_i]\cdot E[X_j]$ since there is independence across any two loci for the within pairs and between pairs, and where the probabilities (assuming equal population sizes) for within-population 1 pairs, within-population 2 pairs, and between-population pairs are *at infinite population size* ¼, ¼, ½ respectively (otherwise, for finite population sizes *m* we have probabilities $\frac{m-1}{4m-2}, \frac{m-1}{4m-2}, \frac{m}{2m-1}$ respectively).

Now, from per-locus probabilities in Tal (2013, eq. 3 and Table 1) we derive the expected values,

$$\begin{aligned}
&E[X_i]\,\big|\, both\ genotypes\ \ from\ pop\, 1 = 4(-p_i^4 + 2p_i^3 - 2p_i^2 + p_i) \\
&E[X_i^2]\,\big|\, both\ genotypes\ \ from\ pop\, 1 = 4p_i(1 - p_i) \\
&E[X_i]\,\big|\, both\ genotypes\ \ from\ pop\ 2 = 4(-q_i^4 + 2q_i^3 - 2q_i^2 + q_i) \qquad (2) \\
&E[X_i^2]\,\big|\, both\ genotypes\ \ from\ pop\ 2 = 4q_i(1 - q_i) \\
&E[X_i]\,\big|\, one\ genotype\ \ from\ pop\, 1\, and\ one\ \ from\ pop\ 2 = 2(2p_i q_i^2 + 2q_i p_i^2 - 2p_i^2 q_i^2 - 4p_i q_i + p_i + q_i) \\
&E[X_i^2]\,\big|\, one\ genotype\ \ from\ pop\, 1\, and\ one\ \ from\ pop\ 2 = 2(p_i^2 + q_i^2 - 4p_i q_i + p_i + q_i)
\end{aligned}$$

So that,

$$
\begin{aligned}
E[D_n^2] = \frac{1}{n^2}\Bigg[& \sum_{i=1}^{n} p_i(1-p_i) + 8\sum_{i \neq j}^{n}(-p_i^4 + 2p_i^3 - 2p_i^2 + p_i)(-p_j^4 + 2p_j^3 - 2p_j^2 + p_j) \\
& + \sum_{i=1}^{n} q_i(1-q_i) + 8\sum_{i \neq j}^{n}(-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)(-q_j^4 + 2q_j^3 - 2q_j^2 + q_j) \\
& + 2\sum_{i=1}^{n}(p_i^2 + q_i^2 - 4p_iq_i + p_i + q_i) \\
& + 8\sum_{i \neq j}^{n}(2p_iq_i^2 + 2q_ip_i^2 - 2p_i^2q_i^2 - 4p_iq_i + p_i + q_i)(2p_jq_j^2 + 2q_jp_j^2 - 2p_j^2q_j^2 - 4p_jq_j + p_j + q_j)\Bigg]
\end{aligned} \tag{3}
$$

Also,

$$
E[D_n]^2 = \left(E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} E[X_i]\right)^2
$$

From Tal (2013, section 3.2) we have the expression for X_i and thus for E[X_i] such that,

$$
E[D_n]^2 = \frac{1}{16n^2}\left(\sum_{i=1}^{n}(p_i + q_i)(2 - p_i - q_i)(4 - (p_i + q_i)(2 - p_i - q_i))\right)^2
$$

So that finally,

$$
\begin{aligned}
& Var[D_n] = E[D_n^2] - E[D_n]^2 \\
& = \frac{1}{n^2}\Bigg[\sum_{i=1}^{n} p_i(1-p_i) + 8\sum_{i \neq j}^{n}(-p_i^4 + 2p_i^3 - 2p_i^2 + p_i)(-p_j^4 + 2p_j^3 - 2p_j^2 + p_j) \\
& + \sum_{i=1}^{n} q_i(1-q_i) + 8\sum_{i \neq j}^{n}(-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)(-q_j^4 + 2q_j^3 - 2q_j^2 + q_j) \\
& + 2\sum_{i=1}^{n}(p_i^2 + q_i^2 - 4p_iq_i + p_i + q_i) \\
& + 8\sum_{i \neq j}^{n}(2p_iq_i^2 + 2q_ip_i^2 - 2p_i^2q_i^2 - 4p_iq_i + p_i + q_i)(2p_jq_j^2 + 2q_jp_j^2 - 2p_j^2q_j^2 - 4p_jq_j + p_j + q_j) \\
& - \frac{1}{16}\left(\sum_{i=1}^{n}(p_i + q_i)(2 - p_i - q_i)(4 - (p_i + q_i)(2 - p_i - q_i))\right)^2\Bigg]
\end{aligned} \tag{4}
$$

Thus we have an explicit formulation for the variance of the pairwise distance distribution of genotypes from two panmictic populations in terms of the allele frequencies across a given number of loci, *n*.

