A Recommendation Systems Approach for Detecting Epistasis in Genomic Signals

Mario Banuelos* and Marissa Hernandez†
Department of Mathematics, California State University, Fresno, Fresno, CA, 93740, USA
* E-mail: mbanuelos22@csufresno.edu
† E-mail: mhernandez98@mail.fresnostate.edu

***Abstract*—There are a variety of methods used to understand and interpret an organism's phenotype, the physical expression of one or more genes. Epistasis, the phenomenon of one mutation affecting the resulting quantitative or qualitative phenotype, is used to assess gene variation in an attempt to find a combination of single nucleotide polymorphisms (SNPs) that contribute to a certain phenotype. Since one SNP rarely completely describes an organism's phenotype, detecting these groups, or coalitions, of mutations without relying on an exponential number of numbers is one of the main challenges in this field. To alleviate these computational bottlenecks, we propose a neighborhood-based collaborative filtering approach by viewing this data with a recommender system formulation. As such, we are able to detect statistically significant higher order SNP interaction phenotypes related to muscle mice genomic variants.**

I. Introduction

The genome of all living organisms are comprised of the basepairs A, G, C, and T. Changes may occur from a variety of reasons. These changes are known as genomic variations and may take the form of longer mutations, known as structural variants (SVs), or changes of a single basepair, referred to as single nucleotide polymorphisms (SNPs) [1]. The latter mutations are known to contribute to genetic diversity and are oftentimes associated with health concerns or disease (e.g. cancer)[2], [3]. In some cases, multiple mutations may be viewed as an interconnected network, or coalition, leading to increased fitness in adaptations such as diet and altitude acclimation [4], [5], [6], [7], [8].

There are a variety of methods used to understand and interpret one's phenotype, but it is rare that one SNP completely describes the physical expression of genes. As such, we focus on detecting epistasis, the interaction and dependency of genetic mutations in an organism. In particular, we note that these effects are usually not additive and are illustrated in Fig. 1. Epistasis is also used to assess gene variation in an attempt to find a combination of single nucleotide polymorphisms that contribute to a certain phenotype [9], [10]. It originally described the masking effect a variant or allele at one locus prevents the variant at another locus from manifesting its effect [11].

Many methods exist to detect these groups of mutations and they are typically divided into quantitative or qualitative phenotypes [12], [13]. Recent methods have also included exploring non-abelian Fourier analysis to create a subset of higher-order coalitions to further consider [14]. The difficulty

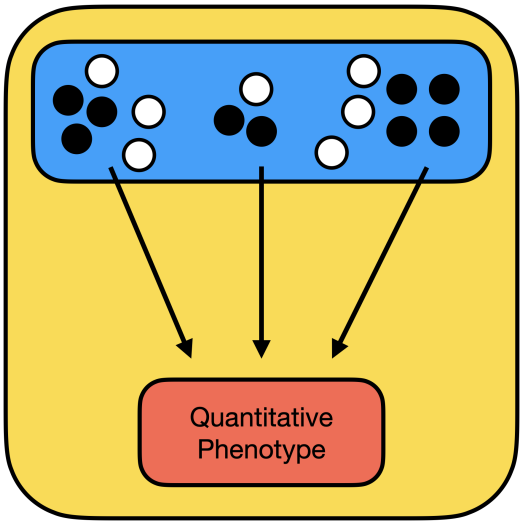

Fig. 1. Illustration of non-additive effects of single nucleotide polymorphisms (SNPs) on quantitative phenotype in an organism. Shaded circles represent mutations, while white circles represent normal basepairs. Groups of SNPs, rather than individual ones, often are responsible for disease and evolutionary adaptations.

with effectively detecting epistasis is due to complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects. Moreover, exhaustive searches within pairs, triples, or higher-order interactions of mutations may result in a computational bottleneck [15]. These types of statistical approaches also are prone to type 1 errors, and Bonferroni corrections do little to alleviate this shortcoming [16].

