A CASE-CASE GENOME-WIDE ASSOCIATION STUDY OF TRISOMY 21

by

Praewpannarai Buddadhumaruk

A.S. in Nursing, Community College of Allegheny County, 2006

B.S. in Biology, Carlow University, 2010

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Praewpannarai Buddadhumaruk

It was defended on

November 30, 2012

and approved by

Thesis Advisor:
Eleanor Feingold, Ph.D., Professor
Department of Human Genetics
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Roslyn A. Stone, Ph.D., Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
John R. Shaffer, Ph.D., Research Assistant Professor
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

A CASE-CASE GENOME-WIDE ASSOCIATION STUDY OF TRISOMY 21

Praewpannarai Buddadhumaruk, **M.S.**

University of Pittsburgh, 2012

The Centers for Disease Control and Prevention (CDC) estimates that about 6,000 babies are born with Down syndrome (DS) each year in the United States—that is one out of every 691 babies, making it the most common birth defect in humans. It is a genetic disorder that results from being born with an extra chromosome 21 (trisomy 21). In addition to mental retardation and abnormal physical development, DS individuals can also suffer from physiological abnormalities such as heart defects and leukemia. They experience Alzheimer-like symptoms, accelerated aging, and their average age at death is at 49 years old. The majority of trisomy 21 cases occurred due to maternal meiotic nondisjunction (NDJ). There are two types of NDJ: M1 NDJ and M2 NDJ, nondisjunction in the meiosis I and meiosis II stages respectively. The two types differ in their etiology: maternal age, recombination pattern, and environmental factors. This thesis is a pilot study using genome-wide association analysis to locate genes that may be responsible for the meiotic error. This study has public health significance because it identifies potential genes associated with DS susceptibility. It will help enable us to tease apart the susceptibility of NDJ into its individual liability factors. Genotype data from 134 Caucasian mothers of maternal NDJ cases were examined using the Cochran-Armitage test for trend. Two genes were identified to potentially be associated with the difference in NDJ error: SNX13 and PRPS1L1. The data were also used to test known genes responsible for altered recombination and Alzheimer's disease and we found suggestive associations between NDJ and RNF212, a

gene thought to be associated with meiotic recombination rate, and between NDJ and PSEN1, an Alzheimer's disease-related gene.

TABLE OF CONTENTS

PREFACE ..... **X**

1.0 INTRODUCTION **1**

1.1 TRISOMY 21 ..... **1**

1.2 GENOME-WIDE ASSOCIATION STUDY **4**

1.2.1 Case-Case Genome-Wide Association Study **7**

1.3 THESIS PROJECT **8**

1.3.1 Overview **8**

1.3.2 Objective **9**

2.0 METHODS AND RESULTS **10**

2.1 STUDY SAMPLE **10**

2.1.1 Genome-Wide Genotyping **10**

2.1.2 Determining Meiotic Nondisjunction Error 'Phenotype' **11**

2.2 DATA MANAGEMENT **12**

2.2.1 Starting Files **12**

2.2.2 PLINK **14**

2.2.3 Binary PED File **14**

2.2.4 Merging Genotype and Phenotype Data **14**

2.2.5 Per-Individual Quality Control **16**

2.2.5.1 Sex Check 16
2.2.5.2 Identification of Duplicated or Related Individuals 17
2.2.6 Per-Marker Quality Control 18
2.3 TEST FOR ASSOCIATION 20
2.3.1 Background 20
2.3.2 Association Analysis 24
2.3.3 Multiple Testing 25
2.3.4 Manhattan Plot 26
2.3.5 Quantile-Quantile Plot 28
2.3.6 LocusZoom 29
2.3.6.1 Scanning for Potential Gene Candidates 29
2.3.6.2 Scanning for Association with Candidate Down Syndrome Genes 31
3.0 DISCUSSION 35
3.1 SAMPLE SIZE AND DATA QUALITY 35
3.2 SIGNIFICANT GENES 37
3.2.1 Genome-Wide Exploration 37
3.2.2 Recombination and Alzheimer ’s Disease Genes 39
BIBLIOGRAPHY 42

LIST OF TABLES

Table 1. Output after executing the missing command shows inconsistent recordings of Individual ID 15

Table 2. Output after running sex discordance check shows incorrect recordings of sex 17

Table 3. Output after computing within-family IBS shows a duplicate sample recorded under a different ID 18

Table 4. Distribution of meiosis I and meiosis II NDJ cases after quality control implementation 20

Table 5. Summary of output after implementing per-marker QC 20

Table 6. 2x3 genotype-based table 23

Table 7. Output from the Cochran-Armitage test for trend 25

LIST OF FIGURES

Figure 1. Manhattan plot of -log (P values) using Haploview..... 27

Figure 2. QQ plot of -log10 (P value) using R 28

Figure 3. Regional plot of area surrounding rs1404414 of -log (P values) using LocusZoom 30

Figure 4. Regional plot of area surrounding PRPS1L1 of -log (P values) using LocusZoom 31

Figure 5. Regional plot of area surrounding RNF212 of -log (P values) using LocusZoom 32

Figure 6. Regional plot of area surrounding PRDM9 of -log (P values) using LocusZoom..... 32

Figure 7. Regional plot of area surrounding APOE of -log (P values) using LocusZoom..... 33

Figure 8. Regional plot of area surrounding PSEN1 of -log (P values) using LocusZoom 33

Figure 9. Regional plot of area surrounding MAPT of -log (P values) using LocusZoom 34

PREFACE

This thesis is written assuming the reader has taken at least a college level introductory biology course and understand the basics of cell division genetic biology.

I would like to give my sincere thanks to Dr. Eleanor Feingold for giving me this project to work on and have been patiently providing me guidance in completing this thesis. I would like to thank Dr. Roslyn Stone and Dr. Lisa Weissfeld for all the help they gave to make my graduation possible this December.

I would like to thank my family for their support—my mom, my dad, and my brother, Pete—and my dearest friend and close confident, Kevin.

This thesis is dedicated to the loving memory of Mickey.

1.0 INTRODUCTION

1.1 TRISOMY 21

Down syndrome (DS) is the most common birth defect in humans (National Association for Down Syndrome, 2012). The Centers for Disease Control and Prevention (CDC) estimates that about 6,000 babies are born with the disorder each year in the United States (Parker et al, 2010). In other words, the disorder occurs in 1 out of every 691 babies. Signs typically seen include mental retardation, hypotonia, and abnormalities of the face, hands, and feet. Individuals with DS also often have congenital heart defects, and are prone to have hearing and vision problems, weakness in the neck joints, and development of thyroid disease and leukemia. Their aging process seems to be accelerated with the average age of death at 49 (Liptak 2008). Heart disease and leukemia account for most deaths. However, it has been discovered that they develop Alzheimer-like lesions in their brains and show signs of memory loss, further lowering of intellect, and personality changes early in age--as early as 30 years old (Kolata, 1985; Liptak, 2008).

Development of DS results when an individual is born with three 21 chromosomes, termed trisomy 21, instead of having only the usual two 21 chromosomes (one inherited from the mother and the other inherited from the father). Trisomy is a type of aneuploidy (having an abnormal number of chromosomes) in which there are three copies of a particular chromosome.

There can be trisomy of chromosome 16, trisomy of chromosome 18, etc. However, trisomies often result in miscarriage. Trisomy 21 is one of the few trisomies in which the babies survive to term, even after 80% of fetuses with trisomy 21 are spontaneously aborted (Hook et al., 1995). 95% of trisomy 21 cases occur due to meiotic nondisjunction (NDJ), and 90% of NDJ error happens in the mother (Sherman et al., 2005). The most significant risk factor for NDJ of chromosome 21 is the age of the mother at the time of conception (Sherman et al., 2005). As a woman ages, the risk for conceiving a child with trisomy 21 increases. The other significant risk factor that is molecularly related is altered meiotic recombination patterns (Sherman et al., 2005).

There are two types of NDJ: M1 NDJ and M2 NDJ, nondisjunction in the meiosis I and meiosis II stages respectively. In M1 NDJ, the paired heterozygous 21 chromosomes fail to separate during the first meiosis stage, while in M2 NDJ, the homologous sister chromatids of one 21 chromosome fail to separate during the second meiosis stage. The etiology for each type of NDJ seems different however. Preliminary data from the Atlanta Down Syndrome Project suggests that the effect of age on NDJ vary between the two types. There is an increasing risk for M2 NDJ when shifted to older maternal ages compared M1 NDJ though this difference is not statistically significant (Sherman et al., 2005). The two types of NDJ have also shown to have different altered recombination pattern: M1 NDJ is associated with lack of exchange or with a single exchange close to the telomere, whereas exchanges occurring very close to the centromere (pericentromeric exchange) increase the risk for M2 NDJ (Lamb et al., 1996, 1997, 2005).

The next questions that could be asked are whether there is a genetic component to maternal meiotic NDJ in trisomy 21 and whether that genetic component is associated in any way with the previously mentioned two risk factors. A disorder is likely genetically influenced if the disorder recurs among genetically related individuals. However, discerning whether there is a

pattern of recurrence risk for NDJ among siblings is complicated by several situations: the survival of the aneuploid fetus, small sibship size among families, and the age when the mother was pregnant with each of the children. Nonetheless, there are case studies reporting the recurrence of trisomy 21 among siblings (Der Kaloustian et al., 1987; Neilsen et al., 1988; Al Awadi et al., 1999).

And so what genes might increase the risk for trisomy 21? The process of cell division requires an interworking among a complex variety of structural cell components, enzymes, and regulatory proteins--all of which are coded by the DNA. Any gene for which variation in the coding results in an erroneously functioning protein in the meiotic machinery, is a clear candidate. With regards to recombination, genes that are associated with recombination frequency or pattern are also potential candidates. There has been some identification of such genes. Kong et al. (2008) identified the gene RNF212 (ring finger protein) from a genome-wide scan for variants associated with recombination rate. Gene ortholog predictions suggest that the mammalian RNF212 is similar to the ZHP-3 gene in *Caenorhabditis elegans* (roundworm) and it is essential for successful recombination between homologous chromosomes (Jantsch et al., 2004). Parvanov et al. (2010), Baudat et al. (2010), and Myers et al. (2010) found that PRDM9 (PR domain containing 9) gene controls the extent to which crossovers occur in preferred chromosomal locations (known as “hotspots”). Could any of these recombination-related genes be responsible for the NDJ in trisomy 21?

