European Journal of Human Genetics (2005), 1–5

www.nature.com/ejhg

ARTICLE

Clustering of haplotypes based on phylogeny: how good a strategy for association testing?

Claire Bardel*[,1], Pierre Darlu[1,2] and Emmanuelle Génin[1,2]

[1]*INSERM U535, Hôpital Paul Brousse, Villejuif, France*

Haplotypes are now widely used in association studies between markers and disease susceptibility locus. However, when a large number of markers are considered, the number of possible haplotypes increases leading to two problems: an increased number of degrees of freedom that may result in a lack of power and the existence of rare haplotypes that may be difficult to take into account in the statistical analysis. In a recent paper, Durrant *et al* proposed a method, CLADHC, to group haplotypes based on distance matrices and showed that this could considerably increase the power of the association test as compared to either single-locus analysis or haplotype analysis without prior grouping. Although the authors considered different one-disease-locus susceptibility models in their simulations, they did not study the impact of the linkage disequilibrium (LD) pattern and of the susceptibility allele frequency on their conclusions. Here, we show, using haplotype data from five regions of the genome of different lengths and with different LD patterns, that, when a single disease susceptibility locus is simulated, the prior grouping of haplotypes based on the algorithm of Durrant *et al* does not increase the power of association testing except in very particular situations of LD patterns and allele frequencies.
European Journal of Human Genetics advance online publication, 23 November 2005; doi:10.1038/sj.ejhg.5201501

Keywords: association test; haplotype phylogeny; cladistic analysis

Introduction

A large number of single nucleotide polymorphisms (SNPs) can now be used to look for an association between a disease and a candidate gene. These markers can either be studied one at a time or jointly in haplotypes. The advantages of haplotypic *versus* single-marker methods have been widely debated in the literature: some studies show that haplotypic tests are more powerful,[1–3] whereas others conclude that single-site analysis should be preferred.[4–6] However, the relative power of these two approaches depends on whether the disease contributing SNPs are among the investigated SNPs or not,[7,8] on the number of disease susceptibility sites,[8] on the disease susceptibility model and on the type of interactions between disease contributing sites.[9,10] The number of SNPs that are considered jointly in haplotypes is also an important parameter since the number of haplotypes increases with the number of investigated SNPs and, consequently, increase the degrees of freedom of tests comparing cases and controls, thus reducing their power. Moreover, as some haplotypes would only be carried by a few individuals, there could be statistical problems owing to small sample sizes making difficult the evaluation of their possible effect on the susceptibility. To face this problem, different strategies have been developed for the grouping of haplotypes. The method of Templeton *et al*[11] consists in building a cladistic phylogenetic tree of the haplotypes and statistically comparing the number of cases and controls carrying haplotypes from the different nested clades. Recently, Durrant *et al*[12] have proposed a different method in which the grouping of haplotypes is based on a

*Correspondence: C Bardel, Génetique Epidémiologique et structure des populations humaines, Hôpital Paul Brousse, Bâtiment Leriche, B.P 1000, 94817 Villejuif Cedex, France. Tel: +(33) 1 45 59 53 68; Fax: +(33) 1 45 59 53 31; E-mail: bardel@vjf.inserm.fr
[2]These authors contributed equally to this work
Received 15 December 2004; revised 18 August 2005; accepted 24 August 2005

distance metric and the association in the different groups is tested by a regression analysis. The method allows the study of large number of SNPs through a sliding window approach and is implemented in the software CLADHC. It differs from the method of Templeton[11] since trees are reconstructed by simple group average linkage (clustering) and rooted, instead of being reconstructed by a parsimony method and unrooted. It also allows the analysis of long haplotypes, whereas the method of Templeton focuses on few SNPs. Based on simulations, Durrant *et al*[12] showed that their clustering method may considerably increase the power to detect an association with a genomic region including a single disease susceptibility locus as compared to single-site or classical haplotype analysis. To perform these simulations, Durrant *et al*[12] considered the haplotype data observed in Caucasians in a 10 Mb region of chromosome 20. To determine if their conclusions remain valid with different genes and linkage disequilibrium (LD) patterns, we present here the results of simulations using real haplotype data from five different genomic regions.

