nature genetics

A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer

David J Hunter[1–4], Peter Kraft[2], Kevin B Jacobs[5], David G Cox[1,2], Meredith Yeager[4,6], Susan E Hankinson[1], Sholom Wacholder[4], Zhaoming Wang[4,6], Robert Welch[4,6], Amy Hutchinson[4,6], Junwen Wang[4,6], Kai Yu[4], Nilanjan Chatterjee[4], Nick Orr[7], Walter C Willett[1,8], Graham A Colditz[9], Regina G Ziegler[4], Christine D Berg[10], Saundra S Buys[11], Catherine A McCarty[12], Heather Spencer Feigelson[13], Eugenia E Calle[13], Michael J Thun[13], Richard B Hayes[4], Margaret Tucker[4], Daniela S Gerhard[14], Joseph F Fraumeni, Jr[4], Robert N Hoover[4], Gilles Thomas[4] & Stephen J Chanock[4,7]

We conducted a genome-wide association study (GWAS) of breast cancer by genotyping 528,173 SNPs in 1,145 postmenopausal women of European ancestry with invasive breast cancer and 1,142 controls. We identified four SNPs in intron 2 of FGFR2 (which encodes a receptor tyrosine kinase and is amplified or overexpressed in some breast cancers) that were highly associated with breast cancer and confirmed this association in 1,776 affected individuals and 2,072 controls from three additional studies. Across the four studies, the association with all four SNPs was highly statistically significant (P_{trend} for the most strongly associated SNP (rs1219648) = 1.1×10^{-10}; population attributable risk = 16%). Four SNPs at other loci most strongly associated with breast cancer in the initial GWAS were not associated in the replication studies. Our summary results from the GWAS are available online in a form that should speed the identification of additional risk loci.

Family history is an established risk factor for breast cancer, yet estimates of the inherited component of the disease are uncertain[1]. Investigation of multiple-case families in which breast cancer segregates with mendelian patterns of inheritance led to the identification of the tumor suppressor genes *BRCA1* and *BRCA2*, which account for a substantial proportion of early-onset breast cancer but a much smaller proportion of late-onset disease[2,3]. Most late-onset cases occur in the absence of a first-degree family history of breast cancer and are often called 'sporadic' cases. In the past, family-based studies have been the primary focus of study in the search for genetic determinants, but with new technologies that enable analysis of hundreds of thousands of SNPs, together with new insights into the structure of variation in the human genome, it is now possible to scan the genome in an agnostic manner in studies of unrelated cases and controls in search of common genetic variants associated with disease risk[4]. One strategy for conducting a GWAS is to analyze early-onset cases, often enriched with cases with a positive family history of the disease, in order to maximize the opportunity to detect inherited causal variants. Already, two such studies, one of diabetes and one of breast cancer, have successfully identified and replicated associations of common genetic variants with these diseases[5,6]. Alternatively, GWA analysis of older subjects may identify common genetic variants associated with sporadic disease, as has been successfully demonstrated for prostate cancer[7,8].

We initially genotyped 1,183 women with postmenopausal invasive breast cancer and 1,185 individually matched controls from the Nurses' Health Study (NHS) cohort[9] using the Illumina HumanHap500 array, as part of the National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMS) Project. Affected individuals were identified from a consecutive series of postmenopausal women, unselected for any other characteristics, who were among the 32,826 cohort members 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. Controls were postmenopausal women who were matched with cases by year of birth and post-menopausal hormone use at blood

[1]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [3]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), US National Institutes of Health (NIH), Department of Health and Human Services (DHHS), Bethesda, Maryland 20892, USA. [5]Bioinformed Consulting Services, Gaithersburg, Maryland 20877, USA. [6]SAIC-Frederick, NCI-FCRDC, Frederick, Maryland 21702, USA. [7]Pediatric Oncology Branch, Center for Cancer Research, NCI, NIH, DHHS, Bethesda, Maryland 20892, USA. [8]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [9]Washington University School of Medicine, St. Louis, Missouri 63130, USA. [10]Division of Cancer Prevention, NCI, NIH, DHHS, Bethesda, Maryland 20892, USA. [11]Department of Internal Medicine, University of Utah, Salt Lake City, Utah 84112, USA. [12]The Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin 54449, USA. [13]Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Georgia 30329, USA. [14]Office of Cancer Genomics, NCI, NIH, DHHS, Bethesda, Maryland 20892, USA. Correspondence should be addressed to D.J.H. (dhunter@hsph.harvard.edu).

