

## Significance testing for direct identity-by-descent mapping

G. R. GRANT<sup>1</sup>, E. MANDUCHI<sup>1</sup>, V. G. CHEUNG<sup>2</sup> AND W. J. EWENS<sup>3</sup>

<sup>1</sup> *Center for Bioinformatics, University of Pennsylvania*

<sup>2</sup> *Department of Pediatrics, University of Pennsylvania*

<sup>3</sup> *Department of Biology, University of Pennsylvania*

(Received 22.3.99. Accepted 30.9.99)

### SUMMARY

Direct identity-by-descent mapping is a technique for narrowing down the location of the gene or genes responsible for a given genetic disease to small segments of the genome. The technique involves DNA comparisons between pairs of affected individuals. The data generated are in the form of matching segments of the genome, representing regions likely to be identical-by-descent (IBD). Regions in the genome over which there are significantly more segments aligned than is expected by chance are taken as candidate regions for the disease gene or genes. Due to the complex geometric nature of the data, significance testing involves certain mathematical difficulties. We present here a new method for measuring this significance. This method introduces a novel statistic and is appropriate whether or not the relationships between the paired individuals are known. We give examples that we have calculated by implementing this method, including an application to real data.

### INTRODUCTION

In this paper we describe a method for testing the statistical significance of data obtained in a direct identical-by-descent (IBD) gene mapping procedure. Direct IBD mapping is a combination of DNA microarray technology and genomic mismatch scanning (GMS); see Nelson *et al.* (1993), Brown (1994), McAllister *et al.* (1998), and Cheung & Nelson (1998) for a description of the technique.

Our focus is on using the technique to map disease genes. This is not a new endeavor. Mirzayans *et al.* (1997) used the method for locating the chromosome region containing the gene for an eye pigmentation anomaly. Important contributions to the linkage theory for this application of the method have been made by Feingold *et al.* (1993), Houwen *et al.* (1994), Thomas *et al.* (1994), Guo (1995), Smalley *et al.* (1996), and Durham & Feingold (1997).

However, all linkage methods advocated and discussed in the references listed above are based on the use of related individuals. By 'related individuals' we mean individuals whose relationship is known to us, and is thus generally close. By contrast, the linkage method we propose here can be used with individuals who are so distantly related that their relationship is not known to us, 'unrelated individuals', as in Cheung *et al.* (1998). The ability to use unrelated individuals has certain advantages over the use of individuals whose relationship is assumed known. First, the relatedness requirement imposes an unnecessary restriction on the data sets that can be used. Second, closely related individuals necessarily share significant portions of the genome IBD, making fine mapping difficult. Affected individuals who are distantly related tend to share comparatively short regions

Correspondence: Dr Gregory R. Grant, Center for Bioinformatics, University of Pennsylvania, 1333 Blockley Hall, 418 Guardian Drive, Philadelphia, PA 19104-6021, USA. Tel: 215 573 3117; Fax: 215 573 3111.

E-mail: ggrant@pcbi.upenn.edu

of the genome near the disease locus (or loci). Our approach thus shares several of the benefits of linkage mapping via association for complex diseases, as described by Risch & Merikangas (1996). However, even if the relatedness of some or all individuals happens to be known, our analysis remains valid.

### *IBD mapping*

Direct IBD mapping is described in detail by Cheung *et al.* (1998). For completeness we outline the procedure here.

Given a target genetic disease, a group of individuals, of possibly unknown relationships, who express the disease is selected from a population isolate (in the case of Cheung *et al.* (1998) the disease is hyperinsulinism and the population isolate is the Ashkenazi Jews). The theory discussed below assumes that the individuals are grouped into independent pairs. Clearly there is a great degree of arbitrariness concerning the choice of pairings. This arbitrariness is removed by considering all possible pairs, but if this is done independence is lost. We will raise these issues again in the Discussion section.

The DNA of the two individuals in any pair is compared for exact matching at a sequence of DNA fragments, where each fragment is on the order of 5–6 kb long. Comparisons are done using the genomic mismatch scanning (GMS) technique. In GMS comparisons, the DNA fragments are enriched by using mismatch repair proteins to remove mismatch-containing heterohybrids. Since the rate of natural polymorphism in the human genome is at least 1 in 1000 basepairs, large genomic regions that are identical in sequence are likely to be IBD (see Nelson *et al.* (1993) and Cheung *et al.* (1998) for further discussion of this issue). The GMS-selected DNA fragments are then mapped by hybridization onto a DNA microarray containing clones arrayed in chromosomal order. Adjoining clones are separated in the genome by long sequences of DNA on the order of one Mb. The clones on the microarray that are hybridized by the IBD DNA fragments are scored as ‘positives’.

We call the set of clones where two individuals are IBD the IBD *profile* of that pair. Under the (null) hypothesis that there is no disease gene in the region of the genome considered, we expect these profiles to be located randomly. If, by contrast, some of the affected individuals have inherited a disease gene (or genes) in this region from a common ancestor, then we expect an overrepresentation of the IBD profiles clustering around the location(s) of the disease gene(s). The tests we consider for linkage are thus tests for significant clustering of the IBD profiles. Since IBD profiles decrease in size through generations, the use of distantly related individuals should identify a reasonably small part of the genome containing the disease gene or genes, if any exists in the region of genome considered.

### *Notation, Data, and Statement of Problem*

Henceforth we use the word ‘genome’ to mean the part of the complete genome considered. We will think of the clones described in the previous section as points, and assume that these points are equally spaced, at unit intervals, along the genome, and numbered  $1, 2, \dots, L$ . A profile is not necessarily connected, that is, it is not necessarily a collection of consecutive clones. We define an IBD *interval*, or simply an interval, of a profile to be a set of consecutive clones in the profile surrounded by clones which are not in the profile. An IBD profile is then the disjointed union of its IBD intervals.

We define the *size* of an IBD profile (resp. interval) as the number of clones in that profile (interval). An IBD profile is then specified by giving the number of IBD intervals it contains, their respective sizes, and their respective positions (usually specified by giving the left endpoint of each interval). We say that an IBD profile (resp. interval) *covers* the clones it contains.

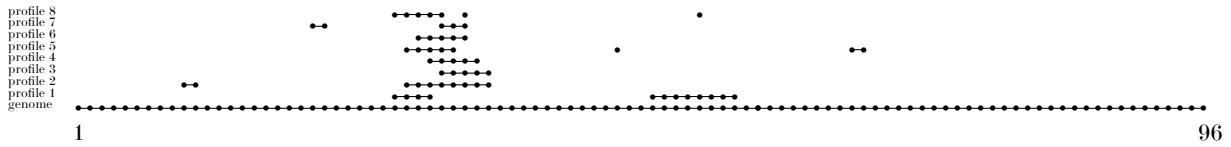


Fig. 1. The data of Cheung *et al.* (1998) consists of eight IBD profiles. The bottom long segment represents the genome, and each successive row corresponds to an IBD profile, that is, to a comparison of two affected individuals.

