European Journal of Human Genetics (2007) 15, 864–871
© 2007 Nature Publishing Group All rights reserved 1018-4813/07 $30.00
www.nature.com/ejhg

ARTICLE

Estimating the odds ratios of Crohn disease for the main CARD15/NOD2 mutations using a conditional maximum likelihood method in pedigrees collected via affected family members

Leigh Pascoe*[,1], Habib Zouali[1], Mourad Sahbatou[1] and Jean-Pierre Hugot[2]

[1]*Fondation Jean Dausset CEPH, Paris, France;* [2]*INSERM Avenir, Université Paris 7, Assistance Publique-Hôpitaux de Paris, Hopital Robert Debré, Paris, France*

Estimation of genotype-specific risks at a disease susceptibility locus is an important question that is best carried out in a prospective study. Nevertheless it is usually desirable to make use of data from the families that have already been collected to identify the susceptibility locus. Assuming that the families have been collected without regard to genotype at the locus in question, most of the information can be extracted by writing the likelihood in terms of the risk for a genotype relative to the standard genotype and conditional on the parental mating type. Parameters may then be estimated by explicit solution of likelihood equations. This method permits estimation of risks for heterozygotes and homozygotes for different alleles, testing of different modes of inheritance and heterogeneity of risk between alleles. It is applicable to risk alleles for any disease locus or incompletely penetrant phenotype. We have used the method to estimate risks of Crohn disease for different CARD/NOD2 15 mutations, using the families originally collected to identify this susceptibility locus. The odds ratio of Crohn disease were, respectively, 1.97 ± 0.85, 3.05 ± * and 4.55 ± 1.34 for the R702W, G9068R and 1007fs heterozygotes and 3.29 ± 0.64, 12.13 ± * and 34.66 ± 12.87 for the corresponding homozygotes. (* Signifies insufficient data to estimate these values.) These results confirm the dosage effect for CARD15/NOD2 mutations and demonstrate that the disease risks are very different in homozygotes. This last observation illustrates the power of this approach, especially for alleles with low or moderate frequency in the general population.
European Journal of Human Genetics (2007) **15,** 864–871; doi:10.1038/sj.ejhg.5201839; published online 25 April 2007

Keywords: odds ratio; Crohn disease; complex disease risk

Introduction

Crohn disease (CD) is a complex genetic disorder with an estimated prevalence of 1/1000 in Western countries. The first susceptibility gene identified for CD was CARD15/NOD2, a gene involved in the innate immune response directed toward bacterial cell wall components.[1,2] More than 30 nonconservative variations have been reported within the gene.[3,4] Among them, three main mutations (R7012W, G908R and 1007fs) represent 82% of the CD-associated variations in the CARD15 gene.[3]

The frequencies of the three main mutations have been estimated to be 0.04, 0.01 and 0.02, respectively, for the R702W, G908R and 1007fs variants in Caucasian populations, with relatively large geographic fluctuations (Hugot *et al.*, unpublished data). In contrast, the same mutation frequencies were estimated to 0.11, 0.06 and 0.11 in a European panel of Caucasian CD patients.[3]

*Correspondence: Dr L Pascoe, Fondation Jean Dausset CEPH, 27 rue Juliette Dodu, 75010 Paris, France.
Tel: +33 1 53725125; Fax: +33 1 53725158;
E-mail: leigh@cephb.fr
Received 29 November 2006; revised 8 March 2007; accepted 22 March 2007; published online 25 April 2007

These mutations have occurred on a common ancestral haplotype which does not by itself confer a higher disease risk.[1,5] In a large panel of CD families we failed to detect chromosomes carrying two different mutations on the same chromosome homolog.[3] Data from the literature confirm this observation.[5] Consequently, we can consider the R702W, G908R and 1007fs variants as identifying distinct haplotypes in the vast majority of cases.

