Department of Computer Science
Series of Publications C
Report C-2010-5

Linkage Disequilibrium between Chromosomes in the Human Genome: Test Statistics and Rapid Computation

Ella Bingham, Mikko Koivisto, Yrjö Leino, Heikki Mannila

University of Helsinki
Finland

Linkage Disequilibrium between Chromosomes in the Human Genome: Test Statistics and Rapid Computation

Ella Bingham, Mikko Koivisto, Yrjö Leino, Heikki Mannila

HIIT Basic Research Unit
Department of Computer Science
P.O. Box 68, FIN-00014 University of Helsinki, Finland
ella.bingham@cs.helsinki.fi
mikko.koivisto@cs.helsinki.fi
yrjo.leino@csc.fi
heikki.mannila@cs.helsinki.fi

Technical report, Series of Publications C, Report C-2010-5
Helsinki, April 2010, 27 pages

Abstract

Linkage disequilibrium (LD) refers to the statistical dependency of the DNA content at nearby locations of the chromosome. Numerous approaches to analyze genome data rely on the well documented fact that LD decays monotonously with the distance of the studied loci. This decay, though noisy and modified by a number of factors, can be attributed to the recombination process, a major source of genetic variation in diploid organisms.

In this work we take first steps toward analyzing the extent of LD between very distant loci, even loci from different chromosomes. This is in contrast to traditional "genome-wide" analyses which merely study the LD within each chromosome separately.

We design several measures of LD, and use them for analyzing the HapMap data. We also consider LD between supermarkers determined by haplotype clusters in windows of a few SNPs.

We report on suggestive pairs of loci where unusually large correlations are observed within all ethnic groups.

We describe how the computations can be arranged in a way that enables an all-pairs analysis of the data, that is, all pairs of loci across all the 22 autosomal chromosomes. This kind of "genome times genome" analysis is computationally very burdensome due to the sheer number of possible pairs. We show ways to make it feasible.

Computing Reviews (1998) Categories and Subject Descriptors:

G.3 Probability and Statistics: Multivariate Statistics, Probabilistic Algorithms, Statistical Computing

I.5.3 Pattern Recognition: Clustering

General Terms:
Algorithms, Experimentation

Additional Key Words and Phrases:
Clustering, Genome Structure, Linkage Disequilibrium, Mixture Models

1 Introduction

In this study we consider dependencies between genomic regions. A widely used measure, linkage disequilibrium (LD), refers to the fact that neighboring areas in a chromosome are dependent of each other. The dependence stems from recombination: when genetic material is switched between the members of a chromosome pair, neighboring locations have a higher probability of staying close to each other. LD between two loci decays as the distance between the loci increases, and very far away loci are typically assumed independent of each other.

Our goal is to find out whether there are distant loci whose contents are correlated with each other. A significant LD between two far-away loci, perhaps across two different chromosomes, suggests that there might be a functional explanation behind it: the two loci affect a phenotype characterstic to the studied population through interactions.

There are some obvious computational problems involved when trying to compare all pairs of loci across all chromosomes. To this end, we develop summary statistics and arrange the computations in an efficient way.

We mostly concentrate on comparing two SNPs at a time. In addition we will consider summary statistics: "supermarkers", clusters of haplotypes in windows of a few SNPs. The clusters are found by a Bernoulli mixture model.

We analyze the HapMap data consisting of 209 individuals from four ethnic groups, and we indeed find some pairs of loci whose mutual dependence is unusually high. We also control against population effects, so that the observed correlations are not just due to one population being different from another, but to, e.g., interactions between genes.

We also briefly touch upon the Perlegen data consisting of three ethnic groups, and demonstrate some small-sample effects that arise.

This paper is organized as follows. We first briefly introduce the data. We then detail the steps of our analysis, in particular the testing for LD between the loci. We also discuss the computational issues and present some solutions that we have found useful. We then present results on which kind of dependencies are found and where. We conclude with a discussion, with pointers to further research.

2 Materials and methods

2.1 The data sets

We analyzed the Hapmap data [The03], http://www.hapmap.org, phase I (March 2005). The data consist of 45 Han Chinese (HCB), 44 Japanese (JPT), 60 European (CEU) and 60 Yoruban (YRI) samples. Among the CEU and YRI populations, the data was originally given in trios consisting of two parents and a child, making 90 samples in both groups, but we discarded the children in order to have independent samples.

We also analyzed the Perlegen data set [HSN+05] that consists of the genotype of 71 individuals: 23 African Americans, 24 European Americans, and 24

Han Chinese. In total the Perlegen data contain 1.6 million SNPs.

In the remainder of the document, we present results on the HapMap data, unless otherwise noted.

Different subpopulations contain partially different markers or SNPs. We have taken markers which are found in all subpopulations. In the HapMap data, the number of SNPs in each chromosome is shown in Table 1.

chr	SNP
1	49221
2	57155
3	41720
4	37514
5	37492
6	42700
7	28413
8	48863
9	38271
10	31131
11	28238
12	26652
13	23514
14	18453
15	16200
16	14949
17	15001
18	25871
19	10838
20	12721
21	13143
22	11900

Table 1: Number of markers per chromosome, HapMap data

In some cases we also preprocess the data such that we compute the minimum allele frequency (MAF) of each marker, and remove the markers for which the MAF is smaller than 5 per cent. A marker whose MAF is very small assumes almost constant values throughout the population and, thus, has a low information content: with a limited sample size, such a marker cannot participate in statistically significant correlations with other markers.

2.2 Outline

Our measures of LD are based on comparing the haplotype distribution at two loci. For each pair of two loci, A and B, we compute the value of a test statistic $s(A, B)$ that measures the statistical dependency of a random individual's genotypes on the loci A and B. We call the pair (A, B) statistically significant if the test statistic $s(A, B)$ exceeds a certain threshold θ. The parameter θ is to be fixed and will be considered in more detail later.

2.3 Statistical significance

Before we start, let us mention a few important issues regarding the validity of our results, in the light of statistical significance.

Our first concern are the populations. As there are four distinct populations in the HapMap data (and three in the Perlegen data), it can be expected that some distant loci are in strong association in the combined data solely due to the population structure. For example, the allele frequencies at a SNP can be very different in different populations. Such associations, however, are not very interesting, for they can be explained by "neutral variation."

We therefore examine associations separately in each of the populations. Because of the small sample size, we restrict our attention to pairs of loci for which the association is relatively strong in all populations.

Another concern worth discussing in the beginning is that of multiple testing. Over the course of all-pairs analysis of loci from two chromosomes we will eventually perform millions of statistical tests. A natural question then is, do we find some statistically significant results merely by chance, even though there is no dependency in the data?

The problem of multiple testing is often tackled by e.g. Bonferroni correction or Fisher's combined probability test. Both of these approaches assume that the tests are independent, which is not the case in our setting.

Instead, we will rely on permutation tests: we randomly permute the individuals in one locus and leave the other locus intact, and test for the dependency between the loci. This is repeated, say, 100 or more times. If a dependency is found in a significant portion of the runs, we can conclude that the observed phenomenon is not a characteristic of the original data, but just a spurious correlation.

Also, in cases where exhaustive permutation tests are too burdensome, we choose a very small value for P, far smaller than the usual P=0.05 limit.

Let us briefly mention a recent paper in which a somewhat similar problem was addressed. Misawa and Kamatani [MK09] discuss the testing of allele frequency differences between case and control populations in genome-wide association studies (GWAS). A common approach is to test for differences in the allele frequencies of every single-nucleotide polymorphism (SNP) between the case and the control populations. Misawa et al [MFY$^+$08] developed haplotype-based algorithms to correct for multiple comparisons. Instead of permutation tests, Misawa and Kamatani developed a set of computer programs for the parallel computation of accurate P values in haplotype-based GWAS.

We now turn to the technical measures of linkage disequilibrium.

2.4 Measures of LD: Contingency tables

Given two loci A and B, we want to map the haplotype data at A and B to a real number that measures the statistical dependency of an individual's genotypes at the loci A and B in the sampled population.

Our measures of dependency are based on the contingency table where the entry at uv counts the number of individuals having genotype u at locus A and

genotype v at locus B. Here the genotypes on A are understood as the possible pairs of haplotypes on A, labelled arbitrarily; similarly for locus B. An example of such a table is given in Table 2. In the table we assume that loci A and B both consist of one SNP only: at locus A the possible alleles are A and a, and at locus B they are B and b.

		locus B			
		BB	Bb	bb	Total
locus A	AA	$O(\mathrm{AA}, \mathrm{BB})$	$O(\mathrm{AA}, \mathrm{Bb})$	$O(\mathrm{AA}, \mathrm{bb})$	$N(\mathrm{AA})$
	Aa	$O(\mathrm{Aa}, \mathrm{BB})$	$O(\mathrm{Aa}, \mathrm{Bb})$	$O(\mathrm{Aa}, \mathrm{bb})$	$N(\mathrm{Aa})$
	aa	$O(\mathrm{aa}, \mathrm{BB})$	$O(\mathrm{aa}, \mathrm{Bb})$	$O(\mathrm{aa}, \mathrm{bb})$	$N(\mathrm{aa})$
	Total	$N(\mathrm{BB})$	$N(\mathrm{Bb})$	$N(\mathrm{bb})$	N

Table 2: A contingency table of possible genotype values. Both loci A and B consist of one diploid SNP. By $O(i, j)$ we denote the number of individuals having genotype i at locus A and genotype j at locus B.

It is a standard procedure to measure the linkage disequilibrium by the deviation from independence assumption in the contingency table. Assuming the loci A and B independent of each other, the uvth entry of the contingency table should be the product of the uth and vth row and column marginals, divided by the total number of individuals. If the entries deviate substantially from these expected counts, we reject the independence assumption and state that there is a dependence between the loci A and B, and this dependence controls the distribution of the entries in the contingency table. Widely used statistical tests include the Chi squared goodness-of-fit test and an exact binomial test. Of these, the Chi squared test is computationally simpler but it assumes that the counts are approximately normally distributed, posing some restrictions on the entries of the table. For this reason, we choose the exact binomial test, to be discussed in the following subsection. Other numerical measures have been devised, too, and we will present one of them in the sequel.

We note that with a sample of N individuals, the number of different possible 3×3 contingency tables with 9 entries is $\binom{N+8}{8}$. For $N = 24$ and $N = 60$, this equals to 10 518 300 and 7 392 009 768 different tables. Accordingly, with current powerful computers, it is quite conceivable to run statistical analysis for the complete ensemble.

