AN INTEGRATIVE PIPELINE FOR MULTI-MODAL DISCOVERY OF DISEASE RELATIONSHIPS

BENJAMIN S. GLICKSBERG[1,2], LI LI[1], WEI-YI CHENG[1], KHADER SHAMEER[1], JÖRG HAKENBERG[1], RAFAEL CASTELLANOS[1], MENG MA[1], LISONG SHI[1], HARDIK SHAH[1], JOEL T. DUDLEY[1,2], RONG CHEN[1]

Department of Genetics and Genomic Sciences[1]
Department of Neuroscience[2]

Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl.
New York City, NY 10029, USA
Email: Rong.Chen@mssm.edu

In the past decade there has been an explosion in genetic research that has resulted in the generation of enormous quantities of disease-related data. In the current study, we have compiled disease risk gene variant information and Electronic Medical Record (EMR) classification codes from various repositories for 305 diseases. Using such data, we developed a pipeline to test for clinical prevalence, gene-variant overlap, and literature presence for all 46,360 unique diseases pairs. To determine whether disease pairs were enriched we systematically employed both Fishers' Exact (medical and literature) and Term Frequency-Inverse Document Frequency (genetics) methodologies to test for enrichment, defining statistical significance at a Bonferonni adjusted threshold of ($p < 1x10^{-6}$) and weighted $q<0.05$ accordingly. We hypothesize that disease pairs that are statistically enriched in medical and genetic spheres, but not so in the literature have the potential to reveal non-obvious connections between clinically disparate phenotypes. Using this pipeline, we identified 2,316 disease pairs that were significantly enriched within an EMR and 213 enriched genetically. Of these, 65 disease pairs were statistically enriched in both, 19 of which are believed to be novel. These identified non-obvious relationships between disease pairs are suggestive of a shared underlying etiology with clinical presentation. Further investigation of uncovered disease-pair relationships has the potential to provide insights into the architecture of complex diseases, and update existing knowledge of risk factors.

1. Introduction

With growing genetic and epidemiological knowledge of diseases, it is becoming increasingly vital to develop tools to integrate the bodies of data to better understand clinically relevant connections among diseases. In many instances disease factors are studied independently, but when integrated they have the potential to reveal new disease understanding and accelerate translational findings.[1] Previous studies have created disease networks such as the "human disease network"[2], which has already been successful in identifying molecular relationships between phenotypically distinct diseases. The researchers formulated disease-specific functional modules by assessing similarity metrics between all disease genes to all genetic disorders, an approach not yet performed at such a large scale. A similar study was extremely successful in associating disease comorbidities within an extensive collection of Electronic Medical Records (EMR) to known genetic variants in complex and Mendelian disorders to infer novel information of disease etiologies[3]. Disease network methods are also utilized for drug repurposing[4], in which drugs that are labeled to treat one disease are repurposed to treat another that it is linked to in the network[5,6,7]. One such analysis utilized a computational network analytical approach to identify that the anticonvulsant topiramate was beneficial in treating irritable bowel disease[8]. Disease networks are also used to discover unidentified components of disease risk[9].

For the current study, we accumulated a unique database of disease-causing gene variants and performed statistical analysis to determine shared genetic architecture between diseases. We then

overlaid an epidemiological enrichment analysis of co-occurrence rates in an EMR database. We hypothesize that disease pairs that have enrichment within both statistical tests are of highest interest for both known and yet unidentified connections. These unidentified connections will identify new and update existing knowledge of risk factors, elucidate disease mechanisms of action, and provide insight on relative environmental and genetic contribution to disease acquisition.

2. Methods

The workflow of the experiment is displayed in Figure 1. The major components of this analytical pipeline include the gathering, organizing, merging, and analyzing of disease-related data from multiple sources. The following sections will detail the process of each.

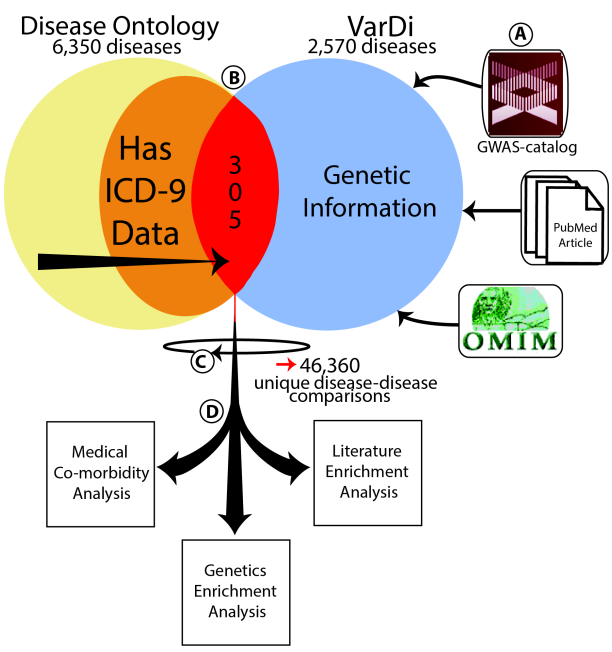

Fig. 1. The overall pipeline for the current experiment. The sources, integration, and analysis of the data are labeled alphabetically and described in detail within the text. *A*) data sources; *B*) data organization and filtering; *C*) unique combinations of disease pairs; *D*) enrichment analyses.

