Polygenic Inherited Predisposition to Breast Cancer

B.A.J. PONDER, A. ANTONIOU, A. DUNNING, D.F. EASTON, AND P.D.P. PHAROAH
Departments of Oncology & Public Health & Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratories, University of Cambridge, Cambridge CB1 8RN, United Kingdom

Our aim is to identify the genes that contribute to susceptibility to breast cancer. There are two potential applications. The first is to cancer biology. The demonstration that a particular gene contributes to susceptibility provides genetic evidence to confirm the importance of genes or pathways whose role in cancer development is already suspected, or it can highlight new genes and new pathways. The second is to defining individual risks. The polygenic model for susceptibility outlined below suggests a wide range of risk within the population. Knowledge of individual genetic makeup may allow prediction of risk and thus inform choices about screening and prevention.

Two predisposing genes are already well known: BRCA1 and BRCA2. Rare, strongly predisposing mutations in these genes give rise to multiple-case families, allowing the identification of the genes by linkage and positional cloning. Subsequently, knowledge of the genes has indeed led to new insights into cancer biology (see, e.g., Venkitaraman 2002), culminating recently in proposals for new therapeutic targets (Farmer et al. 2005), and to gene-based estimates of risk that guide clinical management (Antoniou et al. 2003).

Although important, mutations in BRCA1 and 2 account for only a small part of inherited susceptibility to breast cancer. The size of the total inherited contribution can be inferred from the extent of familial clustering. On average, the risk of breast cancer to the mother or sister of a case is about 1.8-fold higher than the risk across the population. Twin studies suggest that most of this family clustering is due to inheritance rather than shared environment (Lichtenstein et al. 2000). The contribution of BRCA1 and 2 can be estimated by taking a large population-based series of breast cancer cases with details of family history—and thus of the familial excess of breast cancers—and measuring the fraction of the familial excess that is accounted for by families in which there is a BRCA1 or 2 mutation. Such studies indicate a contribution of 15–20% (Easton 1999; Peto et al. 1999; Anglian Breast Study Group 2000). Mutations in other genes, in addition to defining individual risks. The polygenic model (additive, multiplicative) by which their effects combine (Fig. 1).

Two points arise from this figure: (1) The strategy to search for the genes depends on the model. Linkage is appropriate for BRCA-like genes, but not for polygenes (Risch and Merikangas 1996; Risch 2000); association studies are appropriate for common polygenes but not for rare strong genes. (2) If the polygenic model is chosen, the “genetic architecture” will also be critical. One can hope to detect common alleles of moderate effect, but weak, and especially rare, alleles will be difficult. This is discussed further below.

THE POLYGENIC MODEL

The evidence for this model in breast cancer comes from population-based epidemiological studies. Most of the excess familial clustering of breast cancer is found to be distributed across many families, each with small numbers of cases, rather than in a few very extensive fami-
ilies (see, e.g., Cui et al. 2001; Antoniou et al. 2002). This is the pattern expected with a substantial contribution from a polygenic mode of predisposition, where the excess risk decays rapidly with increasing distance of relationship. A Mendelian BRCA-like model of a single mutant gene with strong effects, in contrast, would result in concentration of the excess cases in a few multiple-case families. We have therefore based our gene finding strategy on the polygenic model.

The polygenic model has important implications for the distribution of risk in the population (Pharoah et al. 2002). Risk is normally distributed, and the width of the distribution—which determines the magnitude of the difference in risk between the high- and low-risk groups—is related to the size of the overall genetic effect. For breast cancer, our model derived from a population-based series of cases from the East Anglian region of the UK suggests a 30- to 40-fold difference between women in the highest and lowest 20% of the population. This translates, for example, into the 12% of women most at risk accounting for half of all breast cancers, while the 50% at lowest risk account for only 12%. The public health implications, if individuals could be placed on the spectrum, are clear.

