Clustering and Classification Problems in Genetics Through U-Statistics

Gabriela B. Cybis[a], Marcio Valk[a], Sílvia R.C. Lopes [a]

[a]Department of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

ARTICLE HISTORY
Compiled April 7, 2017

ABSTRACT

Genetic data are frequently categorical and have complex dependence structures that are not always well understood. For this reason, clustering and classification based on genetic data, while highly relevant, are challenging statistical problems. Here we consider a versatile U-statistics based approach for non-parametric clustering that allows for an unconventional way of solving these problems. In this paper we propose a statistical test to assess group homogeneity taking into account multiple testing issues and a clustering algorithm based on dissimilarities within and between groups that highly speeds up the homogeneity test. We also propose a test to verify classification significance of a sample in one of two groups. We present Monte Carlo simulations that evaluate size and power of the proposed tests under different scenarios. Finally, the methodology is applied to three different genetic datasets: global human genetic diversity, breast tumor gene expression and Dengue virus serotypes. These applications showcase this statistical framework's ability to answer diverse biological questions in the high dimension low sample size scenario while adapting to the specificities of the different datatypes.

KEYWORDS
Clustering, Classification, U-Statistics, Genetic Data, Statistical Significance, Microarray gene expression data, Homogeneity test

1. Introduction

The last few decades have seen a tremendous rise in the availability and diversity of genetic data, and with it, a marked increase of statistical methods tailored to answer biological questions. Clustering and classification are at the heart of many of these genetic problems. In this paper we explore a model free approach for clustering and classification of genetic data based on U-statistics that leads to alternative ways of looking at these problems. The methods are versatile enough to be applied to a wide variety of genetic problems and adaptable enough to consider the specificities of different datatypes.

Classical inference in this area generally depends on specific modeling assumptions. However, the complexity of genetic data presents a challenge for parametric multivariate analysis techniques. In fact, details of the data generating processes are not always well understood and modeling them might involve a large number of parameters. In this context, Pinheiro et al. [23] propose an alternative method to test group homo-

CONTACT G. B. Cybis. Email: gabriela.cybis@ufrgs.br

geneity based on the Hamming distance, which gives less emphasis to the likelihood function and more to dissimilarity measures. The test statistic is built upon comparisons of these measures between and within groups, and the test does not require homocedasticity assumptions. Asymptotic normality of the test statistic is obtained through properties of U-statistics.

This approach is particularly appealing in the high-dimension low-sample size (HDLSS) scenario prevalent in genetics for not requiring the computationally intensive covariance matrix inversions, which in these cases are also frequently singular [28]. In this scenario, Pinheiro et al. [22] show that these dissimilarity measure based statistics belong to a general class of first order degenerate U-statistics. Furthermore, under the hypothesis of homogeneity, martingale properties are available for this class, allowing for asymptotic results. These asymptotic properties hold, even without assumptions of stochastic independence or homogeneity of the marginal probability laws. Furthermore, in the work by Valk and Pinheiro [31] these tests were adapted to the time series framework. The resulting test statistics are asymptotically Gaussian, both for the independent and identically distributed case, as well as for non-identically distributed groups of time-series under mild conditions. These conditions make it possible to deal with different correlation structures. In this paper we explore this U-statistic clustering framework in the context of genetic data.

We first examine the problem of clustering a set of observations into two groups and assessing their significance. While there are many different clustering methods in common use [11], assessing the significance of a particular clustering, specially in the HDLSS scenario, is still a challenging problem. Suzuki and Shimodaira [30] present the R package pvclust that contains an approach inspired in the bootstrap strategy used in phylogeneics to assess confidence in hierarchical clustering. Maitra et al. [19] assess the significance of a sample clustered through k-means, by assuming that clusters are compact and, after some ellipsoidal local transform, are spherical and similar to other clusters. While both methods are not well suited for HDLSS datasets, Liu et al. [18] present a statistical test of clustering focusing on this environment. Their approach, implemented in the R package SigClust [13], can be applied to any clustering method and has a test statistic built on the ratio of the within cluster variation to the total variation. However, they adopt as the null hypothesis a normality assumption, which can be an issue since rejection of the null may be a simple consequence of non-normal data.

To address the issue of assessing significance in clustering, we propose a U statistics based test that takes as the null hypothesis overall group homogeneity. Additionally, when the Euclidean distance is considered, we present a clustering algorithm that represents a significant speed-up for the homogeneity test. This approach thrives in the HDLSS context, and unlike the other methods is model-free and can be applied to both categorical and quantitative data.

Next, we consider the problem of a new element that must be classified in one of two groups. Kalina [14] reviews classification methods for high dimensional genetic data, highlighting that many of the traditional methods are not well suited for HDLSS. The widely used classification methods, such as linear value decomposition, Bayes classifiers and support vector machine methods [11] measure confidence in this classification by attributing a classification probability/score to the assignment of the sample in each group. As an alternative way of looking at this problem, we propose a statistical test to verify whether the new element's classification in one of the groups is statistically significant.

Finally we explore these results in three applications that showcase the versatility of our methods. In the first application, we resolve small discrepancies between different tree classifications of human populations built on SNP frequency data. In the second one, we improve confidence in classification of a patient tumor subtype based on gene expression data through the classification test, which can lead to more reliable disease prognostics. Finally, we explore the genetic diversity of Dengue virus through sequence data, by finding genetically homogeneous clusters.

The paper is organized as follows: in Section 2 we present the basic notions of U-statistics and U-statistics based tests. The U test for group separation, the group homogeneity test as well as the classification test are in this section. Section 3 presents a Monte Carlo simulation study for the classification test in which we consider different sample sizes and separation degrees between the two groups to estimate size and power of the classification test. This section also contains a simulation study for comparative analysis of power and size of the homogeneity test. Section 4 presents three applications of the methodology, and Section 5 presents discussions of our results.

2. U-Statistics Based Tests

U-statistics were introduced by Halmos [9] and Hoeffding [12] and play an important role in estimation theory. Details on the general theory may be found in Denker [6] and Lee [17]. Particularly in this work we are interested in the class of U-statistics of order 2. For a random sample $X_1, \cdots, X_n$ of size $n \geq 2$ sampled from a distribution F_1, suppose there is a symmetric square integrable function $g(\cdot, \cdot)$, such that $\mathbb{E}(g(X_1, X_2)) \equiv \theta(F_1)$. Then the U-statistics with kernel g, defined as

$$U_n = \binom{n}{2}^{-1} \sum_{C_{n,2}} g(X_{i_1}, X_{i_2}), \tag{1}$$

is an unbiased estimator of $\theta(F_1)$, where the above summation is over the set $C_{n,2}$ of all $(n;2)$ combinations of 2 integers, $i_1 < i_2$, chosen from $\{1, 2, \cdots, n\}$.

