COMPUTATIONAL METHODS FOR CALCULATING MEIOTIC RECOMBINATION FROM NUCLEAR PEDIGREES

by

Nandita Mukhopadhyay

B. Tech (Hons.), Indian Institute of Technology, India, 1989

M.S., Case Western Reserve University, 1992

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Nandita Mukhopadhyay

It was defended on

March 29, 2016

and approved by

Daniel E. Weeks, Ph.D., Professor, Human Genetics and Biostatistics
Graduate School of Public Health, University of Pittsburgh

Candace M Kammerer, PhD, Associate Professor, Human Genetics
Graduate School of Public Health, University of Pittsburgh

Mary L. Marazita, Ph.D., Professor of Oral Biology, Human Genetics, and Psychiatry
Graduate School of Public Health and School of Dental Medicine, University of Pittsburgh

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

Eleanor Feingold, Ph.D.

COMPUTATIONAL METHODS FOR CALCULATING MEIOTIC RECOMBINATION FROM NUCLEAR PEDIGREES

Nandita Mukhopadhyay, PhD

University of Pittsburgh, 2016

ABSTRACT

Meiotic recombination is increasingly an important area for research in genetics. Recombination is critical for the proper segregation of chromosomes, and errors in recombination may result in chromosomal abnormalities and non-disjunction. Both the total number and the pattern of recombination events are known to vary genome-wide and from person to person. Using genome-wide genotype data to detect locations of recombination in individuals is the first necessary tool to study recombination. Earlier methods, e.g. CRI-MAP, used linkage-style modeling on three-generation families and sparse microsatellite markers to detect recombination events. More recently, methods using "streaks" of SNPs showing IBD status on dense GWAS SNP data have been used to score recombination locations in sibships. Here, I have developed a new SNP streak method to score recombination locations in pedigree types not previously handled, such as half-sibling pedigrees, and pedigrees with one or more ungenotyped individuals. We implemented our new method as a Python software package, MBFam. This package analyzes family-based genome-wide association datasets, accepting input data as PLINK binary files, a widely used input format for genetic data. The computation steps involve extraction of recombination probands, detection of recombination events, computation of recombination breakpoint locations and the offspring inheriting each recombination event, while accounting for Mendelian inheritance inconsistency errors and proximate double recombinations. MBFam has been extensively tested on the Mac OSX and Linux platforms. For demonstration purposes, this new method was applied to two family-

based GWAS datasets. Recombination intervals scored were used to create sex-specific average recombination counts (ARC) using all new pedigree structures and only the full-sibships. GWASs were conducted for male and female probands for both sets of ARCs. In one of the datasets, the added pedigree structures increased the female proband sample. This new method has the potential to significantly improve sample sizes for recombination studies, eventually leading to a better understanding of the biology of recombination and fertility, and benefitting the design of medical and public health interventions for improving maternal and child health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I dedicate this dissertation to my family, who have always believed that I would ultimately achieve the goal of earning a doctorate, and did their utmost to help me through this long journey.

I am deeply grateful to my advisor and mentor Dr. Eleanor Feingold for providing me the opportunity to work on such an exciting area of research. She was instrumental in my transition to the field of Statistical Genetics from my previous field of study, Computer Science. Starting from my very first Population Genetics course, which she taught, she has always guided and helped me to fulfil both my roles in the University of Pittsburgh, first as a staff researcher and secondly, as a graduate student.

I would like to thank my other committee members, Dr. Candace Kammerer, Dr. Daniel Weeks, and Dr. Mary Marazita for their invaluable guidance and advice. Dr. Weeks, in the capacity of my supervisor at the department of Human Genetics, taught me a great deal about research in Human Genetics, and also provided the first impetus for me to pursue a doctorate. I would like to thank Dr. Marazita, my current supervisor Dr. Manika Govil, and my colleagues at the Center for Craniofacial and Dental Genetics for providing the data upon which this dissertation is based, state-of-art computational resources and facilities, and last, but not the least, a great deal of encouragement and advice, all of which made my years of research exciting and enjoyable.

1.0 OVERVIEW AND SPECIFIC AIMS

Meiotic recombination is important for the proper segregation of chromosomes during meiosis. Errors in recombination result in a variety of chromosomal abnormalities and increased risk of non-disjunction [1, 2]. Both the total number, and the pattern of recombination events have been shown to vary across the human genome and from person to person [3-5]. This variation has been shown to have genetic basis [6]. Moreover, abnormal recombination rates in mothers have been shown to be associated with aneuploidy in their offspring [7, 8].

In order to study recombination in human pedigrees, the first necessary tool is a method for using genome-wide genotype data to detect locations of meiotic recombination breakpoints in individuals referred to as recombination probands. These are individuals who either have multiple genotyped offspring, or a genotyped offspring as well as genotyped parents. In a typical genomewide association study or GWAS of recombination traits, the first step is to identify the recombination probands. Next, their genotyped first-degree relatives are used to detect recombination breakpoints. Third, recombination phenotypic measures are created and used as outcomes within GWASs of recombination on these proband individuals.

The main task accomplished by the method described in this dissertation is to detect and record locations of recombinations in recombination probands. This task is related to but distinct from those addressed by linkage analysis and related methods, which are: identification of shared

chromosomal regions among related individuals, inference of any missing genotypes, or testing relationships.

The earliest method for detecting recombinations used linkage-style modeling on three-generation families and sparse microsatellite markers, e.g. as implemented in CRI-MAP genetic map creation program [9, 10], which use grand-parental phase to detect meiotic breakpoints. In the more recent studies of recombination, the most notable being the study performed by Kong et al. on the Icelandic population [11], phased haplotypes of 71,929 parent-offspring pairs (with all 4 grandparents genotyped) on approximately 30.3 million imputed and genotyped SNPs were used to resolve locations of recombination breakpoints with high accuracy. Phased haplotypes were created using genealogical information.

In practice, however, such large-scale phasing is usually not available for most population-based studies, so methods using densely-spaced GWAS SNP genotypes on nuclear families with two or more children [3, 12]were developed to detect recombination breakpoints based on "streaks" of identity-by-descent (IBD) in smaller, two-generation pedigrees. In SNP-streak, IBD between offspring is calculated separately for each parent's alleles. The IBD values are ordered by SNP position to produce streaks for each specific parent and offspring-pair combination. Switches in the IBD streaks are used to identify meiotic recombination breakpoints within chromosomes inherited by the offspring from each parent.

Once recombinations have been scored in a set of recombination probands, one can define various recombination phenotypes for these probands such as average recombination count, proportion of recombinations/recombinants within hotspots, recombinations/recombinants outside of hotspots etc. The recombination phenotypes are then used in genome-wide association (GWAS) analyses for the purpose of detecting genes that control recombination. Several recombination

GWASs have been performed in the recent past, and potential genes identified [4, 12-14]. Existing recombination studies are described in greater detail in the Background chapter.

Further details of SNP-streak and existing recombination GWASs are provided in the background section.

Here, we present a new SNP streak method for scoring recombination locations that can make use of family structures not previously considered. Our method handles half-sibships and three-generation pedigrees allowing for missing genotypes within such pedigrees. This substantially increases sample sizes for performing GWAS on recombination phenotypes. This new recombination scoring method has been implemented in software as the program MBFam (**M**eiotic **B**reakpoints in **Fam**ilies). In this study, we describe MBFam methodology, and then apply it to real GWAS datasets that include multiplex pedigrees. We also discuss guidelines for running the MBFam program.

1.1 SPECIFIC AIMS

In this work, I have developed methods and software to score recombination locations in pedigree types that have not previously been considered, including half-sibling pedigrees and pedigrees with some individuals ungenotyped. I assume that GWAS data are available and develop "SNP streak" type methods for finding recombination breakpoints. The new method is capable of analyzing extended pedigrees with missing genotypic information, thereby increasing the available sample for GWAS.

Preliminary analytical work on adding half-sibling and three-generation pedigrees was carried out by Dr. Ferdouse Begum [15]. Here, we have built upon Dr. Begum's work, and implemented it as a software program.

1.1.1 Aim 1 – Method development

In the first aim, I develop procedures and algorithms for (i) identification of recombination probands, (ii) examining genotype data of each proband's spouse, offspring and parents (if available) to extract locations of genome-wide recombination breakpoints within his/her offspring, followed by (iii) tabulating all observed recombination breakpoints for a proband by collective examination of his/her offspring's' recombinations, for the subsequent purpose of creating phenotypes.

1.1.2 Aim 2 – Software implementation

The second aim is to implement the methods in software.

Input data: Our method requires dense, genome-wide genotype data for families. Also required are information on familial relationships, genotype panel annotations such as allele labels and locations of variants. Suitable datasets, as described above, are typically gathered for conducting family-based genome-wide association study (GWAS). In my software implementation, I consider a few of the most common study data formats (such as PLINK [16] and linkage-format [10]). Only autosomal chromosomes are handled, i.e. human chromosomes 1-22, as there is no recombination within the non-pseudo autosomal regions of the sex chromosomes. The pseudo-autosomal region of the sex-chromosomes are also excluded from analysis.

MBFam software package: The MBFam software package has been designed to be object-oriented and modular, with a simple set of user-interaction steps and some customization. The programming language of choice is Python, a widely used language in scientific computing. Although Python is not the best language in terms of speed, it provides many other advantages in software design such as an object-oriented computation model, testing and performance evaluation, and portability across operating systems.

Output: Output files are formatted in a fashion to allow for processing with other software, e.g. I use comma-separated-value (CSV) formats extensively for use by Excel and statistical programs such as R. Output files are generated per-chromosome for the greater part, with some exceptions, where genome-wide output is also available.

1.1.3 Aim 3 - Application to real study datasets

In this aim, I analyze two family-based GWAS datasets as follows: a) apply the new recombination calling method to identify recombinations and recombinant offspring, and calculate the average recombination count (ARC) phenotype, b) confirm that the ARC values are in good agreement with the currently existing method, where applicable, and c) conduct GWASs of ARC, testing whether the increase in sample size due to the addition of the new pedigree structures also results in increased power to detect association.

1.1.3.1 Dental caries study: CARIES

The dental caries study (CARIES) data consists of multiplex pedigrees from a cohort from the Center for Oral Health in Appalachia study [17, 18] genotyped as part of the GENEVA project, on the Illumina Human660-Quad Beadchip. These families are ascertained by household, without

regard to any phenotype, and individuals with severe physical and mental disorders are excluded from the study.

For this dataset, the ARC measures obtained from a previous analysis carried out by Ms. Ferdouse Begum using the existing full-sibship based SNP-streak method [19] are available to us. These earlier ARC values will be compared against ARCs created by my method for probands who were included in both analyses, using only their full-sibling offspring.

1.1.3.2 Guatemala orofacial cleft study: OFC

The Guatemala cohort (OFC) is part of a larger multi-center study of oro-facial clefts, a common birth defect worldwide. The pedigrees were ascertained on the basis of cleft-palate and cleft-lip status, and are multi-generational. The genotyping panel is again the Illumina Human660-Quad Beadchip, carried out under the GENEVA project.

2.0 BACKGROUND

In this chapter, an overview of the study of recombination is presented starting from scoring recombinations to running a GWAS of recombination phenotypes. Input data necessary for such a study is described along with terminology and conventions used in the rest of the dissertation. Finally, a literature review of existing methods to score recombinations and recombination GWASs is presented.

2.1 OVERVIEW OF STUDY OF RECOMBINATION

In recent years, the study of meiotic recombination has been geared towards characterizing individual-level recombination patterns and the identification of genes that control these patterns. Meioses and recombination occur only in gametes, not in somatic cells, and cannot be directly observed within the recombination proband. Recombinations transmitted to the proband's offspring can be detected and localized by phasing the genotype data of proband's offspring with respect to his/her parents. Figure 1 presents a simple example of recombination scoring using phased genotypes at 2 adjacent SNPs within a pedigree that includes one of the proband mother's offspring and her parents. For the maternal allele at both SNPs within the offspring, the grandparent from whom the allele was inherited is determined using Mendelian inheritance laws. Recombinations can then be scored by looking for changes in grandparent-of-origin.

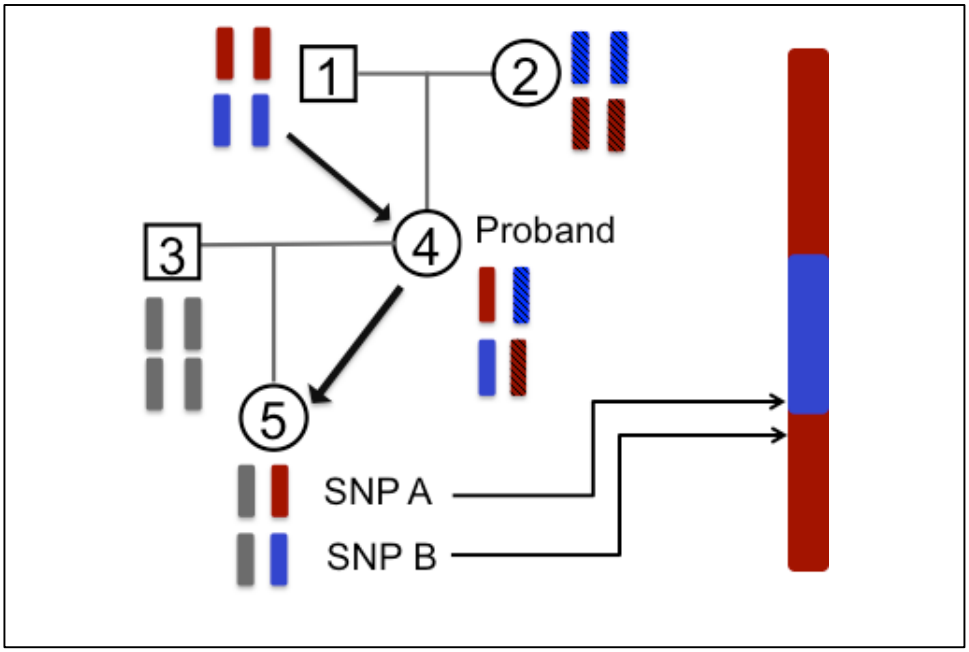


Figure 1. Detecting recombination from pedigree genotypes

Note: Individual 4 is the proband, red and blue shades identify the 2 different SNP alleles, hatching is used to distinguish grandmother's alleles from those of the grandfather, bold black arrows show transmission from grandfather to grandchild through proband mother, transmission is shown for 2 adjacent SNPs. The red-blue bar on the right represents the chromosome inherited by offspring from mother with locations of SNPs A and B as shown.

In this example, both SNPs are assumed to be genotyped for all pedigree members. In addition, the genotypes are informative for determining transmission from grandparent to grandchild. The grandparents' mating type is thus informative, as shown, and the proband's genotypes can be phased. In order to also phase the offspring with respect to her parents (proband and proband's spouse), the spouse should have homozygous genotypes at both SNPs (not shown in diagram). In subsequent sections, a SNP will be referred to as being informative with respect to a particular proband, if status at that SNP is determinate, else that SNP will be labeled uninformative.

Modern genetic data is typically available for closely spaced variants termed as single nucleotide polymorphisms or SNPs, either from a genotyping chip or from sequencing. A SNP-streak plot of grand-parental origin ordered by SNP locations for the pedigree shown in Figure 1 and

one of the autosomal chromosomes is shown in Figure 2. The horizontal axis shows location on the chromosome, and each SNP is represented by one point. In this figure, there are 3 distinct segments of grand-parental origin values, and hence, two recombination events. The methods used in scoring recombinations from GWAS data on 3-generation and other pedigrees are described in subsequent chapters. Once recombination events have been identified in the offspring, these can be used to create phenotypes for the proband and used to conduct an association analysis. In the rest of this dissertation, the term proband recombination will be used to refer to the recombinations observed in the probands' offspring.

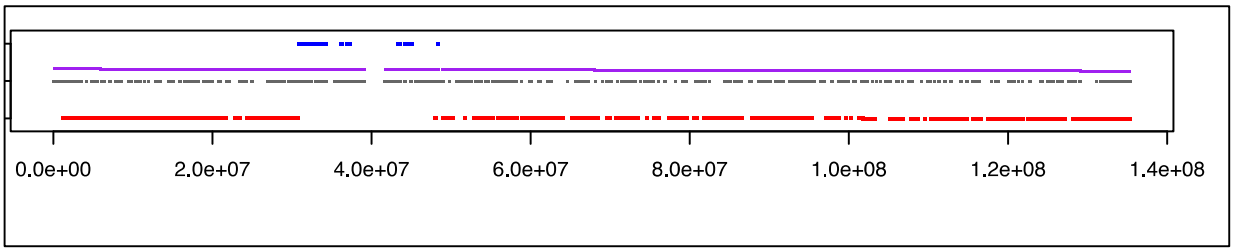


Figure 2. Offspring SNP-streak scan of a typical chromosome corresponding to pedigree shown in Figure 1

Red dots show SNPs for which the offspring allele inherited from proband belongs to grandfather, blue dots show SNPs at which the alleles originated from grandmother, gray and purple points are for ungenotyped and uninformative SNPs respectively, X-axis shows base-pair position along chromosome.

2.2 INTRODUCTION TO DATA AND TERMINOLOGY

In this section I introduce the structure of the data that is used by MBFam for recombination scoring, and define terminology used in the subsequent chapters.

2.2.1 Data

Pedigrees: Recombination scoring requires pedigrees with multiple generations genotyped. Sibling pairs, i.e. a pedigree consisting of at least two offspring and their parents are the smallest pedigree structure suitable for our purpose. The two offspring are not required to be full-siblings. Pedigree relationship information should be provided as a pre-makeped linkage-format file [10], which is the *de facto* standard for many commonly used genetic analysis programs. The sex of individuals is assumed to be present (as mandated by linkage-format), but is ignored by my recombination scoring process. The accuracy of pedigree relationships is paramount, so pedigree information should have been verified beforehand. This is usually not a problem, as several methods to detect pedigree relationship errors are available, such as PREST [20] and KING [21]. These methods use pairwise genetic sharing measures such as IBD and kinship coefficients to detect inconsistencies within the stated relationships. A more recent method by Zeng et al. takes into consideration the spatial distribution of IBDs, using the number of recombinations observed to verify unilineal relatives [22].

Markers: Markers are required to be densely and uniformly spaced. They are assumed to be bi-allelic in my method, as most SNPs are. Marker annotation in the form of a physical map needs to be provided. Incorrect map locations can seriously affect the accuracy of the results.

Genotypes: The genotype data need not be free of Mendelian errors, as detection of such errors is built into the scoring process, however, it is assumed to have been leaned using customary quality control checks used for high-throughput genotype data. The input genotype data can be combined across multiple chromosomes, and does not have to be ordered by genomic location, as long as accurate and complete marker annotation is provided. MBFam separately analyzes each

chromosome based on these annotations. Individuals with missing genotypes are allowed during the scoring procedure, however, only those individuals who have been selected based on acceptable genotyping rates can be used as recombination probands for GWAS of recombination.

