GENOME-WIDE ASSOCIATION STUDIES, FALSE POSITIVES, AND HOW WE INTERPRET THEM

by

Richard J. Biedrzycki

BS Biological Sciences, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of the

Department of Human Genetics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Richard J. Biedrzycki

It was defended on

April 5th 2018

and approved by

Committee Member:
Wei Chen, PhD
Associate Professor, Pediatrics, School of Medicine
Associate Professor, Biostatistics, Graduate School of Public Health
Associate Professor, Human Genetics, Graduate School of Public Health
University of Pittsburgh

Committee Member:
John R. Shaffer, PhD
Assistant Professor, Human Genetics, Graduate School of Public Health
Assistant Professor, Oral Biology, School of Dental Medicine
University of Pittsburgh

Thesis Director:
Daniel E. Weeks PhD
Professor, Human Genetics, Graduate School of Public Health
Professor, Biostatistics, Graduate School of Public Health
University of Pittsburgh

Daniel E. Weeks PhD

GENOME-WIDE ASSOCIATION STUDIES, FALSE POSITIVES, AND HOW WE INTERPRET THEM

Richard J. Biedrzycki, MS

University of Pittsburgh, 2018

ABSTRACT

Genome-wide association studies (GWAS) identify genetic regions that may play a role in the development of phenotypes. Annotation of these significantly associated "peaks" for their connection to the tested phenotype is an important step in the process of GWAS. However, some of these peaks may be false positives. In this study, we analyzed annotators' ability to make convincing connections between synthetic GWAS peaks and the phenotype of interest, where synthetic peaks were taken from a GWAS of a trait genetically uncorrelated to the original trait. We provided five annotators with a mix of three original and six synthetic GWAS peaks of three peak significance categories along with relevant literature search results and asked them to annotate these regions. We asked three annotators to record how strong the evidence was at each peak connecting it to the scanned trait as well as the likelihood of the annotator's further study of that region for its role in the development of the trait. Annotation status was significantly associated with both original/synthetic peak status ($p = 0.0034$) and peak significance category ($p = 0.0112$). The proportion of synthetic peaks annotated as having convincing connections was greater than expected of an "ideal" annotator ($p < 2.2 \times 10^{-16}$). Annotators rated original peaks as having significantly higher strength of evidence ($p_{\text{Bonferroni}} = 0.0348$) and likelihood of further study than synthetic peaks ($p_{\text{Bonferroni}} = 0.0122$). Highly significant peaks had significantly higher strength of evidence than suggestively significant peaks ($p_{\text{Bonferroni}} = 0.0441$).

This study shows the ease with which annotators are able to make convincing connections between synthetic peaks and tested phenotypes. Because of the amount of resources invested into GWAS, it is essential that these studies are performed with care to reduce wasting resources on follow-up of false positives and ensure effective follow-up of true positives for the benefit of public health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my thesis advisor, Dr. Daniel Weeks for being a great mentor whose advice and suggestions were always useful and helped make this project what it is. I would also like to thank my thesis committee members, Dr. John Shaffer and Dr. Wei Chen for their suggestions and recommendations to improve the quality of this thesis.

I would like to thank Dr. Susanne Gollin for her advice and assistance in the writing of this manuscript which made the process much smoother.

I would like to thank all those who contributed their valuable time and effort by collecting and providing me with the data necessary for this project. I would also like to thank the various consortia whose provision of their GWAS summary statistics made this project possible.

1.0 INTRODUCTION

A genome-wide association study (GWAS) is a study design used to find associations between a given trait or disease and loci across the entire genome. During annotation of genome-wide association studies, false positive results can mislead and misdirect researchers, leading to wasted resources, time, and money. Annotation is the process in which researchers evaluate the results they obtained using GWAS by reviewing the literature and using tools and online resources to understand the functions of genes and single nucleotide polymorphisms (SNPs) found within the significantly associated regions. Using this information, they can determine if the traits they are testing for could have convincing connections to genes in the immediate regions of each association peak. However, there is the possibility that some of these association peaks may in fact be false positives and have no true connection to the given trait.

Our goal is to test if GWAS annotators are able to make convincing connections between a given trait and "synthetic peaks," loci from a GWAS for a trait genetically uncorrelated with the given trait. To do this, we created mixtures of "original" peaks and "synthetic" peaks from pairs of given and synthetic GWAS scans using summary statistics from previous consortia studies. We provided annotators with detailed plots of the significantly associated regions and literature search results for genes and SNPs found in these regions. Using these materials as well as other resources they wished to use, annotators were tasked with making connections between the given trait and the associated regions. We expected annotators would be able to make convincing connections between the given trait and the genes for both the “original” peaks and the “synthetic” peaks. As far as we are aware, this is the first study of its type.

We have two primary aims as part of this study. The first aim is to determine how often these “synthetic” peaks are annotated in such a way that they appear convincing and reasonable. This is defined as having plausible biological connections to the etiology of the trait. We will do this by determining the number of synthetic peaks annotated as having convincing connections for each trait and each annotator. Our hypothesis is that annotators will find a significant proportion of synthetic peaks as having strong evidence explaining the association with the given trait.

The second aim is to identify what attributes of the peaks contribute to these connections being made by the annotators. We will do this through both qualitative and quantitative analysis of the annotation.

2.0 BACKGROUND

To better understand the purpose of this study, it is important to understand the background and methodology of GWAS and the annotation process. In the following sections, we discuss the purpose of GWAS, what genetic variants GWAS can identify, and the power of statistical techniques used for analysis. We also discuss the use of annotation to assess associated variants and regions for their potential role in the etiology of the scanned trait. Finally, we discuss causes of false positives and methods used to avoid them for the best possible results. We hope that by including this background information, the purpose and rationale for this study will be made clear.

2.1 GENOME-WIDE ASSOCIATION STUDIES

2.1.1 Rationale for performing genome-wide association studies

Genome-wide association studies (GWAS) are one of a multitude of study designs used to find associations between genes and phenotypes. GWAS have been used to find genetic associations for coronary artery disease (The Coronary Artery Disease Genetics, 2011), lipid levels (Willer et al., 2013), Alzheimer's disease (Lambert et al., 2013), psychiatric disorders such as schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics, 2014), and inflammatory bowel diseases (Liu et al., 2015), to name a few.

GWAS have several defining characteristics that lend them to being an effective and commonly used study design. One such characteristic is their effectiveness in generating hypotheses. When performing a GWAS, an assumption is made that the trait has at least some genetic basis. GWAS do not require *a priori* knowledge as to which loci may play a role in the development of the trait being tested. If one is unsure what genes may be involved, GWAS can find markers in or near genes that may not have been selected if one was performing a candidate gene study, for example. Finding a gene to be associated with a phenotype does not, however, explain how the associated gene may be involved in the development of said phenotype. Once this association is found, future research can be done to determine the mechanism of the association. For example, if one is interested in finding causal variants at the associated locus, they could perform fine-mapping (P. M. Visscher et al., 2017). The locus could also be sequenced to directly identify what the causal variants may be (P. M. Visscher et al., 2017). Researchers could also perform expression analysis to find how the causal variants change expression of the gene (P. M. Visscher et al., 2017).

When performing a study to find associations between phenotypes and genes, it is important to take into account what types of variants one is expecting to find. For GWAS, the main types of variants most often tested are common variants. Specifically, a common variant is often defined as a SNP with a minor allele frequency (MAF) greater than 0.01-0.05. This study design can be traced back to the common disease, common variant (CD-CV) model of genetic variance. The CD-CV model states that common diseases, such as type 2 diabetes and coronary heart disease, attribute much of their genetic variance to variants common in the population.

However, as the field has progressed and more studies have been done examining heritability of these diseases, this model has been shown to not explain a large percentage of the heritability (Gibson, 2012). It is thought that this “missing heritability” may be partly explained by the effects of other variants following the infinitesimal model, which states that many variants of very small effect size contribute to this variance, the rare allele model, which states that rare alleles with large effect sizes contribute to this variance, and the broad sense heritability model, which states that there are gene-environment interactions that contribute to this variance (Gibson, 2012). Due to the nature and design of GWAS, it is not able to effectively detect these types of variants, so other methods like the previously mentioned fine-mapping and sequencing can be used instead. However, these methods are not the focus of this thesis; instead, let us discuss the methods and practices done when one performs a GWAS.

2.1.2 Methods for performing genome-wide association studies

The primary method of GWAS is the genotyping of markers across the genomes of many different individuals. Before one does any genotyping, it is necessary to clarify what traits are being examined and what hypotheses are being tested as part of this study. In GWAS focused on dichotomous traits, defining cases and controls consistently and correctly is an important part of controlling for selection. Selection bias occurs when either the cases or controls are not representative of the populations they have been sampled from. One possible consequence of this is biased marker allele frequency estimation. If the frequencies of the alleles are biased in either the cases or the controls, this can lead to errors during analysis. For example, if the sampled cases have a higher frequency of a marker allele than the general population such that they are identified as having a significantly higher frequency than controls, this could lead to type I error where this marker allele is determined to be associated with the tested trait in the population when there is actually no association.

As part of GWAS, DNA samples are obtained from individuals and genotyped using microarrays that detect markers across the genome. These “SNP chips” can contain hundreds of thousands to millions of different SNP probes that are hybridized to subjects’ DNA. What SNPs are chosen depend on a variety of factors. First, SNPs are chosen to provide a large coverage of the genome. As the SNPs are being used as markers for genes and loci, by adding more SNPs, more loci can be analyzed through the use of linkage disequilibrium (LD) with SNPs that are not directly genotyped. LD is a measure of the correlation between two alleles that has developed due to a lack of recombination between them over many generations. Non-genotyped SNPs may be found in genes near the genotyped markers, and through examining the LD between these SNPs, one can find associations with these genes using these correlations.

When choosing a genotyping chip to be used for a GWAS, the specific loci and regions covered by the chip are worth considering. Some chips may be designed to assess SNPs for genes related to specific traits. The Metabochip, for example, contains markers near genes and loci that play roles in metabolism and anthropometric traits (Voight et al., 2012). Because chips can have different markers, one must take care in choosing a chip that analyzes loci relevant to the study.

It is also important to consider what population is being studied when selecting a chip. As human populations diverged and evolved, allele frequencies of SNPs deviated between populations, and linkage disequilibria shrank in older populations. Due to these changes, SNPs used as markers in some populations may be ineffective as markers in others. These differences are seen in the genotyping chips used in GWAS. A SNP with a large allele frequency may serve as a marker in individuals of European descent and thus be put on a chip. However, other populations, such as those with African ancestry, may have a very low frequency for this allele, causing this allele to be a poor marker for those populations. Furthermore, SNPs that are monomorphic or very low frequency in Europeans may be excluded during chip design, even though such variants might be polymorphic in other populations. Because of this, one must consider what population the genotyping chip was designed for as it may not be able to be used effectively for other populations.

Data quality control is an important part of any study, and there are many different ways it is performed during GWAS. By using these methods, researchers can avoid making errors during their analyses. When the chips are processed and the SNPs genotyped, the time when the SNPs are genotyped can affect the results through batch effects (Turner et al., 2011). Each individual sample can also be checked for sex and chromosomal inconsistencies or abnormalities or overall sample genotyping call rate. If any of these measures show poor quality or problems, those samples can be dropped from further analysis. Specific markers may also be dropped if they fall out of Hardy-Weinberg equilibrium, have too low of a frequency for sufficient power, or are not called at a high enough rate during the genotyping process (Turner et al., 2011).

During GWAS data analysis, there are multiple methods used to plot and examine the data that have been obtained. First, to show the significance of p-values measuring the associations of each SNP with the trait of interest, one can create a Manhattan plot. The Manhattan plot is useful in that it shows on what chromosomes any significantly associated SNPs are found as well as their p-values. An example Manhattan plot can be seen in Figure 1.

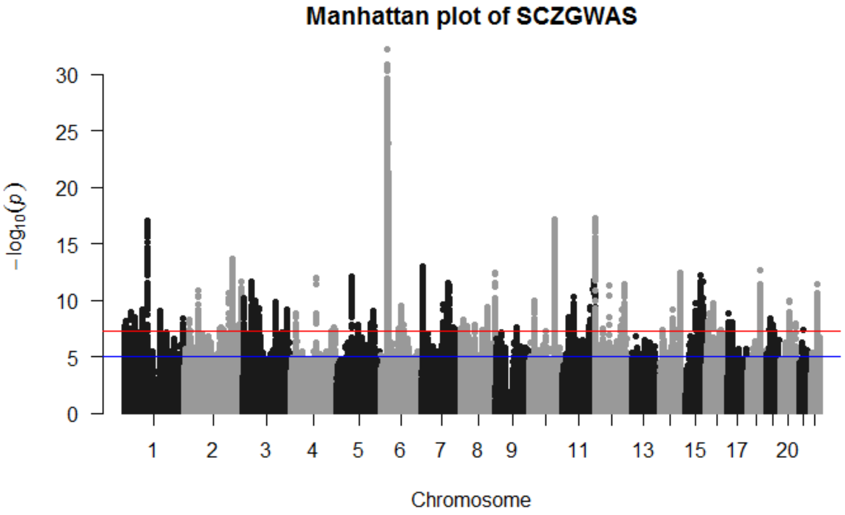

Figure 1. Manhattan plot of SNPs tested for association with schizophrenia using summary statistics obtained from the PGC.
(Schizophrenia Working Group of the Psychiatric Genomics, 2014) The y-axis represents the –log10(p) value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10-5. The red line represents a genome-wide significance level of 5 x 10-8.

