RARE: Evolutionary Feature Engineering for Rare-variant Bin Discovery

Satvik Dasariraju
sdasariraju23@lawrenceville.org
University of Pennsylvania
Philadelphia, PA, USA

Ryan J. Urbanowicz*
ryanurb@upenn.edu
University of Pennsylvania
Philadelphia, PA, USA

ABSTRACT

Features with rare states, such as rare genetic variants, pose a significant challenge for both statistical and machine learning analyses due to limited detection power and uncertainty surrounding the nature of their role (e.g., additive, heterogeneous, or epistatic) in predicting outcomes such as disease phenotype. Rare variant 'bins' (RVBs) hold the potential to increase association detection power. However, previously proposed binning approaches relied on prior-knowledge assumptions, instead of data-driven techniques, and ignored the potential for multivariate interactions. We present the Relevant Association Rare-variant-bin Evolver (RARE), the first evolutionary algorithm for automatically constructing and evaluating RVBs with either univariate or epistatic associations. We evaluate RARE's ability to correctly bin simulated rare-variant associations over a variety of algorithmic and dataset scenarios. Specifically, we examine (1) ability to detect RVBs of univariate effect (with or without noise), (2) using fixed vs. adaptable bins sizes, (3) employing expert knowledge to initialize bins, and (4) ability to detect RVBs interacting with a separate common variant. We present preliminary results demonstrating the feasibility, efficacy, and limitations of this proposed rare-variant feature engineering algorithm.

CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; • **Applied computing** → *Bioinformatics.*

KEYWORDS

rare variants, evolutionary algorithm, feature engineering, binning, Relief-based algorithm, epistasis

ACM Reference Format:
Satvik Dasariraju and Ryan J. Urbanowicz. 2021. RARE: Evolutionary Feature Engineering for Rare-variant Bin Discovery. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3449726.3463174

1 INTRODUCTION

In scientific fields such as genetics, features with rare variant states present considerable difficulty for association analysis. For example, despite genome-wide association studies (GWAS) elucidating the role of 'common' genetic variants in disease associations [26, 28], uncertainty surrounding the contribution of rare genetic variation in complex diseases persists [13, 38]. Rare genetic variants, typically defined as having a minor allele frequency (MAF) less than 0.05 [16], likely play a significant role in explaining missing heritability in complex diseases [39]. However, traditional methods such as GWAS and other univariate association testing are underpowered for the detection of rare variant associations as a result of the low frequency of rare variants [15]. Univariate testing also fails to detect gene-gene interactions (e.g., epistasis), which are recognized as contributors to complex disease heritability [17, 20].

To date, a variety of methods have been proposed to try and address rare variant association analysis challenges. First, burden testing methods, such as the cohort allelic sum test (CAST) described in [22], combine all rare variants within a single genetic region into a single variable (i.e., bin), that is subsequently evaluated with a univariate association test with a class outcome (e.g., healthy vs. disease). Several other rare variant association analysis methods expand upon burden testing, including the combined multivariate and collapsing method, which prevents inclusion of noncausal features [15], and the nonparametric weighted sum test (WST), which adds weights to each rare variant in the bin [19]. One complication is that individual rare variants can have different directions of effect (i.e., protective vs. increasing disease risk). Because of this, burden testing experiences a significant loss of predictive ability when rare variants in the same genetic region hold differing effects on class value [23]. An ideal binning strategy might adapt bins to separate rare variants with different effect directions.

Non-burden testing methods include the sequence kernel association test (SKAT) and its optimized form (SKAT-O) [14, 35]. These utilize a multiple regression model to determine regression coefficients for each rare variant in a genetic region, with the goal of optimizing association to class value while detecting epistatic effects between individual variants. SKAT and SKAT-O improve upon burden testing by eliminating the assumption that every rare variant in a genetic region is causal and enabling modeling of both risk-increasing and protective rare variants [35]. Yet both burden tests and non-burden tests rely upon the assumption that only rare variants from the same genetic region should be binned, so they are unable to flexibly group rare variants from different genetic regions, potentially missing optimally predictive, combinations of rare variants across genetic regions. Another tool for rare variant bin (RVB) discovery is BioBin, which adopts a biological knowledge-guided

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07…$15.00
https://doi.org/10.1145/3449726.3463174

approach to binning rare variants [21]. In BioBin, candidate bins can be based on a variety of biological annotations including predefined genomic elements such as gene, intergenic region, exon, intron, enhancer, promoter, etc.

All previous methods for RVB discovery rely on some form of prior knowledge, limiting the discovery of novel predictive RVBs not defined by existing knowledge. Also, these methods have not taken bin-interactions into account, i.e., potential interactions among unique RVBs or between a RVB and common variant. Furthermore, these methods have focused specifically on RVB in genetics applications with little applicability to outside fields and problem domains that may also struggle with the uncertainty surrounding use of features with rare states, despite the promise of feature binning as a method for generalized feature engineering [6].

Evolutionary algorithms (EAs) are stochastic computing methods inspired by natural selection that seek to optimize the fitness of candidate solutions by repeatedly alternating between evaluation of candidate solutions' fitness and applying genetic operations on parent candidate solutions to create offspring candidate solutions [4, 29]. Many previous works have implemented EAs for feature selection [1, 36], including the GARS algorithm presented in [2], which was demonstrated to outperform other popular feature selection methods. EAs have also been used to group features in [6–8], where features are clustered based on similarity.

