STATISTICAL METHODS AND ANALYSIS FOR HUMAN GENETIC COPY NUMBER VARIATION AND HOMOZYGOSITY MAPPING

by

Xiaojing Zheng

M.S., Tongji Medical School, Huazhong University of Science and Technology, China, 2001

Ph.D., University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Xiaojing Zheng

It was defended on

May 9, 2012

and approved by

Dissertation Advisor: **Eleanor Feingold**, Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: **Gale A. Richardson**, Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

Committee Member: **Bernard J. Devlin**, Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

Committee Member: **Richard D. Day**, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: **Michael Vanyukov**, Professor, Department of Pharmaceutical Sciences, School of Medicine, University of Pittsburgh

Eleanor Feingold, PhD

STATISTICAL METHODS AND ANALYSIS FOR HUMAN GENETIC COPY NUMBER VARIATION AND HOMOZYGOSITY MAPPING

Xiaojing Zheng, PhD

University of Pittsburgh, 2012

Single nucleotide polymorphism (SNP) arrays are used primarily for genetic association studies, with data being analyzed in most cases one SNP at a time. Several other applications of SNP arrays, however, involve integration of data over multiple markers for a single individual. Two such applications of SNP arrays are studies of copy number variants (CNVs) and regions of homozygosity or identity by descent. Hidden Markov models are a common approach to both of these problems, but other methods have been used as well. In this dissertation I address several methodological issues related to these two types of analysis, and also apply the methods to several datasets.

The purpose of my studies in CNVs is to better detect and analyze CNVs. A major concern for all copy number variation (CNV) calling algorithms is their reliability and repeatability. I use family data as a verification standard to evaluate CNV calling strategies and methods. I make recommendations for how CNV calls can be used in genome-wide association studies. I then apply them to analyze CNVs in studies of psychiatric disorders and birth outcomes. Results from these studies have the potential for great public health significance, because they can lead to better understanding of the genetic etiology and eventually to better markers for disease screening and diagnosis.

Homozygosity mapping is a powerful method to map genes for rare recessive disorders. However, current methods are not ideal, especially when using high density SNP array data from consanguineous families. This study develops improved methods for homozygosity mapping using dense SNP data, and thus will improve the ability of geneticists to find genetic causes of rare recessive diseases. Many of these rare disorders are life-threatening; identification of the disease genes may help with early diagnosis and treatment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

To my family and teachers

I would like to express my sincere gratitude to my thesis advisor, Dr. Eleanor Feingold. She gave me unlimited chance and space to realize my ideas, while she was always there to steer me in the right direction. She never told me the solution to a question, but she helped me to find the keys to the kingdom of science. I hope I can help others someday just like what she has done for me. She is the person who really changed the rest of my life.

My deep appreciation goes to my training program director and thesis committee member, Dr. Gale Richardson. She brought me a brand new world - “Psychiatric Epidemiology”. She provided me great opportunities to present my work in conferences and patiently helped me improving my presentation skills. She also helped with shaping and building up my career little by little.

I am grateful to my first year GSR tutor and thesis committee member, Dr. Michael Vanyukov for his support in my first year study and for his inspiration in my dissertation study.

I acknowledge my committee members, Dr. Bernie Devlin and Dr. Richard Day, for their intellectual input during my graduate career and dissertation study.

I would specially like to take this advantage to thank my husband, son, parents and brother. I owe what I'm now to their unconditional love, complete support and always encouragement. They will never know how much that means to me.

I would like to acknowledge all members in Department of Biostatistics, Psychiatry and Human Genetics, especially Jessalynn Oliver, Yolanda E. Duncan and Jeanette Norbut, for being nice and helpful to my study here; Chris Huziak and Ryan Evans for computer technical supports.

Finally I would like to thank the NIMH Psychiatric Epidemiology T32 training grant for providing me the opportunity to be a trainee in this great program.

1.0 INTRODUCTION

1.1 SNP ARRAY

A single nucleotide polymorphism (SNP) is a DNA sequence variation. It refers to a variation at a single DNA nucleotide site in a population. A SNP can be used as a marker in genetic mapping and other studies. A SNP array is a DNA microarray designed for high-throughput genotyping of thousands to millions of SNPs simultaneously. In general, a SNP array will generate fluorescent intensity signals of the two alleles of each SNP. Detailed information about the data that the SNP array produces is described in the next section. Two commonly used makers of SNP arrays are Affymetrix and Illumina.

SNP array data are most often analyzed by looking at a single SNP at a time over many individuals, such as in genome-wide allelic association analysis; but they can also be used for looking at regions or groups of SNPs in a single or a few individuals. Two such applications are CNV (copy number variation) studies and homozygosity mapping. This dissertation addresses those two applications, which are described in more detail below.

1.2 OVERVIEW OF COPY NUMBER VARIATION

Copy number variation (CNV) refers to the deletion or amplification of a segment of DNA, ranging from one kilobase to several megabases in size. CNVs have raised more and more interest in genetics in recent years, since accumulating evidence has shown that CNVs may play an important role in causing disorders [Henrichsen et al., 2009; Cahan et al., 2009; Sebat et al., 2007]. At the same time, high throughput DNA microarrays, which make the whole-genome scanning of CNVs possible, also greatly accelerate the development of CNV studies.

The first generation of DNA arrays for CNV studies was comparative genomic hybridization (CGH) arrays, which were first reported in the 1990's [Solinas-Toldo, 1997; Pinkel 1998]. Test and reference DNAs were labeled with different fluorescent tags and hybridized on a slide spotted with thousands of genomic DNA (BAC, cosmid, or cDNA) clones. The fluorescence ratio between test and reference DNA was then used to infer the genomic gains/losses. However, the low resolution and sparse coverage are the major disadvantages of CGH. The high noise due to spotting process is another problem for it.

The second generation of methods used for CNV studies is SNP arrays. The array resolution is greatly increased over what can be done in array CGH. For example, the Illumina HumanHap550 BeadChip contains over 550K SNP probes in a single array. However, the probes are not uniformly distributed across the genome; they are particularly sparse in CNV regions, due to the difficulties in designing robust polymorphic probes in these regions.

The third generation of methods for CNV detection is SNP arrays with CNV markers. Non-polymorphic probes (CNV markers) were added to chips, which may help with identifying the CNVs in the regions not previous covered by the SNP arrays. For example, the Illumina

HumanHap610 puts together the content from Illumina HumanHap550, containing around 550K SNPs, and 60K CNV probes, while the Illumina Human660W-Quad adds a different set of over 110K CNV markers based on HumanHap550. The current highest-density chips, the Affymetrix 6.0 and Illumina 1M, each have approximately one million SNPs and one million CNV markers. The CNV studies in this dissertation use the earliest versions of the third generation platforms. However, how to accurately identify ("call") the CNVs based on the data from these assays and how the CNV calls can be used in genetic association studies are major statistical concerns for even the best platforms.

CNV calling algorithms can be classified into three categories: 1) one-dimensional segmentation methods that utilize total intensity measurements only; 2) genotype mining approaches that only use SNP genotype information; 3) generalized genotyping approaches that use both genotype calls and total probe intensities [Yau et al., 2008]. All arrays will generate fluorescent intensity signals for the A and B alleles – X and Y intensities for each person and each SNP/marker. In order to infer the location of the CNVs, the information of X and Y intensities must be combined across SNPs/markers. The two statistics used to model the copy number changes and genotype calls are log R ratio (LRR) and B allele frequency (BAF) respectively. LRR=log (X+Y), which measures the total intensity of the two alleles. The assumption for LRR is that human beings contain two copies of genomic DNAs, the genomic gain (amplification) of a DNA segment will lead to a higher LRR at that specific locus; while loss (deletion) will cause lower LRR. BAF = Y/(X+Y), which assays the relative intensity of the two alleles. For example, if the genotype at a locus is AB, the total copy number will be 2 (X+Y=2), and BAF=1/2 (Y=1); if the genotype is ABB, the total copy number will be 3 (X+Y=3), and BAF=2/3 (Y=2).

The one-dimensional segmentation method was originally proposed for analyzing CGH array data. Both Nonparametric methods (SW-ARRAY, Circular Binary Segmentation, RankCopy, et al) [Price et al., 2005; Olshen et al., 2004, LaFramboise et al., 2009] and parametric methods (GADA, ITALICS, CNAG, dChip et al.) [Pique-Regi et al., 2008; Rigaill et al., 2008; Nannya et al., 2005, Zhao et al., 2004] have been widely used in detecting changes of copy number in tumors. A common question they tried to answer is how to best localize the change-point, where copy numbers change between contiguous segments along the chromosome. For example, circular binary segmentation (CBS) uses a binary segmentation procedure to look for breakpoints, and SW-ARRAY uses dynamic programming to search for breakpoints. A major drawback of these methods is that they do not use SNP genotype information, which may reduce the statistical power in identifying CNVs.

The genotype mining algorithm assumes three classes of diploid genotypes (AA, AB or BB) and models genotyping errors, for example: Mendelian errors, departure from Hardy-Weinberg Equilibrium in contiguous regions, contiguous regions of homozygous genotype calls etc. One of the recent examples is Microdel, proposed by Kohler [Kohler and Cutler, 2007]. This method requires parent-offspring trio data, and it is hard to identify the amplifications due to the ambiguous genotype classifications in this situation (for example, AAB can be called as AB or AA).

The generalized genotyping approaches utilize both genotype information and total intensity measurements. For example, Birdsuite [Korn et al., 2008] for Affymetrix SNPs arrays; and QuantiSNP [Colella et al., 2007] and PennCNV [Wang et al., 2007] for Illumina BeadChips, use a Hidden Markov Model (HMM) to predict the copy number state (hidden state) at each probe locus (time point) along the chromosome. The assumption for PennCNV is that the

emission probability of the log R ratio is distributed as a Gaussian mixture; emission probability of BAF is distributed as a Gaussian mixture when the BAF value is between 0 and 1, and a mixture of point mass and truncated normal when the BAF value is 0 or 1. The strengths of PennCNV include that family relationships can be incorporated into CNVs calling and that the likelihood ratio of the copy number state at each marker is available. Also, it runs fast and is very user friendly. However, the common disadvantages for the generalized genotyping approaches are that the programs do not predict genotype information and the detection of change-points may not be very accurate. PennCNV has become the *de facto* standard for CNV calling using the Illumina platform, and almost all of the work proposed for CNV calling in this dissertation is based on PennCNV.

1.3 CNV IN THIS DISSERTATION

Although various statistical methods have been proposed for detection of CNVs, the statistical estimation of CNVs is still unreliable. The lists of CNVs are different with different statistical algorithms. But even with the same calling algorithm, the results can be quite different due to different calling strategies. Chapter 2 evaluates the reliability of CNV calling strategies. My goal was to make recommendations for how CNV calls can be created and filtered for use in genetic association studies. Next, I applied the CNV-calling strategies in genetic association studies in Chapters 3 and 4. Chapter 3 investigates association between CNVs and psychosis in Alzheimer's disease (AD+P). Chapter 4 tries to identify the CNVs associated with adverse birth outcomes (low birth weight, preterm delivery) and maternal smoking.

1.4 HOMOZYGOSITY MAPPING

Methods that look across multiple markers in SNP arrays are also used for homozygosity mapping. Homozygosity mapping is a method for mapping genes for rare recessive disorders in families or populations. It takes advantage of a fact that offspring of a consanguineous marriage are more likely to have rare recessive disorders. Given a consanguineous family with a rare recessive disorder, it is assumed that the disease genes are more likely to be located in the regions where affected individuals have two identical alleles inherited from a common ancestor. Despite the fact that homozygosity mapping is relatively common; computational methods for it are often very ad hoc and/or statistically sub-optimal. The goal of Chapter 5 is to recommend and test methods that can bring more statistical rigor and power to this endeavor. I focus in particular on family data, as opposed to population data, and on dense SNP data rather than microsatellite data, and develop a hidden Markov model that can be used for detecting regions of homozygosity taking into account multiple family members.

2.0 USING FAMILY DATA AS A STANDARD TO EVALUATE COPY NUMBER VARIATION CALLING STRATEGIES FOR GENETIC ASSOCIATION STUDIES

Xiaojing Zheng[1], John R. Shaffer[2], Caitlin P. McHugh[3], Cathy C. Laurie[3], Bjarke Feenstra[4], Mads Melbye[4], Jeffrey C. Murray[5], Mary L. Marazita[2,6], Eleanor Feingold[1,2]

1. Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; 2. Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; 3. Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; 4. Department of Epidemiology Research, Statens Serum Institut , Copenhagen, Denmark, United States of America; 5. Department of Pediatrics, University of Iowa, Iowa City, Iowa, United States of America; 6. Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America.

This chapter has been accepted in a modified form for publication by Genetic Epidemiology 2012. It is presented here with permission of John Wiley and Sons. Citation:

Xiaojing Zheng, John R. Shaffer, Caitlin McHugh, Cathy Laurie, Mads Melbye, Jeffrey C. Murray, Mary L. Marazita, Eleanor Feingold. Using family data as a verification standard to evaluate CNV calling strategies. Genetic Epidemiology. 2012; 36: 253-262.

2.1 ABSTRACT

A major concern for all copy number variation (CNV) detection algorithms is their reliability and repeatability. However, it is difficult to evaluate the reliability of CNV calling strategies due to the lack of gold standard data that would tell us which CNVs are real. We propose that if CNVs are called in duplicate samples, or inherited from parent to child, then these can be considered validated CNVs. We used two large family-based Genome-Wide Association Study (GWAS) datasets from the GENEVA consortium to look at concordance rates of CNV calls between duplicate samples, parent-child pairs, and unrelated pairs. Our goal was to make recommendations for ways to filter and use CNV calls in GWAS datasets that do not include family data. We used PennCNV as our primary CNV-calling algorithm, and tested CNV calls using different datasets and marker sets, and with various filters on CNVs and samples. Using the Illumina core HumanHap550 SNP (single nucleotide polymorphism) set, we saw duplicate concordance rates of approximately 55% and parent-child transmission rates of approximately 28% in our datasets. GC model adjustment and sample quality filtering had little effect on these reliability measures. Stratification on CNV size and DNA sample type did have some effect. Overall, our results show that it is probably not possible to find a CNV calling strategy (including filtering and algorithm) that will give us a set of "reliable" CNV calls using current chip technologies. But if we understand the error process, we can still use CNV calls appropriately in genetic association studies.

Key words: evaluation; CNV calling strategies; family-based GWAS.

2.2 INTRODUCTION

Most investigators performing genome-wide association studies (GWAS) would like to include association tests for CNVs, but low reliability of CNV calls has been a roadblock [Carter, 2007; Lai et al., 2005; Peiffer et al., 2006; Wineinger et al., 2008]. If a sample is genotyped twice, relatively different lists of CNVs can result, and this difference may be exacerbated if two different CNV-finding algorithms are used. Various factors are known to affect the reliability, most notably DNA quality and differences among CNV calling algorithms.

The holy grail of CNV calling for genetic association studies is a procedure that will produce "reliable" CNV calls - at least high specificity if not high sensitivity. (Note that this is somewhat different from the goals of CNV calling for clinical purposes, in which the relative value of sensitivity and specificity might be different). Such a procedure might in theory be achieved by a combination of algorithm choice, data pre-processing, sample filtering, marker sets, and CNV filtering. Typical applications currently in the literature filter by using only samples that have high quality by some metric and only CNVs of a certain length. But it has been very difficult to compare and validate such procedures because of the lack of gold-standard datasets in which CNVs have been molecularly validated. In the absence of datasets with known "right answers," the performance of a calling algorithm on real data cannot be quantified. While simulated data can be useful for this type of investigation, particularly in the early stages of algorithm development, we believe that there is no substitute for assessing performance on fully-complex real data, which is the goal of this study.

The premise of our study is that we can use family data as a substitute for molecular validation. If a CNV is called repeatedly in duplicate samples, or transmitted from parent to

child, then it can be considered validated. This type of validation is not 100% accurate and will not be appropriate for clinical use, but it is sufficient to allow us to estimate error rates and use those estimated error rates to compare CNV calling strategies. We use two large family-based GWAS datasets and compute CNV concordance rates for duplicate samples, parent-child pairs, and unrelated pairs. We then use the concordance rates to evaluate a variety of CNV calling and filtering strategies. Because previous studies have focused on comparing different software packages [Dellinger et al., 2010; Pinto et al., 2011], we focus instead on the role of filtering in CNV calling - which markers, samples, and CNV calls should be used. We primarily report results for the PennCNV package [Wang et al., 2007], which is generally acknowledged to be one of the best for the Illumina (San Diego, California) platform, although we also report some results for genoCN [Sun et al., 2009]. Our goal is to make recommendations for how CNV calls can be created and filtered for use in genetic association studies. Since these studies generally involve unrelated individuals, we do not focus on optimizing calls within families but rather we use our families to understand what the best filtering procedures are for individuals. A secondary goal is to contribute to the literature describing features and distributions of rare CNVs in the human genome.

Our study design is sketched graphically in Figure 1. From the GENEVA dental caries study (http://www.ncbi.nlm.nih.gov/gap?term=geneva), which is a large community-based study of oral health genotyped on the Illumina HumanHap610 chip, we selected 91 duplicate pairs and 752 father-mother-child trios. From the GENEVA preterm delivery study (http://www.ncbi.nlm.nih.gov/gap?term=geneva), we used almost all samples -1782 mother-child pairs genotyped on Illumina Human660W-Quad chip. Of these, 943 pairs were cases of pre-term delivery, 779 pairs were controls and the remaining 60 pairs were neither cases nor

controls. All subjects in the preterm delivery study were from the Danish National Birth Cohort. Since the chips used in these two studies share a core set of 550K SNPs, we started by calculating and comparing the CNV concordance rates in the two datasets using that shared SNP set. We then looked at the concordance rates for the full sets of SNP and CNV markers on each chip. We also looked at the effects of using PennCNV's GC adjustment and filtering out high-variability samples in the dental caries dataset only. Finally, we looked at subsets of data such as amplifications vs. deletions, common vs. rare CNVs, different CNV sizes, and different DNA sample types.

2.3 MATERIALS AND METHODS

2.3.1 Study Populations

Both the dental caries and preterm delivery datasets are part of the GENEVA consortium. In both datasets, GWAS data was used to verify all parent-child relationships. Detailed information on both studies is available from study documents in dbGAP (http://www.ncbi.nlm.nih.gov/gap) [Mailman et al., 2007]. The full dental caries study included four different community-based samples from Western Pennsylvania, West Virginia, and Iowa. Individuals were selected without regard to phenotype, and then were extensively phenotyped for oral health and related traits. We used a subset of the full study: 91 pairs of duplicate samples and 752 complete trio-family samples from two of the four recruitment sites. The pre-term delivery study is a case-control study within a cohort of approximately 1000 mother-child case pairs (cases were defined as

infants<37 weeks of gestation), and 1000 mother-child controls pairs (controls were defined as infants=40 weeks of gestation) from the Danish National Birth Cohort study [Olsen et al., 2001]. 1782 mother/child pairs with complete genotype information were used in this study.

2.3.2 Genotyping and Quality Control

Complete genotyping and data cleaning reports for both studies are available in dbGAP (http://www.ncbi.nlm.nih.gov/gap).The level of genotyping quality was extremely high.

2.3.3 CNV Calls by PennCNV

We generated CNV calls using the PennCNV software (2009Aug27 verion) [Wang et al., 2007]. Each sample was called individually, regardless of family relationships. PennCNV is a Hidden Markov Model (HMM) based method. It uses the log R ratio (LRR) and B allele frequency (BAF) measures computed from the signal intensity files by BeadStudio. To limit analyses to the core HumanHap550 (550K) marker sets in the HumanHap610 and Human660W-Quad chips, we used the hg18 (NCBI 36) "hh550" Population Frequency of B allele (PFB) file during CNV calling. For algorithms with GC model adjustment, we implemented the GC model wave adjustment procedure in PennCNV. For sample filtering, after GC model adjustment, we excluded samples meeting the criterion lrrsd > 0.3. All analysis was restricted to autosomes. The PennCNV trio-based CNV calling feature was not used, since we were interested in assessing quality of calls in individuals. PennCNV did not find any loss-of-heterozygosity in our samples.

2.3.4 CNV Calls by genoCN

We generated the CNV calls using the genoCN package (version genoCN 1.08) in R [Sun et al., 2009]. genoCN is also a HMM-based method using the log R ratio (LRR) and B allele frequency (BAF) measures from the signal intensity files by Beadstudio. Unlike PennCNV, which assumes that the mean value and SD of LRR and BAF for each HMM state are known, genoCN estimates HMM parameters from data. All procedures followed the user guidelines of genoCN.

2.3.5 Calculation of Overlap Quantities

Overlap quantities calculated include duplicate concordance, transmission and inheritance rates and unrelated pair concordance rates. “Overlap” of CNVs was defined as follows (both criteria must be met): a) the overlap length in base pairs is larger than 50% of the length in base pairs of the smaller size CNV, b) copy number (cn) state must be either both deletion or both amplification.

For concordance rate, in each sample pair, say sample A and sample B, we first used sample A as a "template" and counted how many CNV calls in sample A overlapped with ones in sample B. Then we used sample B as a "template" and counted how many overlapped with those called in sample A. We summed the numbers of overlapping CNVs in the two comparisons, and then divided it by sum of CNV calls in sample A and sample B. We restricted the maximum number of overlaps for each CNV in a template sample to one. For example, if a single CNV in sample A overlapped with two different CNVs in sample B, only one overlap was counted; this avoided overcounting of larger CNVs that were broken into smaller pieces by the

algorithm. For the dental caries dataset, the unrelated pair concordance rate was computed from father-mother pairs. For the preterm delivery dataset, the unrelated pair concordance rate was computed separately in mothers and children. The unrelated-pair concordance rate derived from children was highly consistent with the one from mothers, so only the rate from mothers was reported. The concordance rate in table 2-12 is also calculated in the same way.

For the transmission rate, in each parent-child pair, we used CNV calls in the parent as a "template," counted how many of them were also called in the child, and then divided by the total number of CNV calls in the parent. For the inheritance rate, we used CNV calls in the child as a "template," counted how many CNVs in the child overlapped with those in either parent, and then divided by the total number of CNVs in the child.

All concordance and transmission rates were calculated as the average over all pairs, so each pair contributed equally to the mean overlap rate and pairs with especially high or low rates were not excessively influential. All calculation was done in R (version 2.10.1) [R Development Core Team. 2009].

2.3.6 Stratification of CNV calls

Deletion vs. duplication CNVs were defined as CNV calls with cn <2 vs. >2 respectively. Common CNVs were defined as a frequency greater than 2%. Frequencies of CNVs were derived from unrelated individuals. Each CNV was compared with CNVs in other individuals; its frequency was defined as the overlap rate. We restricted the maximum number of overlaps from a pair of samples for each CNV to one.

2.4 RESULTS AND DISCUSSION

2.4.1 CNV Concordance Rates Using the Illumina HumanHap550 SNP Set

In an ideal dataset, duplicate concordance rates would be 100%, transmission rates would be 50%, and inheritance rates would be 100%. Several major factors, however, potentially cause datasets to deviate from this ideal. Most importantly, falsely detected CNVs will cause all of these rates to be below their ideal levels. Failure to detect CNVs (false negatives) will have a similar effect. For both false negatives and false positives, we should consider the possibility that the error is not random - that it could be repeated in duplicate samples or even in parent-child pairs because of sample or sequence similarity. A third important factor is de novo mutations in children, which will not affect duplicate concordance rates or transmission rates, but will affect inheritance rates. Finally, there is the possibility of somatic mutation with age (essentially de novo mutations in parents), which would affect apparent transmission rates but not inheritance or duplicate concordance rates. Because all of these factors are acting simultaneously, it is not possible to estimate them from this type of dataset, but some qualitative conclusions can be drawn, as discussed below, in particular if we are willing to assume that de novo mutations and somatic mutations are rare compared to CNV-calling errors.

The first column of Table 2-1 shows the results for the dental caries dataset using the common HumanHap550 SNP set. The average parent-child transmission rate is 28%, and the duplicate concordance rate is 55%. Father-child transmission rates and mother-child transmission rates are essentially identical. The fact that parent-child transmission rates are just about half of duplicate concordance rates suggests that we are probably not seeing repeated false calls in

duplicates due to sample issues – repeated calls in duplicates are likely to be real. The average unrelated pair concordance rate is 5%, which is presumably primarily accounted for by common CNVs, although a small amount of concordance by chance of rare CNVs and systematic error would also be included. Under simple but very conservative assumptions (such as that almost all CNV calls are false positives) we estimate a completely random concordance rate of about 0.3%. The fact that the inheritance rate of 42% is much less than twice the transmission rate implies that de novo CNVs may account for a non-ignorable proportion of the child CNVs. We note that the average child inheritance rate (42%) in our study is lower than what was reported by K. Wang et al [2007]. They examined "the fraction of CNVs inferred in offspring but not detected in parents (CNV-NDPs)", and found 25.2% of offspring CNVs from HumanHap550 were CNV-NDPs. This may due to differences in sample size, sample quality and sample populations. K. Wang et al. examined CNV-NDPs in the HapMap CEU + YRI offspring, which is a much smaller dataset.

The first column of Table 2-2 shows the corresponding results for the preterm delivery dataset. The concordance rates are very similar to those in the dental caries dataset: duplicate concordance rate 52%, mother-child transmission rate 26%, and unrelated concordance rate 4%. The highly consistent results imply that the findings from our study are not dataset specific and may be reasonably generalizable to other studies, at least for this marker set.

2.4.2 Addition of CNV Markers from the HumanHap610 Chip and the Human660W-Quad Chip

The HumanHap610 chip (dental caries study) and the Human660W Quad chip (pre-term delivery study) each consist of the HumanHap550 SNP set augmented by different sets of CNV probes. The second columns of Tables I and II give results for each study using the full chip for that study. For the HumanHap610 chip, the parent-child transmission rate and inheritance rate are similar to those from the HumanHap550 SNP set, but the average unrelated pair concordance rate is higher: 13% vs. 5%. One of the likely explanations is that the 60K additional CNV probes on the HumanHap610 chip contain more probes for common CNVs, and this is supported by evidence from later analyses (common vs. rare CNVs). Another noticeable difference is that the average duplicate concordance rate for HumanHap610 is much lower than for HumanHap550 (45% vs. 55%). This suggests quite poor performance of the CNV probes on this chip.

The full Human660W-Quad chip performs very differently than the HumanHap610, finding about seven times as many CNVs per sample. It also has much higher concordance and transmission rates, suggesting that the CNV probes have much better performance. The average duplicate concordance rate for the Human660W-Quad is 64%, as compared to 45% for the HumanHap610 and 55% for the HumanHap550. The average unrelated pair concordance rate for the Human660W-Quad is also much higher, 21%, suggesting that the CNV markers on the Human660W-Quad find many common CNVs. This is likely to also be the reason that the mother-child transmission rate is much higher than that on the HumanHap550 (38% vs. 26%).

2.4.3 GC Model Adjustment

PennCNV includes a GC model adjustment feature that adjusts the CNV calls to account for varying GC-content of the chromosome in different locations. Our analyses above included that adjustment, but the third column of Table 2-1 shows an analysis without the GC adjustment. Removing the adjustment increased the number of called CNVs and decreased the reliability measures (compare column 3 to column 2), but only very slightly. We conclude that the GC adjustment probably does improve quality, but does not make a major difference.

2.4.4 Sample Filtering

Column four of Table 2-1 shows an analysis in which we omitted the samples (about 13%) that had the PennCNV variability measure lrrsd (log R ratio standard deviation) greater than 0.3. As with the GC model adjustment, this improved reliability, but only very slightly. It is clearly a good idea in CNV analyses to omit poor-quality samples, but it appears that lrrsd might not be the most useful quality measure.

2.4.5 Deletion vs. Amplification CNVs

We used the dental caries dataset with the GC adjustment and the full HumanHap610 marker set to ask whether concordance rates differed for deletion and amplification CNVs. Table 2-3 shows the results. The number of deletion CNVs is 1.5~2 times that of amplification CNVs, but this does not necessarily reflect frequency in the human genome, since any given CNV calling

algorithm may have higher sensitivity to either deletions or amplifications. It is interesting to note that while duplicate concordance and parent-child transmission rates are higher for deletions, the inheritance rate (percent of the child's CNVs that are inherited from parents) is higher for amplifications. It is possible that this means that de novo deletions are more common in viable offspring than de novo amplifications, but that would clearly merit further investigation.

2.4.6 Common vs. Rare CNVs

Again using the dental caries dataset with the GC adjustment and the full HumanHap610 marker set, we asked whether concordance rates differed for rare and common CNVs. Parent-child and unrelated-pair concordance rates are clearly expected to be higher for common CNVs because of chance matching, but duplicate concordance rates should not be different for rare and common CNVs if the algorithm is equally good at finding both. However, one of the concerns in CNV calling is that common CNVs can be hard to detect, since the deviation of the log R ratio between case and reference is small after normalization.

Results are given in Table 2-4. For the purposes of this analysis we arbitrarily considered a CNV to be common if it occurred in 2% or more of the sample. The concordance rates in unrelated pairs are 3% for rare CNVs and 19% for common CNVs, which confirms that most of the concordance between unrelated individuals is due to CNVs that are common in the population. This may also explain the higher transmission rate in common CNVs than rare ones (32% vs. 20%). The finding that the average duplicate concordance rate in common CNVs is

higher than in rare ones (51% vs. 44%) suggests that PennCNV does not in fact have more difficulty detecting common CNVs.

2.4.7 Samples With High CNV Number

It might be logical to presume that samples with very high CNV numbers are of low quality and that the CNVs called in those samples are not real. To investigate this, we plotted CNV number vs. concordance rate using the dental caries dataset (HumanHap 610 marker set) in Figure 2-2. In general, the concordance/transmission rates tend to decrease with the number of CNV calls, but there are clearly some pairs that have high concordance and/or transmission rates even with more than 100 CNVs. This suggests that while it might be advisable to filter samples with very high numbers of CNV calls out of association studies, there are in fact some individuals who do carry high numbers of real CNVs.

2.4.8 CNV Size

It is often assumed that calls of larger CNVs are more likely to be accurate, and our results using the dental caries dataset (HumanHap 610 marker set) (Table 2-5) support that. We measured the "size" of the CNV by the number of markers rather than the physical length, and found that the shortest CNVs (3 - 5 markers) had only an 18% mean parent-child transmission rate, while the longest (> 54 markers) had a 42% mean parent-child transmission rate. This is a substantial difference, but it is not substantial enough to justify filtering out the smallest CNVs or to justify

assuming that the largest ones are necessarily correct. Thus while our results support the common wisdom, they do not suggest a workable filtering strategy for association studies.

2.4.9 DNA Source

Next, we compared the reliability of CNV calls for different DNA sample types. The dental caries study includes samples from blood, saliva, mouthwash and buccal swabs. To investigate the effect of sample type, we compared transmission rates in pairs with different combinations of sample types using the HumanHap 610 marker set. The resulting transmission rates in the dental caries dataset are shown in Table 2-6. There is no detectable difference in reliability between the blood and saliva samples. The comparative reliability of mouthwash is not conclusive due to the small sample size. A limitation of the results shown in Table 2-6 is that because all parental samples are either blood or saliva, it is difficult to tell from transmission rates if the other sample types are more error-prone. That is, child samples with more false CNVs due to poor DNA quality may not necessarily show lower transmission rates if the true CNVs were also called. Thus in Table 2-7 we also show the number of CNVs called per sample by sample type. The mouthwash, buccal and WGA samples do have significantly more CNVs called per person, and we conclude that it is likely that they have higher false-positive rates than the blood and saliva samples. It is also intriguing that there is a higher number of CNVs per person in the children's saliva and mouthwash samples than in their parents. It appears that children may in general be producing lower-quality saliva samples than adults. By contrast, we saw comparable CNV numbers from blood samples in children and parents in both the dental caries (Table 2-7) and preterm delivery (Table 2-8) datasets.

In the pre-term delivery study most maternal samples were from buffy coat, but a substantial proportion of the infant samples were from dried blood spots. Some of the buffy coat samples and some of the blood spot samples were whole-genome amplified (WGA). Mother-child transmission rates calculated using the Human660W-Quad marker set are listed in Table 2-9 according to the sample type for both mother and child. When both mother and child are buffy coat (no WGA), the transmission rate is 40%. If the child is WGA, it only drops to 31%, so we can infer that most of the real CNVs are still being found in buffy coat WGA. However, when the mother is WGA transmission percentages drop substantially, from which we can infer that the WGA samples are giving us many spurious CNV calls in addition to the real ones. These findings suggest that the WGA samples give us reasonable sensitivity, but very poor specificity.

2.4.10 Age

Another interesting question is that of somatic mutations with age. Using parents only from the two studies separately, we regressed the log of the number of CNVs on the age of the individual. After excluding a few extreme outliers, we found a very small but statistically significant increase in the number of CNVs with age (Figure 2-3). This finding is consistent with suggestions made in previous studies [Martin et al., 1996; Maslov and Vijg, 2009].

2.4.11 Comparison to genoCN

Finally, in order to compare with the performance of PennCNV, we conducted a limited study using another algorithm - genoCN [Sun et al., 2009]. We chose two pairs of duplicates and two

pairs of unrelated samples at random (from samples with lrrsd<0.3) and tested them using genoCN (Tables 2-10 and 2-11). genoCN detected 4~20 fold more CNVs than PennCNV (with GCmodel and HumanHap610 markers). Most of the CNVs called by PennCNV were also called by genoCN (table 2-9), but the duplicate concordance rates for the genoCN calls were much lower than those for PennCNV. From this we can infer that genoCN may give a lot of spurious CNV calls in addition to the real ones (reasonable sensitivity but poor specificity). We also observed that the unrelated pair concordance rates in genoCN were lower than in PennCNV, presumably also due to low specificity.

