Universiteit Leiden
The Netherlands

Opleiding Bioinformatica

Actionable pharmacogenomic variation present in the general population

Nienke Biesot

Supervisors:
Katy Wolstencroft & Sacha Goultiaev

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

05/07/2021

Abstract

People respond differently to drugs because of the variations in DNA that alter the expression of genes that metabolise drugs. Cytochrome P450 (CYP) is a protein that plays a key role in the metabolism of drugs. The CYP class of proteins has more than 50 enzymes that are essential for the metabolism of many medications. In this thesis project I looked at which actionable pharmacogenomics variation is present in the general population by studying publicly available human variation data. I explored how to score and partition groups who are predicted to have sub-optimal responses to treatment. This is done by analysing the Variant Call Format files from the Personal Genome Project UK. Stargazer is to analyse CYP protein variants and Clustal Omega is used to analyse Multiple Sequence Alignments of individual variants at the protein level. To analyse the unknown variants, Variant Effect Predictor is used. This analysis shows that from the individuals studied, there were three different types of responders to a whole range of drugs. Many of the CYP genes showed variations, with the most variation in the CYP2D6 gene, which is responsible for metabolising 25% of all approved drugs. Much of this variation corresponds to the most common variations in Europe. The consequence of these three different phenotypes is that the individuals should have different drug prescriptions from one another.

Contents

1 Introduction

Pharmacogenomics focuses on the identification of genetic variants that influence drug effects. This is often done by alterations in pharmacokinetics (what does the body do to a drug) and pharmacodynamics (what does the drug do to the body) [RE15].

1.1 The situation

By mapping individual variations of a patient, it is possible to suggest personalised treatment, which is the main idea behind personalised medicine. This is necessary because people respond differently to drugs. This can sometimes pose danger when using them, because the drugs are not metabolised (as quickly) or it has little effect on people because the drugs are metabolised to quickly. The situation is that there are differences in DNA that alter the expression or function of proteins, which is the reason for the difference in metabolism. The targeted proteins can contribute significantly to variation in the responses of individuals [ER04]. A good example are the CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4, and CYP3A5 genes, which plays a key role in the metabolism of 90% of the medications [TL07].
Depending on the variation in the genes, you can be a poor-metabolizer, intermediate-metabolizer, extensive-metabolizer (normal) and even ultrarapid- metabolizer. Ultrarapid metabolizers are at an increased risk of therapeutic failure because they metabolize drugs extremely quickly [Gae13]. Poor metabolizers who do not have appreciable gene activity are at risk of dose-dependent adverse drug events. It is therefore important to know what the variation is in the genes that are involved in metabolizing drugs.
A lot of genotyping test are available for these genes. A good example is Genelex, by doing a DNA test Genelex can provide an overview of the different variations present in the genes that metabolise drugs [teac]. The 6 most common variations (*1A,*1C,*1E,*1F,*1J,*1K) for the CYP1A2 gene are tested by Genelex. For the CYP2C9 gene, Genelex can identify the nine most common variations (*2-*6,*8,*11,*13,*15). Genelex identifies the eleven most commen alleles (*2-*10,*12,*17) for the CYP2C19 gene. The CYP2D6 test identifies 25 most common variants ((*2,*2A,*3-*12,*14,*15,*17,*19,*20,*21,*29,*30,*35,*36,*41,*56,*109). For the CYP3A4 gene the variants *1B,*22 can be identified, and for the gene CYP3A5 the variants *3,*6,*7 can be identified [teac].
The aim of the research is to investigate what actionable pharmacogenomics variation present is in the general population and explore how to score and partition groups who are predicted to have sub-optimal responses to treatment. In this project I create automatic methods to examine gene variation data and predict the effects of those variations on drug metabolism. I use data from the Personal Genomes UK database [CCGA+19], which contains open genomic sequence data from 118 healthy individuals. These variations are represented in Variant Call Format (VCF) files. To analyse these variants in the VCF files, Stargazer is used to analyse CYP protein variants and Clustal Omega is used to analyse Multiple Sequence Alignments of individual variants at the protein level. To analyse the unknown variants, Variant Effect Predictor is used.

1.2 Thesis overview

This chapter 1 contains the introduction. Section 2 gives the background on the relevant topics and section 3 gives a description on previous research knowledge. The methods used in this thesis project are explained in section 4. The resulst are shown in section 5. Section 6 discusses the found results and section 7 gives the conclusion of this thesis.

2 Background

2.1 Cytochrome P450

Cytochrome P450 (CYP) is a protein that plays a key role in the metabolism of drugs. Pathways are classified by similar gene sequences, which is assigned as a family number (CYP2) and a subfamily letter (CYP2D) and are then differentiated by a number for the isoform or individual enzyme (CYP2D6) [AMMCHD13]. The CYP class has more than 50 enzymes that are essential for the metabolism of many medications. 90% of the drugs are metabolized by six enzymes: CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4, and CYP3A5 [TL07], table 1 gives an overview of which drugs are involved for the six genes.

Enzyme	**Inhibitors**	**Inducers**
CYP1A2	Amiodarone, cimetidine, ciprofloxacin, fluvoxamine	Carbamazepine, phenobarbital, rifampin, tobacco
CYP2C9	Amiodarone, fluconazole, fluoxetine, metronidazole, ritonavir, trimethoprim/sulfamethoxazole	Carbamazepine, phenobarbital, phenytoin, rifampin
CYPC19	Fluvoxamine, isoniazid, ritonavir	Carbamazepine, phenytoin, rifampin
CYP2D6	Amiodarone, cimetidine, diphenhydramine, fluoxetine, paroxetine, quinidine, ritonavir, terbinafine	No significant inducers
CYP3A4/CYP3A5	Clarithromycin, diltiazem, erythromycin, grapefruit juice, itraconazole, ketoconazole, nefazodone, ritonavir, telithromycin, verapamil	Carbamazepine,Hypericum perforatum, phenobarbital, phenytoin, rifampin

Table 1: Cytochrome P450 Enzymes and their inhibitors (increase drug effect) and inducers (decrease drug effect) [TL07].

Genetic variation in these genes can influence a patient's response to commonly prescribed drug. The process of genotyping is used to determine the genotype. It indicates which alleles are presented in the individual for the CYP genes. Each allele has a name that consists of a star (*) and a number [KNMb], an example of a possible variation in the CYP2D6 genotype is CYP2D6*1/*3. This means that this individual is heterozygous for variation *1 and variation *3.
Based on these star allele variations, the population can be divided into four categories [Gae13].

- Poor metaboliser (PM): no metabolic capacity
- Intermediate metaboliser (IM): reduced metabolic capacity
- Extensive metaboliser (EM): normal metabolic capacity
- Ultra-rapid metaboliser (UM): increased metabolic capacity

To predict the phenotype of an individual, each variation has an activity score. The combination of the activity scores from both the variation from each chromosome determines the phenotype (if there is no specific variation then you have star allele *1 with an activity score 1). For example if an individual has the genotype CYP2D6*1/*9, the activity score for CYP2D6*1 is 1 and the activity score for CYP2D6*9 is 0.5. Combining these activity scores gives an activity score of 1.5. That means that this individual is a normal metaboliser. An activity score of 0.5 indicates decreased activity and not that the activity is half of a normal functioning allele. Table 2 provides an overview of the activity score for each phenotype.

Phenotype	**Activity score**
Poor metaboliser	0
Intermediate metaboliser	0.25 - 1
Extensive metaboliser	1.25 - 2
Ultra-rapid metaboliser	≥ 2

Table 2: Activity score for each phenotype

2.1.1 CYP2D6

Cytochrome P450 2D6 (CYP2D6) is one of the first identified examples and therefore one of the most researched gene. The CYP2D6 gene plays an important role in the metabolism of about 25% of clinically used drugs [Gae13]. The drugs that fall under this are, for example: many antidepressants, antipsychotics and opioids (painkillers) [Gae13]. The CYP2D6 gene locus contains three genes: CYP2D6, CYP2D7 and CYP2D8. CYP2D7 and CYP2D8 are considered pseudogenes (Pseudogenes are nonfunctional segments of DNA that resemble functional genes [Tut12]). All three genes are composed of nine exons and share a high degree of sequence similarity [NTS+19].

The majority of the CYP2D6 variants are single nucleotide variants (SNVs) and indels (short insertions and deletions). These variations may result in an enzyme with functions more or less efficiently than in the normal state. Some of the variants are more complex. For example the variant CYP2D6*5 is characterized by a deletion of the entire CYP2D6 gene. Another example is that numerous CYP2D6 alleles are known to occur as diplications or multiplications, for example CYP2D6*1x2 (x2 means that the gene occurs twice) [NTS+19]. In picture 1 an example of an SNV, deletion and multiplication variation is given.

