User Manual for `VolcanoFinder` v1.0

Derek Setter, Sylvain Mousset, Xiaoheng Cheng, Rasmus Nielsen,
Michael DeGiorgio, Joachim Hermisson

February 24, 2020

1. Introduction

VolcanoFinder is a program to perform genome-wide scans of adaptive introgression by using the spatial distribution of within-species polymorphism and between species substitution in the genome. The current version of the software implements all four model test statistics of Setter and Mousset *et al*. (2019).

If you identify any bugs or issues with the software, then please contact Michael DeGiorgio at mdegiorg@fau.edu to report the issue.

If you use this software, then please cite it as

> D Setter, S Mousset, X Cheng, R Nielsen, M DeGiorgio, J Hermisson (2019) VolcanoFinder: genomics scans for adaptive intogression. *bioRxiv* doi:XXXX.

Note also that much of the code-base used for VolcanoFinder comes from SweepFinder (Nielsen *et al*. 2005) and SweepFinder2 (DeGiorgio *et al*. 2016).

2. Installation

VolcanoFinder should run on a **Linux** system (we will work to extend the functionality to other Unix-based systems in the future). To extract and compile the program, enter on the command line:

```
>tar -xzvf volcanofinder_v1.0.tar.gz
>cd volcanofinder_v1.0
>make
```

3. Input file format

Note: If you have used a Windows operating system to generate the following input files, then you will likely need to run the UNIX command dos2unix on each of the files before using them as input in VolcanoFinder.

3.1. Allele frequency file

The allele frequency input file is tab-delimited and contains a header. Each row represents the allele frequency for a position in the genome, and the rows are ordered by increasing position along the chromosome. There should be a separate allele frequency file for each chromosome when performing a scan for adaptive introgression. At each row, the first column is the physical position on the chromosome, the second column is the allele count (x), the third column is the sample size (n), and the fourth column is an indicator as to whether the site has been polarized (*i.e*., whether it is known that the allele is derived or ancestral). If the site is polarized, then the entry in the folded column should be 0, and the entry in the second column should be the derived allele count x. If the site is not polarized, then the entry in the folded column should be 1. The allele count can take on values $x = 1,2,\ldots,n$. If $x = n$ and the site is polarized, then the site is a substitution (monomorphic and different from the outgroup used to polarize the site). Otherwise, the site is a polymorphism. An example input file is (say for chromosome 6):

position	x	n	folded
460000	9	100	0
460010	100	100	0
460210	30	78	1
463000	1	94	0
…	…	…	…

The first line of the example input file displays the header, which must be identical to this example. The next four rows display allele frequency data for four positions on a chromosome (positions 460000, 460010, 460210, and 463000). Each row indicates the number of derived alleles observed (and the total number of alleles observed) at the given chromosomal position. At position 460000, 9 derived alleles were observed out of 100 total alleles (50 diploid individuals) leading to an observed polymorphism in the sample. At position 460010, 100 derived alleles were observed out of 100 total alleles (50 diploid individuals) leading to an observed substitution in the sample. At position 460210, 30 alleles of one type were observed out of 78 total alleles (39 diploid individuals), leading to another observed polymorphism in the sample. Note that at this row we set folded to 1 because we were not sure whether the allele was derived or ancestral. At position 463000, one derived allele was observed out of 94 total alleles (47 diploid individuals), leading to an observed polymorphism in the sample. If the true sample size was 50 diploid individuals, then positions 460210 and 463000 would be genomic positions with missing data in the sample.

3.2. Empirical unnormalized site frequency spectrum

The empirical unnormalized site frequency spectrum file is tab-delimitated. In a sample of size n alleles, the input file will have n rows. The xth row, $x = 1,2, \dots, n$, represents the fraction of callable sites in the sample that have exactly x derived alleles. By callable sites we mean any site, regardless of whether it was monomorphic ancestral, monomorphic derived, or polymorphic. The first column represents the number of derived alleles x, $x = 1,2, \dots, n$, such that the xth row has value x in the first column. The second row represents the fraction of callable sites with a particular number of derived alleles, such that the xth row has a value in the second row equal to the fraction of callable site with exactly x derived alleles. An example unnormalized site frequency spectrum file is (say for a sample of $n = 6$ alleles):

```
1	0.00156
2	0.00064
3	0.00043
4	0.00032
5	0.00026
6	0.01337
```

This input file states that in our sample of six alleles, the proportion of callable sites with 1, 2, 3, 4, 5, or 6 derived alleles was 0.00156, 0.0064, 0.00043, 0.00032, 0.00026, and 0.01337, respectively.

3.3. User-defined grid file

The user-defined grid input file has a simple format with a single position on each row (there is no header). Each position will specify a genomic location at which the test statistic will be computed. The positions in the user-defined grid file should be spanned by the range of positions in the allele frequency input file. Only those positions in the user-defined grid file will have an adaptive introgression test computed. That is, providing a user-defined grid file overrides the uniform grid option that is default to `VolcanoFinder`. An example user-defined grid input file:

```
460000
460010
460210
463000
…
```

The first four rows indicate that a test for adaptive introgression will be computed at positions 460000, 460010, 460210, and 463000.