Binomial test. The number of counts in each cell of the contingency table is binomially distributed: $O(i, j) \sim \mathrm{Bin}(N, p)$. Here N, the total number of trials, is given by the total number of counts, and p, the probability of falling into cell ij, is given by the product of the ith and jth row and column marginals, divided by N^2.

In the binomial test, we measure the tail probability P of a binomial distribution at the value given in entry ij of the contingency table, assuming the mean of the binomial distribution is given by the expected count in entry ij. The tail probability is computed at each entry of the contingency table, and

the smallest probability is returned. In case the smallest probability is very small, we conclude that the contingency table deviates from the independence assumption.

A possible drawback of the binomial test is that only the cell giving the smallest tail probability is taken into account. Suppose that there are two cells whose tail probability is quite small but not smaller than a predefined limit — we end up ignoring this table although the independence assumption might well be violated. However, due to the sheer number of tests that we will perform, we always prefer to have as small probabilities as possible, to avoid detecting spurious dependencies merely by chance.

Weir's delta. Weir [Wei79] and Weir & Cockerham [WC89] introduce the composite measure of linkage disequilibrium, denoted as Δ, which is based on the number of haplotypes in contrast to the number of genotypes. Hamilton and Cole [HC04] further discuss how to standardize the measure, allowing comparison between populations. Using the notation of Table 2, the value for Δ for loci A and B is given by

$$\begin{aligned}\Delta_{AB} &= (2O(\mathrm{AA},\mathrm{BB}) + O(\mathrm{AA},\mathrm{Bb}) + O(\mathrm{Aa},\mathrm{BB}) + \frac{1}{2}O(\mathrm{Aa},\mathrm{Bb}))/N \\ &\quad -2\Big(N(\mathrm{AA}) + \frac{1}{2}N(\mathrm{Aa})\Big)\Big(N(\mathrm{BB}) + \frac{1}{2}N(\mathrm{Bb})\Big)\Big/N^2\,, \qquad (1)\end{aligned}$$

and as we see, the Δ measure indeed measures the number of the alleles A and B, and not just the number of observed genotypes containing A and B. The Δ measure for loci A and B is symmetric in that although formula (1) used the numbers of observations of alleles A and B, the output would be the same if the formula was written for alleles a and b instead. This symmetry implies that the expected value of Δ over all possible contingency tables with N individuals is 0. In practice, we wish to look at the absolute value of Δ.

The null hypothesis of no linkage disequilibrium corresponds to the case $\Delta_{AB} = 0$, and large absolute values of Δ_{AB} indicate a dependency between loci A and B.

We sometimes prefer to use a *scaled* version of Δ as it is not affected by different population sizes. The scaling is presented in Hamilton and Cole [HC04]: any set of genotype frequencies will fall into one of six possible cases. For each of these cases we can determine the extremum value of Δ. The scaled Δ is obtained simply by dividing the value in formula (1) by the corresponding extremal value.

2.5 Behaviour of Weir's Δ and the binomial test

Let us first compare some basic properties of the test statistics for genotypic LD, irrespective of the data at hand. In the analyses discussed in this section, we consider all possible 3×3 contingency tables for a given number of individuals.

Differences between Δ and P. We examined whether the P value of the binomial test and Weir's Δ can lead to substantially different judgements of the

linkage disequilibrium between two loci. To this end, we consider the distributions of these statistics over all 3×3 contingency tables for $N = 24$ individuals. Perhaps surprisingly, we observed that in many cases, Weir's Δ (either scaled or unscaled) finds no indication of linkage disequilibrium even though the binomial test does (Figure 1): the binomial P value is very small and thus significant, but the Δ is not large and thus not significant. An extreme example of such a contingency table is

	BB	Bb	bb
AA	0	x	0
Aa	6	0	6
aa	0	$12 - x$	0

where x is any integer between 0 and 12: Based on the marginal distributions of the table, the middle entry of the table should have several observations, assuming independence of the genotypes. The binomial test recognizes the apparent deviation from independence. In sharp contrast, Weir's Δ evaluates to 0, since the observations fit perfectly with independence at the *allelic* level. This casts some doubt on the validity of Δ as a measure of genotypic LD.

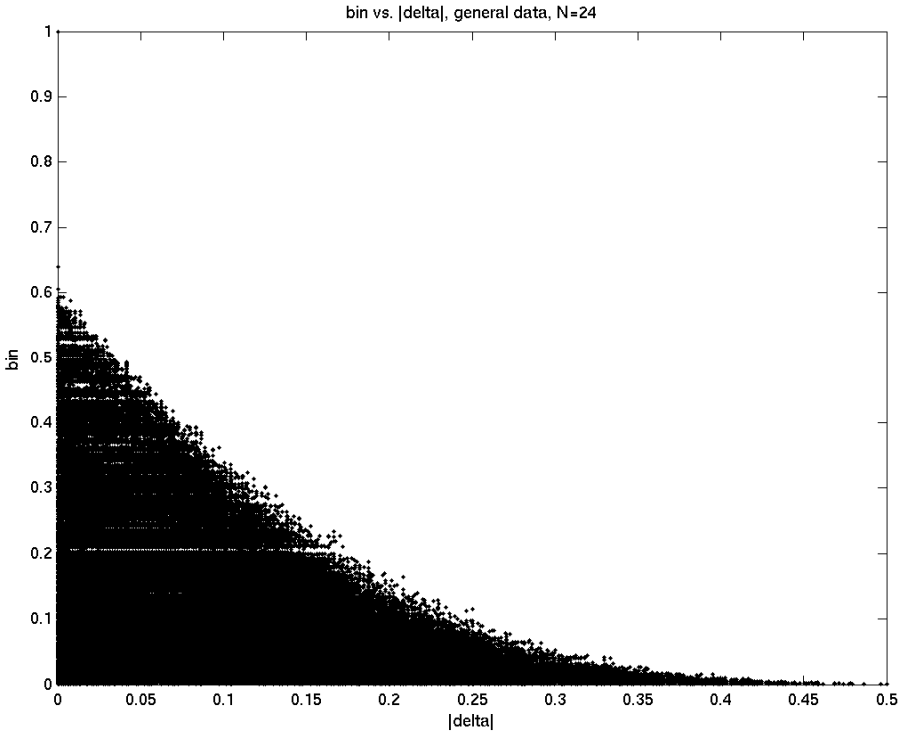


Figure 1: P value of the binomial test (vertical axis) versus Weir's Δ (horizontal axis). Each dot corresponds to one possible contingency table of 24 individuals. The tables in the bottom left corner are such that the binomial test recognizes a deviation from linkage equilibrium, but the Δ measure does not.

Special cases of the contingency table. In our experiments on HapMap data, we will later see that there are two special cases of the contingency table (Table 2) for which statistically significant values of the test statistics are obtained. These are the diagonal and upper-triangular configuration whose general patterns are shown in Tables 3 and 4.

	BB	Bb	bb
AA	N_1	0	0
Aa	0	N_2	0
aa	0	0	N_3

Table 3: Diagonal configuration of the contingency table.

	BB	Bb	bb
AA	N_1	N_2	0
Aa	N_3	0	0
aa	0	0	0

Table 4: Upper triangular configuration of the contingency table.

In the diagonal configuration of Table 3, the unscaled delta equals

$$(2N_1N_3 + \frac{1}{2}N_2(N_1 + N_3))/N^2.$$

The largest absolute value is $\frac{1}{2}$, which is achieved for $N_1 = N_3 = N/2$, $N_2 = 0$. Now, if N_2 is large and N_1 and N_3 are small, then Δ will be small, and the case is not interesting. This agrees well with the interpretation of the binomial test. When N_1 and N_3 are large and N_2 is small, Δ becomes large and the case is regarded interesting. Again, this is in accordance with the binomial test. But using the scaling proposed in [HC04], the scaled Δ is equal to 1 in this configuration, irrespective of the values of N_1, N_2 or N_3.

In the upper triangular configuration of Table 4, the unscaled Δ would give $\Delta = -N_2N_3/(2N^2)$ where $N = N_1+N_2+N_3$. If N_1 is large and N_2 and N_3 are small, Δ is close to zero and interpreted as insignificant. The conclusion with the binomial test is similar. The largest absolute value for unscaled Δ is $1/8$, obtained when $N_1 = 0$ and $N_2 = N_3$. Using the scaling proposed in [HC04], the scaled Δ is again equal to 1 in this configuration, irrespective of the values of N_1, N_2 or N_3. Thus the scaled Δ would always find a substantial linkage disequilibrium in the two configurations depicted in Tables 3 and 4.

The above two cases, diagonal and upper-triangular, only have two free parameters when the total $N = N_1 + N_2 + N_3$ is fixed, and it is thus easy to visualize their behaviour as a function of the two free parameters. This is done in figures 2 to 3 for $N = 60$. We see that the patterns are nonsymmetric, and that their extrema do not always coincide: the smallest binomial P values are obtained at configurations for which the (unscaled) Δ is not extremal.

While the diagonal configuration yields a very significant association, such a pattern may be more easily explained by nonbiologial than biological reasons.

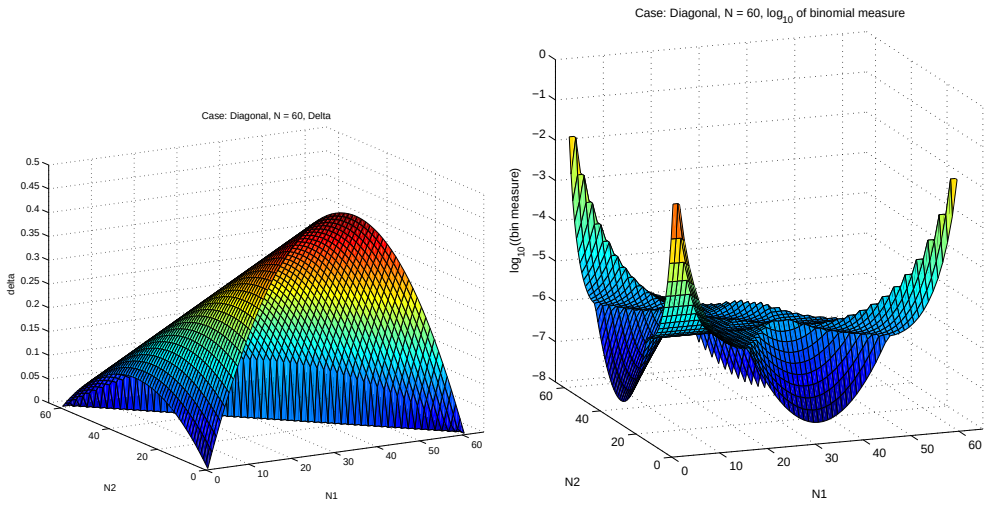


Figure 2: Weir's Δ (left) and log of binomial P value (right) at the diagonal configuration, $N = 60$

