Finding Enrichments of Functional Annotations for Disease-Associated Single-Nucleotide Polymorphisms

Steven Homberg

11/10/13

Abstract

Computational analysis of SNP-disease associations from GWAS as well as functional annotations of the genome enables the calculation of a SNP set's enrichment for a disease. These statistical enrichments can be and are calculated with a variety of statistical techniques, but there is no standard statistical method for calculating enrichments. Several entirely different tests are used by different investigators in the field. These tests can also be conducted with several variations in parameters which also lack a standard. In our investigation, we develop a computational tool for conducting various enrichment calculations and, using breast cancer-associated SNPs from a GWAS catalog as a foreground against all GWAS SNPs as a background, test the tool and analyze the relative performance of the various tests. The computational tool will soon be released to the scientific community as a part of the Bioconductor package. Our analysis shows that, for R2 threshold in LD block construction, values around 0.8-0.9 are preferable to those with more lax and more strict thresholds respectively. We find that block-matching tests yield better results than peak-shifting tests. Finally, we find that, in block-matching tests, block tallying using binary scoring, noting whether or not a block has an annotation only, yields the most meaningful results, while weighting LD r2 threshold has no influence.

1 Introduction

In the human genome, approximately 2% of the genetic material actually codes for protein [1]. The remainder of the genome lies in between genes or in the untranslated introns in the middle of genes. This vast quantity of potentially functional but non-coding genetic material is comparatively not very well understood. A part of the genome not coding for protein might have roles in transcriptional and translational regulation, but the mechanism by which the function is achieved is often unknown. One of the important mechanisms to explore the function of non-coding regions of the genome is the genome-wide association study (GWAS). In a genome-wide association study, many individuals with and without a given trait are examined. Common allelic variants between different study participants are examined in order to find an association between certain variants and a disease or trait. Of particular relevance to the investigations being conducted here are those which identify an association between a single-nucleotide polymorphism and a disease. A single-nucleotide polymorphism (SNP) is a specific type of allelic variant. At a specific point in the genetic sequence, a single nucleotide (A, C, G, or T) is changed to a different nucleotide (substitution), is removed (deletion), or is added between two formerly adjacent nucleotides (insertion). When such a change occurs in a coding region of the genome, the ramifications are more clear: the protein for which that segment of the gene codes is altered because the different nucleotide sequence causes a different amino acid sequence in the polypeptide. The effect in a non-coding segment of the genome, however, is often unknown. If the variant is in between two protein-coding regions, then a transcriptional regulatory motif may have been disrupted, the folding properties of the DNA may have been changed, or the instructions for a non-coding RNA like microRNA may have been altered. If the

variant is instead in the untranslated region (UTR) or an intron, each of which is a part of a gene but is does not code for protein, the change might have an effect on stability regulation, alter the binding site for miRNA, or change the manner in which splicing or translation is regulated [2]. In each case, however, GWASs do not provide information about the mechanism. In this investigation, we seek to form hypotheses about these mutations which fall in regions that regulate transcription.

At the foundation of the enrichment-finding process we use is the information from GWAS, which identifies disease-associated SNPs. Rather than considering each of the published GWAS individually, a compiled catalog detailing all of the important results from GWAS were used to obtain information from GWAS in a uniform, easily parsed manner. The GWAS catalog used was the PheGenI catalog [6]. While GWAS in general are adept at finding associations between common SNPs and diseases, these common SNPs are co-inherited as linkage disequilibrium (LD) blocks. Using an array of common SNPs rather than individual SNPs increases the cost-effectiveness but detracts from the resolution of the studies. In GWAS, a block of SNPs linked by LD is linked with a phenotype rather than each individual SNP, so any of the SNPs in an LD block may be the causal variant, limiting their resolution.

