POPULATION SPECIFIC PHARMACOGENOMICS FOR PRECISION MEDICINE IN GLOBAL HEALTH

TIMOTHY MAN HAY LEE
School of Engineering (Computer Science), Stanford University
Stanford, CA, 94305 USA
timothyl@stanford.edu

ANGELA SHAN LI
School of Engineering (Biomedical Computation), Stanford University
Stanford, CA, 94305, USA
asli@stanford.edu

MAJED MOHAMED MAGZOUB
School of Engineering (Bioengineering), Stanford University
Stanford, CA, 94305, USA
mmagzoub@stnaford.edu

CHRISTINE YIWEN YEH
School of Humanities and Sciences (Biology), Stanford University
School of Medicine (Biomedical Informatics), Stanford University
Stanford, CA, 94305 USA
cyyeh@stanford.edu

Variabilities in individual responses to a given drug can result in health outcomes ranging from complete disease resolution to severe side effects and even mortality. Indeed, the one-drug-fits-all regime is neither effective for treatment of all populations nor cost-effective for health care systems. Some of the heterogeneity in individual patient responses to drugs can be characterized and explained by studying genetic variation across global populations. Therefore, it is necessary to systematically uncover population-specific genetic signatures in order to make informed prioritizations and recommendations of essential medicines.

The advent of sequencing technologies provides newly accessible genomic information from understudied populations. The combination of these data with pharmacogenomic data can associate population genetics to drug response, opening the door for informed drug prescription in global precision medicine. Separate studies have identified single drugs as unsuitable for patients with particular genetic variants. The newly available data will now allow us to systematically identify similar medicines - with high variability in drug response - on a high throughput level.

First, we identified pharmacogenomic SNPs from PharmGKB that are associated with drugs on the World Health organization's Essential List of Medicines. Next, we examined the variant data for these particular SNPs for the 26 subpopulations in the 1000 Genomes database. After that, we determined which of these SNPs are enriched in each subpopulation as compared to the global population. Finally, we used PharmGKB to match the clinical annotations for each SNP and come up with a finalized list of recommendations specific to individual drugs and populations. Ultimately, we believe that this study has the potential to combat poor health outcomes and relieve the economic burden that results from the current one-drug-fits-all model.

1. Introduction

1.1 Problem Statement

Adverse drug reactions are a major concern since they are neither ideal for patient health outcomes, nor cost-effective for global health care systems. In the UK it was reported one year that

6.5% of patients in 18,820 admissions to the National Health Service hospitals were caused by severe adverse drug reactions.[1] Similarly, in Japan, 33% of 3459 adult patients experienced seriously negative responses to drugs, with some reported as life threatening or even fatal.[2] This renders adverse drug reactions between fourth and eighth leading causes of death in many countries.[3]

Especially with the identification of pharmacogenomic biomarkers, there has been increasing evidence that genetic variants are associated with dosage dependent, or high-risk negative responses to drugs. It has also been observed through key examples that the prevalence of these pharmacogenomically significant genetic variants are population specific. For example, abacavir is a potent HIV drug deemed by the WHO as globally "essential". However, patients with the particular rs2395029 SNP exhibit symptoms such as fever, nausea, vomiting, and a life threatening rash in some cases.[4] This SNP occurs 2% in East Asian populations, 7% in Western Europeans and up to 20% in South Asian populations.[5] These numbers are not ideal for clinical practice nor for pharmaceutical drugs, especially under the current one-drug-fits-all regime that global health systems adopt.

Unfortunately, many other drugs such as abacavir exist - ones with severe drug effects associated with genetic variance. And with the current under-representation of nations in drug development, there may still be many other drugs on the market unknown to cause poor health outcomes specific to population-based variation. Hence, we identified that there is much need to uncover population-specific genetic signatures to make medical recommendations of essential medicines.

1.2 Opportunity

Our ability to determine population-specific recommendations of essential medicines depends on the fusion of three large datasets, described below. This intersection ultimately merges recent advances in pharmacogenomics with the abundance of genomic data from specific populations around the world. It also helps focus the recommendations on drugs that are most crucial to all health care systems.

