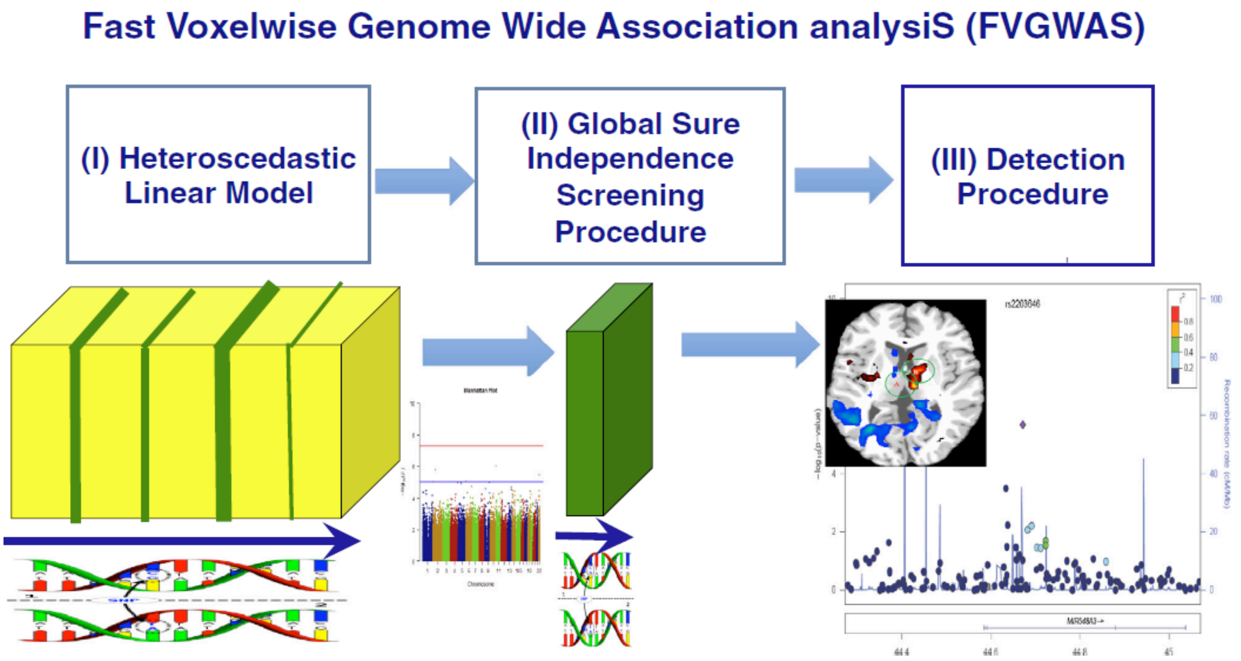
FVGWAS-3.0 Manual

Hongtu Zhu @ UNC BIAS
Chao Huang @ UNC BIAS
Nov 8, 2015

More and more large-scale imaging genetic studies are being widely conducted to collect a rich set of imaging, genetic, and clinical data to detect putative genes for complexly inherited neuropsychiatric and neurodegenerative disorders. Several major big-data challenges arise from testing genome-wide associations with signals at millions of locations in the brain from thousands of subjects. The aim of this software is to efficiently carry out whole-genome analyses of whole-brain data. FVGWAS consists of three components including a heteroscedastic linear model, a global sure independence screening (GSIS) procedure, and a detection procedure based on wild bootstrap methods. In the latest version, we implemented the option to “divide and conquer” large genotype datasets, which further improves the efficiency. Our FVGWAS may be a valuable statistical toolbox for large-scale imaging genetic analysis as the field is rapidly advancing with ultra-high-resolution imaging and whole-genome sequencing.

