nature genetics

A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*)

Gilles Thomas[1], Kevin B Jacobs[1–3], Peter Kraft[4], Meredith Yeager[1,3], Sholom Wacholder[1], David G Cox[4,5], Susan E Hankinson[5], Amy Hutchinson[1,3], Zhaoming Wang[1,3], Kai Yu[1], Nilanjan Chatterjee[1], Montserrat Garcia-Closas[1], Jesus Gonzalez-Bosquet[1], Ludmila Prokunina-Olsson[1], Nick Orr[1], Walter C Willett[5,6], Graham A Colditz[7], Regina G Ziegler[1], Christine D Berg[8], Saundra S Buys[9], Catherine A McCarty[10], Heather Spencer Feigelson[11], Eugenia E Calle[11], Michael J Thun[11], Ryan Diver[11], Ross Prentice[12], Rebecca Jackson[13], Charles Kooperberg[12], Rowan Chlebowski[14], Jolanta Lissowska[15], Beata Peplonska[16], Louise A Brinton[1], Alice Sigurdson[1], Michele Doody[1], Parveen Bhatti[1], Bruce H Alexander[17], Julie Buring[18], I-Min Lee[18], Lars J Vatten[19], Kristian Hveem[19], Merethe Kumle[20], Richard B Hayes[1], Margaret Tucker[1], Daniela S Gerhard[21], Joseph F Fraumeni Jr[1], Robert N Hoover[1], Stephen J Chanock[1] & David J Hunter[1,4–6,22]

We conducted a three-stage genome-wide association study (GWAS) of breast cancer in 9,770 cases and 10,799 controls in the Cancer Genetic Markers of Susceptibility (CGEMS) initiative. In stage 1, we genotyped 528,173 SNPs in 1,145 cases of invasive breast cancer and 1,142 controls. In stage 2, we analyzed 24,909 top SNPs in 4,547 cases and 4,434 controls. In stage 3, we investigated 21 loci in 4,078 cases and 5,223 controls. Two new loci achieved genome-wide significance. A pericentromeric SNP on chromosome 1p11.2 (rs11249433; $P = 6.74 \times 10^{-10}$ adjusted genotype test, 2 degrees of freedom) resides in a large linkage disequilibrium block neighboring *NOTCH2* and *FCGR1B*; this signal was stronger for estrogen-receptor–positive tumors. A second SNP on chromosome 14q24.1 (rs999737; $P = 1.74 \times 10^{-7}$) localizes to *RAD51L1*, a gene in the homologous recombination DNA repair pathway. We also confirmed associations with loci on chromosomes 2q35, 5p12, 5q11.2, 8q24, 10q26 and 16q12.1.

Epidemiologic investigation of breast cancer has identified a number of environmental and lifestyle risk factors[1]. Breast cancer is nearly twice as frequent in first-degree relatives of women with the disease than in relatives of women without this history, suggesting an important contribution of inherited susceptibility. Established variants from before the GWAS era account for a small fraction of sporadic breast cancers. These include rare, high-penetrance germline mutations segregating in high-risk pedigrees, notably in the *BRCA1* and *BRCA2* genes[2,3], and a handful of rare susceptibility variants with lower penetrance identified in DNA repair and apoptosis genes[4–8]. Only one common variant with a minor allele frequency larger than 5% (*CASP8*) was found using the candidate gene approach[9].

Genome-wide association studies have identified multiple new common genetic variants influencing breast cancer risk. Easton *et al.* genotyped 390 cases enriched for a family history of breast cancer and 364 controls with 227,876 SNPs and followed the top 10,405 SNPs in a

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA. [2]Bioinformed Consulting Services, Gaithersburg, Maryland, USA. [3]Core Genotyping Facility, Advanced Technology Program, SAIC-Frederick Inc., NCI-Frederick, Frederick, Maryland, USA. [4]Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [5]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [6]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. [7]Washington University School of Medicine, St. Louis, Missouri, USA. [8]Division of Cancer Prevention, NCI, NIH, DHHS, Bethesda, Maryland, USA. [9]Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA. [10]The Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA. [11]Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Georgia, USA. [12]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [13]Division of Diabetes, Endocrinology and Metabolism, The Ohio State University Medical Center, Columbus, Ohio, USA. [14]Harbor-University of California at Los Angeles Medical Center, Torrance, California, USA. [15]Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland. [16]Nofer Institute of Occupational Medicine, Łódź, Poland. [17]Division of Environmental Health Science, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA. [18]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [19]Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway. [20]Institute of Community Medicine, University of Tromso, Tromso, Norway. [21]Office of Cancer Genomics, NCI, NIH, DHHS Bethesda, Maryland, USA. [22]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. Correspondence should be addressed to D.J.H. (dhunter@hsph.harvard.edu).