If one wants to take a closer look at a particular region of interest, such as a region containing the scan's most significant SNP, a LocusZoom plot is an effective tool (Pruim et al., 2010). In a LocusZoom plot, a SNP is selected and the area surrounding it on the chromosome is displayed. Genes within this region are displayed as well as other genotyped and imputed SNPs and recombination peaks. Additionally, each SNP is color-coded according to its LD with the selected SNP. These features of LocusZoom plots all contribute to their usefulness in interpreting GWAS data. Knowing what genes are found in the region can help one understand why the SNP was found to be associated. The provided LD and recombination rates can also help narrow down what genes in the region are worth examining. For example, SNPs in high LD with the peak SNP are more likely to be causal if they are found in functional or regulatory regions. An example LocusZoom plot can be seen in Figure 2.

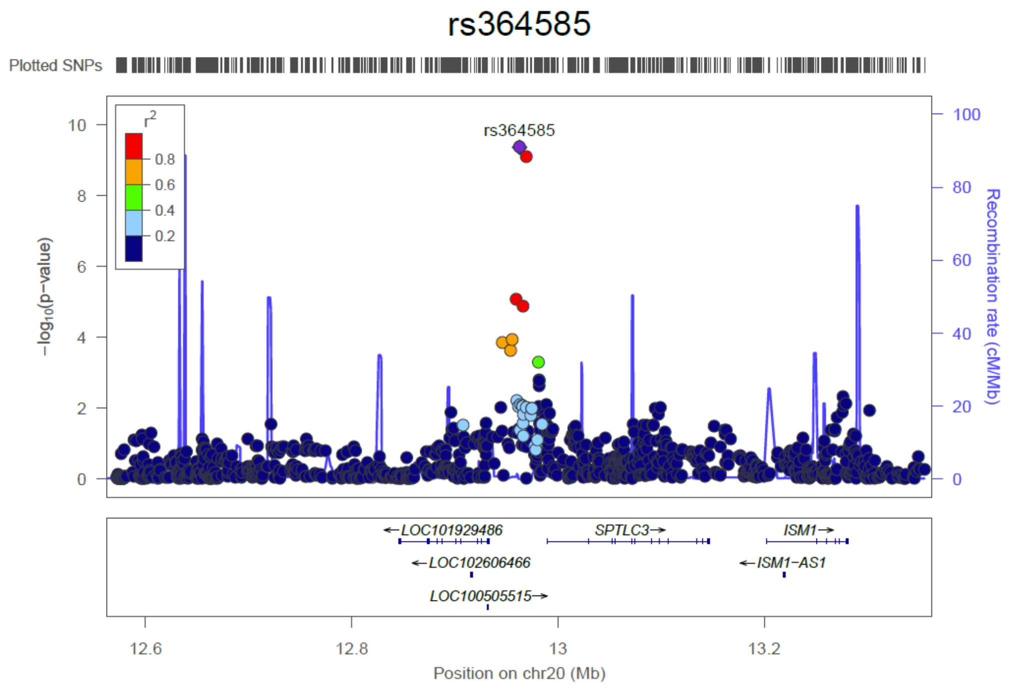

Figure 2. A LocusZoom plot of a 400 kb region surrounding the schizophrenia-associated SNP rs364585 using summary statistics from the PGC.
(Schizophrenia Working Group of the Psychiatric Genomics, 2014) Each dot represents a SNP with its color corresponding to its LD, measured in r^2 to rs364585. The y-axis represents the –$\log_{10}$(p) value of each tested SNP, with the x-axis being location on the chromosome. The lines represent the recombination rate, measured in cMMb, at those loci. Genes found within this region are shown in the bottom box.

Another useful method of visualizing and analyzing GWAS data is the quantile-quantile, or Q-Q, plot. The Q-Q plot can provide information on two main aspects of GWAS data: whether the statistical testing is well-controlled for challenges such as population stratification and whether there is any association. Q-Q plots measure and compare the p-values expected to be seen when testing for association and those actually observed. If there is population stratification that was not adjusted for, the λ statistic will be inflated, indicating genomic inflation due to the genetic relationships between the individuals sampled. When there are significant associations found between markers and the trait of interest, points at the end of the Q-Q plot will rise off the line. Using this information, one can be more certain of the results being accurate when performing further analysis. An example Q-Q plot can be seen in figure 3.

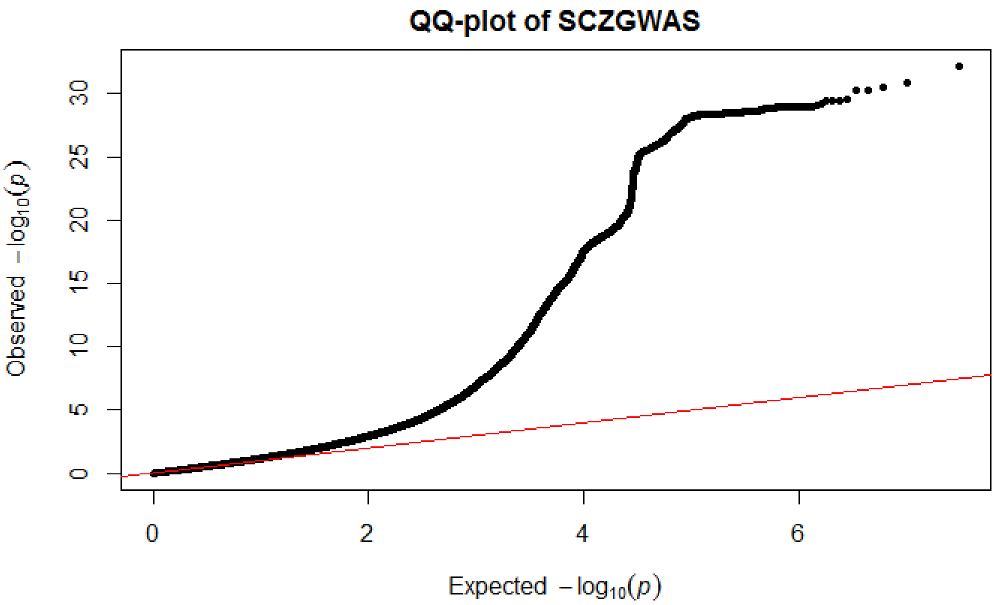

Figure 3. Quantile-quantile (Q-Q) plot for SNPs tested for association with schizophrenia using summary statistics obtained from the Psychiatric Genomics Consortium (PGC).
(Schizophrenia Working Group of the Psychiatric Genomics, 2014) Each dot represents a SNP. The x-axis represents the expected –log$_{10}$(p) value. The y-axis represents the observed -log$_{10}$(p) value. The red line represents the pattern of -log$_{10}$(p) value if no SNP has a significant genetic association with the trait.

2.1.3 Statistical analysis of genome-wide association studies

GWAS is not that different from other genetic association study designs in terms of the methods used but the use of proper statistical methods is necessary for ensuring high statistical power and low rates of false positives to find true genetic associations. Statistical power is the probability of rejecting the null hypothesis of no genetic association when the alternative hypothesis is true, that is, there is a genetic association. How much statistical power a test has depends on various factors. One factor is what kind of genetic model is being assumed for the alleles and genes being tested. If the genetic model differs from how the variants or alleles actually affect the trait of interest, power will be lost. Typically, when performing a GWAS, the additive allele model is used where each additional allele affects the trait of interest, increasing disease risk or protection in an additive manner. However, not every allele is additive so power will be lost if the allele being tested affects genetic variance in a non-additive manner.

Before further continuing into the discussion of GWAS and statistical power, the theory and formulations of power need to be mentioned. As previously stated, statistical power is the probability of rejecting the null hypothesis when the alternative is true. A GWAS can be a case-control study where the outcome has been measured, such as the presence of some trait or disease, and the frequency of an allele is compared between the cases and controls. For quantitative traits such as BMI, case and control status may not be defined and instead the GWAS may follow a cross-sectional design. The significance of the difference between the frequencies can be measured using either a chi-squared test or a trend test and can serve as marking an association between that gene and trait. To determine the effect size of any significant alleles found, regression can be done and any important covariates, such as age, sex, or principal components of ancestry, can be included to take into account their effect on the trait of interest.

The number of markers genotyped can also have an effect on the power of the study. Most SNPs that are genotyped during a GWAS are not causative of the associations found; rather, they are in LD with other SNPs in the same region that may cause the effect that drives the association. When more markers are tested across the genome, there will be larger coverage of the genome and more associated SNPs may be detected. If the coverage is too low due to a lack of SNPs, associated regions could be missed during analysis, reducing the power of the study.

Another factor that can affect the power of a GWAS is the heterogeneity of the trait of interest. Heterogeneity is when there may be different forms, causes, or etiologies for a particular trait or disease. Genetic heterogeneity comes in multiple types such as locus heterogeneity where multiple genes in different regions may play a role in disease development, and allelic heterogeneity where different variants in the same region can have different effects on disease development. Due to its reliance on LD and focus on specific alleles, GWAS is not very robust to allelic heterogeneity. Studied traits may also be affected by environmental factors, so it is important that when one is choosing individuals to form the study's sample that these factors are taken into account to avoid any biases that could occur. Environmental factors and other alleles drive the rate of phenocopies, cases that are not due to the tested variant. If the assumed phenocopy rate differs from the actual phenocopy rate, the study could lose power or these factors could cause bias in the results.

2.2 ANNOTATION OF GENES AND GENOMES

Once results have been obtained and significant associations have been found, determining what the results mean is necessary to understand how the genetic associations seen could affect the trait of interest. Through this process, known as annotation, information about the specific associated regions and variants can be collected and analyzed to make inferences about the effects of variation in the region. However, the results obtained may not be straightforward due to a variety of factors. The markers genotyped may not be directly involved in the etiology of the trait; rather, they may be in LD with the variants that are. If the association is in a region with a high number of genes and high LD with the marker, determining which gene is involved can be difficult as the high LD makes it hard to clearly say what gene may be implicated. An example of this would be the major histocompatibility complex, a region containing many genes involved in immune functions which has very high LD such that identifying what associated variant may be causal in this region is near impossible.

Many GWAS results return variants found in non-coding regions which are hard to evaluate and annotate as it is not always clear how the change affects development of the trait (Tak & Farnham, 2015). It is thought that if these variants are causal, they likely affect the expression and regulation of nearby genes which leads to the development of the trait (Tak & Farnham, 2015). By using a combination of different annotation methods, one can attempt to create a reasonable interpretation of the results.

2.2.1 Literature review of significantly associated genes

Performing a literature review is one of the main methods used to determine the possible roles an associated gene or region may have in the etiology of the trait. If the trait or disease that is being studied has been examined in previous studies, the results from those studies can be compared to one's own results to see if the same or similar associations have been found before. If associations have been found previously, this can serve as a possible validation of results obtained in the current study and could serve as a sign that the associations are accurate. The literature can also provide information on what roles the genes may play such as what cellular and physiological activities they are involved in or what diseases they have been previously implicated in.

2.2.2 Functional and regulatory annotation of genetic variants

There are two main ways of annotating significantly associated genetics variants: identifying changes in function of gene products and identifying changes in expression and regulation of genes. Variants found within coding regions can be much easier to annotate as changing the coding sequence can change the structure, and thus function, of the protein; however, less than 10% of significantly associated SNPs found by GWAS are found within coding regions, and most SNPs that are in LD with the associated marker SNPs are also found in non-coding regions, making it difficult to understand what role they may play in the etiology of the phenotype (Tak & Farnham, 2015). These non-coding regions can contain various regulatory sites where alterations can lead to changes in expression of genes which may have an effect on phenotype. Some of the sites that may be examined are promoters, regions where cellular machinery binds to initiate transcription; DNAse hypersensitivity sites, regions where DNA is more accessible and open, allowing transcription factors to bind; histone modification sites, regions where compounds are added to histone proteins, changing chromatin structure and changing the accessibility of the DNA; transcription factor binding sites, sequences where transcription factors bind to regulate transcription; and CpG islands, regions where cytosines are methylated, often in promoters, and expression is downregulated (Tak & Farnham, 2015). Variants in these sites can lead to changes in expression such as promoters not recognizing cellular machinery, DNAse hypersensitivity sites closing up, transcription factors not being recognized due to motif changes, or CpG islands no longer being methylated. SNPs that are found in regions such as these that cause changes in expression are called expression quantitative trait loci (eQTLs) (Tak & Farnham, 2015).

The goal of expression analysis is to determine how genes are expressed in various tissues and at different times in the body to better understand the role of those genes. By using expression analysis, GWAS associated SNPs can be evaluated to see if their association with a trait has a plausible explanation. For example, one may expect to see a schizophrenia-associated gene expressed in the brain. If this gene is not expressed in the brain, its association with schizophrenia may be a false positive or it may be associated with schizophrenia through an unrelated pathway.

2.3 FALSE POSITIVES AND METHODS TO AVOID THEM

False positives have always been a concern when performing GWAS. Due to the high number of tests performed with millions of loci analyzed, if a traditional significance level of 0.05 is used, there is a high likelihood of seeing many significant results due to chance alone. Because of this, a Bonferroni correction for one million independent loci is typically applied such that to reach significance, the p value must be less than 5×10^{-8}. Using a Bonferroni correction based on the number of loci tested, Ioannidis et al. (2011) showed that ~99% of the many suspected disease loci observed during candidate gene studies were unable to be replicated. Johnson et al. (2017) compared schizophrenia candidate gene loci with those identified by the Psychiatric Genomics Consortium (PGC) and found that the majority of these previously associated loci were not enriched in the more powerful PGC study.

