

Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations

Christopher A. Haiman^{1,*}, Chris Hsu¹, Paul I.W. de Bakker², Melissa Frasco¹, Xin Sheng¹, David Van Den Berg¹, John T. Casagrande¹, Laurence N. Kolonel³, Loic Le Marchand³, Susan E. Hankinson⁴, Jiali Han⁴, Alison M. Dunning⁵, Karen A. Pooley⁵, Matthew L. Freedman^{2,6}, David J. Hunter⁷, Anna H. Wu¹, Daniel O. Stram¹ and Brian E. Henderson¹

¹Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA, ²Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA, ³Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, HI 96813, USA, ⁴Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA 02115, USA, ⁵Cancer Research UK, Department of Oncology, Strangeways Research Laboratory, University of Cambridge, UK, ⁶Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA and ⁷Epidemiology Department, Harvard School of Public Health, Boston, MA 02115, USA

Received September 21, 2007; Revised November 12, 2007; Accepted November 29, 2007

Genetic association studies of multiple populations investigate a wider range of risk alleles than studies of a single ethnic group. In this study, we developed a multiethnic tagging strategy, exploiting differences in linkage disequilibrium (LD) structure between populations, to comprehensively capture common genetic variation across 60 genes spanning multiple DNA repair pathways, in five racial/ethnic populations. Over 2600 SNPs were genotyped in each population and single- and multi-marker predictors of common alleles were selected to capture the LD patterns specific to each group. Coding variants ($n = 211$) were genotyped to test whether combinations of putative functional variants in DNA repair pathway genes could have cumulative effects on risk. Tests of association were conducted in a multiethnic breast cancer study (2093 cases and 2303 controls), with validation of the top allelic associations ($P \leq 0.01$) performed in additional studies of 6483 cases and 7309 controls. A variant in the *FANCA* gene (rs1061646, 0.15–0.68 frequency across populations) was associated with risk in the initial study ($P = 0.0052$), and in the replication studies ($P = 0.032$). In a combined analysis (8556 cases and 9605 controls), this SNP yielded an 8% increase in risk per allele. Combinations of coding variants in these genes were not associated with breast cancer and together, these data suggest that common variation in these DNA repair pathway genes are not strongly associated with breast cancer risk. The methods utilized in this study, applied to multiple populations, provide a framework for testing in association studies in diverse populations.

INTRODUCTION

The ability to repair and faithfully replicate DNA is crucial and multiple mechanisms have evolved to maintain

genomic integrity. Deficiencies in many DNA-damage response and repair processes lead to highly penetrant genetic disorders, many of them with cancer as a predominant

*To whom correspondence should be addressed. Tel: +1 3234427755; Fax: +1 3234427749; Email: haiman@usc.edu

phenotype, such as xeroderma pigmentosum, Werner's syndrome, Fanconi anemia (FA) and Bloom syndrome (1).

Inter-individual differences in DNA repair capacity, as determined from *in vitro* assays, has been reported in multiple studies, with diminished repair capacity associated with an increased risk of breast cancer (2,3). Rare germline variants in genes involved in sensing and/or repairing DNA damage, such as *BRCA1*, *BRCA2*, *ATM*, *FANC* genes and *CHEK2* are established markers of breast cancer susceptibility and highlight the importance of aberrant DNA repair activity in the pathogenesis of the disease (4). The extent to which common forms of genetic variation in DNA repair-related pathway genes underlie differences in repair competence, and breast cancer risk, however, remains unclear.

Given the clear involvement of DNA repair in familial breast cancer, it is hypothesized that common coding and non-coding variation in genes in DNA repair pathways may contribute risk to breast cancer. Resequencing of large numbers of DNA repair genes has been performed to catalogue coding variants, with over 30% of identified variants predicted, based on *in silico* methods, to have altered activities (5). Studies conducted to investigate the role of common variation in candidate DNA repair-related pathway genes in relation to breast cancer risk have been limited in scope, focusing on only a small number of genes and/or coding variants, and associations have not been shown to replicate across multiple studies and racial/ethnic populations (6–10).

Empirical data generated by the HapMap Project (11; www.hapmap.org/) and others (12–14), supports the selection of informative markers (tag SNPs) to efficiently test common heterozygosity for association with disease risk. In this study, we applied such an approach to comprehensively examine common variation in coding and non-coding regions across 60 DNA repair-related pathway genes for association with breast cancer risk. These pathways/genes included direct reversion repair (*MGMT*), base excision repair (*APE1*, *LIG3*, *NEIL1*, *NEIL2*, *OGG1*, *PARP1*, *XRCC1*), nucleotide excision repair (*XPA*, *ERCC3*, *XPC*, *ERCC2*, *ERCC4*, *ERCC5*, *ERCC1*, *LIG1*, *ERCC6*, *ERCC8*, *RPA1*, *RPA2*, *RPA3*), double-strand break (DSB) repair via (i) homologous recombination (*RAD50*, *RAD51*, *RAD52*, *XRCC2*, *XRCC3*, *NBS*, *MRE11A*) or (ii) non-homologous end-joining (*XRCC4*, *XRCC5*, *XRCC6*, *DCLRE1C*, *PRKDC*, *LIG4*), DNA polymerases, nucleases and helicases (*POLB*, *POLD1*, *POLE*, *POLI*, *POLK*, *PCNA*, *FEN1*, *BLM*), DNA-cross-link repair (*FANCA*, *FANCC*, *FANCD2*, *FANCE*, *FANCF*, *FANCG*), mismatch repair (*MSH2*, *MSH3*, *MSH6*, *MLH1*, *MLH3*, *PMS1*, *PMS2*) and genes involved in DNA damage recognition and response (*ATM*, *ATR*, *CHEK1*, *CHEK2*, *TP53*) (15).

High-density SNP genotyping across each locus was initially performed in a multiethnic panel to characterize linkage disequilibrium (LD) patterns, and a multiethnic tagging strategy was utilized to comprehensively capture common variation in all populations. We also analyzed the independent and combined effects of 211 coding SNPs at these candidate loci. Association testing with breast cancer risk was conducted in a large case-control study of African Americans, Native Hawaiians, Japanese Americans, Latinos and European Americans in the Multiethnic Cohort Study (MEC; 2093

cases and 2303 controls) (16). Validation of the top allelic associations was performed in large replication studies that included an additional 6483 cases and 7309 controls.

