Statistical Applications in Genetics and Molecular Biology

Volume 7, Issue 1

2008

Article 16

A SNP Streak Model for the Identification of Genetic Regions Identical-by-descent

Gregory Leibon^{*}

Daniel N. Rockmore[†]

Martin R. Pollak[‡]

^{*}Dartmouth College, gleibon@gmail.com

[†]Dartmouth College, dnrockmore@gmail.com

[‡]Harvard Medical School and Brigham and Women's Hospital, mpollak@rics.bwh.harvard.edu

A SNP Streak Model for the Identification of Genetic Regions Identical-by-descent*

Gregory Leibon, Daniel N. Rockmore, and Martin R. Pollak

Abstract

The availability of very dense genetic maps is changing in a fundamental way the methods used to identify the genetic basis of both rare and common inherited traits. The ability to directly compare the genomes of two related individuals and quickly identify those regions that are inherited identical-by-descent (IBD) from a recent common ancestor would be of utility in a wide range of genetic mapping methods. Here, we describe a simple method for using dense SNP maps to identify regions of the genome likely to be inherited IBD by family members. This method is based on identifying obligate recombination events and examining the pattern of distribution of such events along the genetic map. Specifically, we use the length of a consecutive set of biallelic markers that have a high probability of having avoided such obligate recombination events. This “SNP streak” is derived from subsets of samples within a pedigree and allows us to make statistical inferences about the ancestry of the region(s) containing stretches of markers with these properties. We show that the use of subsets of more than two samples has the advantage of identifying shorter shared subsegments as significant. This mitigates the effects of errors in SNP calls. We provide specific examples of microarray-based SNP data, using a family with a complex pedigree and with a rare form of inherited kidney disease, to illustrate this approach.

KEYWORDS: genetics, SNPs, identity-by-descent, haplotype, FSGS

*G.L. and D.R. supported in part by NIH grant GM075310. M.P. supported in part by NIH grant DK54931. Acknowledgments: The authors thank the referees for their close reading of the manuscript and in particular, the referee who brought to our attention the nearly simultaneous work of Thomas, et al. (2008) and whose comments of comparison are included in the subsection “Relation to previous work.”

1 Introduction

Family-based methods for the study of inherited disease, whether linkage analyses in extended pedigrees with Mendelian phenotypes, or affected relative pair methods to identify loci responsible for complex phenotypes, ultimately seek to identify regions of the genome shared identical-by-descent (IBD) between specific individuals. For example, in a typical linkage analysis in a pedigree with a dominant, rare phenotype, the investigator aims to find a haplotype shared in affected family members by virtue of inheritance from a recent common founder. In a large pedigree with 100% penetrance of a rare autosomal dominant trait and low phenocopy number, identifying a disease-associated region by linkage analysis is equivalent to identifying a region shared IBD by all affected individuals (but no unaffected individual). The availability of extremely dense samplings of a genome, obtained from maps of informative single nucleotide polymorphisms (SNPs) (e.g. Affymetrix SNPChips, Illumina BeadChips) allow for new analyses that exploit the high density (see [Matsuzaki, et al. (2004)-1] and [Matsuzaki, et al. (2004)-2]).

In this paper we show that given SNP data from a set of individuals in a complex pedigree, the identification of “streaks,” defined as sequences of consecutive loci in the SNP map in which no two samples have genotypes AA and BB, respectively, provides a simple but powerful heuristic for determining candidate IBD regions. Note that a streak is not “broken” if one genotype is AA and the other AB (or BB and AB or AB and AB). The underlying principle we employ is simple: if two related individuals share one contiguous chromosomal segment by inheritance from a recent common ancestor, then within such a region, we will, by definition, see no recombination events. If the SNP map used for genotyping is sufficiently dense and polymorphic, it is highly likely that such recombination events, if present, will break a “streak” of SNPs. In other words, if a region is not inherited IBD from a common founder in the individuals that we are comparing, we should be able to identify obligate recombination events within this region. Specifically, within such a region, we should find markers where the two individuals have no alleles in common (e.g., are genotype AA and BB). We illustrate the technique on a complex pedigree obtained from a family whose members share a rare form of kidney disease.

The basic idea that we employ is simple - we consider all possible subsets of samples and look for SNP streaks common to all samples in the subset whose length exceeds some threshold (that is set to avoid false positives). Our approach is straightforward to apply, and we derive a method for assessing the statistical significance of long streaks.