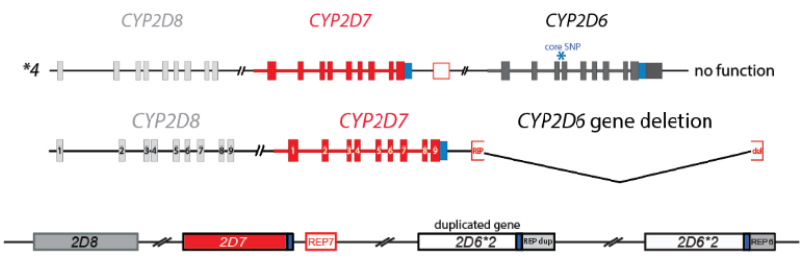

Figure 1: An example of a SNV, deletion and duplication variation. https://www.pharmvar.org/gene-support/Variation_CYP2D6.pdf

2.2 Data

2.2.1 Personal Genome Project UK

The Personal Genome Project UK (PGP UK) provides open genome, trait, and health data. Sharing data is critical to scientific progress, but has been hampered by traditional research practices, and of the concerns about sharing your genome data publicly. Their approach is to invite willing participants to openly share their personal genome data [CCGA+19]. As previously mentioned, the data that can be used are the VCF files form 118 volunteers.

2.2.2 Reference genome

A reference genome can be seen for example as a digital nucleic acid sequence database. It is a representative of the whole sequence in one individual organism of a species. A reference genome is assembled from the sequencing of DNA from a number of individual donors. When an individual genome is sequenced and compared to the reference sequence, the reference is seen as 'normal'. By identifying the variations in the individuals sequence, it is possible to see if these variations can cause problems. The reference genome for the Pharmacogene Variation Consortium (PharmVar) files and Stargazer tool (see section 3.2 and 3.3) are GRch37 and the reference genome for the VCF files created by the PGP-UK is humanG1Kv37. The reference genome humanG1Kv37 is equivalent to b37 (b37 is a human genome reference based on GRCh37), with the exception that it does not contain the decoy sequence for human herpesvirus 4 type 1 [Teab]. The positions that stand in the PharmVar files cannot be compared with the positions in the PGP-UK VCF files, because of the different referene genome. Therefore it is important to convert genomic coordinates between different reference genomes. To do a liftover from the coordinates a chain file is needed. A chain file describes a pairwise alignment between two reference assemblies. The section 4.1.1 explains how this can be done.

2.2.3 Variant Call Format

A VCF file contains meta-data lines (meta-data is information about the information standing in the VCF file), a header line and then the data lines. The data lines contains information of the position where there is a variant [LHW+09]. VCF simply records where there are differences

between the individual genome and the reference genome.

3 Related work

3.1 Worldwide distribution of cytochrome P450 alleles

A lot of research has been done on the different variations in the CYP genes. The worldwide distribution of the CYP genes has also been looked at. These studies have shown that there is a lot of difference between the most common variations of different ethnic backgrounds [ZISL17]. Figure 2 gives an overview of the the worldwide distribution for the CYP2C9, CYP2C19, CYP2D6, CYP3A4 and CYP3A5 genes. As for the CYP1A2 gene, the variant *1F is most common in all the different ethnic backgrounds [ZISL17].

Figure 2: An overview of the worldwide distribution for the CYP2C9, CYP2C19, CYP2D6, CYP3A4 and CYP3A5 genes [ZISL17]

3.2 Pharmacogene Variation Consortium

Pharmvar is a repository for pharmacogene variation. The focus from Pharmvar is mainly on the haplotype structure and allelic variation [NTS+19]. PharmVar maps sequence variations for genes in the CYP family, to genomic and transcript reference sequences, for example the GRch37 genome build. In other words, all known variations of the CYP genes are available through the website (https://www.pharmvar.org/genes). The variations are also described in VCF files.

Pharmvar provides an overview and summary of the CYP genes genetic variation. This information provided by PharmVar is used by the Pharmacogenomics Knowledgebase (PharmGKB) and the Clinical Pharmacogenetics Implementation Consortium (CPIC).
PharmGKB provides clinical guidelines and drug labels, potentially clinically actionable gene-drug associations and genotype-phenotype relationships [WCMH+12]. The information that is available for the CYP genes are an allele definition table, allele functionality table and diplotype-phenotype table (https://www.pharmgkb.org/page/cyp2d6RefMaterials).
The goal of CPIC is to address the barrier to clinical implementation of pharmacogenetic tests by creating freely available clinical practice guidelines [RK11]. The guidelines for the CYP genes can be found here: https://cpicpgx.org/guidelines/.

3.3 Stargazer

Stargazer is a bioinformatic tool, that can be used for identifying star alleles in PGx (pharmacogenes) genes [bLWP+18]. The way stargazer does the identification is by detecting SNVs, indels and SVs (structural variations).

Stargazer has four tools that can be used [bLWP+18]:

- Genotype: predicts star alleles for a chosen target gene from genomic data
- Setup: provide files that are necessary for the use of Stargazer
- View: performs secondary analyses of genotype calling
- Pipeline: end-to-end solutions for genotyping pipeline

The tool that can be used in this research is the genotype tool.

4 Methods

In the workflow diagram 3 an overview of the methods used in the project is given. First a liftover for the VCF files is needed using the CrossMap tool, after that the VCF files are used for the Stargazer tool to predict the star alleles. The output files from Stargazer are then used to help with a multiple sequence alignment together with literature about the variations and the reference SNP report. The Stargazer output is also used together with the lifted VCF files to make a variant overview in the genome browser IGV. Variant Effect Predictor is used for analysing the variants that were not included in any of the star allele variants. These steps are explained in more detail in the subsections 4.1 and 4.2.

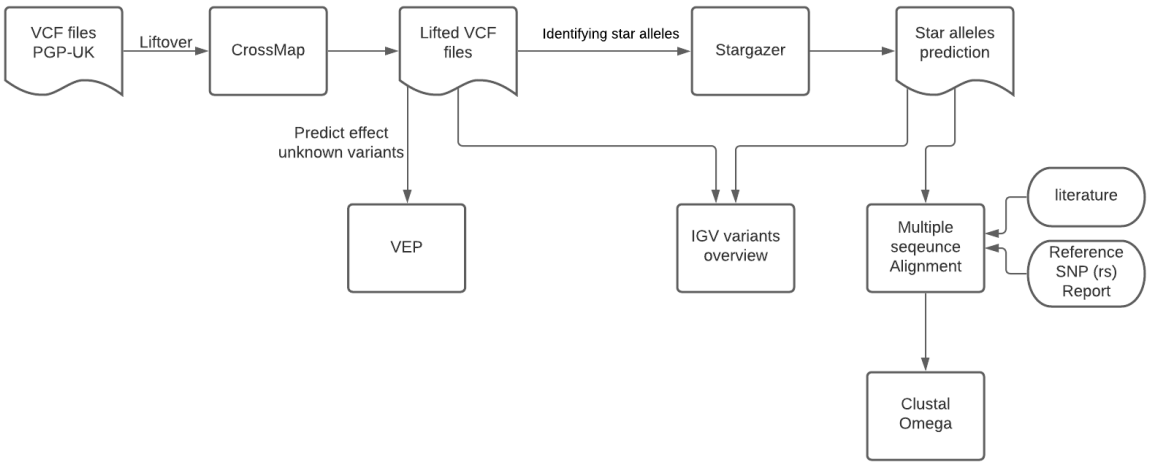

Figure 3: A workflow diagram for the methods used

4.1 Data preprocessing

In the following subsections, it is explained how the VCF files have been edited for analysis. The steps taken are automated by writing a bash script which can be found in section 7 or can be found here: https://git.liacs.nl/s1940023/thesis_files/-/tree/master/Thesis_files. For each subsection explained, this step is applied to all VCF files contained in a folder.

4.1.1 Liftover

To convert coordinates from one assembly to another, many resources can be used. The liftover tool used in this thesis project is CrossMap.
CrossMap is designed to liftover genome coordinates between assemblies [ZSW+13]. CrossMap is written in Python and C. Source code and a comprehensive user's manual are freely available. To convert coordinates from a VCF file different arguments are needed:

- Chain file
- Input file
- Output file
- Reference genome

In figure 4 an overview of how CrossMap works is given. An interval tree is used to lift over the given interval to the new genome build.
The chain file used for the liftover can be found here: https://gatk.broadinstitute.org/hc/en-us/articles/360035890811 and the reference genome can be found here: https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/.
An example of a command line for the liftover:

```
$ CrossMap.py vcf b37tohg19.chain human1.vcf hg19.fa human1.lifted.vcf
```

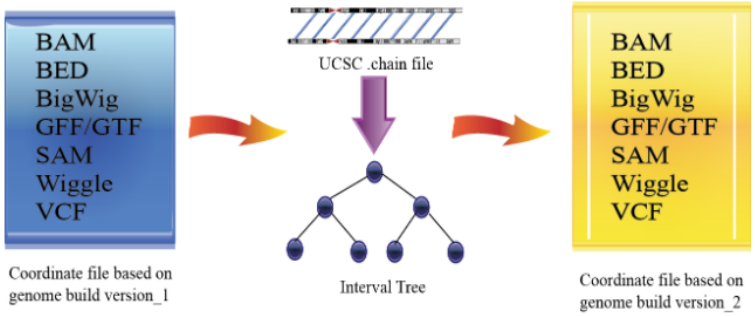

Figure 4: An overview of how CrossMap works [ZSW+13].

