The Apportionment of Total Genetic Variation by Categorical Analysis of Variance

Tsung Fei Khang*  Von Bing Yap†

*National University of Singapore, mrtfkhang@yahoo.com
†National University of Singapore, stayapvb@nus.edu.sg
The Apportionment of Total Genetic Variation by Categorical Analysis of Variance∗

Tsung Fei Khang and Von Bing Yap

Abstract

We wish to suggest the categorical analysis of variance as a means of quantifying the proportion of total genetic variation attributed to different sources of variation. This method potentially challenges researchers to rethink conclusions derived from a well-known method known as the analysis of molecular variance (AMOVA). The CATANOVA framework allows explicit definition, and estimation, of two measures of genetic differentiation. These parameters form the subject of interest in many research programmes, but are often confused with the correlation measures defined in AMOVA, which cannot be interpreted as relative contributions of particular sources of variation. Through a simulation approach, we show that under certain conditions, researchers who use AMOVA to estimate these measures of genetic differentiation may attribute more than justified amounts of total variation to population labels. Moreover, the two measures can also lead to incongruent conclusions regarding the genetic structure of the populations of interest. Fortunately, one of the two measures seems robust to variations in relative sample sizes used. Its merits are illustrated in this paper using mitochondrial haplotype and amplified fragment length polymorphism (AFLP) data.

KEYWORDS: AFLP, AMOVA, ANOVA, binary data, CATANOVA, categorical data, measure of genetic differentiation, mitochondrial haplotype

*We would like to thank Sharon Sim, Dr. Yu Dahui and Dr. Lene Rostgaard Nielsen for providing the AFLP data sets for analysis. Two anonymous reviewers gave constructive comments that resulted in substantial improvement to form and content.
1 Introduction

The apportionment of total genetic variation to its respective sources of variation informs us about the relative importance of these sources. Such information can provide an objective basis for guiding research programmes. In conservation biology, it may be possible to preserve a significant portion of the total genetic variation in a species by protecting one or two chosen populations, if most genetic variation resides within populations. On the other hand, the loss of even a single population may prove disastrous for a species's evolutionary potential if substantial genetic variation resides between populations. Anthropology has also benefitted from insights obtained from the study of how total genetic variation is partitioned to within and between ethnic groups. Using electrophoretic gene frequency data, Lewontin (1974) presented evidence that ethnic labels only account for a small proportion of total genetic variation in humans. Since then, overwhelming evidence in support of this finding has accrued (see Chakraborty, 1982; Witherspoon et al., 2007).

In the study of molecular genetic variation, Excoffier et al. (1992) introduced a method known as the analysis of molecular variance (AMOVA) for analysing multilocus molecular variation. The AMOVA is an extension of previous studies (see Cockerham, 1969, 1973; Weir 1996) that use a similar analysis of variance (ANOVA) approach to analyse population genetic structure. Its parameters of interest are known as "Φ-statistics", which measure "haplotypic correlation". A software known as Arlequin (Excoffier et al., 2005), now in its third version, supports the implementation of AMOVA.

Our extensive survey reveals that most researchers interpret the resultant Φ-statistics as the proportion of total genetic variation that is explained by a particular source of variation (see for example, Tero et al., 2003; Honnay et al., 2006; Scheinfeldt et al., 2007). This discrepancy between what researchers are interested in knowing, and what AMOVA actually offers, seems to have been overlooked. Indeed, whereas proportions are bounded within the unit interval, in principle no such restriction applies to the Φ-statistics, which can take negative values. The latter cases are said to be possible in outcrossing species where genes from different populations can be more related to each other than genes from within the same population (see Excoffier, 2009). If this explanation is incompatible with the known biology of the species of interest, researchers often pragmatically interpret negative estimates of Φ-statistics as indication that a particular source of variation contributes nothing to total variation.

In this paper, we wish to present evidence that the categorical analysis of variance (CATANOVA; Light and Margolin, 1971) is the appropriate frame-
work for answering the question of apportioning total genetic variation. To
acquaint readers with the necessary background on CATANOVA, AMOVA
and $G_{ST}$ (a widely used measure of allelic differentiation among populations;
see Nei, 1973 and Nei, 1987), we have prepared an appendix. In the results
section, we show that the CATANOVA $R^2$ statistic, which is analogous to
ANOVA’s, estimates $\gamma$ under simple random sampling. Because this param-
eter can measure genetic differentiation at the allelic, genotypic, or haplotypic
level, the result provides an ANOVA perspective to Nei’s $G_{ST}$. Using sim-
ulations based on actual data, we highlight a situation where two candidate
parameters for summarising population genetic structure: $\bar{\gamma}$ (the average of
locus-specific $\gamma$), or $\gamma^M$ (a variant of $\gamma$ with equal weight for each locus),
can potentially lead to conflicting interpretation of population genetic structure.
Finally, we discuss the relevance of this finding in applied work.

2 Methods

2.1 A Truncation Algorithm for Removing Correlated
Binary Variables

The main purpose of the following algorithm is not so much for practical use,
as it is for theoretical discussion. Its aim is to chiefly illustrate potential
discrepancy that can arise if correlated variables are used in a method that
assumes all variables to be independent.

As input, the truncation algorithm takes in the correlation of two binary
variables, which is given by (see Bishop et al., 1975)

$$r = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1+}p_{2+}p_{+1}p_{+2}}}$$

(1)

where $p_{ij}$ are the proportions of $ij$-type event in a $2 \times 2$ table for the two
binary variables; and $p_{i+}, p_{+j}$ are the row and column marginal proportions,
respectively. Although zero correlation does not imply independence in gen-
eral, it is true for the case of binary variables. The proof is simple: from (1),
zero-correlation implies that the covariance $p_{11}p_{22} - p_{12}p_{21} = 0$. Since a $2 \times 2$
table with fixed marginals has only one degree of freedom, we just need to
show that $p_{i+}p_{j+} = p_{22}$, which involves only simple algebra.

To obtain a subset of independent binary variables from an initial set $L$
variables (loci), we used a test statistic
\[ H_{\text{max}} = \max_{1 \leq i \leq L} \{H_i\}, \]

where

\[ H_i = \sum_{g=1}^{G} \sum_{j=1}^{L} s_g |r_{ij}(g)|. \]  

(2)

Here, \(|r_{ij}(g)|\) is the absolute value of the correlation between the \(i\)th and \(j\)th variables (the \(ij\)-cell of the sample correlation matrix) in the \(g\)th population, and \(s_g\) denotes the relative sample size of the \(g\)th population. A similar use of the absolute correlation in the study of gene co-expression networks is given in Zhang and Horvath (2005). If the \(i\)th locus is more or less uncorrelated with other loci, then (2) will be close to 1, which is its lower bound.

The vector of proportion of enzyme restriction cuts at each site is given by

\[ \mathbf{p}^{(g)} = (p_1^{(g)}, p_2^{(g)}, \ldots, p_L^{(g)}). \]

Since the \(p_i^{(g)}\) are unknown, we replaced them with their sample estimates \(\hat{p}_i^{(g)}\). For the \(g\)th population, we performed Bernoulli sampling with “success” (presence of restriction enzyme cut) probability \(\hat{p}_i^{(g)}\) in each individual at the \(i\)th locus. We simulated the distribution of \(H_{\text{max}}\) (500 iterations) under the null hypothesis that all entries in \(\mathbf{p}^{(g)}\) are mutually independent. Against this null distribution, we computed the approximate \(p\)-value as the proportion of simulated \(H_{\text{max}}\) values that exceed the observed \(H_{\text{max}}\) from the data set. If the \(p\)-value is less than 0.05, we truncated the locus that gives \(H_{\text{max}}\) from the set of \(L\) loci in all three populations. We then iterated this procedure using the truncated set of loci until we obtained a final set with \(p\)-value greater than 0.05.

