Computational Toxicology 25 (2023) 100261

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology

Exploring genetic influences on adverse outcome pathways using heuristic simulation and graph data science

Joseph D. Romano [a,b,c], Liang Mei [d], Jonathan Senn [d], Jason H. Moore [e], Holly M. Mortensen [f,*]

[a] Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States
[b] Center of Excellence in Environmental Toxicology, University of Pennsylvania, Philadelphia, PA, United States
[c] Department of Biostatistics, Epidemiology, & Informatics, University of Pennsylvania, Philadelphia, PA, United States
[d] Oak Ridge Associated Universities, Oak Ridge, TN, United States
[e] Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States
[f] United States Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, Research Triangle Park, NC, United States

ARTICLE INFO

Keywords:
Adverse outcome pathway
Liver cancer, genetic programming
Graph data science

ABSTRACT

Adverse outcome pathways provide a powerful tool for understanding the biological signaling cascades that lead to disease outcomes following toxicity. The framework outlines downstream responses known as key events, culminating in a clinically significant adverse outcome as a final result of the toxic exposure. Here we use the AOP framework combined with artificial intelligence methods to gain novel insights into genetic mechanisms that underlie toxicity-mediated adverse health outcomes. Specifically, we focus on liver cancer as a case study with diverse underlying mechanisms that are clinically significant. Our approach uses two complementary AI techniques: Generative modeling via automated machine learning and genetic algorithms, and graph machine learning. We used data from the US Environmental Protection Agency's Adverse Outcome Pathway Database (AOP-DB; aopdb.epa.gov) and the UK Biobank's genetic data repository. We use the AOP-DB to extract disease-specific AOPs and build graph neural networks used in our final analyses. We use the UK Biobank to retrieve real-world genotype and phenotype data, where genotypes are based on single nucleotide polymorphism data extracted from the AOP-DB, and phenotypes are case/control cohorts for the disease of interest (liver cancer) corresponding to those adverse outcome pathways. We also use propensity score matching to appropriately sample based on important covariates (demographics, comorbidities, and social deprivation indices) and to balance the case and control populations in our machine language training/testing datasets. Finally, we describe a novel putative risk factor for LC that depends on genetic variation in both the aryl-hydrocarbon receptor (*AHR*) and ATP binding cassette subfamily B member 11 (*ABCB11*) genes.

1. Introduction

Informatics and computational methods have revolutionized biomedical research and enabled scientists to explore questions that are either infeasible or impossible through traditional experimentation alone [32]. In environmental health and toxicology, common computational tasks include building and training models that predict various chemical properties, conducting statistical analysis of observational and epidemiological data to better understand exposure-related health outcomes, and performing network analyses to discover key processes in biochemical pathways, among others [17,38]. Despite the successes made using these methods, some key deficiencies have become

Abbreviations: AI, Artificial Intelligence; AOP, Adverse Outcome Pathway; AOP-DB, US Environmental Protection Agency's Adverse Outcome Pathway Database; AOP-KB, Adverse Outcome Pathway Knowledge Base; COVID-19, Coronavirus Disease 2019; EAGMST, Extended Advisory Group on Molecular Screening and Toxicogenomics; EPA, US Environmental Protection Agency; eQTL, Expression Quantitative Trait Locus; GP, Genetic Programming; Gtex, Genotype-Tissue expression project; GWAS, Genome-Wide Association Study; HIBACHI, Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions; KE, Key Event; LC, Liver Cancer; MIE, Molecular Initiating Event; OECD, Organisation for Economic Co-operation and Development; PSM, Propensity Score Matching; SNP, Single Nucleotide Polymorphism; UK:, United Kingdom; UKBB, UK Biobank.

* Corresponding author.
E-mail address: mortensen.holly@epa.gov (H.M. Mortensen).

https://doi.org/10.1016/j.comtox.2023.100261
Received 3 October 2022; Received in revised form 6 January 2023; Accepted 23 January 2023
Available online 25 January 2023

apparent in toxicological research, such as a lack of richly structured, multimodal biomedical data describing chemicals and the biological systems that respond to chemical exposure [31] and a paucity of novel methods for discovering new knowledge from these complex data resources [36]. In this paper, we employ both to gain new insights into a phenomenon of growing interest: the influence of genetics on susceptibility to an adverse outcome following specific chemical exposures.

Adverse Outcome Pathways (AOPs) are pathway-like descriptions that outline the mechanistic associations between molecular exposure events and higher-order clinical and population-level outcomes that may arise from the exposure [2,26]. AOPs consist of molecular initiating events (MIEs), key events (KEs), and adverse outcomes (AOs). By definition, a KE is any internal step within an AOP at some level of biological organization, and an MIE is a particular kind of KE that both initiates an AOP and is comprised of a molecular interaction between a toxicant and a body component. AOPs are classified according to their respective health outcomes, and AOPs associated with similar outcomes often overlap to create an 'AOP network.' An AOP's set of KEs can include genetic polymorphisms that are associated with higher risk to the adverse outcome. For example, colon cancer AOPs include 53 unique SNP associations originally derived from GWAS [37]. This study will attempt to look at the influence genetic phenomena have on susceptibility to adverse outcomes after specific chemical exposures using AOPs as a framework for reference.

Methodologically, one area in particular that has experienced rapid growth, and holds great promise in all areas of biomedicine, is artificial intelligence (AI). AI broadly aims to construct computational systems that make intelligent decisions based on available data, knowledge, and/or human input. The scope of what comprises AI is broad, and usually nebulously defined. In this paper, we explore two areas within AI: Evolutionary algorithms and graph data science. Evolutionary algorithms are a family of algorithms that imitate processes found in biological evolution to optimize a system (e.g., a predictive model, a symbolic mathematical equation, or even another algorithm). Unsurprisingly, evolutionary computation is often used in computational biology, for example, in the context of simulating natural systems or processes [10,11,22] and building machine learning classifiers that perform well on a specific task [16,23,28]. Graph data science refers to the quantitative analysis of *graphs* – sometimes known as networks (e.g., biological networks), and comprised of a set of nodes connected by a set of edges that define relationships between those nodes [7,27]. Some tasks within graph data science involve community detection [9], identification of the shortest paths linking two nodes in a graph [12], determining 'hub nodes' that play critical roles in the global structure of a graph [7,41], and using computational algorithms that yield quantitative understandings of the behavior and characteristics of a given graph [1,15]. Since AOPs can be represented as graphs, graph data science provides a powerful set of tools for discovering properties of AOPs that are not obvious through manual inspection.

