Association between Common Variation in 120 Candidate Genes and Breast Cancer Risk

Paul D. P. Pharoah[1,2*], **Jonathan Tyrer**[1], **Alison M. Dunning**[1], **Douglas F. Easton**[2], **Bruce A. J. Ponder**[1], **SEARCH Investigators**

1 Department of Oncology, University of Cambridge, Cambridge, United Kingdom, **2** Strangeways Research Laboratory, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

Association studies in candidate genes have been widely used to search for common low penetrance susceptibility alleles, but few definite associations have been established. We have conducted association studies in breast cancer using an empirical single nucleotide polymorphism (SNP) tagging approach to capture common genetic variation in genes that are candidates for breast cancer based on their known function. We genotyped 710 SNPs in 120 candidate genes in up to 4,400 breast cancer cases and 4,400 controls using a staged design. Correction for population stratification was done using the genomic control method, on the basis of data from 280 genomic control SNPs. Evidence for association with each SNP was assessed using a Cochran–Armitage trend test (*p*-trend) and a two-degrees of freedom χ^2 test for heterogeneity (*p*-het). The most significant single SNP (*p*-trend $= 8 \times 10^{-5}$) was not significant at a nominal 5% level after adjusting for population stratification and multiple testing. To evaluate the overall evidence for an excess of positive associations over the proportion expected by chance, we applied two global tests: the admixture maximum likelihood (AML) test and the rank truncated product (RTP) test corrected for population stratification. The admixture maximum likelihood experiment-wise test for association was significant for both the heterogeneity test ($p = 0.0031$) and the trend test ($p = 0.017$), but no association was observed using the rank truncated product method for either the heterogeneity test or the trend test ($p = 0.12$ and $p = 0.24$, respectively). Genes in the cell-cycle control pathway and genes involved in steroid hormone metabolism and signalling were the main contributors to the association. These results suggest that a proportion of SNPs in these candidate genes are associated with breast cancer risk, but that the effects of individual SNPs is likely to be small. Large sample sizes from multicentre collaboration will be needed to identify associated SNPs with certainty.

Citation: Pharoah PDP, Tyrer J, Dunning AM, Easton DF, Ponder BAJ, et al. (2007) Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genet 3(3): e42. doi:10.1371/journal.pgen.0030042

Introduction

Breast cancer tends to cluster in families, with disease being approximately 2-fold more common in first-degree relatives of cases [1]. The higher rate of most cancers in the monozygotic twins of cases than in dizygotic twins or siblings suggests that most of the familial clustering is the result of inherited genetic variation rather than lifestyle or environmental factors [2]. Some of this clustering occurs as part of specific familial breast cancer syndromes where disease results from single alleles conferring a high risk. However, such alleles are rare in the population, and highly penetrant variants of *BRCA1* and *BRCA2* account for less than 20% of the genetic risk of breast cancer with other rarer high penetrance genes such as *TP53, ATM,* and *PTEN* counting for less than 5% [3]. Despite extensive efforts, linkage studies have failed to map further more *BRCA*-like highly penetrant cancer susceptibility genes [4]. Together with data on patterns of familial occurrence of cancer that exclude cases because of known high-risk genes, this argues strongly that most genetic susceptibility results from the combined effects of many genetic variants, each of which have a modest effect individually [5]. Family-based linkage studies have been the foundation for the many successes in mapping of genes associated with Mendelian disorders, but they lack power to detect alleles conferring moderate risks that are likely to be the norm in complex disease. The main alternative to linkage studies for disease gene mapping is the association study, in which the frequency of a genetic variant in diseased individuals (cases) and individuals without the disease (controls) are compared [6,7]. Association studies for disease genes are generally based on the "common variant: common disease" hypothesis [8]. Allelic association is present when the distribution of genotypes differs in cases and controls. Such an association provides evidence that the locus under study, or a neighbouring locus, is related to disease susceptibility.

Considerable research effort has been put into the search

Editor: Gonçalo Abecasis, University of Michigan, United States of America

Received September 22, 2006; **Accepted** February 2, 2007; **Published** March 16, 2007

A previous version of this article appeared as an Early Online Release on February 2, 2007 (doi:10.1371/journal.pgen.0030042.eor).

