DOI: 10.1093/jnci/djq563
Advance Access publication on January 24, 2011.

ARTICLE

Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study

Olivia Fletcher, Nichola Johnson, Nick Orr, Fay J. Hosking, Lorna J. Gibson, Kate Walker, Diana Zelenika, Ivo Gut, Simon Heath, Claire Palles, Ben Coupland, Peter Broderick, Minouk Schoemaker, Michael Jones, Jill Williamson, Sarah Chilcott-Burns, Katarzyna Tomczyk, Gemma Simpson, Kevin B. Jacobs, Stephen J. Chanock, David J. Hunter, Ian P. Tomlinson, Anthony Swerdlow, Alan Ashworth, Gillian Ross, Isabel dos Santos Silva, Mark Lathrop, Richard S. Houlston, Julian Peto

Manuscript received December 9, 2009; revised December 14, 2010; accepted December 17, 2010.

Correspondence to: Julian Peto, DSc, MSc, Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel St, London, WC1E 7HT, UK (e-mail: julian.peto@lshtm.ac.uk).

Background Genome-wide association studies have identified several common genetic variants associated with breast cancer risk. It is likely, however, that a substantial proportion of such loci have not yet been discovered.

Methods We compared 296 114 tagging single-nucleotide polymorphisms in 1694 breast cancer case subjects (92% with two primary cancers or at least two affected first-degree relatives) and 2365 control subjects, with validation in three independent series totaling 11 880 case subjects and 12 487 control subjects. Odds ratios (ORs) and associated 95% confidence intervals (CIs) in each stage and all stages combined were calculated using unconditional logistic regression. Heterogeneity was evaluated with Cochran Q and I^2 statistics. All statistical tests were two-sided.

Results We identified a novel risk locus for breast cancer at 9q31.2 (rs865686: OR = 0.89, 95% CI = 0.85 to 0.92, $P = 1.75 \times 10^{-10}$). This single-nucleotide polymorphism maps to a gene desert, the nearest genes being Kruppel-like factor 4 (*KLF4*, 636 kb centromeric), RAD23 homolog B (*RAD23B*, 794 kb centromeric), and actin-like 7A (*ACTL7A*, 736 kb telomeric). We also identified two variants (rs3734805 and rs9383938) mapping to 6q25.1 estrogen receptor 1 (*ESR1*), which were associated with breast cancer in subjects of northern European ancestry (rs3734805: OR = 1.19, 95% CI = 1.11 to 1.27, $P = 1.35 \times 10^{-7}$; rs9383938: OR = 1.18, 95% CI = 1.11 to 1.26, $P = 1.41 \times 10^{-7}$). A variant mapping to 10q26.13, approximately 300 kb telomeric to the established risk locus within the second intron of *FGFR2*, was also associated with breast cancer risk, although not at genome-wide statistical significance (rs10510102: OR = 1.12, 95% CI = 1.07 to 1.17, $P = 1.58 \times 10^{-6}$).

Conclusions These findings provide further evidence on the role of genetic variation in the etiology of breast cancer. Fine mapping will be needed to identify causal variants and to determine their functional effects.

J Natl Cancer Inst 2011;103:425–435

Many breast cancers arise in a genetically susceptible minority of women (1), most of whom do not carry mutations in breast cancer 1 (*BRCA*)*1* or *BRCA2* (2). The familial excess in risk not accounted for by *BRCA1* or *BRCA2* is plausibly explained by a polygenic model in which a large number of 'low-penetrance' variants act in combination to cause wide variation in risk in the population (3). Twelve regions containing at least one common susceptibility locus have been previously discovered through genome-wide association (GWA) studies (4–10) (Table 1). It is likely, however, that a high proportion of susceptibility loci have not yet been detected. We conducted a GWA study to identify further susceptibility loci. To enhance the statistical power of our study, we included in our sample a high proportion of case subjects with two primary breast cancers or a family history of the disease (11,12).

Methods

Study Subjects: Stage 1 GWA

Our stage 1 GWA study was based on genotyping 1766 prevalent case subjects taking part in the British Breast Cancer (BBC) Study (13). A total of 1170 case patients were ascertained through the English and Scottish cancer registries. Thirty-four of these 1170 case patients were subsequently excluded as part of post-genotyping quality control (see below for details), leaving 1136 case patients for analysis (Figure 1). Briefly, registry records were used to identify women whose first breast cancer was diagnosed before age 65 years in 1971 or later. Those with two sequential or simultaneous primary breast cancer registrations and an equal number with a single primary breast cancer were invited to participate in the study. Of the 1136 case patients who were included in the final

CONTEXTS AND CAVEATS

Prior knowledge

The familial excess in breast cancer risk not accounted for by known genetic variants such as *BRCA1* or *BRCA2* may be attributable to an unknown number of variants that have not yet been discovered.

Study design

To identify additional breast cancer susceptibility loci, 296 114 tagging single-nucleotide polymorphisms were compared in a genome-wide association study of 1694 breast cancer case patients (92% with two primary cancers or at least two affected first-degree relatives) and 2365 control subjects.

Contribution

A new locus associated with lower risk for breast cancer was identified at chromosomal locus 9q31.2. Two variants mapping to the estrogen receptor 1 region were also found to be statistically significantly associated with increased risk of breast cancer in subjects of northern European ancestry.

Implications

Although genome-wide association studies have identified multiple common genetic variants associated with breast cancer risk, larger studies and combined analyses and studies of non-European populations could identify further low-penetrance variants that may act in combination to cause wide variation in risk.

Limitations

A limitation of the study was lack of statistical power to detect common variants conferring lower relative risks of breast cancer or minor allele frequencies of less than 10%, even if they were associated with substantial effects.

From the Editors

analysis, 1038 (91.4%) had two primary breast cancers, 17 (1.5%) had a family history of breast cancer, and the remaining 81 (7.1%) had a single primary breast cancer and no known family history. Recruitment to this study is ongoing. Case patients included in this analysis were recruited between November 2001 and September 2008.

Five hundred ninety-six case patients diagnosed after 1967 and before age 71 years were recruited between January 2005 and September 2008 through National Cancer Research Network breast cancer clinics. Thirty-eight of these 596 case patients were subsequently excluded as part of post-genotyping quality control (see below for details), leaving 558 for analysis (Figure 1). The 558 who were included in the final analysis comprised 390 (69.9%) with two primary breast cancers, 151 (27.1%) with a family history of breast cancer, and 17 (3.0%) with a single primary breast cancer and no known family history. Recruitment to this study is ongoing. Mean age at diagnosis of case subjects was 49.3 years (SD = 8.8). Control subjects were 927 disease-free individuals from the UK Colorectal Tumor Gene Identification (CORGI) study (14) and 1438 individuals from the British 1958 Birth Cohort (http://www.cls.ioe.ac.uk/studies.asp?section=000100020003) (15,16) (Figure 1). CORGI control subjects were recruited between January 1998 and December 2005. Mean age at blood draw was 57.8 years (SD = 12.3). All case and control subjects were British residents.

Replication Series: Cancer Genetic Markers of Susceptibility Stage 1

To select single-nucleotide polymorphisms (SNPs) for replication in follow-up studies, we combined our stage 1 data with publicly available data from stage 1 of the Cancer Genetic Markers of Susceptibility (CGEMS) study (5) (Figure 1). Full details of the CGEMS stage 1 GWA have been reported previously (http://cgems.cancer.gov/) (5). Briefly, 1145 incident breast cancer case subjects and 1142 control subjects from the Nurses' Health Study were genotyped using Illumina HumanHap 550K arrays (Illumina Inc, San Diego, CA). Case subjects were a consecutive series of postmenopausal women selected among the 32 826 members of the Nurses' Health Study who gave a blood sample in 1989–1990 and had not been previously diagnosed with breast cancer but who were subsequently diagnosed before June 1, 2004. Control subjects were postmenopausal women who were matched to case patients by year of birth and postmenopausal hormone use at blood draw and who were not diagnosed with breast cancer during follow-up. Mean age at diagnosis of case subjects was 65.6 years (SD = 6.7), and mean age at blood draw in control subjects was 58.4 years (SD = 6.4). All case and control subjects reported being of European ancestry.

