Journal of Human Genetics (2011) 56, 383–389

www.nature.com/jhg

ORIGINAL ARTICLE

Exploration of the disease locus by a careful evaluation of the likelihood polynomial for pedigree data

Yuki Sugaya[1] and Ritei Shibata[2]

It is demonstrated through two case studies that a careful evaluation of the likelihood polynomial results in a more accurate localization of disease locus. The evaluation of the likelihood function as a polynomial enables more flexible exploration of the disease locus. Visualization by a contour plot of the function on a unit square of paternal and maternal recombination fractions along with a superimposed ellipsoid of the Fisher information matrix helps us to find a more accurate localization of the disease locus.

Journal of Human Genetics (2011) **56,** 383–389; doi:10.1038/jhg.2011.24; published online 17 March 2011

Keywords: Fisher information; likelihood polynomial; pedigree analysis; sex-specific recombination fraction

INTRODUCTION

The aim of this paper is to demonstrate through two case studies that a careful evaluation of the likelihood polynomial yields a better result in a linkage analysis using pedigree data. A simple visualization of the likelihood on the whole region $[0,1]\times[0,1]$ of paternal and maternal recombination fractions helps us understand more accurately what the pedigree data tell us about the disease locus. Superimposing the Fisher information matrix to the contour plot also helps us to see the reliability of the estimates of the fractions. We will use the probability inheritance algorithm[1] to evaluate the likelihood as a polynomial of recombination fractions:

$$L(\theta_0,\theta_1)=\sum_{i,j}\gamma_{ij}\theta_0^i\theta_1^j,$$

where θ_0 and θ_1 are the paternal recombination fraction and the maternal recombination fraction, respectively. The introduction of different recombination fractions for male and female plays an important role when seeking a more accurate mapping of the disease locus, as is already pointed out.[2–5] The same is also demonstrated in this paper.

MATERIALS AND METHODS

Pedigree data

We have analyzed two real pedigree data to show the importance of careful evaluation of the likelihood polynomial.

Primary open-angle glaucoma data. The data used in Case study 1 is the primary open-angle glaucoma pedigree data,[6] in which the markers are placed on chromosome 5q. As in Pang *et al.*,[6] we have used disease allele frequency 0.0001 and an autosomal dominant mode of inheritance with one liability class. The penetrances for a homozygote without the disease allele, a heterozygote and a homozygote with the disease allele were set at 0, 1 and 1, respectively. The marker allele frequencies were estimated from the given data.

Familial juvenile hyperuricemic nephropathy data. The data used in Case study 2 is the familial juvenile hyperuricemic nephropathy (FJHN) pedigree data.[7] The use of the individual genotype data has been approved by the institutional ethics committees of the Keio University and the Tokyo Women's Medical University. Informed consent was obtained from each of the subjects. For FJHN, the disease gene has already been identified as uromodulin (*UMOD*; GenBank accession no. NM_003361) on chromosome 16p.[8] We have analyzed the 81 markers on chromosome 16p that are used in the analysis by Kudo *et al.*[9] The pedigree consists of 65 individuals, but only 58 descendants are analyzed because seven ancestors have no effect on the maximization of the likelihood since their marker genotypes are not available (NA). We also changed the affected status of IV-21 to non-affected since Kudo *et al.*[9] reported that it is a phenocopy. We assume that the mode of inheritance is autosomal dominant with one liability class with penetrance 0, 0.95 and 0.95. The marker allele frequencies were estimated from the given data. The disease allele frequency was assumed to be 0.0001, same as in Hart *et al.*[8]

Visual validation on a unit square

The likelihood function for pedigree data can be obtained as a polynomial by the probability inheritance algorithm.[1] The idea behind the algorithm is that the probability of affected status and marker genotypes of the ancestor is inherited to their descendants along with the inheritance of the haplotype. The likelihood is reduced generation by generation into the likelihood of the ancestor starting from a terminal sibling until it is reduced into the haplotype frequencies of the founder. The evaluation of the likelihood is then executed back to the terminal sibling. The likelihood of the ancestor is polynomial of recombination fractions; therefore, it is enough that the descendants inherit the coefficients of the polynomial. This is in contrast to the existing algorithms,[10–15] in which the likelihood has to be numerically evaluated for each value of recombination fractions. More details of this algorithm and its implementation on R are available from http://stat.math.keio.ac.jp/~sugaya/PIA/index.html. The obtained polynomial is useful for drawing the two-dimensional contour on the unit square $[0,1]\times[0,1]$ with a superimposed ellipsoid for the Fisher information matrix for each marker, as well as for finding the maximum likelihood

