Tools for analysis of clonal populations in R

Zhian N. Kamvar[1] Jonah C. Brooks[2] Niklaus J. Grünwald[1,3]

1: Department of Botany and Plant Pathology, Oregon State University
2: College of Electrical Engineering and Computer Science, Oregon State University
3:Horticultural Crops Research Laboratory, USDA Agricultural Research Service, Corvallis, OR, USA

More about poppr

github.com/grunwaldlab/poppr#readme

Code for this poster

github.com/grunwaldlab/poppr-poster-aps-2016

Motivation

Knowledge of the population dynamics and evolution of plant pathogens allows inferences on evolutionary processes involved in their adaptation to hosts, pesticides, and other environmental pressures. With the advent of high-throughput sequencing technologies, obtaining genome-wide data has become easier and cheaper than ever before with techniques such as genotyping-by-sequencing (GBS).

We present here an overview of the R package poppr as it pertains to traditional and high-throughput population genetic data with select applications (Kamvar et al., 2015).

poppR — Population Genetics in R

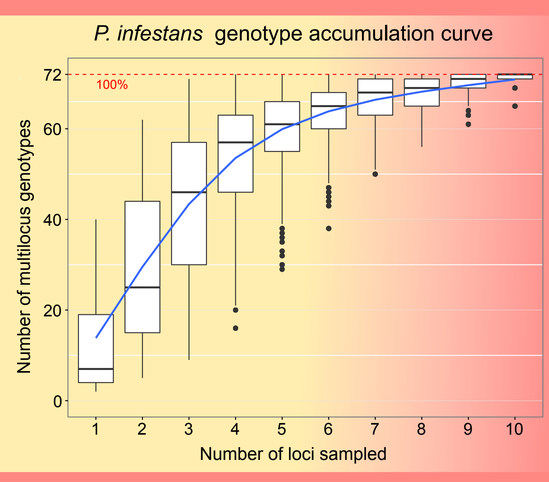
Genotype Accumulation Curve

One recent addition to poppr is the genotype accumulation curve (GAC). When starting analysis with a handful of markers, it's important to check the number of markers you have is sufficient to detect the diversity of mulitlocus genotypes (MLGs). This is visualised by the GAC (S1), which randomly samples n loci from your data (x axis), reporting the number of MLGs observed (y axis).

SSR Data

(Phytophthora infestans: 86 genotypes, 11 SSRs)

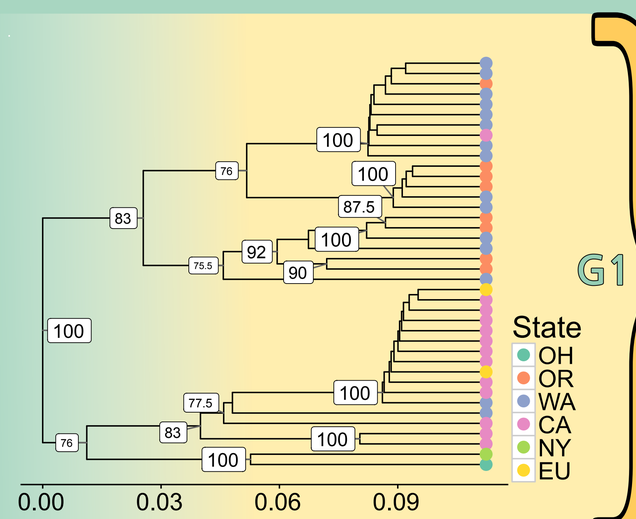
S1



GBS (SNP) Data

(Phytophthora rubi: 40 genotypes, 43,414 SNPs)

