Computer simulation of linkage and heterogeneity in tuberous sclerosis: a critical evaluation of the collaborative family data

L A J Janssen, L A Sandkuijl, J R Sampson, D J J Halley

Abstract

The existence of locus heterogeneity for a genetic disease may complicate linkage studies considerably, especially when very few large families with the disease are available. In this situation a modest collection of families is unlikely to be sufficient for successful localisation of one or more disease genes. Recently, eight research groups working on tuberous sclerosis (TSC) brought together linkage data pertaining to the candidate chromosomes 9, 11, and 12 for a large group of families. In a series of simulation studies we determined the probability of detecting linkage and linkage heterogeneity in this set of families.

On average TSC families are very small; in most cases there are fewer than two informative meioses. The size distribution of chromosome 9 linked families was similar to that of non-linked families. This indicates that a dramatic difference in the clinical severity of major genetic forms of TSC is unlikely.

The results of our simulation studies show that this set of families can generate highly significant evidence for linkage and heterogeneity. When two TSC genes are equally common, the strongest evidence for linkage and heterogeneity could be obtained using a method based on the incorporation of multiple candidate regions in a single analysis, with an average lod score of 24.27.

(*J Med Genet* 1992;29:867-74)

under locus heterogeneity, it is necessary to identify the subset of families that show linkage to a given chromosomal region. For large families that can generate significant evidence for linkage when studied individually this can be done directly. When the average family size is small, however, this distinction cannot be made in a straightforward manner, and it is intuitively clear that a large number of families will be required for detection of one or several of the responsible genes. This has been confirmed in theoretical studies using computer simulation and analytical methods.⁴⁵

Various statistical methods have become available to make optimal use of the linkage information in a series of (relatively) small families. The most commonly applied test is the admixture test,⁶⁷ which involves simultaneous estimation of the location(s) of the responsible genes and of the proportion of families segregating for each of these genes. This method has been instrumental in detecting several loci causing retinitis pigmentosa on the X chromosome.⁸

For tuberous sclerosis, a neurocutaneous disorder characterised by widespread hamartosis, linkage studies have yielded conflicting results. While some studies provided evidence for a locus close to the ABO blood group gene on chromosome 9,^{9,10} other studies could not confirm this linkage, and indicated chromosomes 11 and 12 as possible alternative locations.^{11,12} Considerable efforts have been made to combine linkage data from the groups participating in the TSC consortium and to analyse the data with a variety of statistical methods. Unfortunately, these studies have not entirely resolved the controversy, although consistent evidence has emerged for a TSC1 gene on the long arm of chromosome 9 and for the existence of locus heterogeneity.¹³⁻¹⁵

For the simultaneous evaluation of the different chromosomal regions that may be involved in TSC, we have developed an extension of the admixture approach,^{67,16} which we have previously called the imaginary chromosome (IC) approach. In a single analysis, this method will evaluate linkage results for all relevant chromosomal regions. For a given family, this allows simultaneous evaluation of positive evidence for one of the regions and negative linkage information for the other regions. While this approach has the advantage of making maximal use of all available information, it shares with some other statistical methods the disadvantage of not being transparent. Thus, it is not immediately apparent

Department of Clinical Genetics, Academic Hospital Rotterdam Dijkzigt and Erasmus University Rotterdam, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.
L A J Janssen
L A Sandkuijl
D J J Halley

Institute of Medical Genetics, University Hospital of Wales, Cardiff CF4 4XN.
J R Sampson
L A Sandkuijl

Correspondence to Dr Janssen.

Received 4 August 1992.
Revised version accepted 30 September 1992.

During the past decade, linkage analysis has been successfully applied to localise the genes responsible for many different inherited diseases.¹ The availability of a large series of highly polymorphic markers throughout the genome has considerably facilitated linkage mapping. For a number of diseases, however, linkage results have been reported that have not been confirmed by further studies, as was initially the case for Charcot-Marie-Tooth disease, and is still the case for a number of psychiatric disorders including schizophrenia and manic depression.²³ Locus heterogeneity has been suggested as a possible explanation for such differing results. While it is commonly recognised that locus heterogeneity may complicate linkage studies, it has not precluded the accurate localisation of major genes for polycystic kidney disease and, eventually, Charcot-Marie-Tooth disease. For mapping

how much evidence a given data set can potentially yield in an analysis of linkage and heterogeneity which includes several chromosomal regions simultaneously. Also, it is not entirely clear what level of evidence is to be regarded as 'significant', the criterion that researchers are most interested in. Ott⁷ suggested that the likelihood ratio favouring linkage or heterogeneity should always be reported, leaving the decision about whether a certain likelihood ratio is to be regarded as significant to the individual researchers. Indeed, uniform guidelines are hard to define, as the statistical behaviour of combined tests of linkage and heterogeneity in multipoint linkage analysis has not been adequately investigated.