Another group of genes to consider are genes related to the putative accelerated aging process among mothers of DS individuals. It has been repeatedly postulated over the years that it is the ‘biological age,” not the chronological age, of the mother that is the risk factor for aneuploidy in offspring (Emanuel et al., 1972; Brook et al, 1984). This hypothesis would help

explain the birth of DS babies among young mothers. Evidence supporting this thought includes the high prevalence of grey hair in young mothers of DS children (Emanuel et al., 1972) and their fivefold increased risk for Alzheimer's disease (AD) (Schupf et al., 1994, 2001). Because of this relatedness between AD and DS, researchers are reviewing genes responsible for AD among mothers of DS children. Avramopoulos et al. (1996) examined the distribution of apolipoprotein E (APOE) and found that the frequency of the APOE allele that is susceptible for AD among young mothers with M2 error is 30.0%, which is significantly higher than the frequency among older mothers with M2 error (13.0%, p-value=0.03). A rare form of early onset AD is caused by a mutation in the presenilin-1 (PSEN1) gene (Sherrington et al., 1995). This gene codes for the presenilin protein which is localize in the nuclear membrane, kinetochores, and centrosomes suggesting that this protein may play a role in chromosome organization and segregation (Li et al., 1997). It has also been found that M2 mothers has increased frequency of the AD susceptible allele of PSEN1 (70.8%) in comparison to 52.7% in M1 mothers (p-value < 0.01) (Petersen et al., 2000). Another AD gene worth considering is the MAPT (microtubule-associated protein tau) gene, which codes for the microtubule-associated protein tau that forms the hallmark neurofibrillary tangles in AD histology. The MAPT gene is located in a genomic inversion region that has been shown to be involved in recombination rate (Zody et al., 2008; Fledel-Alon et al., 2011).

1.2 GENOME-WIDE ASSOCIATION STUDY

A genome-wide association study (GWAS or GWA study) is an approach that involves rapidly scanning genetic markers, usually single-nucleotide polymorphisms (SNPs), across complete sets

of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Such study is useful in finding potential genes that contribute to common but complex diseases such as cancer, diabetes, heart disease and mental illnesses. This type of study became possible after the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005. The projects accomplished in mapping the entire human genome and determined the 3 billion basepair sequences of human DNA identifying approximately 20,000-25,000 genes (U.S. Department of Energy Genome Program, 2012; The International HapMap Consortium, 2003).

According to the National Human Genome Research Institute (2011), there are about 10 million common SNPs in the human genome. Scanning this large number of SNPs for the entire genome in a large sample of study subjects would be time-and-resource intensive. Fortunately, there is a shortcut that can reduce the workload by 30-fold. Upon the completion of the International HapMap Project, it was found that those 10 million SNPs form into clusters (called "haplotypes"). Because of this clustering, a few SNPs can be selected to represent each cluster and the entire genome can be well-represented by simply using as low as 300,000 chosen SNPs (The International HapMap Consortium, 2003).

The most frequently used GWA study design is case-control. The frequencies of each genotype (aa, Aa, and AA) of the tested SNP among subjects with the disease of interest are compared to the frequencies in those disease-free. The comparison is made for each SNP. Under the null hypothesis of no association with the disease, the ratio is similar within each genotype group. However, if a particular genotype of the SNP is more frequent in the disease than the control group, then that SNP is said to be "associated" with the disease.

GWA study is very different from other type of applied statistical research studies. Typically statistical studies have an a priori hypothesis being tested and if the topic is genetically related, only one or a few genetic regions are being tested. In contrast, GWA study usually does not have pre-specified genes to be tested and a large number of statistical tests are being done (one for each SNP). Thus GWA study is considered a hypothesis generating study or exploratory study. It is not used to established causal relationship. Rather, it identifies potential genetic candidates out of the entire genome for future investigation.

Pearson and Manolio (2008) outlined the parts of a typical GWA study: (1) selecting a large number of individuals with the outward characteristic (phenotype) of interest and a suitable comparison group; (2) isolating DNA from these individuals, running the genetic assays (genotyping), and reviewing the obtained data to ensure high genotyping quality; (3) computing the statistical tests for associations between the phenotype and SNPs that has passed set quality control thresholds; and (4) replicating putative associations in an independent population sample or performing an experiment that would test biologically test for potential causuality.

A case-control design is often easier and less expensive to conduct than designs such as a prospective cohort study. However, this design also carries the most assumptions, which if not met, can lead to substantial biases and spurious associations. The assumptions are: (1) case and control subjects are drawn from the same population; (2) the sampled case subjects are representative of all cases of the disease, else the limitations on diagnostic specificity and representativeness are clearly specified; (3) genomic and epidemiologic data are collected similarly in cases and controls; and (4) the differences in allele frequencies is due to the outcome of interest, not the difference in background between cases and controls (Pearson and Manolio

2008). It can clearly be seen from the above list of assumptions that proper sampling of the cases and controls are crucial to maintain the validity of the study results.

1.2.1 Case-Case Genome-Wide Association Study

The case and the control groups are somewhat arbitrary defined. As such, a novel way to make comparisons is to compare a group of one disease type against a group of another disease type, or do a "case-case comparison."

There are theoretical reasons why comparing between two groups of cases, rather than between a case and a control, has benefits. Some advantages are the elimination of selection biases and reduced background noise (Curtis et al., 2011). Typically, controls are recruited from a different source population than the cases. Cases often are derived from clinical settings whereas controls are obtained from a non-clinical setting. There may be unforeseen differences between them other than just the disease status. By using only the cases however, all subjects are recruited from the same clinical setting. The two case groups might be more similar in terms of their geographical background, social class, and factors influencing presentation to services. Having well-matched samples reduces background noise and it is expected to increase both the power and specificity of the association study.

By comparing the genome data between two sets of cases, we are able to pick up genetic markers that are associated with the differentiation between the two disease groups. Identifying genes that divides the disease from the control is medically crucial. However, it is sometimes important as well to identify what makes disease D type A different from disease D type B. There could be differences in etiology between the two. Perhaps one type is easily preventable, curable, or manageable than the other. But we will not be aware of such information if we do not

analyze the data in this manner. Of course, the caveat here is that the case-case analysis would not be able to identify disease-causing genetic components shared between the two types.

1.3 THESIS PROJECT

1.3.1 Overview

This thesis focuses performs genome-wide association analysis to locate genes that may influence the difference in maternal meiosis NDJ errors between M1 and M2. It is a case-case analysis, comparing among mothers of DS children. In addition to reducing the background noise from examining a more homogenous study sample, there has been compelling evidence that the etiology of M1 and M2 NDJ is distinct, and we wish to locate the genetic factors potentially responsible for these distinct etiologies. Though this case-case GWA study will efficiently pick up genes that are associated with the individual types of NDJ, it will miss any genes that are common to both NDJ types.

As previously mentioned, a GWA study has many parts and many specialized skills are required to complete the whole study. Only the data management and statistical analyses aspects are completed for the purpose of this thesis project and explicitly described. Sampling of study cases and processing of genetic assays were done by another party prior to the start of this project.

1.3.2 Objective

Risk factors for maternal NDJ have been identified in past studies to include advanced maternal age and altered recombination. While there have been identification of genes likely responsible for recombination, there has yet been any study linking between any particular recombination genes with NDJ and trisomy 21. Therefore this is a pilot study. It is also exploratory in nature due to the application of genome-wide association analysis. There are no pre-defined hypotheses to be tested but the direction of the analysis can be stated within two aims:

Aim1: Is the primary analysis. It involves scanning the entire genome of the study sample and identify for any possible genes that may be responsible for the separate etiologies of M1 and M2 NDJ.

Aim2: Utilizing the already analyzed data, we review the five candidate genes specifically discussed in section 1.1 above. The five genes candidates are RNF212, PRDM9, APOE, PSEN1, and MAPT. The first two genes (RNF212 and PRDM9) are recombination genes, while the next two (APOE and PSEN1) are Alzheimer's disease related genes. MAPT is a special case of gene that may be related to both recombination rate and Alzheimer's disease.

The importance of understanding the causes of NDJ cannot be over-stated. NDJ in general, is the leading cause of pregnancy loss and among, live births--it is the leading genetic cause of intellectual and developmental disabilities and birth defects. This thesis project will enable us to tease apart the susceptibility of human NDJ into its individual liability factors.

2.0 METHODS AND RESULTS

2.1 STUDY SAMPLE

2.1.1 Genome-Wide Genotyping

Genotyping was carried out on biological parents of Down syndrome children. This collection of biological assays was a part of HL092981 (Zwick PI) to identify copy-number variations (CNVs) that are associated with congenital heart defects among individuals with Down syndrome (Hedge et al., 2008). Families were identified through a birth defect surveillance system and included live born infants with documented trisomy 21. They were recruited by the Atlanta Down Syndrome Project (ADSP, birth years 1989-1999, ascertaining cases from the 5-county metropolitan Atlanta area) and the National Down Syndrome Project (NDSP, birth years 2000-2004, ascertaining cases from 6 national sites).

DNA samples were genotyped using the Affymetrix® Genome-Wide Human SNP 6.0 Array chips per Affymetrix standard protocol (Affymetrix, Santa Clara, California, USA). Initial quality control (QC) measures were implemented in the Zwick lab. For a sample to pass QC and hence be included in the downstream analysis, each sample had to have 86% call rate at the minimum, 0.4 contrast quality control and gender concordance, which are all Affymetrix recommended parameters. After this initial quality control, there were a total of 383 genotyped

samples. The Affymetrix® Genome-Wide Human SNP 6.0 Array contains 906,600 selected markers from the entire human genome. Genotype information of the parents was electronically stored in the text data file, *parents.ped*.

2.1.2 Determining Meiotic Nondisjunction Error 'Phenotype'

The parental source of the third chromosome 21 was determined by examining parental and child genotypes on chromosome 21.. This was accomplished using both STR and SNP genotyping methods. If a maternal origin was established, then the stage of meiotic error (meiosis I vs. meiosis II) was determined using genetic markers located in the pericentromeric region (13,615,252 bp - 16,784,299 bp) of 21q. If both maternal chromosomes had the same allele for all the aforementioned genetic markers (homozygous), then it was inferred that the error happened in meiosis II. If some of the alleles for one maternal chromosome were different than the other one (heterozygous), then the error was classified as meiosis I.

The parental source, stage of meiotic error and race of 187 mothers and 185 fathers (189 unique families) were stored on a Microsoft Excel spreadsheet named *Phenotype Variables.xlsx*. Note that the number of samples in the genotype data did not exactly match the number of parents recorded on the Excel spreadsheet.

2.2 DATA MANAGEMENT

Due to the massive nature of genetic data, it was not easy to simply view the data and check the genotype files. Thus this section explains the setup of data storage and how data management was accomplished.