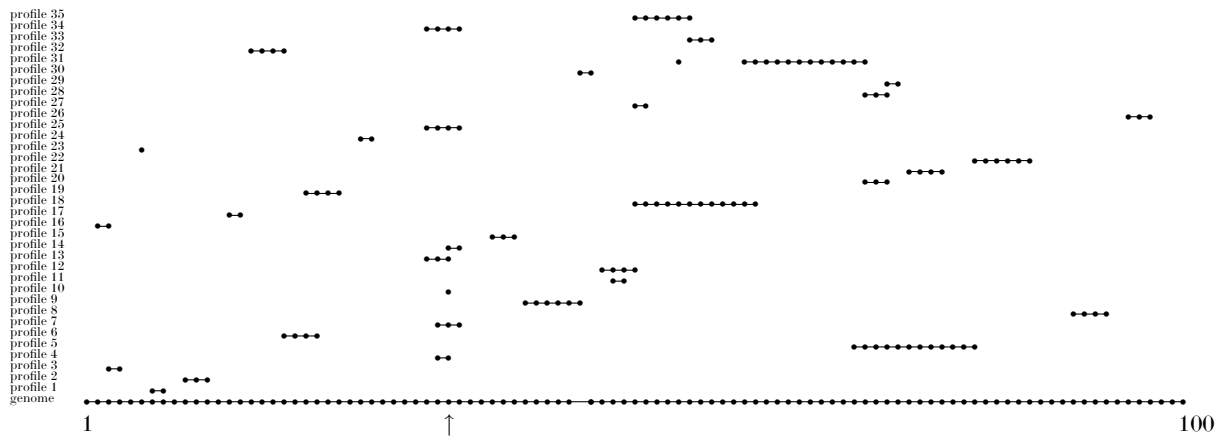


Fig. 2. An example of 35 IBD profiles (each consisting of one interval only) for which the significance of the clustering above the point indicated by the arrow is not clear to the naked eye.

We denote the number of IBD profiles in the data by  $k$ . Any such profile is then a subset of  $\{1, 2, \dots, L\}$ .

The data of Cheung *et al.* (1998), shown in Figure 1, illustrate these definitions. There are  $k = 8$  paired comparisons, leading to eight IBD profiles, each displayed as horizontal segments: profile 1 consists of two intervals of sizes 4 and 8 and left endpoint positions 27 and 49 respectively, profile 2 consists of two intervals of sizes 2 and 8 and left endpoint positions 9 and 28 respectively, etc.

As mentioned above, our statistical procedure tests for significant clustering of IBD profiles over one or more locations in the genome. If there are locations in the genome at which *all*  $k$  pairs of subjects are IBD, then significance will probably be evident to the naked eye. But if there is a substantial amount of noise in the data, with some pairs of affected individuals being IBD in some region (or regions) not containing the disease gene, the significance may be unclear. Clones that happen to be identical in sequence but not IBD will also add to the noise.

For example, in Figure 2 (where  $L = 100$  and  $k = 35$ ) there are seven short (connected) IBD profiles covering one point on the chromosome (indicated by the arrow). Is this significant under some reasonable measure of clustering? We will later apply the method we propose to answer this question. We will also apply our method to analyze the data of Cheung *et al.* (1998) and will confirm the conservative  $p$ -value that they calculate.

As mentioned above, our approach does not assume knowledge of the various relationships between the individuals in the various pairs. Instead the size and the number of intervals of the IBD profile defined by any pair of individuals are taken as a surrogate for the degree of relatedness of those two individuals. Because the number of intervals constituting each of the various IBD profiles as well as

the sizes of these intervals are taken as given, the significance associated with any test is calculated conditional on these values.

There are two broad approaches to hypothesis testing in applied statistics. The first uses likelihood ratios and in effect rejects the null hypothesis if the probability of the observed data under that hypothesis is sufficiently small compared to the probability of those data under the alternative hypothesis. The second, *which we adopt here*, is to decide on some test statistic which appears to distinguish between null and alternative hypotheses, and to reject the null hypothesis if the observed value of that statistic is sufficiently extreme as judged by its null hypothesis distribution. We now discuss in turn the null hypothesis setting and the choice of a test statistic.

#### *Null hypothesis*

Null hypothesis properties are found by standard randomization procedures. If the genome does not contain a disease gene (the null hypothesis), all allowable rearrangements of the IBD profiles are equally likely. An *allowable rearrangement* of the profiles in the data is one in which the number of intervals in each profile and their sizes are left unchanged.

Any proposed test statistic, such as the ‘max’ statistic discussed in the next section and the various footprint statistics discussed later, will have a probability distribution under the null hypothesis defined by these equally likely rearrangements. We call this the *randomization distribution* of the test statistic. The assessment of the significance of the observed value of this statistic is then determined by reference to its randomization distribution.

For a connected IBD profile (i.e. one consisting only of one interval), all possible locations on the genome are equally likely. However, the probability distribution of the location of an interval coming from a disconnected IBD profile is not quite uniform, as locations near the ends become more likely. When the size of the IBD profile is small compared to the size of the genome, the probability distribution of the location of any of its intervals differs negligibly from uniform. For example, in the case of the data of Cheung *et al.* (1998) in figure 1, there are 89 possible locations on the genome for the left endpoint of the interval of size 8 of profile 1. The null hypothesis probabilities of these locations do not differ by more than 0.0006, and are in fact identical for locations 5–85.

Thus, when the size of an IBD profile is small, the distribution of the location of any of its intervals can be approximated well by a uniform distribution. When the size of the profile is large, the distribution of the location of any of its intervals should not be approximated in this way. In any case this distribution can be quickly obtained computationally.

With the null hypothesis framework in hand, we now turn to the discussion of test statistics.

#### *The ‘max’ statistic $M$*

For each clone  $u$  in the genome ( $u = 1, 2, \dots, L$ ) let  $Y_u$  be the number of IBD profiles (out of  $k$ ) which cover this clone. Clearly  $0 \leq Y_u \leq k$ . Define the ‘max’ statistic  $M$  as

$$M = \max_{u=1,2,\dots,L} Y_u.$$

This statistic is at the core of the statistical tests of Feingold *et al.* (1993), Smalley *et al.* (1996) and Cheung *et al.* (1998). To assess the significance of an observed value of  $M$  we must compute or approximate the upper tail probabilities of its null hypothesis distribution. In Appendix A we give exact null hypothesis formulae for  $\text{Prob}(M = k)$  and  $\text{Prob}(M = k - 1)$  in the case where each IBD profile is connected, i.e. is an interval.