Aside from GWAS, the other development enabling enrichment analysis is the increasingly extensive noncoding functional annotation of the genome. Annotations of the genome range from the simple to the complex. Some simply identify a SNP's placement relative to gene models, identifying whether a SNP is in a coding exon, intron, intergenic region, or untranslated region. Other annotations predict the effect of a SNP on polypeptide, including nonsense mutations, missense mutations, and synonymous mutations. Less straightforward are

annotations of the SNP's behavior within enhancer and promoter regions which promote the binding of proteins which in turn promote transcription of nearby genes. Enhancers and promoters contian regulatory motifs, which are sequences of nucleotides which allow protein binding or serve some other significant function. The action of these regulatory motifs is achieved by regulating other genes, rather than by acting independently. Thus, SNPs may modulate enhancer or promoter activity through alteration of a regulatory motif, creating a difference in transcription. Information about these annotations of SNPs is being discovered and published. The combination of these two links, from annotations to SNPs and from SNPs to diseases or traits, allow the computational techniques used here to be implemented to draw a link between the annotations and traits, thus suggesting possible causal means by which these SNPs affect a trait or disease. These possible causal links can then serve as the foundation for further biological investigation into the sources of otherwise largely enigmatic diseases.

Prior work in the field indicates the potential for success in conducting such computational analyses. Significant associations have been found between functional annotations and disease using GWAS using a variety of statistical tests [2]. However, there is no consensus on the proper method to find these enrichments [3,4], and no standard tools available to the scientific community which can compare these methods. When comparing the frequencies, the characteristics of the set of SNPs associated with the disease need to be taken into account when building the null distribution against which to test the statistic, but there are several characteristics which may need to be accounted for in building the null model. The proper process correcting for these confounding factors has yet to be determined for use at large in the enrichment search process. Similarly, the ideal way to account for LD between GWAS

SNPs and potentiall causal SNPs is also unknown. This investigation seeks to explore several of these issues.

In conclusion, the investigation seeks to apply recent developments in GWAS and genomic annotation to discover novel statistical associations between disease-associated SNPs and regulatory functions, and to evaluate different enrichment tests. Conducting this investigation provides even greater potential to discover significant enrichments of functional annotations now more than ever. The progress of annotation of the genome has progressed further, allowing access to a much richer set of annotations relative to those of previous studies. For example, rather than looking at single histone modifications, HMMs that deliniate chromatin states are increasingly available. Regulatory maps for over 100 cell types are now available from the NHGRI Roadmap Epigenome Project [7]. In addition, the ENCODE project has discovered a large number of novel regulatory motifs [8]. As a result, the set of annotations through which to search for an enrichment is greatly expanded, providing greater opportunity to find strong and potentially biologically significant enrichments for these newly available annotations. These can then be used to evaluate the process and enable further research involving statistical enrichment to be even more effective.

2 Methods

The first step in the process of finding enrichments was the gathering of SNPs from GWAS which are associated with various diseases. Tabulated separately for the various diseases considered, the sets of SNPs associated with each disease in the GWAS catalog available from the National Human Genome Research Institute (NHGRI) and the PheGenI catalog were extracted and grouped [3]. Each of these sets of SNPs associated with a given disease, however,

constitute only those which have been established directly by a particular genome-wide association study to have an association with the given disease. GWAS only idientifies one "tag" or "index" SNP for each LD block, so LD information is necessary to find all of the relationships between diseases and annotations.

Therefore, LD information from the 1000 Genomes project was used [9]. These SNPs are then each used to generate the associated variant set (AVS) of the disease [3]. For each catalog SNP associated with a disease or trait, the cluster of SNPs in linkage disequilibrium (LD) with the tag SNP is reconstructed [5]. Subsequently, rather than compiling information about the annotations of each SNP individually, annotations are compiled collectively for each cluster of SNPs in linkage disequilibrium with the identifying tag SNP. Although each SNP in a cluster is associated with the same disease as the head SNP, not every SNP associated with a given disease appears in the GWAS catalog with such an association because the GWAS catalog only contains tag SNPs. Furthermore, because all of the SNPs in an LD cluster with a disease-associated head-SNP are in some capacity associated with the disease themselves, it is unclear which SNPs are the causal variant, if any. Each SNP in the LD cluster of a head SNP contributes annotations to the set of annotations of the LD cluster rather than individually. If two SNPs in LD do happen to each appear in the GWAS catalog as associated with the disease being considered, then one is arbitrarily chosen as the head SNP of the shared LD block (LD pruning). This ensures independence. After generating the AVS, the annotations of each cluster in the AVS are then determined by intersecting the locations of the blocks of SNPs with the locations of annotations downloaded as .BED files containing the annotation information for the feature type tested [7-9]. Annotations in several categories, specifically gene model annotations from dbSNP, Dnase

