**SambaR main function options**

**28-11-2020**

This document contains an overview of the main function options of the software SambaR, a R pipeline for rapid population genetic analyses on biallelic SNP datasets. SambaR can be downloaded from *https://github.com/mennodejong1986/SambaR* and subsequently loaded into R using the source command. The pipeline consists of 7 main functions: getpackages, importdata, filterdata, findstructure, calcdistance, calcdiversity, and selectionanalyses. The getpackages(), importdata() and filterdata() functions should be executed sequentially. Afterwards, the remaining functions can be executed in any preferred order. More detailed instructions are provided in the SambaR manual.

**getpackages()** *(install and load all packages needed for the SambaR pipeline)*

|  |  |  |
| --- | --- | --- |
| **argument** | **Default** | **explanation** |
| myrepos | "http://cran.us.r-project.org" | CRAN repository |
| mylib | NULL | If you do not have root access (administrator rights) on your computer, you need to specify the path to an existing directory where you want install the packages. For example, if you want to install the packages in the directory ‘home/userX/Rpackages’, you would type: getpackages(myrepos='http://cran.us.r-project.org',mylib=‘home/userX/Rpackages’). SambaR will automatically add the specified directory to the library paths, so that R knows where to find the installed packages. |
| noupdates | TRUE | Suppress package updates. |
| do\_halt | TRUE | Abort if an error is encountered; if FALSE, jump to next package. |
| silent | TRUE | Suppress verbose |

**importdata()** *(import RAW/BIM files into R, as well as optional files with additional info)*

|  |  |  |
| --- | --- | --- |
| **argument** | **default** | **explanation** |
| inputprefix | NULL | Prefix of input RAW and BIM file (i.e. without .raw and .bim extension). |
| sumstatsfile | FALSE | If TRUE, the function expects to find in the working directory a file called 'sumstats.tsv' generated by the STACKS refmap pipeline. You should include the sumstats.tsv file if you called your SNPs using the STACKS refmap pipeline, regardless of whether you created your PED/MAP files from a genepop file or vcf file. A genepop file does not contain information on the positions of your SNPs, and as result the first two columns of the MAP file wil contain zero’s only. A vcf file does contain information about the positions of your SNPs, but the vcf file created by STACKS shows the position of your SNPs in the STACKS loci rather than position in the reference genome |
| snpinfofile | NULL | Name of tab separated file with additional loci information. Number of lines should equal number of lines of BIM file. |
| samplefile | NULL | Name of optional input file with population assignment. At minimum the file should contain two tab-separated columns called 'name' and 'pop'. The name column contains the sample name (identical to names specified in PED-file) and the pop column contains the population names (max 10 characters, no underscores or spaces). Population subdivision can be specified in an optional third column called 'pop2'. This sample file is not needed if the populations are defined in the first column of the PED-file. |
| geofile | NULL | Name of optional input file with geographical coordinates (longitude and latitude) of the samples. Input file should contain three tab-separated columns called ‘sample’, ‘longitude’ and ‘latitude’. Sample names in the geofile should be identical to sample names in PED-file. The coordinates should be in decimal degrees (e.g. New York: lat = 40.714, lon = -74.006). |
| depthfile | FALSE | If TRUE, the function expects to find in the working directory the files 'out.idepth' and 'out.ldepth.mean', generated with the software vcftools using the --depth and --site-mean-depth flags. (Note that a vcf file created with ‘plink --recode vcf’ does not contain read depth information and therefore cannot be to used to calculate mean read depth.) |
| nchroms | NULL | Number of chromosomes of the reference genome assembly (optional). |
| colourvector | NULL | Optional vector with colour names, which will be assigned to your populations in alphabetical order. The vector should have the following syntax: c("blue","darkgreen","darkred"), and the length of the vector should be equal or greather than the number of populations defined in the input data. |
| silent | TRUE | Suppress verbose |