The calculations for the probabilities that  $M = k - 2$ ,  $M = k - 3$ , ... are increasingly complex. When, as in practice will usually be the case, the profiles are disconnected, the corresponding

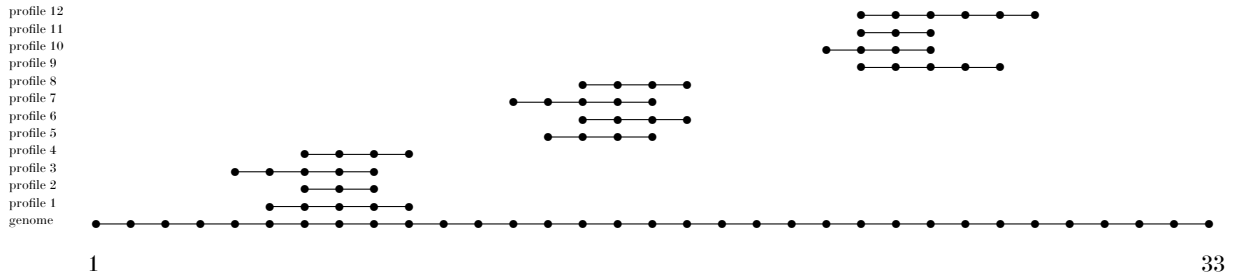


Fig. 3. An example of twelve connected IBD profiles, forming three stacks of four. The value of the ‘max’ statistic is 4 and the footprint is 19.

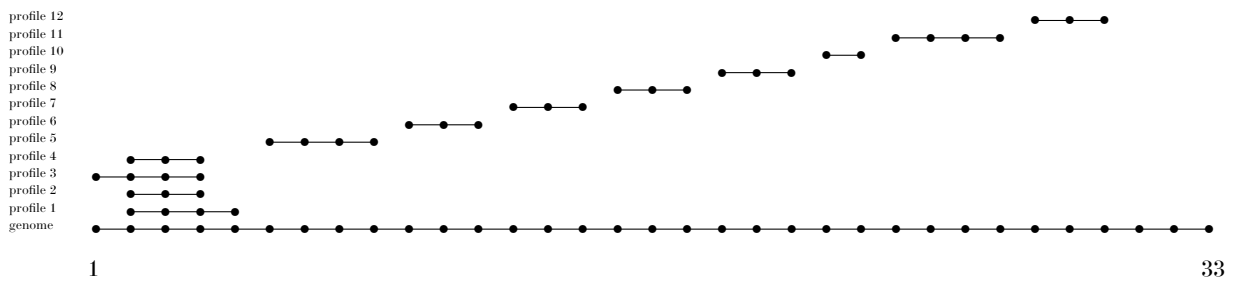


Fig. 4. An example of twelve connected IBD profiles, for which the value of the ‘max’ statistic is 4 and the footprint is 30.

calculations become even more complicated. In such cases it seems that simulation is the best way to approximate the distribution of  $M$ .

For the data in Cheung *et al.* (1998) not all of the IBD profiles are connected, and  $M = k - 1$ . The statistical analysis in that paper used a Bonferroni correction to give a conservative  $p$ -value of  $3 \times 10^{-4}$ . Using simulation, we have found an estimate of  $3 \times 10^{-5}$  for the  $p$ -value of their data; one tenth of their conservative estimate.

While the simplicity of the statistic  $M$  is appealing, it is easy to construct cases where it might not give sufficiently informative results. An example is given by the hypothetical data in Figure 3. For these data  $M = 4$ . A value  $M = 4$  when  $k = 12$  is not particularly significant. However the fact that there are three stacks of four is much stronger evidence for the existence of disease genes in the genome (alternative hypothesis) than, for example, the data in figure 4, for which  $M$  also equals 4.

While data as extreme as those depicted in Figure 3 will probably not arise in practice, they are sufficient to illustrate a limitation with the statistic  $M$ . The  $M$  statistic will see no difference between the two configurations in Figures 3 and 4. This is a realistic concern for direct IBD mapping data, as there may be several genes responsible for the disease in question, in which case data approximating those shown in Figure 3 could likely arise. For this reason we turn to a second statistic whose use overcomes this problem.

*The ‘footprint’ (the connected case)*

For ease of exposition we start by analyzing the case in which all IBD profiles are connected. We will consider the general case in a later section.

Let  $\mathcal{C}$  be a collection of  $k$  connected IBD profiles (so each profile in  $\mathcal{C}$  is an interval). We define the *footprint*  $F(\mathcal{C})$  of this collection as the number of clones in the genome that are covered by at

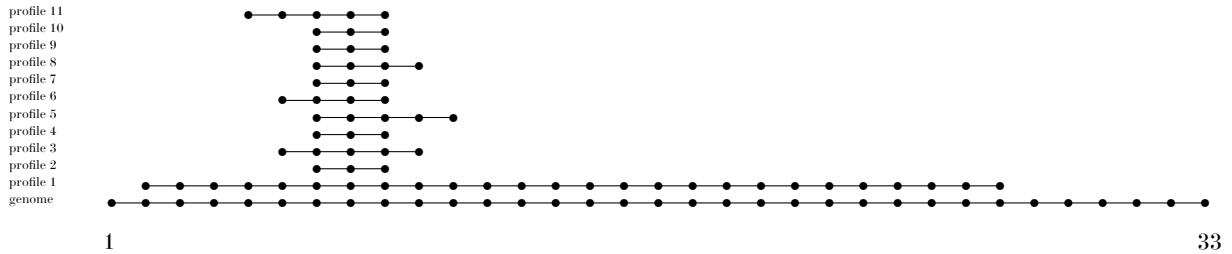


Fig. 5. A collection of eleven connected IBD profiles, with evident clustering, but large footprint.

least one of the profiles in  $\mathcal{C}$ . In this section we discuss the properties of  $F(\mathcal{C})$ . This will lead to the introduction of a new test statistic, the normalized footprint, which will be the basis of our method to test for significance in IBD mapping.

In each of Figures 3 and 4, there are twelve connected IBD profiles. In Figure 3 the footprint is small and supports the hypothesis that there are three possible locations for disease genes. In figure 4 the footprint is large and there appears to be no significant evidence of a disease gene in the region sampled. In general, sufficiently small values of  $F(\mathcal{C})$  support the hypothesis that a disease gene or disease genes occur in the genome. Thus, unlike  $M$ ,  $F(\mathcal{C})$  has the potential to serve as a test statistic for linkage which is more sensitive to differences such as those illustrated in Figures 3 and 4.

However, a more subtle approach than the direct use of  $F(\mathcal{C})$  is necessary. The reason for this is illustrated by the hypothetical data in Figure 5. The data in this figure are clearly significant; however, this significance is not picked up by the footprint because of the long single IBD profile (profile 1), which might have been caused by the pairing of two closely related individuals.

We resolve this problem by considering the collection of footprints defined by all possible subsets of size  $m$  of  $\mathcal{C}$ , where  $m$  varies from 2 to  $k$ . This will lead, for example, to a significance for the data in figure 5 deriving from the subset consisting of the ten short IBD intervals.