hypersensitive regions, presence in regulatory motifs, and binding by regulatory proteins in ChIP

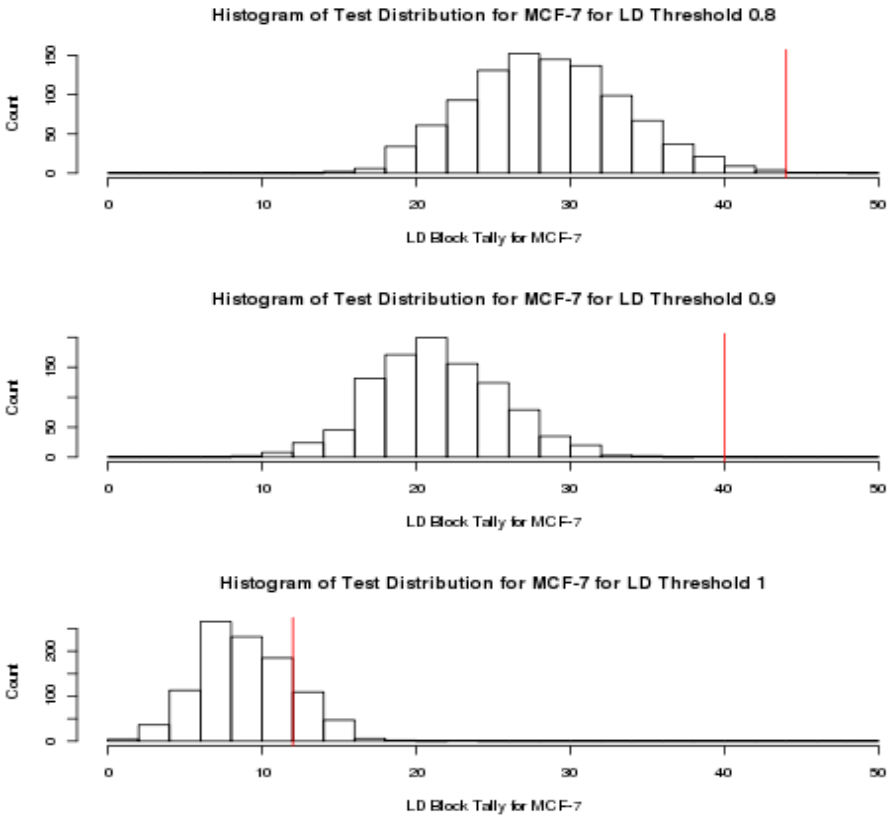

Figure 1: Test distributions for DNase hypersensitivity in the MCF-7 cell line in a global sampling procedure with scoring type 0 (binary, whether or not the LD block has the annotation or not) and without LD R2 weighting. The three histograms correspond to three LD thresholds 0.8, 0.9, and 1 at which the test was conducted. Each histogram shows the distribution of the tallies for the annotation in each of the 1000 samples. The red lines indicate the annotation's tally in the set of breast cancer-associated SNPs.

experiments, were all compiled and tabulated separately.

