Genetics and population analysis

EpiGEN: an epistasis simulation pipeline

David B. Blumenthal [1,*], Lorenzo Viola [1], Markus List [1], Jan Baumbach [1], Paolo Tieri [2], and Tim Kacprowski [1,*]

[1] Technical University of Munich, Chair of Experimental Bioinformatics, 85354 Freising, Germany
[2] CNR National Research Council, IAC Institute for Applied Computing, 00185 Rome, Italy

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Simulated data is crucial for evaluating epistasis detection tools in genome-wide association studies. Existing simulators are limited, as they do not account for linkage disequilibrium (LD), support limited interaction models of single nucleotide polymorphisms (SNPs) and only dichotomous phenotypes, or depend on proprietary software. In contrast, EpiGEN supports SNP interactions of arbitrary order, produces realistic LD patterns, and can generate both categorical and quantitative phenotypes.

Availability and Implementation: EpiGEN is implemented in Python 3 and is freely available at https://github.com/baumbachlab/epigen.

Contact: david.blumenthal@wzw.tum.de, tim.kacprowski@wzw.tum.de

Supplementary information: A detailed user guide for EpiGEN is available at *Bioinformatics* online.

1 Introduction

The goal of *genome-wide association studies* (GWAS) is to link genetic variants to phenotypic traits of interest, most commonly a disease (Bush and Moore, 2012). More specifically, GWAS usually look for *biallelic single nucleotid polymorphisms* (SNPs) that are predictive of the phenotype. While thousands of SNPs have been associated with diseases since the early 2000s (MacArthur *et al.*, 2017), they can account only for a fraction of the investigated traits' heritability. The most common hypothesis is that the missing heritability can be explained by *epistasis*, i. e., by interactions between SNPs that are jointly predictive of the phenotype but individually have little or no effect (Manolio *et al.*, 2009). Various epistasis detection tools have been proposed (Jing and Shen, 2015; Niel *et al.*, 2015; Chatelain *et al.*, 2018; Ansarifar and Wang, 2019; Chattopadhyay and Lu, 2019; Cao *et al.*, 2020). For evaluation, availability of simulated data is crucial.

Popular epistasis simulators include GWAsimulator (Li and Li, 2008), HAPGEN2 (Su *et al.*, 2011), GAMETES (Urbanowicz *et al.*, 2012), EpiSIM (Shang *et al.*, 2013), simGWA (Yang and Gu, 2013), and TriadSim (Shi *et al.*, 2018). However, all of them are limited in different aspects (cf. Table 1). Most notably, none of them can generate arbitrary (i. e., dichotomous, categorical, and quantitative) phenotypes or simulate observation bias. Moreover, many tools can simulate only pairwise interactions, are restricted to specific epistasis models (e. g., multiplicative interactions), and are not straight-forward to use as they require third-party input files.

Further approaches for simulating epistasis have been described by Chatelain *et al.* (2018) and Id-Lahoucine *et al.* (2019) but lack of publicly available implementations. Moreover, in (Chatelain *et al.*, 2018), only dichotomous phenotypes are supported, while the algorithm described by Id-Lahoucine *et al.* (2019) does not generate phenotypes but positive, negative, or neutral directions of epistatic selection. There are also various tools for simulating genotype data without phenotypes (Siragusa *et al.*, 2019; Peng *et al.*, 2019; Juan *et al.*, 2020), which could substitute HAPGEN2 in the first step of the simulation pipeline (cf. Sec. 2.1).

Table 1. Features of EpiGEN and existing epistasis simulators.