[1]School of Fundamental Science and Technology, Keio University, Yokohama, Japan and [2]Department of Mathematics, Keio University, Yokohama, Japan
Correspondence: Dr Y Sugaya, School of Fundamental Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa 223-8522, Japan.
E-mail: sugaya@stat.math.keio.ac.jp
Received 6 October 2010; revised 12 January 2011; accepted 7 February 2011; published online 17 March 2011

estimate of $\boldsymbol{\theta}=(\theta_0, \theta_1)$ using a Newton–Raphson type algorithm. Fisher information is helpful to know the reliability of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ even when it falls in the feasible region $[0,0.5]\times[0,0.5]$. The Fisher information matrix

$$I(\boldsymbol{\theta}) = E(-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta}))$$

at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ provides the amount of information that can be extracted from the given data using the maximum likelihood principle. We will use

$$-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad (1)$$

as an estimate of $I(\hat{\boldsymbol{\theta}})$ since it is very complicated to exactly evaluate functional $I(\boldsymbol{\theta})$. The evaluation of (1) is straightforward in our case because we have already obtained the functional form of $L(\boldsymbol{\theta})$ as a polynomial. The Fisher information is displayed together with the contour plot by an ellipsoid with axes proportional to the eigenvalues of $I(\boldsymbol{\theta})$ in the direction of the eigenvectors of $I(\boldsymbol{\theta})$. Thus, we can see that the maximum likelihood estimate is reliable if the size of the ellipsoid is relatively large in the direction of each coordinate. Thomas[16] has drawn the two-dimensional contour on the half square $[0,0.5]\times[0,0.5]$ by using a contour-drawing package CONICON3,[17] evaluating values and values of the derivatives on grid points, but the contour is an approximation of the surface. It is impossible to calculate Fisher information and to display it together with the obtained contour plot for visual validation of the likelihood.

RESULTS

Case study 1: primary open-angle glaucoma data

First we show the result by conventional linkage analysis. Figure 1 shows the curves of the LOD scores $\log_{10}(L(\theta)/L(0.5))$ for each marker, where $0\leqslant\theta=\theta_0=\theta_1\leqslant 0.5$ is assumed. Our estimates of θ are summarized in Table 1 and are more precise than those in Pang *et al.*[6] because of our functional evaluation of the likelihood. A natural consequence from these estimates is that the disease locus would be around D5S2098. In fact, the disease locus suggested by Pang *et al.* is around this marker.

However, a different picture emerges from the sex-specific linkage analysis, particularly when using the probability inheritance algorithm. The likelihood function is obtained as a polynomial of the

Table 1 Maximum likelihood estimates of the common recombination fraction for the POAG pedigree data

Marker	θ
D5S2501	0.4
D5S2051	0.5
D5S2027	0.5
D5S2065	0.2
D5S1384	0.2
D5S471	0.21
D5S503	0.2
D5S2098	0.07
D5S2011	0.18
D5S638	0.15
D5S2090	0.5
D5S2014	0.17

Abbreviation: POAG, primary open-angle glaucoma.

