European Journal of Human Genetics (2015) 23, 1357–1363
© 2015 Macmillan Publishers Limited All rights reserved 1018-4813/15
www.nature.com/ejhg

ARTICLE

Where is the causal variant? On the advantage of the family design over the case–control design in genetic association studies

Claire Dandine-Roulland*,[1,2] and Hervé Perdry[1,2]

Many associated single-nucleotide polymorphisms (SNPs) have been identified by association studies for numerous diseases. However, the association between a SNP and a disease can result from a causal variant in linkage disequilibrium (LD) with the considered SNP. Assuming that the true causal variant is among the genotyped SNPs, other authors demonstrated that the power to discriminate between it and other SNPs in LD is low. Here, we propose to take advantage of the information provided by family data to improve the inference on the causal variant: we exploit the linkage information provided by affected sib pairs to discriminate the causal variant from the associated SNPs. The family-based approach improves discrimination power requiring up to five times less individuals than its case–control equivalent. However, the main advantage of family design is the possibility to carry out the procedure one step further: the linkage information allows inference on causal variants, which are not genotyped but in LD with tag-SNPs displaying association, which is impossible with case–control design. By means of Bayesian methods, we estimate the LD between the observed SNPs and an unobserved causal variant, as well as the allelic odds ratio at the unobserved causal variant. The proposed procedure is illustrated on a multiple sclerosis (MS) family data set including genotypes of SNPs in IL2RA, confirming the advantage of using a family design to identify causal variants. The results of our method on this data suggest the existence of two distinct causal variants in this gene for the MS.

European Journal of Human Genetics (2015) **23,** 1357–1363; doi:10.1038/ejhg.2014.284; published online 14 January 2015

INTRODUCTION

Association studies aim to identify variants associated with a disease, usually focusing on single-nucleotide polymorphisms (SNPs). They are able to detect the variants with modest effect, which are implied in complex diseases, contrarily to linkage analysis.[1] In genome-wide association studies (GWAS), the considered variants are tag-SNPs, which capture most common SNPs of the genome through linkage disequilibrium (LD).[2] However, the association between a SNP and a disease does not prove the causality link between the two: the association can result from a causal effect of the SNP itself or from the LD with another causal variant. Consequently, a significant association signal indicates a set of correlated variants associated with the disease. Discriminating between the causal variant and variants in LD with it by using case–control data was addressed by Udler *et al.*[3] Under the hypothesis that the causal variant is among the genotyped SNPs, the proposed method allows to select a minimal subset of potentially causal SNPs among disease-associated variants. Family data convey more information than case–control data, and their use can improve the performance of this selection process; moreover, family data allow to address a limitation of the discrimination method with case–control data, which is that the causal variant among genotyped SNP.

Here, we propose a method exploiting family data to select a minimal subset of associated SNPs and to make inference on putative causal variants in LD with those SNPs. This method uses an association framework, which takes advantage on the linkage information existing in affected sib-pairs (ASPs) data.[4] The first step is to select a minimal subset of potentially causal SNPs among disease-associated variants; assuming that the causal SNP is among the genotyped SNPs, we compare the performance of this discrimination step with the method using case–control data.[3] The second step of the method addresses the situation where the causal variant is not directly genotyped, but is in LD with genotyped SNPs. In this situation, case–control data do not convey enough information to make the difference between a SNP in strong LD with the unobserved causal variant, and a truly causal variant. Using a sample of ASPs, the number of alleles that shared identical-by-descent (IBD) by the two affected siblings allows to make inference about the causal variant, to estimate its allelic frequencies, the allelic odds ratio (OR) and the LD between it and an observed SNP.

The advantage of the family method is illustrated on a sample of multiple sclerosis (MS) data. MS is a chronic autoimmune neurological disease of the central nervous system, which affects about 1–2 per 1000 people in Europe and North America.[5] It is manifested by demyelination of nerve fibers in the brain, spinal cord and optic nerve. The disease is progressive and may lead to the loss of walking and eventually death. It is a multifactorial disease and has environmental and genetic factors. Several associations with genes involved in the immune response have been found. In particular, in the literature, there are associations with various genes in the human leukocyte antigen (HLA) region, and various non-HLA genes, for example, CD58, IL2RA and IL7R.[6] Our data set consists of french ASP and

[1]UMR-S 669, Université Paris-Sud 11, Villejuif, France; [2]U669, INSERM, Villejuif, France
*Correspondence: C Dandine-Roulland, UMR-S 669,Université Paris-Sud, Hôpital Paul Brousse Bâtiment Inserm 15/16, 16 avenue Paul-Vaillant-Couturier, Villejuif, F-94807, France. Tel: +331 45 59 53 64; Fax: +331 45 59 53 50; E-mail: claire.dandine-roulland@inserm.fr
Received 23 July 2014; revised 3 December 2014; accepted 4 December 2014; published online 14 January 2015

controls data from a previous study[7] and collected through REFGENSEP. It comprehends the genotypes of 26 SNPs in IL2RA on the chromosome 10. Several studies find association between MS and this gene in Caucasian populations.[8–13] The method using case–control data selects a minimal subset of seven associated SNPs, which reduces to three SNPs when using family data. The second step shows that none of these SNPs is causal, and that the association signal is due to at least two different ungenotyped variants in the region.

MATERIALS AND METHODS

In the first two paragraphs below, we give an overview of the method proposed by Udler *et al*[3] to identify causal variants in case–control association studies. The reader is referred to the original paper for details. In the third paragraph, we present the discrimination method that uses family data in two steps; selection of a subset of associated SNPs and inference about a putative causal variant not genotyped in the sample. In the last paragraph, we describe the MS data set used to illustrate this approach.