After generating the clusters and determining their annotations, tallies were calculated as the basis for measuring the enrichment or depletion of annotations with respect to background. The tallies are calculated in one of three ways: as the number of clusters with a given annotation, to be taken as a ratio with the total number of clusters; as the number of total instances of an annotation among all clusters, to be taken as a ratio with the total number of annotation instances; and as a proportion of the the SNPs in each block exhibiting an annotation, summed across all blocks. In order to determine significance of the enrichment, the proportion needs to be compared to a null distribution obtained by a sampling procedure. In this investigation, two

measures of background frequency were used. In the first, a global sampling procedure, information about the background frequency is garnered in a manner similar to the process for the disease-associated SNPs. A typical commercial SNP array, Affymetrix 500K, provided a large list of tag SNPs from which GWAS results are often generated, and for each SNP an LD cluster was generated in the same manner as for the disease-associated SNPs. Annotations of each type were similarly compiled and tabulated for each cluster. This in turn gives a background frequency for each annotation type against which to compare the tallies for the disease-associated SNPs. Null distributions were generated by taking groups of 1000 samples from the background LD clusters, matching each block in the foreground SNP set with an LD block in the background sharing similar values for controlling factors, like block length, allelic frequency, and distance to the nearest TSS, which could influence the annotation tallies. The second measure of background frequency involved a process of local sampling often called peak shifting. Rather than matching blocks in a background SNP set, the same distribution of SNPs in each block used in the original analysis are shifted to take on different positions within a window around their true positions, maintaining the same positions relative to each other, with the counts of the annotations intersected at the new shifted locations used to obtain tallies. Taking tallies at each of several different shifts generates the distribution against which to test. The local sampling process, rather than controlling for particular factors through binning as in the block match process, controls for the relative distribution of the SNPs directly as well as the relative distributions of annotations in the locale of the SNPs of interest.

In each case, the counts of the number of clusters with the annotation in question were totaled for each of the samples or shifts, and these counts were then used as the basis of the null

distribution. The null distributions were observed to be approximately normal as expected by CLT; therefore, a normal probability density function was fitted to the mean and standard deviation of the generated distribution and this normal model was then used to obtain a p-value.

Personal Contributions

Because of the computationally intensive nature of the process, all of the computation and statistical testing is completed by computer, and even then parallelization to split the computation across several computers was necessary. Consequently, much of the project was the creation of a tool consisting of scripts written in the R language which enabled tabulating the appropriate information from databases, conducting the necessary searches and manipulations, and performing statistical tests Aside from basic familiarization with the computer system on which I was working and with a few standard techniques for improving computational efficiency from my mentor, I completed this independently. I wrote all of the scripts for the tool and conducted all of the tests using the tool on sample data without aid.

3 Results/Discussion

Enrichment Search Tool

The first step necessary for conducting the investigation into the enrichments and statistical processes is actually to be able to calculate enrichments. In order to do this, a tool was developed to conduct each step of the search process from from database searching to scripted data manipulation to statistical processes. Taking as input an arbitrary set of SNPs, the tool conducts all of the necessary procedures to find enrichments for the given SNPs, including constructing the LD blocks around each SNP, intersecting the blocks with preloaded or custom sets of annotations, and conducting global or local sampling trials for tests on the SNP set using

these preloaded or custom annotation sets. Additionally, the tool provides features to prune sets of associations from the literature on LD, narrowing the list only to those which are LD-independent up to a certain threshold. The tool consists of a series of functions using the statistical language R and will be released to the scientific community as a part of the Bioconductor group of open source biological computation packages for R.

Analysis of the Process

The search tool created as the basis of the investigation can be used to calculate the enrichments of arbitrary SNP sets with respect to arbitrary annotation sets using a variety of configurable statistical tests. We can therefore use the tool to address the questions involved in comparing the results of the various tests and configurations to determine which give more meaningful results. Specifically, we explore changes in p-value as a function of several parameters: change in the type of test between local and global sampling; change in the r^2 threshold used in determining LD blocks; different scoring procedures counting the binary presence of an annotation in each block, the raw number of an annotation's occurrence in each block, or the proportion of each block's annotation set an annotation comprises; and change in the weighting of the level of LD linking a SNP and therefore its annotations to each block. Examining the results of tests conducted on the sets of LD blocks constructed for the various configurations can give insight into which configurations are preferable to use in the procedure. For our investigation, we use as our SNP set the set of SNPs associated with breast neoplasms, as compiled in the PheGenI GWAS catalog [6]. For background annotations, we use ENCODE DNase hypersensitivity experiments across several human cell types, conserved regulatory

motifs, as well as ENCODE protein ChIP-seq experiments. These are the annotations which we include preloaded in the R computational tool.