	GWAsimulator	HAPGEN2	GAMETES	EpiSIM	simGWA	TriadSim	**EpiGEN**
generates arbitrary phenotypes	✗	✗	✗	✗	✗	✗	✓
simulates observation bias	✗	✗	✗	✗	✗	✗	✓
generates realistic LD patterns	✓	✓	✗	✓	✓	✓	✓
supports MAF specification	✗	✗	✓	✓	✓	✗	✓
simulates higher-order interactions	✓	✗	✓	✗	✓	✓	✓
allows arbitrary epistasis models	✗	✗	✓	✗	✓	✗	✓
requires no proprietary software	✓	✓	✓	✗	✓	✓	✓
requires no third-party input files	✗	✗	✓	✓	✗	✓	✓

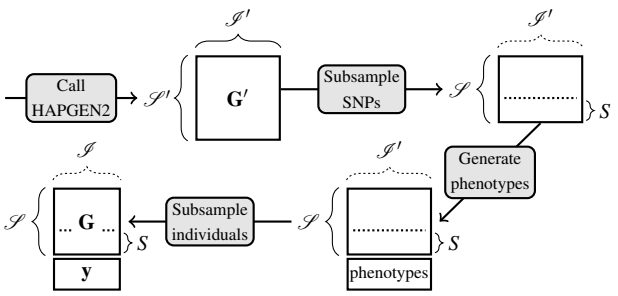


Fig. 1. EpiGEN's simulation pipeline. $\mathbf{G}'$: genotype corpus generated by HAPGEN2; $\mathscr{S}'$ and $\mathscr{I}'$: $\mathbf{G}'$'s sets of SNPs and individuals; $\mathbf{G}$: final genotype matrix; $\mathscr{S} \subseteq \mathscr{S}'$ and $\mathscr{I} \subseteq \mathscr{I}'$: $\mathbf{G}$'s sets of SNPs and individuals; $\mathbf{y}$: phenotype vector; $S \subseteq \mathscr{S}$: set of disease SNPs.

Since existing tools lack numerous features important for comprehensive simulation of epistasis, we have developed EpiGEN—a generically applicable, easy-to-deploy and feature-rich epistasis simulation pipeline.

2 Simulation pipeline

EpiGEN generates a genotype matrix $\mathbf{G} = (g_{s,i}) \in \{0,1,2\}^{|\mathscr{S}|\times|\mathscr{I}|}$, a disease SNP set $S \subset \mathscr{S}$, and a phenotype vector $\mathbf{y} = (y_i) \in \mathbb{P}^{|\mathscr{I}|}$, where $\mathbb{P} = \mathbb{R}$ for quantitative phenotypes and $\mathbb{P} = \{0,\dots,c-1\}$ for discrete phenotypes with c categories. The entry $g_{s,i}$ of $\mathbf{G}$ encodes individual i's number of minor alleles at SNP s. EpiGEN's simulation pipeline (Fig. 1) consists of four sub-routines, described in detail in the following sections:

1. Call HAPGEN2 to generate genotype corpus $\mathbf{G}' \in \{0,1,2\}^{|\mathscr{S}'|\times|\mathscr{I}'|}$.
2. Subsample the SNP sets $\mathscr{S} \subseteq \mathscr{S}'$ and $S \subseteq \mathscr{S}$.
3. Generate the phenotypes by applying an epistasis model to S.
4. Subsample the individuals $\mathscr{I} \subseteq \mathscr{I}'$.

2.1 Generating the genotype corpus

EpiGEN requires (a) a set of chromosomes for which $\mathbf{G}'$ should be generated; (b) a HapMap 3 population code (Altshuler *et al.*, 2010); (c) the number $N_{\mathscr{I}'}$ of individuals in the corpus. For each chromosome, EpiGEN generates a genotype corpus for $N_{\mathscr{I}'}$ individuals by calling HAPGEN2 without disease SNPs for the selected HapMap 3 reference panel. This ensures that LD is modeled adequately. The joint corpus $\mathbf{G}'$ is then obtained as the concatenation of the corpora for the chromosomes. Since this step is by far the computationally most expensive sub-routine, we ship EpiGEN with pre-computed corpora for all chromosomes, all population codes, and 10^4 individuals (around 20 GB download size). Of course, the user can also generate her own corpora. Note that, in principle, HAPGEN2 could be substituted by any genotype simulator that produces realistic LD patterns.