Relation to previous work

Our consideration of SNP streaks among subsets generalizes that of the very interesting work of Miyazawa, et al. [Miyazawa, et al. (2007)] which considers SNP streaks of homozygous loci (the so-called “homozygosity haplotype”) in the sample genomes shared by pairs of samples. Our work differs in two ways. First, by not restricting to the homozygosity haplotype (HH) we are able to use more of the full information contained in the SNP data. Second, by using subsets of more than two samples we are able to exploit the fact that there is a lower threshold for coincidental sharing among larger subsets of samples, a threshold that diminishes rapidly as the number of samples (i.e., the subset size) increases. In other words, a pairwise comparison will always require a longer SNP streak for significance than a comparison among several - more than two - samples. Thus, if one processes a collection of SNP samples two at a time, broken streaks simply due to SNP errors - which are more likely in longer sequences - will cause one to eliminate samples and regions that otherwise would have been detected in a search for shorter streaks that would be indicated by the use of larger subsets. Notice, also that this sort of reasoning is advantageous even when there are no errors, since a sample may be considered IBD in a short region that is consistent with multiple sharing, but ignored (not long enough to be considered likely to be IBD) if considered only when compared against a single sample. On the flip side, the higher threshold required for pairs can also be a disadvantage if the error rate on the SNP data is high. In this case an error could induce a false negative via the introduction of a spurious break in an actual SNP streak.

In the course of the review of this paper, the paper [Thomas, et al. (2008)] appeared in print. As has been pointed out by the referee, the streaks statistic introduced therein is identical to that contained herein. That said, there are differences between the two papers and we quote here from the (very thorough) referee report (suitably modified to incorporate our bibliographic style): “The innovation in this current paper is to extend the methods of [Miyazawa, et al. (2007)] for homozygous sharing to evaluating statistical significance for heterozygous sharing. This has the advantage of being applicable when the pedigree structure is unsure, or has paths back to multiple common ancestors. Although the simulation method in [Thomas, et al. (2008)] is general, the theoretical results they use apply only when there is a single common ancestor or common ancestral pair. The works of [Donnelly (1983)] and [Cannings (2003)] apply to arbitrary pedigrees, but concern general modeling issues and don’t address the P-value calculations directly (although they could be developed to do so). None of [Donnelly (1983)], [Thomas, et al. (1994)], [Cannings (2003)] or [Thomas, et al. (2008)] allow for uncertainty in the pedigree structure. The current paper also makes more extensive use of sharing among sub-

sets of the cases.”

Motivation

The motivation in considering these SNP streaks is twofold. On the one hand, it has been pointed out elsewhere that the identification of regions IBD from common ancestors in a pedigree could be incorporated into any of a number of parametric or non-parametric methods to help determine candidate disease regions [Thomas, et al. (2008)]. Our primary motivation however, derives from the challenges inherent in performing linkage analyses in large and complex pedigrees. In multigenerational families, even those harboring rare gene variants with large phenotypic effects, linkage analysis can turn into a very complicated and lengthy computational problem whose performance is very sensitive to the errors that can occur in inheritance model, genetic map details, and phenotyping. In addition, accurate multipoint linkage analysis requires assumptions regarding recombination frequencies and linkage disequilibrium between neighboring markers.

Streak analysis does not suffer from any of these systematic difficulties. At the same time, it is still as useful as traditional methods. For example, in the case of a family with several members affected with a rare inherited disease, the identification of every genetic region inherited IBD from a common ancestor in each of the affected individuals would pick out the same candidate regions as an “affecteds only” genome-wide linkage analysis. Moreover, the identification of regions inherited IBD possesses a certain intuitive simplicity, not found in linkage analysis. For example, if only a few regions of the genome are shared IBD by all affected members of a kindred, then all such regions should be considered “candidate” regions. If we see that a region is shared IBD by all but, say, one affected family member, then we gain a sense of how unlikely it is that such a region will harbor the genetic variant of interest: such a region is one misdiagnosis or phenotype away from being IBD in all affected members.