Two points must be kept in mind when adopting this approach. First, suppose for example that the data consist of  $k = 30$  connected profiles (so each profile is a single interval). It is possible that, say, ten of these are closely aligned and, considered on their own, would define a significantly small footprint. However when considered as the ‘best’ set of ten out of the thirty intervals, this significance might disappear. We must ask if the footprint of the ten intervals is significantly smaller than the expected *smallest* footprint of ten intervals out of thirty. While we illustrate this point here in the case where all IBD profiles are connected, the same problem obviously still exists in the general case.

Second, the footprint should not be used directly for comparisons between two subsets, each with the same number of intervals but of differing sizes in each subset. For example, if five intervals in one subset each have size two, and five intervals in another subset each have size ten, and if the short intervals do not overlap at all and the long intervals align perfectly, the footprints of the two sets of five will be equal (see figure 6). But the five long perfectly aligned intervals are obviously more significant than the five short unaligned ones. This argument leads us to the normalized footprint.

*The ‘normalized footprint’ (the connected case)*

Given a set  $\mathcal{S}$  of  $m$  connected IBD profiles, we define the *normalized footprint* of  $\mathcal{S}$  by

$$NF(\mathcal{S}) = \frac{F(\mathcal{S})}{E(F)}, \tag{1}$$

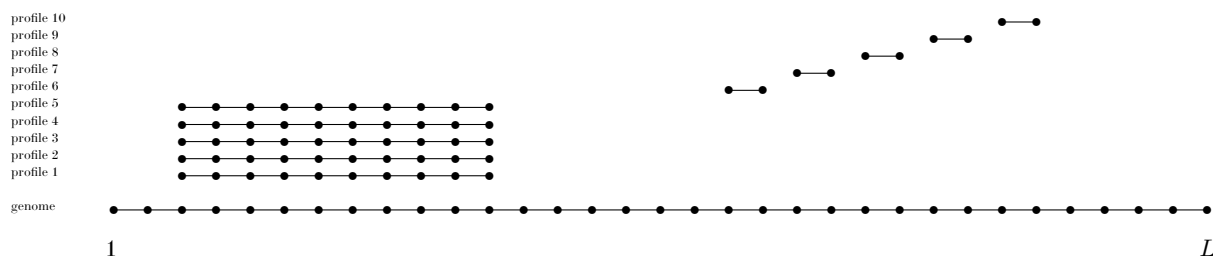


Fig. 6. A collection of ten connected IBD profiles. The footprint of the subset consisting of the five perfectly aligned intervals is 10. The footprint of the subset consisting of the remaining five intervals is also 10.

where  $F(\mathcal{S})$  is the number of clones covered by at least one interval in  $\mathcal{S}$  and  $E(F)$  is the null hypothesis expected value of  $F$ , obtained by considering all allowable rearrangements of the  $m$  profiles (which are just intervals in this case). The formula for  $E(F)$  will be given in the theorem in Appendix B. The normalized footprint allows a meaningful comparison of subsets of IBD intervals of different sizes. Smaller normalized footprints indicate more alignment than larger normalized footprints, independently of the sizes of the intervals involved.

This is illustrated as follows. Suppose that in Figure 6,  $L = 100$ . Under randomization, the expected footprint of the five short intervals is 9.606 (from (B 1) and (B 8) in Appendix B). This value arises since, under most randomizations, the intervals do not overlap, leading to a footprint of 10. The few randomizations where they do overlap reduce the expected footprint to 9.606. The normalized footprint is thus  $10/9.606 = 1.04$ . The expected footprint of the five long intervals is 40.49. This is somewhat less than 50, the sum of the interval sizes, since in many randomizations the intervals overlap. For these five long intervals, the normalized footprint is thus  $10/40.49 = 0.247$ , less than a quarter of that for the five short intervals.

*General case (possibly disconnected profiles)*

We now state our general approach to the problem. Let  $\mathcal{C}$  be any collection of  $k$  (possibly disconnected) IBD profiles. For each  $m = 2, 3, \dots, k$ , we define an  $m$ -sample of  $\mathcal{C}$  as a subset of  $m$  IBD intervals from  $\mathcal{C}$ , where no two intervals are from the same profile.

If  $\mathcal{S}$  is an  $m$ -sample of  $\mathcal{C}$ , we define its normalized footprint by the same formula as in (1). Here, as before,  $F(\mathcal{S})$  is the number of clones covered by at least one interval in  $\mathcal{S}$  and  $E(F)$  is the null hypothesis expected value of  $F$ , obtained by considering all allowable rearrangements of the  $m$  profiles containing the intervals in  $\mathcal{S}$ . The calculation of  $E(F)$  is described in the theorem in Appendix B.

For each  $m = 2, 3, \dots, k$ , let the statistic  $NF_m(\mathcal{C})$  be the *smallest observed* normalized footprint over all  $m$ -samples of  $\mathcal{C}$ . Then this observed statistic will have a  $p$ -value, determined by the randomization distribution of  $NF_m$ . Call this  $p$ -value  $X_m(\mathcal{C})$ . We now calculate in turn  $X_2(\mathcal{C}), X_3(\mathcal{C}), \dots, X_k(\mathcal{C})$ , and from this we calculate  $X(\mathcal{C})$ , the minimum of these  $p$ -values. The significance of the observed minimum  $p$ -value can be obtained by referring it to the randomization distribution of  $X$ . We define the data to be significant if the observed value  $X(\mathcal{C})$  is significant (at the desired level) as determined by this distribution.

More generally, for any  $m = 2, 3, \dots, k$  and for any  $m$ -sample  $\mathcal{S}$  of  $\mathcal{C}$ , let  $Y(\mathcal{S})$  be the  $p$ -value of  $NF(\mathcal{S})$  with respect to the randomization distribution of  $NF_m$ . Then we (conservatively) define the  $m$ -sample  $\mathcal{S}$  to be significant if  $Y(\mathcal{S})$  is significant (at the desired level) with respect to the

randomization distribution of  $X$ . If  $\mathcal{S}$  is significant, the union of clones covered by at least one interval in  $\mathcal{S}$  indicates a significant region (or regions if disconnected).

Every significant sample leads to a significant region (or significant regions) in the genome. For any such sample, those points over which there are the most intervals should be considered the most likely regions for the location(s) of the disease gene(s) within the significant region.

This solves the problem of defining the significance of the data and of the samples, and given an arbitrary amount of computing power we could calculate this significance. In practice, however, since the number of samples grows exponentially with  $k$ , an approximation becomes necessary in the calculation of the  $NF_m(\mathcal{C})$  when  $k$  is large. We overcome this by using an algorithm which involves a heuristic step. This algorithm, which is linear in  $k$ , is described in detail in the next section. The soundness of the heuristic step is also discussed.

