**SPECIFIC AIMS**

Genome-wide associations studies (GWAS) have identified >15,000 single nucleotide polymorphisms (SNP) for complex diseases. For most diseases, however, SNPs are located discretely throughout the genome each with a very small effect size. It remains elusive how these SNPs collectively perturb the biological system into a pathophysiological disease state, or more complicatedly, a comorbid disease state.

Classic genetic approaches, such as logistic regression, search for epistatic SNPs whose cumulative effect is non-additive from. These approaches generate only a few epistatic interactions between SNPs across all complex diseases because they search up to 1 trillion pairwise combinations of SNPs, thus requiring a vast number of samples to reach sufficient power. To circumvent this barrier, we and a few others have integrated external knowledge to reduce the size of hypotheses, such as functional annotations of the disease genes and downstream genes affected by SNPs. We have demonstrated that, although distantly located, SNPs associated with the same or comorbid diseases are likely to share similar biological mechanisms. For instance, multiple SNPs locating on chromosomes 6 and 21 commonly perturb immune response, which bridges the genetic mechanisms of rheumatoid arthritis. We called these SNPs cooperative and observed many epistases from them. Identifying these cooperative SNPs will greatly reduce the search space and allow better biological interpretability for the epistatic relationships identified in GWAS results.

Very few studies have reported these cooperative mechanisms even with the abundant functional data accumulated by the Encyclopedia of DNA Elements (ENCODE) project. Our longstanding goal is to identify the driving cooperative mechanisms of SNPs for the general scope of complex diseases and their comorbidities through integrative analyses of big omics data. **We hypothesize that multiple factor analysis of ENCODE data can unveil the cooperative and epistatic relationship of SNPs associated with the same or comorbid complex diseases.** Examples of these hidden factors commonly underlying ENCODE data include key transcription factors (TF) that modulate the chromatin accessibility and modification, and gene activation and expression of a cell. Our specific aims are:

***Aim 1: Integrate ENCODE data to prioritize SNP pairs with cooperative mechanisms for each complex disease****.* Identifying multiple cooperative SNPs allows for more accurate diagnosis biomarkers. **We hypothesize that cooperative SNPs of the same complex disease are similar to each other in some hidden, unrelated factors that determine the results of the multiple omics assays in ENCODE**. We will integrate major types of omics data in ENCODE by multiple factor analysis (MFA) to identify cooperative SNPs, such as chromatin modification and TF binding, for about 300 diseases in NHGRI GWAS catalog. We plan to validate the top 10 SNP pairs in Electronic Medical Records and Genomics **(**EMERGE) dataset. The dataset comprises both genotypic and phenotypic data for thousands of patients thus allowing for patient level validation of epistatic disease biomarkers.

**Aim 2: Integrate ENCODE data to determine the genetic basis for pairs of comorbid complex diseases**. Complex diseases, such as obesity and diabetes, are often comorbid in patients. Comorbidity increases the risk of mortality and rehospitalization, which complicates the treatment and burdens patients. From our published work and preliminary results, we have shown that comorbid complex diseases are more likely to associate with disease genes of similar functions. However, only SNPs within coding genes have been studied for their effects to comorbidity, while we know little about the roles of intergenic SNPs in disease comorbidity. Also, our knowledge about genetic drivers of comorbidity is limited to DNA regions rather than specific causal SNPs. Here, we intend to understand the role of intergenic SNPs to comorbidity. **We hypothesize that common genetic mechanisms underlie many comorbid complex diseases, including common SNPs and genes, and commonly perturbed mechanisms by cooperative SNPs**. We will integrate ENCODE data to quantify the pairwise functional similarity among 600 complex diseases. Then, we will examine the concordance between the disease similarity and the comorbidity in two independent electronic medical record datasets for which we have already obtained access. Again, using EMERGE dataset, we will assess the effect sizes of disease comorbidities for the top shared or cooperative SNPs. Validated SNPs of disease comorbidities are tentative biomarkers to predict future comorbid diseases based on existing patient conditions.

Successful completion of this grant will enhance our understanding to the causal mechanisms of complex diseases (Aim 1) and their comorbidities (Aim 2). More importantly, this work will generate new mechanistic biomarkers for disease diagnoses and disease progression.

**Strategy**

**Significance**

Complex diseases occur frequently and exist in a large proportion of the population. They are often comorbid with each other and complicate the treatment. Inappropriate and insufficient management of complex diseases and their comorbidities greatly increases the costs both for patients and medical providers. Genetics have been demonstrated to play an important role in the prevalence of complex diseases and their comorbidity [1], in addition to environmental exposure and socio-economical factors.

Genome-wide association studies (GWAS) have unveiled thousands of single nucleotide polymorphisms (SNPs) that have significant associations with disease phenotypes [2]. Recent functional studies such as Encyclopedia of DNA Elements (ENCODE) [3] and expression Quantitative Trait Loci (eQTL) [4] indicate that these SNP regions not only correlate, but also functionally contribute to these diseases [3] and their associated comorbidities [5]. However, each SNP only has a small effect size, and fails to explain a substantial proportion of heritability, even when their additive effects are taken into consideration. Several factors may cause the missing heritability, among which are the genetic interactions of these SNPs [6] and SNP-environment interactions [7]. Classic genetic approaches only focused the effects of single SNPs due to the vast number of possible combinations, which are measured in the trillions [11]. Recent approaches for genetic interactions, such as logistical regression [8], Lasso [9], and machine learning approaches [10],still have difficulties in detecting the interactions from the large set of combinations and have only generated very small number of epistaic results to date.

On the other hand, some studies have implied or reported a few shared or linked mechanisms between SNPs associated with complex diseases [12] and comorbid diseases [13, 14], such as cooperative mechanisms between SNPs of proximity and linkage disequilibrium [15], and shared or interacting transcription factor binding to SNPs associated with the same diseases [12]. This enables a functional study of SNPs and their epistases. However, the comprehensive picture of biological perturbation mechanisms by SNPs for complex diseases and their comorbidities remains vague, probably due to lack of integration between genetic studies and functional studies, and lack of integration of latest results in GWAS for disease comorbidities. Furthermore, the genetic mechanisms of comorbidity have been rarely studied, which prevent effective intervention to reduce the cost of health care based on personal genomics [7] providing an untapped source of potential interventions for the reduction in health care costs and the improvement of health outcomes.