Where `b37tohg19.chain` is the chain file, `human1.vcf` is the input (vcf) file, `hg19.fa` is the reference genome and `human1.lifted.vcf` is the output (vcf) file

4.1.2 Tabix

To view a VCF file in the genome browser IGV [RTW+11], a tabix has to be created. This can be done very easily by the following two command lines [Li]:

```
$ bgzip -c human1.lifted.vcf > human1.lifted.vcf.gz
$ tabix -p human1.lifted.vcf.gz
```

4.2 Data analysing

4.2.1 Identifying star alleles

As described previously, the tool from Stargazer that can be used for analaysing the star alleles is Genotype. To go from a VCF file as input to a predicted phenotype as output, several steps are required in the algorithm. In figure 5 the different steps are shown.
The first step is to predict the star alleles from the input VCF file based on the SNVs/indels. Beagle (Beagle is a software package for imputing ungenotyped markers and for phasing genotypes) is used to haplotype phase heterozygous variants for the target gene [BZB18]. The phased haplotypes are then matched with Stargazer to star alleles with a translation table. The translation table that is used are the tables from the Pharmacogene Variation Consortium (see section 3.2).
The next step is the detection of SVs. This is done by using a target GATK-DepthOfCoverage format file (The GATK-DepthOfCoverage tool is used to generate coverage summeray information [Teaa]). Stargazer uses these files to convert read depth for the target gene to copy number by performing intra- and inter-sample normalizations. By using a control GDF file Stargazer then automates detection of SVs with changepoint, an R package.
If the prediction of the star alleles and de detection of SVs are done, then the next step is the identification of diplotypes. If there are samples without SVs, Stargazer determines the diplotype from the target gene by combining the star allele that are used by the phased haplotype.
The last step is the assignment of predicted phenotypes. This is done by translating the diplotypes into an activity score. If the activity score is known, a phenotype can be assigned to it (see section

2.1).

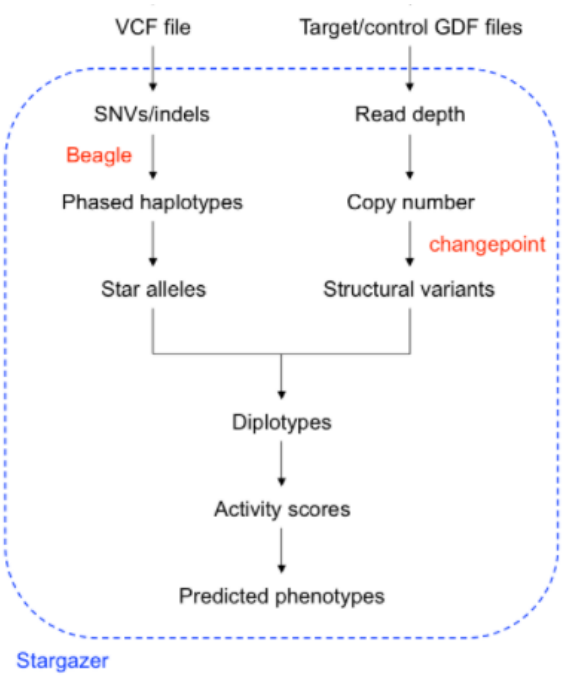

Figure 5: A schematic diagram of the Stargazer target gene (CYP2D6) genotyping pipeline, [bLWP+18].

The Genotype tool can be used by the command line:

```
$ stargazer.py genotype -o output_CYP2D6 -d chip -t CYP2D6
--vcf human1.lifted.vcf
```

Where `output_CYP2D6` is the output file, `chip` is the input data, `CYP2D6` is the target gene name and `human1.lifted.vcf` is the input file.

Table 3 provides an overview from the output file.

4.2.2 IGV variant overview

To make an overview of the different types of variants found in the genes, the genome browser IGV is used. Pharmvar provides the type of variant for the star alleles, for example a missense variant or a frameshift variant. These variants can be indicated with a color.

4.2.3 Variant Effect Predictor (VEP)

A lot of the variants in the different CYP genes were not described in any of the star allele variants or sub star allele variants on Pharmvar. To see if these variants result in change of the protein sequence the tool VEP [MGH+16] is used. The effect of the variants on the genes,

Header	Description
hap1_main	Main star allele for the 1st haplotype
hap2_main	Main star allele for the 2nd haplotype
hap1_cand	Candidate star alleles for the 1st haplotype
hap2_cand	Candidate star alleles for the 2nd haplotype
hap1_score	Activity score for the 1st haplotype
hap2_score	Activity score for the 2nd haplotype
dip_score	Combined activity score for both haplotypes
phenotype	Predicted phenotype based on the diplotype activity score
hap1_sv	SV call for the 1st haplotype
hap2_sv	SV call for the 2nd haplotype
hap1_main_core	Core SNPs of the 1st haplotype's main star allele
hap2_main_core	Core SNPs of the 2nd haplotype's main star allele

Table 3: An overview of an output file from stargazer with the header used for the analysing [bLWP+18].

transcripts and protein sequence are determined by VEP. VEP can be found on this webiste: https://grch37.ensembl.org/Homo_sapiens/Tools/VEP.

4.2.4 Multiple Sequence Alignment

Classification of the population is based on the activity score that the variation has. It is therefore interesting to see what the variation looks like at protein level. This can be done by performing a Multiple Sequence Alignment (MSA). The algorithm used for this is Clustal Omega [SWD+11] and can be found on this website: https://www.ebi.ac.uk/Tools/msa/clustalo/.

The first step of the algorithm is producing a pairwise alignment by using a k-tuple analysis. The next step is clustering the sequences by using the method mBed. This is followed by the k-means clustering method. Followed up by constructing the guide tree using the UPGMA method. The final step is producing the multiple sequence alignment by using the HHalign package [DDS13].

For most of the variations, Pharmvar provides the location of the variation and how this variation change the amino acid. This amino acid can be changed in the reference protein sequence (the reference protein sequence can be found on the website https://www.ncbi.nlm.nih.gov/). For the variants that are more difficult, for example a splicing defect, it is harder to determine what this looks like at the protein level. By looking at the literature and dbSNP reports, the protein sequence was discovered.

5 Results

The following subsections show the results per gene. For each gene, a summary can be seen of the output of the Stargazer tool, the type of variant overview in the genome browser IGV and the Multiple Sequence Alignment of the most interesting alignment. The output files from Stargazer and the Multiple Sequence Alignment files can be found here: https://git.liacs.nl/s1940023/thesis_files/-/tree/master/Thesis_files

5.1 CYP1A2

In table 4 an overview is given of the output of the Stargazer tool for the gene CYP1A2. All individuals have variant *1F, whose activity score is unknown. It is therefore not possible to know the phenotype for these individuals. For the individuals heterozygous for the variant, it can be seen that the activity score for the first haplotype 1.0 is. The individuals can not be poor metabolizers because of this. If there is no duplication of the gene, the individuals can not be ultrarapid metabolizers, because the highest activity score the variation can have is 1 and only if there is an repetition of the gene the activity score can be higher than 1. Because then you sum up the activity score for each gene duplication. For the individual who is homozygous for this variant, it can not be said that this individual is not a poor metaboliser. As mentioned in section 3.1 the most common veriant in Europe for the CYP1A2 gene is *1F, which can also be seen in the table.

CYP1A2						
	Star allele for the 1st haplotype	Star allele for the 2nd haplotype	Activity score 1st haplotype	Activity score 2nd haplotype	Combined activity score	Phenotype
Human1	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 2	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 3	*1F	*1F	Unknown	Unknown	Unknown	Unknown
Human 4	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 5	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 6	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 7	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 8	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 9	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 10	*1	*1F	1.0	Unknown	Unknown	Unknown
Human 11	*1	*1F	1.0	Unknown	Unknown	Unknown

Table 4: Variations, activity score and phenotype for the 11 humans for gene CYP1A2

Figure 7 shows the variants that occur in the CYP1A2 gene region. Each variant is assigned a color, the meaning of the color can be found in figure 6. Variant *1F is an intron variant that has a low impact. In addition to this variant, the individuals also have a synonymous variant that has no impact on the protein sequence. Many of these variants that the individuals have are not described in the star alleles. For this, VEP was used to see whether these variants can have an impact on the protein activity. VEP gave the output that the variants are a transcript variant occurring within an intron, with a modefier impact (this is not only for the CYP1A2 gene, also for the other genes VEP gave this output for the undescribed variants). A modefier impact means that non-coding variants or variants affecting non-coding genes are hard to predict or there is no evidence of impact [MGH+16].