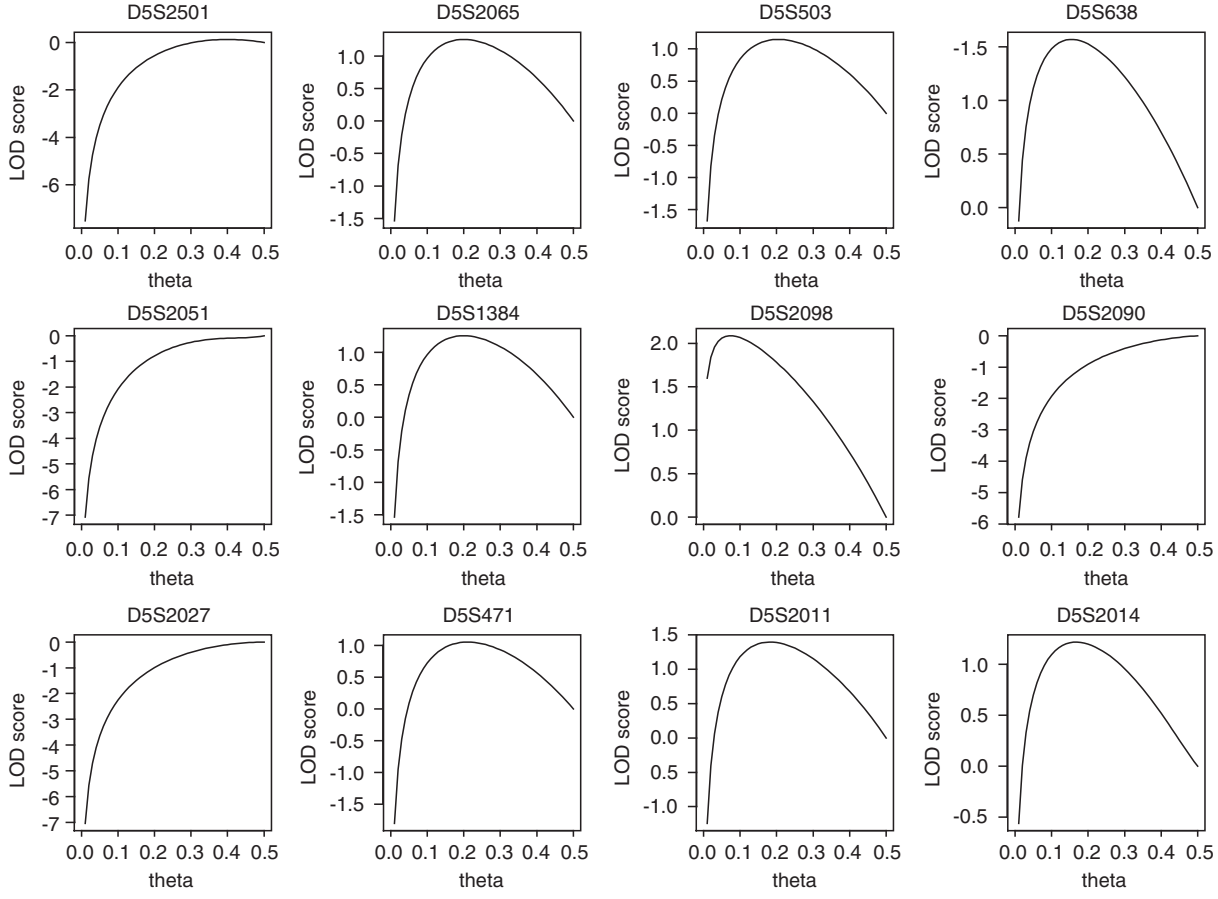

Figure 1 Curves of the LOD score for the primary open-angle glaucoma pedigree data.

paternal and maternal recombination fractions (θ_0, θ_1) between an unknown disease locus and a marker locus. The orders of each likelihood polynomial in terms of θ_0 and θ_1 are listed in Table 2 and vary with the number of homozygotes in the pedigree. The contour plots on the unit square $\{(\theta_0, \theta_1); 0 \leqslant \theta_0, \theta_1 \leqslant 1\}$ are given in Figure 2; these are arranged from the top to the bottom and from the left to the right in the order of the marker locations. We observe that such contour plots on the unit square are more informative than those on the region of feasible recombination fractions, $\{(\theta_0, \theta_1); 0 \leqslant \theta_0, \theta_1 \leqslant 0.5\}$. In fact, the maximum likelihood estimates of $\boldsymbol{\theta}=(\theta_0, \theta_1)$ exist outside the feasible region for the first three markers and the penultimate marker. Although it is unrealistic to deal with recombination fractions >0.5, such a value can appear as an estimate, particularly when penetrance values such as 0, 1 and 1 are assumed. This

Table 2 Orders of the likelihood polynomial for the POAG pedigree data

Marker	*Order of* θ_0	*Order of* θ_1
D5S2501	10	5
D5S2051	12	5
D5S2027	5	5
D5S2065	10	5
D5S1384	10	5
D5S471	10	5
D5S503	10	5
D5S2098	12	5
D5S2011	12	5
D5S638	12	5
D5S2090	3	5
D5S2014	12	5

Abbreviation: POAG, primary open-angle glaucoma.

