Constraints for genetic association studies imposed by attributable fraction and familial risk

Kari Hemminki1,2,∗ and Justo Lorenzo Bermejo1

1Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany and 2Center for Family Medicine, Karolinska Institute, 141 83 Huddinge, Sweden

∗To whom correspondence should be addressed.
Email: k.hemminki@dkfz.de

Candidate gene studies have become very popular but some of their implicit constraints, such as the familial risk and the population attributable fraction (PAF) conferred by the gene under study, are poorly understood. We model here these parameters for susceptibility genes in terms of genotype relative risk (GRR), allele frequency and statistical power in simulated genetic association studies, assuming 500 or 2000 case–control pairs and different modes of inheritance. The results show that the common association studies on genes with minor allele frequency >10% have sufficient power to detect disease-causing variants conferring PAFs >10%, which can be compared to known genes, such as BRCA1 with a PAF of 1.8%. Yet, common low-risk variants confer low familial relative risks (FRRs), typically <1.1. The models show that candidate gene studies may be able to identify genes conferring close to 100% of the PAF, but they may not explain the empirical FRRs. In order to explain FRRs, rare, high-penetrant genes or interacting combinations of common variants need to be uncovered. However, the candidate gene studies for common alleles do not target this class of genes. The results may challenge the common disease–common variant hypothesis, which posits common variants with low GRRs and large PAFs, however failing to accommodate the empirical FRRs.

Introduction

The strategy for dissecting genetics of complex diseases has been debated, particularly regarding the assumptions of the common disease–common variant (CDCV) hypothesis (1–6). Some statistical issues of concern in candidate gene and genome-wide studies have been sample size, level of significance, marker densities, population stratification and replication of results (2,5,7,8). However, little attention has been paid to the population attributable fractions (PAFs) and the familial relative risks (FRR) conferred by the tested gene, both of which limit the domain of the possible genotype relative risks (GRRs) and allele frequencies, and vice versa. We illustrate these concepts in Figure 1. The attractive aspects of considering PAFs include, first, the possibility to compare with PAFs of known genes, as shown here, and, second, the independence of PAFs on the number of unobserved genes and their interactions with the gene under study. The PAF for the gene under study merely states the contribution of the measured gene to disease etiology, regardless of unmeasured genetic effects. The combined PAFs of the independent susceptibility genes cannot exceed 100%, but yet they have to be able to explain completely the familial aggregation observed in family studies (9). Obviously, frequencies of variant alleles vary between populations and the corresponding PAFs vary accordingly.

Familial aggregation is usually measured as FRR, which compares the risk of disease for relatives of patients to that for the general population. PAF and FRR address some underpinnings of the CDCV paradigm and they should be used a priori, to assess the feasibility of association studies and a posteriori, to evaluate the consistency of the results with population based data (10,11). The present study explores the constraints posed by allele frequency (q), GRR, PAF, FRR and statistical power on genetic association studies. Along the presentation, we show how the high-penetrant breast cancer gene BRCA1 and the novel low risk, relatively common chromosome 8 prostate cancersusceptibility locus DG8S737 (12) fit into the calculations.

Methods

The developed models denote the frequency of a susceptibility allele A by q, the relative risk for variant homozygotes compared to wild-type genotypes by GRR, the proportion of cases attributable to the susceptibility alleles by PAF, the familial relative risk by FRR, the prevalence of the disease in the population by κ and the prevalence of the disease among individuals with wild-type genotypes by f. We explored first the relationship between q and GRR for a fixed PAF. The following calculations are based on a dominant model, which assumes that the risk of heterozygotes (Aa) equals that of the wild-type homozygotes, for the additive model, in which the risk of heterozygotes equals that of the wild-type homozygotes, and for the multiplicative model, where the risk for heterozygotes is GRR2. The probability that an individual in the population is wild-type homozygote (G = aa) and he is affected (D = 1) is:

\[ Pr(G = aa, D = 1) = (1 - q)^2 f \]

Similarly,

\[ Pr(G = Aa, D = 1) = 2q(1 - q)GRRf \]

\[ Pr(G = AA, D = 1) = q^2 \ GRRf \]

The prevalence of the disease in the population is then:

\[ \kappa = Pr(G = aa, D = 1) + Pr(G = Aa, D = 1) + Pr(G = AA, D = 1), \]

and the PAF is:

\[ PAF = (\kappa - f) / \kappa \]