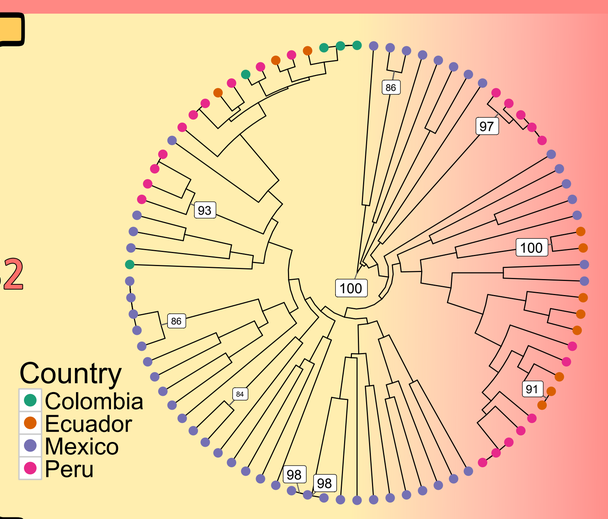
G1



Flexible bootstrap analysis

Bootstrapping dendrograms is a way to assess the internal consistency of your data. Poppr provides the function aboot(), which stands for "any boot". This function will create and bootstrap dendrograms for any data type including SNPs (G1) and SSR (S2), with any genetic distance, from individuals to populations.

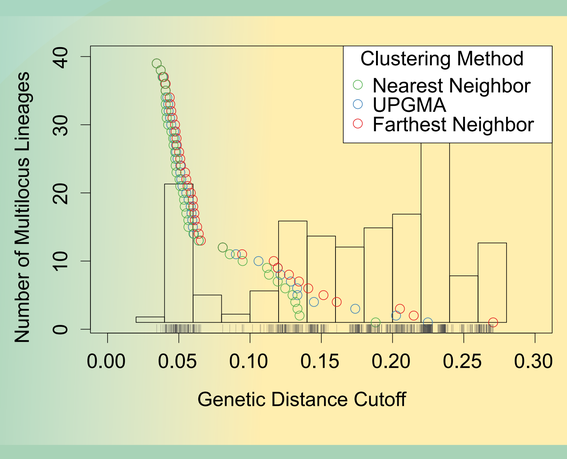
S2



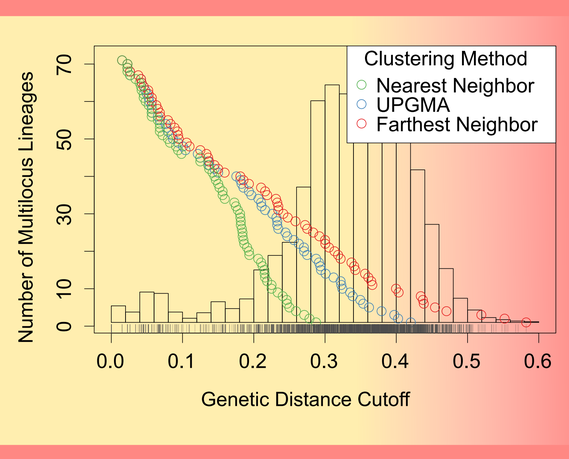
Define Multilocus Lineages by Genetic Distance

Multilocus genotypes (MLGs) are initially defined as the unique combination of alleles across all loci. New MLGs are created when this combination differs, which could be caused by genotyping error. New in poppr v2 is the ability to use genetic distance to identify and collapse similar MLGs into multilocus lineages (MLLs) for any data type using three different clustering methods, Nearest Neighbor, Average Neighbor (UPGMA), and Farthest Neighbor (G2, S3).

