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 with a very small effect size. It remains elusive how these SNPs interplay and 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 SNPswhose cumulative effect is non-additive from the individual ones. These approaches generate only a few epistatic interactions between SNPs across all complex diseases because they search from up to 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 ones and observed many epistases from them. Identifying these cooperative SNPs will greatly reduce the search space and allows better biological interpretability for the epistatic relationships 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. On this regard, 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 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 on multiple 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 600 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 showed 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 in DNA regions rather than specific causal SNPs. Here, we also 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 we can access. Again, using EMERGE dataset, we will assess the effect sizes of disease comorbidities for top ten 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.

Approach

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

The functional connections among multiple SNPs associated with the same complex disease are large unknown. This aim is to fill this gap by integrative study of ENCODE functional annotations of DNA regions of these SNPs.

Problem statement:
Preliminary results: The functional linkage among multiple independent SNPs associated for the same complex diseases is largely unknown. Our preliminary studies on eQTL data of 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 in their downstream genes perturbed by these SNPs. 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 many statistical significant cooperative SNPs have epistatic effects. Since eQTL associations cannot indicate causal relationship, we endeavor to identify causal SNPs and their cooperative biological mechanisms at play. We extend our study on SNPs to their high linkage disequilibrium regions (LD; correlation r^2>0.8) since SNPs in the LD regions are associated with the same diseases equivalently with the proxy SNPs tested in GWAS. Our preliminary results on these SNPs using multiple factor analysis (MFA) on chromatin accessibility and transcription factor binding of B-lymphocyte cell line GM12878 indicate cooperative SNPs with LD tend to cluster together for the same disease, or same disease class (Figure 1). The preliminary work also demonstrated the feasibility of the MFA algorithm on genome-scale calculation (running time < 2 minutes) and unveiling of SNPs with similar binding to 5 transcriptional factors for the same disease.

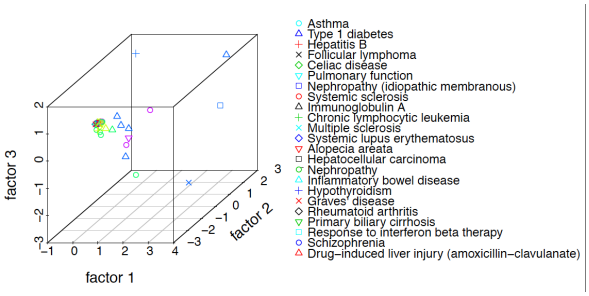

Figure 1. **Example MFA results**. First three factors of disease-associated SNPs on chromosome 6 from five replicates of chromatin accessibility data and five TFs (16 replicates in total). Although many SNPs are clustered together due to sparse function on these five TFs, SNPs (in blue triangles) associated with inflammatory bowl diseases are close with each in the first three factors, indicating similar functions among five TFs.

Hypothesis and rationale: The interplay mechanisms of multiple independent SNPs for complex diseases are largely unknown. Classic genetics methods are unable to detect epistasis among candidate SNPs due to vast number of combinations (up to trillion pairs). eQTL data allows unveiling the functional linkage among SNPs for complex disease but are unable to distinguish confounders and causal ones. ENCODE project generated abundant functional annotations for DNA on multiple scales and are thus promising to unveil the causal SNPs and their interactions. **Our objective in this aim 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 that is available in R to identify the unknown driving factors and measure the similarity (or their equivalent distance) on these factors for any pair of candidate SNP pairs by using multiple scale assays on hundreds of cell lines in ENCODE. The rationale of this aim is that the successful completion will fill the gap of unknown systematic mechanisms of complex diseases, generate many testable biomarkers for diagnosis of complex diseases. Upon the completion of this aim, it is our expectation that we will generate hundreds of functionally cooperative SNPs for complex diseases, which would allow high throughput discovery and validation of the epistasis for the first time.