2.2 Analysis of Empirical Data

Two types of published data: mtDNA haplotype data, and amplified fragment length polymorphism (AFLP; Vos et al., 1995; see also Meudt and Clark, 2007) nuclear genotype data were used to illustrate important results in this paper.
1. MtDNA Haplotype Data

In many multicellular organisms, the mtDNA is uniparentally inherited (usually along maternal lines), and therefore haploid in nature. Depending on the DNA sequence encountered by the restriction enzymes used, DNA breakage may or may not occur at various sites of the mtDNA molecule. The pattern of breakage thus provides a genetic signature that allows inference of the closeness or remoteness of ancestry. A mtDNA haplotype with \( L \) polymorphic loci (variables) can be represented as a vector of binary variables (1 for presence of breakage; 0 for absence). If the haplotypes are treated as categories, then \( 2^L \) unique haplotypes are possible. The latter could be large even for modest \( L \). In practice, only a small subset of them are present in a population.

We used two mtDNA data sets to illustrate the univariate approach (haplotypes as categories), and conflicts in interpretation of population genetic structure due to choice of \( \gamma \) or \( \gamma^M \) as the parameter of interest (details in Section 3). The first one is Johnson et al.’s (1983) mtDNA data set (see Appendix). The data consist of mtDNA haplotypes obtained using a set of five restriction enzymes: \( \text{BamHI, HpaI, HaeII, AvaII and MspI} \). These restriction enzymes collectively induce breaks at a total of 25 loci, 23 of which are polymorphic, generating a total of 35 unique haplotypes. Assuming stratified random sampling, samples were taken from populations in Africa (Africans, \( N_1 = 74 \)), Europe (Caucasians, \( N_2 = 50 \)), East Asia (Orientals, \( N_3 = 46 \)) and North America (American aborigines). We considered only the first three populations. The second data set comes from Harihara et al.’s (1988; see Appendix) study of mtDNA variation in five Asian populations: Japanese (\( N_1 = 74 \)), Korean (\( N_2 = 64 \)), Ainu (\( N_3 = 48 \)), Aeta (\( N_4 = 37 \)) and Vedda (\( N_5 = 20 \)). The set of restriction enzymes used: \( \text{AvaII, HincII, HpaI, MspI, PstI, PvuII, XbaI and XhoI} \) generated a total of 20 unique haplotypes.

To quickly assess the correlation between loci, we used a “heat plot” to summarise graphically the sample correlation matrix for each of the three populations. The heat plot replaces entries in the sample correlation matrix with suitable color tones ranging from beige to red. Each tone (in increasing sharpness) corresponds to correlation magnitudes in the set of intervals \( \{ [0, 0.1), [0.1, 0.2), \ldots, [0.9, 1] \} \).
2. AFLP Data

The AFLP is a molecular biology technique that is capable of harvesting genotype information from large numbers of loci. Presumably, the method samples the latter randomly in the genome of an organism. Because this method appears to score highly polymorphic loci and does not require prior sequence knowledge (see Lowe et al., 2004), it is ideally suited for preliminary studies of the population genetic structure of a species. The AFLP methodology, however, suffers from genotype ambiguity induced by dominance. Assuming a two-allele locus, only two phenotypes can be observed instead of three, since the heterozygous state cannot be differentiated from the dominant homozygous state. These two states are visualised as a band on a polyacrylamide gel, whereas the null homozygote state is indicated by band absence.

We used three AFLP data sets to study the relation between $\bar{\gamma}$ and $\gamma^M$ via simulation. The first AFLP data set comes from a study of the pearl oyster *Pinctada fucata* (Yu and Chu, 2006) at three locations: China ($N_1 = 28$), Japan ($N_2 = 19$) and Australia ($N_3 = 15$). The total number of loci is 184. The second one comes from a study of a lowland rainforest tree species *Calophyllum ferrugineum* at five locations in Singapore (Sim, 2007). These locations are the Singapore Botanic Gardens ($N_1 = 42$), Upper Pierce ($N_2 = 5$), Bukit Timah Nature Reserve ($N_3 = 32$), MacRitchie Reservoir ($N_4 = 41$) and Upper Seletar ($N_5 = 8$). The total number of loci is 159. The third one comes from a study of the endemic plant *Scalesia affinis* in seven populations at the Galápagos Islands (Nielsen, 2004). Four of them are located in Isabela Island ($N_1 = 18, N_2 = 15, N_3 = 16, N_4 = 20$), one in Santa Cruz Island ($N_5 = 5$), and the rest in Floreana Island ($N_6 = 22, N_7 = 28$). Where data at particular loci were missing (assumed to be at random) for an individual, we imputed them using Bernoulli random variables with “success” (band presence) probability estimated using the remaining individuals in the population. The total number of loci is 157.

For each data set, we first estimated $\bar{\gamma}$, $\sigma_\gamma$ (standard deviation of locus-specific $\gamma$) and $\gamma^M$ in the data sets using CATANOVA and AMOVA. Then, we assumed that $\gamma^M = \hat{\gamma}_C^M$ and $\bar{\gamma} = \bar{\hat{\gamma}}_C$, where the subscript C denotes estimation by CATANOVA. Next, we generated 1000 simulated data by resampling individuals within populations with replacement, and then estimated $\bar{\hat{\gamma}}$, $\sigma_{\hat{\gamma}}$ and $\hat{\gamma}^M$ in each simulated data. To assess the correlation between $\bar{\hat{\gamma}}$ and $\hat{\gamma}^M$, we plotted their estimates (computed from the simulated data) against each other. We also compared the root mean squared error (RMSE) of $\bar{\hat{\gamma}}_C$ and $\bar{\hat{\gamma}}_A$, where the subscript A denotes estimation by AMOVA. To investigate the effect
of unbalanced relative sample sizes, we repeated the simulation for Sim’s and Nielsen’s data, using different groupings of populations. For Sim’s data, we combined the Bukit Timah Nature Reserve, MacRitchie Reservoir and Upper Seletar populations. For Nielsen’s, we grouped the populations according to the three islands. Table 1 shows the variance of relative sample sizes under these scenarios.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Sample sizes</th>
<th>Total</th>
<th>Variance of relative sample sizes (×100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yu and Chu (2006)</td>
<td>N = (28,15,19)</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>Sim (2007)</td>
<td>N = (42,5,32,41,8)</td>
<td>128</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>N = (42,5,81)</td>
<td>128</td>
<td>9</td>
</tr>
<tr>
<td>Nielsen (2004)</td>
<td>N = (18,15,16,20,5,22,28)</td>
<td>124</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>N = (69,5,50)</td>
<td>124</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1 Variance of relative sample sizes in the three AFLP data sets.

2.3 Computation

We used R (Version 2.6.1; R Core Development Team, 2007) to perform all necessary computations. The R codes can be obtained by e-mail request.