1.2.1 Pharmacogenomics Knowledge Base (PharmGKB)

PharmGKB[6] is a web-based pharmacogenomic resource that matches clinical information on drug responses with curated knowledge on genetic variation. Specifically, PharmGKB now makes available variant annotations curated manually from high profile publications in the scientific public domain. These annotations are associations between a single variant and a drug phenotype from a single or multiple publications. Moreover, PharmGKB also offers clinical annotations built on multiple variant annotations - i.e genotype-based summaries of the clinical impact of a genomic variants. Many of these annotations include ones on polymorphisms that affect the differential expression of drug mediating enzymes, drug transporters, and drug target genes - explaining many of the variability to drug response due to genetic variation.

1.2.2 1000 Genomes Project

Finalized in 2015, data from the 1000 genomes project includes genomes of 2,504 individuals from 26 populations across the globe. These populations are also grouped into 5 global superpopulations - American (4), African (7), East Asian (5), South Asian (5), and European (5). Using this data we will identify which variants (SNPs) are enriched in a particular population as compared to the global average.

1.2.3 List of Essential Medicines from the World Health Organization (WHO)

The "WHO Model List of Essential Medicines" outlines a list of medicines required for a basic health-care system, optimized for cost, safety, and relevance to public health. Each drug listed has a list of possible dosage forms and age or weight restrictions. It is continually revised every two years to adapt to the continuously changing healthcare systems around the world. This list allows us to focus our list of recommendations on medicines that are most widely used around the world.

2. Methods

An overview of our methodology is as follows: first, we identified pharmacogenomic SNPs from PharmGKB that are associated with drugs on the World Health Organization's Essential List of Medicines. Next, we examined the variant data for these particular SNPs for the 26 subpopulations in the 1000 Genomes database. After that, we determined which of these SNPs are enriched in each subpopulation as compared to a global reference. Finally, we used PharmGKB to match the clinical annotations for each SNP and come up with a finalized list of recommendations specific to individual drugs and populations.

2.1 Extracting List of Drugs from the World Health Organization's List of Essential Medicines

2.1.1 Extracting and Parsing World Health Organization's List of Essential Medicines

By parsing the World Health Organization's "List of Essential Medicines", we were able to procure a list of drugs that are currently generalized to all populations and will have population-specific recommendations based on our analyses. The World Health Organization prepares the "List of Essential Medicines" document in a pdf format for distribution, but due to the unstructured nature of the presented data, we extracted this list from an Excel spreadsheet entitled "Comparative table of medicines on the WHO Essential Medicines lists from 1977 to present".[7] The excel format gave it more structure, allowing easier parsability. We extracted only those drugs that were found in the most current (19th) version of the list. In addition to the list of drugs, the functional category of each drug was also procured.

2.1.2 Intersection with PharmGKB Pharmacogenomically significant SNPs

We first obtained the "Variant, Gene and Drug Relationship data" from PharmGKB, a data file of all literature-curated pharmacogenomic relationships between types of variants, genes, drugs and diseases. We then used the binary search accession algorithm in the R package "data.table" to derive the variant-drug entries from this data. Next, we took this list and filtered down to entries that involved drugs from the WHO Essential Medicines.The PharmGKB ranked all relationship

entries as either “not associated”, “ambiguous” or “associated”, we extracted relationships that were either “ambiguous” or “associated” first to increase the sensitivity of our analytical model. This section resulted in a data table of SNP-drug relationships, accessible by SNP rs numbers and generic drug names.

2.2 Processing the VCF files from 1000 Genomes

2.2.1 Data Set Overview

The 1000 Genomes Data was gathered by sequencing of a diverse set of individuals representing 26 different populations. Individuals were sequenced by both whole-genome and targeted exome sequencing, with a mean depth of 7.4× and 65.7× respectively.[8] Corresponding to the number of times a nucleotide is read during the sequencing process, a depth this large increases likelihood of detection for low frequency variants. The 1000 Genomes Project presents its data in Variant Call Format (VCF files), which contain the genotype for each identified variant - or SNP - for each individual. The 1000 Genomes Consortium used the following workflow to generate their data: (1) align reads to a reference genome GRCH37 to get allele counts, (2) use counts and a combination of machine-learning classifiers and sequence analysis tools to generate the genotype and calculate likelihood of variation at each reported each locus, (3) filter out unlikely candidates. Our analysis made only use of only variants whose likelihood of assertion passed a phred quality score cutoff of 100.

The project identified 84.7 million single nucleotide polymorphisms, accounting for 80M of the 100M variants catalogued in The Single Nucleotide Polymorphism Database (dbSNP). Of the 40M novel variants identified 24% belong to South Asian populations and 28% to African populations, highlighting the projects unprecedented characterization of understudied populations.

