The Equidistance Index of Population Structure

by

Yaron Granot, Omri Tal, Saharon Rosset, and Karl Skorecki

Preprint no.: 28 2016
RESEARCH ARTICLE

The Equidistance Index of Population Structure

Yaron Granot*1, Omri Tal2, Saharon Rosset3, and Karl Skorecki1

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 FST, 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 FST) 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 FST is also a function of internal diversity, it does not directly reflect the strength of separation between population clusters. For any given FST value, clusters can be tighter (more panmictic) or looser (more stratified), and in this respect higher FST does not always imply stronger differentiation. Finally, FST 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 EST, 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 FST and EST and found substantial differences in ranking order. In some cases examined, most notably among isolated Amazonian tribes, EST ranking seems more consistent with demographic, phylogeographic and linguistic measures of classification compared to FST. Thus, EST may at times outperform FST in identifying evolutionarily significant differentiation.

Keywords: FST, 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 1A), 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 1B), reflecting the much higher \( E_{ST} \).

![Figure 1](image)

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

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

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 ($F_{ST}=0.12$) or as the Bantu are from the Mongola ($F_{ST}=0.13$). 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 4A).

![Figure 4](https://example.com/figure4.png)

**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 4B), which is further accentuated by removing the inter-individual component (Figure 4C) and stretching the Yoruba-Russian tree to match the level of structure observed in the Surui-Karitiana tree (Figure 4D). 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\textsubscript{ST} and the Dissimilarity Fraction

The dissimilarity fraction, \( \omega \), 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 \( \omega=0 \) indicating that individuals are always closer to members of their own population and \( \omega=0.5 \) 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 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 (\( \omega=0 \)) between the Adygei and Uygur HGDP samples, but some overlap (\( \omega > 0 \)) between Mayans and Surui, despite a 4x higher F\textsubscript{ST} (Figure 5). Thus, F\textsubscript{ST} and the dissimilarity fraction (\( \omega \)) are not necessarily congruent. The E\textsubscript{ST} values for these two population pairs are more consistent with \( \omega \), 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\textsubscript{ST} vs. genetic similarity in various population pairs.](image)

Pairwise distances are colored red or blue within populations and black between populations. (A) Even at a relatively low F\textsubscript{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\textsubscript{ST} of 0.09, there is substantial overlap between Maya-Maya (red) and Maya-Surui (black) samples. E\textsubscript{ST} values are more consistent with the within-vs.-between population overlap and the dissimilarity fraction (\( \omega \)).

Summary and Conclusions

The main distinction between F\textsubscript{ST} and E\textsubscript{ST} is that F\textsubscript{ST} partitions diversity, whereas E\textsubscript{ST} partitions structure within and between populations. F\textsubscript{ST} is more sensitive to effective population size, while E\textsubscript{ST} is more sensitive to outliers, though this is largely mitigated by using E\textsubscript{ST} median rather than E\textsubscript{ST} mean (see Materials and Methods). F\textsubscript{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\textsubscript{ST}, but it does demonstrate a conceptual disconnect between F\textsubscript{ST} and clustering. Sewall Wright proposed a series of arbitrary F\textsubscript{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](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):
\[ F_{ST} = 1 - \frac{S}{T} \]  

(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:

\[ E_{ST} = 1 - \frac{SD_S}{SD_T} \]  

(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: \( E_{STmin} \), \( E_{STmedian} \), and \( E_{STmax} \) (Figure S10). All four estimators use the same basic formula, with only the type of \( SD_S \) differing among estimators. In \( E_{STmin} \), \( E_{STmedian} \), and \( E_{STmax} \), \( SD_S \) 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. \( E_{STmin} \) uses the smallest individual \( SD_S \) from each population, i.e., the SD of the most panmictic sample, \( E_{STmedian} \) uses the median individual \( SD_S \), and \( E_{STmax} \) 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 \( E_{STmedian} \) (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, \( E_{STmax} \) may be preferable, as long as at least one individual has no close relatives among the samples. \( E_{ST} \) values, especially \( E_{STmin} \) 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:

\[ E_{BT} = 1 - \frac{SD_B}{SD_T} \]  

(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,
\[ \text{Var}[D_n] = E[D_n^2] - E[D_n]^2 \]

