Am. J. Hum. Genet. 68:154–159, 2001

Use of Closely Related Affected Individuals for the Genetic Study of Complex Diseases in Founder Populations

C. Bourgain,[1] E. Génin,[1] P. Holopainen,[2] K. Mustalahti,[3,4] M. Mäki,[3,4] J. Partanen,[2] and F. Clerget-Darpoux[1]

[1]Unité de Recherche d'Epidémiologie Génétique, INSERM U535, Kremlin-Bicêtre, France; [2]Department of Tissue Typing, FRC Blood Transfusion Service, Helsinki; and [3]Institute of Medical Technology, University of Tampere, and [4]Department of Pediatrics, Tampere University Hospital, Tampere, Finland

We propose a method, the maximum identity length contrast (MILC) statistic, to locate genetic risk factors for complex diseases in founder populations. The MILC approach compares the identity length of parental haplotypes that are transmitted to affected offspring with the identity length of those that are not transmitted to affected offspring. Initially, the statistical properties of the method were assessed using randomly selected affected individuals with unknown relationship. Because both nuclear families with multiple affected sibs and large pedigrees are often available in founder populations, we performed simulations to investigate the properties of the MILC statistic in the presence of closely related affected individuals. The simulation showed that the use of closely related affected individuals greatly enhances the power of the statistic. For a given sample size and type I error, the use of affected sib pairs, instead of affected individuals randomly selected from the population, could increase the power by a factor of two. This increase was related to an increase of kinship-coefficient contrast between haplotype groups when closely related individuals were considered. The MILC approach allows the simultaneous use of affected individuals from a founder population and affected individuals with any kind of relationship, close or remote. We used the MILC approach to analyze the role of HLA in celiac disease and showed that the effect of HLA may be detected with the MILC approach by typing only 11 affected individuals, who were part of a single large Finnish pedigree.

Introduction

Founder populations have been extensively used to map and clone genes involved in rare, monogenic diseases, and specific methods have been developed (Hästbacka et al. 1992; Houwen et al. 1994). It has been suggested that such populations, deriving from a small number of individuals and remaining relatively isolated after foundation, could also be useful for the study of complex diseases (Lander and Schork 1994).

We have recently developed a method, the maximum identity length contrast (MILC) statistic (Bourgain et al. 2000), for the study of multifactorial diseases in isolated populations. The major characteristic of the MILC statistic is that, unlike most methods proposed so far for the study of complex diseases in founder populations (Lazzeroni 1998; Service et al. 1999; McPeek and Strahs 1999), MILC does not make the assumption that all carriers of a susceptibility allele inherited it from a single common ancestor. Although this assumption may be true for the rare mutations involved in monogenic diseases, it is likely to be irrelevant for many genetic risk factors involved in complex diseases. Indeed, during the last 20 years, epidemiological studies of complex diseases have resulted in a body of convincing arguments that susceptibility alleles are not rare but common (Bourgain et al. 2000).

The MILC method relies on the fact that, even though not all affected individuals carrying the susceptibility allele have received it from the same single ancestor, they have a stronger kinship coefficient at the disease locus than do nonaffected individuals. Affected individuals are thus more likely to share common haplotypes in the vicinity of the disease locus than are control subjects.

The MILC method looks for an excess of identity length in parental haplotypes transmitted to affected offspring (transmitted haplotypes) compared with those not transmitted (nontransmitted). In Bourgain et al. (2000), we reported results of our studies of the power of this statistic when applied to randomly selected affected individuals from a founder population.

When genealogical records are available, large pedigrees comprising multiple affected individuals may be reconstructed in founder populations (Hovatta et al. 1999; Gasparini et al. 2000; Roks et al. 2000). Fur-

Received September 14, 2000; accepted for publication November 6, 2000; electronically published November 30, 2000.

