Detecting Clustered Independent Rare Variant Associations Using Genetic Algorithms

Mauricio Guevara Souza, Edgar E. Vallejo, and Karol Estrada

Abstract—The availability of an increasing collection of sequencing data provides the opportunity to study genetic variation with an unprecedented level of detail. There is much interest in uncovering the role of rare variants and their contribution to disease. However, detecting associations of rare variants with small minor allele frequencies (MAF) and modest effects remains a challenge for rare variant association methods. Due to this low signal-to-noise ratio, most methods are underpowered to detect associations even when conducting rare variant association tests at the gene level. We present a new method for detecting rare variant associations. The algorithm consists of two steps. In the first step, a genetic algorithm searches for a promising genomic region containing a collection of genes with causal rare variants. In the second step, a genetic algorithm aims at removing false positives from the located genomic region. We tested the proposed method with a collection of datasets obtained from real exome data. The proposed method possesses sufficient power for detecting associations of rare variants with complex phenotypes. This method can be used for studying the contribution of rare variants with complex disease, particularly in cases where single-variant or gene-based tests are underpowered.

Index Terms—Genetic rare variants, rare variant association studies, SKAT, genetic algorithms, complex disease

1 Introduction

Recent advances in sequencing technologies have revolutionized human genetics research. The availability of an increasing collection of human genomic sequences enables the study of genome variation to an unprecedented level of detail. These studies hold the promise to transform our understanding of genomic variation and its contribution to human disease.

Preliminary studies of human variation on large genome samples have discovered a high abundance of rare genetic variants in the human genome [1]. Recent studies of both exome sequencing and a combination of whole-genome sequencing and imputation have started to identify a collection of rare genetic variants associated with different human complex diseases and traits [2], [3], [4].

Genome-Wide Association Studies (GWAS) have contributed to our greater understanding of the role of common genetic variation in complex disease [5]. In effect, GWAS have been capable of implicating thousands of common variants to hundreds of complex phenotypes [6]. However, the first generation of GWAS arrays were designed to capture common (minor allele frequency (MAF) > 0.05) genetic markers and therefore have limited power for identifying associations of rare genetic variants (MAF < 0.01) with disease [7]. More recent GWAS arrays have been designed to capture rare variation, but they are still unable to identify novel variants.

The single association variant approach that is commonly used in GWAS would require sample sizes of the order of hundreds of thousands in order to possess sufficient power to detect associations of rare variants with modest effects. Therefore, several methods for Rare-Variant Association Studies (RVAS) have been proposed in recent years [7], [8]. The aim of these methods is to increase the power of detecting associations by using either collapsing strategies or methods derived from the C-alpha variance components test [9].

Even though these strategies seem promising in principle, the power of most RVAS methods ($\sim 20\%$) still possesses an ample room for improvement. In recent experiments, no RVAS existing method have shown to perform best in all situations [10].

Alternative strategies such as using an ensemble of these methods have been proposed toward solving this problem [11], [12]. Alternative strategies have relied on unsupervised learning for identifying genes associated with complex diseases [13], [14], [15], [16], [17]. However, detecting associations under low signal-to-noise ratio scenarios remains a challenge for rare-variant association methods.

The current dominant approach for conducting RVAS is the gene-based test in which the variants of a gene are collectively tested in order to increase the power for detecting associations. Recent studies have explored with collapsing genetic variants in pairs of genes [18], pathways [19], [20], and multiple genes [21] to increase the signal-to-noise ratio for detecting rare variant associations with promising results.

As the sample size of GWAS studies have increased, it has been evident that several GWAS loci have shown

- *M. G. Souza is with Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Bioinformatica y Diagnostico Clinico, Monterrey, NL 64710, Mexico. E-mail: mauricio.guevara.souza@gmail.com.*
- *E. E. Vallejo, deceased, was with Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Bioinformatica y Diagnostico Clinico, Monterrey, NL 64710, Mexico. E-mail: vallejo@itesm.mx.*
- *K. Estrada is with the Division of Graduate Professional Studies, Brandeis University, Waltham, MA 02453 USA. E-mail: jestradag@gmail.com.*

Manuscript received 6 Feb. 2019; revised 27 June 2019; accepted 19 July 2019. Date of publication 12 Aug. 2019; date of current version 3 June 2021. (Corresponding author: Mauricio Guevara Souza.) Digital Object Identifier no. 10.1109/TCBB.2019.2930505

statistical evidence of having secondary signals [22],[23],[24],[25]. In the latest GWAS of adult height, it was observed a significant clustering of signals ($P < 1 \times 10^{-4}$), from 423 loci; it was found that 90, 26 and 31 loci contained 2, 3 and ≥ 4 independent signals, respectively [25]. Current GWAS methods cannot differentiate between a locus harboring multiple genes each one with independent GWAS signals, or alternatively, a locus where a single gene has multiple independent signals.

We therefore created a novel method that would allow us to test the hypothesis of multiple rare variants associated with the phenotype occurring in different genes in the same locus. This approach would in principle allow us to increase the power of detecting an association under the hypothesis described above and to accommodate for the study of epistatic effects at the gene level. To our knowledge, the proposed approach consisting of testing rare variant associations using multiple genes in the same locus has not been previously reported in the literature.