Stage 2

Stage 2 of this study comprised 4829 prevalent breast cancer case subjects ascertained through the National Cancer Research Network breast cancer clinics (n = 3105) as part of the BBC study or through the Royal Marsden Hospital (RMH) (n = 1724). Twenty-five BBC case subjects were subsequently excluded as part of post-genotyping quality control (see below for details), leaving 3080 for analysis. None of the RMH case subjects were excluded (Figure 1). A majority of BBC case subjects were selected for a genetic predisposition to breast cancer; 436 (14.12%) had had two primary breast cancers and 2542 (82.5%) had at least one first-degree relative affected with breast or ovarian cancer. The remaining 102 (3.3%) were case patients with a single primary breast cancer and no known family history of the disease. RMH case subjects were consecutive case patients unselected for any other characteristics recruited from May 2000 to January 2007. Mean age at diagnosis was 52.2 years (SD = 9.3) for BBC case subjects and 55.1 years (SD = 11.4) for RMH case subjects. Control subjects were friends and nonblood relatives of breast cancer case subjects ascertained through the case subjects (n = 2906) and additional healthy women who were participating in a randomized trial of mammographic screening at younger ages (17) (n = 1046). Four control subjects who were recruited through the BBC study and 12 of the control subjects from the Mammography Oestrogens and Growth Factors study were subsequently excluded during post-genotyping quality control (see below for details). This left a total of 3936 stage 2 control subjects for analysis (Figure 1). BBC control subjects were recruited between January 2002 and September 2008, and mean age at blood draw was 47.2 years (SD = 11.7). Control subjects in the Mammography Oestrogens and Growth Factors study were recruited between January 2001 and January

Table 1. Results for single-nucleotide polymorphisms (SNPs) reported in previous genome-wide association studies*

SNP, allele† (MAF)	Cytoband (gene)	Stage 1‡	CGEMS 1§	Stage 2‖			Overall¶			
		P_{trend}	P_{trend}	P_{trend}	OR_{het} (95% CI)	OR_{hom} (95% CI)	OR_{all} (95% CI)	P_{trend}	P_{het} (I^2)	P_{fit}
rs1338704#, A,G (0.48)	2q35	2.04×10^{-3}	2.99×10^{-3}	2.02×10^{-6}	0.83 (0.77 to 0.90)	0.75 (0.68 to 0.82)	0.86 (0.82 to 0.90)	1.83×10^{-10}	.87 (0%)	.20
rs4973768, C,T (0.49)	3p24.1 (*SLC4A7*)	7.33×10^{-4}	1.29×10^{-1}	2.03×10^{-5}	1.17 (1.08 to 1.26)	1.29 (1.18 to 1.42)	1.14 (1.09 to 1.19)	2.34×10^{-8}	.71 (0%)	.38
rs4415084, C,T (0.42)	5p12	3.96×10^{-4}	4.72×10^{-3}	2.72×10^{-6}	1.23 (1.14 to 1.32)	1.33 (1.21 to 1.47)	1.17 (1.11 to 1.22)	7.62×10^{-11}	.91 (0%)	.07
rs6900157,# T,C (0.34)	6q25.1	3.22×10^{-3}	1.08×10^{-1}	—	1.14 (1.03 to 1.27)	1.28 (1.09 to 1.50)	1.13 (1.05 to 1.22)	1.01×10^{-3}	.65 (0%)	.76
rs1562430,# A,G (0.40)	8q24.21	6.22×10^{-3}	1.19×10^{-2}	2.64×10^{-8}	0.84 (0.78 to 0.90)	0.74 (0.67 to 0.81)	0.86 (0.82 to 0.90)	3.15×10^{-11}	.71 (0%)	.59
rs1219648, A,G (0.42)	10q26.13 (*FGFR2*)	8.31×10^{-8}	1.74×10^{-6}	2.29×10^{-19}	1.31 (1.24 to 1.39)	1.72 (1.60 to 1.86)	1.31 (1.25 to 1.37)	1.41×10^{-30}	.80 (0%)	.84
rs999737, C,T (0.24)	14q24.1	1.27×10^{-1}	1.28×10^{-2}	—	0.94 (0.85 to 1.05)	0.69 (0.55 to 0.87)	0.89 (0.82 to 0.97)	6.57×10^{-3}	.29 (11.2%)	.07
rs4783780,# A,C (0.49)	16q12.1 (*TOX3*)	2.96×10^{-5}	6.80×10^{-1}	2.48×10^{-7}	1.14 (1.06 to 1.23)	1.34 (1.22 to 1.46)	1.16 (1.10 to 1.21)	4.42×10^{-10}	.07 (62.4%)	.71
rs3112612,# C,T (0.43)	16q12.1 (*TOX3*)	1.44×10^{-3}	4.08×10^{-1}	6.72×10^{-8}	1.18 (1.10 to 1.27)	1.31 (1.19 to 1.44)	1.15 (1.10 to 1.21)	3.96×10^{-10}	.21 (36.9%)	.37

* SNPs at four previously reported loci were not tagged or were not successfully genotyped in stage 1: rs11249433 (1p11.2), rs889312 (5q11.2), rs3817198 (11p15.5), and rs6504950 (17q23.2) (4,8,9). Odds ratios (ORs) and associated 95% confidence intervals (CIs) in each stage and all stages combined were calculated using unconditional logistic regression. All statistical tests were two-sided. CGEMS = Cancer Genetic Markers of Susceptibility study; *FGFR2* = fibroblast growth factor receptor 2; I^2 = percentage of variance that is between stages; MAF = minor allele frequency; OR_{all} = allelic odds ratio; OR_{het} = heterozygote odds ratio; OR_{hom} = homozygote odds ratio; P_{fit} = P for departure from multiplicative model; P_{het} = P for Cochran Q test for heterogeneity; P_{trend} = P for null hypothesis of allelic odds ratio = 1.0; *SLC4A7* = solute carrier family 4, sodium bicarbonate transporter, member 7; *TOX3* = TOX high-mobility group box family member 3.

† In order major, minor.

‡ N = 1694 case patients, 2365 control subjects.

§ N = 1145 case patients, 1142 control subjects.

‖ N = 4804 case patients, 3936 control subjects.

¶ N = 7643 case patients, 7443 control subjects.

rs6900157 was selected as a proxy for rs2046210 (r^2 = 0.92, D' = 1.00) (10), rs1562430 as a proxy for rs13281615 (r^2 = 0.42, D' = 0.95) (4), rs4783780 and rs3112612 as proxies for rs12443621 (r^2 = 0.97, D' = 1.0 and r^2 = 0.53, D' = 0.91, respectively) (4).

2005 and were aged between 44 and 55 years at blood draw (mean age = 47.2 years, SD = 1.2). All case and control subjects were British residents, and all control subjects were free from breast cancer at the time of recruitment.