2.2.2 Glossary of terms

The table below defines terminology used in the rest of this dissertation, along with examples in some cases for better clarity.

Table 1. Glossary of terms

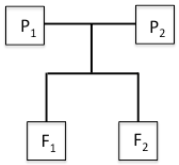
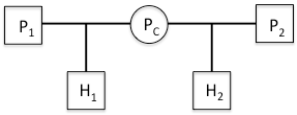
Term	Meaning	Illustrative instance, value or explanation as appropriate
Recombination proband	An individual whose recombinations can be scored	The proband is typically the parent in a two-generation pedigree or the middle generation in a three-generation pedigree
Proband sub-pedigree (PSP)	Sub-pedigree surrounding a proband, whose genotypes are used to score proband's recombinations	Includes spouse and offspring, and parents where available
Sibling pair proband sub-pedigree PSP-FS	A proband sub-pedigree consisting of a sib-pair and their parents	
Half-sibling pair proband sub-pedigree PSP-HS	A proband sub-pedigree consisting of a pair of half-siblings and all three parents	

Table 1 Contd.

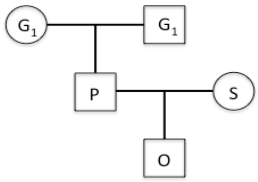
Term	Meaning	Illustrative instance, value or explanation as appropriate
3-generation proband sub-pedigree PSP-3G	A proband sub-pedigree consisting of the two parents, spouse, and an offspring of the proband	
Genotype Configuration	Ordered set of genotypes of a PSP at a SNP, ordered by pedigree-member; each configuration is mapped to a *status*	Example of a PSP-HS configuration: (P_1:1/1, P_C:1/2, P_2:2/2, H_1:1/2, H_2:1/2) Here, parent P_1's genotype is 1/1, P_C's genotype is 1/2, P_2's genotype is 2/2, offspring H_1's genotype is 1/2 and H_2's genotype is 1/2.
Complete (genotype) configuration	Genotype configuration for which all genotypes are available	The above example is a complete configuration.
Incomplete (genotype) configuration	Genotype configuration with one or more unknown genotypes	e.g. for a PSP-HS: (P1:1/1, P_C:1/2, P_2:0/0, H_1:1/2, H_2:1/2) where "0/0" represents the unknown genotype
Status	(i) IBD status of an offspring-pair at a given SNP in PSP-HS and PSP-FS (ii) grand-parental origin of allele inherited from proband in PSP-3G	Status values: S, D, 1, 2, U, I, N (see below for meaning of status values)
Informative configuration, status	Status for which IBD can be determined	IBD: **S**, same; **D**, different GP-origin: **1**, grandparent G1; **2**, grandparent G2

Table 1 Contd.

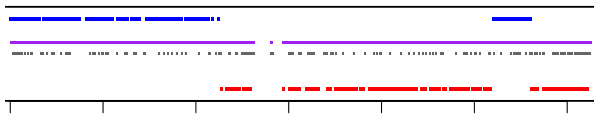
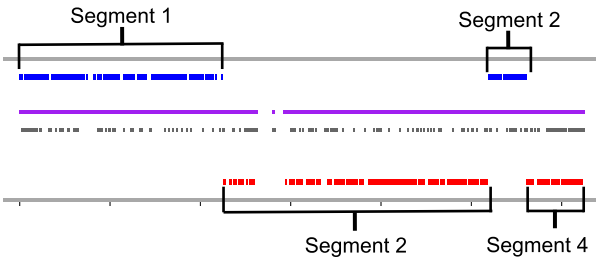
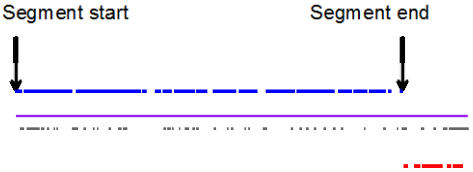
Term	Meaning	Illustrative instance, value or explanation as appropriate
Uninformative configuration, status	Mendelian inconsistent	Status value: **I**
	Unknown if status is ambiguous	Status value: **U**
	Not available if PSP genotypes are missing	Status value: **N**
Status scan	Statuses produced by a PSP, ordered by SNP	 Scan of a chromosome; each point depicts status at a SNP, blue:S/1, red: D/2, purple:U, gray: N
Segments	Sections of a status scan where informative statuses are identical (blue/red); segments are numbered left to right within each scan.	 Segments are always non-overlapping
Segment start, end	BP Position of first and last informative status within a segment measure	 Same as the start and end SNPs' map locations
Segment size	Number of informative statuses in a segment	Positive Integer

Table 1 Contd.

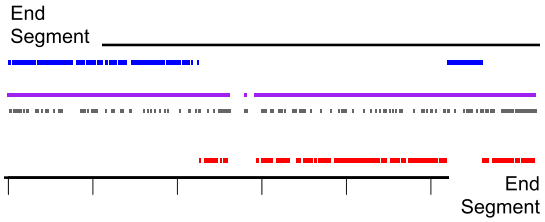
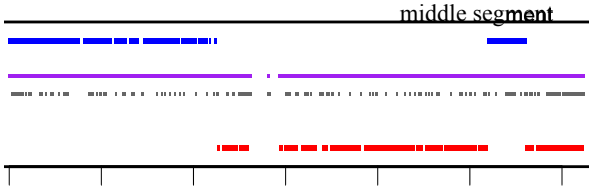
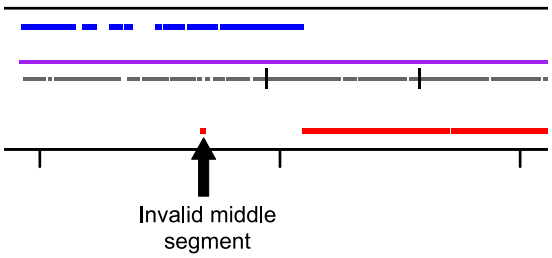
Term	Meaning	Illustrative instance, value or explanation as appropriate
Segment length	BP difference between segment start and end	Same as the distance between segment start and end SNPs
End segment	The first and last segments of a status scan (when ordered left to right by segment start positions)	 1st and 4th segments are end segments
Middle segment	Scan segments excluding the end segments	 2nd and 4th segments are middle segments
Segment filter	Minimum number of informative statuses in a segment to be considered a valid segment for scoring recombinations	**Default value** 10 SNPS (see definition of segment validity for applying filters)
Valid/Invalid segment	Segment satisfying/failing filtration criteria	

Table 1 Contd.

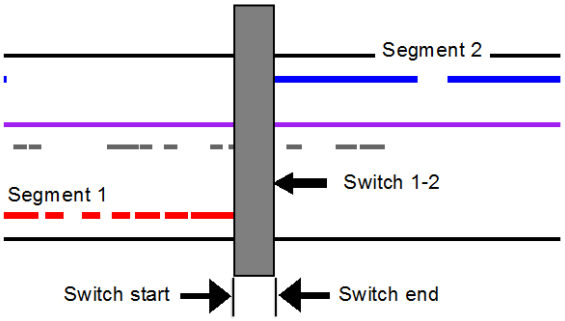
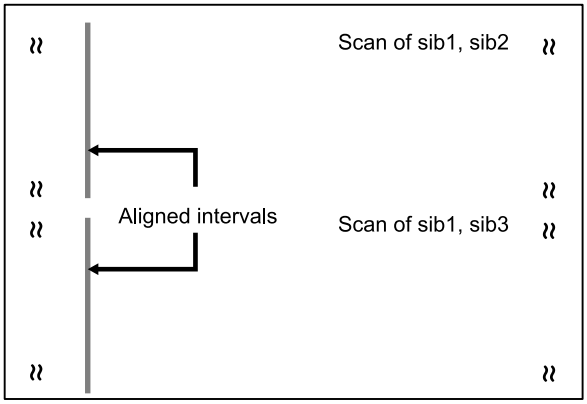
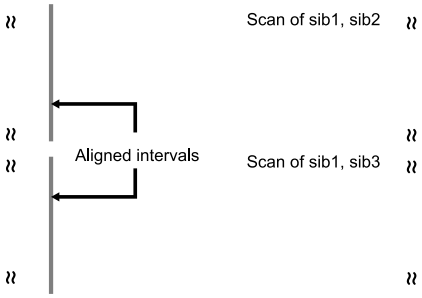
Term	Meaning	Illustrative instance, value or explanation as appropriate
Switch; **Switch interval start, end**	Interval region between two adjacent segments; **start** = left segment's **end** position; **end** = right segment's **start** position	
Aligned switch intervals; exact vs. approximate alignment	For scans from sibships > 2, sets of intervals from the separate status scans are **exactly** aligned if their switch **start** and **end** positions are identical; **approximately** aligned using other alignment criteria.	Intervals from sibship scans are exactly aligned; intervals from half-sibship scans may be approximately aligned. 
Offspring Participation count (for offspring in sibships)	The number of times an offspring contributed to a set of aligned intervals	 Participation counts for the 2 aligned intervals {sib1:2, sib2:1, sib3:1}

Table 1 Contd.

Term	Meaning	Illustrative instance, value or explanation as appropriate
Recombination interval	Pair of start and end positions representing a region with an odd number of recombination events	Each recombination interval is inferred from one or more switch intervals after the pairwise switch intervals belonging to a sibship or half-sibship have been aligned.

2.3 EXISTING RECOMBINATION DETECTION METHODS

In this section, three categories of existing methods for detecting recombinations are reviewed (i) the maximum-likelihood method CRI-MAP and its variants used to create genetic maps for linkage analysis, (ii) haplotyping methods, and (iii) SNP-streak. Detection of recombination events is equivalent to identifying intervals along a chromosome within which a recombination event occurred.

2.3.1 CRI-MAP

CRI-MAP is a tool to construct genetic maps from extended pedigree genotype data available on multiple co-dominant marker loci [9]. From an assumed ordering of the marker-loci, CRI-MAP uses multipoint maximum likelihood estimation of the average frequency of recombination crossovers based on the observed genotype data. Missing genotypes are filled in using Mendelian inheritance rules for non-founders, where possible. If inference of the actual genotype is not possible, it includes all the possible genotypes and their probabilities in the likelihood calculation. For missing founder genotypes, the population-based allele frequencies are used as probabilities, whereas, for missing non-founder genotypes, Mendelian segregation ratios are used.

CRI-MAP was developed at a time when only a very few marker loci were available, e.g. the first genetic map was created to include 60 RFLP loci on human chromosome 7 in 25 CEPH families, subsequently extended to include approximately 400 microsatellite markers spanning the genome. CRI-MAP makes use of numerical approximation (such as quasi-Newton optimization) and/or iterative search techniques (EM, or expectation-maximization, and Markov-chain based models) to speed up calculations, depending on the pedigree sizes and number of marker loci. Subsequent enhancements allowed parallel execution of these mapping functions on a distributed computing network, in order to overcome the computational complexity. CRI-MAP can also be used to compute haplotypes.

2.3.2 Methods for phasing and haplotyping entire chromosomes

Haplotyping methods were developed to improve the accuracy of linkage analysis. Thus, most modern-day linkage analysis software such as Merlin [23], Mendel [24]and SimWalk2 [25] to name a few, provide the capability to generate phased haplotypes for all individuals within a pedigree based on maximum-likelihood estimates. Table 2 below summarizes the method and output of several programs that can carry out haplotyping on pedigree data. All programs except Hapi handle extended pedigrees, i.e. multi- pedigrees that also contain related individuals besides parents and their offspring.

The phased haplotypes produced by these methods could be utilized with additional computation to score recombinations within recombination probands, however, all these methods are highly computation-intensive when applied to modern-day dense genotype data to multi-generational pedigrees. A previous study by Coop et al. [3] judged haplotyping software to be inadequate for scoring recombinations as they do not make provisions for genotyping errors that

may result in double-recombinations, i.e. putative crossovers occurring within a very short distance of each other.

Table 2. Haplotyping software for pedigrees

Method	Method	Output
Merlin [23], SimWalk2[25], Mendel [24], and Superlink [26]	Maximum likelihood of observed pedigree data	Phased haplotypes consistent with all pedigree individuals
Zaplo [27] and PedPhase [28]	0-recombinant/ Minimum-recombinant	Most common haplotypes
Hapi [29]	Minimum recombinant, Maximum-likelihood	Phased haplotypes for nuclear pedigrees

More recent studies of recombination performed by Kong et al. within the Icelandic population have used long-range phasing (LRP) to map recombination events [4, 11]. The LRP method was developed by Kong et al. to compute phase for SNP data using pedigrees [30]. It examines each individual in turn, classifying the other pedigree members into 2 groups of surrogate relatives including parents, representing maternal and paternal lineages respectively. The alleles are then assigned to one of these lineages using inference rules based on the Erdös distance between the individual and the surrogate relative taking into consideration whether the genotypes are heterozygous or homozygous. This process is also illustrated in detail in the context of an enhanced LRP method, LRPHL1 developed by Hickey et al [31]. In the Icelandic studies, phased haplotypes of 71,929 parent-offspring pairs (with all 4 grandparents genotyped) on approximately 30.3 million imputed and genotyped SNPs were used to resolve locations of recombination breakpoints with high

accuracy. However, such a large-scale pedigree-based genotype data along with complete genealogical records are rarely available for any population.

2.3.3 SNP-streak

SNP-streak methods use a heuristic algorithm to score recombinations on sibships, consisting of parents who are the recombination probands by looking at IBD of all genotyped offspring collectively. This method was developed to create a high-resolution map of crossovers to examine the variation in individual-level recombination patterns by Coop et al [3] and Chowdhury et al. [12], so named as it was based on analyzing "streaks" of IBD statuses of sibling-pairs, ordered by physical map location.

Figure 3 illustrates how IBD observed among the proband's offspring can be used to infer recombination events in the proband. For a sibship, both parents are potential recombination probands, and the concept behind using IBD status to infer crossover events has been shown for the mother. The offspring's chromosomes are represented as combinations of the maternal grandparents' chromosomes. Since grandparents are not available to determine phase in the mother, the offspring's chromosome cannot be thus labeled in reality. However, IBD between the offspring can be used to determine the phase switches due to meiosis in the mother. In figure 3, the black lines and arrows map underlying crossover events that switch the phase of the mother's chromosome from one of her parents to the other to switches in IBD status of the offspring.

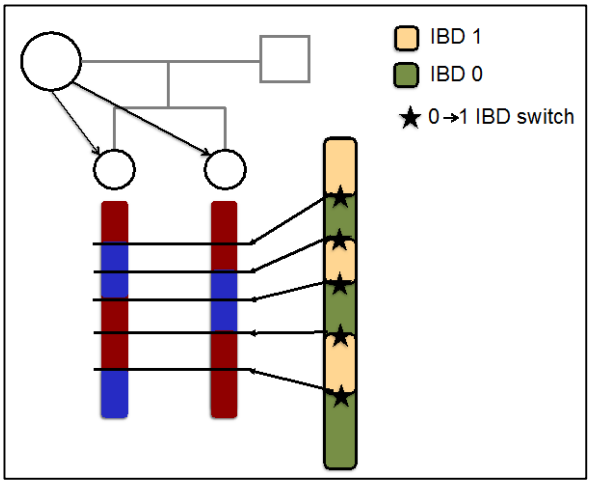


Figure 3. Use of identity-by-descent(IBD) to detect recombination

Note: Here, the red/blue regions represent unobserved phase within the mother circled in black, the green-yellow bar on the right shows IBD between the two offspring, lines indicate the correspondence between IBD switch and phase switch due to meiotic recombination.

To score recombinations for one of the parents from 2 or more offspring within a full-sibship, the SNP-streak algorithm first selects one of the offspring as the **template offspring**. The other **non-template** offspring are assigned IBD statuses at genotyped SNPs to indicate whether the allele inherited from the parent under consideration is the same or different as the template offspring. Switch locations where this IBD status changes from one to the other are identified in all the non-template offspring. At any given SNP location, if switches at that location are detected in a majority of the non-template offspring, the template offspring is assigned a recombination corresponding to that location. The compilation of recombination events across all observed offspring is used to create recombination phenotypes for the proband. Since the majority rule cannot be applied to a sibling-pair, each one is considered as equally likely to have the recombination. This does not affect the recombination phenotypes defined for the proband, which is the main target of a recombination study.

In order to eliminate double recombinations, multiple switches with fewer than a specified number of informative SNPs between them are flagged, (e.g. Coop et al's method sets the number SNPs to 5). If an odd number of switches were flagged, these switches are combined into one switch interval, else all flagged switches are discarded. For each template offspring, the algorithm then tests whether a majority (usually all) of non-template offspring show a switch in IBD status, assigning a recombination to that offspring and labeling that offspring as the recombinant if this is true. The exception is a sibling-pair, where only the first sibling is analyzed as a template. This is repeated for each IBD status switch on a chromosome. The flowchart in figure 4 summarizes the recombination scoring algorithm developed jointly by Coop and Chowdhury.

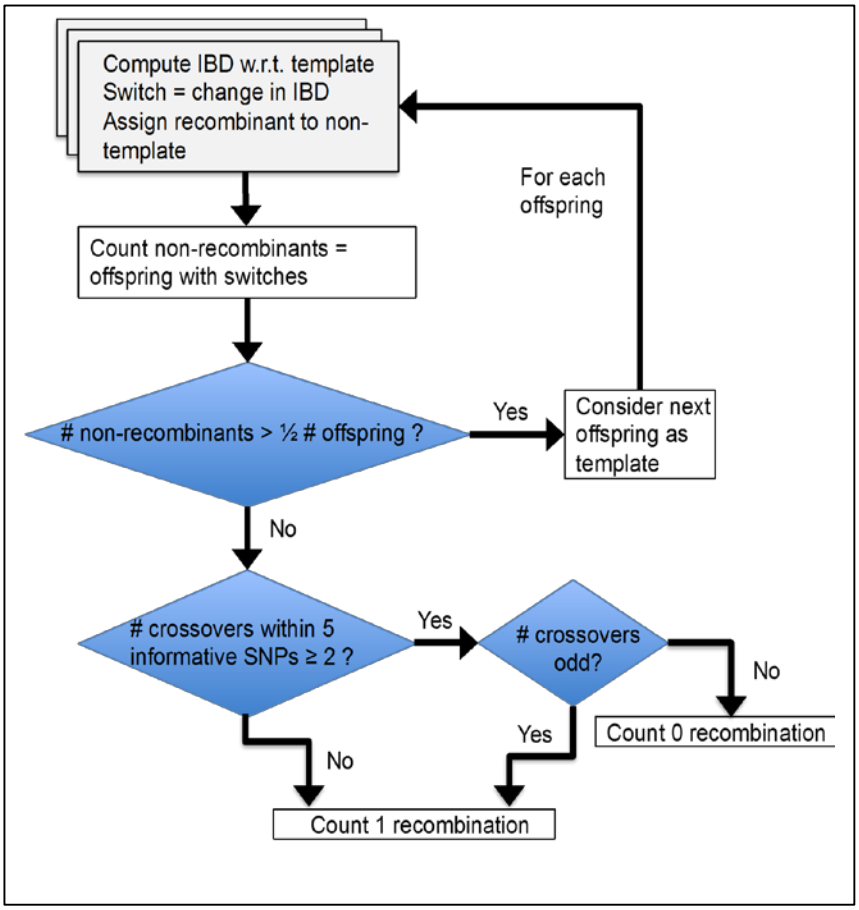


Figure 4. Flow-chart showing the steps of the existing SNP-streak method using full-sibships.

