ERCC4 Associated with Breast Cancer Risk: A Two-Stage Case-Control Study Using High-throughput Genotyping

Roger Laughlin Milne,† Gloria Ribas,‡ Anna González-Neira,† Rainer Fagerholm,† Antonio Salas,† Emilio González,† Joaquín Dopazo,§ Heli Nevanlinna,⊥ Mercedes Robledo,⊥ and Javier Benítez†

†National Genotyping Centre, ‡Human Genetics Group, §Hereditary Endocrine Cancer Group, Human Cancer Genetics Programme, Spanish National Cancer Centre, Madrid, Spain; †Department of Obstetrics and Gynecology, Helsinki University Central Hospital, Helsinki, Finland; ‡Unidad de Genetica, Instituto de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, Galicia, Spain; and †Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia, Spain

Abstract

The failure of linkage studies to identify further high-penetrance susceptibility genes for breast cancer points to a polygenic model, with more common variants having modest effects on risk, as the most likely candidate. We have carried out a two-stage case-control study in two European populations to identify low-penetrance genes for breast cancer using high-throughput genotyping. Single-nucleotide polymorphisms (SNPs) were selected across preselected cancer-related genes, choosing tagSNPs and functional variants where possible. In stage 1, genotype frequencies for 640 SNPs in 111 genes were compared between 864 breast cancer cases and 845 controls from the Spanish population. In stage 2, candidate SNPs identified in stage 1 (nominal \( P < 0.01 \)) were tested in a Finnish series of 884 cases and 1,104 controls. Of the 10 candidate SNPs in seven genes identified in stage 1, one (rs744154) on intron 1 of ERCC4, a gene belonging to the nucleotide excision repair pathway, was associated with recessive protection from breast cancer after adjustment for multiple testing in stage 2 (odds ratio, 0.57; Bonferroni-adjusted \( P = 0.04 \)). After considering potential functional SNPs in the region of high linkage disequilibrium that extends across the entire gene and upstream into the promoter region, we concluded that rs744154 itself could be causal. Although intronic, it is located on the first intron, in a region that is highly conserved across species, and could therefore be functionally important. This study suggests that common intronic variation in ERCC4 is associated with protection from breast cancer. (Cancer Res 2006; 66(19): 9420-7)

Introduction

The ability to detect the genetic determinants of complex diseases, where many environmental and genetic factors interplay to cause disease, has until recently been very limited. Although traditional linkage analysis has successfully identified several genes causing predominantly Mendelian (monogenic) diseases, the genetic etiology of complex diseases remains unclear. It seems most plausible that more common genetic variants with modest effects on disease risk cause the bulk of this unexplained risk and that these have greater implications for public health because of their higher frequency in the general population (1–7). Association studies have for some time now attempted to identify more common causal variants by studying potentially functional polymorphisms in candidate genes; however, very few findings have been verified by replication in independent studies (1).

An alternative approach is based on the argument that disease-associated variants with modest effects might be distributed proportionately between coding and noncoding sequences of the genome (8, 9). Recent developments in high-throughput genotyping technology have made genotyping up to hundreds of thousands of marker single-nucleotide polymorphisms (SNPs) throughout the genome a possibility, both logistically and economically (9), and studies are now beginning to emerge applying these to a range of complex diseases using case-control studies (10–16). Marker SNPs are ideally chosen to maximally capture the common variation across the genome. The idea behind this approach is that associations will be detected either directly with causal variants, if genotyped, or indirectly with markers in linkage disequilibrium with causal variants (8, 17).

Association studies that test a large number of SNPs are expensive and can lead to false-positive associations if multiple testing is not adequately accounted for. At the same time, correcting for a large number of tests requires large sample sizes to maintain adequate statistical power and therefore avoid false-negative associations. In addition, confirmation of associations identified in such studies by replication in independent and adequately sized samples is essential (9). These considerations present an economic and logistical challenge to investigators seeking to produce quality research in this field. Two-stage study designs have been proposed as an efficient means of addressing these challenges (9, 18). Under this design, all SNPs are genotyped in a case-control series at stage 1 (discovery) and a reduced set of candidate SNP selected based on unadjusted \( P \) values. In stage 2 (replication), only candidate SNPs are genotyped and tested for associations in an independent case-control series, thereby reducing both genotyping costs and the number of tests to be corrected for. Both the discovery and replication studies should be of sufficient size to avoid false-positive and false-negative findings (19).

Breast cancer is a complex disease for which very little of the genetic cause is known. Although several rare, high-penetrance genes have been identified (BRCA1 and BRCA2 being the most common), these explain only a minority (30%) of familial breast cancers and a negligible proportion of sporadic breast cancers (20, 21). A polygenic model, with more common variants having modest effects on breast cancer risk, has therefore been suggested (4, 9).

We have carried out a two-stage case-control study in two European populations to identify low-penetrance genes for breast cancer, focusing on SNPs in 111 preselected candidate cancer-related genes.