Crucially, we would like to prove that at the limit, the pairwise distance variance is asymptotically above zero *if and only if* the population has internal structure; i.e., if under our model with any F_{ST}>0,

$$S = lim_{n\to\infty} Var[D_n] > 0$$

We will proceed by deriving an explicit expression for *S*. Consider an *equivalent setting* comprised of three random variables *W*,*Y* and *Z*, which represent the pairwise distances of genotypes within population 1, within population 2 and between populations 1 and 2, respectively. We sample *n* values X_i from just *one* of these distributions, by first flipping a 3-sided coin to decide from which: with a probability α for *W*, a probability *β* for *Y* and a probability γ for *Z*. Once the distribution was selected, the sampling of X_i is done *i.i.d*. Note that due to the randomized choice of the distribution from which to sample *all* the X_i, they are identically distributed but *not* independent. Now we set,

$$D_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We would like to get an expression for *S* in terms of the expectations of *W*, *Y*, *Z* and α, *β*, γ, where,

$$S = lim_{n\to\infty} Var(D_n)$$

From the law of total variance,

$$Var(D_n) = Var(E[S_n \mid B]) + E[Var(S_n \mid B)] = Var(U_{XYZ}) + \frac{1}{n} C \cdot Var(S_n)$$

where *B* here is a categorical random variable that describes from which of distributions *W*, *Y*, *Z* we are sampling from, with probabilities α, *β*, γ respectively, and where U_{XYZ} is a discrete random variable taking the values of μ_W=*E*[*W*], μ_Y=*E*[*Y*], μ_Z=*E*[*Z*] with corresponding probabilities α, *β*, γ respectively. Hence at the limit *n*→∞ we have,

$$S = Var(U_{XYZ}) = \alpha(\mu_W - \mu)^2 + \beta(\mu_Y - \mu)^2 + \gamma(\mu_Z - \mu)^2 \qquad (5)$$

$$\mu = \alpha\mu_W + \beta\mu_Y + \gamma\mu_Z$$

and *S*=0 *if and only if* the three means are equal, i.e., $\mu_W=\mu_Y=\mu_Z$.

Now consider three *sequences* of random variables W_i, Y_i, Z_i, *i*:1...*n*, instead of the three single random variables, and sample *n* values from one of these sequences (again according to the prior probabilities α, *β*, γ). Once the sequence is selected, these samples are independent but now *not identically distributed*. We would again like to find *S*, and more importantly, the condition for which it is zero, this time in terms of *E*[W_i], *E*[Y_i], *E*[Z_i] (and the prior probabilities). Sampling from a sequence with fixed probabilities just defines a new mixture distribution -- so the problem gets reduced to the one already solved. Therefore U_{XYZ} is now defined by the three limits (since we have derived *S* in Eq. (5) at the limit *n*→∞),

$$\mu_W = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[W_i],\ \mu_Y = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[Y_i],\ \mu_Z = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[Z_i] \qquad (6)$$

with probabilities α, *β*, γ.

Crucially, this sampling scenario corresponds to our original setting of formulating the variance of the genetic distance of genotypes sampled from the total population, given the sequencing of an infinite number of loci, where μ_W and μ_Y in Eq. (6) represent the two within-population pairwise distance means and μ_Z the total-population mean (derived below), and where the respective probabilities are as in Eq. (1), α=¼, *β*=¼, γ= ½, assuming *infinite population size*. Again, *S*=0 *if and only if* these means are equal, i.e., μ_W=μ_Y=μ_Z. Let us analyze the conditions for these equalities, given the corresponding formulations of the pairwise distance means. First, using the additivity of expectations,