An additional problem that we observed in both algorithms was that the largest CNVs were not "called" as single units, but were broken into several reported smaller CNVs. This problem was worse in genoCN than in PennCNV.

2.5 CONCLUSIONS

In summary, CNV association studies have been of great interest lately, but a key problem is how to identify a set of reliable CNVs. Molecular validation is not feasible for GWAS-sized datasets, and without gold-standard data it has been quite difficult to compare CNV calling algorithms to make recommendations for the best ones to use in association studies. We have taken advantage of two large family-based GWAS studies to use inheritance as a substitute for molecular validation and ask questions about what kind of sample, SNP, and CNV filtering leads to the most reliable CNV calls. While many authors have previously reported concordance rates

for CNV calls in duplicate samples, we hope that by also looking at very large samples of parent-child pairs we have added depth to that picture.

We found several classes of samples that clearly have low reliability and should be filtered out of CNV association studies, including any with whole-genome amplification and any with excessive numbers of called CNVs. These results are quite concordant with conclusions of previous authors. But filtering out these samples did not result in high reliability rates in the remaining samples, an issue that we believe has not received adequate attention previously. The most prognostic variable we looked at was CNV size, but even that did not guarantee high reliability for large CNVs nor low reliability for small CNVs. Thus we suggest that the common strategy of using only the largest CNV calls and assuming they are correct is excessively crude and probably quite detrimental to statistical power.

Overall, we conclude from our data that it is probably not possible to find a CNV calling strategy that will give us a set of "reliable" CNV calls using current chip technologies. For now, CNV calls will need to be understood as having high error rates. But if we understand and model the features of that error process, we can still use them appropriately in genetic association studies. In particular, the most critical issue will be to make sure that cases and controls are well matched on any features that we know affect CNV call reliability rates, such as DNA sample type.

We also made some contributions to the growing picture of what “normal” variability in copy number means for the human genome. In particular, we found a subset of individuals who carry a fairly high load of rare CNVs (100 or more) that appear from inheritance rates to be real. We also found a modest increase in the number of CNVs with age, suggesting a non-trivial rate of somatic mutation, although this clearly bears further study. Finally, we found some intriguing

results related to the relative inheritance rates of deletions vs. amplifications, which would be interesting to follow up further.

2.6 ACKNOWLEDGMENTS

The work of XZ was supported by T32MH015169. The work of JRS, MLM, and EF was supported by U01DE018903 and U01HG004423. The work of BF, MM, and JCM was supported by U01HG004423. The work of CPM and CCL was supported by U01HG004446. Genotyping was performed by the Johns Hopkins University (JHU) Center for Inherited Disease Research (CIDR) through contract HHSN268200782096C. Dental caries subjects were collected by the Center for Oral Health Research in Appalachia (PI M. Marazita, a collaboration of the University of Pittsburgh and West Virginia University funded by NIDCR R01-DE 014899) and the Iowa Fluoride Study and the Iowa Bone Development Study (PI S. Levy), funded by NIDCR R01-DE09551 and R01-DE12101, respectively). Pre-term birth subjects were a part of the Danish National Birth Cohort (DNBC), which was established with the support of a major grant from the Danish National Research Foundation. Additional support for the DNBC has been provided by the Danish Pharmacist's Fund, the Egmont Foundation, the March of Dimes Birth Defects Foundation, The Augustinus Foundation, and the Health Fund of the Danish Health Insurance Societies.

2.7 REFERENCES

Carter NP. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39: S16–S21.

Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 38: e105.

Lai WR, Johnson MD, Kucherlapati R, Park PJ. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21: 3763-3770.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181-1186.

Martin GM, Ogburn CE, Colgin LM, Gown AM, Edland SD, Monnat RJ Jr. 1996. Somatic Mutations Are Frequent and Increase with Age in Human Kidney Epithelial Cells. Hum Mol Genet 5: 215- 221.

Maslov AY, Vijg J. 2009. Genome instability, cancer and aging. Biochim Biophys Acta 1790: 963-969.

Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Andersen AM, Taxbøl D, Hansen KD, Juhl M, Schow TB, Sørensen HT, Andresen J, Mortensen EL, Olesen AW, Søndergaard C. 2001. The Danish National Birth Cohort. Its background, structure and aim. Scand J Public Health 29: 300-307.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136-1148.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nature Biotechnology 29: 512-521.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, Yu T, Kristensen VN, Perou CM. 2009. Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res 37: 5365-5377.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

Wineinger NE, Kennedy RE, Erickson SW, Wojczynski MK, Bruder CE, Tiwari HK. 2008. Statistical issues in the analysis of DNA Copy Number Variations. Int J Comput Biol Drug Des 1: 368–395.

2.8 TABLES AND FIGURES

Table 2-1. Mean overlap quantities (± SEM) in the dental caries dataset.

	HumanHap550	HumanHap610		
	GC*	GC	non-GC	GC+filtering**
Num of duplicate samples (avg. CNV/sample)***	182 (18.6)	182 (79.4)	182 (61.1)	162 (54.1)
Num of non-duplicate samples(avg. CNV/sample)	1736 (26.6)	1736 (68.5)	1736 (92.9)	1512 (54.9)
Duplicate concordance rate	0.55(±0.02)	0.45(±0.02)	0.43(±0.02)	0.48(±0.02)
Unrelated pair concordance rate****	0.05(±0.003)	0.13(±0.004)	0.11(±0.004)	0.14(±0.004)
Father-child transmission rate	0.28(±0.006)	0.28(±0.005)	0.27(±0.005)	0.31(±0.005)
Mother-child transmission rate	0.28(±0.006)	0.27(±0.005)	0.26(±0.005)	0.31(±0.005)
Child inheritance rate	0.42(±0.009)	0.40(±0.008)	0.36(±0.008)	0.45(±0.008)

All mean overlap quantities were calculated as the average over pairs.
*GC model adjustment procedure in PennCNV.
** Samples were filtered by the criterion: LRR standard deviation (sd) > 0.3.
*** Number of duplicate samples (average number of CNVs per sample).
**** The unrelated pair concordance rate was calculated among father-mother pairs.

Table 2-2. Mean overlap quantities (± SEM) in the preterm delivery dataset.

	Hap550	Human660W-Quad
	GC	GC
Num of dup samples (avg. CNV/sample)	40 (21)	40 (383)
Num of non-dup samples(avg. CNV/sample)	3564 (48.8)	3564 (438.6)
Duplicate concordance rate	0.52 (±0.06)	0.64(±0.02)
Unrelated pair concordance rate*	0.04(±0.002)	0.21(±0.002)
Mother-child transmission rate	0.26 (±0.004)	0.38(±0.002)

* The unrelated pair concordance rate was derived from mothers. The rate derived from children was very similar.

Table 2-3. Mean overlap quantities (±SEM) in deletion vs. amplification CNVs.

	Deletion	**Amplification**
Avg CNVs* in dup samples/person	36.1	24.9
Avg CNVs in non-dup samples/person	46.6	21.9
Duplicate concordance rate	0.51 (±0.02)	0.40 (±0.02)
Unrelated pair concordance rate	0.11 (±0.004)	0.14 (±0.007)
Father-child transmission rate	0.32 (±0.006)	0.25 (±0.007)
Mother-child transmission rate	0.30 (±0.006)	0.27 (±0.007)
Child inheritance rate	0.41 (±0.009)	0.46 (±0.009)

***** Avg CNVs: average number of CNVs

Table 2-4. Mean overlap quantities (± SEM) in common vs. rare CNVs.

	Common	**Rare**
Avg CNVs in dup samples/person *	41.7	19.4
Avg CNVs in non-dup samples/person **	34	23.9
Duplicate concordance rate	0.51 (±0.03)	0.44 (±0.04)
Unrelated pair concordance rate	0.19 (±0.006)	0.03 (±0.003)
Father-child transmission rate	0.32 (±0.006)	0.20 (±0.007)
Mother-child transmission rate	0.31 (±0.006)	0.21 (±0.007)

Only CNVs from unrelated subjects were used to infer the common vs. rare CNVs.
* Total sample number of duplicates was 66.
** Total sample number of non-duplicate subjects was 984.

Table 2-5. Mean overlap quantities (± SEM) by size of CNV call.

	Size of CNV call					
	3-5 SNPs	**6-10 SNPs**	**11-22 SNPs**	**23-54 SNPs**	**> 54 SNPs**	**All**
Avg num of CNVs/person*	13	19.2	18.6	13.4	3.5	67.6
Duplicate concordance rate	0.31 (±0.02)	0.49 (±0.02)	0.48 (±0.02)	0.55 (±0.03)	0.64 (±0.04)	0.45 (±0.02)
Unrelated pair concordance rate	0.07 (±0.004)	0.11 (±0.004)	0.12 (±0.005)	0.23 (±0.008)	0.16 (±0.01)	0.13 (±0.004)
Father-child transmission rate	0.18 (±0.007)	0.29 (±0.006)	0.30 (±0.008)	0.41 (±0.01)	0.42 (±0.02)	0.28 (±0.005)
Mother-child transmission rate	0.19 (±0.007)	0.27 (±0.006)	0.30 (±0.008)	0.40 (±0.01)	0.43 (±0.02)	0.27 (±0.005)

*In all 1736 non-duplicate samples.

Table 2-6. Mean parent-child transmission rate (± SEM) by sample type in the dental caries dataset.

Sample sources		**Num of sample pairs**	**Father-child transmission**	**Mother-child transmission**
Parent	**Child**			
Mouthwash	Mouthwash	9	0.22 (±0.05)	0.34 (±0.05)
Saliva	Saliva	98	0.27 (±0.01)	0.29 (±0.01)
Blood	Blood	349	0.28 (±0.008)	0.27 (±0.007)
Blood	Buccal	89	0.24 (±0.01)	0.23(±0.01)
Blood	Saliva	51	0.30 (±0.02)	0.26 (±0.02)
Blood	Mouthwash	40	0.32 (±0.02)	0.32 (±0.02)
Blood	WGA	10	0.32(±0.06)	0.32(±0.06)

Table 2-7. Mean number of CNVs called per sample (± SEM) by sample type in the dental caries dataset.

	Child		Father		Mother	
	sample num	CNV num/sample	sample num	CNV num/sample	sample num	CNV num/sample
Blood	349	61.5±3.3	539	55.9±2.1	539	55.7±1.9
Saliva	150	82.0±6.7	98	58.4±3.8	98	51.1±3.8
Mouthwash	49	120.7±23.1	9	71.8±19.4	9	49±11.0
Buccal	89	109.8±9.3				
WGA	10	153.6±38.4				

Table 2-8. Mean number of CNVs called per sample (± SEM) in buffy coat blood samples in the preterm delivery dataset.

	Child		Mother	
	sample num	CNV num/sample	sample num	CNV num/sample
Buffy coat	1257	379.2±2.9	1257	380.8±2.3

Table 2-9. Mean mother-child transmission rate (± SEM) by sample type in the preterm delivery dataset.

Sample source		Num of sample pairs	Mother-child transmission
mother	child		
Buffy coat	Buffy coat	1347	0.40 (±0.003)
Buffy coat	Blood spot	346	0.35 (±0.004)
Buffy coat	Buffy coat WGA	52	0.31 (±0.008)
Buffy coat WGA	Buffy coat	13	0.06 (±0.006)
Buffy coat WGA	Blood spot	18	0.14 (±0.01)
All		1782	0.38 (±0.002)

Table 2-10. Comparison of duplicate concordance rate among two pairs of duplicate samples from the dental caries dataset using PennCNV and genoCN.

	Subject ids		Num of CNVs		Num of concordant CNV	Duplicate concordance rate
	dup1	dup2	dup1	dup2		
PennCNV	175040850	9942	88	23	16	0.28
	175043297	9950	37	37	24	0.65
genoCN	175040850	9942	540	616	67	0.12
	175043297	9950	129	101	52	0.45

Table 2-11. Unrelated concordance rate among two pairs of father-mother samples from the dental caries dataset using PennCNV and genoCN.

	Subjects id		Num of CNVs		Num of concordant CNV	Unrelated concordance rate
	father	mother	father	mother		
PennCNV	175192256	175049618	64	63	7	0.11
	175133191	175097605	58	51	10	0.18
genoCN	175192256	175049618	195	507	21	0.06
	175133191	175097605	322	304	44	0.14

Table 2-12. Concordance rate between PennCNV and genoCN for each person.

Subject id	Num of CNVs/Sample		concordance rate
	PennCNV	genoCN	
175040850	88	540	0.19
9942	23	616	0.07
175043297	37	129	0.29
9950	37	101	0.33
175192256	64	195	0.52
175049618	63	507	0.19
175133191	58	322	0.26
175097605	51	304	0.25

Legends for figures

Figure 2-1. Summary of study design.

Figure 2-2. Relationship between number of CNV calls per sample and concordance rate in the dental caries dataset.

* All x–axes are log scale.

(A) shows relation between average number of CNV calls per pair of duplicate samples and duplicate concordance rate. (B) shows relation between average number of CNV calls per pair of unrelated samples and unrelated concordance rate. (C) shows relation between number of CNV calls in each father and father-child transmission rate. (D) shows relation between number of CNV calls in each mother and mother-child transmission rate.

Figure 2-3. Relationship between age and number of CNV calls in each adult (log scale) in the dental caries dataset.

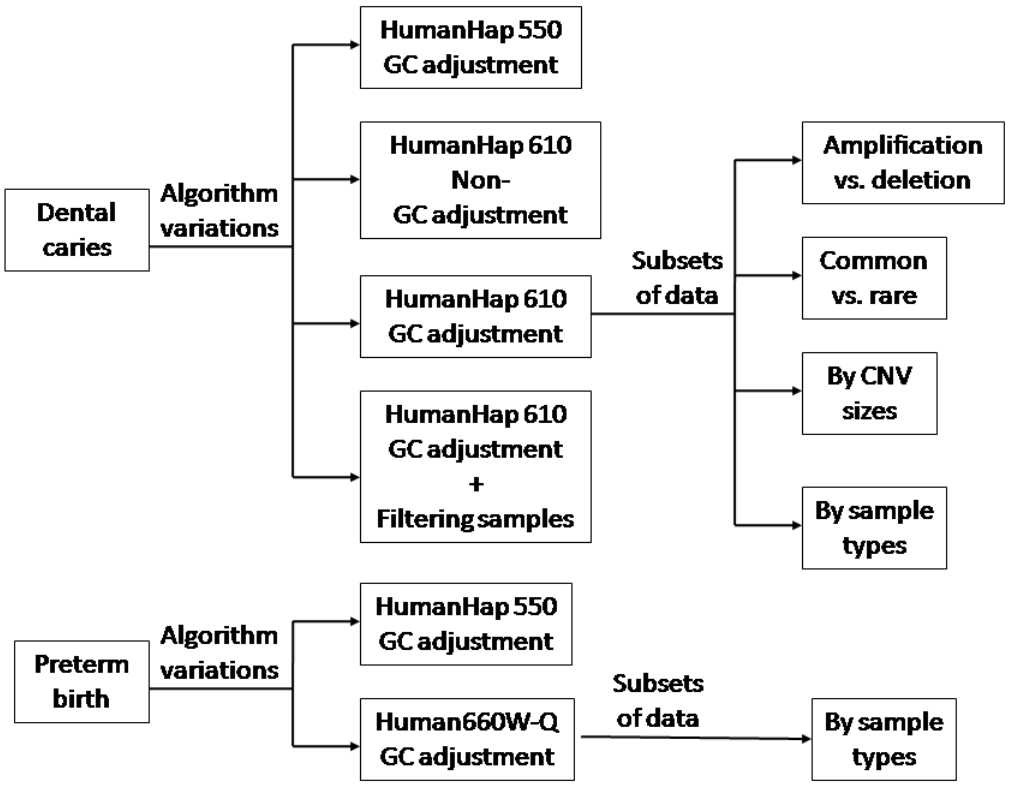


Figure 2-1. Summary of study design.

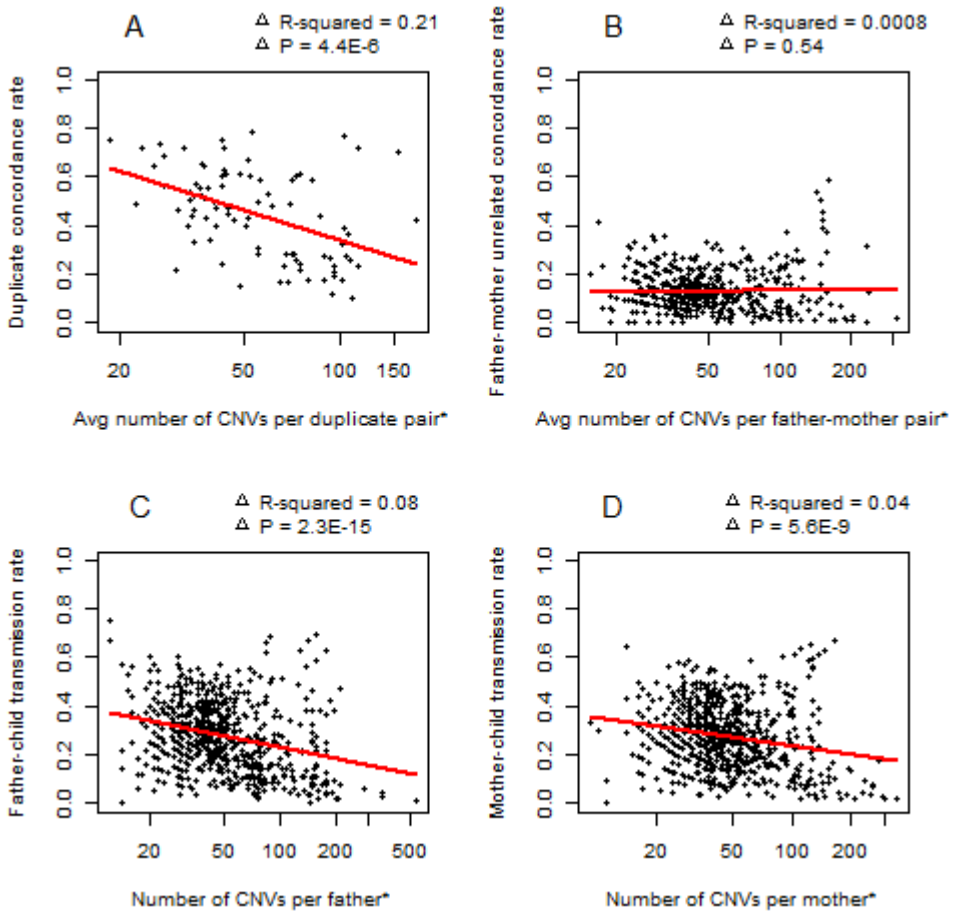


Figure 2-2. Relationship between number of CNV calls per sample and concordance rate in the dental caries dataset.

***** All x–axes are log scale. (A) shows relation between average number of CNV calls per pair of duplicate samples and duplicate concordance rate. (B) shows relation between average number of CNV calls per pair of unrelated samples and unrelated concordance rate. (C) shows relation between number of CNV calls in each father and father-child transmission rate. (D) shows relation between number of CNV calls in each mother and mother-child transmission rate.

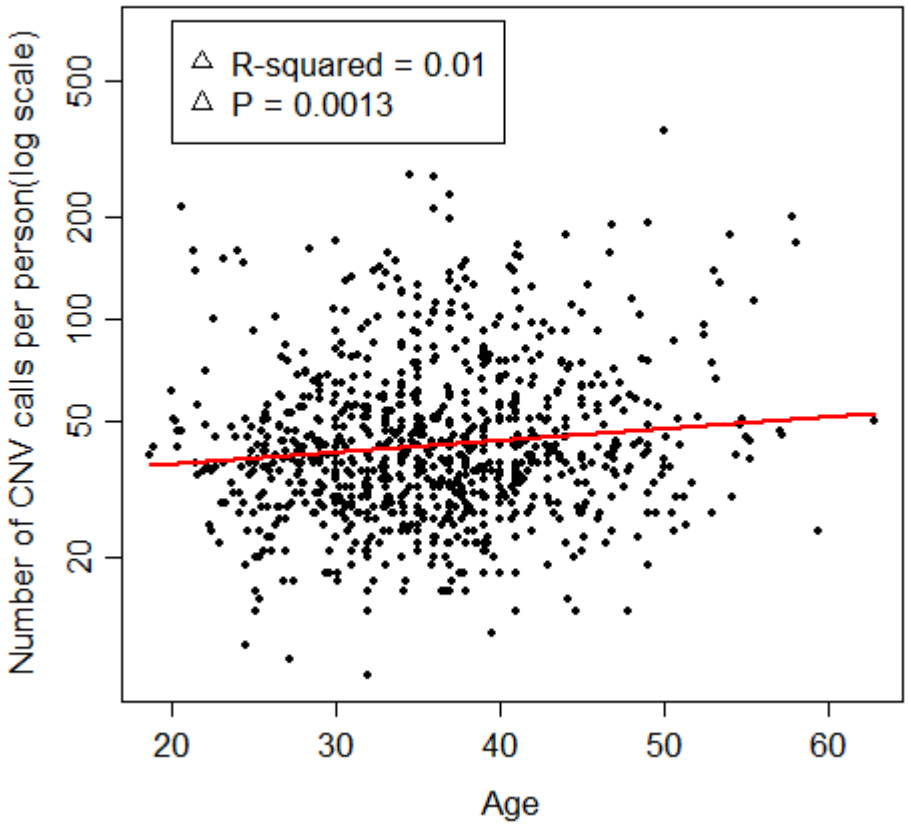


Figure 2-3. Relationship between age and number of CNV calls in each adult (log scale) in the dental caries dataset.

3.0 DNA COPY NUMBER VARIANTS LINKED TO AUTISM AND SCHIZOPHRENIA ARE ALSO ASSOCIATED WITH PSYCHOSIS IN ALZHEIMER DISEASE

Xiaojing Zheng [1], M. Ilyas Kamboh [2], M. Michael Barmada [2], Robert A. Sweet [3], F. Yesim Demirci [2], Eleanor Feingold [1,2] .

1. Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA

2. Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA

3. Alzheimer's Disease Research Center, University of Pittsburgh, Pittsburgh, PA

3.1 ABSTRACT

Objectives: Alzheimer disease (AD) is the most common form of dementia. 40-60% of AD patients develop psychosis (AD+P), a subtype of disease with more deleterious outcomes. AD+P is highly heritable with unknown genetic etiology. It has been proposed that schizophrenia, autism and AD+P may share risk genes; however conflicting results have been reported from SNP association studies. Recent evidence showed that schizophrenia and autism share some rare

copy number variants (CNVs). However it is unknown whether those CNVs are also linked to AD+P.

Methods: We searched for CNVs associated with AD+P in 7 CNV regions that have been previously identified across autism and schizophrenia, using the Illumina Omni1-Quad SNP array. We also screen the rest of genome by genome-wide association study to nominate candidate CNVs for AD+P. CNVs were inferred using PennCNV.

Results: A 16p11.2 duplication CNV was identified in 2 of 440 AD+P subjects, but not in 136 AD without psychosis, 593 AD with intermediate psychosis, or 855 non-AD individuals. The frequency of this duplication CNV in AD+P is similar to that in schizophrenia. We also found a small CNV on 3q29 which is within *PAK2*, one of the most interesting candidate genes for schizophrenia in 3q29. We did not find meaningful CNVs in the other five reported CNV regions. In the rest of the whole genome, we did not find any CNV that reaches genome-wide significance. The CNVs that are in the top most significant association list are all common ones.

Conclusions: In conclusion, we are the first to report that AD+P shares rare risk CNVs on 16p11.2 and 3q29 with schizophrenia and autism. Although rare, these CNVs may have important functions in the development of psychosis.

3.2 INTRODUCTION

Alzheimer disease (AD) is the most common form of dementia. About 4.5 million (12.8%) people age 65 years or older in US have AD [Hebert et al, 2003]. Although the incidence of AD

increases with age, AD is not a process of normal aging. AD patients have typical pathological changes in the brain.

Family and twin studies have shown that genetic factors play an important role in AD [Kauwe et al, 2007 and Bergen et al, 1997]. Depending on age of onset, AD can be divided into two subgroups: early onset AD (EOAD) and late onset AD (LOAD) with different genetic etiology. EOAD occurs typically before age 60 years. It is rare, and accounts for less than 1% of all AD cases [Campion yet al, 1999]. EOAD is a familial AD with autosomal dominant transmission. Mutations with complete penetrance in three genes *AβPP* (amyloid-β protein precursor), *PSEN1* (presenilin-1) and *PSEN2* (presenilin-2) have largely contributed to EOAD [Goate et al, 1991; Sherrington et al, 1995; Rogaev et al, 1995; De Strooper et al, 1998]. However, none of the three genes is significantly associated with LOAD. LOAD is sporadic and not inherited in Mendelian fashion; but it is highly inherited with heritability up to 79% [Gatz et al, 2006]. Currently, only *ApoE* (on chromosome 19) has been consistently reported to be associated with LOAD [Goedert et al, 2006; Khachaturian et al, 2004; Corder et al, 1993; Saunders et al, 1993]. The majority of the genes related with LOAD have not been identified.

40-60% of LOAD patients develop psychosis (AD+P) [Sweet RA et al, 2000; Farber NB et al, 2000; Forstl H et al, 1994]. Primary features of psychosis are delusions and hallucinations [Burns et al, 1990; DeMichele-Sweet et al, 2010]. Delusions are false beliefs held that are not consistent with reality; hallucinations are false perceptions without a stimulus - visual or auditory-based delusions. Although AD+P is less likely to be an early symptom of AD or secondary to more severe AD [Lyketsos CG 2000; Paulsen JS 2000; Sweet RA 2000； Jeste DV 1992； Ballard CG 1997], it may connect with cognitive decline and serve as a marker for more severe cognitive dysfunction [Paulsen JS 2000, Rockwell E 1994].

Familial aggregation studies strongly suggested the role of genetic factors in AD+P [Bacanu et al, 2005; Sweet et al, 2002; Tustall et al, 2000]. Sweet [Sweet et al, 2003] proposed that AD+P represents a distinct subtype with deleterious outcome and more homogeneous genetic etiology. Linkage and association studies identified some susceptibility loci (chromosome 2p, 6q, 8p, 7, 15, and 21) and candidate genes; however, none of them have been consistently associated with AD+P. Many studies have investigated the *APOE* ε4 allele, the well documented risk factor for LOAD, but most of them found no evidence for an effect of *APOE* ε4 on AD+P as compared to AD-P [Lopez et al, 1997].

It is of great interest that some susceptibility loci and candidate genes for AD+P have been shown to increase the risk of schizophrenia (SCZ). For example, *NRG1* (neuregulin1) on 8p and *CHRNA7* (cholinergic receptor, nicotinic, alpha 7) on chromosome 15 have been found to be linked to and associated with SCZ and AD+P [Go et al, 2005; Carson et al, 2008], although the association of mutations in these two genes with AD+P has not been confirmed by replicate studies yet. Sweet et al [2003] also found that AD+P was associated with similar biological changes in specific brain areas to those in individuals with idiopathic psychosis of SCZ. Recent studies [Prestia 2011; Horesh et al, 2011] compared gene expression profiles in cases vs. controls of AD and SCZ, and reported that AD and SCZ may share certain molecular background. Actually the shared susceptibility loci between AD and other psychiatric disorders (SCZ, bipolar disorder, alcoholism) were observed by Zubenko over 10 years ago [2000], however, the underlying mechanisms for the pleiotropic effects is unclear. Sweet et al [2003] have hypothesized three possible pathways to AD and AD+P: 1) some shared genes modify the course of neurodevelopment disturbances (eg. in SCZ) and process of neurodegenerative illness (eg. AD), and increase the risk of psychosis in those disorders. These modifier genes would

affect individuals who already have AD, and are therefore distinct from those that contribute to the risk for AD. 2) AD+P and AD have their origins in the same genetic mechanisms; 3) Genetic factors attributed to SCZ also lead to AD+P.

If the findings about association of SNPs in AD+P candidate genes (eg. *NRG1* and *CHRNA7*) with AD+P and SCZ can be confirmed, it will be in support of the first pathway; but a roadblock to making any conclusive inference on this pathway is the inconsistent results from SNP association studies. This may due to small sample size, varied phenotypic definition of AD+P across studies, population stratification and allelic heterogeneity.

Compared to SNP association analysis, CNV (copy number variation) studies in psychiatric disorders have achieved significant progress in the past a few years. A CNV is a structural variation, which can be a deletion or duplication of a segment of DNA (size 1kb ~ several Mb). Heinzen et al recently [2010] conducted the first genome-wide scan of CNVs in LOAD in 331 LOAD cases and 368 controls. Although nothing was statistically significant in this study, they found an interesting rare duplication CNV in *CHRAN7*, which has much higher frequency (2%) in cases than in controls (0.3%). However no one has done genome-wide CNV analysis in AD+P. Cumulative evidence shows that rare CNVs (freq<1%) may be more important in behavior disorders, such as SCZ and autism, than common CNVs. Current studies in autism and SCZ have shown that autism and SCZ share several rare CNVs. Some of these are also shared with various intellectual disability (ID) syndromes. For example, 3q29 deletion was identified in children with ID and autism, as well as in adults with SCZ [Quintero-Rivera et al, 2010]. Moreno-De-Luca, et al. recently summarized the CNV studies in autism and SCZ, and reported 7 recurrent CNVs across autism and SCZ [Moreno-De-Luca et al, 2010], which are located in chr1q21.1, chr3q29, chr15q13.3, chr16p11.2, chr16p13.11, chr17q12, and chr22q11.2

respectively (Table 3-1). The 7 CNVs share two common features: large size and rare frequency. Each of them contains several tens to hundreds of genes; and the frequencies in cases and controls may be less than 0.5%, and 0.05% respectively.

These CNV findings lead us to hypothesize that distinct psychiatric disorders may be caused by the same or similar genetic variants, perhaps influenced by different environmental modifiers. We therefore designed this study to examine whether AD+P shares risk CNVs with autism and SCZ. We specifically searched for CNVs for AD+P in the above 7 reported shared CNV regions across autism and SCZ, using the Illumina Omni1-Quad SNP array. We also screen the rest of genome by genome-wide association study to nominate new candidate CNVs for AD+P.

Table 3-1. 7 recurrent CNVs across ASD and SCZ reported by Moreno-De-Luca et al.

Chr	Chromosome Regions	CNV	CNV Starting Position(bp)[a]	CNV Ending Position(bp)
1	q21.1	Deletion	144,963,73	145,864,377
3	q29	Deletion	197,244,288	198,830,238
15	q13.3	Deletion	28,698,632	30,234,007
16	p11.2	Duplication	29,557,553	30,107,434
16	p13.11	Duplication	15,421,876	16,200,195
17	q12	Deletion	31,893,783	33,277,865
22	q11.2	Deletion	17,412,646	19,797,314

[a] Human genome assembly build 36/hg18.

3.3 MATERIALS AND METHODS

3.3.1 Study Populations

AD cases and controls were recruited through the University of Pittsburgh Alzheimer's Disease Research Center. Controls met the criteria for being free of dementia using the Mini Mental State Exam and the Alzheimer's Disease Assessment Cognitive Scale. Cases were those with diagonosis of either probable or definite AD according to criteria set by the National Institute of Neurological and Communicative Disorders and Stroke – Alzheimer's Disease and Related Disorders Association and the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). All cases have age of onset of at least 60 years. AD+P was ascertained if any of the CERAD behavior rating scale items for psychotic features were rated as occurring three or more times in the past month at any visit. AD-P subjects were defined as scores of zero on the same items at all visits. Subjects with scores in between 0 and 3 on those items were classified as "indeterminate psychosis". Exclusion criteria included previous history of SCZ, mood disorders, bipolar disease, unipolar disease, or anxiety disorder.

3.3.2 CNV Calling

DNAs from all subjects were genotyped using the Illumina Omni-Quad array. 2249 samples were retained after quality control. All samples with missing genotype rates >= 0.02 were removed from the study. All of this cleaning work was done as a part of the original GWAS study before we obtained the data for CNV analysis.

We generated CNV calls using the PennCNV software (2009Aug27 verion) [Wang et al., 2007]. PennCNV is a Hidden Markov Model (HMM) based method. It uses the log R ratio (LRR) and B allele frequency (BAF) measures computed from the signal intensity files by Beadstudio to detect the CNVs. We used the GC model wave adjustment procedure in PennCNV. After GC model adjustment, we filtered the samples that met the criterion of LRR standard deviation ≥ 0.3 (Table 3-2). All procedures followed the user guidelines of PennCNV and those developed in Chapter 2 of this dissertation. Human NCBI Build 36 (hg18) was used for this study.

3.3.3 Statistical Analysis of CNVs

CNVs with copy number >2 were defined as duplications; while those with copy number <2 were considered deletions. We conducted genome-wide association analysis of AD+P with CNVs using logistic regression. Amplification and deletion CNVs were coded as dummy variables with normal copy number as the reference. The regression model is: logit (Y) = α + $\beta_1 X_1 + \beta_2 X_2$. X_1 = 1 if and only if copy number <2, otherwise equals zero; X_2 = 1 if and only if copy number > 2, otherwise equals zero. We adjusted for the covariates age, sex, and principle components of population stratification in the regression model. We also specifically searched for rare CNVs in the 7 recurrent CNVs identified across ASD and SCZ summarized by Moreno-De-Luca et al. The criterion for rare CNVs in this search was the occurrence of the deletion and duplication CNV in AD-P and no-AD controls ≤ 1 at each of three or more consecutive markers. We used the Cochran–Armitage test for trend in SAS (version 9.2) [SAS Institute Inc., Cary,

NC] to calculate the exact permutation p-value for a trend of the three AD groups (AD-P, AD indeterminate P and AD+P).

Table 3-2. Sample sizes for each study group before and after filtering by LRR deviation.

