

COLORECTAL CANCER: HOW EFFECTIVELY CAN WE FIND THE GENETIC LINKS

**Tamie Henderson and Ian Saunders
CSIRO Mathematical and Information Sciences**

CMIS Technical Report No 05/92

1. Introduction

Colorectal Cancer (CRC) is the second-leading cause of cancer-related deaths in Australia. The number of deaths could be reduced if diagnosis of this disease occurred earlier. One way to do this would be to locate the genes which cause CRC, so that a DNA scan will be able to indicate the likelihood of the disease developing. 25% of CRC are hereditary, that is they have a genetic link, of these only 5-6% of the genetic causes are actually known. In the past research have only analysed segments of our DNA, but new technology now allow the entire DNA to be scanned. CSIRO's preventative health flagship project is using this technology and CMIS is developing ways to analyse the data. Two of these methods, Allele Sharing and SibRave Analysis, are designed to locate the regions of the genome where a disease gene may lie. The cost of obtaining the data for this analysis is quite expensive, so prior study need to be done to determine how effective these methods will be in locating these disease genes.

2. Mathematical Model

To generate each sibling pair, information was obtained from SNP10kInfo.txt, which contained data about 10,552 SNPs along the 22 autosomal chromosomes. The relevant information obtained for each SNP was the recombination frequency (theta values) and the allele frequency (PrA). The recombination frequencies were estimated as 10^{-8} multiplied by the number of nucleotides between adjacent SNPs. In addition to this information a gene was placed into the genome half-way between two SNPs (400 and 401) to model a disease gene.

The disease penetrance and the frequency of the disease gene allele were chosen depending on four models to be analysed, and for each of these models the number of sib-pairs to be analysed in each group and the number of siblings affected in each pair were varied.

The haplotypes for the parents were randomly generated based upon the allele frequencies, and the locations of crossovers between a pair of chromosomes were determined in relation to the theta values. Two siblings were then generated based on this information. Sibling one always initially received alleles from the first pair in each parent's two chromosomes, and this was randomised for sibling two. Their disease status was generated at random according to the disease penetrance for this genotype at the disease gene. A new set of parents were generated for each sibling pair and only the information about the sib-pairs where one or two of the pair were affected were stored.

For each sibling pair chosen the number of alleles identical by descent (IBD) was determined at each SNP location. As the genotypes of the parents and the crossover points were known, the true IBD statuses (trueIBDs) could be determined. This information may not be known for the actual siblings to be analysed, so the Viterbi algorithm was used to estimate the IBD status (estIBDs) at each SNP locus.

For each model, 50 000 one affected and two affected sib-pairs were generated. This was done for when the disease gene was and wasn't genetically linked. Sib-pairs were then randomly placed into groups of a given size. For each group of siblings, the ratio of 0, 1 and 2 IBD values were determined at each SNP locus. According to Mendelian segregation rules if there is no genetic link then the ratios of 0, 1 and 2 IBD statuses would be 25:50:25 respectively. If the disease gene is genetically linked then these ratios will vary at SNPs near the disease gene. For example, if both of the siblings have the disease then we would expect to see a greater similarity around this location in the genome, hence more than 25% of siblings like this would have 2 IBD. If only one of the siblings was affected then there would be less similarity, hence more 0 IBD. For the Allele Sharing method the number of 0 and 2 IBD statuses at each SNP were used to calculate the p-values. Only the 0 and 2 IBD statuses were used because when there is a genetic link the greatest deviation from the Mendelian ratios is observed in the proportions of these two. Hence the expected ratio of 1 to 2 IBD statuses at each SNP is 50:50.

To determine the p-value for the deviation of the ratios observed at each SNP from the expected 50:50 the log likelihood ratio was used where,

x_0 = the number of 0 IBD for the group of sib-pairs with one affected
 m_1 = the total number of 0 and 2 IBD for all the one affected sib-pairs
 x_2 = the number of 2 IBD for the sib-pairs with two affected siblings
 m_2 = the total number of 0 and 2 IBD for all the two affected sib-pairs

from these a Z statistic was calculated as

$$Z = \frac{wght[1]x_0 + wght[2]x_2 - \frac{1}{2}(wght[1]m_0 + wght[2]m_1)}{\frac{1}{2}(\sqrt{(wght[1])^2 m_0 + (wght[2])^2 m_1})}$$

where $wght[1] = 1$ and $wght[2] = 3$ were chosen to give good power properties for a range of likely alternatives. Under the null hypothesis of no genetic link, Z will have a standard normal distribution. (See Appendix 1 for further derivations)

The p-values for each SNP locus were then determined from the standard normal distribution and scaled to negative log base 10. After obtaining the p-values, it was determined which SNPs were above a chosen critical level, and appropriate measurements were taken for the SNPs that were above this level. A critical level of 2 was chosen, so any SNP which had a p-value less than 0.01 was taken to be significant. This implies that for each SNP there is only a 1% chance of it being critical if it is not linked to the disease gene.