2.2 Subsampling the SNPs

EpiGEN samples SNP sets $\mathscr{S}$ and S that respect user-defined side constraints. To that end, EpiGEN requires (a) the set S of disease SNPs or the size d of S and a range $r_1 \subseteq [0,1]$ of acceptable MAFs for all disease SNPs $s \in S$; (b) the number $N_{\mathscr{S}}$ of SNPs in the final genotype matrix $\mathbf{G}$; (c) a range $r_2 \subseteq [0,1]$ of acceptable MAFs for all SNPs $s \in \mathscr{S} \setminus S$. If S is not provided by the user, EpiGEN samples S by randomly selecting d SNPs from $\mathscr{S}'$ whose MAFs fall into the range r_1. Subsequently, EpiGEN initializes $\mathscr{S} \coloneqq S$ and then extends $\mathscr{S}$ by sampling $N_{\mathscr{S}} - d$ SNPs from $\mathscr{S}' \setminus S$ whose MAFs fall into the range r_2. Both ranges are extended dynamically if too few SNPs satisfy these constraints.

2.3 Generating the phenotypes

Once the SNP sets $\mathscr{S}$ and S have been selected, EpiGEN generates phenotypes for all individuals by applying a user-specified d-dimensional *epistasis model* $M : \{0,1,2\}^d \to \mathscr{D}(\mathbb{P})$. $\mathscr{D}(\mathbb{P})$ denotes a set of suitable probability distributions for phenotypes from $\mathbb{P}$. If $\mathbb{P} = \{0,\dots,c-1\}$, we define $\mathscr{D}(\mathbb{P})$ as the set of all categorical distributions with c categories; if $\mathbb{P} = \mathbb{R}$, $\mathscr{D}(\mathbb{P})$ is the set of all normal distributions. For each individual $i \in \mathscr{I}'$, the phenotype y_i is drawn from the distribution $M(\mathbf{g}_S(i))$, where $\mathbf{g}_S(i) \coloneqq (g_{s,i})_{s\in S}$ is i's genotype at the SNPs contained in S. Epistasis models hence generalize penetrance tables, which are often used to model epistatic interaction in the case $\mathbb{P} = \{0,1\}$ (Wang *et al.*, 2010; Urbanowicz *et al.*, 2012; Shang *et al.*, 2013; Jing and Shen, 2015; Cao *et al.*, 2020).

The epistasis model M can be specified either (a) by a full extensional definition or (b) as a combination of parameterized risk models for dichotomous and quantitative phenotypes. If $\mathbb{P} = \mathbb{R}$, M is extensionally defined via parameters $\mu_{\mathbf{g}}$ and $\sigma_{\mathbf{g}}$ of the normal distribution $M(\mathbf{g})$ for each genotype $\mathbf{g} \in \{0,1,2\}^d$. If $\mathbb{P} = \{0,\dots,c-1\}$, probability vectors $\mathbf{p}_{\mathbf{g}} \in \mathbb{R}^c_{\geq 0}$ for all $\mathbf{g} \in \{0,1,2\}^d$ must be provided. If $c = 2$, these vectors can optionally be generated with tools such as GAMETES designed for this purpose.

For dichotomous and quantitative phenotypes, EpiGEN alternatively provides the option to specify M via a set $\mathscr{M}$ of parametrized d-dimensional *risk models* $R : \{0,1,2\}^d \to \mathbb{R}_{>0}$ (explained below). Given $\mathscr{M}$, EpiGEN defines the joint epistasis model M as follows: For each genotype $\mathbf{g} \in \{0,1,2\}^d$, let $\mathscr{M}(\mathbf{g}) \coloneqq \prod_{R\in\mathscr{M}} R(\mathbf{g})$ be the joint risk of the models contained in $\mathscr{M}$. If $\mathbb{P} = \{0,1\}$, $\mathscr{M}(\mathbf{g})$ represents the disease odd of genotype $\mathbf{g}$, and the vector of probabilities of the categorical distribution $M(\mathbf{g})$ is defined as $\mathbf{p}_{\mathbf{g}} \coloneqq (1 - p_{\mathbf{g}_1}, p_{\mathbf{g}_1})$, where $p_{\mathbf{g}_1} \coloneqq \mathscr{M}(\mathbf{g})/(1+\mathscr{M}(\mathbf{g}))$. If $\mathbb{P} = \mathbb{R}$, $\mathscr{M}(\mathbf{g})$ is interpreted as the mean μ_g of the normal distribution $M(\mathbf{g})$. The standard deviations are globally set to $\sigma_{\mathbf{g}} \coloneqq \sigma$, where σ is provided by the user.