Discrimination procedures

Consider n highly correlated variants in a genomic region. Under the hypothesis that one of these variants is causal, the aim is to select a subset of these variants that is likely to contain the causal variant. The method relies on Bayesian principles: if L_i for $i=1,\dots,n$ is the likelihood that the ith variant is the causal variant, the variants selected are those of index i such that

$$\frac{\max_j L_j}{L_i} > K,$$

or, equivalently,

$$2\ln\left(\max_j L_j\right) - 2\ln L_i > k,$$

where $k=\ln(K^2)=2\ln K$. Following Udler *et al*,[3] we take $K=100$ (ie, $k=2\ln 100\approx 9.21$), which is interpreted as excluding variants with odds greater than 100:1.

For example, for two SNPs A and B, the SNP B is not retained if $2\ln L_{\mathrm{A}}-2\ln L_{\mathrm{B}}>k$. As asymptotically

$$2\ln L_A - 2\ln L_B \approx Y_A^2 - Y_B^2$$

where Y_A^2 and Y_B^2 are association test statistics corresponding to the likelihoods L_{A} and L_{B}, such as the Armitage Trend Test statistic[14–16] for case–control design, and a score statistic adapted to the family design,[4] this is equivalent to not retaining B is

$$Y_A^2 - Y_B^2 > k \tag{1}$$

Both these association statistics Y are approximately standard normal, $Y\sim N(0,1)$, under the hypothesis of no association with the disease. Otherwise, it is approximately decentered normal: assuming that the causal variant for the disease is A with ψ its per-allele OR,

$$Y_A \sim N(\eta,1)$$

where η is a decentered parameter, which will depend on the sample sizes, on the allele frequencies and on ψ. This parameter η also depends on the association statistic test used. Being in LD with the variant A, the variant B is also associated with the disease. Then,

$$Y_B \sim N(r\eta,1) \text{ and } \mathrm{cov}(\mathrm{Y_A},\mathrm{Y_B}) \approx r,$$

where r is the correlation coefficient between the two variants, measuring the intensity of the LD. The distribution of the discrimination statistic (equation 1) is approximately

$$Y_A^2 - Y_B^2 \sim N\left(\eta^2\left(1-r^2\right), 4\left(1-r^2\right)\left(1+\eta^2\right)\right)$$

(see details in section 1 of Supplementary Information).

Then, we can rely the power of discrimination $1-\beta$ with the decentered parameter η by

$$\eta^2\left(1-r^2\right) - z_{1-\beta}\sqrt{4(1+\eta^2)(1-r^2)} = k \tag{2}$$

where $z_{1-\beta}$ is the quantile of level $1-\beta$ of the standard normal distribution.

Discrimination with case–control data

Here we consider case–control data: the association statistic is the Armitage statistic.[14–16] Udler *et al* show that in this case

$$\eta = \frac{\sqrt{2f_a f_A}(\psi-1)}{\sqrt{\frac{\left(f_a\psi+f_A\right)}{m}+\frac{\psi}{n}}} \tag{3}$$

where, m and n are the number of controls and cases, f_{A} and f_{a} the frequencies of the reference and alternative alleles, A and a, and ψ the per-allele OR of a. The demonstration is given in section 2 of Supplementary Information. Then, if we assume that the number of controls and cases are equal, the total sample size needed to achieve power $1-\beta$ is

$$n+m = \frac{\eta^2\left(f_A + f_a\psi\right)^2 + \psi}{f_a f_A(\psi-1)^2} \tag{4}$$

where η^2 can be computed from β using equation 2.

Family design

Here, we propose a method in two steps. The first step is the selection of a subset of variants that is likely to contain the causal variant using the same discrimination procedure that in the case–control design, but using an association statistic designed for family data. We compute the power of discrimination of this procedure, assuming that the causal SNP is among the genotyped SNPs.

The second step uses the selected variants to make inference on causal variants in LD with them, relying on Bayesian principles. This step allows to retrieve information on a causal variant even if it is not genotyped.

First step: discrimination with family data

First, we use the same discrimination statistic (equation 1) based on a statistic Y, which has been proposed for ASPs and controls.[4] The data considered include genotypes of controls, genotypes of the index cases and the number of IBD alleles in each sib-pair. Hereafter, we denote the three possible genotypes by the number of alternative alleles: 0, 1 and 2.

We denote n_{ki} as the number of ASPs in which the index genotype is k and the number of IBD alleles is i, m_k as the number of controls with genotype k, and n and m as the total number of ASPs and controls. The association statistic is $Y=U/\sqrt{\sigma^2}$ where U is the score

$$U = \left(\sum_{k,i\in\{0,1,2\}}(2+i)n_{ki}\right)\hat{f} + \frac{1}{2}\sum_{k,i\in\{0,1,2\}}(2+i)kn_{ki}$$

with

$$\hat{f} = \frac{1}{2(m+n)}\left(\sum\nolimits_{i\in\{0,1,2\}} n_{1i} + 2\sum\nolimits_{i\in\{0,1,2\}} n_{2i} + m_1 + 2m_2\right)$$

the estimator of the alternative allele frequency, and

$$\sigma^2 = \frac{\frac{1}{4}\times\left(1-\hat{f}\right)\hat{f}(19m+n-1)n}{n+m}$$

the estimator of the variance of U under the hypothesis of no association. In absence of association, the distribution of Y is standard normal $Y\sim N(0,1)$.

