Published in final edited form as:

Genet Epidemiol. 2010 February ; 34(2): 119–124. doi:10.1002/gepi.20440.

Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium

Alun Thomas*

Department of Biomedical Informatics University of Utah

Abstract

We describe methods and programs for simulating the genotypes of individuals in a pedigree at large numbers of linked loci when the alleles of the founders are under linkage disequilibrium. Both simulation and estimation of linkage disequilibrium models are shown to be feasible on a genome wide scale. The methods are applied to evaluating the statistical significance of streaks of loci at which sets of related individuals share a common allele. The effects of properly allowing for linkage disequilibrium are shown to be important as they explain many of the large observations. This is illustrated by re analysis of a previously reported linkage of prostate cancer to chromosome 1p23.

Keywords

Genetic mapping; graphical modelling; identity by descent

1 Introduction

Dense single nucleotide polymorphism, or *SNP*, genotype assays offer both great opportunities and challenges to statistical genetics. The quantity, quality, and cheapness of the assays can allow powerful methods for precise gene localizations. However, the quantity of the data makes more involved forms of analysis intractable on a genome wide scale and the density of the loci assayed requires detailed modelling of the structure of the chromosomes. In particular, linkage disequilibrium, or *LD*, could for the most part be safely neglected when using micro satellite genotypes, whereas it is critical to model it accurately with dense SNP assays.

Although these challenges are encountered with population samples, they are even more problematic with pedigree data. Linkage analysis is a powerful statistical method that has been used very successfully to map disease genes, but it is computationally intense and intractable to carry out on one million marker genotype data. Moreover, adapting it to model LD greatly increases the computational requirements (Thomas 2007). This has led to interest in simpler, more tractable analyses for family data that rely on the quantity of the data rather on statistical efficiency for their power. Thomas et al. (2008) introduced a mapping strategy based on counting long runs of loci at which individuals share alleles identical by state (*IBS*). Leiby et al. (2008) also considered the same statistic which they called the *SNP streak* statistic. Long runs of shared alleles should be rare in unrelated individuals, whereas in relatives within a pedigree, or in samples from founder populations, these long streaks indicate underlying sharing of genomic segments identical by descent

* Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA. alun@genepi.med.utah.edu, +1 801 587 9303 (voice), +1 801 581 6052 (fax)..

(IBD) from a common ancestor. If the individuals sharing are selected to have some trait, and they are sufficiently distantly related that probability of random sharing is very low, then there is strong evidence that the shared segment contains a gene affecting the trait.

Houwen et al. (1994) and Heath et al. (2001) both studied relatively isolated founder populations to identify a small number of related cases who shared common chromosomal segments which they used to map disease genes. However, neither of these approaches incorporated precise pedigree relationships between the cases. Chapman & Thompson (2002) and te Meerman & Van der Meulen (1997) examined shared chromosomal segments in a founder population and showed how these are affected by the time since the founding of the population, population growth, genetic drift, selection and population subdivision. The streak statistic is similar to the haplotype sharing statistics of Van der Meulen & te Meerman (1997) and Beckmann et al. (2005), however, these more complicated statistics require genotyping of close relatives to estimate phase, are based on combining pairwise comparisons, and are applied in populations samples rather than in extended pedigrees. Bourgain et al. (2001) applied similar methods to extended pedigrees, but this again required knowing phase and combining pairwise comparisons. Other streak statistics have also been suggested, for example, Miyazawa et al. (2007) considered streaks of SNP loci at which individuals share homozygously. Other approaches exploiting IBD in pedigrees includes the homozygosity-by-descent method of Abney et al. (2002) which has been used to map recessive traits in known, very large, inbred populations (Newman et al. 2003).

Genomic sharing in a pedigree was modelled by Donnelly (1983) as a random walk on the vertices of a hypercube from which the distribution of number and length of genomic segments shared genome wide by an arbitrary set of relatives can be obtained. Cannings (2003) also derived results for this model. These results, however, describe the underlying process and do not account for observed genetic data. Both Thomas et al. (2008) and Leibon et al. (2008) addressed this problem using the streak statistic, although they differed in how they evaluated the statistical significance of observed streaks. Leibon et al. (2008) extended the theoretical results of Miyazawa et al. (2007) for homozygous sharing to the case where only one haplotype is shared, whereas, Thomas et al. (2008) used multi locus gene drop simulation. Both methods assumed that the loci being assessed were in linkage equilibrium, but both sets of authors also acknowledged that this assumption was inappropriate and likely to lead to underestimates of any p-value, and possible false positive results. It is this limitation that we seek to address here.

