European Journal of Human Genetics (2002) 11, 179–184
© 2002 Nature Publishing Group All rights reserved 1018–4813/02 $25.00

www.nature.com/ejhg

ARTICLE

Clustering of Crohn's disease within affected sibships

Jean-Pierre Hugot*,[1,2], Jean-Pierre Cézard[2], Jean-Frédéric Colombel[3], J Belaiche and the GETAID[4], Sven Almer[5], Curt Tysk[6], Sean Montague[7], Miquel Gassull[8], Steen Christensen[9], Yigael Finkel[10], Corinne Gower-Rousseau[3], R Modigliani[11], Habib Zouali[1], Suzanne Lesage[1], Mathias Chamaillard[1], Jeanne Macry[12], Gilles Thomas[1] and Jean-Marc Victor[13]

[1]Fondation Jean Dausset CEPH, Paris, France; [2]European Working Group on the Genetics of IBD, Department of Paediatric Gastroenterology, Hôpital Robert Debré, Paris, France; [3]Registre EPIMAD, Hôpital Calmette, Lille, France; [4]Department of Gastroenterology, CHU de Liège, Belgium; [5]Division of Gastroenterology and Hepatology, IHM, Linköping Universitet, Linköping, Sweden; [6]Department of Gastroenterology, Örebro Medical Center Hospital, Örebro, Sweden; [7]Department of Gastroenterology, Adelaide & Meath Hospital, Dublin, Ireland; [8]Department of Gastroenterology, Hospital Universitari Germans Trias i Pujol, Badalona, Spain; [9]Department of Gastroenterology Herlev Hospital, Herlev, Denmark; [10]Department of Gastroenterology, Karolinska Children's Hospital, Stockholm, Sweden; [11]Department of Gastroenterology, Hôpital Saint Louis, Paris, France; [12]INSERM U458, Paris, France; [13]CNRS-UMR 7600, Université P. et M. Curie, Paris, France

Crohn's disease (CD) is a complex genetic disorder for which aetiology is unknown. Recently, genetic factors for susceptibility have been described. Several genetic loci have been mapped and partially explain the familial aggregations of the disease. However, environmental factors may also contribute to these aggregations. We considered that if the role of non-genetic factors was negligible, CD patients would be randomly distributed in sibships with multiple affected siblings. On the other hand if there was a significant environmental contribution, the siblings would be affected non-randomly over exposure status. In order to test this hypothesis, we studied 102 sibships with two or more affected siblings. A statistical test, named Cluster of Affected Sibling Test or CAST, was developed, based on the exact calculation of the probability of observing a given number of clusters of affected siblings in multiplex families. The null hypothesis of a random distribution of affected siblings was rejected (*P*=0,005). The observed excess of affected sibling clusters indicates that birth order influences the disease status. Considering that an adjacent order of birth is a global estimate of environmental sharing, this observation strongly suggests that environmental factors contribute to the observed familial aggregations of the disease. This observation provides evidence that familial CD is a relevant tool for further studies of environmental factors and gene-environment interaction. More generally, the CAST statistics may be widely applicable to estimate the involvement of environmental factors in the aetiology of other binary traits which may be observed in multiple members of the same sibship.
European Journal of Human Genetics (2002) **11,** 179–184. doi:10.1038/sj.ejhg.5200932

Keywords: affected siblings; clustering; Crohn disease; complex genetic disorders

Introduction

Crohn's disease (CD, MIM266600) is an inflammatory bowel disease (IBD) that is frequent in the Western world. The annual incidence of CD ranges from 1,7/10^5 to 24.3/10^5.[1] The greatest frequency occurs in the third decade.[2] To date the aetiology of CD is unknown but it is generally

*Correspondence: J-P Hugot, Fondation Jean Dausset/CEPH, 27, rue Juliette Dodu, 75010 Paris, France. Tel: (33) 1 53 72 50 24; Fax: (33) 1 53 72 50 58; E-mail: hugot@cephb.fr

Received 17 August 2001; revised 8 November 2002; accepted 12 November 2002

considered to result from a complex interplay between genetic and environmental factors.[3]

A genetic predisposition to CD was initially suggested by ethnic and familial aggregation of the disease (for review see reference[3]). Twin studies which demonstrated that the concordance rate for the disease was higher in monozygotic twins than in dizygotic twins also argued for a genetic factor.[4–6] A genetic predisposition to CD was firmly established when susceptibility loci were mapped using linkage studies. To date, at least seven susceptibility loci have been localised, on chromosomes 1, 5, 6, 12, 14, 16 and 19.[7–14] More recently, we identified the IBD1 susceptibility locus as the CARD15/NOD2 gene using a positional cloning approach.[15] This result was also obtained by a candidate gene strategy.[16] As expected, CARD15 does not explain the entire genetic predisposition considering that CARD15 variants are observed in no more than 50% of CD patients.