Continuing to treat each SNP independently to investigate disease physiology fails to capitalize on GWAS data because the effect of each individual SNP is minor. Continued treatment of each disease as genetically independent and each type of study as separate also prevents translating the genetic results to a biological interpretation and/or clinical application. Our research will address this gap in knowledge by identifying common or cooperative mechanisms among distantly located, genetically independent SNPs, for the same or comorbid diseases. **We hypothesize that multiple factors analysis (MFA) of ENCODE data can unveil the cooperative and epistatic mechanisms of SNPs underlying the same or comorbid complex diseases.** The rationale for this hypothesis lies in that causal SNPs should perturb some key factors critical to a disease or comorbid disease, which in turn drive assay detectable changes of biological entities at multiple related scales, which can be observed by a variety of assays, including epigenetic effects from environmental pressure. Therefore we propose to integrate the abundant, multiple-scale genomics and epigenetics data about DNAs in the ENCODE repository, by balancing the effects from multiple scales and efficiently handling many missing assays. We will develop an algorithm that combines the advantages of the statistical method of MFA and classic machine learning techniques, to mine the information buried in the complexity of existing datasets. We will also validate the top relationships in a large volume database consisting of both phenotype and genotype data at the single patient level.

In summary, the completion of this project will be significant because this method will result in the discovery of causal biological mechanisms behind complex diseases and their associated comorbidities. The results could also serve as biomarkers for disease diagnoses and disease progression with better accuracy than current biomarkers based on additive effects of SNPs, and will provide generalized approaches for data integration, big data analysis, and translational medicine, which fit well with recent NIH initiatives.

**Innovation**

With the advance of high-throughput techniques, a barrier for developing treatments for disease lies in the development of effective and efficient data integration. The integration demand not only arises from multiple scale assays of a single project like ENCODE, and The Cancer Genome Atlas (TCGA) [16], but also from knowledge integration from several projects such as various GWAS projects, GTEx [4], and ENCODE. Multiple factor analysis has provided a framework for many novel mathematical methods [17, 18] but has only recently been used in data integration of multiple scale data [19]. Also, our proposed combination of multiple factor analysis and machine learning techniques for epistasis studies, such as K-nearest neighbor and clustering, has never been attempted. The method combining multiple factor analysis and clustering techniques is not only applicable to epistatic analysis, but also general enough for other omics data integration from a variety of applications, such as disease subtype identification, signature pathway analysis, and prognosis and survival analysis. We will provide source code for implementing our methods for other potential. In terms of the epistatic underpinnings of complex disease and comorbidity, the integrative analysis of functional data in ENCODE can: 1) dramatically reduce the search space by only testing SNPs with known associations to diseases, and 2) benefit future biological validation and clinical intervention.

**Approach**

***Aim 1: Integrate ENCODE data to prioritize SNP pairs with cooperative mechanisms for each complex disease***

|  |
| --- |
| Macintosh HD:Users:haiquan:Dropbox:study:IMB521:write up:figures:best_results_2overlap_3_snps.pdf |
| **Figure 1**. **Example MFA results**. First three factors (>77% variances) of disease-associated SNPs on chromosome 6 from five TFs (16 replicates in total) and five replicates of chromatin accessibility data. Although many SNPs are clustered together due to sparse functionality on these five TFs, SNPs associated with systemic sclerosis, hypothyroidism are close with each other in the first three factors, indicating similar functions among the five TFs. |

*Hypothesis and rationale*: The cooperative mechanisms of multiple independent SNPs for complex diseases are largely unknown. Classic genetic methods are unable to detect epistasis among candidate SNPs due to the vast number of combinations (up to several trillion pairs). Expression Quantitative Trait Loci (eQTL; Table 1) data allows unveiling the functional linkage among SNPs for complex disease but are unable to distinguish between confounders and causal SNPs. The ENCODE project generated abundant functional annotations for DNA on multiple scales and is thus promising to unveil the causal SNPs and their interactions. **Our objective is to identify functional and cooperative SNPs for complex diseases for more accurate diagnosis**. *We hypothesize that cooperative SNPs of the same complex disease should be similar to each other on multiple hidden, unrelated factors that drive the results of the multiple omics assays in ENCODE.* We will use multiple factor analysis to identify the unknown driving factors and measure the similarity in these factors for any pair of candidate SNP pairs by using multiple scale assays on hundreds of cell lines in the ENCODE dataset. The successful completion of this aim will fill the gap of unknown systematic mechanisms of complex diseases, and generate many testable biomarkers for diagnosis of complex diseases. Upon the completion of this aim, it is our expectation that we will identify hundreds of functionally cooperative SNPs for complex diseases, allowing high throughput discovery and validation of epistasis for the first time.

*Preliminary results:* The functional linkage among multiple independent SNPs associated with the same complex diseases is largely unknown. Our preliminary studies of eQTL data from lymphoblastoid cell lines suggested that SNPs associated with the same diseases are more likely to be similar with each other in the biological processes and molecular functions of their perturbed downstream genes. These cooperative mechanisms between SNPs may contribute to the causal mechanisms of complex diseases. Indeed, using both classic genetics and machine learning methods on case-control patient data, we found statistically significant cooperative SNPs have epistatic effects in rheumatoid arthritis (p=0.02), Alzheimer’s diseases (p=0.046), and bladder cancer (0.039). Since eQTL associations cannot indicate a causal relationship, we endeavored to identify causal SNPs and their cooperative biological mechanisms at play. We extended our study on individual SNPs to their high linkage disequilibrium regions (LD; correlation r2>0.8) since SNPs in the LD regions of proxy SNPs are almost equivalent in association with any disease. Our preliminary results indicated disease associated SNPs on chromosome 6 tend to cluster together for the same disease, or same disease class, using multiple factor analysis (MFA) of the chromatin accessibility and transcription factor binding data from the B-lymphocyte cell line GM12878 (Fig 1). The preliminary work suggested similar binding activity of the clustered SNPs on the five transcriptional factors for the same disease, and demonstrated the feasibility of the MFA algorithm on genome-scale computing (running time < 2 minutes).