Received 6 November 2008; accepted 29 January 2009; published online 29 March 2009; doi:10.1038/ng.353

Table 1 Three-stage study design

	Controls	Cases
Stage 1 (528,173 SNPs)		
NHS1	1,142	1,145
Stage 2 (30,278 SNPs)		
CPSII	543	535
PBCS1	506	669
PLCO	975	948
WHI	2,410	2,395
Stage 2 total	4,434	4,547
Stage 3 (24 SNPs)		
CONOR	498	516
WHS	701	696
NHS2	1,243	619
USRT	998	780
PBCS2	1,783	1,467
Stage 3 total	5,223	4,078
Stages 1–3 combined total	10,799	9,770

Nine studies participated in this multistage GWAS. A portion (26.6%, corresponding to 669 cases and 506 controls, designated as PBSC1) of the Polish Breast Cancer Study (PBCS) was genotyped using the custom iSelect Infinium (Illumina), and the remaining samples (73.4%, corresponding to 1,467 cases and 1,783 controls, designated as PBSC2) were genotyped in stage 3.

second-stage replication study; they then selected 30 SNPs for a much larger third stage (primarily using case-control studies of unrelated subjects), and identified five associated loci (10q26 (*FGFR2*), 16q12.1 (*TOX3*), 5q11.2 (*MAP3K1*), 8q24 and 11p15.5 (*LSP1*))[10]. In the initial report from the NCI Cancer Genetic Markers of Susceptibility (CGEMS) initiative, based on a follow-up of the top ten SNPs from the stage 1 GWAS, we independently identified SNPs in intron 2 of *FGFR2* as associated with breast cancer[11]. Subsequently, the *FGFR2* locus was confirmed in an Icelandic population[12], and another locus at 2q35 was associated with estrogen receptor (ER)-positive breast cancer[12]. Finally, combined analysis of a promising signal using the three published GWAS led to the identification of an additional locus on 5p12 (ref. 13). Power calculations based on the sample sizes of the three GWAS suggest that each has limited power to detect the low observed relative risks (1.1–1.3 per allele) at conventional levels of genome-wide significance ($P < 5 \times 10^{-7}$)[14]. Thus, it is likely a high proportion of susceptibility loci have not yet been detected.

In stage 1 of CGEMS, we genotyped 1,145 cases of invasive breast cancer in postmenopausal women of European ancestry and 1,142 matched controls nested within the prospective Nurses' Health Study cohort[11]. We used 528,173 SNPs estimated to be correlated with an $r^2 > 0.8$ to approximately 90% of the common HapMap Phase II SNPs. We report here a further follow-up: in stage 2, we attempted to genotype 30,448 SNPs in 4,547 cases and 4,434 controls from four different studies (**Table 1**). These SNPs were selected using a stepwise procedure; most were chosen by a hypothesis-free (agnostic) strategy, whereas approximately one-fifth were selected by alternative approaches reported in **Supplementary Methods** online or described below.

For stage 2, we first selected 22,136 SNPs that had a P value <0.05 in a logistic regression model using a two degrees-of-freedom (d.f.) score test with indicator variables for heterozygous and homozygous carriers and four continuous variables representing principal components of population stratification. We chose the 2-d.f. score test because it makes minimal assumptions for the underlying genetic model. We complemented this set of SNPs with 2,773 SNPs with $P < 0.06$ in tests of dominant, recessive or multiplicative models that were not already included by virtue of their P value in the score test (each test has 1 d.f.; see **Supplementary Methods**). In the 'agnostic' category, SNPs in strong linkage disequilibrium ($r^2 \geq 0.8$) were removed. We selected an additional 1,436 'agnostic' SNPs not included in the two previous criteria on the basis of a 2-SNP test that conditioned each SNP on a neighboring SNP, if this improved the P value relative to the single-SNP statistics by an order of magnitude. Loci previously established by GWAS were further explored with a dense set of 1,711 SNPs. We included 3,788 SNPs in candidate genes from proposed pathways or identified in an analysis of suggested interaction with variants in intron 2 of *FGFR2*. To monitor population stratification, we included 1,508 SNPs with low pairwise linkage disequilibrium[15].

