UNIVERSITY OF CALIFORNIA

Los Angeles

Measuring Genetic Contribution in a Drug-Disease Network

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Bioengineering

by

Rayo Suseno

2023

ABSTRACT OF THE THESIS

Measuring Genetic Contribution in a Drug-Disease Network

by

Rayo Suseno

Master of Science in Bioengineering

University of California, Los Angeles, 2023

Professor Jennifer L. Wilson, Chair

A disease can be caused by multiple mechanisms including genetic mutations, proteomic abnormalities, or abnormal mRNA transcription. While the Human Genome Project aims to understand the genetic basis of complex diseases, not all relevant variants are druggable targets. Therefore, a network approach can be useful for understanding if druggable proteins are close to gene variants. Previous development of PathFX, a network algorithm that identifies relationships between drugs and diseases, has given us more comprehension on genes / proteins that might be responsible for drug effects on disease and side effect phenotypes. While PathFX was constructed using a combination of gene expression, gene mutation, proteomics information, and other resources, the goal of the project is to understand how much a drug-phenotype relationship is driven by genetics within the PathFX network. We quantify the genetic component by finding the overlap between the PathFX network and the GWAS catalog data. We discovered that drugs connected to three disease groups: psychological disorders, autoimmune diseases, and inflammatory bowel diseases have greater enrichment of genetic information than other drug-

disease pairs. Our quantitative assessment may uncover considerations for new drug developments, given that a disease is more genetically or less genetically enriched.

Keywords: GWAS, NAA, Overlap, Drug Development, Genetic.

The thesis of Rayo Suseno is approved.

Aaron S. Meyer

Matteo Pellegrini

Jennifer L. Wilson, Committee Chair

University of California, Los Angeles

2023

Table of Content

1 Introduction

One categorization of a disease is whether it is monogenic or polygenic. A monogenic disease, typically known as Mendelian disease, is caused by a single mutation in the subject's genome. An example is cystic fibrosis (CF), where a single mutation in the CFTR gene drives the progression of the disease. On the other hand, a polygenic disease is usually not only driven by multiple genes, but also affected by other factors from the environment[6]. This disease category is far more prevalent in our society and has much more social and economic impact[10]. The development of the Human Genome Project has propelled scientists to better understand the genetic component of complex diseases through initiatives like Genome-Wide Association Studies (GWAS)[9,11].

GWAS is designed to identify genetic variants that are associated with phenotypes of interest. The initiative has been growing exponentially lately in parallel with the growth in sequencing technology. By 2020, there had been 4,300 GWAS papers published with about 55,000 loci of interest[9]. However, numerous problems persist in GWAS studies, hindering its implementation in the drug discovery pipeline. First, GWAS often reports a number of genomic region candidates instead of useful causal variants or specific genes[3]. Second, a typical GWAS will fail to detect variants that are weakly associated with a given disease and may miss variants with true associations to the disease. Lastly, GWAS hits are not always druggable - they cannot be altered using standard drug tools such as small molecules or antibody therapies[2].

The development of network-assisted analysis (NAA) tools made it possible to address this problem, as well as other healthcare-related computational problems such as finding new drug targets[7,16]. Weaker variants are often closely associated within the protein-protein interaction network and studying these variants in the context of a network has helped expand disease pathways. They have also found connections between druggable targets and disease-associated

variants suggesting that novel treatments may leverage indirect altering of disease-variants instead of directly manipulating them. These results suggest that patient genetic variation converges on network submodules but is not localized to single, gene-drivers of disease.

One of the more recent developments of an NAA tool is PathFX, an algorithm used to construct drug pathways using protein-protein interaction (PPI) information obtained from various databases[16]. In contrast to target-driven understanding of drug effects, PathFX aims to understand drug-phenotype associations using druggable targets and proteins downstream of targets. Currently, PathFX leverages gene-disease phenotype associations from multiple databases to define drug-disease relationships. PathFX used a culmination of DisGeNet, Phenotype Genotype Integrator (PheGenI), ClinVar, Online Mendelian Inheritance in Man (OMIM), PheWAS, PharmGKB, and GWAS[16]. From these aggregated sources, PathFX would selectively pick associations depending on the threshold that was set in the algorithm. PathFX associations can arise through the confluence of data sources since genes that are targeted by the drug are then further connected to interacting proteins using a protein-protein interaction network such as STRING and iRefWeb[16].

These sources aggregate information from multiple experiments and databases and may include mutated, over-expressed, or hyperphosphorylated proteins and genes. These aggregated pathways are useful for computational pathways prediction; however, the true disease pathway is often unknown. Additionally, it's unclear which data - gene expression, gene mutation, proteomics - is the dominant contributor to a drug-phenotype relationship in the network.

In this project, we investigate the extent to which PathFX-defined disease modules are enriched for gene variant data from the GWAS catalog and use this enrichment score to rank PathFX disease pathways that are most aligned with genetic definitions of disease. Understanding how much the PathFX drug-phenotype network is driven by its genetic component can potentially assist drug development processes and guide the use of genetic information in the drug discovery pipeline.