There have been several other methods that have been implemented to reduce the number of false positives. One such method is a two-stage GWAS study design consisting of a “discovery” stage and a “replication” or “validation” stage. In the “discovery” stage, the entire genome is scanned for significantly associated loci. In the “replication” stage, these significantly associated loci are then tested again in a similar population as validation. By using this study design, researchers can narrow down the number of loci they examine, reducing the number of tests performed, increasing the power to detect associated variants, and confirm if these variants were associated due to chance or not. However, many loci fail to replicate in these validation stages. This can be partially attributed to the “winner’s curse”. The winner’s curse is a phenomenon where thresholding, such as reaching a particular significance level, leads to inflated effect sizes and biased results near the threshold (Ioannidis, Thomas, & Daly, 2009; Palmer & Pe’er, 2017). Palmer & Pe’er (2017) showed that employing statistical corrections of the winner’s curse improved the replication rates of loci significantly associated at the discovery stage. They also found that loci did not replicate well if the replication sample had different continental ancestry or used an overall as opposed to a per-locus sample size, which would be expected as allele frequencies may differ between populations, leading to potential bias, and using total sample size would exaggerate power (Palmer & Pe’er, 2017).

3.0 METHODS

3.1 SELECTION OF TRAITS

As we were interested in whether annotators could find convincing connections between synthetic peaks and scanned GWAS traits, we selected traits that were not genetically correlated with each other. Genetic correlation was defined as genetic covariance normalized by SNP heritabilities (Bulik-Sullivan et al., 2015). Using genetic correlation data from Bulik-Sullivan et al. (2015), we created genetically uncorrelated trait pairs with LD score regression correlation scores <0.10, producing the following pairs: schizophrenia (SCZ) and low-density lipoprotein (LDL); triglycerides (TG) and Crohn's disease (CD); high-density lipoprotein (HDL) and ulcerative colitis (UC); and coronary artery disease (CAD) and Alzheimer's disease (AD). We obtained GWAS summary statistics for these traits from the Psychiatric Genomics Consortium (PGC) for SCZ, the Global Lipid Genetics Consortium (GLGC) for HDL, LDL, and TG, the Coronary Artery Disease (C4D) Genetics Consortium for CAD, the International IBD Genetics Consortium (IIBDGC) for UC and CD, and the International Genomics of Alzheimer's Project (IGAP) for AD (Lambert et al., 2013; Liu et al., 2015; Schizophrenia Working Group of the Psychiatric Genomics, 2014; The Coronary Artery Disease Genetics, 2011; Willer et al., 2013). We randomly selected which trait would be used as the scan trait and which trait would provide synthetic peaks. In doing so, LDL, CD, and UC were chosen to provide these synthetic peaks. To examine if trait selection in each pair made a difference in what genes and regions were selected as having convincing connections, CAD and AD were chosen to provide each other's respective synthetic peaks, with 2 annotators annotating for (CAD, AD) and 1 annotating for (AD, CAD), where the first trait provided original peaks and the second trait provided synthetic peaks. Our hypothesis is that annotators would still make convincing connections between synthetic peaks and the scanned trait but the genes and regions they identify as noteworthy may be different for CAD and AD.

3.2 DATA CLEANING AND PEAK SELECTION

We acquired summary statistics from the consortium websites and selected peaks with at least 500 kb separating each locus. Three peak significance categories were established: (1) "highly significant" for peaks with a p-value less than 1 x 10^{-15}, (2) "moderately significant" for peaks with a p-value between 5 x 10^{-8} and 1 x 10^{-15}, and (3) "suggestively significant" for peaks with a p-value between 1 x 10^{-5} and 5 x 10^{-8}. We excluded any peak SNPs with an imputation INFO score less than 0.90, and we excluded peak SNPs that were indel variants. We excluded peaks with minor allele frequency less than 0.05 except for SCZ where a cutoff of 0.10 was used instead to match the methods used by the PGC (Schizophrenia Working Group of the Psychiatric Genomics, 2014). For "highly significant" and "moderately significant" peaks, we excluded peaks that were not reported in the original studies. We did not do this for "suggestively significant" peaks because it is quite likely that suggestively significant peaks were not highlighted in the original studies. For each trait pair, nine peaks were selected, with one of each type of significance for the scanned trait and two for each type of significance for the synthetic peaks, excluding AD which had eight due to a lack of “moderately significant” CAD peaks to be used as synthetic peaks. Data cleaning and analysis was performed in R (R Core Team, 2017). We created regional association plots of each peak using LocusZoom with windows of 400 kb and using the European hg19 build from the 1000 Genomes Project to provide LD information (Pruim et al., 2010).

3.3 LITERATURE REVIEW AND ANNOTATION OF SIGNIFICANT HITS

We performed automated literature review searches of PubMed, PubMed Central, and Google Scholar using the R packages “RISmed”, “rvest”, and “data.table” (Dowle, 2017; Kovalchik, 2017; Wickham, 2016). Up to 1000, 20, and 10 results, respectively from each source, were obtained and recorded in spreadsheets. For each peak, our search query was the name of the peak SNP and the scanned trait; for example, “rs4393438 AND schizophrenia”. We also queried the name of genes in close proximity to the SNP or in LD with the SNP and the scanned trait; for example, “RASA3 AND schizophrenia”.

We recruited three volunteers to act as annotators. Two annotators were human genetics students and one was an epidemiology student obtaining a certificate in human genetics. Additionally, I annotated 4 scans and Dr. Weeks annotated 2 scans. What traits each participant annotated can be seen in Table 1.

Table 1. Traits annotated by each participant, with "+" indicating they annotated this trait, and "-" indicating they did not.

	Scanned Trait (Original, Synthetic)				
Annotator	(SCZ, LDL)	(TG, CD)	(HDL, UC)	(CAD, AD)	(AD, CAD)
1	+	+	+	-	+
2	+	+	-	-	-
3	+	+	+	+	-
4	+	+	+	+	-
5	+	+	-	-	-

Annotators were asked to use the provided literature search results and LocusZoom plots to annotate each peak within a scan as having convincing connections with the scanned trait or not and record the results within provided summary sheets. Annotators were free to carry out additional literature and database searches on their own. The provided summary sheets stated the scanned trait and were divided up into multiple pages containing LocusZoom plots of each selected peak region centered on the most significant SNP as well as space for their paragraph used to describe the results they found. The instructions provided to the annotators can be seen in Appendix B. Three annotators were also asked to rate the strength of evidence for association of a particular peak on a scale of 0 to 3 with 0 being no evidence and 3 being very strong evidence. They were also asked to rate the likelihood of further study of those particular peaks on a scale of 0 to 3 with 0 being no likelihood and 3 being a very high likelihood.

3.4 ANALYSIS OF ANNOTATION

Using the summary sheets filled out by the annotators, we recorded whether the annotator found a peak to have convincing connections with the scanned trait. Peaks that annotators found to have convincing connections were stated to be positively annotated. We also recorded what genes they noted as being of interest, whether the genes were selected due to expression or function, what cellular functions and pathways were of interest, how many citations they used, and, for annotators that were asked, strength of evidence and likelihood of further study of the peak.

To determine if there was an association between a peak being annotated as having convincing connections with the scanned trait and original/synthetic peak status, we performed Fisher's exact test for each individual annotator and Pearson's chi-squared test for all annotators combined. We did the same for peak significance category. We then performed logistic regression, regressing positive annotation status over peak significance categories.

We performed chi-squared proportion tests of original/synthetic peak status and peak annotation status with a null distribution assuming perfect annotation status with all synthetic peaks having no convincing connections found and all original peaks having convincing connections found.

To determine if there was an association between peak significance category and strength of evidence of association and likelihood of further study, we performed the Kruskal-Wallis test for the three annotators who were asked to answer this and all three of these annotators combined. We also performed Wilcoxon rank-sum tests for original/synthetic peak type and strength of evidence and likelihood of further study. We then performed polychotomous multinomial logistic regression to assess the effect of peak significance category and original/synthetic peak status on the strength of evidence and the likelihood of further study. For peak significance category, “highly significant” and a rating of 0 were used as reference. For original/synthetic peak status, synthetic status and a rating of 0 were used as reference.

4.0 RESULTS

Five participants were asked to use provided literature search results as well as any other desired sources to annotate whether genes at associated loci identified through GWAS had convincing connections to the scanned trait. The traits each participant annotated can be seen in Table 1. Each scan consisted of nine peaks, excluding AD which had eight, divided into three peak significance categories: suggestively significant, moderately significant, or highly significant. Each category contained three peaks: one peak using summary statistics from a consortium studying the scanned trait, and two peaks using summary statistics from a consortium studying a genetically uncorrelated trait, excluding AD which had only one moderately significant synthetic peak in its scan.

How participants annotated each peak can be seen in Figures 4-8. The number of peaks where convincing connections were found by peak significance category and peak type can be seen in Table 2.

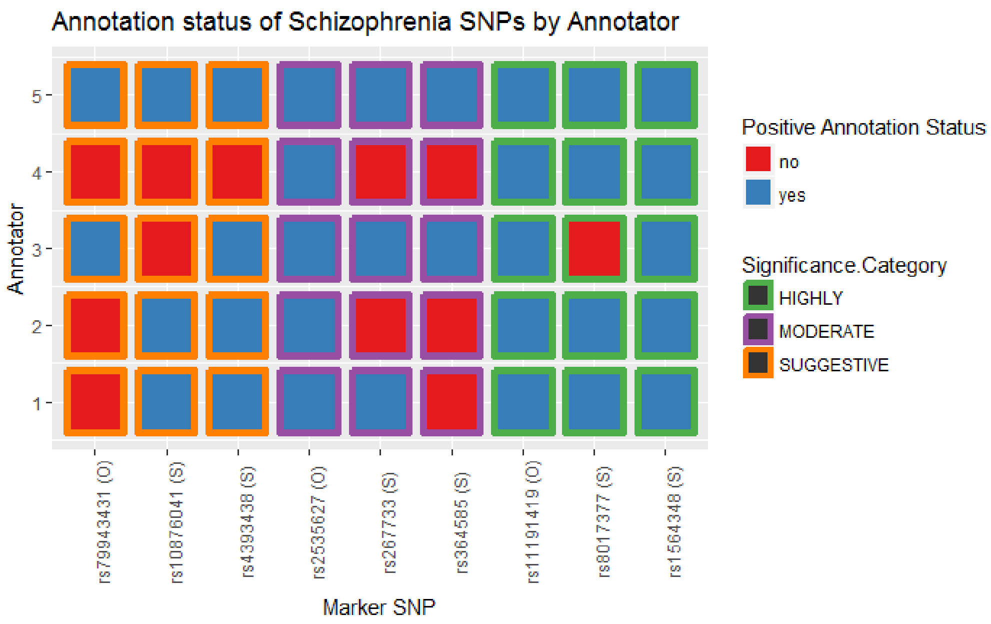

Figure 4. Annotation status of 9 SNPs for showing evidence for schizophrenia association divided by annotator (y axis), peak significance category (border color), and whether the peak was from the schizophrenia GWAS (O) or LDL GWAS (S).

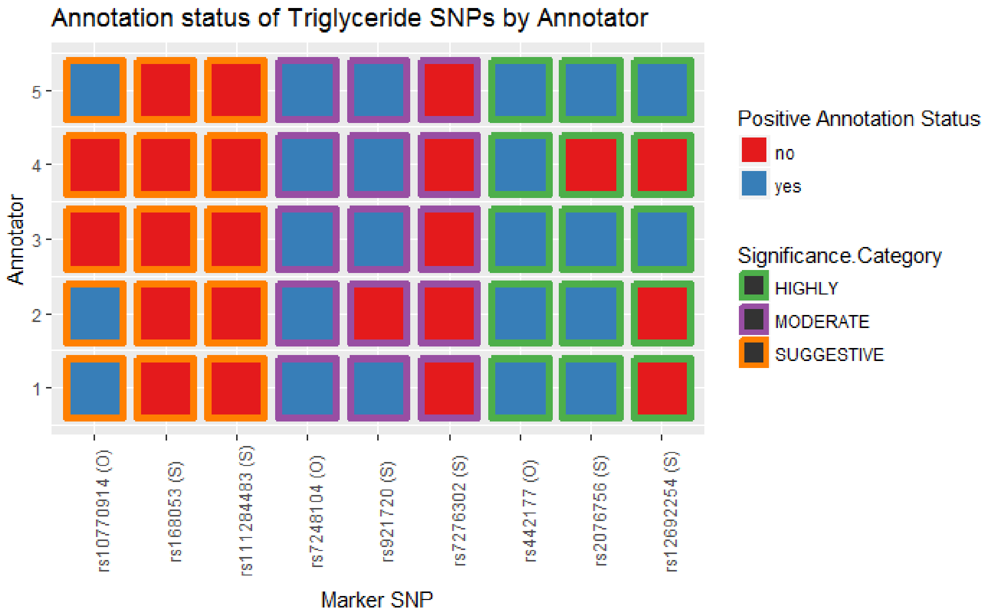

Figure 5. Annotation status of 9 SNPs for showing evidence for triglyceride association divided by annotator (y axis), peak significance category (border color), and whether the peak was from the triglycerides GWAS (O) or Crohn’s disease GWAS (S).

