Hypothesis-Driven Genome Wide Association Study Using Stratified False Discovery Rate Control and Functional Epigenomic Data Version 1.0

Pipeline manual

Table of contents

1. Project Description

Name: Hypothesis-Driven Genome Wide Association Study (HD-GWAS) using stratified false discovery rate (sFDR) control and functional epigenomic data.

Authors

D. Giovana Carrasco-González
Bachelor in Technology
National University of Mexico (UNAM)
Querétaro, Qro. 76230 México

Jessica Dennis
PhD Candidate
Dalla Lana School of Public Health, University of Toronto
Toronto, ON M5T 3M7 Canada

Alejandra Medina-Rivera,
PhD, Associate Researcher
International Laboratory for Human Genome Research, National University of Mexico (UNAM)
Querétaro, Qro. 76230 México

1.1 Introduction

GWAS have identified many loci that predispose to common, complex diseases [Welter, et al. 2014]. The statistical threshold for declaring genome-wide significance, however, is strict, and many more trait-associated loci likely lie below this threshold, and could be identified from prioritized GWAS.

The HD-GWAS pipeline described herein prioritizes SNPs in functional epigenomic regions using stratified false discovery rate (sFDR) [Sun, et al. 2006] and/or weighted false discovery rate (wFDR) [Roeder, et al. 2006] control. Both the sFDR and wFDR are statistical methods to incorporate a prior hypothesis into a GWAS. The sFDR consists of assigning SNPs to high- and low-priority strata based on external information [Sun, et al. 2006]. If the prior is informative, power increases, whereas if the hypothesis is non-informative, the power of the HD-GWAS is the same as in an un-prioritized GWAS. The wFDR consists of assigning continuous weights to GWAS SNPs, where the weights are derived from external information [Roeder, et al. 2006]. The wFDR is as powerful as the sFDR if the prior is informative, but if the prior is misleading, the wFDR HD-GWAS is less powerful than an un-prioritized GWAS.

Functional epigenomic regions are enriched for trait-associated SNPs, especially when the epignomic region and trait share the same tissue specificity [Maurano, et al. 2012]. Such regions therefore contain SNPs that are obvious candidates for prioritization in HD-GWAS.

The analysis routines in this HD-GWAS pipeline were developed for an HD-GWAS of tissue factor pathway inhibitor plasma (TFPI) levels. We prioritized GWAS SNPs in active enhancers and promoters in vascular endothelial cells, a primary cell type that expresses *TFPI*. We observed a marginal but promising enrichment of TFPI plasma level-associated SNPs in these regulatory regions, and our work suggests that an HD-GWAS strategy with cell type-specific functional epigenomic data may prove valuable for other phenotypes as well.

1.2 Objective

To develop a user-friendly pipeline for running an HD-GWAS analysis using sFDR and/or wFDR control with a functional epigenomic data.

1.3 Description

The HD-GWAS pipeline contains Python and R scripts:

-Python script: This script receives the input files and additional information given by the user, verifies their formatting, and manages the analysis steps. The python script also allows users to access the pipeline's help function from the command line.

- sFDR and wFDR analysis scripts (R scripts): Depending on the analysis specified by the user (sFDR or wFDR), the script will perform the requested statistical analysis and provide results in .txt format in the folder specified by the user or in a default folder created by the program.

