Flexible methods for estimating genetic distances from nucleotide data

Simon Joly[1,*], David Bryant[2], and Peter J. Lockhart[3]

[1] *Institut de recherche en biologie végétale, Montreal Botanical Garden, Montréal, Canada;*
[2] *Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand;*
[3] *Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand;*
[*] *To whom correspondance should be addressed: Institut de recherche en biologie végétale, 4101 Sherbrooke East, Montréal (QC) H1X 2B2, Canada; +1 514.872.0344; joly.simon@gmail.com.*

April 11, 2014

Abstract

- With the increasing use of massively parallel sequencing approaches in evolutionary biology, the need for fast and accurate methods suitable to investigate genetic structure and evolutionary history are more important than ever. We propose new distance measures for estimating genetic distances between individuals when allelic variation, gene dosage and recombination could compromise standard approaches.
- We present four distance measures based on single nucleotide polymorphisms (SNP) and evaluate them against previously published measures using coalescent-based simulations. Simulations were used to test (*i*) whether the measures give unbiased and accurate distance estimates, (*ii*) if they can accurately identify the genomic mixture of hybrid individuals and (*iii*) if they give precise (low variance) estimates.
- The results showed that the SNP-based GENPOFAD distance we propose appears to work well in the widest circumstances. It was the most accurate method for estimating genetic distances and is also relatively good at estimating the genomic mixture of hybrid individuals.
- Our simulations provide benchmarks to compare the performance of different distance measures in specific situations.

Key-words: Single nucleotide polymorphisms (SNPs), genetic distances, polyploidy, hybridization, population genomics, coalescent, simulations.

Introduction

The last few decades have witnessed a methodological revolution in the field of population genetics. Model-based likelihood approaches have been propelled to the forefront of species and population level studies (e.g. Beaumont and Rannala 2004; Beaumont et al. 2002; Huelsenbeck et al. 2001). These changes have been made possible by the remarkable advances in computing technology and the application of computationally intensive Monte Carlo methodology.

But even these sophisticated methods are facing critical challenges confronted by the overwhelming amount of data generated by massively parallel sequencing technologies. In many cases, state-of-the-art approaches in terms of models and methods cannot always accommodate population genomics data. Consequently, quick and rapid approaches that allow for investigations of patterns and processes still have their utility in this discipline.

Our objective is to present new, flexible, and robust distance measures for estimating genetic distances from single nucleotide polymorphisms (SNPs) data. We focus on the estimation of distances between individuals (or organisms), even though the distances could certainly be useful in many other circumstances. There are good reasons to focus at the level of individuals rather than populations or species. Individuals are central to biology. Measurements based on morphology, spatial positioning, or genetics are generally performed at the individual level. Individuals are also the fundamental units of natural selection, the central concept of evolutionary biology. And finally, estimates of genetic relatedness between individuals can reveal correlations between genetic and phenotypic distances, spatial genetic structure across a landscape, species boundaries, and could be used for genetic or phylogenetic diversity (PD) surveys.

Although obtaining genetic distances among individuals seems relatively straightforward, there can be several complicating factors. One is the presence of SNPs among gene copies in non-haploid individuals. Polyploidy, which is defined by the presence of more than two genome copies in a nucleus, leads to further complexities. Not only is there the potential presence of more than two character states for each nucleotide, there is also the potential for non-conventional segregation of chromosomes. Finally, recombination along chromosomes renders the problem of calculating distances between organisms even more complex. Given the importance of estimating genetic distances between individuals and the increasing availability of genome-wide sequence data, we think that this issue deserves further investigation.

Only a few approaches, generally motivated by very different research questions, have been proposed to handle SNPs, polyploidy or recombination. Although not based on sequence data, Bruvo et al. (2004) proposed an interesting approach to deal with ploidy level variation for estimating the distances between individuals from microsatellites data that could be generalized to sequence data. Their method consisted in comparing directly the alleles of one individual with that of another, while accounting for the "missing alleles" in comparisons between ploidy levels. Joly and Bruneau (2006) proposed the POFAD algorithm to estimate the genetic distance of individuals from allelic sequence information. Their idea for comparing homozygotes and heterozygotes could be seen as comparing alleles that share a most common recent ancestor. However,

their implementation could not be applied to polyploid organisms. Later, Göker and Grimm (2008) proposed different methods to estimate distances between "populations" using, among others, community ecology statistics such as Shannon's entropy or Euclidean distances. Although not originally designed for the problem we address here, they could nevertheless be relevant if one considers an individual as a "population" of sequences. Their approaches could be applied to individuals of mixed ploidy levels, but they did not deal with the potential presence of recombination.

Here, we propose four methods for estimating genetic distances between individuals from nucleotide sequence data. One of these is an adaptation of Nei's genetic distance (Nei et al. 1983) for this specific problem, but the three other methods are novel. All methods are very general in that they can be applied to individuals of any ploidy level, but also when individuals have different ploidy levels. We first describe in detail the challenges involved in estimating genetic distances between individuals. We then describe the new methods and compare them and others using simulations. We finish by making recommendations on the use of distance measures in different contexts.

