Abstract

Imputation of unknown genotypes is becoming a standard procedure in exploratory genetic association studies. Imputation is accomplished by comparing observed data from the study population to reference panels of individuals who are from a genetically similar population and genotyped at a dense set of polymorphic sites. Linkage disequilibrium within the reference panels is used to construct haplotypes and extrapolate allelic correlations in the test sample. Imputation has been shown to be accurate for the inference of genotypes at unobserved SNPs, as well as for quality control measures at genotyped locations. Imputing genotypes also allows cohorts that were genotyped on different platforms to be combined in a joint or meta-analysis. One of the most widely used imputation software packages is MaCH (http://csg.sph.umich.edu//abecasis/mach/). MaCH uses a powerful and accurate Markov chain-based algorithm, however its usability is lacking. MaCHTools allows the user to streamline their workflow with MaCH through input file specification, error checking, and QC measures. MaCHTools began as a series of Java scripts used to check input files and QC raw data as an initial step before imputing additional genotypes in MaCH. This set of scripts became invaluable to the GWAS workflow, but they were unpolished and ill-suited for public release to benefit the scientific community. This project aimed to bundle the scripts into a single executable program that provides a graphical user interface (GUI) to facilitate use by students and researchers to aid in streamlining the GWAS workflow. Additional functionalities include more efficient launching of jobs to compute clusters and compatibility with different Linux job handlers, the ability to easily switch between different GWAS projects including switching between different genotype data and reference datasets, more simplistic specification of parameters and thresholds, and several other usability improvements.

The GWAS workflow that includes dataset preparation with MaCHTools coupled with haplotype estimation and imputation with MaCH was validated by replicating results from a published study of the genetic basis of Alzheimer's endophenotypes in the Texas Alzheimer's Research and Care Consortium. A similar analysis was then performed to determine the genetic basis of D, a latent variable that represents the dementing process.

MaCHTools: Additional functionality for the imputation software MaCH

DISSERTATION

Presented to the Graduate Council of the
University of North Texas Health Science Center
At Fort Worth
In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

By

Jeffrey S. Mitchel, Jr., BS, MS

Fort Worth, TX

November 18, 2016

Acknowledgements

I have many acknowledgements to make to those in both my professional and personal networks that have suffered with me on this road. Firstly, Dr. Robert Barber, my major professor, supported me through the bulk of my graduate work. I watched Bob succeed in his career while he consistently set me up to succeed in mine. He put me on airplanes and drove me across the country to get me in front of people that have been and will continue to be incredible mentors and friends for life. Dr. Kirk Wilhelmsen at UNC Chapel Hill has been a tremendous resource, role model, and friend. Kirk, bless his patient heart, literally taught me, among many things, how to read and write. I arrived at UNTHSC fluent in English, but computers don't speak English. Kirk took it upon himself to teach me Java, which in today's world is arguably more important than English, as Java is spoken by more computers than there are people, and computers are often more tolerable to talk to. When I got to school, I wanted to learn to code and to apply software development to genetics, and without Kirk, I would have never accomplished that. Dr. Nicole Phillips was just Mrs. Nicole Phillips when I first started working with her. I'm thankful for her constant support and patience as I struggled through many of the same struggles she dealt with in bioinformatics, and it's been fun to watch her succeed in her professional life. My committee member Fan Zhang taught me bioinformatics and machine learning early on and constantly supported me with various software tools along the way. Dr. Jeff Tilson at RENCI tolerated me mutilating MaCHTools, which he used extensively before I repeatedly broke it. His patience and willingness to help was invaluable. Drs. Don Royall and Ray Palmer at UT Health Science Center in San Antonio provided D scores for TARCC participants. I need to acknowledge Dr. Rhonda Roby for having the foresight to pair me with Bob, and my university member Dr. Ann Schreihofer for ensuring that I am fairly treated by my committee. The Neurobiology of Aging Fellowship supported me for two years of my work.

The decision to pursue a PhD is a selfish one because you lean on so many other people along the way. I started the program with a girlfriend, advanced to engagement, and I will graduate with a wife, Anna. Anna has always supported and encouraged me. She has always been impressed by me and tolerated me when I was less than enjoyable to be around. I also want to acknowledge the support of my mom, dad, and sister who are at this point sure I will just remain in school forever. This journey would have been much darker and more perilous without you all.

Table of Contents

List of Tables

List of Illustrations

Chapter I

INTRODUCTION AND LITERATURE REVIEW

Alzheimer's disease (AD) is the most common form of age-related neurodegenerative dementia and one of the most serious public health issues in the United States. Among individuals 65 years and older, there is an estimated prevalence of AD ranging between 6-12%. According to the Alzheimer's Association, over 5 million Americans were living with a diagnosis of late onset AD in 2012, and this figure is expected to double to 10 million over the next 25 years. AD is the 6th leading cause of death in America, and the 5th leading cause of death among those over 65. AD is a large financial burden in the US, with costs exceeding $183 billion annually, and an additional $210 billion in unpaid care provided by the friends and family of patients. There is no cure or care that can be given to patients to remedy or slow the progress of the disease. All therapies provide only symptomatic relief.

AD is a progressive, and eventually fatal, neurodegenerative disease. Its diagnosis is speculative based on symptoms, and only confirmed post-mortem by the presence of extracellular plaques formed from cleaved amyloid precursor protein and intracellular neurofibrillary tangles caused by hyperphosphorylated microtubule associated protein tau[1]. However, it has been shown that levels of amyloid beta peptide and phosphorylated tau in the cerebrospinal fluid have been consistent with autopsy findings[2]. These plaques and tangles are thought to be neurotoxic, and result in the progressive loss of neurons and synapses. After the disease has run its course, the brain is an atrophic version of its former, healthy state. AD initially presents as benign short term and spatial memory loss, often termed mild cognitive impairment (MCI)[3]. MCI can only be detected by careful examination and testing, and does not interfere with daily activities[4]. As the disease progresses, additional symptoms may include irritability, aggression, confusion, loss of long term memory, and language problems[5]. Eventually, the patients will be completely unable to care for themselves. The neuropathologies associated with AD begin in the hippocampus and spread to the cerebral cortex and subcortical regions[6]. Neurodegeneration can be seen in the parietal lobes, frontal cortex, and cingulate gyrus[7]. AD progression can be visualized by MRI and PET analyses to document the atrophy of these brain regions[8].

Broadly, there are two forms of AD: familial, or early onset; and sporadic, or late onset. The familial form of AD is a rare form, only representing 5% of the disease burden. This form of the disease is inherited in a Mendelian dominant manner[9]. In contrast, while genetic variation plays a significant role in the development in late-onset AD, non-genetic, environmental factors are also important. This review will focus on late onset AD (LOAD).

LOAD Genetics

As described earlier, LOAD is etiologically heterogeneous, increases in prevalence with age with a lifetime risk of 1 in 10, and results from many genetic and environmental factors[4]. A great number of these genetic variants have been studied, however over 50% of the genetic variation remains unidentified[10]. The estimated heritability for AD is between 50-75%[11]. Until recently, a single gene had been associated with an increased risk of AD. The epsilon 4 allele of apolipoprotein E (*APOE*4) has been indicated as a reliable risk factor for AD with an odds ratio ranging from 10-20 in homozygotes when compared to *APOE*4 negative individuals[12].

The *APOE* protein is a 299 amino acid protein that is synthesized in the liver, but it is also synthesized in the nervous system by astroglia and microglia. It is involved in the transport of lipids, lipoproteins, fat-soluble vitamins, and cholesterol into the blood. *APOE* is a ligand for several receptors involved in lipid metabolism, including the low-density lipoprotein (LDL), LDL-related protein (LRP), and very-low-density lipoprotein (VLDL)[13]. These receptors are preferentially expressed in neurons[14]. The nature of the *APOE* variation exists at two loci in the 3,597 nucleotide gene. At residue 112 and 158, the E4 isoform contains an arginine/arginine residue, respectively. In contrast, the E2 allele exists as a cysteine/cysteine residue at these loci, and exhibits protective effects against AD. The E3 alleles consists of an arginine residue at the 112 position and a cysteine residue at position 158[15].

In the human population, the E4 allele frequency varies widely by population, with the highest frequencies found in higher and lower latitudes. The E4 allele has a lower frequency near the Latin and African equators[16]. The effect size for *APOE*4 is one of the largest among all multifactorial, complex diseases. As stated earlier, homozygotes are 10-20 times more likely to develop AD when compared to *APOE*4 negative individuals[12]. This risk varies with sex, with women being more susceptible[17]. However, the *APOE*4 allele exhibits incomplete penetrance, as *APOE4* positive individuals often to not develop AD[18].

The clusterin (*CLU*) gene has also been associated with AD risk, with variant rs11136000 showing protective effects (OR = 0.92)[19]. Clusterin is a widely expressed apolipoprotein that is thought to have heat shock protein-like chaperone properties. The *CLU* gene is located on chromosome 8, and is evolutionarily conserved across mammalian taxa. The gene is 16Kb of DNA, and contains 9 exons. It is very similar to *APOE*, and is often referred to as *APOJ*[20]. Clusterin is similar to *APOE* in that they are abundant in the brain, particularly areas associated with AD and the cerebrospinal fluid. Both proteins also are involved in the clearance of beta-amyloid plaques. Clusterin's association with AD risk has been pinpointed to a T/C SNP, rs11136000. Population substructure for the rs11136000 single nucleotide polymorphism, or SNP, varies by ethnicity, with higher frequencies of the C allele in American (63%), Asian (80%), and European (61%) populations, and low frequencies in African populations (41%)[21].

Phosphatidylinositol binding clathrin assembly protein (*PICALM*) is a widely expressed protein that is involved in retrieval of membranes in the synaptic vesicle. This protein functions via clathrin-mediated endocytosis, which is a critical step in the movement of many proteins and lipids, namely the internalization of uncleaved amyloid precursor protein. This protein was first studied in association with several forms of leukemia resulting from a chromosomal translocation[22]. *PICALM*'s association with AD is protective and results from a T/C SNP, rs3851179 (OR = 0.8)[19]. The frequency of the minor allele (T) in the population is 0.33. The T allele is most common in individuals of Asian descent (42%) and least common in those of African descent (11%)[21].

In contrast with the previously outlined AD genes that lie within lipid metabolism network, complement receptor 1 (*CR1*) is a gene within the inflammation network that has been associated with AD risk. *CR1,* also known as CD35, is a member of the complement activation family, and is expressed on erythrocytes, leukocytes, and splenic follicular dendritic cells[23]. *CR1* is the receptor for C3b and C4b in humans, and therefore is the principle component in the clearance of opsonized immune complexes[24]. *CR1* is often implicated as a negative regulator of the complement cascade. *CR1*'s association with AD risk is two-fold. Firstly, there is a significant A/G SNP, rs6656401, with a minor allele frequency of 0.09 (A)[21]. This allele is over represented in those of European descent (19%), and particularly rare in those of African descent (1%). This allele increases risk of AD with an odds ratio of 1.2[25]. Secondly, as individuals age, levels of *CR1* decrease. This may cause a detriment to the clearance of beta-amyloid plaques via the complement cascade.

GALP, which encodes galanin-like protein, was associated with AD in a 2007 genome-wide association study (GWAS, further discussed in the next section). The associated variant, rs3745833, creates a non-synonymous amino acid substitution at the 72nd residue (Ile72Met) in exon 4. The common minor allele C, which has a minor allele frequency of 48% in Caucasians increases AD risk by 10%. Galanin-like protein is a neurotransmitter than binds to galanin receptors 1, 2, and 3, prevent long-term potentiation in the hippocampus. Therefore, overexpression of *GALP* and similar peptides could exacerbate AD symptoms[26].

PGBD1 was also associated with AD in the same 2007 GWAS via a significant coding SNP, rs3800324. This SNP codes for a non-synonymous mutation in exon 5 and is relatively rare with a MAF of 6%. Rs3800324, which increases AD risk by 20%, is expressed in the brain, however its function is not fully understood. It is believed that *PGBD1* interacts with a nearby gene that encodes a zinc-finger protein, which plays a role in transcriptional regulation[26].

TKN1, or tyrosine-kinase, non-receptor 1, was previously known as thirty-eight negative kinase 1'. SNP rs1554948 was found to be associated with AD risk in a protective manner, reducing risk by 15%. The minor allele is present at a frequency of ~48% in the population. Functionally, this gene is involved in activating tumor necrosis factor alpha, or TNFα, which is involved in programmed cell death. *TKN1* and rs1554948 were one of the strongest hits in the 2007 study by Grupe *et al* (p-value = $2x10^{-4}$)[26].

GAB2, which encodes GRB2-associated binding protein 2, was found to be associated with AD in a later study in 2007. SNP rs10793294 confers a reduced risk of ~50%. *GAB2* is a highly conserved protein scaffolding gene that affects kinases that may be responsible for phosphorylating tau, and therefore the generation the neurofibrillary tangles[27,28].

A set (rs10868366, rs7019241) of intronic SNPs near GOLM1 was associated with AD risk in a 2008 GWAS. GOLM1 encodes golgi membrane protein 1, however the functional effects of these SNPs has not yet been postulated[29].

FAM113B encodes family with sequence similarity 113, member B. Not much is known about this protein, however rs11610206 was associated with AD risk in a 2009 study. The authors theorized that a proximal gene encoding a vitamin D receptor may underlie the association, however that gene is 600kb from the signal and not in linkage disequilibrium with it[30].

PCDH11X encodes protocadherin 11 X-linked. While this gene is highly conserved, there are no proven functional consequences of the AD-associated SNP, rs2573905. It has been postulated that protocadherins are substrates for gamma secretase, and may compete with amyloid-precursor protein for gamma secretase[31].

Additional AD-associated genes with smaller effect sizes and more rare minor alleles discovered by GWAS include *ACAN, BCR, CTSS, EBF3, FAM63A*[26], *GWA_14q32.13, GWA_15q21.2, GWA_7p15.2, GWA_9p24.3*[26,29], *LMNA, LOC651924, MYH13, PCK, TRAK2*, and *UBD*[26].

Genome-Wide Association Studies

Over the past two decades, the multitude of AD etiologies has been associated with over 2000 genes. However, it wasn't until recently that two large-scale genome-wide association studies (GWAS) were utilized to replicate associations between LOAD and genes other than *APOE*. These studies found associations in the aforementioned genes, *CLU*, *PICALM*, *CR1*, and others[25,32]. The development of large-scale genotyping platforms, along with the availability of large SNP databases has enabled researchers to conduct massive studies that test between hundreds of thousands and millions of SNPs, while calculating their statistical significance and predictive power for the trait of interest.

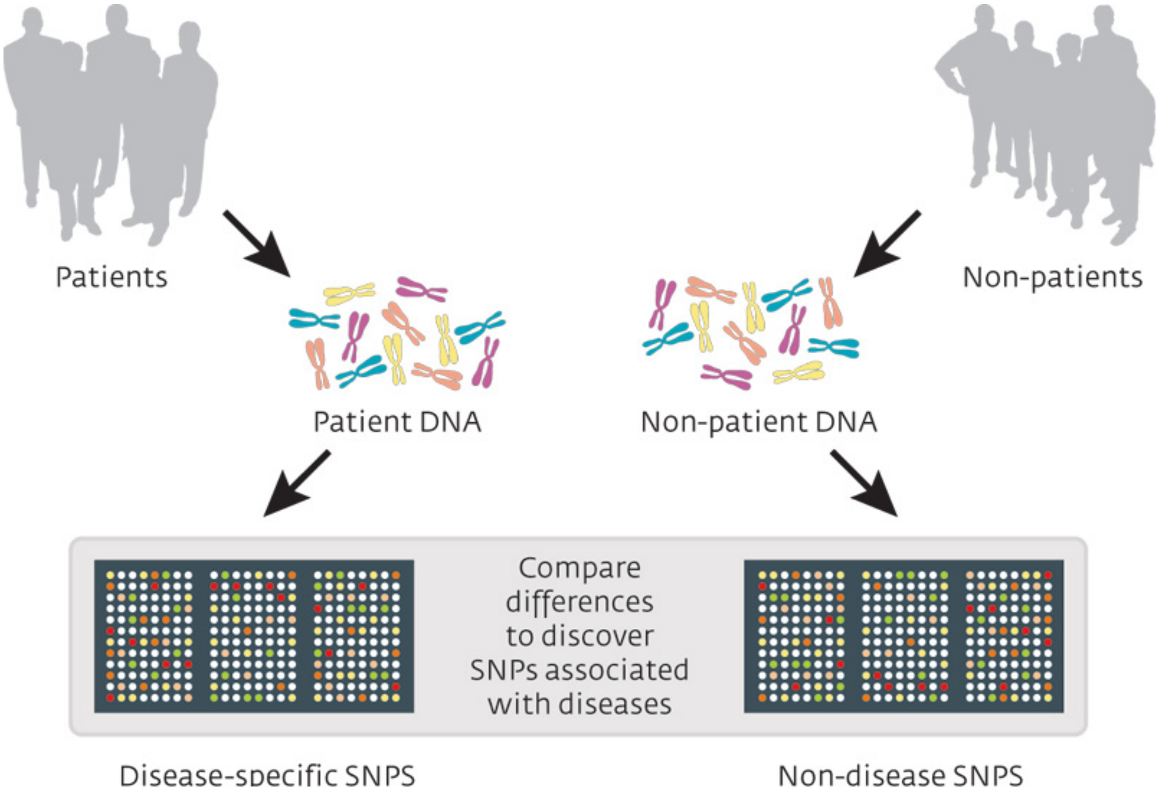


Figure 1. GWAS diagram. Genotype data for patients, or cases are compared with that of non-patients, or controls to ascertain SNPs associated with disease

The first GWAS was completed in 2005, and aimed to associate SNPs with age-related macular degeneration. This study consisted of 96 cases and 50 controls, and despite a very small sample size, the authors were able to detect two significant SNPs with altered allele frequencies between the two study groups.

Typically, GWA studies consist of cases, participants that do not exhibit the trait of interest, and controls, or participants that have the trait of interest. These two groups are ideally as similar as possible, only differing in disease status. The process of selecting cases revolves around enrichment for disease-associated alleles. Including extreme or familial cases as well as attempting to reduce heterogeneity of the phenotype within the cohort often accomplishes this[33]. These practices will, in theory, increase power in the cohort, however poorly understood genetic architectures of complex diseases can make implementing these practices difficult[34]. The most common problem when assembling a control group is misclassification of the individuals, or latent diagnoses in the control individuals later in the study. Control participants require intensive screening to be sure that they do not exhibit the trait of interest. Another pitfall is selecting ‘hypernormal’ participants that are not representative of the population. For example, selecting extremely underweight individuals for a control group in an obesity study may cause false positives associated with medical conditions related to being underweight. Depending on the disease of interests, covariates to control will differ, but some common examples include age, sex, and education. One covariate of importance is population substructure. If a disease that is more prevalent in one population, the case group will be enriched for that population, while the control group will be lacking. Because of this discrepancy in population substructure of the case-control groups, SNPs that are of higher frequency in in the case population may be falsely associated with the disease, when in reality they are only associated with the susceptible lineage. The most common way to accomplish this is by utilizing principle component analysis, which combines the dimensions of a multidimensional data set into principle components. In this case, population substructure is combined into a single principle component, and this component is controlled for in the analysis[35].

GWA studies that are not studying disease status may not employ a case-control design. Some diseases or traits are continuous variables, and association statistics can be calculated using regression analyses instead of Chi-squared tests used in case-control studies. Examples of these phenotypes include age-of-onset for an age-related disease, or conditions with a broad range of severity, given that there is an accurate and consistent method of categorizing these heterogeneous phenotypes.

Once the cohort is assembled and the type of study is determined, the participants are genotyped on a commercially available array. The two main providers of such arrays are Illumina and Affymetrix, however there are other smaller manufacturers. These assays genotype patients at a set number of loci that can range from 500,000 to 5,000,000 or more SNPs[36]. There are different types of assays that focus on loci known to be related to a particular phenotype, while other assays aim for genome-wide coverage.

Each step in the GWAS workflow carries an emphasis on reducing error and bias. The introduction of error or bias at any step can cause extreme values of the association statistics, which will create false positives or globally inflated p-values. After careful assembly of the cohort, the next scrupulous step is genotyping and the subsequent quality control measures implemented to ensure the genotype data are clean and accurate. Raw data from a genotyping assay is typically converted or ‘called’ using an automated algorithm for computing posterior probabilities for each possible genotype given the data from the assay[37]. An important quality control step is to exclude loci with low genotyping success or call rates [38]. Once all well-called genotypes are collected, further quality control measures can be performed on the data. SNPs that deviate from Hardy-Weinberg equilibrium (HWE) are excluded, however there is some debate as to the efficacy or necessity of this practice. Some groups argue that typical GWAS cohorts are underpowered for detecting poor genotype qualities using HWE, while others suggest only excluding SNPs that show extreme departures from equilibrium[39,40]. Further quality control measures include detecting ancestry that deviates from that reported by the participants, duplicate or inverted/swapped samples, cross contamination, or data that suggests patterns of relatedness among the participants[38].

Following thorough quality control measures, association statistics can be generated for each SNP based on the phenotype data observed for each participant. An invaluable tool for this process is PLINK, which, among many other things, generates the association statistics for binary and continuous traits. Data can be visualized using QQ plots, which plot each p-value against the p-value expected by random chance. The observed p-values should approximately match the expected p-values, except the select few extremely low p-values that represent the significant SNPs resolved in the analysis. QQ plots can help detect inflated p-vales and false positives that would indicate error and bias somewhere in the pipeline. The Manhattan plot, named for its resemblance of a cityscape, is a dot plot that is useful for visualizing GWAS results along the axis of the genome, from chromosome 1 to 22. P-values are –log transformed so that dots that lie toward the top of the plot represent the most significant hits.

Further investigation of the significant hits seen in the Manhattan plots can be visualized in a more local context with LocusZoom, which is a Manhattan plot that is localized to a user-defined window on either side of the signal, allowing the visualization of nearby genes.

As mentioned, there are a number of pitfalls that can impart bias and error along the GWAS workflow, and another pitfall of GWAS is the limited coverage of the genome. While a large genotyping array can interrogate 5 million SNPs, there are tens of millions of SNPs in the human genome. This creates a problem of limited coverage of the possible variability that could contribute to disease risk. While sequencing may capture this additional variability, it is expensive and time consuming. Another way to capture this additional variability is through genetic imputation, discussed in the next section.

Genetic Imputation

Imputation is the practice of estimating missing or additional data using observed data. It is a practice common to all disciplines that utilize large datasets, and a common application in statistics would be to impute a missing data point by simply taking the mean of the two adjacent data points. A simplified example is the popular game show Wheel of Fortune, where contestants use given letters in the puzzle to impute the missing letters in order to solve the puzzle. A key part of this practice is the context given with which the additional data are estimated. In the Wheel of Fortune example, this context is the letters that are initially given to start the puzzle. Without context, a much larger proportion of the alphabet could theoretically fit in the puzzle, however the context allows for a narrowing of possibilities to increase the likelihood of an accurate estimation. In genetics, this context is a combination of the observed genotypes from an array and a reference sequence typed at a dense set of markers.

Imputation of unknown genotypes is becoming a standard procedure in exploratory genetic association studies. Genetic imputation is accomplished by comparing observed data from the study population to reference panels of individuals who are from a genetically similar population and genotyped at a dense set of polymorphic sites. Linkage disequilibrium within the reference panels is used to construct haplotypes and extrapolate allelic correlations in the test sample[41]. Imputation has been shown to be accurate for the inference of genotypes at unobserved SNPs, as well as for quality control measures to validate and correct data at genotyped locations[42]. Imputing genotypes also allows cohorts that were genotyped on different platforms to be combined in a joint or meta-analysis[42,43].

Genetic imputation takes advantage of the fact that humans share long stretches of DNA from distant ancestors[44,45]. While there are over 7.5 billion people on the planet, the effective population size in regard to genetic variation is roughly 5,000, due to the genetic similarity among the human population[45]. These long stretches of DNA that are passed from mother and father to offspring are called haplotypes, and genetic imputation begins with estimation of these haplotypes. Mapping the location of haplotypes in a cohort is important to determine which SNPs are commonly inherited together. Two SNPs that are inherited together are said to be in linkage disequilibrium, or linked.