3 Results

3.1 Relation between CATANOVA $R^2$ and $G_{ST}$

Suppose we are interested in $G$ populations of the same species, and we have a simple random sample from the aggregate population. Assume that the data are haploid, with alleles as categories. Consider the simplest case where SST is partitioned to only two component sources of variation: SSB and SSW. As $N \to \infty$, we have $N_j/N \to w_j$, and $N_{ij}/N_j \to p_{ij}$, so that

$$N_i+ = \frac{N_i}{N} = \sum_{j=1}^{G} \frac{N_j}{N} \frac{N_{ij}}{N_j} \to \sum_{j=1}^{G} w_j p_{ij} = \bar{p}_i.$$  

Thus, from (7), $2 \frac{\text{SST}}{N}$ converges to

$$1 - \sum_{i=1}^{I} \bar{p}_i^2 = \sum_{l \neq m} \bar{p}_l \bar{p}_m.$$  

(3)
where \( l, m = 1, 2, \ldots, I \); from (9), \( 2 \frac{SSB}{N} \) converges to

\[
\sum_{j=1}^{G} \sum_{i=1}^{I} w_j \bar{p}_{ij}^2 - \sum_{i=1}^{I} \bar{p}_i^2.
\]

(4)

These results imply that \( R^2 = \frac{SSB}{SST} \) converges to the parameter

\[
\gamma = \frac{\sum_{j=1}^{G} \sum_{i=1}^{I} w_j \bar{p}_{ij}^2 - \sum_{i=1}^{I} \bar{p}_i^2 \sum_{l \neq m} \bar{p}_l \bar{p}_m}{\sum_{l \neq m} \bar{p}_l \bar{p}_m},
\]

which is \( G_{ST} \). Therefore, \( R^2 \) is an estimator of \( G_{ST} \) provided that the categories are alleles. More often, the categories correspond to genotypes. In diploids, a two-allele locus has three unique genotypes, and an \( s \)-allele locus has \( s(s+1)/2 \). In this case, \( \gamma \) no longer measures allelic differentiation like \( G_{ST} \), but measures genotypic differentiation. Similarly, if the categories are haplotypes, then \( \gamma \) is a measure of haplotypic differentiation. Since genotype proportions are functions of allele frequencies, \( G_{ST} \) is correlated with \( \gamma \). Hence, the latter may be useful as a proxy parameter.

In practice, stratified random sampling often substitutes for simple random sampling. As a result, the relative sample sizes do not properly estimate \( w_j \), hence \( R^2 \) cannot be a consistent estimator of \( G_{ST} \) or \( \gamma \). Even so, \( R^2 \) may still be useful in a qualitative sense - when \( p_{ij} \) are more or less similar for all \( i = 1, 2, \ldots, I \) and \( j = 1, 2, \ldots, G \), \( \gamma \) is close to zero even if \( w_j \) differ substantially between the populations. Thus, small values of \( R^2 \) obtained under stratified random sampling can be reasonably interpreted as evidence for low genetic differentiation between populations. On the other hand, when \( p_{ij} \) differ substantially, the magnitude of \( w_j \) can impact the final value of \( \gamma \). Consequently, it is less obvious how one should interpret moderate to large values of \( R^2 \) with stratified random sampling.

When multiple loci (\( L > 1 \)) are available, there are two possible ways of summarising the data. One way is to consider the distribution of locus-specific \( \gamma \), using its mean (\( \bar{\gamma} \)) and standard deviation (\( \sigma_\gamma \)) for data summary. We only need to assume that the loci have been randomly scored; they need not be assumed independent. Estimation of \( \bar{\gamma} \) is straightforward using the sample mean of \( L \) locus-specific \( R^2 \), and its sampling variance can be estimated by bootstrapping individuals within subpopulation (Van Dongen, 1995). Alternatively, we can summarise multilocus information using just a single parameter,
\[
\gamma^M = \frac{\sum_{k=1}^L \left( \sum_{j=1}^G \sum_{i=1}^I w_j p_{ijk}^2 - \sum_{i=1}^I \bar{p}_{ik}^2 \right)}{\sum_{k=1}^L \sum_{l \neq m} \bar{p}_{lk} \bar{p}_{mk}},
\]

where \(l, m = 1, 2, \ldots, I\), and the numerator and denominator are the same as those in \(\gamma\) except that they are now summed over all \(L\) loci. This parameter gives equal weight to all loci, thus implicitly assumes loci independence. In many cases, however, this assumption will not be true because of genetic linkage. It seems natural that the CATANOVA estimator of \(\gamma^M\) should be

\[
\hat{\gamma}^M_C = \frac{\sum_{i=1}^L \text{SSB}(i)}{\sum_{i=1}^L \text{SST}(i)}.
\]

From (17) and (18), the AMOVA estimator of \(\gamma^M\) is given by

\[
\hat{\gamma}^M_A = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_c^2}.
\]

### 3.2 Application of the Truncation Procedure

Figure 1 shows heat plots of the sample correlation matrix in each population for Johnson et al.’s data, before truncation.
Figure 2 Heat plots of the sample correlation matrix (truncated to 16 loci) for African, Caucasian and Oriental populations.

Figure 3 The simulated distributions of $H_{\text{max}}$ under the null hypothesis of independence among loci in each truncation step. The vertical dotted lines indicate the position of observed $H_{\text{max}}$ from the data. The $p$-value is indicated in a histogram’s main panel.

After truncation, only 16 loci remain. Figure 2 shows heat plots of the sample correlation matrix for these 16 loci. The contrast with Figure 1 is sharp - after truncation, the remaining loci are almost uncorrelated in each population. Figure 3 shows distributions of $H_{\text{max}}$ simulated under the null hypothesis of independence among loci, during each round of truncation.
3.3 Comparison between CATANOVA and AMOVA: Theory

Let us first consider the single locus case. Under the sampling scheme described in Section 3.1, the parameter of interest is locus-specific $\gamma$. In light of this, AMOVA is just another means of estimating $\gamma$. Estimates given by the CATANOVA and AMOVA estimators are correlated because they both use (7) and (9). The AMOVA estimator of $\gamma (\hat{\gamma}_A)$, however, returns an estimate that is higher than the one given by CATANOVA ($\hat{\gamma}_C = R^2$), if the following inequality is satisfied.

Proposition 1. Let there be $G$ populations of interest, and denote the sample size of the $j$th population ($j = 1, 2, \ldots, G$) as $N_j$. If the $F$-ratio: $\text{MSB}/\text{MSW} \geq N(N - G)(\sum_{j=1}^{G} N_j^2 - NG)^{-1}$, then $\hat{\gamma}_A \geq \hat{\gamma}_C$.

Proof. See Appendix.

It follows that AMOVA returns a lower estimate compared to CATANOVA if the inequality sign in (19) reverses. For the case of balanced design ($N_j = t$ for all $j$), inequality (19) becomes

$$F\text{-ratio} \geq G \left( \frac{t - 1}{t - G} \right) = G + \frac{G(G - 1)}{t - G}.$$ 

If $t \geq G^2$ - a condition satisfied in many practical situations unless $G$ is large, then the term $G(G - 1)/(t - G) \leq 1$. Therefore, checking whether the $F$-ratio exceeds $G + 1$ provides a quick way of knowing when $\hat{\gamma}_A \geq \hat{\gamma}_C$.

When $\gamma$ is small, AMOVA can return negative estimates of $\gamma$ as (18) is not guaranteed to be positive. Conversely, CATANOVA always returns a positive estimate, obviating the need to rationalise negative estimates that seem out of place when one is interested in discussing apportionment of total variation to different sources of variation. Note that both $\hat{\gamma}_A$ and $\hat{\gamma}_C$ are biased estimators of $\gamma$, since in general, the expectation of ratios is not the ratio of expectations. Whereas $\hat{\gamma}_C = R^2$ is a consistent estimator of $\gamma$ under simple random sampling (Section 3.1), $\hat{\gamma}_A$ is not.

Proposition 2. The AMOVA estimator of $\gamma (\hat{\gamma}_A)$ is not consistent.