2.2.1 Starting Files

The PED file, *parents.ped*, held the actual genotyped data and subject identification variables. Information was arranged in 383 lines (for 383 samples) and 1,813,206 columns. The type of information and order of the columns had to be arranged in a specific pattern as followed:

Column1 = Family ID (number ranges from 171002-9915166)

Column2 = Individual ID [either 10 or 20 (10=father, 20=mother)]

Column3 = Paternal ID

Column4 = Maternal ID

Column5 = Sex (1=male, 2=female)

Column6 = Phenotype [represented as 1, 2, or 0 (1=unaffected, 2=affected, 0=missing)]

Column7+8 = genotype pair at SNP1 (represented as a pair of bases, one column for each allele)

Column9+10 = genotype pair at SNP2

…

Column1813205+1813206 = genotype pair at SNP906600

The MAP file, *parents.map*, was a file to be used concurrently with *parents.ped*. It contained the chromosomal positions of each SNP that has been genotyped in the PED file and had following column arrangement:

Column1 = Chromosome#

Column2 = rs# (or SNP identifier)

Column3 = Genetic distance (in morgans)

Column4 = Physical base-pair position (in bp units)

The primary outcome for the analysis in this thesis project was the meiosis stage where chromosomal nondisjunction had occurred [meiosis I (M1) vs. meiosis II (M2)]. Thus meiosis stage was the phenotype of interest in this research. This information, however, was stored separately on the Microsoft Excel spreadsheet, *Phenotype Variables.xlsx*, and it needed to be extracted and linked to the genotype files before we ran any statistical analyses.

The subpopulation used for this project was limited to only Caucasian mothers of Down syndrome children whose third 21 chromosome had come from the mother. Family IDs, Individual IDs, and phenotype values of 143 Caucasian mothers were extracted from *Phenotype Variables.xlsx* and formatted in the manner below:

Column1 = Family ID (number ranges from 171002-9915166)

Column2 = Individual ID [either 10 or 20 (10=father, 20=mother)]

Column3 = Phenotype [either 1 or 2 (1=M1, 2=M2)]

The file was saved as *meiosis.txt* and served as the ‘alternative phenotype’ for the PED/MAP files.

2.2.2 PLINK

Because of the magnitude of DNA data, specialized software was needed for the extensive computation. PLINK is a free, open-source program designed to perform a range of basic, large-scale whole genome association analyses (Purcell et al., 2007). The program can be downloaded and instructions on how to execute data management-related and analytical tasks can be found at http://pngu.mgh.harvard.edu/~purcell/plink/. It was also the primary software package being used for this thesis research and it is the standard software for almost all GWA study analysis. PLINK is a command-line program written in C/C++ and it is executable via MS-DOS under the Windows platform. All commands involve typing plink at the command prompt. Output files are in a standard plain text format with various possibilities of file extension names depending on the executed command.

2.2.3 Binary PED File

While analyses could be directly done using the MAP/PED files, the duo files were instead converted into a binary PED file (*.bed) to further cut down computation time by using the following PLINK command:

```
plink --ped parents.ped.txt --map parents.map --make-bed
      --out parents
```

2.2.4 Merging Genotype and Phenotype Data

The new genotype BED file was then merged with the meiosis phenotype file. PLINK gave a preliminary descriptive statistics of the merged data below. PLINK reported that after matching

the IDs between the genotype BED file and alternative phenotype text file, there were only 118 matched participants. This number was lower than the expected count from the phenotype file.

```
plink --bfile parents --pheno meiosis.txt --out merge

Reading alternate phenotype from [ meiosis.txt ]
118 individuals with non-missing alternate phenotype
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
35 cases, 83 controls and 6 missing
Before frequency and genotyping pruning, there are 906600 SNPs
```

In order to figure out why the discrepancy, we executed a missingness query using the following command:

```
plink --bfile parents --pheno meiosis.txt --missing --out missing
```

A list of Family IDs and Individual IDs of individuals whose data were missing was generated. We noted that there was a mismatch in Individuals IDs. In the phenotype file, all the Individual IDs for mothers were coded 20. The output in **Table 1** shows that the genotype data have duplicate arrays for some individuals. The Individual IDs for them were not recorded as 20, but were recorded as a 20 with a tag that signified the duplicated sample.

Table 1. Output after executing the missing command shows inconsistent recordings of Individual ID

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
1710033	20	N	10583	906312	0.01168
1710034	10	Y	45406	906600	0.05008
1710034	20_dup1	Y	24808	906312	0.02737
1710034	20_dup2	Y	47716	906312	0.05265
1710038	10	Y	7988	906600	0.008811
1710038	20	N	12369	906312	0.01365
1710042	10_dup2	Y	55953	906600	0.06172
1710042	10_dup1	Y	28126	906600	0.03102

FID=Family ID, IID=Individual ID, MISS_PHENO=indicating whether phenotype is missing or not, N_MISS=number of missing SNPs, N_GENO=number of non-obligatory missing genotypes, F_MISS=proportion of missing SNPs

To solve this discrepancy, Individual IDs for the mothers in the outcome file were corrected by us to match with the Individual IDs in the genotype file. Only one genotype data was allowed per mother. Therefore, if there were duplicate arrays, we kept only the most recent

array in the data set. After this correction, a new phenotype file *meiosis2.txt* was created. The sample size was now at 137 mothers.

```
plink --bfile parents --pheno meiosis2.txt --out merge2

Reading alternate phenotype from [ meiosis2.txt ]
137 individuals with non-missing alternate phenotype
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
42 cases, 95 controls and 246 missing
Before frequency and genotyping pruning, there are 906600 SNPs
```

2.2.5 Per-Individual Quality Control

To reduce the potential for both false-positive and false-negative associations, quality control measures needed to be taken. Per-individual QC is recommended to be performed first before conducting QC on a 'per-marker' basis to maximize the number of markers remaining in the study (Anderson et al., 2010).

2.2.5.1 Sex Check

We checked the reported sex of the individuals against their actual genetic data makeup. This was done by examining the sex chromosomes. A female has two X chromosomes whereas a male has one X chromosome and one Y chromosome. There are many genes unique to the Y chromosome. However, a certain region on the Y chromosome called "the pseudo-autosomal region" contains genes that are common to both sex chromosomes. Therefore if an individual is male, then the genotyping of the Y chromosome area not in the pseudo-autosomal region would not be heterozygous (Anderson et al., 2010).

After running the sex-check command in PLINK, there were 3 individuals that are genetically males but were reported as females. Their IDs were 5510067.20, 7210113.20, and

9910657.20 as shown in **Table 2**. Note that a male call is made by PLINK if F is more than 0.8; a female call is made if F is less than 0.2 (Purcell et al., 2009).

```
plink --bfile parents --pheno meiosis2.txt --check-sex
      --out sexcheck
```

Table 2. Output after running sex discordance check shows incorrect recordings of sex

FID	IID	PEDSEX	SNPSEX	STATUS	F
4710041	10	1	1	OK	0.984
4710041	20	1	2	PROBLEM	0.0508
5510067	10	1	1	OK	0.9843
5510067	20	2	1	PROBLEM	0.9855
7210113	20	2	1	PROBLEM	0.9848
7210159	10	1	2	PROBLEM	0.104
7210159	20	2	2	OK	0.1055
9910657	10	1	1	OK	0.986
9910657	20_dup3	2	1	PROBLEM	0.9857
9910657	20_dup1	2	1	PROBLEM	0.9857
9910657	20_dup2	2	1	PROBLEM	0.9858

FID=family ID, IID=individual ID, PEDSEX=the recorded sex in the PED file,
SNPSEX=sex determined using SNP/genetic data, STATUS=indicate whether there is an error or not,
F=X chromosome inbreeding (homozygosity) estimate

After this analysis, we imputed the correct sex information and a new BED file was generated, *parents2.bed*. Because the analysis for this project did not involve males, data for IDs: 510067.20, 7210113.20, and 9910657.20 were removed from the revised alternative phenotype file, named *meiosis3.txt*.

```
plink --bfile parents --impute-sex --make-bed --out parents2
```

2.2.5.2 Identification of Duplicated or Related Individuals

The next step was to identify duplicated individuals which were recorded by mistake. For each pair of parents within the same family, a kinship metric (identity by state, IBS) was calculated based on the average proportion of alleles shared in common at genotyped SNPs (excluding the sex chromosomes) (Purcell et al., 2007). The IBS output is displayed in **Table 3**.

Table 3. Output after computing within-family IBS shows a duplicate sample recorded under a different ID

FID1	IID1	FID2	IID2	RT	EZ	Z0	Z1	Z2	PI_HAT	PHE	DST	PPC	RATIO
4710041	10	4710041	20	OT	0	0.9213	0.0657	0.0130	0.0459	-1	0.785929	0.9997	2.2298
5510067	10	5510067	20	OT	0	0.9460	0.0166	0.0374	0.0457	-1	0.787197	1.0000	2.2936
7210159	10	7210159	20	OT	0	0.0000	0.0024	0.9976	0.9988	-1	0.999666	1.0000	NA
9910657	10	9910657	20_dup3	OT	0	0.9695	0.0000	0.0305	0.0305	-1	0.783736	0.8690	2.0716
9910657	10	9910657	20_dup1	OT	0	0.9606	0.0302	0.0092	0.0243	-1	0.782070	0.9818	2.1358
9910657	10	9910657	20_dup2	OT	0	0.9572	0.0322	0.0106	0.0267	-1	0.782549	0.9742	2.1260
9910657	20_dup3	9910657	20_dup1	OT	0	0.0004	0.0343	0.9653	0.9825	-1	0.995190	1.0000	2393.0000
9910657	20_dup3	9910657	20_dup2	OT	0	0.0008	0.0363	0.9628	0.9810	-1	0.994807	1.0000	2393.5000
9910657	20_dup1	9910657	20_dup2	OT	0	0.0007	0.0532	0.9461	0.9727	-1	0.992519	1.0000	1600.3333

FID1=family ID for the first individual, IID1=individual ID for the first individual,
FID2=family ID for the second individual, IID2=individual ID for the second individual,
RT=relationship type given PED file, EZ=expected IBD sharing given PED file, Z0=P(IBD=0),
Z1=p(IBD=1), Z2=P(IBD=2), PI_HAT=P(IBD=2)+0.5*P(IBD=1) or proportional IBD,
PHE=pairwise phenotypic code (1, 0, -1= AA, AU, and UU pairs),
DST=IBS distance (IBS2 + 0.5*IBS1)/(N SNP pairs), PPC=IBS binomial test,
RATIO=of HETHET:IBS 0 SNPs (expected value is 2)

The column DST contains the computed IBS distance. Duplicates were defined to have IBS > 0.98 (Anderson et al., 2010). The output confirmed that the two supposedly distinct individuals listed under Family ID 7210159 were actually the same person. Since the ID for 7210159.20 was already listed in *meiosis3.txt* phenotype file, there was no need to make any more changes to it.

2.2.6 Per-Marker Quality Control

Per-marker QC essentially involved filtering our genotype data and removing suboptimal individuals and SNPs. Individuals that were missing too much genetic data (maximum individual missingness rate, MIND) and SNPs that were missing from too many samples (GENO) were not able to provide sufficient information in the statistical analyses. In fact, their presence could give a false-positive result and mask the true association (Anderson et al., 2010). Because of our small sample size, we were willing to accept individuals with up to 10% missing genotype information and SNPs with no more than 10% missingness across samples.

