Log-linear Modelling of SNP Haplotype Blocks

Adrian P Mander[1] and Aruna Bansal[1]

[1]GlaxoSmithKline, New Frontiers Science Park, Harlow, Essex, UK

Running Title: Log-linear Models of Haplotype Blocks

Corresponding Author: Adrian Mander MSc PhD, Glaxosmithkline, Mail Code HW8133, New Frontiers Science Park (South), Harlow, Essex, CM19 5AW. Tel: +44 1279 63 1203.

ABSTRACT

Algorithms for defining the presence and limits of haplotype blocks have been presented previously, using measures of linkage disequilibrium. In the current manuscript, a new log-linear framework is proposed for the identification of blocks. Our approach allows a reparameterisation and formalisation of existing methods, including a means to statistically test, via the likelihood ratio test, for the presence of block-like structure. Our method was applied to a data set of 76 SNP markers spanning a genomic interval of length 2.7Mb. Obvious blocks were verified by our approach. In addition, evidence for sub-structure within blocks was also detected, suggesting that blocks, if they exist, may be far more complex than previously assumed. It is hoped that our approach will provide the basis for a formal statistical evaluation of blocks and will facilitate comparisons among data sets.

Key Words: Haplotype blocks, linkage disequilibrium, SNPs, log-linear models, EM algorithm

INTRODUCTION

The abundance of single nucleotide polymorphisms (SNPs) and the limited power of single-locus analysis has led to increased use of haplotype-inference methods such as Clark's algorithm [Clark, 1990], the expectation-maximisation algorithm [Excoffier and Slatkin, 1995] and iterative-sampling algorithms to resolve phase ambiguity by both coalescent and non-coalescent models [Stephens et al., 2001], [Niu et al., 2002].

Recent studies [Daly et al., 2001], [Jeffreys et al., 2001], [Patil et al., 2001][Gabriel et al., 2002], [Twells et al., 2003], have shown that the human genome can be viewed in terms of haplotype blocks, given by discrete regions of high linkage disequilibrium (LD), and separated by shorter regions of low LD. Haplotype block identification has been conducted via evaluation of measures of LD, such as Lewontin's D`, as well as by methods of directly assessing evidence of recombination [Schwartz et al., 2003]. The corollary of the block concept was that a small proportion of the SNPs, the 'haplotype tagging' SNPs, should be sufficient to capture the majority of the haplotype structure contained in blocks genome-wide [Johnson et al., 2001].

We introduce a novel application of log-linear modelling, to establish block structure. Log-linear models have been used to form the basis of Bayesian priors in resolving phase [Morris et al., 2003], and to model different levels of linkage disequilibria with phase known [Huttley and Wilson, 2000]. We show that not only can the log-linear model describe the discrete islands of LD [Goldstein, 2001], but it can also identify smaller sub-fragments of high LD. Furthermore, we demonstrate the use of likelihood ratio testing, to statistically evaluate specific patterns of LD.

.

MATERIALS

The methods described below were applied to a data set consisting of a random sample of 150 subjects from the PRESTO (Prevention of REStenosis with Tranilast and its Outcomes) study [Holmes et al., 2000, Danoff et al]. These were genotyped across 76 SNPs spanning approximately 2.7Mb within and around the gene UGT1A1. These data and their analyses are described in detail elsewhere [Xu et al, in preparation].

METHODS

LOG-LINEAR APPROACH

Log-linear modelling provides a unified approach for jointly modelling the patterns of LD and haplotype diversity. Lack of diversity leads to fewer parameters in the statistically optimal model, and, due to collinearity, the number of parameters is never greater than the number of haplotypes. Traditionally, haplotype frequencies have been estimated under a saturated model, in which all loci and interactions are represented. However, the necessarily high number of parameters often leads to problems for fine-mapping and for translating the results into inferences on the patterns of LD.

One solution has been to reduce the high-dimensionality of the saturated model by fitting intermediate models. These contain more parameters than a model of complete linkage equilibrium but fewer parameters than the saturated model [Chiano and Clayton 1998], [Mander,

2001]. The intermediate log-linear model has been used to reduce the area of genetic association in the HLA region [Bitti et al., 2001] and it is this, the intermediate model, that is of interest when trying to describe the pattern of LD within genomic regions. In the current manuscript, we show how such models provide the framework for quantifying the patterns of LD and for defining haplotype blocks.

NOTATION

The SNPs are assumed to be ordered, but physical inter-marker distances are ignored. The i^{th} SNP is given by l_i. The log-linear models use the Wilkinson and Rogers notation [Wilkinson and Rogers, 1973], where factor variables (SNPs) are combined by "+", the independence symbol, and "*", the interaction symbol. For example, l_1+l_2 denotes independence between the first and second SNP, and l_3*l_4 denotes interaction between the 3rd and 4th.

