Script Name: dosage_converter_script.sh

Description: Converts dosage vcf files and info files from imputed data into MACH files for use in downstream association analyses.

Required Input: imputation results gzipped and ending in the following prefixes: dose.vcf.gz and info.gz One dose.vcf.gz and one info.gz per chromosome

Expected Output: This script should generate three files:

- prefix.mach.dose, (dosefile variable)
- prefix.mach.info, and (markfile variable)
- prefix.posFile.txt (posfile variable)

All three are required for GENESIS association analysis.

General Use Instructions

- for use in on a UNIX/Linux command line; no interactive sessions!
- Please do no change anything in the code except for the variable values in STEP 2, and do not change the variable names in STEP 2. Changing anything else in the code may cause errors or inaccuracies in the logic that was built
- This code had been optimized by chromosome, therefore, this script will need to be run for every chromosome and can be run in parallel across all chromosomes.

IMPORTANT!! Please go through the checklist and make sure all the following is completed before proceeding:

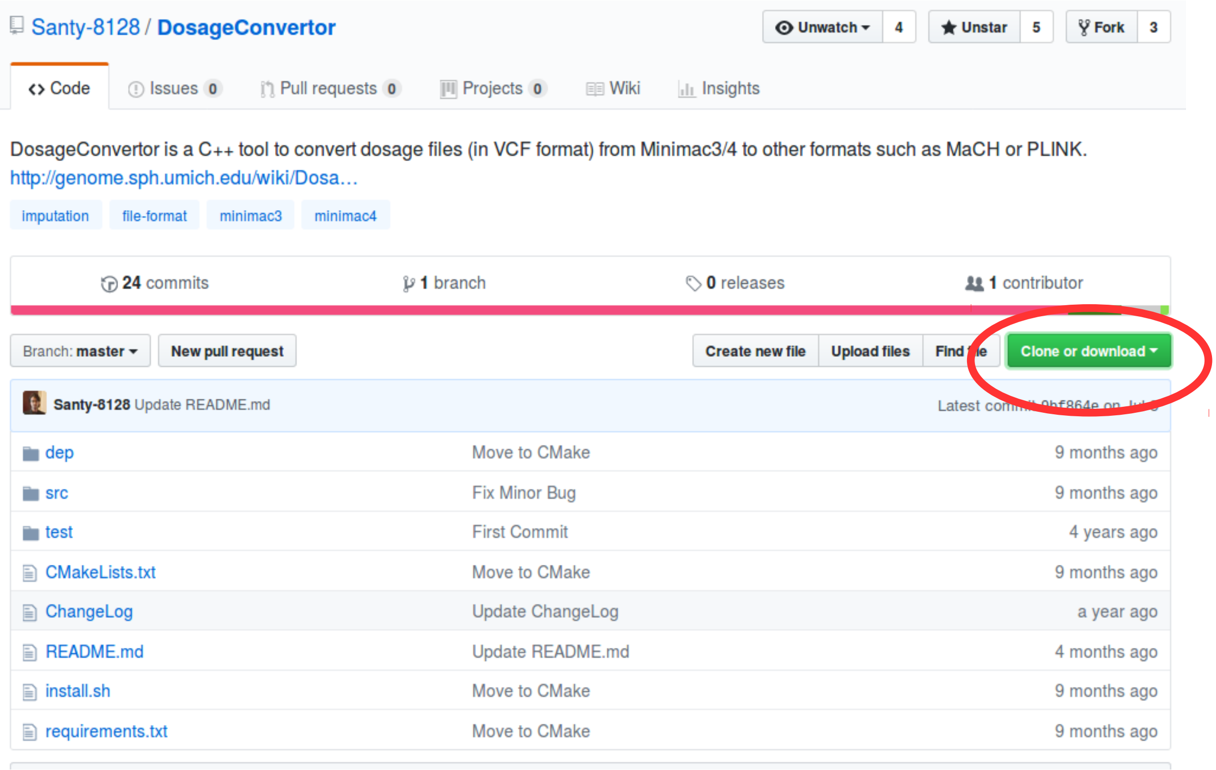
- ✔ Successfully run GENESIS_setup_ANALYSIS_standalone.R and have location of saved data object
- ✔ Successfully run PC_covariates_standalone.R and have a list of pcs that should be included as covariates in your association analysis model
- ✔ Imputed data on the same sample set used in the previous 2 scripts above
- ✔ Sample order of GENESIS scanAnnot is in the same order as the samples in imputed files **– order is critical! It does not use header or sample names to check this, completely positional!**

Step 1:

- First double check that all the items listed above are completed and checked. **I cannot emphasize how important it is that the sample order of the GENESIS scanAnnot is the exact same order as the imputation files since the MACH files will also mimic this same exact sample ordering!!**

- Clone the DosageConvertor software from github and follow installation instructions listed in the README.md and on their front-facing github page:

