A Threshold of Disequilibrium Parameters to Identify Haplotype Blocks on Biallelic Models

Makoto TOMITA[*†], Ryo TAKEMURA[†] and Naoyuki KAMATANI[‡†]

Abstract

A domain in which recombination does not often occur, but in which linkage disequilibrium is present, is known as a "haplotype block" or "LD block". There are many methods for identifying haplotype blocks using disequilibrium parameters, particularly, the method of Kamatani *et al*. . Although their method has a high calculation time requirement, it is advantageous among current methods when we analyzeing large amounts of genotype data. We thought that the cause of the extended calculation time lies in the threshold of D' in the initial condition to identify haplotype blocks. Therefore, here, we report on a method for identifying a more appropriate a haplotype block, which greatly shortens the calculation time by setting the optimal threshold of D'.

1 Introduction

Linkage disequilibrium analysis is one method in statistical genetics. In this method, "linkage" refers to the state of the relationship among loci accumulating over 2-3 generations in a biallelic model, two alleles inherited from one parent show a strong tendency to stay together, as do those from the other parent and "linkage disequilibrium" describes the state in which the alleles in the last generation have a distribution reflecting ancestral recombination events
Methods of trait-mapping based on theories of linkage disequilibrium analysis have developed quickly in recent years. For DNA sequences, there are domain "hotspot" where recombination events occurred often. On the other hand, there are domains where recombination does not often occur, and yet maintained linkage disequilibrium is present, a phenomenon known as "haplotype block", or "LD block". Determining the value of D', a disequilibrium parameter, has been an important step in identifying the haplotype block, up until now. However, D' is supported better supported experientially than it is theoretically. In this report, we consider the relationship between the identification of haplotype block and D'.

1.1 Recent studies

Gabriel *et al*. (2002) defined "strong linkage disequilibrium" as the state where the upper bound of 95% confidence interval of D' exceeds 0.95. Zhu *et al*. (2003) evaluated haplotypes in which each relative frequency is more than 0.04.

Furthermore, the ideas of Gabriel's and Zhu's methods were combined to identify haplotype blocks in another paper (Kamatani *et al*. 2004). Their detailed procedures are as follows.

[*]Department of Mathematical Sciences and Information Engineering, Faculty of Mathematical Sciences and Information Engineering, Nanzan University, 27 Seirei-cho, Seto, Aichi, 489-0863, Japan, E-mail: tomita@nanzan-u.ac.jp, Tel: +81-561-89-2000

[†]Algorithm Development Sub-Team, Genome Diversity Team, Integrated Database Group, Japan Biological Information Research Center, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo, 135-0064, Japan

[‡]Division of Genomic Medicine, Department of Advanced Biomedical Engineering and Science, Tokyo Women 's Medical University, Tokyo, 162-8666, Japan

Step 1 Since the loci with minor allele frequencies of less than 0.1 are likely to have been generated by recent mutations, they were excluded from the genotype data for the determination of haplotype blocks.

Step 2 Initial satellites of a block were constructed using a pair of adjacent SNPs with $D' \geq 0.9$ for all pairs. This is designated as the "minimal block".

Step 3 Using the satellite block, possible extension of the block to an adjacent SNP in ether of direction is examined. If haplotype heterozygosity is unchanged by the extension in one direction, then the block is extended to that direction. If the extension increases the haplotype heterozygosity, then the SNP before the extension is considered as the end of that block. Judging by whether or not to add a locus into a block, it is possible to estimate whether a haplotype has a cumulative relative frequency of 0.95 (or 0.9), constituting a major haplotype.

1.2 Problem of threshold

In the above system for the identification of haplotype blocks, the initial conditions for satellite blocks are $D' \geq 0.9$ as determined by experimental instinct. Alternative conditions also should be considered. The initial conditions are understood to be considerably severe. Therefore, short mininum blocks may be identified in Step 2, but based on these strict conditions, the calculation time for the identification of blocks may become large since the satellite blocks are very short. We therefore examined alternative thresholds for the construction of initial satellite blocks in the following section.

2 Using a threshold of disequilibrium parameters

Imagine that there are 2 linked biallelic loci. Let the major and the minor alleles at locus 1 be 1 and 2 with relative frequencies of p_1 and p_2, respectively, and let those at locus 2 be 3 and 4 with relative frequencies of p_3 and p_4, respectively. Without losing generality, we can assume that $p_1 \geq p_3 \geq 0.5$. Under these conditions, $p_1 \geq p_2$, $p_3 \geq p_4$. Let haplotype frequencies concerning the two loci be as follows.

		locus 2	
		allele 3	allele 4
locus 1			
	allele 1	p_{13}	p_{23}
	allele 2	p_{14}	p_{24}

Table 1: 4 haplotypes on 2 loci.