The incidence of CD has increased in numerous areas in Europe and North America in the second part of the 20th century.[17,18] Furthermore, the concordance rate among monozygotic twins ranges from 25 to 58.3% only. These observations argue for an environmental element in causation.[4–6] Among the large number of environmental factors that have changed in the Western way of life since the second world war, diet, infections and early events in childhood, measles vaccination, hygiene, contraceptive pills and tobacco have been proposed as important. However, to date, cigarette smoking is the only risk factor clearly established for CD.[19–21]

It is not known if environmental and genetic factors interact or if they have independent effects. We tried to deal with this question by studying sibships previously used to demonstrate linkage in genome-wide searches.[11] In a genetically inherited disease, affected siblings are expected to be randomly distributed within the sibship. On the other hand if there is an environmental contribution to the disease, it is expected that sibs who share many environmental factors would be more often concordant for the disease than sibs substantially different in age or parity. We considered the birth order as a global estimate of a shared environment, making the hypothesis that consecutive siblings in a sibship are more likely to be in close contact than those more widely spaced. We thus examined the distribution of affected siblings within sibships segregating for CD and we developed an original procedure to test whether the disease was randomly or non-randomly distributed in the sibships.

Patients and methods

Patients and families

Five hundred multiplex IBD families were identified through a large European consortium on the genetics of IBD. CD only families were included in the study and families with at least one ulcerative colitis or unclassified colitis were excluded. A part of this family set has been used in linkage studies and allowed us to map the IBD1 gene on chromosome 16[11] and to further identify CARD15.[15] Diagnostic criteria have been previously defined.[22]

For each family, a pedigree was drawn that included the date of birth, and the disease status (healthy or affected) for each family member. Families with twins were discarded because of completely shared birth order. Similarly, because we looked at a particular distribution of the disease status within sibships, families in which all siblings were affected were excluded. Thus, a total of 102 CD sibships with at least two affected siblings and one healthy sibling remained for this study. Sibships are described in Table 1 and Figure 1.

Distribution of the disease status within the sibships and the Clustering of Affected Siblings Test (CAST)

The 102 sibships were classed as 'consecutive' or 'not consecutive' according to the birth order of the affected siblings: a sibship was said 'consecutive' if all the affected siblings were born in consecutive order. When one or more healthy sibling separated affected siblings, the sibship was classed as 'non consecutive'. According to this classification, a random variable (rv) X was attributed to every sibship as follows: X=1 for a 'consecutive' sibship, X=0 for a 'non consecutive' sibship.

First consider a given sibship of n siblings with p affected siblings and n-p healthy siblings. This sibship is characterised by the paired values (n,p). For the i^{th} born sibling, let:

Table 1 Characteristics of the studied family set

Number of siblings per family (n)	*Number of affected siblings per family* (p)	*Number of (n,p) families* $N_{(n,p)}$	*Observed number of consecutive families* $M_{(n,p)}$
3	2	37	28
4	2	16	7
4	3	2	2
5	2	7	3
5	3	5	4
5	4	1	1
6	2	7	4
6	3	1	1
7	2	7	2
7	3	2	1
7	5	1	0
8	2	4	2
8	3	1	0
8	4	1	0
9	2	1	1
9	3	1	0
10	2	2	2
11	7	1	0
12	2	1	0
13	2	2	0
13	3	1	0
13	4	1	0
Total		102	58

The 102 Crohn's disease sibships were classed according to the total number of siblings (n) and the numbers of affected siblings (p). For each (n,p) class, the number $N_{(n,p)}$ of corresponding sibships is indicated. The number $M_{(n,p)}$ of consecutive sibships is also mentioned. A sibship is considered as consecutive if all the affected siblings are consecutive for the order of birth (see text).