Consider a second random sample $Y_1, \cdots, Y_m$ of size $m \geq 2$, drawn independently from a distribution F_2 belonging to the same family of distributions as F_1, and let $\theta(F_1, F_2)$ be an unknown estimable parameter in the Hoeffding's sense (12). Then if there exists a function $d : \mathbb{R}^2 \to \mathbb{R}$, where $\mathbb{E}(d(X, Y)) \equiv \theta(F_1, F_2)$, being a distance function such as the Euclidean one, the parameter $\theta(F_1, F_2)$ will be a functional distance between distributions F_1 and F_2. For multivariate categorical and/or quantitative random variables where, for each i-th sequence, for $i \in \{1, \cdots, n\}$, let $\mathbf{X}_i = (X_{i_1}, \cdots, X_{i_L})'$ be an L-vector and let n be the sample size or the total number of sequences. Let $\theta(F_1^\ell, F_2^\ell)$ be a similar functional of the ℓ-th marginal distribution F_g^ℓ, for $\ell = 1, \cdots, L$ and $g \in \{1, 2\}$. Then assume there exists an order 2 symmetric kernel $\phi(\cdot, \cdot)$ such that

$$\theta(F_1^\ell, F_2^\ell) = \int \int \phi(x_1, x_2) dF_1^\ell(x_1) dF_2^\ell(x_2). \tag{2}$$

Therefore, $\theta(F_1^\ell, F_2^\ell)$ satisfies

$$\theta(F_1^\ell, F_2^\ell) \geq \frac{1}{2} \{\theta(F_1^\ell, F_1^\ell) + \theta(F_2^\ell, F_2^\ell)\},$$

for all F_1, F_2 and $\ell = 1, \cdots, L$. If we assume that $\theta(\cdot,\cdot)$ is a convex linear function of the marginal distributions, this implies that

$$\theta(F_1, F_2) \geq \frac{1}{2}\{\theta(F_1, F_1) + \theta(F_2, F_2)\}, \tag{3}$$

for all distributions F_1 and F_2, where equality sign holds whenever $\mathbb{E}(X) = \mathbb{E}(Y)$.

For our purpose we shall consider two groups, that is, $G = 2$ (although, the theory holds for $G \geq 2$). We shall also consider two multivariate categorical and/or quantitative samples of L-vectors drawn from distributions F_1 and F_2 that are L-dimensional distributions defined on a common probability space. The aim is to test the homogeneity of groups with respect to their diversity measures. The test is based on the functional distance $\theta(\cdot,\cdot)$ as defined in (3), where its sample version is a generalized U-statistics. In this multivariate setup, let $(\mathbf{X}_{g1}, \cdots, \mathbf{X}_{gn_g})$ denote the vector of n_g observations in the g-th group of size n_g, for any $g \in \{1, 2\}$. Therefore,

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}), \tag{4}$$

is the g-th generalized U-statistics, for $g \in \{1, 2\}$, with kernel $\phi(\mathbf{x}, \mathbf{y})$. In others words, $U_{n_g}^{(g)}$ is the estimator of the functional distance based on distances within groups of samples drawn from the distribution F_g, for any $g \in \{1, 2\}$. Similarly, the generalized U-statistics

$$U_{n_1,n_2}^{(1,2)} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(\mathbf{X}_{1i}, \mathbf{X}_{2j}) \tag{5}$$

is an unbiased estimator of $\theta(F_1, F_2)$, and satisfies (3). Pinheiro et al. [23] consider the following sub-group decomposition for the combined sample U_n

$$U_n = \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)} + \frac{n_1 n_2}{n(n-1)}(2U_{n_1 n_2}^{(1,2)} - U_{n_1}^{(1)} - U_{n_2}^{(2)}) = W_n + B_n, \tag{6}$$

where $n = n_1 + n_2$ is the sample size. Pinheiro et al. [22] show that B_n is in the class of degenerate U-statistics (called quasi U-statistics) where the asymptotic distribution is normal with convergence rates L and/or n, even if the assumption of stochastic independence between samples does not hold. Adapting the results in Pinheiro et al. [22] to the context of time series, Valk and Pinheiro [31] develop methods for classification and clustering analysis for stationary time series.

2.1. *U test for Group Separation*

We consider G_1 and G_2 two groups of samples and employ the U test for group separation to assess whether these groups constitute statistically significant separate clusters. Each group is assumed to be homogeneous in distribution. The null hypothesis states that both groups are not separate, coming from the same probability distribution, while the alternative hypothesis states that they are in fact separate groups.

The test statistics for the U test is defined as

$$B_n \quad = \quad \frac{n_1 n_2}{n(n-1)}(2U^{(1,2)}_{n_1 n_2} - U^{(1)}_{n_1} - U^{(2)}_{n_2}), \tag{7}$$

where $U^{(1)}_{n_1}$ and $U^{(2)}_{n_2}$ are U-statistics associated to within group dissimilarities, as defined in (4), and $U^{(1,2)}_{n_1 n_2}$ is the U-statistic associated to between group dissimilarities as defined in (5).

Under few regularity conditions, found in Pinheiro et al. [22], B_n is asymptotically normally distributed. The test statistic compares weighted distances between and within groups. Thus, from property (3), under the null hypothesis, $\mathbb{E}(B_n) = 0$, since all samples are generated from the same distribution. Under the alternative, $\mathbb{E}(B_n) \geq 0$, since distances between groups are expected to be larger than distances within groups. However, due to the fact that the variance of B_n is unknown, we employ a resampling procedure akin to permutation tests to obtain the test statistic distribution under the null hypothesis and to assess the statistical significance (23).

2.2. *Assessing Group Homogeneity*

The main assumption for applying the U test is homogeneity for each group. In order to verify group homogeneity, Valk and Pinheiro [31] employ a combinatorial procedure. For each possible arrangement of all group elements in two subgroups, the U test is applied. If the null hypothesis of group homogeneity is rejected for at least one of the arrangements, then the group is considered non-homogeneous. This procedure can only be applied if the group has at least 4 elements, since we can only consider arrangements where each subgroup has at least two elements.

When testing in-group homogeneity for large group sizes, the number of possible assignments of all n elements in 2 subgroups (that is, $2^{n-1} - n - 1$) soon becomes an important computational issue. To reduce the computational effort of assessing overall group homogeneity we attempt to identify the subgroup configuration that best separates the two groups. That is, if we accept the null hypothesis for the subgrouping with this configuration, then all other arrangements will also necessarily be homogeneous. Thus, with this strategy, we need to apply the U test only once.

Note that under H_0 the statistic B_n is asymptotically normal with zero mean. Therefore, $B_n/\sqrt{\text{Var}(B_n)}$ is asymptotically standard normal and the group configuration that maximizes this function will also have the smallest p-value in the U test. Thus, to test for overall group homogeneity we propose a clustering algorithm that finds the group configuration S_1 and S_2 that minimizes the objective function

$$f(S_1, S_2) = \frac{-B_n}{\sqrt{\text{Var}(B_n)}}, \tag{8}$$

where S_1 and S_2 are the sets of observation indexes in the two groups with, respectively, n_1 and n_2 elements. Here $n = n_1 + n_2$ is the total number of elements in $\Omega = S_1 \cup S_2$. Then the whole group is considered heterogeneous if and only if we reject H_0 in the U test for this configuration.

To evaluate the objective function $f(\cdot, \cdot)$, given in (8), we must estimate $\text{Var}(B_n)$. In the Web Appendix A, when $\phi(\cdot, \cdot)$ is the Euclidean distance, we compute the variance of B_n for the independent and identically distributed case. We show that the variance of B_n under the hypothesis of group homogeneity is given by

$$\mathrm{Var}(B_n) = \frac{n_1 n_2}{n^2(n-1)^2}\left[\frac{2n^2-6n+4}{(n_1-1)(n_2-1)}\right]\sigma^4 = C(n,n_1)\sigma^4, \tag{9}$$

where σ^4 depends only on the covariance structure of the i.i.d. vectors $X_1, \cdots, X_n$ and $C(n, n_1)$ depends only on the overall sample size and number of elements in the first group.

