

ARTICLE

Comparison of family based haplotype methods using intragenic SNPs in candidate genes

Catherine Bourgain^{*,1,2}, Emmanuelle Genin¹ and Françoise Clerget-Darpoux¹

¹Unité de Recherche d'Epidémiologie Génétique, INSERM U535, Kremlin-Bicêtre, France

The choice of an optimal marker strategy while analysing intragenic SNPs is presently of crucial importance, given the increasing amount of available data. Classical case/control association studies or family based association tests such as the TDT are very popular. However, as these methods are not able to analyse multiple markers simultaneously, different extensions have been proposed in order to use multiple markers. In the present study, the efficiency of five family based haplotypic methods to detect the role of candidate genes is evaluated and compared between them and with the classical single point TDT. Simulations of intragenic SNP maps are performed in recently founded populations. One or several SNPs are assumed to be the functional polymorphisms following different genetic models. Different modes of SNP combinations underlying the genetic susceptibility (epistasis or heterogeneity) are considered. Whereas haplotypic methods perform better in situations of heterogeneity, the TDT remains the most powerful approach in epistasis models as long as the marginal effect of one the SNPs involved in the susceptibility remains important. Haplotypic methods perform better than the TDT when the marginal effect of each SNP is small. Given the similar characteristics of intragenic LD in both old large populations and recently founded populations, in particular the weak correlation between LD and distance, our results are not likely to be specific to founder populations and can be generalized.

European Journal of Human Genetics (2002) 10, 313–319. doi:10.1038/sj.ejhg.5200808

Keywords: haplotype analysis; complex disease; SNP; linkage disequilibrium; power study

Introduction

Since linkage studies do not allow the fine mapping of genes underlying multifactorial diseases,¹ candidate gene strategies are increasingly used. Efforts have concentrated on the construction of high density biallelic marker (SNPs) maps² and all frequent SNPs may now be identified within candidate genes.

In this context, family based association tests, as the TDT,³ are very popular. However, as the TDT is not able to analyse

multiple markers simultaneously, different extensions have been proposed^{4–6} in order to use multiple markers. One of them⁴ uses the information on identity length among haplotypes of affected individuals. The rationale is that, as argued by the authors, if a response variable tends to be high in one location, it will also tend to be high in nearby locations.

New methods that are not strictly speaking extensions of the TDT have also been proposed such as the Haplotype Pattern Mining method (HPM⁷) or the Maximum Identity Length Contrast statistic (MILC⁸). Whereas HPM looks for haplotype patterns associated with the disease, MILC searches for an excess of haplotype identity length among affected individuals. Contrarily to other methods also using multiple markers simultaneously,^{9,10} HPM and MILC do not suppose that most of the affected individuals carry a unique ancestral mutation. They may thus be used in more general and various contexts.

*Correspondence: C Bourgain, INSERM U535, Batiment Gregory Pincus, 80 rue du Général Leclerc, 94276 Le Kremlin-Bicêtre Cedex, France.
Tel: +33 01 49 59 53 30; Fax: +33 01 49 59 53 31;

E-mail: bourgain@kb.inserm.fr

²Current address: Department of Human Genetics, The University of Chicago, Chicago, USA

Received 15 January 2002; revised 18 March 2002; accepted 20 March 2002

The existence of similarities among haplotypes and more generally the power of haplotype based methods are highly correlated with the characteristics of linkage disequilibrium (LD) among markers in the chromosomal region considered.

LD studies in different parts of the human genome and in a wide range of populations, large or isolated, are now available.^{11–13} Two major characteristics of intragenic LD can be drawn from these studies. Intragenic LD is highly variable and physical distances cannot fully explain this variability.¹⁴ Indeed, as suggested by Jorde,¹⁵ recombination events are rare at this level and do not balance the stochastic LD created by other mechanisms, mainly population admixture and selection in large populations, or genetic drift in founder populations.

Very few genetic risk factors for multifactorial diseases have been identified so far. In the rare cases where a locus has been found, alleles at greater risk seem to be rather common (e.g. ApoE and Alzheimer, HLA and autoimmune diseases). The polymorphisms associated with the disease susceptibility have been described in terms of alleles, rarely resulting from a change at a single SNP, but rather from a combination of several SNPs within the gene. In this way, the different ApoE alleles result from the combination of two SNPs in coding regions (codon 112 and 158), each leading to a change in amino acid sequence.¹⁶ A recent study¹⁷ suggests a more complex model of combination between three SNPs to explain the role of the calpain-10 gene in the susceptibility to NIDDM.

