The Pennsylvania State University

The Graduate School

Huck Institute of Life Science

USING LARGE-SCALE GENOMICS DATA TO UNDERSTAND THE GENETIC BASIS OF COMPLEX TRAITS

A Dissertation in

Bioinformatics and Genomics

by

Ruowang Li

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

December 2016

The dissertation of Ruowang Li was reviewed and approved* by the following:

Marylyn D. Ritchie
Professor of Biochemistry and Molecular Biology
Dissertation Advisor
Chair of Committee

Le Bao
Assistant Professor of Statistics

Ross C. Hardison
T. Ming Chu Professor of Biochemistry and Molecular Biology

Shaun Mahony
Assistant Professor of Biochemistry & Molecular Biology

Peter Hudson
Director, The Huck Institutes of the Life Sciences
Willaman Professor of Biology

*Signatures are on file in the Graduate School

ABSTRACT

With the arrival of big data in genetics in the past decade, the field has experienced drastic changes. One game-changing breakthrough in genetics was the invention of genotyping and sequencing technology that allows researchers to examining single nucleotide polymorphisms (SNPs) across the entire genome. The other major breakthrough was the identification of haplotypes of common alleles in major human populations, which permitted the design of genotyping assays that effectively cover entire human genomes at a resolution appropriate for genetic mapping. Together, these technology breakthroughs have permitted researchers to carry out Genome Wide Association Studies (GWAS) on a wide range of traits including, for example, height and disease status. With GWAS, causal SNPs have been identified for some Mendelian traits, but for more complex genetic traits, the genetic heritability explained by the associated SNPs are low. In addition, high-throughput technologies to generate other types of -omics data such as gene expression, DNA methylation, and protein levels data have also emerged recently. How to best utilize the SNP data and other multi-omics data to understand genetic traits is one of the most important questions in the field today.

With the increasing prevalence of multi-omics data, new types of analysis schemes and tools are needed to handle the additional complexity of the data. In particular, two areas of method development are in great need. First, statistical methods employed by GWAS do not consider the potential interacting relationships among genetic loci. Thus, methods that can explore the joint effect between multiple genetic loci or genetic factors could unveil new associations. Second, different types of –omics data may give distinctive representations of the overall biological system. By combining multi-omics data, we could potentially aggregate non-overlapping information from each individual data types. Thus, the focus of this dissertation is on developing and improving computational methods that can jointly model multiple types of

genomics data. First, an evaluation of an existing method, grammatical evolution neural network, was conducted to identify the optimal algorithm settings for the detection of genetic associations. It was found that under certain algorithm settings, the neural networks have been restricted to one-layer simple network. Using a parameter sweep approach, the analysis identified optimal settings that allow for building more flexible network structures. Then, the algorithm was applied to integrate multi-omics data to model drug-induced cytotoxicity for a number of cancer drugs. By combining different types of –omics data including SNPs, gene expression and methylation levels, we were able to model a higher portion of the observed variability than any individual data type alone. However, one drawback of the existing neural network approach is the limited interpretability. To this end, a new algorithm based on Bayesian Networks was created. One novelty of the approach is the ability to independently fit a distinct Bayesian Network for each categories of a phenotype. This allows for identifying category specific interactions as well as common interactions across different categories. Analysis using simulated SNP data has shown that the Bayesian Network approach outperformed the Neural Network approach in many settings, particularly in situation where the data contains multiple interacting loci. When applied to a type 2 diabetes dataset, the algorithm was able to identify distinctive interaction patterns between cases and controls. Ultimately, the goal of this dissertation has been to fully take advantage of the newly available data to understand the genetic basis of complex traits.

TABLE OF CONTENTS

* Portions of the chapters are from published manuscripts for which Ruowang Li is the primary author

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Through my journey as a graduate student, I am very fortunate to be guided by a great advisor and supported by encouraging friends and family. This dissertation would not have been possible without the help of so many of you.

Foremost, I would like to express my sincerest gratitude to my advisor, Dr. Marylyn Ritchie. You have provided me countless opportunities to be a successful PhD Student. As a scientist, you inspired me with cutting-edge research, encouraged me to think independently, and taught me problem solving skills. As a mentor, you helped me to establish project collaborations, supported me to attend scientific meetings, and always encouraged me through successes and struggles. I am so honored and grateful to have you as my advisor and I look forward to continue learning from you in the future. I also thank my committee members, Dr. Le Bao, Dr. Ross Hardison, and Dr. Shaun Mahony for their constructive feedbacks and suggestions throughout my graduate studies. I have learned a great deal from each one of you.

I have also been surrounded by talented and helpful colleagues in the Ritchie lab. I'd like to thank Dr. Dokyoon Kim not only for all the help and discussions, but also for our regular meetings at the big bowl. I also thank Scott Dudek for implementing and maintaining ATHENA, especially for the GEBN algorithm. Scott has been my go to person for any computational related questions. I also thank Dr. Emily Holzinger for introducing and guiding me through my first ATHENA project. I would also thank Dr. Molly Hall for our discussions in the past year. I also appreciate the help from Blair Zhang and Victoria Li for the GEBN project. Thank you also to other programmers and graduate student John Wallace, Anurag and Shefali Verma, Anna Okula,

and Alex Frase for being always helpful to me. Finally, I would like to thank our staff members Suzy Unger and Donna McMinn for their administrative support.

I would also like to thank my collaborators. Dr. Eileen Dolan and Dr. Heather Wheeler have generously shared their data and time and again having discussions on the project. I would also like to thank Dr. Le Bao for his significant contributions to the GEBN project and for helping me with many statistical questions. I am also grateful to be supported by the fellowship from the National Science Foundation.

I also thank Dr. Cooduvalli Shashikant, who oversees the BG program, for regularly checking with me regarding my research, future career plans, and many other issues. I also thank him for giving me the opportunity to organize the BG retreat. I also appreciate the opportunities given by Dr. Frank Pugh and Dr. Michael Axtell to rotate in their labs. I also want to thank many current and former BG students who have helped me tremendously, Dr. Garam Han, Dr. Rohit Reja, Dr.Melissa Wilson Sayres, and Dr. Jihye Park.

I especially want to thank my parents, Elizabeth and Francis, for the sacrifices they made for me. My mom was the first college student in her town and my dad became an English teacher through self-study. My parents placed great emphasis on my education when I grew up, I feel lucky to have such an amazing parents. Finally, I want to thank my girlfriend, Jing, for being my friend and companion despite the long distance between us. Thank you for bringing joy to my life.

Lastly, I want to thank God for all the abilities you have given me.

Chapter 1*[1]

Introduction

Background

With the rapid development and increasing prevalence of high-throughput multi-dimensional genomics data, genetic research on complex human traits has seen its natural transition to genomics, which concerns with genome-wide interrogation of trait influencing genetic factors. With the arrival of big data in genetics research, the field is facing unprecedented challenges to utilize and develop appropriate analytical approaches to keep up with the increasing growth of data. Thus, the goal for my dissertation is to develop and apply new computational tools that can maximize the utilization of "big data" in genetics.

The straightforward interpretation of "Big data" pertains to the volume of datasets. However, big data is also characterized by three additional Vs: variety, velocity, and veracity(1). In the past decade, we have seen volume of datasets increasing from a few hundred to millions of independent variables. The growth was also observed in a variety of data including, but not limited to, sequencing data, imaging data, and electronic health record data. For sequencing data alone, double the amount of data are being produced every seven months(2). Accompanied with the increasing amount of data is the increase of biases and noise in the data. Thus, more attention is needed to deal with the veracity of the data. Recently, it has been estimated that the computing resources needed to analyze genomics data will exceed that of YouTube, Twitter and astronomy

[1] Portions of the chapter are under preparation for journal submission

by 2025(2). As such, the field is experiencing an unusual growth that needs to be met with commensurate tools.

The significant growth of data in genetics was achieved by the maturation of genome-wide genotyping and sequencing technology, which has led to the development of Genome-Wide Association Studies (GWAS). In the context of genetics research, GWAS systematically evaluates common single nucleotide polymorphisms (SNPs) throughout the genome for association with various phenotypes, such as height(3) and disease status(4). Compared with the candidate gene approach, GWAS excels in discovery of novel associations and the ability to systematically test a much larger number of hypotheses. Many previous research projects have reported successful GWAS studies (5–11). However, while GWAS have identified the most strongly associated individual variants, much more information could still be gained from the data. In addition, association between observable phenotypes with other intermediate phenotypes such as gene expression(12), epigenetic variations(13,14), and protein variations(15) are starting to be explored.

Various computational methods have been developed to handle the increasingly immense and complex data in genomics. Owing to the complex architecture of genetic traits, there is not an off-the-shelf method that is suitable for all types of data and analysis. Different methods are needed for performing association analysis, identifying genetic epistasis, incorporating prior biological knowledge, and carrying out system genomics analysis. In this chapter, I will describe the principles of various analytical methods commonly used in genomics research, the strength and weakness of these approaches, and some tools that implement these strategies.

Association analysis

The simplest and most commonly used analysis strategy in genomics research is the association analysis. As its name suggests, association analysis aims to identify factor(s) that are marginally associated with genetic traits. Example outcomes include whether or not having a disease (case-control), height (quantitative trait), and different levels of drug response (categorical). In statistical terms, association analysis tests the null hypothesis that no factor(s) are associated with the phenotype. Under this null hypothesis, if the observed data lies in the extreme ends of the null distribution, the null hypothesis can be rejected and we can conclude a statistically significant association. Depending on the types of the phenotype, different statistical tests can be used (Table 1-1).

Table **1-1**: Common statistical tests used in association analysis.

Phenotype Variable	Statistical test	References
Binary (case-control)	Pearson 2df, Fisher, Cochran-Armitage, Logistic regression	(16–18)
Continuous (quantitative trait)	Linear regression, ANOVA	(19,20)
Categorical (high, intermediate, low metabolizers)	Multinomial regression, proportional odds regression	(21,22)

The most common types of association analysis are candidate gene studies and genome-wide association studies (23). A genetic study with the candidate gene approach typically interrogates up to hundreds of SNPs within carefully selected biological candidate genes. As such, the most important factor in determining the success of candidate gene study relies on the selection of genes. Selection of candidate genes is generally based on manual search of previously published literature; however, there are also computational methods that can

automatically select candidate genes, for example, based on gene-gene interactions(24). GWAS differs from the candidate gene approach in both scale and interpretation. A typical GWAS includes tens-of-thousands to up to millions of SNPs genotyped in both genic and nongenic regions. Because of this, it can potentially identify novel associations that are not explored by the candidate gene approach. GWAS is also largely "hypothesis free" as opposed to the candidate gene approach, which tests for specific genes' association with the phenotype. Notable examples of GWAS include studies on interferon-α(25–27), statin-induced myopathy(8), and various chemotherapeutic drug-induced responses(20,28–30). In addition, association analysis using other –omics data are also gaining traction. Gene expression has been used to link gene targets to complex phenotypic traits(31–33). Metabotype have been shown to affect treatment response to SSRIs, Lithium, Aspirin and clopidogrel(34). Other data types including protein levels(35), epigenetics(36–38), and copy number variations(39,40) have all demonstrated the power of association analysis in genetic research. However, the major challenge in genome-wide association analysis is the burden of multiple-testing penalties. The basic idea is that given a fixed significance threshold α, the more number of tests being conducted, the more chances of obtaining a type I error. For example, using the typical α of 0.05, conducting 100 tests could have 99.4% probability of finding one significant SNP by random chance (P(at least one type I error) = 1-P(no type I error) = $1-0.95^{100}$=0.994). Thus, the α threshold needs to be lowered to control the number of false positive results in genome-wide studies. The most stringent method for adjusting the α threshold is the Bonferroni correction. In order to achieve the expected α level for all tests, we need to estimate a new α' threshold that satisfy $\alpha = 1-(1-\alpha')^n$, which leads to the Bonferroni corrected α' that is approximately equal to α /n. For example, to achieve an α of 0.05 for 1 million independent tests, the α' needs to be $5x10^{-8}$. The Bonferroni correction is overtly stringent because genome-wide SNPs are not independent, thus n is in fact smaller than the number of total SNPs and is population specific(41). An alternative and less conservative approach to Bonferroni

is the false discovery rate (FDR) approach. FDR controls the proportion of false positive tests among all positive tests. Outside of the frequentist paradigm, Bayesian approaches can avoid the multiple-testing adjustment because the prior probability of associations do not depend on the number of tests performed(42).

Despite the simplicity of the association analysis, it still occupies an important position in the field today. The parsimonious nature of the model enables easy interpretation and replication. Because association analysis is carried out one variable at a time, the analysis can easily be parallelized. As the size of datasets become in the range of Gigabytes and Terabytes, association analysis may be the only viable analysis that can be systematically carried out and compared across different studies.

A number of software packages have been created to perform association analysis. Many have the ability to include covariates such as age and gender. Several mixed model based methods can also automatically adjust for population sub-structures. Popular software for association analysis include Plink(43), PLATO(44), GenABEL(45), GEMMA(46), and FaST-LMM(47).

Epistasis

Epistasis can be interpreted differently under different contexts, but most falls under the distinction between biological and statistical epistasis(48). Briefly, biological epistasis represents the physical interactions among molecules in an organism, while statistical epistasis can be seen as mathematical deviation from additivity in the collected data. Figure 1-1 is an example of statistical epistasis, which is the focus of this section.

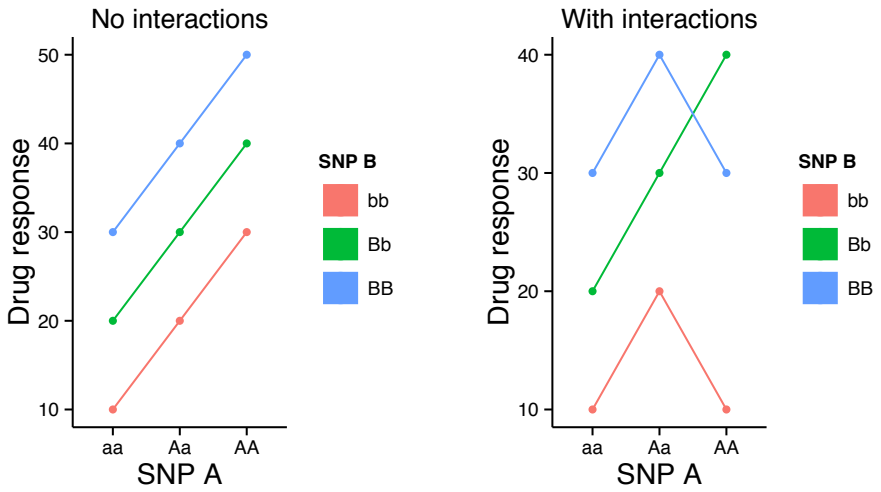


Figure 1-1. Epistasis between two SNPs. Panel A shows the effect of two SNPs on a drug response without epistatic effect. Panel B shows the situation where the drug response depends on the epistasis effect of two SNPs.

One question remaining unanswered is whether SNPs affect genetic traits independently and additively (Figure 1-1A), two assumptions of association studies, or whether the SNPs' effects are dependent on each other (Figure 1-1B). Exploring epistasis, or gene-gene interactions, is a worthwhile complementary strategy in genetics research due to its successes in model organisms(49); however, mapping epistasis is no doubt computationally challenging. Imaging there are N SNPs of interest. Evaluating all pair-wise interactions would equate to (N choose 2) comparisons, thus a typical GWAS of 500,000 SNPs would results in (500,000 choose 2) or $1.2x10^{11}$ comparisons, a difficult yet feasible analysis. However, whole genome sequencing data can easily generate tens of millions of variables, a dataset with 10 million SNPs would equate to $5x10^{13}$ or 50,000 billion pairwise comparisons. Using a lower estimate, if a GWAS analysis takes one hour of computational time, 50,000 billion comparisons would take about one hundred million computer hours. Despite the obvious challenges to identify epistasis from large-scale datasets, there have been increased efforts in methodology development in this area.

Variable filtering is a common strategy to reduce the search space for identifying epistasis. Examples of filtering strategies include hypothesis-driven tests of interaction on the basis of external knowledge (e.g candidate gene, pathways) or hypothesis-free filtering based on the data characteristics (e.g. minor allele frequency threshold or SNPs main effects(50–52)). Filtering strategy not only can speed up the analysis, it can also lead to increased power due to the lowered multiple testing penalties. However, it is also prone to increase false negatives because the causal variants can be filtered out.

Whether or not variable filtering is performed, a regression model can be used to test the existence of epistasis. The most common way to test for interactions using regression models is to include main effect as well as interaction terms in the model then test if the interaction term's coefficient equals to zero(53). i.e Testing two way interactions involving two SNPs would take form of a regression equation $Y = b0 + b1(SNP1) + b2(SNP) + b3(SNP1*SNP2) + e$ with the null hypothesis being tested is $b3=0$. It should be noted that testing all possible interaction models would dramatically increase the multiple testing penalties. For higher order interactions, one of the most widely used methods is Multifactor Dimensionality Reduction (MDR)(54,55). In a two-way interaction model involving di-allelic loci, MDR reclassifies the nine possible genotype combinations into "high risk" and "low risk" groups based on the case and control ratios of each genotype combination. Similarly, any n-way interaction models can be reduced to a one-dimensional model, effectively lowering the number of parameters of statistical tests. MDR has been widely applied to human complex diseases (56,57). Variations of MDR have also been developed to be applicable to continuous phenotypes and have the ability to include covariates(58). Many other approaches in the realm of machine learning and Bayesian methods that were also developed to detect epistatic interactions are shown in Table 1-2.

Table **1-2**: A partial list of data analysis methods of epistasis.

Approach	Methods	Software and/or tools
Statistical	Regression	Plink(43)
	Regression	Plato(44)
	Regression	BOOST(59)
	Regression	EPIBLASTER(60)
	Regression	eCEO(61)
	Bayesian	BEAM(62)
Machine Learning	Neural Networks	ATHENA(63)
	Nonparametric	MDR(54)
	Bayesian Networks	ATHENA(64)
	Bayesian Neural Networks	BNN(65)
	Random Forest	Random Jungle(66), Ranger(67)
	RELIEF	ReliefF/Turf(68,69)

Biological knowledge

With the arrival of low cost and easily generated big data, the typical genetics research has transitioned from candidate gene studies into genome-wide studies. As alluded to previously, the main challenges in the analysis methods of big data are the computational burden of exploring of a large search space and with it, the increased multiple hypothesis penalty. As a result, many

variable selection techniques have been developed to reduce the number of total variables. Typically, the hypothesis-free variable selection techniques assess the associations of predictor variables and the outcome variable. The draw back of using the outcome variable during variable selection is that the subsequent analyses are no longer independent, complicating the multiple hypothesis adjustment calculation. Alternatively, prior biological knowledge can be incorporated into genetics research to help generate and prioritize hypothesis (Table 1-3). Huang et al. used Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) to annotate genes and found significant Drug–gene-pathway relationships associated with drug response(70). Lin et al. designed an automatic system to extract previously published drug-related experiments from Gene Expression Omnibus (GEO) that can be used for meta-analysis(71). A tool named Biofilter has been developed to provide a platform to integrate multiple publically available biological databases(72) and has been applied to study virologic failure in efavirenz-containing treatment(73). Incorporating prior biological knowledge can aid in results interpretation; however, researchers should also keep in mind that prior knowledge is limited to what is known and published. This may bias the interpretation and limit the opportunity to identify novel findings.

Table **1-3**: Biological knowledge databases in pharmacological research.

Databases	Information	URL
KEGG: Kyoto Encyclopedia of Genes and Genomes(74)	Biological pathway, diseases, drugs, chemical substances.	http://www.genome.jp/kegg/
GO: Gene Ontology(75)	Gene, gene products, annotations.	http://geneontology.org/
GEO: Gene Expression Omnibus(76)	Genomics data repository	http://www.ncbi.nlm.nih.gov/geo/
PharmGKB: The Pharmacogenomics Knowledgebase(77)	Genomics, phenotype, and clinical information	https://www.pharmgkb.org/
Biofilter(72)	Variants annotation and filtering	https://ritchielab.psu.edu/research/research-areas/expert-knowledge-bioinformatics/methods/biofilter

Systems genomics

In addition to the advancement in generation of DNA sequence data, a whole range of other multi-omics data including gene expression levels(78,79), epigenetic profiles(80), and proteomics(81) have populated the field of genetics (Figure 1-2).

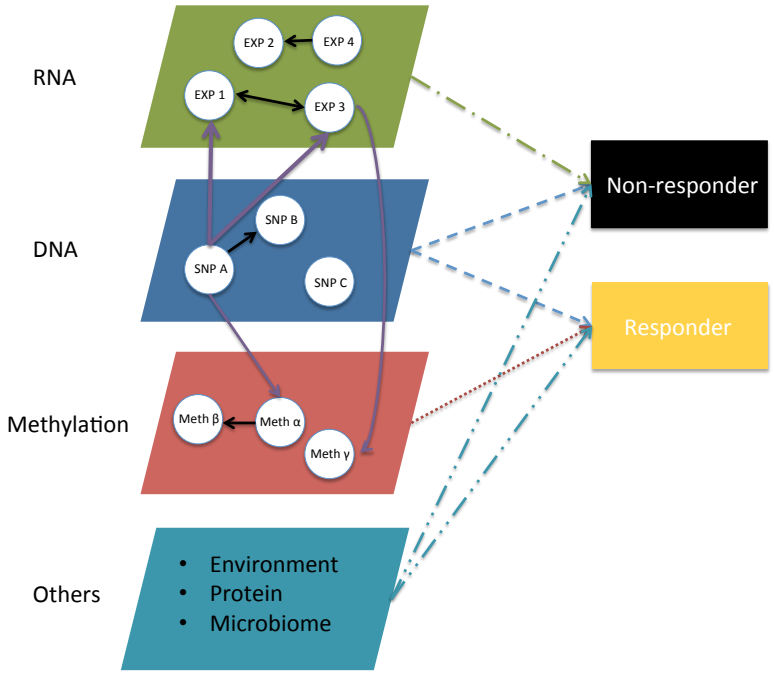


Figure 1-2: A system genomics view of genomics. Complex biological and statistical relationship exists between and within each level of genomics data. Only a comprehensive analysis of all data may reveal the true determinant of the genetic outcome, e.g. drug response.

How to best handle these additional sources of data poses serious challenges due to the unique characteristics associated with each of the data type in addition to the computational challenges due to the dataset size and the computational complexity due to the combinatorics and very large search space. Previous works have showed that a system genomics approach can better interrogate the genetic and phenotype associations than analysis methods based on a single data

type(82–84). Due to the heterogeneity of the data types, methods used in systems genomics approaches are extremely varied. To determine the appropriate approaches, the researcher needs to first formulate the biological question into an analytical one. As an example, if the biological question is to determine subtypes of a disease, the analytical problem would be, if the data is stored as individuals x features (N x P), divide N into X groups such that p vectors in the same group are closer to each other than in other groups according to some distance measure D. N, P, X, and D then becomes important criteria for selecting the appropriate method (Figure 1-3).

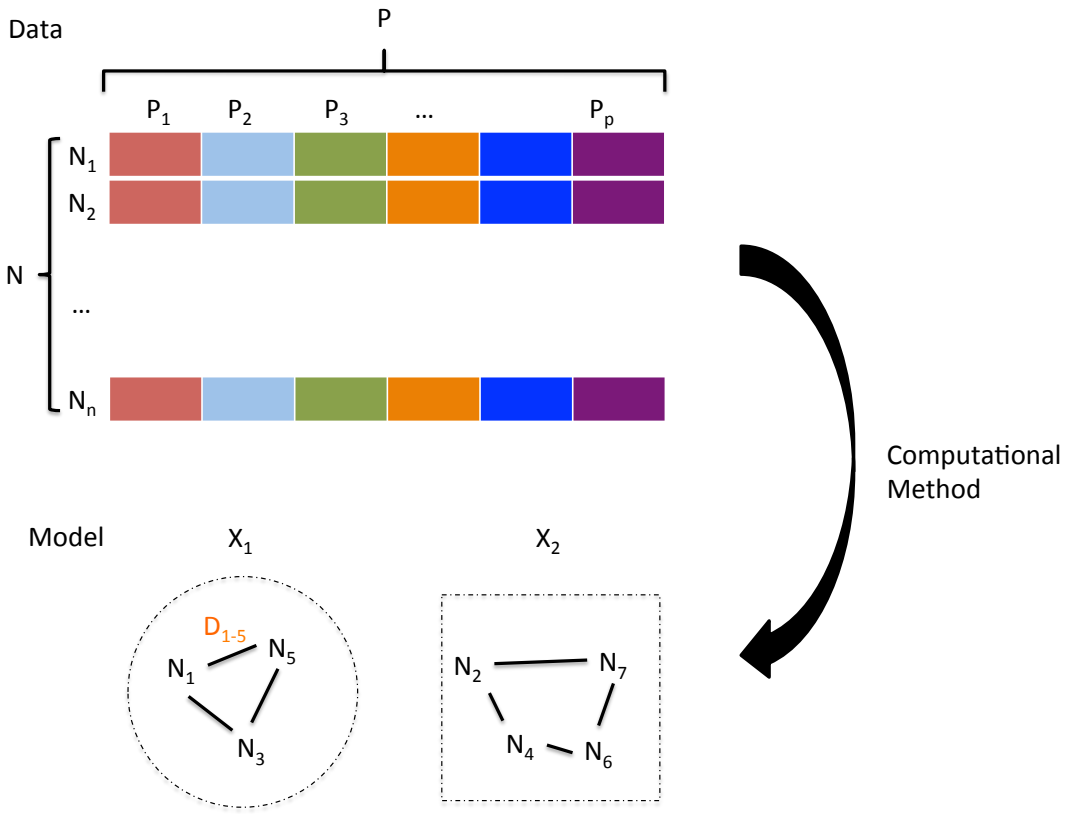


Figure 1-3:Analytical representation of identifying subtypes of a disease. N represents a population of individuals with a disease. P is a vector of measurements taken on each individual, e.g. SNPs or gene expression. The goal is to cluster N into X clusters according to the similarities between P features. The similarities are measured using a distance measure D. The resulting X clusters represent subtypes of the disease.

Generally speaking, system genomics approaches can be further divided into unsupervised or supervised methods. In the framework of unsupervised learning, the goal is to uncover patterns in the data without the presence of a phenotype variable. An important application of unsupervised learning is to infer subtypes of a disease. For instance, iCluster(85) is a method that can integratively cluster samples based on multiple source of genomics data. It achieves this by first building a latent model for each type of data, then a K-means clustering algorithm is applied on the latent models to cluster the samples. Using SNPs, gene expression and Copy number variations (CNVs), iCluster was able to come up with new subtypes of breast cancer(86). Unsupervised learning has also been used in genetics to identify potential confounding factors in the data. Visscher et al. applied principal component analysis to a pharmacogenomics study in Canada and they were able to infer patients' ancestry from a set of key biotransformation gene loci(87). Galvan et al. used a clustering method, AWclust(88), to group patients based on their ancestry and found that opioid-mediated pain relief is not associated with cluster memberships(89).

In the realm of supervised learning, system genomics approach aim to build models that are predictive of the phenotype variable. Traditional association analysis of each individual data type separately is not optimal because information that is shared across multiple data types is not being explored. In real biological systems, it is likely that multiple layers of complexity are underlying the observed phenotype. Being able to model this complexity using multiple sources of data offers unprecedented opportunity for a more comprehensive view of the system. Integrative analysis of multi-omics data can be broadly categorized into multi-stage analysis and meta-dimensional analysis(90,91).

Multi-staged analysis

Multi-staged analysis, as its name suggests, aims to divide data analysis into multiple steps, and signals are enriched with each step of the analysis. The main objective of the multi-staged approach is to divide the analysis into multiple steps to find associations first between the different data types, then subsequently between the data types and the trait or phenotype of interest. Examples of multi-staged analyses are shown in Figure 1-4 and described below.

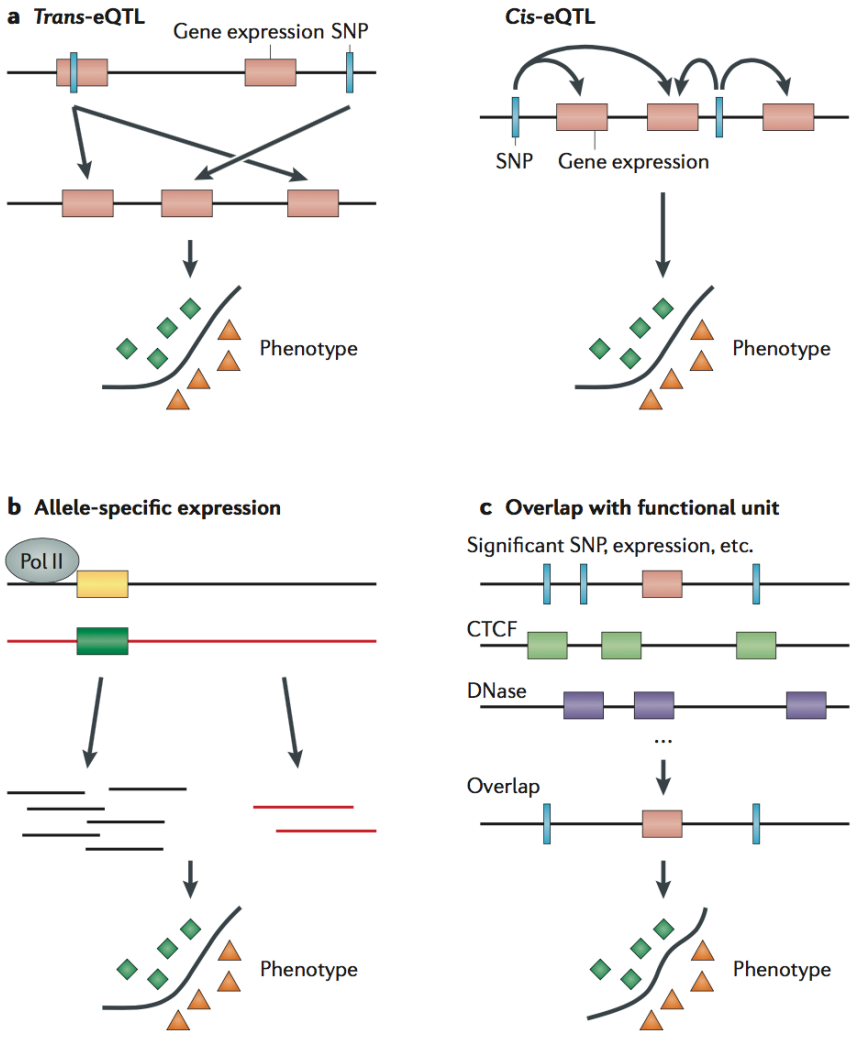


Figure 1-4: Categorization of multi-staged analysis. Multi-staged analysis can be divided into three categories. a | Analysis of expression quantitative trait loci (eQTLs) analysis involves the identification of genetic variation associated with measures of quantitative gene expression. b |

Allele-specific expression involves the analysis of whether the maternal or paternal allele is preferentially expressed, followed by the association of this allele with *cis*-element variations and epigenetic modifications. c | Domain knowledge overlap involves a two-step analysis in which an initial association analysis is performed at the single-nucleotide polymorphism (SNP) or gene expression variable followed by the annotation of the significant associations with knowledge generated by other biological experiments. This approach enables the selection of association results with functional data to corroborate the association. CTCF, CCCTC-binding factor; Pol II, RNA polymerase II. (Source: Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet)

The most commonly used genomic variation integration technique so far has been a three-stage or triangle method(91). In the triangle method, the following steps are taken.

1. SNPs are associated with the phenotype and filtered based on a genome-wide significance threshold.

2. SNPs deemed significant from step 1 are then tested for association with another level of omic data. For example, one option is to look for the association of SNPs with gene expression levels. These SNPs are called expression quantitative trait loci (eQTLs). Alternatively, methylation QTLs (mQTLs; which are SNPs associated with DNA methylation levels), metabolite QTLs (which are SNPs associated with metabolite levels) and protein QTLs (pQTLs; which are SNPs associated with protein levels or other molecular traits such as long non-coding RNA and miRNA) could be used.

3. Omic data used in step 2 are then tested for correlation with the phenotype of interest.

Different methods of analysis can be used to implement this triangle approach, including linear or logistic regression (depending on a continuous or a binary dependent variable, respectively). The rationale of this approach is based on the assumption that genetic variations are the foundation of all other molecular variations. The triangle approach has been used, for example, in studies of chemotherapeutic drug response in HapMap cell lines, in which significant eQTLs were tested for correlation with the drug response(20,92,93). The difficulty of triangle-based

methods comes when a relatively arbitrary threshold, generally a *P* value, is used to identify the significant associations for further analyses. As the *P* value threshold also needs to be adjusted for the number of tests being carried out to combat multiple testing problems, there is likely to be a large number of false-negative SNPs, eQTLs, mQTLs and pQTLs being filtered out. This approach is often used to find SNPs associated with both a gene expression trait or a methylation level and the phenotype of interest to focus on functional SNPs.

Some researchers have begun to develop causal inference association approaches. For example, Schadt *et al.* have introduced a multistep approach to identify key drivers of complex traits that exploit the naturally occurring DNA variation observed in populations(94). DNA variation is tested for association with gene expression, and gene expression traits are then ordered relative to one another. Analyses then determine whether DNA variants that lead to variation in relative transcript abundances are supported statistically as an independent, causative or reactive function[43] using maximum likelihood approaches. These causal approaches(94,95) allow the dissection of the genotype-to-phenotype process in a clear, linear manner.

Another approach that links genomic variations to transcript levels is called allele-specific expression (ASE). In diploid organisms, one of the two alleles is preferentially expressed in some genes(96). ASE variants are associated with *cis*-element variations and epigenetic modifications(97). The first step of ASE approaches is to distinguish the gene product of one parental allele from the product of the other parental allele. Next, an analysis to associate the allele with gene expression (eQTLs) or methylation (mQTLs) can be carried out to compare the two alleles. Finally, the resulting alleles can be tested for correlation with a phenotype or an outcome of interest. The practicality of this approach depends on the extra resources used for experimentally tagging the two alleles and the subsequent mapping of the alleles. ASE and other extended methods such as allele-specific transcript structure (ASTS), which looks at the frequency of expression of splice transcripts that are allele-specific have been used to identify

functional variation(98) and protein–DNA(99) interactions in humans. This allele- specific approach has also been used in other contexts. For example, several groups have explored allele-specific analysis in chromatin state(100) and histone modification(101). More allele-specific applications are likely to emerge as we continue to observe these allele-specific effects.