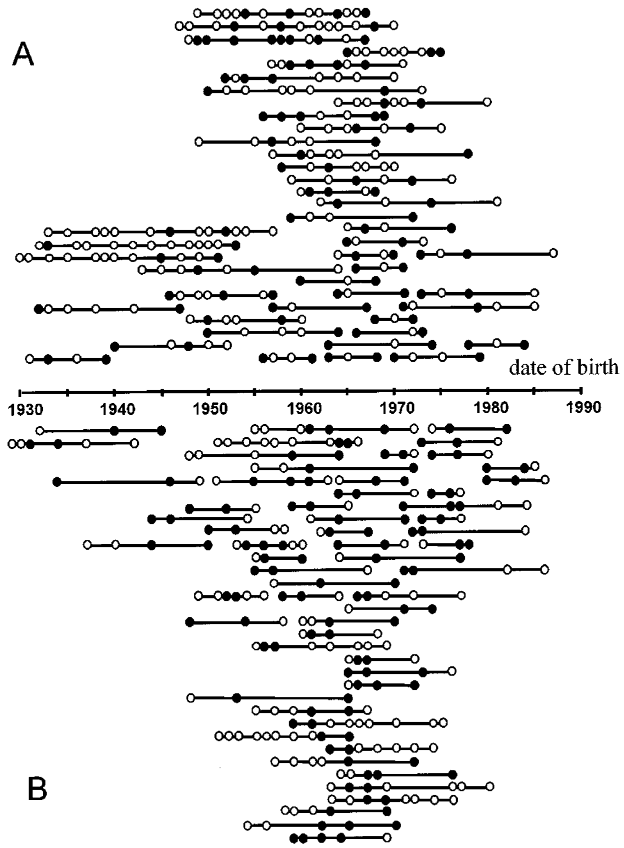

Figure 1 Schematic representation of the sibship sample. Each of the 102 studied sibships is represented by a solid line. The siblings are indicated on the sibship line by open circles (healthy siblings) or black circles (CD patients) drawn at their corresponding dates of birth. (**A**) Sibships where the affected siblings are not all consecutive for birth order (*n*=44). (**B**) Sibships where all the affected siblings are clustering (*n*=58).

$$\begin{aligned} Y_i &= 0 \text{ if healthy} \\ Y_i &= 1 \text{ if affected} \end{aligned} \tag{1}$$

If the disease status is identically distributed within the sibship, ie if the probability that the i$^{\text{th}}$ sibling is affected does not depend on the rank i (the n binary rv Y_i are Bernoulli rv), the probability for a sibship (n, p) to be classified as 'consecutive' is

$$P(X = 1) = \frac{n \quad p+1}{C_n^p} \tag{2}$$

and its probability to be classified as 'non consecutive' is

$$P(X = 0) = 1 \quad P(X = 1) \tag{3}$$

The probability generating function (pgf) of the binary rv X defined as usual as

$$\varphi(X) = P(X = 0) + P(X = 1)X \tag{4}$$

is a polynomial of the first degree in X equal to

$$\varphi(X) = 1 \quad \frac{n \quad p+1}{C_n^p} + \frac{n \quad p+1}{C_n^p} X \tag{5}$$

Then consider the set of N sibships. The N rv X_j (for j from 1 to N) are assumed to be independent. Their sum

$$S = \sum_{j=1}^{N} X_j \tag{6}$$

is again a rv. It is well known that the pgf of S is the product of the pgf of the N rv X_j. The pgf of S is therefore a polynomial of the N$^{\text{th}}$ degree in S equal to:

$$\Phi(S) = \prod_{(n,p)} [P_{(n,p)} S + 1 \quad P_{(n,p)}]^{N_{n,p}} \tag{7}$$

where $P_{(n,p)}$=$(n-p+1)/C^p{}_n$ and $N_{(n,p)}$ is the number of (n,p) sibships.

The probability of observing k 'consecutive' sibships in a given sample of families is equal to the coefficient of the k$^{\text{th}}$ degree of the polynomial Φ(S).