1. Schematic overview of FVGWAS



Fast Voxelwise Genome Wide Association analysiS (FVGWAS)

(I) Heteroscedastic Linear Model

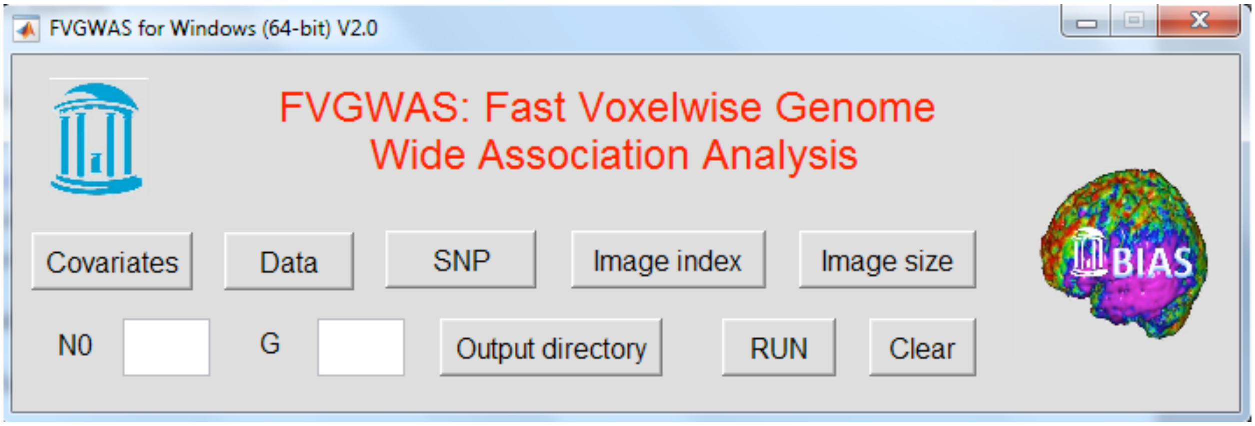
(II) Global Sure Independence Screening Procedure

(III) Detection Procedure

2. FVGWAS pipeline description

For FVGWAS, we provide two different version: interface and server-based pipeline. If you have a small data set, e.g., ROI data and SNP data from one chromosome, you can use the interface on your local desktop. If you have a large data set, e.g., whole brain imaging data and imputed SNP data, you can use the server-based pipeline which can also help you run all the tasks via parallel computing.

2.1. Interface-based FVGWAS


FVGWAS for Windows (64-bit) V2.0
FVGWAS: Fast Voxelwise Genome Wide Association Analysis
Covariates
Data
SNP
Image index
Image size
N0
G
Output directory
RUN
Clear
BIAS

Covariates

n*d design matrix. n is the sample size and d is the number of covariates

Data

Image data (n*V matrix). V is the number of voxels (whole brain analysis) or the number of ROIs (volumetric analysis)

SNP

Genetic data (n*C matrix). C is the number of SNPs

Image index

V*1 vector. The index of voxels within the brain mask (Applicable only for whole brain analysis)

Image size

The size of image data, e.g., 1*3 vector means 3-D image data (Applicable only for whole brain analysis)

N0

The number of top SNPs

G

The number of bootstrap samples

Output directory

Set up the directory, which is used to store all the results

RUN

Start running all the tasks

Clear

Clear all the input information on the interface

2.2 Server-based FVGWAS Pipeline

If you have a big data set and prefer to run FVGWAS on the workstations or servers. There are five MATLAB functions , which can help you accomplish all the tasks in FVGWAS.

A. FVGWAS_GSIS_server.m

This function is to run global sure independent screening (GSIS) procedure. After running this function, you will get the raw p values for all the loci with MAF no less than 0.05.

B. FVGWAS_Manhattanplot_server.m

This function is to summarize all the p values and SNP information into one file, which will be used as an input file in plotting Manhattan plot and QQ plot. After running this function, you will get a file including the p values, SNP names, chromosome IDs, and BP values for all loci.

C. FVGWAS_Bootstrap_server.m

This function is to run the bootstrapping in the detection procedure. After running this function, you will get the bootstrap test statistics, which will be used to approximate null distribution of the proposed test statistics. In this step, because the bootstrapping requires a lot of memory, we adopt to run part of the code via parallel computing. If you have multiple cores on the server, you can finish this step very efficiently. When running large datasets, it's possible to run out of memory. Please use "divide and conquer" algorithm if that happened.