A total of 30,278 SNPs (92.1%) provided genotypes according to our quality control metrics (**Supplementary Methods**). We removed women with greater than 20% admixture of non-European origin using STRUCTURE[16]. We conducted a principal component analysis (PCA) using the SNPs chosen to monitor population stratification and observed minimal evidence of population stratification; the distribution of the P values for the association statistics with a 2-d.f. test unadjusted for population heterogeneity was close to the expected distribution under the null hypothesis[17]. The inflation factor $\lambda = 1.010$ was reduced to 1.009 when the first four principal components were included as covariates. We carried out joint analysis of the genotypes[18] in the first and second stages using a multinomial regression analysis (2-d.f. test) adjusted for age, study design and population stratification.

Table 2 Results for previously reported loci

				Genotype P[a]			Combined			
Chromosome band	Proposed candidate	SNP ID[b]	Risk allele (freq.)[c]	Stage 1	Stage 2	Stage 3	Controls/ cases	Genotype P	OR het (95% CI)	OR hom (95% CI)
10q26.13	*FGFR2*	rs2981579	T (41%)	4.36×10^{-5}	1.22×10^{-6}	–	5,283/5,439	1.79×10^{-10}	1.17 (1.07–1.27)	1.46 (1.30–1.62)
16q12.1	*TOX3*	rs3803662	T (27%)	5.30×10^{-2}	6.82×10^{-9}	–	5,281/5,434	1.11×10^{-9}	1.16 (1.07–1.27)	1.55 (1.34–1.78)
5q11.2	*MAP3K1*	rs16886165	G (15%)	3.10×10^{-2}	1.17×10^{-5}	–	5,283/5,440	5.00×10^{-7}	1.23 (1.12–1.35)	1.65 (1.30–2.10)
8q24.21		rs1562430	A (57%)	1.44×10^{-2}	4.74×10^{-4}	–	5,285/5,440	1.28×10^{-5}	0.84 (0.77–0.92)	0.79 (0.71–0.89)
2q35		rs13387042	A (51%)	1.10×10^{-2}	1.48×10^{-6}	–	5,285/5,433	2.10×10^{-8}	0.80 (0.73–0.87)	0.74 (0.67–0.83)
11p15.5	*LSP1*	rs3817198	C (32%)	5.36×10^{-1}	1.16×10^{-1}	4.34×10^{-1}	10,316/9,408	6.51×10^{-2}	1.02 (0.96–1.08)	1.12 (1.02–1.23)
5p12		rs4415084	T (41%)	1.50×10^{-3}	1.60×10^{-2}	1.60×10^{-2}	10,293/9,367	4.53×10^{-5}	1.09 (1.03–1.17)	1.20 (1.11–1.31)
5p12		rs10941679	G (26%)	–	–	5.50×10^{-3}	5,490/4,575	5.50×10^{-3}	1.12 (1.03–1.22)	1.20 (1.03–1.41)

Results of genotype and trend tests (both adjusted and unadjusted) are presented in **Supplementary Table 1**.
[a]Adjusted genotype test with 2 d.f. [b]SNP ID corresponds to dbSNP ID. [c]Estimated from controls in the combined (stages 1–3) analysis.

Table 3 Newly examined SNPs

Chromosome band	Proposed candidate	SNP ID[a]	Risk allele (freq)[b]	Genotype *P*			Combined (stages 1–3)			
				Stage 1	Stage 2	Stage 3	Controls/ cases	Genotype *P*	OR het (95% CI)	OR hom (95% CI)
1p11.2		rs11249433	C (39%)	1.86×10^{-3}	1.11×10^{-3}	1.49×10^{-5}	10,263/9,335	6.74×10^{-10}	1.16 (1.09–1.24)	1.30 (1.19–1.41)
14q24.1	*RAD51L1*	rs999737	C (76%)	1.31×10^{-2}	6.18×10^{-5}	3.49×10^{-2}	10,298/9,395	1.74×10^{-7}	0.94 (0.88–0.99)	0.70 (0.62–0.80)
5p12	*MRPS30*	rs7716600	A (22%)	5.01×10^{-3}	7.66×10^{-5}	2.18×10^{-2}	10,321/9,400	2.20×10^{-5}	1.10 (1.04–1.17)	1.28 (1.13–1.45)
5p12	*MRPS30*	rs2067980	G (16%)	1.63×10^{-2}	5.75×10^{-4}	6.14×10^{-1}	10,309/9,391	1.24×10^{-3}	1.08 (1.02–1.15)	1.29 (1.09–1.52)

The two additional 5p12 markers were chosen to explore the region previously reported[13]. One SNP assay for rs930395 was not designed adequately, so a surrogate with $r^2 = 1.0$ was substituted, rs7716600.
[a]SNP ID corresponds to dbSNP ID. [b]Estimated from controls in the combined (stages 1–3) analysis.