2.2.2 Data Preprocessing

This data spanning these 84.7 million SNPs from 2506 individuals, was partitioned into vcf files for the 22 autosomal chromosomes, the sex chromosomes.[8] To clean the data we utilized VCFtools, a free programing program package designed for working with the complex genetic variation information in VCF files. Using the perl modules and a binary executables we transform the data such that we could feed into, statistical computing software, R. First we filtered the contents of VCF files by SNP specific rsid tags to only contain variants identified in our list as pharmacogenomically significant and associated with essential medicines (see section 2.1.2). All of the SNPs in our list passed quality cutoffs and contained no missing samples.

Next we merged our 23 filtered VCF files into one aggregate representing the genotype information for all our SNPs of interest from the 2504 individuals sequenced. We then transformed the contents of this format from VCF to Comma-Separated Values format. When transferring over the information we kept only info fields that that were relevant for our analysis, e.g. rsID, reference allele, alternate allele, and summary allele frequency.

After reading the data into R, we then split the set by subpopulation. Using supplementary sample info provided by the 1000 Genomes Project we made lists of all individual in each population and subset for our subpopulations. We now had 27 sets of data corresponding to the 26 subpopulations and a global aggregate.

2.2.3 Data Processing

In our preprocessed data genotypes were encoded as alleles values separated by "|", where the allele values are 0 for the reference allele and 1 for the alternate allele found. If the site is multiallelic alternate allele were recorded as numbers 2 and higher. Our variants were almost entirely biallelic, with only 91 entries (1 SNP, 1 individual) that did not match the reference nor first alternate allele. These entries were removed from our analysis set. Next we converted these genotypes from strings into integers 0, 1, or 2 corresponding to the number of non-reference alleles. To generate allele frequencies for each population we took the sum over all individuals - the total number of non-reference alleles in the population - and divided by 2 times the number of individuals - the total allele count. As a validation we compared the frequencies we calculated for our aggregate data set to the allele frequencies reported in the summary information in the original VCF files.

2.2.4 Bootstrapping

After processing was complete, we had between around 100 and 150 individuals for each of the subpopulations. In order to increase the power of our statistical tests, without sacrificing the specificity of our individual populations, we decided to use bootstrapping to increase the sample size of each population to 1000. We sampled with replacement from the pool of genotypes of existing individuals (0, 1, or 2) using the *sample* function in R. At the end of this step, each of the 26 populations had 1000 individuals (the number of original patients and bootstrapped samples combined).

2.3 Determining which SNPs were enriched in each subpopulation

With the genotypes from the sequenced genomes of the 1000 genomes project of each SNP that had a relationship with a drug and the subpopulation attribute of each patient, we used a series of statistical tests to characterize which SNPs were enriched in a population. Specifically, a SNP was enriched if the minor allele frequency of a subpopulation was statistically significantly higher than that of the global population. We simplified the model of the analysis by defining the minor allele (also called the non-reference allele) of the SNP as the critical allele of the genotype that the drug recommendation is based for. We used three different tests to characterize the relationship: chi-squared test, hypergeometric test, and bootstrapping to obtain a distribution of global minor allele frequencies.

2.3.1 Chi-Squared Test

For each SNP and each subpopulation, we generated a contingency table of allele counts from the set of alleles from all patients. The two categorical variables were whether the allele was the minor allele and whether the allele was from the subpopulation being tested. We used the chi-squared test of independence on this contingency table to determine whether these two categorical variables were independent from each other, meaning that the proportion of minor allele of a subpopulation does not have a significant difference with that of the rest of the population. The test uses a

chi-squared statistic, calculated from the normalized sum of squared deviations between the observed and expected frequencies (calculated with the marginal probabilities by assuming independence). This test assumes that the sample size is sufficiently large and the samples are picked randomly. To account for the case for when the minor allele frequency is actually depleted in the subpopulation, and the chi-square test concludes dependence between the variables, we used the Pearson residual of the observed number of minor alleles in the subpopulation. This measurement computes the difference between the observed count and the expected value, and if positive, the SNP is enriched in the population.