In the present study, the interest of different family based haplotypic methods was evaluated to detect the role of candidate genes using intragenic SNPs and compare them to the classical single point TDT. The power is computed using population based simulations where recently founded populations are considered. Contrarily to a study by Akey *et al*,¹⁸ we do not model a rare unique ancestral mutation shared by most affected individuals and leading to a simple LD pattern decreasing with distance. We rather consider frequent polymorphisms resulting from various SNP combinations as genetic risk factors and a more complex LD pattern. Furthermore, as all frequent SNPs may now be known within a gene, the present study is conducted considering functional polymorphisms as part of the marker map.

Methods

Different methods compared

All the methods used in this study consider case-parent triads, where controls are parental alleles non transmitted to affected offspring. One single point approach and five haplotype based methods are considered.

TDT³ The TDT is a test for linkage and association that is robust to population stratification. Each SNP is tested independently. The test is performed using the GTDT

statistic,¹⁹ as implemented in Gassoc. Note that in a biallelic context, GTDT is equivalent to the classical TDT.

Global TDT⁴ The first way to use multiple markers simultaneously is simply to consider each haplotype as a particular allele and to perform a multiallelic TDT. Haplotypes that can not be unambiguously deduced are discarded. The distribution of this statistic under the null hypothesis is evaluated by simulations.

'Zhao global TDT⁶ Zhao *et al.* have proposed an extension of the multiallelic TDT approach for haplotypic data that takes into account families with ambiguities in phase assignment. Haplotype frequencies are estimated from parental genotypes through an EM algorithm. These frequencies are then used to weigh all possible haplotype combinations in ambiguous families. The distribution of this statistic under the null hypothesis is evaluated by simulations.

Geary Moran test⁴ In order to reduce the degree of freedom of multiallelic TDT tests, Clayton and Jones have proposed to group haplotypes showing similarities. Similarity between two haplotypes is defined as the length, around a focal point, of the contiguous region over which they are identical by state. Length squared may also be used as a similarity measure. However as this latter statistic always gave smaller power than the one using length, it was not considered in the present study. Without a prior idea on the location of the focal point, all SNPs should be considered as focal points one after the other, a test being performed for each. In situations where haplotypes cannot be unambiguously deduced, they are discarded. The testing procedure is the same as for the global TDT described above.

Haplotype pattern mining method (HPM)⁷ This method searches for recurrent haplotype pattern associated with the disease phenotype. A pattern is defined as a group of alleles at adjacent loci, some of them possibly ignored (referred as gaps). The maximum length of the patterns and the maximum number of gaps per pattern are fixed by the user. The association between a wide range of patterns and the disease is tested by χ^2 tests. A *P*-value is then computed for each marker of the haplotype using simulations. This method uses an information on haplotype similarity, defined as a length of identity, however because it allows for gaps, similarity at non-contiguous markers may also be used. In situations where haplotypes cannot be unambiguously deduced, alleles at the ambiguous loci are considered as missing.

Maximum identity length contrast statistic⁸ (MILC) Contrarily to the methods previously described, MILC does not directly contrast the haplotype frequencies between the transmitted and non transmitted groups. MILC contrasts

the mean length of haplotype identity among all transmitted haplotypes with the mean length of haplotype identity among all non transmitted haplotypes. The test is based on the maximum of this contrast among all markers of the haplotype. The exact *P*-value associated to the maximum contrast is computed using a resampling procedure. In situations where haplotypes can not be unambiguously deduced, alleles at the ambiguous loci are considered as missing.

Simulation models

Simulations are performed using the GENOOM software.²⁰

Population model Populations originating from 100 individuals, 10 generations ago, with a number of children per couple randomly drawn from a geometric distribution of mean 3, are considered. Each individual is represented by a pair of chromosomes carrying the disease susceptibility gene.

Typing context A set of eight tightly linked intragenic SNPs numbered SNP1 to SNP8 is considered. No recombination is modelled among them. Linkage disequilibrium among the SNPs may be created along the population history through genetic drift. However SNPs are supposed to be already in LD in the large population from which the 100 founders come from. To create this initial LD, the haplotypes of the founders are randomly drawn from an infinite population with the following standardized LD values.