Here, we propose a novel approach to gain understanding of the mechanisms underlying genetic influences on toxic adverse outcomes, without the inclusion of associated case-control information, that leverages these two areas of AI, and subsequently evaluates the approach in the context of toxicity-mediated adverse outcome pathways involved in liver cancer (LC). Briefly, we train interpretable generative models to construct synthetic datasets resembling real-world LC AOP genotype data via the HIBACHI software, and introspect the best models produced by HIBACHI (Heuristic Identification of Biological Architectures for simulating Complex Hierarchical genetic Interactions) for the most prominent AOP SNPs that influence LC outcomes. HIBACHI is a command line utility based on genetic programming (GP) that generates (synthetic) datasets with interactions between input features [24,25]. It uses the $(\mu + \lambda)$ evolutionary algorithm [6] to construct trees of primitive mathematical operations that can represent interactions between independent variables. For example, when applied to genetic data, these feature interactions may represent epistasis or mechanisms underlying polygenic traits. HIBACHI can take an existing dataset – referred to in the context of GP as a *model* – as input, which is then used to evaluate the fitness of candidate output datasets. Our hypothesis is that HIBACHI can create synthetic datasets of SNPs involved in AOPs that behave the same as real data for the same AOPs. This will allow us to explore the interpretable generative models used to create the synthetic data, which gives insights into interactions between specific features in the real data used to train HIBACHI. Conceptually, this process can be likened to a brute-force version of symbolic regression [18] that avoids pitfalls arising from statistical analyses on genetic data with complex interactions between features [40]. Importantly, this approach utilizes genomic and phenotypic data from real-world populations, combined with information and knowledge sourced from publicly available, open access databases describing mechanisms of toxicity. Our methods are generalizable to other diseases of interest and provide a new framework for toxicologists to explore genetic mechanisms that underlie toxic adverse outcome susceptibility.

2. Methods

2.1. Data sources

Our analysis uses data from the US Environmental Protection Agency's Adverse Outcome Pathway Database (AOP-DB) and the UK Biobank (UKBB). The AOP-DB provides a formal structure for AOPs and their contained key events, as well as the relationships and associations between key events, genes (and their variants), metabolic pathways, diseases, and other relationships of toxicological interest. Data in the AOP-DB are aggregated from third-party public databases, including automated data pulls from the AOP-Wiki [26], as part of the OECD-supported EAGMST AOP-KB sub-group effort.

The UKBB is a large collection of longitudinal genetic, clinical, and demographic data on more than 500,000 adult volunteers in the UK, and is available to the international research community via application (https://biobank.ndph.ox.ac.uk/showcase/index.cgi) [29,35]. These data are suitable for observational analysis of a vast array of clinical phenomena. Here, we utilized data on single nucleotide polymorphisms (SNPs), disease diagnoses, and relevant demographics data collected through extensive patient questionnaires. We use SNPs to establish genotypes that are implicated in AOPs relevant to the toxic outcome of interest, diagnoses to construct case and control cohorts, and demographic data to balance cohorts with respect to a number of demographic and clinical covariates of interest. All UKBB data used in this study are from the current data release as of November 2020.

2.2. Obtaining genotypes for cohort patients

In this study, we focus on LC as a clinical endpoint of interest, but our methods are generalizable to other diseases. Although there are several major subtypes of LC, we treat it as a single disease phenotype, due both to a lack of granularity in established LC AOPs, as well as to provide a larger training dataset for the HIBACHI program. To find genetic variants that play a role in the etiology of LC, we retrieve AOPs related to LC and extract SNPs annotated to key events within those AOPs. Using the AOP-DB, we query AOP titles, organ specificity annotations, and event components (KEs and MIEs) for presence of the terms "liver" and "hepatocellular" to fetch AOPs related to LC. These AOPs, MIEs, and KEs are listed in-detail in Table S1. We then manually remove any AOPs describing hepatic steatosis – a disease that, while a known risk factor for LC, has a distinctly different underlying etiology (Schulz et al. 2015). Using these identified AOPs, we retrieve annotations to the EntrezGene database via associations present in the AOP-DB's "AOP_gene" table [26]. In creating the AOP-DB, SNPs associated with KEs were originally obtained from the GTEX v7 Single Tissue eQTL dataset [13] and from the GWAS v1.0.2 All Associations dataset (https://www.ebi.ac.uk/gwas/docs/methods/criteria).

Finally, we assess overlap between the AOP SNPs and SNPs included in the UKBB genetic data. For every SNP we identify in the AOP data, we obtain genotypes at that locus for all patients in the cohorts defined below, and encode them in an additive model, (e.g., homozygous major allele is "0″, heterozygous is "1", and homozygous minor allele is "2") since this format is easily consumed by downstream analysis tools (e.g., HIBACHI). All AOP SNPs not included in the UKBB data were omitted from consideration in downstream analyses. It should be noted that all LC/SNP associations are determined using expert-curated biomedical knowledge originally mined from the AOP-Wiki, and are therefore independent from any observational biases that may be present in the UK Biobank genotype data.

2.3. Phenotyping and assembling patient cohorts

To assemble cohorts for statistical modeling of our toxic outcome of interest, we retrieved pertinent data from the UKBB [35]. We first filter all patients in the UKBB based on availability of SNPs included in our AOP network. Using the set of SNPs identified above (SNPs found in both AOPs and the UKBB), we retrieve unique identifiers for patients with that set of SNPs available. To construct raw (unbalanced) case and control cohorts, we then separated this set of patients into those with a diagnosis of LC (based on presence of the ICD-9 code prefix "C22") and without LC (all others).

Because many environmental factors can act as confounding variables in observational analyses of complex diseases, these confounders need to be balanced in the case and control cohorts to minimize the risk of predictive models learning to distinguish patients based on the confounding variables rather than the presence or absence of the disease of interest. In the case of this study, the predictive model is the output of HIBACHI's genetic programming algorithm, and the disease of interest is LC. Although there are several strategies for producing balanced cohorts, we used the propensity score matching (PSM) method. Briefly, PSM involves training a logistic regression model where input features are the confounders and the output is the *propensity score*, or probability of being a member of the treatment (case) group [5]. This logistic regression model is then used to match each sample in the case cohort to one or more samples in the control cohort based on having similar propensity scores. The resulting cohorts have an (approximately) balanced distribution of propensity scores within each possible value across all confounders. For confounders with continuous rather than categorical values (e.g., age), values are binned into equally sized groups across the range of values prior to matching. In doing so, PSM minimizes the estimation bias contributed by each confounding variable to the overall predictions of a model trained on the balanced dataset.

In this study, we performed PSM on the raw case and control cohorts using the following confounding features: age at recruitment, sex, ethnicity, and Townsend deprivation index (a composite measure of material deprivation within a population, incorporating employment status, car ownership, home ownership, and household overcrowding) [39]. Each of these is a known demographic confounder for LC risk, and the data is provided by UKBB questionnaire data available for all patients. Additionally, inclusion of the Townsend index helps to ensure generalizability of study results across socioeconomic groups, particularly those with historically poor access to quality healthcare. We also included diabetes status (presence of the ICD-9 code prefix "E1") as a cofounder, as diabetes is a significant risk factor for LC [19]. We used the Pymatch library (https://github.com/benmiroglio/pymatch) to construct the propensity score model, perform the matching procedure, and visualize confounder imbalance before and after matching. Since LC is a relatively rare diagnosis in the UKBB data, we increased the size of our dataset for training HIBACHI by matching 4 control patients to each case patient. We specified a propensity score similarity threshold of $1*10^{-4}$ – the smallest value that retains 100 % of the LC cohort.

2.4. Exploring genetic contributions to toxicity using genetic programming

We ran HIBACHI (available on GitHub at https://github.com/EpistasisLab/hibachi) on an input model consisting of patients in the PSM-balanced case and control cohorts constructed using the method described above. Specifically, we retrieved the LC SNPs of interest (described above) for each of the patients in the balanced case and control cohorts and used those to construct a feature matrix (in the 0,1,2 format, representing an additive or ordinal genetic model) with LC outcome being the binary target variable. We then trained HIBACHI on this LC dataset, with algorithm metaparameters of 100 generations of evolution and a population size of 100. Since HIBACHI outputs both a synthetic dataset with the characteristics of the training dataset as well as the generative model used to construct that dataset, we inspected both in order to explore genetic mechanisms that may govern susceptibility to LC following toxic exposures. To account for potential linkage disequilibrium (LD) between implicated SNPs, we computed pairwise R^2 and D' values between all implicated SNPs (i.e., showing up more than once in the learned generative models) using the LDpair module in the National Cancer Institute's LDlink toolkit [20]. Any pair of SNPs in statistically significant LD should be treated as suspect if they occur in the same generative model.

