Using Identity-by-descent Information in Affected Sib Pairs to Increase the Efficiency of Genetic Association Studies

Jacqueline Wicks, Susan A. Treloar, and Nicholas G. Martin
Queensland Institute of Medical Research, Brisbane, Australia

The aim of this study was to determine whether identity-by-descent (IBD) information for affected sib pairs (ASPs) can be used to select a sample of cases for a genetic case-control study which will provide more power for detecting association with loci in a known linkage region. By modeling the expected frequency of the disease allele in ASPs showing IBD sharing of 0, 1, or 2 alleles, and considering additive, recessive, and dominant disease models, we show that cases selected from IBD 2 families are best for this purpose, followed by those selected from IBD 1 families; least useful are cases selected from IBD 0 families.

When a linkage study for a disease has been conducted and regions showing evidence of linkage have been identified, researchers will seek to use association methods to fine map disease genes. One means of doing this is through a case-control study, which might use individual genotyping or pooled DNA from cases and controls. Affected individuals from the initial linkage study can be used to provide an independent sample of cases if one affected individual per family is used. In the case of ASPs, if one sib is selected from each ASP, this will provide an independent sample.

For ASPs used in a linkage study, IBD information will be available in the location of the variant being studied for association. IBD status might provide a means of selecting a sample of affected individuals whose allele frequency distribution at a susceptibility locus *provides a greater contrast* with that of a random sample of controls. The sample of controls might be either a population-based random sample or a random sample of unaffected individuals from the population. Such a selection criterion for the cases, based on the IBD information in ASPs, would provide a sample of cases with a higher frequency of disease variants and thus result in increased power to detect association.

Methods

The distribution of disease variants among a sample of cases in which one is selected from each ASP can be modeled for ASPs that are IBD 0, 1 or 2 in the region of the variant studied for association. We model the distribution of genotypes in ASPs assuming a bi-allelic susceptibility locus and IBD status of either 0, 1 or 2. Then the expected allele frequencies among samples from each of the three groups are calculated, where in each case, one affected sib is chosen at random from each ASP. Then we consider the frequency of the disease allele for IBD 0, 1 and 2 ASPs under additive, recessive and dominant disease models, and graph the frequency under varying values of the population frequency of the disease variant.

Mathematical Model

Let the two alleles at the locus being studied for association be denoted A_1 and A_2. We assume that the allele A_1 is the susceptibility allele. The model is specified by the population frequency of the alleles, which we respectively denote by p_1 and p_2 (where $p_1 + p_2 = 1$), and by the penetrances for each of the three possible genotypes $A_1\ A_1$, $A_1\ A_2$, and $A_2\ A_2$, which we respectively denote by f_{11}, f_{12}, and f_{22}.

The frequency of the alleles at the susceptibility locus for a sample of cases extracted from either IBD 0, 1, or 2 ASPs can be calculated using the above parameters under the assumptions of Hardy-Weinberg proportions and random mating in the population at large.

The first stage involves calculating the frequencies for the possible genotype combinations for two children, conditional on IBD status and on their both being affected. The genotype combinations are:

$A_1\ A_1$ & $A_1\ A_1$
$A_1\ A_1$ & $A_1\ A_2$
$A_1\ A_1$ & $A_2\ A_2$
$A_1\ A_2$ & $A_1\ A_2$
$A_1\ A_2$ & $A_2\ A_2$
$A_2\ A_2$ & $A_2\ A_2$

Received 31 July, 2003; accepted 5 January 2004.