3.4. User-defined D-value grid file

The input file for the user-defined grid of genetic distance D values has a simple format, with the first row indicating the number of genetic distance D values to evaluate, and each subsequent row indicating a distinct genetic distance D to evaluate. Only those genetic distances D in the user-defined grid file will be used in tests for adaptive introgression. That is, providing a user-defined D-value grid file overrides the genetic other procedures for cycling across genetic distances that are default to `VolcanoFinder`. An example user-defined D-value grid input file:

```
5
0.0015
0.0030
0.0045
0.0060
0.0075
```

The first row indicate that the user is providing a set of five genetic distance D values, with the subsequent five rows indicating that $D \in \{0.0015, 0.0030, 0.0045, 0.0060, 0.0075\}$.

4. Helper files (especially useful for large sample sizes)

`VolcanoFinder` generates a lookup table for its probability distribution prior to performing a scan for adaptive introgression. However, this computation can be computationally burdensome, and so it may make analyses substantially faster if the probability distribution was calculated once. This speedup will be particularly apparent when computing `VolcanoFinder` on many replicate simulated datasets, or across a genome that has been broken into many separate analysis blocks (see section 5.2).

4.1. Compute the probability lookup table (`-p`)

To compute the probability distribution lookup table, use the `-p` option. The command is:

```
./VolcanoFinder -p SpectFile D P MODEL nmin nmax xmin xmax LookupPrefix
```

where `SpectFile` is an input file containing the empirical unnormalized site frequency spectrum defined in section 3.2, `D` is a user-defined number indicating the genetic distance (D of Setter and Mousset *et al*. [2019]) between the sampled population and donor introgressing population, `P` takes on a value of 0 if fixed differences are not polarized (see Setter and Mousset et al. [2019] for details) and any other value if they are polarized, MODEL takes on values 1 or 2, indicating either Model 1 or Model 2 of Setter and Mousset *et al*. (2019), `nmin` and `nmax` are the respective minimum and maximum sample sizes in future allele frequency input files (see section 3.1), `xmin` and `xmax` are the respective minimum and maximum number of derived alleles in future input files, and `LookupPrefix` is the prefix to the file for which the relevant probability lookup table information will printed. The two files generated by this command will be names `LookupPrefix_dvalues` and `LookupPrefix_lookuptable`. Note that if the value for `D` is negative (*i.e*., $D < 0$), then instead `VolcanoFinder` cycles over a grid of valid values for the genetic distance D and chooses the value that leads to the greatest likelihood ratio, with the values equal to $D \in \{\hat{\theta}_\pi, 2\hat{\theta}_\pi, \ldots, k\hat{\theta}_\pi\}$ under Model 1 or $D \in \{2\hat{\theta}_L, 3\hat{\theta}_L, \ldots, k\hat{\theta}_L\}$ under Model 2, where k is chosen to be the maximum integer such that $D < D_o$ for polarized fixed differences and $D < 2D_o$ for non-polarized fixed differences, where $2D_o$ is the genetic distance between the ingroup and the outgroup sequence and is also computed internally in the software from the unnormalized site frequency spectrum (see Setter and Mousset *et al*. [2019] for details), where $\hat{\theta}_\pi$ is Tajima's estimate of the population-scaled mutation rate θ, and where $\hat{\theta}_L$ is an unbiased estimator of θ. The quantities D_o, $\hat{\theta}_\pi$, and $\hat{\theta}_L$ are all computed internally in the software from the unnormalized site frequency spectrum (see Setter and Mousset *et al*. [2019] for details). Moreover, if the value for `D` is a non-numeric string indicating the name of a

filename (*e.g.*, `UserDValueGrid.txt`), then instead `VolcanoFinder` will use the user-defined D-value grid with the specified filename (*e.g.*, `UserDValueGrid.txt`), which is defined in section 3.4.

5. Scanning for adaptive introgression

5.1. Scan for adaptive introgression (`-i`, `-ig`,`-iu`,`-pi`, `-pig`, or `-piu`)

To perform a scan for adaptive introgression, use the `-i` option. The command to perform this scan is

```
./VolcanoFinder -i G FreqFile SpectFile D P MODEL OutFile
```