2 Methodology

2.1 GWAS Data Preparation

We downloaded and pre-processed the NHGRI-EBI Catalog of human GWAS before being used in our analysis. Since every row is a published study, we first aggregated rows that share the same disease name to reduce redundancy. We then extracted the three data columns: disease name, mapped genes, and reported genes. Since this catalog is a culmination of different studies, there is no consistent way of delimiting the genes. We split the genes into lists using three delimiters: comma, semi-colon, and dash. After removing some null and invalid gene entries, we finally curated a dictionary by mapping each disease as the key to its respective gene list as the value.

2.2 Quantification of Disease Genetic Component

The goal of this step was to quantify the extent to which a PathFX disease phenotype is genetically driven. Every disease in the GWAS and PathFX database have been mapped to their respective genes. Calculating the genetic component was done by comparing the two lists that map to the same disease. To better illustrate this process, let us assume that disease X is mapped to four genes in PathFX (gene A, B, C, and D). In GWAS, disease X maps to two genes, A and C. Since gene A and gene C appear in both PathFX and GWAS gene list, it means that there are two overlapping genes. Therefore, we can say that disease X is 50% genetically enriched.

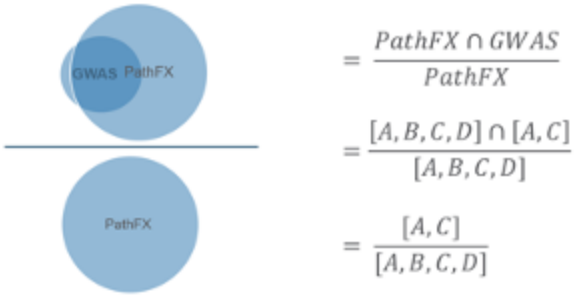

Figure 1. A schematic representation of the genetic measurement calculation to identify how much a disease in PathFX is driven by genetics

2.3 Phenotype Matching

Disease phenotype databases, including GWAS and others contained in PathFX, use a variety of names to describe phenotypes, posing a challenge for cross-database phenotype comparison. For instance, “cancer” can also be represented using the word “carcinoma” or “metastatic tumor”. We initially tackled this problem by performing a simple word count comparison, as written on Section 2.3.1 String Lexical Similarity. However, after a preliminary check on the result, we turn to UMLS MetaMap for a better phenotype matching process, described in Section 2.3.2.

2.3.1 String Lexical Similarity

To compare GWAS and PathFX phenotypes, we first used a string lexical similarity analysis, which simply counts the number of overlapping words between the two phrases. Since the length of strings vary from one disease to the other, we used a ratio-based comparison. Below are actual phenotypes found in both the PathFX and GWAS dataset.

GWAS: “*Squamous cell carcinoma of lung*

PathFX: “*Squamous cell lung carcinoma*”

The word “squamous”, “cell”, “lung”, and “carcinoma” match, totaling to 4 matches. Since GWAS has the longer string length, we set 5 as our length of reference. Four out of five words match; therefore, it is 80% similar. A threshold of 50% is chosen since most phenotypes are fairly short. That being said, the two strings above are deemed similar enough.

While this technique was helpful to determine whether we would obtain any overlapping signal, it does not consider the actual meaning of the phrase. For instance, “*Alzheimer’s disease*” and “*hypertensive disease*” would be considered a match, since they both share the word *disease* in them. The term *disease* is such a generic term that it would match to two phenotypes who are not even remotely related.

2.3.2 Converting GWAS phenotypes to CUI using MetaMap

As an attempt to better match the phenotypes, we make use of the information that PathFX phenotypes were already mapped to Concept Unique Identifiers (CUI), by converting GWAS phenotypes to CUI terms. CUI terms are maintained by the Unified Medical Language System (UMLS), an inventory for biomedical and clinical terms and/or concepts[12]. Each concept is enumerated via Concept Unique Identifier (CUI). To retrieve CUI of a given concept, we utilized MetaMap, a program that maps biomedical text to the most-likely UMLS Metathesaurus CUI term[1].

One advantage of converting phenotypes using MetaMap and matching with UMLS CUI is that many biological ontologies map directly to UMLS, making it one of the largest repositories available. This would also assist us in matching phenotypes that lexical similarity would otherwise fail to categorize. For instance, “*Alzheimer’s disease*” and “*hypertensive disease”* would no longer be considered a match since they map to

different CUIs, C0002395 and C0020538 respectively. This method also lets us introduce contexts in our matching process. Phrases like "*kidney diseases"* from PathFX and "*nephropathy"* from GWAS would never be a match if we were to use lexical similarity. According to UMLS, however, both terms map to C0022658 defined as "a nonspecific term referring to disease or damage of the kidneys" thus granting us a match.

