# How computational thought experiments can improve our understanding of the genetic architecture of common human diseases

Jason H. Moore[1], Randal S. Olson[1], Peter Schmitt[1], Yong Chen[1] and Elisabetta Manduchi[1]

[1]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
jhmoore@upenn.edu

## Abstract

Susceptibility to common human diseases such as cancer is influenced by many genetic and environmental factors that work together in a complex manner. The state-of-the-art is to perform a genome-wide association study (GWAS) that measures millions of single-nucleotide polymorphisms (SNPs) throughout the genome followed by a one-SNP-at-a-time statistical analysis to detect univariate associations. This approach has identified thousands of genetic risk factors for hundreds of diseases. However, the genetic risk factors detected have very small effect sizes and collectively explain very little of the overall heritability of the disease. Nonetheless, it is assumed that the genetic component of risk is due to many independent risk factors that contribute additively. The fact that many genetic risk factors with small effects can be detected is taken as evidence to support this notion. It is our working hypothesis that the genetic architecture of common diseases is partly driven by non-additive interactions. To test this hypothesis, we developed a heuristic simulation-based method for conducting thought experiments about the complexity of genetic architecture. We show that a genetic architecture driven by complex interactions is highly consistent with the magnitude and distribution of univariate effects seen in real data. We compare our results with measures of univariate and interactions effects from two large-scale GWAS studies of sporadic breast cancer and find evidence to support our hypothesis that is consistent with the results of our thought experiment.

## Introduction

Genetic architecture has been defined as the number of DNA sequence variations (i.e. genetic variants) that influence a phenotype such as presence or absence of disease, the frequency of the alleles and genotypes at those variants in the population, and the role that those variants play in determining the phenotype (Weiss 1995). There are many different ways in which a particular genetic variant can influence phenotypic variability (Thornton-Wells et al. 2004). The variant can have an effect that is independent of all other factors and thus additive. The variant can have an effect in some subjects but not others (i.e. genetic heterogeneity). The variant effect can be dependent on one or more environmental factors such as diet or smoking (i.e. gene-environment interaction). The variant effect can also be dependent on one or more other genetic factors (i.e. gene-gene interaction). We focus here on gene-gene interaction or epistasis as an important component of genetic architecture.

The word epistasis was coined by Bateson in the early 1900s to describe observed deviations from Mendelian segregation ratios (Bateson 1909). The literal translation is one gene standing upon another gene to influence a phenotype. There is widespread evidence of epistasis in model systems including bacteria (Remold and Lenski 2004), yeast (Yadav and Sinha 2018), worms (Gaertner et al. 2012), and flies (Mackay 2014). Epistasis is expected to be a ubiquitous component of genetic architecture in humans (Moore 2003) but has been more difficult to study because of the increase in complexity and the limited ability to experiments in controlled environments and with controlled genetic backgrounds (Jansen 2003). Despite these challenges, epistasis has been detected in population-based studies of human disease with evidence of replication. For example, epistatic effects on lipid variability have been detected and replicated in human populations (De et al. 2017; Holzinger et al. 2017). Verma detected and replicated epistatic effects on susceptibility to glaucoma (Verma et al. 2016). These studies are becoming more numerous as the computational and statistical methods required to detect epistasis improve.

Despite the importance of epistasis, the prevailing approach is to measure genetic variants on a genome-wide scale and then assess them individually as independent risk factors for common human diseases. This genome-wide association study (GWAS) approach has been successful at identifying a number of genetic risk factors but most have very small effect sizes. For example, Michailidou identified 65 new genetic variants for sporadic breast cancer using the GWAS approach in a sample of more than 200,000 women (Michailidou et al. 2017). However, these new variants have effect sizes with odds ratios on the order of 1.1 or less and thus individually do not account for much of the risk of breast cancer. What has been observed for GWAS is that the effect sizes trail off in an exponential fashion such that only a few variants have a moderate effect on risk while most have a very small effect. The success of GWAS has emboldened the idea of the infinitesimal model that assumes that phenotypic variation is the result of many variants each with a small additive effect (Barton et al. 2017). This model discounts complexities such as epistasis and gene-environment interaction.

