

## ARTICLE

# Familial cosegregation of rare genetic variants with disease in complex disorders

Ingo Helbig<sup>1</sup>, Susan E Hodge<sup>2,3,5</sup> and Ruth Ottman<sup>\*,3,4</sup>

Family-based designs are increasingly being used for identification of rare variants in complex disorders. This paper addresses two questions related to the utility of these designs. First, under what circumstances are rare disease-related variants expected to cosegregate with disease in families? Second, under what circumstances is a disease–variant association expected to be greater in studies restricted to familial cases than in studies of unselected cases? To investigate these questions, we developed a probability model of disease causation involving two loci. To address cosegregation, we examined the probability that an affected first-degree relative of a variant-carrying proband would also carry the variant. We find that this probability increases with increasing odds ratio (OR) for the variant, but declines with increasing sibling recurrence risk ratio ( $\lambda_s$ ). For example, under reasonable assumptions, the 15q13.3 microdeletion in idiopathic generalized epilepsy, with an OR estimate of 68 in large case–control studies, is expected to be present in >95% of affected first-degree relatives of variant-carrying probands. However, for a variant with OR = 5, the probability an affected relative has the variant ranges from 82% (when  $\lambda_s = 2$ ) to 58% (when  $\lambda_s = 50$ ). We also find that restriction of a study to familial cases does not necessarily increase a rare variant's association with disease, especially if  $\lambda_s$  is high and the variant contributes little to overall disease familial aggregation. These findings provide guidance for the design of family-based studies of rare variants in complex disorders.

*European Journal of Human Genetics* (2013) 21, 444–450; doi:10.1038/ejhg.2012.194; published online 26 September 2012

**Keywords:** rare variant; epidemiology; familial risk; microdeletion; epilepsy; genome sequencing

## INTRODUCTION

The genetic architecture of human disease includes a spectrum ranging from rare monogenic variants with very strong effects to common variants with small effects on the disease phenotype. Variants in the upper end of this spectrum have traditionally been investigated through linkage analysis in rare Mendelian families, whereas those at the opposite end have been investigated in genome-wide association studies (GWAS). The effect sizes of variants uncovered in GWAS have usually been very small, making it increasingly evident that a substantial proportion of the heritability of common diseases remains unexplained.<sup>1</sup> New molecular analysis methods such as massively parallel sequencing are likely to uncover a plethora of variants located between these extremes, that is, rare genetic variants with modest to high effect sizes. Although little is known about these variants, recurrent microdeletions recently discovered in neurodevelopmental disorders allow for a first insight into their properties, both in case–control and family studies.

Here, we address two questions related to the utility of family-based designs for the identification of rare variants in complex disorders. First, under what circumstances are rare disease-related variants expected to cosegregate with disease in families containing multiple affected individuals? Second, under what circumstances is the association of a variant with disease expected to be greater in affected individuals with an affected sibling than in unselected affected individuals?

We use as an example the relationship of the 15q13.3 microdeletion to idiopathic generalized epilepsy (IGE). The 15q13.3 microdeletion is implicated in several neurodevelopmental or neuropsychiatric disorders, including intellectual disability, autism, schizophrenia, and IGE.<sup>2–5</sup> In contrast to many other neurodevelopmental disorders, IGE is a distinct but relatively mild phenotype with a substantially increased familial risk (approximately eightfold in siblings of affected individuals<sup>6</sup>). The 15q13.3 microdeletion appears to confer a higher risk for IGE than for other neurodevelopmental disorders,<sup>7</sup> and a current estimate of the odds ratio (OR) for this variant in individuals with IGE compared with unaffected individuals is 68 (95% confidence interval 29–181).<sup>8</sup> Given this very high OR, evaluation of the effect of the 15q13.3 microdeletion in the families of deletion carriers is of interest.<sup>2,8,9</sup> However, some previous studies of the families of IGE probands with the 15q13.3 microdeletion have had results that appear counterintuitive: despite the high OR from case–control studies, this variant did not appear to segregate consistently with disease in families.<sup>4,8,10</sup> In the families of probands with the variant, some unaffected relatives have been found to carry the variant, a result easily explained by reduced penetrance. However, the variant has also been found to be absent in relatives who were affected, which is much more difficult to explain. This phenomenon has also been noted in the past, for example, in multiplex families with intellectual disability.<sup>11</sup> The findings have led some authors to conclude that in general, rare variants that contribute to risk for

<sup>1</sup>Department of Neuropediatrics, Christian-Albrechts-University Kiel, Kiel, Germany; <sup>2</sup>Department of Psychiatry, College of Physicians and Surgeons; and Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA; <sup>3</sup>Division of Epidemiology, New York State Psychiatric Institute, New York, NY, USA; <sup>4</sup>GH Sergievsky Center and Department of Neurology, College of Physicians and Surgeons; and Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

\*Correspondence: Dr R Ottman, GH Sergievsky Center and Department of Neurology, College of Physicians and Surgeons; and Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, 10032, USA. Tel: (212) 305 7892; Fax: (212) 305 2526; E-mail: ro6@columbia.edu

<sup>5</sup>Current address: Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital & The Ohio State University, Columbus, OH, USA  
Received 3 May 2012; revised 25 July 2012; accepted 26 July 2012; published online 26 September 2012

epilepsies with complex inheritance should not be expected to segregate with disease in families.<sup>8,12</sup>