RESULTS

In this study, common genetic variation across the 60 genes was thoroughly captured. We characterized the LD patterns at each candidate locus by densely genotyping 2627 SNPs in our MEC reference panel of five populations, with an average SNP spacing of 2.5 kb (range across populations 2.27–2.68 kb; Table 1, and Supplementary Material, Tables S1 and S2). For these 60 genes, 1367 tag SNPs (which included 211 coding SNPs), and defined multi-marker tests (1373) (17), captured $\geq 95\%$ of common SNPs genotyped in each population (based on an $r^2 \geq 0.8$; Table 1 and Supplementary Material, Tables S3 and S4). The mean maximum r^2 between the tag SNPs and multi-marker haplotypes, and the non-tags was excellent, ranging from 0.94 to 0.98 across populations. This panel also captured 91% of common SNPs with $r^2 \geq 0.8$ (mean maximum r^2 of 0.94) in the CEU HapMap population, which includes phase II HapMap data (Rel#21/phaseII Jul 06) and additional SNPs examined in this study ($n = 3741$). A summary of the tag SNP coverage by gene and population is provided in Supplementary Material, Table S4.

In the ethnic-pooled case-control analysis in the MEC (2074 cases and 2297 controls), the distribution of P -values for the tag SNPs (1367) was fairly consistent with expectation. However, we did observe some excess P -values in the range from 0.01 to 0.001 (Supplementary Material, Figure 1). This was more pronounced when the imputations were used than when only the tag SNPs were tested for association, which at least partly reflects that most of the imputations are actually based on a single SNP (i.e. the 'same' P -value appears multiple times for a tag SNP that tags lots of SNPs). Principal components was used to examine potential population stratification (see Materials and Methods) (18). In this analysis, none of the eigenvectors were found to be significantly associated with risk and results were similar following adjustment for population stratification. In the MEC, we observed nominally significant associations at $P \leq 0.05$ with 120 SNPs (4.4%; Supplementary Material, Table S5). Thirty-nine SNPs (which included coding SNPs *F390L* in *MSH2* and *L868P* in *BLM*) were significant at $P \leq 0.01$ (1.4%). Three of these (including *F390L* in *MSH2*, $P = 0.0019$) were rare (<1%) in four of the five ethnic groups and were not studied further. Of the remaining 36 SNPs, we selected 15 (including *L868P* in *BLM*) to examine in the replication studies. The remaining 21 SNPs were in strong LD with these 15 tags in each MEC population (pairwise $r^2 \geq 0.77$); the mean pairwise r^2 for a given SNP ranged from 0.84 to 1.0. The odds ratios for these 15 SNPs are presented in Table 2. Adjustment for population stratification did not affect the overall SNPs selected for replication or the magnitude of the associations (Supplementary Material, Table S5).

In the replication phase, we selected 12 common SNPs in MEC Japanese for follow-up in an Asian-American breast cancer study (Supplementary Material, Table S6) (19). Only SNP rs1061646 in *FANCA* (OR = 1.17; 95% CI, 1.01–1.36;

Table 1. The coverage of common variation and the percentage of common SNPs captured by single-SNP and multi-marker tests in each population

Population	No. of SNPs used in LD characterization ^a	Average spacing (kb) of common SNPs ($\geq 5\%$) genotyped in LD characterization (range across genes)	Percentage of common SNPs captured by the tag SNPs+multi-marker tests ^b	Mean maximal r^2 for all genes (range across genes)
African American	2273	2.27 (1.00–7.49)	95%	0.94 (0.86–1.0)
Japanese American	1964	2.68 (1.04–6.86)	98%	0.98 (0.79–1.0)
Latino	2201	2.35 (1.08–6.52)	97%	0.97 (0.77–1.0)
Native Hawaiian	2131	2.46 (1.08–8.42)	98%	0.98 (0.92–1.0)
European Americans	2122	2.57 (1.08–17.46)	97%	0.98 (0.75–1.0)
CEU HapMap	3741 ^c	1.79 (0.47–11.89)	91%	0.94 (0.70–1.0)

^aSNPs in *NBS* are only included for the CEU HapMap population (Data Rel#16/phase I Mar05).

^b $r^2 \geq 0.80$.

^cIncludes common SNPs ($\geq 5\%$) from HapMap as well as SNPs genotyped in 20 of the 30 CEU HapMap trios in this study.

$P = 0.043$; Table 3) was found to be nominally associated with risk, with the direction of the association consistent across the Chinese, Japanese and Filipino populations, as well as with four of the five MEC populations. SNP rs1061646 and three SNPs with P -values < 0.01 that were less common in the MEC Japanese, were next examined among European Americans from the Nurses' Health Study cohort (8). Again, SNP rs1061646 was nominally associated with risk (OR = 1.14; 95% CI, 1.01–1.28; $P = 0.032$; Table 3), with the direction and magnitude of the effect consistent with the previous two studies. However, this SNP was not found to be significantly associated with increased risk in the SEARCH study (20), a third large replication study comprised of European Whites (OR = 1.02; 95% CI, 0.96–1.09; $P = 0.57$), or among the total of 5739 cases and 6321 controls of European ancestry in the replication studies combined ($P = 0.13$; Table 3). In a pooled analysis of all three replication studies among all ethnicities, SNP rs1061646 was only modestly associated with risk (OR = 1.06; 95% CI, 1.01–1.12; $P = 0.032$). Including data from the MEC, this SNP was associated with an 8% increase in risk per allele ($P = 0.0014$; Table 3), however this association was no longer significant after accounting for performing > 1400 statistical tests in this study (a corrected α of 5×10^{-5}). There was no significant departure from an allele dosage effect ($P = 0.08$). The association was similar in an analysis limited to cases with a first-degree family history of breast cancer ($n = 1230$ cases, 14.4%; OR = 1.06; 95% CI, 0.97–1.17), earlier onset cases (< 55 years) and earlier onset cases with a family history (data not shown).

Combinations of common and/or rare coding variants in candidate DNA repair pathway genes have been hypothesized to influence breast cancer susceptibility (21). In an analysis of coding variation in these 60 genes, we tested 211 variants that were seen at least once in one of the five MEC populations, with 25 (12%) being unique to only one population. Seventy-seven SNPs had a minor allele frequencies (MAF) $> 1\%$ in the combined MEC sample and 34 SNPs had a MAF $> 1\%$ in all populations (93 in AAs, 49 in JAs, 58 in NHs, 67 in LAs and 75 in WHs). We observed no significant associations between combinations of variants and breast cancer risk when summing the number of variants within or across pathways. As shown in Table 4, the results did not change when variants were

grouped based on their predicted functional status from SIFT (22) or PolyPhen (23) ('Damaging' or 'Intolerant'; Supplementary Material, Table S7) or MAF. We also did not observe a significant cumulative effect for variants in *BRCA1*, *BRCA2*, *ATM*, *TP53* and *CHEK2*, which had been previously reported in a study of women with two primary breast cancers (21). The results did not change when limiting the analysis to cases with advanced disease ($n = 549$) or those with a first-degree family history of breast cancer (360 cases; data not shown).