There were two primary objectives of the present study. Our first goal was to carry out a computational thought experiment to determine whether the distribution of effect sizes seen in GWAS could be consistent with a complex genetic architecture driven by epistasis. We used a novel

computational discovery engine driven by genetic programming to evolve simulation models that produce data and analytical results consistent with the stated objective. Our second goal was to carry out an analysis of real GWAS data from population-based studies of sporadic breast cancer to determine whether types of complex patterns observed in the simulation studies were also found in real data.

## Methods

### Computational thought experiments

The primary goal of our study is to test the assumption that the observed exponential decay of GWAS effect sizes implies that genetic architecture for common human diseases is simply a sum of independent genetic risk factors (i.e. the infinitesimal model). To test this hypothesis, we modified the recently developed Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) method to generate simulation models and data that match the univariate effects observed in GWAS as closely as possible while maximizing gene-gene interactions.

There are five components to our HIBACHI simulation method. The first is the metaphor for the hierarchical biological framework that transmits information from genotype at the DNA sequence level through biomolecular interactions at the gene, cell, and pathway levels to a clinical endpoint. The second is the mathematical framework that generates the genotype to phenotype relationship or pattern. The third is the liability threshold model that is used to define disease status. The fourth is the genetic programming (GP) methods for the discovery of high-order models. The final component is a modification of the fitness function to carry out the thought experiments described here. We describe each of these in turn. The first three components are descriptions included in previous work on HIBACHI (Moore et al. 2015). The fourth component is included recent work on incorporating genetic programming into HIBACHI (Moore et al. 2018). Both sets of descriptions are included here for completeness. The last component is new and describes the additions for the present study. We present each of these in turn below. All simulated data and models are available upon request.

**A biology-based framework for simulation of complex biological systems.** The goal of this component is to provide a biological framework or scaffold to serve as a metaphor for genetic variants and their phenotypic relationships propagated through a hierarchical set of mathematical functions. The prototype for HIBACHI was developed previously and used a fixed architecture that was based on genetic effects at the gene level (Moore et al. 2015). We describe this framework first and then discuss how this relates to the new approach that uses GP to discover both the wiring diagram and the mathematical functions. The initial HIBACHI framework started with protein-coding gene (i.e. mRNA gene) with a single non-synonymous genetic variant that is assumed to change an amino acid. Upstream of the mRNA gene are a promoter region with a single regulatory variant and an enhancer region with a single regulatory variant. Also included in our initial framework are two genes that code for transcription factors that bind to the regulatory region. We included a protein-coding variant in the gene that codes for each transcription factor. We also included a single variant in a microRNA gene that participates in post-translational regulation. In total, this structure allowed for six genetic variants (coded 0, 1, 2) all influencing a protein product as a quantitative trait. In addition, we included an environmental factor (coded -2, -1, 0, 1, 2) to allow for non-genetic variation in the phenotypic values. It is important to note that this particular biological framework was a preliminary proof of concept. The goal of the present study is to allow this framework to vary as part of the search for models meeting certain objectives using GP. The metaphor still holds but the new GP-based systems allows for much greater flexibility in the size and shape of the models being generated. Other metaphors such as electronic health record (EHR) data could also be used here.

**A mathematical framework for simulation of complex patterns.** The goal of this component is to provide a flexible mathematical framework for combining features to produce an endpoint. Using the genetics metaphor, genotypic and non-genotypic values to produce phenotypic values. Each biology-based locus feeds into a mathematical function whose result is carried forward to the next function. For example, one transcription factor locus combines with the enhancer locus through a function whose result then combines with the second transcription factor. The result of this operation combines with the locus at the promoter. This result combines with the coding variant in the gene. This result combines with the microRNA locus. This result combines with the environmental factor to produce a protein product. Thus, the protein expression value is dependent on mathematical functions of the six loci and the environmental factor. This produces a distribution with several to millions of possible phenotypic values for most combinations of functions that can then be used with the liability threshold model described below to generate disease status. Cases and controls can then be sampled from this distribution or the continuous output can be used directly as a quantitative trait.

For each run the user can specify a set of mathematical function to use to build the models. Examples of basic functions include addition, subtraction, multiplication, division, modulus, and modulus-2. Logical functions include greater than, less than, AND, OR, and XOR. Bitwise functions include bitwise AND, bitwise OR, and bitwise XOR. Unary functions include absolute value, NOT, factorial, left and right. Large functions include power, log, permute, and choose. Miscellaneous functions include minimum and maximum. Functions such as XOR are known to produce patterns of epistasis (Li and Reich 2000; Moore and Williams 2009).

