Genome-Wide Identity-by-Descent Sharing Among CEPH Siblings

Alain Gagnon,[1]* Jan Beise,[2] and J.W. Vaupel[2]

[1]*Aging and Health Research Centre, Population Studies Centre, University of Western Ontario, London, Ontario, Canada*
[2]*Max Planck Institute for Demographic Research, Rostock, Germany*

The concept of genetic identity-by-descent (IBD) has markedly advanced our understanding of the genetic similarity among relatives and triggered a number of developments in epidemiological genetics. However, no empirical measure of this relatedness throughout the whole human genome has yet been published. Analyzing highly polymorphic genetic variations from the *Centre d'études du polymorphisme humain* (CEPH) database, we report the first genome-wide estimation of the mean and variation in IBD sharing among siblings. From 1,522 microsatellite markers spaced at an average of 2.3 cM on 498 sibling pairs, we estimated a mean of 0.4994 and a standard deviation of 0.0395. In order to account for the impact of varying chromosomal lengths and recombination rates, the analysis was also performed at the chromosomal and marker levels and for paternal and maternal DNA separately. Based on the variation, we estimate an ''effective number of segregating loci'' of around 80 for sibling pairs over the whole genome (i.e., the number of loci that would yield the same standard deviation in IBD sharing if all loci were segregating *independently*). Finally, we briefly assess the impact of genotyping errors on IBD estimations, compare our results to published theoretical and simulated expectations, and discuss some implications of our findings. *Genet. Epidemiol.* 2005. © 2005 Wiley-Liss, Inc.

Key words: microsatellites; genomic maps; recombination rate; effective number of loci; typing errors

Contract grant sponsor: Max Planck Society.
*Correspondence to: Alain Gagnon, Department of Sociology, Room 5326, Social Sciences Centre, University of Western Ontario, London, Ontario, Canada, N6A 5C2. E-mail: agagnon4@uwo.ca
Received 18 February 2005; Accepted 12 June 2005
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20090

INTRODUCTION

Finding genomic similarities among related individuals remains one of the fundamental challenges in human genetics. The identity-by-descent (IBD) approach provides a natural way to assess such similarities. When two genes at a given locus are inherited from a common ancestor, they are said to be identical-by-descent. While parent and offspring *exactly* share 50% of their genes on autosomal loci, two siblings share the same proportion *on average*. At any locus, they may share *no alleles* (if both parents transmitted different alleles), *one allele* (if one parent transmitted the same allele to each offspring), or *2 alleles* (if each parent transmitted the same allele to each offspring). As the contribution from the male and the female parent are independent, the respective probabilities of these events are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. The locus-specific expectation of IBD in sib-pairs is thus $\frac{1}{2}$ and the standard deviation around this value is $\sqrt{1/8}$.

Independently introduced by Cotterman [1940] and Malécot [1941], the concept of IBD lies at the very heart of many epidemiological and mapping studies in genetics. It has triggered the development of quantitative trait loci (QTL) mapping [Haseman and Elston, 1972; Wang and Elston, 2005], as well as other methods, such as linkage analyses based on allele sharing [Suarez et al., 1978] and the multipoint interval mapping [Fulker et al., 1995; Rijsdijk and Sham, 2002]. The concept has also generated useful theoretical work with respect to the genetic identity among relatives. Suarez et al. [1979] used simulations to estimate the variability in sib-pairs genetic identity, while Risch and Lange [1979a,b] derived a theoretical expectation of this variability based on the probability generating function. Using the ''recombination index,'' Rasmuson [1993] derived the variability within kinships involving different kinds of relatives. She also addressed the relevance of IBD sharing in sociobiology, and speculated that kin recognition patterns may

require the assessment of genetic identity over a large number of loci. Recent developments in forensic analyses using DNA, which mostly base the identification of relatives through identity-by-state sharing (see for instance [Leclair et al., 2004; Presciuttini et al., 2002]), can also benefit from IBD assessments in the determination of population substructure through coancestry coefficients [Weir, 1994]. The most comprehensive theoretical studies on IBD sharing are probably those of Guo [1994, 1995, 1996]. Based on the work of K.P. Donnelly [1983], who has modeled each chromosome in an offspring as a two-state Markov chain, Guo has provided an extension of the concept to a group of relatives, as well as valuable computation methods and various applications.

Though many have addressed the theoretical distribution of IBD sharing over the whole genome, no empirical study on the subject has yet been published. This is understandable in light of technical difficulties and the massive scale of the genome, which prohibited any such accomplishment until the late 1990s. Recent advances in biotechnology have changed that situation. Two detailed comprehensive genetic maps have been published based on the CEPH and deCode families [Broman et al., 1998; Kong et al., 2002]. The development of single-nucleotide polymorphisms (SNP) linkage maps [Matise et al., 2003] and whole-genome scan of complex diseases based on these polymorphisms [John et al., 2004; Middleton et al., 2004; Pato et al., 2005] is now possible.

Using data on microsatellite markers from the eight ''CEPH reference families,'' we present the first genome-wide estimation of IBD sharing among pairs of siblings. We provide estimations for the mean and the standard deviation of this variable at different levels of observation (genome, chromosome, and marker levels), and compare these empirical estimations to published theoretical expectations and simulations. Furthermore, as several studies have shown that the recombination rate is higher in female gametes than in male gametes [Broman et al., 1998; Jensen-Seaman et al., 2004; Kong et al., 2002; Matise et al., 2003; Yu et al., 2001], all analyses were performed for maternal and paternal DNA separately, as well as for combined (maternal+paternal) DNA. Finally, we briefly assessed the impact of genotyping errors on IBD estimations from simulations of different rates of errors. Our aim is to furnish a baseline account of the variation in a dataset for which, a priori, there is no reason to suspect any deviation from random proportions.