	AD+P	AD (intermediate P)	AD-P	No-AD controls
Before filtering	496	639	156	958
After filtering	440	593	136	855

3.4 RESULTS

3.4.1 Genome-Wide Association Analysis

Using logistic regression, we calculated the p-value for the regression coefficient beta2 in the above model (amplification CNVs) at each marker; then summarized the genome-wide results of p-values in Figure 3-1. The top panel of Figure 3-1 is a Manhattan plot, which shows the -log10 (p-value) from the genome-wide scan. The bottom panel of Figure 3-1 is a Q-Q plot, which provides a graphical view of how observed p-values and expected p-values are similarly or differently distributed. Similarly, we summarized the genome-wide p-values for the regression coefficients beta2 (deletion CNVs) in Figure 3-2 and for the whole model (dummy variables-amplification and deletion CNVs as a set) in Figure 3-3 respectively.

Figure 3-1 shows that all of the amplification CNVs have p values > $e10^{-6}$, which suggests that none of them may be significant after adjustment of multiple comparisons genome-wide. However, the tests at each marker are highly correlated (much more so than in a genome-wide association study), so the Q-Q plot needs to be interpreted with care. The peak association

signals appear on chromosomes 9, 14, 16, 19 and 20 (P-value < e10-^{4}). Figure 3-2 shows peak association signals for deletion CNVs appear on chromosomes 4, 9, and 18. The peak association signal in Figure 3-3 is located in chromosome 16 (P<10e-4).

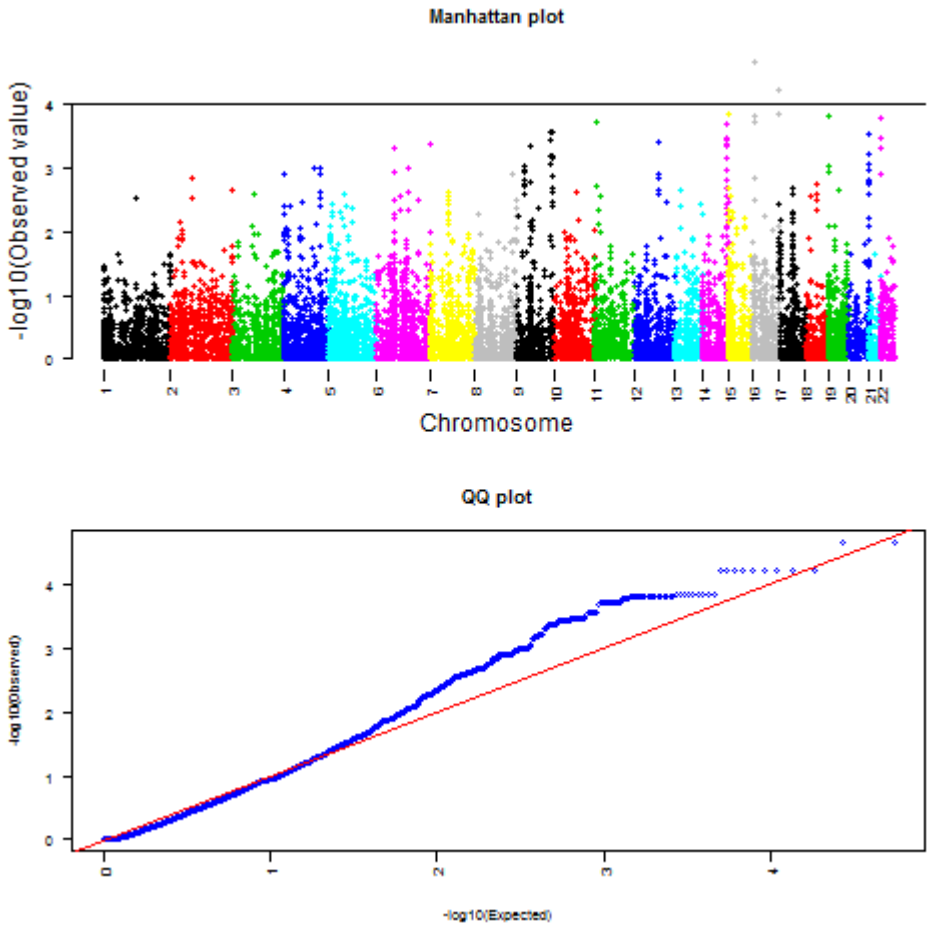


Figure 3-1. Manhattan and QQ plot for amplification CNVs

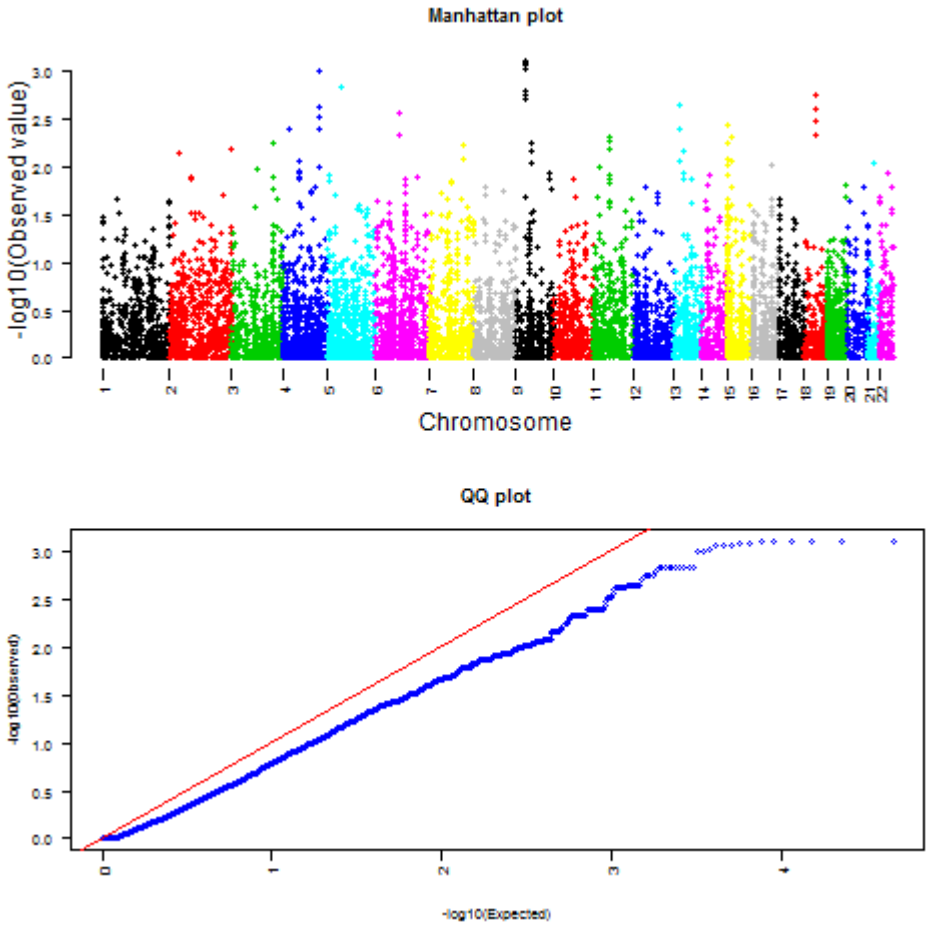


Figure 3-2. Manhattan and QQ plot for deletion CNVs

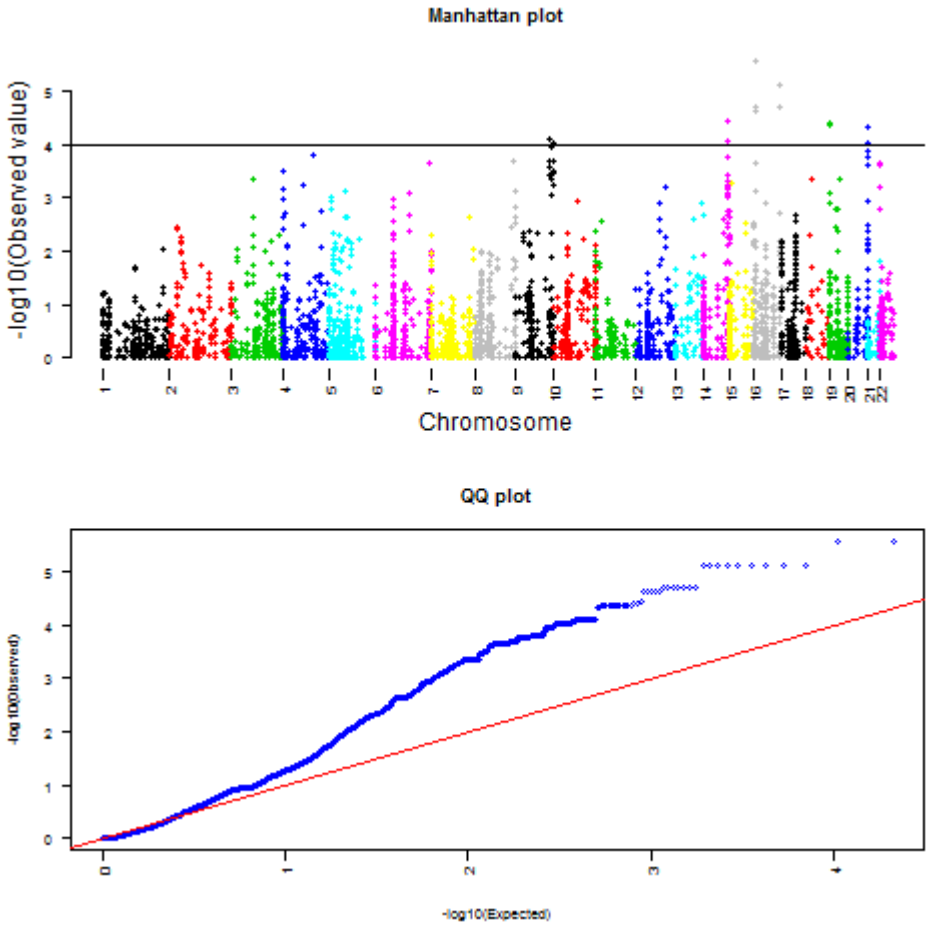


Figure 3-3. Manhattan and Q-Q plot for the whole model

Table 3-3 lists the genes that have CNVs with peak association signals (the highest -$\log_{10}$P values in Manhattan plots) in Figures 3-1, 3-2, and 3-3. These genes may be interesting for AD+P studies. However, the CNVs in all those genes are common CNVs (frequency >5%), which make them less likely to be disease genes with major effects.

Table 3-3. Genes located within association peaks in Manhattan plots (Figures 3-1, 3-2, and 3-3)

Chr	Gene symbols		
	Amplification CNVs	Deletion CNVs	All CNVs
4		—	
9	SET, CAMSAP1	—	
14	KIAA0284, PLD, AHNAK2, CDCA4,GPR132, JAG2, BRF1, PACS2		
16	SOX8, ZFPM1		ZFPM1
18		LOC284260	
19	APC2		
20	TAF4		

3.4.2 Association Analysis in 7 Recurrent CNV Regions across ASD and SCZ

We therefore specifically searched for rare CNVs in the 7 recurrent CNV regions proposed by Moreno-De-Luca et al. (see Table 3-1). All 7 CNVs are large in size and are very rare. We found interesting CNVs in two of the regions: 16p11.2 and 3q29.

3.4.2.1 16p11.2

We found one duplication CNV (copy number = 3) in 16p11.2. It was identified in 2 of 440 AD+P subjects, but not in 136 AD-P, nor in 593 AD with intermediate psychosis, or 855 non-AD controls. This is a very large (> 0.5 Mb) and rare duplication CNV. It is almost completely overlapping with the reported CNV in autism and SCZ, as shown in Figure 3-4. The frequency of this duplication CNV in 16p11.2 in AD+P is similar to that in SCZ; the comparison is listed in Table 3-4.

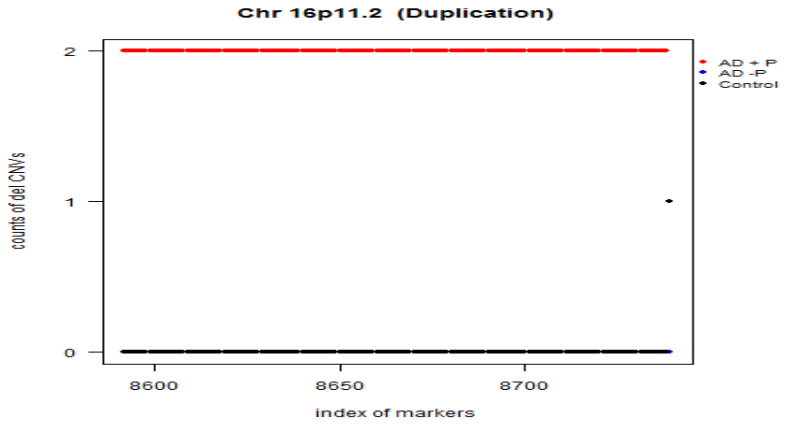


Figure 3-4. The counts of duplication CNVs at 16p11.2 in three sample groups (AD+P, AD-P and non-AD controls).

The X-axis is the index of markers that are located in reported CNV regions in ASD and SCZ; each dot represents one marker. The Y-axis is the count of individuals who carry the duplication CNV at each marker in that region in the three sample groups separately. Black represents non-AD controls; blue represents AD-P; and red represents AD+P.

Table 3-4. Comparison of the duplication CNV in 16p11.2 identified in AD+P and SCZ.

	CNV Position	Number of total subjects	Number of subjects with CNV	Freq of CNV in cases
AD+P	29,554,843-30,105,652	440	2	0.00455
AD (intermediate P)		593	0	
AD-P		136	0	
No-AD controls		855	0	
SCZ	29,557,553-30,107,434	4551	21	0.00461
No-SCZ controls		6391	2	

3.4.2.2 3q29

Figure 3-5 shows the CNV findings in 3q29. In this plot, we found no one carried the large deletion CNV in 3q29 as reported in SCZ. We then specifically looked at CNVs that are very rare in the AD-P and No-AD groups, which (in Figure 3-5) should be regions where the value on the y-axis at each of three or more consecutive markers (points) in AD-P (blue) and No-AD (black) subjects ≤ 1. We did identify a small deletion CNV (~ 28 kb) in that region, as shown by the arrow in Figure 3-5. This CNV is very rare; neither 136 AD-P nor 855 No-AD subjects had it. However 7 of 440 AD+P and 4 of 593 AD with intermediate psychosis subjects carried this CNV. We next conducted trend tests of the three AD groups (AD+P, AD with indeterminate P and "controls"). We chose the "control" group in terms of two different biological models. One model assumes that the genetic variants by themselves do not cause psychosis, but they may increase the risk of psychosis when interacting with AD. Therefore, we may still see the genetic variants in non-AD subjects. An alternative model is that the DNA copy number variant is a direct causal risk factor for AD+P, which is different from risk factors for AD-P; so we would not expect to see those CNVs in non-AD individuals. Under the former model, we only used the AD-P as controls; the one -sided exact p-value from the trend test is 0.0155. Under the latter model, we combined AD-P and No-AD as controls; the p-value is 1.44E-4.

This small deletion CNV, identified in seven AD+P and four AD with intermediate P subjects, is located exactly within gene *PAK2*. The deletions in the different individuals are not identical, but each of them deletes some or all of *PAK2*. They range in size from approximately 10 to approximately 50 kilobases, with those lengths being approximate due to the limited resolution of the SNP array.

The CNVs detected in other two AD with intermediate P subjects are longer. Both of them include *PAK2* and *PIGX*, the gene next to the 5' region of *PAK2*; one of them also included *RNU6-42* and *SENP5*, genes close to the 3' region of *PAK2*. Detailed information about the CNVs in these samples is summarized in Table 3-5. The start and end position of CNVs were estimated by PennCNV, not validated by experiments. So the breakpoints of CNVs may not be accurate.

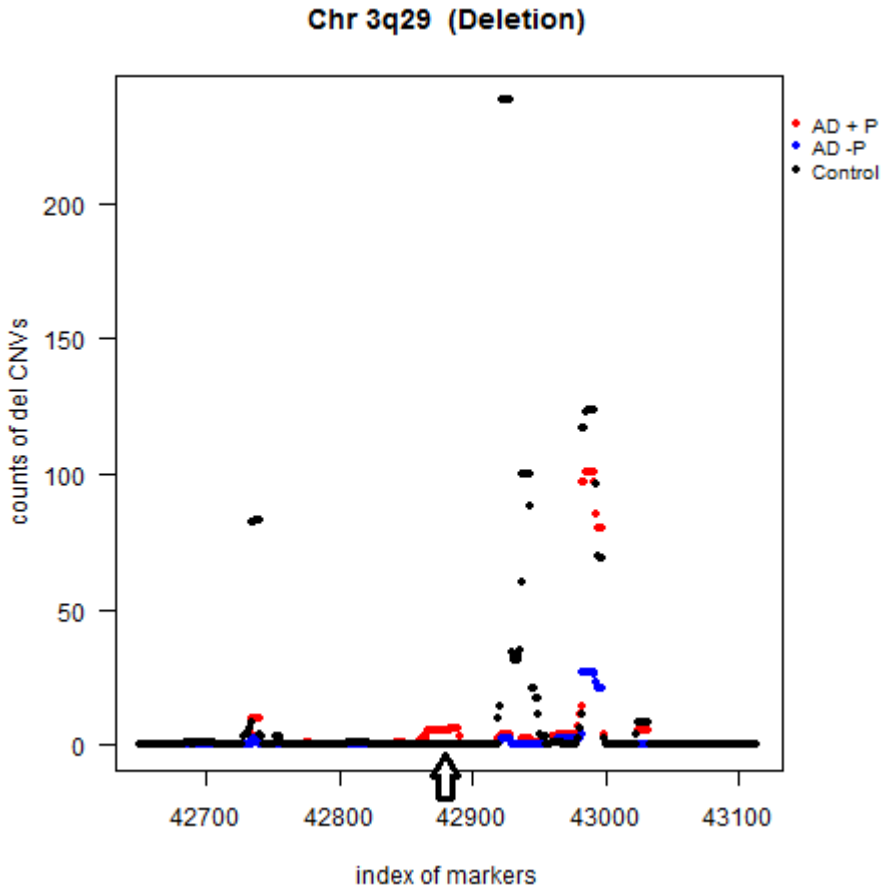


Figure 3-5. The counts of deletion CNVs at 3q29 in three sample groups (AD+P, AD-P and non-AD controls).

The X-axis is the index of markers that are located in reported CNV regions in ASD and SCZ; each dot represents one marker. The Y-axis is the count of individuals who carry the duplication CNV at each marker in that region in the three sample groups separately. Black represents non-AD controls; blue represents AD-P; and red represents AD+P. The region pointed by an arrow is located in *PAK2*.

Table 3-5. Detailed information of the CNV in 3q29 in subjects of AD+P and AD intermediate P

Subjects	CNV start genomic position	CNV end genomic position	CNV start location	CNV end location
C 1	197,946,898	197,959,516	5' region of *PAK2*	Intron 1 of *PAK2*
C 2	198,017,717	198,045,849	Intron 5 of *PAK2*	3' region of *PAK2*
C 3	198,030,771	198,045,849	Intron 12 of *PAK2*	3' region of *PAK2*
C 4	198,038,417	198,045,849	Intron 13 of *PAK2*	3' region of *PAK2*
C 5	198,038,417	198,043,562	Intron 13 of *PAK2*	Intron 15 of *PAK2*
C 6	198,039,349	198,043,562	Intron 14 of *PAK2*	Intron 15 of *PAK2*
C 7	198,041,310	198,043,562	Intron 15 of *PAK2*	Intron 15 of *PAK2*
I 1	197,914,350	197,959,516	5' region of *PIGX*	Intron 1 of *PAK2*
I 2	197,914,350	198,150,251	5' region of *PIGX*	3' of *SENP5*
I 3	197,978,358	198,039,349	Intron 1 of *PAK2*	Intron 14 of *PAK2*
I 4	198,030,771	198,045,849	Intron 12 of *PAK2*	3' region of *PAK2*

3.4.2.3 Other Five CNV Regions

Results in the other five CNV regions are shown in Figure 3-6. We did not find meaningful CNVs in the rest of the five reported CNV regions that are recurrent in SCZ and autism.

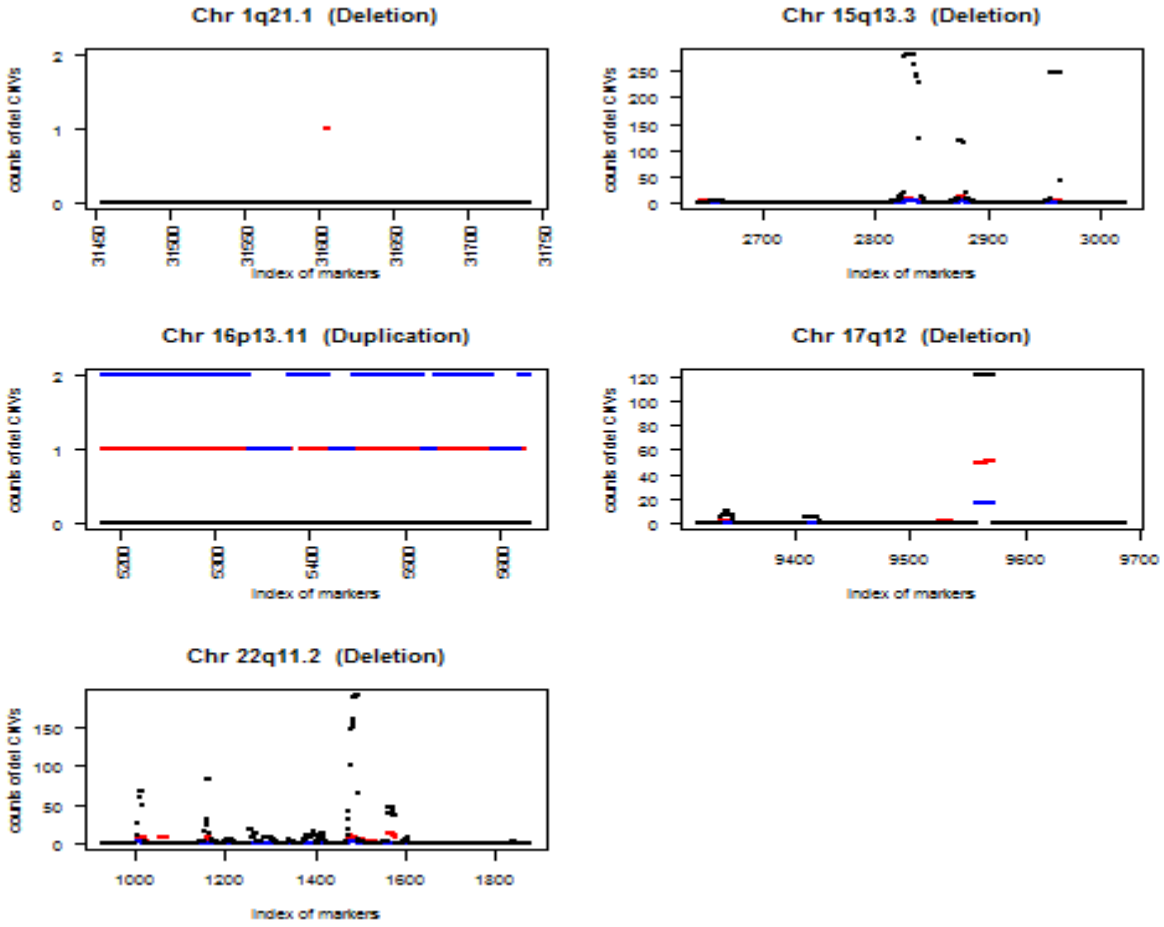


Figure 3-6. The counts of CNVs in other five CNV regions by three sample groups. Red points represent AD+P, blue points represent AD-P and black ones indicate non-AD controls. X-axis is the index of markers that are located in reported CNV regions in ASD and SCZ; each dot represented one marker. Y-axis is the count of individuals who carry duplication CNV at each marker in that region in three sample groups separately. The black one represents non-AD controls; the blue one represents AD-P and the red one represents AD+P.

3.5 DISCUSSION

We are the first to report that AD+P shares a rare risk CNV region on 16p11.2 with SCZ and autism. Its frequency in AD+P is similar to that in SCZ. Several discrete phenotypes, such as SCZ, autism, seizure and mental retardation have been well documented [McCarthy et al, 2009; Weiss et al, 2008; Guilmatre et al, 2009; Bedoyan et al, 2010] to be associated with the microduplication of 16p11.2; therefore they are considered as 16p11.2 duplication syndromes. Our finding actually extends the range of 16p11.2 duplication syndromes to psychosis in Alzheimer disease. It implies that distinct psychiatric disorders may be caused by the same or similar genetic variants, perhaps influenced by different genetic and/or environmental factors. Therefore, it partially supports one of the hypothesized mechanisms proposed by Sweet [2010] that neurodevelopmental disorders and neurodegenerative disorders (eg. AD) may share some common disease modifier genes, which may be involved in the development of psychosis in different disorder processes.

We did not find anyone who had the large deletion CNV (863kb) in 3q29 as reported [Quintero-Rivera et al, 2010; Willatt et al, 2005; Mulle et al, 2010] in SCZ, but we did find a small deletion CNV in *PAK2*, p21 protein (Cdc42/Rac)-activated kinase 2, was considered as one of the most interesting candidate genes in 3q29 for SCZ [Willatt et al, 2005]. Full length *PAK2* stimulates cell survival and cell growth, and may regulate the apoptotic events in the dying cell. *PAK2* has 15 exons; its protein product contains two major functional domains: a regulatory domain and a kinase domain. The regulatory domain located before exon 7. The kinase domain, required for the kinase activity of *PAK2*, starts from the end of exon 7 (Amino Acid 227) and lasts untill the end of whole Amino Acid sequence. We found that the deletion in *PAK2* is a one-

copy deletion, so it won't completely disrupt the function of *PAK2*; instead it may have the potential to decrease the kinase activity of *PAK2* and reduce the cell survival and/or cell growth, which may therefore contribute to the psychosis in degenerative disorder like AD. However, it is unclear how much a one copy deletion can influence a gene's expression or its protein activity. This hypothesis needs to be validated by experiments at the mRNA or protein level.

This deletion CNV in *PAK2* has not been reported in any previous SCZ and autism CNV studies. One of the likely explanations is that the resolutions of the markers in some studies were low; another possible explanation is that most studies only look for large CNVs and filter small CNVs from further analysis, since large CNVs are considered more likely to be real and small CNVs are very difficult to experimentally validate. But in Chapter 2 we showed that small CNVs are not always so unreliable; they should not be ignored in downstream analysis. If this PAK2 deletion can be validated by experiments and other independent studies, we can narrow down the genes responsible for AD+P, and even provide a clue for mechanisms of brain dysfunction. We have initiated such validation studies, but do not have results to report at this time.

In conclusion, although rare, those CNVs we identified in AD+P may have important functions in the development of psychosis. Identification of these CNVs can help with understanding the mechanisms of psychosis disorders. In addition, these CNVs have the potential to be used in clinical practice for screening, diagnosis, disease classification or genetic testing.

3.6 REFERENCES

Bacanu SA, Devlin B, Chowdari KV, DeKosky ST, Nimgaonkar VL, Sweet RA. Heritability of psychosis in Alzheimer disease. Am J Geriatr Psychiatry. 2005; 13: 624–627.

Ballard CG, O'Brien JT, Coope B, Wilcock G. Psychotic symptoms in dementia and the rate of cognitive decline.J Am Geriatr Soc 1997; 45: 1031–1032.

Bedoyan JK, Kumar RA, Sudi J, Silverstein F, Ackley T, Iyer RK, Christian SL, Martin DM. Duplication 16p11.2 in a child with infantile seizure disorder. Am J Med Genet A. 2010; 152A: 1567-74.

Bergen A, Engedel K and Kringlen A. The Role of Heredity in Late Onset Alzheimer's disease and Vascular Dementia Archives of General Psychiatry 1997; 54: 264-270.

Burns A, Jacoby R, Levy R. Psychiatric phenomena in Alzheimer's disease. Br J Psychiatry1990; 157: 72-94.

Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, Thomas-Anterion C, Michon A, Martin C, Charbonnier F, Raux G, Camuzat A, Penet C, Mesnage V, Martinez M, Clerget-Darpoux F, Brice A, Frebourg T. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. Am J Hum Genet. 1999; 65: 664-70.

Carson R, Craig D, Hart D, Todd S, McGuinness B, Johnston JA, O'Neill FA, Ritchie CW, Passmore AP. Genetic variation in the alpha 7 nicotinic acetylcholine receptor is associated with delusional symptoms in Alzheimer's disease. Neuromolecular Med. 2008;10:377–384.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993; 261: 921-3.

DeMichele-Sweet MA, Sweet RA. Genetics of psychosis in Alzheimer's disease: a review. J Alzheimers Dis. 2010; 19: 761-80.

De Strooper B, Saftig P, Craessaerts K, Vanderstichele H, Guhde G, Annaert W, Von Figura K, Van Leuven F. Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. Nature. 1998; 391: 387–390.

Farber NB, Rubin EH, Newhouse PA, Kinscherf DA, Miller JP, Morris JC *et al*. Increased neocortical neurofibrillary tangle density in subjects with Alzheimer's disease. Arch Gen Psychiatry. 2000; 57: 1165–1173.

Forstl H, Burns A, Levy R, Cairns N. Neuropathological correlates of psychotic phenomena in confirmed Alzheimer's disease. Br J Psychiatry. 1994; 165: 53–59.

Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006; 63:168-74.

Ghebranious N, Giampietro PF, Wesbrook FP, Rezkalla SH. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. Am J Med Genet A. 2007; 143: 1462–71.

Go RC, Perry RT, Wiener H, Bassett SS, Blacker D, Devlin B, Sweet RA. Neuregulin-1 polymorphism in late onset Alzheimer's disease families with psychoses. Am J Med Genet B Neuropsychiatr Genet. 2005; 139B: 28-32.

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer’s disease. Nature. 1991; 349: 704–706.

Goedert M, Spillantini MG. A century of Alzheimer's disease. Science. 2006; 314: 777-81.

Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, Drouin-Garraud V, Layet V, Rosier A, Briault S, Bonnet-Brilhault F, Laumonnier F, Odent S, Le Vacon G, Joly-Helas G, David V, Bendavid C, Pinoit JM, Henry C, Impallomeni C, Germano E, Tortorella G, Di Rosa G, Barthelemy C, Andres C, Faivre L, Frébourg T, Saugier Veber P, Campion D. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. Arch Gen Psychiatry. 2009 Sep;66(9):947-56.

Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA. Alzheimer Disease in the US Population. Prevalence Estimates Using the 2000 Census. Arch Neurol. 2003; 60: 1119-1122.

Heinzen EL, Need AC, Hayden KM, Chiba-Falek O, Roses AD, Strittmatter WJ, Burke JR, Hulette CM, Welsh-Bohmer KA, Goldstein DB. Genome-wide scan of copy number variation in late-onset Alzheimer's disease. J Alzheimers Dis. 2010; 19: 69-77.

Horesh Y, Katsel P, Haroutunian V, Domany E. Gene expression signature is shared by patients with Alzheimer's disease and schizophrenia at the superior temporal gyrus. Eur J Neurol. 2011; 18: 410-24.

Jeste DV, Wragg RE, Salmon DP, Harris MJ, Thal LJ. Cognitive deficits of patients with Alzheimer's disease with and without delusions. Am J Psychiatry 1992; 149: 184–189.

Kauwe J and Goate A. Molecular Genetics in Alzheimer's Disease. Neurobiology of Alzheimer's disease Eds. Dawbern, D. and Allen, S. Oxford University Press 2007; 59-80.

Khachaturian AS, Corcoran CD, Mayer LS, Zandi PP, Breitner JC. Apolipoprotein E epsilon4 count affects age at onset of Alzheimer disease, but not lifetime susceptibility: The Cache County Study. Arch Gen Psychiatry. 2004; 61: 518–24.

Lopez OL, Kamboh MI, Becker JT, Kaufer DI, DeKosky ST. The apolipoprotein E e4 allele is not associated with psychiatric symptoms or extrapyramidal signs in probable Alzheimer's disease. Neurology 1997; 49: 794–797.

Lyketsos CG, Steinberg M, Tschanz JT, Norton MC, Steffens DC, Breitner JCS. Mental and behavioral disturbances in dementia: findings from the Cache County study on memory in aging. Am J Psychiatry 2000; 157: 708–714.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, Grozeva D, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Roccanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimäki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nöthen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR; Wellcome Trust Case Control Consortium, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009; 41: 1223-7.

Moreno-De-Luca D; SGENE Consortium, Mulle JG; Simons Simplex Collection Genetics Consortium, Kaminsky EB, Sanders SJ; GeneSTAR, Myers SM, Adam MP, Pakula AT, Eisenhauer NJ, Uhas K, Weik L, Guy L, Care ME, Morel CF, Boni C, Salbert BA, Chandrareddy A, Demmer LA, Chow EW, Surti U, Aradhya S, Pickering DL, Golden DM, Sanger WG, Aston E, Brothman AR, Gliem TJ, Thorland EC, Ackley T, Iyer R, Huang S, Barber JC, Crolla JA, Warren ST, Martin CL, Ledbetter DH. Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. Am J Hum Genet. 2010; 87: 618-30.

Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC, Sobreira NL, Valle D, Rudd MK, Satten G, Cutler DJ, Pulver AE, Warren ST. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet. 2010 Aug 13; 87: 229-36.

Paulsen JS, Salmon DP, Thal L, Romero R, Weisstein-Jenkins C, Galasko D *et al*. Incidence of and risk factors for hallucinations and delusions in patients with probable Alzheimer's disease. Neurology 2000; 54: 1965–1971.

Prestia A. Alzheimer's Disease and Schizophrenia: Evidence of a Specific, Shared Molecular Background. Future Neurology. 2011;6:17-21.

