The GRC Whole Exome Sequencing Annotation and Report Generation Pipeline (WES ARGP): A Fast Start Guide

Miles Benton

18 July, 2016

(This is very much still a work in progress! Feedback and suggestions are most welcome.)

This is a quick start guide to get up and running with the GRC whole exome sequencing (WES) pipeline.

UPDATE [18-07-2016]: the whole process of VCF annotation, filtering and report generation has been overhauled and automated. All that is required from users is an initial provision of:

- sampleID (i.e. DG1051, DG934, etc.)
- runID (this is the run number generated by the Proton)
- genome build (i.e. hg19, hg38, . . .)
- name of the Tier0 gene list (commonly `gene_lists/diagnostic_panel.txt`)
- name of the Tier1 gene list (commonly `gene_lists/test_genes_updated.txt`)
- a unique directory name (an example might be `sampleID_date`, i.e. `DG1051_160718`)

Before running the pipeline

Access taurus

The GRC **W**hole **E**xome **S**equencing **A**nnotation and **R**eport **G**eneration **P**ipeline (**WES ARGP**) is located, maintained and run from the work-station server taurus. As this is a headless server, users will have to remote in to run the pipeline.

Remote Access

If you are working from a PC located on campus and connected to the QUT network you won't have to worry about the following.

Make sure you are logged into the QUT network if you are working remotely. If you are having trouble with this try looking at the following QUT IT resources:

- https://secure.qut.edu.au/itservices/qut/qutservices/qutnetwork/qutsas/
- https://ithelp.qut.edu.au/portal/app/portlets/results/viewsolution.jsp?solutionid=041400820540658
- https://sas.qut.edu.au/

taurus (Linux server) details

You will need these details to log into taurus.

Username: `grcnata`
Password: `Choo2geez=ai0g`
taurus address/host name: `taurus.ihbi.qut.edu.au`
Port: 22 (you may or may not need this, port 22 is default in most applications)

Linux

Use ssh:

```
ssh grcnata@taurus.ihbi.qut.edu.au
```

You will be asked to enter the password provided above.

Windows (not tested)

Putty: https://www.putty.org/

WinSCP: https://winscp.net/

FileZilla: https://filezilla-project.org/

MacOS (not tested)

Use ssh: (MacOS is based on UNIX, so `ssh` should work)

```
ssh grcnata@taurus.ihbi.qut.edu.au
```

You will be asked to enter the password provided above.

Putty: https://www.putty.org/

FileZilla: https://filezilla-project.org/

Setting up and running the WES ARGP

The pipeline is designed to run on Linux machines, the following steps are run at the command line (bash).

Transfer of Exome data to taurus

NOTE: this step will most likely need to be automated in some form. How are people currently transferring the VCF and quality files to the `grcnata/raw_wes_files` at the moment?

Running the WES ARGP script

All that is required to initiate the pipeline on a given sample is following these steps:

1. run the pipeline script by typing in the following:

```
./WESdiag_pipeline.sh
```

2. provide the sampleID followed by `Enter`
3. provide the runID followed by `Enter`
4. provide the genome build followed by `Enter`
5. provide the Tier0 gene list followed by `Enter`
6. provide the Tier1 gene list followed by `Enter`
7. provide a unique directory name (an example might be `sampleID_date`, i.e. `DG1051_160718`) followed by `Enter`
8. confirm the details are correct:
    - if correct, select `1` followed by `Enter` to proceed.
    - If the details are incorrect select `2` followed by `Enter` to exit the script.

A note on run time

The whole process from start to finish will usually take ~**20 minutes**. This can vary depending on the size and amount of variant data, but should never be more than **30 minutes**.

Download the report

Initially the easiest way for users to download the word document report might be by using FileZila as it provides a graphical user interface. This looks like a file explorer/directory structure interface which can be easily navigated and files can downloaded to the users local machine.

checklist / to-do list

- ~~further automate the pipeline~~
- ~~add an option to define directory name separately~~
  - ~~this will allow multiple runs on the same sample and won't create issues with same name directories~~
- explore data transfer automation (for vcf and quality files, as well as the final report and tables)
  - this may require adding a step to compress the final report directory and transfer it to both the diagnostics team as well as backup/long-term storage.
  - additionally it would be nice to be able to transfer data from Proton to taurus using the information provided by the user at the start of the pipeline, i.e. user provides sampleID and runID and this is wrapped in a script to locate the required files on the Proton server (or taurus) and copy them to the `grcnata/wes_raw_data` directory.