SNPs that show little variation, for example 1 of out 100 people, are not useful statistically and are typically removed from a GWAS analysis. The common standard minor allele frequency (MAF) is about 1-2% (Anderson et al., 2010). In this project, a MAF threshold of 1% was used.

Lastly, SNPs that showed extensive deviation from Hardy-Weinberg equilibrium (HWE) were excluded because they are typically indicative of a genotype calling error. The threshold for this parameter ranges between 0.001 to 5.7×10^{-7} (Anderson et al., 2010). For this project, we set the threshold value for HWE to 0.001.

We filtered the data based on these pre-specified parameters and the following output was obtained:

```
plink --bfile parents2 --pheno meiosis3.txt --mind 0.1 --geno 0.1
--maf 0.01 --hwe 0.001 --out summstat

Before frequency and genotyping pruning, there are 906600 SNPs
0 of 383 individuals removed for low genotyping ( MIND > 0.1 )
2240 markers to be excluded based on HWE test ( p <= 0.001 )
      895 markers failed HWE test in cases
      2239 markers failed HWE test in controls
Total genotyping rate in remaining individuals is 0.977565
42538 SNPs failed missingness test ( GENO > 0.1 )
105229 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 763771 SNPs
After filtering, 40 cases, 94 controls and 249 missing
```

Afterwards, we were left with 134 Caucasian mothers. The number of cases of meiosis I and meiosis II nondisjunction is summarized in **Table 4**. **Table 5** summarizes the information from the per-marker QC.

Table 4. Distribution of meiosis I and meiosis II NDJ cases after quality control implementation

Stage of nondisjunction error	Count (%)
M1 (phenotype=1)	94 (70.15)
M2 (phenotype=2)	40 (29.85)
TOTAL	134

Table 5. Summary of output after implementing per-marker QC

	Set parameter	Number removed	Total number	Unit
Maximum individual missingness rate (MIND)	0.1	0	134	individuals
Genotyping rate (GENO)	0.1	42538	906600	SNPs
Minor allele frequency (MAF)	0.01	105229	906600	SNPs
Hardy Weinberg equilibrium (HWE)	0.001	2240	906600	SNPs
controls (M1)	---	2239	906600	SNPs
cases (M2)	---	895	906600	SNPs
After per-marker QC	MIND=0.1 GENO=0.1 MAF=0.01 HWE=0.001	---	**Total remain** 134 763771	 individuals SNPs

2.3 TEST FOR ASSOCIATION

2.3.1 Background

The idea behind association analysis was to look through each SNP one by one, testing to see if there was a difference in the frequency of genotypes seen between the two case groups: M1 vs. M2. If this difference was statistically significant, then that SNP would be said to be associated with the phenotype. The method of statistically testing the frequencies was through a chi-squared

test. Under the null hypothesis of no association, we would expect the genotype frequencies between the two groups to be the same.

There are different types of association analyses:

1. Basic allelic test for association

The unit of this test is the allele rather than the genotype. A genotype consists of two alleles, AA, Aa, or aa. The associations of the phenotype with these individual alleles are then tested. This test has the advantage in that the sample size would double. However, it is not the best test for association in humans as it ignores the overall genotype of the two chromosomes.

2. Genotypic tests

The units for these tests are the genotypes.

a. Basic genotypic test

It is a 3×2 tabulation. There are 3 genotypes--homozygote aa, heterozygote Aa, and homozygote AA--and 2 comparison groups (case/control or case/case). This test has (2-1)×(3-1) = 2 degrees of freedom. The extra degree of freedom means that a larger chi-squared value is needed to obtain the same p-value in comparison to a 2×2 contingency table. However, this is the most basic test and makes no assumptions about the genetic model.

b. Additive model

This model assumes that having two copies of the minor allele (AA genotype) has twice the effect of having a single copy of the

minor allele (Aa genotype). This test has 1 degree of freedom. It is also known as the 'Cochran-Armitage test for trend.'

c. Dominant Model

This model assumes that an effect on phenotype is only seen if you have at least one copy of the minor allele (either genotype Aa or AA). This test also has 1 degree of freedom.

d. Recessive model

This model assumes that an effect on phenotype is only seen if you have two copies of the minor allele (genotype AA only). This test also has 1 degree of freedom.

It is recommended that if the underlying genetic model is unknown, then the additive model should be used (Clarke et al. 2011). Kuo and Feingold (2009) discussed in their paper that recessive and dominant loci are quite rare in whole genome-scanning and the model with the best statistical power is an intermediate between the recessive model and dominant model--the additive model. For these reasons, the type of model selected for our association analysis is the additive model or the Cochran-Armitage test for trend.

Table 6 below illustrates how the genetic information is grouped for statistical analysis. For each SNP, the number of the minor allele A is counted from the subject's genotype (the subject's genotype can only be one of the following: AA, Aa, or aa). If the subject has genotype aa, then the number of minor allele is determined to be 0. If, however, the subject has genotype Aa, then the number of minor allele is determined to be 1, and so on.

Table 6. 2x3 genotype-based table

Group	Number of Minor Allele			Total
	0	1	2	
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

The general format of the test statistic is

$$T = \sum_{i=0}^{2} x_i(Sr_i - Rs_i)$$

$$= N(r_1 + 2r_2) - R(n_1 + 2n_2) \qquad (1)$$

where x_i are weights. For the Cochran-Armitage test for trend trend, it specifically has a weight of x=(0, 1, 2). The second line of equation 1 shows the equation after this specific weight has been applied. The null hypothesis can be expressed as:

H_0: All entries in the table are proportional or

$$\Pr(case|minor\ allele = 0) = \Pr(case|minor\ allele = 1) = \Pr(case|minor\ allele = 2) \quad (2)$$

This test statistic has the expected value

$$E(T) = E[E(T|R,S)] = E(0) = 0 \qquad (3)$$

and variance under the H_0 of

$$Var(T) = \frac{RS}{N}\left[\sum_{i=0}^{2} x_i^2 n_i(N - n_i) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j n_i n_j\right]$$

$$= \frac{SR}{N}[N(n_1 + 4n_2) - (n_1 + 2n_2)^2] \qquad (4)$$

Therefore the Cochran-Armitage trend test statistic is

$$\chi^2_{trend} = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]} \overset{H_0}{\sim} \chi^2_1 \qquad (5)$$

which follows the χ^2 distribution of 1 degrees of freedom (Cochran, 1954; Armitage, 1955; Kuo and Feingold, 2010).

2.3.2 Association Analysis

The Cochran-Armitage test for trend was used to analyze the genetic data. Per-marker QC was done concurrently with the statistical procedure.

```
plink --bfile parents2 --pheno meiosis3.txt --mind 0.1 --geno 0.1
      --maf 0.01 --hwe 0.001 --model --adjust --model-trend
      --out assoctest
```

The output from the trend test was generated as a text file, *assoctest.model.trend.adjusted*. The output was later opened using Microsoft Excel and sorted in ascending order by their unadjusted p-values (UNADJ). **Table 7** lists the first 20 significant SNPs.

Table 7. Output from the Cochran-Armitage test for trend

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
7	rs1404414	2.45E-06	2.73E-06	1	1	0.841	0.841	0.7285	1
4	rs6854711	3.90E-06	4.32E-06	1	1	0.9463	0.9463	0.7285	1
7	rs2691616	4.16E-06	4.61E-06	1	1	0.9559	0.9559	0.7285	1
2	rs1189634	6.95E-06	7.65E-06	1	1	0.9945	0.9945	0.7285	1
7	rs4316058	7.76E-06	8.54E-06	1	1	0.997	0.997	0.7285	1
7	rs2723497	8.25E-06	9.07E-06	1	1	0.9979	0.9979	0.7285	1
7	rs1830006	9.65E-06	1.06E-05	1	1	0.9993	0.9993	0.7285	1
7	rs2691609	9.78E-06	1.07E-05	1	1	0.9993	0.9993	0.7285	1
7	rs1028270	1.00E-05	1.10E-05	1	1	0.9994	0.9994	0.7285	1
7	rs6973480	1.14E-05	1.25E-05	1	1	0.9998	0.9998	0.7285	1
12	rs1085023	1.23E-05	1.35E-05	1	1	0.9999	0.9999	0.7285	1
12	rs1559836	1.23E-05	1.35E-05	1	1	0.9999	0.9999	0.7285	1
7	rs4543433	1.33E-05	1.46E-05	1	1	1	1	0.7285	1
15	rs1720570	1.36E-05	1.49E-05	1	1	1	1	0.7285	1
14	rs1709986	1.64E-05	1.79E-05	1	1	1	1	0.8042	1
9	rs1328673	1.72E-05	1.88E-05	1	1	1	1	0.8042	1
17	rs7207659	2.20E-05	2.40E-05	1	1	1	1	0.8929	1
7	rs4142995	2.22E-05	2.42E-05	1	1	1	1	0.8929	1
7	rs1176886	2.26E-05	2.47E-05	1	1	1	1	0.8929	1
7	rs6970593	2.56E-05	2.79E-05	1	1	1	1	0.9597	1

CHR=chromosome number, SNP=SNP identifier, UNADJ=unadjusted p-value,
GC=genomic-control corrected p-value, BONF=Bonferroni single-step adjusted p-value,
HOLM=Holm (1979) step-down adjusted p-value, SIDAK_SS=Sidak sinle-step adjusted p-value,
SIDAK_SD=Sidak step-down adjusted p-value,
FDR_BH=Benjamin & Hochberg (1995) step-up false discovery rate control,
FDR_BY=Benjamin & Yekutieli (2001) step-up false discovery rate control

2.3.3 Multiple Testing

Because of the multiple association tests being done, one for each SNP, controlling for multiple comparisons should be heavily considered. Type I error is the probability of wrongly rejecting the null hypothesis when the null hypothesis itself is actually true. It is also termed “the false-positive rate.” The conventional p-value of <0.05 used as the threshold of level of significance cannot be applied to this type of study. If so, an association study of 1 million SNPs will only result in 50,000 SNPs being associated with the difference in phenotype. This interpretation of the output is not very helpful.

Both Bonferroni and Sidak adjusting methods are very conservative and assume independence. Holm is a correction method similar but less stringent to the former two. Another method, the false discovery rate (FDR) procedure, controls for the expected proportion of false positives among significant SNPs (Clarke et al. 2011). In GWA studies, the most commonly used correction method is still the Bonferroni method (Pearson et al., 2008). Journals tend to accept GWAS findings at the 5×10^{-7} and sometimes even at the 5×10^{-8} significance level (Clarke et al. 2011).

However, all of these aforementioned methods do not account for the dependence among SNPs (Clarke et al. 2011). SNPs that are near one another are correlated. In addition, this project was a pilot study examining potential genetic candidates accountable for the difference in the nondisjunction stage. The sample size of this study was very small as well, with 94 M1 cases vs. 40 M2 cases. Applying the stringent p-value of 5×10^{-8} or even 5×10^{-7}, would dismiss any potential genetic markers that may truly affected the meiosis mechanism.