We attempted to reconcile these findings by determining the conditions under which familial cosegregation of a rare variant with disease is expected. To address this problem, we developed a probability model based on simple assumptions about the factors other than the rare variant that influence disease risk in the families of affected probands who carry the variant. Our model assumes that disease risk is entirely attributed to the effects of two genetic loci, one of which is the rare variant under consideration. We operationalized the concept of familial cosegregation by considering the probability that the variant is present in an affected sibling of a variant-carrying proband. We estimated this probability under our model, based on the variant's frequency in the general population and in affected individuals (used to compute the OR), disease frequency, and sibling recurrence risk of the disease. We then used the same model to estimate the OR for the variant that would be expected in a study of affected individuals with affected siblings (ie,  $OR_{\text{fhs}}$ , defined below) and compared it with the usual OR (without subscript) in a study of unselected affected individuals.

## MATERIALS AND METHODS

### Parameter definitions

$P(D=1)$  denotes the frequency of the disease in the general population, which is estimated to be 0.005 for IGE, based on an estimated 3% lifetime risk of all epilepsy,<sup>13</sup> ~15–20% of which is IGE.<sup>14</sup>

$P(G=1)$  denotes the frequency of the variant (in heterozygous or homozygous state) in the general population. For the 15q13.3 microdeletion  $P(G=1)$  has been estimated to be 0.0002 in the Icelandic population,<sup>3,5,15</sup> but might be lower in other European populations.<sup>8</sup> For our estimates, we assume a frequency of 0.0002.

$P(G=1|D=1)$  is the frequency of the variant in cases, that is, the probability of the variant given that an individual is affected. 15q13.3 microdeletions have been identified in ~1% of patients with IGE, so that  $P(G=1|D=1) = 0.01$ . This frequency is probably significantly higher than in other neuropsychiatric disorders.

The OR for the variant, estimated from case–control data, is

$$OR = \frac{P(G=1|D=1)[1 - P(G=1|D=0)]}{[1 - P(G=1|D=1)]P(G=1|D=0)} \quad (1)$$

As noted above, for the 15q13.3 microdeletion in IGE, the OR has been estimated as 68.<sup>8</sup>

$P(D=1|G=1)$  refers to the penetrance of the variant, that is, the probability that an individual is affected given that he/she is a carrier of the variant.  $P(D=1|G=1)$  can be derived from the above parameters according to Bayes' theorem:

$$P(D=1|G=1) = \frac{P(G=1|D=1)P(D=1)}{P(G=1)} \quad (2)$$

In the example of the 15q13.3 microdeletion, penetrance =  $0.01 \times 0.005 / 0.0002 = 0.25$ .

### Probability that the variant is present in an affected sibling of a variant-carrying proband

For calculations pertaining to familial risk, we use subscripts *s* and *p* to refer to the disease and variant frequencies in a sibling or proband, respectively. Hence, the recurrence risk in the sibling of an affected proband is  $P(D_s=1|D_p=1)$ , and  $\lambda_s$ , the sibling recurrence risk ratio,<sup>16</sup> is  $P(D_s=1|D_p=1)/P(D_p=1)$ . For probands with IGE,  $\lambda_s$  has been estimated as 8.4 when only siblings with IGE are considered,<sup>6</sup> and 3–5 when siblings with any type of epilepsy are considered.<sup>6,17,18</sup> As our analyses are based on an assumed population risk of IGE specifically (rather than of all epilepsy), we assume  $\lambda_s = 8$ . This corresponds to a sibling recurrence risk of IGE of ~4%.

Using the notation above, the probability that a specific variant under consideration is present in an affected sibling of a variant-carrying proband is  $P(G_s=1|D_s=1, G_p=1, D_p=1)$ , and from the definition of conditional probability:

$$P(G_s=1|D_s=1, G_p=1, D_p=1) = \frac{P(D_s=1, G_s=1, G_p=1, D_p=1)}{P(D_s=1, G_p=1, D_p=1)} \quad (3)$$

To solve equation (3), we need to make some assumptions about the factors other than the genotype at the G locus that influence disease risk in the sibling of a proband who carries the variant. To model these factors, we assume disease risk involves two loci, G and H, where G is the locus under consideration so far, and H is another, unlinked and unknown gene. We assume each locus has two alleles in Hardy–Weinberg equilibrium, the two loci are not in linkage disequilibrium, and each locus is dominant with respect to disease risk. We allow for reduced penetrance of the susceptible genotypes (GG or Gg, and HH or Hh) but assume zero penetrance in the normal homozygote. We further assume that these two loci account for all of the disease risk in the population; that is, disease risk is 0 in individuals with the low-risk genotypes at both loci. The penetrance matrix for the two loci is shown in Table 1A.

We let  $p = \text{freq}(GG \text{ or } Gg \text{ genotype})$  and  $v = \text{freq}(G \text{ allele}) = (1 - \sqrt{1-p})$ . As the G locus in this example represents the 15q13.3 microdeletion, from our previous formulation this implies  $p = P(G=1) = 0.0002$  and  $v = 0.0001$ . Similarly, we let  $q = \text{freq}(HH \text{ or } Hh \text{ genotype})$  and  $w = \text{freq}(H \text{ allele}) = (1 - \sqrt{1-q})$ .