2 Results and Discussion

Method

The essential idea we employ is as follows. Assume that the set of SNPs genotyped in two related individuals (denoted 1 and 2) is reasonably dense and that most of the SNPs in this set are heterozygous in most populations. In this situation, if a genomic region is inherited IBD from a recent common ancestor by all members of some

Figure 1: Illustration of region (marked M) inherited IBD in 4 DNA samples. Streak-breaking SNPs are indicated by italicized letters.

sub-sample and no de novo mutations have occurred that have altered SNP alleles, this region will contain no SNPs such that individual 1 from this sub-sample has genotype AA while individual 2 from this sub-sample has genotype BB. A (maximal) contiguous sequence of such SNPs is called a “streak.” Figure 1 illustrates the situation in which we have four samples that possess a genomic region IBD from a common ancestor. Conversely, a sufficiently large region that is not inherited IBD from a common ancestor should contain SNPs such that individual 1 is genotype AA and individual 2 is genotype BB. Once again, we note that for us, a streak does not mean an exact match of genotypes across samples, and across a consecutive set of loci in the haplotype, but rather that at no locus do we observe AA in one genome and BB in another. I.e., we permit heterozygosity, but not unequal homozygosity.

Notice that concluding a region is IBD with reasonable certainty depends on the streak being *sufficiently long*. Thus, if the pedigree tree(s) through which the allele descends is too large, we will be unlikely to detect such a streak, while if this tree is sufficiently small we will detect it. To back up this intuition, we need to make rigorous a choice of a *threshold length* or *cut-off value*, such that any streak whose length exceeds the threshold should be viewed as a likely candidate for housing an IBD region. The threshold length depends on several parameters, whose definition and/or derivation we give below.

As we have remarked, we see our work as a generalization of homozygosity haplotype approach introduced by Miyazawa et al. [Miyazawa, et al. (2007)]. For ease of comparison as well as understanding, we will now proceed to adapt the terminology and methodology of [Miyazawa, et al. (2007)] from the setting of pairwise comparisons of homozygous loci to arbitrary subsets of potentially heterozygous loci.

Definition of terms

- *Compatible locus* – We say that a genome locus is *compatible* (with respect to some collection of samples) if we do not have AA and BB at that locus for any two samples in the collection.
- *SNP streak* – We say that a consecutive set of SNPs from a collection of samples form a (*SNP*) *streak* if every locus in the sequence is compatible for the collection of samples.
- *IBD region* – As in [Miyazawa, et al. (2007)], we call a region that is identical by descent relative to a recent common ancestor an *IBD region*.
- *compatible haplotype region* – Continuing in this vein, we refer to a candidate IBD region that comes from a SNP streak as a *region with a conserved compatible haplotype* or *compatible haplotype region*. (This adapts the notion of *RCHH* (region with a conserved homozygosity haplotype) in [Miyazawa, et al. (2007)]).

SNP frequencies

Let p denote the average frequency of the most common (major) SNP allele, as measured from the data. (This is to be compared with the use in [Miyazawa, et al. (2007)] of the average frequency of a major allele, $\hat{F}_{\text{major allele}}$, as derived from the Affymetrix look-up table.) In our data $p \approx 0.85$. Let $q = 1 - p$.

Let M denote the number of samples considered. Let p_M denote the probability that for a given SNP, the M samples are consistent by chance, assuming (as in [Miyazawa, et al. (2007)]) that the production of SNPs occurs independently in samples and with the average empirically determined frequency p . Assuming this, we have

$$p_M = (p^2 + 2pq)^M + (q^2 + 2pq)^M - (2pq)^M.$$

Let $q_M = 1 - p_M$.

Let N_b denote the expected total number of SNPs in a sample of size M which by chance will be inconsistent. Using the above assumptions $N_b = Sq_M$ where S is the number of SNPs.¹ (Compare N_b with $N_{\text{mismatched pairs}}$ in [Miyazawa, et al. (2007)].)

Let N_c denote the critical length (cut-off) in base pairs for a SNP streak to be labeled as a compatible haplotype. Let c denote the critical length in Morgans required for a labeling of compatible haplotype. As in [Miyazawa, et al. (2007)], we use the approximation $c = \frac{LN_c}{S}$ where $L \approx 34$ (i.e., L is the Morgan length of the autosome).

Notice, these assumptions are very rough. In the complex pedigrees we have encountered, the pedigree itself is only an approximation. We have found that the variance of the P-values described below is largely accounted for by pedigree variance, and when a pedigree is believed to be rough, the above assumptions are reasonable for approximation. If the pedigree itself is well understood one can sharpen these assumptions and compute more accurate P-values as described in [Thomas, et al. (2008)] (see also [Donnelly (1983)], [Cannings (2003)], and [Thomas, et al. (1994)]). Moreover, using the tools in [Thomas, et al. (2008)], it is possible to account for the broad range of heterozygosity and the non-uniform distribution of SNPs through the genome. A careful evaluation of the most useful degree of approximation would require assigning a prior (in the Bayesian sense - see e.g., [Duda, et al. (2000)]) on the space of pedigrees, an important task which to our knowledge has not been dealt with in the literature.