Approach

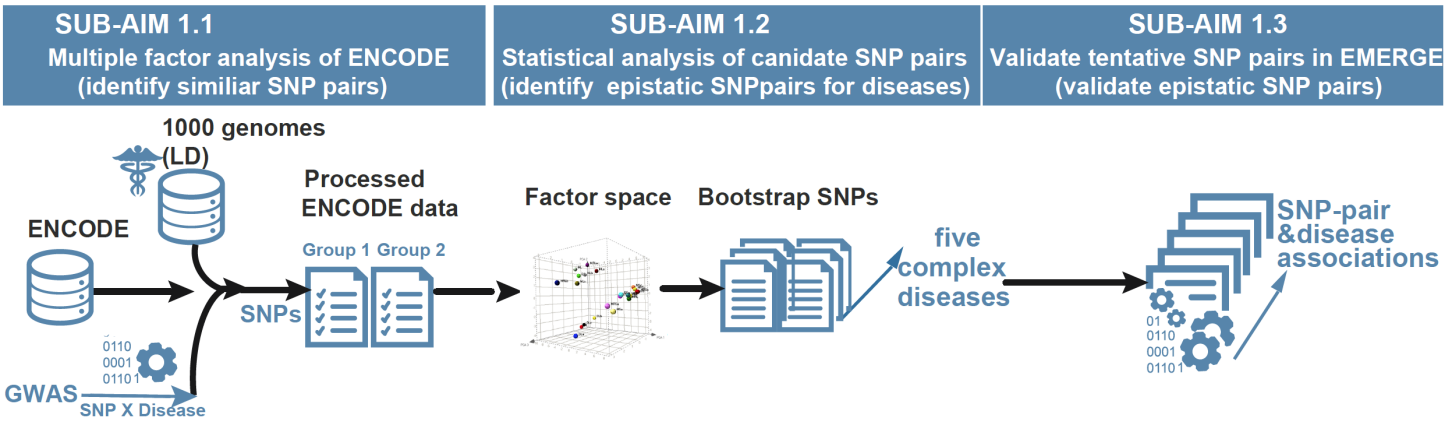

Fig. 2. Processing flow of Aim 1 for a cell type. It consists of three steps: 1) Project multi-scale ENCODE data of a cell line to a data matrix to the SNP level. 2) Map studied SNPs of the cell line to several major factors using multiple factor analysis. 3) Average the distance of a SNP pair on all cells of the same type and conduct bootstrap for statistical significance.

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 ENCODE functional repository (Figure 2). First, we will download disease associated SNPs from National Human Genome Research Institute (NHGRI) GWAS catalog. This catalog consists of the latest results across almost all the GWAS studies. It comprise of 1200+ diseases and 15,000 SNPs (lead SNPs) up to 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 (r^2>0.8) to above 15K SNPs. SNPs with strong linkage disequilibrium to lead SNPs are also likely to associate with the diseases, some of which are even more likely to play functional roles than the corresponding lead SNPs [Ref]. Third, we will collect major assays from ENCODE data repository, including but not limited to the DNASE I hypersensitive site assay, histone modification and transcription binding by Chip-seq, and RNA-seq. Although comprising multiple assays on over one hundred cell lines, functional assays in ENCODE were highly sparse, relying on the investigators' interests and resources. We tailor our approach specifically for this challenge. We will employ the 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 on DNA, many of which span a small region, to the SNPs these signals cover. We hypothesize that the function of the SNPs are very related to the signal around, since the SNP, specifically the minor allele, will perturb the signals of the regions in most cases if the SNP is functional. Finally, for a specific cell line, we generate a matrix for all studied lead and LD SNPs, consisting groups of available assays for these SNPs, with each assay in a column (Figure 2).

We will apply multiple factor analysis on above SNP matrices. We will combine the evidence from multiple cell lines of the same type but handle distinct separately, including Sub Aim 1.2. For a specific cell line, we will apply multiple factor analysis on multiple groups of assays for the function similarity of any pair of SNPs of our interest. Multiple factor analysis (MFA) is an extension of factor analysis that is a classic methodology in multiple variable analyses. While factor analysis is widely used in computational biology [ref], MFA has been infrequently used epistasis analysis and data integration. However, it is an unexplored tool perfect for the integratively epistatic analysis. Considering the scenario 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) and represent the activities of all assayed TFs. Multiple factor analysis extends this methodology and deals with distinct groups of assays and aims to avoid the dominance from a single group and balances the influence across all groups. For instance, histone modification is crucial to the function of DNAs and is also related to transcription factors. In the joint study of histone modifications and transcription

factor binding, multiple factor analysis identifies common factors that determine both histone modification and transcription factor activity assays. MFA assumes any single assay from each group is a linear combination of the common factors, and it can also represent a SNP as linear combinations of the common factors, mathematically equivalent to a point in the space of common factors.