Address for correspondence: Jacqueline Wicks, Queensland Institute of Medical Research, Brisbane Qld 4029, Australia. Email: jackiW@qimr.edu.au

The probabilities for these are calculated in the Appendix. From these the expected proportion of A_1 alleles in a sample of alleles comprising the genotype of one sib chosen at random from each ASP of a given IBD status is given by

$$\begin{aligned}
&\Pr(A_1\,A_1\ \&\ A_1\,A_1 \mid \text{ASP \& IBD status}) \times 1 \\
+\ &\Pr(A_1\,A_1\ \&\ A_1\,A_2 \mid \text{ASP \& IBD status}) \times 0.75 \\
+\ &\Pr(A_1\,A_1\ \&\ A_2\,A_2 \mid \text{ASP \& IBD status}) \times 0.5 \\
+\ &\Pr(A_1\,A_2\ \&\ A_1\,A_2 \mid \text{ASP \& IBD status}) \times 0.5 \\
+\ &\Pr(A_1\,A_2\ \&\ A_2\,A_2 \mid \text{ASP \& IBD status}) \times 0.25 \\
+\ &\Pr(A_2\,A_2\ \&\ A_2\,A_2 \mid \text{ASP \& IBD status}) \times 0.
\end{aligned}$$

Explicit formulas for the probabilities in terms of disease allele frequencies and penetrances are derived in the Appendix. Using these formulas, it is possible to calculate the expected allele frequencies for samples taken from IBD 0, 1 and 2 families under different disease models, and thus get an idea of the differences in allele frequencies between these groups that we might expect to find in practice.

In Table 1 we give the forms for three genotype risk models: additive, recessive and dominant. These are parameterised in terms of the penetrance of the least susceptible genotype $A_2\,A_2$, which we denote by f, and α, which is the risk to carriers of two susceptibility alleles relative to carriers of no susceptibility alleles. We note that in the formulas for the above probabilities, α is the only relevant genotype risk parameter because f cancels out. The other relevant parameter is the frequency in the population at large of the susceptibility allele A_1, which is p_1.

Results

Using the parameters α and p_1, we can plot the expected frequency of A_1 alleles in samples of cases taken, one per family, from IBD 0, 1 and 2 families. This makes it possible to directly assess the usefulness of the different IBD families in providing cases for a case-control study, since greater deviation from the population frequency of the allele (or the frequency in a sample of unaffected individuals) would mean that the sample would give greater power to detect association.

It is also important to consider the frequency of the susceptibility allele that we would expect to see in a population-based independent sample of cases, so that this can be compared to the frequencies in the ASP-derived samples. However, it is interesting to note that under a model which assumes Hardy-Weinberg proportions and random mating in the population at large, the allele frequency in a random population-based sample of cases is actually identical to that found in a sample taken from IBD 0 ASP families.

Plots for the frequency of the susceptibility allele for different population frequencies of this allele (i.e., frequencies in population-based controls) are given in Figure 1 for additive, recessive and dominant disease models. These plots use the value $\alpha = 2$, which means that carriers of two disease alleles are assumed to be twice as likely as carriers of no disease alleles to be affected by the disease. Thus this is an example of a disease locus of very modest effect. In each figure, the highest solid line represents the frequency when the cases are selected from IBD 2 families; the middle solid line gives the same for IBD 1 families and the lowest solid line gives the same for IBD 0 families. The dashed diagonal line represents the expected frequency in population-based controls.

Under all disease models, cases from IBD 2 families are the most useful because they have the highest expected proportion of disease alleles. Cases from IBD 1 families are the next best and cases from IBD 0 families are the least useful. Cases from IBD 0 families have the same expected frequency of the disease allele as random cases from the population at large. The increases in the expected frequency of the disease allele, as can be seen from Figure 1, are substantial. For example, if the population frequency of the disease allele is 0.1, then under the additive disease model the frequency when cases are selected from IBD 0, 1, and 2 families are 0.141, 0.168, and 0.193 respectively. Under a recessive disease model these values are 0.109, 0.114, and 0.126 respectively and under a dominant disease model the values are 0.168, 0.214, and 0.255 respectively.

For larger values of α, the effect is more dramatic. For example under an additive disease model with $\alpha = 3$ the frequency of the disease allele when IBD 0, 1, and 2 ASPs are used to provide the cases are 0.175, 0.229, and 0.278 respectively. Under a recessive disease model the values are 0.118, 0.129, and 0.167 respectively, and under a dominant model are 0.217, 0.229, and 0.357 respectively.