SEARCHING FOR THE GENES UNDER A POLYGENIC MODEL

Provided certain conditions (outlined below) are met, association studies are the most efficient design. In an association study, DNA sequence variants (generally in the form of single-nucleotide polymorphisms [SNPs]) in a series of candidate genes are compared for frequency in a set of cases and of controls, which should be as closely matched for genetic background as possible (Cardon and Bell 2001). The straightforward case is where the variant to be tested is functionally active and thought to be directly causally relevant to the disease. In this case, a consistent finding that the variant occurs at a higher frequency in cases than in controls is presumptive evidence of cause. Usually, however, no clear candidate causative variant is known. In this case, the common variants in the gene (those with minor allele frequency greater than 5–10%) are identified as completely as possible by sequencing, and a subset is chosen, on the basis of allele frequency and linkage disequilibrium relationships, that will most efficiently report on the others. These are “tagging SNPs” (for more detailed discussion, see Chapman et al. 2003; Carlson et al. 2004; Wang et al. 2005). The intention is that one or more of the tagging SNPs, or a haplotype defined by them, will act as a surrogate for the putative but unknown disease allele.

For this strategy to succeed, even if the right candidate genes have been selected, several conditions must be fulfilled. These relate to the genetic architecture of disease predisposition, and they lie at the heart of the problem of finding these genes. First, the effect of any allele must be large enough for the association to be detected within the power of the study. It is usual to express the size of effect in terms of the relative risk of disease conferred by the allele, but of course, the allele frequency is an important determinant of its effect as well. We therefore prefer to use as our measure the percentage of the excess familial risk (the genetic variance) that is explained. Because this combines the effects of allele frequency and relative risk, power is fairly stable across a range of allele frequencies (Fig. 2). Note that even in the most easily detected plausible (dominant, or additive [co-dominant]) models, sample sizes of several thousands are needed to obtain significance values that provide some confidence of true association for common alleles that account for 1% of the genetic variance (Antoniou and Easton 2003; Pharoah et al. 2004). Given the practicalities of assembling and testing very large case/control sets, this indicates the constraints on the size of effect that can be searched for.

Second, for good power in such a study, the disease alleles must also be common, with minor allele frequencies ideally of at least 5–10%, and comparable to those of the “tagging” variants used to report them. Statistical power
is lost if the tagging variants are at a markedly different frequency from the alleles they are to tag. Current association studies are therefore based on common tagging variants, because the search is predicated on the existence of similarly common disease alleles, and because the search for rare alleles has very low statistical power.

Third, the tagging variant and the disease allele should have arisen only once in the history of the population. Variants that have arisen more than once probably have occurred on different chromosomes, which will destroy the co-inheritance of the tag and disease alleles (see Fig. 3).

In summary, the association study strategy is based on the “common variant common disease” hypothesis, which states that a substantial component of predisposition is attributable to common variants which are likely to be ancestral in the population, rather than to multiple individually rare variants recently arisen from new mutation. The likely validity of this hypothesis, and thus of the association-study-based search for common disease genes, has been the subject of a great deal of argument (Reich and Lander 2001; Wright and Hastie 2001; Pritchard and Cox 2002; Smith and Lusis 2002; Botstein and Risch 2003; Lohmueller et al. 2003), none of it conclusive. Probably, some disease alleles are common and ancient, and others are rare and new. A reasonable way to find out the situation in breast cancer would seem to be: Do the experiment.

RESULTS

We have based our studies of breast cancer over the past 5 years on a variety of candidate genes and pathways, including steroid hormone metabolism and signaling, DNA repair, cell cycle and checkpoints, growth factors, carcinogen metabolism, tumor microenvironment, methylation, and candidate genes from mouse and rat models. To date, we have examined 525 SNPs in 110 genes, tagged with increasing efficiency as both experimental design and information on SNP variants have improved. The case/control set we have studied comprises 4,600 cases of invasive breast cancer diagnosed below age 70 ascertained through the Anglian Cancer Registry from the East Anglian region of the UK, and 4,600 controls from the EPIC cohort study based in the same region (Day et al. 1999). Simple epidemiological and clinical data are available for all cases, along with paraffin blocks of tumors from a large subset; the EPIC controls have extensive epidemiological data, frozen serum, and a range of phenotypes including mammographic data and levels of serum hormones and growth factors.