Three different kinds of risk models are available, namely, baseline models R^{bas}_{α}, marginal models $R^{\text{mar}}_{\alpha,t,s}$, and interaction models $R^{\text{int}}_{\alpha,t',S'}$, where $\alpha \in \mathbb{R}_{>0}$, $t \in \{a,d,r\}$, $t' \in \{m,e,d,r\}$, $s \in S$, and $S' \subseteq S$ are parameters provided by the user. R^{bas}_{α} models the baseline risk, i. e., is simply defined as $R^{\text{bas}}_{\alpha}(\mathbf{g}_S(i)) \coloneqq \alpha$. At most one baseline model can be included into $\mathscr{M}$. $R^{\text{mar}}_{s,a,\alpha}$, $R^{\text{mar}}_{s,d,\alpha}$, and $R^{\text{mar}}_{s,r,\alpha}$ model, respectively, additive, dominant, and recessive marginal effects. They are defined as

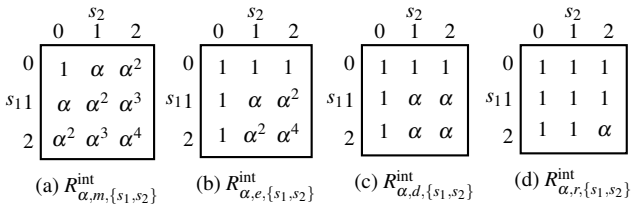
$$R^{\text{mar}}_{\alpha,a,s}(\mathbf{g}_S(i)) \coloneqq [g_{s,i} = 0] + \frac{\alpha \cdot g_{s,i}}{2} \cdot [g_{s,i} > 0]$$
$$R^{\text{mar}}_{\alpha,d,s}(\mathbf{g}_S(i)) \coloneqq [g_{s,i} = 0] + \alpha \cdot [g_{s,i} > 0]$$
$$R^{\text{mar}}_{\alpha,r,s}(\mathbf{g}_S(i)) \coloneqq [g_{s,i} < 2] + \alpha \cdot [g_{s,i} = 2],$$

where $[\texttt{true}] \coloneqq 1$ and $[\texttt{false}] \coloneqq 0$. For each SNP $s \in S$, at most one marginal model can be included into $\mathscr{M}$. Finally, the interaction models $R^{\text{int}}_{\alpha,m,S'}$, $R^{\text{int}}_{\alpha,e,S'}$, $R^{\text{int}}_{\alpha,d,S'}$, and $R^{\text{int}}_{\alpha,r,S'}$ model, respectively, multiplicative, exponential, joint-dominant, and joint-recessive interaction. At most one interaction model per SNP subset $S' \subseteq S$ can be included into $\mathscr{M}$. The interaction models are visualized in Fig. 2 and formally defined as follows:

$$R^{\text{int}}_{\alpha,m,S'}(\mathbf{g}_S(i)) \coloneqq \alpha^{\sum_{s\in S'} g_{s,i}}$$
$$R^{\text{int}}_{\alpha,e,S'}(\mathbf{g}_S(i)) \coloneqq \alpha^{\prod_{s\in S'} g_{s,i}}$$
$$R^{\text{int}}_{\alpha,d,S'}(\mathbf{g}_S(i)) \coloneqq \left(1 - \prod_{s\in S'} [g_{s,i} > 0]\right) + \alpha \cdot \prod_{s\in S'} [g_{s,i} > 0]$$
$$R^{\text{int}}_{\alpha,r,S'}(\mathbf{g}_S(i)) \coloneqq \left(1 - \prod_{s\in S'} [g_{s,i} = 2]\right) + \alpha \cdot \prod_{s\in S'} [g_{s,i} = 2]$$

Further interaction models can easily be implemented by the user. Detailed instructions can be found in EpiGEN's user guide.

