ARTICLE

# Using Genomic Inbreeding Coefficient Estimates for Homozygosity Mapping of Rare Recessive Traits: Application to Taybi-Linder Syndrome

Anne-Louise Leutenegger, Audrey Labalme, Emmanuelle Génin, Annick Toutain, Elisabeth Steichen, Françoise Clerget-Darpoux, and Patrick Edery

The use of inbred patients whose exact genealogy may not be available is of primary interest in mapping genes involved in rare recessive diseases. We show here that this can be achieved by estimating inbreeding coefficients from the patients' genomic information and using these estimates to perform homozygosity mapping. We show the interest of the approach by mapping a gene for Taybi-Linder syndrome to chromosome 2q, with the use of a key patient with no genealogical information.

Affected offspring from consanguineous marriages may be of particular interest in mapping genes involved in recessive diseases. The disease locus is, indeed, likely to be found in a region where affected individuals have received twice the same ancestral allele (identical by descent [IBD]). In such a region, the alleles at polymorphic loci surrounding the disease locus are likely to be also IBD. The patient is said to be autozygous in such a region.

Lander and Botstein[1] proposed a method, referred to as "homozygosity mapping," that consists of searching for a region of the genome that is autozygous in inbred individuals affected by a given disease. They showed that, to quantify the evidence of linkage provided by such a region, a LOD score could be computed for the marker observations by comparing the likelihood of being at the disease locus with the likelihood of being at a random point on the genome. Calculation of the latter likelihood requires that, for each affected inbred individual, the chance of having two IBD alleles at a locus randomly sampled on the individual's genome is known. By definition,[2,3] this value is the individual's inbreeding coefficient (*F*). Efficient algorithms[4–7] based on the known genealogy have been previously developed to compute *F*.

However, information on genealogy may not be accurate or may even be lacking, especially for populations in which marriages between relatives are very frequent, making relationships very complex. Miano et al.[8] reported some pitfalls in homozygosity mapping. One of them was potential LOD score inflation and hence potential increase in false linkage evidence because of underestimation of the extent of inbreeding in the affected individual or, equivalently, of the extent of kinship between the parents of a patient. More generally in linkage analysis, underestimation of the parental relationships may lead to an increase in type I error.[9]

Here, we propose to estimate *F* from each individual's genomic information (by FEstim) as presented by Leutenegger et al.[10] and to use this genomic *F* to control for parental relationships in the LOD score computation. Hence, to perform linkage analysis when parental relationships are poorly known, we introduce a new homozygosity mapping statistic, FLOD. This statistic allows investigators to include inbred patients in homozygosity mapping without having any knowledge of their genealogy.

We show the advantage of FLOD over the usual homozygosity mapping LOD score (HMLOD) by mapping the first locus for an autosomal recessive disease, Taybi-Linder syndrome, with the use of a key patient without any genealogical information.

## Methods

### *Estimation of the Genomic Inbreeding Coefficient by FEstim*

We have proposed a new method, FEstim, to estimate the inbreeding coefficient *F* of an individual by use of the individual's genomic data.[10] Our method does not require any knowledge of the parental relationships. Instead, it uses only information on genotypes at random markers throughout the individual's genome, which allows an estimation of the proportion of the genome that is autozygous. The observed marker genotypes are modeled by a hidden Markov chain that depends on *F* and on the rate of change of IBD status per cM. These are both estimated by maximum likelihood, with the intermarker genetic distances and marker-allele frequencies specified. It is worth noting that the reliability of our estimator depends on the informativity at each point of the genome—that is, on the density and heterozygosity rate of the markers.

FEstim gives more-specific information about an individual's genome than the genealogy does, because it better reflects the true proportion of the individual's genome that is autozygous. Indeed, the genealogical *F* is the expected value of the genomic

From the INSERM U535, Villejuif, France (A.L.L.; E.G.; F.C.-D.); INSERM U679 (A.L.L.), Université Paris 6 (A.L.L.), Hôpital Pitié-Salpétrière (A.L.L.), and Université Paris 11 (E.G.; F.C.-D.), Paris; Hospices Civils de Lyon (A.L.; P.E.) and EA 3739, Université Claude Bernard–Lyon1 (P.E.), Lyon, France; Centre Hospitalier et Universitaire de Tours, Tours, France (A.T.); and Universität Klinik für Kinder- und Jugendheilkunde, Innsbruck, Austria (E.S.)

Received February 13, 2006; accepted for publication March 27, 2006; electronically published April 28, 2006.