Received 26 April; accepted 18 May; published online 27 May 2007; doi:10.1038/ng2075

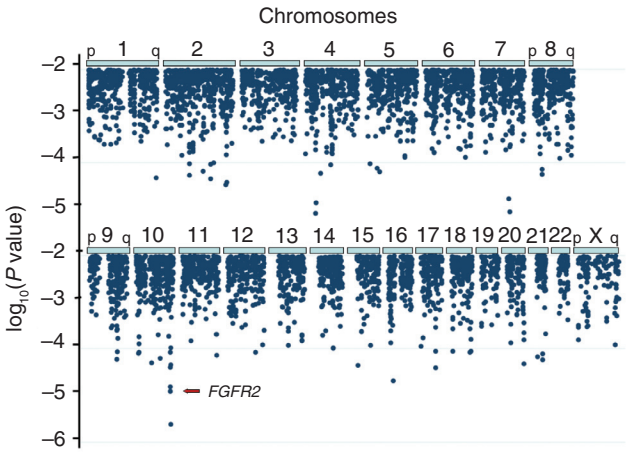

Figure 1 Summary of GWAS results by chromosome. Association with breast cancer was determined for 528,173 SNPs among 1,145 women with postmenopausal breast cancer and 1,142 controls. The *x* axis represents position on each chromosome from p terminus (left) to q terminus (right); the *y* axis shows the *P* value on a logarithmic scale. Only *P* values < 10^{-2} are shown.

draw and who were not diagnosed with breast cancer during follow-up. All cases and controls were self-described as being of European ancestry. We removed 59 samples (30 cases and 29 controls) from the analysis because of completion rates of less than 90% for the 528,173 SNPs that passed quality control. We removed an additional 18 samples (5 cases and 13 controls) because of unclear identity (or possible contamination) and removed four samples (three cases and one control) because of evidence of intercontinental admixture (**Supplementary Fig. 1** online). Thus, we performed GWA analysis on 1,145 affected individuals and 1,142 controls.

We analyzed each locus in a logistic regression model using a two–degree of freedom score test with indicator variables for heterozygous and homozygote carriers of the variant allele (for rare variants, we collapsed heterozygote and homozygote carrier categories; see **Supplementary Methods** online). The distribution of the observed *P* values does not show any suggestion of distortion due to population stratification or other sources of bias or due to distortion of Type I error rates (**Supplementary Fig. 2** online). When we adjusted for matching factors and the top three principal components from an analysis of genetic covariance, the overall distribution of *P* values did not change significantly (Komolgorov-Smirnov, $P = 0.34$). All loci with unadjusted $P < 2 \times 10^{-5}$ maintained this level of significance after adjustment. This suggests that these associations are not due to population stratification.

The GWAS identified several genomic locations as potentially associated with breast cancer (**Fig. 1**). Of 528,173 SNPs tested, two of the most significant *P* values (rs1219648 and rs2420946; **Table 1**) were in intron 2 (**Fig. 2**) of *FGFR2*, encoding a receptor tyrosine kinase previously shown to be important in mammary gland development and neoplasia[10]; an additional two SNPs in *FGFR2* (rs11200014 and rs2981579) were among the 16 most extreme *P* values from the unadjusted analysis. Modeling all pairwise combinations of the four SNPs and their interactions, as well as haplotypes of the four SNPs, suggested that all four were similar with respect to their association with breast cancer risk, consistent with the very high degree of linkage disequilibrium (each pairwise $D' > 0.95$ and $r^2 > 0.84$). None of the other 31 SNPs at the *FGFR2* locus was in strong linkage disequilibrium (r^2) with the SNPs in intron 2, and none was associated with breast cancer risk (**Fig. 2**). Computational analysis of haplotypes (see Methods) indicated four common haplotypes; the AAGT haplotype (the most common risk haplotype) was present in 43.6% of chromosomes from affected individuals and 36.9% of control chromosomes in the NHS (**Table 2**).