Materials and methods

Description of the data

Data from the Variation Discovery Resource Project[13]

Various genes are sequenced in 23 unrelated European individuals. SNPs are identified within these genes and the most likely haplotypes are reconstructed using Phase v2.0.[14,15] Three genes are studied here:

- Interleukine 13 (IL13): 6919 base pairs (bp) on the chromosome 5q31: 12 bi-allelic loci are kept, defining 14 different haplotypes;
- Plasminogen Activator Urokinase (PLAU): 9274 bp on the chromosome 10q24: 16 bi-allelic loci are kept, defining 10 different haplotypes;
- Tumor Necrosis Factor (TNF): 4830 bp on the chromosome 6p21.3: 10 bi-allelic loci are kept, defining six different haplotypes.

Data from chromosome 20 (HapMap project) A 500 kb region of chromosome 20 sequenced for 30 CEPH trios is randomly chosen (from position 48362908 to 48862907). The most likely haplotypes are reconstructed using Phase v2.0.[14,15] Two different sets of SNPs are studied:

- CHR20_1 (461 kb): 13 randomly chosen SNPs, defining 37 different haplotypes
- CHR20_2 (442 kb): 12 randomly chosen SNPs, defining 38 different haplotypes

Data on the CARD15 region The data set[16] includes 232 families with two affected children and their parents genotyped for 13 SNPs covering 140 kb in the CARD15 region. Haplotypes were reconstructed using GENEHUNTER 2.0b.[17] Haplotypes with missing data were removed and 531 control haplotypes (parental haplotypes non-transmitted to the children) were kept for the analysis. These data include 88 different haplotypes.

The pairwise LD (r^2 value) for these six data sets obtained with GOLD software[18] are presented as a Supplementary figure.

The three tests

The data were analyzed using the CLADHC software,[12] kindly provided by Caroline Durrant. This program considers overlapping sliding windows of SNPs across the haplotypes. In each window, three association tests are performed:

- A single-locus allele-based analysis using Pearson's χ^2 test.
- An haplotype-based logistic regression without grouping of haplotypes referred to as T[h].
- An haplotype-based logistic regression analysis with clustering of haplotypes: a tree of the haplotypes is reconstructed using a distance method and a statistics is calculated at each level of the tree. The statistics at the level maximizing the evidence of a disease-marker association is then retained. This test is referred to as T[MAX].

For each window, the program also provides a significant threshold calculated using the Bonferroni correction. The single-locus test is corrected for the number of SNPs, T[h] is corrected for the number of windows and T[MAX] is corrected for the number of windows and for the number of levels in the tree.

The simulation process

We start by selecting a site as the disease susceptibility (DS) site in the studied gene. In the following, we will assume that the minor allele at this locus is the one that confers the highest risk of disease (DS allele). To generate the genotype of each individual, pairs of haplotypes are randomly sampled with replacement and the disease status is obtained by applying the penetrance f_2, f_1 and f_0 associated to the different genotypes (2, 1 or 0 DS alleles) at the DS locus. The chosen penetrance values correspond to an heterozygote genotype relative risk (GRR) of 1.5 and to homozygote GRRs of 2, 5 and 10, respectively. We keep sampling with replacement until we obtain 1000 cases and 1000 controls. This constitutes a replicate and 1000 such replicates are simulated for each of the studied penetrance vectors. CLADHC[12] was applied on the simulated data either with the DS site being kept or removed. The powers of the three tests (single-locus, T[h] and T[MAX]) to detect

an association are compared as in Durrant *et al*.[12] Windows of six markers are considered for T[h] and T[MAX] tests. Type I errors were evaluated by simulations under the null hypothesis of no association, assigning identical values of the penetrances to all three genotypes ($f_2 = f_1 = f_0 = 0.50$) for both cases and controls. However, since the number of performed tests vary from one method to another, different type I errors are obtained. Therefore, to compare the powers of the methods, the data were reanalyzed with different nominal values, until the observed type I errors equal 1% for all the three tests.