Herein, we introduce the Relevant Association Rare-variant-bin Evolver (RARE), an EA feature engineering approach for the flexible discovery of candidate RVBs, particularly useful in datasets with uncertainty surrounding the role of rare variants. Each RVB constitutes a newly engineered feature that improves the power to detect rare variant associations with a target class outcome, and can be more effectively utilized in downstream machine learning predictive modeling than the original rare variants themselves. To the best of our knowledge, RARE is the first 'adaptive' binning strategy, uniquely designed to identify predictive rare variant combinations with either a univariate or epistatic association with outcome. Since RARE does not require experts to define which variants to include in a candidate bin, this approach is generalizable to non-genetics applications seeking to make better use of features with rare states, such as near-zero variance predictors [12]. However, if expert defined candidate bins are available, RARE has also been designed to initialize its EA search using them. In this paper we seek to answer the following questions about RARE: (1) can the algorithm correctly identify RVBs having a univariate association with outcome (with or without noise), (2) what are the trade-offs between fixed vs. adaptable bin sizes, (3) does expert knowledge bin initialization improve bin discovery, and (4) can RARE correctly identify RVBs that have an epistatic interaction with a separate common variant that predicts outcome. Other contributions of this work include two Rare Variant Data Simulators (RVDSs) that can be used to test rare variant analysis tools like RARE.

In the following sections, we (1) describe the RARE algorithm and the RVDSs, (2) discuss results evaluating RARE across simulated data scenarios, and (3) draw conclusions with future directions.

2 METHODS

In this section, we detail the RARE algorithm as well as the rare variant data simulators developed for this study. Further, we describe the simulation study design for evaluating the performance of RARE. Open source code for RARE and the RVDSs can be found at https://github.com/UrbsLab/RARE.

2.1 The RARE Algorithm

RARE is an EA that constructs bins i.e., candidate groups of features with rare states, seeking to optimize the relevant bin association with outcome. Figure 1 describes the main steps and components of the RARE algorithm including: (1) preprocessing, (2) bin initialization, (3) evolutionary cycles involving bin fitness evaluation and genetic operators, (4) final bin evaluation, and (5) outputs and bin summary. RARE was designed to identify one or more candidate RVBs from the evolving bin population. To simplify analyses in this study we focus on the top performing bin identified, however RARE can also be applied to identify a set of candidate bins as engineered features for further analysis.

2.1.1 Preprocessing. After loading the target dataset, RARE begins by separating rare from common variant features. Rare state features will later be considered for inclusion in bins, while common state features are only utilized when evaluating bins for potential epistatic interactions. To separate rare variant from common variant features, RARE first calculates the MAF of each feature. In GWAS data, features are single nucleotide polymorphisms (SNPs) encoded as (0,1,2) corresponding to (AA,Aa,aa) genotype states. This study focuses on features encoded as SNPs, however many other variable types and encodings are possible. Assuming this SNP encoding, MAFs are calculated as the sum of feature values divided by twice the number of instances, i.e., the frequency of the 'a' allele. Rare variant features and common features are subsequently separated based on a user-specified MAF cutoff, typically 0.05 in GWAS [16]. Any feature with an MAF = 0 (e.g., zero variance predictors) are removed because they offer no predictive ability [12]. For non-genetic or non-SNP data, RARE assumes that MAF indicates each feature's frequency of nonzero values relative to other states. Given this assumption, the same MAF calculation can be applied to other feature types and datasets.

2.1.2 Bin Initialization. RARE offers both random initialization and the capability to import expert knowledge (EK) derived bins as an evolutionary starting point. By default, random initialization generates bins with different sizes to promote diversity of discovery and avoid assumptions about optimal bin size. Initialization with a fixed bin size will be discussed later in section 2.1.7. Random initialization is carried out based on the number of rare features in the dataset (F) and the following user-specified parameters: the maximum number of bins in the population (B), the minimum number of features per bin (M), and the maximum number of bins that can contain each rare feature (C_{max}), specified to promote bin diversity. First, RARE randomly selects a number $C_n \in [1, C_{max}]$, inclusive, for each of F rare features to select the number of bins containing each feature. Next M features are assigned to each bin. Then the remaining $\sum_{n=1}^{F} C_n - M \times B$ rare features are randomly distributed to bins. Lastly, any feature duplicates are deleted from respective

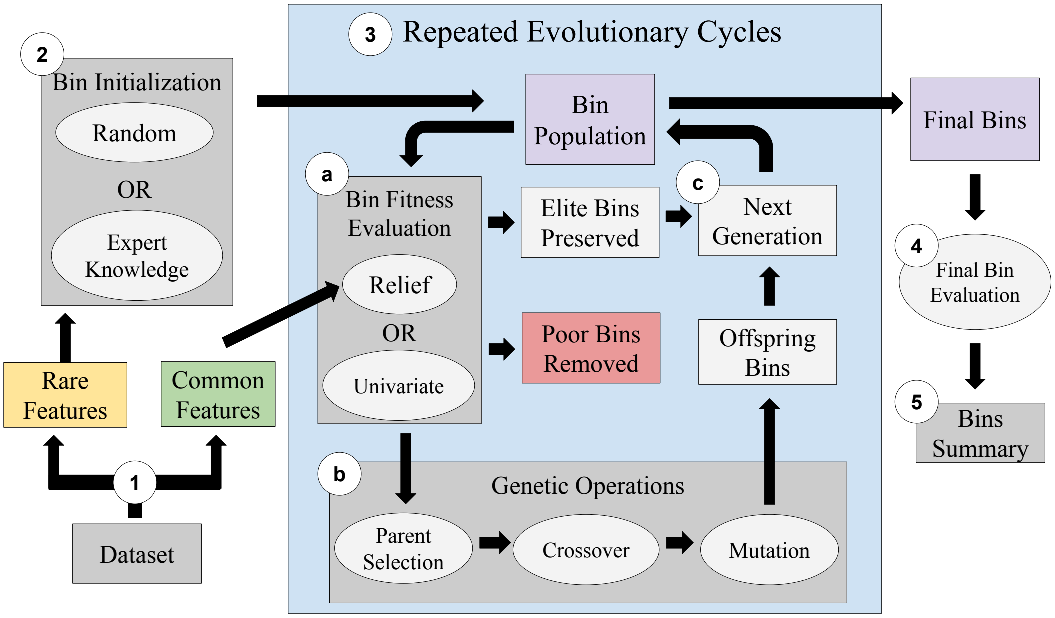


Figure 1: Schematic diagram of RARE algorithm including (1) data preprocessing, (2) bin initialization, (3) evolutionary cycles consisting of bin fitness evaluation, genetic operations, and creation of the next generation, (4) final bin evaluation, and (5) summary of the top bins.

bins and replaced with randomly chosen replacement features. The following formula can be used to calculate the expected value of average bin size:

$$E(\text{average bin size}) = \frac{F \times \frac{1+C_{max}}{2}}{B} \quad (1)$$

As an alternative to random initialization, EK bins, i.e., predetermined rare variant bins that are believed or known to be informative, can be imported to form the initial bin population. Notably, regardless of the initialization method, each bin represents a candidate engineered feature. Thus, for downstream bin evaluation, each bin's rare features' values for an instance are summed to determine the engineered feature's state for that instance. In this way, the population of candidate bins is transformed into a new dataset for the bin fitness evaluation phase.

