BIOC 158/258

Homework 7: Final Project

PLM

Discovering Gene-Gene Interactions to Uncover the Missing Heritability in Genome-Wide Association Studies

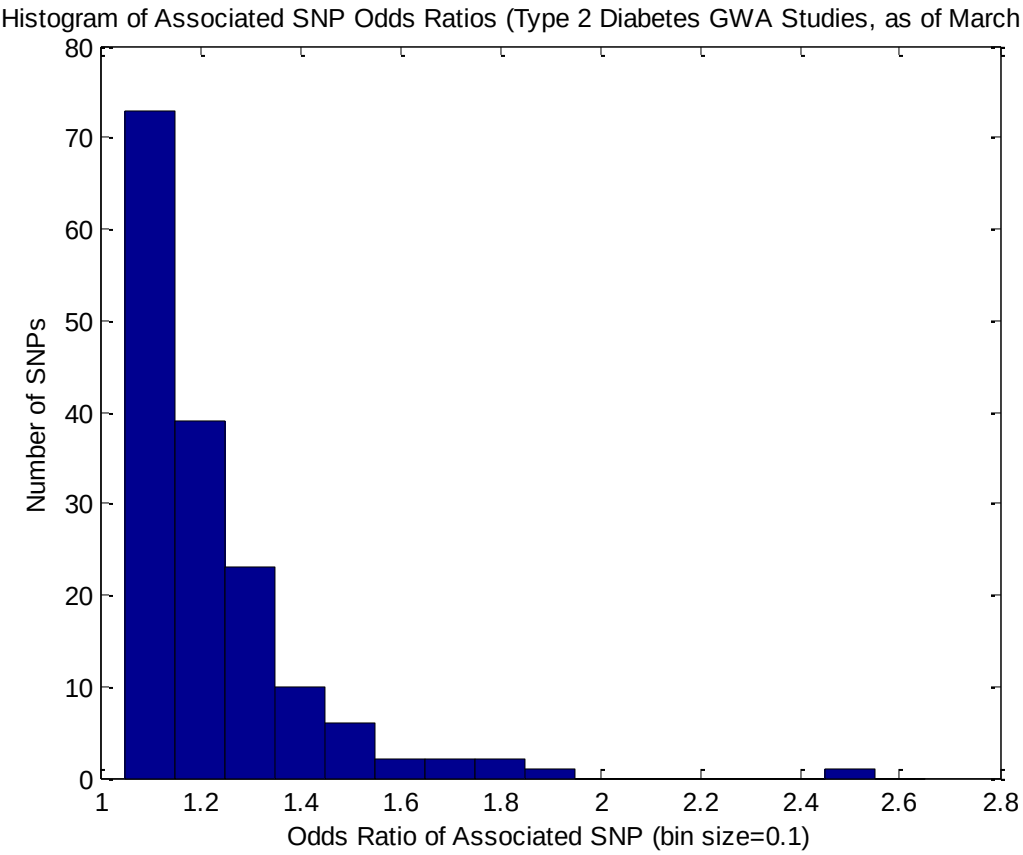
0. Background

Over the past 5 years, many genome-wide association studies (GWASs) [1] have been performed to try to discover the genetic risk factors of complex (non-Mendelian) diseases, with the goal of ultimately uncovering the genetic causes of these diseases. With this knowledge, we should be able to provide genetic tests for complex diseases, and later use our knowledge of genetic associations with disease to produce novel preventive and/or curative medicine.