Variation	Color
Missense high impact	
Missense low impact	
Splicing defect high impact	
Frameshift high impact	
Deletion moderate impact	
Intron high impact	
Intron low impact	
Synonymous	
Homozygous which is not described in any of the variants	
Heterozygous which is not described in any of the variants	

Figure 6: color legend for the variation

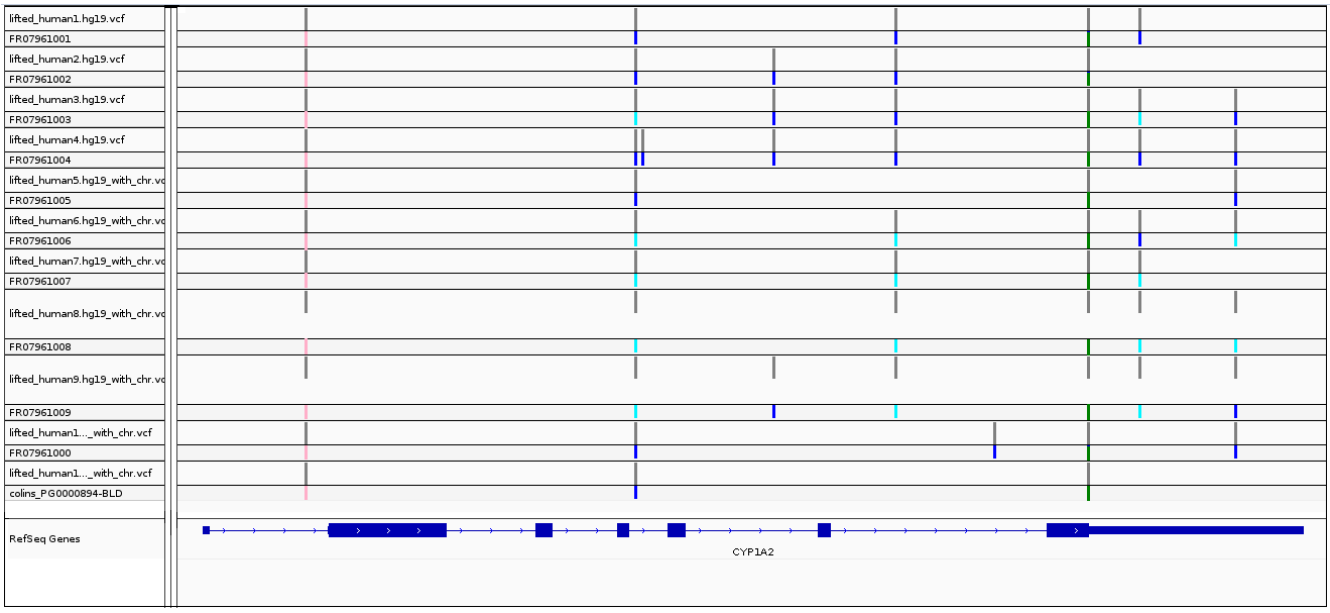

Figure 7: Types of variation in the CYP1A2 gene

As can been seen in figure 8 the variation has no impact on the protein sequence [NLA+02].

```
CYP1A2      MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
1           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
2           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
3           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
4           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
5           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
6           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
7           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
8           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
9           MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
10          MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
11          MALSQSVPFSATELLLASAIFCLVFWVLKGLRPRVPKGLKSPPEPWGWPLLGHVLTLGKN  60
            ************************************************************
```

Figure 8: MSA for the CYP1A2 gene

5.2 CYP2C9

Table 5 shows an overview from the Stargazer output for the gene CYP2C9. For the individuals homozygous for variant *2, this results in them having an intermediate phenotype. This is because the activity score of the variant is 0.5 and the combined activity score is therefore 1.0. The individuals who are heterozygous for the variant have a normal phenotype, because this individuals also have variant *1 with an activity score of 1.0. In figure 2 it can be seen that the most common variant in the CYP2C9 gene, is *2. This is also reflected in the table. The most common variant after that is *3. This variant does not occur in the individuals.

CYP2C9						
	Star allele for the 1st haplotype	Star allele for the 2nd haplotype	Activity score 1st haplotype	Activity score 2nd haplotype	Combined activity score	Phenotype
Human 1	*2	*2	0.5	0.5	1.0	Intermediate
Human 2	*1	*1	1.0	1.0	2.0	Normal
Human 3	*1	*1	1.0	1.0	2.0	Normal
Human 4	*1	*1	1.0	1.0	2.0	Normal
Human 5	*1	*2	1.0	0.5	1.5	Normal
Human 6	*1	*1	1.0	1.0	2.0	Normal
Human 7	*1	*1	1.0	1.0	2.0	Normal
Human 8	*1	*1	1.0	1.0	2.0	Normal
Human 9	*2	*2	0.5	0.5	1.0	Intermediate
Human 10	*2	*2	0.5	0.5	1.0	Intermediate
Human 11	*1	*2	1.0	0.5	1.5	Normal

Table 5: Variations, activity score and phenotype for the 11 humans for gene CYP2C9

Figure 9 shows that variant *2 is a missense variant with a high impact. In addition, it can be seen that the individuals also have many intron variants. The impact of these variants on the protein sequence is also unknown.
Figure 10 shows the missense variant *2. At position 144 amino acid R (arginine) has been replaced by amino acid C (cysteine).

5.3 CYP2C19

The most common variant in Europe in the CYP2C19 gene is *17. This variant does not appear in the individuals, which can be seen in table 6. The most common variant after that is variant *2. This variant does occur in individuals. The difference with variant *2 of the CYP2C9 gene and variant *2 of the CYP2C19 gene, is that variant *2 of the CYP2C19 gene has no activity. The individuals having variant *2 are heterozygous for the variant. In addition, they have variant *1 with an activity score of 1.0, therefore the individuals have an intermediate phenotype and no poor phenotype.
Star allele *2 has two variants. One variant is an intron variant that has a high impact, the other variant causes a splicing defect with a high impact. This can be seen in figure 11. The individuals have a synonymous variant in addition to the variants of *2. Similar to the CYP2C9 gene, individuals have many variations in the CYP2C19 region. Also with these variants, VEP gives the output that it is an intron variant or a non-coding variant.

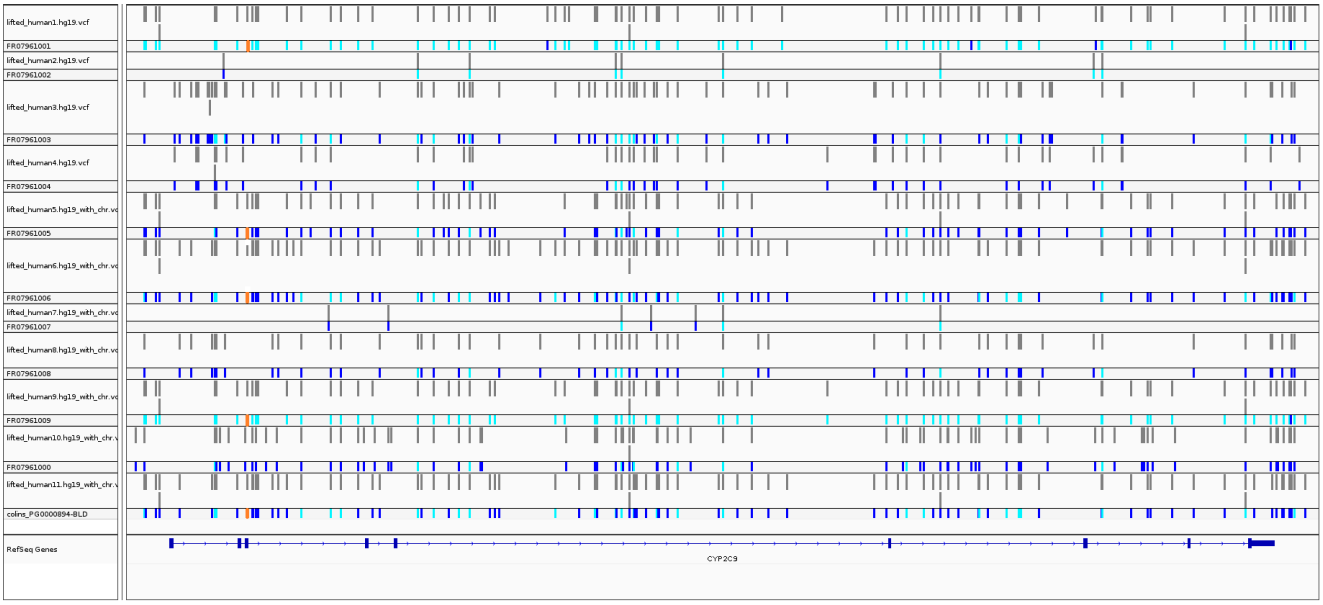

Figure 9: Types of variation in the CYP2C9 gene

```
CYP2C9      KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
1           KEIRRFSLMTLRNFGMGKRSIEDCVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
2           KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
3           KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
4           KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
5           KEIRRFSLMTLRNFGMGKRSIEDCVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
6           KEIRRFSLMTLRNFGMGKRSIEDCVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
7           KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
8           KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
9           KEIRRFSLMTLRNFGMGKRSIEDCVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
10          KEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
11          KEIRRFSLMTLRNFGMGKRSIEDCVQEEARCLVEELRKTKASPCDPTFILGCAPCNVICS  180
            ********************** ************************************
```