To further explore the association signal observed for *FGFR2* in the NHS, we performed analyses using inferred ancestral recombination graphs (ARGs) (**Supplementary Methods**). This involves estimating a simple approximation to the distribution of possible genealogies relating the haplotypes of the affected individuals and controls. Using the Margarita program[11], we inferred ARGs for 81-SNP haplotypes spanning *FGFR2* and its flanking regions (from position 123225862 to position 123471190 on NCBI build35, as shown in **Supplementary Fig. 3** online). For every ARG, a putative risk mutation was placed on the marginal genealogy at each SNP position by maximizing the association between the mutation and disease status. We evaluated the significance of this observed association using a maximum of 10^6 permutations on the phenotypes. The permutation *P* value was consistently higher than 0.05 over the entire region, with the exception of the 20-kb segment located between 123.32 Mb and 123.34 Mb in intron 2 of *FGFR2*

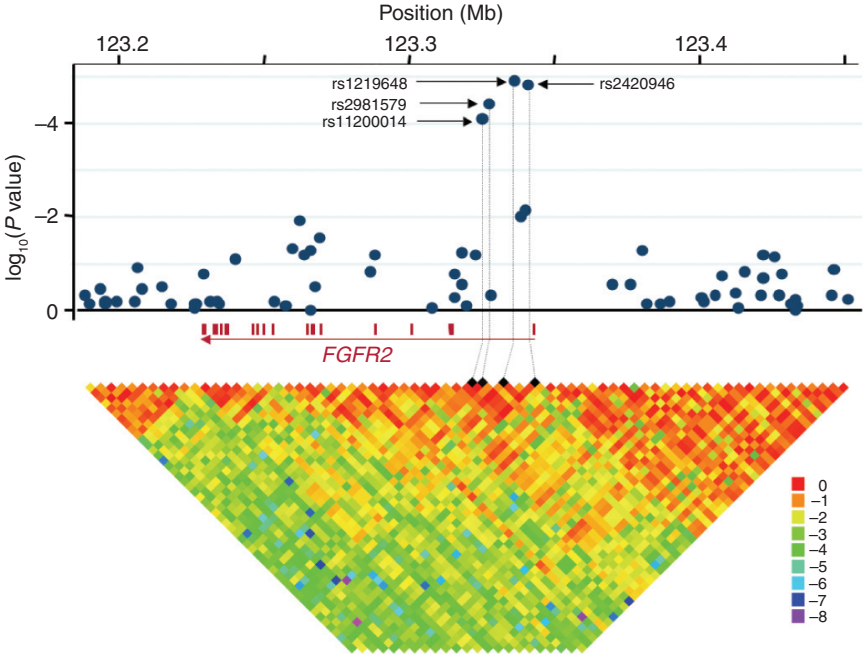

Figure 2 Association analysis of SNPs across *FGFR2*. The upper panel shows *P* values for association testing drawn from the GWAS covering *FGFR2* and 100 kb 5′ upstream of it. The analysis was based on the two–degree of freedom test corrected for age and the three first principal components of population stratification (**Supplementary Methods**). The lower panel shows estimates of the square of the correlation coefficient (r^2) calculated for each pairwise comparison of SNPs. The $\log_{10}(r^2)$ values are color coded according to the scale at the right. The four filled black diamonds indicate the four SNPs most strongly associated with breast cancer risk.

