Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Alex Wang Date

Genome-wide association study in two populations to determine genetic variants associated with *Toxoplasma gondii* infection and relationship to schizophrenia risk.

By

Alex Wang

Master of Public Health

Department of Global Epidemiology

_____

Brad D. Pearce, Ph.D.

Faculty Thesis Advisor

Genome-wide association study in two populations to determine genetic variants associated with *Toxoplasma gondii* infection and relationship to schizophrenia risk.

By

Alex Wang

Bachelors of Science
Emory University
2012

Faculty Thesis Advisor: Brad D. Pearce, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2016

Abstract

Genome-wide association study in two populations to determine genetic variants associated with *Toxoplasma gondii* infection and relationship to schizophrenia risk.

By Alex Wang

According to a World Health Organization's Global Burden of Disease report, schizophrenia is one of the leading causes of years lost due to disability and is characterized by periods of psychosis, social withdrawal, and disorganized thought patterns and behavior. Like many neurological disorders, it is difficult to pinpoint the reason behind the manifestation of the affliction. However, there is a growing body of evidence that suggests a genetic linkage, with studies showing increased risk from 1% in the general population to over 40% in monozygotic twins. Despite these findings, genetic predisposition is not enough to illicit the onset of the disease phenotype. Instead, it is believed to interact with a second level environmental or generalized stressor to precipitate the development of the disease. One such stressor has been suggested to be infection with a protozoan parasite known as *Toxoplasma gondii (T. gondii)*.

Two genetically distinct populations were used in genome-wide association analyses to determine genes that may increase susceptibility to infection with the protozoan parasite and possibly increase risk of developing disorders such as schizophrenia. To conduct the analyses we chose two different outcome variables, one continuous and the other dichotomous classifications of toxoplasmosis infection. From the analyses, a list of single nucleotide polymorphisms was obtained and corresponding genes were identified. Among the top SNPs found in our dichotomous analyses, a variant associated with the gene *CHIAP2* and *CHIA* (rs10857870, p= 5.36E-06) was found. This encodes for a protein called chitinase that plays a role in defense against *T. gondii* cyst formation. Once the threshold value of $p<0.001$ was used to identify corresponding genes, a small number of genes were found to overlap in prediction of *T. gondii* infection between the two populations, among them were the genes found initially in AJ population: *FHIT*, *ALK* and *RBFOX1.* Though no direct linkage to increased susceptibility to infection was identified, pathways of interest that relate to cytokine regulation, transcript level alterations and chitinase activity may hold potential for future research.

Genome-wide association study in two populations to determine genetic variants associated with *Toxoplasma gondii* infection and relationship to schizophrenia risk.

By

Alex Wang

Bachelors of Science
Emory University
2012

Faculty Thesis Advisor: Brad D. Pearce, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2016

Table of Contents

Introduction

Schizophrenia

Schizophrenia (SCZ) is one of the world's most serious and complex brain disorders. While the incidence rate is relatively low (15.2/100,000), the chronic state that follows from early adult onset has led the global prevalence of SCZ to become relatively high (7.2/1,000) (1,2). Combining this epidemiological information with the results from the WHO's Global Burden of Disease study, which showed SCZ as one of the leading causes of years lost due to disability, it is increasingly important to better understand the etiology of this disorder (3).

This condition is most frequently characterized by periods of psychosis, disorganized thought patterns and behavior, with a general state of avolition, anhedonia, and social withdrawal (3,4). Besides these behavioral symptoms, patients suffering from SCZ also show deficits in their neurocognitive abilities. Studies have demonstrated that those with chronic SCZ showed impairments that range from one and a half to two standard deviations below their healthy counterparts on several key cognitive abilities, including working memory, verbal memory, attention, motor speed, and executive functions and verbal fluency (5–7).

Like many complex neurological disorders, it is difficult to pinpoint the root cause of this multifactorial disorder; however, it has shown that one of the greatest risk factor is a positive family history (2,8). While the risk of disease in the general population is below 1% globally, it increases to 6.5% in first degree relatives of the patients and to more than 40% in monozygotic twins (9,10). Results obtained from family, adoption and twin studies suggest that this risk reflects the genetic proximity between relatives and the family member afflicted with SCZ. Despite extensive genomic studies, individual genes that

consistently predispose individuals to an increased risk of developing SCZ have yet to be discovered (11–14).

Among the many genes implicated in SCZ, results from a recent study by Singh *et al.* connects the loss-of-function variant in *SETD1A* with a spectrum of severe developmental disorders that include SCZ (15). Nevertheless, the loss-of-function variants of *SETD1A* are very rare events, and clinical heterogeneity is still observed in carriers.

Endophenotypes

When a very specific and specialized phenotype occurs, there may be a single gene or a small set of genes responsible for producing this trait. Yet when a phenotype as complex as a psychiatric disorder like schizophrenia arises, where symptoms overlap with those of other mental disorders, multiple genes may contribute in an additive fashion to increase the total risk or vulnerability of an individual to develop the disorder. These genes are thought to predispose individuals and interact with second level environmental or generalized stressors to precipitate the onset of disorders. According to current theories of SCZ etiology, if the genetic load is strong enough, a relatively mild stressor may result in the development of the disorder. In contrast, if the genetic risk is considered mild, it may take a larger "second hit" to initiate the cascade in the central nervous system that result in the disease phenotype (16–18). Studies of clinically unaffected relatives of patients suffering from SCZ indicate an intermediate state of impairment termed "endophenotypes." Essentially, an endophenotype is a neuro-behavioral trait that shares characteristics with the psychiatric illness but is thought to be closer to underlying biology that the complex psychiatric illness.

Thus endophenotype research has focused often on family members of people with SCZ, and has attempted to tease-out intermediate or partial level of impairment that occurs

in many functional domains, including neurophysiology, metabolic functioning, verbal reasoning, neuro-cognition and sensory motor gating (16,19). When the clinical presentation of a disorder like SCZ is so complex, having the ability to perform laboratory-based tests on endophenotypes provides researchers the ability to better identify specific genetic variants that make an individual vulnerable to developing the disease phenotype. One such endophenotype that holds promise in learning more about the etiology of SCZ lies in deficits seen in sensory motor gating (20–22). Neuropsychological tests such as assessment of P50 suppression and pre-pulse inhibition of startle responses have been developed to assess this ability in subjects (23).

Toxoplasmosis

From our understanding of endophenotypes, it seems highly likely that there are many genes that confer risk in the general population, and patients are likely to have inherited several risk genes that interact with each other and the environment to bring about the development of SCZ once a threshold is crossed (2). This theory has sparked renewed interest in the possible roles of infectious pathogens as environmental agents in some cases of SCZ. One particular pathogen of interest that is becoming more prominent in SCZ research is the intracellular protozoan parasite *Toxoplasma gondii (T. gondii*).

This infectious microbe is known to be one of the most successful parasites with a varied worldwide prevalence (11). In the United States alone, the incidence rate of infection is estimated to be 1.1 million persons each year, with toxoplasmosis being the second leading cause of deaths attributable to food-borne illness and a leading contributor to loss of quality-adjusted life years (24). Felines have been identified as the only definitive host for *T. gondii*, but the parasite is capable of infecting almost all warm-blooded mammals through various transmission methods (11,25). Only when the parasite reaches the intestinal tract

of its feline host does it become capable of sexual reproduction (26). When these parasites are ingested by an intermediate host, which include human beings, the oocytes undergo asexual reproduction to form small cysts in muscle, liver and neuronal tissue, and are capable of residing in a latent stage potentially for the rest of the host's life (27).

It is well known that infections of *T. gondii* can alter behavioral and cognitive abilities in rodents. Studies from the late 1970s by Piekarski and Witting reported that infection with the parasite caused learning impairment in both mice and rats, and memory impairment in only mice (28,29). This research paved the way for the development of the "manipulation hypothesis" which theorizes that a parasite has the ability to alter the behavior of its host to improve its own transmission rate (30,31). There have been studies that have shown that the hypothesis holds true in other mammalian species. When chimpanzee or other primate species were infected with the parasite, it was shown that these infected individuals lost their innate aversion towards the urine of leopards, the primates' only natural predator (32,33).

Latent toxoplasmosis and congenital outcomes

Moving from animal models, studies have started to look into the effects of *T. gondii* on human behavior and characteristics. It is estimated that prenatal infections with *T. gondii* and rubella are responsible for 2-3% of all cases of cognitive deficiency globally (34). Congenital infection with the parasite, especially early on in pregnancy have been link to intracranial calcifications, mental developmental delays, seizures and retinal damage (35,36). Less well known are the long-term effects of congenital infection with the parasite that occur late in the pregnancy which can remain dormant at birth. A case-control study from Brazil reported that among 450 mentally handicapped and 395 healthy children, approximately 55% of the cases and 29% of the controls tested sero-positive for *T. gondii*;

the study also showed that maternal exposure to cats and contact with soil suspected to contain *T. gondii* were associated with increased risk of mental developmental disability (37). Research groups in the United States have reported similar finds of lower IQ in children and chorioretinitis following latent infections, while other groups who followed a similar cohort in Europe reported no lowering in IQ or other significant side-effects (38–40).

Latent toxoplasmosis and mental health disorders

Recent studies have shown that there is an association between SCZ *and T. gondii* infection. A few meta-analyses have been carried out to determine the association of *T. gondii* on the risk of developing SCZ, one such analysis performed by Arias *et al.* found a significant association (OR=2.70; CI=1.34-4.42; p=0.005) between SCZ and the presence of *T. gondii* infection when comparing the results of eight different studies (41). Another carried out by Torrey *et al.* identified 23 studies of 42 that met their inclusion criteria and found the prevalence of *T. gondii* antibodies in individuals with SCZ were significantly higher than antibodies found in controls with an OR of 2.73 (42).

There has been evidence that individuals who show higher levels of *T. gondii* IgG antibodies have significantly greater risk of developing SCZ and even show more severe symptoms of psychoses compared to their health control counterparts (43). A finding that supports the theory that a "second hit" is required to trigger the onset of a disorder from an endophenotype. In one study conducted on military personnel*, T. gondii* IgG antibody levels were collected pre- and post-diagnosis of SCZ and collected data showed that there was a strong positive association with an overall hazard ratio of 1.24 (44). Other studies report an association between IgG antibody titer levels and the development of other disorders such as anxiety, depression and bipolar disorder. A study by Groër *et al.* found that higher titers

of IgG *T. gondii* antibodies in pregnant women positively related to the development of anxiety and depression during pregnancy (45). Hamdani and colleagues showed in a French cohort that the group who were sero-positive for IgG antibodies had an OR of 2.7 of having bipolar disorder compared to the sero-negative group when controlling for age (46). A study conducted by Pearce *et al.* also found that there was a significant relationship between *T. gondii* sero-prevalence and bipolar disorder type I among respondents to the third National Health and Nutrition Survey (47).

In addition to the development of mental disorders, infections with *T. gondii* may present an increased risk of mortality among individuals suffering from SCZ. A prospective study conducted by Dickerson *et al.* followed 358 individuals for a period of up to 5 years that were diagnosed with SCZ and were tested for *T. gondii* antibodies. They found a significant difference in mortality rates of those who were sero-positive (mortality rate= 8.6%) compared to sero-negative SCZ patients (mortality rate= 1.7%) (48). The study looked at mortality due to natural causes; however, other literature points to the possibility of an association between *T. gondii* infection and increased rates of suicide (49–51). One such study conducted by Yagmur *et al.* measured IgG and IgM antibody levels in a Turkish population that consisted of 200 cases of suicide attempts and 200 health controls. What they found was that there was a statistically significant difference (P=0.004) in the antibody levels between suicide attempters and the controls (50). Another study relating infection to suicide rates conducted by Ling and colleagues obtained results that suggested a positive relationship between rates of infection with *T. gondii* and suicide in women of postmenopausal age (52).

Toxoplasmosis and cognitive function

Human beings are far from the optimal intermediate host for *T. gondii* as they are rarely preyed upon by felines. But studies have shown infections with the parasite can cause behavioral alterations in humans and have been termed 'parasitic constraint' (53). This constraint may have detrimental effects on reaction time, memory and cognitive abilities. It was found by Flegr *et al.* that infected individuals performed significantly worse compared to their uninfected counterparts on a standard computerized test and appeared to lose focus more quickly (33). Another study by a different group led by Flegr even found that individuals infected with latent toxoplasmosis have increased risk of traffic accidents compared to non-infected controls (54). Havlícek *et al.* showed through a double blind study that there was a deterioration in psychomotor performance in subjects with latent toxoplasmosis compared to their healthy control counterparts. Based on this finding, the researchers suggest that psychomotor slowing is a slow and cumulative process in infected individuals (55).

More recently, Pearce *et al.* showed infected subjects scored in the worse quartiles on cognitive tests measuring visual response times, coding ability and learning/ memory compared to controls when adjusting for various demographic and health-related covariates (56). They also presented findings of significant slowing in acoustic startle response with the greatest increase in latency occurring in the *T. gondii* sero-positive schizophrenic group (57). An independent study carried out on a different population by Příplatová. *et al.* found that subjects infected with *T. gondii*, especially men, had significantly prolonged reaction times to simple startle signals compared to controls (58). The inclusion of acoustic pre-pulse shorten reaction times of all subjects, with stronger effects in male subjects and increased with duration of infection.

Genetics and toxoplasmosis

While there have been many studies that have tried to find genetic variants associated with psychiatric disorders like schizophrenia and bipolar disorder, it is only recently that there has been an increased interest in finding links between genetic variants, such as single nucleotide polymorphisms (SNPs), and an individual's susceptibility to infectious diseases, and the detrimental outcomes that associated with them. Table 1 provides a summary of some of the published findings on genes related *T. gondii* infection among in human studies.

Table 1: Summary of previous literature on genes associated with *T. gondii* infection.

Paper	Authors	Gene(s)	Findings
Candidate gene analysis of ocular toxoplasmosis in Brazil: evidence for a role for toll-like receptor 9 (TLR9) (59)	Peixoto-Rangel *et al.*	*TLR9*	Authors reported markers of TLR2, TLR5, and TLR9 that were found to have an association with an ocular disease associated with *T. gondii* infection, with SNPs from TLR9 having an effect on transcript.
Evidence for associations between the purinergic receptor P2X7 (*P2RX7*) and toxoplasmosis (60)	Jamieson *et al.*	*P2RX7*	Authors found an association between variants found in the protein coding gene of P2RX7 to be associated with pro-inflammatory responses that occur in response to congenital infection.
Genetic and Epigenetic Factors at COL2A1 and ABCA4Influence Clinical Outcome in Congenital Toxoplasmosis (61)	Jamieson *et al.*	COL2A1, ABCA4	Found novel gene variants COL2A1 and ABCA4 to be associated with clinical outcomes of congenital toxoplasmosis.
Infection and Inflammation in Schizophrenia and Bipolar Disorder: A Genome Wide Study for Interactions with Genetic Variation (62)	Avramopoulos *et al.*	*SGK1*	Observed a genetic variant that was shows strong biological plausibility. SGK1 region encodes for an mTORC2-dependent regulator utilized in the differentiation and function of T cells.

Though not many genetic studies have been conducted that examine how polymorphisms play a role in altering human susceptibility to *T. gondii*, there is some literature that proposes a link between host genic variation and infection. Based on a small family-based study in Brazil, Peixoto-Rangel *et al.* found that *TLR9* polymorphisms associated with ocular toxoplasmosis (59). This gene is known to play a role in pathogen recognition and activation of innate immunity, the authors hypothesized that there is a possibility that interactions with the parasite is causing an over-active pro-inflammatory response which results in the presentation of an ocular disease. Another study conducted by Jamieson *et al.* found that polymorphisms at *P2RX7*, which codes for the $P2X_7$ protein and activates interleukin activity, influenced susceptibility to toxoplasmosis in a North American family study cohort that contained children who presented symptoms associated with congenital *T. gondii* infection (60). A different group lead by Jamieson also found an association between polymorphisms found in the genes *COL2A1* and *ABCA4* to clinical presentation of congenital infection with the protozoan parasite (61). The gene *COL2A1* encodes type II collagen which is found in the vitreous humor, cornea, sclera, lens, ciliary body, retinal pigment epithelium, and retina of the eye. *ABCA4* encodes a retina-specific ATP-binding cassette transporter protein that is located at the rim of the photoreceptor outer segment disc. Avramopoulos and colleagues found an association between *SGK1*, a gene that encodes for an mTORC2-dependent regulator required in the differentiation and function of T cells, and toxoplasmosis infection among those suffering from schizophrenia (62). *SGK1* is also seen to contribute to the regulation of diverse cerebral function which may hint at a role in the development of mental disorders.

Despite the promising findings that have been published that relate genetic variants in immune responses to both cerebral and ocular presentation of toxoplasmosis, there still is limited information available, with a significant gap in the field of infectious disease

genetic analysis. Most notably the lack of genome-wide association studies (GWAS) that assess whole genomes of affected patients. Hence, to our knowledge there has been no reported genome-wide association study of *T. gondii* susceptibility in any population. Only a few of these GWAS studies have been published to date, one of which conducted by Børglum *et al.* looked at the genome-wide interaction between cytomegalovirus and SCZ which was able to identified new possible loci that were not previously suspected to be involved in SCZ (3,63). Another recent genome-wide study conducted by Avramopoulos *et al.* that was mentioned previously and listed in table 1, looked into infection and how these can interact with genetics to increase the risk of developing SCZ or bipolar disorder. This study was performed with the same cohort and GWAS data that is used as a discovery dataset in the current study. However, the study by Avramopoulos *et al.* did not examine the polymorphisms that associate with susceptibility to *T. gondii* per se, but rather only examined SNPs that statistically interact with infection to modify risk of mental illness.

By determining the common genetic variants (single nucleotide polymorphisms) that are associated with schizophrenia and *T. gondii* in two study populations, we set out to evaluate first whether there are genetic polymorphisms that associate with *T. gondii* infection, and secondly, determine whether there is a significant gene-infection interaction that could increase the risk of schizophrenia. To do this, we analyzed from a genome-wide association study (GWAS) an initial population that will act as the discovery population. This allowed us to identify putative *T. gondii* susceptibility genes, which can be used to test for association with infection in a smaller replication sample.

The discovery population is relative genetically homogenous and of Central European descent (Appendix Figure 1). This contrasts to the replication population that largely consists of individuals of African American descent (Appendix Figure 2). The hope is

that we will find genes that span across the genetically heterogeneous populations that increase the susceptibility of infection with *T. gondii* and reveal pathways in which infection with the protozoan parasite causes an increase in the risk of developing schizophrenia.

Methods

Participants

Subject recruitment for the Ashkenazi Jewish (AJ) population, the discovery dataset, is outlined by Avramopoulos and colleagues (62). Subjects diagnosed with SCZ, or schizotypal personality disorder, were recruited over a 15 year period (1996-2011) through various forms of recruitment methods including advertisements, talks and a study website. Patients were interviewed in-person and diagnosed through a consensus procedure outlined in the paper (62). Screened controls who had no history of depression, mania, psychosis, psychiatric hospitalization, depression, or suicide attempts were recruited over a four year period (2003-2007) at Jewish community centers, synagogues and professional meetings. All cases and controls were ascertained reporting four grandparents of known AJ decent. A total of 790 subjects were included in the analysis.

Participants for the Grady Trauma Project (GTP), the replication dataset, were recruited from the primary care, obstetrics and gynecology, and diabetes clinics of Grady Memorial Hospital in downtown Atlanta. Patients in these clinics were recruited from the waiting rooms Monday-Friday during regular clinic hours. Volunteers and staff at GTP were instructed to approach patients in waiting areas in a random fashion and participation eligibility was determined following initial contact. To be included, participants must be 18 to 65 years of age, able to communicate with the interviewer in English, willing to undergo the initial interview, and willing to provide a saliva sample. Following patient eligibility, study procedures including time commitment, monetary compensation were explained in detail after which verbal and written informed consent was collected. All participants are assigned a study identification number for de-identification purposes (64).

We selected the GTP sub-sample based on the completeness of phenotypic data available (including startle data which is used as an endophenotype for SCZ and Post-Traumatic Stress Disorder (PTSD)). SAS software, ver. 9.4 was used to clean the GTP dataset. The starting dataset consisted of a total of 650 subjects. The dataset was cleaned step by step outlined in Figure 1. Following the cleaning steps, there were a total of 363 subjects eligible to be included in the analysis. Of these 363 individuals, 77 subjects were missing SNP data which resulted in a final cohort of 286 individuals.

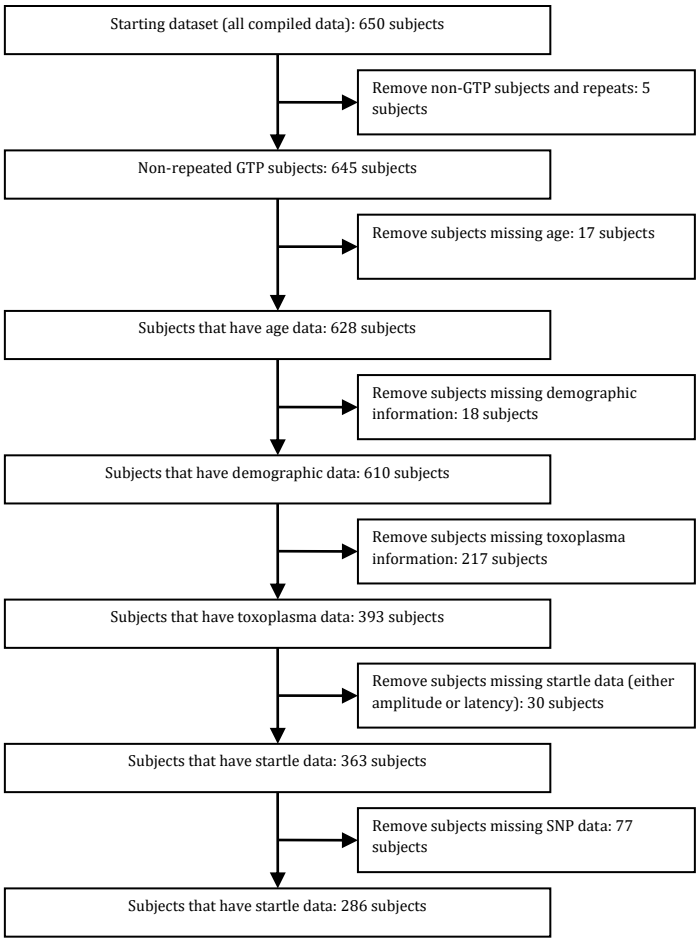


Figure 1: Flow chart to show the cleaning steps for the GTP dataset.

Immunoassay measurement

Immunoassay measurement for antibodies of toxoplasmosis in the AJ population was outlined by Avramopoulos and colleagues (62). Dilute concentrations of plasma were applied to antigens immobilized on the wells of microtiter plates and quantified using anti-human IgG and corresponding substrate. Reagents were obtained from IBL Laboratories, Hamburg Germany.

Sera specimens from participants recruited to the Grady Trauma Project underwent assessment for *Toxoplasma* IgG antibodies in compliance with manufacturer's instructions (Bio-Rad, Redmond, WA). Sero-positivity was determined through a measure of a weakly positive calibrator index value and its absorbance at 450nm which was then multiplied by the absorbance of the sample to find the sample index value. An index value greater than 0.9 was indicative of *Toxoplasma* sero-positivity. For all sero-positive subjects, a discrete titer was determined using a three point curve of the blank, the weakly positive calibrator, and a strongly positive calibrator, in accordance with the manufacturer's instructions.

For participants that received an index score that fell between being sero-positive and sero-negative, they were given a score of "equivocal". Typically when subjects receive this status, the infection may have occurred too recently to produce a robust IgG response and may still be in the acute phase. To properly evaluate the levels of IgG antibody, a follow-up visit is requested after a set duration of time for subjects to return and have their antibody levels measured again. However, since this was not able to be performed, we chose to classify those with equivocal status as sero-positives.

Genotyping

Protocol for genotyping in AJ dataset is outlined by Avramopoulos and colleagues (62). Genotyping was performed with Affymetrix Human Genome-Wide SNP Array 6.0 at Emory University. Genotypes were called using the corrected robust linear mixture model (CRLMM), an algorithm for preprocessing and genotype calling of Affymetrix SNP array data.

Genotyping for GTP was conducted using Illumina's HumanOmni1-Quad or OmniExpress BeadChips (Illumina, Inc., San Diego, CA) (65). Standard quality control of the

genome-wide data was performed using PLINK ver. 1.9 (available in the public domain at http://pngu.mgh.harvard.edu/purcell/plink/).

Genotype data cleaning

Genotype data cleaning was performed on both populations using the software package PLINK ver. 1.9. We followed the cleaning steps outline in the Psychiatric GWAS consortium (66): 1) remove SNPs with more than 5% missing data, 2) remove subjects with more than 2% missing data after the first step of SNP removal 3) remove SNPs with more than 2% missing data after the removal of the subjects 4) remove SNPs out of Hardy Weinberg equilibrium at $p<10^{-6}$ to account for the multiple tests. Further excluded SNPs with minor allele frequency<0.05 as our relatively small sample is likely to introduce artifacts in the presence of even small deviations of the traits from normality.

Genome-wide association study

GWAS analyses were conducted using PLINK ver. 1.9 (67). Related individuals (p-value >0.125) were removed before calculating principal components. Hapmaps and principal component plots were created to ensure no outliers would affect association analyses. Principal components were calculated using R ver. 3.2.2 (available in the public domain http://www.R-project.org/) and included into the association analyses as covariates (Appendix Figures 3 & 4) (68). Age and sex were also included as covariates in the analyses. The obtained *Toxoplasma* sero-results of both populations showed right skewed distributions and values were therefore logarithmically transformed using SAS software to represent a more normal distribution (Appendix Figure 5). Manhattan and Q-Q plots were plotted using PLINK and R software packages.

Obtained SNP data was then run through QIAGEN's Ingenuity® Pathway Analysis software (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) to retrieve known

SNPs by position or by association with a coding gene (69). The software was able to map regions within 2 Kb (base-pairs) upstream or 0.5 Kb downstream of the SNP. Those variants that were not identified using this tool were input into the National Center for Biotechnology Information's Variation Viewer (http://www.ncbi.nlm.nih.gov/variation/view/) to confirm the genetic variants were not in coding regions (70).

Once functional SNPs were identified, regional association plots of the chromosomes were obtained by inputting chromosomal SNP data into Single Nucleotide polymorphisms-annotator (SNiPA) (http://snipa.helmholtz-muenchen.de/snipa/index.php?task=regional_association_plot) (71). Along with the plots, a legend is provided to better understand the annotations that come with the plot. Direct transcript effects were designated for variants in coding sequences, frame-shift variants, missense variants, initiator codon variants, or the gain or loss of a stop codon. Direct regulatory effects refers to variants that are able to directly alter the level of a transcript. Putative regulatory effects refer to variants that are present in regulatory regions, are located within transcription factor binding sites, or are located up or downstream from a 3' or 5' gene variant. Putative transcript effects refer to SNPs that are in untranslated regions both either 3' or 5', may be considered non-coding variants or exons, or may be present in splice regions or possibly nonsense-mediated decay transcript variants. More information regarding the terminology used in these legends can be found at the SNiPA website (http://snipa.helmholtz-muenchen.de/snipa/index.php?task=supplement).

After the completion of the association analyses, genes found to overlap in the two populations were run through QIAGEN's Ingenuity® Pathway Analysis software to create

biological prediction causal network maps to help better understand how overlapping genes associate with one another in functionality.

Results

Demographic Information

Table 2 highlights some of the available demographic information of both populations. The mean age of the two populations were fairly similar with the discovery dataset having a slightly older population at 48.3 years old (SD=13.4) compared to the replication dataset that had a mean age of 41.8 years old (SD=11.5). Overall, the Ashkenazi Jewish population consisted of 30.6% females and contained 128 individuals who tested positive for *T. gondii* antibodies. The median IgG antibody score for the population was 0.0691 (IQR=0.0370-0.137). Scores above 0.9 were considered positive for infection. The GTP dataset contained a greater percentage of female participants (67.5%) compared to the other population but this difference can be attributed to the method in which participants were recruited at Grady Memorial Hospital. In the GTP population 34 participants were classified as being positive for *T. gondii*. The median IgG antibody score subjects received was 0.0925 (IQR=0.0552-0.198).