Escofier and Pages developed multiple factor analysis in the 1980s [ref]. It uses principal component analysis on each group and first eigen value of each group to weight the influence of all variables (e.g. assays) in the group to avoid dominance of a single group [ref]. Several R packages provided implementation of the algorithm such as FactoMineR, PCAmixdata, and ade4. As demonstrated in our preliminary results, FactoMineR is scalable to genome scale and big data analysis. Moreover, it can integrate groups with various number of assays, including a single assay as done in chromatin accessibility and RNA-seq. Using FactoMineR, we will quantify the ENCODE function similarity of any SNP pairs described in preprocessing using distance that is inverse with similarity. The algorithm consists of the following five steps for each cell type.

1) Conduct MFA on each matrix consisting of multiple types of ENCODE assays for a cell line, using top important factors that account for at least 75% of the variances in multiple scale datasets.
2) Project each SNP corresponding a row of the matrix into a point of the five common factors identified across all groups. FactoMineR allows this mapping by generating the coordinator of the any row in the common factor space as a part of output.
3) Calculate the Euclidean distance between any pair of SNPs in the original matrix
4) If multiple cell lines of the same cell type are available in ENCODE, as were for lymphoblastoid cells, calculate step 1 to 3 separately for each cell line and average the distance yielded by all cell lines of the same type for each SNP pair, even if the types of assays vary across different cell lines.
5) For different types of cells, conduct step 1 to 4 separately and develop a distance for each type of cells.

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 in the pair, which is upon further statistical evaluation in SA 1.2 to avoid finding relationships observable by chance.

SA 1.2 Infer the causal and cooperative mechanisms of SNPs for complex diseases

Although we can sort the distance of SNP pairs to prioritize those independent pairs with small distance, these distance values are not normalized, and it is difficult to determine a meaningful cutoff to identify cooperative SNPs. As shown in Figure 1, the SNPs around original point with the coordinator (0,0) are likely to happen by chance. Thus, we should evaluate the statistical significance of the distance value of any SNP pair, and only prioritize those statistically significant pairs as the hypothesized mechanism pairs for their associated complex diseases.

We will conduct 1000 to 100,000 resamplings of SNPs, one of the typical empirical statistics, to infer the statistical significance of a pair. The approach is similar to that of K-nearest neighbor, but using the statistical significance to automatic determine the best K for each SNP and balance the best K for a pair of SNP. Also, it controls the effect of linkage disequilibrium to the distance and non-functional SNPs. It consists of four steps as follows.

1) For each pair of SNPs, we fix one SNP and random resampling the other SNPs from the set of studied SNPs by 1000 to 100,000 times, upon significance requirement and resource availability. We only sample the SNPs from those within the same LD value range, tentatively 10 ranges equally divided from LD value 0 to 1 but will be varied to balance the power and computational efficiency. The distance in the common factor space between the resampled SNP and the fixed SNP will be calculated and the proportion of resampled SNPs yielded smaller or equal distance is calculated as the nominal p-value with respect to the fixed SNP.
2) For above-mentioned pair of SNPs, we fix the other SNP and conduct similar procedures for the other SNP. We average the two nominal p-values and yield the nominal p-value for the pair of SNPs.
3) We perform the above procedures similarly for all SNP pairs and correct multiple comparisons by false discovery rate (FDR).
4) Keep SNP pairs with FDR<0.05 as the prioritized cooperative SNPs for the same complex diseases, or same disease class, if both SNPs in the pair being associated with the disease or class.