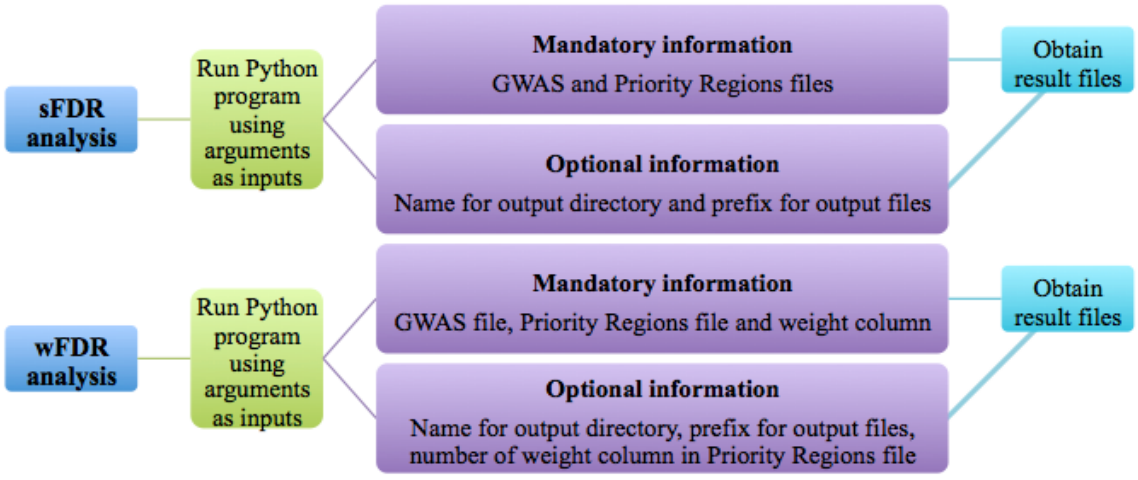

Fig. 1 Program Structure

2. Requirements

A unix based terminal (Ej: MacOS, Linux)

bedTools version 2.17 (or greater)

Python version 2.7

1.4 GB free disk space

HD-GWAS Package including demo files and correct result files

OPTIONAL CHOICE: Set up the directory containing the scripts in the unix PATH (see section 2.5). If you wish to skip the assignation of the PATH variable, you must run the next line in order to locate your HD_GWAS_Package:

```
cd /CORE DIRECTORY/HD_GWAS_Package-master/Scripts/
```

NOTE: "CORE DIRECTORY" is the full name of the directory that contains the HD_GWAS_Package.

2.1 bedTools installation

BedTools facilitates genome arithmetic with functions that allow users to intersect, merge, count, complement and shuffle genomic intervals in files with formats such as BAM, BED, GFF/GTF, VCF. Also, bedTools runs in the command line on UNIX, LINUX and Apple OS X operating systems.
For more information, and to review the installation process, read the bedTools manual: http://bedtools.readthedocs.org/en/latest/content/quick-start.html

2.2 Python

The required Python version can be downloaded using the next link http://www.python.org/downloads/

2.3 R Scripts from the package

The downloaded package must have a directory called "Scripts" that contains the next files:

- hdgwas_analysis.py
- run.hd.gwas.singlesnp.sfdr.R
- run.hd.gwas.singlesnp.wfdr.R

2.4 Demo files and Correct Results from the package

The following test files are provided in the directory "Demo_files":

- gwas.bed
- prioritized_regions.bed

The "CorrectResults" folder contains the results that you should obtain by running a demo.

- result_sFDR_singlesnp.sfdr.top1000.txt
- result_sFDR_singlesnp.sfdr.txt
- result_sFDR_snpList.txt
- result_wFDR_singlesnp.wfdr.top1000.txt
- result_wFDR_singlesnp.wfdr.txt
- result_wFDR_snpList.peaks.txt

2.5 Setting up the PATH (Optional)

To run the programs without mentioning the full directory path, in the command line, type:

```
open ~/.bash_profile
```

This will open a .txt file in which you must add the next lines:

################HD_GWAS_PACKAGE DIRECTORY

```
export HD_GWAS=/CORE DIRECTORY/
```

```
export PATH=${PATH}:${HD_GWAS}
```

NOTE: "CORE DIRECTORY" is the full name of the directory that contains the HD_GWAS_Package.

Afterwards, write the next command:

```
source ~/.bash_profile
```

Finally, echo the full path in the command line:

```
echo $PATH
```

After these steps, the tool is ready to use.

