Artificial selection experiments for association in model organisms

Suyash Shringarpure Eric Xing

August 2012
CMU-ML-12-104

Carnegie Mellon

Artificial selection experiments for association in model organisms

Suyash Shringarpure **Eric Xing**

August 2012
CMU-ML-12-104

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Genetic association studies have been used to examine the genetic basis of many diseases. They have found genomic markers which contribute to risk for a number of diseases. However, genetic association studies have failed to explain the large genetic contribution to complex traits such as height.

In this report, we examine the feasiblity of using artificial selection experiments on model organisms (specifically, *Drosophila melanogaster*) to improve the performance of genetic association methods and understand the nature of genetic associations better. We use simulated artificial selection experiments on *Drosophila melanogaster* to generate genotype data and perform association using sparse regression methods. We demonstrate that this approach improves the accuracy of association methods at recovering causal polymorphisms for a range of allele frequencies and effect sizes.

Keywords: Artificial selection, genetic association, sparse regression.

1 Introduction

Many diseases (and other traits) are inherited familially. Genetic association studies involve finding the polymorphisms in whole-genome genetic data that are responsible for specific phenotypic traits (often diseases) [Consortium, 2007]. Many candidate polymorphisms in the genome have been found to have effect in diseases such as diabetes [Saxena et al., 2007, Sladek et al., 2007], Crohn's disease [Libioulle et al., 2007], prostate cancer [Thomas et al., 2008], breast cancer [Antoniou et al., 2008, Eeles et al., 2008]. These and many other genome-wide association studies have helped improve our understanding of many diseases. However, they have been unable to explain the large fraction of genetic contribution to phenotypic diversity of traits such as height.

In this report, we examine how artificial selection experiments on model organisms can be used to improve association studies. Section 2 examines the the current status of association studies and the challenges they face. Section 3 describes the history of artificial selection experiments and summarizes existing methods of genetic association. We explain the artificial selection approach in more detail in Section 4. Section 5 contains the results of preliminary experiments on simulated data from *Drosophila melanogaster* and Section 6 discusses future work in this direction.

2 Genetic association studies

As outlined earlier, association studies have vastly improved our understanding of many complex traits. Underlying these studies is the "common disease, common variant" assumption, which hypothesizes that common diseases are affected by common allelic variants present in more than 1-5 % of the population [Collins et al., 1997, Pritchard, 2001]. Current SNP chips capture millions of such variants and thus provide a convenient way of setting up such studies.

To analyze the utility of association studies, it is helpful to analyze quantitative traits in terms of their heritability. The heritability of a trait is defined as the fraction of the phenotypic variance of the trait that can be explained by additive genetic factors [Hindorff et al., 2009]. Another important idea that helps our understanding of association studies is the effect of a allele variant, or the increase in the risk of having the disease due to the presence of the variant. The aim of association studies is to find alleles that have some (large or small) effect and account for as much of the heritability of the trait as possible.

2.1 Current status of genetic association studies

Genetic association studies are usually set up in one of three different ways:

Familial studies In familial studies, pedigrees with a known history of a particular disease are genotyped. This avoids the problem of population stratification (an important problem occuring in association studies that will be

explained in more detail in Section 2.2). However, the restriction on the individuals that can be included in the study limits the power of the method in finding associations.

Case-control design Case-control studies involve a comparison between the genotypes of two sets of individuals characterized by presence or absence of the phenotype of interest. Cases are a group of individuals who exhibit the phenotype of interest (a diease or a complex trait). Controls are individuals who do not show prevalence of the phenotype. The underlying assumption is that genotypic differences (in terms of the frequency of certain allelic variants) between cases and controls are likely to be at markers which are causally related to the phenotype being studied.

In most recent studies, association studies are set up using a case-control design.

Population cohorts Rather than designate two different sets of individuals as cases and controls, population cohorts follow a single set of individuals over a longer period of time, collecting phenotypic information for multiple traits. This limits the number of "cases" for a particular disease that might be present in the cohort, but the resulting data includes a lot of longitudinal information about multiple phenotypes that can be useful for other studies of diseases. In particular, environmental and lifestyle information about the cohort can also be used to study the effect of epigenetic factors on various traits [Wong et al., 2004]. It also allows for the study of pleiotropy [Cordell and Clayton, 2005], which is the phenomenon by which a single allelic variant of a gene can affect multiple traits, which may or may not be known to be functionally related.

For certain diseases such as age-related macular degeneration, it has been found that only a few common variants having large effects account for most of the heritability of the trait. Scenarios such as these are conducive to analysis by genome-wide association studies. In many other diseases, most common variants only add small increments to the disease risk and explain only a small percentage of heritability. An example of such a trait is human height, with an estimated heritability of 80%. Genome-wide association studies have indicated $\sim$40 loci that might be associated with human height, but they explain only 5% of the phenotypic variance of human height. Similar problems have been encountered when trying to explain the heritability of other complex traits using association studies. Below we discuss some more of the challenges that are faced when performing association studies.

