ARTICLE

Homozygosity Haplotype Allows a Genomewide Search for the Autosomal Segments Shared among Patients

Hitoshi Miyazawa,* Masaaki Kato,* Takuya Awata, Masakazu Kohda, Hiroyasu Iwasa, Nobuyuki Koyama, Tomoaki Tanaka, Huqun, Shunei Kyo, Yasushi Okazaki, and Koichi Hagiwara

A promising strategy for identifying disease susceptibility genes for both single- and multiple-gene diseases is to search patients' autosomes for shared chromosomal segments derived from a common ancestor. Such segments are characterized by the distinct identity of their haplotype. The methods and algorithms currently available have only a limited capability for determining a high-resolution haplotype genomewide. We herein introduce the homozygosity haplotype (HH), a haplotype described by the homozygous SNPs that are easily obtained from high-density SNP genotyping data. The HH represents haplotypes of both copies of homologous autosomes, allowing for direct comparisons of the autosomes among multiple patients and enabling the identification of the shared segments. The HH successfully detected the shared segments from members of a large family with Marfan syndrome, which is an autosomal dominant, single-gene disease. It also detected the shared segments from patients with model multigene diseases originating with common ancestors who lived 10–25 generations ago. The HH is therefore considered to be useful for the identification of disease susceptibility genes in both single- and multiple-gene diseases.

Current genetic approaches focus on the identification of disease susceptibility genes (hereafter referred to as "disease genes") by exploiting the cosegregation of the disease phenotype over generations with a disease gene as well as a set of polymorphic marker types in its neighborhood (i.e., the haplotype). In a large family including multiple patients with a specific disease, the disease gene is usually derived from a single ancestor. On the basis of this assumption, haplotype analysis[1] or linkage analysis[2] has been used to find the gene. In affected-sib-pair analysis, a sib pair affected with the same disease is considered to share the same disease gene, inherited from their parents. The gene is searched for by looking at the genetic markers shared by the pair.[1,3] In whole-genome association studies, researchers try to capture segments containing disease-risk alleles derived from a limited number of very ancient ancestors where a haplotype block is the ultimate unit of search.[4–6] Because the haplotype contains the canonical information for every approach, determination of the haplotype is considered to greatly simplify the analyses.[7] However, the haplotype is not easy to identify in diploid organisms such as humans, because the genotypes of polymorphic markers are obtained as a mixture of those of the two alleles. Although many methods have been developed to reconstruct the haplotype,[1,7–9] their capabilities are limited. It is currently not possible, at least on a genomewide basis, to obtain haplotype information from an arbitrary subject or to compare two unrelated subjects in order to search for chromosomal segments sharing the same haplotype. In this study, we introduce the homozygosity haplotype (HH), which overcomes a part of this problem. The HH is a form of haplotype described by the homozygous SNPs, and, therefore, it is easily obtained genomewide. Using a family affected with Marfan syndrome (MIM 154700) and patients with model multigene diseases, we demonstrate how HH analysis allows the identification of the location of disease genes.

Material and Methods

Definition of Terms

Homozygosity haplotype (HH).—An HH is a haplotype described by only homozygous SNPs and is obtained by the deletion of heterozygous SNPs (fig. 1*A*i), leaving only the homozygous SNPs (fig. 1*A*ii). At this point, the haplotype of each chromosome is uniquely determined, because all SNPs are homozygous (fig. 1*A*iii). Note that both copies of homologous autosomes have the same HH over their entire length.

Comparable SNP (compSNP).—A compSNP is a SNP that is homozygous in two subjects (fig. 1*B*). We can compare the HHs between two subjects by use of the compSNPs (fig. 1*C*).

Region with a conserved HH (RCHH).—An RCHH is a run of compSNPs matched for allelic type, the genetic length of which is longer than the cutoff value (fig. 1*C*). An RCHH is bounded by either a mismatched compSNP(s) or by the end(s) of an autosome. The RCHHs shared by multiple subjects are the overlap of the RCHHs for each subject pair (fig. 1*D*).

Region from a common ancestor (RCA).—An RCA is an autosomal

Department of Respiratory Medicine (H.M.; N.K.; T.T.; H.; K.H.), Department of Cardiovascular Surgery (M. Kato; S.K.), Division of Endocrinology and Diabetes, Department of Medicine (T.A.), Division of Radioisotope Laboratory, Biomedical Research Center (T.A.), Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine (M. Kohda; H.I.; Y.O.), and Division of Translational Research, Research Center for Genomic Medicine (M. Kohda; H.I.; Y.O.), Saitama Medical University

Received January 23, 2007; accepted for publication March 16, 2007; electronically published May 2, 2007.

Address for correspondence and reprints: Prof. Koichi Hagiwara, Department of Respiratory Medicine, Saitama Medical University, 38 Morohongo, Moroyama-machi, Saitama 350-0495, Japan. E-mail: hagiwark@saitama-med.ac.jp

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2007;80:1090–1102. © 2007 by The American Society of Human Genetics. All rights reserved. 0002-9297/2007/8006-0009$15.00
DOI: 10.1086/518176

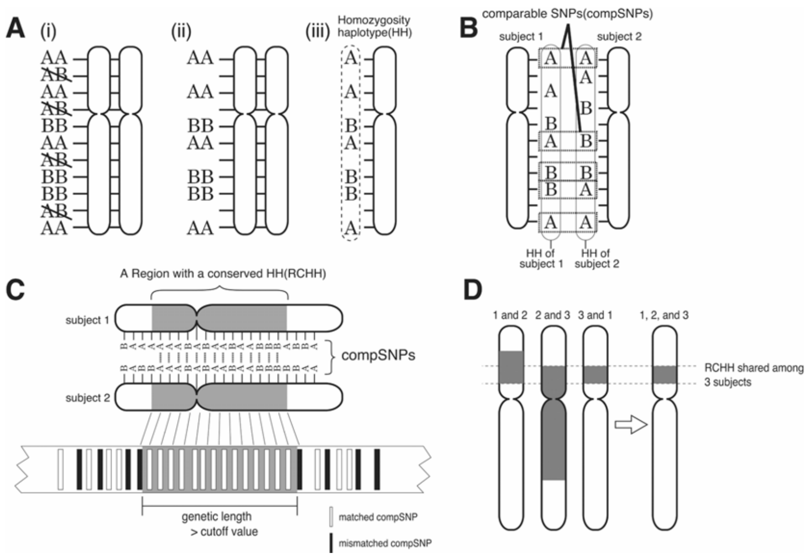


Figure 1. HH analysis. *A,* "A" denotes the major allelic type for each SNP. "B" denotes the minor allelic type for each SNP. *B,* Definition of a compSNP. *C,* Definition of an RCHH. An RCHH has a genetic length longer than the cutoff value. *D,* The definitions of an RCHH shared among multiple patients.

region where subjects share a chromosomal segment derived from a common ancestor (i.e., a segment identical by descent) (fig. 2*A*). In an RCA, subjects share the same segment on one or both copies of their homologous autosomes and thus share the same HH. Conversely, when subjects have the HH in a region, it suggests the presence of an RCA. Note that the RCA is unknown, and its presence is merely predicted through the RCHHs (see the section entitled "The RCHHs, False Negatives, Type A False Positives, and Type B False Positives").

The average genetic length of the RCAs decreases over generations. Figure 2*B* is a model pedigree with common ancestors (A and B). Two descendants (M and N), who are *m* and *n* generations removed from their common ancestors, share the RCAs derived from A and B. Assuming that the spouses (shown in gray shapes in fig. 2*B*) are not the descendants of A or B, then RCA(*m*,*n*) is the ratio of the total genetic length of the A- or B-derived RCA to the entire length of the autosomes. It is expressed as

$$\mathrm{RCA}(m,n{:}m \geqslant n) = \begin{cases} 2^{-m+1} & m \geqslant 1, n = 0 \\ \dfrac{3}{4} & m = 1, n = 1 \\ 2^{-m-n+2} & \text{otherwise.} \end{cases} \qquad (1)$$

A detailed description of the deduction of equation (1) is given in appendix A. Note that RCA(1,0) is equal to 1, indicating that a parent and a child (i.e., $m = 1$, $n = 0$) share the RCAs over the entire lengths of their autosomes.

Crossover Model and Data Analysis

We used the Haldane's Poisson process model[10] for the occurrence of crossovers and performed all calculations on the basis of this model. Information on SNPs used by the 500K GeneChips Mapping Array Set (Affymetrix) was summarized in the GeneChip annotation files (4/13/2006 version; see Affymetrix Web site), where, for each SNP, the genetic distance from the telomere of the short arm of the chromosome was obtained by interpolation from the sex-averaged data by deCODE Genetics.[11] The genetic length of an RCHH is the genetic distance between its bounding compSNPs.

We restricted our analysis to a total of 492,554 SNPs that had assigned dbSNP refIDs (see National Center for Biology Information Web site). The computer programs were written in the C programming language and were compiled by the GNU C compiler 4.0 (see the GNU Compiler Collection Web site). The program is available from our Web site and from the Saitama Medical University Web site.