LIKELIHOOD RATIO TEST

The p-value resulting from a likelihood ratio test (LRT) was used to measure the strength of LD. In the case of two SNPs, the LRT was calculated by comparing the log-likelihood of the model l_1*l_2 (in which there is LD between the SNPs) to the log-likelihood of the model l_1+l_2 (in which the two SNPs are in linkage equilibrium). The LRT was performed using *hapipf* [Mander, 2001], a function implemented in STATA [StataCorp. 2001]. As discussed below, it includes two options, *mv* and *mvdel* for the handling of missing data.

MISSING DATA

The methods described are applicable to unphased genotype data, with or without missing data. In the presence of missing data, two assumptions were explored. The first was that the data were missing completely at random (MCAR) [Little and Rubin, 1987]. Under this assumption, it was

appropriate to delete all subjects that had any missing data. Our *hapipf* option to do this was *mvdel*. Despite its ease of application, it was recognised that the assumption was likely to be broken with real data. The second, less stringent assumption was that values were missing at random (MAR). Under this assumption, the missing alleles were imputed via the EM algorithm, utilising observed alleles where possible. Our function to follow this second approach was *mv*. The relative impact of the two assumptions is discussed below.

APPLICATION TO HAPLOTYPE BLOCKS

By one frequently used definition, a group of two or more consecutive markers are considered to constitute a haplotype block if the following hold: the endpoint markers are in high LD, and the number of pairs in strong LD is at least 19 times the number of pairs of markers "with historical evidence of recombination", as measured by the magnitude of D', [Gabriel et al., 2002]. In log-linear model terms, an analogous approach might be a model containing at least 19 times as many interaction terms as main effects, conditional upon a significant interaction between the end SNPs. The only difference would be that the interaction terms, from the log-linear model, are estimated jointly using orthogonal parameters, making the '19' criterion more stringent. As an aside, it is noted that a more robust approach might be to apply the LRT to a comparison of the saturated model with the model of complete linkage equilibrium (LE), as this would automatically penalise the use of too many parameters by the loss in power of the LRT.

Wall and Pritchard (2003) described three criteria for assessing haplotype blocks derived using pair-wise measures LD. The first, 'coverage' gives the proportion of the genome spanned by

blocks. The second and third, as described below, can be described in terms of log-linear models.

THE HOLE CRITERION

One desirable feature of our approach is that haplotype blocks can be identified even if they contain holes. Holes arise when the outermost SNPs are not in strong LD with a SNP or multiple SNPs that lie in between [Wall and Pritchard, 2003]. To translate this to a log-linear framework, consider a triplet of markers l_1, l_2 and l_3. If l_1 and l_3 show high LD, but intervening pairs (l_1,l_2 and l_2,l_3) do not show high LD, as can happen with low frequency SNPs, then the situation may be described by the model $l_1*l_3+l_2$.

This representation can be extended to a fourth SNP, l_4, in a similar fashion. The latter model would contain a term representing interactions among the SNPs in strong LD, but no interactions with the "hole" SNP. For example if the hole occurs at SNP2 (variable l_2), then one model describing the block would be $l_1*l_3*l_4+l_2$. Alternatively, if the three-way interaction is not needed, then another suitable model might be $l_1*l_3+l_3*l_4+l_1*l_4+l_2$, where, again, SNP2 ($l_2$) is independent of the other 3 SNPs.

OVERLAPPING BLOCKS

Again, as defined by Wall and Pritchard (2003), a feature of certain blocks is the presence of SNPs that are assignable to more than one block. In the simplest case of 4 SNPs (l_1-l_4) and two overlapping independent blocks $l_1*l_2*l_3$ and $l_2*l_3*l_4$, the model may be given by

$l_1*l_2*l_3+l_2*l_3*l_4$. In real data, there may be a combination of holes and overlapping blocks. For example, the same model with a hole at SNP2 for the first block is $l_1*l_3+l_2*l_3*l_4$. The method is highly generalisable.

SEARCH ALGORITHMS

In our application, all of the log-linear models were fitted using an EM algorithm that resolves phase [Mander, 2001]. A dataset containing a large number of SNPs can be very computer intensive to analyse, and slow to estimate haplotype frequencies. In order to apply our methodology for haplotype block detection, two algorithms were implemented that alllow efficiency gains by requiring fewer SNPs per model. The first was used to identify haplotype block edges and the second was used to find the most parsimonious model by a step-wise approach.

EDGE DETECTION ALGORITHM

The first algorithm analyses a set of *"n"* ordered SNPs, labelled as $l_1, l_2, \ldots, l_n$. A window size for analysis is chosen at the start of the algorithm, and this can later be varied as part of a sensitivity analysis. In order to detect an edge within the window, the saturated model is compared to a model in which the LD can be described by two blocks. The latter model has a "+" symbol between a pair of neighbouring SNPs in an otherwise saturated model. In order to cover all possible positions of the haplotype block edge, a 'sliding window' of models is fitted across the marker set of interest..

For example, with a window size of 4 and working left to right, the following models may be compared using a likelihood ratio test.

$l_1+l_2*l_3*l_4$ vs $l_1*l_2*l_3*l_4$

$l_1*l_2+l_3*l_4$ vs $l_1*l_2*l_3*l_4$

$l_2*l_3+l_4*l_5$ vs $l_2*l_3*l_4*l_5$

$l_3*l_4+l_5*l_6$ vs $l_3*l_4*l_5*l_6$

and so on.