Abbreviations: AML, admixture maximum likelihood; MAF, minor allele frequency; RTP, rank truncated product; SNP, single nucleotide polymorphism

* To whom correspondence should be addressed. E-mail: paul1@srl.cam.ac.uk

Author Summary

The polygenic model of cancer susceptibility suggests that multiple alleles contribute to the excess familial risk of most common cancers. Candidate gene association studies have been a commonly used approach in the search for such alleles. We have investigated over 700 common variants in genes that are candidates for breast cancer susceptibility in a large case-control study of breast cancer, but no single variant was identified at an appropriate level of statistical significance. The purpose of this study was to consider these data as a whole, using a novel method, the admixture maximum likelihood test, to test the hypothesis that a proportion (unknown) of the variants we investigated are associated with breast cancer. After adjusting for population substructure, we found evidence for association that was robust to all but the most extreme assumptions about the degree of population stratification. Genes in the cell-cycle control and steroid hormone metabolism and signalling pathways were the main contributors. These results suggest that a proportion of single nucleotide polymorphisms (SNPs) in these candidate genes are associated with breast cancer risk, but that the effects of individual SNPs are likely to be small. Large sample sizes from multicentre collaboration will be needed to identify associated SNPs with certainty.

for low to moderate penetrance breast cancer susceptibility alleles over the past ten years. Most early association studies were based on testing candidate functional polymorphisms in candidate genes, but recently more emphasis has been placed on an empirical approach in which a minimal set of "tagging" single nucleotide polymorphisms (SNPs) that efficiently captures all the common genetic variation in a gene is assayed [9]. In addition, high throughput genotyping technologies have made it possible to assess multiple candidate genes. The analysis of such studies inevitably involves a large number of statistical tests, and there has been much debate about how to analyse the totality of such data, and, in particular, how (or indeed whether) to carry out a correction for multiple hypothesis testing. Most approaches to this problem have considered this as a hypothesis-testing problem, in which the aim is to control the overall "experiment-wise" type I error. Thus, the null hypothesis is that there is no association between the disease and any SNPs in the set, and the aim is to test whether this global null hypothesis of no association can be rejected. A variety of methods has been proposed to test the global null hypothesis [10–18]. Recently we developed a novel method, the admixture maximum likelihood (AML) test, which estimates both the proportion of associated SNPs and their typical effect size [19]. We compared the power of the AML method with several previously proposed approaches by simulation and found that the maximum likelihood approach performed similarly to or better than all other tests across a wide range of scenarios for the alternative hypothesis. The rank truncated product (RTP) method also had good power, though somewhat inferior to the maximum likelihood approach in most cases. A simple Bonferroni correction performed best only when the number of associated SNPs was small.

We have been carrying out association studies in breast cancer for over a decade, and over the past five years most of our work has focussed on a comprehensive tagging approach in candidate genes using a two-stage study design. We now have data from up to 4,400 cases and 4,400 controls on 710 common variants in 117 candidate genes. Here we evaluate the evidence for associations between this set of SNPs and breast cancer using the AML and RTP methods

Results

Data were available for 710 SNPs in 120 genes. Genotype frequencies for cases and controls are shown in Table S3. Based on the trend test for association, 53 SNPs (7.5%) were significant at the 5% level, 17 (2.4%) at the 1% level, and one (0.15%) at the 0.1% level. Only one SNP, in the estrogen receptor α gene (trend test $\chi^2 = 15.6$, $p = 8 \times 10^{-5}$) reaches the $p < 0.0001$ level, which has been suggested an appropriate threshold for candidate gene studies [9]. However, it failed to reach this threshold after adjusting for population stratification by genomic control (p-trend adjusted = 0.00023). Nor was it significant at the 5% level after adjustment for multiple testing using a permutation test that allows for correlation between SNPs tested (the equivalent of a Bonferroni correction for independent hypotheses). Figure 1 shows the Q–Q plots for the univariate trend test using only the first case-control set. The Q–Q plot based on the test statistics adjusted for genomic control follows the line of equivalence for the first 600 SNPs and then starts to deviate as would be expected if a modest proportion of SNPs were associated with disease.

The AML experiment-wise test for association was significant for both the heterogeneity test ($p = 0.0031$) and the trend test ($p = 0.017$), but no association was observed using the RTP method for either the heterogeneity test or the trend test ($p = 0.12$ and $p = 0.24$; respectively). Table 1 shows the results of the AML experiment-wise tests for the complete set of SNPs and for sets of SNPs categorised according to gene functional group. The test for overall association was significant for SNPs in the cell-cycle control genes (p-heterogeneity = 0.019, p-trend = 0.035), steroid hormone signalling and metabolism genes (p-heterogeneity = 0.010, p-trend = 0.0080), and the heterogeneous set of genes categorised as "other" (p-heterogeneity = 0.0068, p-trend = 0.12).