We used MetaMap Batch, a version of MetaMap that accommodates query jobs, to convert GWAS phenotypes to CUI terms. As an input, we prepared a newline-delimited text file with a disease on each line. To treat each line as an input rather than free text, we turned on the `--sldi` option that stands for single-line-delimited-input. Additionally, since most of our entries are not complete sentences, we also turned on the `--term_processing` option to process each input as a short text fragment rather than a complete sentence. Lastly and most importantly, we utilized the `--show_cuis` option so that each input can be mapped to its corresponding CUI.

To simplify our database, we retained the first CUI match of the MetaMap outputs. Parsing this output gave us 5,910 GWAS phenotypes mapped to their most-likely CUI. We then converted the GWAS phenotype-CUI dictionary to a GWAS CUI-genes structure for an easier matching process. Multiple GWAS phenotypes can match to the same CUI term, forcing us to consider how to merge genes from distinct but similar GWAS phenotypes. To prevent data loss, we took the union of all genes and assigned them to the common CUI term.

As an example, C0456962 maps to 3 phenotypes and their own unique set of genes as shown in Table 1 below. We consolidated all the genes and assigned them to C0456962 to obtain the following CUI-genes information: *'C0456962': ['LINC01058', 'CTPS1', 'SCMH1', 'SLFNL1', 'HMGB1', 'SLFNL1-AS1', 'LINC00426', 'KATNAL1', 'APOOP5', 'NR', 'DUXAP11', 'CSNK1G3', 'LINC02147', 'NR', 'LINC01774'].* We obtained

a total of 2,604 CUI to genotype relationships, which can be found under Supplemental File 3.

GWAS Phenotype	GWAS Genes
Rapid response to perioperative phenylephrine (change in mean arterial pressure)	['LINC01058', 'CTPS1', 'SCMH1', 'SLFNL1', 'HMGB1', 'SLFNL1-AS1', 'LINC00426', 'KATNAL1']
Rapid response to perioperative phenylephrine (change in diastolic arterial pressure)	['APOOP5', 'NR', 'DUXAP11']
Rapid response to perioperative phenylephrine (change in systolic arterial pressure)	['CSNK1G3', 'LINC02147', 'NR', 'LINC01774']

Table 1. A list of phenotypes and genes that maps to CUI C0456962. Multiple GWAS phenotypes may map to one CUI, causing us to merge genes from all 3 phenotypes.

2.4 PathFX Approved Indication Genetic Enrichment

Because we were running multiple analyses of drug effects, we generated a pipeline to analyze all drugs available in DrugBank Version 5.1.6 (released with Wilson et al 2021). This release contained 7012 DrugBank identifiers, which we provided to the PathFX algorithm. Briefly, PathFX generates a protein-protein interaction (PPI) network around drug targets based on the amount and quality of evidence supporting the PPIs. Next, PathFX uses a modified Fisher's exact test to discover biological phenotypes associated with the drug's network (full description in Wilson et al 2018).

For our first analysis, we sought to understand the extent to which GWAS genes are associated with approved drug networks and used a published data table from Wilson et al 2018. As a means to characterize PathFX's performance, Wilson et al. benchmarked its output to a number of approved drugs and quantified the algorithm's sensitivity for predicting a drug's intended-to-treat disease. Wilson et al. curated a list of drugs and their approved disease phenotypes. For example, PathFX suggests that acebutolol can be used to treat "*hypertensive disease*", in agreement with the fact that acebutolol is indeed marketed and approved to treat

"*essential hypertension*". With each of these predictions, PathFX also provided a list of network genes that contributed to the prediction. We then take these genes and quantify its genetic component by intersecting them with the GWAS genes that match the appropriate phenotype.

2.5 Statistical Analysis

The shared genes that we found between PathFX and GWAS dataset might still potentially be due to chance, especially since some genes are quite common. For instance, CYP2D6 is a metabolizer that is responsible for the pharmacokinetics of a drug, therefore, is shared among different drug targets. To ensure that our result did not occur by chance, we performed a statistical analysis to determine whether the number of genes that are shared between the PathFX network and GWAS are statistically significant.

Our approach is to introduce a permutation test by matching a PathFX disease with a random GWAS disease. For every PathFX phenotype in the Approved Indication network, we paired it up with a random GWAS disease, bypassing the phenotype matching process. We then performed the same overlap calculation to determine the number of shared genes between two randomly picked phenotypes. As an example, we calculated the number of shared genes between "*schizophrenia*" and "*Rapid functional decline in sporadic amyotrophic lateral sclerosis*", which expectedly yielded 0 shared genes between them. We performed this random phenotype simulation 100 times. Since there are 2,871 rows in the Approved Indication network, we ended up with 287,100 simulated values as the simulated distribution. We also saved the original distribution with 2,871 values in it.

