A Stochastic Search Algorithm for Finding Multi-SNP Effects Using Nuclear Families

Clarice R. Weinberg, Min Shi, Alison Wise, David M. Umbach, Juno Krahn, Yuanyuan Li, Leping Li

Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, DHHS,

Research Triangle Park, NC 27709

Abstract

Biologic systems typically involve failure-resistant redundancy, and phenotypes such as birth defects may occur through joint effects of exposures and multiple genetic variants. Such joint effects may produce a weak signal in a GWAS analysis that only considers single-SNP associations. We describe an approach that uses case-parent triads and applies an "evolutionary" algorithm to stochastically search the large sample space that considers all sets of a given size of potentially-interacting SNPs. We assess the performance of our algorithm using simulated but realistic GWAS data from dbGaP based on families with the birth defect oral cleft. We spike in specific multi-SNP causal effects and then try to recover those causative complexes *de novo*. Initial simulations using our method show promising results in scenarios that involve two sets of four interacting SNPs, each of which has a modest attributable fraction against a sporadic case background.

Introduction

Identification of causative SNPs is a challenge when SNPs have small marginal effects and p values must be adjusted for thousands of tests. Individual SNPs with small marginal effects may have large effects jointly. The use of a stochastic search algorithm such as the Evolutionary Algorithm (EA) allows us to fully explore the search space; here it will be 10,000 SNPs choose 4, which is about 4 x 10^{14}.

Creating a simulated population

We first needed to develop a way to simulate families with realistic linkage disequilibrium structure. We started with real family data, of which we have ~ 2000 trios. We created each simulated trio as follows:

1. For each case trio create a hypothetical "complement" trio, with the same parents and an offspring who carries the non-transmitted parental alleles at each locus.
2. Break each of the 22 autosomes into fourths to form 88 case trio segments and 88 complement trio segments for each trio.
3. For each of the 88 trio segments, randomly select that segment from either a case or complement trio from a randomly selected family. Glue the 88 segments back together to form a 22-chromosome pseudo-trio.

Random selection of either case-trio and complement-trio obliterates existing signals related to clefting.

An example of a simulated chromosome for a simulated trio is show below:

Case:	case 10	complement 51	case 123	case 47
Mother:	mother 10	mother 51	mother 123	mother 47
Father:	father 10	father 51	father 123	father 47

Hiding the needle(s) in the haystack

Next, we allowed each simulated family to be at risk of disease depending on the offspring's SNPs. We assumed there were two causative 4-SNP sets, with the following probabilities:

P(child has disease| child has all 4 SNPs in pathway 1) = 0.4
P(child has disease| child has all 4 SNPs in pathway 2) = 0.4
P(child has disease| child has neither complete set) = 0.00166

We discarded families when the offspring does not have the disease. We repeated this process to generate 1000 trios in which the offspring has the disease. Only 33 cases have pathway 1 and 36 have pathway 2.

An Evolutionary Algorithm

The EA creates pseudo-*populations. Each undergoes random mutation and reproduction* based on a defined fitness score. "Individuals" in each EA population are sets of SNPs (here size 4). To calculate fitness we compute a difference vector by taking the difference at each locus between the number of copies carried by the case and the number not transmitted, i.e. carried by the *complement*. We also calculate the negative of each difference vector. We measure the evidence for a joint effect of the SNPs in a given set, using a nearest neighbor classification that seeks to differentiate case-minus-complement vectors from complement-minus-case vectors. If a SNP set is unrelated to risk, the two kinds of difference vectors will be intermingled. In each of 200 runs, 600 *individual SNP* sets evolve. At the end of each population's 500-generation evolution we record the set with the best fitness and record its fitness score.

K nearest neighbor (KNN) classification: using a "city block" distance metric, a case is judged correctly classified if all 5 of its nearest neighbors are case - complement and none is a complement - case.