Quintero-Rivera F, Sharifi-Hannauer P, Martinez-Agosto JA. Autistic and psychiatric findings associated with the 3q29 microdeletion syndrome: case report and review. Am J Med Genet A. 2010; 152A: 2459-67.

Rockwell E, Jackson E, Vilke G, Jeste DV. A study of delusions in a large cohort of Alzheimer's disease patients. Am J Geriatr Psychiatry 1994; 2: 157–164.

Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang Y, Chi H, Lin C, Holman K, Tsuda T, et al. Familial Alzheimer’s disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer’s disease type 3 gene. Nature. 1995; 376: 775–778.

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ, Hulette C, Crain B, Goldgaber D, Roses AD. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993; 43: 1467-72.

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer’s disease. Nature. 1995; 375: 754–760.

Sweet RA, Hamilton RL, Lopez OL, Klunk WE, Wisniewski SR, Kaufer DI *et al*. Psychotic symptoms in Alzheimer's disease are not associated with more severe neuropathologic features. Int Psychogeriatr 2000; 12: 547–558.

Sweet RA, Nimgaonkar VL, Devlin B, Lopez OL, DeKosky ST. Increased familial risk of the psychotic phenotype of Alzheimer disease. Neurology. 2002; 58: 907–911.

Sweet RA, Nimgaonkar VL, Devlin B, Jeste DV. Psychotic symptoms in Alzheimer disease: evidence for a distinct phenotype. Mol Psychiatry. 2003; 8: 383-92.

Tunstall N, Owen MJ, Williams J, Rice F, Carty S, Lillystone S, Fraser L, Kehoe P, Neill D, Rudrasingham V, Sham P, Lovestone S. Familial influence on variation in age of onset and behavioural phenotype in Alzheimer’s disease. Br J Psychiatry. 2000; 176: 156–159.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ; Autism Consortium. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med. 2008; 358: 667–75.

Willatt L, Cox J, Barber J, Cabanas ED, Collins A, Donnai D, FitzPatrick DR, Maher E, Martin H, Parnau J, Pindar L, Ramsay J, Shaw-Smith C, Sistermans EA, Tettenborn M, Trump D, de Vries BB, Walker K, Raymond FL. 3q29 microdeletion syndrome: clinical and molecular characterization of a new syndrome. Am J Hum Genet. 2005; 77:154-60.

Zubenko GS. Do susceptibility loci contribute to the expression of more than one mental disorder? A view from the genetics of Alzheimer's disease. Mol Psychiatry. 2000; 5:131-6.

4.0 CNVS, BIRTH OUTCOMES AND MATERNAL SMOKING IN A PRETERM BIRTH CASE-CONTROL STUDY

4.1 ABSTRACT

Objectives: Adverse birth outcomes such as preterm delivery increase infant mortality and can have lifelong health consequences. They are likely the results of interactions between genetic factors and maternal or fetal environmental factors. However, the genetic determinants are unclear. The purpose of this study is to search for genetic copy number variations (CNVs) that are associated with preterm delivery and low birth weight, and to investigate whether those CNVs have direct or indirect effects on smoking-induced adverse birth outcomes.

Methods: A large preterm birth case-control dataset including 1937 mothers was used as the primary dataset. We first examined the association of smoking with birth outcomes; next, we examined the association of CNVs with birth outcomes in candidate genes related to *GSTT1*, which harbors a known CNV that has been shown to be associated with birth outcomes in smokers; next, we tested the association of CNVs in *GSTT1/GSTT2* with smoking; finally we analyzed the rest of the genome to nominate candidate CNVs that are associated with birth outcomes in smokers and non-smokers separately. We also analyzed genome-wide to identify CNVs associated with smoking.

Results: We confirmed the association of smoking with low birth weight and preterm delivery (PTD) in the preterm birth dataset. We were not able to confirm the association of the known CNV in *GSTT1* with birth outcomes, because our arrays do not contain markers in that CNV region. However, we found four other CNVs in *GSTT1/GSTT2* that are associated with birth outcomes in smokers and/or non smokers. We also found two CNVs that are associated with smoking in the preterm birth dataset and in a replication dataset. We nominated several candidate genes for smoking and birth outcomes by genome-wide scan.

4.2 INTRODUCTION

Low birth weight (LBW) refers to birth weight less than 2500g [Kramer MS, 1987], which accounts for about 16% of all live-borns in the world [de Onis et al., 1998]. Birth weight is regulated by two major processes: duration of gestation and intrauterine growth rate. Preterm delivery (PTD), the birth of a baby with less than 37 weeks gestational age, is responsible for one-third to two-thirds of infants with LBW [Arifeen et al., 2000; Martin et al., 2007]. PTD and LBW each can increase the risk of fetal mortality and infant mortality. However, the causes of PTD and LBW are not clear. Multiple factors may contribute to the development of LBW and/or PTD, including genetics, environmental and other factors (e.g. demographic, obstetric, nutritional factors and maternal morbidity during pregnancy) [Kramer MS, 1987].

One of the important environmental factors in birth outcomes is tobacco smoking. Maternal tobacco smoking is the single largest prenatal risk factor for a number of different problems. It may reduce the mean birth weight and increase the risk of PTD and intrauterine

growth restriction [Asmussen et al., 1975; Asmussen et al., 1977; Goldenberg et al., 2007; Kjell et al., 2007; Ronco et al., 2005; Nilsen et al., 1984]. One likely pathogenic pathway of smoking-induced adverse birth outcomes may be associated with the metabolism of the tobacco compound PAH (polycyclic aromatic hydrocarbons) [Perera et al., 2005; Tsui et al., 2008, Wu et al., 2010]. PAH is one of the most important carcinogenic compounds, and is detoxified in a two-stage process. PAHs are converted into procarcinogen in the first stage, which is then conjugated into excretal metabolites in the second stage. The conjugation is catalyzed by the gene *GSTT1* (Glutathione S-transferase theta 1), which belongs to the theta class of GSTs. The class members, *GSTT1* and *GSTT2*, are located in human chromosome 22. They are about 50kb away from each other, with a GSTT pseudogene *GSTTP1* located between them. *GSTT1* and *GSTT2* share 55% amino acid sequence identity and both are considered to have a detoxification role [Coggan et al., 1998]. It has been very well established that a common deletion in *GSTT1* is associated with modifying the effect of maternal smoking on birth outcomes. Smokers have lower mean birth weight infants compared to nonsmokers; however the reduction varies according to the *GSTT1* polymorphism. The mean birth weight decreased dramatically in smokers with the *GSTT1* null genotype compared to smokers with the *GSTT1* wild-type genotype [Aagaard-Tillery et al., 2010; Grazuleviciene et al., 2009; Wang et al., 2002; Wu et al., 2007].

Linkage and SNP association studies have identified some genes that are associated with adverse birth outcomes. However, copy number variation (CNV) studies have rarely been conducted in low birth weight and preterm delivery. In this study we set out to do that, taking the following steps. All analyses of birth weight were performed in the controls (term births) only.

1) Confirm the association between smoking and birth outcomes (PTD and LBW) in our dataset.

2) Confirm the association of *GSTT1*/*GSTT2* CNVs with birth outcomes in smokers, and test whether it is also seen in non-smokers.

3) Test for any direct association between *GSTT1*/*GSTT2* CNVs and smoking in both the preterm birth dataset and a replication dataset (dental caries dataset).

4) Genome scans to identify other CNVs that are associated with smoking in both datasets.

5) Analyze the rest of the genome to nominate candidate CNVs that are associated with birth outcomes in smokers and non-smokers separately.

4.3 MATERIALS AND METHODS

4.3.1 Study Populations

Both the preterm birth and dental caries datasets are part of the GENEVA (Gene Environment Association studies) consortium. Detailed information on both studies is available from study documents in dbGaP (http://www.ncbi.nlm.nih.gov/gap) [Mailman et al., 2007]. The preterm birth study is a case-control study of approximately 1000 mother-child case pairs (cases were defined as infants<37 weeks of gestation), and 1000 mother-child controls pairs (controls were defined as infants=40 weeks of gestation) from the Danish National Birth Cohort study [Olsen et al., 2001]. From the GENEVA preterm birth study

(http://www.ncbi.nlm.nih.gov/gap?term=geneva), we used almost all maternal samples -1937 mothers genotyped on Illumina Human660W-Quad chip. Of these, 893 were cases of PTD, 978 were controls, and 66 were neither cases nor controls (37 ≤ infants < 40 weeks of gestation). The smoking phenotype we used was "any smoking" or "non-smoker" during pregnancy in the preterm birth dataset. The replication dataset for smoking is the GENEVA dental caries study (http://www.ncbi.nlm.nih.gov/gap?term=geneva), which is a large community-based study of oral health genotyped on the Illumina HumanHap610 chip. The full dental caries study included four different community-based samples from Western Pennsylvania, West Virginia, and Iowa. Individuals were selected without regard to phenotype, and then were extensively phenotyped for oral health and related traits. A subset of the full study, 1313 adults with complete smoking status, was used in our study. The smoking phenotype was defined as "any smoking" or "non-smoker" in lifetime. The characteristics of these two datasets are summarized in Table 4-1.

Table 4-1. Characteristics of two datasets

	Preterm Birth	**Dental Caries**
Recruitment	Case-control study within a cohort	Community-based families
Sample size	1937 genotyped mothers	1313 adults with complete smoking status
Illumina Chips	Human660W-Quad	Humanhap 610
Number of markers	660K	610K
Source of DNA	Blood with different origins	Blood, saliva and mouthwash
Smoking	"Any smoke"or "Non-smoke" during pregnancy	"Any smoke"or "Non-smoke" in life time
Birth outcomes	Yes	NA

4.3.2 Genotyping and Quality Control

Complete genotyping and data cleaning reports for both studies are available in dbGAP (http://www.ncbi.nlm.nih.gov/gap).The level of genotyping quality was extremely high.

4.3.3 CNV Calls by PennCNV

We generated CNV calls using the PennCNV software (2009Aug27 version) [Wang et al., 2007]. We used the GC model wave adjustment procedure in PennCNV. Samples were filtered after GC model adjustment using the criterion lrrsd > 0.3; 1617 out of 1937 preterm birth samples and 1261 out of 1313 dental caries samples remained. All analysis was restricted to autosomes. All procedures followed the user guidelines of PennCNV and those developed in Chapter 2. Human genome build 36 was used for this study. CNVs with copy number >2 were defined as amplifications, while those with copy number <2 were considered as deletions.

4.3.4 Statistical Analysis

We first examined the association of smoking with PTD and birth weight by logistic regression and linear regression respectively. We next examined the association of CNVs in *GSTT1*, *GSTTP1* and *GSTT2* with the continuous outcome variable birth weight using linear regression, and with binary outcomes smoking and PTD using logistic regression. We finally screened the CNVs genome-wide for association with smoking using a chi-square-test and for birth weight using a t test. If the p-values were smaller than a pre-set threshold at three or more consecutive

markers, the region was considered as a significant CNV, and the smallest p-value was reported as the p-value for that CNV. The boundaries of the CNV were defined as the first and last markers in that region with p-value < 0.05. All calculations were completed in R (version 2.10.1) [R Development Core Team. 2009].

4.4 RESULTS

4.4.1 Association between Smoking and Birth Outcomes

The goal of our first analysis was to confirm the well-known association of smoking with birth weight and PTD. We examined this using linear regression for birth weight in controls (term births) and logistic regression for PTD. Controls are all exactly 40 weeks gestation. The results are summarized in Table 4-2. Smoking is a significant predictor for birth weight in term births (p=2.01E-10); the mean birth weight in the smoking group is 277.87g lower than in the non-smoking group. In addition, smoking is a significant predictor for PTD (odds ratio = 1.27, p=0.028).

Table 4-2. Relationships between smoking and birth outcomes in preterm birth dataset

Outcomes	Predictors	Coefficients (Odds Ratio)	P-value
Birth weight in term births	Smoking	-277.87	2.01E-10
PTD	Smoking	0.24 (1.27)	0.028

4.4.2 Association between *GSTT1* / *GSTT2* and Birth Weight, Stratified by Smoking

We next took a closer look at the CNVs in the regions from *GSTT2* to *GSTT1*. We checked the association of those CNVs with birth weight in controls. We then stratified the controls by smoking status, and analyzed the association in smoking and non-smoking controls separately. Table 4-3 summarizes the association of those CNVs with birth weight in different control groups.

Table 4-3. Association of CNVs in the regions from *GSTT2* to *GSTT1* with birth weight in different control groups in preterm birth data

					Non-PTD mothers (Controls)		Non-PTD smoking mothers (Smoking controls)		Non-PTD Non-smoking mothers (Non-smoking controls)	
Gene Symbol	CNV	Start Position	End Position	Num of SNP	Change of mean BW	Pvalue	Change of mean BW	Pvalue	Change of mean BW	Pvalue
GSTT2	Del	22618460	22624684	10	-82.6	0.0239	-122.8	0.0384	-41.4	0.3521
GSTT2	Amp	22653131	22666327	20	115.9	0.0162	324.1	0.0078	74.1	0.2066
GSTTP1	Del	22672338	22687707	18	167.2	0.0048	322.7	0.0046	74.2	0.2246
GSTT1	Amp	22695041	22697104	16	-115.7	0.0210	20.0	0.8402	-144.9	0.0080

We could not replicate the known association of a deletion CNV in *GSTT1* with low birth weight in our dataset, because no markers in that CNV region were included in our arrays. However, we found a duplication CNV in *GSTT1* that decreases birthweight, but only in non-smokers. We also found three CNVs in genes other than *GSTT1* which are associated with birth weight. Two CNVS are located in *GSTT2*; one is a deletion and the other one is a duplication. Term birth smoking mothers who carry the deletion CNV in *GSTT2* have significantly lower mean infant birth weight than smokers with normal copy number. On the other hand, smokers who carry the amplification CNV show relatively higher mean infant birth weight than those with normal copy number; the mean infant birth weight is 324.1g higher in smokers with this duplication CNV

than in smokers without this duplication CNV. We also found one deletion CNV in *GSTTP1* which has very similar effect as the duplication CNV in *GSTT2*.

4.4.3 Association between *GSTT1* / *GSTT2* and PTD, Stratified by Smoking

We then examined the association of the CNVs in the regions from *GSTT2* to *GSTT1* with PTD. We also stratified the samples by smoking status, and analyzed the association in smoking and non-smoking PTD mothers separately. Other than those 4 CNVs (in Table 4-3) associated with birth weight, we did not find any new CNV in the region from *GSTT2* to *GSTT1* that is significantly associated with PTD. We therefore summarize the association of those 4 CNVs with PTD in different sample groups in Table 4-4.

Table 4-4. Association of CNVs in the regions from *GSTT2* to *GSTT1* with PTD in mothers stratified by smoking state in preterm birth data

				All mothers		Smoking mothers		Non-smoking mothers	
Gene Symbol	CNV	Start Position	End Position	Odds Ratio	P-value	Odds Ratio	P-value	Odds Ratio	P-value
GSTT2	Del	22618460	22624684	0.78	0.0794	0.48	0.0032	0.94	0.7484
GSTT2	Amp	22653131	22666327	1.06	0.7423	1.72	0.0885	0.79	0.2331
GSTTP1	Del	22672338	22687707	1.16	0.4522	1.92	0.1077	1.11	0.5651
GSTT1	Amp	22695041	22697104	1.07	0.6605	0.60	0.0861	1.31	0.1774

The duplication CNV in *GSTT1* is not significantly associated with PTD. Smoking does not statistically significantly modify the relationship, but this bears further study since the p-value for smokers is 0.0885. Smokers with the deletion CNV in *GSTT2* have significantly reduced risk (odds ratio 0.48) of PTD compared to smokers without this CNV. Smokers with the duplication CNV in *GSTT2* have marginally increased risk (odds ratio 1.72) of PTD compared to smokers with normal copy number. The deletion CNV in *GSTTP1* is not significantly associated

with PTD. There is no statistically significant evidence that smoking modifies the effect, but the p-value for smokers is 0.1077.

4.4.4 Association between *GSTT1* / *GSTT2* and Smoking in Two Datasets

In order to further examine the relationship between smoking, GSTT1, and birth outcomes, we used logistic regression to test the relationship between CNVs in the region from *GSTT2* to *GSTT1* and smoking in two independent datasets: the preterm birth dataset and the dental caries dataset. By prior hypothesis, variants in GSTT1 and/or GSTT2 should modify the effect of smoking on birth outcomes, but should not be associated with smoking itself. However, we detected strong associations between CNVs in this region and smoking. The results are summarized in Table 4-5. Note that the dental caries dataset does not include markers covering out GSTT2 CNVs, so those could not be tested in that dataset.

Table 4-5. Association of CNVs in the region from *GSTT2* to *GSTT1* with smoking in two datasets

		Preterm birth dataset						Dental caries dataset			
				All mothers		Non-PTD mothers		All subjects			
Gene Symbol	CNV	Start Position	End Position	Odds Ratio	P-value	Odds Ratio	P-value	Start Position	End Position	Odds Ratio	P-value
GSTT2	Del	22618460	22624684	1.55	0.0005	8.54	0.0001				
GSTT2	Amp	22653131	22666327	1.26	0.2222	0.57	0.0446				
GSTTP1	Del	22672338	22687707	0.70	0.0520	0.52	0.0409	22664948	22668071	0.13	0.0218
GSTT1	Amp	22695041	22697104	1.28	0.1379	1.76	0.0134	22688572	22717669	0.44	0.0198

In the preterm birth dataset, we found that the deletion CNV in *GSTT2* is associated with significantly increased risk of smoking (odds ratio=8.54) in non-PTD mothers (p=0.0001). We have already known non-PTD smokers with this CNV have significant (p=0.024) lower mean

birth weight compared with those without this deletion. This deletion is also significantly (p=0.0032) associated with lower risk of PTD (odds ratio=0.48) in smokers.

Another CNV in *GSTT2* is a duplication. We found that it is significantly (p=0.0446) associated with decreased risk of smoking (odds ratio=0.57) in non-PTD mothers. We have already found non-PTD smokers who carry this CNV have significantly higher mean birth weight (p=0.0078) compared to those with wild genotype; it did not significantly influence the risk of PTD.

The third CNV is a deletion, located in gene *GSTTP1*. It had almost the same relationship with birth outcomes and smoking as the above amplification CNV in *GSTT2*, and the decreased risk of smoking for individuals with this deletion CNV in GSTTP1 is replicated by caries dataset (odds ratio=0.13, p=0.0218). The fourth CNV is a duplication CNV in *GSTT1*. It is significantly associated with smoking in both datasets but with contradictory odds ratio in two datasets. It increased the risk of smoking in non-PTD mothers in preterm birth dataset, but decreased the risk of smoking in dental caries dataset. For this CNV, we have already known that it is associated with decreased birth weight in non-PTD non-smokers, and it was not significantly associated with PTD.

4.4.5 Genome-Wide Scan to Identify CNVs for Smoking in Two Datasets

Next, we screened genome-wide for the candidate CNVs for smoking in those two datasets. Logistic regression was used for the binary outcome of smoking. Amplification and deletion CNVs were coded as dummy variables, with normal copy number serving as the reference. Table 4-6 lists genes in 21 CNVs which were significantly (p<0.002) associated with smoking in the preterm delivery dataset; Table 4-7 lists genes in nine CNVs

which were significantly ($p<0.004$) associated with smoking in the dental caries dataset. Genes *PDZD2*, *GOLPH3*, and *HLA-B* were significantly associated with smoking in both datasets.

Table 4-6. CNVs significantly ($p<0.002$) associated with maternal smoking in preterm birth data

Chr	CNV		Gene Symbol	P-value*
1	Amplification		*NME7*	0.0011
			CFHR4	0.0007
2	Amplification		*ZC3H6*	0.0012
3	Amplification		*CDV3*	0.0004
		Deletion	*ROBO2*	0.0014
		Deletion	*UROC1*	0.0015
5	Amplification		*PDZD2,GOLPH3*	2.93E-05
6		Deletion	*HLA-B*	0.0016
9	Amplification		*DOCK8*	0.0004
10	Amplification		*CAMK2G*	0.0009
14	Amplification		*RNASE2(3)*	0.0002
	Amplification		*FAM30A*	0.0020
16	Amplification		*LUC7L*	0.0007
	Amplification		*ATXN2L*	0.0014
17	Amplification		*KIAA0753*	0.0019
	Amplification		*FASN,CCDC57,FLJ23754*	0.0014
		Deletion	*cr597597*	0.0005
19	Amplification		*ZBTB7A, MAP2K2*	0.0009
		Deletion	*GNG7*	0.0018
22	Amplification		*MTMR3*	0.0006
		Deletion	*GSTT2*	0.0015

***** Smallest p-value in a reported CNV region.

Table 4-7. CNVs significantly (p<0.004) associated with maternal smoking in dental caries data

Chr	CNV	Gene Symbol	P-value*
2	Deletion	*c2orf27,MGC50273*	0.0028
5	Amplification	*PDZD2,GOLPH3*	0.0036
6	Deletion	*HLA-B*	4.55E-6
7	Deletion	*OR2A1,FKSG35,FLJ43692*	0.0002
8	Deletion	*ADAM5P,tMDC*	0.0023
10	Deletion	*CTNNA3*	0.0008
13	Deletion	*DQ586768*	0.0009
15	Amplification	*SGK269*	0.0028

***** Smallest p-value in a reported CNV region.

4.4.6 Genome-Wide Scan to Identify CNVs for Birth Weight (Term Births), Stratified by Smoking

We next used t-test to search the rest of the genome to nominate candidate CNVs that are associated with birth weight in controls (term births). We stratified the controls by smoking status, and compared the genes in those CNVs that are significantly associated with birth weight in smokers and non-smokers. Table 4-8 lists the genes that are associated with birth weight only in smokers, not in non-smokers.

Table 4-8. Genes in CNVs which are significantly (P <0.003) associated with birth weight (term births) only in smokers in preterm birth dataset

Chr	CNV	Gene Symbol	P-value
4	Amp	*BC039519*	0.0008
5	Amp	*TPPP*	0.0011
5	Amp	*CWF19L2*	0.0012
13	Del	*STARD13*	6.08E-05
13	Del	*CORL2*	0.0006
20	Amp	*MYO5B*	0.0025

4.4.7 Genome-Wide Scan to Identify CNVs for PTD, Stratified by Smoking

We next searched for CNVs for PTD genome-wide in the preterm birth dataset using the chi-square test. We conducted the analysis in smokers and non-smokers separately, and compared the results. Table 4-9 summarizes the CNVs that are significantly (p<0.003) associated with PTD only in smokers, but not in non-smokers.

Table 4-9. Genes in CNVs significantly (P<0.003) associated with PTD in smokers only in preterm birth dataset

Chr	CNV	Gene Symbol	P-value
1	Del	*KCNAB2(CHD5,RPL22,ICMT,PCCMT,ACOT7,BACH,HES3)*	0.000188
2	Del	*SNED1 (AK05589, BC040629)*	0.00017
9	Del	*NOTCH1*	0.001097
10	Del	*MMRN2,SNCG*	0.002711
11	Amp	*SLC35F2*	0.001517
	Del	*TRPM5(CD81,TSSC4)*	0.000339
14	Del	*IGHE*	0.001097
16	Amp	*RAB11FIP3*	0.001628
	Amp	*WWOX*	0.001978
	Del	*SOX8,SSTR5,C1QTNF8*	0.001978
	Del	*KCTD5*	0.001978
19	Del	*FAM148C,SHC2,ODF3L2*	0.001978
20	Del	*NTSR1,OGFR,COL9A3,TCFL5,DIDO1,SLC17A9,BHLHE23*	0.001978
22	Del	*GSTT2*	0.00247

4.5 DISCUSSION

We first checked the relationship between smoking and birth outcomes in our preterm birth dataset. The findings that smoking is significantly associated both PTD and LBW were consistent with other reports [Chan et al., 2001; Horta et al., 1997.].

To confirm the association of *GSTT1/GSTT2* with birth outcomes, we next conducted CNV association analysis in regions from *GSTT2* to *GSTT1*, stratifying on smoking status. We also tested the association of CNVs in this region with smoking in two datasets. All the tests for association with birth weight were completed in controls, who have exactly 40 weeks gestation. We identified 4 CNVs in the region: two CNVs (one deletion and one duplication) in *GSTT2*, one deletion in *GSTTP1* and one duplication in *GSTT1*. The deletion CNV in *GSTT2* (size 8kb) significantly decreased birth weight in smokers; but not in non-smokers; it also lowered the risk of PTD in smokers and increased the risk of smoking in term birth mothers. One of the most likely explanations is that the GST gene family has a role in detoxification of tobacco. A deletion CNV in *GSTT2* may delay the metabolism of nicotine, elongate the addiction effect of nicotine, and increases the risk of smoking. The toxicity in smokers with deletion CNV in *GSTT2* may be not severe enough to cause PTD but may induce slow fetal growth and lead to low birth weight. This may also explain the finding of a duplication CNV in *GSTT2*. Smokers with this duplication CNV (size 13kb) in *GSTT2* had significant higher birth weight than smokers with normal copy number. This duplication CNV is associated with decreased risk of smoking in term birth mothers, and is not significantly related with PTD. The consistent findings of the two CNVs in *GSTT2* suggest that *GSTT2* may have an effect on smoking related low birth weight. Interestingly, we also found a deletion CNV (size 15kb) in *GSTTP1*, which has almost the same

effects, but smaller magnitude, on birth outcomes and smoking as the amplification CNV in *GSTT2*. It increases the birth weight in smokers with this CNV as compared to smokers with normal copy number; it also reduces the risk of smoking. Moreover, its effect on decreasing the risk of smoking was replicated in the dental caries dataset. We also identified one duplication CNV in *GSTT1*, which was significantly associated with smoking in both datasets. However the odds ratios were reversed in the two datasets. It increased the risk of smoking in non-PTD mothers in the preterm birth dataset, and decreased the risk of smoking in adults in the dental caries dataset. This is likely due to selection bias. *GSTT1* is a detoxification gene; smoking mothers with a duplication CNV in *GSTT2* may suffer less toxicity; while smoking mothers with normal copy number may suffer more toxicity. Severe toxicity may lead to abortion or quit smoking. Therefore some smoking mothers with normal copy number may not be selected in the preterm delivery study. While in dental caries dataset, the duplication CNV in *GSTT2* may accelerate the metabolism of nicotine, reduce its addiction, and reduces the risk of initial smoking. We also tried to compare our CNV results in *GSTT1* and *GSTT2* with findings of *GSTT1* and *GSTT2* from other studies. A deletion in *GSTT1* [Aagaard-Tillery et al., 2010; Grazuleviciene et al., 2009; Wang et al., 2002; Wu et al., 2007] has already been reported to modify the effect of smoking on birth weight; smokers with this null genotype of *GSTT1* had lower mean birth weight than those with control genotype of *GSTT1*. This deletion polymorphism in *GSTT1* is located between physical genomic position 22700kb to 22710kb on chromosome 22, which unfortunately was not covered by HumanHap660 and HumanHap610 chips used in our study. Wang et al [Wang et al., 2008] found a SNP rs1622002 (genomic position 22630580) in *GSTT2* which is associated with metabolism of major tobacco carcinogen PAH, however this SNP was not included in our chips. We compared the association results

from CNVs to that from SNPs, to investigate whether the CNVs were in linkage disequilibrium with any SNP that is associated with smoking and birth outcomes. Interestingly, all SNPs in CNV regions of *GSTT1* and *GSTT2* identified in our study were excluded from SNP association analysis due to out of Hardy-Weinberg equilibrium or missing call. This suggests that the CNVs we identified in *GSTT1/GSTT2* are very likely to be real.

Next, we searched for other CNVs that are associated with smoking by genome-wide screening in two datasets. Genes *PDZD2, GOLPH3, and HLA-B* are significantly associated with smoking in both datasets. Consistent results suggested that we should further explore the role of those genes in smoking. However, smoking is a complex process, which may include initiation, persistence and cessation of tobacco use. We are not sure in what stage of smoking those genes are involved.

Next, we conducted a genome-wide association study to identify the CNVs for birth outcomes (birth weight and PTD) in smokers and non-smokers respectively. All association tests for birth weight were conducted in controls (term births). We were especially interested in genes that are associated with birth outcomes only in smokers, not in non-smokers. Those genes may interact with smoking to influence the birth outcomes. It is noticeable that none of the genes associated with birth weight (term births) in smokers only overlapped with the genes that are associated with PTD in smokers only. Two likely hypotheses for mechanisms of slow intrauterine growth and PTD are: they share common genetic factors but may have different phenotypes due to modification of environmental factors like smoking; or they have different genetic backgrounds. Our results support the latter hypothesis.

There are several limitations in our study. First, the missingness of smoking state data in PTD cases and controls is not completely random. Seven out of 985 PTD controls and 30 out of

924 PTD cases have missing smoking states; however the number of individuals with missing smoke state is very small, which should not significantly influence the results of our study. Second, we used any smoke in pregnancy in preterm birth data and any smoke in life time in dental caries dataset as a smoking variable, which did not consider the quantitative trait of smoking amount and influence of smoking in different trimesters, and therefore may be less informative. Third, we did not adjust some confounding factors for birth weight, such as demographic, psychosocial and obstetric factors. Fourth, we focused on the effects of maternal genetics on birth outcomes; we ignored the impact of fetal genetics on the birth outcomes.

Despite these limitations, we believe our study sheds light on the CNV studies in adverse birth outcomes. We conducted candidate gene analysis in *GSTT1/GSTT2*, and found four new CNVs that are associated with birth weight and PTD, through or not through the interaction with smoking. We thoroughly screened the CNVs for smoking and birth outcomes genome-wide, and identified several strong candidate genes. Also, the consistent findings in two large-scale GENEVA datasets make the candidate genes for smoking very interesting. Further studies are warranted to further investigate their roles in smoking and birth outcomes, especially the linkage disequilibrium among the CNVs in *GSTT1*/*GSTT2*.

4.6 REFERENCES

Aagaard-Tillery K, Spong CY, Thom E, Sibai B, Wendel G Jr, Wenstrom K, Samuels P, Simhan H, Sorokin Y, Miodovnik M, Meis P, O'Sullivan MJ, Conway D, Wapner RJ; Eunice Kennedy Shriver National Institute of Child Health, Human Development (NICHD) Maternal-Fetal Medicine Units Network (MFMU). Pharmacogenomics of maternal tobacco use: metabolic gene polymorphisms and risk of adverse pregnancy outcomes. Obstet Gynecol. 2010; 115: 568-77.

Arifeen SE, Black RE, Caulfield LE, Antelman G, Baqui AH, Nahar Q, Alamgir S, Mahmud H. Infant growth patterns in the slums of Dhaka in relation to birth weight, intrauterine growth retardation, and prematurity. Am J Clin Nutr. 2000; 72: 1010-7.

Asmussen I, Kjeldsen K. Intimal ultrastructure of human umbilical arteries. Observations on arteries from newborn children of smoking and nonsmoking mothers. Circ Res 1975; 36: 579-89.

Asmussen I. Ultrastructure of the human placenta at term. Observations on placentas from newborn children of smoking and non-smoking mothers. Acta Obstet Gynecol Scand 1977; 56: 119-26.

Chan A, Keane RJ, Robinson JS. The contribution of maternal smoking to preterm birth, small for gestational age and low birthweight among Aboriginal and non-Aboriginal births in South Australia. Med J Aust. 2001;174:389-93.

Coggan M, Whitbread L, Whittington A, Board P. Structure and organization of the human theta-class glutathione S-transferase and D-dopachrome tautomerase gene complex. Biochem J. 1998; 334 (Pt 3):617-23.

de Onis M, Blössner M, Villar J. Levels and patterns of intrauterine growth retardation in developing countries. Eur J Clin Nutr. 1998; 52 Suppl 1: S5-15.

Goldenberg RL and Culhane JF. Low birth weight in the United States. *Am J Clin Nutr* 2007; 85: 584S-590S.

Grazuleviciene R, Danileviciute A, Nadisauskiene R, Vencloviene J. Maternal smoking, GSTM1 and GSTT1 polymorphism and susceptibility to adverse pregnancy outcomes. Int J Environ Res Public Health. 2009; 6: 1282-97.

Horta BL, Victora CG, Menezes AM, Halpern R, Barros FC. Low birthweight, preterm births and intrauterine growth retardation in relation to maternal smoking. Paediatr Perinat Epidemiol. 1997;11:140-51.

Kjell Haram K, Svendsen E, Myking O. Growth Restriction: Etiology, Maternal and Neonatal Outcome. A Review. Current Women's Health Reviews, 2007; 3: 145-160.

Kramer MS. Intrauterine growth and gestation determinants. Pediatrics 1987; 80: 502−511.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181-1186.

Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Kirmeyer S, Munson ML; Centers for Disease Control and Prevention National Center for Health Statistics National Vital Statistics System. Births: final data for 2005. Natl Vital Stat Rep. 2007; 56: 1-103.

Nilsen ST, Sagen N, Kim HC, Bergsjø P. Smoking, hemoglobin levels, and birth weights in normal pregnancies. Am J Obstet Gynecol 1984; 148: 752-8.

Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Andersen AM, Taxbøl D, Hansen KD, Juhl M, Schow TB, Sørensen HT, Andresen J, Mortensen EL, Olesen AW, Søndergaard C. 2001. The Danish National Birth Cohort. Its background, structure and aim. Scand J Public Health 29: 300-307.

Perera FP, Tang D, Rauh V, Lester K, Tsai WY, Tu YH, Weiss L, Hoepner L, King J, Del Priore G, Lederman SA 2005 Relationships among polycyclic aromatic hydrocarbon-DNA adducts, proximity to the World Trade Center, and effects on fetal growth. Environ Health Perspect 113:1062–7.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ronco AM, Arguello G, Munoz L, Gras N, Llanos M. Metals content in placentas from moderate cigarette consumers: correlation with newborn birth weight. Biometals 2005; 18: 233-41.