Once we obtained both the simulated and real distributions of the number of shared genes, we performed a Mann-Whitney U test, also known as the Wilcoxon Rank Sum test: a statistical tool to assess whether two sampled groups are likely to come from the same population. While there exists many other statistical methods to answer that question, we are particularly interested

in Mann-Whitney because our data is quite heavily skewed and both groups share a similar distribution shape, which are two of the main assumptions in a Mann-Whitney test. We imported `mannwhitneyu` from `scipy.stats` and passed in both real and simulated distributions to obtain the U and p-value.

2.6 Pipeline Overview

To summarize our analysis, we started by parsing the Catalog of Human GWAS before using UMLS MetaMap to map GWAS phenotypes to UMLS CUI identifies, creating a CUI-to-genotype data structure. We also ran PathFX for all DrugBank drugs and filtered them using a list of approved drugs. For drugs where PathFX predicted the intend-to-treat indication, we retained the network genes associated with the disease; this yielded 2,871 rows where drugs could be associated with more than one disease-relevant phenotype. After matching the phenotypes and calculating the gene overlap between PathFX and GWAS, we obtained the genetic enrichment table, which includes information such as GWAS genes for the matched phenotype, shared genes, name of GWAS phenotype that matches to the approved phenotypes, CUI (Supplemental File 1).

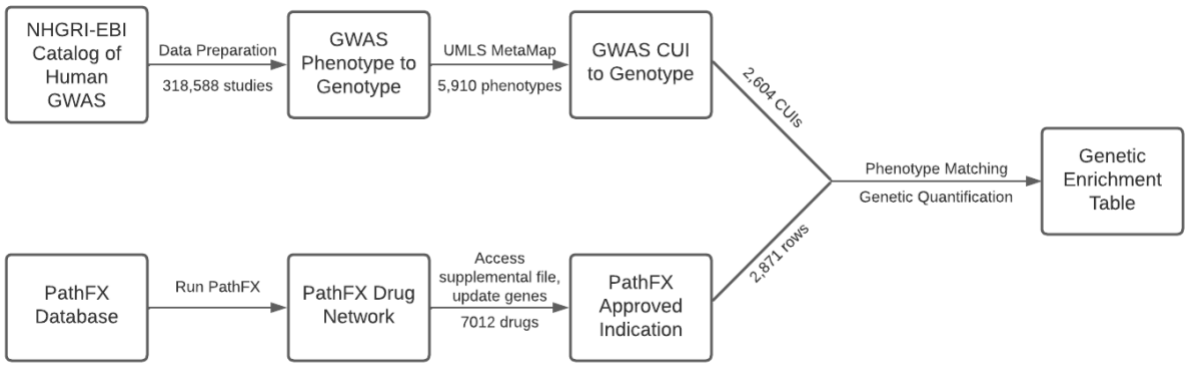

Figure 2. A high-level overview of our analysis to measure genetic contribution in PathFX network by using the NHGRI-EBI GWAS Catalog.

3 Results

3.1 GWAS catalog overview

The NHGRI-EBI Catalog of human GWAS is a data source of 318,588 non-unique published studies with information such as publication date, first author, initial sample size, etc[11]. After preparing the data for our purposes, we were left with 5,910 phenotypes mapped to a total of 282,009 genes consisting of 38,860 unique genes. The median number of phenotype-genotype associations was 9 genotypes for each phenotype. Interestingly, phenotypes with the most associations include phenotypic traits and are not necessarily associated with disease phenotypes. For instance, “height” had 4,380 distinct genes, and “body mass index” had 3,293 associations. This reflects that the GWAS catalog seeks to understand human traits in addition to disease.

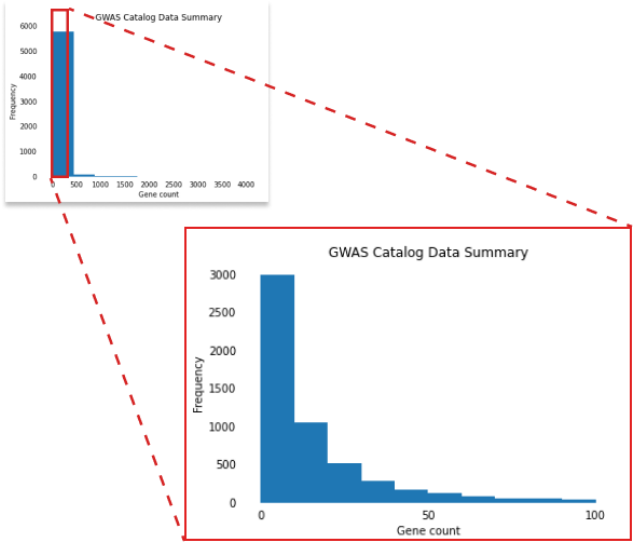

Figure 3. An overview of the NHGRI-EBI GWAS Catalog data. The x-axis shows how many genes are associated with a phenotype. A majority of GWAS phenotypes have fewer than 10 genes associated with it (scaled histogram).