It can be seen that in the first window, the first alternative model has a non-centred edge, as it arises at the start of the sequence of interest. It has therefore fewer degrees of freedom than the subsequent windows shown.

If the p-value for the likelihood ratio test falls below the nominal threshold of p=0.05, then the saturated model is rejected, and an edge of a haplotype block has been detected. For example, if the saturated model is rejected at the 5% level for the comparison between $l_1*l_2+l_3*l_4$ and $l_1*l_2*l_3*l_4$, then there is an edge between SNP 2 and 3 or there is an edge at position 2.5.

It is noted that if a window size of two is chosen, then the method is equivalent to using pair-wise methods. Furthermore, as window size increases, the number of degrees of freedom (df) also increases. For example, a window size of 8 leads to a 225 df LRT ($=2^8-1-(2^4-1)-(2^4-1)$). In this way, power decreases with increasing window size.

SENSITIVITY ANALYSIS FOR EDGE DETECTION

Varying the window size allows some optimisation, and the optimal window size depends upon the size of the dataset. As shown later, for 150 subjects, a window size of 8 SNPs did not detect smaller blocks. In general however, it is acknowledged that Edge Detection, as defined here, relies upon the alternative model ('two blocks') being a reasonable approximation to the data. It is expected that if the true LD pattern encompasses non-distinct blocks or overlapping blocks, then some power of detection will be lost. In these instances, further investigation by the second (stepwise) algorithm below will be of value.

FORWARD STEPWISE ALGORITHM

The second search algorithm inolves a forward stepwise approach to determining the most parsimonious model of LD. Starting with a model of complete linkage equilibrium (LE), higher order LD terms are added sequentially to the model. This continues until a model is found that describes the observed pattern of LD using fewest parameters.

This approach can be used to detect patterns within a haplotype block and it provides a better description of the haplotype blocks in genomic areas in which the first ('Edge Detection') algorithm demonstrated evidence of a relationship between window length and the haplotype blocks detected.

The algorithm examines a window of *'n'* SNPs. In order to preserve efficiency of the EM algorithm, fewer than 8 SNPs is practical. The first step is to estimate the log-likelihood under the linkage equilibrium model $l_1+l_2+...+l_n$. Then, every pair-wise SNP interaction term is added to this model and the likelihood ratio test is re-evaluated. The most significant interaction

term is then added to the base model and the process repeats. A nominal p-value of p=0.05 is chosen to compare new models to the LE model. Once no more pair-wise interactions are significant, the algorithm proceeds to the next order of interaction terms, and so on. This approach accommodates the fact that pair-wise interactions can occur over greater distances than contiguous pairs and that LD does not decay monotonically with distance. Interest focuses on the "simplest" model to describe the pattern of LD. The algorithm continues until the highest interaction term is evaluated, the saturated model.

Holes, overlapping blocks, distant LD or contiguous blocks will all be detected in this very powerful set of models. At each step, the number of degrees of freedom is minimised in the sequence of likelihood ratio tests.

EXAMPLE OF THE STEPWISE APPROACH

The sequence of models for 3 SNPs, when only the l_2*l_3 interaction is significant, is,

$l_1+l_2+l_3$

$l_1*l_2+l_3$

$l_1*l_3+l_2$

$l_1+l_2*l_3$

$l_1*l_2+l_2*l_3$

$l_1*l_3+l_2*l_3$

$l_1*l_2*l_3$

In our experience of applying this algorithm, most haplotype blocks can be divided into smaller sub-blocks. One model of a sub-block might be $l_1*l_4+l_2*l_3$, in which the larger haplotype block l_1-l_4 has a sub-block l_2-l_3. In practice, there may be other pair-wise interactions , e.g. l_1*l3, that obscure this pattern.

RESULTS

All pair-wise |D'| statistics for the 76 SNPs were produced using STATA and the *pwld* command (available from David Clayton's website, http:\\www-gene.cimr.cam.ac.uk/clayton). Figure 1 displays estimates of all the pair-wise statistics. A few areas had very high |D'| values, given by the black squares, and these indicated a block-like structure for the first 41 SNPs. By contrast, the final 35 SNPs did not appear to have a clear block structure.

APPLICATION OF EDGE DETECTION

The edge detection algorithm was applied with window widths of 2, 4, 6 and 8. Even numbers were selected in order that the "+" sign in the alternative model would always be in the middle of the model. For each pair of neighbouring SNPs, a likelihood ratio test p-value was calculated and the p-value was assigned an inter-marker position. For example, in the test $l_1*l_2+l_3*l_4$ versus $l_1*l_2*l_3*l_4$ the "+" term is at position 2.5 i.e. between SNP 2 and SNP 3. Graphs of the p-values against inter-marker position are given in Figure 2 for every test evaluated. In this representation, flat horizontal lines with low p-value indicate blocks; interruptions denote edges of blocks. An example of a block, when using a window width of 4 SNPs is SNP1-SNP5 in Figure 2. Using these data, it was observed that the length of the haplotype block depended upon the window size. Generally the longer the window, the shorter the haplotype block, suggesting the need to apply our stepwise algorithm for clarification.