The RCHHs, False Negatives, Type A False Positives, and Type B False Positives

When subjects who have common ancestors suffer from the same disease, the RCAs are the candidate regions in which to look for the disease gene. Because many RCAs are contained in the RCHHs, we established an algorithm that detects the RCHHs, thereby allowing us to identify the disease gene. As described, an RCHH is defined when a run of type-matched compSNPs is longer than the cutoff value (fig. 3*A*). Many RCHHs contain the RCAs; however, some do not. We defined three types of errors (fig. 3*B*). The false negatives are the RCAs that are not contained in the RCHHs. The type A false positives are the RCHHs that do not contain the RCAs. The type B false positives are the spaces between a containing RCHH and a contained RCA. The equations to calculate each of these errors are given in appendix A. Before the analysis, we calculated the ratios of the false negatives to the total length of the RCAs and of the type A and type B false positives to the entire length of the autosomes for a range of cutoff values, and we selected a value that minimizes the influence of errors.

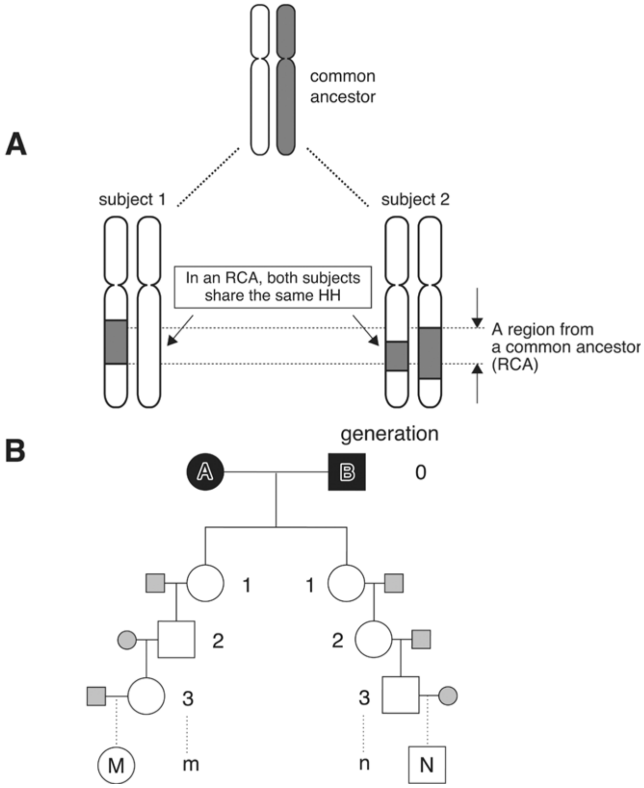

common ancestor
A
subject 1
subject 2
In an RCA, both subjects share the same HH
A region from a common ancestor (RCA)
B
generation
A
B
0
1
1
2
2
3
3
M
m
n
N

Figure 2. The RCA. *A,* The definition of an RCA. Gray regions are those derived from a single ancestral autosome from a single ancestor (segments identical by descent). A subject may have two copies of the segments in case inbreeding exists. *B,* A model pedigree. "A" and "B" denote the common ancestors. M and N are *m* and *n* generations away from the common ancestors. Direct offspring are shown by unfilled shapes. The spouse of each offspring is shown as a gray shape.

Human Subjects

This study was approved by the institutional review board of Saitama Medical University. All DNA samples were purified from peripheral blood drawn after written informed consent had been obtained. A family that included multiple patients with Marfan syndrome was genotyped, as were 46 unrelated subjects. In addition, the genotyping data from 45 unrelated Japanese subjects who had been enrolled in the International HapMap project (see International HapMap Web site) were obtained from the Affymetrix Web site (an average of 199,400 SNPs per subject had confidence values <.05).

Genotyping

Genomewide SNP genotypings were performed using the 500K GeneChips Mapping Array Set (i.e., the GeneChip Human Mapping 250K Nsp Array plus the GeneChip Human Mapping 250K Sty Array) (Affymetrix) or either of the two arrays. (Hereafter, the 500K GeneChips Mapping Array Sets will be abbreviated as "500k GeneChips" and the GeneChip Human Mapping 250K Nsp Array as "250k GeneChips.") 500k GeneChips was used for the analysis of the family with Marfan syndrome. 250k GeneChips was used for the multigene disease simulation.

Pools of Subjects

In the multigene disease simulations, genotyping data of patients who share an RCA at a specific position were constructed by replacing that part of their genotyping data with the genotyping data of a specific subject who acts as a common ancestor. The length of the replaced segment (*x,* in centimorgans) was taken at random from an exponential distribution with a probability density function of

$$f(x) = \lambda e^{-\lambda x} \; ,$$

$$\lambda = \frac{m}{100} \; , \tag{2}$$

where *m* is the age, in generations, of the common ancestor.

Statistical Analysis

The numbers of subjects who share an RCHH at a given position on an autosome were compared between the patient pool and the control pool. The assumption was made that

$$u_0 = \frac{\hat{P}_1^* - \hat{P}_2^*}{\sqrt{\hat{P}^*(1-\hat{P}^*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

has a standard normal distribution, where $\hat{P}_1^* = \frac{x_1+0.5}{n_1+1}$, $\hat{P}_2^* = \frac{x_2+0.5}{n_2+1}$, $\hat{P}^* = \frac{x_1+x_2+0.5}{n_1+n_2+1}$, x_1, and x_2 are the numbers of subjects sharing RCHHs in the patient pool and the control pool, respectively, and n_1 and n_2 are the total numbers of subjects in the patient pool and the control pool, respectively. The *P* value was calculated by

$$P = \int_{u_0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \; .$$

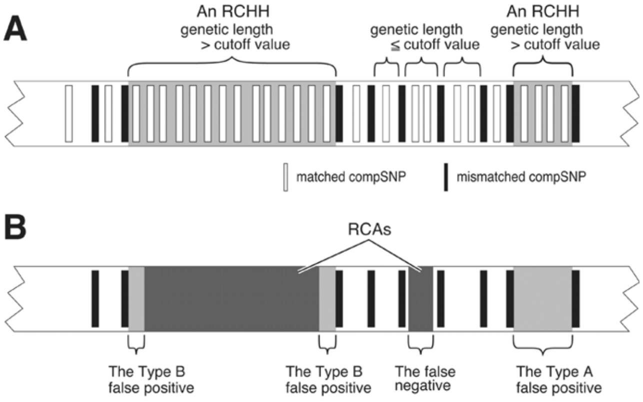


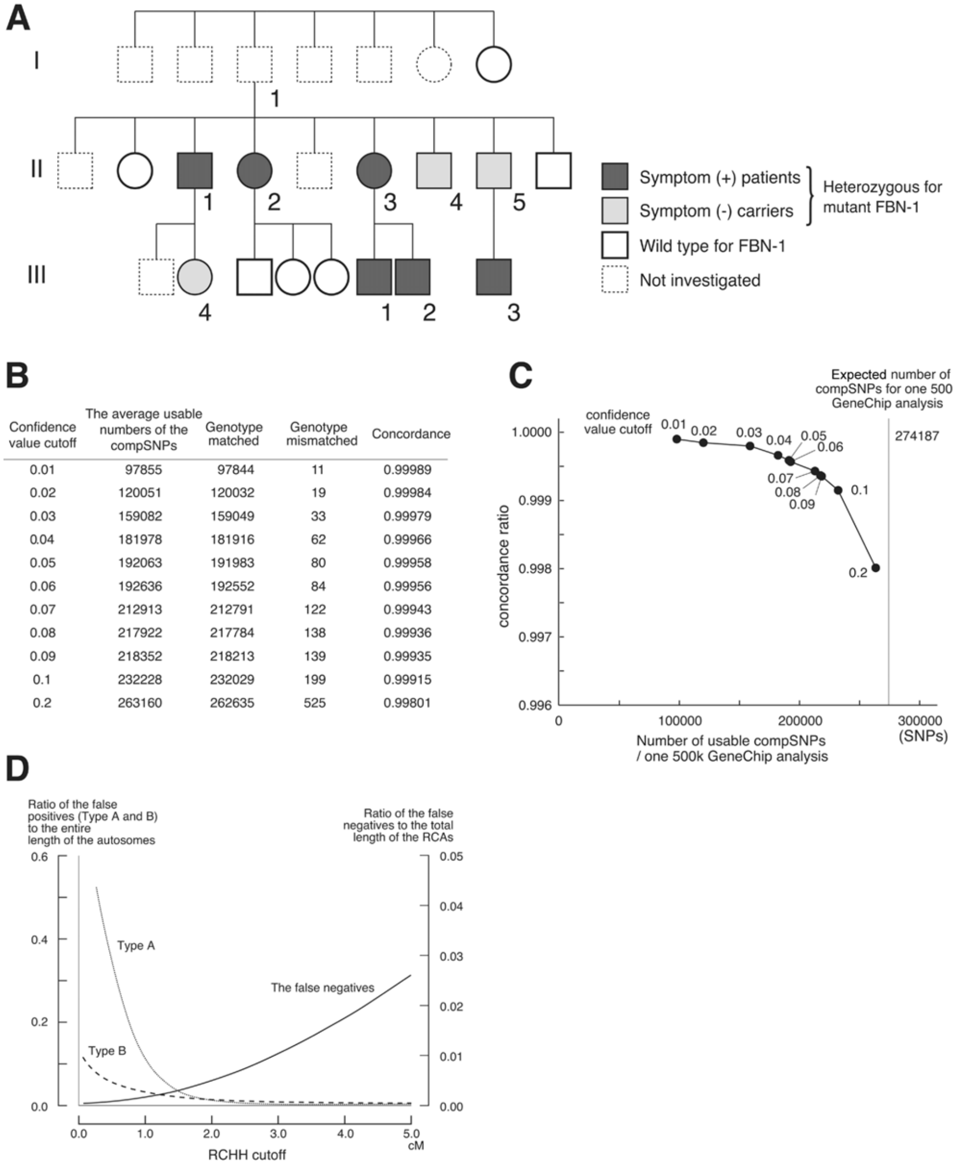
Figure 3. RCHHs, false negatives, type A false positives, and type B false positives. *A,* Detection of an RCHH. An RCHH has a genetic length longer than the cutoff value. *B,* Relationship of an RCHH and an RCA. The RCAs are overlaid and shown by dark gray boxes, and the RCHHs are shown by light gray boxes. Three types of errors are defined.