Other studies have integrated functional and pathway information that is generated and consolidated by initiatives such as the Encyclopedia of DNA Elements (ENCODE)(102) and the Kyoto Encyclopedia of Genes and Genomes (KEGG)(74) to select and annotate significant results. In this approach, the genomic regions of interest are inputs. Various software and databases can be used to determine whether the regions are within pathways and/or overlapping with functional units, such as transcription factor binding, hypermethylated or hypomethylated regions, DNase sensitivity and regulatory motifs. For example, a researcher may take a collection of genotyped SNPs and annotate them with domain knowledge from multiple public database resources. The subsequent list of SNPs that have functional annotations can then be taken into the next stage, during which they are associated with other omic data, such as gene expression data (from microarray or RNA-seq) or metabolomic data. The resulting SNPs that have functional annotations and that are associated with other omic data can then be evaluated for correlation with a phenotype or an outcome of interest. This approach can be similar to the triangle approach mentioned above, with the exception that there is another step of annotating the variants and only taking those with functional annotations to the next stage of analysis. Adding information from diverse data sets can substantially increase our knowledge of our data; however, we are also limited and biased by current knowledge.

Even though multi-staged analysis uses both linear and nonlinear analytical mathematics to understand the relationship between two different types of data, there are clear limitations. For example, if complex traits are the result of a combination of DNA sequence variants, gene expression variability, methylation states and protein structure or expression changes that occurs

simultaneously along with environmental perturbations rather than in a stepwise linear model, the multi-staged approach will fail to effectively model the complex trait. However, when the relationship between genotype and phenotype can be modeled in a linear manner, as is the case for SNPs associated with metabolites and subsequently associated with phenotypes, for example, a multi-staged analysis would be applicable.

Meta-dimensional analysis[2]

Meta-dimensional analysis combines multiple data types in a simultaneous analysis(91,103,104) and is broadly categorized into three approaches: concatenation-based integration, transformation-based integration and model-based integration (Figure 1-5).

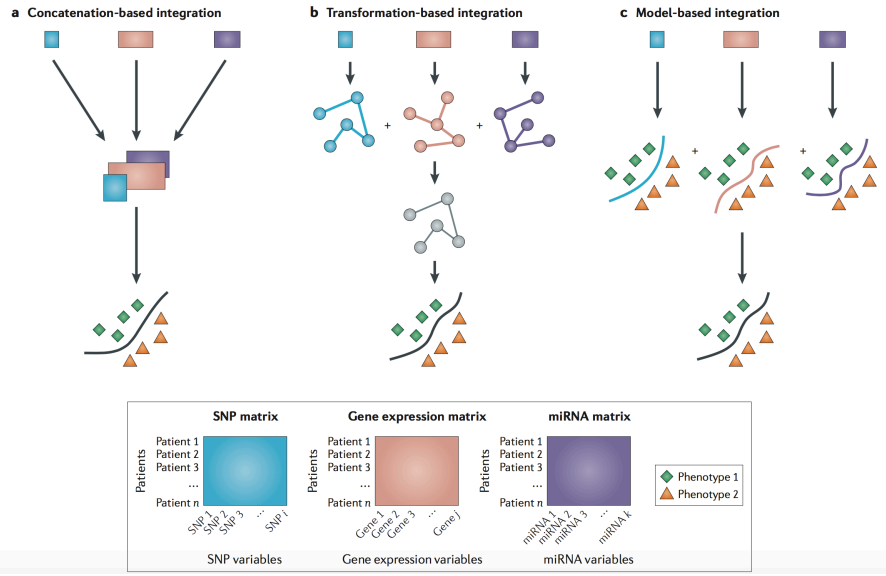


Figure 1-5: Categorization of meta-dimensional analysis. Meta-dimensional analysis can be divided into three categories. a | Concatenation-based integration involves combining data sets

[2] Ruowang Li is the secondary author for this section, which was adapted from Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet.

from different data types at the raw or processed data level before modelling and analysis. b | Transformation-based integration involves performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices. c | Model-based integration is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest. miRNA, microRNA; SNP, single-nucleotide polymorphism. (Source: Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet)

Concatenation-based integration combines multiple data matrices for each sample into one large input matrix before constructing a model. One advantage of concatenation-based integration is that, after it is determined how to combine the variables into one matrix, it is relatively easy to use any statistical method for continuous and categorical data for analysis. For example, Fridley et al.(105) performed concatenation-based integration by incorporating multiple types of genomic data into an association analysis with a complex phenotype using a Bayesian modelling strategy. Data from SNPs and mRNA gene expression were combined into a single data matrix, and the joint relationship of mRNA gene expression and SNP genotypes was then modeled using a Bayesian integrative model to predict a quantitative phenotype (for example, drug cytotoxicity). Mankoo *et al.*(106) predicted time to recurrence and survival in ovarian cancer using copy number alteration, methylation, miRNA and gene expression data using a multivariate Cox LASSO (least absolute shrinkage and selection operator) model. This strategy involves performing variable selection via LASSO, rather than a stepwise method, and then modelling the selected set of variables in a Cox regression. The other main advantage of this approach is that concatenation-based integration is particularly useful for considering interactions between different types of genomic data. For example, if the underlying model that one is trying to detect is a SNP interacting with metabolite to explain disease risk and if the two variables are not combined into one model, then the effect may be missed. This approach has been used to

combine SNP and gene expression data to predict high-density lipoprotein cholesterol levels(107,108), and to identify interactions between copy number alteration, methylation, miRNA and gene expression data associated with cancer clinical outcomes(109).

The challenge with concatenation-based integration is identifying the best approach for combining multiple matrices that include data from different scales in a meaningful way. For example, SNP data contain 0, 1 or 2 as values corresponding to the copies of a specific allele per individual; copy number data may consist of –2, –1, 0, 1 or 2 as values corresponding to copy number status in a given genetic region (although they can also be continuous-scale data); and DNA methylation profiles report between 0 and 1 for CpG loci. Identifying a way to appropriately integrate or combine these data without biases driven by data type can be challenging. Furthermore, this form of data integration can inflate high-dimensionality for the data, with the number of samples being smaller than the number of measurements for each sample(110). Thus, concatenation-based integration is only suitable if the appropriate way to assemble the data matrix for analysis is determined. Subsequently, statistical or computational models can be used to analyse the data matrix to consider interactions between different types of genomic data. Data reduction strategies as described above may be needed, depending on the number of variables in the data matrix. If there are too many variables, the analysis may not be computationally feasible; therefore, performing data reduction to limit the number of variables would be required to make this analysis possible.

The second approach, transformation-based integration, combines multiple data sets after transforming each data type into an intermediate form, such as a graph or a kernel matrix (a symmetrical and positive semi-definite matrix that represents the relative positions of all samples conducted by valid kernel functions). Multiple graphs or kernels can then be merged before elaborating any models (Figure 1-5). The transformation-based integration approach has the advantage of preserving data-type-specific properties from each data set when each type of data is

transformed into an appropriate intermediate representation. In addition, this approach can be used to integrate many types of data, including continuous or categorical values and sequence data, as long as the data contain a unifying feature, such as patient identifiers linking data types. Moreover, the transformation-based integration approach is robust to different data measurement scales.

For example, Lanckriet *et al.*(111) proposed kernel-based integration for protein function prediction with multiple types of heterogeneous data, including amino acid sequences, hydropathy profiles, gene expression data and known protein–protein interactions, and Borgwardt *et al.*(112) combined structural, sequential and chemical information into one graph model for predicting protein function via graph kernels. By contrast, Tsuda *et al.*(113) and Shin *et al.*(114) predicted protein function with multiple networks using graph-based semi-supervised learning. Kim *et al.*(103) proposed a graph-based integration framework for predicting cancer clinical outcomes using copy number alteration, methylation, miRNA and gene expression data. The disadvantage of transformation-based integration is that identifying interactions between different types of data (such as a SNP and gene expression interaction) can be difficult if the separate transformation of the original feature space changes the ability to detect the interaction effect. Each data type is transformed independently, which can make it more difficult to detect some effects. The goal is to perform a data transformation that maintains the majority of the data-type-specific properties so that these types of interaction effects are not missed. Thus, transformation-based integration is suitable if there is a relevant intermediate representation, such as a kernel or graph, for each genomic data type, and the goal is to preserve data-type-specific properties while integrating them.

Model-based integration, the third meta-dimensional approach, encompasses methods in which multiple models are generated using the different types of data as training sets, and a final model is then generated from the multiple models created during the training phase, preserving

data-specific properties. This approach can combine predictive models from different types of data. For example, model-based integration may allow the integration of data sets in which each data type is collected from a different set of patients but all patients have the same disease or phenotype. If the goal is to identify genetic, genomic and proteomic associations with ovarian cancer, data sets could be extracted from the public domain, where DNA sequence data may be available on five sets of patient samples, microarray data on eight sets of patient samples, and proteomic data on two sets of patient samples. Model-based integration would allow the independent analysis of each of the 15 data sets, followed by an integration of the top models from each data set to look for integrative models. This is an area of future work for the Analysis Tool for Heritable and Environmental Network Associations (ATHENA) methodology(107,115,116). ATHENA is a suite of analysis tools for performing systems genomic analyses to integrate different omic data and look for association with clinical outcomes. Model-based integration has been performed with ATHENA to look for associations between copy number alterations, methylation, microRNA and gene expression with ovarian cancer survival(115). A neural network model was constructed for each data type (such as copy number aberration and methylation) separately, and the four resulting models were then analysed to create an integrative model. As another example, a majority voting approach was used to predict drug resistance of HIV protease mutants(117) using structural features of the HIV protease–drug inhibitor complex and DNA sequence variants. In most cases, the variables from the top models are combined in a subsequent analysis. In addition, ensemble classifiers — such as predicted secondary structure, hydrophobicity, van der Waals volume, polarity, polarizability and pseudo-amino acid composition — have been used to predict protein fold recognition(118). The resulting models (from each data type) were combined in a weighted voting scheme to determine the fold of the protein. Finally, network- based approaches have been developed in which a Bayesian network is constructed using gene expression data, metabolomic data and SNP genotype data,

followed by integration to construct probabilistic causal networks(119–121). In each of these model-based integration examples, a model is built on each data type individually, and the models are then combined in some meaningful way to detect integrative models.

It is important to note that model-based integration requires a specific hypothesis and analysis for each data type, and a mechanism to combine the resulting models in a meaningful way. Consider a data set of cancer tumour tissue and normal tissue with DNA sequence, methylation and metabolomic data measured. Each of the three data types can be analysed for association with cancer. The resultant DNA sequence model, methylation model and metabolomics model can then be integrated to identify a meta-dimensional model. As the only variables that are incorporated into the integrative analysis are the ones that are detected in the data- type-specific modelling process, it is possible to miss some of the interactions between different data types if they do not have effects to identify within the data type. For example, if there is a pattern of methylation and another pattern of protein expression that are not associated with the outcome independently but only associated through their interaction, then their effects will be missed in model-based integration. Moreover, these forms of ensemble-based approaches are well known for overfitting(122). Therefore, model-based integration is particularly suitable if each genomic data type is extremely heterogeneous, such that combining the data matrix (concatenation-based integration) or performing data transformation to a common intermediate format (transformation-based integration) is not possible.

Future direction

There is a strong hope that genetics research can lead to personalized treatment in the future. Many analytical approaches have been developed to take advantage of the increasing breath and depth of the next generation genomics data. Despite the surge of new computational

methods development, it is safe to say that no method can comprehensively decipher all of the data on its own. The genetic architecture of complex traits is likely formed by the combination of individual and/or interactive effects from multiple sources of molecular factors. As an example, the NCI-DREAM drug sensitivity prediction challenge asked participants to use multi-omics data to predict drug sensitivity in breast cancer cell lines. The top performer used a Bayesian multitask multiple kernel learning method, which combines kernelized regression, multiview learning, multitask learning, and Bayesian inference(123). Thus, an ensemble approach where results from association analysis, epistasis, and system genomics study that are jointly analyzed would be immensely useful for the success of future genomics research (Figure 1-6).

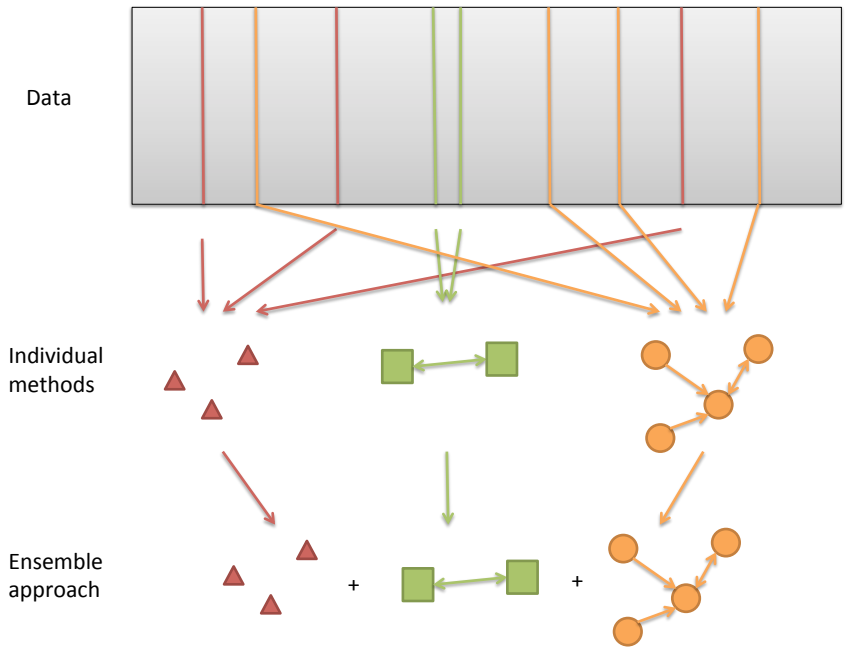


Figure 1-6:An ensemble approach to genomics research. Underneath a genomics data, there are potential causal individual genetic variants (red), epistatic interactions (green), and regulatory gene networks (orange). Individual methods are designed to capture one aspect of the total true signal. An ensemble approach can leverage the power of multiple approaches

Therefore, in relation to the above-mentioned challenges in this field, the main contributions of my dissertation are: 1. Evaluating and improving an existing tool (ATHENA) that has been developed to analyzed multi-omics data (Chapter 2). 2. Applying ATHENA to analyze multi-omics datasets in relation to a chemotherapeutic-induced drug response (Chapter 3). 3. Developing an ensemble analysis approach that combines multiple sources of information to understand a complex genetic trait (Chapter 4). 4. Developing a new Bayesian Network algorithm that improves upon the existing methods implemented in ATHENA and evaluating the algorithm on simulated as well as real biological data (Chapter 5). Finally, I will summarize and discuss future directions (Chapter 6).

Chapter 2*[3]

Evaluation of parameter contribution to neural network size and fitness in ATHENA for genetic analysis

Abstract

The vast amount of available genomics data provides us an unprecedented ability to survey the entire genome and search for the genetic determinants of complex diseases. Until now, Genome-wide association studies have been the predominant method to associate DNA variations to disease traits. GWAS have successfully uncovered many genetic variants associated with complex diseases when the effect loci are strongly associated with the trait. However, methods for studying interaction effects among multiple loci are still lacking. Established machine learning methods such as the grammatical evolution neural networks (GENN) can be adapted to help us uncover the missing interaction effects that are not captured by GWAS studies. We used an implementation of GENN distributed in the software package ATHENA (Analysis Tool for Heritable and Environmental Network Associations) to investigate the effects of multiple GENN parameters and data noise levels on model detection and network structure. We concluded that the models produced by GENN were greatly affected by algorithm parameters and data noise levels. We also produced complex, multi-layer networks that were not produced in the previous study. In summary, GENN can produce complex, multi-layered networks when the data require it for higher fitness and when the parameter settings allow for a wide search of the complex model space.

[3] Adapted from Li R, Holzinger ER, Dudek SM, Ritchie MD. Evaluation of Parameter Contribution to Neural Network Size and Fitness in ATHENA for Genetic Analysis. Genet Program Theory Pract XI. 2014

Introduction

With the rapid advancement of the genomics field, huge amounts of biological data are being generated. One beneficiary of the technology advancement is the exponential growth of genotype data, which measures single nucleotide variations across the genome. In fact, recent assays provide up to 5-million SNVs on every person for relatively low cost. The most widely used method for analyzing the genotyping data has been the genome-wide association study (GWAS). GWAS employs a statistical test at each individual locus across the genome to determine whether the locus is associated with the outcome (phenotype). GWAS has successfully revealed the genetic determinants for many complex human diseases(124). However, analysis of the genotyping data has been a great challenge because of the complex underlying relationships that exist in the data and many of the current GWAS analyses do not test the effect of interactions among multiple loci on the phenotype. Performing an exhaustive search of all possible combinations of loci is also not feasible with the current computational power, as the combinatorics for 5-million SNVs explodes. To mine the missing genetic variations of the data, we utilized and modified machine-learning techniques to search for interactions among genetic factors(125,126). We developed the ATHENA package, which utilizes grammatical evolution neural networks (GENN), to uncover the genetic network models underlying the disease or phenotype.

Previously, we reported that the final network models produced by grammatical evolution neural networks (GENN) with three variables were simple, 1-layer neural networks(127). Various parameter optimizations were attempted, but yielded little change in the depth of the neural networks. While 1-layer networks decrease the probability of finding false positives, it also raises the question of whether GENN can successfully model more complex interactions that are typically found in genomics data. Theoretically, GENN models should be quite capable of

building multilayer network structures, but empiric evidence based on (127) led to concern. Specifically, three hypotheses were considered: 1) the GENN grammar is biased toward 1-layer networks, 2) the parameters selected for the GENN analysis was not capable of building multi-layer networks, 3) the data simulation models did not warrant multi-layer networks.

The goal of this study is to examine these hypotheses and determine if one or more of them are true. To further examine these ideas, we assessed factors that may influence the network structure and true model detection. We used two types of simulated effect models to investigate the effect of grammar, population size, number of generations, and maximum grammar tree depth on network structure and detection power. Our findings suggest that the previously reported bias towards 1-layer network size was not due to a limitation of the GENN algorithm or an error in the program. Rather, combinations of grammar and maximum network depth affect network structure and detection power. Furthermore, the amount of noise (or non-informative variables) in the dataset also plays a role in networks identified by GENN. As such, it was a combination of parameter optimization and data simulation models that led to the results in (127). In this chapter, we will explain the compendium of simulations performed to address this question, demonstrate the results, and discuss future directions.

Grammatical Evolution Neural Networks

Neural networks (NNs) were designed to imitate neurons in the brain so that the networks can process information in parallel. NNs are widely used data mining methods in scientific research to detect underlying models in data to predict the desired outcome. NNs consist of nodes that can receive inputs from other nodes or from external independent variables. Each input is associated with a coefficient (or weight) which is multiplied and then the NN processes the weighted inputs through some activation function to produce an output signal (128). Generally,

the most popular method for training feed-forward multilayer NN is the gradient descent algorithm back-propagation (BPNN). BPNN randomly initializes weights associated with each node and gradually adjusts the weights with the goal of minimizing an error function (128). However, if the underlying fitness landscape is unknown, BPNN is an insufficient optimization method. In genomics studies of complex diseases, the fitness landscape is always unknown and complex. Thus, in order to avoid defining the fitness landscape a priori, a method has been proposed to apply genetic programming to optimize the structures and weights of the NN (129). A version of genetic programming neural networks (GPNN) has been implemented specifically for genetic association studies (130).

Grammatical evolution neural networks (GENN), an extension of GPNN, uses GE as the evolutionary algorithm. GE is a type of genetic programming (131,132) that uses Backus-Naur Form(BNF) grammar to create a model based on a genetic algorithm. The grammar translates an array of bits into a model, e.g. NN, based on its set of rules. At each generation, the finesses of the NNs are evaluated and the fittest networks are more likely to be selected to reproduce in the following generation. The genetic algorithm evolves for a specified number of generations and outputs the most optimal solution in the final generation (Figure 2-1). GENN optimizes variable selections, node coefficients, the number of hidden layers, and the number of nodes per layer simultaneously (116), so it can be applied to any data sets regardless of their underlying fitness landscape.

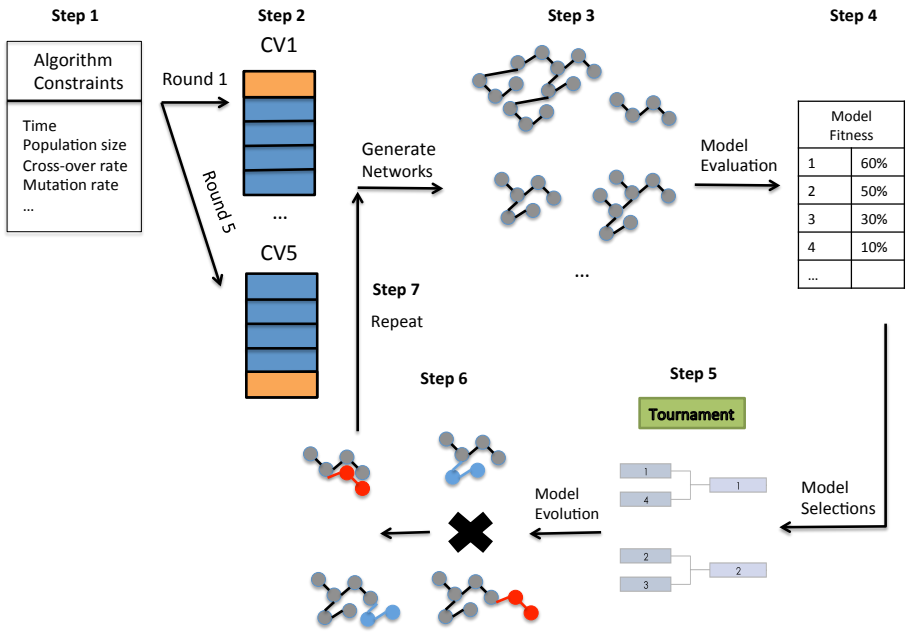


Figure 2-1: Schematic of GENN algorithm

Methods

Data Simulation

We simulated the XOR model because it is commonly used as a benchmark for neural network models; in addition, it also describes a potential type of epistasis interaction that may be observed in biological data. There are also several bioengineering and biochemical applications that have utilized XOR relationship in designing the experiment (133,134). In addition, in the XOR model, neither of the two predictor variables have a main effect; rather, interactions

between the variables determine the outcome. Thus, the model cannot be solved with 1-layer additive neural networks. Lastly, the result of experimenting with the XOR model can be extended to genotyping data because of the similarities of data formats. We simulated datasets under the XOR model with two possible outcomes (such as case and control). The model is detailed in Table 1. Random variables were generated using genomeSIMLA (135) such that both outcomes contain 1000 individuals (for a total of 2000 individuals in the dataset).

Table **2-1**: Description of the XOR model

Phenotype	VARIABLE1	VARIABLE2
0	0	0
1	1	0
1	0	1
0	1	1

In order to detect the effect of noise on network structure and detection power, we simulated two different types of datasets. The first type consisted of only two predictor variables with no noise variables (xor). The second type consisted of 100 predictor variables – two functional variables and 98 non-functional, or noise, variables (xor+noise). For both types of datasets, the two functional variables perfectly predicted the binary outcome using the XOR model shown in (Table 2-1). While this is unrealistic for complex trait epistasis in biology, it gives us a clean benchmark to explore these hypotheses.

ATHENA

The Analysis Tool for Heritable and Environmental Network Associations (ATHENA) is a versatile software package that includes various analysis techniques. One of the modeling

methods in ATHENA is GENN, which uses grammatical evolution to optimize artificial neural networks (ANNs). The GENN algorithm has previously been described in detail (127). The algorithm is briefly described as follows:

Step 1: The data is equally divided into 5 parts with 4/5 for training and 1/5 for testing. Different non-overlapping training and testing data are used for 5-fold cross validations.

Step 2: Under population size constraint, a random population of binary strings are generated to be ANNs using a Backus-Naur grammar. The ANNs are guaranteed to be functional per sensible initialization (131,132). During sensible initialization, an expression tree is created using the specified grammar by randomly selecting grammar rules to construct the tree. The software recursively checks the expression tree to make sure the selected rule would not make the expression tree exceed the maximum depth (Maxdepth) allowed. Half of the expression trees are built to the maximum grammar tree depth and the other half are built with a random depth less than the maximum depth. Finally, the expression trees are converted into corresponding codons. This step concurrently occurs at all demes (computer CPUs).

Step 3: All ANNs are evaluated with training data and the solutions with highest balanced accuracies are selected for crossover and reproduction. The new population is composed of mutated original solutions and new random solutions.

Step 4: Step 3 is repeated until it reaches the set generation number. Migrations of the best solutions occur at specified intervals between CPUs.

Step 5: The best solution at the final generation is tested on the testing data and the balanced accuracy is recorded.

Step 6: Steps 2-5 are repeated each time with a different set of training and testing data.

We ran the GENN algorithm within ATHENA using the parameters settings in (Table 2-2).

The parameters for generations, population size, and maxdepth are selected through empirical studies. Different grammar sets are designed to test the effects of different search spaces: *add* (linear), *bool2* (linear, logical), *bool*, (linear, logical, multiplicative). Each combination of the values for the varying parameters (36 unique combinations) was assessed using 10 xor datasets and 10 xor+noise datasets for a total of 720 experiments.

Table **2-2**: ATHENA GENN parameters

	Parameter	Values	Description
Constant parameters	Demes	16	Number of demes. Evolution occurs within each deme on an individual node with scheduled migration between demes
	Migrations	Every 25 generations	Networks with the highest fitness migrate between demes
	Probability of cross-over	0.9	Probability of exchanging genetic material between solutions
	Probability of mutation	0.01	Probability of random mutation
	Fitness Metric	Balanced accuracy	(sensitivity + specificity) / 2
	Selection	Roulette	The probability of selection is the proportional to its fitness
	Min genome size	20	Minimum size of binary genome at initialization
	Max genome size	15000	Maximum size of binary genome at initialization
	Backpropagation	Every 25 generations	Frequency of back propagation
	Cross-validation	5 fold	4/5th of the data was used as training and 1/5th of the data was used as testing
Varying parameters	Grammar	Bool	Operators: {+, -, *, /, OR, NOR, AND, NAND}
		Bool2	Operators: {+, -, OR, NOR, AND, NAND}
		Add	Operators: {+, -}
	Generations	300, 550, 800	Number of generations
	Deme Population Size	1000, 2000	The number of solutions per deme
	Maxdepth	9, 12	Maximum depth of grammar tree
Datasets	Dataset Type	xor, xor+noise	Type of dataset used to test the effect of the varying parameters.

Maxdepth would more likely to produce a more complex grammar tree, which in turn correlates with the higher complexity of the neural networks. With a Maxdepth of 9, the average

accuracies of models for both types of datasets were approximately 75%-85% with an average neural network depth of 1. When the Maxdepth was 12, the average accuracies for the xor datasets reached 100%, indicating that the maximum depth of the grammar tree was correlated with model detection. Higher Maxdepth also allowed production of multilayer neural networks (Figure 2-2).

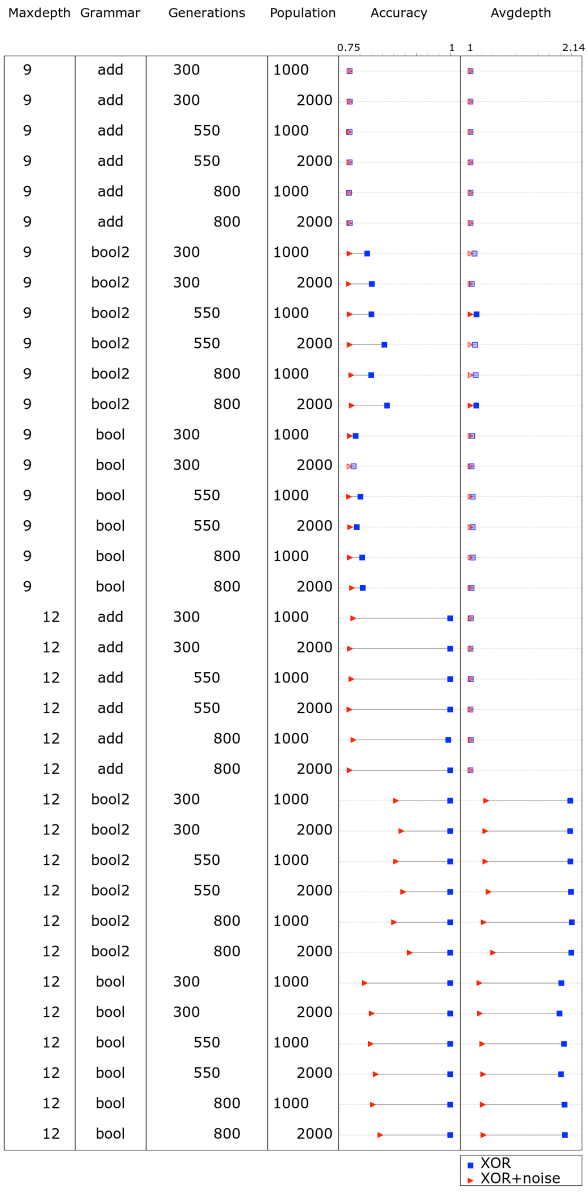


Figure 2-2:Average accuracy and average depth of neural networks

GENN grammar specifies how the nodes connect to each other, e.g. multiplying grammar multiplies the inputs and then feed it into the activation function. We compared three sets of grammars in this study: *Bool* {+, -, *, /, OR, NOR, AND, NAND}, *bool2* {+, -, OR, NOR, AND, NAND}, and *add* {+, -}. Upon random initializing of the population, each individual was generated based on its grammar set. Thus, individuals had access to different building blocks when they have different grammars. Models built with *bool2* grammar achieved higher accuracy than that of *bool* and *add* grammar. Using *bool2* or *bool* grammar in conjunction with Maxdepth of 12 allowed the production of multi-layer neural networks (Figure 2-2).

The noise level of the datasets fell into two distinct groups. One group of data set only contained two functional variables (xor), which can perfectly predict the simulated XOR model. The other group contained the same two variables and 98 randomly generated noise variables (xor+noise). The differences between the two groups are clear as shown in Figure 2-2. In all cases, data sets without noise signals achieved equal or higher prediction accuracies and generated more complex neural networks than xor+noise data sets.

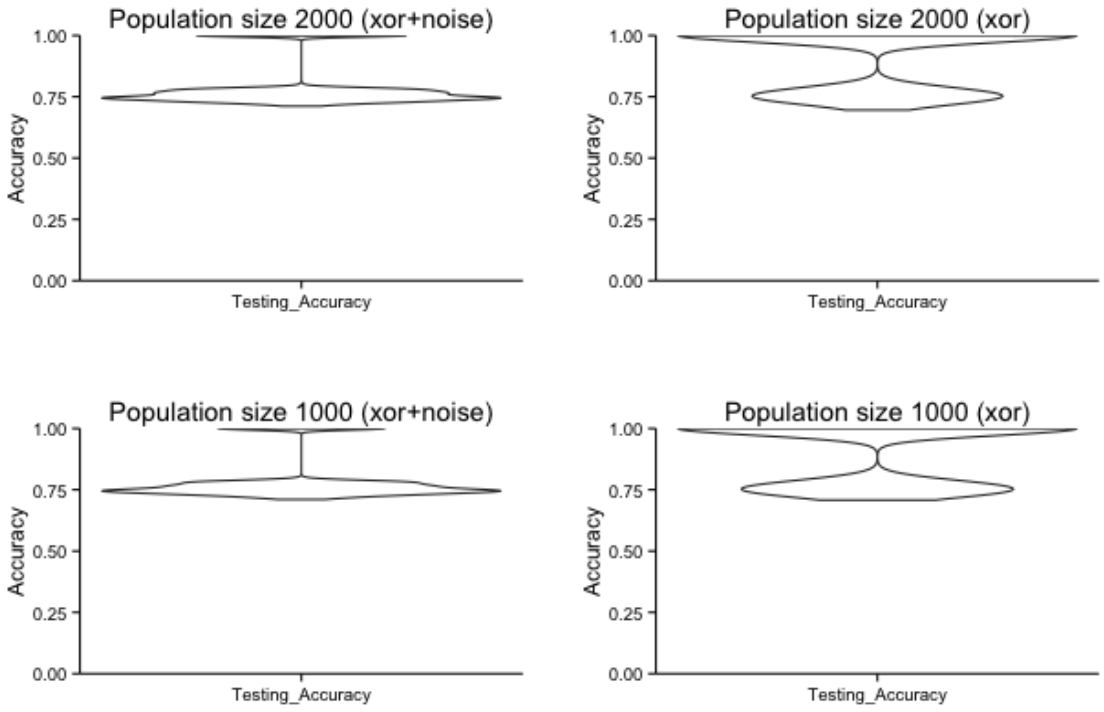


Figure 2-3:Comparison of population sizes on model detection

To further understand the effect of each GENN parameter on model detection, we compared the accuracies for each set of GENN parameters on two types of models. Xor data set produced more accurate models than that of xor+noise data sets regardless of the population size. Also, there was no clear difference between population size of 1000 and 2000 on model detection (Figure 2-3). Similarly with population size, longer generations of evolution did not affect model detection. The prediction accuracies were only affected by data noise level (Figure 2-4). The maximum depth of grammar tree significantly improved model detection for both types of data. In particular, under Maxdepth of 12, xor data achieved perfect predictability in almost all data sets (Figure 2-5). In the order of *add* grammar, *bool* grammar, and *bool2* grammar, model detection has shown an improvement in both data types. This is evident by the increasingly higher number of models at 100% accuracy in the above order (Figure 2-6).