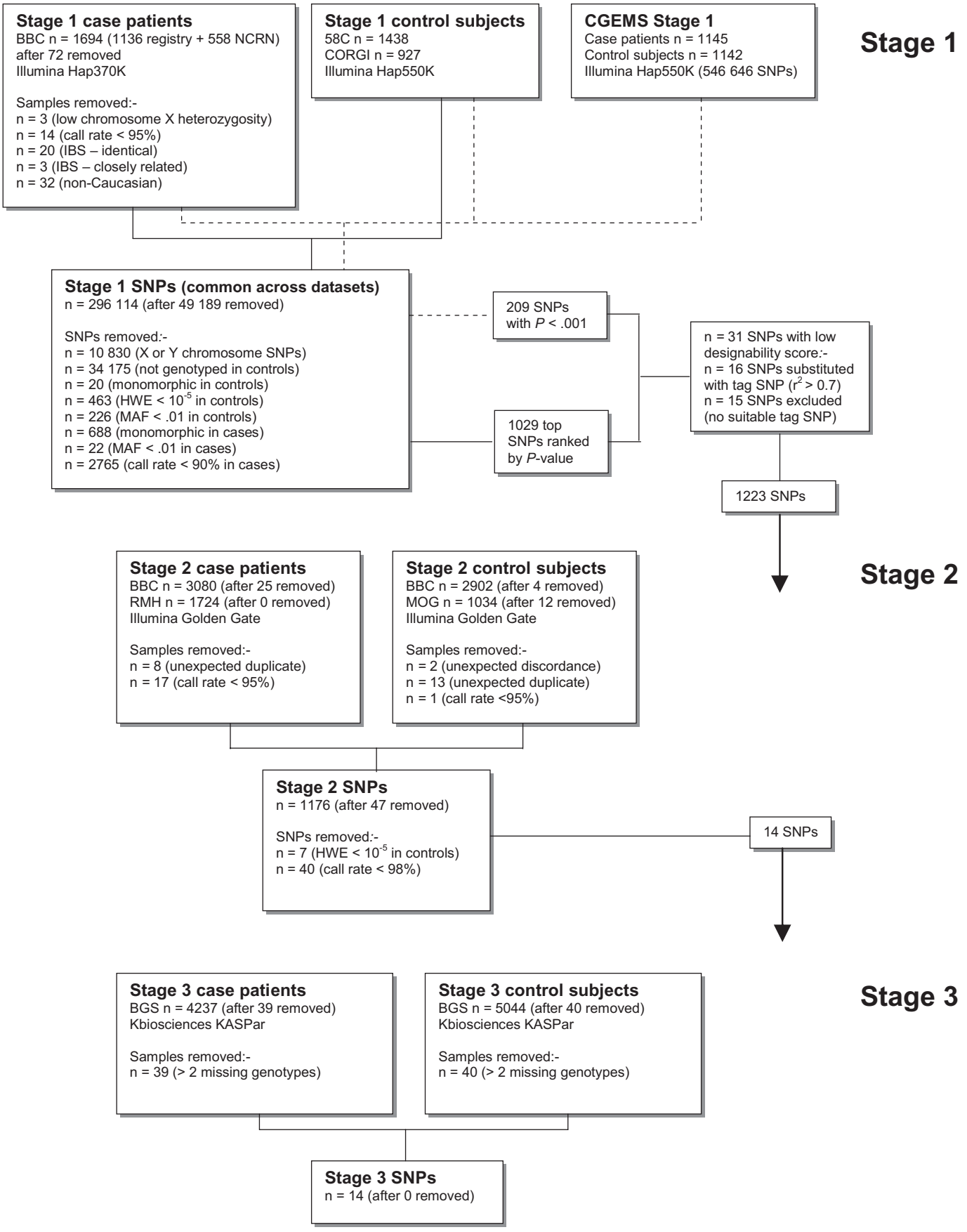
Stage 3

Stage 3 comprised 4276 case subjects and 5084 control subjects from the Breakthrough Generations Study (BGS). Thirty nine case subjects and 40 control subjects were subsequently excluded as part of post-genotyping quality control (see below for details), leaving 4237 stage 3 case subjects and 5044 stage 3 control subjects for inclusion in the final analysis (Figure 1). BGS (http://www.breakthroughgenerations.org.uk/) is a cohort study of more than 100 000 women from the UK general population from whom detailed questionnaires and blood samples have been collected to investigate risk factors for breast cancer. Participants included in this analysis were recruited between January 2003 and October 2008. The 100 000 study participants included prevalent breast cancer case subjects diagnosed before entry into the study. Breast cancer diagnosis was self-reported. Mean age at diagnosis was 50.9 years (SD = 9.3). Control subjects were frequency matched to case patients on year of entry to the study, source of recruitment (supporters, self-referral, nomination by other participants), and date of birth. Mean age at blood draw was 57.9 years (SD = 9.3). All case and control subjects included in the analysis were British residents, and all control subjects were free from breast cancer at the time of recruitment.

Collection of blood samples and questionnaire information from case and control subjects was undertaken with informed consent and relevant ethical review board approval (BBC and BGS, South East Multi-centre Research Ethics Committee; CORGI, Southampton and South West Hampshire Research Ethics Committee; RMH, Royal Marsden National Health Service Foundation Trust Research Ethics Committee) in accordance with the tenets of the Declaration of Helsinki.

Genotyping

DNA was extracted in-house (QIAamp DNA Blood mini kits; Qiagen, Crawley West Sussex, UK) and commercially with a proprietary chloroform extraction, ethanol precipitation-based method (Tepnel Life Sciences, Manchester, UK) using conventional methodologies, and quantified using PicoGreen (Invitrogen, Carlsbad, CA). Stage 1 genotyping was carried out at the Centre National de Génotypage, Cedex, France (http://www.cng.fr/en/welcome/index.html). Barcoded DNA samples were received in standard tubes together with sample information and were subjected to stringent quality control. Concentration, fragmentation, and response to polymerase chain reaction were determined. Processing was carried out under full LIMS control in a fully automated Illumina BeadLab equipped with eight Tecan liquid handling robots, six Illumina BeadArray readers, and two Illumina iScans. Genotyping was carried out with Illumina Human 370K Quad arrays (n = 1729 samples) or, for a small number of the samples (n = 37), with Illumina Human 370K Duo arrays (www.illumina.com). Raw data were analyzed using GTS Image (www.illumina.com) and extracted for statistical analysis. Stage 2 genotyping was conducted using Illumina Golden Gate


Stage 1 case patients
BBC n = 1694 (1136 registry + 558 NCRN) after 72 removed
Illumina Hap370K
Samples removed:-
n = 3 (low chromosome X heterozygosity)
n = 14 (call rate < 95%)
n = 20 (IBS – identical)
n = 3 (IBS – closely related)
n = 32 (non-Caucasian)
Stage 1 control subjects
58C n = 1438
CORGI n = 927
Illumina Hap550K
CGEMS Stage 1
Case patients n = 1145
Control subjects n = 1142
Illumina Hap550K (546 646 SNPs)
Stage 1
Stage 1 SNPs (common across datasets)
n = 296 114 (after 49 189 removed)
SNPs removed:-
n = 10 830 (X or Y chromosome SNPs)
n = 34 175 (not genotyped in controls)
n = 20 (monomorphic in controls)
n = 463 (HWE < 10-5 in controls)
n = 226 (MAF < .01 in controls)
n = 688 (monomorphic in cases)
n = 22 (MAF < .01 in cases)
n = 2765 (call rate < 90% in cases)
209 SNPs with P < .001
1029 top SNPs ranked by P-value
n = 31 SNPs with low designability score:-
n = 16 SNPs substituted with tag SNP (r2 > 0.7)
n = 15 SNPs excluded (no suitable tag SNP)
1223 SNPs
Stage 2 case patients
BBC n = 3080 (after 25 removed)
RMH n = 1724 (after 0 removed)
Illumina Golden Gate
Samples removed:-
n = 8 (unexpected duplicate)
n = 17 (call rate < 95%)
Stage 2 control subjects
BBC n = 2902 (after 4 removed)
MOG n = 1034 (after 12 removed)
Illumina Golden Gate
Samples removed:-
n = 2 (unexpected discordance)
n = 13 (unexpected duplicate)
n = 1 (call rate <95%)
Stage 2
Stage 2 SNPs
n = 1176 (after 47 removed)
SNPs removed:-
n = 7 (HWE < 10-5 in controls)
n = 40 (call rate < 98%)
14 SNPs
Stage 3 case patients
BGS n = 4237 (after 39 removed)
Kbiosciences KASPar
Samples removed:-
n = 39 (> 2 missing genotypes)
Stage 3 control subjects
BGS n = 5044 (after 40 removed)
Kbiosciences KASPar
Samples removed:-
n = 40 (> 2 missing genotypes)
Stage 3
Stage 3 SNPs
n = 14 (after 0 removed)

(continued)

(www.illumina.com) technology and was performed by Tepnel Life Sciences. Liquid handling for stage 3 samples was carried out using a Star workstation (Hamilton Robotics, Bonaduz, Switzerland), and genotyping was conducted using competitive allele-specific polymerase chain reaction KASPar chemistry (KBiosciences Ltd, Hertfordshire, UK; http://kbioscience.co.uk/).