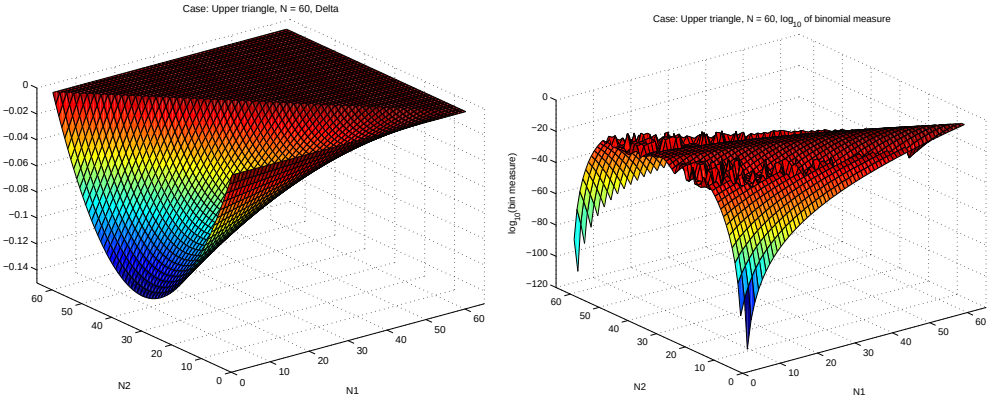


Figure 3: Weir's Δ (left) and log of binomial P value (right) at the upper triangular configuration, $N = 60$

Indeed, if the counts are zero outside the diagonal, the two markers are identical, as if one marker was read as two different markers. Thus, a marker database error could be a plausible explanation for an extreme diagonal pattern.

The upper triangular configuration, on the other hand, suggests that a genotype (or individual) is not viable if it carries two or more copies of the rarer alleles (a and b); or, in other words, the genotype is viable only if it is homozygous with respect to the more frequent allele (A and B) in one or both markers.

Distributions of Δ and P. Apart from the diagonal and upper-triangular configurations of the contingency table, it might be interesting to study the properties of the 3×3 contingency tables in general. The details are presented in Appendix B in which we show that a major proportion of all possible (theoret-

ical) contingency tables output a large Δ or a small binomial P value, indicating linkage disequilibrium. However, in real-world data, the cases of large Δ or small P are much more infrequent, as the genotypes in different chromosomes are most often independent of each other.

2.6 Clustering haplotypes

We also hypothesized that an individual SNP contains a limited amount of information, and we thus designed haplotype-based statistics that collect information from several adjacent SNPs. Consequently, our measures of LD would be based on comparing the haplotype distribution at two loci, each locus specified by a window of a few SNP markers.

In this setting, we proceed as follows. For each locus A consisting of d consecutive SNPs, we group the haplotypes on A into k internally homogeneous classes, called superalleles. An unordered pair of superalleles defines the corresponding supergenotype. Here d and k are parameters to be fixed.

Then for each pair of two loci, A and B, we compute the value of the test statistic $s(A, B)$ that measures the statistical dependency of a random individual's supergenotypes on the loci A and B. The test statistics are the same as in the single-SNP setting described previously in Section 2.4.

We model the haplotype population by a Bernoulli mixture [EH81]. The details of the clustering of the haplotypes via a Bernoulli mixture model are presented in Appendix A.

2.7 Computational issues

The setting. Our input data covered almost 630000 SNPs for HapMap and roughly 1.6 million SNPs for Perlegen data, which translate, respectively, into 200 billion and over one trillion (10^{12}) pairs of SNPs for which the LD measures had to be evaluated. The numbers may appear unpleasantly large at first sight, especially since we had to go through these billion or trillion pairs four times (for four separate populations in HapMap data; in Perlegen the number of populations was three). One should remember that a modern computer can in principle handle several billion additions or multiplications per second under optimum conditions. In practice, this kind of efficiency can only be achieved if the data used for calculations are available in the processor core, and usually this is not true. Instead, data have to be transferred from main memory to processor core through a hierarchical system of faster but smaller cache memories.

Not only delays in receiving data, but also delays in the instruction queue can hamper rapid computations. Complex conditional structures in the inner loops slow down any program. Another related issue is the programming language used. Compiled languages (C, Fortran 90) yield usually much faster programs than languages or scripts interpreted one line or command at a time (e.g., Matlab or R).

Therefore, in order to carry out any computations as fast as possible it is vital to organize both the data structures and the arithmetic operations so that

the processor cores (almost) always have the necessary variables and instructions at disposal.

Computing the association for single-SNP genotypes. Parallelization of LD calculations is very straightforward, for the problem is data parallel in a natural way. In a two-dimensional processor grid the SNP data for one chromosome is distributed along one dimension and the SNP data for another chromosome along the second dimension.

The evaluation of Weir's Δ is rather simple, even when scaled. The evaluation of the binomial measure is a more complicated matter: In principle, one would have to evaluate the binomial cumulative distribution function twice (for both tails) for each of the nine elements in the 3×3 table of observed genotype combinations. For over one trillion SNP pairs this would mean close to twenty trillion function evaluations. The number of these evaluations can be reduced by two techniques. Firstly, one can form a table of the function values for every possible combination of the number of observed cases versus the probability given by the marginal distributions. Admittedly, this is feasible only when the size of the population is rather small, because the size of the table will grow as the cube of the size of the population — in the HapMap data, we could not resort to this technique. Secondly, because the cumulative distribution function is monotonous, it is possible to calculate a set of critical numbers of observations. The actual function values are searched only when it is clear by comparison with these critical limits that a significant case has been found.

The computations were run on a somewhat outdated Sun Fire 25K platform with 1.2 GHz UltraSparc IV processors. With 32 processors it took between 2 to 5 days to complete one sweep where each SNP was paired with every other SNP four times (once for each population). The exact amount of time depended on the test measure used (binomial test was more time consuming than Weir's Δ) and the minority allele frequency threshold — we often discarded SNP's whose minority allele frequency was lower than, say, 5 per cent. With higher threshold a smaller number of SNPs was taken into consideration, and accordingly the computation was faster.

On equivalent markers and distributed computation. It may well happen that the genomic data are identical for several loci, i.e., the sequences of 0's and 1's corresponding to our biallelic markers agree through the whole population for two or more loci. The data are thus in a sense partitioned into equivalence classes. This offers a seemingly simple way to speed up the calculations: Instead of having to find out the value of an LD-measure for every SNP pair, it will suffice to compute the measure only once for each equivalence class and then use the result for all markers in that particular class. Of course, maintaining an index set over the markers belonging to the classes takes some computational effort as well.

It turns out, however, that this computational shortcut introduces a severe problem. The data are distributed between several processors. When advantage is taken of the equivalence classes, the set of loci assigned to a particular

processor is no longer contiguous; furthermore, for different ethnic populations, the partitioning into equivalence classes will be different. This means that for different populations, the data on a processor will not correspond to the same set of loci. Therefore, when trying to form between two populations the intersection of the sets of SNP pairs with a significant amount of LD, one will have to compare the result sets between every processor. This will take a considerable amount of message passing – the more processors, the more message passing – and thus slow down the overall computation. Thus, with a large number of processors it will probably be more efficient to ignore the advantage given by the marker equivalence classes.

3 Results

In Section 2.4 and Appendix B we studied and compared some basic properties of the test statistics for genotypic LD, independent of the data at hand.

We now report on our findings in the all-pairs analyses of HapMap and Perlegen data sets. We start by studying how the LD decays as a function of the physical distance between the loci in Section 3.1. We then demonstrate how strong LD is found, and where, in later sections.

3.1 Decay of genotypic linkage disequilibrium

We examined how the statistical dependency of single-SNP genotypes decays as a function of the physical distance between the SNPs. We observe that LD decays rapidly: the binomial P values increase and thus become less significant, and Weir's scaled Δ decreases. Both measures reach a more or less constant level at distances around 100 kb and larger: Figures 4 and 5 (Perlegen data), 6 and 7 (HapMap data).

A peculiar observation is that in Perlegen data, the genotypic LD tends to be relatively small in general, even at short distances like 1 kb. This is observed in particular in the binomial P values (Figure 4) and unscaled Δ (not shown). This is in contrast to what one would expect, and can perhaps be explained by one or more of the following facts: we consider the dependency of *unphased* genotypes; the haplotypes that build up the genotype are not chosen independently; the Perlegen data has quite small populations, and the number of loci having a very small MAF is large, thus as a result, LD cannot be observed due to noise.

The scaled Δ is not as badly affected by a small population size and small MAF values (Figure 5). Also, the phenomenon was not seen with any LD measure studied in the HapMap data set which is substantially larger.

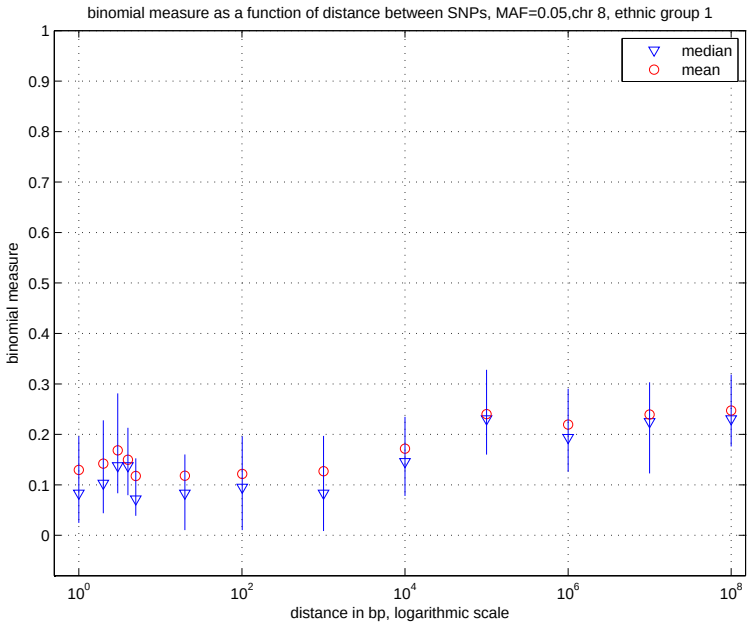


Figure 4: Genotypic LD as a function of the distance between SNPs. Vertical axis: The median, mean, and 75% quantiles for the P value of the binomial test. Horizontal axis: distance in base pairs, on a logarithmic scale. Perlegen data, chromosome 8, ethnic group 1, MAF = 0.05.