Linkage disequilibrium parameter D is defined by the following equation.

$$D = p_{13} - p_1 p_3$$

The following equations hold.

$$\begin{aligned} p_{13} &= p_1 p_3 + D, \qquad (2.1) \\ p_{14} &= p_1 p_4 - D, \\ p_{23} &= p_2 p_3 - D, \\ p_{24} &= p_2 p_4 + D \qquad (2.2) \end{aligned}$$

Here, we propose $f = p_{13} + p_{24}$ as a parameter to measure the linkage disequilibrium.

From equations (2.1) and (2.2), we get

$$f = p_{13} + p_{24} = p_1p_3 + p_2p_4 + 2D.$$

Since, $p_2 = 1 - p_1$ and $p_4 = 1 - p_3$,

$$f = p_1p_3 + (1 - p_1)(1 - p_3) + 2D.$$

From above equation, we get

$$D = \frac{p_1(1 - p_3) + (1 - p_1)p_3 + f - 1}{2}. \tag{2.3}$$

Since the range of D changes with allele frequencies, the parameter $D' = D/D_{\max}$ is used, where

$$D_{\max} = \begin{cases} \min[p_1(1 - p_3), (1 - p_1)p_3] & \text{if } D \geq 0 \\ \max[-p_1p_3, -(1 - p_1)(1 - p_3)] & \text{otherwise} \end{cases}$$

The above equations can easily be written as

$$D_{\max} = \begin{cases} (1 - p_1)p_3 & D \geq 0 \\ -(1 - p_1)(1 - p_3) & D < 0 \end{cases}$$

since $p_1(1 - p_3) \geq (1 - p_1)p_3$ and $-p_1p_3 \leq -(1 - p_1)(1 - p_3)$ if $p_1 > p_3 > 0.5$.
Then,

$$D' = \frac{p_1(1 - p_3) + (1 - p_1)p_3 + f - 1}{2D_{\max}}. \tag{2.4}$$

There is a very close relationship between the cumulative relative frequency f of haplotypes (equation (2.4)) and the value of D'. For example, when $D' = 0.9$ and each allele relative frequency is 0.5, then $f = 0.95$ analytically. The similarity between p_1 and p_3 when $f = 0.95$, D' is close to 1. Moreover, when p_1 and p_3 are not similarity, this relationship also collapses. (See Figure 1)

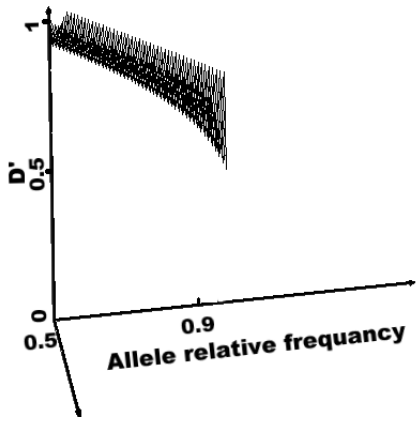

Figure 1: A behavior of D' ($f = 0.95$)

Here, let f_0 as a following equation,

$$f_0 = \max(p_{13} + p_{24}, p_{14} + p_{23}). \tag{2.5}$$

Then, we want to think the range f that can be taken on the condition with enough linkage disequilibrium. The proportion w_2 of haplotypes {1-3, 2-4} is a parameter that we want to give.

It's assumed that above haplotypes are following a uniform distribution with each allele relative frequancy. Then we get

$$p_{13} = p_1 w_2, \tag{2.6}$$
$$p_{24} = (1 - p_3) w_2, \tag{2.7}$$

when $f_0 = p_{13} + p_{24}$. Let $f_2 = (p_1 + 1 - p_3)w_2$ we want to set a parameter. (e.g. w_2=0.95) We want to take $D'(f_2)$ as the threshold of D' as in Step 2 of Section 1.1. This means that we do not take the condition $D' \geq 0.9$, but that we take the condition (we call this "unit of haplotype block".)

$$|D'| \geq |D'(f_2)| \tag{2.8}$$

where, $D' = p_{13}p_{24} - p_{14}p_{23}$ is given by observed data.