MATERIALS AND METHODS

IBD ESTIMATIONS

We analyzed the proportion of alleles shared by siblings at highly polymorphic microsatellite loci of the 22 autosomal chromosomes among the eight large reference families of the CEPH. We focused on the ''core'' CEPH families because the genotypes of these families have been extensively studied and are thus less likely to contain typing errors. The average number of children in the pedigrees is 11.75, and the total number of sib-pairs 517. In sibships larger than two ($s > 2$) as in the CEPH families, only $(s-1)$ of the $s(s-1)/2$ sib-pairs may be considered as independent [Collins and Morton, 1995]. Thus, the effective number of sib-pairs in our study is approximately equal to $8(12-1)=88$. Although this non-independence results in an asymmetrical restriction in the range of IBD values, it does not affect the variability of IBD scores in sibships [Suarez et al., 1979].

All genotypes were downloaded from the CEPH Genotype database browser V2.0b. We used the so-called ''AFM markers,'' which offer maximum accuracy among the CEPH markers (personal communication from Mourad Sahbatou of the CEPH). These dinucleotide repeats (AC)n markers were used in the construction of genetic maps (e.g., Genethon) and were thus submitted to numerous quality control checks [Dib et al., 1996]. In order to minimize the ambiguity concerning the phase, only markers having a heterozygosity of at least 60% were first selected ($N = 4{,}015$). Because an uneven distribution of distances between markers would result in an artificial increase of the genome-wide variance in genetic identity, we trimmed this first set in such a way as to obtain a subset of the most heterozygous markers separated by at least 1 cM and at most 5 cM. This resulted in a total number of markers of $N = 1{,}522$, interspaced by 2.3 cM on average. The Marshfield Map [Broman et al., 1998] was used to estimate the distances between the markers. Like the Genethon map, this more recent map is also based on the CEPH families, but comprises essentially tri- and tetranucleotide markers [Weber and Broman, 2001].

All IBD measures were performed by the ''-ibd'' and ''-extended'' functions of the ''Multipoint Engine for Rapid Likelihood Inferences'' [Abecasis et al., 2002]. ''MERLIN'' can rapidly solve for phase ambiguity by taking into account the information contained at the surrounding markers.

The algorithm generates accurate probabilities that the siblings share 0, 1, or 2 genes at the ambiguous locus provided that the phase is known at nearby loci. Such ''multipoint'' analysis, however, can buffer out a sizeable amount of the standard deviation of IBD sharing, an undesirable outcome given our main goal, i.e., to provide the most accurate estimations of the mean and the variation of actual IBD sharing at the genome level. Consequently, for each marker, we selected only the sibling pairs for which the phase was known without any ambiguity. Note that all allele matching estimates refer to IBD, not to identity by state. The IBD status was determined from the parents' genotypes uniquely.

THE IMPACT OF GENOTYPING ERRORS ON IBD ESTIMATIONS

Estimates of the rate of genotyping errors range from 0.5% [Nicolae and Cox, 2002] and 1% [Ott, 1999] to 3% [Brzustowicz et al., 1993]. Some errors may be easily detected when the genotypes of the parents, along with the genotype of the siblings in focus, are known. In such situations, the occurrence of an allele absent in the parental generation clearly indicates a slip-up in the genotyping. Fortunately, the AFM markers of the CEPH database that we used were already cleaned for these simple Mendelian errors. If multipoint data are available, additional unlikely genotypes can also be detected. The ''-error'' procedure implemented in Merlin finds genotypes that imply a recombination pattern that is not supported by the surrounding markers [Abecasis et al., 2002]. In the present analysis, some inconsistencies of that kind were found, and the unlikely genotypes were removed using the ''-pedwipe'' command also implemented in Merlin. That being said, we found no inconsistencies on chromosome 8, 14, 15, 16, and 17 and very few on the others, perhaps with the exception of chromosome 13, for which 4 individuals had an unlikely genotype at marker D13S1274.

Having minimized the occurrence of genotyping errors in our data as much as possible, we nevertheless attempted to assess their general impact. Although errors can be a major nuisance in linkage analysis [Mitchell et al., 2003; Zou and Zhao, 2004], it remains to be determined whether or not they significantly affect genome-wide IBD estimations. We put forth a simple analysis that replicates the prevailing conditions when the multipoint procedure is not performed. At any locus, an allele may be confused with the other allele present in the same parent, an eventuality that is compatible with Mendelian inheritance and may easily go undetected if the phase at nearby loci is not taken into account. Starting from a reduced sample of our data (with a minimum of 345, a maximum of 517, and an average of 452.55 markers), we simulated such swapping of parental alleles in 5 replicates. To simplify, only one allele per marker was susceptible of being changed. A genotyping error rate of 1%, which falls within reported values, was first simulated. Higher rates of 5 and 10% (an extreme, unlikely case scenario) were also simulated in order to fully assess the range of the impact of errors. The IBD estimation analysis that thus followed was identical to the one performed on the original, non-simulated data (except that, of course, the ''-error'' and ''-pedwipe'' procedure were not performed on the reduced ''experimental'' dataset).