Tsui HC, Wu HD, Lin CJ, Wang RY, Chiu HT, Cheng YC, Chiu TH, Wu FY. Prenatal smoking exposure and neonatal DNA damage in relation to birth outcomes. Pediatr Res. 2008; 64:131-4.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

Wang S, Chanock S, Tang D, Li Z, Jedrychowski W, Perera FP. Assessment of interactions between PAH exposure and genetic polymorphisms on PAH-DNA adducts in African

American, Dominican, and Caucasian mothers and newborns. Cancer Epidemiol Biomarkers Prev. 2008; 17: 405-13.

Wang X, Zuckerman B, Pearson C, Kaufman G, Chen C, Wang G, Niu T, Wise PH, Bauchner H, Xu X. Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. JAMA. 2002; 287: 195-202.

Wu FY, Wu HD, Yang HL, Kuo HW, Ying JC, Lin CJ, Yang CC, Lin LY, Chiu TH, Lai JS. Associations among genetic susceptibility, DNA damage, and pregnancy outcomes of expectant mothers exposed to environmental tobacco smoke. Sci Total Environ. 2007; 386: 124-33.

Wu J, Hou H, Ritz B, Chen Y. Exposure to polycyclic aromatic hydrocarbons and missed abortion in early pregnancy in a Chinese population. Sci Total Environ. 2010; 408: 2312-8.

5.0 METHODS FOR HOMOZYGOSITY MAPPING IN INBRED FAMILIES COMBINING DENSE SNP DATA WITH A NON-PARAMETRIC LINKAGE ANALYSIS PARADIGM

Homozygosity mapping is a method for mapping genes for recessive diseases in inbred families or populations. It relies on the idea that an individual affected by a recessive disease is likely to have both chromosomes identical at or near the region of the disease gene. This identity of the two chromosomes in an individual is due to the fact that they are inherited from a common ancestor, also known as homozygosity by descent (HBD). Another fundamental for homozygosity mapping is that the regions of HBD in distinct affected individuals are likely to be shared identical by descent (IBD) between those individuals. The disease genes are therefore more likely to be at loci which are HBD within an affected individual and IBD across affected individuals.

Depending on the relationships of the study subjects, homozygosity mapping can be conducted in two types of samples: family data or population samples with undocumented relationships. Homozygosity mapping in consanguineous or inbred families, which refers to parents who have at least one common ancestor within a few generations, is a very common method for mapping recessive disorders [e.g. Saar et al, 1999; Brooks et al, 2005; Knight et al, 2008; Mochida et al, 2010; Rizel et al, 2011; Kousar et al, 2011]. For example, Brooks et al.

mapped a locus for Goldberg-Shprintzen syndrome, a rare recessive disorder, by searching for homozygous chromosome regions in an inbred family with several consanguineous loops. They used a panel of microsatellite markers (381 markers) to cover the whole autosomal genome, with an average spacing of 10 cM. They used software to check the genotypes of each marker in affected individuals. Regions of homozygosity were reported if two consecutive markers were homozygous across affected individuals. They then refined the identified region by genotyping additional 11 markers within it. Finally they conducted linkage analysis to see if the homozygous region was also linked with the disorder (i.e. IBD between affected individuals). In this study, the authors searched for homozygosity by state (HBS) first, and then tested whether the affected individuals shared more IBD in those HBS regions.

In addition to inbred families, homozygosity mapping can also be done in putative outbred families within a population isolate [e.g Winick et al, 1999; Kahrizi et al, 2009]. For example, Winick et al. (1999) successfully identified a locus for a very rare recessive disorder in three Pingelapese kindreds by homozygosity mapping. The Pingelapese live in isolated small islands, and are assumed to have had fewer than 20 founders about 220 years ago. Although participants in this study were not from known consanguineous families, they were distantly related. Using a DNA pooling strategy, the authors first searched for markers that shifted toward homozygosity in the affected DNA pool, which is indicated by a reduction in the number of alleles in the affected vs. the unaffected pool. The nominated regions were then genotyped in all individuals and a linkage analysis was conducted. In the linkage peak regions, they manually looked for the homozygous haplotypes across all affected individuals. Thus, in this study they searched for HBS at each marker separately in affected vs. unaffected pools first, followed by examining IBD sharing in those HBS regions, and finally checked the HBS of the haplotypes.

In addition to using family data, some studies have also conducted homozygosity mapping in population samples [e.g. Cao et al, 2006; Spiegel et al, 2009; Browning et al, 2010]. The assumption of this type of study is that although the study subjects have no documented relationship, they may still share common ancestors many generations ago. The expected length of HBD segments in this type of study is therefore extremely short compared with those in family data.

Despite the fact that homozygosity mapping is relatively common, computational methods for it are often very ad hoc and/or statistically sub-optimal. The goal of our study is to recommend and test methods (often combinations of existing methods) that can bring more statistical rigor and power to this endeavor. We focus in particular on family data, as opposed to population data, and on dense SNP data rather than microsatellite data.

Three major steps can be considered for any method of homozygosity mapping, and we explore improvements to all three. The first step is to estimate IBD/HBD pairwise and/or overall in study subjects. The second step is to calculate an IBD/HBD sharing statistic that will indicate which regions of the genome are the most likely to harbor the disease gene. The third step is to calculate a p-value for the statistic, which is only relevant for larger sample sizes and not for studies of one or a few families. For IBD/HBD estimation, we first review current methods and then propose two new methods for estimating IBD/HBD in families. For calculating a statistic, we suggest a scoring paradigm based on non-parametric linkage analysis. We also propose a parametric alternative, and compare it with the non-parametric one. For calculating p-values, we first discuss whether it is necessary to calculate a p-value, and then discuss how to calculate it depending on the sample size. Finally, we apply all of the above to homozygosity mapping in two real pedigrees.

The organization of this chapter is as follows. In section 5.1, we discuss current methods for homozygosity mapping. In section 5.2, we introduce the datasets that have been used in this project. In section 5.3, we use a simple nuclear family to explain some basic principles and our methods. In section 5.4, we discuss in theory how these principles might be extended to more complex pedigrees. In section 5.5, we demonstrate how our methods can be adapted to analyze real data for two pedigrees. In section 5.6, we discuss the strong and weak points of our methods.

5.1 CURRENT METHODS FOR HOMOZYGOSITY MAPPING

In the following paragraphs, we will introduce current methods for homozygostity mapping, breaking each down into the three steps discussed above. The algorithms for IBD/HBD estimation can be classified into two groups. One is likelihood based methods using Hidden Markov Models (HMM), such as MERLIN and BEAGLE; however MERLIN and BEAGLE are doing very different things with likelihoods. MERLIN is a pedigree likelihood on a known pedigree, i.e. finding recombination and IBD. Beagle is a HMM on an individual, looking for historical recombination without a known pedigree. The other one is SNP “streak” based methods, which look for long runs of homozygogous genotypes in contiguous markers, such as PLINK and HomozygosityMapper. As for steps two and three, usually linkage analysis generates parametric and/or non-parametric statistics and calculates the p-value, while the other methods do not calculate rigorous statistics. These characteristics of the different programs are summarized in Table 5-1 and described in more detail below.

MERLIN (Multipoint Engine for Rapid Likelihood Inference) [Abecasis et al, 2002] is an analysis package for family data. For homozygosity mapping, many studies [Winick JD et al., 1999; Garshasbi M et al., 2006] used MERLIN or other linkage analysis software to estimate IBD across affected individuals. They then searched for the regions where the affected individuals shared significantly higher IBD. Finally they refined the regions by manually checking for homologous haplotypes within each affected individual.

To estimate IBD, MERLIN uses the Lander-Green algorithm, which uses a HMM and assumes linkage equilibrium between markers. MERLIN cannot directly work with high-density SNP data, because the markers are in linkage disequilibrium (LD) and because of computational limitations on the number of markers in can handle. To handle LD in high-density SNP data and reduce the number of markers, MERLIN used a method of clustering of markers. This method assumes no recombination within clusters and no linkage disequilibrium between clusters. However the suitability of this method depends on the availability of either HapMap data or other large datasets to allow estimation of LD patterns.

There are several additional limitations in the usefulness of MERLIN for estimating IBD/HBD using dense SNPs for homozygosity mapping. These include the following. 1) MERLIN does not model genotyping errors. It considers the genotyping error as a double recombination and needs to be manually fixed. 2) MERLIN requires that the pedigree structure be known and specified; however the pedigree structure is not always known. 3) It estimates the number of alleles shared IBD among relatives in a pedigree instead of HBD; HBD has to be manually checked for. 4) The computational load increase linearly with numbers of markers, but exponentially with the number of pedigree members. Even with a recently released MERLIN (1.1.2), the maximum pedigree size is limited to 24 bits which is calculated by $2N - F$; N is the

number of non-founders, *F* is the number of founders. Therefore the large pedigrees, especially for those with a high level of inbreeding, have to be split into smaller subunits. However, splitting families can lead to serious loss of information.

MERLIN can calculate a p-value based on the above statistic. Assuming a large sample approximation, linkage analysis methods can use either a z test for non-parametric statistics (NPL all and NPL pairs) or a chi-square test for parametric statistics (LOD score) to calculate a p-value. However, both statistics are for IBD sharing, not for HBD.

HomozygosityMapper [Seelowet al, 2009] is a commonly used software package for homozygosity mapping in population or family data, which uses a SNP streak based method to estimate IBD/HBD. It reports a score, but it does not specify what that score is and how it is calculated. Also, it does not calculate a p-value.

To estimate IBD/HBD, HomozygosityMapper searches for long runs of HBS (in order to identify HBD) in a row. In another word, it screens all samples for blocks of homozygous genotypes in contiguous markers, where the frequency of homozygosity is different in affecteds and unaffecteds. It allows for genotyping errors because the runs of homozygosity do not need to be perfect. However, it focuses on identifying HBD within an individual; it ignores IBD between individuals. Therefore, the HBD regions shared by affected individuals reported by HomozygosityMapper may contain opposite homozygous genotypes, such as AA and BB for a given locus. HomozygosityMapper does not consider family structure and allele frequencies; all individuals are considered separately. The selection of window sizes may influence the estimation, but complete information is not available for this software. It presumably does not account for LD among the markers, though this is less necessary in a SNP streak method than in a likelihood-based method.

PLINK [Purcell et al, 2007] is another commonly used software package for homozygosity mapping in population or family data, which also uses a SNP streak based method to estimate IBD/HBD. PLINK calculates neither an IBD/HBD sharing statistic nor a p-value.

To estimate HBD, PLINK first detects HBS within each individual using sliding windows, then pair-wise compares the segments in all samples to identify the matched overlapping regions of IBD. PLINK allows for genotyping errors. Also it pays attention to allelic matching between individuals. However, it is not sufficient for IBD/HBD detection in consanguineous families. Firstly, it was designed for population data. It ignores the relationship among individuals. Secondly, the selection of window sizes and thresholds for HBS/HBD regions within an individual are arbitrary, which may be affected by the SNP density and expected size of homozygous segments. It also ignores the varied distances between SNPs. It allows for genotyping errors, but this can result in failure to detect small segments of HBD. Thirdly, it does not thoroughly estimate IBD after detection of HBD; it only pools the allele matching results for those individuals who are HBD at a locus.

BEAGLE [Browning et al, 2010] is a software package that estimates IBD/HBD in population data pairwise. It is a likelihood based algorithm, which estimates haplotypes in each individual to model both LD between markers and IBD between individuals using a modified HMM. It assumes that neither the affected individuals nor their parents are related; therefore it might not work optimally in consanguineous family data. BEAGLE does not calculate an IBD/HBD sharing statistics or a p-value for the statistic.

Table 5-1. Commonly used software in homozygosity mapping

Software	Step 1. Estimate IBD/HBD					Step 2. Calculate a statistics	Step 3. Calculate a p-value
	Data type	Algorithms	Stratigies	Model LD	Limitation		
MERLIN	Family	Likelihood based	IBD estimation for linkage analysis; then detection of homozygous haplotypes.	No	Not suitable for complex inbred family; linkage analysis is less powerful than homozygosity mapping; complete pedigree data is needed.	Yes, for IBD	Yes, for IBD
Homozygosity Mapper	Population	SNP streak	HBD estimation in multiple individuals simultaneously; then association test for disorder.	No	alleles are not matched in estimation of HBD/IBD	Unknown	No
PLINK	Population	SNP streak	HBD estimation for each individual; then pairwise comparison.	No	Estimate HBD based on single individual, family relationship cannot be incorporated.	No	No
BEAGLE	Population	Likelihood based	Haplotype based estimation of IBD/HBD, then pairwise comparison.	Yes	need to estimate haplotype first	No	No

5.2 DATASETS USED IN THIS STUDY

We used both simulated and real datasets in this study. We first used a dense SNP (GWAS) dataset, the GENEVA dental caries dataset, to develop and test the methods for estimation of IBD. The phenotype data in this dataset were not used, only genotype and pedigree data. We then used simulated datasets to further test the performance of our methods. Finally, we applied our methods to identify disease genes in two real inbred pedigrees that were genotyped on a SNP linkage panel (approximately 6,000 markers).

5.2.1 Geneva Dental Caries Dataset

The GENEVA dental caries study (http://www.ncbi.nlm.nih.gov/gap?term=geneva) is a large community-based study of oral health genotyped on the Illumina HumanHap610 chip. We selected samples from nuclear families with two or more offspring. We used this dataset to 1) visualize the pattern of IBD configurations across siblings from a nuclear family using plots; 2) roughly check whether the estimated IBD configurations from our methods were consistent with what we observed in the plots. 3) train our methods for IBD estimation and optimize the parameters.

5.2.2 Simulated Datasets

The dental caries dataset is helpful to develop our methods for IBD estimation. However, it does not allow us to examine whether our methods can correctly detect the boundaries of the IBD configurations, since the true IBD configurations in that dataset are unknown. In addition, it is not sufficient for developing methods for HBD estimation in inbred families, since parents do not have any documented relationships. To solve these problems, we simulated high density phased datasets to construct various IBD and HBD configurations.

We first retrieved phased genome-wide genotyping data from a pair of parents (ID NA12748, NA12749) in HapMap Phase III. (http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/CEU/TRIOS/). Next, we simulated the recombination events in meiosis at some previously selected breakpoints

within each parent, and generated the simulated offspring samples with known IBD sharing configurations at exact loci (Figure 5-1).

To observe the influence of marker density on estimation of IBD/HBD, we tested two sets of markers. One is Illumina HumanHap610K (over 610K markers), the other is Illumina 6K linkage panel (over 6K markers). We retrieved the allele frequencies of those markers from the PennCNV PFB (Population frequency of B allele) file: hhall.hg18.pfb, which contains estimates of population (European) allele frequencies for more than 1M markers from the Illumina chips.

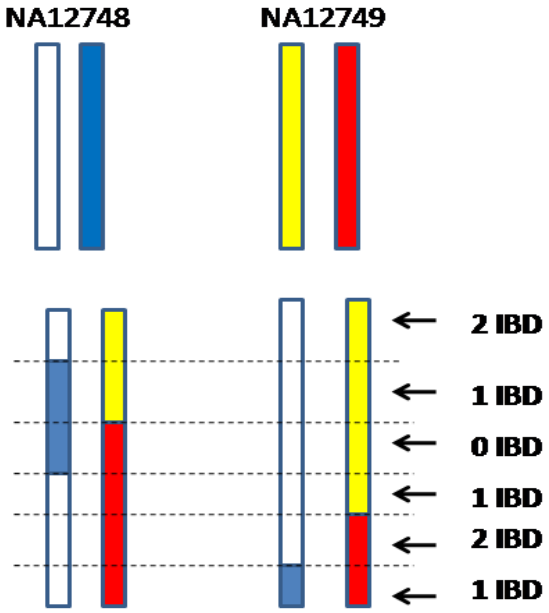


Figure 5-1. Construction of simulated IBD siblings

5.2.3 Inbred Pedigree Dataset

We applied our methods to map disease genes in two real inbred pedigrees. The two pedigrees are shown in Figures 5-2 and 5-3. Both pedigrees are for the same autosomal recessive disorder,

although they are from somewhat different populations. All the samples with a star (*) were genotyped using Human Linkage-24 BeadChip (Infinium assay).

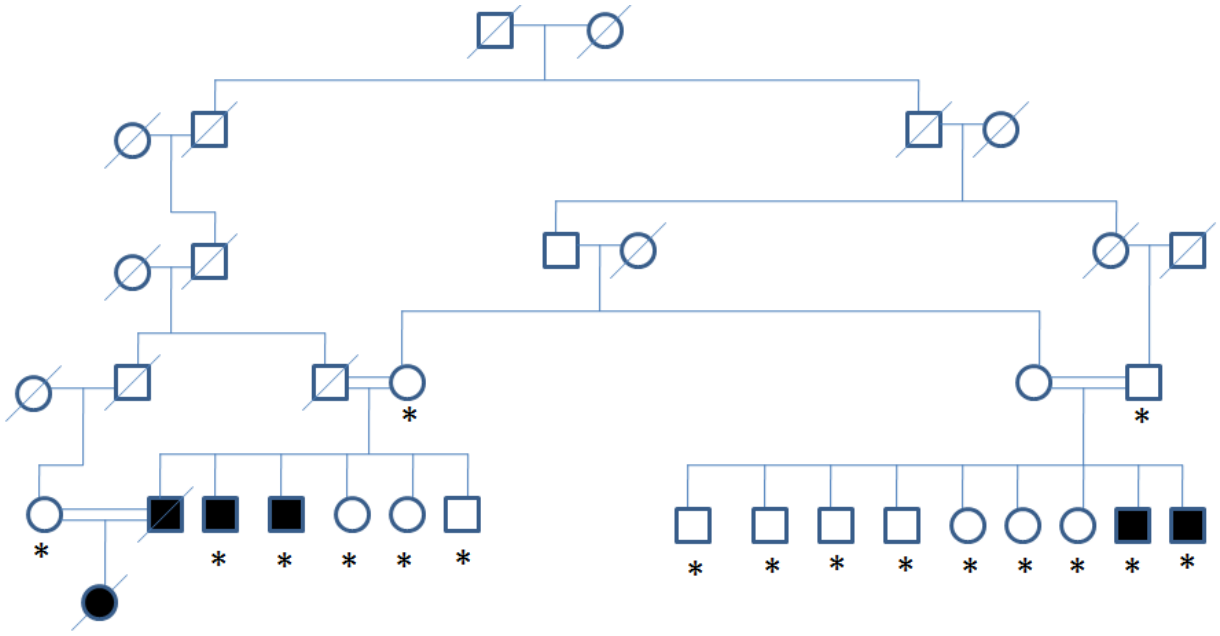


Figure 5-2. Inbred pedigree 1

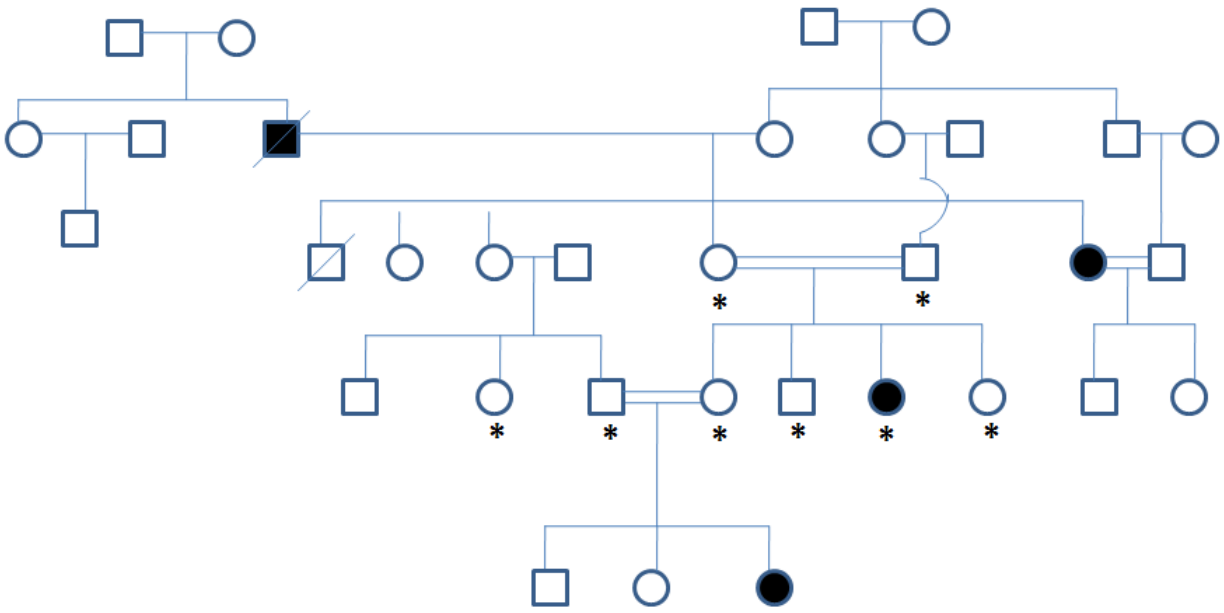


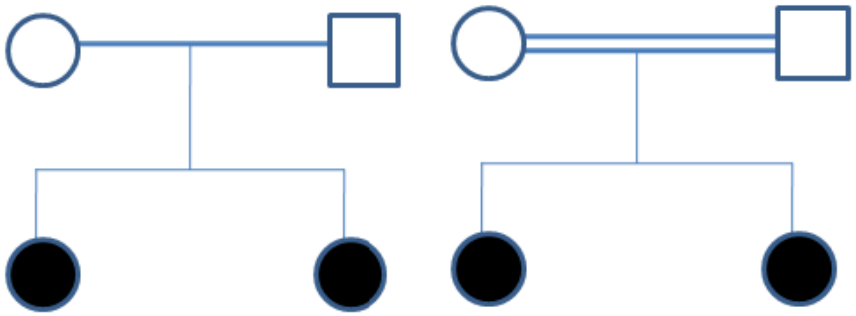
Figure 5-3. Inbred pedigree 2

5.3 OUR METHODS FOR SINGLE NUCLEAR FAMILY DATA WITH A PAIR OF AFFECTED SIBLINGS

We propose a procedure which integrates three steps for homozygosity mapping: estimation of IBD/HBD, calculation of a sharing statistic, and calculation of a p-value for the statistic. In this section we use the example of a single nuclear family with a pair of affected siblings to introduce our methods for each of the three steps.

To illustrate our methods for IBD/HBD estimation, we divide our descriptions into three parts: 1) theoretical model (section 5.3.1 and 5.3.2); 2) simulation analysis (section 5.3.3 and 5.3.4); and 3) real data analysis (section 5.3.5 and 5.3.6). To better understand the methods we

propose for IBD+HBD estimation, we first introduce IBD estimation (in a single family with unrelated parents, Figure 5-4A); and then joint estimation of IBD/HBD (in a single family with related parents, Figure 5-4B). After IBD/HBD estimation, we show how to calculate a sharing statistic (section 5.3.7) and calculate a p-value (section 5.3.8).



(A) (B)

Figure 5-4. (A) Single nuclear family with unaffected unrelated parents and a pair of affected children. (B) Single nuclear family with unaffected closely related parents (relationship unknown) and a pair of affected children.

5.3.1 Methods for Estimation of IBD

Our goal is to use SNP data from a whole chromosome in a nuclear family and estimate locations of meiotic recombination, or equivalently, estimate IBD between the pair of siblings at each point on the chromosome. At each point on the chromosome, two siblings have 1/4 chance to share 2 IBD, 1/2 chance to share 1 IBD, 1/4 chance to share 0 IBD. If no recombination happens during meiosis, the IBD region shared by two siblings should be a whole chromosome long. However, the regions of IBD are usually broken by recombination. The average recombination rate is one per morgan (approximately 10^8 base pairs). A standard Poisson process proposed by Haldane [Ott 1985] has been used to model the recombination event.

In practice, IBD must be estimated in some way from identity by state (IBS), or allele matching. Table 5-2 lists the possible configurations of IBS (identity by state) for a pair of siblings at a given locus. Table 5-3 summarizes all possible IBD states for the pair of siblings, and shows the corresponding possible IBS states for each IBD state. For example, for 2 alleles IBD, the IBS configurations can only be AA / AA or AB /AB; however, for 1 allele IBD, the IBS configurations can be AA / AB, AA /AA or AB /AB.

Table 5-2. IBS configurations between a pair of siblings

IBS states	Observed IBS configurations *	
4	AA	AA
3	AA	AB
2	AA	BB
1	AB	AB

*AA, AB and BB in IBS configurations represent observed SNP genotypes at a given locus.

Table 5-3. IBD configurations and all possible corresponding IBS states between a pair of siblings

Hidden IBD states	IBD configurations *	Corresponding IBS states
a	13 13	4,1
b	13 14	4,3,1
c	13 24	4,3,2,1

* 13, 14, 24 in IBD configurations represent allele sharing status, given parents IBD configuration is 12 and 34 at a given locus. Same number represents same allele, different numbers indicate different allele.

If we plot the IBS configurations for each marker along a chromosome as in Figure 5-5, we can view combinations of IBS states that can be used to infer the IBD states. For example, at the left hand side of Figure 5-5 we observe the simultaneous existence of IBS states 1, 2, 3, and 4. They are not literally simultaneous – any given marker is in only one state - but markers in those four states are densely interspersed, giving the impression of parallel lines in the plot. This corresponds to IBD state c (0 alleles shared IBD between the sibs). The estimated IBD states for Figure 5-5 are c, b, a, b, c, b, moving from left to right. However, the estimation based on visual inspection of graphs is rough; it may be influenced by genotyping errors and marker densities. Also the breakpoints of the IBD states are difficult to determine. To efficiently and accurately estimate the hidden IBD configurations, we developed two algorithmic methods. One is a SNP streak method; the other is a HMM based method.

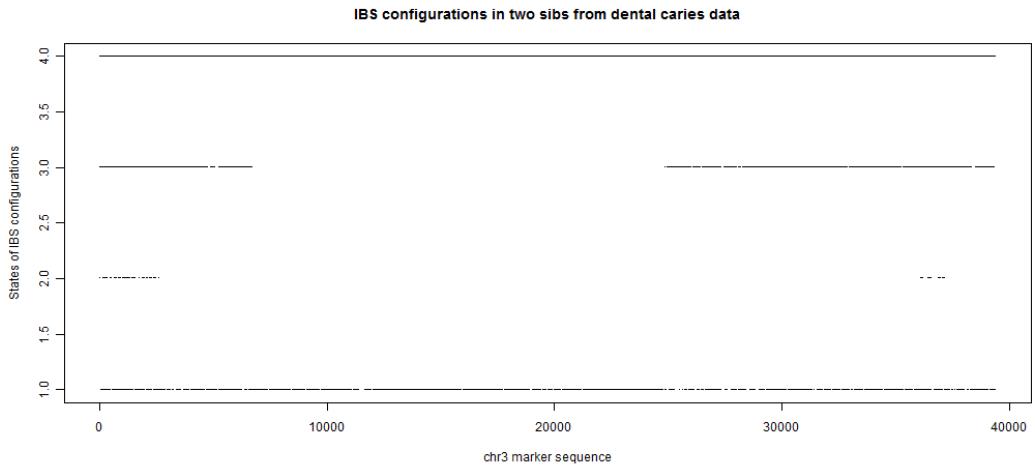


Figure 5-5. IBS configurations on chromosome 3 for two siblings from dental caries dataset

Our SNP streak method takes advantage of two features to define the breakpoints of various IBD states. It is shown in Table 5-3 that IBS states 4 and 1 are common to all IBD

configurations; IBS state 3 is common to IBD configurations b and c; while IBS state 2 is unique to IBD configuration c. Also, under the assumption that the recombination events in a pair of siblings follow a Poison process, two recombinations cannot occur simultaneously at any marker. So IBD state can only switch between “a” and “b”, or “b” and “c”. Taking advantage of these two features, the appearance or disappearance of IBS state “3” can serve as a breakpoint between IBD state “a” and state “b”; while the appearance or disappearance of IBS state “2” can serve as a breakpoint between IBD state “b” and state “c”.

The SNP streak method uses a sliding window, with window size depending on the density of SNPs in the data, to estimate the IBD. We allow for genotyping errors in the following way. For any given window, we first look for genotyping errors. If the occurrence of IBS state “2” or “3” is only once in a given window, we check a pre-set number of markers (the number of markers depends on the selected error rate) before and after that marker; if the singular IBS state does not appear in any of those markers, it is considered as a genotyping error. After excluding genotyping errors, we use following algorithm to detect IBD. Detailed information is described in Appendix C.

1) In the initial window, we get an IBD state call according to IBS states in that window; we then assign that IBD state to all markers on the chromosome.

2) Next, we slide the window to find the breakpoint for IBD state change.

3) Once we find a breakpoint, we will replace the IBD state of markers after that breakpoint (including that point) with the new IBD state.

4) We then repeat procedure 2), until the end of the markers.

The SNP streak algorithm does not use information on SNP allele frequencies. In addition, the choice of window size is arbitrary. Selection of a large window size may miss some small 2 IBD regions; a small window size will generate some false 2 IBD regions.

Instead of using a SNP streak method as described above, it is also possible to use hidden Markov models to find regions of IBD. The Markov model is a statistical technique that models a Markov process, where the probability of observing a particular state at a particular time point only depends on the state at the previous time point. In a hidden Markov model, the states cannot be directly observed, therefore are "hidden." In this project, we used a discrete form first order hidden Markov model (HMM) [Rabiner et al, 1989; Petrushin 2000, Nikolai Shokhirev 2010] to identify the IBD configurations between a pair of siblings. The literal form of the model assumes that the markers are not in LD, but we account for LD by adjusting model parameters.

A hidden Markov model requires specifying the hidden states, the transition probabilities of the true Markov model, the observed states, the probability distribution relating the hidden and observed states (emission probabilities) and the initial probability. The hidden state in our algorithm is the IBD sharing configuration at a given locus. These are listed in Table 5-3 column 2. The observed states are the IBS configurations. These are summarized in Table 5-2. The time points are the SNP markers ordered by their physical location on a chromosome. Although the hidden state is not observable, we can calculate the emission probability according the allele frequency at a marker. The emission probabilities are listed in Table 5-4 under the assumption that the two parents are not inbred and are unrelated (i.e. the four parental alleles are

independent. q is the minor allele frequency at a given SNP marker, $p = 1 - q$, and ε is the total genotyping error rate for those individuals.

Table 5-4. Emission probabilities (Probabilities of IBS states given an IBD state) of HMM in a pair of siblings.

IBD state	IBS state			
	1	2	3	4
a	$2pq\,(1 - \varepsilon)$	0	ε	$(p^2 + q^2)(1 - \varepsilon)$
b	$(p^2q + pq^2)(1 - \varepsilon) + 2/3(p^2q + pq^2)\varepsilon$	$2/3(p^2q + pq^2)\varepsilon$	$(2p^2q + 2pq^2)(1 - \varepsilon) + (p^3 + q^3)\varepsilon + (p^2q + pq^2)\varepsilon$	$(p^3 + q^3)(1 - \varepsilon) + 2/3(p^2q + pq^2)\varepsilon$
c	$(2p^2q^2)\,(1 - \varepsilon) + 4/3(p^3q + pq^3)\varepsilon$	$(4\,p^2q^2)(1 - \varepsilon) + 4/3(p^3q + pq^3)\varepsilon$	$(4p^3q + 4pq^3)(1 - \varepsilon) + (p^4 + q^4)\varepsilon + 6p^2q^2\varepsilon$	$(p^4 + q^4)(1 - \varepsilon) + 4/3(p^3q + pq^3)\varepsilon$

The transition probability describes the underlying Markov process - the probability of having an IBD sharing state change between two adjacent SNPs. The transitions of the IBD sharing states are due to homologous recombination during meiosis. Let λ be the recombination rate in morgans. If we assume equal genetic distance between markers, the transition probabilities would be as in Table 5-5. However, in real SNP arrays, the distance between markers varies. In general, IBD state is unlikely to change for SNPs that are nearby but is more likely to change for SNPs that are far apart. To accommodate heterogeneous distances into the HMM, Marioni [Marioni et al. 2006] proposed a modified transition matrix. It was designed to model the non-linear changes in the probabilistic structure of the transition matrix caused by the heterogeneous distance. However in our model, there is an approximately linear relationship between the changes of recombination rate and adjacent SNP distances, so we simplified

Marioni's equation. Let d_i denote the physical distance (base pairs) between two adjacent SNPs i and i +1. We assume the recombination rate is one per Morgan, and assume 100 Mb is approximately equivalent to one Morgan. Let D be a constant that was set as 100 Mb, the transition probability can be modeled as in Table 5-6. The hidden IBD states were then identified using the Viterbi algorithm [Forney et al, 1973; Shokhirev N, 2010]. All computation were programmed and completed in R version 2.10.1 [R Development Core Team, 2008].

The use of the hidden Markov model also requires specifying the initial probabilities of the Markov process. We used equal probability for each IBD configuration as the initial probability; in our example of a pair of siblings, it was 1/3, 1/3, and 1/3. An alternative initial probability could be the probability of each IBD configuration given the relationship of the pedigree under the null hypothesis of no linkage of the locus with disease; in our example, for 2 IBD, 1 IBD and 0 IBD, it would be 1/4, 1/2, and 1/4.

We tested both the HMM method and the SNP streak method on the dental caries dataset. The estimated IBD states by both methods were consistent with what we observed based on Figure 5-5. However, we cannot judge whether they are completely correct in this dataset, because we don't know the truth. We therefore use simulated data to examine that later.