If problems are encountered, add the following instruction to the command line:

```
chmod a+x /CORE DIRECTORY/HD_GWAS_Package-
master/Scripts/hdgwas_analysis.py
```

3. Input, output and optional files

The python program receives the required and optional inputs, checks the files and, if the paths and the formats are correct, sends the information to the R scripts to run the specified analysis.

Input files should be given with a hyphen (-) followed by the file type using the arguments described in Table 1. Users are encouraged to use full file paths in order to avoid errors.

Mandatory files are the GWAS file and the Priority Regions file. The optional inputs are: the name for the output directory, a prefix for result files, and the column number for the weight to be used from the Priority Regions file. If the name of the output directory, and the prefix aren't given by the user, the default name "Results" will be used.

The user can select the type of analysis to be performed using the option "task", which can be: only sFDR analysis, wFDR analysis using the weight column variable from Priority Regions file, or both sFDR and wFDR analyses.

3.1 Input files and options

Table 1. Description of the correct format for input files.

Input	Argument	Possible Values	Mandatory or Optional
GWAS file	-gwas	Full path of .bed file	Mandatory
Priority Regions file	-priority_regions	Full path of .bed file or .broadPeak file	Mandatory
Output Directory that will be created	-out_dir	Full path of output directory provided by user	Optional
Prefix for result files in the output directory	-prefix	Name provided by user	Optional
Analysis task(s)	-task	sFDR or wFDR	Mandatory

Weight column	-column	Column number with weight information from the Priority Regions file	Mandatory only for the wFDR analysis

3.2 .bed file and .broadPreak file

For .bed and .broadPeak files, only the first four columns from the GWAS file and the first three columns from the priority regions file are required for running the analysis, any extra columns will be ignored. The column order is crucial for the analysis, and the files must not contain headers. An example of each file is provided in Table 2 and Table 3.

Table 2. GWAS file format.

chr1	100010764	100010765	rs1416885
chr1	100010855	100010856	rs1416884
chr1	100010998	100010999	rs1416883
chr1	100011242	100011243	rs4000172
chr1	100011252	100011253	rs1416882

Table 3. Priority regions file format.

chr8	134264537	134265502
chr8	23019467	23020633
chr8	129759541	129760947
chr8	40703924	40704262
chr8	122328112	122328822

The correct format for the first four columns of the GWAS file is: chromosome name (e.g., “chr6”), bp position, bp position +1, and SNP ID (must have alphanumeric characters). If these items aren’t in the files, an error will appear. Given the size of GWAS files, the program checks only the first five rows of the file and assumes the rest is correct.

The correct format for the first three columns of the priority regions file is: chromosome names (e.g., “chr6”), region start position (in bp), and region end position (in bp). If the file will be used for a wFDR analysis, the quantitative weight variable should also be included in this file in the fifth column.

3.3 Output .RData file

R scripts save results and intermediate data in .RData files, which are R workspace files that can be used to perform data acquisition and error troubleshooting.

4. Running the demo

Define the core directory as the directory that contains the downloaded package.

4.1 Examples

The help can be accessed using the following command:

```
hdgwas_analysis.py –h
```

To run an analysis using the demo files, use the following command:

```
python hdgwas_analysis.py –gwas  /FULL/PATH/FOR/GWAS/FILE –
priority_regions /FULL/PATH/FOR/PRIORITY/REGIONS/FILE –out_dir
NAME_OUTPUT_DIRECTORY –prefix PREFIX_NAME_FOR_GENERATED_FILES –
task SPECIFIC_ANALYSIS –column COLUMN_NUMBER_wFDR_ANALYSIS
```

It is not necessary to keep the order in the arguments.