We reanalysed the data after excluding the most significant SNP (*ESR1* rs3020314) and two SNPs (*CASP8* rs1045485 and *TGFB1* rs1982073), which have been confirmed as being associated with breast cancer in pooled data from up to 20 studies in the Breast Cancer Association Consortium [20]. After excluding these SNPs the test for both heterogeneity and trend remained significant ($p = 0.0046$ and $p = 0.031$, respectively). We also reanalysed the data after removing all 80 tSNPs in these genes. Only the heterogeneity test remained significant ($p = 0.015$, p-trend = 0.12). The test for the steroid hormone metabolism pathway remained significant after removing the most significant single SNP (p-heterogeneity = 0.018, p-trend = 0.012) and all SNPs in *ESR1* (p-heterogeneity = 0.10, p-trend = 0.044). The significance of the tests for association of "other" genes became borderline after removing the most single significant SNP in *CASP8* (p-heterogeneity = 0.0065, p-trend = 0.13) and after removing all three SNPs in *CASP8* (p-heterogeneity = 0.0058, p-trend = 0.12).

Data from the genomic control SNPs indicate some evidence of inflation of the test statistics. Given that λ, the measure of bias due to population stratification, is estimated with error, we repeated the AML tests assuming more extreme levels of bias. We used the estimate of the variance

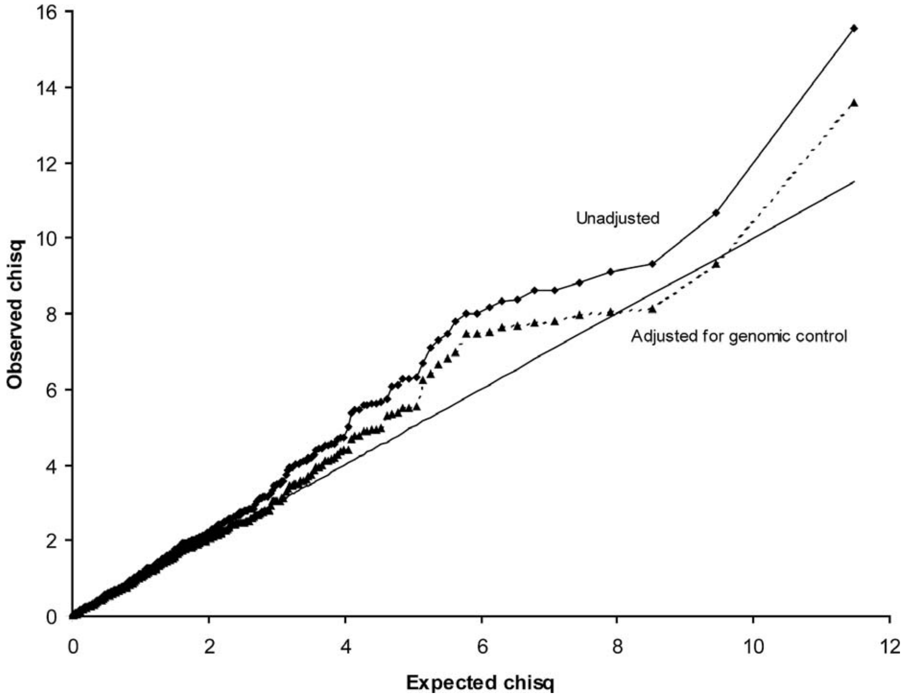

Figure 1. Q–Q Plot for Association of 710 SNPs in Candidate Genes with Breast Cancer Based on First Stage Data
doi:10.1371/journal.pgen.0030042.g001

(σ^2) and the inflation parameter (λ) to carry out sensitivity analyses with the inflation parameter set at $\lambda + \sigma$ and $\lambda + 2\sigma$. The test for heterogeneity remained significant for an inflation factors of $\lambda + \sigma$ ($p = 0.012$), but not for the most extreme estimate of the bias (inflation parameter $= \lambda + 2\sigma$, $p = 0.060$). The AML trend test was not significant under either scenario ($p = 0.070$ and $p = 0.20$, respectively). Other methods to deal with population stratification have been suggested, each of which has some advantages and disadvantages. The method of structured association uses the genomic control SNPs to assign individuals to specific subpopulations, which can then be used in stratified analyses. This method works well for a small number of discrete of subpopulations but less well for more complex population structures [21]. Gorroochurn and colleagues suggested that the χ^2_{test} statistics come from a noncentral χ^2 distribution, and that the noncentrality parameter can be estimated directly using the genomic control SNPs [22]. We adapted this method and estimated the noncentrality parameter (η) by maximum likelihood. The AML tests for association obtained using this method were similar to those used in the genomic control method (data not shown).