As can be seen for $\alpha = 2$ in Figure 1, the increase in disease allele frequency under a recessive model is smaller for alleles of low frequency and higher for alleles of high frequency. This is reversed under a dominant model. Under an additive model the increases are more constant across the range of values for the disease allele frequency in the population.

Table 1

Penetrance Models

Disease Model	Penetrances		
	f_{22}	f_{12}	f_{11}
Additive	f	$\frac{\alpha+1}{2} f$	αf
Recessive	f	f	αf
Dominant	f	αf	αf

Note: Under these additive, recessive and dominant disease models, the penetrance for the A_1A_1 and A_2A_2 genotypes do not vary (f and αf respectively). Only the penetrance of the heterozygote (f_{12}) changes with the disease model.

Discussion

A random sample of cases for a case-control study can be selected from families recruited for a linkage study by selecting one case per family. For ASP families we

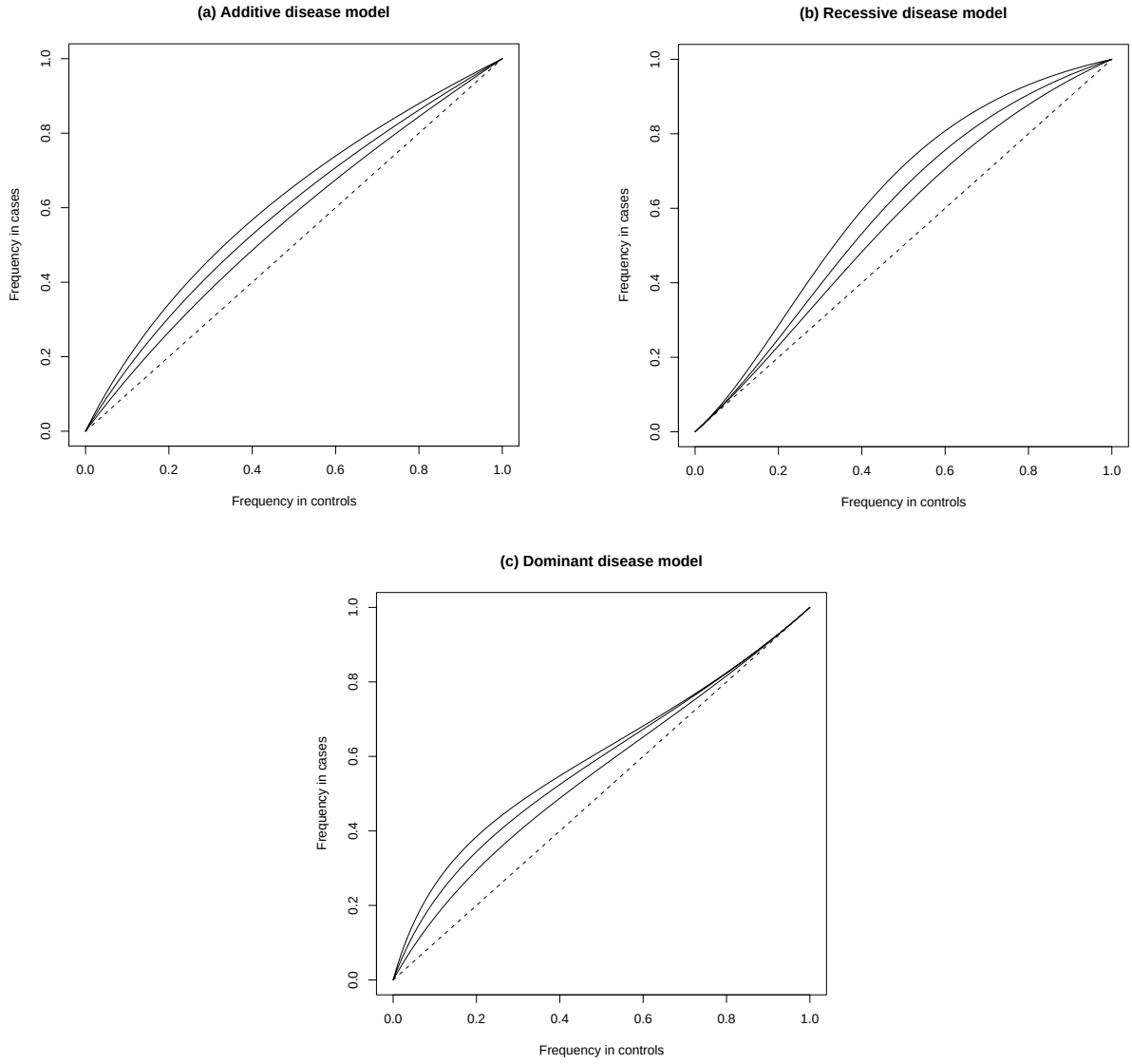

Figure 1

Frequency of disease allele in cases selected according to IBD status. In each of (a), (b) and (c), the curves show IBD 2 (top curve), IBD 1 (middle curve), and IBD 0 (bottom curve). The diagonal line indicates the frequency in unselected population-based controls. Graphs are shown for (a) additive, (b) recessive, and (c) dominant disease models with $\alpha = 2$ (i.e. genotype relative risk for the A_1A_1 genotype is twice that for the A_2A_2 genotype).

have explored the question of whether IBD information for ASPs can be used to select cases which will have a higher frequency of disease variants and thus provide more power in a case-control study. What we have shown is that under additive, recessive and dominant disease models, ASPs with increased IBD sharing provide better cases for this purpose, regardless of the population frequency of the disease variant. Thus, if a subset of cases is required for a case-control study, ASPs with higher IBD sharing values should be used first. Even cases from IBD 0 families have the same disease allele frequency as random population-based cases, so researchers should not hesitate to use these also if resources permit.