D. FVGWAS_Detection_server.m

This function is used obtain adjusted p values for all loci with MAF no less than 0.05 in the detection procedure. After running this function, you will get the significant locus-voxel pairs and significant clusters for top SNPs found in GSIS step.

2.3. Server-based FVGWAS Pipeline with divide and conquer strategy

We implemented a "divide and conquer" strategy to split large dataset and perform the analysis more efficiently in terms of both time cost and memory cost. There are seven MATLAB functions and one bash script, which can help you accomplish the tasks in FVGWAS.

A. FVGWAS_Parameter_DVD.m

This function is to set all input information, including the name of image data file, covariate file, genotype file, image index file, image size file, output directory and all required parameters for FVGWAS.

B. FVGWAS_GSIS_server_DVD.m

This function is to run global sure independent screening (GSIS) procedure. Splitting option is available by specifying "snpc", the size of SNP subset. After running this function, you will get the p values for all the loci with MAF no less than 0.05.

C. FVGWAS_Manhattanplot_server_DVD.m

This function is to summarize all the p values and SNP information into one file, which will be used as an input file in plotting Manhattan plot and QQ plot. After running this function, you will get a file including the p values, SNP names, chromosome IDs, and BP values for all loci.

D. split.scp

This script is to split genotype dataset into several unjoint subsets of small size. The maximum number of snps in each subset is specified by parameter Bsize.

E. FVGWAS_Bootstrap_Top_server_DVD.m

This function is to run the bootstrapping on each split dataset in the detection procedure. After running this function, you will get bootstrap test statistics for SNP subsets. In this step, you are able to run for all subsets simultaneously with much less time & memory cost on each single core. If you have multiple cores on the server, you can finish this step very efficiently.

F. FVGWAS_Bootstrap_Cmb_server_DVD.m

This function is to combine bootstrapping results from in the detection procedure. After running this function, you will get the bootstrap test statistics for the whole dataset, which will be used to approximate null distribution of the proposed test statistics.

G. FVGWAS_Detection_server_DVD.m

This function is to run the bootstrapping in the detection procedure. After running this function, you will get the significant locus-voxel pairs and significant clusters for top SNPs found in GSIS step.

2.4. GWAS Results Display

This function (**FVGWAS_ManPlot.R**) is to plot Manhattan plot and QQ plot with R package (**qqman**). The file generated in function B is treated as the input file in this step.

3. Examples

We consider both ROI volumes and RAVENS maps to illustrate the wide applicability of FVGWAS. We carried out two different FVGWAS analyses: one is to use the volumes of 93 ROIs as multivariate phenotypic vectors and the other is to use RAVENS maps as whole-brain phenotypic vectors. In both analyses, we included an intercept, gender, age, whole brain volume, and the top 5 principal component scores in SNPs as the covariates. Then, we tested the additive effect of each of 501,584 SNPs on either 93 ROI volumes or RAVENS maps. For the volumetric analysis, we run all the tasks via the interface pipeline, while for the Ravens map data, we conduct the analysis via server-based pipeline.

3.1. GUI example: ROI volume analysis

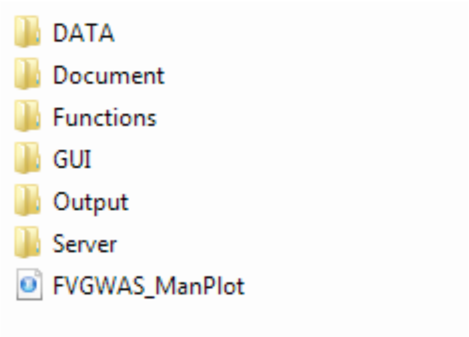
1. Download this package and unzip the file;

