**Practical assignment – Module 1 -Session 1**

**Module topic: GWAS Recap and PostGWAS Overview**

**Contact session title: Brief overview of GWAS and GWAS workflow**

**Trainer: Shaun Aron and Segun Fatumo**

**Participant:** Adeyemo Oluwatosin Maryam

**Date:** 20th Feb. 2023

**The H3ABioNet GWAS Workflow**

**Introduction**

The analysis of millions of Single Nucleotide Variants is a complex process which relies heavily on extracting statistically significant information from large datasets. The analysis process can be broken down into a number of steps which can be executed sequentially. Some steps will need to be run multiple times to optimise the parameters and thresholds used based on the data. The first step involves the conversion of the raw intensity data files into a more manageable data format understood by a program called Plink. Illumina produces files with SNPs generated on the TOP/BOTTOM or Forward strand. A program called *call2plink* can be used on these raw files to create the correctly formatted Plink files. This step can be complicated and relies on the availability of an additional good quality intensity dataset that can be used as a reference to improve the accuracy of variants called in the current dataset, as well as, additional files containing information about the current SNP array you are using. We will not be running this part of the workflow in this practical but you are free to read about it further on the Github page noted below. The second and third steps that we will be running in this practical include the quality control and association analysis of an example dataset. Some additional postGWAS analysis methods are also available through the workflow. The workflows have been built using a workflow language called Nextflow <https://www.nextflow.io/> . The use of Nextflow together with software like Docker <https://www.docker.com/> or Singularity <https://docs.sylabs.io/guides/latest/user-guide/> allows for all the tools and dependencies for the analysis to be packaged and available for analysis locally on your computer or a compute cluster. The workflow has been developed to run on any Unix-based command line system

**Tools used in this session**

During this session you will be required to follow the instructions on the Github page for the workflow in order to run the GWAS QC and Association parts of the workflow on an example dataset.

Main H3ABioNet GitHub page: <https://github.com/h3abionet/h3agwas>

Tutorial Page: <https://github.com/h3abionet/h3agwas-examples>

**Please note**

* **Hand-in information** please upload your completed assignment to the Vula ‘Assignments’ tab. Take note of the final hand-in date for each assignment, which will be indicated on Vula.

**Task 1: Getting started with the workflow**

**Task 1**: In order to get started with the workflow please navigate to the Workflow tutorial page: <https://github.com/h3abionet/h3agwas-examples> . This page provides you with step-by-step instructions on how run the workflow on the example dataset. To run the workflow we will first need to install Nextflow. This can be done following the steps below:

It only needs two easy steps:

1. Download the executable package by copying and pasting either one of the following commands in your terminal window:

wget -qO- https://get.nextflow.io | bash

Or, if you prefer *curl*:

curl -s https://get.nextflow.io | bash

This will create the nextflow main executable file in the current directory.

1. Make the binary executable on your system by running

chmod +x nextflow

1. Optionally, move the nextflow file to a directory accessible by your $PATH variable (this is only required to avoid remembering and typing the full path to nextflow each time you need to run it).

mv nextflow /home/username/bin \*username is your username

Nextflow requires Java 11 up to 18 to run. Ensure that your system has this version of Java installed. Please contact your system administrator if you are having issues with this. On Ubuntu a simple command can be used to install version 11 of Java

sudo apt-get install openjdk-11-jdk

We will also require Singularity to run the association analysis. General instructions for installing Singularity can be found here:

<https://singularity-tutorial.github.io/01-installation/>

Please contact your system administrator should you have troubles with installing Singularity.

There are two ways main ways to run the workflow. You can pull and run the workflow directly using Nextflow or you can use Git to clone the workflow and then run it from the cloned directory. Using the first approach always ensures that you are using the most updated version of the workflow. Let us try the first way. Navigate to a new directory that you want to work in and run the Nextflow command.

cd gwas\_analysis

nextflow pull h3abionet/h3agwas

This command should take some time when run for the first time to update Nextflow and the workflow. It will give you an error if the incorrect version of Java is installed. Let us rather manage it with Git as we will need to copy over the files for analysis as well.

Navigate to where you want to install the workflow and run the git clone command. This command will take some time as all the files in the Github repository will be copied to your computer.

git clone <https://github.com/h3abionet/h3agwas.git>

[shaun@n32 test]$ git clone https://github.com/h3abionet/h3agwas.git

Cloning into 'h3agwas'...

remote: Enumerating objects: 6403, done.

remote: Counting objects: 100% (745/745), done.

remote: Compressing objects: 100% (295/295), done.

remote: Total 6403 (delta 465), reused 717 (delta 446), pack-reused 5658

Receiving objects: 100% (6403/6403), 31.67 MiB | 7.62 MiB/s, done.

Resolving deltas: 100% (4200/4200), done.