3. Results

3.1. AOPs and SNPs associated with liver cancer

Our initial query for LC AOPs finds 16 liver related AOPs and 189 SNPs associated with these AOPs. AOPs 1, 37, 41, 46, 107, 108, and 117 are specific, describing a particular etiology of LC or hepatocellular carcinoma, while the other AOPs describe LC in a more general context. Interestingly, the AOPs describing liver fibrosis, hepatotoxicity, and liver injury contain no SNP associations, although a number of these AOPs are still under development. The AOPs that feature SNP associations often specify a primary gene, inhibitor, or activator that plays a key role in the AOP, such as ABCB11 for AOP 27, PPARα for AOP 37, AHR for AOP 41, and AFB1 for AOP 46. Five AOPs in this list were derived from rodent experimental data (AOPs 37, 41, 107, 108, and 117), while the rest are based upon human-derived evidence. A full list of LC AOPs and their associated SNPs (including omitted AOPs related to hepatic steatosis) is given in Supplemental Information (Table S1). Figs. 1 and 2.

3.2. UKBB liver cancer cohort characteristics

Of the 189 SNPs identified within the AOPs, 25 are represented in the UKBB genotype data (Table 1). The remaining 164 SNPs may be missing due to limited coverage of genotyping panels, semantic inconsistencies between the AOP-DB and UKBB variant nomenclature, or other issues. We identified 488,377 patients with genotypes available for these SNPs. Of these patients, 580 had an LC diagnosis. We then generated balanced case and control cohorts using the propensity score matching method described above. To ensure that the matching procedure was effective, we generated plots for case/control covariate ratios both before and after the matching and used the chi-square test for independence to verify that these ratios are significantly different. For every covariate included in PSM, the difference before and after matching was highly significant, indicating that the dataset was highly unbalanced before PSM, and well-balanced after PSM. Recall that "balanced" in terms of PSM does not necessarily mean equal – rather, the counts of patients within each demographic group were sampled in a way that minimizes estimation bias contributed from each model covariate. For example, the most prevalent ethnicity in our dataset by far is "White – British", in both the original and PSM-balanced datasets. Full details and visualizations of PSM are provided in Supplementary Information. The final, balanced dataset includes 2,895 patients (579 cases, 2,316 controls) with approximately equal distributions of all covariates in the two cohorts.

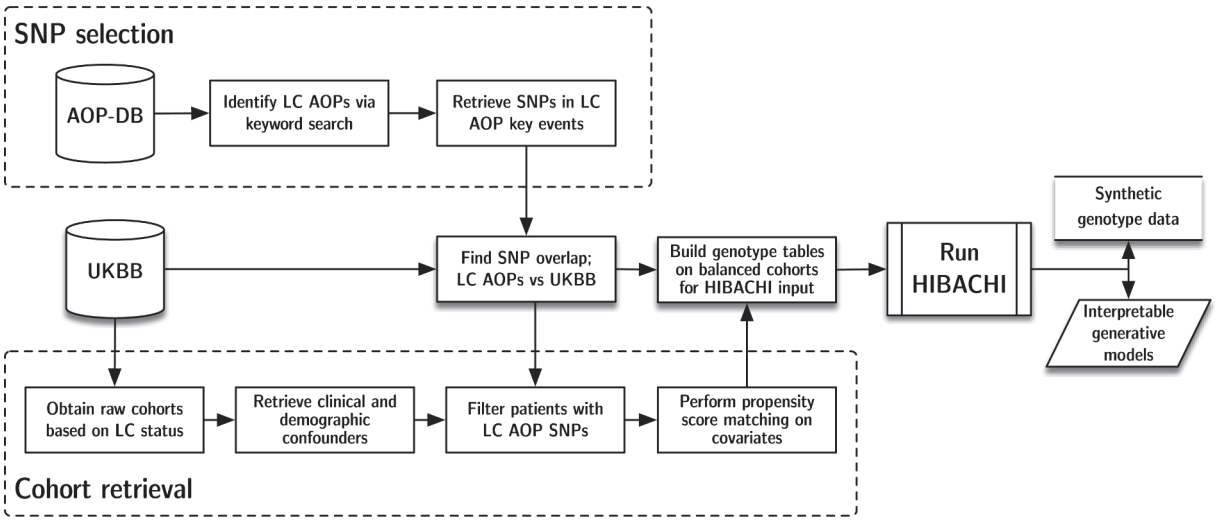

Fig. 1. Building balanced cohorts for learning interactions between AOP key events using HIBACHI.

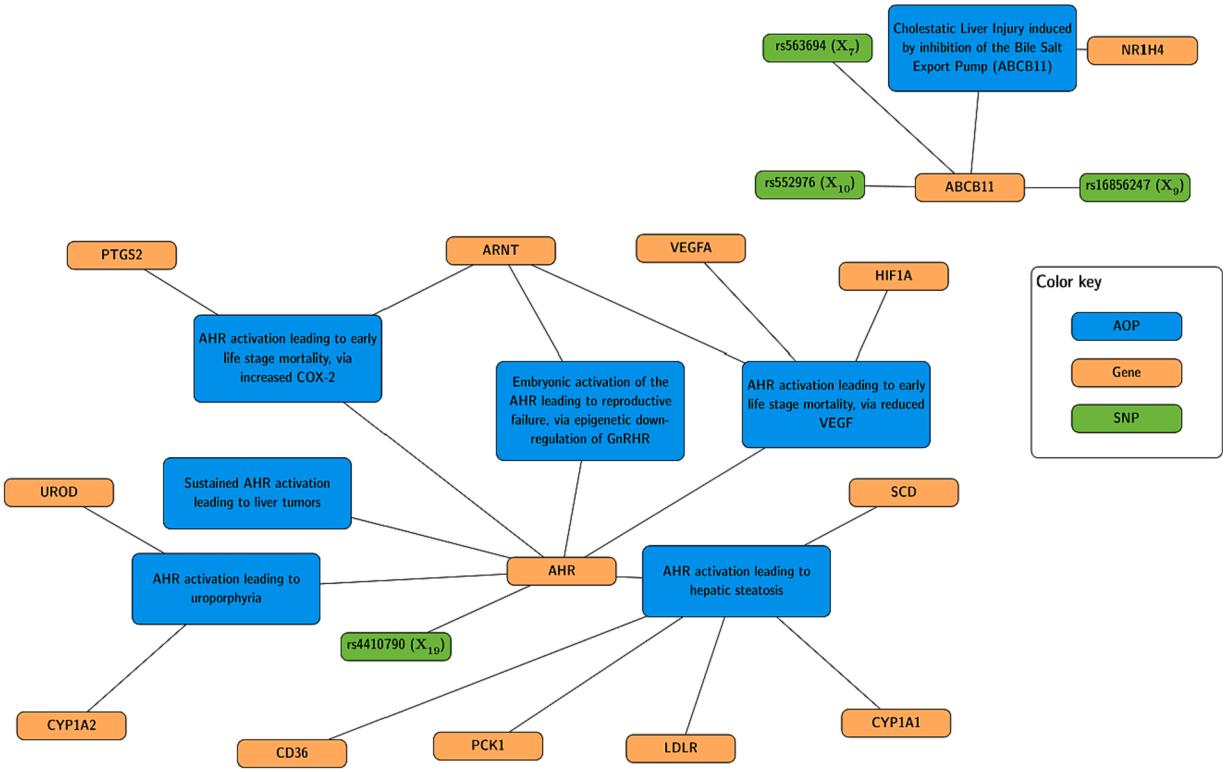

Fig. 2. Network diagram highlighting the SNPs found in HIBACHI's most fit models, along with the network context of their associated genes and AOPs. Note that SNPs not highlighted by the HIBACHI models are omitted. A full network of all LC AOPs along with their full sets of associated genes and SNPs can be found in Supplemental Figure S1.