We have developed and tested an algorithm for conducting rare-variant association studies that test a collection of genes within a genomic region for detecting association of rare variants with a phenotype. The proposed method consists of a two-step genetic algorithm. The first step scans each chromosome of the genome with the aim of locating a promising fixed-length region using a rare-variant association test. This step of our search procedure produced a collection of genes that are included in the identified promising genomic region. The second step selectively tests groups of genes in the promising region in order to reduce the number of false positives. Genetic algorithms are computational search procedures that have been previously proposed to address a variety of problems in computational biology and bioinformatics [26]

We conducted series of computational experiments for assessing the performance of the proposed method for detecting associations of rare-variants with dichotomous phenotypes. We used an extensive collection of simulated datasets created from real exome data including a small group of genes with causal rare variants. Experimental results indicate that the proposed algorithm is capable of detecting associations between multiple genes each one possessing independent modest effects. This associations would not be detected by using either single-variant or gene-based RVAS.

2 Methods

In order to tests our hypothesis, a series of experiments were conducted on a collection of datasets generated from real exome data. The aim of these experiments was to explore whether the proposed method was capable of detecting associations of groups of genes possessing causal rare-variants with a dichotomous phenotype. Particularly, the main focus of this work was on identifying associations for which conventional single rare variant or gene-based rare variant association methods would fail to detect.

2.1 Algorithm

The search procedure devised for this study was based on genetic algorithms. Genetic algorithms are computational search procedures that use a collection of operators that resemble mechanisms from genetics and natural selection. Genetic algorithms have been applied successfully to approximate a variety of search problems in computational biology and bioinformatics [26].

The proposed algorithm is a two step procedure. The first step aimed to locate a promising region in a chromosome which neighborhood collects a group of consecutive genes that are good candidates to possess causal rare variants. The second step searched this region thoroughly in order to collect true positives and remove false positives from the group of genes. The algorithm conducted the search in each chromosome at a time. Therefore, an iteration on each chromosome was required for a whole-exome scan.

The identification of the promising region was performed by a genetic algorithm. The individuals of this genetic algorithm represented a valid random genomic position within a chromosome. The algorithm then collected all of the genes included in the fixed-size neighborhood (10 Mb) centered at this genomic position. This group of genes correponds to a solution represented by an individual in the population of the genetic algorithm. The fitness of each individual in the population of the genetic algorithm was calculated by conducting an SKAT test on the group of genes associated to the individual. The genetic algorithm then iterated a number of generations until no further fitness improvement of the best individual in the population was observed.

SKAT is a supervised, flexible and computationally efficient regression method to test for association between genetic variants in a region and a continuous or dichotomous trait while easily adjusting for covariates. As a score-based variance-component test, SKAT can quickly calculate p values analytically by fitting the null model containing only the covariates, and so can easily be applied to genome-wide data [27].

We used the Sequence Kernel Association Tests (SKAT) as the rare variant association method in this investigation. SKAT has been established as a gold standard in gene-based RVAS due to its abilities to model both deleterious and protective variants, and to allow for the inclusion of covariates [28]. However, the proposed algorithm could be easily extended to accommodate alternative rare-variant tests such as MiST, SKAT-O, KBAC, etc., if required.

Even though a neighborhood size of 10 Mb was used for sampling the causal genes for the generation of the datasets, we decided to constraint the proposed algorithm to search on a neighborhood of 6 Mb. This decision was made in order to accommodate for the uncertainty of the location of the causal genes and for computational considerations. As a consequence, it would be extremely complicated for the proposed method to locate a promising genomic regions in which the causal genes are located further apart in the chromosome.

Table 1 shows the parameters used in our experiments for the first step of the algorithm. These parameters were determined empirically from preliminary experiments.

The search procedure described above produced a group of all genes included in the fixed-length windows of the identified genomic region. We expected that the set of genes contained some of the causal genes and potentially, a group of false positives. In order to remove these false positives

TABLE 1
Parameters Used by the Search Algorithm of the First Step of the Proposed Method

Parameter	Value
Window Size	6 Mb
Representation	Binary
Chromosome Length	Variable
Crossover Probability	0.6
Mutation Probability	0.2
Generations	50
Population Size	200

TABLE 2
Parameters Used by the Search Algorithm of the Second Step of the Proposed Method

Parameter	Value
Representation	Binary
Chromosome Length	Variable
Crossover Probability	0.6
Mutation Probability	0.2
Generations	100
Population Size	100

we devised the second step of the proposed method. This stage aimed at identifying the subset of the genes that produced the best result on the test.

The identification of the best subset was conducted by a genetic algorithm. The individuals of these genetic algorithm represented a valid subset of genes from the identified promising genomic region. The fitness of each individual in the population of the genetic algorithm was calculated by conducting an SKAT test on the subset of genes associated to the individual.

The genetic algorithm then iterated a number of generations until no further fitness improvement of the best individual in the population was observed for 3 generations. In this case, the algorithm proceeded to the next step.

Table 2 shows the parameters used in our experiments for the second step of the algorithm. These parameters were determined empirically from preliminary experiments.

The second step of the algorithm produced a list of genes, together with the SKAT test calculated for each gene and for the group of genes on the list. This is the final result of the procedure.

For the implementation of our algorithm, we used a genetic algorithm framework available for R [29].

The source code of the algorithm is available at `https://github.com/mguevarasouza/RVASGA` with example data and instructions to run the algorithm.

3 Results

3.1 Datasets

The datasets used in this investigation were collected using the SEQPower package [30]. This software enables the generation of datasets using different models that are useful to evaluate the performance of RVAS methods. Additional features of SEQPower include statistical power analysis and sample size estimation for sequence-based association studies.