For IL13, PLAU, TNF, CHR20_1 and CHR20_2, all the SNPs are considered as the susceptibility site in turn, except when two sites are in complete LD. In this latter case, only one of the two sites is studied. For the CARD15 region, as the computation time is really longer due to the larger haplotypic diversity in the data set, only nine SNPs out of the 13 are analyzed.

Results

Results of the power computations are presented in Table 1 and in Figures 1 and 2. In Table 1, the power to detect an association is presented for site 13 of the CARD15 region. As expected, the power of the three tests is higher when the homozygote GRR is high, and when the susceptibility site is included in the analysis. The same results are obtained for all the sites on the four genes.

The power of the three tests for an homozygote GRR of 2 is presented in Figures 1 and 2 for different values of DS allele frequency and for different maximum linkage disequilibrium ($LD_{\max}$, based on the r^2 values) between the susceptibility site and another site. Whatever the test, we can see that the power increases when the DS allele frequency and the $LD_{\max}$ increase.

The difference in power for the three tests can be tested by a Friedman two-way analysis of variance. As our data sets are heterogeneous, we test separately the data sets corresponding to long sequences (CHR20_l, CHR20_2 and CARD15) and to short sequences (IL13, PLAU and TNF). The two groups are referred to as LS and SS, respectively. When the susceptibility site is present (Figure 1), we find that the power of the three tests is statistically different at the 1% level for LS and SS. As expected in this case, the single-locus test is more powerful than the two others since the use of haplotypes increases the number of tests without adding any information. The difference in power between the two haplotypic tests is not significant at the 5% level for both LS and SS. When the susceptibility site is removed (Figure 2), the difference in power between the three tests is significant at the 1% level for LS and at the 5% level for SS. The pairwise comparisons of the three tests show that T[h] and T[MAX] are not significantly different at the 5% level for both LS and SS and that the single-locus test is more powerful than the two other tests. The power of the single-locus test is particularly high when $LD_{\max}$ is high. However, it should be noted that for the haplotypic tests, the results in the overlapping windows are highly correlated. Thus, they may suffer a greater loss of power due to the Bonferroni correction than the single-locus test. With a less conservative correction for multiple testing, the difference of power between the single-locus test and the haplotypic tests should be reduced.

For an homozygote GRR of 5, the powers are very high for all the three tests (around 100% for 66% of the tested sites), and no significant difference is observed between the three tests.

In this study, as in Durrant *et al*,[12] we assume that haplotypes can be reconstructed without ambiguity. If this is not the case, haplotype uncertainty should be taken into account in the haplotypic tests and their power will be reduced. The power of the single-locus test will not be affected and thus the difference in power between the tests will be increased.

Discussion

The results obtained in this study turn out to be very different from those published by Durrant *et al*,[12] although the same software is used. For the same GRR values, our

Table 1 Power of the three tests obtained over 1000 simulations for different penetrance vectors in the CARD15 region. The chosen susceptibility site is SNP13

GRR[a]		*Power with susceptibility site (%)*[b]			*Power without susceptibility site (%)*[b]		
Hom	*Het*	*Single locus*	*T[h]*	*T[MAX]*	*Single locus*	*T[h]*	*T[MAX]*
2	1.5	**81.3**	46	39.6	**13.7***	8.6	**11.8***
5	1.5	**99.8**	94.2*	93.6*	**52.9**	31.1	38.3
10	1.5	**100***	**100***	**100***	**96.5***	87.7	**94.8***

The maximum LD between the susceptibility site (SNP 13) and another site equals 0.23. The frequency of the susceptibility allele equals 0.11.
[a]GRR: genotype relative risks; Hom: homozygote; Het: heterozygote; The penetrance for homozygotes carriers of 0 disease susceptibility allele was set to 0.01.
[b]Powers (%) were evaluated on 1000 simulations. Type I error = 1%. T[MAX] refers to as the grouping method proposed by Durrant *et al*[12] and T[h] to the haplotypic test without prior grouping. For each penetrance vector, the best power is indicated in bold.
*Indicates that there is no statistical difference between the values (at the 5% level).

