Published in final edited form as:

Ann Hum Genet. 2008 March ; 72(Pt 2): 279–287. doi:10.1111/j.1469-1809.2007.00406.x.

Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays

Alun Thomas^{*,†}, Nicola J Camp^{*}, James M Farnham^{*}, Kristina Allen-Brady^{*}, and Lisa A Cannon-Albright^{*}

^{*}Department of Biomedical Informatics, University of Utah

Abstract

We examine the utility of high density genotype assays for predisposition gene localization using extended pedigrees. Results for the distribution of the number and length of genomic segments shared identically by descent among relatives previously derived in the context of genomic mismatch scanning are reviewed in the context of dense single nucleotide polymorphism maps. We use long runs of loci at which cases share a common allele identically by state to localize hypothesized predisposition genes. The distribution of such runs under the hypothesis of no genetic effect is evaluated by simulation. Methods are illustrated by analysis of an extended prostate cancer pedigree previously reported to show significant linkage to chromosome 1p23. Our analysis establishes that runs of simple single locus statistics can be powerful, tractable and robust for finding DNA shared between relatives, and that extended pedigrees offer powerful designs for gene detection based on these statistics.

Keywords

Candidate region; identity by descent; identity by state; prostate cancer; pedigree analysis

1 Introduction

The recently developed ability to genotype dense single nucleotide polymorphism (SNP) marker sets on accurate analytical platforms, coupled with relatively inexpensive costs and high efficiency is changing the nature of genetic analysis. As SNPs are far more abundant than conventional micro satellite markers, they have the capacity to give more precise and sure localization (Kruglyak 1997). To date, analyses of dense SNP genome wide scans using pedigree data have been accomplished by linkage approaches, however, for even moderately sized pedigrees multi locus linkage analysis is tractable only by Markov chain Monte Carlo methods (Thomas et al. 2000, Wijsman et al. 2006), and the number of loci in current SNP assays creates an immense computational burden. In addition to this, the sensitivity of linkage analysis to linkage disequilibrium (*LD*) (John et al. 2004, Amos et al. 2006) and the difficulties of modeling *LD* in linkage analysis, even by Markov chain Monte Carlo integration (Thomas 2007), make alternative approaches very attractive.

Rather than the complete likelihood approach for arbitrarily structured pedigrees that linkage analysis accomplishes, we consider only sets of cases related by a single common ancestor or ancestral couple. Localization is based on the assumption that regions shared identically by

[†]Corresponding author. 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)..

descent (*IBD*) from a common ancestor indicate regions that are likely candidates for a predisposing gene. Since regions shared IBD must also be shared identically by state (*IBS*), runs of loci at which individuals share a common allele will tend to be longer when there is underlying IBD than when there is not. We develop this into a simple approach for localizing predisposition genes for a trait segregating in an extended pedigree. The distribution of runs of IBS loci, and hence statistical significance tests, are evaluated by simulation.

We briefly review relevant literature on IBD sharing in pedigrees, outline our IBS statistic and tests, and illustrate our approach with the analysis of a SNP assay of 109,299 loci for 8 related prostate cancer cases taken from an extended Utah family previously reported to give a lod score of 3.1 for linkage to chromosome 1p23. We discuss the implications of our approach and, in particular, the future work needed for further development.

2 Methods

IBD sharing in pedigrees

There is considerable literature on IBD sharing in statistical genetics, beginning with Fisher's *junction* theory (Fisher 1949, Fisher 1954), a junction being defined as a point on a chromosome where DNA inherited from two distinct ancestral chromosomes meets. Donnelly (1983) modeled the common inheritance of ancestral chromosomal segments as a random walk over the vertices of a hypercube, where each dimension corresponds to a meiosis in the pedigree. Particular states such as, for example, where descendants share a segment IBD, correspond to particular sets of vertices. Cannings (2003) also derived results for this model. Houwen et al. (1994) and Heath et al. (2001) both used relatively isolated founder populations to identify a small number of distantly related cases who shared common chromosomal segments, which they used to map disease genes. However, neither of these approaches incorporated precise pedigree relationships between cases in their methodology and both used micro satellite markers which bypasses the complexity of dense SNP maps. Chapman & Thompson (2002) and te Meerman & Van der Meulen (1997) examined the length of an ancestral chromosomal segment in founder populations and found that the segment length is dependent on time since founder population, population growth, genetic drift, limited negative selection, and population subdivision.