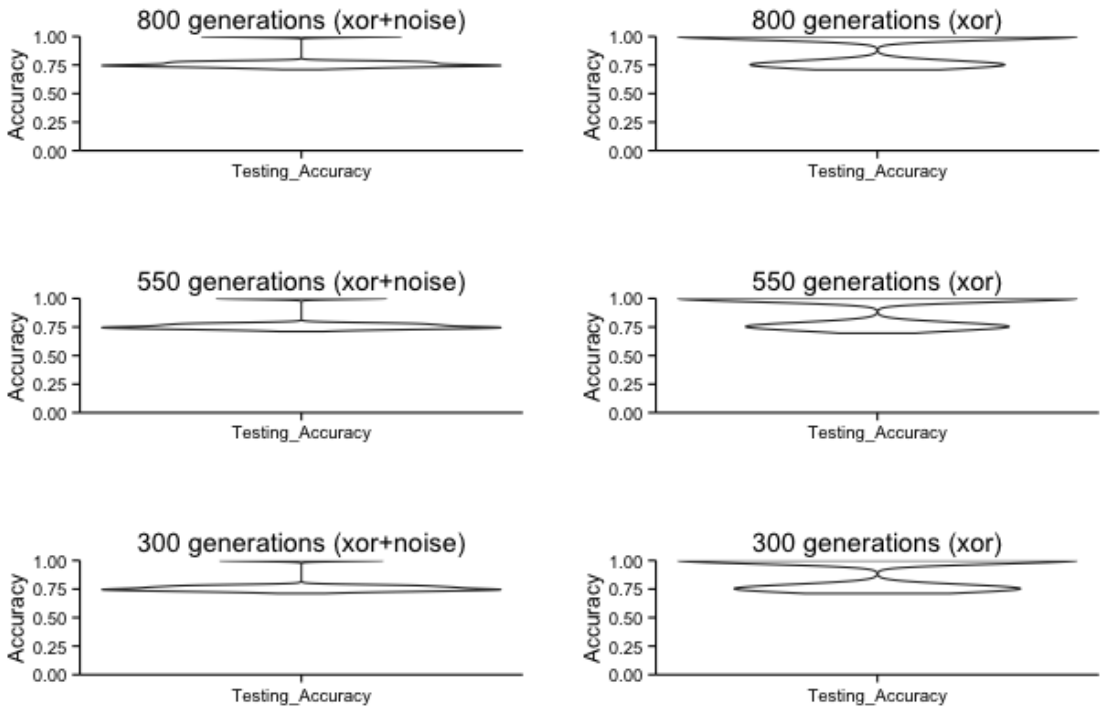


Figure 2-4:Comparison of generations on model detection

From these results, we drew three main conclusions. First, the addition of noise variables results in overall lower testing accuracy, which is not surprising and in some ways serves as a positive control experiment. Second, higher maximum grammar tree depth resulted in more predictive models for both dataset types. Third, grammar types have an effect on accurate modeling for both dataset types.

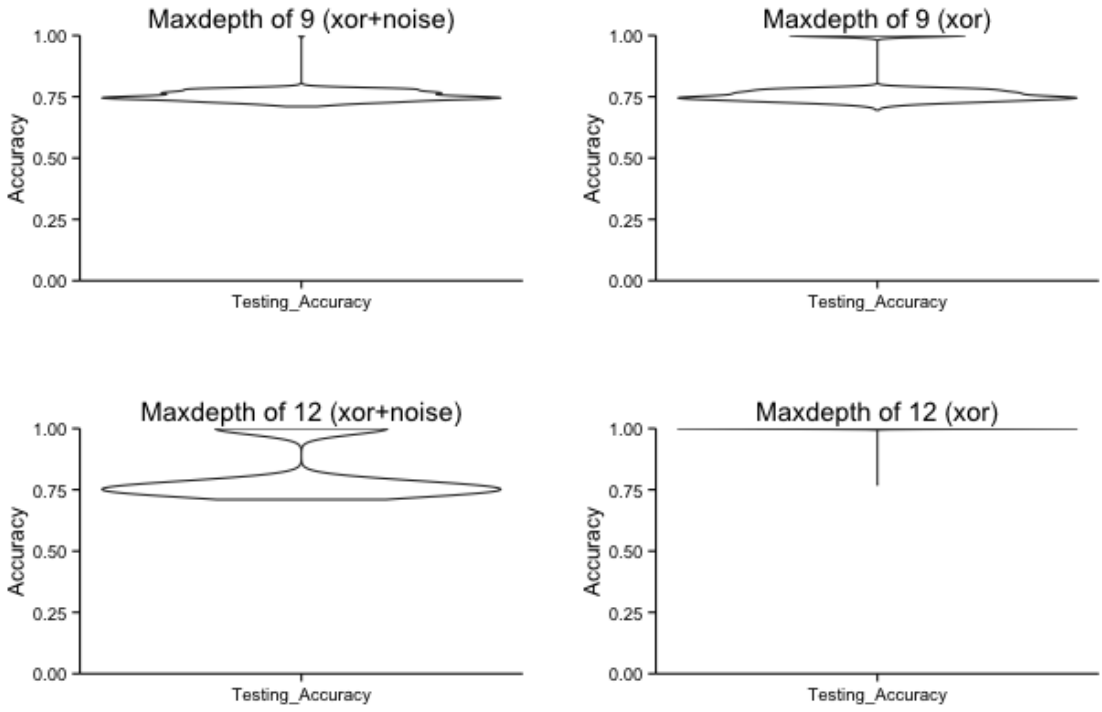


Figure 2-5:Comparison of maximum depth of grammar tree on model detection

To further investigate the factors that influence network size, we studied the effect of each parameter across generations. The average depths of the neural network for xor datasets were higher than that of xor+noise datasets as shown by darker shades at the top portion of the panels. However, there was no clear difference of network depth between the two population sizes (Figure 2-7). Similarly, xor dataset produced more complex neural networks than xor+noise dataset regardless of the number of generations. The depth of the network stayed relatively flat after the initial oscilla- tion (Figure 2-8). The maximum depth of grammar tree had a clear effect on the network structure. With a grammar Maxdepth of 9, GENN produced mostly single layer neural net- works. Increasing the grammar Maxdepth to 12 resulted in more complex neural networks (Figure 2-9). With *add* grammar, neither data set produced complex neural networks. *Add* grammar also differs from the other two grammars in that the depth drops very quickly and

never recovered. Both *bool* and *bool2* grammar were able to produce multilayer neural networks (Figure 2-10).

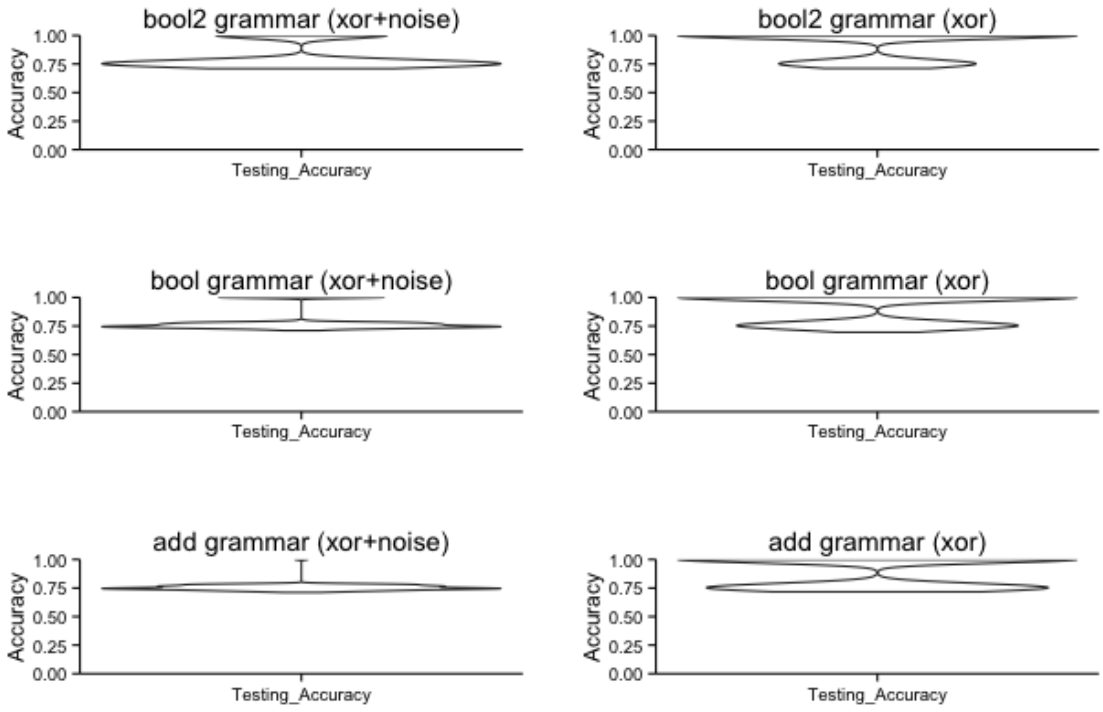


Figure 2-6:Comparison of grammars on model detection

As previously reported, the average depth of the networks decreased at the beginning of the evolution. Except for experiments where the maximum depth was 9 or with *add* grammar, the average neural network depth increased after the initial drop. Datasets without noise variables generally had higher average network depth, indicating that the noise level was a strong determinant of network depth. Higher maximum depth, *bool* and *bool2* grammar were also correlated with higher average network depths.

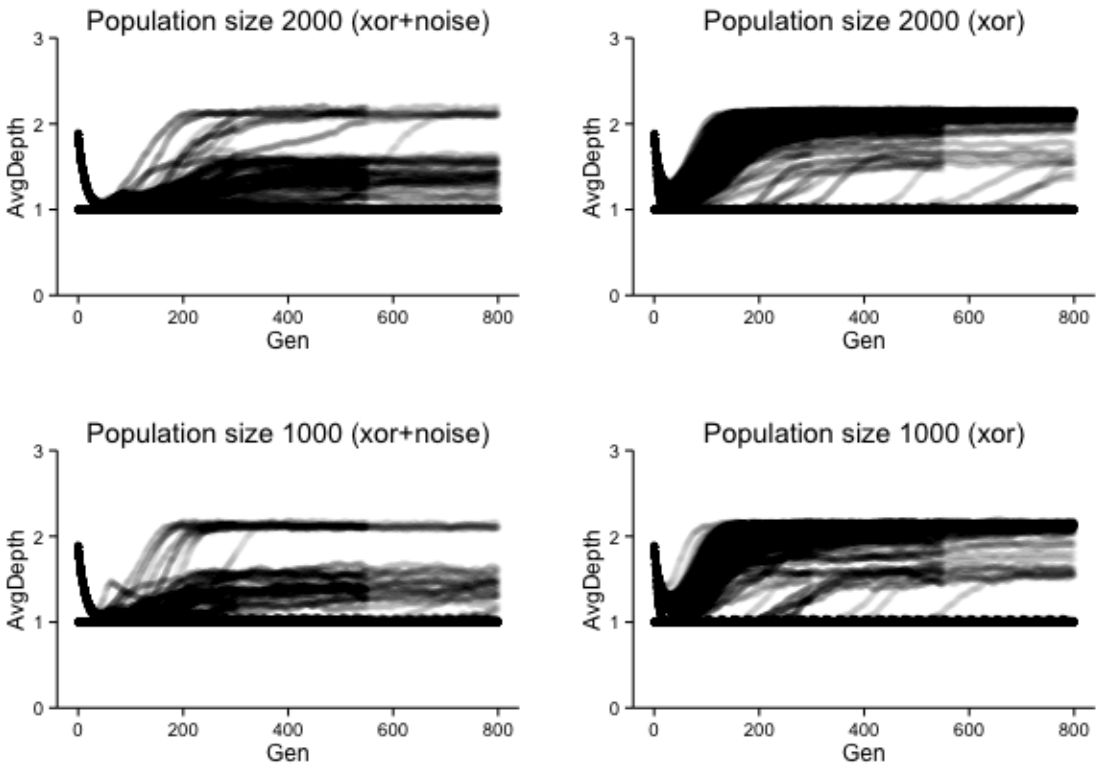


Figure 2-7:Comparison of population sizes on average network sizes

Discussion

In the previous report, we observed that GENN preferentially produced 1-layer neural networks based on the three variables simulated in the model. The reason for producing 1-layer networks remained unclear as to whether it was due to the inherited characteristics of GENN, the GENN parameter settings, or the complexity of the simulated datasets. While simple 1-layer networks offer the advantage of better interpretability, it might not be sufficient to model complex biological processes because there are non-additive epistasis effects in many biological

systems (136,137). To be sure that GENN can produce multilayer network if it is required by data, we experimented with GENN parameters to understand under which parameter combinations can we obtain multilayer networks. We believe that parameter settings should not favor either 1-layer networks or multilayer networks because the evolution process will determine the most fit model structures. However, when the parameters settings are limiting or unnecessarily expanding the search space, it will have a great impact on the generated models. Most importantly, our goal is to apply GENN on biological data with unknown underlying relationships. Through this experimentation, we can eliminate biases due to parameter settings and have more confidence in our result. In this study, we simulated two types of xor models: xor and xor+noise to test the effect of various GENN parameters on model detection and network structures. These results have shown that the noise level in a dataset was the most significant determinant of model detection and network depth. In noisy data, GENN has to perform

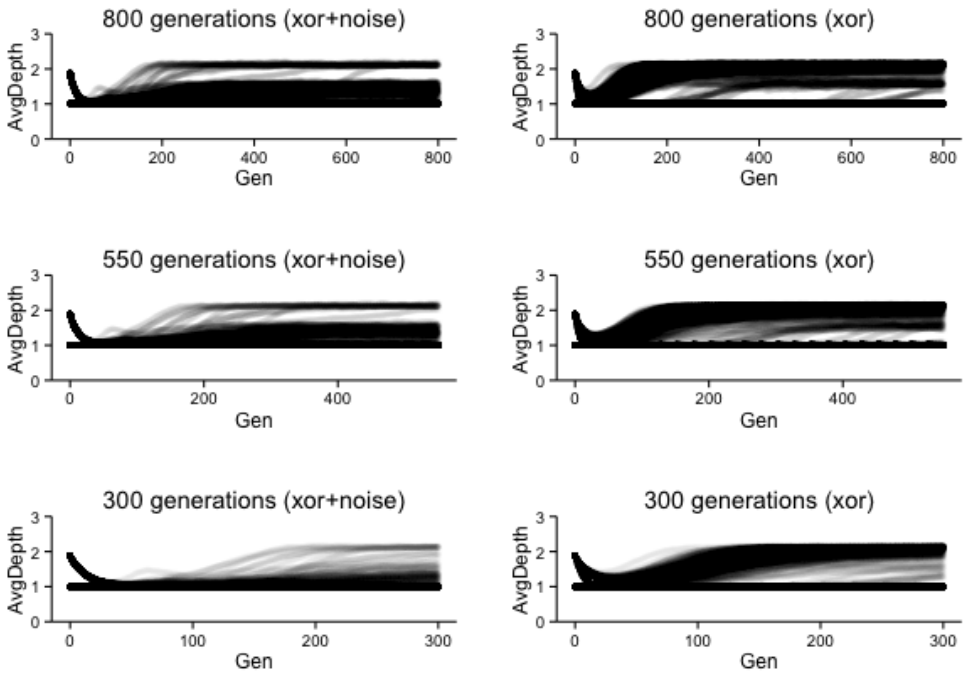


Figure 2-8:Comparison of generations on average network sizes

both variable selection and network modeling as opposed to only network modeling in the noiseless data. The existence of the noise variables decreased the probability of true variable detection, which in turn made it more difficult to maintain multilayer networks through evolution because most of the variables are non-informative. Maximum depth of the grammar tree was also an important factor in model detection. In our simulations, lower Maxdepth limited the possibility of producing multilayer networks, which in turn limited the ability of GENN to produce multilayer network structures necessary to detect interaction effects. However, the one-layer model still correctly identified one of the simulated variables in most simulation settings, which resulted in higher than 50% accuracy. Lastly, different grammars also affected model detection and network depth. *Bool* and *bool2* grammar outperformed *add* grammar because they included more operators that can detect interaction relationships among variables. However, having more operators did not produce the best model as evidenced by the better modeling with *bool2* grammar compared to *bool* grammar. If the additional operators did not add more informative variable relationships, it will unnecessarily increase the search space, which could explain the lower accuracies produced by *bool* grammar. In conclusion, we have determined that GENN does not inherently produce 1-layer networks. The combination of noise level, maximum grammar tree depth and grammar determines the model detection and network sizes in GENN. Due to the simplicity of the XOR model, the highest average NNs depth is only around 2. In future experimentation, we can simulate more complex models and use higher Maxdepth value. Software optimization might be needed because higher Maxdepth significantly requires more computational resources during evolution process.

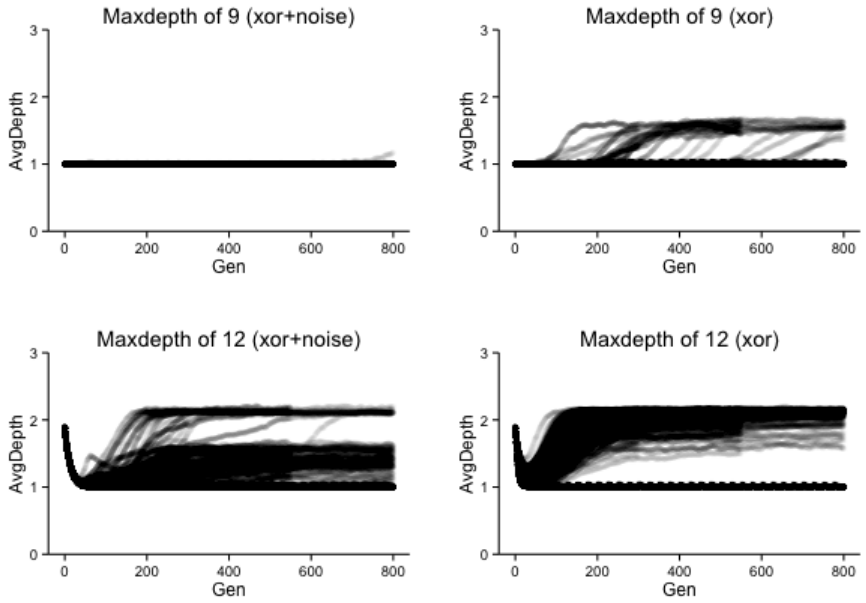


Figure 2-9:Comparison of maximum depth of the grammar tree on average network sizes

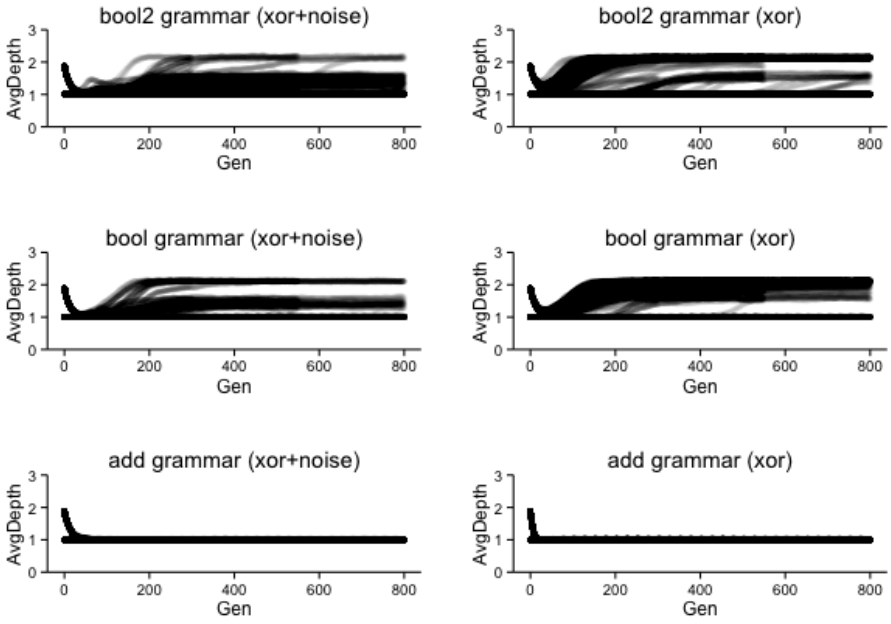


Figure 2-10:Comparison of grammars on average network sizes

Chapter 3*[4]

An integrated analysis of genome-wide DNA methylation and genetic variants underlying etoposide-induced cytotoxicity in European and African populations

Abstract

Genetic variations among individuals account for a large portion of variability in drug response. The underlying mechanism of the variability is still not known, but it is expected to comprise of a wide range of genetic factors that interact and communicate with each other. Here, we present an integrated genome-wide approach to uncover the interactions among genetic factors that can model some of the inter-individual variation in drug response. The International HapMap consortium generated genotyping data on human lymphoblastoid cell lines of (Center d'Etude du Polymorphisme Humain population - CEU) European descent and (Yoruba population - YRI) African descent. Using genome-wide analysis, Huang et al. identified SNPs that are associated with etoposide, a chemotherapeutic drug, response on the cell lines. Using the same lymphoblastoid cell lines, Fraser et al. generated genome-wide methylation profiles for gene promoter regions. We evaluated associations between candidate SNPs generated by Huang et al and genome-wide methylation sites. The analysis identified a set of methylation sites that are associated with etoposide related SNPs. Using the set of methylation sites and the candidate SNPs, we built an integrated model for etoposide response observed in CEU and YRI cell lines. This integrated method can be extended to combine any number of genomics data types to model many phenotypes of interest.

[4] Adapted from Li R, Kim D, Dudek SM, Ritchie MD. An integrated analysis of genome-wide DNA methylation and genetic variants underlying etoposide-induced cytotoxicity in European and African populations. Springer Berlin Heidelberg; 2014

Introduction

Genome-wide analysis is a step forward from candidate gene based approaches because it reduces biases associated with candidates' selections. While candidate gene approaches have successfully identified genes involved in cellular mechanisms of drugs, they failed to uncover interactive relationships among the genetic factors that may be explaining much of the variations in drug effects. The cellular susceptibility of the drug is potentially affected by multiple genetic components through non-linear interactions among the components. However, due to the exponential increases of computational calculations when modeling interactive relationships, most research have been focused on finding linear models associated with drug response(20,28,93,138,139). To uncover the unsolved variances, we propose an integrated genome-wide analysis that identifies interactions among genetic factors from multiple types of genomic data to model the drug response.

The International HapMap Consortium genotyped cell lines of various population groups including trios of European descent (CEU) and Yoruba descent (YRI)(140). Because these cell lines are publicly available, they have also been used to study methylation patterns at gene promoter regions(141). Together, genotype variations and methylation levels enable us to study the relationship between these genetic components and drug responses. Previously, through genome-wide analysis, Huang et al. have identified a set of genetic variants that are associated with chemotherapeutic drug induced cytotoxicity in CEU and YRI cell lines, respectively(20). We used the set of SNPs as dependent variables and methylation levels as independent variables and applied regression models for each unique SNP-methylation combination. We identified SNPs that are correlated with methylation levels, or methylation quantitative trait loci (mQTLs), across the genome using publicly available genome-wide methylation data, generated on the same cell lines(141). Together, using the genetic variants and correlated methylation levels at gene

promoters, we found interactive genetic models that can explain a portion of variability in chemotherapeutic drug response in CEU and YRI cell lines. The integrative models achieved higher explanatory power of drug response in these cell lines than previously published linear models.

Etoposide is a topoisomerase II inhibitor(142) and is used in treatment of cancers including testicular cancer, lung cancer, germinal cancer, endometrial carcinoma, and Kaposi's sarcoma. Treatment with etoposide can lead to severe side effects such as fatigue, bone marrow suppression, diarrhea and acute promyelocytic leukemia(143–145). Thus, our goal is to identify SNPs and methylation interactions that can best model the differential etoposide responses in CEU and YRI cell lines. This result paves the way for better understanding of genetic components involved in drug responses, which is a necessary step towards personalized drug prescription for cancer patients.

Methods

Genetic variants correlated with etoposide IC_{50}

CEU and YRI population, respectively. The inhibition of cell line growth is measured as IC_{50}, which is the drug concentration required to stop cell growth by 50%. The method for identifying the SNPs is as follows. A total of 87 and 89 cell lines from HapMap CEU and YRI populations, respectively, were exposed to increasing concentrations of etoposide. SNP genotypes were obtained from the International Hapmap website (HapMap.org) (release 21). Genotyping errors and extreme outliers were removed and only SNPs within 10kb up or downstream of a gene were retained. Quantitative transmission disequilibrium test (QTDT) analysis was performed on Box-Cox transformed IC_{50} values and filtered SNPs with sex as a covariate. Using

$p < 0.0001$ as threshold for significance, 122 and 51 SNPs were significantly associated with etoposide IC_{50} in CEU and YRI, respectively(20). The associated SNPs were used for subsequent downstream analysis.

Candidate SNPs and methylation levels association

Gene promoter regions methylation data were generated by Fraser et al.(141). The data was downloaded from Gene Expression Omnibus database, accession number [GSE27146]. A total of 84 CEU and all (89) YRI cell lines that were tested for etoposide response were used to measure promoter region methylation levels. Over all, methylation levels at 27,578 CpG sites near transcription start sites were measured using the quantitative BeadChip assay (Illumina, San Diego, CA, USA). Several steps, which are described in detail in Fraser et al.(141). were taken to account for the background noise. Briefly, first, the average background intensity was subtracted from the raw intensity to adjust for sample variations. Then, to minimize batch effects of different arrays, background adjusted raw data were quantile normalized(141).

Regression models were used to test for possible candidate SNPs and methylation level association. Significant CEU and YRI SNPs were tested for their association with methylation in the same respective population. To remove the effect of gender, sex was used as a covariate in the regression model. Using a p-value cut off of 0.0003, 1109 methylation-SNP pairs were significantly associated for CEU and 270 methylation-SNP pairs were significant for YRI, of which 385 and 176 methylation sites were unique, respectively.

Interactive model of SNPs and methylation levels to predict etoposide IC_{50}

We used ATHENA as described in Chapter 2. Additional steps taken here are:

Step 7: SNPs and methylation probes that appear in at least 3 out 5 cross validation models are saved as consistent variables

Step 8: All consistent variables will be modeled over the entire dataset and results in a final model

The fitness of the model aims to measure how well the model can explain the etoposide drug response, a continuous value. We used R-squared as our fitness metric to represent the percentage of drug response explained by the model. The drug response predicted by the model is scaled using the sigmoid function so that the value is between 0 and 1. As a result, we also scaled the original drug response to be between 0 and 1 using min-max scaling, where

$$Normalized\ D_i = \frac{D_i - \min(D)}{\max(D) - \min(D)} (1)$$

And the R^2 is calculated as:

$$R^2 = \frac{\sum_i^n (D_{predict\ i} - \overline{D})^2}{\sum_i^n (D_i - \overline{D})^2} (2)$$

$$\mathrm{D}_i = the\ ith\ value\ of\ drug\ response$$

The final model is an artificial neural network (ANN). ANNs are widely used in data mining field to predict desired outcome. ANNs consist of nodes of input and an output. Each input node is associated with a weight and the weight is generally determined through back-propagation(128). ANNs can have multiple layers, which make it possible for input nodes to have interactive relationships among themselves. Traditionally, the structure of the network and the input variables need to be defined before optimizing the network. However, this is not the case

for genetic analysis because neither the fitness landscape nor the correct variables are known. Evolutionary algorithms can eliminate this deficiency as the network structure and correct variables are evolved automatically, driven by the data(129).

If the variables in the model contain missing values, the samples contain missing values will be removed for that evaluation. To eliminate sample loss, missing values in the SNP genotype data were replaced with 0, making the particular SNP homozygous for its corresponding sample. For 84 CEU samples, there were 176 missing values within 122 SNPs; and for 89 YRI samples there were 86 missing values within 51 SNPs. The replacement represents less than 2% of the data.

Result

To ensure validity of the result, each analysis was repeated with a different random seed and GENN population size (Table 3-1).

Table **3-1**: GENN parameter settings.

Parameter	Sample analysis	
Number of processors	16	16
Population size/ processor	20000	3000
Number of generations	2000	2000
Number of migrations	40	40
Crossover probability	0.9	0.9
Mutation probability	0.01	0.01
Random seed	Random1 /Random 2	Random 3/ Random 4

Using GENN to identify the most informative SNPs to predict etoposide response, the analysis resulted in several SNPs that consistently appeared in different cross validations with different random seeds and population sizes (Table 3-2). Each cross validation returned a SNP interaction model that was found to be the best for a subset of cell lines. SNPs that appeared in three out of five cross validations were considered to be interesting.

Table **3-2**: Associated SNPs and methylation in the best model

Probe name	Population	Chromosome	Host gene ID
rs647955	YRI	Chr1	C1QB
rs2605593	YRI	Chr11	C11orf75
rs6944165	YRI	Chr7	LOC647017
rs16905691	YRI	Chr10	PCDH15
cg21931212	YRI	Chr12	C12orf57
rs403029	CEU	Chr10	GATA3
rs1884679	CEU	Chr14	SLC24A4
rs2607839	CEU	Chr10	GRID1
rs9299075	CEU	Chr9	PTPRD

For YRI, the interesting SNPs were rs4770877, rs9730073, rs16905691, rs12113878, and rs9507577. We then integrated SNPs and methylation data so that we could explore interactions between SNPs and methylation levels. Using the same criteria, we identified SNPs rs647955, rs2605593, rs6944165, rs16905691 and methylation probe cg21931212 were consistently associated with etoposide. Sex was included as an input variable, but it was not incorporated in the fittest model. When we analyzed all of the consistent SNPs and methylation probes together, rs647955, rs2605593, rs9730073, rs12113878, rs16905691, and cg21931212 were selected in the final model. The r-square for the final model was $R^2 = 53.75\%$, indicating that the model can explain around 54% etoposide IC_{50} variations in the YRI population. Our model outperformed previous linear SNPs model, which attained a R^2 around 40%(20). Figure 3-1 shows the interaction model between SNPs and methylation for YRI population.

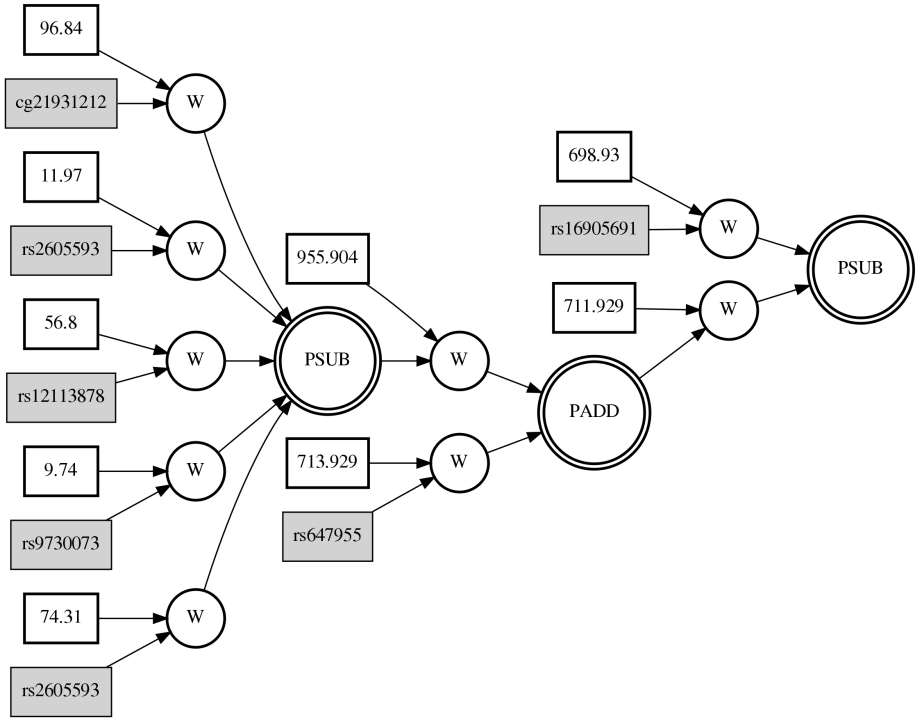


Figure 3-1: Final model of SNPs and methylation interactions to predict etoposide IC_{50} in YRI (w: multiplication between constant and variable, PADD: additive node, PSUB: subtractive node)

We did not identify any consistent SNPs and methylation interactions in CEU population. For SNPs only interactions, rs403029, rs1884679, rs2607839, and rs9299075 showed consistent association with etoposide IC_{50}. We again used all of the consistent variables as input to train the final model and the final model included all four SNPs. The r-square for the final model was R^2 = 46.16%, indicating that the model can explain 46% etoposide IC_{50} variations in the CEU population. Figure 3-2 shows the SNPs interaction model for etoposide IC_{50} in CEU.

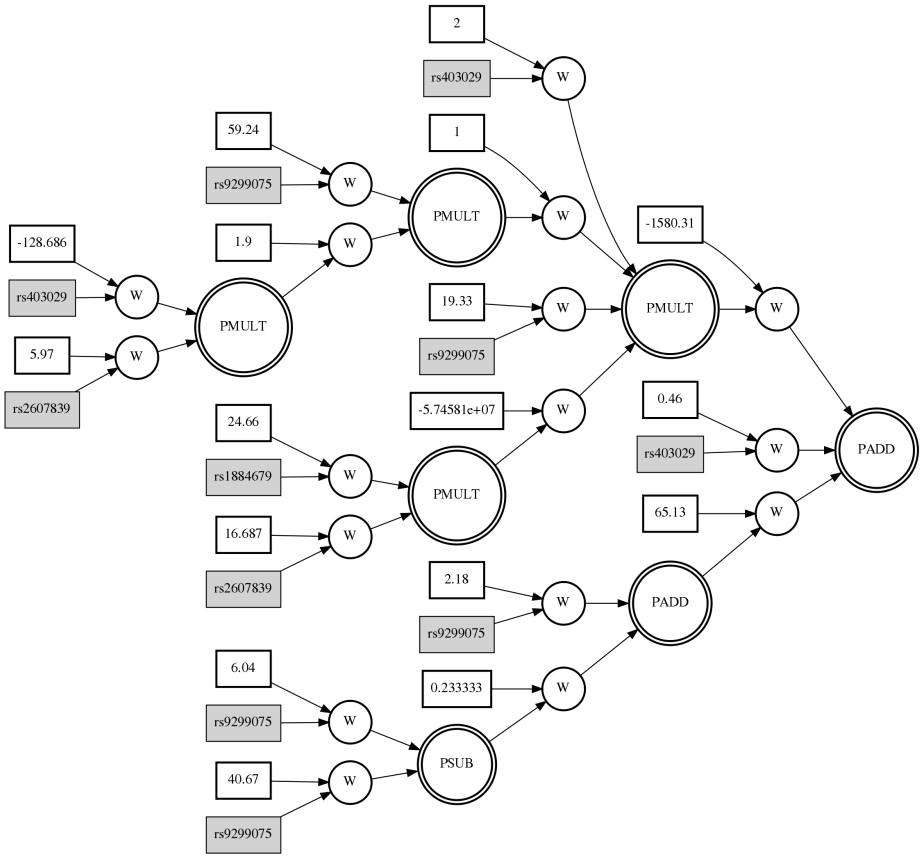


Figure 3-2: Final model of SNPs interactions to predict etoposide IC_{50} in CEU (w: multiplication between constant and variable, PADD: additive node, PSUB: subtractive node, PMULT: multiplicative node)

Conclusion and Discussion

In this study, we explored interaction relationships among SNPs and between SNPs and methylation levels to model etoposide IC_{50} on HapMap CEU and YRI cell lines. The integrated genome-wide approach demonstrated the ability to combine multiple types of genomics data and identify interactive relationships within and between data sources. Due to the small sample size in this study, the results should be viewed as a proof-of-concept or pilot project for this type of data integration. Future directions will evaluate alternative data fusion techniques with ATHENA on multi-omics data to build meta-dimensional models.