2. Start Matlab and set the following directory as the working folder

C:\Users\Administrator\Desktop\FVGWAS-3.0

Where C:\Users\Administrator\Desktop is the directory where you put the unzipped folder FVGWAS-3.0

3. In the folder FVGWAS-3.0, you can find six different folders and one R script.



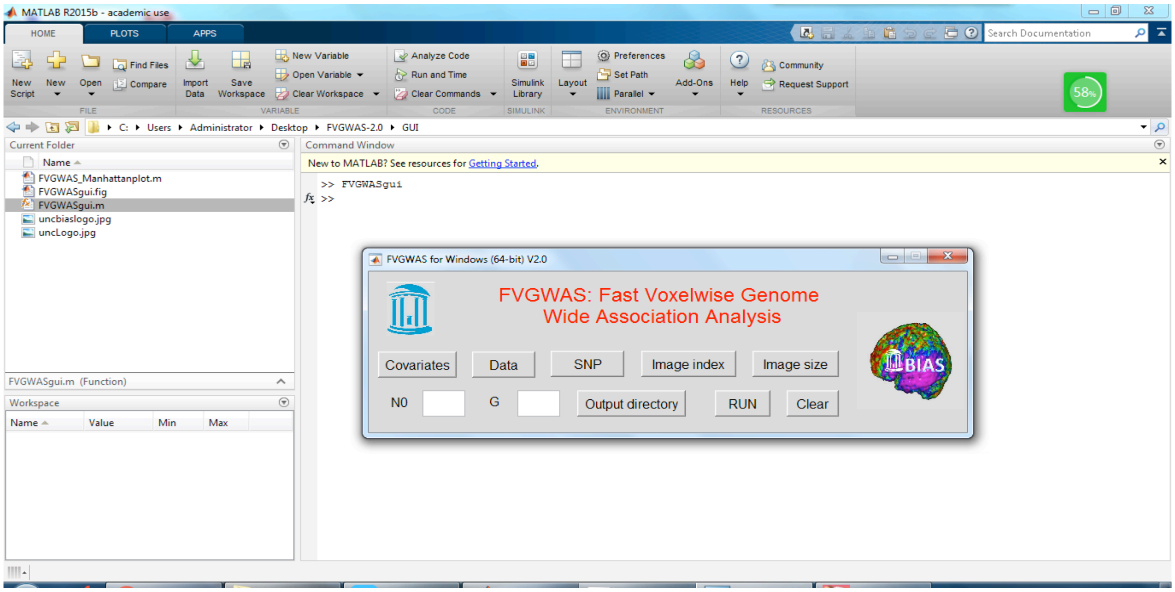
DATA
Document
Functions
GUI
Output
Server
FVGWAS_ManPlot

In the folder DATA, you can find the volumetric data set in the sub-folder volumetric. There are four files in that folder. Covariate_Data.mat includes all the demographic information; Image_Data.mat includes all the ROI thickness information; SNP_Data.txt includes the genetic data for all subjects; and SNP_info.map contains the SNP information including SNP names, chromosome IDs, and BP values for all loci.

All the codes are stored in the folder Functions.

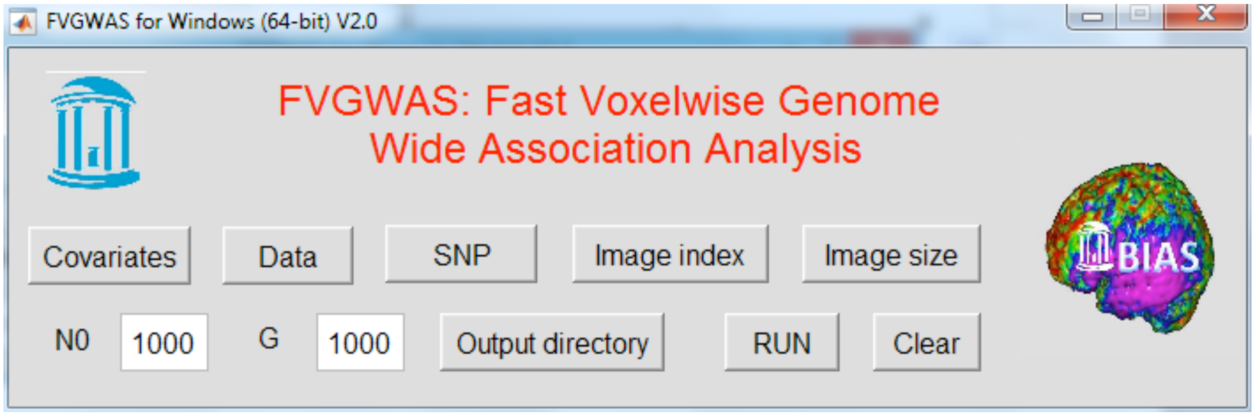
The interface can be found in the folder GUI.

