Machine learning methods for genotype assignment

Jeremy Georges-Filteau

Master of science

McGill School of Computer Science

McGill University, Montreal, Canada

July 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Science, Computer Science

Acknowledgments

I consider myself quite privileged to have done my master's work under the supervision of Prof. Mathieu Blanchette. His genuine interest, positivity and insightful conversations are uplifting and intellectually stimulating. It is always inspiring to work with someone who clearly cares for what he does and about those under his guidance. Most importantly, through his mentorship I have gained confidence and pride in my ability to conduct and share research.

Thanks to Prof. Richard Hamelin for providing the S.musiva data and Prof. Roger C. Lévesque for providing the gipsy moth data.

Thanks to Genome Canada and the Biosafe project for the funding making this research possible.

Thanks to everyone in the Computational Genomics Lab for their everyday help and interesting discussions, making for a great work environment.

Finally, thanks to my parents and my partner Jeisson for their support, encouragement, patience and love.

Abstract

Invasive species are an ongoing concern for countries in which natural resources play a vital economic and social role. In Canada, species such as the Asian long-horned beetle, Dutch elm disease, sudden oak death and the Asian gypsy moth threaten forests and the sectors of industry that profit from them. The economic risk is estimated at up to $800M annually. Machine learning methods that quickly and accurately determine the taxon, geographic origin, and pathogenic fitness of biological samples from genomics data would constitute a valuable tool for risk reduction.

In this thesis, we reviewed concepts of population genetics, phylogenetic networks, genotype data and current methods for genetic population assignment. Having identified a number of the shortcomings of current methods, we propose a new machine learning approach called Mycorrhiza aimed at predicting the geographical origin of a sample from its genotype in which phylogenetic networks are used as feature engineering tools, followed by a Random Forests classifier. The classification accuracy of our method was compared to widely used assessment tests or mixture analysis methods in population genetics such as STRUCTURE and Admixture, as well as a variant where a PCA is used in place of the phylogenetic network. Multiple published SNP, microsatellite or consensus sequence datasets with wide ranges in size, geographical distribution and populations were used for this purpose.

The phylogenetic network and PCA methods show a marked improvement in classification accuracy and definable advantages compared to the existing approaches.

As is to be expected, STRUCTURE and Admixture fall short on almost all datasets with a considerable deviation from the Hardy Weinberg equilibrium. The same can be said for Admixture on datasets with a large expected heterozygosity. Moreover, Mycorrhiza consistently estimates mixture proportions more accurately than the PCA variant. Our approach will be useful in the rapid and accurate prediction of geographical origin from genotype samples without the restrictions inherent to currently used methods.

Résumé

Les espèces invasives sont une préoccupation constante pour les pays dans lesquels les ressources naturelles jouent un rôle économique et social essentiel. Au Canada, des espèces telles que le longicorne asiatique, la maladie hollandaise de l'orme, la mort subite du chêne et la spongieuse asiatique menacent les forêts et les secteurs industriels qui en profitent. Le risque économique est estimé à 800 millions de dollars par année. Des méthodes d'apprentissage machine qui déterminent rapidement et précisément le taxon, l'origine géographique et la capacité pathogène des échantillons biologiques à partir de données génomiques constitueraient un outil précieux pour la réduction des risques.

Dans cette thèse, nous avons examiné des concepts de la génétique des populations, des réseaux phylogénétiques, des données génotypiques et des méthodes actuelles d'attribution de population. Après avoir identifié un certain nombre de lacunes des méthodes actuelles, nous proposons une nouvelle approche d'apprentissage machine nommée Mycorrhiza visant à prédire l'origine géographique d'un échantillon à partir de son génotype dans laquelle des réseaux phylogénétiques sont utilisés comme une étape de transformation des données suivie d'un classificateur forêt d’arbres décisionnels. La précision de classification de notre méthode a été comparée à des méthodes d’assignation génétique ou d'analyse de mixture largement utilisées en génétique des populations, telles que STRUCTURE et Admixture, ainsi qu’à celle d'une variante où une analyse en composantes principales est utilisée à la place du réseau

phylogénétique. Des jeux de données de SNP, de microsatellites ou de séquences consensus publiées avec de larges gammes de taille, distribution géographique et de populations ont été utilisées à cette fin.

Le réseau phylogénétique et les méthodes APC montrent une nette amélioration de la précision de classification et des avantages définissables par rapport aux approches existantes. Comme on peut s'y attendre, STRUCTURE et Admixture échouent sur presque tous les jeux de données avec un écart moyen important par rapport à l'équilibre de Hardy Weinberg. La même chose peut être dite pour Admixture sur des ensembles de données avec une grande hétérozygotie attendue. De plus, l'approche du réseau phylogénétique estime les proportions du mélange avec systématiquement plus de précision que la variante APC. Enfin, l'approche phylogénétique gagne en précision par rapport à la variante APC lorsque les prédictions résultant de multiples sous-ensembles ordonnés des données sont moyennées.

Contribution of authors

The candidate designed and implemented the algorithms, performed the computational analyses and wrote the entire thesis under the supervision of Prof. Mathieu Blanchette. He should be considered the primary author.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Invasive species are of growing concern for countries in which natural resources play a vital economic and social role [1]. Their ecological and economic impact has been known for many decades and the expansion of global trade only aggravates the problem [2]. In Canada, where forests are an important portion of industrial and touristic revenue, species such as the Asian long-horned beetle, Dutch elm disease, sudden oak death and the Asian gypsy moth pose an immediate threat to the economy and jobs. The economic risk is estimated at up to $34.5 billion CND annually [3]. In the United States of America this also ranges in the billions [1,2]. The implementation of countermeasures is indispensable to protect not only the prosperity of a large portion of the population, but also an invaluable natural wealth.

Alien species enter Canada through a multitude of anthropogenic and natural vectors. In some cases, containment measures can be put in place by the appropriate government agencies if the taxon, geographical origin and introduction pathway are determined rapidly. However, current methods of risk assessment relying on phenotypic traits alone are highly inadequate for this purpose as different strains of the same species can be indistinguishable in appearance. Moreover, these can be highly dependent on sex or environmental pressures [4]. Methods based on genomics are currently too slow or costly for large-scale implementation. The development of rapid

and inexpensive genomic methods accurate in the determination of taxon, geographic origin, and pathogenic fitness would therefore be a valuable tool for risk reduction. The BioSAFE project aims to address these goals by implementing a risk mitigation system on an unprecedented scale. Central to this effort is the development of machine learning methods for the rapid and accurate classification of outbreak samples from genomics data.

Within the context of the BioSAFE project, the goal of this research was to develop and/or evaluate the accuracy of various machine learning methods for predicting the geographic origin of specimens from their genotype. In the field of population genetics the tools for this are referred to as assignment tools or methods [5]. Accordingly, the tools and algorithms designed for this purpose are referred to as assignments tests or assignment methods.

We shall consider for our study the use of existing population genetics tools, weigh their strengths and shortcomings, examine their implementation from a machine learning point of view and explore avenues for new algorithms based on our findings. As we are interested in ancestry and, consequently evolutionary events happening on a relatively short timescale, reticulate events such as hybridization, recombination and lateral gene transfer must be taken into account throughout our analysis and algorithm development.

1.2 Problem definition

The problem is defined as follows. We are given a number of biological samples for which the genotype and population of origin, or geographical sampling location, is known. Let M be the number of loci forming the genotype, K the number of discrete

populations we wish to account for, and U the number of sampled individuals of known origin. We wish to train a machine learning model from the genotype and origin of the U known samples. This model will take as input the genotype of a sample of unknown origin and output a probability distribution L over the K populations.

Chapter 2

Literature review

2.1 Population genetics

Population genetics is a branch of evolutionary biology that studies the genetic composition of populations at many different scales [6,7]. The theory and methods established in the field serve to answer a wide range of biological questions in a number of different contexts and generalization levels.

2.1.1 Populations

Consequently, a number of different definitions are used for the term “population”, some of which are subjective [6,8]. In fact, as we shall later consider, these can sometimes be purely arbitrary. Additionally, linguistic, cultural or physical characters can also be used as measures of differentiation or similarity between individuals to infer populations.

In general, a distinction can be made between statistical definitions of populations, which simply refer to an aggregate of entities from which we draw samples and make inferences, and biological definitions, which refer to collections of individuals that share genomic or phenotypic attributes [6]. For the purpose of this research, we are interested in geographical sampling location of specimens and thus view the populations as aggregates delimited by boundaries we have yet to define, such as country or state borders. We will later discuss how these boundaries also correspond to populations in the genetic sense and both definitions can be used interchangeably for our purpose. Furthermore, we are mainly concerned with population observed on an

ecological timescale, rather than an evolutionary one, on the order of one to a thousand generations [5,6,9].

To avoid ambiguities, unless specified, we will use the term “origin” to refer to geographical origin, in other words “the totality of individual observations about which inferences are to be made, existing within a specified sampling area limited in space and time” and the term “natural population” to refer to a population that “can only be bounded by natural ecological or genetic barriers” [6]. Therefore, the term “population” alone will be used in a more general sense and will refer to both concepts previously stated. Additionally, the term “specimen” will refer to a single individual and the term “sample” to a set of specimens.

We shall now give an overview of a few concepts in populations genetics such as hybridization, introgression, heterozygosity and indices of population differentiation that will be referred to later in our study.

2.1.2 Hybridization

Hybridization can be defined simply as the interbreeding of individuals from two populations or groups of populations distinguishable by one or more heritable character [10,11]. The resulting offspring must be viable and fertile for the term to be applicable. Hybridization is generally deemed rare at the level of the individuals, but widespread at the level of the species [12]. In other words, hybrid individuals are rare, but the number of species that hybridize is high and the evolutionary consequences are important. The process can result from both anthropogenic and natural causes, often from environmental disturbances causing habitats to overlap or the creation of artificial bridges between habitats.

The result of hybridization is admixture. As such, the genome of an admixed or hybrid individual is the mixture of alleles from different ancestries. In rare cases where no admixture is present, each individual can only have originated from only a single population. In the case where there is admixture each individual can be a mixture of two or more populations.

With the intensification of global trade and travel, chances for hybridization between populations or species otherwise separated by impassible natural barriers are becoming more common [13]. This is not without ecological consequences. In fact, hybridization and other reticulate evolutionary events have been shown to be one of the mechanisms stimulating invasiveness and to influence pathogenicity of a number of species [14,15]. Moreover, backcrossed individuals are often impossible to differentiate morphologically from the parental populations [12]. As such, tests for hybridization are almost exclusively based on genetics.

2.1.3 Introgression

Introgression, or introgressive hybridization, is the incorporation of alleles or genes from one species, or population, into the gene pool of another as a result of hybridization [10]. Importantly, a locus is said to be introgressed relative to another. In other words, two recognizably distinct and persistent populations must exist for the term to be applicable. Understandably, the genetic boundary between species or populations in which introgression has occurred does not have to extend over the whole genome. As such, these intra- or interspecific boundaries are sometimes said to be “semipermeable”, where some loci are more likely to introgress according to selective pressures. Introgression is also referred to as gene flow between populations.

2.1.4 Population statistics

The Hardy-Weinberg model [16,17] is one of the simplest models in population genetics. A number of more complex models incorporate it in their derivation. Simply put, the model describes allele and genotype frequencies in a population for loci under no evolutionary pressure. The model makes a number of simplifying assumption about the mechanisms of producing gametes [4]. Notably, a single isolated population is assumed and mating within this population is random. We show here how to calculate the Hardy-Weinberg equilibrium from allelic frequencies.

For simplicity, let us limit ourselves to bi-allelic loci. Let p be the frequency of allele A and q the frequency of allele a in the total population, with $p + q = 1$. Then the genotype frequencies are expected to be:

$$G_{AA} = p^2$$

$$G_{aa} = q^2$$

$$G_{Aa} = 2pq$$

where G_{AA} and G_{aa} are the frequencies of homozygous genotypes and G_{Aa} is the frequency of heterozygous genotypes. Deviation from these values is generally the results of evolutionary events such as selective pressure or migration between populations. We can extend this model to cases where there can be more than two alleles at any given locus to calculate the expected heterozygosity for the total population H_e, according to Nei's unbiased estimate:

$$H_e = \frac{1}{M} \sum_{m=1}^{M} \frac{2N(1 - \sum_i x_{mi}^2)}{2N - 1}$$

where x_{mi} is the frequency of allele i at locus m and I is the total number of alleles at locus m [18]. The same equation can be applied to each of the K sub-populations

individually to calculate H_e^k. The observed heterozygosity for the total population H_o and for each of the K subpopulations H_o^k can simply be calculated as the average ratio of heterozygous loci for each individual. Deviation from the Hardy-Weinberg equilibrium can be calculated as the average inbreeding coefficient over all subpopulations F_{IS}, given by:

$$\Delta HW = F_{IS} = \frac{H_e^S - H_o^S}{H_e^S}$$

where H_e^S and H_o^S are the averages of H_e^k and H_o^k over all subpopulations [19].

2.1.4.1 F-Statistics

The three F-statistics introduced by Wright are measures of population differentiation used to explain the structure of genetic variation within and among diploid subpopulations of a total population [20]. Originally, these measures were defined as correlations between uniting gametes. The first, the departure from panmixia in the total population is defined as “the correlation between gametes within an individual relative to the entire population” and is noted as F_{IT} [20,21]. The second, the genetic divergence among subdivisions is defined as “the correlation between gametes within an individual relative to the subpopulation to which it belongs” and is noted as F_{IS} [20,21]. The third, the departures from panmixia within subpopulations is defined as “the correlation between gametes chosen randomly from within the same subpopulation relative to the entire population” and is noted as F_{ST} [20,21]. The latter, also known as the fixation index, is widely used in population genetics studies and a number of other fields including disease association and forensic science [20]. The F-statistics are hard to calculate in practice and are usually estimated in a number of ways.

The F_{IS} is directly related to the Hardy-Weinberg equilibrium. For a given subpopulation k, it can be estimated as:

$$F_{IS}^k \simeq \Delta HW^k = \frac{H_e^k - H_o^k}{H_e^k}$$

where H_e^k is the expected heterozygosity and H_o^k the observed heterozygosity for subpopulation k [19]. As such the average F_{IS} over all subpopulations is estimated as:

$$F_{IS} \simeq \frac{1}{K} \sum_k \frac{H_e^k - H_o^k}{H_e^k}$$

The F_{ST} reflects the distribution of allelic frequencies among populations and is directly related to their variance. In other words, the larger the value of F_{ST}, the greater the allelic differences within each population. A simple way to estimate this parameter for a population k is given by:

$$F_{ST}^k = \frac{H_e^T - H_e^k}{H_e^k}$$

where H_e^T is the expected heterozygosity of the total population [19]. The average F_{ST} over all subpopulations can be calculated in the same way as for the F_{IS}. As we shall now see, assignment methods based on Bayesian analysis make use of these population statistics and assumptions to infer population structure.

2.2 Assignment methods and mixture analysis

2.2.1 Assignment methods

The goal of assignment methods is to classify individuals, more specifically their genotype, into a number of defined or undefined populations from which they could have originated. They can serve to solve a wide range of biological problems and questions. In some studies the goal is to infer population structure, and consequently

the most probable value for K, from a number of individuals for which the geographical sampling location may be known [5,22]. In others, it is to classify a number of individuals of unknown geographical origin into previously defined populations.