2.2 Challenges in genetic association studies

Population stratification Case-control studies are based on the assumption that genotype differences between cases and controls are likely to be causually

related to the phenotype. However, if there is unidentified population stratification between the cases and controls, this assumption does not hold true. If the cases disproportionately represent a genetic population in comparison to the controls, then any SNP with allele frequencies differing between the cases and controls will (incorrectly) be found to be associated with the phenotype, when it is only truly associated with distinguishing case or control status. A variety of methods have been proposed to identify and correct for population stratification in association studies [Price et al., 2006, Pritchard et al., 2000, Puniyani et al., 2010, Roeder et al., 1998].

Insufficient sample size It has been suggested that the partial success of genetic association studies could be a result of not sampling enough individuals. Small sample sizes could result in rigorous tests of statistical significance failing to identify variants of small or moderate effects as causal. Recent work by Yang et al. [2010] suggests that increasing sample sizes identifies new SNPs that allow us to explain up to 40% of the heritability of human height. While this is a significant improvement, it still accounts for only half of the estimated heritability of the trait.

Single locus association statistics Many traditional tests for association are single-locus tests for statistical significance. Due to the large number of statistical tests that have to be performed for all genotyped SNPs, a correction factor must be applied to the test statistic to avoid false positives. A commonly used correction is the Bonferroni correction, by which the test statistic is reduced by a factor of the number of SNPs. This assumes that all the tests performed are independent. However, due to linkage disequilibrium, SNPs are correlated and therefore the tests are not independent of each other. The Bonferroni correction, therefore, is too conservative and ignores weak associations.

Effect size distribution The early genome-wide association studies have been able to identify candidate SNPs that have large effects. The undiscovered causual variants are likely to have smaller effects. Therefore finding newer candidate loci in association studies is likely to be a harder problem [Park et al., 2010].

Comon disease, rare variants Current SNP chips capture variation only at loci where the minor allele frequency (MAF) is between 1-5%. However, low frequency (MAF $\leq$ 1%) variants and rare variants (MAF $\leq$ 0.01 %) are not captured. Since many traits are multifactorial, a relatively small number of rare variants with moderate effect could account for a large percentage of the trait heritability.

3 Related work

Genetic association and artificial selection have both been widely studied. Below, we briefly describe some of the related work in both these areas.

Historically, breeding of plants and animals after domestication can be considered to be artificial selection. The first recorded artificial selection experiments were performed only after 1945 [Hill and Caballero, 1992]. They were used to show that almost any quantitative trait could be altered, that the response was due to change in gene frequencies and that many genes must be involved. Artificial selection experiments have been used for understanding trait variation, estimating genetic covariances, correlations among traits [Garland Jr, 2003, Hill and Caballero, 1992].

Traditional methods for genetic analysis of diseases used techniques such as linkage analysis of candidate markers or genes and quantitative trait locus (QTL) mapping using one marker and one phenotype at a time [Easton et al., 1993], followed by a correction for multiple hypothesis testing [Benjamini and Hochberg, 1995, Storey and Tibshirani, 2003]. Recently, methods have been developed that enhance power by allowing analysis of multiple markers at once [Balding, 2006]. Methods such as eigenanalysis [Price et al., 2006] and regression [Cordell and Clayton, 2002] can perform simultaneous analysis of multiple markers for associations. Mixed models such as EMMA [Kang et al., 2008] extend the regression framework to model the association problem (with confounding variables) as a linear mixed model.

Model organisms have been used to finding genetic associations in many earlier studies, using various experimental methods such as developing transgenic animals or gene knock-out experiments. Artificial selection experiments have also attempted to use allele frequency data to find markers associated with traits. To our knowledge, this work represents the first attempt at using artificial selection experiments to perform genetic association using genotype data. Potentially, this will allow us better control over the data, and avoid problems such as low causal allele frequencies or population stratification.

4 Proposed approach

We propose an artificial selection setup for finding genetic associations. Artificial selection experiments belong to a class of experiments know as laboratory selection [Hill and Caballero, 1992]. Laboratory selection experiments are a useful tool for studying evolution. They can be used to answer questions about adaptations, trait associations, etc [Garland Jr, 2003]. In artificial selection experiments, individuals are chosen to propagate the next generation if they express particular values of a desired phenotypic trait. These experiments allow the experimenter more control over the selection experiment.