Thomas (2009a) and Thomas (2009b) developed and described methods and programs for modelling LD using restricted types of graphical models. These models are tractable so that both estimating them and simulating from them can be done on hundreds of thousands of loci. Given appropriate controls we can estimate LD models on a genome wide scale, and given a pedigree structure we can simulate founder haplotypes from the models and use the multi locus gene drop method to simulate genotypes for the entire pedigree. In what follows we apply these methods to a re analysis of the prostate cancer linkage to chromosome 1p23 reported by Camp et al. (2005) and previously analyzed using SNP streaks by Thomas et al. (2008). We show that appropriate modelling of LD is not only feasible on this scale, but that it is essential for accurately assessing statistical significance. While our focus here is on streak statistics observed in extended pedigrees, we note that the simulation method can also be applied to evaluate the significance the other statistics mentioned above, and to sampling designs using parent-offspring triplets, nuclear families or independent samples.

2 Materials and methods

2.1 SNP streak statistics

Figure 1 shows a pedigree connecting 8 men with prostate cancer. These individuals were selected from a larger extended pedigree from a study of prostate cancer in Utah. These cases are connected by a total of 15 meioses to each of 2 recent common ancestors. Assuming that the total genetic length of the 22 autosomes is 35 Morgans (Broman et al. 1998), the probability that all 8 cases share any genetic segments IBD is 0.067 (Thomas et al. 2008).

Each of the 8 cases was genotyped by the Center for Inherited Disease Research, using the Illumina 110K panel (<http://www.illumina.com>). Of the total of 109299 loci analyzed, 9819 were on chromosome 1. In order to avoid spurious runs of sharing due to low heterozygosity, we discarded SNPs for which the heterozygosity score, as assessed from the controls described below, was less than 0.2. This left 8016 loci in the analysis, evenly spread over chromosome 1 with the exception of a large gap at the centromere.

At each locus i we counted the number of observations of each genotype: n_{11}^i , n_{12}^i and n_{22}^i so that $n_{11}^i + n_{12}^i + n_{22}^i \leq n=8$, with inequality when there were some missing genotypes.

Then we calculated the sharing statistic at the i th locus as

$$S_i = n - \min(n_{11}^i, n_{22}^i).$$

Then, again at each locus, we define $R_i(t)$ to be the longest run containing the locus for which the values of S_i are at least t . We took $t = n$ and $t = n - 1$ for this study, but sharing in smaller sets may be appropriate if more cases are considered.

These statistics were calculated using a program written by the author which, like all the programs described here, is freely available from the author's web site, <http://bioinformatics.med.utah.edu/~alun>. The program for calculating SNP streaks is named SGS, for shared genomic segments, and is called as follows:

```
java SGS input.par input.ped > output
```

where

- input.par is a LINKAGE format parameter file describing the genetic loci.
- input.ped is a LINKAGE format pedigree file giving the pedigree structure and the genotypes of any assayed individuals. The individuals among who sharing is to be counted must have their proband status set to 1.
- output is a text file with one line for each marker in the input. Each gives the marker name followed by S_i , $R_i(n)$, $R_i(n-1)$, $R_i(n-2)$, and $R_i(n-3)$.

The maxima of the $R_i(t)$ statistics over the whole of the data being analyzed are also output to the screen.

Full details of the LINKAGE file formats can be had on the web at <http://linkage.rockefeller.edu>. This format is intended primarily for linkage and segregation analysis and so the distances between loci are specified as recombination fractions. Since it is useful also to have the physical location of the marker, we make the local convention of encoding this in the marker's name so that it is printed as part of the output from SGS. Note

that all file names given here and below are just arbitrary examples, they are not required names.

2.2 LD model estimation

In order to estimate the LD structure of chromosome 1 we used the same control data that was described by Thomas et al. (2008). These are 52 Utah CEPH controls that were included as part of the 120 control sample set genotyped by Illumina for the same set of SNPs.