2.3.2 Hypergeometric Test

We used another test based on the hypergeometric distribution to answer whether an SNP was enriched in a population. The hypergeometric distribution models a scenario of sampling without replacement, where out of a finite population of size N and K "successes" (whether an individual is a success or not is a binary categorical variable), what is the probability that k "successes" are drawn from n samples. The corresponding test uses this distribution to calculate the statistical significance of drawing k "successes" In relation to our test. To use this test, we treated the subpopulation's alleles as a random sample of the alleles of the global population (where a success is defined as a minor allele). If found to be statistically significant, the subpopulation's alleles were not drawn from the same distribution as the global population, but in fact the proportion of minor alleles is overrepresented in the subpopulation. We use a one-tailed significance value of 0.05 to determine that the SNP is enriched in that subpopulation. The density function is as follows (Equation 1):

$$P(X = k) = \frac{\binom{K}{X}\binom{N-K}{n-x}}{\binom{N}{n}} \quad (1)$$

N = the total number of alleles from the global population
K = the number of minor alleles from the global population
n = the total number of alleles from the subpopulation
k = the number of minor alleles from the subpopulation

2.3.3 Bootstrapping

We used bootstrapping in order to estimate the distribution of global minor allele frequencies, using this to construct hypotheses test for the subpopulation minor allele frequency. To estimate this distribution for each SNP, we create bootstrap samples from the set of 2500 aggregated samples. Since each population had 1000 total individuals, we also took samples of size 1000 from the global aggregated set with replacement. For each sample, we calculated a global minor allele frequency. We then repeated this process 1000 times to create a distribution of the global minor allele frequency for that SNP. For each SNP and subpopulation, the hypothesis test is performed, comparing the subpopulation's minor allele frequency with the distribution of global minor allele frequencies. This technique is particularly useful for cases when the distribution of an estimator is intractable, and becomes asymptotically consistent, i.e. approaches the true distribution with

infinite bootstrap samples, and becomes asymptotically more accurate than distributions derived from assumptions of normality.

To account for multiple hypothesis testing, each p-value was converted to a q-value with a false discovery rate of 0.05. All significance tests were performed at a level of 0.05. The population specific pharmacogenomic SNPs that passed all three tests of significance were deemed as enriched in the subpopulation, and mapped back to their clinical annotations to create drug recommendations.

2.4 Mapping population-specific pharmacogenomic SNPs to their clinical annotations

In order to map population specific pharmacogenomic SNPs back to their clinical annotations, we used the Variant and Clinical Annotations data from PharmGKB. Regular expressions parsing techniques were used to process the data into a data table easily accessioned by drug chemical IDs and rs numbers. We used the list of SNPs from section 2.3 to extract SNP-drug pairs with clinical annotations.

3. Results

3.1 Pharmacogenomically SNPs associated with Essential Medicines

Out of the data 153,115 entries, we obtained 9,063 SNP-drug relationships using the binary search accession algorithm in the R package "data.table". Then, using the list of essential drugs processed from the WHO data, we were able to derive 4,294 SNP-drug relationships that had annotations pertaining to the drugs off the list of essential medicines. We extracted the entries that were "ambiguous" and "associated"; this yielded 3,976 SNP-drug relationships comprised of 1,010 unique SNPs.

3.2 Non-Reference allele frequencies in individual populations

After intersecting the List of Essential Medicines with the SNP-drug interaction pairs on PharmGKB, we found that there were 1,010 pharmacogenomic SNPs associated with drugs from the List of Essential Medicines. For each of these SNPs, we calculated and plotted a histogram of the non-reference allele frequency for all SNPs for each population. **Figure 1** shows such a histogram from a population of Mexican ancestry and another of Nigerian ancestry. As shown in the figure, for many of the pharmacogenomic SNPs, the non-reference allele actually appears quite often within a population. Although the non-reference allele is not always the allele that is associated with disease / drug, these distributions show that the allele that is indicative of drug response can occur quite often in different populations. This ensures that certain recommendations have the potential of having widespread impact.