Post-Genotyping Quality Control

In stage 1, we genotyped 345 303 tagging SNPs in 1766 breast cancer case subjects. A total of 300 298 autosomal SNPs with call rates greater than 90% in control subjects were represented in both case and control subjects (Figure 1), of which 297 533 (99.1%) were satisfactorily genotyped in case subjects, with mean individual sample call rates of 99.1% (call rate is defined as the proportion of samples for which a genotype can be assigned). We excluded 708 SNPs that were monomorphic in either case or control subjects and 248 with a minor allele frequency (MAF) less than 1%. Deviation of the genotype frequencies in the control subjects from those expected under Hardy–Weinberg equilibrium was assessed by a χ^2 test or Fisher exact test where an expected cell count was less than 5. Four hundred sixty-three SNPs that showed extreme deviation in Hardy–Weinberg equilibrium ($P < 10^{-5}$ in control subjects) were excluded, leaving 296 114 SNPs for analysis (Figure 1).

The results of an identity-by-state analysis were used as a first step to determine outliers or related individuals for elimination. For any pair with allele sharing of more than 80%, the sample generating the lowest call rate was excluded from further analysis (n = 23). An ancestry analysis was carried out using the EIGENSTRAT 2.0 software (http://genepath.med.harvard.edu/~reich/Software.htm). Haplotype Map Project (HapMap, http://www.hapmap.org/ data (CEU [Centre d'Etude du Polymorphisme Humain collection], YRI [Yoruba population in Ibadan, Nigeria], JPT [Japanese population in Tokyo, Japan], and CHB [Han Chinese population in Beijing, China]) and samples of reference Europeans were used as representatives of European, West African, and East Asian populations to infer ancestry-informative principal components, which were then projected onto the case and control samples. The first two principal components for each individual were plotted, and any individual not present in the main CEU cluster (ie, outside 5% from cluster centroids) was excluded from subsequent analyses (n = 32; Supplementary Figure 1, available online). Fourteen samples that had completion rates of less than 95% and three samples with low chromosome X heterozygosity were also excluded making a total of 72 case samples that were excluded from the analysis (Figure 1). Association between each SNP and risk of breast cancer was assessed using the Cochran–Armitage trend test. The adequacy of the case–control matching and possibility of differential genotyping of case and control subjects were formally evaluated using Q–Q plots of test statistics (Supplementary Figure 2, available online).

In stage 2, we selected 1223 SNPs for follow-up genotyping with Illumina Golden Gate technology using two strategies: 1) 1020 SNPs were selected on the basis of Armitage trend test *P* values from our stage 1 and 2) a further 203 were selected based on our stage 1 and CGEMS stage 1 combined. The CGEMS case subjects are likely to differ from those included in our stage 1 scan because they had later disease onset (they were all postmenopausal) and were not selected for having a genetic predisposition to breast cancer. Combining our data with publicly available CGEMS data, however, increased the statistical power of our stage 1 study to detect alleles associated with breast cancer risk in both populations. Genotyped samples were 4829 case subjects recruited through the BBC study or through the RMH and 3952 control subjects. Because these samples were ascertained from a number of sources, genotype clustering, and preliminary post-genotyping quality control were performed separately according to the source of DNA. We excluded 40 SNPs on the basis of call rates less than 98%, and seven that showed extreme deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-5}$) were also dropped, leaving a total of 1176 SNP assays (96%) that were successful by these criteria. Similarly to stage 1, we removed samples with call rates less than 95% (n = 18). We identified duplicate samples and closely related individuals by considering all possible pairs of samples and determining the pairwise genotype concordance rate. Samples were subsequently categorized as expected or unexpected duplicates if the pairwise concordance rate was 80% or more (unexpected duplicates, n = 11, Figure 1). Individual sets of samples were merged, and we identified an additional 10 pairs of unexpected duplicates, representing individuals who had participated in two separate studies; one member of each duplicate pair was then excluded (n = 5). Thus, a total of 41 samples (25 case samples and 16 control samples) were excluded, leaving 4804 case subjects and 3936 control subjects for analysis. The overall completion rate by sample was greater than 99.9%.

At stage 3, after excluding SNPs correlated with those identified in previous GWA studies ($r^2 > 0.2$), we genotyped the 14 most statistically significant SNPs from combined analysis of stage 1, CGEMS, and stage 2 in a further 4276 case subjects and 5084 control subjects from the BGS (Figure 1). Seventy-nine individuals who returned no-calls at two or more loci (39 case patients and 40 control subjects) were excluded, leaving 4027 case subjects and

Figure 1 (continued).

Figure 1. Patient and single-nucleotide polymorphism (SNP) exclusion schema. Numbers of case patients and control subjects, numbers of SNPs analyzed, and genotyping platform for each stage are shown. The top ranked 1020 SNPs from stage 1 (**solid lines**) and an additional 203 SNPs associated with breast cancer risk ($P < .001$) from stage 1 and CGEMS combined (**dashed lines**) were taken through to stage 2. Numbers of samples and SNPs, after quality control exclusions, and reasons for exclusion are shown. 58C = 1958 British Birth Cohort; BBC = British Breast Cancer study; BGS = Breakthrough Generations Study; CGEMS = Cancer Genetic Markers of Susceptibility; CORGI = Colorectal Tumor Gene Identification study; FH = family history of breast cancer; HWE = Hardy–Weinberg equilibrium; IBS = identical by state; MAF = minor allele frequency; MOG = Mammography, Oestrogens and Growth factor study; NCRN = National Cancer Research Network; RMH = Royal Marsden Hospital breast cancer cases .

5044 control subjects for analysis (Figure 1). The overall completion rate by sample was 99.0%. Overall, 5% duplicate pairs were included in the study to assess genotyping concordance, and the concordance rate was 99.7%.

Statistical Methods

Statistical analyses were performed using GLU version 1.0a6 (code.google.com/p/glu-genetics/), R version 2.6 (http://www.r-project.org/), STATA version 10 (College Station, TX), and PLINK version 1.05 (http://pngu.mgh.harvard.edu/~purcell/plink/). All *P* values reported are two-sided. Odds ratios and associated 95% confidence intervals (CIs) in each stage and all stages combined were calculated using unconditional logistic regression (adjusted for stage in combined analyses). Odds ratios for each locus were determined by fitting multiplicative and unconstrained genetic models. *P* values were estimated using likelihood ratio tests with either 1 or 2 *df* for multiplicative and unconstrained models, respectively. Cochran *Q* statistic to test for heterogeneity and the I^2 statistic (18) to quantify the proportion of the total variation due to heterogeneity were calculated. The relationship between age at menarche and genotype was assessed using linear regression adjusting for case–control status and stage. To assess the relationship between age at diagnosis and genotype, age group–specific odds ratios were calculated using unconditional logistic regression. Case patients in each age group were compared with all control subjects, adjusted for stage in the combined analysis. Case-only unconditional logistic regression was used to test for a trend with age group. *P* values were estimated using likelihood ratio tests with 1 *df*. Under the multiplicative polygenic model (3), the variance of the log(risk) in the population is 2log(FRR), where FRR is the familial relative risk in a patient's mother and sisters. The contribution to this overall variance from each allele at a susceptibility locus with MAF q conferring a (relative) risk R is

$$q \times (1 - q) \times [\log(R)]^2,$$

so the contribution to the overall variance is doubled for the two alleles.

