To obtain the estimated enrichments in introgressed segments follow these steps. Note that the scripts are intended for reproducing our results only. If you want to use the scripts for another purpose it is at your own risks. Please contact David Enard at davidenard@gmail.com for further information.

1) Run the matching_permutations.pl script.
It is the main script in the analysis. What it does is to give a list of (i) VIPs that are matched with at least 3 non-VIPs with similar genomic properties and (ii) lists of random control non-VIPs.

Example of use:

```
./matching_permutations.pl /Users/davidenard/Dropbox/plosgenetics_github/
ensembl_protcoding_genes_v83.txt VIPs_file.txt recombination.txt
factors_table_EUR.txt nearest_vip_distance.txt intervals_EUR.txt 100
matched_VIPs.txt matched_nonVIPs.txt 500000 0.0001
```

Arguments:

$indir : the directory where the script is copied together with all the files needed to run it (see other arguments).

$valid_file : a file that contains the Ensembl identifiers of genes to include in the analysis. For example we used it to restrict the analysis to Ensembl protein-coding genes from Ensembl version 83. Use ensembl_protcoding_genes_v83.txt

$vip_file : file with information on which genes are VIPs and which genes are non-VIPs. Use VIPs_file.txt

$rec_file: a file with gene-specific recombination rates. Use recombination.txt. The rates in the recombination.txt file are in cM per 200kb.

$factors_table_file: the table that includes all the factors to control for. Use factors_table_EUR.txt for Europe and factors_table_ASN.txt for East Asia. 1st column: Ensembl gene ID. 2nd column: CDS density. 3rd column: DNASEI density. 4th column: FUNSEQ. 5th column: GC content. 6th column: recombination. 7th column: Tajima's D. 8th: PhastCons conserved element density.

$distance_file: file with for each Ensembl gene the distance to the nearest VIP. Use nearest_vip_distance.txt.

$intervals_file: file with the lower and upper bound values used to define the tolerance intervals of each controlled genomic factor for matching. The factor at the first line in the file must correspond to the factor at the first column in $factors_table_file, the second line to the second column and so on. The script does not check if there are as many lines in $intervals_file as there are factor columns in $factors_table_file so be careful with this. Use intervals_ASN.txt for East Asia and intervals_EUR.txt for Europe.

$iterations : the number of control sets of matched non-VIPs to generate.

$out_file_1 : list of matched VIPs, i.e. those VIPs with three or more matched control non-VIPs. matched _VIPs.txt in the example.

$out_file_2 : control sets of matched non-VIPs. One set per line. Matched_nonVIPs.txt in the example.

$dist : minimal distance between a non-VIP and the nearest VIP for the non-VIP to be included among controls. Set at 500000 for the analysis.

$cutrec: minimum recombination threshold. Only genes with recombination rate above $cutrec are included in the analysis. Since recombination rates in recombination.txt are in cM per 200kb, set $cutrec at 0.3 for a threshold of 1.5 cM/Mb. To use all genes with recombination information we set $cutrec at 0.0001 (0.0001 cM per 200kb or 0.0005 cM/Mb).

2) Use get_introgressed_genes.pl to prepare file introgressed_genes.txt

In the file introgressed_genes.txt, genes within an introgressed segment above the desired frequency and length threshold are associated with a "1" while other genes are associated with a "0".

Example 1:

```
./get_introgressed_genes.pl ensembl_protcoding_genes_v83.txt
all_neand_ASN_cutoff0.05.txt introgressed_genes.txt 100000
```

→ prepares genes in introgressed segments at frequencies larger than 5% (cutoff0.05) and longer than 100kb in East Asia.

Example 2:

```
./get_introgressed_genes.pl ensembl_protcoding_genes_v83.txt
all_neand_EUR_cutoff0.15.txt introgressed_genes.txt 150000
```

→ prepares genes in introgressed segments at frequencies larger than 15% (cutoff0.15) and longer than 150kb in Europe.

3) Measure the number of segments that overlap VIPs and control sets of non-VIPs with get_neand_clusters.pl

Usage: `./get_neand_clusters.pl ensembl_gene_coords.txt matched_VIPs.txt matched_nonVIPs.txt introgressed_genes.txt 1`

Output will be in the following order:

- -Enrichment ratio
- -Observed number of introgressed segments at VIPs
- -Expected average number of introgressed segments at VIPs
- -Confidence interval lower bound

-Confidence interval upper bound
-P-value