$D'_{18}=0.6875$, $D'_{23}=0.0625$, $D'_{45}=0.625$, $D'_{46}=0.5625$, $D'_{56}=0.625$, $D'_{456}=0.5312$. D'_{xy} is the D' value between SNP x and SNP y as defined by Lewontin²¹: $D'=D/D_{max}$ if $D>0$ and $D'=D/D_{min}$ if $D<0$.

$D=f(1-1)-f_x(1)f_y(1)$ with $f(1-1)$ frequency of haplotype 1-1, $f_x(1)$ frequency of allele 1 at SNP x and $f_y(1)$ frequency of allele 1 at SNP y . D_{max} and D_{min} are the maximum and minimum values achievable for D given the allele frequencies.

For the simulation Models 5 and 6 (see below) when three non adjacent SNPs are involved, the same D' values were considered but involving different SNPs $D'_{17}=0.6875$, $D'_{34}=0.0625$, $D'_{25}=0.625$, $D'_{28}=0.5625$, $D'_{58}=0.625$, $D'_{258}=0.5312$.

Allele frequencies in the original population of founders are the same for the eight SNPs, either 0.5/0.5 or 0.8/0.2.

As an illustration of LD patterns resulting from such simulation models, the distribution of pairwise LD in two different population replicates is presented on Figure 1. LD may be strong even between the most distant markers of the map.

Genetic model Different genetic models underlying the disease susceptibility are considered. For all of them, parameters are chosen to fit an overall disease prevalence in the population of 5%. Situations where the functional polymorphism corresponds to 1, 2 or 3 SNPs included in the map are modelled.

Models 1 and 2 The functional polymorphism is a single SNP (SNP5). The allele at greater risk, allele 2, has a frequency of 0.2, 0.5 or 0.8. Two different penetrance sets are considered.

Model 1: $f_{2/2}=f$, $f_{1/2}=0.2f$, $f_{1/1}=0.1f$. Model 2: $f_{2/2}=f$, $f_{1/2}=f$, $f_{1/1}=0.1f$. $f_{x/y}$ is the probability of being affected given genotype x/y .

Model 3 Two SNPs are assumed to be involved following a heterogeneity model. The eight SNPs of the map have a frequency of 0.5. The first SNP (SNP3) is involved in the susceptibility for 50% of the affected individuals. The penetrances are as in Model 1. The second SNP (SNP6) is involved in the susceptibility for the remaining 50% of affected individuals (also penetrances as in Model 1). Such a model may for instance result from a gene by environment interaction.

Models 4, 5 and 6 Three epistasis models are considered. The eight SNPs have a frequency of 0.5 for these three models. Model 4 : two SNPs are involved. They are either adjacent (SNP4 and SNP5) or non adjacent (SNP3 and SNP6).

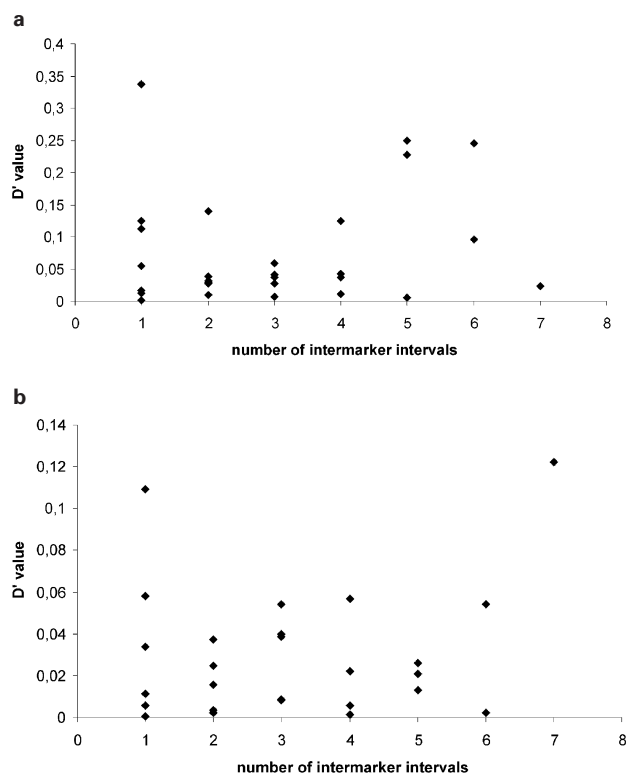


Figure 1 Linkage disequilibrium as a function of the number of intermarker intervals in a random population replicate. D' values are computed on samples of 340 individuals randomly drawn from each population replicate. Population replicate with SNP frequencies of (a) 0.2/0.8 and (b) 0.5/0.5.

