GWA tutorial

This GitHub repository provides several tutorials about techniques used to analyze genetic data. Underneath this README we have provided a step-by-step guide to help researchers without experience in Unix to complete these tutorials succesfully. For reseachers familiar with Unix this README will probably be sufficient.

We have made scripts available for:

- 1. All essential GWAS QC steps along with scripts for data visualization.
- 2. Dealing with population stratification, using 1000 genomes as a reference.
- 3. Association analyses of GWAS data.
- 4. Polygenic risk score (PRS) analyses.

The scripts downloadable from this GitHub page can be seen purely as tutorials and used for educational purposes, but can also be used as a template for analyzing your own data. All scripts/tutorials from this GitHub page use freely downloadable data, commands to download the necessary data can be found in the scripts.

Content:

- 1_QC_GWAS.zip
- 2_Population_stratification.zip
- 3_Association_GWAS
- 4_ PRS.doc

How to use the tutorials on this page: The tutorials are designed to run on an UNIX/Linux computer/server. The first 3 tutorials contain both *.text as *.R scripts. The main scripts for performing these tutorials are the *.text scripts (respectively for the first 3 tutorials: 1_Main_script_QC_GWAS, 2_Main_script_MDS.txt, and 3_Main_script_association_GWAS.txt). These script will execute the *.R scripts, when those are placed in the same directory. Note, without placing all files belonging to a specific tutorial in the same directory the tutorials cannot be completed. Furthermore, the first 3 tutorials are not independent; they should be followed in the order given above, according to their number. For example, the files generated at the end of tutorial 1 are essential in performing tutorial 2. Therefore, those files should be moved/copied to the directory in which tutorial 2 is executed. In addition, the files from tutorial 2 are essential for tutorial 3. The fourth tutorial (4_ PRS.doc) is a MS Word document, and runs independently of the previous 3 tutorials.

All scripts are developed for UNIX/Linux computer resources, and all commands should be typed/pasted at the shell prompt.

Note: The *.zip files contain multiple files, in order to successfully complete the tutorials it is essential to download all files from the *.zip files and upload them to your working directory. To pull all tutorials to your computer simply use the following command: git clone https://github.com/MareesAT/GWA_tutorial.git . Alternatively, you can manually open the *.zip folders and PRS.doc file by clicking on the folder/file followed by clicking on "View Raw".

Contact:

Please email Andries Marees (a.t.marees@vu.nl) for questions

Additional material

Once you completed the current tutorial we recommend you to visit https://github.com/AngelaMinaVargas/eMAGMA-tutorial This Github repository guides the steps to use eMAGMA.
eMAGMA is a post-GWAS analysis, that conducts eQTL informed gene-based tests by assigning SNPs to tissue-specific eGenes.

Step-by-step-guide for this tutorial

Step-by-step-guide for researches new to Unix and/or genetic analyses.

Introduction

The tutorial consist of four separate parts. The first three are dependent of each other and can only be performed in consecutive order, starting from the first (1_QC_GWAS.zip), then the second (2_Population_stratification.zip, followed by the third (3_Association_GWAS). The fourth part (4_ PRS.doc) can be performed independently.

The Unix commands provided in this guide should be typed/copy-and-pasted after the prompt ($ or >) on your Unix machine. Note, the ">" in front of the commands should not be copy-and-pasted. Only what comes after the ">".

We assume that you have read the accompanying article "A tutorial on conducting Genome-Wide-Association Studies: Quality control and statistical analysis " (https://www.ncbi.nlm.nih.gov/pubmed/29484742), which should provide you with a basic theoretical understanding of the type of analyses covered in this tutorial.

This step-by-step guide serves researchers who have none or very little experience with Unix, by helping them through the Unix commands in preparation of the tutorial.

Preparation

Step 1) The current set of tutorials on this GitHub page are based on a GNU/Linux-based computer, therefore:

- Make sure you have access to a GNU/Linux-based computer resource.
- Create a directory where you plan to conduct the analysis.

Execute the command below (copy-and-paste without the prompt: > and without the {}).

```
mkdir {name_for_your_directory}
```

Step 2) Download the files from the GitHub page

- Change the directory of your Unix machine to the created directory from step 1.

Execute the command below

```
cd HOME/{user}/{path/name_for_your_directory}
git clone https://github.com/MareesAT/GWA_tutorial.git
```

- Unzip the folder of the first tutorial and move into the newly created directory.

Execute the commands below

```
unzip 1_QC_GWAS.zip cd 1_QC_GWAS
```

Step 3) This tutorial requires the open-source programming language R and the open-source whole genome association analysis toolset PLINK version 1.07 (all commands also work with PLINK2). If these programs are not already installed on your computer they can be downloaded from respectively: https://www.r-project.org/ http://zzz.bwh.harvard.edu/plink/ https://www.cog-genomics.org/plink2

- We recommend using the newest versions. These websites will guide you through the installation process.
- Congratulations everything is set up to start the tutorial!

Execution of tutorial 1

Step 4) Once you've created a directory in which you have downloaded and unzipped the folder: 1_QC_GWAS.zip, you are ready to start the first part of the actual tutorial. All steps of this tutorial will be excecuted using the commands from the main script: 1_Main_script_QC_GWAS.txt, the only thing necessary in completing the tutorial is copy-and-paste the commands from the main script at the prompt of your Unix device. Note, make sure you are in the directory containing all files, which is the directory after the last command of step 2. There is no need to open the other files manually.

There are two ways to use the main script:

Option 1

- If you are a novice user, we recommend opening 1_Main_script_QC_GWAS.txt in WordPad or Notepad on your Windows computer.

Option 2

- Alternatively, 1_Main_script_QC_GWAS.txt can be opened using an Unix text editor, for example vi.

Open the main script with vi :

```
vi 1_Main_script_QC_GWAS.txt
```

- This enables you to read the script within the Unix environment and copy the command lines from it.

To exit vi and return to your directory use:

```
:q
```

- From there, using either option 1 or 2, you can read the information given at every step of script “, 1_Main_script_QC_GWAS.txt” and copy-paste the commands after the prompt on your Unix machine.

Note, if R or PLINK are installed in a directory other than your working directory please specify the path to the excecutables in the given script. Alternatively, you can copy the executables of the programs to your working directory. For example, by using: cp {path/program name} {path/directiory}. However, when using a cluster computer, commands such a "module load plink", and "module load R" will suffice, regardless of directory.

For more information of using R and PLINK in a Unix/Linux environment we refer to: http://zzz.bwh.harvard.edu/plink/download.shtml#nixs

Execution of tutorial 2&3

- Unzip the tutorial folder of choice as described in step 2.
- Use the output file from the last tutorial as input for the tutorial you want to start.

The command below can be used to copy the file to another directory

```
cp {path/directory/file} {path/directory}
```

- Use 2_Main_script_MDS.txt for the second tutorial and 3_Main_script_association_GWAS.txt for the third tutorial.

Execution of tutorial 4

4_ PRS.doc works independently from the other tutorials. After downloading 4_ PRS.doc, you can run the script, without the need for unzipping, in a directory of choice.