Address for correspondence and reprints: Dr. Catherine Bourgain, INSERM U535, Batiment Gregory Pincus, 80 rue du Général Leclerc, 94276 Le Kremlin-Bicêtre Cedex France. E-mail: bourgain@kb.inserm.fr

0002-9297/2001/6801-0015$02.00

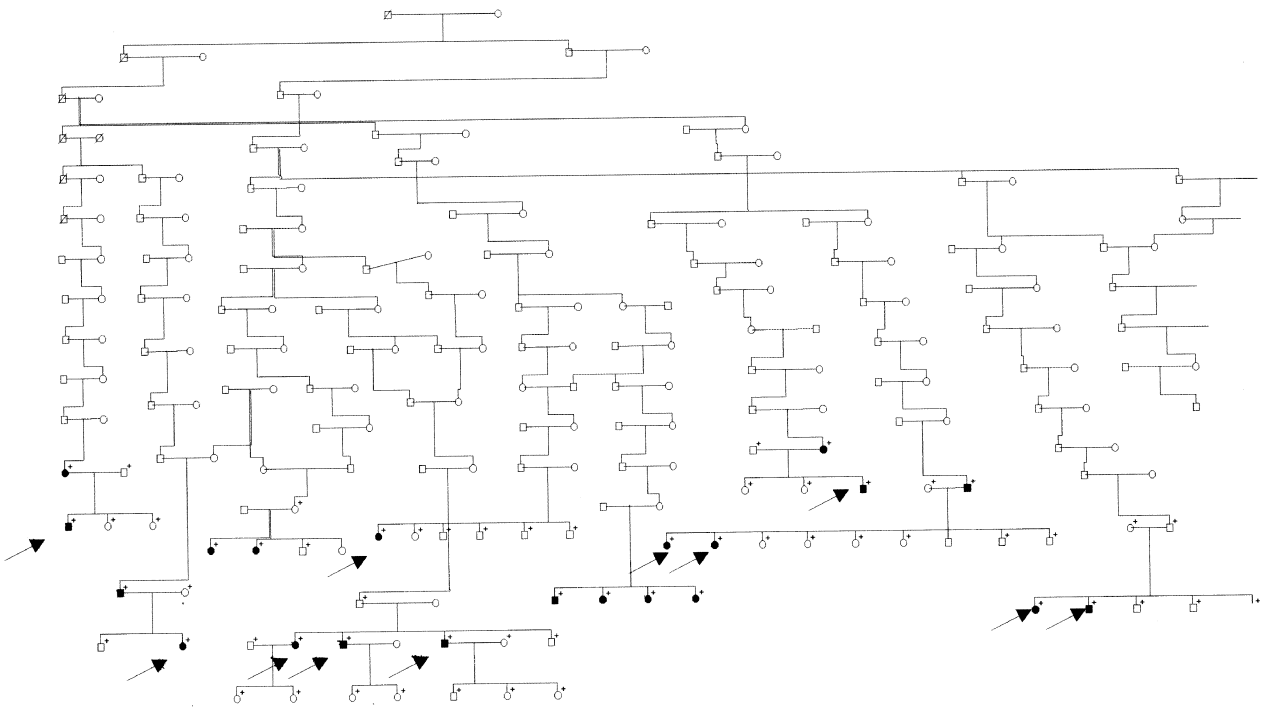

Figure 1 Pedigree of the Kolliskaira family. Plus signs (+) indicate individuals available for genotyping. Arrows indicate affected individuals included in the analysis.

thermore, nuclear families with large sibships may be observed in such populations. For instance, mean family size is eight in the Hutterite population (Ober et al. 1999). It is therefore interesting to evaluate the power of the MILC method, using closely related affected individuals.

Here we report results of a simulation study comparing the power of the MILC method using random affected individuals and using affected sib pairs in founder populations. The usefulness of the method was then illustrated with an analysis of the role of HLA in the determination of celiac disease (CD), using 11 affected individuals in a large Finnish pedigree.

Subjects and Methods

The MILC Statistic

The principle of the MILC method (Bourgain et al. 2000) is to search for an excess of haplotype identity among affected individuals. This excess is expected to signal the presence of a genetic risk factor.

