Robustness of Inference of Haplotype Block Structure

Running title: Robustness of Haplotype Block Inference

Russell Schwartz, Celera Genomics, Informatics Research, Rockville, MD 20850 USA

(current address: Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 USA)

Bjarni V. Halldórsson, Celera Genomics, Informatics Research, Rockville, MD 20850 USA

Vineet Bafna, Celera Genomics, Informatics Research, Rockville, MD 20850 USA

(current address: The Center for the Advancement of Genomics, Rockville, MD 20850 USA)

Andrew G. Clark, Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853 USA

Sorin Istrail*, Celera Genomics, Informatics Research, Rockville, MD 20850 USA

*corresponding author:

Sorin Istrail

45 West Gude Drive

Rockville, MD 20850 USA

phone: (240) 453-3668

fax: (240) 453-3324

email: Sorin.Istrail@celera.com

Keywords: haplotype block, SNP, recombination, linkage disequilibrium, diversity

Abstract

In this report, we examine the validity of the haplotype block concept by comparing block decompositions derived from public data sets by variants of several leading methods of block detection. We first develop a statistical method for assessing the concordance of two block decompositions. We then assess the robustness of inferred haplotype blocks to the specific detection method chosen, to arbitrary choices made in the block-detection algorithms, and to the sample analyzed. Although the block decompositions show levels of concordance that are very unlikely by chance, the absolute magnitude of the concordance may be low enough to limit the utility of the inference. For purposes of SNP selection, it seems likely that methods that do not arbitrarily impose block boundaries among correlated SNPs might perform better than block-based methods.

1 Introduction

Single Nucleotides Polymorphisms (SNPs) in the genome show great promise as predictors of disease, but the large number of known SNPs presents a significant challenge in locating those meaningfully correlated with disease phenotypes. The detection of a haplotype block structure to the human genome (Daly et al. 2001; Jeffreys et al. 2001; Johnson et al. 2001; Patil et al. 2001) — in which the genome is large made up of regions of low diversity, each of which can be characterized by a small number of SNPs — presents a possible way to reduce the complexity of the problem. However, whether construction of haplotype blocks is the most efficient way to reduce the number of SNPs that one needs to type to represent the extant genetic variation remains a controversial topic.

Many methods have been suggested for defining block structures, which can be roughly classified into three groups. One group consists of linkage disequilibrium (LD) methods, such as that of Gabriel et al. (2002), which define blocks so as to enforce generally high pairwise LD within blocks and generally low pairwise LD between blocks. Another group consists of diversity-based methods, such as that of Patil et al. (2001), which define blocks so as to enforce low sequence diversity, by some diversity measure, within each block. Finally, there are methods that look for direct evidence of recombination, such as the four-gamete test applied by Hudson and Kaplan (1985), defining blocks as apparently recombination-free regions. In addition to distinct block definitions, blocking methods can differ according to the optimization criterion by which the best block decomposition is chosen among all possible decompositions satisfying the block test. Two important optimization criteria, which were examined by Zhang et al. (2002) in a paper describing a general algorithm for efficient block decomposition, are minimization of the number of blocks consistent with the block definition and minimization of the number of SNPs needed to characterize all sequences within each block.

Here we assess the merits of haplotype block inference by examining robustness of the block concept to multiple variants in block detection strategies. If haplotype blocks are genuinely capturing islands of low diversity separated by recombination sites, as they are generally represented by proponents, then the various methods proposed in the literature for locating blocks ought to derive essentially the same decompositions each time they are applied. We assess this prediction by developing a statistic for comparing block decompositions and applying it to decompositions derived on publicly available phase-known datasets using variants of the commonly used block detection methods.

In the remainder of this paper, we describe our empirical analysis of the robustness of the block concept. We first present our methodology, describing computational methods used for calculating block boundaries and presenting a statistical method for comparing two block decompositions. We then describe results of an analysis assessing the robustness of block methods to variants in the block detection protocol. In particular, we examine robustness to changes in block definition, optimization criterion, and population sample, as well as robustness to arbitrary choices made in the algorithms. We conclude that different block decompositions of a single genetic region tend to be far more consistent than can be explained by chance, but that the absolute similarity is nonetheless frequently small. It therefore appears that while there is a common underlying structure to haplotype blocks which all methods detect, that structure is less rigidly defined than individual block decompositions might suggest.