The risk of CD for people carrying one or more of the three main mutations has been estimated by case–control studies. In the first reports, the odds ratios (OR) of CD were estimated to be from 1.5 to 2.6 for heterozygous people and from 17.6 to 44 for homozygous or compound heterozygous individuals.[1,2,6] More recently, a meta-analysis has refined the initial estimates to be 2.20 (CI 95%: 1.84–2.62), 2.99 (CI 95%: 2.38–3.74) and 4.09 (CI 95%: 3.23–5.18) for, respectively, the R702W, G908R and 1007fs mutation carriers.[7] In the same meta-analysis, the OR for double dose mutation carriers (homozygous or compound heterozygous) was of 17.1 (CI 95%: 10.7–27.2).[7]

These data suggest that susceptibility is largely recessive in nature, but with a partial penetrance of mutations in heterozygote individuals. Because the three main mutations do not resume all the genetic diversity of the gene, it has been postulated that a second undetected mutation may be found in patients carrying only one of the three main variants. However, only a minority of heterozygous patients seem to carry an additional genetic variant[3,4] confirming the increased disease risk for the different CARD15/NOD2 heterozygous genotypes.

The OR estimated by case–control studies are known to be highly sensitive to recruitment biases. They also suffer from a limited statistical power in the case of rare disease-associated mutations, as is the case here – especially for homozygous and compound heterozygous individuals. For example, the limited number of double mutants in healthy controls makes defining the risk of CD difficult for each of the CARD15/NOD2 mutant homozygotes.

The study of human genetic disease often proceeds by the collection of families of individuals with members that are affected by the disease in question. These families may be identified by single probands, by multiple-affected members or from a specialized clinic. The families are typically extended by including first- or second-degree relatives, with additional branches of the family being sought if there are reports of affected individuals. This process leads to an overrepresentation of affected individuals in the families compared to the population prevalence, known in the literature as ascertainment bias. Ascertainment bias can make estimation of population risks and genetic parameters difficult or impossible, particularly if the recruitment criteria are poorly defined and special methods must be employed.

The overabundance of affected individuals can be accommodated in a genetic linkage analysis, which is often conditioned on the pedigree structure or uses only affected relatives for the analysis. Linkage analysis of genetic markers in the pedigree can indicate chromosomal regions involved in susceptibility to the disease. From there a gene conferring susceptibility to the disease may be identified by sequencing of candidate genes in the region, either with or without an intervening study of association of the disease to particular genetic markers, as was the case for the CARD15 gene and CD. The identification of mutations in a gene, followed by functional studies showing impaired activity in the encoded protein is usually sufficient to prove the involvement of the gene in the disease etiology.

Having identified a susceptibility gene for a complex disease, such as CD, we may then want to estimate the risks associated with a particular allele or genotype of the identified susceptibility locus. Typically we want to know the probability of becoming affected by a certain age, given the genotype at the locus. Subsequently, it may be desired to assign specific risks associated with different alleles of the identified gene or to test for any heterogeneity of risk between different alleles that do not completely abrogate function.

Collections of affected individuals and their genotypes contain information regarding the genotype-specific risks. However it is not straightforward to use this information given the biased nature of typical data-collecting methods. While some information can be obtained from the segregation of the alleles and disease in the families, the haphazard method of collection and extension of the families makes inference of the risk difficult.

Here we present a statistical framework for estimating the genotype-specific risks for disease in such families using the method of maximum likelihood conditioned on the mating type of the parents. Testing of specific hypotheses is also possible by a likelihood ratio test. The likelihood equations can be solved analytically, yielding relative risk estimates and standard errors for heterozygotes and homozygotes for specific alleles.