DISCUSSION

To our knowledge, this is the most comprehensive evaluation of common and coding variation in these 60 candidate DNA-damage repair and response pathway genes in relation with breast cancer risk in multiple ethnic populations. This study was specifically designed to have good power to identify common (i.e. 'pan-ethnic') alleles that contribute to breast cancer risk. We had 80% power to detect effects as low as 1.4 for an allele with a MAF of 10% [assuming it is tagged with an r^2 of 0.95 (the average mean maximal r^2), a type 1 error rate of 10^{-5} and a log-additive effect on risk]. The power to detect similar associations for less common alleles with MAFs of 5% was lower ($\sim 48\%$).

The multiethnic design of this study is valuable in identifying a wider range of risk alleles than are studies focusing on a single ethnic group. Although variants which are only common in a single group are likely to be missed in this study, this is compensated for by increased power to detect variants common in several but not all ethnic groups and, arguably more importantly, by the ability to utilize differences in LD structure between ethnic groups to aid in the localization of causal alleles (13). The generally weaker LD seen in individuals of African ancestry provides greater resolution in localizing causal alleles than do studies just using a single non-African sample, but also of importance are the differences between all ethnic groups in the set of SNPs predicted by each tag SNP. In this study, we developed a prediction-based association approach which consists of testing predictors of all SNPs seen in the multi-ethnic SNP characterization panels by using single and multi-marker tests appropriately

Table 2. The most significant SNP associations with breast cancer risk in the MEC

Gene, marker (SNP) ^a	AA (426/453)		JA (554/565)		Population (no. of cases/no. of controls)		EA (532/568)		All groups (2074/2297) ^b		P/P ^c _{Het}
	L/A	L/A (420/419)	L/A	L/A (420/419)	NH	NH (142/292)	EA (532/568)	EA (532/568)	All groups (2074/2297) ^b	All groups (2074/2297) ^b	
<i>BLM</i> , rs8037430 (C/T)	0.93 (0.76–1.12), 48%	0.96 (0.78–1.19), 22%	0.98 (0.76–1.27), 19%	0.93 (0.63–1.37), 20%	0.70 (0.58–0.85), 32%	0.87 (0.79–0.96)	0.0063/0.12				
<i>BLM</i> (L868P), rs11852361 (C/T)	0.89 (0.59–1.35), 5%	<, <1%	0.83 (0.53–1.30), 5%	1.91 (0.65–5.66), 1%	0.50 (0.34–0.72), 8%	0.73 (0.58–0.92)	0.0069/0.053				
<i>CHEK1</i> , rs12276635 (C/T)	0.80 (0.63–1.01), 22%	0.79 (0.34–1.81), 1%	0.76 (0.53–1.09), 8%	1.33 (0.56–3.17), 3%	0.78 (0.56–1.09), 7%	0.80 (0.68–0.94)	0.0063/0.93				
<i>CHEK1</i> , rs3731459 (C/T)	0.69 (0.51–0.94), 12%	<, <1%	0.73 (0.48–1.11), 7%	0.82 (0.25–2.65), 2%	0.72 (0.59–0.88), 7%	0.72 (0.59–0.88)	0.0012/0.90				
<i>FANCA</i> , rs1061646 (A/G)	1.09 (0.91–1.31), 45%	0.83 (0.65–1.07), 15%	1.17 (0.97–1.42), 42%	1.30 (0.98–1.72), 60%	1.30 (1.08–1.56), 68%	1.14 (1.04–1.25)	0.0073/0.058				
<i>FANCA</i> , rs1800330 (C/T)	1.07 (0.89–1.30), 45%	0.82 (0.63–1.06), 13%	1.22 (1.01–1.48), 41%	1.32 (1.00–1.76), 61%	1.27 (1.06–1.53), 67%	1.14 (1.04–1.25)	0.0052/0.043				
<i>FANCA</i> , rs4713858 (G/A)	0.89 (0.65–1.22), 12%	1.12 (0.91–1.37), 22%	0.71 (0.53–0.94), 17%	0.74 (0.53–1.04), 34%	0.60 (0.47–0.75), 24%	0.81 (0.72–0.91)	0.00050/0.0014				
<i>MSH2</i> , rs10179950 (T/C)	0.81 (0.65–1.00), 29%	0.99 (0.82–1.20), 26%	0.79 (0.65–0.97), 61%	1.18 (0.87–1.59), 36%	0.85 (0.71–1.01), 63%	0.89 (0.81–0.97)	0.0085/0.17				
<i>MSH2</i> , rs6982453 (C/T)	0.72 (0.59–0.87), 55%	0.84 (0.68–1.04), 23%	1.00 (0.83–1.22), 47%	0.85 (0.64–1.13), 40%	0.90 (0.76–1.06), 53%	0.86 (0.79–0.94)	0.0012/0.19				
<i>RAD51</i> , rs2304579 (A/G)	1.15 (0.86–1.55), 10%	1.41 (1.07–1.85), 9%	1.10 (0.78–1.55), 8%	1.25 (0.83–1.91), 13%	1.20 (0.87–1.65), 7%	1.23 (1.07–1.42)	0.0041/0.83				
<i>RP42</i> , rs3766396 (A/G)	0.75 (0.61–0.91), 59%	0.94 (0.79–1.12), 44%	0.70 (0.56–0.86), 40%	0.75 (0.52–1.08), 55%	1.07 (0.88–1.30), 32%	0.85 (0.78–0.94)	0.0010/0.020				
<i>RP42</i> , rs4313418 (A/C)	0.82 (0.68–1.00), 38%	0.97 (0.82–1.15), 59%	0.83 (0.68–1.01), 37%	0.85 (0.63–1.13), 60%	0.85 (0.70–1.03), 31%	0.87 (0.80–0.96)	0.0030/0.74				
<i>XRCC4</i> , rs10050830 (T/G)	0.74 (0.61–0.90), 48%	0.91 (0.73–1.12), 19%	0.97 (0.73–1.29), 13%	0.74 (0.42–1.30), 9%	0.70 (0.45–1.08), 5%	0.82 (0.73–0.92)	0.0010/0.39				
<i>XRCC4</i> , rs4703950 (C/A)	0.76 (0.63–0.92), 50%	0.95 (0.80–1.14), 67%	0.92 (0.75–1.14), 32%	1.04 (0.77–1.39), 56%	0.79 (0.60–1.01), 13%	0.88 (0.79–0.96)	0.0064/0.27				
<i>XRCC4</i> , rs10057194 (A/G)	0.79 (0.65–0.96), 40%	0.85 (0.68–1.07), 17%	0.87 (0.65–1.16), 14%	0.72 (0.41–1.30), 9%	0.81 (0.57–1.15), 7%	0.82 (0.73–0.92)	0.00090/0.96				

Each square gives odds ratios (and 95% confidence intervals) for allele dosage effects along with the minor allele frequency in controls (minor allele is based on all populations combined); AA, African Americans; JA, Japanese Americans; L/A, Latinos; NH, Native Hawaiians; EA, European Americans.