We matched 4 controls to each case, to help compensate for the relative rarity of LC in the overall patient population.

3.3. Using HIBACHI to infer AOP-related genetic interactions

The 7 best (i.e., having the highest fitness score on the balanced input dataset) models found by HIBACHI are shown in Table 2. Since our response variable (LC) is encoded as a binary target in the dataset (1 = LC, 0 = no LC), the models generally only produce binary outcomes. Note that specific motifs are replicated across several of the best models, which are indicative of robust relationships between specific SNPs that influence risk, as well as the evolutionary nature of HIBACHI's algorithm–the most fit models in each generation 'survive' and are subject to refinement by mutation in subsequent generations. Refer to the Discussion section for a more complete interpretation of the interactions suggested by the most fit models. The 4 SNPs that appear repeatedly in the 7 most fit models are X_7 (rs563694), X_{10} (rs552976), X_{19} (rs4410790), and X_9 (rs16856247).

Table 1
Each relevant SNP, along with respective gene, associated AOP_id, and how each SNP was represented in the HIBACHI program.

SNP	Gene (Hugo)	AOP_id	Hibachi identifier
rs2025516	NR1I3	107	X_1
rs4073054	NR1I3	107	X_2
rs115624142	NR1I3	107	X_3
rs116791819	NR1I3	107	X_4
rs12069336	NR1I3	107	X_5
rs72884586	ABCB11	27	X_6
rs563694	ABCB11	27	X_7
rs569805	ABCB11	27	X_8
rs16856247	ABCB11	27	X_9
rs552976	ABCB11	27	X_{10}
rs2287623	ABCB11	27	X_{11}
rs16856332	ABCB11	27	X_{12}
rs10172795	ABCB11	27	X_{13}
rs117263259	AHR	41	X_{14}
rs71540771	AHR	41	X_{15}
rs117132860	AHR	41	X_{16}
rs4476901	AHR	41	X_{17}
rs115256444	AHR	41	X_{18}
rs4410790	AHR	41	X_{19}
rs6968865	AHR	41	X_{20}
rs12670403	AHR	41	X_{21}
rs11109969	NR1H4	27	X_{22}
rs1625895	TP53	46	X_{23}
rs4253772	PPARA	37	X_{24}
rs5031002	AR	117	X_{25}

3.4. Estimating independence of identified SNPs via linkage disequilibrium

Of the 4 well-represented SNPs in the HIBACHI models, only X_7 (rs563694) and X_{10} (rs552976) were found to be in linkage disequilibrium (LD) with a r^2 of 0.419 and a D' of 0.868 [33]. Therefore, motifs involving both X_7 and X_{10} (which are present in the top 4 models in Table 2) should be treated as suspect, since the SNPs will tend to segregate together. Nonetheless, their presence in the top 4 models is evidence that HIBACHI indeed detects meaningful patterns in the SNP data and incorporates those patterns into its learned generative models. Full results of the LD analysis are given in Supplemental Table S3.

4. Discussion

The 4 SNPs implicated by HIBACHI are members of 2 AOPs: *Cholestatic Liver Injury induced by Inhibition of the Bile Salt Export Pump (ABCB11)*, and *Sustained AhR Activation leading to Rodent Liver Tumors*. Since these two AOPs directly implicate key roles played by the *Abcb11* and *Ahr* genes, these can be thought of as the central mediators of genetic risk to toxicity-induced LC. However, although these genes may be the most important in terms of disease etiology, the HIBACHI-identified SNPs may instead serve as regulatory mechanisms that influence the tendency of those genes to result in a disease state. The implications of this finding could impact many areas of research, including suggesting new therapeutic targets for treatment/prevention, previously unknown stressors, or even new subtypes of LC (e.g., hepatocellular carcinoma, cholangiocarcinoma, etc.) with different etiologies and progression of disease. This study shows how genetic programming can be leveraged to create new hypotheses for future targeted investigations. Although we focused particularly on LC, our approach should be easily generalizable to other diseases, given adequate AOP coverage for the disease and sufficient observational data to construct the respective cohorts.

Although in this study we demonstrate the ability of genetic programming and heuristic simulation to gain insights into the genetic mechanisms underlying toxicity risk, we have not yet explored the influence of specific stressors (i.e., toxicants) on these genetic mechanisms. For example, a genetic factor might influence whether an AOP is triggered by a certain stressor, but not other stressors. The key proteins involved in the two AOPs we describe above (ABCB11 and AHR) are well-studied and many ligands have been established. Currently known stressors for ABCB11 include cholestasis-inducing drugs (e.g., cyclosporine A, rifampicin, others) [30]. AHR has many known stressors, including the two diverse families known as the halogenated and polycyclic aromatic hydrocarbons [14]. As of now, no stressors are formally encoded for these two AOPs in either the AOP-DB or the AOP-Wiki, but we expect these data will be completed as data curation efforts for computational toxicology continue to mature.

4.1. Generative models are suggestive of epistatic interactions in conferring LC risk

As we discussed previously, the generative models produced by HIBACHI (see Table 2) are interpretable mathematical models that can be used to generate synthetic data with similar characteristics to the training data. These models can be likened to symbolic regression models, albeit computed using a brute-force search process with evolutionary refinement rather than via convex optimization. Therefore, specific operations in the highest ranked generative models should correspond to robust patterns that distinguish cases (LC) from non-cases (no LC) patients in the training dataset. When considered with the results of our linkage disequilibrium analysis, the most common motif in the most-fit models is (X_{10} *XOR* X_7). Although these two genes are indeed in LD, the influential role they play suggests one or both could be highly significant in conferring LC risk, when considered in conjunction with the other LC AOP SNPs that appear in the most fit models.

Table 2
Most fit generative models learned by HIBACHI, trained on the propensity score matched UKBB genotype dataset. Along with the model, HIBACHI produces a synthetic version of the training dataset constructed using that model. Higher fitness scores indicate better approximation of the training dataset. Arithmetic operations are applied to the values (0, 1, or 2) comprising the input dataset – for example, "X_{10}!" indicates "the factorial of the value representing SNP X_{10}". "XOR" and "AND" are logical Boolean operations, and 'mod' is the modulo operation.