Abbreviations: CDCV, common disease–common variant; FRR, familial relative risk; GRR, genotype relative risk; PAF, population attributable fractions; SNP, single nucleotide polymorphism.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
We also investigated the relationship between $q$ and the power of a case-control study of the susceptibility gene for a fixed PAF. Note that the expected distribution of cases according to genotype is given by:

$$Pr(G = aa | D = 1) = \frac{Pr(G = aa)}{k}$$

$$Pr(G = Aa | D = 1) = \frac{Pr(G = Aa)}{k}$$

$$Pr(G = AA | D = 1) = \frac{Pr(G = AA)}{k}$$

Analogous calculations permitted to derive the expected distribution among controls. In order to estimate the power of a case-control study, genotypes were simulated for 500 cases/500 controls and 2000 cases/2000 controls according to their expected distributions using the R function RUNIF. We assumed Hardy–Weinberg equilibrium and $k = 5\%$, but the results were practically independent of the prevalence of the disease in the population. For each allele frequency between 1 and 90%, 10,000 datasets were generated and the association between genotype and disease was analyzed by logistic regression using the Wald tests. The power of the study was estimated as the proportion of simulated datasets, which resulted in significant genotype effects at the 5% confidence level.

Since the FRR reflects the effects of susceptibility genes transmitted from parents to their offspring, we analyzed the relationship between $q$ and the FRR for a fixed GRR. The familial relative risk for parents and their children is given by:

$$FRR = 1 + \frac{1}{2} \frac{V_a}{K^2}$$

where $V_a$ is the additive genetic variance divided by $q^2$ and $K = \kappa f$.

$$V_a = 2q(1 - q)(1 - GRR)$$

under a dominant model, under additive inheritance $V_a = 2q(1 - q)(1 - GRR)^2$, and $V_a = 2p(1 - q)$ $[q(1 - GRR)]^2$ in the recessive model (13,14). Data on GRR and $q$ for BRCA1 mutations were taken from the literature (15). Results on the prostate cancer susceptibility locus DG8S737 with a confirmed effect in several populations have been recently published (12). The code for calculation of PAF, FRR and power using the free software environment R (www.r-project.org) is provided in the supplementary material.

PAFs and FRRs were also calculated for multiple interacting genes. We present here the formulas for two dominant alleles and multiplicative gene-gene interactions, their extension to other models/additional genes is straightforward. We know that:

$$PAF = (\kappa - f)/\kappa = [(\kappa f) - 1]/(\kappa f)$$

Since the FRR reflects the effects of susceptibility genes transmitted from parents to their offspring, we analyzed the relationship between $q$ and the FRR for a fixed GRR. The familial relative risk for parents and their children is given by:

$$FRR = 1 + \frac{1}{2} \frac{V_a}{K^2}$$

where $V_a$ is the additive genetic variance divided by $q^2$ and $K = \kappa f$.

$$V_a = 2q(1 - q)(1 - GRR)$$

under a dominant model, under additive inheritance $V_a = 2q(1 - q)(1 - GRR)^2$, and $V_a = 2p(1 - q)$ $[q(1 - GRR)]^2$ in the recessive model (13,14). Data on GRR and $q$ for BRCA1 mutations were taken from the literature (15). Results on the prostate cancer susceptibility locus DG8S737 with a confirmed effect in several populations have been recently published (12). The code for calculation of PAF, FRR and power using the free software environment R (www.r-project.org) is provided in the supplementary material.

PAFs and FRRs were also calculated for multiple interacting genes. We present here the formulas for two dominant alleles and multiplicative gene-gene interactions, their extension to other models/additional genes is straightforward. We know that:

$$PAF = (\kappa - f)/\kappa = [(\kappa f) - 1]/(\kappa f)$$
where (x/y) is the population prevalence divided by the prevalence among wild type individuals. Similarly, it is easy to show that

\[
\text{FRR} = \frac{Pr(P = 1, D = 1)}{Pr(P = 1, D = 1)} = \left(\frac{2}{q}\right)^2
\]

where \( Pr(P = 1, D = 1) \) denotes the probability that both the parent \( P \) and his/her offspring \( D \) are affected. We have seen that:

\[
Pr(G = aa, D = 1) = (1 - q)^2 = a_1
\]

\[
Pr(G = AA, D = 1) = 2q(1 - q)GRR = a_2
\]

\[
Pr(G = aa, D = 1) = q^2 GRR = a_3
\]

If we consider two unlinked genes, the term needed to calculate the PAF is:

\[
\sum_{i=1,3} \sum_{j=1,2} a_i b_j
\]

