How Computational Experiments Can Improve Our Understanding of the Genetic Architecture of Common Human Diseases

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 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 interaction effects from two large-scale GWASs of sporadic breast cancer and find evidence to support our hypothesis that is consistent with the results of our computational experiment.

1 Introduction

Genetic architecture has been defined to include the genetic variants that influence a phenotype, such as the 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 [43]. There are many different ways in which a particular genetic variant can influence phenotypic variability [36]. 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., \textit{variant-environment interaction}). The variant effect can also be dependent on one or more other genetic factors (i.e., \textit{variant-variant interaction}). A variant-variant interaction can stem from a gene-gene interaction, or from an interaction between a gene and a regulatory element, or between two regulatory elements, and so forth. We focus here on a particular type of genetic variant, namely single-nucleotide polymorphisms (SNPs), and on SNP-SNP interactions, or \textit{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 [4]. The etymological meaning is one gene standing upon another gene to influence a phenotype. There is widespread evidence of epistasis in model organisms, including bacteria [34], yeast [44], worms [12], and flies [21]. Epistasis is expected to be a ubiquitous component of genetic architecture in humans [23], but has been more difficult to study because of the increase in biological and computational complexity and our limited ability to perform experiments in controlled environments and with controlled genetic backgrounds [18]. 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 [8, 14]. Verma detected and replicated epistatic effects on susceptibility to glaucoma [40]. 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 SNPs on a genome-wide scale and then assess them individually as independent risk factors for common human diseases. This univariate approach in genome-wide association studies (GWASs) has been successful at identifying a number of genetic risk factors, but most have very small effect sizes across a population. For example, Michailidou et al. identified 65 new SNPs for sporadic breast cancer using the univariate GWAS approach in a sample of more than 200,000 women [22]. However, these new SNPs 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 GWASs is that the effect sizes trail off in an exponential fashion such that only a few SNPs have a moderate effect on risk while most have a very small effect. The success of GWASs has encouraged the idea of an \textit{infinitesimal} model that assumes that phenotypic variation is the result of many SNPs, each with a small additive effect [3]. This model discounts complexities such as epistasis and gene-environment interactions.

There were two primary objectives in the present study. Our first goal was to carry out a computational experiment to determine whether the distribution of effect sizes seen in a 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 the types of complex patterns observed in the simulation studies were also found in real data.

2 Methods

2.1 Computational Experiments

The primary goal of our study is to test the assumption that the observed exponential decay of univariate GWAS effect sizes implies that the 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 method of heuristic identification of biological architectures for simulating complex hierarchical interactions (HIBACHI) to generate simulation models and data that match the univariate effects observed in GWASs as closely as possible while maximizing SNP-SNP interactions.

There are five components to our HIBACHI simulation method. The first is the metaphor for the hierarchical biological framework that transmits information from a 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.
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 experiments described here. In what follows we first illustrate the workflow realized by combining these components; then we describe each of these in turn. The first three components are descriptions included in previous work on HIBACHI [24]. The fourth component is included in recent work on incorporating genetic programming into HIBACHI [27]. Both sets of descriptions are included here for completeness. The last component is new and describes the additions for the present study. All simulated data and models are available upon request.

2.2 Workflow Description
Figure 1 illustrates the workflow for our computational experiment. We first simulated 10 features (i.e., SNPs) independently, each with three genotypes (coded 0, 1, 2) of equal frequency, for 1000 virtual individuals. This input data set was fixed and used for every generation in every run of HIBACHI. Each run used either \( n = 10 \) or 50 generations, and for each choice of \( n \) we performed 50 runs, each corresponding to a different choice of random seeds. For each run we used an initial population size of 1000 candidate solutions or functions (generated using the random seeds). Each function in the population corresponds to a wiring architecture of features and simpler functions (operators). We use the term “model” to indicate the wiring architecture representing such a function. For each generation of a given run and for each function in that generation, a liability threshold was used to assign a class label to each individual from the input data set. This label was 1 if the value of the function for that individual was above the third quartile in the distribution across all 1000 individuals, and 0 otherwise. Thus, each function in the population yields a resulting output data set consisting of the input data set augmented with the class labels, which represents a complete simulated data set. These simulated data sets were then used to determine the three fitness values (described below) that were employed in the Pareto optimization to select the functions used by the genetic program to derive the next generation. At the end of each run of HIBACHI, we returned the best functions from the final Pareto front and limited the results to those corresponding to models that contained all 10 features. Each such function yields a simulated data set, as described above.