Using preliminary work of Sanda & Ford (1986), Nelson et al. (1993) developed techniques and analysis methods for molecular genomic mismatch scanning (*GMS*). In *GMS*, long stretches of hybridized DNA from two related individuals identify IBD regions. When several pairs of individuals affected by a disease share the same IBD region, it becomes a candidate region for a shared disease predisposition gene. Thomas et al. (1994) extended statistical analyses of *GMS* data from IBD sharing for two related individuals to IBD sharing among multiple affected individuals in a pedigree. Although the *GMS* method has clear implications for gene mapping, it never realized its potential because the laboratory procedures are complex, subject to substantial background noise, and not suitable for scaling to high throughput systems. With the recent availability of dense SNP assays, however, allelic differences that are indirectly assayed in *GMS* can be directly assayed using SNP genotypes.

Two results derived by Thomas et al. (1994) are relevant here. The first concerns the probability distribution of the number of distinct segments shared IBD among a set of relatives, the second concerns the length of any such shared segment. Consider a set of individuals all descended from a common ancestor or common ancestral couple. Let d be the number of meioses that connect all of these individuals to a common ancestor, and let a be the number of common ancestors: 1 for a single ancestor, 2 for an ancestral pair. For instance, the pedigree shown in figure 1 has $d = 15$ and $a = 2$. Let k be the number of chromosomes being considered and let λ be the total number of recombination events expected over these chromosomes. For example,

for a complete genome scan of the human autosomes $k = 22$ and $\lambda \approx 35$ (Broman et al. 1998). The number of distinct chromosomal regions shared IBD by all the individuals is approximately distributed as a Poisson random variable with mean

$$\frac{d\lambda + k}{2^{d-a}}. \quad (1)$$

An intuitive derivation of this result is straightforward: each of the d meioses creates, on average, λ junctions, which, with the chromosomal breakpoints, give a total of $d\lambda + k$ stretches of contiguous DNA which segregate as a unit, and in a different way to the adjacent units. For

each of these units, the probability that it is transmitted to all of the descendants is $\frac{1}{2^d}$. If there is a common ancestral pair then there are 4 possible sources for the IBD segment, but only 2 if there is a single, multiply mated ancestor. The expected number of IBD segments shared by all descendants is simply the product of these three terms.

Also, if we assume that the underlying recombinations at each meiosis occur as independent Poisson processes, the length of any shared IBD segment is Exponentially distributed with

mean $\frac{1}{d}$ Morgans. This assumes that we can neglect the effects of truncation due to reaching the ends of chromosomes. These effects should be small when d is reasonably large.

Note that the distances are genetic distances so variation in recombination rates through the genome are irrelevant until we map to the physical domain. Note also that since the process marking all junctions is made by overlaying the d independent processes for each meiosis, departures from the Poisson assumptions for the individual processes are diluted in the overlaid process. Hence, the effect of interference is minimal, and the above Poisson and Exponential results are robust to it.

The situation is more complicated when all the individuals considered are full sibs since it is then possible that they share more than one chromosome IBD at any particular locus. A similar complexity occurs more generally in looped pedigrees. While the general case can be handled both by Donnelly's approach and by the simulation methods we describe below, the predominant structure available in our data resource is that addressed by Thomas et al. (1994) and illustrated in figure 1. We focus on the simpler case of a single common ancestor, or ancestral couple, for the remainder of this work. Note that in considering only the closest common ancestor of the cases, we are assuming that the effects of other, more distant, relationships are negligible.