Results

Analysis of a Family with Marfan Syndrome

To investigate the utility of the HH in a family analysis, we studied a family that has multiple patients with Marfan syndrome. Marfan syndrome is an autosomal dominant disease characterized by an abnormality in the connective tissue. Mutations of either the fibrillin-1 gene (*FBN-1*) on 15q21.1 or of the TGF-β type II receptor gene (*TGFBR2*) on 3p24.2 are known to be the cause of this syndrome.[12] The family had been studied, and six symptomatic members and three asymptomatic carriers had a heterozygous 1879C→T (R627C) mutation in *FBN-1* (fig. 4*A*). Subject I-1 is considered to be the common ancestor for the disease gene. The questions we posed were as follows: (i) Could the HH identify the region containing *FBN-1* by the data from six symptomatic members? and (ii) Could the HH further narrow the region with the inclusion into the analysis of three asymptomatic carriers?

Before beginning our analysis, we checked the accuracy of the genotyping data. Equation (1) indicates that a parent and a child share the same HH along the entire length of their autosomes. Therefore, the ratio of the number of the matched compSNPs to the total number of compSNPs indicates the accuracy of genotyping. We studied three pairs: II-1 and III-4, II-3 and III-1, and II-5 and III-3. In a GeneChip analysis, the genotyping result for each SNP is



Confidence value cutoff	The average usable numbers of the compSNPs	Genotype matched	Genotype mismatched	Concordance
0.01	97855	97844	11	0.99989
0.02	120051	120032	19	0.99984
0.03	159082	159049	33	0.99979
0.04	181978	181916	62	0.99966
0.05	192063	191983	80	0.99958
0.06	192636	192552	84	0.99956
0.07	212913	212791	122	0.99943
0.08	217922	217784	138	0.99936
0.09	218352	218213	139	0.99935
0.1	232228	232029	199	0.99915
0.2	263160	262635	525	0.99801

Figure 4. Accuracy of the GeneChip data. *A,* Pedigree of a family with Marfan syndrome. *B,* Relationship between the confidence-value cutoffs and the concordance ratios. *C,* Schematic presentation of the data in panel B. *D,* False negatives between two subjects in generation III and the average ratio of the type A false positives and the type B false positives to the entire length of the autosomes for all combinations of subjects were plotted for a range of RCHH cutoffs.

accompanied by a confidence value; the smaller the confidence values, the more reliable the data. The concordance ratios of the compSNPs in three parent-child pairs for a range of confidence-value cutoffs are shown in figure 4*B* and 4*C*. We chose a confidence-value cutoff of .05.

Secondly, we determined the cutoff value for defining the RCHHs (hereafter, an "RCHH cutoff"). Figure 4*D* shows the relationship of the RCHH cutoffs and three types of errors. We chose 3.0 cM because it gave small rates of type A and type B false positives with an acceptable value for the false negatives.

We then analyzed six symptomatic patients. In figure 5, we present the result stepwise. Patients II-3 and III-1 are a parent-child pair who share RCHHs over the entire length of the autosomes (fig. 5*A*). II-1 and II-2 are siblings whose RCHHs occupy 81% of the autosomes (eq. [1] predicted 75%) (fig. 5*B*). II-2 and III-1 are an aunt and nephew whose RCHHs occupy 56% of the autosomes (eq. [1] predicted 50%) (fig. 5*C*). III-2 and III-3 are first cousins whose RCHHs occupy 39% of the autosomes (eq. [1] predicted 25%) (fig. 5*D*). The RCHHs conserved among all symptomatic members contained 96% of the total length of the RCA (calculated from table A1), and they did indeed contain *FBN-1* (fig. 5*E*). The inclusion of asymptomatic carriers (II-4, II-5, and III-4) (see fig. 4*A*) did further narrow the RCHHs (fig. 5*F*). These results demonstrate that HH analysis is both efficient and intuitive for identifying the location of disease genes in a large family.

Simulation of a Multigene Disease

Each multigene disease has a specific genetic structure. Some are considered to be a collection of single-gene diseases of which the phenotypes are indistinguishable from each other. In others, several genes working together are required to produce symptoms.[13] In either case, a subgroup of patients may share a disease gene from a common ancestor.

To investigate the utility of the HH in multigene diseases, we investigated a model multigene disease (fig. 6*A*). Here, SNP *rs16823424* (the 100,000th SNP on 500k GeneChip) is the location of the disease gene. In this region, 15 patients share an RCA derived from a common ancestor who lived 10 generations ago. The genotyping data around *rs16823424* in these 15 patients were replaced with the genotyping data at the corresponding position from a specific person who, in this analysis, acts as the common ancestor (fig. 6*B*). The lengths of the replacements were taken at random from an exponential distribution (eq. [2]: $m = 10$). Therefore, comparison of two patients corresponds with the situation $m = n = 10$ in fig. 2*B*. The patient pool included these 15 subjects together with 30 unrelated subjects (fig. 6*C*). The control pool consists of 45 unrelated Japanese samples obtained from the Affymetrix Web site. Our aim was to identify the *rs16823424* region. The strategy was as follows. In step 1, we divided autosomal regions into minute regions. In step 2, using the patient pool, we identified the HH shared by the greatest number of subjects for each region (i.e., the most common HH). We then concatenated the most common HHs for each region into a virtual HH for the entire autosome. A virtual subject who has this virtual HH was named "the representative" (step 1 in fig. 6*C*). The subjects were counted who shared the RCHHs with "the representative" in both the patient pool and the control pool for each region (step 2 in fig. 6*C*). Finally, the differences between the pools were expressed by *P* values. The candidate region for the disease gene is the region that has the lowest *P* value (i.e., the greatest $-\log_{10}(P)$ value) in the entire autosome.

Before the analysis, we determined an appropriate RCHH cutoff (fig. 6*D*). Here, the false negatives were plotted for several ages of common ancestors (the ages are expressed by *m* and *n*). As the number of generations increases, the length of the RCA shortens, increasing the difficulty of its detection with increasingly high *m* and *n* values. Because an RCHH cutoff of 5 cM was considered suitable for $m = n = 10$, this value is used hereafter. The false negatives to the total length of the RCAs decreases as we include more SNPs in the analysis. This will be discussed later.

We then performed the analysis. Figure 6*E* is a densitogram of the $-\log_{10}(P)$ value; the denser the areas are, the higher the significance. The *rs16823424* region provided a $-\log_{10}(P)$ of 4.48, and was the only region with a $-\log_{10}(P) > 3.0$ (i.e., $P < .001$). The greatest $-\log_{10}(P)$ outside of the *rs16823424* region was 2.92, which provides the background of the analysis.

Next, we investigated the detection limit. For each number from 7 to 15, we constructed 100 patient pools in which that number of patients, out of 45, shared an RCA at the *rs16823424* region. When the number is <9, the background value overwhelmed the signal in most of the analyses (fig. 6*F*). Therefore, 10 of 45 patients (22%) was the detection limit of this analysis.