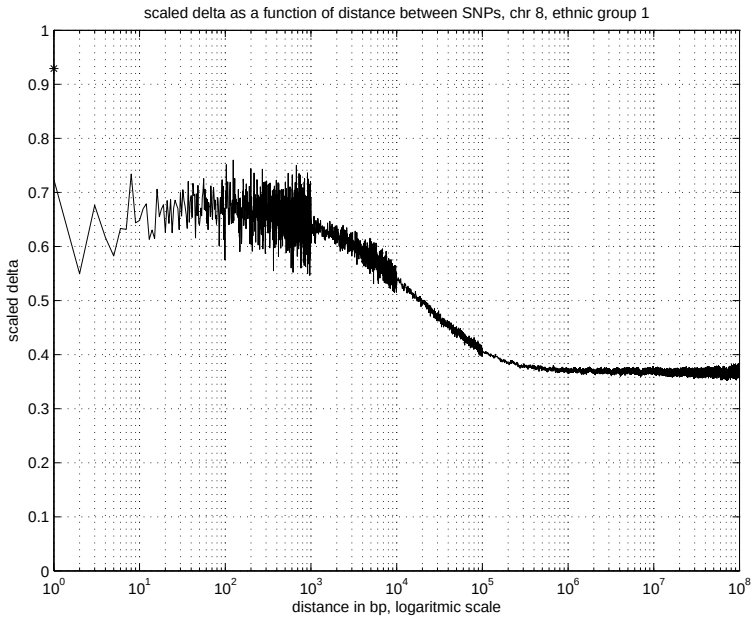


Figure 5: Genotypic LD as a function of the distance between SNPs. Vertical axis: Weir's scaled Δ. Horizontal axis: distance in base pairs, on a logarithmic scale. Perlegen data, chromosome 8, ethnic group 1, MAF = 0.00.

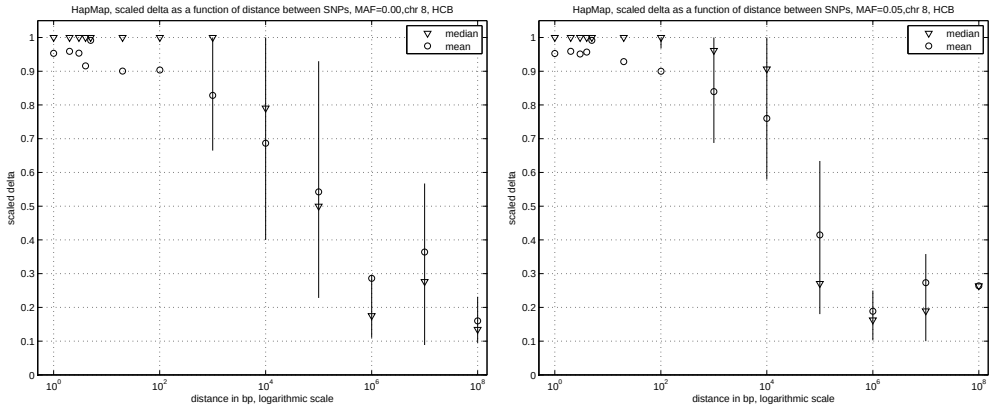


Figure 6: Histogram of scaled Δ. Horizontal axis: distance in base pairs, on a logarithmic scale. MAF, minimum allele frequency, 0.00 (left) or 0.05 (right). HapMap data, chromosome 8, ethnic group HCB.

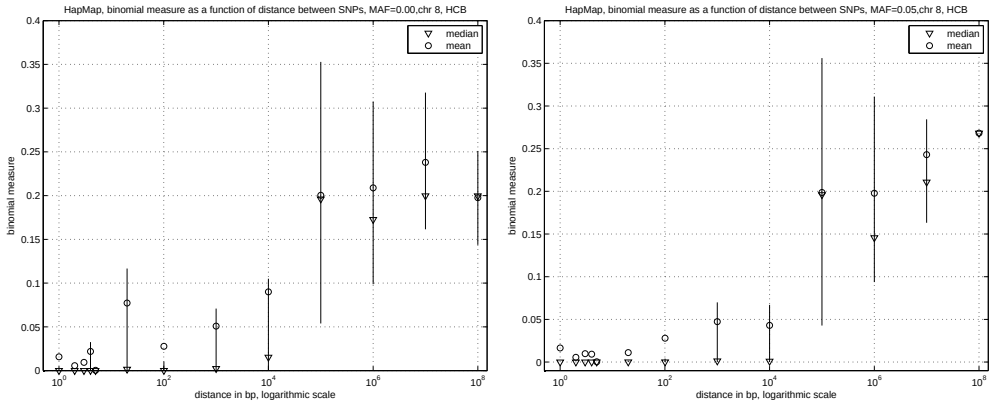


Figure 7: Histogram of binomial P values. Horizontal axis: distance in base pairs, on a logarithmic scale. MAF, minimum allele frequency, 0.00 (left) or 0.05 (right). HapMap data, chromosome 8, ethnic group HCB.

3.2 All-pairs analyses between chromosomes

We have run pairwise analyses between all loci in all chromosomes, measuring the values of the test statistics: Weir's Δ and binomial P values. In this section we will present various aspects of those analyses. We start with a discussion on how to choose the threshold of interestingness for the test statistics, and take one pair of chromosomes as an example. Later we will also report our findings on other pairs of chromosomes.

3.2.1 Choosing the threshold; chromosomes 6 vs 9 as an example.

We computed both Weir's Δ and the binomial measure for all pairs of individual SNPs and for all pairs of 4 consecutive SNPs over the Perlegen and HapMap data sets. We examined the number of locus pairs for different thresholds for the test statistics, counting only locus pairs for which the threshold was exceeded in all ethnic groups and thus a significant linkage disequilibrium was inferred. For Weir's Δ, thresholds of 0.1 and larger were considered, and for the binomial P value, thresholds of P = 0.04 and smaller.

As an example, we show results on chromosomes 6 versus 9. Figure 8 shows the results on Perlegen data when one member of the pair is from chromosome 6 and the other from chromosome 9. Similarly, Figure 9 shows the results on HapMap data.

From these figures we can learn something regarding the choice of the threshold: if the threshold is too strict (large Δ or small binomial P), then no signs of LD are found. Feasible values for the threshold seem to be around 0.15 and smaller for Δ. For the binomial test, in HapMap data (Figure 9), one might choose to use thresholds of P = 0.0001 or larger, as there are several pairs of loci found using such thresholds. In addition, in both data sets, there are significantly more pairs of loci for which thresholds of about P = 0.02 to P = 0.04 are exceeded, but such P values are perhaps not small enough to be interesting: in the course of millions of tests we easily encounter some spurious correlations, as discussed in Section 2.3.

The case $d = 4$ in the figures corresponds to using 4 consecutive alleles which were grouped using the procedure described in Section 2.6: the haplotypes were grouped via Bernoulli mixture modelling into "superalleles" which in turn were tested against linkage disequilibrium using our standard measures. We can see that the independence assumption is again violated in several cases. An exception is the case of HapMap data and Weir's unscaled Δ for which no signs of LD were found using the grouped superalleles.

We also studied choosing the threshold when the data set contained all loci from chromosomes 8 versus 9. In this case (not shown), suitable thresholds for Weir's Δ were again found to be around 0.15 or smaller for Δ. For the binomial P value, interesting values were only found if the threshold was set around P = 0.02 or larger, but it is questionable if this is significant enough to be interesting. It might thus be that there are no strong violations of linkage equilibrium between chromosomes 8 and 9.

When considering the value of the threshold, a natural question is whether

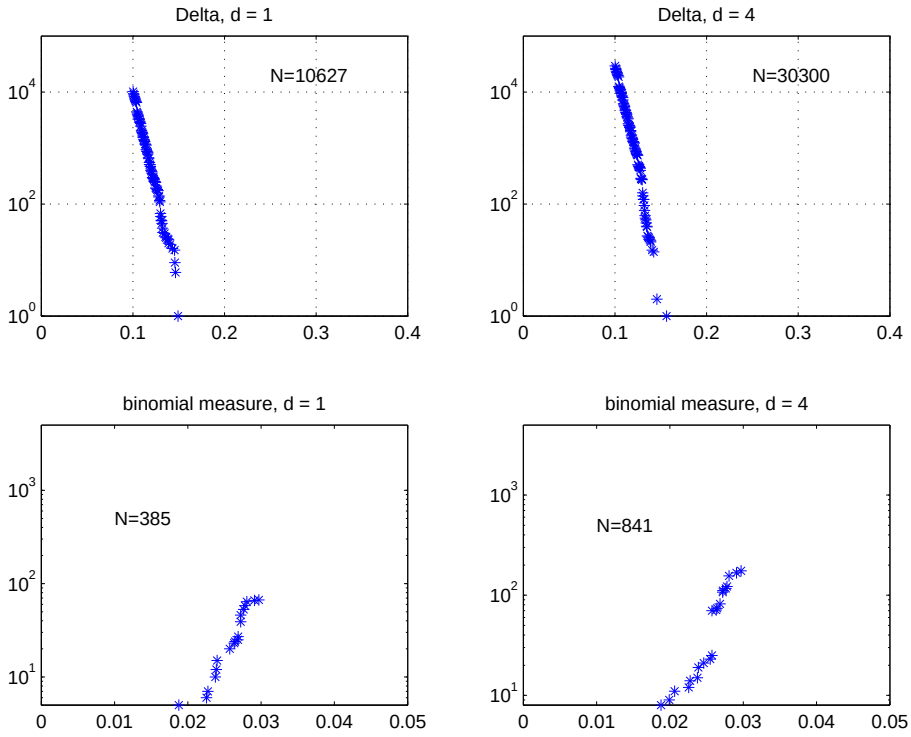


Figure 8: Cumulative distribution (in a logarithmic scale) of the number of pairs of loci for which Weir's unscaled Δ (top) or binomial P value (bottom) exceeds a threshold (horizontal axis) in every ethnic group. Left: loci of 1 SNP, right: loci of 4 SNPs. N = maximum number of pairs exceeding the threshold. In each locus pair, one locus is from chromosome 6 and the other from chromosome 9. Perlegen data.

the chosen threshold is strict enough. It is easy to say that the threshold is too strict if no LD is found, or that the threshold is too lazy if almost all pairs exhibit LD, but these borderlines are not enough. For the binomial test, the chosen threshold has a direct statistical interpretation, but for Weir's Δ this is not the case.