There are two main types of assignments methods: distance and model-based [8]. With distance-based methods, a pairwise dissimilarity matrix is calculated for specimens based on some metric. Standard clustering or PCA based methods can then be applied to the result. With model-based methods, specimens are assumed to be drawn from some parametric model. The parameters corresponding to each cluster in this model and population membership of the specimens must be inferred simultaneously using some statistical model. Examples of model-based methods include maximum-likelihood or Bayesian approaches.

A further distinction between assignments methods is in the estimation of local [23,24] or global parameters [25]. In this study, we mostly concentrate on global methods, which estimate average ancestry over the whole genome. Nonetheless, we could also be interested in inferring ancestry for subsets or partitions of the genome, such as individual chromosomes or genes.

2.2.2 Algorithms and methods

2.2.2.1 Bayesian methods

STRUCTURE is a model-based clustering method for population assignment and inference of admixture proportions from multilocus genotype data [8,22]. It is currently, without a doubt, the most widely used algorithm for inferring population structure and is typically the first step in analyzing population genetics data sets [22].

The model, in which the number of populations K is user-defined, is based on Bayesian probabilities. For this, each of the K populations are characterized by specific allelic frequencies at each locus. Importantly, the model makes two strong assumption about the data. The first is that all loci are unlinked and at linkage equilibrium. The second is that the populations are at Hardy-Weinberg equilibrium. We will return to these assumptions and discuss their implications later in section 2.2.2.1.1. Instead, we first present the basic model and inference algorithm for STRUCTURE, taken from the original publication [8], to give the reader a general understanding of the algorithm and why it depends on the Hardy-Weinberg equilibrium. Let it be noted that the algorithm briefly described below is for the model without admixture.

As previously stated, let N be the number of individuals and M the number of loci and K the number of populations. Furthermore, let V be the number of samples for which the population of origin is unknown, and $N = U + V$ be the total number of samples. Let us first assume that the individuals are diploid and that their origin or ancestry is unknown, such that $N = V$. Furthermore, let X be the genotypes, Y be the populations of origin and Z the allelic frequencies in the populations. We have:

$(x_m^{(n,1)}, x_m^{(n,2)})$	the genotype of individual n at locus m
	where $n = 1, 2, ..., N$ and $m = 1, 2, ..., M$
y_n	the origin or ancestry of individual n
z_{kmi}	the frequency of allele i at locus m in population k

where I_m is the number of distinct alleles at locus m. Accordingly, the probability of having sampled a specific allele x for a specimen, knowing its source population and the allelic frequencies for this population at locus m, is given by $Pr(x_m^{(n,a)} = i|Z, Y) = z_{y_n mi}$.

For simplification let us assume that before observing the genotypes, the prior probability that an individual came from any of the populations is $Pr(y_n = k) = 1/K$. Finally, the allelic frequencies are simply initialized according to a Dirichlet distribution, such that $z_{km} \sim \mathcal{D}(\lambda_1, \lambda_2, ..., \lambda_I)$ with $\lambda_1 = \lambda_2 = ... = \lambda_I = 1.0$, which gives a uniform distribution.

The parameters of the model are then updated for a predetermined number of iterations according to a standard Markov-Chain Monte-Carlo procedure. Typical runs are set to 100 000 iterations to detect finer population structure and to obtain accurate estimates of the allelic frequencies. Nonetheless, as few as 1 000 iterations can give meaningful results in some cases [22].

The model with admixture further introduces a parameter Q representing the admixture proportions and the existing parameter Y is expanded to allow for each observed loci of a specimen to have been derived from a different population. Moreover, the model can be expanded to allow some individuals to have been sourced from a specific population with high probability. This allows the learning process to be supervised for these individuals. This is especially useful when very few loci are used or when population structure is known to be weak [25].

The original model disregarded the possibility of correlation between loci. In other words, the values in y within each individual were assumed independent. It was later updated to account for some, but certainly not all, linkage disequilibrium [26]. As such, the authors note that the model still depends on the presence of several weakly linked, or unlinked genetic regions. As a consequence, it is necessary that markers be taken from many different genomic regions and that linkage remains moderate. A few

additional improvements were made to the model and algorithm, but these are of minor importance for our purpose and will not be discussed here [27,28].

2.2.2.2 Shortcoming of STRUCTURE and Bayesian models in general

A number of studies have identified shortcomings to the STRUCTURE statistical model and inference algorithm. Unsurprisingly, some are related to linkage disequilibrium and deviations from the Hardy-Weinberg equilibrium, but other stem from unbalanced sampling sizes, mutation rates and selective pressure [22]. In some cases, unsampled “ghost” populations can lead to incorrect inferences about population structure [6,29]. Furthermore, the ability of Bayesian methods in general to detect population structure depends on accurately estimating allelic frequencies, which in turn depends on large sample sizes and large quantities of markers [22]. We present here a few empirical studies on the performance of STRUCTURE faced with purposefully designed datasets.

In a study based on simulated data, for which the phylogenetic relationships between populations are known, it was noted that STRUCTURE failed to recognize populations partitions accurately when sample sizes are unbalanced [30]. In some extremely unbalanced sampling cases, it even failed to resolve the correct number of populations. Moreover, STRUCTURE tended to incorrectly merge phylogenetically distant populations when both had smaller sample sizes compared to other closer populations. Finally, authors noted that due to the stochastic nature of the algorithm, results varied considerably between runs.

Another study based on simulated data came to the same conclusions [31]. Again, STRUCTURE failed to recognize the correct number of populations when the

sampling scheme was unbalanced. The authors went as far as to suggest that published studies based on data with clearly uneven sampling sizes should be reevaluated. Here the most sampled populations ended up being differentiated from each other and the least samples clustered together, although correction could be applied.

2.2.2.3 Admixture and other Bayesian methods

Admixture is another popular Bayesian method for population inference and assignment based on the same model. However, it uses maximum likelihood and a block relaxation approach to estimate the model parameters [25,32]. This yields considerable improvements in computational efficiency and, the authors argue, comparable accuracy in estimating mixture proportions [25,33]. A number of other Bayesian methods are available; however, most were developed for specific problems or questions and lack the wide-scale applicability of STRUCTURE or Admixture [22].

2.2.2.4 Distance based and other methods

In the simplest form of distance-based methods, individuals are assigned to the closest population based on some genetic measure of distance, averaged over all members of the given population [34]. The k-nearest neighbor algorithm has also been adapted for the purpose of genetic assignment [35]. With this approach, an unknown individual is assigned to the majority vote of the k closest individuals of known origin (note that k here refers to the number of neighbors, not populations).

The developers of the widely used assignment algorithm STRUCTURE however state that: “the clusters identified [by distance based methods] may be heavily dependent on both the distance measure and graphical representation chosen; it is

difficult to assess how confident we should be that the clusters obtained in this way are meaningful; and it is difficult to incorporate additional information such as the geographic sampling locations of individuals." [8]. Some efforts have been made to overcome these limitations by relying on genetic distances rather than allelic frequencies, without considerable success [34].

2.2.2.5 Principal component analysis approaches

PCA and PCA based methods have been used to cluster individuals of similar genetic ancestry together and are generally computationally efficient even on large dataset [25]. However, as we shall later discuss in Chapter 3, these methods are not without their shortcomings. Some population characteristics that will cause these methods to fail include the presence of closely related subpopulations or a distant subpopulation.

2.3 Machine learning and genomics data

Genomics data possesses complex underlying structure that is, most often, difficult to capture to make inferences [36]. True signal is often obscured by limited samples sizes, noise or high dimensionality. With this in mind, in a recent workshop on genomic data analysis, expert contributors shared several key ideas to explore novel approaches to deal with current roadblocks [36]. (1) The first is that inherent properties of the data can help in choosing unsupervised learning methods to expose hidden structure. (2) The second is combining different data types and sources is beneficial. (3) The third is that giving meaning to the output of machine learning models, in analogy to a p-value for example, is not straightforward, but nonetheless possible. (4) The fourth is

that results from complex machine learning models are difficult to interpret. (5) Finally, and unsurprisingly, computational efficiency is still an issue.

2.3.1 Challenges of using SNP data for genotype assignment

In this study, we are mostly concerned with SNP, microsatellite and other similar types of data. SNP data poses the problem of high dimensionality. With most such genomics datasets the number of features is considerably larger than the number of specimens. Other challenges posed by large-scale SNP datasets include redundancy and the wide range of data formats employed in different fields [37].

As a consequence, information-conserving ways of reducing the feature space must be employed, but no single method is applicable and appropriate to all questions being asked from the data. The methods used to preprocess SNP data for geographical origin prediction will not necessarily provide satisfying results for disease prediction, for example. We can easily assume that solving these problems relies on capturing very different patterns in the data.

A considerable number of methods used to reduce the dimensionality of genomics data have been published. Also known as feature space reduction methods, these can either be categorized as feature selection or feature engineering. Here we present a few such methods that have been applied to SNP data for a number of different problem types.

2.3.1.1 Informative SNP subset selection

Feature selections methods are mainly divided into filter, wrapper and embedding types [37]. Filter methods are in essence unsupervised learning (or are said to not incorporate learning at all). These methods select a subset of features based on

some statistical score reflecting their correlation with the outcome variable. Filter methods are generally fast but ignore possible interactions between loci [37]. Wrapper methods are simply supervised learning, essentially ranking and shortlisting loci that will be used to build the final classifier. Embedding methods are a variation on the latter, where feature selection is done simultaneously with learning the classification task. Here we present two cases of supervised SNP selection methods applied to geographical origin classification and disease prediction.

A combination of PCA and Random Forests has been used successfully to select a subset of informative SNPs for population origin classification [38]. PCA was first applied to the whole dataset and the first two components were used to reduce the number of SNPs. To achieve this, a score is calculated for each marker by squaring and its values along the principal components and summing them. A number of the highest scoring markers are then conserved for each autosome. Following this, a Random Forest classifier was trained on the data to further reduce the number of SNPs based on the mean Gini index and the mean accuracy decrease.

Random Forests have also been used to select SNPs in Genome-wide association studies. In one such approach Random Forests classifiers are repetitively trained on permuted sets of SNPs and their labels to calculate a p-value from the Gini indices [39]. Only SNPs below a certain threshold are kept and used to train a final Random Forests classifier that will serve to make predictions on new specimens. In this study, the limited subsets of informative SNPs outperformed the full set of features in classification of Alzheimer's and Parkinson's patients. It must be noted however that the process is computationally inefficient.

2.3.1.2 Feature space engineering for SNP data

SNP selection methods have been shown to be effective in reducing the dimensionality of the data, however they are prone to overfitting and tend to be computationally inefficient if done in a supervised manner [40,41]. Compressing the entirety of the data, while preserving a maximum of information, generally provides a reduced feature space that allows good generalization to new samples. Importantly, this should be done in a way mindful to the pattern we hope to capture in the data. In the case of geographical origin prediction or population assignment we can expect phylogenetic relationships between the samples to be of interest.

2.4 Phylogenetic networks

Phylogenetic trees are widely used to represent and analyze evolutionary relationships at every taxonomic rank. However, evolution rarely occurs in a perfectly tree-like manner, making other less restrictive models and heuristics, such as a network, more appropriate for complex evolutionary histories [42]. In fact, reticulate evolution, whereby new lineages with novel combinations of phenotypes are created, often precedes patterns of vertical descent with modification and most of the time multiple evolutionary mechanisms are active simultaneously [43]. Moreover, it has been suggested that the very concept of a tree of life is inappropriate to describe prokaryotic evolution [44].

Phylogenetic networks, on the contrary, are a generalization of phylogenetic trees under which other types of networks better suited to account for evolutionary events such as hybridization, horizontal gene transfer, recombination, symbiosis and symbiogenesis are also included [43,45]. Many different network types, differing from

trees only by the presence of reticulation nodes, fit under the definition of a phylogenetic network, but a major distinction can be made in the type representation they offer of evolutionary history [42,45].

Reticulate networks, generally depicted as a rooted phylogenetic tree with additional edges offer an explicit representation of evolutionary history where each node represents an ancestral species (see Figure 1). Their inference is however difficult and algorithms for inferring them are not practical and widely used [46].

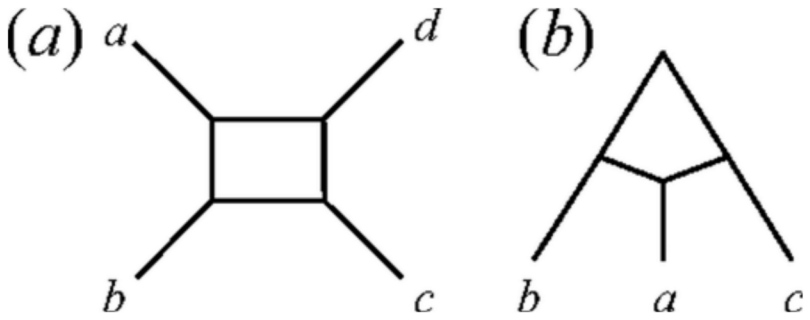


Figure 1 (a) Unrooted phylogenetic network, in which internal nodes do not necessarily represent an ancestral state. (b) rooted phylogenetic network, in which internal nodes represent an ancestral species. [47]

Split networks, on the other hand, usually drawn unrooted, offer an implicit representation of evolutionary history where each node does not necessarily represent an ancestral species or individual (see Figure 1) [45]. Such networks can be better intuited if seen to be expressing affinity relationships, rather than genealogical ones, wherein genetically similar individuals are arranged to be in close proximity [42,48]. A number of well-established algorithms are available for computing them. Here we will concentrate on split networks and the set of splits describing them.

Split networks are in fact a graphical representation of an underlying split system Σ composed of a number of splits. A split S is simply the weighted bipartition of the taxon set $\mathcal{X}$ into non−empty and disjoint subsets A and B [46,49,50]. The components A and B are equivalent to those obtained from the deletion of a single edge of a phylogenetic tree T [46,49,50]. The notation $S = A|B$ is used to represent a split. The size of a split is defined as $size(S) = min\{|A|, |B|\}$ and a split of size one is said to be trivial. It must be noted that many different split networks can represent the same set of splits, but the contrary is not true. A weighted split system, in which each split is assigned a positive weight corresponding to a measure of phylogenetic distance, contains all the information needed to build a split network.

2.4.1 Splits systems

The distinction between a split system describing a phylogenetic tree and one describing a phylogenetic network is that the former is said to be compatible, while the latter is said to be incompatible, weakly compatible, or circular [46]. Two splits $S_1 = A_1|B_1$ and $S_2 = A_2|B_2$ are compatible if one of the intersections $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$ and $B_1 \cap B_2$ is empty. In other words, a pair of compatible splits could be represented by, or combined into, a single tree without reticulation nodes. A set of splits Σ is said to be compatible if all pairs of split it contains are compatible and there exists an unrooted phylogenetic tree T that represents Σ.

Phylogenetics networks are described by sets of splits that may or may not be compatible. However, in practice, allowing for full incompatibility in a set of splits tends to produce unnecessarily complicated networks [46]. Weakly compatible and circular sets of splits were introduced to prevent this and have the added benefit of producing

planar, or close to planar, outer-labeled networks. Such split systems cannot, of course, be represented as a phylogenetic tree. Other types of split systems have also been defined, such as octahedral split systems [51], but here we concentrate on weakly compatible and circular split systems.