Detection of Multiple Targets and the Effect of Age of Common Ancestors

We next simulated a multigene disease with three different causative genes: one at *rs16823424* (the 100,000th SNP on 500k GeneChips), one at *rs4473885* (the 200,000th SNP), and one at *rs11200928* (the 300,000th SNP). The ages of the common ancestors were $m = n = 15$, 20, and 25. We generated 100 sets of 45 subjects. Subjects 1–15 had a segment replaced with that of a specific person (the common ancestor) for a length taken at random from an exponential distribution corresponding to $m = n = 15$. Subjects 16–30 had a segment replaced corresponding to $m = n = 20$. Subjects 31–45 had a segment replaced corresponding to $m = n = 25$ (fig. 7*A*). The analysis was performed with an RCHH cutoff of 5.0 cM. Figure 7*B* demonstrates the detection of three targets simultaneously. The detection limits were 10 (22%) ($m = n = 15$), 13

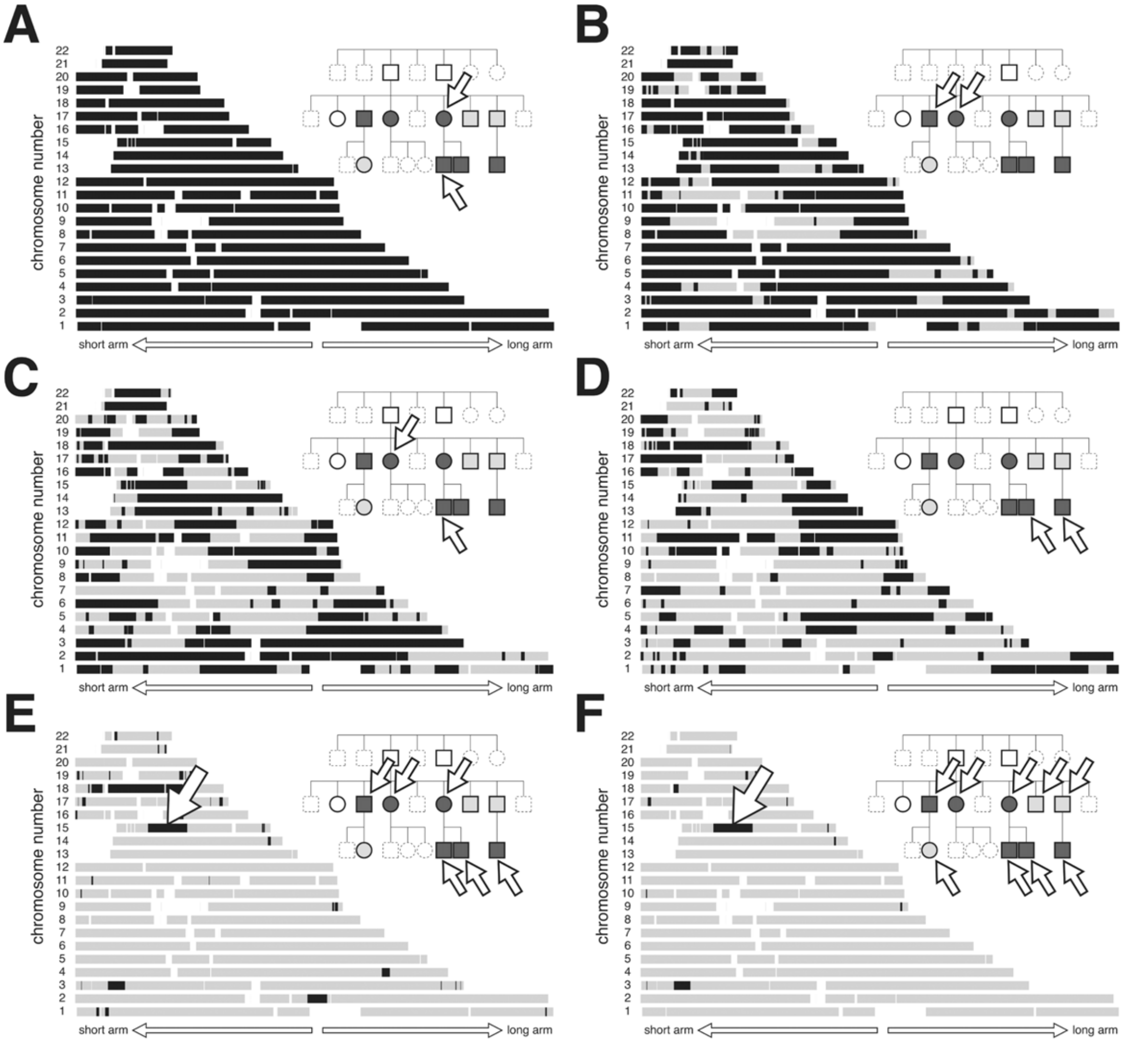


Figure 5. Family analysis. Identification of the candidate regions for the disease gene for a family with Marfan syndrome. The RCHHs are shown in black, whereas the other autosomal regions are shown in gray. *A,* The RCHHs between a parent and child, indicated by arrows, are shown. *B,* The RCHHs between siblings. *C,* The RCHHs between an aunt and nephew. *D,* The RCHHs between first cousins. *E,* The RCHHs for all symptomatic members. The *FBN-1* gene is located in the RCHH (19.6 cM in length) indicated by a large arrow. *F,* The RCHHs for all nine members (six symptomatic members and three asymptomatic carriers) who have a mutation in the *FBN-1* gene.

(29%) ($m = n = 20$) and 13 (29%) ($m = n = 25$) of 45 patients.

Discussion

In this study, we introduce HH analysis. Both copies of the homologous autosomes have the same HH and thus can be handled as if they were a single chromosome with a single HH. This enables the direct comparison of the autosomes between two individuals and thereby enables a search for a shared ancestral segment.

Because HH analysis looks for the ancestral segments, both dominant and recessive genes can be detected. The analysis is nonparametric—that is, it does not require the information from the pedigree. The patient pool contains only affected subjects, and therefore information on penetrance is not necessary. All these characteristics make the design and the interpretation of analyses simple.

Another characteristic of HH analysis is the simplicity of the algorithm, and therefore the calculation may be performed on many personal computers. The calculation for a family with Marfan syndrome that contains nine subjects (fig. 5) is completed in 6 s on our laptop computer. The analysis composed of two pools containing 45 subjects each (fig. 6*E*) took 5 min. The calculation time is proportional to the square of the number of subjects, and thus analyses with a larger number of subjects are not difficult to perform.

We used the classical Haldane's Poisson model for the calculations. In an actual situation, the crossovers do not occur randomly along the length of the autosomes. One of the major causes of the deviation from the model is