Etoposide is a widely used cancer drug for testicular cancer, lung cancer, germinal cancer, endometrial carcinoma, and Kaposi's sarcoma. However, the drug also has severe side effects for the patients(143–145). Better understanding of the mechanism of the drug is a crucial step towards personalized prescription of the drug based on patients' genetic makeup. Genetic variations are the most fundamental and the most widely studied genetic factor in relation to the drug response, as Huang et al. previously reported that a group of SNPs were correlated with etoposide IC_{50} in CEU and YRI population. Using the correlated SNPs, they built a linear additive model to explain the variability of IC_{50} in the two populations. Stemming from their multi-genic model, it is logical to hypothesize that etoposide's cellular mechanism could also be comprised of interactive relationship among SNPs. Recent study also suggested that phenotype associated SNPs tend to fall into function-associated regions(146). Methylation pattern is an important marker for DNA regulatory functions and this led us to explore the interactive relationships between SNPs and methylation levels. Using SNPs and correlated methylation levels, we were able to identify several SNPs and methylation sites that consistently appeared in our models. We applied GENN on these consistent variables to build a final model for each population. For YRI population, we built an interactive model between SNPs and methylation and achieved a R^2 of 54%, exceeding models that only examined linear additive relationships between SNPs. For CEU population, we only identified consistent interactive SNPs variables. Our interaction model with four SNPs resulted in a R^2 of 46%, slightly lower than previously reported R^2 of 55%; potentially due to less number of variables in our model. Based on these results, our genome-wide integrative analysis identified novel interaction relationships between SNPs and methylation sites. This approach can be extended to integrate any number of genomics data to predict or classify a wide range of phenotypes of interest.

Modeling genetic interactions is a complex task, especially when there are a large number of variables. Models produced by GENN are dependent on parameter settings, but they generally

contained around ten variables. Evaluating all possible combinations of interactions is impossible given the current computational power, so GENN uses a guided random search to make the search more feasible. In addition, there is variability between samples partitioned in each cross validation. As a result, the fittest model in each cross validations may suffer from inadequate modeling and may not be applicable to other subsets of data. Thus, we utilized a strict requirement to minimize this bias by only keeping variables that appeared in at least 3 out of 5 cross validations, ensuring that the true signal is strong and applicable to different subsets of the data. The trade off of this approach is increased number of false negatives. This is evident when we evaluated SNPs and methylation interactions in CEU population. Because there is a higher number of SNPs and correlated methylation probes in CEU compared to YRI population, the search space exponentially increased. As a result, when modeling interactions between SNPs and methylation, there were many SNPs and methylation probes that appeared in 2 out of 5 cross validations, but none appeared in at least 3. We could potentially miss some true signals by employing a strict consistency requirement, but we are also more confident about our true signals. For YRI population, our final interactive model of SNPs and methylation resulted in a R^2 of 54%, exceeding the previously reported 40% identified in linear model(20).

One should be aware that the interactive relationships produced by GENN are only statistical relationships. Our model uncovered potential genetic variants and methylation sites that could be further validated by functional studies. Some of the genetic variants are unknown but others are found to be relevant through literature search. Genetic variant of rs647955 is located in the C1QB gene. C1QB is known to be involved in systemic lupus erythematosus, an autoimmune disease(147). The function of variant rs9730073 is not known, but it was also selected by Huang et al. as one of the final four SNPs used in their linear model for YRI. SNP rs12113878 is located within KLRG2 gene. KLRG2 gene has been found to be associated with prostate cancer aggressiveness and is expressed on subsets of NK/T cells(148). Interestingly, rs16905691 is

associated with the PCDH15 gene, which is also expressed on NK/T cells. NK/T cells are known to play a key role in defense against tumor development(149). Methylation probe cg21931212 lies in C12orf57 gene, which has no known functions. However, recent genome-wide studies have identified the gene to be associated with brain and vision development(150,151). CEU model SNPs rs9299075 and rs1884679 are located in genes PTPRD and SLC24A4, respectively and both are associated with tumor suppression and identification(152,153). Many of our modeled genetic variants are associated with cancer and development, which is related to etoposide's drug mechanism. Further study is needed to confirm these relationships.

There are limitations to this study that warrants more future studies on the subject. The study separately analyzed etoposide's response on CEU and YRI cell lines. Previous report has shown that the two cell lines behaved similarly under etoposide(20). Future analysis plan should include combining the two populations with race adjustment in order to find generalizable models across different cell lines, which will also greatly increase sample size and thus statistical power. It is also known that some of the model SNPs and methylations have linkage disequilibrium (LD) or correlation with each other. Exploring these related genetic factors could reveal more insights on etoposide response. eQTL analysis using gene expressions generated on the HapMap cell lines has also shown significant associations with various chemotherapeutic drug responses(20,139). It would be interesting to integrate gene expression data as well as methylation data to model the etoposide response in future analysis. Lastly, the method for imputing missing SNPs could incorporate LD information in the future. However, the result in this study should not be affected because less than 2% of the data was missing.

The ultimate goal of this study is to identify potential models that can model etoposide drug or toxicity response in order to better prescribe treatments to patients and improve clinical knowledge of the treatment. The integrated analysis used in this study has shown that it can identify novel interactions among genetic factors. This approach can also be applied to uncover

genetic factors underlying a wide range of other phenotype and diseases. For example, we can use integration analysis to see if similar genetic models are underlying different chemotherapeutic drugs. In the following chapter, the integration analysis was applied to five chemotherapeutic drugs in two populations in order to comparatively analyze the genetic factors influencing cytotoxic response.

Chapter 4*[5]

Integration of genetic and functional genomics data to uncover chemotherapeutic induced cytotoxicity

Abstract

Identifying genetic variants associated with chemotherapeutic induced toxicity is an important step towards personalized treatment of cancer patients. However, annotating and interpreting the associated genetic variants remain challenging because each associated variant is a surrogate for many other variants in the same region. The issue is further complicated when investigating patterns of associated variants with multiple drugs. In this chapter, we used biological knowledge to annotate and compare genetic variants associated with cellular sensitivity to mechanistically distinct chemotherapeutic drugs, including platinating agents (cisplatin, carboplatin), capecitabine, cytarabine, and paclitaxel. Top SNPs from genome wide association studies of cellular sensitivity to each drug in lymphoblastoid cell lines derived from populations of European (CEU) and African (YRI) descent were analyzed for their enrichment in biological pathways and processes. We annotated genetic variants using higher-level biological annotations in efforts to group variants into more interpretable biological modules. Using the higher-level annotations, we observed distinct biological modules associated with cell line populations as well as classes of chemotherapeutic drugs. We also integrated genetic variants and gene expression variables to build predictive models for chemotherapeutic drug cytotoxicity and prioritized the network models based on the enrichment of DNA regulatory data. Several biological annotations, often encompassing different SNPs, were replicated in independent datasets. By using biological

[5] This chapter has been submitted for journal publication

knowledge and DNA regulatory information, we propose a novel approach for jointly analyzing genetic variants associated with multiple chemotherapeutic drugs.

Introduction

A better understanding of genetic variation contributing to cellular sensitivity to chemotherapeutic drugs can lead to more precise and personalized treatment of cancer patients(154). Lymphoblastoid cell lines (LCLs) have been established as a model system to study the genetic components of drug-induced cytotoxicity by measuring cell growth inhibition following drug exposure(155). Previous genome-wide association analyses (GWAS) have identified numerous genetic variants and gene expression variables associated with drug cytotoxicity(20,28,29,93). However, a comprehensive study of multiple drugs in different populations can reveal new insights into the genetic susceptibility of cytotoxicity.

We studied genetic factors associated with cytotoxicity of five mechanistically distinct chemotherapeutic drugs: cisplatin, carboplatin, capecitabine, cytarabine, and paclitaxel. Cytotoxicity was measured for all drugs in two HapMap populations: Utah Residents with European ancestry (CEU) and African individuals from Yoruba in Ibadan, Nigeria (YRI). Platinum-based compounds, including cisplatin and carboplatin are the most widely applied group of cytotoxic drugs worldwide, used to treat head and neck, testicular, lung, endometrial and ovarian cancers(156–158). Capecitabine is mainly used to treat colorectal and breast cancers(159). Patients with acute myeloid leukemia have long been treated with cytarabine(160). Paclitaxel is commonly used for the treatment of lung, breast, and ovarian cancers(161). Previous studies have shown that drugs in the same class have common genetic loci associated with drug induced cytotoxicity, for example, cisplatin and carboplatin(29). An individual's ancestral background has also been linked to differential risks for cytotoxicity(162). Thus, a more

comprehensive understanding of the distinct and shared genetic components associated with cytotoxicity between drugs and populations would be valuable to identify new treatment options.

However, a molecular understanding of individual genetic variations is challenging because there are a large number of genetic variations that can be associated with drug cytotoxicity and each variant is a surrogate for many other variants in the same region. To address these issues, we evaluated genetic variants using higher-level biological annotations in efforts to group variants into more interpretable biological modules. Comparing CEU to YRI, we found population specific annotations for each drug. Within individual populations, we observed drugs that treat similar types of cancers are enriched for the same biological annotations. In some cases, we identified similar biological annotations across CEU and YRI, as well as across multiple drugs.

Previous studies relied on GWAS to identify genetic variants that have the strongest independent genetic effects on drug-induced cytotoxicity and incorporated gene expression levels through studies of expression quantitative trait loci (eQTL) analysis(163). This work led to the important observation that pharmacological GWAS SNPs are enriched in eQTLs for many cytotoxic drugs(139). While the eQTL method can capture a linear relationship between SNPs and gene expression, it omits the possibility that interactions among SNPs or gene expression could also play a crucial role in drug cytotoxicity. To identify these non-linear interactions, we applied a grammatical evolution neural network (GENN) algorithm to build interaction networks consisting of SNPs and gene expression variables. Although the identification of associated SNPs and gene expression variables is an important first step in understanding drug cytotoxicity, a challenge remains on how to interpret the functional relevance of the interaction models. It has been shown that many regulatory elements can aid in identifying important functional SNPs(124,164). To this end, we used DNAseI and genome segmentation data published by the ENCODE consortium to prioritize the network models. Our studies suggest that combining

genetic and functional genomics information could be a useful approach for interpreting genetic factors contributing to chemotherapeutic drug responses.

Methods

Genetic variants and gene expression data

Genetic variants data for Utah residents with Northern and Western European ancestry (CEU) and African individuals from the Yoruba in Ibadan, Nigeria (YRI) were downloaded from the 1000 Genome project (phase1_release_v3.20101123)(165). RNAseq gene expressions on the same individuals were downloaded from the gEUVADIS project(98).The gene expression data were normalized by library depth and transcripts length (RPKM). Gene expressions with 0 counts in more than half the samples were removed and technical variations were adjusted by PEER normalization. The detailed normalization process was described in(98).

Cytotoxicity data

Lymphoblastoid cell lines from HapMap phase 1 CEU and YRI populations were treated with increasing concentrations of capecitabine(166), carboplatin(93), cisplatin(28), cytarabine(92), and paclitaxel(167) as previously reported. For carboplatin and cisplatin, their IC_{50}, concentration required to inhibit 50% of the cell growth, were calculated and log2 transformed to normality. The areas under the survival curve (AUC) were calculated for capecitabine, cytarabine, and paclitaxel. All AUC values were also log2 transformed to allow for normal distribution. For replication studies, HapMap phase 3 YRI and CEU cell lines were treated with four of the drugs: capecitabine, carboplatin, cisplatin, and cytarabine.

Quality control for genetic variants and gene expression data

SNP data were first transformed into a variant call format (VCF) format. Only SNP data from the autosomes were used for the GWAS analyses. To minimize error accompanied with the sequencing technology, only SNPs with 100% call rate were retained using GATK(168). To remove extreme outliers and increase statistical power, we limited our analysis to SNPs that have all three possible genotypes and each genotype has at least 2 representing samples. Between 2.7 and 4.7 million SNPs have passed the quality control. Gene expressions were filtered so that 90% samples have non-zero expression values. This resulted in around 20,000 gene expression probes being retained (Table 4-1).

Table **4-1**: SNP and gene expression quality control (QC)

Drugs	Population	SNP QCed (million)	Expression QCed
Cisplatin	CEU	3.87	19,919
	YRI	4.69	20,380
Carboplatin	CEU	3.87	19,923
	YRI	4.64	20,427
Cytarabine	CEU	3.87	19,911
	YRI	4.68	20,380
Capecitabine	CEU	3.88	19,859
	YRI	4.66	20,421
Paclitaxel	CEU	2.71	19,683
	YRI	2.99	20,045

GWAS analyses of drug susceptibility

In order to perform subsequent integration analyses using genetic variants and gene expression data, only samples that are common between cytotoxicity data, 1000 Genome genetic variants data, and gEUVADIS gene expression data were used for GWAS analyses. As a result, the number of samples is different for each drug (Table 4-2) and all of the study samples are unrelated. To control for potential confounding effects due to population structure, SNPs that passed quality control criteria were first LD-pruned (--indep 50 5 2) using PLINK software(43). The principal components of the pruned SNP data were estimated using Eigenstrat(169). Along with individual's gender, significant principal components (2 or 3) were adjusted in the association analysis for each SNP. For gene expression data, Individual's sex was adjusted for each expression probe.

Table **4-2**: Genotype and gene expression associations with chemotherapeutic drugs

Drugs	Population	Discovery LCLs	Discovery Associated SNPs	Discovery Associated Expression	Hapmap 3 Replication LCLs	Replicated SNPs
Cisplatin	CEU	72	1945	121	40	324
	YRI	77	2157	76a	46	270
Carboplatin	CEU	72	2530	169a	40	304
	YRI	75	2364	194	44	248
Cytarabine	CEU	72	2156	126	40	276
	YRI	77	2749	106a	46	725
Capecitabine	CEU	73	2014	65a	40	137
	YRI	76	2485	295	46	306
Paclitaxel	CEU	29	1230	94	NA	NA
	YRI	29	1466	80	NA	NA

[a] denotes $p<0.0005$

Functional meta-analysis of associated SNPs

To determine the biological annotations that are associated across populations and drugs, we used Biofilter (v2.2)(72) to separately map the associated SNPs of each cytotoxicity phenotype to functional groups including genes regions, Pfam, GO term, KEGG pathway, and Reactome. Then, for each of the functional groups, we investigated whether any of its functional terms were shared in multiple populations and drugs. To evaluate the significance of the sharing, we carried out one thousand permutation tests, where we permuted each drug's cytotoxicity and performed GWA on the permuted outcome. If less than 5 out of 1000 permutations resulted in equal or larger number of sharing for a function term, the term was deemed significant ($p < 0.005$). After permutation, 63 genes, 35 GO terms, 2 KEGG pathways, 12 Pfam, and 39 Reactome were determined to be significant.

Integration analysis using ATHENA

We used ATHENA as described in Chapter 2.

Linkage disequilibrium patterns exist in the associated SNPs because many are proximately located. Even though they may have distinct biological functions, they are indistinguishable in regards to their association with cytotoxicity because they are highly correlated. To reduce the correlated signals resulting from LD, for each cytotoxicity phenotype, pairwise LD among all associated SNPs were estimated. $r^2 > 0.7$ was used as a threshold to form LD clusters among the associated SNPs and if a cluster has more than one SNP, the SNP that is the most significantly associated with cytotoxicity was selected as the tag SNP for the cluster. To reduce multi-collinearity in the gene expression data, Pearson correlation was calculated for all

possible gene pairs. Genes that have correlation coefficient $r > 0.8$ were grouped into a cluster. One gene from each cluster was selected as the tag gene for the cluster.

We first used ATHENA to perform variable selections on tagging SNPs and gene expressions. SNPs and gene expressions were integrated together to build neural networks that model the data. We selected SNPs and gene expressions that were included in a minimum of 2 out of 5 models built from different cross validations. The variable selection step did not take into consideration of the testing R squared to avoid over-fitting. Using the selected SNPs and gene expressions, we used ATHENA to build five models, one for each cross validation, for each cytotoxicity phenotype.

Using functional data to prioritize Neural Network models

In order to distinguish Neural Network models that have similar predictable power of cytotoxicity, we utilized functional data produced by the ENCODE project(146) to quantify the functional relevance of each model. We downloaded 128 DNase-I hypersensitivity samples from the ENCODE project (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/combined_peaks/). The data contains merged DNAse-I peaks from UW and Duke that passed FDR 1% cutoff. Genome segmentations of six ENCODE cell lines was obtained from (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/). We used the combined segmentations calls based on the consensus calling of ChromHMM and Segway algorithms. The combined segmentations splits the genome into non-overlapping regions of CTCF enriched element, enhancer, weak enhancer, promoter flanking region, promoter region including TSS, transcribed region, and repressed region. For

every SNP in the neural network model, we determined whether it is located in DNase-I hypersensitive regions or genome segmentation regions across all cell types. Because the network models only include tagging SNPs, we also determined the functional region overlaps for SNPs that are in LD with the tagging SNP. The functional score for each model is calculated as the sum of overlap for each individual SNP, normalized by the model size. In case when SNPs in LD with the tagging SNP has a higher number of overlaps, the tagging SNP was replaced with the LD SNP. In order to select the final model, we first selected 3 models that have the best prediction accuracy (R^2). Of those, we selected the model with highest functional score as the final model. Once we had the final model, we used SNPs and gene expressions to separately build SNP and gene expression only models. In case the models have negative R^2 value, the R^2 value was replaced with 0. The mean of testing R^2s for SNP and gene expressions models are shown in Table 4-4.

Results

Chemotherapeutic drug genetic associations

Cell growth inhibition was measured previously on unrelated CEU and YRI LCLs following treatment with increasing concentrations of cisplatin(28), carboplatin(93), cytarabine(92), capecitabine(166) or paclitaxel(170). Their dose-dependent inhibition was calculated as IC50, concentration required to inhibit 50% of cell growth, or AUC, area under the survival-drug concentration curve for up to 77 LCLs (Table 4-2).

Genome-wide SNP data for the LCLs were obtained from the 1000 Genomes Project (http://www.1000genomes.org/) and were evaluated for their association with each drug's cytotoxicity. We adjusted for sex and significant principal components of ancestry (2 or 3) in the

linear regression model. We identified between 1,230 and 2,749 SNPs significantly associated with each drug response at $P < 0.0005$, respectively (Table 4-2). Gene expression levels for the LCLs, measured by RNA-Seq, were downloaded from the gEUVADIS consortium (http://www.geuvadis.org/). Normalized RPKM (reads per kilobase per million) values for ~20,000 genes were tested for association with each drug's IC50 or AUC. To keep the number of associated genes similar across drugs, we used $P < 0.005$ or $P < 0.0005$ to select candidate genes. We identified between 65 and 295 genes whose expression levels were associated with drug outcome (Table 4-2). A list of all associated SNPs and gene expression levels can be found in the supplemental materials (Appendix Table 4-1:5, page 115).

To replicate the SNP associations, we exposed an independent set of HapMap phase 3 LCLs to four of the five chemotherapeutic drugs: cisplatin, carboplatin, cytarabine, and capecitabine. We performed an association analysis on the independent LCLs and using the same p-value threshold ($P < 0.0005$), we replicated between 137 and 725 SNPs that were associated in the original samples (Table 4-2).

Pan-drug analysis of associated SNPs reveals distinct patterns of functional enrichment

To get a better understanding of the biological processes involved in the differential cytotoxicity, we annotated all SNPs that are associated with each drug response using gene regions, KEGG pathways, GO terms, REACTOME, and protein families (Pfam) using Biofilter(72). We observed that many biological annotations were shared across different drugs and/or populations. To remove annotations that were shared due to random chance, we performed a permutation test (1000x) for each drug's IC50 or AUC. Using the permuted IC50 or AUC, we identified associated SNPs using the same criteria as our original analysis. For each permutation,

we calculated how many times an annotation is shared across the drug and population. We then removed any annotations that are over-represented in the permutations ($P < 0.005$).

Cellular sensitivity to drugs is a broad phenotype that include cell cycle arrest, cell damage, and cell death through apoptotic and non-apoptotic mechanisms(171,172). Cytarabine, cisplatin and paclitaxel were evaluated for chemotherapeutic-induced apoptosis because they cause a significant increase in cellular caspase-3/7 activation, a measure of apoptosis(167). Table 4-3 lists the drug, population and sample size for this phenotype.

Table **4-3**: Apoptosis phenotype measured in LCLs

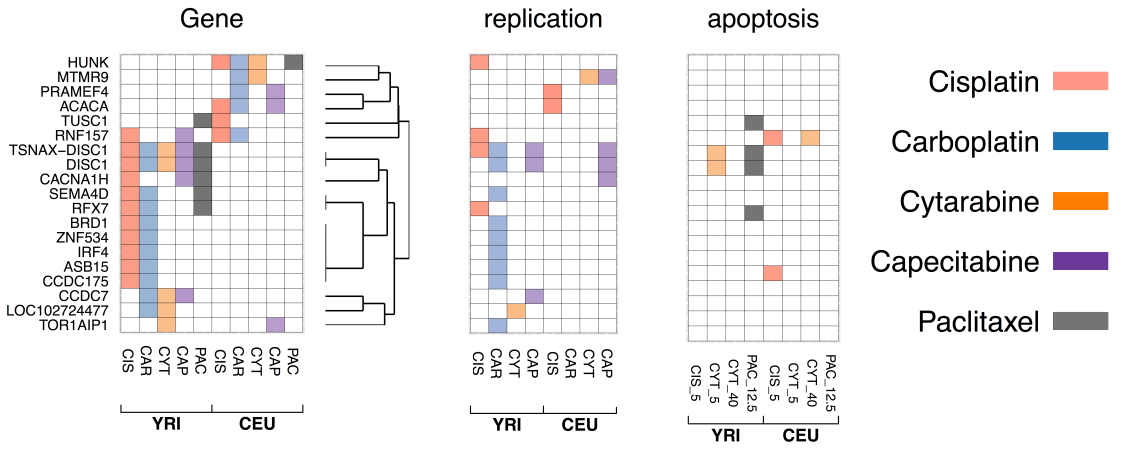
Drug	Population	Sample size
Cytarabine 5uM	CEU	30
	YRI	35
Cytarabine 40uM	CEU	30
	YRI	35
Cisplatin 5uM	CEU	30
	YRI	35
Paclitaxel 12.5nM	CEU	30
	YRI	35

We identified SNPs that are associated with drug induced caspase 3/7 activation (S2 Table) and mapped them using biological annotations. To obtain the most stringent list of biological annotations that are shared between different drugs and populations, we kept only the annotations that passed the permutation test and were also identified in the replication or apoptosis dataset (Figure 4-1).

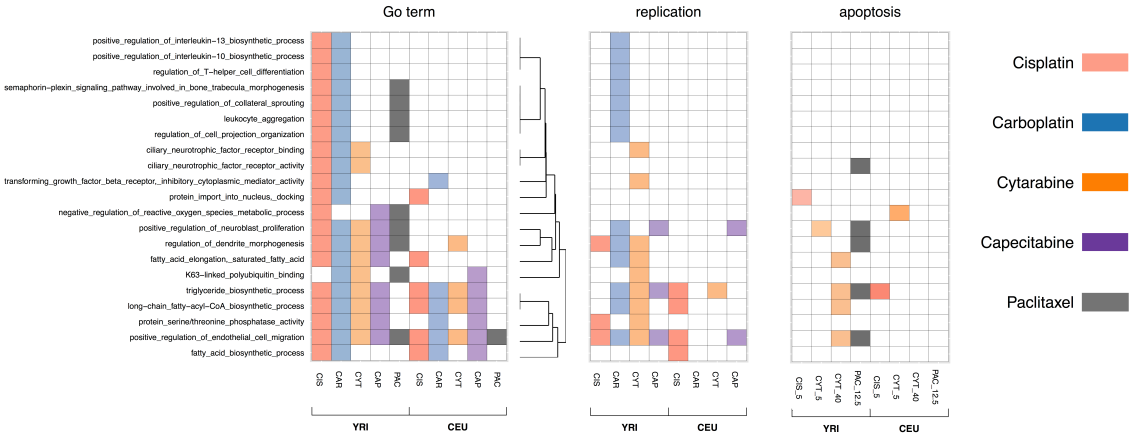
When we compared the associated functional annotations across CEU and YRI LCLs, we observed that some annotations are population specific. For gene annotations, a group of genes including *HUNK*, *MTMR9*, P*RAMEF4*, and *ACACA* were only associated in the CEU population for at lease 2 chemotherapeutics (Figure 4-1a). Meanwhile, *Spermatogenesis family BioT2*, *GNS1/SUR4 family*, *Translin family*, and *Leukotriene A4 hydrolase C-terminal* in pfam (Figure 4-1e), *IKK* related terms in REACTOME (Figure 4-1d), and several neuronal development and leukocytes GO terms (Figure 4-1b) were only identified in the YRI population. On the other hand, there is a common group of functional terms associated in both CEU and YRI populations. This group consists of mostly fatty acid related functional terms clustered together in GO term, REACTOME, and KEGG pathway. One notable example is the *NF-kappa B signaling pathway* in the KEGG pathway. This pathway was associated with all of the drugs in both populations (Figure 4-1c).

Within each population, we observed that drugs within the same class have similar associated annotation patterns. In particular, cisplatin and carboplatin, both platinating agents have many functional annotations in common. Cytarabine and capecitabine, both antimetabolites, have a number of overlapping annotations (Figure 4-1).

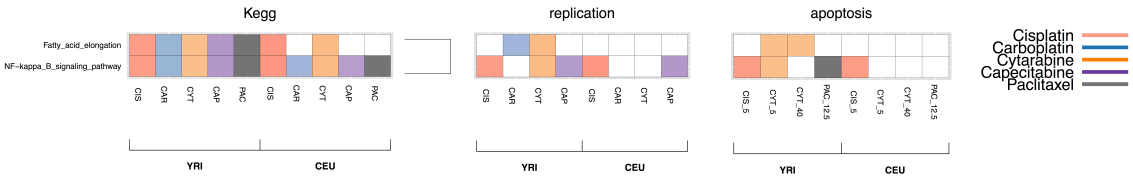
We also observed overlapping annotations between drug cytotoxicity and apoptosis. *TSNAX-DISC1* and *DISC1* gene was associated with cytarabine and paclitaxel for both cell cytotoxicity and apoptosis. A number of triglyceride and fatty acid GO terms and REACTOME pathways were shared for cytarabine, paclitaxel and cisplatin. Both *Fatty acid elongation* and *NF-kappa B signaling pathway* in KEGG are enriched for both processes. In Pfam, GNS1/SUR4 family, Translin family, and RFX DNA binding domain were enriched for cytotoxicity and apoptosis.


Gene
replication
apoptosis
HUNK
MTMR9
PRAMEF4
ACACA
TUSC1
RNF157
TSNAX−DISC1
DISC1
CACNA1H
SEMA4D
RFX7
BRD1
ZNF534
IRF4
ASB15
CCDC175
CCDC7
LOC102724477
TOR1AIP1
CIS CAR CYT CAP PAC CIS CAR CYT CAP PAC
YRI CEU
CIS CAR CYT CAP CIS CAR CYT CAP
YRI CEU
CIS_5 CYT_5 CYT_40 PAC_12.5 CIS_5 CYT_5 CYT_40 PAC_12.5
YRI CEU
Cisplatin
Carboplatin
Cytarabine
Capecitabine
Paclitaxel

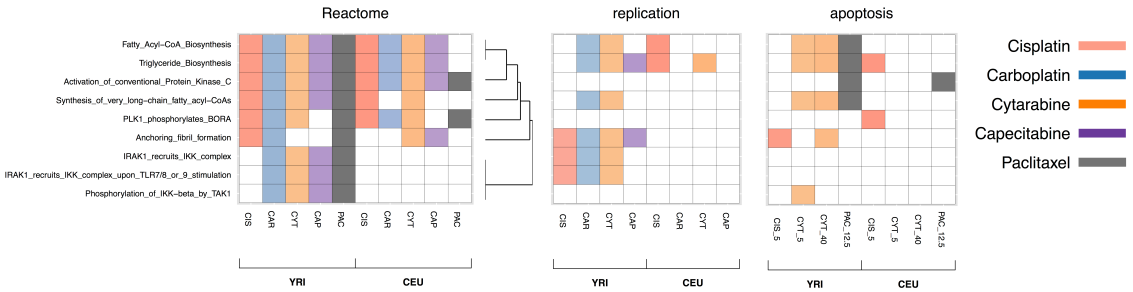
a


Go term
replication
apoptosis
positive_regulation_of_interleukin−13_biosynthetic_process
positive_regulation_of_interleukin−10_biosynthetic_process
regulation_of_T−helper_cell_differentiation
semaphorin−plexin_signaling_pathway_involved_in_bone_trabecula_morphogenesis
positive_regulation_of_collateral_sprouting
leukocyte_aggregation
regulation_of_cell_projection_organization
ciliary_neurotrophic_factor_receptor_binding
ciliary_neurotrophic_factor_receptor_activity
transforming_growth_factor_beta_receptor,_inhibitory_cytoplasmic_mediator_activity
protein_import_into_nucleus,_docking
negative_regulation_of_reactive_oxygen_species_metabolic_process
positive_regulation_of_neuroblast_proliferation
regulation_of_dendrite_morphogenesis
fatty_acid_elongation,_saturated_fatty_acid
K63−linked_polyubiquitin_binding
triglyceride_biosynthetic_process
long−chain_fatty−acyl−CoA_biosynthetic_process
protein_serine/threonine_phosphatase_activity
positive_regulation_of_endothelial_cell_migration
fatty_acid_biosynthetic_process
CIS CAR CYT CAP PAC CIS CAR CYT CAP PAC
YRI CEU
CIS CAR CYT CAP CIS CAR CYT CAP
YRI CEU
CIS_5 CYT_5 CYT_40 PAC_12.5 CIS_5 CYT_5 CYT_40 PAC_12.5
YRI CEU
Cisplatin
Carboplatin
Cytarabine
Capecitabine
Paclitaxel

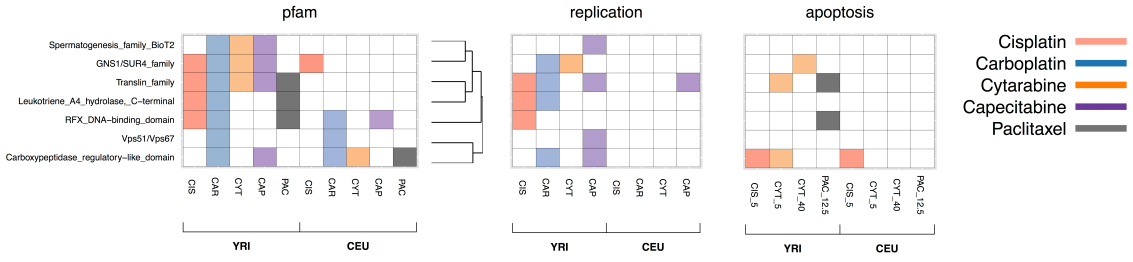
b



c



d



e

Figure 4-1. Pan-drug analysis of functional annotations. For each drug in CEU and YRI, associated SNPs were mapped to various functional annotations. A colored square indicates SNP(s) were mapped to that functional term (Cisplatin: Red, Carboplatin: Blue, Cytarabine: Orange, Capecitabine: Purple, Paclitaxel: Black). Only functional terms that have significant

enrichment across drugs and populations (permutation analysis $p < 0.005$) were shown. Functional terms were grouped using hierarchical clustering according to its enrichment. a. Gene, b. GO term, c. KEGG pathway, d. REACTOME, e. Pfam

Network modeling identified interactions between SNPs and gene expression variables important in cytotoxicity

Starting with the SNPs and gene expression variables that were associated with each drug's cytotoxicity, we calculated pairwise correlations among SNPs or gene expression. Using cutoffs of $r2 > 0.7$ for SNPs and Pearson's $r > 0.8$ for gene expression, we grouped SNPs and gene expression variables that are highly correlated to the same clusters. To reduce multi-colinearity for the network analysis, we selected one tag SNP or tag expression that had the highest association with cytotoxicity to represent each cluster. We integrated the tag SNPs and gene expressions using GENN and built interaction network models for each drug and population combinations.

Using ENCODE data to prioritize network models

It is possible that a number of network models can be similarly predictive for each drug's cytotoxicity. To prioritize these models, we selected the model that contains variables with evidence of functional relevance from ENCODE. Previous studies suggest that SNPs that lie in the open chromatin and regulatory regions are more likely to be functional(146). Thus, we used DNAseI hypersensitivity sites from 124 cell lines and genome segmentation data from 6 cell lines produced by the ENCODE project to give functional relevance for each model. The DNAseI data marks genomic regions that are not occupied by heterochromatin and the genome segmentation data divides the genome into enhancer, transcription start sites, promoter-flanking regions, CTCF

binding sites, and repressed regions. For each network model, we first identify the full set of features by including SNPs that are in the same clusters as the tag SNPs in the model. We then calculated a functional score for each feature that is proportion to the number of functional elements it overlaps with in all of the cell lines. The final score for a network model is the summation of the individual score for each feature normalized by network size (Figure 4-2). Using the functional score, we were able to prioritize models that have similar predictive power in terms of R^2 (amount of variability explained by the model) and identified one final model for each drug and population (Table 4-4 & Appendix Figure 4:1-10, page 105).

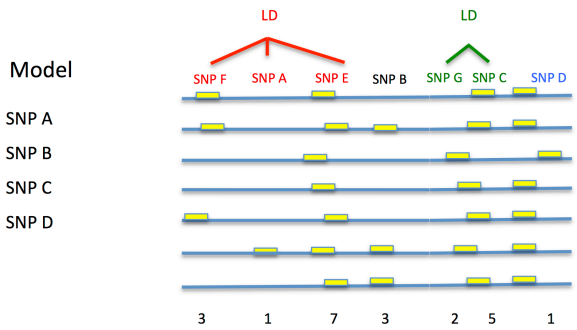


Figure 4-2: Schematic for functional score calculation. Functional score of a model is calculated as the sum of scores of individual SNP or SNPs in LD normalized by the model size. Individual score was determined by its positional overlap with functional regions. In this example, yellow squares represent DNAseI or genome segmentation regions. The score for a network model of SNP A, B, C, D is (7+3+5+1)/4 = 4.

