Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2012 February ; 21(2): 262–272. doi:
10.1158/1055-9965.EPI-11-0719.

Estimating causal effects of genetic risk variants for breast cancer using marker data from bilateral and familial cases

Frank Dudbridge[1], **Olivia Fletcher**[2], **Kate Walker**[1], **Nichola Johnson**[2], **Nick Orr**[2], **Isabel dos Santos Silva**[1], and **Julian Peto**[1]

[1]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

[2]Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, Fulham Road, London, SW3 6JB, UK

Abstract

Background—Cases with a family history are enriched for genetic risk variants, and the power of association studies can be improved by selecting cases with a family history of disease. However in recent genome-wide association scans utilising familial sampling, the excess relative risk for familial cases is less than predicted when compared to unselected cases. This can be explained by incomplete linkage disequilibrium between the tested marker and the underlying causal variant.

Methods—We show that the allele frequency and effect size of the underlying causal variant can be estimated by combining marker data from studies that ascertain cases based on different family histories. This allows us to learn about the genetic architecture of a complex trait, without having identified any causal variants. We consider several validated common marker alleles for breast cancer, using our own study of high risk, predominantly bilateral cases, cases preferentially selected to have at least two affected first or second degree relatives, and published estimates of relative risk from standard case/control studies.

Results—To obtain realistic estimates and to accommodate some prior beliefs, we use Bayesian estimation to infer that the causal variants are probably common, with minor allele frequency >5%, and have small effects, with relative risk around 1.2.

Conclusion—These results strongly support the common disease common variant hypothesis for these specific loci associated with breast cancer.

Impact—Our results agree with recent assertions that synthetic associations of rare variants are unlikely to account for most associations seen in genome-wide studies.

Keywords

Breast cancer; family history; ascertainment bias; causal variant; synthetic association

Introduction

Familial cases of disease are likely to segregate genetic risk variants, and therefore offer increased efficiency over sporadic cases for detecting new variants(1-2). Large collections of familial cases ascertained through genetics clinics(3), and of cases with two primary

Corresponding author: Frank Dudbridge, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. Tel: +44 20 7927 2025. frank.dudbridge@lshtm.ac.uk.

cancers ascertained through cancer registries(4) have been developed specifically for genetic studies of breast cancer. The increased efficiency of studies of these “genetically enriched” cases results from the fact that, for example, the allele frequency difference between cases with one affected first-degree relative and population controls is about 1.5 times greater than in unselected cases, and there is an approximately twofold higher difference for cases with two affected first-degree relatives(1). In breast cancer there is the possibility of bilateral disease, among cases of which there is also a twofold higher difference. Recently, we(5) and others(6) have exploited the genetic enrichment of bilateral and familial cases in genome-wide association scans (GWAS) and thereby identified novel single nucleotide polymorphisms (SNPs) associated with breast cancer.

For a given ascertainment scheme the excess risk, that is the increase in effect size compared to a study of sporadic cases and population controls, can be predicted for variants having a direct causal effect on disease. However, the relationship for marker SNPs is less clear owing to incomplete linkage disequilibrium (LD) between the marker and the causal variant. Significant deviation of the excess risk from its expected value has been observed in familial cases of breast cancer(6), and here we demonstrate similar deviation in bilateral cases. The observed excess risk is generally lower than that predicted for the causal variant, implying that the gain in efficiency is less than had been anticipated.

Here we give expressions for the relative risk for marker genotypes in bilateral and familial cases, in terms of the marker and causal genotype frequencies, LD between marker and causal variant, and the relative risk of the causal variant. We show that the attenuation of the excess risk observed in recent studies can be explained by realistic models of LD and causal relative risks.

We then note that, given the marker relative risks in at least three distinct sampling schemes, it is possible to infer the genotype frequency and relative risk of the causal variant, even if the identity of that variant is unknown. This bears on recent debates over whether the numerous SNP associations emerging from GWAS are primarily driven by rare variants with larger effects than those estimated by GWAS, or by common variants with similar properties to the tag SNPs used in GWAS. The former scenario would explain more heritability than the latter(7), giving a partial account of the missing heritability problem(8). Recently, motivated by earlier observations that rare variation could explain common disease(9), Dickson et al argued that rare variations could stochastically occur in coupling with a common SNP allele, creating “synthetic” association of the common variant(10). But this hypothesis has been challenged on theoretical and empirical grounds, with the evidence to date pointing to a greater role for common causal variation(11-12). These arguments are based on simulations and whole-genome averages, and there has been little work on determining whether specific causal variants are rare or common, as this would normally require complete re-sequencing of associated regions.

Here we address this issue for ten specific loci that have been consistently associated with breast cancer, by inferring the allele frequency and relative risk of the causal variants from the marker relative risks estimated from studies of sporadic, bilateral and familial cases. To allow for sampling variation and to accommodate some prior beliefs, we perform a Bayesian estimation of causal effects to show that the variants underlying these established associations are probably common and have similar relative risks to those observed for the marker SNPs. This is in agreement with recent simulation results(13) and, for the first time, provides explicit support for the common disease common variant (CDCV) hypothesis applied to breast cancer. The approach we develop can be applied to various familial sampling schemes, including those based on concordant or discordant sib pairs, including twins, and is not limited to the studies of breast cancer described here.

Materials and Methods

Subjects

Cases and controls were selected from the British Breast Cancer Study (BBCS), details of which have been published previously (4-5). For this analysis we included 1695 cases, 1564 of whom had two sequential or simultaneous primary breast cancers, and 131 of whom had at least two affected first degree relatives. The excess relative risk among cases with a second primary and those with two affected first degree relatives has been shown to be equivalent(1) so the subsequent analysis assumes that all cases are bilateral. 2001 controls were ascertained as friends and non-blood relatives of the cases. All cases and controls were of self-reported white Caucasian ancestry and all controls were free from breast cancer at enrolment in the study. Collection of blood samples and questionnaire information from case and control subjects was undertaken with informed consent and in accordance with the tenets of the Declaration of Helsinki.

Genotyping

DNA was extracted from blood samples using conventional methodologies and quantified using PicoGreen (Invitrogen). Genotype data from the BBCS was obtained for ten SNPs that have been reproducibly associated with breast cancer by GWAS (table 1). Genotypes for rs13387042, rs4973768 and rs6504950 have been included in previous publications(14-15). Additional genotyping of rs1338740 and *de novo* genotyping of three SNPs (rs10941679, rs2046210, rs11249433) was carried out using using Taqman® nuclease assay, with reagents designed by Applied Biosystems as Assays-by-Design™ and genotyping performed using the ABI PRISM 7900HT Sequence Detection System according to manufacturer's instructions. For four other SNPs (rs889312, rs2981582, rs3817198, rs3803662) genotyping was performed by KBioscience Ltd, using their proprietary in house system (KASPar), a competitive allele specific PCR SNP genotyping system that uses FRET quencher cassette oligos.