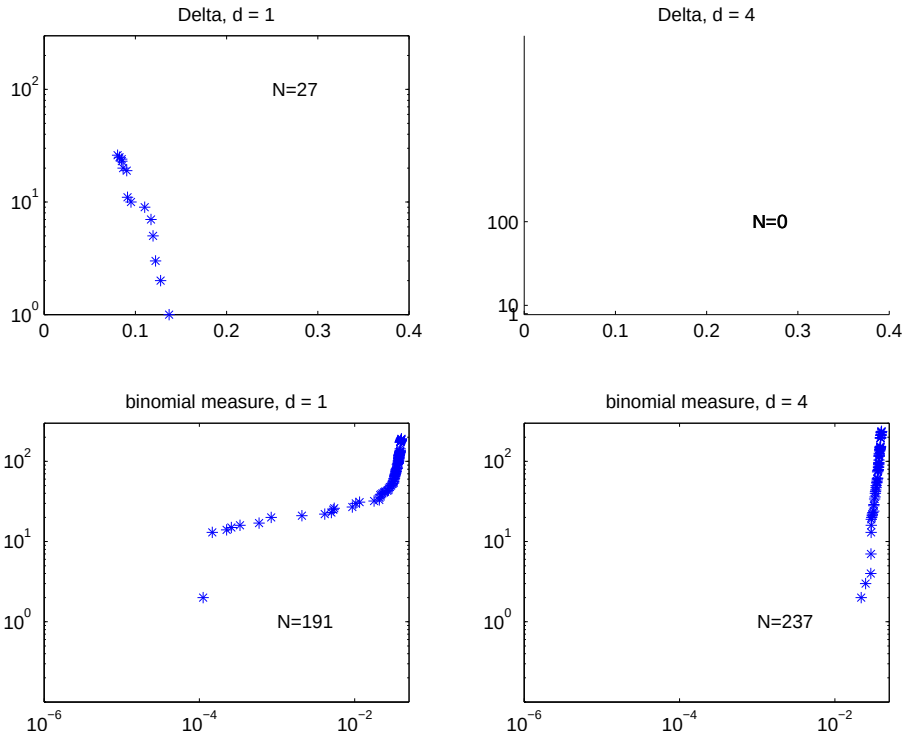


Figure 9: Cumulative distribution (in a logarithmic scale) of the number of pairs of loci for which Weir's unscaled Δ (top) or binomial P value (bottom) exceeds a threshold for every ethnic group. Left: loci of 1 SNP, right: loci of 4 SNPs. N = maximum number of pairs exceeding the threshold. In each locus pair, one locus is from chromosome 6 and the other from chromosome 9. HapMap data.

3.2.2 Findings of high LD

Above we found that there are a few marker pairs in chromosomes 6 versus 9 whose linkage disequilibrium is significantly larger than that of two markers on average. For these pairs, the LD is almost as high as for marker pairs located close to each other in one chromosome.

We ran the analyses through all pairs of chromosomes in HapMap data, and found several cases, where marker pairs picked from certain regions within the chromosomes show unusually high degree of linkage disequilibrium. Both Weir's Δ (unscaled) and the binomial measure typically detect the same regions. The marker pairs with most significant values of LD in these regions are shown in Tables5 and 6. In the tables, we have omitted some marker pairs which are almost the same as a pair listed in the table, to avoid unnecessary repetitions of almost-identical findings.

The marker pairs listed in Tables 5 and 6 obey a "diagonal" configuration such as the one depicted in Table 3. It would be interesting to examine, to what extent these associations can be explained merely due to erroneous marker location information rather than truly biological phenomena; a more detailed study is, however, beyond the scope of the present work.

In Table 5 we have used Weir's unscaled Δ instead of the scaled one. The scaling would require storing nontrivial amounts of measurements in the memory during the computations. Also, setting the threshold proved to be infeasible when the scaled Δ was used (not shown): we got a very large number of significant marker pairs, 0.02 to 0.04 per cent of all marker pairs. Often these marker pairs obey the upper triangular configuration depicted in Table 4. But after discarding the markers whose minimum allele frequency is smaller than 0.05 per cent, the number of significant marker pairs found by scaled Δ is much smaller, only $0.2 \cdot 10^{-7}$ to $2 \cdot 10^{-7}$ times the number of marker pairs. These correspond to pairs in which N_2 and N_3, the numbers of heterozygous alleles Aa and Bb in Table 4, are large enough to reach the minimum allele frequency threshold.

In total, in each pair of two chromosomes C_i and C_j, the number of significant marker pairs found using Weir's unscaled Δ is typically 0 to 3 when the markers are from different chromosomes: remember that there are 22 times 21 such chromosome pairs. A few pairs of chromosomes C_i and C_j are exceptional such that the number of significant marker pairs is 20 to 30.

Table 6 shows the marker pairs detected the binomial measure. Each pair of chromosomes contained a few cases for which the binomial P value was between 0.03 and 0.04. In addition, a couple of extreme cases having a P value from 10^{-5} to 10^{-4} were found in some pairs of chromosomes. These P values are similar to those obtained when we take two nearby markers from the same chromosome (in which case the LD is expected to be high). Thus we are able to find some extreme, statistically significant values of the linkage disequilibrium between different chromosomes.

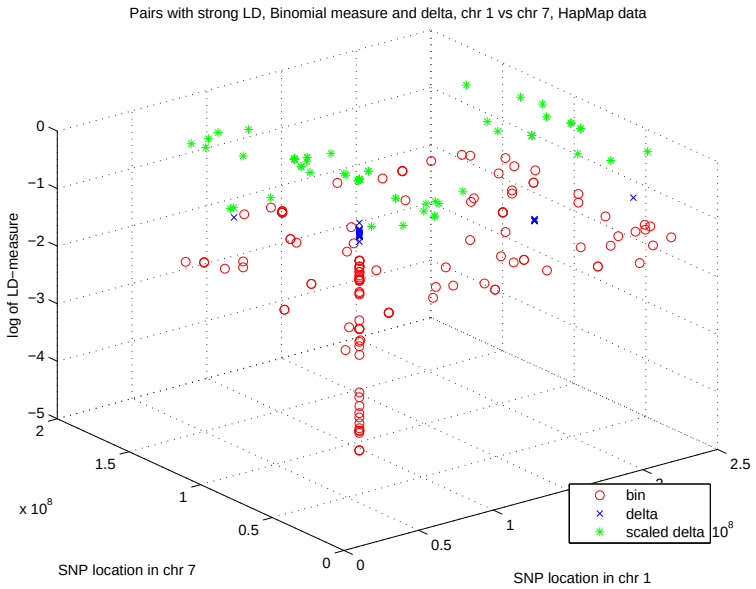
chr	chr			bp	bp	Δ
1	4	48097	24231	240260635	119901820	1.77e-01
1	7	17553	12383	83139601	75977629	1.75e-01
1	13	37391	14169	194523430	74793426	1.34e-01
1	13	1578	18184	8388847	91097772	1.74e-01
1	13	2194	6118	11518948	42261409	1.46e-01
3	4	16851	12593	87016490	57177152	1.69e-01
4	6	17781	8030	85843666	31492413	1.14e-01
4	18	12953	2204	59007364	12085950	1.54e-01
4	18	12953	2205	59007364	12086249	1.54e-01
4	18	12953	2207	59007364	12105509	1.54e-01
4	18	12953	2210	59007364	12121165	1.69e-01
4	18	12953	2212	59007364	12145824	1.34e-01
4	18	12953	2213	59007364	12155155	1.69e-01
4	18	12953	2214	59007364	12168099	1.72e-01
4	18	12953	2216	59007364	12185773	1.69e-01
4	18	12953	2218	59007364	12201483	1.69e-01
4	18	12953	2221	59007364	12207812	1.69e-01
4	18	12953	2222	59007364	12207885	1.69e-01
5	14	10101	7499	51443614	53369029	-1.31e-01
5	14	10100	7502	51436683	53402907	1.22e-01
5	14	10100	7504	51436683	53415605	1.22e-01
5	15	17075	12962	84825180	85670741	1.91e-01
5	15	15805	7883	79078552	59545732	9.78e-02
5	16	8358	5368	39190719	27972673	-8.64e-02
5	17	13446	11925	66318489	67584325	1.66e-01
5	17	13446	11926	66318489	67584377	1.66e-01
6	9	7724	10934	30054632	27480967	1.28e-01
6	9	7740	10934	30118137	27480967	1.22e-01
6	9	7904	14907	31002386	63344625	8.67e-02
6	9	8184	14928	32326215	66475178	9.15e-02
6	9	8185	14928	32328899	66475178	9.15e-02
6	9	8191	14928	32349834	66475178	9.15e-02
6	9	8195	14928	32369264	66475178	9.15e-02
6	9	8197	14928	32372860	66475178	9.15e-02
6	9	8198	14928	32377831	66475178	9.15e-02
6	9	8199	14928	32378952	66475178	9.15e-02
6	9	8205	14928	32407835	66475178	9.15e-02
6	9	8402	13942	33314560	36032008	1.17e-01
6	9	8405	13942	33330653	36032008	-1.43e-01
11	15	11672	9261	58659275	65879592	1.38e-01
11	16	6859	5795	31555909	46763904	1.41e-01
11	16	6864	5795	31596651	46763904	-1.41e-01
11	16	6874	5795	31650758	46763904	-1.42e-01
11	16	6876	5795	31656162	46763904	1.41e-01
11	16	6880	5795	31691152	46763904	-1.41e-01
11	17	2735	14477	12652878	78997868	-1.27e-01
14	15	15415	8713	89759182	63263311	1.82e-01

Table 5: Marker pairs having a significant linkage disequilibrium measured using Weir's Δ. Columns: chromosomes, marker ids, marker locations in the chromosomes measured as base pairs (bp), Δ. HapMap data.