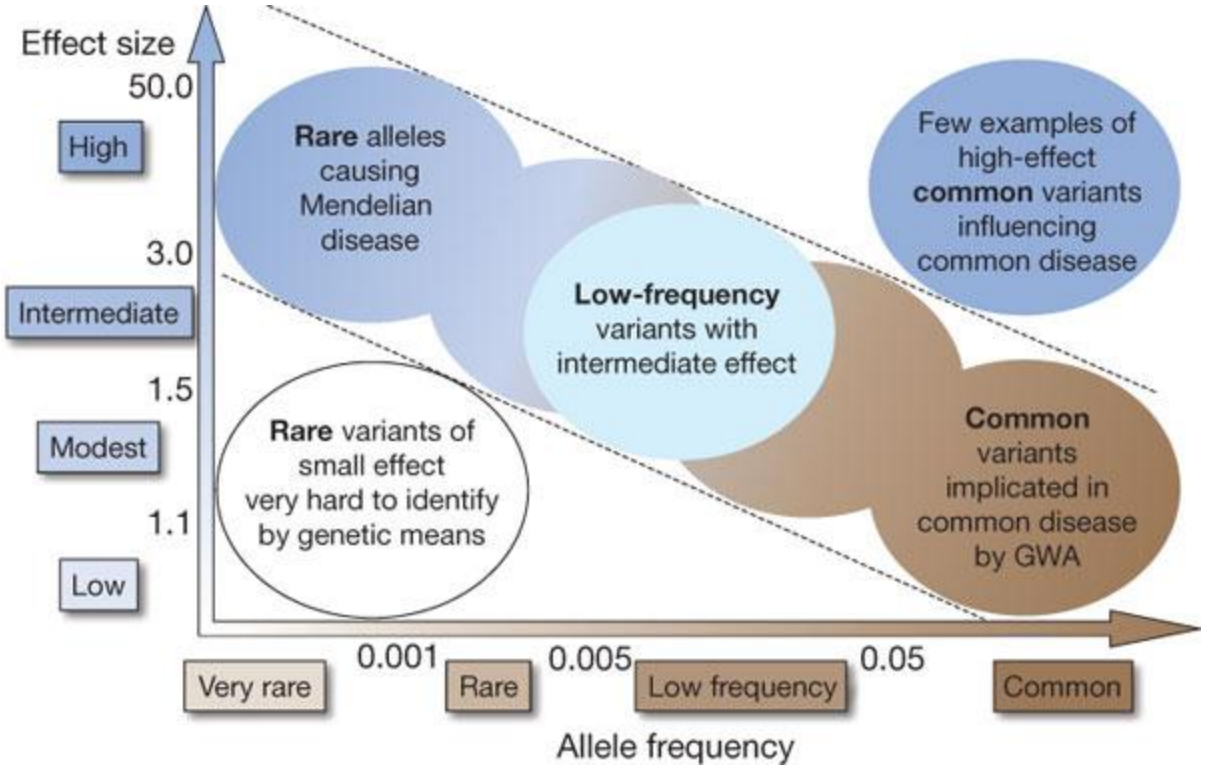
Unfortunately current association studies have largely not been able to explain the majority of the heritability of many diseases [2], and are unable to provide accurate predictions of which individuals will and will not suffer from a particular disease within a certain timeframe. This disappointment is in spite of the fact that GWA studies of important diseases such as cardiovascular disease, type 2 diabetes, Alzheimer's disease and Parkinson's disease, have uncovered 10's of single nucleotide polymorphisms (SNPs) that are associated with these diseases in case-control studies. The odds-ratios of the vast majority of SNPs that have statistical significance have values less than 2, with typically OR < 1.5.

As an example, I recorded, from the NIH GWAS database [1], all the reported odds ratios for risk SNPs found for type 2 diabetes. The database listed 29 papers, and 159 SNPs with statistical significance as risks. (Note that many of the studies replicated the findings of previous studies, and so there are fewer than 159 unique SNPs, and even fewer causative SNPs, since many SNPs are in linkage disequilibrium with others that are also reported.) I created a histogram from this

data, and it quite clearly shows that the majority of reported associations have odds ratios < 1.4, and only one reported association had an odds ratio > 2. As of late 2009, only 6% of the heritability of Type 2 Diabetes had been explained by loci associations found using GWAS [3], despite the discovery of 18 associated SNPs. (The heritability, in this case, was determined from sibling studies.)



Multiple possible explanations for our inability to uncover sufficient genetic risk factors to explain the majority of cases have been hypothesized. The review in ref. [3] focuses on two aspects: rare variants and structural variation. The “rare variants” explanation argues that GWA studies are unable to detect rare or low-frequency variants that have small or intermediate effects, and that if many complex diseases are caused by these kinds of variants, we cannot expect our current GWA studies to be able to uncover them. The landscape of possible variants and their effects, from the review, is shown below.



(Figure taken from ref. [3].)

Several diseases have already been studied by GWA methods with cohorts of tens of thousands of people. We can try to increase the power of our studies by adding further participants, but this is currently costly. (It is certainly possible that in 5-10 years, genome sequencing may become common practice in wealthy nations, and that if this data is shared, we may increase the pool of people with genotype information and disease classification to tens of millions.)