4. Set GUI as the current folder and type FVGWASgui in command window to launch FVGWAS in MATLAB.



```
>> FVGWASgui
fx >>
```

5. Load the data into FVFWAS


FVGWAS for Windows (64-bit) V2.0
FVGWAS: Fast Voxelwise Genome Wide Association Analysis
Covariates
Data
SNP
Image index
Image size
Output directory
RUN
Clear
BIAS

N0 1000 G 1000

Click the button **Covariates**: load “Covariate_Data.mat”

Click the button **Data**: load “Image_Data.mat”

Click the button **SNP**: load “SNP_Data.txt”

Set **N_0**: 1000

Set **G**: 1000

Click the button **Output directory**: e.g. ./FVGWAS-3.0/Output/volumetric

Click the button **RUN**

6. Output files

“GSISresults.mat”

results of GSIS step, variable “pp” contains the -log_10{p-values} of all SNP data, and “SNP_index” contains all the SNP ID which are removed in preprocessing.

“VoxelclusterandSNP.mat”

variable “rawpvalue” is a C*N0 matrix including the raw p-values of top N0 SNPs.

variable “pv” is a C*N0 matrix, including the corrected p-values.

7. Results display

(1) Run “FVGWAS_Manhattanplot.m” in MatLab to generate the GWAS results. Output file: “SNP_GSIS.txt”. One thing should be noted that, in FVGWAS_Manhattanplot.m, there are two parameters should be modified according to your own settings.

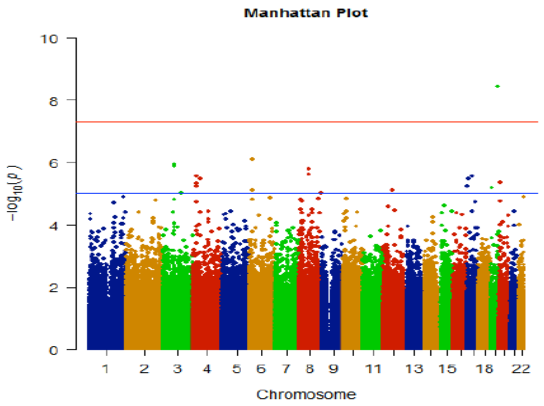
mode='volumetric'; volumetric is the name of the folder where the file SNP_GSIS.txt is saved.

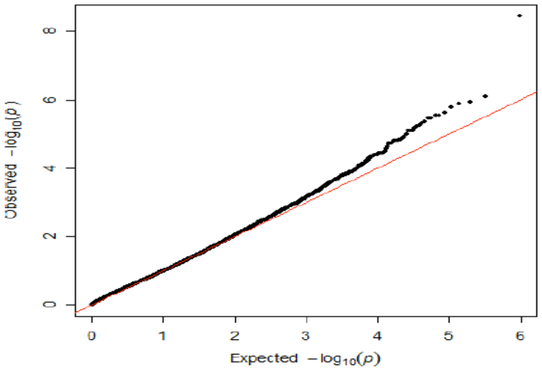
Flag=0; where 0 for ROI-based imaging data; 1 for voxel-based imaging data

(2) Run “FVGWAS_ManPlot.R” in R to plot the Manhattan plot and QQ plot. One thing should be noted that, in FVGWAS_ManPlot.R, there are two parameters should be modified according to your own settings.

statName="volumetric"; volumetric is the name of the folder where the file SNP_GSIS.txt is saved.