**A liability threshold model for biology-based simulation.** We use a liability threshold model to simulate disease from

the distribution of phenotypic values generated from the genotypic values and mathematical functions as described above. The user can select the liability threshold to achieve a particular disease prevalence. More details about the liability model have been provided previously (Moore et al. 2015).

**Genetic programming for model representation and stochastic search.** We selected genetic programming (GP) as our stochastic search engine for several reasons. First, GP uses binary expression trees that are a convenient data structure for representing HIBACHI models. This makes the models very easy to manipulate and evaluate computationally. Flexible representation is a known advantage of GP (Ashlock 2005). Second, GP uses a recombination operator that explicitly swaps subcomponents of expression trees to generate variability in the solutions as part of its iterative search process. This is appealing because HIBACHI models are hierarchical in nature with modular subcomponents representing different biological processes such as transcription factor binding, miRNA regulation of transcription, etc. The ability to mix and match these genomic modules facilitates the development of new models that meet a simulation objective. An introduction to GP has been previously provided in an open-access book for those seeking additional details of the method (Poli et al. 2008). Our initial implementation of the GP approach was for generating HIBACHI models to evaluate and compare machine learning methods (Moore et al. 2018). Our GP implementation used the Distributed Evolutionary Algorithms in Python framework (Fortin et al. 2012) available on GitHub at https://github.com/deap. We describe below our modification of the fitness function to carry out the thought experiments. This new version of HIBACHI is available on GitHub at https://github.com/EpistasisLab/hibachi.

**A fitness function for genetic architecture thought experiments.** A key to GP search is the fitness function that specifies the value or quality of a set of mathematical functions and their wiring (i.e. the model) represented as an expression tree. We had three fitness objectives. Our first objective was to generate simulated data such that the univariate genetic effects matched as closely as possible a set of targets with odds ratios similar to those observed in real data. This fitness component was computed as a sum of absolute differences between the target odds ratios and the odds ratios measured from the solution being evaluated. The odds ratio is a common epidemiologic measure of association between some exposure (e.g. genotype) and presence of disease in a sample of subjects derived from a population-based study. Our second objective was to maximize the three-way gene-gene interactions measured on the entropy scale. This component was computed as a sum of all three-way information gain values. Our third objective was to minimize the complexity of the mathematical models. This component was computed as the number of operators in the mathematical function being evaluated. We balanced these different objectives using multi-objective Pareto optimization. Specifically, we used the Non-dominated Sorting Genetic Algorithm II (NSGA-II) algorithm (Deb et al. 2002).

**Experimental design.** We first simulated 10 features (i.e. genetic variants) each with three genotypes of equal frequency for 1000 virtual human subjects. This data set was used as the input for HIBACHI and used to generate the class labels given a generated mathematical function. We used a liability threshold model such that the subjects with values generated by the function that were above the third quartile (i.e. 25%) were labeled as class 1 and those below as class 0. These resulting data sets were then used to determine the three fitness values that are used in the Pareto optimization to select best models for consideration in the next generation of the genetic program. For the first fitness objective, we used target odds ratios of 1.5, 1.3, 1.2, 1.15, 1.1, 1.09, 1.08, 1.07, 1.06, and 1.05 for each of the 10 features in the input data. These were chosen to be similar to what has been observed in a recent large GWAS study of sporadic breast cancer (Michailidou et al., 2017) and many other common diseases.

We ran the genetic program with a population size of 1000 for 10 and 50 generations. These runs were repeated 50 times with 50 different random seeds. We returned the best models from the final Pareto front and limited the results to those models that contained all 10 genetic variants. The goal of the thought experiment was to first identify mathematical models that minimized the difference between the odds ratios (i.e. univariate effects) of the features from the simulated data and those of the target values representing realistic values observed in large-scale genetic studies. The second goal was to determine whether models consistent with target odds ratios could also have two-way and three-way synergistic interactions that have effect sizes as large or larger than the specified univariate effects. The resulting networks of genetic effects were then compared with networks from real data (see below) from genetic studies of sporadic breast cancer. The conventional wisdom is that the genetic architecture of common diseases is purely due to genetic variants with independent and additive effects. The goal of this thought experiment was to show that additive effects observed in real data are consistent with a complex genetic architecture driven by synergistic interactions.