Note: In the first step, template refers to the current sibling being analyzed within the loop shown in the diagram

Although this existing SNP-streak method addressed only full-sibling nuclear pedigrees, the same principle can be applied to 3-generation pedigrees, by replacing IBD statuses with grand-parental origin. MBFam extends SNP streak methods to half-sibling pedigrees and to 3-generation pedigrees with and without ungenotyped individuals.

2.4 MAPPING GENES THAT INFLUENCE INDIVIDUAL-LEVEL RECOMBINATION

Several studies have conducted GWAS of recombination phenotypes to detect human recombination genes. A few of the prominent studies are summarized in brief below. In this section, a few recombination phenotypes that have been used by these studies are described briefly as well as previous findings with respect to recombination genes.

2.4.1 Recombination phenotypes

The most commonly used recombination phenotype is the Average Recombination Count (ARC), the total number of genomewide recombination intervals identified in all offspring of a proband averaged over the number of offspring. The *RNF212* gene has been associated with ARC. Location-based phenotypes, derived using the location of recombination intervals include, e.g. telomere or centromere usage, i.e., the fraction of crossovers that occurred in telomeric or centromeric regions of the chromosome, and historic hotspot usage, i.e. the proportion of recombination intervals overlapping hotspot regions, or average number per offspring. The converse of hot-spot recombinations can also be used as a phenotype, i.e. the number of recombination

intervals not overlapping hotspots. A third category of phenotypes is derived based on whether the recombination intervals span a known 13-bp long motif that occurs within hotspots identified by Myers et al. [32, 33].

2.4.2 2008 study by Kong et al.

This study performed a genome-wide search for genes using recombination rates (cM/Mb) as the phenotype on 1887 males and 1702 females genotyped on 309,241 SNPs. The recombination rate phenotypic values for these individuals were estimated previously using a 1000 marker microsatellite panel on 5,463 sibships by the authors. This study identified the RNF212 gene located in the 4p16.3 region as being correlated with recombination rates, as well as an inversion in the 17q21.31 region [13].

2.4.3 2009 study by Chowdhury et al.

Here, the recombination phenotype used is the average recombination count or ARC. The authors analyzed 511 females and 511 males the Autism Genetic Research Exchange (AGRE) on 350,000 SNPs. As in the previous study, sibships (with 2 or more siblings) were selected for analysis. Study findings consisted of six associated genetic loci including *RNF212* and the inversion on chromosome 17q21.31 previously reported in the Icelandic population. They also reported 4 new loci in *KIAA1462*, *PDZK1*, *UGCG*, and *NUB1*. The study findings implicated different genes for males vs. females. A replicate sample consisting of 1,633 maternal markers and 1,766 paternal markers in 654 females and 639 males taken from the Framingham Heart Study also produced similar results [12].

2.4.4 2011 study by Fledel-Alon et al.

This study extended the previous study by Chowdhury *et al*. by adding more SNPs and samples as well as more phenotypes, including hotspot use, Myers motif use, and two new recombination phenotypes. The new phenotypes used were the proportion of crossovers lying on telomeric regions as defined by the 20% leftmost or rightmost base-pairs, and for each of the two chromosome arms respectively. This study included 732 families in the Framingham Heart Study, 444 families in an Autism Cohort (AGRE). Nuclear families were extracted from larger pedigree structures to call recombinations. 454,934 and 390,671 SNP markers were analyzed for recombination on the two samples respectively. Subsequently GWAS was performed on 5 recombination phenotypes: mean rate of recombination, rate of crossovers in telomeres, rate of crossover in centromeres, hotspot usage and the Myers motif hotspot usage. *PRDM9* was found to be associated with hotspot usage, *RNF212* with the mean rate in males and females, and 17q21.31 with female mean rate. They failed to replicate the new genes reported in the Chowdhury et al. study. They also confirmed their findings in 163 families from a founder population of Hutterites [14].

2.4.5 2014 study by Kong et al.

This study analyzed the genome-wide recombination rate phenotype within a large data-set comprising 35,927 distinct parents, and 71,929 parent-offspring genotyped at 690,421 SNPs, subsequently phased and imputed at 3,000,000 variant locations. They scored a total of 2,264,323 recombination events, and identified 14 separate variants that influence genome-wide recombination rates. These variants are located in 8 separate regions on chromosomes 1, 4, 5, 14, 17 and 20, within

previously implicated genes and genomic regions such as *PRDM9*, *RNF21*, and the chromosome 17 inversion.

The enormous sample size in this study makes it highly likely that these gene discoveries are true positive associations (p-values range from $1.2e^{-48}$ to $6.5e^{-5}$), at least in the Icelandic population. Thus in this dissertation, I test MBFam by seeing if it can increase the power to find these genes within our datasets [11].

2.4.6 2015 multi-population study by Begum et al. (submitted)

In this study, the authors analyzed three GWAS data sets for five different recombination phenotypes: (i) average recombination counts ARC, (ii) recombination counts in hotspot regions, (iii) proportion of recombination in hotspot regions [14], proportion of recombination in non-hotspot areas and, (v) percentage of recombinations overlapping a 13 base-pair motif found in 40% of hotspots [33]. Genome wide association was run separately for males and females, as well as for both combined within each population. This study replicated associations for several previously reported recombination genes including *RNF212 and PRDM9.* For the non-hotspot recombination phenotype, *PRDM9* was reported to have different effects in males and females. Several new candidate loci were also implicated including regions near the *SPINK6*, *EVC2*, *ARHGAP25*, and *DLGAP2* genes [19].

3.0 MBFAM METHOD

In this chapter, I describe the computational procedures and algorithms developed for my new "SNP streak" method **M**eiotic **B**reakpoints in **Fam**ilies **(MBFam)**. These new procedures and algorithms extend Chowdhury et al. and Coop et al.'s existing SNP-streak method [3, 12] for sibships to general pedigrees genotyped on dense SNP panels, including ungenotyped individuals, thereby substantially increasing sample sizes for performing GWAS on recombination phenotypes. The software implementation details of MBFam is described in detail in the next chapter.

3.1 RECOMBINATION SCORING ALGORITHM

In this section, I first present an outline of the recombination-scoring algorithm, then address each step in detail. Starting with input pedigree structure information, genotype and SNP annotation data, the overall process for calling recombinations is as follows:

1) For each of the three PSP types, for all possible SNP genotype configurations, calculate IBD of PSP's offspring-pair or grandparent-of-origin of PSP's offspring.

2) Extract proband-subpedigrees (PSPs) from multiplex pedigree structures

3) Compute status for each PSP at each SNP by applying the appropriate rules from 1)

4) Create scans of statuses ordered by SNP position; filter poor quality status scans

5) Define status segments and eliminate spurious, short segments (double recombinations) caused by genotyping error

6) Identify status switch intervals for each PSP

7) a) For 3-Gen PSPs, label each switch interval as a recombination interval.

b) For probands who are members of sibpair and half-sibpair PSPs, collectively analyze switch intervals across offspring pairs to identify recombination intervals.

c) Further, for the probands in b) with 3 or more offspring, also identify the recombinant offspring corresponding to each recombination interval.

8) To conduct a GWAS of recombination, an additional last step is to create recombination phenotypes for a proband based on the aggregate of recombinations observed within that proband's offspring identified in the previous step.

3.2 RULES FOR COMPUTING STATUS

In this section the rules used for calculating statuses for each of the proband subpedigrees or PSPs are described for a bi-allelic marker (SNPs are assumed to be bi-allelic). First, the PSP types are described, followed by genotype configurations and the meanings of status labels assigned to the configurations.

Three types of PSPs are analyzed by MBFam, as defined below, and illustrated in Figure 5 (A). First rules are derived assuming a hypothetical bi-allelic SNP, and complete genotype configurations. Next, rules for incomplete genotype configurations are created using a pattern-matching process from the complete configurations. In Dr. Begum's dissertation

1. A three-generation PSP (PSP-3G) consists of the parents, spouse and offspring of the proband. The smallest such pedigree has 5 individuals, the two grandparents (G_1, G_2), the proband parent

(P), spouse of the proband parent (P_S) and the offspring (O), for whom recombination breakpoints are identified.

2. A half-sibpair PSP (PSP-HS) consisting of three parents, one of them, P_C married to the two others, P_1, and P_2, and two offspring, one from each marriage, O_1 and O_2. Only recombinant chromosomes transmitted by the common parent P_C can be detected in a PSP-HS, therefore P_C is the single proband of a PSP-HS.
3. A full-sibpair PSP (PSP-FS) consists of two parents (P_1 and P_2) and two genotyped offspring, O_1, and O_2. Recombinations can be scored for both P_1 and P_2, so a full-sibship PSP has two probands.

Genotype configurations: For each PSP and SNP, the ordered tuple of observed genotypes of the PSP is termed the genotype configuration. The ordering of genotypes within the configurations of the 3 types of PSPs (3-generation, half-sibling pair and full-sibling pair) is shown in Table 3A. A genotype configuration is complete if it does not contain any missing genotypes, else it is incomplete. Each configuration tuple is mapped to a status value consisting of either (1) **IBD**, i.e. whether the same allele was inherited from the proband by a pair of offspring at a SNP, or different alleles, or (2) **GPO,** i.e. which grandparent's allele was transmitted to the offspring via the proband. GPO is calculated only if genotyped grandparents are also available. The different status labels used to denote IBD and GPO are defined in Table 3B. The missing status **N** is assigned to a PSP unless at least 3 of its members are genotyped. The S, D, 1 and 2 statuses are informative, and SNP positions with these statuses are included in the detection of recombination breakpoints. SNP positions with uninformative and inconsistent statuses are ignored.

Table 3. PSP genotype configurations, status labels and definitions

A. Genotype configuration and label by PSP type		
PSP Type	**Order of genotypes in configuration**	**Status Labels**
PSP-3G	(G_1, G_2, P, P_S, O)	1, 2, U, I, N
PSP-HS	(P_1, P_C, P_2, O_1, O_2)	S, D, U, I, N
PSP-FS	(P_1, P_2, O_1, O_2)	S, D, U, I, N

B. PSP Status label definitions	
Status	**Definition**
1	Origin of allele transmitted by proband is grandparent 1
2	Origin of allele transmitted by proband is grandparent 2
S	Same proband allele transmitted to (half) siblings
D	Different proband alleles transmitted to (half) siblings
U	PSP genotyped but uninformative
N	PSP not genotyped
I	Mendelian inconsistent PSP

Illustrations of the use of Mendelian inheritance rules for determining statuses are shown in Figure 6 panels (i) to (iv). In the first 2 examples, the statuses of only complete (i.e. fully typed configurations) are derived. In panels (iii) and (iv), we consider incomplete configurations in which 1 and 2 genotypes are set to missing respectively. The next two sections describe how status is computed within each PSP at a particular SNP.

3.2.1.1 Computing status of a complete PSP configuration

First each genotype configuration is checked to see if the proband is a homozygote; such a configuration is automatically assigned an uninformative status, without any further consideration. Next, each configuration is checked for Mendelian inconsistency, by looking at the genotypes of its

constituent trios. A Mendelian inconsistency in any one of the trios results in flagging the PSP to be inconsistent for that configuration. For the remaining genotype configurations, further inference rules are now applied to compute IBD of GPO in the offspring of each PSP type. For example, in a PSP-3G with the configuration {G_1=*1/1*, G_2=*2/2*, P=*1/2*, P_S=*1/1*, O=*1/1*}, the PSP is assigned a GPO status of 1, meaning that the proband allele's is from grandparent G_1. If the offspring's genotype is instead *1/2*, the offspring's GPO status would be 2. Figure 3 panels (i) and (ii) show the rules for two complete and informative genotype configurations of a PSP-HS, one producing an S (same) status, the other a D (different) status.

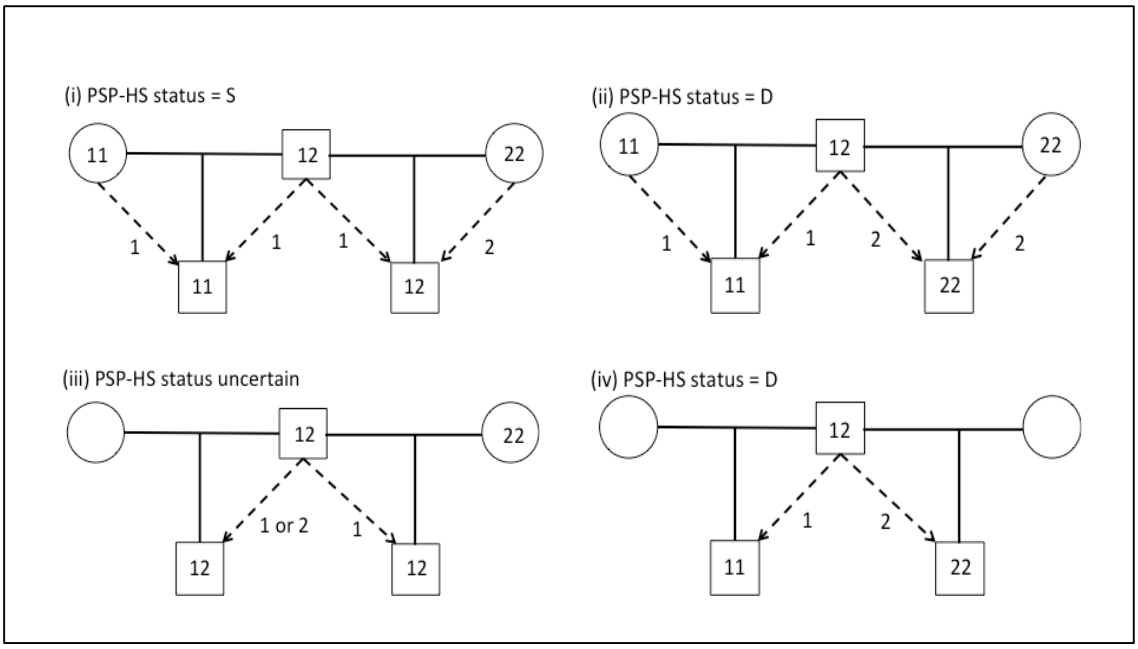


Figure 5. Calculating IBD status for completely genotyped PSP-HS

3.2.1.2 Computing status of an incomplete PSP configuration

Only PSP configurations that contain at least 3 genotyped individuals are scored, otherwise the status **N** (not available) is assigned to incomplete configurations with fewer than 3 individuals genotyped. For each of the remaining incomplete configurations, an extended set of mapping rules are generated as follows: For each unknown genotype, replace it with the 3 possible genotypes in

turn, to produce alternative fully typed configurations. For incomplete configurations missing 1 and 2 genotypes respectively, there are 3 or 9 such complete alternative configurations respectively. Consider the 3 or 9 statuses corresponding to these alternative complete configurations. If any of these 3 or 9 corresponding statuses are unknown, assign the unknown status.

a. If all these alternate statuses are Mendelian inconsistent, assign the Mendelian inconsistent status to the original genotype configuration with missing genotypes.
b. Else, if the Mendelian consistent status values (i.e. those that are U, 1, 2, S and D), are all identical, assign that single status value to the original configuration, otherwise, assign the unknown (U) status.

Two examples of incomplete but informative PSPs are shown in figure 6 panels (iii) and (iv).

Table 4 lists a part of the mapping rules for PSP-HS genotype configurations. The number of informative and Mendelian inconsistent configurations by PSP type is listed in Table 5. Table 4, part (A) consists of complete configurations. There are 918 configurations with missing genotypes for the 5 member PSP-3G and PSP-HS pedigrees and 189 for PSP-FSs. Table 4, part (B) contains 2 instances of incomplete configurations. The 1/2 genotype refers to both 1/2 and 2/1 heterozygotes and the unknown genotype is 0/0. The green and blue highlighted cells in part B refer back to the set of complete configuration rules in part A to be used for the respective incomplete configuration. The first incomplete configuration produces an unknown status as the 3 corresponding complete configurations do not result in a unique known status. Conversely, the second incomplete configuration results in the known status, “different” corresponding to IBD status 0.

The status mapping rules for complete configurations are identical to those derived by Dr. Begum in her dissertation [15]. The derivation for incomplete configurations used here was

developed independently of Dr. Begum's work, as her status rules were not available at the time. Subsequently, the entire set of status rules created by MBFam was verified against those reported in Dr. Begum's dissertation.

Table 4. Rules for half-sibling PSP

Spouse 1	Proband	Spouse 2	Half-sib 1	Half-sib 2	Status
A. Complete configurations					
1/1	1/2	1/1	1/1	1/2	Different
1/1	1/2	2/2	1/1	1/2	Same
1/1	1/2	1/1	1/1	2/2	Inconsistent
1/1	1/2	2/2	2/2	2/2	Inconsistent
1/1	1/2	2/2	2/2	1/1	Inconsistent
1/1	1/2	2/2	1/1	2/2	Different
1/1	1/2	1/2	1/1	2/2	Different
1/1	1/2	1/2	1/2	2/2	Same
1/1	1/2	1/2	1/2	1/2	Unknown
1/1	1/2	1/2	1/1	1/2	Unknown
1/1	1/2	1/1	1/1	1/2	Different
1/1	1/2	2/2	1/1	1/2	Same
1/1	1/2	1/2	1/1	1/2	Unknown
1/1	1/2	1/1	1/1	2/2	Inconsistent
1/1	1/2	1/2	1/1	2/2	Different
1/1	1/2	2/2	1/1	2/2	Different
B. Incomplete configurations					
1/1	1/2	**0/0**	1/1	1/2	**Unknown**
1/1	1/2	**0/0**	1/1	2/2	**Different**

Table 5. Genotype configurations by PSP type, status type and genotyping

PSP type	Fully typed	1 missing	2 missing	Total
	Total, Informative, MI	Total, Informative	Total, Informative	
3-generation	243, 36, 162	405, 42	270, 28	918
Half-sibling	243, 36, 162	405, 60	270, 28	918
Full-sibling	81,16, 52	108, 8		189

3.3 RECOMBINATION PROBANDS

Step 1 in the recombination scoring process is to identify individuals in a GWAS study data set for whom recombinations can be scored. These individuals are referred to as recombination probands. The following sections describe the procedure for identifying recombination probands, and the probands' family members whose genotypes will be subsequently analyzed for calling recombinations.