RESULTS

THE OVERALL DISTRIBUTION OF IBD SHARING

Our analysis provided no evidence for any significant deviation from random expectations. Over 397,831 pairwise comparisons considering each pair of siblings separately for each marker, the mean IBD sharing was 49.984% and the standard deviation 35.34%. These percentages are almost identical to the expectation and the standard deviation of a random variable that can take the values 0, $\frac{1}{2}$, and 1 with the probability $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. As previously stated, such a random variable has a mean of $\frac{1}{2}$ and a standard deviation of $\sqrt{1/8}$ (= 35.36%). A test of proportion using a value of n=8,800 for the number of ''trials'' yielded a test statistic of $Z=-0.030$ and a large p value of 0.976.

Breaking down the tabulations by sibling pairs, and using 600 to 900 markers on the 22 autosomal chromosomes, a mean IBD sharing of 49.94 % and a standard deviation of 3.95% were found for 498 sibling pairs. (The number of sibling pairs is <517 because all pairs for which the IBD status was not known for more than 600 markers were removed.) These figures agree with results obtained from simulations. Simulating a Poisson process for the breaks (and recombination) without interference along each chromosome, a mean of 50% and

a standard deviation of 4.2% were obtained. For maternal and paternal DNA, our estimated standard deviations are, respectively, 4.82% and 6.19%.

Based on the variation in IBD sharing, we can define and estimate an ''effective number of loci'' n_e in the human genome, that is, the number of loci that would yield the same standard deviation in the proportion of genome shared between siblings if all loci were segregating *independently*. When only one locus is considered, the standard deviation of IBD sharing is $\sqrt{1/8}$ or $\sqrt{.125}$. For n_e independent loci, it is $\sqrt{.125/n_e}$. Equating the latter to our estimated s.d. of 3.95%, we find $n_e = 80.2$. In other words, ~ 80 loci segregating independently in the human genome would give the same standard deviation in IBD sharing among siblings as the actual partially linked loci do. An effective number of loci can also be calculated for sex-specific DNA. Noting that the standard deviation would be $\sqrt{.25/n_e}$ in this case, we obtain $n_e \approx 108$ and $n_e \approx 65$, respectively, for maternal and paternal DNA.

Figure 1 shows the distribution of IBD sharing in sibling pairs for combined (maternal+paternal), maternal, and paternal DNA. Two panels are presented with histograms on the same data using different binning sizes. For example, the bars located at the middle of each of the three distributions in Figure 1a (low-resolution panel) includes all pairs of siblings who share between 47.5 and 52.5% of their combined, maternal, and paternal DNA, respectively. In Figure 1b (high-resolution panel), the sizes of the bins are reduced to give more details about the distribution. It can be seen, for example, that exactly 50 sibling pairs shared between 49.5 and 50.5% of their combined (maternal+paternal) DNA.

It is clear that the sharing of paternal alleles has a larger spread than that of maternal alleles. This was expected in as much as the recombination rate (and the number of independent segregating units) is smaller in paternal than in maternal gametes. Note that the sharing of paternal DNA may vary between 30 and 69%, whereas the sharing of maternal DNA is never less than 34%