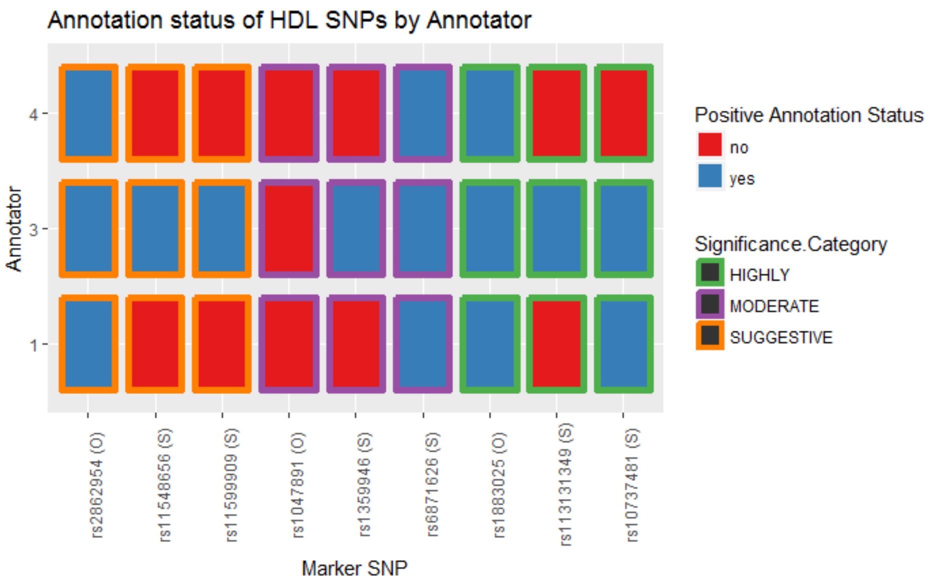

Figure 6. Annotation status of 9 SNPs for showing evidence for triglyceride association divided by annotator (y axis), peak significance category (border color), and whether the peak was from the HDL GWAS (O) or ulcerative colitis GWAS (S).

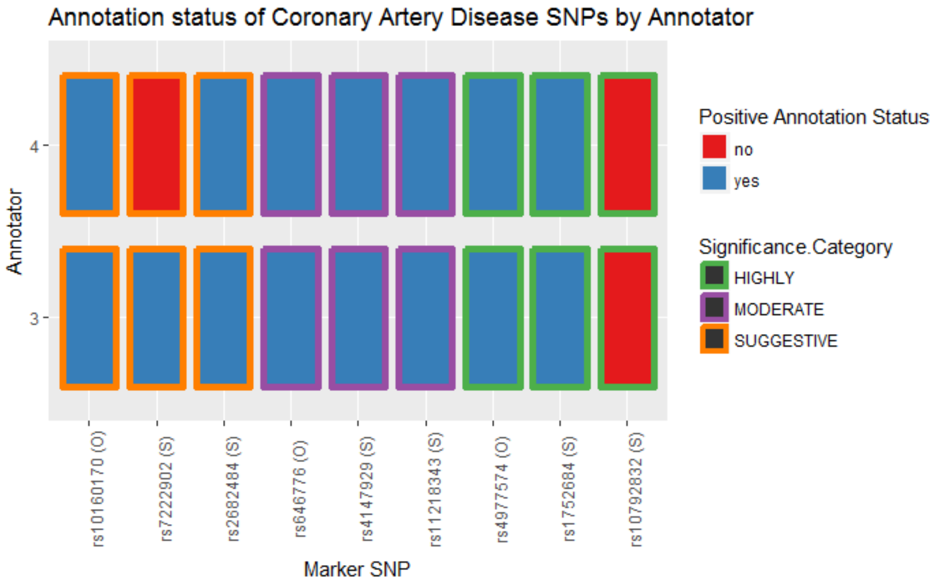

Figure 7. Annotation status of 9 SNPs for showing evidence for coronary artery disease association divided by annotator (y axis), peak significance category (border color), and whether the peak was from the coronary artery disease GWAS (O) or Alzheimer's disease (S).

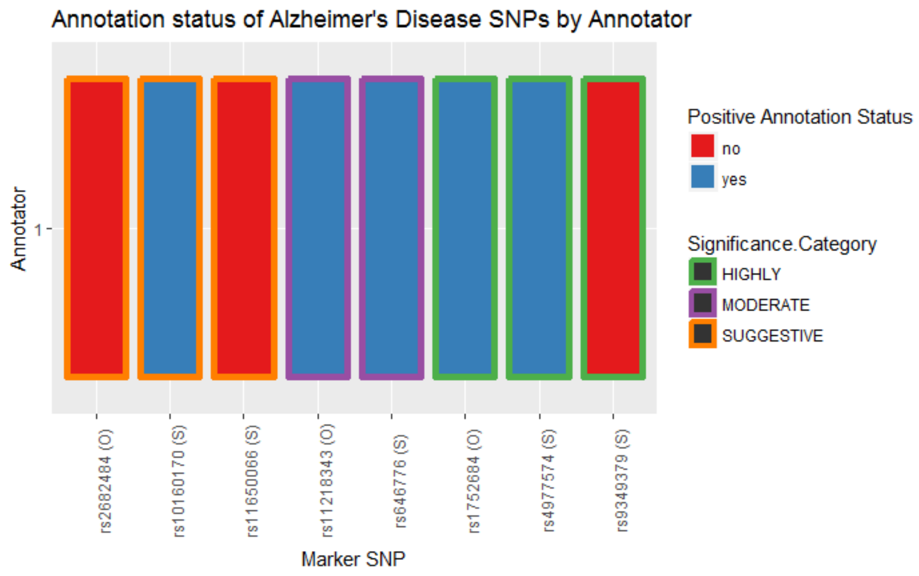

Figure 8. Annotation status of 9 SNPs for showing evidence for Alzheimer's disease association by annotator 1 (y axis), peak significance category (border color), and whether the peak was from the Alzheimer's disease GWAS (O) or coronary artery disease GWAS (S).

Table 2. Positive annotation status by peak significance category and by peak type for all annotators.

	Positive Annotation Status	
Peak Significance Category	**NO**	**YES**
HIGHLY	11 (22.9)	37 (77.1)
MODERATE	16 (34.0)	31 (66.0)
SUGGESTIVE	25 (52.1)	23 (47.9)
	Positive Annotation Status	
Peak Type	**NO**	**YES**
SYNTHETIC	43 (45.3)	52 (54.7)
ORIGINAL	9 (18.75)	39 (81.25)

Annotators 1, 4, and 5 were asked to answer how strong they felt the evidence for association was at each peak as well as the likelihood of continuing to study that region for its role in the scanned trait. The counts and proportions of their responses by peak type and peak significance category can be seen in Figures 9-12.

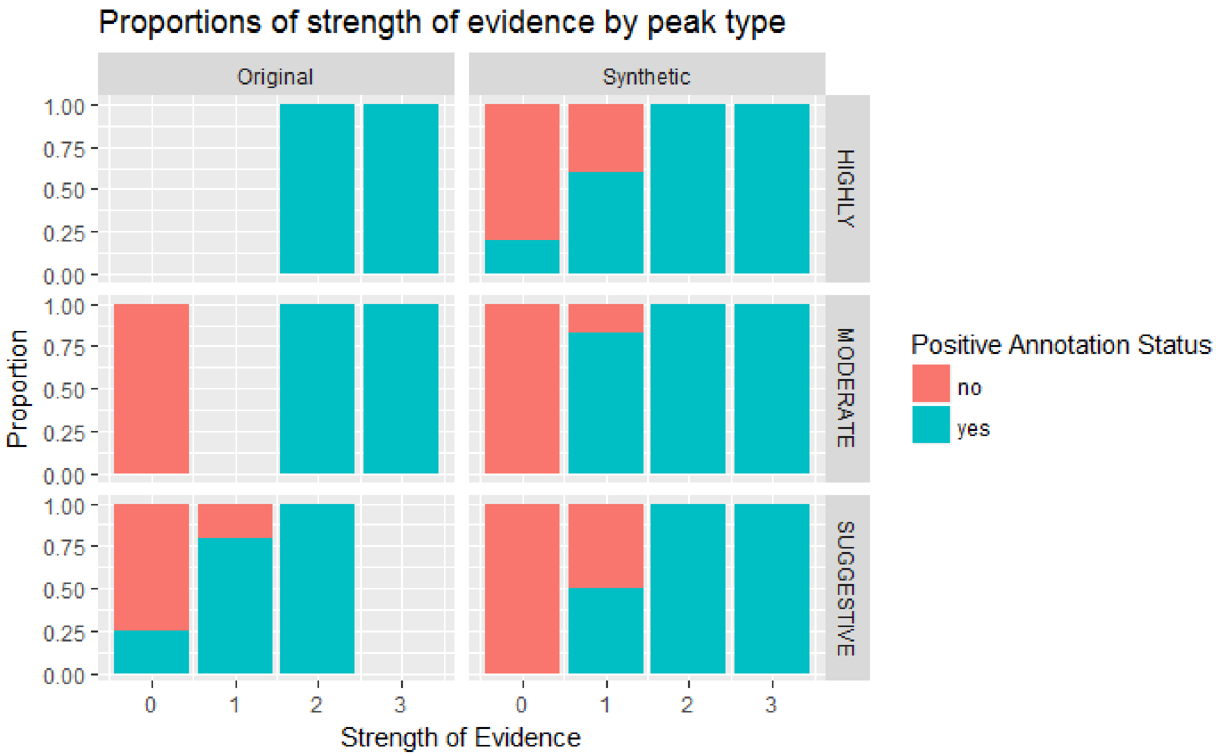

Figure 9. Proportions of strength of evidence scores from annotators 1, 4, and 5 by original or synthetic peak, peak significance category, and positive annotation status. Strength of evidence was rated on a scale of 0 to 3 with 0 being no evidence and 3 being very strong evidence.

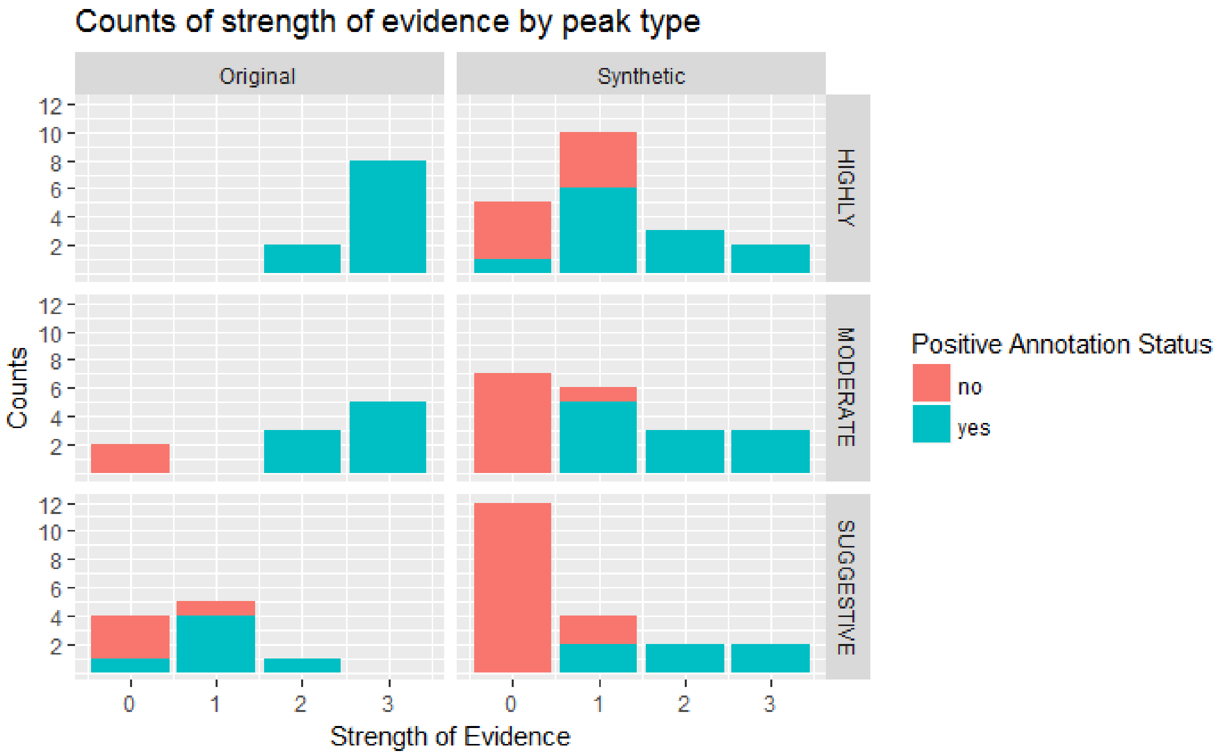

Figure 10. Counts of strength of evidence scores from annotators 1, 4, and 5 by original or synthetic peak, peak significance category, and positive annotation status. Strength of evidence was rated on a scale of 0 to 3 with 0 being no evidence and 3 being very strong evidence.

