Detailed ALFA AF Query Instructions

06/10/2021

**Purpose:**

The purpose of this tool is to extract allele frequencies (AF) and other genotype information from genotype database (ALFA) to generate a flat file of variants of interest and their genotype information.

**Overview:**

The allele frequency query tool was developed to pull AF of specific variants of interest from different genotype databases. This tool eliminates the need to manually look up the allele frequency and genotype details for a long list of SNPs.

The tool uses python to pull ALFA database and bash script with BCFtools to format and pull info from ALFA. This can all be completed using O2.

**Setting up your computer to run the AF query tool:**

1. Create an O2 account with your HMS ID (**NOT** Harvard ID or Harvard Key. This username is lowercase initials, example is gao4). To set up, email [rchelp@hms.harvard.edu](mailto:rchelp@hms.harvard.edu) or sign into HMS and fill out online request form for O2 account.
   1. Harvard’s O2 wiki page will help you navigate O2: <https://harvardmed.atlassian.net/wiki/spaces/O2>
2. Download the following programs onto your desktop:
   1. MobaXterm: <https://mobaxterm.mobatek.net/>
      1. This is where you access your O2 directory and will perform the AF query.
   2. FileZilla (Client): <https://filezilla-project.org/>
      1. This is used for moving files from O2 to your desktop and vice versa. You will use this to initially put the AF query scripts into your O2 directory and to transfer the output files onto your desktop.
3. Log into O2 account in MobaXterm and FileZilla. Both will require DUO authentication.

Host: transfer.rc.hms.harvard.edu

Port: 22

Username: HMS ID (ID you use to login to O2, NOT Harvard ID or Harvard Key. This username is lowercase initials, example is gao4)

Password: HMS password (password you use when logging into O2)

1. Download “AF Query” folder from the Merge Tool git repository (https://github.com/YehGenPedBCH/AF-Query ). There are several ways to get these scripts into your O2 directory. Since there are only 3 files you need on your O2 account, I recommend downloading those onto your computer and then using FileZilla to transfer them to your O2 account.
   1. You will only have to do this **ONCE**. The scripts needed are: ALFA\_pull.py, ALFA\_pull.sh, AF\_query.sh.
   2. FileZilla is a basic drag and drop format from desktop to O2 and vice versa. For more info: <https://harvardmed.atlassian.net/wiki/spaces/O2/pages/1588662157/File+Transfer>
2. Edit shell scripts to output to your O2 directory. You will only need to do this **ONCE**. To get more info about these slurm commands, see next section **Basic bash scripting commands**.

There are two shell scripts you will need to change: ALFA\_pull.sh and master\_AF\_query.sh. You do **NOT** need to alter the ALFA\_pull.py script at all.

In AF\_query directory, in command line:



This will open the ALFA\_pull shell script. You should see this at the top:

A picture containing graphical user interface

Description automatically generated

You need to change the location of STDERR and the email where notifications will be sent to (highlighted in yellow). Once you have done that, press Ctrl X. You will be prompted to “save modified buffer”. Select “Yes” by typing Y, then enter. This will overwrite the original shell script.

Repeat for master\_AF\_query.sh.

**Basic slurm/bash commands:**

For the AF query tool, there are only a few basic slurm commands that you need to know. Harvard’s O2 wiki provides good overview (<https://harvardmed.atlassian.net/wiki/spaces/O2/pages/1586793632/Using+Slurm+Basic>).

***A. Location in O2 directory commands***

1. pwd

Shows where you are in your O2 directory.



1. ls

Shows what is in the directory that you are in. If you want to see size of folders/files do “ls -l”.

Text

Description automatically generated

1. cd

Change directory. This will change your pwd. In the example below, I changed my pwd to AF\_query. The “./” denotes a directory within another.



1. rm

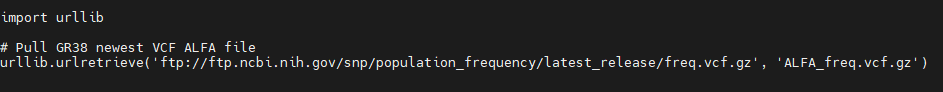
This removes a file. To remove folder, do “rm -r”. Use rm to delete slurm-JOBID.out and hostname-JOBID.err files once you have completed job to keep your O2 directory clean.

***B. Creating text file and editing bash script files***

1. nano

This opens a file and allows you to edit and/or rename as needed.