Call rates for SNPs genotyped as part of this analysis were 99.9% **(**rs1338740), 99.2% (rs10941679), 99.9% (rs2046210), 99.7% (rs11249433), 98.2% (rs889312), 97.8% (rs2981582), 98.9% (rs3817198) and 98.0% (rs3803662) and there was no evidence of deviation from Hardy Weinberg equilibrium in controls, for any of the SNPs (all $P > 0.05$). Duplicate concordance based on a 3.5 % random sample was 100% for all SNPs.

Published data

We obtained summary odds ratio estimates from published studies using cases unselected for a family history(14-19). We also obtained summary estimates from a recent GWAS by Turnbull et al that preferentially selected cases to have at least two affected first or second degree relatives(6). The exact ascertainment criterion was unspecified and we assumed that half the cases had at least two affected first degree relatives and the other half at least two affected second degree relatives. This approximation was made following informal consultation with those authors, but our conclusions turn out to be similar if we assume, say, that each case has at least one affected first degree and one affected second degree relative.

In what follows we approximate the odds ratio by the relative risk, as this is appropriate for a rare disease and leads to simplification in the analysis. Furthermore several published studies used population-based controls, not selected to be disease free, for which the relative risk was estimated directly.

Effect size in familial cases

Let Y denote disease status, Y=1 for disease present and Y=0 for disease absent. Let M be a diallelic marker with genotypes {0, 1, 2} corresponding to the number of minor alleles present. Let D be a diallelic variant with direct causal effect, also with genotypes {0, 1, 2}.

The relative risk of causal genotype d, compared to baseline 0, is

$$\gamma_{D_d}=\frac{\Pr(Y=1|D=d)}{\Pr(Y=1|D=0)} \quad (1.1)$$

and the relative risk of marker genotype m, compared to baseline 0, is

$$\gamma_{M_m}=\frac{\Pr(Y=1|M=m)}{\Pr(Y=1|M=0)}=\frac{\sum_{d\in\{0,1,2\}}\Pr(Y=1|D=d)\Pr(D=d|M=m)}{\sum_{d\in\{0,1,2\}}\Pr(Y=1|D=d)\Pr(D=d|M=0)}$$
$$=\frac{\sum_{d\in\{0,1,2\}}\gamma_{D_d}f_{D|M}(d|m)}{\sum_{d\in\{0,1,2\}}\gamma_{D_d}f_{D|M}(d|0)} \quad (1.2)$$

where $f_{D|M}(d|m)=\Pr(D=d|M=m)$. The marker relative risk thus depends only on the causal relative risks and the conditional distribution of causal genotype given marker genotype, which reflects the LD, or correlation, between the two genotypes.

Among bilateral cases of disease, the marker relative risk is given by

$$\gamma_{M_m;0}=\frac{\sum_{d\in\{0,1,2\}}\gamma^2_{D_d}f_{D|M}(d|m)}{\sum_{d\in\{0,1,2\}}\gamma^2_{D_d}f_{D|M}(d|0)} \quad (1.3)$$

where the subscript of γ indicates that each case has an affected 0-th degree relative. This expression also holds for cases with an affected monozygotic twin. We assume that controls are not selected for their disease status or family history, and that cancers arise independently in each breast, conditional on other individual-level risk factors.

For familial cases, the marker relative risk is obtained by considering the probability that affected relatives share a causal variant identical by descent (IBD). For first-degree relatives the probability is ½ that a causal variant is shared IBD, and the probability that a subject with genotype d is affected and has at least one affected first-degree relative is given by

$$\Pr(Y=1|D=d)\sum_{j=1}^{\infty}\Pr(J=j)\left(1-\left(1-\frac{1}{2}[\Pr(Y=1|D=d)+\Pr(Y=1)]\right)^j\right) \quad (1.4)$$

where J denotes the number of first-degree relatives for the subject. Assuming that disease risks are small, this is approximated by

$$\Pr(Y=1|D=d)\frac{1}{2}[\Pr(Y=1|D=d)+\Pr(Y=1)]\sum_{j=1}^{\infty}j\Pr(J=j) \quad (1.5)$$

so that the marker relative risk is

$$\gamma_{M_m;1}=\frac{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{2}\left(\gamma_{D_d}+S\right)f_{D|M}(d|m)}{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{2}\left(\gamma_{D_d}+S\right)f_{D|M}(d|0)}\qquad(1.6)$$

where $S=\sum_{d\in\{0,1,2\}}\gamma_{D_d}f_D(d)$ with $f_D(d)=\Pr(D=d)$.

In addition to the conditional distribution $f_{D|M}(d|m)$, the marker and causal relative risks are now also related through the causal genotype distribution $f_D(d)$. Similarly, for a case with at least two affected first-degree relatives, the marker relative risk is approximately

$$\gamma_{M_m;1,1}=\frac{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{2}\left(\gamma_{D_d}+S\right)^2f_{D|M}(d|m)}{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{4}\left(\gamma_{D_d}+S\right)^2f_{D|M}(d|0)}\qquad(1.7)$$

Similar arguments give

$$\gamma_{M_m;2}=\frac{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{4}\left(\gamma_{D_d}+3S\right)f_{D|M}(d|m)}{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{4}\left(\gamma_{D_d}+3S\right)f_{D|M}(d|0)}\qquad(1.8)$$

for cases with at least one affected second degree relative, and

$$\gamma_{M_m;2,2}=\frac{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{16}\left(\gamma_{D_d}+3S\right)^2f_{D|M}(d|m)}{\sum_{d\in\{0,1,2\}}\gamma_{D_d}\frac{1}{16}\left(\gamma_{D_d}+3S\right)^2f_{D|M}(d|0)}\qquad(1.9)$$

for cases with at least two affected second degree relatives. We use the geometric mean of (1.7) and (1.9) to model the summary estimates of Turnbull et al(6), denoted by $\gamma_{M;fam}$.

For each case ascertainment scheme we define the excess risk as the ratio of the log relative risk in selected, familial cases to the log relative risk in unselected, sporadic cases. Thus for bilateral cases the excess risk for genotype m is $\frac{\log\left(\gamma_{M_m;0}\right)}{\log\left(\gamma_{M_m}\right)}$, which is 2 when disease and marker genotypes are perfectly correlated.