Address for correspondence and reprints: Dr. Françoise Clerget-Darpoux, INSERM U535, BP 1000, 94817 Villejuif Cedex, France. E-mail: clerget@vjf.inserm.fr

*Am. J. Hum. Genet.* 2006;79:62–66.  0002-9297/2006/7901-0008$15.00

*F.* Since the genome of an individual is of finite size and only represents a small gene sampling, a large variance around this expected value may be observed. Thus, the use of an individual's genealogy to estimate the chance of having two IBD alleles at any random point on the genome may give estimates that are very far from what is actually happening on the genome. For offspring of first cousins, for instance, the probability of sampling an autozygous marker from their genome may be as low as 3% and as high as 12% when its expected value from the genealogy is 6%.[10]

### *New Homozygosity Mapping Statistic FLOD*

We propose to incorporate this genomic *F,* estimated for each affected inbred individual, into the LOD score statistic, instead of *F* estimated from the genealogy. For a given affected inbred individual, we define FLOD at marker *k* as

$$\log_{10}\frac{P(X_k = 1 \mid \mathbf{Y}) + qP(X_k = 0 \mid \mathbf{Y})}{f + q(1-f)} ,$$

where $f$ is the genomic estimate of *F, q* is the disease-allele frequency, $X_k$ is the IBD status at marker $k$ (1 for IBD; 0 for non IBD), and $\mathbf{Y}$ is the observed genotypes at all markers along the individual's genome. The LOD score statistic FLOD is computed using a multipoint method. More details can be found in the work of Leutenegger.[11]

For a sample of independent, affected inbred individuals, the FLOD value for the sample is the sum of each individual FLOD. This statistic enables us to include in a linkage analysis any affected individuals, without requiring any genealogical information.

When siblings of a patient are available for study, FEstim estimates are obtained for all of them, and the median value for the whole sibship can then be used to represent the parental kinship. The FEstim estimation and FLOD computation have been implemented in the FEstim software, which is available on request (leutenegger@vjf.inserm.fr).

### *Taybi-Linder Syndrome Data*

Taybi-Linder syndrome, or microcephalic osteodysplastic primordial dwarfism (MOPD) type I/III (MIM 210710), is a very rare autosomal recessive condition. It is characterized by intrauterine growth retardation, low birth weight, dwarfism, bone dysplasia, facial anomalies, microcephaly, and malformations of the brain.[12] Fewer than 30 patients have been described to date, many of whom died within the 1st year of life because of infectious disease.[13]

Here, we study a sample of four inbred patients, including two siblings (patients 1 and 2). The patients originated from the Mediterranean region: Algeria, Turkey, and Morocco. Clinical reports on the four patients are given elsewhere—for patients 1–3, the manuscript is in preparation, and patient 4 was reported as case 4 by Sigaudy et al.[13] (the other patients described in that article died within the first few mo of life; DNA from only patient 4 was available for the present study). In all cases, inbreeding was present but not well documented. For the affected siblings and patient 3, we had information that the parents were first cousins and that additional relationships were likely. For patient 4, there was no genealogical information at all. Additional members of the nuclear family were available for patients 1 and 2 (three unaffected siblings and their parents) and for patient 3 (her parents). On the other hand, we had no relatives of patient 4. In total, blood samples from 11 individuals were collected. Participants gave informed consent. DNA was extracted from blood lymphocytes by use of standard procedures. A high-density genomewide scan was performed through deCODE services with the use of a 1,000-marker fluorescent-labeled microsatellite screening set that covered the whole genome with an average density of 3.7 cM, where genetic locations were based on the deCODE map.[14] The map used here allows us to get very accurate estimates of the genomic inbreeding coefficient *F* with FEstim. Indeed, with the specific map characteristics, we have, as in the work of Leutenegger et al.,[10] a high correlation of 0.9 between the estimated *F* and the true proportion of genome IBD for offspring of first cousins.

First, we estimated the genomic inbreeding coefficient of all patients and their available siblings, using FEstim. We then computed FLOD for the whole sample with these estimated *F* values. Finally, we computed HMLOD, assuming a first-cousin relationship for the parents of patients 1, 2, and 3 while excluding patient 4 because the calculation could not be done using a standard computer package, such as Allegro.[15] For all LOD score computations, we used a fully penetrant autosomal recessive model and a disease-allele frequency of 0.00001. To check the sensitivity of our results to the disease-allele frequency, we also performed the analysis with a frequency of 0.0001. The LOD score values were identical up to the second decimal place (not shown).