Among the subset of 'consecutive' sibships it is also possible to calculate the probability of finding T sibships where the older (respectively younger) sibling is affected. This probability P(T=k) is the coefficient of the k$^{\text{th}}$ degree of the polynomial Ψ(T) defined as

$$\Psi(T) = \prod_{(n,p)} Q_{(n,p)} T + 1 \quad Q_{(n,p)}]^{M_{(n,p)}} \tag{8}$$

where $Q_{(n,p)}$=$1(n-p+1)$ and $M_{(n,p)}$ is the number of (n,p) 'consecutive' sibships.

We propose a simple procedure to test whether the disease status is identically distributed within the sibships of a given set of families. An identical distribution of the disease status is equivalent to a random distribution of the affected siblings (patients). The null hypothesis:

H_0: 'the disease status is identically distributed within all the sibships' which is equivalent to the statement 'patients are randomly distributed within the sibships' is tested against the alternative hypothesis:

H_1: 'the affected siblings are not randomly distributed in some sibships'.

The statistic upon which the test is based is the number of 'consecutive' sibships. The *exact* probability $P(S \geq k) = \sum_{j=k}^{N} P(S = j)$ can be easily computed with a computer algebra system (ex:Maple$^{\text{R}}$). Because affected siblings are clustered together within a 'consecutive' sibship this new test is called 'Clustering of Affected Siblings Test' and will be quoted by its acronym CAST in the following.

In the same way, a test on the position of the last (or conversely first) affected sibling among the 'consecutive' families is based on the probability $P(T \geq k) = \sum_{j=k}^{N} P(T = j)$ which can be computed exactly as well.

Distribution of the proportion of affected siblings according to their date of birth

The global point of view previously assumed may be supplemented by a local analysis of the distribution of the clusters during the century in order to detect a birth cohort effect. Here again the number of clusters (ie consecutive strings of affected sibs) observed at a given period of time has to be compared with the number of clusters expected according to the number and the size of the families living at that period. A robust statistic is the proportion of affected siblings born at a given time and who belong to a cluster. More precisely we introduce for each date of birth (t) a fraction q(t) defined as follows

$$q(t) = \frac{\textit{number of patients born at t and belonging to a cluster}}{\textit{whole number of patients born at t}} \quad (9)$$

Note that q(t) is nothing but the probability that a CD patient born at a given date will belong to a cluster. The values of q(t) observed on a given sample of families, denoted $q_{obs}(t)$, are then compared to the values $q_{comp}(t)$ computed assuming a random distribution of affected siblings within the sibships. Finally the ratio r(t) defined as

$$r(t) = \frac{q_{obs}(t)}{q_{comp}(t)} \quad (10)$$

is a local measure of an excess of clusters at the time t.

Results

Among the 102 sibships with at least two affected siblings and one healthy sibling, the observed number of 'consecutive' sibships was found to be 58 (Table 1). The trend towards an excess of 'consecutive' sibships was observed for most of the (n, p) classes (Table 1).

According to the null hypothesis H_0 of a uniform distribution of the disease, the mean expected number of 'consecutive' sibships was 46. The computed probability of observing a value equal or higher than 58 was P=0.005. The observed number of 'consecutive' sibships was thus significantly higher than expected and demonstrated that the disease risk was not randomly distributed within the sibships.

In theory, the non-random distribution of the disease within sibships may be explained by the age of the parents at the date of birth of their offspring. The mean maternal (or paternal) age at birth for affected siblings was 27.4 (respectively 31.5) *vs* 27.8 (respectively 31.9) for healthy siblings (NS). Thus, neither the age of the mother nor the age of the father at the date of birth seemed to influence the disease status.

The non-random distribution may also reflect an over-incidence of the disease in special parity classes such as first- or last-born siblings. Out of the 58 'consecutive' sibships, the last born was affected in 21 sibships and the first born in 25. The expected number of families where the first-born sibling would be affected was found to be 22. By symmetry, this was also the expected number of sibships with the last-born affected. Using the above described test on first- (respectively last-) borns in the set of consecutive sibships we found no significant deviation. Thus, the non-random distribution of the disease in sibships was not related to the birth order of affected siblings. In other words, the observed clusters of affected siblings were uniformly distributed within the consecutive sibships (Figure 1).