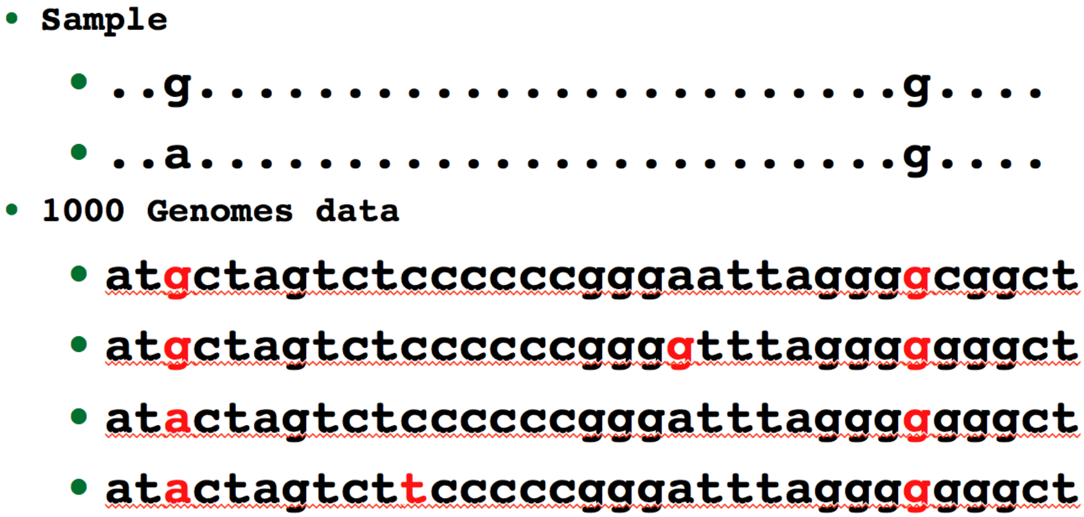
The process of haplotype estimation is the most computationally intense step of the process. It begins by creating a random set of haplotypes that are a mosaic of the reference haplotypes. The haplotypes are then refined in a stepwise hidden Markov process and updated in relation to current state of the haplotypes of the samples around it[46]. Because of the nature of this iterative updating process that relies on the current state of the other haplotypes, this process must be single threaded, as a multithreaded process would create naivety of the state of the adjacent haplotypes if they happen to be updated by a different processing thread. This is one reason that the process is such a long running step.

Once the haplotypes are mapped, the inference of genotypes can begin. This process proceeds base-by-base down the genome. If the base exists in the reference data, but not in the sample, it will be imputed. If it exists in the sample, but not in the reference, it will be skipped. Individual allelic tests are performed to provide an r^2 value that represents imputation accuracy. This calculation encompasses allele frequency, which is an important consideration when determining imputation accuracy. For example, a very rare allele can be 'called' as the major allele every time, and be accurate to a high percentage due solely to the fact that the major allele is present in over 99% of the population. This makes r^2 a much more accurate metric of imputation accuracy than a measure of true positives[46,47].

Imputation outputs include posterior probabilities for the possible alleles. Alleles are 'called' by taking the highest probability allele. Assuming Hardy-Weinberg equilibrium, one can multiply the probabilities of the alleles on each haplotype pair to estimate genotypes. For example, if haplotype 1 consisted of $P(A) = 0.98$ and $P(B) =$

(0.02), and haplotype 2 consisted of P(A) = 0.14 and P(B) = (0.86), the called alleles would be AB. The probability of an AA genotype can be calculated as P(AA) = 0.98 x 0.14 = 0.1372. Allele dosage, which is the number of a certain allele in a genotype, is usually represented as 0, 1 for a heterozygote, and 2 for a homozygote. Imputed allele dosage is calculated by summing the posterior probabilities. For the B allele, allele dosage would be calculated as 0.02 + 0.86 = 0.88.

There are a number of reference samples available for use in imputation analyses, however the current reference set most commonly in use is the 1000 Genomes Project. This project used 2500 samples to assembly 1000 genomes worth of human genetic variation with the aim to represent a master record of the genetic variation with at least 1% minor allele frequency in the human population. There are currently 26 sites across the globe collecting samples that represent a large variety of populations. Because of the limitation on minor allele frequency in reference data, imputation is not accurate for extremely rare alleles in the population[48].



- Sample
  - ..g........................g....
  - ..a........................g....
- 1000 Genomes data
  - atgctagtctccccccgggaattaggggcggct
  - atgctagtctccccccggggtttaggggggggct
  - atactagtctccccccgggatttaggggggggct
  - atactagtcttcccccgggatttaggggggggct

Figure 2. Diagram of the goal of genetic imputation. Imputation aims to estimate the alleles at the missing loci represented by periods by using context provided by the reference sample (1000 Genomes).

After haplotype estimation and imputation, associated statistics are generated for between the trait in question and both observed and imputed SNPs. The MaCH package comes with two executables that perform these calculations for both case-control studies and continuous variables.

There are a number of software packages available for genotype imputation. BEAGLE is a Java program, which is platform independent. It uses a haplotype cluster model in which reference haplotypes are grouped into clusters at each SNP, outputting posterior probabilities of allelic R^2 for the imputed genotypes[49]. IMPUTE is a second package that is available for all major operating systems, and uses a variant of the 'product of approximate conditionals' (PAC) model. MACH, which also uses a variant of the PAC model, is only available on Linux and Mac OSX. Both MACH and IMPUTE output an average maximum of the posterior probabilities[42]. Another program, BIMBAM, is similar to the previously mentioned software in that it imputes unobserved genotypes with a quantitative assessment of uncertainty, however BIMBAM uses a Bayesian regression, as opposed to the common Frequentist *p-value,* to determine association of these genetic variants to the phenotype of interest[50].

Of the available imputation platforms, MaCH was chosen for its compatibility with our operating systems (Linux and OS X), free licensing, its ability to accept linkage input files and HapMap/1000 Genomes references, chromosome-specific processing, strand orientation functionality, and outputs that include posterior probability and allele dosage. MaCH is computationally intensive with large memory requirements, and it also is poorly documented and command line-only. MACH has been shown to be equally as accurate as IMPUTE and more accurate than BEAGLE, however it runs more slowly because it estimates recombination rates from the dataset itself. It also cannot handle multi-allelic markers[42].

MaCHTools Summary

MaCHTools is a front-end/companion for the imputation software MaCH[51]. While robust and powerful for inferring missing genotypes, MaCH would benefit from the addition of functionality, QC, and data-handling capabilities. MaCHTools was created to addresses these needs by directly coupling MaCH with a customizable battery of QC, and data-handling procedures. The resulting toolset enables a high throughput computational pipeline for imputing data across a large number of studies. MaCHTools is an extensible (Java) data management software that facilitates a lengthy imputation procedure and assists users with data segmentation and reassembly that is executed from either command-line or a graphical interface. MaCHTools accepts PLINK[52] formatted inputs and launches and manages parallel computations to several standard cluster job managers such as Slurm. A typical implementation of MaCHTools would proceed first with confirming correct formatting of all inputs files and the existence of one or more reference sets. Then chromosome-specific data are created, filtered, checked for strand ordering, and split into small segments for phasing by MaCH against reference genomes. MaCHTools ligates the resulting haplotype files and prepares them for imputation. MaCH then imputes genotypes and performs an association analysis between the imputed

SNPs and quantitative traits and/or affection status. Finally, MaCHTools concatenates the resulting files and processes them for SQL database importing. MaCHTools also has procedures for subsequent analyses. For example, MaCHTools can facilitate association checks by preparing the necessary files for visually checking genotype call qualities and by identifying the typed SNPs that drive genotype imputation of imputed SNPs that are strongly associated with the dependent variable.

Endophenotypes and Factor VII

Endophenotypes are intermediate traits that are closer to the underlying molecular mechanism than the complex phenotype, and are in principle more likely to be affected by the genetic variation. John Bernard and Kenneth Lewis coined the term in a study that aimed to describe the geographic distribution of grasshoppers. They found that they were unable to explain their geographic distribution based on external "exophenotypes" and posited that it was the grasshoppers' internal "endophenotypes" that determined their distribution. The paper was published in 1966 [53]. The next use of the concept was in psychiatric genetics in a study that aimed to explain the gap between low level genetic variation, such as SNPs, and high-level symptom presentation of complex diseases, such as schizophrenia and bipolar disorder[54]. Discovering genetic and environmental factors contributing to complex human diseases, as well as the development of effective therapies often requires understanding endophenotypes of the disease. For example, discovery of genetic factors contributing to coronary artery disease and the eventual development of effective therapies based on HMG-CoA reductase inhibition was made possible by understanding the endophenotype of hypercholesterolemia[55]. Potential endophenotypes of Alzheimer's disease include quantitative neuroimaging, such as measures of hippocampal atrophy[56-58], or levels of amyloid or tau proteins in the brain or cerebrospinal fluid (CSF)[59-62]. An additional and still evolving source of AD biomarkers is the pool of circulating proteins in the blood[63-66].

Factor VII is a serine protease that is a key member of the coagulation cascade[67]. Along with tissue factor, F7 is responsible for initiating the coagulation cascade. The process begins with release of tissue factor from the external wall of blood vessels following vascular injury. Once inside the circulation, tissue factor binds to F7, which is converted to F7a, leading to conversion of factors IX and X into active proteases; factors IXa and Xa[67]. Factor VII is a vitamin K dependent enzyme and the target of warfarin and other anticoagulants that are used to prevent thrombosis and thromboembolism[68].

Polymorphisms within the F7 gene have not been suggested previously as contributing to AD risk, despite multiple large-scale studies. Nevertheless, a SNP within this region (rs6046) has been associated with variation in risk for cardiovascular disease, venous thrombosis and stroke[69-73]; conditions that are associated with risk for AD and other forms of dementia. The rs6046 polymorphism, which is located in exon 9 and is predicted to cause the substitution of glutamine in place of arginine at amino acid position 353 (R353Q), has been shown to result in reduced levels of F7 activity. The haplotype containing this SNP has been reported as both protective and a risk factor for coagulation related disease phenotypes[71-73].

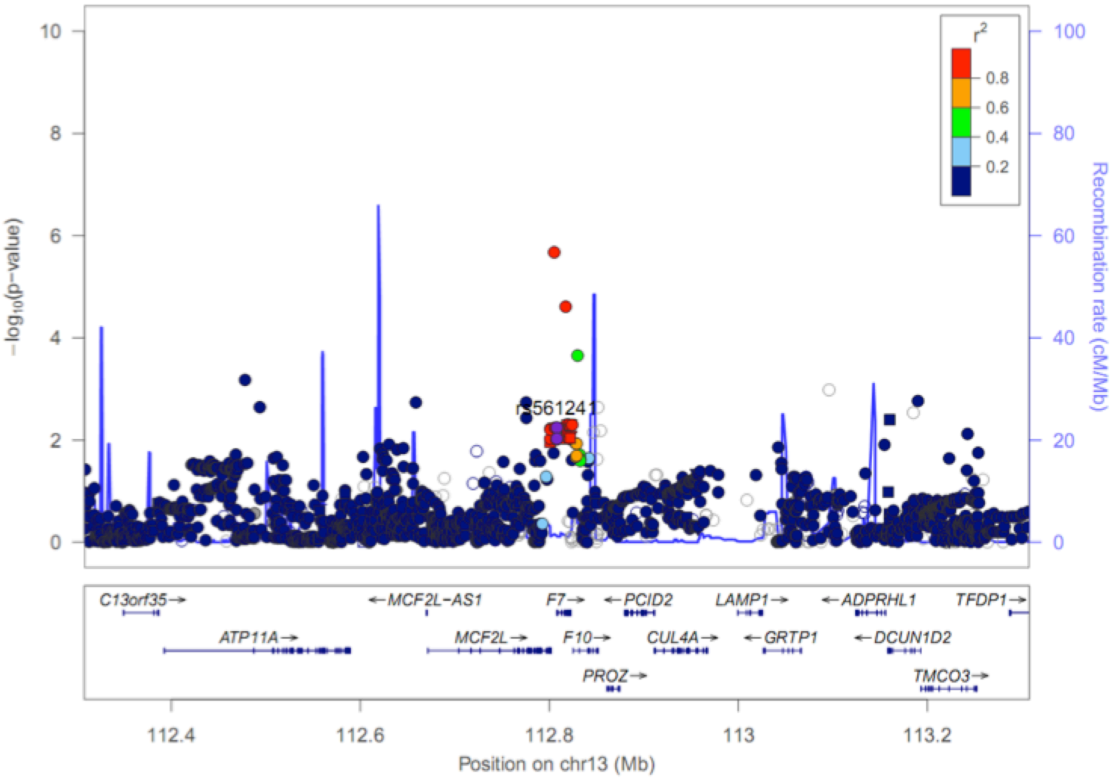


Figure 3. Signal in Factor VII gene in the TARCC cohort. LocusZoom plot of the local signal in the Factor VII gene in association with serum levels of Factor VII. This signal reached genome-wide significance in meta analysis, however the TARCC data to be replicated is presented here.

Royall's δ

A latent variable is a variable that is constructed, or calculated from observed measures. This is in contrast to a variable that is measured. A latent variable is often used because the complexity of the problem demands it. A particularly astute analogy exists in the "Myth of the Cave" from Plato's The Republic. The myth tells of a group of people that are forced to face a wall in a cave. The only things they see are the shadows projected on this wall by things that pass in front of a fire behind them. In science, we try to measure the variables we study as best we can, but we often are only able to measure the 'shadows' and construct a latent variable that represents the otherwise immeasurable variable of interest. We infer these constructs, which are hidden, or latent, from the data we collect[74,75].

δ, or D, is a latent variable that represents cognitive decline based on the condition that a patient must exhibit acquired cognitive impairment, functional disability, and that the disability is related to the cognitive impairment. D is then calculated using cognitive correlates of functional status. D seeks to concentrate Spearman's "g", or G, which is a score for general intelligence, into a score that explains the variance in cognition that related solely to the dementing process[76]. Interestingly, G is highly heritable and therefore may have a genetic component[77].

D has been validated in the Texas Alzheimer's Research and Care Consortium (TARCC), a well-defined AD cohort. In this validation, D is strongly and uniquely associated with dementia severity measured by CDR sum of boxes, and is highly predictive of diagnoses by clinicians[76]. Royall's group then sought to test whether D was associated with cognitive decline in the Freedom House Study (FHS), a longitudinal study of successful aging. The FHS cohort was non-demented at its inception, however it subsequently experienced significant cognitive decline over the course of the study. Royall's group found that D is uniquely associated not only with baseline cognition, but also with longitudinal cognitive change[78].

Project Overview

Genotyping arrays used for genome-wide association studies (GWAS) interrogate only a limited portion of the genome, and therefore do not capture all of the variation that is present. One way to increase the number of variants assayed is to estimate unobserved genotypes. A more in-depth GWAS can then be performed, with a greater chance of detecting novel loci associated with a trait.

Late onset Alzheimer's disease is etiologically heterogeneous, increases in prevalence with age with a lifetime risk of 1 in 10, and results from many genetic and environmental factors[4]. The estimated heritability for AD is between 50-75%[11]. However, despite multiple large independent studies and subsequent meta analyses, over 50% of the genetic variation remains unidentified[10].

According to Don Royall's group at UTSA Health Science center, a latent proxy for dementia severity can be calculated for a patient using cognitive test scores in a structural equation model. The group has named this model 'D', or D. This model distinguishes dementia-relevant variance in cognitive task performance from variance unrelated to general intelligence or the dementing process[76]. Dr. Royall's group has validated this phenotype in multiple cohorts, including TARCC, for which we have (an increasing number of) genotype data.

Another novel approach to studying the genetics of AD is to ascertain quantitative endophenotypes that are associated with AD risk and then look for genetic variants that are associated with those endophenotypes. Endophenotypes are intermediate traits that are closer to the underlying molecular mechanism than the complex phenotype, and are in principle more likely to be affected by the genetic variation. Discovering genetic and environmental factors contributing to complex human diseases, as well as the development of effective therapies often requires understanding endophenotypes of the disease.

Hypothesis

A GWAS analysis pipeline that includes MaCHTools to prepare the data set for imputation provides accurate, repeatable association results while increasing coverage of the genome through imputation with MaCH. Further, there are genetic loci that are significantly associated with 'D', a latent variable for the dementing process.

This hypothesis is tested in collaboration with the Texas Alzheimer's Research and Care Consortium (TARCC), a state-funded collaborative effort between investigators at Baylor College of Medicine, Texas Tech University Health Science Center, University of Texas Southwestern Medical Center, and North Texas Health Science Center. Since 2001, TARCC has enrolled aged subjects classified as NC, MCI and AD. Criteria for categorizing subjects are based upon neurocognitive evaluations, family and/or caregiver interviews and medical history. NC must have normal psychometric test scores and a clinical dementia rating (CDR) score of 0. MCI subjects are classified based upon the Mayo Clinic Alzheimer's Disease Research Criteria. Patients are deemed probable AD according to the NINCDS-ADDA criteria.

Current data analysis methods are becoming the bottleneck in the analysis of large genetic data sets. New software tools for analyzing the data and streamlining these workflows will allow for more effective utilization of the massive data sets that researchers are now generating. Validating MaCHTools on published data and testing a novel phenotype are the first steps in releasing the tool to the public to help ameliorate the analysis bottleneck.

Chapter II

MaCHTools OVERVIEW

MaCHTools began as a series of Java classes designed to be executed individually in a step-wise manner to prepare genotype datasets for imputation with MaCH. MaCH performs only rudimentary error checking based on allele frequencies in the reference haplotypes compared to the allele frequencies in the sample population. The resulting outputs are error prone. Additionally, MaCH jobs are by default launched as one large job, and may require weeks of compute time without parallelization. MaCHTools ameliorates these problems by performing a series of checking steps on the input files, QC measures for the genotypes, splitting haplotype estimation and imputation into smaller, parallel jobs, and resolving ambiguous strand orientation.

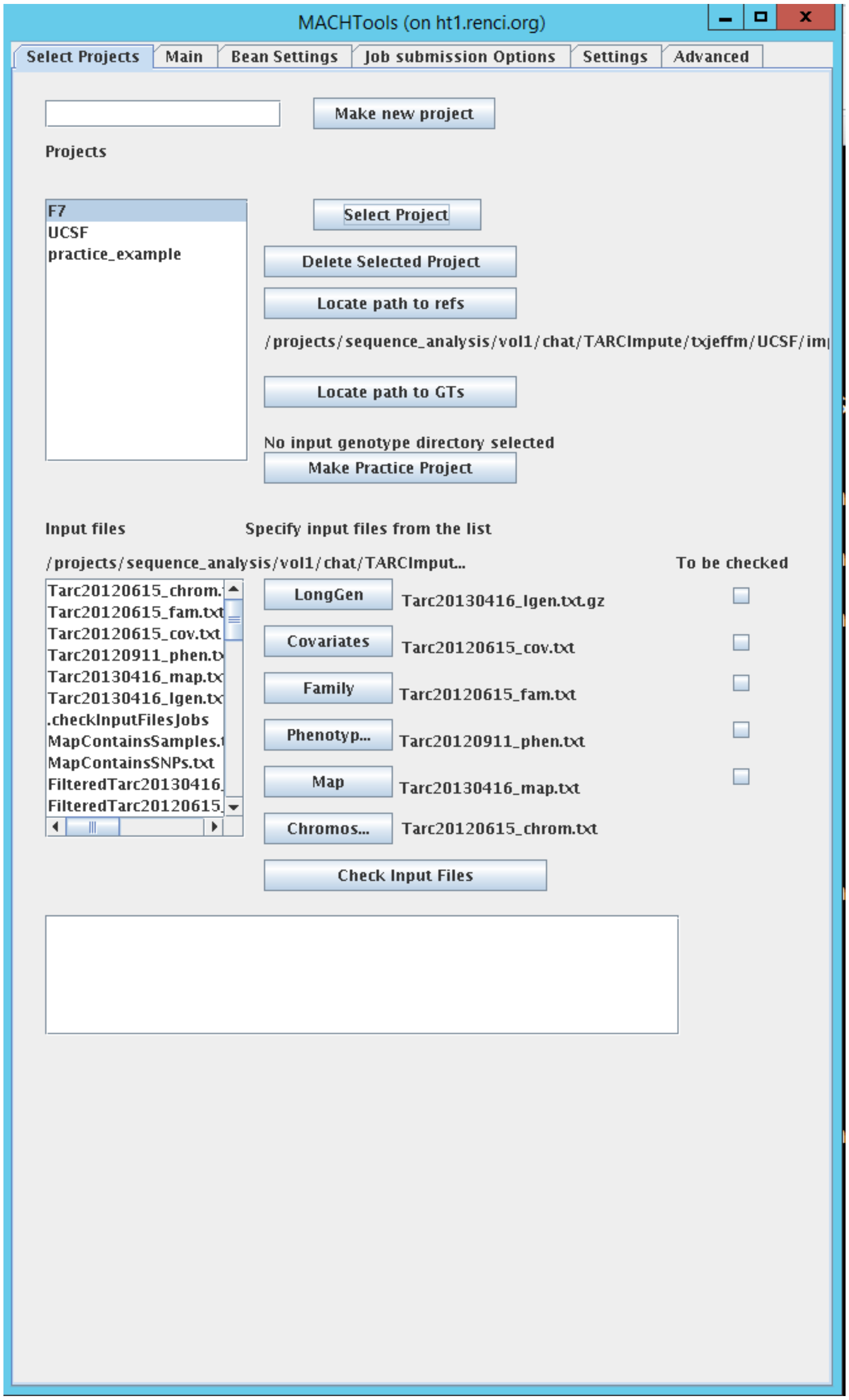
Materials and Methods

MaCHTools was written in the Java using JDK 8.0_25 in Eclipse Java EE IDE 4.5.0 Mars. Dependencies include junit 3.8.1, commons-math 1.2, spring-beans 4.2.4, sprint-context 4.2.4, velocity 1.5, derby 10.4.2.0, commons-lang 2.4, commons-logging 1.1.1, commons-io 1.3.2, spring-core 4.2.4. All dependencies were managed early in the project using Apache Maven 2.2.1, and then using Eclipse's dependency management features in the Mars version. MaCHTools, its dependencies, and its resource files are packaged into a runnable .jar file and executed in a CentOS Linux environment. Jobs for this project were launched to a compute cluster at Renaissance Computing Institute in Chapel Hill, North Carolina. Briefly, the cluster consists of 224 nodes, 3,863 cores, and 27.63 terabytes of RAM. Resource allocation for MaCHTools is 1 node and 1 core per node for all jobs. Pre-haplotype estimation jobs, including haplotype estimation jobs, run well with 32GB RAM per job, and post-haplotype estimation jobs can require up to 96GB per job. This cluster manages user jobs using the Simple Linux Utility for Resource Management, or SLURM.

Results and Discussion

The MaCHTools workflow begins with a project database that stores all saved projects in a given directory. A project-specific database houses the properties, which have defaults but are also editable by the user. These properties include paths to input files and directories, thresholds, and variables. MaCHTools writes a properties file using the key-value pairs in this database for each step for ease of troubleshooting specific steps, in contrast to a master properties file used in previous versions.

MaCHTools utilizes object instances that are managed by the Spring container. These object instances are called ‘beans’, and are essentially recipes for creating instances of Java classes defined by a configuration metadata, in this case an XML file. The first beans that are launched are for checking the completeness and fidelity of the input files. Input files include a long genotype file, family file, phenotype file, covariate file, and map file. This step checks each file for concordance with each other, for example, ensuring that all patients in the family file are present in the long genotype file and that there are genotypes for every patient in the family file.



Projects:
F7
UCSF
practice_example

/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/im

No input genotype directory selected

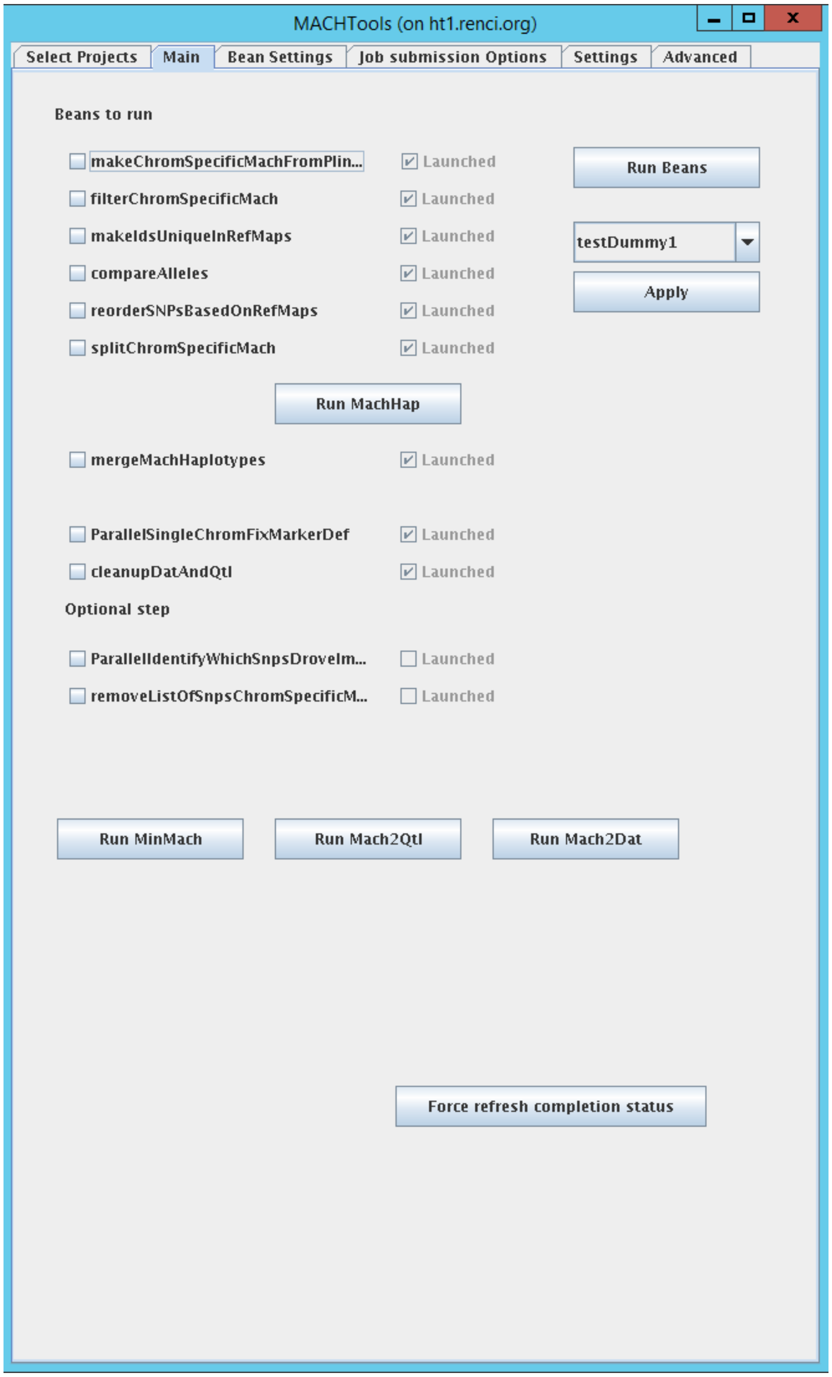
/projects/sequence_analysis/vol1/chat/TARCImput...