Proof. See Appendix.
To estimate $\hat{\gamma}$ with multiple loci, we can use the sample mean of locus-specific $\gamma$ estimated using CATANOVA or AMOVA ($\hat{\gamma}_C$ and $\hat{\gamma}_A$ respectively). If $\hat{\gamma}_A$ is negative in many loci, then $\hat{\gamma}_A$ tends to be lower than $\hat{\gamma}_C$. If the parameter of interest is $\gamma^M$, then in the nonhierarchical case (total SS partitioned into just between and within population SS), $\hat{\Phi}_{ST}$ corresponds to $\hat{\gamma}_A^M$. Both $\hat{\gamma}_A^M$ and $\hat{\gamma}_C^M$ are correlated, and the result in Proposition 1 carries over in the following corollary.

**Corollary 3.** Results in Proposition 1 imply that $\hat{\gamma}_A^M \geq \hat{\gamma}_C^M$, with $SSB$ and $SST$ replaced by $\sum_{i=1}^L SSB(i)$ and $\sum_{i=1}^L SST(i)$, respectively.

### 3.4 Comparison between CATANOVA and AMOVA: Analysis of Empirical Data

1. **MtDNA Haplotype Data**

The univariate CATANOVA analysis of Johnson *et al.*’s mtDNA data set is straightforward, with $\hat{\gamma} = 0.12$. Consider the case where the parameter of interest is $\gamma^M$. Since the $F$-ratios before and after truncation exceed 2.96 (the lower bound of (19)), we have $\hat{\gamma}_A^M > \hat{\gamma}_C^M$ (Table 2). Regardless of estimation method, the estimated $\gamma^M$ is quite large, more so after truncation, indicating that an appreciable amount of total variation in mtDNA haplotypes is accounted for by population labels. Figure 4 shows the proportion of restriction enzyme cuts in all loci for each population, before and after truncation.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$\hat{\gamma}_C^M$</th>
<th>$\hat{\gamma}_A^M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>2</td>
<td>52.9 (30.4)</td>
<td>26.4 (15.2)</td>
<td>0.28 (0.39)</td>
<td>0.36 (0.49)</td>
</tr>
<tr>
<td>Within groups</td>
<td>167</td>
<td>137.8 (47.5)</td>
<td>0.83 (0.28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>169</td>
<td>190.7 (77.9)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Comparison of $\hat{\gamma}_C^M$ and $\hat{\gamma}_A^M$ before and after applying the truncation procedure. Values for the latter are indicated in parentheses. Note the discrepancy in estimated $\gamma^M$ as a consequence of using correlated or independent loci in the analysis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Median $\hat{\gamma}$</th>
<th>$\hat{\sigma}_\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATANOVA</td>
<td>0.04 (0.02)</td>
<td>0.09 (0.07)</td>
</tr>
<tr>
<td>AMOVA</td>
<td>0.04 (0.02)</td>
<td>0.10 (0.08)</td>
</tr>
</tbody>
</table>

Table 3 Summary statistics of the distribution of locus-specific $\hat{\gamma}$ under CATANOVA and AMOVA before and after truncation (in parentheses).

What if we are interested in $\bar{\gamma}$? Figure 5 shows that estimates of locus-specific $\hat{\gamma}_A$ and $\hat{\gamma}_C$ do not differ much (note that $\hat{\gamma}_A > \hat{\gamma}_C$ as seen in the
outliers), and are low in a majority of loci. Both methods subsequently give close estimates of $\bar{\gamma}$ and $\sigma_\gamma$ (Table 3). Whereas truncation of loci leads to rather large changes in $\hat{\gamma}^M$ (0.1 or more), it only changes $\hat{\gamma}$ and $\hat{\sigma}_\gamma$ slightly (at most 0.02).

Figure 4 Proportion of restriction enzyme cuts at each of the 23 mtDNA loci for the African, Caucasian and Oriental populations. The truncated loci (sites) are indicated in red.

Figure 5 Distribution of estimates of locus-specific $\gamma$ under CATANOVA and AMOVA before and after truncation. The numbers beside the outliers are the locus labels.
The preceding analyses show that the three possible ways of analysing mtDNA variation do not necessarily lead to the same conclusion. Both $\gamma$ from the univariate approach and $\bar{\gamma}$ lead to similar inference about the genetic structure of the three human populations - that most variation resides within rather than between populations. The converse, however, is suggested by $\gamma^M$.

2. AFLP Data

The distribution of estimated locus-specific $\gamma$ in each of the three AFLP data sets described in Section 2.2 is given in Figure 6. Table 4 contains CATANOVA and AMOVA estimates of $\bar{\gamma}$, $\sigma_\gamma$ and $\gamma^M$ in the data sets.

![Figure 6 Boxplots of the distribution of estimated locus-specific $\gamma$ in the three AFLP data sets.](image)

By plotting $\hat{\gamma}^M$ against $\hat{\gamma}$ for the simulated data (Figure 7), we have identified several interesting phenomena. First, we note that $\hat{\gamma}^M$ is correlated with $\hat{\gamma}$, regardless of whether AMOVA or CATANOVA is used. The magnitude of correlation appears to vary according to data set, being close to 1 for Yu and Chu’s data set (S1), about 0.9 in Nielsen’s (S3) and about 0.5 in Sim’s (S2). Second, $\hat{\gamma}^M$ can be much larger than $\hat{\gamma}$, particularly when AMOVA is used, and variance of the relative sample sizes is large. This is clearly seen in S2, where on average, $\hat{\gamma}_A^M$ is 0.075 but $\hat{\gamma}_A$ is 0.20; in S3, $\hat{\gamma}_A^M$ is 0.275 but $\hat{\gamma}_A$ is 0.40, on average. On the other hand, when variance of the relative sample sizes is small, interpretation of the relative importance of between population
contribution to total variation using $\hat{\gamma}$ or $\gamma^M$ is likely to be congruent. When estimating $\gamma$, AMOVA always has larger RMSE compared to CATANOVA (Figure 8). These results do not invalidate the use of the $\gamma^M$ parameter, or the AMOVA method of estimating it. Rather, we now understand that while $\gamma^M$ may not be the best parameter of choice, estimating it using AMOVA will probably lead to the correct interpretation of population genetic structure, unless the relative sample sizes are very unbalanced.

<table>
<thead>
<tr>
<th>Data set</th>
<th>$\hat{\gamma}_C$</th>
<th>$\hat{\sigma}_C$</th>
<th>$\hat{\gamma}^M_C$</th>
<th>$\hat{\gamma}_A$</th>
<th>$\hat{\sigma}_A$</th>
<th>$\hat{\gamma}^M_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yu and Chu (2006)</td>
<td>0.070</td>
<td>0.082</td>
<td>0.071</td>
<td>0.054</td>
<td>0.112</td>
<td>0.059</td>
</tr>
<tr>
<td>Sim (2007)</td>
<td>0.087</td>
<td>0.118</td>
<td>0.164</td>
<td>0.080</td>
<td>0.139</td>
<td>0.179</td>
</tr>
<tr>
<td>Nielsen (2004)</td>
<td>0.291</td>
<td>0.245</td>
<td>0.349</td>
<td>0.276</td>
<td>0.267</td>
<td>0.353</td>
</tr>
</tbody>
</table>

Table 4 Estimates of $\hat{\gamma}$, $\sigma_\gamma$ and $\gamma^M$ using CATANOVA and AMOVA for the three AFLP data sets.

Figure 7 Comparison of $\hat{\gamma}^M$ against $\hat{\gamma}$ using AMOVA and CATANOVA in three AFLP data sets. Labels: S1 for data from Yu and Chu (2006); S2 for data from Sim (2007); S3 for data from Nielsen (2004). Symbols: “•” for low relative sample size variance; “+” for high relative sample size variance.
Figure 8 Comparison of RMSE of \( \hat{\gamma}_A \) against \( \hat{\gamma}_C \) for three AFLP data sets. Labels: S1 for data from Yu and Chu (2006); S2 for data from Sim (2007); S3 for data from Nielsen (2004). Symbols: “.” for low relative sample size variance; “+” for high relative sample size variance.

