Brief Communication

Title: `RADpainter` and `fineRADstructure`: population inference from RADseq data

Authors: Milan Malinsky[1,2]*, Emiliano Trucchi[3], Daniel John Lawson[4], Daniel Falush[5]*

Affiliations:

[1]Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK.

[2]Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, CB2 1QN, UK.

[3]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria.

[4]School of Social and Community Medicine, University of Bristol, Bristol, BS8 2BN, UK.

[5]Institute of Life Science , Swansea University, Swansea, SA2 8PP, UK.

*Corresponding authors: Email: millanek@gmail.com (MM), danielfalush@googlemail.com (DF)

Abstract:

Powerful approaches to inferring recent or current population structure based on nearest neighbor haplotype 'coancestry' have so far been inaccessible to users without high quality genome-wide haplotypes. With a boom in non-model organism genomics, there is a pressing need to bring these approaches to communities without access to such data. Here we present `RADpainter`, a new program designed to infer the coancestry matrix from restriction-site-associated DNA sequencing (RADseq) data. We combine this program together with a previously published MCMC clustering algorithm into `fineRADstructure` - a complete, easy to use, and fast population inference package for RADseq data (`https://github.com/millanek/fineRADstructure`).

Main Text:

Understanding of shared ancestry in genetic datasets is often key to their interpretation. The `fineSTRUCTURE` package (Lawson et al. 2012) represents a powerful model-based approach to investigating population structure using genetic data. It offers especially high resolution in inference of recent shared ancestry, as shown for example in the investigations of worldwide human population history (Hellenthal et al. 2014) and of genetic structure of the British population (Leslie et al. 2015). Further advantages when compared with other model-based methods, such as STRUCTURE (Pritchard et al. 2000) and ADMIXTURE (Alexander et al. 2009), include the ability to deal with a very large number of populations, explore relationships between them, and to quantify ancestry sources in each population.

The high resolution of `fineSTRUCTURE` and related methods derives from utilizing haplotype linkage information and from focusing on the most recent coalescence (common ancestry) among the sampled individuals - deriving a 'coancestry matrix', *i.e.* a summary of nearest neighbor haplotype relationships in the dataset. However, the existing pipeline for coancestry matrix inference was designed for large scale human genetic SNP datasets, where chromosomal location of the markers are known, haplotypes are typically assumed to be correctly phased (although it is possible to perform the analysis without this assumption), and missing data also needs to have been imputed. Therefore, these methods have so far been generally inaccessible for investigations beyond model organisms.

Therefore, we have developed `RADpainter`. With no requirements for prior genomic information (*e.g.* no need for a reference genome) and relatively low cost, RADseq data are fuelling a boom in ecological and evolutionary genomics, especially for non-model organisms (Andrews et al. 2016), where questions on population structure and relative ancestry are among the most commonly asked. `RADpainter` is designed to infer the coancestry matrix from RADseq data,

taking full advantage of the sequence of all the SNPs from each RAD locus to find (one or more) closest relatives for each allele. Then information about the nearest neighbours of each individual is summed up into the coancestry similarity matrix. We package `RADpainter` together with the `fineSTRUCTURE` MCMC clustering algorithm into an easy to use population inference package for RADseq data called `fineRADstructure`.

Briefly, the coancestry matrix is calculated as follows: for each RAD locus and each individual (a recipient), we calculate the number of sequence differences (i.e. SNPs) between that individual's allele(s) and the alleles in all other individuals (potential donors). The closest relatives (donors) for each allele are the alleles with the least number of SNPs; an equal proportion of coancestry is then assigned to each 'donor'. The coancestry estimation procedure is outlined in Algorithm 1.