The populations were stratified by schizophrenia or schizotypal personality disorder diagnosis with cases including those who have been diagnosed as such, and controls being those free of any psychiatric diagnoses. In the discovery dataset, there were a total of 519 cases with a mean age of 48.3 years old (SD=12.1) and contained 176 female participants (33.9%). Of those in this schizophrenia case group the IgG antibody score was 0.0659 (IQR=0.0360-0.121) and 68 subjects were sero-positive for toxoplasmosis (13.1%). Considering the control group, there were a total of 271 subjects with a mean age of 57.2 (SD=11.2) and contained 154 female participants (60.5%). In this group the median IgG antibody score was 0.0659 (IQR=0.0360-0.121) and there were a total of 60 subjects who were classified as being sero-positive (22.1%).

In the replication dataset, there were a total of 25 individuals who were classified as schizophrenia cases. Of these individuals the average age was 43 years old (SD=8.90) and included 16 female participants (64.0%). In this group, the median IgG antibody score was 0.130 (IQR=0.0417-0.233) and 4 subjects were classified as being sero-positive for toxoplasmosis.

The control group was considered separately from the group labeled as "others". The "other" group consisted of individuals who were not diagnosed with SCZ or schizotypal personality disorder diagnosis but did contain individuals who had an "other" history of serious psychiatric illness: psychiatric hospitalization, cocaine use and/ or had suffered from other substance abuse. From this population, we extracted individuals who did not have a history of any of these criteria. Due to the high prevalence of PTSD (n=89, 34.9%) and depression (n=84, 32.9%) these diagnoses were omitted in our inclusion criteria as the group would only contain 58 participants. This gave us a group that consisted of 84 individuals. Of these individuals, the average age was 37.5 years old (SD=16.7%) and contained 66 female participants (78.5%). The median IgG antibody score obtained by the group was 0.0926 (IQR= 0.0529-0.232) and 13 of these individuals met the criteria to be classified as sero-positive for toxoplasmosis infection (15.5%).

Table 2: Descriptive statistics of the participants in the Ashkenazi Jewish population (n=790) and the Grady Trauma Project population (n=286).

	Ashkenazi Jews (Discovery)	**Grady Trauma Project (Replication)**
Overall population (n)	**790**	**286***
Demographic Characteristics		
Age in years, mean (SD)	48.3 (13.4)	41.8 (11.5)
Female participants, n (%)	340 (30.6)	193 (67.5)
Toxoplasmosis sero-status		
Positive sero-status, n (%)	128 (16.2)	34 (11.9)
Toxoplasma IgG score, median (IQR)	0.0691 (0.0370, 0.137)	0.0925 (0.0552, 0.198)

Cases (n)	**519**	**25**
Demographic Characteristics		
Age in years, mean (SD)	48.3 (12.1)	43.0 (8.90)
Female participants, n (%)	176 (33.9)	193 (67.5)
Toxoplasmosis sero-status		
Positive sero-status, n (%)	68 (13.1)	4 (16.0)
Toxoplasma IgG score, median (IQR)	0.0659 (0.0360, 0.121)	0.1030 (0.0417, 0.233)
Controls (n)	**271**	**84**
Demographic Characteristics		
Age in years, mean (SD)	57.2 (11.2)	37.5 (16.7)
Female participants, n (%)	164 (60.5)	66 (78.5)
Toxoplasmosis sero-status		
Positive sero-status, n (%)	60 (22.1)	13 (15.5)
Toxoplasma IgG score, median (IQR)	0.0786 (0.0414, 0.219)	0.0926 (0.0529, 0.232)
Others[Δ] (n)		**255**
Demographic Characteristics		
Age in years, mean (SD)		41.8 (11.7)
Female participants, n (%)		175 (61.2)
Toxoplasmosis sero-status		
Positive sero-status, n (%)		30 (11.8)
Toxoplasma IgG score, median (IQR)		0.0915 (0.0580, 0.192)

*6 observations missing SCZ status

[Δ] Others group consists of individuals who may have a history of psychiatric hospitalization, cocaine use or other substance abuse.

Discovery dataset

We performed a genome-wide association analysis in our discovery population with *T. gondii* sero-conversion status as a dichotomous outcome variable to determine genetic variants that may increase an individual's susceptibility to infection. None of the obtained SNPs in the outcome phenotype provided a genome-wide significant result after performing Bonferroni correction when accounting for multiple testing. Once we made the Manhattan plot (Figure 2) we took the results obtained from the analysis and identified SNPs that had a $p<10^{-4}$ to see whether they were located in gene regions.

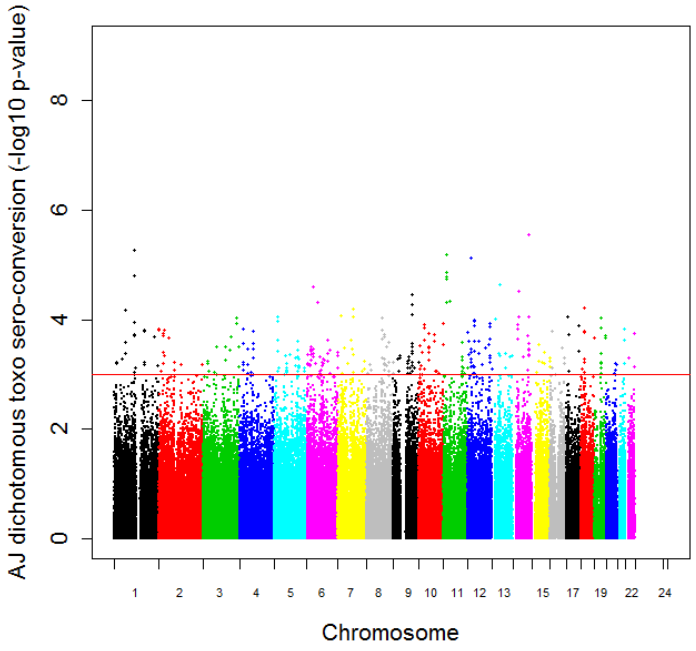
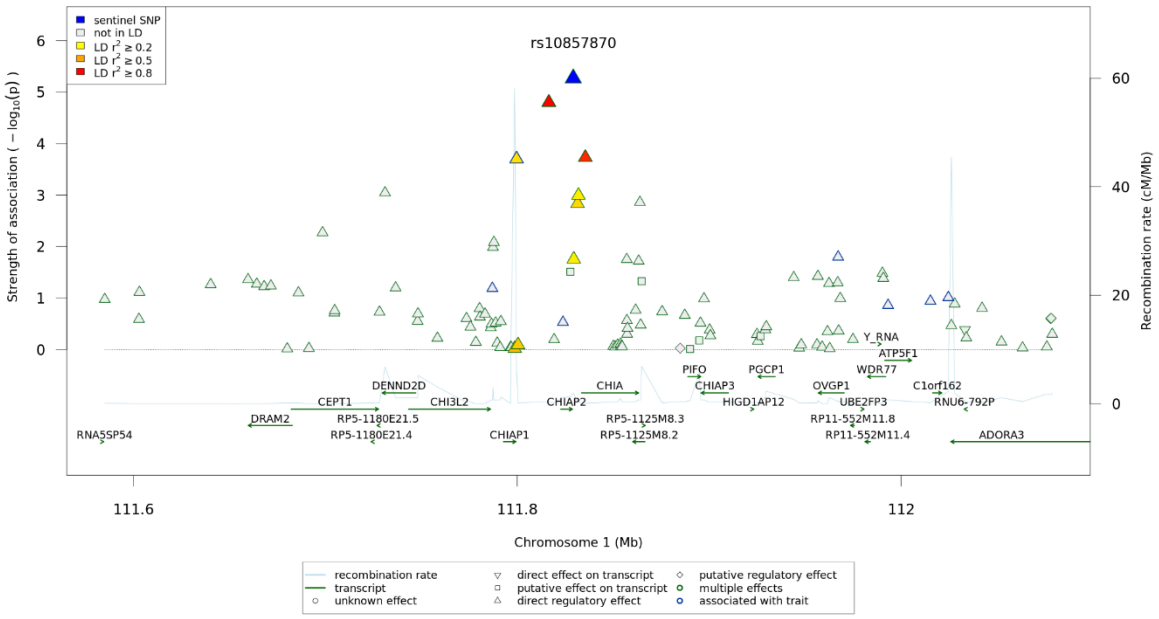


Figure 2: Manhattan plot ($-\log_{10}$ [P-value] genome-wide association plot) of the Ashkenazi Jewish population (discovery dataset) using dichotomous *Toxoplasma* sero-conversion status as the outcome variable. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis.

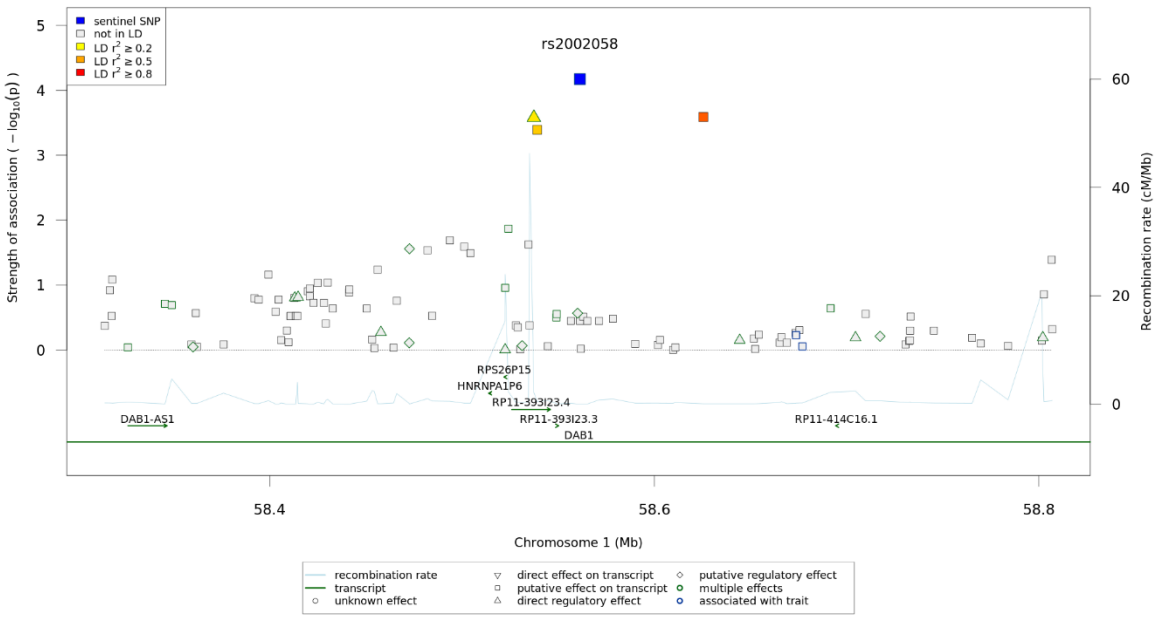
To better define SNPs with nominal significance tagging regions in LD, and hone-in on functional SNPs, we plotted the gene variants using regional association plots (Figure 3). Using these plots we were able to see SNPs of interest resting in coding regions of chromosomes. Within our subset of variants, we were able to find coding regions in seven different chromosomes. The genes that were encoded for by these SNPs included: *CHIAP2*, *NPAS3*, *DAB1*, *CDH18*, *SEMA3D*, *GABBR2*, and *CD44*. The SNP with the most significant association with *T. gondii* infection was rs10857870. This SNP was accompanied by several nominally significant SNPs in LD, and is found to be immediately upstream of *CHIA*.

From the regional plot (Figure 3(e)) it appears that the SNP rs337531 has a direct regulatory effect on the gene *GABBR2* which is required for the formation of GABA receptors

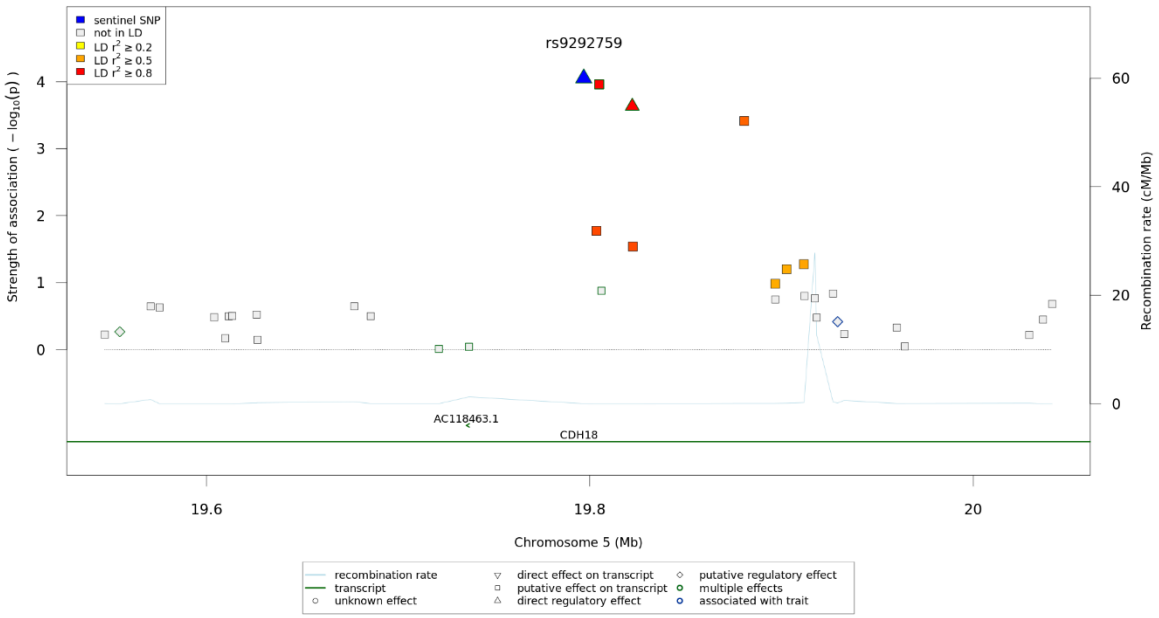
and also plays a role in coupling to G-proteins. Diseases associated with this gene include nicotine dependence and sensory neuropathy (72).



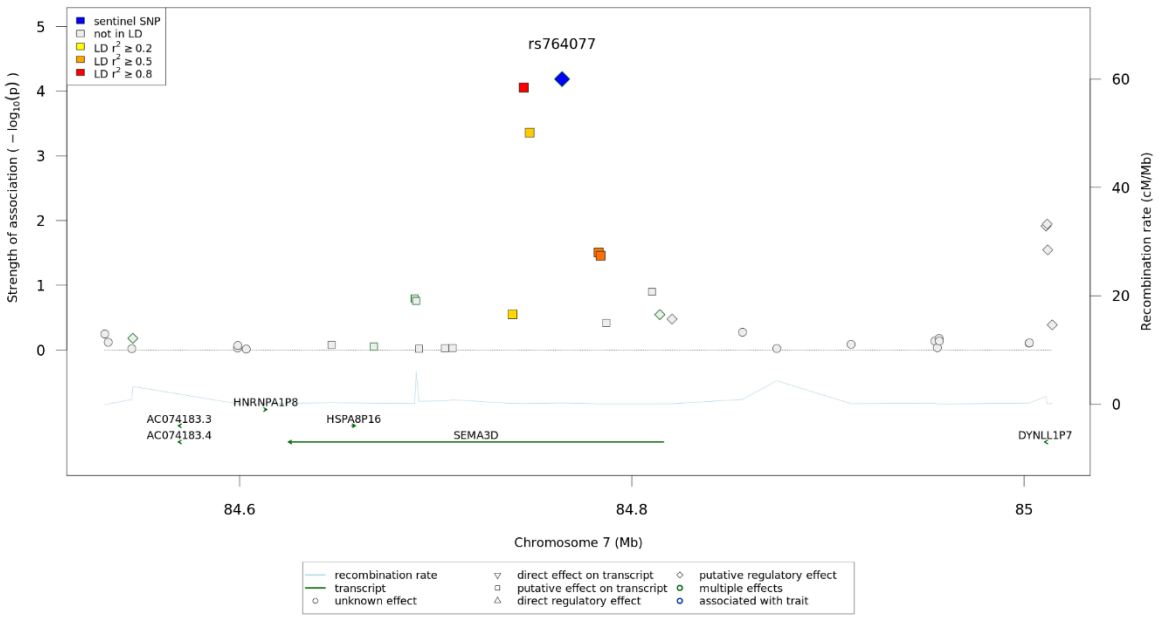
(a) Regional association plot of SNP rs10857870 that is found in gene *CHIAP2* on chromosome 1. Variant appears to have a direct regulatory effect and several associated SNPS that are upstream of the CHIA gene.



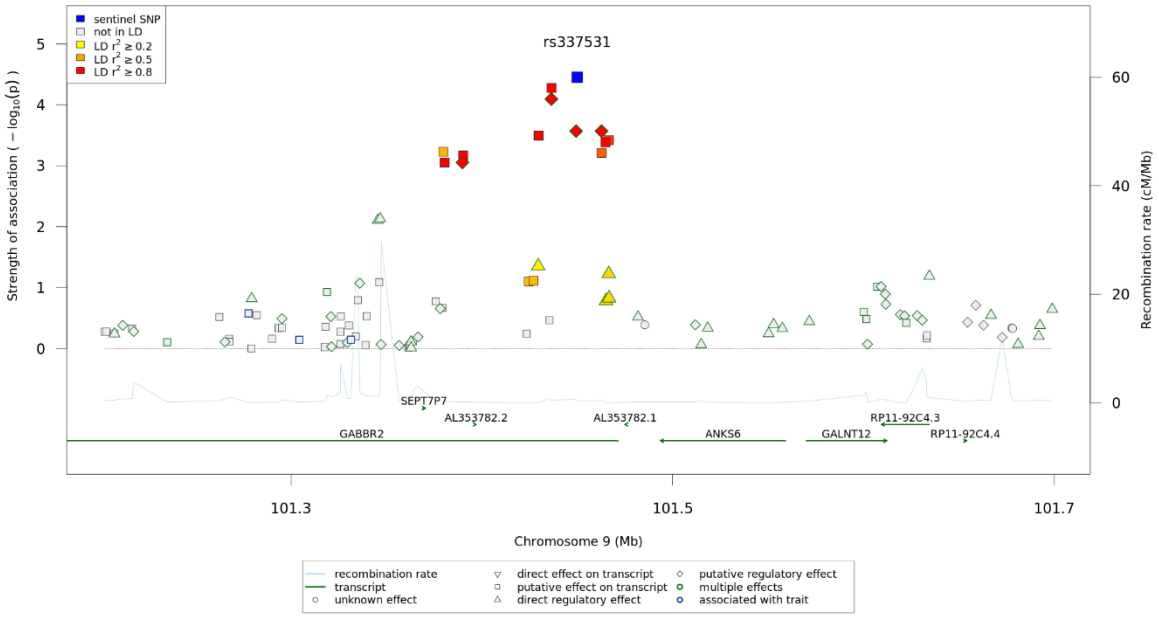
(b) Regional association plot of SNP rs2002058 that is found in gene *DAB1* on chromosome 1. Variant appears to have a putative effect on the transcript.



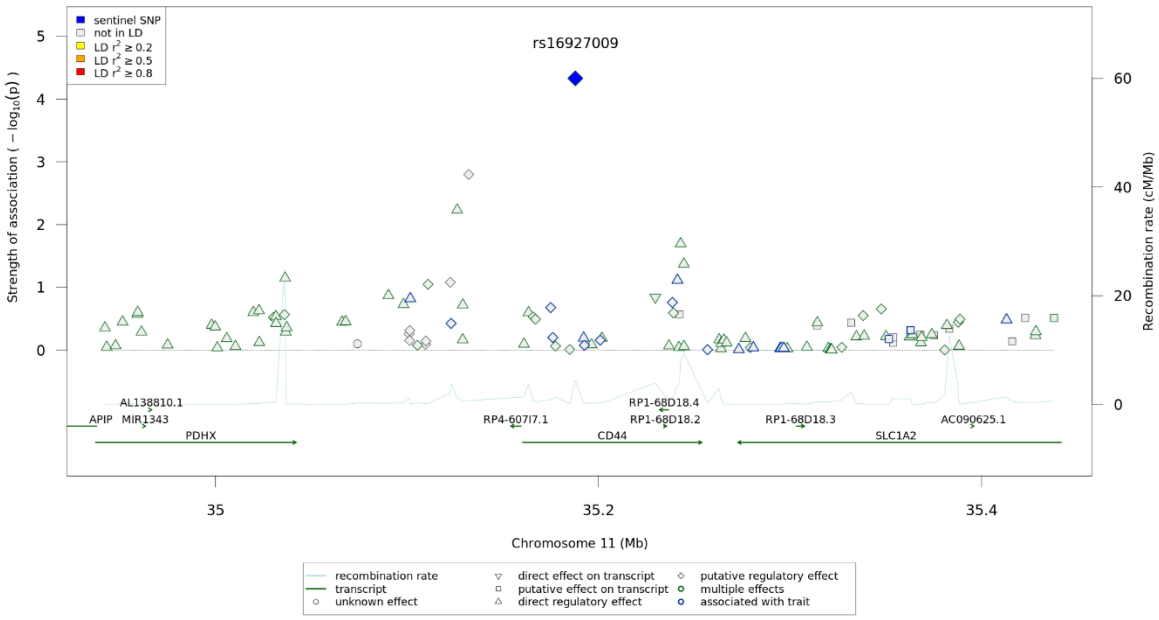
(c) Regional association plot of SNP rs9292759 that is found in close proximity to the *CDH18* gene on chromosome 5. The variant appears to have direct regulatory effect on the gene.



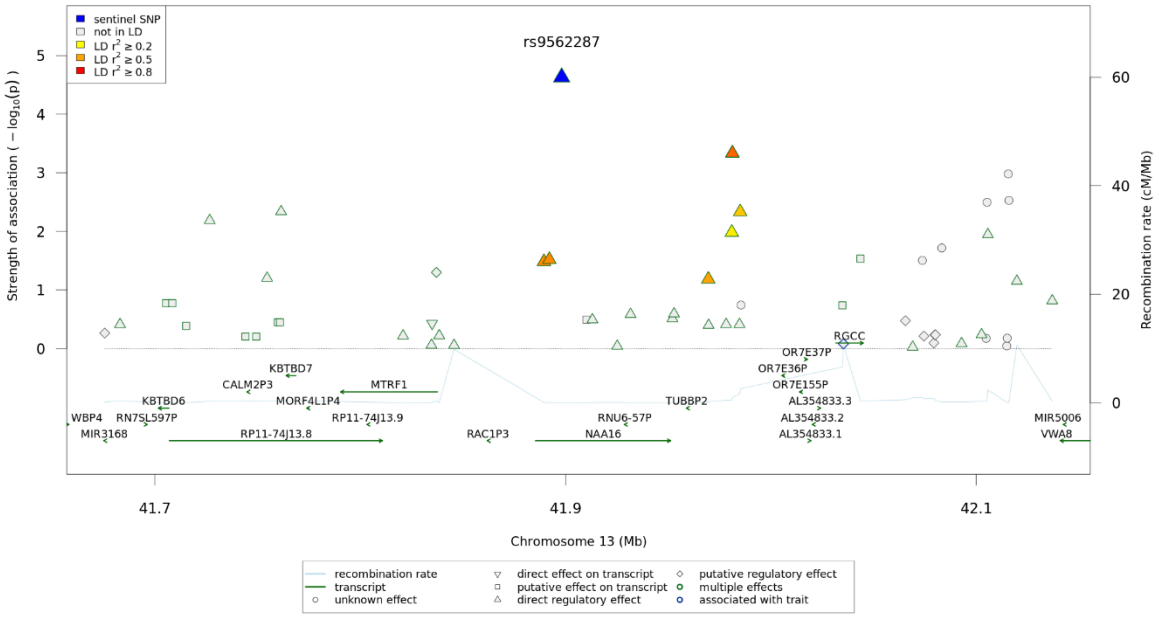
(d) Regional association plot of SNP rs764077 that is found in gene *SEMA3D* on chromosome 7. Variant appears to have a putative regulatory effect on the gene.



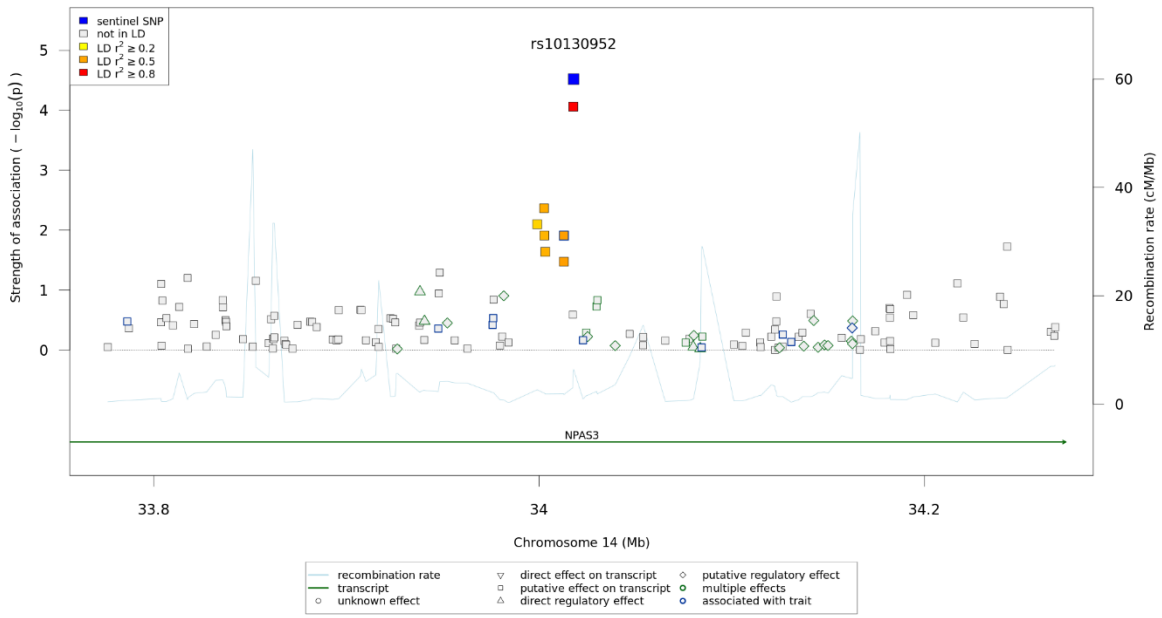
(e) Regional association plot of SNP rs337531 found in gene *GABBR2* on chromosome 9. The variant appears to have a putative effect on the transcript.



(f) Regional association plot of SNP rs16927009 that is found gene *CD44* on chromosome 11. Variant appears to have a putative effect on the transcript obtained from the gene.



(g) Regional association plot of SNP rs9562287 that is found in gene *NAA16* on chromosome 13. Variant appears to have a direct regulatory effect on the gene.



(h) Regional association plot of SNP rs10130952 that is found in the gene *NPAS3* on chromosome 14. This variant appears to have a putative effect on the transcript of the gene.

Figure 3 (a-h): Regional association plot around SNPs with the $p<10^{-4}$ for the Ashkenazi Jewish population when considering sero-conversion status as the outcome variable. The left axis indicates the strength of association ($-\log_{10}$ (P)), the right axis indicates the recombination rate (cM/Mb). The significant p-values are shown on the upper part of the plot. The variant symbols represent functional annotations, and SNPs colors show the pair-wise LD correlations to the sentinel variant based on their r^2. The plot also shows regulatory elements. Chromosome numbers can be found underneath each of the plots. Genes shown here include: (a)C*HIAP2,* (b)*DAB1,* (c)*CDH18,* (d)*SEMA3D,* (e)*GABBR2*, (f)*CD44*, (g)*NAA16*, and (h)*NPAS3*.

We can also see from the plots that the variant rs10857870 is found in a coding region for the gene *CHIAP2* (Figure 3(a)), and may also have a direct regulatory effect on the gene. Despite this only being a pseudo-gene, it is in close proximity to *CHIA,* a gene responsible for the degradation of chitin. The protein coded by *CHIA* can also stimulate interleukin 13 expression, and variations in this gene can lead to asthma susceptibility (73). On chromosome 14, we also can see that there is a SNP (rs10130952) that codes for the gene *NPAS3* (Figure 3(h)). Though this variant is believed the only play a putative effect on the transcript, the encoded protein is localized to the nucleus and is thought play a role as a

regulatory effect on genes involved in neuro-genesis. Abnormalities that affect the coding potential of this gene are associated with mental issues like SCZ and mental retardation (74).

Following the findings from the association plots, we relaxed the threshold to $p<0.001$ to search for other SNPs that may lie in coding regions. This increased our subset of variants from 30 SNP to 346 variants of interest. This group of SNPs was then uploaded into QIAGEN's Ingenuity® Pathway Analysis software to identify which of them may be related to genes that code for proteins (Supplemental Table 1).