Three splits $S_1 = A_1|B_1$, $S_2 = A_2|B_2$ and $S_3 = A_3|B_3$ are said to be weakly compatible if at least one of the intersections $A_1 \cap A_2 \cap A_3$, $A_1 \cap B_2 \cap B_3$, $B_1 \cap A_2 \cap B_3$ and $B_1 \cap B_2 \cap A_3$ is empty. A set of splits Σ is said to be weakly compatible if any three splits it contains are weakly compatible. Interestingly, this set is linearly independent and its cardinality does not exceed $\binom{N}{2}$, where N is the number of taxa [52].

A set of splits Σ on X is said to be circular, if the taxa in X can be placed in a circle such that every split in Σ can be represented by a line through the circle diving the taxa into two sets. Equivalently, if there exists a cyclic permutation $\pi = (x_1, ..., x_n)$ of the taxa in $\mathcal{X}$ such that each split in Σ has the form $S = \{x_p, x_{p+1}, , x_q\}|\mathcal{X} - \{x_p, x_{p+1}, , x_q\}$ and $1 < p \leq q \leq n$ [50]. As such, let $\Sigma^\circ(\pi)$ the set of splits that can be obtained by a bisecting line segment on the plane where the elements of π have been sequentially placed in a circle [49]. Similarly, Σ is circular if $\Sigma \subseteq \Sigma^\circ(\pi)$. There are $(N-1)!$ circular orderings [53]. Unlike weakly compatible sets of splits which can contain crossing edges, circular sets of splits can always be represented as planar, outer labelled networks. Moreover, circular sets of splits are always weakly compatible.

2.4.2 Inferring phylogenetic networks

A phylogenetic network can be inferred from a set of trees [54], quartets [55–58] or pairwise distances. Here we will concentrate on the algorithms used to produce a set

of weakly compatible or circular splits from a pairwise distance matrix or dissimilarity matrix, in this case phyletic distances.

2.4.2.1 Split decomposition

The first method to be introduced is called split decomposition, which decomposes the distance matrix as a sum of weakly compatible split metrics, or weighted splits, plus a residue known as the split-prime residue [46,52,59]. A major part of random noise tends to end up contained in this residue [59]. Split decomposition is however computationally inefficient and thus limited to very small datasets of 100 individuals or less.

Moreover, the algorithm tends to produce overly complicated split systems and phylogenetic networks. This may seem like a purely visual consideration, but akin to the problem of overfitting in machine learning, networks containing more incompatible splits will always fit the data at least as well as purely tree-like split systems [60]. Parsimony is thus an important consideration in the inference of a split system. Take the case of lateral gene transfer, for example. Other, sometimes highly unlikely, evolutionary scenarios such as convergent evolution or multiple independent gene loss events can serve to represent the data equally well [60].

2.4.2.2 The NeighborNet algorithm

The NeighborNet algorithm is an agglomerative method derived from the familiar neighbor joining that produces weighted, circular sets of splits. It tends to produce sets of splits of higher resolution than split decomposition and is more computationally efficient [47,61,62]. Importantly, the algorithm is consistent, producing a network that exactly represents the distances if they are circular and a tree if the distances are

additive [62,63]. It has however been stated that NeighborNet is sensitive to distorted metrics, such as when the sequences used to calculate the dissimilarity matrix are too distantly related [64].

The first step of the algorithm is to produce a linear ordering of the taxa following a process analogous to neighbor joining. It is more appropriate to describe NeighborNet as a greedy algorithm for finding circular split systems that best describes the input distance matrix [65]. Nonetheless, in an agglomerative manner, a criterion is used to merge nodes and the distance matrix is reduced accordingly. The major difference with neighbor joining is that triples of nodes are merged instead of pairs, ultimately producing a linear ordering of the taxa. At this stage, the split system is produced from the ordering by taking all splits that respect circularity. Other similar algorithms use a reverse agglomeration process [66].

Following the production of a circular ordering, split weights are then computed for all splits respecting circularity using a non-negative least squares regression. Finally, splits weighted under a certain threshold are eliminated and the remaining splits form the weighted split system.

2.4.2.3 MC-Net algorithm

The MC-Net algorithm is also a distance-based method that proceeds in a way similar to the NeighborNet algorithm by first finding a circular ordering of the taxa and then weighing the obtained splits [67]. However, MC-Net uses a heuristic to find the circular ordering instead of an agglomerative rule. An initial ordering is first produced following a greedy algorithm. This ordering is then optimized following a standard Monte-Carlo algorithm. The energy function simply defined as the sum of distances

between neighbors in the ordering. Finally, the splits of a weighted in the same way as the Neighbor-Net algorithm.

The authors found the energy score of orderings produced by MC-Net are lower than those produced by NeighborNet. Furthermore, the split systems obtained through MC-Net are generally simpler, containing fewer splits, than those obtained through NeighborNet.

2.4.3 Algorithmic use of phylogenetic networks and trees

The most common application of phylogenetic networks is visualization for the purpose of exploratory data analysis. In this sense, they are a type of agglomerative hierarchical clustering, sometimes referred to as fuzzy clustering, allowing each taxon to simultaneously be a member of many clusters [68]. However, phylogenetic networks and their inference can serve to extract unknown patterns from the data in an unsupervised way, but also to provide a compact representation of a dataset. They have been shown to possess many of the good features of multivariate data summarisation techniques such as PCA, without their known mathematical limitations that produce unwanted artefacts [68]. Here we present a few examples of problems where phylogenetic networks or trees are being used successfully as a data transformation step in various algorithms or methods.

Metagenomics data can serve to discriminate healthy subjects from affected ones in a number of diseases. In a recent study, metagenomics data from the gut microbiota was used to identify and classify Inflammatory Bowel Disease patients [69]. A Convolutional Neural Network (CNN) model was employed for this purpose and the operational taxonomic units present in the microbiota served as features, each sample

presenting different levels. To transform the input space to something recognizable by a CNN, a phylogenetic tree was inferred for all operational taxonomic units and served as a measure of similarity, similar to the concept of neighborhood for pixels in a image.

In another type of problem, phylogenetic diversity and phylogenetic networks were combined to obtain a measure of biodiversity. Split diversity is a score calculated from the set of splits, applicable to a number of taxon selection problems [70]. Other biodiversity or isolation indices considers both the internal structure of the network and the set of subtrees they contain [71,72].

Phylogenetic networks have also been used as a visual aid for manual classification of the Ferritin-like Superfamily proteins [73]. In this case, sequence similarity among types of proteins was too low to allow for inference of function. However, constructing phylogenetic networks from structural alignments of the proteins yielded considerable improvement in the resolution of functional relationships and enabled the authors to classify the proteins in a way that was not possible from sequences alone.

Interestingly, phylogenetic networks have been applied to non-genetic datasets that exhibit evolutionary structure such as languages [74,75] and cultural artefacts [76]. The term phylomemetic has been coined as a combination of meme (ideas and cultural phenomena [77]), and phylogenetics to describe such relationships. Lateral gene transfer and hybridization obviously have parallels in language. These are, it would seem, considerably more common than their biological counterparts [78].

Chapter 3

Mycorrhiza: combining phylogenetic networks and Random Forests for prediction of ancestry from multilocus genotype data

3.1 Preface

In this study we present the result of our efforts to develop a machine learning method aimed at predicting the geographical origin of biological samples in accordance with the requirements of the BioSAFE project. The method which can be applied to a wide range of data types, including SNP and microsatellites, is of trivial runtime considering the use case and performs equivalently, or better than currently accepted and widely used Bayesian methods.

Having reviewed the challenges posed by high dimensional SNP data and the shortcomings of Bayesian methods, we propose an algorithm based on dimensionality reduction and phylogenetic relationships. As has been previously discussed, the ancestry of individual biological samples representing an ecological risk of invasion are most likely marked by reticulate evolutionary events. Consequently, phylogenetic networks are used here as a feature space reduction method that captures the structure inherent to events such as hybridization, lateral gene transfer and recombination. The reduced feature set, in the form of a split system, is then inputted to a Random Forests classifier. We termed our method Mycorrhiza, a word defined “a symbiotic association between a fungus and the roots of a vascular host plant.” [79].

3.2 Abstract

The genotype assignment problem consists of predicting, from the genotype of an individual, which of a known set of populations it originated from. The problem arises in a variety of contexts, including wildlife forensics, invasive species detection, and biodiversity monitoring. Existing approaches perform well under ideal conditions but are sensitive to a variety of common violations of the assumptions they rely on. In this paper, we introduce Mycorrhiza, a machine learning approach for the genotype assignment problem. Our algorithm makes use of phylogenetic networks to engineer features that encode the evolutionary relationships among samples. Those features are then used as input to a Random Forests classifier. The classification accuracy was assessed on multiple published SNP, microsatellite or consensus sequence datasets with wide ranges of size, geographical distribution and population structure. It compared favorably against widely used assessment tests or mixture analysis methods such as STRUCTURE and Admixture, and against another machine-learning based approach using PCA for dimensionality reduction. Mycorrhiza yields particularly significant gains on datasets with a large average FST or deviation from the Hardy Weinberg equilibrium. Moreover, the phylogenetic network approach consistently estimates mixture proportions more accurately than the PCA variant. Mycorrhiza is released as an easy to use open-source python package on GitHub at github.com/jgeofil/mycorrhiza.

3.3 Introduction

3.3.1 Assignment methods

Assignment methods are a group of closely related methods that use genetic information to determine the population membership of individuals from a given species. For this purpose, the term “population” generally refers to a group of individuals in close geographical proximity whose probability of interbreeding is higher than that of interbreeding with other groups [4]. These approaches are thus mostly concerned with events occurring on relatively short timescales, on the order of one to a thousand generations [5,9]. In one version of the problem, called the assignment test, one aims to estimate the probabilities that a multilocus genotype of unknown origin came from each of a fixed set of known populations. This is equivalent to the classification problem in machine learning. In another version, called genetic mixture analysis or genetic stock identification, the objective is to estimate both mixture proportions and posterior source probabilities for each individual. In this paper we present and evaluate a new machine learning algorithm for genetic assignment based in part on phylogenetic networks.

Assignment methods have been used for a variety of applications, including wildlife forensics [80–84], understanding migratory patterns and geographical boundaries for conservation efforts [85] and the identification of hybrid individuals for the management of invasive species [86–90]. Despite their wide range of applications in a variety of fields and a number of well know software tools implementing various algorithms (see below), little consensus exists about their use in classical supervised classification problems. Furthermore, few of these software packages implement machine learning standards and practices, such as cross-validation.

3.3.1.1 Implementations and limitations of assignment methods

Assignment methods have mostly been implemented with frequentist, maximum likelihood or Bayesian analysis algorithms [5]. Most approaches assume that the loci used as features are at Hardy-Weinberg equilibrium and in linkage equilibrium [5,8,26,91]. In other words, these methods essentially clusters individuals in Hardy-Weinberg and linkage equilibrium populations, creating groups with distinct allelic frequencies. In real-world populations, these assumptions are rarely fully satisfied for the total set of available loci. They can even become meaningless if the boundaries of natural populations are forced to accommodate geopolitical borders, rather than reproduction boundaries. This may become necessary when assignment tests are employed as a tool in legal cases opposing countries, in coordinated international recovery efforts, or for the management of invasive species, for example [85,86,92]. Efforts have been made to overcome these limitations by relying on genetic distances rather than allelic frequencies, without considerable success [34].

The widely used STRUCTURE [8,93] program is a model-based, Bayesian clustering method for explicitly inferring population structure and probabilistically assigning individuals to K populations [8,26]. Although originally introduced as an unsupervised clustering approach, the method has since been enhanced with a semi-supervised model in which some of the individuals can be pre-assigned to their known population of origin [22,94]. This capability can also be used to emulate what is known as supervised learning in the machine learning field. Unfortunately, STRUCTURE suffers from its high computational complexity when applied to large SNP datasets, in which case runtime can be on the order of days or even weeks [95]. As an alternative,

FastSTRUCTURE improves the computational efficiency of STRUCTURE using a variational Bayesian framework, thereby making it two orders of magnitude faster [95]. However, unlike STRUCTURE, it cannot account for linkage disequilibrium.

Admixture [33] is another tool for maximum likelihood estimation of individual ancestry. It is based on the same statistical model as STRUCTURE but optimized with a block relaxation algorithm [32]. According to the authors, it is as accurate as STRUCTURE, but with the added advantage of being much more computationally efficient. In practice however, STRUCTURE is sometimes more accurate than Admixture as it can partially account for linkage disequilibrium between markers [26].

Bayesian clustering methods in general come with many known limitations. They are, for example, known to lose their ability to detect subpopulations at very low levels of population differentiation. Some suggest FST values under 0.02 will cause them to fail to resolve all populations [96], while others suggest values in the 0 to 0.05 range [34]. Nonetheless, STRUCTURE has been shown to perform well at levels of population differentiation as low as 0.02 [96]. Studies have also shown that when the parameter K is smaller than the actual number of populations, STRUCTURE can produce clusters inconsistent with their evolutionary history [91]. This can also happen when the size of populations is unbalanced [30,31,91,97].

3.3.1.2 Other approaches

Methods based on Principal component analysis (PCA) or other multivariate analysis methods have long been used as fast and efficient tools to analyse structure in genomics data sets [98–104]. When used simply as a visualization tool, these methods do not however, allow for straightforward interpretation of individual ancestry from the

low-dimensional projection they produce. Alternative clustering methods, such as discriminant analysis of principal components (DAPC) [105], have been developed with the aim of providing fast and flexible exploratory tools that produce easily interpretable results [105]. However, these methods either only allow for hard clustering [105,106] or provide questionable admixture results with soft-clustering [107]. Multivariate analysis approaches are nonetheless prone to mathematical artefacts, leading to spurious conclusions about population structure [108–111].

A number of R packages have been developed with the goal of offering machine learning solutions for genomics. The package Adegenet was developed with the aim of bridging the gap between multivariate data analysis solutions and genomics packages by implementing a number of clustering algorithms such as snapclust and discriminant analysis of principal components [104,112,113]. The package can calculate a number of population statistics and perform spatial genetics analyses [114–116]. It must be noted that, snapclust [113], the algorithm for population assignment implemented in the package, is also based on the same assumptions about Hardy-Weinberg equilibrium as Bayesian methods.

Overall, no existing software addresses all of the shortcomings and limitations mentioned above. In this paper, we set out to develop a new method for genotype assignment and mixture analysis rooted in machine learning principles that would address these problems, while keeping in mind and taking advantage of the phylogenetic structure present in genomics datasets.

3.3.2 Phylogenetic networks as feature engineering

SNP data poses the problem of high dimensionality, otherwise known as the "curse of dimensionality" in the machine learning field [117], where the number of features (SNPs) exceeds the number of training examples (labeled specimens). Unchecked, this often leads to overfitting and poor classifier performance. A first dimensionality reduction approach is to select a subset of features based on some criteria or score. A large number of such supervised and unsupervised feature selection methods have been applied to SNP data [41,118–120]. However, some of these methods are themselves prone to overfitting and it has been demonstrated that this can easily lead to inflated estimation of prediction accuracy. As a matter of fact, is has been noted that a number of genomics studies have overlooked the need for cross-validation [40,41].