Moreover, assuming random mating:

\[
Pr(G_P = aa, G_O = aa, P = 1, D = 1) = 0.5 Pr(G = aa, D = 1)^2 + 0.5 Pr(G = aa, D = 1) Pr(G = aa, D = 1) = b_1
\]

\[
Pr(G_P = aa, G_O = aa, P = 1, D = 1) = 0.5 Pr(G = AA, D = 1)^2 + 0.5 Pr(G = AA, D = 1) Pr(G = aa, D = 1) = b_2
\]

\[
Pr(G_P = AA, G_O = aa, P = 1, D = 1) = 0.5 Pr(G = aa, D = 1)^2 + 0.5 Pr(G = AA, D = 1) Pr(G = aa, D = 1) = b_3
\]

\[
Pr(G_P = AA, G_O = aa, P = 1, D = 1) = 0.5 Pr(G = AA, D = 1)^2 + 0.5 Pr(G = AA, D = 1) Pr(G = aa, D = 1) = b_4
\]

\[
Pr(G_P = aa, G_O = AA, P = 1, D = 1) = 0.5 Pr(G = aa, D = 1)^2 + 0.5 Pr(G = AA, D = 1) Pr(G = aa, D = 1) = b_5
\]

\[
Pr(G_P = AA, G_O = AA, P = 1, D = 1) = 0.5 Pr(G = AA, D = 1)^2 + 0.5 Pr(G = AA, D = 1) Pr(G = aa, D = 1) = b_6
\]

\[
Pr(G_P = AA, G_O = AA, P = 1, D = 1) = Pr(G = AA, D = 1)^2 + 0.5 Pr(G = AA, D = 1) Pr(G = aa, D = 1) = b_7
\]

where \( G_P \) represents the genotype of the parent and \( G_O \) the genotype of his/her offspring. Finally,

\[
Pr(P = 1, D = 1) = \sum_{i=1,3} \sum_{j=1,2} b_i b_j
\]

**Results**

Figure 2 shows the dependence of GRR on \( q \) at various PAFs under three different modes of inheritance. We show a vertical line at \( q = 0.1 \) because many association studies do not test rarer variants. For a dominant gene with a PAF <10%, the GRRs are <2.0 at any \( q > 0.1 \). We have marked the coordinates of the breast cancer-susceptibility gene in Figure 1. BRCA1 mutations with a GRR of \( \sim 10 \) and \( q = 0.001 \) explain a PAF of 1.8%. The novel prostate cancer-susceptibility locus, \( DG8S737 \) with a GRR of \( 1.77 \times 1.77 = 3.13 \) and \( q = 0.078 \) confers a PAF of 11%. A comparison among the three modes of inheritance (from dominant to additive and to recessive) shows a gradual shifting of the curves towards higher GRRs for a fixed allele frequency.

Figure 3 assesses the statistical power to detect a significant GRR (significance level 0.05) in case–control studies using either 500 cases and controls or 2000 cases and controls. Under dominance, a study based on 500 cases and 500 controls reaches an 80% power (shown by the horizontal line) only for rare alleles \( (q < 0.1) \) when the PAF was 5 or 10%; larger studies based on 2000 cases and controls reach an 80% power with \( q = 0.1 \) for a PAF of 5% and with \( q = 0.3 \) for a PAF of 10%. The analysis of 2000 cases and 2000 controls does not provide reasonable power for a PAF of 1%. Using 500 cases and controls, the effect of \( BRCA1 \) mutations is detected with a power of 59% and the effect of \( DG8S737 \) with a power of 93%. Again, the curves for additive and recessive modes shift systematically to the right, indicating higher power at higher allele frequencies.

The relationship between allele frequency and FRR according to a fixed GRR is shown in Figure 4. Dominant genes result in increased FRRs at low allele frequencies; for example a GRR of 10 results in a FRR of \( \sim 2.0 \) for \( q = 0.05 \). The FRRs conferred by \( BRCA1 \) (1.08) and \( DG8S737 \) (1.04) are also represented in the Figure. Again, the curves and their maxima are right-shifted for additive and recessive genes. For an additive gene, a GRR of 10 and \( q = 0.1 \) confers a familial risk of 1.5; under recessive inheritance, GRR = 10 and \( q = 0.4 \) results in FRR = 1.5.