Let c_i be the number of loci for individual i at which the case and the complement differ. Define a weight w_i to be 0 when c_i=0 and otherwise 2^{c_i}. Let the indicator function $I_{(event)}$ be 1 if event is true and 0 otherwise. Our fitness score for each SNP set was:

$$Fitness\ Score = \frac{\sum_i w_i I_{(individual\ i\ correctly\ classified)}}{\sum_i w_i}$$

Step 1: Initialize

Initialize the population by randomly choosing sets of SNPs (here of size 2), calculating the fitness score for each individual.

Generation 1

Individual	SNP 1 index	SNP 2 index	Fitness Score
1	11	210	.523
2	315	7	.419
3	19	573	.638
4	124	953	.112
5	78	31	.201

Step 2: Procreate

Create a new generation by sampling from the previous generation with replacement. The sampling probabilities are proportional to the fitness score, except that the one with the best fitness score is always retained.

Generation 2

Individual	SNP 1 Index	SNP 2 Index	Fitness Score
1	19	573	.638
2	315	7	.419
3	19	573	.638
4	11	210	.523
5	11	210	.523

Step 3: Mutate

Each individual SNP undergoes probability ½ of being mutated, i.e. being replaced by another randomly chosen SNP. Calculate new scores.

Highlighted cells have been switched out.

Generation 2

Individual	SNP 1 Index	SNP 2 Index	Fitness Score
1	19	152	.602
2	93	7	.319
3	413	573	.561
4	11	328	.719
5	11	203	.592

Steps 4 & 5: Repeat

4: Repeat steps 2 and 3. Do this 500 times and note the SNP set with the highest fitness score.

5: Repeat steps 1-4. Do this 200 times to generate independent "populations."

Final generation, population 1

Ind.	SNP 1	SNP 2	Fitness
1	47	213	.843
2	47	115	.812
3	115	63	785
4	11	181	.754
5	11	328	.718

Final generation, population 2

Ind.	SNP 1	SNP 2	Fitness
1	47	115	.812
2	92	213	.798
3	10	413	.761
4	115	213	.704
5	78	31	.693

• • •

Final generation, population 200

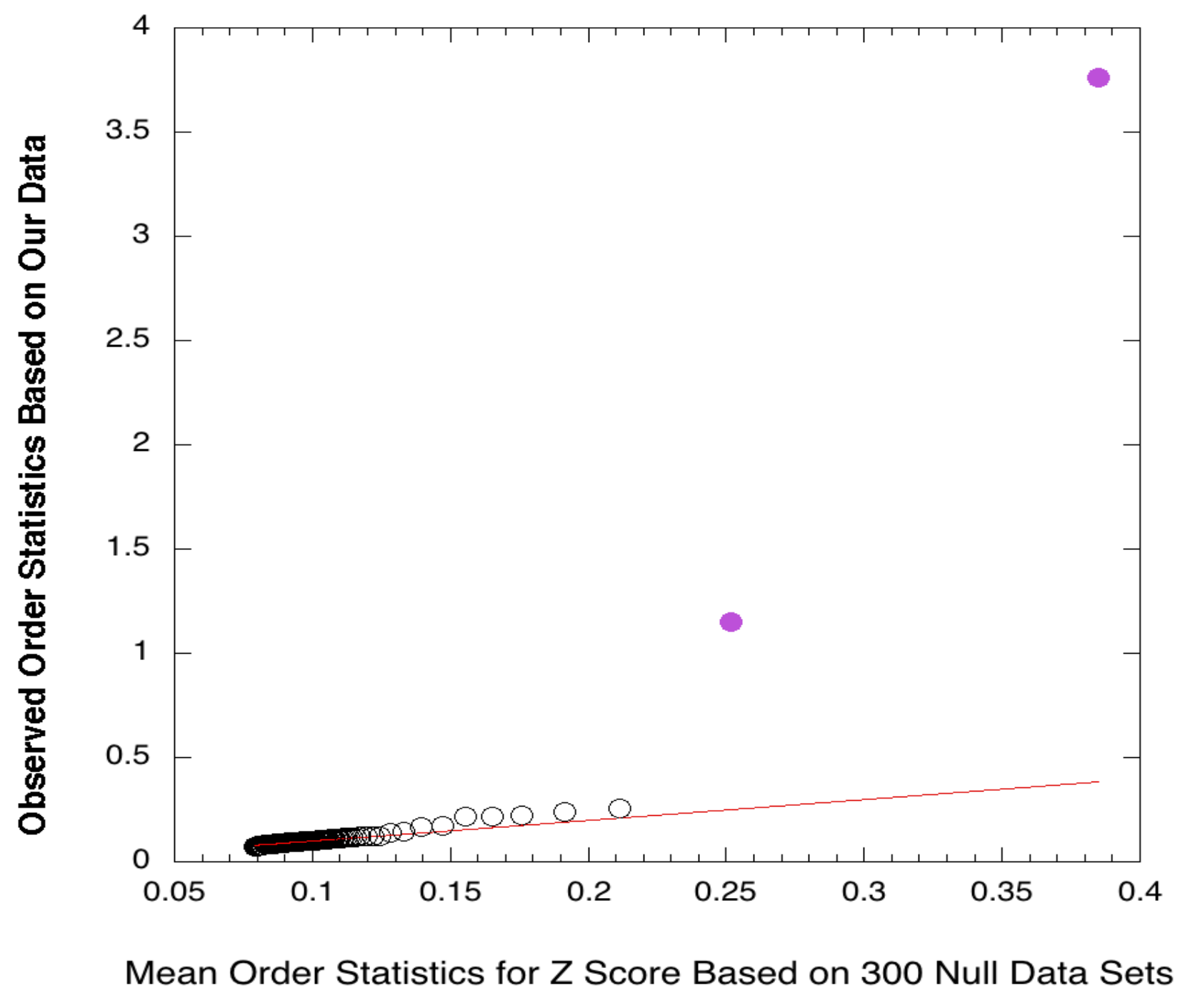
Ind.	SNP 1	SNP 2	Fitness
1	47	213	.843
2	92	213	.798
3	115	60	.781
4	47	953	.734
5	11	320	.712