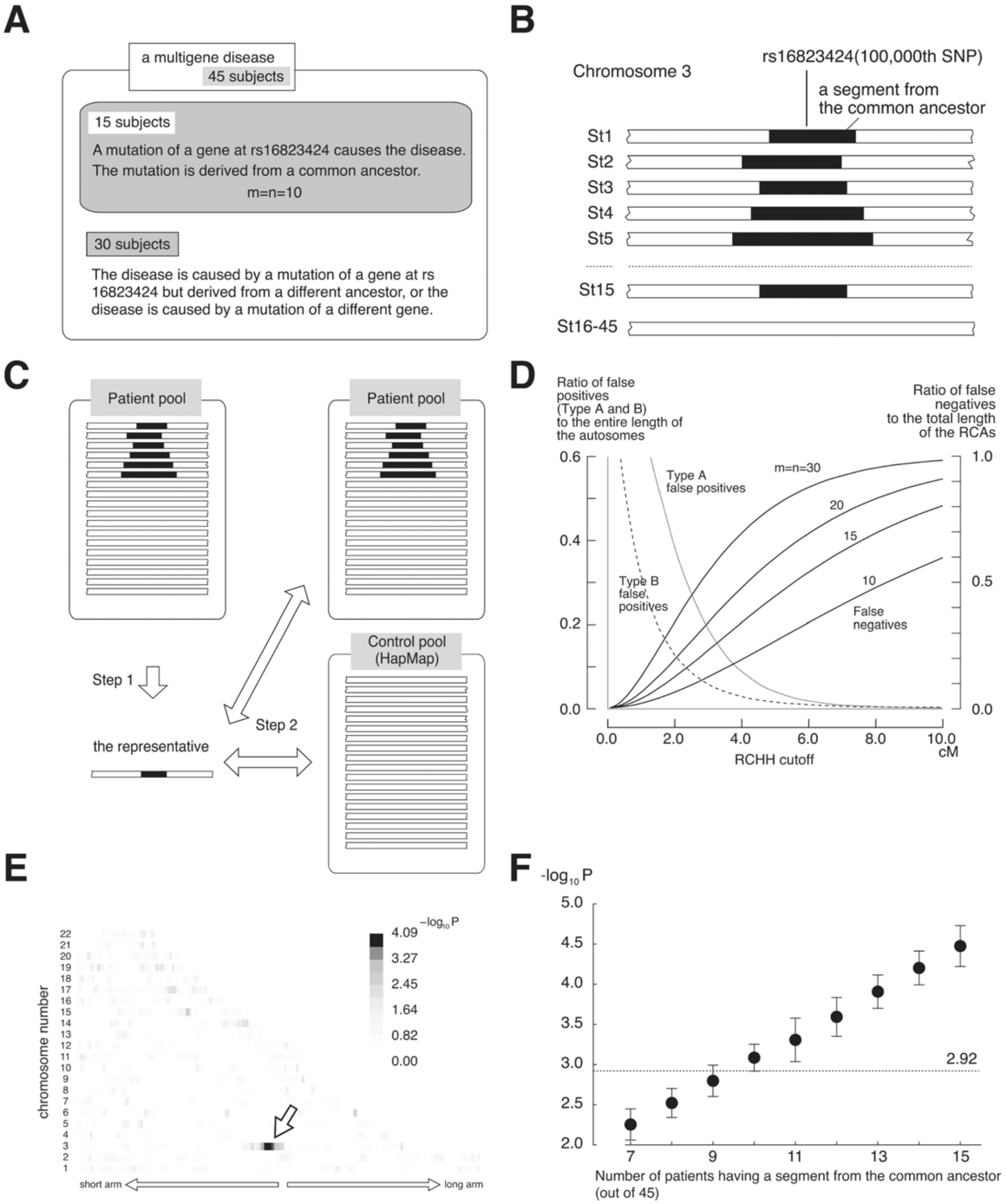


Figure 6. Simulation of a multigene disease. *A,* The structure of a model multigene disease. In 15 patients, a causative mutation of the gene at *rs16823424* is derived from a common ancestor. In the remaining 30 patients, the mutation may be from a different ancestor, or the mutations may occur in a different disease gene(s). *B,* Construction of a patient pool. Black bars indicate segments at the *rs16823424* region taken from the common ancestor. St = subject. *C,* Analysis procedure. *D,* The average ratios of the type A false positives and the type B false positives to the entire length of the autosomes for all combinations of the subjects in the control pool. The ratio of the false negatives to the total length of the RCAs for several values of *m* and *n* are simultaneously shown. *E,* Densitogram of $-\log_{10}(P)$ value for each region of the autosomes. A region that is 2.92 cM in length and contains *rs16823424* (indicated by a white arrow) gave the greatest $-\log_{10}(P)$ value, 4.09. *F,* The distribution of $-\log_{10}(P)$ from the analyses using 100 patient pools for each number of patients is shown as mean ± SD. The highest background value is 2.92 (i.e., the greatest $-\log_{10}(P)$ value outside of the *rs16823424* region).

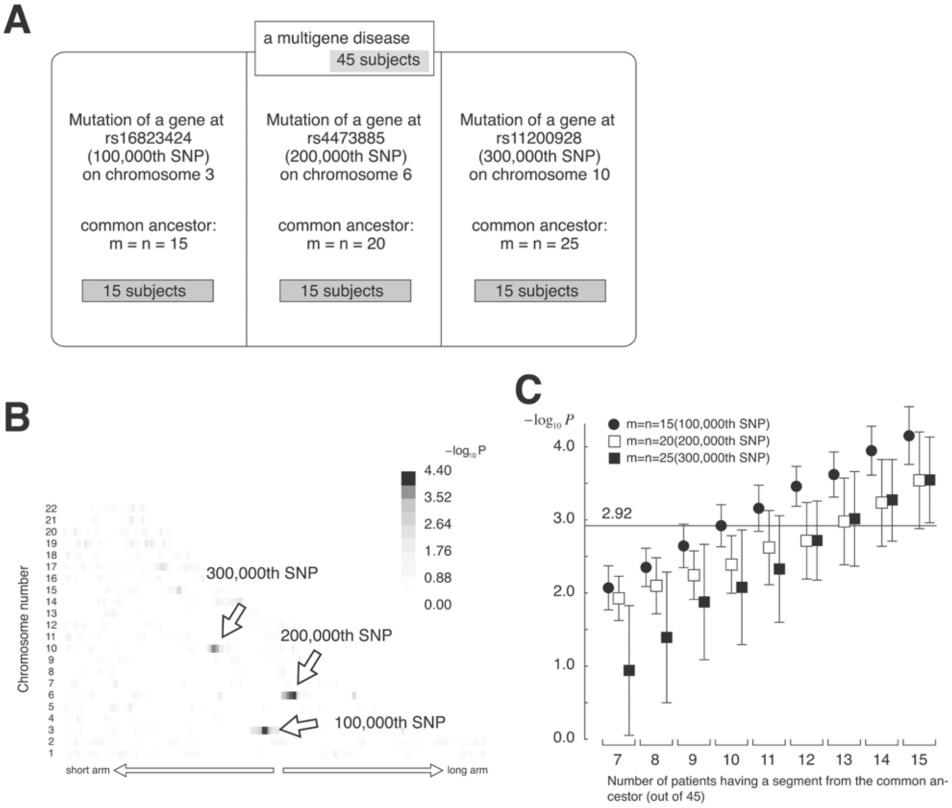


Figure 7. Multiple targets and the effect of the numbers of generations. *A,* Structure of the disease. *B,* An example of simultaneous detection of three regions. White arrows indicate the position of individual SNPs. A 2.56-cM-length region containing 100,000th SNP on chromosome 3, a 0.94-cM-length region containing the 200,000th SNP on chromosome 6 and a 2.55-cM-length containing the 300,000th SNP on chromosome 10 present the greatest $-\log_{10}(P)$ values in their neighborhood, and the values were 4.09, 4.39, and 3.49, respectively. *C,* The distribution of $-\log_{10}(P)$ from the analyses using 100 patient pools. The graph is similar to fig. 6*F,* and the data for three targets (100,000th SNP, 200,000th SNP, and 300,000th SNP) are simultaneously shown. The highest background value is 2.92.

crossover interference.[14] However, the crossover interference suppresses the production of short RCAs and favors the RCHHs in detecting true RCAs, so we made no adjustments for crossover interference in our calculation. For more-detailed discussion, see appendix A. If inbreeding exists in the pedigree, the segments from the common ancestor may be located on both copies of homologous autosomes, as in subject 2 in fig. 2*A*. This increases the average size of the RCAs and reduces the false negative rates. However, the rate of false positives may rise. The detailed information on inbreeding that occurred in previous generations is most often unknown. The practical approach to handling this is to calculate the false positives by use of the actual genotyping data, as illustrated in figure 6*D*, and to determine the RCHH cutoff. This compensates for the lack of information on inbreeding.

The numbers of SNPs used in this study were not sufficient to detect ancestral segments with an age of $m+n>30$ (fig. 6*D*). The number of type A false positives is reduced as the number of SNPs increases (fig. 8). (The rate of type B false positives is heavily dependent on the actual genotyping data and thus was not plotted.) A larger number of SNPs will allow us to use a smaller RCHH cutoff. Figure 8 suggests that the genotyping data of 1,000,000 SNPs may expand the range of analysis to $m+n>60$.

In the model multigene diseases, we used the patient pool containing 45 subjects. However, smaller numbers of subjects worked fine as well. For example, a pool of 18 subjects containing 6 subjects sharing an RCA clearly provided sufficient signal, although with a higher background (data not shown).

The four major methods for the identification of disease genes are the haplotype analysis, the linkage analysis, the sib-pair analysis and the whole-genome association studies.[15] The former two methods target single-gene diseases occurring in families (usually $m+n<6$), whereas the latter two methods target both single-gene and multigene diseases occurring in the general population. When $m+n<3$, HH analysis does not work well. In fact, the HH cannot distinguish parents from children, as shown in figure 5*A*, whereas haplotype analysis or linkage analysis may be able to identify a disease gene from pedigrees com-

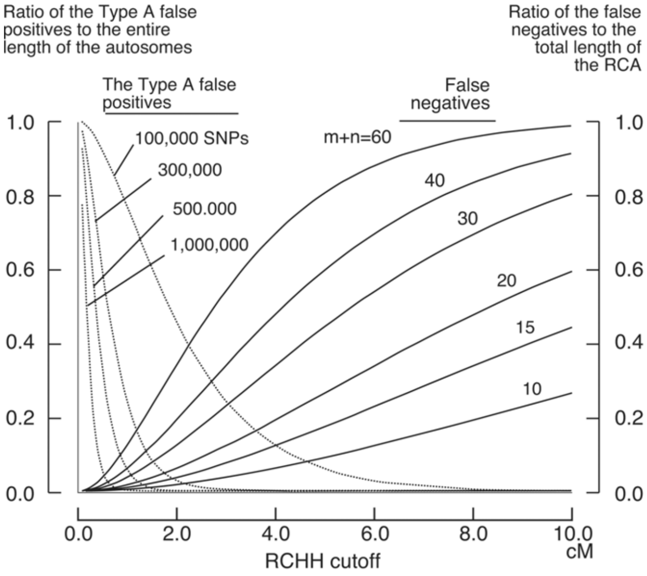


Figure 8. The effect of the numbers of SNPs used and the target generations of the analysis. The type A false positives were plotted for a range of numbers of the SNPs genotyped. The false negatives were plotted for a range of $m + n$ values.

posed of only two generations.[16] For families containing subjects with $m + n \geqslant 3$, HH analysis works well, as shown in figure 5. HH analysis may provide an advantage when $6 \leqslant m + n \leqslant 50$ where the haplotype analysis or the linkage analysis are difficult to perform. HH analysis is considered applicable to sib-pair analysis, where one sib pair provides 3/4 of the entire autosomes as shared regions (see eq. [1]). One attractive application may be for affected-relative-pair analyses.[17] Equation (1) indicates that one second-cousin pair may narrow the candidate autosomal region to 1/16 of the entire length of the autosomes, and three second-cousin pairs may narrow it further to $(1/16)^3 = 1/4{,}096$.