Bioinformatics

Linkage disequilibrium (LD) metrics (r^2 and D') between SNPs reported in HapMap were based on release 27, NCBI B36, and were computed using the Tagzilla module as implemented in GLU version 1.0a6. Association plots were produced using a modified version of SNAP (SNP Annotation and Proxy Search, http://www.broadinstitute.org/mpg/snap/).

Results

In stage 1, after applying quality control filters, 296 114 autosomal SNPs genotyped in 1694 breast cancer case subjects and 2365 control subjects were tested for association with breast cancer risk (Figure 1). There was little evidence of hidden population substructure or differential genotype calling between case and control subjects (inflation factor $\lambda = 1.05$, based on the 95% least statistically significant SNPs; Supplementary Figure 2, available online).

In stage 2, after imposing stringent quality control metrics, 1176 SNP assays genotyped in 4804 case subjects and 3936 control subjects were retained for analysis (Figure 1). Combining stage 1, CGEMS, and stage 2 data, markers in six of the loci identified in previous GWA studies [2q35, 3p24.1 (solute carrier family 4, sodium bicarbonate transporter, member 7 *SLC4A7*), 5p12, 8q24.21, 10q26.13 (fibroblast growth factor receptor 2, *FGFR2*), and 16q12.1 (TOX high-mobility group box family member 3, *TOX3*) (4–10)] were highly statistically significant ($P < 5 \times 10^{-7}$, Table 1) (19).

Joint analysis of combined data from all stages provided evidence of an association between rs865686 and breast cancer risk on the basis of conventionally accepted thresholds for genome-wide statistical significance (19). SNP rs865686, which maps to 9q31.2 (109 928 299 bp), was associated with an allelic odds ratio = 0.89 (95% CI = 0.85 to 0.92; $P = 1.75 \times 10^{-10}$; Table 2 and Figure 2). The association between rs865686 and breast cancer risk was consistent across each case–control series (Table 2; $P_{het} = .33$, $I^2 = 12.8\%$). Odds ratios for heterozygotes and homozygotes were 0.90 (95% CI = 0.86 to 0.96) and 0.77 (95% CI = 0.71 to 0.84), respectively, consistent with a multiplicative model of allelic risk. Two other SNPs that map 6 kb centromeric and 134 kb telomeric to rs865686 (rs667052 at 109 922 211 bp and rs7030526 at 110 062 346 bp, respectively) also showed evidence of association with breast cancer risk ($P = 5.37 \times 10^{-5}$ and 1.03×10^{-5}, respectively; Supplementary Table 1, available online). Although rs667052 is in moderate LD with rs865686 ($r^2 = 0.52$, $D' = 0.99$ in stage 1; Supplementary Table 2, available online), rs7030526 is only weakly correlated with rs865686 ($r^2 = 0.23$, $D' = 0.50$ in stage 1; Supplementary Table 2, available online). The association of rs865686 with breast cancer risk remained highly statistically significant after adjustment for either rs667052 or rs7030526, with adjusted allelic OR = 0.87 (95% CI = 0.83 to 0.92; $P = 6.48 \times 10^{-7}$) and 0.90 (95% CI = 0.86 to 0.94, $P = 2.05 \times 10^{-6}$), respectively. There was, however, no evidence for an independent association between rs667052 and rs7030526 with breast cancer risk after adjusting for rs865686 ($P = .43$ and $P = .14$ for rs667052 and rs7030526, respectively).

The SNP rs865686 localizes to a 17 kb region of LD (109 927 817–109 944 558 bp) on the basis of distribution confidence intervals as defined by Gabriel et al. (21), lacking identifiable genes or predicted transcripts. The nearest genes are Kruppel-like factor 4 (KLF4, 636 kb centromeric), *RAD23B* (794 kb centromeric), and actin-like 7A (*ACTL7A*, 736 kb telomeric). KLF4 is a transcription factor that participates in both tumor suppression and oncogenesis. Interrogation of the Oncomine database (22) has shown a decrease in levels of KLF4 RNA transcripts in breast cancers and an association between *KLF4* expression and estrogen receptor-alpha (ERα) positivity (23). RAD23B, which functions in the nucleotide excision repair pathway and the actin-related protein family, of which ACTL7A is a member, are involved in diverse cellular processes, including vesicular transport, spindle orientation, nuclear migration, and chromatin remodeling (24).

Early age at menarche is an established risk factor for breast cancer (25), and two recent GWA studies have identified an association between 9q31.2 SNPs (rs7861820, rs12684013, rs4452860, rs7028916, and rs2090409) and age at menarche (26,27). These loci map more than 2 Mb from rs865686 and are not correlated with it ($r^2 < 0.01$, $D' < 0.09$ in CEU HapMap phase 2; Supplementary

Table 2. Results for single-nucleotide polymorphisms (SNPs) mapping to 9q31.2, 6q25.1, and 10q26.13 identified in this genome-wide association study and replication series*

SNP, cytoband, location, allele†	Stage	No. of case patients/ control subjects	MAF	OR_{het} (95% CI)	OR_{hom} (95% CI)	OR_{all} (95% CI)	P_{trend}	P_{het} (I^2)
rs865686, 9q31.2, 109 928 299 bp, T, G	Stage 1	1694/2349	0.39	0.85 (0.74 to 0.97)	0.70 (0.58 to 0.86)	0.84 (0.77 to 0.92)	2.28×10^{-4}	
	CGEMS	1135/1134	0.38	0.91 (0.76 to 1.09)	0.68 (0.52 to 0.88)	0.85 (0.75 to 0.96)	7.45×10^{-3}	
	Stage 2	4802/3930	0.38	0.97 (0.88 to 1.06)	0.82 (0.72 to 0.94)	0.92 (0.87 to 0.98)	1.28×10^{-2}	
	Stage 3	4150/4974	0.39	0.87 (0.80 to 0.95)	0.78 (0.69 to 0.89)	0.88 (0.83 to 0.94)	3.21×10^{-5}	
	Combined	11 781/12 387	0.39	0.90 (0.86 to 0.96)	0.77 (0.71 to 0.84)	0.89 (0.85 to 0.92)	1.75×10^{-10}	.33 (12.8%)
rs9383938, 6q25.1, 152 029 050 bp, G, T	Stage 1	1694/2361	0.07	1.37 (1.15 to 1.63)	0.82 (0.36 to 1.87)	1.29 (1.10 to 1.51)	1.96×10^{-3}	
	CGEMS	1145/1142	0.08	1.15 (0.92 to 1.43)	2.74 (1.07 to 7.04)	1.23 (1.01 to 1.51)	4.32×10^{-2}	
	Stage 2	4793/3928	0.08	1.23 (1.10 to 1.38)	1.25 (0.81 to 1.94)	1.21 (1.09 to 1.34)	3.52×10^{-4}	
	Stage 3	4140/4970	0.08	1.08 (0.96 to 1.21)	1.54 (1.00 to 2.36)	1.11 (1.00 to 1.23)	4.32×10^{-2}	
	Combined	11 772/12 401	0.08	1.18 (1.10 to 1.27)	1.40 (1.07 to 1.83)	1.18 (1.11 to 1.26)	1.41×10^{-7}	.44 (0%)
rs3734805, 6q25.1, 151 981 043 bp, A, C	Stage 1	1694/2365	0.07	1.34 (1.13 to 1.60)	0.51 (0.21 to 1.21)	1.22 (1.04 to 1.43)	1.63×10^{-2}	
	CGEMS	1145/1142	0.07	1.22 (0.97 to 1.54)	2.69 (0.96 to 7.59)	1.29 (1.04 to 1.59)	1.82×10^{-2}	
	Stage 2	4803/3935	0.08	1.22 (1.09 to 1.38)	1.60 (0.99 to 2.58)	1.23 (1.11 to 1.37)	1.24×10^{-4}	
	Stage 3	4200/4979	0.08	1.11 (0.99 to 1.25)	1.27 (0.82 to 1.96)	1.11 (1.01 to 1.24)	3.90×10^{-2}	
	Combined	11 842/12 421	0.08	1.20 (1.12 to 1.29)	1.31 (0.99 to 1.73)	1.19 (1.11 to 1.27)	1.35×10^{-7}	.48 (0%)
rs10510102, 10q26.13, 123 615 180 bp, A, G	Stage 1	1690/2351	0.17	1.18 (1.03 to 1.36)	1.60 (1.14 to 2.24)	1.21 (1.09 to 1.36)	7.27×10^{-4}	
	CGEMS	1144/1142	0.18	1.07 (0.89 to 1.28)	1.19 (0.76 to 1.87)	1.08 (0.93 to 1.25)	3.38×10^{-1}	
	Stage 2	4795/3932	0.17	1.18 (1.07 to 1.29)	1.23 (0.96 to 1.56)	1.15 (1.06 to 1.24)	3.97×10^{-4}	
	Stage 3	4211/5011	0.17	1.01 (0.92 to 1.11)	1.33 (1.06 to 1.67)	1.06 (0.99 to 1.15)	1.02×10^{-1}	
	Combined	11 840/12 436	0.17	1.10 (1.04 to 1.17)	1.32 (1.15 to 1.52)	1.12 (1.07 to 1.17)	1.58×10^{-6}	.21 (34.1%)

