MSBI 32400 – LAB 5
LARRY HELSETH, PHD AND JASON EDELSTEIN

February 13, 2019

Exploring (other people's) SNPs

2

- Today we'll download some public 23andMe data from https://opensnp.org/
- From the VM, open Firefox and go to opensnp.org
- Sign in as lhelseth@gmail.com using the VM password (without '!')
- Click on "Data" then "Genotypes" in the top menu, then click on "MsEscue" (ID #**6609**) to download their 23andMe profile
- Click on the "Search" icon in the top menu, then search for "AuriCrow" (ID #**7657**), click on "Users" then download their 23andMe profile
- Create your lab5 folders (lab5/bin, lab5/data, lab5/doc, lab5/source, lab5/results) and move "AuriCrow's" data to your lab5/data folder
- Inspect the genotype file format:
  - less **7657**.23andme.5442
    - Sometimes you have to try unzip -l **7035**.23andme.5442, etc.
  - How many lines of data (grep –v '^#' <file> | wc –l)
  - What are the columns of information (include in your README)

Download software to convert to VCF

3

- Go to your lab5/bin directory and clone software from GitHub:
  - git clone https://github.com/arrogantrobot/23andme2vcf.git
- cd to the new lab5/bin/23andme2vcf directory

```
student@MSBI32400Lab3:/data/lab5/bin
File  Edit  View  Search  Terminal  Help
[student@MSBI32400Lab3 bin]$ pwd
/data/lab5/bin
[student@MSBI32400Lab3 bin]$ git clone https://github.com/arrogantrobot/23andme2vcf.git
Initialized empty Git repository in /data/lab5/bin/23andme2vcf/.git/
remote: Counting objects: 155, done.
remote: Total 155 (delta 0), reused 0 (delta 0), pack-reused 155
Receiving objects: 100% (155/155), 12.96 MiB | 4.89 MiB/s, done.
Resolving deltas: 100% (68/68), done.
[student@MSBI32400Lab3 bin]$
```


Sometimes things don't work…

4

Proxies & firewalls sometimes get in the way…

- If git cloning doesn't work, can ~~always~~ usually just browse to website and download package as ZIP
- Open https://github.com/arrogantrobot/23andme2vcf and Download ZIP
- Use unzip 23andme2vcf.zip to expand in lab5/bin directory as before

Conversion

5

Run this perl script to convert 23andMe file to VCF:

- From the lab5/bin/23andme2vcf folder run:
  perl 23andme2vcf.pl
  /data/lab5/data/**7657.23andme.6002**
  /data/lab5/results/**7657.23andme.6002.**vcf *5*
  (the "5" at the end tells the script which hg19 reference version to use; there are three in the 23andme2vcf folder)
- ➢ Replace "***7657.23andme.6002***" with your genome file's name here and on subsequent slides

Look at the VCF

6

- How many lines were excluded?
- How many lines are there in the VCF (not counting the header)?
  - Put above information in the README you send Jason
- "Genotype"
  - 0/0 = reference/reference
  - 0/1 = reference/alt
  - 1/1 = alt/alt

Another way (*requires full genome*)

7

- Heng Li's bcftools can convert TSV files to VCF
  - https://samtools.github.io/bcftools/howtos/convert.html
  - bcftools convert -c ID,CHROM,POS,AA -s SampleName -f 23andme-ref.fa --tsv2vcf 23andme.txt -Oz -o out.vcf.gz

3.1 GB

```
student@MSBI32400Lab3:~
File  Edit  View  Search  Terminal  Help

TSV conversion:
    --tsv2vcf file
        convert from TSV (tab-separated values) format (such as generated by 23andMe) to VCF. The input file fields can be tab- or space- delimited

    -c, --columns list
        comma-separated list of fields in the input file. In the current version, the fields CHROM, POS, ID, and AA are expected and can appear in
        arbitrary order, columns which should be ignored in the input file can be indicated by "-". The AA field lists alleles on the forward reference
        strand, for example "CC" or "CT" for diploid genotypes or "C" for haploid genotypes (sex chromosomes). Insertions and deletions are not
        supported yet, missing data can be indicated with "--".

    -f, --fasta-ref file
        reference sequence in fasta format. Must be indexed with samtools faidx

    -s, --samples LIST
        list of sample names. See Common Options

    -S, --samples-file FILE
        file of sample names. See Common Options

    Example:

        # Convert 23andme results into VCF
        bcftools convert -c ID,CHROM,POS,AA -s SampleName -f 23andme-ref.fa --tsv2vcf 23andme.txt -Oz -o out.vcf.gz
```

23andMe Quirks

8

- Many proprietary SNP identifiers (not rsIDs)
- Report Indels as I or D instead of showing actual nucleotide (or '-' for deletion)
  - http://www.enlis.com/blog/2015/10/29/reverse-engineering-23andmes-proprietary-insertions-and-deletions/
  - In the latest 23andMe genotyping chip (v4) there are:4,093 total indels and 3,413 of these indels use a 23andMe proprietary identifier (83.3%)