^aAlleles based on forward strand of UCSC Genome Browser (<http://genome.ucsc.edu>); (major allele/minor allele).

^bPooled OR adjusted for age and ethnicity.

^cP-value testing for heterogeneity of allelic effects across populations.

optimized for predicting untyped SNPs in each respective population. Also of importance is our extensive multi-ethnic replication approach. Since the replication studies recruited for these tests are actually considerably greater in size than our original study (6483 total cases versus 2074 total cases in the MEC) and are similarly diverse ethnically, we have extremely good control of type 1 error over the entire study, as well as excellent power to detect risk alleles, especially those which affect more than one ethnic group. We have relied for testing purposes on a log-additive risk model relating the tag SNPs and multi-marker predictors to breast cancer risk. This approach is quite effective at detecting risk alleles with dominant as well as log-additive effects. Even for alleles with recessive effects, the log linear model is useful so long as the recessive alleles are quite common. Rarer recessive alleles are much harder to detect and require sample sizes beyond the scope of this study.

The FA family of proteins participates in homologous recombination repair of DSBs, and mutations in these genes [e.g. *BRCA2* (*FANCD1*), *BRIP1* and *PALB2*] have been associated with increased susceptibility to familial breast cancer (24–27). FANCA is part of a multi-subunit nuclear complex of FA proteins that acts to repair blocks in DNA replication caused by cross-linking (26). SNP rs1061646 is located in intron 42 of the *FANCA* gene at 16q24.3. This region overlaps with the 5'-UTR of *ZNF276* (28) and spans a >175 kb region of strong LD, including *FANCA*, *ZNF276* and other genes. Based on the established role of FA genes in breast cancer susceptibility, the association that we observed is highly plausible, with the chance of observing statistically significant false-positive associations in three consecutive studies being low (~0.000125%). However, the lack of a significant association in the largest replication study from the UK weakens support for the hypothesis that this variant is marking an important susceptibility allele for breast cancer at this locus. It is possible, albeit unlikely, that differences in environmental or ancestral genetic background between the UK and US populations may (partly) explain the different findings.

Previous studies have suggested that combinations of rare coding variants in one or more candidate genes could have cumulative effects on cancer risk (21,29). Johnson *et al.* (21) reported missense variants in the genes *BRCA1*, *BRCA2*, *ATM*, *CHEK2* and *ATM* to be significantly associated with breast cancer risk among cases with bilateral disease ($P = 0.005$), particularly for less common alleles ($MAF < 10\%$; $P = 0.00004$). This hypothesis was not supported however by the findings from the current study in older women. Although we conducted a highly inclusive assessment of coding variants in these candidate genes, there are a number of caveats when interpreting our findings. First, we were unable to interrogate all known coding variants in these genes for technical reasons (assay design or genotyping failure), and there are likely even rarer variants in these genes which are population- or subject-specific that will only be identified through direct resequencing. Second, in this study, we did not enrich for younger cases with a family history of breast cancer or bilateral disease who may be more genetically susceptible and for whom a multiple-variant genetic model may be more probable (30). Lastly, the

Table 3. Association of rs1061646 with breast cancer risk in the replication studies

Replication studies	Ethnicity	Cases/controls	<i>FANCA</i> rs1061646 OR (95% CI) ^a	<i>P</i> -value
LAABS	CH	263/376	1.14 (0.90–1.46), 29%	0.27
LAABS	JA	176/314	1.07 (0.74–1.56), 14%	0.72
LAABS	FL	304/297	1.21 (0.96–1.53), 45%	0.10
Pooled ^b		743/987	1.17 (1.01–1.36)	0.043
NHS ^c	EA	1269/1761	1.14 (1.01–1.28), 70%	0.032
SEARCH ^d	EA	4470/4560	1.02 (0.96–1.09), 70%	0.57
Pooled—EA studies ^e		5739/6321	1.05 (0.99–1.11)	0.13
Pooled—all replication studies ^f		6483/7309	1.06 (1.01–1.12)	0.032
Heterozygotes			1.06 (0.98–1.14)	0.17
Homozygotes			1.12 (1.01–1.14)	0.036
All studies (including the MEC) ^f		8556/9605	1.08 (1.03–1.13)	1.4 × 10 ⁻³
Heterozygotes			1.03 (0.96–1.10)	0.42
Homozygotes			1.18 (1.08–1.30)	4.4 × 10 ⁻⁴

CH, Chinese Americans; JA, Japanese Americans; FL, Filipino Americans; EA, European Americans or European Ancestry.

^aOdds ratios (and 95% confidence intervals) for allele dosage effects along with the risk allele frequency in controls.

^bPooled OR adjusted for age and ethnicity.

^cOR adjusted for the matching variables: date of blood draw, age at blood draw and fasting status at blood draw.

^dOR adjusted for study set (1 or 2).

^eORs adjusted for study.

^fORs adjusted for study and ethnicity.