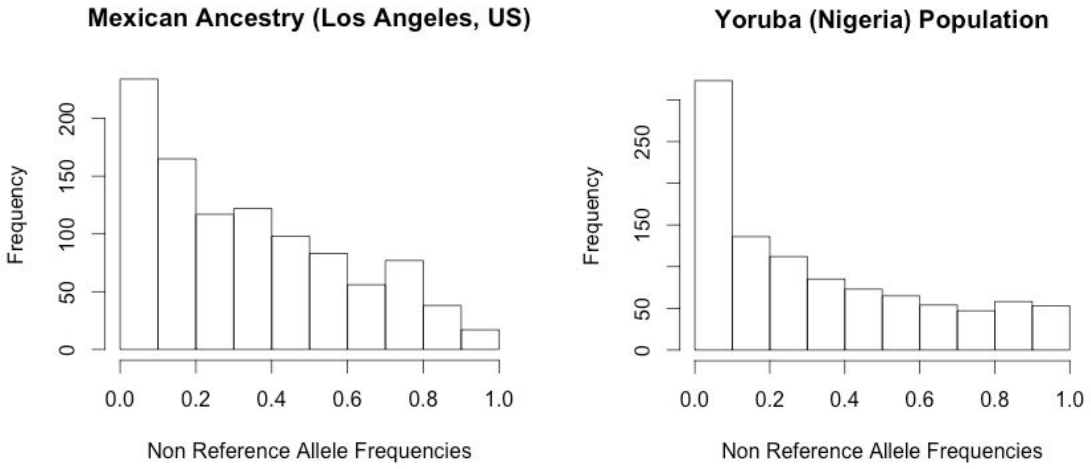

Figure 1 : Histogram of non-reference allele frequencies from two different subpopulations (Mexican and Yoruban)

3.3 Approaches to obtain enriched SNP - subpopulation pairs

In the end, we obtained 6778 instances for when a SNP was deemed to be enriched in a subpopulation by all three statistical tests: Chi-square test of independence, hypergeometric test, and bootstrapping. **Figure 2** shows the intersected number of enriched SNP - subpopulation pairs agreed by the tests. There are 2739 instances of significance deemed only by the test derived by bootstrapping. Many of these may be false positives from using an insufficiently well-approximated distribution of global allele frequencies. To counter this, if more computing power and time were available, complete enumeration of the possible bootstrap samples would decrease the bootstrap resampling variability, and have a better approximation of the distribution.

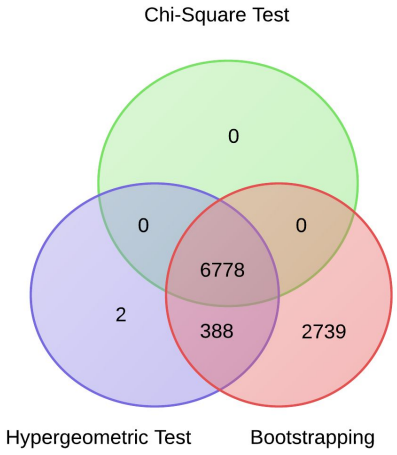

Figure 2 : Venn diagram of number of enriched SNP - subpopulation pairs as determined by each test

3.4 Number of Significant Pharmacogenomic Recommendations In Each Population

A “significant pharmacogenomic recommendation” in this study is defined as a SNP-drug relationship wherein the SNP is identified as enriched in any subpopulation, the drug is listed on the WHO Essential Medicines list, and the relationship has a specific clinical annotation on drug dosage, efficacy or toxicity. **Figure 3** shows the number of recommendations obtained per subpopulation.

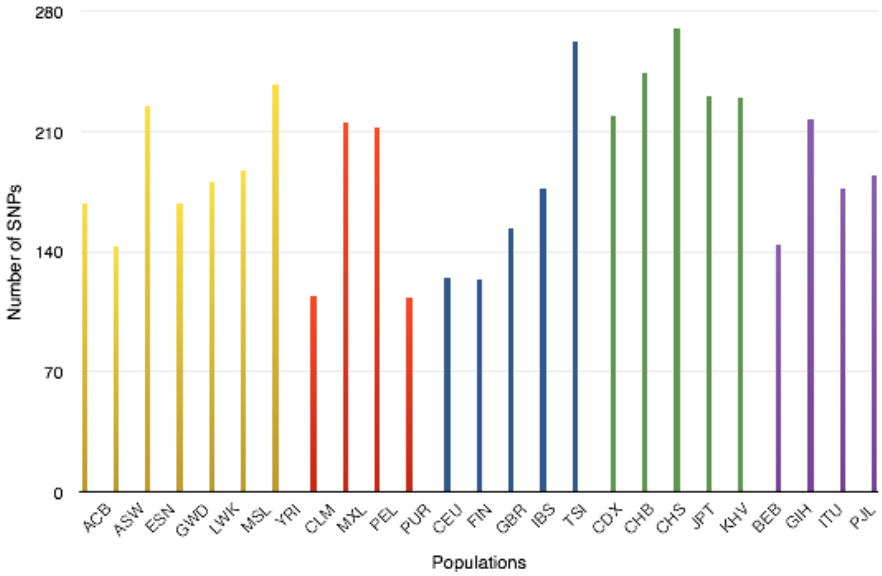

Figure 3 : Number of pharmacogenomically significant recommendations extracted from analytical pipeline