2.1.3 Evolutionary Cycles. Following bin initialization, RARE commences a user-specified number of evolutionary cycles, (i.e., learning iterations), including bin fitness evaluation, genetic operations, and establishment of the next generation of candidate bins. A larger number of cycles generally improves EA performance, but typically, the number of cycles is selected based on time constraints and available computing resources [33].

2.1.4 Bin Fitness Evaluation. At the start of each evolutionary cycle, the fitness of each candidate bin in the population is evaluated. RARE presents two options for fitness evaluation: univariate scoring and Relief-based scoring.

Univariate Fitness. RARE implements univariate fitness evaluation using the chi-square test, a popular filter-based feature selection method in machine learning [9, 25]. This is in line with how RVBs have been evaluated within existing binning methods. Here, the chi-square test statistic serves directly as bin fitness as a myopic quantifier of bin association with outcome.

Relief-based Fitness. Alternatively, RARE can evaluate bin fitness using the MultiSURF algorithm. MultiSURF was implemented within the ReBATE software, a scikit-learn compatible suite of Relief-based feature importance algorithms that are effective at detecting features involved with both univariate and epistatic associations [31, 32]. MultiSURF was previously demonstrated to be the most reliable and effective Relief-based algorithm to date [32]. During each cycle of RARE, MultiSURF is applied at the rule-population level, i.e., the candidate bin population is converted to a dataset of newly engineered features. MultiSURF thus evaluates a given bin in the context of all other bins as well as (optionally) any common variants available. This gives RARE the capability to evaluate bins for not only univariate associations but inter-bin or bin-common variant interactions associated with outcome. While MultiSURF is regarded as being a computationally efficient strategy to detect feature interactions, it scales quadratically with the number of training instances. This necessitates an option in RARE for bin fitness evaluation to take place with a subset of instances.

2.1.5 Genetic Operations. During each evolutionary cycle of RARE, genetic operations take place after bin fitness evaluation. A generation of candidate bins is replaced by the next generation in two ways: preservation of elite bins and creation of offspring bins.

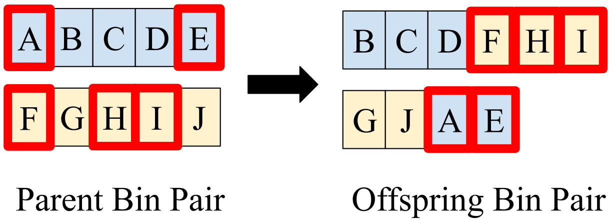


Figure 2: Illustration of RARE's uniform crossover operation. Letters represent rare variant features. During crossover, features in each parent bin are chosen for crossover based on a user-specified crossover probability (these features are outlined in red in the figure). Features chosen for crossover in an offspring bin switch to its paired offspring bin.

Preserving Elite Bins. Elitism is an established addition to the traditional genetic algorithm that preserves a proportion of the highest-scoring solutions of a generation for the next generation [27]. In RARE, the user specifies an elitism parameter, $E \in [0, 1)$ where it is recommended that $E \leq 0.5$, that dictates the proportion of high-scoring bins that are preserved for the next evolutionary cycle. Both elite and non-elite bins are available for parent selection in the next step.

Creating Offspring Bins. Based on the maximum bin population size (B) and the elitism parameter (E), a total of $\lfloor B \times (1 - E) \rfloor$ offspring bins must be created for the next generation. Offspring bins are discovered using the genetic operations of parent selection, mutation, and crossover. A total of $\left\lfloor \frac{B \times (1-E)}{2} \right\rfloor$ pairs of parent bins are chosen using tournament selection, a probabilistic approach where bins with high fitness are more likely to be chosen as parents, and bins with lower fitness still have an opportunity to serve as parents and potentially lead to new fit offspring bins [11].

Each parent bin pair undergoes crossover and mutation to create a corresponding pair of offspring bins. RARE utilizes uniform crossover, as illustrated in Figure 2, where each pair of parent bins is initially copied to create a pair of offspring bins. Then, each feature in each of the two paired offspring bins has a chance to swap between bins with a given user-specified probability. After crossover, a standard mutation operation [3] is applied to each offspring bin, such that each feature in the offspring bin has a user-specified probability of being removed and each feature outside the bin has a proportionate mutation probability of being added to the bin. Since uniform crossover and mutation place no limits on the resulting size of offspring bins, RARE prevents drastic changes in bin size by checking if an offspring bin is over 50% larger than its paired offspring case and redistributing features in such a case.

Offspring bin pairs, created via parent selection, mutation, and crossover, are added to the elite bins to form the next generation. 'Clean up' operations are carried out on this new generation of bins to (1) delete any within bin feature duplicates (replacing any with a randomly selected alternative feature) and (2) delete any candidate bin duplicates (replacing any with a new randomly generated bin). Lastly, the new bins are engineered as features in an evaluation dataset to prepare for bin fitness evaluation at the start of the next evolutionary cycle.