You should now have directory call h3gwas in your folder which contains all necessary scripts to run the analysis. If you cd into the directory and type ls you should be able to see several directories and files. The example data is stored in the sample directory. If you cd into the example data directory and list (ls) the files you should see several additional files. The .bim, .fam, .bed and .phe . Have a look at what each file contains for the files that are human readable.

**Task 1:** What information does each of these files contain?

**The .bim file** contains information about the location of SNPs and it has 6 columns.

The first column is the chromosome number where the SNP is located.

The second column is the SNP name or identifier.

The third column is the position of the SNP in Morgans or centimorgans

The fourth column is the base pair coordinate of the SNP

The fifth and sixth columns are the alleles of the SNP. The fifth is the minor allele while the sixth is the major allele.

**The .fam file** is also readable and it has 6 columns.

The first column is the family ID

The second column is the individual ID

The third and fourth column is the ID of the parent genotype

The fifth column is the sex information (1=male, 2=female, 0=unknown)

The sixth column is the phenotype

**The .bed is unreadable**

**The .phe file contains similar information with the .fam file. Although, .fam file does not have a header, the .phe has a header which are FID, IID, PAT, MAT, SEX, PHE**

**Task 2:** Running the Quality Control component of the workflow

**Task 2:** Instructions

You will now run the qc workflow on the example dataset. This workflow calculates and plots a number of metrics that can be used to assess the quality of both SNPs and individuals in the dataset. If you are running this on a local machine then the following simple command should work provided you are in the uppermost directory of the workflow

nextflow run qc -profile singularity

This should allow the workflow to start running providing you with a list of analyses as they are completed

lN E X T F L O W ~ version 22.10.4

Launching `qc/main.nf` [backstabbing\_agnesi] DSL1 - revision: 7d401873bd

The batch file is 0

Sexinfo available command

WARN: The `echo` directive has been deprecated - use to `debug` instead

0

[- ] process > inMD5 [ 0%] 0 of 1

[- ] process > noSampleSheet [ 0%] 0 of 1

[- ] process > getDuplicateMarkers [ 0%] 0 of 1

[- ] process > removeDuplicateSNPs -

executor > slurm (2)

[- ] process > inMD5 [ 0%] 0 of 1

[28/199092] process > noSampleSheet [ 0%] 0 of 1

[1a/837c4f] process > getDuplicateMarkers (1) [ 0%] 0 of 1

[- ] process > removeDuplicateSNPs -

If you are running the workflow on a compute cluster, you will need to tell Nextflow which queueing system and containerisation software you are using. This can be done by adding the -profile option to the Nextflow command. Systems supported are listed on the GitHub page and include Torque/PBS, SLURM, Docker, DockerSWARM and Singularity. An example command using Slurm is shown below.

nextflow run qc -profile slurm

nextflow run qc -profile singularity,slurm

Results from the workflow run will be saved to the output directory. In this case, all results are collated and presented in a final output file called out.pdf. Individual raw output files and figures can be found in the other folders in the output directory. If you are working on a cluster without a GUI, copy the out.pdf file back to our local computer and open it with a pdf viewer.

**Task 2:** Browse through the report and take note of the various quality control metrics that have been generated. Try to understand how each one can be used to remove bad quality SNPs and individuals. Mention below two metrics that you would use to remove SNPs of low quality.

**Task 2:** Answer

1. **SNPs call rate**
2. **Hardy Weinberg Equilibrium**

**Task 3:** Examining and changing the parameters for the analysis.

All datasets are not the same and therefore the thresholds used to quality control the data will depend on the dataset. The graphs generated in the report can be used to assess what thresholds to use and how much of your data you will lose at specific thresholds. These thresholds are specified in a file that the workflow reads by default when it is run .The file is called the nextflow.config file and is saved in the same directory as the main.nf Nextflow file in the qc directory. If you open this file you will see a number of variables, paths and parameters set – these are used by default when the workflow is run. To change these default values, you can pass them on the command line like -profile slurm or create a new config file with the values you would like to change written in that file. Part of nextflow.config file shown below.

// Directories

work\_dir = "/$PWD"

input\_dir = "sample"

// Can use S3 too

//input\_dir = "s3://h3abionet/sample"

input\_pat = "sampleA"

output\_dir = "${params.work\_dir}/output"

scripts = "${params.work\_dir}/scripts"

output = "out"

max\_forks = 95

high\_ld\_regions\_fname = ""

sexinfo\_available = true

cut\_het\_high = 0.343

cut\_het\_low = 0.15

cut\_diff\_miss = "0.05"

cut\_maf = "0.01"

cut\_mind = "0.02"

cut\_geno = 0.01

cut\_hwe = 0.008

pi\_hat = 0.11

Command to run qc workflow with config file containing different cut-off thresholds

nextflow run qc -c new.config -profile singularity,slurm

**Task 3:** How many times do you think you will need to run the quality control workflow? Do you think that default values for these thresholds should be used across different datasets from different arrays?