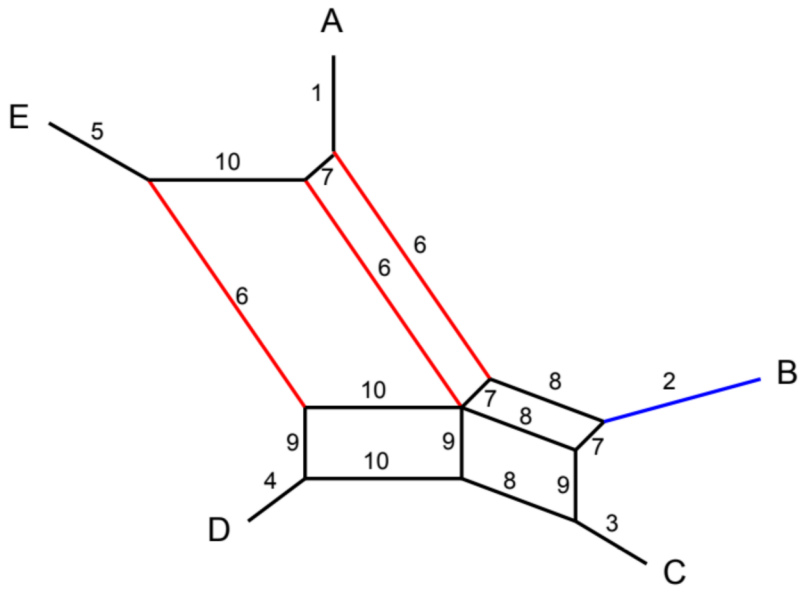
Another possible solution is to project the data into a lower dimensional space while preserving as much information as possible. PCA is probably the best known dimensionality reduction method and is commonly applied to genotype data [121]. However, PCA was not designed to account for phylogenetic structure in populations.

The familiar phylogenetic tree can undoubtedly be employed as a means of dimensionality reduction (see below). Unfortunately, the model assumes evolutionary histories dominated by speciation and descent with modification. This is not appropriate in population genetics settings where populations are inter-fertile and exchange genetic material. Phylogenetic networks have been introduced to capture and represent non tree-like evolution [42,45]. Those that allow for hybrid nodes, or reticulations, are better suited to account for more complex evolutionary events such as hybridization, horizontal

gene transfer and recombination [52,122,123]. One such type of network, the split network, can be interpreted as a combinatorial generalization of unrooted phylogenetic trees [45]. Since split networks are at the core of Mycorrhiza, we provide additional background information on this type of network.

Split networks. Both phylogenetic trees and networks are composed of a set of splits, referred to as a "split system". A split S, of the form $S = A|B$, is a bipartition of the set of taxa (or specimens) into two nonempty subsets A and B [25]. The distinction between a split system describing a phylogenetic tree and one describing a phylogenetic network is in the rules that the set of splits must satisfy. Two splits $S_1 = A_1|B_1$ and $S_2 = A_2|B_2$ are compatible if one of the intersections $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$ and $B_1 \cap B_2$ is empty. A set of splits Σ is said to be compatible if all pairs of split it contains are compatible; in that case, there exists an unrooted phylogenetic tree T that represents Σ. Phylogenetic networks are based on a set of splits that are not necessarily compatible, but may instead be weakly compatible, or circular [26]. Three splits $S_1 = A_1|B_1$, $S_2 = A_2|B_2$ and $S_3 = A_3|B_3$ are said to be weakly compatible if at least one of the intersections $A_1 \cap A_2 \cap A_3$, $A_1 \cap B_2 \cap B_3$, $B_1 \cap A_2 \cap B_3$ and $B_1 \cap B_2 \cap A_3$ is empty. A set of splits Σ is said to be weakly compatible if any three splits it contains are weakly compatible. A set of splits Σ on X is said to be circular, if the taxa in X can be placed in a circle such that every split in Σ can be represented by a line through the circle diving the taxa into two sets. Unlike weakly compatible sets of splits which can contain crossing edges, circular sets of splits can always be represented as planar, outer labelled networks [47,124,125]. Moreover, circular sets of splits are always weakly compatible.

a



b

	A	B	C	D	E
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	0	0	0	1
7	1	1	0	0	0
8	0	1	1	0	0
9	0	0	1	1	0
10	0	0	0	1	1

Figure 2 - Example phylogenetic network and corresponding split system. a) Phylogenetic network. The taxa are denoted by the letters A to E and splits are numbered from 1 to 10. Each unique split is represented by one line (trivial split, e.g. split 2, in blue) or multiple parallel lines (e.g. split 6, in red). b) Split system. Each taxon is placed on either side of the splits, as indicated by a binary flag.

A split network, or phylogenetic network, is a graphical representation of an underlying split system. An example phylogenetic network, along with the associated split system represented by a binary matrix is shown in Figure 2. Samples are represented by the letters A to E and splits by the number 1 to 10. Samples with the

same boolean flag (0 or 1) are placed on the same side of the split. For example, in split 6 samples A and E are on one side of the bipartition and samples B, C and D are on the other. Splits for which one of the subsets is of size one (e.g. split 2, in blue) are said to be trivial. Trivial splits represent genetic divergence that is unique to the lone sample.

In the graphical representation of a split system, each split is represented by one or more parallel edges. Split 6, for example, is represented by the 3 red edges. Note that an edge can be associated with only one split. In most cases a weight, corresponding to a dissimilarity measure is associated to each split. This dissimilarity measure can be a measure of evolutionary distance for example. The evolutionary distance between two taxa is thus the sum of the weights of all splits that place these taxa in different subsets.

Split networks can be inferred from pairwise distances between the set of taxa. The NeighborNet algorithm will produce a split system that perfectly fits the input distance matrix, provided it is circularly decomposable [63,65]. The running time of NeighborNet is $O(n^3)$, making it is suitable for large data sets with thousands of samples and millions of loci.

Phylogenetic networks are generally used for visualisation, data mining and exploratory data analysis as a means of fuzzy clustering [68]. Interestingly, there also exists a few recent cases of network inference being used as a data transformation algorithm [69–72]. This makes sense considering that the splits of a circular split system are linearly independent and very little information is lost when resolving a distance matrix [126]. We therefore propose the use of split system decomposition as a feature

space reduction method. For our purpose, the reduced feature set is then inputted to a Random Forest classifier for geographical origin prediction. We named this approach Mycorrhiza, a term defined as the symbiotic association between a fungal network and the roots of host plants [127].

We first compared the assignment accuracy of our approach with the commonly used Bayesian assignment methods STRUCTURE and Admixture and a variant where split system decomposition is replaced by PCA. We then analysed the effect of marker type, sample size, number of populations, population differentiation and other population statistics on classification accuracy. Overall, we show that Mycorrhiza is more accurate than competing approaches, in particular in cases where population differentiation is high. Finally, we considered the implications of computational efficiency and flexibility of the different methods.

3.4 Methods

Mycorrhiza is composed of two main steps. In the first step, a phylogenetic split system is inferred from the genotype data of all individuals (of both unknown and known origin). In the second step, the placement in the split system of individuals of known origin, along with their categorical label corresponding to geographical origin, is used to train a Random Forest classifier. Following this, the trained model is used to make predictions from the split placement of individuals of unknown origin.

3.4.1 Step 1: Split system inference

To produce a split system, pairwise genetic distances are calculated for all N individuals over M loci. For SNP and sequence data, the Jukes-Cantor distance was calculated with MEGA-CC [128]. For microsatellite data, the distance was calculated as

the number of loci with different copy number. Following this, a phylogenetic split system is built from the pairwise distances. Based on this matrix, the NeighborNet program from the SplitsTree4 (version 4.14.6) package [45] is used to obtain a circular phylogenetic split system. Only splits with a weight above 10^{-6} are considered. Although both the labeled and unlabeled examples are used to build the split system, this is done in an unsupervised fashion, without knowledge of the categorical population labels. Finally, feature vectors of length $S - N$ are extracted from the S dimensional split system, to be used in the learning and prediction step. Each individual is represented by a binary vector corresponding to its placement on either side of each of the non-trivial splits. In Figure 2, sample A, for example, would be represented as with feature vector $[1, 1, 0, 0, 0]$. Trivial splits are ignored because they do not have any discriminatory power. In practice, a split system can contain up to a few thousand splits, depending on the number of samples and the complexity of the population structure.

3.4.2 Step 2: Training and predictions

The feature vectors built based on the split system in step 1 are used as input to a Random Forest classifier [129]. The scikit-learn implementation of a Random Forest was used for this purpose with default parameter settings, except for the number of estimators which was set to 60 (this value was determined on an ad-hoc basis, see discussion) [129]. Alternate families of classifiers such as fully connected neural networks were also evaluated, but they proved less accurate in this context.

The trained model is then used to make predictions for unlabeled examples, based on their split vector. The output of Mycorrhiza corresponds to the probability that a sample belongs to each of the K populations, as estimated by the random forest

predictor. If the desired output is a hard classification to a single population, each sample is attributed the population label that has the highest probability. Alternately, the probability distribution can be interpreted as the individual's mixture proportions over the K populations.

3.4.3 Partitioned Mycorrhiza

In machine learning, ensemble methods, which combine the output of multiple classifiers trained on different subsets of the features, have been shown to reduce generalization error [130]. This is, in fact, the strategy behind Random Forests [131]. We thus developed a variant of Mycorrhiza called Partitioned Mycorrhiza in which P different Mycorrhiza predictors are trained, each on disjoint subsets of loci. This is also known as feature set partitioning and multiple strategies have been explored on how to build the subsets [132]. We presumed that building the subsets sequentially, preserving the order of the loci in the genome, would help capture finer local phylogenetic structure that would otherwise be lost by using the complete feature set. In fact, ancestry proportions are known to vary along the genome and a number of assignment methods are focused on detecting population structure for individual chromosomal segments [23,24]. The final output is obtained by averaging the P predictions. By default, P is set to 10.

3.4.4 PCA variant

For the PCA variant, the first step of the algorithm is replaced by a standard PCA analysis [133]. Here, the placement of a sample in the *D*-dimensional principal component space yields its feature vector. We used the scikit-learn implementation of PCA for this purpose. After evaluating different choices of values for the number of

components D, it became clear that no improvement in accuracy was obtained beyond $D = 50$, and that the results were quite robust to that parameter. We thus set $D = min(N, 50)$ for all data sets. If needed, this hyper-parameter could easily be optimized on a per-dataset basis, e.g. using standard cross-validation procedures

3.4.5 Implementation and software package

Mycorrhiza is released as an easy to use open-source python package on GitHub at github.com/jgeofil/mycorrhiza. The package depends on the SplitsTree software [134]. Data can be inputted in a number of commonly used formats in population genetics. Partitioning parameters can be used in their default setting or optimized with predefined procedures. Common cross-validation methods are implemented to calculate classification accuracy, but static training and testing sets can also be used. Estimated mixture proportions can be outputted as a text file or as a figure similar to those produced by *distruct* [135,136].

3.4.6 Comparison against STRUCTURE and Admixture

For comparison, we also performed the same classification tasks using Mycorrhiza, Partitioned Mycorrhiza, PCA+RF, Partitioned PCA+RF, Admixture V1.3.0 [32,33] and STRUCTURE V2.3.4 [8,26–28]. Classification accuracy was evaluated using 5-fold cross-validation. For the two Mycorrhiza and PCA variants, dimensionality reduction was performed first, with all training and testing examples, but in an unsupervised manner, followed by standard 5-fold cross-validation of the Random Forests predictors. For Admixture and STRUCTURE, cross-validation was emulated as follows. Both programs include an option for supervised learning, by which the population of origin can be fixed for certain individuals, effectively using them as the

training set. For example, to achieve 5-fold cross-validation, five files are outputted with identical genotype data, but each with a different set of samples for which the supervised learning flag is enabled. The program is then run on each of these files separately. For Admixture, supervised analysis was enabled with the --supervised flag [137], and default settings were used otherwise. For STRUCTURE, supervised analysis was enabled with the POPFLAG and USEPOPINFO flags. These parameters tell the program that the input file contains a column of population identifiers and another column indicating for which samples the population information should be taken into account. The burn-in period was set to 20 000, the number of MCMC repetitions to 100 000 and all other parameters were left in their default state.

3.4.7 Datasets

We collected a number of published geotagged genotype datasets (Table 1). Ebola data was obtained from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) [138]. Rice data was obtained from the Rice Diversity Project website [139]. *A. thaliana* data were produced by the Weigel laboratory at the Max Planck Institute for Developmental Biology [140]. *S. musiva* data were obtained from the authors [141]. Human data were obtained from the 1000 genomes project. All other data were obtained from the Data Dryad database [142].

Most datasets had been quality-controlled for their respective publications and were thus used unchanged. Some authors proposed a selected set of markers deemed to be informative for their classification task. To avoid any possible biases, we instead used the full set of markers. SNP datasets were filtered for a minor allele frequency of

0.05 and disallowing any sites for which data is missing for certain individuals. These are the *A. thaliana*, human, *S. musiva* and rice datasets.

3.4.8 Number of loci and partitioning parameters

To investigate the effect of feature set size on classification accuracy, the number of loci used as input was varied by randomly downsampling to the desired number of loci m. Assignment accuracy was evaluated with 5-fold cross-validation. This was repeated 5 times and the results were averaged. For SNP data, m was varied from 50 to M by powers of 2. For microsatellite data, m was varied from 2 to M by increments of 2. For sequence data, the values of m were chosen more arbitrarily, depending on M which varied greatly between datasets.

For Partitioned Mycorrhiza, the same procedure was applied to evaluate the effect of the number of partitions on classification accuracy. The m loci were further divided into $p$subsets and classification accuracy was averaged over five runs for every combination of m and p. For SNP data, p was set to 1, 2, 10, 50, 100 or 500. For microsatellite data, p was set to every value between 1 and to M. For sequence data, the values of p were again chosen more arbitrarily.

3.4.9 Population statistics

Expected heterozygosity was calculated according to Nei's unbiased estimate [18] for the total population (H_e) and each of the K subpopulations (H_e^k). The estimate is thus given by

$$H_e = \frac{1}{M}\sum_{m}^{M}\frac{2N(1-\sum_i r_{mi}^2)}{2N-1}$$

where M is the number of loci, N is the number of samples and x_{mi} is the frequency of allele i at locus m. Observed heterozygosity was calculated as the average ratio of heterozygous loci for each sub-population. Deviation from the Hardy-Weinberg equilibrium was calculated as the average inbreeding coefficient over all subpopulations, given by

$$\Delta HW = F_{IS} = \frac{H_e^S - H_o^S}{H_e^S}$$

where H_e^S and H_o^S are the averages of H_e^k and H_o^k over all subpopulations [19]. The fixation index was calculated in two ways. The first, appropriate for diploid genotypes, is given by

$$F_{ST} = \frac{H_e^T - H_e^S}{H_e^T}$$

where H_e^T is the expected heterozygosity of the total population [19]. The second, appropriate for both haploid and diploid genotypes, according to Hudson's [143] estimate based on genetic distances, is given by

$$F_{ST} = 1 - \frac{H_w}{H_b}$$

where H_w is the mean number of differences between sequences sampled from the same subpopulation and H_b is the mean number of differences between sequences sampled from different subpopulations.

3.5 Results

We analysed several datasets of different types (SNP, microsatellite, sequence) and ploidy. Summary statistics for these datasets are presented in Table 1. Mycorrhiza,

Partitioned Mycorrhiza, PCA+RF, Partitioned PCA+RF, STRUCTURE and Admixture were applied in turn to each of these datasets to evaluate classification accuracy.