Figure 10: MSA for the CYP2C9 gene

CYP2C19						
	Star allele for the 1st haplotype	Star allele for the 2nd haplotype	Activity score 1st haplotype	Activity score 2nd haplotype	Combined activity score	Phenotype
Human1	*1	*1	1.0	1.0	2.0	Normal
Human 2	*1	*1	1.0	1.0	2.0	Normal
Human 3	*1	*2	1.0	0.0	1.0	Intermediate
Human 4	*1	*1	1.0	1.0	2.0	Normal
Human 5	*1	*1	1.0	1.0	2.0	Normal
Human 6	*1	*2	1.0	0.0	1.0	Intermediate
Human 7	*1	*1	1.0	1.0	2.0	Normal
Human 8	*1	*1	1.0	1.0	2.0	Normal
Human 9	*1	*1	1.0	1.0	2.0	Normal
Human 10	*1	*2	1.0	0.0	1.0	Intermediate
Human 11	*1	*2	1.0	0.0	1.0	Intermediate

Table 6: Variations, activity score and phenotype for the 11 humans for gene CYP2C19

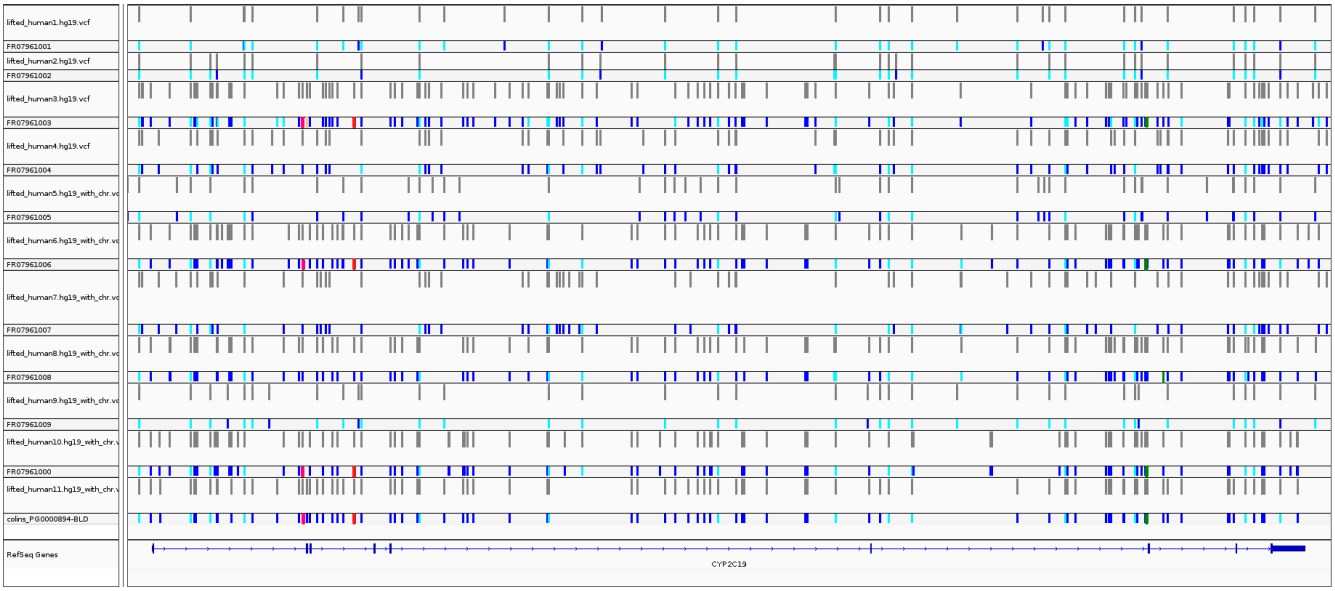

Figure 11: Types of variation in the CYP2C19 gene

Figure 12 shows that the protein seqeunce of variant *2 stops early. This is because the splicing defect variant causes a premature stop codon at postion 215 [dMWB+94].

```
CYP2C19     IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
1           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
2           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQ-------------------------   214
3           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQ-------------------------   214
4           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
5           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
6           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQ-------------------------   214
7           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
8           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
9           IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGTHNKLLKNLAFM   240
10          IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQ-------------------------   214
11          IIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQ-------------------------   214
            ********************************
```

Figure 12: MSA for the CYP2C19 gene

5.4 CYP2D6

The individuals have many different variants in the CYP2D6 gene, which can be seen in table 7. These variants causes three different phenotypes in the individuals. Variant *10 is the only variant in the CYP2D6 gene that has an adjusted activity score of 0.25. The individuals heterozygous for this variant have a normal pehenotype. The individual homozygous for variant *4 has a poor phenotype because the activity score of this variant is 0. Variant *2 has an activity score of 1.0, therefore the individuals heterozygous for this variant have a normal phenotype. The individual heterozygous for variant *6 and *9 has an intermediate phenotype, because the combined activity score is 0.5. Individuals heterozygous for variant *41 have a normal phenotype, because the combined activity score is 1.5. Variant *3 has an activity score of 0, therefore the individual heterozygous for this

variant has an intermediate phenotype.
As can been seen in figure 2, the most common variant in Europe for the CYP2D6 gene is variant *2. three out of eleven people have this variation. The next most common variants are *4, *3, *5, *41, *6 and *9. Most of these variants occur in the individuals. One interesting thing is that variation *10 occurs twice, but variation *10 hardly occurs in European people. But it is the most common variation in the East Asian population with almost 70%. The phenotype data of the PGP-UK does not provide the ethnic background.

CYP2D6						
	Star allele for the 1st haplotype	Star allele for the 2nd haplotype	Activity score 1st haplotype	Activity score 2nd haplotype	Combined activity score	Phenotype
Human 1	*1	*10	1.0	0.25	1.25	Normal
Human 2	*4	*4	0.0	0.0	0.0	Poor
Human 3	*2	*1	1.0	1.0	2.0	Normal
Human 4	*1	*2	1.0	1.0	2.0	Normal
Human 5	*6	*9	0.0	0.5	0.5	Intermediate
Human 6	*1	*1	1.0	1.0	2.0	Normal
Human 7	*1	*41	1.0	0.5	1.5	Normal
Human 8	*1	*41	1.0	0.5	1.5	Normal
Human 9	*1	*2	1.0	1.0	2.0	Normal
Human 10	*1	*10	1.0	0.25	1.25	Normal
Human 11	*1	*3	1.0	0.0	1.0	Intermediate

Table 7: Variations, activity score and phenotype for the 11 humans for gene CYP2D6

Figure 13 shows the different variants of the individuals. Variation *10 is a missense variant with a high impact. Variation *4 causes a splicing defect with a high impact. Variation *6 causes a frameshift with a high impact and variation *9 is a deletion with a moderate impact. Variation *41 causes a splicing defect with a high impact. Variation *2 has two missense variants with a low impact. Variation *3 causes a frameshift with a high impact. In addition to these variants, other variants also occur in the individuals. The impact of these variants are known, because they often occur in the sub star alleles.
Figure 14 shows the protein sequence from the different variants. For variation *10 the amino acid P (proline) is changed to S (serine) at position 34. Human 2 has a protein sequence that stops early and the end of the protein sequene is different. Human 2 has variation *4 which means a change in an intron and results in a different splice recognition site. This leads to a stop codon 38 amino acids after the variant [HKMG90]. Human 11 has variation *3 and the protein sequence also stops early. This is because of a frameshift which leads to a stop codon after the variation [KHK+90]. Human 7 and 8 have variation *41 which leads to a splice defect and results to an early stop codon [RAI04]. Variation *6 has a deletion of amino acid K (lysine) and varation *2 has two changes in amino acid. Amino acid R (arginine) is changed to a C (cysteine) at postion 296 and amino acid S (serine) is changed to a T (threonine) at postion 486.

