*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

*Genetics and population analysis*

# Dsuite - fast D-statistics and related admixture evidence from VCF files

Milan Malinsky*

Zoological Institute, University of Basel, Basel, Switzerland.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Summary:** The D-statistic, also known as the ABBA-BABA statistic, and related statistics are commonly used to assess evidence of gene flow between populations or closely related species. While the calculations are not computationally intensive, currently available implementations require custom file formats and are impractical to evaluate all gene flow hypotheses across datasets that include many populations or species. Dsuite is a fast C++ implementation, allowing genome scale calculations of the D-statistic across all combinations of tens or even hundreds of populations or species directly from a variant call format (VCF) file. Furthermore, the program can estimate the admixture fraction and provide evidence of whether introgression is confined to specific loci. Thus Dsuite facilitates assessment of gene flow across large genomic datasets.

**Availability and implementation:** Source code and documentation are available at: https://github.com/millanek/Dsuite

**Contact:** millanek@gmail.com

## 1 Introduction

There is evidence that admixture between populations and hybridization between species are common and that a bifurcating tree is often insufficient to capture their evolutionary history (Green *et al.*, 2010; Patterson *et al.*, 2012; Tung and Barreiro, 2017; Malinsky *et al.*, 2018; Kozak *et al.*, 2018). The Patterson's D-statistic, first used to detect introgression between modern human and Neanderthal populations (Green *et al.*, 2010; Durand *et al.*, 2011), has since been widely applied and used across a broad range of taxa (Fontaine *et al.*, 2015; vonHoldt *et al.*, 2016; Malinsky *et al.*, 2018; Kozak *et al.*, 2018). The D-statistic and the related estimate of the admixture fraction *f*, also referred to as $f_4$ (Patterson *et al.*, 2012), are easy to calculate and well suited for taking full advantage of genome-scale datasets, while being robust under most demographic scenarios (Durand *et al.*, 2011).

Several programs for calculating D and *f* from genomic data are available, among them ADMIXTOOLS (Patterson *et al.*, 2012), $D_{FOIL}$ (Pease and Hahn, 2015), HyDe (Blischak *et al.*, 2018), and Comp-D (Mussmann *et al.*, 2019). However, what limits their utility is that none of the programs can handle the variant call format (VCF) (Danecek *et al.*, 2011), the standard file format for storing genetic polymorphism data produced by variant callers such as samtools (Li, 2011) and GATK (DePristo *et al.*, 2011). Moreover, as each calculation of D and *f* applies to four populations or taxa, the number of calculations/quartets grows geometrically with the size of the dataset. Therefore, previous studies utilizing D and *f* involved small numbers of populations or taxa, with few exceptions (Malinsky *et al.*, 2018; Kozak *et al.*, 2018). With ever more genomic data becoming available, there is a need for computationally efficient implementation able to handle datasets with tens and even hundreds of taxa. Dsuite addresses the above issues in that it calculates D and *f* statistics directly from VCF files, and is substantially more efficient than other programs. Thus Dsuite further increases the accessibility and applicability of admixture testing based on these statistics.

## 2 Methods and implementation

The statistics calculated by Dsuite apply to biallelic SNPs across four populations or taxa: P1, P2, P3, and O, related by the asymmetric tree (((P1,P2),P3),O). The outgroup O defines the ancestral allele, denoted by A, and the derived allele is denoted by B. The site patterns are ordered such that the pattern BBAA refers to P1 and P2 sharing the derived allele, ABBA to P2 and P3 sharing the derived allele, and BABA to P1 and P3 sharing the derived alleles. Under the null hypothesis, which assumes

no gene flow, the ABBA and BABA patterns are expected to occur with equal frequencies, while a significant deviation from that suggests possible introgression between P3 and either P1 or P2. See (Durand *et al.*, 2011) for more detail.