Table 1 - Summary statistics of the datasets on which assignment methods were tested

Dataset	Type	Ploidy	#Loci	#Pop	#Samples	#Samples / pop		FST	FSTd
						min	max		
Arabidopsis thaliana [140]	SNP	D	458 075	10	979	28	243	0.15	0.24
Brown Rat [144]	SNP	D	32 127	8	185	12	40	0.11	0.32
Gipsy Moth [145]	SNP	D	2 327	8	90	10	12	0.37	0.56
Human [146]	SNP	D	530 973	26	780	30	30	0.10	0.14
Rice [147]	SNP	D	458 475	20	740	20	50	0.16	0.29
Septoria musiva [141]	SNP	H	519 848	8	83	7	19	NA	0.31
Asian Ladybird [148]	STR	D	18	6	1318	87	501	0.03	0.05
Mycosphaerella fijiensis [149]	STR	H	21	21	678	12	66	NA	0.52
Oriental Fruit Moth [150]	STR	D	13	16	376	8	72	0.19	0.29
Yellow Fever Mosquito [151]	STR	D	12	13	1152	30	185	0.16	0.23
Barnacle [152]	SEQ	S	694	12	434	26	57	NA	0.14
Ebola [138]	SEQ	P	18 985	3	794	239	300	NA	0.09
HIV [153]	SEQ	S	3 287	5	628	28	150	NA	0.19
Seabird tick [154]	SEQ	S	456	7	432	5	131	NA	0.62

D = diploid, H = haploid, P=polyploid, S=consensus sequence of diploid organism

3.5.1 Assignment accuracy

Figure 3 shows the accuracies obtained by each tool on each dataset. Overall, Mycorrhiza and Partitioned Mycorrhiza, both of which were run with the same set of default parameters for each dataset, correctly assigned individuals to their populations of origin with greater accuracy than both STRUCTURE and Admixture on the majority of tested datasets. Furthermore, when compared only to their PCA counterparts, Mycorrhiza and Partitioned Mycorrhiza perform similarly or better on all tested datasets.

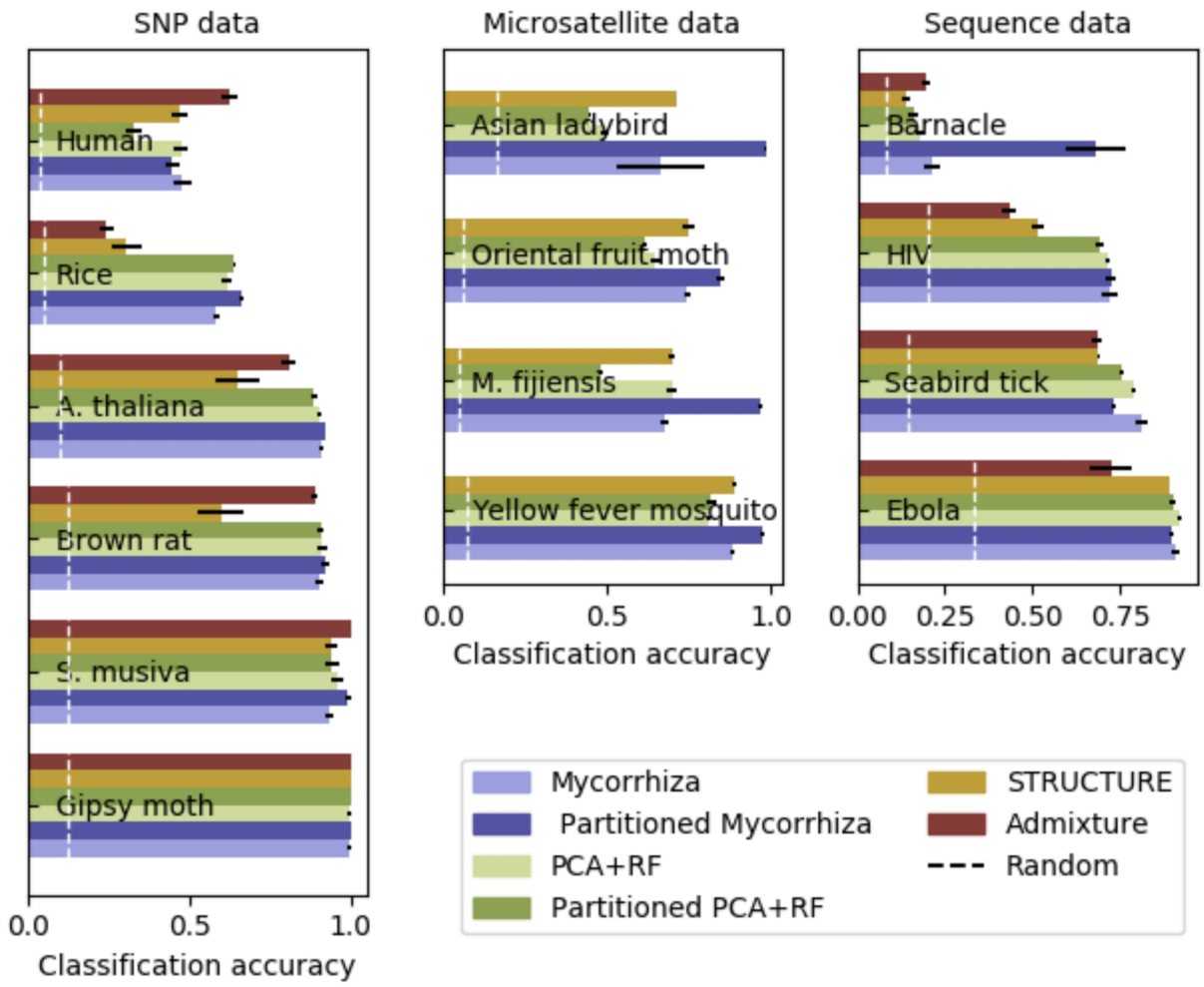


Figure 3 - Classification accuracy estimated using 5-fold cross-validation for the tested assignment methods.

With SNP data, Partitioned Mycorrhiza provided a slight advantage over Mycorrhiza and both PCA counterparts on the *A. thaliana*, Brown rat, *S. musiva* and Rice datasets. Generally, STRUCTURE and Admixture attained lower accuracy, with the exception of the human dataset for which Admixture outperformed our methods. With microsatellite loci, Partitioned Mycorrhiza considerably outperformed all other methods on all tested datasets, by a margin ranging from 9 to 27% in accuracy. Finally, with sequence data, Mycorrhiza, Partitioned Mycorrhiza, PCA+RF and Partitioned

PCA+RF performed nearly equally well on most datasets. The only exception is the Barnacle data, for which Partitioned Mycorrhiza provided a 39% increase in accuracy over the second best method.

3.5.2 Number of loci and partitioning parameters

We evaluated the impact of the number of loci used as input for all methods by executing them on randomly downsampled data sets (Figure 4). With SNP data, the accuracy of Mycorrhiza and PCA+RF (either with or without partitioning), as well as Admixture gradually increases with the number of loci, generally plateauing after 10,000 loci. Interestingly, the number of loci at which the plateau in accuracy is reached with SNP data seems to depend on number of populations present in the dataset.

Results are similar for microsatellites and sequence data, although most datasets are too small to reach a plateau in accuracy, suggesting that the inclusion of additional loci would further improve classification performance. Note that we expect that non-random loci selection would allow reaching a similar accuracy with much fewer loci. Surprisingly, the results obtained with STRUCTURE were somewhat more erratic.

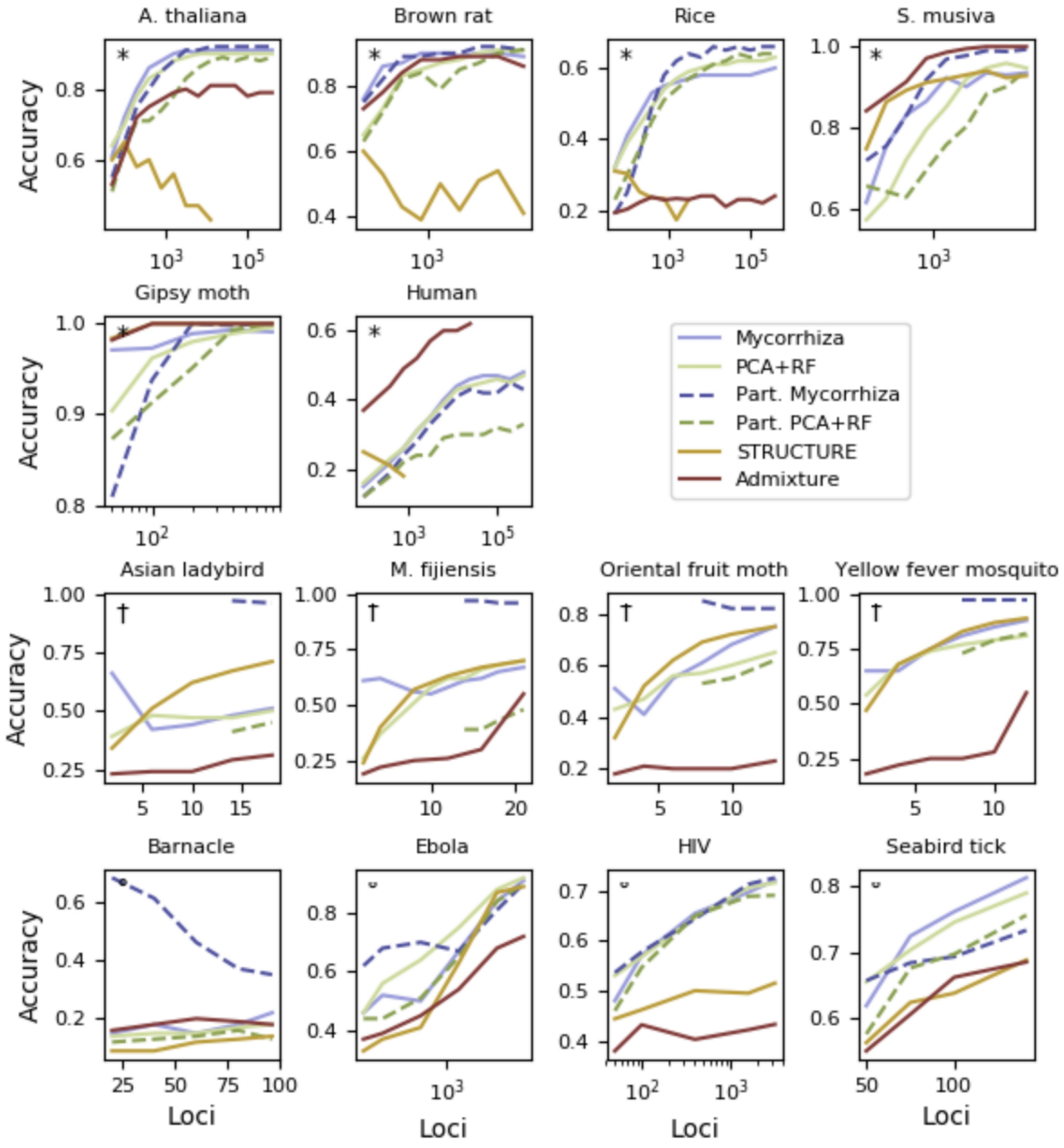


Figure 4 - Accuracy versus the number of randomly selected loci for Mycorrhiza, Partitioned Mycorrhiza, PCA+RF, Partitioned PCA+RF, STRUCTURE and Admixture. SNP (*), microsatellite (†) and sequence (○) datasets. STRUCTURE did not terminate within the allocated time on the A thaliana, rice and human datasets when the number of loci, and model complexity, were too high. Partitioned Mycorrhiza and Partitioned PCA+RF can only be executed with 1 or more locus per partition on the microsatellite datasets.

We also assessed the extent to which loci partitioning improves performance for Mycorrhiza and PCA+RF (Supplementary Figure 1). Setting the number of partitions to between 10 and 50 provided a consistent but moderate improvement in accuracy on nearly all SNP datasets, with the notable exception of the Human dataset, where best results were obtained without partitioning. Accuracy gains obtained by partitioning are most striking for the microsatellite datasets, where a number of partition around 10, corresponding to partitions of only 1 or 2 microsatellites each, yields a 30 to 50% improvement in accuracy. Results for sequence data were not as consistent and did not indicate strong trends, although setting the number of partitions to 10 is near-optimal for all datasets except for the oriental fruit moth. Based on these results, the default number of partitions for Partitioned Mycorrhiza was set to 10 for all data types.

3.5.3 Impact of population structure statistics on prediction accuracy

We next studied how population structure parameters (expected and observed heterozygosity, deviation from the Hardy-Weinberg equilibrium and the fixation index of subpopulations) differentially impact prediction accuracy of both Mycorrhiza and Partitioned Mycorrhiza compared to STRUCTURE and Admixture (Figure 5). As expected, population differentiation and levels of heterozygosity influence the ability of STRUCTURE and Admixture to capture population structure. Both variants of Mycorrhiza performed better than these two tools on SNP datasets for which sub-populations had a high fixation index. The advantage of our methods also increases proportionally to the level of heterozygosity of the total population and deviation from the Hardy-Weinberg equilibrium, but decreases proportionally to the average sub-

population heterozygosity. With microsatellite and sequence data, the trends are not as clear and, consequently, more datasets would be needed to draw conclusions.

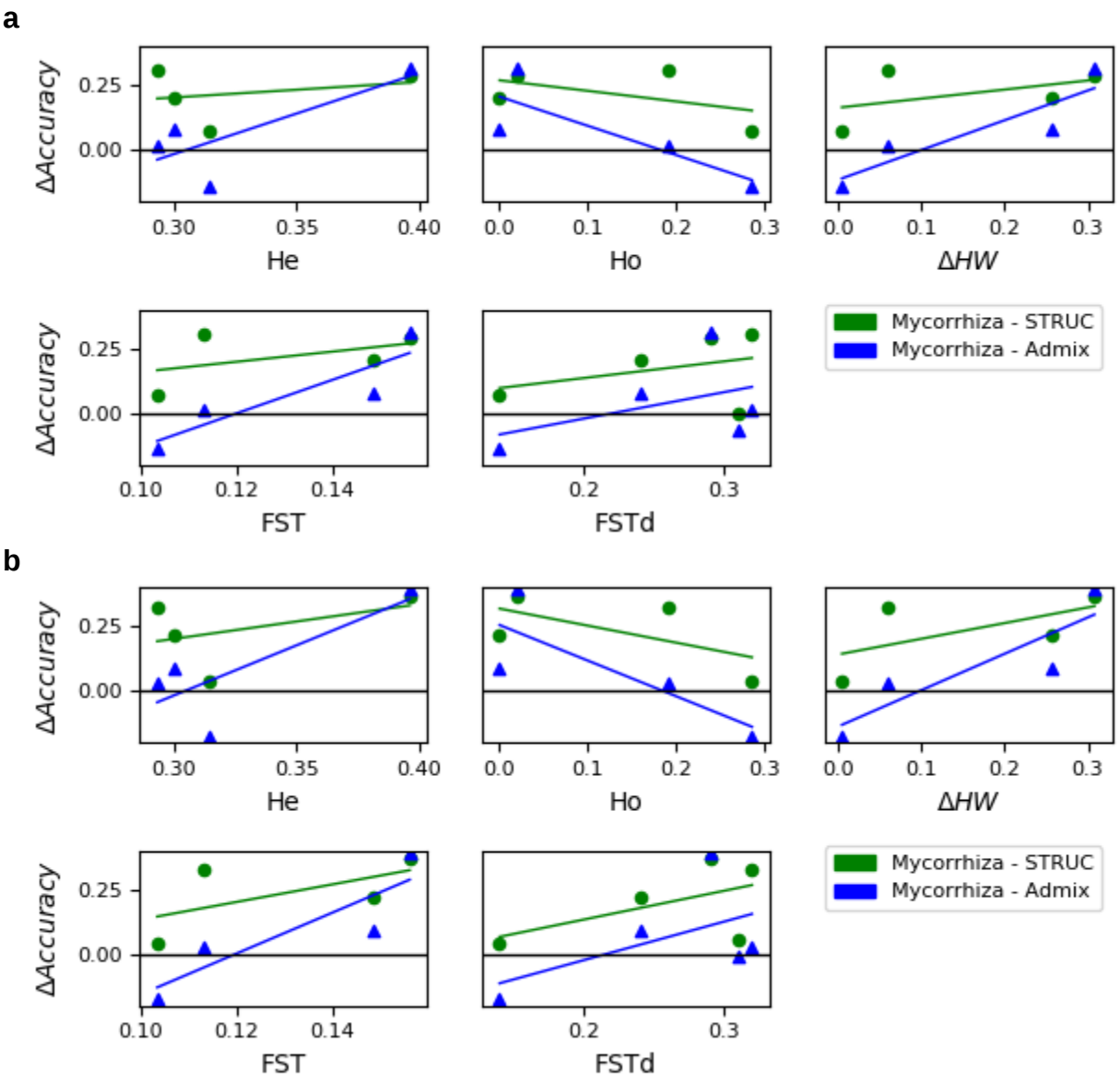


Figure 5 - Difference in assignment accuracy on SNP datasets only between (a) Mycorrhiza and STRUCTURE or Mycorrhiza and Admixture and between (b) Partitioned Mycorrhiza and STRUCTURE or Partitioned Mycorrhiza and Admixture, as a function of the expected heterozygosity (He), observed heterozygosity (Ho) and deviation from the Hardy-Weinberg equilibrium (ΔHW), and average population fixation index, calculated from heterozygosity (FST) and from genetic distances (FSTd).