The simulation results presented in this study suggest that HH analysis demonstrates advantages and may complement whole-genome association studies by detecting genes for common diseases in the following situations. (1) The target population is genetically isolated. (2) The relative risk of the disease gene is moderate to high, and thus the frequencies of the disease-associated HH are expected to be significantly different between the patient pool and the control pool. (3) The common ancestors who brought the disease gene into the population are assumed to have existed within the last several hundred years, thus enabling the detection of the RCAs as RCHHs. (4) The number of the common ancestors who brought the disease gene was small, which limits the number of the disease-associated HHs in the population, and thus the frequencies of some of them may exceed the detection limit shown in figures 6*F* and 7*C*. When these conditions are met, the inclusion of only a few dozen patients may be required to identify the location of the disease gene (figs. 6 and 7). However, the identified regions may be 1–3 cM in length and require more-detailed investigation. Ethnically, geographically, or culturally isolated populations may fulfill these requirements for many diseases. For example, consider the French-Canadian population in Quebec.[18,19] It is known that two-thirds of the genetic pool of the current population of 6 million people is derived from only 2,600 settlers who arrived during the 17th century (i.e., $m = n = 20$, given 20 years per generation). The causative genes for diseases with an incidence of $\leqslant 0.01$ may be derived from only one or two dozen common ancestors. In other words, the number of the disease-associated HH was limited to one or two dozen because of the bottleneck effect caused by the immigration. If random genetic drift is taken into consideration, some of the disease-associated HH may exceed the detection limit of 29% in the patient pool (fig. 7*B*). Therefore, moderate- to high-risk genes for diseases with an incidence of $\leqslant 0.01$ in Quebec fulfills all four requirements and is therefore worth studying by HH analysis, whether the gene is for a single-gene disease or a multigene disease. If any one of the four conditions is not met, HH analysis should not be considered a good choice. The selection of the target population and the target disease are crucial.

In this study, we describe the introduction of the HH and its applications. The HH is easy to obtain and the results are intuitive. Although modern society promotes the movement of people, many countries have a history in which the transfer of people was politically or geographically limited. Patients with a specific disease clustered in a geographical region therefore may inherit a common ancestral disease gene. In such regions, HH analysis may provide a distinct benefit. We believe that HH analysis will therefore facilitate the identification of disease genes both for single-gene and multigene diseases.

Acknowledgments

We thank Prof. Hideshi Kawakami, Hiroshima University, Japan, for his critical review of the manuscript, and Ms. Yukiko Yatsuka, for her valuable technical assistance. This work was supported in part by the following grants-in-aid: (1) for scientific research (No. 18390242) from the Japan Society for the Promotion of Science, (2) for Comprehensive Research on Aging and Health from the Ministry of Health, Labor, and Welfare, Japan, and (3) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, in particular by a Ministry Grant to the Saitama Medical University Research Center for Genomic Medicine.

Appendix A

Deduction of Equation (1)

The calculation to obtain RCA(1,1) is presented as an example (fig. A1). A and B are the common ancestors. m1–

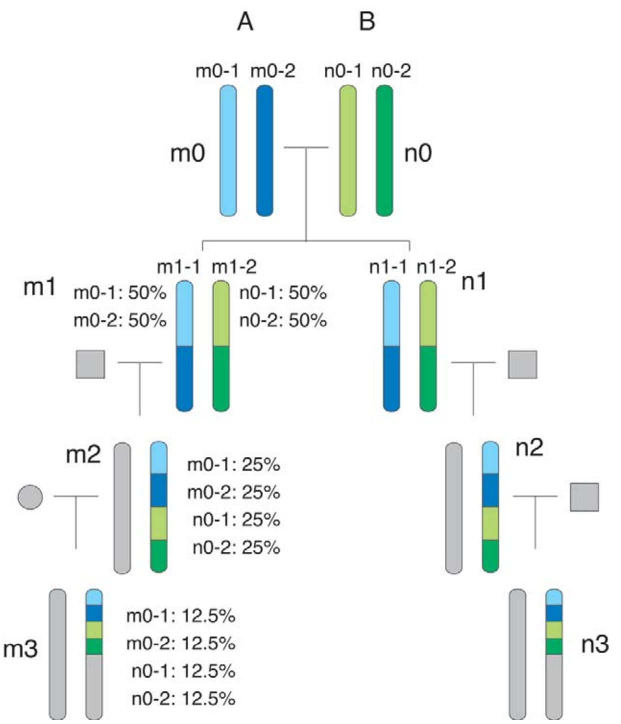


Figure A1. Deduction of equation (1). Gray circles and boxes indicate spouses. Gray areas of the chromosomes come from spouses and do not contain segments from the common ancestors (i.e., A and B).

1 and m1–2 are two copies of homologous chromosomes for subject m1, and n1–1 and n1–2 are two copies of homologous chromosomes for subject n1. One half of m1–1 is from m0–1, and the other half is from m0–2. One half of n1–1 is from n0–1, and the other half is from n0–2. Between subjects m1 and n1, the ratio of the RCA to the entire chromosome is calculated as the probability that m1 and n1 share the same chromosomal segment at a specific position on chromosomes. It can be obtained by subtracting from 1 the probability that m1–1, m1–2, n1–1, and n1–2 all have segments derived from different chromosomes. Therefore,

$$\mathrm{RCA}(1,1) = 1 - \frac{1}{2} \times \frac{1}{2} = \frac{3}{4} \ .$$

RCA(m,n) for other values of m and n were similarly obtained and summarized in equation (1).

Calculation of False Negatives, Type A False Positives, and Type B False Positives

Step 1: Ratio of False Negatives to Total Length of RCAs ($R_{\text{false negatives}}$).—According to the Haldane's Poisson model, the length (x, in centimorgans) of the chromosomal segment derived from an ancestral chromosome in generation m (see fig. 2*B*) has an exponential distribution that has the probability density function

$$f(x) = \lambda e^{-\lambda x} \ ,$$
$$\lambda = \frac{m}{100} \ . \qquad \text{(A1)}$$

First, the union of ancestral chromosomal segments on two homologous chromosomes are taken for each subject. Next, the RCAs are the intersections of these unions between the two subjects. From equation A1, when $m+n$ (see fig. 2*B*) is large enough, $R_{\text{False negatives}}$ for an RCHH cutoff c is approximated by

$$R_{\text{False negatives}} \approx \frac{\int_0^c xf(x)dx}{\int_0^\infty xf(x)dx}$$
$$= 1 - e^{-\lambda c}(1+\lambda c) \ , \qquad \text{(A2)}$$

where

$$f(x) = \lambda e^{-\lambda x}$$
$$\lambda = \frac{m+n}{100} \ .$$

However, when $m+n$ is small, the $R_{\text{False negatives}}$ deviates from the value calculated by equation (A2). We therefore obtained $R_{\text{False negatives}}$ for small values of $m+n$ by the Monte Carlo method, with use of 100,000 pedigrees (table A1). We found that equation (A2) provides good approximations when $m+n>12$ (see table A1; compare the values for $m+n=12$).

Step 2: Ratio of the Type A False Positives to the Entire Autosome ($R_{\text{Type A false positives}}$).—Given that N_{SNP} is the total number of SNPs on a genotyping chip, and P_n and Q_n are the frequencies of the major and minor alleles for the nth SNP, respectively, the average frequencies of the major alleles ($\bar{F}_{\text{major allele}}$) and the minor alleles ($\bar{F}_{\text{minor allele}}$) are

$$\bar{F}_{\text{major allele}} = \frac{\sum_{n=1}^{N_{\text{SNP}}} P_n}{N_{\text{SNP}}}$$

and

$$\bar{F}_{\text{minor allele}} = \frac{\sum_{n=1}^{N_{\text{SNP}}} Q_n}{N_{\text{SNP}}} \ ,$$

respectively. The number of mismatched compSNPs ($N_{\text{mismatched compSNP}}$) is approximated by

$$N_{\text{mismatched compSNP}} \approx \frac{2\left(\bar{F}_{\text{major allele}}\right)^2\left(\bar{F}_{\text{minor allele}}\right)^2 N_{\text{Pt1}}N_{\text{Pt2}}}{N_{SNP}} \ ,$$

where N_{Pt1} and N_{Pt2} are the numbers of SNPs successfully genotyped for Pt1 and Pt2, respectively. $N_{\text{mismatched compSNP}}$ is not a large number. For example, with use of the 500k GeneChips from Affymetrix, $N_{\text{mismatched compSNP}}$ is 22,000 at maximum, spaced at 0.16 cM on average. This spacing is larger in size than most of the haplotype blocks and thus is assumed to be randomly distributed over the entire autosome. The length between two mismatched compSNPs is considered to have an exponential distribution with a density probability function of

$$f(x) = \lambda e^{-\lambda x} \ ,$$

$$\lambda = \frac{N_{\text{mismatched compSNP}}}{L_{\text{autosome}}} \ ,$$

where L_{autosome} is the entire genetic length of the autosomes. Therefore, for the cutoff value *c,*

$$R_{\text{Type A false positives}} = \frac{\int_c^\infty xf(x)dx}{\int_0^\infty xf(x)dx} = (1+\lambda c)e^{-\lambda c} \ .$$

Step 3: Ratio of the Type B False Positives to the Entire Length of the Autosomes ($R_{\text{Type B false positives}}$).—An RCHH containing an RCA is expected to have the type B false positives with a length of $\frac{\text{cut off value}}{2}$ on each end. It is impossible to distinguish RCHHs that contain the RCAs from those that do not (i.e., the type A false positives). We calculated $R_{\text{Type B false positives}}$ under the assumption that every RCHH contains an RCA. Therefore, the $R_{\text{Type B false positives}}$ calculation results in an overestimation, which we consider to be more appropriate than an underestimation when the appropriate RCHH cutoff is being determined.