Table 3 Maximum likelihood estimate of the paternal and maternal recombination fractions for the POAG pedigree data

Marker	θ_0	θ_1
D5S2501	NA	NA
D5S2051	NA	NA
D5S2027	NA	NA
D5S2065	0.2	0.2
D5S1384	0.2	0.2
D5S471	0.2	0.24
D5S503	0.2	0.22
D5S2098	NA	0.2
D5S2011	0.17	0.2
D5S638	0.17	0.07
D5S2090	NA	NA
D5S2014	NA	NA

Abbreviations: NA, not available; POAG, primary open-angle glaucoma.

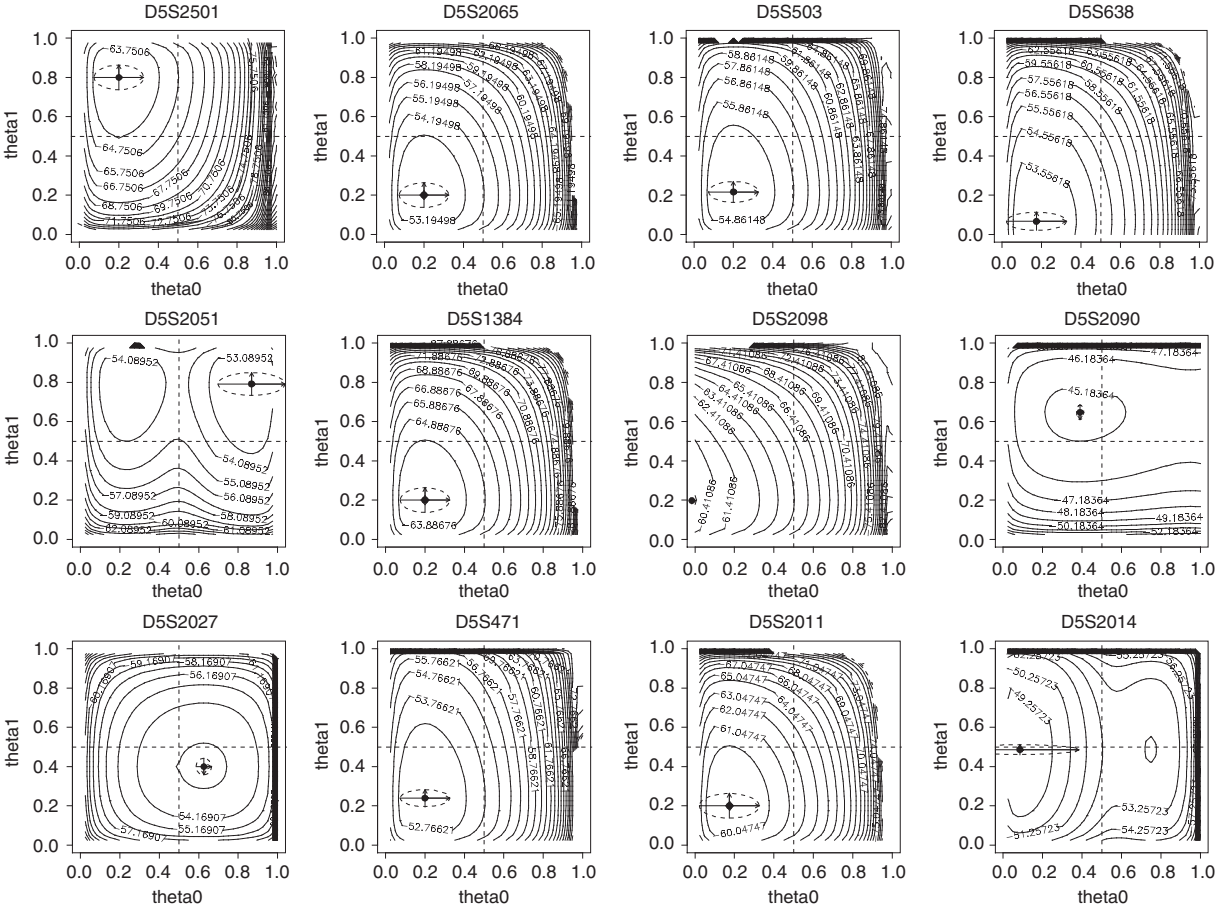

Figure 2 Two-dimensional contour plots of the log likelihood function for the primary open-angle glaucoma pedigree data in which the maximum likelihood estimate is shown by the center of a Fisher information ellipsoid on each plot.