3.3.1 Extraction of probands and proband-subpedigrees

In this step, multi-generation pedigrees are broken down into sub-components as follows. An individual with genotyped parents and a genotyped offspring is a potential recombination proband. An individual whose parents are not present in the study, but who has at least two genotyped offspring is also a potential proband. Thus, within two full-siblings, it is possible to observe recombination breakpoints transmitted from both their parents, within two half-siblings, only recombinations inherited from their common parent can be observed, and within a single

offspring, recombinations inherited from each respective parent can be observed if the corresponding grandparents are also present in the study.

To extract PSPs, first parent-offspring trios are identified and then the trios combined to create the PSPs. For example, two trios with the same pair of parents can be combined to make a PSP-FS, whereas two trios with one common parent can be combined to create a PSP-HS. Figure 5 (B) shows the PSPs extracted by this process from an extended pedigree. Probands with multiple genotyped offspring, but without parents, may participate in multiple full-sibpair and half-sibpair PSPs, e.g. proband 6 in Figure 5B. A proband with genotyped parents as well as 2 or more genotyped offspring can belong to multiple PSPs of all three types.

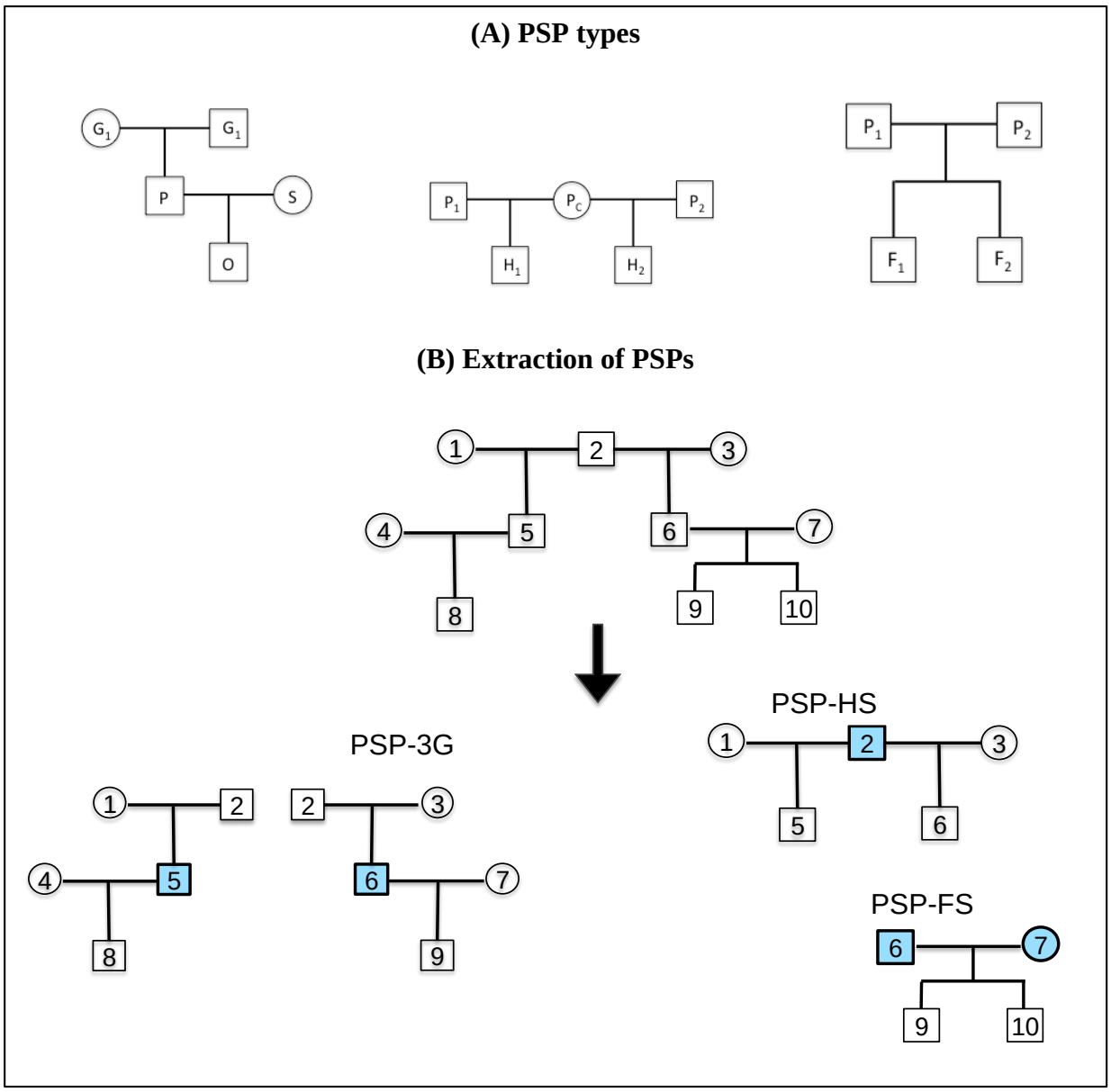


Figure 6. Recombination probands, PSPs and extraction of PSPs from a larger pedigree

Note: In (B) proband 6 is included in two PSPs, three-generation, and the other a full-sibling PSP.

3.4 COMPUTING STATUS SCANS AND SWITCH INTERVALS

In this section, I describe how statuses computed for each PSP and at each SNP are organized to form status scans. The process of filtering out poor quality status scans is also described.

3.4.1 Statuses

In the status computation step, genotypes of the members of each extracted PSP are analyzed to compute IBD and grand-parental statuses at each SNP using the rules derived in section 3.2 to calculate statuses at each SNP. For each SNP and PSP, the genotype configuration matching the PSP members' genotypes at that SNP is identified, and the corresponding status value assigned for that SNP position. Thus, after application of status rules to each SNP and PSP, the result is an array of status labels with values shown in Table 5, each label corresponding to a specific PSP and a specific SNP.

3.4.2 Status scans

A status scan is the list of statuses for a PSP, ordered by physical map position of the SNPs. For convenience, the genome-wide scan is broken up by chromosome, and stored as separate scans for each chromosome. Figure 7 panels (A) and (B) show 2 status scans typical of a GWAS panel on a single chromosome, for a PSP-HS and PSP-3G respectively. In each of the status scans shown in figure 3, the Y-axis represents the type of status (1, 2, S, D, N, I or U). In general, PSP-HS and PSP-FS scans are expected to have fewer U, i.e. unknown statuses, as the number of incomplete but informative configurations are greater than those for the PSP-3Gs.

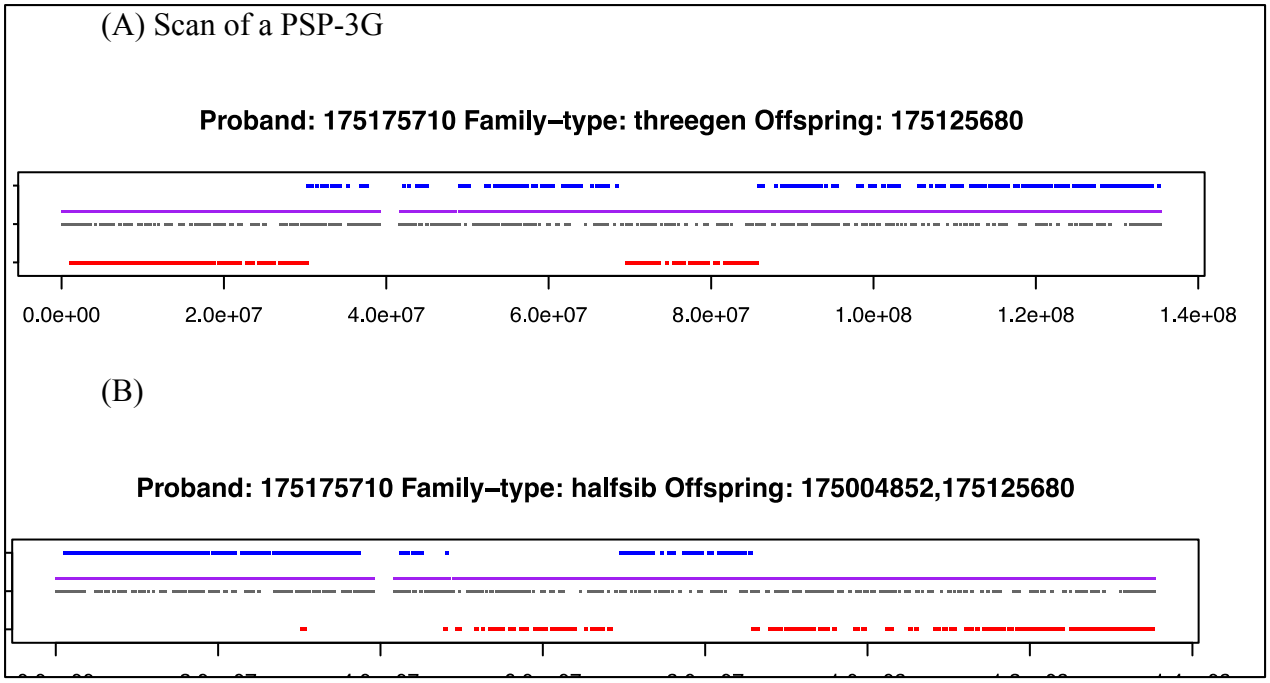


Figure 7. Status scans of (A) PSP-3G and (B) PSP-HS

Note: In (A) blue=GPO status 1, red=GPO status 2; in (B) blue=IBD status S, red=IBD status D; in both, purple=status U (unknown), grey=status N (not available); X-axis shows base-pair location; different types of status values are plotted at different Y-values for clarity.

3.4.3 Filtering poor quality scans

Status scans with large numbers of Mendelian inconsistent statuses are very likely due to the presence of a PSP member with low genotyping quality, and may result in inaccurate recombination calling down the line. Therefore, such PSPs should be excluded from subsequent analyses. The filtering process is described in this section.

To identify poor quality status scans, the total number of Mendelian inconsistent **I** statuses obtained for a PSP across the genome is compared to a pre-defined threshold value. This threshold value should preferably be based on what is observed when analyzing a data-set. Determination of the appropriate threshold value will be discussed in chapter 5, where application to real-data is presented. If the genome-wide count of **I** statuses exceeds this threshold, all scans of a PSP are

excluded from further analysis. The removal of a PSP does not necessarily imply that the proband itself is completely eliminated from recombination scoring, as his/her other PSPs may produce acceptable status scans. MBFam maintains a summary table and histograms of Mendelian inconsistency errors for the PSPs present in the entire GWAS dataset to help in setting the threshold value for retaining status scans.

3.5 STATUS SEGMENT AND SWITCH DEFINITION

This section deals with the partitioning of status scans into status segments, and locating switches in status values. Segments are sections of consecutive statuses within a status scan, in which the informative status values are the same. A status switch refers to the change in one status value to another, which corresponds to a recombination event. A switch is identified by the region between its two adjacent segments; **switch start** is the left segment's **end** position, and **switch end** the right segment's **start** position; the putative recombination is thus localized between the switch start and the switch end. In a PSP-3G, each switch interval on a segment represents the location of a putative recombination event in the proband chromosome inherited by the offspring. In PSP-HS and PSP-FS, a switch interval represents a recombination event in one of the offspring.

3.5.1 Segment definition

In the segment definition step, status scans are divided up into segments ignoring missing, unknown and inconsistent statuses. The value of the status, along with the first and last informative status SNP locations are stored as the segment boundaries, segment lengths computed, and segments

numbered left to right within each scan based on their start and end locations. Segments are strictly non-overlapping, as each SNP location is assigned a unique status value for a PSP.

3.5.2 Segment filtering and selection

In this step, segments are tested using a quality filter in order to decide if two recombinations lie too close by evaluating whether the intervening segment should be considered a real segment. Proximate double recombinations can be artifacts of genotyping errors that are undetected by Mendelian inheritance checks, and should be filtered out from the recombination-calling step. Low-quality segments are those that contain fewer than a pre-determined number of informative statuses, and such segments are assumed to be the result of proximate double recombinations caused by genotyping error.

In our method, different filtering thresholds can be set for the end and middle segments. The reasoning behind this is that, genotyping quality at the ends of the chromosomes may not be as high as that of the rest of the chromosome, therefore, the density of informative SNPs may be lower at the chromosome ends. Figure 8 shows a status scan containing an extremely short segment, which is regarded as a double recombination, presumably caused by genotyping error.

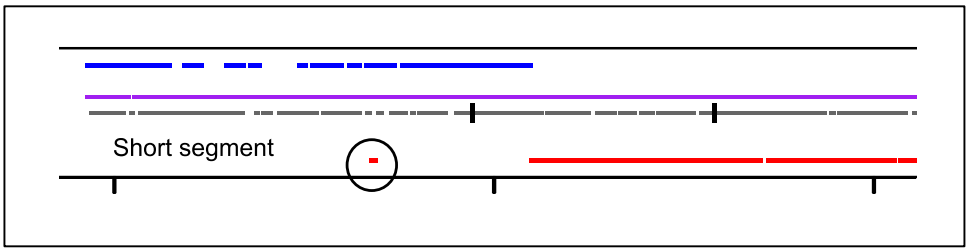


Figure 8. Status scan containing proximate double recombinations

Note: The circled segment fails the size criteria and will be removed leaving behind 3 segments.

3.5.3 Elimination of segments failing selection criteria

Within a middle segment containing fewer than the required number of informative statuses, all status values are set to unknown, which results in this segment being merged into its neighboring segments. Similarly, an end segment failing both the criteria is merged with the segment to its right or left segment depending on whether it is the first or the last segment. In either case, this merging produces fewer, but larger segments for each iteration. This merging is repeated until all segments meet quality requirements.

In Figure 8, due to the density of plotted statuses, there is no visible separation between the consecutive segments, however, recall that every switch in status marks the start of a new segment. So the scan shown in this figure consists of four segments including the one extremely short middle segment. Setting all statuses within the very short segment would result in two segments in the final scan, a blue (1/S) segment followed by a red (2/D) segment as follows. The small red segment is merged into the neighboring blue ones producing a single longer blue segment. The red segment at the q-end remains as is. Thus, removal of each short middle segment results in three segments becoming one.

3.5.4 Switch interval definition

Once the status segments are identified for all PSPs, switch locations are recorded. A switch interval is the region lying between two adjacent segments, containing only uninformative statuses. The **start** of the switch interval corresponds to the left segment's **end** position and the **end** of the switch interval corresponds to the right segment's **start** position. As for segments, switch intervals are non-overlapping. Switch intervals correspond to the locations of putative recombination

breakpoints. Thus, a switch interval is always flanked by two segments of opposite status types (GPO =1/2, IBD=S/D).

3.6 RECOMBINANT AND RECOMBINATION INTERVAL

This section describes the steps involved in identifying recombination locations and recombinant offspring following the creation of status scans, segments and switch intervals, of each individual PSP. PSP-3Gs are handled differently from the other 2 PSPs. In pedigrees containing genotyped individuals from 3 generations, recombination probands may end up in PSP-3G as well as one or both of the other 2. In our method, only PSP-3G scans are used for scoring recombinations for these probands.

3.6.1 Recombination scoring in PSP-3G

In a PSP-3G, segments coincide with the haplotype phase of the offspring with respect to the grandparents' chromosomes. Each status switch therefore, corresponds to a specific recombination event in the intervening parent (proband) within the scan of its PSP. Each PSP-3G of a proband contributes evidence of independent recombination events for the proband; all intervals observed within the offspring of PSP-3Gs represent unique recombination events and can considered to be unique recombination breakpoints. Therefore, if a proband has genotyped parents, only the intervals detected within his or her PSP-3Gs are retained, and those from its other PSPs ignored.

3.6.2 Recombination in PSP-HS and PSP-FS

In sibling-pair status scan (full or half-siblings), a switch in IBD status is due to a recombination inherited by one of the siblings. Other sibling-pairs involving this recombinant offspring may also show switches at corresponding locations, so a single recombinant can manifest itself as multiple status switches in different sibling or half-sibling pairs. Further analysis and interpretation of switches involving the same set of offspring must be carried out in order to identify unique recombination events. In figure 9(A), we illustrate how the one recombination transmitted to offspring 2 results in multiple IBD switches in the status scans of the three offspring pairs.

The 2008 study by Coop et al. on high-resolution mapping of crossovers provides a detailed description of calling recombinations over multiple siblings [3]. In their method, unique recombination intervals are called by comparing the scan of each sibling, designated as the "template" to every other sibling in turn, then each recombination is assigned either to the template, or one of the non-template siblings depending on which assignment produces the larger number of non-recombinant offspring for that interval. If the interval is observed only in the template and one other non-template, the recombination is automatically assigned to the non-template. This procedure implicitly minimizes the number of recombinant offspring.

In our method, we adopt a slightly different procedure. First, switch intervals across all PSP status scans of a proband are examined collectively, and sets of aligned intervals created by matching interval starts and ends, using the approximate alignment procedure described below. Then, for each interval set, a count is created of how many times each offspring appears in the offspring pairs contributing to the interval set, also known as participation count. The recombination is assigned to the offspring with the highest participation count. This procedure is illustrated in figure 9, panel (B).