Pedigree and choice of samples

We now provide a concrete example of the application of this method to a real pedigree. We are working with a set of samples from a large family in which multiple members have focal segmental glomerulosclerosis (FSGS) and/or end-stage kidney disease (ESKD). A rough pedigree inferred from family interviews and containing all the samples is shown in Figure 2. The pedigree shown in Figure 2 is a best guess based on the interviews and known records.

Let T denote the total number of samples considered. In our primary test case $T = 17$. These were the family members in the pedigree in Figure 2 with labels

$$\left\{ \begin{array}{ccccc} 4001, & 75, & 76, & 2034, & 2047, \\ 2077, & 40, & 74, & 91, & 2028, \\ 111, & 2027, & 2037, & 2040, & \\ 2050, & 3057, & 67 & & \end{array} \right\}.$$

Let M denote the number of samples in a specified subset of the T total samples.

¹For our pedigree, genotyped with Affymetric 100K SNP chips, $S \approx 113000$.

Figure 2: Inferred pedigree for FSGS-affected family. Black pedigree symbols denote individuals with definite disease. Checkered symbols indicate individuals with probable disease.

For a given subset of samples of size M , let B_M denote the number of non-founders in a family tree in which the M samples are the only terminal descendants. Notice this depends on the pedigree and the samples. Our $T = 17$ samples were chosen to minimize the variance of B_M with regard to the choice of M samples in our pedigree.² For our $T = 17$ samples, for a tree embedded in our pedigree, we can give a rough estimate of B_M by letting $B_M = 3M$ for any choice of subset. Looking at our likely pedigree, this approximation appears particularly reasonable in the $M = 8$ to $M = 12$ range where our candidate streaks occur (see Tables 2-6). More precise estimates could be given, but our use of this parameter is only for making rough estimates. As described at the end of the previous section, the simplicity of these assumptions should be reasonable when the pedigree is only roughly known.

False negatives and false positives

The discovery of a SNP streak among a collection of samples that exceeds the corresponding threshold is an indication of an IBD region. The limits of the SNP streak

²It is possible to reinforce the belief that a region is IBD and to obtain a more accurate P -value by incorporating other samples. See [Leibon, et al. (2007)] for details.

demarcate a region in the genome that (with high probability) contains an IBD region. Here we compute the probability of a false negative (i.e., the probability that an IBD region shared among the samples is not contained in a compatible haplotype region) as well as the probability of a false positive (concluding the existence of an compatible haplotype region that does not contain an IBD region). Notice, in [Miyazawa, et al. (2007)], the false positive region of an compatible haplotype region that does contain an IBD region, but is not itself IBD, is also explored and called “type B false positive.” As our regions are composed of a very small number of SNP streaks, we only concern ourselves with what is called in [Miyazawa, et al. (2007)] a “type A false positive.” Our analysis of these errors follows that of [Miyazawa, et al. (2007)] with the necessary modifications.

- *False negatives* – For the computation of the probability of a false negative, as in [Miyazawa, et al. (2007)], we use Haldane’s Poisson model. Namely, by adapting the derivation in [Miyazawa, et al. (2007)] and using morgans instead of centimorgans, the Poisson parameter becomes

$$\lambda = B_M$$

and we obtain a ratio of the false negative to the entire genome of

$$R_{\text{False negatives}} = 1 - e^{-\lambda c}(1 + \lambda c). \quad (1)$$

More precisely, $R_{\text{False negatives}}$ is the fraction of the genome that is IBD but is not labeled as such due to the fact that the morgan length of the interval is below our cut-off. An argument to produce this estimate is simple, namely the expected length of an interval that is too short to be detected is $\lambda \int_0^c x e^{-\lambda x} dx = \frac{1}{\lambda}(1 - e^{-\lambda c}(1 + \lambda c))$, while the expected number of intervals per genome is λ . Multiplying these together gives the needed estimate.

- *Type A False Positives* – Once again, we adapt the argument of [Miyazawa, et al. (2007)]. In this case the ratio of the type A false positives to the entire autosome R_A , depends on an estimate of the length between mismatched compatible loci in the sample, which can be taken to be exponential with parameter λ with

$$\lambda = \frac{N_b}{L}.$$

Using an argument as in the previous section, we can then form the estimate

$$R_A = (1 + \lambda c)e^{-\lambda c}. \quad (2)$$

More precisely, the expected length of an interval that is long enough to be considered a false positive is $\lambda \int_c^\infty x e^{-\lambda x} dx = \frac{1}{\lambda} e^{-\lambda c} (1 + \lambda c)$, while the expected number of intervals per genome is λ . Multiplying these together gives the needed estimate.