$$E[D_n] = \frac{1}{n}\sum_{i=1}^{n}E[X_i]$$

we get from Eq. (2) the expressions for any *finite n*,

$$\begin{aligned}
&E[D_n]\,|\,both\ genotypes\ \ from\ pop\,1 = \frac{1}{n}\sum_{i=1}^{n}4(-p_i^4+2p_i^3-2p_i^2+p_i)\\
&E[D_n]\,|\,both\ genotypes\ \ from\ pop\ 2 = \frac{1}{n}\sum_{i=1}^{n}4(-q_i^4+2q_i^3-2q_i^2+q_i) \qquad (7)\\
&E[D_n]\,|\,one\ genotype\ \ from\ pop\,1\,and\ one\ \ from\ pop\ 2\\
&\quad = \frac{1}{n}\sum_{i=1}^{n}2(2p_iq_i^2+2q_ip_i^2-2p_i^2q_i^2-4p_iq_i+p_i+q_i)
\end{aligned}$$

bearing in mind the analysis pertains to *E*[D_n] as *n*→∞. We proceed to examine what can be concluded from the equalities μ_W=μ_Y=μ_Z (the only case where *S*=0) given the means in Eq. (7), about the allele frequencies p_i and q_i for any finite *n* (and this also holds at *n*→∞). Thus we start by explicitely writing the eqaulities (where the 1/*n* canceles out),

$$\begin{aligned}\sum_{i=1}^{n}(-p_i^4+2p_i^3-2p_i^2+p_i)&=\sum_{i=1}^{n}(-q_i^4+2q_i^3-2q_i^2+q_i)\\ \sum_{i=1}^{n}(-p_i^4+2p_i^3-2p_i^2+p_i)&=\sum_{i=1}^{n}(2p_iq_i^2+2q_ip_i^2-2p_i^2q_i^2-4p_iq_i+p_i+q_i)\end{aligned}\qquad(8)$$

To proceed we substitute new variables,

$$x_i=p_i(1-p_i)$$

$$y_i=q_i(1-q_i)$$

Then, the 1st equation in (8) becomes,

$$\sum x_i(1-x_i)=\sum y_i(1-y_i)$$

such that,

$$\sum[x_i(1-x_i)+y_i(1-y_i)]=\sum 2x_i(1-x_i)$$

and using the 2nd equation in (8),

$$=\sum[2p_iq_i(q_i+p_i-p_iq_i-1)+p_i+q_i-2p_iq_i]$$

$$=\sum[-2p_iq_i(1-p_i)(1-q_i)+(p_i-p_i^2)+(q_i-q_i^2)+(p_i^2+q_i^2-2p_iq_i)]$$

again in terms of the new variables

$$=\sum[-2x_iy_i+x_i+y_i+(p_i-q_i)^2]$$

This implies that,

$$\sum[(x_i-y_i)^2+(p_i-q_i)^2]=0.$$

which occurs only if $p_i=q_i$ for all $i=1,\ldots,n$.

Therefore the *asymptotic* variance of the pairwise genetic distances (normalized by number of loci) of genotypes sampled from the combined population, comprising of two subpopulations, is zero *iff* this combined population is essentially a single panmictic population (i.e., p_i=q_i for all i, or F_{ST}=0). Since we have defined E_{ST} in terms of standard deviations rather than variances, we will subsequently consider the *asymptotic standard deviation* SD_{T}, which is simply defined as the square root of the *asymptotic variance*, *S* for the 'total' population. Fig. 1 depicts a numerical simulation of both SD_T and the average within-population SD (SD_W) for our two population model, as a function of the number of SNPs considered. While SD_W converges to zero, SD_T asymptotes to a value greater than zero, revealing the underlying structure.

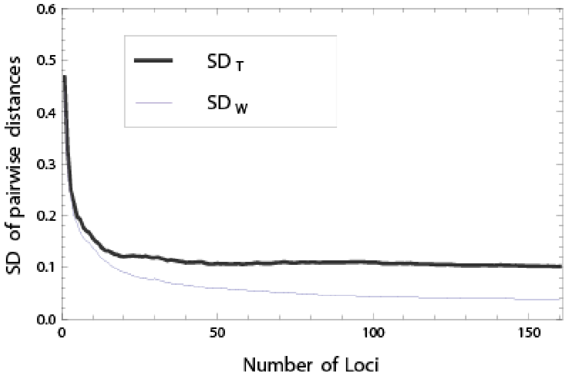

Fig 1. A simulation of SD_T and SD_W under a two-population model demonstrating their divergent behavior with an increasing number of SNP loci. Here SNP frequencies are modeled on *Beta* distributions (as in Tal 2013) with F_{ST}=0.10.

To further substantiate SD_T as a measure of structure, we would like to characterize the relation of SD_T to F_{ST}, both formulated as expressions of allele frequencies from two populations. We will proceed numerically, as our goal here is merely to get a qualitative intuition into the association of the two statistics.