We consider the causal variant A with allele frequencies f_{A} and f_{a} and OR ψ. The association statistic is decentered: $Y_A\sim N(\eta,1)$, where η is approximately

$$\eta = \frac{E(U_A)}{\sqrt{E(\hat{\sigma}_A^2)}}. \tag{5}$$

Formulas for $E(U_A)$ and $E(\hat{\sigma}_A^2)$ (depending on f_{A}, f_{a}, ψ and sample sizes n and m) are given in section 3 of Supplementary Information. Then, we can calculate power of discrimination for a given set of parameters, or total sample size needed to achieve a given power, with equation 2.

Second step: Bayesian inference on the causal variant

Second, we propose to use the SNPs selected by the discrimination step to retrieve information on the causal variants of the region. Let's assume that a variant B in LD with the causal variant A is observed. In this case, we want to make inferences on A, in particular to estimate the LD between A and B, and the OR of A. This task is undoable with case–control data, as an OR for variant B can always be computed, which explains fully the observations under the hypothesis that B is the causal variant. However, the advantage of family data lies in the linkage information provided by the IBD state of the sib-pairs, which allows to discriminate between observations made directly at a causal variant A, and observations made at a variant B in LD with A.

In section 4 of Supplementary Information, we write a likelihood for the family data $L\psi(f_a,f_b,d)$ (where ψ is the OR in A, f_a and f_b the alternative allele frequencies in A and B and d is the LD between A and B). In section 5 of Supplementary Information, we show that all parameters are identifiable, provided that $\psi > 1$ and $d \neq 0$.

This likelihood can be used to define the posterior distribution of parameters, from which we sample using Metropolis–Hastings algorithm[17] (cf section 6 of Supplementary Information for details). In particular, we can estimate the posterior distribution of the disequilibrium $r^2 = d^2/(f_a(1-f_a)f_b(1-f_b))$. We also find simultaneous credibility regions for f_a and f_b, or for ψ and r^2, using the posterior joint density of these parameters as estimated from the values sampled by the Metropolis–Hastings algorithm.

MS data

These two methods of discrimination are illustrated on MS data described in full details in Babron *et al.*[7] This data include 26 tag-SNPs on the IL2RA gene for french families with at least one affected child collected through REFGENSEP. All affected people were reviewed by a board-certified neurologist and diagnosed according to Poser criteria.[18] All individuals signed informed consent in accordance with the European Union and Country Laws and the Helsinki Convention. The sample comprises 522 trio families (one affected with two living parents) and 101 multiplex families (at least two affected sibs).

The trio families are used to create pseudo-control genotypes consisting of the alleles untransmitted by the parents to their affected child. Pseudo-control genotypes are known to represent general population genotypes.[19] ASPs are obtained from multiplex families, randomly selecting two affected sibs in each family. The IBD states are calculated using the software Merlin,[20] which calculates the probability of each IBD state. Only ASPs for which one IBD state has probability higher than 0.8 are kept, assigning the IBD state with probability exceeding 0.8 to them.

Overall, the data set comprises 522 pseudo-controls and 82 ASPs with case index genotypes and IBD states. In addition to applying the family-based descrimination method on the data set, we will also use the case–control method on the 82 index sibs as cases and the 522 pseudo-controls.

RESULTS

Power of the family and case–control discrimination methods

The power of the two discrimination methods depends on the expression of the decentered parameter η (equation 2). In Figure 1, we display η values for an OR ψ varying from 1–5. For all OR, the η parameter in a family design is higher than that of a case–control design.

The total sample size required to achieve 90 power to exclude variants at 100:1 odds assuming an equal number of controls and unrelated cases or ASPs for different values of alternative allele frequency, OR ψ and LD, r^2 is reported in Figure 2. For identical parameters, the family discrimination method needs a smaller sample size than the case–control method. For example, when $\psi = 3$, $r^2 = 0.9$ and $f_{\mathrm{a}} = 0.1$, the case–control method needs the genotypes of 1500 controls and 1500 cases, whereas the family discrimination method needs only the genotypes of 300 controls and 300 sib-pairs (genotype of the index case and IBD state, which can be obtained with a low-density genotyping of the second sib).

Inference on a causal variant with family data

Using Metropolis–Hastings algorithm, we simulate data composed of 1000 ASPs and 1000 controls. The theoretical distribution used for these simulations is described in section 4 of Supplementary Information. The posterior distribution of f_a,f_b,ψ, and $\underline{r}^2$, obtained from 10^7 distribution samples, are displayed in Figures 3 and 4. Each shade of gray represents the credibility region for one level. The lightest gray corresponds to all sampled values.

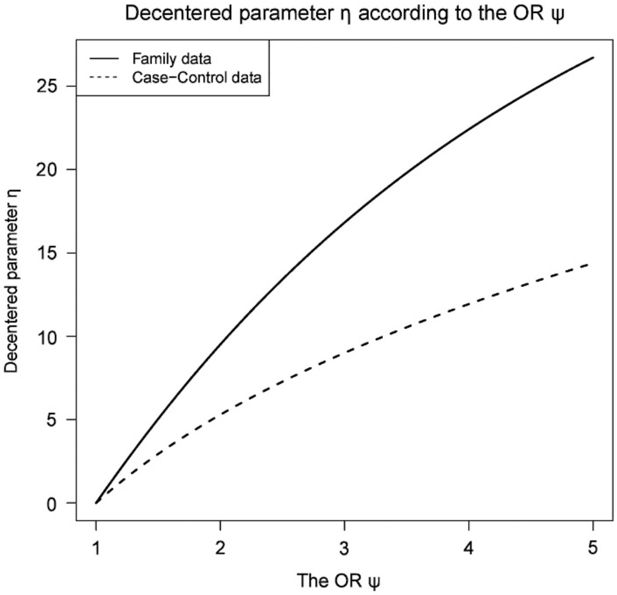

Figure 1 The parameter η according to the OR ψ for 0.1 alternative allele frequency, $r^2 = 0.8$ and 500 cases and 500 controls.