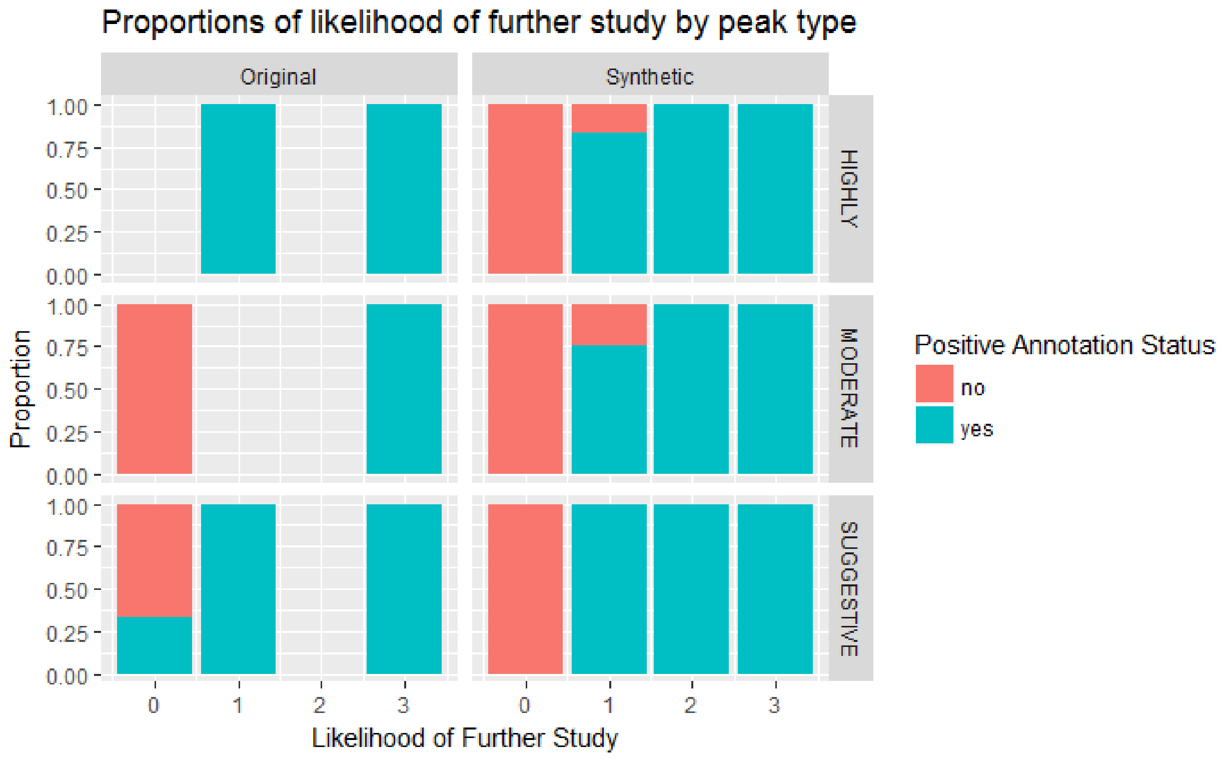

Figure 11. Proportion of likelihood of further study scores from annotators 1, 4, and 5 by original or synthetic peak, significance category, and positive annotation status. Likelihood of further study was rated on a scale of 0 to 3 with 0 being no likelihood and 3 being a very high likelihood.

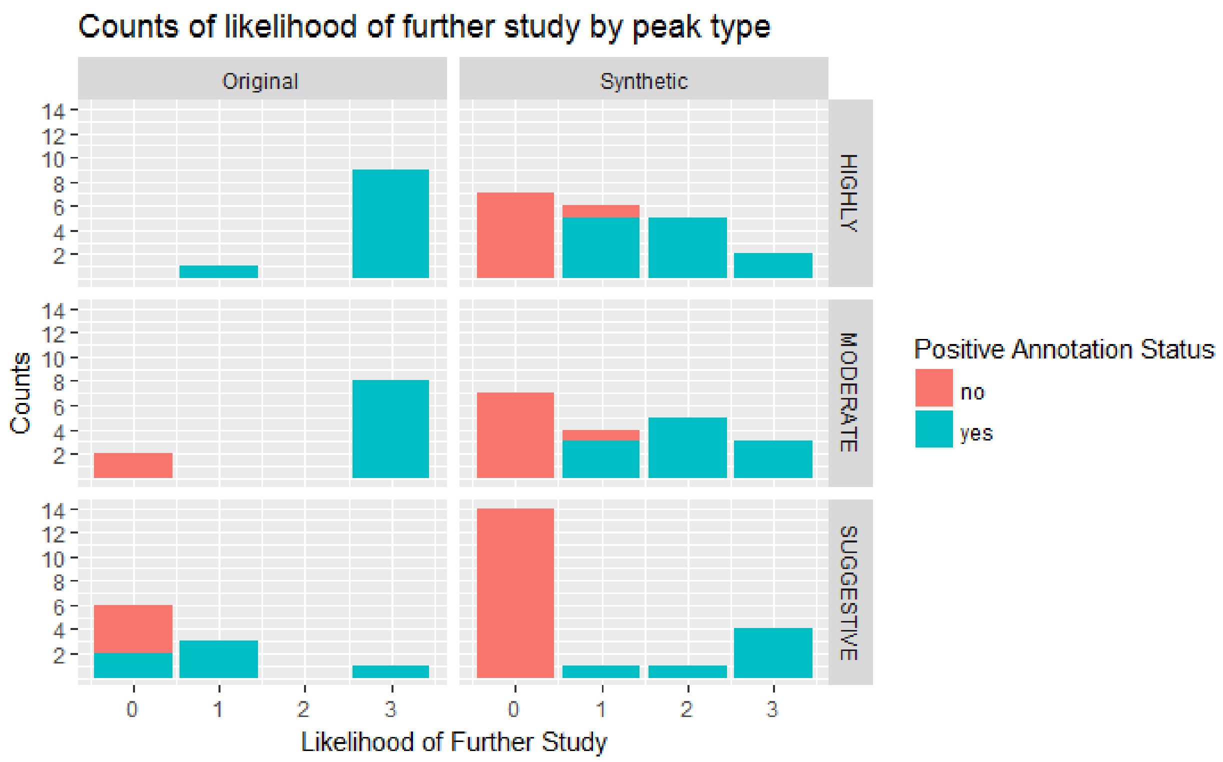

Figure 12. Counts of likelihood of further study scores from annotators 1, 4, and 5 by original or synthetic peak, significance category, and positive annotation status. Likelihood of further study was rated on a scale of 0 to 3 with 0 being no likelihood and 3 being a very high likelihood.

To assess if peak significance category or original/synthetic peak status were associated with positive annotation status, we performed Pearson's chi-squared tests of independence. We found that positive annotation status was significantly associated with original/synthetic peak status (χ^2 = 8.575, df = 1, p = 0.0034). We also found that positive annotation status was significantly associated with peak significance category (χ^2 = 8.986, df = 2, p = 0.0112). Using logistic regression, we found that highly significant peaks had a significantly greater number of convincing connections made by annotators compared to suggestively significant peaks (β = -1.296, z = -2.889, p = 0.003869) and no significant difference in number of convincing connections made compared to moderately significant peaks (β = -0.5516, z = -1.196, p = 0.231659). We performed Fisher's exact tests of independence of positive annotation status and original/synthetic peak type for each annotator but none was significant, possibly from a lack of power due to low sample sizes.

We performed chi-squared proportion tests to see if the number of convincing connections made between peaks and the scanned trait was significantly different compared to if annotators were perfect, i.e could find no convincing connections between synthetic peaks and the scanned trait and found convincing connections between all original peaks and the scanned trait. Annotators found original peaks to have a significantly smaller proportion of convincing connections compared to a perfect annotator ($\chi^2 = 7.847$, df = 1, p = 0.002546); the proportion of original peaks found to have convincing connections was not significantly different from a flawed annotation distribution with one false negative annotation for each scanned trait ($\chi^2 = 1.947$, df = 1, p = 0.1629). They also found synthetic peaks to have a significantly larger proportion of convincing connections compared to a perfect annotator ($\chi^2 = 68.867$, df = 1, $p < 2.2 \times 10^{-16}$); the proportion of synthetic peaks found to have convincing corrections was not significantly different from a flawed distribution with two false positive annotations for each scanned trait after corrections for multiple testing using a Bonferroni correction for four tests ($\chi^2 = 4.763$, df = 1, p = 0.02908, $p_{Bonf} = 0.11632$).

To assess if peak type was associated with the strength of evidence and likelihood of further study, we performed Wilcoxon rank-sum tests. We found that there was a significant difference between original and synthetic peaks for both strength of evidence (U = 518.5, $p < 0.001$) and likelihood of further study (U = 543, p = 0.001775). We performed Kruskal-Wallis tests to determine if peak significance category was associated with strength of evidence (H = 4.709, df = 2, p = 0.09493) and likelihood of further study (H = 6.133, df = 2, p = 0.04659, $p_{Bonf} = 0.09318$), but neither was statistically significant after corrections for multiple testing.

We then performed polychotomous multinomial logistic regression to evaluate the effects of peak significance category and original/synthetic peak type on strength of evidence and likelihood of further study. Annotators also found original peaks to have significantly stronger evidence (β = 2.005, SE = 0.654, t = 3.065, p = 0.0029, p_{Bonf} = 0.0348) and would be significantly more likely to continue studying original peaks compared to synthetic peaks (β = 1.946, SE = 0.572, t = 3.401, p = 0.001016, p_{Bonf} = 0.012192). To adjust for multiple testing, we used a Bonferroni correction for 6 tests. Annotators found highly significant peaks to have significantly stronger evidence than suggestively significant peaks (β = -2.773, SE = 0.929 t = -2.985, df = 79, p = .003677, p_{Bonf} = 0.0221). They also were more likely to recommend highly significant peaks for further study than suggestively significant peaks, but none of the tests was significant after corrections for multiple testing (p > 0.05).

5.0 DISCUSSION

As far as we are aware, this is the first study to analyze annotators' ability to interpret GWAS synthetic peaks. In this study, we have shown that, given literature search results and a mixture of original and synthetic peaks for a scanned trait, annotators are able to make a large number of convincing connections between synthetic peaks and the scanned trait. Annotators were also found to have a higher likelihood of further studying original peaks as well as finding original peaks as having greater strength of evidence compared to synthetic peaks, as expected.

As part of our analysis, we performed several different statistical tests to see how annotators' ability to find a convincing connection between association peaks and scanned traits is affected. After performing chi-squared tests for independence of positive annotation status with peak significance category and with original/synthetic peak status, we found that annotation status was not independent of these two characteristics. This is as we expected as one would assume that more significantly associated peaks would have been studied further and thus have more evidence for annotators to find within the literature. One would also make the fair assumption that peaks that have previously been found to be significantly associated with a trait should be annotated as having convincing connections more often than peaks that have not been found to be significantly associated.

As a way to assess the effectiveness of the participants' annotations, we compared the proportion of annotation status of original and synthetic peaks to the proportion expected under scenarios of perfect annotation as well as flawed annotation and found that the observed proportion of annotations was not significantly different to annotation with two false positives and one false negative. If these synthetic peaks are truly unassociated with the scanned trait, one would expect to see most, if not all, of these peaks annotated as not having evidence explaining the association. One would also expect that if the original peaks are all truly associated with the scanned trait, these peaks would be annotated as having evidence explaining the association. However, these annotators were unable to accurately annotate these peaks, showing that human annotation is not robust to false positive results.

We analyzed the effects of peak significance category and original/synthetic peak status on how strong annotators viewed the evidence and the likelihood that they would further study those peaks for their roles in development of the phenotype. We found that original peaks had significantly different strengths of evidence and likelihood of further study compared to synthetic peaks, with original peaks having significantly more counts of having very strong evidence compared to synthetic peaks (Figures 9-12). This makes sense as one would expect original peaks to have more evidence connecting them to the scanned trait within the literature than the synthetic peaks. Polychotomous multinomial logistic regression showed significant differences between highly significant peaks and suggestively significant peaks for no strength of evidence and very strong evidence. Again, this is as we would expect as suggestively significant peaks are more likely to be false positives than highly significant peaks and thus have less evidence supporting their connections to the scanned trait within the literature. As the suggestively significant peaks do not reach a genome-wide significance level, it is fair to say that they aren't associated at all.

We were interested in whether the selection of the trait was that provided the original peaks affected what the genes annotators found to be of interest. To do this, we compared the annotation of the (CAD, AD) set to that of the (AD, CAD) set, examining whether the genes identified by the annotators were different or not between these reciprocal sets. Of the seven peaks that overlapped between the two sets, annotator 1 found genes with supporting evidence for AD that matched the genes annotators 3 and 4 found for CAD. These SNP regions and their corresponding genes were rs10160170 (*CXCL12*), rs11218343 (*SORL1*), rs646776 (*SORT1*), rs1752684 (*CR1*), and rs4977574 (*CDKN2A*). There are multiple possible explanations for why this occurred. One is that the sample size was too small, having two annotators for CAD and one for AD, and that these matches may not necessarily be seen with a larger sample. Another possible explanation is that CAD and AD are not as genetically uncorrelated as proposed by Bulik-Sullivan et al. which could explain the matches seen at these peaks (Bulik-Sullivan et al., 2015). There is also the possibility that CAD and AD are as uncorrelated as proposed but the subset of the peaks tested here is correlated.

Many current and future GWAS that are being designed are very high-powered, with some having millions of sampled individuals allowing for detection of significant associations of exceedingly small effect sizes. Annotators' ability to find convincing connections for synthetic peaks may still be relevant in these high-powered studies. With such high power, it will not be difficult to find significantly associated variants with effect sizes so small that their clinical relevance may be questionable. For example, if a variant is found to be significantly associated with a trait with an odds ratio of 1.05, how will annotators interpret it? If they are not conservative in their analysis, annotators could spend too much time and resources following up variants that, by themselves, do not have enough of an effect to be interesting.

5.1 LIMITATIONS OF THE STUDY

There were multiple limitations of the study design which could have reduced our power and are worth discussing. The first and largest limitation was the small number of participating annotators: there were only 5 participants and only 3 of those participants annotated 4 total scans. While it would be better to have larger sample sizes, we were still able to obtain some significant results despite this limitation.

Another limitation was the difference in backgrounds of each annotator. Of the five annotators, one was an epidemiology student, three were human genetics students, and one was a professor. Because of this difference in backgrounds, annotation strategies likely varied leading to differences in annotation. However, because of the lack of power due to the small number of peaks each annotator tested, it is not possible to verify this in this study.

This study was also limited in terms of how annotation was performed and what was asked of annotators. Annotators were asked to determine whether a peak showed plausible evidence explaining the association with the scanned trait using provided literature search results. Literature searches are only one part of the annotation process, however. Other annotation methods, such as examining the regulation of annotated genes and the effects of variants within them, were not required of the annotators. As such, the annotation performed may not be an accurate facsimile of typical annotation. Additionally, instructions provided to the annotators could have been more clear and precise in what was asked of them. For example, additional questions asking which genes they found to be of interest, if any, for their connections to the scanned trait would have aided in analyzing their responses. However, in this study, determining whether a convincing connection was made could be difficult. Some annotators were explicit in whether they found a peak to be convincing/worthy of follow-up while others were vague in their writing such that explicitly stating whether they found a convincing connection could be difficult. Because of this, there is the possibility of errors and inaccurate analysis having occurred when the data was recorded.

