SFDR (Stratified False Discovery Rate) Software Documentation

Version 1.6 Feb 7, 2010

Yun Joo Yoo, Shelley B. Bull, Andrew D.Paterson, Daryl Waggott, Lei Sun

1. Overview of the methods

FDR, SFDR, WFDR, q-values, prior information, linkage

False Discover Rate (FDR) method provides more powerful multiple hypothesis testing criteria than the conventional Family-Wise Error Rate (FWER) control method (Benjamini & Hochberg, 1995). FDR method, therefore, is well-suited for Genome-Wide Association (GWA) study of several hundred thousands of SNPs. FDR control is performed collectively to the p-values from all hypotheses, and the critical point for p-values varies for a fixed FDR control level depending on the underlying distribution of true alternative hypothesis. The FDR method can be also performed by obtaining FDR adjusted p-value, i.e. q-value for each hypothesis testing and comparing q-values with FDR control level directly (Storey,2002).

Stratified False Discovery Rate (SFDR) method performs FDR control separately for each stratum which has been decided by prior information or some characteristic of data (Sun et al., 2006). Weighted False Discovery Rate (WFDR) method assigns specific weight for each hypothesis and performs FDR control on weighted p-values (Roeder et al., 2006). These two FDR methods that incorporate prior information are especially suitable for the prior information of Genome-Wide Linkage study results used for the analysis of GWA study.

For SFDR method, the SNPs are divided into two strata using a threshold value for linkage score. If the corresponding linkage score at the location of a SNP is above the threshold, the SNP is assigned to the stratum 1 (high linkage stratum). Otherwise, the SNP is assigned to the stratum 2 (low linkage stratum).
For WFDR method, weight of SNP i is computed as the $w_i=\exp(B \cdot Z_i)/v$ where $v = \sum_i^N \exp(B \cdot Z_i) / N$, N is the number of SNPs and Z_i is the corresponding linkage score at the location of the SNP i.

Since SNP position is in physical base-pair (bp) position, and the linkage results are presented in centi-Morgan(cM) position, the conversion of SNP bp position into cM position is required. The genetic map of bp and cM position for the markers used in the linkage study can be used to interpolate the SNP position. Also, the linkage scores at any position between obtained linkage scores need to be interpolated.

2. About the software

The SFDR program computes q-values and ranks of all SNPs for one or more of selected FDR methods (FDR, SFDR, WFDR). The program only requires the end-results of genome-wide analysis: p-values of SNPs and the position of SNPs. For genome-wide linkage study data, the program requires the information of genome-wide linkage scores at cM positions. Also, the genetic map information of markers in terms of their between bp and cM position needs to be provided. The details about the what the SFDR program can perform are in next chapters.

3. How to run SFDR

To run SFDR program, you can either run the script SFDR.pl (if you already installed a Perl interpreter in you system.) or SFDR.exe (if it is compatible with your OS) with the options followed by the specified value for each option in the command line.

For example, if you want to compute FDR, SFDR, WFDR for a genome wide association data GWA.txt using a genome-wide linkage study result in GWL.txt , a map data MAP.txt store output file in OUTPUT.txt for only the SNP with the rank of q-values until 20, you type:

A. To use perl script SFDR.pl directly

```
SFDR.pl  -assoc GWA.txt -linkage GWL.txt  -map MAP.txt -SFDR  -WFDR -out OUTPUT.txt  -top 20
```

B. To use executive file SFDR.exe file.

```
SFDR  -assoc GWA.txt -linkage GWL.txt  -map MAP.txt -SFDR  -WFDR -out OUTPUT.txt  -top 20
```

Below, I will assume you are using the executive file SFDR.exe file. There are several cases of tasks that SFDR program can do. The example command for those cases are given below

Case 1: GWA data only, only FDR q-values are computed

```
SFDR  -assoc GWA.txt -out OUTPUT.txt  -top 20
```

Case 2: GWA data and GWL data are given. SFDR and/or WFDR q-values are computed.

```
SFDR  -assoc GWA.txt -linkage GWL.txt  -map MAP.txt -SFDR -WFDR  -B 2 -C 1.0  -out OUTPUT.txt  -top 20
```

*Note: -B is for WFDR weighting constant, -C is for SFDR threshold. For default values, see section 4.

Case 3: GWA data already have group assignment or prior values for each SNP, SFDR q-values and/or WFDR q-values are computed

```
SFDR  -assoc GWA.txt -SFDR -out OUTPUT.txt  -top 20
```

*If Prior values are characters, only SFDR q-values will be computed. If Prior values are numeric and >500 categories, only WFDR q-values will be computed.