Individual	Model	Fitness Score
1	$((X_{10}\ \texttt{XOR}\ X_7)\ \texttt{AND}\ (X_{19} - (X_7 \neq ((\log_{X_{10}}(X_2) + X_{19}!) \bmod 2))))$	2.730
2	$((X_{10}\ \texttt{XOR}\ X_7)\ \texttt{AND}\ (X_{19} - (X_7 \neq X_9!)))$	2.280
3	$((X_{10}\ \texttt{XOR}\ X_7)\ \texttt{AND}\ (X_{19} - (X_7 \neq X_{24})))$	2.244
4	$((X_{10}\ \texttt{XOR}\ X_{19})\ \texttt{AND}\ (X_{19} - X_7))$	2.204
5	$(X_{19} + (\neg X_6)) \bmod 2$	1.824
6	$X_9 \neq X_{19}$	1.254
7	$X_{10}!$	0.232

Another SNP that is highly prevalent in the most fit models is X_{19} (rs4410790; within the *AHR* gene). When taken into consideration with X_{10} and X_7 (rs552976 and rs563694, respectively; both within *ABCB11*), the HIBACHI models suggest an epistatic interaction involving both *AHR* (the aryl hydrocarbon receptor; a transcription factor that plays a significant role in detecting and metabolizing xenobiotic chemicals in the liver) [14] and *ABCB11* (which encodes the bile salt export pump protein, a key component in normal, healthy function of the liver) [34]. In each of the top 4 models, HIBACHI yields two motifs–one containing two SNPs within the *ABCB11* gene, and the other containing at least one SNP from the *AHR* gene, along with another, variable number of other SNPs–that are joined by the boolean "AND" operation, meaning that both motifs must evaluate to "1" to result in an outcome of LC in the generative models. In other words, variation in multiple genes involved in the same AOP is required to observe the disease phenotype.

Although these findings are not yet supported by robust experimental evidence, they are highly biologically plausible: *AHR* is a key player in the liver's toxic response, and *ABCB11* governs a central role of the liver; therefore, it would make sense that an interaction between *both* genes helps govern risk for toxicity-mediated liver outcomes. This association could be highly significant clinically, and merits further investigation, either by investigating larger sets of observational data, performing studies in animal models, or both.

Since population-specific prevalence of alleles can affect both the learning of the model (e.g. with populations that have high- or low-prevalence alleles acting as confounders) and the generalizability of results (the study having limited benefit for populations with a low presence of implicated alleles), it is critically important to inspect the prevalence across included populations when interpreting results [8,21]. The 3 SNPs we list above (rs4410790, rs552976, and rs563694) are generally consistent across populations in the 1000 Genotypes Project Phase 3 dataset [4], with the exception of rs55297 in East Asian groups (overall variant allele frequency 0.748; East Asian VAF 0.993) and rs563694 also in East Asian groups (overall VAF 0.158; East Asian VAF 0.026). Therefore, these results may be of limited value to individuals from an East Asian background. However, since self-reported ethnicity was one of the confounders included in propensity score matching, the actual impact of this phenomenon on the HIBACHI model and the interpretation of our results should be minimal. We strongly encourage users of our methodology to carefully inspect population frequencies of alleles, particularly for SNPs that are present in the learned generative model. Whenever possible, users should also choose diverse datasets that are well-annotated with patient demographics and supported by rigorous previous analyses.

4.2. Adjusting cohorts using propensity score matching

Any statistical model learned on observational data is at risk of becoming biased due to the presence of confounding variables. We used the propensity score matching technique–which is a well-established technique in health data research–to select case and control cohorts with similar (almost identical) distributions of a number of important covariates and show in the process that these covariates are significantly unbalanced before the PSM procedure is applied. This leads to two important phenomena. First, there is minimal risk of the interacting SNPs discovered by HIBACHI to reflect associations with the covariates rather than the outcome of interest (in this case, LC). Second, the resulting models should generalize better across different clinical subpopulations. For example, since we included the Townsend deprivation score–a composite measure of material deprivation–and ethnicity in the set of PSM covariates, we can ensure that our case and control cohorts are more socioeconomically and racially balanced than if we did not match on them. Historically, failure to do so has led to study results that do not generalize to underrepresented groups. Beyond these social justice implications, failure to adjust for these factors may lead to constructing a case cohort with less access to high-quality medical care, and therefore worse outcomes or clinical data quality.

4.3. Limitations

Our analysis comes with some limitations, and they are seen through the HIBACHI model results, as the models aren't independent experiments. Some genetic motifs in the models can be artifacts of the evolutionary process rather than meaningful genetic interactions and future HIBACHI analysis will need to account for this. We also need to dig deeper into the biological relationships between the SNPs by running validation experiments. Finally, the underlying premise behind using HIBACHI to perform this analysis is that we hope to capture the 'most fit' models simply by random search followed by refinement using evolutionary algorithms, which can be considered a somewhat 'brute force' approach. A more computationally efficient approach would involve the use of symbolic regression instead of an evolutionary algorithm to explore the search space of all potential generative models. However, we consider our current approach to be both effective and appropriate for this new area of investigation, as the behavior of genetic phenomena (with possible hidden interactions between features) is poorly understood with respect to symbolic regression, and therefore symbolic regression algorithms may not be well-adapted to this task at the current time.

4.4. Future work

We want to explore how HIBACHI works for other adverse outcomes of interest. This is the first time HIBACHI was used to interpret biological relationships from its learned models, and we need to repeat this type of analysis in other scenarios to fully characterize its ability to recognize meaningful biological relationships. The adverse outcome of interest for our future analysis is cardiovascular disease. Cardiovascular disease is of interest because the heart and blood vessels are notably affected as a result of COVID-19 infection, the disease that has caused a global pandemic for over two years. A list of cardiovascular based AOPs and SNPs have been established by queries in the AOP-DB using the same process described in the genotype section of the methods to run in the HIBACHI program for future study. Future analysis should apply the phenotypic data gathering process to different populations. UKBB was our first population of analysis, but we want to apply HIBACHI to other populations as well to explore the robust nature of genetic relationships through patient populations. Another step for future studies is to look at the individual contribution each SNP has for risk of liver cancer through statistical and experimental methods. An example would be to leverage CRISPR-Cas9 nickase technology to selectively edit candidate single nucleotides in cell culture [3] and evaluate the impact of later measurable key events that are predicted to be modulated. This method can therefore quantitate the impact of SNPs that are identified by HIBACHI and provide validation for these computational predictions.

5. Conclusions

In this study, we show that genetic programming and graph data science can be leveraged to uncover patterns of genetic regulation in adverse outcome pathways using real-world observational data. Our approach provides one of the first concrete examples of using HIBACHI – an open-source software tool originally designed to create synthetic datasets with interactions between features – on a task that increases our understanding of biological phenomena. We describe a novel association between variants in the *AHR* and *ABCB11* gene that – when occurring simultaneously – seem to confer increased risk for liver cancer. As a side effect, we also provide a concrete example of using HIBACHI to generate synthetic versions of genetic data, which enables the sharing of genetic data without risks to patient privacy. Furthermore, the technique we use for balancing data with respect to a score of socioeconomic deprivation provides a means for improving social justice in epidemiological

analyses of environmental health. We feel that this study represents the first in a larger body of work exploring how genetic programming can be used to improve our understanding of the genetic mechanisms underlying disease, as well as clinical phenomena resulting from toxic exposures.

