GoShifter v0.2, Manual

GoShifter is written for Python 2.7. It uses the following modules:

- bisect
- subprocess
- chromtree (provided)
- bx.intervals.cluster
- numpy

When testing for enrichment GoShifter uses LD information for the set of provided SNPs. To obtain this information please download precomputed pairwise SNP LD information from (these files are large!):

https://www.broadinstitute.org/~slowikow/tgp/pairwise_ld/

GoShifter is divided into two scripts: 1) **goshifter.py**, which tests for enrichment of a provided set of SNPs with one genomic annotation of interest, and 2) **goshifter.strat.py**, which tests for the significance of an overlap of a provided set of SNPs with annotation A stratifying on secondary, possibly colocalizing annotation B.

1) goshifter.py

Input files:

snpmap – file with mappings for the tested set of SNPs, tab delimited, with columns SNP, Chrom, BP. Chromosome in the format 'chrN'. Must include header. Example:

GoShifter/test_data/bc.snpmappings.hg19.txt

SNP	Chrom	BP
rs10069690	chr5	1279790
rs1045485	chr2	202149589
rs3757318	chr6	151914113
rs10941679	chr5	44706498
rs13281615	chr8	128355618
rs2823093	chr21	16520832
rs17530068	chr6	82193109
rs2380205	chr10	5886734
rs3803662	chr16	52586341

annotation – mappings of the annotation which will be tested for enrichment with the SNP set. Must be in BED format (gzipped), includes Chrom, Start, End columns (no header required). Chromosome in the format 'chrN'. Example:

zcat GoShifter/test_data/UCSF-UBC.Breast_vHMEC.bed.gz

chrY	128031	128231
chrY	142761	142961

```
chrY   231491  231691
chrY   231983  232183
chrY   233430  233630
chrY   285237  285437
chrY   296657  296857
chrY   1318260 1318460
chrY   1459641 1459841
```

```
./goshifter.py --snpmap FILE --annotation FILE --permute INT --ld DIR --out FILE [--rsquared NUM --window NUM --min-shift NUM --max-shift NUM --ld-extend NUM --no-ld]
```

Options:

-h, --help	Print this message and exit.
-v, --version	Print the version and exit.
-s, --snpmap FILE	File with SNP mappings, tab delimited, must include header: SNP, CHR, BP. Chromosomes in format chrN.
-a, --annotation FILE	File with annotations, bed format. No header.
-p, --permute INT	Number of permutations.
-l, --ld DIR	Directory with LD files. LD files must of name: chrN.EUR.ld.bgz
-r, --rsquared NUM	Include LD SNPs at r^2 >= NUM [default: 0.8]
-w, --window NUM	Window size to find LD SNPs [default: 5e5]
-n, --min-shift NUM	Minimum shift [default: False]. Defaults to random shifts.
-x, --max-shift NUM	Maximum shift [default: False]. Defaults to random shifts.
-e, --ld-extend NUM	Fixed value by which to extend LD boundaries [default: False]. Default is to extend LD block 2*median size of annotation.
-n, --no-ld	Do not include SNPs in LD [default: False]. If this is specified the SNPs only index SNP will be tested for enrichment. Note that at the moment you still have to provide a path to directory with LD info.
-o, --out FILE	Write output file.

Example usage:

```
./goshifter.py --snpmap test_data/bc.snpmappings.hg19.txt --annotation test_data/UCSF-UBC.Breast_vHMEC.bed.gz --permute 1000 --ld 1kG-beagle-release3/pairwise_ld/ --out test_data/bc.H3K4me1_vHMEC
```

This will output the message on the screen (and print the *P*-value corresponding to the significance of an overlap). The following output files will be created:

***.enrich** – output file with observed and permuted overlap values

nperm	nSnpOverlap	allSnps	enrichment
0	29	68	0.42647
1	17	68	0.25
2	26	68	0.38235
3	25	68	0.36765
4	23	68	0.33824
5	17	68	0.25
6	21	68	0.30882
7	25	68	0.36765
8	17	68	0.25

nperm = 0 is the observed overlap
nSnpOverlap – number of loci where at least one SNP overlaps an annotation
allSnps – total number of tested loci
enrichment – nSnpOverlap/allSnps

Note, *P*-value is the number of times the "enrichment" is greater or equal to the observed overlap divided by total number of permutations.

***. locusscore** – the likelihood of a locus to overlap an annotation under the null. The smaller the value the more likely a locus overlaps an annotation not by chance. Loci not overlapping any annotation are denoted as "N/A".

locus	overlap	score
rs11780156	1	0.609
rs4808801	1	0.996
rs10069690	N/A	N/A
rs12493607	1	0.888
rs1045485	N/A	N/A
rs204247	1	0.678
rs3757318	N/A	N/A
rs2943559	N/A	N/A
rs11552449	N/A	N/A

***.snpscore** – defines which LD SNPs are overlapping an annotation in the observed data

locus	ld_snp	overlap
rs11780156	rs11780156	1
rs11780156	rs1016578	0
rs11780156	rs12542202	0

rs11780156	rs11776569	0
rs11780156	rs7836152	0
rs11780156	rs10956414	0
rs11780156	rs67397162	0
rs11780156	rs11778142	1
rs11780156	rs11997192	0

2) goshifter.strat.py

Input files:

snpmap – see above

annotation-a – mappings of the primary annotation which will be tested for enrichment with the SNP set. See above for format details for the annotation input file.

annotation-b – mappings of the secondary annotation. Assessment of enrichment for annotation-a will be tested stratifying on this annotation-b. See above for format details for the annotation input file.

Usage:

./goshifter.strat.py --snpmap FILE --annotation-a FILE --annotation-b FILE --permute INT --ld DIR --out FILE [--rsquared NUM --window NUM --min-shift NUM --max-shift NUM --ld-extend NUM --no-ld]

Options, same as above with exception:

-a, --annotation-a FILE	File with primary annotations, bed format. Gzipped. No header.
-b, --annotation-b FILE	File with annotation to stratify on, bed format. Gzipped. No header.

Example usage:

./goshifter_v2_wLicense/goshifter.strat.py --snpmap test_data/bc.snpmappings.hg19.txt --annotation-a test_data/UCSF-UBC.Breast_vHMEC.bed.gz --annotation-b test_data/UCSF-UBC.Breast_Myoepithelial_Cells.bed.gz --permute 1000 --ld 1kG-beagle-release3/pairwise_ld/ --out test_data/bc.H3K4me1_vHMEC_strat_Myoepithelial_Cells

This will output the message on the screen (and print the *P*-value corresponding to the significance of an overlap) and write results to *.enrich (see above for explanation of the format).