The artificial selection experiment setup involves two sets of individuals, a selected group and a control group. The control group is a set of individuals on which no selection is performed. The selected group undergoes selection

according to a regime of selection strength and consistency as chosen by the experimenter. As described earlier, individuals from the selected group are chosen to reproduce to form the next generation if they express particular values of a phenotypic trait. For most traits, selection can be performed to obtain either high values of the trait or low values of the traits. Artificial selection experiments therefore often have two selected sets of individuals, one group selected for high values of the traits and the other selected for low values of the trait. To ensure that the experiment results are due to selection and not due to genetic drift, the experiment is often performed with more than one replicate.

The steps in an artificial selection experiment are:

1. Begin with an initial population of individuals as the current generation.
2. Measure the value of the phenotypic trait chosen for selection in all individuals in the current generation.
3. Individuals whose phenotype value matches a prespecified criterion for the phenotype (for example, trait value larger than an absolute or relative threshold) are chosen to be the parents for the individuals in the next generation.
4. The chosen parents are allowed to mate to produce a new generation of individuals. The number of individuals created is the same as that in the original population.
5. Repeat steps 2-4 with the new population.

Steps 2-5 are performed for the number of generations chosen by the experimenter.

We propose to set up an artificial selection experiment by breeding *Drosophila melanogaster* for a trait of interest. We can then genotype (some of) the generations of individuals created during the artificial selection experiment. The sequenced genotypes and measured phenotypes can then be used for performing association between genotype and phenotype.

4.1 Technical Approach

We will use sparse linear regression from the genotypes on to the phenotypes to recover associations. As described in Section 3, sparse regression methods have previously been successfully used in association studies. The association problem is modeled as a regression problem, with the the genotypes at each locus being the covariates and the phenotype as the dependent variable. The regression coefficients then allow us to estimate the significance of the association between the locus and the phenotype.

We used the lars package by Efron et al. [2004] to perform the sparse regression. For computational efficiency, lars was only run for 10 steps, thus recovering the 10 loci that have maximum association with the phenotype.

5 Preliminary results

For simulating the artificial selection experiment, we used the code provided by Dr. Hudson (personal communication). In the setup, selection and recombination is assumed to occur on females only. We assume that there are no dominance effects and that loci have only additive effects. The parameters to the code are the recombination rate between the ends of the chromosome, the number of ancestors, the number of segregating sites and the number of ancestors, the positions, frequencies and heredities of the QTLs. The recombination probability (between the ends of the chromosome) is set to 0.25 across all experiments, and the number of ancestors is set to 100. Each generation has 400 individuals. From the simulation code, we can extract genotype and phenotype data from any generation of the selection experiment.

5.1 Design of experiments

With the artificial selection experiment setup described above, there are many questions that must be answered when analyzing the utility of the experiment with respect to recovering genotype-phenotype associations. The primary questions we address are:

- What is the effect of using data from different generations on the recovery of associations?
- What is the effect of varying the amount of total heredity of the phenotype?
- How much data do we need to recover associations correctly?
- What is the effect of per-locus heredity on the recovery?
- What is the effect of changing the total number of segregating sites on the recovery?

We will design our experiments to address these questions. The question of effects of heredity is crucial to all our analyses and therefore heredity will be a parameter in all the experiments we design. To address the question of effect of number of segregating sites on recovery, each experiment will be performed at two different settings, with number of segregating sites set to 2,000 and 50,000 in the two cases. The number of QTLs affecting the phenotype is set to two, with heredity shared equally, unless specified otherwise. At each experimental setting, 20 independent selection runs are used to compute the score statistics. To quantify how well associations are recovered, we define a measure of accuracy below.

5.2 Evaluation measure

Association studies generally use p-values as a measure of evaluating how well associations are recovered. P-values are usually computed using permutation

tests, with many $(10^5 - 10^6)$ permutations. However, since we are performing a very large number of simulations, this would be computationally prohibitive. Instead, we define a new measure for evaluating recovery based on our knowledge of the ground truth about the QTLs that affect the phenotype.

Since we use a sparse-regression based approach with only 10 steps, the regression output is a list of 10 loci. The loci are ordered in non-increasing order of their association with the phenotype. Therefore, we use the rank of the known QTLs in this list as an accuracy measure. In particular, the evaluation measure is given by the sum of the ranks of the known QTLs in the list of loci output by the regression. Thus, in a case with two QTLs, the ideal regression output would have the two QTLs ranked 1 and 2 in the output list, thus giving an error measure of 3. Similarly, for a case with three QTLs, the error measure would be at least 6. If a known QTL is not found the the regression output list, a penalty of 10 is added to the score measure.