## Results

FEstim inbreeding coefficient estimates and 95% CIs are shown in figure 1. All inbreeding-coefficient estimates of patients were significantly different from zero. In particular, patient 4 had an estimated *F* of 0.06, which showed that the parents are, in fact, closely related. This patient is thus informative for linkage, whereas previously the patient could not be used in any linkage analysis. For siblings 1 and 2, who have additional siblings available, we observed a large variability of the FEstim estimates for the sibship, from 0.04 to 0.13. The estimate for patient 3 indicates the probable existence of remote consanguinity besides the first-cousin relationship of her parents. Her *F* is estimated to be 0.13, which is significantly ($P < .025$) higher than 0.06, the expected inbreeding coefficient for first-cousin offspring.

First, we used these FEstim estimates to compute the multipoint FLOD statistic on the entire sample of the patients and their available relatives. We obtained a maximum FLOD of 3.28 at the *D2S347* marker. Moreover, an FLOD value >3 was also reached at *D2S2271,* the marker adjacent to *D2S347.* No other chromosomal region gave combined LOD scores >2. A study of the affected individuals' haplotypes on the chromosome 2q region showed meiotic recombinations between centromeric markers *D2S2254* and *D2S347* for patient 3 and between telomeric markers *D2S2271* and *D2S2215* for patients 1 and 2, thus restricting the autozygous candidate region to an ~13-cM interval flanked by *D2S2254* and *D2S2215* on chromosome 2q14.2-2q14.3 (table 1).

To compare our results with the usual homozygosity

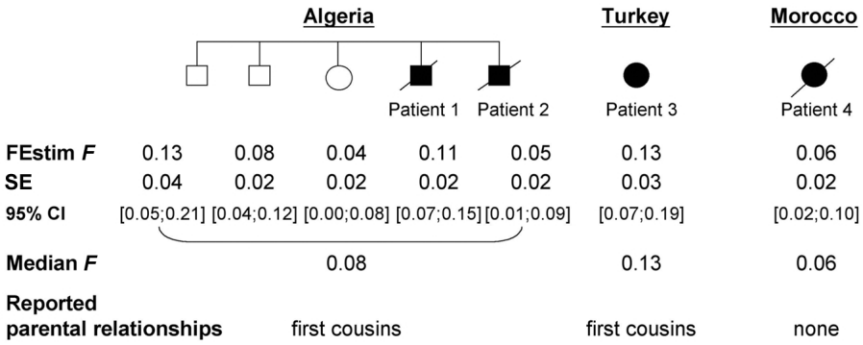


**Figure 1.** FEstim-estimated inbreeding coefficient (*F*), SE, and 95% CI for the patients with Taybi-Linder syndrome and siblings. The median *F* value for each sibship and the reported parental relationships are also specified. The 95% CIs were computed as $F \pm 1.96$ SE.

mapping statistic, we computed HMLOD for patients 1 and 2 (the siblings), patient 3, and their available relatives, assuming a first-cousin relationship for the parents. We did not find evidence of linkage in any part of the genome, but we had some suggestive results, since the combined HMLOD exceeded 2 in two chromosomal regions. We obtained LOD score values of 2.62 on chromosome 2q at *D2S347* and of 2.19 on chromosome 7q at *D7S514*. Thus, no clear-cut linkage could be established by including only these three patients and their available relatives in the linkage study. As can be seen in figure 2, because of our FLOD statistic and the inclusion of patient 4, it was possible, first, to exclude the 7q region, which had an FLOD value of −1.25, and, second, to get a LOD score >3 in the 2q region.

## Discussion

We have mapped the first Taybi-Linder syndrome locus to chromosome 2q, using our genomically controlled homozygosity mapping method. It allowed us to include in the analysis a key patient with no available genealogy. It is also interesting to note that this patient (patient 4), with an estimated genomic *F* of 0.06, is actually more informative for linkage than patient 3, who had an estimated genomic *F* of 0.13.

It is worth noting that, for patient 3, the LOD score values obtained with HMLOD under the assumption that her parents are first cousins were inflated. This statistic reached 1.2 on chromosome 2q; however, with her actual inbreeding level of 0.13, the LOD score should be reduced by $\log_{10}(2)$, thus reaching only 0.9, the observed FLOD value on chromosome 2q.