Identification of causal effects

From equations 1.2-1.9 the marker and causal relative risks are related through the conditional distribution of causal genotype given marker genotype $f_{D|M}(d|m)$ and the unconditional distribution of causal genotypes $f_D(d)$. Given the marker relative risks from a number of study designs, it is possible in principle to solve for the causal effects. For diallelic marker and causal variant, $f_{D|M}(d|m)$ is specified by six free parameters, $f_D(d)$ by two, and there are two causal relative risks γ_{D_d}. All parameters could therefore be identified given the marker relative risks from ten different familial sampling schemes. This is more than is practical to obtain, but with some mild assumptions we can reduce the parameters to a practical number.

Firstly we assume a multiplicative model of disease risk at the causal variant, in which $\gamma_{D_2} = \gamma_{D_1}^2$ This assumption fits most observed marker associations well, and can be expected to extend to causal variants as well(20). We further assume Hardy-Weinberg Equilibrium (HWE) for both marker and causal variant, so that the genotype frequency distributions can be parameterised by allele frequencies. Under these assumptions all the relative risks factorise into allelic terms and we may simply work with D and M taking values in {0,1}. This leaves four free parameters to be identified: $f_{D|M}(1|1)$, $f_{D|M}(1|0)$, $f_D(1)$ and γ_{D1}. Finally, noting that

$$f_{D|M}(1|0) = \left(1 - \frac{f_{D|M}(1|1)\, f_M(1)}{f_D(1)}\right) \cdot \frac{f_D(1)}{f_M(0)} \qquad (1.10)$$

we assume that the marker allele frequencies $f_M(m)$ are known, so that we have just three free parameters to identify. Therefore, the marker relative risks from just three familial ascertainment schemes are needed to solve for the effects of the causal variant. Here, we use studies of unselected cases (equation 1.2, from published data), cases with bilateral disease (equation 1.3, from our data) and an equal mixture of cases with at least two affected first- or second-degree relatives (equations 1.7 and 1.9, from Turnbull et al).

Bayesian inference of causal effects

In practice we cannot solve for $f_{D|M}(1|1)$, $f_D(1)$ and γ_{D1} exactly because estimates of marker relative risks are subject to sampling variation and do not conform to equations 1.2-1.9. Instead we estimate the causal parameters using a likelihood calculated from the estimated marker effects, assuming that they are obtained from independent samples. Letting β denote log relative risk, $\beta = \log(\gamma)$, the likelihood is

$$\begin{aligned} L\left(\beta_D, f_D, f_{D|M}; \widehat{\beta}_M, \widehat{\beta}_{M;0}, \widehat{\beta}_{M;fam}\right) = \; & \phi\left(\widehat{\beta}_M; \beta_M\left(\beta_D, f_D, f_{D|M}\right), \sigma_M^2\right) \cdot \\ & \phi\left(\widehat{\beta}_{M;0}; \beta_{M;0}\left(\beta_D, f_D, f_{D|M}\right), \sigma_{M;0}^2\right) \cdot \\ & \phi\left(\widehat{\beta}_{M;fam}; \beta_{M;fam}\left(\beta_D, f_D, f_{D|M}\right), \sigma_{M;fam}^2\right) \end{aligned} \qquad (1.11)$$

where $\phi(\cdot; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2, the mean terms are obtained from equations 1.2, 1.3, 1.7 and 1.9, and the variances are assumed equal to their sample estimates.

Maximum likelihood estimation of the causal effects is unsatisfactory for two reasons. Firstly, in the data we consider, the likelihood function can be multimodal and difficult to maximise numerically. Secondly, in a number of instances the likelihood is maximised at parameter values that are unrealistic, such as a very high causal relative risk but minimal correlation between causal and marker variant. This scenario is unreasonable because the SNPs studied were selected as the most strongly associated within their regions, and so are likely to have the highest correlation with the causal variant.

For these reasons we perform a Bayesian analysis using the likelihood in (1.11) and the following prior distributions. For the causal minor allele frequency, we follow the distribution of sequence variants observed in the ENCODE regions by the HapMap consortium(21). This strongly favours variants with frequency < 5%, with roughly equal probability given to frequencies > 20%. This distribution is closely approximated by a Beta(0.345,1.058) distribution scaled by 0.5, which we use as our prior for $f_D(1)$ (Table 2).

The causal relative risk is taken to be log-normal with most of the distribution less than 3, the value above which linkage analysis is expected to be more powerful than GWAS(11). We use a log-normal distribution with location 0 and scale dependent on the causal minor allele frequency according to the distribution proposed by Spencer et al(13) (Table 2).

Together with our prior for the allele frequency this represents 80% belief that $\frac{1}{3} \le \gamma_{D_1} \le 3$ and 90% belief that $\frac{1}{5} \le \gamma_{D_1} \le 5$ a strong but not overwhelming prior belief in moderate relative risks.

The prior for $f_{D|M}$ is more difficult to specify because the range of values it can take depends upon the marker and causal allele frequencies. The prior information in this case consists only of imprecise beliefs that the causal and marker genotypes are in strong LD, but this notion is not easy to parameterise and quantify. GWAS are designed on the premise that all common variants are correlated with a tag SNP at say $r^2 \ge 0.8$, but we would like to allow the possibility that causal variants are rare, as reflected in our prior for $f_D(1)$, which implies lower correlation with associated tag SNPs.

Here we quantify the LD between marker and causal variants by the log odds ratio (logOR) for association between the two minor alleles

$$\theta = \log \frac{f_{D|M}(1|1) / f_{D|M}(0|1)}{f_{D|M}(1|0) / f_{D|M}(0|0)} \quad (1.12)$$

This is unbounded for all values of marker and causal allele frequencies, and the values of $f_{D|M}$ can be expressed as functions of the allele frequencies and θ. As the prior for θ we use a normal distribution with mean 6 and variance 1.48, which corresponds to $r^2 \approx 0.8$ when marker and causal alleles have equal frequency of about 0.25, with 90% probability that $0.5 \le r^2 \le 0.9$ in that case. This reflects generally held beliefs about tag SNPs for common causal variants, while allowing strong statistical associations between rare causal variants and common tag SNPs. This prior strongly favours a positive correlation between the minor alleles of the marker and the causal variant: while a negative correlation is possible, it would imply a much lower r^2 between the two alleles, which conflicts with the assumption that the most informative marker SNP has been chosen from the region.