The next step was to run an analysis using a quantitative outcome variable for *T. gondii* infection (sero-intensity) to determine the significant genetic variants, the results from this analysis was plotted on a Manhattan plot to assist in identifying regions of interest (Figure 4). Again, none of the obtained SNPs provided a genome wide significant result after performing Bonferroni correction when accounting for multiple testing.

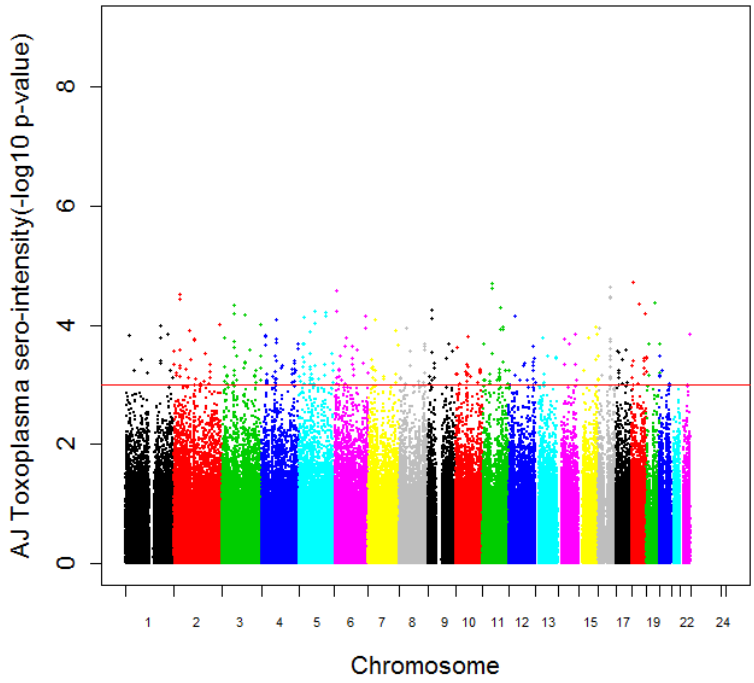
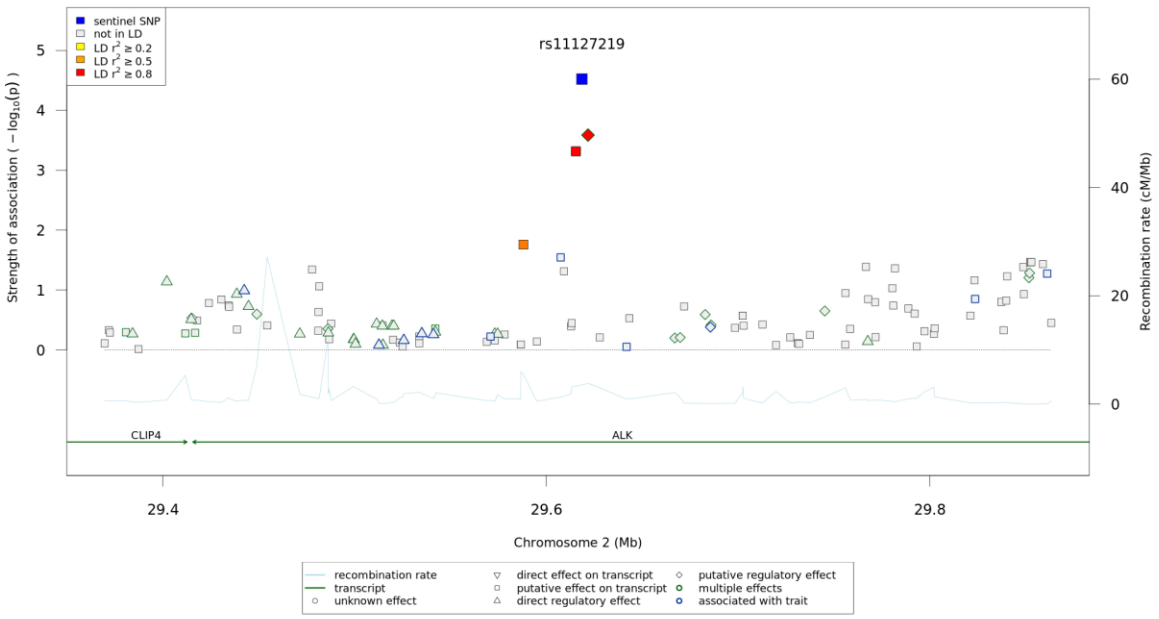
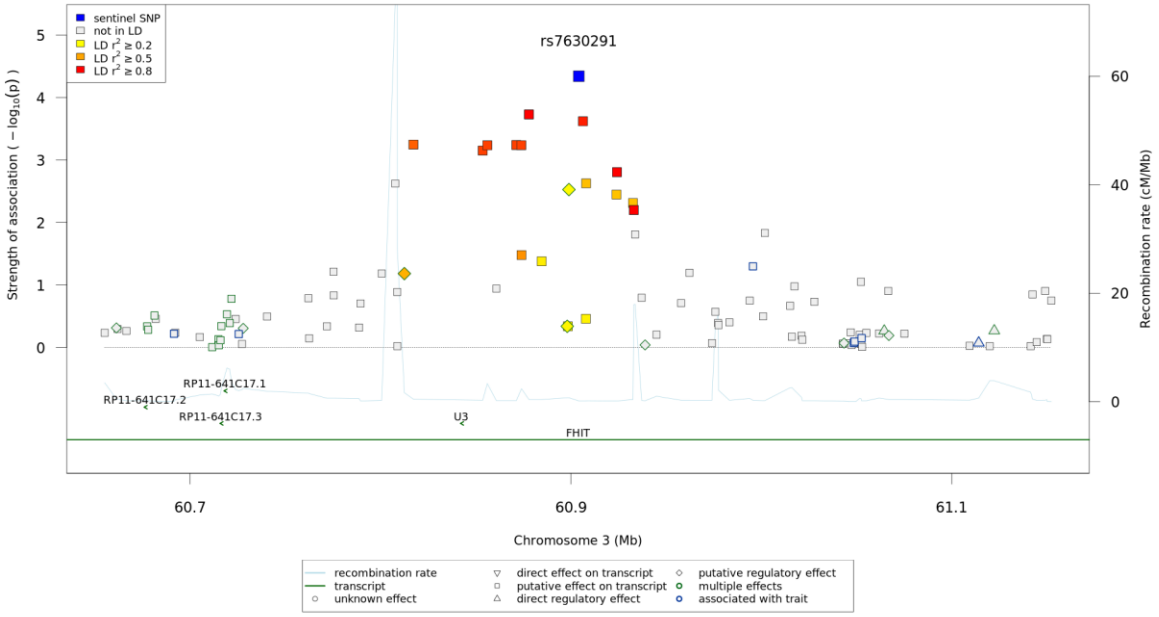


Figure 4: Manhattan plot ($-\log_{10}$ [P-value] genome-wide association plot) of the Ashkenazi Jewish population (discovery dataset) using continuous *Toxoplasma* sero-intensity as the outcome variable. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis.

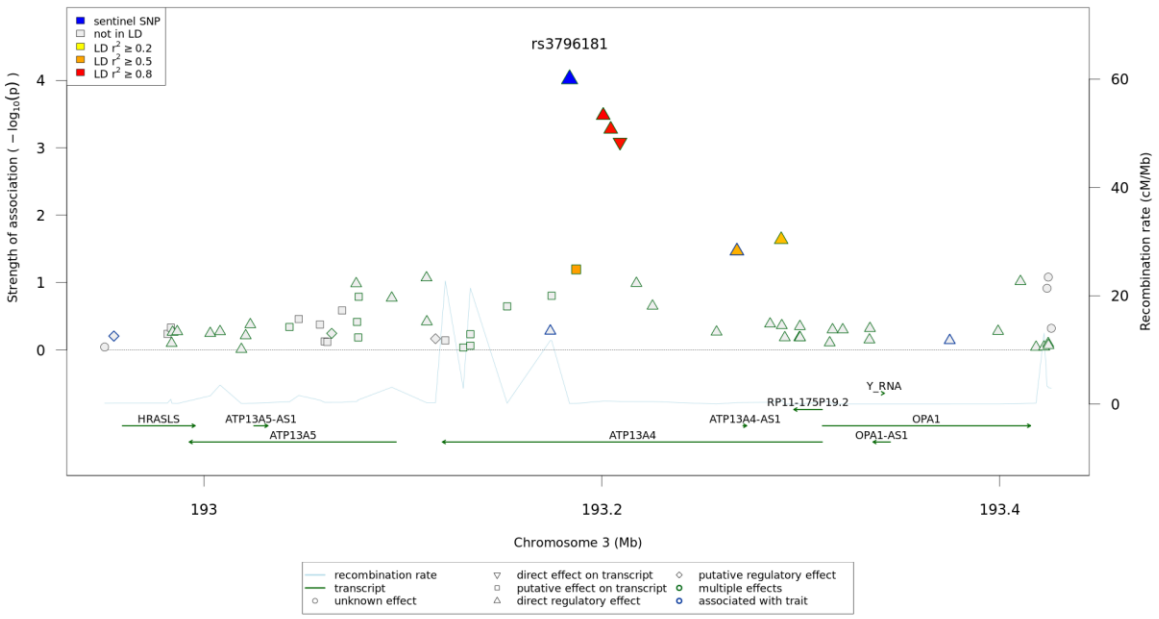
We first focused on all the SNPs that were below our initial threshold of $p<10^{-4}$, this group consisted of 31 SNP variants. Regional plots of these SNPs were created and coding genes were identified (Figure 5). Following this, we relaxed the threshold to p<0.001 to search for other coding regions that may be of interest. This group of 346 SNPs were then uploaded into Ingenuity® IPA software and genes within 2 Kb upstream or 0.5 Kb downstream of the SNP were identified (Supplemental Table 2).



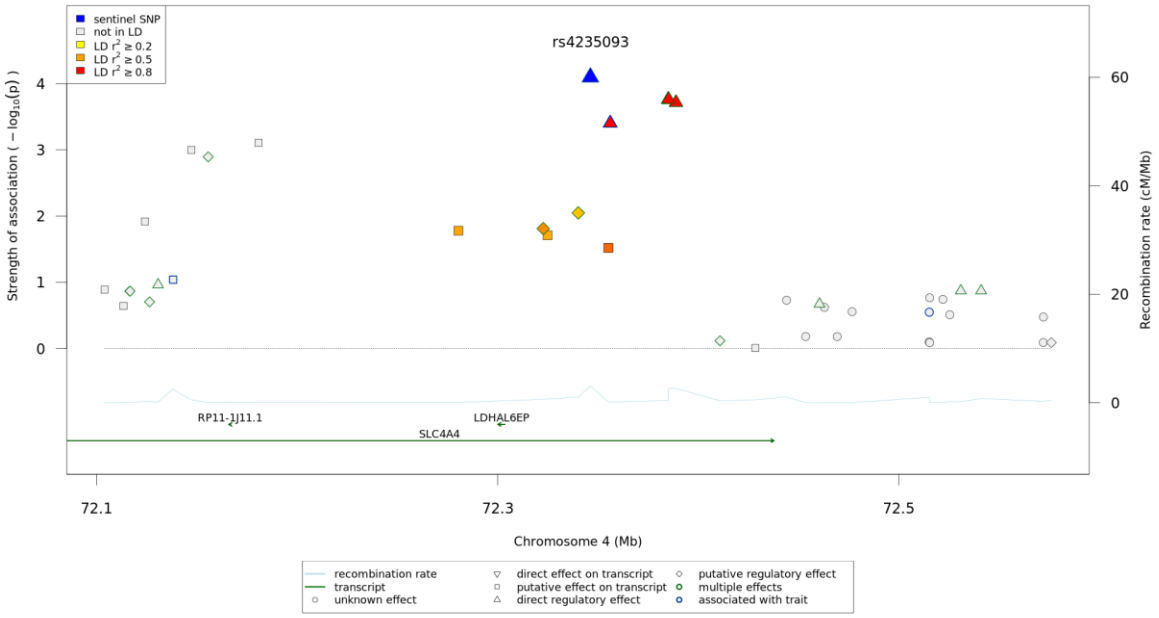
(a) Regional association plot of SNP rs11127219 that is found in the gene *ALK* on chromosome 2. This variant appears to have a putative effect on the transcript of the gene.



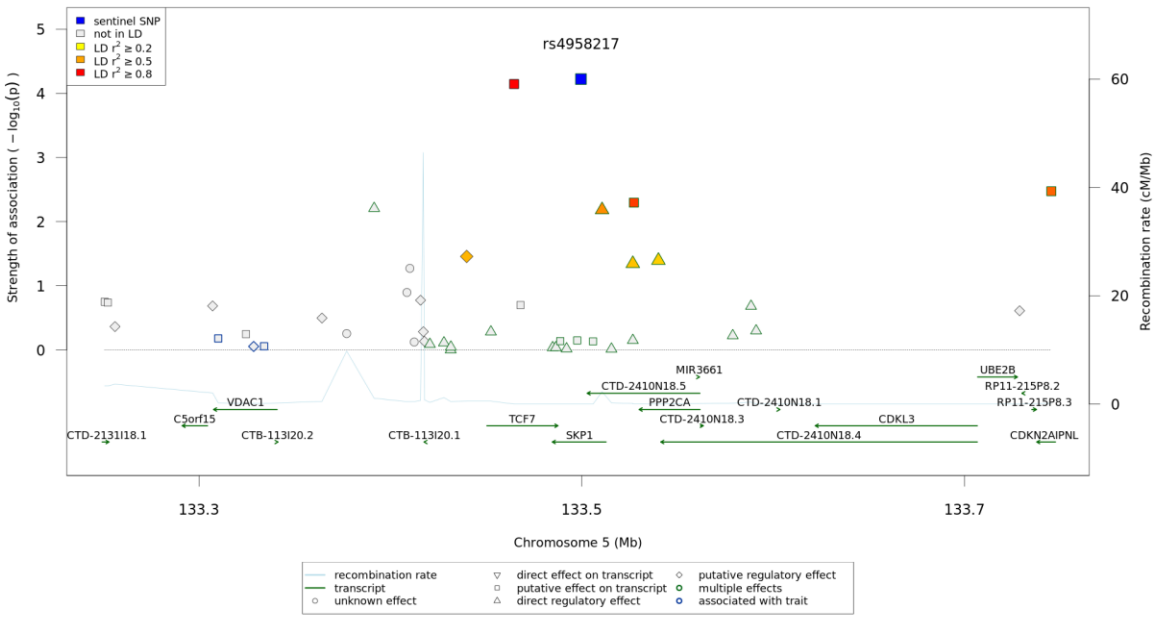
(b) Regional association plot of SNP rs7630291 that is found in the gene *FHIT* on chromosome 3. This variant appears to have a putative effect on the transcript of the gene.



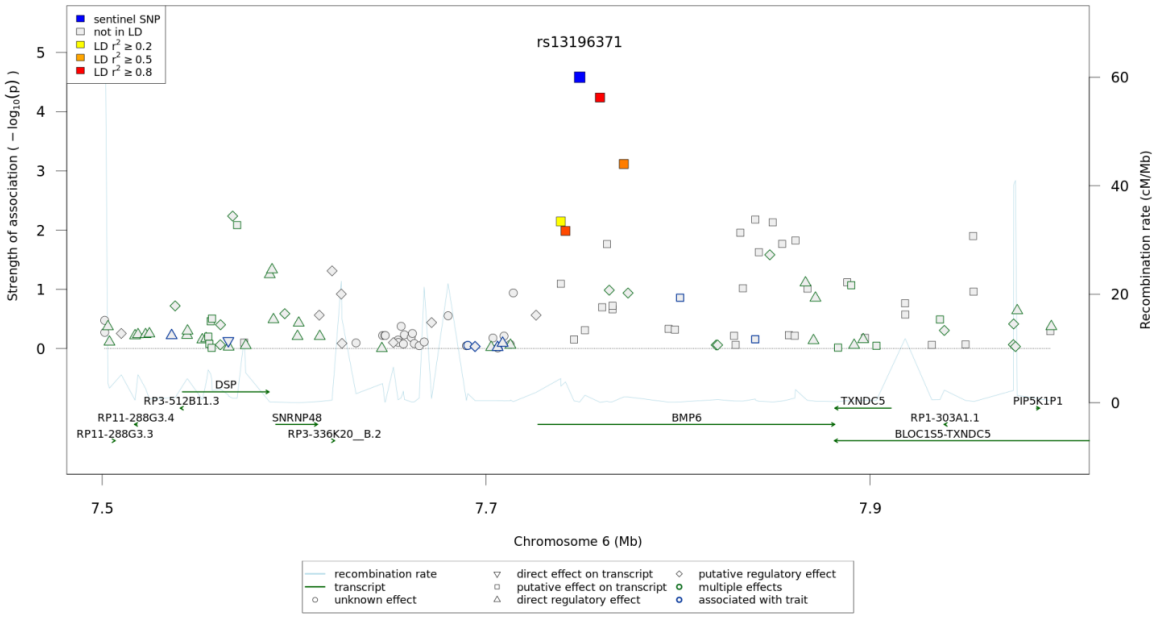
(c) Regional association plot of SNP rs3796181 that is found in the coding region of gene *ATP13A4* on chromosome 3. This variant appears to have a direct regulatory effect on the transcript of the gene.



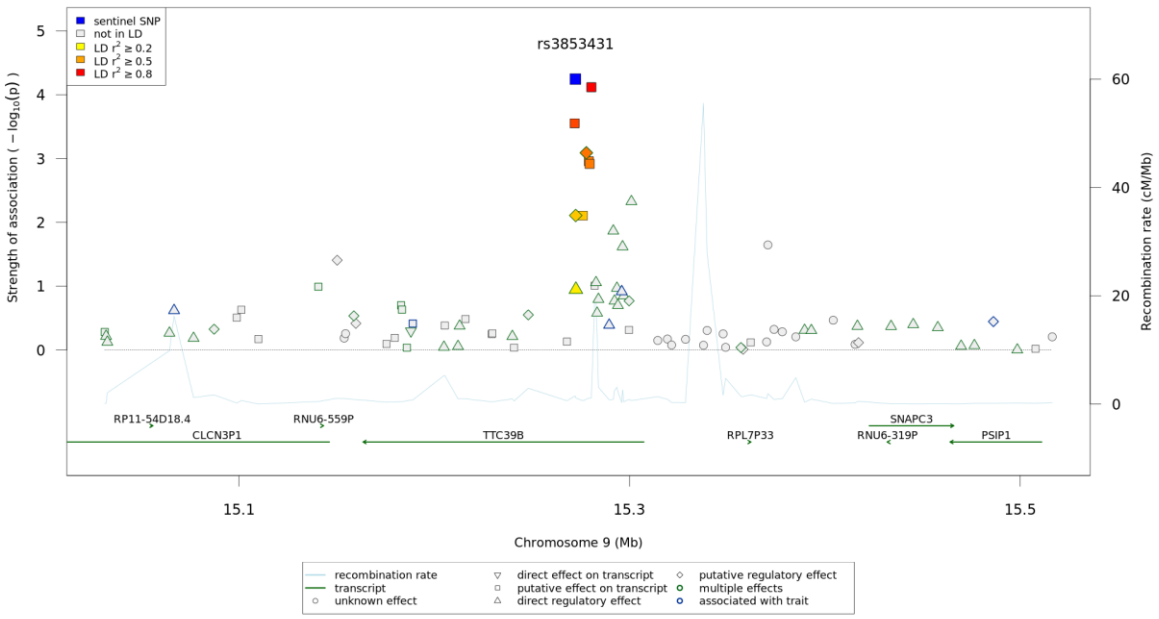
(d) Regional association plot of SNP rs4235093 that is found in the coding region of gene *SLC4A4* on chromosome 4. This variant appears to have a direct regulatory effect on the transcript of the gene.



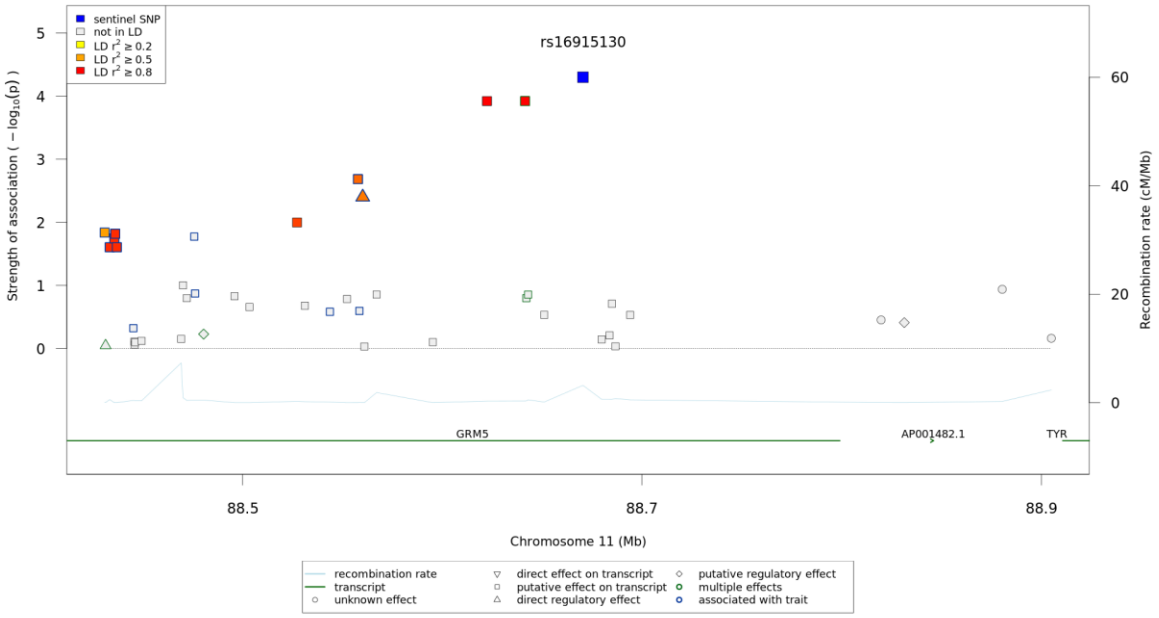
(e) Regional association plot of SNP rs4958217 that is found in the coding region of gene *SKP1* on chromosome 4. This variant appears to have a putative effect on the transcript of the gene.



(f) Regional association plot of SNP rs13196371 that is found in the coding region of gene *BMP6* on chromosome 6. This variant appears to have a putative effect on the transcript of the gene.



(g) Regional association plot of SNP rs3853431 that is found in the coding region of gene *TTC39B* on chromosome 9. This variant appears to have a putative effect on the transcript of the gene.



(h) Regional association plot of SNP rs16915130 that is found in the coding region of gene *GRM5* on chromosome 11. This variant appears to have a putative effect on the transcript of the gene.

Figure 5 (a-h): Regional association plot around SNPs with the p<10^{-4} for the Ashkenazi Jewish population when considering sero-intensity as the outcome variable. The left axis indicates the strength of association (-$\log_{10}$ (P)), the right axis indicates the recombination rate (cM/Mb). The significant p-values are shown on the upper part of the plot. The variant symbols represent

functional annotations, and SNPs colors show the pair-wise LD correlations to the sentinel variant based on their r^2. The plot also shows regulatory elements. Chromosome numbers can be found underneath each of the plots. Genes that are shown here are: (a)*ALK,* (b)*FHIT,* (c)*ATP13A4,* (d)*SLC4A4,* (e)*SKP1,* (f)*BMP6,* (g)*TTC39B,* and (h)*GRM5*.

Looking at the regional association plots obtained from the analysis, we can identify a few genes that appear to have a direct regulatory effect on a gene. The gene *ATP13A4*, is a protein coding gene that is related to pathways that transport glucose, sugars, and cationic molecules (Figure 5(c)). It has been related to the development of benign epilepsy and speech-language disorders (75). Another variant (rs4235093) that is predicted to have a direct regulatory effect on falls into the *SLC4A4* gene region (Figure 5(d)). This gene encodes for sodium bicarbonate co-transporter involved in the regulation of bicarbonate secretion and absorption and intracellular pH (76). Diseases that are related to the dysfunction of the gene include autosomal recessive proximal renal tubular acidosis.

The SNP rs7630291 appears to have a putative effect on the transcript from the gene *FHIT* (Figure 5(b)). This gene encodes for diadenosine 5',5'''-P1,P3-triphosphate hydrolase which is involved in the metabolism of purines (77). The gene contains the fragile site *FRA3B* on chromosome 3, that when damaged by carcinogens can lead to translocations and aberration of the transcripts that have been suggested to be associated with breast and lung cancer. Another SNP of interest represented in Figure 5(h) by rs16915130 lies in the coding region of *GRM5*. This gene is known to encode a member of the G-protein coupled receptor 3 protein family. It is believed that this protein may be involved in the regulation of neural network activity, synaptic plasticity and play a role in the neuro-inflammatory pathways (78).

Once the analyses of the dichotomous and continuous outcomes were completed, we stratified our population based on SCZ status. This was done because we wanted to

determine which of the genetic variants obtained from the analyses were due to interactions with the enriched SCZ population and which of the genetic variants were attributed to increased risk of infection with *T. gondii* itself. The population was separated into cases who are individuals diagnosed with SCZ or schizotypal personality disorder and controls who were free of any psychiatric disorders.

There was a total of 519 cases of individuals diagnosed with SCZ or schizotypal personality disorder, and a total of 271 controls who had no history of depression, mania, psychosis, psychiatric hospitalization, depression, or suicide attempts. We first started the analysis of the stratified population by classifying *T. gondii* sero-conversion status as a dichotomous outcome variable. Based on the obtained results and the Manhattan plots created from this information (Figure 6), there were not many SNPs that met our initial threshold of $p<10^{-4}$.

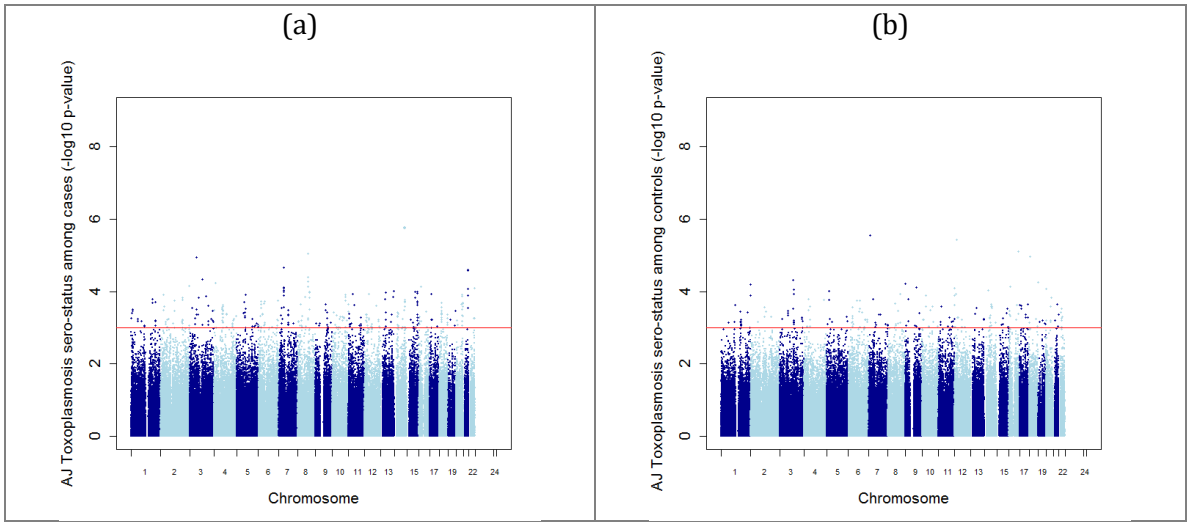


Figure 6: Manhattan plots ($-\log_{10}$ [P-value] genome-wide association plot) of the Ashkenazi Jewish population (discovery dataset) stratified on SCZ diagnosis using a dichotomous *Toxoplasma* sero-conversion status outcome variable. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis. (a) Cases consisted of 600 patients who were diagnosed with SCZ or schizotypal personality disorder. (b) Controls consisted of 508 unrelated individuals who were free of any mental illnesses.

From the group that consisted of only cases, 22 SNPs met this criteria and only six of these SNPs were found in coding regions for the genes: *ARHGEF3, KCNJ6, SHISA9*, and *KIAA0930*. The gene *KCNJ6* encodes a G-protein coupled potassium channel that has been associated with Keppen-Lubinsky Syndrome, a rare condition characterized by severe developmental delay, facial dysmorphism, and intellectual disability (79). A gene like *ARHGEF3* is involved in many cellular processes due to it having a Rho-like GTPase structure. There is evidence indicating that dysfunction in this gene plays a role in the progression of osteoporosis (80).