* Odds ratios (ORs) and associated 95% confidence intervals (CIs) in each stage and all stages combined were calculated using unconditional logistic regression. All statistical tests were two-sided. CGEMS = Cancer Genetic Markers of Susceptibility; I^2 = percentage of variance that is between stages; MAF = minor allele frequency; OR_{all} = allelic odds ratio; OR_{het} = heterozygote odds ratio; OR_{hom}= homozygote odds ratio; P_{het} = P for Cochran Q test for heterogeneity; P_{trend} = P for null hypothesis of allelic OR= 1.0.

† In order major, minor.

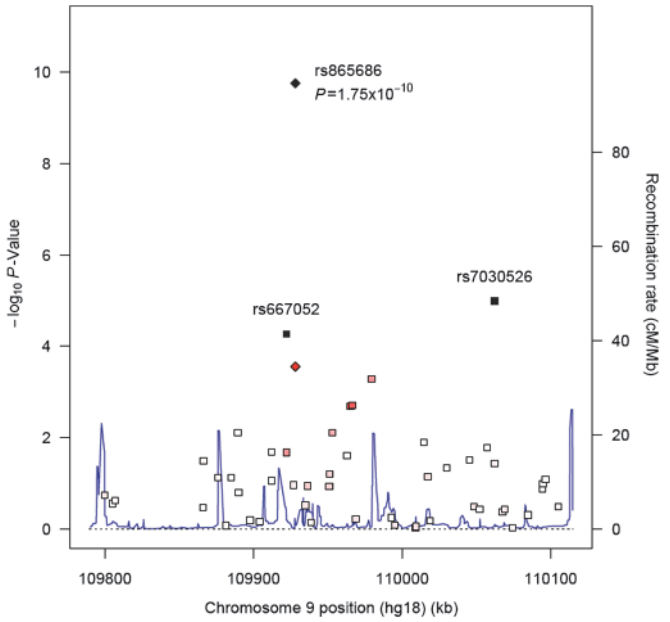


Figure 2. The 9q31.2 breast cancer locus. The local recombination rate (20) is plotted in **blue** over this 309 kb chromosomal segment centered on rs865686. Each **square** represents a single-nucleotide polymorphism found in this locus in the stage 1 genome-wide association study. rs865686 is marked by a **red diamond**. The color intensity of each square reflects the extent of linkage disequilibrium with rs865686—red ($r^2 > 0.8$) through to white ($r^2 < 0.2$). Also shown are the combined analysis results. rs865686 is indicated by a **black diamond**, rs667052 and rs7030526 are indicated by **black squares**. Physical positions are based on build 36 of the human genome. cM/Mb = centiMorgans/megabase.

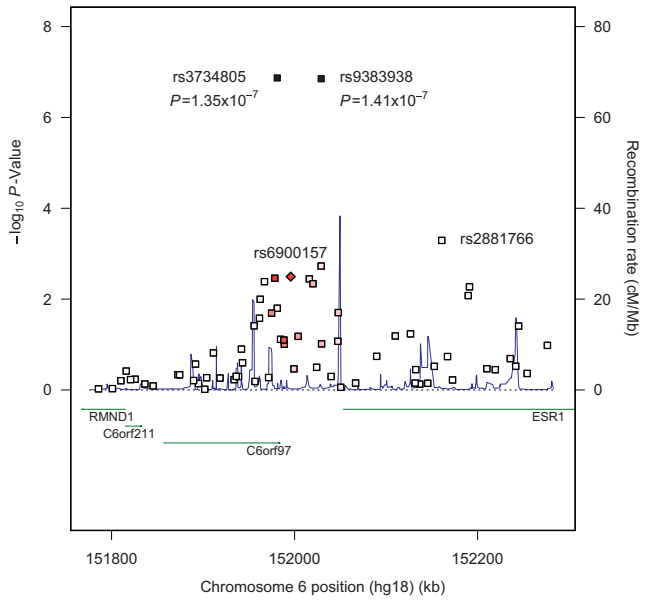


Figure 3. The 6q25.1 breast cancer locus. The local recombination rate is plotted in **blue** over this 491 kb chromosomal segment centered on rs9383938. Each **square** represents a single-nucleotide polymorphism (SNP) found in this locus in the stage 1 genome-wide association study. rs6900157 (a proxy for rs2046210) (10) is marked by a **red diamond** (rs6900157 was not genotyped in stages 2 and 3). The color intensity of each square reflects the extent of linkage disequilibrium with rs6900157 —red ($r^2 > 0.8$) through to white ($r^2 < 0.2$). Also shown are the combined analysis results. rs9383938 and rs3734805 are indicated by **black squares**. Physical positions are based on build 36 of the human genome. One additional SNP, rs2881766, mapping to 152 160 812 bp (165 kb telomeric to rs6900157) showed evidence of an association with breast cancer risk in stage 1 and CGEMS data. After stage 2, the combined evidence for an association was $P = .13$, and this SNP was not taken through to stage 3. CGEMS = Cancer Genetic Markers of Susceptibility study; cM/Mb = centiMorgans/Megabase.

Table 2, available online). An analysis of data from stage 1 (case patients), stage 2 (case patients and control subjects), and stage 3 (case patients and control subjects) provided no evidence that rs865686 is associated with age at menarche (difference in age at menarche per allele = 0.027 years, 95% CI = −0.007 to 0.062; $P = .13$). Three of the SNPs associated with age at menarche (rs12684013, rs4452860, and rs7028916) were genotyped in stage 1 and CGEMS. Adjustment for these three SNPs did not alter the risk of breast cancer associated with rs865686 (unadjusted allelic OR = 0.85, 95% CI = 0.79 to 0.92, $P = 5.00 \times 10^{-5}$; adjusted allelic OR = 0.85, 95% CI = 0.79 to 0.92, $P = 5.00 \times 10^{-5}$ in combined stage 1 and CGEMS data).

The other two strongly suggestive SNPs were rs3734805 (151 981 043 bp, allelic OR = 1.19, 95% CI = 1.11 to 1.27, $P = 1.35 \times 10^{-7}$, $P_{het} = .48$, $I^2 = 0\%$) and rs9383938 (152 029 050 bp, allelic OR = 1.18, 95% CI = 1.11 to 1.26, $P = 1.41 \times 10^{-7}$, $P_{het} = .44$, $I^2 = 0\%$, Table 2). Both SNPs map to 6q25.1 and are in moderate LD ($r^2 = 0.67$, $D' = 0.83$ in stage 1, Supplementary Table 2, available online, and Figure 3). In mutually adjusted analysis, only rs9383938 remained statistically significant (allelic OR = 1.14, 95% CI = 1.01 to 1.28, $P = .03$).