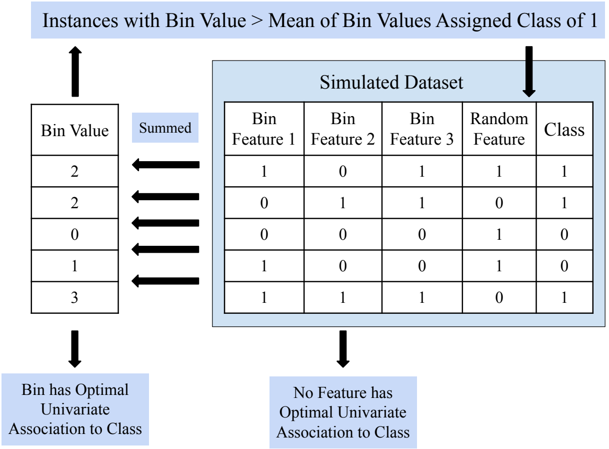
2.1.6 Output and Bins Summary. After the user-specified number of evolutionary cycles are complete, RARE evaluates and outputs the final bins of rare variant features, along with the chi-square or MultiSURF value of each bin. RARE also presents a final engineered feature matrix, where each column represents a bin, to facilitate downstream machine learning. Further, a 'top bins' summary prints a list of features contained in each bin along with the the chi-square or MultiSURF value for each. The top bins summary also reports pertinent information related to each rare variant feature in the bins, such as the feature's MAF and its original chi-square value, which is useful for assessing improved association post-binning.

2.1.7 RARE with Constant Bin Size. In certain problems, the user may know the bin size of the optimal bin solution or the user may want to find the optimal solution for bins of a certain size. Hence, we develop an alternate version of RARE with constant bin size, which differs from the standard version of RARE because (1) all bins are initialized with the same number of features, (2) bin size is held constant throughout the evolutionary cycles, and (3) RARE with constant bin size lacks the ability to discover optimal bin size during evolutionary cycles. To maintain a constant bin size throughout the evolutionary cycles, the uniform crossover operation is modified such that an equal number of features cross over from each offspring in the pair. The mutation operations is modified such that the number of features mutated inside the bin (i.e., features removed) is equal to the number of features mutated outside the bin (i.e., features added).

2.2 Experimental Evaluation of RARE

In order to conduct testing to evaluate RARE, we developed two Rare Variant Data Simulators (RVDSs): functions that create synthetic datasets to simulate different types of relationships in features with rare variant states based on user-specified parameters. RVDSs are applied to simulate datasets with underlying sets of rare-variants that are maximally predictive when combined into an optimal bin. We simulate clean vs. noisy signal, as well as univariate vs. epistatic bin relationships.

2.2.1 RVDS for a Univariate Association Bin. A univariate association RVDS was designed to generate a dataset containing rare variant features such that no single feature can independently predict class value, but a certain combination of features results in a fully penetrant bin, meaning the class label of an instance is completely dependent on the bin value at the instance. The user specifies the number of instances, the total number of rare variant features, the number of rare variant features that belong in the predictive bin, and the minor allele frequency cutoff defining a rare variant feature (e.g., 0.05). Based on these parameters, RVDS for Univariate Association Bin starts by randomly generating values (zero, one, or two) for each of the rare variant features that belong in the predictive bin. The bin value at each instance is calculated by summing the bin's features' values at each instance. Based on the user-specified cutoff metric, either mean or median of bin value across instances, instances with a bin value below the cutoff receive a class value of zero, while other instances receive a class value of


Instances with Bin Value > Mean of Bin Values Assigned Class of 1
Simulated Dataset
Summed
Bin has Optimal Univariate Association to Class
No Feature has Optimal Univariate Association to Class

Bin Value
2
2
0
1
3

Bin Feature 1	Bin Feature 2	Bin Feature 3	Random Feature	Class
1	0	1	1	1
0	1	1	0	1
0	0	0	1	0
1	0	0	1	0
1	1	1	0	1

Figure 3: RVDS for a Univariate Association Bin.

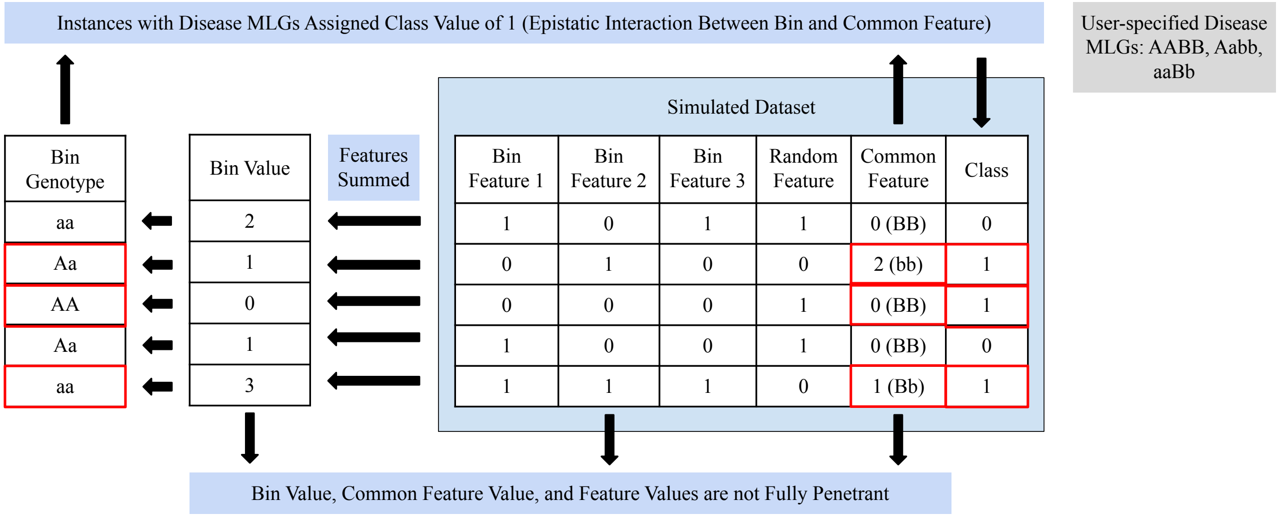
one. The remainder of the features, random features that do not belong in the bin, are randomly generated. This process is illustrated in Figure 3. After all feature values are assigned, a user specified endpoint variation probability, a method of introducing noise, is applied to probabilistically switch the class value of each instance. RVDS for a Univariate Association Bin produces a feature matrix where a certain, user-known combination of rare variant features can be binned for optimal univariate association to class value.