Suppose now that we select our set of relatives to be cases for a disease. If we find that they share at least one segment of DNA IBD anywhere in the genome, the probability that this occurs under the hypothesis that the disease status is independent of genetic events is

$$1 - e^{-\frac{(d\lambda + k)}{2^{d-a}}} \approx \frac{d\lambda + k}{2^{d-a}}. \quad (2)$$

If this is sufficiently small, we would reject the null hypothesis that they share by chance in favour of the hypothesis that there is a gene located in the IBD region affecting the trait. We emphasize that these probabilities are calculated based on the number of chromosomes, k , and the total genetic length, λ , so that no further multiple testing correction is required. Note that under the alternative hypothesis, each of the cases shares the length of the IBD region due to a shared disease predisposition gene, so the shared length is equal to the distance to the first

junction on either side, and is hence the sum of two Exponentials or a Gamma $\left(2, \frac{1}{d}\right)$, which again assumes that chromosome end effects are negligible. This also gives us some power to detect deviations from random segregation.

IBS sharing in pedigrees

Unlike GMS where shared IBD segments may be determined, genome wide SNP scans can only provide information on shared IBS regions. When individuals share a common allele at a contiguous series of SNPs this may correspond to underlying IBD or may have occurred by chance, particularly in a run of SNPs with low minor allele frequency. It may also be due to some combination of both causes. However, since regions IBD must also be IBS, IBS regions that cover IBD regions will generally be longer than those that do not. Thus, we can again test the null hypothesis that a shared IBS segment is independent of any underlying genetic influence on the disease if the length of the segment exceeds some critical value.

As we will show below, IBS sharing closely tracks IBD, so an alternative approach might be to use IBS sharing to infer IBD sharing and then test for a genetic cause using the distributions described above in section 2, taking into account the uncertainty in the inference of IBD. It is more straightforward, however, to test for the genetic cause using the IBS sharing directly, determining the critical value by simulation.

We define a set of n individuals to be IBS at a genotyped locus if they all share a common allele. At any locus i we can define S_i as the largest number of individuals who share an allele and calculate this from the genotype counts (n_{11}, n_{12}, n_{22}) as

$$S_i = n - \min(n_{11}, n_{22}). \quad (3)$$

where 1 and 2 are arbitrary labels for the SNP's alleles. This assumes that the data are without error, but allows for missing values which are, in effect, counted as heterozygotes. Thus, S_i can also be thought of as the largest subset of individuals whose genotypes, if correctly assayed, do not exclude the possibility that they share IBD.

Taken individually, these S_i have a low amount of information, however, we can exploit the density of a SNP assay by looking for runs of consecutive S_i which exceed a given threshold.

To assess the extremity of any observed value of S_i under the null hypothesis in a candidate region, we can compare it with sharing from the rest of the genome, under the assumption that the majority of the genome behaves under the null hypothesis. This approach has the advantage that it can be applied even when the genealogy is not known. Alternatively, when have pedigree data, we can use a simulation scheme based on the model of Donnelly (1983) or Thomas et al. (1994), as follows:

1. Each founder chromosome is represented by a unique identifier applied to an interval of $(0, l)$ where l is the physical length of the region being simulated in bases. The physical distance between loci is maintained as in the observed data.
2. Working through each non founder in birth order we allocate them a chromosome represented by a list of intervals, where each interval has an identifier indicating the founder chromosome from which it is descended. Each set of intervals is derived by recombining the parent's chromosomes, the junctions being determined by a Poisson process with rate λ , the genetic length of the region in Morgans.

3. At each locus, each founder chromosome is allocated an allelic state randomly generated according to specified allele frequencies. This determines the genotypes of the remainder of the pedigree.
4. Genotype counts for the cases and S_i , or other relevant statistics are computed.

Step 2 above assumes that the recombination rate is constant over the region simulated. Sex specific recombination rates, large scale variations such as the tendency for lower recombination rates near centromeres, and small scale variation due to recombination hot spots are issues that we plan to address in future work. Step 3 assumes that the loci are in linkage equilibrium. We make an initial investigation of the effects of LD below, but this will also be a focus of future work.