2.3.4 Manhattan Plot

A Manhattan plot is a type of scatter plot used to display data with a large number of data points. In a GWA study, the SNP location is displayed on the x-axis while the y-axis displays the -log10 of the p-value for that SNP. Therefore the stronger the association, for example p-value = 10^{-7}, the higher the negative log value will be (e.g. 7).

Haploview is a bioinformatics software designed to plot data from statistical genomic studies. gPLINK conveniently links the analyzing capacity of PLINK with the visualizing capacity of Haploview. gPLINK was launched within PLINK by using the command

```
Java -jar gPLINK.jar
```

and the output file, *assoctest.model.trend.adjusted*, was selected in the menu box. After hitting the 'Plot' button, -log10 transformed unadjusted p-values (UNADJ) were selected to be plot on the y-axis. **Figure 1** is the obtained Manhattan plot.

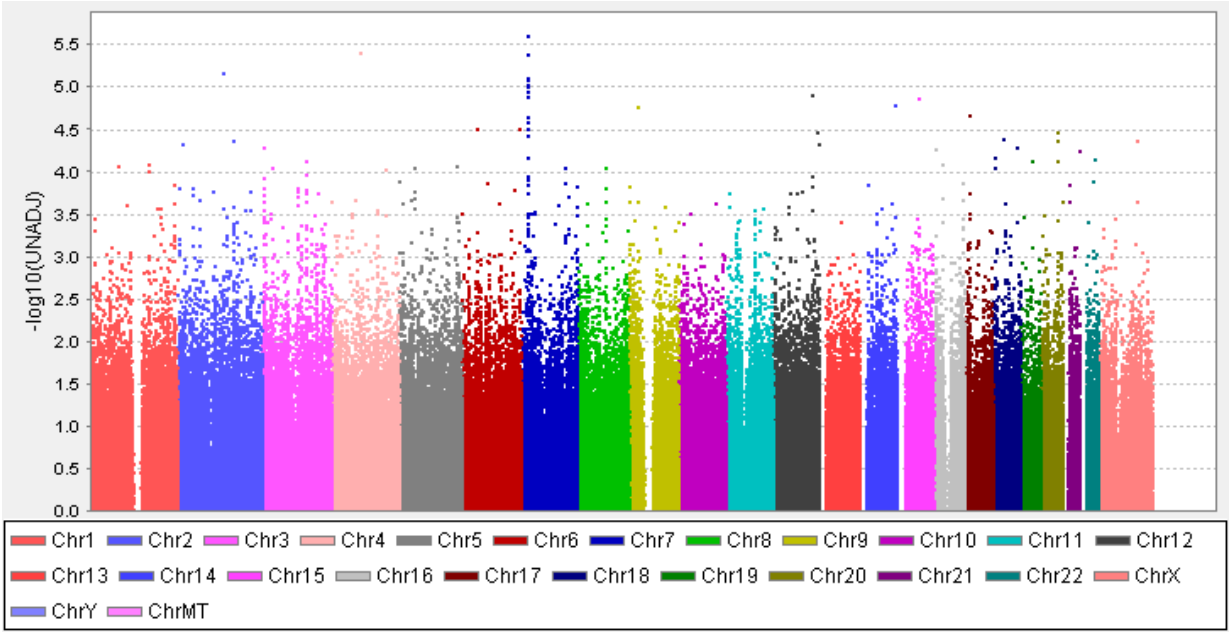

Figure 1. Manhattan plot of -log (P values) using Haploview

The plot conveniently allows us to see if there are multiple associations within a small genomic region. The highest blue dot represents the most significant SNP, rs1404414, which is located on chromosome 7. Note that there are many other SNPs on chromosome 7 that have very small p-values, as small as the p-value of rs1404414. This grouping effect is due to the correlation among nearby SNPs and it is termed "linkage disequilibrium."

2.3.5 Quantile-Quantile Plot

A quantile-quantile plot (QQ plot) serves as a diagnostic tool for checking the type I error. It plots the observed -log10 p-values against the expected -log10 p-values under the null model. The expected p-values under the null follow a uniform distribution. If all points fall on the diagonal line, then none of the observed p-values are different from the expected ones. This would mean that there is no association.

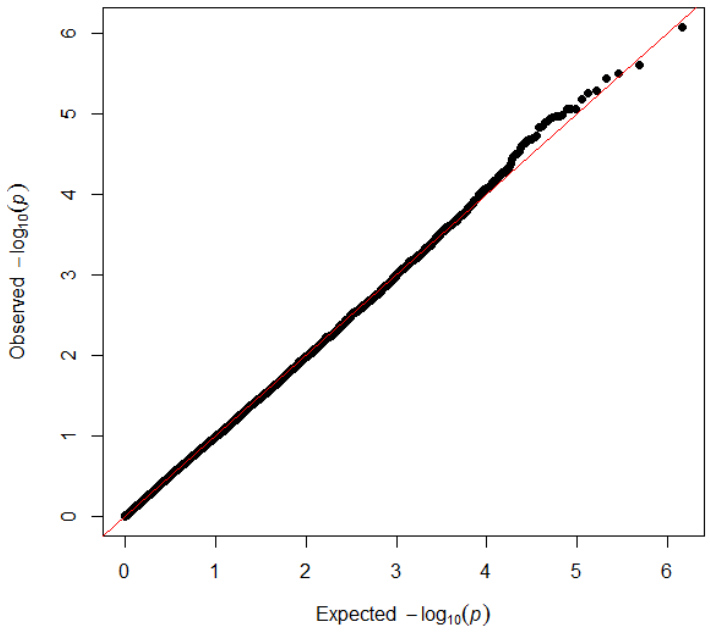

Figure 2. QQ plot of -log10 (P value) using R

The obtained quantile-quantile plot in **Figure 2** shows that most of the points fall on the diagonal line. There is a little deviation at the top right corner which is expected when there is an association between SNPs and the phenotype of our interest.

2.3.6 LocusZoom

LocusZoom is a tool available on the web to plot association results from GWAS and scan to see if any of the SNPs landed on any actual biological genes. Being able to identify the genes that are involved solidifies the association further and it suggests explainable biological mechanism. There are two gene scanning methods that we are utilizing. First we scan for genes near the most significant SNP, rs1404414. In the second method, we scan the output in the area of five prechosen genes that we suspect might be associated with DS.

2.3.6.1 Scanning for Potential Gene Candidates

The output file, *assoctest.model.trend.adjusted*, is first compressed using SecureZIP® for Windows and it is being uploaded to https://statgen.sph.umich.edu/locuszoom/genform.php?type=yourdata.

The information fields were set to accommodate PLINK data. A slight adjustment is being made where the *P-value Column Name* is specified to be **UNADJ**. The *Specify Region to Display* is SNP **rs1404414**. **Figure 3** shows the LocusZoom plot for rs1404414.

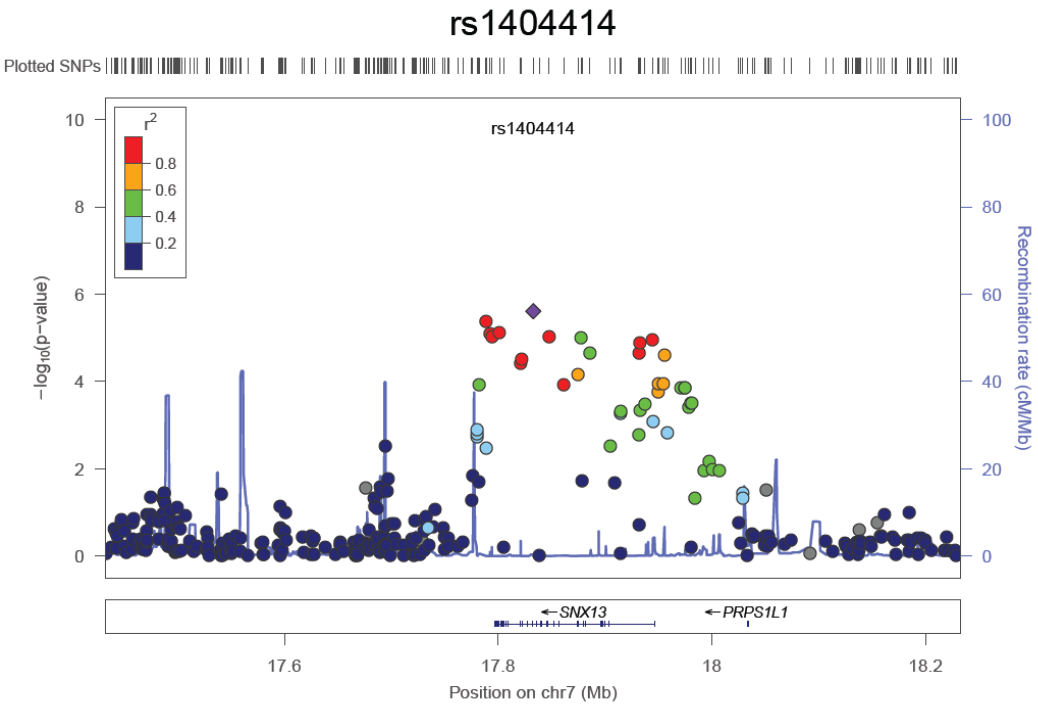

Figure 3. Regional plot of area surrounding rs1404414 of -log (P values) using LocusZoom

The top significant p-values belong SNPs located on the region of the gene SNX13. Another cluster of SNP with a slightly higher p-values seem to fall near gene PRPS1L1. Another LocusZoom plot was obtained. This time the *Specify Region to Display* was Gene **prps1l1**. **Figure 4** shows obtained plot for this gene.