2.1. *Data Sources*

While the data (Figure 1, *A*) for the current project came from various sources (described below), they can be classified into two groups pertaining to the type of information they contributed: medical or genetic. Due to space constraints, every disease is shown within Figure 4.

2.1.1. *Disease Ontology*

Disease Ontology[10] (DO) is an open-source repository for integrated information relating to human disease, including but not limited to: OMIM identifiers, International Classification of Diseases (ICD)-9-CM codes, and Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) codes and has been extensively used in large-scale disease analyses[11]. This repository was especially useful for EMR-related portion of the study, namely the disease to ICD-9 mappings. As there are many criticisms and problems with using the ICD-9-CM classification system[12], particularly when dealing with rare and/or recently discovered diseases, it was necessary to utilize a pre-curated ontology of established and documented mappings for clinical studies[13]. To address this challenge, we filtered from DO all diseases that: 1) had either a direct mapping or an exact synonym match to at least one ICD-9-CM code and 2) present in our genetic information database. At the time of acquisition (June, 2014), DO contained a total of 6,351 unique diseases, with 2,333 of them having at least one ICD-9 map directly or to an exact synonym.

2.1.2. *Genetics Repository – VarDi*

Our curated genotype-phenotype repository, which we have named VarDi, contains phenotype-gene-mutation mappings for 2,570 diseases. This repository combines information from public online resources, specifically GWAS-catalog[14] and the Online Mendelian Inheritance in Man (OMIM)[15] (acquired June, 2014), as well as a proprietary disease-variant database, built through a combination of a Hadoop-based text mining tool and manual curation. This database is comprised of 24,111 Single Nucleotide Polymorphism (SNP) mutations ($p<1x10^{-8}$) associated within 3,661 genes in 893 distinct phenotypes with all associations.

Among the online resources, GWAS-catalog was comprised of 4,831 SNPs ($p< 1x10^{-8}$) within 1,838 genes in 776 phenotypes, although some being traits and not diseases. While OMIM contained 5,082 phenotypes, it is difficult to ascertain how many were distinct as subtypes of the same disease were encoded as separate entries. Nonetheless, these phenotypes encompassed 4,211 mutated genes.

2.2. *Merging Datasets*

With all the disease-gene variant data organized, there were three methods to connect the two repositories, as depicted in Figure 2 and Figure 1, *B*. From the DO repository, almost all diseases had at least one associated CUI number by which diseases in VarDi were matched to. If there was any discrepancy or more than one possible match due to multiple CUI numbers, the one with the closest name was used. Additionally, most DO entries had at least one associated OMIM code. These diseases were labeled by the DO entry and genetic info taken from each OMIM entry individually. Based on criteria listed above, there were 305 unique disease entries in total that were merged from DO and VarDi that we had both genetic data and at least one associated ICD-9 code. Most diseases were compiled using OMIM matching method (242/305), but CUI matching also added some (63/305).

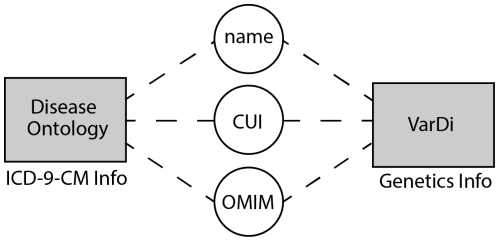

Fig. 2. The three possible components to link Disease Ontology entries to the proprietary database, VarDi.

2.3. *Statistical Analyses*

In the current study, statistical analyses were performed on every possible disease combination within the selected database to test for enrichment of three different components (Figure 1, *C*). From the 305 diseases in the database derived at in the previous section, there were 46,306 possible unique combinations. The three types of statistical analyses (Figure 1, *D*) will be described in detail in the following section.

2.3.1. *Genetics Analysis*

A primary aim of the current study is to elucidate novel genetic determinants or makers of diseases. Diseases that are genetically linked, in this study defined as having disease causing mutations within the same gene, can facilitate stronger understanding of each disease etiology alone and also add knowledge of new risk factors. The overarching rationale for the notion that disease gene variant association between diseases is indicative of a functional relationship is highlighted in Goh et al.'s (2007) work in generating a human disease network. In our database of 305 diseases, we compiled a total of 1,496 unique genes with disease-causing mutations. The

distribution of the number of genes per disease (6.8 genes per disease on average) was not normalized which was expected due to the unbalanced mix of Mendelian and complex diseases. As expected, broader diseases, such as rheumatoid arthritis and schizophrenia, contained more genes than more specific ones, such as restrictive cardiomyopathy.

To determine this relationship between all unique disease pairs, we directly adopted an extremely informative statistical enrichment methodology from Li Li *et al*. (2014)[16]. This study was successful in uncovering risk factors relating to diseases and traits through shared genetic architecture. In their original study[17], they first utilized the Term Frequency-Inverse Document Frequency methodology[18], which weighs the relative frequency of a gene within a disease in proportion to its frequency among all diseases in the database. To test the statistical significance of these scores, we computed a False Discover Rate (FDR) by randomly shuffling (10,000 times) the genes across all diseases. The q-value was calculated as the ratio of the expected number of false positives over the total number of hypotheses tested[19]. For this study, the significance threshold of $q<0.05$ was used.