**Real data analysis.** The goal of the real data analysis was to examine genome-wide association study (GWAS) data available from population-based studies of a common human disease for evidence of the types of patterns that are observed from the simulations. We selected sporadic breast cancer as the disease of interest because numerous GWAS studies have been conducted resulting in dozens of published genetic risk factors (e.g. Michailidou et al., 2017).

We used two publicly available GWAS data sets from the database of Genotypes and Phenotypes or dbGaP that is maintained by the National Institutes of Health, USA (Tryka et al. 2014). The first is from the Cancer Genetic Markers of Susceptibility (CGEMS) Breast Cancer GWAS study with dbGaP accession phs000147.v3.p1. We merged the two available genotype call data sets that each had more than

500,000 measured genetic variants with a total sample size of 2,576 subjects. The second is from the Breast and Prostate Cancer Cohort Consortium (BPC3) GWAS of Aggressive Prostate Cancer and ER- Breast Cancer with dbGaP accession phs000812.v1.p1. We merged the breast cancer data across four consent groups that each had more than 500,000 measured genetic variants with a total sample size of 4,915 subjects. Each of these data sets were from studies of primarily European-derived subjects.

We used PLINK v1.9 (Purcell et al. 2007), bcftools and vcftools (Danecek et al. 2011) to manipulate and filter our data sets, after mapping coordinates to human genome assembly 19 (hg19). We followed the guidelines in (Anderson et al. 2010) for Quality Control (QC), applying the following filters in the listed order:

i. Individuals failing the PLINK '--check-sex' were removed.
ii. Markers with missing-call rate exceeding 0.01 were removed.
iii. Individuals with missing-call rate exceeding 0.01 were removed.
iv. Markers with Minor Allele Frequency (MAF) below 0.05 were removed.
v. Markers with Hardy-Weinberg equilibrium exact test p-value below 0.00001 were removed.
vi. Individuals were filtered based on relatedness (Identity By State>0.125). Then steps (ii)-(v) were repeated.
vii. Markers with significantly different genotype call rates between cases and controls were removed.

For both data sets the first 10 Principal Components (PCs) were then obtained using the PLINK –pca command after Linkage Disequilibrium (LD) pruning (–indep-pairwise 50 5 0.2). Upon using logistic regression to assess association of these PCs with case/control status and evaluating inflation with and without correction for PCs, we established that for the BPC3 data set it was necessary to adjust for the first 6 PCs in downstream analyses. We analyzed the data with Visualization of Statistical Epistasis Analysis (ViSEN) software version 1.0_beta_02 (Hu, Chen, Kiralis, and Moore 2013), using default parameters. ViSEN estimates two-way and three-way interactions using measures from information theory (Hu, Chen, Kiralis, Collins, et al. 2013; Hu et al. 2011; Moore and Hu 2015). Since VISEN does not have a built-in covariate adjustment, we first applied a local case-control subsampling approach (Fithian and Hastie 2014) to data set 2. The main idea is to take a subsample of the original collection of individuals, consisting of the most 'surprising' individuals, essentially those for which the prediction based on the covariates alone does not explain their case-control status well, and use only these individuals in the analyses. After applying this to the BPC3 data, we were left with 2336 subsampled individuals, a size comparable to that the CGEMS data set (2504 after QC). VISEN was run using these individuals and focusing on the 29 SNPs obtained by considering the SNPs with a combined p-value < $5x10^{-8}$ from Supplementary Table 2 of (Michailidou et al., 2017) which were common to our two data sets, after QC.

## Results

### Computational thought experiments

The primary goal of our computational thought experiments using HIBACHI was to determine whether independent and additive effects observed for a complex disease such as sporadic breast cancer are consistent with a complex genetic architecture driven by synergistic interactions or epistasis. We ran HIBACHI 50 times with 50 different random seeds for 10 and 50 generations of the genetic programming algorithm. Of the final Pareto optimal models we kept the ones that used all 10 features in the model. The results show that we can routinely generate models that match the target univariate odds ratios but that also have synergistic interactions with effect sizes as big or bigger than single feature effects. In fact, every Pareto model exhibited synergistic interactions in the presence of independent additive effects. Models generated with 50 generations of the genetic program had stronger interactions (mean information gain = 1.08) than those generated using only 10 generations (mean information gain = 0.97). This difference was highly significant base on a two-sample t-test (p=0.000048) and a nonparametric Mann-Whitney U test (p=0.0001). Further, the models generated using 50 generations had univariate effects closer to the target (mean difference = 0.52) than those generated using only 10 generations (mean difference = 0.74). This difference was also highly significant base on a two-sample t-test (p=0.000389) and a nonparametric Mann-Whitney U test (p=0.0001).