2 Methods and Data

2.1 Finding Block Boundaries

In this paper, we compare variants of the three general block-detection methodologies that have so far appeared in the literature: recombination-based, diversity-based, and linkage-disequilibrium (LD) based. We use the four-gamete test (Hudson and Kaplan 1985) as the simplest example of a recombination-based test. One problem with the four-gamete test is that it is based on the infinite sites model (in which a given site mutates at most once during a sequence's evolutionary history) and can falsely infer recombination events when

that model fails due to recurrent mutations. That problem is not a significant drawback for block construction, however, since sites of recurrent mutation can be reasonably considered to disrupt block patterns even if they do not correspond to recombination sites. For a diversity-based test, we use a generalization of the Patil et al. (2001) test. In their test, a region is a block if at least 80% of the sequences occur in more than one chromosome. This test was developed for a sample of only 20 chromosomes and does not scale well to larger sample sizes as it will tend to yield larger blocks as more chromosomes are studied. We generalized this test by defining a region as a potential block if sequences within that region (haplotypes) accounting for at least 80% of the sampled population each occur in at least 10% of the sample. Finally, we developed a test based on the D' statistic, similar to that used by Gabriel et al. (2002) but with the test of significance tuned to produce more meaningful results on small population samples. In this approach, we consider a set of SNPs to form a block if they are contiguous and the D' value of every pair of SNPs within the block shows significant LD with a P-value of < 0.001, as estimated by simulations using the empirically measured single-site allele frequencies for a given SNP pair and the assumption of complete linkage equilibrium.

We further consider the two common optimization criteria discussed in the introduction: minimizing the total number of blocks and minimizing the number of SNPs needed to characterize the block decomposition of every sequence assuming all sequences contain only observed haplotypes within each block. We solve optimally for each measure using a variant of the dynamic programming algorithm of Zhang et al. (2002) modified to sample uniformly at random from all optimal solutions rather than deterministically choosing a single optimum. All of the methods were implemented in C++ running on DEC Alpha Unix

machines.

2.2 A Statistic for Block Comparison

We use the number of shared block boundaries as a statistic for the similarity of two block partitions. This yields a computationally tractable method for exactly computing the p-value for rejecting the null hypothesis that two block decompositions were chosen at random (uniformly) and independently from one another. If B_1 and B_2 are the number of boundaries in the two partitions, m the number of boundaries shared by the partitions, and S the total number of SNPs, then the probability they share exactly m boundaries under the null hypothesis is

$$\frac{\binom{B_1}{m}\binom{S-1-B_1}{B_2-m}}{\binom{S-1}{B_2}}$$

yielding a p-value (probability of the statistic being at least m) of

$$\sum_{i=m}^{\min(B_1,B_2)} \frac{\binom{B_1}{i}\binom{S-1-B_1}{B_2-i}}{\binom{S-1}{B_2}}.$$

This measure thus allows us to test the hypothesis that two partitions are related and provide a degree of confidence with which we can reject the null hypothesis of independence.

2.3 Data

For evaluation, we rely on two publicly available datasets. The first is the Perlegen chromosome 21 dataset (Patil et al. 2001), which consists of 24,047 SNPs typed on 20 phased chromosomes. This dataset contains a large contiguous set of SNPs providing an excellent

test of blocking algorithms. Although a substantial portion of chromosome 21 was not covered by multi-SNP blocks by any method, it yielded enough long blocks to clearly detect statistically significant concordances between distinct block decompositions. We also use a dataset derived from 71 individuals typed at 88 polymorphic sites in the human lipoprotein lipase (LPL) gene (Nickerson et al. 2000), from which we ignored one multi-alleic site to simplify our analysis. The fewer SNPs in the LPL dataset makes it more manageable for illustrative purposes. In addition, its greater depth of coverage allows us to draw more confident predictions and provides enough individuals to compare results from distinct subsets of the population sample.

3 Results

We first asked whether the concept of blocks is robust to the various block measures. We conducted comparisons by running the available algorithms on the two datasets and comparing the outcomes using the shared boundary statistic described above. Table 1 describes the results when run on the chromosome 21 data of Patil et al. (2001). We note that it is also possible to compare an algorithm to itself because of the fact that the block definitions generally yield many equally good solutions of which the algorithms must choose just one. By independently sampling among the optima, we can compare distinct runs of a single algorithm, revealing what is robust to that single definition as well as providing a baseline for how much solutions derived by distinct methods can possibly coincide with one another. Although the percent similarities between distinct methods or even within any one method are not high in an absolute sense, they are much greater than can be explained by chance.