Tarc20120615_chrom.
Tarc20120615_fam.txt
Tarc20120615_cov.txt
Tarc20120911_phen.t
Tarc20130416_map.tx
Tarc20130416_lgen.tx
.checkInputFilesJobs
MapContainsSamples.
MapContainsSNPs.txt
FilteredTarc20130416
FilteredTarc20120615

LongGen Tarc20130416_lgen.txt.gz □
Covariates Tarc20120615_cov.txt □
Family Tarc20120615_fam.txt □
Phenotyp... Tarc20120911_phen.txt □
Map Tarc20130416_map.txt □
Chromos... Tarc20120615_chrom.txt

Figure 4. MaCHTools GUI Project Selection Tab. Here the user creates and select projects, specifies input files and reference directories, and runs the file checking procedures

Following successful completion of the file checking steps, the next bean creates chromosome-specific data files by reading the long genotype file line-by-line and writing the genotypes to a new file for each chromosome. This allows the longer running steps to be run in parallel by submitting jobs that run on individual chromosomes instead of the entire data set at once.



Beans to run

□ makeChromSpecificMachFromPlin... ☑ Launched

□ filterChromSpecificMach ☑ Launched

□ makeIdsUniqueInRefMaps ☑ Launched

□ compareAlleles ☑ Launched

□ reorderSNPsBasedOnRefMaps ☑ Launched

□ splitChromSpecificMach ☑ Launched

testDummy1

□ mergeMachHaplotypes ☑ Launched

□ ParallelSingleChromFixMarkerDef ☑ Launched

□ cleanupDatAndQtl ☑ Launched

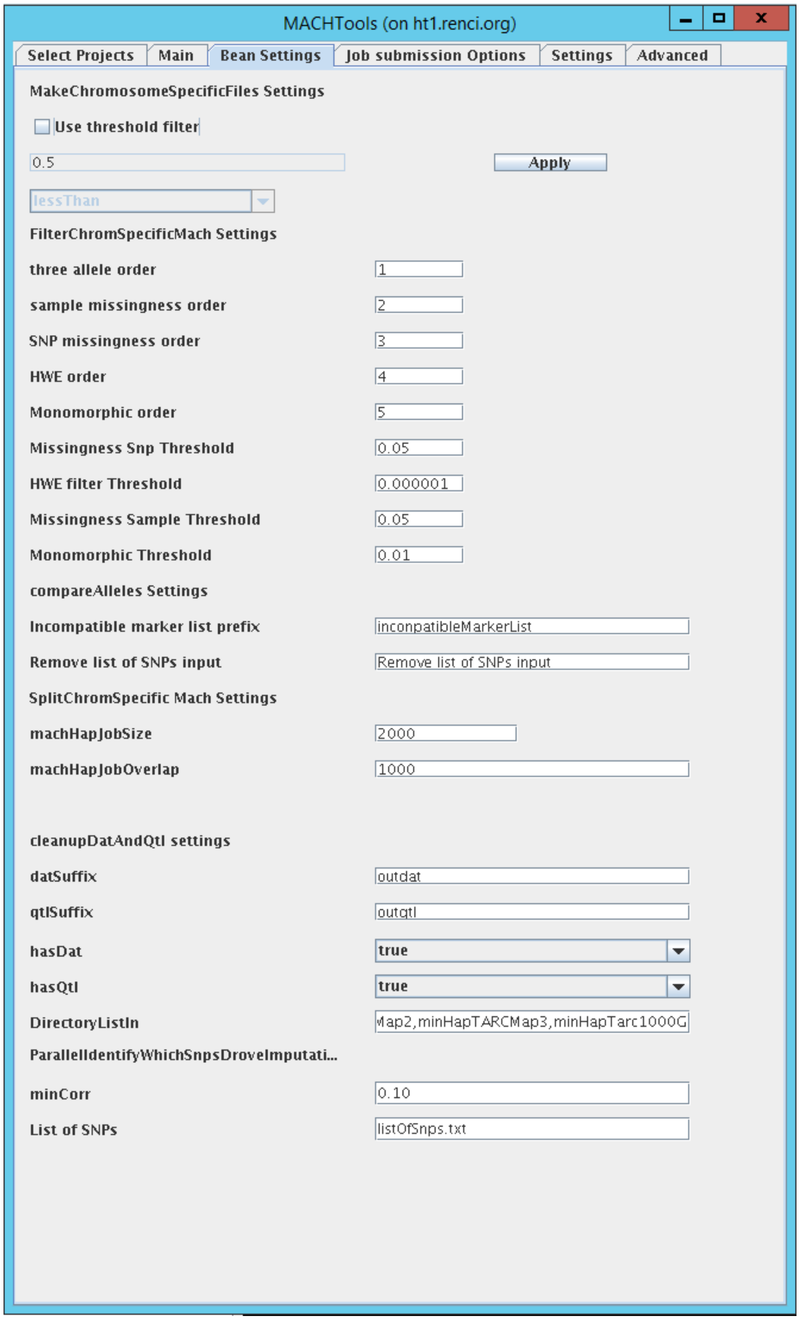
Optional step

□ ParallelIdentifyWhichSnpsDroveIm... □ Launched

□ removeListOfSnpsChromSpecificM... □ Launched

Figure 5. MaCHTools GUI Main tab. Here the user launches the individual MaCHTools beans, haplotype estimation, imputation, and association jobs.

Once the chromosome-specific data files have been written, the next set of beans filter on specified thresholds such as missingness and tests Hardy-Weinberg equilibrium. At this point, the references are read to ensure that each marker is at a unique place on one chromosome before comparing the sample alleles with reference alleles to check for compatibility. If there are any SNPs that have alleles that are incompatible with the reference alleles, they are deleted before haplotypes are estimated.



MakeChromosomeSpecificFiles Settings

□ Use threshold filter

0.5

lessThan

FilterChromSpecificMach Settings

three allele order 1

sample missingness order 2

SNP missingness order 3

HWE order 4

Monomorphic order 5

Missingness Snp Threshold 0.05

HWE filter Threshold 0.000001

Missingness Sample Threshold 0.05

Monomorphic Threshold 0.01

compareAlleles Settings

Incompatible marker list prefix inconpatibleMarkerList

Remove list of SNPs input Remove list of SNPs input

SplitChromSpecific Mach Settings

machHapJobSize 2000

machHapJobOverlap 1000

cleanupDatAndQtl settings

datSuffix outdat

qtlSuffix outqtl

hasDat true

hasQtl true

DirectoryListIn Map2,minHapTARCMap3,minHapTarc1000G

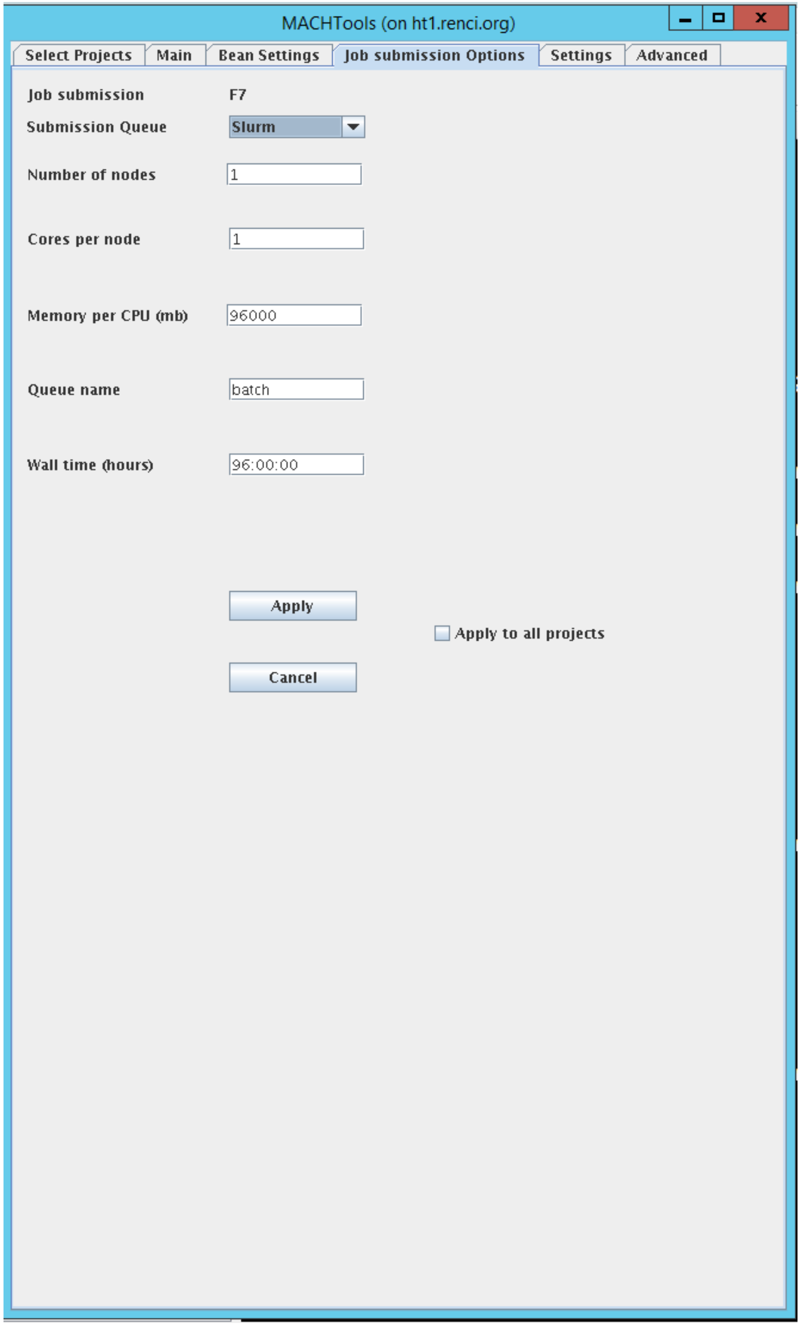
ParallelIdentifyWhichSnpsDroveImputati...

minCorr 0.10

List of SNPs listOfSnps.txt

Figure 6. MaCHTools GUI Bean Settings tab. This tab includes user-defined settings, thresholds, priorities, and file name conventions

MaCHTools then reorders sample SNPs based on the order of the SNPs in the reference and splits each chromosome-specific data file into smaller, more manageable files in preparation for haplotype estimation. This results in approximately 1000 files, which are submitted in parallel for haplotype estimation by MaCH. MaCH generates a random set of haplotypes for the study population as a mosaic of the reference haplotypes and iteratively refines them over a set number of cycles in a hidden Markov chain [46].



Job submission F7

Submission Queue Slurm

Number of nodes 1

Cores per node 1

Memory per CPU (mb) 96000

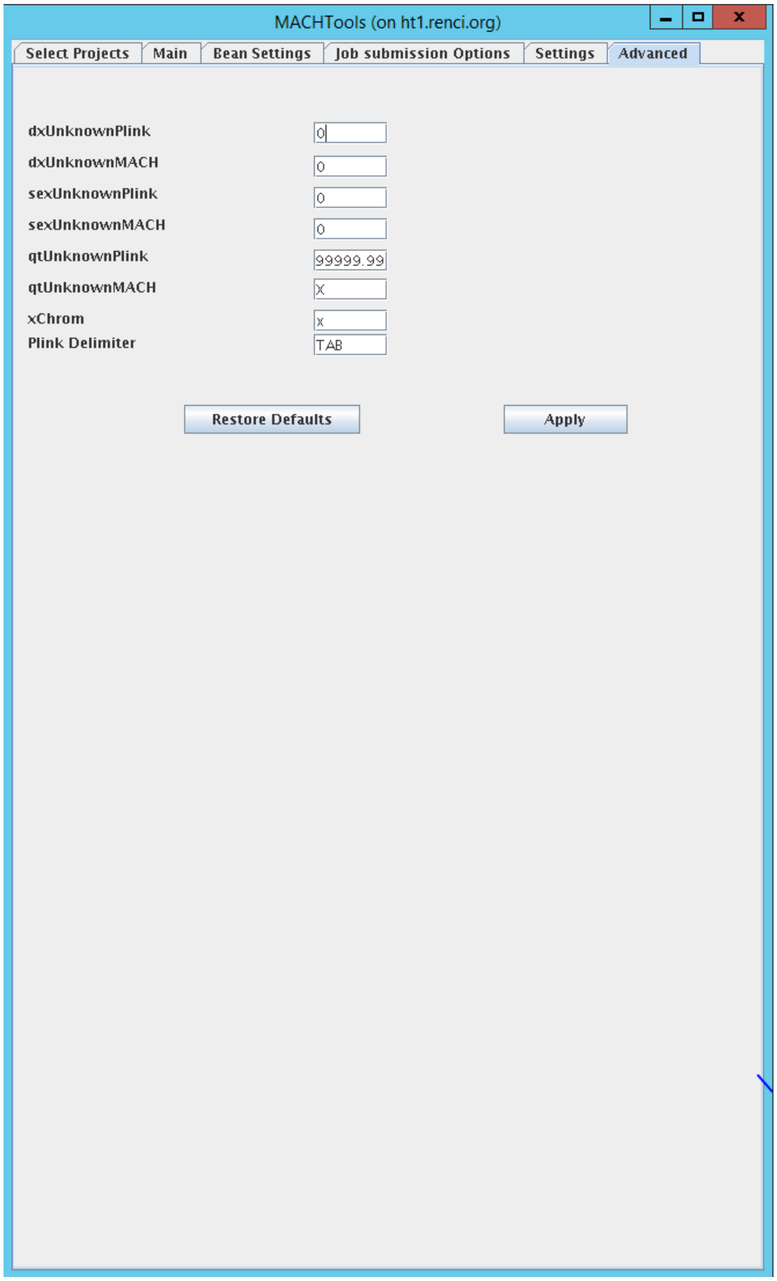
Queue name batch

Wall time (hours) 96:00:00

□ Apply to all projects

Figure 7. MaCHTools GUI Job submission options. This tab is used to specify job submission parameters, specifically in regard to resource allocation and run time limits.

After the completion of haplotype estimation, MaCHTools ligates the haplotypes back together into individual chromosome-specific data files[46]. The final bean before imputation launches jobs in parallel for each chromosome to resolve any strand orientation issues where sample SNPs are coded ambiguously in relation to their reference haplotype. This step is useful for sample SNPs that are not on the reference strand and need to be flipped.



dxUnknownPlink	0
dxUnknownMACH	0
sexUnknownPlink	0
sexUnknownMACH	0
qtUnknownPlink	99999.99
qtUnknownMACH	X
xChrom	x
Plink Delimiter	TAB

Figure 8. MaCHTools GUI Advanced settings tab. These settings are related to how missing values are coded. They are editable by the user but will most often remain unchanged.

This concludes the pre-imputation steps with MaCHTools and at this point, the data are ready for imputation and quantitative or binary trait associations using mach2qtl and/or mach2dat. These jobs can be launched in parallel from the MaCHTools GUI, and their results will be stored in the project directory. A job will be launched for each chromosome, and once complete, MaCHTools' final bean combines the output for each chromosome into one large file and compresses it in gzip format.

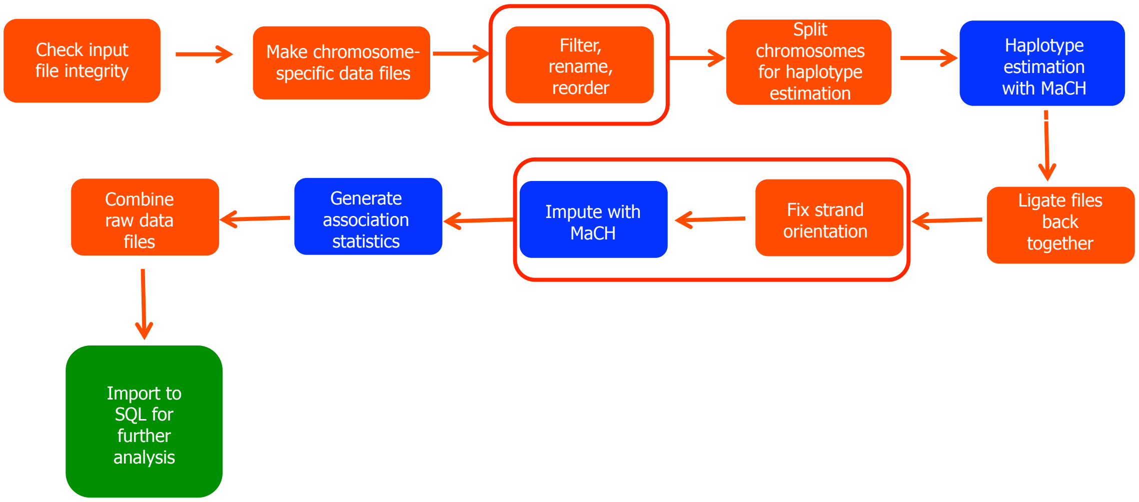

Check input file integrity
Make chromosome-specific data files
Filter, rename, reorder
Split chromosomes for haplotype estimation
Haplotype estimation with MaCH
Ligate files back together
Fix strand orientation
Impute with MaCH
Generate association statistics
Combine raw data files
Import to SQL for further analysis

Figure 9. Summary of MaCHTools workflow. Steps in red are performed by MaCHTools. Steps in blue are performed by MaCH. Steps requiring the user are performed in green. Red outlined steps are more computationally demanding and may require additional resources depending on cohort size.

Chapter II- Noteworthy Results

(1) MaCHTools checks input files, filters, reorders, and compares alleles for more accurate haplotype estimation and imputation

(2) While MaCHTools necessitates several additional steps in the GWAS workflow, the workflow is streamlined, more accurate, and parallelized

Chapter III

VALIDATION OF PUBLISHED ENDOPHENOTYPE GWAS DATA

After a large overhaul of the MaCHTools software, which included changing how properties are handled, implementation of two databases, added compatibility with new job managers, launching jobs from a GUI, among other large changes, it was necessary to validate previously published findings from a prior MaCHTools workflow. MaCHTools was a critical tool in a GWAS of endophenotypes in which a panel of serum proteins was measured in TARCC participants, and the levels of these proteins were used to create a highly accurate predictive model for diagnosing Alzheimer's disease [65] [64]. This study aimed to resolve genetic loci that are significantly associated with the altered levels of these serum proteins, with the ultimate of aim of finding susceptibility loci for Alzheimer's disease [79]. A portion of the results of the analysis from this study is replicated here to validate MaCHTools' accuracy.

Materials and Methods

The data files used for this validation were copied to a new project directory and used as inputs for the MaCHTools workflow. These files were dated from April 16, 2013. The methodologies for generating this data are described elsewhere [80], but briefly, participants were categorized as probable AD, mild cognitive impairment (MCI), or normal control (NC) based on neurocognitive evaluations, family/caregiver interviews, and medical history. NC participants were required to have a clinical dementia rating (CDR) of 0. Patients were classified as either MCI based on the Mayo Clinic Alzheimer's Disease Research Criteria [81], or probable AD according to the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's

Disease and Related Disorders Association (ADRDA) [82]. Protein concentrations were measured from serum either from baseline or from a year-one follow-up exam. After excluding proteins that did not contribute to the O'Bryant et al. screening algorithm in the same direction in two different cohorts (Alzheimers Disease Neuroimaging Initiative), a list of seven proteins included Adiponectin, Beta 2 Microglobulin, Factor VII, Monocyte Chemotactic Protein 1, Pancreatic Polypeptide, Tenascin C, and Vascular Cell Adhesion Molecule 1.

The TARCC cohort was genotyped using the Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA) which includes 906,600 SNP markers. This panel obtain genome-wide coverage. The BirdSeed v2 algorithm was used to call genotypes[83].

Results and Discussion

The current version of MaCHTools was used to process, QC, and prepare these files for imputation with MaCH. Along the way, the outputs for each step were compared to outputs generated in 2013 for the published endophenotype study. All outputs were identical. After imputation, the raw outputs were imported to SQL, and a similar comparison was done to compare the Manhattan tables for the Factor VII phenotype between the 2013 analysis and the analyses performed by the current version of MaCHTools. The most significant p-values were identical. Any discrepancy between p-values for the analyses were seen in very large p-values or far outside the number of significant figures commonly reported in manuscripts, and may be due to the hidden Markov chains involved in imputation. Random haplotype generation followed by iterative refinement is a reliable and accurate process, however exact repeatability may not be possible. Regardless, the final conclusions drawn from each analysis relating to significant loci would be identical.

Chapter III- Noteworthy Results

(1) The current version of MaCHTools replicated published data generated in 2013 using the exact same input files and genotypes

(2) Despite large changes in code, MaCHTools is still able to perform reliable and accurate analyses

Chapter IV

GWAS OF 'D', A LATENT VARIABLE REPRESENTING THE DEMENTING PROCESS

In statistics, variables can be both observed and measured, or calculated. The latter is termed a 'latent variable'. These inferred variables are derived from mathematical equations and are used in many disciplines including psychology, machine learning, and economics. One advantage to a latent variable is that it reduces dimensionality. Variance in an outcome may be described by a multitude of other measured variables in a highly dimensional data set. By inferring a latent variable via a mathematical model that encompasses these measured variables, the variance in an outcome can largely be explained by a single latent variable.

D is a promising latent variable that distinguishes dementia-relevant variance in cognitive task performance from variance unrelated to the dementing process (G'). Effectively, D and G' comprise Spearman's G, or "general intelligence" [78]. G has been shown to be highly heritable[77]. This, coupled with the observations that AD has an estimated heritability between 50-75%[11] leads to the hypothesis that there is a genetic basis for D. D has been validated in the TARCC cohort[76], which has been genotyped and well characterized in regards to AD. TARCC systematically excludes participants that show evidence of non-AD and/or mixed dementias, therefore D scores in this cohort are likely to reflect AD-specific dementia.

D was calculated in TARCC and associated with cytokines and serum biomarkers. Initially, this D homolog did not exhibit consistent factor loadings across ethnicity. A follow-up study calculated a homolog of D that exhibited consistent mean and factor loadings across ethnicities, but associations with biomarkers were strongest in non-Hispanic white participants. The authors concluded that the dementing process is distinct in these two groups, which is been evidenced in other published studies[84-86].

Further, D was calculated in a separate cohort, the University of Kansas Brain Aging Project. In this cohort, D was able to classify patients as either mild cognitive impaired (MCI) or Alzheimer's disease (AD) and rank order their dementia severity. Additionally, the study utilized magnetic resonance imaging (MRI) to associate D score with brain matter loss and found that higher D score, i.e. more advanced dementia, was localized to the default mode network. The default mode network is a network of brain regions that are most active when the brain is at a wakeful rest, such as daydreaming or mind wandering and without focus[87,88].

Lastly, Royall's group investigated the association between D score and Vitamin D binding protein, VDBP. VDBP has been found to be elevated in the cerebrospinal fluid of patients with neurodegenerative disorders[89]. Royall's group found that D is significantly positively associated with levels of VDBP, and they conclude that D mediates the adverse effects of VDBP on cognition. Drawing on the previous study that D is related to the default mode network, they postulate that VDBP effects on AD are mediated through the default mode network, as VDBP is an amyloid-beta scavenger[90,91]. Amyloid-beta deposition can be seen in the default mode network through neuroimaging[92].

Materials and Methods

TARCC participants have been genotyped using the Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA). This analysis includes 690 of these participants.