Table 4 Maximum LOD score for each marker in the FJHN pedigree data

Marker	θ	*Maximum LOD score*
D16S292	0.2	0.58
GATA42E11	0.81	0.42
FJHN-A	0.21	0.67
D16S683	0	1.3
D16S3103	0	0.93
D16S3017	0.39	0.04
D16S3056	0	0.72
#33rp3	0	0.29
FJHN-B	0.09	1.2
#120	0.1	0.91
#118	0.05	3.4
#32rp3	0.09	0.8
#116	0.1	0.95
#241	0	1.19
#238	0	5.54
D16S3041	0	2.2
#236	0	1.02
#233	0	1.09
D16S3036	0	4.79
#122	0	1.78
#30rp3	0	1.62
#123	0	4.81
(*UMOD*)	—	—
D16S773	0	3.92
ac024562a21	0	2.43
ac024562a10	0	3.03
D16S3046	0	4.26
D16S772	0	5.55
D16S403	0	2.45
D16S412	0	3.47
#128	0	2.94
#129	0	0.36
D16S3130	0	1.43
ac002400a4	0	2.52
ac002400a1	0	0.41
ac008870b2	0	1.91
ac008870b1	0	1.6
D16S417	0	0.57
ac002302a4	0	5.84
ac024562a24	0	0.3
ac002299a3	0.38	0.03
ac002299a4	0.06	0.5
D16S420	0	1.8
ac002299a6	0	1.5
ac004125b1	0.02	2.04
ac004125a3	0.16	0.35
ac004125a6	0	2.16
ac004125a9	0	0.33
ac008938a1	0	0.91
ac008938a2	0	2.06
#209	0	0.91
D16S3113	0	4.37
#62rp3	0.03	1.32
D16S401	0	4.02
#215	0	2.4
#216	0.27	0.12
D16S3133	0	2.21
ac008731a2	1	0.27
#217	0	0.89
ac008741a2	0	0.95
ac008741b2	0	1.88
ac008741a6	0	1.03
D16S3068	0.54	0.01
#179	0.31	0.04
#180	0	1.34
ac092141a1	0.09	0.42
ac092141a2	0	1.83
D16S537	0.15	0.92
D16S3131	0.11	1.41
D16S769	0.57	0.01
#226	0.76	0.09
D16S296	0	0.93
#228	0	1.1
#229	0	1.29
#221	0.15	0.36
#63rp3	0.24	0.13
D16S3100	0.44	0.01
D16S3093	0.1	2.17
D16S297	0	0.06
D16S3145	0.4	0.03
#204	0.16	0.2
D16S3225	0.24	0.25

Abbreviations: FJHN, familial juvenile hyperuricemic nephropathy; NA, not available.

Table 5 Orders of the likelihood polynomial for the FJHN pedigree data

Marker	*Order of* θ_0	*Order of* θ_1
D16S292	12	17
D16S3017	13	14
FJHN-B	16	18
#120	12	13
#118	20	25
#32rp3	14	17
#116	12	13
#241	16	13
#238	25	21
D16S3036	24	20
#123	25	21
D16S773	25	22
D16S3046	23	17
D16S772	23	25
#129	15	12
ac002302a4	24	20
ac004125b1	15	16
D16S3113	24	23
#62rp3	16	17
D16S401	24	22
D16S3133	20	22
#217	14	13
ac008741a6	12	15
#180	15	16
ac092141a1	10	15
ac092141a2	14	15
D16S537	13	15
D16S3131	14	16
D16S296	11	15
#221	16	18
#63rp3	14	11
D16S3093	21	22

Abbreviation: FJHN, familial juvenile hyperuricemic nephropathy.