The four-gamete test appears much closer to the diversity- and LD-based tests than either of those is to the other. Furthermore, minimum SNP solutions appear to have more in common with one another than minimum block solutions.

[Approximate location of Table 1.]

For purposes of illustration, we provide visual comparisons of block assignments for the LPL dataset of Nickerson et al.(2000). Due to the much smaller number of SNPs in the LPL dataset compared to the chromosome 21 dataset, pairwise comparisons do not generally yield statistically significant results, although they show comparable percent agreement to the pairwise comparisons. Figure 1 illustrates the correspondence between the different block measures by showing side-by-side comparisons of block boundaries for each. The results appear to show generally poor agreement between block boundaries derived from distinct measures, with somewhat better agreement between distinct runs of a single measure.

[Approximate location of Figure 1]

In order to assess the role of optimization criteria on concordance, we conducted further pairwise comparisons between the two criteria — minimum blocks or minimum SNPs — for each of the three block definitions. Table 2 shows comparisons of the two criteria for each of the three block definitions examined. We see substantially higher correspondence between distinct optimization criteria for a single block definition than we saw between distinct block criteria for a single optimization method. This result suggests that minimizing the number of blocks is a reasonable approximation to minimizing the number of SNPs. Figure 2 illustrates this greater concordance on the LPL data, particularly for the four-gamete method. Furthermore, the figure shows that when block decompositions do disagree, it is often because of a slight slippage in some boundaries rather than a sizable overall change

in block structure.

[Approximate location of Table 2.]

[Approximate location of Figure 2.]

It is also important to assess our ability to detect haplotype blocks for realistic sizes of datasets. We assessed this with the LPL data by dividing the dataset into two subsets, randomly placing individuals into one or the other, then comparing block decompositions drawn from the two samples by a single method. We did not attempt this analysis on the Patil et al. (2001) chromosome 21 dataset because the number of chromosomes would have been too small for the half-size datasets to yield reasonable results. Table 3 describes the results. There was substantially worse agreement between runs on separate half-size samples (each with 71 chromosomes) than we saw between distinct runs on the same full-size sample (with 142 chromosomes). This is discouraging news, because it shows that even a sample of 35 individuals may be too small to reliably capture haplotype block structures. The four-gamete test shows the most robustness to sample and the diversity-based test the least. We attribute the lack of statistical significance of most results primarily to the small number of SNPs in the dataset.

[Approximate location of Table 3.]

Figure 3 illustrates the similarities of the block decompositions. Different block decomposition algorithms show noticeably weaker correspondence than when the same algorithm is run twice on the same full-size dataset. Some slippage of boundaries no longer seems an adequate explanation for the discrepancies. While the block structures have some correspondence, individual boundaries in one no longer generally have clear corresponding boundaries nearby in the other.

[Approximate location of Figure 3.]

4 Discussion

Our results on comparisons between distinct methods support the notion that there is a tendency for the variability in the human genome to be organized into blocks of adjacent sites that share ancestral history. These blocks can be identified crudely by a variety of ad hoc algorithms. Our results show unequivocally that the different block-finding algorithms identify similar structure to an extent that cannot be explained by chance. They also show, however, that the absolute correspondence between block-assignments can differ markedly in response to changes in both block definition and optimization criterion.

Our results appear consistent with two contradictory explanations: that haplotype blocks are a valid scientific reality but require greater sample sizes or better algorithms to detect reliably, or that the population level processes of mutation, drift, and recombination that give rise to haplotype blocks do so in a way that makes it very difficult to settle on a single unified definition that best captures that history. We argue in favor of the latter hypothesis on the grounds that if the blocks were simply and universally well-defined, then current methods ought to be robust to a variety of measures of diversity, linkage disequilibrium, and recombination probability. The fact that these measures do not consistently provide the same block decomposition suggests that the imperfections lie in the blocks themselves and not the methods.