2.3.2. *Medical/Epidemiological Analysis*

To determine if any two diseases were phenotypically linked, specifically if they co-occurred in a patient population more than one would expect by chance, we performed a statistical analysis on the patient pool within Mount Sinai Hospital's (MSH) EMR. The MSH is in a uniquely heterogeneous location and receives patients with a variety of phenotypes from diverse ethnicities. The Mount Sinai Data Warehouse (MSDW)[20], which houses all the clinical data, currently has 3,691,966 unique patients, over 16 million patient visits recorded, over 1.5 billion patient encounters, and 37,456,873 ICD-9 coded diagnoses documented.

Each of the 305 diseases has at least one associated ICD-9 code obtained from DO. In fact, the distribution of ICD-9 codes per disease is highly skewed towards one single code per disease, but some have multiple. While the average is 1.48 code per disease, this is highly affected by "categorical" diseases that encompass a range of codes, such as "gastrointestinal diseases " which has 60 associated codes respectively. Based on ICD-9 convention, each code could be from three to five numbers long with each proceeding number adding to specificity. As every patient is encoded with the most specific code possible (i.e. full five number code), if a disease had a code that was less than five digits, we automatically assigned that disease every possible five digit code extension.

For every possible unique disease pair combination, we performed a one-sided Fisher's exact statistical test to determine co-morbidity enrichment. The amount of patients that were observed to have at least one ICD-9 code from both diseases at any given time were compared with the amount of unique patients that had at least one ICD-9 code for each disease separately amongst a background group of any possible patient from the disease pool (559,708). A disease combination was deemed statistically significant if the resulting Bonferroni corrected *p*-value was less than 0.05.

2.3.3. *Literature Analysis*

To determine how well documented any given disease pair is in the scientific and medical world, a literature enrichment analysis was conducted using a text-mining tool. This tool queries

all abstracts and titles in PubMed for mention of a disease name using Simple Object Access Protocol (SOAP) to access NCBI's Entrez Programming Utilities[21]. A literature enrichment score was determined by performing a one-sided Fisher's exact test on the amount of articles available for each disease pair combination. Each disease was queried in quotes and each disease pair with an 'AND' operator between the two quoted terms to ensure specificity. Specifically, the test compared how many articles returned for the pair to each disease separately (number alone – number together) amongst all unique PubMed articles in the disease space (3,722,357; all diseases queried with an 'OR' operator between quoted terms). A disease combination was deemed statistically significant if the resulting Bonferroni corrected *p*-value was less than 0.05.

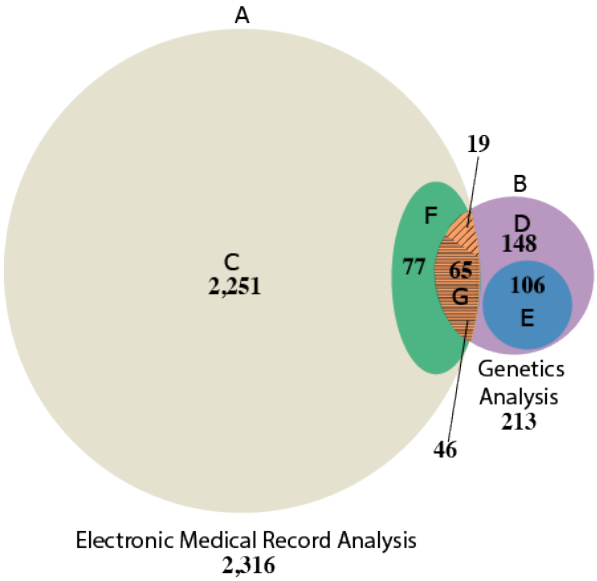

Fig. 3. The distribution of significant disease pair amounts per type of analysis. The letters and shaded regions correspond to subsets of results described in the Results section.

3. Results

Figure 3 displays the distribution of significant disease pair connections for EMR and genetics analyses with a literature score filter for pairs that were significant in both tests. As shown in *A*, there were 2,316 pairs of diseases that had a significant amount of co-morbidities in the EMR, while *B* reflects the 213 significant pairs that were enriched in the genetics analysis. *C* shows 2,251 pairs that were enriched in the EMR, but not genetically, while *D* represents the 148 pairs that were enriched genetically but not in the EMR. Key disease pairs of focus were those that had reached significance criteria in both genetic and medical enrichment analysis (*G*), specifically $q<0.05$ and $p<0.05/46{,}360$ respectively.

The two subsections, *E* and *F*, contain disease pairs that while not significant in both analyses, provide noteworthy results. By definition of the analytical procedure, all disease pairs in *G* must have both genes in common as well as co-morbidities in the EMR. The 106 disease pairs in *E* were not significantly enriched in the EMR, but had at least one co-morbidity instance. The 77 disease pairs in *F*, on the other hand, did not achieve significance in the genetics analysis, but had at least one gene in common.