2.2.2 RVDS for an Epistatic Interaction Bin. The second data simulator, RVDS for an Epistatic Interaction Bin, generates a dataset containing rare variant features and one common feature, such that there exists an epistatic interaction between a bin of rare variant features and the common feature. This RVDS is inspired by the GAMETES software, which generates simulated SNP datasets with user-specified forms of epistatic interactions [30]. This RVDS starts by randomly generating values (zero, one, or two) for each of the rare variant features belonging in the epistatic interaction bin. Bin values at each instance are calculated by summing the bins' feature's values at the instance, and bin values of the instances are categorized into three groups to assign one of three bin 'genotypes' (AA, Aa, aa) to each instance. Similarly, a common feature value of zero, one, or two is assigned to each instance; a common feature value of zero corresponds to a BB common feature 'genotype', one corresponds to a Bb common feature genotype, and two corresponds to a bb common feature genotype. The user also selects which of the nine multi-locus genotypes (MLGs) out of AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb, and aabb should correspond to disease status. For each of the instances whose bin value and common feature value matches an MLG of disease status, the class value will be one, while all other instances are assigned a class value of zero. Similar to the univariate RVDS above, random rare variant features, which do not belong in the bin, have their values randomly generated. The resulting dataset is generated in the form of a feature matrix where certain, user-known rare variant features can be additively grouped in a bin that holds an epistatic interaction with the common feature to predict class. The function also outputs the penetrance and frequency of each of the genotypes and MLGs, where penetrance is defined as an instance's probability of having class value of one (i.e., disease status) [30]. A schematic of RVDS for an Epistatic Interaction Bin is illustrated in Figure 4. This RVDS offers the user flexibility, as the user's choice of which MLGs should correspond to disease status determines the type of relationship between the bin of rare variant features and the common feature. Certain combinations of disease status MLGs result in impure, strict epistatic relationships between the bin and common feature. On the other hand, other choices of disease status MLGs results in no bin value genotype and no common feature genotype being fully penetrant, which is an example of pure, strict epistasis.

2.2.3 Data Simulation Scenarios. Table 1 summarizes the various experimental simulation scenarios applied to test and evaluate RARE. Included is the experiment ID, RVDS association type (i.e., univariate or epistatic with a common variant), magnitude of noise, RARE initialization strategy used, fitness scoring used (i.e., Relief or univariate chi-square), total rare variants simulated in each dataset, the predictive rare variants simulated in each dataset, the MAF cutoff separating rare variant features from common variant features, and the number of instances in the simulated dataset. Experiments 1-4 examine RARE's basic ability to identify RVBs from a randomly initialized population given different degrees of noise, specifically 0, 0.05, 0.1, and 0.5. Experiment 4 (noise = 0.5) has eliminated all associations and serves as our negative control. Experiment 5 (in comparison with 1) compares univariate association RVB detection performance using the more computationally efficient chi-square test fitness. Experiment 6 (in comparison with 1) examines the use of a constant bin size (assuming the optimal bin size is known). Here, the optimal bin size of 10 is applied. Experiment 7 (in comparison with 1) examines how utilization of EK bins (when available) can improve efficiency of bin discovery. We simulate this here by initializing the bins such that they randomly include 50-80 % of the 10 predictive rare variants, and at least 4 non-predictive features. In Experiments 1-7, all simulated datasets contain 1000 instances, 50 total rare variant features, and 10 predictive rare variant features, that should be binned together for optimal univariate association with outcome.

Experiments 8 and 9 use the epistatic RVDS to create simulated datasets with 1000 instances, where 5 rare variant features, out of 15 total rare variant features, belong in a bin modeling a pure, strict epistatic interaction with a single common variable predictive of outcome. In the simulated datasets, AABB, AAbb, AaBb, aaBB, and aabb MLGs are denoted as having disease status and the common feature genotype frequency values are 0.25 for BB, 0.5 for Bb, and 0.25 for bb, such that all bin genotypes have a penetrance of 0.5 (i.e., bin values alone have no univariate association with outcome). Univariate fitness in Experiment 8 is compared to MultiSURF fitness in Experiment 9 to demonstrate the importance of the MultiSURF approach for detecting bin interactions.

Thirty replicates of each of the nine experiments are run (270 total trials). Univariate scoring is always carried out on all instances, while Relief-based scoring is done on a sub-sample of 500 out of the 1000 instances to reduce computational expense and evaluate efficacy of simple subsampling. RARE hyperparameter settings: maximum bin population size = 50, evolutionary cycles = 3000,


Instances with Disease MLGs Assigned Class Value of 1 (Epistatic Interaction Between Bin and Common Feature)
User-specified Disease MLGs: AABB, Aabb, aaBb
Features Summed
Simulated Dataset
Bin Value, Common Feature Value, and Feature Values are not Fully Penetrant

Bin Genotype	Bin Value
aa	2
Aa	1
AA	0
Aa	1
aa	3

Bin Feature 1	Bin Feature 2	Bin Feature 3	Random Feature	Common Feature	Class
1	0	1	1	0 (BB)	0
0	1	0	0	2 (bb)	1
0	0	0	1	0 (BB)	1
1	0	0	1	0 (BB)	0
1	1	1	0	1 (Bb)	1

Figure 4: RVDS for an Epistatic Interaction Bin.