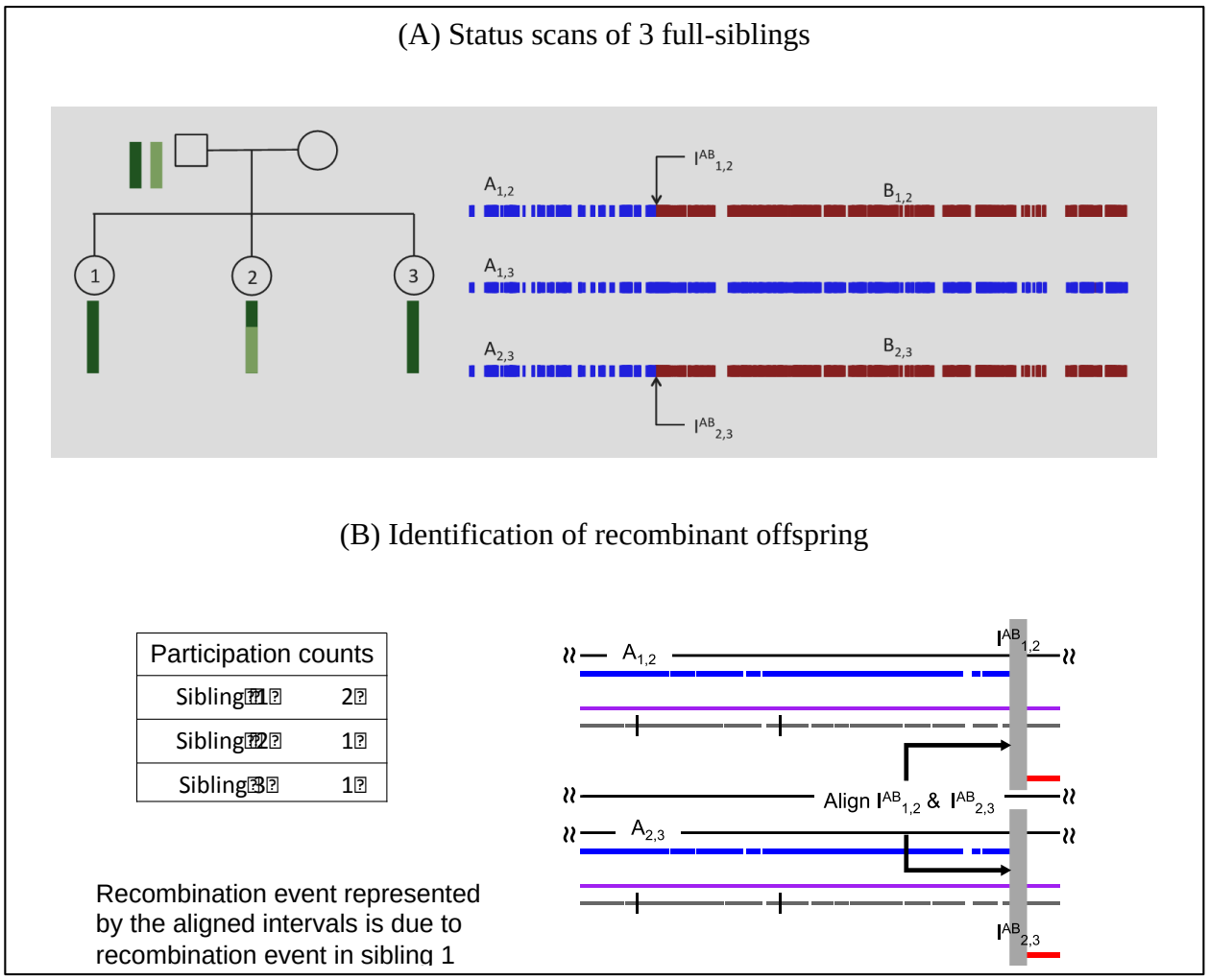


Figure 9. Alignment of status scans to identify recombination and recombinant

Note: (A) Status scans: shaded bars below offspring represent recombinant chromosomes inherited from father, phase in father is shown as light or dark green, in BD status scans blue and red segments consist of S and D statuses respectively. (B) Interval alignment and sibling participation counts, the scan and intervals correspond to those shown in (A).

Assuming that dense genomic scans produce very narrow switch intervals, and that independent recombination events transmitted to multiple offspring are unlikely to occur at the exact same location, exactly one offspring will have a participation count larger than 1; all others' counts should be 1, just as shown in figure 9, panel (B). If an interval is observed within only two siblings, the recombinant cannot be identified (similar to Coop et al.'s method). So, for such an interval, the recombination event is assigned to both the offspring jointly, i.e. when the recombinations are counted for a proband's offspring, each of the two offspring will contribute half a recombination.

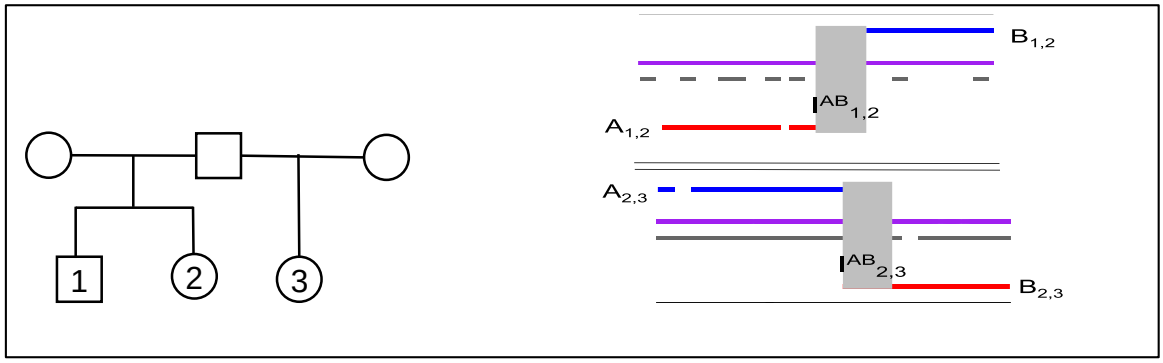


Figure 10. Difference in interval starts and ends of switch intervals from two PSP-HS of a proband

Status scans: $A_{1,2}$ and $B_{1,2}$ are for full-sibling pair 1,2,; $A_{2,3}$ and $B_{2,3}$ are for half-sibling pair 2,3; Interval $I^{AB}_{1,2}$ and $I^{AB}_{2,3}$ correspond to respective switch intervals.

Approximate interval alignment procedure: The boundaries of status scan switch intervals from multiple sibpairs may not be all be at exactly the same locations. There are two reasons:

(1) Differences in pattern of informativeness: Half-sibling pairs come from two separate mating types, therefore, a SNP may have an informative IBD status for one set of offspring, and not for the other. In order to set phase in each offspring, one parent has to have the heterozygous genotype, while the other parent is required to be a homozygote.

(2) Difference in patterns of missingness: Not all the same SNPs are genotyped for all offspring, thus there is variation within the location of missing statuses across scans of the different offspring-pairs of a proband.

Switch interval starts and ends are determined by the locations of informative status labels (S or D) calculated at each SNP, therefore when multiple pairwise scans of a proband's offspring are compared base-pair to- base-pair, the switch intervals corresponding to the same recombination event may not start and end at identical locations. An instance of this variation in interval boundaries is shown in Figure 10. The status scan in Figure 10 has been modified to display both S and D

statuses from a scan as a single row, rather than separately as was the case in previous scans. In MBFam, two intervals are considered to be aligned if there is either partial or complete overlap between them. Thus, switch intervals from multiple status scans from the same sibship (or half-sibship) are grouped into sets of intervals as these scans are analyzed in turn. After all scans have been analyzed, a single recombination interval is defined for each group of switch intervals by taking the shortest interval that can be created, i.e. by taking the maximum of the lefts and minimum of the rights to represent the recombination interval.

In the existing method of Coop et al. and Chowdhury et al. only full-sibships and SNPs with 100% genotyping are included in the analysis, therefore, switch intervals from a single recombination event match exactly with respect to their start and end positions [3, 12]. Our approximate switch interval alignment allows the inclusion of half-siblings as well as the handling of missing genotypes.

3.7 RECOMBINATION PHENOTYPES OF PROBANDS

The unique sets of switch intervals across all the offspring of a proband are identical to the total number of unique recombination events transmitted by the proband to his/her offspring. As described earlier, all switch intervals observed within offspring of PSP-3Gs represent unique recombination events, and do not need alignment. Where the proband is scored through full- and half-sibling PSPs, the aligned switch intervals are the recombination intervals, i.e. regions on the chromosome where a recombination event has occurred.

Recombination phenotypes for the proband are created based on the total number of recombinations observed across all his/her offspring. The most basic phenotype is the average

genome-wide recombination count or ARC obtained by dividing the total number of recombinations observed for a proband by the number of offspring included in the recombination scoring analysis. Others may consist of averaged counts within specific chromosomal regions such as hotspots, telomeric/centromeric regions etc. as described in chapter 2. Genome-wide association of recombination phenotypes is then carried out with the genotyped probands. Thus, we use the genotype data twice: (1) to detect recombination breakpoints along chromosomes and create recombination phenotypes, and (2) to run a genome-wide association analysis with recombination phenotypes.

4.0 SOFTWARE IMPLEMENTATION OF MBFAM

In this chapter, I describe the software implementation of the MBFam recombination scoring method. MBFam is implemented in object-oriented Python with an easy-to-use command-line interface. Graphing functions are implemented in R.

4.1 INPUT OF GWAS DATA AND CONTROL PARAMETERS

MBFam accepts PLINK binary genotype files [16] as its input format, and also stores genotypes in binary format, as 2 bit genotypes, following the convention used within PLINK. Control parameters include locations of input data files, input file format, genome-wide total number of Mendelian inconsistencies allowed per offspring/offspring-pair (default 800), segment quality criteria (by default, set to a minimum of 10 informative SNPs for both end and middle segments), and run label (for identifying output files). Input GWAS data and control parameter values are read in by the **Input Module.** Input and output file names, and control parameter values are supplied to MBFam by means of a control file, similar to the one shown below for a single chromosome.

Control file parameters

```
Map-file = chr16.bim
Input-format = plinkbed
Output-debug? = False
Genotype-file = chr16.bed
Pedigree-file = chr16.fam
Output-prefix = chr16_out
Separate-chromosomes? = True
Min-snp-numbers=10,10
```

Control parameters are provided as (keyword, value) pairs, and white-space is ignored.

4.2 RECOMBINATION SCORING MODULES

The flowchart in figure 12 shows the processing steps and modules implemented within MBFam. Scoring of recombinations is done in several stages, each using output from the previous stage to perform the required computations. Each stage of the recombination scoring process can be run or rerun separately, as output from the earlier stages become available.

Stage 1. **Pedigree Module:** Genotype, pedigree and SNP annotation data are read in, PSPs extracted, and statuses generated using the status rules.

Stage 2. **Status Module:** Status scans for each PSP are ordered by map-position and stored within a separate file that can be used by an R graphing utility. Mendelian inconsistent statuses are counted across the genome for each offspring/offspring-pair and excluded from further analysis if the total count exceeds the allowed number.

Stage 3. **Segment Module:** Clean status scans are traversed to create segments of identical status values. Each segment is checked for validity, and all statuses within it set to unknown if there are fewer than the required number of informative statuses.

Stage 4. **Interval Module:** Switch intervals between segments are defined. PSP-3G switch intervals are not processed further in this step. PSP-FS and PSP-HS switch intervals boundaries are aligned using the alignment tolerance, and unique intervals identified subsequent to alignment. For each unique interval, based on all offspring-pairs showing the interval, count how many times an offspring is part if these pairs (interval participation count).

Stage 5. **Recombination Module:**

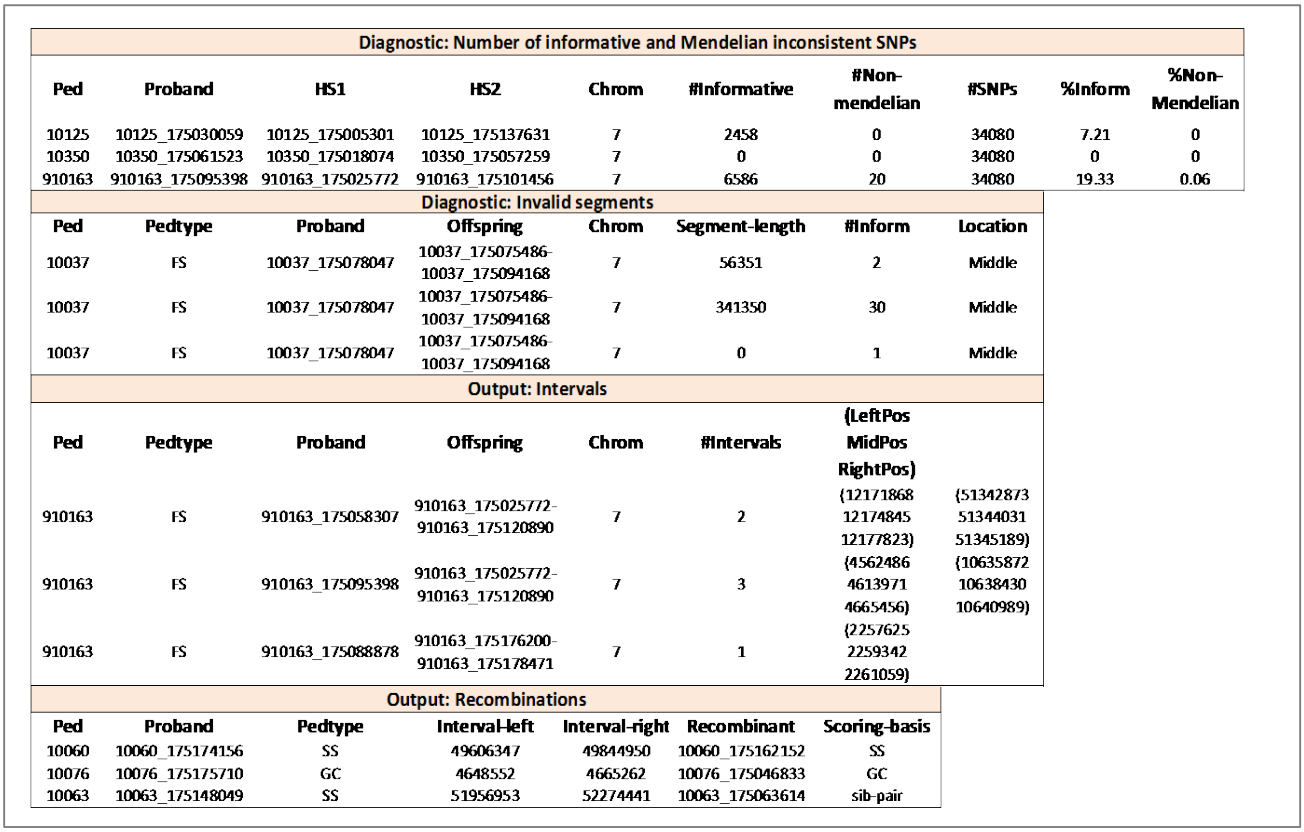
a. Count each switch interval for **PSP-3Gs** as a recombination interval; the offspring showing the interval is a recombinant.

b. Count each aligned interval as a recombination interval. For sibships larger than two, the offspring with the maximum interval participation count is the recombinant for that interval.

c. Count each aligned interval as a recombination event, assign recombination interval to both siblings, and make a note that this interval belongs to a sibling-pair.

4.3 OUTPUT AND DIAGNOSTICS

Output files consist of both recombination-related results such as the status scans, switch intervals and recombination intervals, as well as diagnostics such as genotyping rates, the number/proportion of uninformative and Mendelian inconsistent SNPs for each PSP by chromosome and genomewide, and lists of segments that were discarded for not meeting size criteria. The **Output module** includes methods to produce output during the various computation stages. Recombinant offspring and the locations of recombination intervals are the final output of MBFam. Snippets of each type of output file are reproduced below.



Diagnostic: Number of informative and Mendelian inconsistent SNPs

Ped	Proband	HS1	HS2	Chrom	#Informative	#Non-mendelian	#SNPs	%Inform	%Non-Mendelian
10125	10125_175030059	10125_175005301	10125_175137631	7	2458	0	34080	7.21	0
10350	10350_175061523	10350_175018074	10350_175057259	7	0	0	34080	0	0
910163	910163_175095398	910163_175025772	910163_175101456	7	6586	20	34080	19.33	0.06

Diagnostic: Invalid segments

Ped	Pedtype	Proband	Offspring	Chrom	Segment-length	#Inform	Location
10037	FS	10037_175078047	10037_175075486-10037_175094168	7	56351	2	Middle
10037	FS	10037_175078047	10037_175075486-10037_175094168	7	341350	30	Middle
10037	FS	10037_175078047	10037_175075486-10037_175094168	7	0	1	Middle

Output: Intervals

Ped	Pedtype	Proband	Offspring	Chrom	#Intervals	(LeftPos MidPos RightPos)	
910163	FS	910163_175058307	910163_175025772-910163_175120890	7	2	(12171868 12174845 12177823)	(51342873 51344031 51345189)
910163	FS	910163_175095398	910163_175025772-910163_175120890	7	3	(4562486 4613971 4665456)	(10635872 10638430 10640989)
910163	FS	910163_175088878	910163_175176200-910163_175178471	7	1	(2257625 2259342 2261059)	

Output: Recombinations

Ped	Proband	Pedtype	Interval-left	Interval-right	Recombinant	Scoring-basis
10060	10060_175174156	SS	49606347	49844950	10060_175162152	SS
10076	10076_175175710	GC	4648552	4665262	10076_175046833	GC
10063	10063_175148049	SS	51956953	52274441	10063_175063614	sib-pair

Figure 11. Examples of output produced by MBFam

R functions are provided to create plots of statuses grouped by proband over all his/her PSPs, as well as histograms of segment lengths and other diagnostic output.

An additional **Recombination phenotype module** has been implemented to write a PLINK phenotype file based on the average recombination count (ARC), along with "keep" lists of male and female probands [16] to facilitate sex-specific GWAS of ARC. Alternatively, genomewide total recombination counts for probands and offspring can be written out as comma separated value format files.

4.4 RUNNING MBFAM

MBFam can be run on data combined over multiple chromosomes, and all its computation steps can be executed one chromosome at a time. Further, the user may opt for running the entire recombination scoring process as a single run, or run it in stages, while also interleaving the genomewide diagnostics calculation steps with the computational stages. These alternatives are described below. The flowchart in Figure 12 highlights the various stages of computation.

4.4.1 Single MBFam run

When MBFam is invoked as a single run, all the computations proceed one after another, and diagnostic reports are produced at the end. In this case, the segment size filters for double recombinations) and the default interval alignment tolerance, have to be specified as run parameters inside the control file prior to the run. The current implementation of MBFam does not allow the filtering of PSPs based on levels of genome-wide Mendelian errors in the single-run mode

4.4.2 Multi-stage MBFam runs

The MBFam program can be run separately for each of the four stages shown in Figure 12, status-calling, segments, intervals, and recombination. The counts for Mendelian error levels should be invoked after status-calling as shown. Following the segment calculation, summaries of segment lengths can be examined to determine the size filters for flagging double recombinations. As yet, the alignment tolerance threshold needs to be set prior to interval calling. The reporting of the information required for setting an appropriate tolerance, namely the differences between closely matching intervals is underway.

4.4.3 Chromosome-specific MBFam runs

When running MBFam on data separated by chromosome, multiple control files need to be provided, one for each of the chromosomes. Functions to compile genomewide diagnostic summary tables for Mendelian errors and excluded segments is provided.