Recently, we have analysed the collaborative TSC data set, after rigorous checking for data errors, and after reassessment of diagnoses following uniform criteria. The results, presented in an accompanying paper,¹⁷ indicate existence of a major locus (*TSC1*) on chromosome 9 and at least one other locus elsewhere in the genome. Chromosome 12 might harbour another TSC gene of minor importance. To help interpret these results, we carried out extensive computer simulations, using both the conventional admixture test and its multi-locus extension. In the current study we will address the following questions. How much linkage information is potentially present in the combined TSC families, and how is that information distributed over the families? Is there any difference in effective size between the families that show linkage to the chromosome 9 markers and those that yield negative lod scores? How much evidence for linkage and heterogeneity may one expect to find under various degrees of locus heterogeneity in a data set of this size? The answers to these questions can provide both information about the power and size of the collaborative TSC data set and insight into the performance of the admixture test when using multiple candidate regions.

Materials and methods

FAMILY MATERIAL

In our studies we used 128 families from eight different centres: Irvine, Boston, Houston, Durham (USA), Cardiff, London (UK), Erlangen (Germany), and Rotterdam (The Netherlands). This set of families is identical to that used in the collaborative linkage study.¹⁷ Almost all families used in this study have been described before,^{11 12 16 18-22} although phenotypic data have been fully reviewed with amendment of every subject's affection status for these studies.¹⁷

For our simulation studies we classified all subjects for whom actual marker tests had been carried out as 'available', others were classified as 'unavailable' for DNA analysis. Apart from family structure, diagnostic assessment, and availability, no other family data were used.

COMPUTER SIMULATION

In our simulation studies of linkage and heterogeneity, three separate steps can be distinguished.

(1) Preparation of hypothetical marker data for all persons for whom DNA samples were available. These marker data were generated using the computer program SLINK,²³ which takes the disease status of all family members into account. Penetrance and gene frequency were as used in the collaborative linkage study.¹⁷ We assumed that a hypothetical marker was located at 5% recombination from each of the TSC genes. For each family, 100 distinct replicates were made, each with new hypothetical marker data. Three levels of informativeness were simulated: eight alleles (PIC value 0.86), four alleles (PIC value 0.7), and two alleles (PIC value 0.375). In a separate series of simulations, we generated data for markers with similar informativeness, but unlinked to any of the TSC genes.

From the simulated data for each individual family, simulated versions of the entire data set were created by selecting for each family a linked or unlinked replicate with probabilities α and $1-\alpha$ respectively (where α denotes the assumed proportion of families linked to a particular region).

(2) The lod score calculations, by regular analysis of linkage using the generated marker data. Lod scores were calculated for each replicate, varying the recombination frequency from 0.0 to 0.5 in steps of 0.01. Lod score calculations on these replicates were performed batchwise, using the MLINK option of the LINKAGE package (version 5.03). In these calculations, we applied the same gene frequency as in the collaborative linkage analysis ($1 \cdot 10^{-4}$),¹⁷ while no allowance was made for new mutations.

(3) Heterogeneity analysis using single locus and multilocus versions of the admixture test. All admixture tests that we applied are based on the assumption that disease genes may exist at two or more locations in the genome, but that penetrance and mutation frequency are equal for all genes involved. Lod scores calculated in the previous step were transferred to a slightly modified version of the HOMOG program.⁷ For simultaneous evaluation of several chromosomal regions the HOMOG2 program was used. The necessary input was created by appending a list of lod scores for an unlinked replicate of a given family to a list of lod scores for a linked replicate of the same family or vice versa (probabilities α and $1-\alpha$ respectively). Thereby we created an input file containing two chromosomal regions, each represented by one marker. Each family was linked to either the first or the second region. Together the regions formed a so called 'imaginary chromosome'.

INFORMATION CONTENT OF FAMILIES

In order to get a more precise measure of the usefulness of each family for linkage studies, we calculated the mean lod score (expected lod

score) at 5% recombination over the 100 replicates in simulations of an eight allele marker. The resulting mean $Z(\theta=0.05)$ was divided by 0.215 (the expected lod score at $\theta=0.05$ for one completely informative meiosis). The quotient was termed the 'effective number of informative meioses' (EFNIM), since it enables us to compare families with completely different structures. This measure is closely related to Edwards's 'equivalent observations'²⁴ (appendix).

Results

POWER OF TSC FAMILIES TO DETECT LINKAGE UNDER HOMOGENEITY

We estimated the power to detect linkage under homogeneity by simulating markers with low and high informativeness. We specified a recombination frequency of 5% for the following reasons. (1) The chromosome 9 markers most closely linked to the TSC1 locus show approximately 5% recombination,¹⁷ and (2) in a random genome search it is frequently attempted to test markers that divide the genome into 10 cM intervals, which implies a frequency of recombination between the disease gene and the closest marker of at most 5%.

By simulating a two allele marker we learned what values of the lod score ($Z_{\max}$) might be expected if only regular RFLPs (PIC 0.37) were used for linkage mapping. Over 100 distinct simulations of the entire data set, the two allele marker gave a mean overall lod score of 16.8. The highest lod score value was 24.8, while $Z_{\max}$ exceeded 10.5 in 95% of all replicates.

Simulations on an eight allele marker gave us insight into the lod scores that could be obtained with dinucleotide repeats and other highly informative markers. A similar amount of informativeness can be expected from a map of two, four, and five allele markers in a multipoint analysis as performed on the actual data of the TSC collaborative group.¹⁷

The highest lod score obtained with the eight allele marker was 56.7, the mean $Z_{\max}$ was 41.8, and at least 95% of all replicates showed a $Z_{\max}$ of 33.6 or higher.