3.5.4 Estimation of mixture proportions

Although Mycorrhiza and Partitioned Mycorrhiza was not specifically designed to estimate population mixture proportions, we observe that the population membership probabilities it outputs closely resemble those obtained with STRUCTURE, which is widely acknowledged to be accurate at estimating population mixture proportions. An example output representing mixture proportions obtained by all methods from the *S. musiva* dataset with 800 loci is shown in Figure 6. Notice also how PCA+RF appears to overestimate the contribution of secondary populations, even on datasets for which it performs well in terms of classification accuracy.

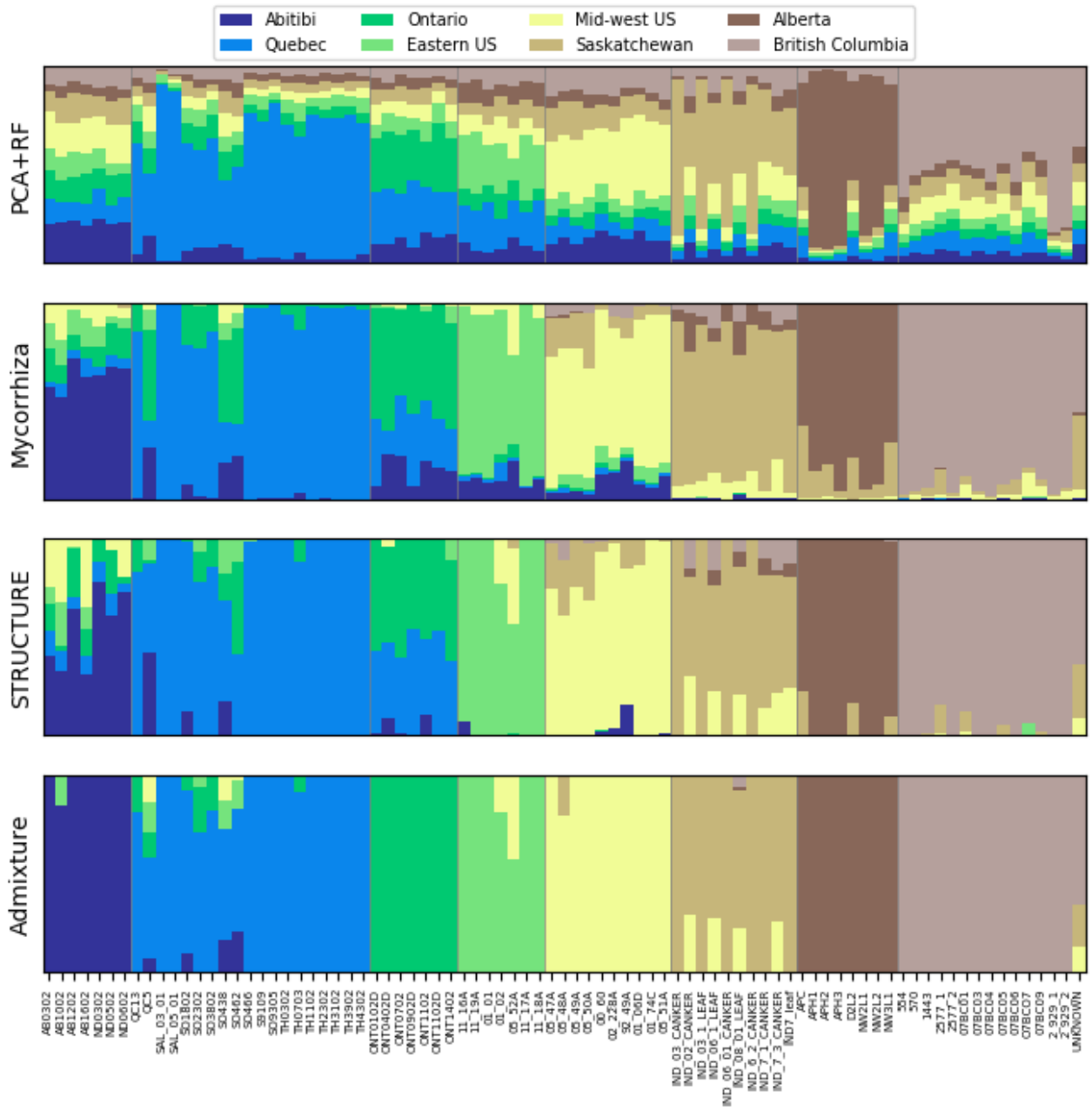


Figure 6 - Representative output of mixture proportions estimated by all methods applied to the S. musiva dataset with 800 loci.

3.5.5 Runtime

The runtime of Mycorrhiza with 5-fold cross-validation on each dataset is presented in Figure 7. Time was calculated with no partitions on a single core of a standard laptop computer (Intel(R) Core(TM) i5-7300HQ CPU @ 2.50GHz). All runtimes were under an hour and most were under 10 minutes. They increase cubically with the number of samples, and linearly with the number of loci, although under approximately 10 000 loci only the former has a significant impact. The number of populations in the datasets had no discernable impact on runtime. Partitioned Mycorrhiza with P partitions takes approximately P times longer to run than Mycorrhiza. For comparison, the runtime of STRUCTURE was at least five times larger, resulting in certain datasets taking several days, weeks or months to analyze. Admixture, is considerably faster than STRUCTURE and was comparable to Mycorrhiza in execution time (under an hour), although slightly slower.

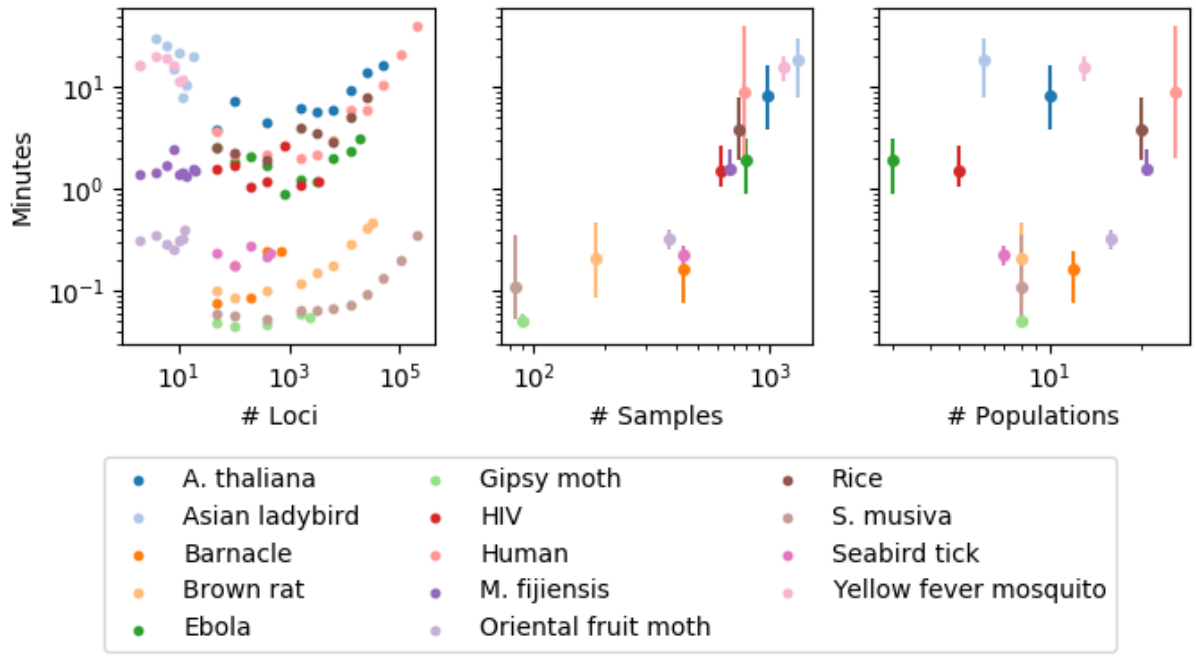


Figure 7 - Runtime of Mycorrhiza on each dataset analyzed, as a function of the number of loci, samples and populations.

3.6 Discussion and Conclusion

We introduce Mycorrhiza, a machine learning method making use of phylogenetic networks to assign multilocus genotypes to their geographical origin.

Mycorrhiza proved to be highly flexible, outperforming or equaling the other methods on all data types. On a side note, this flexibility will likely allow for Mycorrhiza to be applied to other, non-genetic, data types possessing tree-like relationships. Indeed, phylogenetic networks have in fact been used to model relationships between languages [74,75] and cultural artifacts [76].

Mycorrhiza proved to be highly accurate, flexible and robust, outperforming or equaling in accuracy the most popular existing methods (STRUCTURE, Admixture) on

a variety of datasets based on SNPs, microsatellites, and sequence. The consistency of results produced by Mycorrhiza provides a considerable advantage for the method: default parameters will provide near-optimal results in almost all cases. Mycorrhiza's accuracy improves gradually as the number of markers available increases, to eventually plateau, but never decrease. Furthermore, methods such as STRUCTURE and Admixture depend on assumptions about population structure that are often unmet in practice. Mycorrhiza is not directly dependent on these assumptions, and indeed, the data sets where populations exhibit high fixation index or strong deviation from Hardy-Weinberg equilibrium are those where the benefit of Mycorrhiza is most striking. Unsurprisingly, Admixture and, to a lesser extent, STRUCTURE performed poorly on sequence data due to the fact that allelic frequencies of diploid or polyploid genotypes are impossible to estimate from a reduced sequence. Mycorrhiza, being based on distances rather than allelic frequencies, is seemingly less affected by this loss of information. Finally, although it is not explicitly designed for that purpose, Mycorrhiza produces admixture proportion estimates qualitatively similar to those of STRUCTURE, although additional work would be needed to quantify the accuracy of those estimates. Mycorrhiza is thus a tool that is simple and straightforward to use, requiring little or no parameter optimization, and exhibiting a high degree of robustness to the type of data at hand and the parameters of the population structure. Mycorrhiza would, in particular, be a tool of choice when dealing with geopolitically, rather than genetically, defined populations. For example, this is necessary when assignment test must be used in international trade discussions or to coordinate risk mitigation of invasive species between countries.

With the increasing throughput and affordability of DNA sequencing, genotype assignment problems are quickly becoming very large, both in terms of the number of samples and number of loci they contain. Computationally efficient algorithms are thus more necessary than ever. Mycorrhiza's running time, which is dominated by that of the $O(n^3)$ NeighborNet algorithm [61], remains moderate (less than one hour) on even the largest datasets available today. Moreover, improvements to the NeighborNet algorithm are published on a regular basis, including a recent report of 2-fold speed-up and 6-fold reduction in the memory footprint [155]. Notably, unlike for Bayesian methods, the numbers of loci and of populations in the dataset have a negligible effect on runtime. Obviously, running Partitioned Mycorrhiza with a large number of partitions increases running time proportionally, but the process is easily parallelized. Overall, Mycorrhiza not only provides better classification accuracy in most datasets tested, but also reduces computation time considerably.

The difference in accuracy between Mycorrhiza and STRUCTURE is particularly strong for microsatellite data. This could be due to several factors. First is the higher information content of each microsatellite locus, on average approximately 14 alleles per locus for the data used in our study compared to 2 for the SNP datasets. Second, microsatellites are mostly non-coding, whereas SNP are mostly coding. These genetic marker types are thus under very different levels of selective pressure. Lastly, all studies from which the microsatellite data were taken are mainly about population genetics and the markers were chosen accordingly. The same cannot be said about most of the SNP datasets used in this study.

Although we limited our performance comparison to the two most widely used approaches in the field, some of the data sets analyzed here had been previously analyzed by their authors using other tools. This provides additional points of comparison, albeit in a less rigorous framework. Picq *et al.* analyzed their gipsy moth data sets using DAPC [105], reporting 100% classification accuracy with all 2327 SNPs, as well as with a selected set of 48 SNPs. We produced similar results, obtaining perfect accuracy with as few as 200 randomly selected SNPs (99% with 100 randomly selected SNPs and 97% with 50). Results obtained for the yellow fever mosquito microsatellite data set are also consistent with the authors' analyses. Using GeneClass2 [156] on 10 non-problematic loci, the authors reported 87.7% classification accuracy. In our hands, STRUCTURE obtained a similar performance (88.9%), but Partitioned Mycorrhiza did much better, obtaining an accuracy of 98.4%.

3.6.1 Future work

While already fast and accurate, several directions would be worth investigating to improve Mycorrhiza and further broaden its range of applications. Firstly, only phylogenetic networks built using the NeighborNet algorithm were used in this study. Although this algorithm is known to produce well resolved and simple networks in most conditions, it would be interesting to study whether other network building algorithms could perform better, and under what conditions. Similarly, classifiers other than random forests, such as deep neural networks, may be advantageous for data sets with an extremely large number of samples; replacing our random forest predictor by such an alternative would be straightforward. Finally, Mycorrhiza currently makes no use of the

weights assigned to splits by NeighborNet. Those weight could instead be used, e.g. to bias the sampling of the corresponding features by the random forest algorithm.

It is also worth noting that Mycorrhiza is applicable not only to genotypic data, but also to any other type of population-specific traits. All that would be needed is the definition of a suitable phenotypic distance measure. In fact, phylogenetic networks have even been used to model relationships among languages [74,75] and cultural artifacts [76], and Mycorrhiza may be applicable to these data types as well.