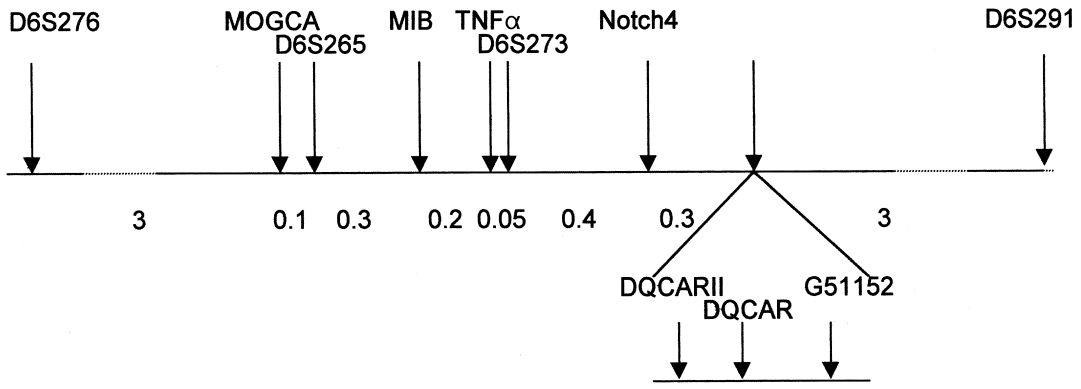

Figure 2 Genetic map of the microsatellites used in the analysis. Distances are shown in centimorgans.

Table 1

Power and Type I Error of the MILC Statistic, at the 5% Level, in Affected Sib Pairs or an Equivalent Number of Random Affected Individuals from the Same Population

Sample Type	Mean Sample Size	Power	Type I Error
Sib	198	64%	5.2%
Random	198	33%	4.9%

NOTE.—Results are based on 2,500 population replicates.

The sample consists of affected individuals and their parents, typed for a set of linked markers. Information on parental genotypes and on offspring genotypes is required. Two groups of haplotypes are contrasted: parental haplotypes transmitted to affected offspring (case group) and parental haplotypes not transmitted (control group). In both groups, all possible haplotype pair comparisons are performed to compute a mean haplotype identity length for each marker of the haplotype. Intermarker genetic distances are used to compute the identity lengths. The difference in mean haplotype identity length between transmitted and nontransmitted haplotypes is then computed for each marker. However, the test that we propose is based on the maximum of this contrast among all markers of the haplotype. The exact *P* value associated with the maximum contrast is computed using an adapted bootstrap procedure.

Simulation Conditions

Populations were simulated with the GENOOM software (Quesneville and Anxolabéhère 1997). Simulation conditions were as in Bourgain et al. 2000. In brief, populations founded by 500 individuals 10 generations ago were simulated. The number of offspring per couple was randomly drawn from a geometric distribution of mean 3. Each individual was represented by two chromosome pairs. Chromosome 1 carried a biallelic disease-susceptibility locus with a frequency of 20% for the susceptibility allele (allele 2) and with penetrances of .03, .06, and .30, respectively, for genotypes 1/1, 1/2, and 2/2. One set of five markers spaced 1 cM apart, with five equifrequent alleles in the original population of founders was also simulated on this chromosome, so that the susceptibility locus was located on the third marker of the set (though being different from this marker). Chromosome 0 only carried the same set of five markers and was used to simulate the absence of a genetic risk factor in the studied region (H0).

Furthermore, two highly polymorphic markers were simulated (HM1 and HM0), one on each chromosome. Both were located on the third marker of the set and had 1,000 different alleles in the 500 founder individuals. These two markers were used to compute kinship coefficients.

Power and Type I Error

Two samples with the same number of affected individuals were extracted from each population replicate. The sib sample consisted of all available nuclear families with one affected sib pair, and the random sample contained an equivalent number of randomly affected individuals with their two parents. From one replicate to another, these samples varied in size: mean sample size was 198 (95% confidence interval [CI] = 133–267). We proceeded this way in order to use the maximum information available from sib pairs in every replicate.