There was also a limitation of what genes were selected for annotation. To allow for enough time for participants to complete annotation, only genes in close proximity to the peak SNP or in LD with the peak SNP were selected for literature searches. Linkage disequilibrium is not constant between populations so which genes would be selected could vary if summary statistics from studies of different populations were selected. While the populations examined in this study were primarily of white European descent, future studies that used a similar design could see different results depending on gene selection.

Literature search results could have also limited the effectiveness of annotation. There were only two types of search queries used: searches for the SNP and the trait and searches for a particular gene and the trait. However, typical annotators may use other search terms in their queries which could give them additional or different information that could affect their interpretations of the gene and region. Additionally, genes were queried using their HUGO gene nomenclature, but because some genes may have different names, information about the same genes using different names could have been lost, possibly affecting the annotators' interpretations.

For the peaks that were selected, we made assumptions about the true associations of each peak. Specifically, we made the assumptions that the synthetic peaks from the genetically uncorrelated trait are not associated with the scanned trait and that the original peaks from the scanned trait are actually associated with the trait. However, there is the possibility that the synthetic peaks selected may in fact be associated with the trait but have not been detected yet or were detected in a study following the consortia whose summary statistics we used. There is also the possibility that the original peaks we selected were false positives and are not actually associated with the scanned trait. Due to the inter-connected nature of biology, we cannot be certain of this. This is particularly important for suggestively significant peaks as they did not reach the typical Bonferroni significance threshold and it is quite possible these peaks simply arose due to chance.

5.2 CONCLUSIONS

To analyze the ability of annotators to find convincing connections between false positive results and the scanned trait of a GWAS, we created synthetic peaks using summary statistics from GWAS of genetically uncorrelated traits and asked annotators to annotate a mix of original peaks and these synthetic peaks. We performed statistical analyses and found that annotators made a larger proportion of convincing connections between synthetic peaks and scanned traits compared to if they annotated them all as false; annotators also made a smaller proportion of convincing connections between original peaks and scanned traits than would be expected if they were able to annotate them all accurately. Annotators also found original peaks to have significantly stronger evidence than synthetic peaks and a significantly increased likelihood of further study. These results show that annotators are easily able to make convincing connections between synthetic peaks and the scanned trait of a GWAS.

5.3 SIGNIFICANCE

GWAS are some of the most commonly done genetic studies in the field. In 2012, it was estimated that ~$250 million had been spent on research performing GWAS up to that point (Peter M. Visscher, Brown, McCarthy, & Yang, 2012). Since then, GWAS has become more widespread with many more studies published. With the large amount of resources invested into GWAS, it is important that studies are performed with the utmost accuracy and care. If improper study procedures are used, there is the possibility of false positives and false negatives which could lead to follow-up or lack thereof for the wrong variants and genes, leading to wasted resources. To ensure that the genetic causes of traits and diseases are properly understood so that we may be better able to treat and manage them, we need accurate results to do so.

This study shows the ease with which annotators are able to make convincing connections between synthetic peaks and scanned traits in GWAS. To ensure that public resources are used as efficiently as possible, this issue needs to be recognized so researchers can work to perform studies effectively without bias or error.

5.4 FUTURE RESEARCH

There are several different ways that this study could be expanded upon and improved in future work. The current study was limited in the number of participants who acted as annotators. As such, the recruitment of additional annotators would help greatly in improving the study's power and allow for additional analytical methods.

Future studies would also gain from more strict and clear annotation rules and requirements for the participants, such as additional questions to evaluate the participant's annotation process and improve consistency between annotators. Additional requirements beyond literature review such as the use of annotation databases like RegulomeDB and FUMA would also allow for better simulations of the typical annotation process (Boyle et al., 2012; Watanabe, Taskesen, van Bochoven, & Posthuma, 2017).

This current study examined only genes within LD or in close proximity of the associated peak; however, due to differing LD structure between populations, different genes may appear to be associated in different populations. As such, although there would be an increase in time to complete annotation, future researchers may find it wise to examine all genes in each significantly associated region to take these differences into account.

Peaks that were selected for this study were chosen based on their associations, or lack thereof, within the consortia whose summary statistics we used. However, if studies that took place after these consortia found associations in synthetic peak genes with the scanned trait, their status as a "false" peak would no longer be valid. To avoid this, as part of peak selection, possible peaks and the genes found within them can be searched in the NHGRI-EBI GWAS Catalog, and if they have been found to be associated with the scanned trait, they can be designated as ineligible to be synthetic peaks (MacArthur et al., 2017).

APPENDIX A: ADDITIONAL FIGURES AND TABLES

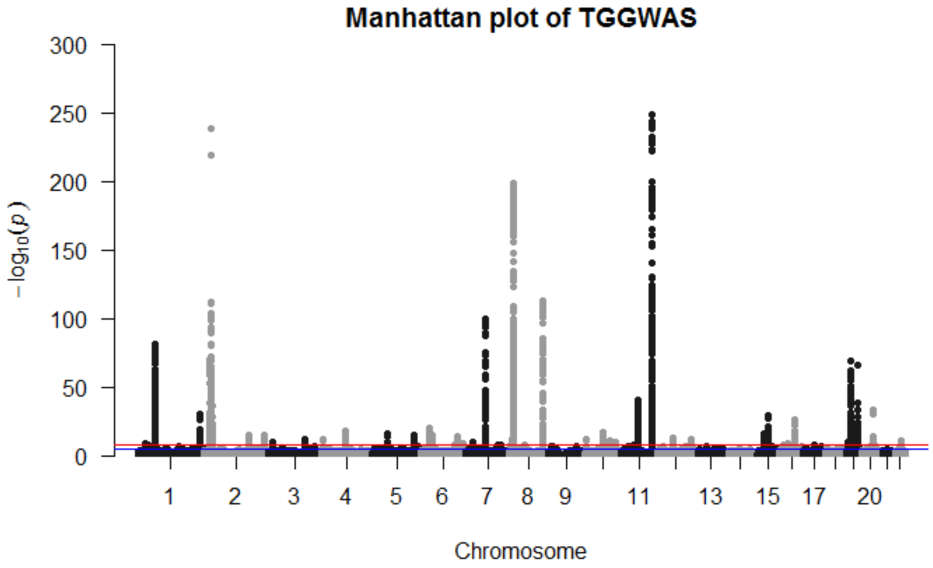

Figure 13. Manhattan plot of SNPs tested for association with triglycerides using summary statistics obtained from the GLGC.

(Willer et al., 2013) The y-axis represents the –log$_{10}$(p) value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

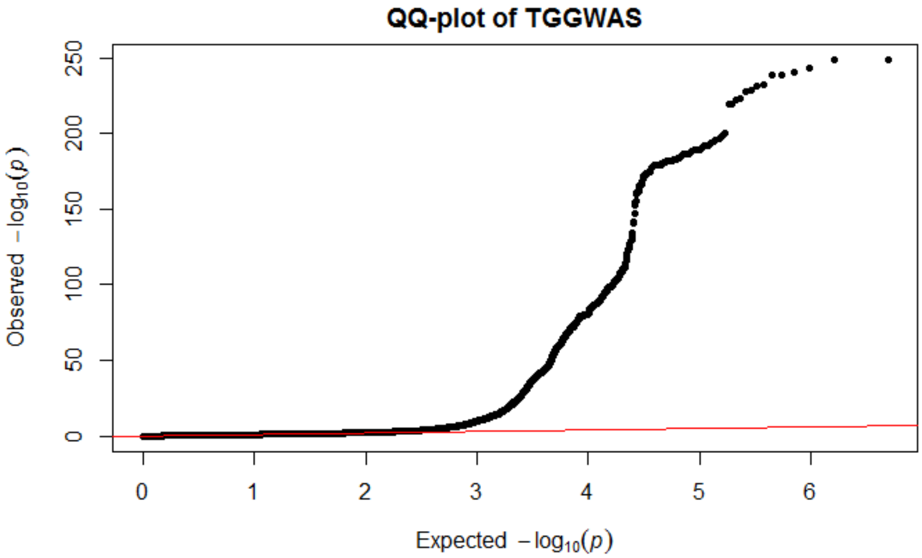

Figure 14. Quantile-quantile (Q-Q) plot for SNPs tested for association with triglycerides using summary statistics obtained from the GLGC.

(Willer et al., 2013) Each dot represents a SNP. The x-axis represents the expected –$\log_{10}$(p) value. The y-axis represents the observed -$\log_{10}$(p) value. The red line represents the pattern of -$\log_{10}$(p) value if no SNP has a significant genetic association with the trait.

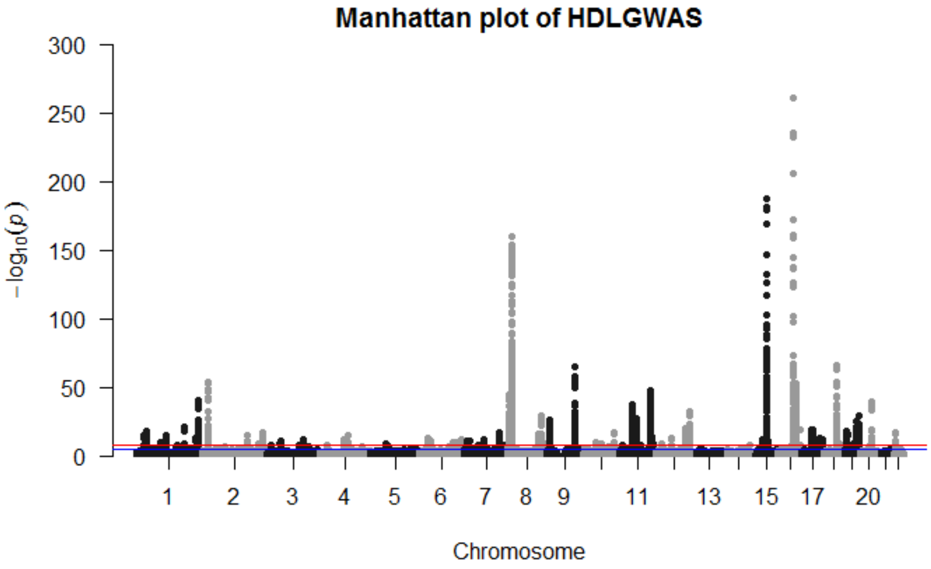

Figure 15. Manhattan plot of SNPs tested for association with HDL using summary statistics obtained from the GLGC.

(Willer et al., 2013) The y-axis represents the –$\log_{10}$(p) value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

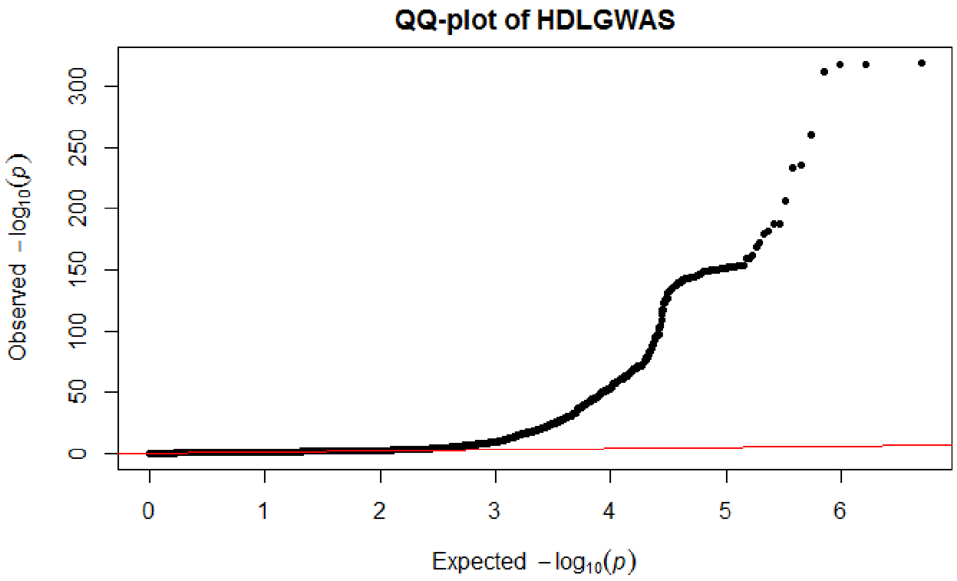

Figure 16. Quantile-quantile (Q-Q) plot for SNPs tested for association with HDL using summary statistics obtained from the GLGC.
(Willer et al., 2013) Each dot represents a SNP. The x-axis represents the expected –log$_{10}$(p) value. The y-axis represents the observed -log$_{10}$(p) value. The red line represents the pattern of -log$_{10}$(p) value if no SNP has a significant genetic association with the trait.