Table 4. Associations of coding variants with breast cancer risk in the MEC populations

Repair Pathway ^b	No. of Genes	All coding SNPs		Predicted to be damaging and/or intolerant ^a		Minor allele frequency ≤10%		Damaging and/or intolerant ^a , and ≤10%	
		No. of SNPs	OR (95% CI) ^c	No. of SNPs	OR (95% CI) ^c	No. of SNPs	OR (95% CI) ^c	No. of SNPs	OR (95% CI) ^c
BER	7	19	1.02 (0.98–1.06)	7	1.02 (0.92–1.12)	14	1.02 (0.93–1.13)	6	0.95 (0.82–1.10)
DSB-HR	7	9	0.98 (0.90–1.06)	4	0.95 (0.73–1.24)	8	0.92 (0.78–1.08)	4	0.95 (0.73–1.24)
DSB-HR ^d	9	33	0.98 (0.94–1.02)	18	0.94 (0.87–1.01)	29	0.96 (0.89–1.01)	17	0.90 (0.80–1.01)
DSB-NHEJ	6	24	1.01 (0.93–1.09)	13	1.07 (0.92–1.23)	23	1.03 (0.94–1.13)	13	1.07 (0.92–1.23)
FA	6	20	1.03 (0.98–1.08)	14	1.00 (0.93–1.07)	17	1.02 (0.93–1.11)	13	1.00 (0.92–1.10)
HEL	1	4	0.91 (0.78–1.06)	3	0.77 (0.62–0.96)	4	0.91 (0.78–1.06)	3	0.77 (0.62–0.96)
MMR	7	37	0.99 (0.96–1.03)	17	0.92 (0.86–1.00)	30	0.96 (0.88–1.04)	16	0.90 (0.79–1.03)
NER	13	48	1.00 (0.97–1.02)	26	1.00 (0.96–1.04)	40	0.99 (0.92–1.07)	22	0.99 (0.89–1.10)
NUC	1	0	–	0	–	0	–	0	–
POL	6	23	1.02 (0.96–1.10)	10	1.05 (0.88–1.26)	21	1.10 (0.96–1.25)	10	1.05 (0.88–1.26)
REV	1	2	0.92 (0.83–1.02)	0	–	1	0.94 (0.76–1.16)	0	–
SIGN/RESP	5	25	0.94 (0.88–1.00)	15	0.96 (0.85–1.08)	23	0.92 (0.84–1.01)	15	0.96 (0.85–1.08)
Johnson <i>et al.</i> ^e	5	46	0.97 (0.94–1.01)	24	0.95 (0.88–1.01)	43	0.96 (0.90–1.02)	23	0.93 (0.85–1.02)
Total ^f	60	211	1.00 (0.98–1.01)	109	0.99 (0.96–1.02)	181	0.99 (0.96–1.02)	102	0.98 (0.94–1.02)

^aPredictions of non-synonymous (NS) and nonsense SNPs from SIFT and PolyPhen programs (see Materials and Methods).

^bThe genes in each pathway and a complete list of all coding SNPs included in these analyses are provided in Supplementary Material, Table S7.

^cEffect per allele adjusted for age and ethnicity.

^dIncludes 24 NS SNPs from our previous studies of *BRCA1* and *BRCA2* in the MEC (36,37); the analysis is limited to 1686 cases and 2000 controls with genotype data for these SNPs.

^eThe combination of coding variants in genes *BRCA1*, *BRCA2*, *TP53*, *CHEK2* and *ATM* genes, as presented by Johnson *et al.* (21).

^fAnalysis of all coding SNPs in the 60 genes evaluated in the present study (not including *BRCA1* or *BRCA2*).

precision of available computational tools used to predict the effects of coding variants and relating putative activity differences to relative risk of a specific phenotype is not clear, and may have resulted in the misclassification and/or the removal of biologically important SNPs from the analysis. These potential shortcomings notwithstanding, we did not observe significant evidence that combinations of putative functional

common and/or rare coding variants in these genes contribute to breast cancer risk.

Since the design of this study knowledge about the DNA repair pathways has increased and a similar study starting now would undoubtedly include more candidate genes. There is now an even more comprehensive reference panel for tag SNP selection (i.e. HapMap Phase II) than we developed

specifically for this study, although we note that coverage of our selected tag SNPs was excellent (91% percent of common SNPs observed in HapMap CEU were captured with $r^2 \geq 0.8$). Further expansion of the candidate gene approach to DNA damage response and repair pathway genes must now be weighed against whole genome scanning (31–33). Focused candidate gene methods can still achieve higher capture of target genetic variation than use of fixed commercial WGA arrays (however the gap is quickly decreasing). A still more serious limitation of the current generation of WGA studies is their almost universal restriction to participants (cases and controls) of European ancestry (31–33) with consequent limitations on the scope of variation to be interrogated and the ability to localize variation. Moving beyond the interrogation of main effects of common genetic variants in target genes to $G \times E$ and ultimately $G \times G$ interactions will need to be considered in a comprehensive manner, especially for genes that act in response to an environmental stimulus (which applies to many DNA repair genes). An approach that considers interactions only for variants that have been shown to have significant main effects has been argued to be suboptimal (34). It is possible that some effects may not have been observed in the MEC or replicated in subsequent studies due to genetic or environmental modifiers.

In conclusion, we found very little evidence for the hypothesis that common alleles in DNA repair pathway genes contribute to breast cancer risk in the general population. We failed to confirm the finding of Johnson *et al.* (21) that the total number of putative deleterious (as predicted *in silico*) coding SNPs in DNA repair pathway genes predict breast cancer risk. The association with SNP rs1061646 in *FANCA* should be considered for replication efforts in other larger studies to better define its role in breast cancer.

MATERIALS AND METHODS

Study population: the Multiethnic Cohort Study

The initial characterization of LD patterns and tag SNP selection and testing was conducted in the Multiethnic Cohort Study (MEC). The MEC consists of over 215 000 men and women in Hawaii and Los Angeles (with additional African-Americans from elsewhere in California) and has been described in detail elsewhere (16). The cohort is comprised predominantly of African Americans (AA), Native Hawaiians (NH), Japanese Americans (JA), Latinos (LA) and European Americans (EA) who entered the study between 1993 and 1996 by completing a 26-page self-administered questionnaire that asked detailed information about dietary habits, demographic factors, personal behaviors, history of prior medical conditions, family history of common cancers, and for women, reproductive history and exogenous hormone use. The participants were between the ages 45 and 75 at enrollment. Incident cancers in the MEC are identified by cohort linkage to population-based cancer Surveillance, Epidemiology and End Results (SEER) registries covering Hawaii and Los Angeles County, and to the California State cancer registry covering all of California. Information on stage of disease at the time of diagnosis is also collected from the cancer registries; women were classified as having

advanced breast cancer when there was evidence of dissemination beyond the breast at diagnosis (SEER stage ≥ 2).

Beginning in 1994, blood samples were collected from incident breast cancer cases and a random sample of MEC participants to serve as a control pool for genetic analyses in the cohort. Controls were frequency matched to cases based on race/ethnicity and age (in 5-year intervals). The breast cancer case–control study in the MEC consists of 2093 invasive breast cancer cases and 2303 controls. The mean ages of the cases and control were 65.0 and 64.2, respectively. This study was approved by the Institutional Review Boards at the University of Southern California and at the University of Hawaii.