Besides the rare allele argument, ref. [3] also highlights that the SNP chip arrays that have thus far primarily been used in GWA studies do not provide information about structural variations in the genome, such as copy number variations, deleted regions, etc. Some studies have now been performed to investigate the association between structural variants and disease, but this work is ongoing, and may uncover connections that partially explain the missing heritability of some diseases.

In ref. [4], three further hypotheses are listed, in addition to those posited in ref. [3]: first, SNP chips do not cover all human SNPs, and so we may simply not be measuring the important SNPs; second, the interaction of genes with environment is typically not considered in GWA studies, and the gene-environment interaction is important (the effect of the environment should be controlled for in measures of heritability too, so this does not explain missing heritability, but it can explain our inability to correctly predict disease based on a person's genome alone); third, the interaction of multiple genes is currently very understudied, and may be important.

1. Gene-gene interactions and missing heritability

The biological pathways that are involved in the phenotypes that characterize complex diseases are typically complicated, and involve many genes. In Mendelian diseases, an important pathway often has no redundancy, and modifying a single gene can result in the disease phenotypes. The variant typically occurs in the coding region, and can, for example, result in the structure of the encoded protein being changed, or the protein being dramatically shortened due to a premature stop codon appearing. We learned in class about how Huntington's disease arises in humans who have an excessive number of repeats of a particular section of the Huntington gene.

However, in complex diseases, the situation is not quite so simple. We have examples of risk alleles that occur in the noncoding regions and cause changes in gene expression, but no changes to protein structure, e.g. for high cholesterol [5]. We can now easily imagine scenarios where multiple variants interact with each other [3, 4]. For example, a crucial biological function may be performed in two different ways, by two different pathways, and disabling or reducing the capacity of only one pathway is insufficient to obtain the disease phenotype. However, if both pathways are debilitated, then the disease phenotype can arise. Alternatively, we can imagine a single pathway that involves a feedback mechanism so it can recover from, for example, reduced or increased gene expression at some stage in the pathway, but that if

two different parts of the pathway have variants that result in changed expression, the feedback mechanisms are unable to compensate, and the disease phenotype arises.

There is therefore at least some biological plausibility for the claim that gene-gene interactions may account for some of the missing heritability of diseases in GWA studies.

As an example, consider two loci, 1 and 2, considered in a GWA study of some disease. Suppose that for all cases, locus 1 has genotype AA and locus 2 has genotype GG. Now suppose that in the controls, there are no probands with locus 1 genotype AA and locus 2 genotype GG; the only combinations that appear are locus 1 = AA and locus 2 = AA, or locus 1 = GG and locus 2 = GG. (I've only made the example so restrictive to simplify it. The example is also valid so long as in the controls, no combinations of locus 1 = AA and locus 2 = GG appear, and the overall proportion of genotypes at each locus is the kept the same for the cases and controls.) In a standard GWA study, where we only search for single locus associations, we will not uncover any association to the disease for locus 1 or locus 2, for the following reason. If we count the number of cases that have genotype AA at locus 1, and do the same for the controls, and then count the number of cases that have genotype GG at locus 1, and do the same for the controls, we will find that an equal proportion of cases and controls have the same genotype. We will therefore conclude that this SNP has an odds ratio of 1 (i.e. no effect on the disease risk). We will make a similar conclusion for locus 2. However, let us now consider what happens if we consider the interaction of locus 1 with locus 2. We will observe that every case has locus 1 = AA and locus 2 = GG, but that no controls have this *combination* of genotypes at these loci.

Of course in a real biological system, we will likely never see such a clear signal from a combination of SNPs, but no signal from the SNP loci when considered individually. Nevertheless, we have now seen how interactions between genes can plausibly lead to only combinations of SNPs yielding associations, and thus how gene-gene interactions may be able to explain some of the missing heritability in complex disease studies.

2. The computational burden of searching for gene-gene interactions

Given that we have a problem with missing heritability in current GWAS results, and that gene-gene interactions provide a plausible idea for how we might uncover better associations, one might reasonably ask why every GWA study isn't already performing tests for gene-gene interactions.

I think that the answer is (at least) two-fold: first, and most importantly, performing gene-gene interaction searches is very computationally expensive, and so long as single-SNP searches continue to yield results, researchers will continue performing these much less computationally demanding analyses; second, preliminary studies on gene-gene interactions have not yielded any particularly exciting results – consequently the potential gains from performing gene-gene interaction searches are seen as being far outweighed by the costs.