In the group that contained controls who were not diagnosed with SCZ or other mental disorders, only 16 SNPs met our criteria of $p<10^{-4}$. Among these, there were variants that were found to code for *NXPH1, DMRT2, AHCTF1, MAF, SLC6A13, NPAS3* and *ZBTB20*.

Once we identified related genes, we relaxed our threshold criteria to include all SNPs with $p<0.001$. The protein coded by the gene *MAF* is a transcription factor that, depending on the site of action can act as a transcriptional activator or repressor. Defects of this gene have been linked to many different forms of cataracts, including juvenile-onset pulverulent cataract (81). The gene *NXPH1* encodes for a secreted protein used as a signaling molecule and can act as a promoter for adhesion between dendrites and axons (82).

After these findings, we once again lowered our threshold to any SNP with $p<0.001$, in the stratified dataset there were a total of 338 SNPs among cases that met this criteria, and 212 SNPs among controls that fell into this range. These SNPs were input into Ingenuity® IPA software and variants in expressed regions were identified (Supplemental Tables 3 & 4).

Following the use of a dichotomous outcome, we conducted the same stratification to separate cases from controls, and performed the association analyses using the continuous toxoplasmosis sero-intensity as the outcome variable. Looking at the Manhattan plots (Figure 7) constructed based on the obtained results, there were a similar number of SNPs among the cases that met our threshold criteria of $p<10^{-4}$ compared to the dichotomous classification. Compared to the 16 SNPs from the dichotomous classification among controls, using sero-intensity as the outcome variable yielded 35 SNPs that fall below the threshold level.

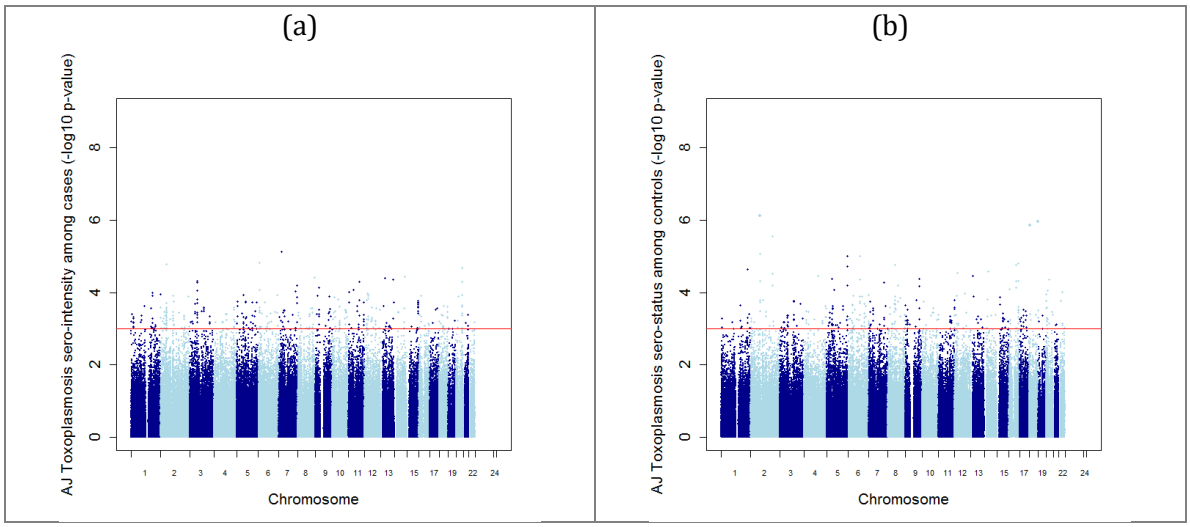


Figure 7: Manhattan plots ($-\log_{10}$ [P-value] genome-wide association plot) of the Ashkenazi Jewish population (discovery dataset) stratified on SCZ diagnosis using a continuous *Toxoplasma* sero-intensity outcome variable. The red line indicates a threshold value of $p<0.001$, those above the line were included in further analysis. (a) Cases consisted of 519 patients who were diagnosed with SCZ or schizotypal personality disorder. (b) Controls consisted of 271 unrelated individuals who were free of any history of depression, mania, psychosis, psychiatric hospitalization, depression, or suicide attempts.

In the case only group 16 SNPs met our initial inclusion criteria, of these only 10 of the variants were found to be in genes. These genes included *OFCC1, PPP1R21, TRAPPC9,*

GRM5, DPP6, BMP6, TTC17, FHIT, and *OR2AG2*. The gene *DPP6* codes for a single pass membrane protein that binds specific voltage-gated potassium channels. Variations in the gene have been associated with increased susceptibility to amyotrophic lateral sclerosis and may even contribute to microcephaly and mental retardation (83). Protein coding genes such as *OFCC1* are associated with diseases including orofacial cleft and chronic tic disorder (84). It is worth noting that the genes GRM5, *BMP6*, and *FHIT* also appeared when we conducted the analysis using the entire discovery population.

When looking at the 35 SNPs in the controls that met the threshold p-value, we found only 10 of the variants were in of genes. These genes were: *KCNIP1, MAF, NPAS3, SOX5, TTC30A, SVEP1, HCN1, MAFTRR, PDE11A,* and *RANBP17*. The gene *SOX5* encodes for a transcription factor that is involved in embryonic development, and *RANBP17* plays a role in transports of protein through nuclear pores. None of these genes appeared among the SNPs below p-value of 10^{-4} during the analysis of the entire population. We proceeded by lowering the threshold value to p=0.001; among cases, 361 SNPs met this inclusion criteria and among controls, 338 SNPs. These were then input into the gene finding software and variants in close proximity to expressed genes were identified (Supplemental Tables 5 & 6).

Following the completion of these analyses, we compiled all the SNPs that had a p<0.001 and looked to see if there were any genes that overlapped between the various tests. Table 2 contains genes found in two or more of the associations. Based on the findings, there appear to be various genes that overlap in two or more of the analyses. Genes that were found in three of the separate analyses were *AGBL1, BOC, CSMD1, FAM110C, FHIT, KCNJ6, LOC101927630, LOC101928505, LOC105371978, LOC105374693, OR2AG2, SKP1, SRRM4*, and *TCF7*. Even more remarkable, there were genes that appeared in four of the association analyses, these genes included *ALK, LINGO2, LOC101929468, LOC105377899,*

NPAS3, PDE1C, RBFOX1, SHISA6, and *SLC4A4*. The gene *ALK* encodes for a receptor tyrosine kinase and has been found to be amplified in a types of tumors such as lymphomas, neuroblastoma and small cell lung cancer (85). Another overlapping gene of interest is *LINGO2*, this is a protein coding gene that has been classified as a susceptibility gene for Parkinson's disease (86). Of the genes that appeared in four of the analyses, *NPAS3* was the only one to appear in all three of the associations using a dichotomous Toxoplasmosis sero-conversion status.

Table 2: Summary of the overlap between genes among two or more of the genome-wide association analyses carried out on the Ashkenazi Jewish population among SNPs with p<0.001. Analyses included the whole population, or was stratified by SCZ status into cases (SCZ or schizotypal personality disorder) or controls (no history of depression, mania, psychosis, psychiatric hospitalization, depression, or suicide attempts).

	Dichotomous Toxoplasmosis sero-conversion status			Continuous Toxoplasmosis sero-intensity		
Gene	**Whole pop**	**Cases**	**Controls**	**Whole pop**	**Cases**	**Controls**
AGBL1			x	x	x	
ALK	x	x		x	x	
ANKRA2		x			x	
ARHGEF10L				x	x	
ARHGEF3		x		x		
ATP13A4	x			x		
BACH2				x	x	
BMP6				x	x	
BOC			x	x		x
BRINP1				x	x	
C5orf17				x		x
CAMK1D		x			x	
CD1C			x			x
CD44	x	x				
CDH18	x			x		
CHIAP2	x	x				
CMYA5				x	x	
COL19A1	x					x
CSMD1	x			x	x	
DAB1	x		x			

DDX60L				x	x	
DMRT2			x			x
DNAH10	x			x		
EML5	x	x				
FAM110C	x			x	x	
FHIT	x			x	x	
GABBR2	x	x				
GRM5				x	x	
HIVEP2			x			x
HPGD	x	x				
IL1RL2				x	x	
ITGA11	x	x				
ITPR2	x			x		
KCNIP1				x		x
KCNJ6	x	x			x	
KIF1B	x	x				
LINC00400*				x	x	
LINC01114*				x	x	
LINGO2	x	x		x	x	
LOC101927412				x	x	
LOC101927630	x			x		x
LOC101928505		x		x	x	
LOC101929028	x	x				
LOC101929468	x		x	x	x	
LOC105369459	x			x		
LOC105370345		x			x	
LOC105370822			x			x
LOC105371307				x		x
LOC105371887	x		x			
LOC105371978			x	x		x
LOC105374693	x			x		x
LOC105374694	x			x		
LOC105375001				x	x	
LOC105375557				x	x	
LOC105376408	x			x		
LOC105377899	x		x	x		x
LOC105377936				x		x
LOC284898			x			x
LPP		x	x			
LSAMP				x	x	
MACROD2				x	x	

MAF			x			x
MAGI2		x		x		
MAST4	x	x				
MUC16			x	x		
MYO1E	x	x				
NAA16	x	x				
NALCN-AS1	x					x
NKAIN3			x			x
NPAS3	x	x	x			x
NRG1		x		x		
NTN4	x	x				
NUP50-AS1		x			x	
OFCC1				x	x	
OR2AG2		x		x	x	
P3H2-AS1			x			x
PDE11A		x				x
PDE1C	x		x	x	x	
PDE4D		x		x		
PEPD			x			x
PPFIBP1	x			x		
PPM1A	x		x			
PPM1H	x		x			
PSMD1				x	x	
RAB3C				x	x	
RBFOX1		x	x	x	x	
SEC16B				x	x	
SETBP1				x	x	
SHISA6	x		x	x		x
SIRPB1				x		x
SKP1	x			x		x
SLC4A4	x	x		x	x	
SLC6A13			x			x
SOX5			x			x
SRRM4	x	x		x		
TCF7	x			x		x
TMEM114	x		x			
TMEM135				x	x	
TPH2				x	x	
TRAF3IP2	x		x			
TRAM2				x	x	
TRIM60		x			x	

TRPC6	x	x				
TTC30A				x		x
VPS33A	x				x	
XYLT1		x			x	
ZBTB20			x			x

* Non-coding DNA sequence, long intergenic non-protein coding RNA.

Replication dataset

Following the analysis of our discovery dataset, we performed genome-wide association analyses using the Grady Trauma Project population. We started by first conducting an association analysis with a dichotomous classification of *T. gondii* sero-conversion as the outcome variable. Once again, none of the SNPs provided a genome-wide significant result after accounting for multiple testing. After creating the Manhattan plot (Figure 8), we were able to visualize regions that may be of interest and combine this with the threshold $p<10^{-4}$ to obtain a list of variants that we then used to determine if they were in close proximity to expressed gene.

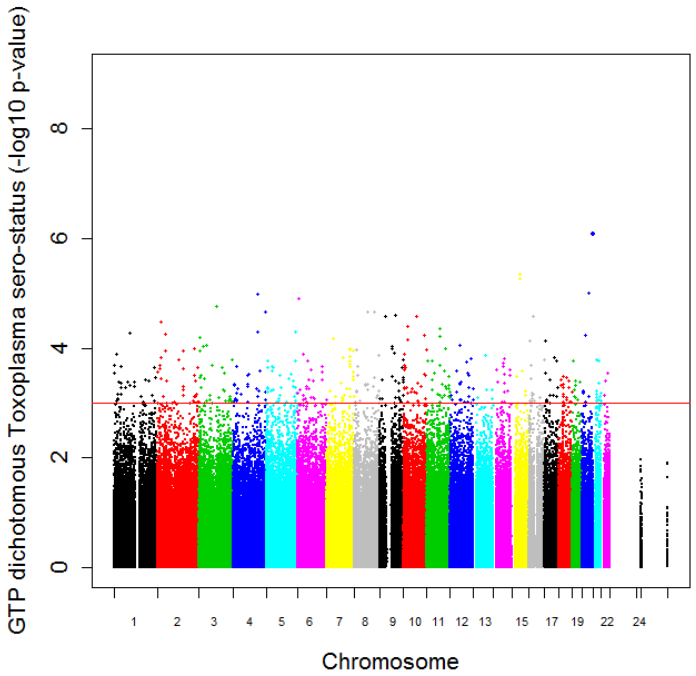
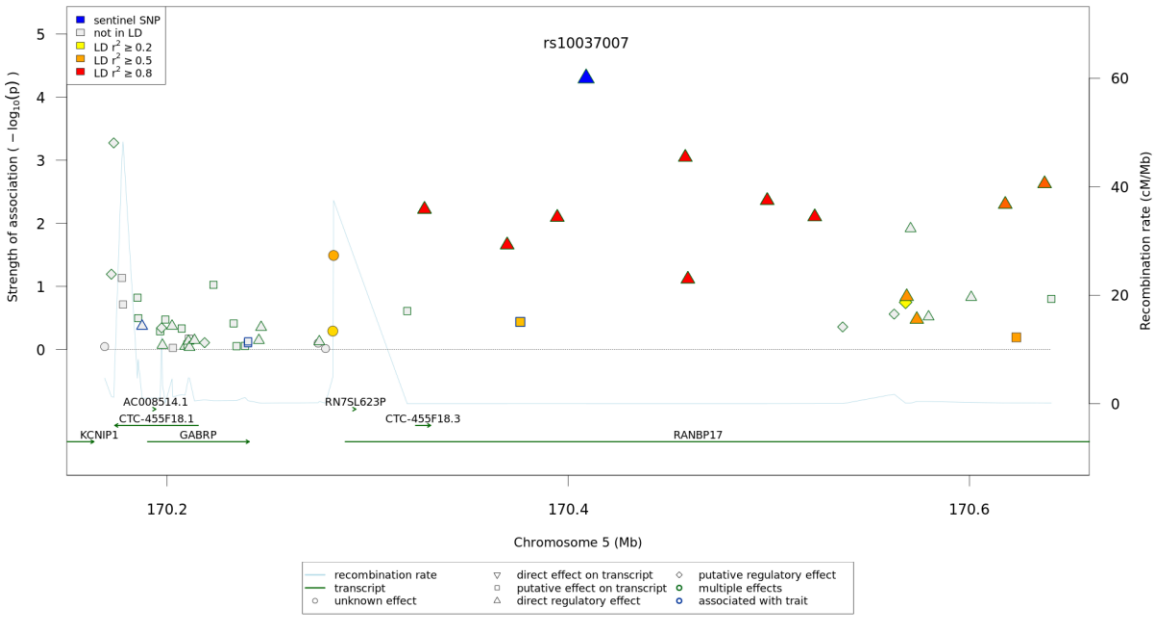
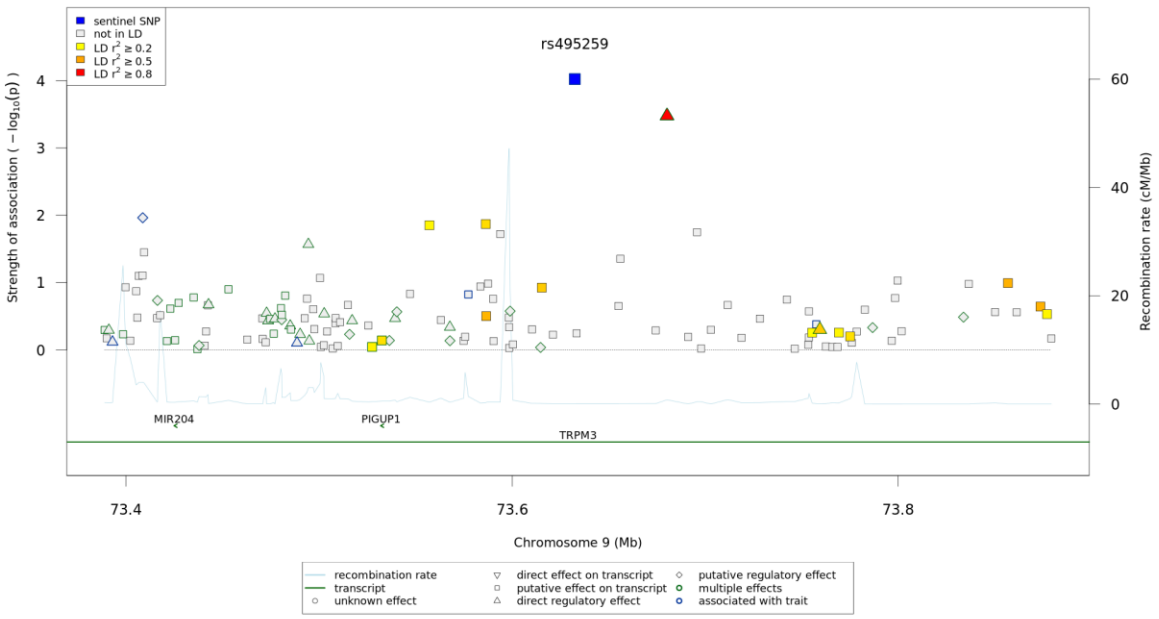


Figure 8: Manhattan plots ($-\log_{10}$ [P-value] genome-wide association plot) of the Grady Trauma Project population (replication dataset) using Toxoplasmosis sero-conversion as the outcome variable. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis.

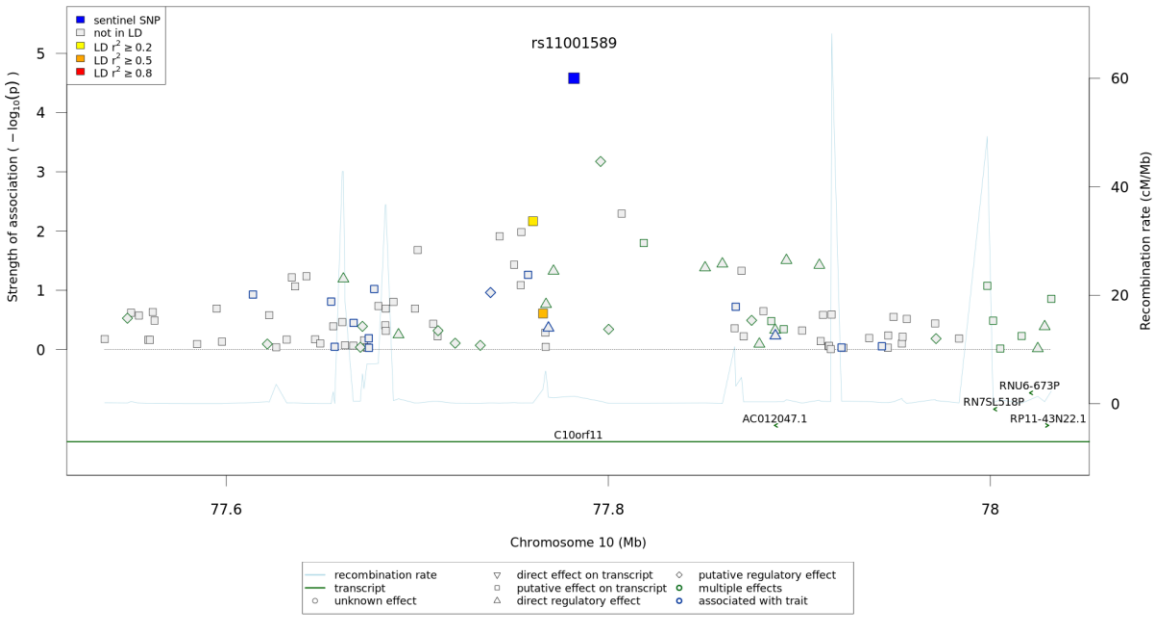
From the 34 SNPs that were below our chosen p-value, only 12 were found in coding regions of the genes *PATL2, SYCP2L, SEMA4D, C10orf11, PRKCB, ARHGAP21, RANBP17, PKN2-AS1, KIF16B, NEBL, ANO10* and *TRPM3*. Regional association plots were then made for the SNPs to visualize their coding regions (Figure 9). From these plots we can see that there are a few expressed genes that have multiple SNPs in linkage disequilibrium.



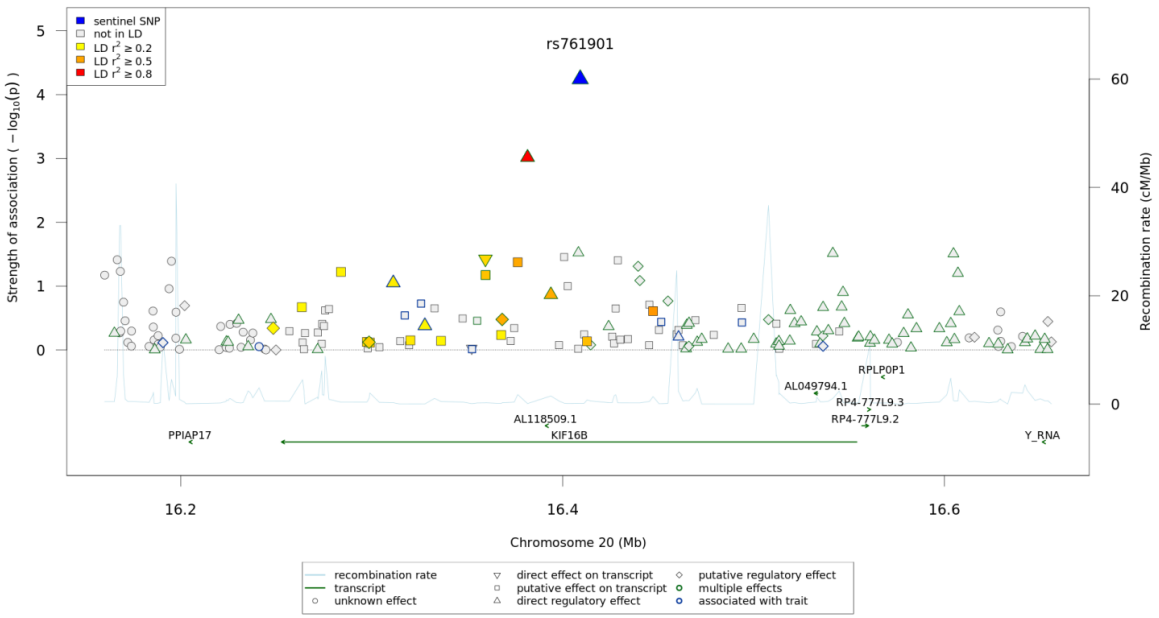
(a) Regional association plot of SNP rs10037007 that is found in the coding region of gene *RANBP17* on chromosome 5. This variant appears to have a direct regulatory effect on gene.



(b) Regional association plot of SNP rs495259 that is found in the coding region of gene *TRPM3* on chromosome 9. This variant appears to have a putative effect on the transcript of the gene.



(c) Regional association plot of SNP rs11001589 that is found in the coding region of gene *C10orf11* on chromosome 10. This variant appears to have a putative effect on the transcript of the gene.



(d) Regional association plot of SNP rs761901 that is found in the coding region of gene *KIF16B* on chromosome 20. This variant appears to have a direct regulatory effect on the gene.

Figure 9 (a-d): Regional association plot around SNPs with the $p<10^{-4}$ for the Grady Trauma Project population when considering sero-conversion status as the outcome variable. The left axis indicates the strength of association ($-\log_{10}$ (P)), the right axis indicates the recombination rate (cM/Mb). The significant p-values are shown on the upper part of the plot. The variant symbols represent functional annotations, and SNPs colors show the pair-wise LD correlations to the sentinel variant based on their r^2. The plot also shows regulatory elements. Chromosome numbers can be found underneath each of the plots. Genes that are shown here are: (a)*RANBP17*, (b)*TRPM3*, (c)*C10orf11*, and (d)*KIF16B*.

From the regional plots we can see that the variant found on the gene *RANBP17*, seems to be quite interesting (Figure 9(a)). Not only is it believed to have a direct regulatory effect on the gene, but all the other SNPs in its proximity seem to have the same effect on the gene. As mentioned earlier, this gene encodes for a protein that assists transport of protein and large RNA through nuclear pores. Another variant that appears to have a direct regulatory effect lies on the gene *KIF16B* (Figure 9(d)). This gene encodes for a protein that is involved in intracellular trafficking, and receptor recycling and degradation (87). Other genes of interest that arose from this analysis include *TRPM3*. This gene encodes for a cation

channel that is important in cellular calcium signaling and may play a role in increasing susceptibility to ocular diseases (88).

Following this, we then considered a threshold level at p<0.001 to search for other SNPs that might be near other coding regions that may be of interest. This gave us a list of 399 SNPs that were then input into Ingenuity® IPA software where variants close to protein coding gene regions were identified, the entire list of genes and SNPs can be found in Supplemental Table 7.

After completing the association analysis with a dichotomous sero-conversion status outcome variable, we wanted to test the association using the continuous sero-intensity variable. The results of the analysis was used to create a Manhattan plot to better visualize possible regions of interest (Figure 10). None of the SNPs taken from the results were significant after performing Bonferroni correction. Regional plots were created to visualize variants in regions and what effect they might have on the gene (Figure 11).