It turns out that when this prior is used for the logOR, the posterior is virtually unchanged from the prior. We noticed a similar behaviour for other priors that strongly favoured a positive correlation between the two alleles, suggesting that the likelihood is almost independent of θ for high values of θ. Therefore the prior for the logOR has little effect on the posterior distributions for the causal relative risk and allele frequency, which are our main parameters of interest. We return to this point in the discussion.

We obtained posterior distributions for γ_{D_1}, $f_D(1)$ and θ using winBUGS with 100,000 samples, discarding the first 10,000 and keeping all other settings at default values. In addition to posterior median, mode and 95% credible intervals, we calculated the posterior probability that the causal minor allele frequency is less than 5% or 1%.

Results

Marker relative risks in bilateral cases

In table 1 we show the relative risks for ten SNPs estimated from our sample of bilateral cases along with recent literature-based estimates for sporadic and familial cases. It is clear that there is an excess relative risk in bilaterals, confirming that bilateral sampling does offer a gain in efficiency. However this excess is systematically less than 2, the value predicted for a causal variant, so that the gain in efficiency is less than had been anticipated.

These observations can be formally tested. Assuming normality of the log relative risk estimates,

$$\frac{\left(c\widehat{\beta}_{M}-\widehat{\beta}_{M;0}\right)^{2}}{c^{2}\sigma_{M}^{2}+\sigma_{M;0}^{2}}\sim\chi_{(1)}^{2} \quad (2.1)$$

for each SNP, if the hypothesis holds that $c\beta_M = \beta_{M;0}$. Summing (2.1) over all ten SNPs gives a χ^2 variable on 10df which can be regarded as a deviance for the parameter c. The maximum likelihood estimate of c from table 1 is 1.41 which is significantly different from both 1 (p=0.027) and 2 (p=0.003). It is of note that no individual SNP had an excess risk significantly less than 2, but the attenuation is significant when considering all SNPs jointly.

Figure 1 illustrates the excess relative risk in bilateral cases for a marker with allele frequency 0.25, for a range of causal relative risks and allele frequency. The logOR between marker and disease minor alleles is 6, although this value has very little effect on the figure. It is noticeable that the excess risk exceeds 2 when the causal allele is less common than the marker, but is less than 2 if the causal allele is more common. The attenuation is greater when the causal relative risk is higher. This pattern roughly holds for all values of marker and causal parameters, and also when the causal variant has dominant or recessive action (results not shown). Since the excess risk is systematically less than 2 for the markers shown in table 1, this suggests that the causal variants might have risk allele frequency at least equal to that of the marker SNPs. However this must be weighed against the greater prior probability that the causal variant is rare, given the distribution of allele frequency in the genome. The following section addresses this question.

Inference of causal effects

We applied Bayesian estimation of causal effects to the estimated marker effects shown in table 1. The marker MAFs are the most accurate currently available and they were assumed to be known exactly. Tables 3 and 4 show summaries of the posterior distributions of the causal allele frequency and relative risk of the ten associated loci. Kernel density plots of these parameters are shown in Figures 2 and 3.

The posterior estimates of the causal allele frequency have a wide range, roughly 0.1-0.4, whereas the estimates of the causal relative risks are all roughly between 1.2-1.3. The evidence points strongly to the causal variants being common. For only three loci, 11p15 (LSP1, rs3817198), 6q25 (rs2046210) and 17q22 (COX11, rs6504950), is there a reasonable probability that the causal allele has frequency less than 5%, and for each of those the probability that it is less than 1% is considerably lower. Neither rs2046210 nor rs6504950 were directly typed in the familial cases(6); instead a SNP in strong LD was used, and this slight conflict of information may have kept the posteriors closer to the prior than otherwise. The generally high probability of common causal variants is in spite the wide credible intervals on the allele frequencies, which arise from the conflict between the prior that strongly favours rare variants, and the data which are more consistent with common variants. This result is consistent with recent work suggesting that causal variants have similar properties to tag SNPs(11-13, 22), and is in line with the CDCV hypothesis.

Sensitivity analysis

While our prior distributions reflect generally held beliefs about causal variants, there are two important questions that can be asked of our approach. Firstly, noting that our posterior distributions consistently indicated common causal variants, would our procedure identify a rare causal variant if one were present? Secondly, as we could not solve for the causal effects exactly, how much information was gained by combining the estimates from three study designs?

To address the first question, we considered a causal allele with frequency 0.1% and relative risk 3.35, and logOR of 9 with a marker allele of frequency 5%. This represents a rare variant with effect size at 1 standard deviation of its corresponding distribution, and allele frequency fairly close to that of a common marker SNP. From equations 1.2-1.9 the predicted marker relative risk is 1.047 in unselected cases, 1.204 in bilateral cases and 1.222 in cases with a mixture of family histories. The r^2 between marker and causal alleles is only 0.019, but the excess risk is greater than 2 in bilateral and familial cases, in line with figure 1.

It is apparent that this configuration differs from the marker effects we observed for breast cancer, yet it represents a rare causal variant that is consistent with our prior beliefs and is strongly associated with the marker. To determine whether we could infer such a causal variant with our approach, we used standard errors of 0.05, 0.06 and 0.03 for $\widehat{\beta}_{M_1}$, $\widehat{\beta}_{M_1;0}$ and $\widehat{\beta}_{M_1;fam}$ respectively, similar to the higher values appearing in table 1. We then sampled $\widehat{\beta}_{M_1}$, $\widehat{\beta}_{M_1;0}$, and $\widehat{\beta}_{M_1;fam}$ from the corresponding normal distributions and inferred posterior distributions for the causal parameters. This procedure was repeated 1000 times.

On average, the posterior median for the causal allele frequency was 0.10 and the mode was 0.0002. The mean probability was 47% that the causal allele frequency was less than 5%, and 36% that it was less than 1%, the corresponding prior probabilities being 46% and 27%. The average posterior median of the causal relative risk was 1.41 and the mode 1.19. Thus the presence of a rare variant could be inferred reliably, although point estimation of its allele frequency and relative risk appears to be heavily biased and the entire posterior distribution ought to be considered when drawing inferences.

To address the second question of whether the familial samples add information to the inference, we repeated the estimation of causal effects using only the marker relative risks for unselected cases. We adjusted their standard errors to reflect the total information provided from the three studies, using an inverse-variance formula

$$\sigma=\left[\sigma_{M}^{-2}+\sigma_{M;0}^{-2}+\sigma_{M;fam}^{-2}\right]^{-\frac{1}{2}} \quad (2.2)$$

In this way we can assess the information contributed specifically by the study designs as distinct from their additional sample size.