We have extensive experiences in conducting such empirical statistical studies, using high performance computing (HPC) both in and out campus. Of note, we have 5% of the access (equivalent to >1000 cores) in the Beagle system owned by the University of Chicago and 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 significances (FDR<0.05) that were exclusively approachable by this big data approach.

Internal evaluation: if the measure of SNP similarity/distance works, similar SNPs should be enriched in the same diseases as compared to distinct diseases, in the same disease class as compared to distinct disease classes, and in LD SNPs as compared to independent SNPs. The similarity of LD SNPs of the same lead SNP may arise in part from confounders of SNP proximity and part from real biology, as demonstrated in literature [ref].

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

SA 1.3 Validate top five complex diseases for their epistasis of cooperative SNPs in EMERGE dataset
Finally in this aim, we seek for validate the cooperative SNPs prioritized from ENCODE data. Electronic Medical Records and Genomics (EMERGE) project endeavors to streamline any GWAS studies by dynamically selecting the patients and controls, and thus ideal for the validation purpose. Coupled with traditional electronic medical record system, it also collects patients' genotypes after a blood test if obtained authorization. Thus, this dataset allows patient level validation of any single or combinatory genetic markers for various diseases. We have built collaborations with key investigators of EMERGE project and will extend that 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, rheumatoid arthritis, Alzheimer's diseases, and bladder cancer, have validated this hypothesis. We are in hope to validate more epistatic SNPs from the results of ENCODE dataset.

We will prioritize 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. The selection of validation diseases will comply with the expertise of EMERGE team because we may need manual curation of disease codes in a validation study. As done before, the medical experts in EMERGE network will determine the criteria of patient inclusion and exclusion for a validation disease. We will match controls from dbGAP as usual. We will employ PLINK software for assessing the epistatic effect size of tentative SNP pairs, upon classic quality control of samples and SNPs and adjusting the covariates such as age and gender. Multiple comparisons of SNP pairs will be corrected by false discovery rate, if necessary.

For validated epistatic SNPs, we will further determine the driving mechanisms that lead to the cooperative mechanisms. For instance, we can search for the mutual 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 experiences on searching for these common underlying mechanisms from ENCODE datasets by our previous studies on eQTL associations.

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

Expected outcomes: We expect 100 to 1000 SNP pairs with causal cooperative mechanisms, based on our preliminary study of eQTL associations from lymphoblastoid cells (LCL). Our eQTL study on LCL cells unveiled more than 100 SNP pairs with similar downstream genes by using 500 diseases and around 2000 SNPs, a smaller and older dataset than what we carry out here. With more SNPs, diseases, and cell types, it is reasonable to expect more positive outcomes. In particular, more cell types should yield more diseases mechanisms that are only specific to these cell types. We also expect half of the prioritized SNP pairs can be validated by EMERGE, based on our explorer on 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 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 (MFA). We will consider other factor analysis methods such as coordinating the rotations of the factors identified from each scale to balance the influence from each scale, if the MFA method does not work well. The current algorithm also assumes the balance influence from different groups, which assume a lot of independence of assays between groups thus may be over-simplified the relationships among the assays of different groups . If this balance-based approach does not work well, we will develop new multiple factor algorithms by setting up more sophistical strategies. Our statistical co-investigators XXX will supervise the development of the new algorithms.. Alternatively, we can try multiple principal component analysis [ref] and multiply non-negative matrix factorization [ref], which has been used in data integration and clustering of multiple scale biological data. Another risk is potentially insufficient power to detect significant cooperative SNPs from already 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 be under power for such huge number of pairwise combinations. If that happens, we will restrict our search to only pairs of SNPs and their LD SNPs 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 cells of the same type, we will conduct the statistical significance for each cell line separately and then combine the significance for the same cell type. Finally, if the bootstrap strategy does not work well, 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. Complex diseases are often comorbid with one another, but the underlying genetic causal mechanism of the disease comorbidity is largely unknown.