In this paper, we propose a recommender system approach, where we use neighborhood-based collaborative filtering techniques on a reduced data matrix to extract the most similar individuals with an extreme quantitative phenotype. We then post-process the most similar users to narrow the candidate groupings of mutation. Although this method has been widely used in product recommendations, this is the first application, to our knowledge, of such a method to study epistasis. In applying this framework, we detect higher-order interactions of single nucleotide polymorphisms by avoiding an exponential number of models. Our method is scalable to larger data, and

we present our results on both simulated and experimental data.

II. Method

We consider a recommendation system framework to detect statistically significant coalitions of single nucleotide polymorphisms (SNPs) on quantitative phenotypes in genomic signals from m individuals. If only one individual has a SNP at a location j, then that locus, or genomic position, will still be included in the data matrix. For simplicity, we focus on up to fourth-order coalitions of SNPs but our method can be expanded to higher-order interactions.

A. Observational Model

We begin by considering the SNP matrix $\mathcal{S}$,

$$\mathcal{S} = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ \vdots \\ I_m \end{array} \begin{array}{c} \begin{array}{ccccc} S_1 & S_2 & \cdots & \cdots & S_n \end{array} \\ \begin{bmatrix} 0/1/2 & \cdots & \cdots & \cdots & 0/1/2 \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \cdots & 0/1/2 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0/1/2 & \cdots & 0/1/2 & \cdots & 0/1/2 \end{bmatrix} \end{array}, \quad (1)$$

where S_i $(i = 1, 2, \ldots, n)$ represents the n SNPs to be considered and I_j $(j = 1, 2, \ldots m)$ represents the m individuals (i.e., I_j represents the jth row and S_i represents the ith column of $\mathcal{S}$, respectively). We note that each entry in $\mathcal{S}$ is either 0, 1, or 2, where 0 indicates no presence of a SNP, 1 indicates one copy of the SNP, and 2 indicates two copies. The quantitative phenotype vector $\vec{P} \in \mathbb{R}^m$ is horizontally stacked to the right of (1), to create the data matrix, $\mathcal{X}$,

$$\mathcal{X} = [\mathcal{S}|P]. \quad (2)$$

We consider the original data matrix as well as a heterozygous variation of (2) in which the presence of a SNP will be considered a 1. We refer to this as the binary, or heterozygous, reduction of (2).

B. Recommender System Approach

We implement a collaborative filtering approach in which we compute the SVD of $\mathcal{X}$ [17], [18]. We denote the low-rank approximation, with rank r, of $\mathcal{X}$ as $\tilde{\mathcal{X}}$. To determine similarity, $w_{t,u}$, between individual t with the highest (or lowest) quantitative phenotype when compared to individual u, we use both the Pearson correlation measure

$$w_{t,u} = \frac{\sum_{j \in \mathcal{S}} (I_{t,j} - \bar{I}_t)(I_{u,j} - \bar{I}_u)}{\sqrt{\sum_{i \in \mathcal{S}} (I_{t,j} - \bar{I}_t)^2 \sum_{i \in \mathcal{S}} (I_{u,j} - \bar{I}_u)^2}},$$

and cosine similarity,

$$w_{t,u} = \cos(I_t, I_u) = \frac{I_t \cdot I_u}{\|I_t\|_2 \times \|I_u\|_2}.$$

Instead of removing duplicate individuals, we avoid the singularity caused from identical individuals in (1) by adding a small amount of Gaussian noise $\mathcal{N}(\mu = 0.005, \sigma^2 = 0.01)$ to (1). This approach results in the k individuals most similar to individual t.

C. Grouping Top SNPs

After extracting the top k individuals most similar to the top (and lowest) individual t through a chosen similarity metric, we continue to group the overlapping mutations shared by these k individuals. We find the highest first-order to fourth-order single nucleotide polymorphisms by combining the total presence of each SNP within the set of the top k individuals and take the mean of each combination.

We summarize our neighborhood epistasis recommendation detection approach in Algorithm 1 below. By default, we report the top 10 most similar individuals.