Note, however, that $\sigma^4 = 4\,(\mathrm{vec}(\Sigma))'\,\mathrm{vec}(\Sigma)$, where Σ is the covariance matrix of each vector X and thus has the order of L^2 parameters. For large values of L, directly estimating the variances and covariances between the vectors components is not a feasible strategy to estimate $\mathrm{Var}(B_n)$. We instead employ the bootstrap technique (7) to estimate $\mathrm{Var}(B_n)$ when the size of G_1 is $n_1 = \lfloor n/2 \rfloor$, where $\lfloor x \rfloor$ means the integer part of x, and we explore the relationship between $\mathrm{Var}(B_n)$ for different group sizes. If we have an estimate for the variance of B_n for $n_1 = i$, $\widehat{\mathrm{Var}_i(B_n)}$, then we can compute

$$\widehat{\mathrm{Var}_j(B_n)} = \frac{C(n,j)}{C(n,i)}\widehat{\mathrm{Var}_i(B_n)}, \tag{10}$$

for all group of size j. To optimize (8) we employ the clustering algorithm in Web Appendix A.

The procedure for assessing group homogeneity proposed by Valk and Pinheiro [31] involves applying the U test for all possible group configurations. For large group sizes, when applying this strategy, we must take into account multiple testing issues.

Here, however, we propose a procedure in which only the group configuration with maximum standardized B_n is tested. Thus we consider an approximation of the distribution of B_n maximum under H_0.

If we assume that the B_n's are independent for different group configurations, then the asymptotic cumulative distribution function of the maximum standardized B_n is given by

$$F_{\max}(x) = \mathbb{P}\left(\max\left(\frac{B_n}{\sqrt{\mathrm{Var}(B_n)}}\right) < x\right) = \Phi(x)^\gamma, \tag{11}$$

where $\Phi(\cdot)^\gamma$ is the standard normal cumulative distribution function at the power γ, with $\gamma = 2^{n-1} - n - 1$. If $F_{\max}(x) > 1 - \alpha$, then we reject the null hypothesis of overall group homogeneity with α significance level.

We note that, when $\phi(\cdot,\cdot)$ is not the Euclidean distance, while the variance of B_n is still constant for each group size, it may not be computed by expression (9). Thus, the procedure described for estimating the different variances based on (10) may not be applied, and we must estimate the variance for each group size with a separate bootstrap. The computational effort incurred is still significantly smaller than individual testing. Once the configuration with the maximum standardized B_n is found, then we can carry out the test for the maximum outlined in expression (11), effectively correcting for multiple testings (see Web Apendix A).

2.3. *Classification Test*

Consider the case where groups G_1 and G_2 are in fact dissimilar, as indicated by rejection of H_0 in the U test. We are interested in whether a new sample $\mathbf{X}^*$ would be classified in group G_1 or G_2. Valk and Pinheiro [31] suggest a comparative approach based on statistics B_1 and B_2, where B_1 is the statistics B_n of (7) when the new sample is classified in group G_1, and B_2 is defined likewise. Note that if $\mathbf{X}^*$ is not well classified in G_2, we might expect the statistic B_2 to be smaller than B_n computed without including the new sample, since this increases the distances within group G_2. Thus, if B_1 is larger than B_2, classifying the new sample in group G_1 produces a better grouping in the sense that distances within the groups are comparatively smaller than distances between groups.

While this procedure gives us an empirical criterion for classification, it does not assess statistical significance. We here propose a classification test based on the difference $D = B_1 - B_2$ to verify if the classification of $\mathbf{X}^*$ in group G_1 is statistically significant. Let μ_{B_1} and μ_{B_2} be, respectively, the expected values of statistics B_1 and B_2, then $\mathbb{E}(D) = \mu_{B_1} - \mu_{B_2} \equiv \mu_D$. The null hypothesis states that $\mathbf{X}^*$ belongs to group G_2, and thus the sample arrangement that produces B_2 is better than the one that produces B_1. The alternative hypothesis states that $\mathbf{X}^*$ is correctly classified in group G_1. The null and alternative hypotheses for this new test are given as

$$H_0 : \mu_D \leq 0 \quad \text{versus} \quad H_1 : \mu_D > 0. \tag{12}$$

However, the full distribution of D is not known, hence we employ the bootstrap technique to assess significance. In order to do this, we obtain samples from the distribution of D under the null hypothesis by assuming that $\mathbf{X}^*$ belongs to group G_2. For each bootstrap iteration, we generate group G_1^b by resampling elements of group G_1, and we generate G_2^b and $\mathbf{X}^b$ by separately resampling elements of $G_2 \cup \mathbf{X}^*$ with replacement. We then compute the test statistic D^b based on the resampled groups. The test rejects the null hypothesis if the test statistic D is larger than the $1 - \alpha$ percentile of the bootstraped distribution.

3. Simulation Studies

In this section we present two Monte Carlo simulation studies. In the first one we analyze the size and power of the homogeneity test proposed in Section 2.2. In the second one we analyze the performance of the classification test proposed in Section 2.3.

3.1. *Size and Power of the Homogeneity Test*

We analyse the size and power of the homogeneity test proposed in Section 2.2, to assess weather a group of samples is homogeneous. For these simulations, we consider a simple model in which the samples are sequences of length 50, generated from independent identically distributed standard multivariate normal distributions, and dissimilarities are measured by the Euclidean distance.

To study the size of the test we simulate under the null hypothesis of homogeneity, for varying group sizes $n \in \{10, 20, 40, 60\}$. We consider the homogeneity test which uses the clustering algorithm given in Web Appendix A to find the configuration with maximum normalized U test statistic and then correct for multiple testing through the max test. We compare these results with the approach of Valk and Pinheiro [31] of multiple U tests, and with this same approach corrected for multiple testing through the Bonferroni correction. For large group sizes, it is not feasible to perform all the $2^{n-1} - n - 1$ tests required to assess homogeneity through the U test approach. For this reason, in order to compare our approach to that of Valk and Pinheiro [31], we estimate size and power of these tests by applying them directly to the group configuration with maximum standardized U test statistic. We also present computational times for the Max test and estimated times based on the total number of combinations for the multiple U test approaches.

Table 1 presents the size of the homogeneity test, measured as the fraction of simulations under the null hypothesis for which H_0 was rejected, considering the theoretical $\alpha = 0.05$. We note that the actual sizes of the tests are greatly affected by group size n (even with multiple testing corrections), which is an expected consequence of the combinatorial approach that underlies our concept of homogeneity. As expected, the simple multiple U test approach performs well only for very small groups, if $n \geq 20$ it will almost certainly reject the null hypothesis. For the group sizes considered in this study, both the Bonferroni correction and our max test achieve suitable size control.

To evaluate the power of the homogeneity test we consider a scenario in which the group is divided into two equal sized subgroups with different mean vectors. Table 2 presents the estimated power, computed as the fraction of simulations under H_1 for which we reject the homogeneity hypothesis. As expected, for all tests the power increases with group sizes and separation between groups. Additionally, the uncorrected U test approach has higher power than the other two tests. However, since this test fails to achieve correct type I error probabilities in most scenarios, we would only recommend its use when the group has up to around 10 elements. We also note that the Max test performs slightly better in terms of power than the Bonferroni corrected U test, in all intermediate scenarios. Thus, when the group has over 20 elements, our results favor the use of the max test. Additional factors that favor the use of the max test in this context are the significant computational savings of the clustering algorithm, and the fact that the max test arises naturally as a test for the maximum standardized U test statistic.