Rather than comparing individual families by their mean lod scores, we decided to describe the families in terms of their 'effective number of informative meioses' (EFNIM). The EFNIM of a family represents the number of informative meioses one may expect to score on average in that family for a marker system with given informativeness. The $Z(\theta=0.05)$ values and EFNIM values for all families are presented in the table (A and B).

EFNIM values for the chromosome 9 linked group of TSC1 families were compared with those obtained for non-chromosome 9 linked families. Assignment of families to either of these two groups was made according to the posterior probabilities for linkage to chromosome 9 as determined in the accompanying study¹⁷ using real linkage data for chromosome 9 markers. The distribution of the families

over 11 EFNIM categories of increasing informativeness is shown in fig 1. There are no marked differences in size (as summarised by EFNIM) between the two groups of TSC families. Another obvious finding is the paucity of large families in both groups. Most families are in the 0–1 EFNIM category. This is illustrated by the mean EFNIM values for both family groups: 1.41 for the chromosome 9 linked families and 1.46 for the non-chromosome 9 linked group.

EXCLUSION POWER OF TSC FAMILIES

In general, exclusion studies are only valid when the mode of inheritance is specified correctly in the analysis. In a real exclusion study for TSC one would therefore have to take into account the locus heterogeneity. In our simulation study we evaluated the power of the families for an uncomplicated exclusion study (that is, under linkage homogeneity), with the sole purpose of comparing the families. As expected, families that contained much information for linkage detection also contributed most information for exclusion (table A and B).

Together, the TSC1 families had the power to exclude linkage over large genetic distances. When an eight allele marker was examined, 95% of the replicates excluded 31 cM or more on either side of the marker. With a two allele marker the exclusion distance decreased to 12 cM.

The non-chromosome 9 linked families were able to exclude similar areas: at least 29 cM for an eight allele marker and at least 11 cM for a two allele marker in 95% of the replicates.

POWER OF TSC FAMILIES TO DETECT HETEROGENEITY

We tested the performance of the families in heterogeneity analyses. We analysed HOMOG input files (containing lod scores from linked and unlinked families) and HOMOG2 input files (containing lod scores from two regions, with each family linked to only one of these). Thus we tried to answer two questions. (1) What is the power of this data set for detection of heterogeneity? (2) What is gained by including information from the alternative region in a two locus heterogeneity analysis?

The interpretation of results of linkage analysis in terms of statistical significance is not always simple. Under homogeneity a lod score of 3.0 or more is accepted as 'significant' evidence for linkage. A lod score of 3.0 corresponds to an odds ratio of 1000:1. This high threshold has a statistical basis ($p < 0.05$ is usually regarded as significant evidence): when two loci are selected at random the chances that they will show linkage are small. It has been estimated that the prior probability of two loci showing true linkage is only 1 in 50. If the theoretical odds ratio of 1000:1 is corrected for the low (1 in 50) prior expectation of finding linkage, the resulting frequency of false positive linkages will be 0.05.

Results of simulation studies presented as lod scores, EFNIM values, and percentages of all replicates. B = Boston, C = Cardiff, L = London, I = Irvine, R = Rotterdam, E = Erlangen, D = Durham, H = Houston. $Z(\max)$ = lod score maximised over θ , $Z(\theta = 5\%)$ = lod score at a recombination fraction of 0.05, EFNIM = effective number of informative meioses (see text).