Algorithm 1: Neighborhood Epistasis Recommendation Detection (N.E.R.D.) Algorithm

```
function N.E.R.D. (X);
  Input  : SNP - Phenotype Data matrix X = [S|P]
  Output: Highest and Lowest SNP groupings
begin
    (optional) Create binary reduction of S.
    Extract t_Highest and t_Lowest individuals from P
    Compute SVD of X + N(0.005, 0.01)
    for t_Highest and t_Lowest do
        Calculate similarity metric for all m individuals
        Determine top 10 similar individuals
        Calculate mean occurrence of SNPs for top 10
         individuals
    end
    Return SNP groupings with highest mean
     occurrence among top individuals with respect to
     t_Highest and t_Lowest.
end
```

III. Results

To validate our method, we implemented our method to detect SNP coalitions on both simulated and a subset of real mice genotype-phenotype data [19]. All experiments were run on a commodity machine with 8 GB of RAM and an Intel i5 processor. In all of our experiments, we use a rank 15 approximation for $\tilde{\mathcal{X}}$. Simulated results followed a similar approach as [14], and our method was able to detect first as well as higher-order interactions in this data (results not displayed).

A. Karst Mice Data

In the validation of our method, we investigated the weight and lean mass of male and female mice in an effort to draw a conclusion between their quantitative phenotypic response and the high order combinations of their SNPs [19]. This previously study provides a candidate set of mutations in

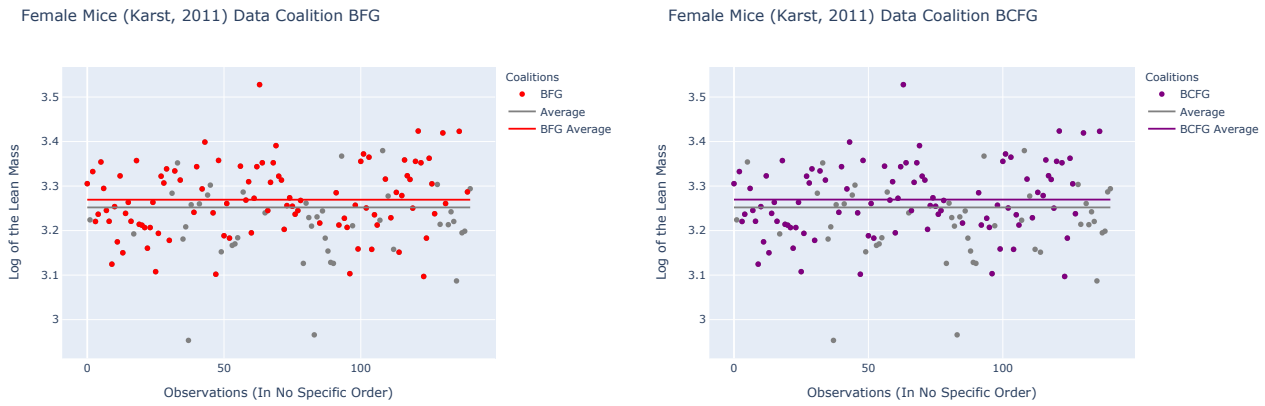

Fig. 2. Mice population with log(lean mass) plotted. *Left.* Female mice with coalition BFG are highlighted in red. *Right.* Female mice with coalition BCFG are plotted in purple. In both examples, our method (using cosine similarity) identifies these groupings of SNPs as ones with a statistically significant difference from the population's lean mass.

muscle mice from intercrossed lines. Here, we focus on the set of SNPs from Chromosome 1 and note that each SNP has a value of 0, 1, or 2, where: 0 indicates no presence of a copy from the parents, 1 indicates 1 copy (haploid) from the parents, and 2 indicates two copies (diploid) from the parents. For simplicity, we alphabetically labeled each SNP as seen in Table I.

