**Explanation of GWDS (Genome Wide Differentiation Scan), implemented in R package SambaR**

*https://github.com/mennodejong1986/SambaR*

GWDS is a SNP-by-SNP interpopulation selection scan which searches for SNPs under (linked) positive selection, by comparing allele frequencies between pairs of populations (or groups of populations), in a dataset of biallelic SNPs. GWDS can be regarded the between-population variant of GWAS.

***STEP 1. SNP by SNP Fisher exact test on allele counts***

For all *n\_loci* SNPs, Fisher exact test p-values are calculated based on 2x2 tables of minor and major allele counts per population, using the function *fisher.test().* These pF-values quantify the difference between the two population, like pairwise Fst-values. For any SNP with pF = 1, the allele frequencies in population 1 are identical to the allele frequencies in population 2. SNPs with pF = 0 have fixated differentially in the two populations (e.g. population 1 is fixed for A, and population 2 for T). Obtained pF-values are converted using a negative natural log:

*fisher\_logscore* <- *-log(fisher.test(allele\_count\_matrix)$p-value*)

*Hypothetical example for a single SNP:*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | major allele count | minor allele count | total allele count (non-missing) |  |  |
| Pop1 | **42** | **22** | 64 |  | Fisher exact test p-value: 0.01914415 |
| Pop2 | **19** | **27** | 46 |  | Negative natural log: **3.955758** |
| *total* | *61* | *49* | *110* |  |  |

***STEP 2. Goodness-of-fit test between observed data and fitted exponential curve***

GWDS assumes that the distribution of obtained log-scores can be described by an exponential curve of which the rate parameter (λ) equals the inverse of the mean (µ) of all log-scores:

*f(x,λ) = λ·e-λ·x in which: λ = 1/ µ*

To check this assumption, a chi-squared goodness of fit test is executed between observed data and *n\_loci* random data points generated using the fitted exponential function, divided into 1 unit bins:

*gwds\_fitteddata <- rexp(n\_loci=n\_loci,rate=1/mean(fisher\_logscores))*

*gwds\_chiout <- chisq.test(cbind(gwds\_logscores,gwds\_fitteddata))*

GWDS issues a warning in case of a poor fit between the fitted curve and the observed log-scores.



***STEP 3. Assign p-values, derive outlier threshold and mark outliers***

Based on the Lewontin-Krakauer axiom (which states that demography (i.e. genetic drift) affects the entire genome roughly uniformly, whereas selection only affects particular parts of the genome), GWDS assumes that SNPs which stands out from the overall distribution, are putatively SNPs under positive selection.

Because the fitted exponential curve (f(x,λ) = λ·e-λ·x)is a probability density function, it can be used to assign p-values to each SNP log-score. These pgwds-values reflect the probability of observing a log-score given the distribution of all other log-scores in the dataset. By default the pgwds-values are corrected for multiple testing using the Bonferroni correction method, but users can also select the Holm or Benjamini-Hochberg correction method. Adjusted pgwds-value below 0.05 (or any other user defined significance threshold) are marked as outliers:

*GWDS\_logpscores <- -log10(1-pexp(fisher\_logscores, λ))*

*GWDS\_bonferroni\_p <- 1-(0.05/n\_loci)*

*GWDS\_bonferroni\_q <- qexp(GWDS\_bonferroni\_p, λ)*

*myoutliers <- GWDS\_logpscores > GWDS\_bonferroni\_q*

GWDS output plots display the logs of Fisher exact test p-values as well as the final GWDS p-values:



***OPTIONAL: thin data prior to STEP2***

As can be inferred from the explanation above, rather than trying to infer the expected null distribution based on a demographic model (such as island models applied in Fdist approaches and Bayescan), GWDS, like OutFLANK, takes the approach of directly inferring the null distribution from the data. This direct approach is less confounded by assumptions – e.g. no need for assumptions about population hierarchy, equal population sizes, levels of gene flow, etc.

One major assumption of GWDS is though that the vast majority of SNPs are neutral, and that therefore the neutral distribution can be inferred from the overall distribution. On a total of thousands of SNPs, one or two SNPs under selection will not massively influence the overall mean, and hence the rate parameter of the fitted exponential distribution.

This assumption is however violated by dense SNP datasets in which many SNPs can represent the same selective sweep. In that case, GWDS offers the option to infer the neutral distribution from a thinned dataset which contains at maximum 1 SNP per 1MB (or other user defined size). (In comparison, OutFLANK deals with this issue by trimming the top 5% of Fst-values before fitting the probability distribution.)