(A) The power of the chromosome 9 linked (TSC1) families

Family No	Original Family No	8 allele marker at $\theta = 5\%$		EFNIM	2 allele marker at 5%	Unlinked 8 allele marker ($\theta = 50\%$)	
		Max $Z(\max)$	Mean $Z(\theta = 5\%)$		Mean $Z(\theta = 5\%)$	Mean $Z(\theta = 5\%)$	Power to exclude > 5 cM (%)
1	B 1	1.35	0.41	1.92	0.18	-0.46	0
2	B 2	0.00	0.00	0.00	0.00	0.00	0
5	B 5	0.70	0.29	1.36	0.12	-0.28	0
8	B 8	1.38	0.42	1.96	0.15	-0.49	0
9	B 9	0.70	0.19	0.89	0.10	-0.25	0
12	B 12	0.73	0.20	0.95	0.07	-0.29	0
14	B 14	0.00	0.00	0.00	0.00	0.00	0
20	B 20	2.22	0.70	3.26	0.31	-0.73	0
21	B 21	0.36	0.10	0.45	0.04	-0.08	0
22	B 22	0.23	0.04	0.18	0.02	-0.05	0
1006	C 6	0.25	0.09	0.42	0.03	-0.05	0
1009	C 9	0.25	0.09	0.40	0.02	-0.08	0
1010	C 10	0.25	0.12	0.57	0.03	-0.07	0
1012	C 12	0.77	0.14	0.67	0.04	-0.08	0
1013	C 13	2.89	1.55	7.20	0.59	-1.53	46
1015	C 15	0.48	0.10	0.49	0.03	-0.17	0
2005	L 5	0.25	0.10	0.48	0.04	-0.09	0
2008	L 8	2.24	0.95	4.40	0.32	-1.07	15
2009	L 9	0.57	0.22	1.01	0.11	-0.26	0
2010	L 10	0.25	0.08	0.38	0.02	-0.09	0
2011	L 11	0.57	0.22	1.04	0.11	-0.28	0
2014	L 14	0.44	0.20	0.92	0.04	-0.21	0
2015	L 15	0.06	0.00	0.00	0.00	-0.00	0
2016	L 16	0.95	0.21	0.97	0.09	-0.35	0
2018	L 18	1.43	0.74	3.42	0.28	-0.83	4
2023	L 23	0.00	0.00	0.00	0.00	0.00	0
2026	L 26	0.70	0.24	1.10	0.08	-0.29	0
3001	I 1	1.58	0.82	3.80	0.31	-0.89	15
3011	I 11	1.16	0.62	2.88	0.21	-0.63	6
3015	I 15	3.23	1.34	6.25	0.47	-1.27	25
3019	I 19	0.24	0.08	0.37	0.03	-0.05	0
3021	I 21	0.24	0.03	0.15	0.01	-0.07	0
3026	I 26	1.60	0.54	2.52	0.10	-0.88	15
3028	I 28	0.25	0.11	0.53	0.05	-0.08	0
3029	I 29	0.00	0.00	0.00	0.00	0.00	0
3033	I 33	0.25	0.09	0.40	0.01	-0.10	0
3101	I 101	0.86	0.34	1.60	0.12	-0.36	0
4046	R 2046	0.00	0.00	0.00	0.00	0.00	0
4067	R 2067	0.25	0.06	0.29	0.02	-0.12	0
4068	R 2068	0.00	0.00	0.00	0.00	0.00	0
4077	R 2077	1.27	0.50	2.31	0.17	-0.55	0
4219	R 1219	2.12	0.78	3.61	0.24	-0.73	5
4221	R 1221	0.83	0.15	0.71	0.04	-0.05	0
4222	R 1222	1.00	0.08	0.36	0.02	-0.01	0
4264	R 1264	0.99	0.27	1.27	0.13	-0.25	0
4467	R 1467	0.04	0.00	0.00	0.00	-0.01	0
5080	E 2080	0.02	0.00	0.00	0.00	-0.00	0
5159	E 3159	0.46	0.07	0.35	0.03	-0.03	0
5240	E 2024	2.03	1.07	4.97	0.37	-0.95	20
5452	E 1452	0.00	0.00	0.00	0.00	0.00	0
5733	E 3733	0.00	0.00	0.00	0.00	0.00	0
7430	D 430	0.88	0.31	1.44	0.11	-0.32	0
7474	D 474	1.45	0.82	3.81	0.40	-0.77	14
7479	D 479	0.99	0.31	1.45	0.08	-0.37	0
7781	D 781	0.25	0.04	0.20	0.02	-0.07	0
7873	D 873	1.57	0.81	3.78	0.37	-0.85	9
7981	D 981	1.99	0.64	2.99	0.24	-0.83	11
9004	H 4	3.83	1.91	8.89	0.84	-2.02	59
9005	H 5	0.57	0.23	1.07	0.05	-0.24	0
9006	H 6	0.06	0.00	0.02	0.00	-0.00	0
9007	H 7	0.83	0.25	1.17	0.09	-0.32	0
9008	H 8	0.25	0.07	0.32	0.02	-0.10	0
9010	H 10	0.07	0.00	0.01	0.00	-0.00	0

When one is examining the possibility of locus heterogeneity, however, no prior expectations can be formulated about whether a disease will show locus heterogeneity or not. Therefore lod scores of 0.834, approximately corresponding to a χ^2 of 3.841 ($p = 0.05$ at 1 df), have been conventionally accepted as 'significant'. The situation changes again, when examining the possibility that the presumed second locus is also located within the region for which markers were tested. The prior expectation for the second locus being in the tested region is low. Therefore, a threshold lod score difference of 3.0 seems a prudent choice. Accordingly, we formulated the following tests of various hypotheses.

H0. No locus exists in the area. (Null hypothesis. The lod score is 0.0 by definition.)

H1. One locus maps in the area tested and there is no heterogeneity. A lod score of at least 3.0 is required.

H2_(1 mapped). One locus maps in the area, heterogeneity exists, but it is assumed that the second locus is not located in the tested region. Both the null hypothesis and the H1 hypothesis should be rejected in favour of this alternative when the lod score exceeds that for H1 by at least 0.834 and the lod score exceeds that for H0 by 3.834 ($= 3.0 + 0.834$).