As we are interested in genotype probabilities in proband–sibling pairs, we can simplify our calculations by recognizing that within each family, the alleles of the parents are transmitted independently to each successive offspring. Hence, we consider all of the possible parental 'mating types' (ie, combinations of genotypes in mother and father) with regard to the G and H loci (Table 2). In our example, the G allele is very rare; hence we restrict attention to genetic parental mating types involving either 0 or 1 G allele (either  $gg \times gg$  or  $Gg \times gg$ ). However, we consider all possible mating types at the H locus. Taking these genetic parental mating types into consideration, equation (3) can be written as:

$$P(G_s=1|D_s=1, G_p=1, D_p=1) = \frac{\sum_{\text{mt}} P(\text{mt})P(G_s=1, D_s=1, G_p=1, D_p=1|\text{mt})}{\sum_{\text{mt}} P(\text{mt})P(D_s=1, G_p=1, D_p=1|\text{mt})} \quad (4)$$

**Table 1 Two-locus penetrance matrix for genotypes G and H**

	G = 1 (high-risk genotype at G locus (GG or Gg))	G = 0 (low-risk genotype at G locus (gg))
(a) General definitions		
H = 1 (high-risk genotype at H locus (HH or Hh))	$f_1$	$f_3$
H = 0 (low-risk genotype at H locus (hh))	$f_2$	0
(b) Values under Risch's heterogeneity model <sup>1</sup>		
H = 1 (high-risk genotype at H locus (HH or Hh))	$f_1 = 1 - (1 - x_1)(1 - y_1)$	$f_3 = 1 - (1 - x_0)(1 - y_1)$
H = 0 (low-risk genotype at H locus (hh))	$f_2 = 1 - (1 - x_1)(1 - y_0)$	$0 = 1 - (1 - x_0)(1 - y_0)$

<sup>1</sup> $x_1, x_0, y_1,$  and  $y_0$  are penetrance 'summands' from Risch.<sup>16</sup> Given our assumption of zero penetrance for individuals with  $G=0, H=0$ , these values imply:  $1 - f_1 = (1 - f_2)(1 - f_3)$ , leading to  $f_1 = f_2 + f_3 - f_2f_3$ .

**Table 2** Components needed to compute probability that the variant is present in an affected sibling of a variant-carrying proband<sup>a</sup>

Parental mating type at G and H loci	P(mating type)	P(D=1 mating type) <sup>b</sup>	P(D=1, G=1 mating type) <sup>c</sup>
Gg × gg, hh × hh	4v(1-v) <sup>3</sup> · (1-w) <sup>4</sup>	0.5f <sub>2</sub>	0.5f <sub>2</sub>
Gg × gg, hh × Hh	4v(1-v) <sup>3</sup> · 4w(1-w) <sup>3</sup>	0.25 (f <sub>1</sub> + f <sub>2</sub> + f <sub>3</sub> )	0.25(f <sub>1</sub> + f <sub>2</sub> )
Gg × gg, hh × HH	4v(1-v) <sup>3</sup> · 2w <sup>2</sup> (1-w) <sup>2</sup>	0.5(f <sub>1</sub> + f <sub>3</sub> )	0.5f <sub>1</sub>
Gg × gg, Hh × Hh	4v(1-v) <sup>3</sup> · 4w <sup>2</sup> (1-w) <sup>2</sup>	0.5(0.75f <sub>1</sub> + 0.25f <sub>2</sub> + 0.75f <sub>3</sub> )	0.5(0.75f <sub>1</sub> + 0.25 f <sub>2</sub> )
Gg × gg, Hh × HH	4v(1-v) <sup>3</sup> · 4w <sup>3</sup> (1-w)	0.5(f <sub>1</sub> + f <sub>3</sub> )	0.5f <sub>1</sub>
Gg × gg, HH × HH	4v(1-v) <sup>3</sup> · w <sup>4</sup>	0.5(f <sub>1</sub> + f <sub>3</sub> )	0.5f <sub>1</sub>
gg × gg, hh × hh	(1-v) <sup>4</sup> · (1-w) <sup>4</sup>	0	0
gg × gg, hh × Hh	(1-v) <sup>4</sup> · 4w(1-w) <sup>3</sup>	0.5f <sub>3</sub>	0
gg × gg, hh × HH	(1-v) <sup>4</sup> · 2w <sup>2</sup> (1-w) <sup>2</sup>	f <sub>3</sub>	0
gg × gg, Hh × Hh	(1-v) <sup>4</sup> · 4w <sup>2</sup> (1-w) <sup>2</sup>	0.75f <sub>3</sub>	0
gg × gg, Hh × HH	(1-v) <sup>4</sup> · 4w <sup>3</sup> (1-w)	f <sub>3</sub>	0
gg × gg, HH × HH	(1-v) <sup>4</sup> · w <sup>4</sup>	f <sub>3</sub>	0

<sup>a</sup>As the G allele is very rare, only mating types involving 0 or 1 G allele are considered.  
<sup>b</sup>P(D=1|mt) = f<sub>1</sub> · P(G=1|mt) P(H=1|mt) + f<sub>2</sub> · P(G=1|mt) · P(H=0|mt) + f<sub>3</sub> · P(G=0|mt) P(H=1|mt).  
<sup>c</sup>P(D=1, G=1|mt) = f<sub>1</sub> · P(G=1|mt) P(H=1|mt) + f<sub>2</sub> · P(G=1|mt) P(H=0|mt).