3.5 Examples of recommendations

The full collection of recommendations is included in the submission along with this manuscript under the file name of “resulting_recommendations_conglomerate.RData”

3.5.1 Warfarin in African subpopulations

All subpopulations in the super population “AFR: African Descent” yielded significant SNP-drug recommendations. The numerous SNP-drug recommendations for warfarin were conflicting in subpopulations YRI, MSL, ESN and ACB, with some enriched SNPs indicating both need for increased dosage and decreased dosage for optimal administration of warfarin. However, in subpopulations LWK, GWD and ASW, all SNP-drug recommendations agreed that more so in these specific populations, patients would most benefit from a higher dosage of warfarin for effective treatment.

Table 1 : Subpopulations in the African superpopulation yield SNP-drug associations to Warfarin. SNP-drug recommendations only agree for LWK, GWD and ASW populations for need of increased dosage of warfarin

Subpopulation Code	Subpopulation Full Name	PharmGKB Conglomerate Recommendation

YRI	Yoruba in Ibadan, Nigeria	Ambiguous
LWK	Luhya in Webuye, Kenya	associated with increased dose of warfarin
GWD	Gambian in Western Divisions in the Gambia	associated with increased dose of warfarin
MSL	Mende in Sierra Leone	Ambiguous
ESN	Esan in Nigeria	Ambiguous
ASW	Americans of African Ancestry in SW USA	associated with decreased dose of warfarin
ACB	African Caribbeans in Barbados	Ambiguous

3.5.2. Recommendations from enriched SNPs that were unique to one subpopulation

Some drug recommendations were unique to particular subpopulations, the two most stark examples are summarized in Table 2 below. The first example is combination therapy of amitriptyline and clomipramine: antidepressant drugs used for patients that are clinically depressed. In the Utah Residents (CEPH) with Northern and Western Ancestry subpopulation (CEU), it was observed that the one particular SNP, rs3892097 (GrCh37) associated with lower dosage of the two drugs are enriched in this population. This allele is observed at 0.0931 frequency in the global population according to 1000 genomes, but at 0.5034 in the CEU population.

The second example is the use of carbamezapine, a drug primarily used in the treatment of epilepsy and neuropathic pain, in the Han Chinese in Beijing population. SNP rs17183814 is associated with complete resistance to carbamezapine and occurs at a frequency of 0.0587 in the global population, but at 0.1707 in the CHB population.

Table 2 : Examples of subpopulation specific recommendations

Subpopulation Code	Subpopulation Full Name	PharmGKB Clinical Recommendation	SNP	Global Minor Allele Frequency	Minor Allele Frequency in Subpopulation
CEU	Utah Residents with Northern and Western Ancestry	associated with decreased dose of amitriptyline and clomipramine	rs3892097	0.0931	0.5034
CHB	Han Chinese in Bejing, China	associated with resistance to carbamazepine	rs17183814	0.0587	0.1707

4. Conclusions and Future Directions

4.1.1 Statistical Models able to reproduce Population Specific SNPs

From the three statistical models we implemented, we obtained "population-specific" SNPs. From the meta-analysis of our statistical methods, it appeared that Chi-Squared non-parametric test was the most stringent - perhaps because we refrained from assuming a prior distribution. It should be

noted that all three tests were able to reproduce the 6778 SNPs identified in the chi-square test - suggesting reproducibility of the 3 models. These statistical attempts however, exhibited that there may still be high variability in statistical definitions and stringency on the mathematical definitions of "population-specific" SNPs. Previous work has reported the use of chosen cutoffs in frequency in the global population versus specific subpopulations. (8)In this body of work, we present three statistical models that may be more robust in identifying population-specific SNPs without setting arbitrary cutoffs. This is, at least in part, supported by our analytical pipeline's ability to validate medical recommendations already known in the literature.

However, we have defined "population specific SNPs" as ones that appear more statistically frequently in particular populations than in the global reference. In continuing work, we aim to also use pair-wise comparisons between populations or comparisons of each subpopulation to a specific reference population in order to obtain variations on "population specific SNPs" to use in our analytical pipeline. This will henceforth allow us to consolidate our definition of "population-specific" variation, bringing us closer to biological confidence in our recommendations.