Requests for reprints: Javier Benitez Human Cancer Genetics Programme, Spanish National Cancer Centre, Melchor Fernández Almagro, 3 E-28029, Madrid, Spain. Phone: 34-91-224-6965; Fax: 34-91-224-6923; E-mail: jbenitez@cnio.es.

©2006 American Association for Cancer Research. doi:10.1158/0008-5472.CAN-06-1418

Cancer Res 2006; 66: (19). October 1, 2006 9420 www.aacrjournals.org
Materials and Methods

Study population—stage 1 Spanish study. Cases were 864 women with breast cancer, recruited between 2000 and 2004 and with mean age at diagnosis of 50 years (range, 23–86 years). Of these, 574 were a consecutive series recruited via three public hospitals in Spain: 124 (22%) from Hospital La Paz and 187 (33%) from the Fundación Jiménez Díaz (both in Madrid, Spain) and 263 (46%) from Hospital Monte Naranco (Oviedo, Spain). The remaining 290 cases were women with at least one first-degree relative also affected with breast cancer (familial cases), who attended the Spanish National Cancer Centre family cancer clinic for genetic testing. They were included to increase the power to detect associations (22, 23). Of these 290 familial cases, the 110 who met criteria for “high-risk families” were tested for mutations in BRCA1 and BRCA2 as previously described (21, 24) and screened for large deletions (by multiplex ligation-dependent probe amplification, using MRC-Holland kits for BRCA1 and BRCA2) and found to be negative. The remaining 180 were tested for three specific deleterious mutations in BRCA1 and three in BRCA2, which are those most frequently observed in the Spanish population (21), and were not found to carry these mutations.

Controls were 845 Spanish women free of breast cancer at ages ranging from 23 to 86 years (mean, 53 years), recruited via the following sources: 442 (52%) from the Menopause Research Centre at the Instituto Palacios (Madrid, Spain), 239 (28%) from the College of Lawyers (Madrid, Spain), 91 (11%) from the National Blood Transfusion Centre (Madrid, Spain), 57 (7%) from the Catalan Institute of Oncology (ICO; Barcelona, Spain), and 16 (2%) from the Centre for the Investigation of Cancer (CIC; Salamanca, Spain). Controls were recruited between 2000 and 2005, and those from the latter two centers were women aged >60 years recruited as part of prospective epidemiologic studies unrelated to cancer, specifically selected for this study to be comparable with older cases on age.

Informed consent was obtained from all participants, and the study was approved by the institutional review board of Hospital La Paz.

Study population—stage 2 Finnish study. The case series of 884 women includes consecutive newly diagnosed breast cancer patients recruited in 1997 to 1998 (622 patients) and 2000 (262 patients) at the Helsinki University Central Hospital (Helsinki, Finland) and covers 79% of all breast cancer patients treated at the Department of Oncology during the collection period (described in detail in refs. 25, 26). Their mean age at diagnosis was 57 years (range, 22–96 years). This series included 214 familial cases (with no ovarian cancers in the family), 66 with a strong family history (three or more first- or second-degree relatives with breast cancer in the family, including the proband), and 148 with one first-degree relative affected with breast cancer. The 622 unselected cases recruited in 1997 to 1998 were screened for 19 deleterious Finnish BRCA1 and BRCA2 mutations as described previously (26), complemented with a more thorough screening of the genes in familial cases (27, 28), and 12 mutation carriers were identified. Among the 262 cases collected in 2000, 11 familial cases were screened for mutations and 1 was found to carry a deleterious BRCA2 mutation.

Eligible controls were a random sample of 28% of blood donors free of breast cancer, attending blood banks in Helsinki in 2003. DNA was available for 1,104 (82%) of these, and they were between 18 and 65 years of age (mean, 41 years). The study was carried out with informed consents from the patients and permissions from the Ethics Committees of the Departments of Oncology and Obstetrics and Gynecology as well as from the Ministry of Social Affairs and Health in Finland.

Candidate gene choice and SNP selection. A total of 112 candidate genes were selected for stage 1 according to the following criteria: genes previously reported to be associated with or known to be involved in cancer and genes involved in cell cycle pathways, DNA repair, cell communication, hormone metabolism, apoptosis, carcinoagen metabolism, cell adhesion, and/or signal transmission. A full list of these genes is provided in Appendix A. SNP selection across each of these genes was carried out using density as the primary criterion, with lower density in regions of higher linkage disequilibrium and higher density in regions of lower linkage disequilibrium and giving priority to tagSNPs defining common haplotypes. In addition, SNPs with potentially functional effects (causing amino acid changes, potentially causing alternative splicing, in the promoter region or in putative transcription factor binding sites) were chosen wherever possible. In general, SNPs selected had minor allele frequencies (MAFs) of at least 10%, with the exception of putative coding SNPs (where available) with a minimum MAF of 5%.