Recall, R_A is the ratio of the size of the genome that is in a state of being false positive to the size of the whole genome. The Poisson approximation in (2) requires q_M to be small, which it need not be for large M . An approximation not requiring this assumption is given by

$$R_A = \frac{\left(N_c + \frac{1}{q_M}\right) S q_M p_M^{N_c}}{S} = (N_c q_M + 1) p_M^{N_c}. \quad (3)$$

Notice $\lambda c = N_c q_M$, hence the approximation (2) satisfies

$$R_A = (N_c q_M + 1) \left(1 - \frac{q_M N_c}{N_c}\right)^{N_c} = (1 + \lambda c) \left(1 - \frac{\lambda c}{N_c}\right)^{N_c}$$

which is $\approx (1 + \lambda c) e^{-\lambda c}$ for λc small relative to N_c .

For large M the false positives are rare, so a second quantity can be used as a more meaningful false positive rate. As in [Miyazawa, et al. (2007)], using the approximation where the SNPs are assumed independent and occurring with the average allele frequency we can estimate

$$\frac{Pr(\exists \text{ compatible haplotype that does not contain an IBD region})}{Pr(\exists \text{ chance streaks of length } \geq N_c)} =$$

The fact that this probability is less than or equal to the expected number of such streaks³ yields

$$Pr(\exists \text{ Chance streaks of length } \geq N_c) \leq S q_M p_M^{N_c}. \quad (4)$$

This approach could be enhanced to compensate for linkage disequilibrium and variability in SNP frequency. Simple changes, such as using a Markov model rather than an independent model for assessing the probability of sharing an allele could be used. One could also use known SNP frequencies and correlation estimates when computing the probabilities. The rough estimates given here

³This is true since $P(A) = E(X)$ where $X \equiv 1$ if A occurs and zero otherwise, and, in general, if $X \leq Y$ then $P(A) \leq E(Y)$. Also, in our case, $E(Y) - P(A)$ is small when the probability of multiple events (chance streaks in our case) is also small.

would be most useful when the pedigree variance is high (as in our example), but when the pedigree variance is low, such modifications would help produce a more accurate assessment of the false positive rate (see also see [Donnelly (1983)], [Cannings (2003)], and [Thomas, et al. (2008)] for various suggestions for incorporating more detailed information).

P-values for regions IBD

In [Miyazawa, et al. (2007)] a *region from a common ancestor* or *RCA* is defined as a homozygous IBD chromosomal segment. Given two descendants who are m and n ($m \geq n$) generations removed from of a pair of founders, the quantity $RCA(m, n)$ for $m \geq n$ is defined therein as the expected ratio of the RCA to total length of the autosome (see Equation (1) in [Miyazawa, et al. (2007)]).

The derivation of $RCA(m, n)$ in [Miyazawa, et al. (2007)] can be extended to a family tree containing all and only the descendants of an ancestral couple in direct lineage with the samples. We thus, for a set B of nonfounders in a family tree, define the quantity $IBD(B)$ as the expected fraction of the autosome shared by B . We consider such trees since we are tracking IBD regions and such a tree represents the path of a genetic region from a recent common founder to our samples.

If B denotes the number of non-founders in the tree (if our founders have a least two children in the tree) then we have:

$$IBD(B) = 2^{-B+2}.$$

This is also the probability that a given point in the genome is *IBD*. Note that calling a pedigree a “tree” simply means that each of the four candidate founding chromosomes determines a tree when we connect the members of the pedigree with the meiosis links capable of passing along this chromosome’s genetic material. The four possible chromosomes corresponds to the four in expression $IBD(B) = 4(2^{-B}) = 2^{-B+2}$.

The computation in (2) is closely related to a P-value. Namely if we have identified an interval as an IBD region, then the P-value is the probability of a randomly chosen interval being an IBD region. Of course, a good notion of P-value should reflect the probability that there is **some** IBD interval **detected** due to chance. This is bounded above by the expected number of such intervals

$$Pr(\exists \text{ IBD}) \leq (LB + 21)e^{-Bc}IBD(B). \quad (5)$$

As in the computation (4), we bound the probability by the expected number of such intervals. The expected number of intervals with different genetic linkage in our

pedigree is $(LB + 21)$ and for each one there is some probability that it is IBD and detected (where detect means that it is longer than our cut-off c). The probability that an interval is IBD and detected equals $P(\text{Detect}|\text{IBD})P(\text{IBD}) = e^{-Bc}IBD(B)$. This gives us our final estimate.