### 3.5 Sensitivity Analysis

With simple random sampling, the relative sample sizes are unbiased estimators of the relative population weight \( w_j \); with stratified random sampling, they are not. To investigate how sensitive estimators of \( \gamma, \bar{\gamma}, \sigma, \) and \( \gamma^M \) are to \( w_j \), we used Johnson et al.’s mtDNA data (with the 16 loci after truncation) as a case study. Assuming that the African, Caucasian and Oriental populations are equally weighted, that the proportions of restriction enzyme cut at each locus are equal to those estimated from data, and that categories correspond to different haplotypes, we have \( \gamma = 0.11, \bar{\gamma} = 0.08, \sigma = 0.18 \) and \( \gamma^M = 0.37 \).

Table 5 shows how different estimates of \( w_j \) affect \( \hat{\gamma}_C \). It appears that \( \hat{\gamma}_C \) can still be close to 0.11, even if one estimate of \( w_j \) differs from 1/3 by as much as 0.33. However, substantial departure is the norm if this difference goes up to as high as 0.63. This result suggests that minor populations should not be included along with major populations if we want to preserve the interpretability of \( \gamma \). Further empirical support comes from analysis of Harihara et al.’s (1988) mtDNA data, which contain two major populations (Japanese and
Korean), while the remaining three are minor aboriginal populations found in Japan (Ainu), Philippines (Aeta) and Sri Lanka (Vedda). Under univariate CATANOVA, $\hat{\gamma}_C = 0.04$ when all five populations are included; with only Japanese and Korean, $\hat{\gamma}_C = 0.004$, which is smaller by a factor of 10. Among the three minor populations where no domination by any population is likely, $\hat{\gamma}_C = 0.07$, which appears reasonable as phylogenetic analysis (Harihara et al., 1988) indicates that these three populations diverged much earlier compared to Japanese and Koreans.

Table 5 Effects of different combinations of estimated $w_j$ on $\hat{\gamma}_C$. The entries in $w$ are in the order: African, Caucasian and Oriental populations.

<table>
<thead>
<tr>
<th>$w = (w_1, w_2, w_3)$</th>
<th>$\hat{\gamma}_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.6, 0.3, 0.1)</td>
<td>0.10</td>
</tr>
<tr>
<td>(0.3, 0.6, 0.1)</td>
<td>0.09</td>
</tr>
<tr>
<td>(0.3, 0.1, 0.6)</td>
<td>0.12</td>
</tr>
<tr>
<td>(0.9, 0.05, 0.05)</td>
<td>0.04</td>
</tr>
<tr>
<td>(0.05, 0.9, 0.05)</td>
<td>0.03</td>
</tr>
<tr>
<td>(0.05, 0.05, 0.9)</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 6 shows the impact of relative sample sizes on $\hat{\gamma}$, $\hat{\sigma}_C$ and $\hat{\gamma}_M$. In almost all combinations of sample sizes considered, $\hat{\gamma}_C$ is close to 0.08, and $\hat{\sigma}_C$ to 0.18, except in two cases where the relative sample sizes are very unbalanced. On the other hand, $\hat{\gamma}_M$ is close to 0.37 only for the first seven combinations of sample sizes. In the rest, the difference ranges from 0.08 to 0.26.

<table>
<thead>
<tr>
<th>$w = (w_1, w_2, w_3)$</th>
<th>$\hat{\gamma}_M^C$</th>
<th>$\hat{\gamma}_C$</th>
<th>$\hat{\sigma}_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.5, 0.25, 0.25)</td>
<td>0.39</td>
<td>0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>(0.25, 0.5, 0.25)</td>
<td>0.33</td>
<td>0.08</td>
<td>0.19</td>
</tr>
<tr>
<td>(0.25, 0.25, 0.5)</td>
<td>0.34</td>
<td>0.08</td>
<td>0.19</td>
</tr>
<tr>
<td>(0.6, 0.1, 0.3)</td>
<td>0.38</td>
<td>0.07</td>
<td>0.16</td>
</tr>
<tr>
<td>(0.6, 0.3, 0.1)</td>
<td>0.38</td>
<td>0.07</td>
<td>0.16</td>
</tr>
<tr>
<td>(0.3, 0.6, 0.1)</td>
<td>0.35</td>
<td>0.08</td>
<td>0.19</td>
</tr>
<tr>
<td>(0.3, 0.1, 0.6)</td>
<td>0.36</td>
<td>0.08</td>
<td>0.19</td>
</tr>
<tr>
<td>(0.1, 0.3, 0.6)</td>
<td>0.21</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>(0.1, 0.6, 0.3)</td>
<td>0.21</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>(0.8, 0.1, 0.1)</td>
<td>0.29</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>(0.1, 0.8, 0.1)</td>
<td>0.19</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>(0.1, 0.1, 0.8)</td>
<td>0.21</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>(0.9, 0.05, 0.05)</td>
<td>0.18</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>(0.05, 0.9, 0.05)</td>
<td>0.11</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>(0.05, 0.05, 0.9)</td>
<td>0.12</td>
<td>0.08</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 6 Effects of different combinations of estimated $w_j$ on $\hat{\gamma}_M^C$, $\hat{\gamma}_C$ and $\hat{\sigma}_C$. The entries in $w$ are in the order: African, Caucasian and Oriental populations.
The effects of estimating $\gamma^M$, $\bar{\gamma}$ and $\sigma_\gamma$ with balanced and unbalanced sample sizes using CATANOVA and AMOVA are shown in Table 7. Using simulated data (1000 iterations) with the specified sample sizes, we computed the final parameter estimates as their mean over the 1000 iterations. The SE of the estimators were computed by taking the standard deviation of their estimates over 1000 iterations. If $\gamma^M$ is of interest, AMOVA returns estimates that are substantially biased upwards compared to CATANOVA, even when the relative sample sizes are the same as the true population weights; note as well the relatively larger SE. Both AMOVA and CATANOVA return estimates that are close, however, if $\bar{\gamma}$ and $\sigma_\gamma$ are of interest.