The mean width of the 95% credible interval for the causal allele frequency was 44%, compared to 43% when using three study designs. For the causal relative risk, the mean width was 2.04 compared to 0.78. The posterior median for the causal allele frequency was on average 0.6 times that when using three study designs, whereas the posterior median for the relative risk was 1.2 times higher. Although these results appear to give more support for a rare causal variant than the three study model, this is due to the reduced information in the data (reflected in the wider credible intervals) to move the estimates away from their prior distributions. Moreover the degree of support is still weak. We see that the use of three study designs allows stronger conclusions to be reached than a single study of equivalent sample size, as a result of the increased number of parameters identifiable by including familial cases.

Discussion

One should be cautious about taking our estimates of causal effects too literally. They are dependent on prior distributions, and have wide credible intervals. Although we have shown that our procedure could infer a rare variant if one were present, point estimates of its allele frequency and relative risk are heavily biased. Several groups are currently engaged in fine mapping and resequencing efforts in the regions studied, which will lead to more direct estimates of causal effect sizes. Thus the quantitative estimates presented here will eventually be redundant, although it will be interesting to compare our estimates to the actual causal effects when known.

Instead we emphasise the qualitative nature of our results, which indicate that most, if not all, associations with breast cancer so far identified by GWAS are likely to be markers for common causal variants with modest effects. This is consistent with the CDCV hypothesis that originally motivated GWAS, but not with recent suggestions that many GWAS hits could be markers for rare causal variants(10). In this respect our results agree with other recent work in support of the CDCV hypothesis. Anderson et al(11) argued that GWAS has low power to find a rare variant that had not already been detected by linkage, and noted examples of resequencing projects that had not identified rare variants underlying a common GWAS hit. This includes currently unpublished work by the Wellcome Trust Case-Control Consortium in which sequencing of 16 regions identified by GWAS did not identify any underlying rare causal variant. Wray et al(12) show that the distribution of risk allele frequencies from currently known GWAS hits is consistent with the majority of these hits arising from common variants. Iles(22) showed that early findings of GWAS have been at loci for which the power is highest, which are indeed the common variants. Since we confined attention to SNPs identified in the first wave of breast cancer GWAS, and have subsequently been replicated, we should expect these loci to be enriched for common variants, and in this respect our results are unsurprising. But in contrast to these other studies we are able to estimate causal effects for specific loci rather than average properties of all causal variants. Our results indicate that these loci are consistent with the general pattern of common causal variation suggested by other work, and our methods can be applied to further markers that emerge from GWAS.

Our approach cannot distinguish between the effect of a single common variant and the average effect of a number of variants with a common total frequency. While such a scenario is theoretically possible for a complex disease(9), Wray et al have argued against this scenario for the loci found to date(12). We cannot lend support to either position here other than to note the fact that all ten SNPs indicated a common causal variant, suggesting that if rare variants do underlie these associations then they do so either in large numbers or not at all.

Several authors (13, 20, 22) have used simulations to estimate the empirical conditional distribution of causal allele frequencies and relative risks, given that a marker was identified by GWAS and subsequently replicated. Our approach to modelling the LD between markers and causal variants is much simpler, but we found this model had little effect on the parameters of interest. We do not explicitly model the process of marker discovery by GWAS, and in that respect our prior is more favourable to rare variants, thus strengthening our conclusion that the causal variants are common.

The use of familial cases in association studies is motivated by the excess relative risk in the ascertained sample compared to a sample of unselected cases. We have shown however that imperfect correlation between markers and causal variants leads to an excess risk in familial cases that differs from the predicted value. The difference could be in either direction, and indeed when the causal and marker variants have similar frequency, the excess risk is higher at the marker than at the causal variant, so that the study design is even more efficient than predicted (figure 1). In our data however there was a systematic attenuation of the excess risk in bilateral cases, similar to observations for familial cases in the study of Turnbull et al(6), which is most consistent with causal variants of higher frequency than the markers. The efficiency of bilateral sampling, while still greater than that of unselected sampling, appears to be less than predicted, and this may have implications for the design of future studies of common genetic risk factors.

Some other mechanisms can also lead to attenuation of the excess risk. We assumed a multiplicative model in which each copy of the risk allele multiplies the disease risk to the same degree, but the true model could be recessive, dominant or more general. We can rewrite equations 1.2-1.9 in terms of recessive or dominant effects: it turns out that under a recessive model the excess risk attenuates at higher causal frequencies than under the multiplicative model, whereas for a dominant model it attenuates at lower frequencies (results not shown). Dominant causal variants could therefore be more consistent with rare variation than the multiplicative model considered, but the relevant probabilities remained low when we assumed this model in our analyses, and for brevity we have omitted these results.

We have also assumed that effects act on the log-risk scale, which is convenient as additional polygenic and environment effects cancel out of relative risk calculations so we need not assume a model for them. If however the effects act on say the logistic or probit scales, then the excess relative risk would be attenuated even at the causal variant. We considered this possibility by allowing for a normally distributed polygenic random effect with mean zero and variance 2log(2), consistent with a sibling relative recurrence of 2(23). Acting on the logistic scale this could reduce the excess relative risk for the causal variant from 2 to 1.8 in bilateral cases, but this is less than the degree of attenuation we observed in our data. Subgroup effects, such as age or tumour subtype specific risks, could also attenuate the marginal excess risk, but we did not observe any such effects in our data.

We have shown that genetic markers of breast cancer have lower excess risk in familial cases than had been predicted, leading to reduced improvements of efficiency in these study designs. However this information can be usefully exploited to estimate the relative risk and allele frequency of the underlying causal variants. Despite using a prior distribution that favours rare variation, we showed that data from bilateral and familial cases strongly imply that the causal variants underlying recent GWAS findings are common with modest effects, in line with other recent work favouring the CDCV hypothesis. We look forward to the outcome of current fine mapping projects to confirm the accuracy of these predictions.

Acknowledgments

Grant support

This work was supported by the Medical Research Council (G1000718 to FD), Cancer Research UK (C150/A5660 and C1178/A3947 to JP and IdSS), and Breakthrough Breast Cancer (OF, NJ, NO). We acknowledge NHS funding to the NIHR Royal Marsden Biomedical Research Centre and the National Cancer Research Network (NCRN).