We finally looked at the distribution of the clusters over time. Figure 2 reports the observed proportion of patients belonging to a cluster of affected siblings and the corresponding computed value assuming a random distribution of the disease within the sibships. In accordance with the previously obtained results using the CAST statistic, the mean observed value of q(t) was higher than the expected one. More interestingly, an excess of clusters of affected siblings was observed at any point between 1930 and 1980. This observation suggested that affected siblings were not clustering around a specific date within the century and did not argue for a cohort effect.

Discussion

In order to detect the contribution of the environment in familial CD, we analysed a large sample of sibships with at least two affected siblings and one healthy sibling. Using a new statistical test which we named 'Clustering of Affected Siblings Test' (acronym CAST), we were able to prove that the affected siblings are not randomly distributed within sibships as would be expected in case of a pure genetic disorder. Acquired genetic variations in the parental germinal cells are also unlikely considering that the disease is not associated with the parent's age at the time of birth. Thus, it can be concluded that the non-random distribution of the disease is very likely related to environmental factors playing a role in families.

In this set of multiplex families we could not find any evidence that the rank of birth was a risk factor. This observation suggests that the environmental exposure is not related to the intrinsic chronological history of the family, but rather to familial exposure to outside factors. The absence of a demonstrated cohort effect in these families suggests that the environmental exposure did not occur at the same time for all the families. On the contrary, starting dates of the exposure appear to be likely different from one family to another.

A gradually increasing incidence of the disease during the second part of the 20th century has been widely reported.[17,18] This finding demonstrates that CD phenotype is also modulated by the environmental exposure in sporadic cases. As for familial CD, no cohort effect has been detected. Thus, observations in sporadic and familial CD suggest that the same environmental risk factor(s) may be involved in both presentations of the disease.

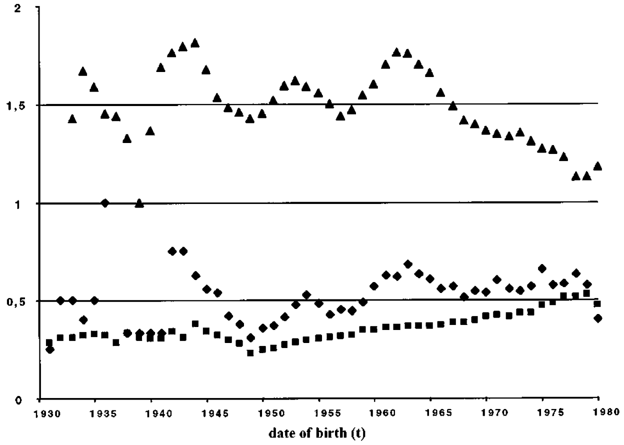

Figure 2 Local analysis of the distribution of the clusters during the century. $q_{obs}(t)$ (diamonds) is the observed proportion of affected siblings born at time t and belonging to a cluster. $q_{comp}(t)$ (squares) is the corresponding computed value under the hypothesis of a disease identically distributed among the siblings. r(t) (triangles) is the ratio of $q_{obs}(t)/q_{comp}(t)$. To smooth out the data, $q_{obs}(t)$ and $q_{comp}(t)$ were averaged over a 5 years window. More precisely

$$q(t) = \frac{\textit{number of patients born within } [t \quad 2,\ t+2] \textit{ and belonging to a cluster}}{\textit{whole number of patients born within } [t \quad 2,\ t+2]}$$

The excess of clusters of affected siblings in this family set which contributed to the localisation and identification of the IBD1 gene[11,15] suggests that both environmental and genetic factors are involved in the familial CD predisposition. Considering that CARD15 is to date the main known genetic factor involved in CD predisposition, we looked at the proportion of families with one or more CARD15 variation(s) in consecutive and non consecutive sibships. The proportion of families with CARD15 mutation(s) was respectively 0.55 and 0.57 (NS). This finding argues against any distinction of CD families with either a genetic determinant or an environmental exposure. Rather, this observation supports the hypothesis that genetic and environmental risk factors combine to control risk and that CD is a true multifactorial disorder. Multiplex families may thus be considered as an efficient tool for studying environmental predisposing factors and gene-environment interactions.

The new test proposed here and called CAST, is based on the calculation of the probability of observing a given number of clusters of affected siblings in multiplex families. To our knowledge, such a proposal has not been addressed in the literature. This approach is very different from a test which would try to correlate the disease status with the interval between the dates of birth of affected siblings. It thus appears as complementary to other correlation statistics which can be proposed to define the exact risk factor involved in the observed phenomenom.