Problems associated with the estimation of distances between individuals

Allelic variation

If the estimation of genetic distances between DNA sequences is straightforward, the potential presence of more than one allele at autosome loci in non-haploid individuals makes it more complex to estimate the genetic distances between individuals, especially when combining information from multiple loci (Joly and Bruneau 2006). Also, one important property of distances that measure overall difference between individuals is that the comparison of a heterozygous individual with itself should have a distance of 0, something that is not necessarily obtained with all existing approaches. For instance, taking the mean pairwise distance between all alleles will not generally give a distance of 0 when comparing an individual with itself.

Polyploidy

Polyploidy brings two other problematic issues: inheritance and gene dosage. Inheritance of diploids is always disomic while it can be either disomic or multisomic in polyploids (Comai 2005). Polyploids are disomic if chromosomes group by pairs at meiosis, one example being homeologous chromosomes in allopolyploids. However, they are multisomic when chromosomes form multivalents. In many cases, inheritance of polyploid taxa is unknown or difficult to determine precisely. Some polyploids are even characterized by a mixture of inheritance modes. For instance, a marker could have mainly disomic inheritance with occasional multisomic inheritance, or different chromosomes could have different modes of inheritance within a genome (Wendel 2000).

Gene dosage is another issue associated with polyploidy (Bruvo et al. 2004). In diploids, gene dosage is obvious: a homozygous individual has two copies of the same

allele and a heterozygous individual has one copy of each allele. In polyploids, it is rare that we know the exact dosage of each allele in the genome. A tetraploid that has the observed nucleotide state 'A' at a position (i.e., it is homozygote) can only have genotype 'AAAA'. However, a tetraploid individual with observed states 'A' and 'T' at a site could have the genotypes 'ATTT', 'AATT', or 'AAAT'. The unknown dosage of these character states makes it more difficult to estimate precisely the genetic distances between polyploids. The situation can become even more complicated when there are more than two character states at a sequence site, a feature that becomes more likely in higher polyploids. Finally, another important feature of the desired distance measure is the capacity to estimate distances between individuals of different ploidy levels (Bruvo et al. 2004).

Distance definitions

We propose four new distance measures to calculate the genetic distance between individuals from sequence data. The main novelty of these proposed measures is that they are all computed at the nucleotide level. Therefore, we define them first at the individual nucleotide site level, and explain later how these distances can be extended to strings of nucleotides, some potentially linked (within loci) and others unlinked. These measures assume that we know the nucleotides present at a given position in an individual but not necessarily gene dosage, which is typical for data obtained from genotyping or sequencing. All proposed distances are bounded between 0 and 1 and have the property that the distance between an individual and itself is 0.

matchstates

This measure looks at each nucleotide present at a given sequence site in one individual and checks if there is a nucleotide in the other individual that matches. More formally, consider a specific sequence site i that might be present in multiple alleles or gene copies in an individual. Let A_X^i be the complete set of nucleotides for individual X at site i and let $|A_X^i|$ be the number of nucleotide states observed for individual X at site i. The MATCHSTATES distance between individual X and individual Y at site i is

$$\text{MATCHSTATES}_{XY}^i := \frac{|A_X^i \Delta A_Y^i|}{|A_X^i| + |A_Y^i|},$$

where $A_X^i \Delta A_Y^i$ denotes the set of elements that belong to either A_X^i or A_Y^i, but not in both.

genpofad

The GENPOFAD measure is named after the POFAD algorithm described by Joly and Bruneau (2006). The GENPOFAD distance can be defined as one minus the ratio of the number of nucleotides shared between two individuals divided by the maximum number

of nucleotides observed in either of the individuals at a given sequence site. Following the notation introduced above,

$$\text{GENPOFAD}^i_{XY} := 1 - \frac{|A^i_X \cap A^i_Y|}{max(|A^i_X|, |A^i_Y|)}.$$

mrca

The MRCA distance measure gives a distance of 0 whenever two individuals share at least one nucleotide at a given site and a distance of 1 otherwise. Formally, the MRCA distance between individual X and individual Y is

$$\text{MRCA}^i_{XY} := \begin{cases} 0 \; if \; |A^i_X \cap A^i_Y| & \neq & \emptyset \\ 1 \; if \; |A^i_X \cap A^i_Y| & = & \emptyset \end{cases}.$$

nei

This distance is the application of Nei's genetic distance (Nei et al. 1983) at the nucleotide level. The frequency of each nucleotide is estimated per site for each individual and then NEI genetic distance between individual X and individual Y for site i is estimated as

$$\text{NEI}^i_{XY} := 1 - \sum_{j}^{A,C,T,G} \sqrt{p^i_{j\in X} p^i_{j\in Y}},$$

where $p^i_{j\in X}$ is the frequency of nucleotide j in individual X at site i. This formula is flexible as it can be easily applied among individuals from different ploidy levels. Gene dosage is assumed to be known, but it can also be used if it is unknown by giving equal weight to each nucleotide present.