Financial support: Medical Research Council (G1000718 to FD), Cancer Research UK (C150/A5660 and C1178/A3947 to JP and IdSS), Breakthrough Breast Cancer (OF, NJ, NO)

References

1. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: Implications for design of association studies. Genet Epidemiol. Nov; 2003 25(3):190–202. [PubMed: 14557987]

2. Begg CB, Berwick M. A note on the estimation of relative risks of rare genetic susceptibility markers. Cancer Epidemiol Biomarkers Prev. Feb; 1997 6(2):99–103. [PubMed: 9037560]

3. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. Aug; 2006 38(8):873–5. [PubMed: 16832357]

4. Johnson N, Fletcher O, Naceur-Lombardelli C, dos Santos Silva I, Ashworth A, Peto J. Interaction between CHEK2*1100delC and other low-penetrance breast-cancer susceptibility genes: a familial study. Lancet. 2005; 366(9496):1554–7. Oct 29-Nov 4. [PubMed: 16257342]

5. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst. Mar 2; 2011 103(5):425–35. [PubMed: 21263130]

6. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet. Jun; 2010 42(6): 504–7. [PubMed: 20453838]

7. Hemminki K, Bermejo JL. Constraints for genetic association studies imposed by attributable fraction and familial risk. Carcinogenesis. Mar; 2007 28(3):648–56. [PubMed: 17012223]

8. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. Oct 8; 2009 461(7265):747–53. [PubMed: 19812666]

9. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. Jul; 2001 69(1):124–37. [PubMed: 11404818]

10. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. Jan.2010 8(1):e1000294. [PubMed: 20126254]

11. Anderson CA, Soranzo N, Zeggini E, Barrett JC. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol. 2011; 9(1):e1000580. [PubMed: 21267062]

12. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 2011; 9(1):e1000579. [PubMed: 21267061]

13. Spencer C, Hechter E, Vukcevic D, Donnelly P. Quantifying the underestimation of relative risks from genome-wide association studies. PLoS Genet. Mar.2011 7(3):e1001337. [PubMed: 21437273]

14. Milne RL, Benitez J, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, et al. Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. J Natl Cancer Inst. Jul 15; 2009 101(14):1012–8. [PubMed: 19567422]

15. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet. May; 2009 41(5): 585–90. [PubMed: 19330027]

16. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. Jun 28; 2007 447(7148):1087–93. [PubMed: 17529967]

17. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet. Jul; 2007 39(7):865–9. [PubMed: 17529974]

18. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet. Mar; 2009 41(3):324–8. [PubMed: 19219042]

19. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1 p11.2 and 14q24.1 (RAD51L1). Nat Genet. May; 2009 41(5):579–84. [PubMed: 19330030]

20. Vukcevic D, Hechter E, Spencer C, Donnelly P. Disease model distortion in association studies. Genet Epidemiol. Mar 17.2011

21. A haplotype map of the human genome. Nature. Oct 27; 2005 437(7063):1299–320. [PubMed: 16255080]

22. Iles MM. What can genome-wide association studies tell us about the genetics of common disease? PLoS Genet. Feb.2008 4(2):e33. [PubMed: 18454206]

23. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. May; 2002 31(1):33–6. [PubMed: 11984562]

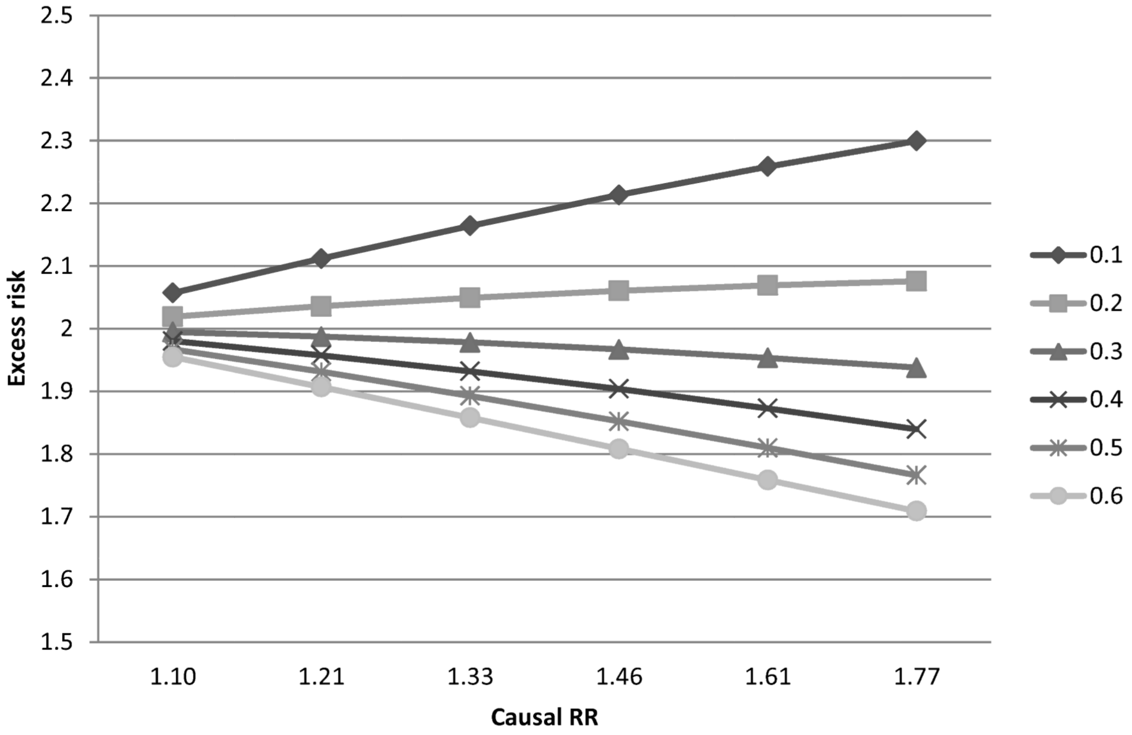

Figure 1.
Excess relative risk in bilateral cases for a marker with allele frequency 0.25. Excess risk is the ratio of the log relative risk in bilateral cases to that in unselected cases. Each line corresponds to a frequency of the minor allele at the causal variant, which is associated to the marker minor allele with log odds ratio of 6.