Materials and methods

Crohn disease family recruitment and genotyping method

CD families have been recruited through a large European consortium working on the genetics of Inflammatory Bowel Diseases (IBD) using classic diagnostic criteria based on clinical, endoscopic and histological findings.[8] In the present study, we analyzed the genotypes of 881 CD patients and 2901 of their unaffected relatives in 776 families; 235 of these families with only CD and healthy members were previously used in the CARD15/NOD2 cloning project.[1]

The three main mutations associated with CD were genotyped as described previously by an allele-specific PCR

procedure for the R702W variant, by a PCR-RFLP procedure for the G908R variant and by allele sizing on an automatic sequencer for the 1007fs variant. The experimental protocols have been reported previously in detail.[3]

Estimation of the disease risk using family data

Consider a locus with two alleles (1 and 2, where 2 is the disease-associated allele) and hence three genotypes, $g_1 = 11$, $g_2 = 12$ and $g_3 = 22$. We define the probability of affection of the ith genotype as $\Pr(A|g_i)$ and, using Bayes theorem,

$$\Pr(A|g_i) = \theta \frac{\Pr(g_i|A)}{P(g_i)}$$

Clearly these formulae cannot be used to estimate the disease risk without knowing the disease prevalence; θ, a parameter we assume is unknown and not able to be estimated from the family data, due to the ascertainment bias. However, if $\Pr(g_i)$ is the probability of g_i according to the Mendelian segregation probabilities, appropriate to the parental mating type, we can estimate the ratios

$$D = \frac{\Pr(A|g_2)}{\Pr(A|g_1)} = \frac{\Pr(g_2|A)}{\Pr(g_1|A)} \frac{\Pr(g_1)}{\Pr(g_2)}$$

$$D' = \frac{\Pr(\bar{A}|g_2)}{\Pr(\bar{A}|g_1)} = \frac{\Pr(g_2|\bar{A})}{\Pr(g_1|\bar{A})} \frac{\Pr(g_1)}{\Pr(g_2)}$$

$$R = \frac{\Pr(A|g_3)}{\Pr(A|g_1)} = \frac{\Pr(g_3|A)}{\Pr(g_1|A)} \frac{\Pr(g_1)}{\Pr(g_3)}$$

and

$$R' = \frac{\Pr(\bar{A}|g_3)}{\Pr(\bar{A}|g_1)} = \frac{\Pr(g_3|\bar{A})}{\Pr(g_1|\bar{A})} \frac{\Pr(g_1)}{\Pr(g_3)}$$

for the risks of affection and nonaffection for genotypes g_2 and g_3 relative to the risks for the reference genotype g_1.

Using these definitions we can write the expectations for the number of genotypes among affected offspring for each mating type as functions of D and R, as shown in Figure 1. Similar expressions can easily be written for the expected number of genotypes among unaffected offspring as functions of D' and R'.

Maximum likelihood method for OR estimates

The likelihood for affected offspring, conditional on the observed mating types, can be written directly in terms of D and R using separate calculations for each mating type and restricting the data to affected individuals only. For the given data set, the probability of observing the data among affected individuals would now be written as

$$P = 0.5^N C D^b (1+D)^{-(a+b)} (2D)^d R^e (1+2D+R)^{-(c+d+e)} D^f R^g (D+R)^{-(f+g)}$$

where C and K (below) are constants that do not affect the likelihood equation and N is the total number of affected individuals considered. Collecting like terms

$$P = K D^{(b+d+f)} R^{(e+g)} (1+D)^{-(a+b)} (D+R)^{-(f+g)} (1+2D+R)^{-(c+d+e)}$$

So

$$L = \log(P) = \text{Const} + (b+d+f)\,Log(D) + (e+g)\,Log(R) - (a+b)\,Log(1+D) - (f+g)\log(D+R) - (c+d+e)\,Log(1+2D+R)$$

Equating the partial derivatives with respect to each parameter to zero we have

$$\frac{\partial L}{\partial D} = \frac{(b+d+f)}{D} - \frac{(a+b)}{(1+D)} - \frac{2(c+d+e)}{(1+2D+R)} - \frac{(f+g)}{(D+R)} = 0 \quad (1)$$

$$\frac{\partial L}{\partial R} = \frac{(e+g)}{R} - \frac{(c+d+e)}{(1+2D+R)} - \frac{(f+g)}{(D+R)} = 0 \quad (2)$$