It was only by using a window of 8 SNPs that a haplotype block edge was detected between SNP22 and SNP24. This pattern was visible in the *pwld* graph (Figure 1), but the |D'| values

were so high that other methods might have considered this a single block. Similarly, all the window sizes detected a block somewhere between SNP1 and SNP5, but only window sizes of greater than 4 SNPs showed that the first SNP should not be included in the block. The largest discrepancies in haplotype block definition with varying window size occured between SNP67 and SNP73. From the *pwld* comparisons, high LD occured between SNP73 and SNP67, without much LD among intervening SNPs. Therefore, in this instance, the variability in predicted structure may be traced to the unsuitability of the alternative hypothesis of 'two hapotype blocks'.

A window width of 8 SNPs did not capture much LD structure downstream of SNP41 because there was not a sufficiently large area of high LD. This lack of detection again was a symptom of an inappropriate alternative hypothesis.

This procedure provided a very fast initial screen, taking about an hour on a 1gigaHz computer. The stepwise algorithm (results below) was more time-consuming to perform, but allowed the unravelling of more detailed structure.

EFFECT OF MISSING DATA

By taking a window size of 8 and the two missing data mechanisms *mv* (assuming MAR) and *mvdel* (assuming MCAR), the results in Figure 3 were obtained.

Major differences between Figure 2 and Figure 3 were seen for markers falling between SNP2 and SNP24, and this was attributable to the assumed missing data mechanism. The percentage of missing values at each point is given in Figure 3 and interestingly, for models that included

the "+" term at position 6.5, about 90% of the data was missing, but only 20% was missing at position 13.5. This suggested that the extent of missingness was not the only factor. Instead, greater impact was from the nature of the missingness; the fact that missing values were not a random sample of the full sample.

APPLICATION OF STEPWISE ANALYSIS

Based upon the results of the first algorithm, two regions appeared to have a more complex pattern of LD than the 'two-block' alternative model suggested. This was implied by the fact that different haplotype blocks were estimated according to window size. These regions were SNP67 to SNP73 and SNP46 to SNP50. The most parsimonious models was found for these two intervals and for an additional three areas of high LD, (SNP1 to SNP5), (SNP24 to SNP28) and (SNP12 to SNP17), to look for any sub-blocks.

STEPWISE ANALYSIS OF SNP67 TO SNP73

A sequence of 198 models was fitted and the output from the early stages - selection up to only the first significant pair-wise interaction - is given below in Table 1. In this analysis, SNP67 is labelled l_1, SNP68 is labelled l_2, and so on.

The likelihood ratio test statistics all have a single degree of freedom and the largest chi-squared value is 36.799, for the model $l_1+l_4+l_5+l_6+l_7+l_2*l_3$ versus a model of LE. This is highly significant, with $p<1 \times 10^{-8}$. The second term was selected after refitting all the models again (output not given) and so on. The most parsimonious model was

$l_5+l_6+l_2*l_3+l_2*l_7+l_1*l_7+l_1*l_3+l_1*l_2*l_4$.

According to this model, SNP71 (l_5) and SNP72 (l_6) were independent of all the other SNPs. The largest possible block detected was SNP67-SNP73, assuming that both SNP71 and SNP72 were "holes", because the model included the interaction term l_1*l_7. However, smaller blocks were also identified, namely SNP67-SNP69 and SNP67-SNP70. Indeed, most of the parameters of the model were within the SNP67-SNP70 block. It was notable that only one three-way interaction term was needed ($l_1*l_2*l_4$, corresponding to SNP67, SNP68 and SNP70), and that SNP69 was not in this interaction despite being related, in a pairwise fashion, to SNP67 and SNP68. SNP69 may define a "hole" in this block as seen from the *pwld* graph in Figure 1.

STEPWISE ANALYSIS OF OTHER SUB-REGIONS

A look at the other regions showed evidence of a large number of overlapping blocks. For SNP24-SNP28, the most parsimonious model was $l_1*l_4+l_4*l_5+l_1*l_3+l_2*l_5+l_2*l_3$. This complicated structure appeared to correspond to two overlapping blocks, l_1-l_4 and l_2-l_5. Similarly, for SNP1-SNP5 the most parsimonious model was $l_2*l_3+l_3*l_5+l_3*l_4+l_1*l_4$, giving a clearer pair of overlapping blocks, SNP1-SNP4 and SNP3-SNP5 (l_1-l_4 and l_3-l_5). For SNP12-SNP17, the most parsimonious model was $l_3*l_4+l_4*l_6+l_1*l_2+l_1*l_4+l_1*l_5$, suggesting another pair of overlapping blocks, consisting of SNPs12-16 and 15-17. Lastly, for SNP46-SNP50, the most parsimonious model was $l_2*l_4+l_1*l_5+l_1*l_2*l_3+l_3*l_4*l_5+l_1*l_3*l_4$. All but one two-way interaction term was significant here. There was clearly a block SNP48-SNP50, but SNP46 was also related to all the other SNPs independently. This pattern was not clear in the *pwld* graph (Figure 1), although it was seen using the Edge Detection algorithm. The large number of

parameters suggests that this is an area of high haplotype diversity and that a haplotype diversity approach to block identification may have missed this structure.