where `G` is a user-defined number of grid points (G test sites are equally spaced across the genomic region spanned by the positions in `FreqFile`) to compute the test statistic, `FreqFile` is the allele frequency input file defined in section 3.1, `SpectFile` is an input file containing the empirical unnormalized site frequency spectrum defined in section 3.2, `D` is a user-defined number indicating the genetic distance (D of Setter and Mousset *et al*. [2019]) between the sampled population and donor introgressing population, `P` takes on a value of 0 if fixed differences are not polarized (see Setter and Mousset et al. [2019] for details) and any other value if they are polarized, MODEL takes on values 1 or 2, indicating either Model 1 or Model 2 of Setter and Mousset *et al*. (2019), and `OutFile` is the name of a file where the results will be printed. Here, `FreqFile` would be for a specific chromosome (or region of the genome) rather than combined across all chromosomes. Note that if the value for `D` is negative (*i.e.*, $D < 0$), then instead `VolcanoFinder` cycles over a grid of valid values for the genetic distance D and choose the value that leads to the greatest likelihood ratio, with the values equal to $D \in \{\hat{\theta}_\pi, 2\hat{\theta}_\pi, \dots, k\hat{\theta}_\pi\}$ under Model 1 or $D \in \{2\hat{\theta}_L, 3\hat{\theta}_L, \dots, k\hat{\theta}_L\}$ under Model 2, where k is chosen to be the maximum integer such that $D < D_o$ for polarized fixed differences and $D < 2D_o$ for non-polarized fixed differences, where $2D_o$ is the genetic distance between the ingroup and the outgroup sequence and is also computed internally in the software from the unnormalized site frequency spectrum (see Setter and Mousset *et al*. [2019] for details), where $\hat{\theta}_\pi$ is Tajima's estimate of the population-scaled mutation rate θ, and where $\hat{\theta}_L$ is an unbiased estimator of θ. The quantities D_o, $\hat{\theta}_\pi$, and $\hat{\theta}_L$ are all computed internally in the software from the unnormalized site frequency spectrum (see Setter and Mousset *et al*. [2019] for details). Moreover, if the value for `D` is a non-numeric string indicating the name of a filename (*e.g.*, `UserDValueGrid.txt`), then instead `VolcanoFinder` will use the user-defined D-value grid with the specified filename (*e.g.*, `UserDValueGrid.txt`), which is defined in section 3.4.

Sometimes it is more convenient to set the spacing between grid points rather than the number of grid points. The user may specify the approximate desired spacing between test sites using the `-ig` option. The command to perform this scan is

```
./VolcanoFinder -ig g FreqFile SpectFile D P MODEL OutFile
```

where `g` is a user-defined space between grid points. For example, if the user desired a test site approximately every one kilobase, then $g = 1000$, representing 1000 nucleotides.

Further, it can often be useful to use a custom grid of test sites rather than a uniform grid. The user may specify this custom grid using the `-iu` option. The command to perform this scan is

```
./VolcanoFinder -iu Gridfile FreqFile SpectFile D P MODEL OutFile
```

where `GridFile` is a user-defined gird input file defined in section 3.4.

If a lookup table has already been computed for the probability distribution using the `-p` command in section 4.1, then scans can be performed conditional on this lookup using the commands

```
./VolcanoFinder -pi G FreqFile SpectFile LookupPrefix OutFile

./VolcanoFinder -pig g FreqFile SpectFile LookupPrefix OutFile

./VolcanoFinder -piu Gridfile FreqFile SpectFile LookupPrefix OutFile
```

where `LookupPrefix` is the prefix to the file for which the relevant probability lookup table information was printed using the `-p` command in section 4.1, and where the command `-pi` functions like `-i`, `-pig` functions like `-ig`, and `-piu` functions like `-iu`.

5.2. Scan for adaptive introgression in blocks (`-bi`, `-big`, `-pbi`, or `-pbig`)

For computational reasons, it is sometimes more convenient to distribute jobs on a cluster with a single scan on a chromosome broken up into multiple sub-scans. The user may specify the number of roughly equally-sized blocks to break a scan of a chromosome into as well as the particular block to scan. This way, a chromosome can be broken up into say 10 blocks, and test sites within each of the 10 blocks can be scanned on separate cores while all tests still have the whole chromosome available to them to compute test site likelihood ratios. To perform a scan for adaptive introgression using blocks in which test sites have been equally spaced across an entire chromosome, use the `-bi` option. The command to perform this scan is

```
./VolcanoFinder -bi G FreqFile SpectFile D P MODEL OutFile BLOCK NBLOCK
```