Preliminary results: Genetics may play an important role in disease comorbidity because clinical practice observed disparity on the occurrence of disease comorbidity across different populations several decades ago [ref]. Recently, others and we have observed enriched shared SNPs and genes in comorbid diseases [ref]. Via integrative analysis of GWAS and gene ontology annotations, we found that very known comorbid diseases are more likely to associated with genes with similar biological functions and involving in common biological processes [ref]. Through electronic medical record (EMR) datasets south California Healthcare Cost and Utilization Project (HCUP), we also observed significant enrichment of comorbidity in the disease pairs with significant disease gene functional similarity (Fig. 3). Since only a small proportion of SNPs locate in coding regions (<5%), we hypothesized that comorbid diseases may have common regulatory underpinnings. Our preliminary results from joint study of transcription factor (TF) binding sites in ENCODE and the HCUP dataset also yielded significant enrichment of shared transcription factors binding to the regions of SNPs associated with pairs of comorbid diseases respectively (Table 1; OR=1.2, p=2.2x10^{-9}).

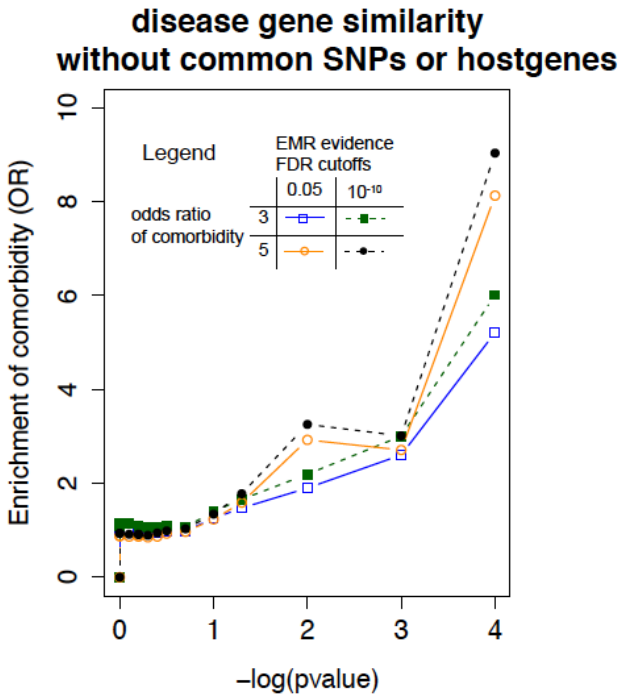

Figure 3: Enrichment of disease comorbidity in HCUP for disease pairs sharing significant disease gene similarity.

Table 1. Enrichment of disease associated SNPs binding to the same transcription factors (TF) for pairs of comorbid diseases unveiled by ENCODE Chip-seq and electronic medical record datasets. We mapped 185 diseases in ICD-9-CM codes and checked every pair of diseases whether they had some associated SNPs binding to the same TF, and whether they are significantly comorbid in south California HCUP dataset at 2011 (Comorbid odds ratio>1.5 and FDR <0.05). The straightforward study unveiled significant enrichment between common TF binding of disease SNPs and disease comorbidity (OR=1.2, p=2.2x10^{-9}; Fisher's exact Test).

	Disease pairs with SNPs binding to a same TF	Disease pairs without SNPs binding to a same TF
Disease comorbidity pairs	6867	2670
Non comorbidity pairs	5051	2395

Hypothesis and rationale: Disease comorbidity influences the treatment of patients of same major disease with distinct comorbidity conditions. It also burdens the patients and medical providers by increasing the risk of mortality, hospitalization, readmission, and emergence visits. Understanding the causes and taking necessary preventive intervention will greatly improve the health care. Yet, the genetic factors and their interactions with environmental exposures that determine the disease comorbidity are largely unknown. With the opportunities brought from high throughput omics and data driven projects, **our long-standing objective is to identify genetic basis of disease comorbidity, by which the perturbation of the physiology of a diseases tend to perturb some related diseases**. *We hypothesize that 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.* Instance of these common or related mechanisms include common associations with two diseases from the same SNP, same gene, and the same pathways. In more complicated but more likely 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. We will investigate these straightforward regulatory mechanisms shared between diseases and quantify the commonality, and more importantly, quantify the overall functional similarity via integrative analysis of ENCODE data. The rationale of this aim is that successful completion will greatly

enhance our understanding to the genetic and epigenetic mechanisms of disease comorbidity, and enable interpreting and predicting disease comorbidity from genetic perspective. At the completion of this aim, it is our expectation that we will generate hundreds of testable biomarkers for comorbid diseases and a substantial number of genetically validated biomarkers, which would allow for the first time, large scale of genetic discovery for disease comorbidity mechanisms.