In Figure 3, the data are simulated under a model with total LD ($r^2 = 1$), alternative allele frequencies $f_a = f_b = 0.2$, and an OR $\psi = 2$ for the causal variant A. The 95 credibility regions of f_a, f_b, ψ, and r^2 are ~(0.17,0.215), (0.08,0.29), (1.8,3) and (0.31), respectively. They contain the true values of parameters, and the mode of the distribution is near to the true values. Note that the allele frequency is best estimated at the variant that is directly observed, which corresponds to a certain amount of uncertainty on r^2.

In Figure 4, the data are simulated with $r^2 = 0.8$, $f_{\mathrm{a}} = 0.435$, $f_{\mathrm{a}} = 0.448$ and $\psi = 3$. The 95 credibility regions of f_{b}, f_{a}, ψ and r^2 are (0.415,0.47), (0.3,0.55), (2.5,5) and (0.5,1), respectively. Again, they contain the true values of parameters. Interestingly, although the causal variant is not directly observed, some inference of its characteristics is possible.

Application to MS data

Discrimination methods. The values of association statistics and *P*-values for the two discrimination methods at all SNPs are displayed in the Table 1. For the SNPs with the smallest *P*-values (SNPs 1–4, 21 and 24), the family-based *P*-values are lower than the case–control ones. However, the use of family data does not decrease *P*-value for all SNPs: for example, the SNP reported in the literature[21–25] as associated with MS, rs2104286 (SNP 23 in our numbering), is not associated using case–control data, and adding the IBD information does not decrease its *P*-value. After Bonferroni correction, the association is significant association only with rs3118470 (SNP 24) for both case–control and family designs, and, with rs12359875 (SNP 1) for family design only.

To apply discrimination methods on these data, we compute the difference of association statistics between the most associated SNP, that is, SNP 24, and others (Table 1). Comparing these values with the threshold $k = 9.210$, we select the set of SNPs 1, 2, 3, 4, 20, 21 and 24 as likely to contain the causal variant, using the case–control

]

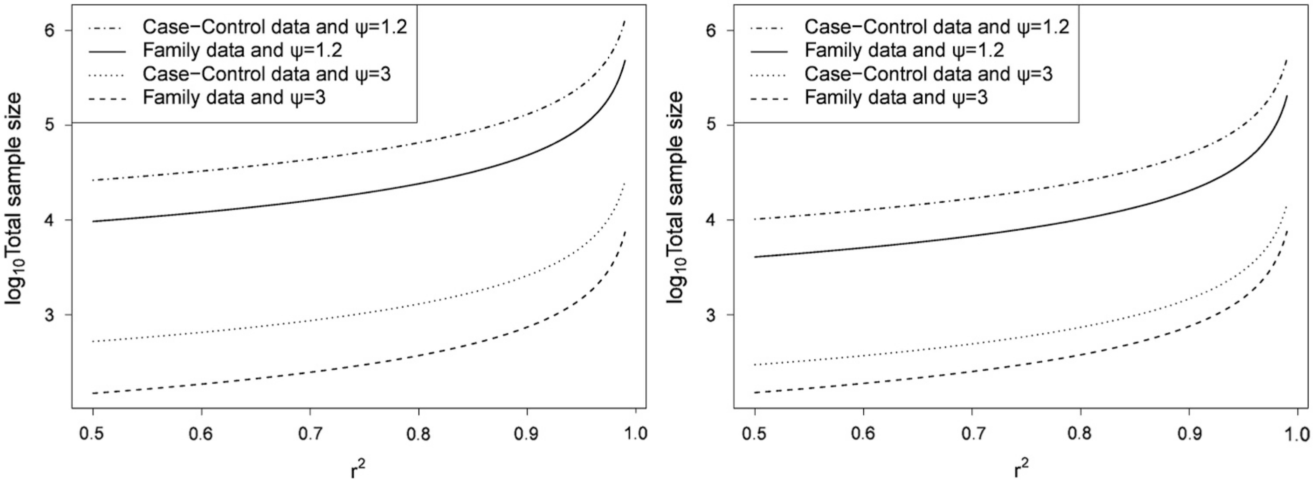

Figure 2 The sample size needed to achieve 90 power to exclude variants at 100:1 odds is plotted as function of r^2, for various values of ψ and for (a) $f_a = 0.1$ and (b) $f_a = 0.5$. The number of cases and controls are assumed to be equivalent.