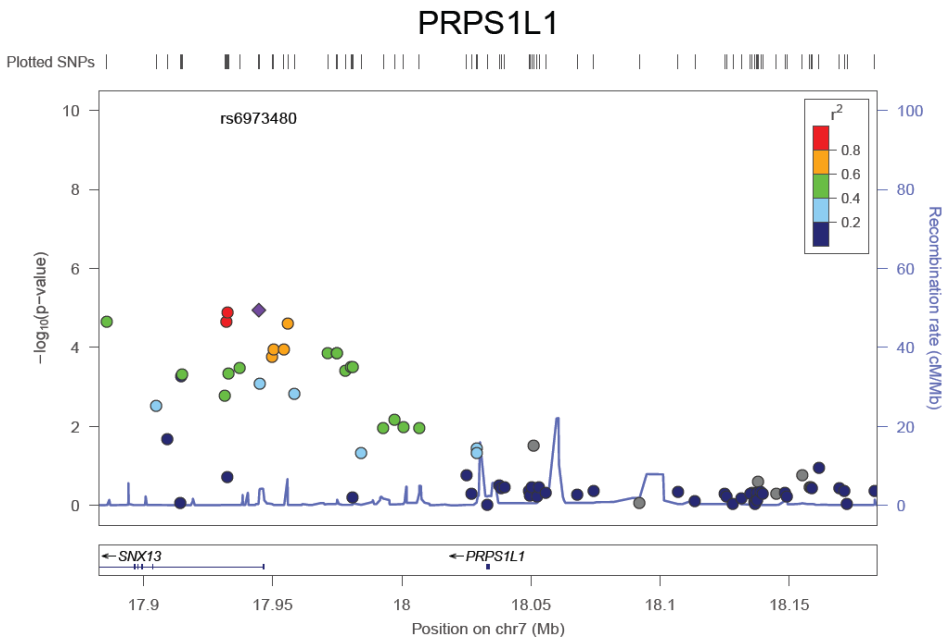

Figure 4. Regional plot of area surrounding PRPS1L1 of -log (P values) using LocusZoom

2.3.6.2 Scanning for Association with Candidate Down Syndrome Genes

The p-values from the trend test were used to examine for possible association with five genes that are candidates for association with Down syndrome. These genes are RNF212, PRDM9, APOE, PSEN1, and MAPT. The regional plots of the p-values for these five genes are shown in **Figure 5 - Figure 9**. They were obtained by simply switching gene names under the *Specify Region to Display* to each of the five genes listed above.

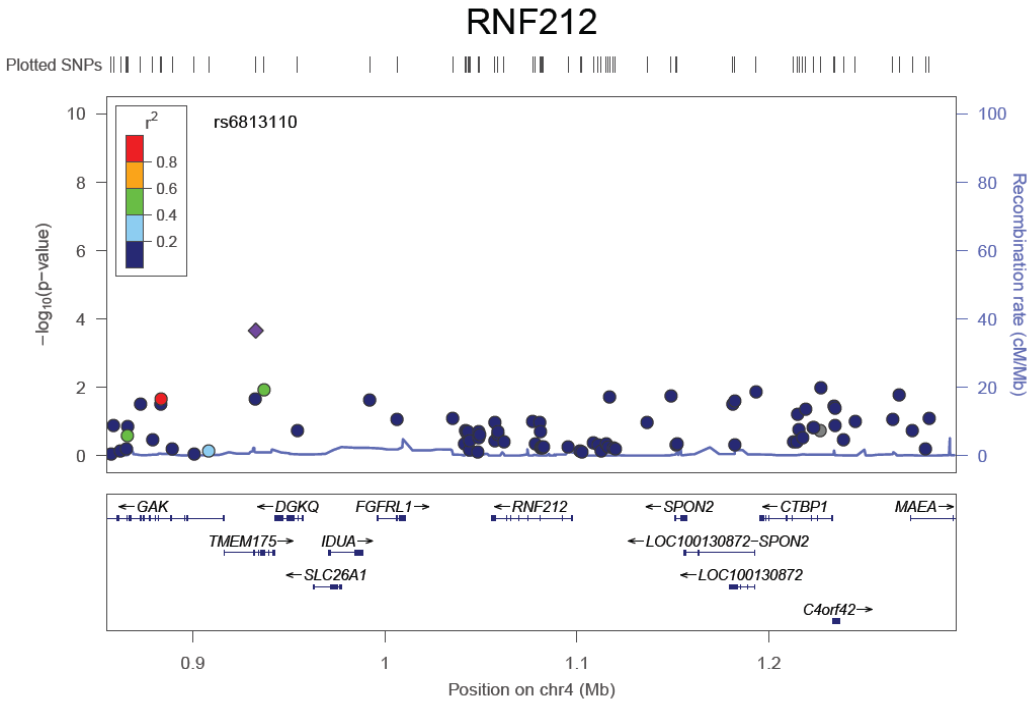

Figure 5. Regional plot of area surrounding RNF212 of -log (P values) using LocusZoom

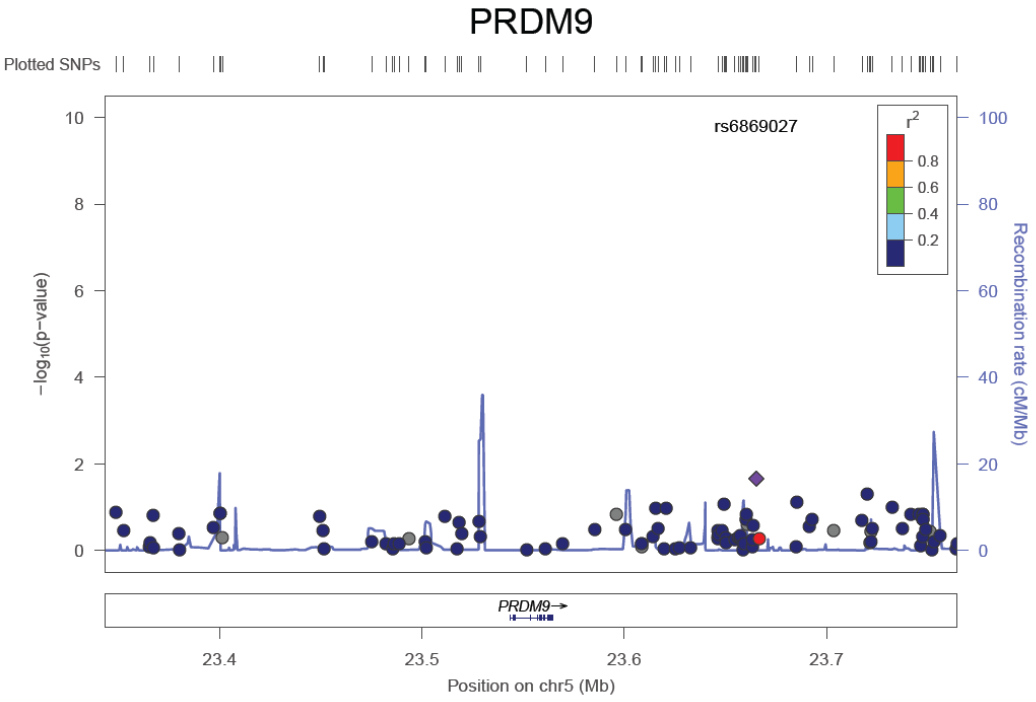

Figure 6. Regional plot of area surrounding PRDM9 of -log (P values) using LocusZoom

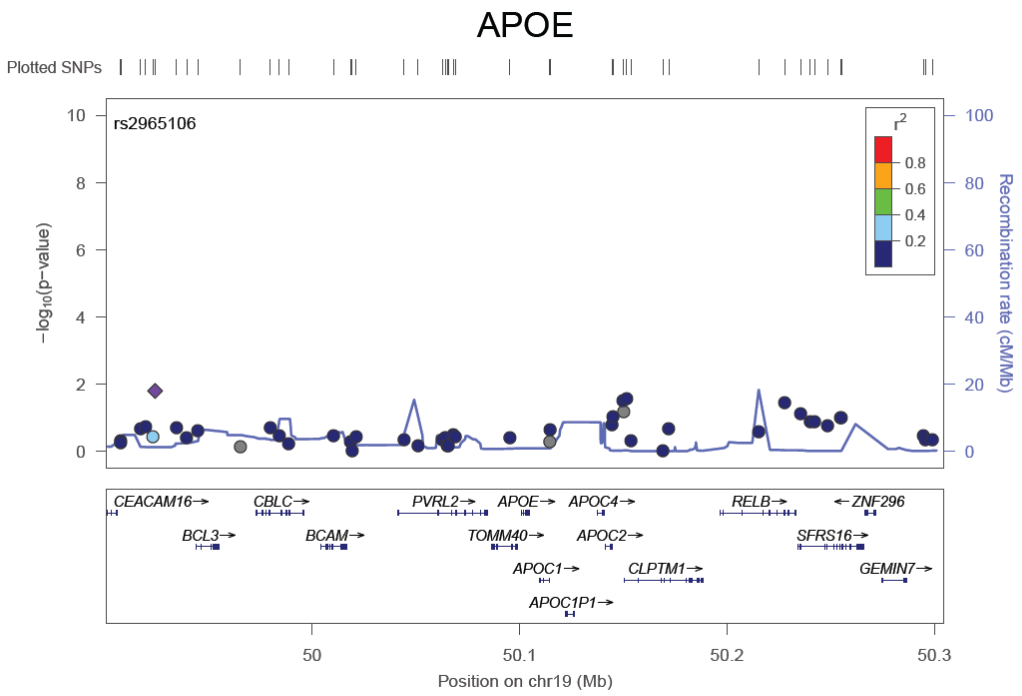

Figure 7. Regional plot of area surrounding APOE of -log (P values) using LocusZoom

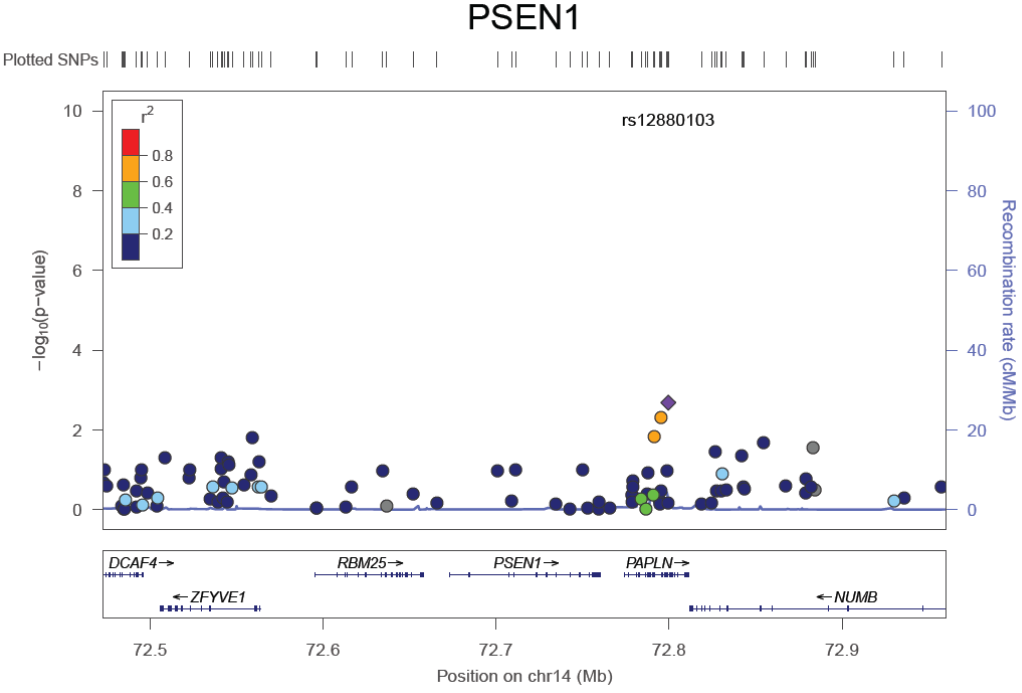

Figure 8. Regional plot of area surrounding PSEN1 of -log (P values) using LocusZoom