Table 1: Experimental Data Simulation Scenarios for Evaluating RARE

ID	RVDS Type	Noise	Initialization	Fitness	Total Rare Variants	Predictive Rare Variants	MAF Cutoff	Instances
1	Univariate	0	Random	MultiSURF	50	10	0.05	1000
2	Univariate	0.05	Random	MultiSURF	50	10	0.05	1000
3	Univariate	0.1	Random	MultiSURF	50	10	0.05	1000
4	Negative Control	0.5	Random	MultiSURF	50	10	0.05	1000
5	Univariate	0	Random	Univariate	50	10	0.05	1000
6	Univariate	0	Random, Constant Size	MultiSURF	50	10	0.05	1000
7	Univariate	0	Partial EK	MultiSURF	50	10	0.05	1000
8	Epistatic	0	Random	Univariate	15	5	0.05	1000
9	Epistatic	0	Random	MultiSURF	15	5	0.05	1000

elitism parameter = 0.4, crossover probability = 0.8, mutation probability = 0.1, are held constant throughout all experiments. Both population size and number of evolutionary cycles are set modestly here to allow for evaluation of the 270 trials, but their increase would be expected to further improve performance in the future with a computational trade-off.

3 RESULTS AND DISCUSSION

Here we summarize the results of applying RARE to the various RVDS datasets. All subsequent results are averages over 30 replicate trials for each experiment. For each scenario we highlight the average percent of 'correct' simulated predictive rare variants appearing in the 'best' bin identified by RARE. The 'best' bin is chosen based on the highest bin chi-square score for all univariate Experiments (1-7), regardless of the fitness metric used. Given that MultiSURF scores are calculated based on a sub-sample of instances, for epistatic Experiments (8-9) the 'best' bin was chosen as the one with the largest percentage of correct rare variant features, with percentage of incorrect features as a tiebreaker. We also report the percentage of incorrect rare variants from the data that were included. Note, these do not sum to 100 since we are examining bin composition with respect to features across the entire dataset. Further, for relevant experiments we report the optimal average chi-square scores obtained on respective RVDS datasets when all 10 user-known, correct rare variants (with no other variants) are included and evaluated as a bin. We contrast this with the average chi square scores of the best bin identified in each experiment with RARE. All p-values were < 0.001 from chi-square test in the 30 replicates for each of Experiments 1-3 and 5-7, while all p-values were > 0.05 from chi-square test in the 30 replicates for Experiment 4 (negative control). The chi-square degrees of freedom depended on the number of unique bin sums across all instances in the dataset. This is why large chi-square values were observed in these analyses.

3.1 Evolving Univariate Association Bins

Table 2 presents average results, along with standard deviation (SD) of each metric across the 30 trials, for univariate bin association Experiments 1-4. Here, RARE applies random initialization and

MultiSURF fitness. With a clean signal, Experiment 1 achieves high correct percentage and low incorrect percentage as well as a near optimal chi-square for the 'best' discovered bin. This is in direct contrast with Experiment 4 (negative control) where correct and incorrect rare variants were almost equally likely to have been included in the best discovered bin, which also yielded an expected low chi-square value. Note that in Experiment 4, the average 'optimal bin' chi-square value is lower than the average chi-square value of RARE's best bin since class values are randomly assigned when noise = 0.5, so it is likely that a combination of randomly generated features have a higher univariate association to outcome than the bin of features that was optimal before noise was applied. Experiments 2 and 3 illustrate RARE's ability to successfully manage a noisy association, despite expected reduced correct percentage and chi-square value, and increased incorrect percentage as noise increased in contrast with Experiment 1.

Table 3 offers a comparison of Experiment 1 to 5 and 6, where different configurations of the RARE algorithm were tested. Specifically, the adoption of the univariate chi-square value for RARE fitness (Experiment 5) yielded optimal RVB discovery across all 30 replicates. The lower performance observed in Experiment 1 is the result of only using half the training instances to reduce computational expense during MultiSURF fitness evaluation, rather than a reflection of MultiSURF's ability to detect univaraiate associations. Next, utilizing a constant bin size (i.e., assumed optimum of 10) in Experiment 6 again improved performance over Experiment 1. This illustrates how precise or approximate knowledge of optimal bin size can be applied to improve RARE performance when available.

Table 4, offers a comparison of Experiment 1 to 7, i.e., random vs EK bin initialization. We observe improved overall performance after all training cycles when adopting available EK bins to initialize the RARE population. Furthermore when we compare the number of evolutionary cycles RARE took to achieve an 80% solution, defined as a bin containing 80% of correct rare variant features and no more than 2 incorrect features, we found EK bin initialization to have dramatically improved RARE efficiency in bin discovery. This is consistent with the previous findings on improving EA efficiency by inputting partial EK [18]. The average run time for 3000 evolutionary cycles with Relief-based scoring in Experiments 1-4 and 6-7 was 24676.83 seconds and the average run time for univariate scoring in Experiment 5 was 338.30 seconds. Clearly, univariate fitness is much more computationally efficient when searching for rare variants bins of univariate effect. However in the next section we highlight the potential benefit of utilizing MultiSURF in RARE fitness for the detection of bin interactions.

3.2 Evolving Epistatic Association Bins

Table 5 presents the percentages of correct and incorrect features binned by RARE in the scenario involving a RVB having a pure interaction with a separate common variable (i.e. Experiments 8 and 9). As a strict, pure epistatic association, the optimal bin chi-square value is negligible (i.e. no univariate association). This was confirmed in experiment 9 which yielded an average top bin chi-square value of 0.68 (average p-value > 0.1). As one would expect, for experiment 8, univariate (chi-square) scoring failed to evolve bins that were informative in combination with the separate common variable since they do not consider multivariate interactions. This was evidenced by the similar inclusion of correct and incorrect rare variants in the best bins. Differently, RARE with MultiSURF (using all bins and the common variant to evaluate fitness), was successful in binning together correct rare variants and excluding incorrect ones. This illustrates the potential for RARE to be applied not only to detect RVBs with a univariate association, but also RVBs that are only informative in the context of other common variants and RVBs.

4 CONCLUSIONS

This study introduces the RARE algorithm, the first EA approach for engineering rare variant feature bins to improve power to detect both univariate or epistatic associations with outcomes. Over a variety of simulation studies, RARE reliably constructs (1) bins with optimal or near-optimal univariate association to outcome and (2) optimal or near-optimal bins involved in an epistatic interaction with a common feature. While existing RVB tools can evaluate user-inputted bins of rare variants, RARE engineers bins from the ground up in a stochastic search for optimal bins. RARE can either discover novel bins from scratch when EK bins are not available or RARE offers the potential to improve upon existing EK bins that may be sub-optimal. This is because, unlike burden and non-burden tests used in genetics, RARE does not assume that solely rare variants belonging to the same genetic region should be grouped. In this work we have also presented two RVDSs that can be applied to generate a variety of rare variant simulation scenarios for testing or comparing other RVB methods in the future.