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

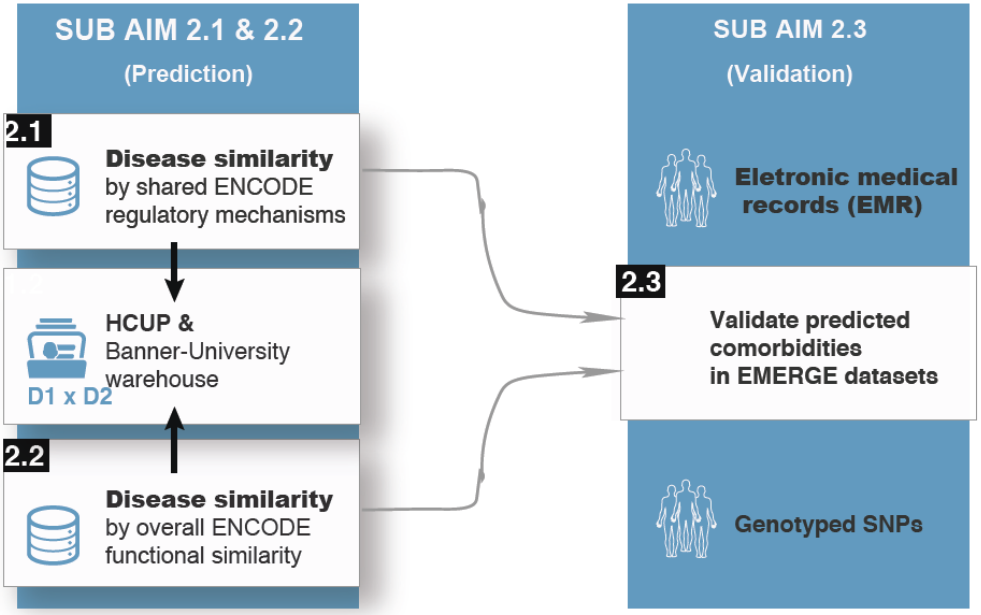

Figure 4: Processing flow of Aim 2. Both SA 2.1 and SA 2.2 use similar strategy for empirical statistical analysis and internal validation by HCUP and Banner-University EMR datasets after quantifying disease similarity. Both share similar strategy in patient level validation in EMRGE dataset.

SA 2.1 Quantify disease similarity by their common regulatory elements in ENCODE
This sub aim will quantify every pair of complex diseases by their common regulator elements. We hypothesize that comorbid diseases are more likely to have common regulatory elements among their associated SNPs, specifically we will investigate the binding to the same TF, binding to the interacting TF, or locating in the anchor regions of a long range chromatin interaction (looping) for a pair of SNPs associated with the two disease respectively. If both SNPs have active perturbed roles, they are likely to perturb the binding affinity of the same, interacting TF, or the affinity of chromatin looping. Then, the perturbation of one disease implies the perturbation of the other. Our preliminary results demonstrate the feasibility of this approach. In this sub aim, we will investigate more common regulatory mechanisms for disease comorbidity and estimate the effects of these regulatory mechanisms to each pair of diseases.

We will investigate about 1200 complex diseases in the latest NHGRI GWAS catalogue, and map them into ICD-9-CM code manually. To quantify the disease comorbidity in term of common regulatory mechanism and validate the predictions, we will merge the same or very similar diseases to the same ICD-9 code and merge multiple billing codes for the same diseases, such as various ICD-9-CM codes for diabetes type 1. We will download the latest version of transcription factor binding and chromatin folding data from ENCODE data repository [ref], and protein interaction data from STRING–DB [ref]. We have extensive experiences in analyzing such data in our previous and ongoing studies [ref], partially shown in the preliminary results.