Figure 1 shows a network diagram of the genetic variants for one of the results generated by HIBACHI from the 50-generation runs. The circular nodes in the network represent the 10 genetic variants or features. Edges between the nodes represent the two-way interaction information (Hu et al., 2011) that is estimated by subtracting out the independent effects. The triangular nodes represent information gain due to pure three-way interactions after subtracting out the two-way and one-way effects (Hu et al., 2013b) and are connected by dashed lines to the three participating genetic variants. The magnitude of the one-way, two-way, and three-way genetic effects are color coded with darker red indicating a bigger effect size. The color of the nodes is a good reference point since these are by design close to the magnitude of the independent and additive effects seen in real data from genetic studies of sporadic breast cancer. Note that many of the synergistic interactions are of greater magnitude. For example, the largest univariate effect has an information gain of 0.019480 while the largest two-way interaction has an information gain of 0.055524 and the largest three-way interaction has an information gain of 0.067746. In fact, there are three two-way interactions and three three-way interactions that exceed the effect size of the genetic variant with the largest univariate effect. This pattern was consistent for every result generated suggesting that HIBACHI could

routinely generate models that satisfied the objectives of the thought experiment. The mathematical function that generated this pattern of interactions is shown in Figure 2.

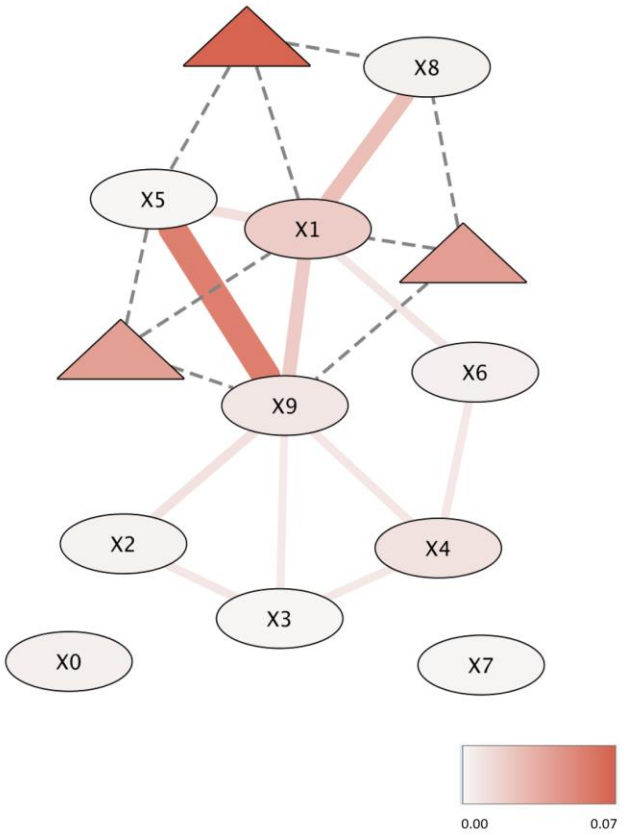


**Figure 1.** Network diagram summarizing the univariate (circles), two-way interaction (edges), and three-way interaction (triangles) effects from the computational thought experiment. Darker shading indicates higher information gain.

### Application to sporadic breast cancer

Figure 3 shows the one-way, two-way, and three-way genetic effects from real GWAS data measured in two different population-based studies of sporadic breast cancer. The first network diagram is from the CGEMS study (Figure 3A) while the second is from the BPC3 study (Figure 3B). Note that both studies exhibit synergistic interactions (edges and triangles) that are as strong or stronger than the independent effects (circular nodes) as indicated by the darker color. This is consistent with what we observed in the thought experiments. It is important to note that the 29 genetic variants shown here were all significant at a genome-wide level ($p < 5x10^{-8}$) that corrects for one million independent statistical tests in the parent study with sample sizes exceeding 200,000 women (Michailidou et al., 2017). Thus, these variants are considered highly likely to be real genetic risk factor signals. The question is not whether these genetic effects are real but rather whether the effects are independent from one another.