An individual could not be considered twice as affected, that is, once as an offspring and again as a parent. In other words, the trios in which the affected offspring was a parent in another trio were discarded. Indeed, if a parent and an offspring are affected, the two trios (child-parents and parent-grandparents) necessarily share one transmitted haplotype (or a recombinant haplotype). The use of both trios would therefore inflate the type I error.

The power of the MILC statistic was studied in the two samples at a nominal value of 5%, computing the statistic on the marker set of chromosome 1 (H1 situation) among 2,500 population replicates. The type I error of MILC was studied by computing the statistic on the marker set of chromosome 0 (H0 situation) over the same replicates.

Kinship Coefficient

The kinship coefficient was calculated as the probability that two randomly selected alleles are identical by descent at a given locus. In a group of haplotypes, the exact kinship coefficient at the third locus was available from the highly polymorphic markers HM1 and HM0 by allele comparisons of all possible haplotype pairs. For chromosome 1, the kinship coefficient measured corresponds to that of the susceptibility locus.

Table 2

Frequencies of the Susceptibility Allele among Transmitted and Nontransmitted Haplotype Groups in Sib Samples and Random Samples

Sample Type	Mean Frequency of Susceptibility Allele[a]		Mean Difference of Allele Frequency
	Transmitted Haplotypes	Nontransmitted Haplotypes	
Sib	.54	.26	.28
Random	.42	.20	.22

[a] Mean values were calculated on the basis of all 2,500 replicates.

Table 3

Mean Kinship Coefficients for Transmitted and Nontransmitted Haplotype Groups in Samples of Affected Sib Pairs and Random Affected Individuals

		Mean Kinship Coefficient under H1[a]			Mean Kinship Coefficient under H0[a]	
Sample Type	No. of Haplotypes	Transmitted Haplotypes	Nontransmitted Haplotypes	Difference	Transmitted Haplotypes	Nontransmitted Haplotypes
Sib	396 (99 pairs)	.0086	.0057	.0029	.0055	.0055
Random	396 (198 subjects)	.0054	.0041	.0013	.0041	.0041

[a] Mean values were calculated on the basis of all 2,500 replicates.

CD

CD is an autoimmune-like disorder characterized by malabsorption and injury to the small intestine due to the sensitivity of intestine to gliadin. CD is a multifactorial disease for which a genetic component is clearly established; the only locus that has been confirmed in independent studies is in the major histocompatibility complex (MHC) on 6p21.3.

Koilliskaira Family

Initially, Finnish families with multiplex CD were recruited through the Finnish Celiac Society by advertising in the patients' national newsletter. A total of 140 families that fulfilled the criteria were identified. Forty of the families were living in the northeastern part of Finland, and their genealogical origins were investigated, to detect interrelated families that might have a common ancestor. The main sources of information were public church and state official registries; the approach was basically similar to the one described by Varilo (1999). Ten families were found to have a common ancestor who lived around 1550 in the Koilliskaira region in northeastern Finland, giving rise to a single large pedigree. This family is hereafter referred to as the Koilliskaira family (fig. 1). The Koilliskaira region was first populated by a small number of settlers in the early 16th century and has remained relatively isolated, with only a low rate of immigration from other parts of Finland (for details see Varilo 1999). The protocol for collection of DNA samples was approved by the ethics committee of the Tampere University Hospital.

Genetic Markers

The following HLA-linked microsatellites were used in this study: D6S276, MOGCA, D6S265, MIB, TNFα, D6S273, Notch4, DQCARII, DQCAR, G51152, and D6S291. Typing was performed as described by Karell et al. (2000). Genetic distances between microsatellites are shown in figure 2 (Foissac and Cambon-Thomsen 1998).

Transmitted and nontransmitted parental haplotypes could be determined for 11 affected individuals from seven nuclear families (four families with one affected offspring, two families with affected sib pairs, and one family with an affected sib trio), either directly or by using information from nonaffected sibs. We excluded seven affected individuals of the pedigree, because of incomplete parental information. Individuals included in the analysis are indicated by arrows in figure 1.