Ctrl X. You will be prompted to “save modified buffer”. Select “Yes” by typing Y, then enter. This will overwrite the original shell script. You can also rename instead.

Text

Description automatically generated

Note: Do **NOT** try to open VCF files using nano. They are zipped because they are so big- only use head command if you want to look at VCF.

1. head

This allows you to see the first 10 lines of a file. Unlike nano, it does not redirect you into the file- you will still be in the command line. You cannot edit files with head though- that needs to be done using nano. Additionally, if you want to see head of a VCF, you need to unzip first (see example below).

A picture containing graphical user interface

Description automatically generated

***C. Submitting a job and checking on job status***

1. sbatch



This submits a job to O2 cluster. Since the sbatch specifications are in the shell scripts, the sbatch commands for AF query are simple.

The shell scripts are set up so that when you submit job, you will be emailed when it enters queue, begins, and ends. If ran correctly, you will receive an email that states job had exit code 0.

If you get an exit code of NOT 0, that means that there was an error. Open the hostname-JOBID.err file in O2 directory using head or nano. This will tell you in more detail what went wrong.

1. squeue -u HMS ID



Checks status of submitted job.

1. scancel JOBID



Cancel submitted job using jobID. Use if you realize job is going to fail/input something incorrectly instead of waiting for it to fail.

**Example:**

In this example, we want to produce a file with the allele frequencies of a list of variants associated with cardiomyopathy in childhood cancer survivors.

1. **Create text file containing SNPs of interest. Name cardio\_variants\_2021-04-23.txt (replace with relevant names as needed for different projects)**

***OPTION A:***

In your AF\_query folder, create a new txt file. On command line:



This will open a blank txt file. Copy and paste in the SNPs of interest, with each SNP having its own line.

Text

Description automatically generated

Press Ctrl X. You will be prompted to “save modified buffer”. Select “Yes” by typing Y, then enter.

***OPTION B:***

Create text file containing SNPs of interest on your desktop. Use FileZilla to transfer to O2 directory.

Check: Make sure the txt file is nonempty and saved in AF\_query folder in O2 for the scripts to run properly.

1. **Download ALFA database.**

In AF\_query folder, submit batch job in command line:



It is not necessary to redownload the ALFA database every time you do a query. It is released on a quarterly basis. Check ALFA website (<https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/>) to see if a new version has been released since your last download. If not, keep using the one you already have!

Check: This will take ~10 minutes to complete. If ran correctly, directory should now have a large file ALFA\_freq\_YEAR-MONTH-DATE.vcf.gz in AF\_query folder. You now have the database that you want to extract info about SNPs from.

1. **Run AF\_query with ALFA database.**

In AF\_query folder, submit batch job in command line:



This script requires two arguments in the command line. After “sbatch master\_AF\_query.sh”, you need to put the ALFA database VCF name in your directory. Then, you put the SNP txt file name in your directory. All arguments are just separated by a space. Basically, you are telling the AF query script to find and pull all the SNPs in the SNPs text file from the ALFA database.

This will output the flat TSV file:

ALFA\_AF\_cardio\_variants\_2021-06-09.tsvI

In command line:

A picture containing text, keyboard

Description automatically generated

Check: This script will take about 30 minutes to complete. Check that the new file is nonempty using nano or head. In the above example, we see with head command that the AF query tool worked properly, and we have the outputs of interest in the final file.

1. **Export the ALFA\_AF\_cardio\_variants\_2021-04-23.tsv flat file using FileZilla.**

It is TSV file format, so when you open it on desktop, it will open in notepad. To switch to Excel, copy and paste it- Excel will recognize the tabulation and will fill the cells automatically.

**Summary Checklist:**

Do ***ONCE*** for initial AF query tool set up:

* Set up O2 account, HMS Duo authentication.
* Download MobaXterm and FileZilla.
* Transfer AF\_query folder to O2 using FileZilla.
* Edit shell scripts, AF\_pull.sh and master\_AF\_query.sh.

Do ***EVERY TIME*** you run AF query tool:

* Create new txt file containing the SNPs you want to get AF info from.
* Run AF\_pull.sh ***if*** a new ALFA version has been released since your last query (<https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/>).
* Run master\_AF\_query.sh with ALFA VCF and new SNP txt file you created.
* Transfer flat file of AF info onto desktop using FileZilla.