Table I shows PAFs and FRRs for various values of \( q \) and GRRs. PAFs increases continuously with the allele frequency. By contrast, the FRR increases with the allele frequency to a maximum, and decreases thereafter. For example, a dominant allele with a GRR of 3 shows the maximal FRR (1.15) at \( q = 0.1 \) and it confers a PAF of 27.5%; when \( q = 0.5 \), the FRR decreases to 1.04, but the PAF increases to 60%. Rare variants explain relatively more of FRR than of PAF (e.g. \( BRCA1 \) mutations: FRR = 1.08, PAF = 1.8%), in contrast to common variants (\( DG8S737 \): FRR = 1.04, PAF = 11%).

Table II shows PAFs and FRRs for some basic interaction models, considering up to 10 genes with \( q = 0.1 \) and GRR = 2. The assumed allele–allele interactions were dominant, additive or recessive; multiplicative and additive gene–gene interactions were considered. Multiplicative interactions of genes with dominant alleles cause the highest PAFs and FRRs; with 10 interacting genes, the PAF was 82.4% and the FRR was 1.65. Additive interactions of genes with dominant alleles and multiplicative/additive interactions of genes with additive alleles contributed relatively more to the PAF than to the FRR. The impact of recessive alleles on PAFs and FRRs was relatively small.

**Discussion**

FRR is an indicator of heritability, assuming that the contribution of shared environmental risk factors to the familial
Fig. 2. Relationship between allele frequency (q) and GRR for a fixed PAF, according to the inheritance mode (A, dominant; B, additive; and C, recessive). The genetic parameters for DG8S737 were calculated assuming a multiplicative model.
Fig. 3. Relationship between allele frequency (q) and statistical power to detect a significant gene–disease association (Type I error = 5%) for a fixed PAF, according to the inheritance mode and the sample size. The genetic parameters for DG8S737 were calculated assuming a multiplicative model.
Fig. 4. Relationship between the allele frequency (q) and the FRR for a fixed GRR, according to the inheritance mode (A, dominant; B, additive; and C, recessive). The genetic parameters for DG8S737 were calculated assuming a multiplicative model.
aggregation of disease is small. FRR for most types of cancer are ~2.0 (16), and it is likely that the familial aggregation of cancer is mostly due to heritable causes (17,18). The FRR of colorectal cancer is ~1.5, when known conditions, such as hereditary non-polyposis colorectal cancer (HNPCC) and familial adenomatous polyposis are excluded (19,20). The combined PAF of the related high penetrance genes, mismatch repair genes and APC for colorectal cancer has been estimated to range from 1 to 3% in Western populations (21).

According to Figure 4, a FRR of 1.5 can be explained by a dominant gene (or the combined effect of many genes) of \( q = 0.1 \) or 0.2; for additive or recessive genes with FRR = 10, the allele frequency should be 0.1 or 0.4, respectively. According to Table I, a dominant gene with the described parameters would show a PAF of 15% (\( q = 0.01 \)) or 76% (\( q = 0.2 \)), compared with a PAF of 47% for an additive gene and a PAF of ~50% for a recessive gene.

Many candidate gene association studies focus on genes for which the minor allele frequency is 0.1 or higher because the likelihood of identifying a significant effect (statistical power) is higher for common alleles, assuming a constant GRR (22,23). In the HapMap Project, the minor allele frequency was restricted to 0.1, GRR = 2. Therefore, a case–control study based on 500 cases and 500 controls only has a reasonable power (>80%) to detect dominant and additive variants with PAFs over 10%, but a recessive gene could be identified even at a lower PAF. If 500 cases and 500 controls were used to investigate the effects of rare variants, the likelihood of detecting the effect of \( BRCA1 \) mutations would be 59%; the \( DG8S737 \) locus would be detected with a power of 93% (Figure 3). In fact, a case–control study based on 500 cases and 500 controls only has a reasonable power (>80%) to detect dominant and additive variants with PAFs over 10%, but a recessive gene could be identified even at a lower PAF. If 500 cases and 500 controls were used to investigate the effects of rare variants, the likelihood of detecting the effect of \( BRCA1 \) mutations would be 59%; the \( DG8S737 \) locus would be detected with a power of 93% (Figure 3).