*Approach*

|  |
| --- |
| **Macintosh HD:Users:haiquan:Dropbox:study:IMB521:write up:figures:Figure2-Aim1_v2.pdf** |
| **Fig. 2. Processing flow of Aim 1 for a cell type**. It consists of three steps: 1) Project multi-scale ENCODE data of a cell line to a data matrix at the SNP level. 2) Map disease associated SNPs of the cell line to several major factors using multiple factor analysis. 3) Average the distance of a SNP pair on all cells of the same type and conduct resampling for statistical significance. |

|  |  |
| --- | --- |
| Table 1: Definitions used in the text | |
| Abbreviations | Descriptions |
| MFA | Multiple factor analysis |
| ENCODE | Encyclopedia of DNA Elements |
| HCUP | Healthcare Cost and Utilization Project |
| EMERGE | Electronic Medical Records and Genomics |
| SNP | Single Nucleotide Polymorphisms |
| GWAS | Genome-wide Association Study |
| eQTL | Expression Quantitative Trait Loci |
| LD | Linkage disequilibrium |
| EMR | Electronic Medical Records |
| OR | Odds ratio |
| FET | Fisher’s Exact Test |
| TF | Transcription factor |

**SA 1.1 Calculate the overall ENCODE similarity for pairs of disease associated SNPs using multiple factor analysis**

We will employ three types of data as input for this aim: SNP-disease associations, SNP linkage disequilibrium, and the ENCODE functional repository (Fig. 2). First, we will download disease associated SNPs from the National Human Genome Research Institute (NHGRI) GWAS catalog. This catalog consists of the latest results across almost all GWAS studies. It comprises over 1200 diseases and 15,000 SNPs (lead SNPs) as of Oct. 2015. Second, we will download data from the 1000 genome project and calculate the linkage disequilibrium for all lead SNPs on major populations, specifically Caucasian, African American, Hispanic, and Asian. Using a pipeline we have already built on the PLINK package, we will extract SNPs with strong disequilibrium (r2>0.8) to the 15K SNPs. SNPs with strong linkage disequilibrium to lead SNPs are also likely to associate with the same diseases of the lead SNPs, some of which may have more functional roles than the corresponding lead SNPs [20]. Third, we will collect major assays from the ENCODE data repository, such as the DNASE I hypersensitive site assay, histone modification and transcription binding by Chip-seq, and RNA-seq. Although comprised of multiple assays on over one hundred cell lines, functional assays in the ENCODE project are highly sparse, a function of the investigators’ interests and resources. We tailor our approach specifically for this challenge. We will employ existing resources as much as possible for each cell type, some of which have multiple individuals such as lymphoblastoid cell lines and several cancer cell lines. For each cell line, we will map the signals with a solution of a small region or single nucleotide level on the stretch of DNA, to individual SNPs these signals may be related. We assume that the function of the SNPs are very related to the proximate signals, since most SNPs’ minor alleles are likely to perturb the signals of the regions if the SNP is functional. Finally, we will generate a matrix for all studied lead and LD SNPs for each specific cell line consisting groups of available assays with each assay in a column (Fig. 2).

We will apply multiple factor analysis on the above SNP matrices. We will combine evidence from multiple cell lines of the same type but handle distinct cell types separately. For a specific cell line, we will apply multiple factor analysis on multiple groups of assays for the function similarity of any pair of disease associated SNPs . Multiple factor analysis (MFA) is a well-established extension of factor analysis in multiple variable analyses. While factor analysis has been used in computational biology [19], MFA has been rarely used in epistasis analysis and data integration. However, it is an unexplored tool perfect for integrative epistatic analysis. In case of multiple transcription factor binding assays of a cell using Chip-seq, some TFs may correlate with each other. Thus, factor analysis can identify those uncorrelated factors (e.g. driving TFs) that can represent the activities of all assayed TFs. Multiple factor analysis extends this methodology. It deals with distinct groups of assays and aims to avoid the dominance from a single group by balancing the influence across all groups. For instance, histone modification is crucial to the function of DNAs and is also related to the activities of transcription factors. In the joint study of histone modification and transcription factor binding, multiple factor analysis identifies common factors that derive both histone modification and transcription factor activity assays. MFA assumes any single assay from each group is a linear combination of the common factors. It can also represent a SNP as a linear combination of the common factors, thus mathematically projecting a point to the space of common factors [21].

Escofier and Pages developed balance-based multiple factor analysis in the 1980s [18]. It conducts principal component analysis on each group separately and then use the first eigenvalue of a group to weight the influence of all variables (e.g. assays) in the group to avoid dominance of the group [18] [21] . Mathematically, a group (k) of assays can be represented as follows (Eq. 1) [21]:

where Xk is a n x mk matrix of n SNPs and mk assays from the group k, σk1 is the largest eigenvalue of Xk. T is the n x R matrix of R factor scores, shared across all groups, Pk is the mk x R loading matrix to R factors, Ek is a n x mk matrix of residuals in this group, and tr means transpose of the loading matrix.

Several R packages provide implementation of the algorithm including: FactoMineR [22], PCAmixdata [23], and ade4 [24]. As demonstrated in our preliminary results, FactoMineR is scalable to the genome scale and suitable for big data analysis. Moreover, it can integrate groups with various numbers of assays, even if there is only a single assay in a group (as seem from chromatin accessibility and RNA-seq data). Using FactoMineR, we will quantify the ENCODE function similarity of any SNP pairs with preprocessed signals using distance measurement that is inverse with the similarity measurement. For each cell type, the MFA based algorithm consists of the following five steps.

1. Conduct MFA on each matrix consisting of multiple types of ENCODE assays from a cell line, using the top common factors that account for at least 75% of the variances of the data matrix.
2. Project each SNP, corresponding to a row of the matrix, into a point on the space consisting of the top common factors identified in Step 1). FactoMineR supports and outputs this mapping for any SNP.
3. Calculate the Euclidean distance between any pair of SNPs in the common factor space
4. For multiple cell lines of a cell type in the ENCODE data, such as lymphoblastoid cells, calculate step 1 to 3 separately for each cell line and average the distance on all cell lines of the same type for each SNP pair, even if the completed assays are different across these cell lines.
5. For different types of cells, conduct step 1 to 4 separately and develop a distance for each cell type respectively.

In short, SA 1.1 will generate a ENCODE distance (equivalent to similarity) score for each pair of trait-associated SNPs, including lead and LD SNPs, for each cell type studied in ENCODE. The score serves as the preliminary measurement of their functional cooperation of the SNP pair.