**filterdata()** *(quality control and data filtering)*

|  |  |  |
| --- | --- | --- |
| **argument** | **default** | **explanation** |
| indmiss | 0.25 | Maximum allowed proportion of missing data per sample. By default, samples with more than 25% missing data will be excluded from subsequent analyses. |
| snpmiss | 0.1 | Maximum allowed proportion of missing data per SNP. By default, SNPs with more than 10% missing data (averaged over samples which passed the indmiss threshold) will be excluded from subsequent analyses. |
| min\_mac | 2 | Minimum allowed number of minor allele copies per SNP. By default, SNPs with contain only 1 copy of the minor allele (only considering samples which passed the indmiss threshold), will be excluded from subsequent analyses. |
| dohefilter | TRUE | If TRUE, the function will remove SNPs with heterozygosity levels which are potentially indicative of paralogs (see paralog\_threshold1 and paralog\_threshold2 argument). These SNPs will be highlighted in red in the 'He\_vs\_maf' plot which is generated by the filterdata function. Excluded from subsequent analyses are SNPs which meet two criteria: He > 2pq + pT1 and p>T2 (in which p and q denote respectively minor and major allele frequency, and in T1 and T2 denote respectively paralog\_threshold1 and paralog\_threshold2. |
| snpdepthfilter | TRUE | If TRUE, the function will remove SNPs with high read depth (if information is available). These SNPs will be highlighed in red in the 'Locusdepth' plot which is generated by the filterdata function. |
| min\_spacing | 500 | Minimum distance between adjacent SNPs (in bp) when thinning the data. In order to avoid linkage disequillibrium issues, SambaR will use this thinned dataset for structure, genetic distance and diversity analyses. |
| nchroms | NULL | Number of chromosomes of the reference genome assembly (optional). If a value is provided, SambaR will generate additional plots showing the number of SNPs per chromosome, and try to determine the X-chromosome. |
| TsTv\_filter | NULL | Can be set to either 'Ts' or 'Tv' in order to exclude respectively transversions or transitions. |
| ychrom | NULL | Name of Y-chromosome (optional). |
| paralog\_threshold1 | 0.5 | Setting to filter out SNPs with high heterozygosity, indicative of paralogs (see dohefilter argument). |
| paralog\_threshold2 | 0.05 | Setting to filter out SNPs with high heterozygosity, indicative of paralogs (see dohefilter argument). |
| silent | TRUE | Suppress verbose |

**findstructure()** *(population structure analyses)*

|  |  |  |
| --- | --- | --- |
| **argument** | **default** | **explanation** |
| Kmax | 6 | Maximum number of clusters to be considered in DAPC and LEA admixture analyses. |
| add\_legend | TRUE | If TRUE, add legend to output plots. |
| pop\_order | NULL | Vector to specify order of populations in admixture plots (other than the default alphabetical and geographical order). For example: if your populations are named ‘America’, ‘Asia’, ’Europe’, you can specify an alternative order from left to right by running the command: findstructure(Kmax=6,pop\_order=c(‘Europe’,’Asia’,’America’). |
| legend\_pos | "bottomright" | String indicating position of legend in ordination plots. Accepted values: bottomright, bottom, bottomleft, right, center, left, topright, top, topleft. |
| legend\_cex | 3 | Legend size |
| symbol\_size | 3 | Symbol size |
| silent | TRUE | Suppress verbose |

**calcdistance()** *(population differentiation statistics)*

|  |  |  |
| --- | --- | --- |
| **argument** | **default** | **explanation** |
| nchroms | NULL | Number of chromosomes of the reference genome assembly (optional). If a value is provided, SambaR will generate karyotype plots showing sliding window genetic distance measures. |
| max\_npop\_stats | 6 | Maximum number of populations accepted for ABBA-BABA analyses. |
| silent | TRUE | Suppress verbose |

**calcdiversity()** *(population genetic diversity statistics)*

|  |  |  |
| --- | --- | --- |
| **argument** | **default** | **explanation** |
| legend\_cex | 2.5 | Legend size |
| nrsites | NULL | Combined length (in bp) of the sequences from which your SNP dataset is obtained, including the sequences which did not contain SNPs. If the SNP data has been generated using whole genome resequencing, an estimate of nrsites can be obtained by generating a (thinned and quality filtered) vcf file which contains both polymorphic and monomorphic sites prior to SNP calling. If the SNP data has been generated with the software STACKS, an estimate of nrsites can be obtained from the populations.log file. The end of this file contains a line stating: ‘Kept ……. loci out,’. (For older versions of STACKS, the line reads: ‘Of these, ……. loci/stacks passed the filters.) Multiplying this number with the length (in basepair) of the trimmed reads (e.g. 110 bp), returns a rough estimate of the number of retained sequenced sites (in bp). If a value is provided to the nrsites argument, the function will calculate estimates of genome wide heterozygosity, nucleotide diversity and Watterson’s theta and Tajima’s D. |
| silent | TRUE | Suppress verbose |