I will focus on the first point now, and will return to the issue of lackluster initial results later.

A typical GWA study uses genotyping arrays that test between 10^5 and 10^6 SNPs. If we want to search for associations with all possible combinations of two SNPs, then we need to perform the analogous case/control tallying that gets done for 10^6 SNPs, with $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ combinations, where *n* is the number of SNPs, e.g. $n = 10^6$, and *k* is the length of the combinations to consider, e.g. $k = 2$ if we want to consider all combinations of 2 SNPs.

This is a very significant problem. Note that if $n = 10^6$ SNPs, just to investigate the effects of two-locus combinations requires the calculation of ~ 5×10^{11} different combinations! The total number of three-locus combinations is ~ 1.7×10^{17}. The exponential growth in problem size as you increase the number of loci in the combinations, *k*, means that it is completely infeasible to perform a brute-force analysis for even modest values of *k*.

Because of the size of GWAS data sets, even the case of searching for associations with pairs of SNPs (k=2) is difficult. Ref. [4] claims that on a single computer circa 2009, their analysis of a single dataset of a case-control study with 17000 participants, would have taken ~ 570 years to complete on pairs of SNPs, based on the amount of time their software took to run each test. This is not yet completely infeasible: the task of assessing each combination is embarrassingly parallel (i.e. the task of testing one combination is independent of the task of testing another combination), and so the analysis software can fairly easily be made to run on a cluster of computers. A 1000-node cluster would thus require only ~ 7 months to complete the analysis, which is expensive, but possible. Another data point on how many computational resources are required is given in ref. [7], which stated that a brute-force study of two-SNP combinations was carried out in ~ 33 hours on a 10-node cluster. They had data from 1000 cases and 1000 controls, and had $n = 3 \times 10^5$ SNPs. Their software was presumably better optimized than that of ref. [4], because we expect the computation time to scale linearly with the number of study participants. The difference is SNPs results only in a ~2.7x increase in the number of tests required. The ~ 33-hour runtime on the cluster would have been ~ 330 hours on a single computer, and would have been ~ 3300 hours on a dataset with 10x more participants, and would have been ~ 8190 hours with the same number of SNPs. This is roughly 1 year, in contrast to the time that ref. [4] claims for the equivalent analysis, which was ~ 570 years!

Nevertheless, this speedup factor obtained by optimizing the code is not particularly helpful for $k > 2$ analyses: for $k = 3$, the runtime will increase by a factor of $\sim 10^5$, and so will be ~ 376 years on the same 10-node cluster. For $k = 4$, the computation time would be 37.6 million years (!).

The conclusion is that brute-force analyses of two-SNP combinations is feasible, albeit expensive, but larger combinations cannot feasibly be tested this way.

3. Statistical power

Besides the (important) issue of computation time, the quick expansion of the search space as we consider combinations of *k* SNPs has another important, detrimental impact. When we consider the association of a combination of two SNPs with a disease, we are performing another statistical test [6]. Therefore, just as it is necessary to perform a correction to our *p*-values in single SNP analyses to account for the fact that each analysis we perform on a single SNP is a single test, so too is it necessary to perform a correction to account for the fact that every combination of SNPs being analyzed is another test. Therefore, for a fixed number of study participants, considering two-SNP combinations in addition to just single SNPs significantly reduces the statistical power of the survey. As we have already covered in the previous section, the number of combinations is very large, and so the number of extra tests being performed is very large. If one uses a Bonferroni correction, which is very conservative, the effect of any two-SNP combination will have to be very large for it to have a corrected p-value that is < 0.05. This loss of power can't merely be dealt with by increasing the number of study participants, since it would be necessary to increase it by a factor of ~ 10^5 for a study with n ~ 5 x 10^5 SNPs. For most current large studies, which have ~ 10^4 to 10^5 participants, there aren't even enough people on the planet to do this!

One approach is to design is a less conservative "multiple comparisons" correction than the Bonferroni correction. This is an approach taken by and described in ref. [4], following a proposal in ref. [8]. What the authors do is perform a correction that takes into account the fact that the tests are not completely independent. In particular, a correlation matrix of the SNPs is computed, and from this, they determine some measure of "how independent" each test (with a pair of SNPs) actually is from all the other tests. The correction of the *p*-values then takes this into account by lowering the factor that each *p*-value should be multiplied by.