where mt refers to the genetic mating type at the G and H loci. Within each parental mating type, each offspring (whether proband or sibling) is independent with regard to whether or not he/she inherits a high-risk genotype or develops disease. Thus, the term in the numerator of equation (4), P(G<sub>s</sub>=1, D<sub>s</sub>=1, G<sub>p</sub>=1, D<sub>p</sub>=1|mt), can be written as P(G<sub>s</sub>=1, D<sub>s</sub>=1|mt) · P(G<sub>p</sub>=1, D<sub>p</sub>=1|mt) = P((G=1, D=1|mt))<sup>2</sup> (ie, the subscripts that refer to proband and sibling can be removed). Similarly, the term in the denominator of equation (4), P(D<sub>s</sub>=1, G<sub>p</sub>=1, D<sub>p</sub>=1|mt) = P(D<sub>s</sub>=1|mt) · P(G<sub>p</sub>=1, D<sub>p</sub>=1|mt) = P(D=1|mt) P(G=1, D=1|mt). Hence,

$$P(G_s = 1 | D_s = 1, G_p = 1, D_p = 1) = \frac{\sum_{mt} P(mt) [P(G=1, D=1 | mt)]^2}{\sum_{mt} P(mt) P(D=1 | mt) P(G=1, D=1 | mt)} \tag{5}$$

where P(G=1, D=1|mt) = P(G=1, H=1, D=1|mt) + P(G=1, H=0, D=1|mt)  
 = P(D=1|G=1, H=1) P(G=1, H=1|mt) + P(D=1|G=1, H=0)  
 P(G=1, H=0|mt)  
 = f<sub>1</sub> · P(G=1|mt) P(H=1|mt) + f<sub>2</sub> · P(G=1|mt) P(H=0|mt).  
 Similar reasoning yields P(D=1|mt) =

$$f_1 \cdot P(G=1 | mt) P(H=1 | mt) + f_2 \cdot P(G=1 | mt) P(H=0 | mt) + f_3 \cdot P(G=0 | mt) P(H=1 | mt)$$

Table 2 shows general formulae for all of the ‘components’ needed to compute the individual terms in the summations in (5).

**OR for the variant in affected individuals with an affected sibling**

Under our model, the probability that the proband has the variant, given that he/she has an affected sibling, is

$$P(G_p = 1 | D_p = 1, D_s = 1) = \frac{\sum_{mt} P(mt) P(G_p = 1, D_p = 1, D_s = 1 | mt)}{\sum_{mt} P(mt) P(D_p = 1 | mt) P(D_s = 1 | mt)} = \frac{\sum_{mt} P(mt) P(G=1, D=1 | mt) P(D=1 | mt)}{\sum_{mt} P(mt) [P(D=1 | mt)]^2} \tag{6}$$

where P(G=1, D=1|mt) and P(D=1|mt) are defined above. We define a new OR, OR<sub>flx</sub>, representing the odds of the variant in cases with an affected sibling vs unaffected controls.

$$OR_{flx} = \frac{P(G_p = 1 | D_p = 1, D_s = 1) [1 - P(G=1 | D=0)]}{[1 - P(G_p = 1 | D_p = 1, D_s = 1)] P(G=1 | D=0)} \tag{7}$$

We use the probability in equation (6) to derive a formula for this OR. Below, we use OR (without a subscript) to refer to the OR defined in equation (1) (ie, the OR in unselected cases vs controls), to distinguish it clearly from OR<sub>flx</sub>, defined in equation (7).

**Finding a numerical solution consistent with the data**

To estimate the probabilities in equations (5) and (7) under specific scenarios, we need to derive reasonable values for f<sub>1</sub>, f<sub>2</sub>, f<sub>3</sub>, and q. To do this, we note that specific algebraic relationships hold, under the assumptions of the model in Table 1. First, the overall disease frequency is equal to:

$$P(D=1) = pqf_1 + p(1-q)f_2 + (1-p)qf_3 = p[qf_1 + (1-q)f_2] + (1-p)qf_3 \tag{8}$$

The penetrance of the G genotype is equal to:

$$P(D=1 | G=1) = qf_1 + (1-q)f_2 \tag{9}$$

From equations (8) and (9), we obtain:

$$qf_3 = \frac{P(D=1) - p \cdot P(D=1 | G=1)}{1 - p}, \tag{10}$$

which can be calculated from our input parameters.

If we can obtain a value for f<sub>3</sub>, we can derive f<sub>1</sub> and f<sub>2</sub> using reasonable assumptions about the relations between G and H in terms of their effect on disease risk. For example, under Risch’s heterogeneity model,<sup>16</sup> f<sub>1</sub> = f<sub>2</sub> + f<sub>3</sub> - f<sub>2</sub>f<sub>3</sub> (Table 1, Part b), and substituting for f<sub>1</sub> in equation (9), we obtain:

$$f_2 = \frac{P(D=1 | G=1) - qf_3}{1 - qf_3} \tag{11}$$

For an alternative model, we assume a model of ‘epistasis’ (in the sense that the combined effects of the G and H loci are greater than additive) in which f<sub>1</sub> = 1.0 (ie, all of the individuals with both G and H are affected). Under this model, P(D=1|G=1) = qf<sub>1</sub> + (1q)f<sub>2</sub>, and hence,

$$f_2 = \frac{P(D=1 | G=1) - q}{1 - q} \tag{12}$$

Second, we note, again, that conditional on mating type, disease occurs independently in the proband and sibling. Hence, the sibling recurrence risk can be expressed as:

$$P(D_s = 1 | D_p = 1) = \frac{\sum_{mt} P(mt) [P(D=1 | mt)]^2}{P(D=1)} \tag{13}$$

To obtain a reasonable value for q, we set q equal to a range of values between qf<sub>3</sub> and 1, which in turn provides values for f<sub>1</sub>, f<sub>2</sub>, and f<sub>3</sub> under the model assumed in equations (11) or (12). Then, we use equation (13) to compute the corresponding sibling recurrence risk for

each value of  $q$ , and select the value consistent with the observed sibling recurrence risk (see below).