However, based on the conditions in Figure 1, where ($f_2 = 0.95$), it is understood that there is a range of D' that is not the gain according to major allele frequencies. We think that $f_2 = 0.95$ or 0.9 constitutes an appropriate condition because 0.95 is a cumulative relative frequency of major haplotypes based on Step 3 of Section 1.1. Thus, adding not only two haplotypes {1-3, 2-4}, but another one haplotype {2-3 or 1-4} as a cumulative relative frequency of haplotypes. After a similar process, the condition is described by the following equation where f_2 is added to p_{23} or p_{14} ,then let $f_3 = \max(1 - p_{23}, 1 - p_{14})$. (We call this "deviation of haplotype".)

$$|D'| \geq \left| \frac{p_1(1 - p_3) + f_3 - 1}{D_{\max}} \right| \tag{2.9}$$

where, D' given by observed data, f_3 is set by user.

Therefore, we think that an optimal initial satellite of a block can be constructed when equations (2.8) and (2.9) for pairwise D' are satisfied for all pairs of loci. Constructing an optimal initial satellite of a block leads to the greatest decrease in estimation time for haplotype in Step 3 of Section 1.1. Consequently, this results in a decrease in the calculation time for identifying haplotype blocks.

2.1 Calculation time

Finally, we considered how calculation time differs in total. We assume that a function indicating the calculation time of inferring the haplotype frequencies is represented by an algorithm linking k loci as $T(k)$. The calculation time of the block identification with the data linked by n loci by the method of Kamatani *et al*. is represented by the following equation

$$f_k(n) = T(n)^2 + (n - s)T(n)^n, \tag{2.10}$$

where s denotes the number of pairs next to each other that are $D' > 0.9$, and s does not depend on n. The calculation time of of our method is given by the following equation

$$f_o(n) = T(n)^2 + g(n), \tag{2.11}$$

where $g(n)$ denotes a function of time for calculating the threshold value for each pair of loci, and it is clearly influenced by the property of the algorithms such that $T(n)^2 > g(n)$.

In order to evaluate the time complexity, equations (2.10) and (2.11) can be rewritten as follows.

$$O(f_k(n)) = nT(n)^n \tag{2.12}$$
$$O(f_o(n)) = T(n)^2 \tag{2.13}$$

We then compare equations (2.12) with (2.13);

$$O(f_k(n)) > O(f_o(n)). \tag{2.14}$$

Therefore, the longer the haplotype block becomes, the calculation time by our method is understood to be shortened greatly compared to that of Kamatani *et al.* .

3 Numerical examples and results

The data of International HapMap Project is then analyzed herein. Three (3) actual data sets were downloaded and regenerated by Furihata *et al.*(2004). A region of the X chromosome is analyzed for 43 or 44 subjects (29 mothers and 15 virtual daughters). The data are unphased data regenerated by them. Figure 2 shows the loci and LD(D') map for data1, Figure 3 shows that for data2. LD maps have been made using the GUI software "Integrated Environmental System for SNPs Data Analysis" (Tomita *et al.*, 2004). Data3 comprises 83 loci ranging from from rs845127 to rs756384; this set was prepared prepared to develop a large data for the purposes of making time comparisons. (The LD map was omitted because it was very large.) The haplotype-block identification was performed on the data by the methods shown in Section 1.1 and our method. The same calculations were run using three versions: our method the method of Kamtani *et al.*, and that of Haploview (Barrett *et al.*, 2005). Figure 2 and Figure 3 show the results of the identifying blocks, as well, and the calculation time of our method was shorter than that of Section 1.1. (See a next section.)