Our strategy has been to aim for 90% power at $p = 10^{-4}$ to detect common variants that account for 1% of the genetic variance. For economy, we have used a 2-stage genotyping strategy in which all SNPs are tested in a first set of 2,300 cases and controls, and those that meet specified statistical criteria ($p < 0.1$ in a test of genotype distribution or a test for trend) are tested in a second similar-sized case/control set drawn from the same sampling frame.

To date, we have found no SNP in any gene that gives us $p < 10^{-4}$ for association with breast cancer, or which accounts for 1% of the genetic variance. Of the first 500 SNPs analyzed, 36 gave a $p$ value for association with breast cancer at $p < 0.05$, and 9 at $p < 0.01$, using the 2 df test for genotype distribution, compared to 25 and 5 expected by chance. Similar, but slightly different, results were obtained using a test for trend.
No individual result at these levels of significance can be relied upon as a true positive rather than a chance result. Given the prior probability that any given variant might contribute to susceptibility, and setting aside the argument that particular genes are, a priori, more likely candidates, a $p$ value of 0.001 in our studies has something like a 50:50 probability of representing a “true” result (Pharoah et al. 2004). We have therefore asked whether, taking the results in aggregate, we can reject the possibility that there is no effect—that is, that none of our tested candidates contributes to breast cancer risk. To address this, we have ranked the $p$ values we have obtained, calculated a cumulative $p$ value as each successive result is added, and used permutation testing (because of the incomplete independence of the individual SNPs) to compare these values with those expected under the hypothesis of no genetic effect. In addition to the global analysis, we have modeled different combinations of numbers, frequency, and relative risks of susceptibility alleles, and asked what power we would have had to detect these effects in the global permutation test. The details are not presented here: In summary, even with our case/control set, which is large in comparison to those that are usually reported, our power proves to be quite poor to detect the presence of genetic effects involving a few alleles that each contribute around 0.5–1% of the total variance, under a variety of plausible scenarios. We therefore remain uncertain whether, and to what extent, our results provide evidence for the presence of susceptibility alleles in our population. We are investigating further statistical methods to address this question.

**HOW TO EVALUATE THE RESULTS?**

The above analysis leaves open the possibility that, among the candidate genes we have tested, there may be some in which common variants do contribute to susceptibility, but that these effects are too weak for us to have detected them. Support for this possibility comes from the clearly positive results with serum hormones and growth factors as “intermediate phenotypes,” described below. Probably none of these genes individually contributes as much as 1% of the genetic variance, and in aggregate, they probably account for less than 3%.

We may have missed alleles of larger effect because of shortcomings in our analysis and study design. Thus, we have not incorporated the possibility of gene interactions into the analyses (Marchini et al. 2005) but have concentrated on main effects, thereby possibly missing larger effects in a subset of individuals. Similarly, we have used a single dichotomous phenotype—breast cancer—whereas breast cancer is heterogeneous in molecular terms and probably in genetic etiology. Quantitative phenotypes, such as serum hormone levels or mammographic density, have greater statistical power. This is illustrated by our results with variants in the CYP19 gene, which is involved in the conversion of testosterone to estradiol, and their effects on serum levels of estradiol and of testosterone (Dunning et al. 2004). We found a highly significant association of the variants in CYP19 with hormone levels, but epidemiological data indicate that these effects on hormone levels would be predicted to correspond to a very small effect on breast cancer risk, consistent with our failure to detect an association of the same CYP19 variants with breast cancer (Fig. 4A,B).