### *Implementation and Applications*

We now describe our heuristic for approximating  $NF_m(\mathcal{C})$ , where  $\mathcal{C}$  is an arbitrary collection of  $k$  IBD profiles. If  $k$  is large, it is not feasible to consider all the samples of  $\mathcal{C}$ . The basis of the heuristic algorithm we use to circumvent this problem is that if  $\mathcal{S}$  is an  $m$ -sample of  $\mathcal{C}$  having the smallest normalized footprint over all  $m$ -samples of  $\mathcal{C}$ , then it will not in general be the case that *all*  $(m-1)$ -subsamples of  $\mathcal{S}$  will have (relatively) large normalized footprints. Based on this observation we can work our way up to  $NF_m(\mathcal{C})$  as follows. Let  $m < k$  and let  $\mathfrak{S}$  be a set of  $m$ -samples of  $\mathcal{C}$ . Denote by  $\mathfrak{B}(\mathfrak{S})$  the set of all  $(m+1)$ -samples of  $\mathcal{C}$  which contain a sample  $\mathcal{S} \in \mathfrak{S}$ . For  $N \geq 1$  and  $\mathfrak{S}$  a set of samples of  $\mathcal{C}$ , denote by  $\mathfrak{S}_N$  the subset of the  $N$  samples in  $\mathfrak{S}$  with the smallest normalized footprints. The algorithm goes as follows:

- Choose ‘cutoffs’  $N \geq 1$  and  $r \geq 2$  (integers to be determined later).
- For  $i = 2, 3, \dots, r, k-r, k-r+1, \dots, k$  compute  $NF_i(\mathcal{C})$  (using (1) and (B1)).
- Let  $\mathcal{S}(r)$  be the set of *all*  $r$ -samples of  $\mathcal{C}$ , and, for  $i = r+1, r+2, \dots, k-r-1$ , let  $\mathcal{S}(i) = \mathfrak{B}(\mathcal{S}(i-1)_N)$ .
- For  $i = r+1, r+2, \dots, k-r-1$ , compute  $NF_i^*(\mathcal{C}) = \min \{NF(\mathcal{S}) : \mathcal{S} \in \mathcal{S}(i)\}$  and use this as an approximation of  $NF_i(\mathcal{C})$ .

The larger  $N$  is, the better an approximation  $NF_i^*(\mathcal{C})$  is for  $NF_i(\mathcal{C})$ . Note that this is a heuristic step. However, experiments that we have run indicate that the algorithm fails on less than 0.1% of the values. Furthermore, the mean of the relevant statistic over samples for which this algorithm fails is higher than the mean over of all samples, and the standard deviation is lower. For the general algorithm to be adversely affected it must fail on one of the small number of values at the lowest extreme (usually one or two values) which, in light of this, appears to be of negligible likelihood. One should choose  $N$  and  $r$  as large as possible, still allowing for completion of the calculation in a reasonable amount of time. For a configuration  $\mathcal{C}$  of  $k = 35$  connected IBD profiles over a genome length of  $L = 100$ , we found that for  $r = 3$  and  $N = 1000$ , the calculation of all the  $NF_m^*(\mathcal{C})$  takes about 20 seconds when efficiently coded in C and run on a 333 Mhz Pentium II.

We now consider the quantity

$$X(\mathcal{C}) = \min_{2 \leq m \leq k} X_m(\mathcal{C})$$

defined in the previous section. Here we use the standard procedure of substituting  $z$ -scores for  $p$ -values. For a fixed  $m$ , we define  $z_m(\mathcal{C})$  as the minimum  $z$ -score of the normalized footprints over the

$m$ -samples of  $\mathcal{C}$ . As the randomization distributions for  $NF_m$  are not expected to vary significantly from  $m = i$  to  $m = i + 1$ ,  $\text{argmin}_m z_m(\mathcal{C})$  should correspond closely to  $\text{argmin}_m X_m(\mathcal{C})$ . Let

$$z(\mathcal{C}) = \min_{2 \leq m \leq k} z_m(\mathcal{C}).$$

In order to compute  $z_m(\mathcal{C})$  and the  $p$ -value of  $z(\mathcal{C})$ , we need to know the mean and variance of  $NF_m$  for each  $m = 2, 3, \dots, k$  and the randomization distribution of  $z$ . We have to use empirical distributions (and maximum likelihood estimators for the relevant means and variances), except for the randomization distribution for  $NF_k$  when all profiles are connected (for this we can compute exactly mean and variance with the help of (1), (B 1), and (B 2)). These empirical distributions were generated by simulating (via a C program which also uses the algorithm described above) random rearrangements of the  $k$  IBD profiles.

We have run this program for  $k = 35$  connected IBD profiles,  $L = 100$ ,  $N = 1000$ , and  $r = 3$ , using 20 000 random rearrangements to get empirical distributions of all the  $NF_m$  and then 10 000 random rearrangements to get an empirical distribution for  $z$ . (This took about 167 h on a 333 Mhz Pentium II.) For the collection  $\mathcal{C}$  as in Figure 2,  $z(\mathcal{C})$  was smaller than  $z(\mathcal{R})$  for all but three of the 10 000 random rearrangements  $\mathcal{R}$  that our program generated. We thus estimate a  $p$ -value for these data of  $3 \times 10^{-4}$ , and further calculation shows that the  $p$ -value is less than  $10^{-3}$  with probability 0.99. Moreover, the 7-sample above the arrow in Figure 2 is the only sample, significant at this level, identified by our algorithm.

We have also run our algorithm on the data of Cheung *et al.* (1998) from Figure 1. Here  $k = 8$ ,  $L = 96$ , and  $n = 72$ . We used 20 000 random rearrangements to get empirical distributions of all  $NF_m$  and then 1 000 000 random rearrangements to get an empirical distribution for  $z$ . We did not make use of  $N$  and  $r$ , since  $k = 8$  is small enough as to allow an exhaustive scanning of all samples. (The running time on a 333 Mhz Pentium II was about 42 h.) For the collection  $\mathcal{C}$  as in Figure 1,  $z(\mathcal{C})$  was smaller than  $z(\mathcal{R})$  for all of the 1 000 000 random rearrangements that our program generated. Using this, we have calculated that the  $p$ -value of these data is less than  $10^{-5}$  with probability 0.9999. Moreover the most significant samples identified by our algorithm consist of eight intervals and specify the region from clone 27 to clone 35 inclusive, agreeing with our visual intuition.

Software for analysis of IBD data will be made available to academic and non-profit institutions upon request to the authors.

#### DISCUSSION

We have introduced a method which uses a novel statistic, the normalized footprint, to detect significant alignments of segments of the genome which are shared IBD between pairs of individuals. Significant alignments identify regions in the genome where the gene(s) responsible for a given disease is (are) likely to be. We showed how this statistic overcomes some of the limitations of another statistic, the 'max' statistic, which has been used in the past to analyze IBD data.

Our method is appropriate to study data from pairs of individuals whose relationship is unknown, thus usually distant. This is in contrast with methods that have been proposed in the literature, where the use of pairs of individuals of known relationship is assumed. Since it is anticipated that an increasing volume of data from individuals whose relationship is unknown will be generated using direct IBD mapping, a method to deal with this situation is necessary. In any case, our method is valid also when the relationship between individuals in a pair is known.