Table 5-5. Transition probabilities of hidden states

	a	b	c
a	1-4λ	4λ	0
b	2λ	1-4λ	2λ
c	0	4λ	1-4λ

Table 5-6. Transition probability of heterogeneous HMM

	a	b	c
a	1 - 4d_i/D	4d_i/D	0
b	2d_i/D	1 - 4d_i/D	2d_i/D
c	0	4d_i/D	1 - 4d_i/D

5.3.2 Methods for Estimation of IBD and HBD Simultaneously

The methods described above can be extended to analyze a two-child nuclear family with related parents (Figure 5-4 B) and estimate both IBD and HBD (IBD+HBD). For a given locus, all possible IBS + HBS configurations for a pair of siblings from inbred parents are listed in Table 5-7. In Table 5-8, we summarize the possible IBD + HBD states and all possible IBS + HBS states for each given IBD + HBD state.

Table 5-7. Observed IBS + HBS configurations

IBS + HBS states	Observed IBS + HBS configurations	
4	AA	AA
3	AA	AB
2	AA	BB
1	AB	AB

Table 5-8. IBD + HBD configurations and Corresponding IBS + HBS states

Hidden IBD + HBD states	IBD + HBD configurations	Corresponding IBS + HBS states
a	13 13	4,1
b	13 14	4,3,1
c	13 24	4,3,2,1
d	11 11	4
e	11 13	4,3
f	11 23	4,3,2

Our SNP streak method for IBD + HBD calling is similar to the one used in IBD estimation for the outbred family above. We define IBD + HBD states according to the IBS + HBS states seen in a window.

Our HMM method described above can also be extended to consider both IBD and HBD. Again, we must specify true states, initial probabilities, and transition probabilities, hidden states, and emission probabilities. The observed states are the homozygosity by genotype state (HBS) + IBS shown in Table 5-7. The hidden states are the HBD+IBD sharing configurations at a given locus, which are listed in Table 5-8. Equal probability for each IBD+HBD configuration is used as the initial probability. The emission probability is summarized in appendix I. Calculation of transition probabilities is more complicated than in the previous example. We therefore describe our derivation for the transition probabilities in detail. The transition process is not a real Markov

process. The current IBD+HBD state in a pair of siblings does depend not only on their previous state of IBD+HBD, but also on the IBD sharing states between parents. The former is determined by recombination events in the parents; the latter is determined by the relationship between the parents and thus by earlier recombination events. However, we use a model that assumes the transitions are Markov. We first calculate the approximate transition matrix of IBD states in parents. For a pair of parents that are first cousins, let ϕ_0 denote the event that parents sharing 0 IBD, ϕ_1 denote that parents sharing 1 IBD, we have $P(\phi_0) = 3/4$, $P(\phi_1) = 1/4$.

Assuming Haldane's model, the g generations after founding, single-path IBD tracts would have approximately exponential length distributions with a mean of $1 / (2g)$ Morgan. For first cousin parents, $g = 2$, the expected IBD length $A = 1/2\ g = 1/4$ Morgan = 100/4 cM. Let B denote the expected length of non-IBD tracts. Since $A / (A + B) = 1/4$, solving it, we get B = 300/4 cM. So the expected non-IBD length B = 300/4 cM. Therefore, in this pair of closely related parents, the transition from IBD to non-IBD is t_0 = 4/100 per CM; the transition from IBD to IBD is $1 - t_0$; the transition from non-IBD to IBD is t_1 = 4/300 per CM; the transition from non-IBD to non-IBD is $1 - t_1$. Feingold [1993] proposed a method to calculate the exact transition probabilities for a function of a Markov chain. Using her method, we got exactly the same results here.

Next, we calculate the IBD+HBD transition matrix for a pair of siblings given the IBD state in the parents. Again according to the Poisson process model, we assume that at any time point, the transition of IBD in parents and recombination in meiosis will not happen at the same time. Let Ω denote the whole transition matrix for a pair of siblings; ϕ_{00} denotes transition from non-IBD to non-IBD in parents; ϕ_{11} denotes transition from IBD to IBD in parents; ϕ_{01} denotes

transition from non-IBD to IBD in parents; ϕ_{10} denotes transition from IBD to non-IBD in parents. Let i denotes a previous IBD+HBD state; j denotes a current state. ω_{ij} denotes the transition from i to j.

$\omega_{ij} = P(\omega_{ij}|\phi_0)\ \ P(\phi_0) + P(\omega_{ij}|\phi_1)\ \ P(\phi_1)$

$= [P(\omega_{ij}|\phi_{00})P(\phi_{00}) + P(\omega_{ij}|\phi_{01})P(\phi_{01})]\ P(\phi_0) + [P(\omega_{ij}|\phi_{11})P(\phi_{11}) + P(\omega_{ij}|\phi_{10})P(\phi_{10})]\ P(\phi_1)$

We can eventually get the transition matrix (Table 5-9). This transition matrix is based on the assumption that the parents are first cousins. We also calculated the transition matrix by assuming the parents are second cousind, and tested both in simulated and real data. We got the same estimated HBD+IBD segments in the children, no matter the assumed relationship of the parents. So our model is robust to this assumption, at least as between first and second cousin relationships.

Table 5-9. Transition probabilities of hidden states

	a	b	c	d	e	f
a	1-3 (1- t_1) d_i/D - (1-t_0) d_i/D - 3/16t_1	3 (1-t_1) d_i/D + 2/3(1- t_0) d_i/D	0	3/16t_1	1/3(1- t_0) d_i/D	0
b	3/2(1-t_1) d_i/D + 1/3(1-t_0) d_i/D	1- 3(1- t_1) d_i/D - 2/3(1- t_0) d_i/D -3/8 t_1	3/2(1- t_1) d_i/D	0	3/8 t_1 + 1/6(1- t_0) d_i/D	1/6(1- t_0) d_i/D
c	0	3(1- t_1) d_i/D + 3/8 t_1	1-3(1- t_1) d_i/D - 3/4 t_1	0	0	3/8 t_1
d	1/4t_0	0	0	1-1/4t_0 - (1-t_0) d_i/D	(1- t_0) d_i/D	0
e	1/4(1- t_0) d_i/D	1/4[(1-t_0) d_i/D + t_0]	0	1/4(1- t_0) d_i/D	1-(1-t_0) d_i/D -1/4t_0	1/4(1- t_0) d_i/D
f	0	1/2(1- t_0) d_i/D	1/4t_0	0	1/2(1- t_0) d_i/D	1- (1-t_0) d_i/D - 1/4t_0

5.3.3 Simulation Study to Compare SNP Streak and HMM Methods for IBD Estimation

In section 5.3.1 we introduced two methods for IBD estimation in a two-child nuclear family with unrelated parents: a HMM-based method and a SNP streak based method. In this section we test their accuracy using simulated data. We did the simulation study qualitatively. We randomly selected the breakpoints for each chromosome and ran the simulation tests across all autosomal chromosomes in human genome. In this section we only show the results on chromosome 3 as an example. For a more rigorous simulation test, we should run the test in a large number of times and report the percentage of correct IBD "calls." Since the model assumes that the markers are in linkage equilibrium, which is not true for high density markers, we evaluate the performance of our methods using two marker sets with different density: Illumina 6K linkage panel and Illumina HumanHap 610K. We discuss performance in the Illumina 6K linkage panel first, and the Illumina HumanHap 610K next. In each marker set, we simulated IBD configurations first; then estimated the IBD configurations using our two methods; and finally compared the estimated IBD with the known simulated IBD configurations to evaluate the accuracy of our methods.

Figure 5-6 is the plot of IBS configurations of simulated sibling pairs for the Illumina 6K linkage panel markers on chromosome 3. We can roughly see from Figure 5-6 that there are 6 different segments of IBS configuration combinations, which may imply 6 different IBD states. The true simulated IBD states are listed in Table 5-10. Inferred IBD by our SNP streak method with window size 15 SNPs (average distance is 11 Mb) shown in Table 5-11. Inferred IBD by

our HMM model is shown in Table 5-12. From the above comparison we can see that both methods work very well for the simulated data with 6K linkage panel SNP markers.

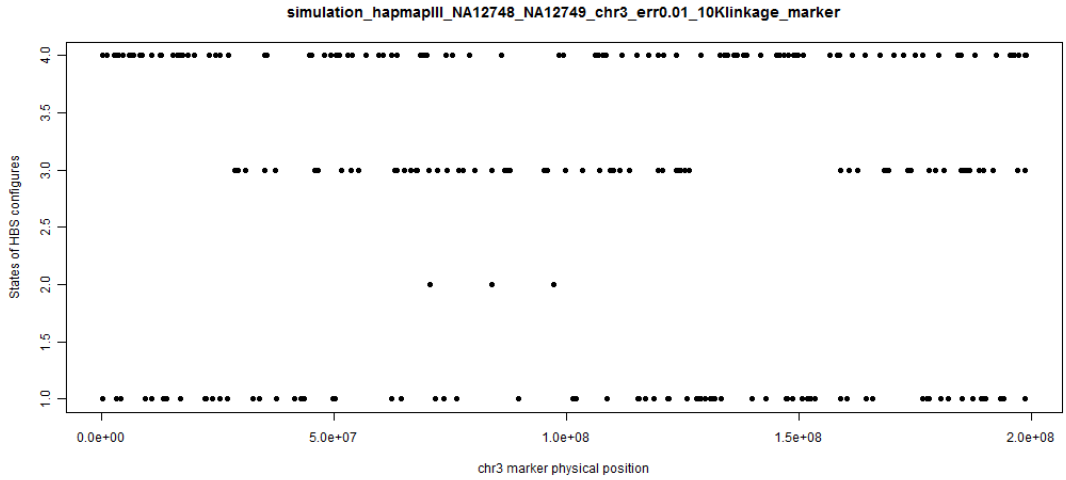


Figure 5-6. The plot of IBS configurations of simulated siblings with Illumina 6K linkage panel markers on chromosome 3.

Table 5-10. The true simulated IBD states

chr	start_snp_index	end_snp_index	IBD_state
3	1	40	1
3	41	80	2
3	81	120	3
3	121	160	2
3	161	200	1
3	201	263	2

Table 5-11. Inferred IBD by SNP streak method

chr	start_snp_index	end_snp_index	IBD_state	start_position	end_position
3	1	40	1	177,033	27,228,974
3	41	91	2	28,649,284	70,395,473
3	92	120	3	70,528,666	97,161,899
3	121	161	2	98,368,882	127,977,995
3	162	200	1	128,119,045	158,772,629
3	201	263	2	158,876,463	198,707,094

Table 5-12. Inferred IBD by our HMM model

chr	start_snp_index	end_snp_index	IBD_state	start_position	end_position
3	1	40	1	177,033	27,228,974
3	41	90	2	28,649,284	69,831,064
3	91	120	3	70,395,473	97,161,899
3	121	160	2	98,368,882	126,294,003
3	161	197	1	127,977,995	156,507,016
3	198	263	2	158,239,353	198,707,094

We next evaluate the performance of those two methods using the Illumina HumanHap 610K markers. This marker set has high density - 610K markers genome-wide. Figure 5-7 shows IBS configurations of simulated sibling pairs using the Illumina HumanHap 610K markers on chromosome 3. We can see from Figure 5-7 that there are 6 different segments of IBS configuration combinations, which may imply 6 different IBD states. Table 5-13 lists the true simulated IBD configurations. Table 5-14 lists the inferred IBD by SNP streak with window size of 400SNPs (average distance is ~2.2 Mb). The HMM model reported a lot of small IBD segments for the high density 610K markers. We therefore have to optimize the parameter D. If we set $D = 10^{21}$, then our HMM method works very well, which actually implies that using the incorrect value for D makes up for the model "error" of assuming no LD. Table 5-15 shows the

inferred IBD by our HMM method after setting D = 10^{21}. The estimated IBDs by the two methods are consistent with the real IBD configurations. We also found that the two choices of initial probability have no effect on the IBD estimation.

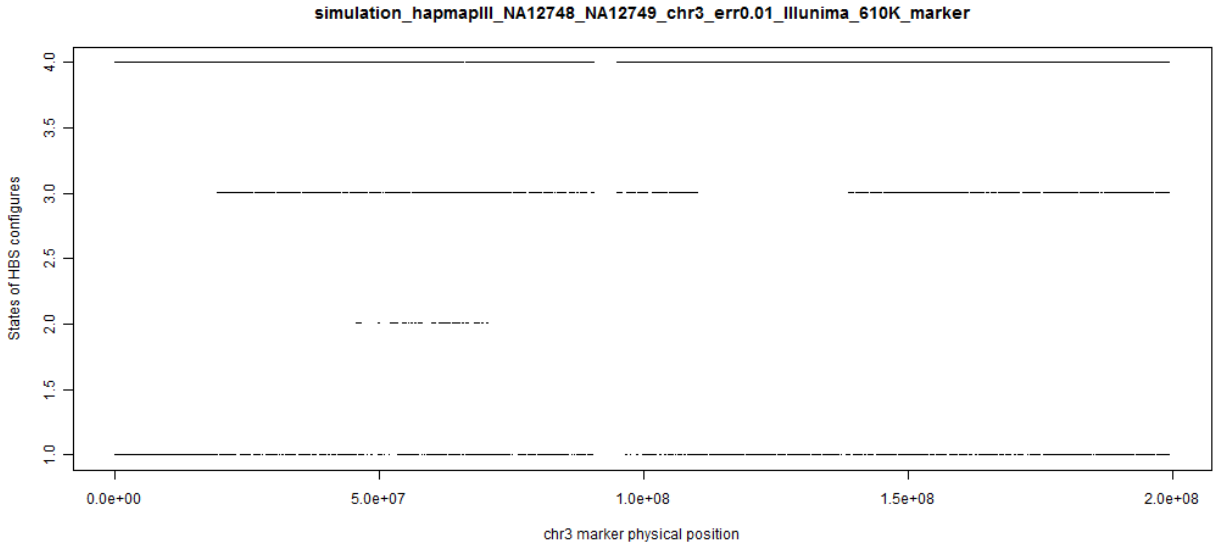


Figure 5-7. The plot of IBS configurations of simulated paired siblings by using Illumina HumanHap 610K markers on chromosome 3.

Table 5-13. The true simulated IBD states

chr	seq_start	seq_end	IBD_state
3	1	5000	1
3	5001	10000	2
3	10001	15000	3
3	15001	20000	2
3	20001	25000	1
3	25001	35684	2

Table 5-14. Inferred IBD by SNP streak method

chr	seq_start	seq_end	IBD_state	start_position	end_position
3	1	5001	1	38,411	19,458,699
3	5002	10098	2	19,487,078	45,785,098
3	10099	14957	3	45,787,518	70,675,771
3	14958	19997	2	70,678,539	110,141,846
3	19998	25042	1	110,145,770	138,682,404
3	25043	35684	2	138,685,339	199,348,860

Table 5-15. Inferred IBD by our HMM method ($D = 10^{21}$)

chr	seq_start	seq_end	IBD_state	start_position	end_position
3	1	4999	1	38,411	19,449,331
3	5000	10086	2	19,455,317	45,737,729
3	10087	14981	3	45,767,213	70,817,227
3	14982	19996	2	70,849,409	110,134,811
3	19997	25036	1	110,141,846	138,642,262
3	25037	35684	2	138,672,290	199,348,860

5.3.4 Simulation Study to Compare Two Methods for IBD+ HBD Estimation

In this section, we use simulated data very similarly to what was done above to evaluate the performance of those two methods on joint estimation of IBD+HBD in the two-sibling nuclear family with related parents. Again we use two sets of markers with different densities. We use Illumina 6K markers first, followed by Illumina HumanHap610K. Again in each marker set, we simulate IBD configurations first; followed by estimation of IBD configurations using our two methods; and finally comparison of the estimated IBD with the known simulated IBD configurations to evaluate the accuracy of our methods.

Figure 5-8 is the plot of IBS + HBS configurations of a simulated sibling pair using the 6K linkage panel markers on chromosome 3. It is relatively hard to tell based on Figure 5-8 how

many different patterns of IBS + HBS configurations exist, because the expected IBD+HBD regions are smaller than previous IBD only regions and the markers are not high density. This actually suggests the necessity of statistical methods for estimation of IBD+HBD.

The true simulated IBD + HBD is listed in the Table 5-16. Inferred IBD + HBD by our SNP streak method with window size 15 SNPs is shown in Table 5-17. Inferred IBD + HBD by our HMM model is shown in Table 5-18. The estimated IBD+HBD by SNP streak and HMM methods have highly consistent results with each other and with the true IBD+HBD configurations, which suggests that both of them work very well with the Illumina 6K marker set.

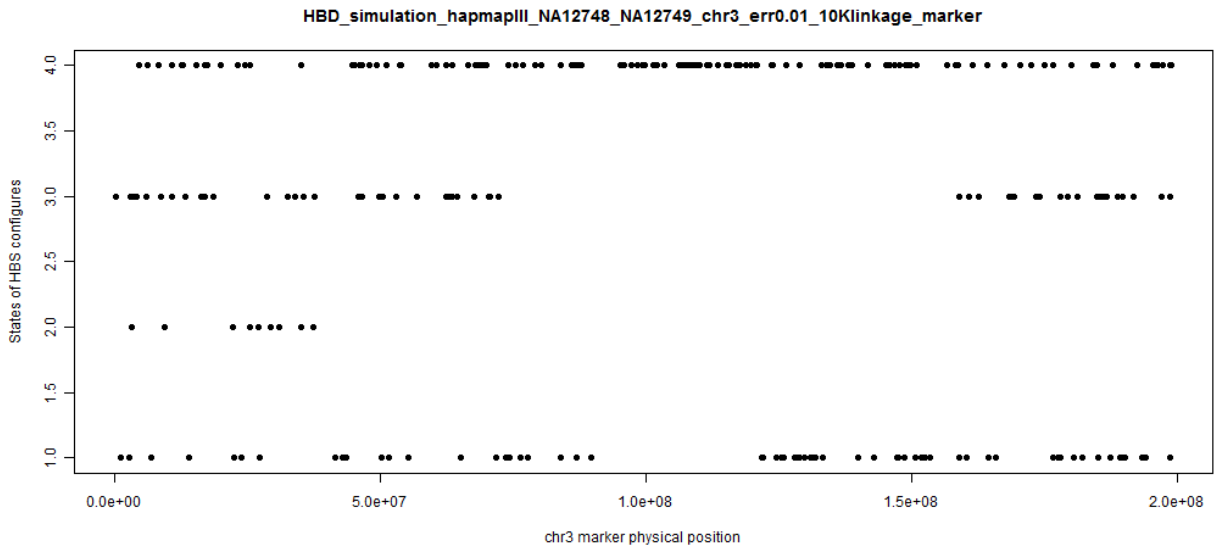


Figure 5-8. The plot of IBS + HBS configurations of simulated siblings with 6K linkage panel markers on chromosome 3.

Table 5-16. The true simulated IBD + HBD states

chr	start_snp_index	end_snp_index	HBD_state
3	1	50	3
3	51	100	2
3	101	120	1
3	121	150	4
3	151	200	1
3	201	263	2

Table 5-17. Inferred IBD + HBD by SNP streak method

chr	start_snp_index	end_snp_index	HBD_state	start_position	end_position
3	1	49	3	177,033	37,286,095
3	50	96	2	37,560,742	72,109,515
3	97	125	1	73,514,520	101,388,492
3	126	150	4	101,916,281	120,791,949
3	151	200	1	121,602,168	158,772,629
3	201	263	2	158,876,463	198,707,094

Table 5-18. Inferred IBD + HBD by HMM method

chr	start_snp_index	end_snp_index	HBD_state	start_position	end_position
3	1	50	3	177,033	37,560,742
3	51	96	2	41,353,624	72,109,515
3	97	116	1	73,514,520	89,589,647
3	117	150	4	95,087,815	120,791,949
3	151	197	1	121,602,168	156,507,016
3	198	263	2	158,239,353	198,707,094

Figure 5-9 is the plot of IBS + HBS configurations of simulated paired siblings by using Illumina 610K markers on chromosome 3. The true simulated IBD + HBD is listed in the Table 5-19. Inferred IBD + HBD by SNP streak method with window size 400 SNPs is shown in Table 5-20. Again, our HMM algorithm reported a lot of small segments of IBD+HBD, due to LD of

high dense markers. After setting the parameter $D = 10^{21}$, the inferred IBD + HBD by our HMM model is shown in Table 5-21. Results from our two methods are consistent with true IBD+HBD configurations in simulated dataset. HMM has a little better prediction of breakpoints.

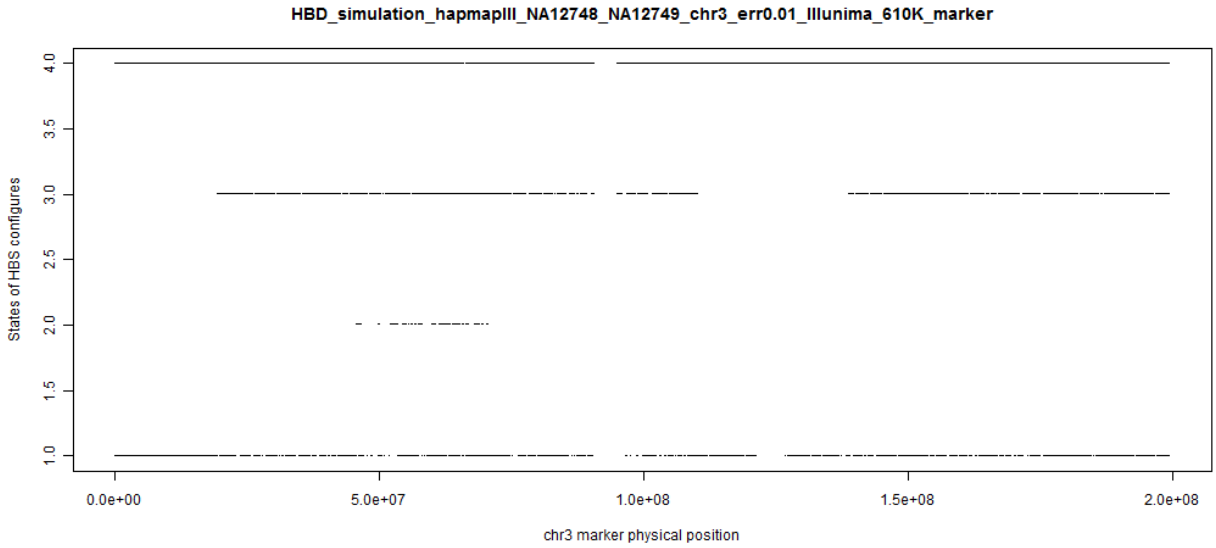


Figure 5-9. The plot of IBS + HBS configurations of simulated siblings with Illumina 610K markers on chromosome 3.

Table 5-19. The true simulated IBD + HBD states

chr	Start_snp_index	End_snp_index	IBD_state
3	1	5000	1
3	5001	10000	2
3	10001	15000	3
3	15001	20000	2
3	20001	22000	1
3	22001	23000	4
3	23001	25000	1
3	25001	35684	2

Table 5-20. Inferred IBD + HBD by SNP streak method

chr	start_snp_index	end_snp_index	HBD_state	start_position	end_position
3	1	5000	1	38,411	19,455,317
3	5001	10097	2	19,458,699	45,784,333
3	10098	14956	3	45,785,098	70,661,075
3	14957	19996	2	70,675,771	110,134,811
3	19997	21991	1	110,141,846	121,306,727
3	21992	23000	4	121,311,202	126,729,053
3	23001	25041	1	126,732,030	138,680,883
3	25042	35684	2	138,682,404	199,348,860

Table 5-21. Inferred IBD + HBD by our HMM model ($D = 10^{21}$)

chr	start_snp_index	end_snp_index	HBD_state	start_position	end_position
3	1	4999	1	38,411	19,449,331
3	5000	10086	2	19,455,317	45,737,729
3	10087	14981	3	45,767,213	70,817,227
3	14982	19996	2	70,849,409	110,134,811
3	19997	21991	1	110,141,846	121,306,727
3	21992	22999	4	121,311,202	126,720,789
3	23000	25036	1	126,729,053	138,642,262
3	25037	35684	2	138,672,290	199,348,860

5.3.5 Inbred Pedigree Data to Compare Two Methods for IBD Estimation

In order to further examine the performance of our two methods for IBD estimation, we used inbred pedigree 1. This data has the merit that it is contributed by many factors and their interactions which may not be modeled in simulated data. But we don't know the true results, so we compared their performance with each other. Compared with dental caries dataset, which used illumine 610K, these inbred pedigrees were genotyped using the Illumina 6K linkage panel array. Figure 5-10 is the plot of IBS configurations for a selected a pair of siblings on chromosome 3 in pedigree 1. The inferred IBD by our SNP streak method with window size 15

SNPs is shown in Table 5-22. Inferred IBD by our HMM model is shown in Table 5-23. The results from the two methods are consistent, except that the first segment of IBD reported by HMM was not reported by the SNP streak method, which is likely due to that segment of IBD is smaller than the window size.

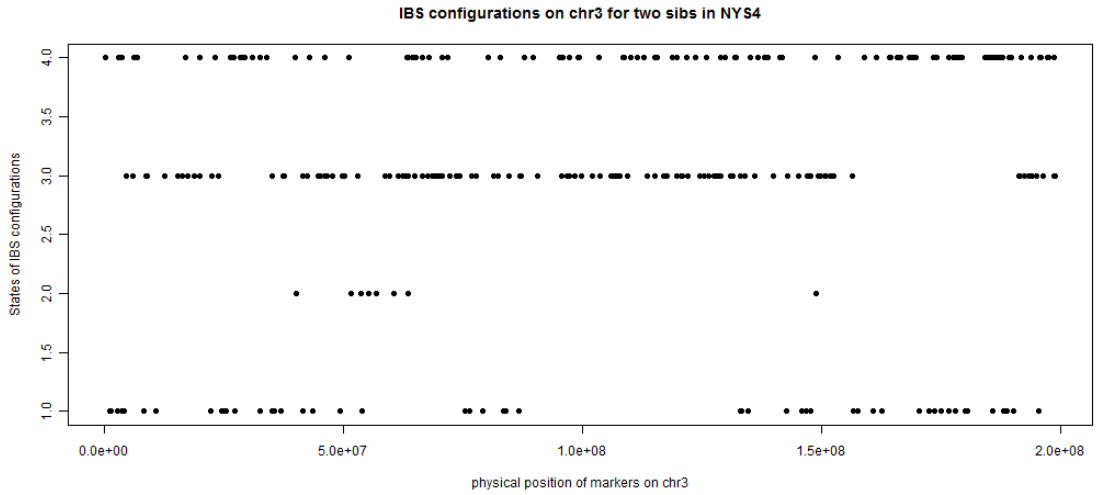


Figure 5-10. IBS configurations of a pair of siblings on chromosome 3 in inbred pedigree 1.

Table 5-22. Inferred IBD by SNP streak method

chr	start_snp_index	end_snp_index	IBD_state
3	1	33	2
3	34	49	1
3	50	72	2
3	73	87	3
3	88	222	2
3	223	274	1
3	275	292	2

Table 5-23. Inferred IBD by HMM method

chr	start_snp_index	end_snp_index	IBD_state	start_position	end_position
3	1	11	1	169,613	4,084,699
3	12	33	2	4,568,187	23,848,605
3	34	47	1	24,449,356	32,530,183
3	48	70	2	33,880,569	49,738,341
3	71	88	3	50,089,518	63,538,921
3	89	223	2	64,402,141	156,507,016
3	224	274	1	157,451,011	190,015,502
3	275	292	2	191,130,782	198,707,094

5.3.6 Inbred Pedigrees Data to Compare Two Methods for IBD+HBD Estimation

Next, we compared the performance of our two methods for jointly estimating IBD+HBD in pedigree 1. The plot of IBS+HBS configurations is same as Figure 5-10. The inferred IBD + HBD by our SNP streak method with window size 11 SNPs is shown in Table 5-24. Inferred IBD + HBD by our HMM model is shown in Table 5-25. It appears that the SNP streaks method missed a small IBD region (SNP index 1-11) at the beginning of chromosome 3, and broke large regions (SNP indices 71~88 and 89 ~ 130 into smaller regions, due to the arbitrary window size and ignorance of allele frequency.

Table 5-24. Inferred IBD + HBD by SNP streak method

chr	start_snp_index	end_snp_index	IBD_state	start_position	end_position
3	1	33	2	169,613	23,848,605
3	34	49	1	24,449,356	35,050,834
3	50	72	2	35,059,970	50,962,234
3	73	76	3	51,556,549	53,884,770
3	77	87	6	55,246,275	63,417,661
3	88	107	5	63,538,921	74,336,367
3	108	123	2	75,317,481	86,643,308
3	124	188	5	86,929,468	132,970,355
3	189	222	2	133,018,071	156,310,480
3	223	274	1	156,507,016	190,015,502
3	275	292	2	191,130,782	198,707,094

Table 5-25. Inferred IBD + HBD by our HMM method

chr	start_snp_index	end_snp_index	IBD_state	start_position	end_position
3	1	11	1	169,613	4,084,699
3	12	33	2	4,568,187	23,848,605
3	34	47	1	24,449,356	32,530,183
3	48	70	2	33,880,569	49,738,341
3	71	88	3	50,089,518	63,538,921
3	89	130	2	64,402,141	90,472,437
3	131	187	5	95,087,815	132,129,169
3	188	223	2	132,970,355	156,507,016
3	224	274	1	157,451,011	190,015,502
3	275	292	2	191,130,782	198,707,094

5.3.7 Calculation of IBD/HBD Sharing Statistics

After estimation of IBD/HBD configurations, we would like a statistic to measure how well the IBD/HBD sharing pattern in the family fits the genetic model. For example, under a simple

Mendelian model, we would expect all affected individuals to share both IBD and HBD, while no unaffected individual should fully share IBD with the affecteds. If we do not expect a simple Mendelian model, we might prefer a statistic that gives the highest score to the sharing configuration described above, but some intermediate score to a configuration in which, say, one unaffected member of the family also shares IBD with the affecteds. We consider both parametric and non-parametric IBD/HBD sharing statistics.

We first propose a non-parametric score function, *S*, for IBD/HBD configurations, in the manner of non-parametric linkage analysis. Let ϕ represent the IBD/HBD configuration of the target individuals, which in our example is the 2 affected siblings. A "perfect" score function for a Mendelian mode of inheritance is: $S(\phi) = 1$ if ϕ is 11 11; otherwise $S(\phi) = 0$. It assumes complete penetance and no phenocopies. Penetrance is defined as the proportion of individuals who carry a disease gene but that develop an observable disease trait. A phenocopy is defined as an individual who does not carry a disease gene but nonetheless displays a disease trait. A "forgiving" score function is a multi-class scoring rule that allows for incomplete penetrance and/or phenocopies. For example, we can score IBD/HBD configuration 11 12 as 1/5 (an arbitrary score, or based on the penetrance rate), instead of 0. So, for any observed IBD/HBD configuration at a specific locus, the non-parametric score statistic $T = S(\phi)$ actually measures the extent of IBD/HBD across the affected individuals compared to unaffected ones. In other words, we incorporate the affectedness information into the score statistic.

We next propose a parametric alternative: a scoring function based on the logarithm of likelihood ratio. Let L_1 be the maximum value of the likelihood of the data. Let L_0 be the

maximum value of the likelihood of the data under the null hypothesis of no linkage of IBD/HBD with the disease, which is just the Mendelian probability of the IBD/HBD configuration. We can form the log-likelihood ratio statistic ln (L_1 / L_0). Let ϕ represent the IBD/HBD configuration of the target individuals - in our example the 2 affected siblings; $\mathcal{R}$ is their relationship in the pedigree; ω denotes the phenotypes of the 2 siblings; f is the penetrance of each genotype. For simplicity, we use an outbred family here as an example. If we assume the parents' IBD configurations are 12 34, then $L_0 = P(\phi_j | \mathcal{R})$ for each IBD configuration class j (j=1, 2, 3) are listed in Table 5-26. For L_1 we have:

$$L_1 = P(\phi_j | \omega, f, \mathcal{R}) = P(\omega | \phi_j, f, \mathcal{R}) P(\phi_j | \mathcal{R}) / P(\omega | f)$$

$$P(\omega|f) = P(\phi_j) P(\omega | \phi_j, f)$$

$$= 1/4 \sum_{j=1}^{4} P[\omega|\phi_{(j=1)}, f] + 1/2 P[\omega|\phi_{(j=2)}, f] + 1/4 P[\omega|\phi_{(j=3)}, f]$$

Let D represent the disease allele; d is the non-disease allele; q is the disease allele frequency; f is the penetrance rate, which refers to a rate of occurrence of a disease among individuals whose genotypes are rare homozygosity. For example, if f is100%, all individuals with DD in a recessive disorder will be affecteds. For simplicity, we assume penetrance of the common homozygote is zero, then $P[\omega | \phi_{(j=1)}, f] = P(\omega | \phi_{(j=1)}$, sib1 carries DD) P(sib1 carries DD) + $P(\omega | \phi_{(j=1)}$, sib1 carries Dd) P (sib1 carries Dd) + $P(\omega | \phi_{(j=1)}$, sib1 carries dd) P (sib1 carries dd). We can easily derive that in our example: $P(\omega | \phi_{(j=1)}$, sib1 carries DD) P (sib1 carries DD) = f^2q^2 ; $P(\omega | \phi_{(j=1)}$, sib1 carries Dd) = $P(\omega | \phi_{(j=1)}$, sib1 carries dd) =0. Therefore, $P[\omega | \phi_{(j=1)}, f] = f^2q^2$. Similarly, we can get the conditional probabilities of affection states for other IBD configurations, and finally get the log likelihood ratio statistic ψ. It is obvious that ψ is

a function of disease allele frequency *q* and penetrance *f*. Since in real world, we do not know the true values for parameters *q* and *f*, we prefer a non-parametric method to a parametric one.