Drugs	Population	R^2			SNPs (LD)	DNAseI	Genome Segmentation	Gene
		Integration	SNP	Expression				
Capecitabine	CEU	67.9	67.9	NA	rs4855025	NA	R, R, R, R, R, R	NA
					rs28444711	NA	R, R, R, R, R, R	
					rs7153327	11	R, R, R, R, R	
					rs75202456	NA	R, R, R, R, R	
					rs1596124	NA	R, R, R, R, R, R	
					rs2570317	NA	R, R, R, R, R, R	
	YRI	64.3	64.3	NA	rs11204113	NA	R, R, R, R, R, R	NA
					rs10760086	NA	R, R, R, R, R, R	
					rs9303059	NA	R, R, R, R, R, R	
					rs9661131	NA	R, R, R, R, R, R	
					rs6671214	NA	CTCF, CTCF, CTCF, CTCF, CTCF, R	
Carboplatin	CEU	60.4	62.1	23.1	rs11233413	9	T, E	TMEM14E
					rs12816395	NA	T, T, R, R, R, R	
					rs79062064	NA	T, T, R, R, R, R	
	YRI	66.2	66.2	NA	rs16823342	NA	R, R, R, R, R, R	NA
					rs2553650	5	WE, R, R, R	
					rs2079192	3	T, T, T, T, WE	
					rs7325063	NA	T, R, R, R, R, R	
					rs916396	NA	T, T, R, R, R, R	
Cisplatin	CEU	66.6	46.3	14.0	rs11715866	NA	T, R, R, R, R, R	FABP6 HCFC1 TAS2R30 ZNF192P1
					rs344946	NA	R, R, R, R, R, R	
					rs11628331	NA	R, R, R, R	
					rs77859257	NA	R, R, R, R, R, R	
					rs557453	NA	T, T, R, R, R, R	
					rs9422887	9	CTCF, CTCF, CTCF, CTCF, CTCF, CTCF	
					rs8074638	5	R, R, R, R, R, R	
					rs557453	NA	T, T, R, R, R, R	
					rs812652	NA	R, R, R, R, R, R	
					rs4750139	5	TSS, TSS, R	
					rs7257166	2	WE, T, R, R, R, R	
	YRI	52.4	36.8	12.7	rs12255911	NA	T, T, T, T, R, R	IL27
					rs6814234	9	WE, T, T, R, R	
					rs10426529	NA	E, R, R, R, R, R	
Cytarabine	CEU	47.7	42.9	0	rs1281461	NA	R, R, R, R, R, R	RP11-463J10.3 IL11RA
					rs2780788	NA	T, R, R, R, R, R	
					rs593525	11	T, T, T	
					rs4910512	2	T, R, R, R	
					rs7962806	NA	R, R, R, R, R, R	
	YRI	72.2	28.2	45.4	rs7666224,	NA	R, R, R, R, R, R	MAB21L3 RP11-134G8.8
					rs9564627	NA	R, R, R, R, R, R	
					rs2216926	NA	R, R, R, R, R, R	
					rs10913404	NA	R, R, R, R, R, R	
Paclitaxel	CEU	67.1	67.1	NA	rs2116796	NA	R, R, R, R, R, R	NA
					rs28634858	2	WE, R, R, R, R, R	
					rs10773683	3	R, R, R, R, R, R	
	YRI	87.8	57.0	19.0	rs10478863	NA	R, R, R, R, R, R	MAPKBP1 LPP
					rs10094960	NA	R, R, R, R, R, R	
					rs446139	NA	R, R, R, R, R, R	
					rs9905351	8	T, T, T, R, R, R	
					rs28570663	NA	R, R, R, R, R, R	
					rs10948390	NA	T, T, R, R, R, R	

Table **4-4**: Network model identified by GENN. For each drug and population, we listed R^2 and variables for integration, snp, and gene expression model. Genome segmentations abbreviations are: Enhancer (E), weak Enhancer (WE), CTCF binding (CTCF), transcribed region (T), repressed region (R), transcription start site (TSS)

Discussion

Understanding a patient's genetic susceptibility to chemotherapeutic drugs will provide important information for precision medicine. Previous studies have evaluated genotype associations to an individual chemotherapeutic drug; however, a comparative study of multiple drugs in different populations could reveal common or unique mechanisms that can be exploited in terms of therapy. Here, we present the first study to analyze the genetic associations of cytotoxicity induced by five chemotherapeutic drugs (cisplatin, carboplatin, capecitabine, cytarabine, and paclitaxel) in LCLs derived from two populations (CEU and YRI). To comparatively analyze the associated genetic variants across multiple drugs in two populations, higher-level biological knowledge was used to group variants into functional modules. We discovered that mechanistically distinct drugs are enriched in the same functional modules such as NF-kB pathway. We also set to identify biomarkers that are predictive of the drug cytotoxicity. To this end, we found that integrated networks of SNP and gene expression performed better than either data type alone. Finally, we used DNA regulatory information to select network models that are both predictive and functionally important.

We performed genome-wide SNP association analysis for each of the 5 drugs in both populations to identify significant genetic associations with drug-induced cytotoxicity. A major challenge to interpreting significant SNP associations across different drugs and populations is that comparing individual SNPs alone can be misleading. A slight change in allele frequency could result in any of the SNPs in linkage disequilibrium to be identified, however SNPs in LD are likely located in the same genes or regions. We, therefore, annotated the associated SNPs to higher-level biological processes using gene regions, GO term, KEGG pathway, REACTOME pathway, and Pfam. We found that biological annotations are considerably different between LCLs derived from individuals of European and African ancestry. Interestingly, ancestry has also been reported to affect gene expression(163), modified cytosines(173) and sensitivity to chemotherapy(174). The disparities might lie in the differences in population susceptibility to cancer, which could also affect cytotoxicity-induced response. *HUNK* and *ACACA* genes were associated only in the CEU population and are both related to breast cancer(175,176) (Figure 1a). A previous report has shown that differences exist between African Americans and European American women in the nature of breast cancer(177). *SEMA4D* and *CCDC7* genes were associated in the YRI population (Figure 4-1a). Expressions of the genes have been reported to correlate with poor outcome in cervical cancer(178,179). In addition, a recent survey has found that African American are more likely to develop cervical and lung cancer(180). These candidate genes could be further validated in their respective population. Several *IKK* related REACTOME pathways were associated with YRI population (Figure 4-1d). *IKK* is a central regulator of NF-kB pathway(181) and activation of NF-kB pathway has been observed in many solid tumors(182). Interestingly, NF-kB pathway is associated in both CEU and YRI population (Figure 4-1c), but *IKK* is only associated with the YRI population. This suggests a possible alternate regulator of NF-kB pathway for cytotoxic response.

Many annotation terms were also associated in both populations. Fatty acid and triglyceride related functional terms were identified in GO term, KEGG pathway, and REACTOME (Figure 4-1b,c, d). In Pfam, GNS1/SUR4 family is also involved in fatty acid elongation systems(183). Fatty acid synthase is an important process for cancer cells to expand and proliferate. High expression of fatty acid synthase was observed in colon, prostate, ovary, breast and endometrium cancers(184,185). Altered growth is one of the direct results of cytotoxic response, so it is likely that fatty acid synthase is also involved in the observed differential drug responses. Positive regulation of endothelial cell migration was associated with all 5 drugs. In addition, it was reported that during metastasis, cancer cells extravasate metastasis sites by attaching to endothelial cells(186). We also observed drugs that were known to treat similar cancers have high overlap of biological annotations. In particular, cisplatin and carboplatin are both platinum compounds that treat lung, head and neck, testicular, and ovarian cancer(29,158). It can be seen that cisplatin and carboplatin have high overlap in all annotations, especially in the YRI population (Figure 4-1).

LCLs' cellular sensitivity to drugs is a broad phenotype that encompasses many sub-phenotypes including drug-induced apoptosis. Cell apoptosis, as measured by caspase activity, was shown to be weakly correlated with cytotoxicity(167). Despite the weak correlation at the phenotypic level, we found that many functional terms enriched for cell cytotoxicity are also associated with cell apoptosis (Figure 4-1), indicating shared biological mechanism for the two responses. As an example, SNPs in RFX2 gene were identified in a clinical trial evaluating paclitaxel-induced neuropathy of breast cancer patients and shown to be functionally important in paclitaxel-induced cytotoxicity using siRNA(170). In our analysis, RFX DNA binding domain was associated with both paclitaxel-induced cytotoxicity and apoptosis (Figure 4-1e).

The integration of SNP and gene expression data yielded higher predictive R^2 than SNP or gene expression data alone (Table 4-4), which supports the potential value for combining multiple types of genomics data(33,90,187). Because we prioritized our model based on overlaps with DNA regulatory regions, many of our models contain SNPs that are located in the DNAseI region and functional genome segmentation regions. This information can provide additional interpretability to our models compared with using R^2 alone.

Due to the small sample size of LCLs in the original analysis, we sought for replication in independent HapMap3 LCLs to confirm our result. Of note, we found a large number of biological annotations were replicated in the independent datasets. Of annotations/drug pairs identified in the discovery analysis, between 15-100% were also significant in the respective HapMap3 replication population. This confirms that the associated SNPs might not be identical between discovery and replication studies, but the underlying biological mechanisms are the same. Our results show that many genetic variants and genes are involved in chemotherapeutic drugs cytotoxicity. By mapping genetic variants to higher-level biological processes, we were able to encapsulate variants in the same genomic region into more informative units. Comparing biological processes groups showed population specific patterns between CEU and YRI. However, as CEU LCLs were derived from an earlier time point(188), further studies are needed to verify whether some of the observed differential patterns might be due to time in culture. Nonetheless, a previous study showed that the cellular proliferation rate was not significantly different between CEU and YRI and no widespread genetic differences on common SNPs were observed between phase 2 and phase 3 YRI LCLs(188). Also, there are common processes across all drugs as well as between drugs that belong to the same class. These results could identify new drug repositioning candidates based on sharing of biological processes. We built predictive network models for drug cytotoxicity that are also functionally relevant. Future work can include additional types of functional data to better reflect the functional relevance of the models.

Chapter 5*[6]

Identification of genetic interaction networks via an evolutionary algorithm evolved Bayesian Network

Abstract

The future of medicine is moving towards the phase of precision medicine, with the goal to prevent and treat diseases by taking inter-individual variability into account. A large part of the variability lies in our genetic makeup. With the fast paced improvement of high-throughput methods for genome sequencing, a tremendous amount of genetics data have already been generated. The next hurdle for precision medicine is to have sufficient computational tools for analyzing large sets of data. Genome-Wide Association Studies (GWAS) have been the primary method to assess the relationship between single nucleotide polymorphisms (SNPs) and disease traits. While GWAS is sufficient in finding individual SNPs with strong main effects, it does not capture potential interactions among multiple SNPs. In many traits, a large proportion of variation remain unexplained by using main effects alone, leaving the door open for exploring the role of genetic interactions. However, identifying genetic interactions in large-scale genomics data poses a challenge even for modern computing. For this study, we present a new algorithm, Grammatical Evolution Bayesian Network (GEBN) that utilizes Bayesian Networks to identify interactions in the data, and at the same time, uses an evolutionary algorithm to reduce the computational cost

[6] Adapted from Li R, Dudek SM, Kim D, Hall MA, Bradford Y, Peissig PL, et al. Identification of genetic interaction networks via an evolutionary algorithm evolved Bayesian network. BioData Mining

associated with network optimization. GEBN excelled in simulation studies where the data contained main effects and interaction effects. We also applied GEBN to a Type 2 diabetes (T2D) dataset obtained from the Marshfield Personalized Medicine Research Project (PMRP). We were able to identify genetic interactions for T2D cases and controls and use information from those interactions to classify T2D samples. We obtained an average testing area under the curve (AUC) of 86.8%. We also identified several interacting genes such as *INADL* and *LPP* that are known to be associated with T2D. Developing the computational tools to explore genetic associations beyond main effects remains a critically important challenge in human genetics. Methods, such as GEBN, demonstrate the utility of considering genetic interactions, as they likely explain some of the missing heritability.

Background

Over the past decade, development in large-scale, high-throughput methods to characterize the human genome has dramatically improved our ability to assess the relationship between an individuals' genome and diseases(189). With the ever-increasing generation of genomic data, development of computational methods necessary to analyze the vast amount of data are becoming increasingly important(90). The genome-wide association study (GWAS) was the pioneering method to interrogate the genotypic and phenotypic relationship and is still being widely used today(124,190). However, despite GWAS' wide success in finding associated SNPs in many common diseases, it lacks the power to detect more complex genetic architectures such as genetic interactions(191). Therefore, a more comprehensive analysis method that can detect both main effects as well as genetic interactions is needed.

Much variability in human diseases and traits remain unexplained by using GWAS alone(191). It is hypothesized that some of the missing variability could stem from complex

genetic interactions that are unexplored by traditional association analysis. Furthermore, studies that do explore genetic interactions are often limited to two-way interactions due to the exponential increase of computational burden associated with higher-way interactions(192). A number of analytic methods have been proposed and implemented to explore interactions using statistical and data mining strategies. For example, MDR(54,55) can exhaustively evaluate all possible n-way interactions for a given n and selects the best model based on cross validations. Network based methods such as Neural Networks (63), (65) and Bayesian Networks(193) use their respective network structures to model interactions. Other notably machine learning methods including random forest(194) and SURF(195) use variable importance score to select potential interacting variables that are predictive of the outcome. However, strategies that employ exhaustive search are difficult to scale up due to the exponentially increasing search space. Machine learning methods are more flexible but they often suffer in model interpretability. Typically, the underlying pattern in data is not known a priori, thus it is important to develop a flexible method to model different types of genetic architecture. In previous chapters, we have used GENN to model complex integrations with some success. However, GENN also suffers from the issue of limited interpretability, especially with complex multi-layer networks. Compared with Neural Networks, Bayesian Networks are more interpretable in their network structures. Thus, we designed a new algorithm that is based on Bayesian Networks.

To capture main effects of genetic variants as well as complex genetic interactions, we created the Grammatical Evolution Bayesian Network (GEBN) algorithm. The algorithm can simultaneously identify marginal effects as well as interaction effects without exponentially increasing the search time. GEBN can also identify interactions that occur between different sets of genetic variants in different groups (i.e. cases and controls). This flexibility allows discovery of non-overlapping genetic architectures in multiple groups. Previous Bayesian Networks

methods to detect genetic interactions(193,196) have been limited to a small set of input SNPs. Here, we specifically chose to implement an evolutionary computation strategy to evolve the structure of the Bayesian Network because it allows us to model a larger number of SNPs while controlling for the computational time.

We implemented GEBN algorithm in the software package ATHENA. We tested the algorithm on various simulation datasets. We also applied GEBN to a case-control dataset for type 2 diabetes obtained from the Marshfield Personalized Medicine Research Project Biobank (Marshfield PMRP)(197). The network models identified novel interaction networks for type 2 diabetes cases and healthy individuals, respectively. Using the interaction networks for the two groups, we built prediction models that have an average AUC of 86%. In the following sections, we describe the GEBN algorithm, data simulations and the application in type 2 diabetes. Our results demonstrate the promise of methods like GEBN.

Methods

Grammatical Evolution Bayesian Network (GEBN)

Bayesian Network is a multivariate modeling method that expresses the relationship of variables through a series of conditional distributions. The use of Bayesian Networks is becoming very important in biology because of their ability to infer biological networks(198), model signaling pathways(199), and classifications(200,201). The current obstacle for the application of Bayesian Networks in large-scale genomics data is the exponentially increase of search space with the increase of input variables. Thus, we used a grammatical evolution (GE) algorithm to evolve Bayesian Networks in order to reduce computational time. GE is a type of genetic

programming(131,132) that uses Backus-Naur Form (BNF) grammar to create a model based on a genetic algorithm. The advantage of GE algorithm lies in its guided random search so that the search space is greatly reduced. The steps of the GE algorithm is the following:

1. Divide the data into five equal parts for cross-validations

2. For each cross validation:

Populations of binary string are randomly generated and translated into functional Bayesian Networks by the grammar. For each individual genome, the binary string is divided into consecutive codons. The codons are then translated according to the grammar (Figure 5-1).

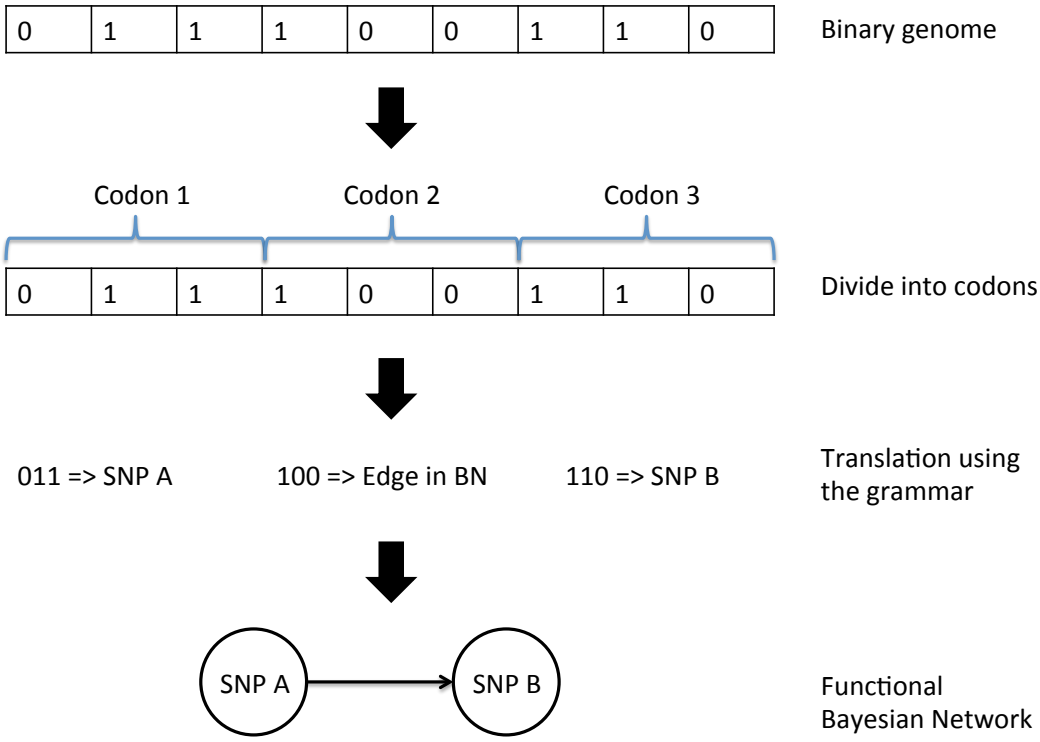


Figure 5-1: Generation of BN using the grammar

3. Calculate the fitness of the Bayesian Networks using the K2 scoring function(202).

$$P(B_s, D) = P(B_s)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!}N_{ij}\prod_{k=1}^{r_i}N_{ijk}!$$

Where D is the dataset, B is Bayesian Network, n is total number of variables, q_i is the number of different values of X_i's parents, r_i is the number of values of X_i. The score calculates the probability of observing the network given the data.

4. Select the Bayesian Networks that have the highest fitness, which will then undergo crossover and mutations. During crossover and mutation, parts of the different Bayesian Networks are exchanged or mutated to create new networks.

5. Repeat 3-4 for a set number of generations

6. Save the best model in the final generation and evaluate it on testing data

The final Bayesian Network is composed of connected and unconnected variables. Variables that are connected in the network are directly dependent with each other, while unconnected variables are conditionally independent. The advantage of GEBN over the more traditional network construction is that it can explore a wider search space, thus more suitable for large-scale genomics data. In addition, using an evolutionary search strategy removes the dependency on human trial and error to create optimal network structures and instead relies on the data and computation along with evolutionary learning to find optimal structures.

Discriminant analysis

The above GEBN method is applied to the case group and the control group independently. To prevent over-fitting, we used Bayesian Information Criteria (BIC)(203) to control the model complexity. The BIC is calculated as:

$$BIC = -2 * \ln(L) + k * \ln(n)$$

Where L is the maximum likelihood of data given a network, k is the number of free parameters, and n is the sample size. We iteratively removed each edge in the case or control network and calculated BIC for the reduced model. If the reduced model had higher BIC value, the edge was retained, and vice versa.

Finally, we used the discriminant analysis to assign an individual into either the case group or the control group. Using Bayes theorem, the probability of the sample belonging to a case group is calculated by:

$$P(Y = Case|Data) = \frac{P(Data|Y = Case) * P(Y = Case)}{P(Data|Y = Case) * P(Y = Case) + P(Data|Y = Control) * P(Y = Control)}$$

Where $P(Y = Case)$ and $P(Y = Control)$ are given by their proportions in the total sample and $P(Data|Y = Case)$ is calculated as:

$$P(Data|Y = Case) = P(Data|Y = Case, Case\ Net) = \prod_{i}^{p} P(G_i|Case\ Net)$$

p = total number of variables. $P(Data|Y = Control)$ was calculated in the same fashion.

Genetic data simulation

To test our approach, we simulated data that contains functional SNP variables with main effects and interaction effects. For main effect simulation, we simulated data that consist of different numbers of functional SNPs with varying degrees of association to a binary outcome.

For interaction effects, we separately simulated a number of interaction effects in case and control groups. We purposely made the interaction effects different in case and control groups to mimic different genetic architectures in two groups (Figure 5-2).

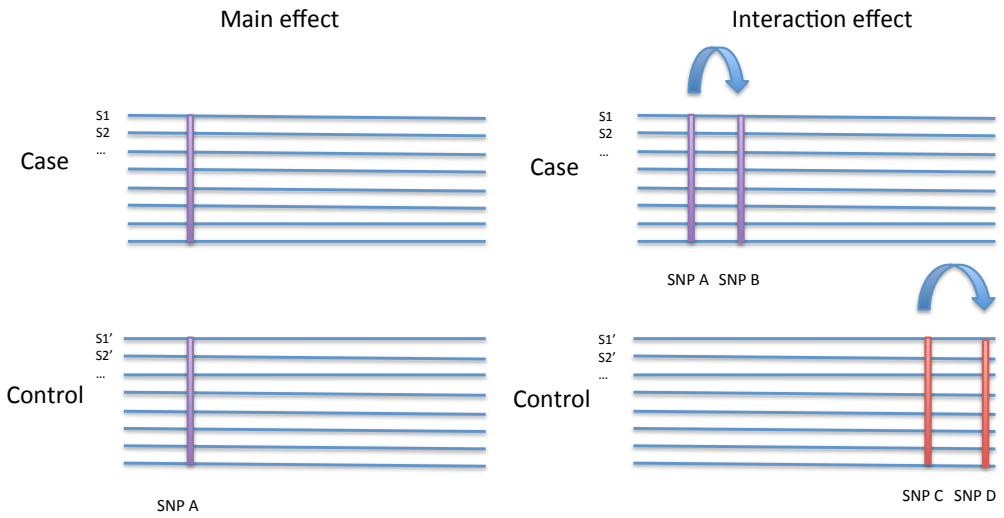


Figure 5-2: Schematic of data simulation. Main effect models have different allele frequencies in case and control datasets at the simulated SNPs. In interaction effect models, cases and control datasets have different simulated interacting SNPs without main effects.

To simulate different degrees of main effect, for a functional SNP, we altered the allele frequencies in the case data (F_{case}) using a weighted average of allele frequencies in the control data ($F_{control}$) and the extreme allele frequencies (F_{effect}) that were defined as (AA=100%, Aa=0%, aa=0%). Thus, the allele frequencies of the functional SNP in the case data is obtained by F_{case} =

$w*F_{effect} + (1-w)*F_{control}$, where w is the weight index – larger w indicating more discrepancy between the case frequencies and control frequencies.

The interaction effects were simulated as follows: Let F_{ind} denotes the joint frequencies of a pair of uncorrelated SNPs, which is calculated as the product of marginal frequencies between SNPs. The correlation can be increased by relocating the frequencies from the off-diagonal to the diagonal in the frequency table, and an extreme case is that only the diagonal have non-zero frequencies, which is denoted by F_{diag}. Different strength of interactions can be simulated by $w*F_{diag} + (1-w)* F_{ind}$.

For each dataset, we used the simulated frequency tables with sampling with replacement to determine the genotype of the functional SNPs. Then, we embedded the functional SNPs into a dataset with random SNPs to make it comparable to real biological datasets. Details of simulation parameters are shown in Table 5-1.

Table **5-1**: Data simulation details

	Functional SNPs in Case data	**Functional SNPs in Control data**	**Weight (W)**	**No. Datasets for each W**	**Total SNPs**	**Sample size**
Main effect	SNP A	SNP A	0.1, 0.5, 0.9	10	100, 500	4000
	SNP A, B, C, D	SNP A, B, C, D	0.1, 0.5, 0.9	10	100, 500	4000
Interaction effect	SNP A * SNP B	None	0.1, 0.5, 0.9	10	100, 500	4000
	SNP A * SNP B SNP C * SNP D	SNP W * SNP X SNP Y * SNP Z	0.1, 0.5, 0.9	10	100, 500	4000

Marshfield PMRP Type 2 Diabetes Dataset

The Marshfield PMRP is a biobank that has collected ~20,000 adult subjects' biological samples and electronic health records(197). We obtained SNPs data of type 2 diabetes cases and controls who were genotyped on Illumina Human660W-Quad BeadChip. We only retained individuals who are European Americans because they account for over 95% of samples and we also removed related samples. For SNP quality control (QC), we kept SNPs that have 100% call rate and minor allele frequency > 5%. The cleaned data consists of 267, 209 SNPs in 800 cases and 2465 controls. We then performed a GWAS using logistic regression to identify a set of candidate SNPs with main effects for GEBN analysis (this is a main effects filtering step(204)). Association analysis was performed while adjusting for sex, median BMI, and birth decade. Case-control status for T2D was determined using Mount Sinai's diabetes algorithm(205) from the Diabetes HTN CKD algorithm(206).

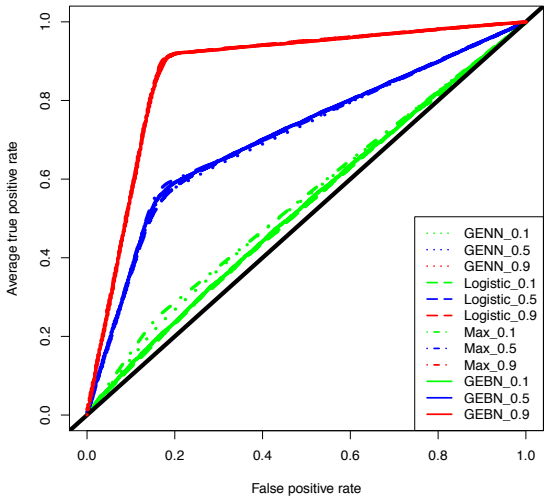
Results and discussion

Simulation Results

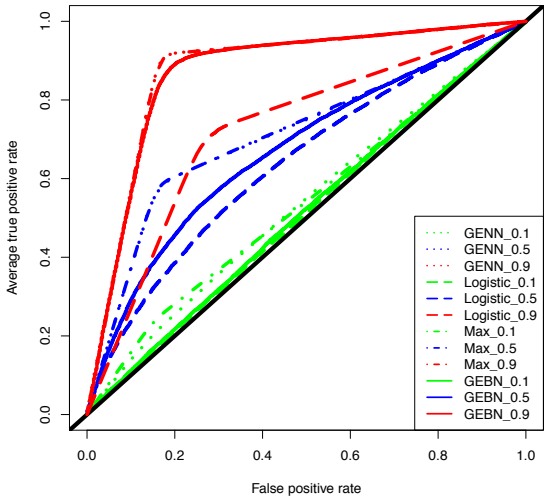
In the simulation study, we compared the performance of GEBN to that of the traditional GWAS approach based on logistic regression and another widely used method for detecting interactions, grammatical evolution neural network (GENN)(107,108). The prediction performance is summarized by the respective receiver operating characteristic (ROC) curves and the area under the curve (AUC). For each setting, we show the prediction performance averaged over 10 simulations. Regression models that include the exact simulated model (MAX) are also used to show the upperbound of prediction performance.

For main effect models, GEBN achieved close to maximum prediction performance in datasets with 100 SNPs. Logistic regression showed similar power, while GENN showed lower power. With 500 SNPs, the performance advantage of GEBN is even more visible (Figure 5-3a-d). The performance of all methods were improved by increasing the number of functional SNPs and increasing the effect size.

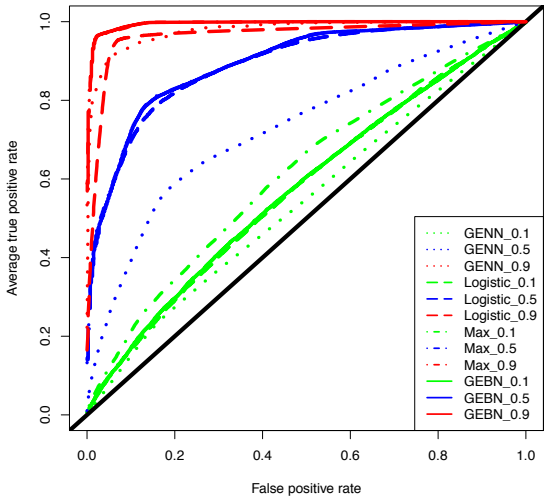
When case data and control data only differ by SNP interactions, logistic regression failed to separate two types with ROC curve fluctuating along the 45 degree line which corresponds to random guesses. GENN showed some power in detect interactions. However, GEBN showed improved ROC especially when the effect size is large (Figure 5-3e-h). The excution time for GEBN depends on the parameter settings. With the current settings of population size of 3000 and 300 generations of evolution, the average running time is 1.5±0.07 hrs for 100 SNPs and 0.97±0.1 hrs for 500 SNPs and the running time is not dependent on the underlying model. The average AUC for all the models are listed in Table 5-2.



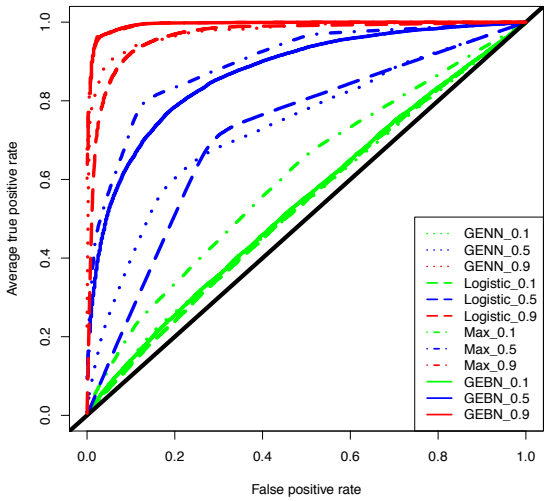
a


Average true positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
GENN_0.1
GENN_0.5
GENN_0.9
Logistic_0.1
Logistic_0.5
Logistic_0.9
Max_0.1
Max_0.5
Max_0.9
GEBN_0.1
GEBN_0.5
GEBN_0.9

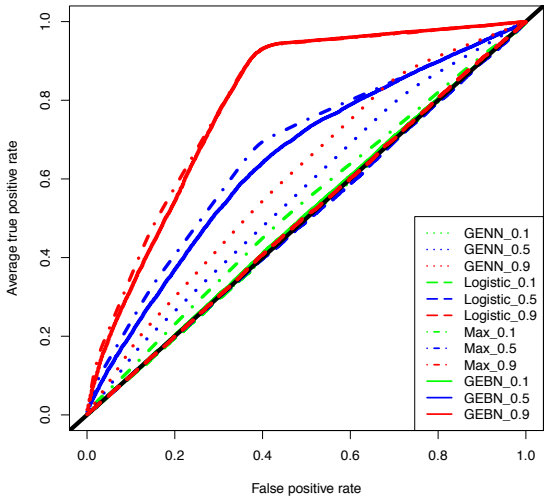
b


Average true positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
GENN_0.1
GENN_0.5
GENN_0.9
Logistic_0.1
Logistic_0.5
Logistic_0.9
Max_0.1
Max_0.5
Max_0.9
GEBN_0.1
GEBN_0.5
GEBN_0.9

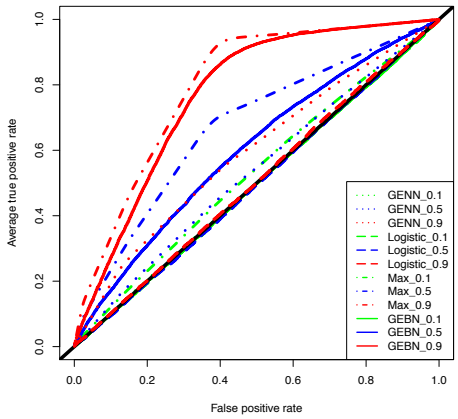
c


Average true positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
GENN_0.1
GENN_0.5
GENN_0.9
Logistic_0.1
Logistic_0.5
Logistic_0.9
Max_0.1
Max_0.5
Max_0.9
GEBN_0.1
GEBN_0.5
GEBN_0.9

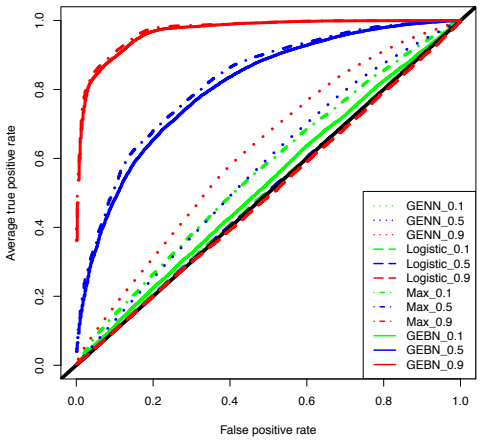
d


Average true positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
GENN_0.1
GENN_0.5
GENN_0.9
Logistic_0.1
Logistic_0.5
Logistic_0.9
Max_0.1
Max_0.5
Max_0.9
GEBN_0.1
GEBN_0.5
GEBN_0.9

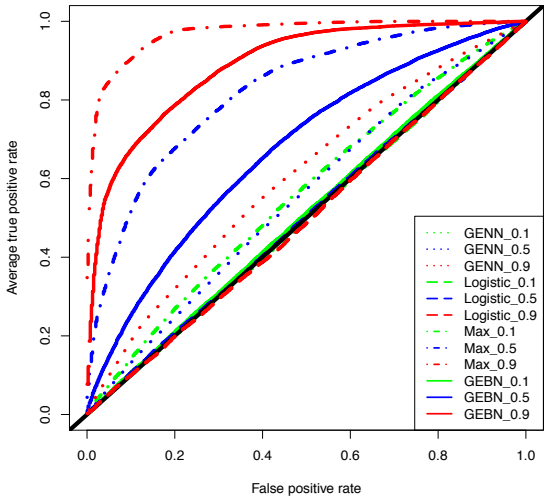
e


Average true positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
GENN_0.1
GENN_0.5
GENN_0.9
Logistic_0.1
Logistic_0.5
Logistic_0.9
Max_0.1
Max_0.5
Max_0.9
GEBN_0.1
GEBN_0.5
GEBN_0.9

f


Average true positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
GENN_0.1
GENN_0.5
GENN_0.9
Logistic_0.1
Logistic_0.5
Logistic_0.9
Max_0.1
Max_0.5
Max_0.9
GEBN_0.1
GEBN_0.5
GEBN_0.9

g



h

Figure 5-3: Simulation results for additive and interaction models using grammatical evolution Bayesian Network (GEBN), grammatical evolution neural network (GENN), logistic regression, and logistic regression with the exact simulated model (MAX). The colors represent different weight indexes (red = 0.9, blue = 0.5, green = 0.1). These weight indices correspond to strength of the simulated effects. a. Main effect model: SNP A (100) b. Main effect model: SNP A (500) c. Main effect model: SNP A, B, C, D (100) d. Main effect model: SNP A, B, C, D (500) e. Interaction model: SNP A<->B (100) f. Interaction: SNP A<->B (500) g. Interaction model: SNP A<->B, C <->D, W<->X, Y<->Z (100) h. Interaction model: SNP A<->B, C <->D, W<->X, Y<->Z (500)

Table **5-2**: Comparison of AUC for GEBN and logistic regression

	Functional SNPs in Case data	Functional SNPs in Control data	Weight (W)	MAX	Regression		GENN		GEBN	
				100	100	500	100	500	100	500
Main effect	SNP A	SNP A	0.1	55	52	51	54	54	53	52
			0.5	71	71	64	70	71	71	67
			0.9	88	88	72	88	88	88	87
	SNP A, B, C, D	SNP A, B, C, D	0.1	61	57	53	54	54	58	54
			0.5	90	89	72	73	73	89	87
			0.9	99	96	87	98	98	99	99
Interaction effect	SNP A * SNP B	None	0.1	53	50	50	50	50	50	50
			0.5	67	50	49	56	53	65	60
			0.9	89	50	50	60	59	80	77
	SNP A * SNP B	SNP W * SNP X	0.1	56	50	50	50	50	52	51
			0.5	82	50	50	57	55	81	67
	SNP C * SNP D	SNP Y * SNP Z	0.9	97	49	50	62	60	97	89

Type 2 diabetes results

We first performed association analysis using logistic regression for 267,209 SNPs associations with type 2 diabetes, using $p < 0.001$ as threshold, we identified 259 SNPs associated with type 2 diabetes. The top associated SNP was rs7903146 (p=2.997e-06), which maps to *TCF7L2* gene. To remove SNPs that are correlated, we used PLINK software(43) to prune the associated SNPs based on linkage disequilibrium (--indep 50 5 2). 202 SNPs remained after LD pruning. We applied GEBN on the 202 SNPs, together with sex, median BMI, and birth decade,

to separately build interaction networks for type 2 diabetes cases and controls. We then used the final network from cases and controls to perform discriminate analysis on the independent testing data. The average prediction AUC of 5-fold cross validation was 86.8% (Figure 5-4).