When analyzing data for individuals whose relationships are known, natural independent pairs can be formed. One limitation of our approach is that, when applied to individuals whose relationships are not known, natural independent pairs cannot be formed. So the choice of the independent pairs

introduces some arbitrariness. To eliminate this arbitrariness one would be forced to use dependent pairs (for example all possible pairings); however this would make the analysis considerably more complicated. The normalized footprint would be an appropriate statistic to use in this case also, however null hypothesis distributions would have to be investigated. This seems likely to add significant complications.

Note that power calculations in the literature, as for example those in Smalley *et al.* (1996), rely upon knowledge of the relationship of the individuals in a pair. We cannot do this. Power also relies on a variety of other parameters. For example Smalley *et al.* (1996) point out that mode of inheritance (a complex concept and possibly not even meaningful for complex diseases), phenocopy rate, disease gene frequency, the recombination fractions and the number of clones, all influence power. So there is no unique meaningful power curve.

However, the use of unrelated individuals, for which our method is applicable, leads to increased power since these individuals tend to have small IBD profiles. In standard linkage analysis applied to complex diseases the problems associated with IBD methods (e.g. affected sib pairs) are now being approached by using tests based on association (e.g. the TDT). Our procedure has many points of similarity with the association approach and indeed was motivated by it. We thus believe that, in a general way, our methods should have good power properties.

#### APPENDIX A

Given  $k$  IBD profiles of respective sizes  $\ell_1, \ell_2, \dots, \ell_k$ , let

$$P = \prod_{i=1}^k (L - \ell_i + 1), \quad (\text{A } 1)$$

and for  $u = 1, 2, \dots, L$ , let

$$N_1(u) = \prod_{i=1}^k \min \{u, \ell_i, L - \ell_i + 1, L + 1 - u\}, \quad (\text{A } 2)$$

and

$$N_2(u) = \prod_{i=1}^k \min \{u, \ell_i - 1, L - \ell_i + 1, L - u\}. \quad (\text{A } 3)$$

**Theorem A.** *If each profile is connected, then*

$$\text{Prob}(M = k) = \frac{1}{P} \left( \sum_{u=1}^L N_1(u) - \sum_{u=1}^{L-1} N_2(u) \right), \quad (\text{A } 4)$$

and

$$\begin{aligned} \text{Prob}(M = k - 1) = & \frac{1}{P} \left[ \sum_{r=1}^k (L - \ell_r + 1) \left( \sum_{u=1}^L \prod_{i \neq r} m_1(u, i) - \sum_{u=1}^{L-1} \prod_{i \neq r} m_2(u, i) \right) \right. \\ & - k \left( \sum_{u=1}^L \prod_{i=1}^k m_1(u, i) - \sum_{u=1}^{L-1} \prod_{i=1}^k m_2(u, i) \right) \\ & \left. - \sum_{1 \leq r < s \leq k} \delta_{rs} \sum_{t=0}^{t_{rs}} \left( \sum_{u=\ell_r}^{L-\ell_s-t} \prod_{i \neq r, s} m(u, i, t) + \sum_{u=\ell_s}^{L-\ell_r-t} \prod_{i \neq r, s} m(u, i, t) \right) \right], \quad (\text{A } 5) \end{aligned}$$

$$\begin{aligned} \text{where } \delta_{rs} &= \begin{cases} 0 & \text{if } \min\{\ell_i : i \neq r, s\} = 1 \\ 1 & \text{if } \min\{\ell_i : i \neq r, s\} > 1, \end{cases} \\ m_1(u, i) &= \min\{u, \ell_i, L - \ell_i + 1, L - u + 1\}, \\ m_2(u, i) &= \min\{u, \ell_i - 1, L - \ell_i + 1, L - u\}, \\ \text{and } m(u, i, t) &= \min\{u, \ell_i - t - 1, L - t - u, L - \ell_i + 1\}. \end{aligned}$$

**Proof of (A 4).** For  $u = 1, 2, \dots, L$  let  $A_u$  be the event that the  $u$ -th clone in the genome is covered by all  $k$  IBD profiles. Then the event ' $M = k$ ' is the same as the event that at least one of  $A_1, A_2, \dots, A_L$  occurs. For each  $1 \leq u_1 < u_2 < \dots < u_h \leq L$ , let  $A_{u_1}A_{u_2} \cdots A_{u_h}$  be the intersection of  $A_{u_1}, A_{u_2}, \dots, A_{u_h}$ . Since each profile is connected, for  $1 \leq u_1 < u_2 < \dots < u_h \leq L$  the event  $A_{u_1}A_{u_2} \cdots A_{u_h}$  is the same as the event  $A_{u_1}A_{u_h}$ . Using this fact and standard inclusion-exclusion formulae (Feller (1968)), after much telescoping and cancellation of terms, we get

$$\text{Prob}(M = k) = \sum_{u=1}^L \text{Prob}(A_u) - \sum_{u=1}^{L-1} \text{Prob}(A_uA_{u+1}). \quad (\text{A } 6)$$

The quantity  $P$ , given by (A 1), is easily seen to be the number of all possible rearrangements of the various IBD profiles. Similarly,  $N_1(u)$ , given by (A 2), is the number of all possible rearrangements of the profiles for which clone  $u$  is covered by all profiles, and  $N_2(u)$ , given by (A 3), is the number of all possible rearrangements of the profiles for which both clones  $u$  and  $u + 1$  are covered by all profiles. Therefore, if the profiles are placed randomly, the ratio  $N_1(u)/P$  is  $\text{Prob}(A_u)$  and  $N_2(u)/P$  is  $\text{Prob}(A_uA_{u+1})$ . This concludes the proof of (A 4).  $\square$

Note that a central component of the above proof is the observation that the event  $A_{u_1}A_{u_2} \cdots A_{u_h}$  is identical to the event  $A_{u_1}A_{u_h}$ . This identity holds because of the assumption that each IBD profile consists of a contiguous sequence of clones, that is, consists of one interval. When this assumption does not hold, the simplifications in the calculation do not arise.