Results

Power Study

Table 1 compares the power and type I error of the MILC statistic obtained using sib samples with those obtained using random samples. With a given sample size, at a type I error level of 5%, the power of the MILC statistic is almost twice as high for sib pairs (64%) as for random affected individuals (33%). Because the type I errors computed under H0 correspond to the fixed significance level (5.2 and 4.9 are within the 1% CI of the fixed significance level of 5%), the test remains valid in the presence of closely related individuals.

The different allelic distributions at the disease locus in the various groups may explain the origin of the observed power difference between the two kinds of samples. Table 2 shows the frequency of the susceptibility allele (allele 2) in haplotype groups for the sib and random samples. In the nontransmitted haplotype group of the random sample, this allele had a frequency of .20. This was expected because this group is representative of the general population. In both the sib sample and

Table 4

Location of the MILC Statistic, and Associated *P* Value, for 11 Affected Individuals of the Koilliskaira Family

No. of Markers[a]	MILC Location	*P*
12	G51152	.009
5	MIB–G51152	.013

[a] Analysis with all 12 markers and with only 5 markers (with markers between MIB and G51152 considered as a unique supermarker).

the random sample, the allele frequency was higher. In particular, the frequency was very high (.54) in the transmitted haplotype group of the sib sample. The contrast between this frequency in transmitted and nontransmitted haplotype groups was stronger in the sib sample than in the random sample. The difference in power may also be explained by comparing the characteristics of the kinship coefficients in the two samples.

Kinship Coefficient

As shown in Bourgain et al. 2000, the power of MILC relies on the existence of a kinship-coefficient contrast between transmitted and nontransmitted haplotype groups. Table 3 presents the kinship coefficients in transmitted and nontransmitted haplotype groups for marker HM1 (H1 situation) and marker HM0 (H0 situation) in the two kinds of sample. Not surprisingly, haplotypes of the sib samples were more related to each other than haplotypes from random samples, regardless of the kind of haplotype group considered. However, a contrast in kinship coefficients between transmitted and nontransmitted haplotype was observed for sib and random samples only in the presence of a susceptibility factor (for marker HM1).

This contrast was stronger for the sib samples (.0029) than for the random samples (.0013), in agreement with the power increase and with the previous observation on the frequency contrast. The greater the kinship-coefficient contrast, the more powerful the MILC method became. A possible explanation for this excess of kinship-coefficient contrast for sib pairs is that the disease-susceptibility allele (allele 2) is more likely to segregate within families with multiple affected individuals. The gain of power is also observed for other kinds of relatives—for example, families that included more than two affected sibs or cousins (data not shown); however, it tends to decrease as the relationships become more remote. In addition, the MILC method allows the combined use of different relatives in a single study.

Application to the Genetic Study of CD

The MILC statistic was first computed with each of the 12 microsatellites considered separately. Because markers MIB to G51152 are very closely linked and in strong linkage disequilibrium (LD) (Karell et al. 2000), a second analysis was performed considering markers MIB to G51152 as a unique "supermarker," thereby calculating the MILC statistic on the basis of five markers. Results are shown in table 4.

The MILC method allowed the detection of a genetic risk factor for CD in the MHC region ($P < .01$), using only 11 affected individuals. Note that for both analyses the MILC was observed for the marker closest to the HLA DQ locus at which the functional role of a heterodimer in CD is well established (Sollid et al. 1989; Mazzili et al. 1992; Clerget-Darpoux et al. 1994).

Discussion

The MILC statistic was designed to search for genetic risk factors for complex diseases in founder populations. Indeed, without making any assumption about the existence of a single ancestor mutation, the MILC method can be used to detect frequent variants. As illustrated in the present report, the MILC method is able to detect the effect of HLA in the susceptibility to CD although the most at-risk HLA genotype has a high frequency in the general population of Finland (~25%; authors' unpublished data).