Conceived as a tool for rare variant association analysis, where there is significant uncertainty surrounding the role of rare variants, RARE can be applied to evolve bins of rare genetic variants to improve prediction of disease phenotypes and elucidate novel associations and interactions between rare variants, potentially contributing to the recognized missing heritability of complex diseases [24, 37]. Beyond genetics, RARE could also be applied to other biological data with rare variant states (e.g., HLA amino acid mismatches for predicting graft failure in organ transplantation) [5, 10, 34]. Similarly RARE could be applied to a variety of data outside biology for binning features with rare feature states, such as near-zero variance predictors [12].

A limitation of this prototype RARE algorithm is the computational expense of MultiSURF-based bin evaluation since Relief-based algorithms scale quadratically with the number of instances. Such algorithms were designed to be run a single time on a given dataset rather than be run repeatedly once per evolutionary cycle, as is done in RARE. Thus, future investigations will explore (1) dual-scoring methods that integrate univariate scoring on the entire dataset and Relief-based scoring on a sample of instances and (2) parallelizable, intelligent instance sampling strategies for applying MultiSURF in RARE. We also plan to greatly expand simulation testing scenarios as well as real world biomedical applications with the aim of discovering predictive bins that reveal novel relationships among rare variants. Further, we expect to expand RARE by evaluating alternative implementation options and ultimately making it available as an open source scikit-learn Python package.

Table 2: Univariate Association Bin Results with Noise

ID	Noise	Correct % (SD)	Incorrect % (SD)	Optimal Chi-square (SD)	RARE Chi-square (SD)
1	0	94.22 (9.85)	0.17 (0.62)	875. (109.62)	855.76 (77.96)
2	0.05	91.56 (10.28)	0.33 (0.85)	690.89 (84.43)	655.17 (52.54)
3	0.1	86.11 (12.44)	0.50 (1.19)	538.25 (58.44)	511.76 (82)
4	0.5	9.00 (8.31)	12.08 (3.77)	1.18 (3.59)	2.53 (1.7)

Table 3: Univariate Association Bin Results with Different RARE Configurations

ID	RARE	Correct % (SD)	Incorrect % (SD)	Optimal Chi-square (SD)	RARE Chi-square (SD)
1	MultiSURF	94.22 (9.85)	0.17 (0.62)	875. (109.62)	855.76 (77.96)
5	Univariate	100.00 (0)	0.00 (0)	889.16 (60.97)	889.16 (60.97)
6	MultiSURF, Constant Size	99.00 (3)	0.25 (0.75)	867.39 (79.14)	866.39 (77.21)

Table 4: Univariate Association Bin Results with Random Vs. Expert Knowledge Initialization

ID	Initialization	Correct % (SD)	Incorrect % (SD)	Optimal Chi-square (SD)	RARE Chi-square (SD)	Cycles (SD)
1	Random	94.22 (9.85)	0.17 (0.62)	875. (109.62)	855.76 (77.96)	286.48 (298.1)
7	Partial EK	99.00 (3.96)	0.08 (0.45)	864.47 (72.03)	862.84 (69.13)	21.57 (68.18)

Table 5: Epistatic Association Bin Results with Univariate vs. MultiSURF Fitness

ID	Fitness Evaluation	Correct % (SD)	Incorrect % (SD)
8	Univariate	33.00 (21.16)	32.33 (14.53)
9	MultiSURF	96.00 (8)	1.00 (3)

ACKNOWLEDGMENTS

This work was supported by the following grant from the National Institute of Health: 1U01AI152960-01. Input from Dr. Loren Gragert of Tulane University School of Medicine and Dr. Malek Kamoun of Penn Medicine was very helpful throughout this work. Special thanks to Robert Zhang of University of Pennsylvania for his guidance and assistance in creating a GitHub repository for RARE. We also extend our gratitude to the Penn Medicine Academic Computing Services for use of their computing resources.

REFERENCES

[1] Nadia Abd-Alsabour. 2014. A review on evolutionary feature selection. In *2014 European Modelling Symposium*. IEEE, 20–26. https://doi.org/10.1109/EMS.2014.28

[2] Mattia Chiesa, Giada Maioli, Gualtiero I Colombo, and Luca Piacentini. 2020. GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. *BMC bioinformatics* 21, 1 (2020), 54. https://doi.org/10.1186/s12859-020-3400-6

[3] Kalyanmoy Deb and Debayan Deb. 2014. Analysing mutation schemes for real-parameter genetic algorithms. *International Journal of Artificial Intelligence and Soft Computing* 4, 1 (2014), 1–28. https://doi.org/10.1504/IJAISC.2014.059280

[4] James A Foster. 2001. Evolutionary computation. *Nature Reviews Genetics* 2, 6 (2001), 428–436. https://doi.org/10.1038/35076523

[5] Loren Gragert, Michael Halagan, and Martin Maiers. 2011. 32-OR: Clustering HLA alleles by sequence feature variant type (SFVT). *Human Immunology* 1, 72 (2011), S177.

