**Significance**

Genome-wide association studies (GWAS) have unveiled thousands of single nucleotide polymorphisms (SNPs) that have statistically significant associations with disease phenotypes[re]. Recent functional studies such as Encyclopedia of DNA Elements (ENCODE) and expression Quantitative Trait Loci (eQTL) indicated that these SNP regions are indeed functionally related to complex diseases and disease comorbidity [ref]. However, these SNPs collectively only explain a small proportion of heritability [ref]. The small proportion of heritability explained by previous studies utilizing individual SNPs could be due to the methodological limitations of these studies which ignore genetic interactions due to insufficient power to detect them from the vast number of combinations [ref]. As a result, each SNP only has a small effect size and fails to explain a substantial proportion of heritability, even with their additive effects.

The continued treating each SNPs independently for the disease physiology makes the current GWAS results almost useless for translation to health care because effect of each SNP is almost neglectable, thus unable to predict any disease and its progression [ref]. If these SNPs indeed cause the disease, they should converge some points at the biological systems and finally perturb the human physiology to a disease state. Therefore, our objective is to identify common or cooperative mechanisms among distantly located, genetically independent SNPs for the same or comorbid diseases. Although some studies have implied or reported a few shared or linked mechanisms for complex diseases [ref] and comorbid diseases [ref], the comprehensive picture of biological perturbation mechanisms by SNPs for complex diseases and their comorbidities remains vague. **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 diseases, which drive detectable changes of biological entities at multiple related scales, which can be observed by a variety of assays. Specifically, we will integrate the abundant, multiple-scale functional data about DNAs in ENCODE repository by balancing the effects from multiple scales and handling many missing assays. We will develop an algorithm that combines the advantages of statistical method of MFA and classic machine learning techniques to mining the data buried in such a comprehensive data-driven project. This method would allow discovery biological mechanisms within the big data for complex disease and comorbidity.

This grant is expected to generate thousands of cooperative SNPs along with the driven mechanisms for complex diseases and disease comorbidity. We will also validate the top relationships in a large volume database consisting of both phenotype and genotype data at single patient level. The contribution of this project will be significant because it will greatly enhance our understanding causal mechanisms of disease and comorbidity by leveraging both genetics and biological mechanisms. The cooperative mechanisms among SNPs have been reported recently for correlated SNPs, but the mechanisms for independent SNPs are yet to be understood. Also, genetics have changed our view of disease classification, such as metabolic syndrome, but how it relates to complex disease comorbidity is only reported in a few diseases [ref]. On successful completion of this project, the understanding to both genetic mechanisms for individual complex disease and disease comorbidity will be greatly refreshed. More importantly, the validated results of this project can serve as biomarkers for disease diagnoses and disease progression with better accuracy than their individual ones. These results will also allow for new invention ways at the downstream genes of key causal cooperative SNPs, which prospectively improve the health care. Finally, this project will provide generalized approaches for data integration, big data analysis, and translational medicine, which fit well with NIH recent prompted directions.

**Innovation**

With the advance of many high-throughput techniques, the barriers of biological and biomedicine diseases lie on effective and efficient data integration, rather than data generation itself. The integration demand not only arises from multiple scale assays of a single project like ENCODE and The Cancer Genome Atlas (TCGA), but also from knowledge integration from several projects such as GWAS, eQTL, and ENCODE. Multiple factor analysis has been employed and referred in developing many novel mathematical methods [ref] but has only recently used in data integration of multiple scale data [ref]. The advance inspires and facilitates the novel application of multiple factor analysis for SNP epistatic analysis. Also, the combination of multiple factor analysis and machine learning techniques, such as K-nearest neighbor and clustering, is novel. Furthermore, the method combining multiple factor analysis and clustering techniques is not only applicable to the epistatic analysis, but also general enough for other omics data integration for an variety of applications, such as disease subtype identification, signature pathway analysis, and prognosis and survival analysis. We will open the source codes of the algorithms for potential applications of this integrative approach. In term of the epistatic underpinnings of complex disease and comorbidity, the integrative analysis of functional data in ENCODE can dramatically reduce the search space and facilitate the future intervention, thus is novel as well. Finally, the systematically investigation of causal biological mechanisms for complex disease comorbidity is also novel both conceptually and technically.