2.3 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 SNPs and their phenotypic relationships propagated through a hierarchical set of operators. The
prototype for HIBACHI was developed previously and used a fixed architecture (Figure 2(a)) that was based on genetic effects at the gene level [24]. By fixed architecture we mean that the wiring diagram was fixed and the GP was used to discover the operators. We describe this framework first and then discuss how it relates to the new approach that uses GP to discover both the wiring diagram and the operators. The initial HIBACHI framework started with a protein-coding gene (i.e., an mRNA gene) with a single nonsynonymous SNP that is assumed to change an amino acid. Upstream of the mRNA gene are a promoter region and an enhancer region, each with a single regulatory SNP. 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 SNP in the gene that codes for each transcription factor. We also included a single SNP in a microRNA gene that participates in post-translational regulation. In total, this structure allowed for six SNPs (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. Thus, this framework expects an input data set where each data point (representing an individual) consists of a vector of seven values. It is important to note that this particular biological framework with a fixed wiring 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 allow 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.

Figure 2. Prototype HIBACHI biology-based framework, with a fixed architecture. In this framework only the operators in the orange cells vary, and each data point is represented by seven features. (a) Example of a model with this architecture and specific operator choices. (b) Distribution of endpoint values from this model across the input data set. Red bars correspond to values above the liability threshold.
2.4 **A Mathematical Framework for Simulation of Complex Patterns**

The goal of this component is to provide a flexible wiring framework for combining features to produce an endpoint. Using the genetics metaphor, features (reflecting biology-based loci) feed into operators, the result of each being carried forward to the next operator. This wiring architecture of features and operators defines therefore a function whose arguments are the features. For example, in the prototype for HIBACHI, one transcription factor SNP was combined with the enhancer SNP through an operator, whose result was then combined with the second transcription factor SNP (via another operator). The result of this operation was combined with the promoter SNP. This result was combined with the coding SNP in the gene. This result was combined with the environmental factor to produce a protein product. Thus, the protein expression value was defined as a function of seven features, namely the six SNPs and the environmental factor.

In the more general framework, each data point will be represented by \( n \) features, giving the genotypic values for an individual, and in this work we used \( n = 10 \). For a given model, corresponding to a function in this framework, evaluation of such a function across the data points in the input data set produces endpoint values, which simulate phenotypic values for these individuals. More precisely, these values can be directly used as representing a quantitative trait, or a liability threshold on their distribution can be used to generate a case-control disease status for each individual, as described below.

For each run of HIBACHI the user can specify a set of operators to use to build the models. Examples of basic operators include addition, subtraction, multiplication, division, modulus, and modulus-2. Logical operators include greater than, less than, AND, OR, and XOR. Bitwise operators include bitwise AND, bitwise OR, and bitwise XOR. Unary operators include absolute value, NOT, factorial, left, and right. Large operators include power, log, permute, and choose. Miscellaneous operators include minimum and maximum. Operators such as XOR are known to produce patterns of epistasis [20, 29]. The choice of operators depends on the desired (biological) complexity—for example, on whether simple assumptions (e.g., additivity) are made or more complex scenarios (e.g., epistasis) are contemplated.

2.5 **A Liability Threshold Model for Biology-Based Simulation**

As indicated above, for a given model, we use a liability threshold to simulate disease status from the distribution of phenotypic values generated from the genotypic values using that model. A threshold is selected, and individuals whose phenotypic values are above this threshold are assigned to class 1, whereas the others are assigned to class 0. The user can select the liability threshold to achieve a particular disease prevalence. In this work we use as threshold the third quartile of the distribution of simulated phenotypic values across all input individuals. More details about the liability model have been provided previously [24].