**SA 1.2 Infer the causal and cooperative mechanisms for disease associated SNPs**

We will identify the causal and cooperative SNP pairs for complex diseases as those with statistically significant small distance in the common factor space. Although we can sort SNP pairs by their distances to prioritize those independent pairs with small distance, these distances are not normalized and confounded by linkage disequilibrium, thus it is difficult to identify cooperative SNPs by a simple and uniform distance cutoff. For instance, unrelated SNPs without any functions are always clustered together for whatever transformation. Therefore, we will evaluate the statistical significance of SNP pair distances calculated in SA 1.1, and only prioritize those statistically significant pairs as the tentative mechanism pairs for their associated complex diseases.

We will conduct *s* (1000 ≤ *s* ≤ 100,000) resamplings of SNPs to infer the statistical significance of a pair distance due to the large number of pairs from disease associated SNPs (>150k). The approach is similar to that of K-nearest neighbor clustering, but uses the statistical significance of SNP pair distances to automatically determine the best K for each SNP and balances the two best K values for a pair of SNPs. Also, this empirical approach will likely filter out non-functional SNPs with small distances observable purely by chance and control the effect of linkage disequilibrium to the distance of SNP pairs. It consists of four steps as follows.

1. For each pair of SNPs (SNP *a* and *b*), we fix one SNP (e.g. SNP *a*) and randomly resample the other SNPs from the set of studied SNPs *s* times(to be determined by resource availability and significance requirement). Additional permutations are required if we are unable to distinguish whether an adjusted p-value is really significant or caused by insufficient resampling. We restrict the random paired SNPs of the fixed SNP (a) from those within the same LD value range of the SNP pair for the fixed SNP (a), tentatively 10 LD ranges equally divided from 0 to 1 (this can be changed to balance the power and computational efficiency of the method). We will calculate the distance between the resampled SNP and the fixed SNP in the common factor space for each resampling SNP and the designate the proportion of resampled SNPs yielded smaller or equal distance as the nominal p-value of the SNP pair by the fixed SNP.
2. For the same SNP pair (SNP *a* and *b*), we fix the other SNP (*b*) and conduct similar procedures for the other SNP. We average the two nominal p-values arising from the two SNPs and obtain the nominal p-value for the SNP pair.
3. We perform the above procedures similarly for all SNP pairs and correct multiple comparisons of all pairs by false discovery rate (FDR).
4. Prioritize SNP pairs with FDR<0.05 as tentative cooperative SNPs if both SNPs in the pairs are associated with the same disease or disease class.

We have extensive experience conducting such empirical statistical studies, using high performance computing (HPC). We have 5% of the access (equivalent to ~1000 cores) in the Beagle system hosted by the University of Chicago and the Argonne National Laboratory. We have conducted 100,000 permutations on eQTL association networks between SNPs and mRNAs to quantify SNP-SNP downstream functional similarity and detected hundreds of cooperative SNP pairs with sufficient statistical significance (FDR<0.05), a result only approachable by our proposed method.

*Internal evaluation*: If the measure of SNP similarity/distance works, significant similar SNPs should be enriched in the same diseases as compared to distinct diseases, and in the same disease class as compared to distinct disease classes. The similarity of LD SNPs may be confounded by the SNP proximity, but the small distances are still expected to enrich in LD SNPs as compared to independent SNPs due to related biology in LD SNPs [15].

On completion of this sub aim, we will prioritize a set of cooperative SNPs pairs for each cell type, as tentative epistatic SNPs for their associated complex diseases. Of note, distinct cell types may yield different cooperative mechanisms for complex diseases, as many diseases only arise from specific types of cells due to their underlying pathophysiology.

**SA 1.3 Validate the epistasis of prioritized cooperative SNPs for the top five complex diseases in the EMERGE dataset**

Finally, we seek to validate the epistasis of the cooperative SNPs prioritized from the ENCODE data. The Electronic Medical Records and Genomics (EMERGE) project [25] endeavors to streamline GWAS studies by dynamically selecting the patients and controls, and thus is ideal for validation purpose. In addition to traditional electronic medical record (EMR) system, it collects patients’ genotypes after a blood test and links the genotype data with the EMR data. Thus, this dataset allows patient level validation of any single or combinatory genetic markers for a large number of diseases. We have collaborated with key investigators of the EMERGE project and plan to extend the collaboration to this project.

*We hypothesize that a substantial proportion of cooperative SNPs are epistatic for their commonly associated complex diseases*. Our preliminary results on three diseases using eQTL associations support this hypothesis (Preliminary Results). We hope to discover more epistatic SNPs from the ENCODE dataset using MFA based approach.

We will prioritize the top five complex diseases with at least one pair of significant cooperative SNPs by our MFA approach (SA 1.1 and 1.2) based on their statistical significance. We will then utilize the expertise of the EMERGE team to confirm the selection of validation diseases. As done before, the medical experts in the EMERGE network will determine the inclusive and exclusive criteria of patient selection for a validation disease. They will also match controls from dbGAP for these diseases. We will employ PLINK software for assessing the epistatic effect size of tentative SNP pairs, using classic quality control processes for samples and SNPs and adjusting the effects from the covariates such as age and gender. We will correct multiple comparisons of SNP pairs by false discovery rate.

For validated epistatic SNPs, we will further determine the driving mechanisms that lead to the cooperative mechanisms. For instance, we can search for the shared mechanisms between the pair of SNPs such as binding to the same transcription factor, interacting factors, and locating in the two anchor regions of a long-range chromatin interaction. We have extensive experience searching for these common underlying mechanisms from the ENCODE datasets from our previous studies on eQTL associations.

Upon completion of this sub aim, we are likely to obtain several causal SNP pairs with epistatic interactions for some diseases. These epistatic effects can be either synergistic or antagonistic. Synergistic SNP pairs can serve as more accurate biomarkers for diagnosis of underlying complex diseases than their corresponding individual biomarkers.

*Expected outcomes*: We expect 100 to 1000 SNP pairs with causal cooperative mechanisms, because our eQTL study on lymphoblastoid cells LCL cells unveiled more than 100 SNP pairs with similar downstream genes by using 500 diseases and around 2000 SNPs. With more SNPs, diseases, and cell types, it is reasonable to expect more positive results. In particular, more cell types should yield more disease mechanisms specific to these cell types than only a single cell type (LCL). We also expect half of the prioritized SNP pairs can be validated by the EMERGE dataset, based on our exploratory eQTL studies. While the large-scale validation of all results is out of the scope of this study, it provides a large number of hypotheses of epistatic disease mechanisms for the research community.