**selectionanalyses()** *(selection scans FstHet, GWDS, OutFLANK and PCadapt)*

|  |  |  |
| --- | --- | --- |
| **argument** | **default** | **Explanation** |
| do\_meta | TRUE | If TRUE, the function will run selection analyses for the metapopulation, considering all populations simultaneously. This option will exclude the selection scan GWDS, as GWDS works for pairwise comparisons only. |
| do\_pairwise | FALSE | If TRUE, the function will run selection analyses for all pairwise population comparisons. Depending on the number of populations, this can result in a high number of analyses. |
| do\_pheno | FALSE | If TRUE, the function will run selection analyses for population pairs which are assigned to different groups (as defined by the inds$type column). The division can be based on either a phenotypic trait or on geographical occurrence. Populations can be split in two groups by adding to the inds dataframe a column called ‘inds$type’ containing TRUE/FALSE values. For example: if your data contains two marine populations (pop1 and pop2) and two freshwater populations (pop3 and pop4), a division between freshwater and marine populations can be specified with the command: *inds$type<-ifelse(inds$pop=="pop1"|inds$pop=="pop2",TRUE,FALSE)*. |
| onlypooled | TRUE | If TRUE, the function will run selection analyses for a pooled population comparison (as defined by inds$type column) and not for all pairwise populations comparisons between both groups. To be used in conjunction with do\_pheno argument. |
| export | NULL | File type of output plots. Accepted values are 'eps','pdf','png', and 'wmf'). |
| export\_data | FALSE | If TRUE, PED and MAP files will be exported which contain either neutral SNPs (marked by none of the selected outlier scans) or outlier SNPs (marked by at least one of the selected outlier scans). |
| overwriteped | TRUE | If FALSE, the function will not generate new PED and MAP files. This can save time if rerunning the function, but can lead to errors if run for the first time or if run with different filter settings. |
| do\_pcadapt | TRUE | If TRUE, the function will run the software PCadapt. |
| do\_outflank | TRUE | If TRUE, the function will run the software OutFLANK. OutFLANK throws errors at times, which can only be circumvented by omitting OutFLANK from the analyses by setting this argument to FALSE. |
| do\_fsthet | FALSE | If TRUE, the function will run the software FstHet. The selection scan ‘Fsthet’ takes a relatively long time to run (like 30 minutes for a dataset of 60 individuals and 50,000 loci), and is therefore excluded from the analyses by default. If TRUE, the function expects to find a genepop file of the original input data in the input data directory (i.e. same directory as where RAW/BIM files are stored.) |
| add\_bayescan | FALSE | If TRUE, the function expects to find Bayescan output files (originally ending on 'baye\_fst.txt' or 'g\_\_fst.txt' but renamed to ‘pop1\_pop2.bayescanout.fst’ or ‘pheno.bayescanout.fst’) in the SambaR inputfiles directory, and will include the information in these files to the output plots and tables. |
| bayescan\_FDR | 0.01 | Bayescan false discovery rate |
| overwrite\_  bayescan | FALSE | If TRUE, and if the function has been run before, data in snps dataframe will be overwritten. |
| pheno\_labels | c(‘pheno1’,  ‘pheno2’) | Optional vector (consisting of two elements) specifying the group division (e.g. c('marine','freshwater'). To be used in conjuction with the do\_pheno argument. Labels will be included in the output plots. |
| my\_correction | ‘bonferroni’ | String which specifies the multiple test correction method. Values can be NULL (default), ‘bonferroni’, ‘holm’, and ‘BH’, the latter denoting Benjamini-Hochberg. If NULL, GWDS will be run using the Bonferroni correction, PCadapt will be run using the Holm correction holm, and OutFLANK will be run using q-values. |
| do\_thin | FALSE | The do\_thin argument applies to the GWDS test, and indicates whether the neutral distribution should be inferred from a thinned dataset. If set to TRUE, the GWDS scan will use one SNP per genomic region to infer the neutral distribution. The size of these genomic regions can be adjusted with the gwdsbinsize argument (default is 1000000 bp). The ‘do\_thin’ option should be set to TRUE only if working with dense SNP datasets. When setting the argument ‘do\_thin’ to TRUE for datasets with relatively low number of SNPs (e.g. RADseq SNP datasets), GWDS will return many false positives. |
| gwdsbinsize | 1000000 | Thinning parameter applied by GWDS selection scan (see do\_thin argument). |
| silent | TRUE | Suppress verbose |