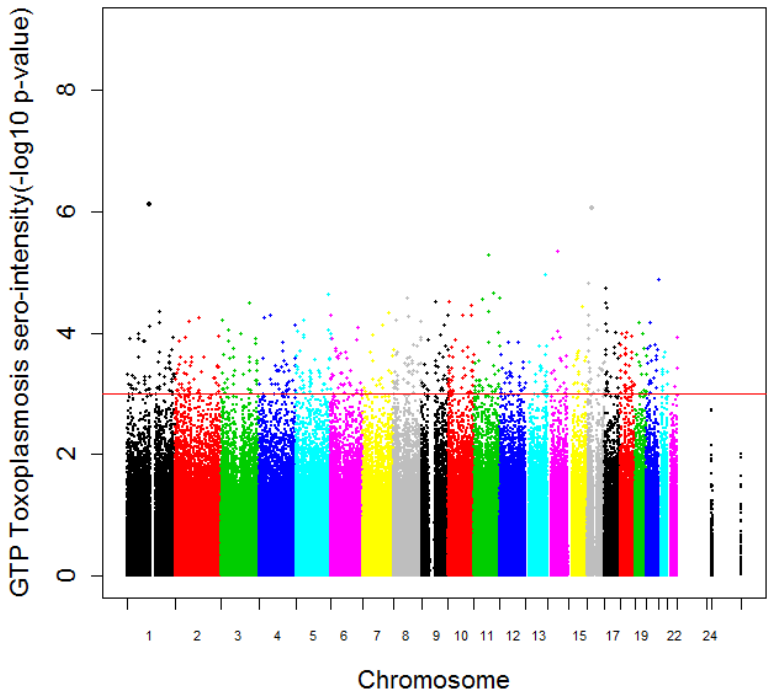
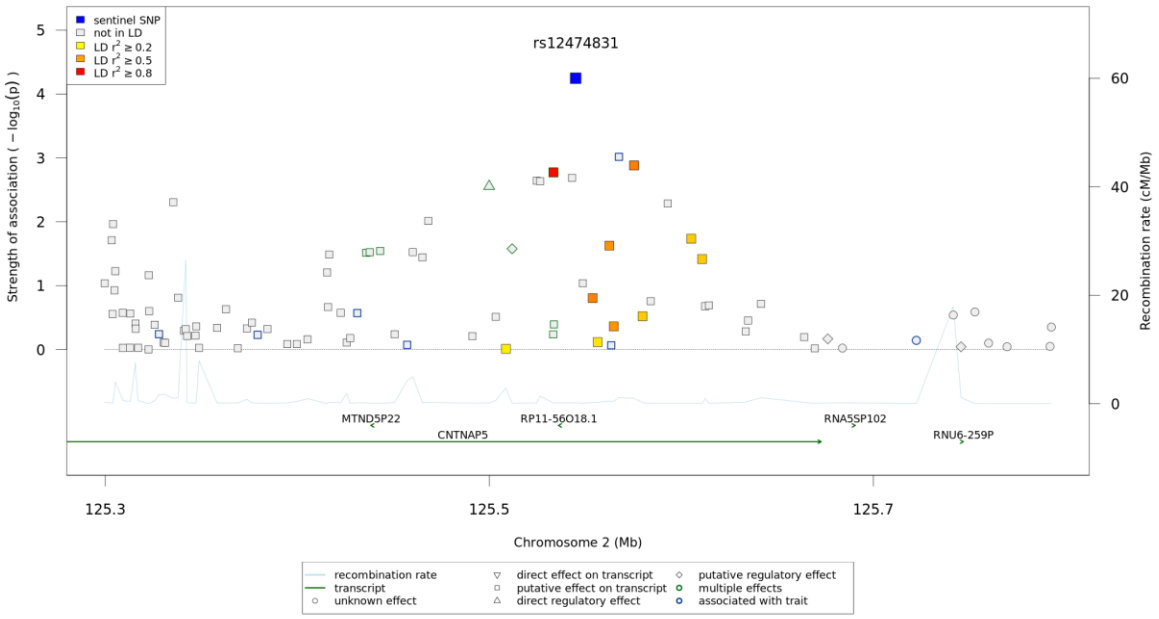
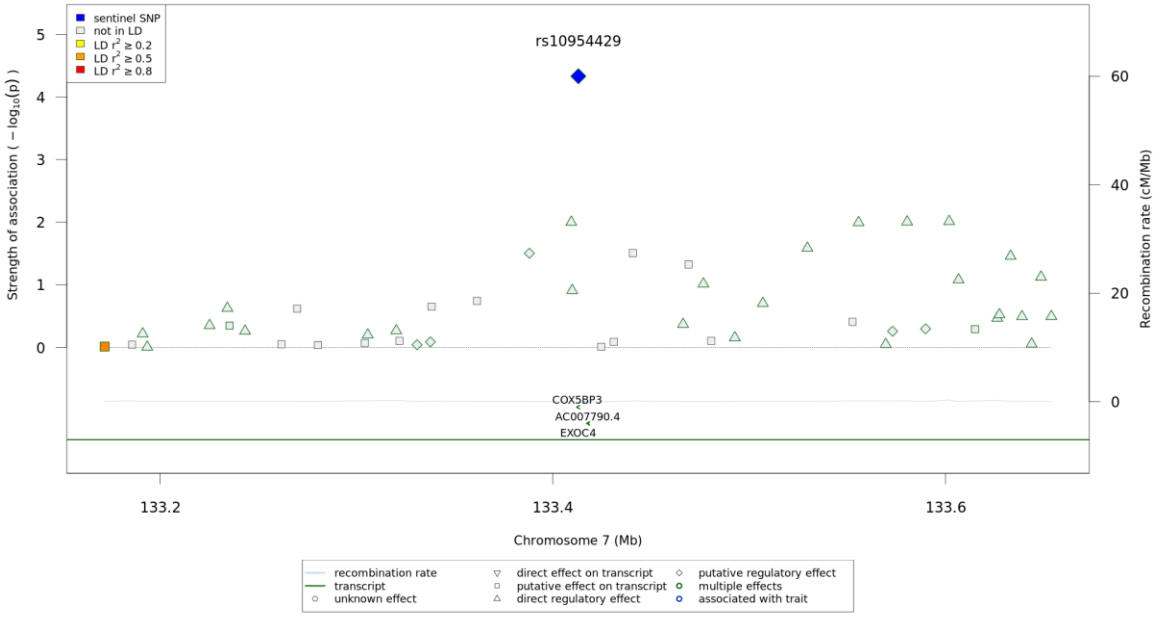


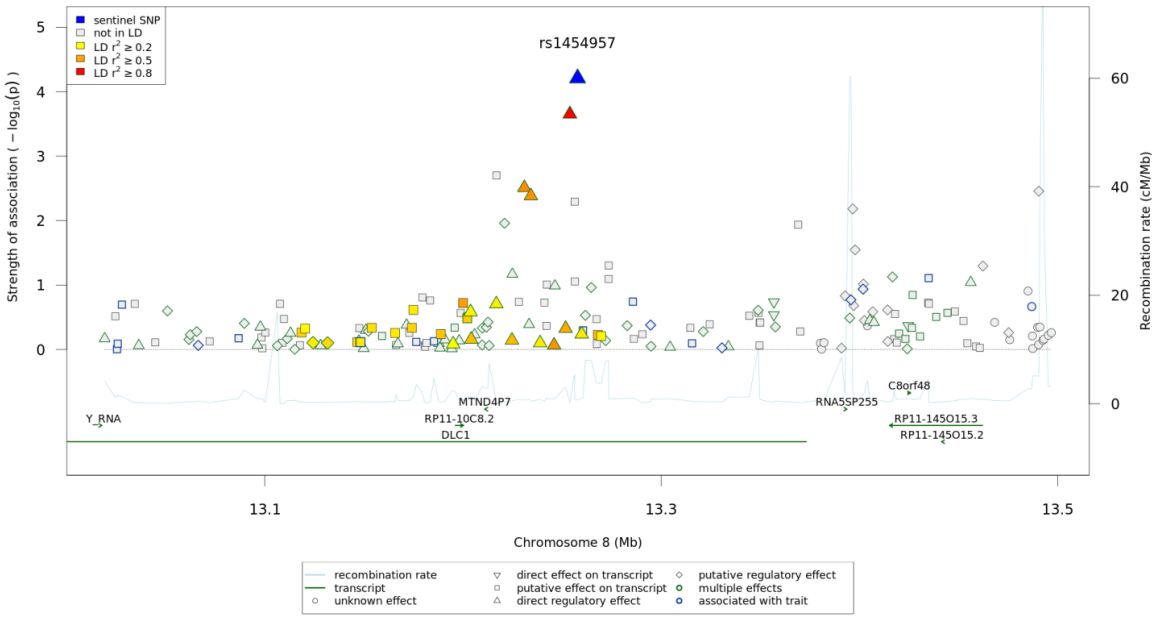
Figure 10: Manhattan plot (−log$_{10}$ [P-value] genome-wide association plot) of the Grady Trauma Project population (replication dataset) using Toxoplasmosis sero-intensity as the continuous outcome variable. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis.



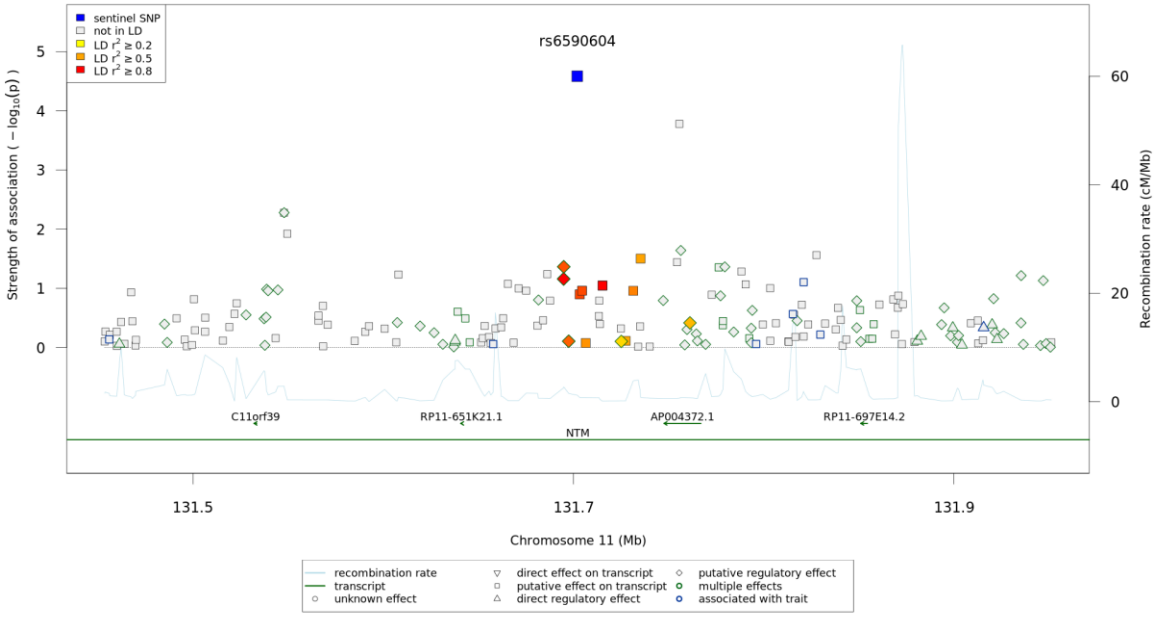
(a) Regional association plot of SNP rs12474831 that is found in the coding region of gene *CNTNAP5* on chromosome 2. This variant appears to have a putative effect on the transcript of the gene.



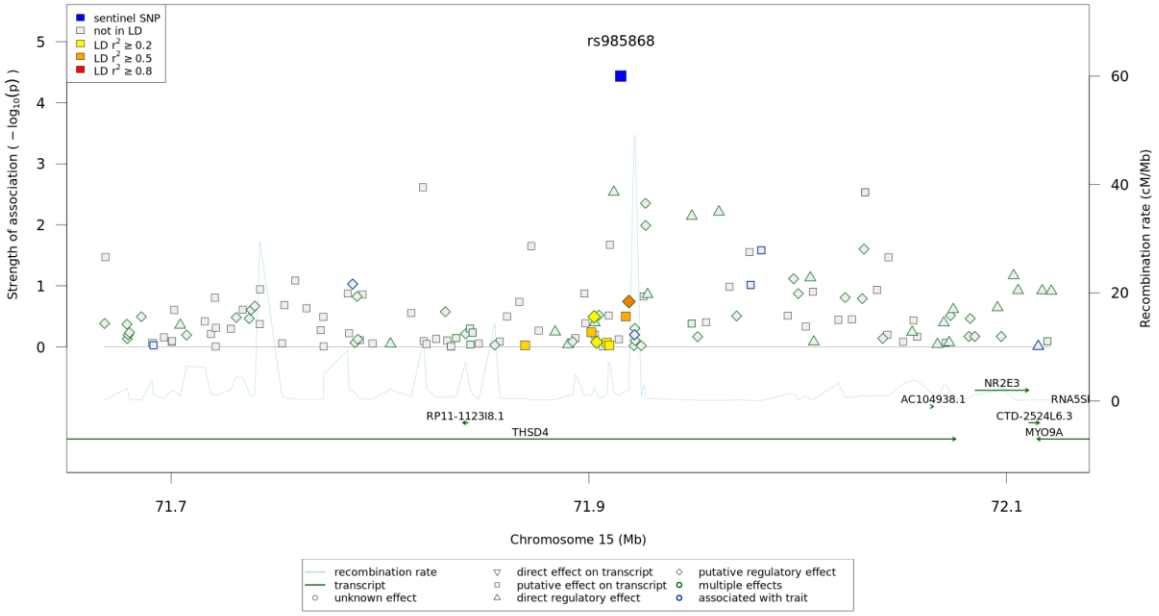
(b) Regional association plot of SNP rs10954429 that is found in the coding region of gene *EXOC4* on chromosome 7. This variant appears to have a putative regulatory effect on the gene.



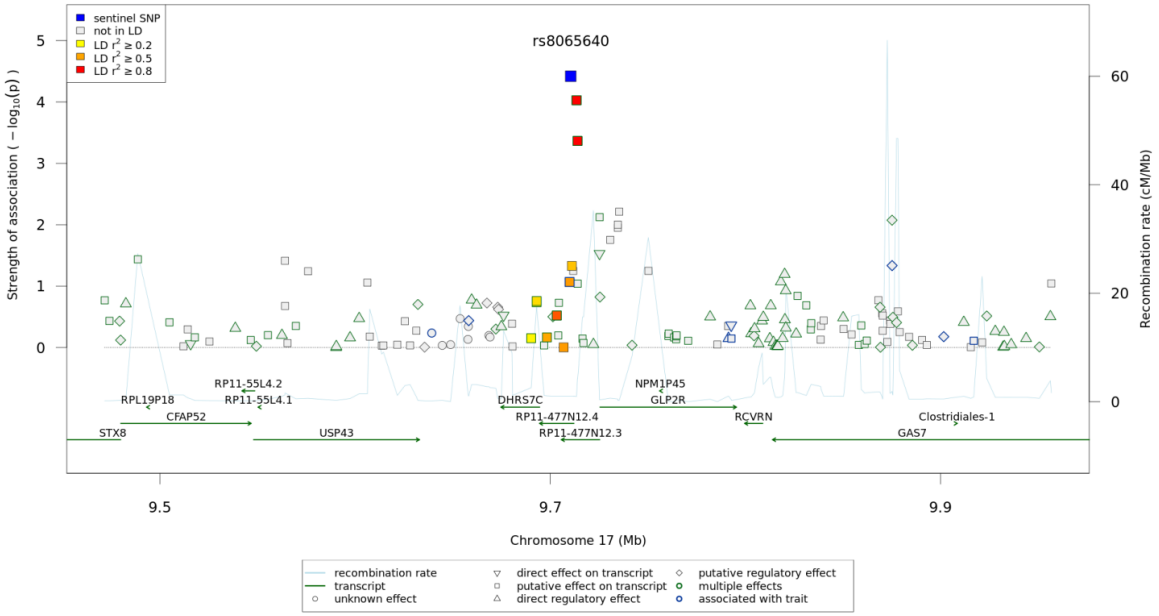
(c) Regional association plot of SNP rs1454957 that is found in the coding region of gene *DLC1* on chromosome 8. This variant appears to have a direct regulatory effect on the gene.



(d) Regional association plot of SNP rs6590604 that is found in the coding region of gene *NTM* on chromosome 11. This variant appears to have a putative effect on transcript of the gene.



(e) Regional association plot of SNP rs985868 that is found in the coding region of gene *THSD4* on chromosome 15. This variant appears to have a putative effect on transcript of the gene.



(f) Regional association plot of SNP rs8065640 that is found in the coding region of gene *GSG1L2* on chromosome 17. This variant appears to have a putative effect on transcript of the gene.

Figure 11 (a-f): Regional association plot around SNPs with the $p<10^{-4}$ for the Grady Trauma Project population when considering sero-intensity as the outcome variable. The left axis indicates the strength of association ($-\log_{10}$ (P)), the right axis indicates the recombination rate (cM/Mb). The significant p-values are shown on the upper part of the plot. The variant symbols represent functional annotations, and SNPs colors show the pair-wise LD correlations to the sentinel variant based on their r^2. The plot also shows regulatory elements. Chromosome numbers can be

found underneath each of the plots. Genes that are shown here are: (a)*CNTNAP5,* (b)*EXOC4,* (c)*DLC1,* (d)*NTM*, (e)*THSD4,* and (f)*GSG1L2*.

Based on these regional plots, many of the candidate genes seem to mostly have a putative effect on the transcripts of genes they are associated with. However, there was a variant that was found to have a direct regulatory effect on the gene, in this case *EXOC4* (Figure 11(b)), a gene the codes a protein that is essential to the exocytic vesicles for specifying docking sites on the plasma membrane (89).

Once we applied the threshold level of $p<0.001$, we were left with 58 SNPs that may be of interest. Of these SNPs, 30 were in close proximity to protein coding genes which include *NTM, CD101, SAMD4A, TRPM3, CASC2, GSG1L2* and *POLD3*. We then proceeded to apply the more liberal threshold value to include any genetic variant that had a $p<0.001$. This yielded 634 SNPs that were of interest. These SNPs were uploaded into Ingenuity® IPA software where coding variants were identified. The complete list can be found in Supplemental Table 8.

Upon completing this analysis, we again decided to stratify the population based on SCZ status to determine which of the genetic variants obtained from the analyses were due to interactions with the enriched SCZ population and which of the genetic variants were attributed to increased risk of infection with *T. gondii* itself.

The population was separated into cases who are individuals diagnosed with SCZ or schizotypal personality disorder and controls who are individuals who have been free of any history of psychiatric hospitalizations, cocaine use, or substance abuse. Once stratified, there were a total of 25 cases and 84 mentally healthy controls ready for analysis.

We proceeded by first carrying out an association analysis using a dichotomous outcome variable, toxoplasmosis sero-conversion status. There were not many SNPs that met our first threshold of $p<10^{-4}$. In fact, among cases there were no variants that met our criteria for inclusion and among controls, there was only one SNP that was under the required p-value. Unfortunately, this SNP did not fall into a coding region on a chromosome. Following this we loosened our threshold level to any SNP with a $p<0.001$. At this level, SNPs obtained from the analysis of the cases still did not yield any variants that met inclusion. Due to the lack of suggestive SNPs, the Manhattan plot for the cases was left out of Figure 12. Instead only the control group is shown, where a total of 137 SNPs were available for further consideration. Once again we read the list of single nucleotide polymorphisms into Ingenuity® IPA software and obtained a list of genes that these variants are in close proximity to (Supplemental Table 9).

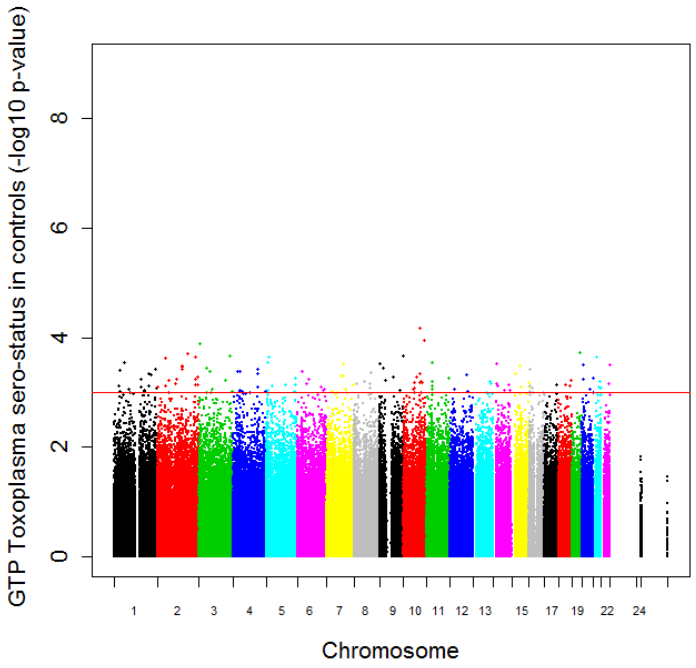


Figure 12: Manhattan plots ($-\log_{10}$ [P-value] genome-wide association plot) of the Grady Trauma Project population (replication dataset) with mentally healthy controls while using the dichotomous *Toxoplasma* sero-conversion status as the outcome variable. Controls consisted of 84 unrelated individuals who were free of any psychiatric hospitalization, cocaine use,

or other substance abuse. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis.

Our next step was to repeat the stratified analysis using a continuous outcome, for this association analysis we used toxoplasmosis sero-intensity as the outcome variable. Like the previous analysis, the population was stratified based on SCZ diagnosis with 25 cases and 84 mentally health controls. Upon completing the analysis, we created a Manhattan plot to visualize regions of interest (Figure 13). Due to the small sample of cases, we chose to omit the Manhattan plot. Based on the plot we can see using a continuous outcome provided more SNPs that could potential be of interest compared to using a dichotomous outcome.

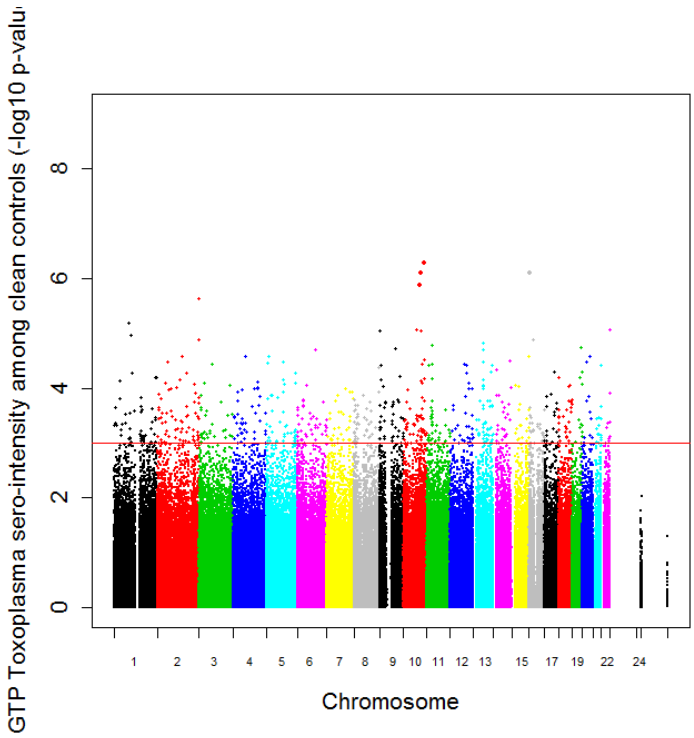


Figure 13: Manhattan plots ($-\log_{10}$ [P-value] genome-wide association plot) of the Grady Trauma Project population (replication dataset) comparing SCZ cases with mentally healthy controls while using the continuous *Toxoplasma* sero-intensity measure as the outcome variable. Controls consisted of 84 unrelated individuals who were free of any psychiatric hospitalization, cocaine use, or other substance abuse. The red line indicates a threshold value of p<0.001, those above the line were included in further analysis.

When applying our threshold level of $p<10^{-4}$ to the results obtained from the association analysis, 85 SNPs in both cases and controls meet the inclusion criteria. Upon closer inspection of these 85 variants among cases, SNPs were found to be related to genes such as *DDX60, CNTNAP2, PTPRT, SHISA6, KDM4B, WDR70,* and *TRPM3*. Some of these genes came up in our previous analyses whereas a new genes appeared such as *DDX60,* which is exhibits antiviral activity through interferon inducible gene expression against hepatitis C virus and vesicular stomatitis virus (90). Another new gene of interest is *CNTNAP2*, that encodes for a protein that functions as an adhesion molecule in the nervous system and has been implicated in multiple neuro-developmental disorders, including Gilles de la Tourette syndrome, schizophrenia, epilepsy, autism, ADHD and mental retardation (91). However, due to the small sample size, the significant SNPs should be considered with extreme caution.

Looking at the SNPs found from the group of only controls, we were able to identify SNPs that coded for gene like *RBFOX1, RBM44, ADGRL2, NCOA3, LRP1B, DCAF17*, and *CPN1*. It is interesting that the gene *RBFOX1* came up in the analysis. This gene encodes for an RNA-binding protein that use involved in regulating alterative splicing. Recent studies have found that this gene is relate to a numerous neuro-developmental disorders including sporadic focal epilepsy and autism spectrum disorder (92).

Once we found this initial set of genes, we reduced our inclusion threshold to any SNP with a $p<0.001$. This meant that among cases, 760 SNPs would be included in the next stage and 709 SNPs from our control group could be analyzed for proximity to genes. We read these variants into Ingenuity® IPA software and obtained a list of genes these SNPs and related to (Supplemental Tables 10 & 11).

Following the last analysis of the GTP population, we went ahead and compiled all the SNPs that fell below p=0.001 and are in genes to see if there were any variants that overlapped between the various association analyses. Table 3 contains genes found in two or more of the associations in our replication dataset. Based on the findings, there appear to be many genes that overlap in two or more of the analyses. It appears that many of the overlapping genes occur between the two analyses that look at the whole population but used different outcome classification such as the genes *AGFG2*, *CDK14*, *ERBB4*, *KIF16B*, *NPR3*, or *SAMD4A*. There were also quite a few genes that occurred in three of the analyses, like *SYT16*, *CDH13*, *PTK2*, or *TENM2*. There were some that appeared in four of the analyses, with genes such as *GALNT11*, *MIR99AHG*, and *PATL2* seen in the whole population analyses and only in the control group. The genes *RYR2* and *SHISA6* were seen in all of the analyses that use classified the outcome as continuous. The most interesting result from this table is that *ASTN2* appeared in all of the analyses.

Table 3: Summary of the overlap between genes among two or more of the genome-wide association analyses carried out on the Grady Trauma Project population among SNPs with p<0.001. Analyses included the whole population, or was stratified by SCZ status into cases (SCZ or schizotypal personality disorder) or controls (no mental issues or substance abuse).

	Dichotomous Toxoplasmosis sero-conversion status			Continuous Toxoplasmosis sero-intensity		
Gene	**Whole pop**	**Cases$^{\Delta}$**	**Controls**	**Whole pop**	**Cases**	**Controls**
A4GALT			x			x
ADARB2	x			x		x
AGFG2	x			x		
AKAP13				x	x	
ALK				x		x
ANKFN1	x				x	
ANO10	x		x	x		
ANTXR2	x			x		
ANXA5				x		x
APBB2					x	x
ARAP1	x				x	
ARHGAP42	x			x	x	
ARHGEF18	x			x		

ARNT			x			x
ASTN2	x		x	x	x	x
C10orf11	x		x	x		x
C8orf37-AS1			x	x		x
CA10	x			x		
CAMTA1				x		x
CCSER1	x			x	x	
CD101	x			x		
CDH13				x	x	x
CDK14	x			x		
CHST9	x			x		
COL24A1			x	x		
COLEC12					x	x
CSMD1				x	x	x
CTIF				x		x
CTNND2			x	x		x
CYBRD1			x			x
DEPDC4	x			x		
DLEC1				x	x	
DNMBP			x			x
DTWD2	x			x		
ELAVL2			x			x
ERBB4	x			x		
EXOC4	x			x		
F2RL1	x			x		
FAM110B			x			x
FGF12			x			x
FGF14			x			x
FHIT	x			x		
FLJ33360	x			x		
FOXP2			x			x
FRMD4A					x	x
FZD10-AS1	x				x	
GALNT11	x		x	x		x
GOT1				x		x
GRM7	x				x	x
GSG1L2	x			x		
HS3ST4	x		x			
INPP4B			x			x
IQSEC1				x		x
KCNH5	x			x		

KCNIP4	x			x		
KIF16B	x			x		
KIRREL3					x	x
KLHL1	x			x		x
LINC00499*	x			x		x
LINC00578*	x			x		
LINC00836*	x			x		
LINC01173*			x			x
LINGO2				x		x
LMO7				x		x
LMX1A				x		x
LOC100128317			x			x
LOC101927598				x		x
LOC101927960	x					x
LOC101929406			x			x
LOC102724418			x			x
LOC105369463	x			x		
LOC105369464	x			x		x
LOC105370003					x	x
LOC105370731			x			x
LOC105371024					x	x
LOC105371070				x		x
LOC105371308				x		x
LOC105371874			x			x
LOC105374008	x			x		
LOC105374524				x		x
LOC105374720			x			x
LOC105374754				x		x
LOC105374919				x	x	
LOC105375855				x		x
LOC105376387					x	x
LOC105376605	x		x			x
LOC105377110	x			x		x
LOC105377161			x			x
LOC105377866	x			x		
LOC339166	x			x		
LOC645177				x	x	
LOC730100				x	x	
LRP1B			x			x
MIR924HG			x			x
MIR99AHG	x		x	x		x

MSI2					x	x
MTHFD1L			x			x
MYH11			x			x
MYO18B					x	x
NCKAP5				x	x	
NCOA3			x			x
NEBL	x			x		
NELL1				x		x
NFIB	x			x		
NKAIN3	x				x	
NMI	x			x	x	
NPR3	x			x		
NTM	x			x	x	
NXN				x	x	
NYAP2	x			x		
OLMALINC	x			x		
PAG1				x		x
PARK2				x	x	x
PATL2	x		x	x		x
PDZRN4				x	x	
PKN2-AS1	x			x		
PLA2G2A	x			x		
POLD3	x			x		
PPARGC1A			x			x
PPEF2				x		x
PRKCB	x			x		
PRKCE				x		x
PRKCH					x	x
PRKG1					x	x
PTER				x		x
PTK2	x			x		x
PTPRD				x	x	
PTPRT				x	x	x
PVT1	x			x	x	
RANBP17	x			x		x
RBFOX1	x		x		x	x
RBMS3				x	x	
RELN				x	x	
RIC1	x		x	x		x
RPA1	x			x		
RPGRIP1	x		x			x

RXRA	x			x		
RYR2	x			x	x	x
SAMD4A	x			x		
SDK1					x	x
SEMA4D	x			x		
SEMA5A			x		x	
SGCZ					x	x
SHISA6	x			x	x	x
SLC39A11					x	x
SMYD2			x			x
SMYD3			x	x		
SORBS2				x	x	
SPRYD4	x			x		
SUPT16H			x			x
SYCP2L	x			x		
SYT16			x		x	x
TBC1D2				x	x	
TBC1D22A			x		x	x
TENM2			x		x	x
TIAM1	x			x		
TRPM3	x			x	x	
TXK			x			x
UBASH3A	x					x
VEZT	x			x		
VPS41				x		x
VPS8	x			x		
XKR6				x		x
XYLT1	x			x		
YIPF7	x					x
ZNF407	x			x		

ᐞ Cases for Dichotomous Toxoplasmosis sero-conversion status was unavailable because there were no SNPs that met the criteria of $p<0.001$.
* Non-coding DNA sequence, long intergenic non-protein coding RNA.