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

\[
E[D_n^2] = \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 = \frac{1}{n^2} \left( \sum_{i=1}^{n} X_i^2 + 2 \sum_{i \neq j} (X_i \cdot X_j) \right) = \frac{1}{n^2} \left( \sum_{i=1}^{n} E[X_i^2] + 2 \sum_{i \neq j} E[X_i \cdot X_j] \right)
\]

\[
= \frac{1}{4} \cdot \frac{1}{n^2} \left( \sum_{i=1}^{n} E[X_i^2] + 2 \sum_{i \neq j} E[X_i \cdot X_j] \right) \left| \text{both genotypes from pop 1} \right.
\]

\[
+ \frac{1}{4} \cdot \frac{1}{n^2} \left( \sum_{i=1}^{n} E[X_i^2] + 2 \sum_{i \neq j} E[X_i \cdot X_j] \right) \left| \text{both genotypes from pop 2} \right.
\]

\[
+ \frac{1}{2} \cdot \frac{1}{n^2} \left( \sum_{i=1}^{n} E[X_i^2] + 2 \sum_{i \neq j} E[X_i \cdot X_j] \right) \left| \text{one genotype from pop 1 and one from pop 2} \right.
\]

\tag{1}

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 \( \frac{1}{4}, \frac{1}{4}, \frac{1}{2} \) 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,

\[
E[X_i] \left| \text{both genotypes from pop 1} \right. = 4(-p^4 + 2p^3 - 2p^2 + p) \\
E[X_i] \left| \text{both genotypes from pop 1} \right. = 4p_i(1-p_i) \\
E[X_i] \left| \text{both genotypes from pop 2} \right. = 4(-q^4 + 2q^3 - 2q^2 + q) \\
E[X_i] \left| \text{both genotypes from pop 2} \right. = 4q_i(1-q_i) \\
E[X_i] \left| \text{one genotype from pop 1 and one from pop 2} \right. = 2(2p_iq_i^2 + 2q_i^2p_i - 2p_i^2q_i - 4p_iq_i + p_i + q_i) \\
E[X_i] \left| \text{one genotype from pop 1 and one from pop 2} \right. = 2(p_i^2 + q_i^2 - 4p_iq_i + p_i + q_i)
\]

So that,
\[
E[D_n^2] = \frac{1}{n^2} \left[ \sum_{i=1}^{n} p_i(1 - p_i) + 8 \sum_{i=j}^{n} (-p_i^4 + 2p_i^3 - 2p_i^2 + p_i)(-p_j^4 + 2p_j^3 - 2p_j^2 + p_j) \\
+ \sum_{i=1}^{n} q_i(1 - q_i) + 8 \sum_{i=j}^{n} (-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)(-q_j^4 + 2q_j^3 - 2q_j^2 + q_j) \\
+ 2 \sum_{i=1}^{n} (p_i^2 + q_i^2 - 4p_iq_i + p_i + q_i) \\
+ 8 \sum_{i=j}^{n} (2p_iq_i^2 + 2q_ip_i^2 - 2p_i^2q_i - 4p_iq_i + p_i + q_i)(2p_jq_j^2 + 2q_jp_j^2 - 2p_j^2q_j - 4p_jq_j + p_j + q_j) \right]
\]

(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,
\[
Var[D_n] = E[D_n^2] - E[D_n]^2
\]

\[
= \frac{1}{n^2} \left[ \sum_{i=1}^{n} p_i(1 - p_i) + 8 \sum_{i=j}^{n} (-p_i^4 + 2p_i^3 - 2p_i^2 + p_i)(-p_j^4 + 2p_j^3 - 2p_j^2 + p_j) \\
+ \sum_{i=1}^{n} q_i(1 - q_i) + 8 \sum_{i=j}^{n} (-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)(-q_j^4 + 2q_j^3 - 2q_j^2 + q_j) \\
+ 2 \sum_{i=1}^{n} (p_i^2 + q_i^2 - 4p_iq_i + p_i + q_i) \\
+ 8 \sum_{i=j}^{n} (2p_iq_i^2 + 2q_ip_i^2 - 2p_i^2q_i - 4p_iq_i + p_i + q_i)(2p_jq_j^2 + 2q_jp_j^2 - 2p_j^2q_j - 4p_jq_j + p_j + q_j) \right] - \frac{1}{16} \left( \sum_{i=1}^{n} (p_i + q_i)(2 - p_i - q_i)(4 - (p_i + q_i)(2 - p_i - q_i)) \right)^2
\]