Thus, if the CAST statistic strongly argues for an environmental factor in familial CD, additional studies taking into account specific environmental exposures (e.g. cigarette smoking), are needed in order to determine which environmental factor(s) is (are) involved. However, the CAST statistic is based on the hypothesis of a close contact between consecutive siblings rather than between more distant siblings. This hypothesis is probably true for environmental exposure in childhood but it appears less valid for environmental exposure later in life. Thus the result of the CAST statistic suggests that the environmental factor(s) play a role during childhood. This hypothesis is in accordance with the young age at onset of the disease for which the maximum of the incidence is observed in the third decade[2] and can be brought together with other works suggesting that hygiene in childhood may be a risk factor for CD.[23]

Some recruitment strategies may interfere with the test. For example, the larger sibships provide more data to test the null hypothesis. Thus, considering that older sibships are usually larger, it can be proposed to recruit older patients. In addition such older sibships would be less subject to misclassification of younger sibs for diseases with delayed age-of-onset. However, these older sibships are less likely to actually have good data on environmental factors and their ability to perform further correlation studies are limited.

As reported here, this test can be used in heterogeneous sets of families with various numbers of siblings, various proportions of affected siblings and whatever the familial predisposition to the disease. It is non-parametric and does not depend on the underlying model of the disease inheritance, including the number of susceptibility genes, the number of environmental factors and the interaction between them. If trivial biases (such as the effects of parental age or delayed age of onset) are discarded, it gives evidence for or against environmental risk exposure in families. It is easy to compute using a computer algebra system such as Maple^R (see Annex). Altogether, these properties of the CAST statistic suggest that it can be applied to a large variety of familial traits for which multiplex family samples are available.

Acknowledgements

We acknowledge for their contribution to family recruitment doctors Joaquin Balanzo, Vibeke Binder, Bruno Bonaz, Yoram Bouhnik, Guillaume Cadiot, Antoine Cortot, Salvatore Cucchiara, Bart Crusius, Jean-Jacques Delchier, Bernard Duclos, Jean-Louis Dupas, Yigael Finkel, Jean-Paul Galmiche, Jean-Pierre Gendre, Denis Golfain, Corinne Gower-Rousseau, Christer Grännö, Denis Heresbach, André Lachaux, Hervé Lautraite, Catherine Lenaerts, Eric Lerebours, Victor Levy, Robert Löfberg, Helmut Malchow, Philippe Marteau, Robert Modigliani, Alain Morali, Francesco Pallone, Salvatore Pena, André Rotenberg, Isabelle Rousseau, Jacques Schmitz, Fergus Shanahan, Iradj Sobhani, Hans Svensson, André Van Gossum, Myriam Van Winckel and Michel Veyrac. This project received financial support *from: the European Community (Contract Biomed 2 n° BMH4-97-2098), the Ministère Français de l'Education Nationale et de la Recherche, the INSERM, the Direction Générale de la santé (convention n° 4T004C), the Association François Aupetit and the Institut de Recherche des Maladies de l'Appareil Digestif and the Swedish Society of Medicine.*

References

1 Shivananda S, Lennard-Jones J, Logan R *et al*: Incidence of inflammatory bowel disease across Europe : is there a difference between North and South? Results of the European collaborative study on inflammatory bowel disease (EC-IBD). *Gut* 1996; **39**: 690–697.

2 Gower-Rousseau C, Salomez JL, Dupas JL *et al*: Incidence of inflammatory bowel disease in Northern France (1988–1990). *Gut* 1994; **35**: 1433–1438.

3 Hugot JP, Zouali H, Lesage S, Thomas G: Etiology of the inflammatory bowel diseases. *Int J Colorectal Dis* 1999; **14**: 2–9.

4 Tysk C, Lindberg E, Jarnerot G, Floderus-Myrhed B: Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* 1988; **29**: 990–996.

5 Thompson NP, Driscoll R, Pounder RE, Wakefield AJ: Genetics versus environment in inflammatory bowel disease: results of a British twin study. *Br Med J* 1996; **312**: 95–96.

6 Orholm M, Binder V, Sorensen TIA, Kyvik KO: Inflammatory bowel disease in a Danish register. *Gut* 1996; **39** (Suppl 3): A187.