Graphical models (Lauritzen 1996) are a broad class of statistical models encompassing Bayesian networks, Markov random fields and probabilistic expert systems. They can tractably model complex relationships between variables. In particular, they can be efficiently estimated from data and used for simulation. The use of graphical models to model LD was first proposed by Thomas & Camp (2004) who developed a method of model estimation from phase known haplotype data. Thomas (2005) extended this to using unphased genotypic data by employing a two stage stochastic search approach. Given an initial model for LD, haplotypes are imputed conditional on the model and the genotype states. The imputed haplotypes are then used for re estimating the model using the Thomas & Camp (2004) method. The haplotypes are then re imputed, and so on. While this approach works well on moderately sized sets of loci, up to a few thousand, it does not scale well beyond this order of magnitude. Thomas (2009a), however, showed that restricting the class of models considered to that of interval graphs greatly improved efficiency without sacrificing modelling properties.

Thomas (2009b) describes an implementation of this approach called IntervalLD which employs a walking window approach so that the program's running time scales linearly with the number of loci being modelled. The program was run as follows:

```
java IntervalLD input.par control.ped > ldmodel
```

where

- input.par is the LINKAGE parameter file described above.
- control.ped is a LINKAGE pedigree file giving the genotypes of the control samples. The controls are unrelated so for each individual the parents are given as 0. If any relationships are specified in the control data, IntervalLD will ignore them and treat the observations as unrelated.
- ldmodel is a text file containing the estimated interval graphical model. The first line gives the number of alleles seen at each locus. This is followed by one line for each locus specifying the conditional distribution of alleles at the locus given the states at the loci that the program has estimated to be relevant. Although the file is human readable, it is primarily intended for input to other programs as described below.

2.3 Gene drop simulation with LD

Gene drop is a simple method for randomly generating genotypes for a set of related individuals. Alleles are randomly assigned to the founders of a pedigree and dropped at random to their offspring and other descendants mimicking Mendelian inheritance. The single locus method was described by MacCluer et al. (1986). The multi locus method is similar, but the inheritances at a genetic locus depend on those at the previous locus and the recombination fraction between them. An implementation of this is given, for instance, by the MERLIN program (Abecasis et al. 2001). The implementation given here differs only in that the alleles for the founders are simulated as haplotypes generated from the distribution specified by the given graphical model. The program SimSGS implements this to simulate

random genotypes to match those in a given input file. This includes matching the pattern of missing data: if a genotype is missing in the real data it will also be missing in the simulated. The program is used as follows:

```
java SimSGS input.par input.ped s ldmodel > output
```

where

- input.par is the same LINKAGE parameter file as used above with SGS and IntervalLD.
- input.ped is the same LINKAGE pedigree file as was used to obtain the observed statistics using SGS. This will specify the pedigree and the genotypes to simulate.
- s is the number of simulations to perform. This parameter is optional, the default value is 1000.
- ldmodel is a graphical model for LD as estimated above using IntervalLD. This is also an optional parameter. If it is omitted the allele frequencies given in input.par are used to simulate data using conventional multi locus gene drop under the assumption of linkage equilibrium.
- output is a text file containing one line for each simulation made. On each line are the maxima over loci i of $R_i(n)$, $R_i(n-1)$, $R_i(n-2)$, and $R_i(n-3)$ for that simulation.

We also have a program SimSGSRegions which is used with the same syntax as SimSGS but the file output gives the number of times each locus is contained in the maximum run simulated.

3 Results

Figure 2 plots (a) $R_i(8)$ and (b) $R_i(7)$ for our case data on chromosome 1. The longest run for all 8 sharing was 64, and for 7 from 8 sharing was 495 occurring at the position marked A in the plots. This corresponds to the location previously found using the same method by Thomas et al. (2008) and to the original linkage peak reported by Camp et al. (2005) for the same family. Other locations with long runs are marked B, C, D and E on these and following plots. The distribution of locus heterozygosity is shown in figure 2(c) as a

cumulative sum chart plotting $\sum_{j=1}^i (h_j - \bar{h})$ against the locus number i where

$h_j = 2p_j(1 - p_j)$, $\bar{h} = \frac{1}{n} \sum_{j=1}^n h_j$, and p_j is the observed frequency of the minor allele at the j th locus.