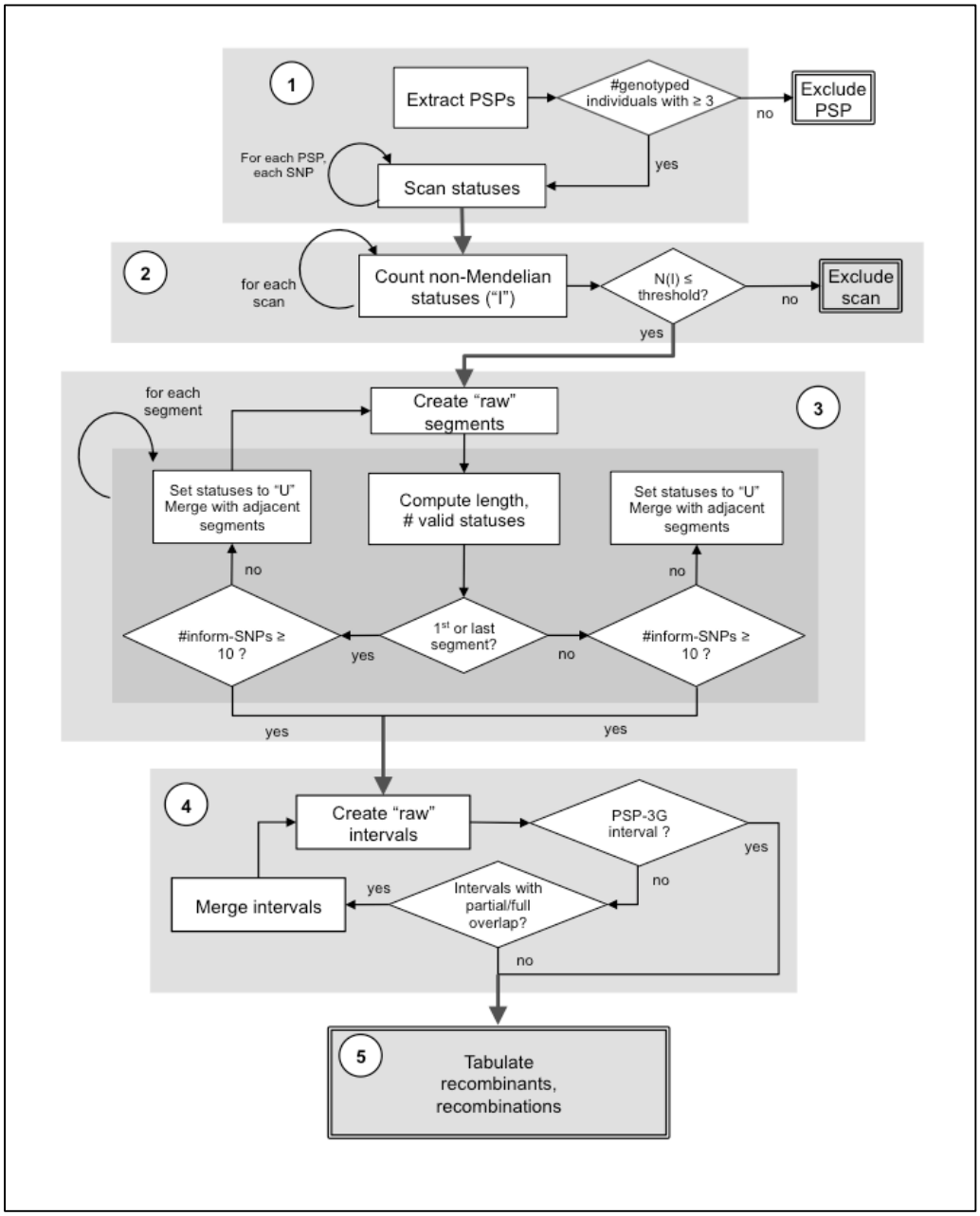


Figure 12. Recombination scoring flowchart

5.0 REAL DATA APPLICATION

In this chapter, we describe the application of MBFam to two population-based GWAS datasets, a Dental Caries (**CARIES**) data set from rural Appalachia and an oro-facial cleft (**OFC**) data set from Guatemala.

The central motivation of this dissertation is that using the new, enhanced SNP-streak recombination calling method increases sample sizes for a recombination GWAS, thereby increasing the power to detect association. Our real data applications are designed to demonstrate this increased sample size for real pedigree collections. For each of the two datasets, we created two analysis sets, one consisting of all probands found by MBFam and the other consisting of only the probands that could be used with previous methods (i.e. full-sib pedigrees with both parents genotyped). We use these two versions of each dataset to perform a GWAS for the ARC phenotype. In addition, we specifically examine the power of each version of the dataset to detect association with the genes reported within the Icelandic population by Kong et al., as we considered these genes to be "true" recombination genes. The Icelandic study involves a large sample (35,927 parents, and 71,929 parent-offspring pairs), and the association p-values reported are highly significant (on the order of 10^{-50}).

The following sections describe the application of MBFam recombination scoring, followed by genome-wide association of the average recombination count (ARC) in the **CARIES** and **OFC** data sets. First, the overall study design used to analyze both datasets is outlined. Then, study-specific details such as the number of probands, offspring used to score recombinations, number of

recombinations scored etc. are presented, along with genome-wide association and candidate gene association results.

5.1 STUDY DESIGN

This section presents an outline of the analyses performed on the two population-based GWAS datasets. Details specific to each data set are presented in the later sections. For each data set, we scored recombinations using our MBFam method and then ran GWASs of the ARC (average genomewide recombination count) phenotype separately for male and female probands. The allowed number of Mendelian inconsistent statuses genome-wide per PSP was set to 800 for the CARIES dataset. The OFC dataset was cleaned of Mendelian errors prior to our analysis, and none were detected during our analysis. Table 6 below gives an overview of the two samples with respect to number of probands, male and female, and number of offspring scored. Note that in table 6 below, probands may belong to more than one PSP.

Table 6. Recombination GWAS sample size for CARIES and OFC

	CARIES	**OFC**
Pedigrees	403	93
Probands	598	163
Male	395	56
Female	203	107
Offspring	993	466
Informative PSPs	1390	707
PSP-FS	1018	548
PSP-HS	319	59
PSP-3G	53	93
Recombinations	50,117	14,660

5.1.1 Comparing recombinations scored using all three PSP structures to using only full sibships

In order to observe the effect of adding in new pedigree sub-structures, two sets of probands were created, the full sample consisting of all three PSP structures referred to as **ALL**, and a subset scored using only PSP-FS structures, referred to as **SIBS**, within each data set. A comparison between the proband sample sizes of ALL vs. SIBS is presented for each dataset. For comparison purposes, the recombination probands present in the two samples have been sub-divided into 5 categories: (1) probands whose PSP type and number of offspring remained unchanged, but for some, combining half-siblings produced additional offspring pairs; (2) probands with parents added in, thereby number of offspring remained constant while PSP type changed, (3) probands from multiple marriages, where at least one pairing had a single offspring genotyped; this offspring was added to the analysis in the ALL set, (4) and (5) probands present only in ALL, who had only one offspring per marriage, and so were not scored within the SIBS set; in category 4, only nuclear families of probands are available, and in category 5, genotyped grand-parents were also available. The first 3 categories are explained in figure 13. In figure 13, recombination probands are identified as colored individuals, and labels are used to identify the common pedigree members across PSPs in either sample, where necessary.

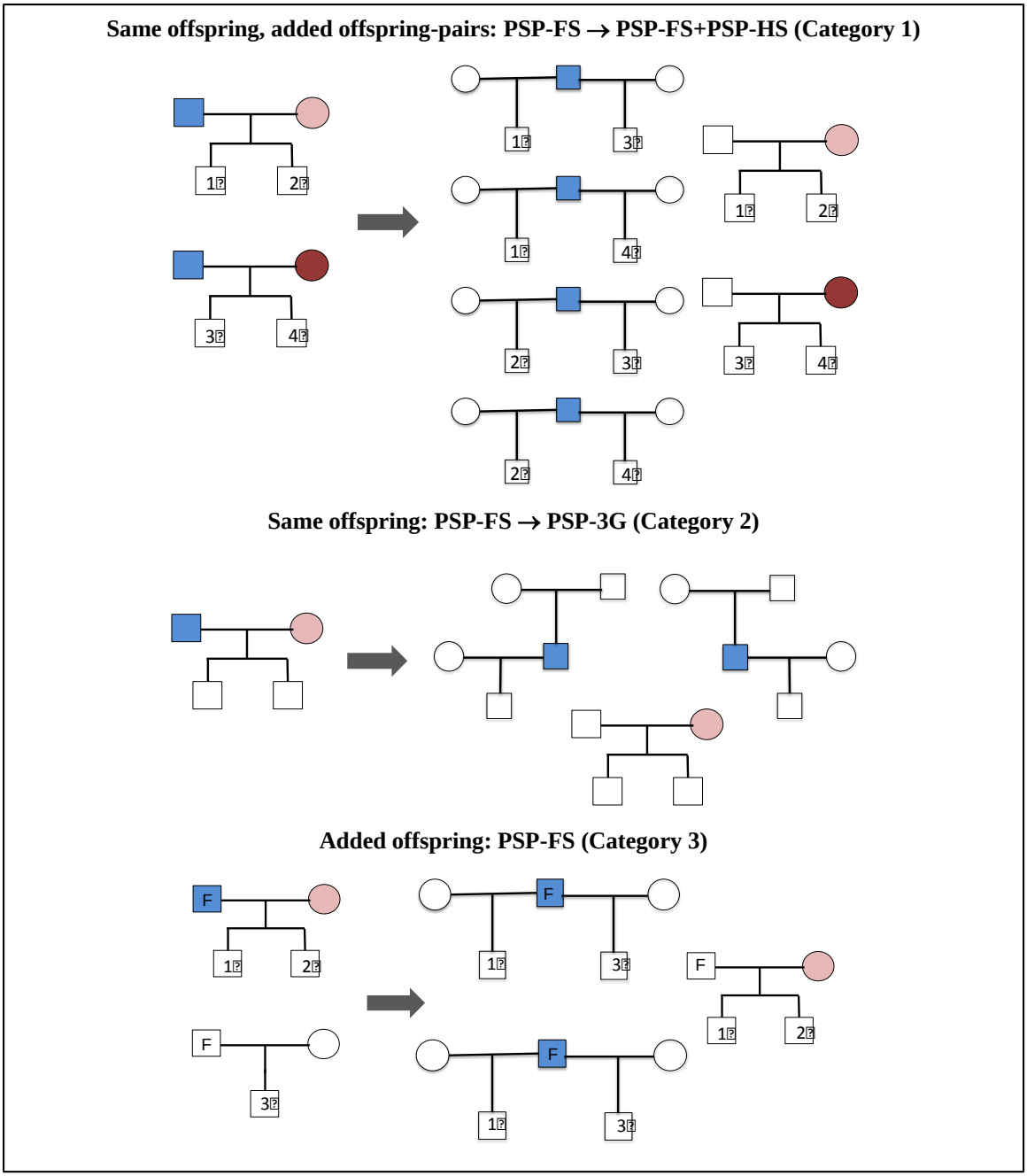


Figure 13. Changes to proband pedigree structure type by adding half-siblings and grandparents

5.1.2 Phenotype, GWAS and candidate-gene association panels

Genome wide associations were carried out on the ALL and SIBS samples in each of the CARIES and OFC study datasets. The average genome wide recombination count (**ARC**) was used

to perform genome-wide **quantitative phenotype** associations using PLINK. For each study and sample, GWASs were run separately within male and female probands. Genotype data for both studies are available for the Illumina Human610_Quadv1_B platform, which consists of approximately 600,000 SNPs. GWAS included only the autosomes.

Candidate SNP association of ARC was carried out using genotyped as well as imputed SNPs using PLINK. Imputed genotypes for these datasets were previously carried out at CIDR using 1000 genomes phase 1 for a reference panel of approximately 5.7 million genome wide variants. From these, we selected SNPs within the candidate gene regions reported by Kong et al. 2014 on their Icelandic population. Using the first SNP reported for each gene reported except the intronic variant, the SNP at or closest to its location was selected from the GWAS panel, and the latter's base-pair location obtained from the UCSC genome browser (NCBI build 36 as appropriate for these datasets). Imputed SNPs within a 500 MB region spanning each SNP were analyzed for association with ARC. Table 7 lists the genes, SNP names and physical locations, and the corresponding candidate regions analyzed.

Table 7. Candidate gene regions for association of imputed SNPs

Chromosome	Gene	SNP	SNP position (bp)	Start SNP bp	End SNP bp
1	MSH4	rs5745459	76,118,161	76,096,617	76,595,451
4	CPLX1	rs7677766	795,890	548,152	1,045,824
4	RNF212	rs4045481	1,090,625	840,980	1,340,353
5	PRDM9	rs6889665	23,532,643	23,282,784	23,781,623
14	C14orf39	rs1254319	60,903,757	60,655,514	61,152,692
14	SMEK1	rs10135595	91,925,027	91,675,267	92,174,914
14	CCNB1IP1	rs1132644	20,784,718	20,534,814	21,034,611
17	CCDC43	rs75502650	42,766,062	42,516,098	43,015,516
17	CHR17INV	rs1724424	43,779,962	43,530,550	44,029,178
20	RAD21L1	rs970084	1,221,171	972,932	1,470,380

5.2 DENTAL CARIES DATASET (CARIES)

The CARIES dataset consists of a cohort collected as part of an oral health study conducted by the Center for Oral Health in Appalachia. The study subjects were assessed for oral health and related environmental factors [17, 18].

5.2.1 Recombination scoring results for ALL and SIBS

There are 598 recombination probands and 993 genotyped offspring in the dental caries data set, belonging to 403 pedigrees. These include 23 probands with genotyped parents available for scoring recombinations, and 575 without any parents in the data. A total of 49,460 recombinations were detected in this data. Table 8 and Figure 13 below contain the comparison of probands, offspring and scored recombinations between the ALL and SIBS analysis sets.

In Table 8 the number of recombinations scored in each analysis set is broken down based on proband types in the 2 samples. The ALL sample consists of 203 male and 395 female probands vs. 194 and 304 for the SIBS respectively. The ALL samples represent increases of 4% and 30% in the number of male and female probands over the SIBS samples respectively. In the first 3 categories of Table 8, some of the PSP-FS probands are reassigned the PSP-HS proband type as a result of combining across multiple marriages. This reassignment is described in the previous section and illustrated in Figure 12. In the first category, although the number of offspring in both the samples is identical, the number of recombinations scored is different due to added full- or half-sibling pairs, as shown in Figure 12, top panel.

Table 8. Probands, offspring and recombinations in CARIES dataset

	SIBS			FULL		
	#Probands	**#Offspring**	**#Recombs**	**#Probands**	**#Offspring**	**#Recombs**
Same number of offspring	439	1,065	33,619	439	1,065	33,681
Male probands	184	452	11,578	184	452	11,578
Female probands	255	613	22,041	255	613	22,103
Same offspring with added parents	5	11	267	5	11	309
Male probands	3	7	148	3	7	167
Female probands	2	4	119	2	4	142
Added offspring	53	127	4,395	53	186	7,536
Male probands	6	16	401	6	23	577
Female probands	47	111	3,994	47	163	6,959
New 3G probands				14	21	884
Male probands				3	3	110
Female probands				11	18	774
New HS probands				87	197	7,707
Male probands				7	14	365
Female probands				80	183	7,342
Total	497	1203	38,281	598	1480	50,117
Total male	193	475	12,127	203	499	12,797
Total female	304	728	26,154	395	981	37,320

Figure 14 shows the histogram of ARCs obtained within the ALL and SIBS sets, classified by sex of the proband, along with the respective empirically fitted density curves. Female probands have a larger mean ARC than male probands, as would be expected. Distributions of ARCs do not appear to be normally distributed.

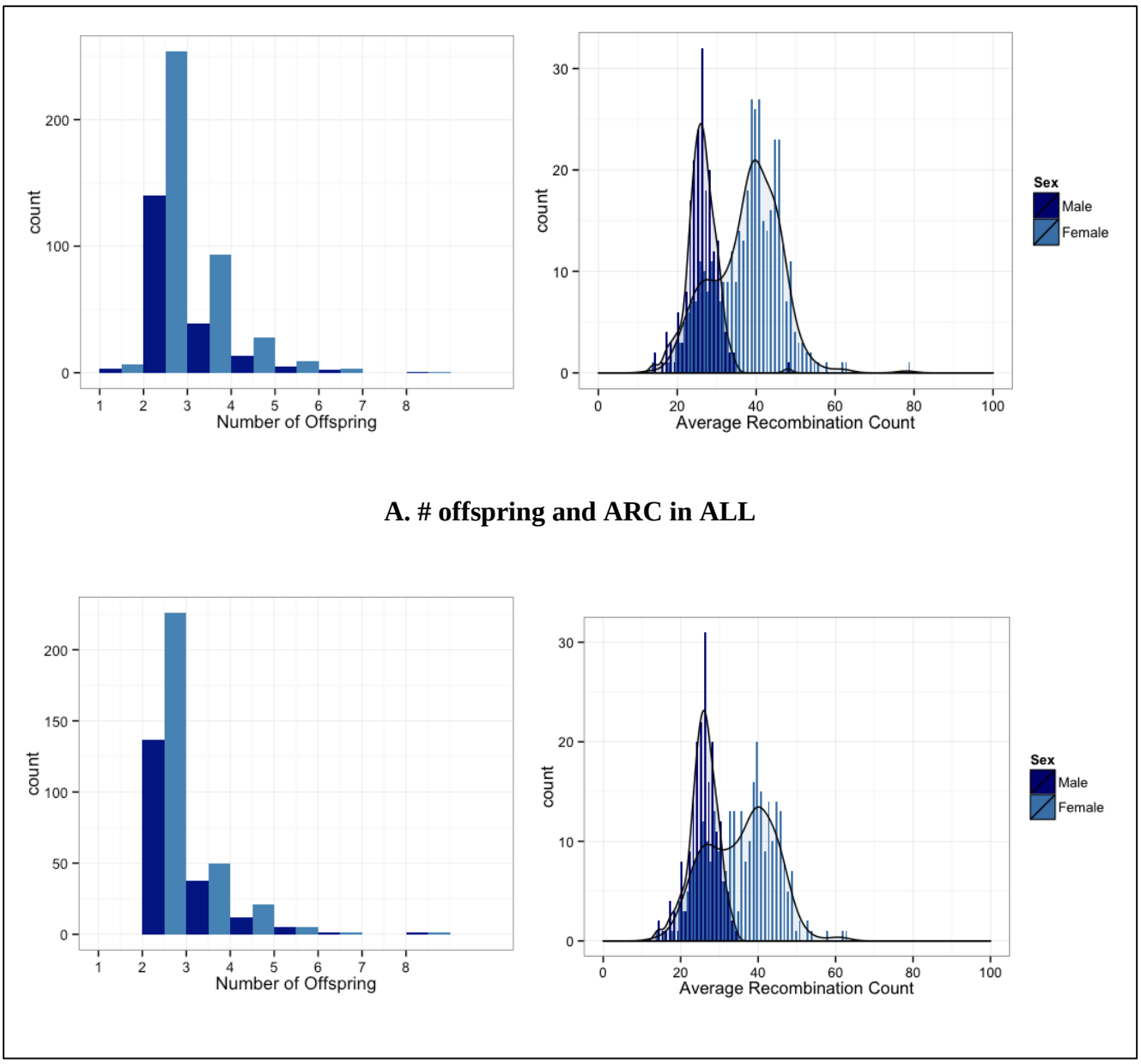


Figure 14. Distribution of number of offspring and ARC in CARIES dataset

5.2.2 Genome wide association results for genotyped SNPs

Figure 14 below shows association p-values for female and male recombination probands in the ALL and SIBS sets respectively. The red horizontal lines represent the respective Bonferroni-corrected significance threshold, approximately 10E-7.01 for both GWASs, gray horizontal lines, a p-value 10^{-5}. P-values above 10^{-5} were observed on chromosomes 1, 7, 10, 13, 17 and 20 for female

probands, and on chromosomes 2, 4, 7, 8, and 20 for male probands. Peak regions observed in males differ from those observed in females. There are no peaks in common between the male and female samples and none met the Bonferroni threshold.