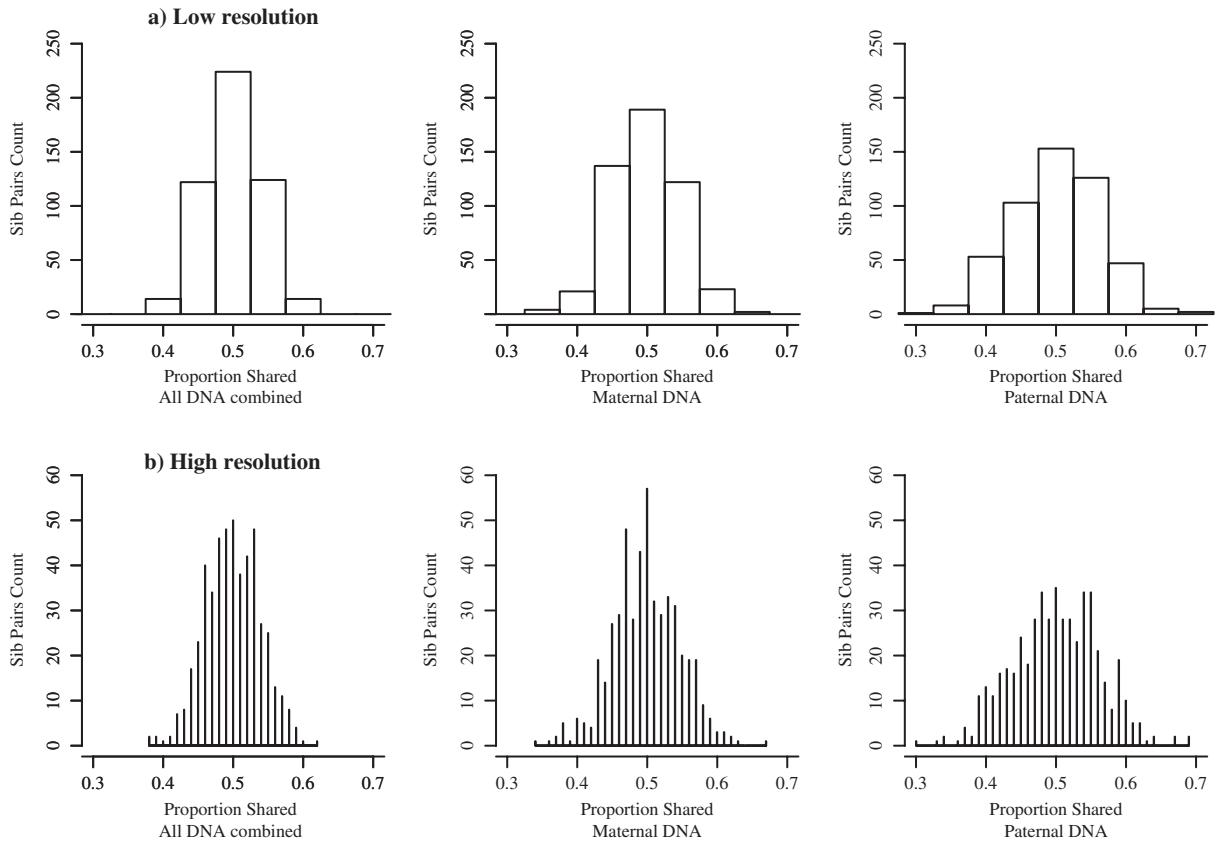

Fig. 1. Frequency distribution of the proportion of genome shared identical by descent among CEPH siblings (n=498 sibling pairs).

or greater than 66%. More generally, the quarter of sib-pairs who share the least DNA share 44.9% of their alleles on average, while those who share the most share 54.8%. The corresponding proportions are 43.8 and 55.9% for maternal DNA, and 42.0 and 57.7% for paternal DNA. The scores are even more spread out at the chromosomal level.

DISTRIBUTION PER CHROMOSOME

Table I lists first, for each chromosome, the mean and the standard deviation of IBD sharing among siblings for all (maternal+paternal) DNA, as well as for maternal and paternal DNA separately. From chromosome 1 to 22, one sees a sizable increase in the standard deviation. For combined (maternal+paternal) DNA, the standard deviation goes from 13.5% on chromosome 1 up to 27.2% on chromosome 22. Again, the most extreme tendencies are seen in paternal DNA, with standard deviations ranging between 20.9% and 38.6%. Table I also displays the effective number of loci for each chromosome for all DNA. For instance, chromosome 1 has about 7 effective loci, while chromosome 22 has less than 2. Note that by summing all the individual n_e pertaining to each chromosome, we obtain 80.3 effective loci, a result almost identical to the one obtained above for the whole genome (80.2; see The Overall Distribution of IBD Sharing). This remarkable property of additivity is not a coincidence and is apparent in Guo's [1996] formula relating chromosomal lengths to variations in IBD sharing.

Table II shows the proportions of siblings who share no DNA or their entire DNA on each chromosome. These proportions provide a way to estimate the total length (TL) of the chromosomes for each sex, and to assess the validity of our results. For example, 8.5% of sibling pairs do not share any paternal DNA on chromosome 19, while about the same proportion (7.9%) shares all. The corresponding figures for maternal DNA are 5.5 and 3.9%. Barring the unlikely events that crossovers repeatedly occur at the same spot, the proportion of sib-pairs sharing all or nothing on a chromosome is the probability of having no recombination at two independent meioses. This probability can be approximated by: $P = e^{-2\,TL}$ so that $TL = -\ln(P)/2$. Thus, the paternal and maternal lengths of chromosome 19 should be approximately equal to $TL = -\ln(16.4\%)/2 = 90.2\,\text{cM}$, and $TL = -\ln(9.4\%)/2 = 118\,\text{cM}$, respectively. The first of these two estimates agrees with deCode's findings (92.6 cM; [Kong et al., 2002]). However,

TABLE I. Proportion (%) and variation (%) of DNA shared IBD in sib-pairs per chromosome

	Combined			Maternal		Paternal	
Chromosome	Mean	Std. dev.	n_e	Mean	Std. dev.	Mean	Std. dev.
1	48.7	13.5	6.8	49.4	16.5	48.1	20.9
2	49.4	14.0	6.4	49.2	16.5	49.6	22.2
3	50.1	14.9	5.6	48.8	19.3	51.3	22.4
4	52.5	16.7	4.5	50.6	19.3	54.4	25.2
5	49.4	16.6	4.5	49.2	19.2	49.7	27.0
6	49.3	18.7	3.6	48.8	22.2	49.8	27.7
7	49.3	16.1	4.8	49.2	18.6	49.4	26.6
8	51.1	18.1	3.8	50.4	19.7	51.8	29.1
9	49.5	18.7	3.6	50.1	22.8	48.8	28.2
10	48.8	17.9	3.9	49.0	20.6	48.6	27.6
11	49.9	19.2	3.4	50.5	22.0	49.3	30.9
12	50.4	17.9	3.9	50.3	19.9	50.5	28.0
13	49.8	21.3	2.8	50.8	26.7	48.8	33.5
14	50.7	21.4	2.7	50.3	23.9	51.1	32.7
15	51.3	21.7	2.7	51.0	25.0	51.6	33.5
16	50.7	20.5	3.0	49.5	23.2	51.9	33.0
17	49.5	20.1	3.1	49.7	23.7	49.3	31.8
18	49.7	20.7	2.9	50.9	22.9	48.6	32.6
19	49.3	21.7	2.6	49.9	26.4	48.6	33.4
20	50.1	24.2	2.1	50.6	26.2	49.6	37.9
21	49.3	26.6	1.8	49.2	36.0	49.5	38.6
22	50.8	27.2	1.7	50.2	33.8	51.4	38.5