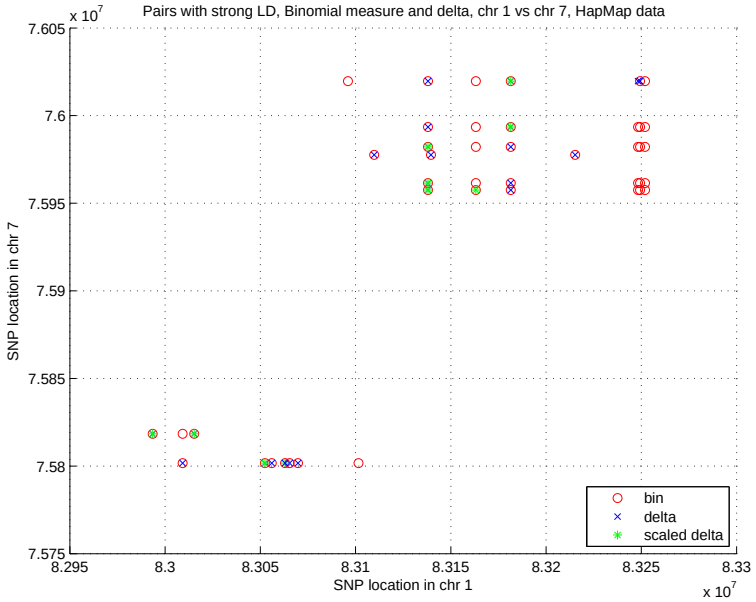
chr	chr	marker	marker	bp	bp	P value
1	4	48097	24231	240260635	119901820	5.37e-05
1	7	17552	12381	83138012	75957562	1.93e-05
1	7	17552	12382	83138012	75961477	1.93e-05
1	7	17552	12384	83138012	75982113	1.93e-05
1	13	37391	14169	194523430	74793426	8.83e-05
1	13	1578	18184	8388847	91097772	5.30e-04
1	13	2194	6118	11518948	42261409	9.26e-04
3	4	16851	12593	87016490	57177152	5.14e-05
4	6	17781	8030	85843666	31492413	9.83e-05
4	18	12953	2204	59007364	12085950	3.78e-05
4	18	12953	2205	59007364	12086249	3.78e-05
4	18	12953	2207	59007364	12105509	3.78e-05
4	18	12953	2210	59007364	12121165	3.78e-05
4	18	12953	2212	59007364	12145824	3.78e-05
4	18	12953	2213	59007364	12155155	3.78e-05
4	18	12953	2214	59007364	12168099	3.78e-05
4	18	12953	2216	59007364	12185773	3.78e-05
4	18	12953	2218	59007364	12201483	3.78e-05
4	18	12953	2221	59007364	12207812	3.78e-05
4	18	12953	2222	59007364	12207885	3.78e-05
5	14	10101	7499	51443614	53369029	1.05e-04
5	15	17075	12962	84825180	85670741	6.42e-05
5	15	15805	7883	79078552	59545732	7.93e-05
5	16	8358	5368	39190719	27972673	7.34e-05
5	16	36920	7061	178018332	53543451	4.81e-04
5	17	13446	11925	66318489	67584325	7.34e-05
5	17	13446	11926	66318489	67584377	7.34e-05
5	17	13446	11929	66318489	67621922	7.34e-05
6	9	7724	10934	30054632	27480967	1.12e-04
6	9	7740	10934	30118137	27480967	1.12e-04
6	9	7904	14907	31002386	63344625	5.00e-03
6	9	8184	14928	32326215	66475178	1.46e-04
6	9	8185	14928	32328899	66475178	1.46e-04
6	9	8191	14928	32349834	66475178	1.46e-04
6	9	8195	14928	32369264	66475178	1.46e-04
6	9	8197	14928	32372860	66475178	1.46e-04
6	9	8198	14928	32377831	66475178	1.46e-04
6	9	8199	14928	32378952	66475178	1.46e-04
6	9	8205	14928	32407835	66475178	1.46e-04
6	9	8402	13942	33314560	36032008	5.85e-04
6	9	8405	13942	33330653	36032008	8.41e-04
11	15	11672	9261	58659275	65879592	7.12e-05
11	16	6859	5795	31555909	46763904	7.93e-05
11	16	6864	5795	31596651	46763904	7.93e-05
11	16	6874	5795	31650758	46763904	7.93e-05
11	16	6876	5795	31656162	46763904	7.93e-05
11	16	6880	5795	31691152	46763904	7.93e-05
11	17	2735	14477	12652878	78997868	6.62e-05
14	15	18159	12116	102561095	81519400	7.53e-05
14	15	18159	12118	102561095	81524629	7.53e-05
14	15	18159	12119	102561095	81527871	7.53e-05
14	15	18159	12120	102561095	81538040	7.53e-05
14	15	18159	12121	102561095	81539432	7.53e-05
14	15	18159	12122	102561095	81542012	7.53e-05
14	15	18159	12124	102561095	81553349	7.53e-05
14	15	18159	12125	102561095	81562882	7.53e-05
14	15	18159	12126	102561095	81565918	7.53e-05
14	15	18159	12127	102561095	81577713	7.53e-05
14	15	18159	12128	102561095	81583855	7.53e-05
14	15	15415	8713	89759182	63263311	1.42e-04

Table 6: Marker pairs having a significant linkage disequilibrium measured using the binomial P value. Columns: chromosomes, marker ids, marker locations in the chromosomes measured as base pairs (bp), binomial tail probability. HapMap data.

Locations of interesting marker pairs along the chromosome. Above we found several marker pairs in exceptionally strong linkage disequilibrium. Let us now visualize their locations in more detail, taking chromosomes 1 and 7 as an example. Chromosomes 1 and 7 contain 49 221 and 28 413 markers, respectively, yielding a total number of about 1.4 billion marker pairs. A marker pair was interpreted as having a significant LD if the binomial measure had a P value smaller than 0.04; this was the case for 150 marker pairs among those 1.4 billion marker pairs. The P value 0.04 is still quite high, and in fact most of those 150 pairs have a P value between 0.01 and 0.04 which is perhaps insignificant. However, there are a few pairs having a P value 10^{-5}. Alternatively, a marker pair was interpreted as having a significant LD if the absolute value of Weir's Δ was 0.08 or larger. In Figure 10 we see the binomial P values and Weir's unscaled Δ for the marker pairs whose LD was deemed significant. Figure 10 (a) is 3-dimensional, having the locations in chromosomes 1 and 7 as the horizontal axes, and logarithmic values of the binomial measure and Weir's Δ in the vertical axis. Small values of the binomial measure are interesting, and large values of Δ; we see that there is a small area in which several extremal values are concentrated, as the binomial measure is very small and Δ is large. Scaled Δ never yields interesting values. The 3-dimensional plot is somewhat difficult for the human eye, and thus we also show a 2-dimensional plot in Figure 10 (b) which is a zoomed-in detail of the interesting area. In (b) the actual numerical values of the LD measures are not visible, but only the locations in chromosome 1 and chromosome 7. There clearly are two regions in which strong LD is observed.



(a)



(b)

Figure 10: LD measures of significant marker pairs, in a logarithmic scale. (a) 3-dimensional plot showing the locations in chromosomes 1 and 7, and the values of the LD measures. There is a small area in which extremal LD is reached. (b) A 2-dimensional projection, zoomed into the interesting area.

3.2.3 Permutation tests

When running through the all-pairs analyses of loci from two chromosomes we will perform millions of statistical tests. It might then be that we encounter some spurious correlations merely by chance, even though there are no real dependencies in the data. To tackle this problem, we rely on permutation tests.

We permuted the ordering of individuals within each subpopulation separately, and then tested for LD using the binomial and Δ measures. Again, we required that the chosen threshold value for LD was exceeded in all subpopulations separately. This makes it quite hard for random associations to pop up: even if we by chance found a significant LD in one marker pair in one subpopulation, the same marker pair should be significant in all subpopulations in order to be reported as interesting.

In interchromosome marker pairs of chromosomes 14 and 15 in HapMap there are a few pairs whose binomial measure is $7.53 \cdot 10^{-5}$ in the original data. We permuted the data 1000 times, and tested if any dependency was found in a significant portion of the runs – if yes, then our original finding would be doubtful. However, in 1000 permutations we did not find any pairs having a binomial measure smaller than 0.01 (for all the three subppoulations simultenously). This gives evidence that the dependencies found in the original data are truly interesting.

3.3 All-pairs analyses within a chromosome

Until now, we have presented *interchromosomal* results in which one member of the pair of loci was from one chromosome, and the other member was from another chromosome. It is interesting to see how significant the linkage disequilibrium is in an *intrachromosomal* case where both members are from the same chromosome.

Along the lines of Section 3.2.1, we counted the number of locus pairs for which our test statistics exceeded certain thresholds in all ethnic groups. Both members of the pair are from chromosome 6.

Figure 11 shows the results at HapMap data. (For Perlegen data, the observations are somewhat similar but as the data set is smaller, we cannot have as small P values as with HapMap data.) When comparing Figures 9 and 11 we see that a larger number of extreme values of the LD measures is obtained when both members of the pair of loci are from the same chromosome. In the figures, this is denoted by N. Also, the extreme values are "more extreme": the maximum values of Weir's Δ are larger (roughly 0.27 vs. 0.13), and the binomial P values are smaller (roughly 10^{-6} vs. 10^{-4}). This is to be expected, as nearby loci often demonstrate LD, and the distance between loci at two distinct chromosomes is always "infinite".

The case of loci consisting of $d = 4$ SNPs, the right-hand side plots of Figure 11, show the behaviour of the Bernoulli mixture model. We can see that the mixture model can indeed find interesting locations, as soon as there exist some.

When running through all chromosomes, the number of significant intrachromosomal marker pairs is from several tens of thousands to a few hundred

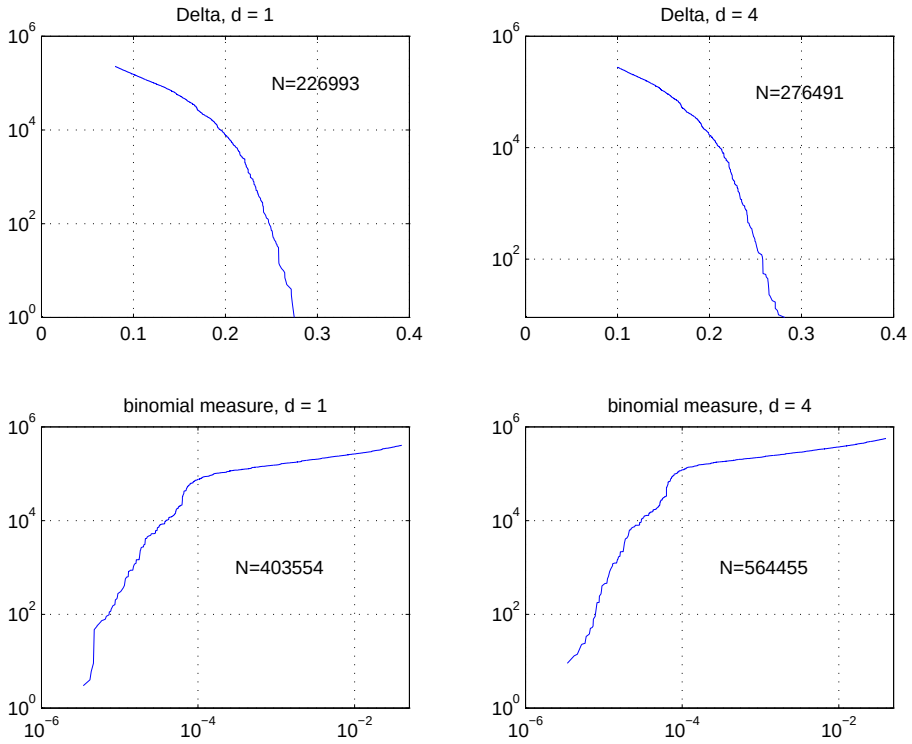


Figure 11: Cumulative distribution (in a logarithmic scale) of the number of pairs of loci for which the Weir's unscaled Δ (top) or binomial P value (bottom) exceeds a threshold for every ethnic group. Left: loci of 1 SNP, right: loci of 4 SNPs. N = maximum number of pairs exceeding the threshold. All loci are from chromosome 6. HapMap data.

thousands, depending on the size of the chromosome. This is of course significantly more than in the interchromosomal case, which was discussed in Section 3.2.2.

