

Project option 1 - Drug annotation of 23andme report

23andme provides the raw data regarding genetic variants for a sample provided by a customer in a tab separated file.


The goal of this class project is to detect variants with drug effects as provided by PharmGKB.

Examples of 23andme files are available at:

<https://github.com/arrogantrobot/23andme2vcf> 
(<https://github.com/arrogantrobot/23andme2vcf>)

For this project use: **23andme_v5_hg19_ref.txt.gz** from the above link.

Details about the format can be found at:

<https://eu.customercare.23andme.com/hc/en-us/articles/115002090907-Raw-Genotype-Data-Technical-Details>  (<https://eu.customercare.23andme.com/hc/en-us/articles/115002090907-Raw-Genotype-Data-Technical-Details>)

<https://samtools.github.io/bcftools/howtos/convert.html> 
(<https://samtools.github.io/bcftools/howtos/convert.html>)

Columns in the 23andme_v5_hg19_ref.txt file are:

CHR: the chromosome number

POS: the position of the variant - position where the difference in nucleotides is found

dbSNP_ID: the variant identifier in the dbSNP database (<https://www.ncbi.nlm.nih.gov/snp/>)
- this field will contain a dot . when the SNP is not available in the dbSNP database

ALLELE: the nucleotides found at that position on the pair of chromosomes

PharmGKB data is available at:

<https://www.pharmgkb.org/downloads>  (<https://www.pharmgkb.org/downloads>)

To connect the variants to the information available in the PharmGKB data we need the **var_drug_ann.tsv** available at the above link in the **variantAnnotations.zip** archive.

The **var_drug_ann.tsv** file contains the variant annotation information in a tab separated format.

The file has a bad line (5487) where a quote is not closed ("The 15-year cumulative).

You can open the file in Jupiter Lab to see the line numbers.

Handle the bad line (either fix it, manually, or skip it programmatically, but only skip that line).

Columns in the var_drug_ann.tsv file are:

Variant Annotation ID: unique ID number for each variant/drug annotation

Variant/Haplotypes: dbSNP ID or haplotype(s)

Gene: HGNC symbol

Drug(s): Drug name

PMID: PubMed identifier

Phenotype Category: options [efficacy, toxicity, dosage, metabolism/PK other]

Significance: yes or no – determined by if the author stated the association was significant

Notes: curator notes field

Sentence: structured sentence

Alleles: variant alleles in annotation

Specialty Population: tags for any special populations this annotation is relevant to (e.g. pediatric)

And other columns: Metabolizer types, isPlural, Is/Is Not associated, Direction of effect, PD/PK terms, Multiple drugs And/or, Population types, Population Phenotypes or diseases, Multiple phenotypes or diseases And/or, Comparison Allele(s) or Genotype(s), Comparison Metabolizer types

The following tasks must be completed for this project:

1. Map/merge the 23andme file and the variant-drug annotation file based on dbSNP_ID (also known as rsID).

The merged result should have the following columns:

dbSNP_ID, GENE_SYMBOL, DRUG_NAME, PMID, PHENOTYPE_CATEGORY, SIGNIFICANCE, NOTES, SENTENCE, ALLELE_PharmGKB (variant alleles in annotation), ALLELE_23andme (variant alleles in 23andme file)

2. Filter the output so that it only contains significant associations (**SIGNIFICANCE** is **yes**) for variants that affect the drug efficacy (**PHENOTYPE_CATEGORY** is **efficacy**).

3. Save the output of the filtering step in a tab-separated file (**23andme_PharmGKB_map.tsv**) with the following columns:

dbSNP_ID, GENE_SYMBOL, DRUG_NAME, NOTES, SENTENCE, ALLELE_PharmGKB, ALLELE_23andme

4. Create a tab separated file (**23andme_PharmGKB_summary.tsv**) with summarized data with the following columns:

GENE_SYMBOL, DRUG_NAME, dbSNP_IDs (list of IDs separated by ";")

5. Plot the distribution (histogram) of the number of drugs associated with a gene, and the number of SNPs for a gene.

6. As a team decide on a new feature to implement that can answer a relevant biological question using these data and implement the feature