The main argument that there may be undiscovered genes of larger effect is that we have tested as candidates only about 100 genes of the 30,000 that are possible; moreover, we have tested those incompletely, especially with respect to putative regulatory variants. Theoretical arguments, as well as empirical data from man and other organisms (for review, see Mackay 2001), suggest that the distribution of sizes of effect for alleles that modify phenotypes or influence susceptibility will be roughly exponential—that is, there will be some alleles of moderate effect, and a “tail” of many alleles of smaller effect. Some alleles of moderate effect have indeed been identified to contribute to susceptibility to other diseases (Lohmueller et al. 2003), in a model system such as Drosophila, where a more complete analysis is possible, a range of effect sizes is seen (Long et al. 2000). To search efficiently for the few alleles with larger effects that may exist, we need to move from a candidate gene approach to an empirical genome-wide scan for association.

**A GENOME-WIDE SCAN**

The aim is to use a set of SNPs that will report as completely as possible on common disease alleles across the entire genome. Some have argued that for reasons of cost and practicality, a genome scan should be targeted at locations where a more complete analysis is possible, a range of effect sizes is seen (Long et al. 2000). To search efficiently for the few alleles with larger effects that may exist, we need to move from a candidate gene approach to an empirical genome-wide scan for association.

---

**Figure 4.** Intermediate phenotypes and prediction of breast cancer risk. Panel A shows levels of serum estradiol (in pmole/liter) in control women, corresponding to genotypes for the 1-c polymorphism in the 3′ UTR of CYP19. There is a significant correlation of hormone level with genotype ($p$ for homogeneity = 0.0086). Panel B shows (a) the predicted effect, from the data in panel A, of CYP19 genotype on estradiol levels, (b) the predicted effect, from published studies, of a twofold increase in estradiol level on breast cancer risk, and (c) derived from these, the predicted effect of CYP19 genotype on breast cancer risk. Risks of this magnitude were used in our analysis of 50 cases and 34,000 controls to demonstrate with 50% power at $p = 10^{-4}$ the association between CYP19 genotype and breast cancer risk. (Data from Dunning et al. 2004.)
common nonsynonymous variants in known coding regions, since these are the most likely to be relevant. However, given the emerging data on interindividual variation in levels of gene expression, the evidence that many of the phenotypic differences between inbred strains of mice are regulatory in origin, and the data both on structural variations in the genome (Shabt et al. 2004) and on the existence of highly conserved eukaryotic regions with possible regulatory functions (Merlo et al. 2005), we have thought it better to adopt a completely unbiased approach and to aim for as wide coverage of the genome as possible, with sufficient power to test the hypothesis that there are at least some alleles of moderate effect.

Over 6 million SNPs have been described across the genome, but it is not, of course, necessary to test them all. Many will be tightly associated with one or several others, which will then be mutually redundant. The first task is to assemble the set of tagging SNPs that is the best compromise between power and cost in reporting on the remainder of the common variation in the genome. The power of one SNP to report on another is expressed as the \( r^2 \) value; an \( r^2 \) of 1 means perfect power, whereas an \( r^2 \) of 0.5 means that the sample size to be tested must be doubled to obtain power equivalent to a SNP with \( r^2 \) of 1. The distribution of maximum \( r^2 \) values with which a set of tagging SNPs reports on all other known common variants provides an expression of the power of the proposed scan with these SNPs (Carlson et al. 2004).

Perlegen Sciences have identified over 2 million SNPs and typed them in single human chromosomes isolated in somatic cell hybrids, thus obtaining unequivocal haplotypes. From these, they defined in 2004 a subset of 266,000 SNPs that they showed empirically would report with \( r^2 \) 0.8 or greater on an estimated 73% of all common SNPs in 152 genes for which resequencing data were available (Hinds et al. 2005). The process of SNP discovery of this set is described in detail in Patil et al. (2001) and Zhang et al. (2002). A substantially larger set of SNPs would have provided some increase in coverage, but at correspondingly greater cost.