After the data are available, we will measure the commonality of each of the three common regulatory mechanisms, and the combined commonality for 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, due to the sparsity of regulatory functions in a specific cell line or a single cell type.

Then, we will assess the statistical significance of the observed common regulatory mechanisms for each pair of diseases because more SNPs associated with a pair of diseases increase the chance of observation of such relationship thus may induce a bias. We will permute the associations between the SNPs and diseases for one

million times, during which we keep the number of associations for each SNP and each disease the same as observed, to control above-mentioned biases. We have extensive experiences on this type of permutations [ref]. From the permutations, we can compute the random number of such regulatory mechanisms for each pair of diseases and check the proportion of permutations 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 potential comorbid disease pairs.

Internal evaluation: To evaluate the accuracy of this approach, we will compare the prioritized disease pairs with the disease comorbidity in HCUP dataset and Banner-University data warehouse. We will assess the comorbidity of any diseases that we computationally assessed in ENCODE in these two datasets separately. HCUP consists of more than 7 million patients. We will use Fisher Exact Test to test the comorbidity of diseases in the two EMR datasets with the control of the potential confounders from gender and age. We will also use Fisher Exact Test to assess the coherence between the genetic and clinical datasets by testing the enrichment of statistically significant regulatory mechanisms in comorbid diseases. 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 complex diseases. We hypothesize that a pair of comorbid diseases in patients are underlying by two functional related sets of SNPs that are 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 (or equivalently distance) for pairs of diseases.

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

1) Collect all SNPs that associated with the two diseases from the NHGRI GWAS catalog, along with their LD SNPs by r^2>0.8 based on the 1000 genome project
2) Reuse the distances of SNP pairs across the two diseases calculated in **Sub Aim 1.1** by multiple factor analysis on a variety of assays in ENCODE
3) For each SNP (s_1) in a disease (without losing generality, let it be disease A), identify the shortest distance to the other disease, defined as the distance to a SNP (s_2) in the other disease (disease B) that has the shortest distance with the former SNP (s_1) among all SNPs associated with the other disease B (See two examples of shortest distances in Fig. 5). Do that reciprocally for every SNP associated with disease B with respect to disease A.
4) 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 above strategy is pretty alike the some distance strategies in second level of hierarchy clustering algorithms since SNPs associated with the same disease can are inseparable and comprise the first level of clustering. If we regard each disease as a conceptual cluster in a space, the shortest distance of each SNP to a disease corresponds to

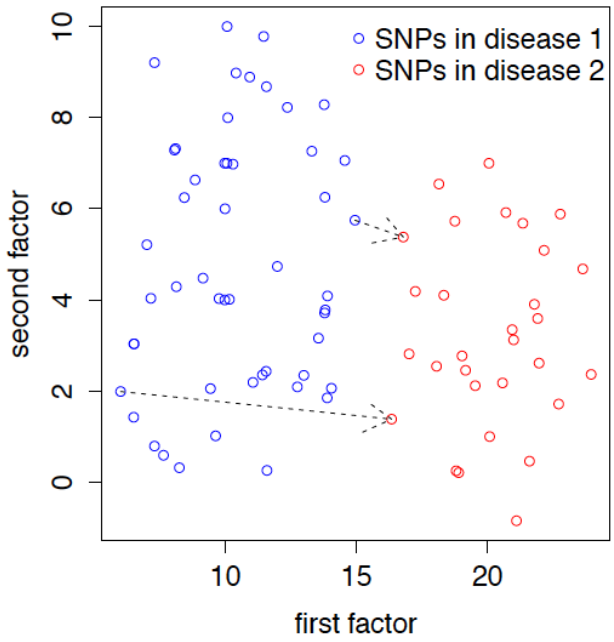

Figure 5: An example of disease distance based on SNP distance to a disease. The distance of a SNP to another disease is defined by the shortest distance between the SNP to any SNP in the other disease. The longer arrow corresponds to the largest distance from a SNP of disease 1 (in blue) to disease 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 line greatly drive the overall distance between the two diseases.