A list of validated SNPs in each gene, along with their MAFs, was compiled using publicly available information in dbSNP build 120. TagSNPs were defined using HapMap CEU genotype data (29) and Haploview software (30). Putative functional SNPs were identified using the bioinformatic tool PupaSNP (31),8 now part of the PupaSuite package. SNPs were also screened for suitability for the Illumina genotyping platform (selecting only those with an assay score >0.6, associated with a high success rate). A final total of 710 SNPs relevant to this study was included in an oligonucleotide pool assay for analysis using the Illumina platform. The average density was 1 SNP every 8.7 kb. All SNPs with associated nominal $P$ values < 0.01 in stage 1 were selected for genotyping in stage 2.

An additional 28 SNPs across the genome were independently selected to be used as markers to assess population stratification in the Spanish subjects. All 28 were chosen to be at least 100 kb from genes.

In post hoc analyses, putative causal SNPs in the promoter region of BRCA1 were identified using both PupaSNP (31) as well as additional phylogenetic analyses. For the latter, human DNA sequence was compared with that of other species using ECR Browser (32),8 which aligns nucleotide sequences using FASTA and screens for evolutionary conserved regions, defined as fragments of at least 200 bp with similarity >75%.

Genotyping. Genomic DNA from Spanish subjects was isolated from peripheral blood lymphocytes using automatic DNA extraction (MagNA Pure, Roche, Mannheim, Germany) according to the manufacturer's recommended protocols. This DNA was quantified using PicoGreen and diluted to a final concentration of 50 ng/μl for genotyping. The DNAs in the Finnish study were isolated by standard procedures using phenol-chloroform extraction and phase-lock gel tubes (Eppendorf AG, Hamburg, Germany), quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE), and diluted to a final concentration of 100 to 200 ng/μl for genotyping.

Genotyping of SNPs in candidate genes in the stage 1 Spanish study was carried out according to the manufacturer’s protocols using the Illumina Bead Array System (Illumina, Inc., San Diego, CA). For stage 1 and all post hoc SNP genotyping, at least one duplicate and one negative control were included per 96-well plate and six samples were duplicated across plates. The total number of duplicates across all plates was 35 (15 cases, 17 controls, and a non-study child-parents triad).

Genotyping of the 10 SNPs in the stage 2, the Finnish study was carried out using AmpliFluor fluorescent genotyping (KBiosciences, Cambridge, United Kingdom).10 For quality control, duplicate samples from 92 (10%) cases and 92 (8%) controls were independently reanalyzed in a blinded fashion.

Genotyping of the 28 marker SNPs to assess population stratification among Spanish subjects was carried out using the MassARRAY genotyping system (Sequenom, Inc., San Diego, CA) following the manufacturer’s instructions.

In post hoc analysis using the Spanish data, genotyping was carried out using Taqman technology (Applied Biosystems, Foster City, CA) for rs1800067 and, rs1649492 and using AmpliFluor for rs3136038 (Fig. 1) following manufacturer’s instructions in both cases.

Statistical analyses. The potential influence of population stratification in the Spanish data was assessed by genotyping a set of 28 unlinked biallelic markers in 343 randomly selected subjects (163 cases and 180 controls).

8 http://www.pupasnp.org/.
10 http://www.kbioscience.co.uk.

www.aacrjournals.org 9421 Cancer Res 2006; 66: (19). October 1, 2006

Downloaded from cancerres.aacrjournals.org on January 15, 2018. © 2006 American Association for Cancer Research.
STRAIT software version 1.1 (34) was used to test for associations between these markers and case-control status under the assumption of no population structure using the method described by Pritchard and Rosenberg (35). The program Structure version 2.0 was also applied to test for population substructure (36). Five independent replicates were run for values of K (the number of inferred clusters) from one to six, with each run consisting of 106 Markov Chain Monte Carlo steps after a burn in of length 100,000. Posterior probabilities for K were estimated based on a log likelihood: ln[P(X|K)], where X denotes the genotypes of the sampled individuals. Ancestry coefficients were estimated for each value of K (37).

Departure from Hardy Weinberg equilibrium (HWE) for all SNPs was tested in controls using the genhwi command in STATA version 8. In the stage 1 analyses, a modified Bonferroni-corrected nominal P-value threshold of 0.05/Ns was used in assessing departure from HWE, where Ns is the “effective number of independent marker loci” after consideration of linkage disequilibrium between SNPs (marker loci) on the same chromosome. Ns was calculated using the formula of Li and Li (38) by applying the web-based program SNPSpD (39, 40) to SNPs on individual chromosomes and summing estimates across chromosomes.

Associations between individual SNPs and breast cancer risk were assessed using unconditional logistic regression, comparing genotype frequencies in cases and controls and estimating odds ratios (OR) using unconditional logistic regression.