Algorithm 1: Coancestry estimation: a simple case with haploid individuals and without missing data

```
Data: For each locus, the sequences of nucleotides that vary between individuals
Result: Estimated coancestry matrix
Init: Initialize the coancestry matrix C, so that C_ij=0 for all i and j
foreach RAD locus do
    foreach individual i // a recipient
    do
        foreach individual j ≠ i // potential donors
        do
            D_ij = number of SNPs between i and j;
        end
        M_i = min_{j≠i}(D_ij); // The minimum of D_ij for all j≠i
        N_Mi = count_{j≠i}(D_ij==M_i); // Count the number of 'donors'
        foreach individual j ≠ i do
            if (D_ij == M_i) // If individuals i and j are closest, j is a 'donor'
            then
                // Assign an equal proportion of coancestry to each donor
                C_ij = C_ij + 1/N_Mi
            end
        end
    end
end
```

Differences in ploidy are handled by averaging coancestry across alleles in the same individual. Practically, individuals are by default assumed to be haploid or diploid (depending on input format). However, we have also implemented an option to handle tetraploid species, a feature likely to be of particular use to the plant research community.

Where the donor and/or the recipient alleles are missing, their coancestry is assumed to be proportional to the amount of coancestry observed between them in the rest of the data (*i.e.* 'missing data' coancestry is shared proportionally to 'observed data' coancestry). Unknown nucleotides (Ns)

can be present within alleles and their positions are ignored in all pairwise sequence comparisons. `RADpainter` outputs missingness (the proportion of missing alleles) per individual. We recommend that outliers should be removed and users should avoid cases with large systematic differences in missingness between putative populations.

The summation across loci assumes perfect linkage within each RAD sequence and frequent recombination between different RAD sequences - each RAD locus is counted as if providing independent evidence with regards to coancestry. However, we at least partially account for any linkage between RAD loci (*e.g.* pairs of loci on the two sides of each restriction site) by adjusting the scaling constant 'c' that is passed on to the `fineSTRUCTURE` clustering algorithm together with the co-ancestry matrix (Lawson et al. 2012).

We applied `fineRADstructure` to a RAD dataset including 120 individuals from 12 populations of the alpine *Heliosperma pusillum* species complex sampled at 1,097 unmapped loci (Trucchi et al. 2016). The complicated network of relationships among these twelve populations belonging to two phylogenetically intertwined species (*H. pusillum*: P, *H. veselskyi*: V) with contrasting ecology and a post-glacial history of divergence in some of the six sampled localities (A to F; Figure 1) makes it an excellent case to study the performance of our approach.

The `fineRADstructure` results (a clustered coancestry matrix; Figure 1) make the presence of twelve populations immediately clear, with substructure suggested in some of the populations. The relationships between some of the populations (A, B, C and D) are clearly not tree-like with strong evidence of heterogeneous gene flow between the species (Trucchi et al. 2016). A variable level of intra-population co-ancestry, likely related to different degree of isolation or effective population size, is also visible across the populations. A major improvements of our approach over previous results is the clear identification, at the same time, of both a global structure produced by historical process and a local structure related to ecological divergence.