Giving us two simultaneous equations in the two unknown parameters that can be solved for the maximum

Parental mating type	g_1		g_2	g_2		g_2	g_2	g_2	g_3
	1|1	x	1|2	1 2|	x	1|2	1|2	x	2|2
Genotype	1|1		1|2	1|1	1|2		2|2	1|2	2|2
Observed	a		b	c	d		e	f	g
Expected	1		D	1	2D		R	D	R
Probability	$\frac{1}{(1+D)}$		$\frac{D}{(1+D)}$	$\frac{1}{(1+2D+R)}$	$\frac{2D}{(1+2D+R)}$		$\frac{R}{(1+2D+R)}$	$\frac{D}{(R+D)}$	$\frac{R}{(R+D)}$

Figure 1 Observed and expected ratios of genotypes among affected offspring from different informative parental mating types at a disease susceptibility locus. The symbols D and R represent the respective risks of developing disease for heterozygotes and homozygotes for a susceptibility mutation, relative to the risk for a normal homozygote.

likelihood estimates. These simultaneous equations lead to the respective maximum likelihood estimates

$$\hat{D} = \frac{b}{a}$$

$$\hat{D} = \frac{d}{2c},\ \hat{R} = \frac{e}{c} \text{ and}$$

$$\hat{D} = \frac{f}{g}\hat{R}$$

when only mating type 1 (when $c, d, e, f, g = 0$), mating type 2 (when $a, b, f, g = 0$) or mating type 3 (when $a, b, c, d, e = 0$) is present, as expected. The latter result shows that mating type 3 only contains information about the relative values of D and R. The estimates obtained when two of the three mating types are present are also easy to obtain. For example if we have data from mating types 1 and 2, the equations reduce to

$$\frac{(b+d)}{D} - \frac{(a+b)}{(1+D)} - \frac{2(c+d+e)}{(1+2D+R)} = 0$$
$$\frac{e}{R} - \frac{(c+d+e)}{(1+2D+R)} = 0$$

which yield estimates

$$\hat{D} = \frac{-a+2b-2c+d+\sqrt{(a-2b+2c-d)^2+8(a+c)(b+d)}}{4(a+c)}$$

$$\hat{R} = \frac{(a+2b+d+\sqrt{(a-2b+2c-d)^2+8(a+c)(b+d)}}{2(a+c)(c+d)}$$

Similar to the information from the first and third mating types, the equations simplify to

$$\frac{-(a+b)}{(1+D)} + \frac{(b+f)}{D} - \frac{(f+g)}{(D+R)} = 0$$
$$\frac{g}{R} - \frac{(f+g)}{(D+R)} = 0$$

yielding solutions

$$D = \frac{b}{a}$$

$$R = \frac{bg}{af}$$

For mating types 2 and 3, the equations reduce to

$$\frac{(d+f)}{D} - \frac{(f+g)}{(D+R)} - \frac{2(c+d+e)}{(1+2D+R)} = 0$$
$$\frac{(e+g)}{R} - \frac{(f+g)}{(D+R)} - \frac{(c+d+e)}{(1+2D+R)} = 0$$

and these have solutions

$$D = \frac{c(3d+2(e+f)+g-\sqrt{c^2(4(d+e)(e+g)+d+2f+g)}}{4c^2}$$

$$R = \frac{-c(d+2f+g)+\sqrt{c^2(4(d+e)(e+g)+(d+2f+g)^2)}}{2c^2}$$

Finally if we have information from all of the mating types, the solutions to the likelihood equations are quite complex to write down, but can easily be produced by modern symbolic manipulation methods. We have used the computer algebra system Mathematica 5.0 (Wolfram Research) to obtain explicit solutions to the general equations that enable us to evaluate the ML estimates for our data (Mathematical notebooks available on request). It is also possible to find these estimates using numerical methods.