4.1.2 Analytical Pipeline uncovers known population specific pharmacogenomic recommendations

It has been long known that genetic variants affect the pharmacology of warfarin. It has been reported, especially, that SNP variants that fall in the gene loci of *CYP2C9* and *VKORC1* affect the dosage requirements in patients. [9] Some studies have performed pharmacogenomic analysis on these particular variants. Many population association studies have reported that SNP variants more prevalent in the African superpopulation (AFR) are associated with the clinical observation that a higher percentage than expected of patients require much higher dosage for pharmacogenomic efficacy. [10]

In our study, we were able to support these conclusions in the literature. In all subpopulations that fall under the AFR superpopulation, we obtained numerous SNPs associated with dosage requirements with regard to warfarin. It was clear that many SNPs associated with clinical response to warfarin appeared at higher frequency within these 7 population. It was interesting to observe however, that for some of these subpopulations, the recommendations were conflicting whilst in others, recommendations were unanimous in confirming higher dosage needed for their respective subpopulations. These results have shown our analytical pipeline's ability to identify pharmacogenomically significant SNPs already known to be specific to the AFR super population. What we have exhibited further however, is refinement of that pharmacogenomic recommendation to evaluate the specific significance in smaller subpopulations - bring us one step more precise in precision medicine.

4.1.3 Novel pharmacogenomic medical recommendations uncovered

Through our analytical model, we were also able to obtain recommendations unique to particular subpopulations. Although these hits need to be validated, we are hopeful that these results (**Table 2**) may potentially provide useful information on the genetic landscape of particular populations with regard to variation in patient response to essential drugs. In this study, we have taken advantage of the "one sentence summary" feature in the clinical annotations from PharmGKB. We

hope to potentially glean more information from the manually curated data in order to make more refined and specific medical recommendations.

4.1.4 Data Visualization and User Interface

After we have optimized and consolidated our results, we hope to be able to visualize all the data and store them in a way that is user friendly and accessible to physicians and researchers around the world.

5. Acknowledgments

We wanted to thank the BMI 212 teaching team for all of their support and help this quarter, and we also wanted to thank Dr. Howard McLeod, our mentor at PGENI (Pharmacogenomics for Every Nation Initiative) who, along with Professor Altman, inspired us to pursue this project.

References

1. Pirmohamed, M. et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ 329, 15–19 (2004).

2. Morimoto, T. et al. Incidence of adverse drug events and medication errors in Japan: the JADE study. J. Gen. Intern. Med. 26, 148–153 (2011).

3. Lazarou, J., Pomeranz, B. H. & Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 279, 1200–1205 (1998).

4. Martin, A. M. et al. Predisposition to abacavir hypersensitivity conferred by HLA-B*5701 and a haplotypic Hsp70-Hom variant. Proc. Natl. Acad. Sci. 101, 4180–4185 (2004).

5. Pai, S. A. & Kshirsagar, N. A Critical Evaluation of Pharmacogenetic Information in Package Inserts for Selected Drugs Marketed in India and its Comparison with US FDA Approved Package Inserts. J. Clin. Pharmacol. (2016). doi:10.1002/jcph.720

6. McDonagh, E. m, Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. Enabling the curation of your pharmacogenetic study. Clin. Pharmacol. Ther. 97, 116–119 (2015).

7. Xls file found in http://www.who.int/entity/medicines/publications/essentialmedicines/EMLsChanges1977_2011.xls?ua=1

8.1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74.

9. Fung, Erik, Nikolaos A. Patsopoulos, Steven M. Belknap, Daniel J. O'Rourke, John F. Robb, Jeffrey L. Anderson, Nicholas W. Shworak, and Jason H. Moore. "Effect of genetic variants,

especially CYP2C9 and VKORC1, on the pharmacology of warfarin." In *Seminars in thrombosis and hemostasis*, vol. 38, no. 8, p. 893. NIH Public Access, 2012.

10. Scott, S. A., Jaremko, M., Lubitz, S. A., Kornreich, R., Halperin, J. L., & Desnick, R. J. (2009). CYP2C9* 8 is prevalent among African-Americans: implications for pharmacogenetic dosing. *Pharmacogenomics*, *10*(8), 1243-1255.