3.2. *Size and Power of the Classification Test*

The performance of the classification test proposed in Section 2.3 is affected by several factors. Critical issues are the effect of the sample size of each group and the degree of separation between groups on the power of the test. In order to answer these issues, we perform some simulations.

The classification test can be applied to any type of data for which dissimilarity measures are available. Due to our interest in genetic data, and the wide use of distance methods for DNA sequences, we chose to simulate aligned DNA sequences, in a situation similar to our Dengue application (see Section 4.3). The data simulation emulates the evolution of sequences along phylogenetic trees. We first generate separate coalescent trees for each group (16) and link the trees through their roots with a branch of length τ multiplied by the root hight of the largest tree. The parameter τ is our proxy measure for the degree of separation between groups. We then simulate the evolution of the $n_1 + n_2 + 1$ DNA sequences along the combined tree using the HKY base substitution model (10).

In order to estimate the size of the classification test we generate DNA sequences under the null hypothesis that the sequence being classified $\mathbf{X}^*$ belongs to group G_2. This is done by generating group G_2 with n_2+1 sequences and randomly assigning one as $\mathbf{X}^*$. We performed 1000 simulations under this scheme and applied the separation U test to each one by using the HKY distance (10). In those simulations where there is a significant separation between the n_1 elements of G_1 and the n_2 elements of G_2, we then apply the classification test of Section 2.3 to assess the statistical significance of classifying $\mathbf{X}^*$ in group G_1, at level $\alpha = 0.05$. The size of the classification test is estimated as the proportion of these simulations in which the null hypothesis is incorrectly rejected. The power of the test is assessed analogously, with simulations performed under H_1, in which $\mathbf{X}^*$ belongs to group G_1.

We perform simulations with varying degrees of between group separation, corresponding to values of $\tau \in \{0.001, 0.5, 1, 2, 4\}$, and group sizes ranging in $n_1 \in \{5, 10, 15, 20\}$, with $n_1 = n_2$. All simulated sequences are 1000 bases long, and the overall evolutionary rate is 0.01. Table 3 shows the estimated size and power for all these simulations.

Our simulations show that the size of the test is close to, or smaller than, $\alpha = 0.05$ in all scenarios, even though the data generating process induces a complex dependency structure between the sequences. For very small groups, however, the estimated size of the test is slightly larger than α. Additionally, our simulations show that the test has very large power for almost all simulated scenarios. Furthermore, increasing the number of elements in both groups leads to power increases for almost all situations, even though the size of the test tends to decrease with group size. This is a good indication of test consistency. Moreover, as expected, increasing the separation τ between groups also leads to power increases. As expected, the more separate the groups are, the easier it is to verify that $\mathbf{X}^*$ in fact belongs to group G_1. However, even when τ is extremely low the test has considerably high power for large group sizes. This is at least partially due to the fact that these simulations use the U test to enforce the assumption that groups G_1 and G_2 are in fact separate, and we only consider for power estimation purposes those cases in which the separation assumption is satisfied.

In future work, we shall consider an extensive study of the effects of dissimilarity measure choices on analyses results. It is important to understand if and how different measures may affect performances for the homogeneity and classification tests. In a different context, the work of (4) compared different types of bases substitution models for DNA sequences through the likelihood ratio test. From the asymptotic theory, the authors proposed a low computational cost estimator for the power of the likelihood ratio test.

4. Applications

In order to showcase our methods, we now present three applications to problems of biological classification based on different types of genetic data.

4.1. *Global human genetic diversity*

The Human Genetic Diversity Project (HGDP) is a collaboration that makes publicly available several datasets of human genetic information. We here consider the HGDP 2002 dataset, that contains data for 377 autossomal microsatelite markers in 1056 individuals from 52 populations (26). These data have been previously considered in different studies to assess the evolutionary relationships among the populations (2,26). Through alternative methodologies, the studies produced tree representations of these relationships that agree in broad strokes, but have discrepancies regarding the placement of several populations. We employ our methodology to help resolve some of the points for which the previous studies disagree.

For the dissimilarity measure in this analysis we consider the fixation index F_{ST}, a commonly used differentiation measure in population genetics (20). We compute pairwise F_{ST} values using the R package POLYSAT (3). For visualization purposes Figure 1(a) presents a map produced from multidimensional scaling of the dissimilarity matrix for all populations in the dataset, in which populations are color coded according to the continent of the origin. As expected, even this low-dimensional representation of the genetic data shows some population separation according to geography. In this analysis, we highlight four sets of populations for which there were classification discrepancies between the analyses in Rosenberg et al. [26] and Chen et al. [2]

The first set consists of the East Asian populations represented in Figure 1(b). The trees produced in both Rosenberg et al. [26] and Chen et al. [2] show separation between the populations represented in orange (group A) and in blue (group B), but disagree concerning the placement of the Japanese and Yakut populations. We first test genetic distances between groups A and B to verify if these in fact represent two separate population clusters. We find that the separation between these populations is non-significant (U test p-value of 0.154). Since the group separation assumption is not satisfied, we cannot apply the classification test. However, B_n statistics of 1.56×10^{-4} and 4.31×10^{-4}, respectively, favor placing the Japanese and Yakut in group B.

Figure 1(c) presents a set of Central-South Asian populations in blue (Group C) and a set of European and Middle Eastern populations in red (group D). Both Rosenberg et al. [26] and Chen et al. [2] indicate separation for these groups, but differ regarding the placement of the Kalash and Uygur populations. To verify if groups C and D represent statistically significant clusters we employ the U test and obtain a p-value of 0.01, indicating group separation. Furthermore, we apply the homogeneity test to both groups C and D, and find that they are in fact homogeneous. We then apply the classification test to both populations, and find that both the Kalash and Uygur populations significantly classify in group C (p-value=0.0380 and 0.0180 respectively).

Two other groups that are also reliably separated in both previous studies are the Middle Eastern populations presented in red in Figure 1(d) (group E) and the European populations, presented in blue (group F). However, the studies diverge on the placement of the Mozabites (located in North Africa but here classified as a Middle Eastern population) whose genetic data place among the European populations in the multidimensional scaling map. The U test for genetic separation of groups E and F indicates that these are not significant population clusters (p-value of 0.854). For this reason, we cannot apply the classification test to assess the placement of the Mozabites. However, a B_n statistic of 1.97×10^{-4} favors placement of the Mozabites in group E.

As proof of principle we choose two populations that are clearly separate: the American populations, shown in blue in Figure 1(e) (group G), and the African populations shown in red (group H). As expected, a testing for separation of these groups yields a highly significant p-value of 0.001. Additionally, since the African San population presents a troubling placement in Chen et al. (2014), we test to see in which group it should be classified. Again, our classification test easily places the African San with the other African populations of group H (p-value 0.0360).

4.2. *Breast Tumor Gene Expression Clusters*

Gene expression data have been successfully used to define tumor subtypes in different types of cancer, and these results have been associated to different clinical outcomes [15,25]. Here we analyze the Norway/Stanford dataset from Sørlie et al. [29], that consists of gene expression data measured by DNA microarrays for 534 genes from 122 breast tissue samples. They use machine learning techniques to classify the samples into five clinically relevant tumor subtypes, based on gene expression profiles, and show that subtype association correlates to survival prognostics. These genes constitute an "intrinsic" gene list selected by Sørlie et al. [29] as good candidates for subtype differentiation. Their procedure consists of first selecting a few tumor samples that are archetypes for each cluster, and then training the classification procedure on these cluster seeds. All the 45 samples that do not belong to any cluster seed are classified according to which subtype they fit in better. We apply our classification test to assess whether cluster assignments are statistically significant, potentially improving the confidence in individual prognostics.