DISCUSSION

This study shows that log-linear modelling can successfully identify areas of high LD and the blocks they encode. This approach allows likelihood ratio testing to be applied, and thus it provides a robust framework within which to investigate haplotype block definitions. The models described elucidate how many SNPs (and subsequent interactions) are needed to describe the pattern of LD within a region. That is they provide the route to 'haplotype tagging' SNPs or htSNPs. At the same time, they identify SNPs that are uninformative, in the sense of failing to add any information about LD within a region.

Earlier evaluations of partial LD models have been made. One group commented on the exceeding complication arising from the inclusion of higher order interactions [McPeek and Strahs, 1999]. Model complexity is indeed an outcome of applying this method to a large window size. However, it is clear that areas of very high LD can more often be described in terms of lower order interactions, with very specific meaning. This benefit is attributable to the orthogonal parameterisation of the log-linear model.

The power of haplotype anlysis has been improved previously by looking at subsets of haplotypes [Fallin et al., 2001]. However, a benefit of our approach is that the dimensionality of the data is preserved while only the dimensionality of the model is reduced. Furthermore, there

is no need to combine rare haplotypes, because of the relationship between haplotype diversity and the number of parameters in the model.

In most studies, the next step, having genotyped in full, the subset of most highly informative markers, is to identify a single SNP that is associated with a disease. This step also can be accommodated within the current framework. Given the most parsimonious model, disease status can be added to the model under the assumption of no association (as an independent factor), and compared to a model in which disease status interacts with all other terms. It is predicted that this approach would be much more powerful than the usual comparison with the saturated model, employed by other EM algorithms.

Our results support the view that areas of high LD probably form discrete haplotype blocks [Schwartz et al., 2003]. However, the likely presence of overlapping sub-blocks, indicates that the true block structure in the human genome may be more complex than the original vision of discrete blocks of uniformly high LD.

ACKNOWLEDGEMENTS

The authors would like to thank members of the Statistics and Programming Department, the Population Genetics Department and the Discovery Genetics Department at GlaxoSmithKline, for helpful discussions and input, particularly Chun-Fang Xu, for generating and sharing the data described here.

REFERENCES

Bitti PP Murgia BS, Ticca A, Ferrai R, Musu L, Piras ML, Puledda E, Campo S, Durando S, Montomoli C, Clayton DG, Mander AP, Bernardinelli L. 2001. Association between the ancestral haplotype HLA A30*B18*DR3 and multiple sclerosis in Central Sardinia. Genetic Epidemiology 20:271-283.

Chiano MN, Clayton DG. 1998. Fine genetic mapping using haplotype analysis and the missing data problem. Ann.Hum.Genet. 62:55-60.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol.Biol.Evol. 7:111-122.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. Nat.Genet. 29:229-232.

Danoff TM, Campbell DA, McCarthy LC, Lewis KF, Cantone KL, Repasch MH, Kazierad DJ, Saunders AM, Spurr NK, Purvis IJ, Roses AD, Xu CF. A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia. In press.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol.Biol.Evol. 12:921-927.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and Alzheimer's disease. Genome Res. 11:143-151.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Goldstein DB. 2001. Islands of linkage disequilibrium. Nat.Genet. 29:109-111.

Holmes D, Fitzgerald P, Goldberg S, LaBlanche J, Lincoff AM, Savage M, Serruys PW, Willerson J, Granett JR, Chan R, Shusterman NH, Poland M. 2000. The PRESTO (Prevention of restenosis with tranilast and its outcomes) protocol: a double-blind, placebo-controlled trial. Am. Heart J. 139: 23-31.

Huttley GA, Wilson SR. 2000. Testing for Concordant Equilibria Between Population Samples. Genetics 156, 2127-2135.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. 29: 217-222.

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, DiGenova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat.Genet. 29:233-237.

Little RJA, Rubin DB. 1987. Statistical analysis with missing data.Wiley: New York.

Mander AP. 2001. Haplotype analysis in population-based association studies. Stata Journal 1:58-75.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. Science 294(5547):1719-23.

McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am.J.Hum.Genet. 65:858-875.

Morris A, Pedder A, Ayres K. 2003. Linkage Disequilibrium assessment via log-linear modeling of SNP haplotype frequencies. Genetic Epidemiology 25:106-114.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. Am.J.Hum.Genet. 70:157-169.

Schwartz R, Halldorsson BV, Bafna V, Clark AG, Istrail S. 2003. Robustness of inference of haploype block structure. J.Comp.Biol. 10:13-19.

StataCorp 2001. *Stata Statistical Software: Release 7.0*. College Station, TX: StataCorp LP.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am.J.Hum.Genet. 68:978-989.