We could at this stage base our statistical tests on either the number of contiguous loci that are IBS, or by the physical genomic distance that they span. Some initial investigation showed that, as expected, there was little difference in effect when the loci were evenly spaced. However, the simple count of loci was more robust to gaps in the genetic map such as near the centromeres. When there were no observed loci that could reveal a lack of IBD sharing, the length statistics became highly inflated. For this reason we base our statistical conclusions on the number of loci in an IBS run, although we report the physical length of any interesting regions found.

This simulation process has been implemented by the authors in a Java program. Java was also used to compute the IBS statistics. All other analysis, including calculating the run length statistics, was done in the R statistical environment (R Development Core Team 2004).

Our approach has broad similarities with the haplotype sharing statistics of Van der Meulen & te Meerman (1997) and Beckmann et al. (2005) in that it aims to identify excess IBD sharing in cases. However, we note that their more complicated statistics require genotyping of close relatives in order to estimate phase, are based on combining pairwise comparisons, and are applied in population samples rather than in extended pedigrees. The work of Bourgain et al. (2001) is more similar to what we present here as it is applied in a very large extended pedigree, but it again requires knowing phase and combines pairwise distances.

3 Example

In a genome wide micro satellite marker linkage scan for prostate cancer predisposition, reported by Camp et al. (2005), a single extended pedigree showed a multi point lod score of 3.1 at chromosome 1p23. This pedigree is shown in part in figure 1. To examine the utility of our methods, DNA from 8 affected individuals, shown numbered in the figure, was submitted to the Center for Inherited Disease Research (*CIDR*), and genotyped using the Illumina 110K panel (<http://www.illumina.com>).

Our illustrative analysis focuses on using our methods to confirm the linkage result for 1p. Since we use a subset of the same pedigree that gave this linkage, it is appropriate to evaluate the significance of our runs statistics on a genome wide level, and this forms the first part of our analysis. We then proceed to evaluate our result as if it were from an independent study to confirm a localization on chromosome arm 1p. This is partially justified by the considerable literature indicating the presence of a prostate cancer susceptibility gene on 1p, although the prior evidence near 1p23 (Xu et al. 2003, Witte et al. 2003, Maier et al. 2005) is weaker than that for loci elsewhere on 1p (Gibbs et al. 1999, Suarez et al. 2000, Matsui et al. 2004). Mainly, however, we do this for the purpose of illustration as it enables us to obtain many more simulations and to focus on the particular structure of this region.

Genome wide analysis

From a total of 109,299 loci, 3,442 were on the sex chromosomes and are not included in this analysis. Figure 2 shows the genome wide results for the lengths and positions of 38,373 runs for which all 8 cases shared a common allele, that is, where $S_i = 8$. Among the longest of these is a run of 79 loci on 1p, spanning 1.96 Mb, which covers exactly the marker at which the peak lod score of 3.1 was observed. Haplotypes reconstructed from the micro satellite data are also consistent with this result, however, the smallest region defined by the micro satellite data is 16.7 Mb, so the SNP data allows us to narrow this considerably.

Using the simulation method described above, we made 10,000 genome wide simulations of the run length statistic. Genetic distances were taken from the Marshfield map (<http://research.marshfieldclinic.org>), and allele frequencies were estimated from 52 Utah CEPH controls that are included as part of the 120 control sample set genotype by Illumina for the same set of SNPs. We found that a run of 79 for $S_i = 8$ was equaled or exceeded only 382 times, giving an empirical p-value of 0.0382. Although significant, we consider this a tentative result, particularly in view of the concerns we discuss below. Nonetheless, it is sufficient to maintain interest in the region.

It is informative to compare the genome wide p-value of 0.0382 for the run of 79 loci with $S_i = 8$ with the probability from equation 2 that there is a shared IBD region among the cases. For $= 35$, $d = 15$, $a = 2$, $k = 22$ this is 0.067. Clearly, there were IBD regions simulated that were not of sufficient length to cover 79 loci or more. This shows that the p-value reflects power not only from the implied existence of a shared IBD region, but also from its length. It also shows that uniform spacing of the SNP coverage in the assay is important.