In this application the dissimilarity measure that we use is the Euclidean distance based on expression levels for the 534 "intrinsic" genes. In order to apply our methodology we must first verify if the seed samples used to define the clusters in fact constitute distinct homogeneous groups. The five subtypes, named Luminal A, Luminal B, Basal, ERBB2+ and Normal-like have between 10 and 27 seed elements, and were all found to be extremely homogeneous. This was assessed using the homogeneity test with the clustering algorithm speed-up and max test correction for multiple testing. For the clusters with larger seed groups, the speed-up of the classification algorithm is paramount to the applicability of the homogeneity test of Section 2.2, since in order to apply the test directly to a group of 27 elements we would need to test 67,108,836 different configurations. Given the homogeneity of all seed groups, we apply the U test to the 10 pair comparisons between the five groups and verify that all groups are in fact separate (with p-values < 0.002). Therefore, all assumptions of the classification test are satisfied.

We now wish to verify if the remaining 45 samples, that do not constitute any cluster seed, can be significantly classified in one of the five clusters. Our classification test can only verify if the classification of a sample is statistically significant when comparing two distinct groups. Since we want to assess significance of classification in one of the 5 groups, we adopt the following heuristic procedure. We first compute the centroid gene expression values for each cluster; then we verify which are the two centroids that are closer to the sample that must be classified. This sample should be classified in the group which has the closest centroid. To assess significance of this classification, we apply the classification test considering the two groups with closest centroids.

Of the 45 samples not assigned to any cluster seed, the classification of 10 was considered statistically significant (with significance level $\alpha = 0.01$): 7 were classified in the Luminal B cluster and 3 in the ERBB2+. None of the significantly classified samples were assigned to the Luminal A, Basal and Normal-like clusters.

In order to evaluate the groups defined by significantly classified samples we perform a survival analysis on different sample groups. Figure 2 presents the Kaplan-Meier analysis of relapse times, when dividing samples into 5 clusters. First, in Figure 2(a), we consider only the samples that constitute the cluster seeds, selected for being typical examples of each subtype. Then, in Figure 2(b), we include the seed samples and those whose classification was considered statistically significant. Finally, in Figure 2(c), we consider the full dataset, classifying all non-seed samples according to proximity to the cluster centroids. We note that separation of the survival curves was improved when considering only the significantly classified samples, in comparison to the full dataset. Additionally, the inclusion of the significantly classified samples had a relatively small effect in the p-value of subgroup survival curve separation when compared to the seed samples alone. Furthermore, the grouping considering significantly classified samples achieved better separation of Luminal B and ERBB2+ survival curves. This indicates that our method only classifies samples that are well within the patterns of each group, obtaining group separation similar to that of only the benchmark samples.

We note, however, that while Sørlie et al. [29] employ a complex machine learning procedure to select the optimal genes for this classification in the specific dataset, we simply consider the whole “intrinsic” gene list (the starting point for their analysis). For this reason, the details of our classification will differ. Our objective here was not to provide a final classification, but to show the usefulness of our methodology in refining the groups. Given a refined list of genes, or a set of weights for the different genes, it is straightforward to adapt our analysis.

4.3. *Dengue virus serotypes*

In recent years, Dengue virus has become a serious epidemiological problem in the Americas, infecting over 2 million people in 2015 alone, it is distributed in almost all countries of the continent (21). Viral sequence data have been used, in a variety of scenarios, to study temporal, geographic and demographic aspects of rapidly evolving pathogens, such as Dengue virus (5,24). Here we analyze the genetic variability of the virus in the Americas between the years of 2007 and 2008 by considering 144 RNA sequences from Allicock et al. [1] sampled in that period. Our purpose in this application is not to map the whole genetic diversity of the virus (for which we would need to consider a wider range of sequences and temporal sampling), but to showcase our methods by identifying clusters of homogeneous genetic variation within the 2007 - 2008 viruses. For this analysis, we consider the HKY distance, which is built upon base substitutions, and differentiates between transition and transversion mutations (10).

Dengue virus has 4 phylogenetically separate serotypes DENV1 - DENV4, all represented in this sample. This is clearly reflected in the heat map of sequence distances between all samples (see Figure 3), which presents 4 clear blocks, one for each serotype. Accordingly, when we apply the homogeneity test to the whole dataset, we obtain a p-value of 0 (up to numerical precision of the Gaussian approximation), indicating a highly heterogeneous group. Additionally, pairwise U tests for group separation indicate that all serotypes in fact constitute distinct groups.

Of more interest is the structure of genetic variance within each group. Figure 4 presents Neighbour-Joining trees (27) for each of the serotypes. Applying the homogeneity test to the sequences of each individual serotype, we verify that all serotypes are composed of heterogeneous sequences ($\alpha = 0.05$).

Through Figure 4-DENV1 we identify three main subgroups in Serotype 1. We applied the homogeneity test to each of the individual subgroups, and only the subgroup composed of sequences from Brazil (in brown) was considered homogeneous (p-value = 0.8271). Even when we remove the 2008 Nicaraguan sequence that stands out in the left group of the DENV1 tree, the group composed mainly of Mexican and Nicaraguan sequences still tests heterogeneous.

The tree for serotype 2 shows that the sequences of DENV2 for the 2007 - 2008 period are divided into two subgroups, none of which is homogeneous according to the homogeneity test. If we remove the 2007 Nicaraguan sample that stands out in the tree, the group in the upper part of the Figure 4-DENV2, composed mainly of Brazilian, Venezuelan and Colombian viruses seems to be divided into two obvious subgroups. However only the blue group of Brazilian and Puerto Rican sequences was considered homogeneous (p-value = 0.2500).

In serotype 3, we identify three major subgroups, however only the green group composed of Brazilian and Argentinian viruses was considered homogeneous (p-value = 0.8340). At the right side of Figure 4-DENV3, there is a group of Nicaraguan viruses that are genetically very similar, but when we applied the homogeneity test to this group, it was considered heterogeneous. This highlights the fact that homogeneity is not merely a result of small distances, but a property of the distance distributions.

Finally, the DENV4 tree divides its viruses into two subgroups. The homogeneity test confirms that both the Peruvian group (purple) and the Venezuelan group (orange) are homogeneous (see Figure 4-DENV4). Additionally the U test for group separation indicates that they are in fact distinct groups (p-value < 0.001).

We have analyzed the genetic variation of Dengue virus between 2007 and 2008 in the Americas, and identified five homogeneous subgroups. As expected, the viruses tend to cluster according to geographical location. We also uncovered a curious pattern in which the Brazilian sequences for serotypes 1 - 3, are all part of homogeneous groups.

5. Discussion

In this paper we explore U statistics based methods to solve clustering and classification problems for genetic data in different biological settings. We propose a classification test for verifying whether the assignment of an individual data point to one of two groups is in fact significant. Additionally, we propose a test to assess group homogeneity, focusing on computational efficiency. Finally, to showcase their versatility, we apply these techniques to three biological problems in which we address distinct clustering and classification questions using different types of genetic data.