G2



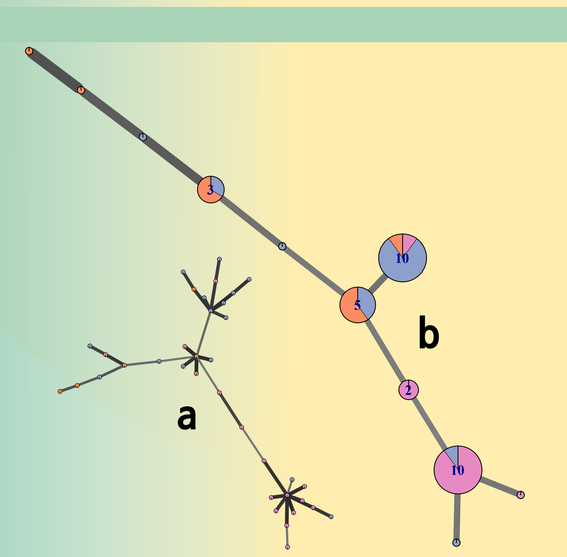
S3



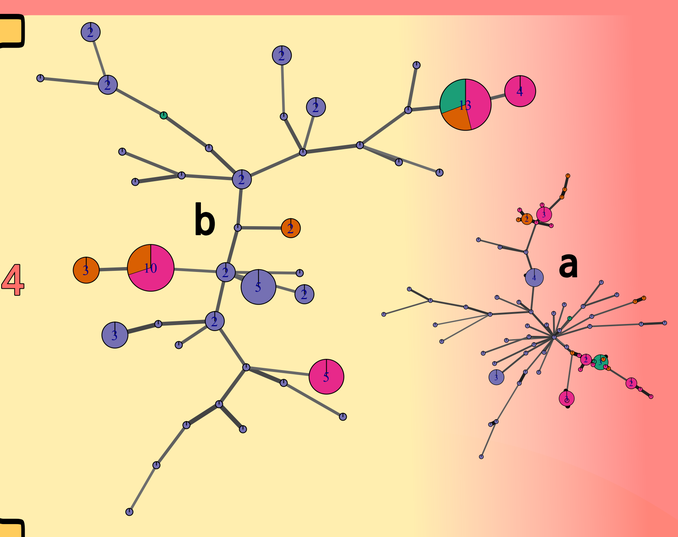
Minimum Spanning Networks with filtered MLGs

Minimum spanning networks are a useful tool for asessing the population structure of clonal organisms by showing the shortest connections between multilocus genotypes (G3-a, S4-a). New to poppr v2 is the ability to create minimum spanning networks showing connections between MLLs (G3-a, S4-a).

G3



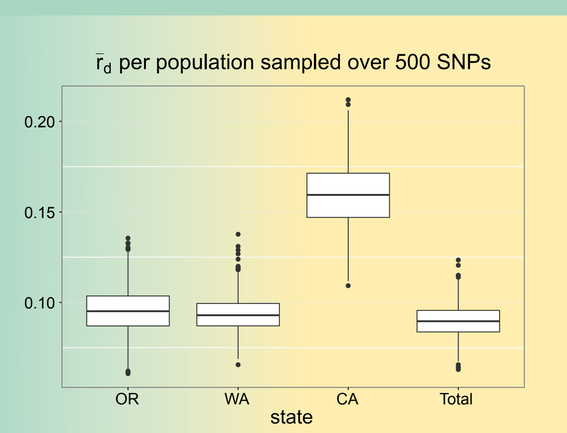
S4



Index of Association for SNPs

The index of association (and it's standardized form) is a measure of mutlilocus linkage disequilibrium, which can serve as a metric for clonality. It was initially implemented for traditional markers such as AFLP or SSR, but with high-throughput sequencing technologies like GBS, assessing linkage across thousands of loci became computationally intensive. We have created a method for quickly asessing the clonality of populations by sampling SNPs within a sliding window or randomly. Fig. G4 was generated by randomly sampling 500 SNPs 1000 times over each population showing that CA samples have a strong signal of clonal reproduction.

G4



Acknowledgements

The GBS example data are kindly provided by Javier F. Tabima and Niklaus J. Grünwald. These data were first filtered with the package **vcfR** before running the analysis (Knaus & Grünwald, 2016). You can find the scripts/tutorials by Brian Knaus at **https://github.com/knausb/vcfR_class**. The SSR data was provided from Erica M. Goss as part of Goss et al. (2014).

Kamvar ZN, Brooks JC and Grünwald NJ (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. Front. Genet. 6:208. doi: 10.3389/fgene.2015.00208

Goss, Erica M., et al. "The Irish potato famine pathogen *Phytophthora infestans* originated in central Mexico rather than the Andes." Proceedings of the National Academy of Sciences 111.24 (2014): 8791-8796. doi: 10.1073/pnas.1401884111

Knaus, Brian J., and Niklaus J. Grunwald. In press. vcfR: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources. doi: 10.1111/1755-0998.12549.