TABLE II. Proportion (%) of sibling pairs sharing no DNA or their entire DNA per chromosome

	Combined		Maternal		Paternal	
Chromosome	Zero	All	Zero	All	Zero	All
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.2	0.2
3	0.0	0.0	0.0	0.0	0.0	0.2
4	0.0	0.0	0.2	0.0	0.8	1.0
5	0.0	0.0	0.8	0.4	0.8	1.4
6	0.0	0.0	1.2	0.2	2.4	3.4
7	0.0	0.0	0.4	0.0	3.0	2.6
8	0.0	0.0	0.4	0.2	3.0	3.2
9	0.0	0.0	0.4	0.6	2.4	2.2
10	0.0	0.0	0.6	0.2	2.4	1.6
11	0.0	0.6	0.0	1.0	2.2	4.0
12	0.0	0.0	0.2	0.0	2.6	2.0
13	0.2	0.8	2.8	4.8	6.0	5.8
14	0.2	0.2	2.6	1.4	3.4	5.2
15	0.0	0.0	2.6	2.6	4.8	3.8
16	0.0	0.0	1.4	0.6	6.6	5.8
17	0.0	0.0	1.2	1.2	6.4	4.6
18	0.0	0.0	1.4	0.8	5.6	4.8
19	0.2	0.2	5.5	3.9	8.5	7.9
20	1.0	1.0	3.4	3.6	16.1	14.7
21	4.2	3.0	18.1	15.3	21.7	21.5
22	2.2	2.6	11.2	10.4	20.7	20.3

our estimated length for the maternal chromosome is smaller than that of deCode (126.8 cM) [Kong et al., 2002]. Random fluctuations may explain this discrepancy, as the total number of *independent* sib-pairs in our sample is approximately equal to 88 (see above). Another likely possibility may be that interference in crossovers and the existence of recombination "hotspots" and "deserts" significantly affect the markers' lengths [Crawford et al., 2004; McVean et al., 2004].

DISTRIBUTION PER MARKER

Figure 2 shows that the overall distribution of IBD sharing among siblings per marker is positively skewed. This skew arises because of the large progeny sizes in the CEPH families. At any given locus, alleles may be all identical in the sibship if parents systematically transmit the same one to their progeny, although the probability for this to occur is low. On the other hand, siblings cannot have alleles that are all different from one another on a given marker because the parents have only 2 alleles to transmit. Consequently, when the progeny exceeds two, the mode and the median are shifted left to the mean. Note the very high paternal IBD sharing for one marker (D-number: D16S3068) on the high-resolution panel in Figure 2 (69.3%).

THE EFFECT OF TYPING ERRORS ON IBD SHARING

Table III shows the results of three simulations corresponding to typing error rates of 1, 5, and 10%. In the initial free-of-errors simulated dataset, the average IBD sharing was 49.94% and the standard deviation 4.15% (the standard deviation is higher than the one calculated on the main dataset because the marker coverage is now more sparse and uneven; remember that the purpose here is to illustrate with a replication of experimental conditions). As expected, the means of the different simulation rounds were very close to the theoretical value of IBD sharing (50%). The