The Representative

The easiest way to compare the patient pool and the control pool is to directly compare the number of patients sharing the RCHHs at the given position (fig. A2*A*). This algorithm usually works fine, but actually this reduces the sensitivity. Assume that, at a specific position, the patient pool has 4 subjects sharing HH1 and has 0 subject sharing HH2. The control pool has 0 subject sharing HH1 and 4 subjects sharing HH2. Although two pools are different in their frequency of HH1, the algorithm shown in figure A2*A* does not detect it.

One of the ways to solve this problem is to have a representative, as shown in fig. A2*B,* as we did in this study. For the actual algorithm, please see the program source code. This algorithm may have difficulty picking up the most common HHs in a region where there is no dominant HH but only many kinds of HHs with low frequencies, which we think does not cause any major problems. We have also provided the source code for an alternative algorithm. The source may be modified according to your uses.

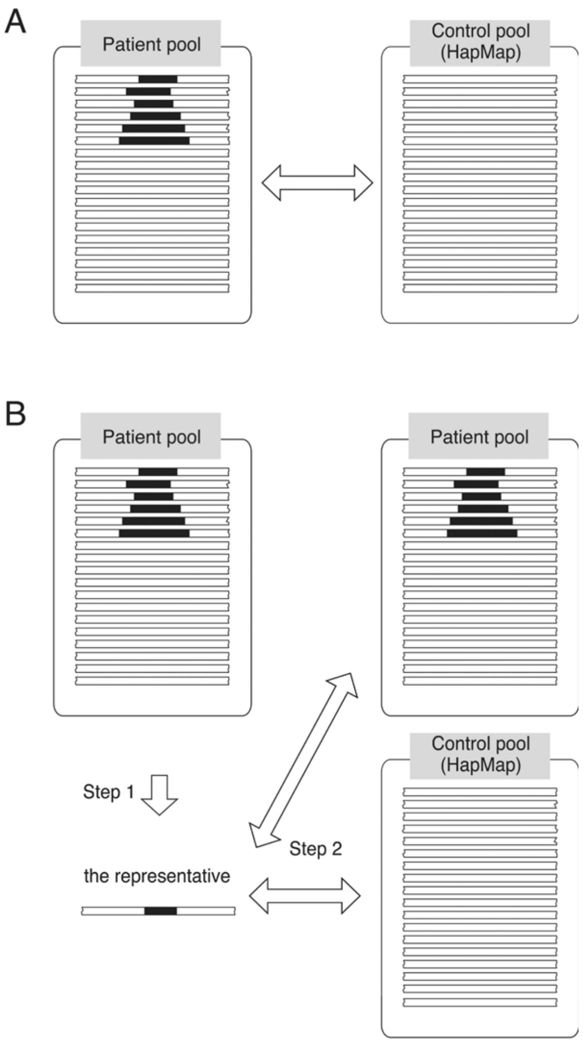


Figure A2. Two strategies for comparing the patient pool with the control pool.

Crossover Interference and the Size of the RCAs

Crossover interference increases the average size of the RCA, and favors the RCHHs in detecting the RCA, which reduces the false negatives. This results in a better performance in HH analysis. Figure A3 shows an RCA in one generation. The size of RCA may be reduced in size in the next generation. The reduction occurs by two processes: (1) crossover occurs in one or both subjects and (2) mul-

Table A1. Ratios of False Negatives to the Total Length of the RCAs for a Range of RCHH Cutoffs