Subsequently, the results in *G* were divided into two sections based on a significant literature enrichment analysis ($p<0.05/46{,}360$). Out of the 65 key pairs, 19 are suggestive of novel findings as they are not established in the literature (diagonal line section) while 46 are established in the literature (horizontal line region). The purpose of this distinction is to facilitate the highlighting of putatively novel disease pairs. Accordingly, we hypothesized that pairs that are not represented in the literature, or have an insignificant literature score ($p>0.05/46{,}360$) are of more interest and should be further pursued. Conversely, pairs with significant literature score ($p<0.05/46{,}360$) theoretically have a recognized link or relation and would be more akin to positive controls. A selection of these results of interest (*G*) is detailed in Table 1, separated by literature significance.

Table 1. **Selected Statistically Enriched Disease Pairs.** All possible 46,360 disease pairs were generated from the enrichment methodologies described above. The disease pairs listed in Table 1 have reached statistical enrichment in both Genetics and EMR analysis (**Gen**. $q<0.05$, **EMR** $p<0.05/46{,}360$ respectively). The top half is of disease pairs that are not enriched in the literature, while the bottom half are of disease pairs that have an enriched literature score (**Lit.** $p<0.05/46{,}360$). All reported p-values pass Bonferroni correction.

Disease Pair		EMR (p)	Gen. (q)	Lit. (p)
Coronary Artery Disease	Hypothyroidism	*0*	*0*	1
Lung Cancer	Nasopharynx Carcinoma	$6.35x10^{-07}$	*0.0476*	1
Lung Cancer	Hepatocellular Carcinoma	$3.16x10^{-11}$	*0.0168*	1
Cerebrovascular Disease	Factor V Deficiency	$3.80x10^{-11}$	*0.0022*	1
Hemorrhagic Thrombocythemia	Hypothyroidism	$6.67x10^{-15}$	*0.0092*	1
Esophageal Cancer	Hypertension	$6.30x10^{-17}$	*0.0270*	1
Esophagitis	Open-angle Glaucoma	$3.40x10^{-21}$	*0.0098*	1
Colorectal Cancer	Hepatocellular Carcinoma	$1.29x10^{-86}$	*0.0001*	1
Coronary Artery Disease	Alcohol Dependence	$1.26x10^{-08}$	*0.0106*	0.9999
Asthma	Sarcoidosis	$3.99x10^{-12}$	*0.0305*	0.9999
Myasthenia Gravis	Hypothyroidism	$9.78x10^{-11}$	*0.0099*	0.9991
Celiac Disease	Hypothyroidism	$1.96x10^{-22}$	*0.0062*	0.9785
Hemochromatosis	Variegate Porphyria	$3.01x10^{-12}$	*0.0078*	0.5262
Alcoholic Cirrhosis	Hepatic Steatosis	$2.86x10^{-97}$	*0.0017*	0.0021
Hypertrophic Cardiomyopathy	Limb Girdle Muscular Dystrophy	$1.04x10^{-12}$	*0.0365*	$3.523x10^{-05}$
Hemorrhagic Thrombocythemia	Myelofibrosis	$1.03x10^{56}$	*0.0033*	$5.99x10^{-09}$
Acute Lymphocytic Leukemia	Aplastic Anemia	$3.32x10^{-34}$	*0.0063*	$9.12x10^{-14}$
Velocardiofacial Syndrome	Tetralogy of Fallot	$4.33x10^{-32}$	*0.0056*	$2.77x10^{-16}$
Vitiligo	Hypothyroidism	$1.55x10^{-09}$	*0.0026*	$2.21x10^{-17}$
Systemic Lupus Erythematosus	Membranous Nephropathy	$1.42x10^{-13}$	*0*	$1.21x10^{-37}$
Ankylosing Spondylitis	Ulcerative Colitis	$3.13x10^{-15}$	*0.0468*	$1.20x10^{-43}$
Chronic Obstructive Pulmonary Disease	Lung Cancer	$2.87x10^{-262}$	*0.0064*	$7.44x10^{-43}$
Systemic Lupus Erythematosus	Myasthenia Gravis	$5.67x10^{-07}$	*0.0008*	$1.22x10^{-57}$
Hepatitis B	Primary Biliary Cirrhosis	$6.18x10^{-18}$	*0.0395*	$2.11x10^{-78}$
Ankylosing Spondylitis	Rheumatoid Arthritis	$7.41x10^{-38}$	*0*	*0*
Coronary Artery Disease	Myocardial Infarction	*0*	*0.0125*	*0*
Crohn's Disease	Ulcerative Colitis	*0*	*0*	*0*
Diabetes Mellitus	Hypertension	*0*	*0.0075*	*0*
Double Outlet Right Ventricle	Tetralogy of Fallot	$8.24x10^{-38}$	*0*	*0*
Systemic Lupus Erythematosus	Rheumatoid Arthritis	$2.88x10^{-263}$	*0*	*0*

3.1. *Subclass Cluster Interpretation*

While all disease pairs resulting in enrichment by either the genetics or EMR analyses (reflected in Figure 3) can be informative, those of highest interest are those that meet criteria in both the genetics and the EMR enrichment tests. The general distribution and amount of connections for each disease in all tests is shown in Figure 4. A disease pair that is enriched in such analyses is not only visible in a clinical population but can be further explored due to known genetic ties. Accordingly, genes that are present in one disease are natural candidates for exploration in the other.