Although the reasoning above is presented in terms of the probability that an affected sibling of a variant-carrying proband will carry the variant, we also performed the same modeling with regard to this probability for an affected offspring or parent. The results were exactly the same as those for an affected sibling; hence we conclude that for the model used here, the results apply to all first-degree relatives and not only siblings.

## RESULTS

### Application to probability that an affected sibling has the 15q13.3 microdeletion

We first used the model described above to estimate the probability that an affected sibling of a proband with the 15q13.3 microdeletion would also carry this variant. To estimate  $f_1$ ,  $f_2$ ,  $f_3$ , and  $q$  for the 15q13.3 microdeletion and IGE, we consider the information we have:  $P = 0.0002$ ,  $P(D = 1) = 0.005$ , and  $P(D = 1|G = 1)$  (ie, the penetrance of the microdeletion) = 0.25. Also, the sibling recurrence risk = 0.04. In this example, equation (10) gives:

$$qf_3 = \frac{0.005(0.0002 \cdot 0.25)}{0.9998} = 0.00495. \quad (14)$$

Under the heterogeneity model in equation (11), this implies  $f_2 = (0.25 - 0.00495)/(1 - 0.00495) = 0.2463$ , and hence  $f_1 = 0.2463 + 0.00495/q - (0.2463 \cdot 0.00495/q)$ . This relationship constrains the value of  $q$  to be  $\geq 0.00495$  (ie, the value of  $qf_3$  derived above); lower values of  $q$  would imply that  $f_1 > 1.0$ .

For the observed recurrence risk of 0.04 for IGE, we obtain an estimate of  $q = 0.066$ , which leads to  $f_1 = 0.303$ ,  $f_2 = 0.246$  (as above), and  $f_3 = 0.075$ . Using these values in equation (5) leads to  $P(G_s = 1|D_s = 1, G_p = 1, D_p = 1) = 97.9\%$ .

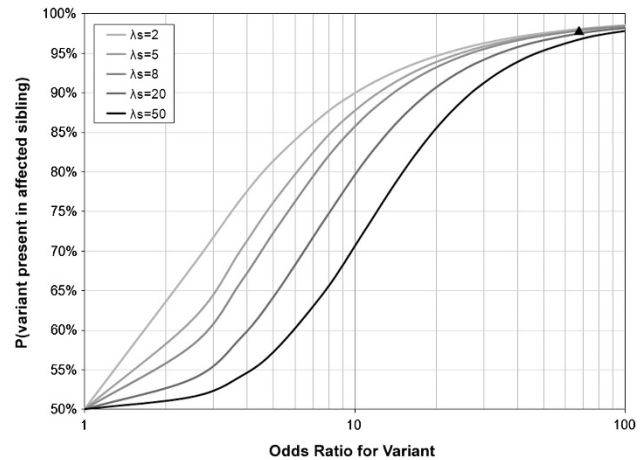
If we assume the epistatic model in equation (12), we obtain  $q = 0.066$  (as before),  $f_1 = 1.0$ ,  $f_2 = 0.197$ ,  $f_3 = 0.075$  (as before), and the probability that an affected sibling has the variant is slightly lower than before: 96.4%, but still very high.

### Extension to other inheritance models

We also considered whether other genetic inheritance models could lead to a different outcome. Our analysis indicated that given our input data for the 15q13.3 microdeletion in IGE – that is,  $P(D = 1) = 0.005$ ,  $P(G = 1) = 0.0002$ ,  $P(D = 1|G = 1) = 0.01$ , and sib recurrence risk = 0.04 – the system is quite constrained. Specifically, the value of  $q$  is constrained to be very close to 0.066, and this in turn implies  $f_3$  is very close to 0.075. Given these values, mating types involving both the rare 15q13.3 variant and other genetic causes (represented by locus H in our example) are extremely rare. In our example, almost all affected individuals who carry the variant come from mating type Gghh  $\times$  gghh, so that the only way their siblings can develop disease is through the effects of G. We conclude that given the assumptions of our model and the input data for the 15q13.3 microdeletion in IGE, the probability that an affected sibling of a variant-carrying proband also carries the variant cannot deviate substantially from the two values reported above, and certainly seems unlikely ever to be <95%.