We made three sets of simulations with which to evaluate the statistical significance of the observations. Two sets of 10000 simulations were made under the assumption of linkage equilibrium, the first assuming allele frequencies of 0.5 for each allele at each locus, the second using allele frequencies estimated from the controls. The first of these approaches is clearly something of a straw man and should not be used in reality, however, it gives us a base line with which to compare subsequent simulations. The third set of simulations was made under LD. The model fitting program IntervalLD uses a stochastic search, so the estimated model will typically differ from run to run. To see whether this had any influence on the results we made 10 independent LD model estimates from the control data and simulated 1000 gene drops from each of them. For each set of simulations we obtained the cumulative probability distribution of the maximum run length for IBS sharing between all 8, and between 7 of 8 cases. These results are shown in figure 3.

Finally, for three sets of simulations as described above we recorded the probability that each locus was contained in the longest run of sharing. The probabilities are plotted against location in figure 4. Note that as the maximum run lengths increase, more loci are covered by them, hence, the areas under the curves shown in figure 4 changes. The locations of the peaks seen in figure 2 are also indicated in these plots.

4 Discussion

Several of the issues raised in the simulation analysis of Thomas et al. (2008) have been addressed here. While the previous simulation program worked on the physical locations of the markers, IntervallLD works on a genetic map as specified by the recombination fractions given in the linkage parameter file. This allows specification of recombination hot and cold spots. The file format also allows different maps for male and female recombinations and the program interprets and uses these appropriately.

More importantly, we are now able to model LD in the simulations. The example presented here had 8016 loci. The average time taken for the 10 runs to estimate models on these loci was just under 400 seconds. However, the program has been demonstrated to scale linearly with the number of loci. Thomas (2009b) ran the program on over 200000 loci for 60 control individuals taking just under 700 minutes to complete the model estimation. As the largest number of loci that need to be considered together in a genome wide analysis for 1 million loci are the approximately 100000 loci on chromosome 1, the program is well able to deal with this scale of analysis. Having estimated a graphical model for LD, the simulations themselves are reasonably quick. Making 1000 gene drops on the pedigree in figure 1 takes just under 300 seconds. Again, the gene drop program has been shown to scale linearly with the number of loci (Thomas 2009b). All the running times given here are for the author's HP laptop that runs Java 1.5.0_02-b09 under Linux. It has 4 Gbytes of memory and two 2.8 GHz central processing units.

Not only is large scale LD modelling feasible, it is also shown here to be necessary. The distributional shifts seen in figures 3 (a) and (b) due to LD are far greater than the shifts obtained by using realistic allele frequencies rather than assuming all alleles are equally frequent. The empirical p-value for a streak of 64 loci on chromosome 1 at which all individuals share an allele is 0.0037 under linkage equilibrium, but 0.034 under LD. The change for the run of 495 at which 7 of 8 share is less dramatic, from 0.01 to 0.012 which reflects the extreme nature of this observation: the length of sharing is beyond the influence of the LD. The empirical p-values under LD here are taken from combining the 10 samples of 1000 observations under the 10 different estimated LD models. However, the overlaid distributions in figure 3 show that although the models are typically different, the haplotypes estimated under them are very similar.

Figure 4 also demonstrates the dramatic effects of LD. Under linkage equilibrium with allele frequencies of 0.5, figure 4 (a) shows that the location of the longest simulated run length is uniformly distributed across the chromosome. When the control allele frequencies are used in the pedigree simulation, the longest run is more often at positions such as that marked B in the figures where there is a run of loci with low heterogeneity. The run of low heterogeneity at point B is demonstrated by the steep descent of the cumulative sum chart in figure 2 (c). Note also that since the longest runs are getting longer, more points are covered by them and there is, therefore, more area under the curve in figure 4 (b) than 4 (a). However, the peaks marked A, C, D and E in figure 2(a) do not correspond to points at which long runs are simulated under linkage equilibrium in figure 4 (b). The change from figure 4 (b) to (c) is even greater. We can now see that under LD, the peaks in the observed data at B, D and E all correspond to positions at which long runs of sharing are simulated.

These peaks cannot, therefore, be taken as evidence of significant sharing due to selection for a common phenotype.

The conclusions for the validity of the linkage of prostate cancer to chromosome 1p23 are still mixed. The p-value for the run of 64 loci at which all 8 cases share on chromosome 1 is 0.034. This is not significant when we allow for selection of the peak on chromosome 1 as the best of all those seen in a genome wide analysis. Neither is the run of 495 loci at which 7 from 8 share. A more detailed inspection of the data in this region shows that it is the same individual who does not share each side of the 64 shared by all. Although there is clearly an underlying genomic segment shared here IBD by these individuals, this is not sufficiently unusual in the pedigree to indicate that the sharing is due to selection for phenotype. On the other hand, we note that the longest runs are rarely simulated at position A. This is unlike positions B, D and E which are clearly peaks due to low heterozygosity and LD, and unlike position C where the run length is near the median of the distribution of the maximum.