Twells RCJ, Mein CA, Phillips MS, Hess JF, Veijola R, Gilbey M, Bright M, Metzker M, Lie BA, Kingsnorth A, Gregory E, Nakagawa Y, Snook H, Wang WY, Masters J, Johnson G, Eaves I, Howson JM, Clayton D, Cordell HJ, Nutland S, Rance H, Carr P, Todd JA. 2003. Haplotype

structure, LD blocks, and uneven recombination within the LRP5 gene. Genome Res. 13:845-855.

Wall JD, Pritchard JK. 2003. Assessing the Performance of the Haplotype Block Model of Linkage Disequilibrium. Am.J.Hum.Genet. 73:502-515.

Wilkinson G, Rogers C. 1973. Symbolic description of factorial models for analysis of variance. Applied Statistics 22:392-399

Xu CF, Lewis KF, Yeo AJ, McCarthy LC, Maguire MF, Caine C, Anwar Z, Danoff TM, Roses AD, Purvis IF. Identification of pharmacogenetic effect by linkage disequilibrium mapping. In preparation.

.

List of Figures

Figure 1

Graph depicting all pair-wise |D'| statistics for the 76 SNPs, produced using STATA and the *pwld* command (available from David Clayton's website, http:\\www-gene.cimr.cam.ac.uk/clayton).

Figure 2

Graph showing p-values derived from the Edge Detection LRT (option *mvdel*), plotted against inter-marker edge position. Results from window widths of 2, 4, 6 and 8 SNPs are shown. Flat areas of the graph with low p-value denote evidence for block structure; interruptions raising above p=0.05 denote block edges.

Figure 3

Graph showing p-values, derived from the Edge Detection LRT, using a window size of 8 SNPs. The results from using the *mv* function (assuming MAR) are given by a broken line. The results from using the Mvdel function (assuming MCAR) are given by a solid line. The proportion of missing data is also plotted for each marker.

Figure 1

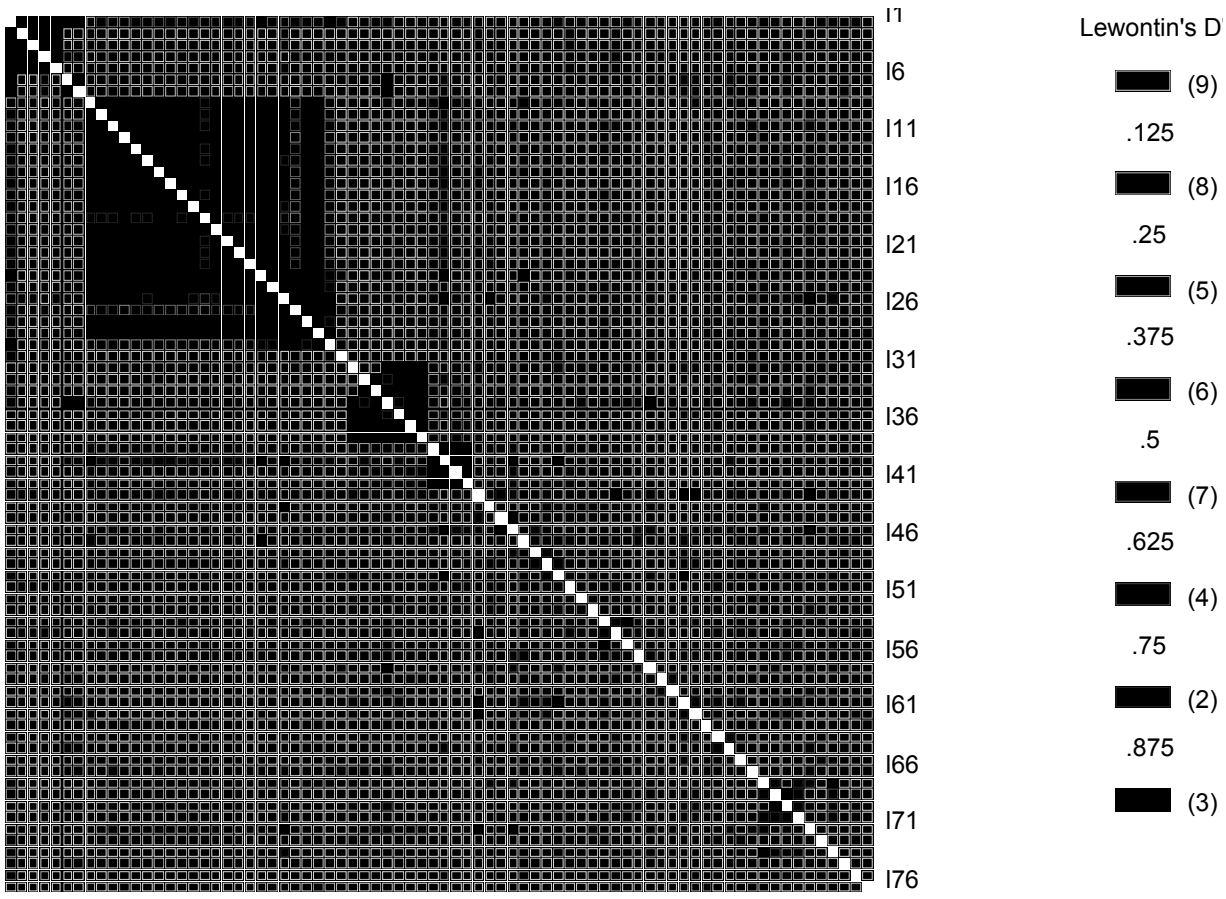
I1
I6
I11
I16
I21
I26
I31
I36
I41
I46
I51
I56
I61
I66
I71
I76
Lewontin's D
(9)
.125
(8)
.25
(5)
.375
(6)
.5
(7)
.625
(4)
.75
(2)
.875
(3)