While simple site pattern counts can be computed for single sequences, the Dsuite implementation works with allele frequency estimates, so multiple individuals can, and ideally should, be included from each population or taxon. Denoting the allele frequency estimate at site $i$ in P1 as $\hat{p}_{i1}$, the following sums are calculated across all $n$ biallelic sites:

$$\begin{aligned} nABBA &= \sum_{i=1}^{n} \left(1-\hat{p}_{i1}\right) * \hat{p}_{i2} * \hat{p}_{i3} * \left(1-\hat{p}_{iO}\right) \\ nBABA &= \sum_{i=1}^{n} \hat{p}_{i1} * \left(1-\hat{p}_{i2}\right) * \hat{p}_{i3} * \left(1-\hat{p}_{iO}\right) \\ nBBAA &= \sum_{i=1}^{n} \hat{p}_{i1} * \hat{p}_{i2} * \left(1-\hat{p}_{i3}\right) * \left(1-\hat{p}_{iO}\right) \end{aligned} \quad (1)$$

### 2.1 The `Dtrios` program

Dsuite does not assume *a priori* knowledge of population or species relationships, only the outgroup has to be specified. The first subprogram, `Dtrios` calculates the sums in equation (1) for all trios of populations or taxa in the dataset. The command produces three types of output. For the first, in a file with the "BBAA.txt" suffix, `Dtrios` attempts to infer the population or species relationships: it orders each trio assuming that the correct tree is the one where the BBAA pattern is more common than the discordant ABBA and BABA patterns, which are assumed to result for example from incomplete lineage sorting, repeated mutation at the same site, or from introgression. In addition P1 and P2 are ordered so that $nABBA >= nBABA$ and, therefore

$$D = \frac{nABBA - nBABA}{nABBA + nBABA} \quad (2)$$

is never negative. The second type of output is the $D_{min}$ score, the minimum D for each trio, regardless of any assumptions about the tree topology (Malinsky *et al.*, 2018). There is no attempt to infer the true tree; instead, the trio is ordered so that the difference between *nABBA* and *nBABA* is minimized. This output is in a file with the "Dmin.txt" suffix. Finally, there is also an option for the user to supply a tree in Newick format to specify known or hypothesized relationships between the populations or species. An output file with the "tree.txt" suffix then contains D values for trios ordered in a way consistent with this tree.

To assess whether D is significantly different from zero, `Dtrios` uses a standard block-jackknife procedure as in (Green *et al.*, 2010; Durand *et al.*, 2011), obtaining an approximately normally distributed standard error. For three types of output, `Dtrios` calculates the Z-scores: $Z = D/std_err(D)$, and outputs the associated p-values. However, when testing more than one trio, users should take into account the multiple testing problem and adjust the p-values accordingly. Although the different D-statistics calculated on the same dataset do not constitute independent tests, users could consider them as such and control for overall false discovery rate, as this is a straightforward and conservative approach.

### 2.2 The `DtriosCombine` program

It is common practice, especially for larger genomic datasets, that VCF files are divided into smaller subsets by genomic regions, e.g. per chromosome. This facilitates the parallelization of many computational workflows. The `DtriosCombine` program enables parallel computation of the D-statistic across genomic regions, by combining the outputs of multiple `Dtrios` runs, summing up the counts in equation (1). It also calculates overall block-jackknife standard error across all regions.

### 2.3 The `Dinvestigate` program

The program `Dinvestigate` is designed to provide further information about trios that have significant D-statistics. `Dinvestigate` estimates the admixture fraction $f$, using $f_G$ as in (Green *et al.*, 2010), and also calculates $f_d$ (Martin *et al.*, 2015) and $f_{dM}$ (Malinsky *et al.*, 2015), two statistics related to $f_G$ but better suited to be applied to genomic windows. The $f_d$ and $f_{dM}$ statistics can be used to assess whether the admixture signal is confined to specific loci and to assist in locating any such loci.

Calculating $f_G$, the Green *et al.* estimate of $f$, requires that P3 be split into two subsets: P3a and P3b. If the dataset contains more than one individual from P3, then individuals are randomly assigned to the two subsets. If there is only one individual from P3, then only SNPs where the individual is homozygous for the derived allele are used.

The program outputs overall $f_G$, and also the more conservative $f_d$, and $f_{dM}$ to the standard output for each trio specified by the user. In addition, it produces a text file for each trio which contains the values of $f_d$ and $f_{dM}$ in sliding genomic windows, with the size of the windows specified by the user as a fixed number of SNPs.