LD R2 Threshold

Analyzing the results of our tests across the r2 thresholds used in determining LD blocks, we see that the the best r2 threshold is an intermediate one. The most strict threshold of 1 produced poor results which failed to distinguish between the highest results. In contrast, the intermediate r2 thresholds of 0.8 and 0.9 exhibit much more meaningful results, with a peak at 0.9. Beyond generalities, we can look in particular to the annotations for DNase hypersensitivity

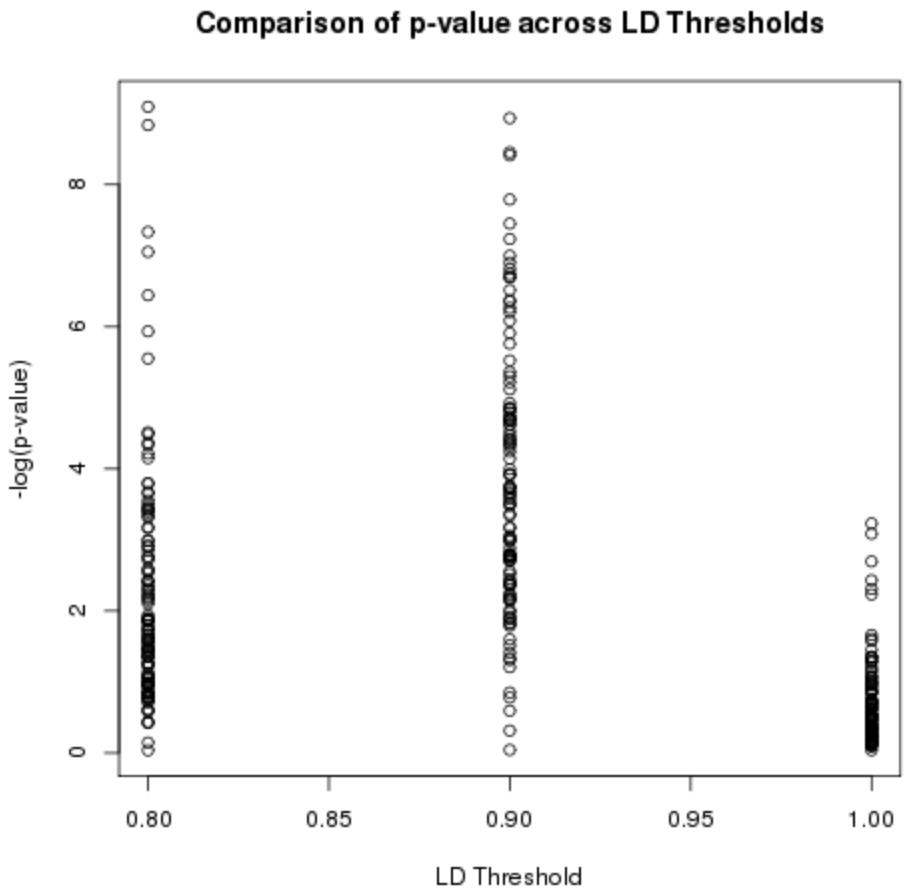

Figure 2: Plot of results of global sampling tests at varying LD thresholds. Each shows all of the results for global sampling with the indicated LD threshold using binary scoring (noting whether or not each block contains the annotation in question) and no LD r2 weighting. It is clear from the plot that the most stringent threshold of 1 provides the weakest results, while the more moderate thresholds of 0.8 and 0.9 give better results.