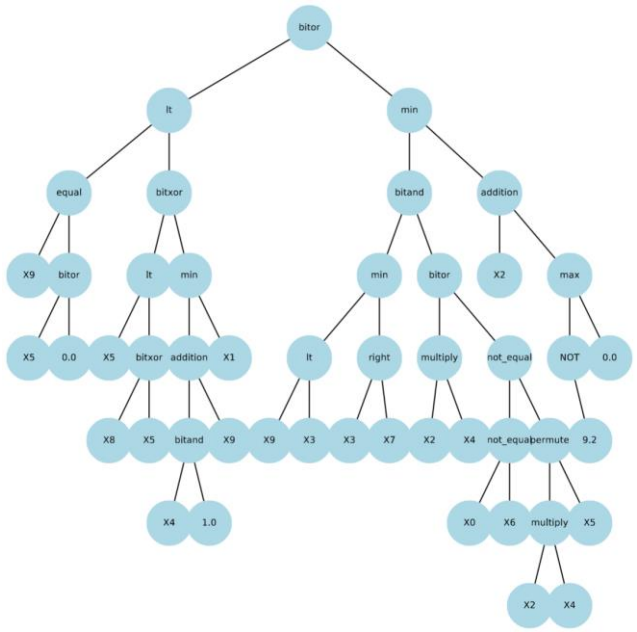


**Figure 2.** Mathematical model for simulating the data that generated the results shown in Figure 1. Each numbered node that starts with X is a genetic variant.

## Discussion

Genome-wide association studies (GWAS) were enabled more than 15 years ago by DNA chip technology that allowed hundreds of thousands of common genetic variants to be measured across the genome in a cost-effective manner (Bush and Moore 2012; Hirschhorn and Daly 2005; Wang et al. 2005). By taking advantage of the correlation structure (i.e. linkage disequilibrium) of genetic variation, these studies are able to capture most of the relevant common variants. Over the past 10 years there have been hundreds of published GWAS studies reporting replicable genetic associations for hundreds of phenotypes including many common human diseases (Visscher et al. 2017). Some of these are being functionally validated in experimental studies and may lead to new drugs. Despite these successes, there is growing recognition that GWAS has not delivered on its promises to uncover the genetic basis of common diseases. For example, the vast majority of genetic risk factors identified with this approach have very small effect sizes with odds ratios of 1.1 or less. This is certainly true for sporadic breast cancer that is the focus of this study (Michailidou et al., 2017). Further, when added together these variants don't explain much of the overall heritability of the disease. Even with much of the common genetic variation being measured, and with sample sizes in the hundreds of thousands, the variants found to date explain less than half of the risk due to inherited genetic variation. This means that much of the heritability of sporadic breast cancer and other common diseases remains a mystery.

Despite the limited success of the GWAS approach there is still a fundamental assumption that genetic variation contributes to disease risk in an additive manner. That is, the

effect of each genetic variant on risk in independent of all the other variants in the genome and all environmental exposures.