Since this is a re analysis of the pedigree that gave the original lod score of 3.1 reported by Camp et al. (2005), it can not serve as an independent assessment of the finding. It does, however, give an indication that the the pattern of IBD sharing in the pedigree is consistent with linkage. For true confirmation for the result more data is needed.

The pedigree design used here is suited to detecting genes with dominant mode of expression. For recessive diseases inbred pedigrees could be used, as could random samples from an inbred population. In this case we would define streaks as runs of loci where individuals share both alleles in common, as described by Miyazawa et al. (2007). It is also informative to look for streaks of loci where sampled individuals are homozygous, but not necessarily for the same haplotype, as this may actually indicate hemizyosity and the presence of a chromosomal deletion. In either case, the statistical significance can be assessed by simulations that need to take LD into account. Thus, the estimation and simulation programs described here are directly applicable.

While this work goes some way to addressing the effects of LD on streak statistics, there are other issues. Perhaps the most pressing is the sensitivity of streak statistics to genotyping error. A single misclassification can, potentially, end a streak. The presence of two longer than average runs adjacent to each other, as occasionally seen in our analyses, would suggest that underlying these is an even longer run that has been broken up by genotyping error. If we take the S_i as our basic statistics, then the problem is essentially one of detecting the change points in their distribution: the more individuals that share a genomic segment, the higher S_i should be on average, even in the presence of genotyping error. One approach to this may be to replace the R_i statistics with more robust methods from the field of process control. For instance, cumulative sum methods, or CUSUM charts, could be used with significance again being assessed using simulation under LD. This is an approach that we wish to investigate in future work.

Acknowledgments

This work was supported by Grant Number R01 GM081417 to Alun Thomas from the NIH National Institutes of General Medical Sciences, and NIH National Cancer Institute grant R01 CA90752 and subcontract from Johns Hopkins University with funds provided by grant R01 CA89600 from the NIH National Cancer Institute to L A Cannon Albright. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of General Medical Sciences, National Cancer Institute or the National Institutes of Health. It was also supported by US Army Medical Research and Materiel Command W81XWH-07-1-0483 to Alun Thomas.

Research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the National Cancer Institute's SEER program with additional support from the Utah State Department of Health and the University of Utah

Partial support for all datasets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute.

This investigation was supported by the Public Health Services grant number M01-RR00064 from the National Center for Research Resources.

Genotyping services were provided by the Center for Inherited Disease Research which is fully funded through a federal contract from the NIH to the Johns Hopkins University, contract number N01-HG-65403.

References

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*. 2001; 30:97–101. [PubMed: 11731797]
- Abney M, Ober C, McPeck MS. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: Fasting serum-insulin levels in the Hutterites. *American Journal of Human Genetics*. 2002; 70:920–934. [PubMed: 11880950]
- Beckmann L, Thomas DC, Fischer C, Chang-Claude J. Haplotype sharing analysis using Mantel statistics. *Human Heredity*. 2005; 59:67–78. [PubMed: 15838176]
- Bourgain C, Genin E, Holopainen P, Musthlahti K, Maki M, Partanen J. Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *American Journal of Human Genetics*. 2001; 68:154–159. [PubMed: 11102286]
- Broman KW, Murray JC, She eld VC, White RL, Weber JL. Comprehensive human genetic maps: individual an sex-specific variation in recombination. *American Journal of Human Genetics*. 1998; 63:861–869. [PubMed: 9718341]
- Camp NJ, Swensen J, Horne BD, Farnham JM, Thomas A, Cannon-Albright LA, Tavtigian SV. Characterizaion of linkage disequilibrium structure, mutation history, and tagging SNPs, and their use in association analyses: ELAC2 and familial early-onset prostate cancer. *Genetic Epidemiology*. 2005; 28:232–243. [PubMed: 15593091]
- Cannings C. The identity by descent process along the chromosome. *Human Heredity*. 2003; 56:126–130. [PubMed: 14614246]
- Chapman NH, Thompson EA. The effect of population history on the lengths of ancestral chromosome segments. *Genetics*. 2002; 162:449–458. [PubMed: 12242253]
- Donnelly KP. The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*. 1983; 23:34–63. [PubMed: 6857549]
- Heath S, Robledo R, Beggs W, Feola G, Parodo C, Rinaldi A, Contu L, Dana D, Stambolian D, Siniscalco M. A novel approach to search for identity by descent in small samples of patients and controls from the same Mendelian breeding unit: a pilot study in myopia. *Human Heredity*. 2001; 52:183–190. [PubMed: 11713414]
- Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB. Genomic screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics*. 1994; 8:380–386. [PubMed: 7894490]
- Lauritzen, SL. *Graphical Models*. Clarendon Press; 1996.
- Leibon G, Rockmore DN, Pollack MR. A SNP streak model for the identification of genetic regions identical-by-descent. *Statistical Applications in Genetics and Molecular Biology*. 2008; 7:16.
- MacCluer JW, Vandeburg JL, Read B, Ryder OA. Pedigree analysis by computer simulation. *Zoo Biology*. 1986; 5:147–160.
- Miyazawa H, Kato M, Awata T, Khoda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, Hagiwara K. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *American Journal of Human Genetics*. 2007; 80:1090–1102. [PubMed: 17503327]