For our first group of experiments, a collection of 150 datasets were produced by SEQPower from real exome data consisting of 6500 individuals of European and American ancestral origin. This data was retrieved from the Exome Variant Server [31]. Each dataset was generated independently using a Population Attributable Risk (PAR) model for case-control study designs.

The generation of each dataset consisted of two steps. First, a baseline dataset describing the genetic variability of 5000 individuals was generated by SEQPower using a PAR model with a NULL effect of detrimental rare variants on the phenotype. In principle, this dataset should not include significant associations of rare variants with the phenotype, even when testing for associations at the gene level. Second, a causal genes dataset consisting of a small collection of genes located in the same genomic neighborhood was generated using SEQPower with a different PAR model. In this case, the effect of detrimental rare variants was set to a range of values to confer them with different effect sizes on the phenotype. The latter dataset was then inserted in the former in order to introduce the causal genes into the appropriate chromosome. Depending on the PAR risk value, the obtained dataset should include significant associations of rare variants with the phenotype.

More specifically, 30 different datasets were prepared for each the 5 values of the Population Attributable Risk for detrimental rare variants parameter used in this study {0.01, 0.02, 0.03, 0.04, 0.05}, for a total of 150 datasets for the experiments. Each dataset consisted of 5 genes randomly sampled from a given position of a chromosome using a genomic region of 10 Mb. These datasets were used to explore the capabilities of the proposed method to identify the collection of genes possessing causal rare variants at different risk values.

In order to speed up the execution of the proposed method, we conducted a whole- exome gene-based SKAT association test in order to exclude the genes possessing a p-value > 0.1 from the baseline dataset.

In order to assess the performance of the proposed method we first conducted a gene-based rare variant association study using SKAT on individual genes (~2,000) of the baseline dataset described above. Fig. 1 shows a QQ-plot of the resulted p-values for this study.

In addition we designed two groups of experiments. In the first group of experiments, a collection of 5 genes of the same genomic neighborhood was randomly selected from one designated chromosome of the baseline dataset and the causal effect was conferred to them using SEQPower with varying effect sizes. These causal genes were then introduced into the baseline dataset. The goal of the proposed algorithm was to identify as many of these causal genes as possible. Fig. 2 shows a QQ-plot of the resulted p-values for a dataset containing causal genes generated with a PAR of 0.03.

In the second group of experiments, we used the same approach as before but this time a group of causal genes were generated and introduced into two different chromosomes. This is a more complex scenario but we decided to also test the algorithm in such conditions because we expect that multi-factorial diseases will have loci distributed across the genome [32].

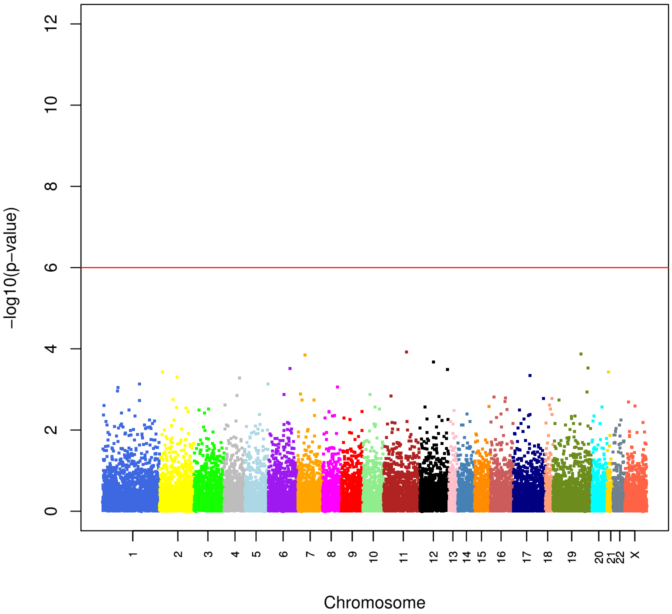

Fig. 1. Manhattan plot for the gene-based RVAS. The dataset was obtained with a population attributable risk of detrimental rare-variants of 0.0. No gene in the baseline dataset produced a statistically significant p-value from the SKAT test.

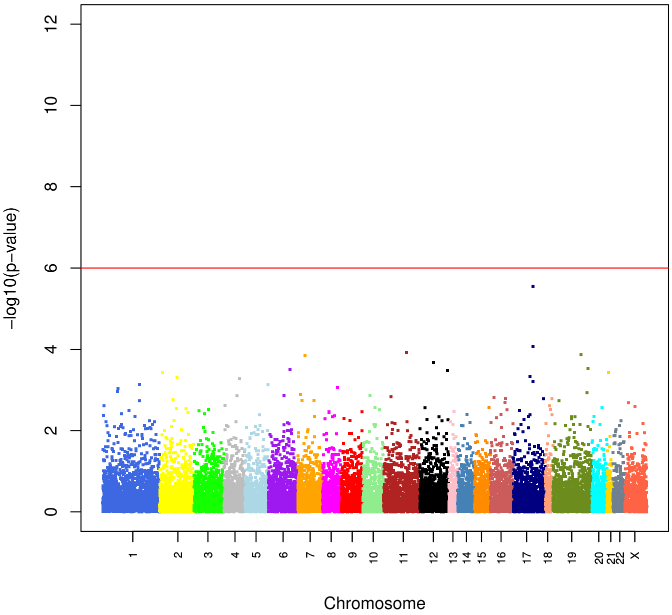

Fig. 2. Manhattan plot for the gene-based RVAS. In this experiment, a collection of five causal genes was inserted on chromosome 17. The effect of the causal genes were obtained for a population attributable risk of 0.03. A few genes show modest associations with the phenotype ($P < 0.0001$) but did not reach statistical significance.