### Relationship to OR and sibling recurrence risk ratio

Figure 1 shows the probability the variant is present in an affected sibling of a variant-carrying proband under the heterogeneity model,<sup>16</sup> as a function of the OR in equation (1) and sibling recurrence risk ratio,  $\lambda_s$ . Although the findings are presented for siblings, as noted above we have also determined that they also apply to other classes of first-degree relatives (affected offspring or parents



**Figure 1** The probability that a rare variant is present in an affected sibling of a variant-carrying proband, as a function of the OR and the sibling recurrence risk ratio,  $\lambda_s$ , where  $\lambda_s$  is the sibling recurrence risk divided by population disease frequency. Disease assumed to be caused by variants at two loci, G and H, with a penetrance matrix described by Risch's heterogeneity model<sup>16</sup> as shown in Table 1b. Population disease frequency assumed to be 0.005 and variant frequency at G locus assumed to be 0.0002. The black triangle indicates the expected result for parameters corresponding to the 15q13.3 microdeletion in IGE.

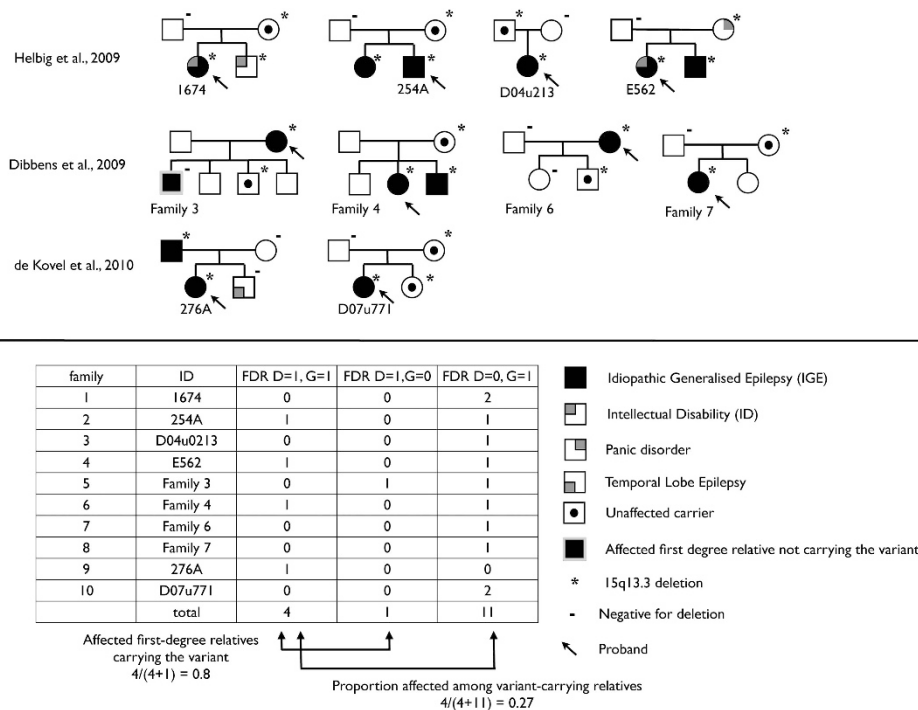
of variant-carrying probands). For a given level of familial aggregation ( $\lambda_s$ ), the probability that an affected sibling has the variant increases with increasing OR for the variant. Also, for a given OR, the probability that an affected sibling has the variant declines with increasing  $\lambda_s$ .

The results in Figure 1 are based on an assumed disease frequency of 0.005 and variant frequency of 0.0002 in the general population, consistent with IGE and the 15q13.3 microdeletion. We evaluated the impact of these assumptions by changing the assumed disease frequency and variant frequency. Changing the disease frequency had virtually no impact on the findings. Increasing the variant frequency in the general population (so that the OR declined) led to a decrease in the probability that the affected sibling had the variant, equivalent to the trend shown in Figure 1.

For many of the other microdeletions implicated in neurodevelopmental disorders, ORs of 5–10 are observed. Although these ORs are substantially higher than those observed in most GWAS, they imply penetrance estimates of only 6–8%.<sup>7</sup> Our results suggest that with ORs of this magnitude, under Risch's heterogeneity model<sup>16</sup> anywhere from 65 to 90% of the affected siblings of a proband who carries the variant would also be expected to carry it (for  $\lambda_s$  ranging from 2 to 20, Figure 1). Hence, in some situations (particularly with low ORs and high  $\lambda_s$ ), lack of clear cosegregation with disease in families is expected. An example of this is the pattern in four IGE families with inherited 15q11.2 microdeletions (OR of 4.9,  $\lambda_s \approx 8$ ), where only three (43%) of seven tested affected first-degree relatives were found to carry the variant.<sup>9</sup>

The figure also reveals that at very low and very high ORs, the sibling recurrence risk ratio has little effect on the probability of interest. For example, for ORs between 1.0 and about 1.3, the probability remains between 50 and 55%, over the whole range of  $\lambda_s$  values considered there, and for ORs over 50, the probability is between 95 and 100% for all those  $\lambda_s$  values. In contrast, for ORs around 5, the probability ranges from 58% (when  $\lambda_s = 50$ ) to 82% (when  $\lambda_s = 2$ ). This suggests that for rare variants with ORs of





**Figure 2** Pedigrees of published IGE multiplex families with inherited 15q13.3 microdeletion.

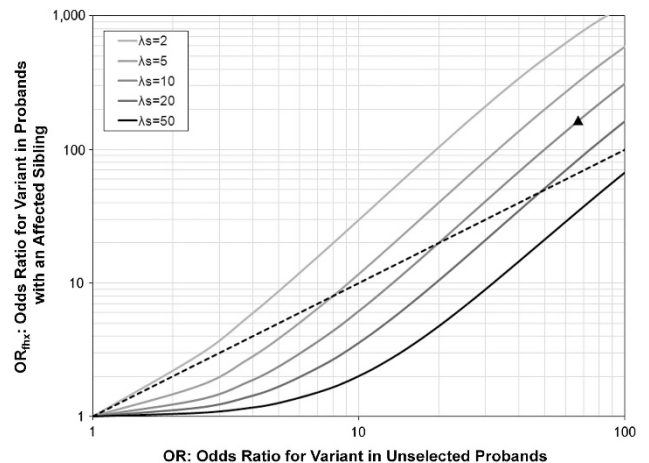
around 5, the likelihood of familial cosegregation with disease is critically dependent on the overall level of familial aggregation.