Penetrances are : $f_{2-2/2-2}=f$, $f_{2-2/x-y}=0.2f$, $f_{x-y/z-t}=0.1f$, where $(x-y) \neq (2-2)$ and $(z-t) \neq (2-2)$. $f_{x-y/z-t}$ is the probability of being affected for an individual with two SNP haplotypes $x-y$ and $z-t$ (genotypes are x/z for the first SNP and y/t for the second SNP).

Finally we modelled a situation where three SNPs are involved. They are either adjacent (SNP4, SNP5 and SNP6) or non adjacent (SNP2, SNP5 and SNP8). Two penetrance sets are considered.

Model 5 : $f_{2-2-2/2-2-2}=f$, $f_{2-2-2/x-y-z}=0.2f$, $f_{x-y-z/t-u-v}=0.1f$ where $(x-y-z) \neq (2-2-2)$ and $(t-u-v) \neq (2-2-2)$

Model 6 : $f_{1-1-1/1-1-1}=f_{2-2-2/2-2-2}=f_{1-1-1/2-2-2}=f$, $f_{1-1-1/x-y-z}=f_{2-2-2/x-y-z}=0.2f$, $f_{x-y-z/t-u-v}=0.1f$ where $(x-y-z) \neq (1-1-1)$ and $(x-y-z) \neq (2-2-2)$ and $(t-u-v) \neq (1-1-1)$ and $(t-u-v) \neq (2-2-2)$. $f_{x-y-z/t-u-v}$ is the probability of being affected for an individual with three SNP haplotypes $x-y-z$ and $t-u-v$ (genotypes are x/t for the first SNP, y/u for the second SNP and z/v for the third SNP).

In Model 5 a single haplotype (2-2-2) is at greater risk whereas two haplotypes (1-1-1 and 2-2-2) have an equivalent higher risk in Model 6.

Power study

Samples of 100 affected individuals and their two parents are randomly drawn from the population replicates and analysed using the different methods. Power is computed as the proportion of replicates in which at least one SNP of the map is shown to be significantly associated (at the 5% level) with the disease.

For the TDT, Geary Moran and HPM methods, a test is performed for each marker of the map. In order to control for multiple testing, simulations are also performed under the null (same typing context, but with no SNP involved in the disease susceptibility). The following thresholds corresponding to a global 5% type I error when a single candidate gene is tested are used for the power computations : 0.008 for the TDT, 0.003 for the Geary Moran test and 0.02 for the HPM method (performed with a maximum pattern length of eight markers and a maximum number of gaps of 2).

Results

One SNP involved in the disease susceptibility (Models 1 and 2)

Power results for the situation where a single SNP of the map is involved in the disease susceptibility, are presented in Table 1. Whatever the marker frequency and genetic model, the TDT, a single point approach, is more powerful than all haplotypic approaches. The functional polymorphism being one of the SNPs, the other markers do not bring any additional interesting information. Haplotypic approaches integrating a useless level of information are penalised.

Surprisingly the results of the different haplotypic approaches are rather close in the different situations, except the global TDT which is clearly less powerful in almost every situation. Differences may still be observed. Note in particular that in the situation where the allele at greater risk is highly frequent (0.8) the power of the HPM method is strongly reduced whereas MILC performs clearly better than all other haplotypic approaches.

Two SNPs involved in the disease susceptibility, Heterogeneity model (Model 3)

The power results for the heterogeneity model involving two SNPs of the map are presented in Table 2. Three haplotypic approaches -HPM, Zhao global TDT and MILC- give much stronger power than the TDT, even though functional polymorphisms are included in the analysis. Note however that relatively to Model 1 and 2, Model 3 leads to a smaller marginal effect for each of the two functional SNPs. Indeed, if the relative penetrances of the different genotypes for the functional polymorphism are such that $f_{2/2}/f_{1/1}=10$ in Model 1, Model 3 roughly corresponds to $f_{2/2}/f_{1/1}=2.8$ for each of the two functional SNPs. Because the TDT only uses single point information, this test is very sensitive to this decrease in marginal effects.

