Exploring the Potential of Direct-To-Consumer Genomic Test Data for Predicting Adverse Drug Events

Patrick Zhang

Brown University, Providence, RI

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the School of Engineering at Brown University

Providence, Rhode Island

May 2018

This thesis by Patrick Zhang is accepted in its present form by the School of Engineering as satisfying the thesis requirements for the degree of Master of Science.

Approved by the Graduate Council

Date _____ _____

Dr. Neil Sarkar, Advisor

Date _____ _____

Dr. Vicki Colvin

Date _____ _____

Dr. Anubhav Tripathi

Acknowledgements

I would like to thank Dr. Neil Sarkar and the Brown School of Engineering for advising and assisting me in the completion of this thesis.

This work was funded in part by National Institutes of Health grants R01LM011364, R01LM011963, and U54GM115677. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Table of Contents

List of Tables

List of Figures

Introduction

Adverse drug events (ADEs) are a significant challenge facing global public health. In addition to causing clinical harm to patients that can range from a minor headache to death, ADEs represent an economic burden in the form of hospital admissions, prolonged stays, and additional treatments. Although exact figures vary between sources, the estimated annual cost in the United States alone is $30.1 billion, where each incident can amount to $3,000 depending on severity[1,2]. Spontaneous reporting systems are valuable resources for post-market pharmacovigilance, presenting collected reports of ADEs and medication errors that are voluntarily submitted by clinicians, healthcare facilities, and patients. The US Food and Drug Administration (FDA) adverse event reporting system (FAERS) is one such database. As with all spontaneous reporting systems, limitations exist for FAERS; a report for a drug and adverse effect does not necessarily demonstrate a causal relationship between them, and not all adverse events for a particular drug may be reported[3]. A recently curated and standardized version of FAERS has been made publicly available, called the Adverse Event Open Learning through Universal Standardization (AEOLUS). Using data originating in FAERS, AEOLUS provides standardized data and correlative statistics about drugs administered for an indication and the adverse outcomes[4].

While an estimated 50% of ADEs are preventable and result from a medication error, many are the result of other factors like genetic variations that lead to a heightened drug sensitivity. Pharmacogenomics focuses on the study of pharmacology in the context of genetics, aiming to develop therapies that maximize efficacy and minimize risk of ADEs. The Pharmacogenomics Knowledge Base (PharmGKB)[5], developed by the Pharmacogenomics Research Network, facilitates exploring the effect of genetic variation on drug response. Many of the examined variants are single nucleotide polymorphisms (SNPs), which are the focus of this study, and SNPs associated with an increased risk of ADEs are of particular interest. The data contained within PharmGKB are the product of utilizing natural language processing techniques on clinical studies from PubMed and verification through manual curation. The vocabulary for reference SNP cluster IDs (RSIDs) and drugs are primarily standardized through dbSNP[6] and DrugBank[7]. An annotation for a given variant indicates that a peer-reviewed article exists containing an association between a gene, drug, and disease. A Singaporean

study evaluated the prevalence of hospitalizations related to ADEs and observed that around 30% of ADEs present at admission were caused by drugs with PharmGKB annotations[8].

The role of pharmacogenomic tests in clinical practice has been expanding, with recent advances in technology that make genome-wide studies both economically and practically feasible. Several publications have demonstrated the potential clinical utility of pharmacogenomic tools, which reduced re-hospitalizations and lowered health care costs by 84% compared to controls[9,10]. Clinical applications of personalized medicine are partially limited by the current shortcomings of genome-wide association studies that include difficulty obtaining large sample sizes, particularly for minority populations[11]. Short of genomic testing being available and affordable to inform every clinical encounter, direct-to-consumer (DTC) genetic testing is an emerging technology that has the potential to fill the gap of available genomic data. For instance, 23andMe is a leading producer of DTC tests and provides consumers with their genetic information without the need for a healthcare professional. This information includes inherited variants associated with risk factors for conditions and hypersensitivities to drugs. Earlier this year, 23andMe received FDA approvals to market their tests to assess the genetic risk for breast cancer and for ten diseases including Parkinson's and late-onset Alzheimer's disease[12]. Among other limitations, however, its genetic tests do not detect all relevant mutations[13]. Its intended use is to prompt counseling from healthcare professionals rather than as a diagnostic tool, but the wealth of information that this accessible technology can provide may have promising utility for research purposes. 23andMe variant profiles for over 777 individuals are currently available at the Harvard Personal Genome Project (PGP)[14].

This exploratory study focused on two objectives. The first was to examine the clinical relevance of DTC reports through a thorough examination of the population data. This was in part achieved by obtaining clusters within the population using unsupervised learning methods. The secondary focus of this study was to explore the variants themselves, considering both locations within the genome and prevalence across DTC reports. While not diagnostic, 23andMe data might serve as an additional source of patient-supplied information for the prescription of drugs and prompt additional, clinical-grade genetic tests when necessary. In particular, co-

occurring variants were used to discover pharmacogenomic associations that may guide future research directions or act as a separate resource by creating a networks of drugs and adverse events.

Methods

Exploration of PGP Patient Profiles

PGP reports were downloadable as text files, containing data in columns corresponding to the RSID, chromosome, position number, and affected alleles. Less common file types like vcf and bam were converted to csv or txt when appropriate. When participants uploaded multiple files, the most recent report was used. In some cases, the most recent profiles included fewer variants than the originals, so in these situations the more complete upload was selected. Six participants uploaded reports from Gene for Good, and eight reports were generated from AncestryDNA, which were kept only if standard 23andMe profiles were not available.