in the MCF-7 cell line as an indicator of the effectiveness of the test. The MCF-7 cell line is known to be associated with breast cancer because it is a breast cancer cell line, so it is expected that MCF-7 should give a strong enrichment using our computational analysis methods. We see a similar trend for MCF-7 as for all of the top enrichments, with a poor enrichment at 1 and with stronger enrichments at 0.8 and 0.9, particularly 0.9, confirming that these intermediate r2 thresholds give the most meaningful results. This matches with what theory predicts as well, as a threshold which is too strict does not include enough information about an associated SNP's ld neighborhood and potentially leaves out causal mutations, while sufficiently lax thresholds begin to approach background frequency for the foreground comparison as more and more of the genome is included in the LD blocks.

Test Type

In general, we see that the global sampling procedure provides much more meaningful results than the local sampling procedure. In the local sampling procedure, very few enrichments are meaningfully distinguished, even at the peak r2 value of 0.9, while in the global sampling the distinctions are much stronger. The MCF-7 annotation specifically exhibits the same trend, strenthening our conclusions.

Score Type for Global Sampling

With score type in the global sampling procedure, we see a pronounced trend of more meaningful results when scoring by giving each block a binary score indicating whether or not it contains the annotation. The scoring using the proportion of a block which is comprised by a given annotation gives distinctly weaker results, followed by the scoring method taking the raw number of annotation occurrences in each block which is the worst by far.

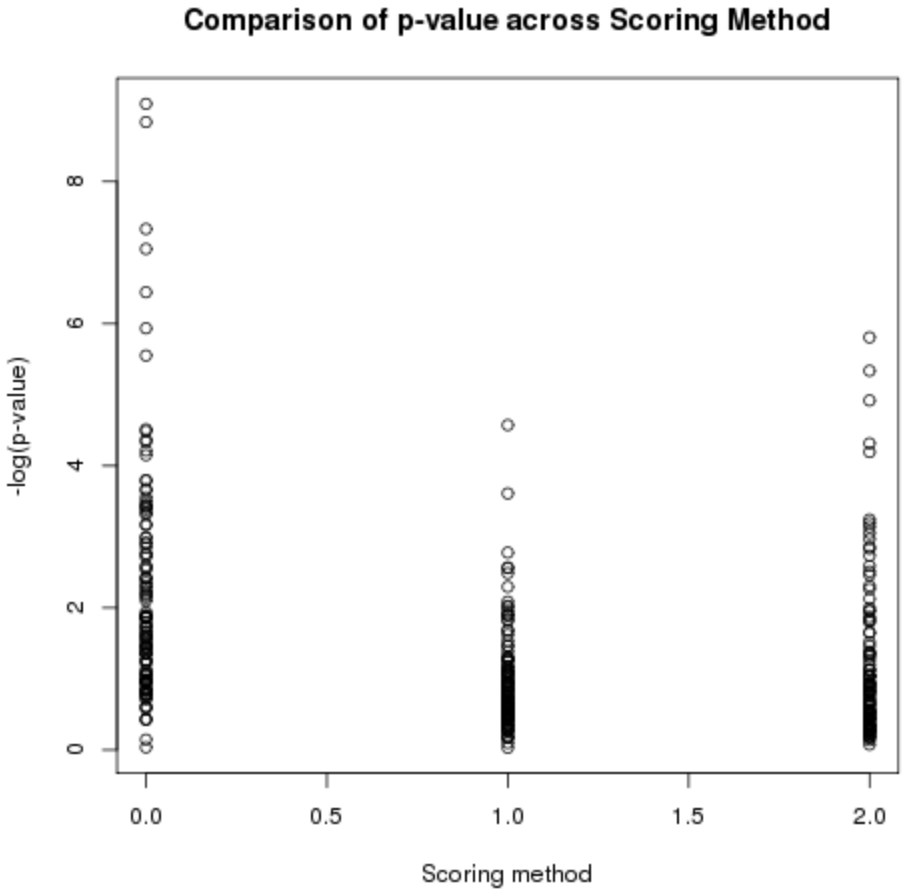

Figure 3: Plot of results for all annotations across different scoring methods. Scoring method 0 indicates binary scoring, with each block counted as 1 or 0 to the tally depending on whether or not the block contains the annotation. Scoring method 1 takes the raw number of instances of the annotation in the block. Scoring method 2 takes the proportion of a block's annotations which are of the given type. Comparing we see that binary scoring gives the strongest results, while proportional scoring is weaker and raw scoring is worst of all.