Assume, for instance, that we want to generate dichotomous phenotypes with baseline risk 0.25 and a joint-dominant interaction of the SNPs s_1 and s_2 with intensity $\alpha = 6$. Let i_1 and i_2 be individuals whose genotypes at



(a) $R^{\text{int}}_{\alpha,m,\{s_1,s_2\}}$ (b) $R^{\text{int}}_{\alpha,e,\{s_1,s_2\}}$ (c) $R^{\text{int}}_{\alpha,d,\{s_1,s_2\}}$ (d) $R^{\text{int}}_{\alpha,r,\{s_1,s_2\}}$

Fig. 2. Available parametrized interaction models for the case $|S'| = 2$: (a) multiplicative, (b) exponential, (c) joint-dominant, and (d) joint-recessive interaction.

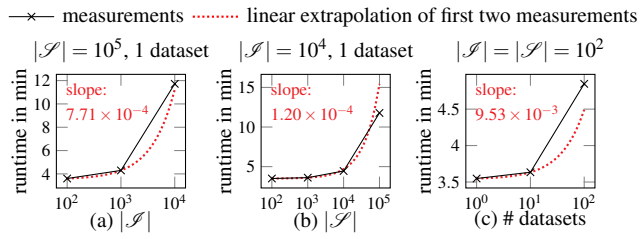


Fig. 3. Runtime vs. number of individuals (a), SNPs (b), and simulated datasets (c).

$S = \{s_1, s_2\}$ equal $\mathbf{g}_S(i_1) = (1,2)$ and $\mathbf{g}_S(i_2) = (2,2)$. Then i_1's and i_2's overall risks are $\mathscr{M}(\mathbf{g}_S(i_1)) = 0.25$ and $\mathscr{M}(\mathbf{g}_S(i_2)) = 1.5$, and i_1's and i_2's phenotypes are drawn from $D_1 = (0.8, 0.2)$ and $D_2 = (0.4, 0.6)$.

2.4 Subsampling the individuals

The final step is to subsample the individuals, i. e., to select the set $\mathscr{I} \subseteq \mathscr{I}'$ of individuals for which the geno- and phenotypes are returned, based on (a) the number $N_{\mathscr{I}} \leq N_{\mathscr{I}'}$ of individuals to be sampled and optionally (b) a target distribution $D \in \mathscr{D}(\mathbb{P})$ modeling observation bias.

If no target distribution is provided, $\mathscr{I}$ is constructed by uniformly sampling $N_{\mathscr{I}}$ individuals from $\mathscr{I}'$. Otherwise, individuals are randomly sampled such that the obtained phenotype distribution is expected to be similar to D. For categorical phenotypes, this is achieved by defining sampling probabilities $p'_i := (p_{y_i} \cdot f(y_i)^{-1}) / \sum_{i' \in \mathscr{I}'} (p_{y_{i'}} \cdot f(y_{i'})^{-1})$ for all $i \in \mathscr{I}'$, where $\mathbf{p} := (p_l)_{l \in \mathbb{P}}$ is the target distribution's probability vector and $f(x)$ is the number individuals $i \in \mathscr{I}'$ with phenotypes $y_i = x$. For quantitative phenotypes, $f(x)$ is the number of individuals $i \in \mathscr{I}'$ whose phenotypes y_i fall into the same bin defined by the 1000-quantiles of D as x. Sampling probabilities are then defined as $p'_i := f(y_i)^{-1} / \sum_{i' \in \mathscr{I}'} f(y_{i'})^{-1}$.