A risk locus for breast cancer mapping to 6q25.1 (*ESR1*), annotated by rs2046210 (151 990 059 bp), has previously been reported by Zheng et al. (10) in a study of primarily Chinese subjects. SNP rs2046210 was not genotyped in our study, but a proxy, rs6900157, which is highly correlated with rs2046210 ($r^2 = 0.92$, $D' = 1.00$ in CEU HapMap phase 2) was genotyped in stage 1 and CGEMS. SNP rs9383938 is only weakly correlated with rs6900157 ($r^2 = 0.13$, $D' = 0.90$ in stage 1, Supplementary Table 2, available online), and, in mutually adjusted analysis of stage 1 and CGEMS data, both SNPs remained statistically significant (rs9383938 OR= 1.20, 95% CI 1.05 to 1.37, $P = .008$ and rs6900157 OR = 1.10, 95% CI = 1.01 to 1.18, $P = .02$). In the analysis by Zheng et al. (10), the allelic odds ratio estimate in subjects of European ancestry was lower than that reported in subjects of Chinese ancestry (OR = 1.15, 95% CI = 1.03 to 1.28, $P = .01$ and OR = 1.29, 95% CI = 1.21 to 1.37, $P = 2.0 \times 10^{-15}$, respectively), suggesting that there may be heterogeneity in the magnitude of the association between these populations (10). To investigate this further, we genotyped rs2046210 in stage 3; our odds ratio estimate (OR = 1.04, 95% CI = 0.97 to 1.10, $P = .26$) did not provide independent evidence of an association with breast cancer risk, although it was consistent with the published odds ratio estimate in subjects of European ancestry (Cochran $Q = 2.6$, $P_{het} = .11$). The LD structure within the region (151 974 868 to 152 029 050 bp), defined by 15 SNPs that were genotyped by Zheng et al. (10) and the two SNPs (rs3734805 and rs9383938) that were associated with breast cancer risk in this GWA, differs between Han Chinese subjects from Beijing (HCB) and Caucasian subjects from the CEU (Supplementary Figure 3, available online). Although the correlation between rs2046210 [Zheng et al. (10)] and rs9383938 (this study) is weak in both

populations (r^2 = 0.28, D' = 0.57 and r^2 = 0.07, D' = 0.78 in HCB and CEU, respectively, HapMap phase 2), they are each correlated with the same single variant (rs9397436) in their respective populations (rs2046210 and rs9397436; r^2 = 0.82, D' = 1.0 in HCB; rs9383938 and rs9397436, r^2 = 0.70, D' = 1.0 in CEU). A similar pattern is observed between rs2046210 [Zheng et al. (10)], rs3734805 (this study), and rs9397436 (Supplementary Figure 3, available online).

The strongest breast cancer association to be discovered through GWA studies maps to 10q26.13 (*FGFR2*), annotated by rs2981582 (4) and rs1219648 (5). SNP rs10510102 (123 615 180 bp), which maps 279 kb telomeric to rs1219648 (123 336 180 bp), was associated with breast cancer risk in this GWA (OR = 1.12, 95% CI = 1.07 to 1.17, $P = 1.58 \times 10^{-6}$, $P_{het} = 0.21$, $I^2 = 34\%$; Table 2 and Figure 4), albeit not at genome-wide statistical significance. rs10510102 is not correlated with rs1219648 (r^2 = 0.006, D' = 0.14 in stage 1), and after adjusting rs10510102 for rs1219648 in data from stage 1, CGEMS and stage 2 combined, the association between rs10510102 and breast cancer risk remained statistically significant (OR = 1.12, 95% CI = 1.05 to 1.19, $P < .001$).

Using inferred ancestral recombination graphs, Hunter et al. (5) demonstrated that a single risk locus at 10q26.13 (*FGFR2*) within a region defined by 123 225 862 to 123 471 190 bp is likely to underlie the association between rs1219648 and breast cancer risk. Fine mapping and functional studies proposed rs2981578 mapping to 123 330 301 bp as the causal basis of this association (28,29). The SNP rs10510102 lies approximately 144 kb telomeric to the region defined by Hunter et al. (5), within intron 8 of arginyltransferase 1 (*ATE1*), an enzyme that plays a role in ubiquitin-dependent protein degradation.

We also assessed the relationship between age at diagnosis and genotype at each of these three loci (9q31.2 [rs865686], 6q25.1 [rs9383938 and rs3734805], and 10q26.13 [rs10510102], using case-only unconditional logistic regression (Supplementary Table 3, available online). There was no trend in odds ratio with age at diagnosis for any of these loci.

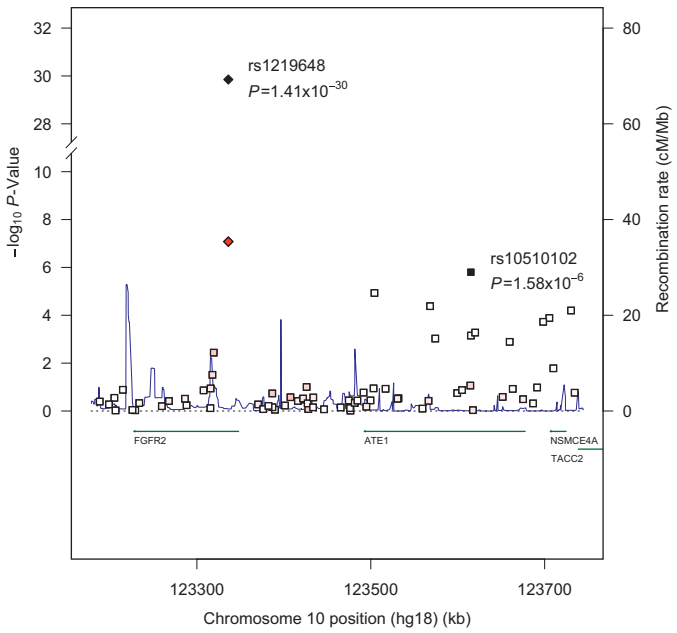


Figure 4. The 10q26.13 breast cancer locus. The local recombination rate is plotted in **light blue** over this 546 kb chromosomal segment that includes both rs1219648 (5) and rs10510102 (this study). Each **square** represents a single-nucleotide polymorphism found in this locus in the stage 1 genome-wide association study. rs1219648 is marked by a **red diamond**. The color intensity of each square reflects the extent of linkage disequilibrium with rs1219648—red ($r^2 > 0.8$) through to white ($r^2 < 0.2$). Also shown are the combined analysis results. rs10510102 (which was genotyped in stage1, CGEMS, stage 2 and stage 3) is indicated by a **black square**, rs1219648 (which was genotyped in stage 1, CGEMS and stage 2) is indicated by a **black diamond**. Physical positions are based on build 36 of the human genome. CGEMS = Cancer Genetic Markers of Susceptibility study; cM/Mb = centiMorgans/Megabase.

Discussion

We have identified a new risk locus for breast cancer at 9q31.2 and provide evidence of an association between variants mapping to 6q25.1 (*ESR1*) and breast cancer risk in subjects of European ancestry. Additional studies will be required to identify causal variants underlying the associations we have observed at 9q31.2, 10q26.13, and 6q25.1 and to determine whether, although only weakly linked, rs9383938 and rs2046210 (*ESR1*) are correlated with the same causal variant. The SNPs rs9383938 and rs3734805 map 24 and 72 kb, respectively, from the 5′ untranslated region of *ESR1* and, given the prior evidence that ESR1 plays a role in breast cancer etiology, it seems likely that both SNPs are correlated with a causal variant that exerts an effect on *ESR1* levels of expression. The SNP rs10510102 (10q26.13) lies within intron 8 of *ATE1*, but the proximity of rs10510102 to *FGFR2*, another gene with clear relevance to breast cancer and a proven association with breast cancer risk (4,5), makes any association between rs10510102 and breast cancer risk likely to be mediated through LD with sequence changes that affect expression of *FGFR2*. Although rs865686 (9q31.2) lies more than 600 kb from the nearest genes, *KLF4* and *RAD23B* are both attractive candidates for mediating an effect on breast cancer risk. Recent functional studies of rs6983267 (a colorectal cancer risk locus mapping to 8q24.21.) have shown that physical interaction between a causal variant and its target (the MYC proto-oncogene) can occur over a large distance (~335 kb) (30,31).