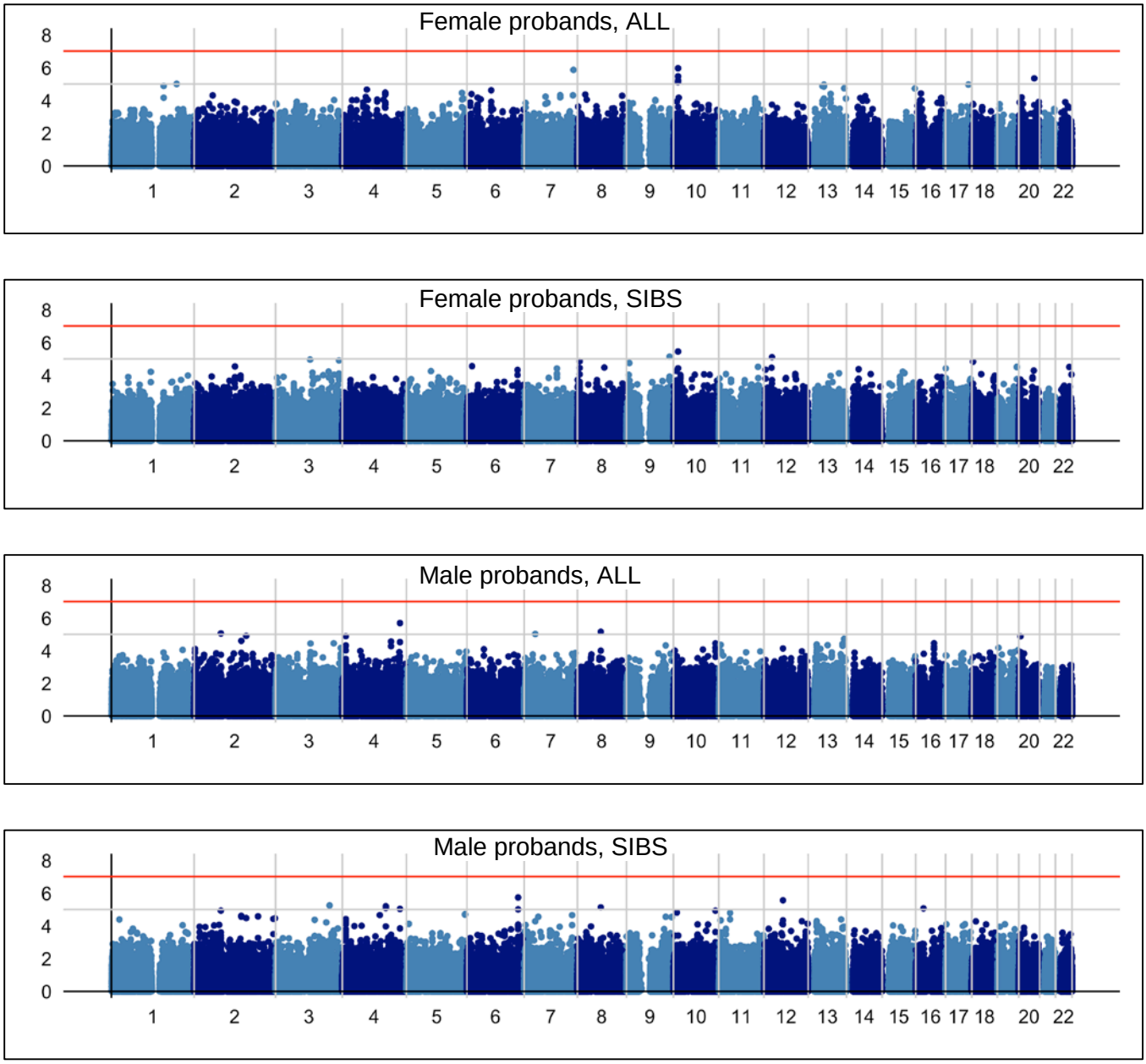


Figure 15. GWAS of male and female proband samples for CARIES dataset

5.2.3 Candidate gene association results for genotyped and imputed SNPs

The association P-values observed within candidate gene regions are presented in figure 15 for the ALL and SIBS analyses. The phenotype is the genome-wide average recombination count or ARC. Green triangles represent the SIBS sample and blue circles, the FULL sample.

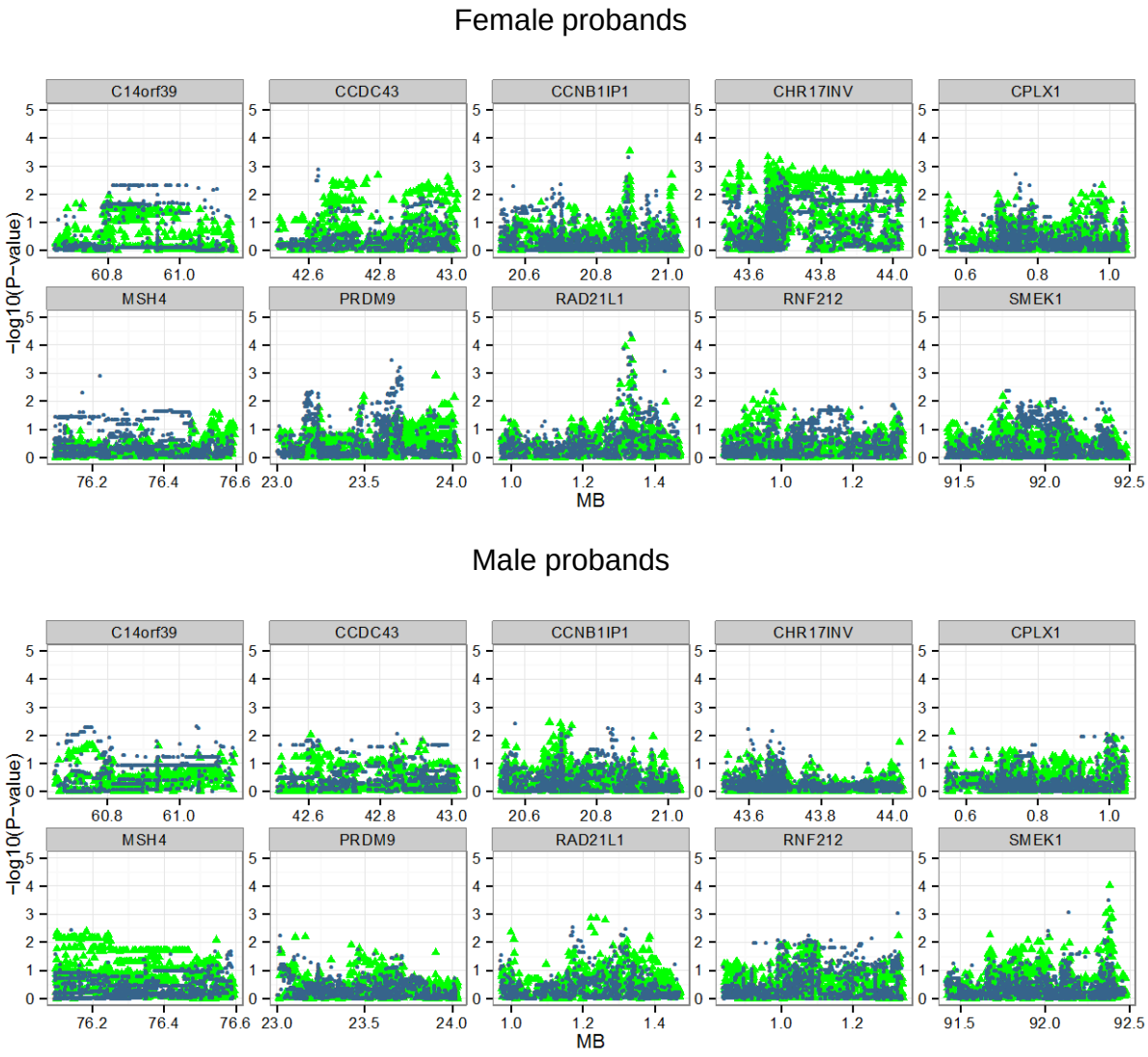


Figure 16. Comparison of ALL vs. SIBS p-values in candidate gene regions in CARIES dataset

Note: blue dots represent p-values for the ALL sample, and green triangles for the SIBS

The highest association (p-value < 10^{-4}) is observed in the *RAD21L1* gene within female probands. The FULL sample produced more significant association p-values over the SIBS sample in seven out of the ten candidate regions (*C14orf39*, *CCDC43*, *CPLX1*, *MSH4*, *PRDM9*, *RAD21L1*, and *SMEK1*) for female probands. The difference is not noticeable in the male probands. This is to be expected, since the female FULL sample is 30% larger than the SIBS sample, whereas the male samples are of almost equal sizes.

5.3 GUATEMALA ORO-FACIAL CLEFT DATASET (OFC)

This dataset is part of a multi-population study of cleft lip and palate, a birth defect. It consists of multi-generation families with non-syndromic individuals affected with cleft lip and/or cleft palate and control families without any history of cleft lip/cleft palate. In the following sections, results of recombination scoring are presented, followed by GWAS and candidate gene association results. For the purposes of GWAS, ancestry principal components available of this dataset from a separate study was used to correct for population admixture.

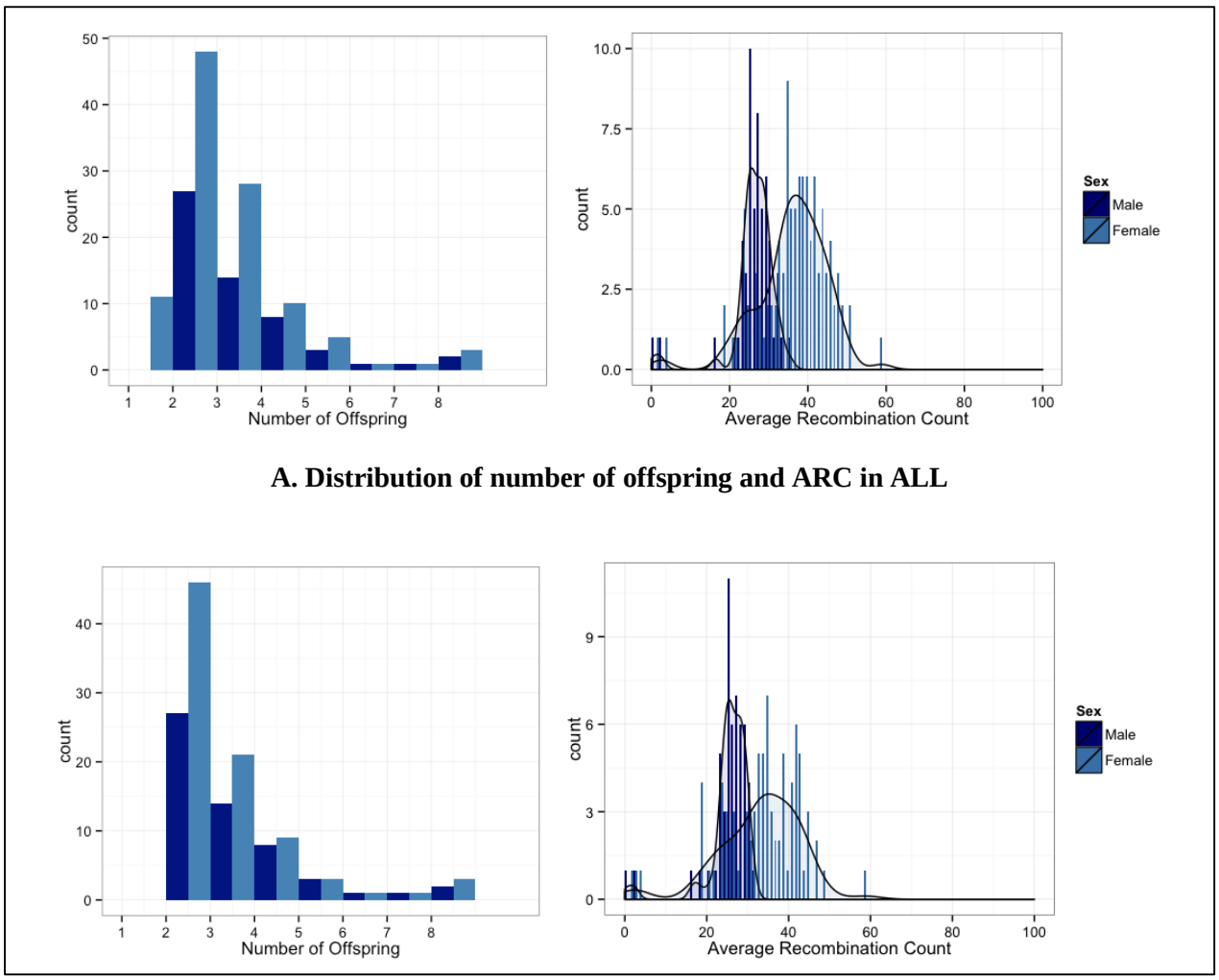
5.3.1 Recombination scoring results

In this dataset there are 163 recombination probands, including 56 male and 107 female probands, and 452 genotyped offspring in the OFC data set, belonging to 93 extended pedigrees. These include 23 probands with genotyped parents available for scoring recombinations, and 575 without any parents in the data. A total of 49,460 recombinations were detected in this data. Table 9 shows the comparison of probands, offspring and PSPs between the FULL and SIBS analysis sets, including addition of new probands and added offspring in the former. Figure 16 shows the

histogram of ARCs obtained within the FULL and SIBS samples, classified by sex of the proband and empirically fitted density curves. On the average males show fewer total genome-wide recombinations than females, as expected.

Table 9. Probands, offspring and recombinations in OFC

	SIBS			FULL		
	#Probands	**#Offspring**	**#Recombs**	**#Probands**	**#Offspring**	**#Recombs**
Same number of offspring	110	329	9,956	110	329	9,938
Male probands	49	154	3,916	49	154	3,907
Female probands	61	175	6,040	61	175	6,031
Same offspring with added parents	18	57	1,892	18	57	2,040
Male probands	6	15	370	6	15	393
Female probands	12	42	1,522	12	42	1,647
Added offspring	10	23	632	10	36	1,360
Male probands						
Female probands	10	23	650	10	36	1,360
New 3G probands				10	10	353
Male probands						
Female probands				10	10	353
New HS probands				12	26	969
Male probands				1	2	52
Female probands				11	24	917
Total	138	409	12,480	160	458	14,660
Total Male	55	169	4,286	56	171	4,352
Total Female	83	240	8,212	104	287	10,308



A. Distribution of number of offspring and ARC in ALL

Figure 17. Distribution of number of offspring and ARC in OFC dataset

5.3.2 GWAS results

There are no genome-wide significant associations due to small sample sizes. Suggestive associations are seen on chromosomes 1, 2, 7, 12, and 20 in female probands. The male proband sample showed evidence of inflation in p-values, and are therefore being investigated further. It is likely that the adjustment for population admixture was inadequate for the male probands.

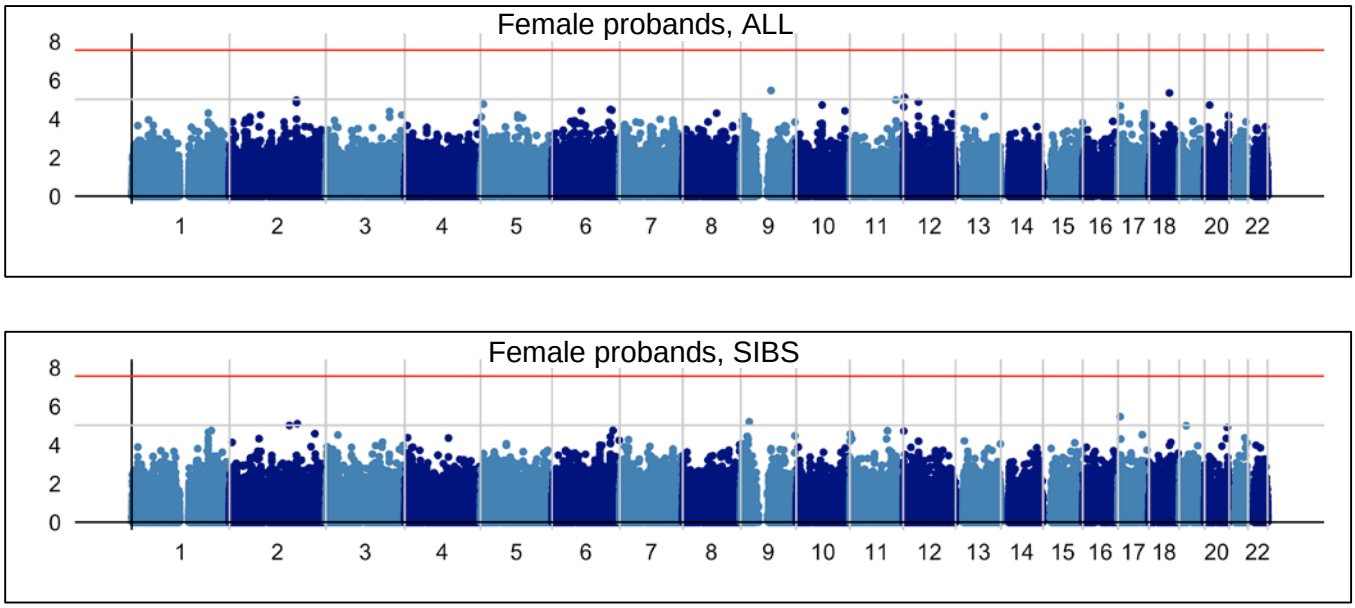


Figure 18. GWAS of FULL and SIBS samples in OFC data set

5.3.3 Candidate gene regions using imputed SNP genotype data

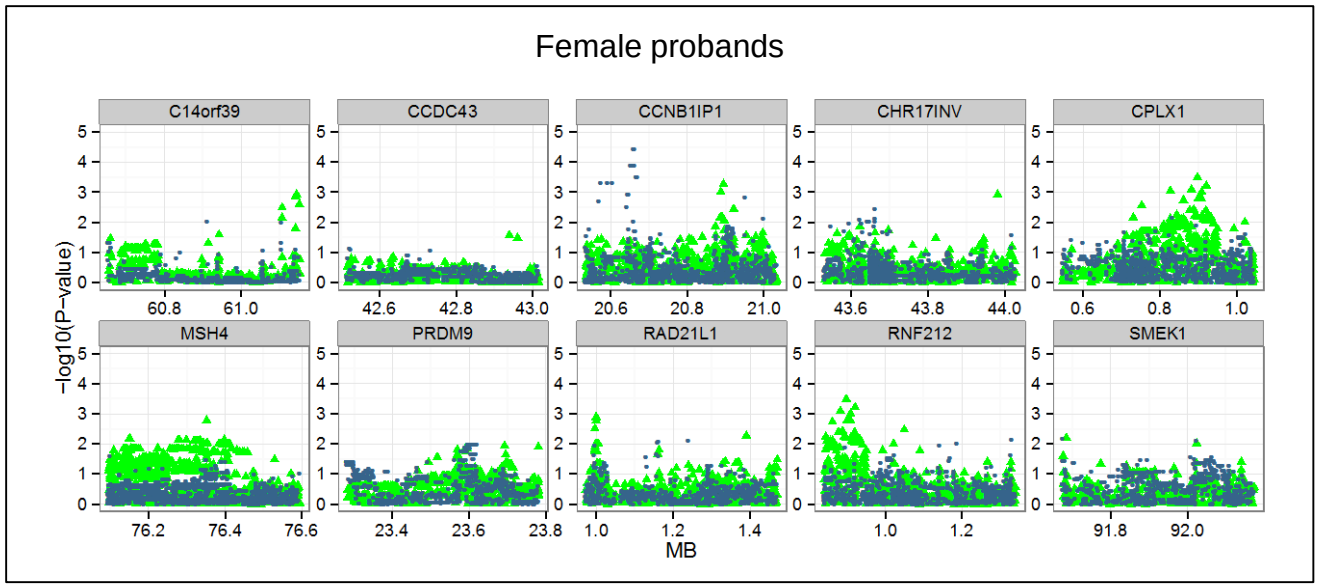


Figure 19. Comparison of ALL vs. SIBS p-values in candidate gene regions in OFC data set

Note: blue dots represent p-values for ALL, and green triangles for SIBS

The highest association (p-value < 10^{-4}) is observed in the *CCNB1IP1* gene within female probands for the FULL sample. There was no noticeable improvement in association p-values with the added probands in the female sample.