*Potential problems & alternative strategies*: The success of this aim may rely on the multiple factor analysis algorithm, which has been frequently used in other fields but not in genetics and translational bioinformatics. The current algorithm is based on principle component analysis (PCA). If this PCA based MFA method does not work well, we will try other factor analysis methods such as coordinating the rotations of the factors identified from each scale to balance the influence from each scale. The current algorithm also assumes the balanced influence from different groups of assays, which is reasonable without prior knowledge about SNPs but may be simplistic. If necessary, we will develop new multiple factor algorithms under the supervision of our co-investigators XXX by automatically training the weights across groups. Alternatively, we can also try multiple principal component analysis [26] and multiply non-negative matrix factorization [27] methods to identify cooperative SNPs. These methods have been used in data integration and clustering of multiple scale biological data.

We risk having insufficient power to detect significant cooperative SNPs from the greatly reduced, but still large, number of combinations. There are about 150,000 trait-associated SNPs and LD SNPs, which lead to ~11 billion combinations. While we have derived significant results from 2 million combinations from 100k permutations, we may lack power for this scale of pairwise combinations. If this happens, we will restrict our search to only SNP pairs commonly associated with the same diseases, or disease classes, which will significantly reduce the search space. If our cell type based method does not work due to excessive heterogeneity among a cell type, we will conduct the statistical control for each cell line separately and then combine the significance for the same cell type using Fisher’s method.

Finally, if the resampling strategy does not yield SNP pairs with expected enrichment, we will assess the statistical significance by permuting assays. We will also try an overall distance or similarity score from all cell types for a SNP pair if cell type specific approach does not work satisfactorily.

**Aim 2. Integrate ENCODE data to determine the genetic basis of comorbid complex diseases**

|  |
| --- |
| Macintosh HD:Users:haiquan:Dropbox:study:IMB521:write up:figures:host_sim_comorbid.pdf |
| **Figure 3: Enrichment of disease comorbidity in HCUP for disease pairs sharing significant disease gene similarity.** We used Fisher’s Exact Test (FET) coupled with false discovery rate to measure the disease comorbidity in HCUP, and took the disease similarity of gene ontology annotations from a recently published study [30]. |

*Hypothesis and rationale*: Disease comorbidity influences the treatment of patients with the same major disease but with distinct comorbidity conditions. It also burdens the patients and medical providers by increasing the risk of mortality, hospitalization, readmission, and emergence visits [28]. Understanding the causes and taking necessary preventive intervention will greatly improve the health care. Yet, the factors that determine the disease comorbidity are largely unknown. Potential factors include genetic factors, environmental exposures, and their interactions. With the opportunities brought from high throughput omics and data driven projects, **our long-standing objective is to identify the genetic basis of disease comorbidity**. *We hypothesize that a disease comorbid in the same patient should have closely related biological mechanisms, thus perturbation from one disease will inevitably disturb the physiology of the other due to the linked genetic basis.* These common or related mechanisms include common associations with two diseases from the same SNP, same gene, and the same pathways [14, 29]. In frequent but more complicated cases, distinct SNPs of the two diseases respectively may involve common regulatory mechanisms, such as SNPs binding to the same TFs, binding to interacting TFs, or locating in anchor regions of a long-range chromatin interaction. For instance, many autoimmune diseases are related to antigen processing and presentation of peptide or polysaccharide antigen via MHC class II, and their susceptible alleles are proximate or shared in regions of chromosome 6, subject to related regulation process such as chromatin folding. These types of related mechanisms have not been studied yet. We will quantify this novel type of shared mechanism and overall ENCODE similarity between diseases and study their predictive capability to clinically observed disease comorbidity through an integrative analysis of ENCODE data. The rationale of this aim is that successful completion will greatly enhance our understanding of the genetic and epigenetic mechanisms of disease comorbidity at the level of molecular mechanisms at play, and enable interpretation and prediction of disease comorbidity from a genetic perspective. At the completion of this aim, it is our expectation that we will generate hundreds of testable biomarkers for comorbid diseases. This will demonstrate, for the first time, the advantage of large-scale genetic discovery for disease comorbidity mechanisms, More importantly, we will generate a substantial number of genetically validated biomarkers, which allows further biological validation.

*Preliminary results:* Recently, we and others have observed enrichment of shared SNPs and genes in comorbid diseases [30]. Via integrative analysis of GWAS and gene ontology annotations, we found that known comorbid diseases are more likely to be associated with coding genes with similar biological functions and involved in common biological processes [30]. We further demonstrated significant enrichment of similar functions and biological processes between disease genes of comorbid diseases in south California Healthcare Cost and Utilization Project (HCUP) Electronic Medical Record (EMR) dataset (Fig. 3). Since disease genes only correspond to a small proportion of coding SNPs (<5%), we hypothesized that comorbid diseases may have common regulatory underpinnings. Our preliminary results unveiled significant enrichment of shared transcription factors binding to the susceptible SNP regions of comorbid diseases via the HCUP and ENCODE datasets (Table 2; OR=1.2, p=2.2x10-9; FET).

|  |  |  |
| --- | --- | --- |
| Disease pairs | with SNPs binding to a same TF | without SNPs binding to a same TF |
| Comorbid pairs | 6867 | 2670 |
| Non-comorbid pairs | 5051 | 2395 |

**Table 2. Enrichment of same transcription factors (TF) binding to SNPs associated with comorbid disease pairs unveiled by ENCODE Chip-seq and electronic medical record datasets**. We mapped 185 diseases to ICD-9-CM codes and checked every pair of diseases whether they are associated with some SNPs binding to the same TF, and whether they are significantly comorbid in south California HCUP dataset in 2011 (Comorbid odds ratio>1.5 and FDR <0.05; FET). This study unveiled significant enrichment between common TF binding to disease SNP regions and disease comorbidity (OR=1.2, p=2.2x10-9; Fisher’s exact Test).