Replication studies

Los Angeles County Asian-American Breast Cancer Case–Control Study (LAABC). This study includes 743 Chinese American (CH), Japanese American (JA) and Filipino American (FL) cases between the ages of 25 and 74 years at the time of diagnosis (on or after January 1995 through December 2001) identified through the Los Angeles County Cancer Surveillance Program (CSP) and the statewide California Cancer Registry, both SEER registries. Controls in this study ($n = 987$) were selected from the neighborhoods where cancer cases resided at the time of diagnosis using a well-established, standard algorithm to identify neighborhood controls (19).

Breast Cancer Case-Control Study in the Nurses' Health Study. The Nurses' Health Study nested breast cancer case–control study (1270 cases and 1762 controls) is derived from 32 826 women who provided a blood sample in 1989 and 1990, and were followed for incident disease until May 31, 2000 (8). Medical records were used to confirm the diagnoses in women who reported a diagnosis of breast cancer on biennial questionnaires. Control subjects were matched to cases based on age, menopausal status, recent hormone replacement therapy and blood-draw specific variables (such as date and time of day).

The SEARCH Study. The SEARCH study is an ongoing population-based study of breast cancer, with cases ascertained through the Eastern Cancer Registry (formerly East Anglian CR) (20). All patients diagnosed with invasive breast cancer <55 years since 1991 and still alive in 1996 (prevalent cases, median age 48 years; Set 1), together with all those diagnosed <70 years between 1996 and the present (incident cases, median age 54 years; Set 2), were eligible to take part. Sixty-seven percent of eligible breast cancer patients returned a questionnaire and 64% provided a blood sample for DNA analysis. The total number of cases available for analysis in this study was 4474, of which 27% were prevalent cases. Controls were randomly selected from the Norfolk component of EPIC (European Prospective Investigation of Cancer) (35) and are broadly similar in age to the cases (aged 42–81 years). This study has been approved by the Eastern Region Multicentre Research Ethics Committee, and all participants gave written informed consent.

Characterization of linkage disequilibrium patterns in DNA repair genes

Our investigation of genetic variation focused on common alleles with MAF $\geq 5\%$ in at least one of the five racial/ethnic populations in the MEC. Since the design of the experiment was initiated prior to the availability of the complete HapMap database, we selected our tag SNPs based on the following set of criteria. We initially surveyed variation across coding and non-coding regions of 59 candidate genes by genotyping a high density of SNPs selected from the public SNP map (dbSNP; www.ncbi.nlm.nih.gov/SNP). We have previously studied *BRCA1* and *BRCA2* using similar methods in the MEC (36,37). In an attempt to cover variation in putative regulatory regions, we considered SNPs 20 kb upstream through 10 kb downstream of each gene. To increase our chances of selecting common SNPs that convert to working assays, we selected SNPs preferentially based on validation status ('two-hit SNPs') reported in dbSNP and the assay design score (≥ 0.6) provided by Illumina (Illumina Inc., San Diego, CA, USA). A total of 2897 SNPs were selected for genotyping, of which 128 were coding variants (non-synonymous or nonsense SNPs) with reported MAF $\geq 5\%$ from dbSNP or from the NIEHS Environmental Genome Project (GeneSNPs; www.genome.utah.edu/genesnps/) as of August, 2004. These SNPs were evaluated in a multiethnic reference panel comprised of 67–70 each of AA, JA, NH, LA and EA breast cancer controls from the MEC and 20 CEPH (Centre d'Etude du Polymorphisme Humain) trios (60 individuals in total) which are a subset of the 30 trios used in the HapMap Project (CEU samples) (11). Nine quality control (QC) duplicates were included to assess genotyping reproducibility. Of the 2897 SNPs genotyped, 92 were monomorphic in all populations and 2627 (93.7%) passed QC filtering criteria (see Supplementary Material, Tables S1 and S2), of which 105 were coding variants.

Tag SNP selection

From the generated LD panel, we selected a set of cosmopolitan tag SNPs to capture all common SNPs observed in each of the five populations represented in the MEC with a high correlation ($r^2 \geq 0.8$) based on pairwise and multi-marker haplotype-based tests (17). For each gene, we began by including all coding SNPs ($n = 105$) as tag SNPs, followed by the selection of additional tags in each population. In total, 1234 SNPs were selected for the 59 genes. Tag and coding SNPs for the *NBS* gene ($n = 23$) were selected using Phase I HapMap data (Data Rel#16, Mar05) for the Yoruba (YRI), Japanese (JPT) and CEU populations. We also selected an additional 58 tags in the CEU population using HapMap data (Data Rel#16/phase I Mar05) that captured common alleles not adequately predicted ($r^2 < 0.8$) by the SNPs examined in this study. Public SNP databases (dbSNP and GeneSNPs) were resurveyed prior to SNP genotyping in the MEC breast cancer case–control samples, at which time (July, 2005) >400 coding variants had been reported (~ 200 with MAFs $\geq 1\%$). An additional 148 coding variants with MAFs as low as 1% (as reported in the databases) were selected, bringing the total number of SNPs to be tested in

the MEC case–control samples to 1463 (Supplementary Material, Table S3).

SNP genotyping

Genotyping in the LD characterization phase and in the MEC case–control samples was conducted using the GoldenGate assay and Illumina BeadArray™ technology in the USC Genomics Center. Genotyping in the replication studies was performed using the TaqMan allelic discrimination assay at the University of Southern California (LAABC), Harvard (NHS) and Cambridge (SEARCH). Primers, probes and assay conditions are available upon request.

Genotyping quality control

We applied strict criteria to maximize genotyping quality of the 1463 SNPs tested in the 2093 breast cancer cases and 2303 controls in the MEC. We excluded 53 SNPs and 25 subjects with more than 25% missing data across subjects or SNPs, respectively. Thirty CEU trios, used by HapMap, and 53 QC replicates were inserted across the DNA plates to evaluate the genotyping error rate. Of the remaining 1410 SNPs, 6 had ≥ 2 QC errors (based on discordant replicates + Mendelian errors in CEU trios) and 14 were out of HWE ($P < 0.01$ among controls in more than one of the MEC populations) and were excluded. Twenty-three SNPs were monomorphic in all populations. For the remaining 1367 SNPs (which included 211 coding SNPs), the average genotype call rate was 98.86% (range 75.28–100%) among the 4371 subjects (2074 cases and 2297 controls). SNPs selected for replication were first genotyped in ≥ 874 case–control samples from the MEC to assess comparability of genotype calls between TaqMan and Illumina platforms. The genotype concordance rate between platforms was 99.4%. Blinded duplicate samples (5–10%) were also included in the replication studies and concordance of these samples was $\geq 99\%$ in all studies.