where `G` is a user-defined number of grid points (G test sites are equally spaced across the genomic region spanned by the positions in `FreqFile`) to compute the test statistic, `FreqFile` is the allele frequency input file defined in section 3.1, `SpectFile` is an input file containing the empirical unnormalized site frequency spectrum defined in section 3.2, `D` is a user-defined number indicating the genetic distance between the sampled population and donor introgressing population, `P` takes on a value of 0 if fixed differences are not polarized (see Setter and Mousset et al. [2019] for details) and any other value if they are polarized, MODEL takes on values 1 or 2, indicating either Model 1 or Model 2 of Setter and Mousset *et al*. (2019), `OutFile` is the prefix of a file where the results will be printed, `BLOCK` is the current block to compute test sites one (with valid values of 1, 2, …, `NBLOCK`), and `NBLOCK` is positive integer representing the number of contiguous blocks in which to break a chromosome. Here, `FreqFile` would be for a specific chromosome (or region of the genome) rather than combined across all chromosomes. The output file produced from scanning with genomic blocks will have the name `OutFile_BLOCK_NBLOCK`, indicating that we have broken this scan into `NBLOCK` blocks and this output file is for the block number `BLOCK`. Note that if the value for `D` is negative (*i.e.*, $D < 0$), then instead `VolcanoFinder` cycles over a grid of valid values for the genetic distance D and chooses the value that leads to the greatest likelihood ratio, with the values equal to $D \in \{\hat{\theta}_\pi, 2\hat{\theta}_\pi, \ldots, k\hat{\theta}_\pi\}$ under Model 1 or $D \in \{2\hat{\theta}_L, 3\hat{\theta}_L, \ldots, k\hat{\theta}_L\}$ under Model 2, where k is chosen to be the maximum integer such that $D < D_o$ for polarized fixed differences and $D < 2D_o$ for non-polarized fixed differences, where $2D_o$ is the genetic distance between the ingroup and the outgroup sequence and is also computed internally in the software from the unnormalized site frequency spectrum (see Setter and Mousset *et al*. [2019] for details), where $\hat{\theta}_\pi$ is Tajima's estimate of the population-scaled mutation rate θ, and where $\hat{\theta}_L$ is an unbiased estimator of θ. The quantities D_o, $\hat{\theta}_\pi$, and $\hat{\theta}_L$ are all computed internally in the software from the unnormalized site frequency spectrum (see Setter and Mousset *et al*. [2019] for details). Moreover, if the value for `D` is a non-numeric string indicating the name of a filename (*e.g.*, `UserDValueGrid.txt`), then instead `VolcanoFinder` will use the user-defined D-value grid with the specified filename (*e.g.*, `UserDValueGrid.txt`), which is defined in section 3.4.

Sometimes it is more convenient to set the spacing between grid points rather than the number of grid points. The user may specify the approximate desired spacing between test sites using the `-ig` option. The command to perform this scan while breaking the chromosome into blocks is

```
./VolcanoFinder -big g FreqFile SpectFile D P MODEL OutFile BLOCK NBLOCK
```

where `g` is a user-defined space between grid points. For example, if the user desired a test site approximately every one kilobase, then $g = 1000$, representing 1000 nucleotides.

If a lookup table has already been computed for the probability distribution using the `-p` command in section 4.1, then scans can be performed conditional on this lookup using the commands

```
./VolcanoFinder -pbi G FreqFile SpectFile LookupPrefix OutFile BLOCK NBLOCK

./VolcanoFinder -pbig g FreqFile SpectFile LookupPrefix OutFile BLOCK NBLOCK
```

where `LookupPrefix` is the prefix to the file for which the relevant probability lookup table information was printed using the `-p` command in section 4.1, and where the command `-pbi` functions like `-bi` and `-pbig` functions like `-big`.

5.3. Merging the output of blocks (`-m`)

When performing scans for adaptive introgression using by breaking a chromosome into blocks as in section 5.2, it is important to be able to merge the output of these blocks. The command to merge a blocks for a given chromosome scan is

```
./VolcanoFinder -m OutFile NBLOCK
```

where `OutFile` is the prefix for all scans as indicated in section 5.2, and `NBLOCK` is the number of blocks that the scan was computed with. This `-m` option will take all files produced from the commands in section 5.2 with names

```
OutFile_1_NBLOCK
OutFile_2_NBLOCK
…
OutFile_NBLOCK_NBLOCK
```

and produce an output file with the name `OutFile` that has all blocks combined.

6. Output file format

The final output file (produced either by the procedures in sections 5.1 or 5.3) is tab-delimited and contains a header. Each row represents the calculation of a log likelihood ratio test statistic for adaptive introgression at a given grid position. The first column is the position on the chromosome of the test site, the second column is the log likelihood ratio test statistic for adaptive introgression, the third columns is the value of α (see Setter and Mousset *et al*. [2019]), and the fourth column is the value of genetic distance D between the sampled population and the donor population. If the genetic distance parameter `D` in sections 5.1 and 5.2 was positive, then D is the value `D` input by the user. If `D` is instead negative, then D is the value that maximizes the log likelihood, where D is computed across a grid of values as indicated in sections 5.1 and 5.2.

7. Examples

The `example_input` directory provides example input files for neutral and adaptive introgression simulations. For the following commands, we assume that executable `VolcanoFinder` is located in the `/usr/bin` directory in a Unix environment. If `VolcanoFinder` is being used in a local `example_input` directory, then `VolcanoFinder` in all examples should have a `./` in front such that it is called using the command `./VolcanoFinder` as in section 5.