[6] Tzung-Pei Hong, Chun-Hao Chen, and Feng-Shih Lin. 2015. Using group genetic algorithm to improve performance of attribute clustering. *Applied Soft Computing* 29 (2015), 371–378. https://doi.org/10.1016/j.asoc.2015.01.001

[7] Tzung-Pei Hong and Yan-Liang Liou. 2007. Attribute clustering in high dimensional feature spaces. In *2007 International Conference on Machine Learning and Cybernetics*, Vol. 4. IEEE, 2286–2289. https://doi.org/10.1109/ICMLC.2007.4370526

[8] Tzung-Pei Hong, Po-Cheng Wang, and Chuan-Kang Ting. 2010. An evolutionary attribute clustering and selection method based on feature similarity. In *IEEE Congress on Evolutionary Computation*. IEEE, 1–5. https://doi.org/10.1109/CEC.2010.5585918

[9] Xin Jin, Anbang Xu, Rongfang Bie, and Ping Guo. 2006. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *International Workshop on Data Mining for Biomedical Applications*. Springer, 106–115. https://doi.org/10.1007/11691730_11

[10] Malek Kamoun, Keith P McCullough, Martin Maiers, Marcelo A Fernandez Vina, Hongzhe Li, Valerie Teal, Alan B Leichtman, and Robert M Merion. 2017. HLA amino acid polymorphisms and kidney allograft survival. *Transplantation* 101, 5 (2017), e170. https://doi.org/10.1097/TP.0000000000001670

[11] Padmavathi Kora and Priyanka Yadlapalli. 2017. Crossover operators in genetic algorithms: A review. *International Journal of Computer Applications* 162, 10 (2017).

[12] Max Kuhn et al. 2008. Building predictive models in R using the caret package. *J Stat Softw* 28, 5 (2008), 1–26. https://doi.org/10.18637/jss.v028.i05

[13] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. 2014. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* 95, 1 (2014), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

[14] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics* 91, 2 (2012), 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007

[15] Bingshan Li and Suzanne M Leal. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* 83, 3 (2008), 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024

[16] Bingshan Li, Dajiang J Liu, and Suzanne M Leal. 2013. Identifying rare variants associated with complex traits via sequencing. *Current protocols in human genetics* 78, 1 (2013), 1–26. https://doi.org/10.1002/0471142905.hg0126s78

[17] Yulun Liu, Jing Huang, Ryan J Urbanowicz, Kun Chen, Elisabetta Manduchi, Casey S Greene, Jason H Moore, Paul Scheet, and Yong Chen. 2020. Embracing study heterogeneity for finding genetic interactions in large-scale research consortia. *Genetic epidemiology* 44, 1 (2020), 52–66. https://doi.org/10.1002/gepi.22262

[18] Qiang Lu, Jun Ren, and Zhiguang Wang. 2016. Using genetic programming with prior formula knowledge to solve symbolic regression problem. *Computational intelligence and neuroscience* 2016 (2016). https://doi.org/10.1155/2016/1021378

[19] Bo Eskerod Madsen and Sharon R Browning. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5, 2 (2009), e1000384. https://doi.org/10.1371/journal.pgen.1000384

[20] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461, 7265 (2009), 747–753. https://doi.org/10.1038/nature08494

[21] Carrie B Moore, John R Wallace, Alex T Frase, Sarah A Pendergrass, and Marylyn D Ritchie. 2013. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC medical genomics* 6, 2 (2013), 1–12. https://doi.org/10.1186/1755-8794-6-S2-S6

[22] Stephan Morgenthaler and William G Thilly. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615, 1-2 (2007), 28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003

[23] Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7, 3 (2011), e1001322. https://doi.org/10.1371/journal.pgen.1001322

[24] Luisa F Pallares. 2019. Genetic Variation: Searching for solutions to the missing heritability problem. *Elife* 8 (2019), e53018. https://doi.org/10.7554/eLife.53018

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Jonathan K Pritchard. 2001. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics* 69, 1 (2001), 124–137. https://doi.org/10.1086/321272

[27] Noraini Mohd Razali, John Geraghty, et al. 2011. Genetic algorithm performance with different selection strategies in solving TSP. In *Proceedings of the world congress on engineering*, Vol. 2. International Association of Engineers Hong Kong, 1–6.

[28] David E Reich and Eric S Lander. 2001. On the allelic spectrum of human disease. *TRENDS in Genetics* 17, 9 (2001), 502–510. https://doi.org/10.1016/S0168-9525(01)02410-6

[29] Andrew N Sloss and Steven Gustafson. 2020. 2019 Evolutionary Algorithms Review. *Genetic Programming Theory and Practice XVII* (2020), 307–344. https://doi.org/10.1007/978-3-030-39958-0_16

[30] Ryan J Urbanowicz, Jeff Kiralis, Nicholas A Sinnott-Armstrong, Tamra Heberling, Jonathan M Fisher, and Jason H Moore. 2012. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData mining* 5, 1 (2012), 1–14. https://doi.org/10.1186/1756-0381-5-16

[31] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. 2018. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics* 85 (2018), 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

[32] Ryan J Urbanowicz, Randal S Olson, Peter Schmitt, Melissa Meeker, and Jason H Moore. 2018. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of biomedical informatics* 85 (2018), 168–188. https://doi.org/10.1016/j.jbi.2018.07.015

[33] Pradnya A Vikhar. 2016. Evolutionary algorithms: A critical review and its future prospects. In *2016 International conference on global trends in signal processing, information computing and communication (ICGTSPICC)*. IEEE, 261–265. https://doi.org/10.1109/ICGTSPICC.2016.7955308

[34] Adam Waring, Andrew Harper, Silvia Salatino, Christopher Kramer, Stefan Neubauer, Kate Thomson, Hugh Watkins, and Martin Farrall. 2020. Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathy. *Journal of Medical Genetics* (2020). https://doi.org/10.1136/jmedgenet-2020-106922

[35] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* 89, 1 (2011), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

[36] Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. 2015. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation* 20, 4 (2015), 606–626. https://doi.org/10.1109/TEVC.2015.2504420

[37] Alexander I Young. 2019. Solving the missing heritability problem. *PLoS genetics* 15, 6 (2019), e1008222. https://doi.org/10.1371/journal.pgen.1008222

[38] Xinyuan Zhang, Anna O Basile, Sarah A Pendergrass, and Marylyn D Ritchie. 2019. Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC bioinformatics* 20, 1 (2019), 1–10. https://doi.org/10.1186/s12859-018-2591-6

[39] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. 2014. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences* 111, 4 (2014), E455–E464. https://doi.org/10.1073/pnas.1322563111