Our results show that the power of the MILC statistic is enhanced in the presence of closely related affected individuals. Indeed, as few as 11 affected individuals, who were part of a large pedigree, were sufficient to detect the role of HLA in CD. The MILC method can thus be considered a useful tool for the study of complex diseases in large pedigrees from founder populations. However, because different kinds of data, such as data from a single affected child or multiple affected sibs, can be used simultaneously with the MILC approach, this method is useful even in the absence of large pedigrees. The information extracted from a population can be maximized.

In the present version of the test, nontransmitted parental haplotypes are required for the statistic. Indeed, seven affected individuals of the Koilliskaira family had to be excluded from the analysis because of the lack of parental information. We would like to stress the importance of collecting parental information when sampling affected individuals. If parents are not available, the typing of nonaffected sibs should help to infer parental haplotypes. The MILC method will be further developed for use in such a situation.

It is also important to note that LD in the HLA region is very strong (Karell et al, 2000), probably because of the strong selection processes that have occurred during human history. This region is thus particularly informative for the MILC statistic (Bourgain et al. 2000), as would be other genomic regions that have undergone a strong selection, leading to large LD. In other regions where LD typically extends over smaller distances, particular attention should be given to the choice of markers that exhibit the maximum level of LD.

References

Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F (2000) Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 64:255–265

Clerget-Darpoux F, Bouguerra F, Kastally R, Semana G, Ba-

bron MC, Debbabi A, Bennaceur B, Eliaou JF (1994) High risk genotypes for celiac disease. C R Acad Sci III 317:931–936

Foissac A, Cambon-Thomsen A (1998) Microsatellites in the HLA region: 1998 update. Tissue Antigens 52:318–52

Gasparini P, Palancia P, Lo Vecchio F, Villela M, Villela A, Urbani M, Grosso G, Abiuso F, Di Giovine M, Bertoldo F, Zelante F (2000) Use of isolated inbred human populations for the study of complex traits: the Carlantino Project. Eur J Hum Genet 8:168

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8:380–386

Hovatta I, Varilo T, Suvisaari J, Terwilliger JD, Ollikainen V, Arajarvi R, Juvonen, Kokko-Sahin ML, Väisänen L, Mannila H, Lönnqvist J, Peltonen L (1999) A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. Am J Hum Genet 65:1114–1124

Karell K, Klinger N, Holopainen P, Levo A, Partanen J (2000) Major histocompatibility complex (MHC)–linked markers in a founder population. Tissue Antigens 56:45–51

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lazzeroni LC (1998) Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am J Hum Genet 62:159–170

McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65: 858–875

Mazzilli MC, Ferrante P, Mariani P, Martone E, Petronzelli F, Triglione P, Bonamico M (1992) A study of Italian pediatric coeliac disease patients confirms that the primary HLA association is to the DQ (a*0501, b*0201) heterodimer. Hum Immunol 33:133–139

Ober C, Hyslop T, Hauck WW (1999) Inbreeding effects on fertility in humans: evidence for reproductive compensation. Am J Hum Genet 64:225–231

Quesneville H and Anxolabéhère D (1997) GENOOM: a simulation package for GENetic Object Oriented Modeling. Ann Hum Genet Suppl 61:543

Roks G, Sandkuijl LA, Van Swieten JC, Snijders PJLM, Van Broeckhoven C, Oostra B, Van Duijn CM (2000) Alzheimer's disease in a genetically isolated population: the GRIP study. Eur J Hum Genet 8:169

Service SK, Temple Lang DW, Freimer NB, Sandkuijl A (1999) Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder population. Am J Hum Genet 64:1728–1738

Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, Thorsby E (1989) Evidence for a primary association of celiac disease to a particular HLA-DQ α/β heterodimer. J Exp Med 169: 34–5–350

Varilo T (1999) The age of the mutations in the Finnish disease heritage: a genealogical and linkage disequilibrium study. PhD thesis, University of Helsinki, Helsinki