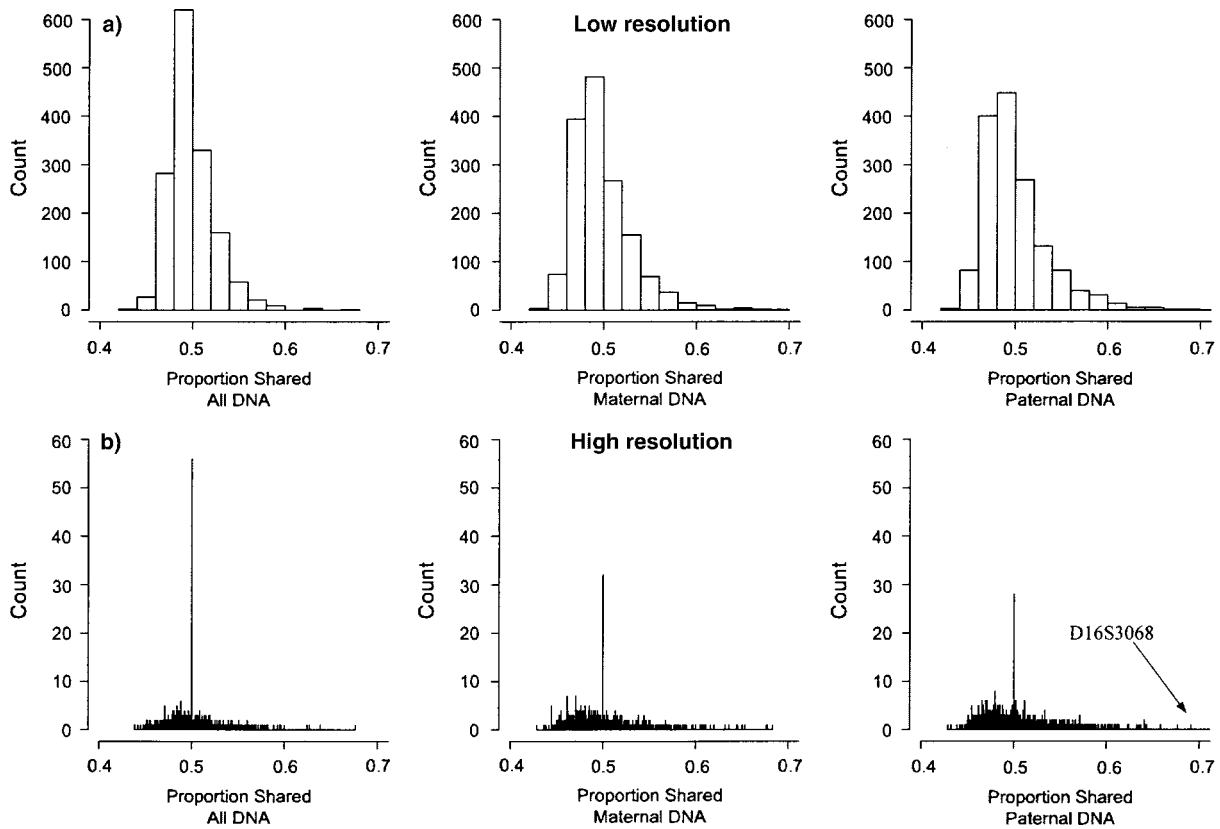

Fig. 2. Frequency distribution of the proportion of DNA shared identical by descent per marker among CEPH siblings (n=1,257 markers).

TABLE III. Mean and standard deviations (%) of IBD sharing among siblings with simulated genotyping error rates of 1, 5, and 10% (from a subset of our data having an initial standard deviation of 4.15% in IBD sharing)

	Genotyping error rate					
	1%		5%		10%	
Simulation round	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
1	49.95	4.12	50.09	3.96	50.01	3.66
2	49.94	4.04	49.96	3.88	49.96	3.65
3	49.95	4.11	49.92	3.91	49.94	3.73
4	49.92	4.12	49.91	3.97	49.95	3.61
5	49.95	4.12	50.00	3.90	49.95	3.63
Mean total	49.94	4.10	49.97	3.92	49.96	3.66

standard deviation, on the other hand, decreased with increasing error rates. Each error can be seen as a virtual double crossing-over event that artificially increases the number of segregating units and thus decreases the standard deviation of IBD sharing. Concomitantly, the recombination fraction is inflated, resulting in increased apparent map distance [Buetow et al., 1994; Mitchell et al., 2003]. In the worse case scenario envisaged, a 10% genotyping error rate brings down the standard deviation from 4.15 to 3.66%, a reduction of about 12%. A more realistic error rate of 1% would slightly reduce the standard deviation from the same expected 4.15% to about 4.10%, representing a 1.2% decrease. At the intermediate level of a 5% error rate, one falls back on our estimated value of the standard deviation, i.e., ~3.95% (3.92% in Table III).

DISCUSSION

Although Mendel largely ignored several genetic mechanisms involved in the determination of traits (dominance, epistasis, interaction, etc.), he accurately described the actual process of gene transmission at the locus level. Genes are truly transmitted randomly with a fair chance of 50%, which results in an average of 50% of DNA sharing among brothers and sisters. However, there is a considerable amount of variation around the mean that has not yet been accurately sized and accounted for. The purpose of this report was to provide a glimpse of the variation that could trigger more advanced studies. Here are some comparisons of our results with estimations reported in the literature and some observations that can be derived from our findings.

The standard deviation in IBD sharing for the 498 sib-pairs of our sample was 3.95%, a score falling within the range of previously published theoretical expectations or simulations (Table IV). The lowest standard deviation found so far is that of Rasmuson [1993] who, using the ''recombination index'' (RI), derived a value of 3.53%. It is not clear how exactly Rasmuson obtained the value 100 for the RI. This is probably an overestimation, which results in a low variation in IBD sharing. The highest standard deviation documented is that of Suarez et al. [1979]. Based on a simulated model of ''chiasma localization,'' Suarez et al. [1979] obtained a value of 5.60%. Such localization may render the loci more dependent on one another, and may thus increase the variance in genome-wide IBD sharing. However, it is hard to imagine how this could have such a large effect. Judging from Guo's [1996] results, the increase in variation associated with the use of a chiasma model appears to be small, if not undetectable (Table IV). The reason for the discrepancy lies elsewhere: Suarez et al. [1979] used Hultén's data [1974] on *male* meiosis to calculate the variance of identity by descent, which has artificially decreased the number of independent meioses for the combined sex autosomes.

TABLE IV. Estimations of the standard deviations of the proportion of the genome shared by siblings in the literature and in this study

Study	Method	Std. dev. (%)	Corresponding n_e
Suarez et al. [1979]	Simulations using male chiasma distribution	5.59	40
Risch and Lange [1979b]	Analytical estimation based on a probability-generating function	4.00	78
Rasmuson [1993]	Estimation based on the recombination index	3.53	100
Guo [1996]	Analytical result based on a two–state Markov chain, using a sex–averaged linkage map	3.67	93
Guo [1996]	Using a chiasma map	3.78	87
Guo [1996]	Using data from deCode's map [Kong et al., 2002] in Guo's formula	3.89	83
This study	Empirical, with CEPH data	3.95	80

In our opinion, the first ''accurate'' estimate of IBD variation was done by Risch and Lange as early as 1979 [Risch and Lange, 1979b]. Developing a probability-generating function for the number of crossovers, and using the relative lengths of each of the 22 autosomal chromosomes from available maps, they estimated a standard deviation of 4.00%. The latter result is closer to our empirical value than the low scores of 3.67 and 3.78% offered by Guo [1996] who used chromosomal lengths amounting to 4,000 cM (adapted from Buetow et al.'s [1994] and Morton's [1991] maps). Plotting the more recent and probably more accurate lengths from the deCode's map (amounting to a total of 3,435 cM) in Guo's formula, one finds a standard deviation of 3.91%, which is consistent with our estimate. Note that the results are almost identical when using the 3,498-cM SNP map from Matise et al.'s [2003] map or the 3,488-cM microsatellite Marshfield map from Broman et al. [1998], i.e., 3.89%.