To better describe the population of reports from PGP, the contents were compared, specifically attempting to find subgroups within the population. The approach presented here was to form groups by population as defined by the 1000 Genomes Project. Allele frequencies by population for each variant were scraped from PharmGKB, and an individual was assigned to one of 5 aggregate and 26 individual populations by examining the genotypes of all variants within his/her genetic report. The aggregate populations were African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). Individual populations were subpopulations within the defined aggregate ones (e.g. Northern Europeans from Utah, or CEU, was a subpopulation of EUR). The preliminary classification model used in this study was Naive Bayes, and the results of this model were compared to clusters formed using principal component analysis (PCA). PCA is a dimensionality reduction algorithm that obtains a new set of uncorrelated values from a set of observations through linear transformations of the original features. The variants within genetic reports were grouped by chromosome, and for each chromosome, the genotypes were value encoded for all individuals and saved as a separate file. For selected chromosomes, clusters were presented in a scatter plot using the first two principal components, where the points were colored by classified populations.

PharmGKB: Variant-drug Associations

Only variant-drug associations previously identified to increase the risk for ADEs were included in this study. The Variant and Clinical Associations file from PharmGKB included this information and referred to the type of association as a phenotype category. The categories were either one or combinations of the following: toxicity, efficacy, dosage, metabolism/PK, and other. Each variant-drug pair additionally contained a summary of the publication's findings. Annotations involving RSIDs were filtered for significance, and the annotations were divided into those that indicated increases and decreases in the incidence of a particular event. First, only the toxicity associations were examined. Language patterns of the summaries were analyzed and a series of regular expressions were devised to obtain a vocabulary for relevant outcomes; for example, outcomes usually were found between the phrases "risk/severity/likelihood of" and a conditional word like "when." Outcomes like "dose reduction" or "non-response" were examples of those mislabeled as toxicity. This vocabulary was then applied to classify toxicity annotations hidden within other categories.

Next, the effect of genetic variation on a drug was summarized. For this study, a group of 42 drugs that treat mental health or psychiatric disorders was selected and obtained from Drugs.com[15]. Localization of ADE-related variants was examined by first joining the Variant and Clinical Associations file with data that included spatial information from dbSNP. Gene maps were scraped from the Online Mendelian Inheritance in Man (OMIM), which contained both cytogenetic locations and positions according to the Genome Reference Consortium human genome (build 37 or GRCh37). Location densities were plotted for selected chromosomes and drugs to visualize regions in which ADE-related variants were found.

Using PGP to Find Associated Variants and Drugs Within a Population

Variants commonly present in individuals within a population were found in this study by matching the 23andMe patient profiles to the annotated PharmGKB RSIDs. The results were one-hot encoded with columns representing RSIDs and rows representing individuals. From this, co-occurring variants were examined by grouping subsets of reports that contained a particular variant. RSIDs were considered related in a particular group if they were present in 90% of the subset, an arbitrary significance level. Overall association between variants was determined by examining all subgroups.

Briefly, for each annotated RSID, the related variants were found. Within these related variants, those that had annotations for the drug of interest were obtained, and lastly with these variants, the occurrence of all drugs were counted. The fraction of RSIDs associated with a given drug was calculated, and higher values suggested that the drug should be closely related to the drug of interest. Since the type of association between drugs captured using this methodology was unknown, multiple means of validation were necessary. DrugBank included a list of interacting drugs, which was compared to the drugs indicated by genetic association. AEOLUS provided similar information on interactions in the form of case reports. Within a case, a drug may be considered a primary suspect, secondary suspect, concomitant, or interacting drug. For cases where the drug of interest was the primary suspect, all interacting drugs were obtained. For cases where the drug of interest was an interacting drug, the primary suspect was considered an interacting drug.

Mapping Variants to Adverse Events

Adverse event information from PharmGKB and AEOLUS was used to map variants to adverse events. AEOLUS contains reports of ADEs without indications of genetic association. For a particular drug, outcomes from all cases in which it was the primary suspect and only administered drug were examined. Edges were drawn between outcomes within a case and represented as pairs of associated ADEs, and the total occurrence of these pairs reflected the degree of correlation. Some of these relationships contained outcomes included in PharmGKB associations, which facilitated a partial mapping of RSIDs directly to ADEs. These annotated ADEs were manually curated for the drug of interest from the annotations file and standardized to the vocabularies used in the AEOLUS database.

Gephi Visualization

Gephi is a visualization software for networks[16]. The application has several force-directed layout algorithms to distribute nodes and edges, and the layouts used in each graph were chosen to most clearly display clusters and associations. The Fruchterman-Reingold layout models nodes and edges with attractive and repulsive forces and distributes them such that the overall energy of the system is minimized. In the force atlas layout, nodes and edges similarly repulse and attract, but the layout favors bringing poorly connected nodes closer to very connected nodes. Lastly, the Yifan Hu layout combines standard methods but treats clusters as a single

node when determining repulsive forces. Gephi offers several statistical measures to assist in quantifying node associations. The modularity of a network measures the connectedness of a graph and its calculation involves comparing the density of edges within a community and those between communities. Communities are indicative of some form of association[17].

Relationships between drugs were visualized using Gephi with edge weights represented by the ratio of associated RSIDs. The classes of drugs were defined using PubChem[18] and the standard anatomical therapeutic chemical (ATC) classification system obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG)[19]. The nodes were colored by drug class or by modularity to best display associations; the node sizes were proportional to its degree, or number of incoming and outgoing edges. Associations between ADEs were visualized using a similar approach, where nodes and edge weights represented outcomes and the correlations between them. The ADEs found in PharmGKB annotations were labeled on the visualization, and the results were filtered by edge weight, using the Yifan Hu layout in combination with the force-directed algorithms. Similarly, the nodes were colored by modularity and sized proportional to its weighted degree.