2.6 **Genetic Programming for Model Representation and Stochastic Search**

We selected GP as our stochastic search engine for several reasons. First, GP uses binary expression trees, which 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 [2]. 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 or miRNA regulation of transcription. 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 [32]. Our initial implementation of the GP approach was for generating HIBACHI models to evaluate and compare machine learning methods [27]. Our GP implementation used the framework of distributed evolutionary algorithms in Python [11], available on GitHub at https://github.com/deap. We describe below our modification of the fitness function to carry out the experiments. This new version of HIBACHI is available on GitHub at https://github.com/EpistasisLab/hibachi.
2.7 A Fitness Function for Genetic Architecture Experiments

A key to GP search is the fitness function that specifies the value or quality of a set of operators 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. In this work, 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 the 10 features. These were chosen to be similar to what has been observed in a recent large GWAS of sporadic breast cancer [22] and many other common diseases. The first fitness component was computed as a sum of absolute differences between the target odds ratios and the odds ratios measured from the simulated data set being evaluated. The odds ratio is a common epidemiologic measure of association between some exposure (e.g., genotype) and the presence of disease in a sample of subjects derived from a population-based study. Our second objective was to maximize the three-way SNP-SNP interactions measured on the entropy scale. This component was computed as a sum of all three-way information gain values [26]. Our third objective was to minimize the complexity of the mathematical models. This component was computed as the number of operators in the model being evaluated. As illustrated in Figure 1, the first two fitness components are evaluated using the simulated data from each model, whereas the third is solely based on the model itself. We balanced these different objectives using multi-objective Pareto optimization. Specifically, we used Non-dominated Sorting Genetic Algorithm II (NSGA-II) [9].

2.8 Comparing Models from the Computational Experiments with Real Data

The first goal of the computational experiment was to identify models that minimized the differences 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. This motivated our choice for the first fitness component. The second goal was to determine whether models consistent with target odds ratios could also have synergistic interactions that have effect sizes as large as or larger than the specified univariate effects. This motivated the choice for the second fitness component. The choice for the third fitness component, represented by complexity, was motivated by the desire to identify models that could be more easily interpreted biologically. The network of one-way, two-way, and three-way SNPs for any such model can then be compared with analogous networks from real data, which in this study were data from genetic studies of sporadic breast cancer. One-way effects reflect association of individual SNPs with the phenotype, two-way effects reflect epistatic associations of pairs of SNPs with the phenotype, and three-way effects reflect epistatic associations of triples of SNPs with the phenotype. The conventional wisdom is that the genetic architecture of common diseases is purely due to genetic variants, such as SNPs, with independent (i.e., one-way) and additive effects. The goal of this experiment was to show that one-way effects observed in real data are consistent with a complex genetic architecture driven by epistatic interactions. We used the Visualization of Statistical Epistasis Analysis (ViSEN) software (version 1.0_beta_02) [16] to compute and display the one-way, two-way, and three-way effects for both computationally generated models and real data. In a ViSEN diagram, the values (displayed as color darkness) in elliptical nodes reflect the magnitude of the one-way effect for the corresponding feature (SNP), those in edges reflect the magnitude of the two-way effects for the SNPs in the two vertices, and those in triangles reflect the magnitude of the three-way effects for the SNPs connected to the triangles. ViSEN estimates two-way and three-way interactions using measures of interaction information from information theory [15, 17, 26].

2.9 Real-Data Analysis

The goal of the real-data analysis was to examine 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 GWASs of it have been conducted, resulting in dozens of published genetic risk factors (e.g., [22]).
We used two publicly available GWAS data sets from the Database of Genotypes and Phenotypes (dbGaP) that is maintained by the National Institutes of Health, USA [37]. The first is from the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer GWAS with dbGaP accession phs000147.v3.p1. We merged the two available genotype call data sets, which 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, which each had more than 500,000 measured SNPs, with a total sample size of 5,261 subjects. Each of these data sets was from studies of primarily European-derived subjects.