D scores are calculated for each participant in waves, each wave being approximately one year apart for a total of five waves. Not every participant has a score for each wave due to the variability of when the participant joined or left the cohort. For each participant, Dr. Don Royall's group using structural equation modeling to derive D. The model includes several parameters including Instrumental Activities of Daily Living 1-5 (IADL), Basic Activities of Daily Living 1-5 (BADL, Boston Naming Test, Controlled Oral Word Association Test, Digit Span Test, Weschler Memory Scale, among others.

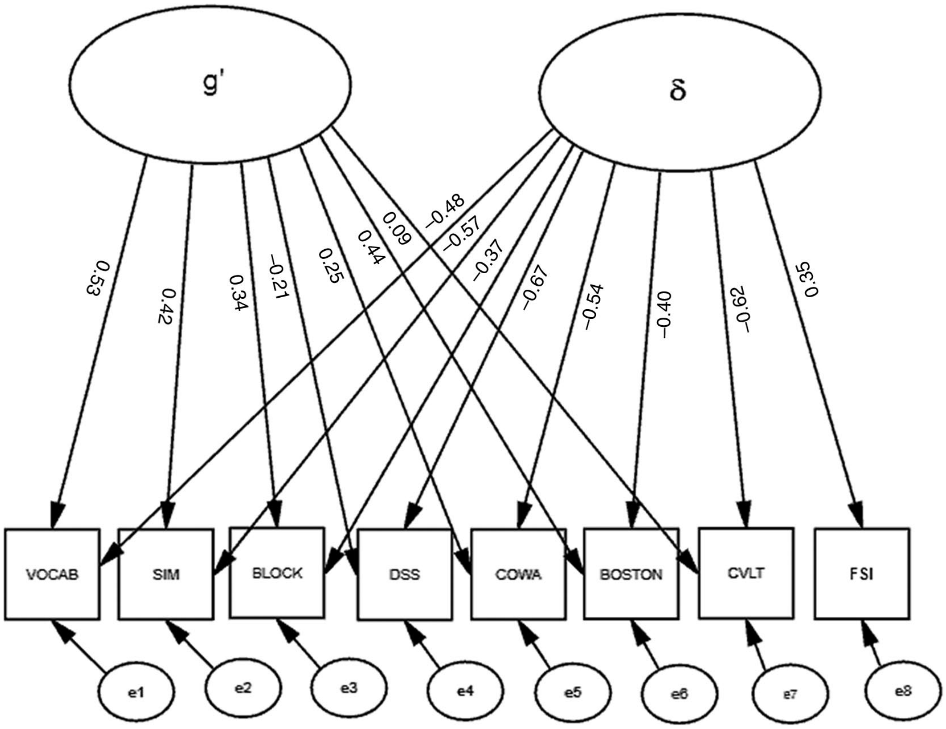


Figure 10. Structural equation model for calculating D. Measured variables are in boxes with small ovals indicated the error in their measurement. G' and D factor loadings represent the correlation between the observed score and the latent variable.

A GWAS pipeline that includes MaCHTools for input file processing and QC and MaCH for haplotype estimation, imputation, and generation of association statistics was performed. A new project was created in the MaCHTools GUI, and the phenotype file was edited to reflect patient D scores provided by Royall's group. The rest of the input files remained unchanged from the 2013 analysis of endophenotypes, which included covariates and adjustments for age, sex, and education, along with principle components to control for population substructure. The genotype missingness threshold was set to exclude any genotypes that were not called in greater than 95% of participants. The test for Hardy-Weinberg equilibrium was set at 0.000001. Monomorphic SNPs were excluded at a 1% threshold. Generation of input files and initial associations of typed SNPs were done with PLINK and Windows SQL Server 2012.

All files were checked, filtered, reordered, and split for haplotype estimation by MaCHTools. After haplotype estimation by MaCH, MaCHTools was used to ligate the haplotypes back together for imputation. MaCHTools then corrected ambiguous SNPs that needed to be flipped to the reference strand, followed by imputation with MaCH. Association statistics were performed by mach2dat and mach2qt.

Results and Discussion

One of the first steps in generating GWAS results is to generate a quantile-quantile, or QQ plot. This plot maps each data point, in this case SNPs on an axis representing the observed p-values vs. the expected p-values due to random chance. The expectation, if the assumptions that the phenotype is normally distributed and the SNPs are in Hardy-Weinberg equilibrium, is that the data points should fall on a somewhat straight line, or X=Y. Any deviation from X=Y would suggest that confounders are at work or the assumptions are not met.

Despite several individual SNPs in each wave reaching genome-wide significance, a full association signal could not be resolved. There are weakly associated

SNPs in linkage disequilibrium with the most significant SNPs, and these may be better resolved in subsequent studies with larger sample sizes. The results are promising but not definitive. The most promising signal was at rs12056944 on chromosome 8. This particular SNP is mapped in genetic wasteland and is not in or near any genes. Rs12056944 does however map to an eQTL, or expression quantitative trait locus, that influences expression of the RP11-62H7.2 gene in thyroid tissue, which encodes ribosomal protein L10[21]. If this signal were validated in a larger cohort, this would be an interesting finding due to the comparison between Alzheimer's disease and metabolic syndromes such as diabetes, with AD often being referred to as type 3 diabetes.[93-95] This signal is weak in wave one, stronger in wave two, and reaches genome-wide significance with a strong shoulder of SNPs in linkage disequilibrium in wave 3. However, this signal drops out in waves four and five.

The likely explanation for this is that waves 4 and 5 have a much lower sample size than waves 1-3, with wave 5 dropping as low as 50 individuals. Another possible reason for the weak associations is that D is a latent variable derived from cognitive measures. While the literature describes D as an endophenotype, endophenotypes can be exhibited at any point along the path from genotype to phenotype. Those closer and endophenotype is to a phenotype, the greater the power required to associate with genotypes. The TARCC cohort provided sufficient sample size to generate strong association signals in the GWAS of serum protein endophenotypes possibly because serum protein levels are endophenotypes that are much closer to genotypes and gene expression than the behavioral and cognitive phenotypes used to derive a D score.

Chapter IV – Noteworthy Results

(1) Using the MaCHTools/MaCH workflow a GWAS was performed on 5 waves of D scores simultaneously

(2) Due to sample size constraints, a significant signal could not be resolved in this analysis. However, there are promising signals that may rise to genome-wide significance in future analyses with larger samples

Chapter V

DISCUSSION

Utility of MaCHTools in complex disease genetics

The current bottleneck in genetics studies has moved from the generation of data to the analysis of data. Data sets are tens to hundreds of gigabytes in size and require entire compute clusters to store and process. This creates a need for new analysis methods and streamlining of our current methods. The ability to process data sets, raw input files, and correct errors and ambiguities in strand orientation in a point-and-click manner allows for parallelization of not only job submissions, but of the users research methods as well. Reducing hands-on time in the GWAS workflow allows the user to focus efforts between steps on other projects, and to quickly back up and re-run analysis that failed or were run incorrectly. Waiting on a large imputation job to run for weeks only to find out the analyses were set up erroneously is a large time sink. Being able to check the status of the smaller chromosomes as they finish in a matter of hours allows the user to cancel the larger chromosome jobs, fix the error and re-launch the jobs with minimal time lost. MaCHTools is presented here in an AD context, but imputation is a common technique used broadly in the field of genetics. Any group that studies complex disease genetics can benefit from utilizing MaCHTools in their GWAS pipelines.

Interrogating the genetics of D using MaCHTools

Despite low sample sizes, particularly in the later waves, there are some promising signals that may be better resolved in future studies that include more participants. The most promising signal was on chromosome 8 with a terminal SNP rs12056944 that was replicated in the first three waves. It's interesting that despite a decrease in sample size with each passing wave, this signal grew stronger. The average lifespan for Alzheimer's disease is five years[96]. At wave three, dementia will have progressed markedly, and therefore may explain the increase in this signal between waves 1 and 3. An interesting follow-up study would be to use change in D over time, or ΔD as the phenotype for a GWAS to understand the genetics of how a patient progresses through the dementing process.

While the terminal SNP in the signal on chromosome 8 does not land in or near any genes, one approach for ascertaining the biological significance of variants is to look at eQTLs, or expression quantitative trait loci. These loci effect how genes are expressed, and they may or may not be in genes themselves. An effective tool is the GTEx project, which aims to provide information about the relationship between tissue-specific gene expression and genetic variants[97]. This tool is invaluable in GWAS because many of the loci that are found to be significant lie outside of protein coding regions, which makes it difficult to ascertain the biological mechanisms of disease relating to the significant variants in the analyses.

The GTEx tool located rs12056944 to an eQTL that effects expression of the RP11-62H7.2 gene. This gene expresses a ribosomal protein, L10 in thyroid tissue. There is a non-zero chance that this signal may be spurious, however a signal with a biological context in thyroid tissue would not be surprising, as Alzheimer's disease has been shown to present through, among others, a metabolic process[98].

Future directions:

TARCC researchers have recently completed an additional round of genotyping on an Illumina MEGA chip for ~2800 participants. Royall's group is continuously curating D scores in the growing TARCC cohort. The analysis of the genetics of D will be repeated using this larger data set in hopes of resolving the weak associations seen in the previous analysis.

Additionally, work is underway on MaCHTools to incorporate the merging of data sets. The ability to merge data sets has eluded researchers due to the difficulty in harmonizing phenotypes, thresholds, and genotype calls. This is particularly difficult when genotype data sets are generated on different platforms, e.g. Illumuna vs. Affymetrix. One of the final steps in the MaCHTools workflow is to flip SNPs that are ambiguous as to which strand they reside on. The code for this step has been revised in a testing version of MaCHTools to include entropy minimization features to harmonize SNP definitions between two data sets. Data set merging will provide large, homogeneous data sets for which to associate novel loci with disease phenotypes and to generate accurate genomic prediction models.

Appendix A- GWAS Analysis Guide

The following manual was written by Dr. Nicole Phillips and Dr. Robert Barber to detail the procedures required to execute the analysis pipeline created by Dr. Kirk Wilhelmsen at UNC Chapel Hill. It has been revised and updated to reflect the latest version of MaCHTools presented here.

Steps:

0. Perform Eigenstrat to determine population structure covariates.

- a. Kirk used smartpca in the Eigensoft suite, the files are very similar to the ones needed for plink.

I. Run plink on genotyped markers (file name: FinalGt)

- a. Make plink files (see Appendix A for example code)
    - i. Generate queries to produce lgen, fam, cov, phen, map tables (see code below)
    - ii. Export tables to file space ** add steps for export**
- b. Move from terminal server to BR with FileZilla
    - i. Open FileZilla
    - ii. Connect (host →: "br0.renci.org"; *User* → <enter UN for terminal server access>; password→ <same PW as for terminal server access>; port → "22")
    - iii. Move files to BR
        1. Navigate to the location of the files in the left pane (terminal server directories)
        2. Navigate to the desired destination in the right pane (the BR directories)
        3. Drag files from terminal server side to the BR side
- c. Open a bash shell create needed directories with needed files
    - i. Open MRemote; connect to BR (right-click BR0 and select connect) Note- if BR connection has not been established, right-click Connections→Add Connection; complete the Config window as shown in Appendix B)
    - ii. In new bash window, *Change Directory* to the TarcImpute directory and *Make a Directory* called impute<date>
        1. cd /projects/sequence_analysis/vol1/chat/TARCImpute
        2. mkdir impute<date>
    - iii. *Change Directory* to the new directory named impute<date>

        1. cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<date>

    iv. *Make a Directory* called plink

        1. mkdir plink

    v. *Change Directory* to the new plink directory and *Copy* needed plink COMMAND files:

        1. cd plink
        2. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/plink/plink
        3. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/plink/plink_maf
        4. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/plink/plink_logistic

    vi. *Go up* a directory to the impute<date> directory and *Copy* over the plink DATA files:

        1. cd .. (or cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<<date>>)
        2. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute<<date>>/ Tarc<<date>>_lgen.txt .
        3. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute<<date>>/ Tarc<<date>>_map.txt .
        4. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/ Tarc20120615_fam.txt .
        5. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/ Tarc20120911_phen.txt .
        6. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/ Tarc20120615_cov.txt .
        7. cp /projects/sequence_analysis/vol1/chat/TARCImpute/impute20121015/ Tarc20120615_chrom.txt .

Copy
newes
of the
map f

These
versio
fam,
and c

d. *Edit* plink COMMAND files using emacs to include new paths and ensure file names and paths are correct.

    i. Make sure you are in the impute<date> directory

        1. cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<date>
        2. Note- a shortcut can be made to TARCImpute in each user's home directory

            a. Ex: home/Niphilli/TARCImpute hyperlinks to projects/sequence_analysis/vol1/chat/TARCImpute

    ii. In shell, enter "emacs ."

1. Using arrow keys, key down to the directory named plink; hit Enter.
2. Using arrow keys, key down to the file named **plink**; hit Enter.
3. Using the arrow keys, modify the path of each file in the command line to reflect the newly created directory, impute<date>, the correct file names (see step 1.c.vi above), and enter the desired output name(after --out <*output_name*>).
4. Exit emacs
    a. Hit Ctrl+x, Ctrl+s to save
    b. Hit Ctrl+x, Ctrl+c to exit emacs

iii. In shell, enter “emacs .”

1. Using arrow keys, key down to the directory named plink; hit Enter.
2. Using arrow keys, key down to the file named **plink_maf**; hit Enter.
3. Using the arrow keys, modify the path of each file in the command line to reflect the newly created directory, impute<date>, the correct file names (see step 1.c.vi above), and enter the desired output name(after --out <*output_name*>).
4. Exit emacs
    a. Hit Ctrl+x, Ctrl+s to save
    b. Hit Ctrl+x, Ctrl+c to exit emacs

iv. In shell, enter “emacs .”

1. Using arrow keys, key down to the directory named plink; hit Enter.
2. Using arrow keys, key down to the file named **plink_logistic**; hit Enter.
3. Using the arrow keys, modify the path of each file in the command line to reflect the newly created directory, impute<date>, the correct file names (see step 1.c.vi above), and enter the desired output name (after --out <*output_name*>).
4. Exit emacs
    a. Hit Ctrl+x, Ctrl+s to save
    b. Hit Ctrl+x, Ctrl+c to exit emacs

e. Run plink command files; this will take 4-5 hours

i. In bash shell, make sure you are in the plink directory

1. cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<date>/plink

ii. Execute the plink files by entering the following commands

1. qsub ./plink

a. generates output: <*output_name*>_dat.assoc.linear

b. generates output for each trait: <*output_name*>.<qtl_pheno_name>_qtl.assoc.linear

2. qsub ./plink_maf

a. generates output: <*output_name*>_qtl.freq

3. qsub ./plink_logistic

a. generates output: <*output_name*>_dat.assoc.logistic

iii. Check on progress using the command qstat

II. Process plink output files

a. Concatenate and process files with sed for importation into db

i. Make sure you are in the plink directory

1. cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<date>/plink

ii. The following commands are performed recursively to build a long command:

1. Concatenate all qtl outputs: grep [0-9] <*output_name*>.*.assoc.linear

a. example: grep [0-9] Tarc20121015_out_qtl.*.assoc.linear

b. The wild card (*) represents any of the quantitative traits include as pheontypes: Adiponectin, Alpha_2_Microglobulin, AOO, etc….

2. Remove output name and "." From filename: ↑|sed 's/<*output_name*>.//g'

3. Remove "assoc.linear:" from filename: ↑|sed 's/.assoc.linear://g'

4. Replace one-or-more spaces with one tab: ↑|sed 's/ \+/\t/g'

a. Note- I found this command online because I could not get it to work with the notes I took on the call; see Call Log notebook, pg 155. The command in my notes said sed 's/ /\t\t/\t/g', performed multiple times.

5. Replace NA with -99999: ↑|sed 's/NA/-99999/g'

6. Drops header line(?): ↑|grep -v CHR

7. Make output: ↑ **>** < *name_for_consolidated_file*>_Plink_QTL.txt

Note: To view the data at any point, add |less to the end of a command; enter "q" to quit out of the less preview.

b. Modify the .frq file

i. Make sure you are in the plink directory

1. cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<date>/plink

ii. The following are performed recursively to build one long command:

1. Remove leading tab: cat *.frq|sed ‘s/^[\t]*//g’
2. Replace one-or-more spaces with one tab: ↑|sed ‘s/ \+/\t/g’
    a. Note- I found this command online because I could not get it to work with the notes I took on the call; see Call Log notebook, pg 155. The command in my notes said sed ‘s/ /\t\t/\t/g’, performed multiple times. Tested and works!
3. Make output: ↑ **>** <*name_for_mod_file*>_Plink_Freq.txt

c. Modify the .dat file

i. Make sure you are in the plink directory

1. cd /projects/sequence_analysis/vol1/chat/TARCImpute/impute<date>/plink

ii. The following are performed recursively to build one long command:

1. Remove leading tab: cat *.dat.assoc.linear|sed ‘s/^[\t]*//g’
2. Replace one-or-more spaces with one tab: ↑|sed ‘s/ \+/\t/g’
    a. Note- I found this command online because I could not get it to work with the notes I took on the call; see Call Log notebook, pg 155. The command in my notes said sed ‘s/ /\t\t/\t/g’, performed multiple times. Tested and works!
3. Make output: ↑ **>** <*name_for_mod_file*>_Plink_DAT.txt

d. Move files back to terminal server file space with FileZilla

i. See Section I.b above for directions on connecting and using FileZilla

e. Import into sql server database

i. Import raw data

ii. Truncate strings, correct field lengths & names & convert numeric values to int or float

Needs to be detailed; see notes pg 177 from 041613.

III. Check clusters for SNPs of interest and a random set of markers across genome (~2000)

a. Make tables for SNP Checker (CheckerList, CheckerOut, CheckerReviewed, CheckerGT)

b. Check Clusters with Eclipse on terminal server

i. See Appendix C for detailed instructions

c. Update EDITED genotypes in appropriate genotype tables in SQL database

Needs to be detailed; see notes from pg 182 from 041613.

IV. Run Plink on EDITED Genotypes and Process plink output files - 2nd cycle

a. Regenerate lgen and map files for plink & imputation (see Appendix A for example code)
   i. Generate queries to produce new lgen and map tables named with the creation date. (do we always have to create a new map table?)
   ii. Export tables to Blue Ridge file space

b. Move from terminal server to BR with FileZilla
   i. See Section I.b above

c. Run plink for qtls, qualitative traits and affection status, determine MAFs
   i. See step I.c.vi and Section I.d above

d. Process plink outputs
   i. See Section II above

V. Run Imputation (can be done concurrently with second round of Plink tests)

i. Create new MaCHTools (MT) projects directory
   1. `mkdir MaCHTools_projects`

ii. Navigate to projects directory and run MaCHTools
   1. cd <<machtools projects directory made in i.>>
   2. Execute `MACHTool MakeProjectDB` from within the MT projects directory made in step .i

iii. Create new MT project
   1. Type name of project in the 'new project' blank
   2. Click 'Make new project' button
   3. Your project will appear in the list of projects and a directory will be made in the MT projects directory made in step .i

iv. Specify reference directory
   1. Click 'Locate path to refs' button
   2. Navigate to *parent* directory of references.
      a. if 1000G directory resides in ~/references, navigate to select ~/references.

v. Specify job submission parameters and settings
   1. See appendix MAKE SCREENSHOT IN APPENDIX

vi. Copy input files to inputFiles directory
   1. `cd NewProjectFolder/inputFiles`
   2. `cp` plink input files
      a. lgen
      b. map
      c. cov
      d. fam
      e. phen

3. Click ‘Select project’ button to refresh inputFiles file list. Designate lgen, map, cov, fam, phen files by selecting them in the list and selecting the appropriate button

vii. Check plink files using MACHTools
   1. FIRST…make sure that the lgen does not have a header line. (use more <filename> to take a look). If so the following sed command will remove the headerline in the source file:
      a. sed -i ‘1d’ <lgen_file.txt>
   2. Check the boxes for the input files that need to be checked and click ‘Check Input Files’ button.
   3. Jobs will launch for each input file to be checked. Outputs reside in directories corresponding to each job within input files directory
      a. ~/inputFiles/.checkLgen, etc.

viii. Make chromosome specific data files with MACHTools
   1. Choose the ‘Main’ tab and check the box for ‘makeChromSpecific…’, then click Run Beans
   2. The ‘Launched’ box will check when the job has been launched to the cluster. Outputs can be checked ~/inputFiles/.makeChromSpecific…/slurm-12345.out
   3. Upon successful completion, set genotype directory
      a. Click ‘Locate Path to GTs’ in Select Project tab and navigate to ~/inputFiles/ChromosomeSpecificData

ix. Filter, reorder and split them for MACH phasing against each reference haplotype set (HAPMap2, HAPMap3, HAP1000G) with MACHTools
   1. Check the box for filterChromSpecificMach and click ‘Run Beans’ button
   2. The ‘Launched’ box will check when the job has been launched to the cluster. Outputs can be checked in ~/inputFiles/.filterChromSpecificMach…/slurm-12345.out
   3. Repeat sequentially (not concurrently) with
      a. makeIdsUniqueInRefMaps
         i. inputFiles/.makeIdsUniqueInRefMaps
      b. compareAlleles
         i. inputFiles/.compareAlleles
      c. reorderSNPs
         i. inputFiles/.reorderSNPs
      d. splitChromSpecificMach
         i. inputFiles/.splitChromSpecificMach

x. Phase with MACH – runHapMACH.sh for each SplitJobs directory created in each of the ChromosomeSpecificData directory
   1. Click Run MachHap

2. NOTE!!! This script submits many jobs to the cluster and places them in queue. It requires a lot of time to run to completion. Use `squeue -u <<username>>` to check the status of jobs. If jobs need to be cancelled from the queue, use `scancel <<jobID>>` or `scancel -u <<username>>` to cancel all your jobs

xi. Use MACHTools to ligate the split haplotype files back together and fix marker strand.
   1. Choose the 'Main' tab and check the box for 'mergeMachHaplotypes', then click Run Beans
   2. The 'Launched' box will check when the job has been launched to the cluster. Outputs can be checked ~/inputFiles/.mergeMachHaplotypes/slurm-12345.out
   3. Choose the 'Main' tab and check the box for ''ParallelFixMarkerDefs…", then click Run Beans
   4. The 'Launched' box will check when the job has been launched to the cluster. Outputs can be checked ~/inputFiles/.planFixMarkerDef/slurm-12345.out

xii. Use runMinMACH.sh to impute genotypes; this makes dose and info files. This submits the jobs to a large memory machine queue; these jobs will run in the background. To see the status of the jobs, use `squeue`
   1. Click Run MinMach button

xiii. Use runMach2QTL.sh to perform association between imputed SNPs and quantitative traits.
   1. Click Run Mach2QTL button

xiv. Use runMach2Dat.sh to perform association between imputed SNPs and affection status.
   1. Click Run Mach2DAT button

xv. Use MACHTools to process and concatenate DAT.gz and QTL.gz files
   1. Choose the 'Main' tab and check the box for 'cleanupDatAndQTL, then click Run Beans
   2. The 'Launched' box will check when the job has been launched to the cluster. Outputs can be checked ~/inputFiles/.cleanupDatAndQtl/slurm-12345.out

xvi.

xvii. Unzip the DAT.gz and QTL.gz files, format and rename

1. In the directory where the DAT.gz and QTL.gz files are created enter the following command to unzip the files: gunzip <name_of_zipped_file.gz>

   a. Example: gunzip DAT.gz

2. Format to replace spaces with one tab and rename the new output.

   a. grep [0-9] QTL|sed ‘s/ \+/\t/g’>*<<name_of_new_QTL.txt>>*

   b. grep [0-9] DAT|sed ‘s/ \+/\t/g’>*<<name_of_new_DAT.txt>>*

xviii. Concatenate and process INFO files generated by minMACH with sed

1. Add Chr column to info files

   a. Add Chr_*.info filename as last column in each info file

      i. for f in Chr_*.info; do sed “s/$/\t$f/” $f > $f.infochrom; done

   b. Strip “Chr_” and “.info” from the file name

      i. for f in Chr_*.info.infochrom; do sed “s/.info//” $f > $f.infochrom

      ii. for f in Chr_*.info.infochrom; do sed “2,$s/Chr_//” $f > $f.infochrom

      iii. Delete the _1 from the Chr_1 column using nano in Chr_1.info.infochrom

2. Need to detail sed commands: concatenate the info files in each Hap directory, then concatenate the resulting three files into the final INFO file; NOTE you need to drop the header lines using either a shell command or in SQL . YOU will not be able to cast the Rsq and MAF as float with the header lines in the table.

   a. JSM used the following:

      i. head -1 Chr_1.info.infochrom > all.info #write the header line

      ii. tail –n +2 –q Chr_*.info >> all.info #write the rest, skipping the headers

3. Move info files to terminal server file space with FileZilla (See step X above)

xix. Import into sql server database

1. Import imputed files: .DAT, .QTL, and .INFO (See Appendix X)

2. Truncate strings , correct length /names and convert numeric values to int or float

   a. Need to detail this out; see pg 177 of notes

VI. Identify imputed SNPs of interest

a. Run a query in sql server to generate SNP table

- i. Uses rsq values from info file to select SNPs that were imputed reliably:

refHap	resqThresholds
Map1000G	0.5
HapMap2	0.3
HapMap3	0.3

- b. Export table to BR with FileZilla
- c. Use MACHTools to identify SNPs that drove imputation
- d. Move list of SNPs to terminal server file space with FileZilla
- e. Import into sql server

VII. Check clustering of SNPs that drove imputation

- a. Generate SNP Checker tables (don't check SNPs already checked)
- b. Check Clusters with Eclipse on terminal server
- c. Update genotypes in appropriate genotype tables in db (see example script in Nicole's folder; also pasted here as an appendix).