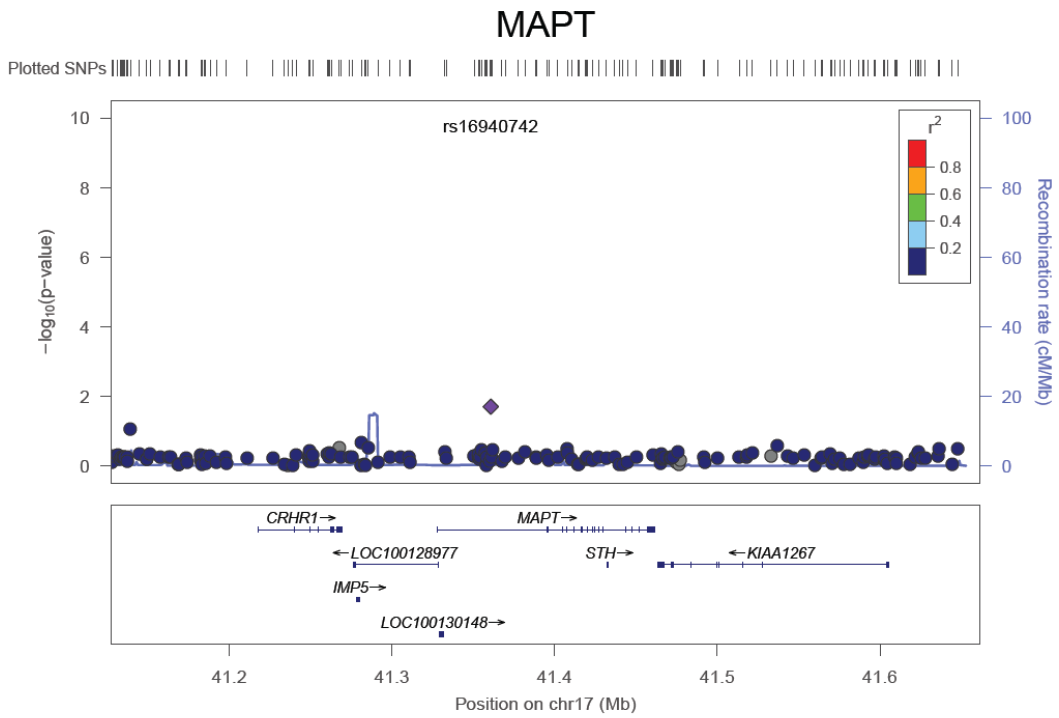

Figure 9. Regional plot of area surrounding MAPT of -log (P values) using LocusZoom

RNF212 regional plot has the highest -log p-value (in other words, the lowest p-value) among the five plots. This is followed by PSEN1 regional plot. Clustering of correlated data points can be observed in these two plots whereas data points in APOE and MAPT regional plots show no pattern. The regional plot of PRDM9 shows a slight clustering and the p-value of the highest point does not even reach the 10^{-2} or 0.01 level.

3.0 DISCUSSION

3.1 SAMPLE SIZE AND DATA QUALITY

The final sample size of the study was 134 Caucasian mothers who passed on the extra chromosome 21 to their DS children. Of these 134, 94 (70%) of them had an NDJ error of chromosome 21 during meiosis I and the remaining 40 mothers (30%) had the NDJ error during meiosis II. The distribution of these cases agrees with past data (Sherman et al, 2005). This sample size is considered relatively low for a GWA study. However, this was expected and unavoidable given that we were a previously-collected study sample to analyze a new pilot research question. Nevertheless, the data had some positive qualities with regards to reducing confounding effects and we were able to identify suggestive candidate genes for NDJ of trisomy 21 and provide supporting evidence for some candidate NDJ-related genes via recombination or AD.

Of course due to the small sample size, statistical power was consequently relatively low and the significance of the evidence for association with the identified SNPs is lower than the standard genome-wide significance using Bonferroni correction method of 5×10^{-8} (Pearson and Manolio, 2008). On the other hand, we had some advantages from using a homogenous cohort of individuals. Both M1 and M2 cases were recruited from the same place using the same data collection method and researchers were not aware of the particular case outcome of each

individual until the genetic data had already been collected and assayed in lab. It could be said that one positive quality of our data despite the sample size was the reduced background noise and confounding effect. This could be confirmed by the QQ plot in **Figure 2**. The plot shows the expected distribution of the association test statistics (x-axis) across the million SNPs compared to the observed values (y-axis). Any deviation from the X=Y line implies a consistent difference between cases and controls (or between cases 1 and cases 2) across the whole genome. Our QQ plot was remarkably ideal showing not a lot of deviation from X=Y line until reaching towards the right-hand end, where the deviation from the X=Y line represents the difference between the M1 and M2 cases implying a true association of the SNPs with NDJ.

The believability of the association between significant SNPs and NDJ lies within the quality of our genotype data. After tremendous effort being put in quality control, we arrived with quite clean genotype data. Good QC should filter out artifacts thus allowing the true association to show through. The Manhattan plot, in addition to being used to identify the location of significance SNPs, could be used to assess the quality of the genotype data and believability of the association. The association would be questionable if the dots representing each SNP were quite scattered about despite locating quite high on the y-axis (having very low p-values). We were satisfied to find that our data had good genotype quality. Our Manhattan plot in **Figure 1** showed neat clustering of SNPs for each chromosome and an obvious peak formed by significant SNPs on chromosome 17 with nearby correlated SNPs showing similar signal strength.

3.2 SIGNIFICANT GENES

3.2.1 Genome-Wide Exploration

This genome-wide case-case study identified several SNPs on chromosome 7 having lowest p-values--the lowest at 2.45×10^{-6} (rs1404414; see **Table 7**). This SNP did not meet the "genome-wide" significance threshold as previously mentioned in section 3.1. Nevertheless, this SNP was found to be correlated ($r^2 > 0.8$) with several other low p-value SNPs clustering in one localize region, the SNX13 gene (see **Figure 3**). This pattern suggested a probable finding, not simply an artifact that happened to have the lowest p-value. If the data set had not been limited to a small number of cases, then this SNP would likely have been significant. The association between SNX13 and NDJ could not be concluded with enough statistical significance. However, we could say that there was a suggested association with the SNX13 gene that we could further explore about.

The purpose of this study was to identify genetic markers for meiotic NDJ and hypothesize the involvement of such genes with NDJ error. Therefore, we examined the characteristics of SNX13. SNX13 gene (also previously known as RGS-PX1 gene) codes for the sorting nexin 13 protein which has both a PHOX and a RGS protein domains thus making it a member of both the sorting nexin (SNX) protein family and the regulator of G protein signaling (RGS) family. Sorting nexin is a large group of proteins localize in the cytoplasm and interact either with the cell membrane directly through their phospholipid-binding PX domain or with protein complexes on the cell membrane through protein-protein interaction (Worby and Dixon, 2002). G proteins, on the other hand, are a family of protein involved in transmitting chemical signal originating from outside the cell inside the cell (Zheng et al., 2001). Because of this

duality, this protein locating within the endosomes may serve as a link between protein signaling and vesicular trafficking (Zheng et al, 2001; Worby and Dixon, 2002). Unfortunately, there has been no knowledge of SNX13 involvement with NDJ or meiosis in general. We could only speculate how SNX13 could be related to NDJ error. Perhaps among those with a particular NDJ type, SNX13 erroneously could not function its usual role in signal transmission or protein trafficking that eventually leads to NDJ during one of the meiosis stage.

While the most significant SNPs was in the SNX13 gene region, we could not ignore the correlation ($0.2 < r^2 < 0.6$; see **Figure 4**) with its neighboring gene, phosphoribosyl pyrophosphate synthetase 1-like 1 (PRPS1L1), and discount the potential significance of the latter gene. PRPS1L1 is 86K basepairs away from SNX13 based on the information available on the Epigenomics database (NCBI, 2012). This gene (recall that it is located on the autosomal chromosome 7) is related to the PRPS1 and PRPS2 genes of the X chromosome (Taira et al., 1989). The protein that PRPS1L1 encodes is highly homologous to phosphoribosylpyrophosphate synthetase encoded by PRPS1 and PRPS2. These enzymes altogether convert the nitrogenous bases (pyrimidine and purine) to their corresponding nucleotides in the making of mRNA, although the involvement of the PRPS1L1 in this function has only been observed in the testis so far (Taira et al., 1990). The same study found an association of this gene with chromosome deletion(Taira et al., 1990). However, there has been no further studies on PRPS1L1 to confirm or deny this relationship. Similar to the circumstance with SNX13, we could only speculate the involvement of PRPS1L1 with NDJ that perhaps a malfunction of the enzyme (wrong conversion of nucleotides or deleting some other important chromosome segments) somehow leads to NDJ in one of the meiosis stage.

The design of our case-case study allowed us to avoid selection biases and maximize the association between SNPs and NDJ. There were still limitations as in all other GWA studies. One in particular was the potential for a false-positive association (Pearson and Manolio, 2008). The association between NDJ and SNX13 (or PRPS1L1) could be true but it could be false as well. This is not saying that completing this study was not fruitful. The benefit of a GWA study is the narrowing down of genes from the entire human genome to a few ones that we could subsequently focus our effort and verify in a follow-up study. Other limitations of this study in particular was the small sample size and the significance of the association not meeting the Bonferroni-adjusted threshold recommended for all GWA studies (Pearson and Manolio, 2008). As such, additional genotyping of a larger sample of NDJ cases should be subsequently done to confirm and prove the significance of the relationship between SNX13/ PRPS1L1 and difference in M1 and M2 NDJ etiologies.

3.2.2 Recombination and Alzheimer 's Disease Genes

In comparing the results from the Cochran-Armitage test for trend with the five gene candidates of RNF212, PRDM9, APOE, PSEN1, and MAPT, we utilized a different significant p-value threshold. For this analysis we were no longer making comparisons among 10 million SNPs. Rather, we made comparisons among a smaller subset of SNPs--those that made up the five candidate genes. A different p-value threshold needed to be constructed instead of using the genome-wide threshold of 5×10^{-8}. The new p-value threshold was computed by utilizing the simple and most conservative Bonferroni correction method.

The Bonferroni correction for multiple comparisons is given in equation 6 below (Bonferroni, 1935, 1936). It gives the new p-value threshold (α_e) to be applied given a pre-

determined point-wise (or single test) threshold (α_p) and the number of independent comparisons being made (*N*).

$$\alpha_e = \frac{\alpha_p}{N} \qquad (6)$$

We approximated that the number of SNPs per gene to be no more than 15. Therefore there were 15 SNPs/gene × 5 genes = 75 SNPs being tested. Using a point-wise p-value of 0.05, the Bonferroni-corrected p-value for this particular testing is then 0.05/75 ≈ 0.000667. In other words, the p-value for determining significance of the five candidate genes would approximately be between 5×10^{-4} and 10^{-4}. Note that we could be a little liberal with the number due to the pilot nature of this study.