For values of the 710 SNP assays, 65 either failed genotyping (no PCR amplification, insufficient intensity for allele calling or no or poor cluster definition) or were found to have nominal P values < 0.01. Of the 28 unlinked marker SNPs studied to test for population stratification in stage 1, none had nominal P values < 0.01. Analysis of these data using Structure suggested no evidence of population stratification, with an estimated posterior probability of approximately one for K = 1, consistent across independent runs. Also consistent with these findings was that, for K = 2, the minimum and maximum ancestry coefficients were 0.45 and 0.55, respectively.

All duplicates both within and between plates genotyped in stage 1 were concordant for all SNPs. Of the 710 SNP assays, 65 either failed genotyping (no PCR amplification, insufficient intensity for cluster separation, or no or poor cluster definition) or were monomorphic. All three SNPs in MAP2K3 failed genotyping, leaving 111 genes with at least one SNP genotyped. A further 5 of the remaining 645 SNPs were found to violate HWE, with nominal P values well below the modified Bonferroni threshold of 0.0001 (based on Ns = 422), and were therefore excluded from further analyses. The number of SNPs successfully genotyped in each gene in stage 1 is included in Appendix A.

The 640 SNPs in 111 genes successfully genotyped and investigated for associations with breast cancer in stage 1 represented the equivalent of an estimated 417 independent loci. A full list of these SNPs, including estimated MAFs for controls, can be found at Bioinformatics Web site. Their genomic positions are summarized in Table 1. Allele frequencies observed in controls were highly consistent with those reported for Centre d’Etude du Polymorphisme Humaine individuals by HapMap (29), with a high positive correlation of 0.91 (43).

Of the 640 SNPs tested in stage 1, 10 were found to have nominal P values associated with the best-fitting model <0.01. These 10 SNPs were from seven genes (one in each of MSH3, MSH6, BUB1B, MLL3, ERCC4, MSH2, and MSH4).
controls. were studied in the stage 2 Finnish series of 884 cases and 1,104
formultipletesting using either method considered. These 10 SNPs putative associations were statistically significant after adjustment
highly consistent for all three (e.g., for rs744154 in
for family history or age at diagnosis), and OR estimates were
women recruited outside of Madrid, and excluding cases selected
repeated under three scenarios (adjusting for age, excluding
variable (multiplicative) codominant model for one. Analyses were
best-fitting model was a single variable one, a dominant model
being the best fit for six, a recessive model for three, and a single-
variable (multiplicative) codominant model for one. Analyses were
repeated under three scenarios (adjusting for age, excluding
women recruited outside of Madrid, and excluding cases selected
for family history or age at diagnosis), and OR estimates were
highly consistent for all three (e.g., for rs744154 in ERCC4, the OR
was 0.64 versus 0.64, 0.60, and 0.64, respectively). None of the
putative associations were statistically significant after adjustment
for multiple testing using either method considered. These 10 SNPs
were studied in the stage 2 Finnish series of 884 cases and 1,104
controls.

and GSTP1 and two in each of BCL2, EGFR, and ERCC4) on seven
individual chromosomes. Results, including the chromosome, gene,
and gene region of each SNP, are summarized in Table 2A. The two
SNPs in BCL2 were in linkage equilibrium (pairwise correlation,
\( r^2 < 0.01 \)), those on EGFR had \( r^2 = 0.18 \), and the two on ERCC4 were
in very high linkage disequilibrium, with \( r^2 = 0.96 \). For all 10 SNPs, the best-fitting model was a single variable one, a dominant model
being the best fit for six, a recessive model for three, and a single-
variable (multiplicative) codominant model for one. Analyses were
repeated under three scenarios (adjusting for age, excluding
women recruited outside of Madrid, and excluding cases selected
for family history or age at diagnosis), and OR estimates were
highly consistent for all three (e.g., for rs744154 in ERCC4, the OR
was 0.64 versus 0.64, 0.60, and 0.64, respectively). None of the
putative associations were statistically significant after adjustment
for multiple testing using either method considered. These 10 SNPs
were studied in the stage 2 Finnish series of 884 cases and 1,104
controls.