For both scenarios we explored the ability of our algorithm to identify associations using different values for the Population Attributable Risk for detrimental rare variants parameter. Specifically, we used values of 0.05, 0.04, 0.03, 0.02 and 0.01 to assess the performance of the proposed method on different signal-to-noise ratio conditions.

3.2 Experiments

In this scenario, as mentioned above, all the causal genes were placed in the same chromosome. All the results presented in this section are the average of 30 independent runs.

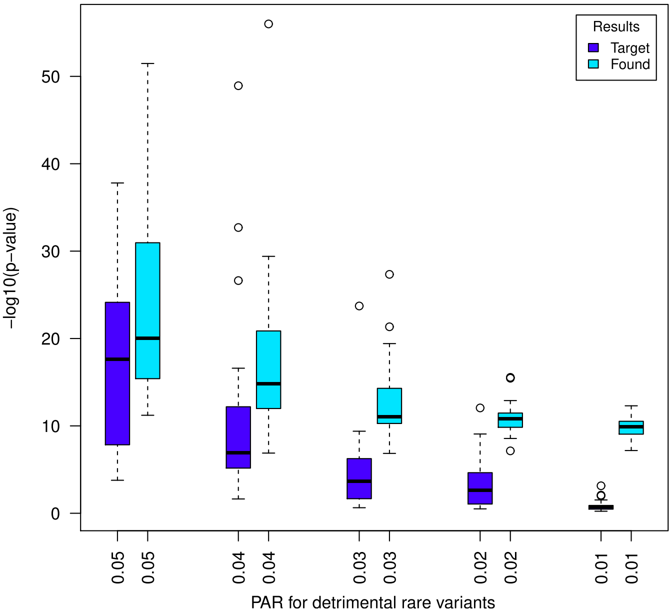

Fig. 3. Boxplot of the results from 30 runs of the proposed method (Found) for each disease model compared to the target p-value that obtained by conducting a multi-gene SKAT test on the five causal genes (Target).The box represents the interquartile range (IQR): 25th to the 75th percentile, whiskers represent the maximum: Q3 + 1.5*IQR and the minimum : Q1 -1.5*IQR values.

3.2.1 SKAT p-Values

After the genetic algorithm completed the search, we calculated the joint p-value using SKAT tests of the list of genes contained in the final solution. It is important to point out that due to the differences in the datasets, the p-values cannot be compared directly between different runs but the average provides a tendency of the behavior of the SKAT tests as the association signal of the causal genes decreases.

The result of these experiments can be seen in Fig. 3. We compared the p-value of the solution yielded by our algorithm with the p-value obtained by testing SKAT on the causal genes exclusively. An improvement on the p-value of the solution is consistently achieved by the proposed algorithm regardless of the risk value of the PAR model.

More specifically, Fig. 4 shows the improvement on the p-value for the 30 experiments of the disease model with PAR = 0.05. In most of the experiments, the proposed method yielded an improvement on the p-value with respect to the joint p-value of the casual genes. The triangles represented in the graphs are the SKAT p-values for the best solution for each of the 30 runs. The triangles pointing upwards represent the solutions that provided a better p-value of the SKAT test than the expected p-value for the joint test for the causal genes exclusively.

3.2.2 Precision and Recall

We also evaluated the trade-off between the precision and recall of the proposed method. As expected, the precision of the method is a decrease function of the risk. In effect, our method is capable of detecting most casual genes when PAR = 0.05, even though the length of the region used by

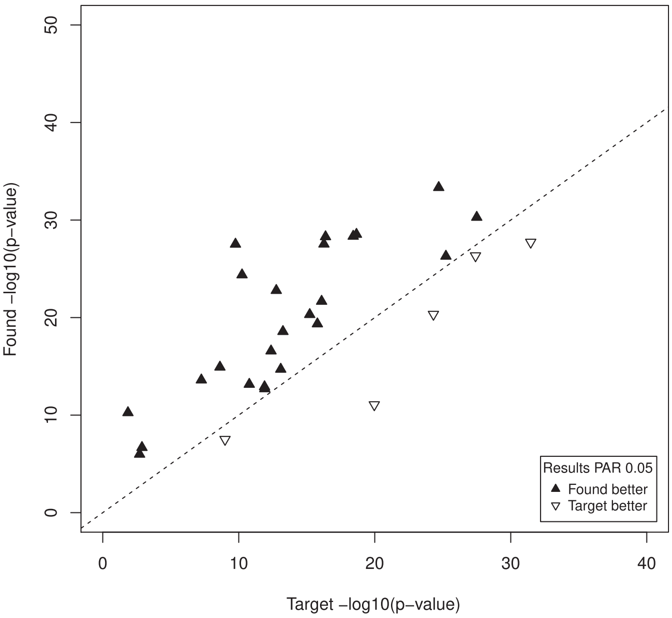

Fig. 4. Results for the 30 experiments with PAR = 0.05. Experiments where the proposed method produced a better p-value than the target p-value are indicated by black triangles and white triangles otherwise.