A p-value was determined for each SNP for when there was a genetic link and when there wasn't. Critical values of the measurements were determined for a series of

measurements from the sib-pair groups where there wasn't a genetic. These critical values were then used to determine the power of this Allele Sharing method in finding the disease gene when it was genetically linked.

3 Tests of Genetic Association

To test if there was a genetic association four different methods of measurements were used for each group of sib-pairs:

1. Total Length – the sum of all the SNPs above a given critical level for the entire genome.
2. Maximum Length - the length of the longest string of consecutive values greater than the critical level.
3. Average – the average string length of consecutive values greater than the critical level.
4. Gene Length – the length of the string of consecutive values greater than the critical level that include the disease gene. If the disease gene is not above the critical value this length was taken as zero.

For each of these methods two units of measurement were used for the lengths:

- The number of SNPs.
- The total of the theta values for the range of critical SNPs. The theta values are the recombination frequencies between to consecutive SNPs. The unit of measurement is the centiMorgan, where 1 cM is equivalent to there being a 1% chance that two genes will be separated due to recombination in a single generation.

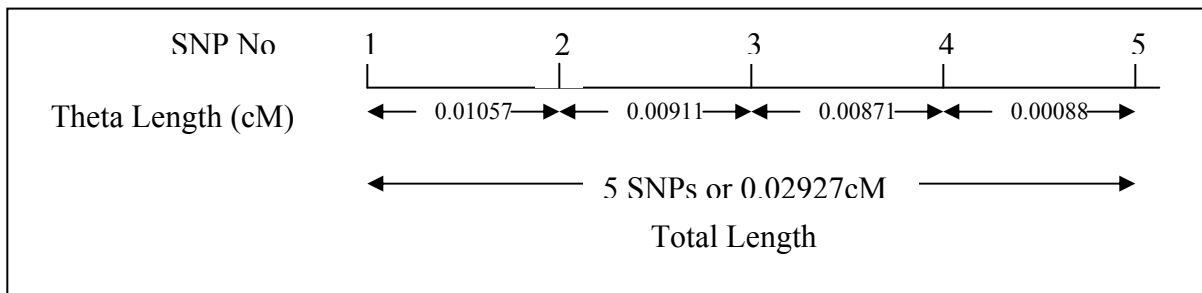


Figure 1 - The first 5 SNPs along chromosome 1 with the corresponding theta lengths

If only one SNP is critical then the length measured in SNPs is 1, however the theta length is 0.

Here is an example of the 4 length measurements:

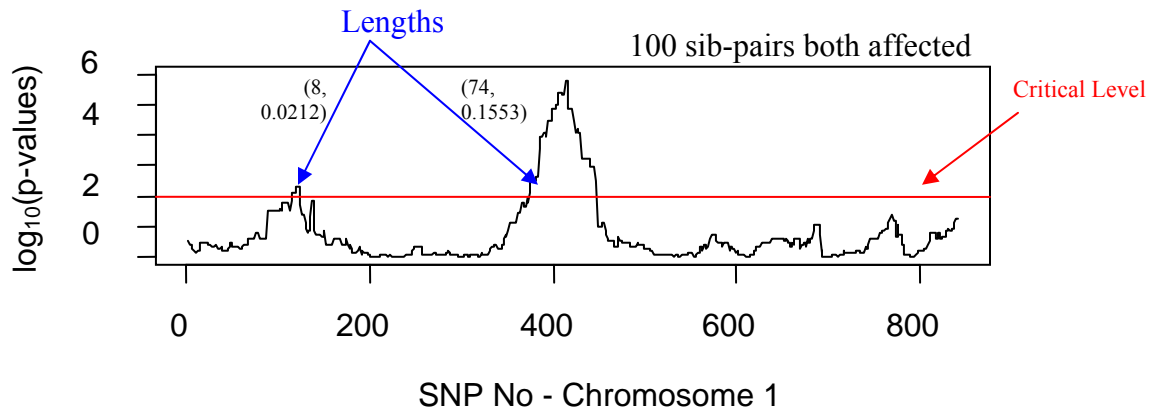


Figure 2 – Two lengths (SNPs, theta) were in the above group of sib-pairs where the disease gene was at position 401: Total Length = (82, 0.1765), Maximum = (74, 0.1553), Average = (41, 0.08825) and Gene Length = (74, 0.1533)

Methods 1 – 3 were used to determine whether or not there is a disease gene genetically linked within a group of sib-pairs. If the disease gene is genetically linked then it would be expected that due to the linkages there will be a greater number of critical SNPs, hence longer lengths around where the disease gene is located. If these methods show that there is a genetic link, then method 4 indicates whether the interval observed contains the disease gene. One other test to identify how well the DS gene would be detected was to consider whether or not the p-value at the DS gene was critical.