LD R2 Weighting for Global Sampling

Finally, we consider the strength of the results when comparing between scoring using weighting of LD r2 values and not using weighting. Looking at the top results for each, we see that there is essentially no change between using weighting and not. As a result, we conclude

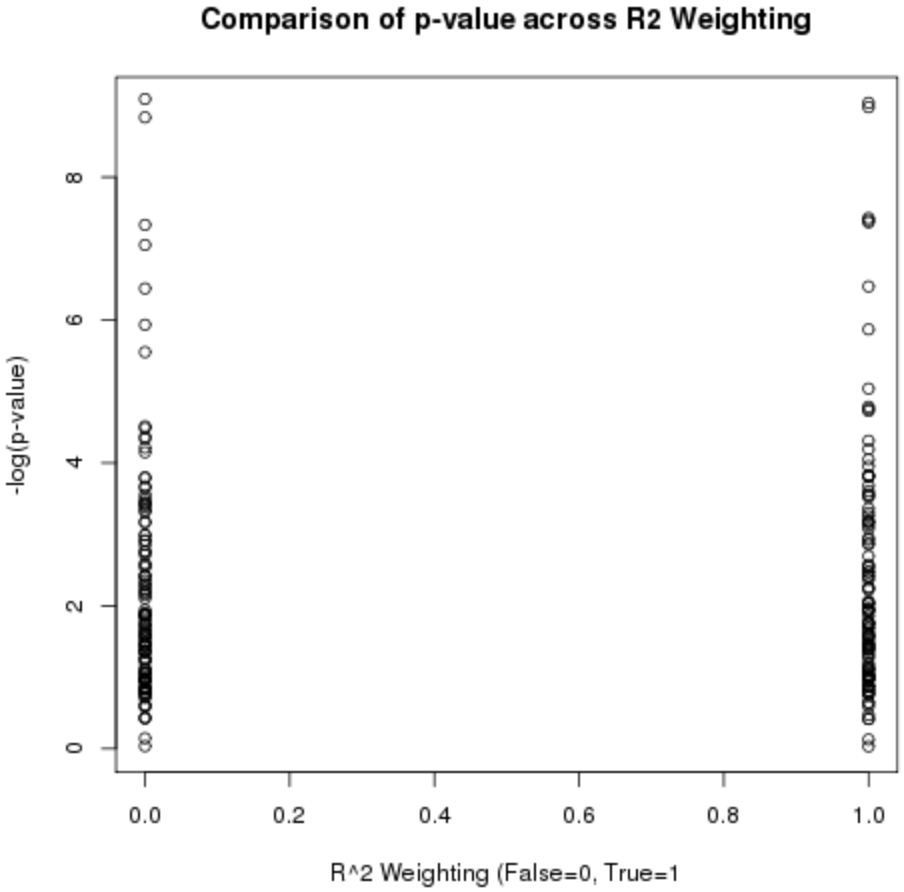

Figure 4: Plot of results with and without weighting of LD r2 values. 0 indicates no r2 weighting, while 1 indicates the use of r2 weighting. The results indicate little appreciable difference between using and not using r2 weighting.

that neither using weighting or not is preferable in conducting these tests.