CRediT authorship contribution statement

Joseph D. Romano: Visualization. **Liang Mei:** Visualization. **Jonathan Senn:** Visualization. **Jason H. Moore:** Software. **Holly M. Mortensen:** Conceptualized the study methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported by the Environmental Protection Agency's National Research Program in Chemical Safety and Sustainability, Adverse Outcome Pathway Discovery and Development (FY22 CSS AOPDD 4.3.2.2). This research has been conducted using data from UK Biobank, a major biomedical database. The work was additionally funded using grant support from the US National Institutes of Health: K99-LM013646 (PI: Romano), R01-AG066833, R01-LM010098, R01-LM013463 (PI: Moore), and P30-ES013508 (PI: Trevor Penning). We would like to thank Dr. Nisha Sipes and Dr. Brian Chorley for providing editorial review of the manuscript prior to submission.

EPA Disclaimer

This manuscript has been reviewed by the Center for Public Health and Environmental Assessment, United States Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency nor does mention of trade names or commercial products constitute endorsement or recommendation for use. The authors declare no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.comtox.2023.100261.

References

[1] T. Aittokallio, Graph-based methods for analysing networks in cell biology, Briefings in Bioinformatics 7 (3) (2006) 243–255, https://doi.org/10.1093/bib/bbl022.

[2] G.T. Ankley, R.S. Bennett, R.J. Erickson, D.J. Hoff, M.W. Hornung, R.D. Johnson, D.R. Mount, J.W. Nichols, C.L. Russom, P.K. Schmieder, J.A. Serrrano, J.E. Tietge, D.L. Villeneuve, Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment, Environmental Toxicology and Chemistry 29 (3) (2010) 730–741, https://doi.org/10.1002/etc.34.

[3] A.V. Anzalone, P.B. Randolph, J.R. Davis, A.A. Sousa, L.W. Koblan, J.M. Levy, P. J. Chen, C. Wilson, G.A. Newby, A. Raguram, D.R. Liu, Search-and-replace genome editing without double-strand breaks or donor DNA, Nature 576 (7785) (2019) 149–157, https://doi.org/10.1038/s41586-019-1711-4.

[4] A. Auton, G.R. Abecasis, D.M. Altshuler, R.M. Durbin, G.R. Abecasis, D.R. Bentley, A. Chakravarti, A.G. Clark, P. Donnelly, E.E. Eichler, P. Flicek, S.B. Gabriel, R. A. Gibbs, E.D. Green, M.E. Hurles, B.M. Knoppers, J.O. Korbel, E.S. Lander, C. Lee, H. Lehrach, E.R. Mardis, G.T. Marth, G.A. McVean, D.A. Nickerson, J.P. Schmidt, S. T. Sherry, J. Wang, R.K. Wilson, R.A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. i. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J.G. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B.o. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y.e. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E.S. Lander, D. M. Altshuler, S.B. Gabriel, N. Gupta, N. Gharani, L.H. Toji, N.P. Gerry, A.M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S.E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W.M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R.E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D.R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M.W. Albrecht, V. S. Amstislavskiy, T.A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E.R. Mardis, R.K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G.A. McVean, R.M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T.M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J.P. Schmidt, C.J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C.L. Campbell, Y.u. Kong, A. Marcketta, R.A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L.J.M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G.T. Marth, E.P. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A.N. Ward, J. Wu, M. Zhang, M. J. Daly, M.A. DePristo, R.E. Handsaker, D.M. Altshuler, E. Banks, G. Bhatia, G. del Angel, S.B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E.S. Lander, S. A. McCarroll, J.C. Nemesh, R.E. Poplin, S.C. Yoon, J. Lihm, V. Makarov, A.G. Clark, S. Gottipati, A. Keinan, J.L. Rodriguez-Flores, J.O. Korbel, T. Rausch, M.H. Fritz, A. M. Stütz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W.M. McLaren, G.R. S. Ritchie, R.E. Smith, D. Zerbino, X. Zheng-Bradley, P.C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D.N. Cooper, E.V. Ball, P.D. Stenson, D.R. Bentley, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E.E. Kenny, M.A. Batzer, M.K. Konkel, J.A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V.S. Amstislavskiy, R. Herwig, E.R. Mardis, L. i. Ding, D.C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. Frederick Willems, J.T. Simpson, M. D. Shriver, J.A. Rosenfeld, C.D. Bustamante, S.B. Montgomery, F.M. De La Vega, J. K. Byrnes, A.W. Carroll, M.K. DeGorter, P. Lacroute, B.K. Maples, A.R. Martin, A. Moreno-Estrada, S.S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, C. Lee, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F.C.L. Hyland, D.W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S.A. Sinari, K. Squire, S.T. Sherry, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E.G. Burchard, R.D. Hernandez, C.R. Gignoux, D. Haussler, S.J. Katzman, W. James Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S.E. Devine, G.R. Abecasis, H. Min Kang, J.M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. Kate Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, G.A. McVean, J.L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T.K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. E. Eichler, B.L. Browning, S.R. Browning, F. Hormozdiari, P.H. Sudmant, E. Khurana, R.M. Durbin, M.E. Hurles, C. Tyler-Smith, C.A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T.M. Keane, S. McCarthy, K. Walter, Y. Xue, M.B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A.O. Harmanci, M. Jin, D. Lee, J. Liu, X. Jasmine Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G.T. Marth, E.P. Garrison, D. Kural, W.-P. Lee, A.N. Ward, J. Wu, M. Zhang, S.A. McCarroll, R. E. Handsaker, D.M. Altshuler, E. Banks, G. del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J.C. Nemesh, K. Shakir, S.C. Yoon, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel, M.H. Fritz, S. Meiers, B. Raeder, T. Rausch, A.M. Stütz, P. Flicek, F. Paolo Casale, L. Clarke, R.E. Smith, O. Stegle, X. Zheng-Bradley, D.R. Bentley, B. Barnes, R. Keira Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E.-W. Lameijer, M.A. Batzer, M.K. Konkel, J.A. Walker, L.i. Ding, I. Hall, K. Ye, P. Lacroute, C. Lee, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D.W. Craig, N. Homer, D. Church, C. Xiao, J. Sebat, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S.E. Devine, E.J. Gardner, G.R. Abecasis, J.M. Kidd, R.E. Mills, G. Dayama, S. Emery, G. Jun, X. Shi, A. Quitadamo, G. Lunter, G.A. McVean, K. Chen, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E.E. Eichler, M.J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P.H. Sudmant, N.F. Parrish, E. Khurana, M.E. Hurles, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M.B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. Jasmine Mu, C. Sisu, J. Zhang, Y. Zhang, R.A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U.S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G.T. Marth, E.P. Garrison, W. Fung Leong, A.N. Ward, G. del Angel, M.A. DePristo, S.B. Gabriel, N. Gupta, C. Hartl, R.E. Poplin, A.G. Clark, J.L. Rodriguez-Flores, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, D.G. MacArthur, E.R. Mardis, R. Fulton, D.C. Koboldt, S. Gravel, C.D. Bustamante, D.W. Craig, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry, C. Xiao, E.T. Dermitzakis, G.R. Abecasis, H. Min Kang, G.A. McVean, M. B. Gerstein, S. Balasubramanian, L. Habegger, H. Yu, P. Flicek, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage, J. Berg Jespersen, H. Horn, S.B. Montgomery, M.K. DeGorter, E. Khurana, C. Tyler-Smith, Y. Chen, V. Colonna, Y. Xue, M.B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, A. Auton, A. Marcketta, R. Desalle, A. Narechania, M.A. Wilson Sayres, E. P. Garrison, R.E. Handsaker, S. Kashin, S.A. McCarroll, J.L. Rodriguez-Flores, P. Flicek, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. Frederick Willems, C.D. Bustamante, F.L. Mendez, G. David Poznik, P.A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G.R. Abecasis, L. Coin, H. Shao, D. Mittelman, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T.