- Newman DL, Abney M, Dytch H, Parry R, McPeck MS, Ober C. Major loci influencing serum triglyceride levels on 2q14 and 9p21 localized by homozygosity-by-descent mapping in a large Hutterite pedigree. *Human Molecular Genetics*. 2003; 12:127–144.
- te Meerman GJ, Van der Meulen MA. Genomic sharing surrounding alleles identical by descent: Effects of genetic drift and population growth. *Genetic Epidemiology*. 1997; 14:1125–1130. [PubMed: 9433635]
- Thomas A. Characterizing allelic associations from unphased diploid data by graphical modeling. *Genetic Epidemiology*. 2005; 29:23–35. [PubMed: 15838847]
- Thomas A. Towards linkage analysis with markers in linkage disequilibrium. *Human Heredity*. 2007; 64:16–26. [PubMed: 17483593]
- Thomas A. Estimation of graphical models whose conditional independence graphs are interval graphs and its application to modeling linkage disequilibrium. *Computational Statistics and Data Analysis*. 2009a; 53:1818–1828. [PubMed: 20161313]
- Thomas A. A method and program for estimating graphical models for linkage disequilibrium that scale linearly with the number of loci, and their application to gene drop simulation. *Bioinformatics*. 2009b In press.
- Thomas A, Camp NJ. Graphical modeling of the joint distribution of alleles at associated loci. *American Journal of Human Genetics*. 2004; 74:1088–1101. [PubMed: 15114533]
- Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Annals of Human Genetics*. 2008; 72:279–287. [PubMed: 18093282]
- Van der Meulen MA, te Meerman GJ. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genetic Epidemiology*. 1997; 14:915–919. [PubMed: 9433600]

Figure 1.
A pedigree connecting 8 men with prostate cancer.

Figure 2.

(a) shows $R_i(8)$, the run lengths where all 8 cases share an allele IBS plotted against the physical location of the loci. (b) shows $R_i(7)$, the run lengths where at least 7 of the 8 cases share an allele IBS. (c) is a cumulative sum chart for the heterozygosities of the loci: areas of steeply increasing or decreasing slope indicate high or low heterozygosity respectively.

Figure 3.

These plots show the cumulative distribution functions for the maximum simulated run lengths for which (a) all 8 cases share an allele IBS and (b) 7 of 8 cases share. In both plots the distribution furthest to the left is that from 10000 simulations when the loci are assumed to be in linkage equilibrium with equal allele frequencies. The distribution in the middle is that for 10000 simulations under linkage equilibrium, but with allele frequencies at each locus estimated from control data. On the right of each plot are overlaid 10 distributions each of 1000 simulations under a linkage disequilibrium model estimated from the control data. The dashed vertical lines show the values observed in the data.

Figure 4.

These plots show the probability, as estimated by simulation, that a locus is contained in the maximum run of sharing among all 8 cases. (a), (b) and (c) show respectively 10000 simulations under linkage equilibrium with equal allele frequencies, 10000 simulations under linkage equilibrium with estimated allele frequencies, and 10×1000 simulations under LD.