5.3 Effect of generations

First, we will analyze the effect of using data from various generations on the recovery of associations. Two natural choices are:

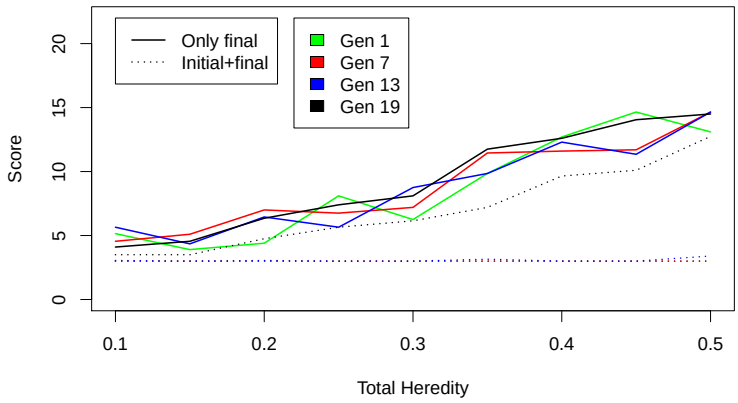
- Use data only from the final generation of the artificial selection.
- Use data from the initial and final generation of the artificial selection. This has the advantage of having double the number of individuals (compared to the previous option) and also having large variation in the phenotypes and genotypes of the two generations.

To examine the effect of using data from various generations on the recovery of associations, we will use variations of the second scheme described above. We will use data from generation X (X = 1/7/13/19) and the final generation to do the regression. We will also compare the results to using data from only the final generation.

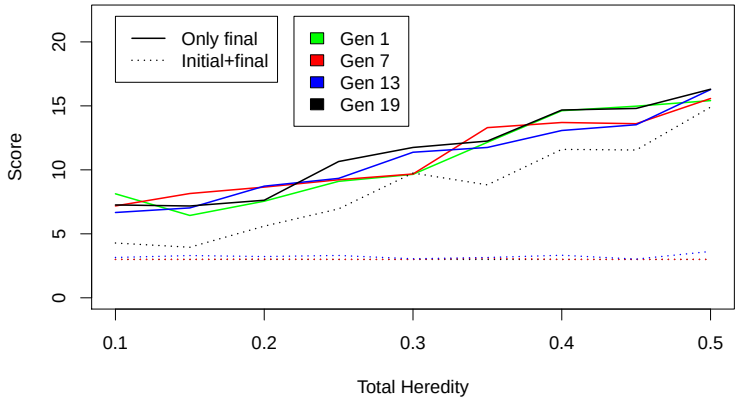
Figures 1(a) and 1(b) show the results of using data from different generations on the recovery. For easy visual interpretation, we have shown only the mean value of the score at each experimental setting.

From the figures, we can see that using data from two generations is almost always better than using data from just the final generation. The only exception is when data from the 19th and 20th generations is used for the regression. In this case, the behavior is almost identical to that seen when using data only from the final generation. Both these observations suggest that genotypic and phenotypic diversity in the data is important for the regression to work well. However, this could also be in part due to the larger number of individuals that the regression involved. We will examine how the number of individuals chosen has an effect on association recovery later.

Another interesting effect we observe is that the performance of the regression with only data from the final generation becomes worse as the total heredity



(a) With 2000 sites



(b) With 50000 sites

Figure 1: Effect of changing the initial generation on the recovery of associations. The solid line shows results when only data from the final generation is used for regression. The dashed line shows results when data from an initial generation and the final generation is used. The different colors indicate different choices of initial generations 1,7,13,19.

of the phenotype increases. This could be because the increased correlation between genotype and phenotype results in reduced variance in the phenotype in the data used for regression. We are still exploring this effect and more analysis

is needed to explain this behavior satisfactorily.

We also observe that the results are not significantly affected even when the number of sites is increased to 50000. There are minor effects on accuracy as seen in the mean scores, but the overall behavior is nearly the same. This is an encouraging sign since the real data will be quite large in size.

5.4 Effect of number of individuals

A possible reason for the good performance of using data from two different generations could be the increase in the number of individual samples. To examine whether the number of individuals has an impact on accuracy, we replicated the experiments using only half the number of individuals (200) from each generation. Figures 2(a) and 2(b) show the results of those experiments. In this case, we show the boxplot of the score statistics for each experimental setting. A small random jitter has been added to the scores for ease of visualization.

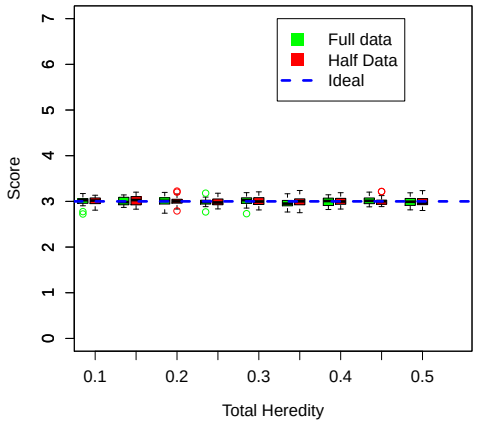
From the figures, we can see that accuracy is not affected even if the data size is halved. This suggests that the improved performance earlier is not just due to an increase in data size. Another thing to note is that accuracy is not significantly affected even if the number of sites is increased to 50000, though there are more errors than in the case with 2000 sites.