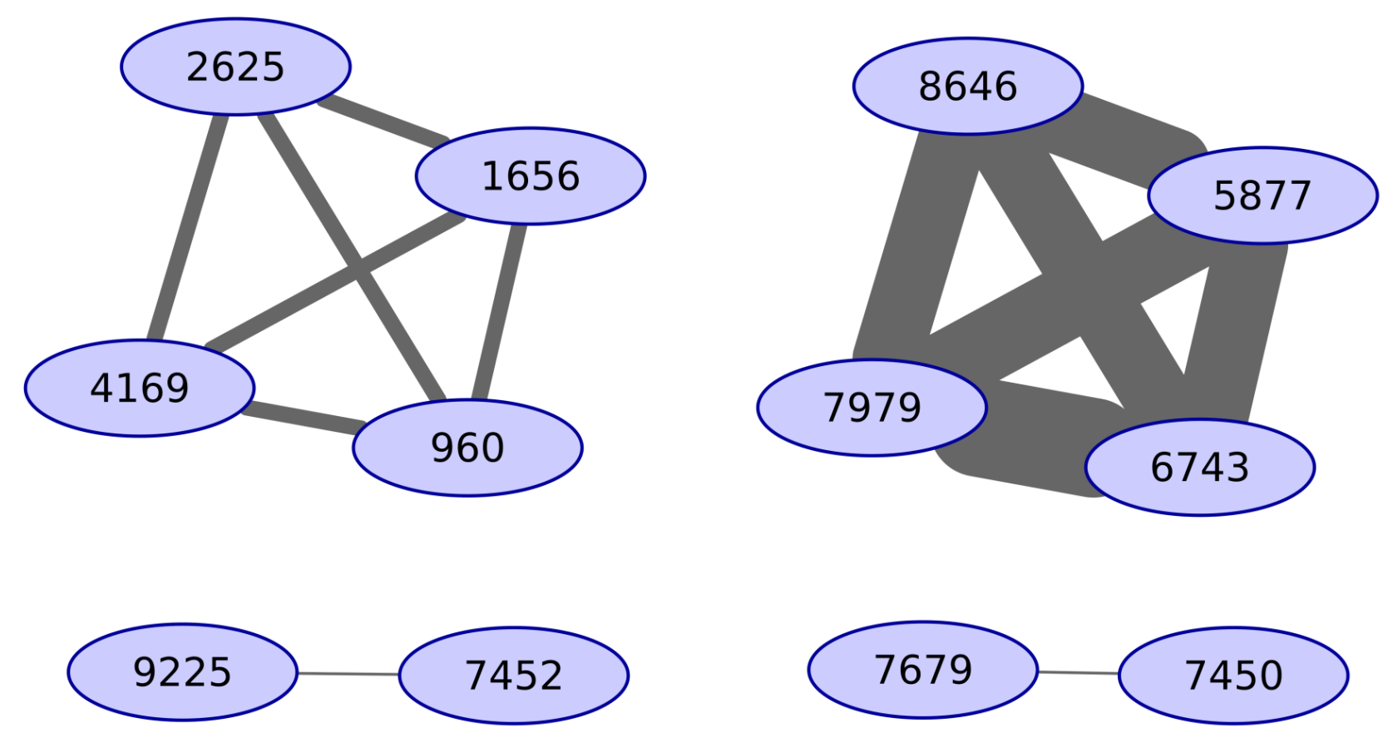
We run the algorithm with populations of 600 individuals. At the end we have 200 independent evolved sets and from each one we record the winning set of SNPs. Note that some sets may appear repeatedly among the 200 selected best sets. We can now assess frequency of sampling of individual SNPs and frequency of sampling of pairs of SNPs, etc.