Through the applications we exemplified how our methodology can be used in different settings. For each dataset, we considered the appropriate dissimilarity measure according to the peculiarities of the individual biological problems. First, in the global human genetic diversity application of Section 4.1, we explored small discrepancies between conflicting hierarchical classifications of human populations by assessing significance of group separation. Then, in the breast tumor application of Section 4.2, we sought to increase confidence in genetically based patient prognostics by assessing significance of tumor subtype classification. Finally, in the Dengue application of Section 4.3, we examined the genetic diversity of the virus to identify clusters of genetically homogeneous strains. The versatility of these methods is in large a consequence of their small reliance on distributional assumptions and their flexibility in considering different dissimilarity measures.

All applications and simulations performed in this paper deal with cases in which the data dimension is larger than the sample size. While the HDLSS scenario is troublesome for many statistical methods, our U statistics based methodology thrives in the $L >> n$ context prevalent in genetics.

In Section 2.2 we present the max test, a homogeneity test based on the approach of Valk and Pinheiro [31]. This is a test for the maximum of the standardized U test statistic over the set of all possible subgroupings, and it arises to control for multiple testing. Additionally, when the Euclidean distance is used, we explore the theoretical variance of the U test statistic to build a clustering algorithm which gives significant computational time savings. Through simulations, we established that the max test adequately controls the type I error as the number of elements in the group increases. Furthermore, we note that, for larger group sizes, the test achieves adequate power, and we thus recommend its use for homogeneity testing with around 20 samples or more. Moreover, when groups reach around 40 or more samples our approach allows for homogeneity testing, since it is not computationally feasible to carry out the alternative multiple U test procedure of Valk and Pinheiro [31]. For smaller group sizes, the overall type I error of the uncorrected multiple U test approach is not largely affected by multiple testing, and should be preferred due to its larger power.

One use of the clustering algorithm developed for the homogeneity test, which we did not explore in this paper, is the clustering of data into two optimal groups. Although this procedure shares conceptual similarities with k-means clustering ($k = 2$), it produces quite different results since it aims to simultaneously minimize within group distances and maximize between group distances.

The Dengue application of Section 4.3 highlights the fact that our concept of homogeneity is not merely a result of small distances between the samples, but a property of the distance distributions. Thus, our method for finding genetically homogeneous groups could be applied to the study of early stages of adaptive radiation, situation in which a group of organisms diversifies very rapidly, which may lead specific evolutionary structures (8). These methods could also be employed in questions regarding the determination of biological species based on genetic variability.

In Section 2.3 we explore the classification criterion of Valk and Pinheiro [31] for classifying a sample $\mathbf{X}^*$ into one of two groups to build a classification test. We employ the bootstrap to assess significance of 2-way classification by comparing the U test statistic B_n, computed with $\mathbf{X}^*$ classified in group G_1, with B_n when $\mathbf{X}^*$ is classified in group G_2. This method is tailored for a situation in which we have two reference groups, and does not naturally extend to settings with more groups, such as the one presented in the breast cancer application of Section 4.2. The choice of the heuristic group centroid procedure for that 5-way classification problem reflects the centroid based algorithm of Sørlie et al. [29]. However, we were not able to verify that this procedure satisfies some desirable properties in 5-way classification. For instance, it is not clear that if we apply the classification test for some pair of the 5 groups and the new sample is significantly classified in one, then it will also be significantly classified in the group with closest centroid. This is mainly due to considerations of different group sizes. However, our heuristic procedure presents a method for assessing statistical significance based on a 2-way classification test that is closely related to the original problem. In order to address these types of problems formally, an n-way classification test based on U statistics should be subject of future work.

Most statistical methods in genetics use simplifying assumptions on the data generating processes, and their impacts on the analyses are not always clear. In contrast, the U statistics model free approach that we employ here assumes only that all samples belonging to a group come from the same distribution, relying on no further marginal distributional assumptions.

In particular, the genetic data dependency structure can be a critical modelling issue. Correlation of genetic data within an individual genome can be a consequence of genetic linkage and functional constraints, while correlations between samples can arise from evolutionary relatedness. In general, these processes are not completely mapped out, and most statistical genetics methods make strong simplifying assumptions regarding these dynamics whenever they are not the focal points of the analyses, since explicitly modeling them can often be prohibitive. The impacts of such assumptions are not always clear. The non-parametric bootstrap approach that we employ for the U test and the classification test implies that most of our methodology is robust to dependency assumptions, as illustrated by the simulations of Section 3.2. However, the asymptotic normality of the test statistic established in Valk and Pinheiro [31] depends on independence between samples and the particular choice of dissimilarity measure. This result underlies our reasoning for the max homogeneity test. Moreover, our clustering procedure (see Section A.1), that largely accelerates the homogeneity test, is built upon observations for the variance of B_n derived under the Euclidean distance as well as between sample independence assumptions. However, at this point, the robustness of our homogeneity test to deviations from this fairly common independence assumption has not been fully quantified. This will be the subject of future work.

Software

R codes are available at
https://github.com/gcybis/UStatistics_ClusteringAndClassification_Biosequences

Supplementary Material

Supplementary material is available online at *http://some_address*.

Funding

S.R.C. Lopes research was partially supported by CNPq-Brazil and by INCT *em Matemática*.