Let's try to annotate the SNPs

9

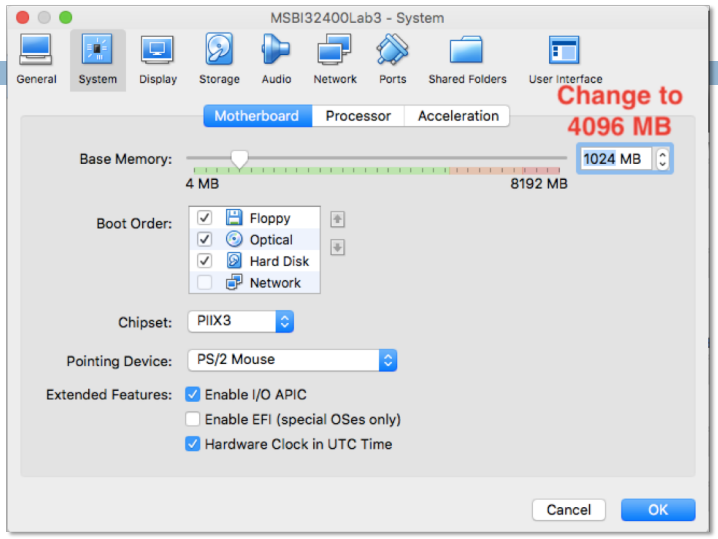
- I've installed a Java application, snpEff, on the VM in /data/snpEff/
  - Cf- http://snpeff.sourceforge.net/
    - https://www.ncbi.nlm.nih.gov/pubmed/?term=22728672
  - Check the command syntax by typing:
    java **-Xmx2G** -jar /data/snpEff/snpEff.jar

Changing memory on VM

10

- Shut down the VM completely
- Expand listing to show details
- Expand **System**
- Change memory as needed
  - NB-Do not assign all your laptop's RAM to the VM!

11



Annotating with RefSeq

12

- Use hg19 reference genome already installed
  - Use canonical sequences
  - Disable logging
- From lab5/results:
  java -Xmx2G -jar /data/snpEff/snpEff.jar eff -canon -noLog hg19 **7657.23andme.6002**.vcf > **7657.23andme.6002**.snpEff.vcf
  - *Be sure to "clean" the above if you paste it into your VM before executing*
  - *Takes ~2 minutes on 4 GB RAM VM*

Check out the new VCF

13

- snpEff adds annotation to each line.
- Read the header fields then look for codes
- Free text search for "stop" to find gain or loss of stop codons, etc.
  - List a few in your README

Expected File Structure

14

student@MSBI32400Lab1:/data/lab5

File Edit View Search Terminal Help

```
[student@MSBI32400Lab1 lab5]$ tree
.
├── bin
│   └── 23andme2vcf
│       ├── 23andme2vcf.pl
│       ├── 23andme_v3_hg19_ref.txt.gz
│       ├── 23andme_v4_hg19_ref.txt.gz
│       ├── LICENSE.txt
│       ├── readme.md
│       └── sites_not_in_reference.txt
├── data
│   └── 5657.23andme.4141
├── doc
│   └── README.md
├── results
│   ├── 5657.23andm3.4141.snpEff.vcf
│   ├── 5657.23andm3.4141.vcf
│   ├── snpEff_genes.txt
│   └── snpEff_summary.html
└── src

6 directories, 12 files
[student@MSBI32400Lab1 lab5]$
```

snpEff Summary Files

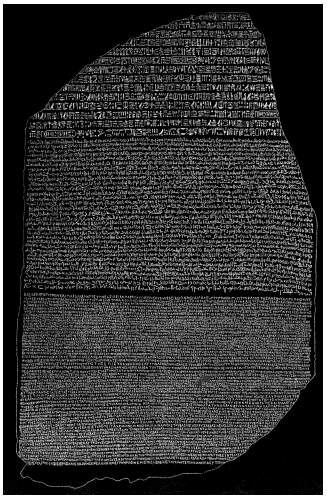
15

- From the lab5/results directory, open the snpEff generated summary file in Firefox:
  firefox snpEff_summary.html
- Review the summaries of the annotation.
  - How many were classified as "stop_lost" and how many as "stop_gained"?
  - Include in your README

Standardize data: what is the variation?