We then contrast the findings of high interchromosome LD, shown in Section 3.2.2, to the values of intrachromosome LD. Consider the LD of two markers which both belong to chromosome 1. The LD of nearby markers is typically high, so we wish to contrast our interchromosome (1 vs 7) findings to intrachromosome (1 vs 1) values. Of all marker pairs in chromosome 1, 0.017 per cent have a P value less than 0.04; this is a significantly larger number than when one marker was from chromosome 1 and another from chromosome 7. Then again, consider the peak value of the interchromosome (1 vs 7) LD – among the intrachromosome (1 vs 1) marker pairs, less than $1/10^6$ are more extreme than the interchromosome peak. Figure 12 shows the sorted values of the binomial measure in intrachromosome (1 vs 1) and interchromosome (1 vs 7) pairs; we indeed see that the extreme interchromosome P values are small also when compared to intrachromosome values. This is indicative of a nontrivial LD between distant loci.

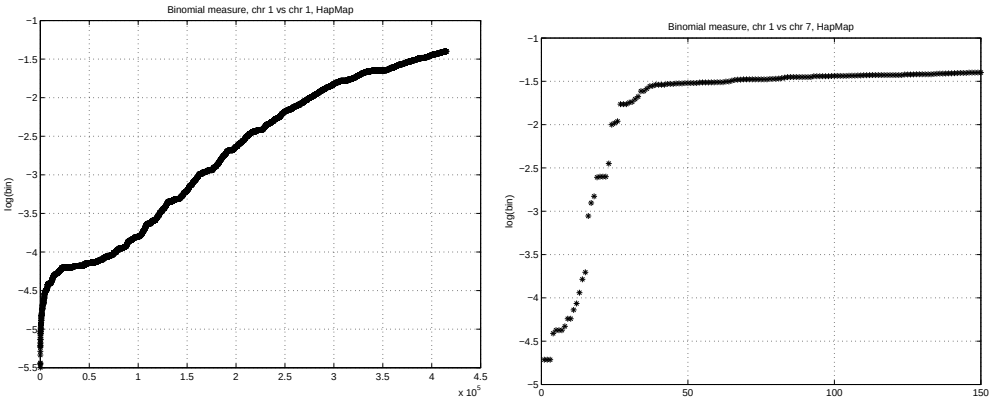


Figure 12: Sorted values of the binomial measure, in a logarithmic scale. Left: markers within chromosome 1. Right: one marker from chromosome 1 and another from chromosome 7.

3.4 Clustering consecutive markers by the Bernoulli mixture model

In the above sections we have reported the LD between two markers located in two different chromosomes and found that there are a few interesting marker pairs. We have also briefly touched upon the case of grouping consecutive markers into "supermarkers". Let us now return to this for a small while.

We took $d = 4$ consecutive markers and clustered them using Bernoulli mixture modeling as described in Section 2.6 into supermarkers. We then tested for the dependence between those. This was repeated for all $d = 4$ length windows in all chromosomes.

Both the binomial measure and Weir's unscaled Δ found two areas containing interesting supergenotypes in HapMap data. The first of these areas is spanned by markers 17531–17545 in chromosome 1 and markers 12366–12369 in chromosome 7. The second area is close to the first one, and spanned by markers 17552–17557 in chromosome 1 and markers 12380–12388 in chromosome 7. These interesting areas were also found at our studies on individual markers reported in Section 3.2, and thus there do not seem to be areas that are interesting *only* as supergenotypes as opposed to original genotypes.

4 Discussion

We have presented methods for analysing the linkage disequilibrium (LD) between distant genetic loci. LD refers to the statistical dependency of the DNA content at two locations. An often documented fact is that LD of nearby loci is high, and it decays monotonously with the distance of the studied loci. This decay can be attributed to the recombination process that is a major source of genetic variation: the more distant the two loci are, the higher the probability that genetic material is switched between the members of a chromosome pair.

We were interested in finding distant loci where the content of the DNA

is correlated not merely due to the structure of the population or the study sample. Finding a significant LD between two far-away loci is indicative of a possible functional dependence: some phenotype characteristic to the population in question is affected by two loci that interact in a non-trivial way, e.g., a person may carry a certain genotype at one locus only if he carries a certain other genotype at the other locus. It is, however, also possible that a significant LD between two far-away loci arises just due to a misplaced marker, that is, for some reason a marker is erroneously annotated with a wrong chromosome number.

To reveal linkage disequilibrium between two distant loci, we introduced a new test statistic, termed the binomial measure, that aims at finding two strongly correlated genotypes, one from each locus. We showed that this new measure can be more powerful than, for instance, Weir's Δ [Wei79, WC89], when the alleles at the two loci are relatively independent. We found some problems in the usage of Δ: Firstly, one can demonstrate by a simple theoretical analysis that it does not always distinguish dependence between two loci. Secondly, we do not have a baseline to which to compare the level of Δ and thus cannot assess when Δ is large enough to be interesting. In contrast, the binomial measure behaves in a way that is easily tractable in statistical terms.

We also described various computational techniques and tricks that enable a genome-wide analysis of all pairs of SNPs or small windows of SNPs. The techniques include parallelisation, precomputation of certain threshold values, "lazy evaluation", and means for efficient memory management.

We reported results concerning the LD between distant loci in the HapMap and Perlegen data sets. The HapMap data consist of four populations and the Perlegen of three populations. The observed distributions of the test statistics suggest that some of the strongest correlations may indicate a non-random, functional association. We chose a threshold for the P values of the binomial test, and required that the threshold was met in each subpopulation separately. This adds to the statistical significance of the observations.

However, it is somewhat nontrivial to estimate the statistical significance of these associations due to a huge number of dependent multiple tests performed and due to the population structure underlying the data.

Permutation testing is one reasonable approach to estimating the extent to which large correlations may occur just by chance. The idea is to draw a large sample, say 100 or 1000, random permutations of the genotypes in one chromosome, within each of the four (HapMap) or three (Perlegen) ethnic groups, while keeping the order of the genotypes fixed in the other chromosome. Then for any fixed threshold of a test statistics, one can compute the proportion of the permuted data sets for which the count of locus pairs exceeding the threshold is larger than the corresponding count for the original data. We performed this in each subpopulation separately, making it quite hard for spurious correlations to appear: even if one marker pair in one subpopulation was by change found significant, the same marker pair should be significant in all subpopulations, in order to be interpreted as interesting. Indeed, in the original data, there were a few marker pairs whose binomial P value was of order 10^{-5}; in 1000 permuted data sets we did not find any P values smaller than 0.01. This gives strong

evidence to the significance of our results.

In addition, we briefly considered the use of "supergenotypes" consisting of a few consecutive markers, constructed using a mixture-model based clustering method. Our original hypothesis was that an individual SNP contains a limited amount of information, and that it would be fruitful to collect information from several adjacent SNPs. Our measures of LD would then be based on comparing the haplotype distribution at two loci, each locus specified not by one marker but a window of a few markers. Somewhat in contrast to our original hypothesis, the findings of high LD were mostly at pairs of single-SNP loci. A few interesting multiple-SNP loci were also discovered, but those loci were also found via the single-SNP studies.

In the Perlegen data, the findings of LD between distant loci were not as strong as in the Hapmap data. For this reason, we also considered the possibility that truly dependent distant regions of the genome are hundreds of kilobases long, spanning tens to hundreds of SNPs. To this end, we considered a summary statistic that aggregates the statistics based on pairs of short windows of SNPs. We found that often in Perlegen data there is a consecutive region within one chromosome, containing markers that are highly dependent with a small number of markers in another chromosome.

Let us also note that if the data were completely random in an uniform manner, then we should have found a much larger number of pairs in high linkage disequilibrium. In contrast, the genomic data is highly structured such that nearby loci are in linkage disequilibrium – as is well known – and far-away loci are not, except for the few interesting cases that we have indicated.

Our methodology can be further extended and tested in a few directions.

First, we described an approach to measure LD between larger genomic regions by aggregating the tests for pairs of loci within the regions. How much one can gain with this approach compared to single locus pairs remains to be studied.

Second, we have used the binomial P values for judging if a contingency table is in linkage equilibrium or not, in other words, if there is a dependence between two variables in the table. We could also use Fisher's exact test whose purpose is to reveal the dependence as well. The margins of the table are fixed. Under the null hypothesis of independence, this leads to the hypergeometric distribution. We compute a P value for the observed table of counts, using our favourite method (binomial, Chi squared, likelihood ratio or other). We also list all tables having the same row and column marginals as our observed table, and compute their P values. We then sum the P values of all those tables whose P values are at most as large as the P value of the observed table. If the sum of these P values is smaller than a predefined limit, say 0.05, we conclude that the observed table violates the independence assumption.

Third, we considered the dependency of unphased genotypes. In a recent paper, Wu et al [WJX08] present a way of measuring the LD when linkage phase information of marker loci for unrelated individuals is unknown. They classify the interaction into *intragametic* interaction of two alleles from different loci on the same haplotype, and *intergametic* interaction of two alleles from different loci on different haplotypes. These interactions will lead into intragametic and

intergametic LD.

Fourth, we do not known whether there exists any true, non-random associations in the real data we analyzed. Thus, it would interesting to simulate data with planted dependencies of varying strength between some distant loci, and to see how large sample size is needed for reliable discovery of the truly dependent loci, at some tolerable false positive rate.

References

[EH81] Brian S. Everitt and David J. Hand. *Finite mixture distributions.* Chapman & Hall, London, 1981.