The variances of the estimators can be obtained from the information matrix of partial second derivatives whose terms are as follows

$$\frac{\partial^2 L}{\partial D^2} = -\frac{(b+d+f)}{D^2} + \frac{(a+b)}{(1+D)^2} + \frac{4(c+d+e)}{(1+2D+R)^2} + \frac{(f+g)}{(1+2D+R)^2}$$

$$\frac{\partial^2 L}{\partial D \partial R} = \frac{2(c+d+e)}{(1+2D+R)^2} + \frac{(f+g)}{(D+R)^2} = \frac{\partial^2 L}{\partial R \partial D}$$

$$\frac{\partial^2 L}{\partial R^2} = -\frac{(e+g)}{R^2} + \frac{(c+d+e)}{(1+2D+R)^2} + \frac{(f+g)}{(D+R)^2}$$

The matrix of variances and covariances is then the inverse of the information matrix whose elements are the second partial derivatives given above. A similar procedure can be followed with the unaffected offspring to obtain D', R', permitting calculation of a relative OR.

$$OR_R = \frac{\Pr(A|g_3)}{\Pr(\bar{A}|g_3)} \frac{\Pr(\bar{A}|g_1)}{\Pr(A|g_1)}$$

and

$$OR_D = \frac{\Pr(A|g_2)}{\Pr(\bar{A}|g_2)} \frac{\Pr(\bar{A}|g_1)}{\Pr(A|g_1)}$$

The variance of the OR can be calculated from the variances of the two estimates forming the ratio by the formula

$$\sigma^2_{x/y} = \frac{y^2\sigma_x^2 + x^2\sigma_y^2}{y^4}$$

Results

Estimating the *D* and *R* parameters in the family data

As mentioned above, molecular genetic analysis failed to detect more than one mutation on the same haplotype in these families.[9] It was thus possible to consider the three mutations as independent events. The data were analyzed in a two-stage procedure. Firstly we analyzed the data for each mutation separately, assuming that it was the only mutation present in the data. This has for effect to slightly overestimate the heterozygous risks for that allele, which

are inflated by the presence of any compound heterozygotes. The separate estimates for the homozygotes can then be compared for their heterogeneity by inspection of the standard errors or using a likelihood ratio statistic. If the risks from different alleles are comparable we can then pool the data from the different alleles reclassifying all alleles as either mutant or normal. This latter procedure yields a combined estimate or average effect of carrying mutant alleles in homozygous or heterozygous form. This procedure was chosen over the more complicated one of assigning separate risks for each heterozygote, homozygote and compound heterozygote, due to the paucity of the data.

Assuming a dichotomous classification of alleles as either mutant or wild type, we can simply count the number of mutations in an individual and assign a genotype 1/1, 1/2 and 2/2 according to the number of mutations.

Data of this type were generated from the full pedigrees using a PERL program that picked out informative matings from a file in linkage format and then counted genotypes in each class (program available on request). The numbers of each genotype (*a*, *b*, *c*, *d*, *e*, *f*, *g*), corresponding to the categories defined in Figure 1, are shown in Table 1 for each of the three main CARD15/NOD2 mutations. The values shown in the final row correspond to data where presence of any of the three alleles was counted as a mutation. ML estimates derived from the data in Table 1 and using the method described above are shown in Table 2.

We contrast the likelihood obtained for each of the individual estimates with the likelihood obtained assuming the mutation has no effect ($D=1$, $R=1$). It can be seen that each of the mutations has a highly significant effect on disease risk (Table 3).

The relative risks of different mutations

The estimates of relative risk shown in Table 2 enable us to assess the effects of the different alleles on the risk of developing disease. *A priori* we might expect that mutations leading to a complete inactivation of the encoded protein would give rise to similar relative risk estimates and could easily be combined. In that case we could simply recode the data to be mutant (any of the inactivating mutations) or normal and estimate the risk from the total data (as in Tables 1 and 3, last line). However, if some of the mutations are miss-sense mutations that potentially encode a partially active protein, it is preferable to test the

Table 1 Observed numbers of affected offspring of the genotypes defined in Figure