Combining the initial scan with the second stage, we found that markers in six of the seven loci identified in previous GWAS studies were strongly associated with breast cancer risk (**Table 2**). SNPs in 2q35, 5p12, 5q11.2 (*MAP3K1*), 8q24, 10q26 (*FGFR2*) and 16q12.1 (*TOX3*) provided strong signals (**Table 2** and **Supplementary Table 1** online); in some cases, an alternative SNP to the originally reported SNP provided a smaller *P* value (see below). The lowest *P* value for a marker at 11p15.5 (*LSP1*, rs3817198) was minimally significant ($P = 3.87 \times 10^{-2}$, trend test with 1 d.f.; **Supplementary Table 1**), but its allele-specific odds ratio was similar to that reported previously (heterozygote OR = 1.04; 95% CI = 1.00–1.09; homozygote OR = 1.09; 95% CI = 1.00–1.19) in our combined three-stage analysis. For the single candidate gene variant that had previously been reported as genome-wide significant, rs1045485 in *CASP8*, the results ($P = 5.47 \times 10^{-2}$, trend test with 1 d.f.) were also consistent with previous findings (heterozygote OR = 0.96; 95% CI = 0.91–1.00; homozygote OR = 0.92; 95% CI = 0.84–1.00). After stage 2, no indication of association ($P_{2df} = 0.50$) was observed for rs2107425 in the *H19* region, previously associated at lower level of significance by Easton *et al.*[10] (reported $P_{trend} = 2 \times 10^{-5}$). A GWAS in American Ashkenazi Jewish women[19] reported a locus on chromosome 6 (rs2180341) with a minor allele frequency of 0.21 and a per-allele OR of 1.41 ($P = 3.0 \times 10^{-8}$). In CGEMS, SNP rs9398840, which is strongly correlated with rs2180341 ($r^2 = 1.0$) in the CEU HapMap population, was not significantly associated ($P_{2df} = 0.58$) and was not taken into stage 2.

Stage 3 included 4,078 cases and 5,223 controls, in which 24 SNPs were genotyped, 21 of which were chosen on the basis of a preliminary combined analysis of the first two stages (**Tables 1** and **2**). Specifically, we examined 16 promising new regions with one SNP each; these associations had the lowest *P* values of the preliminary data build. We further examined two new regions with two SNPs apiece: at 3p24.1, two SNPs (rs724244 and rs4973768) separated by 170 kb ($r^2 = 0.35$) each had low *P* values, and at 1p34.2, because of difficulty in the assay design, we selected two SNPs separated by 40 kb and in strong LD ($r^2 = 0.88$). In the region of 5p12, in which rs4415084 and rs10941679 were recently reported[13], we advanced two more SNPs (rs7716600 and rs2067980) separated by 100 kb ($r^2 = 0.50$) (**Supplementary Fig. 1** online). Thus, we explored the 5p12 region with four SNPs. For stage 3, we also added rs3817198 in *LSP1* to the set because of a previous publication[10].