It is possible that our results are biased because of systematic differences in genotype frequencies between the incident and prevalent cases (survival bias). We therefore carried out a case only analysis to compare genotype frequencies in incident and prevalent cases. There was no evidence of association using the AML test ($p = 0.36$).

Discussion

In the past five years we have evaluated over 700 common genetic variants in 120 genes that are candidates for breast cancer susceptibility in a case-control study of over 4,400 cases and 4,400 controls. Based on univariate analyses no definite susceptibility alleles have emerged from this research effort. Only one SNP significant at $p < 0.0001$ was identified in this dataset, and this became less significant after adjusting for population stratification. This association was not significant at a nominal $p < 0.05$ after adjusting for multiple testing. However, it is not clear that this is an appropriate adjustment in such studies, because the result is then highly dependent on the number of SNPs that happen to have been typed at a given time.

The lack of evidence of association in these SNPs may indicate that these genes we have studied do not harbour any susceptibility variants. An alternative explanation is that one or more of the SNPs is associated with disease, but we do not have sufficient statistical power to detect these with appropriate highly stringent levels of statistical significance. For individual variants, the statistical power of the study depends on the at-risk allele frequency, the risks conferred, and the genetic model. For example, assuming that the causative SNP is tagged with $r^2 = 0.8$, a type I error rate of 10^{-5}, and genotyping success rate of 0.95, the staged study has 67% power to detect a dominant allele with a minor allele frequency (MAF) of 0.05 with an odds ratio of 1.5 or 69% power to detect a dominant allele with MAF of 0.25 with an odds ratio of 1.3. Power to detect recessive alleles is less –39% for an allele with MAF of 0.25 and an odds ratio of 1.5 and 46% for an allele with MAF 0.5 and an odds ratio of 1.3.

We have recently shown that methods that take into account the totality of the data have greater power to detect association than simple methods, such as the Bonferroni correction (or an equivalent such as permutation testing that

Table 1. Results of the Experiment-Wise Tests for 710 Candidate Gene SNPs by Gene Functional Group

Gene Functional Group	Number of Genes	Number of SNPs	Reference[a]	p-Value of Most Significant SNP[b]	AML Heterogeneity	AML Trend
17q21 Amplicon	10	24	[29]	.013	.67	.45
Animal models	13	34	[30]	.0023	.097	.11
Cell-cycle control	18	112	[31]	.0048	.019	.035
DNA repair	20	164	[32–35]	.027	.92	.96
Epigenetic modifiers	13	67	[36]	.0044	.26	.21
Growth factors	8	43	[25,37]	.019	.085	.13
Oxidative damage repair	11	65	[38]	.046	.42	.95
Steroid hormone metabolism and signalling	8	104	[39]	.00023	.010	.0080
Others	19	97	[24]	.012	.0068	.12
Total	120	710			.0031	.017

[a] Cited papers do not necessarily report totality of data for each pathway as some unpublished data.
[b] Based on trend test adjusted for genomic control.

doi:10.1371/journal.pgen.0030042.t001

takes into account correlation of the SNPs), where there are multiple SNPs associated with disease [19]. We therefore tested the hypothesis that subsets of the SNPs we have assessed are associated with breast cancer. Using the AML method, we found evidence for an overall association between common genetic variation in 120 candidate genes and breast cancer. In particular, we found some evidence of an association with SNPs in genes involved in cell-cycle control and steroid hormone metabolism. We also found evidence for population stratification in our study using data from 280 genomic control SNPs and corrected for this in all the analyses. However, the estimate of the inflation factor based on 280 SNPs is imprecise. We therefore carried out a sensitivity analysis using more extreme values of the inflation factor based on its variance. Under all but the most extreme assumptions (i.e., the upper 95% confidence limit of the inflation factor estimate) the global test of association remained significant. We therefore conclude that some proportion of the variants we have investigated are likely to be associated with breast cancer. In support of this conclusion, it is notable that stronger evidence of association has emerged for two of the SNPs we analysed, through collaborative analyses by the Breast Cancer Association Consortium [23]. One of these, in *CASP8,* was originally identified in another study but also showed evidence in our study [24]. The other, in *TGFB1,* was originally identified in a subset of cases and controls from our study [25].