Statistical analysis

To assess the coverage of the tag SNPs, we computed the mean maximum r^2 between all common SNPs genotyped in the LD characterization panel and the selected tag SNPs for each gene and racial/ethnic population, as well as for the CEU population using Phase II data from HapMap (11). For the majority of SNPs evaluated in the LD analysis but not genotyped directly in the case–control study, single tag SNPs served as good pairwise proxies. In addition, we identified multi-marker haplotype tests to predict untyped SNPs (in case these were not adequately captured by pairwise LD). Thus, in our analysis of breast cancer risk, we tested each tag SNP and coding SNP and the selected combinations of 2–3 SNP haplotypes that were identified as predictors for specific SNPs. Because LD patterns may be different between populations, any one tag SNP (or multi-marker haplotype test) can potentially predict a different set of (untyped) SNPs in different populations. Therefore, in each population, we defined specific multi-marker tests based on the LD characterization panel of that population.

Allele dosage effects for tag SNPs and coding SNPs were tested using unconditional logistic regression in a series of age and ethnicity-adjusted analyses. For the purpose of testing SNPs predicted by multi-allelic haplotypes, we utilized a procedure which is equivalent to the haplotype-specific risk testing method of Zaykin *et al.* (38). For each untyped SNP (predicted by a multi-marker predictor), we used the software tagSNPs (39) to form for each individual in the case-control data the best multi-allelic predictor of the number of copies (0, 1 or 2) of that SNP carried given the tag SNP genotypes for that individual [using an EM algorithm (40) and based on haplotype frequencies estimated from the LD characterization panel]. These predicted SNPs were then used in the same logistic regression analysis as were the single SNPs, to compute odds ratios and confidence intervals assuming a log-linear model. Tests of heterogeneity of effects across racial/ethnic groups in the MEC were performed by a likelihood ratio analysis following the inclusion of an interaction term between SNP or haplotype, and ethnicity in the multivariate model.

Population stratification was evaluated using principal components (18). We estimated the relative degree of relatedness between all pairs of individuals in the study using all SNPs for a given subject, computed an adjusted covariance between all SNPs genotyped for each pair of subjects and then computed the eigenvectors (principal components) of the resulting matrix of all pairs of individuals. The first 10 eigenvectors associated with the largest eigenvalues were included as adjustment variables in the logistic regression analysis.

SNPs found to be nominally associated with risk in the ethnic-pooled analyses in the MEC with a P -value ≤ 0.01 and present in more than one racial/ethnic population at a frequency $\geq 1\%$ were selected for replication. SNPs that were common (MAF $> 5\%$) in the MEC Japanese population were first followed up in the Los Angeles County Asian-American Breast Cancer (LAABC) study (19). The remaining SNPs were followed up among European American samples from the Nurses' Health Study (8). Associations in either study ($P < 0.05$) were followed up in the SEARCH study (20). Unconditional logistic regression was used, controlling for age and Asian ancestry (LAABC) to estimate study-specific ORs as well as pooled ORs across the replication studies and across all four studies. We used the Statistical Analysis System (SAS Institute Inc., Cary, NC, Version 9.1) and PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/>) for all analyses.

Analysis of coding SNPs

For the 211 coding variants genotyped in this study, we evaluated their effects on risk by counting the number of alleles for an individual both within and across DNA repair pathways. We grouped the variants in four ways: (i) using all 211 variants examined, (ii) based on the functional nature of the amino acid substitution as predicted by SIFT (22) and PolyPhen (23) (see below), (iii) focusing on less common alleles with MAFs $\leq 10\%$ (based on the minor allele frequency among all MEC populations combined) and (iv) the intersection of (ii) and (iii). We also summed coding variants in the genes *BRCA1*, *BRCA2*, *ATM*, *CHEK2* and *TP53*, as vari-

ants in these genes had been reported by Johnson *et al.* (21) to confer significant risk. For this analysis, we included 24 non-synonymous variants from our previous studies of *BRCA1* and *BRCA2* in the MEC (36,37), and evaluated all 46 coding variants in these five genes in a smaller sample of 1686 cases and 2000 controls.

Functional predictions for each variant were obtained from pre-computed SIFT (http://blocks.fhrc.org/sift/SIFT_dbSNP.html) and PolyPhen databases (<http://genetics.bwh.harvard.edu/pph/data/index.html>) (provided in Supplementary Material, Table S7). For variants not found in these databases (dbSNP build 126), we submitted their NCBI GI # or FASTA sequence along with amino acid substitution and position into the SIFT (<http://blocks.fhrc.org/sift/SIFT.html>) and PolyPhen (<http://coot.embl.de/PolyPhen/>) protein alignment query programs to ascertain prediction scores. Although SIFT/PolyPhen have explicit prediction score cut-offs, we utilized a more relaxed scheme as implemented in several previous studies (5,21,41). For PolyPhen predictions, variants were categorized as 'Damaging' with a prediction score > 1.25 (which included SNPs with scores for 'Probably Damaging', 'Possibly Damaging', and 'Potentially Damaging'). For SIFT predictions, variants were categorized as 'Intolerant' if the score was < 0.1 (which included SNPs with scores for 'Intolerant' and 'Potentially Intolerant'). Logistic regression models were fitted to estimate odds ratios associated with the number of variant alleles treated as linear variables, adjusted for age and race.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

From the University of Southern California, we thank Loreall Pooler and David Wong for their laboratory assistance and Dr Kristine Monroe, Hank Huang and Chiu-Chen Tseng for their technical support. We also thank Hardeep Ranu of the DF/HCC High Throughput Polymorphism Detection Core for sample handling and genotyping of the NHS samples. We are also indebted to the participants in all of these studies.

Conflict of Interest statement. None declared.

FUNDING

This study was supported in part by the Wright Foundation (award to C.A.H.). The Multiethnic Cohort Study was supported by National Cancer Institute (NCI) grants CA63464 and CA54281. Additional support was provided by grants CA17054, HG002790 and HL084705 (D.O.S). The Los Angeles-based case-control study was supported by NCI grant CA17054 and grants from the California Breast Cancer Research Program (1RB-0287, 3PB-0102, 5PB-0018, 10PB-0098). The Nurses' Health Study was supported by NCI grants CA065725, CA098233, CA049449 and CA118447. The SEARCH study was supported by funding from Cancer-Research UK.