Table 5-26. $P(\phi_j \mid \mathcal{R})$ for each IBD configuration class j

IBD configuration			$P(\phi_j \mid \mathcal{R})$
class indicator (j)	Sib1	Sib2	
1	**13**	**13**	1/4
2	**13**	**14**	1/2
3	**13**	**24**	1/4

5.3.8 Calculate a P-Value for the Statistic

If the number of families were large, we could use the nonparametric statistics to test the null hypothesis that the locus has no linkage to the disease. This can be done easily if all pedigrees are identical, and otherwise must be approximated. We first standardize the score statistics T_i in each family i (i= 1, 2, …, N) by equation (1).

$$T_i' = [T_i - E(T_i)] / SE \qquad (1)$$

E (T_i) is the expected value of T_i under the null hypothesis, and SE is the standard deviation of T_i, they can be calculated as followed. $P_i(\phi_j \mid \mathcal{R})$ is the null hypothesis probability of IBD/HBD configuration j in family i.

$$E(T_i) = \sum_{j=1}^{J} S_i(\phi_j) P_i(\phi_j \mid R)$$

$$SE = \text{Var}(T_i) / N$$

$$\text{Var}(T_i) = \text{E}(T_i^2) - [\text{E}(T_i)]^2$$

$$\text{E}(T_i^2) = \sum_{j=1}^{J} [S_i(\phi_j)]^2 P_i(\phi_j \mid R)$$

$$\text{Let } \overline{T}\,c = \sum_{i=1}^{N} T_i' / N$$

If all pedigrees are identical, including the same relationship between the parents, then the T_i's are iid with mean=0 and standard deviation=1. Then if the sample size is large, according to the central limit theorem, $\overline{T}\,c \sim N(0, 1)$. We can easily get the corresponding p-value. When sample size is small, since $\overline{T}\,c$ does not follow a normal distribution, we can calculate an exact p-value based on the exact distribution of the statistics.

5.4 IDEALIZED EXTENSION TO LARGE FAMILIES

The principle of extending our methods to a large complex family is that we want to calculate HBD and IBD simultaneously for all family members, including both affected and unaffected, to maximize statistical information. In theory we can extend the IBD/HBD estimation methods implemented above to any large family, although the computation will become complex. We first illustrate this in larger sibships, using a single nuclear family with 3 children (Figure 5-11) as an example, and then in extended families. Theoretically, estimating IBD/HBD in multiple individuals simultaneously is more powerful than pairwise estimation.

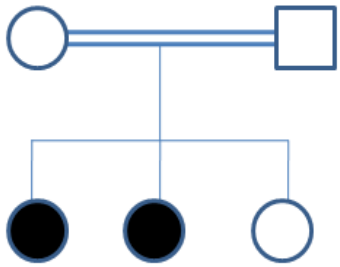


Figure 5-11. Single nuclear family with unaffected related parents (relationship unspecified) and mix of two affected and one unaffected children.

5.4.1 Methods for Estimation of IBD+HBD in Three Siblings Simultaneously

The principle of the SNP streak method in multiple individuals is same as before, but the number of IBS+HBS configurations and IBD+HBD configurations increases. Table 5-27 lists all IBS+HBS states and their corresponding configurations. Table 5-28 summarizes the IBD+HBD states and the corresponding IBS+HBS configurations and states. Two IBD+HBD states “h” and “L” correspond to same set of IBS+HBS states - “6” and “3”, however the probability of being state “h” or “L” given the IBS+HBS states (here is “6” and “3”) is different. In another word, the emission probability is different. Since the SNP streak does not take into account of allele frequencies, it cannot distinguish these, so we prefer not to use this method. Instead we will use HMM for IBD+HBD detection in three children simultaneously.

Again, the HMM method for IBD+HBD is composed of five core elements. The observed IBS+HBS states are listed in Table 5-27, and the hidden IBD+HBD states are summarized in Table 5-28. We used an initial probability that was equal for all IBD+HBD configurations. The emission probabilities we used are listed in appendix II. The principle of

calculation of transition probabilities of hidden states here is same as for a sibling pair. The transition process is actually not a Markov process any more, as we described previously. Again, for simplicity of calculation, we assume the transition of IBD in parents and recombination in meiosis does not happen at the same marker. The model also assumes LD between the markers.

Table 5-27. IBS+HBS states and their corresponding configurations

IBS + HBS states	Observed IBS + HBS configurations		
6	AA	AA	AA
5	AA	AA	AB
4	AA	AA	BB
3	AA	AB	AB
2	AA	AB	BB
1	AB	AB	AB

Table 5-28. IBD+HBD states and the corresponding IBS+HBS configurations and states

Hidden IBD+HBD states	IBD+HBD configurations	All possible IBS+HBS states
a	13 13 13	6,1
b	13 13 14	6,1,5,3
c	13 13 24	6,1,5,3,4
d	13 14 23	6,1,5,3,2
e	11 11 11	6
f	11 11 13	6,5
g	11 11 23	6,5,4
h	11 13 13	6,3
i	11 31 12	6,5,3
j	11 13 23	6,5,3,2
k	11 23 23	6,4,3,2
L	12 31 32	6,3

For example, we assume the parents are first cousins. Let d_i denote the physical distance (base pairs) between two adjacent SNPs i and i +1. If we assume the recombination rate is one per Morgan, and assume 100 Mb is approximately equivalent to one Morgan. Let D be a constant that is set as 100 Mb. The transition probabilities are given in Table 5-29.

Table 5-29. Transition probability of HMM for three siblings simultaneously

	a	b	c	d	e	f	g	h	i	j	k	L
a	$1-9/2(1-t_1)\ di/D - 3/2(1-t_0)\ di/D - 3/16t_1$	$9/2(1-t_1)\ di/D + (1-t_0)\ di/D$	0	0	$3/16\ t_1$	0	0	$1/2(1-t_0)\ di/D$	0	0	0	0
b	$3/4(1-t_1)\ di/D + 1/6(1-t_0)\ di/D$	$1-3(1-t_1)\ di/D - 5/6(1-t_0)\ di/D - 3/8\ t_1 - 1/12\ t_0$	$3/4(1-t_1)\ di/D + 1/12\ t_0$	$3/2(1-t_1)\ di/D$	0	$3/16\ t_1$	0	$3/16\ t_1 + 1/12(1-t_0)\ di/D$	0	$1/6(1-t_0)\ di/D$	$1/12(1-t_0)\ di/D$	$1/3(1-t_0)\ di/D$
c	0	$3/2(1-t_1)\ di/D + 3/16\ t_1$	$1-9/2(1-t_1)\ di/D - 9/16\ t_1$	$3(1-t_1)\ di/D$	0	0	$3/16\ t_1$	0	0	0	$3/16\ t_1$	0
d	0	$3/2(1-t_1)\ di/D$	$3/2(1-t_1)\ di/D$	$1-3(1-t_1)\ di/D - 3/8\ t_1$	0	0	0	0	0	$3/16\ t_1$	0	$3/16\ t_1$
e	$1/4\ t_0$	0	0	0	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	$3/2(1-t_0)\ di/D$	0	0	0	0	0	0
f	0	$1/4\ t_0$	0	0	$1/4(1-t_0)\ di/D$	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	$1/4(1-t_0)\ di/D$	$1/2(1-t_0)\ di/D$	$1/2(1-t_0)\ di/D$	0	0	0
g	0	0	$1/4\ t_0$	0	0	$1/2(1-t_0)\ di/D$	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	0	0	$(1-t_0)\ di/D$	0	0
h	$1/4(1-t_0)\ di/D$	$1/4(1-t_0)\ di/D + 1/4\ t_0$	0	0	0	$1/2(1-t_0)\ di/D$	0	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	0	$1/2(1-t_0)\ di/D$	0	0
i	0	$1/2(1-t_0)\ di/D$	0	$1/4\ t_0$	0	$1/2(1-t_0)\ di/D$	0	0	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	$1/2(1-t_0)\ di/D$	0	0
j	0	$1/4(1-t_0)\ di/D$	0	$1/4\ t_0$	0	0	$1/4(1-t_0)\ di/D$	$1/4(1-t_0)\ di/D$	$1/4(1-t_0)\ di/D$	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	$1/4(1-t_0)\ di/D$	$1/4(1-t_0)\ di/D$
k	0	$1/2(1-t_0)\ di/D$	$1/4\ t_0$	0	0	0	0	0	0	$(1-t_0)\ di/D$	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$	0
L	0	$5/4(1-t_0)\ di/D$	0	$1/4\ t_0$	0	0	0	0	0	$1/4(1-t_0)\ di/D$	0	$1-3/2(1-t_0)\ di/D - 1/4\ t_0$

5.4.2 Simulation Study to Explore HMM Method for Detection of IBD+HBD in Three Siblings Simultaneously

We performed a simulation study to explore the accuracy of the HMM method for IBD+HBD estimation in larger sibships with different densities of markers: Illumina 6K linkage panel and Illumina HumanHap 610K. Again, this simulation study is judged qualitatively as described in section 5.3.3. Figure 5-12 shows the plot for 6K data of IBS+HBS for three siblings with inbred parents. We can find many different streaks of IBS+HBS states in this figure, but it is hard to estimate the IBD+HBD states by viewing the plot. The true simulated IBD+HBD states are listed in Table 5-30. Table 5-31 summarizes the inferred IBD+HBD by our HMM model. Comparison of the two tables shows that our HMM model works very well; it reports results that are consistent with the true IBD + HBD states. However, we found that the HMM model cannot work well in high density (Illumina HumanHap 610K) SNP data to predict the IBD+HBD in three siblings simultaneously, when we set $D = 10^{-21}$ as previously used in a pair of siblings. It broke the large piece into many small ones, which may due to the ignorance of LD in our model.

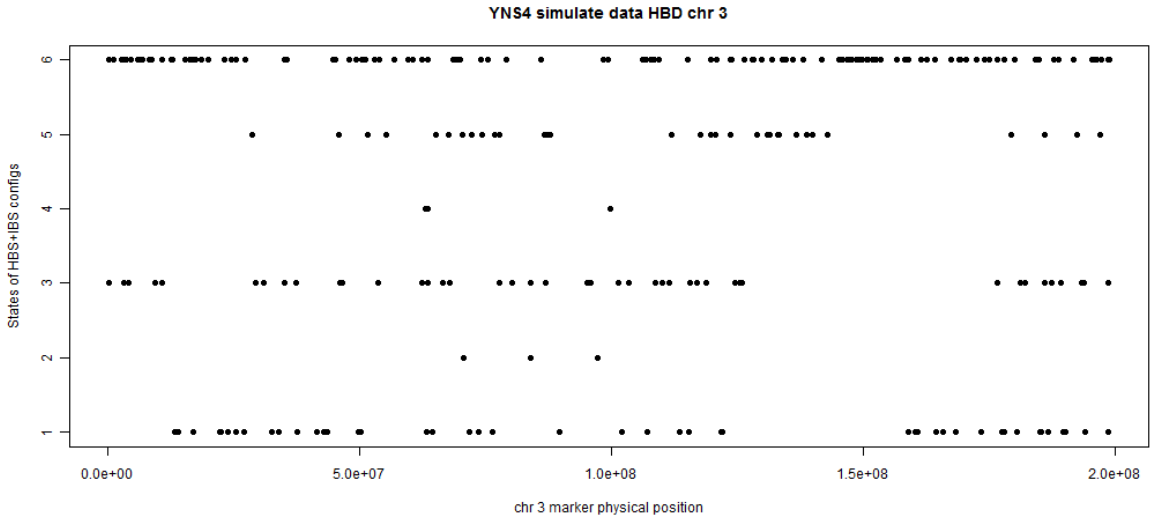


Figure 5-12. IBS+HBS states vs. physical position of 6K linkage panel markers on chromosome 3 for three simulated siblings with inbred parents.

Table 5-30. The true simulated IBD+HBD states

chr	SNP_start_index	SNP_end_index	IBD_state
3	1	20	8
3	21	40	1
3	41	75	2
3	76	90	3
3	91	120	4
3	121	130	3
3	131	160	2
3	161	180	6
3	181	200	5
3	201	220	1
3	221	263	2

Table 5-31. Inferred IBD+HBD by our HMM model in simulated data

chr	SNP_start_index	SNP_end_index	IBD_state	start_position	end_position
3	1	18	8	177,033	10,682,860
3	19	40	1	12,551,845	27,228,974
3	41	75	2	28,649,284	60,440,318
3	76	84	3	62,379,875	64,969,334
3	85	120	4	66,408,991	97,161,899
3	121	127	3	98,368,882	103,343,249
3	128	158	2	106,082,020	125,766,590
3	159	180	6	126,284,920	142,814,254
3	181	200	5	145,136,566	158,772,629
3	201	219	1	158,876,463	175,008,461
3	220	263	2	176,593,117	198,707,094

5.4.3 Inbred Pedigrees Data with 6 K Linkage Panel for Detection of IBD+HBD by HMM in Three Siblings Simultaneously

To further investigate the performance of the HMM method, we randomly picked 3 siblings whose parents are first cousins in pedigree 1, and investigated the performance of our methods. Figure 5-13 is the scatter plot of IBS+HBS states vs. physical position of 6K linkage panel markers on chromosome 3. Table 5-32 summarizes the results from the HMM method. Although it is hard to accurately tell the IBD+HBD states and the break points between them visually from the plot, we can still roughly infer the number of different IBD+HBD segments based on the plot. Comparing the estimation from the HMM method with the plot, we can see consistency on the number of different IBD+HBD states, which implies that our HMM model works for estimation of IBD+HBD in three siblings at the same time in real data.

Our HMM method can be extended to larger families with complicated relationships, but it becomes computationally difficult to list and analyze all possible IBD+HBD configurations and their corresponding IBS+HBS states.

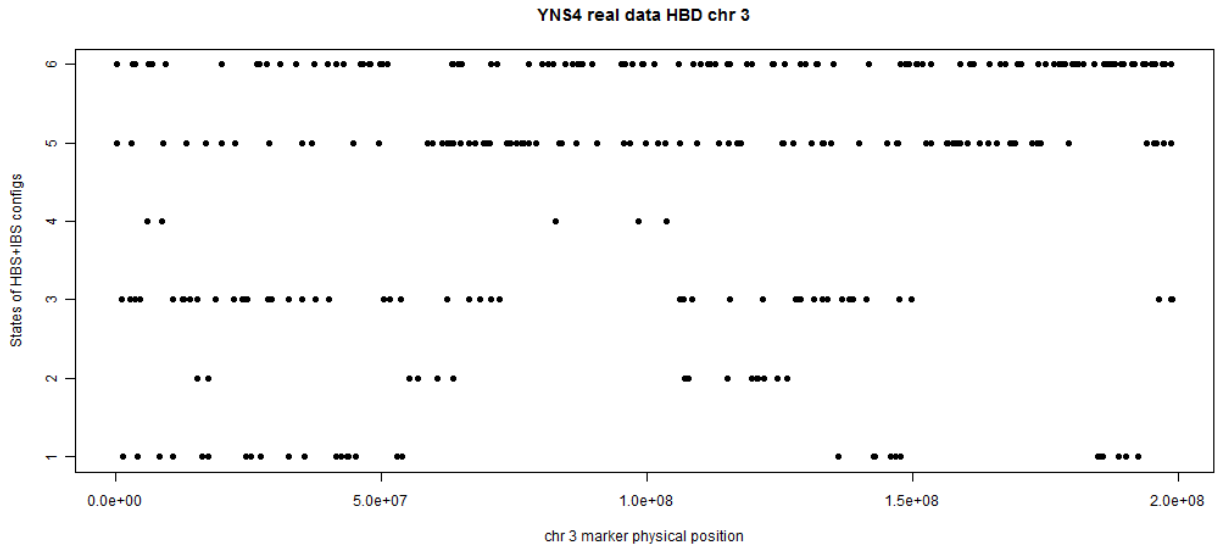


Figure 5-13. IBS+HBS states vs. physical position of 6K linkage panel markers on chromosome 3 for three siblings with inbred parents from real data.

Table 5-32. Inferred IBD+HBD by our HMM model in real data

			IBD+HBD		
chr	start_snp_index	end_snp_index	state	start_position	end_position
3	1	23	3	166,244	10,682,860
3	24	40	4	12,551,845	22,441,845
3	41	88	2	23,848,605	53,884,770
3	89	116	10	55,246,275	72,109,515
3	117	159	7	73,514,520	103,496,947
3	160	196	10	105,934,105	126,294,003
3	197	234	2	127,419,176	149,673,437
3	235	265	6	150,516,770	174,047,370
3	266	280	5	174,955,020	184,078,508
3	281	306	1	184,727,003	193,634,068
3	307	321	2	194,013,895	198,707,094

5.4.4 Calculating an IBD+HBD Sharing Statistic

The methods for creating a statistic (score) for a two-sibling family outlined above can be extended to larger families. Here we show the extension to the three-sibling family with two affected and one unaffected children. We again discuss both non-parametric and parametric scoring functions.

To create a non-parametric score function, let ϕ represent the HBD/IBD configuration of the 2 affected and 1 unaffected siblings. A perfect scoring function $S(\phi) = 1$ if ϕ is 11 11 12, 11 11 22, or 11 11 23; otherwise $S(\phi) = 0$. A forgiving score function will allow incomplete penetrance and/or phenocopies. For example, we can score HBD configuration 11 11 11 as 1/10 (an arbitrary score, or based on penetrance rate), instead of 0. This scoring concept can extend easily to larger families with more than 3 target individuals or to families that combine affected and unaffected individuals.

For a parametric alternative, we again use a score based on likelihood ratio $\ln(L_1 / L_0)$. If we assume the parents' IBD configurations are 12 34, then $L_0 = P(\phi_j \mid \mathcal{R})$ for each IBD configuration class j (j =1, 2, 3, 4) are listed in Table 5-33.

Table 5-33. $P(\phi_j \mid \mathcal{R})$ for each IBD configuration class j

IBD configuration				$P(\phi_j \mid \mathcal{R})$
class indicator (j)	Sib1	Sib2	Sib3	
1	**13**	**13**	**13**	1/16
2	**13**	**13**	**14**	3/8
3	**13**	**13**	**24**	3/16
4	**13**	**14**	**23**	3/8

While $L_1 = P(\phi_j \mid \omega, f, \mathcal{H}) = P(\omega \mid \phi_j, f, \mathcal{H})\, P(\phi_j \mid f, \mathcal{H}) / P(\omega \mid f, \mathcal{H})$

$= P(\omega \mid \phi_j, f)\, P(\phi_j \mid \mathcal{H}) / P(\omega \mid f)$

$P(\omega \mid f) = \sum_{j=1}^{4} P(\phi_j)\, P(\omega \mid \phi_j, f)$

$= 1/16\, P[\omega \mid \phi_{(j=1)}, f] + 3/8\, P[\omega \mid \phi_{(j=2)}, f] + 3/16\, P[\omega \mid \phi_{(j=3)}, f] + 3/8\, P[\omega \mid \phi_{(j=4)}, f]$

Let *D* represent the disease allele; *d* is the non-disease allele; *q* is the disease allele frequency; f is the occurrence rate of a disease among individuals whose genotypes are rare homozygosity. For simplicity, we assume penetrance of the common homozygote is zero, then

$P(\omega \mid \phi_j, f)$ = $P(\omega \mid \phi_j$, sib1 carries DD) P(sib1 carries DD) + $P(\omega \mid \phi_j$, sib1 carries Dd) * P(sib1 carries Dd) + $P(\omega \mid \phi_j$, sib1 carries dd) P(sib1 carries dd)

We can easily derive that in our example

$P(\omega \mid \phi_j$, sib1 carries DD) P(sib1 carries DD) = $f^2(1\text{-}f)\, q^2$

$P(\omega \mid \phi_j$, sib1 carries Dd) = $P(\omega \mid \phi_j$, sib1 carries dd) =0

Therefore, $P[\omega \mid \phi_j$, f] = $f^2(1\text{-f})\, q^2$

Similarly, we can get the conditional affection states for other IBD configurations. We can extend this type of parametric score to more than 3 individuals if we can list all IBD configurations and $P(\phi_j \mid \mathcal{H})$ for each IBD configuration class i under the null hypothesis of no association of the locus with the disease.

As in the previous discussion, we still prefer non-parametric statistics because the disease allele frequency, penetrance rate and phenocopy rate are likely to be unknown.

5.4.5 Calculate a P-Value for The Statistic

Calculation of a p-value for the statistic is similar to what was previously described in section 5.3.8.

5.5 APPLICATION

The methods described above can in theory be extended to an arbitrarily large and complex family, but in practice the computational limitations are significant. In this section we apply the principles outlined above to two large inbred pedigrees that are segregating a Mendelian disorder; this dataset was described in section 5.2.3. Theoretically, we would like to estimate the IBD/HBD overall in each pedigree. However, for efficiency of computation, we take a compromise approach. We first estimate the IBD/HBD in two affected siblings at a time by the HMM method and in non-sib affected pairs by the SNP streak method. For the affected individual who is IBD+HBD with at least one other affected individuals, we check the IBD/HBD between this affected with each of his/her unaffected sibs (two individuals at a time) by HMM. We then combine them to get the overall results.

5.5.1 Pedigree 1

Pedigree 1 is shown in Figure 5-2. This large family contains two related sibships: a first-cousin inbred family and a second-cousin inbred family. We summarize our findings in Table 5-34. We did not find any region in which HBD and IBD are shared across all affected individuals without sharing with any unaffected individuals. However, we did identify a couple of regions with almost perfect sharing patterns. These might be appropriate candidate regions, if we believe that the penetrance is not complete and/or phenocopies exist.

Table 5-34. Summary of IBD+HBD findings in pedigree 1.

Chr	Start position	End position	Shared by
7	39,062,912	51,818,497	all 4 affected, and 2 unaffected individuals
22	38,643,301	45,944,914	3 affected, and 0 unaffected individuals
6	165,642,334	170,734,025	2 affected, and 1 unaffected individuals
2	66,648,337	80,327,668	2 affected, and 1 unaffected individuals

If we use perfect score statistics, the total score for any region in this family is zero. However, if we use forgiving score statistics we can give, for example, a score “1/2” to the IBD+HBD configuration in row one (chr 7) of Table 5-34; a score 1/2 to the configuration in row two (chr22); a score “1/8” to each of the other two configurations (chr6 and chr2). Then chr 7 and chr 22 will be the most interesting regions due to the highest score they get.

Since this is a single pedigree, we will not calculate a p-value; it will not be very informative.

5.5.2 Pedigree 2

Pedigree 2 carries the same recessive disorder as pedigree 1, but comes from a different population. The structure of this family is shown in Figure 5-3. We have genotyping data from seven unaffected ones and two affected living individuals. The relationship between the two affected individuals is complicated, because of several multi-level inbreeding loops. It is thus difficult to apply a HMM model to this pair of individuals. For computational efficiency, we use the SNP streak method to detect IBD+HBD in this non-sibling affected pair. We still use our HMM method to estimate IBD+HBD in each pair of siblings. Finally, we combine the results.

We found three regions that are IBD+HBD in the two affected individuals, but not in the seven unaffected ones. We summarize the findings in Table 5-35.

Table 5-35. Summary of IBD+HBD findings in pedigree 2.

chr	start position	end position
6	107,436,098	123,962,270
7	125,007,188	134,817,061
8	101,541,340	116,719,665

If we use a perfect scoring function as a statistic, we can give score "1" to each of these three regions (chr 6, 7 and 8) in Table 5-35, and score zero to all other regions. It is notable that although the disorder in pedigrees 1 and 2 is same, the origins of the samples are different. We found inconsistent results between pedigrees 1 and 2, which implies that this disorder may be genetically heterogeneous in different populations.

5.6 DISCUSSION

We proposed a rigorous statistical framework for homozygosity mapping in consanguineous family data with high density SNP markers. This procedure contains three steps. For step one - estimation of IBD/HBD - existing methods are not sufficient. They are not optimal for homozygosity mapping in inbred families. Linkage analysis software such as Merlin can work for familial IBD, but it is not sufficient for HBD estimation. We therefore proposed two algorithms, one a SNP streak method and the other an HMM method. The SNP streak algorithm is simple and straightforward; it does not use information on family structures or SNP allele frequencies. However, the selection of a window size is arbitrary, and the method may fail to find some small homozygous regions. Our HMM method employs a number of approximations to the true model, such as assuming no LD and assuming a Markov structure, but despite these approximations it works very well as demonstrated in simulated and real data.

For calculation of an IBD sharing statistic, we suggested a non-parametric scoring statistic (perfect score or forgiving score), which incorporates the affection status into the scores. The forgiving score will allow for incomplete penetrance rate and/or phenocopies. The non-parametric score does not depend on an assumption about disease allele frequency; therefore it is robust to some model misspecification. We also considered a parametric alternative, a likelihood statistic, however it heavily depends on disease allele frequency, penetrance and phenocopy rates, and therefore it is sensitive to selection of model parameters.

Calculation of a p-value for the statistic depends on the sample size of identical families. If the sample size is large, we can use the approximate normal distribution of the mean score to get the p-value. If the sample size is small, we can use an exact test.

All in all, our simulated and real data studies have shown that our procedure is a statistically and computationally efficient method for homozygosity mapping in inbred families. It is difficult to compare results of our methods to those from other methods, because the other methods do such different things, but compared with other methods, there are several strengths of our methods. First, our method is fast. For our real 6K linkage panel data, it took less than 1 minute for our method to infer IBD+HBD genome-wide in a pair of samples. Unlike MERLIN, we incorporated genotyping errors. Also MERLIN requires a known pedigree structure; however our algorithm can accommodate the unknown relationship between the inbred parents by just modeling them as either 1^{st} or 2^{nd} cousins, which does not appear to change the results of inference. But note that we did not test more distant relationships. Our methods are an improvement over HomozygosityMapper in that we will not report runs of homozygosity with unmatched alleles between individuals as HBD regions. Compared with BEAGLE, which does not bear close relationships among the study subjects, our methods can work in inbred families.

Our methods do have several limitations. Our HMM model assumes linkage equilibrium among the markers, which is not true for high density SNP data. In order to make HMM method work in high density SNP data, we have to increase the parameter "D" dramatically to compensate for the model's no-LD assumption. This is not an accurate way to model LD, but it is fast and it works. We failed in estimation of IBD+HBD in three siblings at the same time for high density SNP data with adjusted parameter *D*, but it works fine for linkage panel data. Another limitation of our method is that it becomes complicated if we want to estimate IBD+HBD in many individuals from a large complex family simultaneously; we may need to compromise by estimation of IBD+HBD in a pair or three individuals at a time. In addition, the

transition process is not a real Markov process; we used some approximations. However it works well in both simulated and real data.

In the future, we would like to model LD in our HMM method for inbred family data. We also will do simulation more quantitatively, and draw conclusions about the accuracy of the estimated breakpoints.

5.7 REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30: 97-101.

Brooks AS, Bertoli-Avella AM, Burzynski GM, Breedveld GJ, Osinga J, Boven LG, Hurst JA, Mancini GM, Lequin MH, de Coo RF, Matera I, de Graaff E, Meijers C, Willems PJ, Tibboel D, Oostra BA, Hofstra RM. Homozygous nonsense mutations in KIAA1279 are associated with malformations of the central and enteric nervous systems. Am J Hum Genet. 2005;77:120-26.

Browning SR, Browning BL. High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet. 2010; 86: 526-39.

Cao Q, Zhu Q, Wu ML, Hu WL, Gao M, Si JM. Genetic susceptibility to ulcerative colitis in the Chinese Han ethnic population: association with TNF polymorphisms. Chin Med J (Engl). 2006; 119:1198-203.

Eleanor Feingold. Markov process for modeling and analyzing a new genetic mapping method. J. Appl. Prob. 1993; 30: 766-79.

Forney GD. The Viterbi algorithm. Proceedings of the IEEE. 1973; 61: 268–78.

Garshasbi M, Motazacker MM, Kahrizi K, Behjati F, Abedini SS, Nieh SE, Firouzabadi SG, Becker C, Rüschendorf F, Nürnberg P, Tzschach A, Vazifehmand R, Erdogan F, Ullmann R, Lenzner S, Kuss AW, Ropers HH, Najmabadi H. SNP array-based homozygosity mapping reveals MCPH1 deletion in family with autosomal recessive mental retardation and mild microcephaly. Hum Genet. 2006; 118:708-15.

International HapMap Consortium. The International HapMap Project. Nature. 2003; 18; 426: 789-96.

Kahrizi K, Najmabadi H, Kariminejad R, Jamali P, Malekpour M, Garshasbi M, Ropers HH, Kuss AW, Tzschach A. An autosomal recessive syndrome of severe mental retardation, cataract, coloboma and kyphosis maps to the pericentromeric region of chromosome 4. Eur J Hum Genet. 2009; 17:125-28.

Knight HM, Maclean A, Irfan M, Naeem F, Cass S, Pickard BS, Muir WJ, Blackwood DH, Ayub M. Homozygosity mapping in a family presenting with SCZ, epilepsy and hearing impairment. Eur J Hum Genet. 2008;16:750-8.

Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet. 1997; 61:1179-88.

Kousar R, Hassan MJ, Khan B, Basit S, Mahmood S, Mir A, Ahmad W, Ansar M. Mutations in WDR62 gene in Pakistani families with autosomal recessive primary microcephaly. BMC Neurol. 2011;11:119-25.

Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987; 236: 1567-70.

Marioni JC, Thorne NP, Tavaré S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics. 2006; 22: 1144-6.

Mochida GH, Ganesh VS, Felie JM, Gleason D, Hill RS, Clapham KR, Rakiec D, Tan WH, Akawi N, Al-Saffar M, Partlow JN, Tinschert S, Barkovich AJ, Ali B, Al-Gazali L, Walsh CA. A homozygous mutation in the tight-junction protein JAM3 causes hemorrhagic destruction of the brain, subependymal calcification, and congenital cataracts. Am J Hum Genet. 2010; 87: 882-89.

Nikolai Shokhirev. Hidden Markov model. Last Modified: 02/15/2010; retrieved: 04/16/2011. http://www.shokhirev.com/nikolai/abc/alg/hmm/hmm.html.

Ott J . Analysis of Human Genetic linkage. 1985. Johns Hopkins University Press, Baltimore.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559-75.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria.2009. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989; 77: 257-286.

Rizel L, Safieh C, Shalev SA, Mezer E, Jabaly-Habib H, Ben-Neriah Z, Chervinsky E, Briscoe D, Ben-Yosef T. Novel mutations of MYO7A and USH1G in Israeli Arab families with Usher syndrome type 1. Mol Vis. 2011; 17: 3548-55.

Saar K, Al-Gazali L, Sztriha L, Rueschendorf F, Nur-E-Kamal M, Reis A, Bayoumi R. Homozygosity mapping in families with Joubert syndrome identifies a locus on chromosome 9q34.3 and evidence for genetic heterogeneity. Am J Hum Genet. 1999; 65: 1666-71.

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P. HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Res. 2009; 37: W593-99.

Spiegel R, Shaag A, Mandel H, Reich D, Penyakov M, Hujeirat Y, Saada A, Elpeleg O, Shalev SA. Mutated NDUFS6 is the cause of fatal neonatal lactic acidemia in Caucasus Jews. Eur J Hum Genet. 2009; 17:1200-203.

V. Petrushin. Hidden Markov Models: Fundamentals and Applications. Part 2: Discrete and Continuous Hidden Markov Models, Online Symposium for Electronics Engineers. 2000

Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. Biometrics. 1994; 50: 118-27.

Winick JD, Blundell ML, Galke BL, Salam AA, Leal SM, Karayiorgou M. Homozygosity mapping of the Achromatopsia locus in the Pingelapse. Am J Hum Genet. 1999; 64:1679-85.

6.0 DISCUSSION

6.1 CONCLUSIONS AND CONRIBUTIONS OF THIS WORK

This dissertation is composed of four projects; all of them involve using dense SNP data. The first three projects focused on studies of CNVs; the fourth project was about homozygosity mapping. The core question I addressed in this dissertation is how to look across multiple SNPs to accurately detect some specific DNA regions (CNV or IBD/HBD) in high density SNP arrays, and how to apply those methods to real data analysis of complex diseases, such as CNV studies in mental health and behavioral disorders, and IBD/HBD estimation in rare recessive diseases.

Our first project made recommendations for how CNV calls can be used in genome-wide association studies. A roadblock to comparison of different CNV-calling strategies is the lack of gold standard data to tell us which CNVs are real. We used family data as a verification standard, and proposed that if CNVs are repeatedly called in duplicate samples, or inherited from parent to child, then these can be considered validated CNVs. We used two large family genome wide association study (GWAS) datasets to look at concordance and inheritance rates of CNV calls. This allowed us to draw inferences about the performance of various CNV calling strategies, and the features and distributions of CNVs in the human genome. We found current filtering strategies and the common strategy of using only the largest CNV calls cannot guarantee high

reliability. We concluded from our data that it is probably not possible to find a CNV calling strategy that will give us a set of "reliable" CNV calls using current chip technologies. For now, CNV calls will need to be understood as having high error rates. But if we understand and model the features of that error process, we can still use them appropriately in genetic association studies. In particular, the most critical issue will be to make sure that cases and controls are well matched on any features that we know affect CNV call reliability rates, such as DNA sample type. We also made some contributions to the growing picture of what “normal” variability in copy number means for the human genome.

Our second project was a real data analysis on AD+P. We proposed that different mental health disorders may share some genetic factors and that their expression might be modified by other environmental and genetic factors. Therefore, we specifically searched for CNVs in seven recurrent CNV regions identified in schizophrenia and autism. We are the first to report that AD+P shares a rare risk CNV region on 16p11.2 with schizophrenia and autism. Its frequency in AD+P is similar to that in schizophrenia. We also found a smaller CNV on 3q29 which is within *PAK2,* one of the most interesting candidate genes for schizophrenia in that region. Although rare, these CNVs may have important functions in the development of psychosis. Identification of these CNVs can help with understanding the mechanisms of psychosis disorders. These CNVs have the potential to be used in clinical practice for screening, diagnosis, disease classification or genetic testing. Some CNVs are already in use in pediatric settings. For example, 3q29 is evaluated in children with autism.