<table>
<thead>
<tr>
<th>Sample sizes</th>
<th>$\hat{\gamma}^M_C$</th>
<th>$\hat{\gamma}^M_A$</th>
<th>$\hat{\gamma}_C$</th>
<th>$\hat{\gamma}_A$</th>
<th>$\hat{\sigma}_C$</th>
<th>$\hat{\sigma}_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(22, 22, 22)</td>
<td>0.40 ± 0.07</td>
<td>0.47 ± 0.08</td>
<td>0.08 ± 0.01</td>
<td>0.07 ± 0.01</td>
<td>0.19 ± 0.03</td>
<td>0.20 ± 0.02</td>
</tr>
<tr>
<td>(46, 46, 46)</td>
<td>0.38 ± 0.05</td>
<td>0.47 ± 0.06</td>
<td>0.08 ± 0.01</td>
<td>0.08 ± 0.01</td>
<td>0.18 ± 0.02</td>
<td>0.20 ± 0.01</td>
</tr>
<tr>
<td>(46, 10, 10)</td>
<td>0.37 ± 0.06</td>
<td>0.53 ± 0.07</td>
<td>0.07 ± 0.01</td>
<td>0.08 ± 0.02</td>
<td>0.15 ± 0.02</td>
<td>0.19 ± 0.02</td>
</tr>
<tr>
<td>(10, 46, 10)</td>
<td>0.29 ± 0.07</td>
<td>0.42 ± 0.09</td>
<td>0.08 ± 0.01</td>
<td>0.08 ± 0.02</td>
<td>0.20 ± 0.03</td>
<td>0.23 ± 0.02</td>
</tr>
<tr>
<td>(10, 10, 46)</td>
<td>0.30 ± 0.07</td>
<td>0.44 ± 0.09</td>
<td>0.08 ± 0.02</td>
<td>0.08 ± 0.02</td>
<td>0.20 ± 0.03</td>
<td>0.23 ± 0.02</td>
</tr>
<tr>
<td>(33, 16, 17)</td>
<td>0.42 ± 0.07</td>
<td>0.51 ± 0.07</td>
<td>0.07 ± 0.01</td>
<td>0.07 ± 0.01</td>
<td>0.17 ± 0.02</td>
<td>0.19 ± 0.02</td>
</tr>
<tr>
<td>(16, 33, 17)</td>
<td>0.36 ± 0.07</td>
<td>0.44 ± 0.08</td>
<td>0.08 ± 0.01</td>
<td>0.07 ± 0.01</td>
<td>0.19 ± 0.03</td>
<td>0.21 ± 0.02</td>
</tr>
<tr>
<td>(16, 17, 33)</td>
<td>0.36 ± 0.07</td>
<td>0.45 ± 0.08</td>
<td>0.08 ± 0.01</td>
<td>0.07 ± 0.01</td>
<td>0.19 ± 0.03</td>
<td>0.21 ± 0.02</td>
</tr>
</tbody>
</table>

Table 7: Effects of balanced and unbalanced sample sizes on estimators of $\gamma^M$, $\bar{\gamma}$ and $\sigma_\gamma$ (SE attached). The entries in the vector of sample sizes are in the order: African, Caucasian and Oriental populations.

4 Discussion

In this paper, we have focused our analysis on the observed populations only. As argued by Nei (1987), the observed populations are studied precisely because researchers are interested in inferring parameters associated with their unique evolutionary history. Seen in this light, treating the observed populations like replicate populations as in AMOVA seems unnatural - although this idea is probably useful in experimental settings. The assumptions of CATANOVA are basic and do not need to invoke an unfamiliar linear model that uses categorical response variable. Indeed, this approach has permitted the definition of a coherent parameter $\gamma$, which measures allelic, genotypic or haplotypic differentiation between the populations, depending on whether the categories are alleles, genotypes or haplotypes. If between population variation is large, then $p_{ij}$ are very different across different populations, yielding large $\gamma$. Conversely, $\gamma$ is low if between population variation is small.

Adopting the CATANOVA framework has also allowed us to study the relation between $\bar{\gamma}$ and $\gamma^M$, and to assess their usefulness for describing population
genetic structure under stratified random sampling. From a practical point of view, using CATANOVA or AMOVA to estimate these parameters may not matter much, except when sample sizes are very unbalanced. On the basis of simulation results in Section 3.5, we suggest that researchers consider reporting $\bar{\gamma}$ when analysing either mtDNA or nuclear DNA data. Currently, many published results still use “conventional” AMOVA to estimate the $\Phi$-statistics, which corresponds to $\hat{\gamma}^M$ in the simplest nonhierarchical case. While estimation of locus-specific $\gamma$ using AMOVA (known as “locus-by-locus” AMOVA; see Excoffier, 2009) can be done using Arlequin, few researchers use it (see Adeyemo et al., 2005 for an example), and even then, summary statistics such as the mean and standard deviation of the locus-specific estimates are not reported.

Boxplots of the distribution of estimated locus-specific $\gamma$ are useful for visually spotting potentially important outliers. We are aware of one study (Chakraborty et al., 1977) where estimates of $\hat{F}_{ST}$ and $F_{ST}^M$ were compared. However, because the populations under study did not differ much in their allele frequencies, the two estimates were small and differed little. Other researchers could have been tempted to conclude from such studies that the choice of parameters did not matter, and consequently, opted to report $\gamma^M$. As we have shown in Section 3.5, $\hat{\gamma}$ appears to be relatively unchanged when the population weights are arbitrarily estimated using the relative sample sizes, compared to $\hat{\gamma}^M$. In other words, it is the parameter that is easiest to interpret when major populations are analysed together with minor populations.

It might seem that the present work indicates the unhelpfulness of $\gamma^M$. This is true to some extent, but we think important information can still be gleaned from $\gamma^M$ estimates if researchers bear the following caveats in mind. Indeed, because $\gamma^M$ and $\bar{\gamma}$ are correlated (Figure 7), they often lead to the same inference, particularly when sample sizes are not too unbalanced. First, because the loci independence assumption implicit in $\gamma^M$ is likely to be violated, it is not clear how much bias is induced when correlated loci are used to estimate $\gamma^M$. As shown in Section 3.4, doing so can lead to quite different estimates. Second, large changes to $\hat{\gamma}^M$ as the relative sample sizes change (Table 6) imply that conclusions may depend more on the latter than underlying differences in $p_{ij}$. Third, evidence from simulation results (Figure 7) suggests that $\gamma^M$ exaggerates the contribution of population labels to total genetic variation compared to $\bar{\gamma}$, when variance of the relative sample sizes is large.

Although we have elaborated our findings using a simple nonhierarchical CATANOVA framework, an hierarchical extension is straightforward. Thus,
we have \( \text{SST} = \text{SSBR} + \text{SSWR} + \text{SSW} \), where \( \text{SSBR} \) is the SS between regions, \( \text{SSWR} \) is the SS between populations within regions, and \( \text{SSW} \) is the usual SS within populations. Both \( \text{SST} \) and \( \text{SSW} \) are easy to compute using (7) and (8). To compute \( \text{SSWR} \), we just need to apply results of nonhierarchical CATANOVA on each region, and then sum the required SS across regions. The \( \text{SSBR} \) is computed by complementation. The parameters of interest can be similarly found using methods in Section 3.1.

As it becomes increasingly cheaper to sequence DNA (New Scientist, 2009), it will not be long before almost entire genome of individuals are compared on a routine basis. Many more polymorphic sites - probably several orders of magnitude higher than those generated by AFLP, will be found. With sound grounding in theory, the distributional approach using locus-specific \( \gamma \) should provide a powerful means of understanding single nucleotide polymorphisms variation in populations of a species.

5 Summary

In this paper, we have argued that CATANOVA instead of AMOVA should be used as the analysis of variance framework for analysing molecular variation. Under CATANOVA, the sampling model considers populations under study as the subject of interest, and does not need to assume a linear model with categorical response variable. We have shown that the \( R^2 \) measure of CATANOVA is a consistent estimator of \( \gamma \) under simple random sampling. Thus, where alleles are the categories, we link \( \gamma \) to \( G_{ST} \), a widely-used measure of allelic differentiation among populations. More generally, \( \gamma \) is a measure of allelic, genotypic or haplotypic differentiation, depending on whether categories correspond to alleles, genotypes or haplotypes. With multiple loci, two measures: \( \bar{\gamma} \) and \( \gamma^M \), are possible, and we compared the statistical properties of CATANOVA and AMOVA estimators of these two parameters using a simulation approach. Results suggest that \( \bar{\gamma} \) is a much more robust parameter compared to \( \gamma^M \) with regards to two aspects: when stratified random sampling substitutes for the ideal simple random sampling, and when loci independence cannot be guaranteed. As a result, \( \bar{\gamma} \) is much easier to interpret than \( \gamma^M \). Subsequently, we have shown that the RMSE of the CATANOVA estimator of \( \bar{\gamma} \) is smaller than that of AMOVA.
6 Appendix