W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G.R. Ritchie, Y. Xue, F. Yang, R.A. Gibbs, C. Kovar, D. Kalra, W. Hale, D. Muzny, J.G. Reid, J. Wang, X.u. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D.M. Altshuler, P. Flicek, L. Clarke, X. Zheng-Bradley, D.R. Bentley, A. Cox, S. Humphray, S. Kahn, R. Sudbrak, M.W. Albrecht, M. Lienhard, D. Larson, D.W. Craig, T. Izatt, A.A. Kurdoglu, S.T. Sherry, C. Xiao, D. Haussler, G.R. Abecasis, G.A. McVean, R.M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti, B.M. Knoppers, G.R. Abecasis, K. C. Barnes, C. Beiswanger, E.G. Burchard, C.D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N.P. Gerry, N. Gharani, R.A. Gibbs, C.R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J.S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P.N. Ossorio, M. Parker, A.M. Resch, C.N. Rotimi, C.D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L.H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T.K. Oleksyk, K.C. Barnes, R.A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, P.C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. Sisay Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, Trần Tịnh Hiếu, S.J. Dunstan, N. Thuy Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P.C. Sabeti, J. Schieffelin, D.S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L.D. Brooks, A.L. Felsenfeld, J.E. McEwen, Y. Vaydylevich, E.D. Green, A. Duncanson, M. Dunn, J.A. Schloss, J. Wang, H. Yang, A. Auton, L.D. Brooks, R. M. Durbin, E.P. Garrison, H. Min Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G. A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.

[5] U. Benedetto, S.J. Head, G.D. Angelini, E.H. Blackstone, Statistical primer: Propensity score matching and its alternatives, European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery 53 (6) (2018) 1112–1117, https://doi.org/10.1093/ejcts/ezy167.

[6] H.-G. Beyer, H.-P. Schwefel, Evolution strategies—A comprehensive introduction, Natural Computing 1 (1) (2002) 3–52, https://doi.org/10.1023/A:1015059928466.

[7] B. Bollobás, Modern Graph Theory, Springer, 1998.

[8] D.G. Clayton, N.M. Walker, D.J. Smyth, R. Pask, J.D. Cooper, L.M. Maier, L. J. Smink, A.C. Lam, N.R. Ovington, H.E. Stevens, S. Nutland, J.M.M. Howson, M. Faham, M. Moorhead, H.B. Jones, M. Falkowski, P. Hardenbol, T.D. Willis, J. A. Todd, Population structure, differential bias and genomic control in a large-scale, case-control association study, Nature Genetics 37 (11) (2005) 1243–1246.

[9] S. Fortunato, Community detection in graphs, Physics Reports 486 (3–5) (2010) 75–174, https://doi.org/10.1016/j.physrep.2009.11.002.

[10] P. François, E.D. Siggia, A case study of evolutionary computation of biochemical adaptation, Physical Biology 5 (2) (2008), 026009, https://doi.org/10.1088/1478-3975/5/2/026009.

[11] A.S. Fraser, Monte Carlo analyses of genetic models, Nature 181 (4603) (1958) 208–209, https://doi.org/10.1038/181208a0.

[12] G. Gallo, S. Pallottino, Shortest path algorithms, Annals of Operations Research 13 (1) (1988) 1–79, https://doi.org/10.1007/BF02288320.

[13] GTEx Consortium, E.R. Gamazon, A.V. Segrè, M. van de Bunt, X. Wen, H.S. Xi, F. Hormozdiari, H. Ongen, A. Konkashbaev, E.M. Derks, F. Aguet, J. Quan, D. L. Nicolae, E. Eskin, M. Kellis, G. Getz, M.I. McCarthy, E.T. Dermitzakis, N.J. Cox, K.G. Ardlie, Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation, Nature Genetics 50 (7) (2018) 956–967, https://doi.org/10.1038/s41588-018-0154-4.

[14] O. Hankinson, The aryl hydrocarbon receptor complex, Annual Review of Pharmacology and Toxicology 35 (1995) 307–340, https://doi.org/10.1146/annurev.pa.35.040195.001515.

[15] W. Huber, V.J. Carey, L. Long, S. Falcon, R. Gentleman, Graphs in molecular biology, BMC Bioinformatics 8 (S6) (2007) S8, https://doi.org/10.1186/1471-2105-8-S6-S8.

[16] A.G.J. MacFarlane, M.o. Jamshidi, Tools for intelligent control: Fuzzy controllers, neural networks and genetic algorithms, Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 361 (1809) (2003) 1781–1808.

[17] R.J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R.W. Setzer, I. Shah, D. Villeneuve, E. Weber, Computational Toxicology—A State of the Science Mini Review, Toxicological Sciences 103 (1) (2008) 14–27, https://doi.org/10.1093/toxsci/kfm297.

[18] La Cava, William, Orzechowski, Patryk, Burlacu, Bogdan, de Franca, Fabricio Olivetti, Virgolin, Marco, Jin, Ying, Kommenda, Michael, & Moore, Jason H. (2021, June 6). Contemporary Symbolic Regression Methods and their Relative Performance. *NeurIPS 2021 Track Datasets and Benchmarks (Round 1)*. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). https://openreview.net/forum?id=xVQMrDLyGst.

[19] X. Li, X. Wang, P. Gao, Diabetes Mellitus and Risk of Hepatocellular Carcinoma, BioMed Research International 2017 (2017) 5202684, https://doi.org/10.1155/2017/5202684.

[20] M.J. Machiela, S.J. Chanock, LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants, Bioinformatics (Oxford, England) 31 (21) (2015) 3555–3557, https://doi.org/10.1093/bioinformatics/btv402.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys 54 (6) (2022) 1–35.

[22] R. Miikkulainen, S. Forrest, A biological perspective on evolutionary computation, Nature Machine Intelligence 3 (1) (2021) 9–15, https://doi.org/10.1038/s42256-020-00278-8.

[23] J.F. Miller, Cartesian Genetic Programming, in: J.F. Miller (Ed.), Cartesian Genetic Programming, Springer, Berlin Heidelberg, 2011, pp. 17–34, https://doi.org/10.1007/978-3-642-17310-3_2.

[24] Moore, J. H., Olson, R. S., Schmitt, P., Chen, Y., & Manduchi, E. (2018). How computational thought experiments can improve our understanding of the genetic architecture of common human diseases. *The 2018 Conference on Artificial Life*, 23–30. https://doi.org/10.1162/isal_a_00012.

[25] J.H. Moore, M. Shestov, P. Schmitt, R.S. Olson, 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*, Pacific Symposium on Biocomputing 23 (2018) 259–267.

[26] H.M. Mortensen, J. Senn, T. Levey, P. Langley, A.J. Williams, The 2021 update of the EPA's adverse outcome pathway database, Scientific Data 8 (1) (2021) 169, https://doi.org/10.1038/s41597-021-00962-3.

[27] M. Needham, A.E. Hodler, Graph Algorithms: Practical Examples in Apache Spark and Neo4j, 1st Ed., O'Reilly Media, 2019.

[28] R.S. Olson, R.J. Urbanowicz, P.C. Andrews, N.A. Lavender, L.C. Kidd, J.H. Moore, Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, in: G. Squillero, P. Burelli (Eds.), Applications of Evolutionary Computation, Springer International Publishing, 2016, pp. 123–137, https://doi.org/10.1007/978-3-319-31204-0_9.