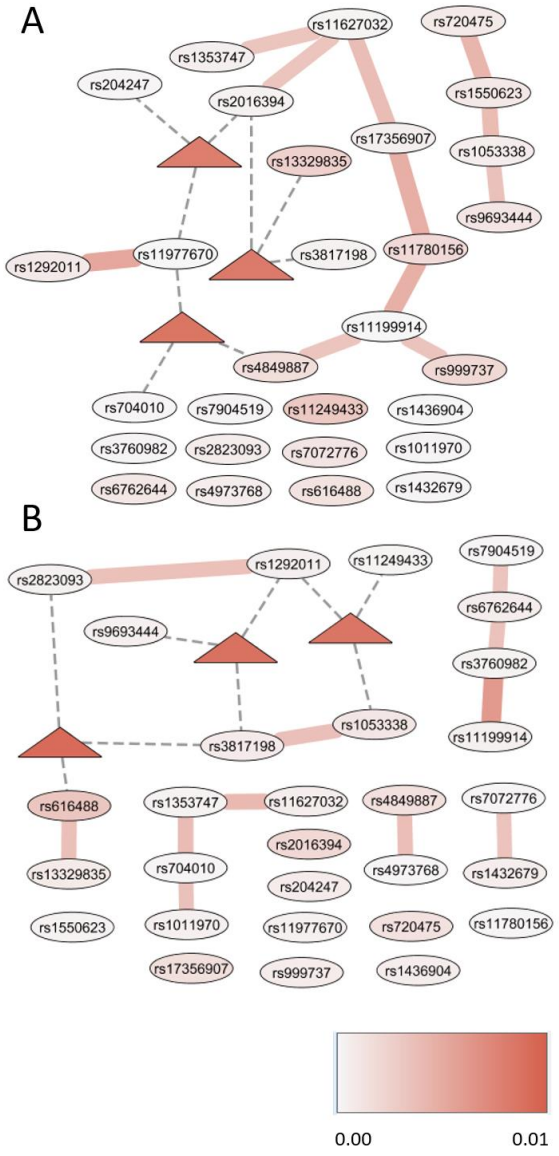


**Figure 3.** Network diagram summarizing the univariate (circles), two-way interaction (edges), and three-way interaction (triangles) effects from the CGEMS (A) and BPC3 (B) breast cancer GWAS data. Darker shading indicates higher information gain. Node labels indicate the RS number of the genetic variant.

This assumption is statistically convenient because it suggests that simple univariate statistics and large sample sizes are all that is needed to identify the remaining missing heritability. A major problem with the independence assumption is that it is not consistent with how biology works (Moore 2003). Molecular and cellular systems are driven by complex biomolecular interactions. This raises the question as to what degree complex interactions at the cellular level translate to complex statistical patterns of interaction among genetic variants at the population level (Moore and Williams 2005). We certainly see widespread evidence of complex genetic interactions in model systems. Humans are many times more complex. Do we really expect them to be simpler with respect to genetic architecture?

The goal of this study was to use computational thought experiments to test the hypothesis that observed independent and additive effects seen in GWAS exclude the possibility that genetic architecture is complex and largely due to synergistic interactions that are more consistent with the complexity of biomolecular interactions seen at the cellular level in individuals. An additional goal was to interrogate GWAS data from two publicly available studies of sporadic breast cancer to explicitly test for pattern of genetic interaction among replicable genetic risk factors. The results of the computational thought experiments show that a simulation approach based on genetic programming can routinely generate mathematical models that specify genetic effects with both independent effects and complex interactions that are of a larger magnitude. These results demonstrate that it is possible to have a complex genetic architecture with GWAS results that match what is observed in the literature. Furthermore, the real data analysis shows evidence of the same kind of genetic interactions from two large GWAS studies of sporadic breast cancer.

We have previously presented the working hypothesis that epistasis is a ubiquitous component of the genetic architecture of common diseases (Moore 2003). This is based on the idea that epistasis has been studied for more than 100 years (Bateson 1909), that many univariate effects don't replicate suggesting their actions are context-dependent, that epistasis is commonly found when investigated using proper computational methods, and that biological and biomedical phenomena are driven by complex biomolecular interactions at the cellular level. This last point is important because we now know a lot about the complexity of processes such as eukaryotic gene regulation that is governed by large numbers of protein-protein, protein-DNA, protein-RNA, DNA-DNA, and RNA-RNA interactions. These biomolecular interactions also connect genes and other non-coding RNA regions to each other forming extensive gene networks. Intra- and inter-genic interactions among biomolecules provide a framework for epistasis observed at the population level through robustness due to stabilizing selection (Cowper-Sal lari et al. 2011). Connecting cellular and population-level epistasis will be important if we are to use genetic and genomic information to improve health (Moore and Williams 2005).

Artificial life and digital biology have an important role to play in helping us understand the genetic architecture of complex traits such as susceptibility to common human diseases. This can include simulation methods such as GAMETES that uses probability to generate patterns of interactions that are associated with a discrete outcome (Urbanowicz, Kiralis, Fisher, et al. 2012; Urbanowicz, Kiralis, Sinnott-Armstrong, et al. 2012). It can also include methods such as AVIDA that evolve digital organisms under natural selection that exhibit epistasis (Lenski et al. 1999). At a minimum, these are good starting points as we develop the computational, statistical, and experimental tools that will

allow us to document epistasis in human populations and at the molecular and cellular levels. Future directions for the current study include allowing for more complex models and exploring additional GWAS data for similar patterns.

## Acknowledgments

This work was supported by National Institutes of Health (USA) grants LM010098 and LM12601.

## References

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5:1564–1573.

Ashlock, D. (2005). *Evolutionary Computation for Modeling and Optimization*. Springer Science & Business Media.

Barton, N. H., Etheridge, A. M., & Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73.

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8:e1002822.