In practice, one issue that researchers might face is that there is often incomplete IBD information available for ASPs in a region; thus, such families cannot be classified as either IBD 0, 1 or 2 with certainty. A strategy to alleviate this uncertainty is to prioritize ASP families for inclusion according to their expected IBD sharing. Thus the families chosen first would be those with expected IBD sharing closest to 2; then families can be selected as far down the list as resources permit. If the whole sample of ASPs is to be employed, then no case selection strategy is required. If evidence for a disease-associated variant is found in such a sample, it would still be useful to consider its frequency in cases selected from the different IBD classes to see whether the predicted pattern of increased frequency with increased IBD sharing can be seen.

Appendix A

Here we calculate the probabilities of each of the possible genotype combinations for ASPs who share 0, 1 or 2 alleles IBD. From these the allele frequencies in the three IBD classes can be calculated.

The general form of probability we are interested in giving a model for in terms of genetic parameters is

$$\Pr(\text{geno1 \& geno2} \mid \text{ASP \& IBD status})$$

where geno1 and geno2 denote the genotypes of the two ASPs, and the IBD status is either 0, 1 or 2.

To derive the model, we note that the above probability is equal to

$$\Pr(\text{geno1 \& geno2} \mid \text{ASP \& IBD status}) \div \Pr(\text{ASP \& IBD status})$$

and use the decompositions

$$\Pr(\text{geno1 \& geno2 \& ASP} \mid \text{IBD status})$$

$$= \Pr(\text{geno1}) \Pr(\text{geno2} \mid \text{geno1 \& IBD status}) \times \Pr(\text{ASP} \mid \text{geno1 \& geno2})$$

and

$$\Pr(\text{ASP} \mid \text{IBD status})$$

$$= \Pr(\text{IBD status} \mid \text{ASP}) \times \Pr(\text{ASP}) \div \Pr(\text{IBD status})$$

Below we give the forms for the probabilities in these decompositions for IBD = 0, 1 and 2, in terms of the population allele frequencies and the penetrances, as geno1 and geno2 range over the possible genotype combinations. The derivation assumes Hardy-Weinberg proportions and random mating in the population at large.

IBD = 2

Putting geno1 = A_1A_1, we have $\Pr(\text{geno1}) = p_1^2$. Then for geno2 = A_1A_1, A_1A_2, and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1 \& IBD} = 2)$ is respectively given by 1, 0, and 0. For geno2 = A_1A_1, $\Pr(\text{ASP} \mid \text{geno1 \& geno2})$ is given by f_{11}^2.