**Proof of (A 5).** For any rearrangement, we denote the profiles by  $\mathfrak{s}_1, \mathfrak{s}_2, \dots, \mathfrak{s}_k$ . Let  $B_r$  be the event that  $M$  is  $k - 1$  and  $\mathfrak{s}_r$  is superfluous to achieve this maximum. For each  $1 \leq r_1 < r_2 < \dots < r_h \leq k$ , let  $B_{r_1}B_{r_2} \cdots B_{r_h}$  be the intersection of  $B_{r_1}, B_{r_2}, \dots, B_{r_h}$ . Given  $r < s$ , if  $B_rB_s$  occurs, then there is no position in the genome covered by both  $\mathfrak{s}_r$  and  $\mathfrak{s}_s$  (if there were, then  $M$  would be  $k$ ). It follows that  $\text{Prob}(B_{r_1}B_{r_2} \cdots B_{r_h}) = 0$  for each  $h > 2$ . So

$$\text{Prob}(M = k - 1) = \sum_{r=1}^k \text{Prob}(B_r) - \sum_{1 \leq r < s \leq k} \text{Prob}(B_rB_s). \quad (\text{A } 7)$$

So we need to compute  $\text{Prob}(B_r)$  and  $\text{Prob}(B_rB_s)$ . Let  $P$  be as in (A 1) and let  $C_r$  be the event that there is a position on the genome covered by all the  $k - 1$  IBD profiles different from  $\mathfrak{s}_r$ . Then

$$\text{Prob}(B_r) = \text{Prob}(C_r) - \text{Prob}(M = k). \quad (\text{A } 8)$$

We already have a formula for  $\text{Prob}(M = k)$ . Using the same formula with  $k$  replaced by  $k - 1$  we get

$$\begin{aligned} \text{Prob}(C_r) &= \frac{(L - \ell_r + 1)}{P} \times \left( \sum_{u=1}^L \prod_{i \neq r} \min\{u, \ell_i, L - \ell_i + 1, L - u + 1\} \right. \\ &\quad \left. - \sum_{u=1}^{L-1} \prod_{i \neq r} \min\{u, \ell_i - 1, L - \ell_i + 1, L - u\} \right). \quad (\text{A } 9) \end{aligned}$$

By replacing these two formulas in (A 8), we get

$$\text{Prob}(B_r) = \frac{1}{P} \left[ (L - \ell_r + 1) \times \left( \sum_{u=1}^L \prod_{i \neq r} \min\{u, \ell_i, L - \ell_i + 1, L - u + 1\} \right. \right. \\ \left. \left. - \sum_{u=1}^{L-1} \prod_{i \neq r} \min\{u, \ell_i - 1, L - \ell_i + 1, L - u\} \right) - \left( \sum_{u=1}^L N_1(u) - \sum_{u=1}^{L-1} N_2(u) \right) \right], \quad (\text{A } 10)$$

where  $N_1(u)$  and  $N_2(u)$  are defined in (A 2) and (A 3). It remains to compute  $\text{Prob}(B_r B_s)$ . To this end, note that  $B_r B_s$  is the event that there is no clone in the genome covered by both  $\mathfrak{s}_r$  and  $\mathfrak{s}_s$  and the remaining  $k - 2$  IBD profiles cover the position (clone) on the genome covered by the right endpoint of the leftmost of these two IBD profiles and the position occupied by the left endpoint of the rightmost one. Also note that  $P(B_r B_s) = 0$  if  $\min\{\ell_i : i \neq r, s\} = 1$ . Hence it remains to consider the case in which  $\min\{\ell_i : i \neq r, s\} > 1$ . In such a case, set  $t_{rs} = \min\{\ell_i : i \neq r, s\}$  and, for each  $t = 0, \dots, t_{rs}$ , let  $S_{r,s}^t$  be the event in which  $\mathfrak{s}_r$  falls to the left of  $\mathfrak{s}_s$  and the number of clones on the genome strictly between the right endpoint of  $\mathfrak{s}_r$  and the left endpoint of  $\mathfrak{s}_s$  is  $t$ , and all of the remaining  $k - 2$  profiles cover both the right endpoint of  $\mathfrak{s}_r$  and the left endpoint of  $\mathfrak{s}_s$ . Similarly, denote by  $D_{r,s}^t$  the event in which  $\mathfrak{s}_r$  falls to the right of  $\mathfrak{s}_s$  and the number of clones on the genome strictly between the right endpoint of  $\mathfrak{s}_s$  and the left endpoint of  $\mathfrak{s}_r$  is  $t$ , and all the remaining  $k - 2$  profiles cover both the right endpoint of  $\mathfrak{s}_s$  and the left endpoint of  $\mathfrak{s}_r$ . Then  $B_r B_s$  is the disjoint union, over all  $t$  from 0 to  $t_{rs}$ , of the  $S_{r,s}^t$  and the  $D_{r,s}^t$ . Thus the number of ways in which  $B_r B_s$  occurs is the sum of two terms. The first is the sum, over  $t = 0, 1, \dots, t_{rs}$ , of the number of ways in which  $S_{r,s}^t$  occurs. The second is the sum, over  $t = 0, 1, \dots, t_{rs}$ , of the number of ways in which  $D_{r,s}^t$  occurs. For  $S_{r,s}^t$  to occur, the right endpoint of  $\mathfrak{s}_r$  must cover a position  $u$  on the genome between  $\ell_r$  and  $L - \ell_s - t$ . Given that this happens and that the left endpoint of  $\mathfrak{s}_s$  is  $t + 1$  positions to the right of  $\mathfrak{s}_r$ ,  $S_{r,s}^t$  will occur if and only if each of the remaining  $k - 2$  profiles covers both the right endpoint of  $\mathfrak{s}_r$  and the left endpoint of  $\mathfrak{s}_s$ . For each of the remaining profiles, if its size is  $\ell_i$ , then the number of ways in which it will do this is  $\min\{u, \ell_i - t - 1, L - u - t, L - \ell_i + 1\}$ . So the number of ways in which  $S_{r,s}^t$  can occur is

$$\sum_{u=\ell_r}^{L-\ell_s-t} \prod_{i \neq r, s} \min\{u, \ell_i - t - 1, L - t - u, L - \ell_i + 1\} \quad (\text{A } 11)$$

In a similar fashion one gets that the number of ways in which  $D_{r,s}^t$  can occur is

$$\sum_{u=\ell_s}^{L-\ell_r-t} \prod_{i \neq r, s} \min\{u, \ell_i - t - 1, L - t - u, L - \ell_i + 1\} \quad (\text{A } 12)$$

Substituting (A 10), (A 11) and (A 12) in (A 7) we get (A 5).  $\square$

#### APPENDIX B

**Theorem B.** Let  $\mathcal{C}$  be a collection of IBD profiles and let  $\mathcal{S}$  be an  $m$ -sample of  $\mathcal{C}$ . Denote by  $\mathfrak{s}_1, \mathfrak{s}_2, \dots, \mathfrak{s}_m$  the intervals in  $\mathcal{S}$ . Then

$$E(F) = \sum_{u=1}^L \left( 1 - \prod_{i=1}^m p_i(u) \right) \quad (\text{B } 1)$$

and

$$\text{Var}(F) = \sum_{u=1}^L \left[ \left( \prod_{i=1}^m p_i(u) \right) \left( 1 - \prod_{i=1}^m p_i(u) \right) \right] + 2 \sum_{u < v} \left[ \prod_{i=1}^m p_i(u, v) - \prod_{i=1}^m (p_i(u) p_i(v)) \right], \quad (\text{B } 2)$$

where  $p_i(u)$  denotes the probability that  $\mathfrak{s}_i$  does not cover clone  $u$ , and  $p_i(u, v)$  denotes the probability that  $\mathfrak{s}_i$  does not cover either clone  $u$  or clone  $v$ .