HomeDir="C:/Users/Administrator/Desktop/FVGWAS-2.0"; which is the directory the folder FVGWAS-2.0 is located.





3.2. Server based example: Voxel-wise Analysis

In the folder DATA, you can find the RavenMap data set in the sub-folder RavenMap. There are six files in that folder. Covariate_Data.mat includes all the demographic information; Image_Data.mat includes all the ROI thickness information; ImageIndex_Data.mat is a V*1 vector, which is the index of voxels within the brain mask; ImageSize_Data.mat is the size of image data, e.g., 1*3 vector means 3-D image data; SNP_Data.txt includes the genetic data for all subjects; and SNP_info.map contains the SNP information including SNP names, chromosome IDs, and BP values for all loci.All the codes are stored in the folder Functions.

The server-based main functions can be found in the folder Server.

1. Set Server as the current folder and run FVGWAS_GSIS_server.m;

2. After Step 1 is finished, run FVGWAS_Bootstrap_server.m

3. After Step 2 is finished, run FVGWAS_Detection_server.m;

4. Result Display

(1) Run “FVGWAS_Manhattanplot_server.m” in MatLab to generate the GWAS results. Output file: “SNP_GSIS.txt”. One thing should be noted that, in FVGWAS_Manhattanplot_server.m, there are two parameters should be modified according to your own settings.

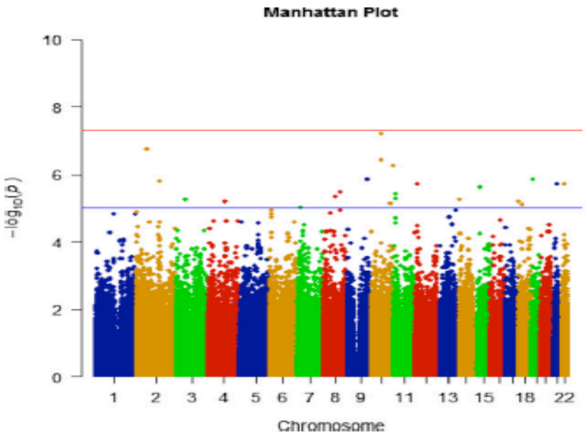
mode='RavenMap'; RavenMap is the name of the folder where the file SNP_GSIS.txt is saved.

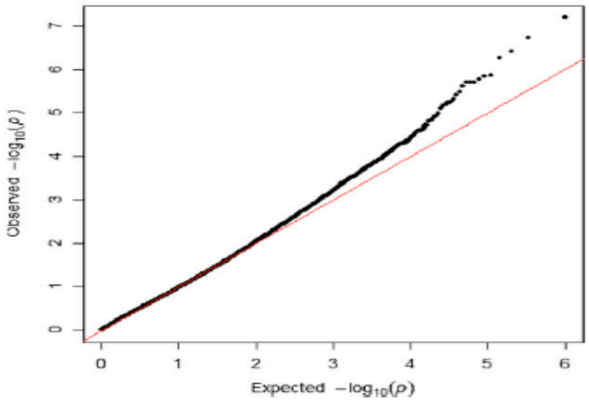
Flag=1; where 0 for ROI-based imaging data; 1 for voxel-based imaging data

(2) Run “FVGWAS_ManPlot.R” in R to plot the Manhattan plot and QQ plot. One thing should be noted that, in FVGWAS_ManPlot.R, there are two parameters should be modified according to your own settings.

statName="RavenMap"; RavenMap is the name of the folder where the file SNP_GSIS.txt is saved.

HomeDir="C:/Users/Administrator/Desktop/FVGWAS-3.0"; which is the directory the folder FVGWAS-3.0 is located.