In conclusion, by combining sophisticated phylogenetic network reconstruction algorithms with machine learning approaches, Mycorrhiza represents a novel solution to the genotype assignment problem. Its accuracy, scalability, and most importantly its robustness to data types and sizes, and properties of underlying population structure, should make it an attractive solution for a wide array of population genetics researchers.

3.7 Supplementary material

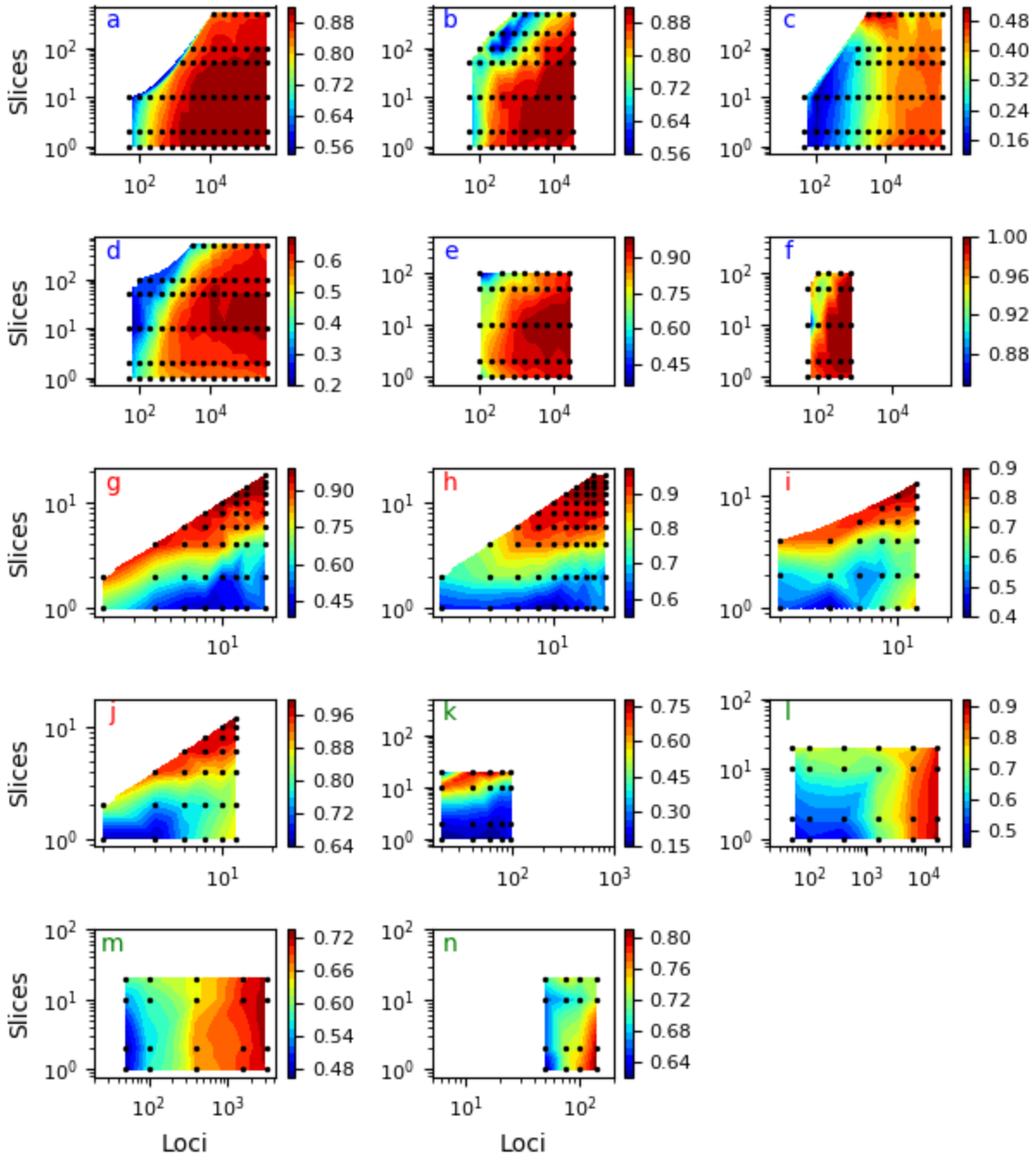


Figure 8 - Linear interpolation of the classification accuracy versus the number of loci and the number of partitions for Mycorrhiza on datasets (a) A. thaliana, (b) Brown rat, (c) Human, (d) rice, (e) S. musiva, (f) Gipsy moth, (g) Asian ladybird, (h) M fijiensis, (i) Oriental fruit moth, (j) Yellow fever mosquito, (k) Barnacle, (l) Ebola, (m) HIV (n) Seabird tick.

Chapter 4

Discussion and conclusion

The BioSAFE project aims to deploy genomics tools to reduce the risk posed by invasive species in Canada. Central to the effort of this project was the development of a machine learning algorithm to predict the geographical origin of biological samples. We presented Mycorrhiza, a machine learning algorithm-based dimensionality reduction of the feature space and phylogenetic relationships.

Mycorrhiza and Partionned Mycorrhiza meet all the requirements put forth by the BioSAFE project by capturing reticulate evolutionary events and being computationally efficient. Moreover, we tested our method for classification accuracy and admixture analysis on a variety of real-world datasets in comparison with two of the most used assessment methods in the field. Having identified the shortcomings and properties of Bayesian assignment algorithms, were able to demonstrate quantitatively in which cases Mycorrhiza possess an advantage in comparison with currently used methods. Moreover, by also replacing the phylogenetic network inference with a PCA analysis, we also demonstrated that Mycorrhiza and Partitioned Mycorrhiza do in fact capture phylogenetic structure, and consequently provide better classification accuracy.

Notably, by replacing model-based algorithms with a model-free algorithm our work inscribes itself in a paradigm shift in the biological sciences engendered by modern developments in the machine learning field [157]. In addition to its role in the BioSAFE project, we released Mycorrhiza as an open-source in hopes of stimulating interest in the method and encouraging further improvements.

4.1 Directions for future work

As previously stated, it would be interesting to determine if Mycorrhiza can be successfully applied to different types of data presenting phylomemetic structure. Obvious examples include languages and cultural artefacts, but we can also imagine how the algorithm could aid in predicting binding affinities of small molecules in drug discovery for example. The reconstruction of evolutionary relationships from sequences in genomic and proteomic databases has already been used to generate leads in drug discovery [158].

In fact, many different examples of phylogenetic structure being harnessed to aid in prediction of certain traits can be found. In one case, it was shown that producing a phylogeny from ethnomedicinal biological responses can aid in the predicting the medicinal properties of plants [159]. In another similar case, the authors found interesting correlations when the medicinal uses of plant families were mapped onto the corresponding molecular phylogeny [160]. In yet another case, the phylogenetic profiles of gene conservation were used to aid in predicting the subcellular localization of proteins [161]. However, in each of these cases phylogenetic trees are employed as opposed to phylogenetic networks. It would thus be interesting to see if Mycorrhiza can be applied successfully to this type of problem or, more generally, phenotype prediction problems from molecular or not-molecular based phylogenies. As a matter of fact, with the need for pathogenicity and risk prediction, the BioSAFE project presents an exceptional opportunity to extend the applicability of Mycorrhiza.

Currently, there are no methods to insert one or more taxa into an existing phylogenetic network. Developments in this direction could prevent the need to retrain

Mycorrhiza on the entirety of the data when predictions need to be done on new examples. In informal tests, we achieved interesting results for the insertion of a taxa in an existing phylogenetic network by employing simulated annealing and evolutionary algorithms. This would allow for a phylogenetic network to be built from only the training samples. The insertion algorithm could then be used to produce a split placement for the testing samples, which would be used as input the trained model to make predictions. The concern here is not to avoid potential overfitting, but for cases where predictions are done “online”, from a stream of new samples. In such cases, eliminating the need to retrain the model for each new sample would provide considerable efficiency improvement.

References

1. Paini DR, Sheppard AW, Cook DC, De Barro PJ, Worner SP, Thomas MB. Global threat to agriculture from invasive species. Proc Natl Acad Sci U S A. 2016;113: 7575–7579.
2. Sakai AK, Allendorf FW, Holt JS, Lodge DM, Molofsky J, With KA, et al. The Population Biology of Invasive Species. Annu Rev Ecol Syst. Annual Reviews; 2001;32: 305–332.
3. Colautti RI, Bailey SA, van Overdijk CDA, Amundsen K, MacIsaac HJ. Characterised and Projected Costs of Nonindigenous Species in Canada. Biol Invasions. 2006;8: 45–59.
4. Templeton AR. Population genetics and microevolutionary theory. Hoboken, NJ: Wiley-Liss; 2006.
5. Manel S, Gaggiotti OE, Waples RS. Assignment methods: matching biological questions with appropriate techniques. Trends Ecol Evol. 2005;20: 136–142.
6. Waples RS, Gaggiotti O. INVITED REVIEW: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol. Wiley Online Library; 2006;15: 1419–1439.
7. Population genetics - Latest research and news | Nature [Internet]. 25 Jun 2018 [cited 25 Jun 2018]. Available: https://www.nature.com/subjects/population-genetics
8. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000-6;155: 945–959.
9. Johnson MTJ, Stinchcombe JR. An emerging synthesis between community ecology and evolutionary biology. Trends Ecol Evol. 2007;22: 250–257.
10. Harrison RG, Larson EL. Hybridization, introgression, and the nature of species boundaries. J Hered. 2014;105 Suppl 1: 795–809.
11. Arnold ML. Natural Hybridization and Evolution. Oxford University Press, USA; 1997.
12. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. Elsevier; 2005;20: 229–237.
13. Mooney HA, Cleland EE. The evolutionary impact of invasive species. Proc Natl Acad Sci U S A. 2001;98: 5446–5451.
14. Dhillon B, Feau N, Aerts AL, Beauseigle S, Bernier L, Copeland A, et al. Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. Proc Natl Acad Sci U S A. 2015;112: 3451–3456.
15. Ellstrand NC, Schierenbeck KA. Hybridization as a stimulus for the evolution of invasiveness in plants? Proc Natl Acad Sci U S A. 2000;97: 7043–7050.
16. Weinberg W. Über den nachweis der vererbung beim menschen. 1908.
17. Hardy GH. MENDELIAN PROPORTIONS IN A MIXED POPULATION. Science. 1908;28: 49–50.
18. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 1978;89: 583–590.
19. Nei M. F-statistics and analysis of gene diversity in subdivided populations. Ann Hum Genet. Wiley Online Library; 1977;41: 225–233.
20. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009;10: 639–650.
21. Yang R-C. ESTIMATING HIERARCHICAL F-STATISTICS. Evolution. 1998;52: 950–956.
22. Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, Lareu MV. An overview of STRUCTURE: applications, parameter settings, and supporting software. Front Genet. 2013;4: 98.
23. Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, et al. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with

admixed ancestry from two or more populations. Hum Biol. 2012;84: 343–364.
24. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5: e1000519.
25. Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, Bruse S. Softwares and methods for estimating genetic ancestry in human populations. Hum Genomics. 2013;7: 1.
26. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164: 1567–1587.
27. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 2009;9: 1322–1332.
28. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 2007;7: 574–578.
29. Falush D, van Dorp L, Lawson D. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots [Internet]. bioRxiv. 2016. p. 066431. doi:10.1101/066431
30. Neophytou C. Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: effects of asymmetric phylogenies and asymmetric sampling schemes. Tree Genet Genomes. Springer Berlin Heidelberg; 2014;10: 273–285.
31. Puechmaille SJ. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Mol Ecol Resour. 2016;16: 608–627.
32. Zhou H, Alexander D, Lange K. A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat Comput. 2011;21: 261–273.
33. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664.
34. Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics. 1999;153: 1989–2000.
35. Degen B, Blanc-Jolivet C, Stierand K, Gillet E. A nearest neighbour approach by genic distance to the assignment of individuals to geographic origin [Internet]. bioRxiv. 2016. p. 087833. doi:10.1101/087833
36. König IR, Auerbach J, Gola D, Held E, Holzinger ER, Legault M-A, et al. Machine learning and data mining in complex genomic data—a review on the lessons learned in Genetic Analysis Workshop 19. BMC genetics. BioMed Central; 2016. p. S1.
37. Liang Y, Kelemen A. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. Stat Surv. 2008;2: 43–60.
38. Bertolini F, Galimberti G, Calò DG, Schiavo G, Matassino D, Fontanesi L. Combined use of principal component analysis and random forests identify population-informative single nucleotide polymorphisms: application in cattle breeds. J Anim Breed Genet. 2015;132: 346–356.
39. Nguyen T-T, Huang J, Wu Q, Nguyen T, Li M. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics. 2015;16 Suppl 2: S5.
40. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23: 2507–2517.
41. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci Rep. 2015;5: 10312.
42. Morrison DA. Is the Tree of Life the Best Metaphor, Model, or Heuristic for Phylogenetics?

Syst Biol. 2014;63: 628–638.
43. Gontier N. Reticulate Evolution Everywhere. In: Gontier N, editor. Reticulate Evolution: Symbiogenesis, Lateral Gene Transfer, Hybridization and Infectious Heredity. Cham: Springer International Publishing; 2015. pp. 1–40.
44. Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, et al. Prokaryotic evolution and the tree of life are two different things. Biol Direct. 2009;4: 34.
45. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 2006;23: 254–267.
46. Huson DH, Rupp R, Scornavacca C. Introduction to phylogenetic networks. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press; 2010. pp. 68–84.
47. Huson DH, Scornavacca C. A survey of combinatorial methods for phylogenetic networks. Genome Biol Evol. 2011;3: 23–35.
48. Networks of affinity rather than genealogy [Internet]. [cited 25 Jun 2018]. Available: http://phylonetworks.blogspot.com/2012/05/networks-of-affinity-rather-than.html
49. Semple C, Steel M. Cyclic permutations and evolutionary trees. Adv Appl Math. 2004;32: 669–680.
50. Huson DH, Rupp R, Scornavacca C. Splits and unrooted phylogenetic networks. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press; 2010. pp. 87–126.
51. Dress AWM, Huber KT, Moulton V. An Exceptional Split Geometry. Ann Comb. Birkhäuser Verlag; 2000;4: 1–11.
52. Bandelt H-J, Dress AWM. A canonical decomposition theory for metrics on a finite set. Adv Math . 1992;92: 47–105.
53. Gambette P, Huber KT, Scholz GE. Bridging the gap between rooted and unrooted phylogenetic networks [Internet]. arXiv [math.CO]. 2015. Available: http://arxiv.org/abs/1511.08387
54. Hassanzadeh R, Eslahchi C, Sung W-K. Constructing phylogenetic supernetworks based on simulated annealing. Mol Phylogenet Evol. 2012;63: 738–744.
55. Yang J, Grünewald S, Xu Y, Wan X-F. Quartet-based methods to reconstruct phylogenetic networks. BMC Syst Biol. 2014;8: 21.
56. Grünewald S, Forslund K, Dress A, Moulton V. QNet: An Agglomerative Method for the Construction of Phylogenetic Networks from Weighted Quartets. Mol Biol Evol. 2007;24: 532–538.
57. Eslahchi C, Hassanzadeh R, Mottaghi E, Habibi M, Pezeshk H, Sadeghi M. Constructing circular phylogenetic networks from weighted quartets using simulated annealing. Math Biosci. 2012;235: 123–127.
58. Yang J, Grünewald S, Wan X-F. Quartet-Net: A Quartet-Based Method to Reconstruct Phylogenetic Networks. Mol Biol Evol. 2013;30: 1206–1217.
59. Bandelt HJ, Dress AW. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol. 1992;1: 242–252.
60. Velasco JD, Sober E. Testing for treeness: lateral gene transfer, phylogenetic inference, and model selection. Biol Philos. 2010;25: 675–687.
61. Bryant D, Moulton V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Mol Biol Evol. 2004;21: 255–265.
62. Bryant D, Moulton V. NeighborNet: An Agglomerative Method for the Construction of Planar Phylogenetic Networks. Algorithms in Bioinformatics. Springer Berlin Heidelberg; 2002. pp. 375–391.
63. Bryant D, Moulton V, Spillner A. Consistency of the Neighbor-Net Algorithm. Algorithms Mol Biol. 2007;2: 8.