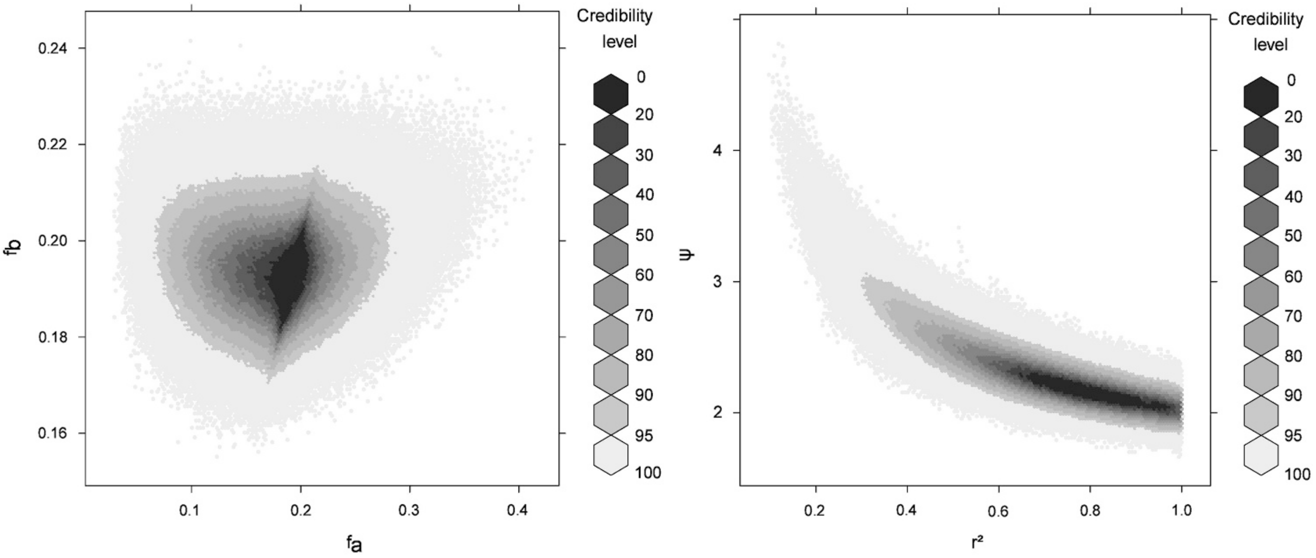

Figure 3 Posterior distributions with Metropolis–Hastings for a simulated sample of 1000 sib-pairs and 1000 controls. The parameters used for simulation are $r^2 = 1$, $f_a = f_b = 0.2$ and $\psi = 2$.

discrimination method, whereas the selected set contains only SNPs 1, 4 and 24, for the family discrimination method.

Metropolis–Hastings on SNPs 24 and 1. Applying the Metropolis–Hastings algorithm on the most associated SNP, that is, SNP 24, the posterior distributions are displayed in Figure 5. The frequency f_a corresponds to the SNP 24 and f_a to the hypothetical causal variant. The 95 credibility region of f_a and OR ψ are (0.24,0.31) and (1.4,3.2), respectively. The LD r^2 is not well estimated, as its 95 credibility region containing almost all possible values. Finally, for the parameter f_a, the disease allele frequency is bimodal, with two modes near 0.3 and 0.8.

Also applying the Metropolis–Hastings algorithm on the second associated SNP, that is, SNP 1, the posterior distributions are displayed in Figure 6. Graphically, the 95 credibility regions of f_b, f_a, ψ and r^2 are ~0.68,0.76), (0.6,0.95), (1.5,5) and (0.1,1), respectively. The mode of f_a is ~0.8.

We have also applied the Metropolis–Hastings algorithm on rs9663421 (SNP 4), which is in the subset of SNPs selected by the family method. The results are similar with those obtained for SNP 1. This is consistent with the observed LD between the SNPs 1 and 4 in our data ($r^2 = 0.85$).

DISCUSSION

Nowadays, research on complex diseases focuses on massive case–control designs, neglecting family designs. However, the joint use of linkage and association information in families allows efficient designs for complex diseases. Using linkage information in association studies results not only in a gain of power in association testing, but also in an increased ability to estimate the risk conferred by the allelic variants, as illustrated in previous papers on Rheumatoid Arthritis[26,27] and MS.[7] The MASC method[28] was developed to exploit all information in family data. The association test from Perdry *et al*[4] is built on the same idea.

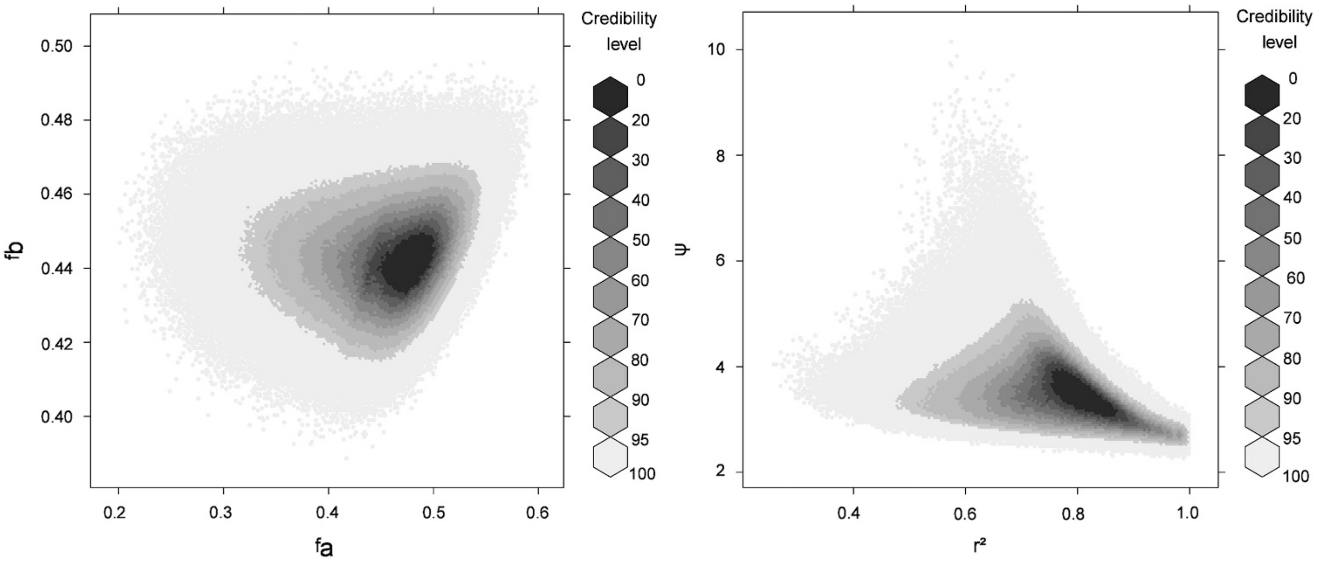

Figure 4 Posterior distributions with Metropolis–Hastings for a simulated sample of 1000 sib-pairs and 1000 controls. The parameters used for simulation are $r^2 = 0.8$, $f_a = 0.435$, $f_b = 0.448$ and $\psi = 3$.