Case 4: GWA data and only limited result of peak linkage region from GWL study. SFDR q-values are computed.

```
SFDR  -assoc GWA.txt -linkpeak Lpeak.txt  -map MAP.txt -SFDR  -out OUTPUT.txt  -top 20
```

4. Details of options

Option	Description	Default Value
-assoc	Genome-wide association study result file of SNP p-values. Mandatory option in any case	None
-linkage	Genome-wide linkage study result file of linkage scores (LOD or NPL).	None
-linkpeak	Genome-wide linkage study result file presented only for linkage peak information	None
-map	Genetic map file that matches physical position (bp) and linkage position (cM) of markers across genome	None
-out	Output file name	Same name as the input file for -assoc with an extension ".out"
-top	Only the SNPs with rank (in any FDR) not more than this value are printed in the output.	10^8
-SFDR	SFDR is performed	No value required
-WFDR	WFDR is performed	No value required
-C	SFDR threshold value	0.5
-B	WFDR weighting constant value	1
-nC	Number of cut point. If more than 1, it reads –C values as much as –nc value (e.g : -nC 2 -C 0.5 1)	1

5. How to prepare Input files

All input data files require header line with column names. SFDR program tries to locate the required column by searching the header. If you have an input file with all the required columns (extra columns are ok), then you need to make sure those columns have correct header that the program will identify. The column names are not case-sensitive, and also the order of the columns can be arbitrary. If the input file does not have the necessary columns with correct names, then SFDR program will produce an error message and stop. All input files are not required to be sorted.

1) GWA input file for –assoc option

Chr- chromosome identifier (This identifier should be consistent with other input files)

Snp – SNP name

Pos – Base-pair position of each SNP

P – Association analysis p-values for each SNP

Prior- group or prior value assignment for SFDR and WFDR, numeric or character (optional)

Example : GWA.txt

snp	chr	pos	p
Snp1	1	995669	0.199
Snp2	1	1038818	0.001
Snp3	1	1066927	0.387
…			

*****Missing data should be designated by symbol ".".

2) GWL input file for –linkage option

Chr- chromosome identifier (This identifier should be consistent with other input files)

cM- cM position for each computed linkage score

LOD (or NPL) – linkage score either as LOD score or NPL score

Example: GWL.txt

chr	cM	LOD
1	0.0	0.0000
1	5.0	0.0076
1	10.0	0.0280
…		

3) Linkage peak information file for –linkpeak option

Chr- chromosome identifier (This identifier should be consistent with other input files)

cM- cM position at peak linkage score

LOD (or NPL) – linkage score either as LOD score or NPL score

Example: Lpeak.txt

chr	cM	LOD
1	130.5	1.72
3	176.0	3.17
11	76.0	2.09
…		

4) Map file for –map option

Chr - chromosome identifier (This identifier should be consistent with other input files)

Marker – marker names relevant to linkage study

bp – physical position of the marker

cM – cM position of the marker

Example: MAP.txt

chr	marker	bp	cM
1	D1S2217	787021	0.00
1	D1S243	2129098	4.47
1	D1S468	3574721	9.23
…			

5. Output files

There will be two output files produced. One is for the q-value and rank outputs (out file), the other is for the summary of the SFDR analysis (log file). If the out file name has been designated by –out option, for example, OUTPUT.txt , then the log file name will be OUTPUT.log. If the out file name has not been designated by –out option and the input association file name was GWA.txt, then the out file name will be GWA.out and the log file name will be GWA.log.

1) Out file

Following information will be produced in the out file for each SNP

chr : chromosome identifier

snp : snp name

pos: bp position of the SNP

p : association p-value of the SNP

q_FDR : FDR q-value

r_FDR : rank of FDR q-value

q_SFDR: SFDR q-value

r_SFDR: rank of SFDR q-value

group: group assignment for SFDR

q_WFDR : WFDR q-value

r_WFDR : rank of WFDR q-value

weight : weight of each SNP computed for WFDR

Z : Interpolated linkage score for each SNP

*out file can be big. You can use –top option to produce only top rank SNPs results (small q-values).

2) Log file

The option values used and the number of SNPs, size of groups for SFDR, weight information for WFDR will be produced.

Example. OUTPUT.log

```
>>>>>>>SFDR v1.3 <<<<<<<<

Association Data File: GAW.txt
Linkage Data File: GWL.txt

NOTE: FDR will be done
NOTE: SFDR will be done.
NOTE: WFDR will be done.
NOTE: Map file MAP.txt will be used for linkage data.
NOTE: Only SNPs with rank <= 10 will be included in the output.

>>FDR summary
# of SNPs : 306284

>>WFDR summary
Weighting constant B: 1
Maximum Weight: 15.1898
Mininum Weight: 0.7728

>>SFDR summary
Strata: 1, 2
Threshold C: 0.5
```

6. Remarks

*You can download the SFDR program from http://www.utstat.toronto.edu/sun/

*Read the error messages and notes printed in the screen while you are running the program. See if these messages fit your intension for the program.

7. Contact information

Lei Sun
Dalla Lana School of Public Health
Department of Statistics
University of Toronto
Phone: (416) 978-7519
E-mail: sun at utstat dot toronto dot edu

Yun Joo Yoo
Samuel Lunenfeld Research Institute
Mount Sinai Hospital, Toronto
Phone: (416) 586-4800 ext 5836
E-amil: yoo at mshri do on dot ca

8. Change log

*Nov 10, 2008 – First public release of version 1.3

*Dec 2. 2008 – A bug related to parsing data corrected and missing data instruction included.

*Feb 7, 2010 – Program modified to incorporate prior information per SNP in the association data file

9. References

Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B. 57, 289-300 (1995)

Roeder, K., Bacanu, S.A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans.Am. J. Hum. Genet. 78, 243-252 (2006)

Storey, J.D. A direct approach to false discovery rates. J. R. Stat. Soc. B.64, 479-498 (2002)

Sun, L., Craiu, R.V., Paterson, A.D. & Bull, S.B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet. Epidemiol. 30, 519-530 (2006)