\( BRCA1 \) was originally identified in linkage studies, but the effects of specific mutations were statistically significant in a case–control study based on 2000 breast cancer patients and 4000 controls from Poland (15), in agreement with the present data. The novel prostate cancer-susceptibility locus \( DG8S737 \) was mapped in 323 Icelandic families with a lod score 2.11 and confirmed in a case–control study on 869 cases and 596 controls. It was subsequently replicated in three other populations, one Swedish and two others European and African American (12). The

### Table I. PAF(%) and FRR according to allele frequency (q) and GRR

<table>
<thead>
<tr>
<th>Allele frequency (q)</th>
<th>GRR</th>
<th>Dominant allele</th>
<th>Multiplicative</th>
<th>Additive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PAF</td>
<td>FRR</td>
<td>PAF</td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2.00</td>
<td>2.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3.00</td>
<td>3.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>5.00</td>
<td>5.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>10.00</td>
<td>10.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### Table II. PAF(%) and FRR according to the number of interacting genes

<table>
<thead>
<tr>
<th>Number of genes</th>
<th>Gene–gene interaction model</th>
<th>Additive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multiplicative</td>
<td>Additive</td>
</tr>
<tr>
<td>654</td>
<td>PAF</td>
<td>FRR</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>0.1</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>0.3</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>0.5</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### Table I.

The assumed parameters were: allele frequency, \( q = 0.1 \), GRR = 2.
predicted power for an association study with 500 cases and 500 controls was 93%, in perfect agreement with the actual data. It should be pointed out that this locus was not detected in 1233 families collected by an international consortium (25). The reasons for the discrepancies in the prostate cancer linkage results are not known, but one of them may be the genetic homogeneity of the Icelandic population.

Our data show that, in concert with the CDCV paradigm, common risk alleles confer high PAFs but modest FRRs. For example, it has been speculated that the five genes HRA51, NAT2, GSTT1, TNF-alpha and MTHFR would explain 54–64% of the cases of colorectal cancer, assuming that the GRRs were replicated (11). According to a large association study, the two genes NAT2 and GSTM1 would explain 31% of cases of bladder cancer (26). All these variants showed small GRRs (≤ 2.0) and thus, they would only marginally contribute to the familial risk. This is a dilemma of the CDCV paradigm: candidate gene studies may be able to identify genes conferring close to 100% of the PAF, but they may not explain the empirical FRRs. In order to explain FRRs, rare, high-penetrant genes need to be uncovered. However, the candidate gene studies for common alleles do not target this class of genes. Even the presently known susceptibility genes have a limited effect on FRRs (in breast cancer: BRCA1 1.08, Figure 4) because the mutant alleles are rare. The true test for the CDCV hypothesis will come in diseases of high FRR, such as multiple sclerosis, for which no high penetrant genes are known to date. The FRR for multiple sclerosis is about 8.0, which is almost impossible to explain by any numbers of low-risk genes, assuming that genes and no environmental sharing through e.g. infective agents explain the risk (27,28). The interacting common variants are expected to be rare and the associated GRRs could be high, thus bringing gene–gene interactions into the realm of rare Mendelian type genes. There are numerous possibilities for such interactions, and we tested some basic models in Table II. Multiplicative gene–gene interactions increase FRRs but they also tend to increase PAF. Little empirical data are available on gene–gene interactions, simply because the sample size requirements for even binary interactions become challenging, and they would vastly exceed those used in the present exercise (5,29,30). Low GRRs pose problems, even if they are formally significant, because the carriers are not enriched among familial cases and, thus, they cannot be properly replicated on familial cases (31,32) and because it will be difficult to provide experimental support in their favor. Furthermore, their significance for individual clinical genetic counseling remains doubtful because of low risk. Testing of rarer alleles may help to find genes of higher risk but larger sample sizes are required. Another option is to use familial cases, which may help to find genes of higher risk but larger sample sizes are required.

The present concepts have many useful applications in practical gene identification work. Before launching on a study, it would be useful to check the empirical FRRs of the particular disease. To explain a high familial risk, such as 8.0 for multiple sclerosis, rare high-penetrant genes need to be found and linkage studies may be the suitable approach. For a low familial risk, such as 1.5 for colorectal cancer, association studies appear justified but sample sizes should also allow for inclusion of rare alleles. After the study, PAFs may be used to assess the relevance of the findings in terms of disease etiology. The PAF reflects the overall effect of the gene under study by taking into account possible interactions with unknown genes. The a posteriori calculation of the FRRs is also advantageous; there are examples from the early genotyping studies reporting such high GRRs for common alleles that the resulting FRRs would exceed the empirical FRRs (33).

Acknowledgements

The Family-Cancer Database was created by linking registries maintained by Statistics Sweden and the Swedish Cancer Register, and supported by the Deutsche Krebshilfe, the Swedish Cancer Society, the Swedish Council for Working Life and Social Research, the EU, LSHC-RT-2004-503465 and EU Food-CT-2005-016320.

Conflict of Interest Statement: None declared.

References


Received July 26, 2006; revised September 13, 2006; accepted September 15, 2006