Using these data, we designed what we believed to be the most cost-effective experiment available at the time, which was to use the 266,000 SNP set to search for any breast cancer susceptibility allele that would account for 2% or more of the total genetic variance. We estimated the power to detect such variants to be 80%. This estimate allows for loss of power from incomplete linkage disequilibrium and allele frequency differences. We analyzed only invasive breast cancer and treated breast cancer as a single diagnosis, not attempting to define subsets to recognize heterogeneity.

To attain this power, we needed to analyze several thousand breast cancer cases and controls, but with 266,000 SNPs in a single experiment, this would amount to well over a billion genotypes, and therefore would be too expensive. Pooling of samples (for example, in sets of 80 to well over a billion genotypes, and therefore would be too expensive. Pooling of samples (for example, in sets of 80) makes effects of the size we thought might be important—relative risks of 1.3—difficult to detect reliably. We used instead a 2-stage strategy (Satagopan 2004) similar to that we had used for our previous candidate gene studies (Antoniou and Easton 2003). In the first stage, all 266,000 SNPs were tested against a small set of 400 breast cancer cases and controls, the cases genetically enriched by having a strong family history, but with no evidence of BRCA1 or 2 mutations, and of one statistical model in which all familial cases (Antoniou and Easton 2003) are correct, these 400 familial cases should be equivalent to about 1,600 unselected cases. The 400 controls taken from the EPIC cohort also provide substantial additional phenotypic data that can be used for genotypic analysis. The 12,000 or so SNPs with the most significant results from the first stage were then to be analyzed in a further, population-based, set of 4,600 cases and of 4,600 EPIC controls not overlapping with the first set. The best SNPs at this stage would be further analyzed as a third stage in a set of over 20,000 cases and controls currently being assembled through an International Consortium. These numbers should provide convincing statistical evidence to identify alleles associated with effects down to 1% of the overall genetic variance; but as for all association studies, the final identification of the functional allele, rather than simply the variant showing association with it, may require much more work.

The first-stage genotyping was completed in mid-summer 2005. The results provide encouragement that technical aspects of the experiment are satisfactory. 84% of the SNPs were scoreable in greater than 70% of individuals; for those SNPs, 95% of all the possible genotypes were obtained, with duplicate genotype reproducibility of 99.9%. There were no systematic differences in results obtained with DNA from each of the 4 U.K. centers that provided familial cases. Moreover, the distribution of \( p \) values for association that we observed was closely similar to that expected if there were no genetic effect, and with no highly positive outliers (the extreme outliers suggests that there are no gross artefacts within the system. The results of the second stage should be available early in 2006.

**RARE AND WEAK ALLELES**

If the genome scan identifies some alleles of moderate effect, further analysis of these and their interactions should enlarge our knowledge of the genetic architecture of the disease. There will still be the question of common, weak alleles and of rare alleles which the association study is not powerful to detect.

The search for common weak alleles requires increased power. This will come from some combination of the following: larger studies; improved phenotyping of breast cancer; probably through molecular profiling of tumors; more sophisticated analysis of gene interactions; and possibly the use of quantitative intermediate phenotypes. With the tumor profiling data, it may, incidentally, be possible to address the interesting question of the relationship between germ-line genetic variation and the somatic genetic profile of the tumor. To what extent is the behavior and natural history of a neoplasm determined or constrained by the germ-line genetic environment? If there is
2. The resolution of phenotypic heterogeneity. In cancer, the genetic architecture of susceptibility. How much is the genetic contribution of rare variants and rare disease alleles in the population unclear but will likely determine the most cost-effective approach. Questions include the numbers of rare sequence variants that will be found, the proportion of these that have pathological significance, and the confidence with which these can be recognized, the extent of allelic heterogeneity; and whether susceptibility alleles are concentrated in rather few genes (suggested by theoretical arguments and the example of the ABC A1 gene and plasma HDL levels; Cohen et al. 2004) or are widely dispersed. Until affordable genome-wide sequencing becomes available, a candidate gene approach might be based on rare variant discovery by resequencing in a few hundred “genetically enriched” individuals, followed by case/control comparisons using allele-specific assays. The choice of a quantifiable intermediate phenotype for which there are good candidate genes, such as some aspect of DNA repair function, combined with a search for rare variants across the candidate gene set, might provide a better-defined first step to investigate allele architecture than the study of a cancer phenotype for which the candidate genes are uncertain.