Allele	a	b	c	d	e	f	g	*Total*
R702W	45	79	4	11	8	0	2	150
G908R	18	45	0	1	2	0	0	68
1007fs	17	61	1	9	14	0	1	106
All alleles	42	90	5	29	49	6	17	238

Table 2 Maximum likelihood disease risk estimates for heterozygotes (*D*) and homozygotes (*R*) relative to the risk of the reference homozygous genotype with standard errors

Allele	$\hat{\mathrm{D}} \pm s.e.$	$\hat{\mathrm{R}} \pm s.e.$	$-2\mathrm{L}$	χ^2_2 *(No effect)*	$\chi^2_1(\hat{\mathrm{R}}\|D=1)$	$\chi^2_1(\hat{\mathrm{D}}\|\mathrm{D}=\mathrm{R})$
R702W	1.71±0.30	2.73±1.18	227.04	11.4**	1.30 NS	1.22 NS
G908R	2.53±0.70	12.13±15.13	80.95	14.60***	9.89**	1.76 NS
1007fs	3.64±0.96	12.06±5.68	133.95	42.11***	8.91**	8.70**
All mutations	2.24±0.39	7.62±1.91	372.94	72.06***	43.97***	36.99***

χ^2-statistics for testing the null hypothesis, heterogeneity between alleles and specific modes of inheritance are also shown
**$P<0.01$; ***$P<0.001$; NS, $P>0.05$.

Table 3 Observed numbers of unaffected offspring of genotypes, as defined in Figure

Allele	a	b	c	d	e	f	g	*Total*
R702W	95	85	9	13	6	0	1	209
G908R	50	41	0	1	0	0	0	92
1007fs	68	60	10	7	3	2	0	150
All alleles	140	108	23	32	19	16	8	346

There are insufficient observations to estimate *R* for the second mutation.

hypothesis that these mutations are of similar effect using a statistical test. We can construct such a test using the difference in log-likelihoods according to the formulae already developed. Suppose we have n data sets, d_i, $i=1, n$, representing data from different alleles. Then we may test the heterogeneity of the estimates from each allele using the statistic

$$X^2_N = \sum_{i=1,n} -2L(d_i|\hat{D}_i, \hat{R}_i) - 2L(\sum_{i=1,n} d_i|\hat{D}, \hat{R})$$

which will have a χ^2 distribution with $N=2(n-1)$ d.f.

If this difference is judged insignificant, we are justified in combining the data to obtain a single estimate of the relative risks. Alternatively we can simply use the standard errors of the individual estimates to judge if they are similar or not.

In our data it seems clear that the G908R and 10007fs have similar effects on disease risk with a 12-fold increase in risk for homozygotes and approximately 3-fold increase in risk for heterozygotes. The R702W mutation appears to incur a lower risk of disease, either in homozygous or heterozygous form. However, only the difference between alleles 1 and 3 is judged significant by the likelihood ratio statistic, which is consistent with the standard errors shown in Table 2 ($\chi^2_{2:1\,vs\,2}=2.33$ NS, $\chi^2_{2:2\,vs\,3}=0.92$ NS, $\chi^2_{2:1\,vs\,3}=8.10$, $P<0.02$). For individual patients it will clearly be preferable to calculate the risk for the specific allelic combination appropriate to their genotype, when more family data are available. The theory for risk estimation with multiple alleles may be developed using the approach presented in this study. We have not attempted that here for all possible allelic combinations due to the paucity of observations.

Estimating the D' and R' parameters in the family data

We also carried out a similar analysis for the unaffected individuals in the families (Table 3), permitting us to calculate the corresponding estimates and standard errors reported in Table 4.

Here $\hat{D}'$ and $\hat{R}'$ are the probabilities of nonaffection, given genotype 1/2 or 2/2, where 2 is the mutated allele, relative to the 1/1 genotype. We are unable to reject the hypothesis that $\hat{D}'=1, \hat{R}'=1$ for the wild-type alleles at the 5% significance level, although it is approaching significance. So there is no significant distortion of segregation among the unaffected offspring. Finally, the ORs for each genotype, with their standard errors, are given in Table 5.