Extension to multiple sites and genes

The extension of all distance measures to many sites within a locus is easily done by taking the average distance over all DNA positions such as

$$d_{XY} = \frac{1}{s} \sum_{i=1}^{s} d^i_{XY},$$

where s is the number of sites and d^i_{XY} is the contribution of site i to the distance. An estimate of standard error is then provided by the standard statistical formula

$$var(d_{XY}) = \frac{1}{s(s-1)} \sum_{i=1}^{s} (d^i_{XY} - d_{XY})^2.$$

In some cases, it might be important to divide nucleotides into different loci, such as when several unlinked genes are sampled throughout the genome, each containing several linked nucleotides. We suggest distances be calculated first across sites within a marker to obtain distance matrices for each marker. Once this is done, one can

compute a genome-wide distance matrix by taking the mean of all marker matrices. In calculating this genome-wide distance matrix, it is possible to scale each individual matrix by dividing the distances of a given matrix by the maximum distances in that matrix. This scaling gives the same weight to all markers whatever their variability, which could be interesting if the markers do not have the same evolution rates (e.g., exons, introns, non-coding regions, etc.). If the nucleotides cannot easily be divided into distinct loci, such as when we have a long contiguous sequence along a chromosome, the average distance over all DNA positions is appropriate because each site is then assumed to represent an independent assessment of the distance between the individuals.

Implementation

All these algorithms are implemented in POFAD version 1.06 (www.plantevolution.org/en/pofad.html). The matchstates algorithm is also implemented in SplitsTree4 (Huson and Bryant 2006).

Simulations

Computer simulations were performed to compare the performance of the distances in different situations. We evaluated three properties of the distance measures. First, we tested if the measures provide an unbiased and accurate estimate of distances between organisms. Second, we investigated how the different distances are able to detect the genomic mixture of hybrid individuals. Third, we evaluated how precise these different measures were. We evaluated our new distance metrics along with other previously published distances of Göker and Grimm (2008) that are relevant in the present context: the MIN distance and the Phylogenetic Bray-Curtis (PBC) distance (see Appendix for mathematical definition). The FRQ and the ENTROPY distance measures of Göker and Grimm (2008) were not investigated because they are not bounded between 0 and 1 and because they are more relevant in a context of host-parasite associations as originally described. Finally, we also evaluated the recent 2ISP method (Potts et al. 2014), even if the distance is not bounded between 0 and 1, as it is similar to our proposed methods (see appendix for mathematical definitions of previously published distances).

Accuracy of distance measures

To investigate whether the distance measures were accurate for estimating distances between individuals, we simulated tetraploid individuals ($2n = 4x$) along a species tree using the coalescent and estimated the genetic distances between individuals that have been evolving for different periods of time. Gene sequences of 1000 bp were simulated using MCcoal (Rannala and Yang 2003) on a species tree where the individuals compared had the following divergence times (τ): 0, 0.0005, 0.001, 0.002, 0.003, and 0.005. The divergence times (τ) represent the expected number of mutations per site from the node in the species tree to the present time. However, the expected divergence times of the sequences between individuals will be greater than the time of species divergence

as the time to coalescence of the sequences in the ancestral species needs to be considered (Nei 1987; Edwards and Beerli 2000; Arbogast et al. 2002). The expected time to coalescence in the ancestral species (population) is equal to $2N$ or $\theta/2$ (Edwards and Beerli 2000). The expected genetic distance is thus twice the coalescence time expectation, which is twice the time since the species divergence plus twice the expectation for the coalescent time in the ancestral population: $d = 2\tau + \theta$. Distance measures were thus compared to this expected sequence divergence, but also with the expected species divergence (2τ). Simulations were performed with two population sizes ($\theta = 0.001$ and $\theta = 0.01$) that were held constant throughout the tree. The larger population size increased the number of polymorphisms in individuals. All simulations were repeated 2000 times.

Estimation of the genomic mixture of hybrids

To investigate how good the different distance measures are at detecting the genomic mixture of hybrid individuals, we estimated and compared the genetic distance of an allopolyploid with its two parents. For this, we simulated an allopolyploid speciation event. Gene copies inherited from one parent in the allopolyploid were then transferred by descent in the allopolyploid species via multisomic inheritance (i.e., they can be assumed to form a panmictic population and simulated with the coalescent), and were evolving independently from the gene copies inherited from the other parent. This allowed us to simulate gene sequences using multi-labeled species trees (see Jones et al. 2013). The parental species were tetraploids whereas the allopolyploid species was either octopolyploid with four gene copies coming from each parent or hexaploid with four copies coming from one parent and two from the other. This allowed us to test two ratios of parental genome contribution in the hybrid.

Gene sequences of 1000 bp were simulated on a species tree as described above with a population size $\theta = 0.001$ and with a divergence time to the two parental species fixed at $\tau = 0.003$. Three different scenarios were investigated for the timing of the allopolyploid event: $\tau = 0$ (in which case it is an immediate descendent of the two parental species), $\tau = 0.001$ or $\tau = 0.002$. To investigate the hybrid mixture of the allopolyploid individual, we estimated an hybrid index that indicates the relative distance of the hybrid from its two parents:

$$I = \frac{d_{AX}}{d_{AX} + d_{BX}},$$

where A and B are the two parents and X the hybrid, and where d_{AX} is the genetic distance between species A and the hybrid. The hybrid index (I) is bounded between 0 and 1 and an index of 0.5 indicates that the hybrid is equally distant to both parents. Cases where both $d_{A,X}$ and $d_{B,X}$ were equal to zero were given $I = 0.5$. All simulations were repeated 2000 times.