There are, however, some discrepancies at the chromosomal level between our results and those obtained with the use of Guo's formula (not shown). In comparison, our standard deviations are slightly lower for longer chromosomes and, conversely, slightly higher for smaller chromosomes. This is probably due to the limitation of our sample. The number of *independent* sib-pairs in the 8 CEPH families is approximately equal to 88. This number of ''trials'' may not be sufficient to allow, on the one hand, a complete unfolding of the variation to the extremes on large chromosomes and, on the other hand, to limit the random fluctuations occurring on the smaller ones. Admittedly, the two biases may have acted in opposite directions. Even on the largest chromosomes, the proportion of sib-pairs who share no DNA or all their DNA should be greater than zero in a sufficiently large sample. Table II shows that it is not the case in our sample. The large CEPH family structures may also asymmetrically constrain the range of variation [Suarez, et al., 1979]. Figure 2 clearly illustrates how this can lead to a high false-positive rate in classic linkage studies. Just by chance, a few markers may be shared by many siblings, while most will display less than 50% of sharing.

An improper marker coverage could have affected our results, but to a relatively small extent. In order to assess this possibility, we made successive random trims of the markers at different random positions, leaving a coarse, uneven distribution of anchors on the map. The results remained unchanged until reaching an average of 4 cM. After this threshold, the standard deviations in IBD sharing increased rapidly over 4%. The inclusion of only one type of marker in our analysis (dinucleotides) may also be seen as a potential source of bias. For instance, dinucleotide repeats are more prone to strand slippage (skipping of repeats during amplification resulting in fragments smaller than original fragments) than trinucleotide or tetranucleotide repeats [Weber and Broman, 2001]. However, owing to higher mutation rates, dinucleotides have higher heterozygosity rates than markers with higher repeat lengths [Chakraborty et al., 1997], a highly desirable property in a study of the variation in IBD sharing. Similarly, although a cost-effective set of tightly spaced SNPs provides superior power to detect linkage than the more widely spaced microsatellites [Evans et al., 2004; Middleton et al., 2004; Pato et al., 2005], the strong reduction in heterozygosity of such markers could result in a highly uneven distribution of loci with unambiguous phase, which would unduly increase the variance in the estimation of genome-wide IBD sharing.

In our ''no errors'' simulation dataset, a sparse and non-uniform distribution of markers was deliberately chosen, resulting in an overestimate of the variations in IBD. Genotyping error rates of 1, 5, and 10% resulted in smaller standard deviations. The intermediate rate of 5% produced a result that closely matched our above-reported final estimate of 3.95%. All these manipulations have left the mean unchanged to 50%, while showing how two counterbalancing biases may cancel each other out so as to finally yield reasonable estimates of the standard deviation. Nevertheless, overall, an unlikely high error rate (5% or more) is needed to appreciably affect the variation in IBD sharing. In a sib-pair study with no parental DNA available, other typing errors would be likely, such as those involving alleles that are absent in the parent generation. Such errors would reduce the mean but increase the standard error of IBD sharing.

Our empirical study may provide useful insights for research involving linkage analyses and human identity assessment in general. For instance, the high proportions of sibling pairs who share no DNA or their entire DNA on small chromosomes (Table II) highlight the difficulties for detecting linkage in those chromosomes. As another example, multi-locus match probabilities in forensic inquiries usually assume indepen-

dence between loci [Ayres, 2000], and typing additional loci on small chromosomes, which contain very few *independent* loci, may rapidly produce decreasing returns [Weir, 1994].

Using the effective number of loci, we can derive the approximate number of markers that one could expect to fall in the critical area at a given level of significance. The variation in IBD that we measured for the sib-pairs is that of a random variable with about 80 independent ''trials.'' Dividing the total number of markers included in the analysis (1,522) by the number of independent loci (80.2), we obtain the size of the ''blocks'' that segregated independently in the genome of the sib-pairs ($=1{,}522/80.2 \approx 19$). The markers contained in such a hypothetical block do not segregate together as parts of a haplotype on a given chromosome; they form a bloc in the sense that they are ''statistically'' linked. At a 95% level of confidence, and with 100 independent sib-pairs, we would expect about $(0.05).80 = 4$ of such blocs, for which the IBD sharing would fall over 56.9% or under 43.1% $(= 0.5 \pm 1.96 \times \sqrt{.125/100})$. In other terms, out of 1,522 markers, 38 would have IBD scores greater than 56.9% and the same amount would have scores lower than 43.1%. These calculations are fairly simple and open the way for a quick assessment of the genome-wide significance threshold for sib-pairs in a descriptive, preliminary phase of a linkage study.