We used PLINK v1.9 [33], BCFtools, and VCFtools [7] to manipulate and filter our data sets, after mapping coordinates to human genome assembly 19 (hg19). We followed the guidelines in [1] 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 six PCs in downstream analyses. We analyzed the data with ViSen, using default parameters. Since ViSen does not have a built-in covariate adjustment, we first applied a local case-control subsampling approach [10] to the BPC3 data set. 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 of 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 $< 5 \times 10^{-8}$ from Supplementary Table 2 of [22], which were genotyped in both of our data sets, after QC. We ran ViSen with default settings, which results in a network displaying the top 11 two-way interactions and the top 3 three-way interactions based on information gain values.

For each data set we generated receiver operating characteristic (ROC) curves and computed areas under the curve (AUCs) based on random forests (RFs) obtained with three different choices of input features: (i) the 29 SNPs only, (ii) the 29 SNPs plus the multifactor dimensionality reduction (MDR) transforms of the 14 two- and three-way top models from ViSen, and (iii) the MDR transforms of the 14 two- and three-way top models for ViSen only. MDR [25] uses constructive induction to define a new attribute as a function of two or more other attributes by transforming a multilocus genotype combination into 1 if the ratio of cases to controls in that group exceeds the overall ratio of cases to controls in the data set, and into 0 otherwise. We used the Python package
scikit-mdr, available at https://github.com/EpistasisLab, to obtain MDR transforms of the 14 two- and three-way models for each data set. Then, for each choice of input features among (i)–(iii) above, we randomly split our data set into a training set (75% of individuals) and a testing set (the remaining 25%); then used the training set to fit a RF via the tree-based optimization tool (TPOT), implemented in the Python package tpot, also available at https://github.com/EpistasisLab, with the AUC as scoring function; and obtained the AUC on the testing set. The TPOT is an automated machine learning (AutoML) tool described in [30, 31, 35].

3 Results

3.1 Computational Experiments

The primary goal of our computational 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. As indicated earlier, we ran HIBACHI 50 times for each of \( n = 10 \) and 50 generations of the genetic programming algorithm. The results show that we can routinely generate final Pareto optimal 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 as measured by the second fitness function (mean sum of three-way information gains = 1.08) than those generated using only 10 generations (mean sum of three-way information gains = 0.97). This difference was highly significant according to 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 sum of absolute differences = 0.52) than those generated using only 10 generations (mean sum of

Figure 3. Network diagram summarizing the one-way (ellipses), two-way interaction (edges), and three-way interaction (triangles) effects from a simulated data set in the computational experiment. Darker shading indicates higher information gain.
absolute differences = 0.74). This difference was also highly significant according to a two-sample t-test ($p = 0.000389$) and a nonparametric Mann-Whitney U-test ($p = 0.0001$).

Figure 3 shows a ViSEN network diagram of the features (representing SNPs) for one of the results generated by HIBACHI from the 50-generation runs. The elliptical nodes in the network represent the 10 SNPs. Edges between the nodes represent the (top 11) two-way interaction information [17] that is estimated by subtracting out the independent effects. The triangular nodes represent the (top 3) information gain due to pure three-way interactions after subtracting out the two-way and one-way effects [16] and are connected by dashed lines to the three participating SNPs. The magnitude of the one-way, two-way, and three-way genetic effects is color coded, with darker red indicating a bigger effect size. 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 SNP with the largest univariate effect. This pattern

Figure 4. Model for simulating the data that generated the results shown in Figure 3. Each numbered node that starts with X is a feature (representing a SNP).
Figure 5. Network diagram summarizing the one-way (ellipses), 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, as per displayed color scale. Node labels indicate standard SNP identifiers.
was consistent for every result generated, suggesting that HIBACHI could routinely generate models that satisfied the objectives of the experiment. The model that generated this pattern of interactions is shown in Figure 4.

### 3.2 Application to Sporadic Breast Cancer

Figure 5 shows the one-way, two-way (top 11), and three-way (top 3) genetic effects from real GWAS data measured in two different population-based studies of sporadic breast cancer for 29 established breast cancer SNPs. The first network diagram is from the CGEMS study (Figure 5(a)), while the second is from the BPC3 study (Figure 5(b)). Note that both studies exhibit synergistic interactions (edges and triangles) that are as strong as or stronger than the independent effects (elliptical nodes), as indicated by the darker color. This is consistent with what we observed in the computational experiments. It is important to note that the 29 SNPs shown here were all significant at a genome-wide level \((p < 5 \times 10^{-8})\) that corrects for one million independent statistical tests in the parent study with sample sizes exceeding 200,000 women [22]. Thus, these SNPs 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.