Our proposed solution to incorporate the genomic *F* of each affected inbred individual, instead of the expected *F* from the genealogy, into the LOD score statistic should make it easier to map autosomal recessive traits. The FLOD statistic has the great advantage of taking into account the actual inbreeding of individuals and its variability while allowing researchers to study a sample of patients without requiring any genealogical information. As for any linkage study, however, it does require a very informative marker map. This was the case here, in which we used a map with an average intermarker distance of 3.7 cM and an average heterozygosity rate of 0.75. A more standard 10-cM microsatellite map would yield less precise *F* estimates.

The approach proposed here is especially well suited to studying patients from populations with a long tradition of marriages between close relatives. For the case of populations with lower levels of inbreeding, such as founder populations, we are working on a typing strategy that would allow this sparse inbreeding to be correctly captured. Indeed, in that case, one needs a high informativity at each point of the genome, which depends on both marker informativity and density. This will likely require mixing SNP and microsatellite markers. However, when

**Table 1. Haplotypic Analysis Showing Candidate Region for the Taybi-Linder Syndrome Locus on Chromosome 2q14**

| | | Algeria | | | | | | | Turkey | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Location (cM) | Marker | Parents | Unaffected Siblings | | | | Patient 1 | Patient 2 | Parents | Patient 3 | Morocco Patient 4 |
| 132.2 | *D2S2254* | 2 2 | 2 2 | 2 2 | 2 2 | 2 2 | 2 2 | 2 2 | 2 2 2 12 | 2 12 | 9 9 |
| 138.1 | *D2S347* | 1 5 | 1 8 | 1 5 | 8 5 | 8 1 | 1 1 | 1 1 | 1 9 6 1 | 1 1 | 1 1 |
| 143.6 | *D2S2271* | 3 3 | 3 4 | 3 3 | 4 3 | 4 3 | 3 3 | 3 3 | 2 6 1 2 | 2 2 | 3 3 |
| 145.4 | *D2S2215* | 5 9 | 9 3 | 9 9 | 3 9 | 3 5 | 9 5 | 9 5 | 4 7 7 4 | 4 4 | 10 10 |

NOTE.—Patients' haplotypes defining the candidate region are shaded.