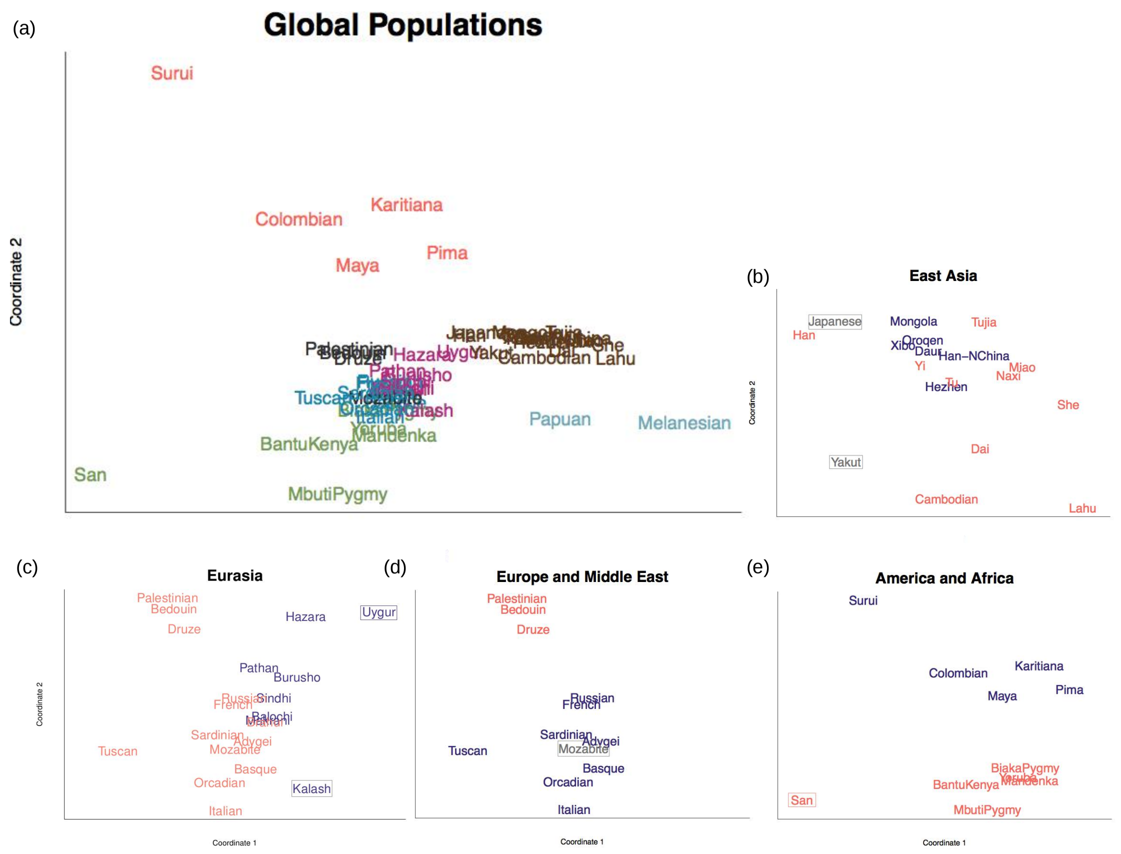

Figure 1. Multidimentional scaling maps from dissimilarity matrix. In (a) all populations of the HGDP 2002 dataset, color coded according to region of origin: red for Americas, green for Africa, dark blue for Europe, black for Middle East, pink for Central-South Asia, brown for East Asia and light blue for Oceania. In (b) East Asian populations from group A (red) and B (blue). In (c) Eurasian populations from groups C (blue) and D (red). In (d) European and Middle Eastern populations from groups E (red) and F (blue). In (e) American and African populations from groups G (blue) and H (red). Populations whose classifications were tested are highlighted by boxes, and those in gray did not yield significant p-values.

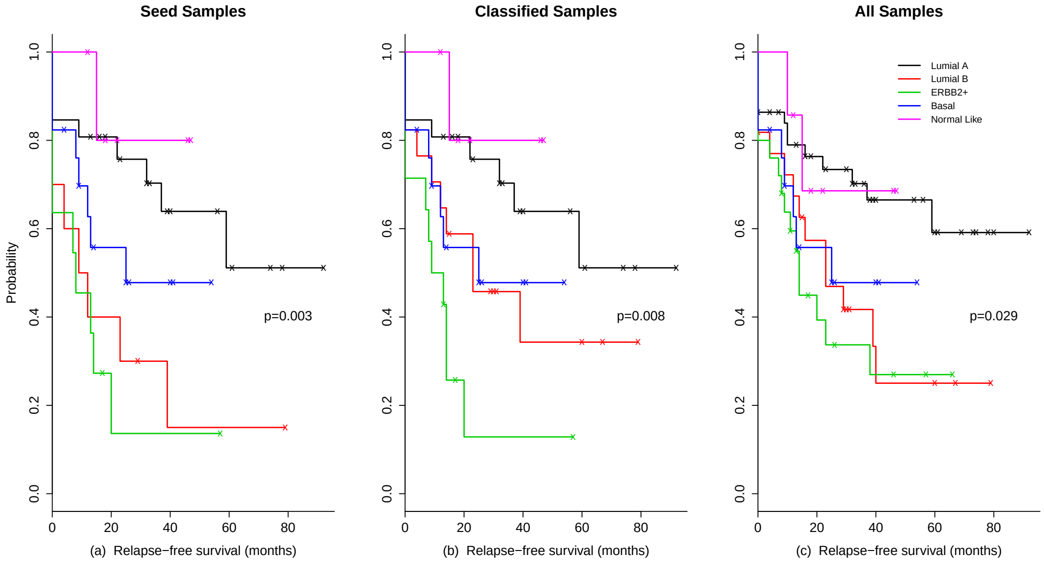

Figure 2. Kaplan-Meier analysis of relapse time. Comparing relapse times for the different clusters considering (a) only the samples in the cluster seeds; (b) cluster seed samples combined with those whose classification was considered statistically significant; (c) the full dataset.

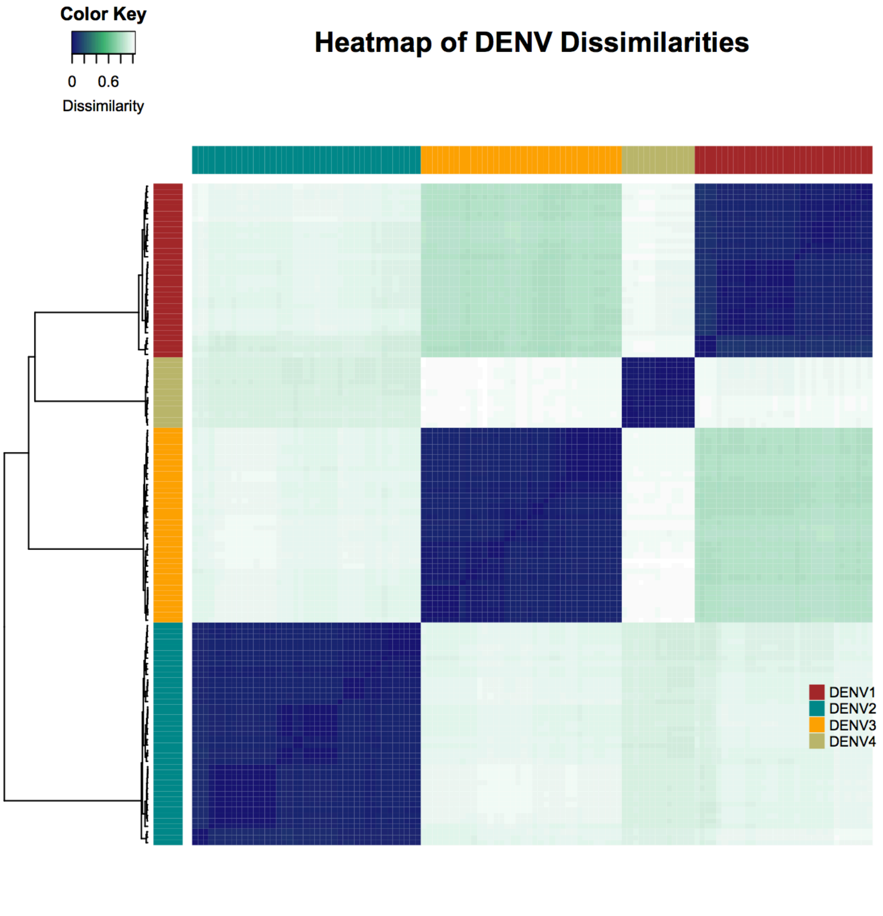

Figure 3. Heatmap of HKY distances between all DENV sequences. Side colors indicate the serotype of each sequence.