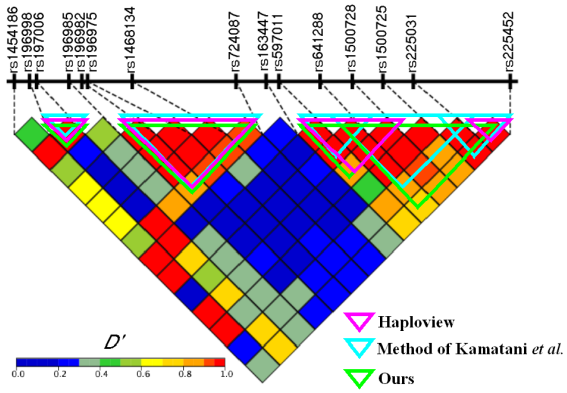

Figure 2: Identifing haplotype blocks on data1

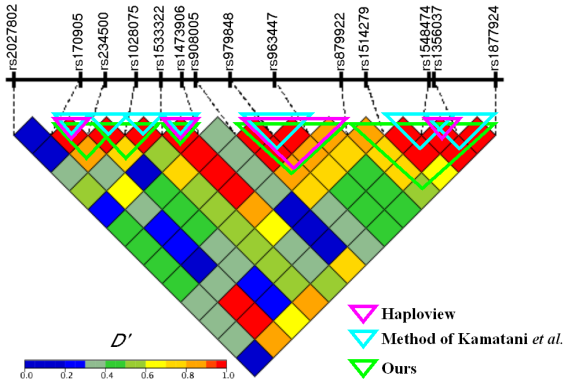

Figure 3: Identifing haplotype blocks on data2

	Haploview	The method of Kamatani *et al*.	Ours
data1	2.145	1.725	0.105
data2	1.763	1.830	0.090
data3	3.434	14.509	2.043

Table 2: Real calculation times by each programs. (seconds)

4 Discussion

The results of the block identification were obtained by each method for three sets of actual data. When the results were obtained, blocks where our method, the method of Kamawani *et al*. and that of Haploview were the identical or differed little were identified. Because it appears to be possible to use the optimized threshold of D', even for data points that differ slightly, this method is convincing.

Next, we compare the calculation time of each program. Table 3 shows the calculation time of each program performed on LapTop (Pentium 600MHz, VineLinux OS). It is understood that there are clear differences in the calculation time as shown in Table 3. When loci become longer, differences in calculation time can be expected to become considerably longer, as shown by calculations using these data sets.

In conclusion, it is understood that almost the same block can be identified by any of these methods, but our proposed method rapidly shortens the calculation time compared with that of Kamatani et al. It is expected that not only the calculation time, but also more appropriate block identification, can be accomplished by optimizing the threshold of D'.

5 Acknowledgment

We wish to express our gratitude to Furihata *et al.* for allowing us access to data. The present study was supported by grants from the New Energy and Industrial Technology Development Organization. It was also supported by grants from Pache Research Subsidy I-A-2, Nanzan University (2006).

References

[1] Barrett J.C., Fry B., Maller J., Daly M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 21(2), 263-265.

[2] Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., 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 Altshuler D. (2002). The structure of haplotype blocks in the human genome. Science. 296, 2225-2229.

[3] Hedrick P.W. (1987), Gametic Disequilibrium Measures: Proceed With Caution. Genetics. 117(2), 331-341.

[4] Kamatani N. edited (2001). Statistical Genetics in Post-Genome (In Japanese). Yodosha in Japan.

[5] Kamatani N., Sekine A., Kitamoto T., Iida A., Saito S., Kogame A., Inoue E., Kawamoto M., Harigai M. and Nakamura Y. (2004). Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. American Journal of Human Genetics. 75(2), 190-203.

[6] Kitamura Y., Moriguchi M., Kaneko H., Morisaki H., Morisaki T., Toyama K. and Kamatani N. (2002). Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm. Annual of Human Genetics. 66(3), 183-193.

[7] Furihata S., Takemura R., Tomita M., Yang W.S., Yasuno K., Nakasige R., Shinkura T., Yasuda T., Imanishi T., Isomura M., Ushijima D., Yanagisawa M., Yotsuji T., Shinohara S., Nomura K., Itakura M. and Kamatani N. (2004). JBIC2004, Tokyo, Japan.

[8] The International HapMap Consortium. (2003). The International HapMap Project, Nature 426(6968), 789-796.

[9] Tomita M. (2002). A Relationship Between a Contingency Table and a Linkage Disequilibrium. Proceedings of the 4th Conference of Asian Regional Section of the International Association for Statistical Computing. 70-73.

[10] Tomita M., Inoue E. and Kamatani N. (2004). Integrated Environmental System for SNPs Data Analysis. Program and Abstracts of The 13th Takeda Science Foundation Symposium on Bioscience. 92.

[11] Zapata C., Carollo C. and Rodriguez S. (2001). Sampling Variance and distribution of the D' measure of overall gemetic diseequilibrium between multiallelic loci. Annual of Human Genetics. 65(4), 395-406.

[12] Zhu X., Yan D., Cooper R.S., Luke A., Ikeda M.A., Chang Y.P., Weder A. and Chakravarti A. (2003). Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. Genome Research. 13, 173-181.