Effect of the number of markers on precision

We also estimated the impact of gene number on precision in the two previous simulation settings. For the precision of the genetic distance estimate, we used the simulations

with $\theta = 0.001$ and the expected distance of 0.01. For the hybrid index, we used the framework of the octopolyploid speciation event at $\tau = 0.001$. In both cases, we evaluated the statistics (distance or hybrid index) with 1, 2, 5, 10, 20, and 40 markers. Distances were estimated 100 times for each scenario and standard deviation among estimates was computed and plotted to investigate the decrease in standard deviation with the number of markers for each method.

Results

Theoretical considerations

Before comparing the different distance methods, it is relevant to note the similarities between the SNP-based methods proposed here and the previously published methods based on whole marker sequences. For example, MRCA is the same as MIN applied to a single nucleotide. As such, it is interesting to compare the performance of this pair of methods in the simulations. Moreover, the GENPOFAD distance is equivalent to the POFAD algorithm of Joly and Bruneau (2006) when applied to a single nucleotide in diploid individuals. For a locus evolving under an infinite site mutation model without recombination, the GENPOFAD distance should give the same distance as POFAD when extended to the whole locus (see below). However, GENPOFAD has the advantage that it could be applied to individuals of any ploidy level.

Distance accuracy

Only GENPOFAD provided an accurate estimate of the sequence divergence ($2\tau + \theta$; Figs. 1, 2). The GENPOFAD estimates were very accurate with small population sizes ($\theta = 0.001$), but tend to provide a slightly underestimated distance for small divergence times with $\theta = 0.01$ (Figs. 1, 2). Moreover, it also underestimated sequence divergence within populations (i.e., when species divergence = 0), suggesting that it is not a very accurate estimator of θ. Nevertheless, it was the best estimator of θ among the methods tested.

Other distance measures had interesting properties. MIN underestimated sequence divergence (Fig. 1), but provided an accurate estimate of the species divergence (Fig. 2). MATCHSTATES and PBC provided similar estimates that fell between the expected sequences divergence and the species divergence. The other estimates either largely overestimated sequence divergence (2ISP, NEI) or underestimated species divergence (MRCA) in all situations (Figs. 1, 2).

Hybrid genetic mixture

Distances measure were evaluated for estimating the intermediacy of hybrid individuals relative to its parents. When the parents contributed an equal number of gene copies, all methods were accurate, but NEI provided the most precise estimate of the hybrid index (Fig. 3). GENPOFAD and 2ISP were the second best methods according to precision,