Results

DTC Genomic Reports

777 DTC reports were downloaded from 23andMe or other sources and used for the analysis. The number of RSIDs in each report ranged from 546,058 to 1,003,774. 89.5% of the 1,586 unique annotated RSIDs were among the variants detected by the 23andMe genetic tests, and 410 out of 466 variants on VIP genes (88.0%) were contained. 497 out of the 523 unique drugs (95.0%) had at least one associated variant within the reports.

To explore groups within all reports, individuals were classified by aggregate and individual populations. A majority of individuals were classified as EUR (95.1%), while AMR and EAS each constituted 1.54% of the sample and AFR and SAS each were 0.09%. The total counts for the individual populations grouped by aggregate population are shown in Figure 1a along with the population definitions. Figure 1b shows scatter plots obtained by plotting the first two principal components after using PCA. The clustering method was

applied to variants on chromosomes 1, 2, 11, and 19 and on a combination of those chromosomes. Chromosome 1 had the greatest number of variants at 154,024, which was ultimately projected to two dimensions that captured approximately 61.0% of the cumulative variance.

a

Aggregate	Individual	
AFR	ACB	6
	ASW	1
AMR	BEB	1
	CLM	6
	MXL	3
	PUR	2
EAS	CHB	4
	CHS	6
	JPT	1
	KHV	1
EUR	CEU	262
	CLM	4
	FIN	9
	GBR	232
	IBS	123
	PUR	11
	TSI	98
SAS	GIH	3
	ITU	1
	PJL	2
	STU	1

Individual Population	Population Definition	Aggregate Population
CHB	Han Chinese in Beijing, China	EAS
JPT	Japanese in Tokyo, Japan	EAS
CHS	Southern Han Chinese	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
CEU	Northern Europeans in Utah (CEPH)	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian Population in Spain	EUR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Caribbeans in Barbados	AFR
MXL	Mexican Ancestry from LA, USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
GIH	Gujarati Indian from Houston, Texas	SAS
PJL	Punjabi from Lahore, Pakistan	SAS
BEB	Bengali from Bangladesh	SAS
STU	Sri Lankan Tamil from the UK	SAS
ITU	Indian Telugu from the UK	SAS

b

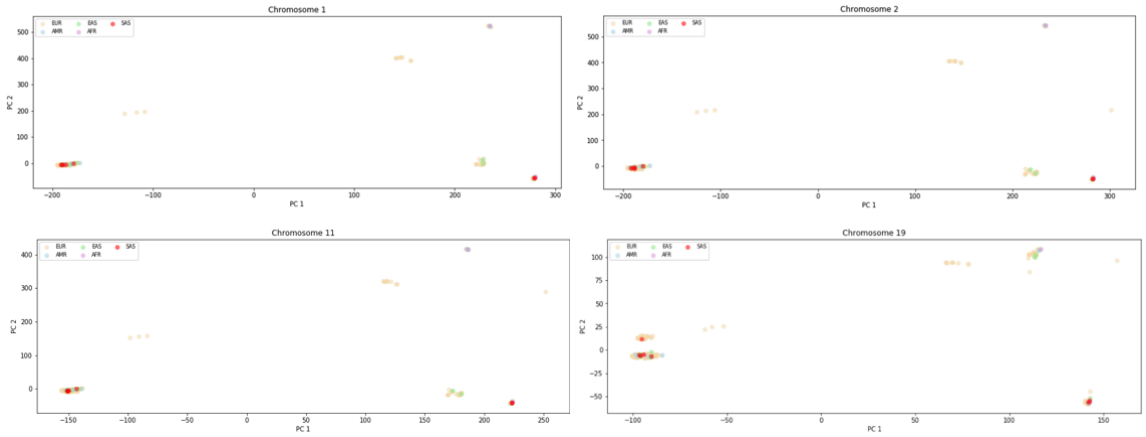

Figure 1. (a) Classification of individuals to individual population using Naïve Bayes with definitions taken from the International Genome Sample Resource (IGSR). **(b)** Clustering of individuals based on variants located on chromosomes 1, 2, 11, and 19. PCA was used to project the data to the first two principal components.

Variant-Drug Annotations

The PharmGKB annotations file contained 2,588 unique variant-drug pairs, capturing 1,586 unique RSIDs and 523 unique drug and drug combinations. PharmGKB contained variant annotations for 987 genes, of which 59 were genes containing Very Important Pharmacogene (VIP) summaries. VIP genes typically have been

reviewed by the FDA and Clinical Pharmacogenetic Implementation Consortium (CPIC) and have been studied in a large number of high-level publications. ABCB1, DPYD, and DRD were three VIP genes with the greatest number of variant annotations. Some variant-drug associations were extensively studied in multiple publications; for example, 68 studies reported genetic associations between clopidogrel and rs4244285 on the CYP2C19 gene, and 38 studies did for warfarin and rs9923231 on VKORC1.

Table 1 summarizes shared variant-drug associations for 26 of the 42 mental health drugs examined in this study, grouped by classifications curated from PubChem. Figure 2 shows location densities for chromosomes 1, 2, 11, and 19 to explore potential association between drug and location of ADE-related variants. Coordinate positions originated from build 37 of the reference genome. Drugs with the greatest number of variants on each chromosome were plotted; for example, antipsychotics, paroxetine, hydrochlorothiazide, olanzapine, and gemcitabine were selected for chromosome 11.