**Sketch of proof.** Consider, for each  $u$  from 1 to  $L$ , the indicator random variable  $F_u$  which is defined to be 1 if the point (clone)  $u$  in the genome is covered by at least one of the  $\mathfrak{s}_i$  and 0 otherwise. So  $F = \sum_{u=1}^L F_u$ .

Let  $p(u)$  be the probability that none of the  $m$  IBD intervals covers the point  $u$ , i.e.  $p(u) = \text{Prob}(F_u = 0)$ . The intervals are placed independently, so

$$p(u) = \prod_{i=1}^m p_i(u) \quad (\text{B 3})$$

Since  $\text{E}(F_u) = 0 \cdot p(u) + 1 \cdot (1 - p(u))$ , we have

$$\text{E}(F) = \sum_{u=1}^L \text{E}(F_u) = \sum_{u=1}^L (1 - p(u)).$$

Substituting (B 3) in the latter equation, we get (B 1). Moreover,

$$\text{Var}(F_u) = \text{E}(F_u^2) - \text{E}(F_u)^2 = 1 - p(u) - (1 - p(u))^2 = p(u)(1 - p(u)). \quad (\text{B 4})$$

Next, we compute  $\text{E}(F_u F_v)$ , for  $1 \leq u < v \leq L$ .  $F_u F_v = 1$  if and only if  $F_u = 1 = F_v$ , and  $F_u F_v = 0$  otherwise. Then

$$\begin{aligned} \text{Prob}(F_u F_v = 0) &= \text{Prob}(F_u = 0) + \text{Prob}(F_v = 0) - \text{Prob}(F_u = 0 = F_v) \\ &= p(u) + p(v) - \text{Prob}(F_u = 0 = F_v). \end{aligned} \quad (\text{B 5})$$

The probability  $p(u, v)$  that none of the  $m$  IBD intervals covers  $u$  or  $v$ , i.e.  $p(u, v) = \text{Prob}(F_u = 0 = F_v)$ , is given by

$$p(u, v) = \prod_{i=1}^m p_i(u, v), \quad (\text{B 6})$$

as the intervals are independent.

Thus, since  $\text{E}(F_u F_v) = 1 - \text{Prob}(F_u F_v = 0)$ , we have

$$\text{Cov}(F_u, F_v) = \text{E}(F_u F_v) - \text{E}(F_u)\text{E}(F_v) = 1 - p(u) - p(v) + p(u, v) - (1 - p(u))(1 - p(v)) = p(u, v) - p(u)p(v).$$

Using this and (B 4), we get

$$\text{Var}(F) = \sum_{u=1}^L \text{Var}(F_u) + 2 \sum_{u < v} \text{Cov}(F_u F_v) = \sum_{u=1}^L p(u)(1 - p(u)) + 2 \sum_{u < v} (p(u, v) - p(u)p(v)). \quad (\text{B 7})$$

Substituting (B 3) and (B 6) in this equation, we obtain (B 2).  $\square$

If  $\mathfrak{s}_i$  comes from a *connected* IBD profile in  $\mathcal{C}$ , and if  $\ell_i$  is the size of  $\mathfrak{s}_i$ , then

$$p_i(u) = \frac{\max\{L - \ell_i + 1 - u, L - 2\ell_i + 1, u - \ell_i, 0\}}{L - \ell_i + 1} \quad (\text{B 8})$$

and

$$p_i(u, v) = \frac{1}{L - \ell_i + 1} \cdot \max\{L - \ell_i + 1 - v, L - 2\ell_i + 1 - u, v - u - \ell_i, L - 2\ell_i + 1 - v + u, 0, L - 3\ell_i + 1, v - 2\ell_i, u - \ell_i\}. \quad (\text{B 9})$$

This can be checked by careful counting.

If  $\mathfrak{s}_i$  comes from a *disconnected* IBD profile whose size is small compared to  $L$ , then (B 8) and (B 9)

are very good approximations, for reasons discussed at the end of the Null Hypothesis section. In any case,  $p_i(u)$  and  $p_i(u, v)$  can be obtained computationally.

We thank Richard S. Spielman, from the Department of Genetics of the University of Pennsylvania, for many important suggestions. We also thank Christian G. Overton, the director of the Computational Biology and Informatics Laboratory in the Center for Bioinformatics of the University of Pennsylvania, for his support. Finally, we thank the referees for their comments.

## REFERENCES

- BROWN, P. O. (1994). Genome scanning methods. *Curr. Opin. Genet. Dev.* **4**, 366–373.
- CHEUNG, V. G., GREGG, J. P., GOGOLIN-EWENS, K. J. *et al.* (1998). Linkage-disequilibrium mapping without genotyping. *Nat. Genet.* **18**, 225–230.
- CHEUNG, V. G. AND NELSON, S. F. (1998). Genomic mismatch scanning identifies human genomic DNA shared identical by descent. *Genomics* **47**, 1–7.
- DURHAM, L. K. AND FEINGOLD, E. (1997). Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am. J. Hum. Genet.* **61**, 830–842.
- FEINGOLD, E., BROWN, P. O., SIEGMUND, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.* **53**, 234–251.
- FELLER, W. (1968). *An introduction to probability theory and its applications*, vol. 1 (3rd edn). New York: J. Wiley & Sons.
- GUO, S. W. (1995). Proportion of genome shared identical by descent by relatives: concept, computation, and applications. *Nat. Genet.* **4**, 11–18.
- HOUWEN, R. H. J., BAHARLOO, S., BLANKENSHIP, K. *et al.* (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.* **8**, 380–386.
- MCALLISTER, L., PENLAND, L., BROWN, P. O. (1998). Enrichment for loci identical-by-descent between pairs of mouse or human genomes by genomic mismatch scanning. *Genomics* **47**, 8–14.
- MIRZAYANS, F., MEARS, A. J., GUO, S. W. *et al.* (1997). Identification of the human chromosomal region containing the iridogoniodysgenesis anomaly locus by genomic-mismatch scanning. *Am. J. Hum. Genet.* **61**, 111–119.
- NELSON, S. F., MCCUSKER, J. H., SANDER, M. A. *et al.* (1993). Genomic mismatch scanning: a new approach to genetic linkage mapping. *Nat. Genet.* **4**, 11–18.
- RISCH, N. AND MERIKANGAS, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.
- SMALLEY, S. L., WOODWARD, J. A., PALMER, C. G. (1996). A general statistical model for detecting complex-trait loci by using affected relative pairs in a genome search. *Am. J. Hum. Genet.* **58**, 844–860.
- THOMAS, A., SKOLNICK, M. H., LEWIS, C. M. (1994). Genomic mismatch scanning in pedigrees. *IMA J. Math. Appl. Med. Biol.* **11**, 1–16.