*Approach:*

This aim consists of three parts: the first two are independent, while the third one is dependent on the first two (Fig. 4).

|  |  |
| --- | --- |
| Macintosh HD:Users:haiquan:Dropbox:study:IMB521:write up:figures:aim2_flow.pdf | **Figure 4: Processing flow of Aim 2**. SA 2.1 and SA 2.2 use distinct strategies in quantifying disease similarity but have similar strategy for empirical statistical analysis and internal validation by HCUP and Banner-University EMR datasets. They also share similar genetic validation strategy using the patient level EMRGE dataset. |

**SA 2.1 Quantify disease similarity by their common regulatory elements in ENCODE**

This sub aim will quantify common regulator elements of every pair of complex diseases. *We hypothesize that comorbid diseases are more likely to have common regulatory elements among their associated SNPs*. Specifically we will investigate three shared biological mechanisms between a pair of SNPs associated with the two diseases respectively: binding to the same TF, binding to interacting TFs, and locating in the anchor regions of a long range chromatin interaction (looping). If both SNPs have a functional effect, they are likely to perturb the binding affinity of the same TF or interacting TFs, or the affinity of chromatin looping. Due to the shared foundation of the SNP pairs, the perturbation of one disease from one SNP implies the perturbation of the other disease of the other SNP in the pair if an individual carries both risk alleles of the SNP pair. Our preliminary results demonstrate the feasibility of this approach. In this sub aim, we will quantify the common regulatory mechanisms for each disease pair associated with multiple SNPs, and estimate the effects of these regulatory mechanisms to the disease comorbidity.

To study the relationship between disease biological similarity and disease comorbidity from EMR, we need establish a mapping between literature disease terms and clinically standardized terms, i.e. International Classification of Disease (ICD; ICD9-CM in practice in US). We will investigate ~1200 complex diseases reported in the latest NHGRI GWAS catalogue and merge the same or very similar diseases. Similarly, we will merge multiple billing codes of the same diseases, such as various ICD-9-CM codes for diabetes type 1. In this regard, we will use the Clinical Classification Software (CCS) system [31] developed by the HCUP project for its wide use and successful application [32]. Then, we will map our merged disease terms into the categorized ICD-9-CM codes using a nature language system developed by our group, coupled with manual curation by clinicians in our team. We will then download the latest data of transcription factor binding and chromatin folding from the ENCODE data repository [3], and download protein-protein interaction (PPI) data from STRING–DB [33]. We have extensive experience in analyzing the ENCODE and PPI data in our previous and ongoing studies [30], as partially demonstrated in the preliminary results.

|  |  |
| --- | --- |
| Macintosh HD:Users:haiquan:Dropbox:study:IMB521:write up:figures:comorbidity_examples.pdf | |
| **Figure 5: An example of disease distance based on SNP distance to a disease.** The distance of a SNP to another disease is defined by the shortest distance between the SNP to any SNP in the other disease. The longer arrow corresponds to the largest distance from a SNP of disease 1 (in blue) to disease 2 (in red), while the shorter arrow corresponds to the shortest distance from a SNP (in blue) of disease 1 to disease 2 (in red). Both lines greatly drive the overall distance between the two diseases. |

After the data are available, we will measure the commonality of each of the three common regulatory mechanisms, as well as the combined commonality of the three mechanisms for each pair of diseases. In brief, for each disease, we will first collect all reported SNPs in the NHGRI GWAS catalog and identify the LD SNPs in studied populations, such as Caucasians and African Americans. We will then check how many pairs of SNPs across the two diseases have the same regulatory mechanisms across all available cell types. Of note, we aggregate TF binding and chromatin looping signals because of the sparsity of regulatory functions in a specific cell line or a single cell type.

We will then assess the statistical significance of the observed common regulatory mechanisms for each pair of diseases because more SNPs of a disease pair increase the chance of observing such relationship thus inducing a bias. We create a bipartite graph of GWAS associations between all studied diseases and their associated SNPs (including LD SNPs). We permute the graph for 100,000 to 1 million times to create a null distribution. In each permutation, the number of diseases with which each SNP is associated retains the same as that in the observed association network, similarly for the number of SNP with which each disease is associated. We shuffle the edges in the GWAS association network to fulfill this constraint and control abovementioned biases (Fig 5B). We have extensive experience with this type of network permutation [30, 34].

From the permutations, we can compute the random number of SNP pairs with such regulatory mechanisms for each pair of diseases, and check the proportion of permutations which yielded more common regulatory mechanisms than observed, which serves as the p-value for the corresponding regulatory mechanism of the disease pair. Finally, we will correct multiple comparisons by false discovery rate for all tested disease pairs, and prioritize those with FDR<0.05 as tentative comorbid disease pairs.

*Internal evaluation*: To evaluate the accuracy of this approach, we will compare the prioritized disease pairs with the disease comorbidity in the HCUP dataset and the Banner-University data warehouse. First, we will assess the comorbidity of any diseases that we computationally modeled in the ENCODE by each EMR dataset. The HCUP dataset we have access consists of more than 7 million patients and is sufficient for estimating the comorbidity effect. We will use Fisher’s Exact Test to test the comorbidity of diseases in each EMR dataset with the control of the potential confounders, such as sex and age. We will also use Fisher’s Exact Test to assess the enrichment of significantly shared biological mechanisms in clinical observed comorbidity. A significant enrichment with enough effect size (OR>2) between the genetic and clinical datasets suggests a potential role of the regulatory mechanisms in disease comorbidity as a whole.

**SA 2.2 Use overall ENCODE similarity of disease pairs to model disease comorbidity**

This sub aim will quantify the overall ENCODE similarity of any pair of complex diseases. *We hypothesize that a pair of comorbid diseases are underlain by two functionally related sets of SNPs associated with the two diseases respectively.* Thus, the perturbation of the physiology of one disease by genetic variations may perturb the other. Therefore, we can employ similar strategy in Aim 1 to first quantify functional similarity of SNP pairs, and then integrate the functional similarity (equivalently functional distance) for pairs of diseases.

For a pair of investigated diseases, denoted as A and B, we will quantify their ENCODE similarity with respect to a cell type as follows:

1. Collect all SNPs that associate with the two diseases from the NHGRI GWAS catalog, along with LD SNPs of these SNPs if their linkage disequilibrium r2>0.8 based on the 1000 genome project.
2. Reuse the distances of SNP pairs across the two diseases calculated from a variety of assays in the ENCODE dataset by multiple factor analysis (**Sub Aim 1.1**)
3. For each SNP (s1) in a disease (i.e. A), identify the shortest distance to the other disease in the pair (i.e. B). The distance of a SNP (s1 of A) to another disease (B) is defined as the distance between the SNP (s1) and a SNP (s2) in the other disease (B) that has the shortest distance with the former SNP (s1) among all SNPs associated with the other disease (disease B) (See two examples of shortest distances in Fig. 5A). Do that reciprocally for every SNP associated with the latter disease (B).