Using less conservative multiple comparisons correction techniques is helpful, but cannot overcome the problem posed by having such a large number of tests.

4. Search space reduction

The main goal in developing a technique to search for gene-gene interactions in GWA studies can be simply described as trying to find a sensible way to reduce the search space. One needs to be able to reduce the number of tests to carry out (i.e. the number of pairs of SNPs that are tested, in the $k = 2$ case) without skipping any tests that would actually yield a statistically significant association. How can one decide, without performing the tests, which ones are unlikely to be fruitful?

Several approaches are reviewed in ref. [6]. The simplest strategy is the following. Make the assumption that if a pair of SNPs would show an association with a disease, then each SNP considered individually would also show an association, albeit a weaker one. Therefore instead of considering all pairs of SNPs, test only pairs of SNPs from the set of SNPs that pass the single-SNP association test with some *p*-value cutoff. The major benefit of this method is that you have a tuning parameter (the single-SNP significance cutoff) that you can use to limit the number of combinations you test, and the approach is very simple. The major disadvantage is that you have to make the assumption that the marginal significance of each SNP is high enough to be detected. For biological scenarios where there is a "dosing" argument, this is perhaps a reasonable assumption. (More explicitly, imagine a scenario where the expression of two different genes can be modified by two different SNPs. If the expression of only gene 1 is low, you are at some lower risk of getting the disease, and likewise if the expression of only gene 2 is low, but if the expression of both genes 1 and 2 is low, then you are at a very high risk of disease, because your body needs at least one or the other protein at some level. In this situation, a single-SNP study should find weak associations of SNPs 1 and 2 individually, but a SNP-pair analysis would find a strong association for their combined effect.) However, not all (or perhaps even many) biological scenarios satisfy this requirement, so this simple approach to reducing the search space is sure to miss some (perhaps many) SNP combinations that do actually have strong associations with disease.

A key observation one can make is that not all combinations of SNPs are equally likely to have an effect, based on our knowledge of the biology of the genes that the SNPs are nearby. We have more knowledge than just the genotyping data from our participants, and we might be able to use it to our advantage. Two examples of this approach are in ref. [4] and in ref. [9]. In ref. [4], the authors use a protein-protein interaction database to reduce the search space. In particular, they used the STRING database, which had ~71000 protein-protein interactions listed at the time of the study. They associated the SNPs with proteins in the following way. For each protein, they used the Ensembl database to find the corresponding gene. They then flagged all the SNPs that were 100 kbp on either side of that gene. They performed a search for associations only on combinations of SNPs where one SNP was associated with a particular protein (let's call it protein 1) in the STRING database, and the other SNP was associated with a another protein, protein 2, and the STRING database listed an interaction between protein 1 and protein 2. In this way they reduced the number of SNP combinations to test from ~1.25 x 10^{11} to ~4 x 10^{6}. Using this approach, the authors were able to re-analyze the Wellcome Trust Case-Control Consortium data from 2007, and found one statistically significant SNP-pair association for each of Crohn's disease, bipolar disorder, hypertension and rheumatoid arthritis.

In ref. [9] the authors used a different approach to finding biologically related genes that might be expected to show gene-gene interactions. They performed an analysis on a human reference genome, and a mouse reference genome, where the goal was to find genes in homologous regions where the frequencies of genotypes was different than would be expected from the predictions of population genetic theory about independent loci that diverge from a common ancestor. If two genes do not interact, then we don't expect to see correlations between them over evolution. The authors were able to produce a candidate set of pairs of genes that may have interactions, and tested them on data from a childhood depression genetic association study. They were able to show for one pair of SNPs that the joint effect of the SNPs was statistically significant, but that the marginal effect of each SNP was insignificant. We see that in

this study, and in the one we discussed from ref. [4], by using biological knowledge, it is possible to find associations that would have been excluded by the naïve search space reduction method of simply filtering out individual SNPs who marginal associations are very weak.