H2_(2 mapped). Two loci exist in the tested area. There are three requirements for significance. There must be a lod score difference with

(B) The power of the non-chromosome 9 linked families

Family No	Original Family No	8 allele marker at $\theta = 5\%$			2 allele marker at 5%		Unlinked 8 allele marker ($\theta = 50\%$)	
		Max Z(max)	Mean Z($\theta = 5\%$)	EFNIM	Mean Z($\theta = 5\%$)	Mean Z($\theta = 5\%$)	Power to exclude > 5 cM (%)	
3	B 3	0.36	0.07	0.34	0.02	-0.09	0	
6	B 6	0.05	0.00	0.00	0.00	-0.00	0	
7	B 7	0.00	0.00	0.00	0.00	0.00	0	
10	B 10	0.87	0.24	1.10	0.11	-0.25	0	
15	B 15	0.25	0.14	0.63	0.03	-0.11	0	
16	B 16	0.04	0.00	0.00	0.00	-0.00	0	
18	B 18	0.48	0.11	0.51	0.04	-0.13	0	
19	B 19	0.04	0.00	0.01	0.00	0.00	0	
1003	C 3	0.73	0.35	1.65	0.11	-0.36	0	
1004	C 4	1.58	0.87	4.07	0.27	-0.84	4	
1005	C 5	0.36	0.05	0.25	0.01	-0.10	0	
1007	C 7	0.56	0.09	0.41	0.03	-0.13	0	
1011	C 11	0.48	0.15	0.70	0.08	-0.08	0	
2001	L 1	6.74	3.17	14.76	1.30	-3.75	90	
2002	L 2	0.04	0.00	0.00	0.00	0.00	0	
2003	L 3	0.57	0.26	1.22	0.09	-0.28	0	
2004	L 4	0.36	0.04	0.20	0.00	-0.10	0	
2006	L 6	0.36	0.10	0.46	0.04	-0.15	0	
2007	L 7	2.63	1.34	6.23	0.66	-1.93	55	
2012	L 12	0.68	0.23	1.06	0.09	-0.13	0	
2013	L 13	0.48	0.15	0.70	0.06	-0.12	0	
2017	L 17	0.25	0.06	0.29	0.01	-0.11	0	
2019	L 19	0.57	0.28	1.30	0.12	-0.26	0	
2020	L 20	2.03	1.03	4.82	0.42	-1.08	18	
2021	L 21	0.25	0.05	0.23	0.01	-0.12	0	
2022	L 22	0.46	0.06	0.26	0.02	-0.09	0	
2024	L 24	0.70	0.32	1.49	0.14	-0.27	0	
2025	L 25	1.08	0.34	1.59	0.16	-0.29	0	
3003	I 3	1.56	0.74	3.46	0.31	-0.89	11	
3004	I 4	0.70	0.26	1.19	0.08	-0.17	0	
3008	I 8	0.67	0.15	0.69	0.05	-0.09	0	
3016	I 16	1.28	0.63	2.95	0.24	-0.67	7	
3020	I 20	0.42	0.08	0.36	0.03	-0.05	0	
3024	I 24	0.46	0.04	0.18	-0.01	-0.15	0	
3034	I 34	0.06	0.00	0.01	0.00	-0.00	0	
4079	R 2079	0.87	0.44	2.04	0.15	-0.39	0	
4197	R 1197	0.04	0.00	0.00	0.00	-0.00	0	
4223	R 1223	0.25	0.11	0.49	0.04	-0.06	0	
5024	E 2024	0.70	0.24	1.10	0.07	-0.42	0	
5246	E 2246	0.25	0.06	0.29	0.02	-0.15	0	
5383	E 3383	0.49	0.16	0.73	0.07	-0.03	0	
5449	E 1449	0.00	0.00	0.00	0.00	0.00	0	
5459	E 1459	0.00	0.00	0.00	0.00	0.00	0	
5462	E 3462	0.00	0.00	0.00	0.00	0.00	0	
5465	E 1465	2.19	0.84	3.92	0.30	-1.01	22	
5468	E 3468	0.00	0.00	0.00	0.00	0.00	0	
5865	E 865	0.64	0.07	0.32	0.05	-0.06	0	
5866	E 4865	0.57	0.23	1.05	0.03	-0.31	0	
7085	D 1085	0.02	0.00	0.00	0.00	-0.00	0	
7111	D 1111	2.29	1.12	5.19	0.44	-1.15	23	
7188	D 1188	0.82	0.33	1.51	0.10	-0.35	0	
7427	D 427	1.11	0.40	1.86	0.22	-0.37	0	
7431	D 431	0.83	0.12	0.57	0.04	-0.06	0	
7437	D 347	3.72	1.97	9.16	0.68	-2.37	64	
7473	D 473	0.48	0.08	0.35	0.01	-0.16	0	
7482	D 482	0.09	0.00	0.01	0.00	0.00	0	
7507	D 507	1.52	0.60	2.80	0.27	-0.77	11	
7821	D 821	0.00	0.00	0.00	0.00	0.00	0	
9001	H 1	0.68	0.11	0.50	0.05	-0.11	0	
9003	H 3	0.98	0.46	2.14	0.19	-0.42	0	
9011	H 11	2.75	1.34	6.24	0.53	-1.44	36	
9012	H 12	0.07	0.00	0.02	0.00	-0.00	0	
9013	H 13	1.70	0.80	3.73	0.32	-1.00	22	
9014	H 14	2.00	1.06	4.95	0.38	-1.00	21	
9015	H 15	0.57	0.26	1.21	0.12	-0.18	0	

$H_{2(1 \text{ mapped})}$ of 3.0, with H_1 of 3.834, and with H_0 of 6.834.