For geno1 = A_1A_2, we have $\Pr(\text{geno1}) = 2p_1p_2$. Then for geno2 = A_1A_1, A_1A_2, and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1 \& IBD} = 2)$ is respectively given by 0, 1, and 0. For geno2 = A_1A_2, $\Pr(\text{ASP} \mid \text{geno1 \& geno2})$ is given by f_{12}^2.

For geno1 = A_2A_2, we have $\Pr(\text{geno1}) = p_2^2$. Then for geno2 = A_1A_1, A_1A_2, and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1 \& IBD} = 2)$ is respectively given by 0, 0, and 1. For geno2 = A_2A_2, $\Pr(\text{ASP} \mid \text{geno1 \& geno2})$ is given by f_{22}^2.

From the above we obtain

$$\Pr(A_1A_1 \,\&\, A_1A_1 \mid \text{ASP} \,\&\, \text{IBD} = 2) = p_1^2 f_{11}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 2)$$
$$\Pr(A_1A_1 \,\&\, A_1A_2 \mid \text{ASP} \,\&\, \text{IBD} = 2) = 0$$
$$\Pr(A_1A_1 \,\&\, A_2A_2 \mid \text{ASP} \,\&\, \text{IBD} = 2) = 0$$
$$\Pr(A_1A_2 \,\&\, A_1A_2 \mid \text{ASP} \,\&\, \text{IBD} = 2) = 2p_1p_2 f_{12}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 2)$$
$$\Pr(A_1A_2 \,\&\, A_2A_2 \mid \text{ASP} \,\&\, \text{IBD} = 2) = 0$$
$$\Pr(A_2A_2 \,\&\, A_2A_2 \mid \text{ASP} \,\&\, \text{IBD} = 2) = p_2^2 f_{22}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 2)$$

where $\Pr(\text{ASP} \mid \text{IBD} = 2)$ is given by $(p_1^2 f_{11}^2 + 2p_1p_2 f_{12}^2 + p_2^2 f_{22}^2) \times \Pr(\text{ASP}) \div \frac{1}{4}$.

IBD = 1

Putting geno1 = A_1A_1, we have $\Pr(\text{geno1}) = p_1^2$. Then for geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1} \,\&\, \text{IBD} = 1)$ is respectively given by p_1, p_2 and 0. For geno2 = A_1A_1 and A_1A_2, $\Pr(\text{ASP} \mid \text{geno1} \,\&\, \text{geno2})$ is respectively given by f_{11}^2 and $f_{11}f_{12}$.

For geno1 = A_1A_2, we have $\Pr(\text{geno1}) = 2p_1p_2$. Then for geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1} \,\&\, \text{IBD} = 1)$ is respectively given by $\frac{1}{2}p_1$, $\frac{1}{2}$ and $\frac{1}{2}p_2$. For geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{ASP} \mid \text{geno1} \,\&\, \text{geno2})$ is respectively given by $f_{11}f_{12}$, f_{12}^2 and $f_{11}f_{12}$.

For geno1 = A_2A_2, we have $\Pr(\text{geno1}) = p_2^2$. Then for geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1} \,\&\, \text{IBD} = 1)$ is respectively given by 0, p_1 and p_2. For geno2 = A_1A_2 and A_2A_2, $\Pr(\text{ASP} \mid \text{geno1} \,\&\, \text{geno2})$ is respectively given by $f_{12}f_{22}$ and f_{22}^2.