Nevertheless, such a model is not systematically advantageous for all haplotypic methods. In particular the Global TDT and the Geary Moran test are very sensitive to the pattern of heterogeneity considered.

Table 1 Power^a (%) at the 5% level of the different statistics, as a function of SNP frequency and genetic model when a single SNP is involved in the disease susceptibility and included in the map

Genetic model		Model 1			Model 2		
SNP frequency ^b		0.2	0.5	0.8	0.2	0.5	0.8
Single point approach	TDT	95	100	96	100	50	6
	Global TDT	75	70	59	73	18	8
	Zhao global TDT	81	94	69	81	23	5
Haplotypic approaches	Geary Moran Test	84	95	68	90	28	5
	HPM	86	90	42	91	26	6
	MILC	83	96	83	89	32	10

^aPower computed on samples of 100 families over 1000 simulation replicates; ^bThe eight SNPs of the map including the SNP involved in the disease susceptibility have the same allele frequency.

Two or three SNPs involved in the disease susceptibility, Epistasis models (Model 4, 5, 6)

Power results when the functional polymorphism corresponds to two or three SNPs of the map interacting on an epistasis model are presented in Table 3.

For the two SNP model (Model 4), TDT and the HPM method have an equivalent power, slightly higher than that of MILC, Geary Moran test and Zhao global TDT. In this situation even though susceptibility depends on the haplotypes at two loci, the marginal effect of each locus remains strong enough to allow a good detection power using the TDT. Model 4 roughly corresponds to a $f_{2/2}/f_{1/1}=3.7$ for both SNPs of the functional polymorphism.

Power is dramatically reduced for the six statistics under Model 5 and even more under Model 6. For Model 5 the results do not strongly differ from those of Model 4. In particular TDT performs better than most haplotypic approaches, even though MILC may be slightly more powerful. For Model 6, the global TDT, Zhao global TDT and MILC perform slightly better. This result could have been predicted as two different haplotypes (1-1-1 and 2-2-2) are at equivalent higher risk in this model. The very low power achieved with this model whatever the method considered, prevents however from a clear advantage for haplotype based methods.

The relative location of the functional SNPs on the map (adjacent or not) seems to have no influence on power except

for the MILC and Geary-Moran tests, two methods using an information on haplotype identity length. The sensitivity of these tests to the relative location of the functional SNPs remains relatively minor. However the power of MILC increases when SNPs are adjacent, so that MILC turns out to be the most powerful method under Models 5 and 6. The Geary-Moran test shows a similar gain in power for Model 4 and 5. The very low power achieved with this test under Model 6 may explain that no gain in power is observed under this latter model.

These results are rather intuitive, as a frequency increase of haplotypes made of non-adjacent loci may not systematically lead to an excess of identity length. Parameters such as the number of markers in-between the functional SNPs and the LD pattern among them are of crucial interest.

In order to evaluate the sensitivity of these results to the pattern of LD among the functional SNPs, the power of the six statistics when the three SNPs involved in the susceptibility are in complete linkage disequilibrium (1-1-1 and 2-2-2 are the only observed haplotypes for these loci) has been computed. As respectively 1-1-1 and 1-1-1 and 2-2-2 are the haplotypes at greater risk in Model 5 and 6, a greater power of haplotypic approaches could be expected with this LD pattern. However, results presented in Table 4 are very similar to those observed with the previous LD pattern (shown in Table 3). Power is very low for all the statistics, MILC and TDT giving slightly better results for Model 5. For such complex models of SNP interactions underlying susceptibility to multifactorial diseases, even a priori advantageous situations may prove to be hard to detect without prior idea on functional SNPs.

Table 2 Power^a (%) at the 5% level of different statistics, when the functional polymorphism corresponds to two SNPs included in the map^b interacting on an heterogeneity model (Model 3)

Single point approach	TDT	66
Haplotypic approaches	Global TDT	24
	Zhao global TDT	91
	Geary Moran Test	37
	HPM	92
	MILC	89

^aPower computed on samples of 100 families over 1000 simulation replicates. ^bThe eight SNPs of the map have the same allele frequency of 0.5, including the SNPs involved in the disease susceptibility.