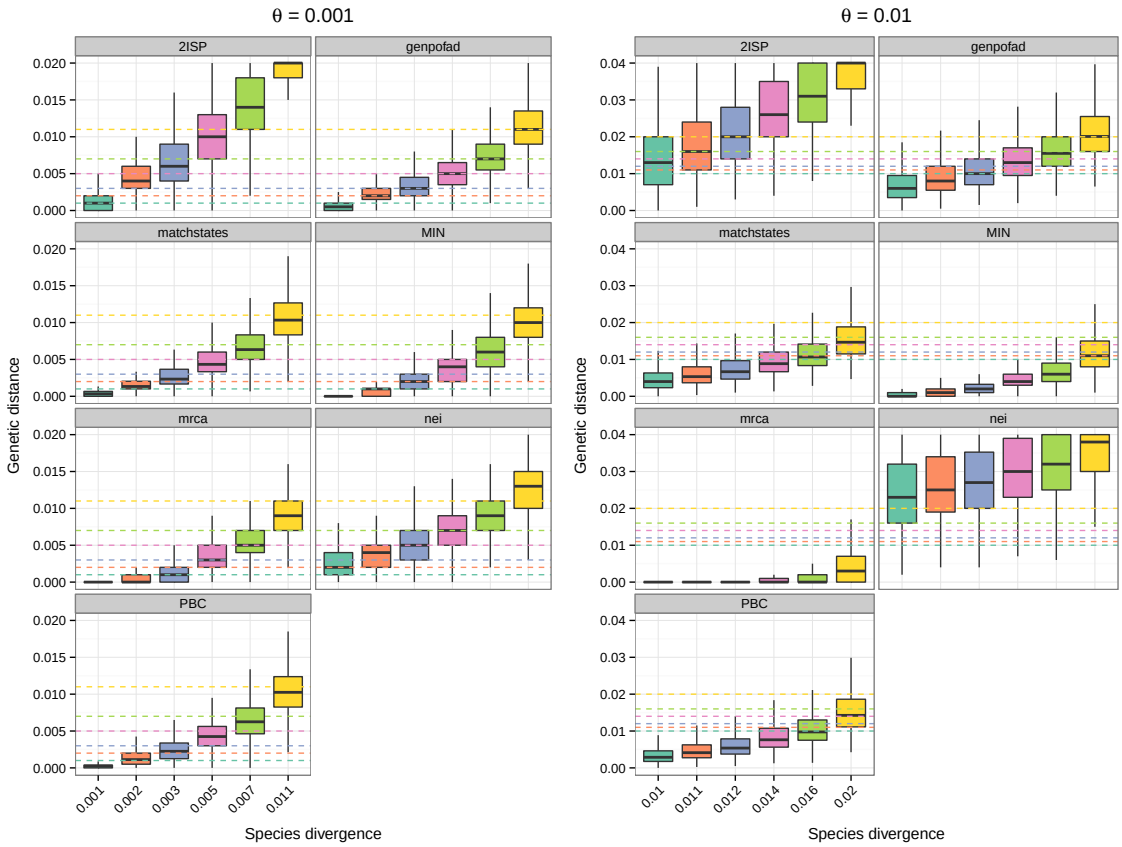

Figure 1: Boxplots showing the estimated divergence for several distance measures, compared to expected sequence divergence ($d = 2\tau + \theta$; dotted lines of the same colour as the boxes). Simulations were performed on a species tree with the coalescent using populations sizes of $\theta = 0.001$ (left panels) or $\theta = 0.01$ (right panels).

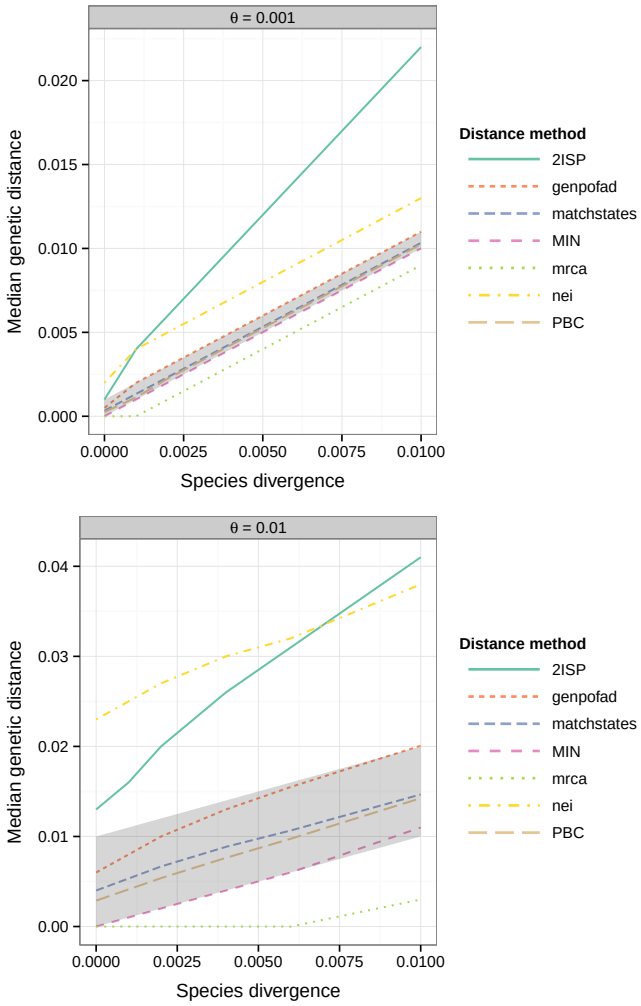

Figure 2: Plots showing the relationship between median estimated sequence divergence for the distance methods and the species divergence used in the simulations, for two populations sizes. The gray area indicates the time range between the expected species divergence ($d = 2\tau$; lower bound) and the expected sequence divergence ($d = 2\tau + \theta$; upper bound).