ACKNOWLEDGMENTS

We thank Mourad Sabbatou of the CEPH in Paris for useful advice on data selection, Benjamin Beall, Edwin John Sullivan, and the two anonymous reviewers for helpful comments on an earlier draft of this paper.

ELECTRONIC DATABASE INFORMATION

All data used in this article were downloaded from the CEPH Genotype database browser V2.0b at http://www.cephb.fr/cephdb/php.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Ayres KL. 2000. A two-locus forensic match probability for subdivided populations. Genetica 108:137–143.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861–869.

Brzustowicz LM, Merette C, Xie X, Townsend L, Gilliam TC, Ott J. 1993. Molecular and statistical approaches to the detection and correction of errors in genotype databases. Am J Hum Genet 53:1137–1145.

Buetow KH, Ludwigsen S, Scherpbier-Heddema T, Quillen J, Murray JC, Sheffield VC, Duyk GM, Weber JL, Weissenbach J, Gyapay G, et al. 1994. Human genetic map. Genome maps V. Wall chart. Science 265:2055–2070.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94:1041–1046.

Collins A, Morton NE. 1995. Nonparametric tests for linkage with dependent sib pairs. Hum Hered 45:311–318.

Cotterman CW. 1940. A Calculus for statistical genetics: Ohio State University.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706. Epub 2004 Jun 6.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154.

Donnelly KP. 1983. The probability that related individuals share some section of genome identical by descent. Theor Popul Biol 23:34–63.

Evans DM, Cardon LR, Morris AP. 2004. Genotype prediction using a dense map of SNPs. Genet Epidemiol 27:375–384.

Fulker DW, Cherny SS, Cardon LR. 1995. Multipoint interval mapping of quantitative trait loci, using sib pairs. Am J Hum Genet 56:1224–1233.

Guo SW. 1994. Computation of identity-by-descent proportions shared by two siblings. Am J Hum Genet 54:1104–1109.

Guo SW. 1995. Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am J Hum Genet 56:1468–1476.

Guo SW. 1996. Variation in genetic identity among relatives. Hum Hered 46:61–70.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19.

Hulten M. 1974. Chiasma distribution at diakinesis in the normal human male. Hereditas 76:55–78.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14:528–538.

John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, et al. 2004. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. Am J Hum Genet 75:54–64.

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.

Leclair B, Fregeau CJ, Bowen KL, Fourney RM. 2004. Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair flight 111 disaster. J Forens Sci 49:939–953.

Malécot G. 1941. Etude mathématique des populations mendéliennes. (Mathemetical studies of Mendelian populations.) Annales de l'Université de Lyon, Sciences (A-4):45–60.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, Chui B, Cohen P, de Toma C, et al. 2003. A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet 73:271–284.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, Brown A, Petryshen TL, Kirby AN, Medeiros H, et al. 2004. Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. Am J Hum Genet 74:886–897.

Mitchell AA, Cutler DJ, Chakravarti A. 2003. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet 72:598–610. Epub 2003 Feb 13.

Morton NE. 1991. Parameters of the human genome. Proc Natl Acad Sci USA 88:7474–7476.

Nicolae DL, Cox NJ. 2002. MERLIN.and the geneticist's stone? Nat Genet 30:3–4.

Ott J. 1999. Analysis of human genetic linkage. Baltimore: Johns Hopkins University Press.

Pato CN, Middleton FA, Gentile KL, Morley CP, Medeiros H, Macedo A, Azevedo MH, Pato MT. 2005. Genetic linkage of bipolar disorder to chromosome 6q22 is a consistent finding in Portuguese subpopulations and may generalize to broader populations. Am J Med Genet B Neuropsychiatr Genet 134: 119–121.

Presciuttini S, Toni C, Tempestini E, Verdiani S, Casarino L, Spinetti I, De Stefano F, Domenici R, Bailey-Wilson JE. 2002. Inferring relationships between pairs of individuals from locus heterozygosities. BMC Genet 3:23.

Rasmuson M. 1993. Variation in genetic identity within kinships. Heredity 70:266–268.

Rijsdijk FV, Sham PC. 2002. Estimation of sib-pair IBD sharing and multipoint polymorphism information content by linear regression. Behav Genet 32:211–220.

Risch N, Lange K. 1979a. An alternative model of recombination and interference. Ann Hum Genet 43:61–70.

Risch N, Lange K. 1979b. Application of a recombination model in calculating the variance of sib pair genetic identity. Ann Hum Genet 43:177–186.

Suarez BK, Reich T, Fishman PM. 1979. Variability in sib pair genetic identity. Hum Hered 29:37–41.

Suarez BK, Rice J, Reich T. 1978. The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94.

Wang T, Elston RC. 2005. Two-level Haseman-Elston regression for general pedigree data analysis. Genet Epidemiol 29:12–22.

Weber JL, Broman KW. 2001. Genotyping for human whole-genome scans: past, present, and future. In: Rao DC, Province MA, editors. Genetic dissection of complex traits. San Diego: Academic Press. p 77–96.

Weir BS. 1994. The effects of inbreeding on forensic calculations. Annu Rev Genet 28:597–621.

Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, et al. 2001. Comparison of human genetic and sequence-based physical maps. Nature 409:951–953.

Zou G, Zhao H. 2004. The impacts of errors in individual genotyping and DNA pooling on association studies. Genet Epidemiol 26:1–10.