Implications

The findings in this investigation include enrichments for several functional annotations with SNPs associated with breast cancer. The enrichments found provide potential causal biological mechanisms for breast cancer-associated inherited mutations. More generally, the tool developed as a part of the investigation to conduct enrichment searches on arbitrary SNP sets has the potential to produce vast numbers of such subjects for further biological investigation as the scientific community uses the tool to find enrichments for various SNP sets and various diseases being investigated. The results of analysis of the various configurations of the statistical tests used in enrichment calculations also provide indication of which parameters and scoring methods are preferable when searching for enrichments.

5 Conclusions/Future Research

From the material in the investigation alone, few conclusions can be drawn independently. The statistical enrichments found using the tool developed to conduct the enrichment search process can safely be concluded, but the true value of the research conducted is in the application of the results garnered as indicators of potentially fruitful areas of future biological investigation. As statistical tests, enrichments only establish associations, but future investigation into areas highlighted by statistical enrichments may be able to establish useful causal links between these functional annotations and diseases. Additionally, conclusions about the advantages of the different configurations for the statistical tests can be used to direct future investigations into enrichments.

Moving forward, there are several directions for further progress. As of yet, the application of the enrichment search process to breast cancer revealed several significant noncoding regulatory enrichments, notably HMEC. This process can be applied to the remainder of the diseases in the GWAS catalog, identifying further annotations associated with various other diseases. Identification, however, is only the first step. For the significant enrichments detected through the process, further investigation can be conducted to find out more about the enrichments. This provides more of a foundation for the search for causal mechanisms for these diseases.

6 Acknowledgments

I'd like to acknowledge those who enabled this project. First, I'd like to thank Professor Kellis of MIT for his help in finding a suitable area for research. I'd also like to thank Dr. Luke

Ward for all of the time and help he gave me in guiding my research process. My parents helped immensely by providing transportation in to MIT to meet with my mentor. Finally, I want to thank the MIT PRIMES program for giving me the opportunity to conduct this research.

7 Citations

[1] Haunun, Holly, Jennifer Bownas, and Kristine Christen, eds. "The Science behind the human genome project basic genetics, genome draft sequence, and post-genome science." *Human genome project information*. U.S. Department of Energy Genome Program, 26 Mar 2008. Web. 19 Apr 2013. <http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml>.

[2] Ward, Lucas, and Manolis Kellis. "Interpreting noncoding genetic variation in complex traits and human disease." *Nature Biotechnology*. 30.11 (2012): 1095-1106. Web. 19 Apr. 2013. <http://compbio.mit.edu/publications/81_Ward_NatureBiotechnology_12.pdf>.

[3] Cowper-Sal.lari, Richard, Xiaoyang Zhang, et al. "Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression."*Nature Genetics*. 44.11 (2012): 1191–1198. Web. 11 May. 2013. <http://www.ncbi.nlm.nih.gov/pubmed/23001124>.

[4] Chromatin marks identify critical cell types for fine mapping complex trait variants G Trynka, C Sandor, B Han, H Xu, BE Stranger, XS Liu… - Nature genetics, 2012

[5] The International HapMap Consortium, . "The International HapMap Project." *Nature*. 426.6968 (2003): n. page. Web. 23 Sep. 2013. <http://www.nature.com/nature/journal/v426/n6968/suppinfo/nature02168_S1.html>.

[6] Ramos, Erin M, Douglas Hoffman, et al. "Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources." *Eur J Hum Genet*. (2013): n. page. Web. 23 Sep. 2013.

[7] Bernstein, Bradley E, John A Stamatoyannopoulos, et al. "The NIH Roadmap Epigenomics Mapping Consortium." *Nat Biotech*. 28.10 (2010): n. page. Web. 23 Sep. 2013. <http://dx.doi.org/10.1038/nbt1010-1045>.

[8] "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project."*Nat Biotech*. 447.7146 (2007): n. page. Web. 23 Sep. 2013. <http://dx.doi.org/10.1038/nature05874>.

[9] "1000 Genomes project." *Nat Biotech*. 26.3 (2008): n. page. Web. 23 Sep. 2013. <http://dx.doi.org/10.1038/nbt0308-256b>.