Discussion

The estimation of the risk of developing a disease for given genotypes at a susceptibility locus is of obvious importance to individuals at risk. This risk estimation is often difficult when using families that have been collected to identify and characterize a susceptibility locus, due to complex ascertainment biases. While prospective studies to estimate these risks are feasible in principle, it is desirable to extract as much information as possible from the sometimes extensive pedigrees that have already been collected. It is equally desirable to estimate risks for homozygotes and heterozygotes that are specific to the allelic combinations occurring in individual patients, to test whether different alleles engender similar risks and to test specific hypotheses about the mechanism of inheritance.

The methods we outline in this paper, using a conditional likelihood estimation, can be used to achieve these aims. We have used the method to estimate risks for three common CARD15/NOD2 mutations conferring an increased risk of CD, using previously collected families with IBD susceptibility. In heterozygotes, these estimates are compatible with those derived from independent case–control studies. In fact the results obtained for heterozygote risks using data from 235 families with this approach are very similar to those observed in a large meta-analysis, which included 8944 CD patients and more than 7000 healthy controls from 42 published studies. It thus appears that the method is more powerful

Table 5 Odds ratios for the risk of CD for heterozygotes and homozygotes relative to the standard genotype, with standard errors

Alleles	*Heterozygotes*	*Homozygotes*
R702W	1.97±0.85	3.29±0.64
G908R	3.05±*	12.13±*
1007fs	4.55±1.34	34.66±12.87
All	3.19±0.72	11.21±1.48

*Insufficient data to estimate these values.

Table 4 Maximum likelihood estimates of the probability of being unaffected for heterozygotes ($\hat{D}'$) and homozygotes ($\hat{R}'$) relative to the probability for the reference genotype, with standard errors

Alleles	$\hat{D}' \pm s.e.$	$\hat{R}' \pm s.e.$	$-2L$	$\hat{D}'=1, \hat{R}'=1$	χ^2_2
R702W	0.87±0.12	0.83±0.37	327.57	328.55	0.98
G908R	0.83±*	*	*	128.93	*
1007fs	0.80±0.13	0.35±0.22	231.40	235.67	4.27
All	0.77±0.09	0.68±0.17	576.48	582.24	5.76 $P=0.056$

*Insufficient data to estimate these values.

than case–control studies, especially for alleles with low or moderate frequency in the general population. This is true *a fortiori* for risk estimates of the different categories of homozygotes, which have previously never been calculated individually.

Using the present method we are able to report for the first time the risks of disease for different homozygous mutations. The disease risks appear different from one mutation to the other. For example, the R702W mutation confers a reduced risk of CD not only in heterozygotes but also in homozygotes. For this mutation, the reported dosage effect is very limited with only a slightly increased risk in homozygotes (3.29 ± 0.64) when compared to heterozygotes (1.97 ± 0.85).

The principal limitation of our approach is that no account has been taken of the age of individual family members. Some of the individuals carrying one or more disease alleles may eventually develop the disease, so the risks cited above are conservative. We ignored this complication after surveying the ages of the individuals concerned, most of whom had passed the typical age of onset for CD. Finally, we did not take into account the rare mutations which are observed in up to 22% of patient chromosomes.[3] Thus while the risk estimates for homozygotes are valid, those for heterozygotes will be slightly biased upward due to the possible presence of the other mutations in the pedigrees.

A better estimate of the disease risks in mutation carriers is useful not only for genetic counseling, but also for the understanding of disease mechanisms. Currently, there is no universally accepted model to explain how the CARD15/NOD2 mutations induce gut lesions. CARD15/NOD2 mutations are known to be unable to respond to the muropeptides.[2,10] More recently, it has also been shown that the mutated proteins fail to localize to the cytoplasmic membrane.[11] However, the relevance of these findings to CD is still unclear. Intuitively, it is expected that the biological defect should be correlated with the disease risk. In a previous study, the R702W was characterized by a strong defect of NF-κB activation after stimulation by the muramyl dipeptide, while the G908R mutation carried a less pronounced deficiency.[12] The risk estimates reported here for those mutations do not support the idea that the response to muropeptides drives the risk of disease.