VIII. Repeat Step IV (above)

IX. Generate output statistics

- a. Calculate lambda for sampled genotypes for each trait of interest (in a single step)
- b. Generate publication figures:
  - i. Generate tables in sql server & export files to BR with FileZilla
    - 1. Find interesting
      - a. FILL IN instructions here to use the Interesting snps template
    - 2. Generate Genome-wide Manhattan plots
      - a. Make required tables in SQL and export
        - i. Open the most recently generated template (e.g., Eotaxin_3_NRP; code is appended below)
        - ii. Save query with name of new trait
        - iii. Do a global find and replace 'Eotaxin_3' for '*TraitName*'

          [NAME TRAIT EXACTLY THE SAME AS IN DATABASE (TARCC vs. ADNI)]

          Do a global find and replace 'Name' for '*User* (your name)'
        - iv. Execute query

        (Comment out all of the rows that say top 1000; used to test code)
        - v. Export Tables (4) - Walk through export Wizard for each file (4X); (to launch Wizard; right click on name of database, right click 'tasks', then select 'export')

Destination = Flat file, browse to Texas_access>PlotsTarcAdni>*TraitName*>*File name*

Select 'Columns in first data row', 'Column delimited Tab', finish

1. *User_Dataset*_Man_*Trait*
2. *User_Dataset*_Man_Trim_*Trait*
3. *User_Dataset*_Plink_*Trait*
4. *User_Dataset*_Imp_*Trait*

(*User* = Nicole, Bob, Ryan, etc; *Dataset* = TARCC or ADNI; *Trait* = new trait ID'ed above)

b. Move files to BR

- i. Open Filezilla (See Step 1b above)
- ii. In left pane, Browse to location of exported Tables: Texas_access>PlotsTarcAdni>*TraitName*>*File name*
- iii. In right pane, Browse to /Home/*User*/TarcPlot/*TraitName*
  (for ADNI data, use AdniPlot, not TarcPlot)
- iv. Drag folders from left to right pane

c. Run the tools for QQ and manhattan plots

- i. Open MRemote
  Connect to BlueRidge (see Appendix B below)
- ii. Type "qsub -I" to request one interactive node
  1. Qsub -I may be busy; use "ssh largemem-0-0" instead
- iii. Change directory (cd) to: /Home/*User*/TarcPlot/*TraitName*
- iv. Run R (Command: 'R' then hit 'enter')
  1. Type R; return
  2. [Source the code
     - a. Type: Source(http://dl.dropbox.com/u/66281/0_Permanenet/qqman.r); return] DEPRECATED
  3. cp .tar from /projects/sequence…/vol1/chat/TARCImpute/niphilli/qqman_0.1.2.tar to directory where you are working (where your Manhattans will be)

4. Run this: install.packages("qqman_0.1.2.tar", repos = NULL, type="source")
5. Then your manhattan function should work.
6. 1. Read data table into R desired_dataset_name<-read.table (“name_of_man_table.txt”, header=T)
7. 2. Run the pdf(“desired_name_of_pdf_output.pdf”) step first to open a pdf for the output.
8. 3. Then run manhatttan(dataset name) to generate the plot.
9. Possible error may crop up due to the order of columns. Awk them into the right order

v. Read incat data once for each file
Command: ‘*TraitName*_Man<-read.table (*“FileName.txt”, header=T)’*
(Example FileName”
*User_Dataset*_Man_*TraitName;*
*User_Dataset*_Man_Trim_*Trait*;
*User_Dataset*_Plink_*TraitName;*
*User_Dataset*_Imp_*TraitName*)

vi. Run pdf commands for Manhattan Plot:

1. Pdf(“OutputFileName.pdf”) hit enter
2. Manhattan(*TraitName*_Man) hit enter
3. Dev.off () hit enter

vii. Run pdf commands for Q-Q Plots:

1. Pdf(“OutputFileName.pdf”) hit enter
2. qq(*TraitName*_plink **$P**) hit enter
3. Dev.off () hit enter

viii. To change working directories in R, use the following command

1. setwd (“<<path_to_new_dir>>”)

3. Run Metal

a. Connect to Blue Ridge via MRemote (see Appendix B below)

b. Change directory (cd) to: /Home/*User*/TarcPlot/*TraitName*

c. Copy Metal script into the new Trait folder

Example command: cp
../Eotaxin_3/metal_Eotaxin_3.sh
../*TraitName*/metal_*TraitName*.sh

d. Open emacs
emacs . hit enter
Scroll down with arrow keys to locate the metal_*TraitName*.sh file
Hit enter
Global find and replace old *TraitName* with new *TraitName*
    i. Hit Ctrl+x, Ctrl+s to save
    ii. Hit Ctrl+x, Ctrl+c to exit emacs

e. Run Metal in bash shell
Command: metal <metal_*TraitName*.sh

f. To look at results;
open emacs
emacs . hit enter
Scroll down with arrow keys to locate the new metal file (METAANALYSIS.tbl) hit enter

g. To write most significant results to a file;
run a grep command
grep 'e-0' METAANALYSIS.tbl >
*TraitName*_SigP_metal.tbl

4. Generate Local Manhattan plots (LocusZoom) for all SNPs in TARCC & ADNI where $p \leq 10^{-6}$ AND any metal results where $p \leq 10^{-7}$

    a. Connect to Blue Ridge via MRemote (see Appendix B below)

    b. Change directory (cd) to:
    /Home/*User*/TarcPlot/*TraitName*

    c. Copy LocusZoom script into the new Trait folder
    Example command: cp
    ../Eotaxin_3/runLocuszoom.sh ../*TraitName*/
    runLocuszoom.sh

    d. To look at results;
    open emacs
    emacs . hit enter
    Scroll down with arrow keys to locate the new LocusZoom file
    hit enter
    Change the text file that LocusZoom references to:
    *User_Dataset*_Man_*TraitName*

Change the SNP reference to the current SNP of interest
Save: Ctrl xs
Close: Ctrl xc

e. Run LocusZoom
Command: cat runLocuszoom.sh

f. Highlight the text that is returned

g. Right click to paste and execute the command (this will take a while to run)

h. To view results:
Open Filezilla
Refresh Blueridge folder
A new folder will appear with the name of the SNP just examined
Folder will contain a pdf file showing local Manhattan plot
Rename folder to indicate if result are from TARCC or ADNI

VI. Check the typed “driver” SNPs that were important for the significant imputation association results

A. Generate list of driver SNPs

a. Modify the MACHTools.xml file

i. Navigate to the directory where the imputation was performed

ii. emacs .

1. Key down to make the following bean active by removing the “<!--“ from before the block and the “-->” after the block (highlighted below). The color of these lines will change from red (inactive) to multicolor (active).

```
<!--        <<ref bean="CalculateCorrForList"/>>-->
```

IMPORTANT NOTE!! The commands at the beginning of this file (that appear before the first line that says <list>) and the commands at the end of this file (that appear after the second line that says </list>) should not be altered and should appear as active at all times. Only the lines between the two lines that say list should ever be commented in or out by the user.

2. Key down close the bottom of the file to the bean id properties for the “CalculateCorrForList” bean; enter the file name for the .txt file which has the list of imputed SNPs of interest (example of filename to change highlighted below). This .txt file should be generated and placed in the same “impute” directory that you are working on.

```
  <bean id="CalculateCorrForList"
class="org.renci.machtool.interpreter.ParallelIdentifyWhichS
npsDroveImputationSingleChromForLIstBasedOnRe
fMapRunnable">
    <property name="minCorr" value="0.10"/>
    <property name="listOfSnps"
value="RBM_metal_imp_SNPs_of_interest.txt"/>
    <property name="properties" ref="configProperties"/>
```

3. Save the .xml file and close.

b. Modify the machTools.properties file

i. Make sure you are still in the imputation directory

ii. emacs .

1. Make sure that the lines that say “projectRefDir=, projectMachHapDir=, and projectMiniMachHapDir=” only have the reference database of interest shown. In previous MACHTool steps for imputation, all three databases were specified, separated by commas; however, for this bean, there must only be one specified.

iii. Save the properties file and close.

c. Run MACHTools and designate a log file name

i. MACHTool MACHTool.xml >*log_file_name*.log

d. Format outputs

i. Add file name as a column

1. For f in Corr_list_of_snps_*.txt; do sed “s/$/\t$f/” $f > $f.corrChrom; done

ii. Strip file extension from that column

1. For f in Corr_list_of_snps_*.txt.corrChrom.corrChrom; do sed “s/.txt//” $f > $f.corrChrom; done

iii. Strip file name leaving chromosome number

1. For f in Corr_list_of_snps_*.txt.corrChrom.corrChrom; do sed “s/Corr_list_of_snps_//” $f > $f.corrChrom; done

iv. Move Chr number to front of first column

1. For f in Corr_list_of_snps_*.txt.corrChrom.corrChrom.corrChrom; do awk b=$4”:”$1; print b, $2, $3 $f.finalchrom; done

v. Remove chrom from header in Corr_list_of_snps_1.txt…finalchrom

vi. Write header to new file

1. Head -1 Corr_list_of_snps.txt….finalchrom > allcorrChrom.txt

vii. Write the rest of the files to allcorrChrom.txt

1. Tail –n +2 –q Corr_list_of_snps_*.txt….finalchrom >> allCorrChrom.txt

I. Making plink files using lgen style (See Long-format filesets In the plink documentation (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml))

- a LGEN file containing genotypes (5 columns, one row per genotype)
- a MAP file containing SNPs (4 columns, one row per SNP)
- a FAM file containing individuals (6 columns, one row per person)

Consider the following example: A MAP file `test.map`

```
1 snp2 0 2
2 snp4 0 4
1 snp1 0 1
1 snp3 0 3
5 snp5 0 1
```

as described above. A FAM file `test.fam`

```
1 1 0 0 1 2
2 1 0 0 2 2
2 2 0 0 1 1
9 1 1 2 0 0
```

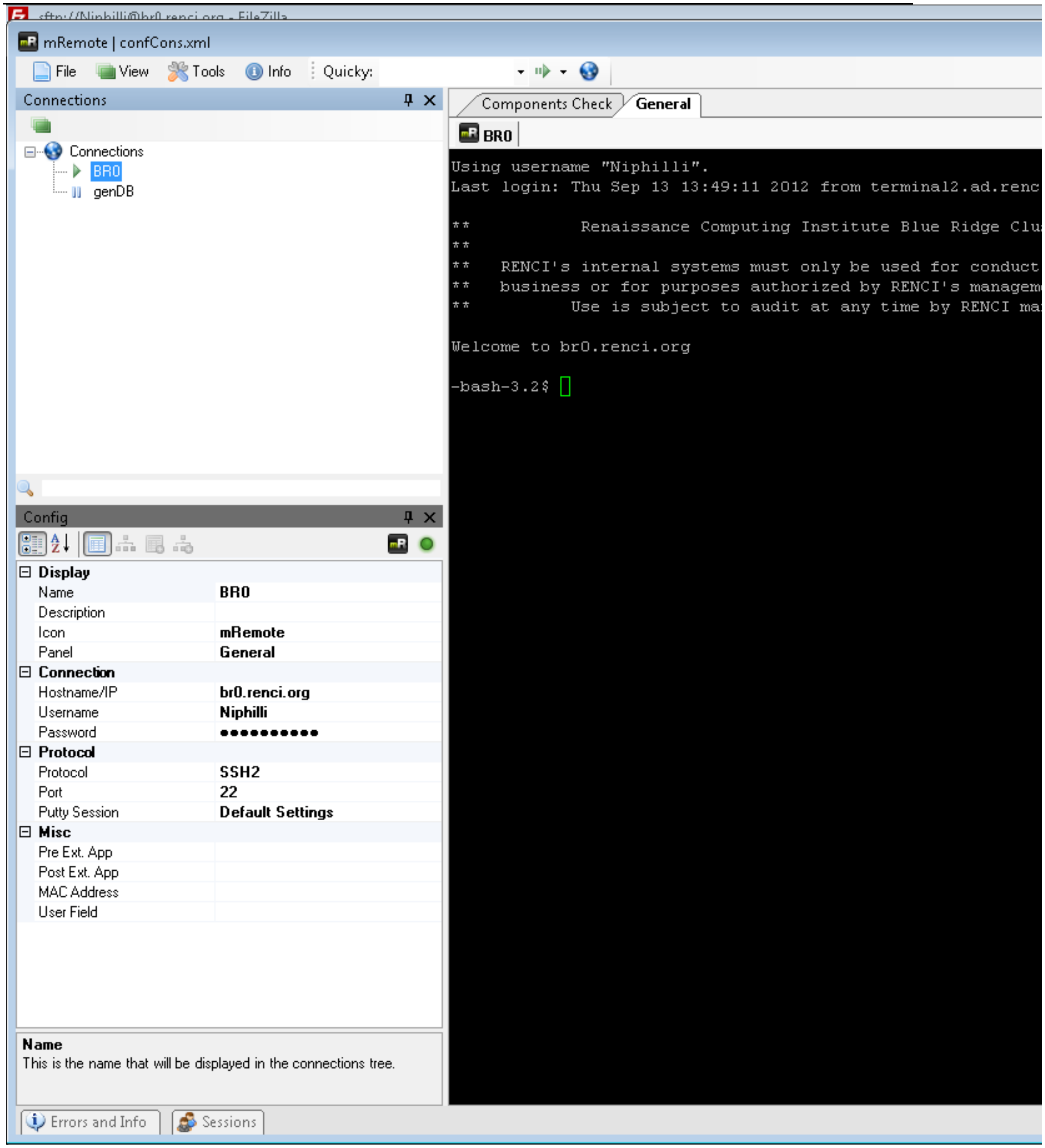
as described below. Finally, an LGEN file, `test.lgen`

```
1 1 snp1 A A
1 1 snp2 A C
1 1 snp3 0 0
2 1 snp1 A A
2 1 snp2 A C
2 1 snp3 0 0
2 1 snp4 A A
2 2 snp1 A A
2 2 snp2 A C
2 2 snp3 0 0
2 2 snp4 A A
```

The columns in the LGEN file are

```
family ID
individual ID
snp ID
allele 1 of this genotype
allele 2 of this genotype
```

Appendix B- Establishing Blue Ridge Server Connection



```
Using username "Niphilli".
Last login: Thu Sep 13 13:49:11 2012 from terminal2.ad.renc

**          Renaissance Computing Institute Blue Ridge Clu
**
**   RENCI's internal systems must only be used for conduct
**   business or for purposes authorized by RENCI's managem
**          Use is subject to audit at any time by RENCI ma

Welcome to br0.renci.org

-bash-3.2$
```

Display	
Name	BR0
Description	
Icon	mRemote
Panel	General
Connection	
Hostname/IP	br0.renci.org
Username	Niphilli
Password	●●●●●●●●●●●
Protocol	
Protocol	SSH2
Port	22
Putty Session	Default Settings
Misc	
Pre Ext. App	
Post Ext. App	
MAC Address	
User Field	

Name
This is the name that will be displayed in the connections tree.

Figure 1- Creating a BR0 connection and open a BR bash shell. Open MRemote; File→New Connection; fill in the configuration as shown. Right Click on the new connection name in the top pane and select "Connect". A new bash shell will open in the right pane (tab named "General").

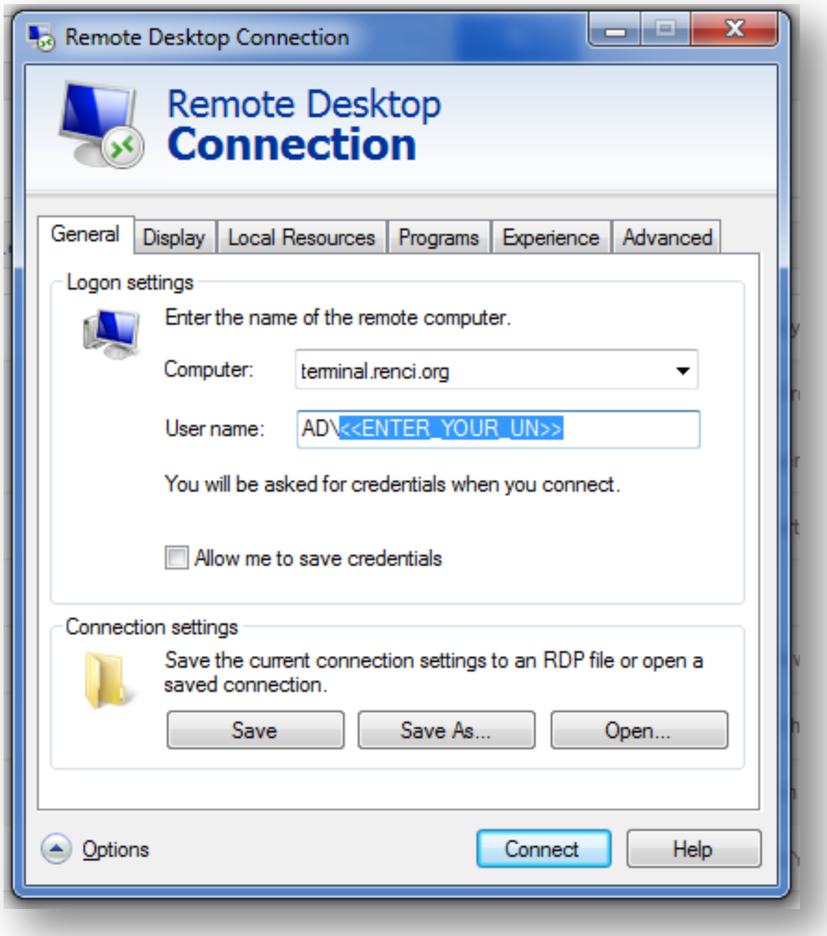
Appendix C- Manual SNP Cluster Checking

Access the Remote Desktop Space

1. Launch Remote Desktop.

   Note: if you do not see the icon in your Start Menu, you may need to search for the application. ***Start→Search→***Type "Remote Desktop" in the field. Double click on the Remote Desktop Connection line in the results window. You can drag and drop this item onto your desktop to create a shortcut to the Remote Desktop Application.

2. Enter the Computer path as shown below. In the User name field, enter "AD\" followed by your RENCI assigned user name.



Remote Desktop Connection

Logon settings

Enter the name of the remote computer.

Computer: terminal.renci.org

User name: AD\<<ENTER_YOUR_UN>>

You will be asked for credentials when you connect.

□ Allow me to save credentials

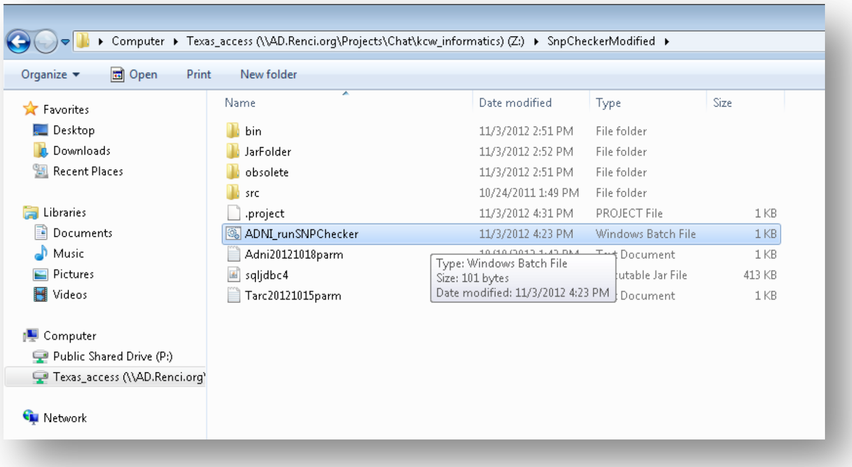
Connection settings

Save the current connection settings to an RDP file or open a saved connection.

3. Click ***Connect***.
4. Once in the remote space, click on your user icon and enter your RENCI account password. Your screen should show the RENCI remote desktop space (a plain black desktop). You can minimize the remote space at anytime and return to your local space by using the blue tab in the top, center of the screen.

Checking SNP Clustering

1. Open a Computer browser. Navigate to the Network location called Texas_Access (Z:\\) then enter the directory called SnpCheckerModified (\Texas_Access\SnpCheckerModified).
2. Double-click the file called ADNI_runSNPChecker (a Windows Batch file), and click **Run** in the Security Warning popup window. See the figure below for the path to the file.

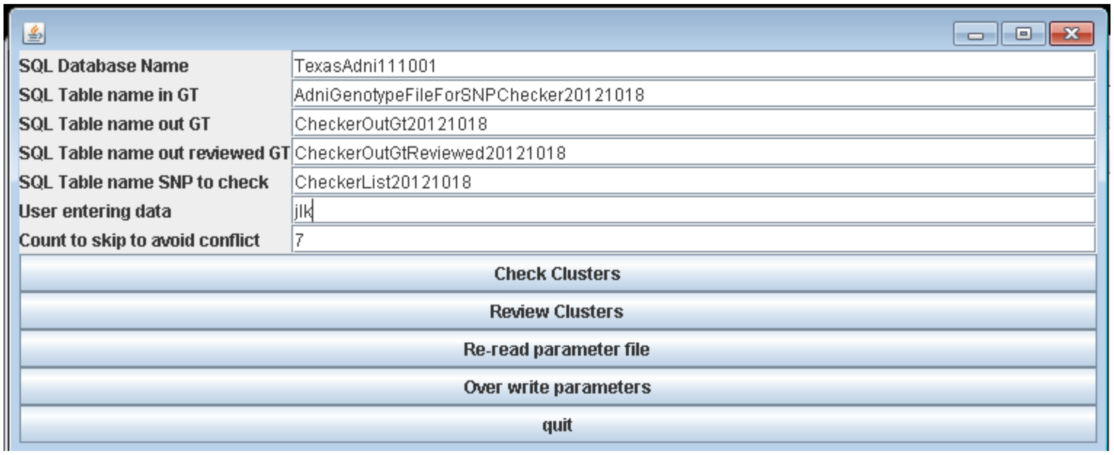


Computer ▸ Texas_access (\\AD.Renci.org\Projects\Chat\kcw_informatics) (Z:) ▸ SnpCheckerModified ▸

Name	Date modified	Type	Size
bin	11/3/2012 2:51 PM	File folder	
JarFolder	11/3/2012 2:52 PM	File folder	
obsolete	11/3/2012 2:51 PM	File folder	
src	10/24/2011 1:49 PM	File folder	
.project	11/3/2012 4:31 PM	PROJECT File	1 KB
ADNI_runSNPChecker	11/3/2012 4:23 PM	Windows Batch File	1 KB
Adni20121018parm	[illegible]	Text Document	1 KB
sqljdbc4		utable Jar File	413 KB
Tarc20121015parm		Document	1 KB

Type: Windows Batch File
Size: 101 bytes
Date modified: 11/3/2012 4:23 PM

3. In the first window, change "kirk" to your initials in the ***User entering data*** field and enter the number 7 in the ***Count to skip to avoid conflict*** field. See the figure below for guidance.