3.3. Example using “divide and conquer” strategy: Voxel-wise Analysis

In the folder DATA, you can find the RavenMap data set in the sub-folder RavenMap. There are six files in that folder. Covariate_Data.mat includes all the demographic information; Image_Data.mat includes all the ROI thickness information; ImageIndex_Data.mat is a V*1 vector, which is the index of voxels within the brain mask; ImageSize_Data.mat is the size of image data, e.g., 1*3 vector means 3-D image data; SNP_Data.txt includes the genetic data for all subjects; and SNP_info.map contains the SNP information including SNP names, chromosome IDs, and BP values for all loci. All the codes are stored in the folder Functions.

The server-based main functions can be found in the folder DVDConquer.

1. Set DVDConquer as the current folder and run FVGWAS_Parameter_DVD.m

2. Run FVGWAS_GSIS_server_DVD.m

3. After Step 2 is finished, run script split.scp. Parameter “Bsize”, the maximum size of each genotype subset, can be changed according to user need.

4. After Step 3 is finished, run FVGWAS_Bootstrap_Top_server_DVD.m with different subset index. For example, if there are 10 genotype subsets, the user can simultaneously run FVGWAS_Bootstrap_Top_server_DVD(1), FVGWAS_Bootstrap_Top_server_DVD(2), … , FVGWAS_Bootstrap_Top_server_DVD (10)

5. After all jobs in Step 4 are finished, run FVGWAS_Bootstrap_Cmb_server_DVD.m

6. After Step 5 is finished, run FVGWAS_Detection_server_DVD.m

7. Result Display

(1) Run “FVGWAS_Manhattanplot_server_DVD.m” in MatLab to generate the GWAS results. Output file: “SNP_GSIS.txt”. One thing should be noted that, in FVGWAS_Manhattanplot_server_DVD.m, there are two parameters should be modified according to your own settings.

mode='RavenMap'; RavenMap is the name of the folder where the file SNP_GSIS.txt is saved.

Flag=1; where 0 for ROI-based imaging data; 1 for voxel-based imaging data

(2) Run “FVGWAS_ManPlot.R” in R to plot the Manhattan plot and QQ plot. One thing should be noted that, in FVGWAS_ManPlot.R, there are two parameters should be modified according to your own settings.

statName="RavenMap"; RavenMap is the name of the folder where the file SNP_GSIS.txt is saved.

HomeDir="C:/Users/Administrator/Desktop/FVGWAS-3.0"; which is the directory the folder FVGWAS-3.0 is located.

4. PS

1. If you want to run volumetric analysis via server-based pipeline. You need to change two parameters in the following four main functions in folder Server

FVGWAS_GSIS_server.m

FVGWAS_Bootstrap_server.m

FVGWAS_Detection_server.m

mode='RavenMap'; should be changed to mode='volumetric';

Flag=1; should be changed to Flag=0;

2. Divide and conquer based pipeline (**Example 3.3**) is summarized by bash script subjob.scp, which can be used directly in linux. You should run only one step each time, following the order in the script and change parameter settings according to the instruction. Also, you might also need to change command to submit matlab jobs based on the server setting.

3. If you want to run volumetric analysis via divide-and-conquer pipeline. You need to change two parameters in split.scp and FVGWAS_Parameter_DVD.m in folder DVDConquer.

mode='RavenMap'; should be changed to mode='volumetric';

Flag=1; should be changed to Flag=0;

If you are using bash script subjob.scp,you should also change mode='RavenMap' to mode='volumetric';

4. Please remain the name of each input file as the one shown in the folder DATA. Otherwise, you also need to modify them in the functions/scripts shown above.

5. If you have a new data set. You can put your data folder into the folder DATA, modify the value of 'mode' in the four functions above, e.g., if the name of the new data folder is "Newdata", you need to change the value of "mode" to mode='Newdata'. Also, if the new data set is voxel-type imaging data, you need to set Flag=1, otherwise, Flag=0.