TABLE I
CHROMOSOME 1 SNPS (ALPHABETICAL LABELS) FROM KARST (2011) MUSCLE MICE

Label	SNP	Label	SNP	Label	SNP
A	rs31194300	F	rs31684041	K	rs3672697
B	rs4222269	G	rs31234127	L	rs31424068
C	rs4222320	H	rs31791013	M	rs32257630
D	rs31991963	I	rs32520046	N	rs31474366
E	rs31886089	J	rs4222579	-	-

We computed the SVD of each set of mice data so we could extract the rank $r = 15$ approximation. We found the mouse with the highest phenotypic response and the mouse with the lowest phenotypic response for lean mass and body weight and applied Algorithm 1 to find the top 10 similar mice and we reported the weighted average of SNPs most occurring in this group. Both Fig. 2 and 3 illustrate the difference between our method's suggested groupings and how those subpopulations differ in lean mass and body weight, respectively. The detected groupings are not reported in [19], but yet may prove to have a positive effect on increased lean mass and body weight.

From Table II, we see that our approach is able to identify statistically significant (p-value < 0.05), first-order, second-order, and higher-order groupings of SNPs. Given data for male and female mice, we were only able to obtain statistically significant coalitions for female mice and this warrants further exploration. In Table III, we see the strongest SNP combinations for first-order, second-order, and third-order for

TABLE II
STATISTICALLY SIGNIFICANT COALITIONS FOR BODY WEIGHT AND LEAN MASS

Coalition	**p-value**		**Sex**
	Body Weight	*Lean Mass*	
BC	0.00041	0.0068	Female
BF	0.00419	0.0116	Female
BG	0.000080	0.00021	Female
CF	0.00467	0.02294	Female
CG	0.000093	0.00058	Female
FG	0.00328	0.00136	Female
EF	0.0126	0.0246	Female
EG	0.00034	0.00061	Female
BCF	0.00408	0.0188	Female
BCG	0.000078	0.00045	Female
BFG	**0.000078**	**0.00021**	Female
CFG	0.000093	0.00058	Female
EFG	0.000339	0.000607	Female
GMN	0.0553	0.0165	Female
IMN	0.1293	0.0432	Female
BCFG	**0.000078**	**0.00045**	Female
BEMN	0.0367	0.0663	Female
EFGH	0.00275	0.00683	Female
EFGI	0.00294	0.00763	Female
EFGJ	0.00188	0.00507	Female
EFGK	0.0103	0.0409	Female
EFGL	0.0033	0.0197	Female
EFGM	0.0112	0.0093	Female
EFGN	0.0019	0.0016	Female
FGMN	0.0553	0.0165	Female

both male and female mice body weights and lean mass. We hypothesize that including a higher rank approximation of the SNP matrix will yield more accurate results. Nevertheless, we are able to greatly reduce the exponential number of models when detecting these higher-order interactions.

One of the drawbacks of this method, similarly to that of other methods, is the issue with calculating the higher-order SNP iterations by taking the mean. We found this method to be quicker than using logistic regression, and it can still

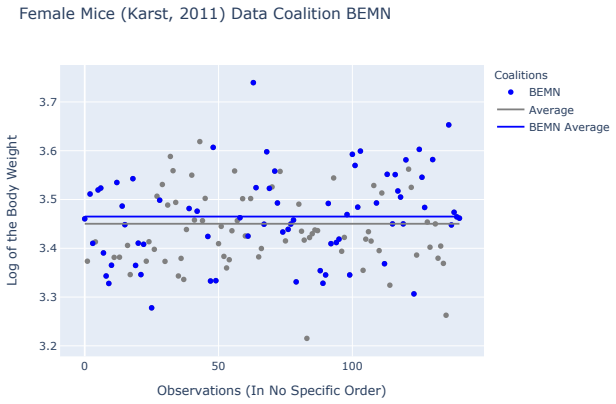

Fig. 3. Mice population with log(body weight) plotted. Female mice with coalition BEMN are highlighted in blue. In this case, our method (using cosine similarity and Pearson correlation) identified this groupings of SNPs which reflect a statistically significant difference from the population's body weight.

TABLE III
STRONGEST COALITIONS OF MAX LEAN MASS DATA

Coalition	Strength	Sex
J	17.0	male
HI	16.0	male
HIK	16.0	male
L	16.0	female
CD	15.0	female
CDE	15.0	female

be implemented rather quickly. Our results on real data also suggest future experimental validation routes for researchers without extensive combinations of mutations.