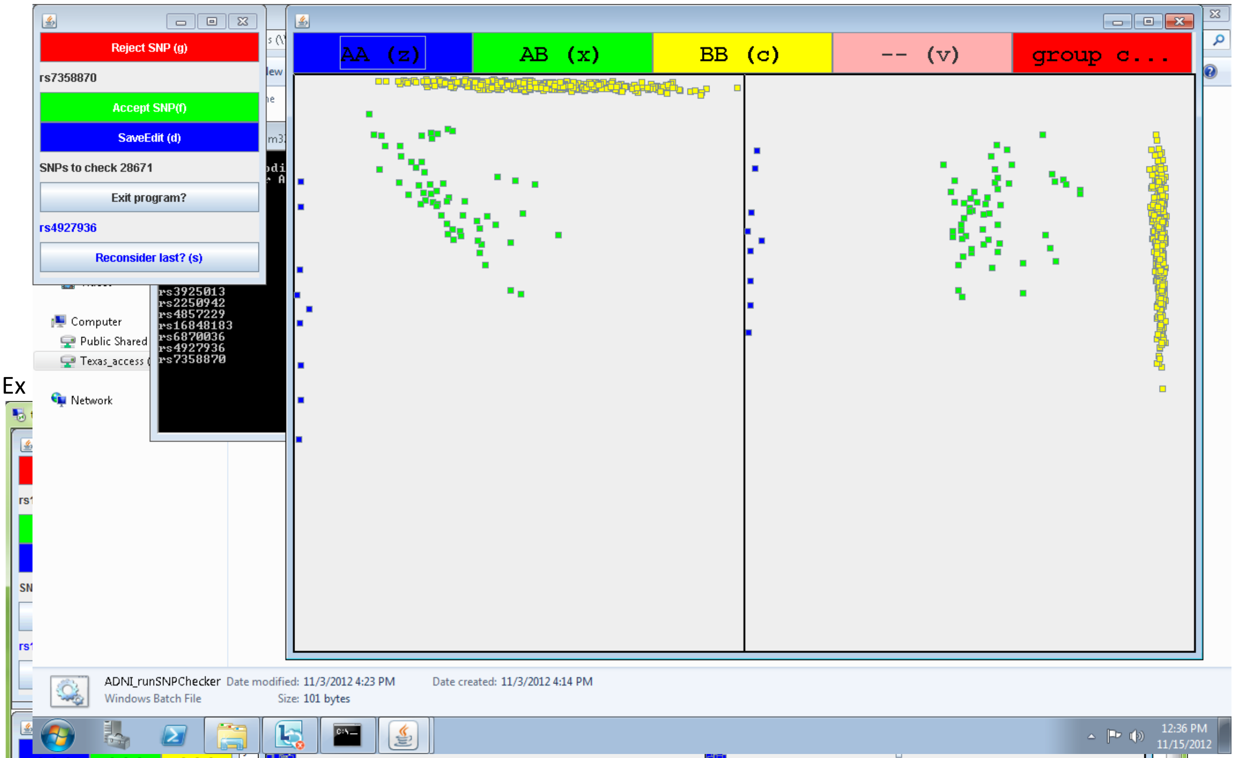
SQL Database Name	TexasAdni111001
SQL Table name in GT	AdniGenotypeFileForSNPChecker20121018
SQL Table name out GT	CheckerOutGt20121018
SQL Table name out reviewed GT	CheckerOutGtReviewed20121018
SQL Table name SNP to check	CheckerList20121018
User entering data	jlk
Count to skip to avoid conflict	7

4. Click ***Check Clusters***.

5. The program will launch and automatically show the first SNP for you to check.
   a. Editing SNPs- Edit the SNP if there are data points that called incorrectly. Note, edits can be made in either view.
      i. Edit many data points at once:
         1. Select a group of data points by dragging a box around the points.
         2. Use the bottom row quick keys to designate the genotype that you wish to call:
            z = blue (AA) x = green (AB) c = yellow (BB) v = clear (do not call)
         3. Enter **g (group change)** to execute the group change and make all selected data points the designated genotype.
      ii. Edit single data points:
         1. Use the bottom row quick keys to designate the genotype that you wish to call:
            z = blue (AA) x = green (AB) c = yellow (BB) v = clear (do not call)
         2. Click on any individual data point to change the point to the designated genotype.
      iii. Once all required edits are made**, Save Edits (quick key = d)**.
   b. **Accept SNP (quick key = f)** if all data points fall into three distinct clusters (homozygous A, homozygous B, and heterozygous AB).
   c. **Reject SNP (quick key = g)** if a SNP does not have distinct clustering of data points/genotypes.
   d. If you need to re-visit the previous SNP, click **Reconsider last? (quick key = s)**, and the last marker will be reloaded without the edits previously made.
6. At any time, you can exit the program by clicking **Exit program?** And the current SNP will not be saved. Log out of the remote space when you are finished checking.

Examples

Ex 1: Accept; no edits required:



rs7358870

SNPs to check 28671

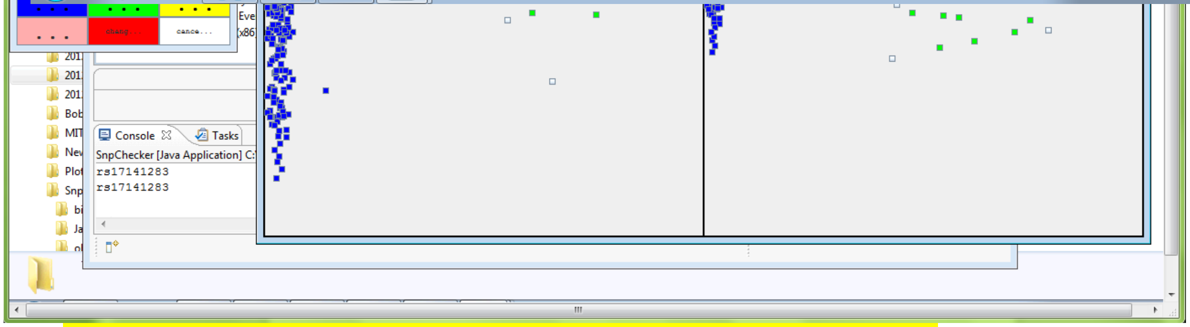
rs4927936

```
rs3925013
rs2250942
rs4857229
rs16848183
rs6870036
rs4927936
rs7358870
```

ADNI_runSNPChecker Date modified: 11/3/2012 4:23 PM Date created: 11/3/2012 4:14 PM
Windows Batch File Size: 101 bytes

12:36 PM
11/15/2012

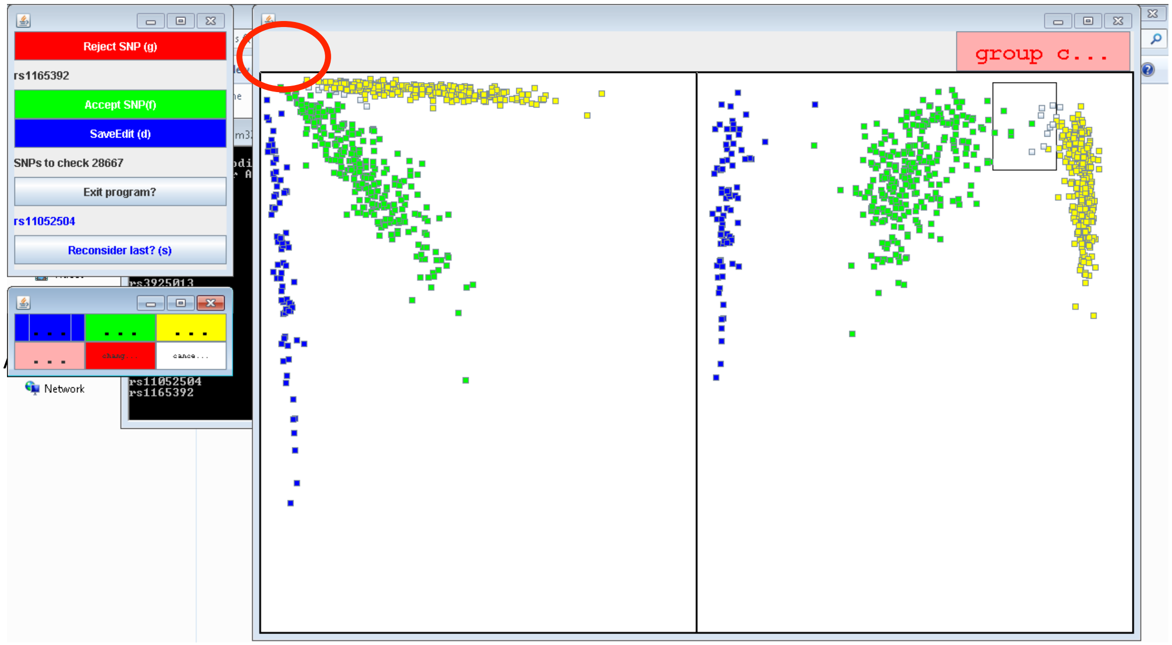
Ex



```
SnpChecker [Java Application] C:
rs17141283
rs17141283
```

Ex 3: Edit; there are many points too called too close to the origin (area circled in red):

Before edits:

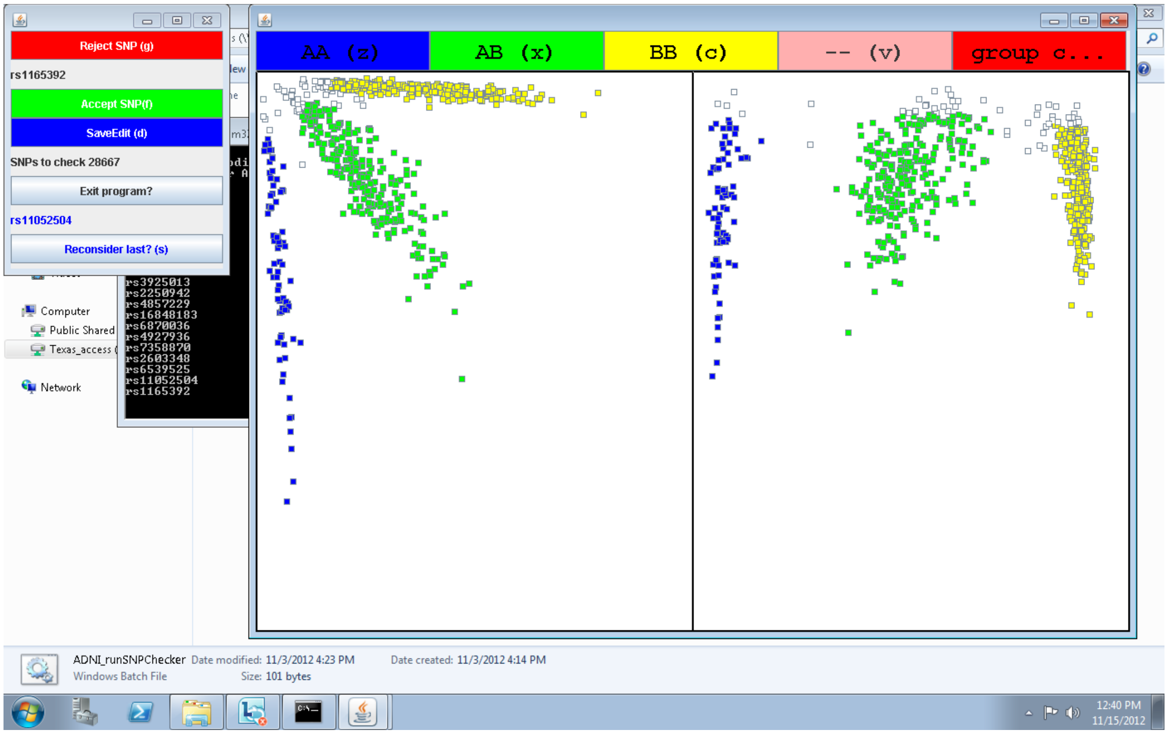


rs1165392

SNPs to check 28667

rs11052504

```
rs3925013
rs11052504
rs1165392
```



rs1165392

SNPs to check 28667

rs11052504

```
rs3925013
rs2250942
rs4857229
rs16848183
rs6870036
rs4927936
rs7358870
rs2603348
rs6539525
rs11052504
rs1165392
```

ADNI_runSNPChecker Date modified: 11/3/2012 4:23 PM Date created: 11/3/2012 4:14 PM
Windows Batch File Size: 101 bytes

12:40 PM
11/15/2012

Appendix B – MaCHTools Example Outputs

Checking input files:

```
INFO:
Running...org.renci.machtoolv4.interpreter.MakeChromSpecificMachFromPlinkRunnable
The chromosome file had 23 lines.
There were 23 unique lines in the chromosome file.
The sample fam had 692 lines
The map file had 903007 lines of which 901454 are being used
Chrom 1 has 70913 snps.
Chrom 2 has 73588 snps.
Chrom 3 has 60373 snps.
Chrom 4 has 55733 snps.
Chrom 5 has 56153 snps.
Chrom 6 has 55981 snps.
Chrom 7 has 46766 snps.
Chrom 8 has 48369 snps.
Chrom 9 has 41228 snps.
Chrom 10 has 47990 snps.
Chrom 11 has 44333 snps.
Chrom 12 has 42333 snps.
Chrom 13 has 34069 snps.
Chrom 14 has 27924 snps.
Chrom 15 has 25958 snps.
Chrom 16 has 27573 snps.
Chrom 17 has 20558 snps.
Chrom 18 has 26401 snps.
Chrom 19 has 11840 snps.
Chrom 20 has 22743 snps.
Chrom 21 has 12496 snps.
Chrom 22 has 11473 snps.
Chrom X has 36659 snps.
Had to create chromsome specific data directory called
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData.
Had to create chromsome specific data directory called
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/TempChromosomeSpecificRawDat
a.
Sorting genotype: dot printed for each 10M genotypes read 1000M/row
.............................................................
Genotype records for chromosome 1 = 47992881
Genotype records for chromosome 2 = 49755287
Genotype records for chromosome 3 = 40773057
Genotype records for chromosome 4 = 37586524
Genotype records for chromosome 5 = 37941035
Genotype records for chromosome 6 = 37858487
Genotype records for chromosome 7 = 31619602
Genotype records for chromosome 8 = 32700978
Genotype records for chromosome 9 = 27852960
Genotype records for chromosome 10 = 32518836
Genotype records for chromosome 11 = 29975294
Genotype records for chromosome 12 = 28629138
Genotype records for chromosome 13 = 23018662
Genotype records for chromosome 14 = 18885474
Genotype records for chromosome 15 = 17595346
Genotype records for chromosome 16 = 18688357
Genotype records for chromosome 17 = 13953762
Genotype records for chromosome 18 = 17852117
Genotype records for chromosome 19 = 8029414
Genotype records for chromosome 20 = 15426694
Genotype records for chromosome 21 = 8462338
Genotype records for chromosome 22 = 7796449
Genotype records for chromosome X = 24879711
...............................................
Warning there were a total of 29489 samples-snp pairs with discordant
genotypes.
A list is written to
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/discordantGenotype01.txt.
Review before deciding to proceed.
total of 47992881 genotypes for chromosome 1 read
.................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_01.ped written with 49071796 genotypes.
......................................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_01.dat written with 70983 lines, including 692 samples and
70913 SNPs.
Also includes 69 traits and covariates.
.................................................
Warning there were a total of 31097 samples-snp pairs with discordant
genotypes.
```

```
A list is written to
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/discordantGenotype02.txt.
Review before deciding to proceed.
total of 49755287 genotypes for chromosome 2 read
..................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_02.ped written with 50922896 genotypes.
.......................................................................
..
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_02.dat written with 73658 lines, including 692 samples and
73588 SNPs.
Also includes 69 traits and covariates.
```

Filtering by user defined QC measures

```
INFO:
Running...org.renci.machtoolv4.interpreter.FilterChromSpecificMachRunna
ble
Created chromsome specific data directory called
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Backu
p_org.
The chromosome file had 23 lines.
There were 23 unique lines in the chromosome file.
There are expected to be 69 traits plus covariates and 70913 SNPs based
on the dat file for chrom 1.
.................................................
Finished reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_01.ped
.................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_01.ped written with 49071796 genotypes.
......................................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_01.dat written with 70983 lines, including 692 samples and
70913 SNPs.
Also includes 69 traits and covariates.
There are expected to be 69 traits plus covariates and 73588 SNPs based
on the dat file for chrom 2.
..................................................
Finished reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_02.ped
..................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_02.ped written with 50922896 genotypes.
.......................................................................
..
```

```
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputatio
n/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificData/Genot
ypeChrom_02.dat written with 73658 lines, including 692 samples and
73588 SNPs.
Also includes 69 traits and covariates.
```

Make SNP IDs unique in references

```
INFO:
Running...org.renci.machtoolv4.interpreter.MakeIdsUniqueInRefMapsRu
nnable
The chromosome file had 23 lines.
There were 23 unique lines in the chromosome file.
Reading ref map for 1000G dir  for chrom 1 had 943778 snps
Reading ref map for 1000G dir  for chrom 2 had 1008166 snps
Reading ref map for 1000G dir  for chrom 3 had 855225 snps
Reading ref map for 1000G dir  for chrom 4 had 876838 snps
Reading ref map for 1000G dir  for chrom 5 had 758761 snps
Reading ref map for 1000G dir  for chrom 6 had 793578 snps
Reading ref map for 1000G dir  for chrom 7 had 712005 snps
Reading ref map for 1000G dir  for chrom 8 had 672841 snps
Reading ref map for 1000G dir  for chrom 9 had 525998 snps
Reading ref map for 1000G dir  for chrom 10 had 609026 snps
Reading ref map for 1000G dir  for chrom 11 had 598263 snps
Reading ref map for 1000G dir  for chrom 12 had 579840 snps
Reading ref map for 1000G dir  for chrom 13 had 433811 snps
Reading ref map for 1000G dir  for chrom 14 had 394379 snps
Reading ref map for 1000G dir  for chrom 15 had 354493 snps
Reading ref map for 1000G dir  for chrom 16 had 378740 snps
Reading ref map for 1000G dir  for chrom 17 had 328899 snps
Reading ref map for 1000G dir  for chrom 18 had 343862 snps
Reading ref map for 1000G dir  for chrom 19 had 281732 snps
Reading ref map for 1000G dir  for chrom 20 had 266037 snps
Reading ref map for 1000G dir  for chrom 21 had 170073 snps
Reading ref map for 1000G dir  for chrom 22 had 170949 snps
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/references/1000G/Chr_X.map does not exist
All ref map SNP names are at unique place on one chromsome.
```

Reorder sample SNPs to match reference SNPs

```
INFO:
Running...org.renci.machtoolv4.interpreter.ReorderSNPsBasedOnRefMap
Runnable
The chromosome file had 23 lines.
There were 23 unique lines in the chromosome file.
Created chromsome specific data directory called
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/projects/sequence_an
alysis/vol1/chat/TARCImpute/txjeffm/UCSF/imputation/MaCHTools_proje
cts/MT4test/F7_validation/inputFiles/ChromosomeSpecificData.
Checking MACH dat files
Checking project map file
```

Compare allele definitions between samples and references

```
INFO:
Running...org.renci.machtoolv4.interpreter.CompareAllelesRunnable
The chromosome file had 23 lines.
There were 23 unique lines in the chromosome file.
There are expected to be 69 traits plus covariates and 70913 SNPs
based on the dat file for chrom 1.
.................................................
Finished reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/GenotypeChrom_01.ped
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/references/1000G/Chr_1.gz
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................

Reading ref map for 1 had 943778 lines.
All ref map SNP names are at unique place on  chromsome 1.
rs41457352 had alleles GT the test data and TA for the ref
haplotypes for chrom 1 for 1000G
There are expected to be 69 traits plus covariates and 73588 SNPs
based on the dat file for chrom 2.
..................................................
Finished reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/GenotypeChrom_02.ped
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/references/1000G/Chr_2.gz
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................

Reading ref map for 2 had 1008166 lines.
All ref map SNP names are at unique place on  chromsome 2.
There are expected to be 69 traits plus covariates and 60373 SNPs
based on the dat file for chrom 3.
.........................................
```

```
Finished reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/GenotypeChrom_03.ped
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/references/1000G/Chr_3.gz
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................

Reading ref map for 3 had 855225 lines.
All ref map SNP names are at unique place on chromosome 3.
```

Split chromosomes into manageable jobs for haplotype estimation

```
INFO:
Running...org.renci.machtoolv4.interpreter.SplitChromSpecificMachRu
nnable
Created chromsome specific data directory called
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs.
The chromosome file had 23 lines.
There were 23 unique lines in the chromosome file.
There are expected to be 69 traits plus covariates and 70913 SNPs
based on the dat file for chrom 1.
.................................................
Finished reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/GenotypeChrom_01.ped
.
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_0.ped written with 1384692 genotypes.
..
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_0.dat written with 2071 lines, including 692
samples and 2001 SNPs.
Also includes 69 traits and covariates.
.
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_1.ped written with 1384692 genotypes.
..
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_1.dat written with 2071 lines, including 692
samples and 2001 SNPs.
Also includes 69 traits and covariates.
```

```
.
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_2.ped written with 1384692 genotypes.
..
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_2.dat written with 2071 lines, including 692
samples and 2001 SNPs.
Also includes 69 traits and covariates.
.
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_3.ped written with 1384692 genotypes.
..
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_3.dat written with 2071 lines, including 692
samples and 2001 SNPs.
Also includes 69 traits and covariates.
.
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_4.ped written with 1384692 genotypes.
..
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_4.dat written with 2071 lines, including 692
samples and 2001 SNPs.
```

Merge MaCH haplotypes back together after haplotype estimation

```
INFO:
Running...org.renci.machtoolv4.interpreter.MergeMachHaplotypesRunna
ble
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_0.dat
There are expected to be 69 traits plus covariates and 2001 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_0.gz
.............
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_1.dat
There are expected to be 69 traits plus covariates and 2001 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_1.gz
```

```
.............
Ligating haps
.............
...
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.dat written with 3071 lines, including 3001
SNPs.
Also includes 69 traits.
Hap file
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.gz written.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.dat
There are expected to be 69 traits plus covariates and 3001 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.gz
.............
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_2.dat
There are expected to be 69 traits plus covariates and 2001 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_2.gz
.............
Ligating haps
.............
....
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.dat written with 4071 lines, including 4001
SNPs.
Also includes 69 traits.
Hap file
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.gz written.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.dat
There are expected to be 69 traits plus covariates and 4001 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_temp.gz
.............
```

```
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_3.dat
There are expected to be 69 traits plus covariates and 2001 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/SplitJobs/Chr_1_3.gz
.............
Ligating haps
.............
.....
```

Resolving ambiguities regarding strand orientation for sample SNPs – Chr 1

```
INFO:
Running...org.renci.machtoolv4.interpreter.SingleChromFixMarkerDefI
nHaplotypesRunnable
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.dat
There are expected to be 69 traits plus covariates and 70913 SNPs
based on the dat file for chrom 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.gz
.............
Reading ref map for 1 had 943778 lines.
All ref map SNP names are at unique place on chromosome 1.
Reading
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/references/1000G/Chr_1.gz
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................

1 incompatible markers
29874 markers with assays from the other strand based on alleles
observed
10112 markers that are ambiguous whether they are coded the same
way as the ref that need to be checked
Switching alleles of test haps to base pairing compliment to make
consistent with ref haps
.............................
Testing best marker configuration of markers where base pairing
rules do not specify which way to code snps
...................................................................
.................................
```

```
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...................................................................
.................................
...........
10112 of 10112 snps that were ambiguously coded were recoded.
...................................................................
...
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.snp written with 70913 lines, including 70913 SNPs.
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.gz backed up to
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Backup
1 is/are being removed because they were incompatible the reference
haplotypes based on observed alleles
 or are gt or AT polymorphisms and the reference genotypes were
monomorphic
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.dat backed up to
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Backup
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.snp backed up to
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Backup
.........Now excluded 9823    rs41457352
.............................................................
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.dat written with 70982 lines, including 70912 SNPs.
Also includes 69 traits.
...................................................................
...
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.snp written with 70912 lines, including 70912 SNPs.
```

```
Hap file
/projects/sequence_analysis/vol1/chat/TARCImpute/txjeffm/UCSF/imput
ation/MaCHTools_projects/MT4test/F7/inputFiles/ChromosomeSpecificDa
ta/Chr_1.gz written.
1 chromsome files changed.
```

Appenix C – Factor VII Validation Table

Trait	**Chromosome**	**SNPs**	**P-value in published analysis**	**P-value in MaCHTools validation**
Factor VII	13	rs7630910 rs1755685 rs2637255 rs1545251 rs6762390 rs12151434 rs9600699 rs7804867 rs207482 rs113114	1.87E-06 2.11E-06 5.38E-06 6.10E-06 6.14E-06 7.71E-06 8.81E-06 9.59E-06 1.37E-05 1.64E-05	1.87E-06 2.11E-06 5.38E-06 6.10E-06 6.14E-06 7.71E-06 8.81E-06 9.59E-06 1.37E-05 1.64E-05

Table 1: Summary of the most significant signals observed in each QTL analysis. Factor VII p-values were pulled from Manhattan tables before meta-analysis. All significant p-values were replicated identically.

.