5.5 Effect of phenotype values

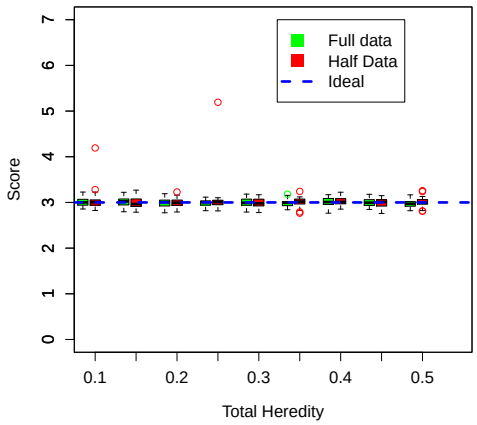
The previous experiment suggests that phenotypic/genotypic diversity, and not the number of individuals, is the important factor in determining the accuracy of the regression. To understand this, we need to examine whether the actual phenotypic values are important for accuracy or if it suffices to replace the phenotype by a dummy variable indicating the generation it comes from. In particular, we will replace the phenotype of individuals in the initial generation by a '0' and the phenotype of individuals from the final generation by a '1', and perform the regression on the modified data. We also perform the regression using half the data to examine the effect of the number of individuals simultaneously.

Figures 3(a) and 3(b) show the results of the experiments. We can see that the associations are still recovered correctly, even when using only half the data from each generation. This suggests that the important factor affecting accuracy is phenotypic diversity, rather than number of individuals. However, the experiment suggests that just the presence of phenotypic diversity, rather than the actual phenotype values, are important for accuracy. We should note, though, that this could be due to our particular choice of a rank-based evaluation measure, and that a more sensitive evaluation measure that actually measures the strength of the association (through a regression coefficient) might show that phenotypic values do have an effect on accuracy.

From the figures, we can also see that the accuracy is higher in the case with 2000 sites than in the case with 50000 sites, as suggested by the larger number of high error values in Figure 3(b). Similarly, the higher error values in the red



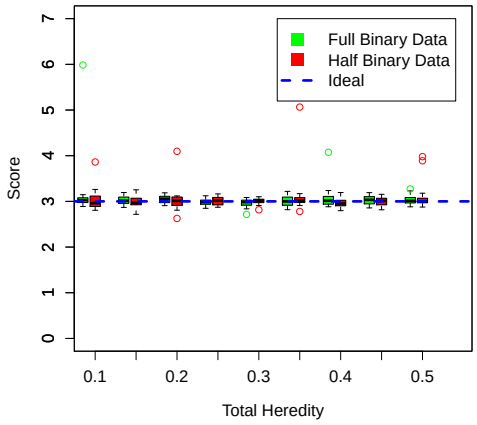
(a) With 2000 sites



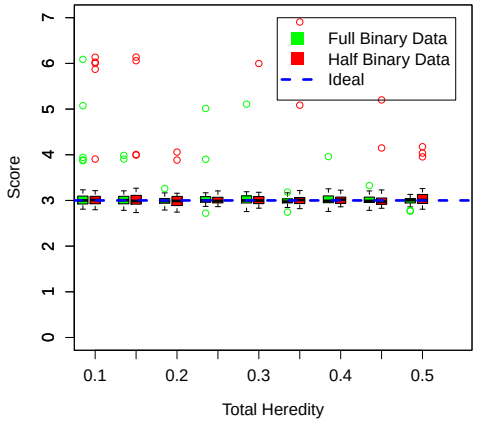
(b) With 50000 sites

Figure 2: Effect of number of individuals on the recovery of associations. The green boxplot shows results with full data - 400 individuals from the initial generation and 400 from the final generation. The red boxplot shows results with only half the number of individuals from each generation. The blue line shows the ideal score.

boxplot (half data per generation) compared to the green boxplot (full data per generation) suggests that the number of individuals does have a small effect on the accuracy.



(a) With 2000 sites



(b) With 50000 sites

Figure 3: Effect of replacing phenotype values by a dummy variable on the recovery of associations. The green boxplot shows results with full data - 400 individuals from the initial generation and 400 from the final generation. The red boxplot shows results with only half the number of individuals from each generation. The blue line shows the ideal score.