In order to allow for the possibility of sporadic incidence of prostate cancer among the cases, we also looked for runs where 7 of the 8 cases shared IBS. As can be seen in figure 3 the longest such run was 619 loci long, occurred at 1p23, and included the 79 loci described above. Moreover, apart from the first 4 and last 1, the non-sharing individual was the same one: individual 7. However, although this was by far the longest of the 12,078 runs of $S_i \geq 7$ seen in the whole genome scan, the next longest being only 213, this result is not statistically significant on a genome wide level (p-value = 0.0874). This suggests considerable skew in the distribution of such run lengths.

Four runs where $S_i = 8$ were longer than the 79 observed on 1p. These were runs of 84, 83, 81 and 83 on chromosomes 10, 12, 17 and 22 respectively. These might also be considered candidate regions for a prostate cancer predisposition gene. They were surrounded by runs where $S_i \geq 7$ of lengths 167, 102, 83 and 85, respectively, thus none of these is robust to the possibility of a sporadic case. However, note that it is not the case that a run where all 8 cases share IBD has to be surrounded by a run where 7 of the 8 share: a recombination that occurs early in the pedigree may cause several cases to stop being IBD with the majority at the same junction.

Candidate region analysis

Assuming, for the sake of illustration, that prior evidence for a prostate cancer susceptibility locus on chromosome arm 1p allows us to restrict attention to this region, we made 100,000 simulations of for the 120 Mb spanning the 5,213 SNP markers here. The genetic distance was taken as 1.5 Morgans.

In our 100,000 simulations, the longest run of IBS on chromosome arm 1p for all 8 cases was greater than or equal to 79 loci only 207 times (empirical p-value=0.00207). The longest run where 7 of 8 cases shared exceeded 619 only 482 times (empirical p-value=0.00482). Figures 4 and 5 show the empirical distribution functions for the IBS run lengths from which these p-

values are calculated. These plots also show the different distributions of longest IBS run when there is and is not underlying IBD sharing. The clear difference between when there is and is not IBD sharing illustrates the considerable power in IBS runs for detecting IBD. The very close relationship between the longest runs of IBS and IBD sharing is also shown in figures 6 and 7 for all 8 cases and for 7 of 8 cases, respectively. Based on the simulations, an IBD region of any length, common to all 8 cases, occurred on chromosome 1p with probability 0.00276, closely matching the theoretical value of 0.00287 derived from equation 2 with $d = 15$, $a = 2$, $\lambda = 1.5$, $k = 1$. In the simulations in which no IBD region was shared by all 8 cases the longest run of loci at which $S_i = 8$ was 106. An IBD region common to 7 of the 8 cases sharing occurred with probability only 0.03452 as estimated by simulation. In the simulations in which no 7 cases shared an IBD region the longest run of loci at which $S_i \geq 7$ was 491. Thus, it is clear both that IBS runs closely track the length of IBD regions, and that in extended pedigrees, long IBS runs occur only rarely without underlying IBD.

Conclusion

There is clear evidence in our data for a shared IBD region in our 8 cases, however, the consequent conclusion that this region must contain a prostate cancer susceptibility gene is marginal. The pedigree we have been able to collect is of a size suitable for a candidate region or confirmatory study. A de novo genome wide scan requires larger pedigrees, as indicated by Thomas et al. (1994) who recommended a pedigree linked by more than 20 meioses. We are currently ascertaining and genotyping prostate cancer pedigrees in excess of this size in our follow up studies.