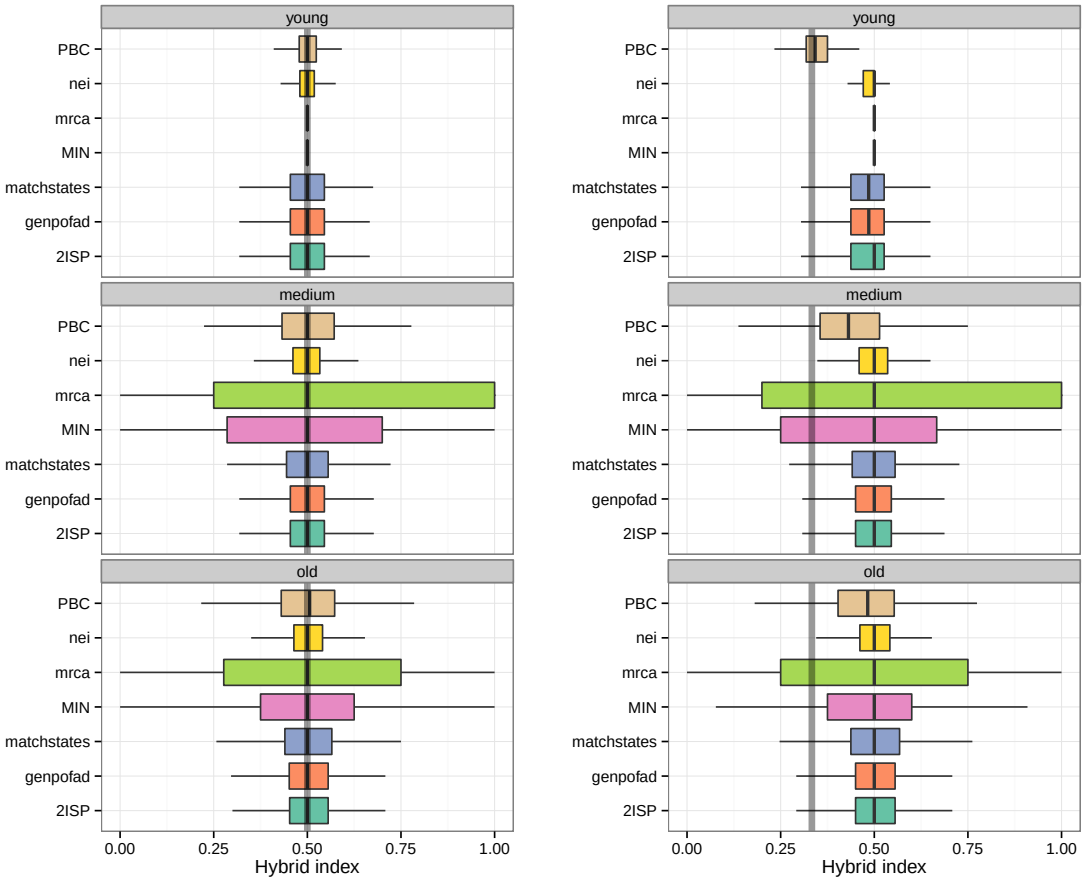

Figure 3: Boxplots showing a hybrid index (i.e., the relative contribution of each parental genome) for the different distance measures and for different time since the allopolyploid (hybridization) speciation event. The gray lines indicates the genomic mixtures that were simulated: one in which each parent contributed equally (1:1) to the allopolyploid (left panels) and another where one parent contributed twice the number of copy (2:1) than the other parent (right panels).

followed very closely by PBC and MATCHSTATES. MRCA and MIN provided imprecise estimates of hybrid index (Fig. 3).

No method provided an accurate hybrid index estimate when one parent contributed twice the number of gene copies as the other (Fig. 3), but some methods performed better than others. PBC was by far the best method, followed by GENPOFAD and MATCHSTATES. As before, MRCA and MIN provided the worst estimates of the hybrid index. Also, if some evidence for an unequal contribution was visible for the young hybrid for GENPOFAD and MATCHSTATES, evidence of unequal parental contribution for older hybrids was only observed with the PBC distance.

Effect of the number of markers on precision

Evaluation of the methods' precision showed different results for the distance accuracy and for the hybrid index simulations. For the estimation of the genetic distance, all

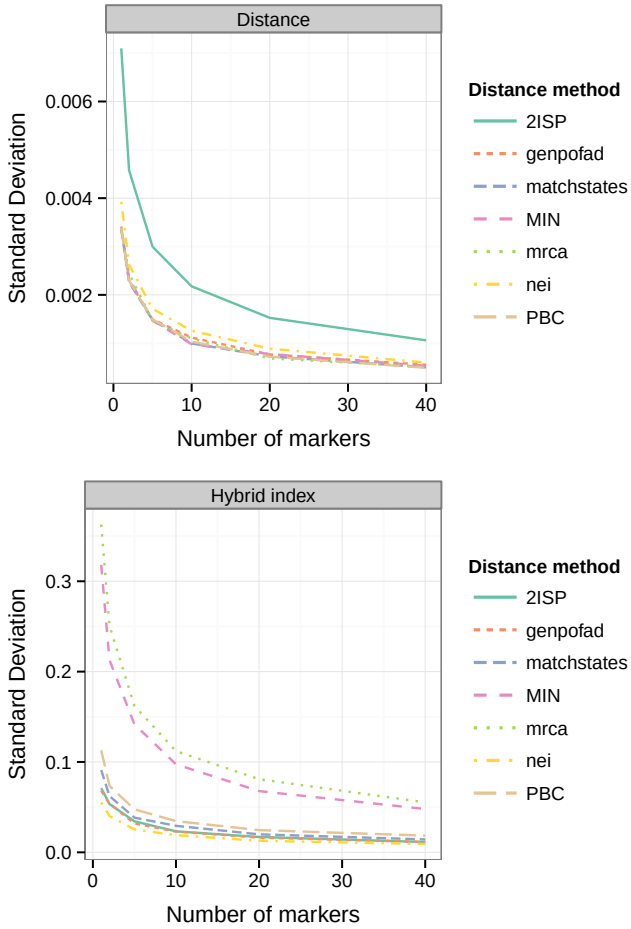

Figure 4: Plot showing the effect of the number of markers used on the precision of the genetic distances (upper panel) and the hybrid index estimates (lower panel). A small standard deviation indicates better precision. The simulation settings for the genetic distance were as for Fig. 1a with expected distance of 0.01 and for the hybrid index they were the same as those for Fig. 2a with a medium timing for the allopolyploid speciation event.

methods showed a similar precision and the increase in precision (decrease in standard deviation among replicates) was similar for the different methods, with the exception of 2ISP that had a much larger error than all others (Fig. 3a). The pattern was different for the precision of the hybrid index. The methods MRCA and MIN were much less precise than the others and they required more markers to converge on stable estimates (Fig. 3b). The remaining methods had a similar precision, although they could be ranked as followed for precision (from best to worst): NEI > GEPOFAD = 2ISP > MATCHSTATES > PBC (Fig. 3b).