Population comparison

Among the Ashkenazi Jewish population, there was a total of 364,227 SNPs that met the criteria for inclusion into the study. In the Grady Trauma Project population, there was a total of 640,669 SNPs that met the required criteria in order to be included in the analyses.

However, since two different platforms were used to obtain genotype information, Affymetrix for the AJ population and Illumina for the GTP population, this mean that only 89,614 SNPs were present in both of the platforms. At the SNP level, using a threshold of p=0.001 meant that there were no SNPs that were significant in both populations for either sero-positivity or sero-intensity. However, given that we are comparing a relatively homogenous population of Ashkenazi Jews of Central European decent with a predominantly African American population, the ancestral origin is likely relevant to which SNPs associate with the phenotype. Therefore, using the approach of a gene set comparison is more biologically relevant. We compiled a list of genes that were seen in multiple analyses for each of the datasets and compared the population to one another to determine whether there was any overlap in genes tagged by SNPs that were found to be associated with *Toxoplasma gondii* infection (Table 4). They were then input into Ingenuity® IPA software for further causal network analyses.

Table 4: A complied list of all the genes that were tagged with SNPs from two or more of the genome-wide association analyses that were found to be associated with T. gondii infection in the two populations with a threshold value of p<0.001. Genes in bold were overlapping the AJ and GTP populations from either continuous or dichotomous classifications of toxoplasmosis infection.

	Ashkenazi Jewish dataset						Grady Trauma Project					
	Dichotomous Toxoplasmosis sero-conversion status			Continuous Toxoplasmosis sero-intensity			Dichotomous Toxoplasmosis sero-conversion status			Continuous Toxoplasmosis sero-intensity		
Gene	**Whole pop**	**Cases**	**Controls**	**Whole pop**	**Cases**	**Controls**	**Whole pop**	**Cases***	**Controls**	**Whole pop**	**Cases**	**Controls**
AGBL1			x	x	x							x
AKAP13		x								x	x	
AKAP6					x					x		
ALK	x	x		x	x					x		x
ANO10		x					x		x	x		
BACH2				x	x						x	
BBS9				x							x	
C19orf18				x						x		
CAMK1D		x			x		x			x		
CDK14						x	x			x		
CMYA5				x	x					x		
CNTNAP2	x										x	
CRTAC1					x				x			
CSMD1	x			x	x					x	x	x
CTNNA2				x					x	x		x
DAPK1				x							x	
DLGAP1			x								x	
DSCAM			x				x					
ESRRG				x							x	
ETV6			x								x	
FAM107B	x										x	
FAM19A1						x					x	
FARS2	x										x	

FGFR2		x									x	
FHIT	x			x	x		x			x		
FHOD3					x						x	
IQGAP2						x				x		
KAZN					x						x	
KCND2		x									x	
KCNJ6	x	x			x						x	
KCNK9						x				x		
KLF5	x									x		
KLHL1					x		x			x		x
LINGO2	x	x		x	x					x		x
LOC101927630	x			x		x	x				x	
LOC101927967						x					x	
LOC101928394					x						x	
LOC101929563	x			x							x	
LOC105370345		x			x					x		
LOC105370822			x			x				x		
LOC730100		x								x	x	
LRRK1	x									x		
MACROD2				x	x					x		
MEIS2	x									x		
MSI2						x					x	x
MYO16			x							x		
NALCN-AS1	x					x				x		
NEBL		x					x			x		
NKAIN3			x			x	x				x	
NPAS3		x	x			x						x
NPSR1-AS1						x					x	
NRG3		x									x	

NRXN1		x									x	
NRXN3	x										x	
PCDH15		x							x			
PDE4D		x		x								x
PKHD1		x					x					
PLCB4						x					x	
PLXNA4	x						x					
PPARGC1A				x					x			x
PPM1A	x		x								x	
PRKCE					x					x		x
PTPRG					x				x			
PTPRT			x							x	x	x
RANBP17						x	x			x		x
RBFOX1		x	x	x	x		x		x		x	x
RBMS3	x									x	x	
RNF219-AS1					x						x	
RYR2	x						x			x	x	x
SETBP1				x	x		x					
SGCZ						x					x	x
SHANK2					x							x
SHISA6	x		x	x		x	x			x	x	x
SMYD3					x				x	x		
SORBS2		x								x	x	
SOX5			x			x					x	
SPATA13					x						x	
TBC1D22A	x								x		x	x
TENM2						x			x		x	x
TMEM135				x	x					x		
TRAPPC9					x		x					

XYLT1		x			x		x			x		
ZNF236				x								x
ZNF606				x								x

* Cases for Dichotomous Toxoplasmosis sero-conversion status was unavailable because there were no SNPs that met the criteria of $p<0.001$.

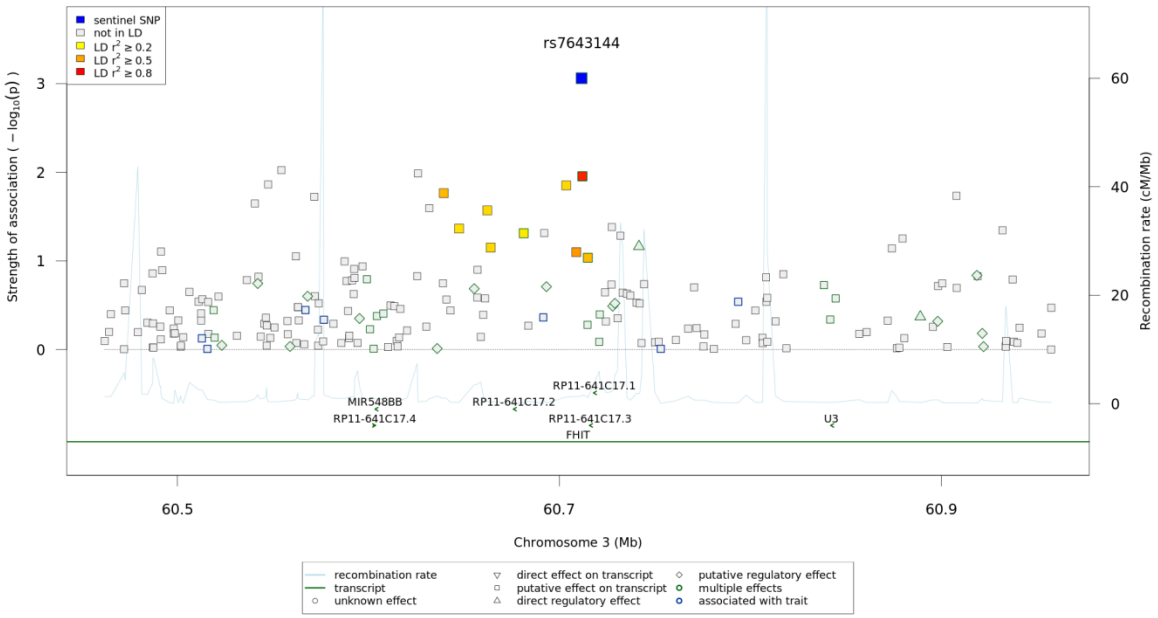
Discussion and Integration of Data

When comparing genes that overlap between the two populations, we must rely on the genome-wide association analyses that were conducted on the entire cohort due to the limited number of cases present in the GTP dataset. In doing so, we hoped to identify *T. gondii* susceptibility genes that can be considered common in the general population, reducing the possibility of an overlap solely due to chance, which may provide insight into a biologically relevant pathway associated with increased susceptibility. When contrasting the results obtained from our dichotomous classification of toxoplasmosis infection in both of the populations only three genes appear in both of the datasets, *FHIT*, *PLXNA4* and *SHISA6*.

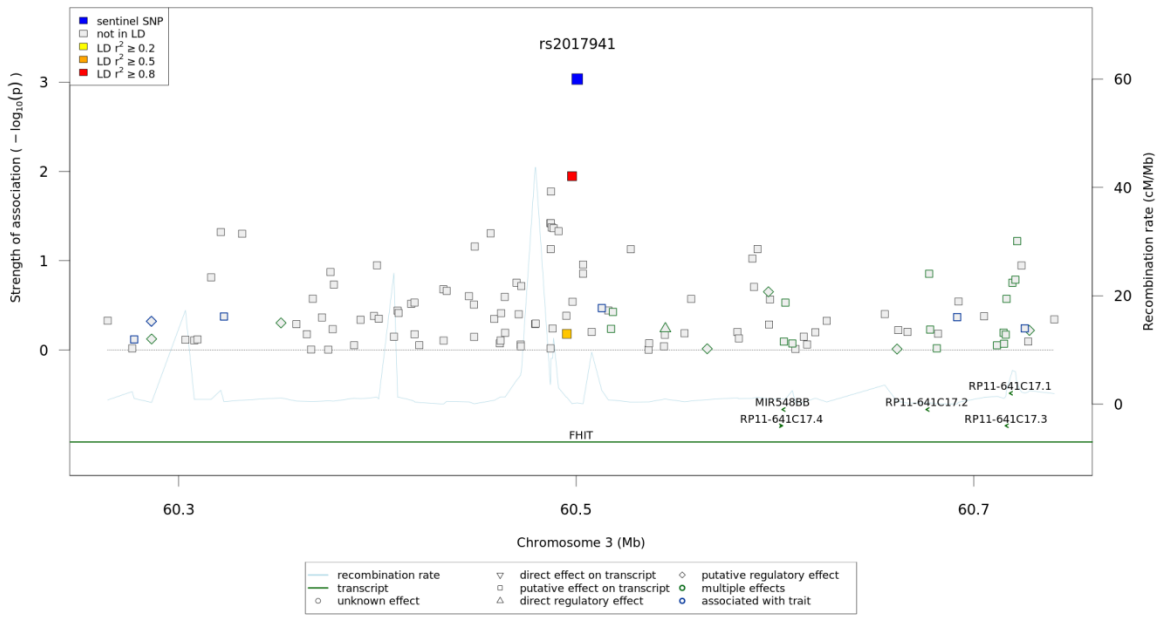
The *PLXNA4* gene encodes for a protein known as plexin-A4 that is believed to transduce signals from semaphorin molecules which then initiate cascades that regulate diverse processes such as axon pruning and repulsion, dendritic attraction and branching, regulation of cell migration, vascular remodeling, and growth cone collapse (93). Through another genome-wide association study by Jun *et al.*, it was found to modulate tau phosphorylation and be significantly associated with Alzheimer's disease (94). One study by Kenny *et al.* looked into the association of *PLXNA4* in an Irish schizophrenic case-control population but failed to find an association (95).

Looking at the results in table 4, we can see that the gene *FHIT* made an appearance in all four of the whole population analyses that used both dichotomous and continuous classifications of the outcome variable. From investigations into genes that appear in our previous regional association plots, we know that the gene *FHIT* encodes for a hydrolase that is involved in the metabolism of purines molecules. Due to the presence of the fragile site FRA3B on chromosome 3, damaged by carcinogens can lead to translocations and

aberrant transcripts. These irregular transcripts have been found in a number of carcinomas but have yet to be linked to mental disorders or infectious diseases. When we plot data from the two populations into regional association plots (Figure 14) we can see that both SNPs appear to be within 200 Kb of one another and both seem to have a putative effect on the transcript of the gene. Though the exact position of the FRA3B instability site is not well known, its believed to span a 500 Kb region along the center of the 1.5 Mb gene for *FHIT* (96). There is a possibility that SNPs obtained from the analyses fall into this common fragile site region.



(a) Regional association plot of SNP rs7643144 that is found in the coding region of gene *FHIT* on chromosome 3 of the AJ population. This variant appears to have a putative effect on transcript of the gene.



(b) Regional association plot of SNP rs2017941 that is found in the coding region of gene *FHIT* on chromosome 3 of the GTP population. This variant appears to have a putative effect on transcript of the gene.

Figure 14 (a-b): Regional association plot around SNPs with the $p<10^{-4}$ for gene *FHIT* when considering sero-conversion status as the outcome variable. The left axis indicates the strength of association ($-\log_{10}$ (P)), the right axis indicates the recombination rate (cM/Mb). The significant p-values are shown on the upper part of the plot. The variant symbols represent functional annotations, and SNPs colors show the pair-wise LD correlations to the sentinel variant based on their r^2. The plot also shows regulatory elements. Chromosome numbers can be found underneath each of the plots.

Inspecting biological network analyses pathways generated through the use of Ingenuity® IPA software we are able to see a computational network of how genes obtained through our analyses are believed to interact with one another (Figure 15). When focusing on the gene *FHIT*, we see that it has a number of connections with genes that were not present in our current analysis but are believed to be in the same pathway. Notice that there is a direct link from the gene to the NFκB complex. The complex has been shown to be integral in controlling transcription of DNA, cell survival and cytokine production (97). Along with this function, it is believed that this complex plays an important role in learning

and memory, with dysfunction of the complex being related to deficits in cognitive function (98). There has been cumulative evidence that has begun to confirm the role of cytokines in conferring susceptibility to schizophrenia and evidence that suggests cytokine regulation also plays an important role in immunopathology of toxoplasmosis (99,100).

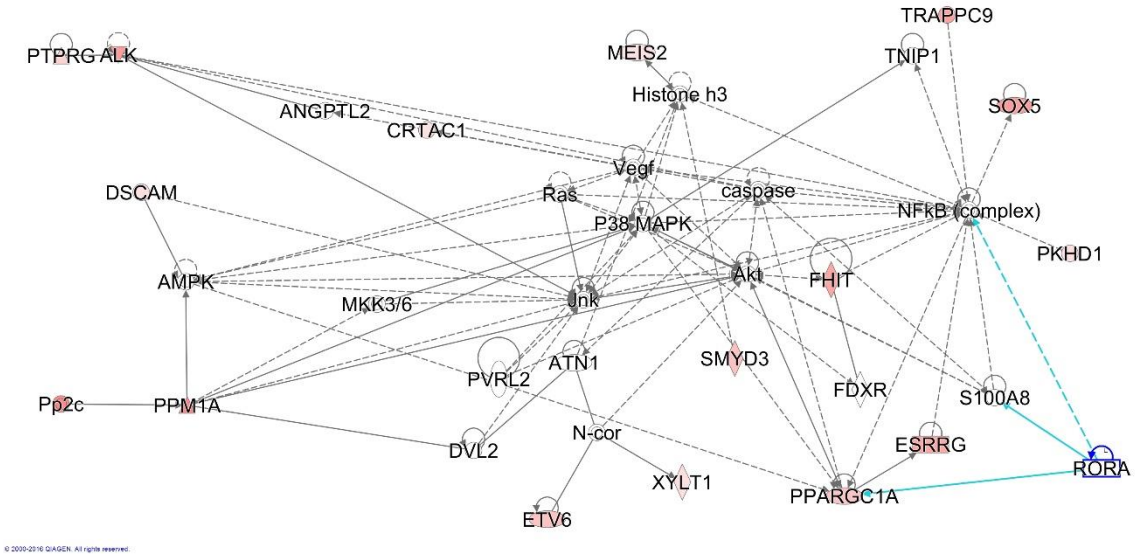


Figure 15: Causal network analyses pathway obtained through Ingenuity® IPA software that shows how genes that are found in both AJ and GTP populations are believed to interact with one another. Genes marked with colored silhouettes are ones obtained through our genome-wide association analyses. Colors intensity correlate with SNPs with lower p-values. Symbols represent classes of protein these genes code for.

Another gene found from our association analyses that appears linked to the NFκB complex is *ALK*. Mentioned previously, the gene encodes for a receptor tyrosine kinase and has been also linked to a number of carcinomas (101). There has been literature that has linked polymorphisms of the gene with increased vulnerability of schizophrenia (102). The protein is expressed preferentially by neurons present in the central and peripheral nervous systems and play important roles in neurodevelopment such as differentiation, proliferation, survival, neurite outgrowth and synaptic formation. Alterations to the

expression of this gene may modify levels of neurotrophic factors in the nervous system such as brain-derived neurotrophic factors and neurotrophin-3, both of which have been implicated in schizophrenia (103).

A gene that was not found through the association analyses but seems to be connected along this computational pathway is the gene that encodes for *RORA*. This gene encodes for a ligand-dependent nuclear receptor that acts as a transcriptional regulator. Only recently has it been identified as a novel candidate gene for neuropsychiatric disorders that include autism spectrum disorder and schizophrenia (104). Although clinically distinct from one another, the two neurological disorders seem to share some common neurological or genetic factors. For example, dendritic spine density have been found in both disorders, with dendritic spine loss observed in schizophrenics but increased spine density was seen among autistic samples (105). The presence of this gene in our network analysis provides a connection to another computational pathway that was observed (Figure 16).

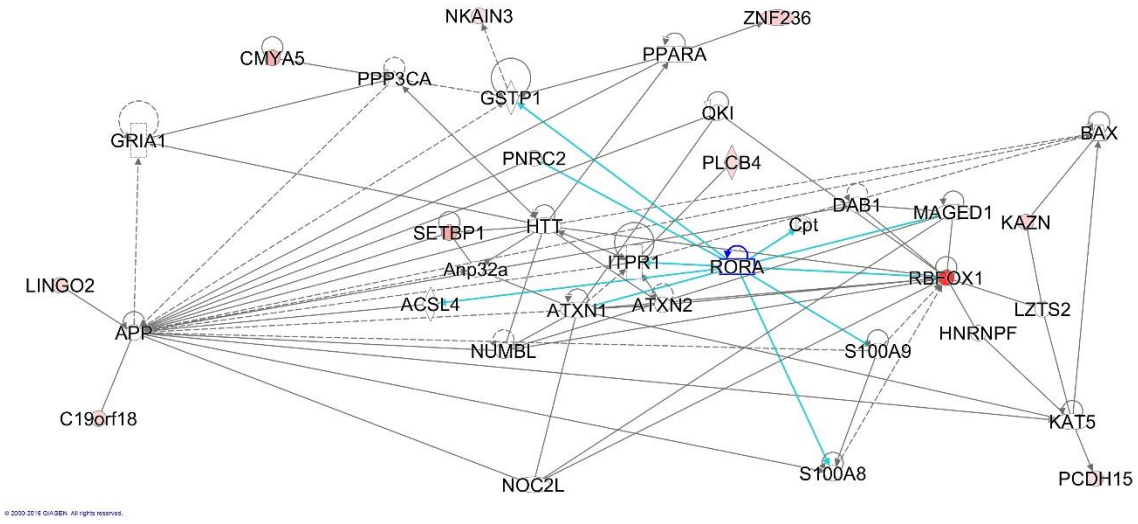


Figure 16: Causal network analyses pathway obtained through Ingenuity® IPA software that shows how genes that are found in both AJ and GTP populations are believed to interact with

one another. Genes marked with colored silhouettes are ones obtained through our genome-wide association analyses. Colors intensity correlate with SNPs with lower p-values. Symbols represent classes of protein these genes code for.

One gene that appears to be significant along this pathway and has a direct connection with the *RORA* gene is the *RBFOX1* gene. From our analyses of regional plots we understand that this gene encodes for an RNA-binding protein that use involved in regulating alterative splicing responsible for widespread effects that both enhance and inhibit the alternative splicing of many cellular pre-mRNAs (106). Deletions and alterations to this gene have been linked to increased risk of neuro-developmental disorders such as autism spectrum disorder, epilepsy (107). The connection between the two genes makes functional sense as both are linked to altering expression or structural aspects related to transcription.

Though quite a few overlapping genes were found between the two distinct populations during the genome-wide association analyses, only a handful were able to be connected together using network modeling. While many of the genes discussed have been related to neuro-developmental disorders and cancer, few have been linked into the field of infectious disease epidemiology and even fewer have been associated with toxoplasmosis infections. A promising avenue of further inquiry would be to assess the role of genes such as *ALK* and *FHIT* on cytokine expression and what specific role these genes play in the regulation of immuno-pathology of toxoplasmosis. Additionally, combining the genetic information obtained through these association analyses with results from a test of cognitive function such as acoustic startle response could highlight genes that may contribute to increased susceptibility to infection which in turn, may increase the risk of developing psychiatric disorders like schizophrenia.

One interesting variant obtained from the association analysis of the AJ dataset when classifying toxoplasmosis status as a dichotomous outcome variable was the top SNP (rs10857870, p= 5.36E-06) that is found in the gene *CHIAP2*. As mentioned in the previous section, this is a pseudo-gene found upstream from *CHIA*, a gene known to encode for the protein chitinase and appears to have a direct regulatory effect on the gene. When the protozoan parasite enters an intermediate host, its life-cycle alternates between a rapidly dividing tachyzoite and the slower dividing bradyzoite, which eventually mature into cysts that remain latent in tissue (108). In humans these cysts are predominantly found in immune protected brain making clearance of the parasite difficult and results in a lifelong infection (109). Chitin is thought to comprise a substantial component of the *T. gondii* cysts wall, and treatment of cysts with chitinase have been shown to cause disruption in the cyst wall. Moreover, alternatively activated macrophages in the brain respond to *T. gondii* infection, and the production of chitinase by these cells is a crucial novel cyst control mechanism in the infected brain (109). Thus, individual genetic variation in chitinase expression may determine susceptibility to *T. gondii*. This insight may pose potential in the future for the development of treatments to combat such infection (110). Importantly, the association of SNPs that coded for this gene was also observed in SCZ subset of the AJ population (p=0.0009353), indicating that the burden of cysts upon the brain may be related to the incidence of the disorder.

Some of the limitations of the current study include having a limited number of cases among the Grady Trauma Project population. This meant we were unable to utilize some of the stratified results obtained from the association analyses. Since this group only consisted of 25 individuals, of which four tested sero-positive for toxoplasmosis infection, the subsample would be minimally informative in a GWAS and therefore was excluded. Another limitation of the study was the classification of individuals who were considered

equivocal after immunoassay measurement. Without a second measurement, these individuals can be subject to misclassification in which case findings of the association analyses may differ.

Additional limitations to the interpretation of the findings in this study is the length of coding sequences. Longer genes means that there is a greater possibility of encountering genetic variants. By approaching our analyses from a gene perspective, the greater number of SNPs in a region, the more emphasis is placed on the gene of interest, which has the possibility of increasing the number of false positives. Approaching the analysis by performing meta-analyses on SNP results may provide results with greater power.

Public Health implications

Parasites pose an important problem all around the world, and disproportionally affects those of lower socio-economic status. By better understanding the genes and pathways that may increase susceptibility to infection with these parasites, we can better understand ways of preventing these infections in the future. Infection with *T. gondii* has been shown to cause mental and cognitive problems after prenatal infections, increase risk of mental health disorders and even alter cognitive abilities in adults. Current literature suggests an association between *T. gondii* infection and the increased risk of developing schizophrenia despite a limited knowledge of the pathways involved. Through a better understanding of the genetics, we hope to gain insight into this connection and one day be able to introduce therapeutic agents that have the ability to reduce the burden of disease globally.

Bibliography

1. Saha S, Chant D, Welham J, et al. A systematic review of the prevalence of schizophrenia. *PLoS Med.* 2005;2(5):0413–0433.
2. Picchioni MM, Murray RM. Schizophrenia. *Bmj* [electronic article]. 2007;335(7610):91–95. (http://www.bmj.com/cgi/doi/10.1136/bmj.39227.616447.BE)
3. Grove J, Børglum AD, Pearce BD. GWAS, Cytomegalovirus Infection, and Schizophrenia. *Curr. Behav. Neurosci. Reports* [electronic article]. 2014;1(4):215–223. (http://link.springer.com/10.1007/s40473-014-0022-1)
4. Wennström M, Nielsen HM, Orhan F, et al. Imbalanced Kynurenine Pathway in Schizophrenia. *Int. J. Tryptophan Res.* 2014;7:1–7.
5. Keefe RSE, Goldberg TE, Harvey PD, et al. The Brief Assessment of Cognition in Schizophrenia: Reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophr. Res.* 2004;68(2-3):283–297.
6. Keefe RSE, Lees-Roitman SE, Dupre RL. Performance of patients with schizophrenia on a pen and paper visuospatial working memory task with short delay. *Schizophr. Res.* 1997;26(1):9–14.
7. Luo X, Li M, Huang L, et al. The Interleukin 3 Gene (IL3) Contributes to Human Brain Volume Variation by Regulating Proliferation and Survival of Neural Progenitors. *PLoS One* [electronic article]. 2012;7(11):e50375. (http://dx.plos.org/10.1371/journal.pone.0050375)
8. Norman RMG, Malla AK. Family history of schizophrenia and the relationship of stress to symptoms : preliminary findings. *Aust. N. Z. J. Psychiatry*. 2001;35(i):217–223.
9. Kendler KS. The Roscommon Family Study. *Arch. Gen. Psychiatry* [electronic article].

1993;50(12):952. (http://archpsyc.jamanetwork.com/article.aspx?articleid=496422)

10. Cardno A, Marshall EJ, Coid B, et al. Heritability Estimates for Psychotic Disorders. *Arch. Gen. Psychiatry*. 1999;56:162–168.
11. Yolken RH, Dickerson FB, Fuller Torrey E. Toxoplasma and schizophrenia. *Parasite Immunol.* [electronic article]. 2009;31(11):706–15. (http://www.ncbi.nlm.nih.gov/pubmed/19825110)
12. Harrison PJ, Law AJ. Neuregulin 1 and Schizophrenia: Genetics, Gene Expression, and Neurobiology. *Biol. Psychiatry*. 2006;60(2):132–140.
13. Morar B, Schwab SG, Albus M, et al. Evaluation of association of SNPs in the TNF alpha gene region with schizophrenia. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* 2007;144(3):318–324.
14. Sarnat JA, Golan R, Greenwald R, et al. Exposure to traffic pollution, acute inflammation and autonomic response in a panel of car commuters. *Environ. Res.* 2014;133:66–76.
15. Singh T, Kurki MI, Curtis D, et al. Rare loss-of-function variants in KMT2F are associated with schizophrenia and developmental disorders. *bioRxiv* [electronic article]. 2016;(http://biorxiv.org/content/early/2016/01/12/036384.abstract)
16. Braff DL, Light GA. The use of neurophysiological endophenotypes to understand the genetic basis of schizophrenia. *Dialogues Clin. Neurosci.* 2005;7(2):125–135.
17. Braff DL, Freedman R. Endophenotypes in Studies of the Genetics of Schizophrenia. *Neuropsychopharmacol. Fifth Gener. Progress.* 2002;703–716.
18. Walters JTR, Owen MJ. Endophenotypes in psychiatric genetics. *Mol. Psychiatry*. 2007;12:886–890.
19. Gottesman II, Gould TD. The Endophenotype Concept in Psychiatry: Etymology and

Strategic Intentions. *Amercian J. Psychiatry*. 2003;160(April):636–645.

20. Braff DL, Grillon C, Geyer MA. Gating and Habituation of the Startle Reflex in Schizophrenic Patients. *Arch. Gen. Psychiatry* [electronic article]. 1992;49(3):206. (http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archpsyc.1992.01820030038005\nhttp://www.scopus.com/inward/record.url?eid=2-s2.0-0026581360&partnerID=tZOtx3y1)
21. Hasenkamp W, Epstein MP, Green A, et al. Heritability of acoustic startle magnitude, prepulse inhibition, and startle latency in schizophrenia and control families. *Psychiatry Res.* 2010;178(2):236–243.
22. Ludewig K, Geyer MA, Etzensberger M, et al. Stability of the acoustic startle reflex, prepulse inhibition, and habituation in schizophrenia. *Schizophr. Res.* 2002;55(1-2):129–137.
23. Cadenhead KS, Swerdlow NR, Shafer KM, et al. Modulation of the startle response and startle laterality in relatives of schizophrenic patients and in subjects with schizotypal personality disorder: Evidence of inhibitory deficits. *Am. J. Psychiatry*. 2000;157(10):1660–1668.
24. Jones JL, Parise ME, Fiore AE. Neglected Parasitic Infections in the United States: Toxoplasmosis. *Am. J. Trop. Med. Hyg.* [electronic article]. 2014;90(5):794–799. (http://www.ajtmh.org/cgi/doi/10.4269/ajtmh.13-0722)
25. Afonso C, Paixão VB, Costa RM. Chronic Toxoplasma infection modifies the structure and the risk of host behavior. *PLoS One* [electronic article]. 2012;7(3):e32489. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3303785&tool=pmcentrez&rendertype=abstract)
26. Sutterland AL, Fond G, Kuin A, et al. Beyond the association. Toxoplasma gondii in schizophrenia, bipolar disorder, and addiction: systematic review and meta-analysis.