5.5 CYP3A4

The most common variant in the CYP3A4 gene in Europe are *22, *3 and *2. As can be seen in table 8, these variants do not occur in the individuals. The individuals, except human 4 are

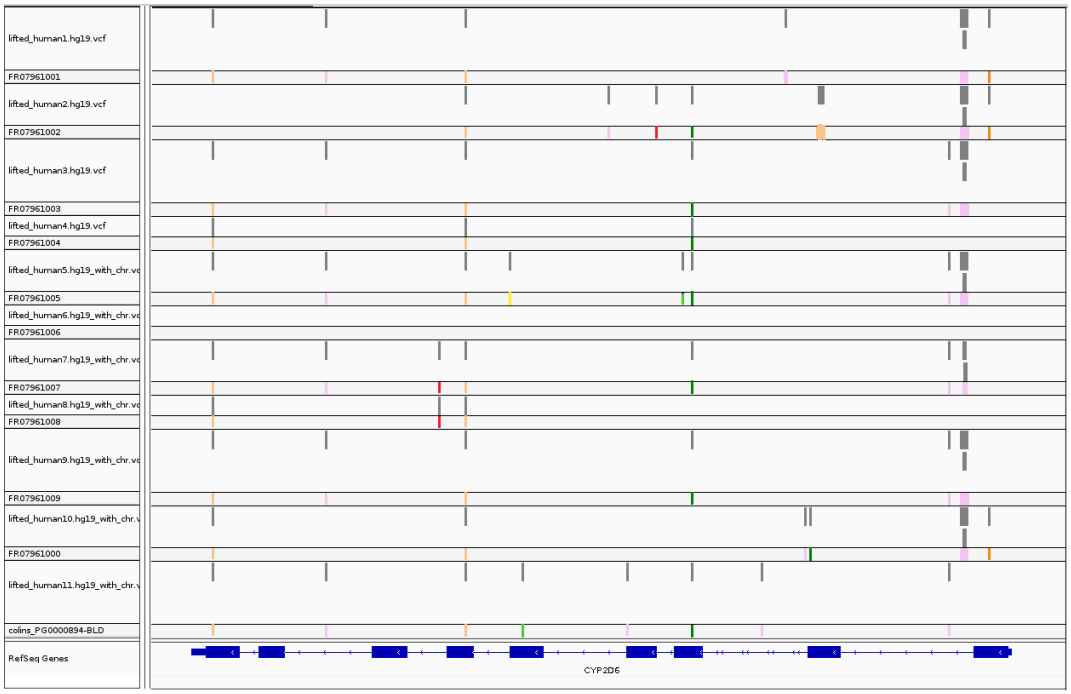

Figure 13: Types of variation in the CYP2D6 gene

heterozygous for variant *1A. The activity score for variant *1A is 1.0 which results in a normal phenotype.

CYP3A4						
	Star allele for the 1st haplotype	Star allele for the 2nd haplotype	Activity score 1st haplotype	Activity score 2nd haplotype	Combined activity score	Phenotype
Human1	*1	*1A	1.0	1.0	2.0	Normal
Human 2	*1	*1A	1.0	1.0	2.0	Normal
Human 3	*1	*1A	1.0	1.0	2.0	Normal
Human 4	*1	*1	1.0	1.0	2.0	Normal
Human 5	*1	*1A	1.0	1.0	2.0	Normal
Human 6	*1	*1A	1.0	1.0	2.0	Normal
Human 7	*1	*1A	1.0	1.0	2.0	Normal
Human 8	*1	*1A	1.0	1.0	2.0	Normal
Human 9	*1	*1A	1.0	1.0	2.0	Normal
Human 10	*1	*1A	1.0	1.0	2.0	Normal
Human 11	*1	*1A	1.0	1.0	2.0	Normal

Table 8: Variations, activity score and phenotype for the 11 humans for gene CYP3A4

Variant *1A is an intron variant with a low impact as can be seen in figure 15. This variant has no impact on the protein sequence [KNMa] as can be seen in figure 16.

5.6 CYP3A5

The most common variant in the CYP3A5 gene for Europe is *3. In table 9 it can be seen that this variation occurs in the individuals. The activity score for variation *3 is 0, therefore the individual

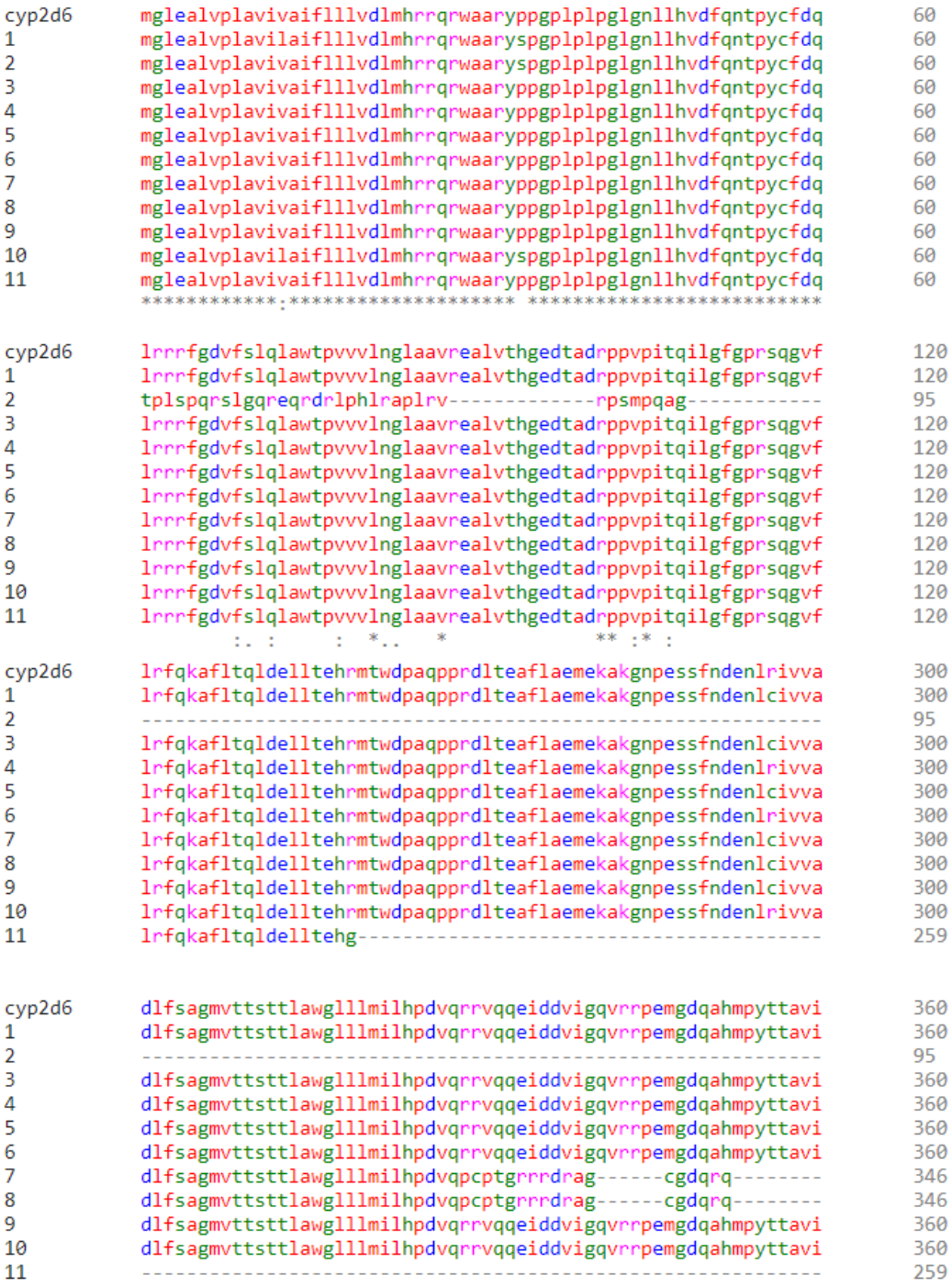

Figure 14: MSA for the CYP2D6 gene

with varation *1 and *3 has a intermediate phenotype. Variation *6 has also an activity score of 0. The individual with varation *3 and *6 has therefore a poor phenotype. Varation *6 is not common in the European people. Variation *6 is common in Africa. One individual has variation *2 which is a rare variation. The activity score of this variation is unknown. As described previously this person has not a poor phenotype and as long as there are no duplicates, no ultrarapid phenotype.

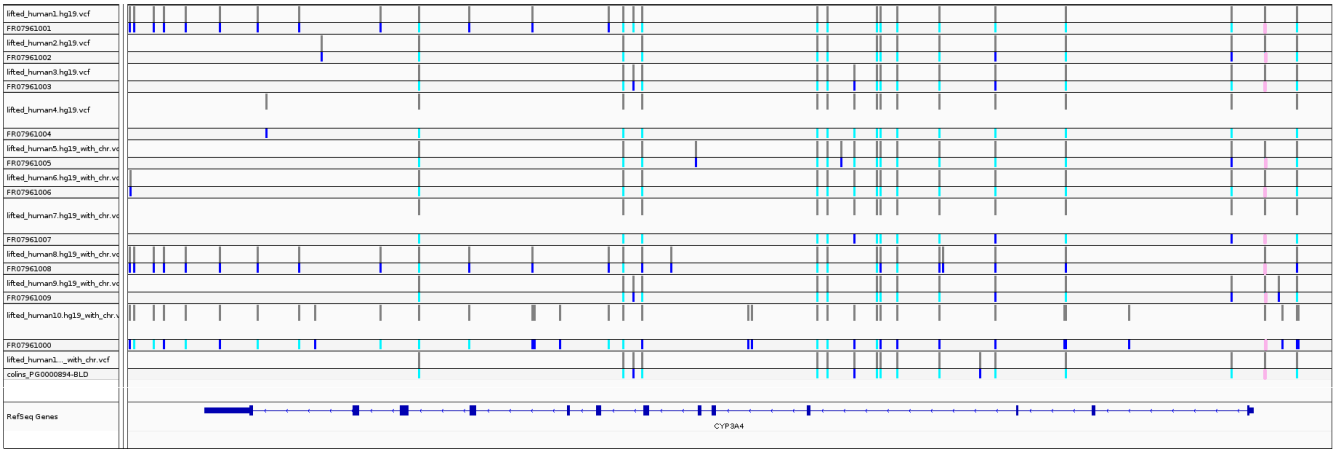

Figure 15: Types of variation in the CYP3A4 gene

```
CYP3A4      MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
1           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
2           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
3           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
4           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
5           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
6           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
7           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
8           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
9           MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
10          MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
11          MDVITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICV  240
            ************************************************************
```