The final class of techniques for search space reduction is that which does not use biological knowledge, but attempts to apply techniques from data mining and machine learning. The field of machine learning has grown rapidly in the last 20-30 years, and now provides a substantial set of tools for automatically finding relationships in data, classifying data, and drawing inferences from data. Many machine learning techniques rely on a "supervised learning" approach, in which an algorithm is given a "training set" to train on (this consists of both the data and its classifications, so that the algorithm has a set of complete examples to learn from), and once the algorithm is trained, it can be given as set of data to classify. Techniques such as neural networks and support vector machines fall in this class. Cordell [6] discusses "multifactor dimensionality reduction", which is also a supervised learning technique. The method proceeds by taking the dataset it is given and divided it in to 10 equal parts. An algorithm is then run to fit a model to 9 of the data buckets. The model is assessed by comparing its output to the data in the 1 bucket that was not used in the fitting. This is repeated for all combinations of 9 buckets of training data and 1 bucket of validation data. At the end of this procedure, the best model is chosen.

Several Bayesian-inspired approaches that have also been reported. For example, in ref. [10], the authors present a method that averages over a set of logistic regression models (or linear models, if the measured phenotype is not binary) that are evaluated against the data (by calculating P(Data|Model)). Each model makes different assumptions about how the loci might interact, and they are averaged in a way that optimizes the ability of the combined model to detect gene-gene interactions in the data. This method has, however, only yielded results in simulations.

5. Conclusions

The missing heritability in GWA studies is a major challenge that the field is faced with. Gene-gene interactions are one plausible explanation of why at least some of the heritability is missing. Over the past 7 years, several methods have been developed to search for gene-gene interactions in GWAS data. Unfortunately the exponential growth of the search space as one considers combinations of SNPs has made it difficult to progress. Brute-force searches of combinations of $k = 2$ SNPs have been conducted in some cases, but most GWAS reports still do not present analyses for SNP combinations (for examples, see ref. [2]). The loss of power caused by performing such a large number of tests is the likely reason why brute-force approaches to finding 2-SNP associations has not been particularly successful.

Methods that dramatically reduce the search space are promising, since they tackle both the computational resources problem of brute-force searches, and also address the reduction in statistical power. Techniques that rely on prior biological data are interesting, but have not yet become popular. In a handful of proof-of-concept tests, they have yielded some statistically significant results. To date, no one has found a large set of hitherto unknown associations using these methods. Currently researchers appear to be more focused in increasing study sizes to improve study power for single-SNP associations, but despite the increases from studies with 1000's of participants to >100000 (in meta-analyses), the missing heritability has not been recovered. This suggests that one or more of the other explanations for the missing heritability is correct.

One promising use of the techniques developed for multi-locus association testing is for better understanding gene-environment interactions, and controlling for population stratification in GWA studies. The data mining and machine learning-based methods could be particular useful here, since they are able to reduce the search space without requiring domain knowledge about the data. In this respect, they are general. There are many environmental variables that we can measure, so being able to reduce the search space will be important. Likewise, being

able to automatically deal with population stratification (rather than having to perform multiple independent GWA studies for different population groups, which reduces the study power) would be a large advance.

References

[1] http://www.genome.gov/gwastudies/

[2] T. Manolio. “Genomewide Association Studies and Assessment of the Risk of Disease.” *NEJM* (2010), 363, pp. 166-176.

[3] T. Manolio, *et al*. “Finding the missing heritability of complex diseases.” *Nature* (2009), 461, pp. 747-753.

[4] M. Emily, *et al*. “Using biological networks to search for interacting loci in genome-wide association studies.” *European Journal of Human Genetics* (2009), 17, pp. 1231-1240.

[5] K. Musunuru, *et al*. “From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus.” *Nature* (2010), 466, pp. 714-719.

[6] H. Cordell. “Detecting gene-gene interactions that underlie human diseases.” *Nature Reviews Genetics* (2009), 10, pp. 392-404.

[7] J. Marchini, *et al*. “Genome-wide strategies for detecting multiple loci that influence complex diseases.” *Nature Genetics* (2005), 37, 4.

[8] J. Li, *et al*. “Adjusting multiple testing in multilocus analyses using the eigenvalues of the correlation matrix.” *Heredity* (2005), 95, pp. 221-227.

[9] Z. Bochdanovits, *et al*. “Genome-Wide Prediction of Functional Gene-Gene Interactions Inferred from Patterns of Genetic Differentiation in Mice and Men.” *PLoS ONE* (2008), 3(2).

[10] T. Ferreira and J. Marchini. “Modeling interactions with known risk loci-a Bayesian model averaging approach.” *Annals of Human Genetics* (2010), 75, pp. 1-9.