Acta Psychiatr. Scand. [electronic article]. 2015;n/a–n/a. (http://doi.wiley.com/10.1111/acps.12423)

27. Dubey JP. Advances in the life cycle of Toxoplasma gondii. *Int. J. Parasitol.* 1998;28(7):1019–1024.
28. Piekarski G. Behavioral alterations caused by parasitic infection in case of latent toxoplasma infection. *Zentralbl Bakteriol Mikrobiol Hyg*. 1981;250:403–406.
29. Witting PA. Learning capacity and memory of normal and Toxoplasma infected laboratory rats and mice. *Z Parasitenkd*. 1979;(61):29–51.
30. Vyas A, Kim S-K, Giacomini N, et al. Behavioral changes induced by Toxoplasma infection of rodents are highly specific to aversion of cat odors. *Proc. Natl. Acad. Sci. U. S. A.* [electronic article]. 2007;104(15):6442–7. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1851063&tool=pmcentrez&rendertype=abstract)
31. Berenreiterová M, Flegr J, Kuběna AA, et al. The distribution of Toxoplasma gondii cysts in the brain of a mouse with latent toxoplasmosis: Implications for the behavioral manipulation hypothesis. *PLoS One*. 2011;6(12).
32. Poirotte C, Kappeler PM, Ngoubangoye B, et al. Morbid attraction to leopard urine in Toxoplasma-infected chimpanzees. *Curr. Biol.* [electronic article]. 2016;26(3):R98–R99. (http://linkinghub.elsevier.com/retrieve/pii/S0960982215015171)
33. Flegr J. Effects of Toxoplasma on human behavior. *Schizophr. Bull.* 2007;33(3):757–760.
34. Stern H, Boothroyd JC, Elek SD, et al. Microbial Causes of Mental Retardation: The Role of Prenatal Infections with Cytomegalovirus, Rubella virus and Toxoplasma. *Lancet*. 1969;(August).
35. Alford Jr CA, Stagno S, Reynolds DW. Congenital toxoplasmosis: clinical, laboratory,

and therapeutic considerations, with special reference to subclinical disease. *Bull. N. Y. Acad. Med.* 1974;50(2):160–181.

36. Wilson CB, Remington JS, Stagno S, et al. Development of Adverse Sequelae in Children Born with Subclinical Congenital Toxoplasma infection. *Pedicatrics*. 1980;66(767):436–438.
37. Caiaffa WT, Chiari CA, Figueiredo ARP, et al. Toxoplasmosis and mental retardation: report of a case-control study. *Mem. Inst. Oswaldo Cruz*. 1993;88(2):253–261.
38. Koppe JG, Loewer-Sieger DH, DE ROEVER-BONNET H. Results of 20-Year Follow-Up of Congenital Toxoplasmosis. *Lancet*. 1986;327(8475):254–256.
39. Koppe JG, Kloosterman GJ. Congenital toxoplasmosis: long-term follow-up. *Padiatr. Padol.* 1982;17(2):171–179.
40. Berrébi A, Assouline C, Bessières M-H, et al. Long-term outcome of children with congenital toxoplasmosis. *Am. J. Obstet. Gynecol.* [electronic article]. 2010;203(6):552.e1–e6. (http://dx.doi.org/10.1016/j.ajog.2010.06.002)
41. Arias I, Sorlozano A, Villegas E, et al. Infectious agents associated with schizophrenia: A meta-analysis. *Schizophr. Res.* [electronic article]. 2012;136(1-3):128–136. (http://dx.doi.org/10.1016/j.schres.2011.10.026)
42. Torrey EF, Bartko JJ, Lun ZR, et al. Antibodies to Toxoplasma gondii in patients with schizophrenia: A meta-analysis. *Schizophr. Bull.* 2007;33(3):729–736.
43. Amminger GP, McGorry PD, Berger GE, et al. Antibodies to Infectious Agents in Individuals at Ultra-High Risk for Psychosis. *Biol. Psychiatry*. 2007;61(10):1215–1217.
44. Niebuhr DW, Millikan AM, Cowan DN, et al. Selected infectious agents and risk of schizophrenia among U.S. military personnel. *Am. J. Psychiatry*. 2008;165(1):99–106.
45. Groër MW, Yolken RH, Xiao J, et al. Prenatal depression and anxiety in Toxoplasma

gondii – positive women. *Am. J. Obstet. Gynecol.* [electronic article]. 2011;204(5):433.e1–433.e7. (http://dx.doi.org/10.1016/j.ajog.2011.01.004)

46. Hamdani N, Daban-Huard C, Lajnef M, et al. Relationship between Toxoplasma gondii infection and bipolar disorder in a French sample. *J. Affect. Disord.* 2013;148(2-3):444–448.

47. Pearce BD, Kruszon-Moran D, Jones JL. The relationship between Toxoplasma gondii infection and mood disorders in the NHANES III. 2012;72(4):290–295.

48. Dickerson F, Boronow J, Stallings C, et al. Toxoplasma gondii in individuals with schizophrenia: Association with clinical and demographic factors and with mortality. *Schizophr. Bull.* 2007;33(3):737–740.

49. Arling T a, Yolken RH, Lapidus M, et al. Toxoplasma gondii antibody titers and history of suicide attempts in patients with recurrent mood disorders. *J. Nerv. Ment. Dis.* 2009;197(12):905–908.

50. Yagmur F, Yazar S, Temel HO, et al. May Toxoplasma gondii increase suicide attempt-preliminary results in Turkish subjects? *Forensic Sci. Int.* [electronic article]. 2010;199(1-3):15–17. (http://dx.doi.org/10.1016/j.forsciint.2010.02.020)

51. Sublette ME, Galfalvy HC, Fuchs D, et al. Plasma kynurenine levels are elevated in suicide attempters with major depressive disorder. *Brain. Behav. Immun.* [electronic article]. 2011;25(6):1272–1278. (http://dx.doi.org/10.1016/j.bbi.2011.05.002)

52. Ling VJ, Lester D, Mortensen PB, et al. Toxoplasma gondii seropositivity and suicide rates in women. *J. Nerv. Ment. Dis.* [electronic article]. 2011;199(7):440–4. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3128543&tool=pmcentrez&rendertype=abstract)

53. Webster JP. Rats, cats, people and parasites: The impact of latent toxoplasmosis on behaviour. *Microbes Infect.* 2001;3(12):1037–1045.

54. Flegr J, Havlícek J, Kodym P, et al. Increased risk of traffic accidents in subjects with latent toxoplasmosis: a retrospective case-control study. *BMC Infect. Dis.* [electronic article]. 2002;2:11. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=117239&tool=pmcentrez&rendertype=abstract)

55. Havlícek J, Gasová ZG, Smith a P, et al. Decrease of psychomotor performance in subjects with latent "asymptomatic" toxoplasmosis. *Parasitology*. 2001;122(Pt 5):515–520.

56. Pearce BD, Kruszon-Moran D, Jones JL. The association of Toxoplasma gondii infection with neurocognitive deficits in a population based analysis. *Soc. Psychiatry Psychiatr. Epidemiol.* 2014;49(6):1001–1010.

57. Pearce BD, Hubbard S, Rivera HN, et al. Toxoplasma gondii exposure affects neural processing speed as measured by acoustic startle latency in schizophrenia and controls. *Schizophr. Res.* [electronic article]. 2013;150(1):258–261. (http://dx.doi.org/10.1016/j.schres.2013.07.028)

58. Příplatová L, Šebánková B, Flegr J. Contrasting Effect of Prepulse Signal on Performance of Toxoplasma-infected and Toxoplasma-free Subjects in and Acoustic Reaction Time Test. *PLoS One*. 2014;9(11).

59. Peixoto-Rangel AL, Miller EN, Castellucci L, et al. Candidate gene analysis of ocular toxoplasmosis in Brazil: evidence for a role for toll-like receptor 9 (TLR9). *Mem. Inst. Oswaldo Cruz*. 2009;104(8):1187–1190.

60. Jamieson S, Peixoto-Rangel A, Hargrave AC, et al. Evidence for associations between the purinergic receptor P2X7 (P2RX7) and toxoplasmosis. *Genes Immun.* 2010;11(5):374–383.

61. Jamieson SE, de Roubaix LA, Cortina-Borja M, et al. Genetic and epigenetic factors at

COL2A1 and ABCA4 influence clinical outcome in congenital toxoplasmosis. *PLoS One*. 2008;3(6).

62. Avramopoulos D, Pearce BD, McGrath J, et al. Infection and Inflammation in Schizophrenia and Bipolar Disorder: A Genome Wide Study for Interactions with Genetic Variation. *PLoS One* [electronic article]. 2015;10(3):e0116696. (http://dx.plos.org/10.1371/journal.pone.0116696)

63. Børglum AD, Demontis D, Grove J, et al. Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Mol. Psychiatry* [electronic article]. 2013;19(3):325–333. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3932405&tool=pmcentrez&rendertype=abstract)

64. Nylocks KM, Michopoulos V, Rothbaum a. O, et al. An angiotensin-converting enzyme (ACE) polymorphism may mitigate the effects of angiotensin-pathway medications on posttraumatic stress symptoms. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* [electronic article]. 2015;168(4):307–315. (http://doi.wiley.com/10.1002/ajmg.b.32313)

65. Wingo AP, Almli LM, Stevens JJ, et al. DICER1 and microRNA regulation in Post-Traumatic Stress Disorder With Comorbid Depression. *Nat. Commun.* [electronic article]. 2015;6:1–12. (http://dx.doi.org/10.1038/ncomms10106)

66. Ripke S, Dushlaine CO, Chambert K, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. …* [electronic article]. 2013;45(10):1–26. (http://www.nature.com/ng/journal/v45/n10/abs/ng.2742.html)

67. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* [electronic article]. 2015;4(1):7. (http://arxiv.org/abs/1410.4803)

68. Team RC. R: A language and environment for statistical computing. 2012;5.

69. Riva A, Kohane IS. SNPper: Retrieval and analysis of human SNPs. *Bioinformatics*. 2002;18(12):1681–1685.

70. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* [electronic article]. 2001;29(1):308–11. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29783&tool=pmcentrez&rendertype=abstract)

71. Arnold M, Raffler J, Pfeufer A, et al. SNiPA: An interactive, genetic variant-centered annotation browser. *Bioinformatics*. 2015;31(8):1334–1336.

72. Li MD, Mangold JE, Seneviratne C, et al. Association and interaction analyses of GABBR1 and GABBR2 with nicotine dependence in European- and African-American populations. *PLoS One*. 2009;4(9).

73. Chatterjee R, Batra J, Das S, et al. Genetic association of acidic mammalian chitinase with atopic asthma and serum total IgE levels. *J. Allergy Clin. Immunol.* 2008;122(1).

74. Pickard BS, Pieper A a, Porteous DJ, et al. The NPAS3 gene--emerging evidence for a role in psychiatric illness. *Ann. Med.* 2006;38(6):439–448.

75. Kwasnicka-Crawford DA, Carson AR, Roberts W, et al. Characterization of a novel cation transporter ATPase gene (ATP13A4) interrupted by 3q25-q29 inversion in an individual with language delay. *Genomics*. 2005;86(2):182–194.

76. Zhu Q, Shao XM, Kao L, et al. Missense mutation T485S alters NBCe1-A electrogenicity causing proximal renal tubular acidosis. *Am. J. Physiol. Cell Physiol.* [electronic article]. 2013;305(4):C392–405. (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3891219&tool=pmcentrez&rendertype=abstract)

77. Waters CE, Saldivar JC, Hosseini SA, et al. The FHIT gene product: tumor suppressor

and genome "caretaker." *Cell. Mol. Life Sci.* 2014;71(23):4577–4587.

78. Zantomio D, Chana G, Laskaris L, et al. Convergent evidence for mGluR5 in synaptic and neuroinflammatory pathways implicated in ASD. *Neurosci. Biobehav. Rev.* [electronic article]. 2015;52:172–177. (http://dx.doi.org/10.1016/j.neubiorev.2015.02.006)

79. Masotti A, Uva P, Davis-Keppen L, et al. Keppen-lubinsky syndrome is caused by mutations in the inwardly rectifying K+ channel encoded by KCNJ6. *Am. J. Hum. Genet.* [electronic article]. 2015;96(2):295–300. (http://dx.doi.org/10.1016/j.ajhg.2014.12.011)

80. Mullin BH, Prince RL, Mamotte C, et al. Further genetic evidence suggesting a role for the RhoGTPase-RhoGEF pathway in osteoporosis. *Bone* [electronic article]. 2009;45(2):387–391. (http://dx.doi.org/10.1016/j.bone.2009.04.254)

81. Vanita V, Guo G, Singh D, et al. Differential effect of cataract-associated mutations in MAF on transactivation of MAF target genes. *Mol. Cell. Biochem.* [electronic article]. 2014;396(1-2):137–145. (http://link.springer.com/10.1007/s11010-014-2150-z)

82. Saykin AJ, Shen L, Foroud TM, et al. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer's Dement.* [electronic article]. 2010;6(3):265–273. (http://dx.doi.org/10.1016/j.jalz.2010.03.013)

83. Liao C, Fu F, Li R, et al. Loss-of-function variation in the DPP6 gene is associated with autosomal dominant microcephaly and mental retardation. *Eur. J. Med. Genet.* 2013;56(9):484–489.

84. Ross AP, Mansilla MA, Choe Y, et al. A Mutation in Mouse Pak1ip1 Causes Orofacial Clefting while Human PAK1IP1 Maps to 6p24 Translocation Breaking Points Associated with Orofacial Clefting. *PLoS One*. 2013;8(7).

85. Zhao J, Zhou J, Chen Z, et al. Different histopathology but the same clonality: ALK rearrangement in a patient with metastatic non-small-cell lung cancer. *Int. J. Clin. Exp. Pathol.* 2015;8(3):3344–3348.

86. Wu YW, Prakash KM, Rong TY, et al. Lingo2 variants associated with essential tremor and Parkinson's disease. *Hum. Genet.* 2011;129(6):611–615.

87. Hoepfner S, Severin F, Cabezas A, et al. Modulation of receptor recycling and degradation by the endosomal kinesin KIF16B. *Cell*. 2005;121(3):437–450.

88. Bennett TM, Mackay DS, Siegfried CJ, et al. Mutation of the melastatin-related cation channel, TRPM3, underlies inherited cataract and glaucoma. *PLoS One*. 2014;9(8).

89. Grindstaff KK, Yeaman C, Anandasabapathy N, et al. Sec6/8 complex is recruited to cell-cell contacts and specifies transport vesicle delivery to the basal-lateral membrane in epithelial cells. *Cell*. 1998;93(5):731–740.

90. Oshiumi H, Miyashita M, Okamoto M, et al. DDX60 Is Involved in RIG-I-Dependent and Independent Antiviral Responses, and Its Function Is Attenuated by Virus-Induced EGFR Activation. *Cell Rep.* [electronic article]. 2015;11(8):1193–1207. (http://dx.doi.org/10.1016/j.celrep.2015.04.047)

91. Chiocchetti a G, Kopp M, Waltes R, et al. Variants of the CNTNAP2 5′ promoter as risk factors for autism spectrum disorders: a genetic and functional approach. *Mol. Psychiatry* [electronic article]. 2014;20(7):839–849. (http://www.nature.com/doifinder/10.1038/mp.2014.103)

92. Lal D, Pernhorst K, Klein KM, et al. Extending the phenotypic spectrum of RBFOX1 deletions: Sporadic focal epilepsy. *Epilepsia*. 2015;56(9):e129–e133.

93. Suto F, Ito K, Uemura M, et al. Plexin-A4 Mediates Axon-Repulsive Activities of Both Secreted and Transmembrane Semaphorins and Plays Roles in Nerve Fiber Guidance. *J. Neurosci.* [electronic article]. 2005;25(14):3628–3637.

(http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4480-04.2005)

94. Jun G, Asai H, Zeldich E, et al. PLXNA4 is Associated with Alzheimer Disease and Modulates Tau Phosphorylation. *Ann. Neurol.* 2014;76(3):379–392.

95. Kenny E, Morris D, Gilks W. Association studies of SEMA6A, SEMA6B, PLXNA2 and PLXNA4 genes in an Irish schizophrenia case-control sample. *Ulster Med. J.* 2008;77(1).

96. Becker N a, Thorland EC, Denison SR, et al. Evidence that instability within the FRA3B region extends four megabases. *Oncogene*. 2002;21(57):8713–8722.

97. Gilmore TD. The Rel/NF-kappaB signal transduction pathway: introduction. *Oncogene*. 1999;18(49):6842–6844.

98. Kaltschmidt B, Ndiaye D, Korte M, et al. NF- κ B Regulates Spatial Memory Formation and Synaptic Plasticity through Protein NF-kB Regulates Spatial Memory Formation and Synaptic Plasticity through Protein Kinase A / CREB Signaling. *Mol. Cell. Biol.* 2006;26(8):2936–2946.

99. Watanabe Y, Someya T, Nawa H. Cytokine hypothesis of schizophrenia pathogenesis: Evidence from human studies and animal models. *Psychiatry Clin. Neurosci.* 2010;85(3):1–10.

100. Gaddi PJ, Yap GS. Cytokine regulation of immunopathology in toxoplasmosis. *Immunol. Cell Biol.* 2007;85(2):155–159.

101. Doyle LA, Mariño-Enriquez A, Fletcher CD, et al. ALK rearrangement and overexpression in epithelioid fibrous histiocytoma. *Mod. Pathol.* 2015;28(7):904–912.

102. Kunugi H, Hashimoto R, Okada T, et al. Possible association between nonsynonymous polymorphisms of the anaplastic lymphoma kinase (ALK) gene and schizophrenia in a Japanese population. *J. Neural Transm.* 2006;113(10):1569–1573.

103. Green MJ, Matheson SL, Shepherd a, et al. Brain-derived neurotrophic factor levels in schizophrenia: a systematic review with meta-analysis. *Mol. Psychiatry*. 2011;16(9):960–972.

104. Devanna P, Vernes SC. A direct molecular link between the autism candidate gene RORa and the schizophrenia candidate MIR137. *Sci. Rep.* 2014;4(5):3994.

105. Glantz LA, Lewis DA. Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. *Arch Gen Psychiatry*. 2000;57(1):65–73.

106. Bill BR, Lowe JK, DyBuncio CT, et al. Orchestration of neurodevelopmental programs by RBFOX1: Implications for autism spectrum disorder. 1st ed. Elsevier Inc.; 2013 251-267 p.

107. Lal D, Trucks H, M??ller RS, et al. Rare exonic deletions of the RBFOX1 gene increase risk of idiopathic generalized epilepsy. *Epilepsia*. 2013;54(2):265–271.

108. Nefoy SBON, Nger IANDMA. Genetic and biochemical analysis of development in T oxoplasma gondii. *Microbiol. Immunol.* 1997;1347–1354.

109. Nance JP, Vannella KM, Worth D, et al. Chitinase Dependent Control of Protozoan Cyst Burden in the Brain. *PLoS Pathog.* 2012;8(11).

110. Sutherland TE, Maizels RM, Allen JE. Chitinases and chitinase-like proteins: Potential therapeutic targets for the treatment of T-helper type 2 allergies. *Clin. Exp. Allergy*. 2009;39(7):943–955.

Appendix

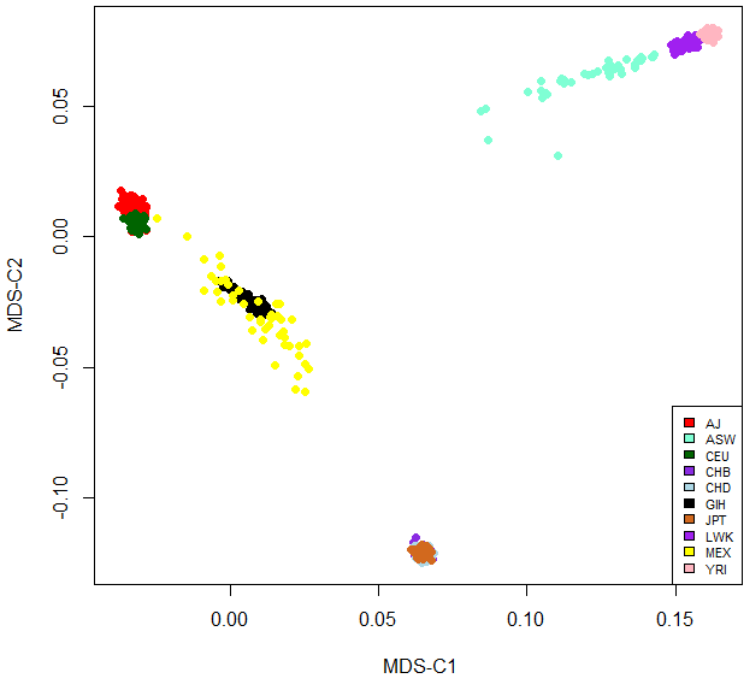


Figure 1: HapMap for Ashkenazi Jewish discovery population (in red) showing the clustering of the subjects. The tight clustering around the CEU (Northern and Western European ancestry) population indicates that no subjects needed to be removed for principal component calculation.

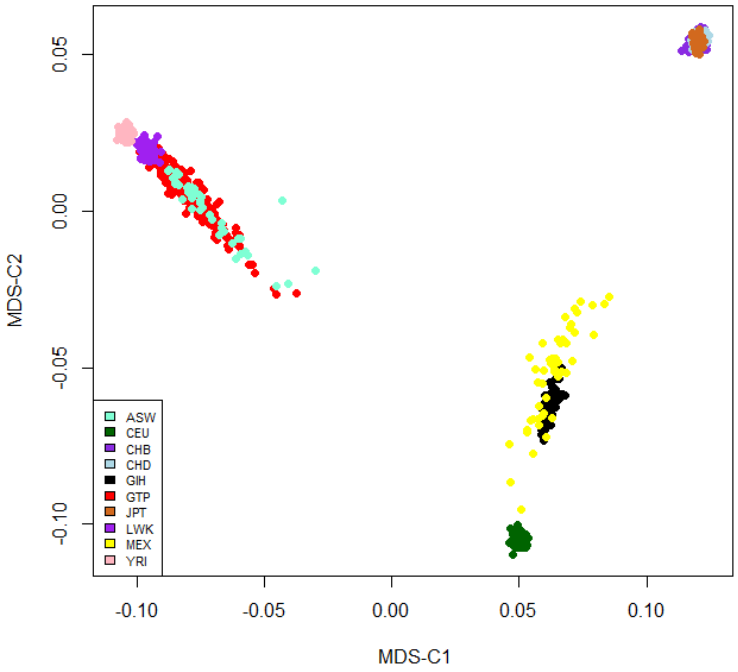


Figure 2: HapMap for Grady Trauma Project replication population (in red) showing the clustering of the subjects. The tight clustering around the ASW (African ancestry) and LWK (Kenya) populations indicates that no subjects needed to be removed for principal component calculation.

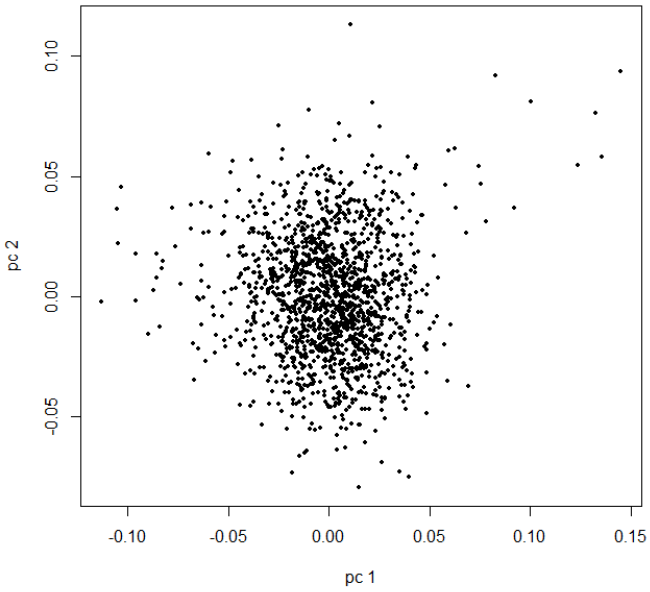


Figure 3: Population structure plot of Ashkenazi Jewish discovery population of PC1 and PC2 indicating the clustering of subjects.

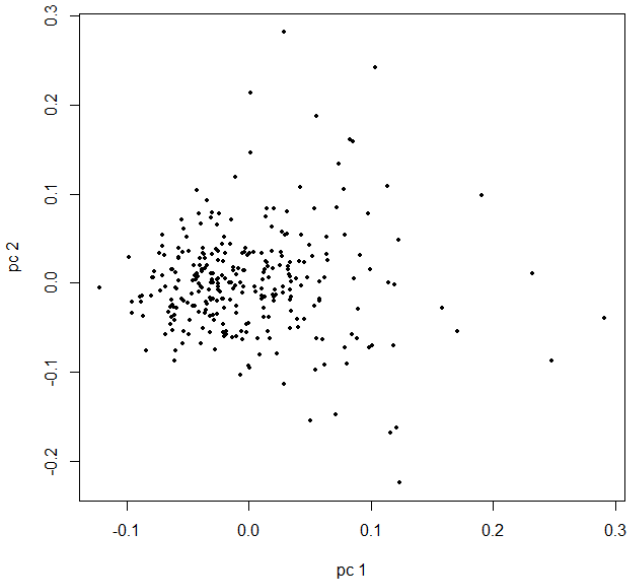


Figure 4: Population structure plot of Grady Trauma Project replication population of PC1 and PC2 indicating the clustering of subjects.

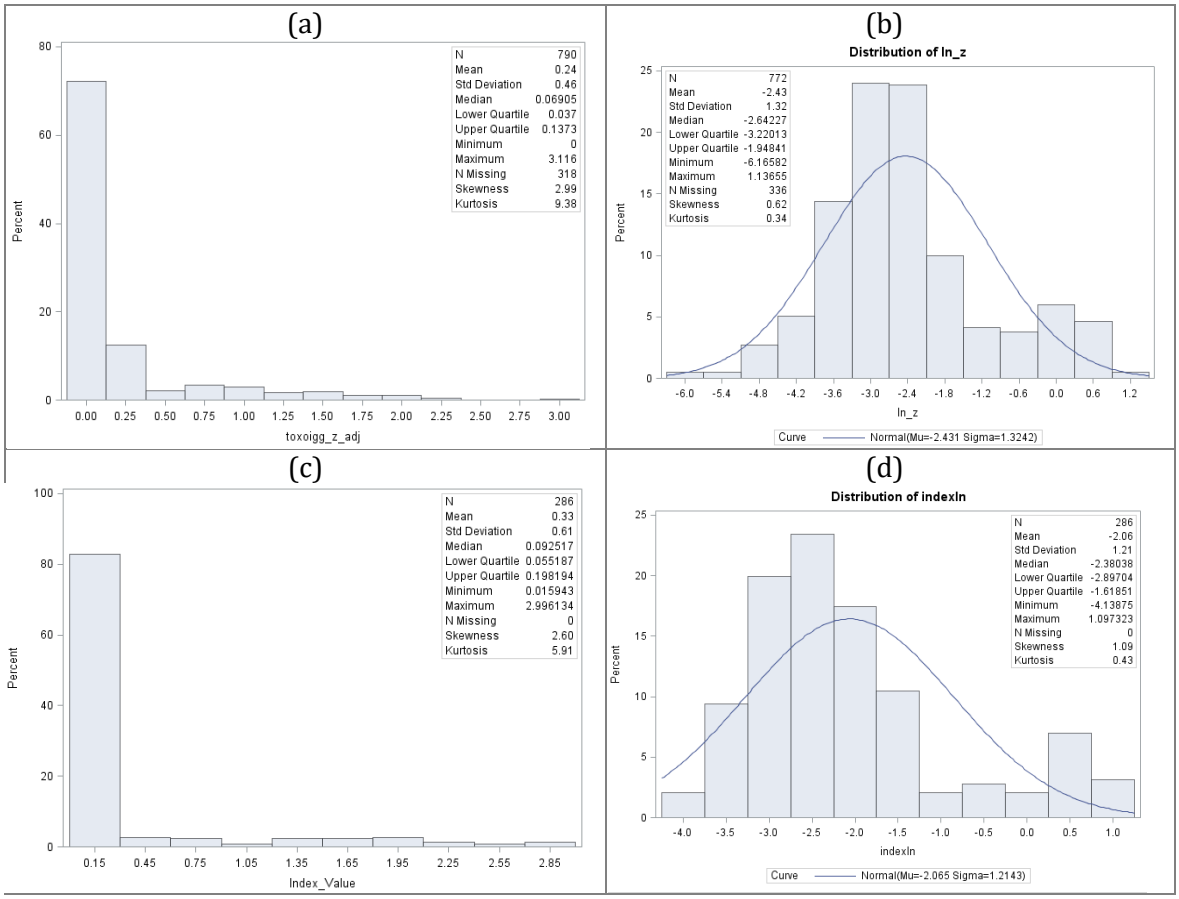


Figure 5: Histograms to show the distribution of *Toxoplasma* sero-conversion scores and natural log transformed toxoplasmosis sero-conversion scores. (a) *Toxoplasma* sero-conversion scores of Ashkenazi Jewish population, showing a clear right skewed distribution. (b) *Toxoplasma* sero-conversion scores of Ashkenazi Jewish population converted using natural logarithm to fit a normal distribution. (c) *Toxoplasma* sero-conversion scores of Grady Trauma Project population, showing a clear right skewed distribution. (d) *Toxoplasma* sero-conversion scores of Grady Trauma Project population converted using natural logarithm to fit a normal distribution.