While mapping the unique 5,910 phenotypes to CUI identifiers, we parsed the phenotype's most-likely CUI from MetaMap (a phenotype may map to multiple CUI terms), meaning that a given phenotype is not a guaranteed unique match to the CUI. The phenotypes that fall under the same CUI are sufficiently similar for our analysis. For instance, the three phenotypes: *"venous thromboembolism", "venous thromboembolism adjusted for sickle cell variant rs77121234-T",* and *"venous thromboembolism (SNP x SNP interaction)"* all map to C18161172. Another example, *"Kidney disease (end stage renal disease vs non-end stage renal disease) in type 1 diabetes", "Renal underexcretion gout",* and *"Renal overload gout"* all map to C0022646. We manually validated some of the other CUI-to-phenotype mappings and discovered our assumption was sufficient for this analysis. These mapping results can be found on Supplemental File 2.

3.2 PathFX phenotypes have relatively few GWAS Genes

Before understanding how genetic information contributes to network drug-phenotype predictions, we first quantified the overlap of GWAS and PathFX phenotypes. The PathFX database contains phenotype-genotype information from various databases like PharmGKB, OMIM, and DisGeneNet[16]. This data includes genetic information from GWAS, proteomic studies, mRNA information, and post-translational modification.

Since this was a preliminary analysis, we did not implement MetaMap to match the phenotype between GWAS and the PathFX database. A simple lexical similarity found that out of 23,080 phenotypes in the PathFX database, 2,860 were a match to GWAS. Of this subset, we found that 626 phenotypes have at least 1 gene in common, 328 of which have exactly only 1 overlap. We summarized the analysis in Figure X below. As shown in the figure, most of these diseases have only 1 or 2 GWAS genes. However, some diseases have a relatively high number of GWAS genes, such as schizophrenia with 287 shared genes, asthma with 163 shared genes,

coronary artery disease with 116 shared genes, psoriasis with 64 shared genes, and attention deficit hyperactivity disorder with 55 shared genes. Overall, this suggests that the majority of PathFX gene-phenotype relationships are not derived from GWAS studies.

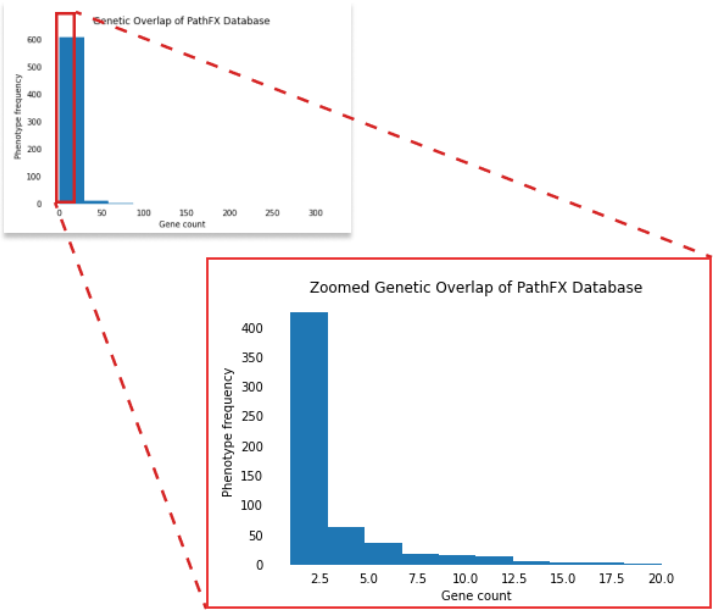

Figure 4. Number of genetic overlaps / shared genes between diseases in the GWAS catalog and the PathFX database. Most diseases have a low number of shared genes between the two data sources.

3.3 GWAS genes are infrequent in PathFX predictions of drug's intend-to-treat diseases

After exploring the extent to which PathFX contained GWAS data, we next sought to understand how the algorithm used this information in predicting drug-phenotype relationships for approved drugs. To narrow our search, we used the supplemental file from the original PathFX publication that contained correctly identified drug-disease pairs from a list of approved drugs. Specifically, this dataset contained 2,871 rows where a predicted disease is sufficiently similar to a drug's intend-to-treat disease (for determination of similarity, see Wilson et al 2018); it was possible to have multiple rows for a single drug because PathFX often contained multiple disease phenotypes that were sufficiently related to a drug's intend-to-treat disease.

Each row has information including, but not limited to, drug name, PathFX predicted phenotype & Concept Unique Identifier (CUI), approved indication phenotype & CUI, and network predicted genes from PathFX. As an example, acebutolol is approved to treat *"hypertensive disease"* (C0020538) and is predicted by PathFX to treat *"essential hypertension"* (C0085580), a sufficiently similar disease phenotype. The PathFX prediction was associated with the acebutolol network genes: ADM, ADRB1, ADRB2, APOB, CALCA, CALCRL, DNM2, GCGR, GNAS, PTH, and RAMP1. Additionally, *"hypertensive disease"* as an approved phenotype also maps to PathFX's *"preeclampsia/eclampsia"* and is an example of how a drug may have multiple relevant pathway predictions from PathFX which causes a disease to potentially appear in multiple rows. The first 2 rows of Table 2 highlight this occurrence, where one drug can be used to treat multiple PathFX predicted phenotypes. The median number of network genes is 11, while the minimum and maximum number of network genes are 1 and 120 respectively.

