

Appendix for: Flexible methods for estimating genetic distances from single nucleotide polymorphisms

Simon Joly^{1,*}, David Bryant², and Peter J. Lockhart³

¹*Institut de recherche en biologie végétale, Montreal Botanical Garden, Montreal, 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.*

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 & 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 & 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 & Grimm (2008) as:

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

PP distance

The PP 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

- Göker, M. & Grimm, G.W. (2008) 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.
- Joly, S., McLenachan, P.A. & Lockhart, P.J. (2009) A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am Nat*, **174**, e54–e70.
- Liu, L., Yu, L., Pearl, D.K. & Edwards, S.V. (2009) Estimating species phylogenies using coalescence times among sequences. *Syst Biol*, **58**, 468–477.
- Mossel, E. & Roch, S. (2010) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *(IEEE/ACM) Trans Comput Biol Bioinformatics*, **7**, 166–171.
- Potts, A.J., Hedderson, T.A. & Grimm, G.W. (2014) Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Syst Biol*, **63**, 467–478.