Figure 16: MSA for the CYP3A4 gene

Figure 17 shows the different types of varation. Variation *2 is a missense variant with low

CYP3A5						
	Star allele for the 1st haplotype	Star allele for the 2nd haplotype	Activity score 1st haplotype	Activity score 2nd haplotype	Combined activity score	Phenotype
Human1	*1	*2	1.0	Unknown	Unknown	Unknown
Human 2	*1	*1	1.0	1.0	2.0	Normal
Human 3	*1	*1	1.0	1.0	2.0	Normal
Human 4	*1	*1	1.0	1.0	2.0	Normal
Human 5	*1	*1	1.0	1.0	2.0	Normal
Human 6	*1	*1	1.0	1.0	2.0	Normal
Human 7	*1	*1	1.0	1.0	2.0	Normal
Human 8	*1	*3	1.0	0.0	1.0	Intermediate
Human 9	*1	*1	1.0	1.0	2.0	Normal
Human 10	*3	*6	0.0	0.0	0.0	Poor
Human 11	*1	*1	1.0	1.0	2.0	Normal

Table 9: Variations, activity score and phenotype for the 11 humans for gene CYP3A5

impact. Variation *3 and *6 causes a splicing defect wiht a high impact.

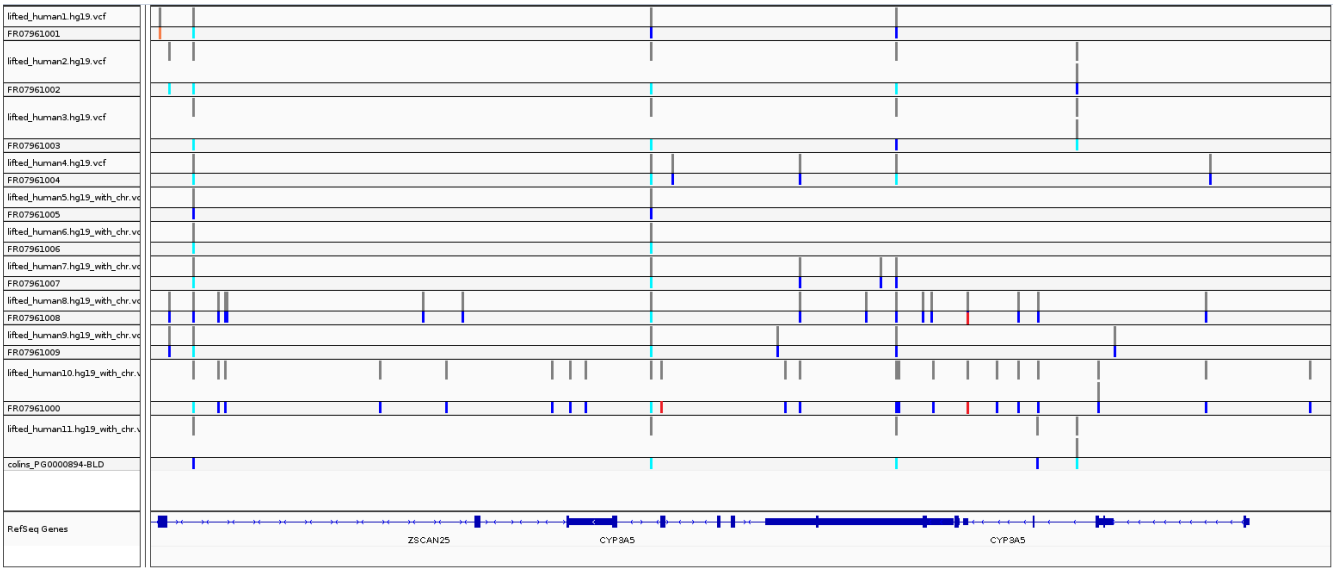

Figure 17: Types of variation in the CYP3A5 gene

As can be seen in figure 18, the protein sequence of variation *3 stops early. This is because the splicing defect generates a protein that is prematurely terminated at amino acid 109 [KZL+01]. Varation *6 misses a part of the protein sequence. This is because the splicing defect correlates with the deletion of exon 7 [KZL+01]. In variation *2, amino acid T (threonine) is replaced by amino acid N (asparagine) at postition 398.

```
CYP3A5      DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
1           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
2           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
3           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
4           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
5           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
6           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
7           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
8           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLG-----------      109
9           DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
10          DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
11          DTECYKKYGKMWGTYEGQLPVLAITDPDVIRTVLVKECYSVFTNRRSLGPVGFMKSAISL      120
            ************************************************

CYP3A5      MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
1           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
2           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
3           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
4           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
5           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
6           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
7           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
8           ------------------------------------------------------------      109
9           MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
10          ----------------------------LKFGFLDPLFLSIILFPFLTPVFEALNVSL      210
11          MDVITGTSFGVNIDSLNNPQDPFVESTKKFLKFGFLDPLFLSIILFPFLTPVFEALNVSL      240
```

Figure 18: MSA for the CYP3A5 gene

6 Discussion

As described in section 5, there are variants found in the CYP genes for the studied individuals. There is mainly variation in the CYP2D6 genes. These variants results that there are three different phenotypes in the studied individuals for a whole range of drugs. Most variants found in the individuals correspond to the most common variants in Europe. But not all the most common variants in Europe are found in the individuals. Two variants (CYP2D6*10 and CYP3A5*6) found are not common in Europe and also a rare variant (CYP3A5*2) has been found.
The two variants found in the individuals that are not common in Europe could be explained by the fact that the UK is a ethnically diverse nation. It is therefore quite possible that these individuals have an East Asian or African background. The ethnic background of the individuals is not described in the phenotype data of the PGP-UK data.

Some variants have an activity score that is unknown. It is therefore not possible to assign a phenotype to these individuals if the activity score is used. Because the individuals that are heterozygous for the variant also have variant *1, it can be said that the individuals can not have a poor phenotype. As long as the gene has no duplicates either, the individuals will not have an ultra rapid phenotype. It is important that these variants are researched to determine the activity of these variants. Without an activity score, no phenotype can be assigned to the individuals with this variant.
In addition, many variants have also been found in the gene region whose effect on the protein activity is unknown. In addition to the star alleles variants that the individuals have, these variants could also influence the activity score. For the CYP2D6 gene, these variants were described in sub star alleles, but not for the other CYP genes. More research needs to be done on these possible new star alleles, in order to conclude whether these variants have an impact on the protein activity. It could be that the individuals who have a normal or intermediate phenotype actually have a poor phenotype because they have an undescribed variant that causes reduced activity.

A combined activity score of 1, can be caused by various variations. For example an individual can be heterozygous for a variant with an activity score of 1 and be heterozygous for a variant with an activity score of 0. Or an individual can be homozygous for a variant with an activity score of 0.5. This means that the first individual has one functioning gene and one gene with no function. For the second individual it means two genes with a decreased activity. As mentioned earlier an activity score of 0.5 indicates decreased activity and not that the activity is half of a normal functioning allele. It has functional activity somewhere between no function and full function [GDJ+18]. So how much activity the protein has compared to a normal functioning protein is not known with an activity score of 0.5. There are still large intra- and inter-individual variability in CYP genes activity within a given genotype group [GDJ+18]. This could be due to the fact that an activity score of 1.0 can be obtained in different ways and the total activity is not the same. CPIC's drug prescription guidelines do not differentiate within a phenotype group. It should be investigated whether activity in a phenotype group is the same. If not, then there should be other guidelines for drug prescriptions that also take into account differences in activity in the same phenotype group.

As mentioned before, genome data is available for 118 volunteers. However, only 11 individuals have been looked at. In future work, the remaining individuals can be studied. This would

require changes to the data preprocessing and analysis workflow to accomplish. This is because the format used in the VCF files is different from the VCF files of the first 11 individuals. The data from these volunteers came later and perhaps a different method was used to obtain the VCF files. Although only 11 individuals were looked at, there is a lot of variation in the CYP genes.