compels us to focus on the inheritances from the heterozygote of the affected individuals in terms of the underlying marker. For example, the two affected females are only focused on for the first and

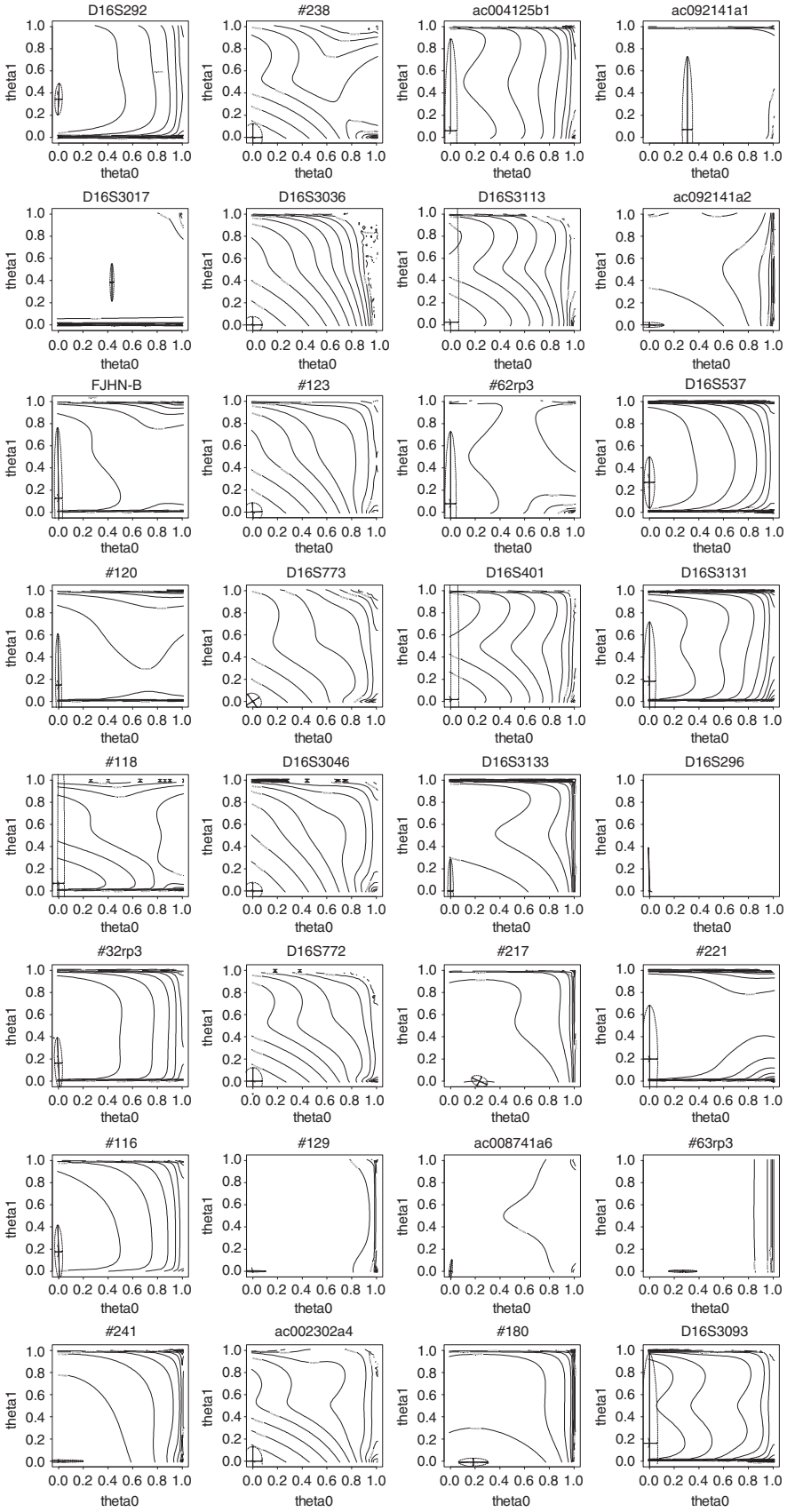

Figure 3 Two-dimensional contour plots of the likelihood function for the familial juvenile hyperuricemic nephropathy (FJHN) pedigree data.