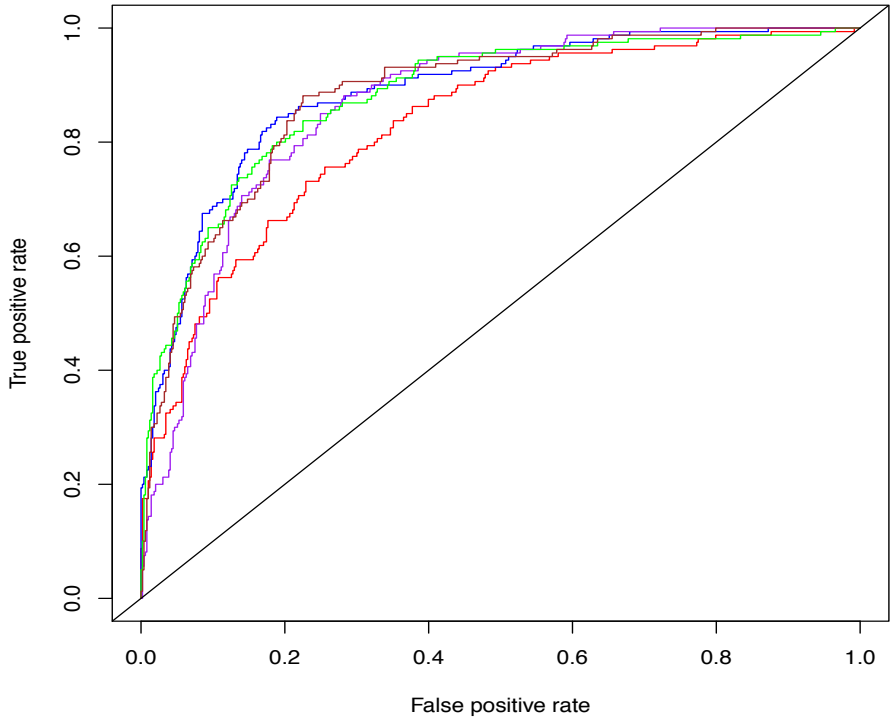
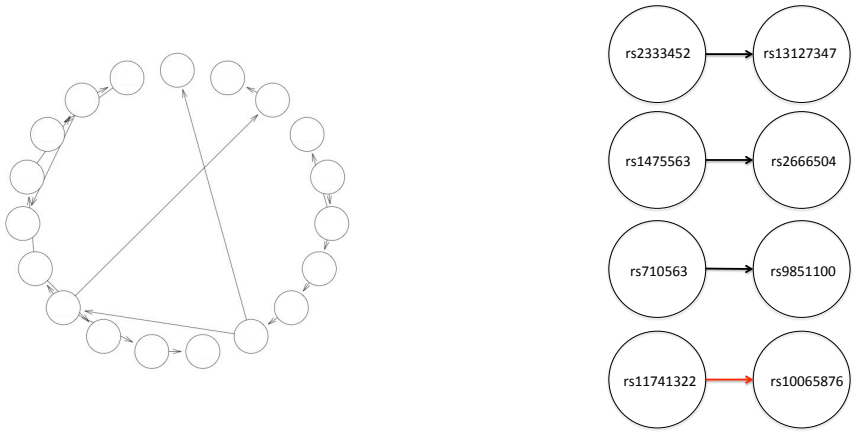


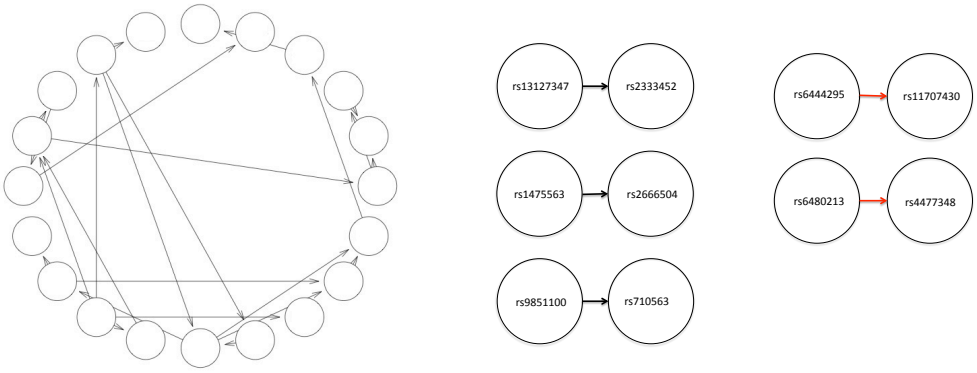
Figure 5-4: Testing ROC curve for type 2 diabetes. Each color represents a single cross-validation.

Figure 5-5 shows the best Bayesian Network models for cases and controls. The AUC for the best model was 88.7%. The networks also include the rest of the SNPs as marginal variables, but for clarity, they were not shown. The cases and controls share there common interactions: rs13127347 and rs2333452, rs9851100 and rs710563 (both in *P3H2* gene), rs2666504 and rs1475563 (*INADL* gene*)*. There was also one unique interaction for cases, which is rs10065876

and rs11741322 and two for controls, which are rs4477348 and rs6480213 (both in *CTNNA3* gene), and rs11707430 and rs6444295 (both in *LPP* gene).



a.



b.

Figure 5-5: Best Bayesian Network models for cases and controls. Left panel shows network structure before BIC pruning. Right panel shows network structure after BIC pruning, and the red edges indicate interactions only found in the case data or the control data, but not both cases and

controls. a. Case data network b. Control data network

Conclusions

In this study, we presented a novel algorithm that can efficiently capture marginal and interaction effects present in the genetic data. We demonstrated in simulation data that GEBN performed equal or better than the standard GWAS analysis method using logistic regression as well as GENN on data with only main effect functional SNPs. In data with interacting SNPs, logistic regression failed to capture the true model which is shown by the ~50% AUC (Figure 5-2). GENN was able to capture simulated interactions, however, the predictive power were significantly lower than the MAX models, which gives the upper bound of prediction performance. On the other hand, GEBN were able to separately identify the unique interactions in cases and controls and use that information to distinguish the two groups. The performance of GEBN was close to the maximum prediction power in data with 100 and 500 SNPs. One concern was that GEBN can potentially over fit the data because networks were trained separately for each group. However, our testing AUCs showed that we did not over fit the model. As a further validation, GEBN was applied to the xor dataset from Chapter 2. The xor dataset contains two simulated xor variables and 98 noise variables. GEBN was able to achieve 100% power in all datasets.

Using main effect filtering followed by GEBN analysis, we replicated canonical associations and also identified novel genetic interactions for type 2 diabetes. The most significant association was rs7903146, which is located in the *TCF7L2* gene. We also identified rs12255372, which is in LD with rs7903146, as a significant association. *TCF7L2* gene has been

implicated for type 2 diabetes in many studies(4,207). We limited the network analysis to the top 202 associated SNPs because it is a comparable size to our simulation study. It is interesting that the top case and control networks have common as well as unique edges. The common edges include two non-coding SNPs on chromosome 4, two SNPs within *P3H2* gene and one SNP in *INADL* gene and one SNP in the non-coding region of chromosome 1. The *INADL* gene is part of the hippo signaling pathway(74). The pathway has been shown to regulate pancreas development(208) and adipocyte development(209). Interestingly, a prior study has found that *INADL* was associated with children's weight(210). It is difficult to interpret the unique interaction for case group because both of the SNPs are located in non-coding regions. These could be further analyzed by looking into the ENCODE and GTEx regulatory data for possible functions. For controls, *CTNNA3* were found to be associated with Alzheimer(211) and heart disease(212). *LPP* gene has shown a robust association with type 2 diabetes in multiple ethnicities as well as combined meta-analysis(213). Taken together, we have shown that GEBN have identified several known genes associated with type 2 diabetes. Using logistic regression, we also obtained a similar prediction AUC of 86.5%. The similar performance was mostly due to the candidate SNPs were selected using a main effect filtering. Despite the similarity in the AUCs, GEBN was able to identify more complex genetic structures in diabetes cases and controls than logistic regression.

This paper presents the first step of the algorithm development that aims to address the pressing need for tools to identify complex relationships within the genetics data. Due to the flexibility of the Bayesian networks, the algorithm could be applied to datasets with more than two outcomes. For example, drug response phenotypes might be categorized as high responder, low responder, and non-responder. This would be possible to analyze with GEBN.

The utility of GEBN will be even greater in those settings because traditional statistical approaches are generally limited to binary outcomes. We also plan to integrate other –omics data such as transcriptomic and methylomic data into the network. The potential interactions between factors from different data types could reveal novel biological insights not seen at any individual data alone. The ultimate goal of individually identifying networks for different groups or subtypes of disease is to more precisely understand the disease so that we can improve detection and treatment of the disease. The method presented in this chapter will help further elucidate the complex biological relationship present in the genetics data.

Chapter 6 Conclusion

While the pace of data generation in genetic research has skyrocketed in recent years, the development of appropriate tools has comparably lacked behind. Therefore, this dissertation was particularly focused in the areas of methods development for analyzing large-scale genomics dataset. In this chapter, the author will provide summaries for the previous chapters and provide several future development plans.

Chapter 1 has provided an important rationale for this dissertation, namely, the biological system is a complex system and it needs diverse datasets and methods to model it. Genomics data is no longer restricted to genotyping or sequencing data. As Figure 6-1 shows, multiple layers of data already exist between the genome level to the phoneme level. The additional layers of data allow researchers to paint a fuller picture of the genetic association to the phenotype.

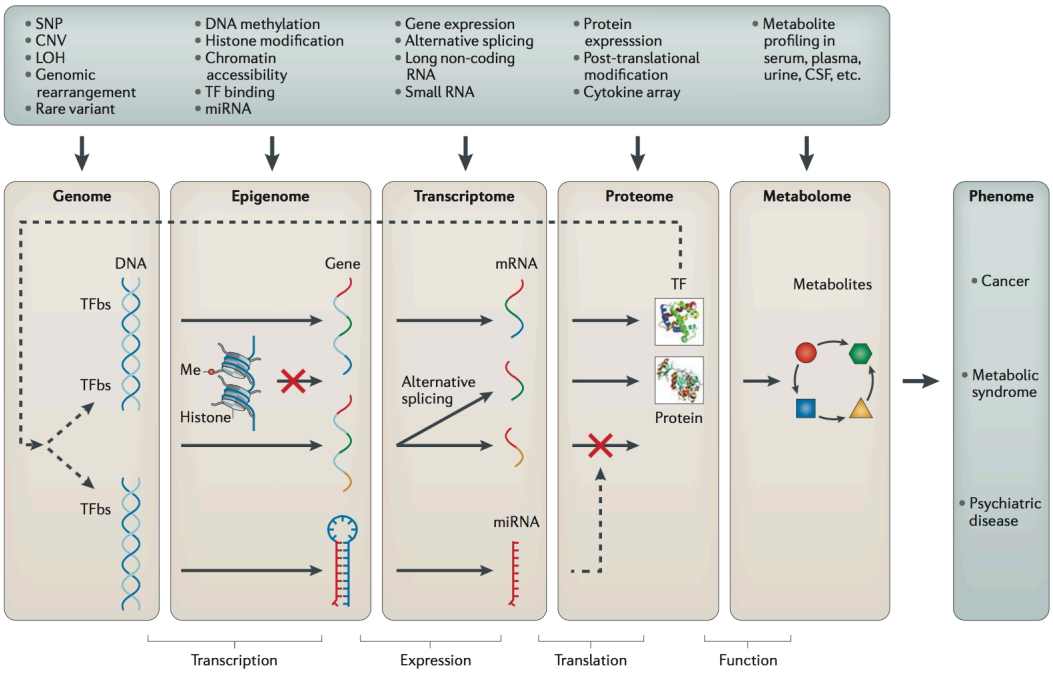


Figure 6-1 Biological systems multi-omics from the genome, epigenome, transcriptome, proteome and metabolome to the phenome. Heterogeneous genomic data exist within and between levels, for example, single-nucleotide polymorphism (SNP), copy number variation (CNV), loss of heterozygosity (LOH) and genomic rearrangement, such as translocation, at the genome level; DNA methylation, histone modification, chromatin accessibility, transcription factor (TF) binding and micro RNA (miRNA) at the epigenome level; gene expression and alternative splicing at the transcriptome level; protein expression and post-translational modification at the proteome level; and metabolite profiling at the metabolome level. Arrows indicate the flow of genetic information from the genome level to the metabolome level and, ultimately, to the phenome level. The red crosses indicate inactivation of transcription or translation. CSF, cerebrospinal fluid; Me, methylation; TFBS, transcription factor-binding site. (Source: Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet)

The most widely used analysis technique in detecting associations between genotype and phenotype is the GWAS. However, with the addition of other layers of data, GWAS is no longer sufficient on its own. Chapter 1 reviewed several analysis approaches in addition to association analysis including: identifying epistasis interactions, using prior biological knowledge, and performing system genomics analysis. The author concluded that an ensemble approach that can integrate multiple methods would likely to be the most successful in the future.

One ensemble approach, ATHENA, was the focus of Chapter 2. A neural network approach (GENN) implemented in ATHENA was consistently producing simple 1-layer networks in many different datasets. One drawback of the small network is the reduced ability to model complex relationships that are commonly present in the genomics data. Using a simulated XOR model that requires multi-layer neural network to model, the author found the optimal algorithm settings that would allow GENN to build multi-layer networks.

Using the knowledge gained from Chapter 2, the author applied GENN to a pharmacological phenotype, etoposide IC_{50}. The analysis integrated two types of data: SNPs and DNA methylations to predict drug responses on CEU and YRI cell lines. The analysis identified networks of SNPs and methylation levels that are predictive of the drug response. The study, however, was limited by the lack of replications and comparisons with other drug response phenotypes.

With the limitations of the previous study in mind, a more comprehensive analysis on drug-induced cytotoxicity was carried out in Chapter 4. The new study included 5 different drug's IC_{50}, a related pharmacological phenotype, cell apoptosis, and an independent set of samples for replication. The analysis also featured an ensemble approach that included using biological databases, ENCODE annotations, and network analysis of genetic factors to compare models across drugs and cell lines. This approach is also not limited to pharmacological phenotypes; any traits or phenotypes can be interchanged with IC_{50}.

Neural Network (NN) models used in the previous chapters have been widely applied to many genetic traits. However, it has several limitations with the primary being poor interpretability. In Chapter 5, a Bayesian Network (BN) based algorithm was created to address this issue. Bayesian Networks has better interpretability than Neural Networks as the network edges in BN represent direct relationship among variables. The disadvantage of Bayesian Networks is that it requires much more computational time to model. To circumvent this problem,

an evolution algorithm approach was fused with BN to create GEBN. GEBN was tested in simulation data and applied to a type 2 diabetes dataset.

Currently, a new version of Bayesian Network algorithm is being developed. The new algorithm Genetic Algorithm Bayesian Networks (GABN) differs from GEBN in one important aspect (Figure 6-2). GABN would able to identify multiple sub-connected networks in the data compared to only one for GEBN.

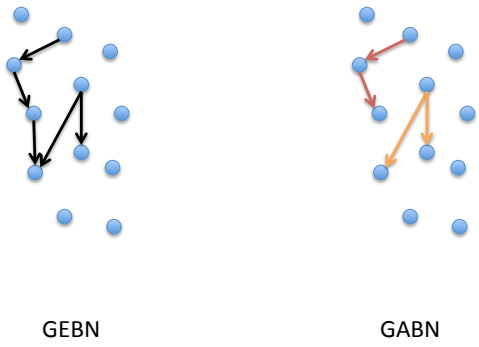
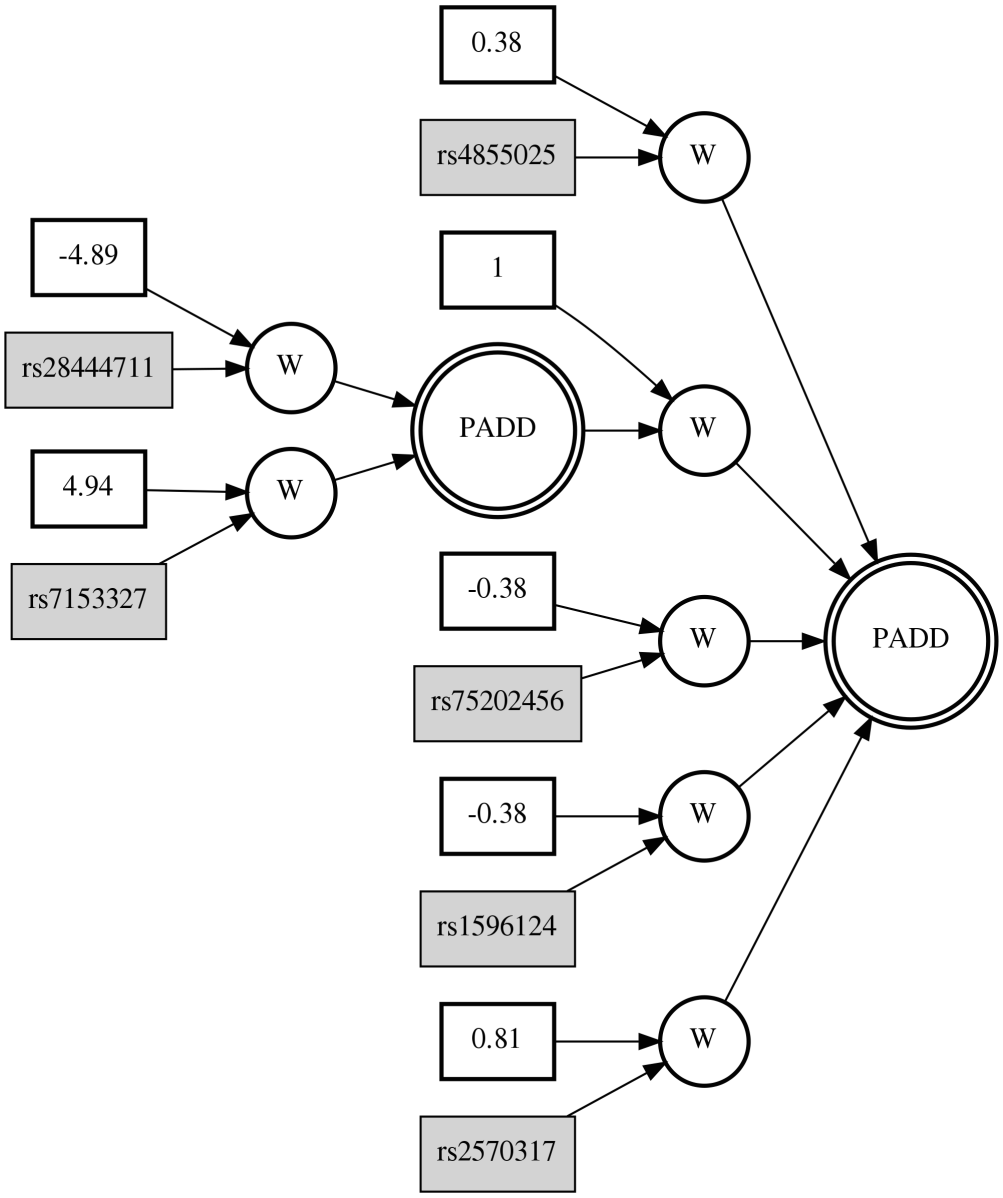


Figure 6-2 Differences between GEBN and GABN. GEBN only allows identifying one connected network, while GABN allows multiple sub-networks. In this example, the two sub-networks were forced to be connected in GEBN

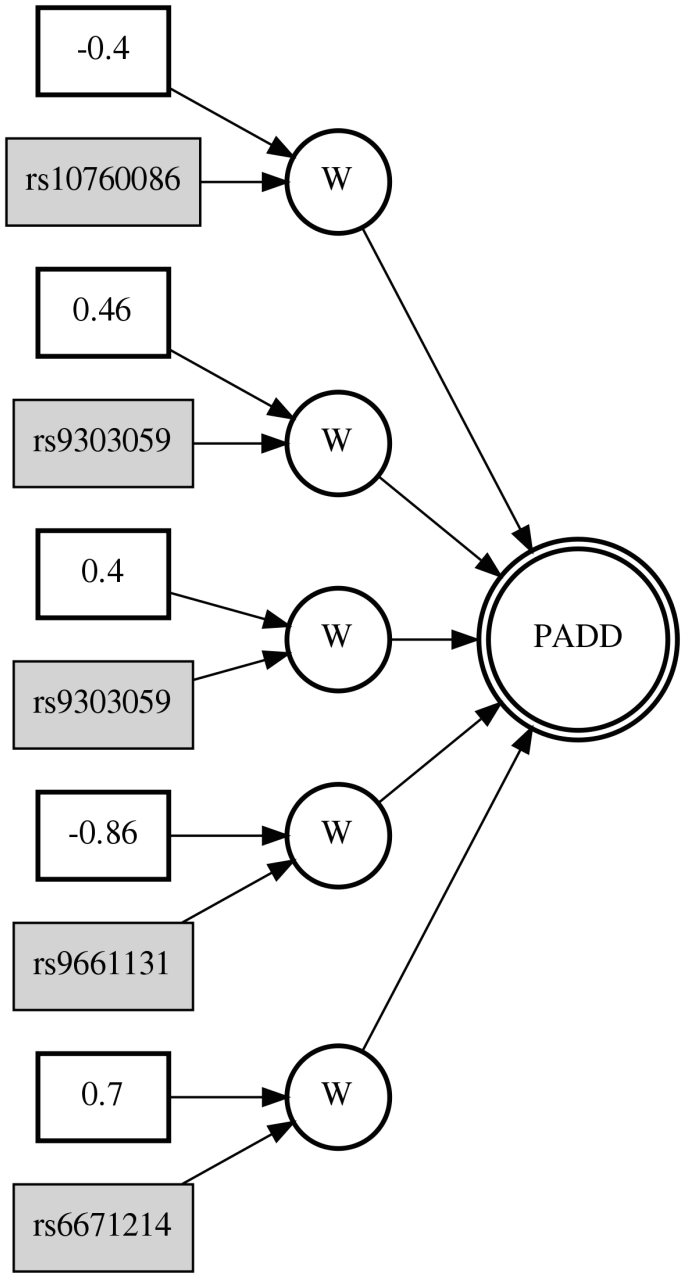
Other areas of future development include the ability to use prior knowledge when constructing edges of the networks, the capacity to include continuous variables into the network, and the ability to model multi-level phenotypes, such as cancer subtypes.

In summary, this dissertation represents only a small step towards using multi-omics data to understand genetic traits. While many diverse data types have been used, two other important data: copy number data and rare genetic variants data have not been explored. It is for certain that additional data sets will be discovered and implemented in the future. However, the central theme should remain unchanged, that is, in order to understand the complexity of the biological system, information from multiple facets of the biological system needs to be integrated and interrogated.

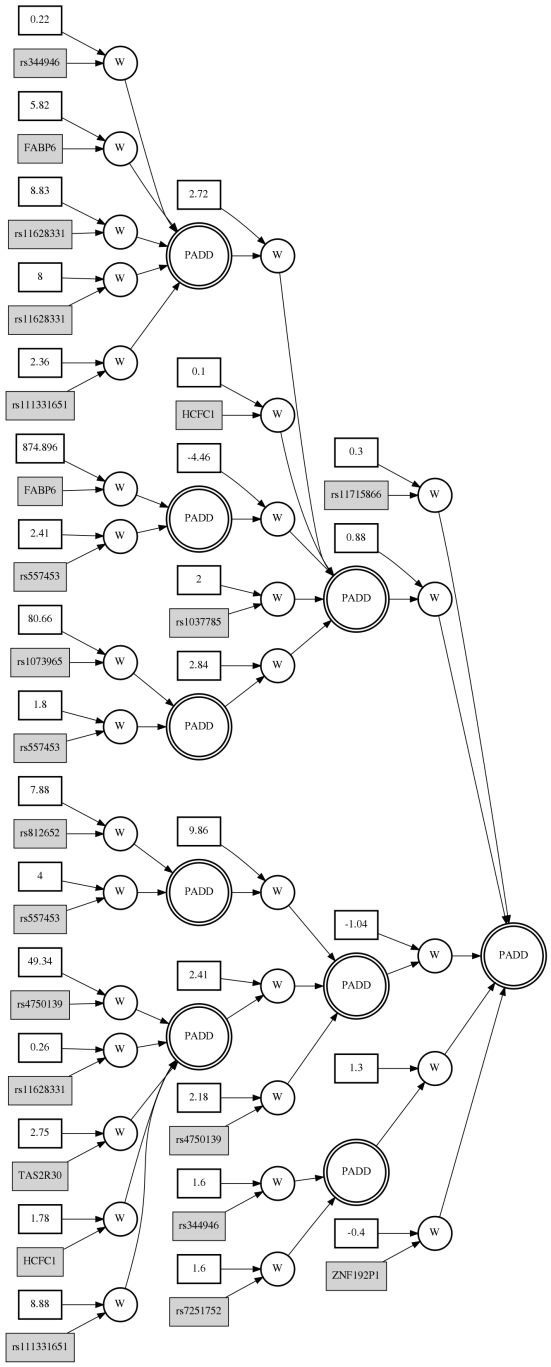
Appendix



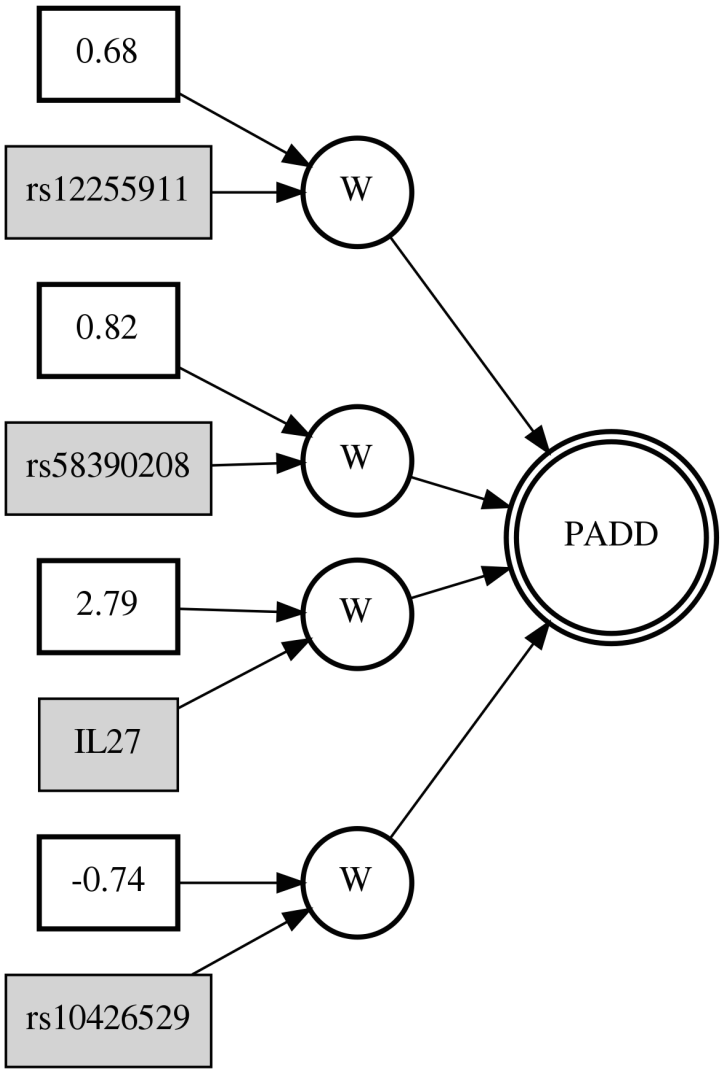
Appendix Figure 4-1: Neural Network model for capecitabine chemotherapeutic response in CEU. W is a weight node, PADD is an addition activation node



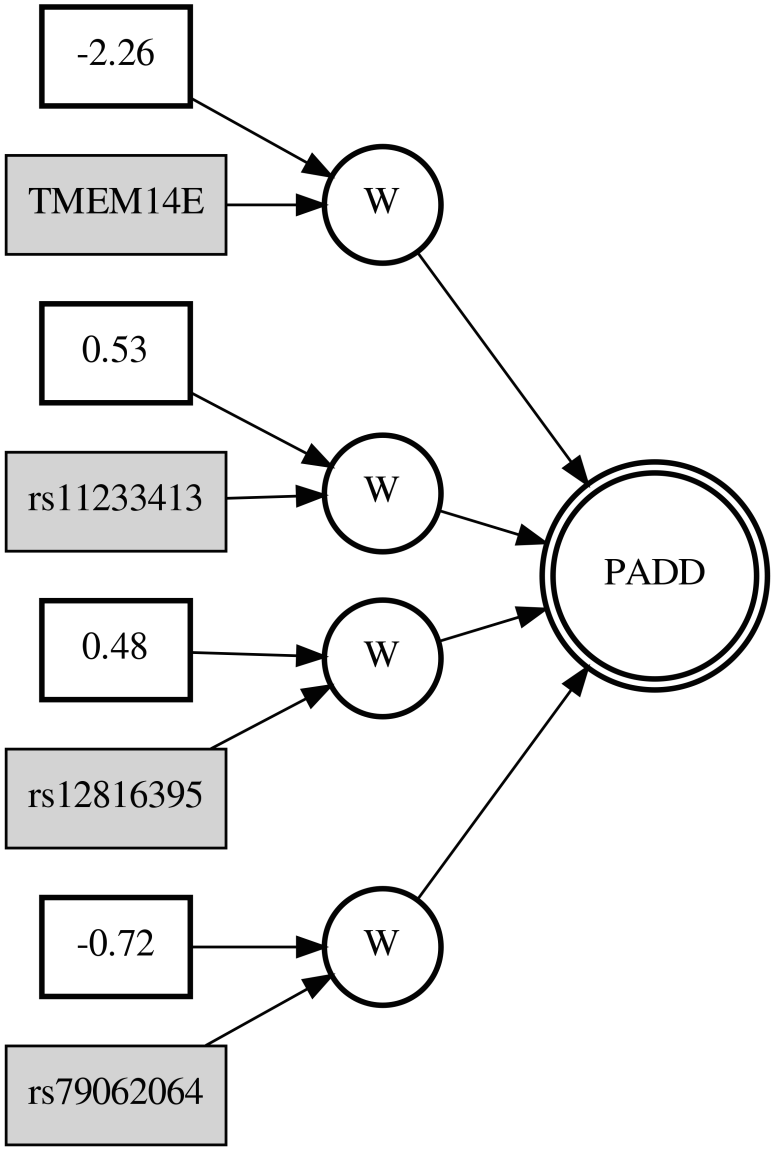
Appendix Figure 4-2: Neural Network model for capecitabine chemotherapeutic response in YRI. W is a weight node, PADD is an addition activation node