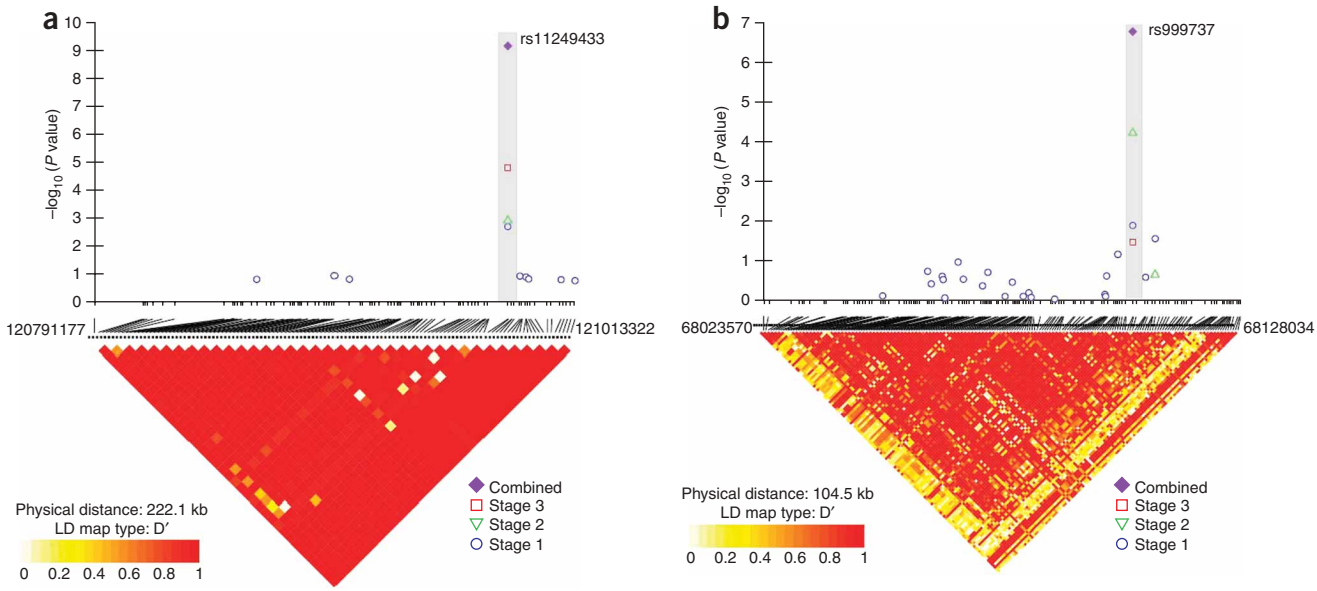

Figure 1 Linkage disequilibrium plots of two newly discovered loci. Both panels present LD plots (using *D′*) based on SNPs with MAF > 5% using HapMap stage 2 individuals of European background ($n = 60$ unrelated individuals). Above the plots are the results of the three individual stages and the combined analysis for the SNPs reaching genome-wide significance. (**a**) Chromosome 1 region marked by rs11249433 and bounded by SNPs between 120,400,700 and 121,060,765. Note that one side is closely anchored to the centromere whereas the region distal to the centromere is bounded by a SNP desert of approximately 220 kb. (**b**) Chromosome 14q24.1 region marked by rs999737. The block resides in the intron between two exons, of which the last has been observed in one of the three splice variants. The SNP is located in an intron exclusive to the longest predicted transcript of *RAD51L1*.

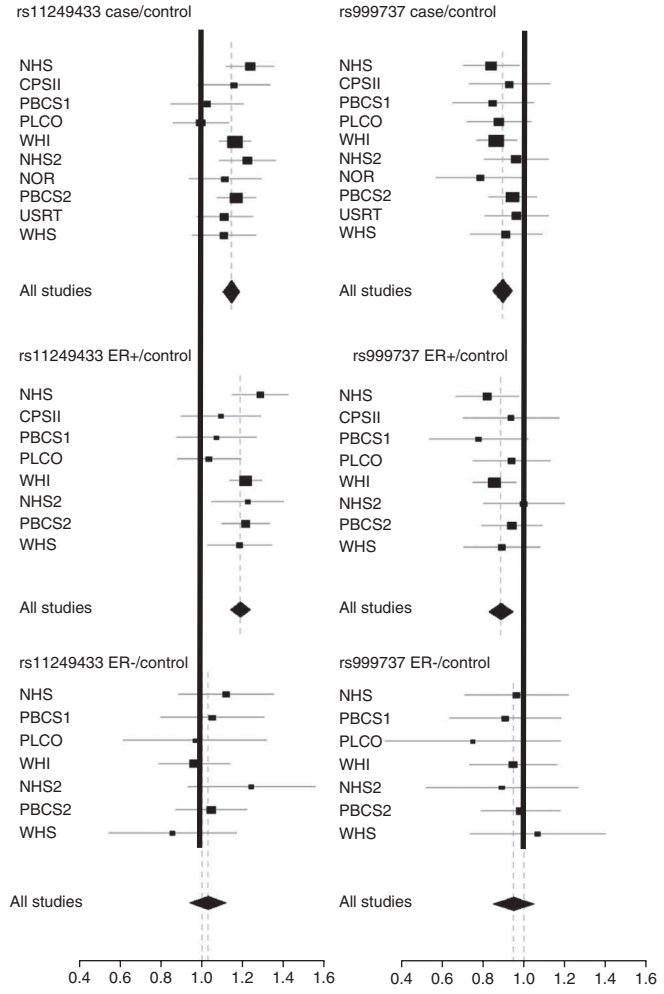

Figure 2 Forest plots for overall, ER-positive and ER-negative analyses for rs1124933 and rs999737. Results of the overall pooled analysis and case-control analyses for ER-positive and ER-negative cases were generated using a trend test with 1 d.f. Plots include per-allele odd ratios (log additive/multiplicative model) for each study. For the overall analysis, the *P* heterogeneity values are $P = 0.44$ for rs1124933 and $P = 0.79$ for rs999737. Data were available for ER status in 6,586 cases.