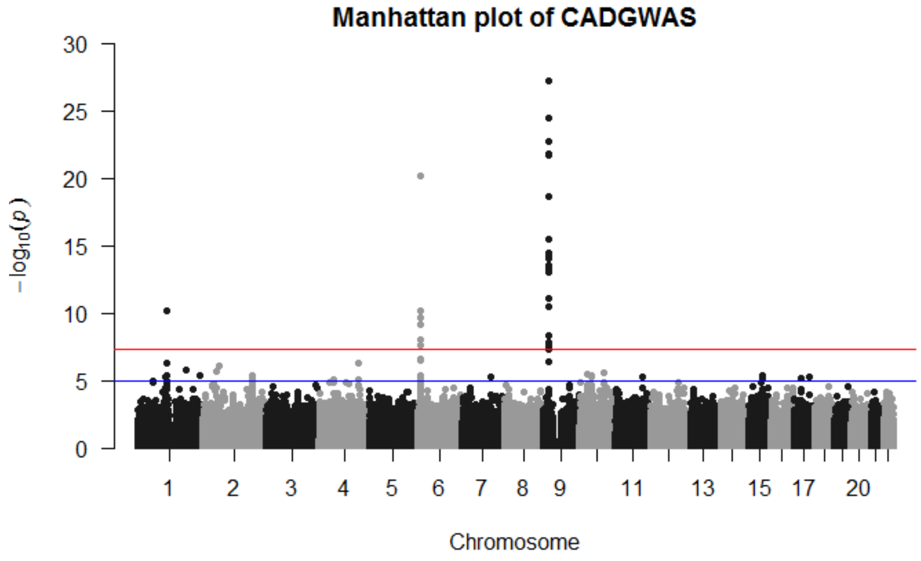

Figure 17. Manhattan plot of SNPs tested for association with coronary artery disease using summary statistics obtained from the C4D.
(The Coronary Artery Disease Genetics, 2011) The y-axis represents the –log$_{10}$(p) value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

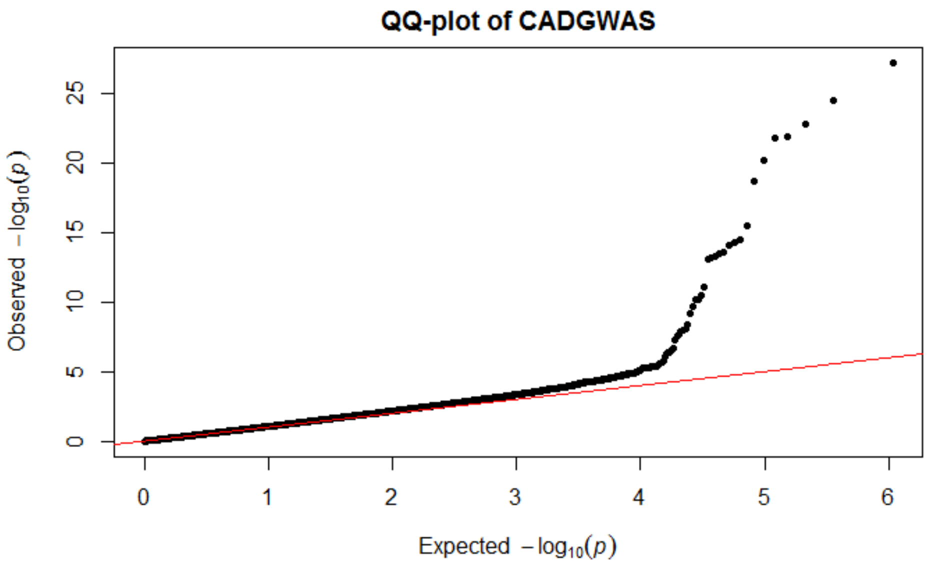

Figure 18. Quantile-quantile (Q-Q) plot for SNPs tested for association with coronary artery disease using summary statistics obtained from the C4D consortium.

(The Coronary Artery Disease Genetics, 2011) Each dot represents a SNP. The x-axis represents the expected –log$_{10}$(p) value. The y-axis represents the observed -log$_{10}$(p) value. The red line represents the pattern of -log$_{10}$(p) value if no SNP has a significant genetic association with the trait.

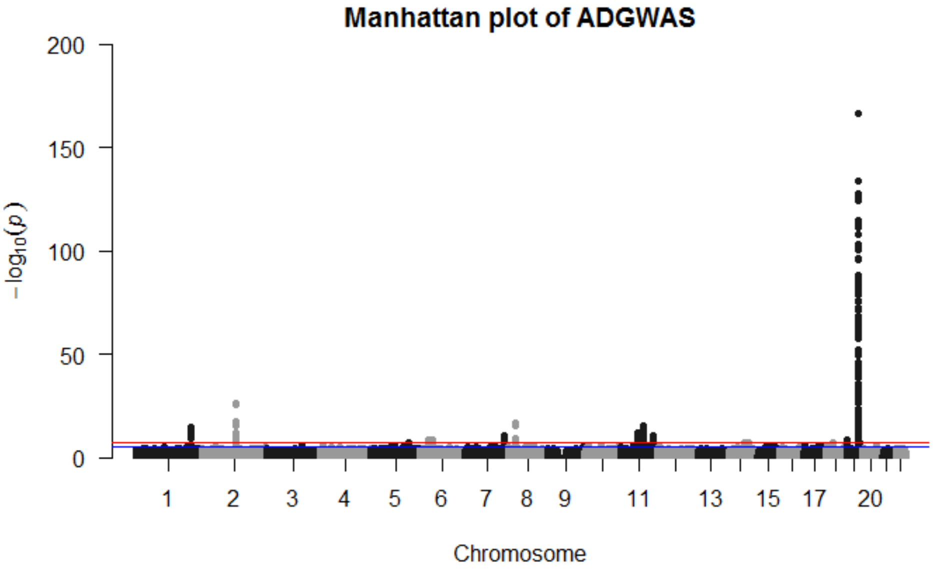

Figure 19. Manhattan plot of SNPs tested for association with Alzheimer's disease using summary statistics obtained from the IGAP consortium.

(Lambert et al., 2013) The y-axis represents the –log$_{10}$(p) value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

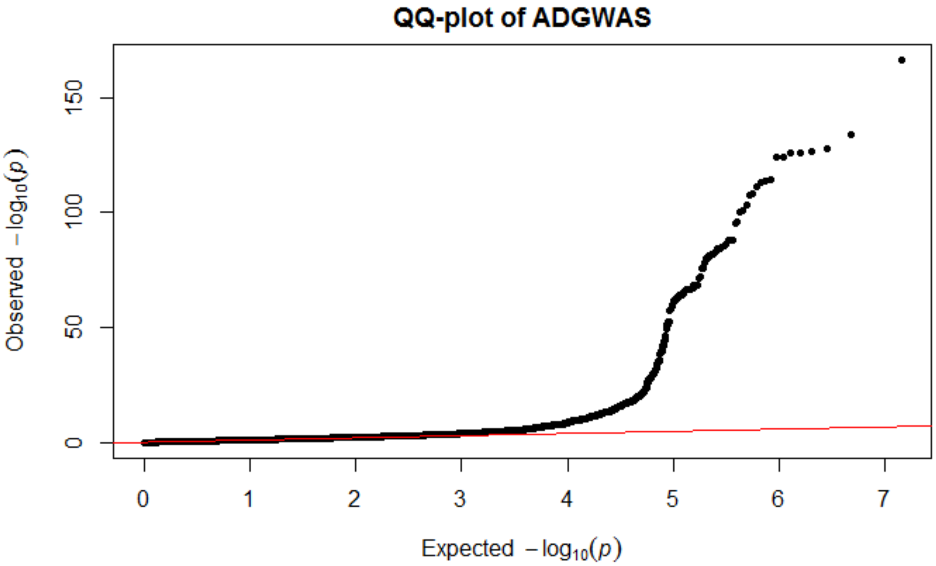

Figure 20. Quantile-quantile (Q-Q) plot for SNPs tested for association with Alzheimer's disease using summary statistics obtained from the IGAP consortium.
(Lambert et al., 2013) Each dot represents a SNP. The x-axis represents the expected $-\log_{10}(p)$ value. The y-axis represents the observed $-\log_{10}(p)$ value. The red line represents the pattern of $-\log_{10}(p)$ value if no SNP has a significant genetic association with the trait.

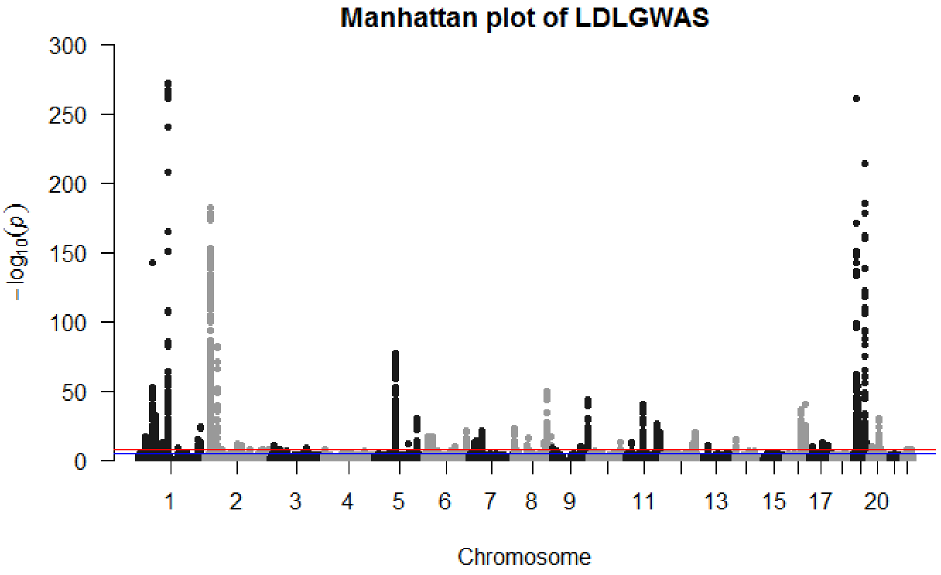

Figure 21. Manhattan plot of SNPs tested for association with LDL using summary statistics obtained from the GLGC.
(Willer et al., 2013) The y-axis represents the $-\log_{10}(p)$ value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

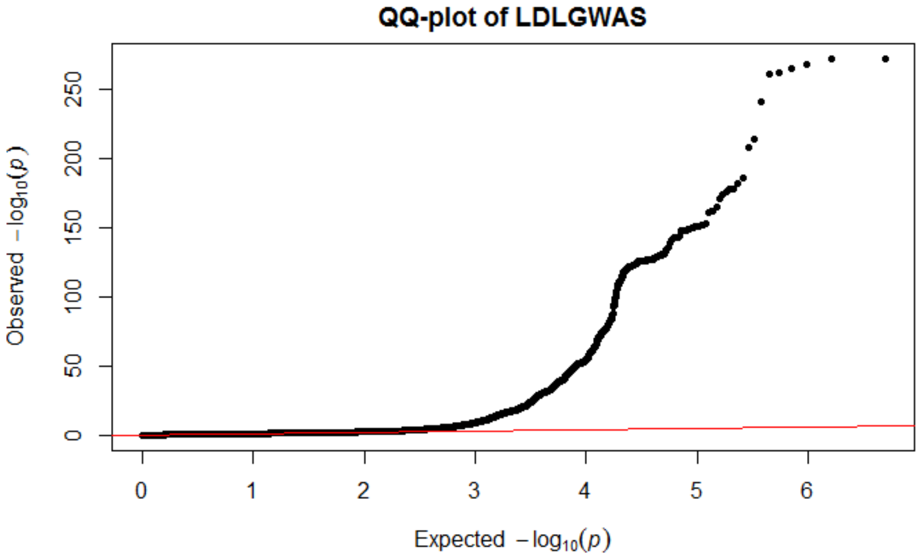

Figure 22. Quantile-quantile (Q-Q) plot for SNPs tested for association with LDL using summary statistics obtained from the GLGC.

(Willer et al., 2013) Each dot represents a SNP. The x-axis represents the expected –log$_{10}$(p) value. The y-axis represents the observed -log$_{10}$(p) value. The red line represents the pattern of -log$_{10}$(p) value if no SNP has a significant genetic association with the trait.

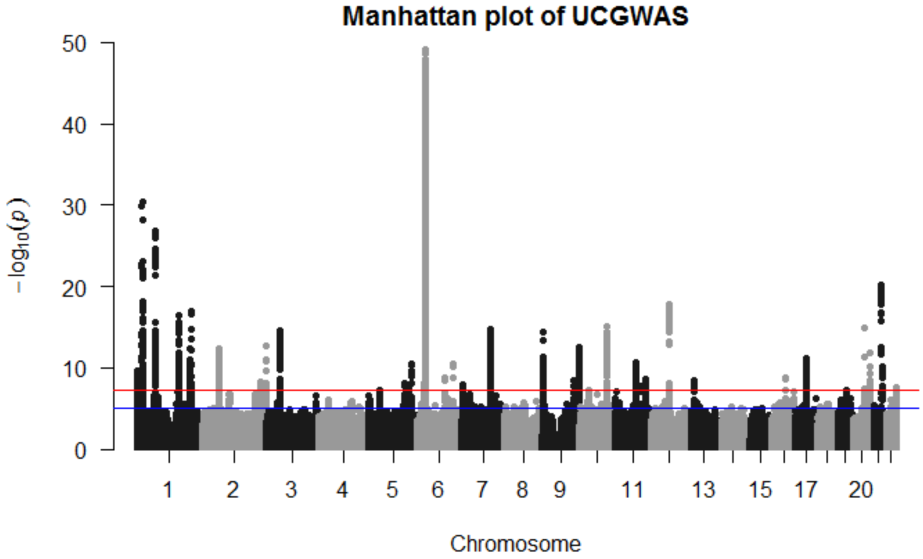

Figure 23. Manhattan plot of SNPs tested for association with ulcerative colitis using summary statistics obtained from the IIBDGC.