3 Scalability

Fig. 3 shows the runtime behavior of EpiGEN for varying numbers of individuals, SNPs, and simulated datasets, when run from one of the pre-computed genotype corpora shipped with EpiGEN. The tests were run on a MacBook Pro 2019 with a 2.8 GHz quad-core Intel i7 processor and 16 GB of main memory running macOS Catalina. EpiGEN can generate epistasis data with 10^5 SNPs and 10^4 individuals in around 12 minutes; and increasing $|\mathscr{I}|$, $|\mathscr{S}|$, and the number of datasets slows down EpiGEN only very moderately (note the logarithmic scale of the x-axes). Generating the genotype corpora shipped with EpiGEN took 49 minutes, on average.

4 Conclusions

EpiGEN is a generic epistasis simulation pipeline that can generate dichotomous, categorical, and quantitative phenotypes, can simulate observation bias, and allows the user to select SNPs based on their MAFs. Epistasis models can be provided via extensional definitions or be specified in terms of parametrized risk models. In future work, we will use EpiGEN for systematically evaluating new and existing epistasis detection tools.

Funding

This work was supported by COST (European Cooperation in Science and Technology) project OpenMultiMed [CA15120]; L. V. has been partially supported by the "Tornosubito" exchange project of the Lazio Region.

References

Altshuler, D. M. *et al.* (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311), 52–58.

Ansarifar, J. and Wang, L. (2019). New algorithms for detecting multi-effect and multi-way epistatic interactions. *Bioinformatics*, **35**(24), 5078–5085.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLOS Comp. Bio.*, **8**(12), e1002822.

Cao, X. *et al.* (2020). DualWMDR: Detecting epistatic interaction with dual screening and multifactor dimensionality reduction. *Hum. Mutat.*, **41**(3), 719–734.

Chatelain, C. *et al.* (2018). Performance of epistasis detection methods in semi-simulated GWAS. *BMC Bioinform.*, **19**(1), 231.

Chattopadhyay, A. and Lu, T.-P. (2019). Gene-gene interaction: the curse of dimensionality. *Ann. Transl. Med.*, **7**(24), 813.

Id-Lahoucine, S. *et al.* (2019). Screening for epistatic selection signatures: A simulation study. *Sci. Rep.*, **9**(1), 1026:1–1026:5.

Jing, P.-J. and Shen, H.-B. (2015). MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics*, **31**(5), 634–641.

Juan, L. *et al.* (2020). PGsim: A comprehensive and highly customizable personal genome simulator. *Front. Bioeng. Biotechnol.*, **8**, 28.

Li, C. and Li, M. (2008). GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, **24**(1), 140–142.

MacArthur, J. *et al.* (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.

Manolio, T. A. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–53.

Niel, C. *et al.* (2015). A survey about methods dedicated to epistasis detection. *Front. Genet.*, **6**, 285.

Peng, B. *et al.* (2019). Genetic simulation resources and the GSR certification program. *Bioinformatics*, **35**(4), 709–710.

Shang, J. *et al.* (2013). EpiSIM: simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis. *Genes Genom.*, **35**(3), 305–316.

Shi, M. *et al.* (2018). Simulating autosomal genotypes with realistic linkage disequilibrium and a spiked-in genetic effect. *BMC Bioinform.*, **19**(1).

Siragusa, E. *et al.* (2019). Linear time algorithms to construct populations fitting multiple constraint distributions at genomic scales. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **16**(4), 1132–1142.

Su, Z. *et al.* (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**(16), 2304–2305.

Urbanowicz, R. J. *et al.* (2012). GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Min.*, **5**(1), 16.

Wang, X. *et al.* (2010). The meaning of interaction. *Human Hered.*, **70**(4), 269–277.

Yang, W. and Gu, C. C. (2013). A whole-genome simulator capable of modeling high-order epistasis for complex disease. *Genet. Epidemiol.*, **37**(7), 686–694.