All the 213 genetically enriched pairs (Figure 3, *B*) are interesting as they can inform the molecular mechanisms behind the relationship and the role of the gene variants in the diseases themselves. For instance, Cystic Fibrosis Transmembrane Regulator (CFTR) gene mutations have been implicated in both cystic fibrosis and bronchiectasis, which has facilitated better understanding of bronchiectasis etiopathogenesis[22]. This disease pair was genetically and clinically enriched in our results (EMR $p=3.06x10^{-48}$, Gen. $q=0.0043$). Further analysis of clustering diseases into groups with common genetic overlap can inform both new disease risk approximations based on genetic testing as well as better biological insight into the mechanisms of action[23]. Additionally, the 106 pairs that did not reach significance threshold in the EMR analysis but had at least one co-morbidity instance (as seen in Figure 3, *F*) are still worth consideration. If

they were trending towards significance, clinical enrichment might be achieved if a larger data set was used. In fact, 39 of the pairs had significant *p*-values ($p<0.05$), but did not pass Bonferroni correction. It is certainly possible that sample size or geography concealed additional co-morbidity instances that otherwise would have made the connection significant.

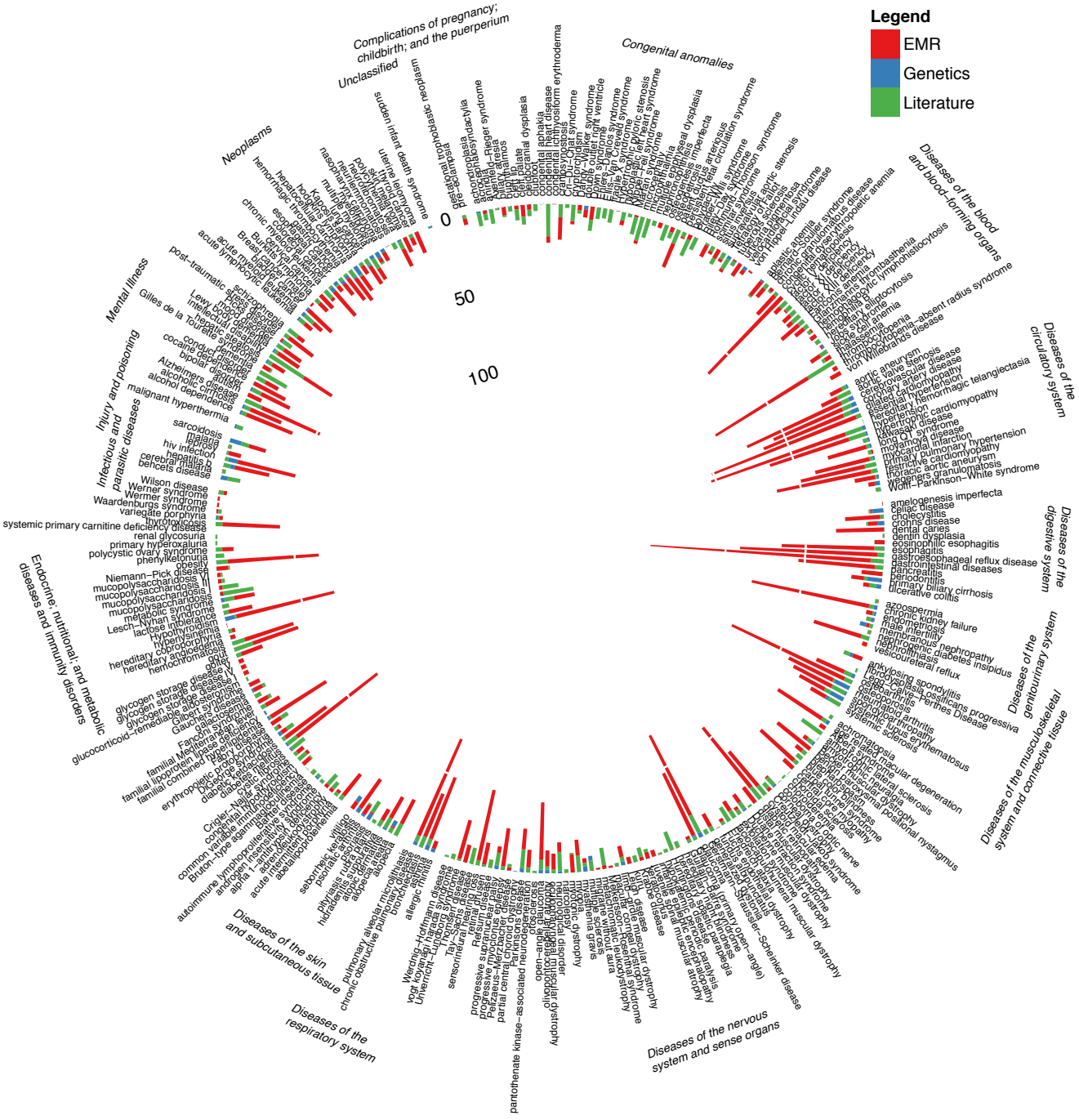

Fig. 4. The overall distribution and amount of significant pairs of each disease for EMR (red), genetics (blue), and literature (green) enrichments. The maximum amount of connections for each disease in each test is 304 (paired with every other disease). The stacked bar is the total number of connections for each test. White demarcations within stacked bar correspond to discrete, annotated y-axis numerical values.