The method used in this paper is not specific to IBD, applying to any disease with incomplete penetrance and a risk locus. It can be extended to multiple alleles and mating types if sufficient data are available. Here we did not separately analyze the risk in compound heterozygotes after considering the limited number of each genotype in our data set. However, such an extension of the method can be easily developed.

As with the transmission distortion test,[13] our method relies on the unequal representation of susceptibility and normal alleles among affected individuals. However the emphasis is on estimating a specific genotype risk, using data with strong ascertainment biases that is more appropriate for genetic counseling. In this sense it resembles the genotype relative risk first proposed by Schaid and Sommer[14] as an extension of the haplotype relative risk.[15] Although the emphasis in the former study was also on testing for association with disease rather than risk estimation at a known susceptibility locus, the resulting procedure is similar to that presented in this paper. In our approach the different alleles at a locus are compared for heterogeneity before making combined or separate risk estimates. The likelihood equations have also been solved explicitly to evaluate the maximum likelihood estimates and likelihood ratio statistics are recommended to test the various hypotheses.

In principle this procedure could be used as a more general test for the effect of an anonymous marker allele, or haplotype, on disease risk in a genome scan of anonymous genetic markers. However, the test would be subject to the usual provisos regarding multiple testing and false positives and would require both genetic linkage and association to the disease. We consider that it is most suitable for testing a candidate locus where the role of the gene in disease susceptibility is known or is suspected *a priori*.

Acknowledgements

This work was supported by the fondation Jean Dausset/CEPH, the Institut National de la Santé et de la Recherche Médicale, la Fondation pour la Recherche Médicale and la Mairie de Paris.

References

1 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.
2 Ogura Y, Bonen DK, Inohara N *et al*: A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001; **411**: 603–606.
3 Lesage S, Zouali H, Cezard JP *et al*: CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 2002; **70**: 845–857.
4 King K, Sheikh MF, Cuthbert AP *et al*: Mutation, selection, and evolution of the Crohn disease susceptibility gene CARD15. *Hum Mutat* 2006; **27**: 44–54.
5 Vermeire S, Wild G, Kocher K *et al*: CARD15 genetic variation in a Quebec population: prevalence, genotype–phenotype relationship, and haplotype structure. *Am J Hum Genet* 2002; **71**: 74–83.
6 Hampe J, Cuthbert A, Croucher PJ *et al*: Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* 2001; **357**: 1925–1928.
7 Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP: Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* 2004; **99**: 2393–2404.
8 Lennard-Jones JE: Classification of inflammatory bowel disease. *Scand J Gastroenterol Suppl* 1989; **170**: 2–6 (Discussion 16–19).
9 Hugot JP: Role of NOD2 gene in Crohn's disease. *Gastroenterol Clin Biol* 2002; **26**: 13–15.

10 Girardin SE, Hugot JP, Sansonetti PJ: Lessons from Nod2 studies: towards a link between Crohn's disease and bacterial sensing. *Trends Immunol* 2003; **24**: 652–658.
11 Barnich N, Aguirre JE, Reinecker HC, Xavier R, Podolsky DK: Membrane recruitment of NOD2 in intestinal epithelial cells is essential for nuclear fac. *J Cell Biol* 2005; **170**: 21–26.
12 Chamaillard M, Philpott D, Girardin SE *et al*: Gene-environment interaction modulated by allelic heterogeneity in inflammatory diseases. *Proc Natl Acad Sci USA* 2003; **100**: 3455–3460.
13 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.
14 Schaid DJ, Sommer SS: Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 1993; **53**: 1114–1126.
15 Falk CT, Rubinstein P: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 1987; **51** (Part 3): 227–233.