Table 1. Shared variants for atypical antipsychotics (except haloperidol, which is a typical antipsychotic), anticonvulsants (AC), selective serotonin reuptake inhibitors (SSRI), and tricyclic antidepressants (TCA). The presence of a variant-drug annotation is denoted by a dot

		rs489693	rs3813929	rs324420	rs17782313	rs518147	rs1414334	rs3780412	rs3780413
Antipsychotics	haloperidol*	•	•	•		•			
	aripiprazole	•		•					
	clozapine	•	•	•	•	•	•	•	•
	iloperidone		•						
	olanzapine	•	•	•	•	•	•	•	•
	paliperidone	•			•				
	quetiapine	•	•	•	•				
	risperidone	•	•	•	•	•	•	•	•
	ziprasidone	•	•						

		rs2606345	rs2844665	rs3094188	rs3130501	rs3130931	rs3815087	rs3815087
AC	carbamazepine	•	•	•	•	•	•	•
	phenytoin	•	•	•	•	•	•	•
	valproic acid	•						

		rs3892097	rs2032582	rs130058	rs1360780	rs2032583	rs2235040
SSRI	citalopram					•	•
	escitalopram						
	fluvoxamine					•	•
	nefazodone		•	•	•		
	ondansetron		•				
	paroxetine		•	•	•	•	•
	sertraline					•	•
	venlafaxine		•	•	•	•	•
TCA	amitriptyline	•					
	clomipramine	•	•	•	•		
	doxepin	•					
	imipramine	•					
	maprotiline	•					
	nortriptyline	•					

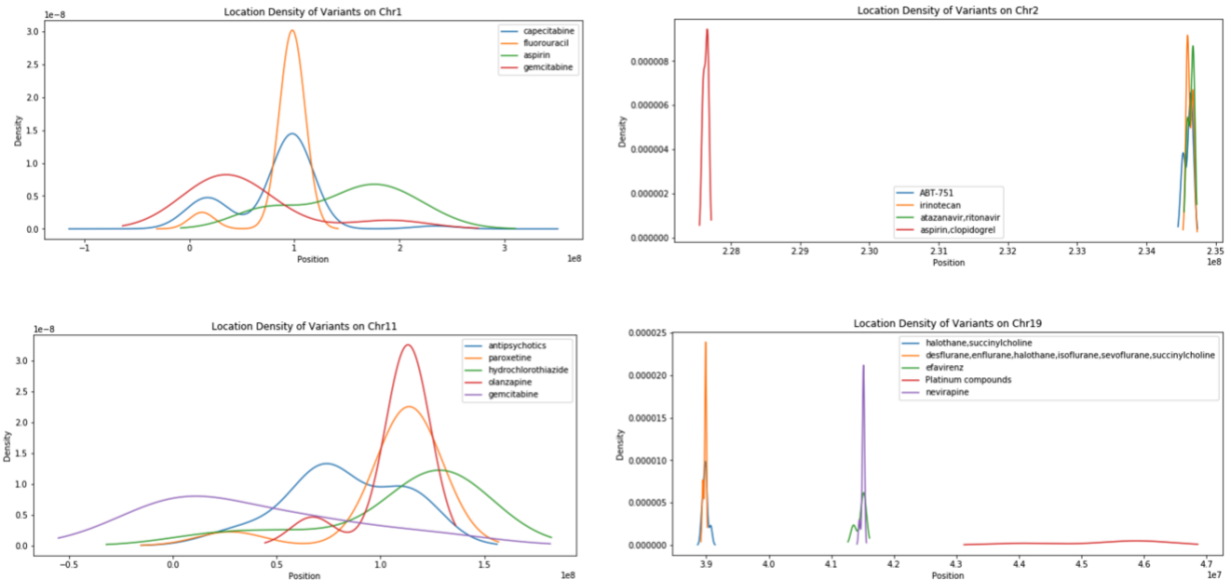

Figure 2. Location densities for ADE-related variants on chromosomes 1, 2, 11, and 19. The four or five drugs with the greatest number of variants on each chromosome were plotted.

Associated Drugs

Drugs that treat mental illnesses were chosen for this analysis. The network of drug associations is shown in Figure 3a, where the nodes represent drugs and are colored by modularity clusters. The labeled nodes are those identified from Drugs.com, and the non-labeled are those that were only present in PharmGKB annotations. Therefore, non-labeled drugs connected by edges or in the same cluster are those with a possible association determined through shared RSIDs. For example, the labeled nodes within the green community are exclusively selective serotonin reuptake inhibitor (SSRI) drugs (ecitalopram, citalopram, fluvoxamine, and sertraline) and phenothiazine antipsychotics (chlorpromazine, fluphenazine, thioridazine, and trifluoperazine). Non-labeled nodes within the cluster include the drug class antipsychotics and milnacipran, which is a serotonin-norepinephrine reuptake inhibitor. Other nodes in the cluster were caffeine, fenofibrate, terbinafine, and ticlopidine. The latter three drugs had only singular connections to sertraline in the cluster.

Paroxetine was chosen to perform a closer analysis on associated drugs for a specific use case. The resulting network around paroxetine is shown in Figure 3b, and 34 drugs were found to be related, in addition to two

drug classes, antipsychotics, and taxanes. One reason for association may be drug interactions, and DrugBank provides an extensive list of interactions obtained from drug labels and scientific literature. For paroxetine, DrugBank lists 779 interacting drugs, although a majority of them cite outcomes that were excluded from this study such as a decrease in serum concentration of a drug. Thirty out of the thirty-four 23andMe predicted drug interactions were listed as interacting drugs in DrugBank, and five of these had outcomes that matched the types of toxicities reported in the PharmGKB annotations. The drugs associated to paroxetine were also compared to those found in reported cases from AEOLUS. Twenty-four drugs were obtained, and every drug was listed among the interacting drugs on DrugBank, while thirteen of them had toxicity-related outcomes. There was no overlap between associated drugs found through AEOLUS and 23andMe genetic reports.