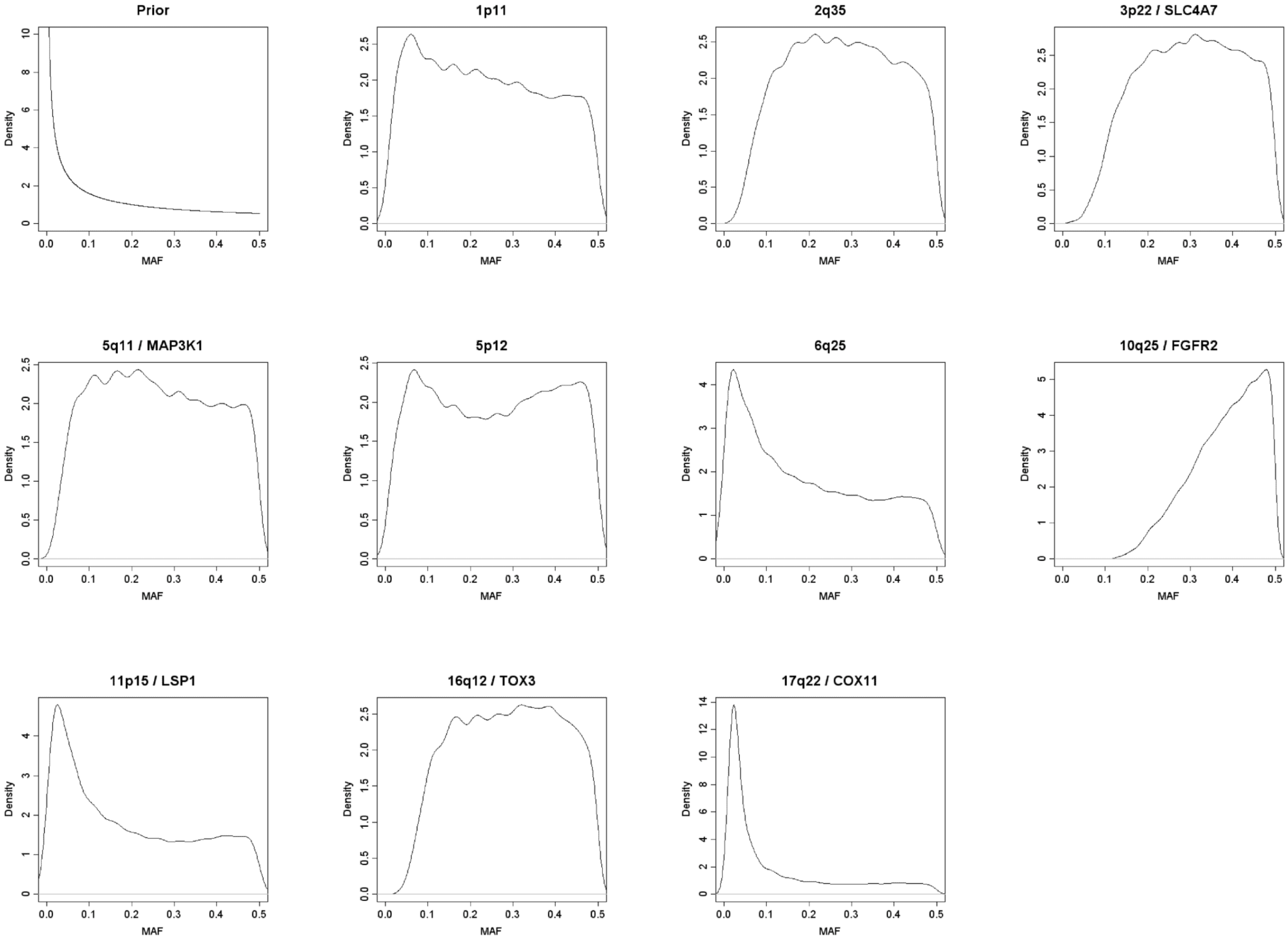

Figure 2.
Kernel density plots of prior and posterior distributions of the causal allele frequency at ten breast cancer loci.