We tested these hypotheses in heterogeneity analyses using simulated data sets containing a certain proportion of linked families (α of 10%, 30%, 50%, 70%, and 90%). These analyses made use of data for a four allele marker system, since this closely resembles the combined informativeness of true marker maps in the collaborative data (when expected and observed lod scores were compared). Each series consisted of 100 replicas and therefore involved 100 runs of the HOMOG program.

By combining the simulated data for a linked and an unlinked marker into an imaginary chromosome, we also created a typical HOMOG2 problem, with two loci to be mapped within the tested area.

For $\alpha = 0.5$, a mean lod score favouring $H_{2(2 \text{ mapped})}$ over H_0 of 24.27 was obtained. These results indicate that the family material is highly suitable for any type of linkage or heterogeneity analysis (fig 2). When at least 50% of the families were assumed to be linked, it was even possible to detect linkage under the (false) assumption of locus homogeneity in 94/100 attempts. For most α values the imaginary chromosome approach was more powerful than conventional HOMOG analysis.

We also studied the precision of the estimates obtained for the recombination fraction and the proportion of linked families. It was counted how often the correct values for α (true $\alpha \pm 0.1$) and θ (between 2 and 8%) were found in replicates that yielded significant evidence for heterogeneity or linkage or both. The α and

Figure 1 Family distribution by 'effective number of informative meioses' (EFNIM). The distribution of all 63 TSC1 families is shown in front (hatched), the 65 non-chromosome 9 linked families are shown behind (solid).

Figure 2 Power to detect linkage and heterogeneity. Broken line ($H1$) = power to detect linkage under the assumption of homogeneity ($Z > 3.0$), $+-+$ ($H2_{(1 \text{ mapped})}$) = power to detect linkage and heterogeneity if only one locus is known (see text), $*-*$ ($H2_{(2 \text{ mapped})}$) = power to detect linkage and heterogeneity if two loci are known, applying the imaginary chromosome approach (see text).

θ values obtained from the one locus and two locus heterogeneity tests were quite accurate (figs 3 and 4). However, for very low true α s the precision of the tested methods was found to be insufficient.

Discussion

SUITABILITY OF FAMILY MATERIAL

Our family material was found to be very suitable for heterogeneity analysis. The family structures provide sufficient information both

for detection of linkage and exclusion of linkage. Furthermore family structure is very similar in the chromosome 9 linked and unlinked groups.

The suitability of our family set has been confirmed by heterogeneity analyses of the collaborative data set.¹⁷ For major genes, detection of linkage and heterogeneity will not be problematical. The precision of the resulting values for α and θ are acceptable. However, mapping genes responsible for TSC in a small minority of families will be difficult, particularly when the α value approaches 10% or less. Under these circumstances the additional power offered by the imaginary chromosome approach is very useful, although still only 31/100 analyses reached significance.

This may be important with respect to the non-significant chromosome 12 findings of the collaborative linkage study.¹⁷ Our findings indicate that existence of a minor locus may only rarely yield significant evidence against homogeneity. As long as reasonable θ values emerge, it may be wise to continue the study of a putative minor locus. Linkage methods, however, can contribute little to such a study. Other avenues, such as the $t(3;12)$ translocation in the case of TSC3,¹² will have to be explored.

POWER OF THE METHODOLOGICAL APPROACH

Overall the imaginary chromosome approach has been shown to be a powerful method for mapping loci when locus heterogeneity occurs. If multiple candidate regions have been identified it seems sensible to analyse these simultaneously. Only when the α exceeds 0.8 does the conventional HOMOG approach ($H2_{(1 \text{ mapped})}$) perform better than the imaginary chromosome approach. Since a high α implies that a small number of families are assigned to the alternative locus, we may presume that the power of these families is insufficient to meet the required lod score difference of 3.0 between $H2_{(2 \text{ mapped})}$ and $H2_{(1 \text{ mapped})}$, as defined above.

Although we found the HOMOG analyses on our data to be quite satisfactory, we disagree with the thresholds normally accepted for significance levels. Older versions of the HOMOG programs used to indicate results as significant ($p < 0.05$) when the difference in lod score for hypothesis $H2_{(2 \text{ mapped})}$ versus $H1$ is 1.000 or more; in more recent publications,⁸ odds of 50:1 or 100:1 in favour of $H2_{(2 \text{ mapped})}$ have been regarded as significant. We feel that a lod difference of 3.834 is more appropriate (evidence for a second localisation should be at least as convincing as evidence required for a first localisation).

Our revision of lod score difference required for localisation of new TSC genes is relevant to the previous apparent support for the putative chromosome 11 TSC locus.¹⁶ The results obtained using conventional criteria for significance are apparently misleading. In contrast the new results of the collaborative study do not provide any support for a TSC2 locus on chromosome 11.¹⁷

Figure 3 Precision of θ estimates (θ obtained = true $\theta \pm 3\%$) under heterogeneity. Only results supported by significant lod scores were evaluated. H1: broken line, $H2_{(1 \text{ mapped})}$: +—+, $H2_{(2 \text{ mapped})}$: *—*.