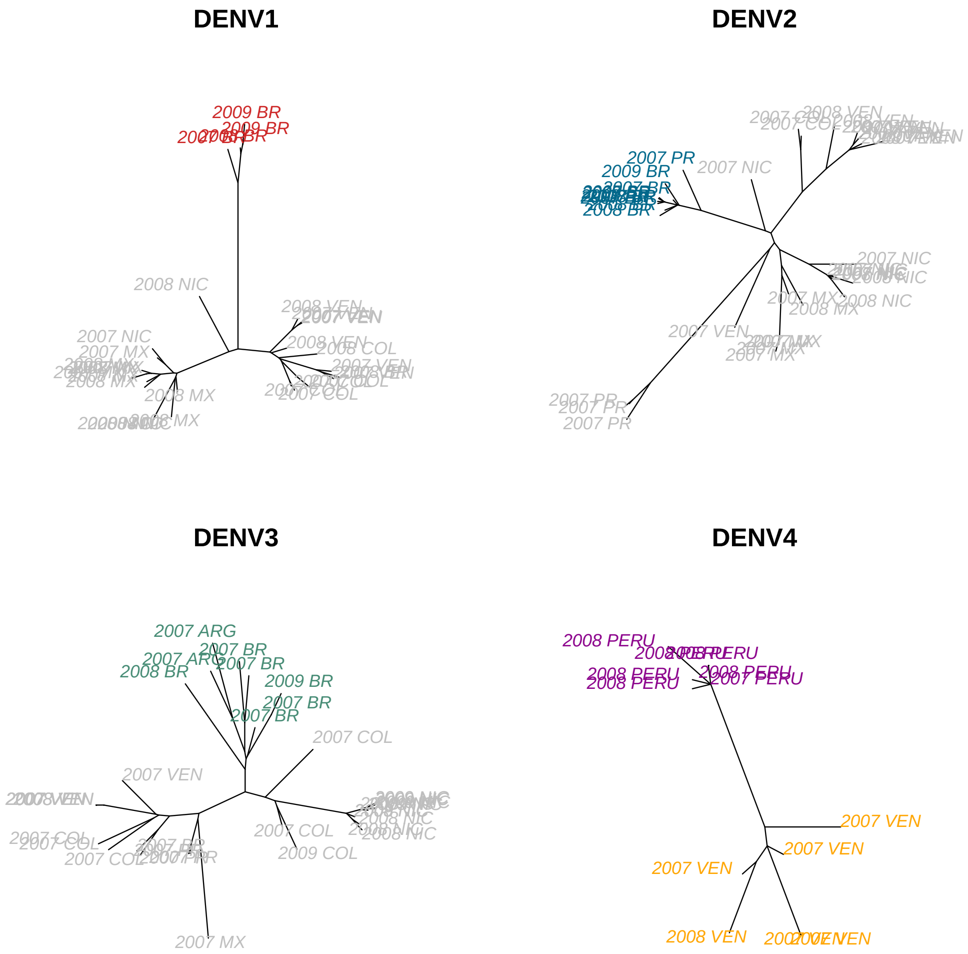

Figure 4. Neighbour-Joining tree for each Dengue virus serotype (DENV1-DENV4). Sequences are labelled according to year and country of isolation. Coloured (non-grey) labels indicate homogeneous clusters (p-value > 0.05).

Table 1. Size and computational time (in seconds) of the homogeneity test for different group sizes n.

	U Test			Max Test	
n	uncorrected size	Bonferroni size	time	size	time
10	0.022	0.000	69.30	0.000	3.45
20	0.997	0.000	2.38×10^5	0.000	11.44
40	1.000	0.003	9.15×10^{11}	0.003	44.47
60	1.000	0.060	2.15×10^{18}	0.079	103.20

Table 2. Power of the homogeneity test for different group sizes n_1 and n_2 and different group separation.

$n_1 \times n_2$	U Test		Max Test
	uncorrected	Bonferroni	
$\mu_1 = 0, \quad \mu_2 = 0.33$			
5 × 5	0.034	0	0
10 × 10	0.980	0	0
20 × 20	0.999	0.018	0.020
30 × 30	1.000	0.415	0.465
$\mu_1 = 0, \quad \mu_2 = 0.66$			
5 × 5	0.537	0	0.001
10 × 10	0.995	0.386	0.445
20 × 20	1.000	0.998	1.000
30 × 30	1.000	1.000	1.000
$\mu_1 = 0, \quad \mu_2 = 1$			
5 × 5	0.994	0.312	0.398
10 × 10	1.000	0.998	1.000
20 × 20	1.000	1.000	1.000
30 × 30	-	-	-

Table 3. Estimated size and power of the classification test for varying degrees of separation τ between groups.

Group sizes	τ									
$n_1 \times n_2$	0.001		0.5		1		2		4	
	Size	Power	Size	Power	Size	Power	Size	Power	Size	Power
5×5	0.058	0.450	0.066	0.842	0.071	0.905	0.074	0.927	0.079	0.913
10×10	0.052	0.946	0.043	0.999	0.049	0.999	0.035	1.000	0.054	1.000
15×15	0.026	0.986	0.052	0.998	0.038	1.000	0.028	1.000	0.042	1.000
20×20	0.035	0.997	0.037	0.999	0.035	1.000	0.035	1.000	0.030	1.000

References

[1] Allicock, O. M., Lemey, P., Tatem, A. J., Pybus, O. G., Bennett, S. N., Mueller, B. A., Suchard, M. A., Foster, J. E., Rambaut, A., and Carrington, C. V. (2012). Phylogeography and population dynamics of dengue viruses in the americas. *Molecular Biology and Evolution*, 29(6):1533–1543.

[2] Chen, G. K., Chi, E. C., Ranola, J. M. O., and Lange, K. (2015). Convex clustering: an attractive alternative to herarchical clustering. *PLoS Computational Biology*, 11(5):e1004228.

[3] Clark, L. V. and Jasieniuk, M. (2011). Polysat: an r package for polyploid microsatellite analysis. *Molecular Ecology Resources*, 11(3):562–566.

[4] Cybis, G. B., Lopes, S. R., and Pinheiro, H. P. (2011). Power of the likelihood ratio test for models of dna base substitution. *Journal of Applied Statistics*, 38(12):2723–2737.

[5] Cybis, G. B., Sinsheimer, J. S., Lemey, P., and Suchard, M. A. (2013). Graph hierarchies for phylogeography. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1614):20120206.

[6] Denker, M. (1985). *Asymptotic Distribution Theory in Nonparametric Statistics.* Braunschweig: Springer.

[7] Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap. *Monograps on Statistics and Applied Probability*, 57.

[8] Gavrilets, S. and Losos, J. B. (2009). Adaptive radiation: contrasting theory with data. *Science*, 323(5915):732–737.

[9] Halmos, P. R. (1946). The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43.

[10] Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22(2):160–174.

[11] Hastie, T., Tibshirani, R., and Friedman, J. (2013). The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics (*.

[12] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.

[13] Huang, H., Liu, Y., Yuan, M., and Marron, J. S. (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993.

[14] Kalina, J. (2014). Classification methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, 34(1):10–18.

[15] Kapp, A. V., Jeffrey, S. S., Langerød, A., Børresen-Dale, A.-L., Han, W., Noh, D.-Y., Bukholm, I. R., Nicolau, M., Brown, P. O., and Tibshirani, R. (2006). Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(1):1.

[16] Kingman, J. F. (1982). The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248.

[17] Lee, J. (1990). *U-statistics: Theory and Practice.* New York: Marcel Dekker.

[18] Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. (2008). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293.

[19] Maitra, R., Melnykov, V., and Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, 107(497):378–392.

[20] Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323.

[21] PAHO/WHO (2016). Number of reported cases of dengue and severe dengue (sd) in the americas, by country: Figures for 2015 (to week noted by each country). *http://www.paho.org/*.

[22] Pinheiro, A., Sen, P. K., and Pinheiro, H. P. (2009). Decomposability of high-dimensional diversity measures: Quasi–statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, 100(8):1645 – 1656.

[23] Pinheiro, H. P., Pinheiro, A., and Sen, P. K. (2005). Comparison of genomic sequences using the hamming distance. *Journal of Statistical Planning and Inference*, 130(1):325–339.

[24] Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071.

[25] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154.

[26] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

[27] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

[28] Sen, P. K. (2006). Robust statistical inference for high-dimensional data models with application to genomics. *Austrian Journal of Statistics*, 35(2/3):197–214.

[29] Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423.

[30] Suzuki, R. and Shimodaira, H. (2006). Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542.

[31] Valk, M. and Pinheiro, A. (2012). Time-series clustering via quasi u-statistics. *Journal of Time Series Analysis*, 33(4):608–619.