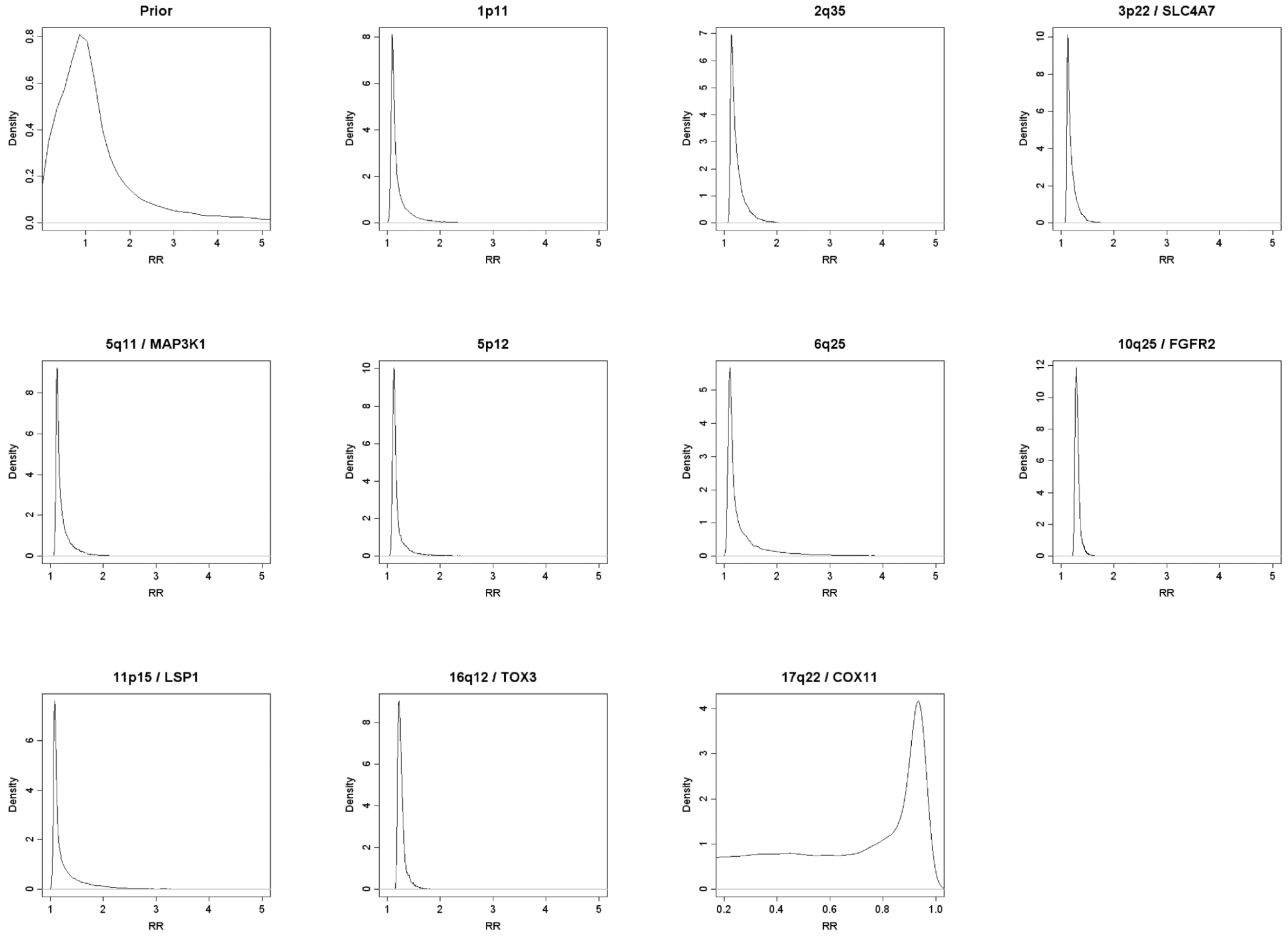

Figure 3.
Kernel density plots of prior and posterior distributions of the causal relative risk at ten breast cancer loci.

Table 1

Per-allele relative risks of associated SNPs considered in this study.

SNP	Band	Gene	Minor allele frequency	Unselected cases	Familial cases	Bilateral cases
rs11249433	1p11		0.39	1.16 (1.09-1.24)[a]	1.08 (1.02-1.15)	1.22 (1.09-1.35)
rs13387042	2q35		0.49	1.12 (1.09-1.15)[b]	1.21 (1.14-1.29)	1.09 (0.97-1.21)
rs4973768	3p22	SLC4A7	0.46	1.11 (1.08-1.13)[c]	1.16 (1.10-1.24)	1.11 (0.99-1.23)
rs10941679	5p12		0.25	1.19 (1.11-1.28)[d]	1.11 (1.04-1.19)[†]	1.29 (1.15-1.43)
rs889312	5q11	MAP3K1	0.38	1.12 (1.08-1.16)[e]	1.22 (1.14-1.30)	1.11 (1.01-1.21)
rs2046210	6q25		0.34	1.15 (1.03-1.28)[f]	1.15 (1.08-1.22)[†]	1.12 (0.99-1.25)
rs2981582	10q25	FGFR2	0.38	1.26 (1.22-1.29)[e]	1.43 (1.35-1.53)	1.39 (1.29-1.49)
rs3817198	11p15	LSP1	0.3	1.07 (1.04-1.11)[e]	1.12 (1.05-1.19)	1.10 (1.00-1.20)
rs3803662	16q12	TOX3	0.25	1.19 (1.15-1.23)[e]	1.30 (1.22-1.39)	1.41 (1.30-1.52)
rs6504950	17q22	COX11	0.27	0.95 (0.92-0.97)[c]	0.92 (0.86-0.99)[†]	0.93 (0.79-1.07)

Estimates for unselected cases are taken from

[a] ref (19)

[b] ref (14)

[c] ref (15)

[d] ref (17)

[e] ref (16)

[ƒ] ref (18). Estimates for familial cases are taken from ref (6). Estimates for bilateral cases are taken from this study (see Methods).

[†] Relative risk for a proximal SNP in LD (r^2>0.75) with the lead SNP.

Table 2

Prior distribution of minor allele frequency (MAF) and standard deviation of log relative risk. “ENCODE frequency” gives the proportion of variants in the ENCODE regions having MAF in the stated range. “SD of log RR” gives the standard deviation of the log relative risk for SNPs according to the model of Spencer et al. (13)

Minor allele frequency (%)	ENCODE frequency (%)	SD of log RR
0-5	46	1.21
5-10	13	0.84
10-15	10	0.61
15-20	6	0.46
20-25	5	0.36
25-30	5	0.3
30-35	4	0.25
35-40	4	0.23
40-45	4	0.21
45-50	3	0.2

Table 3

Posterior distribution summaries for causal minor allele frequencies, estimated from 100,000 MCMC samples. MAF, minor allele frequency. CI, credible interval. Pr(<5% (1%)), probability that causal allele frequency is less than 5% (1%).

Band	Gene	Marker SNP	Marker MAF	Median	Mode	95%CI	Pr(<5%)	Pr(<1%)
(Prior)				0.063	0.00	10^{-5}-0.46	0.46	0.27
1p11		rs11249433	0.39	0.23	0.06	0.021-0.48	0.093	0.0060
2q35		rs13387042	0.49	0.28	0.25	0.069-0.49	0.0071	$<10^{-5}$
3p22	SLC4A7	rs4973768	0.46	0.30	0.31	0.097-0.49	0.0017	$<10^{-5}$
5p12		rs10941679	0.25	0.26	0.07	0.023-0.49	0.067	0.0034
5q11	MAP3K1	rs889312	0.38	0.25	0.21	0.045-0.49	0.032	$<10^{-5}$
6q25		rs2046210	0.34	0.16	0.02	0.0055-0.48	0.22	0.051
10q25	FGFR2	rs2981582	0.38	0.40	0.48	0.21-0.49	$<10^{-5}$	$<10^{-5}$
11p15	LSP1	rs3817198	0.3	0.16	0.03	0.0074-0.48	0.23	0.038
16q12	TOX3	rs3803662	0.25	0.29	0.38	0.084-0.49	0.0018	$<10^{-5}$
17q22	COX11	rs6504950	0.27	0.051	0.02	0.013-0.46	0.49	0.0033

Table 4

Posterior distribution summaries for causal relative risks, estimated from 100,000 MCMC samples. RR, relative risk. CI, credible interval.

Band	Gene	Marker SNP	Marker RR	Median	Mode	95%CI
(Prior)				1.00	1.00	0.14-7.24
1p11		rs11249433	1.16	1.13	1.09	1.06-1.88
2q35		rs13387042	1.12	1.19	1.14	1.11-1.67
3p22	SLC4A7	rs4973768	1.11	1.16	1.12	1.10-1.46
5p12		rs10941679	1.19	1.14	1.12	1.08-1.75
5q11	MAP3K1	rs889312	1.12	1.16	1.13	1.10-1.71
6q25		rs2046210	1.15	1.17	1.11	1.07-3.12
10q25	FGFR2	rs2981582	1.26	1.30	1.29	1.25-1.43
11p15	LSP1	rs3817198	1.07	1.13	1.08	1.06-2.48
16q12	TOX3	rs3803662	1.19	1.24	1.22	1.18-1.50
17q22	COX11	rs6504950	0.95	0.73	0.93	0.095-0.95