Appendix D – D Plots

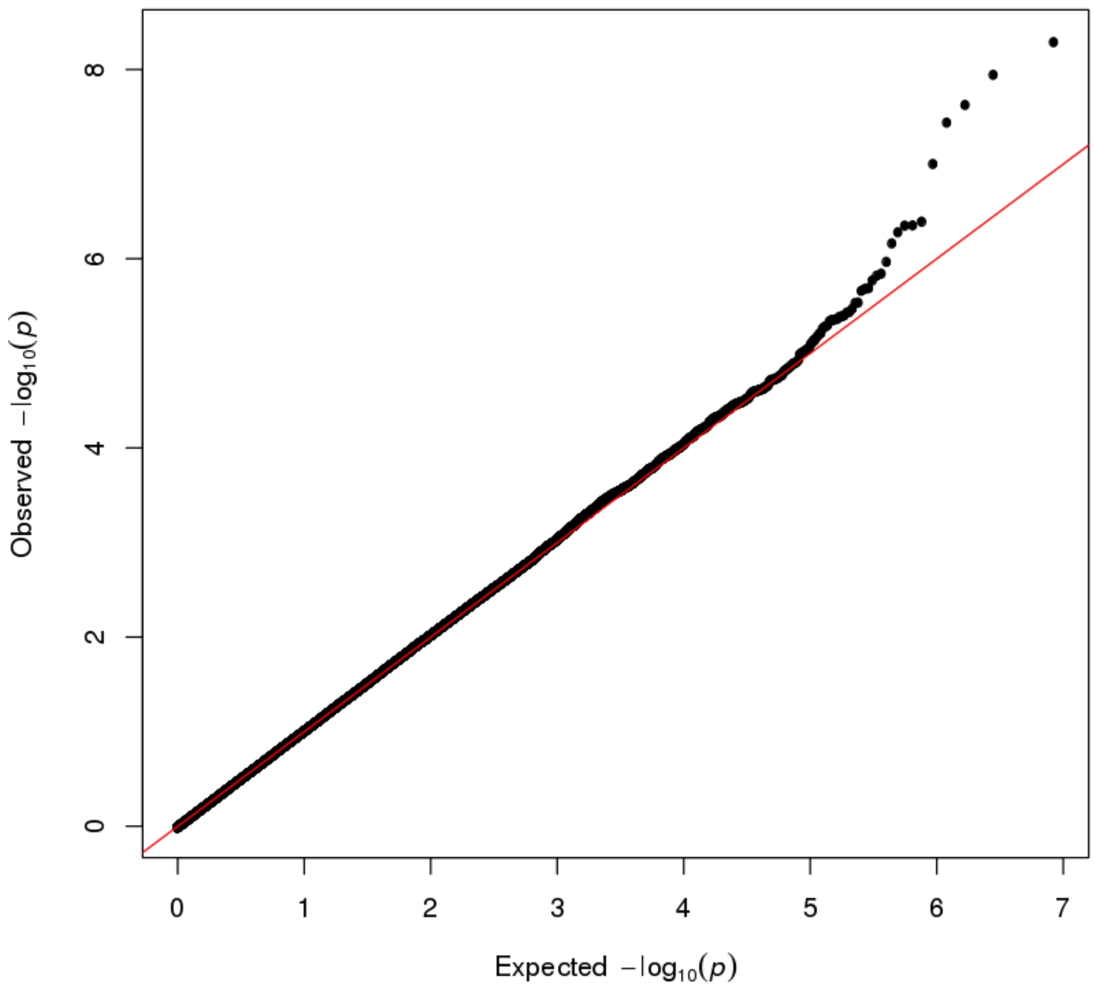


Figure 11. QQ plot of expected vs. observed p-values in GWAS of D using participants with D scores recorded during wave 1.

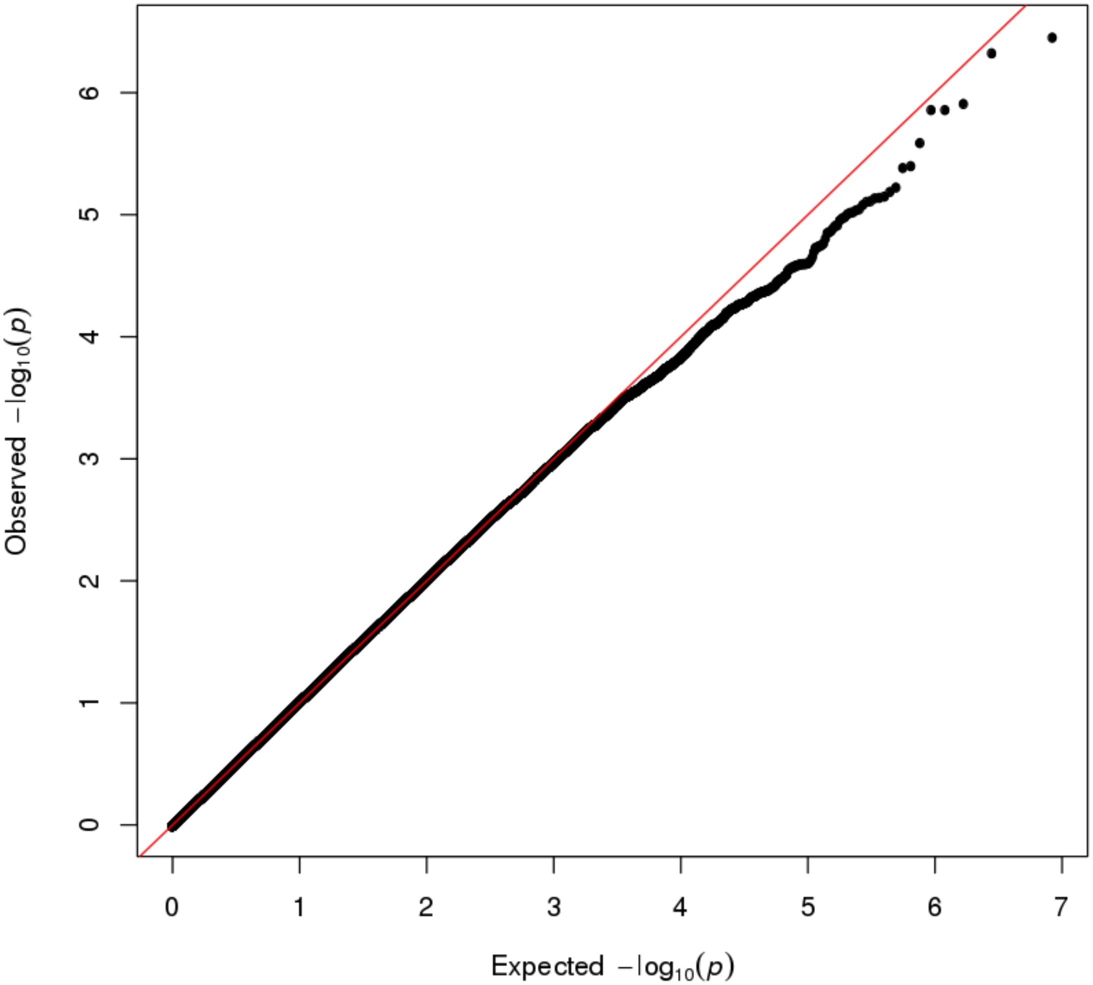


Figure 11. QQ plot of expected vs. observed p-values in GWAS of D using participants with D scores recorded during wave 2.

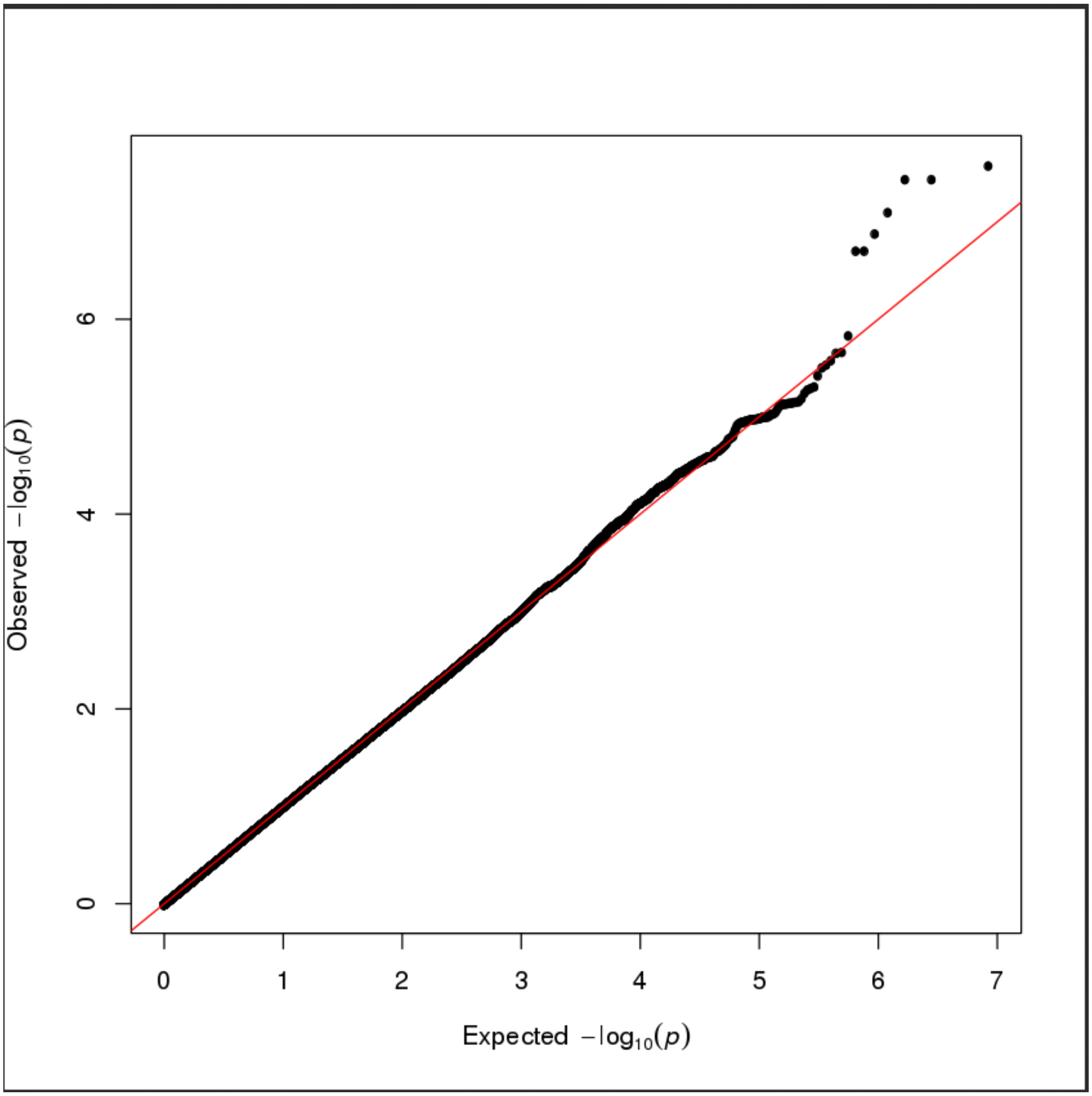


Figure 12. QQ plot of expected vs. observed p-values in GWAS of D using participants with D scores recorded during wave 3.

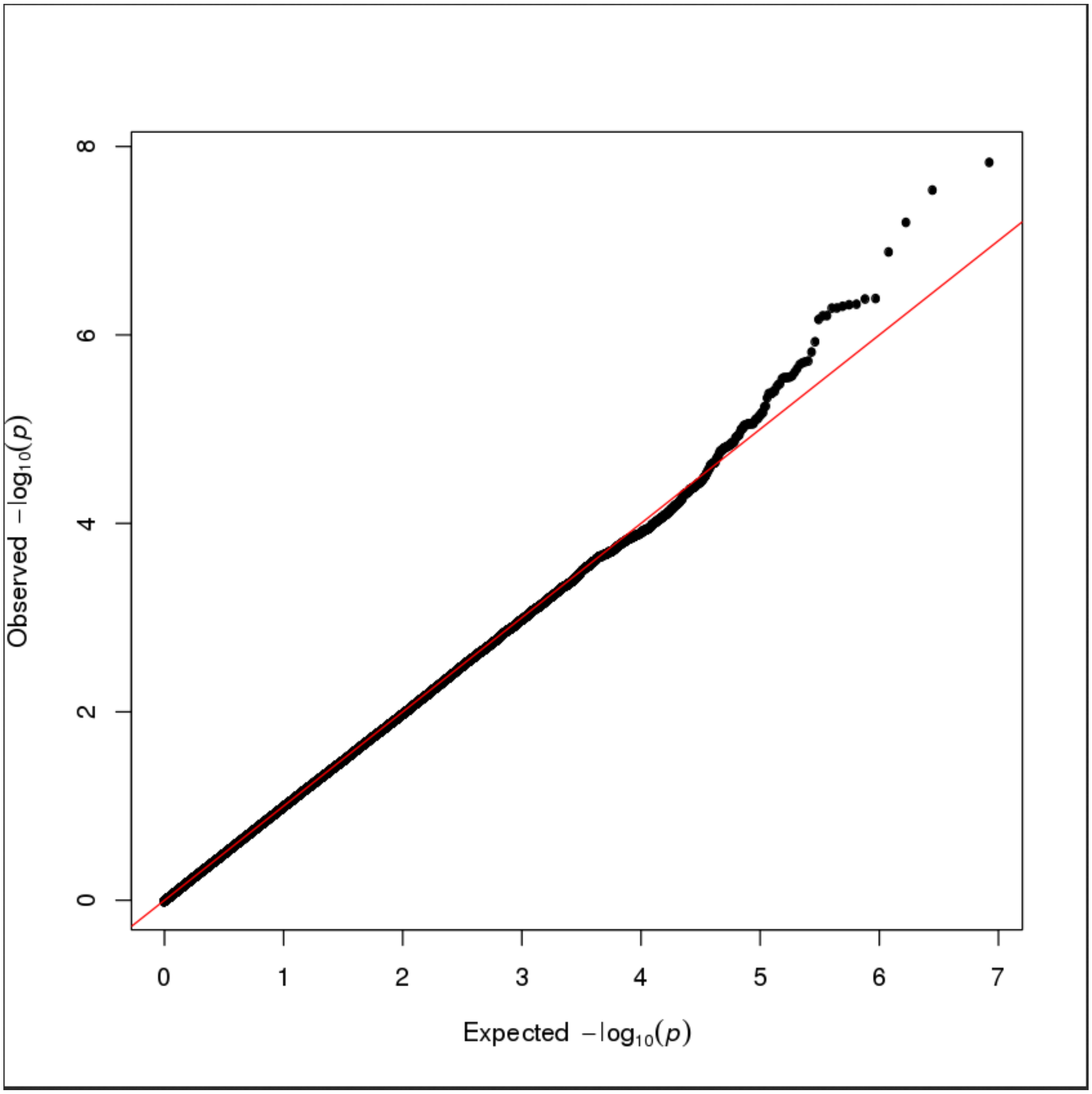


Figure 13. QQ plot of expected vs. observed p-values in GWAS of D using participants with D scores recorded during wave 4.

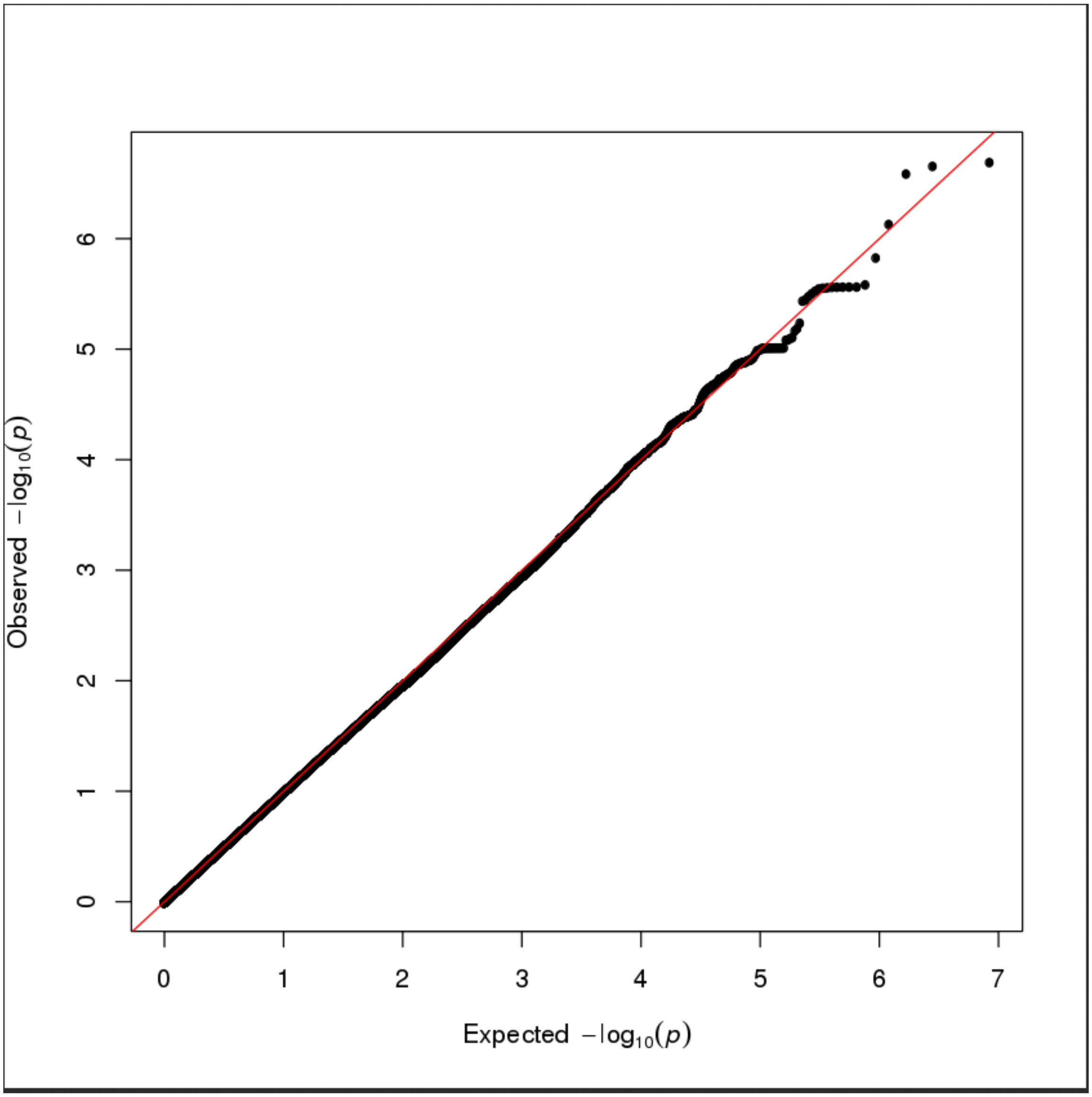


Figure 14. QQ plot of expected vs. observed p-values in GWAS of D using participants with D scores recorded during wave 5.

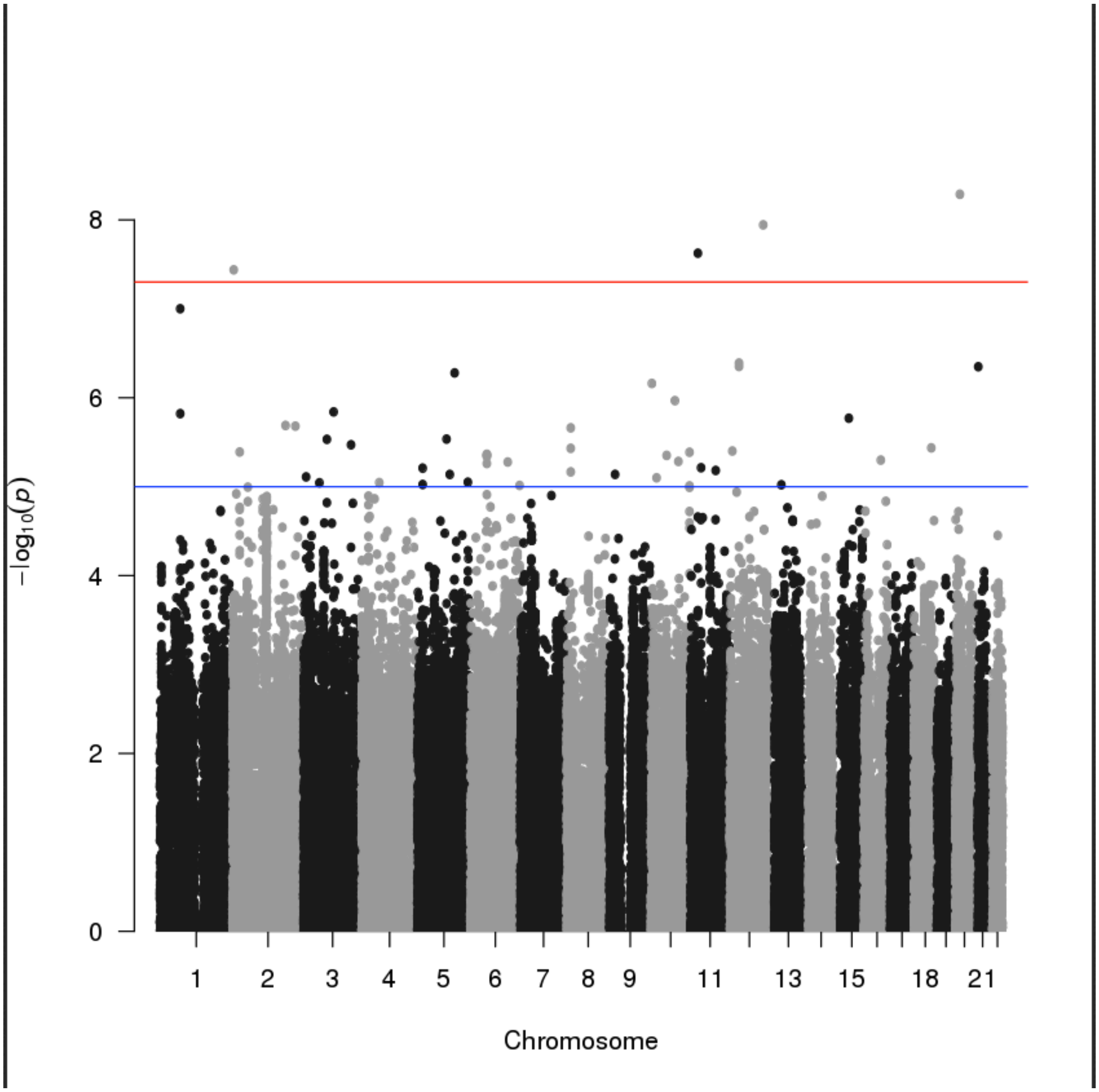


Figure 15. Manhattan plot of –log p-values across the 22 chromosomes in the GWAS of D using participants with D scores recorded during wave 1.

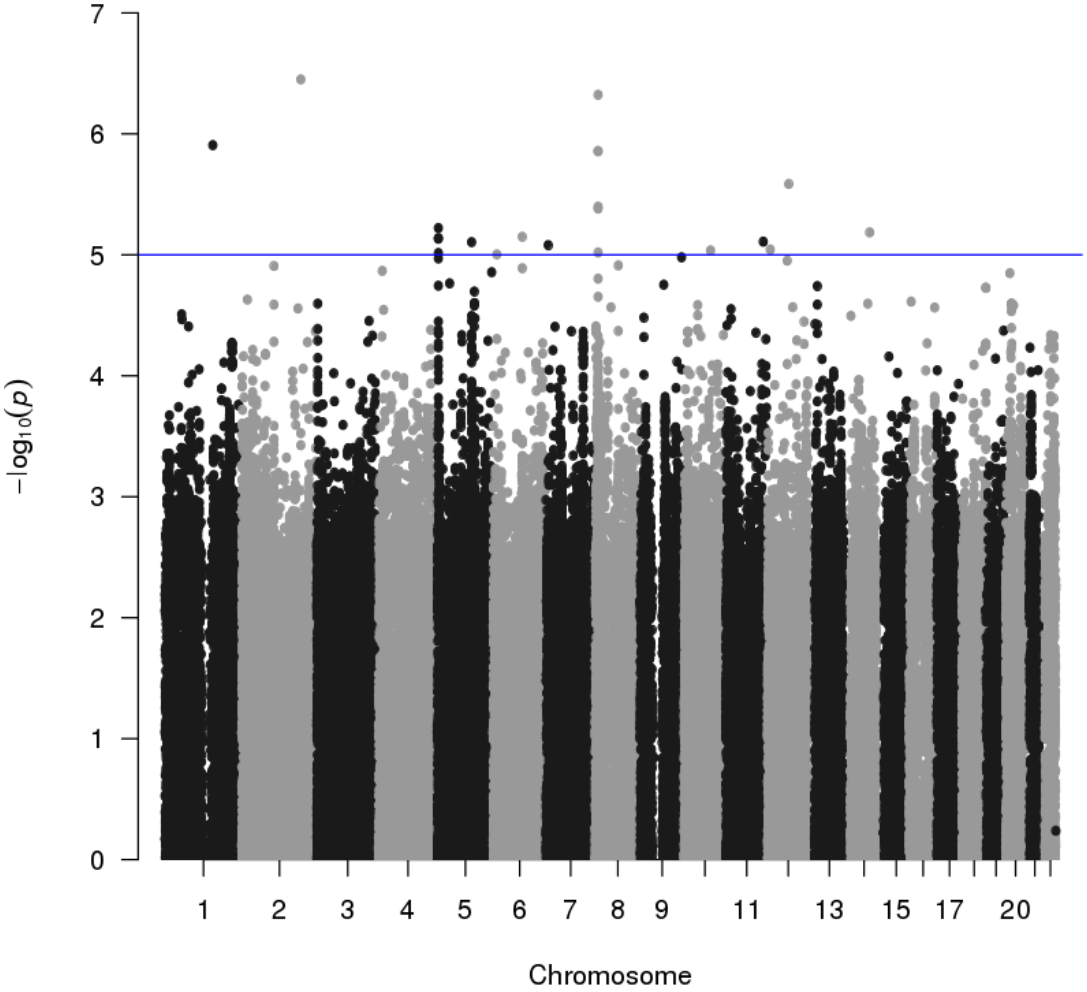


Figure 16. Manhattan plot of –log p-values across the 22 chromosomes in the GWAS of D using participants with D scores recorded during wave 2.