Cowper-Sal lari, R., Cole, M. D., Karagas, M. R., Lupien, M., & Moore, J. H. (2011). Layers of epistasis: genome-wide regulatory networks and network approaches to genome-wide association studies. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, 3:513–526.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, 27:2156–2158.

De, R., Verma, S. S., Holzinger, E., Hall, M., Burt, A., Carrell, D. S., et al. (2017). Identifying gene-gene interactions that are highly associated with four quantitative lipid traits across multiple cohorts. *Human Genetics*, 136:165–178.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6:182–197.

Fithian, W., & Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of Statistics*, 42:1693–1724.

Fortin, F.-A., Rainville, F.-M. D., Gardner, M.-A., Parizeau, M., & Gagné, C. (2012). DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research*, 13:2171−2175.

Gaertner, B. E., Parmenter, M. D., Rockman, M. V., Kruglyak, L., & Phillips, P. C. (2012). More than the sum of its parts: a complex epistatic network underlies natural variation in thermal preference behavior in Caenorhabditis elegans. *Genetics*, 192:1533–1542.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, 6:95–108.

Holzinger, E. R., Verma, S. S., Moore, C. B., Hall, M., De, R., Gilbert-Diamond, D., et al. (2017). Discovery and replication of SNP-SNP interactions for quantitative lipid traits in over 60,000 individuals. *BioData Mining*, 10:25.

Hu, T., Chen, Y., Kiralis, J. W., Collins, R. L., Wejse, C., Sirugo, G., et al. (2013). An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, 20:630–636.

Hu, T., Chen, Y., Kiralis, J. W., & Moore, J. H. (2013). ViSEN: methodology and software for visualization of statistical epistasis networks. *Genetic Epidemiology*, 37:283–285.

Hu, T., Sinnott-Armstrong, N. A., Kiralis, J. W., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2011). Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC bioinformatics*, 12:364.

Jansen, R. C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews. Genetics*, 4:145–151.

Lenski, R. E., Ofria, C., Collier, T. C., & Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400:661–664.

Li, W., & Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50:334–349.

Mackay, T. F. C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews. Genetics*, 15:22–33.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551:92–94.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56:73–82.

Moore, J. H., Amos, R., Kiralis, J., & Andrews, P. C. (2015). Heuristic identification of biological architectures for simulating complex hierarchical genetic interactions. *Genetic Epidemiology*, 39:25–34.

Moore, J. H., & Hu, T. (2015). Epistasis analysis using information theory. *Methods in Molecular Biology*, 1253:257–268.

Moore, J. H., Shestov, M., Schmitt, P., & Olson, R. S. (2018). A heuristic method for simulating open-data of arbitrary complexity that can be used to compare and evaluate machine learning methods. *Pacific Symposium on Biocomputing*, 23:259–267.

Moore, J. H., & Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*, 27:637–646.

Moore, J. H., & Williams, S. M. (2009). Epistasis and its implications for personal genetics. *American Journal of Human Genetics*, 85:309–320.

Poli, R., Langdon, W. B., & McPhee, N. F. (2008). *A Field Guide to Genetic Programming*. Lulu Enterprises, UK Ltd.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81:559–575.

Remold, S. K., & Lenski, R. E. (2004). Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. *Nature Genetics*, 36:423–426.

Thornton-Wells, T. A., Moore, J. H., & Haines, J. L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *Trends in Genetics*, 20:640–647.

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., et al. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, 42:D975-979.

Urbanowicz, R. J., Kiralis, J., Fisher, J. M., & Moore, J. H. (2012). Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection. *BioData Mining*, 5:15.

Urbanowicz, R. J., Kiralis, J., Sinnott-Armstrong, N. A., Heberling, T., Fisher, J. M., & Moore, J. H. (2012). GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining*, 5:16.

Verma, S. S., Cooke Bailey, J. N., Lucas, A., Bradford, Y., Linneman, J. G., Hauser, M. A., et al. (2016). Epistatic Gene-Based Interaction Analyses for Glaucoma in eMERGE and NEIGHBOR Consortium. *PLoS genetics*, 12:e1006186.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS

Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101:5–22.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews. Genetics*, 6:109–118.

Weiss, K. M. (1995). *Genetic Variation and Human Disease: Principles and Evolutionary Approaches*. Cambridge: Cambridge University Press.

Yadav, A., & Sinha, H. (2018). Gene-gene and gene-environment interactions in complex traits in yeast. *Yeast*, in press.