We have from Eqs. (5), (6), (7), that asymptotically as $n \rightarrow \infty$, or practically under a high number of SNP loci,

$$SD_T = \sqrt{S} = \sqrt{\tfrac{1}{4}(\mu_W - \mu)^2 + \tfrac{1}{4}(\mu_Y - \mu)^2 + \tfrac{1}{2}(\mu_Z - \mu)^2} \tag{9}$$

where,

$$\mu = \tfrac{1}{4}\mu_W + \tfrac{1}{4}\mu_Y + \tfrac{1}{2}\mu_Z$$

$$\mu_W = lim_{n\rightarrow\infty} \frac{1}{n}\sum_{i=1}^{n} 4(-p_i^4 + 2p_i^3 - 2p_i^2 + p_i)$$

$$\mu_Y = lim_{n\rightarrow\infty} \frac{1}{n}\sum_{i=1}^{n} 4(-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)$$

$$\mu_Z = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}2(2p_iq_i^2+2q_ip_i^2-2p_i^2q_i^2-4p_iq_i+p_i+q_i)$$

And from Tal (2013, Eq. 10) we use the most common expression for F_{ST} across any number of n SNPs,

$$F_{ST}=\frac{\sum_{i=1}^{n}(p_i-q_i)^2}{\sum_{i=1}^{n}(p_i+q_i)(2-p_i-q_i)} \qquad (10)$$

Under the standard assumption that SNP frequencies are modeled on a *Beta* distribution with parameters deriving from some historical process (see Tal 2013; Gao and Martin, 2009) we sample a large number of sets of SNP frequencies for two populations, each set generated from two *Beta* distributions with some randomized parameters. For each set we compute the pair SD_T (Eq. 9) and F_{ST} (Eq. 10) to generate a scatter plot of their association. Fig. 2 is a typical instance of such simulation, and demonstrates that the correlation of the two statistics is quite substantial,

$$0 << \rho[SD_T, F_{ST}] < 1$$

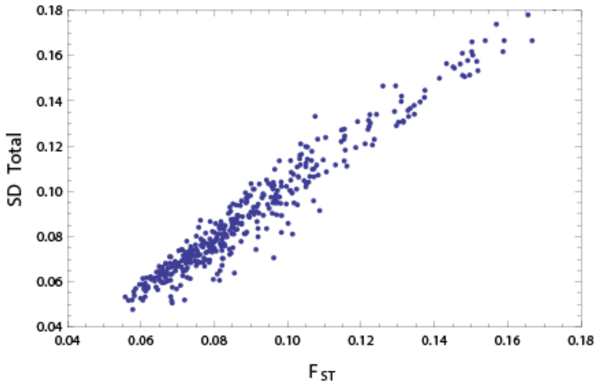

Fig 2. A scatter plot indicating a high positive correlation the two statistics SD_T and F_{ST}. Each dot represents the two statistics computed for data sampled from a two-population model with 1000 SNPs, and allele frequencies from Beta distributions. The Pearson product-moment correlation coefficient is here roughly 0.97. Note also the regression is approximately 1, implying that SD_T and F_{ST} take on very similar values.

A further perspective into SD_T as an unsupervised measure of internal structure is afforded by a qualitative comparison with principal component analysis (PCA) plots on data generated by the model. PCA is an *unsupervised* technique used to emphasize variation and bring out strong patterns in a dataset, essentially a dimensionality reduction procedure. It can be used as a 'preprocessing' stage for clustering high-dimensionality data, such as characteristic of population genetic samples. In such setting, the first principal components tend to also extract the substructure of the data – revealing the existence of clusters in the data (Patterson et al. 2006). But more crucially to our goals, the dispersion of clusters revealed by PCA is highly associated with internal structure, i.e., departures from panmixia, with increasing number of loci (and asymptotically, panmictic clusters would diminish to a single dot). This property is congruent with the convergence of SD_T to some value strictly greater than zero for non-panmictic populations. This is depicted in the four PCA plots of the same populations under increasing SNP count in Fig 3A-D.

