File S2: Recommendations for the use of SDpop

SDpop is designed to be applicable to a wide range of organisms and using different types of sequencing data (e.g. RNAseq, DNA-reseq, RADseq) as input. Cleaning, mapping, genotyping and filtering of the data can thus be done in different ways to obtain input files (in the standard vcf format). The results, of course, will depend on the quality of the data and the bio-informatic pipelines used, and it's not possible to tailor a standard pipeline. Consider, for example, the case in which no genome of the species studied is available: one could have the choice either to map on a well-assembled genome of a closely related species, or to produce one's own reference genome, but this choice cannot be prescribed without any prior knowledge about the species.

The main requirement for SDpop is that the individuals, which should be sexed, are sampled from a panmictic population. Sampling from one local population seems the most appropriate strategy to ensure that sufficient gene flow has occurred between the individuals, and many population genetic tools are available to verify this (e.g., principal component analysis of genetic variation, clustering; Pritchard *et al.* 2000).

Some output parameters of SDpop can be used to assess suitability of the data and the quality of the upstream data treatment. E.g., a high percentage of SNPs inferred as haploid indicates that rare alleles are more often found to be homozygous then expected, which could happen if the samples come from a highly spatially structured population, or if alternative alleles are often missed, either through mapping biases or through monoallelic expression in RNAseq data. Or, there could be a high percentage of paralogous SNPs, which could indicate polyploidy, a genome duplication with respect to the reference genome, and possibly other problems. Thus, users of SDpop should have a close look at the general mapping statistics, the genome-wide estimates of heterozygosity in all samples, and possible population structure (e.g. by performing a principal component analysis of genetic variation in the samples).

We recommend that for species with known sex chromosomes, the genome or transcriptome of the homogametic sex is used as a reference for mapping, in order to correctly identify gametologous genes as such. However, SDpop can also be used to detect the sex chromosome system in species for which such knowledge is lacking. An obvious solution would be use two different references, one for each sex. In other cases, a high-quality assembly might be available for one sex only, which happens to be the heterogametic sex, and the assembly might contain both gametologous copies of some genes. These could be identified and removed from the reference; an example of such procedure is given in Käfer *et al.* (2020) where additional coverage data from DNA-resequencing was used to remove either Z- or W-specific parts of the reference genome.

As gametology, which yields the most powerful signal, is close to paralogy, we recommend that no prior filtering against paralogous sequences is performed. For this reason, we included a paralogous segregation type in SDpop, so the method is able to distinguish it from gametology. Of course, paralogy could also occur for gametologous genes (e.g. a gene duplication present on the sex chromosomes and an autosome), and such cases cannot be distinguished from standard paralogy in the current model. Thus, if many genes or SNPs are inferred as paralogous, the power to detect sex-linkage is reduced, and in that case, it could be worth to finetune the mapping algorithm.

We further recommend that the contig- or gene-wise posterior probabilities are used to identify sex-linked regions. Aggregating the site-wise likelihoods by calculating the geometric mean is a much more robust procedure than focusing on single sites, as the geometric mean gives more weight to sites that are informative (i.e. having a large difference in the likelihood for each of the segregation types). This would help researchers separating noise from signal. For transcriptome or exome data, where the unit of study is a gene, contig or exon, such contig-wise posterior probabilities are naturally calculated. When the data have larger scaffolds or even pseudo-molecules (chromosomes) as units, these could be cut into smaller windows for the calculation of contig-wise posterior probabilities. When the genotyped units only have one or a few SNPs (e.g. RADseq, GBS), we recommend using more individuals and higher thresholds for the inference of sex-linkage.

Finally, we stress that SDpop's inference is based only on SNP segregation patterns. Thus, other statistics can be used for the assessment of the inferences. For gametologous genes, heterozygosity in the heterogametic sex should be higher than in the homogametic sex, and higher than the genome-wide heterozygosity. With DNA sequencing data, hemizygosity should result in lower coverage in the heterogametic sex. This allows researchers to perform cross-validation of the results using independent data.

Literature Cited

Käfer, J., A. Bewick, A. Andres-Robin, G. Lapetoule, A. Harkess, *et al.*, 2020 A derived ZW chromosome system in *Amborella trichopoda*, the sister species to all other extant flowering plants. bioRxiv .

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