In our applications we assume that we have chosen a set of T samples such that each subset of size M corresponds to a pedigree with roughly B_M non-founders independent of the choice of the subset of size M . Hence the probability that we find an *IBD* region among the subsets of size M is bounded by the expected number of such intervals

$$Pr(\exists M, \text{IBD region among a subset of size } M) \leq \binom{T}{M} Prob(\exists \text{ IBD}). \quad (6)$$

Notice, the multiple testing effect producing the factor of $\binom{T}{M}$ suggests that one would like to focus on as large a subset as possible. In our example in Section 2, we started with all $T = 17$ samples and worked our way down, finding our first significant streak at $M = 11$. Also, notice this approximation assumes the pedigree is a tree with independent founders, which is only a **very** rough approximation in our case.

***P*-values for compatible haplotype regions**

The probability that we detect a compatible haplotype region is less than the probability that we detect an *IBD* region plus the probability that we see a SNP streak of length greater than N_c .

Also, $Pr(\exists M, \text{Chance streaks of length } \geq N_c \text{ among a subset of size } M)$ is less than the expected number of compatible haplotype regions, so by our estimate $Pr(\exists \text{ chance streaks of length } \geq N_c)$ in equation (4) we see

$$Pr(\exists M, \text{Chance streaks of length } \geq N_c \text{ among a subset of size } M) \leq \binom{T}{M} Sq_M p_M^{N_c}.$$

This enables a computation of a *P*-value as

$$\begin{aligned} P(\exists \text{ compatible haplotype}) &= Pr(\exists M, \text{IBD among a subset of size } M) \\ &\quad + Pr(\exists \text{ Chance streak of length } \geq N_c) \end{aligned}$$

so that

$$P\text{-val} \approx \binom{T}{M} (4(LB + 21)e^{-B(\ln(2)+c)} + Sq_M p_M^{N_c}). \quad (7)$$

Example

Analysis of a family with kidney disease

Now to consider the data. Unfortunately the false negative rate is very high for 113,000 SNPs. For example, for a set of size $M = 12$ we would require a cut-off of 33 for our streak length in order to have the false negative rate be below 0.05. However that does not prevent us from trying to find relevant extreme events (i.e., small P-value). We can use the P-value from equation (7) to gauge this. Table 1 shows the cut-off values that guarantee a P-value of less than 0.05 as a function of M . We include what the cut-offs would look like for 500k and 1M SNPs.

# SNPs	M	5	6	7	8	9	10	11
113k	N_c	1992	1437	998	634	322	105	91
	c	0.599	0.432	0.3	0.191	0.097	0.032	0.027
500k	N_c	8814	6356	4413	2804	1422	202	97
	c	0.599	0.432	0.3	0.191	0.097	0.014	0.007
1M	N_c	17628	12712	8826	5607	2844	403	100
	c	0.599	0.432	0.3	0.191	0.097	0.014	0.003
# SNPs	M	12	13	14	15	16	17	
113k	N_c	81	72	63	54	45	35	
	c	0.024	0.022	0.019	0.016	0.014	0.011	
500k	N_c	86	77	67	58	49	39	
	c	0.006	0.005	0.005	0.004	0.003	0.003	
1M	N_c	89	79	70	61	51	41	
	c	0.003	0.003	0.002	0.002	0.002	0.001	

Table 1: Cut-off values for significant streaks for FGFM as a function of subset size. This table depends on having $T = 17$ total samples and approximating B_M with $3M$, nothing else about the FGFM family. Columns are indexed according to the value of M (first row). The next two rows correspond to our 113,000 SNPs, while the next two rows correspond to a 500k similarly informative SNPs, and the final two rows correspond to 1M similarly informative SNPs.

As the data shows (Tables 2–6) we find that that for $M \leq 11$ there are instances

of statistically significant events. For $M \geq 12$ the probability

$$Pr(\exists \text{ Chance streak of length } \geq N_c)$$

is quite high. Thus, the probability that these streaks contain an IBD region with respect to all M samples is likely to be very low.⁴ For small M the significance of long streaks is mediated by the high probability

$$Pr(\exists M, \text{ IBD among } M).$$

Hence, these streaks are very likely to contain an *IBD* region, but such IBD regions should be considered as unremarkable events due to chance alone. An example is the streak of length 265 on chromosome 22 involving 8 samples. Tables 2–6 provide a listing of significant streaks with their positions on the genome.