Given that only a restricted number of SNPs were evaluated in replication stages, the main limitation of our study is statistical power of the overall design to harvest either common variants conferring relative risks of breast cancer of less than 1.10 or variants with minor allele frequencies of less than 10%, even if they were associated with substantial effects. For example at a significance threshold of $P < .003$ (the threshold for progressing a SNP to stage 2 replication), we had 50% statistical power to detect a variant with MAF of 20% and an odds ratio of 1.08 or a variant with MAF of 10% and an odds ratio of 1.11

An important eventual aim of GWA studies is to account for a large enough proportion of the excess FRR to identify women who would benefit from more intensive screening or prophylaxis. The conventional assumption that the FRR is about 2 implies that the known susceptibility loci for breast cancer (8), including those reported more recently (9,10) and this study, account for 7%–8% of this overall variation, and this may increase when all the functional variants at these loci have been identified. The

FRR in the mother, sisters, and daughters of a woman diagnosed with breast cancer before age 40 years is more than 5 below age 40 years but falls to 1.4 above age 60 years (32). The weak or absent trend with age in the allelic odds ratio for the variant at 9q31.2 and other known susceptibility loci (4) thus suggests that above age 60 years, when most breast cancers are diagnosed, variants in known loci may already account for 20% or more of genetic variation, even if the observed FRR in older women is due entirely to genetic effects. The additional contribution of known SNPs to a risk prediction model based on family history and nongenetic risk factors, however, is minor for women of any age (33).

In conclusion, GWA studies have identified multiple common genetic variants associated with breast cancer risk. Given the size of GWA studies that have been conducted to date, and the coverage current arrays afford, it is unlikely that there are many more common disease loci with MAFs >20% in European populations that have stronger effects than those already identified. It is likely, however, that there are many more susceptibility loci, conferring odds ratios of 1.05–1.10 that have not yet been detected. Larger studies, combined analyses across multiple scans, GWA scans in non-European populations, and scans of well-defined breast cancer subtypes may lead to the identification of additional loci. In addition, current estimates of effect sizes are likely to be conservative as the effect of causal variants will typically be larger than the associations detected through tag SNPs. Identifying the causal variants and determining their functional consequences will be challenging, but such insights have the potential to provide a greater understanding of cancer biology and suggest potential targets for therapeutic and preventive measures.

References

1. Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet.* 2000;26(4):411–414.
2. Peto J, Collins N, Barfoot R, et al. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J Natl Cancer Inst.* 1999;91(11):943–949.
3. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet.* 2002;31(1):33–36.
4. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007; 447(7148):1087–1093.
5. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* 2007;39(7):870–874.
6. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet.* 2007;39(7):865–869.
7. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet.* 2008;40(6):703–706.
8. Ahmed S, Thomas G, Ghoussaini M, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet.* 2009;41(5): 585–590.
9. Thomas G, Jacobs KB, Kraft P, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet.* 2009;41(5):579–584.
10. Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet.* 2009; 41(3):324–328.
11. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol.* 2003;25(3): 190–202.
12. Fletcher O, Johnson N, Palles C, et al. Inconsistent association between the STK15 F31I genetic polymorphism and breast cancer risk. *J Natl Cancer Inst.* 2006;98(14):1014–1018.
13. Johnson N, Fletcher O, Naceur-Lombardelli C, dos Santos Silva I, Ashworth A, Peto J. Interaction between CHEK2*1100delC and other low-penetrance breast-cancer susceptibility genes: a familial study. *Lancet.* 2005;366(9496):1554–1557.
14. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet.* 2007;39(8):984–988.
15. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol.* 2006;35(1):34–41.
16. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447(7145):661–678.
17. Moss S, Thomas I, Evans A, Thomas B, Johns L. Randomised controlled trial of mammographic screening in women from age 40: results of screening in the first 10 years. *Br J Cancer.* 2005;92(5):949–954.
18. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557–560.
19. Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype-phenotype associations. *Nature.* 2007;447(7145):655–660.
20. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science.* 2004;304(5670):581–584.
21. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science.* 2002;296(5576):2225–2229.
22. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia.* 2004;6(1):1–6.
23. Akaogi K, Nakajima Y, Ito I, et al. KLF4 suppresses estrogen-dependent breast cancer growth by inhibiting the transcriptional activity of ERalpha. *Oncogene.* 2009;28(32):2894–2902.
24. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2005;33(database issue):D54–D58.
25. Bernstein L. Epidemiology of endocrine-related risk factors for breast cancer. *J Mammary Gland Biol Neoplasia.* 2002;7(1):3–15.
26. He C, Kraft P, Chen C, et al. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat Genet.* 2009;41(6):724–728.
27. Perry JR, Stolk L, Franceschini N, et al. Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat Genet.* 2009;41(6):648–650.
28. Meyer KB, Maia AT, O'Reilly M, et al. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol.* 2008;6(5):e108.
29. Udler MS, Meyer KB, Pooley KA, et al. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum Mol Genet.* 2009;18(9):1692–1703.
30. Pomerantz MM, Ahmadiyeh N, Jia L, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet.* 2009;41(8):882–884.
31. Tuupanen S, Turunen M, Lehtonen R, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet.* 2009;41(8):885–890.
32. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet.* 2001;358(9291):1389–1399.
33. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *J Natl Cancer Inst.* 2009;101(13):959–963.

Funding

Cancer Research UK (C150/A5660 and C1178/A3947 to J.P. and I.d.S.S.); Breakthrough Breast Cancer (A.A. and O.F.); the Institut National de Cancer (M.L.); the Cridlan Trust (G.R.). We acknowledge National Health Service funding to the NIHR Biomedical Research Centre and the National Cancer

Research Network (NCRN). Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

Notes

O. Fletcher and N. Johnson contributed equally to this work.This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of these data is available from www.wtccc.org.uk. We also used genome-wide association data from the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer study. A full list of the investigators who contributed to the generation of the CGEMS data is available from http://cgems.cancer.gov/. We are grateful to all the patients and control subjects for their participation. We thank the clinicians and other hospital staff, cancer registries, and study staff who contributed to the blood sample and data collection for the British Breast Cancer study, Breakthrough Generations Study, and Mammography Oestrogens and Growth Factors study. The sponsor had no role in the design of the study; the collection, analysis and interpretation of the data; the writing of the article; and the decision to submit the article for publication.

Affiliations of authors: Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK (OF, NJ, NO, CP, BC, JW, SC-B, KT, GS, AA); Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, UK (FJH, PB, RSH); Non-communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, UK (LJG, KW, IdSS, JP); Centre National de Génotypage, IG/CEA, Evry Cedex, France (DZ, IG, SH, ML); Section of Epidemiology, Institute of Cancer Research, Sutton, Surrey, UK (MS, MJ, AS); Core Genotyping Facility, Advanced Technology Program, SAIC-Frederick Inc, National Cancer Institute at Frederick, Frederick, MD (KBJ); Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD (KBJ, SJC); BioInformed LLC, Gaithersburg, MD (KBJ); Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA (DJH); Molecular and Population Genetics, Wellcome Trust Centre for Human Genetics, Oxford, UK (IPT); The Royal Marsden NHS Foundation Trust, London, UK (GR); Foundation Jean Dausett-CEPH, Paris, France (ML); Cancer Research UK Genetics and Epidemiology Group, Institute of Cancer Research, Sutton, Surrey, UK (JP).