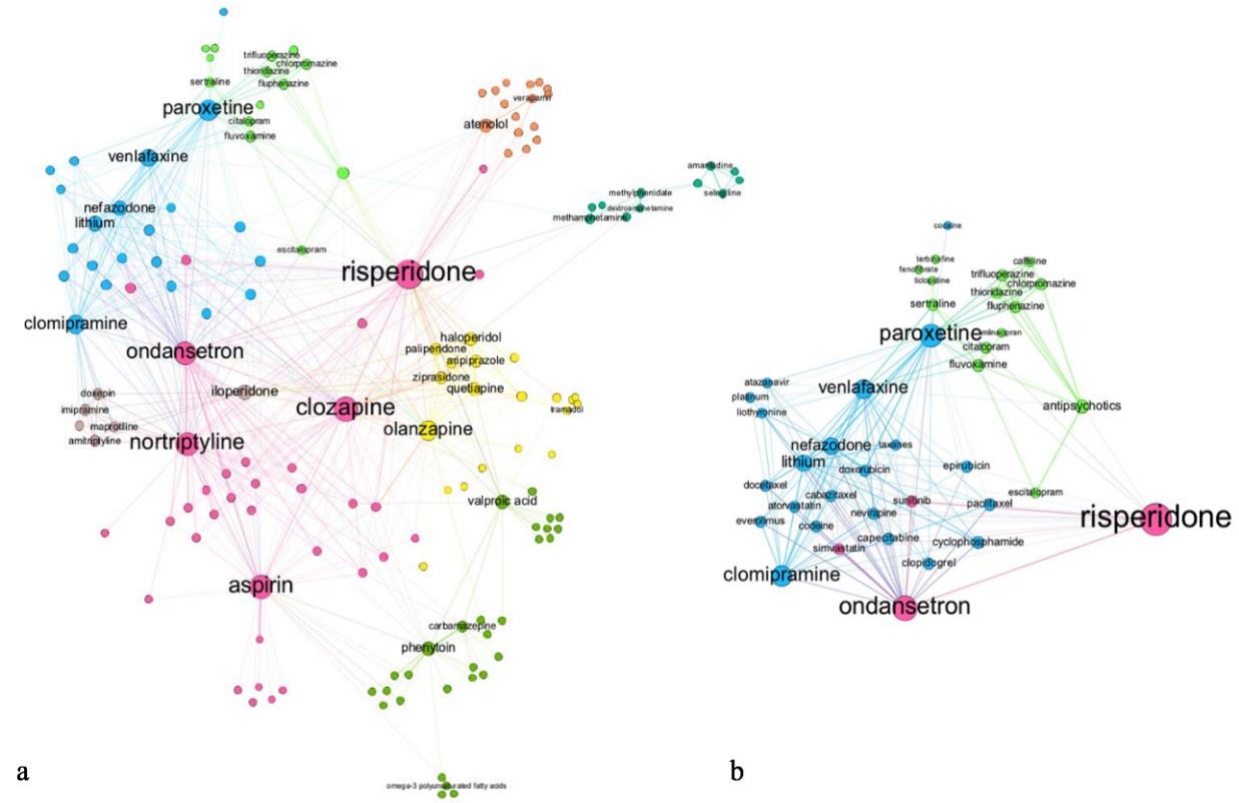

Figure 3. Network of mental health drugs based on pharmacogenetics. (**a)** An overall network that was distributed using a force-directed algorithm, and the nodes were colored by modularity to indicate clusters (**b)** Closer figure of the network of drugs associated with paroxetine.

Variant-ADE Associations

Adverse events for paroxetine were examined within the AEOLUS database, and the resulting associations are graphed in Figure 4a. ADEs with a PharmGKB variant annotation were labeled; nausea, fatigue, and suicidal ideation were found to occur the most frequently and had the greatest weighted degrees. The network was filtered by edge weight to limit the associations to those connected to one or more of the three outcomes (Figure 3b). While many of these ADEs were linked to both nausea and fatigue, two communities are distinguishable. Notably, dizziness, vomiting, and headache were clustered with nausea; and tremors, vertigo, and confusional state were grouped with fatigue. From these results, there was no clear relationship between the variants. Nausea and vomiting each had one variant RSID; patients with rs762551 had an increased risk of fatigue, and the variant is located on chromosome 15, affecting CYP1A2. rs1176744 was associated with discontinuation syndrome and nausea, and it is located on chromosome 11, affecting HTR3B[5].