(Liu et al., 2015) The y-axis represents the –log$_{10}$(p) value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

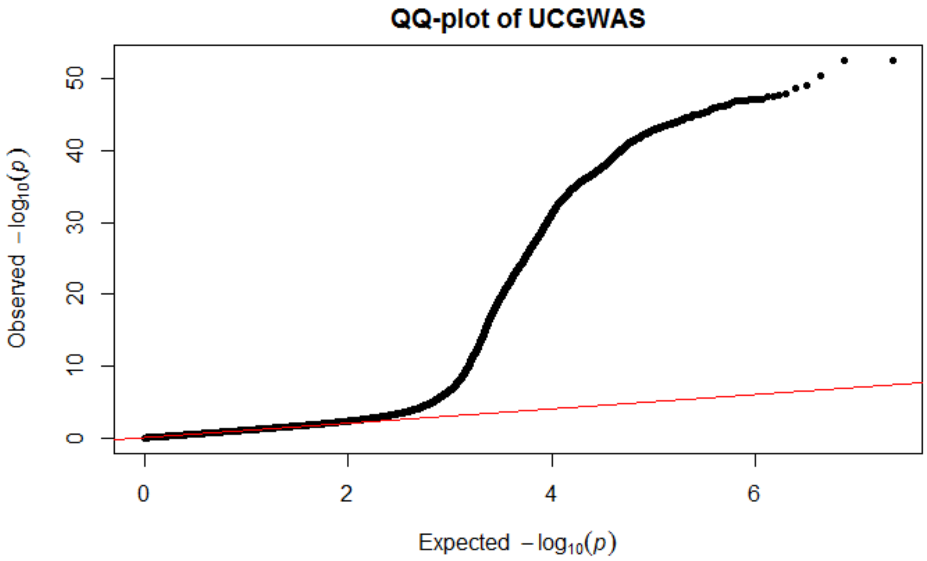

Figure 24. Quantile-quantile (Q-Q) plot for SNPs tested for association with ulcerative colitis using summary statistics obtained from the IIBDGC.
(Liu et al., 2015) Each dot represents a SNP. The x-axis represents the expected $-\log_{10}(p)$ value. The y-axis represents the observed $-\log_{10}(p)$ value. The red line represents the pattern of $-\log_{10}(p)$ value if no SNP has a significant genetic association with the trait.

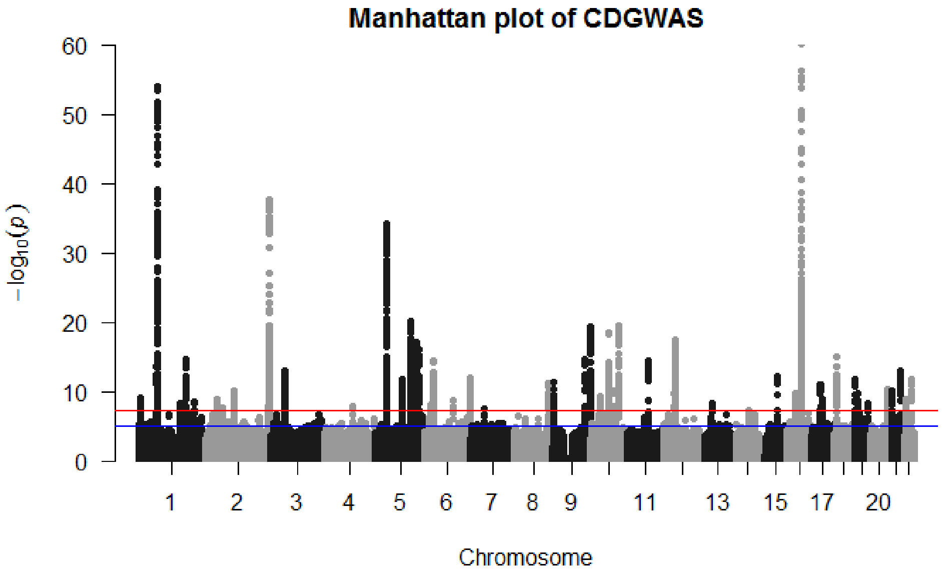

Figure 25. Manhattan plot of SNPs tested for association with Crohn's disease using summary statistics obtained from the IIBDGC.
(Liu et al., 2015) The y-axis represents the $-\log_{10}(p)$ value of each tested SNP divided across chromosomes, represented on the x-axis. The blue line represents a locus-wide significance level of 1 x 10^{-5}. The red line represents a genome-wide significance level of 5 x 10^{-8}.

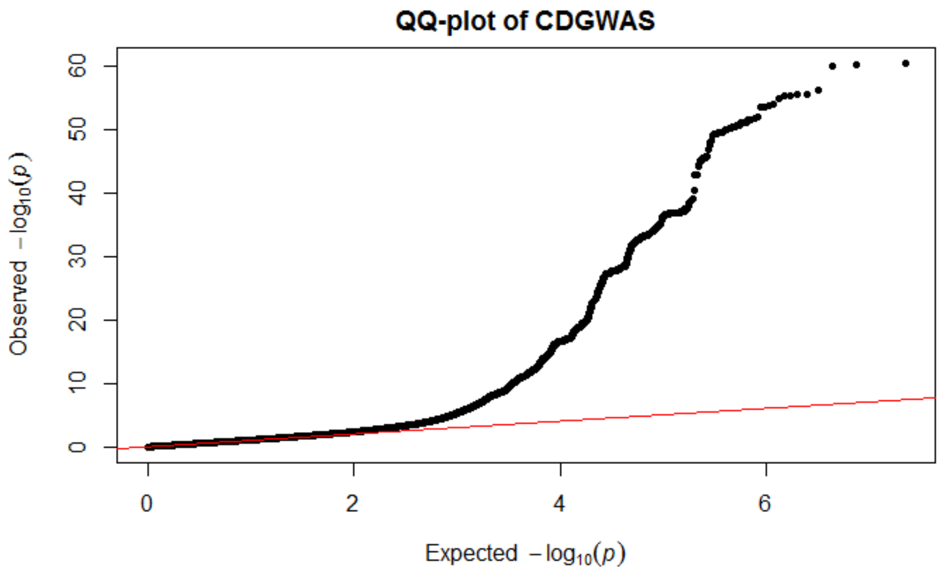

Figure 26. Quantile-quantile (Q-Q) plot for SNPs tested for association with Crohn’s disease using summary statistics obtained from the IIBDGC.

(Liu et al., 2015) Each dot represents a SNP. The x-axis represents the expected –$\log_{10}$(p) value. The y-axis represents the observed -$\log_{10}$(p) value. The red line represents the pattern of -$\log_{10}$(p) value if no SNP has a significant genetic association with the trait.

Table 3. Counts of strength of evidence and likelihood of further study by peak significance category, peak type, and annotator.

		Annotator 1					Annotator 4					Annotator 5			
		Strength of Evidence													
		0	1	2	3		0	1	2	3		0	1	2	3
Significiance Category	**HIGHLY**	3	2	3	4		2	5	1	4		0	3	1	2
	MODERATE	4	2	2	3		4	3	3	2		1	1	1	3
	SUGGESTIVE	7	2	2	1		9	2	1	0		0	5	0	1
		Likelihood of Further Study													
		0	1	2	3		0	1	2	3		0	1	2	3
Significiance Category	**HIGHLY**	2	2	2	6		5	3	1	3		0	2	2	2
	MODERATE	3	2	1	5		5	1	3	3		1	1	1	3
	SUGGESTIVE	8	1	0	3		10	1	0	1		2	2	1	1
		Strength of Evidence													
		0	1	2	3		0	1	2	3		0	1	2	3
Peak Type	**SYNTHETIC**	11	4	4	4		12	9	2	1		1	7	2	2
	ORIGINAL	3	2	3	4		3	1	3	5		0	2	0	4
		Likelihood of Further Study													
		0	1	2	3		0	1	2	3		0	1	2	3
Peak Type	**SYNTHETIC**	9	4	3	7		16	4	4	0		3	3	4	2
	ORIGINAL	4	1	0	7		4	1	0	7		0	2	0	4

Table 4. Positive annotation status by peak significance category, peak type, and annotator.

		Annotator 1			Annotator 2			Annotator 3			Annotator 4			Annotator 5	
		Positive Annotation Status													
		No	Yes		No	Yes		No	Yes		No	Yes		No	Yes
Peak Significance Category	**HIGHLY**	3	9		1	5		2	10		5	7		0	6
	MODERATE	4	7		4	2		2	10		5	7		1	5
	SUGGESTIVE	7	5		3	3		4	8		9	3		2	4
		Positive Annotation Status													
		No	Yes		No	Yes		No	Yes		No	Yes		No	Yes
Peak Type	**SYNTHETIC**	11	12		7	5		6	18		16	8		3	9
	ORIGINAL	3	9		1	5		2	10		3	9		0	6

APPENDIX B: ANNOTATION INSTRUCTIONS

The following pages contain the annotation instructions provided to each participating annotator. Subcategory 4 was only provided to annotators 1, 4, and 5.

Background

As a member of this research project, you will be acting as an annotator of genome-wide association study (GWAS) results, seeking to create plausible functional/biological 'stories' for selected top hits from the GWAS. You will be provided with the results of a GWAS for a given phenotype in the form of nine (9) LocusZoom plots of SNPs that reached genome-wide significance and SNPs that reached suggestive significance. You will also be provided with literature search results from Pubmed, PubMed Central, and Google Scholar in .csv format for the selected SNPs and genes within each LocusZoom plot. These will contain the title of the publication, abstract (if available), and authors (if available).

Instructions

1. We will provide you with the literature searches, LocusZoom plots, and phenotype.
2. You may examine the LocusZoom plots and literature searches for any evidence or suggestions of association.

3. Using a template provided to you, write a summary of evidence you have found that supports an association between the genes/SNPs and the given phenotype, citing the sources of this evidence. You are trying to make a plausible "story" as to how each SNP may biologically or functionally be related to the trait via possibly 'interesting' genes in the immediate region of the SNP.
    a. Written below is an example paragraph in a style similar to what we would like to see
        i. "We found that previous authors have found associations between rs1234567 and BMI (Test et al. 2009, Source et al. 2012, Professor et al. 2016). Researcher et al. showed that mice with increased expression of the Wght gene showed a significant increase in BMI (2011). However, because the WGHT gene, where rs1234567 is found in an intron, is in very high LD with rs7654321 in the nearby GRWTH gene (r^2 >0.9), it is difficult to determine if this association is primarily due to WGHT or GRWTH
4. After writing the summary paragraph, you will be asked two brief questions about what you found in the region. These questions will follow this format:

On a scale of 0 to 3, how strong is the evidence that this region is associated with this trait, with 0 being no evidence and 3 being very strong evidence? Please bold or circle your choice.

0 1 2 3

On a scale of 0 to 3, how likely would you be to continue studying this region for its association with this trait, with 0 being very unlikely and 3 being very likely? Please bold or circle your choice.

0 1 2 3

BIBLIOGRAPHY

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., . . . Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research, 22*(9), 1790-1797. doi: 10.1101/gr.137323.112

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., . . . Neale, B. M. (2015). An Atlas of Genetic Correlations across Human Diseases and Traits. *Nature genetics, 47*(11), 1236-1241. doi: 10.1038/ng.3406

Dowle, M. A. S. (2017). data.table: Extension of `data.frame` (Version 1.10.4). https://CRAN.R-project.org/package=data.table: CRAN.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat Rev Genet, 13*(2), 135-145.

Ioannidis, J. P. A., Tarone, R., & McLaughlin, J. K. (2011). The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology, 22*(4), 450-456. doi: 10.1097/EDE.0b013e31821b506e

Ioannidis, J. P. A., Thomas, G., & Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet, 10*(5), 318-329.

Johnson, E. C., Border, R., Melroy-Greif, W. E., de Leeuw, C. A., Ehringer, M. A., & Keller, M. C. (2017). No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes. *Biological Psychiatry, 82*(10), 702-708. doi: 10.1016/j.biopsych.2017.06.033

Kovalchik, S. (2017). Download Content from NCBI Databases (Version 2.1.7) [Package]. CRAN: CRAN.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., . . . Amouyel, P. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. *Nature genetics, 45*(12), 1452-1458. doi: 10.1038/ng.2802

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., . . . Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics, 47*(9), 979-986. doi: 10.1038/ng.3359

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., . . . Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research, 45*(Database issue), D896-D901. doi: 10.1093/nar/gkw1133

Palmer, C., & Pe’er, I. (2017). Statistical correction of the Winner’s Curse explains replication variability in quantitative trait genome-wide association studies. *PLOS Genetics, 13*(7), e1006916. doi: 10.1371/journal.pgen.1006916

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., . . . Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics, 26*(18), 2336-2337. doi: 10.1093/bioinformatics/btq419

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421-427. doi: 10.1038/nature13595 http://www.nature.com/nature/journal/v511/n7510/abs/nature13595.html#supplementary-information

Tak, Y. G., & Farnham, P. J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin, 8*, 57. doi: 10.1186/s13072-015-0050-4

The Coronary Artery Disease Genetics, C. (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *43*, 339. doi: 10.1038/ng.782 https://www.nature.com/articles/ng.782#supplementary-information

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., . . . Ritchie, M. D. (2011). Quality Control Procedures for Genome Wide Association Studies. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.], CHAPTER*, Unit1.19-Unit11.19. doi: 10.1002/0471142905.hg0119s68

Visscher, Peter M., Brown, Matthew A., McCarthy, Mark I., & Yang, J. (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics, 90*(1), 7-24. doi: https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet, 101*(1), 5-22. doi: 10.1016/j.ajhg.2017.06.005

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., . . . Boehnke, M. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLOS Genetics, 8*(8), e1002793. doi: 10.1371/journal.pgen.1002793

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications, 8*(1), 1826. doi: 10.1038/s41467-017-01261-5

Wickham, H. (2016). Easily Harvest (Scrape) Web Pages (Version 0.3.2). CRAN: CRAN.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., . . . The Global Lipids Genetics, C. (2013). Discovery and Refinement of Loci Associated with Lipid Levels. *Nature genetics, 45*(11), 1274-1283. doi: 10.1038/ng.2797