### Comparing the theoretical predictions to existing data

To compare our theoretical predictions with existing data on the segregation pattern of the 15q13.3 microdeletion, we tabulated data from informative families with IGE, that is, families in which first-degree relatives were genotyped, clinical information was available, and the 15q13.3 microdeletion was identified in the proband and confirmed to be inherited as opposed to *de novo*.<sup>2,8,9</sup> Information was available on 10 families that met these criteria (Figure 2). In these 10 families, 5 first-degree relatives were affected with IGE and four of them (80%) carried the variant. Conversely, the variant was present in 15 first-degree relatives, of whom 4 (27%) were affected with IGE. Three first-degree relatives with intellectual disability, panic disorder, or temporal lobe epilepsy who carried the variant were excluded from this analysis. The observed proportion of first-degree relatives with IGE who carried the variant was lower than expected (80% vs 97.9% expected, excluding *de novo* mutations). However, this is based on very small numbers (only five relatives); the exact *P*-value for the comparison is 0.101. Thus, from the available data on the 15q13.3 microdeletion, it would not be correct to conclude that familial cosegregation is not observed.

### Impact of having an affected sibling on the OR of a rare variant: $OR_{fHX}$ vs OR

Next, we investigated whether the  $OR_{fHX}$  is expected to be higher than the OR, that is, do we expect a stronger disease-variant association in a study restricted to probands with an affected sibling than in a study of unselected probands? The answer to this question would obviously influence the selection of familial vs sporadic cases for genetic studies. Contrary to our expectation, we found that the  $OR_{fHX}$  is not always higher than the OR (Figure 3).



**Figure 3** The expected OR for the variant in cases with an affected sibling vs unaffected controls ( $OR_{fHX}$ , y axis), and in unselected cases vs unaffected controls (OR, x axis). Disease assumed to be caused by variants at two loci, G and H, with a penetrance matrix described by Risch's heterogeneity model<sup>16</sup> as shown in Table 1b. Population disease frequency assumed to be 0.005 and variant frequency at G locus assumed to be 0.0002. The dashed black line indicates equal  $OR_{fHX}$  and OR. The black triangle indicates the expected result for parameters corresponding to the 15q13.3 microdeletion in IGE.

When the level of familial aggregation was relatively low (eg,  $\lambda_s = 2$ ), the  $OR_{fHX}$  was higher than the OR over the full range of ORs examined. For higher values of  $\lambda_s$ , however, under some conditions the  $OR_{fHX}$  was predicted to be lower than the OR. For example, our model predicts that if the OR = 5, the  $OR_{fHX}$  is expected to increase to 9 in a study of a disease with  $\lambda_s = 2$ , but to decrease to 2.8 in a study of a disease with  $\lambda_s = 8$ .

## DISCUSSION

We developed a simple probability-based model to explore the expected behavior of rare variants in families containing multiple affected individuals. The results have important implications with regard to the utility of family-based designs for detecting these variants.

First, the magnitude of a variant's effect on disease risk (measured in terms of the OR) strongly predicts whether or not it cosegregates with disease in families. This suggests (consistent with<sup>19</sup>) that tests of cosegregation in family data will not be very useful for identifying variants of small effect. However, for variants with ORs of 30 or higher, cosegregation with disease in families is expected: almost all affected first-degree relatives of probands who carry the variant are also expected to carry it. Although the actual numbers we report in Figure 1 depend on the specific assumptions used in our model, we believe the qualitative conclusions will hold even if some of those assumptions are relaxed.

The results in Figure 1 also provide quantitative estimates of the magnitude of the OR that would lead an investigator to expect cosegregation. If case-control data indicate an OR greater than  $\sim 30$ , but family data do not show cosegregation, the data are inconsistent. Either the OR is overestimated, for example, by 'winner's curse' or an inappropriate control group, or the test of cosegregation is methodologically flawed, for example, by selection bias, phenotypic misclassification, or small sample size. In our example of the 15q13.3 microdeletion in IGE, some authors have concluded, based on the scanty data available, that the variant does not cosegregate. We demonstrate that the number of families studied is too small to provide a valid test.

Second, for variants with a modest effect on disease risk (OR 2–20), the overall level of familial aggregation influences the likelihood that an affected first-degree relative of a variant-carrying proband will also have the variant. If  $\lambda_s$  is high, variants with ORs in this range may be absent in a substantial proportion of the affected first-degree relatives of variant-carrying probands. For example, that proportion could be as high as one-third when the OR is 5 and the sibling recurrence risk ratio is as high as 20. This implies that in studies using next-generation sequencing, filtering strategies that restrict attention to variants shared by affected first-degree relatives may miss variants with a modest effect on disease risk, especially in highly familial disorders.