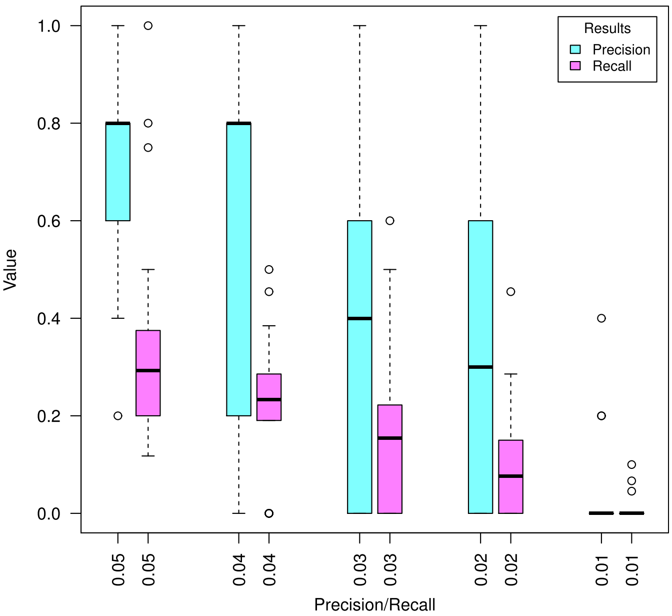

Fig. 5. Boxplot of precision and recall for the collection of 30 experiments for each risk value considered in this study. Results are shown by decreasing values of the population attributable risk. The box represents the interquartile range (IQR): 25th to the 75th percentile, whiskers represent the maximum: Q3 + 1.5*IQR and the minimum : Q1 -1.5*IQR values.

the method is smaller than the region used for sampling the casual genes. However, as the risk is decreased, the proposed method becomes decreasingly capable of detecting the casual genes. In addition, the proposed method produced a list of genes that includes a number of false positives. In our experiments, recall was 30 percent in the best case. These results are shown in Fig. 5.

We were also interested in confirming if the improvement of the p-value yielded by the proposed method was correlated with the number of genes of the result list. This is to be expected as SKAT aggregates the signal of the list of genes in order to increased the power for detecting causal genes. Fig. 6 shows that the genes count of the solution produced by our method was not positively correlated with to the obtained p-value. This is more evident in disease models where the risk is high to moderate.

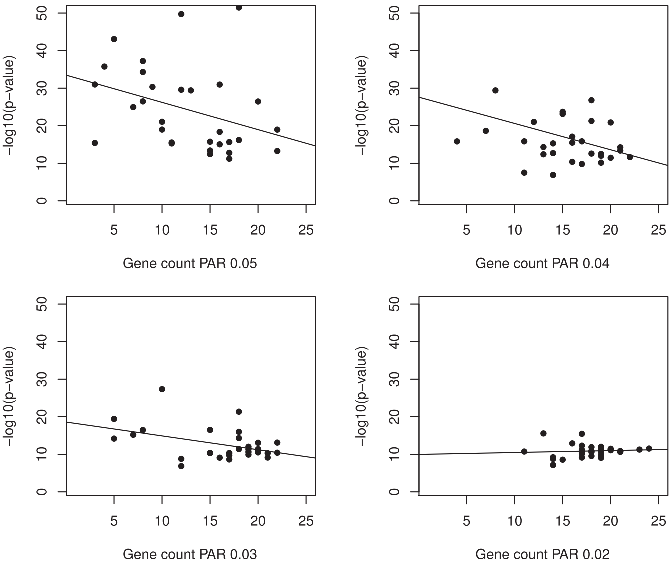

Fig. 6. Number of genes in the solution produced by the proposed method and the related p-value. Each figure shows the 30 experiments for PAR = 0.05 to 0.02 in decreasing order. The case of PAR = 0.01 is not informative.

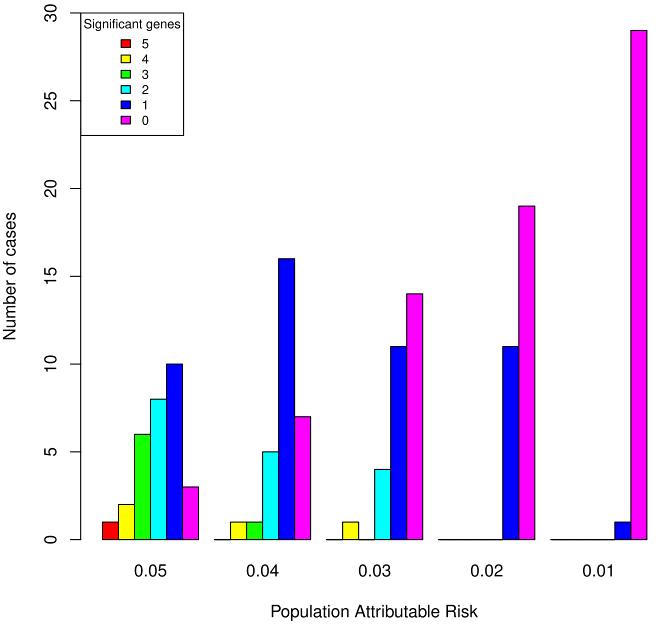

Fig. 7. Number of SKAT significant genes ($P < 2.5 \times 10^{-6}$) in the group of 5 causal genes. Our interest was focused in the magenta bar, which correspond to experiments in which none of the 5 causal genes is significant with respect to individual SKAT test.