The p-values were initially calculated using a χ^2 statistic for the observed IBD proportions at each SNP location for 0, 1 and 2 IBDs, where the null hypothesis was that there was no genetic link, therefore the ratio of 0, 1 and 2 IBDs were 25:50:25 respectively. This method worked accurately when either all the sib-pairs had only one sibling affected or when all pairs had both the siblings affected. When there was a combination of the number of affected siblings in each pair two χ^2 statistics needed to be calculated, one for the group of sib-pairs where only 1 was affected and another for the group of sib-pairs where they were both affected. It was found that these two χ^2 statistics could not just be added together as if the two groups were independent, as if we had a group of 1 Affected sib-pairs and then we added some 2 Affected sib-pairs to this group the power would actually drop.

3 Critical Values for each test

The critical values were determined by firstly taking 10,000 groups for each model where the group size and proportion of sib-pairs that were 1 affected was varied, where the disease gene wasn't genetically linked. For each of the 10,000 groups each of the four measurements were taken for both the number of SNPs and the theta lengths. For each measurement and group size the 90th, 95th and 99th percentiles were calculated. The mean, standard deviation and median were also calculated to look at the distribution of

these measurements. A 95% confidence interval was obtained about the 95th percentile using ordered statistics.

When there is no genetic link there should be no significant difference between the four models tested for a given number of sib-pairs, therefore for each group size the 90th, 95th and 99th critical values were determined by taking the median over all the given percentile values of the models. The critical values can be view in Appendix 2.

4 Alternative Models Used

The models used were:

DS Genotype	Strong	Moderate	Weak	HNPCC
pAffected[AA]	0.7	0.4	0.18	0.8
pAffected[AB]	0.7	0.4	0.18	0.8
pAffected[BB]	0.035	0.0296	0.0167	0.004
Freq. A Alleles	0.05	0.1	0.3	0.00125
Prevalence	0.1	0.1	0.1	0.006

Table 1 – Each model represents 4 possible disease genes. pAffected[AA], pAffected[AB] and pAffected[BB] represent the probability of developing CRC given the genotype, Freq. A Alleles is the probability of the disease gene being passed onto the next generation, and the prevalence is the population prevalence for the disease.

For the strong, moderate and weak models the population prevalence was set at 10% since the siblings that are going to be selected for this method of analysis have a known family history of CRC, and within this population there is a 10% prevalence of the disease¹. It is expected that the disease gene is dominant so the probability of developing the disease given genotypes AA and AB were assigned the same value. The values of AA and the frequency of the A allele were determined for the first 3 models to ensure a range of values that contains high, medium and low penetrance of the disease gene. The value for BB was determined by solving

$$prev = (Pr A)^2 pAffected[AA] + 2 Pr A(1 - Pr A) pAffected[AB] + (1 - Pr A)^2 pAffected[BB]$$

The HNPCC model was chosen where the parameters are related to a gene (MLH1) which causes a hereditary colorectal cancer known as hereditary non-polyposis colorectal cancer. This model was chosen to determine how effectively the code detects a known disease gene.

These parameters were determined by:

- 6% of CRC cases are HNPCC²

¹ Weitz (2005, p183); Peel (2000, p 1517)

² Lynch & de la Chappelle (2003, p.925); Peel (2000, p 1517); Aaltonen(1998, p.1482)

- 0.25% of people in the population carry the MLH1 or MSH2 mutation, where 0.20% have been diagnosed with HNPCC and 0.05% who have not.
- 80% of carriers develop the disease

Hence the population frequency for each mutant allele MLH1 and MLH2 was assumed to be $0.25\%/2 = 0.125\%$. As there is a 6% lifetime risk of CRC in the whole population, 10% of 6% is 0.6% which is the population prevalence of HNPCC for families chosen with a history of HNPCC.

For each of the given models we also looked at different number of sib-pairs per group and different proportions of 1 affected sib-pairs and 2 affected sib-pairs. The number of sib-pairs analysed per group were 100 and 250. For each group size the proportion of one or two affected varied as follows: all 1 affected, 90% 1 Affected /10% 2 Affected, 50% 1 Affected/50% 2 Affected, and all 2 Affected.