`link/URL:  https://github.com/Santy-8128/DosageConvertor`



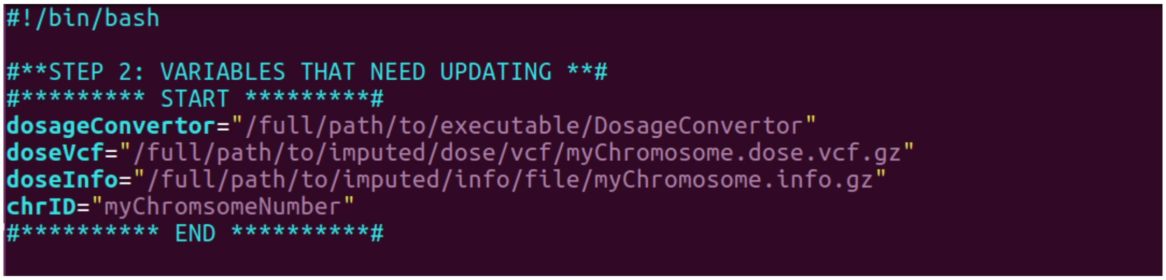
DosageConvertor is a C++ tool to convert dosage files (in VCF format) from Minimac3/4 to other formats such as MaCH or PLINK.
http://genome.sph.umich.edu/wiki/Dosa…

Santy-8128 Update README.md — Latest commit 0bf864e on Jul 5

dep	Move to CMake	9 months ago
src	Fix Minor Bug	9 months ago
test	First Commit	4 years ago
CMakeLists.txt	Move to CMake	9 months ago
ChangeLog	Update ChangeLog	a year ago
README.md	Update README.md	4 months ago
install.sh	Move to CMake	9 months ago
requirements.txt	Move to CMake	9 months ago

Step 2: Variables that need updating

- Open the dosage_converter_script.sh shell script in a text editor of your choice. This is the only step where code should be changed.
- There are 4 variables where the values need to be updated (Note, do not change the actual names of the variables)
- Since this is a shell script, it is critical that there are no spaces directly before or after the "=" sign and that all the variable values listed here are between quotation marks.
- Also, please do not modify or remove the very first line, `#!/bin/bash`, since this line will tell the computer which interpreter should be used to read the code.



```
#!/bin/bash

#**STEP 2: VARIABLES THAT NEED UPDATING **#
#********** START *********#
dosageConvertor="/full/path/to/executable/DosageConvertor"
doseVcf="/full/path/to/imputed/dose/vcf/myChromosome.dose.vcf.gz"
doseInfo="/full/path/to/imputed/info/file/myChromosome.info.gz"
chrID="myChromsomeNumber"
#*********** END **********#
```

Variable Name	Type	Definition
dosageConvertor	string	Full path to the installed dosageConverter executable
doseVcf	string	Full path to a single chromosome dose vcf file gzipped generated from imputation. Usually this file will look similar to this: chr1.dose.vcf.gz
doseInfo	string	Full path to a single chromosome info file gzipped generated from imputation. Usually this file will look similar to this: chr1.info.gz
chrID	string	A string of the chromosome ID that is running. This serves as a file prefix. Examples can include: chr1, 1, chromosome1, etc… please no whitespaces or special characters

- After the variable values have been updated, save the file.

Steps 3 and 4 below are explanations of what the code is doing, however, no changes should be made to these steps in the script.

Step 3: Decompress files and run DosageConvertor

- No changes to the code need to be made here.

Explanation of code:

This is decompress the gzip imputation files and store the output with the same file name, except dropping the .gz suffix.

After all files have been decompressed, dosageConvertor is called which takes in the decompressed dose.vcf file and the .info file an inputs. The `--prefix` parameter is set to the variable in STEP2 called chrID, meaning the output of the MACH files will begin with that variable value.

The two files generated here will populate the path variables (dosefile and markfile) in the GENESIS_Association_Analysis_standalone.R

Step 4: Generate SNP position file for GENESIS

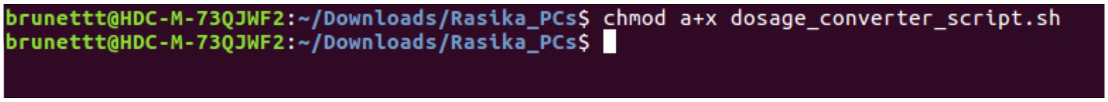
- No changes to the code need to be made here.

Explanation of code:

This generates the SNP position file, which will populate the path variable, posfile, in the GENESIS_Association_Analysis_standalone.R script.

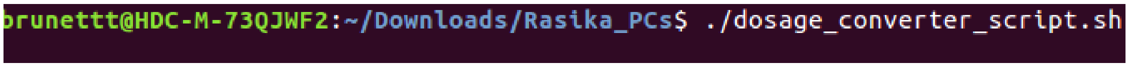
Running the Shell Script

- Open a terminal or command prompt and change the permission of the shell script to be an executable by running the following command:



```
brunettt@HDC-M-73QJWF2:~/Downloads/Rasika_PCs$ chmod a+x dosage_converter_script.sh
brunettt@HDC-M-73QJWF2:~/Downloads/Rasika_PCs$ 
```

- Now, run the script by running the following command:



```
brunettt@HDC-M-73QJWF2:~/Downloads/Rasika_PCs$ ./dosage_converter_script.sh
```

- Continue to run the script, x times until all chromosomes have been converted to MACH files.