**Approach**

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

|  |
| --- |
|  |
| **Figure 1**. **Example MFA results**. First two factors of disease-associated SNPs on chromosome 6 from five replicates of chromatin accessibility data and five TFs (16 replicates in total). The clusters of SNPs are LD SNPs or independent cooperative SNPs. |

*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 r2>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 efficiency of the MFA algorithm by using within 2 minutes for a genome-wide analysis on a desktop computer.

*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 an in house built pipeline built on the PLINK package, we will extract SNPs with strong disequilibrium (r2>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. We found relatively few functional assays in ENCODE even though these assays were based on multiple cell lines and over 100 cell lines. 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. 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 proximate signal , since the SNP for the minor allele is 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 for these SNPs with each assay in a column (Figure 2).

We will apply multiple factor analysis on the above SNP matrices. We will combine the evidence from multiple cell lines of the same type but handle distinct cell lines 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 SNPs of our interest. Multiple factor analysis (MFA) is a well-established extension of factor analysis in multiple variable analyses. While factor analysis is widely used in computational biology [ref], MFA has been infrequently used for epistasis analysis and data integration. 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, instead, 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 five most important factors. Five factors usually contain the majority (>85%) of information in multiple scale datasets and thus being a default setting in the MFA package.
2. Project each SNP corresponding to 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 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 exist within a single cell type, as were for lymphoblastoid cells, calculate step 1 to 3 separately 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.

**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 cut them to identify cooperative SNPs. As shown in Figure 1, the SNPs around original point with the coordinate (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 statistically significant pairs as the hypothesized mechanism for their associated complex diseases.

We will conduct bootstrap on SNPs, one of the typical empirical statistics, to infer the statistical significance of a pair. It consists of four steps as follows.

1. For a given set of lead and LD SNPs (see SA 1.1), create 1000 to 100,000 bootstraps, upon resource availability and significance requirement. In brief, in each bootstrap, we pick up a random SNP from the SNP set for each SNP and put it back after each random resampling (sampling with replacement).
2. For each bootstrap, conduct MFA as in SA 1.1 on a random matrix that consists of the real assayed values of the each sampled SNP. We obtain a random distance for each pair of SNPs in each bootstrap. Of note, the value may be averaged values from multiple cell lines of the same type.
3. For each SNP pair, check how many random bootstraps out of all bootstraps yielding smaller distance than that of the real data; the proportion of such bootstraps becomes the statistical significance (p-value) of the SNP pair for the cell type.
4. Correct multiple comparisons for tested SNP pairs, using false discovery rate (FDR). 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 in SNP functional similarity study by eQTL associations and detected hundreds of cooperative pairs with sufficient statistical significances 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). 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 assumes the balance influence from different groups, which may be too simple. If necessary, we will seek for collaboration with our statistical co-investigators to develop more sophisticated algorithm. Alternatively, we can try multiple principal component analysis [ref] and multiply non-negative matrix factorization, 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 relevant SNP pairs associated with the same diseases, or disease classes, which will significantly reduce the search space. Finally, if the boostrap 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.