7.1. Neutral simulation with a non-admixed background

This example applies to example input file `psvf_2300_0001.txt`, which should be used together with unnormalized site frequency spectrum file `spectvf_2300.txt`, which is from a non-admixed genomic background. This sample (with size $n = 40$ alleles) is aimed to be a neutral reference without admixture in the genomic background. To run `VolcanoFinder` with model 1 on these input files over a grid of 800 equally-spaced test sites while simultaneously inferring the maximum genetic divergence D with a putative donor population that has polarized fixed differences, use the command

```
VolcanoFinder -i 800 psvf_2300_0001.txt spectvf_2300.txt -1 1 1 vf_2300_0001.out
```

where the output of the scan is stored in `vf_2300_0001.out`. Figure 1 shows plots of the maximum likelihood test statistic, $-\log_{10}(\alpha)$, and maximum D as a function of test site location.

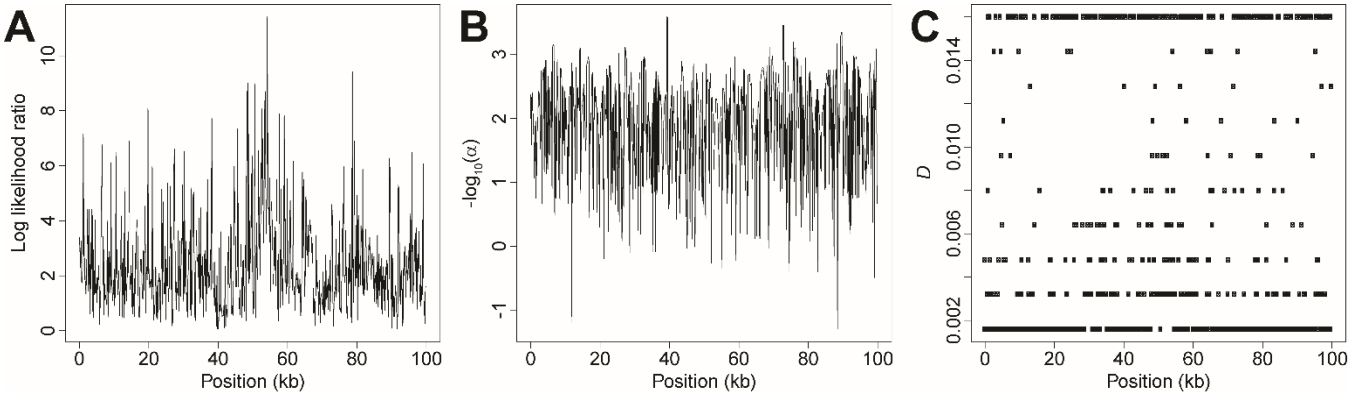

Figure 1: `VolcanoFinder` applied to data simulated under neutrality with a non-admixed genomic background. (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

7.2. Neutral simulation with an admixed background (example 1)

This example applies to example input file `psvf_2377_0001.txt`, which should be used together with unnormalized site frequency spectrum file `spectvf_2377.txt`, which is from an admixed genomic background. This sample (with size $n = 40$ alleles) is aimed to be a neutral reference with admixture in the genomic background. The admixture is from a donor population that diverged at $T_d = 5.5$ (see Setter and Mousset *et al*. [2019] for details). To run `VolcanoFinder` with model 1 on these input files over a grid of 800 equally-spaced test sites while simultaneously inferring the maximum genetic divergence D with a putative donor population that has polarized fixed differences, use the command

```
VolcanoFinder -i 800 psvf_2377_0001.txt spectvf_2377.txt -1 1 1 vf_2377_0001.out
```

where the output of the scan is stored in `vf_2377_0001.out`. Figure 2 shows plots of the maximum likelihood test statistic, $-\log_{10}(\alpha)$, and maximum D as a function of test site location.

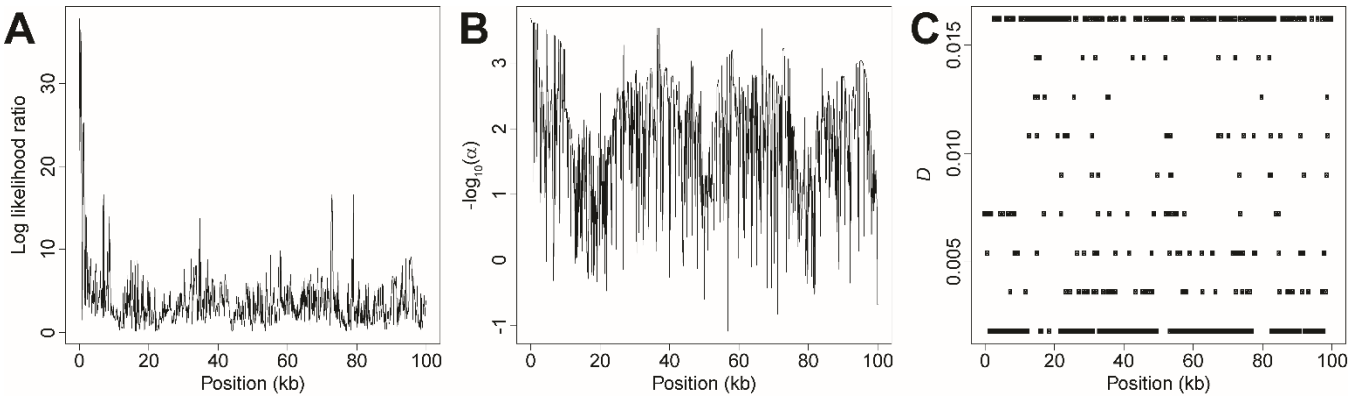

Figure 2: `VolcanoFinder` applied to data simulated under neutrality with an admixed genomic background. (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