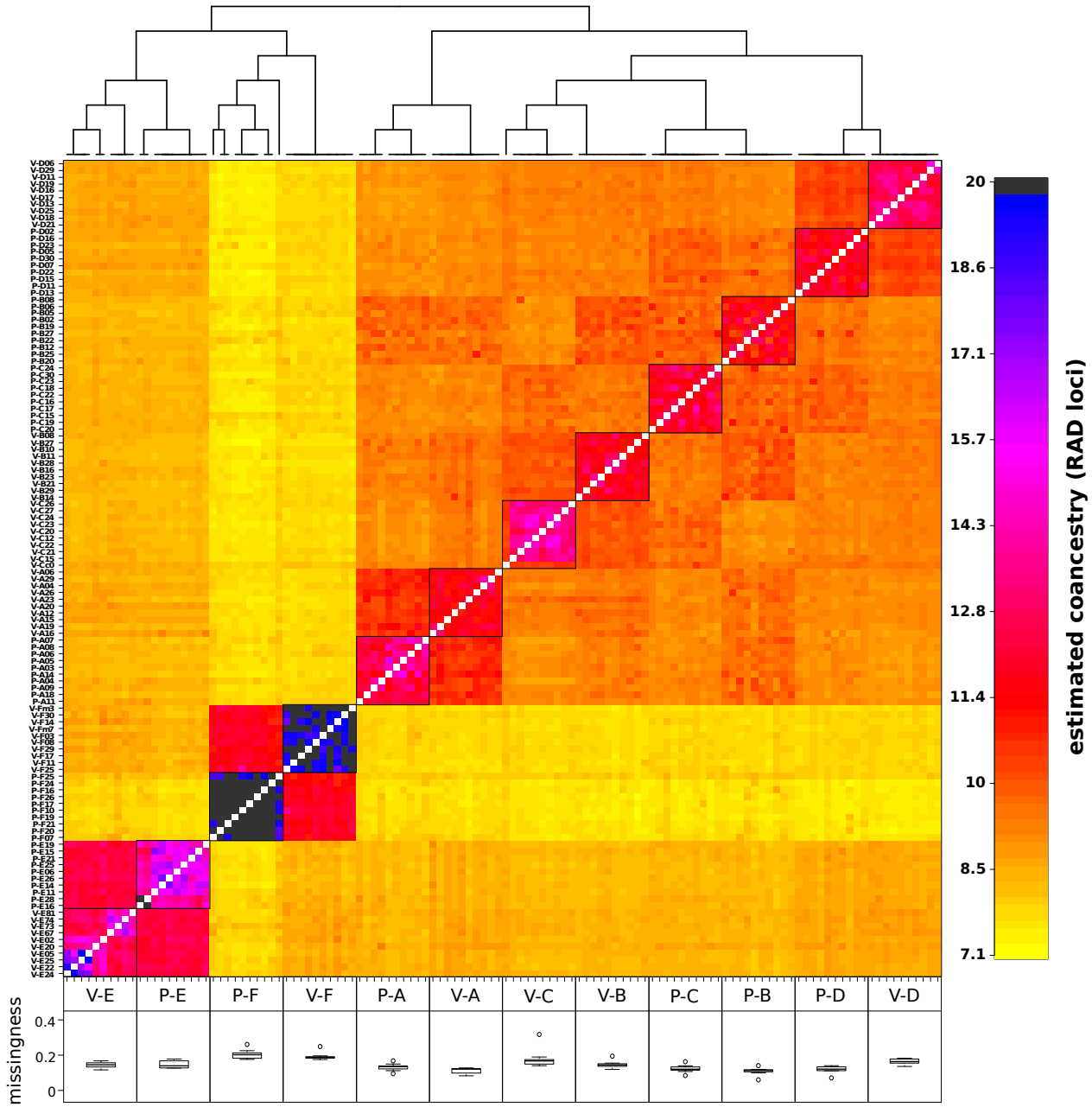

Figure 1: Clustered fineRADstructure coancestry matrix. Individuals within the twelve populations corresponding to the two sibling species (*H. pusillum*: P, *H. veselskyi*: V) and six localities (A to F) share more coancestry with each other than between populations. Hierarchical structure among and between localities is clearly inferred - populations at localities B and C cluster by species, whereas populations at localities A, D, E, and F cluster by locality. Missingness is shown below each population label, confirming that it does not contain strong outliers and does not vary systematically between inferred populations.

Acknowledgements:

This work was supported by the Medical Research Council (MR/M501608/1 to D.F), the Austrian Climate Research Programme (ACRP5-EpiChange-KR12AC5K01286, directed by Peter Schönswetter, to E.T.) and the Wellcome Trust (097677/Z/11/Z to M.M.)

References:

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genes Dev. 19:1655–1664.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17:81–92.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343:747–751.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control Consortium 2, et al. 2015. The fine-scale genetic structure of the British population. Nature 519:309–314.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Trucchi E, Frajman B, Haverkamp T, Schönswetter P., Paun O. 2016. Genomic and Metagenomic Analyses Reveal Parallel Ecological Divergence in *Heliosperma pusillum* (Caryophyllaceae). BioRxiv, doi: http://dx.doi.org/10.1101/044354.