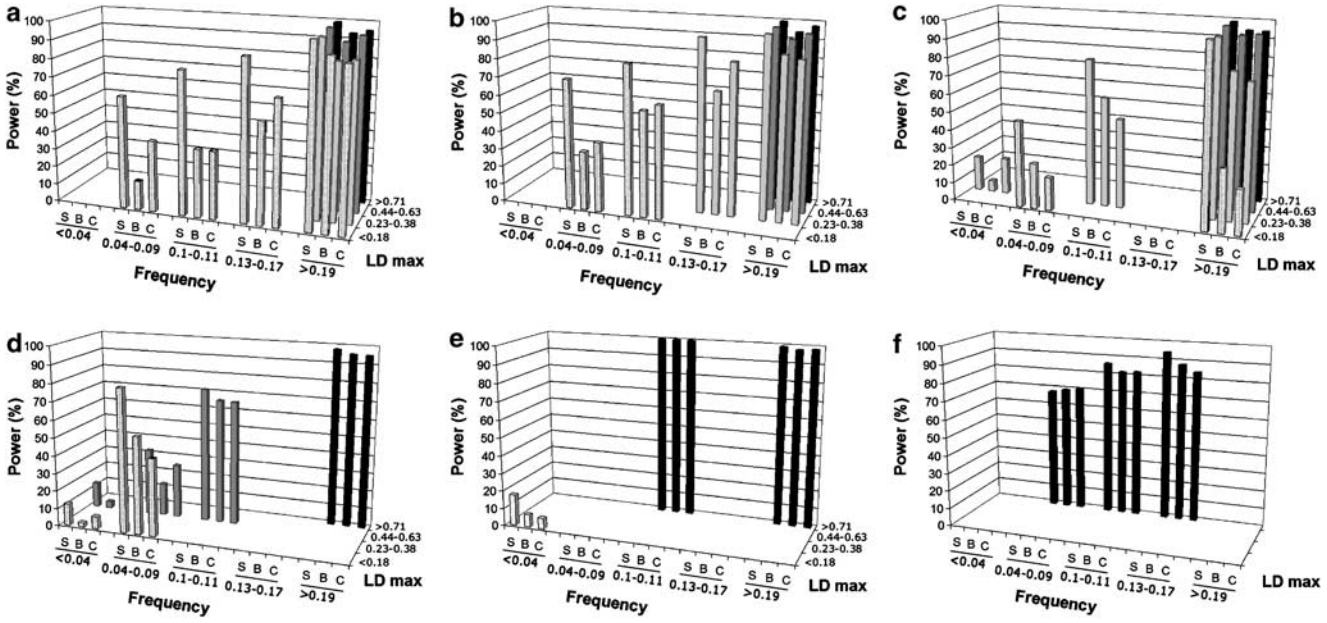

Figure 1 Power of the three tests (S: single-locus test, B: T[h] test, C: T[MAX]) test according to the frequency of the DS allele and to the LD_{max} between the DS site and another site. The DS site is kept in the analysis. (a) CHR20_1, (b) CHR20_2, (c) CARD15, (d) IL13, (e) PLAU, (f) TNF.