CONCLUSIONS

Despite the growing recognition of the importance of polygenic susceptibility in cancer and in other diseases such as diabetes and hypertension, the search for the predisposing genes has so far been less successful than was hoped (Lohmueller et al. 2003). This disappointment stems in large part from a failure to appreciate the difficulties. Critical determinants of success are likely to include the following:

1. The genetic architecture of susceptibility. How much is due to alleles of sufficiently large effect that we can hope to identify them in the human population?

2. The resolution of phenotypic heterogeneity. In cancer, there are many ways to slice this cake, and we don’t know which are best. If, as many believe, the cases to be included in association studies should be subclassified at the outset by molecular features of their tumor, the assembly of sufficiently large sets of cases with this information represents a formidable challenge.

3. The analysis of interactions. Although it can be argued that most predisposing alleles of sufficient impact to be of interest should be detectable by main effects, this is not always so—for example, in the case of interactions involving a rare allele. New statistical methods and, probably, better understanding of biology will also be needed to tackle the problems of higher-dimensional interactions.

4. The extent to which the search is focused—for cost-effectiveness—on non-synonymous coding variants to the exclusion of regulatory variants including structural variations in extragenic regions, and possible epigenetic effects. All possibilities are still to be considered.

Progress over the next few years is likely to come from four directions: larger, and better defined, sample collections; improved genotyping and sequencing technologies; increasing databases of human genetic variation; and new statistical methods in complex genetics. Of these, possibly the most critical are the sample collections. These are expensive to set up and maintain, and funders need to understand their importance. The recent investments in large population cohorts in several countries may ultimately—in 10 or 20 years and more—provide valuable resources for the study of complex genetics and gene-environment interaction, but should not be a reason to neglect funding for the targeted case/control collections that are needed now to initiate gene discovery. The size of case/control sets that is needed will require international consortia and an ethos of sharing and collaboration perhaps rather different from the competitive days of Mendelian gene discovery.

When the genes are found, the first applications will almost certainly be to cancer biology: indications of new pathways, or genetic evidence to support the relevance of those already implicated. It is possible that a single common allele might contribute a sufficient fraction of disease incidence as to be itself a target for prevention (Fig. 5). The construction of individual risk profiles from multiple genotypes is potentially an important goal, but in breast cancer at least, still a distant prospect.

Figure 5. Attributable risks of common predisposing alleles. The vertical axis shows the population-attributable fraction—that is, the fraction of disease incidence that would be avoided if the predisposing allele were neutralized or removed from the population—for different combinations of frequency and relative risk of a dominant susceptibility allele. Thus, an allele with frequency 20% and a modest relative risk of around 1.5 has an attributable fraction of 15–20%. Note, for comparison, that the strongly predisposing mutant alleles of BRCA1 and 2, with population frequencies 0.02%, have very small attributable fractions because of their rarity.
PREDISPOSITION TO BREAST CANCER

ACKNOWLEDGMENTS

Work in the author’s laboratory is funded by Cancer Research UK and E.A.P. is a Principal Fellow, and P.D.P.P. is a Senior Clinical Fellow of Cancer Research UK.

REFERENCES


Controversy over the Croesus code. Science 303:589.


Polygenic Inherited Predisposition to Breast Cancer


Cold Spring Harb Symp Quant Biol 2005 70: 35-41
Access the most recent version at doi:10.1101/sqb.2005.70.029

References
This article cites 38 articles, 7 of which can be accessed free at:
http://symposium.cshlp.org/content/70/35.full.html#ref-list-1

License

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Cold Spring Harbor Symposia on Quantitative Biology go to:
http://symposium.cshlp.org/subscriptions

Copyright 2005, Cold Spring Harbor Laboratory Press