The results of stage 3 are notable for only four SNPs. Two previously unreported SNPs (rs11249433 in the pericentromeric region of chromosome 1, and rs999737 in *RAD51L1* (RAD51-like 1) on chromosome 14q24.1) reached genome-wide significance in the combined analysis of all three stages (**Table 3**). In addition, two of the SNPs in 5p12 (rs7716600 and rs2067980) confirmed the previously reported locus[13].

The combined joint adjusted analysis of the genome-wide scan plus two follow-up stages provide conclusive statistical significance for an association with rs11249433, located in the pericentromeric region of the short arm of chromosome 1 ($P = 6.74 \times 10^{-10}$) (**Table 3** and **Supplementary Fig. 1**). Pericentromeric regions are known to be recombination-poor, and thus it is not surprising that rs11249433 maps to a large block of linkage disequilibrium. The definition of the block is difficult to determine for two reasons: (i) its close proximity to the centromere and (ii) the presence of a SNP desert of approximately 220 kb immediately distal to the block (**Fig. 1a**). The block contains several pseudogenes and a member of the highly paralogous low-affinity Fc gamma receptor family, *FCGR1B*. Distal to the SNP desert is the promoter of *NOTCH2*, a gene recently shown to be associated with type 2 diabetes[20]. Some epidemiological studies have suggested an association between type 2 diabetes and postmenopausal breast cancer[21]. Further mapping and subsequent functional work is required to provide biological plausibility for the association signal observed with rs11249433.

The second newly discovered marker, rs999737, is in *RAD51L1* (also known as *RAD51B*) ($P = 1.74 \times 10^{-7}$), a gene on chromosome 14q24.1 in a prior candidate pathway for breast cancer susceptibility, the double-strand break repair and homologous-recombination pathway (**Table 3**). The SNP maps to a 70-kb LD block defined by two recombination hot spots entirely contained within intron 12 of the gene (**Fig. 1b** and **Supplementary Fig. 2** online). Its gene product is one of five paralogs that interact directly with the product of the *RAD51* gene, which catalyzes key reactions in homologous recombination[22]. A polymorphism in the 5′ UTR of *RAD51* has recently been identified as a genetic modifier of outcome in women with deleterious *BRCA2* mutations[23]. A copy number variant on chromosome 14q24.1 that includes *RAD51L1* has been observed repeatedly in pedigrees with Li-Fraumeni syndrome, suggesting a possible contribution of this locus to the spectrum of cancers (including breast cancer) observed in this hereditary syndrome[24]. Further work is warranted to dissect the genetic signal and investigate potential functional variants.

Tumor estrogen receptor (ER) status was available for 6,386 cases[25]. **Figure 2** shows the results of the analysis for the two newly discovered SNPs, rs11249433 (chromosome 1) and rs999737 (chromosome 14) by ER status. The association with rs11249433 is more apparent for ER-positive than ER-negative breast cancer (**Supplementary Tables 2–4** online). The observed difference was significant in a case-case comparison ($P_{trend} = 0.001$), suggesting that the chromosome 1 locus is more important in ER-positive breast cancer susceptibility. Although there was also some evidence for a stronger association with ER-positive disease for the chromosome 14 SNP, rs999737, it was not significant ($P_{trend} = 0.20$). An analysis stratified by age did not demonstrate any significant differences for the two SNPs, although it should be emphasized that most cases are postmenopausal.

Given the initial genome coverage using the Illumina HumanHap500 platform and our sample size, it is unlikely that many more common loci with relative risks comparable to *FGFR2* will be discovered among Europeans. We confirmed strong association signals for six previously reported genomic regions and identified new associations at genome-wide significance for markers on chromosome 1p11.2 and 14q24.1. In addition, although we provide supportive evidence for two loci previously associated with genome-wide significance, namely, 2p24.1 (*CASP8*) and 11p15.5 (*LSP1*), these data reinforce that very large datasets are required to identify at genome-wide significance levels loci with small estimated per-allele effect sizes. Moreover, our study suggests the value of combining scans for discovery with subsequent follow-up in large datasets[9–11].