The EMR enrichment analysis produced 2,316 significant disease pairs (Figure 3, *A*). With over 3.5 million patients and over 1.5 billion encounters in the MSH EMR, it is reasonable to expect the amount of connections. A quick glance at Figure 3 shows that there were over 10x more significant disease pairs in the EMR enrichment compared to genetics enrichment. This is to be expected, as much remains much unknown about the genetic etiology of diseases. All disease pairs that have EMR enrichment (*A*), both with (*G*) and without a genetic connection enrichment (*C*), could have genetic links that have not shown up because the genes themselves have not yet been implicated in the diseases. By keeping our genetic repository as updated as possible will help to expose these "hidden" genetic connections. It is clear that this indicates the need to further identify genetic etiologies of diseases.

Another hypothesis to explain the disease pair enrichment in the EMR alone is that some unrelated factor, namely environment, is responsible for the co-occurrences. This case is certainly possible for all EMR-enriched diseases (*A*), even ones that have shared genetic links (*C*). The manifestation of alcohol dependence and post-traumatic stress disorder (EMR p=6.17x10^{166}, Gen. q=1) is clearly dependent on context and situational history.

A hybrid model, which is reflective of the more realistic scenario of gene-by-environment interactions, can be especially used in cases of EMR, but not genetics, enrichment but where the disease pair shares at least one mutated gene, as in the 77 pairs displayed in Figure 3, *F*. One disease pair that reasonably follows this logic is coronary artery disease and obesity, which we found to be enriched in the EMR (p=0), but not genetically (q=0.6025), although they did have two mutated gene in common. Although a genetic element is believed to be involved in the co-expression of the disease pair[24], more emphasis has been placed on obesity itself being a high risk factor for coronary artery disease.[25] While genetic components for aspects of obesity are known[26] and suggest high heritability[27], environmental factors such as lifestyle and psychological mechanisms are also known to contribute largely to disease acquisition[28]. Accordingly, components such as these for one disease will invariably affect the co-occurrence for both.

4. Discussion

Our integrative analysis pipeline identified many unidentified disease pairs that suggest novel hypothesis for clinical and biological follow-up. We gathered data for 305 diseases from various repositories and performed three statistical enrichment analyses on genetic variant overlap, clinical presentation, and literature presence on each disease pair for a total of 46,360 unique comparisons. The analyses determined whether shared genes, co-occurrence rates, and common mentions in online articles occurred between a disease pair more often than by chance. Between the three analyses in the pipeline, we produced many interesting pairs including novel associations, which have illuminating implications based on the sub-cluster to which they belong.

The general distribution of the number of significant pairs one disease has for all three analyses can be seen in Figure 4. As expected from the proportion of significant pairs in each test, the EMR enrichment analysis is overrepresented compared to the others. This display is a convenient way to view hub diseases, or diseases that have a large amount of connections. These hub diseases are particularly interesting because the sheer amount of information associated with them can reveal intricate patterns between risk factors. The current study has produced many such hub diseases that require further exploration. A brief analysis on one candidate hub disease result is encouraging of meaningful disease pairs.

Hypothyroidism is an endocrine disorder in which the thyroid gland does not produce enough thyroid hormone. Beyond the typical symptoms, there is converging evidence of some of subsequent effects that are important to understand for proper treatment of the condition. In the current study, hypothyroidism is genetically and epidemiologically linked to many diseases, particularly autoimmune diseases listed in Table 1. Of particular interest is the connection to coronary artery disease (EMR p=0, Gen. q=0, Lit. p=1), where variations in five genes are common to both (*PTPN11*, *ATXN2*, *SH2B3*, *NAA25*, and *C12orf51*). While the literature score implies an unknown connection, there are recent studies that highlight the cardiovascular effects[29] of thyroid hormone levels and gland function. Specifically, hypothyroidism has been implicated as a risk factor for coronary artery disease[30], impairing cardiac function through mechanisms such as impaired vasodilation[31], which can lead to atherosclerosis[32]. Interpretation of findings such as these will be explored in subsequent sections.

4.1. *Efficacy of the literature analysis*

The literature enrichment portion of the study was used less as an independent measure and more as a way to highlight significant disease pairs that are relatively unknown to direct focus. In this sense, the literature enrichment score was useful and effective. Out of the 65 pairs that were significant in EMR and genetics analysis, 46 were and 19 were not enriched in the literature (p cut-off = 0.05/46,360). Based on our hypothesis, the former group would be already established and can serve as positive controls while the latter would be strong candidates for new findings. In fact, all disease pairs that were of the same characterization root (*i.e.* distal muscular dystrophy and limb-girdle muscular dystrophy) were captured in this filter. Overall the distinction was helpful but there were obvious oversights in each group. After manual review, almost all of the 46 literature enriched pairs had links or were related. Conversely, while many disease pairs in the non-enriched group had known links or relationship, the majority of the unrecognized connections were in this section. These indicated pairs will be discussed in the next section.