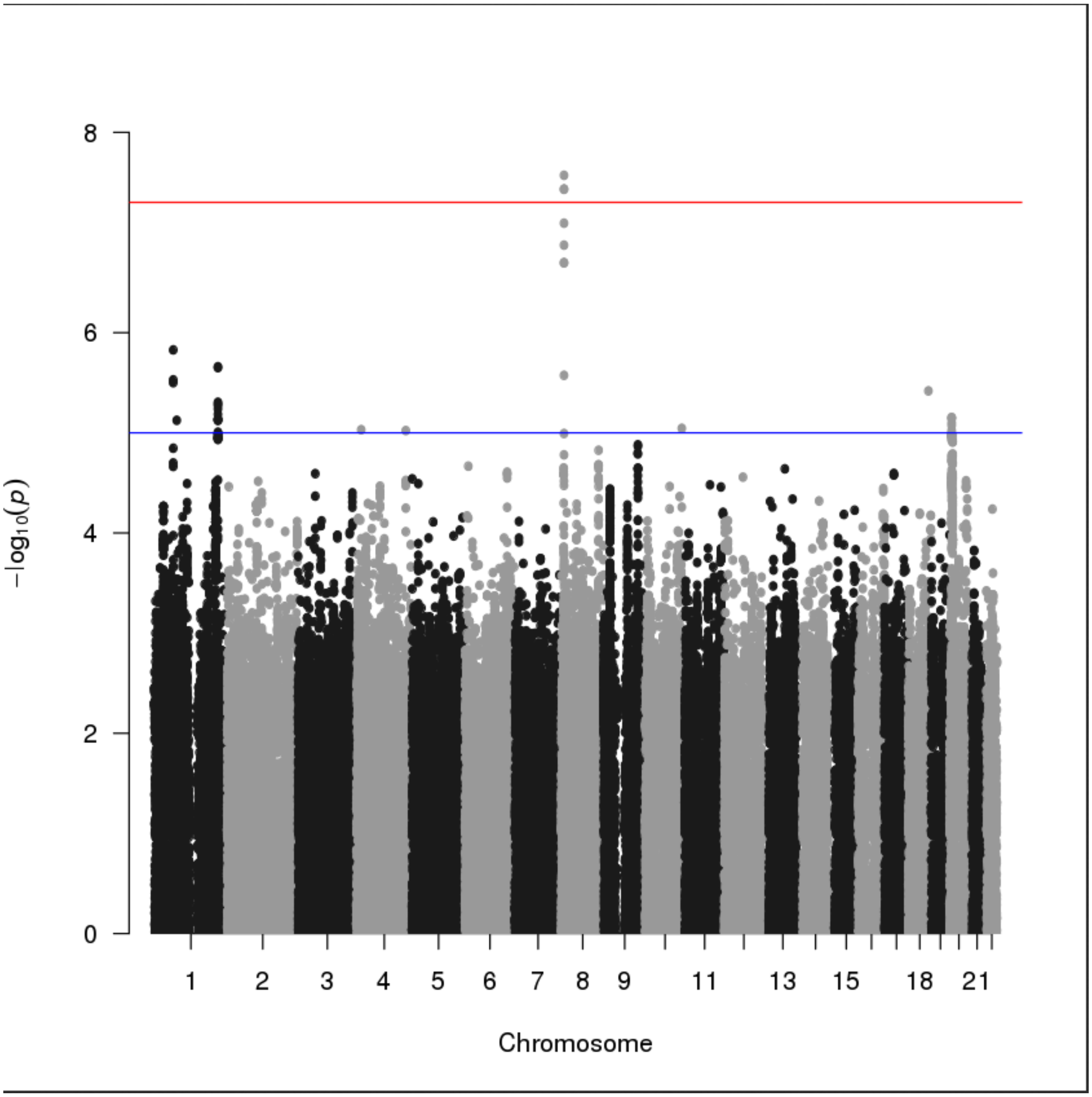


Figure 17. Manhattan plot of –log p-values across the 22 chromosomes in the GWAS of D using participants with D scores recorded during wave 3.

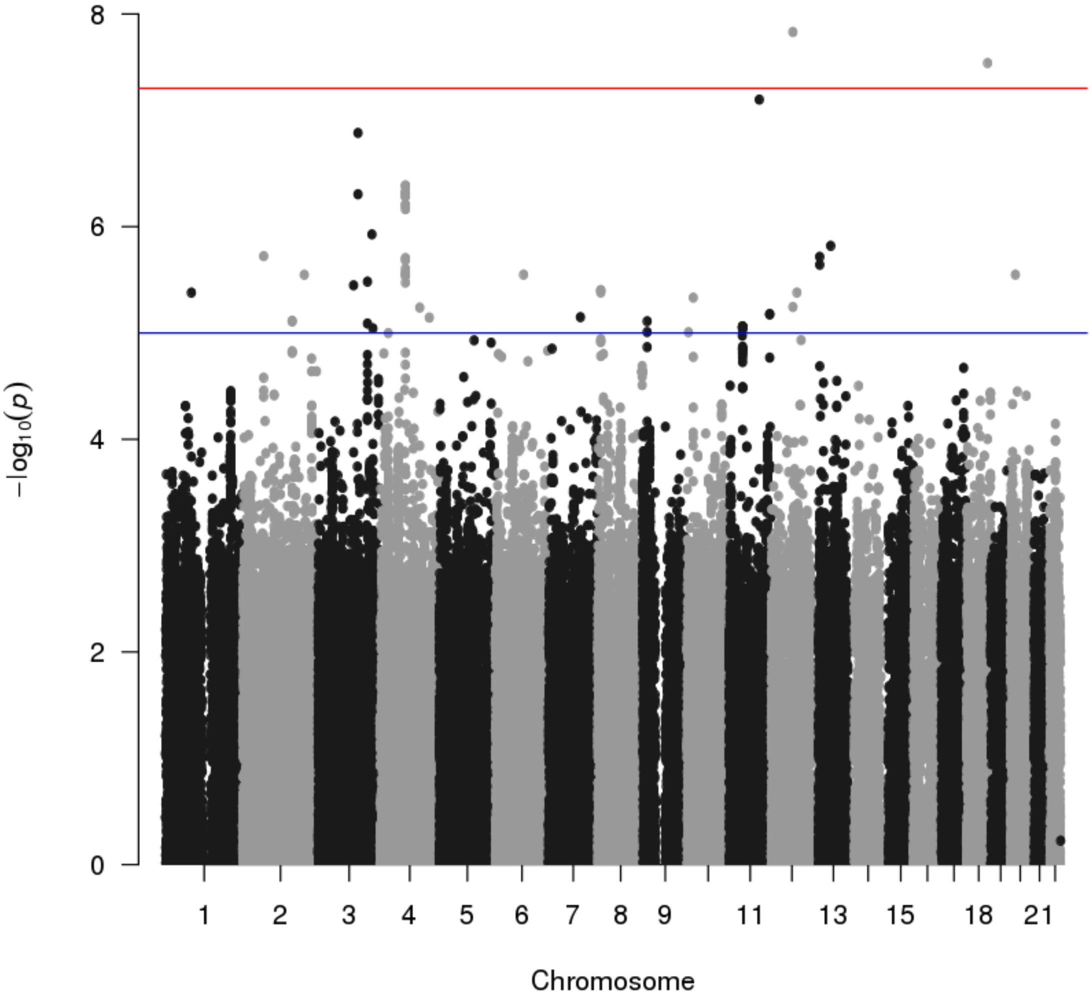


Figure 18. Manhattan plot of –log p-values across the 22 chromosomes in the GWAS of D using participants with D scores recorded during wave 4.

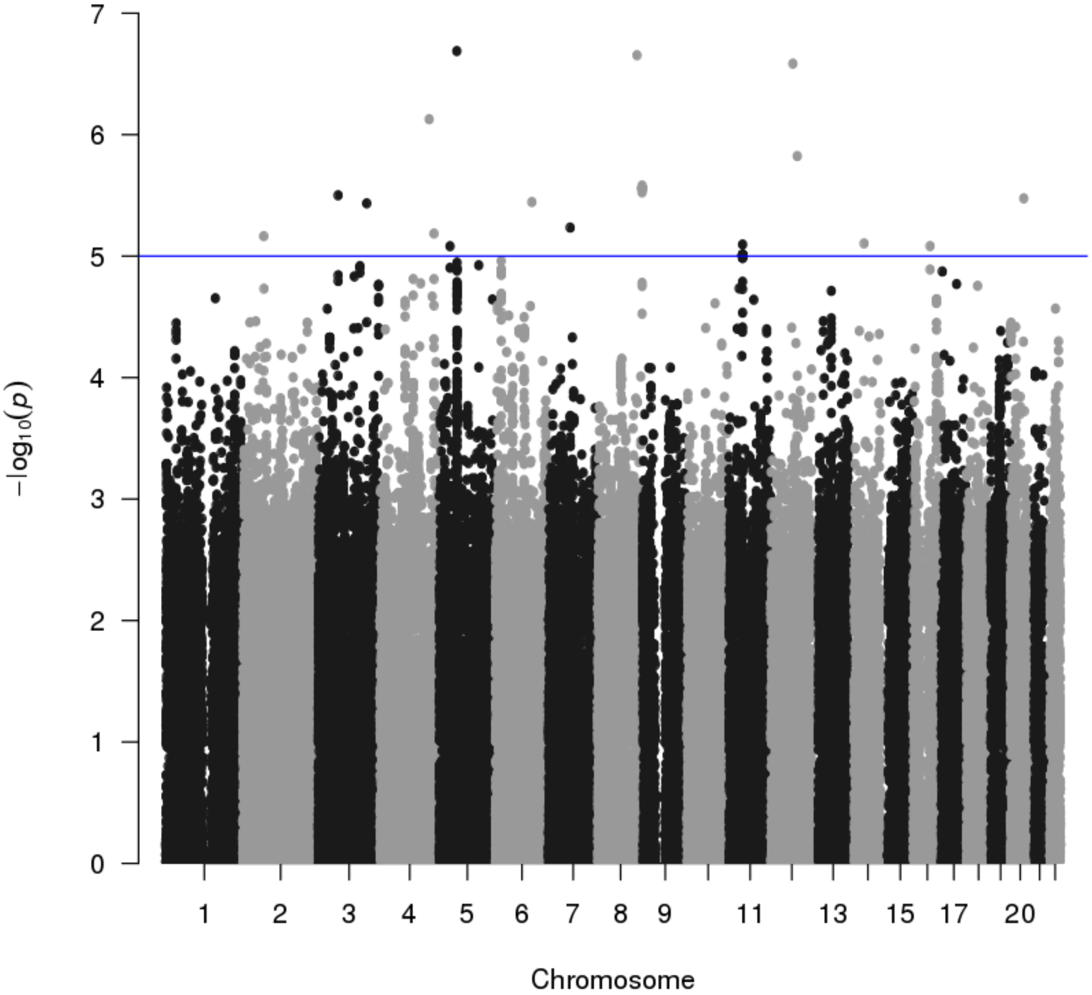


Figure 19. Manhattan plot of –log p-values across the 22 chromosomes in the GWAS of D using participants with D scores recorded during wave 5.

References

1. Arai, H. *et al.* Pathobiology of Alzheimer's disease and biomarker development. *Nippon Yakurigaku Zasshi* **135**, 3-7.
2. Voyle, N. *et al.* Genetic Risk as a Marker of Amyloid-beta and Tau Burden in Cerebrospinal Fluid. *J Alzheimers Dis* (2016).
3. Sunderland, T. *et al.* Clock drawing in Alzheimer's disease. A novel measure of dementia severity. *J Am Geriatr Soc* **37**, 725-9 (1989).
4. Rao, A.T., Degnan, A.J. & Levy, L.M. Genetics of Alzheimer disease. *AJNR Am J Neuroradiol* **35**, 457-8 (2014).
5. Pratico, D. Alzheimer's disease and the quest for its biological measures. *J Alzheimers Dis* **33**, S237-41 (2013).
6. Hall, A.M., Moore, R.Y., Lopez, O.L., Kuller, L. & Becker, J.T. Basal forebrain atrophy is a presymptomatic marker for Alzheimer's disease. *Alzheimers Dement* **4**, 271-9 (2008).
7. Sydykova, D. *et al.* Fiber connections between the cerebral cortex and the corpus callosum in Alzheimer's disease: a diffusion tensor imaging and voxel-based morphometry study. *Cereb Cortex* **17**, 2276-82 (2007).
8. Nordberg, A. *et al.* A European multicentre PET study of fibrillar amyloid in Alzheimer's disease. *Eur J Nucl Med Mol Imaging* **40**, 104-14 (2013).
9. Chau, D.M., Crump, C.J., Villa, J.C., Scheinberg, D.A. & Li, Y.M. Familial Alzheimer disease presenilin-1 mutations alter the active site conformation of gamma-secretase. *J Biol Chem* **287**, 17288-96 (2012).
10. Ridge, P.G., Mukherjee, S., Crane, P.K. & Kauwe, J.S. Alzheimer's disease: analyzing the missing heritability. *PLoS ONE* **8**, e79771 (2013).
11. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168-74 (2006).
12. Schmechel, D.E. *et al.* Increased amyloid beta-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late-onset Alzheimer disease. *Proc Natl Acad Sci U S A* **90**, 9649-53 (1993).
13. Bertram, L. *et al.* The LDLR locus in Alzheimer's disease: a family-based study and meta-analysis of case-control data. *Neurobiol Aging* **28**, 18 e1-4 (2007).
14. Poirier, J. Apolipoprotein E in the brain and its role in Alzheimer's disease. *Journal of Psychiatry & Neuroscience* **21**, 128-134 (1996).
15. Leoni, V. The effect of apolipoprotein E (ApoE) genotype on biomarkers of amyloidogenesis, tau pathology and neurodegeneration in Alzheimer's disease. *Clin Chem Lab Med* **49**, 375-83 (2011).
16. Raichlen, D.A. & Alexander, G.E. Exercise, APOE genotype, and the evolution of the human lifespan. *Trends Neurosci* **37**, 247-55 (2014).
17. Alizadeh, B.Z. *et al.* HFE variants, APOE and Alzheimer's disease: findings from the population-based Rotterdam study. *Neurobiol Aging* **30**, 330-2 (2009).
18. Tanzi, R.E. *et al.* The gene defects responsible for familial Alzheimer's disease. *Neurobiol Dis* **3**, 159-68 (1996).
19. Carrasquillo, M.M. *et al.* Replication of CLU, CR1, and PICALM associations with alzheimer disease. *Arch Neurol* **67**, 961-4 (2010).

20. Aronow, B.J., Lund, S.D., Brown, T.L., Harmony, J.A. & Witte, D.P. Apolipoprotein J expression at fluid-tissue interfaces: potential role in barrier cytoprotection. *Proc Natl Acad Sci U S A* **90**, 725-9 (1993).

21. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res* **42**, D749-55 (2014).

22. Klebig, M.L. *et al.* Mutations in the clathrin-assembly gene Picalm are responsible for the hematopoietic and iron metabolism abnormalities in fit1 mice. *Proc Natl Acad Sci U S A* **100**, 8360-5 (2003).

23. Corneveaux, J.J. *et al.* Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. *Hum Mol Genet* **19**, 3295-301 (2010).

24. Barral, S. *et al.* Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory. *Neurology* **78**, 1464-71 (2012).

25. Lambert, J.C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* **41**, 1094-9 (2009).

26. Grupe, A. *et al.* Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Hum Mol Genet* **16**, 865-73 (2007).

27. Coon, K.D. *et al.* A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry* **68**, 613-8 (2007).

28. Reiman, E.M. *et al.* GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron* **54**, 713-20 (2007).

29. Li, H. *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* **65**, 45-53 (2008).

30. Beecham, G.W. *et al.* Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. *Am J Hum Genet* **84**, 35-43 (2009).

31. Carrasquillo, M.M. *et al.* Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet* **41**, 192-8 (2009).

32. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-93 (2009).

33. Li, M., Boehnke, M. & Abecasis, G.R. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet* **78**, 778-92 (2006).

34. Howson, J.M., Barratt, B.J., Todd, J.A. & Cordell, H.J. Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci. *Genet Epidemiol* **29**, 51-67 (2005).

35. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).

36. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics* **37**, 549-554 (2005).

37. Hong, H.X. *et al.* Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples. *Bmc Bioinformatics* **9**, - (2008).

38. Consortium, W.T.C.C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).

39. Cox, D.G. & Kraft, P. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum Hered* **61**, 10-4 (2006).

40. Wittke-Thompson, J.K., Pluzhnikov, A. & Cox, N.J. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 967-86 (2005).

41. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

42. Ellinghaus, D., Schreiber, S., Franke, A. & Nothnagel, M. Current software for genotype imputation. *Hum Genomics* **3**, 371-80 (2009).

43. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).

44. Ryan, S.G. Human genetic variation. *Pharmacogenomics.* **3**, 9-11 (2002).

45. Browning, S.R. & Browning, B.L. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am J Hum Genet* **97**, 404-18 (2015).

46. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis Gç, R. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genet Epidemiol* **34**, 816-34 (2010).

47. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).

48. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

49. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).

50. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).

51. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic Epidemiology* **34**, 816-834 (2010).

52. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* **81**, 559-575 (2007).

53. John, B. & Lewis, K.R. Chromosome variability and geographic distribution in insects. *Science* **152**, 711-21 (1966).

54. Gottesman, II & Gould, T.D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**, 636-45 (2003).

55. Brown, M.S. & Goldstein, J.L. A receptor-mediated pathway for cholesterol homeostasis. *Science* **232**, 34-47 (1986).

56. den Heijer, T. *et al.* A 10-year follow-up of hippocampal volume on magnetic resonance imaging in early dementia and cognitive decline. *Brain* **133**, 1163-72.

57. Murray, A.D. *et al.* The balance between cognitive reserve and brain imaging biomarkers of cerebrovascular and Alzheimer's diseases. *Brain* **134**, 3687-96 (2011).

58. Melville, S.A. *et al.* Multiple loci influencing hippocampal degeneration identified by genome scan. *Ann Neurol* **72**, 65-75 (2012).

59. Nagy, Z. *et al.* Relative roles of plaques and tangles in the dementia of Alzheimer's disease: correlations using three sets of neuropathological criteria. *Dementia* **6**, 21-31 (1995).

60. Shaw, L.M. *et al.* Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol* **65**, 403-13 (2009).

61. Blennow, K. Cerebrospinal fluid protein biomarkers for Alzheimer's disease. *NeuroRx* **1**, 213-25 (2004).

62. Trojanowski, J.Q. *et al.* Update on the biomarker core of the Alzheimer's Disease Neuroimaging Initiative subjects. *Alzheimers Dement* **6**, 230-8 (2010).

63. Kim, S. *et al.* Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PLoS ONE* **8**, e70269 (2013).

64. O'Bryant, S.E. *et al.* A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI. *PLoS ONE* **6**, e28092 (2011).

65. O'Bryant, S.E. *et al.* A serum protein-based algorithm for the detection of Alzheimer disease. *Arch Neurol* **67**, 1077-81 (2010).

66. O'Bryant, S.E. *et al.* A blood-based algorithm for the detection of Alzheimer's disease. *Dement Geriatr Cogn Disord* **32**, 55-62 (2011).

67. Mackman, N., Tilley, R.E. & Key, N.S. Role of the extrinsic pathway of blood coagulation in hemostasis and thrombosis. *Arteriosclerosis, Thrombosis, and Vascular Biology* **27**, 1687-1693 (2007).

68. Broderick, J. *et al.* Guidelines for the management of spontaneous intracerebral hemorrhage in adults: 2007 Update. Guideline from the American Heart Association/American Stroke Association Stroke Council, high blood pressure research council, and the quality of care and outcomes in research interdisciplinary working group. *Stroke* **38**, 2001-2023 (2007).

69. Drenos, F. *et al.* Integrated associations of genotypes with multiple blood biomarkers linked to coronary heart disease risk. *Hum Mol Genet* **18**, 2305-16 (2009).

70. Fujimaki, T. *et al.* Association of genetic variants with myocardial infarction in Japanese individuals with chronic kidney disease. *Thromb Haemost* **101**, 963-8 (2009).

71. Ken-Dror, G. *et al.* Haplotype and genotype effects of the F7 gene on circulating factor VII, coagulation activation markers and incident coronary heart disease in UK men. *J Thromb Haemost* **8**, 2394-403 (2010).

72. Zakai, N.A. *et al.* Association of coagulation-related and inflammation-related genes and factor VIIc levels with stroke: the Cardiovascular Health Study. *J Thromb Haemost* **9**, 267-74 (2011).

73. Zee, R.Y. *et al.* An evaluation of candidate genes of inflammation and thrombosis in relation to the risk of venous thromboembolism: The Women's Genome Health Study. *Circ Cardiovasc Genet* **2**, 57-62 (2009).

74. Plato & Jowett, B. *Plato's the republic*, (The Modern library, New York, 1941).

75. Spearman, C. " General Intelligence," objectively determined and measured. *The American Journal of Psychology* **15**, 201-292 (1904).

76. Royall, D.R., Palmer, R.F. & O'Bryant, S.E. Validation of a latent variable representing the dementing process. *J Alzheimers Dis* **30**, 639-49 (2012).

77. Bouchard, T.J., Jr. Genetic influence on human intelligence (Spearman's g): how much? *Ann Hum Biol* **36**, 527-44 (2009).

78. Royall, D.R. & Palmer, R.F. Getting Past "g": testing a new model of dementing processes in persons without dementia. *J Neuropsychiatry Clin Neurosci* **24**, 37-46 (2012).

79. Barber, R.C. *et al.* Can Genetic Analysis of Putative Blood Alzheimer's Disease Biomarkers Lead to Identification of Susceptibility Loci? *PLoS One* **10**, e0142360 (2015).

80. Waring S, O.B.S., Reisch JS, Diaz-Arrastia R, Knebl J, Doody R, et al. The Texas alzheimer's Research Consortium longitudinal research cohort: Study design and baseline characteristics. *Texas Public Health Journal* **60**, 9-13 (2008).

81. Petersen, R.C. Mild cognitive impairment as a diagnostic entity. *J Intern Med* **256**, 183-94 (2004).

82. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-44 (1984).

83. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-60 (2008).

84. Tang, M.X. *et al.* The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA* **279**, 751-5 (1998).

85. Maestre, G. *et al.* Apolipoprotein E and Alzheimer's disease: ethnic variation in genotypic risks. *Ann Neurol* **37**, 254-9 (1995).

86. Royall, D.R. & Palmer, R.F. Ethnicity moderates dementia's biomarkers. *J Alzheimers Dis* **43**, 275-87 (2015).

87. Royall, D.R., Palmer, R.F., Vidoni, E.D., Honea, R.A. & Burns, J.M. The Default Mode Network and Related Right Hemisphere Structures may be the Key Substrates of Dementia. *J Alzheimers Dis* (2012).

88. Greicius, M.D., Srivastava, G., Reiss, A.L. & Menon, V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc Natl Acad Sci U S A* **101**, 4637-42 (2004).

89. Gressner, O.A. *et al.* Questioning the role of actinfree Gc-Globulin as actin scavenger in neurodegenerative central nervous system disease: relationship to S-100B levels and blood-brain barrier function. *Clin Chim Acta* **400**, 86-90 (2009).

90. Lee, W.M. & Galbraith, R.M. The extracellular actin-scavenger system and actin toxicity. *N Engl J Med* **326**, 1335-41 (1992).

91. Haddad, J.G., Harper, K.D., Guoth, M., Pietra, G.G. & Sanger, J.W. Angiopathic consequences of saturating the plasma scavenger system for actin. *Proc Natl Acad Sci U S A* **87**, 1381-5 (1990).

92. Bishnoi, R.J., Palmer, R.F. & Royall, D.R. Vitamin D binding protein as a serum biomarker of Alzheimer's disease. *J Alzheimers Dis* **43**, 37-45 (2015).

93. Breteler, M.M.B., Bots, M.L., Ott, A. & Hofman, A. Risk factors for vascular disease and dementia. *Haemostasis* **28**, 167-173 (1998).

94. Polidori, M.C., Pientka, L. & Mecocci, P. A Review of the Major Vascular Risk Factors Related to Alzheimer's Disease. *J Alzheimers Dis* (2012).

95. Kim, I. *et al.* A Relationship Between Alzheimer's Disease and Type 2 Diabetes Mellitus Through the Measurement of Serum Amyloid-beta Autoantibodies. *Journal of Alzheimers Disease* **19**, 1371-1376 (2010).

96. Terry, R.D. Alzheimer's disease and the aging brain. *J Geriatr Psychiatry Neurol* **19**, 125-8 (2006).

97. Carithers, L.J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* **13**, 311-9 (2015).

98. Raffaitin, C. *et al.* Metabolic syndrome and risk for incident Alzheimer's disease or vascular dementia: the Three-City Study. *Diabetes Care* **32**, 169-74 (2009).