To date, GWAS for breast cancer have been conducted among women of European ancestry, mainly with ER-positive tumors. Well-designed scans in other populations, and of women with ER-negative tumors, should yield additional loci, some of which could be population specific. The evidence for two new associations presented here pinpoints genomic regions that could elucidate previously unknown etiologic pathways contributing to the development of breast cancer. Carriage of the multiple loci reported so far, together with additional loci yet to be identified, should refine estimates of the risk of sporadic breast cancer associated with multiple inherited genetic loci, although the clinical utility of these estimates has yet to be determined[26,27].

METHODS

Initial genotyping for genome-wide scan. Briefly, this study reports the follow-up genotyping of studies based on the previously reported genome-wide scan conducted in the prospective Nurses' Health Study using the Human Hap500 Infinium Assay (Illumina) in 1,145 cases of women with postmenopausal breast cancer and 1,142 controls[11]. The details are reported elsewhere[11]. Quality control metrics included removal of samples with call rates under 90% and SNP assays with call rates under 95%. Subjects with more than 15% admixture of non-European background were removed from the analysis.

Replication samples. In stage 2, we genotyped 30,278 SNPs in four follow-up studies of women of European background with breast cancer totaling 4,547 cases and 4,434 controls drawn from the American Cancer Society Cancer Prevention Study II, the Prostate, Lung, Colon and Ovarian Screening Trial, part of the available Polish Breast Cancer Study, and the observational arm of the Women's Health Initiative. In stage 3, we genotyped 24 SNPs in 4,078 cases of breast cancer in women of European background and 5,223 controls drawn from the CONOR Norwegian cohort, the remaining cases and controls of the Polish Breast Cancer Study, the US Radiologic Technologists Study, the Nurses' Health Study II and the Women's Health Study. These studies were approved by the appropriate institutional review boards, and informed consent was obtained from all subjects.

Replication genotyping. In stages 2 and 3, we genotyped 18,282 unique subjects (excluding validation samples and study duplicates) passing sample handling quality control metrics in the Core Genotyping Facility of the National Cancer Institute. For NHS II and WHS, the 24 SNPs of stage 3 were genotyped at the DF/HCC Genotyping Core at the Harvard School of Public Health. Stage 2 samples were genotyped using a custom-designed iSelect assay from Illumina with content described above; 9,804 samples were attempted (including known duplicates). Using quality control measures, we removed samples with call rates under 90% and SNPs with call rates less than 95%. Fitness for Hardy-Weinberg proportion was assessed for each SNP in unique control subjects only but was not used to exclude SNP assays (see **Supplementary Methods**). In Stage 3, we genotyped 9,301 unique subjects for 24 TaqMan assays (ABI) selected by the criteria described above using custom-designed assays that were subsequently optimized in the SNP500 Cancer initiative.

A small fraction (less than 2%) of subjects who were successfully genotyped in stage 2 were excluded from analysis because of one or more of the following reasons: (i) unanticipated interstudy or intrastudy duplicates; (ii) unanticipated non-European admixture of greater than 20% (for example, African or East Asian; notably, in stage 1, the threshold for non-European admixture was 15%); or (iii) incomplete covariate data.

In stage 2, a total of 16,715 discordant genotypes were detected out of a possible 7,255,923 genotype comparisons (237 duplicate pairs and one triplicate), yielding a discordance rate of 0.23%. Infinium cluster plots for notable SNPs are included in **Supplementary Methods**.

For the 24 SNPs analyzed in stage 3, we validated genotype calls determined by Infinium HumanHap500 and custom iSelect assay by comparing TaqMan results in the entire Polish Breast Cancer Study. For the 1,110 samples genotyped with both platforms, the overall concordance rate was 99.52% (see **Supplementary Methods**).

The individual genotype data for the stage 1 CGEMS GWAS in 1,145 cases and 1,142 controls, and the aggregate data for stages 1, 2 and 3, are available to researchers registered after approval by the NCI Data Access Committee (DAC) through the CGEMS portal (see URLs section below).

Analysis. For the follow-up replication studies, all single-SNP analyses were conducted using unconditional logistic regression, adjusted for age in ten-year intervals and study. For stages 1 and 2, four continuous covariates were included to account for population heterogeneity based on principal component analysis of genotype correlations. Separate analyses were conducted according to the individual studies, the pooled replication studies in stage 2 and stage 3 and for all studies combined. Genotype effects were modeled individually, and a single-SNP score test with 2 d.f. was computed. To enable comparison with other published GWAS, we also conducted a Cochran-Armitage trend test. To explore a possible difference in effect between ER-positive and ER-negative breast cancer, we conducted separate analyses for ER-positive and ER-negative cases, using a trend test with 1 d.f.