Table 1 Association test statistics and *P*-values; discrimination statistics of all SNPs with SNP 24

		Association test statistics and P-*values*				*Discrimination statistic with SNP 24*	
SNP name		*Case–control*	P-*value*	*Family*	P-*value*	*Case–control*	*Family*
SNP 1	rs12359875	8.75	0.0031	**9.63**	**0.0019**	**2.617**	**5.383**
SNP 2	rs12722605	4.04	0.044	4.88	0.027	**7.323**	10.133
SNP 3	rs12244380	2.25	0.13	3.11	0.078	**9.117**	11.910
SNP 4	rs9663421	5.28	0.022	7.62	0.0058	**6.086**	**7.399**
SNP 5	rs12722596	1.00	0.32	1.70	0.19	10.364	13.315
SNP 6	rs2386841	0.04	0.84	0.01	0.93	11.324	15.009
SNP 7	rs12722588	1.33	0.25	0.64	0.42	10.031	14.374
SNP 8	rs2076846	0.10	0.75	0.21	0.65	11.264	14.805
SNP 9	rs12722561	0.13	0.72	0.27	0.60	11.235	14.743
SNP 10	rs6602392	0.20	0.65	0.26	0.61	11.163	14.757
SNP 11	rs7072398	0.29	0.59	0.53	0.47	11.074	14.484
SNP 12	rs11256456	0.05	0.82	0.002	0.96	11.316	15.014
SNP 13	rs11256457	0.01	0.92	0.17	0.68	11.356	14.843
SNP 14	rs4749924	1.97	0.16	2.11	0.15	9.398	12.904
SNP 15	rs11598648	0.02	0.88	0.31	0.58	11.342	14.702
SNP 16	rs11256497	0.04	0.85	0.13	0.71	11.325	14.883
SNP 17	rs791587	2.03	0.15	3.05	0.081	9.339	11.967
SNP 18	rs791589	0.38	0.54	0.16	0.69	10.989	14.860
SNP 19	rs791590	0.37	0.54	0.61	0.44	10.995	14.409
SNP 20	rs10905669	2.31	0.13	2.39	0.12	**9.051**	12.629
SNP 21	rs2476491	3.72	0.054	4.13	0.042	**7.642**	10.889
SNP 22	rs2256774	0.85	0.36	1.48	0.22	10.513	13.534
SNP 23	rs2104286	1.26	0.26	0.35	0.55	10.110	14.663
SNP 24	rs3118470	**11.37**	**0.00075**	**15.02**	**0.00011**	**0.000**	**0.000**
SNP 25	rs12722489	0.33	0.57	0.44	0.51	11.035	14.581
SNP 26	rs12722486	0.10	0.74	0.004	0.95	11.259	15.013

The bold numbers represent the significant tests. We used Bonferroni correction for association tests and *k*=2ln(100) for discrimination methods.

In this paper, we have shown that sib-pairs provide a gain of power to discriminate between several SNPs associated with a disease. For example, with an OR of 3, $r^2 = 0.9$ and 0.1 alternative allele frequency, the family method needs five times less individuals than the case–control method to achieve similar power. This illustrates well the gain of information provided by family data, as the sib-pairs test uses simultaneously association information comparing control and index cases and linkage information through the IBD. Note that in many

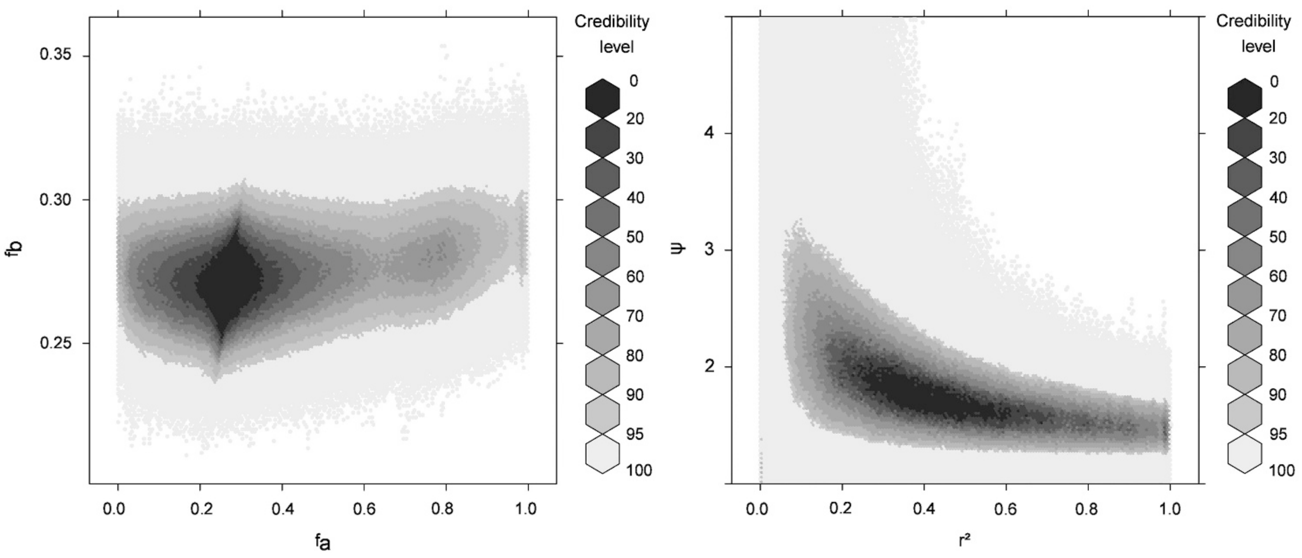

Figure 5 Posterior distributions with Metropolis–Hastings using SNP 24 (rs3118470).