For all 10 SNPs studied in the Finnish case-control series in
stage 2, the concordance rate between duplicate samples in the
quality control analyses was 100%. None of the SNPs presented
evidence of departure from HWE, although one (rs8191439 in
GSTP1) was monomorphic. For the remaining nine, representing
7.9 independent loci, the two SNPs (rs744154 and rs3136079) in
ERCC4 had unadjusted \( P \) values < 0.05 (Table 2B), whereas for the
other seven the difference in allele frequencies of cases versus
controls was in the opposite direction to that observed in stage 1.
For both SNPs in ERCC4, the best-fitting model was recessive, as
observed in stage 1, with OR estimates also highly consistent with
results from stage 1 (OR, 0.57 for rs744154 and 0.66 for rs3136079;
Tables 2B and 3). Although these two SNPs were in very high
linkage disequilibrium in the Finnish sample as well (\( r^2 = 0.92 \)),
one of them (rs744154) remained statistically significant after
correction for multiple testing (Bonferroni-adjusted \( P = 0.04 \)). This
was the case regardless of the correction method used, although
the association with rs3136079 was marginal for FDR <15%
(\( P = 0.02 \) versus a threshold of 0.02). The adjusted \( P \)-value of 0.04
was confirmed using the permutation test (\( P = 0.044 \)). The OR
estimate and unadjusted \( P \)-value for rs774154 were even lower
after adjustment for age (OR, 0.51) and also when women aged >60
years were excluded (OR, 0.45), when known mutation carriers
were excluded (OR, 0.56), and when cases selected for family
history were excluded (OR, 0.56). The OR estimate for the recessive
effect of rs744154 from both studies (stages 1 and 2) combined was
0.61 (unadjusted \( P = 0.002 \)).

The observed association could be due to rs744154 directly or
to it being in linkage disequilibrium with another functional SNP.
HapMap data suggest that the entire gene forms a high linkage

### Table 1. Distribution of the 640 SNPs genotyped in stage 1 in terms of genomic position

<table>
<thead>
<tr>
<th>Region of gene</th>
<th>No. SNPs</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>87</td>
<td>14</td>
</tr>
<tr>
<td>5' upstream</td>
<td>74</td>
<td>12</td>
</tr>
<tr>
<td>3' downstream</td>
<td>77</td>
<td>12</td>
</tr>
<tr>
<td>Intrinsic</td>
<td>402</td>
<td>63</td>
</tr>
<tr>
<td>Total</td>
<td>640</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2.

#### A. Summary information about the 10 candidate SNPs identified in stage 1, including location in the genome, most likely model, estimated OR, and corresponding, unadjusted \( P \)-value

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>Chromosome</th>
<th>Region</th>
<th>Model</th>
<th>OR</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1800932</td>
<td>MSH6</td>
<td>2</td>
<td>Coding</td>
<td>Dominant</td>
<td>0.75</td>
<td>0.004</td>
</tr>
<tr>
<td>rs3776978</td>
<td>MSH3</td>
<td>5</td>
<td>Intrinsic</td>
<td>Dominant</td>
<td>1.40</td>
<td>0.004</td>
</tr>
<tr>
<td>rs917879</td>
<td>EGFR</td>
<td>7</td>
<td>Intrinsic</td>
<td>Codominant</td>
<td>1.26*</td>
<td>0.001</td>
</tr>
<tr>
<td>rs984654</td>
<td>EGFR</td>
<td>7</td>
<td>Intrinsic</td>
<td>Dominant</td>
<td>0.71</td>
<td>0.001</td>
</tr>
<tr>
<td>rs8191439</td>
<td>GSTP1</td>
<td>11</td>
<td>5' upstream</td>
<td>Dominant</td>
<td>0.33</td>
<td>0.007</td>
</tr>
<tr>
<td>rs3759839</td>
<td>BUB1B</td>
<td>15</td>
<td>5' upstream</td>
<td>Dominant</td>
<td>1.34</td>
<td>0.008</td>
</tr>
<tr>
<td>rs744154</td>
<td>ERCC4</td>
<td>16</td>
<td>Intrinsic</td>
<td>Recessive</td>
<td>0.64</td>
<td>0.007</td>
</tr>
<tr>
<td>rs3136079</td>
<td>ERCC4</td>
<td>16</td>
<td>Intrinsic</td>
<td>Recessive</td>
<td>0.64</td>
<td>0.006</td>
</tr>
<tr>
<td>rs156483</td>
<td>BCL2</td>
<td>18</td>
<td>3’ downstream</td>
<td>Recessive</td>
<td>1.95</td>
<td>0.004</td>
</tr>
<tr>
<td>rs987802</td>
<td>BCL2</td>
<td>18</td>
<td>Intrinsic</td>
<td>Dominant</td>
<td>1.38</td>
<td>0.004</td>
</tr>
</tbody>
</table>

#### B. Summary information about the two SNPs in ERCC4 replicated in stage 2, including location in the genome, most likely model, estimated OR, and corresponding, unadjusted one-sided \( P \)-value

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>Location</th>
<th>Model</th>
<th>OR</th>
<th>( P )</th>
<th>TH  (^1)</th>
<th>TH  (^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs744154</td>
<td>ERCC4</td>
<td>Intron 1</td>
<td>Recessive</td>
<td>0.57</td>
<td>0.005</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>rs3136079</td>
<td>ERCC4</td>
<td>Intron 2</td>
<td>Recessive</td>
<td>0.66</td>
<td>0.02</td>
<td>0.006</td>
<td>0.007</td>
</tr>
</tbody>
</table>