Drug	PathFX Predicted Phenotype, CUI	Approved Indication, CUI	PathFX Network Genes	GWAS Genes	Shared Genes	Direct Drug Target
Acebutolol	Essential hypertension, C0085580	Hypertensive disease, C0020538	ADM, ADRB1, ADRB2, …	LYPLAL1-AS1, MTHFR, PGPEP1, …	ADRB1, GNAS	ADRB1
Acebutolol	Preeclampsia/eclampsia 1, C0032914	Hypertensive disease, C0020538	PTGER2, APOB, EGF, …	LYPLAL1-AS1, MTHFR, PGPEP1, …	None	None
Amantadine	Parkinsonism, C0242422	Parkinson Disease, C0030567	DRD2, DDC, APP, …	CLRN3, CACNA1b, NR3C2, …	None	None
Amantadine	Parkinsonism, C0242422	Parkinson Disease, Postencephalitic, C0030568	DRD2, DDC, APP, …	None	None	None
Amantadine	CNS disorder, C0007682	Parkinsonian Disorders, C002422	APP, C5, C5AR1, …	DAAM1, LINC01500, SP1, …	None	None
Quetiapine	Depressive Disorder, C0011581	Schizophrenia, C0036341	ABCB1, ADCYAP1, …	FGF8, VN1, ABCB1, …	ABCB1, CYP2D6, DRD2, …	DRD2, CYP2D6, HTR3A, …

Table 2. A subset of the Approved Indication Table with added information on the last 3 columns. The “Shared Genes” column is an intersection between “PathFX Network Genes” and “GWAS Genes”. The “Direct Drug Target” column is identifying which gene in the “Shared Genes” column is the drug’s direct target.

Of the *“essential hypertension”* genes, we found two - ADRB1 and GNAS - that were also associated with “*hypertension*” in the GWAS catalog. This demonstrated that PathFX used 2/11 GWAS-associated network genes to predict approved drug-phenotype associations. We next sought to understand this phenomenon more generally across PathFX predictions We performed Phenotype Matching (Section 2.3) to the PathFX dataset and found 1,705 (out of 2,871) drug-phenotype associations where the PathFX phenotype matched a GWAS entry (Table 2, full data

in Supplemental File 1). From the 1,705 associations, we found 659 associations with at least one GWAS gene. Grouping the 659 associations by disease group, we found 34 unique disease groups with an average of 2 GWAS genes. To understand drug-phenotype associations with the greatest number of GWAS genes, we truncated our table to associations with 5 GWAS genes. Filtering the table yields us with 71 associations with 11 unique disease groups listed in Table 3 below. For example, inflammatory bowel and related diseases had relatively high numbers of GWAS genes: PathFX predictions for inflammatory bowel disease (IBD), ulcerative colitis and crohn's disease drugs had 7.2, 5, and 7.33 GWAS genes on average, respectively. Relative to the number of network genes used to support a PathFX prediction (a median of 11 genes), the number of GWAS genes represents a small portion of these genes (a median of 1 gene). Taken together, this suggests that GWAS studies are underrepresented in network predictions.

Disease Group	GWAS-PathFX Average Overlap Count
Asthma	6.44
Bipolar Disorder	5.00
Crohn Disease	5.00
Diabetes Mellitus, Non-Insulin-Dependent	5.80
Inflammatory Bowel Diseases	7.20
Multiple Sclerosis	6.00
Nicotine Dependence	5.50
Parkinson Disease	6.83
Rheumatoid Arthritis	6.67
Schizophrenia	8.00
Ulcerative Colitis	7.33

Table 3. Diseases that are most genetically enriched within the PathFX's approved indication network. The second column is calculated by taking the average gene overlaps among a disease's multiple rows.

3.4 GWAS genes are rarely direct drug targets

There is already evidence that GWAS studies don't often recover druggable targets, limiting the utility of these studies for finding therapeutic targets[2]. Instead, some studies have emphasized finding druggable targets with functional relationships, including network associations via protein-protein interactions, to GWAS hits. After discovering GWAS genes in PathFX networks, we also sought to understand how many of these GWAS genes were direct drug targets. Out of 659 associations where GWAS and PathFX share at least one gene, 34% (225 associations) have at least one direct drug target. In the case of the anti-hypertensive drug, acebutolol, the PathFX network contains two GWAS genes: ADRB1 and GNAS (**Table 2**). ADRB1 is a receptor predominantly located in the heart that is meant to control epinephrine level. It has reportedly been involved in multiple heart failure incidences. GNAS, on the other hand, is a locus that provides instruction for a stimulatory alpha subunit of a guanine nucleotide binding protein. We found that ADRB1 is a direct target of acebutolol, whereas GNAS is a downstream network protein added by PathFX (first row and last column of Table 2). Our analysis confirms that GWAS genes were infrequently drug targets and were more often found in proteins downstream of drug targets, yet they are still infrequently associated with approved drug-phenotype associations.