Table 6 Summary of the result for the FJHN pedigree data

Marker	*Position (Kb)*	θ_0	θ_1	*Maximum LOD score*
D16S292	13447	NA	0.34	0.96
D16S3017	16760	NA	0.38	0.04
FJHN-B	18632	NA	0.13	1.37
#120	18632	NA	0.15	1.08
#118	18692	NA	0.07	3.53
#32rp3	18698	NA	0.16	0.99
#116	18710	NA	0.18	1.15
#241	18771	0	NA	1.19
#238	18811	0	0	5.54
D16S3036	18954	0	0	4.79
#123	19104	0	0	4.81
(*UMOD*)	19750	—	—	—
D16S773	20689	0	0	3.92
D16S3046	20895	0	0	4.26
D16S772	20980	0	0	5.55
#129	23358	0	NA	0.36
ac002302a4	24039	0	0	5.84
ac004125b1	24360	NA	0.06	2.12
D16S3113	24697	NA	0.02	4.4
#62rp3	24715	NA	0.08	1.42
D16S401	24785	NA	0.02	4.04
D16S3133	NA	NA	0	2.21
#217	25017	0.23	0	0.97
ac008741a6	25436	NA	0	1.03
#180	25810	0.18	NA	1.44
ac092141a1	25825	NA	0.07	0.47
ac092141a2	25885	0	NA	1.83
D16S537	25897	NA	0.27	1.38
D16S3131	26102	NA	0.18	1.75
D16S296	26593	NA	0	0.93
#221	26657	NA	0.2	0.54
#63rp3	26658	0.27	NA	0.18
D16S3093	NA	NA	0.16	2.57

Abbreviations: FJHN, familial juvenile hyperuricemic nephropathy; NA, not available.

penultimate markers, and as a result, the pattern of disease inheritance significantly increases the estimate of θ_1 toward 1. The same also happens for the second marker where the affected males and females are focused on, and also for the third marker where the affected males are focused on. Hence, it would be natural to set the estimates of both θ_0 and θ_1 as NA if one of them exceeds 0.5, since nothing definite can be said from such markers. It is worth noting that the likelihood may take its maximum at θ_0=0.5 or θ_1=0.5 for such markers if (θ_0, θ_1) is restricted to the half-square $[0,0.5]\times[0,0.5]$ as in Thomas,[16] and a misleading answer may be obtained that the disease locus is far from the underlying marker.

Among the estimates of $\boldsymbol{\theta}$ for non-NA markers 4–10, the estimate $\hat{\theta}_0$ for marker 8 is unreliable from the viewpoint of Fisher information and hence it is also regarded as NA. Further, for the last marker, Fisher information is very small for the maternal recombination fraction, although the estimated recombination fractions are 0.08 and 0.49.

Table 3 summarizes the above information. We can see from Table 3 that the estimated recombination fractions are 0.20 or 0.17 except for the marker D5S638. This observation suggests that the disease locus would not be in the region between the given markers D5S2065 and D5S2011. This region seems to be forming a block that inherits as a whole, generation by generation. In fact, the disease gene *WDR36* (GenBank accession no. NM_139281) identified by Monemi *et al.*[18] is outside of the region in the centromeric direction. This case study shows the importance of carefully looking at a two-dimensional picture of the likelihood function.

Case study 2: FJHN data

Table 4 gives the list of the estimated recombination fractions θ with the LOD scores for all 81 markers. The markers with small θ and high LOD scores (>3) are scattered over. It shows that the conventional procedure does not provide a clear indication of the disease locus.

Only 32 markers remained as 'non-NA markers' after the same criterion is applied as in the previous case study. The orders of all likelihood polynomials for such markers are listed in Table 5. The contour plots for these 32 markers are shown in Figure 3. The plots are arranged from the top to the bottom and from the left to the right in the order of the marker locations. Table 6 gives us a numerical summary. *UMOD* is also listed in the first column as a reference. The second column in the table indicates the physical positions of the markers on chromosome 16p in the kilobase pairs. It is easily seen from Table 6 that the estimates of both θ_0 and θ_1 are non-NA and that the LOD scores are high for a block from marker #238 to marker ac002302a4, with the exception of #129. This observation and the fact that both the estimated values are 0, except for marker #129 suggest that the disease locus exists in this block. In fact, the disease locus *UMOD* resides in the middle of this block. It was hard to identify this block using Table 4 obtained from the conventional procedure.