\(^{*}\)OR per copy of the minor allele.  
\(^{1}\)Bonferroni-adjusted \( P \)-value threshold (based on 7.9 independent loci and a global type I error rate of 5%).  
\(^{2}\)\( P \)-value threshold to maintain a FDR <5% (based on nine SNPs tested).
disequilibrium block, including the promoter region (Fig. 1). We therefore searched the entire gene for coding SNPs and found only one, rs1800067 (R415Q) on exon 8, with MAF >0.05. We genotyped rs1800067 in 543 sporadic cases and 560 controls and found no evidence of an association, with the per-allele OR estimate (OR, 0.89; \( P = 0.4 \)) in the opposite direction to that previously reported (44, 45). Although the putative causal SNP could lie anywhere in the extensive region of high linkage disequilibrium that includes rs744154, that the association with rs3136679 on intron 2 did not reach statistical significance in stage 2, that the results from analysis of the SNPs on each of introns 8 and 9 and exon 11 at stage 1 did not reach the established \( P \)-value threshold (Table 3), and that the one candidate causal SNP (rs1800067) on exon 8 seemed unassociated together suggest that the causative SNP might most likely lie in the upstream end or promoter region of the gene (Fig. 1).

We carried out phylogenetic analyses and found that sequences in the first intron and in the promoter region are highly conserved across species, highlighting that SNPs lying therein (including rs744154) could be of functional importance and therefore cause the observed protection from breast cancer. We screened these regions to identify additional SNPs with allele frequencies >5% located at potential transcription factor binding sites and/or in highly conserved sequences. We selected an insertion/deletion polymorphism (rs11337253) and a SNP (rs11649492) located 0.6 and 4.6 kb upstream of \( \text{ERCC4} \), respectively, as candidate causal loci and genotyped them in the Spanish case-control series. PupaSNP predicts that rs11337253 is located at an HNF1 transcription factor binding region, whereas rs11649492 lies in a sequence that is highly conserved in both dog and mouse. The insertion/deletion polymorphism (rs11337253) could not be genotyped due to difficulty designing probes for Taqman or Amplifluor, given the poly(A) flanking sequence (9A). We instead genotyped rs3136038 located just 85 bp upstream of rs11337253.

Table 3 summarizes the comparison of genotype distributions between cases and controls for all SNPs studied in \( \text{ERCC4} \).

<table>
<thead>
<tr>
<th>SNP (position(^*)), location (^1), sample</th>
<th>N (^1)</th>
<th>MAF(^1)</th>
<th>Alleles(^1)</th>
<th>Het.(^1) OR (( P ))</th>
<th>Hom.(^{**}) OR (( P ))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNPs initially studied in stage 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs744154 (13,922,582), intron 1</td>
<td>1,700</td>
<td>0.32</td>
<td>CG</td>
<td>0.92 (0.4)</td>
<td>0.62 (0.005)</td>
</tr>
<tr>
<td>Stage 1 (Spanish study)</td>
<td>1,900</td>
<td>0.23</td>
<td>CG</td>
<td>1.01</td>
<td>0.57</td>
</tr>
<tr>
<td>Stage 1 (Finnish study)</td>
<td>1,697</td>
<td>0.33</td>
<td>TG</td>
<td>0.92 (0.4)</td>
<td>0.62 (0.004)</td>
</tr>
<tr>
<td>Stage 1 (Finnish study)</td>
<td>1,895</td>
<td>0.24</td>
<td>TG</td>
<td>1.03</td>
<td>0.67</td>
</tr>
<tr>
<td>rs3136079 (13,926,738), intron 2</td>
<td>1,695</td>
<td>0.38</td>
<td>CT</td>
<td>0.93 (0.5)</td>
<td>0.71 (0.03)</td>
</tr>
<tr>
<td>Stage 1 (Spanish study)</td>
<td>1,597</td>
<td>0.39</td>
<td>GA</td>
<td>0.92 (0.5)</td>
<td>0.69 (0.02)</td>
</tr>
<tr>
<td>rs3136202 (13,945,675), intron 9</td>
<td>1,680</td>
<td>0.32</td>
<td>TC</td>
<td>0.90 (0.3)</td>
<td>0.67 (0.02)</td>
</tr>
<tr>
<td>Stage 1 (Spanish study)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1799801 (13,949,459), exon 10 (^{++})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 1 (Spanish study)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SNPs studied post hoc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs11649492 (13,916,919), promoter</td>
<td>1,622</td>
<td>0.38</td>
<td>GC</td>
<td>0.88 (0.2)</td>
<td>0.83 (0.2)</td>
</tr>
<tr>
<td>Spanish study (post hoc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3136038 (13,920,880), promoter</td>
<td>1,583</td>
<td>0.37</td>
<td>CT</td>
<td>1.04 (0.7)</td>
<td>0.78 (0.1)</td>
</tr>
<tr>
<td>Spanish study (post hoc)</td>
<td>1,103(^{1})</td>
<td>0.11</td>
<td>GA</td>
<td>0.82 (0.2)</td>
<td>1.33 (0.6)</td>
</tr>
<tr>
<td>rs1800067 (13,936,534), exon 8 (^{++})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish study (post hoc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Note: Two-sided \( P \) values for stage 1 only.  
\(^*\)Chromosomal position (dbSNP125).  
\(^1\)Location in, or relative to, the gene.  
\(^2\)Number of subjects with genotypes.  
\(^3\)Minor allele frequency among controls.  
\(^4\)Minor allele in bold.  
\(^*\)Comparing heterozygotes with homozygotes in the more common allele.  
\(^**\)Comparing homozygotes in the rare allele with homozygotes in the more common allele.  
\(^{++}\)Synonymous coding SNP.  
\(^{1}\)Nonsynonymous coding SNP.  
\(^{1}\)Studied in a subsample of cases and controls.