IV. CONCLUSIONS

In this paper we present a recommender system approach to detecting epistasis in many individuals across SNPs. In particular, we showed how our method successfully identifies statistically significant higher order interactions of genomic mutations in mice data. By framing quantitative phenotypes as a rating, our binary reduction also successfully identifies the most similar individuals. Obtaining these values and combinations yields candidate groupings by comparing those SNPs with the quantitative phenotypic response. This and similar approaches may also have personalized health benefits when other -omic data sets are incorporated by detecting important groupings of contributing factors to disease and health. In future studies, we plan on systematically investigating higher-order interactions while varying the number of principal components, incorporating genome-wide association studies (GWAS) data in scaling our approach, and extend our model comparisons to other methods.

ACKNOWLEDGMENT

M.B. and M.H. thank California State University, Fresno and the College of Science and Mathematics for funding support.

REFERENCES

[1] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.
[2] I. Martincorena and P. J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.
[3] Rocio Acuna-Hidalgo, Joris A Veltman, and Alexander Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome biology*, 17(1):1–19, 2016.
[4] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.
[5] Wen Huang, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert RH Anholt, Julien F Ayroles, Laura Duncan, Katherine W Jordan, Faye Lawrence, Michael M Magwire, et al. Epistasis dominates the genetic architecture of drosophila quantitative traits. *Proceedings of the National Academy of Sciences*, 109(39):15553–15559, 2012.
[6] Matteo Fumagalli, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, Pascale Gerbault, Line Skotte, Allan Linneberg, et al. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1347, 2015.
[7] C. Jeong, D. B. Witonsky, B. Basnyat, M. Neupane, C. M. Beall, G. Childs, S. R. Craig, J. Novembre, and A. Di Rienzo. Detecting past and ongoing natural selection among ethnically tibetan women at high altitude in nepal. *PLOS Genetics*, 14(9):1–30, 09 2018.
[8] Guido A Gnecchi-Ruscone, Paolo Abondio, Sara De Fanti, Stefania Sarno, Mingma G Sherpa, Phurba T Sherpa, Giorgio Marinelli, Luca Natali, Marco Di Marcello, Davide Peluzzi, et al. Evidence of polygenic adaptation to high altitude from tibetan and sherpa genomes. *Genome biology and evolution*, 10(11):2919–2930, 2018.
[9] Thomas F Hansen. Why epistasis is important for selection and adaptation. *Evolution*, 67(12):3501–3511, 2013.
[10] Trudy FC Mackay and Jason H Moore. Why epistasis is important for tackling complex human disease genetics. *Genome medicine*, 6(6):42, 2014.
[11] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.
[12] Asko Mäki-Tanila and William G Hill. Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367, 2014.
[13] Clement Niel, Christine Sinoquet, Christian Dina, and Ghislain Rocheleau. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6:285, 2015.
[14] David Uminsky, Mario Banuelos, Lillian González-Albino, Rosa Garza, and Sylvia Akueze Nwakanma. Detecting higher order genomic variant interactions with spectral analysis. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.
[15] Junliang Shang, Yingxia Sun, Jin-Xing Liu, Junfeng Xia, Junying Zhang, and Chun-Hou Zheng. Cinoedv: a co-information based method for detecting and visualizing n-order epistatic interactions. *BMC bioinformatics*, 17(1):1–15, 2016.
[16] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392, 2009.
[17] Prem Melville and Vikas Sindhwani. Recommender systems. *Encyclopedia of machine learning*, 1:829–838, 2010.
[18] John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. *arXiv preprint arXiv:1301.7363*, 2013.
[19] S. Kärst, R. Cheng, A. O. Schmitt, H. Yang, F. P. M. De Villena, A. A. Palmer, and G. A. Brockmann. Genetic determinants for intramuscular fat content and water-holding capacity in mice selected for high muscle mass. *Mammalian genome*, 22(9-10):530, 2011.