One clear limitation to this methodology is the syntax regarding the search criteria. Specifically, we did not filter negative mentions of the diseases, *i.e.* "disease *x*, but not disease *y*," which is actually the opposite of the intended effect. This issue will be addressed in future iterations of the pipeline. The inherent specificity of a disease name can also confound enrichment. Coronary artery disease and alcohol dependence, for instance, were found to be non-enriched in the literature, suggesting that the connection between them is not well documented. The search term 'alcohol dependence' could not be the best representation of the phenotype. Searching 'Coronary artery disease' and 'alcoholism' in PubMed produced 106 articles, instead of six yielded by the former query. This issue will only affect a small subset of the diseases, however. Further refining of the tool can rectify these issues via search optimization protocols.

4.2. *Efficacy of the EMR enrichment analysis*

It is important to note, however, that there are some clear factors that might have confounded this EMR analysis and led to spurious correlations. First, while MSH is in a location with a uniquely heterogeneous population, almost all patients are from the New York City area, which can introduce geographic disease bias. Another potential biasing factor has to do with limitations of how diseases are encoded and recorded. Some diseases are so rare or recently discovered that distinct codes do not exist for them. Crigler-Najjar syndrome and Gilbert's syndrome, for instance,

are similar but different disorders that both are encoded by the ICD-9-CM code 277.4, "Disorders of bilirubin excretion." Their high level of co-occurrence in the EMR (exact overlap to be specific) is a confounder due to lack of disease specificity and does not have any actual applicability. Unfortunately this issue cannot be readily addressed but will undoubtedly be less problematic when ICD-10 codes can be used instead.

4.3. *Notable Significant Disease Pairs*

For each of these key 65 pairs (Figure 3, *G*), extensive further evaluation was conducted to determine impact and relevance. Due to space constraints, a representative subset of notable pairs is displayed in Table 1. As mentioned, these results were split into clusters based on literature enrichment scores that provided a somewhat useful first pass filter to highlight previously unexplored disease pairs.

4.3.1. *Disease pairs that are well established act as strong positive controls for the pipeline*

Disease pairs across all disease categories that were enriched in the literature (and a few that were not) produced positive control connections that are well established and some that have recently been discovered. The relationship between Coronary artery disease and myocardial infarction is well established[33], and it is encouraging that our pipeline produced strong significance in all three analyses for this pair (EMR p=0, Gen. q=0.0125, Lit. p=0). Similarly, ulcerative colitis and Crohn's disease are known to be highly related[34] and showed up to be strongly significant in all three analyses (EMR p=0, Gen. q=0, Lit. p=0). Diabetes mellitus and Hypertension are also well linked and have been shown to co-occur more often than expected by chance[35] (EMR p=0, Gen. q=0.0075, Lit. p=0). Chronic obstructive pulmonary disease (COPD) and lung cancer are both common diseases amongst smokers[36] and patients with COPD are at increased risk for developing lung cancer[37] and have a robust link in our results (EMR $p=2.87\text{x}10^{-262}$, Gen. q=0.0064, Lit. $p=7.44\text{x}10^{-43}$). Furthermore, recent studies have offered hypotheses to explain the common origins to these "anatomic and functionally disparate diseases."[38] This is a clear example of how the current pipeline can be used to provide context for both genetics and epidemiological prevalence. Tetralogy of Fallot and velocardiofacial syndrome have both genetic links[39] and prevalent co-occurrences[40] and were enriched in our results (EMR $p=4.33\text{x}10^{-32}$, Gen. q=0.0056, Lit. $p=2.77\text{x}10^{-16}$).

4.3.2. *Unknown aspects of disease pairs are candidates for further analysis*

One of the best possible uses for this pipeline is for identifying candidate gene variants that can explain the link between diseases that have prevalent comorbidities. Esophageal cancer and hypertension are observed in patients[4142], but little is known about how they are related. As expected, this pair was not enriched in the literature, but had strong EMR and genetic overlap (EMR $p=6.30\text{x}10^{-17}$, Gen. q=0.027, Lit. p=1). Similarly, esophagitis and open-angle glaucoma do not often show up together in the literature but there is also very little, if any, information documenting any sort of connection between the two, yet our pipeline found a connection (EMR $p=3.40\text{x}10^{-21}$ Gen. q=0.0098, Lit. p=1.

5. Conclusions and Future Directions

The current study combined clinical, genetics, and literature analytical methods to create a pipeline to identify key disease pairs of interest. With this initial iteration of the pipeline, we identified 2,316 and 213 disease pairs that were enriched in the EMR and for shared genetic variants, with 65 in both. Using the analytical approaches listed above, we are able to infer new insights about mechanistic origin, molecular pathways, and risk factors for such pairs. A component to easily adapt to the pipeline is to stratify data based on ethnicity for disease expression to determine if disease pair prevalence is uniform or specific to particular ancestries. We also have categorical data of disease type that we will perform enrichment analysis on to establish if certain connections are more common in certain disease classes. More specific future direction plans are listed briefly in the final section.

5.1 *Predicting disease risk through temporal co-morbidity analysis*

A large component of Li *et al*.'s work is determining the time course of disease onset in related pairs to identify which is the casual risk factor. For all medically enriched pairs in our pipeline, we plan on incorporating the time line for the disease manifestation to see which is a risk for the other. For instance, we would hypothesize, based on current knowledge,[43] that obesity would manifest before diabetes.