These result does not contradict the evidence for recombination hotspots in the genome (Jeffreys et al. 2001), but do suggest that hotspots do not in general lead to a well defined

block structure; this may be either because existing computational methods cannot detect them reliably or because they do not account for a sufficiently large fraction of the total recombination in the genome. Nor do our results contradict the notion that there are genuine regions of low diversity, high linkage disequilibrium, or low historic recombination in the human genome. Rather, they argue that such regions are not as sharply defined as the concept of discrete blocks would imply. Nonetheless, the fact that blocks, however defined, do allow one to significantly compress the information in SNP datasets suggests that they can be quite useful for the primary task of representing SNP datasets by an informative subset of the SNPs. The inconsistency of block decompositions, however, coupled with the correlation of SNPs among adjacent blocks, suggests that other methods might perform better than haplotype blocks. In particular, we suggest considering methods for reducing haplotype redundancy, such as the Alignment Algorithm of Schwartz et al. (2000), that detect conserved haplotype regions without requiring a global block structure.

5 Acknowledgements

We would like to thank Francis Kalush, Francisco de la Vega, Ross Lippert, Michele Cargill, Kit Lau and Mark Adams for many valuable discussions and technical suggestions about SNPs and haplotypes data analysis challenges, and for sharing the Applera Resequencing Project data with us.

References

Daly, M., Rioux, J., Schaffner, S., and Hudson, T. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29,229–232.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., and Altschuler, D. 2002. The structure of haplotype blocks in the human genome. *Science* 296,2225–2229.

Hudson, R., Kaplan, N. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111,147–164.

Jeffreys, A., Kauppi, L., Neumann, R. 2001. Intensely punctute meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29,217–222.

Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S.C., Clayton, D.G., and Todd, J.A. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29,233–237.

Nickerson, D.A., Taylor, S.L., Fullerton, S.M., Weiss, K.M., Clark, A.G., Stengaard, J.H., Salomaa, V., Boerwinkle, E., Sing, C.F. 2000. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res.* 10,1532–1545.

Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., Nguyen, B.T., Norris, M.C.,

Sheehan, J.B., Shen, N., Stern, D., Stokowski, R.P., Thomas, D.J., Trulson, M.O., Vyas, K.R., Frazer, K.A., Fodor, S.P., and Cox, D.R. 2001. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* 294,1719–1722.

Schwartz, R., Clark, A.G., and Istrail, S. Methods for inferring block-wise ancestral history from haploid sequences, 44–59. In Guigó, R., and Gusfield, D., eds., *Algorithms in Bioinformatics. Lecture Notes in Computer Science 2452*, Springer, Berlin.

Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. 2002. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* 99,7335–7339.

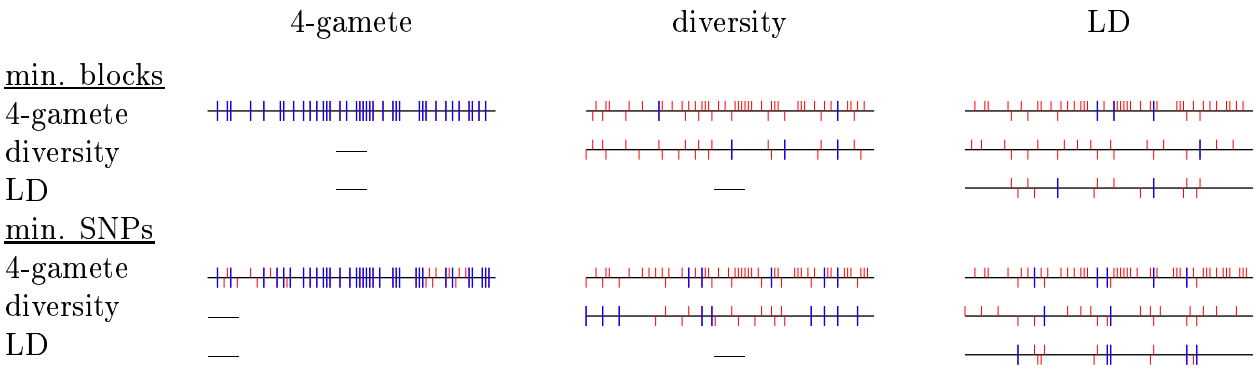

Figure 1: Pairwise comparison of block boundaries for the different block definitions for the LPL dataset of Nickerson et al. Each image shows the block boundaries as vertical lines, with boundaries above the horizontal line coming from the first method and those below the horizontal line coming from the second method. Those boundaries appearing above and below the horizontal line are shared by both methods and are drawn thicker in order to highlight them.