[HC04] D. C. Hamilton and D. E. C. Cole. Standardizing a composite measure of linkage disequilibrium. *Annals of Human Genetics*, 68(3):234–239, May 2004. doi:10.1046/j.1529-8817.2004.00056.x.

[HSN+05] David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox. Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079, February 2005. DOI: 10.1126/science.1105436.

[MFY+08] Kazuharu Misawa, Shoogo Fujii, Toshimasa Yamazaki, Atsushi Takahashi, Junichi Takasaki, Masao Yanagisawa, Yozo Ohnishi, Yusuke Nakamura, and Naoyuki Kamatani. New correction algorithms for multiple comparisons in case-control multilocus association studies based on haplotypes and diplotype configurations. *Journal of Human Genetics*, 53:789âĂŞ801, 2008.

[MK09] Kazuharu Misawa and Naoyuki Kamatani. Parahaplo: A program package for haplotype-based whole-genome association study using parallel computing. *Source Code for Biology and Medicine*, 4:7, 2009.

[The03] The International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, December 2003.

[WC89] B. S. Weir and C. C. Cockerham. Complete characterization of disequilibrium at two loci. In M. W. Feldman, editor, *Mathematical Evolutionary Theory*, pages 86–110. Princeton University Press, 1989.

[Wei79] B. S. Weir. Inferences about linkage disequilibrium. *Biometrics*, 35:235–254, March 1979.

[WJX08] Xuesen Wu, L. Jin, and Momiao Xiong. Composite measure of linkage disequilibrium fro testing interactino between unlinked loci. *European Journal of Human Genetics*, 16:644–651, 2008.

A Clustering of haplotypes via Bernoulli mixture modeling

To test our hypothesis that an individual SNP contains a limited amount of information, we designed haplotype-based statistics that collect information from several adjacent SNPs. In this setting, our measures of LD are based on comparing the haplotype distribution at two loci, each locus specified by a window of a few SNP markers.

For each locus A consisting of d consecutive SNPs, we group the haplotypes on A into k internally homogeneous classes, called superalleles. An unordered pair of superalleles defines the corresponding supergenotype. Then for each pair of two loci, A and B, we compute the value of the test statistic $s(A, B)$ that measures the LD between supergenotypes on the loci A and B.

We model the haplotype population by a Bernoulli mixture [EH81].

Bernoulli mixture model. Consider a locus A consisting of d consecutive SNPs. Let x_{2i-1} and x_{2i} denote the ith individual's paternal and maternal haplotypes over the d SNPs, for $i = 1, \ldots, n$. At every SNP, let the alleles take values from $\{0, 1\}$. For brevity, let x_h denote a haplotype, be it maternal or paternal, and $h = 1, \ldots, 2n$.

We clustered the $2n$ haplotypes into k groups via mixture modelling. We modelled the haplotype population by a Bernoulli mixture [EH81] with k components, that is, we assigned haplotype x_h a probability

$$p(x_h; \alpha, \beta) = \sum_{c=1}^{k} \alpha_c \prod_{j=1}^{d} \beta_{cj}^{x_{hj}} (1 - \beta_{cj})^{1 - x_{hj}} ,$$

where α consists of the mixture proportions $\alpha_1, \ldots, \alpha_k$, that sum up to unity, and where β consists of the probabilities β_{cj} that a haplotype assigned to the cth component has allele 1 at the jth SNP. A standard expectation–maximization algorithm was employed for setting the parameters α and β so as to (locally) maximize the likelihood $\prod_h p(x_h; \alpha, \beta)$ for the observed haplotypes.

Independence assumption in a multivariate mixture model. In a multivariate mixture model such as the one presented above, the underlying assumption is that the attributes (here, SNPs) are independent of each other (given the component in the mixture). This simplifying assumption allows us to write the joint probability of the attributes x_{hj} in a product form $\prod_j \beta_{cj}^{x_{hj}} (1 - \beta_{cj})^{1 - x_{hj}}$. However, in haplotype data, it is well known that neighboring SNPs depend on each other in a more complicated fashion. Thus, the model can viewed as a computational convient approximation that is expected to give a good fit if the haplotype clusters are tight: each β_{cj} is close to 0 or 1, in which case the haplotype distribution within a cluster can be (mostly) explained by genotyping errors.

Estimating the model and assigning supergenotypes. In the EM algorithm, the mixture proportions α_c are iteratively updated by

$$\alpha_c = \frac{1}{n} \sum_{h=1}^{2n} r_{ch} \,,$$

where r_{ch} is the posterior probability of component c having created haplotype x_h, and its update rule is in turn

$$r_{ch} = \frac{\alpha_c \prod_j \beta_{cj}^{x_{hj}} (1-\beta_{cj})^{1-x_{hj}}}{\sum_c \alpha_c \prod_j \beta_{cj}^{x_{hj}} (1-\beta_{cj})^{1-x_{hj}}} \,.$$

The update equation for the parameter β_{cj} is

$$\beta_{cj} = \frac{\sum_{h=1}^{2n} r_{ch} I(x_{hj}=1)}{\sum_{h=1}^{2n} r_{ch}}$$

where $I(x_{hj}=1)$ is an indicator function, taking value 1 when $x_{hj}=1$ and 0 otherwise.

Finally, after the EM algorithm has converged, we assign each haplotype x_h the cluster c for which the posterior probability r_{ch} is the largest over $c = 1,\ldots,k$. Each haplotype x_h is thus assigned into one of k groups, which we call a *superallele*, reflecting the fact that it takes values in $\{1,\ldots,k\}$, instead of $\{0,1\}$ as in the case of the original alleles. Thus the haplotypes x_{2i-1} and x_{2i} determine a *supergenotype* s_i as an unordered pair of maternal and paternal superalleles, denoted as $\{u,v\}$, with $1 \le u \le v \le k$. Note that a supergenotype is spanned by d SNPs.

B Distributions of Δ and binomial P in a general case

Apart from the diagonal and upper-triangular configurations of the contingency table, it is interesting to study the properties of the 3×3 contingency tables in general. Here we show that a major proportion of all possible (theoretical) contingency tables output a large Δ or a small binomial P value, indicating linkage disequilibrium. However, in real-world data, the cases of large Δ or small P are much more infrequent, as the genotypes in different chromosomes are most often independent of each other.

The expected value for Δ over all possible contingency tables with N individuals is 0.083 for $N = 24$, 0.078 for $N = 45$ and 0.077 for $N = 60$. The straightforward definition of Δ renders an analytical calculation of the standard deviation possible, and the result is

$$\sigma = \sqrt{\frac{(11N+2)(N-1)(N+9)}{1320N^3}}.$$

The range of Δ is $[-0.5, 0.5]$.

A major proportion of all possible (theoretical) contingency tables thus output a large Δ, but in real-world data, the cases of large Δ are much more

infrequent. In Figure 13 (left panel) we present the cumulative distribution function of the absolute value of Δ, computed over all contingency tables with a population of $N = 60$. In the HapMap experiments presented in Section 3.2 we concluded that "interesting" values of Δ typically are 0.1 or larger. A computational analysis shows that the fraction of tables with $|\Delta| > 0.1$ is 0.295. So, with a genuinely random uniform distribution of contigency tables, we should expect that approximately 30 percent of tables show strong LD in the sense that $|\Delta|$ exceeds the threshold value of 0.1. In real-world data, the situation is quite different.

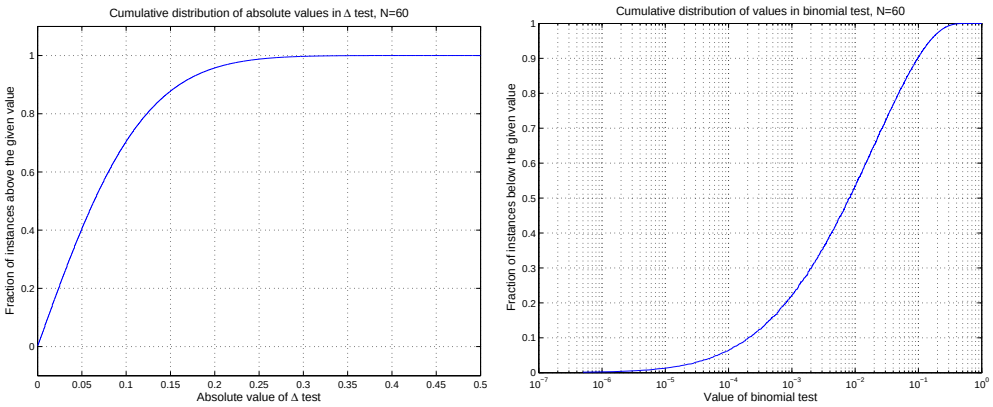


Figure 13: Cumulative distribution of $|\Delta|$ (left) and binomial P (right) computed over all 3×3 contingency tables in a population of $N = 60$ individuals

Similarly, a major proportion of all possible contingency tables output a small (and thus "significant") P value of the binomial test, although real-world data is typically closer to the null hypothesis of no linkage disequilibrium. We have computed numerically the expected value and standard deviation for the binomial test over all possible contigency table configurations with certain fixed number of people. For $N = 24$, the expected value of P is $\mu = 0.098$ and the standard deviation is $\sigma = 0.087$; for $N = 45$ the results are $\mu = 0.048$ and $\sigma = 0.068$, and for $N = 60$, $\mu = 0.032$ and $\sigma = 0.056$. Because of its probabilistic nature, the range of the binomial test is $\mathrm{P} \in (0, 1]$. Note that we have deliberately chosen not to consider the cases where the binomial distribution returns zero probability.

Figure 13 (right panel) shows the cumulative distribution of P, computed over all contingency tables with $N = 60$.

Naturally, the exact values of the cumulative distribution function depend on the size of the population, but with $N = 45$ the results are almost identical to $N = 60$. If we regard our computational task as a Bernoulli process consisting of a series of four independent tests, the share of SNP pairs passing the process should be of the order $0.3^4 = 0.0081$. As demonstrated in Section 3.2, this is not at all the case. Of course, the four populations are not genuinely independent. Also, the principles on which the sampling of SNPs is based affects the results. A more careful look at the distribution of the number of three possible genotypes (AA, Aa and aa) shows that the SNP data really is not uniformly distributed.

Typically, only the number of the more common homozygote AA is even nearly uniformly distributed, *i.e.*, in a population of size N, the number of persons with this genotype gets any of the values $0, 1, ..., N$ with an almost equal probability $1/(N+1)$. The minor allele homozygote aa, on the other hand, is much more likely to appear only in a few persons, or not at all, than in a large number of people. The heterozygote case Aa is most likely to appear in a moderate number of people, while the cases with a very small or a very large number of heterozygotes are rare.