SNPs in the RNF212 and PSEN1 regional plots both showed linkage clustering and had p-values meeting this constructed threshold (see **Figure 5** and **Figure 8**). While the SNPs were not exactly located on the RNF212 and PSEN1 genes, the significant SNPs were near these gene regions enough to suspect RNF212 and PSEN1 involvement due to the highly correlated nature of genes within the same neighborhood. As previously discussed in section 1.1, RNF212 was found to affect recombination rate (Kong et al., 2008). M1 NDJ and M2 NDJ cases had different recombination rate; M1 had reduced to no recombination (Lamb et al., 1996, 1997, 2005). PSEN1 was a gene, in which its mutation was observed among a rare form of early onset AD (Sherrington et al., 1995). This gene was also thought to be involved in chromosome organization and segregation (Li et al., 1997). Petersen et al. (2000) found increased frequency of PSEN1 susceptible allele among M2 cases and that difference in distribution was reflected in our discovery of possible association between PSEN1 and meiotic NDJ of trisomy 21.

The obtained regional plots for RNF212 and PSEN1 were not perfect--the significant SNPs not being located exactly on each respective gene region. Probable explanations for the

imperfection included the small sample size of subjects and the type of genotype assay chips being used. We could not iterate enough the importance of a follow-up study to verify the findings from this project using a much larger sample size. A systemic way of genotyping the collect data using different kinds of assay chips could also be attempted.

BIBLIOGRAPHY

Al Awadi SA, Naguib KK, Bastaki L, Gouda S, Mohammed FM, Abulhasan SJ, Al-Ateeqi WA, Krishna Murthy DS: Down's syndrome in Kuwait: recurrent familial trisomy 21 in sibs. Med Principles Pract 8:156-163 (1999).

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: Data quality control in genetic case-control association studies. Nat Protoc 5:1564–1573 (2010).

Armitage P: Tests for Linear Trends in Proportions and Frequencies. Biometrics 11:375–386 (1955).

Avramopoulos D, Mikkelsen M, Vassilopoulos D, Grigoriadou M, Petersen MB: Apolipoprotein E allele distribution in parents of Down's syndrome children. Lancet 347:862-865 (1996).

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B: PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 324:836 (2010).

Bonferroni CE: Il calcolo delle assicurazioni su gruppi di teste, chapter "Studi in Onore del Professore Salvatore ortu Carboni". Rome, Italy: p. 13–60 (1935).

Bonferroni CE: Teoria statistica delle classi e calcolo delle probabilitá. ubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3–62 (1936).

Brook JD, Gosden RG, Chandley AC: Maternal ageing and aneuploid embryos--evidence from the mouse that biological and not chronological age is the important influence. Hum Genet 66:41-45 (1984).

Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: Basic statistical analysis in genetic case-control studies. Nat Protoc 6:121-133 (2011).

Cochran WG: Some Methods for Strengthening the Common $\chi 2$ Tests. Biometrics 10:417-451 (1954).

Curtis D, Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Lawrence J, Anjorin A, Choudhury K, Datta SR, Puri V, Krasucki R, Pimm J, Thirumalai S, Quested D, Gurling HM: Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. Psychiatr Genet 21:1-4 (2011).

Der Kaloustian VM, Masri R, Khudr A, Talj F, Libbus B, Nabulsi M, Khouri FP: Down syndrome in two siblings with 47,XY, +21 and 46,XY/46,XY,-21, +t(21q;21q). Hum Genet 75:97 (1987).

Emanuel I, Sever LE, Milham S Jr, Thuline HC: Accelerated ageing in young mothers of children with Down's syndrome. Lancet 2:361-363 (1972).

Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, Przeworski M: Variation in human recombination rates and its genetic determinants. PLoS One. 6:e20321 (2011). doi:10.1371/journal.pone.0020321

Hegde MR, Chin EL, Mulle JG, Okou DT, Warren ST, Zwick ME: Microarray-based mutation detection in the dystrophin gene. Hum Mutat 29:1091-1099 (2008).

Hook EB, Mutton DE, Ide R, Alberman E, Bobrow M: The natural history of Down syndrome conceptuses diagnosed prenatally that are not electively terminated. Am J Hum Genet 57:875-881 (1995).

International HapMap Consortium: The International HapMap Project. Nature 426:789-796 (2003).

Jantsch V, Pasierbek P, Mueller MM, Schweizer D, Jantsch M, Loidl J: Targeted gene knockout reveals a role in meiotic recombination for ZHP-3, a Zip3-related protein in *Caenorhabditis elegans*. Mol Cell Biol 24:7998 (2004).

Kolata G: Down syndrome--Alzheimer's linked. Science 230:1152-1153 (1985).

Kuo CL, Feingold E: What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol 35:246-253 (2010).

Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, Jonsdottir GM, Gudjonsson SA, Sverrisson S, Thorlacius T, Jonasdottir A, Hardarson GA, Palsson ST, Frigge ML, Gulcher JR, Thorsteinsdottir U, Stefansson K: Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science 319:1398-1401 (2008).

Lamb NE, Freeman SB, Savage-Austin A, Pettay D, Taft L, Hersey J, Gu Y, Shen J, Saker D, May KM, Avramopoulos D, Petersen MB, Hallberg A, Mikkelsen M, Hassold TJ, Sherman SL: Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II. Nat Genet 14:400–405 (1996).

Lamb NE, Feingold E, Savage A, Avramopoulos D, Freeman S, Gu Y, Hallberg A, Hersey J, Karadima G, Pettay D, Saker D, Shen J, Taft L, Mikkelsen M, Petersen MB, Hassold T, Sherman SL: Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjunction of chromosome 21. Hum Mol Genet 6:1391–1399 (1997).

Lamb NE, Yu K, Shaffer J, Feingold E, Sherman SL: An association between maternal age and meiotic recombination for trisomy 21. Am J Hum Genet 76:91–99 (2005).

Li J, Xu M, Zhou H, Ma J, Potter H: Alzheimer presenilins in the nuclear membrane, interphase kinetochores, and centrosomes suggest a role in chromosome segregation. Cell 90:917-927 (1997).

Liptak GS: Down syndrome (trisomy 21; trisomy G). The MERCK Manual Home Health Handbook (2008). Retrieved December 13, 2012, from http://www.merckmanuals.com/home/childrens_health_issues/chromosomal_and_genetic_abnormalities/down_syndrome_trisomy_21_trisomy_g.html?qt=&sc=&alt=

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P: Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327:876-879 (2010).

National Association for Down Syndrome: Facts about Down syndrome (2012). Retrieved December 13, 2012, from http://www.nads.org/pages_new/facts.html

National Center for Biotechnology Information (NCBI): Epigenomics. Retrieved December 20, 2012, from http://www.ncbi.nlm.nih.gov/epigenomics/

National Human Genome Research Institute (NHGRI): Whole Genome Association Studies (2011). Retrieved November 9, 2012, from http://www.genome.gov/17516714

National Human Genome Research Institute (NHGRI): Genome-Wide Association Studies (2011). Retrieved November 9, 2012, from http://www.genome.gov/20019523#gwas-1

Neilsen KG, Poulsen H, Mikkelsen M, Steuber E: Multiple recurrence of trisomy 21 Down syndrome. Hum Genet 78:103-105 (1988).

Parker SE, Mai CT, Canfield MA, Rickard R, Wang Y, Meyer RE, Anderson P, Mason CA, Collins JS, Kirby RS, Correa A, National Birth Defects Prevention Network: Updated National Birth Prevalence Estimates for Selected Birth Defects in the United States, 2004-2006. Birth Defects Res A Clin Mol Teratol 88:1008-1016 (2010).

Parvanov ED, Petkov PM, Paigen K: Prdm9 controls activation of mammalian recombination hotspots. Science 327:835 (2010).

Pearson TA, Manolio TA: How to interpret a genome-wide association study. JAMA 299:1335-1344 (2008).

Petersen MB, Karadima G, Samaritaki M, Avramopoulos D, Vassilopoulos D, Mikkelsen M: Association between presenilin-1 polymorphism and maternal meiosis II errors in Down syndrome. Am J Med Genet 93:366-372 (2000).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559-575 (2007).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: Plink...whole genome association analysis toolset (2009). Retrived September 19, 2012, from http://pngu.mgh.harvard.edu/~purcell/plink/

Schupf N, Kapell D, Lee JH, Ottman R, Mayeux R: Increased risk of Alzheimer's disease in mothers of adults with Down's syndrome. Lancet 344:353-356 (1994).

Schupf N, Kapell D, Nightingale B, Lee JH, Mohlenhoff J, Bewley S, Ottman R, Mayeux R: Specificity of the fivefold increase in AD in mothers of adults with Down syndrome. Neurology 57:979-984 (2001).

Sherman SL. Freeman, SB, Allen EG, Lamb, NE: Risk factors for nondisjunction of trisomy 21. Cytogenetic Genome Res 111:273-280 (2005).

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, Tsuda T, Mar L, Foncin JF, Bruni AC, Montesi MP, Sorbi S, Rainero I, Pinessi L, Nee L, Chumakov I, Pollen D, Brookes A, Sanseau P, Polinsky RJ, Wasco W, Da Silva HA, Haines JL, Perkicak-Vance MA, Tanzi RE, Roses AD, Fraser PE, Rommens JM, St George-Hyslop PH: Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 375:754-760 (1995).

Taira M, Kudoh J, Minoshima S, Iizasa T, Shimada H, Shimizu Y, Tatibana M, Shimizu N: Localization of human phosphoribosylpyrophosphate synthetase subunit I and II genes (PRPS1 and PRPS2) to different regions of the X chromosome and assignment of two PRPS1-related genes to autosomes. Somat Cell Mol Genet 15:29-37 (1989).

Taira M, Iizasa T, Shimada H, Kudoh J, Shimizu N, Tatibana M: A human testis-specific mRNA for phosphoribosylpyrophosphate synthetase that initiates from a non-AUG codon. J Biol Chem 265:16491-16497 (1990).

U.S. Department of Energy Genome Program: Human Genome Project information (2012). Retrieved November 9, 2012, from http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

Worby CA, Dixon JE: Sorting out the cellular functions of sorting nexins. Nat Rev Mol Cell Biol 3:919-931 (2002).

Zheng B, Ma YC, Ostrom RS, Lavoie C, Gill GN, Insel PA, Huang XY, Farquhar MG: RGS-PX1, a GAP for GalphaS and sorting nexin in vesicular trafficking. Science 294:1939-1942 (2001).

Zody MC, Jiang Z, Fung H, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A, Chen L, Wallis J, Glasscock J, Wilson RK, Reily AD, Duckworth J, Ventura M, Hardy J, Warren WC, Eichler EE: Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat Genet 40:1076-1083 (2008).