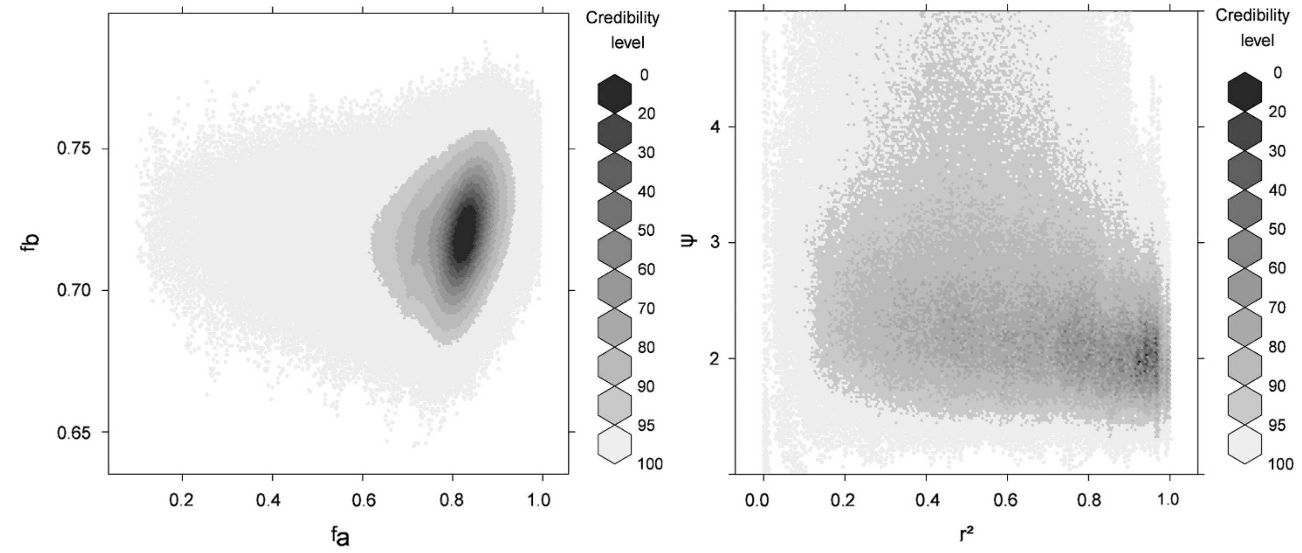

Figure 6 Posterior distributions with Metropolis–Hastings using SNP 1 (rs12359875).

cases, the IBD information is already available from previous linkage studies that have been performed using the same sib-pair sample. If it is not available, it can be obtained through low-density genotyping, which has a negligible cost as compared with the high-density genotyping of the index cases.

Moreover, Udler *et al*'s method for case–control data assumes that the causal variant is genotyped, which is unlikely to be true when using tag-SNPs. Imputation methods, which have been widely used in GWAS, can help to reach a fine enough mapping scale. Nevertheless, we have shown that with family data, thank to the IBD information, we can capture information on the unobserved causal variants through the linked observed variants. This was first done formally, by proving the identifiability of the parameters (section 5 of Supplementary Information). This allows in theory to assess whether the observed variant is the causal variant, or if it is only in LD with the causal variant. Sampling from the posterior distribution of the different parameters allows to obtain credibility regions for allele frequencies, per-allele risk and LD between observed and causal variants. We showed on simulated data that these credibility regions are reasonably accurate.

These methods were illustrated on a real data set, consisting of IL2RA genotypes on MS cases and controls, which were first considered in Babron *et al.*[7] Both discrimination methods provide subsets of SNPs, which likely contain the causal variants. However, the subset obtained using case–control data contains seven SNPs, whereas the subset obtained using the family method contains only three SNPs. This reduced size illustrates the increase of discrimination power. In addition, the family data were additionally used to estimate the risk allele frequency and the OR of a putative causal variant in IL2RA: when using rs12359875 (SNP 1), the method infers a causal variant with a risk allele frequency ~ 0.8, likely different from SNP 1; the value of the LD between SNP 1 and the causal variant and the allelic OR are not well identified. When using rs3118470 (SNP 24), the posterior distribution of the risk allele frequency of the causal variant is bimodal, with one mode ~ 0.8 and another ~ 0.3. This pleads in favor of the existence of a second causal variant with a risk allele frequency near 0.3, in LD with SNP 24 but not with SNP 1, whereas the first causal variant with a risk allele frequency near 0.8 is in LD with both SNPs. In the previous study,[7] the association signal in IL2RA was best captured by an haplotype of rs2256774 (SNP 22) and rs3118470 (SNP 24). Besides that we agree on the fact that the association signal is not due to a single SNP in the region, it is difficult to compare our results with their results, as our approach does not allow to consider several markers at the same time. Also, we considered only a multiplicative risk model, which was not the case in the previous study.[7]

Our approaches could be extended in these directions: considering haplotypes instead of isolated SNPs, droping the multiplicative risk hypothesis, and allowing for multiple disease alleles. If this can improve the disease model inference, a compromise has, however, to be found between the complexity of the model considered and the amount of available information. Considering larger nuclear families or even multiplex families can be a solution to improve the precision of the inference. Another interesting strategy would be to use the difference of LD pattern between cases and controls in the vicinity of the causal variant[29] at the same time as the IBD information.