In Tables 2–6 we give the results for all subsets of sizes 8–12. There were no significant streaks involving more than 11 out of 17 of our samples. The tables are of a common form: column 1, (labeled L) gives the length of the streak, column 2 gives the chromosome, column 3 gives the associated P-value for the compatible haplotype, columns 4 and 5 give the start and end position of the streak. The remaining 17 columns index the patient samples: a “1” in a given column indicates the presence of that patient in the sample subset and a “0” indicates the absence. Notice we find 10 regions that have a P-value smaller than the traditional 0.05 cut off, but the streaks are often redundant. Namely we often see the phenomenon of shorter streaks shared among more samples contained within longer streaks shared among fewer samples. Presumably this is an indication of recombinations that successively cut down the size of the region.

L	Ch	P-Val	St	End	Subset
68	5	> 0.05	57.5	58.1	{3,5,7,8,9,11,12,13,14,15,16,17}
61	5	> 0.05	108.4	109.5	{1,2,5,6,7,9,10,11,13,14,15,17}

Table 2: Streak data for subsets of size 12

3 Conclusion

In this paper, we have extended and generalized the recently published work of Miyazawa et al. [Miyazawa, et al. (2007)]. We use dense SNP genotyping infor-

⁴This suggests that this disease is not “Mendelian” - i.e., a single defect in one gene does not appear to be the cause of this disease.

L	Ch	P-Val	St	End	Subset
162	5	0.013	43.4	52.2	{3,5,7,8,9,12,13,14,15,16,17}
82	5	> 0.05	57.3	58.1	{3,5,7,8,9,11,12,13,14,15,17}
77	5	> 0.05	62.5	64.1	{3,5,7,8,9,12,13,14,15,16,17}

Table 3: Streak data for subsets of size 11

L	Ch	P-Val	St	End	Subset
606	5	0.0003	54.2	65.1	{3,5,7,8,9,12,13,14,15,17}
237	5	0.01	43.4	53.8	{3,5,7,8,12,13,14,15,16,17}
144	7	0.02	134.5	137.6	{3,4,5,6,7,10,13,14,16,17}
82	2	> 0.05	52.6	53.7	{1,3,4,5,8,9,11,13,14,16}
82	5	> 0.05	59.1	60.9	{1,3,5,6,8,9,13,14,15,17}
81	4	> 0.05	94.1	96.0	{3,4,5,6,7,10,11,12,16,17}

Table 4: Streak data for subsets of size 10

mation (such as that obtained from the new array-based genome wide panels from Affymetrix and Illumina) to identify regions of the genome highly likely to be inherited identical-by-descent in groups of related individuals. We believe this to be of particular utility in identifying putative disease loci in large complex families.

In essence, our method is based on the notion that with very dense, highly informative genetic maps, obligate recombination events can be identified, even if haplotype phases are unknown. Specifically, if some individuals are genotype AA at a marker where others are BB, they cannot have inherited this locus IBD from a common ancestor (ignoring the rare possibility of a new mutation altering the genotype). If we look at two or more individuals, if a sufficiently long “streak” of SNPs is observed lacking such obligate recombination events, then we can make conclude that such a region is in fact IBD.

We give reasonable estimates of the sensitivity of this approach for the detection of an IBD region by identification of compatible haplotype regions, as well as estimates of the statistical significance of finding such a compatible haplotype region. We discuss a specific example of a complex pedigree in which multiple family members have kidney disease, and identify regions of the genome highly likely to be shared IBD in a significant subset of affected family members.

The HH approach of Miyazawa et al. [Miyazawa, et al. (2007)] and the generalization described here have certain features which make this methodology attractive in the analysis of large families with Mendelian or oligogenic disease. It is computa-

Leibon et al.: Identification of IBD regions					
L	Ch	P-Val	St	End	Subset
859	5	0.0006	43.4	65.1	{3,5,7,8,12,13,14,15,17}
639	5	0.004	54.2	66.0	{5,7,8,9,12,13,14,15,17}
327	7	0.048	131.4	137.6	{3,4,5,6,10,13,14,16,17}
198	5	> 0.05	42.0	52.3	{3,5,7,8,9,12,13,14,15}
195	7	> 0.05	134.5	140.2	{3,5,6,7,10,13,14,16,17}
156	5	> 0.05	61.1	64.1	{5,7,8,9,12,14,15,16,17}
95	14	> 0.05	78.3	79.5	{2,3,7,8,9,13,14,15,17}
94	7	> 0.05	132.4	133.9	{3,4,5,6,7,10,13,14,17}

Table 5: Streak data for subsets of size 9

tionally much simpler than linkage analysis. Unlike linkage analysis, it is reasonably forgiving of errors in pedigree structure.