Average the shortest distance obtained from Step 3) for every SNP in both diseases, and use the average shortest distance between SNPs of the two diseases as the overall distance for the two diseases. The distance of two diseases is mathematically denoted as follows (Eq 2):



where dist (si, s­j) was calculated in SA1.1, and SNPs(A) is the set of SNPs associated with disease A (similarily for disease B) and || is the cardinality (number of elements) of a set.

The above similarity measurement problem can be regarded as a part of special hierarchichal clustering problem. In this clustering problem, each disease is a conceptual cluster consisting of the associated SNPs. These SNPs are inseparable, automatically comprising the first level of clustering, but they can belong to multiple disease clusters, which make this problem different from classic clustering. The above algorithm corresponds to the second level of clustering, evaluating the distance of each disease pair and clustering those with small distance. During this course, the shortest distance of a SNP to another disease corresponds to the distance to the closest boundary of the disease cluster consisting of its associated SNPs (shorter arrow in Fig. 5A). As a result, only disease clusters locating closely enough will get small distance and will be sorted out. A disease cluster must be condensed and avoid dispersion, i.e. with small diameter, in order to cluster with another disease cluster. This is because the distant end points (longer arrow in Fig. 5A) of a disease cluster relative to another disease cluster will yield a large shortest distance, thereby penalizing the overall distance between the two diseases.

For a given cell type, we will generate the ENCODE distance for every pair of complex diseases using the above method. Since diseases are associated with various numbers of SNPs, simply ranking the distance across all disease pairs will favor disease pairs with small number of SNPs, thus introducing biases. In this regard, we seek an empirical statistical solution exactly the same as that of SA 2.1. We will conduct 1000 to 1 million permutations to the GWAS association network. For each permutation, we calculate the random disease ENCODE similarity for each disease pair and then we count the proportion of permutations yielding smaller distance than that yielded from observed network. We use that as the p-value of the disease pair on this cell type, and correct multiple comparisons for all tested disease pairs using false discovery rate (FDR). Finally, we prioritized the disease pairs with a certain FDR cutoff, e.g. FDR < 0.05. We have extensive experience conducting such permutations and prioritization [30, 34].

*Internal evaluation*: As did in SA 2.1, we will use our internal electronic medical record datasets HCUP and Banner-university warehouse to validate the accuracy of this approach. We will measure disease comorbidity similarly as in SA 2.1, i.e., assessing the enrichment of co-occurrence of two diseases on patients while controlling the effects of gender and sex. We will then test the enrichment of the ENCODE derived disease distance/similarity and clinically observed comorbidity in the two EMR datasets. We will assert the effectiveness of this approach if significant enrichment with a clinically meaningful effect size (odds ratio >2) is obtained across the genetic and clinical datasets. Furthermore, we will also compare the effect size of this comprehensive approach to the specific mechanism based on the approach described in SA 2.1.

**SA 2.3 Validate top ten prioritized disease comorbidities from ENCODE in EMERGE dataset**

We will validate five prioritized disease pairs from SA 2.1 and SA 2.2, respectively, based on the statistical significance and effect size of these disease pairs (FDR<0.05) as well as the availability of expertise in the EMERGE network. *We hypothesize that disease pairs with strong common mechanisms (SA 2.1) and strong functional coherence (SA 2.2) in ENCODE are more likely to have comorbidity in patients due to the genetic dependence between these diseases.* We will use classical genetic approaches to conduct the genetic validation, and tailor strategies particularly for this comorbidity problem.

For each pair of tentative comorbid diseases, we will determine the rules of selecting patients from the EMERGE data warehouse. We will select patients with only one disease, with both diseases (comorbidity patients), and with none of the diseases based on billing codes. At least two separate billing codes for the same diseases are required to assert the diagnosis of a disease. Patients with uncertain diseases by ambiguous codes will be filtered from both cases and controls. We have already established an early collaboration with the EMERGE investigators.

We will use a directional model to test the effect size of disease comorbidity with respect to shared functional SNPs or a cooperative SNP pair. This directional model distinguishes the comorbidity from a disease A (prior distance) to another disease B (posterior disease) from the inverse direction, i.e. from the disease B (prior disease) to A (posterior disease). First, we will identify the shared SNPs or independent SNPs that mostly contribute the similarity/distance of the tentative comorbid diseases. Then, we will employ logistic regression to test the effect size of acquiring a pair of comorbid diseases after diagnosed with a disease. We will regard the shared SNPs or cooperative SNPs as dependent variables, and more importantly, regard the interactions between the tested SNPs (shared or cooperative pair) and the prior disease as a dependent variable as well, and regard the prior disease as a covariate. We will test effect size of the interaction term, under the control of age and sex as covariates. We also test the other direction as well and treat the two tests as separate tests as the disease progression for the two cases are different. We believe we can reach sufficient power for the comorbidities among common diseases if the effect size is moderate (e.g. OR=1.5), considering the large number of samples (>2000 for most common diseases) available in the EMERGE dataset (Table 3).

**Table 3: Power analysis by Quanto**. The table shows the number of cases (comorbidities) required to reach a power of 80% when 5 times of controls are matched, for expected effect size of 1.5 (OR) and case risk of 1%. Rows and columns show the minor allele frequency (MAF) of SNP1 and SNP2, respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | MAF of SNP2 | | |
|  |  | 0.1 | 0.2 | 0.3 |
| MAF of SNP1 | 0.1 | 666 | 376 | 288 |
| 0.2 | 376 | 217 | 170 |
| 0.3 | 288 | 170 | 136 |

Upon completion of this sub aim, we are likely to validate about half of disease pairs. For validated disease comorbidities, the contributing shared functional SNPs or cooperative SNP pairs are promising as tentative biomarkers that merit further biological validation.

*Expected outcomes*: We expect to investigate 300-500 diseases with both genetic relationships and clinical data available. From these diseases, we expect over 1000 disease pairs will be concordant between their genetic foundation and comorbidity in medical records, based on our prior experience and preliminary results. We also expect that the disease pairs with observed comorbidity in EMR are cell type specific. For instance, the data from lymphoblastoid cell lines usually yield relationship between autoimmune diseases. We will be particularly interested with unexpected comorbidities across organ systems since they may be caused by similar biological mechanisms that are difficult to discover. Finally, we expect about half of top disease pairs can be validated at the patient level in the EMERGE dataset based on our experience in studying cooperative mechanisms in eQTL data.

