Title: A Bayesian Network Approach for Assessing Factors in Complex Disease Etiology

Introduction and Problem Statement: Complex diseases arise from the cumulative and joint effects of numerous genetic as well as environmental factors. Acquiring a better understanding of the degree to which these factors, either independently or combinatorially, contribute to disease etiology is crucial for successful health care planning and clinical intervention. Due to the inefficiency of observational and experimental epidemiological studies, sophisticated computational tools are needed to discern the complex interactions among these factors that dictate disease expression [1,2]. I propose a computational, Bayesian network approach to infer relationships among interacting genetic and environmental factors in the expression of 26* complex diseases using exome sequencing and medical record data from the Personal Genomes Project [3], the International HapMap Project, the 1000 Genomes Project and other available datasets and resources. I also propose a novel conservation-weighted-sum grouping method for collapsing genetic variants based on orthologous sequence conservation, and a novel gene-burden score to incorporate functional annotation information into the network. Predictive features selected by the network will be implemented on a support vector machine classifier to construct a disease phenotype prediction model. I will then systematically explore factor combinations in network construction to evaluate the resulting feature selection and subsequent impact on prediction performance to quantitatively estimate factor contributions to disease etiology.

Hypothesis: I hypothesize that networks constructed on solely genetic or environmental factors will lead to poor classifier prediction performance. *Key Question 1*: What combination of factors results in the optimal classifier prediction performance? *Key Question 2*: What are the underlying relationships among factors that dictate disease expression?

Methods: Variant filtering: To reduce computational load, all genetic variants from exome sequencing will be filtered to remove variants unlikely to be associated with disease using methodology based on my previous research on synergistic epistasis. All variants will be filtered against Yoruba data from 1000 Genomes to remove variants accumulated before the out-of-Africa migration. Resulting variants will be filtered against an appropriate background population based on demographic information to remove population-specific common variants.

Conservation-weighted-sum scores: The resulting sample size of variants per gene will likely be very small. Therefore group-wise analysis must be employed to reduce noise and improve statistical power. I propose a conservation-weighted sum (CWS) grouping method for collapsing variants v_1 through v_m in gene i: $CWS_i = \sum_{j=1}^{m} I_{ma}(v_j)/H(v_j)$ where I_{ma} is the number of minor alleles in variant v_j. $I_{ma} \in \{0,1\}$ if variant v_j is known to act dominantly or recessively, else $I_{ma} \in \{0,1,2\}$. H is the column entropy at the sequence position of variant v_j in the orthologous sequence alignment for gene i. The orthologous sequence alignment will be constructed using orthologous up to the optimal phylogenetic depth limit determined using methodology from my previous research. In this manner, each variant will be weighted by its position conservation to emphasize well-conserved variants. CWS scores will be generated in the same manner for pathways. CWS scores will be benchmarked against the more standard weight-sum score where each variant is weighted by its minor allele frequency to emphasize rare variants [4]. Both methods assume unidirectional, additive impact of variants.

Gene-burden scores: Variants confer differing effects based on their functional impact. Therefore incorporating functional annotation information into the network will be crucial for accurate

* Complex diseases are chosen based on available phenotypic information from the Personal Genomes Project [3]

disease phenotype prediction [1]. I propose incorporating functional annotation information using gene-burden scores for each gene i: $G_i = (1-b)(S_i + \alpha B_i) + bD_i$ where b is the CWS score for all deleterious variants in causing the disease phenotype and α is an empirically derived non-negative scalar. S_i, B_i, and D_i represent the CWS scores for synonymous, non-synonymous benign, and non-synonymous damaging variants in gene i respectively. Functional annotations will be obtained from databases such as dbSNP and HGMD if available or predicted using Polyphen2 [6]. Pathway-burden scores will be generated in the same manner.

Network Construction: Assuming continuous distributions of factors, linear Gaussian conditional densities will be used to specify distributions on the network. The network will be constructed on nodes containing variables for demographic and environment variables, quantitative traits, gene and pathway-burden scores, and disease phenotypes. The network will be optimized using standard Monte Carlo Markov chain methods including the Metropolis–Hastings algorithm [2,5]. Confidence scores for each edge will be determined by nonparametric bootstrapping.

Disease Phenotype Prediction: A disease phenotype prediction model will be constructed using a support vector machine classifier trained to optimize on the association of feature scores selected by the network with known disease phenotypes. Leave-one-out cross validation, standard AUC statistics, and ROC analysis will be used to assess classifier performance.

Disease Contribution: Factors will be systematically withheld from network reconstruction thereby generating a new set of features for classifier training and model validation. The resulting decrease or increase in classifier performance will be quantified as the factor's disease contribution. To ensure sufficient statistical power, factors may be withheld in groups.

Anticipated Results or Findings: I anticipate that the features resulting in the optimal classifier prediction performance will contain both genetic and environmental factors (*Key Question 1*). Gene-trait and pathway-trait edges inferred by the network should reflect known empirical relationships. Many environmental variables such as smoking may be inferred to directly affect disease phenotype. Edges formed in the network may reveal unexpected underlying relationships between factors (*Key Question 2*). However, if too many spurious edges are formed, certain directed edges may be restricted based on logical interpretability of edge relations.

Expected Significance and Broader Impacts: My research will provide a quantitative estimate of the degree to which genetic and environmental factors contribute to disease and will enhance the scientific understanding of disease etiology. Identification of novel underlying relationships between factors may provide new leads for further empirical analysis and future drug development or repurposing to improve disease treatment. My research will also provide further validation of the utility of bioinformatics and engineering principles in medicine and healthcare. Furthermore, my novel CWS and gene-burden scores may be integrated into other networks or prediction models thereby contributing to computational methodology and advancing bioengineering knowledge. A paper on the results and methodology of this research will be written for submission to relevant conferences and a peer-reviewed journal.

Literature Citations: (1) Kang J, et al. (2011) *Use of Bayesian networks to dissect the complexity of genetic disease: application to the Genetic Analysis Workshop 17 simulated data*. BMC Proceedings, 5:S37. (2) Needham CJ, et al. (2007) *A Primer on Learning in Bayesian Networks for Computational Biology.* PLoS Comput Biol 3(8): e129. (3) Church GM (2005) *The Personal Genome Project*. Molecular Systems Biology 1:2005.0030. (4) Madsen BE & Browning SR (2009) *A groupwise association test for rare mutations using a weighted sum statistic*. PLoS Genet, 5:e1000384. (5) Ben-Gal I, et al. (2007) *Bayesian Networks*. Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons. (6) Adzhubei IA, et al. (2010) *A method and server for predicting damaging missense mutations*. Nat Methods 7(4): 248-249.