Annotated code

Data Cleaning

```
libname toxo "H:\My Documents\Thesis Project\SAS program";

proc contents data=toxo.compiled;
run;
proc contents data=toxo.ressler;
run;

proc sort data= toxo.compiled nodupkey;
by SID;
run;
```

```
proc sort data=toxo.compiled nodupkey;
by SID;
run;

data toxo.gtpdata;
merge toxo.ressler toxo.compiled;
by SID;
run;

proc contents data=toxo.gtpdata;
run;

proc freq data=toxo.gtpdata;
tables age/ list;
where age ge 1;
run;

*create a copy of dataset for removal of duplicate and non-GTP subjects;
proc export data=toxo.gtpdata dbms=xlsx            outfile="H:\My Documents\Thesis Project\SAS
program\gtpdata.xlsx" replace;
run;

*remove all subjects missing age variable;
data cleanstep1;
set toxo.gtpdata;
if age=. then delete;
run;
*total of 628 subjects in cleanstep1;

proc freq data=cleanstep1;
tables education*employment*income/list;
where education ge 0;
where employment ge 0;
where income ge 0;
run;

*remove all subjects who are missing demographic data;
data cleanstep2;
set cleanstep1;
if race_ethnic=. then delete;
if education=. then delete;
if income=. then delete;
run;
*total of 610 subjects in cleanstep2;

*remove all sujects who are missing toxoplasma data;
data cleanstep3;
set cleanstep2;
if Tox=" " then delete;
run;
*total of 393 subjects in cleanstep3;

*remove all subjects who are missing startle data;
data cleanstep4;
set cleanstep3;
```

```
if amp1_7=. then delete;
if Lat1_7=. then delete;
run;
*total of 363 subjects in cleanstep4;

*create a copy of cleaned datastep;
proc export data=toxo.gtpdata dbms=xlsx outfile="H:\My Documents\Thesis Project\SAS program\clean_gtp.xlsx" replace;
run;
```

Discovery dataset GWAS

Calculating principal components

plink:

Remove related individuals (removing bipolar patients)
plink --bfile bp_scz_epigen_genotype_clean_100112 --genome --min 0.1 --out genome_aj
#create text file with individuals who should be excluded
plink --bfile bp_scz_epigen_genotype_clean_100112 --remove AJ_remove_bp.txt --make-bed --out AJ_gwas
#re-check to make sure no related individuals are included
plink --bfile AJ_gwas --genome --min 0.1 --out genome_aj_nobp

Remove SNPs not HWE & missing p<0.05
plink --bfile bp_scz_epigen_genotype_clean_100112 --pheno AJsczpheno.txt --pheno-name toxosero --make-bed --out aj_hwemiss
plink --bfile aj_hwemiss --hardy --out aj_hardy
2 unaffected with p-value < 10^{-6}
plink --bfile aj_hwemiss --test-missing –out aj_missing
#3203 with missing p-value <0.05
plink --bfile AJ_gwas --exclude aj_miss.txt --make-bed --out AJ_gwas_nomiss

Prune data:
Extract autosomal DNA; remove sex chromosomes
write AJ_gwas.bim
plink --bfile AJ_gwas_nomiss --exclude AJ_gwas_exclude.txt --make-bed --out AJ_gwas_autosomal

Independent-pairwise analysis
plink --bfile AJ_gwas_autosomal --indep-pairwise 50 5 0.25 --out AJ_gwas_autosomal_pruned
plink --bfile AJ_gwas_autosomal --extract AJ_gwas_autosomal_pruned.prune.in --make-bed --out AJ_gwas_autosomal_pruned
plink --bfile AJ_gwas_autosomal_pruned --geno 0.0 --make-bed --out AJ_gwas_autosomal_pruned_geno0.0

```
# Extract pruned autosomal genos to .raw file
plink --bfile AJ_gwas_autosomal_pruned_geno0.0 --recodeA --out AJ_PC
```

R:

```
install.packages("qqman")
install.packages("plyr")

prune<-read.table("AJ_PC.raw",h=T)
prune[1:5, 1:5]
dim(prune)
# 1521 32231

#insert FID;IID row name
fid<-prune$FID
iid<-prune$IID
names <-paste(fid, iid, sep=";")
row.names(prune)=names
prune[1:8, 1:9]

#have only genos (remove first 6 columns)
taims<-prune[,c(-1:-6)]
dim(taims)
# 1521 32225

taims[1:5, 1:5]

aims<-na.omit(taims)
dims=dim(aims)
dim(aims)
# 1521 32225

aims[1:10, 1:5]

#double check for missing data
which(is.na(aims))
#0

#calculate principal components
#number of people
n=dims[1]
#number of genos
m=dims[2]
#scale function scales and centers columns of the matrix
```

```
g=scale( aims[1:n,1:m])
out=prcomp( x=t(g), center=FALSE, scale.=FALSE )
dim(out$rotation)
# 1521 1521

#plot to see population structure, good to look at all pop structures (first 10)
plot( x=out$rotation[,1], y=out$rotation[,2],pch=16,xlab="pc 1",ylab="pc 2", cex=0.5)

#creating PC file
temp<-out$rotation
temp[1:5,1:5]
#create IID file
IID<-rownames(temp)
#should have same number of subjects
length(IID)
# 1521
# adds FID, IID back into file
PC<-cbind.data.frame(names, temp)
PC[1:5, 1:5]

top10<- PC[,1:11]
top10[1:5,]
write.csv(top10, "top10PC_AJ_nobp.csv")

# extract PC1 & PC2 from file and place into excel spreadsheet. Add age and sex into
covariate file.
master<-read.csv("aj_master.csv", as.is=TRUE)
master[1:5,1:5]
dim(master)
# 2746 36

PC<-read.csv("top10PC_AJ_nobp.csv", as.is=TRUE)
PC[1:5,]
dim(PC)
# 1521 12

merge.data2<-merge(master, PC, by=c("FID", "IID"))
merge.data2 [1:5, 1:5]
dim(merge.data2)
# 1172 46
write.csv(merge.data2, "AJ_PC_nobp_merge.csv")
```

HapMap

plink:

write AJ_gwas_autosomal_pruned.bim
#remove AT-GC alleles
plink --bfile AJ_gwas_autosomal_pruned --exclude AG_CT_SNPs.txt --make-bed --out AJ_gwas_auto_pruned_noCG_AT
write AJ_gwas_auto_pruned_noCG_AT.bim
#save as text file

plink --bfile hapmap_9pop_genome_noAArelated --extract AJ_gwas_auto_pruned_noCG_AT.txt --make-bed --out hapmap_9pop_local
write hapmap_9pop_local.bim
#save as text file

plink --bfile AJ_gwas_auto_pruned_noCG_AT --extract hapmap_9pop_local.txt --make-bed --out AJ_gwas_auto_pruned_local

plink --bfile hapmap_9pop_local --bmerge AJ_gwas_auto_pruned_local.bed AJ_gwas_auto_pruned_local.bim AJ_gwas_auto_pruned_local.fam --make-bed --out merge_hapmap_AJ

perform MDS analysis
plink --bfile merge_hapmap_AJ --cluster --mds-plot 10 --out HAPMAP_AJ_MDS

R:

mds1<- read.csv("HAMAP_AJ_MDS.csv", header=T)
mds1[1:3,]
table<-table(mds1[,2])
library(plyr)
count(table)
AJ 1521
ASW 43
CEU 112
CHB 84
CHD 85
GIH 88
JPT 86
LWK 79
MEX 50
YRI 110

# Define colors for different populations
colors = c(rep("red", 1521),rep("aquamarine", 43),rep("darkgreen",

```
112),rep("blueviolet",84),rep("lightblue", 85), rep("black", 88),rep("chocolate", 86),
rep("purple", 79),rep("yellow", 50),rep("lightpink",110))

# Plot HapMap
plot(mds1$C1, mds1$C2, col=colors, pch=16,  ylab="MDS-C2", xlab=" MDS-C1")

# Include legend into plot
legend("bottomright", c("AJ","ASW", "CEU","CHB", "CHD", "GIH","JPT","LWK","MEX","YRI"),
fill=c("red","aquamarine","darkgreen","blueviolet","lightblue","black","chocolate","purple","
yellow",
"lightpink"),cex=0.7)
```

Associations' analyses

AJ GWAS toxo assoc (binary) w PC:

plink:

```
plink --bfile bp_scz_epigen_genotype_clean_100112 --covar AJ_PC_nobp.txt --covar-name
PC1, PC2, sex, age --pheno AJsczpheno.txt --pheno-name toxosero --logistic --hide-covar --
out ajtoxo
```

R:

```
library(qqman)
ajtoxo<-read.table("ajtoxo.assoc.logistic", header=TRUE)

# create Q-Q plot
qq(ajtoxo$P)

pvalues=sort(ajtoxo$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")

# Manhattan Plot
score=-log10(ajtoxo$P)

newdata=cbind(ajtoxo, score)
newdata[1:5,]

chr=newdata$CHR
```

```
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9), ylab="AJ
dichotomous toxo sero-conversion (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```

<u>AJ GWAS toxo assoc (quantitative) w PC:</u>

plink:

```
plink --bfile bp_scz_epigen_genotype_clean_100112 --covar AJ_PC_nobp.txt --covar-name
PC1, PC2, sex, age --pheno AJsczpheno.txt --pheno-name ln_z --linear --hide-covar --out
ajqtoxo
```

R:

```
library(qqman)
```

```
ajqtoxo<-read.table("ajqtoxo.assoc.linear", header=TRUE)

# create Q-Q plot
qq(ajqtoxo$P)

pvalues=sort(ajqtoxo$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")


# Manhattan Plot
score=-log10(ajqtoxo$P)

newdata=cbind(ajqtoxo, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9), ylab="AJ
Toxoplasma sero-intensity(-log10 p-value)",
```

```
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)

#Add threshold line
abline(h=3, col="red")
```

AJ GWAS toxo seroconversion status stratified by outcome (SCZ vs control):

Control:

plink:

plink --bfile bp_scz_epigen_genotype_clean_100112 --covar AJ_PC_nobp.txt --covar-name PC1, PC2, sex, age --pheno ajsczpheno-cntrl.txt --pheno-name toxosero --logistic --hide-covar --out ajtoxocntrl

R:

```
library(qqman)
ajtoxocntrl<-read.table("ajtoxocntrl.assoc.logistic", header=TRUE)

# create Q-Q plot
qq(ajtoxocntrl$P)

pvalues=sort(ajtoxocntrl$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")
```

```
# Manhattan Plot
score=-log10(ajtoxocntrl$P)

newdata=cbind(ajtoxocntrl, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr
chr.color[chr.color==2|chr.color==4|chr.color==6|chr.color==8|chr.color==10|chr.color==1
2|chr.color==14|chr.color==16|chr.color==18|chr.color==20|chr.color==22|chr.color==24]
="lightblue"
chr.color[chr.color==1|chr.color==3|chr.color==5|chr.color==7|chr.color==9|chr.color==11
|chr.color==13|chr.color==15|chr.color==17|chr.color==19|chr.color==21|chr.color==23|c
hr.color==25]="darkblue"

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr.color,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9),
ylab=" AJ Toxoplasmosis sero-status among controls (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}
```

```
labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```

Case:

plink:

plink --bfile bp_scz_epigen_genotype_clean_100112 --covar AJ_PC_nobp.txt --covar-name PC1, PC2, sex, age --pheno ajsczpheno-case.txt --pheno-name toxosero --logistic --hide-covar --out ajtoxocase

R:

```
library(qqman)
ajtoxocase<-read.table("ajtoxocase.assoc.logistic", header=TRUE)

# create Q-Q plot
qq(ajtoxocase$P)

pvalues=sort(ajtoxocase$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")

# Manhattan Plot
score=-log10(ajtoxocase$P)

newdata=cbind(ajtoxocase, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25
```

```
#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr
chr.color[chr.color==2|chr.color==4|chr.color==6|chr.color==8|chr.color==10|chr.color==12|chr.color==14|chr.color==16|chr.color==18|chr.color==20|chr.color==22|chr.color==24]="lightblue"
chr.color[chr.color==1|chr.color==3|chr.color==5|chr.color==7|chr.color==9|chr.color==11|chr.color==13|chr.color==15|chr.color==17|chr.color==19|chr.color==21|chr.color==23|chr.color==25]="darkblue"

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr.color,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9),
ylab="AJ Toxoplasmosis sero-status among cases (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```

AJ GWAS toxo assoc (quantitative) w PC stratified by outcome (SCZ vs control):
Control:

plink:

```
plink --bfile bp_scz_epigen_genotype_clean_100112 --covar AJ_PC_nobp.txt --covar-name
PC1, PC2, sex, age --pheno ajsczpheno-cntrl.txt --pheno-name ln_z --linear --hide-covar --out
ajqtoxocntrl

library(qqman)
ajqtoxocntrl<-read.table("ajqtoxocntrl.assoc.linear", header=TRUE)

# create Q-Q plot
qq(ajqtoxocntrl$P)
```

```
pvalues=sort(ajqtoxocntrl$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")

# Manhattan Plot
score=-log10(ajqtoxocntrl$P)

newdata=cbind(ajqtoxocntrl, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr
chr.color[chr.color==2|chr.color==4|chr.color==6|chr.color==8|chr.color==10|chr.color==1
2|chr.color==14|chr.color==16|chr.color==18|chr.color==20|chr.color==22|chr.color==24]
="lightblue"
chr.color[chr.color==1|chr.color==3|chr.color==5|chr.color==7|chr.color==9|chr.color==11
|chr.color==13|chr.color==15|chr.color==17|chr.color==19|chr.color==21|chr.color==23|c
hr.color==25]="darkblue"

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr.color,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9),
ylab="AJ Toxoplasmosis sero-status among controls (-log10 p-value)",
```

```
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```

Case:

plink:

```
plink --bfile bp_scz_epigen_genotype_clean_100112 --covar AJ_PC_nobp.txt --covar-name
PC1, PC2, sex, age --pheno ajsczpheno-case.txt --pheno-name ln_z --linear --hide-covar --out
ajqtoxocase
```

R:

```
library(qqman)
ajqtoxocase<-read.table("ajqtoxocase.assoc.linear", header=TRUE)

# create Q-Q plot
qq(ajqtoxocase$P)

pvalues=sort(ajqtoxocase$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")

# Manhattan Plot
score=-log10(ajqtoxocase$P)

newdata=cbind(ajqtoxocase, score)
```

```
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr
chr.color[chr.color==2|chr.color==4|chr.color==6|chr.color==8|chr.color==10|chr.color==1
2|chr.color==14|chr.color==16|chr.color==18|chr.color==20|chr.color==22|chr.color==24]
="lightblue"
chr.color[chr.color==1|chr.color==3|chr.color==5|chr.color==7|chr.color==9|chr.color==11
|chr.color==13|chr.color==15|chr.color==17|chr.color==19|chr.color==21|chr.color==23|c
hr.color==25]="darkblue"

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr.color,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9),
ylab="AJ Toxoplasmosis sero-intensity among cases (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```

Replication dataset GWAS

Calculating principal components

plink:
Checked against omniquad PC, found to be different so had to calculate own PC
Remove related individuals (removing bipolar patients)
plink --bfile toxo_GTP_gwas --genome --min 0.1 --out genome_gtp
No relatedness found in the dataset

Remove SNPs not HWE & missing p<0.05
plink --bfile toxo_GTP_gwas --pheno phenogtp286.txt --pheno-name Toxo --make-bed --out gtp_hwemiss
plink --bfile gtp_hwemiss --hardy --out gtp_hardy
1 unaffected SNP falls outside HWE
plink --bfile gtp_hwemiss --test-missing --out gtp_missing
#1459 with missing p-value <0.05
plink --bfile toxo_GTP_gwas --exclude gtp_miss.txt --make-bed --out GTP_gwas_nomiss

Extract autosomal DNA; remove sex chromosomes
Create text file only containing sex chromosomes to exclude.
write toxo_GTP_gwas.bim
plink --bfile GTP_gwas_nomiss --exclude GTP_gwas_exclude.txt --make-bed --out GTP_gwas_autosomal

Independent-pairwise analysis
plink --bfile GTP_gwas_autosomal --indep-pairwise 50 5 0.05 --out GTP_gwas_autosomal_pruned
plink --bfile GTP_gwas_autosomal --extract GTP_gwas_autosomal_pruned.prune.in --make-bed --out GTP_gwas_autosomal_pruned
plink --bfile GTP_gwas_autosomal_pruned --geno 0.0 --make-bed --out GTP_gwas_autosomal_pruned_geno0.0

Extract pruned autosomal genos to .raw file
plink --bfile GTP_gwas_autosomal_pruned_geno0.0 --recodeA --out GTP_PC

R:
install.packages("qqman")

prune<-read.table("GTP_PC.raw",h=T)
prune[1:5, 1:5]
dim(prune)
# 286 52497

```
#make SIDs row names
row.names(prune)=prune$IID
prune[1:8, 1:9]

#have only genos (remove first 6 columns)
taims<-prune[,c(-1:-6)]
dim(taims)
# 286 52491

taims[1:5, 1:5]

aims<-na.omit(taims)
dims=dim(aims)
dim(aims)
# 286 52491

aims[1:10, 1:5]

#used to double check non-missing data
which(is.na(aims))
#0

# calculate principal components
#number of people
n=dims[1]
#number of genos
m=dims[2]
#scale function scales and centers columns of the matrix
g=scale( aims[1:n,1:m])
out=prcomp( x=t(g), center=FALSE, scale.=FALSE )
dim(out$rotation)
# 286 286

#plot to see population structure
plot( x=out$rotation[,1], y=out$rotation[,2],pch=16,xlab="pc 1",ylab="pc 2", cex=0.5)

#creating PC file
temp<-out$rotation
temp[1:5,1:5]
#create IID file
IID<-rownames(temp)
#should have same number of subjects
length(IID)
```

```
# 286
# adds IID back into file
PC<-cbind.data.frame(IID, temp)
PC[1:5, 1:5]

top10<- PC[,1:11]
top10[1:5,]

#add FID back into file
sid<-prune[,1:2]
sid[1:5,]
top10PC_GTP <- merge(sid, top10, by="IID")
top10PC_GTP[1:4, 1:8]
write.csv(top10PC_GTP, "top10PC_GTP.csv")

# extract PC1 & PC2 from file and place into excel spreadsheet. Add age and sex into
covariate file.
```

HapMap

plink:

```
write GTP_gwas_autosomal_pruned.bim
# remove AT-GC alleles
plink --bfile GTP_gwas_autosomal_pruned --exclude GTP_AG_CT_SNPs.txt --make-bed --out
GTP_gwas_auto_pruned_noCG_AT
write GTP_gwas_auto_pruned_noCG_AT.bim
# save as text file

plink --bfile hapmap_9pop_genome_noAArelated --extract
GTP_gwas_auto_pruned_noCG_AT.txt --make-bed --out GTP_hapmap_9pop_local
write GTP_hapmap_9pop_local.bim
# save as text file

plink --bfile GTP_gwas_auto_pruned_noCG_AT --extract GTP_hapmap_9pop_local.txt --make-
bed --out hapmap_GTP

plink --bfile GTP_hapmap_9pop_local --bmerge hapmap_GTP.bed hapmap_GTP.bim
hapmap_GTP.fam --make-bed --out merge_hapmap_GTP

# error occurred so we need to flip
plink --bfile GTP_hapmap_9pop_local --flip merge_hapmap_GTP-merge.missnp --make-bed -
-out GTP_flipped

plink --bfile GTP_flipped --bmerge hapmap_GTP.bed hapmap_GTP.bim hapmap_GTP.fam --
```

make-bed --out merge_hapmap_GTP

perform MDs analysis
plink --bfile merge_hapmap_GTP --cluster --mds-plot 10 --out HAPMAP_GTP_MDS

R:
mds1<- read.csv("HAPMAP_GTP_MDS.csv", header=T)
mds1[1:3,]
table<-table(mds1[,2])
install.packages('plyr')
library(plyr)
count(table)
ASW 43
CEU 112
CHB 84
CHD 85
GIH 88
GTP 286
JPT 86
LWK 79
MEX 50
YRI 110

# Define colors for different populations
colors = c(rep("red", 286),rep("aquamarine", 43),rep("darkgreen", 112),rep("blueviolet",84),rep("lightblue", 85), rep("black", 88),rep("chocolate", 86), rep("purple", 79),rep("yellow", 50),rep("lightpink",110))

# Plot HapMap
plot(mds1$C1, mds1$C2, col=colors, pch=16, ylab="MDS-C2", xlab=" MDS-C1")

# Include legend into plot
legend("bottomleft", c("ASW", "CEU","CHB", "CHD", "GIH","GTP","JPT","LWK","MEX","YRI"), fill=c("aquamarine","darkgreen","blueviolet","lightblue","black","red","chocolate","purple","yellow","lightpink"),cex=0.7)

Associations' analyses
GTP GWAS toxo assoc (binary) w PC:
plink:
plink --bfile toxo_GTP_gwas --covar GTP_PC.txt --covar-name PC1, PC2, sex, age --pheno phenogtp286.txt --pheno-name Toxo --logistic --hide-covar --out gtptoxo

R:
gtptoxo<-read.table("gtptoxo.assoc.logistic", header=TRUE)

```
# create Q-Q plot
library(qqman)
qq(gtptoxo$P)

pvalues=sort(gtptoxo$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")


# Manhattan Plot
score=-log10(gtptoxo$P)

newdata=cbind(gtptoxo, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9), ylab="GTP
dichotomous Toxoplasma sero-status (-log10 p-value)",
```

```
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)

#Add threshold line
abline(h=3, col="red")
```

GTP GWAS toxo assoc (quantitative) w PC:

plink:

```
plink --bfile toxo_GTP_gwas --covar GTP_PC.txt --covar-name PC1, PC2, sex, age --pheno
phenogtp286.txt --pheno-name indexln --linear --hide-covar --out gtpqtoxo
```

R:

```
gtpqtoxo<-read.table("gtpqtoxo.assoc.linear", header=TRUE)

# create Q-Q plot
library(qqman)
qq(gtpqtoxo$P)

pvalues=sort(gtpqtoxo$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")


# Manhattan Plot
score=-log10(gtpqtoxo$P)
```

```
newdata=cbind(gtpqtoxo, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9), ylab="GTP
Toxoplasmosis sero-intensity(-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)

#Add threshold line
abline(h=3, col="red")
```

GTP GWAS toxo seroconversion (binary) stratified on SCZ status:

Controls (clean):

plink:

```
plink --bfile toxo_GTP_gwas --covar GTP_PC.txt --covar-name PC1, PC2, sex, age --pheno
phenogtp286-cntrl.txt --pheno-name Toxo --logistic --hide-covar --out gtptoxocntrl
```

R:

```
gtptoxocntrl <-read.table("gtptoxocntrl.assoc.logistic", header=TRUE)

# create Q-Q plot
library(qqman)
qq(gtptoxocntrl$P)
pvalues=sort(gtptoxocntrl$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")

# Manhattan Plot
score=-log10(gtptoxocntrl$P)

newdata=cbind(gtptoxocntrl, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}
```

```
chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9), ylab=" GTP
Toxoplasma sero-status in clean controls (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)

#Add threshold line
abline(h=3, col="red")
```

Case:

plink:

```
plink --bfile toxo_GTP_gwas --covar GTP_PC.txt --covar-name PC1, PC2, sex, age --pheno
phenogtp286-case.txt --pheno-name Toxo --logistic --hide-covar --out gtptoxocase
```

R:

```
gtptoxocase<-read.table("gtptoxocase.assoc.logistic", header=TRUE)

# create Q-Q plot
library(qqman)
qq(gtptoxocase$P)

pvalues=sort(gtptoxocase$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
```

```
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")


# Manhattan Plot
score=-log10(gtptoxocase$P)

newdata=cbind(gtptoxocase, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr
chr.color[chr.color==2|chr.color==4|chr.color==6|chr.color==8|chr.color==10|chr.color==1
2|chr.color==14|chr.color==16|chr.color==18|chr.color==20|chr.color==22|chr.color==24]
="lightblue"
chr.color[chr.color==1|chr.color==3|chr.color==5|chr.color==7|chr.color==9|chr.color==11
|chr.color==13|chr.color==15|chr.color==17|chr.color==19|chr.color==21|chr.color==23|c
hr.color==25]="darkblue"

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr.color,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9),
ylab="GTP Toxoplasma sero-status among cases (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
```

```
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```

GTP GWAS toxo assoc (quantitative) w PC:
Controls (clean):

plink:

```
plink --bfile toxo_GTP_gwas --covar GTP_PC.txt --covar-name PC1, PC2, sex, age --pheno
phenogtp286-cntrl.txt --pheno-name indexln --linear --hide-covar --out gtpqtoxocntrl
```

R:

```
gtpqtoxocntrl<-read.table("gtpqtoxocntrl.assoc.linear", header=TRUE)

# create Q-Q plot
library(qqman)
qq(gtptoxocntrl$P)

pvalues=sort(gtpqtoxocntrl$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")

# Manhattan Plot
score=-log10(gtpqtoxocntrl$P)

newdata=cbind(gtpqtoxocntrl, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
```

```
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9), ylab="GTP
Toxoplasma sero-intensity among clean controls (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)

#Add threshold line
abline(h=3, col="red")
```

cases:

plink:

```
plink --bfile toxo_GTP_gwas --covar GTP_PC.txt --covar-name PC1, PC2, sex, age --pheno
phenogtp286-case.txt --pheno-name Indexln --linear --hide-covar --out gtpqtoxocase
```

R:

```
gtpqtoxocase<-read.table("gtpqtoxocase.assoc.linear", header=TRUE)

# create Q-Q plot
library(qqman)
qq(gtpqtoxocase$P)

pvalues=sort(gtpqtoxocase$P)
n=length(pvalues)
q=(1:n)/(n+1)
log10.pvalues=-log10(pvalues)
```

```
log10.q=-log10(q)

#create limits
lq=length(q)
qq=1:lq
upper=qbeta(.95,qq,lq-qq+1)
lower=qbeta(.05,qq,lq-qq+1)
points(-log10(q),-sort(log10(upper),na.last=FALSE),type='l',lty=2, col="green")
points(-log10(q),-sort(log10(lower),na.last=FALSE),type='l',lty=2, col="green")


# Manhattan Plot
score=-log10(gtpqtoxocase$P)

newdata=cbind(gtpqtoxocase, score)
newdata[1:5,]

chr=newdata$CHR
pos=newdata$BP
score=newdata$score

# Sort everything by chr and pos.
k=order(chr,pos)
chr=chr[k]
pos=pos[k]
score=score[k]

# Define range of chromosomes
cmax=25

#Define genomewide position vector
genomepos=pos
for (i in 2:cmax) {
  genomepos[chr==i]=genomepos[chr==i]+max(genomepos[chr==i-1])+100
}

chr.color=chr

#Basic Manhattan plot without x axis
plot(genomepos,score,col=chr.color,pch=19, cex=.2+(score*.05),xaxt='n', ylim=c(0,9),
ylab="GTP Toxoplasma sero-intensity among cases (-log10 p-value)",
xlab="Chromosome", cex.lab=1.3, cex.axis=1.3)

#Add x axis
meds=matrix(0,cmax)
labs=matrix(NA,cmax)
tix=matrix(0,cmax+1)
labs_null=matrix(NA,cmax+1)

for (i in 1:cmax) {
```

```
  tix[i+1]=max(genomepos[chr==i])
  meds[i]=median(genomepos[chr==i])
  labs[i]=i
}

labs[23]="X"
axis(side=1,at=meds,lwd.ticks=0,labels=labs, cex.axis=0.7)
axis(side=1,at=tix,lwd.ticks=1,labels=labs_null)
```