Calculation Method and Variable(s)			Ratio of False Negatives to Total Length of the RCAs, by RCHH Cutoff (in cM)																													
			.2	.4	.6	.8	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
Monte Carlo																																
m+*n*	*m*	*n*																														
2	2	0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.001	.001	.001	.001	.001	.001	.002	.002	.002	.002	.003	.003	.003	.003	.004	.004	.004	.004	.005	.005
	1	1	.000	.000	.000	.000	.000	.000	.001	.001	.001	.001	.001	.001	.002	.002	.002	.003	.003	.003	.004	.004	.004	.005	.005	.005	.006	.006	.007	.007	.008	.008
3	3	0	.000	.000	.000	.000	.000	.001	.001	.001	.001	.002	.002	.002	.003	.003	.003	.004	.004	.005	.005	.006	.007	.007	.008	.008	.009	.010	.011	.012	.012	.013
	2	1	.000	.000	.000	.000	.001	.001	.001	.001	.002	.002	.003	.003	.004	.004	.005	.005	.006	.007	.008	.008	.009	.010	.011	.012	.013	.014	.015	.016	.017	.018
4	4	0	.000	.000	.000	.000	.001	.001	.001	.002	.002	.003	.003	.004	.005	.005	.006	.007	.008	.009	.010	.011	.012	.013	.014	.015	.017	.018	.019	.021	.022	.024
	3	1	.000	.000	.000	.001	.001	.001	.002	.002	.003	.004	.004	.005	.006	.007	.008	.009	.010	.011	.013	.014	.015	.017	.018	.019	.021	.022	.024	.026	.028	.029
	2	2	.000	.000	.000	.001	.001	.002	.002	.003	.004	.004	.005	.006	.007	.009	.010	.011	.013	.014	.015	.017	.019	.020	.022	.024	.026	.028	.030	.032	.034	.036
5	5	0	.000	.000	.000	.001	.001	.002	.002	.003	.004	.004	.005	.006	.007	.008	.010	.011	.012	.014	.015	.017	.019	.020	.022	.024	.026	.028	.030	.032	.034	.037
	4	1	.000	.000	.001	.001	.001	.002	.003	.004	.004	.005	.007	.008	.009	.010	.012	.013	.015	.017	.019	.020	.022	.025	.027	.029	.031	.034	.036	.039	.041	.044
	3	2	.000	.000	.001	.001	.002	.002	.003	.004	.005	.006	.008	.009	.011	.012	.014	.016	.018	.020	.022	.024	.026	.029	.031	.034	.036	.039	.042	.044	.047	.050
6	6	0	.000	.000	.001	.001	.002	.002	.003	.004	.005	.006	.008	.009	.010	.012	.014	.016	.018	.020	.022	.024	.026	.029	.031	.034	.036	.039	.041	.044	.047	.050
	5	1	.000	.000	.001	.001	.002	.003	.004	.005	.006	.007	.009	.011	.013	.014	.016	.018	.020	.023	.025	.028	.030	.033	.036	.039	.042	.045	.049	.052	.056	.059
	4	2	.000	.000	.001	.001	.002	.003	.004	.006	.007	.009	.010	.012	.014	.016	.018	.021	.023	.026	.029	.032	.035	.038	.041	.044	.048	.051	.055	.059	.062	.066
	3	3	.000	.000	.001	.001	.002	.003	.004	.006	.007	.009	.010	.012	.014	.016	.019	.021	.024	.026	.029	.032	.035	.038	.042	.045	.048	.052	.055	.059	.063	.066
7	6	1	.000	.000	.001	.002	.003	.004	.005	.006	.008	.010	.011	.013	.016	.018	.021	.023	.026	.029	.033	.036	.040	.043	.046	.050	.054	.058	.063	.067	.071	.075
	5	2	.000	.000	.001	.002	.003	.004	.006	.007	.009	.011	.013	.016	.018	.021	.024	.027	.030	.034	.037	.041	.044	.048	.052	.056	.061	.065	.070	.075	.080	.084
	4	3	.000	.000	.001	.002	.003	.004	.006	.007	.009	.011	.013	.016	.018	.021	.024	.026	.030	.033	.037	.040	.044	.048	.052	.056	.061	.065	.069	.074	.078	.083
8	6	2	.000	.001	.001	.003	.004	.005	.007	.009	.012	.014	.017	.020	.023	.026	.029	.033	.037	.041	.046	.050	.055	.060	.064	.069	.075	.081	.086	.092	.097	.103
	5	3	.000	.001	.001	.002	.004	.005	.007	.009	.011	.014	.017	.019	.022	.026	.029	.032	.036	.040	.045	.050	.054	.059	.063	.069	.074	.080	.085	.091	.096	.101
	4	4	.000	.001	.001	.002	.004	.005	.007	.009	.011	.014	.017	.019	.022	.026	.029	.032	.036	.040	.045	.049	.054	.059	.063	.068	.073	.079	.084	.089	.094	.099
9	6	3	.000	.001	.002	.003	.005	.007	.009	.012	.015	.017	.020	.024	.027	.031	.035	.039	.044	.049	.055	.061	.067	.073	.079	.085	.092	.098	.104	.111	.117	.124
	5	4	.000	.001	.002	.003	.004	.006	.009	.011	.013	.016	.019	.022	.026	.030	.034	.038	.043	.047	.052	.057	.063	.069	.074	.081	.088	.094	.100	.105	.111	.118
10	6	4	.000	.001	.002	.004	.006	.008	.011	.014	.017	.021	.024	.028	.032	.036	.041	.046	.051	.056	.062	.068	.074	.081	.090	.097	.104	.111	.119	.126	.134	.142
	5	5	.000	.001	.002	.003	.005	.007	.011	.014	.016	.019	.023	.027	.031	.035	.040	.045	.051	.056	.062	.068	.074	.081	.088	.095	.102	.110	.117	.124	.131	.138
11	6	5	.000	.001	.002	.004	.007	.009	.013	.016	.019	.023	.027	.032	.037	.042	.047	.053	.060	.066	.072	.080	.089	.096	.103	.110	.117	.125	.133	.141	.150	.158
12	6	6	.000	.001	.003	.005	.008	.010	.014	.019	.023	.027	.032	.037	.042	.049	.056	.063	.069	.077	.084	.092	.101	.111	.120	.128	.137	.145	.152	.159	.166	.173
Eq. (A1)																																
m+*n*	*m*	*n*																														
12			.000	.001	.002	.004	.007	.009	.013	.016	.020	.025	.029	.034	.040	.045	.051	.057	.064	.070	.077	.084	.091	.099	.106	.114	.122	.130	.138	.146	.154	.163
13			.000	.001	.003	.005	.008	.011	.015	.019	.023	.028	.034	.040	.046	.052	.059	.066	.073	.081	.088	.096	.104	.113	.121	.130	.139	.148	.157	.166	.175	.184
14			.000	.002	.003	.006	.009	.013	.017	.022	.027	.033	.039	.045	.052	.059	.067	.075	.083	.091	.100	.109	.118	.127	.137	.146	.156	.166	.175	.185	.196	.206
15			.000	.002	.004	.007	.010	.014	.019	.025	.031	.037	.044	.051	.059	.067	.075	.084	.093	.103	.112	.122	.132	.142	.152	.163	.173	.184	.195	.206	.217	.228
16			.000	.002	.004	.008	.012	.016	.022	.028	.034	.041	.049	.057	.066	.075	.084	.094	.104	.114	.125	.135	.146	.157	.168	.180	.191	.203	.214	.226	.238	.250
17			.001	.002	.005	.008	.013	.018	.024	.031	.038	.046	.055	.064	.073	.083	.093	.104	.115	.126	.137	.149	.161	.173	.185	.197	.209	.222	.234	.247	.259	.272
18			.001	.002	.005	.009	.014	.020	.027	.034	.042	.051	.060	.070	.081	.091	.103	.114	.126	.138	.150	.163	.175	.188	.201	.214	.228	.241	.254	.267	.280	.294
19			.001	.003	.006	.010	.016	.022	.030	.038	.047	.056	.066	.077	.088	.100	.112	.125	.137	.150	.163	.177	.190	.204	.218	.232	.246	.260	.274	.288	.302	.316
20			.001	.003	.007	.012	.018	.025	.033	.041	.051	.062	.073	.084	.096	.109	.122	.135	.149	.163	.177	.191	.206	.220	.235	.250	.264	.279	.294	.308	.323	.337
25			.001	.005	.010	.018	.027	.037	.049	.062	.075	.090	.106	.122	.139	.156	.173	.191	.209	.228	.246	.264	.283	.301	.319	.337	.355	.373	.391	.408	.425	.442
30			.002	.007	.014	.025	.037	.051	.067	.084	.103	.122	.142	.163	.184	.206	.228	.250	.272	.294	.316	.337	.359	.380	.401	.422	.442	.462	.482	.501	.519	.537
35			.002	.009	.019	.033	.049	.067	.087	.109	.132	.156	.180	.206	.231	.257	.283	.308	.334	.359	.384	.408	.432	.455	.478	.501	.522	.543	.563	.583	.602	.620
40			.003	.012	.025	.041	.062	.084	.109	.135	.163	.191	.220	.250	.279	.308	.337	.366	.394	.422	.449	.475	.501	.525	.549	.572	.594	.615	.636	.655	.674	.692
50			.005	.018	.037	.062	.090	.122	.156	.191	.228	.264	.301	.337	.373	.408	.442	.475	.507	.537	.566	.594	.620	.645	.669	.692	.713	.733	.751	.769	.785	.801
60			.007	.025	.051	.084	.122	.163	.206	.250	.294	.337	.380	.422	.462	.501	.537	.572	.605	.636	.665	.692	.717	.740	.762	.782	.801	.818	.834	.849	.862	.874

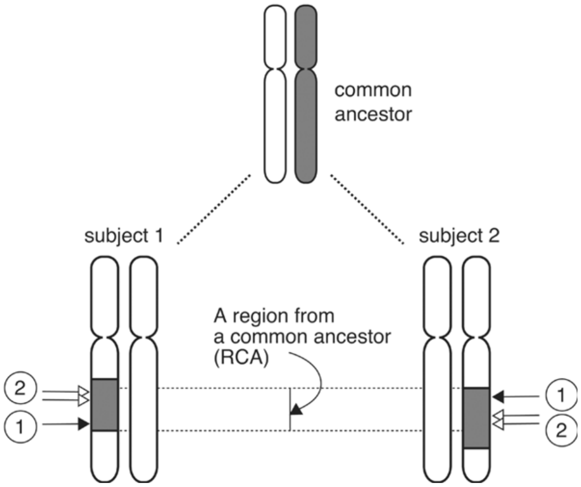


Figure A3. Effects of crossover interference. Process numbers are enclosed in the circle. Processes 1 and 2 both reduce the length of the RCAs in the next generation. Process 1 is independent of and process 2 is dependent on the crossover interference. *Gray area,* shared segments derived from a common ancestor.

tiple crossovers occur in one or both subjects. Although occurrence of process 2 may be suppressed by crossover interference, process 1 is independent of the interference and is not suppressed. Moreover, as the size of the shared segments from the common ancestor (shown in gray in fig. A3) shortens over generations, multiple crossovers in a single RCA become less frequent, even without crossover interference, and process 1 becomes the main determinant of the size of the RCAs. Therefore, crossover interference has a limited effect on HH analysis, and so we chose not to make any adjustment in the algorithm.

Web Resources

URLs for data presented herein are as follows:

Affymetrix, http://www.affymetrix.com/index.affx
GCC, the GNU Compiler Collection, http://gcc.gnu.org/
K.H.'s Web site, http://homepage.mac.com/hagiwark/FileSharing1.html (for HH analysis program)
International HapMap Project, http://www.hapmap.org/
National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
Saitama Medical University, http://www.saitama-med.ac.jp/genome/TR/hh_analysis_program_and_manual.zip (for HH analysis program and manual)

References

1. Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810
2. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318
3. Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454
4. International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
5. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596–601
6. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
7. Gillanders EM, Pearson JV, Sorant JM, Trent JM, O'Connell JR, Bailey-Wilson JE (2006) The value of molecular haplotypes in a family-based linkage study. Am J Hum Genet 79:458–468
8. Amos CI, Dawson DV, Elston RC (1990) The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. Am J Hum Genet 47:842–853
9. Zhang K, Zhao H (2006) A comparison of several methods for haplotype frequency estimation and haplotype reconstruction for tightly linked markers from general pedigrees. Genet Epidemiol 30:423–437
10. Haldane JBS (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8:299–309
11. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247
12. Hayward C, Brock DJ (1997) Fibrillin-1 mutations in Marfan syndrome and other type-1 fibrillinopathies. Hum Mutat 10: 415–423
13. Pritchard DJ, Korf BR (2003) Medical genetics at a glance. Blackwell Publishing, Birmingham, United Kingdom
14. Sturtevant AH (1915) The behavior of chromosomes as studied through linkage. Z Indukt Abstammungs-Vererbungsl 13: 234–287
15. Strachan T, Read A (2003) Human molecular genetics. Garland Science/Taylor & Francis Group, Oxfordshire, United Kingdom
16. Shore EM, Xu M, Feldman GJ, Fenstermacher DA, Brown MA, Kaplan FS (2006) A recurrent mutation in the BMP type I receptor ACVR1 causes inherited and sporadic fibrodysplasia ossificans progressiva. Nat Genet 38:525–527
17. Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241
18. Heyer E, Tremblay M (1995) Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet 56:970–978
19. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, Mitchell GA (2005) Population history and its impact on medical genetics in Quebec. Clin Genet 68:287–301