Discussion

The choice of an optimal marker strategy while analysing intragenic SNPs is presently of crucial importance, given the increasing amount of available data.

Neither the methods used nor the situations modelled in the present study are exhaustive. Indeed, the models underlying the susceptibility to a multifactorial disease as well as its intragenic LD pattern is likely to be different for each gene. Furthermore, if SNPs are an important source of genome

Table 3 Power^a (%) at the 5% level of different statistics when the functional polymorphism corresponds to two or three SNPs included in the map^b interacting on an epistasis model. The SNPs are either adjacent or non adjacent

	Genetic model SNP location	Model 4		Model 5		Model 6	
		Adjacent	Non-adjacent	Adjacent	Non-adjacent	Adjacent	Non-adjacent
Single point approach	TDT	91	91	24	24	6	7
Haplotypic approaches	Global TDT	68	69	11	12	11	9
	Zhao global TDT	82	83	10	11	9	10
	Geary Moran Test	83	80	16	12	5	6
	HPM	89	89	14	12	7	7
	MILC	83	79	27	22	15	10

^aPower computed on samples of 100 families over 1000 simulation replicates. ^bThe eight SNPs of the map have the same allele frequency of 0.5, including the SNPs involved in the disease susceptibility.

Table 4 Power^a (%) at the 5% level of different statistics when three SNPs^b in complete linkage disequilibrium are involved in the disease susceptibility following an epistasis model. The SNPs are either adjacent or non adjacent

	Genetic model SNP location	Model 5		Model 6	
		Adjacent	Non-adjacent	Adjacent	Non-adjacent
Single point approach	TDT	21	21	6	6
Haplotypic approaches	Global TDT	13	11	7	9
	Zhao global TDT	11	9	5	5
	Geary Moran	15	15	5	7
	HPM	15	13	5	5
	MILC	22	22	7	9

^aPower computed on samples of 100 families over 1000 simulation replicates. Thresholds corresponding to a global 5% type I error for this LD pattern were 0.01 for the TDT, 0.003 for the Geary Moran test and 0.02 for the HPM method. ^bThe eight SNPs of the map have the same allele frequency of 0.5, including the SNPs involved in the disease susceptibility.

variability, other kinds of polymorphisms (CA repeat) may also be involved in the genetic susceptibility for some diseases.

Even if no general rules can be drawn from our comparative study, the different situations considered, allow us to enlighten interesting points regarding intra-genetic SNP analysis.

Although different family based haplotypic approaches show close results in the different situations considered, three of them (MILC, HPM and Zhao global TDT) tend to give better power. In particular situations where functional polymorphisms are contiguous, the MILC method, using information on haplotype identity length, gives higher power than the other haplotypic methods.

When functional polymorphisms are available from genotyping, the TDT may be more powerful than all or most haplotypic approaches tested. In particular, a haplotype based genetic susceptibility does not imply that haplotypic approaches are more powerful than single point approaches. When several SNPs are involved in the susceptibility, the marginal effect of one of them may be strong enough to allow a better detection power when testing each marker separately rather than all together. Conversely the additional information brought by use of haplotypes may not counterbalance the cost of these tests in terms of increase in degree of freedom. This may be particularly true when there is no prior idea on the functional SNPs so that many uninformative SNPs are tested in the analysis.

However, single point approaches may not systematically be the most powerful approaches. For instance, models of heterogeneity may lead to weak overall marginal effects of the SNPs involved in the susceptibility, drastically reducing the power of the TDT whereas haplotypic approaches like Zhao global TDT, the HPM or the MILC method, remain powerful. This result is rather interesting as heterogeneity models are likely to be frequent in genetic susceptibility to multifactorial diseases as consequences of gene x gene or gene x environment interactions. Particular types of functional polymorphisms, such as combinations of more than two SNPs, may also lead to weak marginal effects of single SNPs. In the paper describing their extension of the TDT, Zhao *et al.*⁶

used such a model (functional polymorphism resulting from the interaction of three available SNPs) to show that their approach performs better than the TDT. Considering a model 'close' to theirs (Model 6) we also found haplotypic approaches to be slightly more powerful than the TDT.