3.5 GWAS genes in PathFX Network are more frequent than random

Because GWAS genes were infrequently associated with PathFX predictions, we next estimated the chance of discovering a GWAS gene in a PathFX prediction. After performing 100 simulations of randomly matching PathFX and GWAS phenotypes, we ran a Mann-Whitney U test and obtained a p-value of 0.0 and a U-value of 491,436,079. Since the null hypothesis is that the two distributions (real and simulated) are likely to derive from the same population, we can reject them after finding a p-value of less than 0.05, thus accepting the alternative hypothesis that the occurrence of GWAS genes in PathFX network predictions is greater than expected by chance. This suggests that GWAS genes may have meaningful associations with network drug-disease associations.

4 Conclusion

The development of sequencing technologies gave rise to studies like GWAS which aims to identify genetic variants that might be responsible for a disease (and useful for guiding treatment). While these studies have been proven useful in many cases, it has not been able to recover many druggable targets[2]. As a better alternative for drug developments, scientists have turned to network models to better understand drug effects. Network approaches have improved the ability to connect disease associated variants to druggable targets[14]. In our network model of interest, PathFX, it's unclear which data - gene expression, gene mutation, proteomics - is the dominant contributor to a drug-phenotype relationship in the network. In this study, we measured how much a drug-phenotype relationship within the PathFX network is driven by genetics. We used the NHGRI-EBI GWAS Catalog as the basis of genetic information.

We discovered that PathFX as a network model generally does not heavily rely on genetic information from GWAS. This was done by identifying what genes are shared between the ones in the PathFX network and the ones in the GWAS catalog. While the median of PathFX network genes is 11, the median number of shared genes is 1, signifying that GWAS genetic information infrequently contributes to drug network predictions. Despite the overall low contribution, the two disease groups with the highest genetic component are autoimmune and psychological disorders. Both disease groups are complex and polygenic, meaning that it has been shown to require activation/deactivation of multiple genes. Our analyses seem to recover information that autoimmune disease and psychological disorder are more genetically enriched than other disease groups.

Diseases with the most overlaps

While our findings have mostly found that GWAS are underrepresented in PathFX's predictions, it might still be valuable to identify pathways that are most genetically enriched for some disease areas. The two disease groups whose network associations had a greater number of GWAS genes - as observed in Table 3 - are autoimmune diseases (multiple sclerosis, Parkinson's disease, rheumatoid arthritis, Crohn's disease (CD), ulcerative colitis (UC), inflammatory bowel disease) and psychological disorders (schizophrenia and bipolar disorder).

We found that schizophrenia has the most shared genes of 287, drastically higher than the median shared gene of 1. This could reflect increased interest in the disease; Schizophrenia has a total of 40 unique publications, much higher than the median of 1 publication for a disease. It could also reflect the strength of the gene variation in disease progression. Trubetskoy et al. reported 120 genes that are likely to drive the development of schizophrenia[15]. Schizophrenia also had a relatively high overlap between GWAS genes and PathFX network genes and had greater overlap than random. This suggests that diseases with a higher number of disease-associated variants in their networks may result from having more documented variants.

In comparison, of the autoimmune diseases, multiple sclerosis and crohn disease have a total of 20 and 23 publications respectively, showing that there's also a greater interest in understanding the genetic component of autoimmune diseases. GWAS was able to identify that the IL23R locus was associated with several autoimmune conditions[3]. This discovery has led to various developments of monoclonal antibodies targeting the interleukin-23 (IL-23) protein, one of the key proteins responsible for pathogenesis in CD and UC[3,13]. Interestingly, despite the fact that IL-23 has been used as a treatment target, it does not appear frequently in the PathFX network genes. This may suggest that network prediction mainly does not rely on genetic information, even if the gene has been shown to play an integral role in the disease development. This also shows that approved drugs are associated with proteins aside from the top proteins

discovered and validated by GWAS studies - instead, they tend to impact multiple neighboring proteins to ultimately achieve a similar effect.

Limitations

Decisions were bound to be made when performing scientific analyses. One decision that we had to make was to aggregate genes collected from different phenotypes. While we have demonstrated in section 3.1 that this assumption gave us a sufficient result, we would like to acknowledge that there are hundreds of other methods that may be used to better match phenotypes. We realized that there are natural language researchers tackling this issue and would like to admit that this is not the focus of our analysis. This method was chosen to best fit our data structure and currently existing pipeline. Additionally, we would like to acknowledge that while GWAS encompasses a great deal of genetic information, its low p-value cutoff standard tends to recover a few genes[8], which may be one reason why they share a very low number of genes with the PathFX network. Future consideration of including additional data sources such as DisGeNet or PharmGKB may help us better understand the overall contribution of genetic information to the PathFX network.