3.3 Detecting Causal Genes with Low Effect Sizes

We were particularly interested in assessing the power of the proposed algorithm when confronted to experiments including exclusively causal genes with low effect sizes; that is, those experiments in which none of the 5 causal

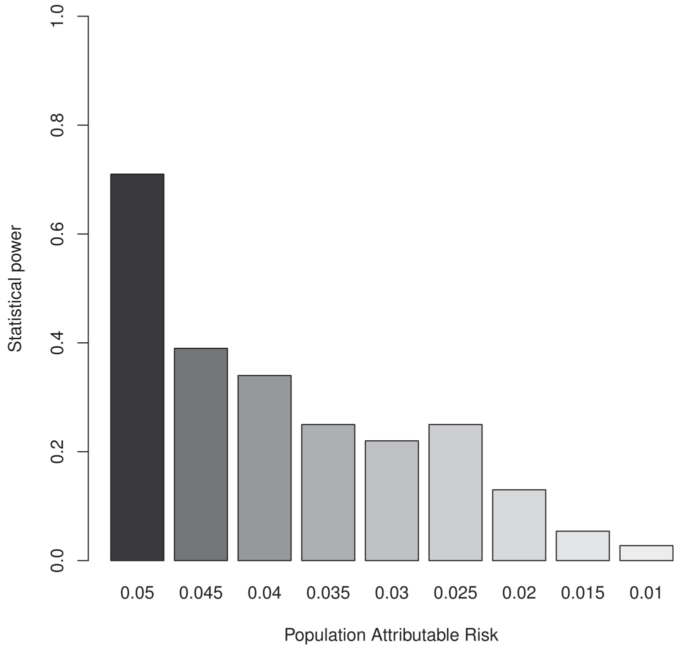

Fig. 8. Statistical power of the proposed method for detecting genes with low effect sizes. Here, we show the power of our algorithm on the most difficult scenarios where no single gene passed the SKAT $P < 2.5 \times 10^{-6}$ significance threshold.

genes would be detected using a gene-based whole-exome scan. Therefore, we identified the experiments in which all of the individual SKAT tests of the causal genes did not produce a significant p-value after the Bonferroni correction (2.5×10^{-6}). Fig. 7 shows the different possibilities of the number of significant causal genes with respect to the SKAT test. It can be seen that the proportion of significant causal genes in the experiment decreases as the risk value is decreased. We hypothesized that our method would still be capable of detecting the causal genes in these situations.

In order to assess the statistical power with better resolution, we conducted additional experiments for risk values of 0.05, 0.045, 0.04, 0.035, 0.03, 0.025, 0.02, 0.015 and 0.01. Specifically, we ran the proposed algorithm only on the experiments in which none of the causal genes is significant. After the execution of the algorithm, we calculated the statistical power obtained in the experiments of the different risk values as shown in Fig. 8.

We compared our genetic algorithm with a stochastic hill climbing algorithm for the 30 experiments with a risk value of 0.05. Overall, the genetic algorithm considerably outperformed the simpler search algorithm at locating a promising region in the genome that potentially includes the causal genes.

4 Discussion

Overall, the proposed method was capable of identifying the genes possessing rare variants that contribute to the disease phenotype. This is specially evident when the PAR of detrimental rare variants is close to 0.05. However, the performance of the algorithm degrades as this risk value is reduced. This is expected as the algorithm is confounded by non-causal genes that are stochastically associated with the trait.

In spite of the limitation that the size of the region used for searching for the causal genes (6 Mb) was considerably smaller than the region used to sample those genes from the baseline dataset (10 Mb), the proposed method was capable of detecting a fair proportion of these genes. However, the procedure also produces a collection of false positives. These false positives are genes surrounding the causal genes that hold a moderate association with the phenotype by chance. Completely removing these genes from the solution proved to be challenging because the inclusion of these genes typically improves the quality of the solution due to their additive effects on the SKAT test.

Comparing the results of the proposed method with the conventional whole-exome scan that searches for individual genes associated with the phenotype produced mixed results. On the one hand, conventional RVAS are incapable of detecting most of the causal genes, using a threshold p-value of 2.5×10^-6 (after Bonferroni correction of $\sim$ 20,000 tests). The reason is that due to the procedure used to generate the datasets, the effect sizes of the causal genes is often marginal. On the other hand, conventional RVAS does not produce false positives during the scan. Therefore, there is a trade-off when using the proposed method against gene-based association tests: precision is typically improved but recall is worsened due to the presence of false positives.

We therefore focused in conducting experiments in which the whole-exome gene-based scan would not be able to identify any of the causal genes. This is the case in which the SKAT tests applied individually to each of the causal genes did not produce a significant p-value. The proposed method was capable of detecting the genomic region that includes the causal genes with varying degrees of statistical power, depending on the risk value, reaching 71 percent of power with PAR = 0.05 As a consequence, we believe that our method can be used as a hypothesis generating step that would require further follow-up and replication for detecting associations of rare variants in a genomic region and a phenotype of interest. Association studies testing individual variants for associations in the promising genomic region can follow afterwards.

The comparisons performed using the p-values demonstrate that our method often obtain a combination of genes that produce a better p-value. This is emphasized when the effect sizes of causal genes is low. In general, the ability of the proposed method to detect causal genes depends on the risk conferred by the genetic rare variants. That is, the lower the risk, the harder for the proposed method to detect the causal genes. Therefore, the algorithm is confounded with the presence of non causal genes associated with the phenotype by chance. That is, the algorithm is guided by the most promising genomic region which typically does not include the causal genes.

The datasets used for the experiments presented here were generated using the Population Attributable Risk Model (PAR) for simulating the genotypes. This model confers an effect on the variants that is inversely proportional to its minimum allele frequency (MAF). That is, the rarer the variant, the larger the effect on the phenotype. An implication of this model is that some of the selected genes to be used as causal genes do not show a strong association with the phenotype. In this context, a whole-exome gene-based scan would produce type II errors. In contrast with the gene-based scan, our algorithm is often capable of detecting those genes but at the expense of producing type I errors.