7 Cho JH, Nicolae DL, Gold LH *et al*: Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, 4q: evidence for epistasis between 1p and IBD1. *Proc Natl Acad Sci USA* 1998; **95**: 7502–7507.

8 Cho JH, Nicolae DL, Ramos R *et al*: Linkage and linkage disequilibrium in chromosome band 1p36 in American Chaldeans with inflammatory bowel disease. *Hum Mol Genet* 2000; **9**: 1425–1432.

9 Duerr RH, Barmada MM, Zhang L, Pfutzer R, Weeks DE: High-density genome scan in Crohn disease shows confirmed linkage to chromosome 14q11-12. *Am J Hum Genet* 2000; **66**: 1857–1862.

10 Hampe J, Shaw SH, Saiz R *et al*: Linkage of inflammatory bowel disease to Human chromosome 6p. *Am J Hum Genet* 1999; **65**: 1647–1655.

11 Hugot JP, Laurent-Puig P, Gower-Rousseau C *et al*: Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996; **379**: 821–823.

12 Rioux JD, Daly MJ, Silverberg MS *et al*: Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 2001; **29**: 223–228.

13 Rioux JD, Silverberg MS, Daly MJ *et al*: Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *Am J Hum Genet* 2000; **66**: 1863–1870.

14 Satsangi J, Parkes M, Louis E *et al*: Two stage genome wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosome 3, 7 and 12. *Nature Genet* 1996; **14**: 199–202.

15 Hugot JP, Chamaillard M, Zouali H *et al*: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**: 599–603.

16 Ogura Y, Bonen DK, Inohara N *et al*: A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001; **411**: 537–539.

17 Rose JDR, Roberts GM, Williams G, Mayberry JF, Rhodes J: Cardiff Crohn's disease jubilee: the incidence over 50 years. *Gut* 1998; **29**: 346–351.

18 Kyle J: Crohn's disease in the Northeastern and Northern Isles of Scoland: an epidemiological review. *Gastroenterology* 1992; **103**: 392–399.

19 Persson PG, Ahlbom A, Hellers G: Inflammatory bowel disease and tobacco smoke—a case control study. *Gut* 1990; **31**: 1377–1381.

20 Sutherland LR, Ramcharan S, Bryant H, Fick G: Effect of cigarette smoking on recurrence of Crohn's disease. *Gastroenterology* 1990; **98**: 1123–1128.

21 Lindberg E, Järnerot G, Huiltfeld B: Smoking in Crohn's disease: effect on localisation and clinical course. *Gut* 1992; **33**: 779–782.

22 Lennard-Jones JE: Classification of inflammatory bowel disease. *Scand J Gastroenterol* 1989; **24** (Suppl 170): 2–6.

23 Gent AE, Hellier MD, Grace RH, Swarbrick ET, Coggon D: Inflammatory bowel disease and domestic hygiene in infancy. *Lancet* 1994; **343**: 766–767.

Appendix

Maple procedure for CAST:

```
>restart:
>G=unapply((x-y+1)/binomial(x,y)*X+1-(x-y+1)/binomial
(x,y),x,y):
  PSI:= G(3,2)^37*G(4,2)^16*G(5,2)^7*G(6,2)^7*G(7,2)^7
    *G(8,2)^4*G(9,2)*G(10,2)^2*G(12,2)*G(13,2)^2
    *G(4,3)^2*G(5,3)^5*G(6,3)*G(7,3)^2*G(8,3)*G(9,3)
    *G(13,3)
    *G(5,4)*G(8,4)*G(13,4)
    *G(7,5)
    *G(11,7):
>evalf(expand(PSI)):
>for i from 0 to degree(PSI) do D.i=evalf(coeff(PSI,X,i)) od:
  Dmax=0:
  for i from 0 to degree(PSI) do
    If D.i>Dmax then index=i:
    Dmax=D.i:
    Fi
  od:
  mean:=sum('i*D.i','i'=1..degree(PSI));# mean expected
number of 'consecutive' sibships according to H0
  indexmax:=index;# indexmax is the mode of the distribution
P(S=k)
  Proba:= sum('D.i','i'=58..degree(PSI));# value of P(S≥58)
```

Results:

```
mean=45.75
  indexmax=46
  Proba=0.005271
```