In intragenic context, where LD is not a simple decreasing function of the physical distance, the use of haplotype identity length does not seem to bring an additional information as useful as in systematic screening strategies, where we showed that this information could help to infer the kinship coefficient⁸—except when functional polymorphisms are adjacent. Other kinds of information, such as phylogenetic relationships between haplotypes²² could be of great interest to enhance the power of haplotypic approaches. However, methods using this kind of information are still facing limits²³ and deserve further development.

Founder populations were simulated in the present study. Aside from the LD initially present among the founder individuals, LD in such isolated and expanding populations is generated by the genetic drift occurring during the first generations when the populations remain of small sizes. Recombination however tends to disrupt this drift generated LD. The extent to which LD may be expected in these populations is highly variable, depending on population history, chromosomal regions considered, but also simply by chance given that drift is a stochastic process.^{24–26} At the intragenic level where recombination is rare, LD created during the population history is likely to be observed and not to be a simple decreasing function of the distance.¹⁵

Intragenic LD has also been reported in large populations. If different processes are responsible for the LD pattern in large populations and in founder populations (genetic drift has a greater importance in founder populations than in large populations, where population admixture may have a greater impact), these different mechanisms (drift, admixture, selection) may lead to equivalent stochastic patterns of LD.

The inter SNP LD pattern considered in our study and results presented are thus likely to be unspecific to founder populations and can certainly be generalized to large human populations.

Acknowledgments

Catherine Bourgain acknowledges financial support from the Fondation pour la Recherche Médicale.

References

- 1 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* 1994; **30**: 2037–2048.
- 2 Collins FS, Guyer MS, Charkravarti A: Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; **278**: 1580–1581.
- 3 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am J Hum Genet* 1993; **52**: 506–516.
- 4 Clayton D, Jones H: Transmission/Disequilibrium Tests for Extended Marker Haplotypes. *Am J Hum Genet* 1999; **65**: 1161–1169.
- 5 Dudbridge F, Koeleman B, Todd J, Clayton D: Unbiased application of the Transmission/disequilibrium test to multi-locus haplotypes. *Am J Hum Genet* 2000; **66**: 2009–2012.
- 6 Zhao H, Zhang S, Merikangas K *et al*: Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 2000; **67**: 936–946.
- 7 Toivonen H, Onkamo P, Vasko K *et al*: Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet* 1997; **67**: 133–145.
- 8 Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* 2000; **64**: 255–265.
- 9 McPeck MS, Strahs A: Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* 1999; **65**: 858–875.
- 10 Service SK, Temple Lang DW, Freimer NB, Sandkuijl A: Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder population. *Am J Hum Genet* 1999; **64**: 1728–1738.
- 11 Kidd KK, Morar B, Castiglione CM *et al*: A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* 1998; **103**: 211–227.
- 12 Goddard KA, Hopkins PJ, Hall JM, Witte JS: Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* 2000; **66**: 216–234.
- 13 Reich D, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **441**: 199–204.
- 14 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.
- 15 Jorde LB: Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* 1995; **56**: 11–14.
- 16 Hanlon C, Rubinsztein DC: Arginine residue at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans. *Atherosclerosis* 1995; **112**: 85–90.
- 17 Horikawa Y, Oda N, Cox NJ *et al*: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000; **26**: 163–175.
- 18 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.
- 19 Schaid DJ: General score tests for association of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996; **13**: 423–449.
- 20 Quesneville H, Anxolabéhère D: GENOOM: a simulation package for GENetic Object Oriented Modeling. In the Proceedings of the European Mathematical Genetics Meeting. *Ann Hum Genet* 1997; **61**: 543.
- 21 Lewontin R: The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 1964; **49**: 49–67.
- 22 Templeton AR: A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and apoprotein E locus. *Genetics* 1995; **140**: 403–409.
- 23 Darlu P, Genin E: Cladistic analysis of haplotypes as an attempt to detect disease susceptibility locus. *Genet Epidemiol* 2001; **21**: S601–S607.
- 24 Terwilliger JD, Zollner S, Laan M, Paabo S: Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum Hered* 1998; **48**: 138–154.
- 25 Jorde LB, Watkins WS, Kere J, Nyman D, Eriksson AW: Gene mapping in isolated populations: new roles for old friends? *Hum Hered* 2000; **50**: 57–65.
- 26 Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L: Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2000; **8**: 604–612.