Supplemental File

File 1: Genetic Enrichment Table of approved indication drugs (*genetic_enrichment.csv)*

- ***Drug***: the name of the drug of interest
- ***PathFX Predicted Phenotype****:* phenotype that PathFX predicted to be treatable by *Drug*
- ***PathFX Predicted Phenotype CUI***: concept unique identifier (CUI) of the phenotype above
- ***Approved Indication***: phenotype that has been approved to be treated by *Drug* and similar to the *PathFX Predicted Phenotype*
- ***Approved Indication CUI****:* CUI of the phenotype above
- ***Network Genes****:* genes that appear in the PathFX network for *Drug*
- ***GWAS Genes***: aggregate of genes that appear in the GWAS catalog for *GWAS Phenotypes*
- ***GWAS Phenotypes****:* phenotypes that share the same CUI with the *Approved Indication CUI.* Multiple phenotypes are aggregated and delimited with a semicolon
- ***Shared Genes***: genes that appear in both *Network Genes* and *GWAS Genes*
- ***Shared Genes count***: number of *Shared Genes*
- ***Direct Drug Target***: genes/proteins that are direct target of *Drug*
- ***Direct Overlap***: genes that appear in both *Shared Genes* and *Direct Drug Target*
- ***Direct Overlap Count***: number of *Direct Overlap*

File 2: GWAS CUI to phenotypes (*cui_to_phens.pkl*)
Contains a Python dictionary with a CUI as the key and all phenotypes that maps to the CUI (delimited by semicolon) as the value

File 3: GWAS CUI to genotypes (*cui_to_genes.pkl*)
Contains a Python dictionary with a CUI as the key and all genes that belongs to the phenotype that corresponds to the CUI as the value

To access File 2 and File 3, use:

```
import pickle as pkl
filename = 'cui_to_phens.pkl' #change accordingly
with open(filename, 'rb') as f:
      loaded_dictionary = pkl.load(f)
```

References

1 Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, 17–21. https://pubmed.ncbi.nlm.nih.gov/11825149/ PMID: 11825149; PMCID: PMC2243666

2 Cao, C., & Moult, J. (2014). GWAS and drug targets. *BMC Genomics*, *15*(S4). https://doi.org/10.1186/1471-2164-15-s4-s5

3 Dugger, S. A., Platt, A., & Goldstein, D. B. (2017). Drug development in the era of precision medicine. *Nature Reviews Drug Discovery*, *17*(3), 183–196. https://doi.org/10.1038/nrd.2017.226

4 Henriksen, M. G., Nordgaard, J., & Jansson, L. B. (2017). Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. *Frontiers in Human Neuroscience*, *11*(322). https://doi.org/10.3389/fnhum.2017.00322

5 Ikegawa, S. (2012). A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going. *Genomics & Informatics*, *10*(4), 220. https://doi.org/10.5808/gi.2012.10.4.220

6 Jackson, M., Marks, L., May, G. H. W., & Wilson, Joanna B. (2018). The genetic basis of disease. *Essays in Biochemistry*, *62*(5), 643–723. https://doi.org/10.1042/ebc20170053

7 Jia, P., & Zhao, Z. (2013). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human Genetics*, *133*(2), 125–138. https://doi.org/10.1007/s00439-013-1377-1

8 Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2010). dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, *27*(1), 95–102. https://doi.org/10.1093/bioinformatics/btq615

9 Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19653-5

10 Lvovs, D., Favorova, O. O., & Favorov, A. V. (2012). A Polygenic Approach to the Study of Polygenic Diseases. *Acta Naturae*, *4*(3), 59–71. https://doi.org/10.32607/20758251-2012-4-3-59-71

11 MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

12 Mcinnes, B., & Pedersen, T. (2007). *Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain Article in AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium ·*. AMIA.

13 Parigi, T. L., Iacucci, M., & Ghosh, S. (2022). Blockade of IL-23: What is in the Pipeline? *Journal of Crohn's and Colitis*, *16*(Supplement_2), ii64–ii72. https://doi.org/10.1093/ecco-jcc/jjab185

14 Pech, R., Hao, D., Po, M., & Zhou, T. (2017). *Predicting drug-target interactions via sparse learning*.

15 Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., & Wu, Y. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508. https://doi.org/10.1038/s41586-022-04434-5

16 Wilson, J. L., Racz, R., Liu, T., Adeniyi, O., Sun, J., Ramamoorthy, A., Pacanowski, M., & Altman, R. (2018). PathFX provides mechanistic insights into drug efficacy and safety for regulatory review and therapeutic development. *PLOS Computational Biology, 14*(12), e1006614. https://doi.org/10.1371/journal.pcbi.1006614