Breast Cancer Candidate Gene Association Study

were only three haplotypes with estimated frequency >1%, the two most common (CTCGT and GGTAC) being yin yang (46) and accounting for 62% and 29% of all haplotypes, respectively, reflecting the high linkage disequilibrium observed across the entire gene. GGTAC includes the minor allele of each SNP and was associated with reduced breast cancer risk compared with CTCGT (per copy OR, 0.84; \( P = 0.02 \)). Results did not change substantially when these analyses were repeated, including rs3136038, located 85 bp from the insertion-deletion polymorphism in the promoter (yin yang haplotypes accounting for 91% of all haplotypes and TGGTAC having OR of 0.83; \( P = 0.01 \)). These estimates are practically identical to the per-allele risk estimated for rs744154 alone (OR, 0.83; \( P = 0.01 \)).

Discussion

In this study, a total of 640 SNPs in 111 cancer-related genes was successfully genotyped in 864 cases and 845 controls, identifying 10 candidate SNPs in 7 genes based on an arbitrary \( P \)-value threshold of 0.01. After testing these 10 SNPs in an independent Finnish series with 884 cases and 1,104 controls, one SNP (rs744154) on intron 1 of \( ERCC4 \) was found to be significantly associated with breast cancer (adjusted \( P = 0.04 \)). It seems to act in a recessive manner, with the minor allele protecting against breast cancer (combined OR, 0.61; unadjusted \( P = 0.0002 \)). Further analyses of SNPs in the promoter region of \( ERCC4 \) and in the same high linkage disequilibrium block as rs744154 using the Spanish data did not identify an alternative candidate causal locus.

Only one study to date has published results from the application of high-throughput genotyping to the study of breast cancer (12, 47, 48), studying 25,494 marker SNPs in 14,000 to 16,000 genes among 254 breast cancer cases and 268 controls. In the first two publications (12, 48), the 52 SNPs with nominal \( P \) values < 0.05 were selected and studied in 368 cases and 330 controls from Germany and Australia. One SNP in \( ICAM5 \) was considered consistently associated with case-control status (unadjusted \( P < 0.05 \)) in the original and replication series and another in \( NuMA \) in the original and pooled series (but not in the replication series alone). Their third publication (47) applied more relaxed criteria to declare associations and identified \( DPF3 \) as an additional candidate gene for further study. None of the SNPs studied were statistically significantly associated with breast cancer after correction for multiple testing.

The present study applies this two-stage candidate SNP selection and replication approach (9) but with much larger samples, both enhanced for family history of disease among cases and each from a distinct European population, and the replicated result was statistically significant after correction for multiple testing.

\( ERCC4 \), also known as \( XPF \), is involved in the nucleotide excision repair pathway and is linked to susceptibility to xeroderma pigmentosum, a rare recessive syndrome that includes photosensitivity and malignant tumor development (49). \( ERCC4 \) plays an important role in recombination repair, mismatch repair, and possibly immunoglobulin class switching because of its unique function in damage site recognition (50). Several studies have investigated associations between polymorphisms in \( ERCC4 \) and breast cancer risk (44, 45, 51). Smith et al. (45) and Mechanic et al. (44) both found that, for the nonsynonymous coding SNP on exon 8, rs1800067 (Fig. 1), the percentage of rare (AA) homozygotes in rs1800067 was higher for cases than controls (among Whites/Caucasians), although in neither was this statistically significant.

Lee et al. (51) found no evidence of an association with breast cancer with the synonymous coding SNP in exon 11, rs1799801 (Fig. 1); however, they suggested that carriers of both the variant in this SNP and Asp312Asn in \( ERCC2 \) might be at increased breast cancer risk. Our results (for rs1799801 studied at stage 1 and rs1800067 studied post hoc) do not support these findings, as per-allele \( P \) OR estimates were in the opposite directions (respectively) for both SNPs (Table 3).