Thus, our data provide further evidence for the existence of common low penetrance variants. The most efficient way to identify such variants is not clear, and considerable research funding and effort is currently being focussed on an empirical genome-wide approach rather than the candidate gene approach that has generally been the norm. The relative merits of the two approaches have not yet been defined, but our data suggest that the candidate gene approach may still be useful. However, our results also highlight the fact that alleles with modest effects, i.e., those conferring relative risks of >1.5, are likely to be the exception, and multicentre collaborations will be needed to generate adequate sample sizes.

Materials and Methods

Study participants. Patients were drawn from Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH), an ongoing population-based study, with individuals ascertained through the East Anglian Cancer Registry. All patients diagnosed with invasive breast cancer below age 55 years since 1991 and still alive in 1996 (prevalent cases, median age 48 years), together with all those diagnosed below age 70 years between 1996 and the present (incident cases, median age 54 years) are eligible to take part. As of 1 August 2005 there have been 12,767 eligible patients. Of these, 2,284 were not contacted because their general practitioner did not respond or thought that it would be inappropriate to contact the patient. Of the 10,583 patients who were contacted, 67% have returned a questionnaire, and 64% provided a blood sample for DNA analysis. Eligible patients who did not take part in the study were similar to participants except, as might be expected, the proportion of clinical stage III/IV cases was somewhat higher in nonparticipants (10% versus 5%). Female controls were randomly selected from the Norfolk component of the European Prospective Investigation of Cancer (EPIC). EPIC is a prospective study of diet and cancer being carried out in nine European countries. The EPIC–Norfolk cohort comprises 25,000 individuals resident in Norfolk, East Anglia—the same region from which the cases have been recruited. Controls are not matched to cases, but are broadly similar in age (42–81 years). The ethnic background of both cases and controls as reported on the questionnaires is similar, with >98% being white. The study is approved by the Eastern Region Multicentre Research Ethics Committee, and all patients gave written informed consent.

The total number of cases used in genetic analyses was 4,473, of whom 27% are prevalent cases. The samples were split into two sets in order to save DNA and reduced genotyping costs: the first set ($n =$ 2,270 cases and 2,280 controls) was genotyped for all SNPs, and the second set ($n =$ 2,203 cases and 2,280 controls) were then tested for those SNPs that showed marginally significant associations in set 1 (p-heterogeneity or p-trend < 0.1). Two SNPs were genotyped in stage 2 as a result of a multimarker haplotype association (see below). This staged approach substantially reduces genotyping costs without significantly affecting statistical power. Cases were randomly selected for set 1 from the first 3,500 recruited, with set 2 comprising the remainder of these plus the next 974 incident cases recruited. As the prevalent cases were recruited first, the proportion of prevalent cases was somewhat higher in set 1 than set 2 (33% versus 20%). Median age at diagnosis is similar in both sets (51 and 52 years old, respectively). There was no significant difference in the morphology, histopathological grade, or clinical stage of the cases by set or by prevalent/incident status.

Candidate gene and SNP selection. We selected genes that encode proteins in cellular pathways that are likely to be involved in breast carcinogenesis. The major pathways we studied were steroid hormone metabolism and signalling, double strand break DNA repair, oxidative damage repair, epigenetic modifiers, and cell-cycle control. We also tested genes in the 17q21 region commonly amplified in breast tumours, several genes that have been found to be important in a variety of animal models of cancer, and some carcinogen metabolism genes. For some pathways, only a small subset of genes was selected for study. For the purpose of subgroup analysis, SNPs in these genes were categorised as "other" together with the SNPs in carcinogen metabolism genes. The genes and the number of SNPs assayed for each are shown in Table S1. The principal hypothesis underlying our approach is that there are one or more common SNPs in the genes of interest that are associated with an altered risk of breast cancer. We therefore aimed to identify a set of tagging SNPs (tSNPs) that efficiently tags all the known common variants (MAF $>$ 0.05) and is likely to tag most of the unknown common variants. We used data from the International HapMap project (http://www.hapmap.org) or resequencing data from the National Institute of Environmental Health Sciences Environmental Genome Project (EGP) (http://www.niehs.nih.gov/envgenom/home.htm). The details of the methodology for tag SNP selection varied over time, but broadly speaking we have aimed to define a set of tagging SNPs such that all known common variants are correlated with a tSNP with r^2 of $>$0.8. Some SNPs are poorly correlated with other single SNPs but may be efficiently tagged by a haplotype defined by multiple SNPs, thus reducing the number of tagging SNPs needed [26]. As an alternative, therefore, we aimed for the correlation between each SNP and a haplotype of tagging SNPs to be at $>$0.8. For some genes, little information on the occurrence of common variants was available at the time the gene was studied, and the SNPs were selected for analysis based on predicted functional effects; Table S1 indicates which genes have been comprehensively tagged. We also obtained genotype data for our cases and controls for 280 randomly selected, unlinked SNPs, which were genotyped as part of an ongoing genome-wide association study (see Table S2). Data on these SNPs were used to adjust for population stratification using the genomic control method.