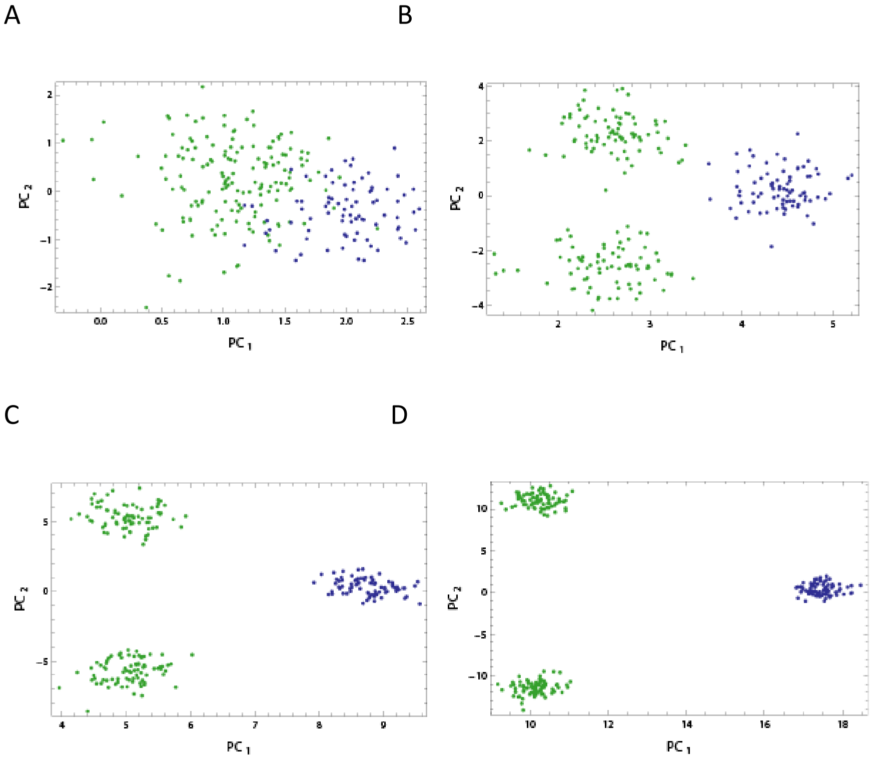

Fig 3. PCA plots demonstrating the much pronounced decrease in SD_T for the panmictic population (blue) relative to the structured one with an internal F_{ST}=0.005 (green), as the number of SNP loci processed by the PCA scheme is increased. A: 1K SNPs | B: 5K SNPs | C: 20K SNPs | D: 60K SNPs.

References

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L (2002) A human genome diversity cell line panel. Science 296 (5566): 261–2.

Darwin C (1871) The Descent of Man and Selection in Relation to Sex. London: John Murray.

Edwards AWF (2003) Human genetic diversity: Lewontin's fallacy, Bioessays, 25, 798–801.

Gao X, Martin ER (2009) Using allele sharing distance for detecting human population stratification. Hum Hered, 68, 182–191.

Hedrick PW (2005) A standardized genetic differentiation measure. Evolution, 59, 1633–1638.

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132 (2), 583–589.

Jorde LB, Wooding SP (2004) Genetic variation, classification, and "race." Nat Genet, 36, S28–32.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. Mol Ecol, 17, 4015–4026.

Lewontin RC (1972) The apportionment of human diversity. Evol Biol, 6, 381–98.

Mitton JB (1977) Genetic Differentiation of Races of Man as Judged by Single-Locus and Multilocus Analyses. Amer Nat, 111 (978), 203–212.

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA, 70, 3321–3323.

Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, et al. (2009) Genetic structure of Europeans: a view from the North-East. PLoS One 4: e5472.

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2: 2074–2093. doi:10.1371/journal.pgen.0020190.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, e Zhivotovsky LA, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet, 70: 841–847.

Tal O (2013) Two complementary perspectives on inter-individual genetic distance. Biosystems, 111: 18–36.

Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, Gregersen PK, et al. (2008) Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays. PLoS ONE, 3, (12): e3862.

Waples RS (1991) Definition of 'species' under the Endangered Species Act: Application to Pacific salmon. U.S. Department of Commerce NOAA Technical Memorandum, NMFS, F/NWC–194.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38 (6): 1358.

Whitlock M (2011) G'st and D do not replace Fst. Mol Ecol 20:1083–109.

Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, et al. (2007) Genetic Similarities Within and Between Human Populations. Genetics 176(1): 351–9. pmid:17339205 doi: 10.1534/genetics.106.067355.

Wright S (1978) Evolution and the Genetics of Populations. Vol. 4, Variability Within and Among Natural Populations. University of Chicago Press, Chicago.