[29] L.J. Palmer, UK Biobank: Bank on it, Lancet (London, England) 369 (9578) (2007) 1980–1982, https://doi.org/10.1016/S0140-6736(07)60924-6.

[30] J.M. Pedersen, P. Matsson, C.A.S. Bergström, J. Hoogstraate, A. Norén, E. L. LeCluyse, P. Artursson, Early identification of clinically relevant drug interactions with the human bile salt export pump (BSEP/ABCB11), Toxicological Sciences: An Official Journal of the Society of Toxicology 136 (2) (2013) 328–343, https://doi.org/10.1093/toxsci/kft197.

[31] J.D. Romano, Y. Hao, J.H. Moore, T.M. Penning, Automating Predictive Toxicology Using ComptoxAI, Chemical Research in Toxicology 35 (8) (2022) 1370–1382, https://doi.org/10.1021/acs.chemrestox.2c00074.

[32] I.N. Sarkar, A.J. Butte, Y.A. Lussier, P. Tarczy-Hornoch, L. Ohno-Machado, Translational bioinformatics: Linking knowledge across biological and clinical realms: Figure 1, Journal of the American Medical Informatics Association 18 (4) (2011) 354–357, https://doi.org/10.1136/amiajnl-2011-000245.

[33] M. Slatkin, Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future, Nature Reviews Genetics 9 (6) (2008) 477–485, https://doi.org/10.1038/nrg2361.

[34] B. Stieger, Y. Meier, P.J. Meier, The bile salt export pump, Pflügers Archiv - European Journal of Physiology 453 (5) (2007) 611–620, https://doi.org/10.1007/s00424-006-0152-8.

[35] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age, PLOS Medicine 12 (3) (2015) e1001779.

[36] I.V. Tetko, G. Klambauer, D.-A. Clevert, I. Shah, E. Benfenati, Artificial Intelligence Meets Toxicology, Chemical Research in Toxicology 35 (8) (2022) 1289–1290, https://doi.org/10.1021/acs.chemrestox.2c00196.

[37] The PRACTICAL consortium, Law, M. Timofeeva, C. Fernandez-Rozadilla, P. Broderick, J. Studd, J. Fernandez-Tajes, S. Farrington, V. Svinti, C. Palles, G. Orlando, A. Sud, A. Holroyd, S. Penegar, E. Theodoratou, P. Vaughan-Shaw, H. Campbell, L. Zgaga, C. Hayward, A. Campbell, S. Harris, I.J. Deary, J. Starr, L. Gatcombe, M. Pinna, S. Briggs, L. Martin, E. Jaeger, A. Sharma-Oates, J. East, S. Leedham, R. Arnold, E. Johnstone, H. Wang, D. Kerr, R. Kerr, T. Maughan, R. Kaplan, N. Al-Tassan, K. Palin, U.A. Hänninen, T. Cajuso, T. Tanskanen, J. Kondelin, E. Kaasinen, A.-P. Sarin, J.G. Eriksson, H. Rissanen, P. Knekt, E. Pukkala, P. Jousilahti, V. Salomaa, S. Ripatti, A. Palotie, L. Renkonen-Sinisalo, A. Lepistö, J. Böhm, J.-P. Mecklin, D.D. Buchanan, A.-K. Win, J. Hopper, M. E. Jenkins, N.M. Lindor, P.A. Newcomb, S. Gallinger, D. Duggan, G. Casey, P. Hoffmann, M.M. Nöthen, K.-H. Jöckel, D.F. Easton, P.D.P. Pharoah, J. Peto, F. Canzian, A. Swerdlow, R.A. Eeles, Z. Kote-Jarai, K. Muir, N. Pashayan, B. E. Henderson, C.A. Haiman, F.R. Schumacher, A.A. Al Olama, S. Benlloch, S. I. Berndt, D.V. Conti, F. Wiklund, S. Chanock, S. Gapstur, V.L. Stevens, C. M. Tangen, J. Batra, J. Clements, H. Gronberg, J. Schleutker, D. Albanes, A. Wolk, C. West, L. Mucci, G. Cancel-Tassin, S. Koutros, K.D. Sorensen, E.M. Grindedal, D. E. Neal, F.C. Hamdy, J.L. Donovan, R.C. Travis, R.J. Hamilton, S.A. Ingles, B. S. Rosenstein, Y.-J. Lu, G.G. Giles, A.S. Kibel, A. Vega, M. Kogevinas, K.L. Penney, J.Y. Park, J.L. Stanford, C. Cybulski, B.G. Nordestgaard, C. Maier, J. Kim, E. M. John, M.R. Teixeira, S.L. Neuhausen, K. De Ruyck, A. Razack, L.F. Newcomb, M. Gamulin, R. Kaneva, N. Usmani, F. Claessens, P.A. Townsend, M. Gago-Dominguez, M.J. Roobol, F. Menegaux, K.-T. Khaw, L. Cannon-Albright, H. Pandha, S.N. Thibodeau, A. Harkin, K. Allan, J. McQueen, J. Paul, T. Iveson, M. Saunders, K. Butterbach, J. Chang-Claude, M. Hoffmeister, H. Brenner, I. Kirac, P. Matošević, P. Hofer, S. Brezina, A. Gsur, J.P. Cheadle, L.A. Aaltonen, I. Tomlinson, R.S. Houlston, M.G. Dunlop, Association analyses identify 31 new risk loci for colorectal cancer susceptibility, Nature Communications 10 (1) (2019), https://doi.org/10.1038/s41467-019-09775-w.

[38] R.S. Thomas, T. Bahadori, T.J. Buckley, J. Cowden, C. Deisenroth, K.L. Dionisio, J. B. Frithsen, C.M. Grulke, M.R. Gwinn, J.A. Harrill, M. Higuchi, K.A. Houck, M. F. Hughes, E.S. Hunter, K.K. Isaacs, R.S. Judson, T.B. Knudsen, J.C. Lambert, M. Linnenbrink, T.M. Martin, S.R. Newton, S. Padilla, G. Patlewicz, K. Paul-Friedman, K.A. Phillips, A.M. Richard, R. Sams, T.J. Shafer, R.W. Setzer, I. Shah, J. E. Simmons, S.O. Simmons, A. Singh, J.R. Sobus, M. Strynar, A. Swank, R. Tornero-Valez, E.M. Ulrich, D.L. Villeneuve, J.F. Wambaugh, B.A. Wetmore, A.J. Williams, The Next Generation Blueprint of Computational Toxicology at the U.S.

Environmental Protection Agency. *Toxicological Sciences: An Official Journal of the Society of*, Toxicology 169 (2) (2019) 317–332.
[39] P. Townsend, Deprivation, Journal of Social Policy 16 (2) (1987) 125–146, https://doi.org/10.1017/S0047279400020341.
[40] Urbanowicz, R. J., Barney, N., White, B. C., & Moore, J. H. (2008). Mask functions for the symbolic modeling of epistasis using genetic programming. *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation - GECCO '08*, 339. https://doi.org/10.1145/1389095.1389154.
[41] M.P. van den Heuvel, O. Sporns, Network hubs in the human brain, Trends in Cognitive Sciences 17 (12) (2013) 683–696, https://doi.org/10.1016/j.tics.2013.09.012.