Our third project was a study of CNVs, smoking and birth outcomes. We thoroughly screened the genome for CNVs associated with smoking and birth outcomes, and identified several strong candidate genes. Also, the consistent findings in two large-scale GENEVA

datasets make the candidate genes for smoking very interesting. A previously-reported CNV in GSTT1 associated with birth outcomes in smokers was not covered by this chip. However, we did find other CNVs in GSTT1 and GSTT2 respectively, which are associated with birth weight in smokers. We don't know if these CNVs are in LD with the previously-reported one.

Our fourth project developed statistical methods for homozygosity mapping in family data using dense SNP arrays. Statistical methods for homozygosity mapping have traditionally been ad hoc and suboptimal. Our objective was to propose more rigorous statistical approaches to this problem, with the goal of improved gene-finding. We proposed two different mathematical approaches for finding the regions of the genome that are most likely to harbor the genes we are looking for. We also proposed a more rigorous framework for statistical evaluation of those regions. We tested our methods in both simulated and real data. We successfully identified the HBD regions in simulated family data. We found several potentially disease-causing regions in two real pedigrees. Compared with current methods, our methods add statistical rigor, and are a great improvement over simple visual inspection methods that are more commonly used.

From the four studies described above, we found that HMM is a good algorithm to look across multiple markers at a time in high density SNP arrays, whether for CNV calling or homozygosity mapping. However several factors may influence the performance of HMM. One important factor is data quality. For CNV calling, different filtering methods and strategies have been proposed in order to improve the data quality; however, none of the methods is optimal. Another factor that is a barrier to making HMM models work for this kind of data is LD. Most common software for homozygosity mapping does not try to model LD; instead it prunes or clustesr markers, which may reduce the power to detect short regions. BEAGLE does model both

LD between markers and IBD between a pair of unrelated individuals. However, it does not model the relationship of a family in HMM; it assumes that neither affected individuals nor their parents are related. So how to model LD between markers and IBD/HBD in multiple inbred family members simultaneously is be an open question.

6.2 FUTURE WORK AND OPEN QUESTIONS

There are some open questions in each of the projects. In our first project, we found a subset of individuals who carry a fairly high load of rare CNVs (100 or more) that appear from inheritance rates to be real. However, the mechanism is unknown. We also found a modest increase in the number of CNVs with age, suggesting a non-trivial rate of somatic mutation, although this clearly bears further study. Finally, we found some intriguing results related to the relative inheritance rates of deletions vs. amplifications, which would be interesting to follow up further.

The CNV findings in project two were all identified by statistical methods. We are conducting molecular experiments to validate these CNVs, especially in 3q29. If validated, this will be a major finding, since a small duplication CNV in 3q29 has not been reported in literature.

In project three, we are not sure whether the association of CNVs with smoking is real. The association could due to an artifact of our selection on successful birth outcomes, since smokers who have fetal deaths were not included in this study. Replicate studies using independent data are warranted.

In project four, the direction of future work is to develop methods that can model LD and IBD/HBD simultaneously in family data using dense SNP array.

APPENDIX A

EMISSION PROBABILITY IN HIDDEN MARKOV MODEL FOR IBD+HBD IN A PAIR OF CHILDREN

Table A-1 lists the probabilities of configurations of HBS+IBS (emission probabilities) conditional on the configurations of HBD+IBD for a pair of siblings.

Table A-1. Probability of configuration of HBS+IBS conditional on the configuration of HBD+IBD for a pair of siblings

index of HBD +IBD configuration	index of HBS + IBS configuration			
	1	2	3	4
a	$2pq\ (1-\varepsilon)$	0	ε	$(p^2+q^2)\ (1-\varepsilon)$
b	$(p^2q+pq^2)\ (1-\varepsilon)$ $+2/3(p^2q+pq^2)\ \varepsilon$	$2/3(p^2q+pq^2)\ \varepsilon$	$(2p^2q+2pq^2)\ (1-\varepsilon) + (p^3+q^3)\ \varepsilon$ $+ (p^2q+pq^2)\ \varepsilon$	$(p^3+q^3)\ (1-\varepsilon) +2/3$ $(p^2q+pq^2)\varepsilon$
c	$(2p^2q^2)\ (1-\varepsilon)+$ $4/3(p^3q+pq^3)\varepsilon$	$(4p^2q^2)\ (1-\varepsilon)$ $+4/3(p^3q+pq^3)\varepsilon$	$(4p^3q+4pq^3)\ (1-\varepsilon) + (p^4+q^4)\ \varepsilon$ $+6p^2q^2\varepsilon$	$(p^4+q^4)\ (1-\varepsilon) +$ $4/3(p^3q+pq^3)\varepsilon$
d	0	0	ε	$1-\varepsilon$
e	$2/3pq\varepsilon$	$2/3pq\varepsilon$	$2\ pq\ (1-\varepsilon) +(p^2+q^2)\varepsilon$	$(p^2+q^2)(1-\varepsilon)+$ $2/3pq\varepsilon$
f	$2/3(p^2q+pq^2)\ \varepsilon$	$(p^2q+pq^2)\ (1-\varepsilon)$ $+2/3(p^2q+pq^2)\ \varepsilon$	$(2p^2q+2pq^2)\ (1-\varepsilon) +(p^3+q^3)$ $\varepsilon+(p^2q+pq^2)\ \varepsilon$	$(p^3+q^3)\ (1-\varepsilon)$ $+2/3(p^2q+pq^2)\ \varepsilon$

APPENDIX B

EMISSION PROBABILITY IN HMM FOR DETECTION OF HBD IN 3 CHILDREN SIMULTANEOUSLY

Table B-1 lists the probabilities of configurations of HBS+IBS (emission probabilities) conditional on the configurations of HBD+IBD for three children simultaneously.

Table B-1. Probability of configuration of HBS+IBS conditional on the configuration of HBD+IBD for three children simultaneously

index of HBD +IBD configuration	index of HBS + IBS configuration					
	1	2	3	4	5	6
a	$2pq\,(1-\varepsilon)$	0	$2pq\varepsilon$	0	$(p^2+q^2)\,\varepsilon$	$(p^2+q^2)\,(1-\varepsilon)$
b	$(p^2q+pq^2)\,(1-\varepsilon)+1/3(p^2q+pq^2)\varepsilon$	$1/3(\,p^2q+pq^2)\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+4/3(p^2q+pq^2)\varepsilon$	$1/3\,(p^2q+pq^2)\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+(p^3+q^3)\varepsilon+1/3(p^2q+pq^2)\varepsilon$	$(p^3+q^3)\,(1-\varepsilon)+1/3\,(p^2q+pq^2)\varepsilon$
c	$(4p^2q^2)\,(1-\varepsilon)+2/3\,(\,p^2q+pq^2)\varepsilon$	$2/3(\,p^2q+pq^2)\varepsilon+p^2q^2\varepsilon$	$2(p^2q+pq^2)\,(1-\varepsilon)\;+\;2/3\,(\,p^2q+pq^2)\varepsilon+4p^2q^2\varepsilon$	$2\,p^2q^2\,(1-\varepsilon)\;+\;2/3(\,p^2q+pq^2)\varepsilon$	$2(p^3q+pq^3)\,(1-\varepsilon)+(p^4+q^4)\,\varepsilon+2/3(p^3q+pq^3)\,\varepsilon+p^2q^2\varepsilon$	$(p^4+q^4)\,(1-\varepsilon)+2/3(p^3q+pq^3)\varepsilon$
d	$2\,p^2q^2\,(1-\varepsilon)\;+\;2/3(p^3q+pq^3+\,p^2q^2)\varepsilon$	$2\,p^2q^2\,(1-\varepsilon)\;+\;2/3(p^3q+pq^3+\,p^2q^2)\varepsilon$	$2/3(p^3q+pq^3+\,p^2q^2)\,(1-\varepsilon)+2/3(p^3q+pq^3)\,\varepsilon+2\,p^2q^2\varepsilon$	$2/3(p^3q+pq^3)\varepsilon+p^2q^2\varepsilon$	$2(p^3q+pq^3\,(1-\varepsilon)+(p^4+q^4)\varepsilon+2/3(p^3q+pq^3+\,p^2q^2)\varepsilon$	$(p^4+q^4)\,(1-\varepsilon)+2/3(p^3q+pq^3)\varepsilon$
e	0	0	0	0	ε	$(1-\varepsilon)$
f	0	0	$2/3pq\varepsilon$	$2/3pq\varepsilon$	$2\,pq\,(1-\varepsilon)+(p^2+q^2)\,\varepsilon$	$(p^2+q^2)(1-\varepsilon)+2/3pq\varepsilon$
g	0	0	$2/3(\,p^2q+pq^2)\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+2/3(\,p^2q+pq^2)\varepsilon$	$2(\,p^2q+pq^2)\,(1-\varepsilon)+(p^3+q^3)\,\varepsilon+(\,p^2q+pq^2)\varepsilon$	$(p^3+q^3)\,(1-\varepsilon)+2/3(\,p^2q+pq^2)\varepsilon$
h	$2/3pq\varepsilon$	$2/3pq\varepsilon$	$2pq\,(1-\varepsilon)$	0	$(p^2+q^2+\,2/3pq)\varepsilon$	$(p^2+q^2)\,(1-\varepsilon)$
i	$1/3(p^2q+pq^2)\,\varepsilon$	$1/3(p^2q+pq^2)\,\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+2/3\,(p^2q+pq^2)\varepsilon$	$2/3(p^2q+pq^2)\,\varepsilon$	$2(p^2q+pq^2)\,(1-\varepsilon\,)+(p^3+q^3)\,\varepsilon+1/3(p^2q+pq^2)\,\varepsilon$	$(p^3+q^3)\,(1-\varepsilon)+2/3(p^2q+pq^2)\,\varepsilon$
j	$1/3\,pq\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+1/3\,pq\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+5/6\,pq\varepsilon$	$5/6\,pq\varepsilon$	$(p^2q+pq^2)\,(1-\varepsilon)+(p^3+q^3)\varepsilon+1/3\,pq\varepsilon$	$(p^3+q^3)\,(1-\varepsilon)+1/3\,pq\varepsilon$
k	$2/3\,(p^2q+pq^2)\varepsilon$	$7/6\,(p^2q+pq^2)\varepsilon$	$2(p^2q+pq^2)\,(1-\varepsilon)$	$(p^2q+pq^2)\,(1-\varepsilon)$	$(p^3+q^3)\varepsilon+7/6\,(p^2q+pq^2)\varepsilon$	$(p^3+q^3)\,(1-\varepsilon)$
l	$(p^2q+pq^2)\varepsilon$	$(p^2q+pq^2)\varepsilon$	$3(p^2q+pq^2)\,(1-\varepsilon)$	0	$(p^3+q^3)\varepsilon+(p^2q+pq^2)\,\varepsilon$	$(p^3+q^3)\,(1-\varepsilon)$

APPENDIX C

PROCEDURES FOR SNP STREAK METHOD

1) Calculate the IBS+HBS states among the assigned family members at each SNP on a chromosome and generate a sequence of IBS+HBS states;

2) Select the window size, the sliding size and the genotyping error rate;

3) Check whether any singular IBS+HBS state in a given window is a genotyping error. For example, we can test 25 SNPs before and after the marker with singular IBS+HBS state. If the IBS+HBS state is still singular in those 50 SNPs, it is considered as a genotyping error.

4) Generate a null sequence of IBD+HBD states with length = all makers in a chromosome.

5) Initial window: estimate the IBD+HBD state according to the IBS+HBS states in the initial window; fill the null IBD+HBD sequence with this estimated IBD+HBD state.

6) Sliding windows: slide the window to find the break points - the last SNP in the first window with newly appeared IBS+HBS states; or the first SNP in the first window with newly disappeared IBS+HBS states. Re-estimate the IBD+HBD states based on the IBS+HBS state in the window starting from the break point, and replace the IBD+HBD states for all markers on and after the break point with the re-estimated one.

BIBLIOGRAPHY

Aagaard-Tillery K, Spong CY, Thom E, Sibai B, Wendel G Jr, Wenstrom K, Samuels P, Simhan H, Sorokin Y, Miodovnik M, Meis P, O'Sullivan MJ, Conway D, Wapner RJ; Eunice Kennedy Shriver National Institute of Child Health, Human Development (NICHD) Maternal-Fetal Medicine Units Network (MFMU). Pharmacogenomics of maternal tobacco use: metabolic gene polymorphisms and risk of adverse pregnancy outcomes. Obstet Gynecol. 2010; 115: 568-77.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30: 97-101.

Arifeen SE, Black RE, Caulfield LE, Antelman G, Baqui AH, Nahar Q, Alamgir S, Mahmud H. Infant growth patterns in the slums of Dhaka in relation to birth weight, intrauterine growth retardation, and prematurity. Am J Clin Nutr. 2000; 72: 1010-7.

Asmussen I, Kjeldsen K. Intimal ultrastructure of human umbilical arteries. Observations on arteries from newborn children of smoking and nonsmoking mothers. Circ Res 1975; 36: 579-89.

Asmussen I. Ultrastructure of the human placenta at term. Observations on placentas from newborn children of smoking and non-smoking mothers. Acta Obstet Gynecol Scand 1977; 56: 119-26.

Bacanu SA, Devlin B, Chowdari KV, DeKosky ST, Nimgaonkar VL, Sweet RA. Heritability of psychosis in Alzheimer disease. Am J Geriatr Psychiatry. 2005; 13: 624–627.

Ballard CG, O'Brien JT, Coope B, Wilcock G. Psychotic symptoms in dementia and the rate of cognitive decline.J Am Geriatr Soc 1997; 45: 1031–1032.

Bedoyan JK, Kumar RA, Sudi J, Silverstein F, Ackley T, Iyer RK, Christian SL, Martin DM. Duplication 16p11.2 in a child with infantile seizure disorder. Am J Med Genet A. 2010; 152A: 1567-74.

Bergen A, Engedel K and Kringlen A. The Role of Heredity in Late Onset Alzheimer's disease and Vascular Dementia Archives of General Psychiatry 1997; 54: 264-270.

Brooks AS, Bertoli-Avella AM, Burzynski GM, Breedveld GJ, Osinga J, Boven LG, Hurst JA, Mancini GM, Lequin MH, de Coo RF, Matera I, de Graaff E, Meijers C, Willems PJ,

Tibboel D, Oostra BA, Hofstra RM. Homozygous nonsense mutations in KIAA1279 are associated with malformations of the central and enteric nervous systems. Am J Hum Genet. 2005;77:120-26.

Browning SR, Browning BL. High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet. 2010; 86: 526-39.

Burns A, Jacoby R, Levy R. Psychiatric phenomena in Alzheimer's disease. Br J Psychiatry1990; 157: 72-94.

Cahan P, Li Y, Izumi M et al. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nat Genet. 2009; 41: 430–437.

Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, Thomas-Anterion C, Michon A, Martin C, Charbonnier F, Raux G, Camuzat A, Penet C, Mesnage V, Martinez M, Clerget-Darpoux F, Brice A, Frebourg T. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. Am J Hum Genet. 1999; 65: 664-70.

Cao Q, Zhu Q, Wu ML, Hu WL, Gao M, Si JM. Genetic susceptibility to ulcerative colitis in the Chinese Han ethnic population: association with TNF polymorphisms. Chin Med J (Engl). 2006; 119:1198-203.

Carson R, Craig D, Hart D, Todd S, McGuinness B, Johnston JA, O'Neill FA, Ritchie CW, Passmore AP. Genetic variation in the alpha 7 nicotinic acetylcholine receptor is associated with delusional symptoms in Alzheimer's disease. Neuromolecular Med. 2008;10:377–384.

Carter NP. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39: S16–S21.

Chan A, Keane RJ, Robinson JS. The contribution of maternal smoking to preterm birth, small for gestational age and low birthweight among Aboriginal and non-Aboriginal births in South Australia. Med J Aust. 2001;174:389-93.

Coggan M, Whitbread L, Whittington A, Board P. Structure and organization of the human theta-class glutathione S-transferase and D-dopachrome tautomerase gene complex. Biochem J. 1998; 334 (Pt 3):617-23.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007; 35: 2013-25.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993; 261: 921-3.

de Onis M, Blössner M, Villar J. Levels and patterns of intrauterine growth retardation in developing countries. Eur J Clin Nutr. 1998; 52 Suppl 1: S5-15.

Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 38: e105.

DeMichele-Sweet MA, Sweet RA. Genetics of psychosis in Alzheimer's disease: a review. J Alzheimers Dis. 2010; 19: 761-80.

De Strooper B, Saftig P, Craessaerts K, Vanderstichele H, Guhde G, Annaert W, Von Figura K, Van Leuven F. Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. Nature. 1998; 391: 387–390.

Feingold E. Markov process for modeling and analyzing a new genetic mapping method. J. Appl. Prob. 1993; 30: 766-79.

Farber NB, Rubin EH, Newhouse PA, Kinscherf DA, Miller JP, Morris JC *et al*. Increased neocortical neurofibrillary tangle density in subjects with Alzheimer's disease. Arch Gen Psychiatry. 2000; 57: 1165–1173.

Forney GD. The Viterbi algorithm. Proceedings of the IEEE. 1973; 61: 268–78.

Forstl H, Burns A, Levy R, Cairns N. Neuropathological correlates of psychotic phenomena in confirmed Alzheimer's disease. Br J Psychiatry. 1994; 165: 53–59.

Garshasbi M, Motazacker MM, Kahrizi K, Behjati F, Abedini SS, Nieh SE, Firouzabadi SG, Becker C, Rüschendorf F, Nürnberg P, Tzschach A, Vazifehmand R, Erdogan F, Ullmann R, Lenzner S, Kuss AW, Ropers HH, Najmabadi H. SNP array-based homozygosity mapping reveals MCPH1 deletion in family with autosomal recessive mental retardation and mild microcephaly. Hum Genet. 2006; 118:708-15.

Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006; 63:168-74.

Ghebranious N, Giampietro PF, Wesbrook FP, Rezkalla SH. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. Am J Med Genet A. 2007; 143: 1462–71.

Goldenberg RL and Culhane JF. Low birth weight in the United States. *Am J Clin Nutr* 2007; 85: 584S-590S.

Go RC, Perry RT, Wiener H, Bassett SS, Blacker D, Devlin B, Sweet RA. Neuregulin-1 polymorphism in late onset Alzheimer's disease families with psychoses. Am J Med Genet B Neuropsychiatr Genet. 2005; 139B: 28-32.

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature. 1991; 349: 704–706.

Goedert M, Spillantini MG. A century of Alzheimer's disease. Science. 2006; 314: 777-81.

Grazuleviciene R, Danileviciute A, Nadisauskiene R, Vencloviene J. Maternal smoking, GSTM1 and GSTT1 polymorphism and susceptibility to adverse pregnancy outcomes. Int J Environ Res Public Health. 2009; 6: 1282-97.

Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, Drouin-Garraud V, Layet V, Rosier A, Briault S, Bonnet-Brilhault F, Laumonnier F, Odent S, Le Vacon G, Joly-Helas G, David V, Bendavid C, Pinoit JM, Henry C, Impallomeni C, Germano E, Tortorella G, Di Rosa G, Barthelemy C, Andres C, Faivre L, Frébourg T, Saugier Veber P, Campion D. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. Arch Gen Psychiatry. 2009 Sep;66(9):947-56.

Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA. Alzheimer Disease in the US Population. Prevalence Estimates Using the 2000 Census. Arch Neurol. 2003; 60: 1119-1122.

Heinzen EL, Need AC, Hayden KM, Chiba-Falek O, Roses AD, Strittmatter WJ, Burke JR, Hulette CM, Welsh-Bohmer KA, Goldstein DB. Genome-wide scan of copy number variation in late-onset Alzheimer's disease. J Alzheimers Dis. 2010; 19: 69-77.

Henrichsen CN, Vinckenbosch N, Zollner S et al. Segmental copy number variation shapes tissue transcriptomes. Nat Genet. 2009; 41: 424–429.

Horesh Y, Katsel P, Haroutunian V, Domany E. Gene expression signature is shared by patients with Alzheimer's disease and schizophrenia at the superior temporal gyrus. Eur J Neurol. 2011; 18: 410-24.

Horta BL, Victora CG, Menezes AM, Halpern R, Barros FC. Low birthweight, preterm births and intrauterine growth retardation in relation to maternal smoking. Paediatr Perinat Epidemiol. 1997;11:140-51.

International HapMap Consortium. The International HapMap Project. Nature. 2003; 18; 426: 789-96.

Jeste DV, Wragg RE, Salmon DP, Harris MJ, Thal LJ. Cognitive deficits of patients with Alzheimer's disease with and without delusions. Am J Psychiatry 1992; 149: 184–189.

Kahrizi K, Najmabadi H, Kariminejad R, Jamali P, Malekpour M, Garshasbi M, Ropers HH, Kuss AW, Tzschach A. An autosomal recessive syndrome of severe mental retardation, cataract, coloboma and kyphosis maps to the pericentromeric region of chromosome 4. Eur J Hum Genet. 2009; 17:125-28.

Kauwe J and Goate A. Molecular Genetics in Alzheimer's Disease. Neurobiology of Alzheimer's disease Eds. Dawbern, D. and Allen, S. Oxford University Press 2007; 59-80.

Khachaturian AS, Corcoran CD, Mayer LS, Zandi PP, Breitner JC. Apolipoprotein E epsilon4 count affects age at onset of Alzheimer disease, but not lifetime susceptibility: The Cache County Study. Arch Gen Psychiatry. 2004; 61: 518–24.

Kjell Haram K, Svendsen E, Myking O. Growth Restriction: Etiology, Maternal and Neonatal Outcome. A Review. Current Women's Health Reviews, 2007; 3: 145-160.

Knight HM, Maclean A, Irfan M, Naeem F, Cass S, Pickard BS, Muir WJ, Blackwood DH, Ayub M. Homozygosity mapping in a family presenting with SCZ, epilepsy and hearing impairment. Eur J Hum Genet. 2008;16:750-8.

Kohler JR, Cutler DJ. Simultaneous Discovery and Testing of Deletions for Disease Association in SNP Genotyping Studies. Am J Hum Genet. 2007; 81:684-99.

Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet. 1997; 61:1179-88.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008; 40: 1253-60.

Kousar R, Hassan MJ, Khan B, Basit S, Mahmood S, Mir A, Ahmad W, Ansar M. Mutations in WDR62 gene in Pakistani families with autosomal recessive primary microcephaly. BMC Neurol. 2011;11:119-25.

Kramer MS. Intrauterine growth and gestation determinants. Pediatrics 1987; 80: 502−511.

LaFramboise T, Winckler W, Thomas RK. A flexible rank-based framework for detecting copy number aberrations from array data. Bioinformatics. 2009; 25: 722-728.

Lai WR, Johnson MD, Kucherlapati R, Park PJ. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21: 3763-3770.

Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987; 236: 1567-70.

Lopez OL, Kamboh MI, Becker JT, Kaufer DI, DeKosky ST. The apolipoprotein E e4 allele is not associated with psychiatric symptoms or extrapyramidal signs in probable Alzheimer's disease. Neurology 1997; 49: 794–797.

Lyketsos CG, Steinberg M, Tschanz JT, Norton MC, Steffens DC, Breitner JCS. Mental and behavioral disturbances in dementia: findings from the Cache County study on memory in aging. Am J Psychiatry 2000; 157: 708–714.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181-1186.

Marioni JC, Thorne NP, Tavaré S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics. 2006; 22: 1144-6.

Martin GM, Ogburn CE, Colgin LM, Gown AM, Edland SD, Monnat RJ Jr. 1996. Somatic Mutations Are Frequent and Increase with Age in Human Kidney Epithelial Cells. Hum Mol Genet 5: 215- 221.

Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Kirmeyer S, Munson ML; Centers for Disease Control and Prevention National Center for Health Statistics National Vital Statistics System. Births: final data for 2005. Natl Vital Stat Rep. 2007; 56: 1-103.

Maslov AY, Vijg J. 2009. Genome instability, cancer and aging. Biochim Biophys Acta 1790: 963-969.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, Grozeva D, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Roccanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimäki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nöthen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR; Wellcome Trust Case Control Consortium, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009; 41: 1223-7.

Mochida GH, Ganesh VS, Felie JM, Gleason D, Hill RS, Clapham KR, Rakiec D, Tan WH, Akawi N, Al-Saffar M, Partlow JN, Tinschert S, Barkovich AJ, Ali B, Al-Gazali L, Walsh CA. A homozygous mutation in the tight-junction protein JAM3 causes hemorrhagic destruction of the brain, subependymal calcification, and congenital cataracts. Am J Hum Genet. 2010; 87: 882-89.

Moreno-De-Luca D; SGENE Consortium, Mulle JG; Simons Simplex Collection Genetics Consortium, Kaminsky EB, Sanders SJ; GeneSTAR, Myers SM, Adam MP, Pakula AT, Eisenhauer NJ, Uhas K, Weik L, Guy L, Care ME, Morel CF, Boni C, Salbert BA, Chandrareddy A, Demmer LA, Chow EW, Surti U, Aradhya S, Pickering DL, Golden DM, Sanger WG, Aston E, Brothman AR, Gliem TJ, Thorland EC, Ackley T, Iyer R, Huang S, Barber JC, Crolla JA, Warren ST, Martin CL, Ledbetter DH. Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. Am J Hum Genet. 2010; 87: 618-30.

Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC, Sobreira NL, Valle D, Rudd MK, Satten G, Cutler DJ, Pulver AE, Warren ST. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet. 2010 Aug 13; 87: 229-36.

Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, et al: A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 2005; 65: 6071–6079.

Nikolai Shokhirev. Hidden Markov model. Last Modified: 02/15/2010; retrieved: 04/16/2011.

http://www.shokhirev.com/nikolai/abc/alg/hmm/hmm.html.

Nilsen ST, Sagen N, Kim HC, Bergsjø P. Smoking, hemoglobin levels, and birth weights in normal pregnancies. Am J Obstet Gynecol 1984; 148: 752-8.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004; 5: 557-72.

Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Andersen AM, Taxbøl D, Hansen KD, Juhl M, Schow TB, Sørensen HT, Andresen J, Mortensen EL, Olesen AW, Søndergaard C. 2001. The Danish National Birth Cohort. Its background, structure and aim. Scand J Public Health 29: 300-307.

Ott J . Analysis of Human Genetic linkage. 1985. Johns Hopkins University Press, Baltimore.

Paulsen JS, Salmon DP, Thal L, Romero R, Weisstein-Jenkins C, Galasko D *et al*. Incidence of and risk factors for hallucinations and delusions in patients with probable Alzheimer's disease. Neurology 2000; 54: 1965–1971.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136-1148.

Perera FP, Tang D, Rauh V, Lester K, Tsai WY, Tu YH, Weiss L, Hoepner L, King J, Del Priore G, Lederman SA 2005 Relationships among polycyclic aromatic hydrocarbon-DNA adducts, proximity to the World Trade Center, and effects on fetal growth. Environ Health Perspect 113:1062–7.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nature Biotechnology 29: 512-521.

Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics. 2008; 24:309–318.

Prestia A. Alzheimer's Disease and Schizophrenia: Evidence of a Specific, Shared Molecular Background. Future Neurology. 2011;6:17-21.

Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ, Smith L, Greenfield A, Tiganescu A, Buckle V, Ventress N, Ayyub H, Salhan A, Pedraza-Diaz S, Broxholme J, Ragoussis J, Higgs DR, Flint J, Knight SJ. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. Nucleic Acids Res. 2005; 33: 3455–3464.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559-75.

Quintero-Rivera F, Sharifi-Hannauer P, Martinez-Agosto JA. Autistic and psychiatric findings associated with the 3q29 microdeletion syndrome: case report and review. Am J Med Genet A. 2010; 152A: 2459-67.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989; 77: 257-286.

Rigaill G, Hupé P, Almeida A, La Rosa P, Meyniel JP, Decraene C, Barillot E. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics. 2008; 24: 768–74.

Rizel L, Safieh C, Shalev SA, Mezer E, Jabaly-Habib H, Ben-Neriah Z, Chervinsky E, Briscoe D, Ben-Yosef T. Novel mutations of MYO7A and USH1G in Israeli Arab families with Usher syndrome type 1. Mol Vis. 2011; 17: 3548-55.

Rockwell E, Jackson E, Vilke G, Jeste DV. A study of delusions in a large cohort of Alzheimer's disease patients. Am J Geriatr Psychiatry 1994; 2: 157–164.

Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang Y, Chi H, Lin C, Holman K, Tsuda T, et al. Familial Alzheimer’s disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer’s disease type 3 gene. Nature. 1995; 376: 775–778.

Ronco AM, Arguello G, Munoz L, Gras N, Llanos M. Metals content in placentas from moderate cigarette consumers: correlation with newborn birth weight. Biometals 2005; 18: 233-41.

Saar K, Al-Gazali L, Sztriha L, Rueschendorf F, Nur-E-Kamal M, Reis A, Bayoumi R. Homozygosity mapping in families with Joubert syndrome identifies a locus on chromosome 9q34.3 and evidence for genetic heterogeneity. Am J Hum Genet. 1999; 65: 1666-71.

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ, Hulette C, Crain B, Goldgaber D, Roses AD. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993; 43: 1467-72.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. Strong Association of De Novo Copy Number Mutations with Autism. Science 2007; 316:4 45-449.

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P. HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Res. 2009; 37: W593-99.

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature. 1995; 375: 754–760.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosom. Cancer. 1997; 20: 399–407.

Spiegel R, Shaag A, Mandel H, Reich D, Penyakov M, Hujeirat Y, Saada A, Elpeleg O, Shalev SA. Mutated NDUFS6 is the cause of fatal neonatal lactic acidemia in Caucasus Jews. Eur J Hum Genet. 2009; 17:1200-203.

Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, Yu T, Kristensen VN, Perou CM. 2009. Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res 37: 5365-5377.

Sweet RA, Hamilton RL, Lopez OL, Klunk WE, Wisniewski SR, Kaufer DI *et al*. Psychotic symptoms in Alzheimer's disease are not associated with more severe neuropathologic features. Int Psychogeriatr 2000; 12: 547–558.

Sweet RA, Nimgaonkar VL, Devlin B, Lopez OL, DeKosky ST. Increased familial risk of the psychotic phenotype of Alzheimer disease. Neurology. 2002; 58: 907–911.

Sweet RA, Nimgaonkar VL, Devlin B, Jeste DV. Psychotic symptoms in Alzheimer disease: evidence for a distinct phenotype. Mol Psychiatry. 2003; 8: 383-92.

Tsui HC, Wu HD, Lin CJ, Wang RY, Chiu HT, Cheng YC, Chiu TH, Wu FY. Prenatal smoking exposure and neonatal DNA damage in relation to birth outcomes. Pediatr Res. 2008; 64:131-4.

Tunstall N, Owen MJ, Williams J, Rice F, Carty S, Lillystone S, Fraser L, Kehoe P, Neill D, Rudrasingham V, Sham P, Lovestone S. Familial influence on variation in age of onset and behavioural phenotype in Alzheimer's disease. Br J Psychiatry. 2000; 176: 156–159.

V. Petrushin. Hidden Markov Models: Fundamentals and Applications. Part 2: Discrete and Continuous Hidden Markov Models, Online Symposium for Electronics Engineers. 2000

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

Wang S, Chanock S, Tang D, Li Z, Jedrychowski W, Perera FP. Assessment of interactions between PAH exposure and genetic polymorphisms on PAH-DNA adducts in African American, Dominican, and Caucasian mothers and newborns. Cancer Epidemiol Biomarkers Prev. 2008; 17: 405-13.

Wang X, Zuckerman B, Pearson C, Kaufman G, Chen C, Wang G, Niu T, Wise PH, Bauchner H, Xu X. Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. JAMA. 2002; 287: 195-202.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ; Autism Consortium. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med. 2008; 358: 667–75.

Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. Biometrics. 1994; 50: 118-27.

Willatt L, Cox J, Barber J, Cabanas ED, Collins A, Donnai D, FitzPatrick DR, Maher E, Martin H, Parnau J, Pindar L, Ramsay J, Shaw-Smith C, Sistermans EA, Tettenborn M, Trump D, de Vries BB, Walker K, Raymond FL. 3q29 microdeletion syndrome: clinical and molecular characterization of a new syndrome. Am J Hum Genet. 2005; 77:154-60.

Wineinger NE, Kennedy RE, Erickson SW, Wojczynski MK, Bruder CE, Tiwari HK. 2008. Statistical issues in the analysis of DNA Copy Number Variations. Int J Comput Biol Drug Des 1: 368–395.

Winick JD, Blundell ML, Galke BL, Salam AA, Leal SM, Karayiorgou M. Homozygosity mapping of the Achromatopsia locus in the Pingelapse. Am J Hum Genet. 1999; 64:1679-85.

Wu FY, Wu HD, Yang HL, Kuo HW, Ying JC, Lin CJ, Yang CC, Lin LY, Chiu TH, Lai JS. Associations among genetic susceptibility, DNA damage, and pregnancy outcomes of expectant mothers exposed to environmental tobacco smoke. Sci Total Environ. 2007; 386: 124-33.

Wu J, Hou H, Ritz B, Chen Y. Exposure to polycyclic aromatic hydrocarbons and missed abortion in early pregnancy in a Chinese population. Sci Total Environ. 2010; 408: 2312-8.

Yau C, Holmes CC. CNV discovery using SNP genotyping arrays. Cytogenet Genome Res. 2008; 123: 307–312.

Zhao X, Li C, Paez JG, Chin K, Janne PA, et al: An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res. 2004; 64:3060–3071.

Zubenko GS. Do susceptibility loci contribute to the expression of more than one mental disorder? A view from the genetics of Alzheimer's disease. Mol Psychiatry. 2000; 5:131-6.