In total there were 16 different combinations to consider for when the disease gene was and wasn't genetically linked – 4 models, 2 group sizes and 4 different proportions of the number affected.

For the given models I only looked at the first 841 SNPs (Chromosome 1) and the DS gene was placed in between the 400th and 401st SNP.

When there was no genetic link the probability of developing CRC is independent of the genotype at this location, hence the probably of developing the disease given your genotype was all set to the value of the population prevalence. So in the first three models $p_{\text{Affected}}[\text{AA}] = p_{\text{Affected}}[\text{AB}] = p_{\text{Affected}}[\text{BB}] = 0.1$

6. Power of test against each alternative model

To determine the effectiveness of the Alleles Sharing method, 1000 sib-pairs where the disease gene was genetically link were generated for each of the 16 combinations of models. For each different model the power was determined for the different measurements given the critical values calculated (90, 95 and 99). The power was also determined for the number of times the length around the disease gene exceeded the critical value for maximum length. Appendix 3 contains all the power results.

The powers of all the different length measurements for each model were similar in strength. However the maximum length measured in the number of SNPs provided the best power overall. It was also statistically better to measure the length around the disease gene in SNPs, rather than centiMorgans.

Model	No. Sib-pairs	Proportion
Strong	100	0
	100	0.5
	250	0*
	250	0.9
	250	0.5**
Moderate	100	0
	250	0*
	250	0.5**
Weak	-	-
HNPCC	100	0*
	250	0*
	100	0.5**
	250	0.9
	250	0.5*

Table 2 – Models that obtain good power results, where * represents the best results and ** the second best.

A model was considered to have good power results if the power at the 90th percentile critical value was greater than 80%. Under this definition the Weak model did not have any good results for the two sizes. The best results for this model was when there was 250 families with all the siblings affected, the power was 0.76, 0.653 and 0.336 for the 90, 95 and 99 critical values respectively. For all the other combinations for this model the power was less than 0.5 at all levels. This was not unexpected as the prevalence of this disease allele was quite low in this model.

The HNPCC model gave low powers (less than 0.225) when there were no siblings that were 2 affected. Due to this, the power was looked at for when there was only 15 and 25 siblings where both siblings were affected for the HNPCC model. The critical values for these group sizes were calculated in the same way as the critical values. It was found that reasonable power results were even obtained for only 25 sib-pairs. The power and critical values for these results can be found in Appendix 4.

By looking at the powers for all the different measurements the power for each one was very similar, however it was marginally better to use the maximum length measured in the SNPs as opposed to centiMorgans.

To determine whether it would be beneficial to take the genotypes of the parents where possible so that the true IBDs could be obtained, these same methods were performed with the true IBDs for the Strong and Moderate models. In the results there was no significant improvement in the powers; hence it will be sufficient to use the estimated IBDs. (Appendix 8)

7. SibRave for Sib-Pair data

The second method was SibRave Analysis, which was developed by Harri Kiiveri of CMIS in Perth, which uses the HG (“Holy Grail”) code to find a small set of useful predictors from a very large set. This code was adapted for sib-pair analysis to find regions of the genome that are linked with the disease due to the IBD statuses at each SNP location of a group of sib-pairs.

The same random sib-pairs chosen for each group in the Allele Sharing were analysed using this method, however only the groups where there was a combination of both 1 and 2 affected sib-pairs could be analysed due to the way the method works. In this analysis the disease gene was also placed between the 400th and 401st SNP, however it was removed from the IBD statuses before they were analysed to best represents a real life situation. Firstly the number of predictors that were selected for each model was determined to see how well this method would select predictors when there was a disease gene present. The second measurement taken was the minimum distance from the disease gene (400.5) to the predictors that were chosen. It was expected that if there was a disease gene then the predictors would be closer to this point.

8. Results of apply SibRave

The results found from the SibRave Analysis can be found in Appendix 5. Firstly, the proportion of sib-pair groups in which predictors were chosen was analysed. The proportion of groups that had predictors chosen when there was no genetic link was similar to the corresponding model when there was a genetic link. When 100 sib-pairs were used where 90 had 1 sibling affected in each pair, the proportion of groups with predictors chosen lied within range of 0.75 and 0.781 when there was no genetic link, and between 0.719 and 0.756 for when there was a genetic link. For all four models less sib-pair groups had predictors chosen for when there was a genetic link. The median number of predictors chosen for each group was the same for each model. When 100 sib-pairs were used when 50% were 1 Affected, the proportion of groups which had predictors chosen lied between 0.922 and 0.97, where all the genetically linked models were had slightly higher proportions. For all 100 sib-pair models the median of the minimum distance to the disease gene was lower for the genetically linked models. However the standard error was quite large so there was no significant difference between them.