7.3. Neutral simulation with an admixed background (example 2)

This example applies to example input file `psvf_2393_0001.txt`, which should be used together with unnormalized site frequency spectrum file `spectvf_2393.txt`, which is from an admixed genomic background. This sample (with size $n = 40$ alleles) is aimed to be a neutral reference with admixture in the genomic background. The admixture is from a donor population that diverged at $T_d = 5.5$ (see Setter and Mousset *et al*. [2019] for details). To run `VolcanoFinder` with model 1 on these input files over a grid of 800 equally-spaced test sites while simultaneously inferring the maximum genetic divergence D with a putative donor population that has polarized fixed differences, use the command

```
VolcanoFinder -i 800 psvf_2393_0001.txt spectvf_2393.txt -1 1 1 vf_2393_0001.out
```

where the output of the scan is stored in `vf_2393_0001.out`. Figure 3 shows plots of the maximum likelihood test statistic, $-\log_{10}(\alpha)$, and maximum D as a function of test site location.

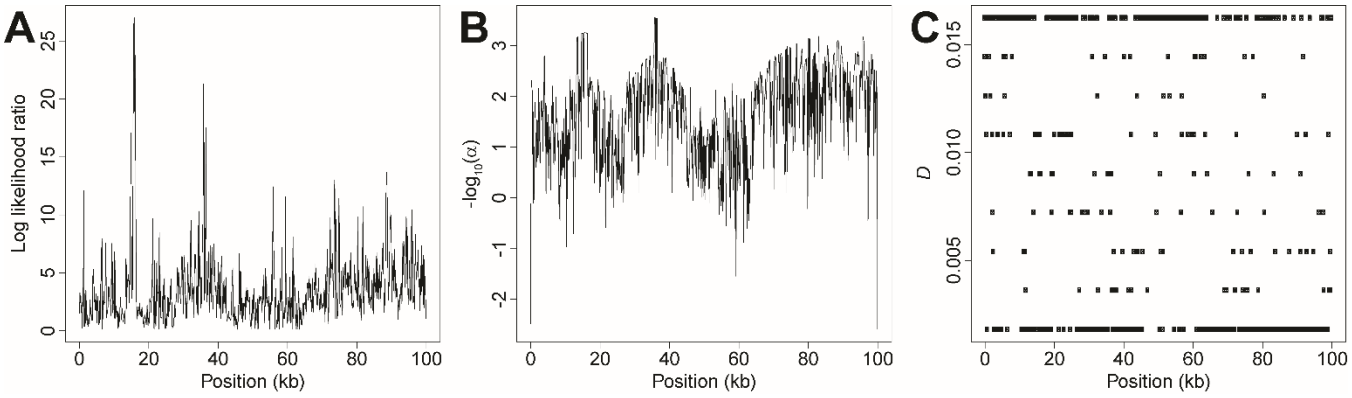

Figure 3: `VolcanoFinder` applied to data simulated under neutrality with an admixed genomic background . (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

7.4. Selection simulation with or without an admixed background (example 1)

This example applies to example input file `psvf_2277_0615.txt`, which should be used together with unnormalized site frequency spectrum files `spectvf_2300.txt` (which is from a non-admixed genomic background) or `spectvf_2377.txt` (which is from an admixed genomic background). This sample (with size $n = 40$ alleles) was generated with a selection model of a hard introgression sweep, with the age of the sweep of $T_s = 0.0$, a population-scaled selection coefficient of $2Ns = 500$, from a donor population that diverged at $T_d = 5.5$ (see Setter and Mousset *et al*. [2019] for details). To run `VolcanoFinder` with model 1 on these

input files over a grid of 800 equally-spaced test sites while simultaneously inferring the maximum genetic divergence D with a putative donor population that has polarized fixed differences, use the commands

```
VolcanoFinder -i 800 psvf_2277_0615.txt spectvf_2300.txt -1 1 1 vf_2277_0615_2300.out

VolcanoFinder -i 800 psvf_2277_0615.txt spectvf_2377.txt -1 1 1 vf_2277_0615_2377.out
```

where the output of the scan is stored in `vf_2277_0615_2300.out` when using the unadmixed background or in `vf_2277_0615_2377.out` when using the admixed background. Figures 4 and 5 show plots of the maximum likelihood test statistic, $-\log_{10}(\alpha)$, and maximum D as a function of test site location for the unadmixed and admixed backgrounds, respectively.