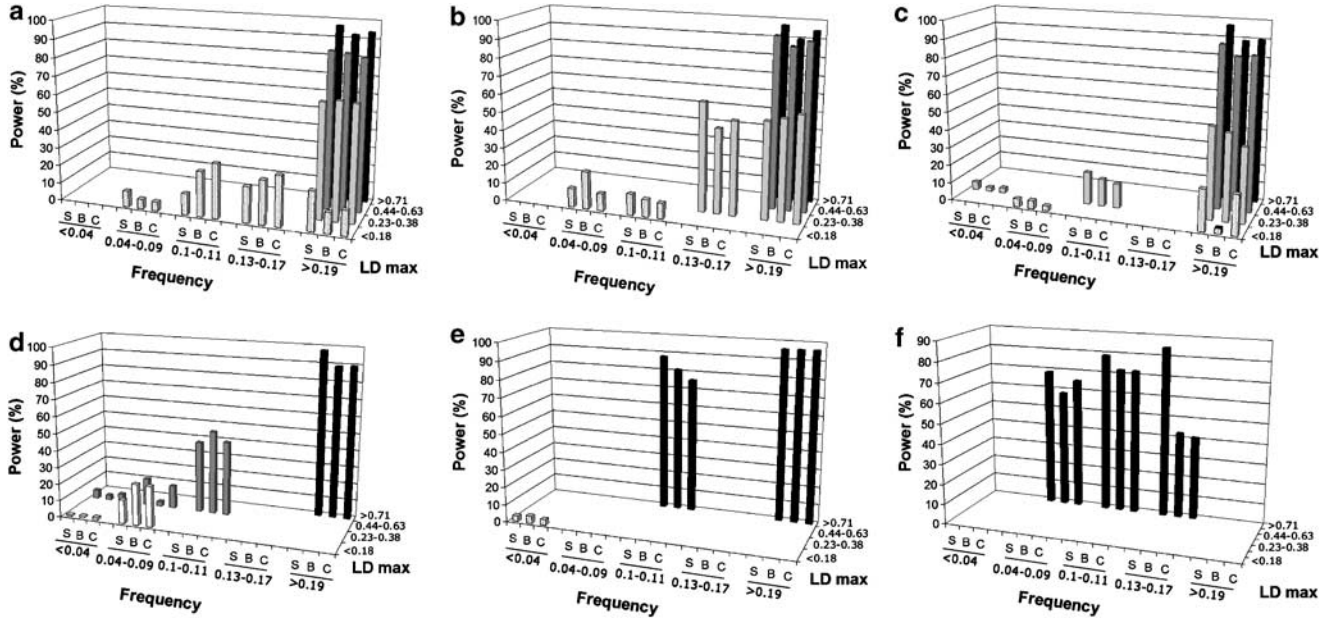

Figure 2 Power of the three tests (S: single-locus test, B: T[h] test, C: T[MAX]) test according to the frequency of the DS allele and to the LD_{mas} between the DS site and another site. The DS site is removed before analysis, (a) CHR20_1, (b) CHR20_2, (c) CARD15, (d) IL13, (e) PLAU, (f) TNF.

power estimates are definitively higher than theirs. A possible explanation may be the large size of their analyzed region (10 Mb region, 5216 markers): although they do not specify the number of markers included in the 'flanking region' used to evaluate the power, we can assume that they are numerous and that a strong correction for multiple testing is used, thus decreasing the power of all the three methods. Our results show that the method of Durrant *et al* does not generally lead to a statistically significant gain in power compared to both single-locus and T[h] tests, T[MAX] being the most powerful test only for six sites out of the 57 sites tested. One must note that

CLADHC is designed to analyze long sequences (several Mb), and thus, in our simulations, we may not be under the optimal conditions, especially for the three short sequences IL13, PLAU and TNF. For the CHR20 data for which distances between markers are close to the ones in the example given by Durrant *et al*[12] (17 of their 23 studied markers are in a 600 kb region of the CFTR gene) we indeed find that the performance of T[MAX] compared to T[h] and to the single-locus test is better than in the other studied regions. However, even then, T[MAX] has the highest power for only five sites out of 25. For the remaining 20 sites, there is a lack of power when using T[MAX] instead of T[h] or a single-locus test. The relative power of the three tests seems to vary with the LD_{max} and with the frequency of the DS allele: all sites but one for which T[MAX] performs better than the two other tests have moderate LD_{max} (<0.6) and moderate frequency of the DS allele (<0.27). The difference between our results and Durrant *et al's* may be explained by the simulation process they use to generate their sample of haplotypes: they are obtained after a particular processing of the observed data that might lead to extreme LD patterns. In our simulations, whole haplotypes are sampled from real data sets, thus keeping the observed LD between the studied locus.