Generating the null distribution by permutation

- We randomly flip the case/complement labels for the 1000 families in the study and run the entire EA for the permuted data, which has been generated to be null.
- We calculate a score, Z, based on multiplying the fitness of a selected set by how many times it was selected into the list of 200 winners.
- We re-permute the data repeating the EA analysis and calculate a new list of Z scores 300 times.

Results of the EA for SNP sets of size 4





We used the null results to estimate the means of the order statistics for Z scores under the null. Under the null those would be close to the observed values of the order statistics. The two pathways (needles in the haystack) are shown in purple.

We use the software, Cytoscape, to explore possible networks based on the frequency of selection of pairs into the list of 200 winning SNP sets.

Ongoing work

- Assess performance under more complex scenarios
- Develop approaches to estimate a FDR

Acknowledgements

Datasets were obtained from dbGaP at www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000094.v1.p1. Funding support for the study entitled "International Consortium to Identify Genes and Interactions Controlling Oral Clefts" was provided by several previous grants from the National Institute of Dental and Craniofacial Research (NIDCR). Data and samples were drawn from several studies awarded to members of this consortium. Funding to support original data collection, previous genotyping and analysis came from several sources to individual investigators. Funding for individual investigators include: R21-DE-013707 and R01-DE-014581 (Beaty); R37-DE-08559 and P50-DE-016215 (Murray, Marazita) and the Iowa Comprehensive Program to Investigate Craniofacial and Dental Anomalies (Murray); R01-DE-016148 (Marazita), R21-DE-016930 (Marazita); R01-DE-013939 (Scott). Parts of this research were supported in part by the Intramural Research Program at the NIH, National Institute of Environmental Health Sciences (Wilcox, Lie). Additional recruitment was supported by the Smile Train Foundation for recruitment in China (Jabs, Beaty, Shi) and a grant from the Korean goverment (Jee).

The genome-wide association study, also known as the Cleft Consortium, is part of the Gene Environment Association Studies (GENEVA) program of the trans-NIH Genes, Environment and Health Initiative (GEI) supported by U01-DE- 018993. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is filly funded through a federal contract from the National Institutes of Health (NIH) to the John Hopkins University, contract number HHSN268200782096C. Funds for genotyping were provided by the NIDCR through CIDR's NIH contract. Assistance with genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01-HG-004446) and by the National Center for Biotechnology Information (NCBI).