Discussion

With the increasing use of massively parallel sequencing approaches in evolutionary biology, fast, accurate, and precise methods to investigate genetic structure and evolu-

tionary history are required. Concatenation approaches are known to be inconsistent in some circumstances (Degnan and Rosenberg 2006; Salter Kubatko and Degnan 2007) and fully Bayesian approaches to population/species reconstruction (e.g. Heled and Drummond 2010; Liu et al. 2009) are computationally demanding with large number of markers. If faster coalescent alternatives exist for genomic studies (Bryant et al. 2012), distance measures nevertheless remain an interesting strategy, especially given the consistent properties of some indices (Liu et al. 2009; Mossel and Roch 2010).

Until now, the toolset of distance measures was limited for studying the relationships of individuals. Overcoming this shortcoming is critical given that individuals are the fundamental unit for many studies at the species level. The main problems encountered at this level are those of allelic variation and polyploidy. However, the potential presence of recombination in the nuclear genome and the SNP based nature of many contemporaneous studies represent further challenges. We thus present here new distance measures that all have the property that they are estimated at the nucleotide level in order to alleviate these biological complexities.

Advantages of SNP-based distances

Interestingly, SNP-based distances do not suffer from the comparison with whole-sequence distances in our simulations. This is relevant because the simulation of long (1000 bp) sequences without recombination should advantage distances estimated on whole sequences. To the contrary, the most accurate method for estimating genetic distances was a SNP-based method. Clearly, one can expect SNP-based methods to rapidly gain an advantage over whole sequence methods in the presence of recombination. In many empirical studies that use large numbers of markers, it is indeed very difficult to rule out completely the presence of recombination, especially if markers are long. If recombination should not affect the performance of SNP-based methods, it will affect those based on whole sequences. SNP-based methods are thus expected to be particularly useful given the increasing abundance of genome-scale studies based on whole genomes or reduced-representation sequencing data.

Another important factor to consider is the length of markers. Massively parallel sequencing technologies generally result in markers of small sequence lengths. With such data, we expect that the relative advantage of distance measures based on the whole marker sequence to decrease with decreasing sequence length. Indeed, we can have an idea of that effect when going from 1000 bp sequences to SNP data by comparing the distances MIN and MRCA as MRCA is identical to MIN applied to a single SNP. Consequently, SNP-based methods are particularly well suited for SNP-based studies or for studies using short length markers.

Importance of gene dosage information

Of the methods evaluated here, two can actually take into account exact gene dosage information if known: PBC and NEI. One would expect this type of information to be particularly important for estimating unequal genomic mixtures in hybrid individuals. This actually seems to be the case for PBC that was the best method according to this

criteria. However, NEI did not appear to benefit from gene dosage information in the same situation. Our results tend to show, however, that gene dosage information is not critical for good performance in all situations. This is especially true for the estimation of genetic distances where the best method did not use gene dosage information. This is a very encouraging result given that such information is rarely known precisely in genomic studies involving polyploids.

Method performances

In term of genetic distance accuracy, the best method was GENPOFAD, a SNP-based method. It provided very accurate estimates of sequence divergence at small population sizes ($\theta = 0.001$), even if the estimates were slightly biased at larger population sizes ($\theta = 0.01$). It was also found to provide a slightly underestimation of θ in populations, even though it was still better than all other methods in this aspect.

The minimum allelic distance between individuals (MIN) provided an accurate estimate of the species divergence time, which is an interesting property. This observation concurs with previous studies that have shown this measure to be a consistent estimator of species distances in certain situations (Mossel and Roch 2010; DeGiorgio and Degnan 2014). However, the simulations showed that this measure performs poorly when it comes to estimating the genomic mixture of individuals, both in terms of accuracy and precision. Interestingly, two distance measures provided estimates that fell between the expected sequence divergence and the species divergence, that is between $2\tau + \theta$ and 2τ. These are the MATCHSTATES and the PBC methods.

Regarding hybrid mixture estimates, the best method was clearly PBC that was the only method to be close to accurate when estimating unequal contribution of the parents in the young age hybrid. Moreover, evidence for unequal contribution remained even for older hybrids, whereas that signal was lost for all other methods. Note that this assumes that we now the exact number of copies in the hybrid (i.e., gene dosage), an information that might not be always available in empirical datasets and that could affect the performance of the PBC distance. Among other methods, GENPOFAD and MATCHSTATES were slightly better as they showed slight evidence for the unequal parental contributions for the young hybrid and they provided precise estimates. The methods MIN and MRCA were not precise and did not detect unequal parental contributions. This is not surprising as these methods essentially ignore polymorphisms by considering only the most similar nucleotides (MRCA) or alleles (MIN).