4 Discussion

This approach to predisposition gene localization is new, and there are several issues to address in subsequent work. Foremost of these is LD, as correlations between alleles at proximal loci will increase run lengths under the null hypothesis of no genetic effect. To make an initial evaluation of the effects of LD, we analyzed genotype data for 60 unrelated CEPH Utah individuals from HapMap for the chromosome 1p region of interest. Using a similar overall distribution of r^2 statistics as that found among the unrelated CEPH Utah individuals, we created a model of correlated alleles which were distributed to the founders in the simulation analysis. We found that the empirical p-value for a shared run length of 79 loci IBS for all 8 cases increased greatly from 0.00207 to 0.20793, but the p-value for the shared run length of 619 loci IBS for 7 out of 8 cases increased only from 0.00482 to 0.00638. The difference in effect is presumably because 619 is in the upper tail of the length distribution even when there is IBD sharing among 7 from 8 cases. Further work is required to include more appropriate and realistic LD models, such as those of Griffiths & Marjoram (1996), Morton & Collins (2002) or Thomas & Camp (2004), in the simulation. In the same vein, estimates of variable recombination rates need to be accounted for by a non linear translation from the genetic to physical domains.

When subsets of the cases are considered, the number of meioses lost can vary. In our example of IBS sharing among any 7 out of 8 cases compared to sharing among all 8 cases, either one or two meioses were lost, that is, there were 13 or 14 meioses in the reduced pedigree compared to the original 15. In a single averaged statistic, as used here, there will be more statistical power to detect subsets which retain fewer meioses. In other data sets where there are bigger differences in the number of meioses lost, a statistic T_i , say, equal to the largest number of meioses separating a set of cases who share a common allele at locus i will be a better basis for hypothesis testing, although slightly more involved to calculate. Under perfect observation of IBD regions and with no sporadic cases, Thomas et al. (1994) showed that a single pedigree with 21 meioses was enough to detect linkage with a genome wide scan. In order to allow for

observed IBS instead of IBD, and for sporadic cases reducing the number of meioses, pedigrees with meiosis count d in the 25 to 30 range are probably needed.

Much of the appeal of this approach is that the power available in a single pedigree obviates the need to consider genetic heterogeneity of the phenotype. However, it is also straightforward to combine data from independent pedigrees by finding regions that co-segregate in them. Note that this does not lead to a test for allelic association unless we specify that the alleles shared in the different pedigrees are the same.

Given the structure of the pedigree in figure 1, it is impossible to detect genotyping errors in a diallelic marker by looking for violations of Mendelian segregation. Our analysis should, however, be extended to allow for error because a single misclassification of a heterozygote as a homozygote can prematurely end a run of IBS sharing. Requiring multiple mismatches before ending a run is one way, alternatively, we can find statistics based on the locus by locus posterior distributions for the inheritance states, which are tractable by the usual peeling method (Cannings et al. 1978). Finding runs of high values can be accomplished using, for example, cumulative sum charts.

Representative population allele frequency estimates are essential for the simulation analysis. In this study, the CEPH Utah individuals genotyped on the same panel of markers as our Utah prostate cancer pedigree fortuitously provided representative population frequency estimates. However, for studies in other populations it may be necessary to use allele frequency estimates from the pedigrees themselves. To explore the effects of this, we repeated our analysis with allele frequencies estimated from the pedigree and a single parent offspring triplet additionally genotyped by CIDR as a control, using the naive unbiased estimator. The effects of this change in allele frequency estimates were minimal. The p -value for the 79 loci shared by all 8 cases changed to 0.00202 while that for the 619 loci shared by 7 of 8 became 0.00511. Again, further work is needed to better quantify the sensitivity to allele frequency estimates.

The central question we have considered here is whether we can infer a shared region containing a predisposition gene from unexpectedly long runs of IBS sharing among distantly related affected individuals. Although our empirical tests assess them jointly, this breaks down into two separable issues. The first issue is whether IBS sharing is sufficient to conclude that there must be underlying IBD sharing. Figures 6 and 7 clearly show that in our simulations IBS runs closely match the underlying IBD, and the evenness of coverage, polymorphic content and quality of assay of the SNP panel are certainly adequate to make this analysis feasible. The clear difference between distribution of run length when there is, and is not, underlying IBD sharing demonstrates the power to detect IBD from IBS. Given that we can infer IBD sharing, the second issue is whether it is sufficiently unexpected under random segregation that we can conclude such sharing must be due to an underlying genetic cause that resulted in the selection of the cases. This is where the power of extended pedigrees is most important. While at first glance it appears that analysis using sets of relatives introduces unnecessary complexity, it is in the balance between the two central issues that the elegance of the extended pedigree design is apparent. The length of any region shared IBD by a set of relatives decreases slowly,

$O\left(\frac{1}{d}\right)$, as d , the number of meioses connecting them, increases. Thus, there is a relatively large target to be covered by the SNPs in the assay. Conversely, the probability under random

segregation that there exists any shared IBD region decreases very quickly, $O\left(\frac{1}{2^d}\right)$, hence, for a sufficiently informative pedigree any detected sharing is likely to be significant. In short: big target, little noise.