It is important to point out that the power of the proposed algorithm relies crucially on the election of the underlying algorithm. We used SKAT during the experiments based on the results reported in the literature and on computational considerations [10], [28]. However, our method is sufficiently flexible to allow the inclusion of alternative rare variant association methods. Further, considering an ensemble of such methods would also be possible.

Scanning the whole-exome for associations between rare variants and phenotypes using large regions exhaustively is prohibitive from the computational perspective. Our method search for promising regions using an heuristic method, then tries to remove false positives on the candidate solutions, providing both efficacy and efficiency in the procedure.

We used dichotomous traits in our experiments. An immediate extension of this work will be to explore on the performance of the proposed method when quantitative traits are considered. For those cases, appropriate genetic models should be used for data generation, such as the Linear Models for Quantitative traits.

Once the proposed method have demonstrated to be capable of producing consistent results on a representative set of problems, we expect to apply the algorithm to the identification of novel loci associated with disease phenotypes. Particularly, we expect to use the algorithm for conducting rare variant association studies for complex diseases and to report the results in future publications.

Additionally, we plan to explore with the use of alternative gene-based association methods such as MiST, SKAT-O, or KBAC. In recent experiments, these methods performed better than SKAT at detecting rare variants associated with both dichotomous and quantitative traits in a variety of scenarios [10].

We expect that these experiments would contribute to the better understanding of the capabilities and limitations of the different gene-base rare variant association methods. This knowledge would provide the basis for identifying the problems in which a specific method perform best.

Genetic algorithms are effective optimization procedures. However, search methods such as Differential Evolution, Particle Swarm Optimization, among others, have also showed to produce competitive results with respect to genetic algorithms in similar problems. Therefore, it would be worthwhile to explore which search method performs better, not only in terms of the quality of the solutions, but also with respect to the required computational costs.

5 Conclusion

This work demonstrated the use of a novel method for identifying associations of rare variants with disease phenotypes. The proposed method is capable of detecting promising genomic regions containing a collection of causal rare variants distributed among different genes within a genomic neighborhood using a multiple gene rare variant association test. In addition, we showed the use of this method for detecting multiple loci in different chromosomes. Overall, we believe that the proposed method hold much promise for contributing to the discovery of novel association of rare variants with complex phenotypes.

Acknowledgments

The authors thank our colleagues from the Bioinformatics for Clinical Diagnosis Program, Escuela de Medicina y Ciencias de la Salud, Tecnologico de Monterrey, for their valuable comments on this work. They also thank Charles E. Taylor for his valuable feedback for improving this manuscript. Finally, they thank Nancy Estanislao, Luis Santiago Luevano, and Jorge A. Zuniga for conducting additional experiments not reported here that were useful for validation of the proposed model. The co-authors of this paper mourn the sudden passing of coauthor Edgar Vallejo, a good friend, respected colleague, mentor, and excellent teacher with outstanding contributions to the research of bioinformatics in Mexico. This article is dedicated to his memory.

References

[1] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser, "An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people," *Sci.*, vol. 337, no. 6090, pp. 100–104, Jul. 2012.

[2] "Coding variation in ANGPTL4, LPL, and SVEP1 and the risk of coronary disease," *New England J. Medicine*, vol. 374, no. 12, pp. 1134–1144, Mar. 2016.

[3] L. M. Polfus, R. A. Gibbs, and E. Boerwinkle, "Coronary heart disease and genetic variants with low phospholipase A2 activity," *New England J. Medicine*, vol. 372, no. 3, pp. 295–296, Jan. 2015.

[4] H.-F. Zheng, V. Forgetta, Y.-H. Hsu, K. Estrada, et al., "Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture," *Nature*, vol. 526, no. 7571, pp. 112–117, Oct. 2015.

[5] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinf.*, vol. 26, no. 4, pp. 445–455, Feb. 2010.

[6] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, pp. 1001–1006, Jan. 2014.

[7] J. Asimit and E. Zeggini, "Rare variant association analysis methods for complex traits," *Annu. Rev. Genetics*, vol. 44, no. 1, pp. 293–308, Dec. 2010.

[8] V. Bansal, O. Libiger, A. Torkamani, and N. J. Schork, "Statistical analysis strategies for association studies involving rare variants," *Nature Rev. Genetics*, vol. 11, no. 11, pp. 773–785, Nov. 2010.

[9] B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly, "Testing for an unusual distribution of rare variants," *PLoS Genetics*, vol. 7, no. 3, Mar. 2011, Art. no. e1 001 322+.

[10] L. Moutsianas, V. Agarwala, C. Fuchsberger, J. Flannick, M. A. Rivas, K. J. Gaulton, P. K. Albers, GoT2D Consortium, G. McVean, M. Boehnke, D. Altshuler, and M. I. McCarthy, "The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease," *PLoS Genetics*, vol. 11, no. 4, Apr. 2015, Art. no. e1005165.

[11] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C.-L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, and W. Sieh, "Revel: An ensemble method for predicting the pathogenicity of rare missense variants," *Amer. J. Human Genetics*, vol. 99, no. 4, pp. 877–885, Oct. 2016.