Table 1 Six SNPs with the smallest *P* values of the 528,173 tested for association among 1,145 cases of postmenopausal invasive breast cancer and 1,142 controls

SNP, chromosome	Location[a]	Gene, MAF[b]	OR_{het}	OR_{hom}	OR_{trend}	P_{2df}[c]	P_{trend}[c]
rs10510126, 10q	124992475	0.13	0.59 (0.48,0.72)	0.59 (0.26,1.34)	0.62 (0.51,0.75)	2.4×10^{-6}	7.1×10^{-7}
rs12505080, 4p	37171906	0.27	1.22 (1.02,1.45)	0.51 (0.35,0.73)	0.93 (0.81,1.06)	8.1×10^{-6}	0.30
rs17157903, 7q	103221987	*RELN*, 0.12	1.60 (1.31,1.95)	0.77 (0.42,1.41)	1.35 (1.14,1.60)	8.8×10^{-6}	0.00060
rs1219648, 10q	123336180	*FGFR2*, 0.39	1.23 (1.02,1.48)	1.79 (1.41,2.28)	1.32 (1.17,1.49)	1.2×10^{-5}	3.2×10^{-6}
rs7696175, 4p	38643552	*TLR1*, *TLR6*, 0.44	1.39 (1.15,1.68)	0.86 (0.67,1.09)	0.98 (0.87,1.10)	1.5×10^{-5}	0.68
rs2420946, 10q	123341314	*FGFR2*, 0.38	1.24 (1.03,1.50)	1.79 (1.40,2.28)	1.32 (1.17,1.49)	1.5×10^{-5}	3.5×10^{-6}

[a]From NCBI genome build 35. [b]Minor allele frequency among Nurses' Health Study controls. [c]From analyses adjusting for age, matching factors (see Methods) and three eigenvectors of the principal components identified by Eigenstrat. P_{2df} values were obtained by a score test with two degrees of freedom (df).

(**Supplementary Fig. 3**). In this segment, all marginal trees demonstrated association with a *P* value lower than 3×10^{-3}. The permutation *P* values for four notable SNPs were all smaller than 2×10^{-5}, and the frequency of the inferred mutation was similar for all four SNP positions (**Supplementary Table 1** online). These results suggest that there is a single risk locus in the *FGFR2* region.

Table 2 Haplotypes for four SNPs in intron 2 of *FGFR2* and association with breast cancer risk

Haplotype[a]	Cases	Controls	OR	95% c.i.	*P*
Nurses' Health Study					
G-G-A-C	1,195	1,348	1.0		
A-A-G-T	998	842	1.33	1.18–1.50	3×10^{-6}
A-A-A-C	40	47	0.96	0.62–1.48	0.85
A-A-G-C	24	19	1.40	0.76–2.57	0.28
Rare <1%	33	28	1.36	0.81–2.29	0.24
Nurses' Health Study 2					
G-G-A-C	295	667	1.0		
A-A-G-T	276	474	1.34	1.09–1.65	0.0061
A-A-A-C	7	21	0.76	0.30–1.90	0.55
A-A-G-C	13	20	1.50	0.70–3.21	0.30
Rare <1%	12	16	1.96	0.90–4.28	0.09
PLCO study					
G-G-A-C	994	1,140	1.0		
A-A-G-T	795	728	1.13	0.99–1.29	0.064
A-A-A-C	32	24	1.44	0.83–2.47	0.19
A-A-G-C	11	24	0.47	0.23–0.97	0.041
Rare <1%	21	34	0.63	0.36–1.10	0.11
ACS CPS-II					
G-G-A-C	583	664	1.0		
A-A-G-T	482	406	1.38	1.16–1.65	0.00040
A-A-A-C	21	22	1.07	0.59–1.94	0.82
A-A-G-C	6	10	0.67	0.23–1.89	0.45
Rare <1%	18	10	1.98	0.91–4.31	0.084
Pooled across studies					
G-G-A-C	3,068	3,718	1.0		
A-A-G-T	2,551	2,450	1.26	1.17–1.35	6×10^{-10}
A-A-A-C	102	114	1.09	0.83–1.43	0.55
A-A-G-C	54	74	0.88	0.61–1.27	0.50
Rare<1%	107	151	0.87	0.68–1.12	0.28

[a]For each study, haplotypes are indicated (from top to bottom) for SNPs rs11200014, rs2981579, rs1219648 and rs2420946, respectively.