6.1 Background on $G_{ST}$

Many species of plants and animals do not exist as a single population, but form several more or less distinct populations according to some particular subdivision criterion, such as geographical location (see Hartl and Clark, 1989). The latter is an important determinant of the amount of gene flow that occurs among these populations. In conjunction with genetic drift, sufficient genetic differences may eventually build up between the populations, leading to the emergence of new species. To study this type of evolutionary phenomenon in a species, it is necessary to have a measure of genetic differentiation between the different populations. Wright (1943, 1951) first proposed a parameter known as $F_{ST}$ (see also Nei 1987) for quantifying allelic differentiation between populations. It was first defined for the case of a two-allele locus as

$$F_{ST} = \frac{\sum_{j=1}^{G} w_j (p_j - \bar{p})^2}{\bar{p}(1 - \bar{p})},$$

where $w_j$ is the relative population size (weight) of the $j$th population, $p_j$ is the allele frequency for either one of the two alleles in the $j$th population, and $\bar{p} = \sum_{j=1}^{G} w_j p_j$. If the allele frequencies do not differ much for all $j = 1, 2, \ldots, G$, then $F_{ST}$ is close to zero; otherwise, $F_{ST}$ tends to be large (its upper bound is 1). Nei and Chesser (1983) discussed the estimation of $F_{ST}$ in detail.

A natural extension of $F_{ST}$ to more than two alleles is $G_{ST}$ (Nei, 1973; Nei, 1987), which is defined as

$$G_{ST} = \frac{\sum_{j=1}^{G} \sum_{i=1}^{I} w_j p_{ij}^2 - \sum_{i=1}^{I} \bar{p}_i^2}{\sum_{i \neq m} \bar{p}_i \bar{p}_m},$$

where $i$ indexes the alleles, and $l, m = 1, 2, \ldots, I$. When $I = 2$, $G_{ST} = F_{ST}$. If multiple loci are used, then the distribution of locus-specific $G_{ST}$ is available, and its characteristics can be summarised using the mean and the variance.

6.2 Background on CATANOVA

An analysis of variance (ANOVA) framework for handling categorical data was first proposed by Light and Margolin (1971; see also Bishop et al., 1975), who coined the term CATANOVA (categorical ANOVA). The (univariate)
CATANOVA was originally developed in the context of testing the null hypothesis of homogeneity of proportions among populations. In this regard, it competes with the standard Pearson chi-squared test. An extension of the CATANOVA to multivariate case is given in Pinheiro et al. (2000).

The foundation of CATANOVA rests on an appropriate definition of sum of squares (SS) for categorical data. Gini’s (1912) definition of the SS, which is based on the identity that the SS for \( N \) quantitative measurements (the sample size) is equal to the sum of all squared pairwise differences divided by \( 2N \), proves suitable. The identity is given by

\[
SS = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2, \tag{5}
\]

where \((x_i - x_j)^2\) is 0 if both categorical variables \(x_i\) and \(x_j\) belong to the same category, or 1 if they do not. Equation (5) can be simplified to

\[
SS = \frac{N}{2} - \frac{1}{2N} \sum_{i=1}^{I} N_i^2, \tag{6}
\]

where \(N_i\) is the count of the \(i\)th category. Using this result, Light and Margolin proceeded to define the total SS as

\[
SST = \frac{N}{2} - \frac{1}{2N} \sum_{i=1}^{I} N_i^2, \tag{7}
\]

where \(N_i = \sum_{j=1}^{G} N_{ij}\), with \(j\) indexing the populations, and \(N = \sum_{i=1}^{I} N_i\). The within population SS, by repeated application of (6) to each of the \(G\) populations and then summing them, is

\[
SSW = \frac{N}{2} - \frac{1}{2} \sum_{j=1}^{G} \sum_{i=1}^{I} \frac{N_{ij}^2}{N_j}, \tag{8}
\]

where \(N_j = \sum_{i=1}^{I} N_{ij}\). The between population SS is defined as \(SSB = SST - SSW\), leading to
SSB = \frac{1}{2} \sum_{j=1}^{G} \sum_{i=1}^{I} \frac{N_{ij}^2}{N_j} - \frac{1}{2N} \sum_{i=1}^{I} N_{i+}^2. \tag{9}

Light and Margolin then proposed a test statistic for testing the null hypothesis of homogeneity of proportions among populations,

\[ C = (N - 1)(I - 1) \frac{SSB}{SST} = (N - 1)(I - 1)R^2. \tag{10} \]

Light and Margolin proposed \( R^2 \) as the categorical analogue of the proportion of total variation explained by population labels. The \( C \)-statistic (10) is asymptotically chi-squared distributed with \((I - 1)(G - 1)\) degrees of freedom. Interestingly, under the null hypothesis SST is asymptotically independent of SSB, whereas in standard ANOVA, SSB and SSW are independent. Depending on the alternative hypotheses, tests based on (10) may have more, or less, power than the standard Pearson chi-squared test. Both tests, however, generally lead to the same decision. Margolin and Light (1974) made a comparative study of the small sample properties of these two tests. Note that no linear model is assumed in CATANOVA.

### 6.3 Background on AMOVA

We describe AMOVA in the context of partitioning genetic variation in mtDNA haplotypes, as per Excoffier et al. (1992). It is assumed that the \( G \) populations under study are a random sample from a universe of populations. Therefore, inference is centered on parameters that characterise this universe, rather than the sample at hand. The data come in the form of haplotypes, which can be conveniently represented as vectors of binary variables indicating the presence or absence of restriction enzyme cuts at particular sites in the mtDNA molecule.

Excoffier et al. considered the following hierarchical random effects linear model

\[ X_{tuv} = p + a_t + b_{tu} + c_{tuv}, \tag{11} \]

where \( X_{tuv} \) is the response variable (the haplotype) of the \( v \)th individual in the \( u \)th population in the \( t \)th region, and \( p \) is its mathematical expectation. The random effects due to regions (\( a \)), populations within regions (\( b \)), and individuals within populations (\( c \)), have variance components given by \( \sigma_a^2, \sigma_b^2 \)
and $\sigma_c^2$, respectively. With this model, they proceeded to define the parameters ($\Phi$-statistics)

$$\Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_c^2},$$  

(12)

$$\Phi_{CT} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_c^2},$$  

(13)

$$\Phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}.$$  

(14)

Note that these parameters are not interpreted as the relative contribution of a source (or sources) of variation in the usual sense, but as correlation measures in Excoffier et al. (1992). To estimate the variance components, they proposed a so-called “Euclidean distance metric”,

$$\delta^2_{jk} = (x_j - x_k)'W(x_j - x_k),$$  

(15)

where $x_j$ is the haplotype vector (of length $L$, for $L$ loci) of individual $j$, and $x_k$ of individual $k$. The matrix $W$ assigns differential weights to all $L$ loci. They then defined the SS as

$$SS^* = \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{N} \delta^2_{jk}.$$  

In general, it will not be clear how to specify $W$. Therefore, the latter is commonly equated to the identity matrix $I$. This implies that all variables are assumed independent; equation (15) then simplifies to

$$\delta^2_{jk} = \sum_{i=1}^{L} (x_{ij} - x_{ik})^2,$$

where the summand is 1 if $x_{ij}$ and $x_{ij}$ belong to different alleles (the categories), and 0 otherwise. The corresponding SS is
\[ SS^* = \sum_{i=1}^{L} \sum_{j=1}^{N} \sum_{k=1}^{N} \frac{1}{2N} (x_{ij} - x_{ik})^2. \] (16)

It turns out that (16) is an extension of Gini’s SS to \( L \) independent loci. The special case of \( L = 1 \) leads to (7) if \( N \) is the total sample size; and to (8) if \( N \) is the sample size for the \( j \)th population and (16) is then summed over all \( G \) populations. Explicitly, the SST and SSW in AMOVA are given by

\[ SST^* = \sum_{i=1}^{L} SST(i), \]
\[ SSW^* = \sum_{i=1}^{L} SSW(i), \]

and \( SSB^* = SST^* - SSW^* \).