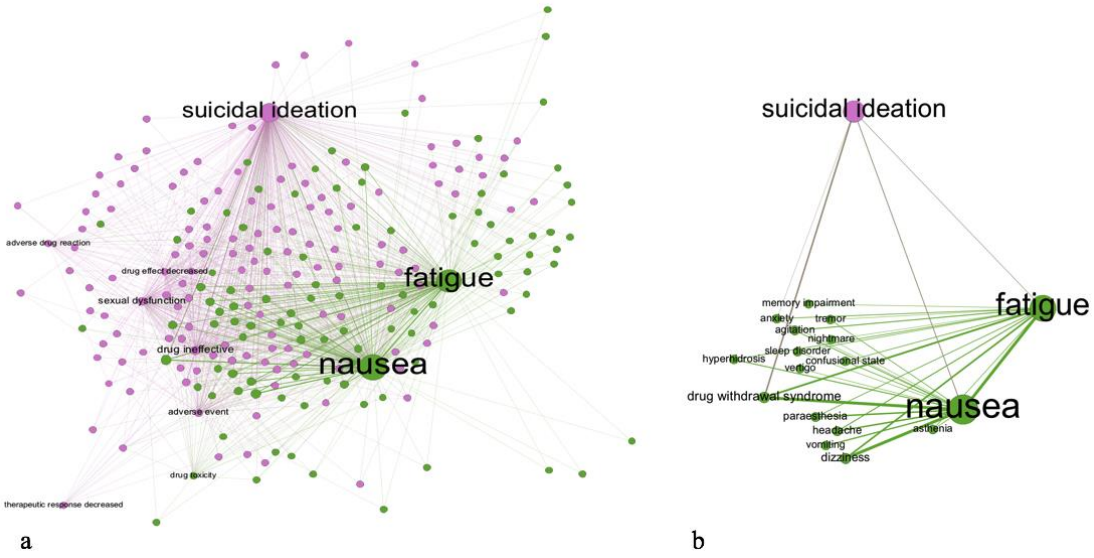

Figure 4. Network of adverse events associated with paroxetine: (**a)** Network of ADEs reported in cases where paroxetine was the only drug administered. (**b)** Network of ADEs related to nausea, fatigue, and suicidal ideation.

Discussion

Exploration of PGP Patient Reports

In this study, Naïve Bayes was the introductory model used to classify an individual's population, where features were the variant genotypes. The underlying assumption for this algorithm is that the features are independent. For this problem, the features are the genotypes for all variants in the genetic report, but independence may be violated, particularly for variants located in the same region of a chromosome. Without the true demographic information of the individuals, it is impossible to completely evaluate the results, but the aggregate and individual population predictions were compared to assess whether the results were reasonable. Namely, a reasonable result here was one in which the predicted individual population belonged to the predicted aggregate population. Sixteen individuals (2.06%) had inconsistent results. Several individuals with a EUR aggregate population were designated as Puerto Rican (PUR) or Colombian (CLM), both of which belonging to the AMR aggregate. There was one instance where an AMR individual was predicted to be Bengali (BEB), belonging to the SAS subpopulation. The analyses presented for associated drugs may include ethnic biases related to the population being predominantly European. Since the significance of single mutations can differ between populations, future work should study trends in each population to capture possible variations between ethnic groups. However, additional genetic data would be required since the currently available data is lacking, particularly for non-European populations.

Due to the ambiguity of self-reported ethnicities, PCA was used to obtain groups within the population of genetic reports, which were then compared to the assigned populations. The chromosomes were chosen to match those used in the plots of location density. Although questionable as a form of validation, the comparison was a useful exercise to search for agreement between the two methods. The clusters did not seem to be separated by aggregate populations particularly since a majority of individuals were classified as EUR. Nevertheless, the scatter plots did suggest that there were groups separable by genetic profile. The plot for chromosome 19 showed the most deviation; when PCA was performed on combined data from the four chromosomes, the resulting plot was similar to those for chromosomes 1, 2, and 11, which masked the

differences seen for chromosome 19. This suggests that clusters formed for each chromosome depend on the genes and variants located on each and that the underlying significance may be chromosome-specific.

Of particular interest is whether ADE-related variants co-occur within the population, and forming clusters within the population of reports may be an informative start. While co-occurring variants were used in creating the pharmacogenomic drug network, this was based only on variant identities while disregarding variant genotypes. The co-occurrence of variants with relevant genotypes associated with a condition or ADE has yet to be successfully studied. One future approach would be to search within the previously formed clusters to discover groups of co-occurring variants; it is quite possible that groups of variants exist since as few as two dimensions captured over sixty percent of the cumulative variance for over one hundred thousand mutations. Another approach would be to examine co-occurrence for variants associated with a particular drug or adverse outcome. For the latter, grouping by ADE is necessary since one genotype for a particular SNP may be associated with a decreased risk of one ADE, yet another genotype may be associated with an increase risk of a different ADE. To do this, additional natural language processing techniques should be tested to standardize adverse event names.

Location of Variants

The location density plot in Figure 1 supports two ideas. The first is that similar drugs may be associated with variants that are located in similar regions of a chromosome, which is a reasonable conclusion particularly if the variants are located on a gene that affects a metabolic pathway. Each plot in the figure supports this, but chromosomes 2 and 19 demonstrate sharper localizations of variants for the selected drugs. For example, efavirenz and nevirapine are non-nucleoside reverse transcriptase inhibitors that treat HIV-1 infections with aligned peaks on chromosome 19. The two drugs had eleven and nine ADE-related variants, respectively, of which four were common to both. Table 1 similarly demonstrated that an ADE-related variant may affect multiple drugs, usually within the same class of drugs. Platinum compounds displayed no significant peak on chromosome 19, which agrees with a secondary observation, namely that locations of ADE-related variants may be distributed across distant loci. As another example on chromosome 1, which is the largest in the human genome, capecitabine and fluorouracil have bimodal distributions with peaks located approximately 10^8 units

apart. A similar phenomenon was observed for antipsychotics and olanzapine on chromosome 11, although with a shorter distance separating the peaks.

Associated Drugs

Population 23andMe data combined with PharmGKB annotations is a potentially useful resource in pharmacogenomics by indicating related RSIDs and, consequently, drugs. The network of drugs built in this work (Figure 2) was based only on genetic and population data. Drug associations may be examined through shared pathways or structural similarities, but a genetic association may encompass those relationships or involve alternative ones. One example is drugs that have a combined therapeutic effect but are chemically different. For example, cyclophosphamide is an alkylating antineoplastic agent with similar affected genes to cisplatin, a platinum-based agent without an alkylating group, and studies have shown encouraging results for combination treatments using the two drugs[20]. In a clinical setting, knowledge of drugs associated with genetic profiles may be important when prescribing alternate lines of treatment for an indication. For example, carboplatin and cisplatin are platinum-based chemotherapy drugs that share structured indications on DrugBank; while analogues, the drugs have no shared genes, and therefore one treatment may be prescribed instead of the other to mitigate ADEs.