Genotyping methods. We genotyped all samples using the ABI PRISM 7900 sequence detection system or "Taqman" (Applied Biosystems, http://www.appliedbiosystems.com). Genomic DNA for set 1 samples was whole genome amplified by primer extension preamplification (PEP, protocol available on request). Genotype calling between PEP-amplified DNA and native genomic DNA was compared for eight Taqman assays, and the concordance was 100%. We carried out PCR on 10 ng of whole-genome amplified genomic DNA for set 1 and native genomic DNA for set 2 using TaqMan universal PCR master mix (Applied Biosystems), forward and reverse primers, and FAM- and VIC- labelled probes designed by Applied Biosystems (ABI Assay-by-Designs) in a 5-μl reaction. We read the completed PCRs on an ABI PRISM 7900 Sequence Detector in end-point mode using the Allelic Discrimination Sequence Detector software (Applied Biosystems). Cases and controls were arrayed together in 12 384-well plates, and a 13th plate contained eight duplicate samples from each of the 12 plates to ensure a good quality of genotyping. Each 384-well plate included two nontemplate controls. Concordance for duplicate samples was $>$98% for all assays. Failed genotypes were not repeated (the rate for failed genotypes did not exceed 8.3% for any of the SNPs under study). Genomic control SNPs were genotyped by Perlegen Sciences (http://www.perlegen.com) using an oligonucleotide array methodology.

Statistical methods. Association between disease and genotype for each SNP was assessed using two tests, the one-degree of freedom Cochran–Armitage trend test and the general two-degrees of freedom χ^2 test (heterogeneity test). Results for all tests were summarised in Q–Q plots, in which the ordered test statistics are plotted against the expected statistic given the rank.

To assess the overall evidence for an excess of associations, we applied two approaches, the AML and RTP methods, which are described in detail elsewhere [17,19]. In brief, the AML method formulates the alternative hypothesis in terms of the probability that a given SNP is associated with disease (α) and a measure effect size. When a SNP is associated with disease, the calculated χ^2 statistic will be distributed, asymptotically, as a noncentral χ^2 distribution with the usual degrees of freedom and a noncentrality parameter η. The noncentrality parameter is a measure of the size of effect of the SNP, is dependent on sample size, and is closely related to the contribution of the SNP to the genetic variance of the trait. If η is assumed to be the same for each associated SNP, then both α and η can be estimated by maximum likelihood, and a test of the null hypothesis can then be derived as a likelihood ratio test. Where (as is the case here) some SNPs are correlated, the full likelihood is no longer straightforward, but pseudo-maximum likelihood estimates can still be generated by the same procedure, as if the SNPs were independent. Statistical significance can then be determined by simulation. The AML method was applied to both the trend and heterogeneity tests. The RTP is simply the product of the K (arbitrary) most significant p-values from L hypothesis tests [17]. A limitation of the RTP is the need to select a truncation point. While this may be straightforward in the context of a genome-wide study where it reduces the exploratory hypotheses to a defined candidate set, it is rather arbitrary in the context of candidate gene studies. For the purpose of these analyses we chose K = 5. We adjusted all analyses for cryptic population stratification using the method described by Devlin and Roeder [27]. An inflation factor (λ) was estimated from the mean of the X^2_{trend} statistics generated on 280 unlinked, randomly selected SNPs typed on a subset of 4,037 cases and 4,012 controls as part of a separate genome-wide association study. A list of these SNPs is provided in the supplementary material. The average call rate was 99.0% in cases and 98.9% in controls. The inflation of the test statistic, adjusted for sample size, was estimated to be 1.15 (95% CI 0.94–1.36) for the trend test and 1.05 (95% CI 0.92–1.19) for the heterogeneity test.