Figure 2

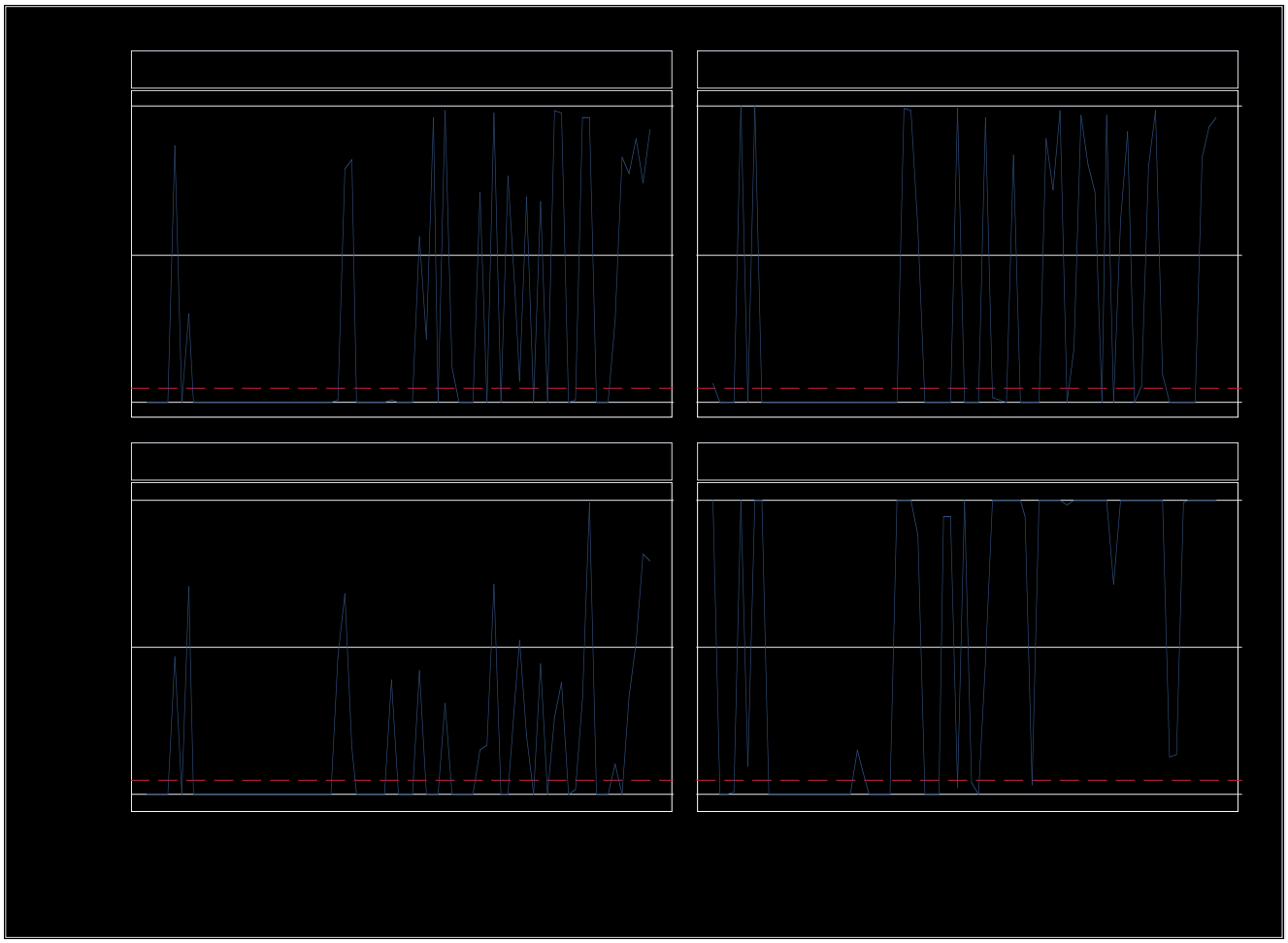

Figure 3

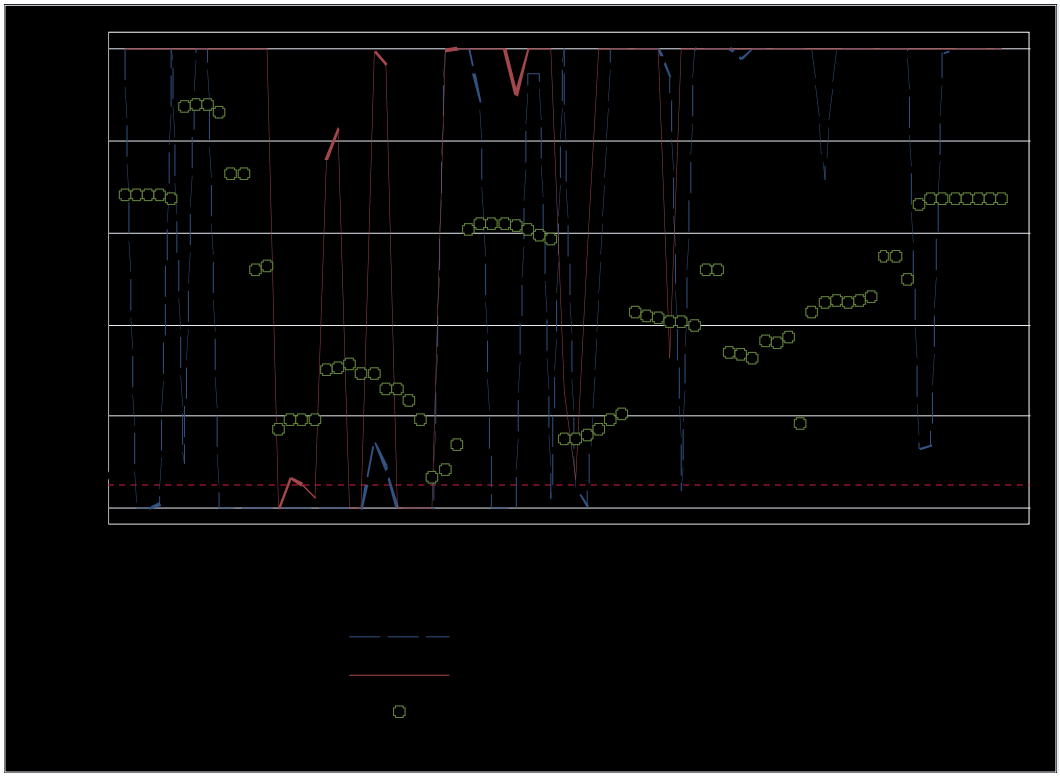

Table 1

First part of the sequence of models fitted for the Stepwise Analysis of SNP67 to SNP73. Output of the likelihood ratio test (LRT) is given, up to selection of the first significant pair-wise interaction. Each test had one degree of freedom.

Model	Loglikelihood	LRT statistic compared to LE	LRT p-value
$l_1+l_2+l_3+l_4+l_5+l_6+l_7$	-337.90253	N/A	N/A
$l_1+l_2+l_3+l_4+l_5+l_6*l_7$	-337.86635	.7879168	.07237
$l_1+l_2+l_3+l_4+l_6+l_5*l_7$	-337.90155	.96461774	.00196778
$l_1+l_2+l_3+l_4+l_7+l_5*l_6$	-337.4069	.31943159	.99127237
$l_1+l_2+l_3+l_5+l_6+l_4*l_7$	-333.16955	.00209319	9.4659637
$l_1+l_2+l_3+l_5+l_7+l_4*l_6$	-337.54557	.39814589	.71391891
$l_1+l_2+l_3+l_6+l_7+l_4*l_5$	-337.54187	.39571111	.72132048
$l_1+l_2+l_4+l_5+l_6+l_3*l_7$	-336.15043	.06121324	3.5042048