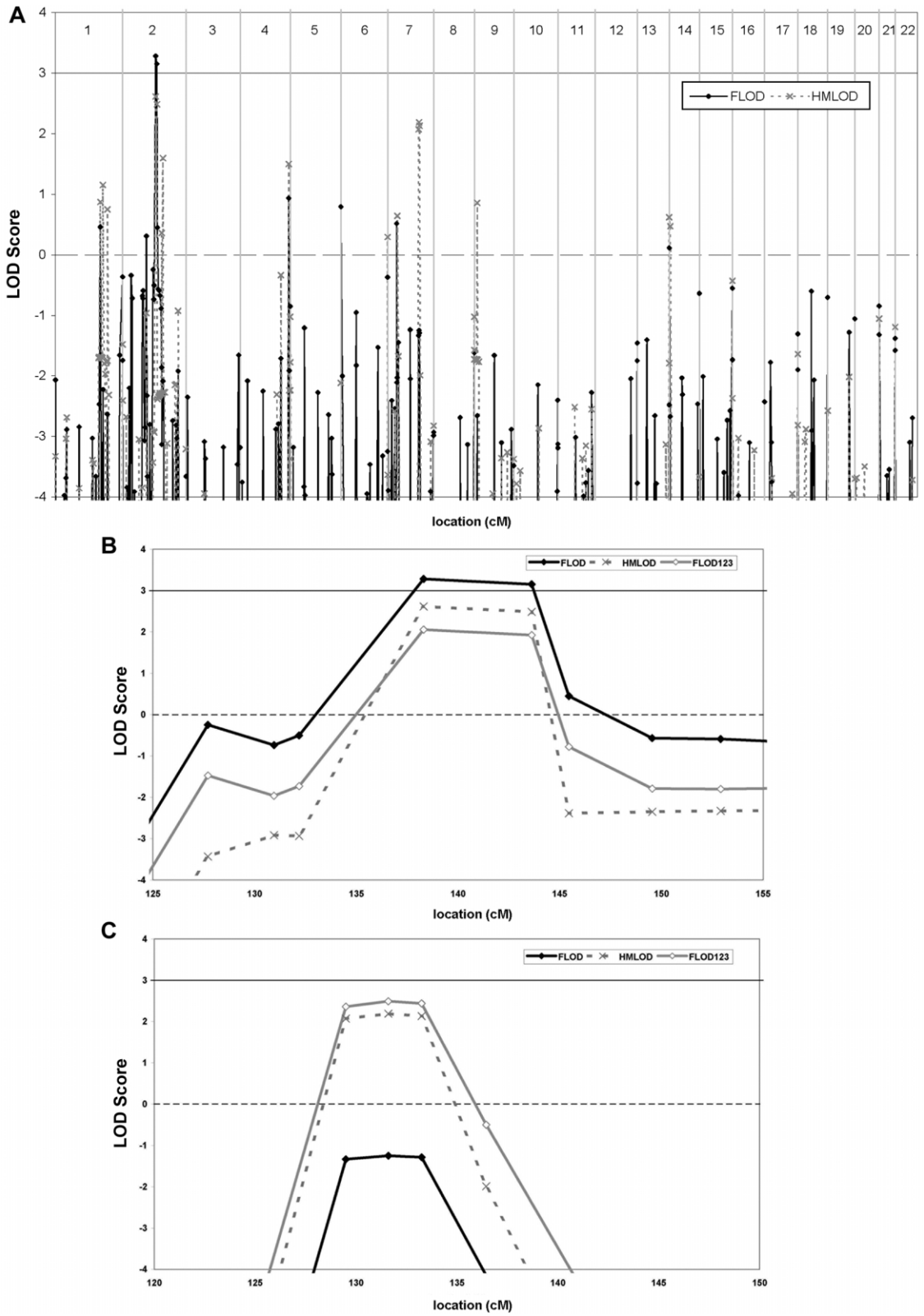


**Figure 2.** LOD score plots of FLOD (*black lines*) and HMLOD (*dashed gray lines*) over the whole genome (*A*), chromosome 2q (*B*), and chromosome 7q (*C*) for the Taybi-Linder syndrome data. On the genomewide plot (*A*), the chromosome numbers are written at the top. *B* and *C,* FLOD123 (*solid gray lines*) represents FLOD computed for patients 1, 2, and 3 and their available relatives only (not patient 4). The solid horizontal lines represent a LOD score of 3.

very dense marker maps are used, it is important to take into account the linkage disequilibrium that may be present between marker alleles. Finally, we currently are refining the disease locus interval and are undertaking a candidate-gene strategy with the aim of identifying the Taybi-Linder syndrome–causing gene itself.


### Acknowledgments

We are grateful to the family members. We also thank Marie-Claude Babron for fruitful discussions. This work was supported by the Unité de Formation et de Recherche Lyon-Nord 2005, the Hospices Civils de Lyon (contracts HCL 1999, HCL 2001, and PHRC 01.099), and the Fondation pour la Recherche Médicale (projet ARS 2.13).


### Web Resource

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for MOPD type I/III)

### References


1. Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570
2. Wright S (1922) Coefficient of inbreeding and relationship. Am Nat 56:330–338
3. Malécot G (1948) Les mathématiques de l'hérédité. Masson, Paris
4. Jacquard A (1966) Logique du calcul des coefficients d'identité entre deux individus. Population (French Edition) 21:751–776
5. Denniston C (1968) Probability and genetic relationship. PhD thesis, University of Wisconsin, Madison
6. Cockerman C (1971) Higher order probability functions of identity of alleles by descent. Genetics 69:235–246
7. Karigl G (1981) A recursive algorithm for the calculation of identity coefficients. Ann Hum Genet 45:299–305
8. Miano MG, Jacobson SG, Carothers A, Hanson I, Teague P, Lovell J, Cideciyan AV, Haider N, Stone EM, Sheffield VC, Wright AF (2000) Pitfalls in homozygosity mapping. Am J Hum Genet 67:1348–1351
9. Leutenegger AL, Genin E, Thompson EA, Clerget-Darpoux F (2002) Impact of parental relationships in maximum lod score affected sib-pair method. Genet Epidemiol 23:413–425
10. Leutenegger AL, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA (2003) Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73:516–523
11. Leutenegger AL (2003) Estimation of random genome sharing: consequences for linkage. PhD thesis, Université Paris 11, Paris, and University of Washington, Seattle
12. Taybi H, Linder D (1967) Congenital familial dwarfism with cephaloskeletal dysplasia. Radiology 89:275–281
13. Sigaudy S, Toutain A, Moncla A, Fredouille C, Bourliere B, Ayme S, Philip N (1998) Microcephalic osteodysplastic primordial dwarfism Taybi-Linder type: report of four cases and review of the literature. Am J Med Genet 80:16–24
14. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247
15. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12–13