Association tests for individual SNPs will not be independent if the markers are in linkage disequilibrium, and the application of both methods needs to allow for the correlation structure of the data. Simulations, based on permuting case-control status whilst retaining the correlation structure among makers, provide a robust approach for obtaining significance levels for these global tests. However, permutation testing is complicated by the use of a staged study design where only those SNPs significant in the first stage data are genotyped for the complete set of cases and controls. We allowed for this using the method proposed by Dudbridge [28] in which a subset of the first stage date is used as the simulated first stage data, selecting markers on the basis of that subset, and using the remainder of the first stage as the simulated second stage.

Supporting Information

Table S1. Genes Investigated as Candidates for Breast Cancer Susceptibility by Pathway

Found at doi:10.1371/journal.pgen.0030042.st001 (137 KB DOC).

Table S2. Control and Case Genotype Frequencies for 280 Genomic Control SNPs

Found at doi:10.1371/journal.pgen.0030042.st002 (521 KB DOC).

Table S3. Genotype Frequencies in Controls and Cases for 710 Candidate Gene SNPs

Found at doi:10.1371/journal.pgen.0030042.st003 (1.0 MB DOC).

Acknowledgments

We thank: all the participants who participated in these studies, the EPIC management team (K.-T. Khaw, S. Bingham, and N. Wareham) for access to control DNA, the Eastern Cancer Registry and Intelligence Unit, and collaborators Michael Gould and Mark Sternlicht. PDPP is a Senior Clinical Research Fellow of Cancer Research United Kingdom, DFE is a Principal Fellow, and BAJP is a Gibb Fellow.

Additional authors have contributed to tag SNP selection and genotyping, bioinformatics and data analysis, case recruitment and data management, control recruitment and data management, and are listed below. These additional SEARCH investigators include the following: Department of Oncology, University of Cambridge, Cambridge, United Kingdom: Jean Abraham, Ali Al-Zahrani, Caroline Baynes, Fiona Blows, Arancha Cebrian, Don Conroy, Gary Dew, Kristy Driver, Helen Field, Katie Gregory, Patricia Harrington, Fabienne Lesueur, Craig Luccarini, Karen Pooley, Serena Scollen, Clare Jordan, Hannah Munday, Barbara Perkins, Mitul Shah, and Judy West; Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom: Antonis Antoniou, Paula Smith, Deborah Thompson, Karen Redman, and Robert Luben; MRC Environmental Epidemiology Unit, Cambridge, United Kingdom: Nicholas Wareham; Department of Oncology, University of Wisconsin, Madison, Illinois, United States: Stephanie Hesselson.

Author contributions. PDPP, JT, AMD, DFE, and BAJP conceived and designed the experiments and wrote the paper. PDPP and AMD performed the experiments. PDPP, JT, and DFE analysed the data. JT, AMD, and DFE contributed reagents/materials/analysis tools.

Funding. This work was funded by Cancer Research United Kingdom and Breast Cancer Research.

Competing interests. The authors have declared that no competing interests exist.