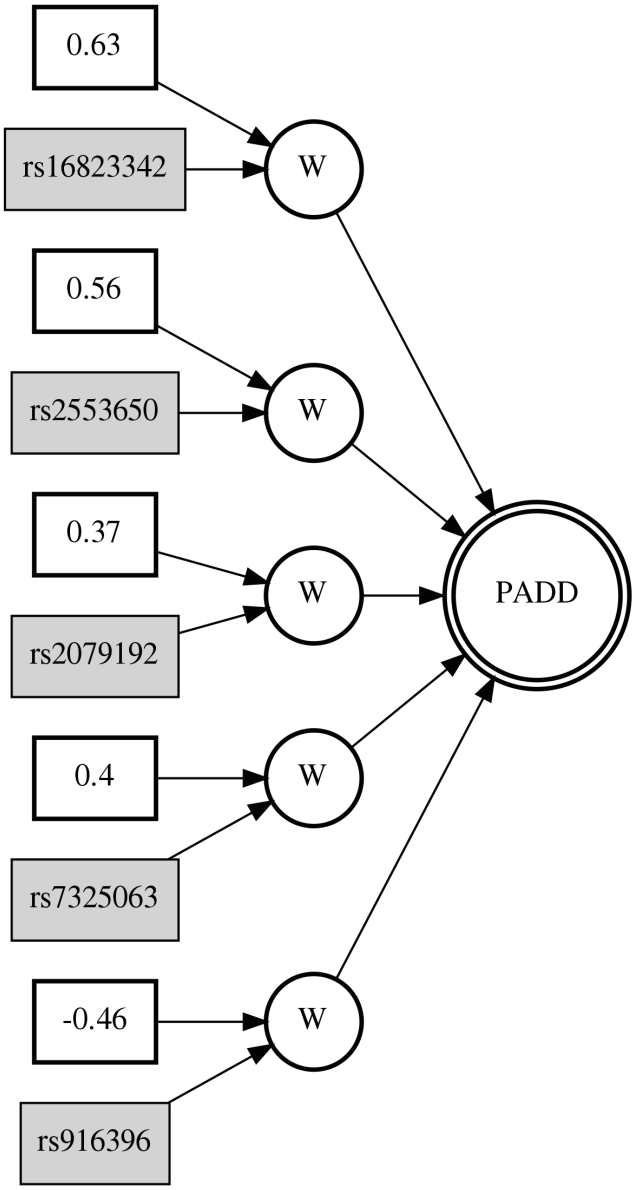
Appendix Figure 4-3: Neural Network model for cisplatin chemotherapeutic response in CEU. W is a weight node, PADD is an addition activation node



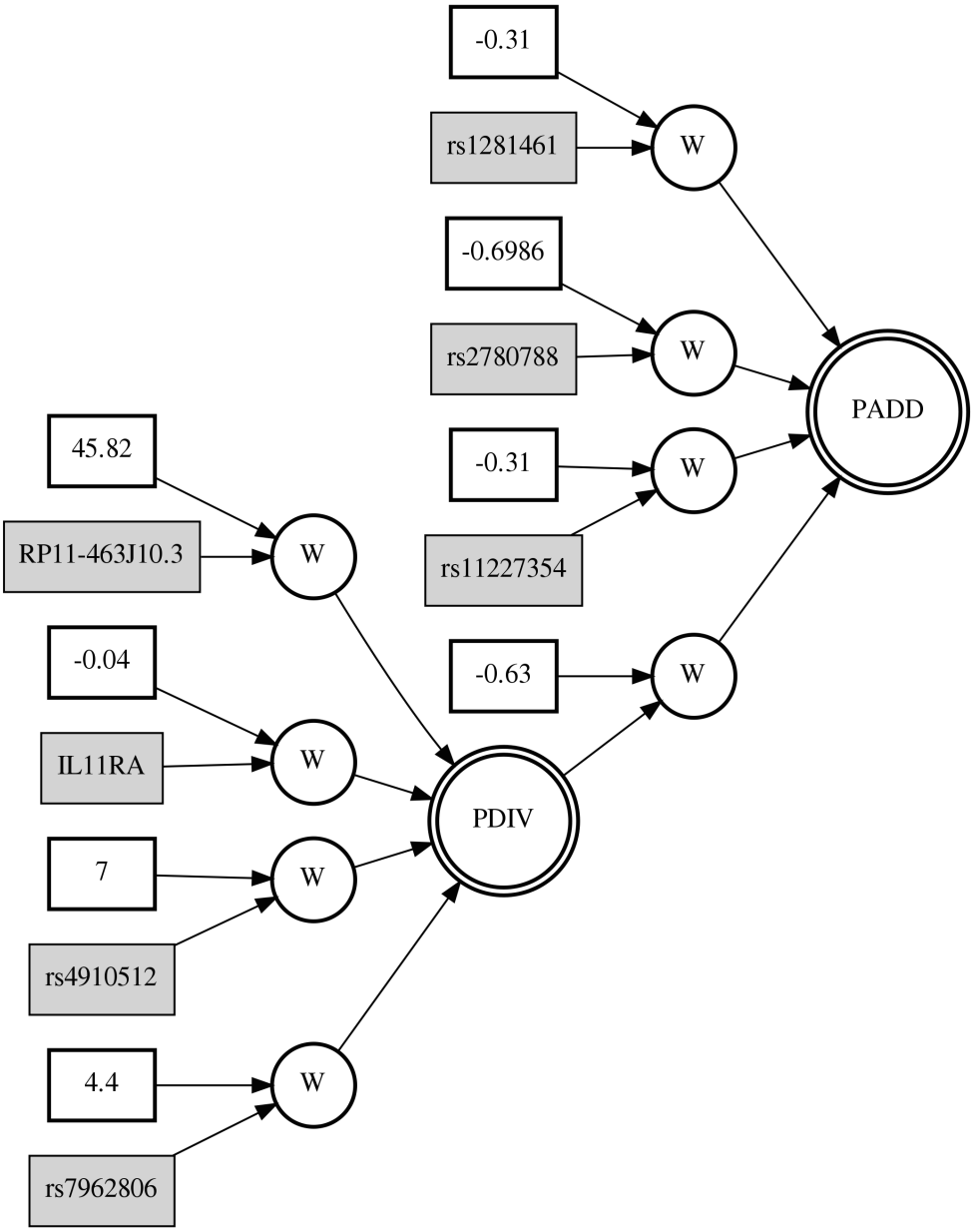
Appendix Figure 4-4: Neural Network model for cisplatin chemotherapeutic response in YRI. W is a weight node, PADD is an addition activation node



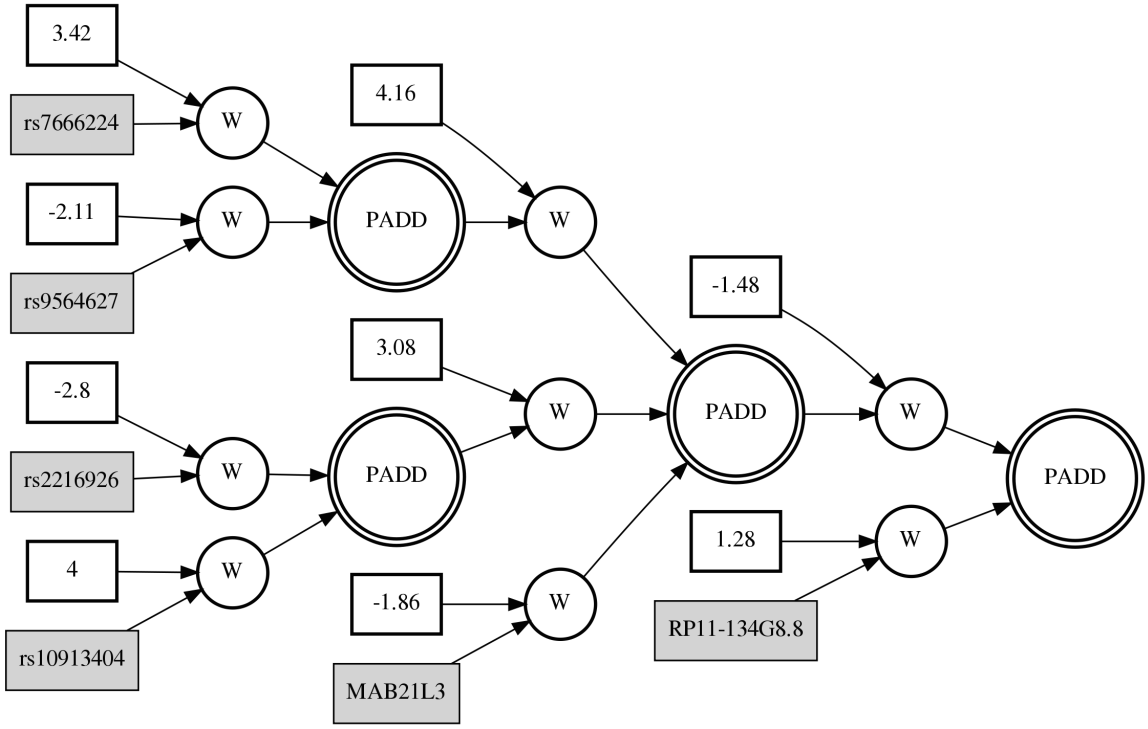
Appendix Figure 4-5: Neural Network model for carboplatin chemotherapeutic response in CEU. W is a weight node, PADD is an addition activation node



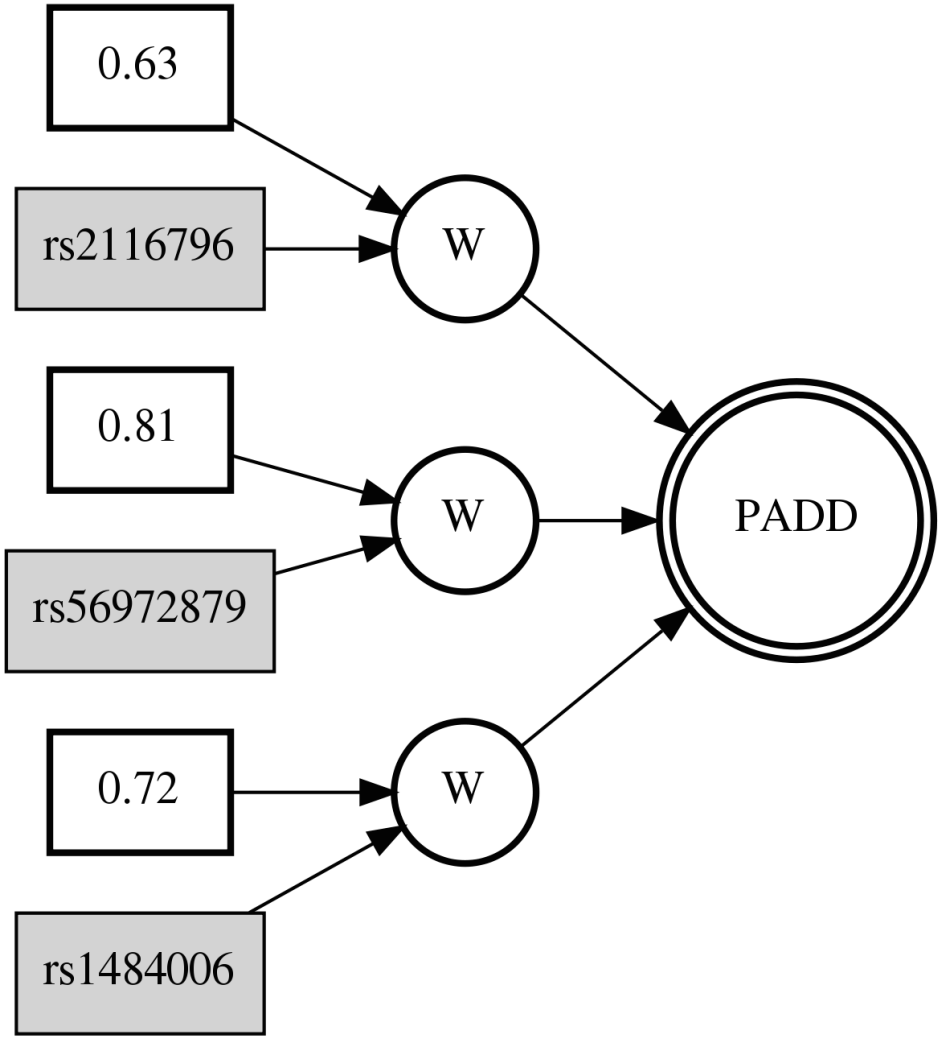
Appendix Figure 4-6: Neural Network model for carboplatin chemotherapeutic response in YRI. W is a weight node, PADD is an addition activation node



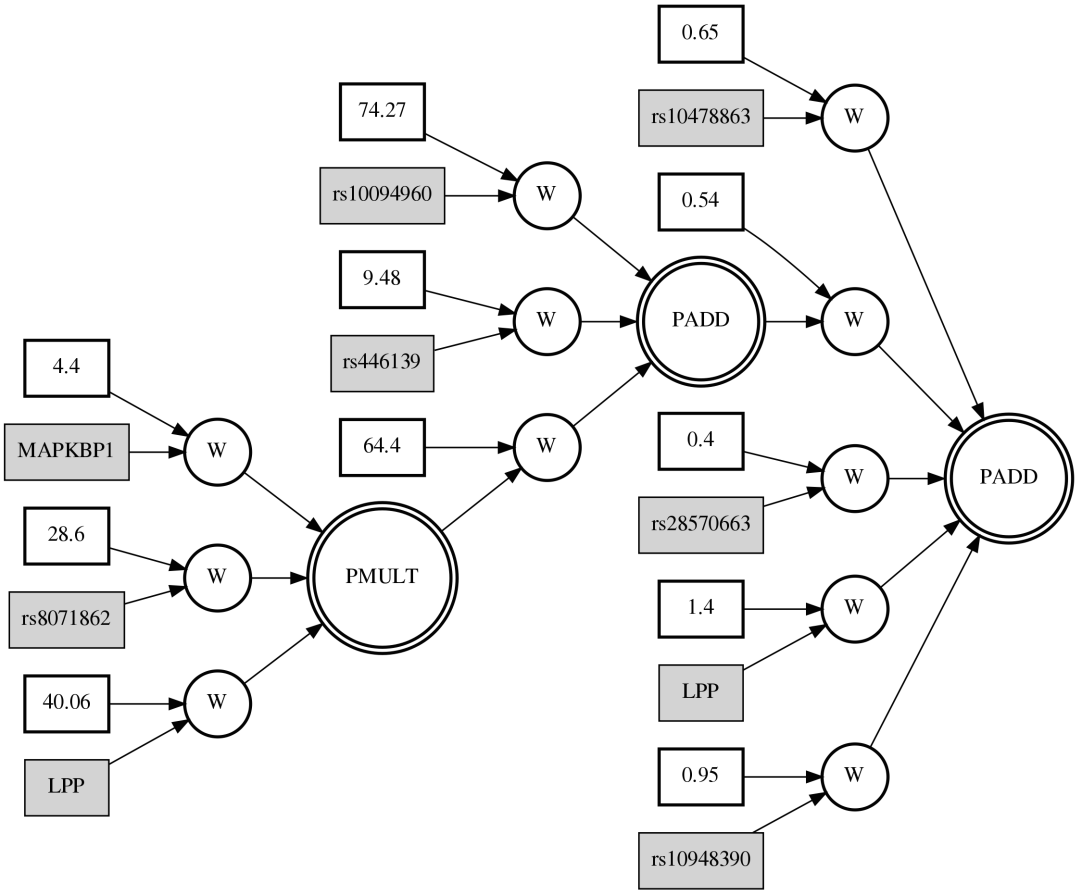
Appendix Figure 4-7: Neural Network model for cytarabine chemotherapeutic response in CEU. W is a weight node, PADD is an addition activation node, PDIV is a division node



Appendix Figure 4-8: Neural Network model for cytarabine chemotherapeutic response in YRI. W is a weight node, PADD is an addition activation node



Appendix Figure 4-9: Neural Network model for paclitaxel chemotherapeutic response in CEU. W is a weight node, PADD is an addition activation node



Appendix Figure 4-10: Neural Network model for paclitaxel chemotherapeutic response in YRI. W is a weight node, PADD is an addition activation node, PMULT is a multiplication node

Appendix Table **4**: Associated SNPs and gene expression for chemotherapeutic drugs

Deposited at https://scholarsphere.psu.edu/collections/ns064615m

Reference

1. IBM. Big Data & Analytics Hub [Internet]. Available from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
2. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLOS Biol [Internet]. Public Library of Science; 2015 Jul 7 [cited 2016 Jul 7];13(7):e1002195. Available from: http://dx.plos.org/10.1371/journal.pbio.1002195
3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet [Internet]. Nature Publishing Group; 2010 Jul 20 [cited 2016 Jul 13];42(7):565–9. Available from: http://www.nature.com/doifinder/10.1038/ng.608
4. Chandak GR, Janipalli CS, Bhaskar S, Kulkarni SR, Mohankrishna P, Hattersley AT, et al. Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population. Diabetologia [Internet]. 2007 Jan [cited 2015 Jul 23];50(1):63–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17093941
5. Ng KP, Hillmer AM, Chuah CTH, Juan WC, Ko TK, Teo ASM, et al. A common BIM deletion polymorphism mediates intrinsic resistance and inferior responses to tyrosine kinase inhibitors in cancer. Nat Med [Internet]. 2012 Apr [cited 2016 Jun 17];18(4):521–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22426421
6. Baldwin RM, Owzar K, Zembutsu H, Chhibber A, Kubo M, Jiang C, et al. A genome-wide association study identifies novel loci for paclitaxel-induced sensory peripheral neuropathy in CALGB 40101. Clin Cancer Res [Internet]. 2012 Sep 15 [cited 2016 Jun 17];18(18):5099–109. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22843789

7. Innocenti F, Owzar K, Cox NL, Evans P, Kubo M, Zembutsu H, et al. A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in CALGB 80303. Clin Cancer Res [Internet]. 2012 Jan 15 [cited 2016 Jun 17];18(2):577–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22142827

8. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, et al. SLCO1B1 Variants and Statin-Induced Myopathy — A Genomewide Study. N Engl J Med [Internet]. 2008 Aug 21 [cited 2015 Oct 14];359(8):789–99. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18650507

9. Teichert M, Eijgelsheim M, Rivadeneira F, Uitterlinden AG, van Schaik RHN, Hofman A, et al. A genome-wide association study of acenocoumarol maintenance dosage. Hum Mol Genet [Internet]. 2009 Oct 1 [cited 2016 Jun 17];18(19):3758–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19578179

10. Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet [Internet]. 2009 Mar [cited 2016 Jun 17];5(3):e1000433. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19300499

11. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, Schwarz UI, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood [Internet]. 2008 Aug 15 [cited 2016 Jun 17];112(4):1022–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18535201

12. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci U S A [Internet]. 2010 May 18 [cited 2015 Dec 11];107(20):9287–92. Available from: http://www.pnas.org/content/107/20/9287.abstract

13. Pisanu C, Papadima EM, Del Zompo M, Squassina A. Understanding the molecular

mechanisms underlying mood stabilizer treatments in bipolar disorder: Potential involvement of epigenetics. Neurosci Lett. 2016;

14. Reynolds GP, Fachim HA. Does DNA methylation influence the effects of psychiatric drugs? Epigenomics [Internet]. 2016 Mar [cited 2016 Jul 8];8(3):309–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26918935

15. Stark AL, Hause RJ, Gorsic LK, Antao NN, Wong SS, Chung SH, et al. Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy. PLoS Genet [Internet]. 2014 Apr [cited 2016 Jul 8];10(4):e1004192. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24699359

16. *SLCO1B1* Variants and Statin-Induced Myopathy — A Genomewide Study. N Engl J Med [Internet]. 2008 Aug 21 [cited 2016 Jun 17];359(8):789–99. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMoa0801936

17. Zhang JE, Jorgensen AL, Alfirevic A, Williamson PR, Toh CH, Park BK, et al. Effects of CYP4F2 genetic polymorphisms and haplotypes on clinical outcomes in patients initiated on warfarin therapy. Pharmacogenet Genomics [Internet]. 2009 Oct [cited 2016 Jul 10];19(10):781–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19741565

18. Mosley JD, Shaffer CM, Van Driest SL, Weeke PE, Wells QS, Karnes JH, et al. A genome-wide association study identifies variants in KCNIP4 associated with ACE inhibitor-induced cough. Pharmacogenomics J [Internet]. Nature Publishing Group; 2016 Jun 14 [cited 2016 Jul 10];16(3):231–7. Available from: http://www.nature.com/doifinder/10.1038/tpj.2015.51

19. Gamazon ER, Huang RS, Dolan ME, Cox NJ, Gonzalez E, Kulkarni H, et al. Copy number polymorphisms and anticancer pharmacogenomics. Genome Biol [Internet]. BioMed Central; 2011 [cited 2016 Jul 10];12(5):R46. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-5-r46

20. Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark T a, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A [Internet]. 2007 Jun 5;104(23):9758–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1887589&tool=pmcentrez&rendertype=abstract

21. Wang D, Guo Y, Wrighton SA, Cooke GE, Sadee W. Intronic polymorphism in CYP3A4 affects hepatic expression and response to statin drugs. Pharmacogenomics J [Internet]. Nature Publishing Group; 2011 Aug 13 [cited 2016 Jul 10];11(4):274–86. Available from: http://www.nature.com/doifinder/10.1038/tpj.2010.28

22. Suarez-Kurtz G, Vargens DD, Santoro AB, Hutz MH, de Moraes ME, Pena SDJ, et al. Global pharmacogenomics: distribution of CYP3A5 polymorphisms and phenotypes in the Brazilian population. PLoS One [Internet]. Public Library of Science; 2014 [cited 2016 Jul 10];9(1):e83472. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24427273

23. Peters BJM, Rodin AS, de Boer A, Maitland-van der Zee A-H. Methodological and statistical issues in pharmacogenomics. J Pharm Pharmacol [Internet]. 2010 Feb [cited 2016 May 10];62(2):161–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20487194

24. Hansen NT, Brunak S, Altman RB. Generating genome-scale candidate gene lists for pharmacogenomics. Clin Pharmacol Ther [Internet]. 2009 Aug [cited 2016 May 12];86(2):183–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2729176&tool=pmcentrez&rendertype=abstract

25. Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet [Internet]. 2009 Oct [cited 2016 Mar 15];41(10):1100–4. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/19749758

26. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, et al. Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet [Internet]. 2009 Oct [cited 2016 Mar 15];41(10):1105–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19749757

27. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna K V., Urban TJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature [Internet]. Macmillan Publishers Limited. All rights reserved; 2009 Aug 16 [cited 2015 Apr 12];461(7262):399–401. Available from: http://dx.doi.org/10.1038/nature08309

28. Huang RS, Duan S, Shukla SJ, Kistner EO, Clark T a, Chen TX, et al. Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. Am J Hum Genet [Internet]. 2007 Sep [cited 2013 Mar 11];81(3):427–37. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950832&tool=pmcentrez&rendertype=abstract

29. Wheeler HE, Gamazon ER, Stark a L, O'Donnell PH, Gorsic LK, Huang RS, et al. Genome-wide meta-analysis identifies variants associated with platinating agent susceptibility across populations. Pharmacogenomics J [Internet]. 2011 Aug 16 [cited 2015 Apr 12];13(1):35–43. Available from: http://www.nature.com/doifinder/10.1038/tpj.2011.38

30. Gamazon ER, Lamba JK, Pounds S, Stark AL, Wheeler HE, Cao X, et al. Comprehensive genetic analysis of cytarabine sensitivity in a cell-based model identifies polymorphisms associated with outcome in AML patients. Blood [Internet]. 2013 May 23 [cited 2013 Aug 19];121(21):4366–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23538338

31. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of

gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet [Internet]. 2012 May [cited 2016 Mar 28];44(5):502–10. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3437404&tool=pmcentrez&rendertype=abstract

32. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet [Internet]. Public Library of Science; 2011 May 26 [cited 2016 Mar 2];7(5):e1002078. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002078

33. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet [Internet]. 2010 Apr [cited 2015 Oct 8];6(4):e1000888. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2848547&tool=pmcentrez&rendertype=abstract

34. Kaddurah-Daouk R, Weinshilboum R. Metabolomic Signatures for Drug Response Phenotypes: Pharmacometabolomics Enables Precision Medicine. Clin Pharmacol Ther [Internet]. 2015 Jul [cited 2016 Apr 14];98(1):71–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25871646

35. Hause RJ, Stark AL, Antao NN, Gorsic LK, Chung SH, Brown CD, et al. Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. Am J Hum Genet [Internet]. 2014 Aug 7 [cited 2016 Apr 8];95(2):194–208. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4129400&tool=pmcentrez&rendertype=abstract

36. Kurita M, Holloway T, García-Bea A, Kozlenkov A, Friedman AK, Moreno JL, et al. HDAC2 regulates atypical antipsychotic responses through the modulation of mGlu2 promoter activity. Nat Neurosci [Internet]. 2012 Oct [cited 2016 Mar 31];15(9):1245–54. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431440&tool=pmcentrez&rendertype=abstract

37. Boks MP, de Jong NM, Kas MJH, Vinkers CH, Fernandes C, Kahn RS, et al. Current status and future prospects for epigenetic psychopharmacology. Epigenetics [Internet]. 2012 Jan 1 [cited 2016 Apr 2];7(1):20–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3329498&tool=pmcentrez&rendertype=abstract

38. Csoka AB, Szyf M. Epigenetic side-effects of common pharmaceuticals: a potential new field in medicine and pharmacology. Med Hypotheses [Internet]. 2009 Nov [cited 2016 Mar 16];73(5):770–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19501473

39. He Y, Hoskins JM, McLeod HL. Copy number variants in pharmacogenetic genes. Trends Mol Med [Internet]. 2011 May [cited 2016 Apr 8];17(5):244–51. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3092840&tool=pmcentrez&rendertype=abstract

40. Kim I-W, Han N, Kim MG, Kim T, Oh JM. Copy number variability analysis of pharmacogenes in patients with lymphoma, leukemia, hepatocellular, and lung carcinoma using The Cancer Genome Atlas data. Pharmacogenet Genomics [Internet]. 2015 Jan [cited 2016 Apr 8];25(1):1–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25379720

41. Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao Y-T, et al. Addressing population-specific multiple testing burdens in genetic association studies. Ann Hum

Genet [Internet]. 2015 Mar [cited 2016 May 15];79(2):136–47. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4334751&tool=pmcentrez&rendertype=abstract

42. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet [Internet]. Nature Publishing Group; 2009 Oct [cited 2014 Jul 16];10(10):681–90. Available from: http://dx.doi.org/10.1038/nrg2615

43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet [Internet]. 2007 Sep [cited 2014 Jul 10];81(3):559–75. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838&tool=pmcentrez&rendertype=abstract

44. Ritchie M. Plato [Internet]. Available from: https://ritchielab.psu.edu/plato

45. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics [Internet]. Oxford University Press; 2007 May 15 [cited 2016 Jul 10];23(10):1294–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17384015

46. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet [Internet]. Nature Publishing Group; 2012 Jun 17 [cited 2016 Jul 10];44(7):821–4. Available from: http://www.nature.com/doifinder/10.1038/ng.2310

47. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods [Internet]. Nature Publishing Group; 2011 Sep 4 [cited 2016 Jul 10];8(10):833–5. Available from: http://www.nature.com/doifinder/10.1038/nmeth.1681

48. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays [Internet].

2005 Jun [cited 2016 Apr 13];27(6):637–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15892116

49. Mackay TFC. Epistasis and quantitative traits : using model organisms to study gene – gene interactions. Nat Publ Gr [Internet]. Nature Publishing Group; 2013;15(1):22–33. Available from: http://dx.doi.org/10.1038/nrg3627

50. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet [Internet]. 2005 Apr [cited 2016 Apr 9];37(4):413–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15793588

51. Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, et al. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Ann Hum Genet [Internet]. 2000 Sep [cited 2016 Apr 9];64(Pt 5):413–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11281279

52. Millstein J, Conti D V, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet [Internet]. 2006 Jan [cited 2016 Apr 9];78(1):15–27. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1380213&tool=pmcentrez&rendertype=abstract

53. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet [Internet]. Nature Publishing Group; 2009 Jun [cited 2016 Feb 20];10(6):392–404. Available from: http://dx.doi.org/10.1038/nrg2579

54. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet [Internet]. 2001 Jul [cited 2016 Feb 10];69(1):138–47. Available from: http://www.sciencedirect.com/science/article/pii/S0002929707614530

55. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics [Internet]. 2003 Feb 12 [cited 2016 Mar 3];19(3):376–82. Available from: http://bioinformatics.oxfordjournals.org/content/19/3/376.short

56. Kim M-K, Moore JH, Kim J-K, Cho K-H, Cho Y-W, Kim Y-S, et al. Evidence for epistatic interactions in antiepileptic drug resistance. J Hum Genet [Internet]. The Japan Society of Human Genetics; 2011 Jan [cited 2016 Apr 9];56(1):71–6. Available from: http://dx.doi.org/10.1038/jhg.2010.151

57. Ritchie MD, Motsinger AA. Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. Pharmacogenomics [Internet]. Future Medicine Ltd London, UK; 2005 Dec 21 [cited 2016 Apr 9];6(8):823–34. Available from: http://www.futuremedicine.com/doi/abs/10.2217/14622416.6.8.823?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed

58. Lou X-Y, Chen G-B, Yan L, Ma JZ, Zhu J, Elston RC, et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet [Internet]. 2007 Jun [cited 2016 May 26];80(6):1125–37. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1867100&tool=pmcentrez&rendertype=abstract

59. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet [Internet]. 2010 Sep 10 [cited 2016 May 9];87(3):325–40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2933337&tool=pmcentrez&rendertype=abstract

60. Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, Erhardt-Lehmann A, et al. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. Eur J Hum Genet [Internet]. 2011 Apr [cited 2016 May 26];19(4):465–71. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3060319&tool=pmcentrez&rendertype=abstract

61. Wang Z, Wang Y, Tan K-L, Wong L, Agrawal D. eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study. Bioinformatics [Internet]. 2011 Apr 15 [cited 2016 Apr 19];27(8):1045–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21367868

62. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nat Genet [Internet]. Nature Publishing Group; 2007 Sep [cited 2016 Apr 9];39(9):1167–73. Available from: http://dx.doi.org/10.1038/ng2110

63. Holzinger ER, Dudek SM, Frase AT, Pendergrass S a, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. Bioinformatics [Internet]. 2013 Oct 27;1–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24149050

64. Li R, Dudek SM, Kim D, Hall MA, Bradford Y, Peissig PL, et al. Identification of genetic interaction networks via an evolutionary algorithm evolved Bayesian network. BioData Min [Internet]. BioMed Central; 2016 Jan 10 [cited 2016 May 16];9(1):18. Available from: http://biodatamining.biomedcentral.com/articles/10.1186/s13040-016-0094-4

65. Beam AL, Motsinger-Reif A, Doyle J. Bayesian neural networks for detecting epistasis in genetic association studies. BMC Bioinformatics [Internet]. BioMed Central; 2014 Jan 21 [cited 2016 Mar 3];15(1):368. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0368-0

66. Schwarz DF, König IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics [Internet]. 2010 Jul 15 [cited 2016 May 26];26(14):1752–8. Available from: http://bioinformatics.oxfordjournals.org/content/26/14/1752.short

67. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. arXiv Prepr [Internet]. 2015; Available from: http://arxiv.org/abs/1508.04409

68. Moore JH, White BC. Tuning ReliefF for genome-wide genetic analysis. Springer-Verlag; 2007 Apr 11 [cited 2016 Apr 9];166–75. Available from: http://dl.acm.org/citation.cfm?id=1761486.1761502

69. Robnik-Šikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Mach Learn [Internet]. Kluwer Academic Publishers; [cited 2016 Apr 9];53(1-2):23–69. Available from: http://link.springer.com/article/10.1023/A%3A1025667309714

70. Huang R, Wallqvist A, Thanki N, Covell DG. Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute's anticancer screen and drug mechanism of action. Pharmacogenomics J [Internet]. 2005 Jan 16 [cited 2016 Apr 9];5(6):381–99. Available from: http://dx.doi.org/10.1038/sj.tpj.6500331

71. Lin Y-A, Chiang A, Lin R, Yao P, Chen R, Butte AJ. Methodologies for extracting functional pharmacogenomic experiments from international repository. AMIA Annu Symp Proc [Internet]. 2007 Jan [cited 2016 Apr 9];463–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655846&tool=pmcentrez&rendertype=abstract

72. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput [Internet]. 2009 Jan [cited 2015 Aug 15];368–79. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859610&tool=pmcentrez&rendertype=abstract

73. BJ G, ES T, PJ M, PI DB, DW H, GK R, et al. Use of biological knowledge to inform the analysis of gene-gene interactions involved in modulating virologic failure with efavirenz-containing treatment regimens in ART-naive ACTG clinical trials participants. Pac Symp Biocomput [Internet]. [cited 2016 Apr 10];253–64. Available from: http://europepmc.org/articles/PMC3094912

74. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2014 Jul 10];28(1):27–30. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract

75. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet [Internet]. 2000 May [cited 2014 Jul 10];25(1):25–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract

76. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res [Internet]. 2002 Jan 1 [cited 2016 Jul 10];30(1):207–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11752295

77. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res [Internet]. 2002 Jan 1 [cited 2016 Jul 10];30(1):163–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11752281

78. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet [Internet]. 2009 Jan [cited 2014 Jul 10];10(1):57–63. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract

79. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011 Feb [cited 2014 Jul 9];12(2):87–98. Available from: http://dx.doi.org/10.1038/nrg2934

80. Laird PW. Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet [Internet]. Nature Publishing Group; 2010 Mar 2 [cited 2014 Jul 9];11(3):191–203. Available from: http://dx.doi.org/10.1038/nrg2732

81. Altelaar AFM, Munoz J, Heck AJR. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet [Internet]. 2013 Jan [cited 2016 Mar 29];14(1):35–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23207911

82. Richardson S, Tseng GC, Sun W. Statistical Methods in Integrative Genomics. Annu Rev Stat Its Appl [Internet]. Annual Reviews; 2016 Jun 1 [cited 2016 Jun 2];3(1):181–209. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-041715-033506

83. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics [Internet]. 2009 Jan [cited 2016 Apr 18];2009. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2950414&tool=pmcentrez&rendertype=abstract

84. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet [Internet]. 2010 Jul [cited 2016 May 9];11(7):476–86. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3321268&tool=pmcentrez&rendertype=abstract

85. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics [Internet]. 2009 Nov 15 [cited 2016 Jun 9];25(22):2906–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19759197

86. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature [Internet]. 2012 Jun 21 [cited 2016 Jun 9];486(7403):346–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22522925

87. Visscher H, Ross CJD, Dubé M-P, Brown AMK, Phillips MS, Carleton BC, et al. Application of principal component analysis to pharmacogenomic studies in Canada. Pharmacogenomics J [Internet]. Nature Publishing Group; 2009 Dec 4 [cited 2016 Jun 9];9(6):362–72. Available from: http://www.nature.com/doifinder/10.1038/tpj.2009.36

88. Gao X, Starmer JD, Lander E, Schork N, Risch N, Marchini J, et al. AWclust: point-and-click software for non-parametric population structure analysis. BMC Bioinformatics [Internet]. BioMed Central; 2008 [cited 2016 Jun 12];9(1):77. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-77

89. Galvan A, Fladvad T, Skorpen F, Gao X, Klepstad P, Kaasa S, et al. Genetic clustering of European cancer patients indicates that opioid-mediated pain relief is independent of ancestry. Pharmacogenomics J [Internet]. Nature Publishing Group; 2012 Oct 12 [cited 2016 Jun 12];12(5):412–6. Available from: http://www.nature.com/doifinder/10.1038/tpj.2011.27

90. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015 Jan 13 [cited 2015 Jan 13];16(2):85–97. Available from: http://dx.doi.org/10.1038/nrg3868

91. Holzinger E, Ritchie M. integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. Pharmacogenomics [Internet]. 2012 [cited 2012 Sep 24];213–22. Available from: http://www.futuremedicine.com/doi/pdf/10.2217/pgs.11.145

92. Hartford CM, Duan S, Delaney SM, Mi S, Kistner EO, Lamba JK, et al. Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. Blood [Internet]. 2009 Mar 5 [cited 2013 Apr 5];113(10):2145–53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2652364&tool=pmcentrez&rendertype=abstract

93. Huang RS, Duan S, Kistner EO, Hartford CM, Dolan ME. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Mol Cancer Ther [Internet]. 2008 Sep [cited 2013 May 16];7(9):3038–46. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2743011&tool=pmcentrez&rendertype=abstract

94. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet [Internet]. 2005 Jul [cited 2012 Jul 20];37(7):710–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2841396&tool=pmcentrez&rendertype=abstract

95. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol [Internet]. Nature Publishing Group; 2013 Jan 20 [cited 2016 Jul 14];31(2):142–7. Available from: http://www.nature.com/doifinder/10.1038/nbt.2487

96. Khan Z, Bloom JS, Amini S, Singh M, Perlman DH, Caudy AA, et al. Quantitative

measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. Mol Syst Biol [Internet]. 2012 [cited 2016 Jul 14];8:602. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22893000

97. Wei X, Wang X. A Computational Workflow to Identify Allele-specific Expression and Epigenetic Modification in Maize. Genomics Proteomics Bioinformatics. 2013;11(4):247–52.

98. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen P a C, Monlong J, Rivas M a, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature [Internet]. 2013 Sep 26 [cited 2014 Jan 20];501(7468):506–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24037378

99. Maynard ND, Chen J, Stuart RK, Fan J-B, Ren B. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. Nat Methods [Internet]. 2008 Apr [cited 2014 Jan 14];5(4):307–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18345007

100. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, et al. Extensive variation in chromatin states across humans. Science [Internet]. 2013 Nov 8 [cited 2016 Jul 14];342(6159):750–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24136358

101. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. Science [Internet]. American Association for the Advancement of Science; 2013 Nov 8 [cited 2016 Jul 14];342(6159):747–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24136359

102. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science [Internet]. 2004 Oct 22 [cited 2016 Jul 14];306(5696):636–40. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/15499007

103. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. J Biomed Inform [Internet]. 2012 Dec [cited 2016 Mar 3];45(6):1191–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22910106

104. Holzinger ER, Dudek SM, Torstenson EC, Ritchie MD. ATHENA Optimization : The Effect of Initial Parameter Settings across Different Genetic Models. 2011;48–58.

105. Fridley BL, Lund S, Jenkins GD, Wang L. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet Epidemiol [Internet]. 2012 May [cited 2016 Apr 11];36(4):352–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3894829&tool=pmcentrez&rendertype=abstract

106. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to Recurrence and Survival in Serous Ovarian Tumors Predicted from Integrated Genomic Profiles. Deb S, editor. PLoS One [Internet]. Public Library of Science; 2011 Nov 3 [cited 2016 Jul 14];6(11):e24709. Available from: http://dx.plos.org/10.1371/journal.pone.0024709

107. Holzinger ER, Dudek SM, Frase AT, Pendergrass S a, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. Bioinformatics [Internet]. 2014 Mar 1 [cited 2014 Aug 5];30(5):698–705. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24149050

108. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD. ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. Pac Symp Biocomput [Internet]. 2013 Jan [cited 2013 Jul 30];385–96. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3587764&tool=pmcentrez&rendertype=abstract

109. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData Min [Internet]. BioData Mining; 2013 Jan [cited 2014 Aug 5];6(1):23. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3912499&tool=pmcentrez&rendertype=abstract

110. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer [Internet]. Nature Publishing Group; 2008 Jan [cited 2016 Jul 14];8(1):37–49. Available from: http://www.nature.com/doifinder/10.1038/nrc2294

111. Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics [Internet]. Oxford University Press; 2004 Nov 1 [cited 2016 Jul 14];20(16):2626–35. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bth294

112. Borgwardt KM, Ong CS, Schönauer S, Vishwanathan SVN, Smola AJ, Kriegel H-P. Protein function prediction via graph kernels. Bioinformatics [Internet]. Oxford University Press; 2005 Jun [cited 2016 Jul 14];21 Suppl 1(suppl 1):i47–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15961493

113. Tsuda K, Shin H, Schölkopf B. Fast protein classification with multiple networks. Bioinformatics [Internet]. Oxford University Press; 2005 Sep 1 [cited 2016 Jul 14];21 Suppl 2(suppl 2):ii59–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16204126

114. Shin H, Lisewski AM, Lichtarge O. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics [Internet]. 2007 Dec 1 [cited 2016 Jul 14];23(23):3217–24. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/17977886

115. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData Min [Internet]. 2013 Dec 20 [cited 2014 Jan 15];6(1):23. Available from: http://www.biodatamining.org/content/6/1/23

116. Turner SD, Dudek SM, Ritchie MD. ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. BioData Min [Internet]. 2010 Jan [cited 2013 Jul 30];3(1):5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2955681&tool=pmcentrez&rendertype=abstract

117. Drăghici S, Potter RB. Predicting HIV drug resistance with neural networks. Bioinformatics [Internet]. 2003 Jan [cited 2016 Jul 14];19(1):98–107. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12499299

118. Shen H-B, Chou K-C. Ensemble classifier for protein fold pattern recognition. Bioinformatics [Internet]. Oxford University Press; 2006 Jul 15 [cited 2016 Jul 14];22(14):1717–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16672258

119. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell [Internet]. 2010 Dec 10 [cited 2013 Mar 6];143(6):1005–17. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013278&tool=pmcentrez&rendertype=abstract

120. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet [Internet]. 2008 Jul [cited 2012 Jul 19];40(7):854–61. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2573859&tool=pmcentrez&rendertype=abstract

121. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, et al. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. PLoS Biol [Internet]. 2012 Jan [cited 2012 Jul 17];10(4):e1001301. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3317911&tool=pmcentrez&rendertype=abstract

122. Maclin R, Opitz D. Popular Ensemble Methods: An Empirical Study. J Artif Intell Res [Internet]. 2011 Jun 1 [cited 2016 Jul 14]; Available from: http://arxiv.org/abs/1106.0257

123. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol [Internet]. Nature Publishing Group; 2014 Jun 1 [cited 2016 Jul 11];32(12):1202–12. Available from: http://www.nature.com/doifinder/10.1038/nbt.2877

124. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A [Internet]. 2009 Jun 9 [cited 2014 Dec 22];106(23):9362–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2687147&tool=pmcentrez&rendertype=abstract

125. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol [Internet]. 2008 May [cited 2016 Jul 11];32(4):325–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18265411

126. Breiman L. Random Forests. Mach Learn [Internet]. Kluwer Academic Publishers; [cited

2016 Apr 9];45(1):5–32. Available from: http://link.springer.com/article/10.1023/A%3A1010933404324

127. Ritchie MD, Holzinger ER, Dudek SM, Frase AT, Chalise P, Fridley B. Meta-Dimensional Analysis of Phenotypes Using the Analysis Tool for Heritable and Environmental Network Associations (ATHENA): Challenges with Building Large Networks. Springer New York; 2013 [cited 2016 Jul 11]. p. 103–15. Available from: http://link.springer.com/10.1007/978-1-4614-6846-2_8

128. Skapura DM. Building neural networks. ACM Press/Addison-Wesley Publishing Co.; 1995 Dec 1 [cited 2013 Jul 30]; Available from: http://dl.acm.org/citation.cfm?id=217718

129. Koza JR, Rice JP. Genetic generation of both the weights and architecture for a neural network. IJCNN-91-Seattle International Joint Conference on Neural Networks [Internet]. IEEE; 1991 [cited 2013 Jul 30]. p. 397–404. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=155366

130. Motsinger AA, Lee SL, Mellick G, Ritchie MD. GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics [Internet]. 2006 Jan [cited 2014 Feb 9];7(1):39. Available from: http://www.biomedcentral.com/1471-2105/7/39

131. O'Neill M, Ryan C. Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language [Internet]. Springer; 2003 edition; 2003 [cited 2014 Apr 7]. 144 p. Available from: http://www.amazon.com/Grammatical-Evolution-Evolutionary-Automatic-Programming/dp/1402074441

132. O'Neill M, Ryan C. Grammatical evolution. IEEE Trans Evol Comput [Internet]. IEEE; 2001 [cited 2014 Mar 22];5(4):349–58. Available from: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=942529

133. Pearson B, Lau KH, DeLoache W, Penumetcha P, Rinker VG, Allen A, et al. Bacterial

Hash Function Using DNA-Based XOR Logic Reveals Unexpected Behavior of the LuxR Promoter. Interdiscip Bio Cent [Internet]. 2011 Jul 18 [cited 2014 Mar 20];3(3):1–8. Available from: http://www.ibc7.org/article/journal_v.php?sid=265

134. Privman V, Zhou J, Halámek J, Katz E. Realization and properties of biochemical-computing biocatalytic XOR gate based on signal change. J Phys Chem B [Internet]. 2010 Oct 28 [cited 2014 Apr 7];114(42):13601–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20882987

135. Edwards TL, Bush WS, Turner SD, Dudek SM, Torstenson ES, Schmidt M, et al. Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008 [cited 2016 Jul 11]. p. 24–35. Available from: http://link.springer.com/10.1007/978-3-540-78757-0_3

136. Gibson G, Riley-Berger R, Harshman L, Kopp A, Vacha S, Nuzhdin S, et al. Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. Genetics [Internet]. 2004 Aug [cited 2016 Jul 12];167(4):1791–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15342517

137. Andrew AS, Hu T, Gu J, Gui J, Ye Y, Marsit CJ, et al. HSD3B and gene-gene interactions in a pathway-based analysis of genetic susceptibility to bladder cancer. PLoS One [Internet]. 2012 [cited 2016 Jul 12];7(12):e51301. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23284679

138. Huang RS, Duan S, Kistner EO, Bleibel WK, Delaney SM, Fackenthal DL, et al. Genetic variants contributing to daunorubicin-induced cytotoxicity. Cancer Res [Internet]. 2008 May 1 [cited 2013 May 16];68(9):3161–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2714371&tool=pmcentrez&rendertype=abstract

139. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci U S A [Internet]. 2010 May 18 [cited 2013 Apr 5];107(20):9287–92. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889115&tool=pmcentrez&rendertype=abstract

140. International T, Consortium H. A haplotype map of the human genome. Nature [Internet]. 2005 Oct 27 [cited 2013 Jul 30];437(7063):1299–320. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1880871&tool=pmcentrez&rendertype=abstract

141. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome Biol [Internet]. BioMed Central Ltd; 2012 Jan [cited 2013 May 23];13(2):R8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3334571&tool=pmcentrez&rendertype=abstract

142. Sinha BK, Haim N, Dusre L, Kerrigan D, Pommier Y. DNA strand breaks produced by etoposide (VP-16,213) in sensitive and resistant human breast tumor cells: implications for the mechanism of action. Cancer Res [Internet]. 1988 Sep 15 [cited 2013 Jul 30];48(18):5096–100. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2842045

143. Mistry AR, Felix CA, Whitmarsh RJ, Mason A, Reiter A, Cassinat B, et al. DNA topoisomerase II in therapy-related acute promyelocytic leukemia. N Engl J Med [Internet]. 2005 Apr 14 [cited 2013 Jul 30];352(15):1529–38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15829534

144. Ratain MJ, Kaminer LS, Bitran JD, Larson RA, Le Beau MM, Skosey C, et al. Acute nonlymphocytic leukemia following etoposide and cisplatin combination chemotherapy for advanced non-small-cell carcinoma of the lung. Blood [Internet]. 1987 Nov [cited

2013 Jul 30];70(5):1412–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2822173

145. Thomson.Micromedex. Drug Inf Heal Care Prof 24th ed. 2004;Volume 1.(Content Reviewed by the United States Pharmacopeial Convention, Inc. Greenwood Village, CO.):1326.

146. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012 Sep 6 [cited 2012 Nov 1];489(7414):57–74. Available from: http://dx.doi.org/10.1038/nature11247

147. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet [Internet]. 2013 Oct [cited 2014 Jan 21];45(10):1238–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24013639

148. Liu X, Cheng I, Plummer SJ, Suarez BK, Casey G, Catalona WJ, et al. Fine-mapping of prostate cancer aggressiveness loci on chromosome 7q22-35. Prostate [Internet]. 2011 May 15 [cited 2013 Jul 31];71(7):682–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3027848&tool=pmcentrez&rendertype=abstract

149. Rouget-Quermalet V, Giustiniani J, Marie-Cardine A, Beaud G, Besnard F, Loyaux D, et al. Protocadherin 15 (PCDH15): a new secreted isoform and a potential marker for NK/T cell lymphomas. Oncogene [Internet]. 2006 May 4 [cited 2013 Jul 31];25(19):2807–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16369489

150. Salih MA, Tzschach A, Oystreck DT, Hassan HH, AlDrees A, Elmalik SA, et al. A newly recognized autosomal recessive syndrome affecting neurologic function and vision. Am J

Med Genet A [Internet]. 2013 Jul [cited 2013 Jul 31];161(6):1207–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23633300

151. Akizu N, Shembesh NM, Ben-Omran T, Bastaki L, Al-Tawari A, Zaki MS, et al. Whole-exome sequencing identifies mutated c12orf57 in recessive corpus callosum hypoplasia. Am J Hum Genet [Internet]. 2013 Mar 7 [cited 2013 Jul 31];92(3):392–400. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23453666

152. Jiang Y, Janku F, Subbiah V, Angelo LS, Naing A, Anderson PM, et al. Germline PTPRD Mutations in Ewing Sarcoma: Biologic and Clinical Implications. Oncotarget [Internet]. 2013 Jun [cited 2013 Aug 1];4(6):884–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23800680

153. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature [Internet]. 2009 Oct 8 [cited 2013 Aug 1];461(7265):809–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19812674

154. FACT SHEET: Investing in the National Cancer Moonshot | whitehouse.gov [Internet]. [cited 2016 Apr 22]. Available from: https://www.whitehouse.gov/the-press-office/2016/02/01/fact-sheet-investing-national-cancer-moonshot

155. Wheeler HE, Dolan ME. Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. 2012;13(1):55–70.

156. Borghaei H, Langer CJ, Millenson M, Ruth KJ, Litwin S, Tuttle H, et al. Phase II Study of Paclitaxel, Carboplatin, and Cetuximab as First Line Treatment, for Patients with Advanced Non-small Cell Lung Cancer (NSCLC). J Thorac Oncol [Internet]. 2008 Nov [cited 2015 Jul 2];3(11):1286–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18978564

157. McWhinney SR, Goldberg RM, McLeod HL. Platinum neurotoxicity pharmacogenetics.

Mol Cancer Ther [Internet]. 2009 Jan [cited 2015 Jul 2];8(1):10–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2651829&tool=pmcentrez&rendertype=abstract

158. Rabik CA, Dolan ME. Molecular mechanisms of resistance and toxicity associated with platinating agents. Cancer Treat Rev [Internet]. 2007 Feb [cited 2015 Feb 13];33(1):9–23. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1855222&tool=pmcentrez&rendertype=abstract

159. Cassidy J, Saltz L, Twelves C, Van Cutsem E, Hoff P, Kang Y, et al. Efficacy of capecitabine versus 5-fluorouracil in colorectal and gastric cancers: a meta-analysis of individual data from 6171 patients. Ann Oncol [Internet]. 2011 Dec [cited 2015 Jul 2];22(12):2604–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21415237

160. Kumar CC. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. Genes Cancer [Internet]. 2011 Feb [cited 2015 May 17];2(2):95–107. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3111245&tool=pmcentrez&rendertype=abstract

161. Rowinsky EK, Wright M, Monsarrat B, Donehower RC. Clinical pharmacology and metabolism of Taxol (paclitaxel): update 1993. Ann Oncol [Internet]. 1994 Jan [cited 2015 Jul 2];5 Suppl 6:S7–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7865438

162. Huang RS, Kistner EO, Bleibel WK, Shukla SJ, Dolan ME. Effect of population and gender on chemotherapeutic agent-induced cytotoxicity. Mol Cancer Ther [Internet]. 2007 Jan [cited 2014 Feb 3];6(1):31–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669540&tool=pmcentrez&re

ndertype=abstract

163. Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, et al. Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet [Internet]. 2008 Mar [cited 2015 Jun 14];82(3):631–40. Available from: http://www.sciencedirect.com/science/article/pii/S0002929708001365

164. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet [Internet]. Public Library of Science; 2013 Apr 25 [cited 2015 Jul 3];9(4):e1003449. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003449

165. Abecasis GR, Auton A, Brooks LD, DePristo M a, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature [Internet]. 2012 Nov 1 [cited 2014 Jan 20];491(7422):56–65. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3498066&tool=pmcentrez&rendertype=abstract

166. O'Donnell PH, Stark AL, Gamazon ER, Wheeler HE, McIlwee BE, Gorsic L, et al. Identification of novel germline polymorphisms governing capecitabine sensitivity. Cancer [Internet]. 2012 Aug 15 [cited 2015 Oct 15];118(16):4063–73. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3413892&tool=pmcentrez&rendertype=abstract

167. Wen Y, Gorsic LK, Wheeler HE, Ziliak DM, Huang RS, Dolan ME. Chemotherapeutic-induced apoptosis: a phenotype for pharmacogenomics studies. Pharmacogenet Genomics [Internet]. 2011 Aug [cited 2015 Jul 3];21(8):476–88. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3134538&tool=pmcentrez&rendertype=abstract

168. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res [Internet]. 2010 Sep [cited 2014 Jul 9];20(9):1297–303. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract

169. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. 2006;38(8):904–9.

170. Wheeler HE, Gamazon ER, Wing C, Njiaju UO, Njoku C, Baldwin RM, et al. Integration of cell line and clinical trial genome-wide analyses supports a polygenic architecture of Paclitaxel-induced sensory peripheral neuropathy. Clin Cancer Res [Internet]. 2013 Jan 15 [cited 2015 Dec 10];19(2):491–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549006&tool=pmcentrez&rendertype=abstract

171. Ricci MS. Chemotherapeutic Approaches for Targeting Cell Death Pathways. Oncologist [Internet]. 2006 Apr 1 [cited 2015 Sep 7];11(4):342–57. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3132471&tool=pmcentrez&rendertype=abstract

172. Brown JM, Attardi LD. The role of apoptosis in cancer development and treatment response. Nat Rev Cancer [Internet]. 2005 Mar [cited 2016 Jan 20];5(3):231–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15738985

173. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-wide variation of Cytosine modifications between European and african populations and the implications for complex traits. Genetics [Internet]. 2013 Aug [cited 2013 Aug

9];194(4):987–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23792949

174. Huang RS, Kistner EO, Bleibel WK, Shukla SJ, Dolan ME. Effect of population and gender on chemotherapeutic agent-induced cytotoxicity. Mol Cancer Ther [Internet]. 2007 Jan [cited 2013 Jul 18];6(1):31–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669540&tool=pmcentrez&rendertype=abstract

175. Quintela-Fandino M, Arpaia E, Brenner D, Goh T, Yeung FA, Blaser H, et al. HUNK suppresses metastasis of basal type breast cancers by disrupting the interaction between PP2A and cofilin-1. Proc Natl Acad Sci U S A [Internet]. 2010 Feb 9 [cited 2016 Jan 20];107(6):2622–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2823890&tool=pmcentrez&rendertype=abstract

176. Sinilnikova OM, McKay JD, Tavtigian S V, Canzian F, DeSilva D, Biessy C, et al. Haplotype-based analysis of common variation in the acetyl-coA carboxylase alpha gene and breast cancer risk: a case-control study nested within the European Prospective Investigation into Cancer and Nutrition. Cancer Epidemiol Biomarkers Prev [Internet]. 2007 Mar [cited 2016 Jan 20];16(3):409–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17372234

177. Amend K, Hicks D, Ambrosone CB. Breast cancer in African-American women: differences in tumor biology from European-American women. Cancer Res [Internet]. 2006 Sep 1 [cited 2015 Jul 3];66(17):8327–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16951137

178. Liu H, Yang Y, Xiao J, Yang S, Liu Y, Kang W, et al. Semaphorin 4D expression is associated with a poor clinical outcome in cervical cancer patients. Microvasc Res [Internet]. 2014 May [cited 2015 Dec 11];93:1–8. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/24603190

179. Shen Y-M, He X, Deng H-X, Xie Y-P, Wang C-T, Wei Y-Q, et al. Overexpression of the hBiot2 gene is associated with development of human cervical cancer. Oncol Rep [Internet]. 2011 Jan [cited 2015 Dec 11];25(1):75–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21109960

180. Rositch AF, Nowak RG, Gravitt PE. Increased age and race-specific incidence of cervical cancer after correction for hysterectomy prevalence in the United States from 2000 to 2009. Cancer [Internet]. 2014 Jul 1 [cited 2015 Jul 3];120(13):2032–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4073302&tool=pmcentrez&rendertype=abstract

181. Israël A. The IKK complex, a central regulator of NF-kappaB activation. Cold Spring Harb Perspect Biol [Internet]. 2010 Mar [cited 2015 Feb 5];2(3):a000158. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2829958&tool=pmcentrez&rendertype=abstract

182. Karin M. NF- B as a Critical Link Between Inflammation and Cancer. Cold Spring Harb Perspect Biol [Internet]. 2009 Sep 30 [cited 2015 Jun 28];1(5):a000141–a000141. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2773649&tool=pmcentrez&rendertype=abstract

183. Oh CS, Toke DA, Mandala S, Martin CE. ELO2 and ELO3, homologues of the Saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. J Biol Chem [Internet]. 1997 Jul 11 [cited 2015 Jul 3];272(28):17376–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9211877

184. Kuhajda FP. Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. Nutrition [Internet]. 2000 Mar [cited 2015 Jul 3];16(3):202–8. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/10705076

185. Kuhajda FP, Pizer ES, Li JN, Mani NS, Frehywot GL, Townsend CA. Synthesis and antitumor activity of an inhibitor of fatty acid synthase. Proc Natl Acad Sci U S A [Internet]. 2000 Mar 28 [cited 2015 Jul 3];97(7):3450–4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16260&tool=pmcentrez&rendertype=abstract

186. Reymond N, d'Água BB, Ridley AJ. Crossing the endothelial barrier during metastasis. Nat Rev Cancer [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013 Dec [cited 2015 Jun 3];13(12):858–70. Available from: http://dx.doi.org/10.1038/nrc3628

187. Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy V V, Dolan ME, Huang RS, et al. Poly-omic prediction of complex traits: OmicKriging. Genet Epidemiol [Internet]. 2014 Jul [cited 2016 Feb 15];38(5):402–15. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4072756&tool=pmcentrez&rendertype=abstract

188. Stark AL, Zhang W, Zhou T, O'Donnell PH, Beiswanger CM, Huang RS, et al. Population differences in the rate of proliferation of international HapMap cell lines. Am J Hum Genet [Internet]. 2010 Dec 10 [cited 2016 Feb 18];87(6):829–33. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2997375&tool=pmcentrez&rendertype=abstract

189. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet [Internet]. Annual Reviews; 2008 Jan 3 [cited 2014 Jul 9];9:387–402. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.9.081307.164359

190. Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of

arrhythmia risk. Circulation [Internet]. 2013 Apr 2 [cited 2015 Jul 23];127(13):1377–85. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3713791&tool=pmcentrez&rendertype=abstract

191. Maher B. Personal genomes: The case of the missing heritability. Nature [Internet]. Nature Publishing Group; 2008 Nov 6 [cited 2015 Jan 9];456(7218):18–21. Available from: http://www.nature.com/news/2008/081105/full/456018a.html

192. Hall MA, Verma SS, Wallace J, Lucas A, Berg RL, Connolly J, et al. Biology-Driven Gene-Gene Interaction Analysis of Age-Related Cataract in the eMERGE Network. Genet Epidemiol [Internet]. 2015 Jul [cited 2015 Jul 23];39(5):376–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25982363

193. Jiang X, Barmada MM, Visweswaran S. Identifying genetic interactions in genome-wide data using Bayesian networks. Genet Epidemiol [Internet]. 2010 Sep [cited 2016 Feb 16];34(6):575–81. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3931553&tool=pmcentrez&rendertype=abstract

194. Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, et al. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics [Internet]. BioMed Central; 2012 Jan 15 [cited 2016 Feb 23];13(1):164. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-164

195. Greene CS, Penrod NM, Kiralis J, Moore JH. Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Min [Internet]. 2009 Jan [cited 2016 Mar 7];2(1):5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2761303&tool=pmcentrez&re

ndertype=abstract

196. Han B, Park M, Chen X. A Markov blanket-based method for detecting causal SNPs in GWAS. BMC Bioinformatics [Internet]. 2010 Jan [cited 2016 Mar 21];11 Suppl 3:S5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2863064&tool=pmcentrez&rendertype=abstract

197. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Per Med [Internet]. Future Medicine Ltd London, UK; 2005 Mar 24 [cited 2015 Jul 23];2(1):49–79. Available from: http://www.futuremedicine.com/doi/abs/10.1517/17410541.2.1.49

198. Friedman N. Inferring cellular networks using probabilistic graphical models. Science [Internet]. 2004 Feb 6 [cited 2014 Apr 7];303(5659):799–805. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14764868

199. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science [Internet]. 2005 Apr 22 [cited 2014 Mar 20];308(5721):523–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15845847

200. Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR. Insights into protein-protein interfaces using a Bayesian network prediction method. J Mol Biol [Internet]. 2006 Sep 15 [cited 2014 Apr 7];362(2):365–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16919296

201. Cooper GF, Hennings-yeomans P, Visweswaran S, Barmada M. An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data. 2010;127–31.

202. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks

from data. Mach Learn [Internet]. 1992 Oct [cited 2014 Apr 7];9(4):309–47. Available from: http://link.springer.com/10.1007/BF00994110

203. Schwarz G. Estimating the Dimension of a Model. Ann Stat [Internet]. Institute of Mathematical Statistics; 1978 Mar 1 [cited 2015 Jul 24];6(2):461–4. Available from: http://projecteuclid.org/euclid.aos/1176344136

204. Sun X, Lu Q, Mukherjee S, Mukheerjee S, Crane PK, Elston R, et al. Analysis pipeline for the epistasis search - statistical versus biological filtering. Front Genet [Internet]. 2014 Jan [cited 2016 Mar 1];5:106. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4012196&tool=pmcentrez&rendertype=abstract

205. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc [Internet]. Jan [cited 2015 Jul 27];19(2):212–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3277617&tool=pmcentrez&rendertype=abstract

206. Nadkarni GN, Gottesman O, Linneman JG, Chase H, Berg RL, Farouk S, et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. AMIA Annu Symp Proc [Internet]. 2014 Jan [cited 2015 Jul 27];2014:907–16. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4419875&tool=pmcentrez&rendertype=abstract

207. Gloyn AL, Braun M, Rorsman P. Type 2 diabetes susceptibility gene TCF7L2 and its role in beta-cell function. Diabetes [Internet]. 2009 Apr [cited 2015 Jul 23];58(4):800–2. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2661580&tool=pmcentrez&rendertype=abstract

208. George NM, Day CE, Boerner BP, Johnson RL, Sarvetnick NE. Hippo signaling regulates pancreas development through inactivation of Yap. Mol Cell Biol [Internet]. 2012 Dec [cited 2015 Jul 23];32(24):5116–28. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3510525&tool=pmcentrez&rendertype=abstract

209. An Y, Kang Q, Zhao Y, Hu X, Li N. Lats2 modulates adipocyte proliferation and differentiation via hippo signaling. PLoS One [Internet]. Public Library of Science; 2013 Jan 16 [cited 2015 Jul 23];8(8):e72042. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0072042

210. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS One [Internet]. Public Library of Science; 2012 Jan 14 [cited 2015 Jun 13];7(12):e51954. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0051954

211. Miyashita A, Arai H, Asada T, Imagawa M, Matsubara E, Shoji M, et al. Genetic association of CTNNA3 with late-onset Alzheimer's disease in females. Hum Mol Genet [Internet]. 2007 Dec 1 [cited 2015 Jul 23];16(23):2854–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17761686

212. van Hengel J, Calore M, Bauce B, Dazzo E, Mazzotti E, De Bortoli M, et al. Mutations in the area composita protein αT-catenin are associated with arrhythmogenic right ventricular cardiomyopathy. Eur Heart J [Internet]. 2013 Jan [cited 2015 Jul 23];34(3):201–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23136403

213. Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, et al. Genome-wide

trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Mar [cited 2015 May 28];46(3):234–44. Available from: http://dx.doi.org/10.1038/ng.2897

VITA
Ruowang Li

Education

2011-Present	The Pennsylvania State University	**PhD in Bioinformatics and Genomics** Advisor: Dr. Marylyn D Ritchie
2014-Present	The Pennsylvania State University	**Master in Applied Statistics**
2007-2011	Worcester Polytechnic Institute	**B.S., Biology & Biotechnology** **Minor in Computer Science** with High Distinction

Scholarships, Awards & Fellowships

2013-2016	National Science Foundation Graduate Fellowship (GRFP)
2011-2013	Graham Endowed Fellowship
2011-2012	Huck Institute of Life Sciences Fellowship
2011-2012	Excellence in Graduate Recruitment award
2011	University Graduate Fellowship
2011	Provost's Major Qualifying Project Best Thesis award
2007-2011	Chemistry and Biochemistry Fellowship

Selected Publications

1. **Li, R.,** Dudek, S., Kim, D., Hall, A.M., Bradford, Y., Peissig, P., Brilliant, M., Linneman, J., McCarty, A.C., Bao, L., Ritchie, D.M. (2016). Identification of genetic interaction networks via an evolutionary algorithm evolved bayesian network. *BioData Mining* doi: 10.1186/s13040-016-0094-4
2. **Li, R.**, Kim, D., Dudek, S., Wheeler, H., Dolan, E., Ritchie, D.M. Integration of genetic and functional genomics data to uncover chemotherapeutic induced cytotoxicity. (submitted)
3. Kim, D., **Li, R.**, Dudek, S., Ritchie, D.M. (2015). Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. *Journal of Biomedical Informatics*. doi:10.1016/j.jbi.2015.05.019
4. Ritchie, D.M., Holzinger, R.E., **Li, R.**, Pendergrass, A.S., Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16(2), 85–97.
5. Kim, D., **Li, R.,** Dudek, S., Wallace, R.J., Ritchie, D.M. (2015). Binning somatic mutations based on biological knowledge for predicting survival: an application in renal cell carcinoma. *Pacific Symposium on Biocomputing*, 96–107.
6. Kim, D., **Li, R.**, Dudek, S., Frase, A.A., Pendergrass, A.S., Ritchie, D.M. (2014). Knowledge-driven genomic interactions: an application in ovarian cancer. *BioData Mining* doi:10.1186/1756-0381-7-20
7. **Li, R.,** Kim, D., Dudek, S., Ritchie, D.M. (2014) An Integrated Analysis of Genome-Wide DNA Methylation and Genetic Variants Underlying Etoposide-Induced Cytotoxicity in European and African Populations. *Applications of Evolutionary Computation (Vol. 8602).* doi:10.1007/978-3-662-45523-4
8. Kim, D., **Li, R.**, Dudek, S., Ritchie, D.M. (2013) ATHENA: Identifying interactions between different levels of genomic data using grammatical evolution neural network. *BioData mining* doi:10.1186/1756-0381-6-23
9. **Li, R.,** Holzinger, R.E., Dudek, S., Ritchie, D.M. (2013). Evaluation of parameter contribution to neural network size and fitness in ATHENA for genetic analysis. *Genetic Programming Theory And Practice. Springer*
10. Yildirim, O., **Li, R.**, Hung, J., Chen, B.P., Dong, X., Ee, L., Weng, Z., Rando, J.O., Fazzio, G.T. (2011). Mbd3/NURD Complex Regulates Expression of 5-Hydroxymethylcytosine Marked Genes in Embryonic Stem Cells. *Cell*. 147 (2011), pp. 1498–1510.
11. Carone, R.B., Fauquier, L., Habib, N., Shea, M.J., Hart, E.C., **Li, R.**, Bock, C., Li, C., Zamore, D.P., Meissner, A., Weng, Z., Hofmann, A.H., Friedman, N., Rando, J.O. (2010). Paternally-induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell.* 143, 1084–1096.

Oral Presentations

1. Integration of genetic and functional genomics data to uncover chemotherapeutic induced cytotoxicity. P-STAR Pharmacogenomics Analysis Workshop. State College, PA, 2015
2. An integrated analysis of genome-wide DNA methylation and genetic variants underlying etoposide-induced cytotoxicity in European and African populations. Granada, Spain, 2014.
3. A genome-wide integrated analysis of chemotherapeutic-induced cytotoxicity in European and African populations. P-STAR Pharmacogenomics Analysis Workshop. Nashville, TN, 2013.
4. Evaluation of parameter contribution to neural network size and fitness in ATHENA for genetic analysis. Genetic Programming Theory And Practice Conference. Ann Arbor, Michigan, 2013
5. Integration of "-omics" data. Epistasis Discovery in Genetic Epidemiology Conference. Key West, Florida, 2013.

Book

Editor, Big Data Analytics in Bioinformatics and Healthcare. S.l.: Medical Infor Science Igi, 2014. Print. ISBN

1466666110