The power of another phylogeny-based association test has also been investigated by Seltman *et al*[19] who extended the method of Templeton[11] to case-parent trios data. However they use a simulation process different from the one used here and in Durrant *et al*.[12] Indeed, rather than choosing a susceptibility site, Seltman *et al*[19] choose groups of at-risk haplotypes on the tree and assume that all the other haplotypes are only carried by control individuals. This may give an advantage to the evolutionary-based method, especially when the at-risk haplotypes are all grouped in a clade.

To conclude, the distance-based grouping of haplotypes as described in Durrant *et al*[12] will usually result in a lack of power as compared to other association tests except for very particular patterns of LD and DS frequency where a slight gain in power may be obtained. However, as suggested in Durrant *et al*,[12] in Seltman *et al*[19] and in Bardel *et al*,[20] the various clustering or phylogeny-based methods might be more powerful when more than one susceptibility site are involved in the disease and be more efficient to precisely localize these susceptibility sites along the haplotypes. Further investigations should confirm these points.

Acknowledgements

We thank Jean-Pierre Hugot, Habib Zouali and Suzanne Lesage from the foundation, Jean Dausset for providing us with Crohn data and Caroline Durrant for kindly providing us with the software CLADHC. We also thank the Runtime team from the LaBRI for letting us run a part of our simulations on their cluster and two anonymous reviewers for their helpful comments.

References

1 Akey J, Jin L, Xiong M: Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* 2001; **9**: 291–300.
2 Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002; **53**: 79–91.
3 Zhang K, Calabrese P, Nordborg M, Sun F: Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002; **71**: 1386–1394.
4 Long AD, Langley CH: The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999; **9**: 720–731.
5 Kaplan N, Morris R: Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet Epidemiol* 2001; **20**: 432–457.
6 Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B: Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* 2005; **28**: 207–219.
7 Judson R, Stephens JC: Notes from the SNP vs. haplotype front. *Pharmacogenomics* 2001; **2**: 7–10.
8 Bader JS: The relative power of SNPS and haplotype as genetic markers for association tests. *Pharmacogenomics* 2001; **2**: 11–24.
9 Culverhouse R, Suarez BK, Lin J, Reich T: A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 2002; **70**: 461–471.
10 Jannot AS, Essioux L, Reese M, Clerget-Darpoux F: Improved use of SNP information to detect the role of genes. *Genet Epidemiol* 2003; **25**: 158–167.
11 Templeton AR, Boerwinkle E, Sing CF: A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* 1987; **117**: 343–351.
12 Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP: Linkage disequilibrium mapping via cladistic analysis of single-nucleotique polymorphism haplotypes. *Am J Hum Genet* 2004; **75**: 35–43.
13 SeattleSNPs. NHLBI Program for Genomic Applications, UW-FH-CRC, Seattle, WA. URL: http://pga.gs.washington.edu [accessed October 2004].
14 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.
15 Stephens M, Donnelly P: A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**: 1162–1169.
16 Hugot JP, Chamaillard M, Zouali H *et al*: Association of NOD2 leucin-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**: 599–603.
17 Daly MJ, Kruglyak L, Pratt S *et al*: Genehunter, version 2.1, 2001.
18 Abecasis GR, Cookson WO: GOLD-graphical overview of linkage disequilibrium. *Bioinformatics* 2000; **16**: 182–183.
19 Seltman H, Roeder K, Devlin B: Transmission/Disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 2001; **68**: 1250–1263.
20 Bardel C, Danjean V, Hugot JP, Darlu P, Genin E: On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics* 2005; **6**.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)