7 Conclusion

So to conclude, even though I only looked at 11 people there were three different types of responders to a whole range of drugs. The consequence of these three different phenotypes is that the individuals must have different drug prescriptions from one another. The people with a normal phenotype can use normal drug prescriptions. But people with an intermediate phenotype will have to receive an adjusted drug prescription, because they do not metabolize the drugs quickly. The people with a poor phenotype will have to be prescribed a different drug because they don't metabolize some drugs. It is important that when people take drugs that are metabolized by these genes, they do a genotyping test to see what variations they have. More use should be made of genomics in all areas of healthcare. There should also be genotype tests not only for the most common variants, but also for variants that are less known. Because in these eleven individuals there is also an individual with the rare variant *2 in the CYP3A5 gene. This would not have been discovered with the Genelex genotype test. By continuing to research the variations in these genes and continually improve the genotyping tests, it will become a lot easier for us to get personalized medicines.

Appendix

```
#!/bin/bash

temp=1
temp2=1

#unzip the VCF files
for filename in *.gz;
do
 gzip -d $filename
done

#Liftover
for filename in *.vcf;
do
 CrossMap.py vcf b37tohg19.chain $filename ucsc.hg19.fasta lifted_human$temp.hg19.vcf
 temp+=1
done

#Stargazer
for filename in *.hg19.vcf;
do
 python3 stargazer.py genotype -o output_human$teller.hg19.CYP1A2 -d chip
 -t CYP1A2 --vcf $filename
 python3 stargazer.py genotype -o output_human$teller.hg19.CYP2C9 -d chip
 -t CYP2C9 --vcf $filename
 python3 stargazer.py genotype -o output_human$teller.hg19.CYPC19 -d chip
 -t CYPC19 --vcf $filename
 python3 stargazer.py genotype -o output_human$teller.hg19.CYP2D6 -d chip
 -t CYP2D6 --vcf $filename
 python3 stargazer.py genotype -o output_human$teller.hg19.CYP3A4 -d chip
 -t CYP3A4 --vcf $filename
 python3 stargazer.py genotype -o output_human$teller.hg19.CYP3A5 -d chip
 -t CYP3A5 --vcf $filename
 temp2+=1
done

#create tabix
for filename in *.hg19.vcf;
do
 bgzip -c $filename > $filename.gz
 tabix -p vcf $filename.gz
done
```

References

[AMMCHD13] BCOP Anne M. McDonnell, PharmD and BCPS Cathyyen H. Dang, PharmD. Basic review of the cytochrome p450 system. *Journal of the Advanced Practitioner in Oncology*, 4(4), August 2013.

[bLWP+18] Seung been Lee, Marsha M. Wheeler, Karynne Patterson, Sean McGee, Rachel Dalton, Erica L. Woodahl, Andrea Gaedigk, Kenneth E. Thummel, and Deborah A. Nickerson. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2d6 as a model. *Genetics in Medicine*, 21(2):361–372, June 2018.

[BZB18] Brian L. Browning, Ying Zhou, and Sharon R. Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, September 2018.

[CCGA+19] Olga Chervova, Lucia Conde, José Afonso Guerra-Assunção, Ismail Moghul, Amy P. Webster, Alison Berner, Elizabeth Larose Cadieux, Yuan Tian, Vitaly Voloshin, Tiago F. Jesus, Rifat Hamoudi, Javier Herrero, and Stephan Beck. The personal genome project-UK, an open access resource of human multi-omics data. *Scientific Data*, 6(1), October 2019.

[DDS13] Jurate Daugelaite, Aisling O' Driscoll, and Roy D. Sleator. An overview of multiple sequence alignments and cloud computing in bioinformatics. *ISRN Biomathematics*, 2013:1–14, August 2013.

[dMWB+94] S.M. de Morais, G.R. Wilkinson, J. Blaisdell, K. Nakamura, U.A. Meyer, and J.A. Goldstein. The major genetic defect responsible for the polymorphism of s-mephenytoin metabolism in humans. *Journal of Biological Chemistry*, 269(22):15419–15422, June 1994.

[ER04] William E. Evans and Mary V. Relling. Moving towards individualized medicine with pharmacogenomics. *Nature*, 429(6990):464–468, May 2004.

[Gae13] Andrea Gaedigk. Complexities ofCYP2d6gene analysis and interpretation. *International Review of Psychiatry*, 25(5):534–553, October 2013.

[GDJ+18] Andrea Gaedigk, Jean Dinh, Hyunyoung Jeong, Bhagwat Prasad, and J. Leeder. Ten years' experience with the CYP2d6 activity score: A perspective on future investigations to improve clinical predictions for precision therapeutics. *Journal of Personalized Medicine*, 8(2):15, April 2018.

[HKMG90] N Hanioka, S Kimura, UA Meyer, and FJ Gonzalez. The human cyp2d locus associated with a common genetic defect in drug oxidation: a g1934—-a base change in intron 3 of a mutant cyp2d6 allele results in an aberrant 3' splice recognition site. *American journal of human genetics*, 47(6):994—1001, December 1990.

[KHK+90] M Kagimoto, M Heim, K Kagimoto, T Zeugin, and U A Meyer. Multiple mutations of the human cytochrome p450iid6 gene (CYP2d6) in poor metabolizers of debrisoquine. study of the functional significance of individual mutations by expression of chimeric genes. *Journal of Biological Chemistry*, 265(28):17209–17214, October 1990.

[KNMa] KNMP. Algemene achtergrondtekst farmacogenetica – cyp3a4. Available at https://www.knmp.nl/downloads/g-standaard/farmacogenetica/cyp3a4-v2.pdf.

[KNMb] KNMP. General background text pharmacogenetics - cyp2d6. Available at https://www.knmp.nl/downloads/g-standaard/farmacogenetica/english-background-information/cyp2d6-english-2020.pdf.

[KZL+01] Peter Kuehl, Jiong Zhang, Yvonne Lin, Jatinder Lamba, Mahfoud Assem, John Schuetz, Paul B. Watkins, Ann Daly, Steven A. Wrighton, Stephen D. Hall, Patrick Maurel, Mary Relling, Cynthia Brimer, Kazuto Yasuda, Raman Venkataramanan, Stephen Strom, Kenneth Thummel, Mark S. Boguski, and Erin Schuetz. Sequence diversity in CYP3a promoters and characterization of the genetic basis of polymorphic CYP3a5 expression. *Nature Genetics*, 27(4):383–391, April 2001.

[LHW+09] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin and. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, June 2009.

[Li] Heng Li. Samtools. Available at http://www.htslib.org/doc/tabix.html.

[MGH+16] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(1), June 2016.

[NLA+02] Anna Nordmark, Stefan Lundgren, Birgitta Ask, Fredrik Granath, and Anders Rane. The effect of the CYP1a2 $*1f mutationonCYP1a2inducibilityinpregnantwomen.BritishJournalofClinicalPharmaco$ $504 - -510, November2002.$

[NTS+19] Charity Nofziger, Amy J. Turner, Katrin Sangkuhl, Michelle Whirl-Carrillo, José A.G. Agúndez, John L. Black, Henry M. Dunnenberger, Gualberto Ruano, Martin A. Kennedy, Michael S. Phillips, Houda Hachad, Teri E. Klein, and Andrea Gaedigk. PharmVar GeneFocus: CYP2d6. *Clinical Pharmacology & Therapeutics*, 107(1):154–170, December 2019.

[RAI04] S RAIMUNDO. A novel intronic mutation, 2988ga, with high predictivity for impaired function of cytochrome p450 2d6 in white subjects$*1.ClinicalPharmacology\&Therapeutics, 76(2) : 128--138, August2004.$

[RE15] Mary V. Relling and William E. Evans. Pharmacogenomics in the clinic. *Nature*, 526(7573):343–350, October 2015.

[RK11] M V Relling and T E Klein. CPIC: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3):464–467, January 2011.

[RTW+11] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, January 2011.

[SWD+11] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1):539, January 2011.

[Teaa] GATK Team. Depthofcoverage (beta). Available at https://gatk.broadinstitute.org/hc/en-us/articles/360041851491-DepthOfCoverage-BETA-.

[Teab] GATK Team. Grch37 hg19 b37 humang1kv37 - human reference discrepancies. Available at https://gatk.broadinstitute.org/hc/en-us/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-Reference-Discrepancies.

[teac] Genelex team. Genelex pgx testing. Available at https://www.genelex.com/test-menu/.

[TL07] AMY PRICE TOM LYNCH. The effect of cytochrome p450 metabolism on drug response, interactions, and adverse effects. *American Family Physician*, (3), aug 2007.

[Tut12] Yusuf Tutar. Pseudogenes. *Comparative and Functional Genomics*, 2012:1–4, 2012.

[WCMH+12] M Whirl-Carrillo, E M McDonagh, J M Hebert, L Gong, K Sangkuhl, C F Thorn, R B Altman, and T E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, October 2012.

[ZISL17] Y Zhou, M Ingelman-Sundberg, and VM Lauschke. Worldwide distribution of cytochrome p450 alleles: A meta-analysis of population-scale sequencing projects. *Clinical Pharmacology & Therapeutics*, 102(4):688–700, May 2017.

[ZSW+13] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, December 2013.