Example sFDR analysis:

```
python hdgwas_analysis.py –gwas /CORE DIRECTORY/HD_GWAS_Package-
master/Demo_files/gwas.bed –priority_regions /CORE
DIRECTORY/HD_GWAS_Package-
master/Demo_files/prioritized_regions.bed –prefix result_sFDR –
task sFDR –out_dir DemoTest
```

Example wFDR analysis:

```
python hdgwas_analysis.py –gwas /CORE DIRECTORY/HD_GWAS_Package-
master/Demo_files/gwas.bed –priority_regions /CORE
DIRECTORY/HD_GWAS_Package-
master/Demo_files/prioritized_regions.bed –out_dir DemoTest –
prefix result_wFDR –task wFDR –column 3
```

Results from running the demo can be checked against the DemoTest results supplied with the package by running the following script:

```
diff  -s  /CORE  DIRECTORY/HD_GWAS_Package-master/CorrectResults
/CORE DIRECTORY/HD_GWAS_Package-master/DemoTest
```

4.2 Possible errors

Table 4. Possible errors and solutions.

Case	Description	Solution
Please add GWAS file	GWAS file is not given in the argument line	Write full path for GWAS file
Incorrect GWAS directory	GWAS directory doesn't exist in the referred path	Check referred path for GWAS file
Incorrect GWAS filename	GWAS file doesn't exist in the referred path	Check GWAS file existence
Please add Priority Regions file	Priority Regions file is not given or doesn't exist in the referred path	Write full path for Priority Regions file
Incorrect Priority Regions directory	Priority Regions directory doesn't exist in the referred path	Check referred path for Priority Regions file
Incorrect Priority Regions filename	Priority Regions file doesn't exist in the referred path	Check Priority Regions file existence
Please add missing arguments	GWAS file and Priority Regions file are not given or don't exist in the referred path	Write full path for mandatory arguments
GWAS with incorrect format	The first five lines of this document were checked by Python and they appear to be incorrect	Check the format of this file (section 3.2)
Priority Regions with incorrect format	The first five lines of this document were checked by Python and they appear to be incorrect	Check the format of this file (section 3.2)
Warning for repeating a prefix	Prefix is repeated so the file	Change the name by writing

	will be overwritten	the whole code line again
Column weight isn't a number or doesn't match the number of columns that Priority Regions file has	Wrong input for weight column in Priority Regions file	Check correct input for this variable
ERROR when running HD GWAS instructions	Python program can't run R scripts	Check if all R scripts are in the source file
Task is not specified	The analysis is not specified by the user	Repeat complete command, specifying the desired analysis
For wFDR analysis: There is no input for the weight column	Weight column was not specified and wFDR analysis can't be run	Add weight column

5. Tips

- Keep in mind your CORE DIRECTORY in which you download the HD-GWAS Package in order to avoid errors when running the demo.

- Try to avoid repetitive names for the output directory and the result files in order to avoid confusion.

- In case you find difficulties in adding the full path of your files, select each file and copy its name. Afterwards, paste the information in the command line and you will have the complete path of that file (this is possible only from the default terminal).

6. References

Collins, A. L., Kim, Y., Sklar, P., Consortium, I. S., O'Donovan, M. C., & Sullivan, P. F. (2012). Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results. Psychological Medicine , 42, 607–616.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J and others. 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337(6099):1190-5.

Roeder K, Bacanu SA, Wasserman L, Devlin B. 2006. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78(2):243-52.

Schork, N., Fallin, D., & Lanchbury, S. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. Clinical genetics , 58, 250-264.

Sun, L., Craiu, R. V., & Paterson, A. D. (2006). Stratified False Discovery Control for Large-Scale Hypothesis Testing with Application to Genome-Wide Association Studies. Genetic Epidemiology , 30, 519-530.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L and others. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42(Database issue):D1001-6.

Xing, C., Cohen, J. C., & Boer, E. (2010). A Weighted False Discovery Rate Control Procedure Reveals Alleles at FOXA2 that Influence Fasting Glucose Levels. The American Journal of Human Genetics , 86, 440-446.