$l_1+l_2+l_4+l_5+l_7+l_3*l_6$	-337.84583	.73629892	.11340632
$l_1+l_2+l_4+l_6+l_7+l_3*l_5$	-337.65884	.4850981	.48738058
$l_1+l_2+l_5+l_6+l_7+l_3*l_4$	-329.69117	.00005067	16.422719
$l_1+l_3+l_4+l_5+l_6+l_2*l_7$	-329.55224	.00004377	16.700579
$l_1+l_3+l_4+l_5+l_7+l_2*l_6$	-337.89887	.93177705	.00732894
$l_1+l_3+l_4+l_6+l_7+l_2*l_5$	-337.42341	.32762735	.95825335
$l_1+l_3+l_5+l_6+l_7+l_2*l_4$	-320.05663	2.311e-09	35.691807
$l_1+l_4+l_5+l_6+l_7+l_2*l_3$	-319.50278	1.309e-09	36.799506
$l_2+l_3+l_4+l_5+l_6+l_1*l_7$	-330.32187	.0000987	15.161334
$l_2+l_3+l_4+l_5+l_7+l_1*l_6$	-337.46949	.35204097	.86608668
$l_2+l_3+l_4+l_6+l_7+l_1*l_5$	-336.01851	.05224042	3.7680529
$l_2+l_3+l_5+l_6+l_7+l_1*l_4$	-334.89927	.01425307	6.0065251
$l_2+l_4+l_5+l_6+l_7+l_1*l_3$	-337.81308	.67231003	.17891188
$l_3+l_4+l_5+l_6+l_7+l_1*l_2$	-334.12801	.00600418	7.5490473