Acknowledgments

This work was supported by Grants Number RO1 GM070710 and RO1 GM081417 to Alun Thomas from the National Institutes of General Medical Sciences, National Cancer Institute grants K07 CA098364 to Nicola Camp, and RO1 CA90752, RO1 CA102422, and RO1 CA89600 to Lisa Cannon-Albright, the latter being via a subcontract from Johns Hopkins University. 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 Material Command W81XWH-07-1-0483 to Alun Thomas.

Some data for this research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the NCI, 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

- Amos CI, Chen WV, Lee A, Li W, Kern M, Lundsten R, Batliwalla F, Wener M, Remmers E, Kastner DA, Criswell LA, Seldina MF, Gregersen PK. High density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Genetic and Immunity* 2006;7:277–286.
- 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, Sheffield 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, Farnham JM, Cannon-Albright LA. Genomic search for prostate cancer predisposition loci in Utah pedigrees. *The Prostate* 2005;65:365–374. [PubMed: 16037989]
- Cannings C. The identity by descent process along the chromosome. *Human Heredity* 2003;56:126–130. [PubMed: 14614246]
- Cannings C, Thompson EA, Skolnick MH. Probability functions on complex pedigrees. *Annals of Applied Probability* 1978;10:26–61.
- 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]
- Fisher, RA. The theory of inbreeding. Oliver and Boyd; Edinburgh: 1949.
- Fisher RA. A fuller theory of 'junctions' in inbreeding. *Heredity* 1954;8:187–197.
- Gibbs M, Stanford JL, McIndoe RA, Jarvik GP, Kolb S, Goode EL, Chakrabarti L, Schuster EF, Buckley VA, Miller EL, Brandzel S, Li S, Hood L, Ostrander EA. Evidence for a rare prostate cancer-susceptibility locus at chromosome 1p36. *American Journal of Human Genetics* 1999;64:774–787.
- Griffiths RC, Marjoram P. Ancestral inference of samples of DNA sequences with recombination. *Journal of Computational Biology* 1996;3:479–502. [PubMed: 9018600]
- 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]
- John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, Eyre S, Jones KW, Ollier W, Silman A, Gibson N, Worthington J, Kennedy GC. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *American Journal of Human Genetics* 2004;75:54–64. [PubMed: 15154113]
- Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics* 1997;17:21–24. [PubMed: 9288093]
- Maier C, Herkommer K, Hoegel J, Vogel W, Paiss T. A genomewide linkage analysis for prostate cancer susceptibility genes in families from Germany. *European Journal of Human Genetics* 2005;13:352–360. [PubMed: 15536476]
- Matsui H, Suzuki K, Ohtake N, Nakata S, Takeuchi T, Yamanaka H, Inoue I. Genomewide linkage analysis of familial prostate cancer in the Japanese population. *Journal of Human Genetics* 2004;49:9–15. [PubMed: 14666403]
- Morton NE, Collins A. Toward positional cloning with SNPs. *Current Opinion in Molecular Therapeutics* 2002;4:259–264.
- Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO. Genomic mismatch scanning: A new approach to genetic linkage mapping. *Nature Genetics* 1993;4:11–18. [PubMed: 8513319]
- R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2004. ISBN 3-900051-07-0
- Sanda AI, Ford JP. Genomic analysis I: inheritance units and genetic selection in the rapid discovery of locus linked dna makers. *Nucleic Acids Research* 1986;14:7265–7283. [PubMed: 3763405]
- Suarez BK, Lin J, Burmester JK, Broman KW, Weber JL, Banerjee TK, Goddard KAB, Witte JS, Elston RC, Catalona WJ. A genome screen of multiplex sibships with prostate cancer. *American Journal of Human Genetics* 2000;66:933–944. [PubMed: 10712208]
- 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. Towards linkage analysis with markers in linkage disequilibrium. *Human Heredity* 2007;64:16–26. [PubMed: 17483593]
- 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, Gutin A, Abkevich V, Bansal A. Multipoint linkage analysis by blocked Gibbs sampling. *Statistics and Computing* 2000;10:259–269.
- Thomas A, Skolnick MH, Lewis CM. Genomic mismatch scanning in pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology* 1994;11:1–16. [PubMed: 8057038]
- 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]
- Wijsman EM, Rothstein JH, Thompson EA. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics* 2006;79:846–858. [PubMed: 17033961]
- Witte JS, Suarez BK, Thiel B, Lin J, Yu A, Banerjee TK, Burmester JK, Casey G, Catalona WJ. Geneome-wide scan of brothers: replication and fine mapping of prostate cancer susceptibility and aggressiveness loci. *The Prostate* 2003;57:298–308. [PubMed: 14601026]
- Xu J, Gillanders EM, Isaacs SD, Chang BL, Wiley KE, Zheng SL, Jones M, Gildea D, Riedesel E, Albertus J, Freas-Lutz D, Markey C, Meyers DA, Walsh PC, Trent JM, Isaacs WB. Geneome-wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families. *The Prostate* 2003;57:320–325. [PubMed: 14601028]