## 3 Performance

I assessed the computational performance of Dsuite by calculating the D-statistic across the dataset of 73 species of Lake Malawi cichlid fishes from (Malinsky *et al.*, 2018). These 73 species are represented in the dataset by 131 individuals. The outgroup is one individual from a Lake Tanganyika cichlid *Neolamprologus brichardi*. For the tests, I used 610,333 SNPs that were called on scaffold_0, the largest scaffold of the reference genome assembly.

**Table 1.** Computational efficiency of D-statistic estimation on 610,333 SNPs in 73 species of Lake Malawi cichlid fishes.

| Software | Options | Peak memory | Run time |
|---|---|---|---|
| Dsuite `Dtrios` | none | 73MB | 113m5s |
| Admixtools `qpDstat` | blgsize: 0.01 | 27,212MB | 125m2s |
| HyDe `run_hyde.py` | none | 178MB | 231m38s |
| Comp-D* | -d -H -b10 | 8,300MB | 24hours+ |

*Comp-D cannot use allele frequencies calculated across multiple individuals, so only one individual per species included.

Table 1 shows a comparison of run time and memory requirements for calculating the D-statistic in all directions across all combinations of the 73 species. This corresponds to 62,196 trios and 186,588 D-statistics. Dsuite was the most efficient of the programs, both in terms of memory requirements and in terms of run time, being the only software to finish in under two hours. Admixtools was comparably fast, but required more than 27GB of memory, ~370x more than Dsuite. HyDe was more memory efficient, but the run time was longer, at just under four hours. Finally, Comp-D required substantial amount of memory at 8.3GB and also took by far the longest to complete the analysis; I cancelled the run after 24hours, with only ~3,000 trios completed.

While the Dsuite analysis was run directly on the VCF file, all other software required format conversion. For Admixtools, I first obtained

data in the PED format using vcftools v0.1.12b (Danecek *et al.*, 2011) with the `--plink` option, and then translated these into the software-specific EIGENSTRAT format using the `convertf` program, which is included in the Admixtools package. Data conversion into the PHYLIP input format for HyDe and Comp-D was done using the vcf2phylip script (Ortiz, 2019). The conversions added approximately ten minutes to each of these workflows, in addition to the run times shown in Table 1.

## Acknowledgements

I would like to thank Richard Durbin, Michael Matschiner, Walter Salzburger, and Hannes Svardal for useful discussions and comments.

## Funding

This work has been supported by EMBO, grant ALTF 456-2016 to MM, and , and the Swiss National Science Foundation (SNF), grant 176039 to Walter Salzburger..

*Conflict of Interest:* none declared.

## References

Blischak,P.D. *et al.* (2018) HyDe: A Python Package for Genome-Scale Hybridization Detection. *Syst. Biol.*, **67**, 821–829.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

DePristo,M.A.M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Durand,E.Y. *et al.* (2011) Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, **28**, 2239–2252.

Fontaine,M.C. *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**, 1258524.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Kozak,K.M. *et al.* (2018) Genome-wide admixture is common across the Heliconius radiation. *bioRxiv*, 414201.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Malinsky,M. *et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, **350**, 1493–1498.

Malinsky,M. *et al.* (2018) Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, **457**, 830.

Martin,S.H. *et al.* (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.*, **32**, 244–257.

Mussmann,S.M. *et al.* (2019) Comp-D: a program for comprehensive computation of D-statistics and population summaries of reticulated evolution. *Conservation Genet Resour*, **16**, 1–5.

Ortiz,E.M. (2019) vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis.

Patterson,N. *et al.* (2012) Ancient Admixture in Human History. *Genetics*, **192**, 1065–1093.

Pease,J.B. and Hahn,M.W. (2015) Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Syst. Biol.*, **64**, 651–662.

Tung,J. and Barreiro,L.B. (2017) The contribution of admixture to primate evolution. *Curr. Opin. Genet. Dev.*, **47**, 61–68.

vonHoldt,B.M. *et al.* (2016) Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Sci Adv*, **2**, e1501714.