The above results give

$$\Pr(A_1A_1 \,\&\, A_1A_1 \mid \text{ASP} \,\&\, \text{IBD} = 1) = p_1^3 f_{11}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 1)$$
$$\Pr(A_1A_1 \,\&\, A_1A_2 \mid \text{ASP} \,\&\, \text{IBD} = 1) = 2p_1^2 p_2 f_{11}f_{12} \div \Pr(\text{ASP} \mid \text{IBD} = 1)$$
$$\Pr(A_1A_1 \,\&\, A_2A_2 \mid \text{ASP} \,\&\, \text{IBD} = 1) = 0$$
$$\Pr(A_1A_2 \,\&\, A_1A_2 \mid \text{ASP} \,\&\, \text{IBD} = 1) = p_1p_2 f_{12}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 1)$$
$$\Pr(A_1A_2 \,\&\, A_2A_2 \mid \text{ASP} \,\&\, \text{IBD} = 1) = 2p_1p_2^2 f_{12}f_{22} \div \Pr(\text{ASP} \mid \text{IBD} = 1)$$
$$\Pr(A_2A_2 \,\&\, A_2A_2 \mid \text{ASP} \,\&\, \text{IBD} = 1) = p_2^3 f_{22}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 1)$$

where $\Pr(\text{ASP} \mid \text{IBD} = 1)$ is given by

$$\{p_1^2 f_{11}(p_1 f_{11} + p_2 f_{12}) + p_1 p_2 f_{12}(p_1 f_{11} + f_{12} + p_2 f_{22}) + p_2^2 f_{22}(p_1 f_{12} + p_2 f_{22})\} \times \Pr(\text{ASP}) \div \tfrac{1}{2}.$$

IBD = 0

Putting geno1 = A_1A_1, we have $\Pr(\text{geno1}) = p_1^2$. Then for geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1 \& IBD} = 0)$ is respectively given by p_1^2, $2p_1p_2$ and p_2^2. For geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{ASP} \mid \text{geno1 \& geno2})$ is respectively given by f_{11}^2, $f_{11}f_{12}$ and $f_{11}f_{22}$.

For geno1 = A_1A_2, $\Pr(\text{geno1}) = 2p_1p_2$. Then for geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1 \& IBD} = 0)$ is respectively given by p_1^2, $2p_1p_2$ and p_2^2. For geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{ASP} \mid \text{geno1 \& geno2})$ is respectively given by $f_{11}f_{12}$, f_{12}^2 and $f_{12}f_{22}$.

For geno1 = A_2A_2, $\Pr(\text{geno1}) = p_2^2$. Then for geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{geno2} \mid \text{geno1 \& IBD} = 0)$ is respectively given by p_1^2, $2p_1p_2$ and p_2^2. For geno2 = A_1A_1, A_1A_2 and A_2A_2, $\Pr(\text{ASP} \mid \text{geno1 \& geno2})$ is respectively given by $f_{11}f_{22}$, $f_{12}f_{22}$ and f_{22}^2.

These results give

$$\begin{aligned}
\Pr(A_1A_1 \,\&\, A_1A_1 \mid \text{ASP \& IBD} = 0) &= p_1^4 f_{11}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 0)\\
\Pr(A_1A_1 \,\&\, A_1A_2 \mid \text{ASP \& IBD} = 0) &= 4p_1^3 p_2 f_{11} f_{12} \div \Pr(\text{ASP} \mid \text{IBD} = 0)\\
\Pr(A_1A_1 \,\&\, A_2A_2 \mid \text{ASP \& IBD} = 0) &= 2p_1^2 p_2^2 f_{11} f_{22} \div \Pr(\text{ASP} \mid \text{IBD} = 0)\\
\Pr(A_1A_2 \,\&\, A_1A_2 \mid \text{ASP \& IBD} = 0) &= 4p_1^2 p_2^2 f_{12}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 0)\\
\Pr(A_1A_2 \,\&\, A_2A_2 \mid \text{ASP \& IBD} = 0) &= 4p_1 p_2^3 f_{12} f_{22} \div \Pr(\text{ASP} \mid \text{IBD} = 0)\\
\Pr(A_2A_2 \,\&\, A_2A_2 \mid \text{ASP \& IBD} = 0) &= p_2^4 f_{22}^2 \div \Pr(\text{ASP} \mid \text{IBD} = 0)
\end{aligned}$$

where $\Pr(\text{ASP} \mid \text{IBD} = 0)$ is given by $(p_1^2 f_{11} + 2p_1p_2 f_{12} + p_2^2 f_{22})^2 \times \Pr(\text{ASP}) \div \tfrac{1}{4}$.