Perhaps the best recommendation we can provide is to use the GENPOFAD distance in general as this is the most accurate method in terms of expected genetic distance and given that it is relatively good at estimating genomic mixture between individuals. Moreover, its performance will not be affected by the presence of recombination or if only short markers are available. In cases where species divergence times are of interest and in absence of recombination, then the MIN distance is of great interest. Finally, if gene dosage is known and genomic admixture is of main interest, then the PBC distance is the best choice if recombination is absent. In any case, we hope that this study and the simulation framework we propose for comparing the performance of distance measures will stimulate the development and testing of further SNP-based distance

measures.

Appendix

Definition of previously published distance measures

In the following definitions based on whole markers sequences, A_X represents the complete set of alleles for individual X and $|A_X|$ is the number of alleles observed for individual X. Also, let d_{ij} be the genetic distance between alleles i and j.

MIN distance

The MIN distance was proposed by Göker and Grimm (2008) in the present context, but it had been often used in other contexts as well (e.g. Joly et al. 2009; Liu et al. 2009; Mossel and Roch 2010). It can be described as:

$$\text{MIN}_{XY} := min(d_{ij}|i \in A_X, j \in A_Y).$$

Phylogenetic Bray-Curtis distance (PBC)

The PBC distance was defined by Göker and Grimm (2008) as:

$$\text{PBC}_{XY} := \frac{\sum_{i \in A_X} min(d_{ij}|j \in A_Y) + \sum_{j \in A_Y} min(d_{ij}|i \in A_X)}{|A_X| + |A_Y|}.$$

2ISP distance

The 2ISP distance is a nucleotide-based distance (Potts et al. 2014). It estimates the distance between nucleotides using the step-matrix presented in Figure 1 of Potts et al. (2014).

References

B. S. Arbogast, S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slowinski. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics*, 33(1):707–740, 2002. doi: 10.1146/annurev.ecolsys.33.010802.150500.

M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature reviews*, 5:251–261, 2004.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

R. Bruvo, N. K. Michiels, T. G. D'souza, and H. Schulenburg. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology*, 13:2101–2106, 2004.

D. Bryant, R. Bouckaert, J. Felsenstein, N. Rosenberg, and A. RoyChoudhury. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 19(8):1917–1932, 2012.

L. Comai. The advantages and disadvantages of being polyploid. *Nat Rev Genet*, 6 (11):836–846, 2005. doi: 10.1038/nrg1711.

M. DeGiorgio and J. H. Degnan. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst Biol*, 63(1):66–82, 2014. doi: 10.1093/sysbio/syt059.

J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68, 2006.

S. V. Edwards and P. Beerli. Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, 54(6):1839–1854, 2000.

M. Göker and G. W. Grimm. General functions to transform associate data to host data, and their use in phylogenetic inference from sequences with intra-individual variability. *BMC Evolutionary Biology*, 8:86, 2008.

J. Heled and A. J. Drummond. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*, 27(3):570–580, Mar. 2010. doi: 10.1093/molbev/msp274.

J. P. Huelsenbeck, F. Ronquist, R. Neilsen, and J. P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294:2310–2314, 2001.

D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

S. Joly and A. Bruneau. Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from rosa in north america. *Systematic Biology*, 55(4):623–636, 2006.

S. Joly, P. A. McLenachan, and P. J. Lockhart. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am. Nat.*, 174(2):e54–e70, 2009.

G. Jones, S. Sagitov, and B. Oxelman. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst Biol*, 62(3):467–478, May 2013. doi: 10.1093/sysbio/syt012.

L. Liu, L. Yu, D. K. Pearl, and S. V. Edwards. Estimating species phylogenies using coalescence times among sequences. *Syst Biol*, 58(5):468–477, Oct. 2009. doi: 10.1093/sysbio/syp031.

E. Mossel and S. Roch. Incomplete lineage sorting: consistent phylogeny sstimation from multiple loci. *(IEEE/ACM) Trans. Comput. Biol. Bioinformatics*, 7(1):166–171, Jan. 2010. doi: 10.1109/TCBB.2008.66.

M. Nei. *Molecular Evolutionary Genetics*. Columbia University Press, 1987. ISBN 9780231063210.

M. Nei, F. Tajima, and Y. Tateno. Accuracy of estimated phylogenetic trees from molecular data. *J Mol Evol*, 19(2):153–170, Mar. 1983. doi: 10.1007/BF02300753.

A. J. Potts, T. A. Hedderson, and G. W. Grimm. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Syst Biol*, 63(1):467–478, 2014. doi: 10.1093/sysbio/syt052.

B. Rannala and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656, 2003.

L. Salter Kubatko and J. H. Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24, 2007.

J. F. Wendel. Genome evolution in polyploids. *Plant molecular Biology*, 42:225–249, 2000.