5.4 COMPARISON OF MBFAM TO EXISITNG SNP-STREAK

In order to assess the accuracy of recombination scoring, calls made by MBFam were compared to those reported by Ms. Begum in her dissertation, scored on the CARIES data set. In Ms. Begum's dissertation, genome-wide average recombination counts (ARC) were created based on recombination intervals reported by the sibship-based method developed by Chowdhury et al. and Coop et al.[3, 12]. For this comparison, only those probands present in Ms. Begum's analysis who were scored using the same number of offspring in both analyses are selected. Within MBFam, recombinations for the selected probands are scored using only the offspring who are full-siblings. Figure 19 shows average recombination counts calculated by our method (ARC_{MBFam}) plotted against those calculated by the method implemented by Coop et al (ARC_{Coop}). The 45° line is drawn in for reference.

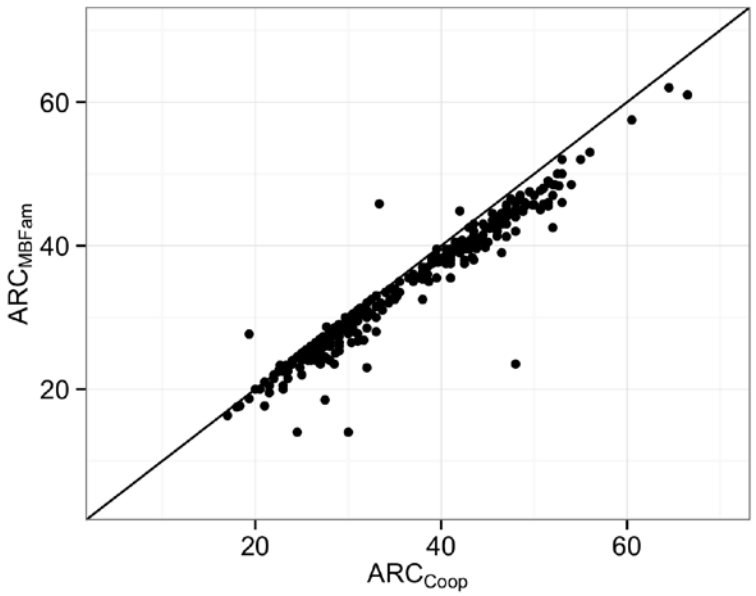


Figure 20. Comparison of ARC between COOP and MBFAM

In general, our MBFam recombination counts are smaller than those produced by Coop et al.'s method, although, in some cases, MBFam calls a larger number of recombinations. Plausible reasons for this difference are:

1) Inclusion of missing genotype data: Our method attempts to call IBD statuses for SNPs for which one parent may be untyped, whereas such SNPs are excluded from IBD calculation in Coop et al.'s method. To determine unique recombination intervals, my method also allows approximate alignment of intervals, thereby affecting the number of recombinations scored.
2) Elimination of proximate double recombinations: The limitations placed by the respective programs on required segment lengths in order to detect and remove proximate double recombinations are different. A more stringent filtering mechanism was used in our analysis; middle segments were deemed acceptable only if they contained at least 10 informative SNPs,

whereas the method used by Coop et al. accepted segments that contained at least 5 informative SNPs, irrespective of the physical length of the segment.

6.0 SUMMARY, CONCLUSION AND FUTURE DIRECTION

6.1 SUMMARY OF DISSERTATION PROJECTS

This aim of this dissertation work was to develop an enhanced method to detect meiotic recombination breakpoints, in order to make best use densely genotyped SNP data gathered on general pedigrees, with the intention of increasing sample sizes available for a recombination GWAS. To achieve this goal, we divided our work into three sub-aims. In the first sub-aim, we developed methodology to extend the current SNP-streak based recombination scoring paradigm that is limited to completely genotyped full sibships. The new method was developed to handle new pedigree structures, namely half-siblings and grandparents, and allow for missing genotypes within the new pedigree structures, as well as full-sibships. The second sub-aim consisted of software implementation, and in the third sub-aim, we applied the new method on two pedigree-based genome-wide association datasets.

In the first part of the project, we developed the methodology needed to extract three types of recombination probands proband pedigree substructures (full-sibpair, half-sibpair, and three-generation) for scoring recombinations, derived Mendelian inheritance-based rules to score identity-by-descent (IBD) on half-sibling pedigree structures as well as grandparent-of-origin (GOO) statuses for offspring three-generation pedigree structures, and finally, methods to detect recombination locations and recombinant offspring. In chapters 1 and 2, I introduce the concepts of study of recombination and describe prior work. This current method is described in detail in chapter 3.

The second part of the project involved implementation of the new methodology as a software program. This program is implements classes and methods in Python for input GWAS and

SNP annotation data in PLINK-format, internal representation of pedigree structures and individuals' genotypes, IBD and GOO status rules for the three pedigree substructures, scans of IBD and GOO statuses, and recombination switches. Limited functionality is also implemented to create recombination phenotypes. Output is provided in the form of recombination locations by proband, offspring and chromosome. Diagnostics, such as Mendelian inconsistencies and double recombinations, are provided, and plots of status scans, combined with recombination locations can be created using R functions provided for this purpose. Details of the architecture of the MBFam scoring software as outlined above are presented in chapter 4.

The third part of the dissertation work consists of applying our new method to two family-based GWAS datasets, one ascertained for dental caries consisting of subjects from the Appalachian region, and the second, a study of oro-facial clefting in subjects from rural Guatemala. The two sets of study data were analyzed for recombination breakpoints followed by genome-wide association analyses of the average recombination count (ARC) phenotype. Association results were compared to those reported by previous studies. We compared two samples within each study, those scored using only full siblings, and those scored using the other two pedigree structures as well. Although both studies yield modest sample sizes of recombination probands, there was some evidence of improved association results using the enhanced structures. Real data analyses are described in chapter 5.

6.2 STRENGTHS AND LIMITATIONS

In this section, I discuss the strengths and limitations of our method of scoring recombinations, as well as the characteristics of data that are desirable for this method to perform well.

6.2.1 Strengths

Our method of SNP-streak based recombination scoring is able to accurately and quickly detect recombinations on genotype data. It makes use of inexpensive rule-based exact computations on informative SNPs, followed by the application of heuristics on this subset to identify recombinations. It has a built-in framework to tailor the heuristics to the observed data, and is, therefore, more robust to variation across data sets. Finally, it makes maximal use available pedigree structures, and can handle missing genotypes.

6.2.2 Limitations

Our method assumes that SNPs are densely spaced, and also that the distribution of informative vs. uninformative or ungenotyped SNP locations are fairly uniform across the chromosome for genotyped probands and their PSPs. If these conditions are not met, MBFam may fail to call recombinations correctly. With widely spaced SNPs, some recombination events may go undetected, and with non-random patterns of informativeness, the resulting calls may be biased for certain regions. However, with the modern genotyping panels, and with outbred populations, this is not expected to happen.

A second limitation arises where a status scan contains long stretches without informative statuses. Figure 21 below shows a scan belonging to a full-sibling pair with only one parent genotyped. With only one parent genotyped, genotype configurations indicating the “S” (red) status are rare. However, the absence of blue points over most of the chromosome indicates that it is highly likely that there is at least one “D” segment between the two “S” segments and hence, at least 2 more

switch intervals, besides the one evident interval at the end of the chromosome. Currently, MBFam is unable to detect recombinations from such scans, as it counts a single blue segment spanning most of the chromosome. There are several sibships with only a single parent genotyped in both datasets analyzed, and their total recombination counts may not be accurate due to some recombination intervals not being detected. This inaccuracy may explain why the larger samples (ALL) did not show improvement in power to detect association. We plan to develop heuristics for calling recombination intervals using scans such as these.

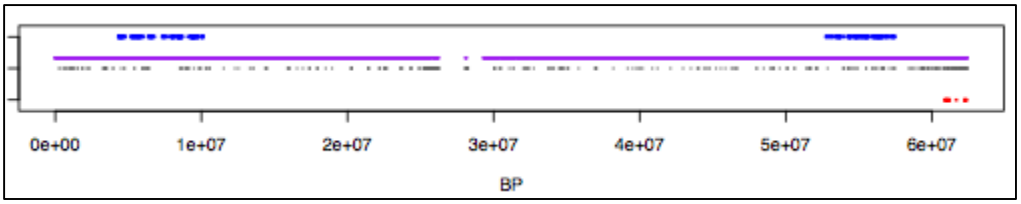


Figure 21. Status scan with a large region of uninformative statuses
Note: "S" statuses colored blue, and red statuses, "D" statuses colored red

6.2.3 Special case: Trisomy 21 dataset

Although MBFam was designed to analyze dense genome-wide data, we applied our method to a trisomy data set, which was genotyped on a linkage panel approximately 6,000 SNPs. This study is described briefly below.

Sample: This study involves the families of affected offspring with trisomy 21. The family units analyzed included the normal offspring, their parents and maternal grandparents. Genotyping was performed at the Center for Inherited Disease Research (CIDR) using the Golden Gate linkage panel consisting of 6,056 SNP markers spaced on average 0.63 cM apart across the genome.

Recombination calling: Recombination calling was performed using only 3-generation PSPs, therefore, all switch intervals detected were recorded as recombination intervals.

Comparison with Merlin-generated haplotypes: Merlin's haplotyping option was run to generate segments transmitted by the maternal grand-parents to the offspring, and the boundary locations of segments selected as the switch intervals, and the number of intervals on chromosome 1 compared between the two methods for each proband. There were 205 probands with recombination calls for both methods, of these, 191 were in agreement. The differences between the 2 sets were extra recombinations at the ends of the chromosomes in one case or the other.

In this dataset, the presence of three-generation pedigrees made it possible to score recombinations even though the SNP spacing was very sparse.

6.3 CONCLUSIONS

In conclusion, our method increases the sample-size for GWAS of recombination, by allowing the detection of recombinations on new pedigree structures and dense SNP panels that are not handled by existing methods. The candidate gene associations on the CARIES dataset shows evidence for improved detection of recombination genes. We have implemented our method as an easy-to-use and platform independent software package, which has been thoroughly tested on two real datasets. The real data applications did not conclusively show improved power for association in the form of more significant p-values due to the reasons discussed above: (a) small gains in sample sizes and (b) unusual status scans that need special handling to correctly detect switch intervals .

6.4 FUTURE DIRECTION

There are several directions in which MBFam can be improved in the future. First, the accuracy of recombination detection can be improved by combining recombination information from all the possible PSPs of each proband. Currently probands are scored on the basis of genotyped parents (if available) as 3-gen PSPs, for each offspring separately, or using their offspring, if parents are not available. For a proband with both parents and multiple offspring genotyped, combining across 3-gen and sibship-PSPs may improve the accuracy of recombination calls. Secondly, for large chromosome segments that are not informative for recombination, a likelihood model can be adopted to infer a minimum number of crossovers, based on known recombination rates. The software itself can be made more functional by incorporating the creation of a wider array of recombination phenotypes, such as counts/percentages in hotspot vs. non-hotspot regions.

Our application was limited to two real study datasets. A thorough test of performance needs to include simulated data, if which the results are known *a prioi*. The design of simulated datasets appropriate for evaluation of recombination call accuracy presents another area of investigation. While the datasets analyzed were comparatively small, and the association mostly inconclusive, we plan to apply our method to a larger, multi-ethnic study consisting of multi-generational pedigrees (OFC reference). This will allow us to investigate whether genes influencing recombination differ by population, but also to detect if patterns of individual-level recombinations differ by population.

APPENDIX

MBFam PROGRAM CODE

BIBLIOGRAPHY

1. Hassold, T. and P. Hunt, *To err (meiotically) is human: the genesis of human aneuploidy.* Nat Rev Genet, 2001. **2**(4): p. 280-91.
2. Lamb, N.E., S.L. Sherman, and T.J. Hassold, *Effect of meiotic recombination on the production of aneuploid gametes in humans.* Cytogenet Genome Res, 2005. **111**(3-4): p. 250-5.
3. Coop, G., et al., *High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans.* Science, 2008. **319**(5868): p. 1395-8.
4. Kong, A., et al., *Fine-scale recombination rate differences between sexes, populations and individuals.* Nature, 2010. **467**(7319): p. 1099-103.
5. Sherman, S.L., N.E. Lamb, and E. Feingold, *Relationship of recombination patterns and maternal age among non-disjoined chromosomes 21.* Biochem Soc Trans, 2006. **34**(Pt 4): p. 578-80.
6. Baudat, F., et al., *PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice.* Science, 2010. **327**(5967): p. 836-40.
7. Middlebrooks, C.D., et al., *Evidence for dysregulation of genome-wide recombination in oocytes with nondisjoined chromosomes 21.* Hum Mol Genet, 2014. **23**(2): p. 408-17.
8. Lamb, N.E., et al., *Association between maternal age and meiotic recombination for trisomy 21.* Am J Hum Genet, 2005. **76**(1): p. 91-9.
9. Lander, E.S. and P. Green, *Construction of multilocus genetic linkage maps in humans.* Proc Natl Acad Sci U S A, 1987. **84**(8): p. 2363-7.
10. Terwilliger, J. and J. Ott, *Handbook of Human Genetic Linkage*. 1994, Baltimore and London: The Johns Hopkins University Press.
11. Kong, A., et al., *Common and low-frequency variants associated with genome-wide recombination rate.* Nat Genet, 2014. **46**(1): p. 11-6.
12. Chowdhury, R., et al., *Genetic analysis of variation in human meiotic recombination.* PLoS Genet, 2009. **5**(9): p. e1000648.
13. Kong, A., et al., *Sequence variants in the RNF212 gene associate with genome-wide recombination rate.* Science, 2008. **319**(5868): p. 1398-401.
14. Fledel-Alon, A., et al., *Variation in human recombination rates and its genetic determinants.* PLoS One, 2011. **6**(6): p. e20321.
15. Begum, F., *GWAS Meta-Analysis: Methodology and Application to Human Meiotic Recombination*. 2013, University of Pittsburgh.
16. Purcell, S. *PLINK v1.07*. Available from: http://pngu.mgh.harvard.edu/purcell/plink/.

17. Polk, D.E., et al., *Study protocol of the Center for Oral Health Research in Appalachia (COHRA) etiology study.* BMC Oral Health, 2008. **8**: p. 18.

18. Wang, X., et al., *Genes and their effects on dental caries may differ between primary and permanent dentitions.* Caries Res, 2010. **44**(3): p. 277-84.

19. Begum, F., et al., *Genome-wide association study of meiotic recombination phenotypes.* Submitted, 2016.

20. Sun, L., K. Wilder, and M.S. McPeek, *Enhanced pedigree error detection.* Hum Hered, 2002. **54**(2): p. 99-110.

21. Manichaikul, A., et al., *Robust relationship inference in genome-wide association studies.* Bioinformatics, 2010. **26**(22): p. 2867-73.

22. Zeng, Z., et al., *A Pipeline for Classifying Relationships Using Dense SNP/SNV Data and Putative Pedigree Information.* Genet Epidemiol, 2016. **40**(2): p. 161-71.

23. Abecasis, G.R., et al., *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees.* Nat Genet, 2002. **30**(1): p. 97-101.

24. Lange, K., et al., *Mendel: the Swiss army knife of genetic analysis programs.* Bioinformatics, 2013. **29**(12): p. 1568-70.

25. Sobel, E., H. Sengul, and D.E. Weeks, *Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees.* Hum Hered, 2001. **52**(3): p. 121-31.

26. Fishelson, M., N. Dovgolevsky, and D. Geiger, *Maximum likelihood haplotyping for general pedigrees.* Hum Hered, 2005. **59**(1): p. 41-60.

27. O'Connell, J.R., *Zero-recombinant haplotyping: applications to fine mapping using SNPs.* Genet Epidemiol, 2000. **19 Suppl 1**: p. S64-70.

28. Li, J. and T. Jiang, *Efficient inference of haplotypes from genotypes on a pedigree.* J Bioinform Comput Biol, 2003. **1**(1): p. 41-69.

29. Williams, A.L., et al., *Rapid haplotype inference for nuclear families.* Genome Biol, 2010. **11**(10): p. R108.

30. Kong, A., et al., *Detection of sharing by descent, long-range phasing and haplotype imputation.* Nat Genet, 2008. **40**(9): p. 1068-75.

31. Hickey, J.M., et al., *A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes.* Genet Sel Evol, 2011. **43**(1): p. 12.

32. Myers, S., et al., *Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination.* Science, 2010. **327**(5967): p. 876-9.

33. Myers, S., et al., *A common sequence motif associated with recombination hot spots and genome instability in humans.* Nat Genet, 2008. **40**(9): p. 1124-9.