*Potential problems & alternative strategies*: Although the approach is very challenging, we do not expect major barriers to the ultimate goal of developing biomarkers for predicting disease comorbidity. It is expected, however, that the simple hypothesis in SA 2.1 may yield moderate enrichment as only a few mechanisms are considered, although the roadmap to disease comorbidity is much clearer than that in SA 2.2. In contrast, we expect that the comprehensive model in SA 2.2 will lead to a higher enrichment. Although unlikely, if only moderate enrichment is obtained in SA 2.2, we will try alternative approaches such as various similarity measurements used in clustering algorithms, or median distance rather than average distance. We also foresee a substantial effect size of the tentative disease pairs in the EMERGE dataset. However, it is possible, although unlikely for common SNPs, that low prevalence of a comorbidity may make the validation under powered. Table 3 estimates the possibility. Moreover, we will estimate the power of candidate disease pairs prior to the validation during the selection of our candidate pairs. We will use propensity scores [35] to increase the power and reduce confounding if it is indeed under-powered. In addition, we will exploit different strategies in handling cell lines when calculating the ENCODE similarity/distance, such as combining all cell lines, and first treating each cell line separately and then searching for the consensus across cell lines. The last concern may be the potential risk caused by the dependence on Aim 1.1 of SA 2.2. Our preliminary results in Aim 1 demonstrate the feasibility of using multiple factor analysis, thus the calculation of SNP distance by this approach is feasible for both aims. Moreover, the success of Aim 2 is by no means dependent on the success of Aim 1. so we can validate our results only from SA 2.1. Even if SA 2.1 only yields moderate enrichment results, we can still pick up the most confident disease pairs for patient level validation. In short, the risk of the partial dependency is minimal.

**References**

1. Lin, P. and A. Shuldiner, *Rethinking the genetic basis for comorbidity of schizophrenia and type 2 diabetes.* Schizophrenia research, 2010. **123**(2): p. 234-243.

2. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic acids research, 2014. **42**(D1): p. D1001-D1006.

3. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

4. Lonsdale, J., et al., *The genotype-tissue expression (GTEx) project.* Nature genetics, 2013. **45**(6): p. 580-585.

5. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-1195.

6. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proceedings of the National Academy of Sciences, 2012. **109**(4): p. 1193-1198.

7. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

8. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

9. Wu, T.T., et al., *Genome-wide association analysis by lasso penalized logistic regression.* Bioinformatics, 2009. **25**(6): p. 714-721.

10. McKinney, B.A., et al., *Machine learning for detecting gene-gene interactions.* Applied bioinformatics, 2006. **5**(2): p. 77-88.

11. Hahn, L.W., M.D. Ritchie, and J.H. Moore, *Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions.* Bioinformatics, 2003. **19**(3): p. 376-382.

12. Karczewski, K.J., et al., *Systematic functional regulatory assessment of disease-associated variants.* Proceedings of the National Academy of Sciences, 2013. **110**(23): p. 9607-9612.

13. Li, L., et al., *Identification of type 2 diabetes subgroups through topological analysis of patient similarity.* Science translational medicine, 2015. **7**(311): p. 311ra174-311ra174.

14. Li, L., et al., *Disease risk factors identified through shared genetic architecture and electronic medical records.* Science translational medicine, 2014. **6**(234): p. 234ra57-234ra57.

15. Corradin, O., et al., *Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits.* Genome research, 2014. **24**(1): p. 1-13.

16. Weinstein, J.N., et al., *The cancer genome atlas pan-cancer analysis project.* Nature genetics, 2013. **45**(10): p. 1113-1120.

17. Thurstone, L.L., *Multiple factor analysis.* Psychological Review, 1931. **38**(5): p. 406.

18. Escofier, B. and J. Pages, *Multiple factor analysis (AFMULT package).* Computational statistics & data analysis, 1994. **18**(1): p. 121-140.

19. De Tayrac, M., et al., *Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach.* BMC genomics, 2009. **10**(1): p. 32.

20. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome research, 2012. **22**(9): p. 1790-1797.

21. Van Deun, K., et al., *A structured overview of simultaneous component based data integration.* BMC bioinformatics, 2009. **10**(1): p. 246.

22. Lê, S., J. Josse, and F. Husson, *FactoMineR: an R package for multivariate analysis.* Journal of statistical software, 2008. **25**(1): p. 1-18.

23. Chavent, M., et al., *Package ‘PCAmixdata’.* 2014.

24. Thioulouse, J., et al., *ADE-4: a multivariate analysis and graphical display software.* Statistics and computing, 1997. **7**(1): p. 75-83.

25. Gottesman, O., et al., *The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future.* Genetics in Medicine, 2013. **15**(10): p. 761-771.

26. Jolliffe, I., *Principal component analysis*. 2002: Wiley Online Library.

27. Zhang, S., et al., *Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.* Nucleic acids research, 2012: p. gks725.

28. Lee, Y.J., et al. *COPD Hospitalization Risk Increased with Distinct Patterns of Multiple Systems Comorbidities Unveiled by Network Modeling*. in *AMIA Annual Symposium Proceedings*. 2014. American Medical Informatics Association.

29. Melamed, R.D., et al., *Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes.* Nature communications, 2015. **6**.

30. Li, H., et al., *Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory.* Journal of the American Medical Informatics Association, 2012. **19**(2): p. 295-305.

31. Elixhauser, A., C. Steiner, and L. Palmer, *Clinical classifications software (CCS).* Book Clinical Classifications Software (CCS)(Editor ed^ eds), 2008.

32. NIS, H.N.I.S., *Healthcare cost and utilization project (HCUP).* 2006.

33. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic acids research, 2011. **39**(suppl 1): p. D561-D568.

34. Lee, Y., et al., *Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases.* Journal of the American Medical Informatics Association, 2013. **20**(4): p. 619-629.

35. Widom, C.S., K. DuMont, and S.J. Czaja, *A prospective investigation of major depressive disorder and comorbidity in abused and neglected children grown up.* Archives of general psychiatry, 2007. **64**(1): p. 49-56.