Figure 1.

An extended Utah pedigree connecting 8 men with prostate cancer. The affected individuals are numbered. In the interests of keeping the pedigree unidentifiable, the sexes of the connecting ancestors are not specified and other relatives are not shown. The numbers under the affected individuals give the genotypes for 6 micro satellite markers which lie within the region of 619 SNP markers where 7 of the 8 individuals share IBS. The marker outlined in a horizontal box lies within the region of 79 SNP markers where all 8 cases share IBS. The vertical boxes show the reconstructed common micro satellite haplotype.

Figure 2.

Runs of loci where $S_i = 8$, that is where the possibility of all 8 cases sharing IBD is not excluded by the SNP genotypes. Each chromosome is represented by a block of length proportional to physical size. The bars representing the runs cover all the bases between the markers in each run. Spaces within a chromosome correspond to centromeric gaps in the marker map. At this resolution not all the gaps are large enough to show up.

Figure 3.

Runs of loci where $S_i \geq 7$, that is where the possibility of at least 7 of 8 cases sharing IBD is not excluded by the SNP genotypes. The format is the same as for figure 2.

Figure 4.

Empirical distribution functions of run lengths where all 8 cases share IBS estimated from 100,000 simulations. The dashed line is the overall distribution, which is a mixture of the run lengths when there is, or is not, underlying IBD sharing. These component distributions are shown as dotted and solid lines respectively. The vertical line at 79 indicates the observed longest run length for chromosome arm 1p. The overall distribution is used to assess empirical p-values.

Figure 5.

Empirical distribution functions of run lengths where 7 out of 8 cases share IBS estimated from 100,000 simulations. The dashed line is the overall distribution, which is a mixture of the run lengths when there is, or is not, underlying IBD sharing. These component distributions are shown as dotted and solid lines respectively. The vertical line at 619 indicates the observed longest run length for chromosome arm 1p. The overall distribution is used to assess empirical p-values.

Figure 6. Length of the longest run of loci IBD in all 8 cases in 100,000 simulations of segregation of chromosome arm 1p against the length of the longest run of loci IBS for all 8 cases. The horizontal line shows the observed value of 79 loci. All points must lie on or above the diagonal line.

Figure 7.

Length of the longest run of loci IBD in 7 of 8 cases in 100,000 simulations of segregation of chromosome arm 1p, against the longest run of loci IBS for 7 of 8 cases. The horizontal line shows the observed value of 619 loci. All points must lie on or above the diagonal line.