We have generalized the results of [Miyazawa, et al. (2007)] to incorporate comparisons between multiple individuals, not just pairwise comparisons. As we have shown above, this provides greater power to identify significant, short IBD regions.

As the platforms for dense SNP genotyping continue to evolve into providing even denser ($> 1,000,000$ SNPs per genome) coverage, the sensitivity of this method will continue to improve, as statistical significance of a streak of a given length will increase, and the risk of missing an IBD region decreases.

L	Ch	P-Val	St	End	Subset
894	5	0.008	42.0	65.1	{3,5,7,8,12,13,14,15}
892	5	0.008	43.4	66.0	{5,7,8,12,13,14,15,17}
737	5	0.02	54.2	67.7	{5,7,8,9,12,14,15,17}
461	5	> 0.05	38.5	52.3	{3,5,7,8,9,12,14,15}
404	7	> 0.05	128.7	137.6	{3,4,5,6,10,13,14,16}
404	7	> 0.05	128.7	137.6	{3,4,6,10,13,14,16,17}
378	7	> 0.05	131.4	140.2	{3,5,6,10,13,14,16,17}
265	22	> 0.05	25.9	34.9	{2,4,5,8,9,10,16,17}
238	7	> 0.05	134.5	141.5	{3,5,7,10,13,14,16,17}
223	14	> 0.05	62.5	69.1	{2,3,7,8,9,14,15,17}
120	5	> 0.05	61.8	64.1	{3,5,9,13,14,15,16,17}
118	14	> 0.05	78.3	80.1	{2,3,7,8,9,14,15,17}
105	8	> 0.05	139.0	146.1	{3,5,6,8,10,13,14,16}
102	7	> 0.05	125.5	126.7	{2,7,8,9,10,14,15,16}

Table 6: Streak data for subsets of size 8

References

- [Cannings (2003)] Cannings, C. (2003) The identity by descent process along the chromosome. *Human Heredity*, **56**, 126–130.
- [Donnelly (1983)] Donnelly, K. P. (1983) The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*, **23**, 34–63.
- [Duda, et al. (2000)] Duda, R. O., P. E. Hart, and D. G. Stork (2000) *Pattern Classification (second edition)*, Wiley, NY.
- [Leibon, et al. (2007)] Leibon, G., D. Rockmore, and M. R. Pollak (2007) A simple computational method for the identification of disease-associated loci in complex, incomplete pedigrees, arXiv:0710.5625v1 [q-bio.GN]
- [Matsuzaki, et al. (2004)-1] Matsuzaki, H., S. Dong, H. Loi, X. Di, G. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. Yang, G. C. Kennedy, T. A. Webster, S. Cawley, P. S. Walsh, K. W. Jones, S. P. Fodor, R. Mei (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods*. Nov;**1**(2), 109–11.

- [Matsuzaki, et al. (2004)-2] Matsuzaki, H., H. Loi, S. Dong, Y. Y. Tsai, J. Fang, J. Law, X. Di, W. M. Liu, G. Yang, G. Liu, J. Hunag, G. C. Kennedy, T. B. Ryder, G. A. Marcus, P. S. Walsh, M. D. Shriver, J. M. Puck, K. W. Jones, R. Mei (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.*, Mar;**14**(3), 414–25.
- [Miyazawa, et al. (2007)] Miyazawa, H., M. Kato, T. Awata, M. Kohda, H. Iwasa, N. Koyama, T. Tanaka, Huqun, S. Kyo, Y. Okazaki, and K. Hagiwara (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *The American Journal of Human Genetics*, **80**, 1090–1102.
- [Ott (1991)] Ott, J. (1991) *Analysis of Human Genetic Linkage (revised edition)*, Johns Hopkins University Press, Baltimore.
- [Thomas, et al. (1994)] Thomas, A., M. H. Skolnick, and C. M. Lewis (1994) Genomic mismatch scanning in pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology*, **11**(1), 1–16.
- [Thomas, et al. (2008)] Thomas, A., N. J. Camp, J. M. Farnham, K. Allen-Brady, L. A. Cannon-Albright (2008) Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Annals of Human Genetics*, **72** (2) , 279–287.