Despite the categorical nature of haplotype data, Excoffier et al. estimated the variance components using results derived under assumption that the data are continuous. In the simplest case, total variation is partitioned to only two sources of variation: between and within populations \( (\sigma^2_b = 0) \). Subsequently, only one \( \Phi \)-statistic is necessary since (14) is zero, and (12) is equivalent to (13). Thus, \( \Phi_{ST} \) is said to describe the correlation of random haplotypes between populations, relative to that of random pairs of haplotypes drawn from the total population. Table 8 shows a tabulation AMOVA results. The variance components are estimated by equating the means squares to their corresponding expectations, and then solving the resultant equations. These are given by (see Searle et al. 1992)

\[ \hat{\sigma}^2_c = MSW, \]
\[ \hat{\sigma}^2_a = \frac{MSB - MSW}{n''}, \] (18)

where

\[ n'' = \frac{1}{G-1} \left( N - \sum_{j=1}^{G} \frac{N_j^2}{N} \right). \]
The moment estimator of $\Phi_{ST}$ is obtained by substituting $\hat{\sigma}_a^2$ for $\sigma_a^2$, and $\hat{\sigma}_c^2$ for $\sigma_c^2$ in (12). The AMOVA can be extended to diploid nuclear data (Peakall et al., 1995; Michalakis and Excoffier, 1996), with genotypes as categories.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between population</td>
<td>$G - 1$</td>
<td>SSB$^*$</td>
<td>MSB = SSB$^*$ / $(G-1)$</td>
<td>$\sigma_c^2 + n''\sigma_a^2$</td>
</tr>
<tr>
<td>Within population</td>
<td>$N - G$</td>
<td>SSW$^*$</td>
<td>MSW = SSW$^*$ / $(N-G)$</td>
<td>$\sigma_c^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td>SST$^*$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8 The one-way random effects AMOVA table. Abbreviations: degrees of freedom (df); sum of squares (SS); mean square (MS); expected mean square (EMS).

It is clear that the linear model (11) plays an important role in AMOVA - it provides the basis for defining and estimating the $\Phi$-statistics. Its theoretical justification, however, appears to rest on shaky ground. Whereas (11) is well-defined in the case of a continuous response variable, it is unclear when the latter is categorical. Because the “variance components” can take negative values, their interpretation is not straightforward, as in the case of a continuous response variable.

### 6.4 Proof for Proposition 1

It suffices to show that $1 - \hat{\gamma}_A \leq 1 - \hat{\gamma}_C$.

\[
1 - \hat{\gamma}_A = \frac{\hat{\sigma}_c^2}{\sigma_a^2 + \sigma_c^2}
\]

\[
= \text{MSW} \left\{ \text{MSW} + \frac{1}{n''}[\text{MSB} - \text{MSW}] \right\}^{-1}
\]

\[
= \text{SSW} \left\{ \text{SSW} + \frac{N - G}{n''} \left\{ \frac{\text{SSB}}{G - 1} - \frac{\text{SSW}}{N - G} \right\} \right\}^{-1}
\]

\[
= \text{SSW} \left\{ \text{SST} + \frac{N - G}{n''} \left\{ \left( \frac{1}{G - 1} - \frac{n''}{N - G} \right) \text{SSB} - \frac{\text{SSW}}{N - G} \right\} \right\}^{-1}
\]

\[
\leq \frac{\text{SSW}}{\text{SST}}.
\]

if and only if

\[
\left( \frac{1}{G - 1} - \frac{n''}{N - G} \right) \text{SSB} - \frac{\text{SSW}}{N - G} \geq 0,
\]
that is,
\[
\frac{SSB}{SSW} \geq \frac{N(G - 1)}{\sum_{j=1}^{G} N_j^2 - NG},
\]
or equivalently,
\[
F\text{-ratio} = \frac{MSB}{MSW} \geq \frac{N(N - G)}{\sum_{j=1}^{G} N_j^2 - NG}.
\] (19)

\[\square\]

6.5 Proof of Proposition 2

From Proposition 1, it suffices to show that
\[
\frac{2}{N} \left( \frac{SST + \frac{N - G}{n''} \left\{ \left( \frac{1}{G - 1} - \frac{n''}{N - G} \right) SSB - SSW \right\} }{N - G} \right) \]
does not converge to (3), except in the special case where all \(p_{ij}\) are equal. To see this, we note that the term involving SSW in (20) converges to 0 as \(N \to \infty\). It remains to check the second term,
\[
\left( \frac{N - G}{(G - 1)n''} - 1 \right) \frac{2}{N} \frac{SSB}{N} = \left( \frac{1 - G/N}{1 - \sum_{j=1}^{G} (N_j/N)^2} - 1 \right) \frac{2}{N} \frac{SSB}{N}
\]
\[
\to \frac{\sum_{j=1}^{G} w_j^2}{1 - \sum_{j=1}^{G} w_j^2} \left( \sum_{j=1}^{G} \sum_{i=1}^{L} w_j p_{ij}^2 - \sum_{i=1}^{L} \bar{p}_i^2 \right) \neq 0,
\]
where we used (4) in the last step. \[\square\]
6.6 Haplotype Data

<table>
<thead>
<tr>
<th>Morph</th>
<th>Japanese</th>
<th>Korean</th>
<th>Ainu</th>
<th>Aeta</th>
<th>Vedda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57</td>
<td>52</td>
<td>42</td>
<td>33</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
<td>64</td>
<td>48</td>
<td>37</td>
<td>20</td>
</tr>
</tbody>
</table>

The 35 different haplotypes (morphs) observed in Johnson et al.'s (1983) study. The presence of a restriction enzyme cut is indicated as 1, and 0 otherwise.
Morph | African | Caucasian | Oriental
---|---|---|---
1  | 11  | 29  | 32
2  | 14  | 0   | 0
3  | 12  | 0   | 0
4  | 9   | 0   | 0
5  | 8   | 0   | 0
6  | 0   | 6   | 1
7  | 6   | 0   | 0
8  | 2   | 0   | 2
9  | 0   | 0   | 4
10 | 4   | 0   | 0
11 | 0   | 3   | 0
12 | 0   | 0   | 2
13 | 0   | 0   | 2
14 | 2   | 0   | 0
15 | 0   | 1   | 0
16 | 0   | 1   | 0
17 | 0   | 1   | 0
18 | 0   | 1   | 0
19 | 0   | 1   | 0
20 | 0   | 1   | 0
21 | 0   | 1   | 0
22 | 0   | 1   | 0
23 | 0   | 1   | 0
24 | 0   | 1   | 0
25 | 0   | 1   | 0
26 | 0   | 1   | 0
27 | 0   | 0   | 1
28 | 0   | 0   | 1
29 | 0   | 0   | 1
30 | 1   | 0   | 0
31 | 1   | 0   | 0
32 | 1   | 0   | 0
33 | 1   | 0   | 0
34 | 1   | 0   | 0
35 | 1   | 0   | 0

Total  | 74  | 50  | 46

The distribution of counts of different haplotypes (morphs) in the African, Caucasian and Oriental samples. Data from Johnson et al. (1983).
References