As the second stage of a multistage design, we plan to genotype the SNPs with the most extreme *P* values corresponding to at least 5% of the SNPs in the initial GWAS. In the present study, for the six most significant SNPs in the GWAS (two in *FGFR2*, four at other loci) and for the two additional SNPs that appeared to define the *FGFR2* risk haplotype, we attempted to replicate the initial associations in the GWAS in an additional 1,776 affected individuals and 2,072 controls from breast cancer case-control studies nested in three prospective cohorts: the Nurses' Health Study 2 (NHS2)[12]; the Prostate, Lung, Colorectal and Ovary Cancer Screening Trial (PLCO) Cohort[13] and the American Cancer Society Cancer Prevention Study-II (CPS-II)[14] (**Table 3**). For the SNP in *FGFR2* most strongly associated with breast cancer in the GWAS (rs1219648), the pooled *P* value across all four studies was 4.2×10^{-10} for the two–degree of freedom model and 1.1×10^{-10} for the Cochran-Armitage test for trend. Both *P* values are lower than a threshold for genome-wide significance based on the conservative Bonferroni correction for 528,173 tests with a nominal $\alpha = 0.05$. There was no statistical evidence of heterogeneity of the genotype-specific odds ratios across the studies. The pooled estimate of the odds ratios across all studies, compared with homozygotes for the wild-type alleles, was 1.20 (95% confidence interval (c.i.), 1.07–1.42) for heterozygotes, and 1.64 (95% c.i., 1.42–1.90) for homozygotes for the variant alleles. Across the four studies, the SNP with the strongest association (rs1219648) was associated with a population attributable risk (PAR) of 16%[15]. In each of the three replication studies, and in all studies combined, the AAGT haplotype was the only common haplotype significantly associated with risk of breast cancer (**Table 2**). The associations

Table 3 Association of rs1219648 in the NHS, NHS2, PLCO, ACS CPS-II and pooled across studies

	Allele frequency (%)					
Study population (cases/controls)	Cases	Controls	OR_{het} (95% c.i.)	OR_{homo} (95% c.i.)	P_{2df}	P_{trend}
NHS (1,145/1,142)	45.5	38.5	1.24 (1.03–1.49)	1.81 (1.42–2.30)	8×10^{-6}	2.0×10^{-6}
NHS2 (302/594)	48.2	40.6	1.28 (0.92–1.76)	1.92 (1.28–2.87)	0.007	0.002
PLCO (919/922)	44.5	41.5	1.07 (0.87–1.23)	1.23 (0.95–1.60)	0.29	0.13
ACS CPS-II (555/556)	45.0	37.4	1.30 (1.40–2.92)	2.02 (1.40–2.92)	9×10^{-4}	0.0002
Pooled estimates (2,921/3,214)			1.20 (1.07–1.42)	1.64 (1.42–1.90)	4×10^{-10}	1.1×10^{-10}

were not significantly different across categories of age at diagnosis of breast cancer. In the NHS2, a study involving mainly premenopausal women, the associations were equivalent to those in the other three studies comprising postmenopausal cases. None of the four SNPs at other chromosomal loci was associated with increased risk in the pooled replication studies (**Supplementary Table 2** online).

FGFR2 is a tumor suppressor gene that is amplified and overexpressed in breast cancer[10,16]. Furthermore, alternatively spliced variants of this gene result in differential signal transduction and transformation of mammary epithelial cell lines. Further work is needed to identify the causal variant at this locus.

A large, three-stage GWAS of breast cancer using the Perlegen platform as the initial genome scan has identified SNPs in *FGFR2* as the strongest of its reported associations[6]. These SNPs were also in intron 2; their association was originally detected with rs2981582, which has an r^2 of 1.0 with rs1219648 and rs2420946, r^2 of 0.97 with rs2981579 and r^2 of 0.96 with rs11200014 in the HapMap CEU samples, which indicates that we have detected essentially the same association. That study used genotypes from 390 breast cancer cases under age 60, selected to have a strong family history or bilaterality of breast cancer, in their genome-wide first stage. Our study focused on later-onset, 'sporadic' cases. Such consecutive series of cases may be more easily obtained for many diseases and may be more generalizable to the most common forms of the disease, so it is reassuring that the use of unselected cases resulted in the identification of the same principal locus.