64. Roch S, Wang K-C. Circular Networks from Distorted Metrics [Internet]. arXiv [q-bio.PE]. 2017. Available: http://arxiv.org/abs/1707.05722
65. Levy D, Pachter L. The neighbor-net algorithm. Adv Appl Math. 2011;47: 240–258.
66. Balvočūtė M, Spillner A, Moulton V. FlatNJ: A Novel Network-Based Approach to Visualize Evolutionary and Biogeographical Relationships. Syst Biol. 2014;63: 383–396.
67. Eslahchi C, Habibi M, Hassanzadeh R, Mottaghi E. MC-Net: a method for the construction of phylogenetic networks based on the Monte-Carlo method. BMC Evol Biol. 2010;10: 254.
68. Morrison DA. Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis. WIREs Data Mining Knowl Discov. 2014;4: 296–312.
69. Fioravanti D, Giarratano Y, Maggio V, Agostinelli C, Chierici M, Jurman G, et al. Phylogenetic convolutional neural networks in metagenomics. BMC Bioinformatics. 2018;19: 49.
70. Chernomor O, Klaere S, von Haeseler A, Minh BQ. Split Diversity: Measuring and Optimizing Biodiversity Using Phylogenetic Split Networks. In: Pellens R, Grandcolas P, editors. Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis. Cham: Springer International Publishing; 2016. pp. 173–195.
71. Wicke K, Fischer M. Phylogenetic diversity and biodiversity indices on phylogenetic networks [Internet]. arXiv [q-bio.PE]. 2017. Available: http://arxiv.org/abs/1706.05279
72. Volkmann L, Martyn I, Moulton V, Spillner A, Mooers AO. Prioritizing populations for conservation using phylogenetic networks. PLoS One. 2014;9: e88945.
73. Lundin D, Poole AM, Sjöberg B-M, Högbom M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. J Biol Chem. 2012;287: 20565–20575.
74. Greenhill SJ, Atkinson QD, Meade A, Gray RD. The shape and tempo of language evolution. Proc Biol Sci. 2010;277: 2443–2450.
75. Gray RD, Bryant D, Greenhill SJ. On the shape and fabric of human history. Philos Trans R Soc Lond B Biol Sci. 2010;365: 3923–3933.
76. Howe CJ, Windram HF. Phylomemetics--evolutionary analysis beyond the gene. PLoS Biol. 2011;9: e1001069.
77. Dawkins R. The Selfish Gene : 30th Anniversary Edition--with a new Introduction by the Author. Oxford University Press, USA; 2006.
78. Greenhill SJ, Currie TE, Gray RD. Does horizontal transmission invalidate cultural phylogenies? Proc Biol Sci. 2009;276: 2299–2306.
79. Ainsworth GC. Ainsworth & Bisby's Dictionary of the Fungi. CABI; 2008.
80. Larraín MA, Díaz NF, Lamas C, Uribe C, Araneda C. Traceability of mussel (Mytilus chilensis) in southern Chile using microsatellite molecular markers and assignment algorithms. Exploratory survey. Food Res Int. 2014;62: 104–110.
81. Schwartz TS, Karl SA. Population Genetic Assignment of Confiscated Gopher Tortoises. J Wildl Manage. The Wildlife Society; 2008;72: 254–259.
82. Glover KA, Hansen MM, Skaala Ø. Identifying the source of farmed escaped Atlantic salmon (Salmo salar): Bayesian clustering analysis increases accuracy of assignment. Aquaculture. 2009;290: 37–46.
83. Lorenzini R, Cabras P, Fanelli R, Carboni GL. Wildlife molecular forensics: identification of the Sardinian mouflon using STR profiling and the Bayesian assignment test. Forensic Sci Int Genet. 2011;5: 345–349.
84. Millions DG, Swanson BJ. An Application of Manel's Model: Detecting Bobcat Poaching in Michigan. Wildl Soc Bull. The Wildlife Society; 2006;34: 150–155.
85. Stewart KR, James MC, Roden - Journal of Animal … S, 2013. Assignment tests, telemetry and tag-recapture data converge to identify natal origins of leatherback turtles foraging in Atlantic Canadian waters. Wiley Online Library. 2013; Available:

http://onlinelibrary.wiley.com/doi/10.1111/1365-2656.12056/full
86. Ibañez-Justicia A, Gloria-Soria A, den Hartog W, Dik M, Jacobs F, Stroo A. The first detected airline introductions of yellow fever mosquitoes (Aedes aegypti) to Europe, at Schiphol International airport, the Netherlands. Parasit Vectors. 2017;10: 603.
87. Johansson ML, Dufour BA, Wellband KW, Corkum LD, MacIsaac HJ, Heath DD. Human-mediated and natural dispersal of an invasive fish in the eastern Great Lakes. Heredity . 2018; doi:10.1038/s41437-017-0038-x
88. Larraín MA, Zbawicka M, Araneda C, Gardner J, Wenne R. Native and invasive taxa on the Pacific coast of South America: Impacts on aquaculture, traceability and biodiversity of blue mussels (Mytilus spp.). Evol Appl. Wiley Online Library; 2018;11: 298–311.
89. Michalecka M, Masny S, Leroy T, Puławska J. Population structure of Venturia inaequalis, a causal agent of apple scab, in response to heterogeneous apple tree cultivation. BMC Evol Biol. 2018;18: 5.
90. Dauphinais JD, Miller LM, Swanson RG, Sorensen PW. Source–sink dynamics explain the distribution and persistence of an invasive population of common carp across a model Midwestern watershed. Biol Invasions. Springer International Publishing; 2018; 1–16.
91. Kalinowski ST. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity . 2011;106: 625–632.
92. Perrings C, Burgiel S, Lonsdale M, Mooney H, Williamson M. International cooperation in the solution to trade-related invasive species risks. Ann N Y Acad Sci. Wiley Online Library; 2010;1195: 198–212.
93. Novembre J. Pritchard, Stephens, and Donnelly on Population Structure. Genetics. 2016;204: 391–393.
94. Jonathan K. Pritchard, Xiaoquan Wen, Daniel Falush. Documentation for structure software [Internet]. Department of Human Genetics University of Chicago, Department of Statistics University of Oxford; 2010 Feb. Available: https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pdf
95. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197: 573–589.
96. Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet. Springer Netherlands; 2006;7: 295–302.
97. Wang J. The computer program structure for assigning individuals to populations: easy to use but easier to misuse. Mol Ecol Resour. 2017;17: 981–990.
98. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909.
99. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008;40: 646–649.
100. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science. 1978;201: 786–792.
101. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2: e190.
102. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009;5: e1000686.
103. Jombart T, Pontier D, Dufour A-B. Genetic markers in the playground of multivariate analysis. Heredity . 2009;102: 330–341.
104. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers.

Bioinformatics. 2008;24: 1403–1405.
105. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11: 94.
106. Liu N, Zhao H. A non-parametric approach to population structure inference using multilocus genotypes. Hum Genomics. 2006;2: 353–364.
107. Lee C, Abdool A, Huang C-H. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics. 2009;10 Suppl 1: S73.
108. Distortions and artifacts in Principal Components Analysis for analysis of genome data [Internet]. [cited 8 Jun 2018]. Available: http://phylonetworks.blogspot.com/2012/12/distortions-and-artifacts-in-pca.html
109. Continued misuse of PCA in genomics studies [Internet]. [cited 8 Jun 2018]. Available: https://phylonetworks.blogspot.com/2016/05/continued-misuse-of-pca-in-genomics.html
110. Grimm D. Comparing neighbour-nets and PCA graphs – the example of Mediterranean oaks [Internet]. [cited 8 Jun 2018]. Available: http://phylonetworks.blogspot.com/2018/03/comparing-neighbour-nets-and-pca-graphs.html
111. Networks can outperform PCA ordinations in phylogenetic analysis [Internet]. [cited 8 Jun 2018]. Available: http://phylonetworks.blogspot.com/2012/12/networks-can-outperform-pca-ordinations.html
112. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27: 3070–3071.
113. Beugin M-P, Gayet T, Pontier D, Devillard S, Jombart T. A fast likelihood solution to the genetic clustering problem. Hansen T, editor. Methods Ecol Evol. 2018;9: 1006–1016.
114. Guillot G, Leblois R, Coulon A, Frantz AC. Statistical methods in spatial genetics. Mol Ecol. 2009;18: 4734–4756.
115. Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol. 2003;18: 189–197.
116. Balkenhol N, Waits LP, Dezzani RJ. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. Ecography . 2009;32: 818–830.
117. Keogh E, Mueen A. Curse of Dimensionality. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US; 2017. pp. 314–315.
118. Kooperberg C, LeBlanc M, Obenchain V. Risk prediction using genome-wide association studies. Genet Epidemiol. 2010;34: 643–652.
119. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 2009;18: 3525–3531.
120. Huang L-C, Hsu S-Y, Lin E. A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. J Transl Med. 2009;7: 81.
121. Vlachos M. Dimensionality Reduction. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US; 2017. pp. 354–361.
122. Bandelt HJ, Forster P, Sykes BC, Richards MB. Mitochondrial portraits of human populations using median networks. Genetics. 1995;141: 743–753.
123. Hendy MD, Penny D. Spectral analysis of phylogenetic data. J Classification. 1993;10: 5–24.
124. Gambette P, Huson DH. Improved layout of phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform. 2008;5: 472–479.
125. Spillner A, Nguyen B, Moulton V. Constructing and Drawing Regular Planar Split Networks. IEEE/ACM Trans Comput Biol Bioinform. 2012;9: 395–407.

126. Bryant D, Dress A. Linearly independent split systems. European J Combin. 2007;28: 1814–1831.
127. Frank B. Ueber die auf Wurzelsymbiose beruhende Ernährung gewisser Bäume durch unterirdische Pilze. 1885.
128. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics. 2012;28: 2685–2686.
129. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.
130. Opitz D, Maclin R. Popular Ensemble Methods: An Empirical Study. 1. 1999;11: 169–198.
131. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20: 832–844.
132. Rokach L. Genetic algorithm-based feature set partitioning for classification problems. Pattern Recognit. 2008;41: 1676–1700.
133. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. Warwick & York; 1933;24: 417.
134. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 1998;14: 68–73.
135. distruct: graphical display of population structure [Internet]. [cited 15 May 2018]. Available: https://web.stanford.edu/group/rosenberglab/distruct.html
136. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002;298: 2381–2385.
137. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12: 246.
138. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 2012;40: D593–8.
139. Rice Diversity [Internet]. [cited 25 Mar 2018]. Available: http://www.ricediversity.org/index.cfm
140. [No title] [Internet]. [cited 5 May 2018]. Available: http://1001genomes.org
141. Sakalidis ML, Feau N, Dhillon B, Hamelin RC. Genetic patterns reveal historical and contemporary dispersal of a tree pathogen. Biol Invasions. 2016;18: 1781–1799.
142. Dryad Digital Repository - Dryad [Internet]. [cited 12 Feb 2018]. Available: https://datadryad.org/
143. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132: 583–589.
144. Puckett EE, Park J, Combs M, Blum MJ, Bryant JE, Caccone A, et al. Global population divergence and admixture of the brown rat (Rattus norvegicus). Proc Biol Sci. 2016;283. doi:10.1098/rspb.2016.1762
145. Picq S, Keena M, Havill N, Stewart D, Pouliot E, Boyle B, et al. Assessing the potential of genotyping-by-sequencing-derived single nucleotide polymorphisms to identify the geographic origins of intercepted gypsy moth (Lymantria dispar) specimens: A proof-of-concept study. Evol Appl. 2018;11: 325–339.
146. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526: 68–74.
147. McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, et al. Open access resources for genome-wide association mapping in rice. Nat Commun. 2016;7: 10532.
148. Lombaert E, Guillemaud T, Lundgren J, Koch R, Facon B, Grez A, et al. Complementarity of statistical treatments to reconstruct worldwide routes of invasion: the case of the Asian

ladybird Harmonia axyridis. Mol Ecol. 2014;23: 5979–5997.
149. Robert S, Ravigne V, Zapater M-F, Abadie C, Carlier J. Contrasting introduction scenarios among continents in the worldwide invasion of the banana fungal pathogen Mycosphaerella fijiensis. Mol Ecol. 2012;21: 1098–1114.
150. Kirk H, Dorn S, Mazzi D. Worldwide population genetic structure of the oriental fruit moth (Grapholita molesta), a globally invasive pest. BMC Ecol. 2013;13: 12.
151. Brown JE, McBride CS, Johnson P, Ritchie S, Paupy C, Bossin H, et al. Worldwide patterns of genetic differentiation imply multiple “domestications” of Aedes aegypti, a major vector of human diseases. Proc Biol Sci. 2011;278: 2446–2454.
152. Wrange A-L, Charrier G, Thonig A, Alm Rosenblad M, Blomberg A, Havenhand JN, et al. The Story of a Hitchhiker: Population Genetic Patterns in the Invasive Barnacle Balanus(Amphibalanus) improvisus Darwin 1854. PLoS One. 2016;11: e0147082.
153. Foley B, Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, Rambaut A, Wolinsky S, and Korber B, Eds. HIV Sequence Compendium 2017. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM. 2017; doi:LA-UR 17-25240
154. Dietrich M, Kempf F, Boulinier T, McCoy KD. Tracing the colonization and diversification of the worldwide seabird ectoparasite Ixodes uriae. Mol Ecol. 2014;23: 3292–3305.
155. Porter J. Fast NeighborNet: Improving the Speed of the Neighbor-Net Phylogenetic Network Algorithm with Multithreading and a Relaxed Search Strategy [Internet]. bioRxiv. 2018. p. 283424. doi:10.1101/283424
156. Piry S, Alapetite A, Cornuet J-M, Paetkau D, Baudouin L, Estoup A. GENECLASS2: a software for genetic assignment and first-generation migrant detection. J Hered. 2004;95: 536–539.
157. Gao C, Sun H, Wang T, Tang M, Bohnen NI, Müller MLTM, et al. Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson’s Disease. Sci Rep. 2018;8: 7129.
158. Ashton JC. Phylogenetic methods in drug discovery. Curr Drug Discov Technol. 2013;10: 255–262.
159. Ernst M, Saslis-Lagoudakis CH, Grace OM, Nilsson N, Simonsen HT, Horn JW, et al. Evolutionary prediction of medicinal properties in the genus Euphorbia L. Sci Rep. 2016;6: 30531.
160. Guzman E, Molina J. The predictive utility of the plant phylogeny in identifying sources of cardiovascular drugs. Pharm Biol. 2018;56: 154–164.
161. Blum T, Briesemeister S, Kohlbacher O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinformatics. 2009;10: 274.