DISCUSSION

We have demonstrated that a careful validation of the likelihood provides a more reliable result. For the validation to be effective, a functional evaluation of the likelihood is useful and the contour plot of the likelihood on the unit square $[0,1]\times[0,1]$ of paternal and maternal recombination fractions is helpful. An overplotted ellipsoid of the Fisher information matrix is also useful to rule out any unreliable estimate. Validation of this method by any other sets of pedigree data will be reported together with further application to multipoint linkage analysis elsewhere, even though the two-point linkage analysis is enough to localize the disease locus in our case studies.

1 Sugaya, Y. & Shibata, R. Probability inheritance algorithm and its application. *The 52nd Annual Meeting of the Japan Society of Human Genetics*, The Japan Society of Human Genetics, Tokyo, Japan, 115 (2007).
2 Daw, E. W., Thompson, E. A. & Wijsman, E. M. Bias in multipoint linkage analysis arising from map misspecification. *Genet. Epidemiol.* **19,** 366–380 (2000).
3 Wu, R., Xing, M. C., Wu, S. S. & Zeng, Z. B. Linkage mapping of sex-specific differences. *Genet. Res.* **79,** 85–96 (2002).
4 Feenstra, B., Greenberg, D. A. & Hodge, S. E. Using LOD scores to detect sex differences in male-female recombination fractions. *Hum. Hered.* **57,** 100–108 (2004).
5 Fingerlin, T. E., Abecasis, G. R. & Boehnke, M. Using sex-average genetic maps in multipoint linkage analysis when identity-by-decent status is incompletely known. *Genet. Epidemiol.* **30,** 384–396 (2006).
6 Pang, C. P., Fan, B. J., Canlas, O., Wang, D. Y., Dubois, S., Tam, P. O. *et al.* A genome-wide scan maps a novel juvenile-onset primary open angle glaucoma locus to chromosome 5q. *Mol. Vis.* **12,** 85–92 (2006).
7 Kamatani, N., Moritani, M., Yamanaka, H., Takeuchi, F., Hosoya, T. & Itakura, M. Localization of a gene for familial juvenile hyperuricemic nephropathy causing under-excretion-type gout to 16p12 by genome-wide linkage analysis of a large family. *Arthritis Rheum.* **43,** 925–929 (2000).
8 Hart, T. C., Gorry, M. C., Hart, P. S., Woodard, A. S., Shihabi, Z., Sandhu, J. *et al.* Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy. *J. Med. Genet.* **39,** 882–892 (2002).
9 Kudo, E., Kamatani, N., Tezuka, O., Taniguchi, A., Yamanaka, H., Yabe, S. *et al.* Familial juvenile hyperuricemic nephropathy: detection of mutations in the uromodulin gene in five Japanese families. *Kidney Int.* **65,** 1589–1597 (2004).
10 Elston, R. C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21,** 523–542 (1971).

11 Ott, J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage study. *Am. J. Hum. Genet.* **26,** 588–597 (1974).
12 Lange, K. & Elston, R. C. Extensions to pedigree analysis. *Hum. Hered.* **25,** 95–105 (1975).
13 Cannings, C., Thompson, E. A. & Skolnick, M. H. Probability functions on complex pedigrees. *Appl. Prob.* **10,** 26–61 (1978).
14 Lander, E. S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Genetics* **84,** 2363–2367 (1987).
15 Fishelson, M. & Geiger, D. Exact genetic linkage computations for general pedigrees. *Bioinfomatics* **18,** 189–198 (2002).
16 Thomas, A. Gene hunting with gradients of likelihoods. *J. R. Statist. Soc. B* **53,** 3–26 (1991).
17 Sibson, R. CONICON 3 handbook (*University of Bath* 1987).
18 Monemi, S., Spaeth, G., Dasilva, A., Popinchalk, S., Illitchev, E., Liebmann, J. *et al.* Identification of a novel adult-onset primary open-angle glaucoma (POAG) gene on 5q22.1. *Hum. Mol. Genet.* **14,** 725–733 (2005).