5.6 Effect of total heredity and number of QTLs

All the previous experiments involved two QTLs determining the phenotype. In this experiment, we shall examine the effect of changing the number of QTLs

that contribute a fixed value of heredity. For this experiment, we shall vary (total) heredity values from 0.1 to 0.5 in steps of 0.1. The number of QTLs will vary from 2 to 7. For each (total heredity value,number of QTLs) pair, the heredity will be shared equally among the QTLs. Figure 4 shows the ideal score graph for the various experimental settings. In the ideal case, the score measure will be constant for a particular number of QTLs regardless of the total heredity.

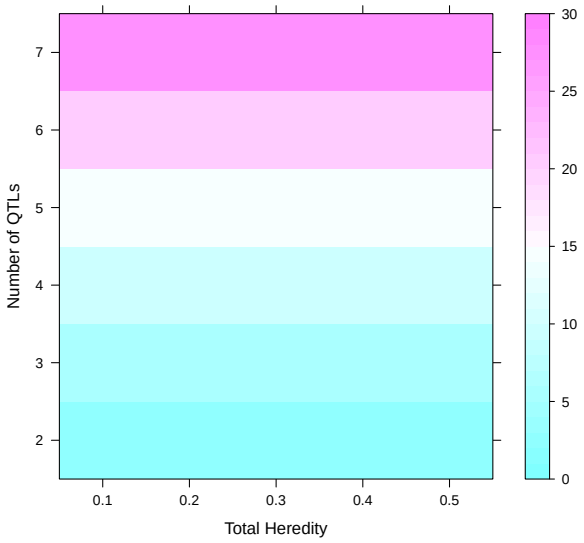
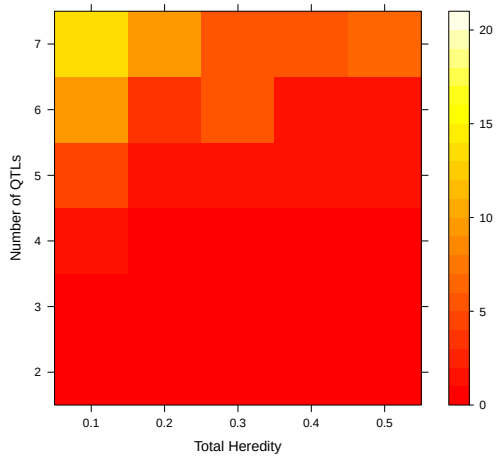


Figure 4: The ideal score graph for varying the number of QTLs and total heredity. The score increases as the number of QTLs increases regardless of the total heredity.

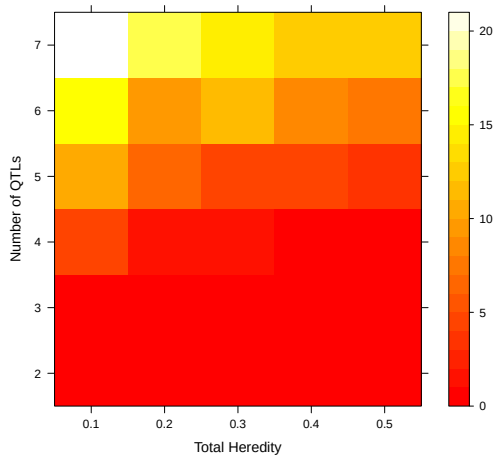
For the two cases of 2000 and 50000 sites, we shall report the error as the deviation of the mean score from the ideal score. The deviations will therefore be non-negative, and a lower deviation would be desirable. Figures 5(a) and 5(b) show the results of the experiment for the two cases.

From the figure, we can see the effect of changing the total heredity and number of QTLs. In this experiment, we can clearly see the effect of the total number of sites on the association recovery. In the absence of such an effect, both effects would be identical, or at least similar, but we see a significant difference between them, with accuracy being noticeably worse in the case with more sites.

In both figures, we can see that the deviation increases (accuracy decreases) as either the number of QTLs increases or the total heredity decreases. It is also interesting to note that total heredity and the number of QTLs are both important here, rather than just the heredity per QTL. This can be seen in



(a) With 2000 sites



(b) With 50000 sites

Figure 5: Effect of changing total heredity and number of QTLs. The effect is measured as a deviation from the ideal score, and a lower deviation is desirable. Each color value indicates a particular numerical value of the deviation. Both graphs are plotted with the same color scale for ease of comparison. Red colors indicate low deviation and lighter colors indicate higher deviation.

Figure 5(b) by observing that the deviation is different for the cases of (total heredity=0.4,number of QTLs=6) and (total heredity=0.2,number of QTLs=3).

6 Discussion

The above experiments suggest that regression methods are effective in recovering associations for a considerable range of heredity values and number of QTLs, though we must be careful when we believe that the total heredity is low and the number of QTLs is high. Phenotypic diversity is the important factor in recovering associations accurately, though the number of individuals and the total number of sites also have an effect of the accuracy. The current set of experiments suggest that we could use regression methods for quantitative traits believed to have additive interactions between loci.