(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 $F_{st}>0$, 

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

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 $\alpha$ for $W$, a probability $\beta$ for $Y$ and a probability $\gamma$ 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 $X_n$, they are identically distributed but not independent. Now we set,

$$D_n = \frac{1}{n} \sum_1^n X_i$$

We would like to get an expression for $S$ in terms of the expectations of $W, Y, Z$ and $\alpha, \beta, \gamma$, where,

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

From the law of total variance,

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

where $B$ here is a categorical random variable that describes from which of distributions $W, Y, Z$ we are sampling from, with probabilities $\alpha, \beta, \gamma$ respectively, and where $U_{XYZ}$ is a discrete random variable taking the values of $\mu_w=E[W], \mu_y=E[Y], \mu_z=E[Z]$ with corresponding probabilities $\alpha, \beta, \gamma$ respectively. Hence at the limit $n \to \infty$ we have,

$$S = \text{Var}(U_{XYZ}) = \alpha(\mu_w - \mu)^2 + \beta(\mu_y - \mu)^2 + \gamma(\mu_z - \mu)^2$$

(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 $\alpha, \beta, \gamma$). 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 $n \to \infty$),

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

(6)

with probabilities $\alpha, \beta, \gamma$.

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 $\mu_W$ and $\mu_Y$ in Eq. (6) represent the two within-population pairwise distance means and $\mu_Z$ the total-population mean (derived below), and where the respective probabilities are as in Eq. (1), $\alpha=\frac{1}{4}, \beta=\frac{1}{4}, \gamma=\frac{1}{2}$, assuming infinite population size. Again, $S=0$ if and only if these means are equal, i.e., $\mu_W=\mu_Y=\mu_Z$. 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 $n$,

$$
E[D_n] \mid \text{both genotypes from pop 1} = \frac{1}{n} \sum_{i=1}^{n} 4(-p_i^4 + 2p_i^3 - 2p_i^2 + p_i)
$$

$$
E[D_n] \mid \text{both genotypes from pop 2} = \frac{1}{n} \sum_{i=1}^{n} 4(-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)
$$

(7)

$$
E[D_n] \mid \text{one genotype from pop 1 and one from pop 2}
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} 2(2p_i q_i^2 + 2q_i p_i^2 - 2p_i^2 q_i^2 - 4p_i q_i + p_i + q_i)
$$

bearing in mind the analysis pertains to $E[D_n]$ as $n \to \infty$. We proceed to examine what can be concluded from the equalities $\mu_W=\mu_Y=\mu_Z$ (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 $n \to \infty$). Thus we start by explicitly writing the equalities (where the $1/n$ cancels out),
\[
\sum_{i=1}^{n} (-p_i^4 + 2p_i^3 - 2p_i^2 + p_i) = \sum_{i=1}^{n} (-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)
\]

\[
\sum_{i=1}^{n} (-p_i^4 + 2p_i^3 - 2p_i^2 + p_i) = \sum_{i=1}^{n} (2p_i q_i^2 + 2q_i p_i^2 - 2p_i^2 q_i^2 - 4p_i q_i + p_i + q_i)
\]

(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),

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

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

again in terms of the new variables

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

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.

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 $n \to \infty$, or practically under a high number of SNP loci,

$$SD_T = \sqrt{S} = \sqrt{\left( \frac{1}{3} (\mu_W - \mu)^2 + \frac{1}{3} (\mu_Y - \mu)^2 + \frac{1}{2} (\mu_Z - \mu)^2 \right)}$$  \hspace{1cm} (9)$$

where,

$$\mu = \frac{1}{4} \mu_w + \frac{1}{4} \mu_Y + \frac{1}{2} \mu_Z$$

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

$$\mu_Y = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} 4(-q_i^4 + 2q_i^3 - 2q_i^2 + q_i)$$
\[
\mu_Z = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left( 2p_i q_i^2 + 2q_i p_i^2 - 2p_i^2 q_i^2 - 4p_i q_i + p_i + q_i \right)
\]

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

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

Under the standard assumption that SNP frequencies are modeled on a \textit{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 \textit{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,

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

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