**There is no specific amount of times needed to run the quality control workflow. It should be carried out until when the dataset is certified to be of high quality for subsequent analyses.**

**I think that the parameters and threshold for each dataset is different.**

You will be running the association analysis on a new dataset. To copy the datasets, make sure you are in the h3agwas directory and run the command below:

git clone https://github.com/h3abionet/h3agwas-examples.git

**Task 4:** You should now have a directory called h3agwas-examples in your current directory (use the “ls” command to check). Below is an example of running the qc analysis on a dataset found in data/array\_plk/array and specifying the parameters via the command line instead of providing a modified config file.

nextflow run h3abionet/h3agwas/qc/main.nf --input\_dir data/array\_plk --input\_pat array --output\_dir qc --output array\_qc \

--phenotype data/pheno/pheno\_test.all --pheno\_col phenoqc\_ql \

--case\_control data/pheno/pheno\_test.all --case\_control\_col Sex \

--batch data/pheno/pheno\_test.all --batch\_col batch \

-profile singularity

Try running the command above and see if you can get it to work. Be careful with copying and pasting this into the terminal as some characters may change. i.e. the -- and - . The command should be on a single line indicated by the “\” characters at the end of each line. When running singularity for the first time, Nextflow will pull the necessary containers for the software required to run the workflow. This will take some time for the first run.

**Task 4:** What do each of the parameters in the command above refer to:

--input\_dir: This specifies the input directory where the input data for the workflow is located.

--input \_pat: This specifies a pattern for identifying the input data files within the input directory

--output\_dir: This specifies the output directory where the results of the workflow will be written.

--output: This specifies the name of the output file(s) generated by the workflow

--case\_control: This indicates the case-control information of the input data.

--case\_control-col: This specifies the column name that contains the case-control information.

--batch: This indicates the batch information of the workflow to perform batch correction.

--batch\_col: This specifies the column name that contains the batch information.

--profile: This specifies the environment for the workflow

--resume: This indicates that the workflow should be resumed from a previous run, if one exists.

Extract of workflow output

N E X T F L O W ~ version 22.10.4

NOTE: Your local project version looks outdated - a different revision is available in the remote repository [f27c45d7f6]

Launching `https://github.com/h3abionet/h3agwas` [happy\_descartes] DSL1 - revision: e2b662b184 [master]

The batch file is data/pheno/pheno\_test.all

Sexinfo available command

WARN: The `echo` directive has been deprecated - use to `debug` instead

0

executor > slurm (2)

[b9/dcebda] process > inMD5 (1) [ 0%] 0 of 1

[cc/2c9b0d] process > noSampleSheet [ 0%] 0 of 1

[- ] process > getDuplicateMarkers [ 0%] 0 of 1

[- ] process > removeDuplicateSNPs -

[- ] process > getX -

[- ] process > analyseX

**Task 5:** Running a simple association analysis

Once you are happy with the quality control of the data, you can use the next workflow to run a simple association analysis on your data. Depending on the type of data you have, you will run a simple linear or logistic regression or more complex mixed model approaches. There are several types you can run using the workflow and you can read more about the different types here: <https://github.com/h3abionet/h3agwas/tree/master/assoc>

Workflow needs at least the PLINK files (see --input\_dir and --input\_pat), phenotype file (see --data) and 1 or more phenotypes (--phenotype)

Here's a simple example, assuming you've done the QC example above and haven't changed working directory. We take the QCed data as genotype input, and the data in data/pheno/pheno\_test.all as the source of the phenotype data. There are two columns in the phenotype file (pheno\_qt1, pheno\_qt2) and we'll test for associations against both. We specify which directory and base file name to use for output. We do linear regression. We use singularity.

**You can now try and run the association test below as one single command.**

nextflow run h3abionet/h3agwas/assoc/main.nf --input\_dir qc/ --input\_pat array\_qc --data data/pheno/pheno\_test.all --pheno pheno\_qt1,pheno\_qt2 --output\_dir assoc1 --output qt12 --linear 1 -profile singularity

This will run a simple association test based on the two phenotypes in the phenotype data file using the array\_qc plink files. The output files will be saved to the assoc1 directory and a pdf report is generated with a summary of the results. The pdf report can open using a pdf viewer. The workflow can be used to run number of additional complex types of association analyses using various input file formats. See the GitHub page for more options.

**Task 5**: Have a look at the report and state briefly what main information is provided.

**Task 5**: Answer

**The Q-Q plot shows that there is significant SNPs in the first and second phenotypes. The Manhattan plot indicates that the significant SNPs can be found in chromosome 1 and 10**