Below we list some of the questions we will try and address with experiments on real and simulated data.

Identifying causative SNPs The goal is to develop a method for identifying causative SNPs contributing to a complex (fitness-related) trait. Artificial selection experiments require more labor than conventional GWAS or QTL mapping studies. However, the benefit would be increased control over confounding factors such as population structure, as well as higher resolution than can be obtained in other settings.

Power across a range of initial allele frequencies In QTL mapping, only those sites that actually differ between the two lines in the cross have the potential to be identified. In GWAS, low frequency alleles are nearly impossible to identify - the bias is strongly towards intermediate frequency alleles. The ability to identify these causative SNPs across a spectrum of initial allele frequencies would therefore be an important one. We would need to evaluate, through more simulations, how statistical power varies across a range of initial causative allele frequencies.

The distribution of effects Yang et al. [2010] suggest that effect sizes vary across a large spectrum. We would need to examine how well the method works for different effect sizes of the causal SNPs.

Modeling epistasis Many studies have indicated that epistasis (interaction between SNPs) might have an important effect on complex traits. Epistasis has also been suggested as a possible cause for the modest success of GWAS in understanding complex traits [Balding, 2006, Manolio et al., 2009]. Extending the method to allow for epistatic effects would therefore be an important technical question.

We will address these questions through large-scale simulations at data sizes comparable to that from genome-wide sequencing in *Drosophila melanogaster*.

References

A. C. Antoniou, A. B. Spurdle, O. M. Sinilnikova, S. Healey, K. A. Pooley, R. K. Schmutzler, B. Versmold, C. Engel, A. Meindl, N. Arnold, W. Hofmann, C. Sutter, D. Niederacher, H. Deissler, T. Caldes, K. Kämpjärvi, H. Nevanlinna, J. Simard, J. Beesley, X. Chen, S. L. Neuhausen, T. R. Rebbeck, T. Wagner, H. T. Lynch, C. Isaacs, J. Weitzel, P. A. Ganz, M. B. Daly, G. Tomlinson, O. I. Olopade, J. L. Blum, F. J. Couch, P. Peterlongo, S. Manoukian, M. Barile, P. Radice, C. I. Szabo, L. H. M. Pereira, M. H. Greene, G. Rennert, F. Lejbkowicz, O. Barnett-Griness, I. L. Andrulis, H. Ozcelik, A.-M. Gerdes, M. A. Caligo, Y. Laitman, B. Kaufman, R. Milgrom, E. Friedman, S. M. Domchek, K. L. Nathanson, A. Osorio, G. Llort, R. L. Milne, J. Benítez, U. Hamann, F. B. L. Hogervorst, P. Manders, M. J. L. Ligtenberg, A. M. W. van den Ouweland, S. Peock, M. Cook, R. Platte, D. G. Evans, R. Eeles, G. Pichert, C. Chu, D. Eccles, R. Davidson, F. Douglas, A. K. Godwin, L. Barjhoux, S. Mazoyer, H. Sobol, V. Bourdon, F. Eisinger, A. Chompret, C. Capoulade, B. Bressac-de Paillerets, G. M. Lenoir, M. Gauthier-Villars, C. Houdayer, D. Stoppa-Lyonnet, G. Chenevix-Trench, and D. F. Easton. Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. *American journal of human genetics*, 82(4):937–948, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18355772`.

D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16983374`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL `http://www.jstor.org/stable/2346101`.

F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human DNA sequence variation. *Science New York NY*, 278(5343): 1580–1581, 1997. URL `http://www.ncbi.nlm.nih.gov/pubmed/9411782`.

T. W. T. C. C. Consortium. Genome-wide association study of 14 , 000 cases of seven common diseases and. *Nature*, 447(June), 2007. doi: 10.1038/nature05911.

H. J. Cordell and D. G. Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *The American Journal of Human Genetics*, 70(1):124–141, 2002. URL `http://www.ncbi.nlm.nih.gov/pubmed/11719900`.

H. J. Cordell and D. G. Clayton. Genetic association studies. *Lance*, 366(9491): 1121–1131, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/16182901`.

D. F. Easton, D. T. Bishop, D. Ford, and G. P. Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *The American Journal of Human Genetics*, 52(4):678–701, 1993.