We focused on the most highly statistically significant associations from our GWAS, identifying variants in *FGFR2* as reproducibly associated with breast cancer. Because a subset of true associations would be weakly associated with outcome in any given GWAS, large-scale replication is necessary for confirmation, and some true associations may be missed if they are not carried forward into replication studies[17]. Multistage designs in which potentially associated SNPs from the first stage are carried into additional studies are an economical, scientifically sound approach to cope with the present cost of high-throughput genotyping[18]. In this regard, the precomputed rankings and *P* values for all the SNPs included in the GWAS conducted in the NHS are freely available from our website (http://cgems.cancer.gov) for others to use in subsequent studies of women with breast cancer.

METHODS

Study populations. For detailed descriptions of the component studies, see **Supplementary Methods**. The study protocol was approved by the Institutional Review Board of the Brigham and Women's Hospital. Informed consent was obtained from all patients.

Genotyping and quality control for NHS. DNA samples were received from the NHS biological repository and visually inspected for adequate fluid in individual tubes. Three measurements of quantification were performed according to the standard procedures at the Core Genotyping Facility of the National Cancer Institute[7], which include pico-green analysis, optical density spectrophotometry and real-time PCR (see URL below). Samples were also analyzed for 15 short tandem repeats and the Amelogenin marker in the Identifiler Assay (ABI). All samples that advanced to the genotype analysis step were successfully called at no fewer than 13 of the 15 microsatellite markers. After final review and sample handling, a total of 1,183 DNA samples from affected individuals and 1,185 DNA samples from controls were selected for genotyping in CGEMS. Ninety-three DNAs were aliquoted twice, and five DNAs were aliquoted three times, resulting in the addition of 103 redundant DNAs from the NHS used for quality control. Finally, 23 external quality control DNA samples were added. Thus, genotyping was attempted on a total of 2,494 DNA samples. Genotyping of the CGEMS Breast Cancer Study was performed at the NCI Core Genotyping Facility using the Sentrix HumanHap550 genotyping assay according to the manufacturer's protocol.

Initial assessment of sample completion rates. A total of 555,352 SNP genotype assays were attempted on the 2,494 DNA samples using the Illumina HumanHap550 chip. Whenever the completion rate for a sample was below 90%, the sample was assayed a second time. Samples that did not meet the 90% completion threshold after a second attempt were excluded from further analysis. We excluded 59 samples from NHS (30 cases and 29 controls) from further analysis based on these criteria, which left 2,435 DNAs for the subsequent analyses.

Assessment of SNP call rates. A total of 8,706 SNPs (~1.57% overall) failed to provide accurate genotype results owing to either a lack of calls or low call rates (<90%). We performed further quality control analysis on the remaining 546,646 SNPs. An additional 18,473 SNPs with an observed low minor allele frequency (MAF) (<1%) were dropped from the association analysis; thus, 528,173 SNPs (95.1%) were maintained in the subsequent analyses. The genotyping of the SNPs with high call rate on the 2,412 NHS DNAs with high completion rate generated 1.27 billion genotype calls. For this set of SNPs and samples, the percentage of missing data was <1% (see **Supplementary Methods** for a table of completion rates).

Concordance rate. The genotype concordance rate for SNP assays was evaluated using the 93 pairs of known duplicated DNAs from the NHS. These pairs of DNAs were separate aliquots from the same DNA preparation; all met quality control criteria required for the other DNAs, thereby providing reliable data for comparison. Analysis of the discrepancies within these pairs of DNA uncovered results similar to those of the Centre d'Etude du Polymorphisme Humain (CEPH) DNA duplicates reported in the prostate cancer CGEMS GWAS[7]. We observed an average concordance rate of 99.985% (50,820,003 concordant genotype calls out of 50,827,468 comparisons). We did not remove any SNPs or DNA from the search for association as a result of this analysis.

Deviation from Hardy–Weinberg proportions in control DNA. Genotype data for all SNPs were tested for deviation from Hardy-Weinberg proportions; the analysis was conducted in the NHS control group. We observed significant deviations for 26,476 SNPs (5.01% of 528,173 SNPs) at the level of $P = 0.05$ and for 6,778 SNPs (1.28%) at $P = 0.01$. However, we did not exclude any of these SNPs from analysis, as the tests for association applied to such data are valid in the presence of departure from Hardy-Weinberg proportions (though with potentially reduced power when these deviations are due to systematic genotyping errors with equal effects among affected individuals and controls).