References

1. Collaborative group on hormonal factors in breast cancer (2001) Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet 358: 1389–1399.
2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer– analyses of cohorts of twins from Sweden, Denmark and Finland. N Engl J Med 343: 78–85.
3. Easton DF (1999) How many more breast cancer predisposition genes are there. Breast Cancer Res 1: 14–17.
4. Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, et al. (2006) A genome wide linkage search for breast cancer susceptibility genes. Genes Chromosomes Cancer 45: 646–655.
5. Antoniou AC, Pharoah PDP, McMullen G, Day NE, Stratton MR, et al. (2002) A comprehensive model for familial breast cancer incorporating *BRCA1*, *BRCA2* and other genes. Br J Cancer 86: 76–83.
6. Risch N (2000) Searching for genetic determinants in the new millenium. Nature 405: 847–856.
7. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2: 91–99.
8. Chakravarti A (1999) Population genetics–making sense out of sequence. Nat Genet 21: 56–60.
9. Pharoah PDP, Dunning AM, Ponder BAJ, Easton DF (2004) Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer 4: 850–860.
10. Fisher RA (1932) Statistical methods for research workers. Edinburgh: Oliver and Boyd. 307 p.
11. Wilkinson B (1951) A statistical consideration in psychological research. Psychol Bull 48: 156–158.
12. Sidak Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 78: 626–633.
13. Edgington ES (1972) An additive model for combining probability values from independent experiments. J Psychol 80: 351–363.
14. Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751–754.
15. Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11: 2115–2119.
16. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002) Truncated product method for combining P-values. Genet Epidemiol 22: 170–185.
17. Dudbridge F, Koeleman BP (2003) Rank truncated product of P-values, with application to genomewide association scans. Genet Epidemiol 25: 360–366.
18. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005) Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76: 780–793.
19. Tyrer J, Pharoah PDP, Easton DF (2006) The admixture maximum likelihood test: A novel experiment-wise test of association between disease and multiple SNPs. Genet Epidemiol 30: 636–643.
20. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian SP, Reed MWR, et al. (2007) A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet. E-pub 11 February 2007.
21. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181.
22. Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA (2006) Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol 30: 277–289.
23. Breast Cancer Association Consortium (2006) Commonly studied single-nucleotide polymorphisms and breast cancer: Results from the Breast Cancer Association Consortium. J Natl Cancer Inst 98: 1382–1396.
24. MacPherson G, Healey CS, Teare MD, Balasubramanian SP, Reed MW, et al. (2004) Association of a common variant of the CASP8 gene with reduced risk of breast cancer. J Natl Cancer Inst 96: 1866–1869.
25. Dunning AM, Ellis PD, McBride S, Kirschenlohr HL, Healey CS, et al. (2003) A transforming growth factorbeta1 signal peptide variant increases secretion in vitro and is associated with increased incidence of invasive breast cancer. Cancer Res 63: 2610–2615.
26. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37: 1217–1223.
27. Devlin B, Roeder K, Bacanu SA (2001) Unbiased methods for population-based association studies. Genet Epidemiol 21: 273–284.
28. Dudbridge F (2006) A note on permutation tests in multistage association scans. Am J Hum Genet 78: 1094–1095; author reply 1096.
29. Benusiglio P, Lesueur F, Luccarini C, Conroy DM, Shah M, et al. (2005) Common polymorphisms in *ERBB2* and risk of breast cancer in a white population: A case-control study. Breast Cancer Res 7: R204–R209.
30. Lesueur F, Pharoah PD, Laing S, Ahmed S, Jordan C, et al. (2005) Allelic association of the human homologue of the mouse modifier Ptprj with breast cancer. Hum Mol Genet 14: 2349–2356.
31. Lesueur F, Song H, Ahmed S, Luccarini C, Jordan C, et al. (2006) Single-nucleotide polymorphisms in the RB1 gene and association with breast cancer in the British population. Br J Cancer 94: 1921–1926.
32. Kuschel B, Auranen A, McBride S, Novik KL, Antoniou A, et al. (2002) Variants in DNA double strand break repair genes and breast cancer susceptibility. Hum Mol Genet 11: 1399–1407.
33. Kuschel B, Chenevix-Trench G, Spurdle AB, Chen X, Hopper JL, et al. (2005) Common polymorphisms in ERCC2 (Xeroderma pigmentosum D) are not associated with breast cancer risk. Cancer Epidemiol Biomarkers Prev 14: 1828–1831.
34. Kuschel B, Auranen A, Gregory CS, Day NE, Easton DF, et al. (2003) Common polymorphisms in *CHEK2* (checkpoint kinase 2) are not associated with breast cancer risk. Cancer Epidemiol Biomarkers Prev 12: 809–812.
35. Healey CS, Dunning AM, Dawn Teare M, Chase D, Parker L, et al. (2000) A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. Nat Genet 26: 362–364.
36. Cebrian A, Pharoah PD, Ahmed S, Ropero S, Fraga MF, et al. (2006) Genetic variants in epigenetic genes and breast cancer risk. Carcinogenesis 27: 1661–1669.
37. Al-Zahrani A, Sandhu MS, Luben RN, Thompson D, Baynes C, et al. (2006) IGF1 and IGFBP3 tagging polymorphisms are associated with circulating levels of IGF1, IGFBP3 and risk of breast cancer. Hum Mol Genet 15: 1–10.
38. Cebrian A, Pharoah PD, Ahmed S, Smith PL, Luccarini C, et al. (2006) Tagging single-nucleotide polymorphisms in antioxidant defense enzymes and susceptibility to breast cancer. Cancer Res 66: 1225–1233.
39. Healey CS, Dunning AM, Durocher F, Teare D, Pharoah PD, et al. (2000) Polymorphisms in the human aromatase cytochrome P450 gene (CYP19) and breast cancer risk. Carcinogenesis 21: 189–193.