Identifying variants helps both to improve disease risk prediction and to uncover biological mechanisms involved in human diseases. A better statistical modeling of the effect of the variants in an associated genomic region is a crucial step on this way. In this regard, family design should not be neglected.

Software

The proposed method is accessible in the R package ASPBay available on the Comprehensive R Archive Network (CRAN).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We wish to thank gratefully Françoise Clerget-Darpoux for many fruitful discussions. We thank all the patients who generously participated in and the physicians constituting the REFGENSEP (Réseau français de la génétique de la sclérose en plaques). We also thank all the authors of Babron *et al*[7] for their help on example data set, in particular Marie-Claude Babron. We also thank Rémi Kazma for a careful reading of the manuscript. Hervé Perdry was partly funded by a grant from ARSEP fondation (Fondation pour l'Aide à la Recherche sur la Sclérose en Plaques).

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Manolio T: Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010; **363**: 166–176.
3 Udler MS, Tyrer J, Easton DF: Evaluating the power to discriminate between highly correlated snps in genetic association studies. *Genet Epidemiol* 2010; **34**: 463–468.
4 Perdry H, Müller-Myhsok B, Clerget-Darpoux F: Using affected sib-pairs to uncover rare disease variants. *Hum Hered* 2012; **74**: 129–141.
5 Compston A, Confavreux C, Lassmann H *et al*: *McAlpine's Multiple Sclerosis*. Churchill Livingstone Elsevier, 2005.
6 Goris A, Pauwels I, Dubois B: Progress in multiple sclerosis genetics. *Curr Genomics* 2012; **13**: 646.
7 Babron MC, Perdry H, Handel AE *et al*: Determination of the real effect of genes identified in GWAS: the example of IL2RA in multiple sclerosis. *Am J Hum Genet* 2011; **20**: 321–325.
8 Matesanz F, Caro-Maldonado A, Fedetz M *et al*: IL2RA/CD2 polymorphisms contribute to multiple sclerosis susceptibility. *J Neurol* 2007; **254**: 682–684.
9 Ramagopalan SV, Anderson C, Sadovnick AD, Ebers GC, Matesanz F *et al*: Genomewide study of multiple sclerosis. *N Engl J Med* 2007; **357**: 2199–2200.
10 Hafler DA, Compston A, Sawcer S *et al*: Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 2007; **357**: 851–862.
11 Ramagopalan SV, Anderson C, Sadovnick AD, Ebers GC: Genomewide study of multiple sclerosis. *N Engl J Med* 2007; **357**: 2199–2200.
12 Rubio J, Stankovich J, Field J *et al*: Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in australians. *Genes Immun* 2008; **9**: 624–630.
13 Weber F, Fontaine B, Cournu-Rebeix I *et al*: IL2RA and IL7RA genes confer susceptibility for multiple sclerosis in two independent european populations. *Genes Immun* 2008; **9**: 259–263.
14 Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11**: 375–386.
15 Sasieni PD: From genotypes to genes: doubling the sample size. *Biometrics* 1997; **53**: 1253–1261.
16 Slager S, Schaid D: Case-control studies of genetic markers: Power and sample size approximations for armitage's test for trend. *Hum Hered* 2001; **52**: 149–153.
17 Robert CP, Casella G: *Monte Carlo Statistical Methods.* New York: Springer, 2004, Vol 319.
18 Poser CM, Paty DW, Scheinberg L *et al*: New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 1983; **13**: 227–231.
19 Thomson G: Mapping disease genes: family-based association studies. *Am J Hum Genets* 1995; **57**: 487.
20 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2001; **30**: 97–101.
21 Tröster AI: Refining genetic associations in multiple sclerosis. *Neurology* 2006; **66**: 1830–1836.
22 Alcina A, Fedetz M, Ndagire D *et al*: IL2RA/CD25 gene polymorphisms: uneven association with multiple sclerosis (ms) and type 1 diabetes (t1d). *PLoS One* 2009; **4**: e4137.
23 Dendrou CA, Plagnol V, Fung E *et al*: Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat Genet* 2009; **41**: 1011–1015.
24 Maier LM, Anderson DE, Severson CA *et al*: Soluble IL2RA levels in multiple sclerosis subjects and the effect of soluble IL-2RA on immune responses. *J Immunol* 2009; **182**: 1541–1547.
25 Maier LM, Lowe CE, Cooper J *et al*: IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genet* 2009; **5**: e1000322.
26 Binet J, Auquier A, Dighiero G *et al*: A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* 1981; **48**: 198–206.
27 Bourgey M, Perdry H, Clerget-Darpoux F: Modeling the effect of PTPN22 in rheumatoid arthritis. *BMC Proc* 2007;**1**:S37.
28 Clerget-Darpoux F, Babron M, Prum B, Lathrop G, Deschamps I, Hors J: A new method to test genetic models in HLA associated diseases: the MASC method. *Ann Hum Genet* 1988; **52**: 247–258.
29 Bochdanovits Z, Simón-Sánchez J, Jonker M, Hoogendijk WJ, van der Vaart A, Heutink P: Accurate prediction of a minimal region around a genetic association signal that contains the causal variant. *Eur J Hum Genet* 2013; **22**: 238–242.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)