16

607008.0001
985A>G
985A>G (K304E)
985A>G (K329E)
A985G
ACADM, LYS304GLU
K304E
K304E (985 A->G)
K304E (K329E)
K304E only
K329E
K329E(985A>G)
LYS304GLU
Mutation c.985A>G (p.K304E)
c.985A>G
c.985A>G (p.K304E)
c.985A>G (p.Lys304Glu
c985A>G
includes: K304E (985A>G)
p.K304E
p.Lys329Glu
previously known as p.Lys329Glu
Analysis of ACADM 985A>G mutation



NC_000001.10:g.76226846A>
GNG_007045.1:g.41804A>G
NM_000016.4:c.985A>G
NP_000007.1:p.Lys329Glu
ACADM:c.985A>G
rs77931234:A>G

- LRG accessions reported when public
- GRCh38 or GRCh37

Source: Donna Maglott, NCBI

Let's see if we get more using Clinvar

17

- Use snpEff's companion package, SnpSift to add annotation to existing .snpEff.vcf
- Download latest GRCh37 VCF (.vcf.gz + .vcf.gz.tbi) from **Clinvar's FTP site** (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20190211.vcf.gz & ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20190211.vcf.gz.tbi)
- Move both files to /data/snpEff/data/hg19/clinvar/
  - Need to create that directory first
  - Don't gunzip files; use "zcat clinvar_20190211.vcf.gz| more" to look at header if curious
    - If you did gunzip either download again OR bgzip clinvar_20190211.vcf && tabix –p vcf clinvar_20190211.vcf.gz

MSBI 32400 Lab 5 2/13/2019

Annotate snpEff.vcf with SnpSift

18

- java -Xmx2G -jar /data/snpEff/SnpSift.jar annotate -noLog /data/snpEff/data/hg19/clinvar/clinvar_20190211.vcf.gz /data/lab5/results/**7657.23andme.6002**.snpEff.vcf > /data/lab5/results/**7657.23andme.6002**.clinvar.snpEff.vcf

➢ Added line breaks for clarity
  - Above multiline break is "<space>\" then Return
  - When pass to Linux, line breaks are removed

```
java -Xmx2G -jar /data/snpEff/SnpSift.jar annotate -noLog /data/snpEff/data/hg19/clinvar/
clinvar_20180701.vcf.gz \7657.23andme.6002.snpEff.vcf > \7657.23andme.6002.clinvar.snpEff.vcf
```

Remove

Inspect the new VCF

19

- Search for lines containing CLNSIG=Pathogenic
  - Check the new header for key
- Include a few rsIDs and gene names in the README file you send to Jason from SNPs flagged as Pathogenic

PharmGKB

20

http://www.pharmgkb.org/

- Let's look up translatation table for TPMT
- Look at different tabs and pathway diagram
- Let's download their translation table for TPMT

Pharmacogenomics

21

- Search the original 23andMe data file for the following rsIDs:
  - rs1800462, rs1142345, rs1800460 & rs1800584
- Determine the "star allele" status for TPMT from the 23andMe data using the translation table shown on the next slide
  - Remember, each person has two alleles so should have two stars (like "*1/*1", etc.)
- Include this in the README file you send to Jason

TPMT Translation Table

22

TPMT	rs1800462	rs1142345	rs1800460	rs1800584
*1	C	T	C	C
*2	G	T	C	C
*3A	C	C	T	C
*3B	C	T	T	C
*3C	C	C	C	C
*4	C	T	C	T

Let's interpret the SNP data on-line

23

- http://genotation.stanford.edu/
- Developed by members of the Training Program in Biomedical Informatics at Stanford University
  - http://psb.stanford.edu/psb-online/proceedings/psb12/karczewski.pdf
- Browser upload the plain text 23andMe file (not ZIP) in top right corner ("Begin Exploring"), then assume European
- The 23andMe data will not be sent to any server, it remains on your computer.

Review their PGX interpretation

24

- Click on Clinical, then Pharmacogenomics, then Show my Common PGx Variants
  - Maintained by people behind PharmGKB.org
  - What is their interpretation of TPMT status?
  - Include this in the README file you send to Jason
- If interested in more, check out:
  - http://www.23andyou.com/3rdparty
  - http://thegeneticgenealogist.com/2013/09/22/what-else-can-i-do-with-my-dna-test-results/

If interested, check out "Disease"

25

- Several predictors, including GWAS Variants
- Derived from NCBI/EBI GWAS Catalog
  - http://www.ebi.ac.uk/gwas/

GENOtation

26

- http://genotation.stanford.edu/
- Explore the **Clinical, Sports, Traits** and **Ancestry** tabs
  - Describe the Clinical/Lung Function results
  - What is your CYP2C19 genotype and how does it affect function?
  - Does your genotype have any risk alleles associated with Caffeine consumption?
  - Does your genotype have any risk for motion sickness?
  - Compare some predicted traits with AuriCrow's self-reported traits (see "Variations" tab at https://opensnp.org/users/7657)

Homework

27

- Please submit the README with the file information requested above before next class through Canvas or e-mail Jason (jasone@uchicago.edu) with "**Lab #5**" in the subject line