The observed association with the intronic SNP, rs744154, in this study would be explained by either rs744154 being directly protective or it being in linkage disequilibrium with another protective variant. HapMap data suggest that the entire \( ERCC4 \) gene forms a high linkage disequilibrium block, including the promoter region (Fig. 1). We studied rs1800067 (R415Q), the only nonsynonymous coding SNP identified in \( ERCC4 \) with an allele frequency >1%, and found no evidence of an association with breast cancer. Having observed that the four stage 1 SNPs located downstream of rs744154 showed weaker evidence of association with breast cancer, we focused on the promoter region and on potential transcription factor binding sites and highly conserved regions in particular, considering these strong candidates to contain functionally important SNPs (9). Associations were not detected for either of two additional SNPs located in highly conserved sequences in this region. One of those (rs3136038) was just 85 bp upstream of a candidate insertion-deletion polymorphism (rs11337253), which we were unable to genotype directly (Fig. 1); however, the lack of association and high linkage disequilibrium observed between all SNPs studied suggests that the latter is unlikely to be causal.

As mentioned above, rs744154, though intronic, could itself be causal. It is plausible that disease-associated variants with modest effects will be distributed proportionately between coding and noncoding sequences of the genome (8), and conserved noncoding regions in particular are often functionally important (9, 52). Indeed, several studies have found functional intronic variants associated with disease (53, 54). Furthermore, various studies have identified transcriptional regulation elements on the first intron of human genes (55–57). Rs744154 is located in a sequence of intron 1 that is highly (>90%) conserved in \( Canis familiaris \), indicating a potential functional role. Functional studies are being planned to clarify this.

Population stratification is unlikely to have confounded the observed association of rs744154 with breast cancer risk. Although the power to detect population stratification in the subsample analyzed was limited, Structure convincingly identified only one population stratum among cases and controls. Results were also highly consistent when women recruited outside of Madrid were excluded. That the association between rs744154 in \( ERCC4 \) was replicated in an independent study of a distinct European population confirms that population stratification is unlikely to have been influential. The inclusion of cases selected for family history of disease in stage 1 also seems not to have influenced the findings of this study because the OR estimates for rs744154 from both samples did not change when selected cases were excluded.

In summary, we have conducted a two-stage case-control study, first screening SNPs in 111 cancer-related genes in a large Spanish case-control series and then validating associations in a large Finnish series. A SNP (rs744154) on intron 1 of \( ERCC4 \) was associated with breast cancer risk after adjustment for multiple testing in stage 2, indicating that common variation in \( ERCC4 \) is associated with protection from the disease.
Appendix A. List of the 112 preselected cancer-related genes, in alphabetical order, with the number of successfully genotyped SNPs per gene in parenthesis


SOS1 (3), STAT4 (1), STK6 (4), TERT (3), TNFRSF6 (10), TNFSF6 (3), TP53L (16), TRAF1 (1), TRAF6 (2), VEGF (7), WRN (11), XPA (3), XPC (6), XRCC1 (7), XRCC2 (5), XRCC3 (2), and XRCC4 (16). *None of the three SNPs in MAP2K3 were successfully genotyped, leaving 111 genes studied in cases and controls.

Acknowledgments

Received 4/19/2006; revised 6/26/2006; accepted 7/20/2006.

Grant support: The Spanish component of this work was funded by the Genome Spain Foundation. The Finnish study was financially supported by the Helsinki University Central Hospital Research Fund, the Academy of Finland (110660), the Finnish Cancer Society, and the Sigrid Juselius Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Jos´e Ignacio Arias (Hospital Monte Naranco), Pilar Zamora (Hospital La Paz), ´Alvaro Ruibal (Fundaci´on Jim´enez D´ıaz), Santiago Palacios (Instituto Palacios), Silvia de Sanjose (ICO), and Rogelio Gareondo Sarmiento (CIC) for the use of Spanish samples and controls; Charo Alfonso, Christian Torrentes, Alicia Barroso, Victoria Fern´andez, Rocio Let´on, and ´Fatima Mercadillo for their technical assistance in Spain; and Drs. Hannaleena Errola and Carl Blomqvist as well as RN Nina Paalakka for their kind help with the patient contacts and sample collection.

References

7. Pharoah PD, Antoniou AC, Rosso S, Z后面的内容被截断了，无法完整阅读。
**ERCC4 Associated with Breast Cancer Risk: A Two-Stage Case-Control Study Using High-throughput Genotyping**

Roger Laughlin Milne, Gloria Ribas, Anna González-Neira, et al.


**Updated version**
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/66/19/9420

**Cited articles**
This article cites 51 articles, 9 of which you can access for free at:
http://cancerres.aacrjournals.org/content/66/19/9420.full#ref-list-1

**Citing articles**
This article has been cited by 7 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/66/19/9420.full#related-urls

**E-mail alerts**
Sign up to receive free email-alerts related to this article or journal.

**Reprints and Subscriptions**
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

**Permissions**
To request permission to re-use all or part of this article, use this link http://cancerres.aacrjournals.org/content/66/19/9420.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.