R. A. Eeles, Z. Kote-Jarai, G. G. Giles, A. A. A. Olama, M. Guy, S. K. Jugurnauth, S. Mulholland, D. A. Leongamornlert, S. M. Edwards, J. Morrison, H. I. Field, M. C. Southey, G. Severi, J. L. Donovan, F. C. Hamdy, D. P. Dearnaley, K. R. Muir, C. Smith, M. Bagnato, A. T. Ardern-Jones, A. L. Hall, L. T. O'Brien, B. N. Gehr-Swain, R. A. Wilkinson, A. Cox, S. Lewis, P. M. Brown, S. G. Jhavar, M. Tymrakiewicz, A. Lophatananon, S. L. Bryant, A. Horwich, R. A. Huddart, V. S. Khoo, C. C. Parker, C. J. Woodhouse, A. Thompson, T. Christmas, C. Ogden, C. Fisher, C. Jamieson, C. S. Cooper, D. R. English, J. L. Hopper, D. E. Neal, and D. F. Easton. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics*, 40(3):316–321, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18264097`.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. ISSN 00905364. doi: 10.1214/009053604000000067. URL `http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1083178935/`.

T. Garland Jr. *Selection experiments: an under-utilized tool in biomechanics and organismal biology*, chapter 3, pages 23–57. BIOS Scientific Publishers, 2003.

W. G. Hill and A. Caballero. Artificial Selection Experiments. *Annual Review of Ecology and Systematics*, 23(1):287–310, 1992. ISSN 00664162. doi: 10.1146/annurev.es.23.110192.001443. URL `http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.es.23.110192.001443`.

L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 106(23):9362–9367, 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19474294`.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18385116`.

C. Libioulle, E. Louis, S. Hansoul, C. Sandor, F. Farnir, D. Franchimont, S. Vermeire, O. Dewit, M. de Vos, A. Dixon, B. Demarche, I. Gut, S. Heath, M. Foglio, L. Liang, D. Laukens, M. Mni, D. Zelenika, A. Van Gossum, P. Rutgeerts, J. Belaiche, M. Lathrop, and M. Georges. Novel Crohn disease

locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genetics*, 3(4):e58, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17447842.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, Oct. 2009. ISSN 1476-4687. doi: 10.1038/nature08494. URL http://dx.doi.org/10.1038/nature08494.

J.-H. Park, S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, 42(7):570–575, 2010. ISSN 15461718. doi: 10.1038/ng.610. URL http://www.ncbi.nlm.nih.gov/pubmed/20562874.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16862161.

J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics*, 69(1):124–137, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11404818.

J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10827107.

K. Puniyani, S. Kim, and E. P. Xing. Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics*, 26(12):i208–i216, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq191. URL http://www.bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq191.

K. Roeder, M. Escoar, J. B. Kadane, and I. Balazs. Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, 85(2):269, 1998.

R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. W. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, T. E. Hughes, L. Groop, D. Altshuler, P. Almgren, J. C. Florez, J. Meyer, K. Ardlie, K. Bengtsson Boström, B. Isomaa, G. Lettre, U. Lindblad, H. N. Lyon, O. Melander, C. Newton-Cheh, P. Nilsson, M. Orho-Melander, L. Rå stam, E. K. Speliotes, M.-R. Taskinen, T. Tuomi, C. Guiducci, A. Berglund, J. Carlson, L. Gianniny, R. Hackett, L. Hall, J. Holmkvist, E. Laurila,

M. Sjögren, M. Sterner, A. Surti, M. Svensson, M. Svensson, R. Tewhey, B. Blumenstiel, M. Parkin, M. Defelice, R. Barry, W. Brodeur, J. Camarata, N. Chia, M. Fava, J. Gibbons, B. Handsaker, C. Healy, K. Nguyen, C. Gates, C. Sougnez, D. Gage, M. Nizzari, S. B. Gabriel, G.-W. Chirn, Q. Ma, H. Parikh, D. Richardson, D. Ricke, and S. Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science New York NY*, 316(5829):1331–1336, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17463246`.

R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130): 881–885, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17293876`.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/12883005`.

G. Thomas, K. B. Jacobs, M. Yeager, P. Kraft, S. Wacholder, N. Orr, K. Yu, N. Chatterjee, R. Welch, A. Hutchinson, A. Crenshaw, G. Cancel-Tassin, B. J. Staats, Z. Wang, J. Gonzalez-Bosquet, J. Fang, X. Deng, S. I. Berndt, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, O. Cussenot, A. Valeri, G. L. Andriole, E. D. Crawford, M. Tucker, D. S. Gerhard, J. F. Fraumeni, R. Hoover, R. B. Hayes, D. J. Hunter, and S. J. Chanock. Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, 40(3):310–315, 2008. ISSN 15461718. doi: 10.1038/ng.91. URL `http://www.nature.com/ng/journal/v40/n3/abs/ng.91.html`.

M. Y. Wong, N. E. Day, J. A. Luan, and N. J. Wareham. Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Statistics in Medicine*, 23(6):987–998, 2004.

J. Yang, B. Benyamin, B. P. Mcevoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42 (June):2010–2010, 2010. ISSN 10614036. doi: 10.1038/ng.608. URL `http://www.nature.com/doifinder/10.1038/ng.608`.

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA15213

Carnegie Mellon