Looking at the RF built either using the 29 SNPs only, or using the 14 features obtained from the MDR transforms of the two- and three-way interactions, or using a combination of all 43 features, we note how the higher-level models improve the AUCs, which are essentially close to random when based solely on the one-way models. Table 1 reports the AUC in each of these three cases for each data set.

### 4 Discussion

Genome-wide association studies (GWASs) were enabled more than 15 years ago by DNA chip technology that allowed hundreds of thousands of common SNPs to be measured across the genome in a cost-effective manner [5, 13, 42]. 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 SNPs. Over the past 10 years there have been hundreds of published GWASs reporting replicable genetic associations for hundreds of phenotypes, including many common human diseases [41]. Typically, these studies report main effects, that is, SNPs that are individually statistically associated with the phenotypes, obtained from carrying out univariate analyses. Some of these are being functionally validated in experimental studies and may lead to new drugs. Despite these successes, there is growing recognition that the GWAS approach 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, which is the focus of this study [22]. Further, when added together, these SNPs do not explain much of the overall heritability of the disease. Even with much of the

<table>
<thead>
<tr>
<th>Data set</th>
<th>Features</th>
<th>Number of features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGEMS</td>
<td>1-way only</td>
<td>29</td>
<td>0.49</td>
</tr>
<tr>
<td>CGEMS</td>
<td>2- and 3-way only</td>
<td>14</td>
<td>0.59</td>
</tr>
<tr>
<td>CGEMS</td>
<td>1-, 2-, and 3-way</td>
<td>33</td>
<td>0.57</td>
</tr>
<tr>
<td>BPC3</td>
<td>1-way only</td>
<td>29</td>
<td>0.5</td>
</tr>
<tr>
<td>BPC3</td>
<td>2- and 3-way only</td>
<td>14</td>
<td>0.62</td>
</tr>
<tr>
<td>BPC3</td>
<td>1-, 2-, and 3-way</td>
<td>33</td>
<td>0.59</td>
</tr>
</tbody>
</table>
common genetic variation being measured, and with sample sizes in the hundreds of thousands, the SNPs 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 univariate 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 SNP on risk is independent of all the other SNPs in the genome and all environmental exposures. 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 [23]. Molecular and cellular systems are driven by complex biomolecular interactions. This raises the question to what degree complex interactions at the cellular level translate to complex statistical patterns of interaction among SNPs at the population level [28]. 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 experiments to test the hypothesis that observed independent and additive effects seen in GWASs 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 patterns of genetic interaction among replicable genetic risk factors. The results of the computational 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 GWASs of sporadic breast cancer. For the latter we took as our starting point a collection of 95 main effect SNPs identified by harnessing power through a combined meta-analysis of several GWASs [22] involving over 200,000 women. Twenty-nine of these SNPs were genotyped in the two breast cancer GWASs to which we had access, these with a sample size 100-fold smaller (just over 2000 women after QC). The predictive power of these 29 features was in effect random when we used them to build RFs out of our data sets. But the top two- and three-way interactions among these SNPs in each of our two data sets had a stronger association with the trait and larger AUC when used to build RFs.

We have previously presented the working hypothesis that epistasis is a ubiquitous component of the genetic architecture of common diseases [23]. This is based on the fact that epistasis has been studied for more than 100 years [4]; that many univariate effects do not 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, which 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 noncoding RNA regions to each other, forming extensive gene networks. Intra- and intergenic interactions among biomolecules provide a framework for epistasis observed at the population level through robustness due to stabilizing selection [6]. Connecting cellular and population-level epistasis will be important if we are to use genetic and genomic information to improve health [28].

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, which uses probability to generate patterns of interactions that are associated with a discrete outcome [38, 39]. It can also include methods such as AVIDA that evolve digital organisms under natural selection that exhibit epistasis [19]. 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 AI116794, LM010098, and LM012601.

**References**