The modularity of the network was dependent on edge weights between drugs. For two drugs to be strongly correlated using this methodology, the associated variants for both must commonly co-occur in the patient population. While variants on different chromosomes have no association by definition of being unlinked, they may have a nonrandom tendency to be co-inherited, referred to as linkage disequilibrium. One reason for this is a shared function between the variants, which are then associated during selection. Although the typical use for this quantity is for locations on the same chromosome, the original calculation allowed for the consideration of different chromosome[21]. When looking at co-occurring variants affecting weight gain, the analysis demonstrated that variants on different chromosomes may have a nonrandom tendency to co-occur. When the steps were repeated for drugs causing neutropenia, it was observed that some variants on different chromosomes co-occurred while others did not. For example, RSIDs on chromosomes 7 and 12 affected sensitivity to clozapine but did not co-occur in the two largest sets, whereas RSIDs on chromosomes 7 and 13

did for valganciclovir. As previously mentioned, these analyses were performed only for variant identities, and additional examination of individual genotypes is needed.

Successful grouping of drugs, which is assessed in detail below, would indicate that annotated RSIDs are captured in 23andMe data, supporting the potential clinical utility of DTC tests. Table 1 shows that drugs in the same classification tend to share variant annotations and are expected to be grouped in the same modularity class. Measures of precision and recall were calculated by comparing the clustering for drugs with drug classes. Each cluster represented a drug classification, and true positive was defined as a drug whose node that was correctly colored. For example, the green cluster in Figure 2 corresponded to phenothiazine antipsychotics and the adjacent blue cluster was SSRIs. The average precision of this methodology was 0.701, and the recall was 0.752. Classes that had fewer than three drugs, among other exceptions, were not included in the averaged values. The green modularity class mentioned in the results lowered these values due to the coverage of two drug classifications that otherwise were clustered with high precision.

Quantifying precision and recall for this application is difficult because the nature of drug association is uncertain. Here, they were calculated according to drug classification, which is an imperfect measure since drugs with different classifications may share other similarities like drug interactions or chemical structure. For example, iloperidone, an atypical antipsychotic, was clustered with several tricyclic antidepressants due to drug interactions, which were included in the list of "moderate" interactions in the drug's boxed warning[19]. Conversely, some drugs of different classifications were grouped in the same modularity class like atenolol and verapamil. While the former is a beta blocker and the latter is a calcium channel blocker, they share one annotated RSID and are both prescribed for angina and high blood pressure.

The levels of precision and recall obtained in this study support the proposed methodology for discovering potential drug associations. Returning to the green cluster from above, fenofibrate, terbinafine, and ticlopidine were grouped drugs that lack clear similarities in structure or indication. It would be interesting to investigate if these unrelated drugs would be removed from the modularity cluster with the introduction of more data or if there is an underlying genetic association that might make an individual prone to ADEs from both drugs. Caffeine was also an associated drug, and it has known effects on psychiatric symptoms and potential

interactions with antipsychotic medications[23,24]. Although this association has been documented in preliminary scientific studies, this is an example of discoveries that may have interesting research or clinical implications.

As mentioned earlier, drug interactions are one possible association depicted by the clustering. One attempt to verify this association was through DrugBank. Paroxetine was chosen for this purpose because its associations were concentrated in two separate clusters. There was no overlap between possibly related drugs through variants and through AEOLUS case reports. The number of drug interactions on DrugBank is rather large, so the list should be curated for significance or another means of verification may be necessary. If drug interactions found through genetic variants can be verified, associations that are unexpected may be have clinical and research potential.

Variant-ADE associations in this study focused only on the outcomes for paroxetine, but in the future the network of ADEs should be mapped across a larger set of drugs to obtain a more comprehensive network. Many of the outcomes shown in Figure 2a had one or two connecting edges; the addition of more nodes and edges would facilitate the formation of more distinct clusters. It is interesting that less distinct groups were already formed from this limited demonstration; for example, headache, vomiting, and dizziness commonly occur alongside feelings of nausea. Expanding the network of associated outcomes to include a variety of drugs would further investigate the notion that ADEs have characteristic pathways that genetic variants may affect. There was no obvious genetic correlation between fatigue, nausea, and suicidal ideation when treated with paroxetine. The variants were located on different chromosomes and affected unrelated genes. An interesting observation is that the variant associated with paroxetine and fatigue affects CYP1A2, and the CYP1A2 enzyme metabolizes caffeine. A recent study demonstrated an increased serum concentration of paroxetine with the coadministration of caffeine[23].

Potential Clinical Utility of DTC-Derived Data

One of the principal aims of this study was to assess the potential clinical utility for DTC genetic testing like 23andMe. These genetic tests have the potential to increase the accessibility of personal genetic data, which can be used as an additional source of information that is more comprehensive and consistent than what is normally provided in patient history forms. 88% of annotated variants and 95% of drugs were contained within

all genetic profiles, which suggests that a majority of variant-drug associations are included in the 23andMe screening. As mentioned previously, the variant-drug associations from PharmGKB were not fully captured in the annotations file due to the limited rule-based approach used in this feasibility study. The categories of some annotations were inconsistent with the findings contained within the summaries. For instance, some associations of toxicity were related to dosage reductions, while an example association within the dosage category involved an increased risk of neutropenia. In particular, an alternate method is necessary to capture these inconsistencies that involve outcomes outside of the curated vocabulary.