[12] F. Serafino, G. Pio, and M. Ceci, "Ensemble learning for multi-type classification in heterogeneous networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2326–2339, Dec. 2018, doi: 10.1109/TKDE.2018.2822307

[13] P. Barbiero, A. Bertotti, G. Ciravegna, G. Cirrincione, E. Pasero, and E. Piccolo, *Supervised Gene Identification in Colorectal Cancer*. Cham, Switzerland: Springer, 2019, pp. 243–251. [Online]. Available: https://doi.org/10.1007/978-3-319-95095-2_23

[14] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomput.*, vol. 256, no. C, pp. 56–62, Sep. 2017. [Online]. Available: https://doi.org/10.1016/j.neucom.2016.07.080

[15] Y. Liu, H. Lu, K. Yan, H. Xia, and C. An, "Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification," *Comput. Intell. Neuroscience*, vol. 2016, 2016, Art. no. 19.

[16] H. Lu, L. Yang, K. Yan, Y. Xue, and Z. Gao, "A cost-sensitive rotation forest algorithm for gene expression data classification," *Neurocomput.*, vol. 228, no. C, pp. 270–276, Mar. 2017. [Online]. Available: https://doi.org/10.1016/j.neucom.2016.09.077

[17] T. Mori, H. Ngouv, M. Hayashida, T. Akutsu, and J. C. Nacher, "ncrna-disease association prediction based on sequence information and tripartite network," *BMC Syst. Biol.*, vol. 12, no. 1, Apr. 2018, Art. no. 37. [Online]. Available: https://doi.org/10.1186/s12918-018-0527-4

[18] S. Papadimitriou, A. Gazzo, N. Versbraegen, C. Nachtegael, J. Aerts, Y. Moreau, S. Van Dooren, A. Nowé, G. Smits, and T. Lenaerts, "Predicting disease-causing variant combinations," *Proc. Nat. Academy Sci. United States America*, vol. 116, no. 24, pp. 11 878–11 887, 2019. [Online]. Available: https://www.pnas.org/content/116/24/11878

[19] S. Lee, S. Choi, Y. J. Kim, B.-J. Kim, T.-G. Consortium, H. Hwang, and T. Park, "Pathway-based approach using hierarchical components of collapsed rare variants," *Bioinf.*, vol. 32, no. 17, pp. i586–i594, 2016. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw425

[20] X. Zhan, N. Zhao, A. Plantinga, T. A. Thornton, K. N. Conneely, M. P. Epstein, and M. C. Wu, "Powerful genetic association analysis for common or rare variants with high-dimensional structured traits," *Genetics*, vol. 206, no. 4, pp. 1779–1790, Aug. 2017, 28642271[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28642271

[21] Z. Wang, Q. Sha, S. Fang, K. Zhang, and S. Zhang, "Testing an optimally weighted combination of common and/or rare variants with multiple traits," *PloS One*, vol. 13, no. 7, pp. e0 201 186–e0 201 186, Jul. 2018, 30048520[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30048520

[22] K. Estrada, U. Styrkarsdottir, E. Evangelou, et al., "Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture," *Nature Genetics*, vol. 44, no. 5, pp. 491–501, May 2012.

[23] A. E. Locke, B. Kahali, S. I. Berndt, et al., "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, pp. 197–206, Feb. 2015.

[24] D. Shungin and T. W. Winkler, "New genetic loci link adipose and insulin biology to body fat distribution," *Nature*, vol. 518, no. 7538, pp. 187–196, Feb. 2015.

[25] A. R. Wood, T. Esko, J. Yang, et al., "Defining the role of common variation in the genomic and biological architecture of adult human height," *Nature Genetics*, vol. 46, no. 11, pp. 1173–1186, Nov. 2014.

[26] J. K. Estrada-Gil, J. C. Fernandez-Lopez, E. Hernandez-Lemus, I. Silva-Zolezzi, A. Hidalgo-Miranda, G. Jimenez-Sanchez, and E. E. Vallejo-Clemente, "GPDTI: A genetic programming decision tree induction method to find epistatic effects in common complex diseases," *Bioinf.*, vol. 23, no. 13, pp. i167–174, Jul. 2007.

[27] M. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *Amer. J. Human Genetics*, vol. 89, pp. 82–93, 2011.

[28] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *Amer. J. Human Genetics*, vol. 89, no. 1, pp. 82–93, Jul. 2011.

[29] L. Scrucca, "Ga: A package for genetic algorithms in R," *J. Statistical Softw. Articles*, vol. 53, no. 4, pp. 1–37, 2013.

[30] G. T. Wang, B. Li, R. P. Lyn Santos-Cortez, B. Peng, and S. M. Leal, "Power analysis and sample size estimation for sequence-based association studies," *Bioinf.*, vol. 30, no. 16, pp. 2377–2378, 2014.

[31] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Sci.*, vol. 337, no. 6090, pp. 64–69, Jul. 2012.

[32] S. B. Gabriel, "The structure of haplotype blocks in the human genome," *Sci.*, vol. 296, no. 5576, pp. 2225–2229, May 2002.

Mauricio Guevara Souza received the PhD degree in computer science from the Tecnologico de Monterrey. His current research interests include the application of artificial intelligence to problems in biology and medicine. Also, he has developed computer models to study alternatives for disease vector population replacement.

Edgar E. Vallejo was an associate professor at the Tecnologico de Monterrey, where he held positions with the Computer Science Department and the Bioinformatics Department. His past research interests included deep learning and bioinformatics. His research was supported by Microsoft (2010) and Google (2016-2018), among others.

Karol Estrada is an adjunct professor of statistical genetics at Brandeis University. He is also the head of statistical genetics at Biomarin, San Rafael, CA. His research interests focus on the translation of genetic discoveries into novel drug targets. Previously, he was appointed at Biogen, the Broad Institute of Harvard-MIT, and the Massachusetts General Hospital.