Final sample selection for association analysis. For all DNA samples, the frequency of heterozygote loci on the X chromosome was compatible with a female origin. We excluded 18 DNA samples (5 affected individuals, 13 controls) with unclear identity (they could not be mapped back unambiguously to previous

genotype results from these samples). Subsequent inspection of the genotype concordance rate between pairs of DNA samples did not disclose unexpected duplicates. Finally, based on two analyses with two independent sets of 7,050 and 7,061 randomly selected SNPs with very low linkage disequilibrium ($r^2 < 0.01$) using the STRUCTURE program[19], four subjects (three affected individuals and one control) were estimated to be of admixed origin with >15% either Asian or West African ancestry. These four subjects were also removed from subsequent analyses. Thus, the search for associations was performed on a final set of 2,287 unique subjects that included 1,145 affected individuals and 1,142 controls.

Statistical analysis. For the initial scan in the NHS, we analyzed each locus using logistic regression. When the rare homozygote genotype was observed more than 15 times, we regressed disease status on indicator variables for heterozygous and homozygote carriers of the variant allele. Otherwise, the rare homozygote genotype was collapsed with the heterozygote genotype, and we regressed disease status on variant-allele carrier status. This aggregation of genotypes was performed for 64,589 SNPs. We calculated unadjusted score tests for genetic association as well as two adjusted score tests. The first adjusted test controlled for age categories (ages <55, 55–59, 60–64, 65–69, 70–74 and >74) and hormone replacement therapy use. The second controlled for age, hormone replacement therapy use and three eigenvectors from the principal component analysis of genetic covariance. The latter were included in the logistic regression as continuous covariates.

We assessed the importance of a second SNP conditional on another SNP using nested likelihood ratio tests. We compared the logistic regression model with genotypic indicator variables for the SNP to the model with indicator variables for each multilocus genotype. None of the four *FGFR2* SNPs showed any evidence of association with risk of breast cancer after adjusting for any of the other three *FGFR2* SNPs.

For the four *FGFR2* SNPs, haplotype frequencies and expected haplotype counts for each individual were estimated using a simple expectation-maximization algorithm (implemented in SAS PROC HAPLOTYPE). Haplotype association analyses were performed using the expectation-substitution technique[20].

Assessment of population stratification. Two independent sets of 7,050 and 7,061 SNPs with very low linkage disequilibrium ($r^2 < 0.01$) were analyzed using the STRUCTURE[19] program to determine if subjects had an admixed origin with >15% either Asian or West African ancestry (based on HapMap II data)[21,22]. The pooled case and control DNAs were analyzed using a set of 14,111 SNPs with very low pairwise linkage disequilibrium ($r^2 < 0.01$) using the procedure described in ref. 23. Testing for significance using the Tracy-Widom statistics[23] showed four significant principal components at the level of $P < 0.05$. Inspection of the distribution of the DNAs in the space defined by these components showed little difference between affected individuals and controls. Nevertheless, borderline statistically significant differences in this distribution for local groups in the space defined by the first three components led us to retain these components in the statistical analysis. We did not observe any difference with the fourth and higher components, which we did not retain in the analysis.

Genotyping in replication studies. The same TaqMan assays developed for rs10510126, rs1219648, rs17157903, rs2420946, rs7696175, rs12505080, rs11200014 and rs2981579 were performed at the NCI Core Genotyping Facility and at the Dana Farber/Harvard Cancer Center Polymorphism Detection Core (primers and probe sequences are available on request).

Statistical analysis of replication studies. For the individual replication studies, we used unconditional logistic regression to fit codominant and additive genetic risk models. For pooled analyses of multiple studies, we used unconditional logistic regression with separate baseline odds parameters for each study. We also adjusted for age in 5-year intervals. Effect modification by age at diagnosis (comparing >65 years versus <65 years or >55 years versus <55 years) and by menopausal status at diagnosis was assessed using nested logistic regression models. For these analyses, controls' ages were set to the ages at diagnosis of the matched cases (NHS, NHS2, ACS) or age at censoring (PLCO). We calculated PAR using the method of ref. 15.