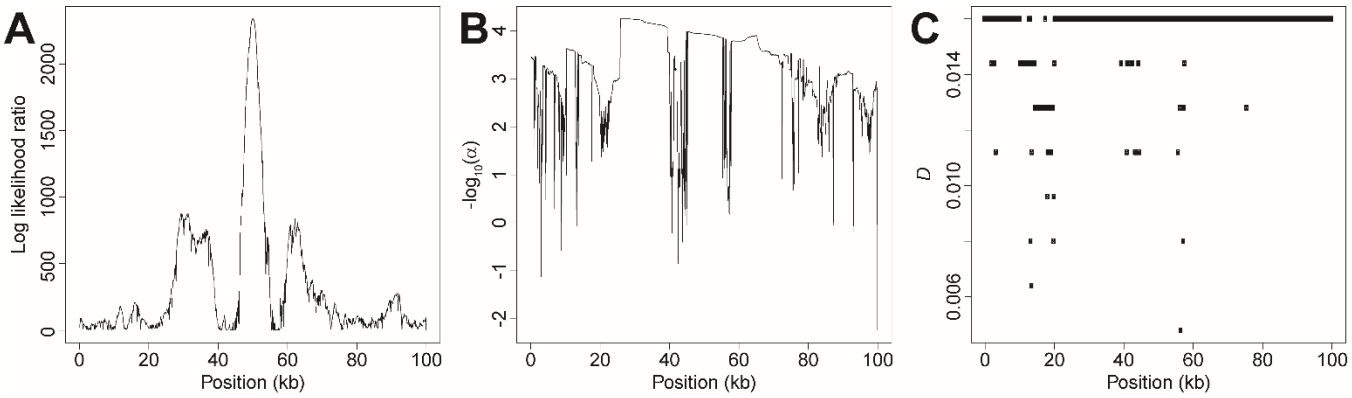

Figure 4: `VolcanoFinder` applied to data simulated under adaptive introgression with a non-admixed genomic background. (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

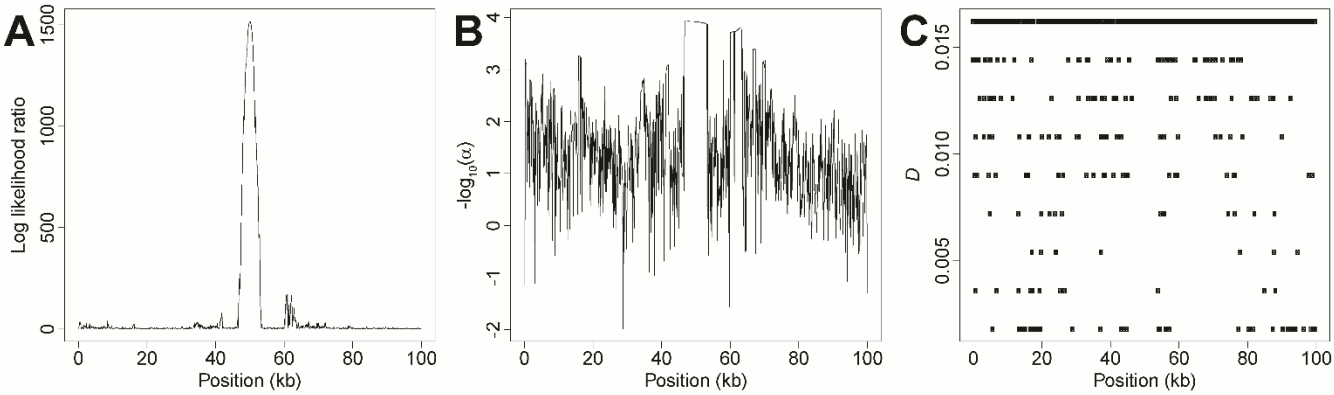

Figure 5: `VolcanoFinder` applied to data simulated under adaptive introgression with an admixed genomic background. (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

7.5. Selection simulation with or without an admixed background (example 2)

This example applies to example input file `psvf_2293_0242.txt`, which should be used together with unnormalized site frequency spectrum files `spectvf_2300.txt` (which is from a non-admixed genomic background) or `spectvf_2393.txt` (which is from an admixed genomic background). This sample (with size $n = 40$ alleles) was generated with a selection model of a hard introgression sweep, with the age of the sweep of $T_s = 0.0$, a population-scaled selection coefficient of $2Ns = 500$, from a donor population that diverged at $T_d = 5.5$ (see Setter and Mousset *et al*. [2019] for details). To run `VolcanoFinder` with model 1 on these input files over a grid of 800 equally-spaced test sites while simultaneously inferring the maximum genetic divergence D with a putative donor population that has polarized fixed differences, use the commands

```
VolcanoFinder -i 800 psvf_2293_0242.txt spectvf_2300.txt -1 1 1 vf_2293_0242_2300.out

VolcanoFinder -i 800 psvf_2293_0242.txt spectvf_2393.txt -1 1 1 vf_2293_0242_2393.out
```

where the output of the scan is stored in `vf_2293_0242_2300.out` when using the unadmixed background or in `vf_2293_0242_2393.out` when using the admixed background. Figures 6 and 7 show plots of the maximum likelihood test statistic, $-\log_{10}(\alpha)$, and maximum D as a function of test site location for the unadmixed and admixed backgrounds, respectively.