the distance to the closest boundary of the disease cluster consisting of its associated SNPs (shorter arrow in Fig. 5). As a result, only disease clusters locating closely enough will get small distance and sort out. Meanwhile, the disease cluster should be condense enough and avoiding too dispersed, i.e. with large diameter, since the distant end points (longer arrow in Fig. 5) of a disease relative to the middle of the two clusters will yield largest shortest distance to the other disease cluster, as a result, penalize the overall distance of 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 apparent biases. On this regard, we seek for an empirical statistical solution similar with that in 2.1. We create a bipartite graph of GWAS associations consisting of all studied SNPs (and their LD SNPs) and diseases. 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 network, similarly for the number of SNP with which each disease is associated. For each permutation, we calculate the random disease ENCODE similarity 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 the multiple comparisons for all test 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 experiences on such permutations [ref].

Internal evaluation: Similarly as did in SA 2.1, we use our internal electronic medical record datasets HCUP and Banner-university warehouse to validate the accuracy of this approach. We measure the comorbidity similarly as above, i.e., using enrichment of co-occurrence on patients between two diseases but control the effects of gender and sex. We then test the enrichment of ENCODE derived disease distance/similarity and clinically observed comorbidity in the two EMR datasets. We will confirm the effectiveness of this approach if significant enrichment is observed with enough effect size (odds ratio >2) across the genetic and clinical datasets. In addition, we will also compare the effect size of this comprehensive ENCODE approach to the specific mechanism based approach 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) and availability of the expertise in 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 patient level due to the genetic dependence between these diseases. We will use classic genetic approach to conduct the genetic validation, and tailor strategies particularly for the comorbidity problem.

For each pair of candidate 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 none of the diseases based on billing codes. At least two separate billing codes for the same diseases are required to assert a disease. Patients with uncertain diseases with ambiguous codes will be filtered from both cases and controls. We have established sufficient experiences in the early collaboration with EMERGE investigators.

We will use a directional model to test the effect size of disease comorbidity with respect a shared functional SNPs or a cooperative SNP pair. First, we will identify the shared SNPs or independent SNPs that mostly contribute the similarity/distance. Then, we will employ logistic regression to test the effect size of acquiring a comorbid disease after diagnosis with a disease. We will regard the shared SNPs or cooperative SNPs as the dependent variable, 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 them as separate test as the disease progression for the two cases are different.

Upon complete of this sub aim, we are likely to validate more than half of disease pairs. For validated disease comorbidity, the contributing shared functional SNPs or cooperative SNP pairs are promising for using as tentative biomarkers that merit further biological validation.

Expected outcomes: We expect to investigate 300-500 diseases with both genetic relationship and clinical data available. From these diseases, we hope to obtain over 1000 disease pairs that will be concordant between genetics and medical practice, based on our experiences 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 underlying by similar biological mechanisms. Finally, we expect at about half of top disease pairs can be validated in EMERGE at patient level based on our experiences in study cooperative mechanisms in eQTL data.

Potential problems & alternative strategies: Although the approach is very challenging, we do not expect major barriers that could fail the ultimate goal of developing tentative 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 incorporated. In contrast, we expect that comprehensive modeling will lead to a much higher enrichment. Although unlikely, if only moderate enrichment is obtained in SA 2.2, we will try alternative approaches such as various similarity measurement used in clustering algorithms. We also foresee a substantial effect size of the candidate disease pairs in the EMERGE dataset but it is possible, although unlikely for common SNPs, that low prevalence of a comorbidity may make the validation under power. We will consider this effect when selecting our candidate pairs by estimating the power of a candidate disease pairs before validation. We will also use propensity scores to increase power and reduce confounders if it does be under power. In addition, we will try different strategies in handling cell lines when calculating the ENCODE similarity/distance, such as combining all cell lines and treating each cell line separately and searching for the consensus across cell lines. The last concern may be the potential dependence on Aim 1.1. Our preliminary results in Aim 1 demonstrate the feasibility of this approach 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. Even that happen, in rare case, we still have alternative approach to validate our results from SA 2.1, which is completely independent with Aim 1. Even SA 2.1 yielding moderate enrichment results, we can still pick up the most confident disease pairs for patient level validation.