Figure 4 Precision of α estimates (α obtained = true $\alpha \pm 10\%$) under heterogeneity. Only results supported by significant lod scores were evaluated. H1: broken line, $H2_{(1 \text{ mapped})}$: +—+, $H2_{(2 \text{ mapped})}$: *—*.

DESCRIPTION OF FAMILIES USING THE

'EFFECTIVE NUMBER OF INFORMATIVE MEIOSES'

We have used the 'effective number of informative meioses' for describing family size. The EFNIM value gives a much better description of the family size than does the number of generations, the number of affected, or the number of relatives. In contrast to the simulated lod score itself, it provides us with an instant image of the family size. We propose the use of EFNIM values to describe families in cases where it is not feasible to show the pedigrees themselves.

The EFNIM distribution of the TSC1 (chromosome 9 linked) and non-TSC1 families showed that there is no obvious difference in family size. This implies that there is no marked difference in the biological fitness associated with the different genetic types of tuberous sclerosis.

We would like to thank the TSC collaborative group, particularly Moyra Smith, Jonathan Haines, Jean Amos, David Kwiatkowski, Priscilla Short, Hope Northrup, Susan Blanton, Sue Povey, Mari Wyn Burley, J Michael Connor, Mark Nellist, Phillip Brook-Carter, Paul Fleury, Arjenne Hesselings-Janssen, Senno Verhoef, Ray Kandt, Raimund Fahsold, Hans-Dieter Rott, Margaret Pericak-Vance, Corinne Merkens, Anneke Maat-Kievit, and Dick Lindhout, for their part in the enormous effort to establish this very useful set of families. This study was funded by the Netherlands Praevention Fund (grant 28-1723) and the National Tuberous Sclerosis Association of the USA (grant 91-03).

Appendix Informativeness of families for linkage.

In linkage studies, the analysis of simple phase known families is straightforward: one can count recombinants and non-recombinants and no complicated statistical analysis is required. When some family members are not available for analysis, or when phase is not known, the calculations can become extremely involved. The evidence for linkage is then summarised via the best estimate of the recombination frequency and the corresponding maximum lod score. To ease the interpretation of lod scores, Edwards²⁴ calculated the so called 'equivalent observations' (n), that is, the number of recombinants and non-recombinants which would give the same lod score.

$$n = \frac{Z_{\max}}{\log 2 + \theta \log \theta + (1 - \theta) \log (1 - \theta)} \quad \text{if } \theta > 0 \quad (1)$$

$$n = \frac{Z_{\max}}{\log 2} \quad \text{if } \theta = 0 \quad (2)$$

According to this equation, a data set yielding a lod score of 0.419 at $\theta = 0.2$, for instance, is equivalent to five meioses, four of which are non-recombinants. It should be noted that this measure is calculated for an entire data set; it is only relevant when, in the calculations, the best estimate of the recombination frequency obtained in that same data set was used. To illustrate this point, a data set containing five completely informative meioses (four non-recombinants) will yield exactly five equivalent observations when a recombination frequency of 0.2 is assumed, but it will yield 25 equivalent observations when analysed under $\theta = 0.4$, or only two when $\theta = 0.1$. This also implies that equivalent observations are only additive over data sets when obtained for the same value of θ .

While the number of equivalent observations is a very convenient measure to summarise a given data set, it cannot be used directly to compare families with respect to their informativeness for disease mapping for several reasons. Firstly, the calculation of the number of equivalent observations is based on the best estimate of θ in the entire data set. When the ratio non-recombinants:recombinants in an individual family deviates from the overall θ ,

the number of equivalent observations from that family will not accurately reflect its information content, as became apparent in the previous example. Secondly, the calculated number of equivalent observations depends on the actual marker segregation in the family: when several parents are by chance homozygous for the marker, the information content of that family may seem small. For a fair comparison between families, the effects of chance marker segregation have to be eliminated.

In our TSC families, we simulated marker segregation for markers with various degrees of informativeness, linked to a putative TSC locus with 5% recombination. For the calculation of EFNIM values, lod scores were calculated in these replicates for that same value of θ . Frequently, the maximum lod score in a particular replicate occurred at some other value of θ ; the mean lod score over all replicates, however, peaked at a value of 5%. That mean lod score was used to compare families via the calculated EFNIM (effective number of informative meioses). The EFNIM was calculated by

$$\text{EFNIM} = \frac{\bar{Z}\theta}{\log 2 + \theta \log \theta + (1 - \theta) \log (1 - \theta)} \quad (3)$$

with $\theta = 0.05$

This equation is closely related to Edwards's formula, but here the assumed recombination frequency is used in the calculations rather than the best estimate of the recombination frequency. Also, the calculations are not carried out for each simulated replicate separately, but rather using the mean lod score over all replicates. Because EFNIM values are based on a large number of simulations (in this case 100), chance fluctuations of marker informativeness reduce to the average informativeness of the marker as characterised by the PIC value. The EFNIM values calculated here were obtained for a marker with eight alleles (PIC 0.86). EFNIM values for a less informative marker (with a PIC value of, say, 0.375) can be approximated via

$$\text{EFNIM}_{(\text{PIC}_1)} \sim \text{EFNIM}_{(\text{PIC}_2)} \times \frac{\text{PIC}_1}{\text{PIC}_2} \quad (4)$$

How accurate this approximation is depends on the actual family structure: when many persons have not been tested, reconstruction of their marker genotypes is only possible for highly polymorphic markers. For such families, less polymorphic markers will yield smaller EFNIM values than calculated by equation 4.