Third, the overall level of familial aggregation of the disorder is also an important consideration in decisions about whether or not to use familial samples for the detection of rare variants (Figure 3). When  $\lambda_s = 2$ , the expected  $OR_{\text{fthx}}$  was higher than the usual OR, regardless of the OR in unselected cases. However, when the level of familial aggregation was higher ( $\lambda_s \geq 5$ ), the  $OR_{\text{fthx}}$  in a study using familial cases was predicted to be lower than the OR in a study using unselected cases, unless the OR in unselected cases was very high. The explanation for this pattern is that when  $\lambda_s$  is relatively high, the rare variant contributes little to the overall disease familial aggregation, so that selection of familial cases leads to an increased likelihood that cases have other genetic causes of disease (represented by H in our model), and a reduced likelihood that they have the rare variant under consideration. These findings are similar to our previous results, which showed that study designs using cases with an affected sibling increase power to detect a rare variant when  $\lambda_s = 2$ , but not when  $\lambda_s$  is higher.<sup>20</sup>

The results in Figure 3 do not necessarily argue against the utility of study designs using samples of familial cases, however. They pertain to the OR (or  $OR_{\text{fthx}}$ ) for a specific rare variant, whereas most studies

aimed at gene discovery involve testing for an association with any of the variants involved, rather than with a single candidate variant. Our results do imply, however, that selection of cases with an affected sibling does not have the same effect on all of the genetic variants that contribute to risk. Instead, the use of familial cases selectively increases the frequency of the variants with the greatest contribution to disease familial aggregation. In fact, if familial aggregation resulted primarily from shared environmental factors, we would predict that a sample of cases with an affected sibling would have reduced frequencies of all of the genetic variants involved.

In the context of a Mendelian disorder, affected family members who do not carry the variant segregating in the family are considered 'phenocopies.' Our findings show that such phenocopies are expected to be much more frequent for susceptibility variants that have ORs in the 5–10 range than for those with higher ORs. Hence, in complex disorders caused by variants with modest effects on disease risk, substantial genetic heterogeneity may be observed, even among closely related individuals within the same family. Application of new strategies for gene identification in complex disorders, such as massively parallel sequencing, need to take this heterogeneity into account.<sup>21</sup>

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (R01-NS043472, R01-NS036319, U01-NS053998, R03-NS065346, RC2-NS070344, U01-NS077276, U01-NS077367, and R01-NS078419 to R.O.), the National Institute of Mental Health (R01-MH048858 and P50-MH090966 (Conte Center, J Gingrich PI) to SEH), and the German Research Foundation (HE 5413/3-1 within the Eurocores program EuroEPINOMICS – Genetics of Rare Epilepsy Syndromes to IH).

- 1 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
- 2 Helbig I, Mefford HC, Sharp AJ *et al*: 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet* 2009; **41**: 160–162.
- 3 International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; **455**: 237–241.
- 4 Sharp AJ, Mefford HC, Li K *et al*: A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* 2008; **40**: 322–328.
- 5 Stefansson H, Rujescu D, Cichon S *et al*: Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; **455**: 232–236.
- 6 Hemminki K, Li X, Johansson SE, Sundquist K, Sundquist J: Familial risks for epilepsy among siblings based on hospitalizations in Sweden. *Neuroepidemiology* 2006; **27**: 67–73.
- 7 Vassos E, Collier DA, Holden S *et al*: Penetrance for copy number variants associated with schizophrenia. *Hum Mol Genet* 2010; **19**: 3477–3481.
- 8 Dibbens LM, Mullen S, Helbig I *et al*: Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance. *Hum Mol Genet* 2009; **18**: 3626–3631.
- 9 de Kovel CG, Trucks H, Helbig I *et al*: Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* 2010; **133**: 23–32.
- 10 van Bon BW, Mefford HC, Menten B *et al*: Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *J Med Genet* 2009; **46**: 511–523.
- 11 Geck J, Shoubridge C, Corbett M: The genetic landscape of intellectual disability arising from chromosome X. *Trends Genet* 2009; **25**: 308–316.
- 12 Dibbens LM, Heron SE, Mulley JC: A polygenic heterogeneity model for common epilepsies with complex genetics. *Genes Brain Behav* 2007; **6**: 593–597.
- 13 Hesdorffer DC, Logroscino G, Benn EK, Katri N, Cascino G, Hauser WA: Estimating risk for developing epilepsy: a population-based study in Rochester, Minnesota. *Neurology* 2011; **76**: 23–27.

- 14 Jallon P, Latour P: Epidemiology of idiopathic generalized epilepsies. *Epilepsia* 2005; **46**(Suppl 9): 10–14.
- 15 Kirov G, Grozeva D, Norton N *et al*: Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 2009; **18**: 1497–1503.
- 16 Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 1990; **46**: 222–228.
- 17 Ottman R, Lee JH, Hauser WA, Risch N: Are generalized and localization-related epilepsies genetically distinct? *Arch Neurol* 1998; **55**: 339–344.
- 18 Bianchi A, Viaggi S, Chioffi E: Family study of epilepsy in first degree relatives: data from the Italian Episcreeen Study. *Seizure* 2003; **12**: 203–210.
- 19 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
- 20 Ionita-Laza I, Ottman R: Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 2011; **189**: 1061–1068.
- 21 McClellan J, King MC: Genetic heterogeneity in human disease. *Cell* 2010; **141**: 210–217.