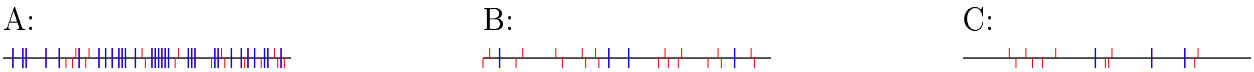

Figure 2: Pairwise comparison of block boundaries for the different optimization methods for the LPL dataset of Nickerson et al. The comparisons shown are A: 4-gamete B: diversity-based C: LD-based

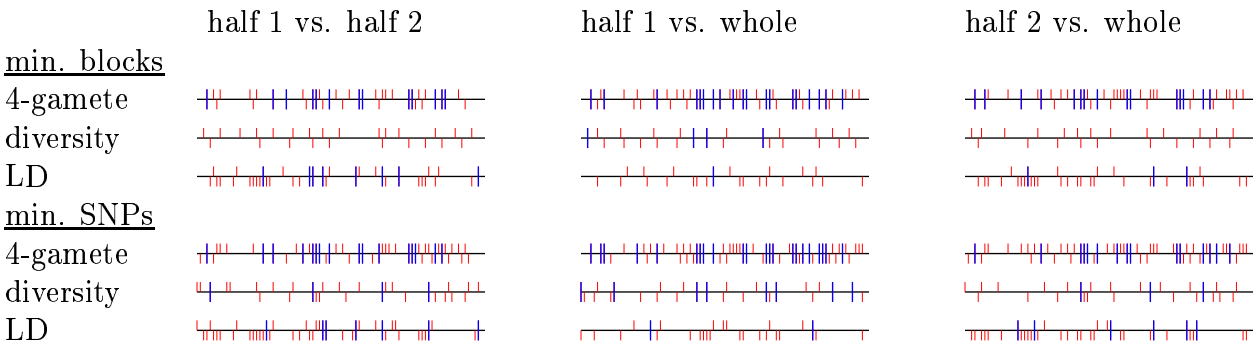

Figure 3: Comparisons of block assignments using distinct population samples for a single method. Each row shows, for a single block definition and optimization criterion, a comparison of boundaries derived from one part of the partitioned dataset to boundaries derived from the other part followed by comparisons of each of the partial-dataset solutions to the full-dataset solution.

	4-gamete	diversity	LD
min. blocks			
4-gamete	$45.3\%/7.25 \times 10^{-1206}$	$15.7\%/3.02 \times 10^{-162}$	$12.3\%/1.70 \times 10^{-18}$
diversity	-/-	$33.8\%/2.97 \times 10^{-626}$	$7.19\%/3.89 \times 10^{-5}$
LD	-/-	-/-	$54.3\%/3.09 \times 10^{-1761}$
min. SNPs			
4-gamete	$65.3\%/2.16 \times 10^{-2276}$	$22.6\%/5.30 \times 10^{-366}$	$16.0\%/1.03 \times 10^{-54}$
diversity	-/-	$51.5\%/6.71 \times 10^{-1222}$	$9.70\%/6.14 \times 10^{-23}$
LD	-/-	-/-	$70.5\%/7.67 \times 10^{-2921}$

Table 1: Comparisons of block definitions on the chromosome 21 dataset of Patil et al. minimizing blocks. Each element of the matrix gives the percentage of block boundaries assigned by either method that are shared by both, followed by the p-value of the overlap. Elements comparing a method to itself show values for two distinct runs of the same algorithm each of which is choosing an optimal solution uniformly at random.

Tests	% Identity	P-value
4-gamete	41.4%	4.51×10^{-1038}
diversity-based	29.9%	6.75×10^{-526}
LD-based	49.8%	7.19×10^{-1513}

Table 2: Comparisons of minimal block to minimal SNP optimization criteria for each block definition on the chromosome 21 dataset of Patil et al. Each element of the table gives the percentage of block boundaries assigned by either method that are shared by both, followed by the p-value of the overlap.

Tests	minimizing blocks		minimizing SNPs	
	% Identity	P-value	% Identity	P-value
4-gamete	35.1%	0.00287	34.0%	0.024
diversity-based	0.00%	1.00	13.8%	0.372
LD-based	22.9%	0.072	16.3%	0.569

Table 3: Comparisons of runs of a single block assignment algorithm on two halves of a balanced randomly selected partition of the LPL dataset of Nickerson et al.