URLs. CGEMS Project: http://cgems.cancer.gov; HapMap: http://hapmap.org/; NHS: http://nurseshealthstudy.org; standard procedures at the Core Genotyping Facility of the NCI: http://cgf.nci.nih.gov/dnaquant.cfm.

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

ACKNOWLEDGMENTS

We thank B. Egan, L. Egan, H. Judge Ellis, H. Ranu and P. Soule for assistance, and we thank the participants in the Nurses' Health Studies. We thank P. Prorok (Division of Cancer Prevention, National Cancer Institute); the Screening Center investigators and staff of PLCO; T. Riley, C. Williams and staff (Information Management Services, Inc.); B. O'Brien and staff (Westat, Inc.) and B. Kopp, T. Sheehy and staff (SAIC-Frederick). We acknowledge the study participants for their contributions in making this study possible. We thank C. Lichtman for data management and the participants on the CPS-II. We thank M. Minichiello for providing the Margarita program and for discussions. We acknowledge D. Easton and colleagues for sharing prepublication results. The Nurses' Health Studies are supported by US NIH grants CA65725, CA87969, CA49449, CA67262, CA50385 and 5UO1CA098233. The ACS study is supported by UO1 CA098710. The PLCO study is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

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

1. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).
2. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with *BRCA1* or *BRCA2* mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* **72**, 1117–1130 (2003).
3. Risch, H.A. *et al.* Population *BRCA1* and *BRCA2* mutation frequencies and cancer penetrances: a kin-cohort study in Ontario, Canada. *J. Natl. Cancer Inst.* **98**, 1694–1706 (2006).
4. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
5. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
6. Easton, D.F. *et al.* A genome-wide association study identifies multiple breast cancer susceptibility loci. *Nature*, advance online publication 27 May 2007 (doi:10.1038/nature05887).
7. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
8. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
9. Tworoger, S.S., Eliassen, A.H., Sluss, P. & Hankinson, S.E. A prospective study of plasma prolactin concentrations and risk of premenopausal and postmenopausal breast cancer. *J. Clin. Oncol.* **25**, 1482–1488 (2007).
10. Grose, R. & Dickson, C. Fibroblast growth factor signaling in tumorigenesis. *Cytokine Growth Factor Rev.* **16**, 179–186 (2005).
11. Minichiello, M.J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
12. Eliassen, A.H., Tworoger, S.S., Mantzoros, C.S., Pollak, M.N. & Hankinson, S.E. Circulating insulin and c-peptide levels and risk of breast cancer among predominately premenopausal women. *Cancer Epidemiol. Biomarkers Prev.* **16**, 161–164 (2007).
13. Hayes, R.B. *et al.* Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control. Clin. Trials* **21**, 349S–355S (2000).
14. Stevens, V.L., Rodriguez, C., Pavluck, A.L., Thun, M.J. & Calle, E.E. Association of polymorphisms in the paraoxonase 1 gene with breast cancer incidence in the CPS-II Nutrition Cohort. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1226–1228 (2006).
15. Bruzzi, P., Green, S.B., Byar, D.P., Brinton, L.A. & Schairer, C. Estimating the population attributable risk for multiple risk factors using case-control data. *Am. J. Epidemiol.* **122**, 904–914 (1985).
16. Moffa, A.B. & Ethier, S.P. Differential signal transduction of alternatively spliced FGFR2 variants expressed in human mammary epithelial cells. *J. Cell. Physiol.* **210**, 720–731 (2007).
17. Chanock, S.J. *et al.* What constitutes replication of a genotype-phenotype association? Summary of an NCI-NHGRI working group. *Nature* (in the press).
18. Wang, H., Thomas, D.C., Pe'er, I. & Stram, D.O. Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* **30**, 356–368 (2006).
19. 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).
20. Kraft, P., Cox, D.G., Paynter, R.A., Hunter, D. & De Vivo, I. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genet. Epidemiol.* **28**, 261–272 (2005).
21. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
22. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
23. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).