REFERENCES

- Hamosh, A., Scott, A.F., Amberger, J., Valle, D. and McKusick, V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **15**, 57–61.
- Kennedy, D.O., Agrawal, M., Shen, J., Terry, M.B., Zhang, F.F., Senie, R.T., Motykiewicz, G. and Santella, R.M. (2005) DNA repair capacity of lymphoblastoid cell lines from sisters discordant for breast cancer. *J. Natl Cancer Inst.*, **97**, 127–132.
- Roberts, S.A., Spreadborough, A.R., Bulman, B., Barber, J.B., Evans, D.G. and Scott, D. (1999) Heritability of cellular radiosensitivity: a marker of low-penetrance predisposition genes in breast cancer? *Am. J. Hum. Genet.*, **65**, 784–794.
- Walsh, T. and King, M.C. (2007) Ten genes for inherited breast cancer. *Cancer Cell*, **11**, 103–105.
- Xi, T., Jones, I.M. and Mohrenweiser, H.W. (2004) Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*, **83**, 970–979.
- Garcia-Closas, M., Egan, K.M., Newcomb, P.A., Brinton, L.A., Titus-Ernstoff, L., Chanock, S., Welch, R., Lissowska, J., Peplonska, B., Szeszenia-Dabrowska, N. *et al.* (2006) Polymorphisms in DNA double-strand break repair genes and risk of breast cancer: two population-based studies in USA and Poland, and meta-analyses. *Hum. Genet.*, **119**, 376–388.
- Goode, E.L., Ulrich, C.M. and Potter, J.D. (2002) Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 1513–1530.
- Han, J., Hankinson, S.E., Ranu, H., De Vivo, I. and Hunter, D.J. (2004) Polymorphisms in DNA double-strand break repair genes and breast cancer risk in the Nurses' Health Study. *Carcinogenesis*, **25**, 189–195.
- Kuschel, B., Auranen, A., McBride, S., Novik, K.L., Antoniou, A., Lipscombe, J.M., Day, N.E., Easton, D.F., Ponder, B.A., Pharoah, P.D. *et al.* (2002) Variants in DNA double-strand break repair genes and breast cancer susceptibility. *Hum. Mol. Genet.*, **11**, 1399–1407.
- Zhang, Y., Newcomb, P.A., Egan, K.M., Titus-Ernstoff, L., Chanock, S., Welch, R., Brinton, L.A., Lissowska, J., Bardin-Mikolajczak, A., Peplonska, B. *et al.* (2006) Genetic polymorphisms in base-excision repair pathway genes and risk of breast cancer. *Cancer Epidemiol. Biomarkers Prev.*, **15**, 353–358.
- The International FapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
- Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
- Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J. *et al.* (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.*, **39**, 638–644.
- Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
- Wood, R.D., Mitchell, M., Sgouros, J. and Lindahl, T. (2001) Human DNA repair genes. *Science*, **291**, 1284–1289.
- Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M., Wilkens, L.R., Pike, M.C., Stram, D.O., Monroe, K.R., Earle, M.E. and Nagamine, F.S. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.*, **151**, 346–357.
- de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
- Wu, A.H., Yu, M.C., Tseng, C.C. and Pike, M.C. (2007) Body size, hormone therapy and risk of breast cancer in Asian-American women. *Int. J. Cancer*, **120**, 844–852.
- Lesueur, F., Pharoah, P.D., Laing, S., Ahmed, S., Jordan, C., Smith, P.L., Luben, R., Wareham, N.J., Easton, D.F., Dunning, A.M. *et al.* (2005) Allelic association of the human homologue of the mouse modifier Ptpj with breast cancer. *Hum. Mol. Genet.*, **14**, 2349–2356.
- Johnson, N., Fletcher, O., Palles, C., Rudd, M., Webb, E., Sellick, G., dos Santos Silva, I., McCormack, V., Gibson, L., Fraser, A. *et al.* (2007) Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility. *Hum. Mol. Genet.*, **16**, 1051–1057.
- Ng, P.C. and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.
- Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
- Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T. *et al.* (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.*, **39**, 165–167.
- Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K. *et al.* (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.*, **38**, 1239–1241.
- Wang, W. (2007) Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. *Nat. Rev. Genet.*, Epub September 4.
- Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D. *et al.* (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. *Science*, **265**, 2088–2090.
- Wong, J.C., Gokgoz, N., Alon, N., Andrusis, I.L. and Buchwald, M. (2003) Cloning and mutation analysis of ZFP276 as a candidate tumor suppressor in breast cancer. *J. Hum. Genet.*, **48**, 668–671.
- Fearnhead, N.S., Wilding, J.L., Winney, B., Tonks, S., Bartlett, S., Bicknell, D.C., Tomlinson, I.P., Mortensen, N.J. and Bodmer, W.F. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA*, **101**, 15992–15997.
- Pharoah, P.D., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F. and Ponder, B.A. (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.*, **31**, 33–36.
- Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struwing, J.P., Morrison, J., Field, H., Luben, R. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.
- Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.
- Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A. *et al.* (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **39**, 865–869.
- Marchini, J., Donnelly, P. and Cardon, L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.
- Riboli, E. and Kaaks, R. (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.*, **26** (Suppl. 1), S6–S14.
- Freedman, M.L., Penney, K.L., Stram, D.O., Le Marchand, L., Hirschhorn, J.N., Kolonel, L.N., Altshuler, D., Henderson, B.E. and Haiman, C.A. (2004) Common variation in BRCA2 and breast cancer risk: a haplotype-based analysis in the Multiethnic Cohort. *Hum. Mol. Genet.*, **13**, 2431–2441.
- Freedman, M.L., Penney, K.L., Stram, D.O., Riley, S., McKean-Cowdin, R., Le Marchand, L., Altshuler, D. and Haiman, C.A. (2005) A haplotype-based case-control study of BRCA1 and sporadic breast cancer risk. *Cancer Res.*, **65**, 7516–7522.
- Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J. and Ehm, M.G. (2002) Testing association of statistically inferred

- haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, **53**, 79–91.
39. Stram, D.O., Haiman, C.A., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E. and Pike, M.C. (2003) Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.*, **55**, 27–36.
40. Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
41. Nakken, S., Alseth, I. and Rognes, T. (2007) Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. *Neuroscience*, **145**, 1273–1279.