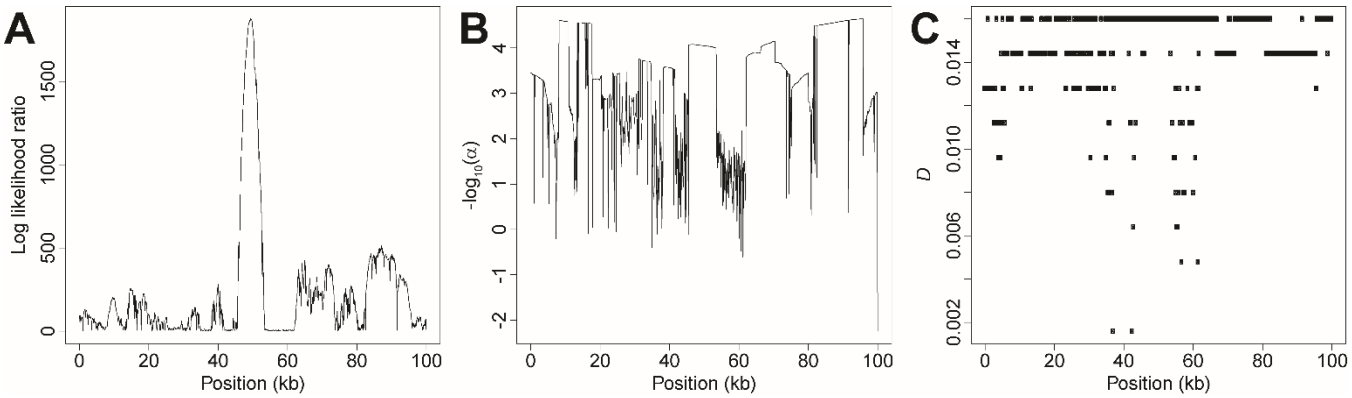

Figure 6: `VolcanoFinder` applied to data simulated under adaptive introgression with a non-admixed genomic background. (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

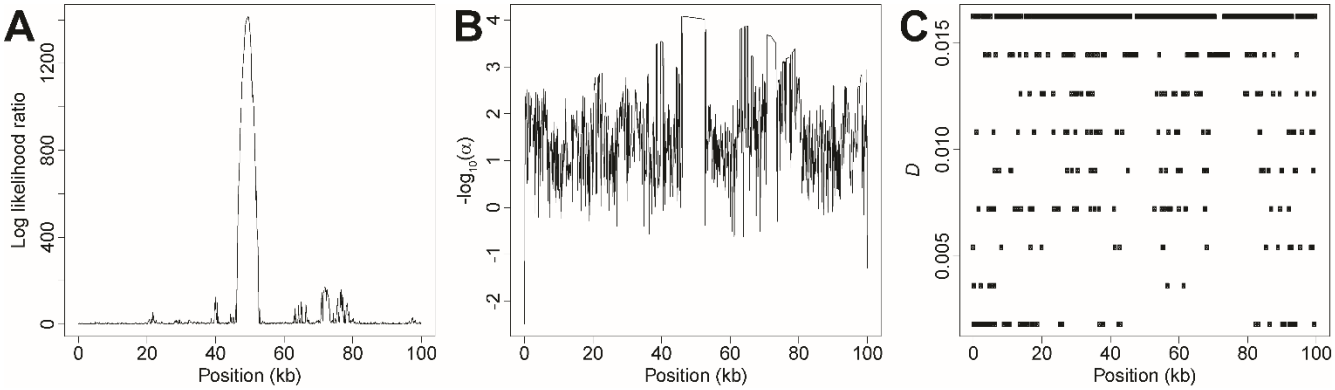

Figure 7: `VolcanoFinder` applied to data simulated under adaptive introgression with an admixed genomic background. (**A**) Maximum log likelihood ratio test statistic. (**B**) Maximum $-\log_{10}(\alpha)$. (**C**) Maximum D.

8. References

R Nielsen, S Williamson, Y Kim, MJ Hubisz, AG Clark, C Bustmante (2005) Genomic scans for selective sweeps using SNP data. *Genome Res* 15:1566-1575.

M DeGiorgio, CD Huber, MJ Hubisz, I Hellmann, R Nielsen (2016) *SweepFinder2*: Increased sensitivity, robustness, and flexibility. Bioinformatics 32:1895-1897.

D Setter, S Mousset, X Cheng, R Nielsen, M DeGiorgio, J Hermisson (2019) VolcanoFinder: genomics scans for adaptive intogression. *bioRxiv* doi:XXXX.