When there were 250 sib-pairs in each group the proportion of groups in which predictors were chosen did increase however there was still no major difference between whether the disease gene was genetically linked or not. The proportion of groups in which predictors for when there was a genetic link ranged from 0.958-0.992 and the when there wasn’t 0.952-0.993. The median number of predictors were again very similar and sometimes the same. The medium of the minimum distance to the predictors chosen was always lower for the genetically linked models. Larger differences were seen for the Stronger and HNPCC models.

Overall there was no way to determine whether or not a model was genetically linked from the number of predictors chosen. One benefit of this method was that if it was known that there was a genetic link then the interval of SNPs that need to be looked in to find the disease gene is usually smaller than the Allele Sharing method, for when there is 250 families.

[Note: The results of applying SibRave only when the Allele Sharing gives a significant result should be discussed here]

9. Discussion: implication for study design

The Allele Sharing method could effectively locate regions containing the disease genes when the disease genes have moderate to strong prevalence. However, for the weak model, even with 250 sib-pairs with all the siblings being affected the disease gene could still not be effectively located. It was also found with the HNPCC model the region in which the disease gene lied could be effectively when only 25 two affected sib-pairs, and reasonable powers were also obtained for 15 two affected sib-pairs.

With the SibRave Analysis, by looking at the proportion of groups in which predictors were chosen and the number of predictors for each group, it could not be determined whether or not there was a genetic link for both group sizes. When there was only 100 sib-pairs there was not any significant difference between the means for the minimum distance to the disease gene. There was a greater difference when there were 250 sib-pairs, and similar to the Allele Sharing this wasn't evident in the Weak model. In the SibRave Analysis, each SNP location had two chances of being selected. The number of predictors chosen was taken to be the number of these "paired" SNPs, divided by two. This will need to be looked into further as if 1 SNP is chosen this is actually taken as 0.5 predictors, and 1 predictor could actually present 2 different SNPs chosen.

Due to time, I was only able to look at one critical level of 2, the first chromosome and two sib-pair group sizes. Further analysis should look at different group sizes, different critical levels and also the condition where there is more than one disease gene. It would also be necessary to analyse the entire genome as opposed to only the first chromosome, and place the disease gene in different locations. In the Alleles Sharing method the IBDs of the disease gene genotype was also analysed in the results, however it would be more realistic to remove this before analysis was done.

10. Appendix:

Appendix 1 – Further information on the calculations of the Z statistic

Under the null hypothesis of no genetic link,

$$(x_0, x_1, x_2) \sim \text{Multinomial}(n, (1/4, 1/2, 1/4))$$

while under the alternative

$$(x_0, x_1, x_2) \sim \text{Multinomial}(n, (p_0, p_1, p_2))$$

where p_0 , p_1 and p_2 are determined by the model for the genetic link. For the models of interest, $p_1 \approx 0.5$, so that x_1 carries little information for distinguishing the models.

Influence is thus best based on the conditional distribution of x_0 given x_0+x_2 which is Binomial(x_0+x_2 , $1/2$) under the null hypothesis, and Binomial(x_0+x_2 , $p_0/(p_0+p_2)$) under the alternative.

For pairs with 2 Affected, $p_0 < p_2$, while for the pairs with 1 Affected, $p_0 > p_2$, hence a reasonable test statistic for the alternative is $w_1x_0+w_2x_2$, which will tend to be large when the genetic is strong.

For a particular alternative model, the optimal weights are proportional to

$$w_1 = \ln\left(\frac{p_0}{1-p_0}\right)$$

$$w_2 = \ln\left(\frac{q_2}{1-q_2}\right)$$

where $p_0 = \text{Pr}(0 \text{ IBD} \mid 1 \text{ affected})$
 $p_2 = \text{Pr}(2 \text{ IBD} \mid 2 \text{ affected})$

For the four models investigated the best weights:

Strong Model	1:2.8
Moderate Model	1:4.24
Weak Model	1:6.63
HNPCC Model	1:12.35

References

1. Weitz J, *Lancet*, 2005; 365, 153-165.
2. Peel D J, *Journal of the National Cancer Institute*, Vol 92, No. 18, September 20, 2000; 1517-1522
3. Gutt...
4. Aaltonen L, *The New England Journal of Medicine*, Vol. 338, No. 21, May 21, 1998; 1481-1487
5. Kemp Z, *Human Molecular Genetics*, 2004, Vol. 13, Review Issue 2; 177-185
6. Lynch H and de la Chapelle A, *The New England journal of Medicine*, 2003;348:919-932