5.2 *Utilizing connections between diseases for drug related analysis*

Another useful type of information that can be derived from this analysis is the case where the treatment for one disease can possibly cause the other. While this pipeline shows clinical presentation relationships, it does not tease out how the disease pairs are related. Systemic lupus erythematosus (SLE) and myasthenia gravis (MG) have an association, albeit a rare one, with a possible common biological explanation along with clinical co-presentation.[44] A review analyzed the association between these two diseases and offered the possibility that hydroxychloroquine, a drug typically used for the treatment of SLE, could have induced MG, at least in one patient.[45] Incorporating patient medication information into our pipeline, something that is feasible, in future iterations will be able to uncover these possible scenarios.

6. Acknowledgments

We would like to thank Mount Sinai's Scientific Computing Team for server cluster IT support. This research was supported by Institutional Grants of Rong Chen and Joel Dudley.

References

1. N. H. Shah, *et al.*, *BMC Bioinformatics*, **10** (2009).
2. K. I. Goh, *et al.*, *Proc Natl Acad Sci U S A* **104**, 8685 (2007).
3. D. R. Blair, *et al., Cell*, **155**, 70 (2013)
4. S. Suthram, *et al.*, *PLoS Comput Biol*, **6** (2010).
5. F. Iorio, *et al.*, *Proc Natl Acad Sci U S A*, **107**, 14621 (2010).

6. J. T. Dudley, *et al.*, *J Cardiovasc Transl Res*, **3**, 438 (2010).
7. L. Yang and P. Agarwal, *PLoS One*, **6** (2011).
8. J. T. Dudley, *et al.*, *Sci Transl Med*, **3**, 96 (2011).
9. A. Goris and A. Liston, *Cold Spring Harb Perspect Biol*, **4** (2012).
10. L. M. Schriml, *et al.*, *Nucleic Acids Res*, **40**, 940 (2012).
11. K. Shameer and R. Sowdhamini, *J Clin Bioinforma*, **2**, 8 (2012).
12. A. Hazlewood, *American Health Information Management Association* (2003).
13. K. Shameer, *et al.*, *Hum Genet*, **133**, 95 (2014).
14. L. A. Hindorff, A Catalog of Published Genome-Wide Association Studies [http://genome.gov/gwastudies].
15. Online Mendelian Inheritance in Man, OMIM [http://omim.org].
16. L. Li, *et al. Sci Transl Med*, **30**, 57 (2014).
17. L. Li, *et al.*, *Pac Symp Biocomput*, 224 (2013).
18. H. C. Wu, *et al.*, *Acm Transactions on Information Systems*, **26**, 13 (2008).
19. J. D. Storey and R. Tibshirani, *Proc Acad Sci U S A*, **100**, 9440 (2003).
20. Mount Sinai Data Warehouse [https://msdw.mountsinai.org/].
21. E. Sayers, E-utilities (2009) [http://eutils.ncbi.nlm.nih.gov/].
22. P. F. Pignatti, *et al.*, *Hum Mol Genet*, **4**, 635 (1995).
23. C. Cotsapas, *et al.*, *PLoS Genet*, **7** (2011).
24. J. D. Brunzell, *Arterioscler Throb Vasc Biol*, **4**, 180 (1984).
25. R. H. Eckel and R. M. Krauss, *Circulation*, **97**, 2099 (1998).
26. C. Bouchard and L. Pérusse, *Annu Rev Neur*, **13**, 337 (1993).
27. J. Wardle, *et al.*, *Am J Clin Nutr*, **87**, 398 (2008).
28. J. O. Hill and J. C. Peters, *Science*, **280**, 1371 (1998).
29. B. Biondi and I. Klein, *Endocrine*, **24**, 1 (2004).
30. O. Mayer Jr., *et al.*, *Vasc Heaith Risk Manag*, **2**, 499 (2006).
31. J. Lekakis, *et al.*, *Thyroid*, **7**, 411 (1997).
32. Y. Zhang, *et al., Anterioscler Thromb Vasc Biol* (2014).
33. E. G. Nabel and E. Braunwald, *N Engl J Med*, **366**, 54 (2012).
34. J. D. Doecke, *et al.*, *Inflamm Bowel Dis*, **19**, 240 (2013).
35. M. Epstein and J. R. Sowers, *Hypertension*, **19**, 403 (1992).
36. E. Potton, F. McCaughan, and S. James, *Resp Med: COPD Update*, **5**, 34 (2009).
37. S. Raviv, *et al.*, *Am J Respir Crit Care Med*, **183**, 1138 (2011).
38. A. M. Houghton, M. Mouded, and S. D. Shapiro, *Nat Med*, **14**, 1023 (2008).
39. F. Amati, *et al.*, *Hum Genet*, **95**, 479 (1995).
40. D. Young, R. J. Shprintzen, R. B. Goldberg, *Am J Cardiol*, **46**, 643 (1980).
41. L. B. Koppert, *et al.*, *Eur J Gastroenterol Hepatol*, **16**, 681 (2004).
42. T. Kato, *et al.*, *Hepatogastroenterology*, **48**, 1656 (2001).
43. J. M. Chan, *et al.*, *Diabetes Care*, **17**, 961 (1994).
44. G.Vaiopoulos, *et al.*, *Postgrad Med J*, **70**, 741 (1994).
45. M. Jallouli, *et al.*, *J Neurol*, **259**, 1290 (, 2012)