Informatics. We used GLU (Genotyping Library and Utilities version 1.0), a suite of tools available as an open-source application for management, storage and analysis of GWAS data. STRUCTURE and EIGENSTRAT programs were used to assess population heterogeneity (see URLs below).

URLs. CGEMS portal, http://cgems.cancer.gov/; CGF, http://cgf.nci.nih.gov/; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; GLU, http://code.google.com/p/glu-genetics/; SNP500Cancer, http://snp500cancer.nci.nih.gov/; STRUCTURE, http://pritch.bsd.uchicago.edu/structure.html; Tagzilla, http://tagzilla.nci.nih.gov/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The Nurses' Health Studies are supported by US National Institutes of Health grants CA65725, CA87969, CA49449, CA67262, CA50385 and 5UO1CA098233. We thank B. Egan, L. Egan, H. Judge Ellis, H. Ranu and P. Soule for assistance, and the participants in the Nurses' Health Studies. The WHI program is supported by contracts from the National Heart, Lung and Blood Institute, NIH. We thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at http://www.whi.org. The ACS study is supported by UO1 CA098710. We thank C. Lichtman for data management and the participants on the CPS-II. The US Radiologic Technologists Study (USRT) is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS. The PLCO study is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. We thank P. Prorok, Division of Cancer Prevention, National Cancer Institute, the Screening Center investigators and staff of the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), and T. Sheehy and staff at SAIC-Frederick. We acknowledge the study participants for their contributions to making this study possible. We thank the radiologic technologists who participated in the study; J. Reid of the American Registry of Radiologic Technologists for continued support of the study; D. Kampa and A. Iwan of the University of Minnesota for study coordination and data collection; B. Kopp and staff at SAIC-Frederick for biospecimen processing; and L. Bowen of Information Management Systems for data management.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Colditz, G.A., Baer, H.J. & Tamimi, R.M. in *Cancer Epidemiology and Prevention* (eds. Schottenfeld, D. & Fraumeni, J.F.) Breast Cancer, 995–1012 (Oxford University Press, New York, 2006).
2. Miki, Y. *et al.* A strong candidate for the breast and ovarian-cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).
3. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378**, 789–792 (1995).
4. Rahman, N. *et al. PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
5. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nat. Genet.* **31**, 55–59 (2002).
6. Erkko, H. *et al.* A recurrent mutation in *PALB2* in Finnish cancer families. *Nature* **446**, 316–319 (2007).
7. Renwick, A. *et al. ATM* mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
8. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38**, 1239–1241 (2006).
9. Cox, A. *et al.* A common coding variant in *CASP8* is associated with breast cancer risk. *Nat. Genet.* **39**, 352–358 (2007).
10. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
11. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).

12. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).
13. Stacey, S.N. *et al.* Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **40**, 703–706 (2008).
14. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
15. Yu, K. *et al.* Population substructure and control selection in genome-wide association studies. *PLoS ONE* **3**, e2551 (2008).
16. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
17. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
18. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
19. Gold, B. *et al.* Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc. Natl. Acad. Sci. USA* **105**, 4340–4345 (2008).
20. Staiger, H. *et al.* Novel meta-analysis-derived type 2 diabetes risk loci do not determine prediabetic phenotypes. *PLoS ONE* **3**, e3019 (2008).
21. Xue, F. & Michels, K.B. Diabetes, metabolic syndrome, and breast cancer: a review of the current evidence. *Am. J. Clin. Nutr.* **86**, s823–s835 (2007).
22. Li, X. & Heyer, W.D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res.* **18**, 99–113 (2008).
23. Antoniou, A.C. *et al. RAD51* 135G→C modifies breast cancer risk among *BRCA2* mutation carriers: results from a combined analysis of 19 studies. *Am. J. Hum. Genet.* **81**, 1186–1200 (2007).
24. Shlien, A. *et al.* Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc. Natl. Acad. Sci. USA* **105**, 11264–11269 (2008).
25. Garcia-Closas, M. *et al.* Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.* **4**, e1000054 (2008).
26. Pharoah, P.D., Antoniou, A.C., Easton, D.F. & Ponder, B.A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
27. Pepe, M.S. & Janes, H.E. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J. Natl. Cancer Inst.* **100**, 978–979 (2008).