While consumer genetic tests are not intended for diagnostic purposes, the information obtained from a population perspective may have clinical utility. Given the results of this study, 23andMe profiles could be used to alert patients and medical practitioners of potential ADEs, prompting a more comprehensive genetic test in the clinic. At the minimum, a screening process could include rapid comparison of a DTC profile like 23andMe to a knowledgebase of RSIDs associated with drug toxicity (e.g., sourced from PharmGKB). The genetic profile for an individual that indicates an increased chance for an adverse event to one drug might also suggest that ADEs are probable for another, whether it is due to drug interactions or a separate affected pathway. The drug and ADE relationships examined in this study demonstrate that valid associations may be obtained by examining genetic information at a population level. That being said, more work is required to validate the findings and to expand the analyses beyond a single drug and drug group.

Limitations

One of the major shortcomings Lu et al. discussed was the significant genetic variations between different populations. Here, individuals were assigned populations using allele frequencies calculated based on self-reported ethnicities, which raises separate concerns of ethnicity having confounding cultural factors or race being a social construct. More importantly, the population data used in this study lacked labelled demographic information, and as a result many of the analyses performed were from an unsupervised learning perspective. Lastly, the sample size of genetic reports was on the smaller size for a population study, particularly if further research should be performed on the smaller clusters formed. The addition of data that is labelled and in greater quantity could facilitate the study of the concepts presented here in a more rigorous manner.

Conclusion

This study demonstrated that 23andMe genetic reports test for a majority of variants with clinical annotations for drug hypersensitivity, which is important to note if this technology will be used as a clinical tool to flag the potential need for full diagnostics. The FDA has established special controls that need to be met to demonstrate safety and effectiveness for genetic tests that assess risk for conditions. Among these controls are expectations of clinical performance and labeling, so beginning with VIP-designated genes that affect drug pathways might be a promising approach to obtain approval for assessing risk for drug sensitivity.

Preliminary results also suggest that associations between drugs can be obtained by examining genetic profiles at a population level. The pharmacogenomic network presented here was partly dependent on ADE-related variants co-occurring within reports. From a research perspective, these associations could guide future research directions or serve as an additional pharmacogenetic resource. In future work, co-occurring variants will be pursued further.

References

1. Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. J Pharmacol Pharmacother. 2013; 4(Suppl 1): S73-7.
2. Hug BL, Keohane C, Seger DL, Yoon C, Bates DW. The costs of adverse drug events in community hospitals. Jt Comm J Qual Patient Saf. 2012; 38(3): 120-6.
3. Questions and Answers on FDA's Adverse Event Reporting System (FAERS) [Internet]. US Food and Drug Administration. 2017 [cited 29 August 2017]. Available from: https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm
4. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016; 3: 160026.
5. 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 (2012) 92(4): 414-417.
6. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1; 29(1): 308-11.
7. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014 Jan 1; 42(1): D1091-7.
8. Chan SL, Ang X, Sani LL, Ng HY, Winther MD, Liu JJ et al. Prevalence and characteristics of adverse drug reactions at admission to hospital: a prospective observational study. Br J Clin Pharmacol. 2016; 82: 1636-46.
9. Elliott LS, Henderson JC, Neradilek MB, Moyer NA, Ashcraft KC, Thirumaran RK. Clinical impact of pharmacogenetic profiling with a clinical decision support tool in polypharmacy home health patients: A prospective pilot randomized controlled trial. Plos One. 2017; 12(2).
10. Olson MC, Maciel A, Gariepy JF, Cullors A, Saldivar JS, Taylor D, et al. Clinical impact of pharmacogenetic-guided treatment for patients exhibiting neuropsychiatric Disorders: a randomized controlled trial. Prim Care Companion CNS Disord. 2017; 19(2): 16m02036.
11. Giacomini KM, Yee SW, Mushiroda T, Weinshilboum RM, Ratain MJ, Kubo M. Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. Nat Rev Drug Discov. 2017;16(1):1.
12. FDA authorizes, with special controls, direct-to-consumer test that reports three mutations in the BRCA breast cancer genes [Internet]. U.S. Food & Drug Administration. 2018 [cited 8 April 2018]. Available from: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm599560.htm
13. Lu M, Lewis CM, Traylor M. Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe. BMC Med Genomics. 2017; 10(1): 47.

14. Ball M, Bobe J, Chou M, Clegg T, Estep P, Lunshof J et al. Harvard Personal Genome Project: lessons from participatory public research. Genome Medicine. 2014; 6(2): 10.
15. Mental Health Disorders [Internet]. Drugs.com. 2017 [cited 26 September 2017]. Available from: https://www.drugs.com/mental-health.html
16. Bastian M, Heymann S, Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
17. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks.
18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 Jan 4; 44(D1): D1202-13.
19. Iloperidone - Drug Summary [Internet]. Prescriber's Digital Reference. 2017 [cited 6 January 2018]. Available from: http://www.pdr.net/drug-summary/Fanapt-iloperidone-429#topPage
20. Dasari S, and Tchounwou PB. Cisplatin in cancer therapy: molecular mechanisms of action. Eur J Pharmacol. 2014 Oct 5; 0: 364-378.
21. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun; 9(6): 477–485.
22. Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).
23. Szopa A, Poleszak E, Wyska E, Serefko A, Wosko S, Wlaz A, et al. Caffeine enhances the antidepressant-like activity of common antidepressant drugs in the forced swim test in mice. Naunyn Schmiedebergs Arch Pharmacol. 2016; 389: 211–221.
24. Broderick PJ, Benjamin AB, Dennis LW. Caffeine and psychiatric medication interactions: a review. J Okla State Med Assoc. 2005 Aug; 98(8): 380-4.