As two examples, consider families 1 and 1013 in the table (A).

Family 1 yielded an EFNIM of 1.92 for a marker with PIC=0.86. Using (4), a marker

with PIC=0.375 (two alleles with equal frequencies) should yield an EFNIM of 0.84. When calculated from the mean lod score for the marker with two alleles the actual EFNIM is indeed 0.84.

For family 1013, however, the predicted EFNIM for the two allele marker is 3.14, while the observed EFNIM is 2.74. In this large family, analysis with a marker with eight alleles allows reconstruction of missing genotypes which is frequently not possible using a two allele system.

- McKusick VA. Current trends in mapping human genes. *FASEB J* 1991;5:12-20.
- St Clair D, Blackwood D, Muir W, et al. No linkage of chromosome 5q11-q13 markers to schizophrenia in Scottish families. *Nature* 1989;339:305-9.
- Kelsoe JR, Ginns EI, Egeland JA, et al. Re-evaluation of the linkage relationship between chromosome 11p loci and the gene for bipolar affective disorder in the Old Order Amish. *Nature* 1989;342:238-43.
- Narod SA. Power of the admixture test to detect genetic heterogeneity. *Genet Epidemiol* 1991;8:209-16.
- Martinez M, Goldin LR. Power of the linkage test for a heterogeneous disorder due to two independent inherited causes: a simulation study. *Genet Epidemiol* 1990;7:219-30.
- Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet* 1963;27:175-82.
- Ott J. *Analysis of human genetic linkage*. Revised edition. Baltimore: Johns Hopkins University Press, 1991.
- Ott J, Bhattacharya S, Chen JD, et al. Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. *Proc Natl Acad Sci USA* 1990;87:701-4.
- Connor JM, Yates JRW, Mann L, Aitken DA, Stephenson JBP. Tuberous sclerosis: analysis of linkage to red cell and plasma protein markers. *Cytogenet Cell Genet* 1987;44:63-4.
- Fryer AE, Chalmers A, Connor JM, et al. Evidence that the gene for tuberous sclerosis is on chromosome 9. *Lancet* 1987;i:659-61.
- Smith M, Smalley S, Cantor R, et al. Mapping of a gene determining tuberous sclerosis to human chromosome 11q14-q23. *Genomics* 1990;6:105-14.
- Fahsold R, Rott HD, Lorenz P. A third gene locus for tuberous sclerosis is closely linked to the phenylalanine hydroxylase gene locus. *Hum Genet* 1991;88:85-90.
- Haines J, Amos J, Attwood J, et al. Genetic heterogeneity in tuberous sclerosis: study of a large collaborative dataset. *Ann NY Acad Sci* 1991;615:256-64.
- Povey S, Attwood J, Janssen LAJ, et al. An attempt to map two genes for tuberous sclerosis using novel two-point methods. *Ann NY Acad Sci* 1991;615:298-305.
- Janssen LAJ, Povey S, Attwood J, et al. A comparative study on genetic heterogeneity in tuberous sclerosis: evidence for one gene on 9q34 and a second gene on 11q22-23. *Ann NY Acad Sci* 1991;615:306-15.
- Janssen LAJ, Sandkuijl LA, Merckens EC, et al. Genetic heterogeneity in tuberous sclerosis. *Genomics* 1990;8:237-42.
- Sampson JR, Janssen LAJ, Sandkuijl LA, et al. Linkage investigation of three putative tuberous sclerosis determining loci on chromosomes 9q, 11q, and 12q. *J Med Genet* 1992;29:861-6.
- Kandt RS, Pericak-Vance MA, Hung W-Y, et al. Linkage studies in tuberous sclerosis: chromosome 9?, 11? or maybe 14!. *Ann NY Acad Sci* 1991;615:284-97.
- Sampson JR, Yates JRW, Pirrit LA, et al. Evidence for genetic heterogeneity in tuberous sclerosis. *J Med Genet* 1989;26:511-6.
- Northrup H, Kwiatkowski DJ, Roach ES, et al. Evidence for genetic heterogeneity in tuberous sclerosis: one locus on chromosome 9 and at least one locus elsewhere. *Am J Hum Genet* 1992;51:709-20.
- Haines JL, Short MP, Kwiatkowski DJ, et al. Localization of one gene for tuberous sclerosis within 9q32-9q34, and further evidence for heterogeneity. *Am J Hum Genet* 1991;49:764-72.
- Burley MW, Attwood J, Kwiatkowski D, Povey S. The search for TSC1 on chromosome 9q34. *Ann Hum Genet* (in press).
- Ott J. Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci USA* 1989;86:4175-8.
- Edwards JH. The interpretation of lods in linkage analysis. *Cytogenet Cell Genet* 1976;16:289-93.