Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression

Hatef Darabi,1,119 Karen McCue,2,119 Jonathan Beesley,2 Kyriaki Michailidou,3 Silje Nord,4,5 Siddhartha Kar,6 Keith Humphreys,1 Deborah Thompson,1 Maya Ghousaini,6 Manjeet K. Bolla,3 Joe Dennis,3 Qin Wang,3 Sander Canisius,7 Christopher G. Scott,8 Carmel Apicella,9 John L. Hopper,9 Melissa C. Southey,10 Jennifer Stone,11 Annegien Broeks,7 Marjanka K. Schmidt,7 Rodney J. Scott,12,13 Artitaya Lophatananon,14 Kenneth Muir,14,15 Matthias W. Beckmann,16 Arif B. Ekici,17 Peter A. Fasching,16,18 Katharina Heusinger,16 Isabel dos-Santos-Silva,19 Julian Peto,19 Ian Tomlinson,20 Elinor J. Sawyer,21 Barbara Burwinkel,22,23 Frederik Marme,22,24 Pascal Guenel,25,26 Thérèse Truong,25,26 Stig E. Bojesen,27,28,29 Henrik Flyger,30 Javier Benitez,31,32,33 Anna González-Neira,31 (Author list continued on next page)

Genome-wide association studies have identified SNPs near ZNF365 at 10q21.2 that are associated with both breast cancer risk and mammographic density. To identify the most likely causal SNPs, we fine mapped the association signal by genotyping 428 SNPs across the region in 89,050 European and 12,893 Asian case and control subjects from the Breast Cancer Association Consortium. We identified four independent sets of correlated, highly trait-associated variants (iCHAVs), three of which were located within ZNF365. The most strongly risk-associated SNP, rs10995201 in iCHAV1, showed clear evidence of association with both estrogen receptor (ER)-positive (OR = 0.85 [0.82–0.88]) and ER-negative (OR = 0.87 [0.82–0.91]) disease, and was also the SNP most strongly associated with percent mammographic density. iCHAV2 (lead SNP, chr10: 64,258,684:D) and iCHAV3 (lead SNP, rs7922449) were also associated with ER-positive (OR = 0.93 [0.91–0.95] and OR = 1.06 [1.03–1.09]) and ER-negative (OR = 0.95 [0.91–0.98] and OR = 1.08 [1.04–1.13]) disease. There was weaker evidence for iCHAV4, located 5’ of ADO, associated only with ER-positive breast cancer (OR = 0.93 [0.90–0.96]). We found 12, 17, 18, and 2 candidate causal SNPs for breast cancer in iCHAVs 1–4, respectively. Chromosome conformation capture analysis showed that iCHAV2 interacts with the ZNF365 and NRBF2 (more than 600 kb away) promoters in normal and cancerous breast epithelial cells. Luciferase assays did not identify SNPs that affect transactivation of ZNF365, but identified a protective haplotype in iCHAV2, associated with silencing of the NRBF2 promoter, implicating this gene in the etiology of breast cancer.

Introduction

Breast cancer is one of the most commonly occurring epithelial malignancies in women, with an estimated 1.7 million new cases and more than 520,000 deaths annually worldwide.1,2 Familial aggregation and twin studies have shown a substantial contribution of inherited susceptibility to breast cancer.3 Genome-wide association studies (GWASs) provide a powerful approach to identify common disease alleles. In a two-stage genome-wide association study conducted in European descendants, Turnbull et al.4 identified five new susceptibility loci.
Hoda Anton-Culver,"\textsuperscript{34} Susan L. Neuhausen,"\textsuperscript{45} Volker Arndt,"\textsuperscript{56} Hermann Brenner,"\textsuperscript{56,37,38} Christoph Engel,"\textsuperscript{59} Alfons Meindl,"\textsuperscript{41,42} Rita K. Schmutzler,"\textsuperscript{41,42,43} German Consortium of Hereditary Breast and Ovarian Cancer, Norbert Arnold,"\textsuperscript{44} Hiltrud Brauch,"\textsuperscript{38,45,46} Ute Hamann,"\textsuperscript{47} Jenny Chang-Claude,"\textsuperscript{48} Sofia Khan,"\textsuperscript{49} Heli Nevanlinna,"\textsuperscript{50} Hitkem Ito,"\textsuperscript{50} Keitaro Matsuo,"\textsuperscript{51} Natalia V. Bogdanova,"\textsuperscript{52} Thilo Dörk,"\textsuperscript{53} Annika Lindblom,"\textsuperscript{54} Sara Margolin,"\textsuperscript{55} kConFab/AOCS Investigators, Veli-Matti Kosma,"\textsuperscript{56,57,58} Arto Mannermaa,"\textsuperscript{56,57,58} Chiu-chen Tseng,"\textsuperscript{59} Anna H. Wu,"\textsuperscript{59} Giuseppe Floris,"\textsuperscript{60} Diether Lambrechts,"\textsuperscript{61,62} Anja Rudolph,"\textsuperscript{48} Paolo Peterlongo,"\textsuperscript{63} Paolo Radice,"\textsuperscript{64} Fergus J. Couch,"\textsuperscript{65} Celine Vachon,"\textsuperscript{8} Graham G. Giles,"\textsuperscript{9,66} Catriona McLean,"\textsuperscript{67} Roger L. Milne,"\textsuperscript{9,66} Pierre-Antoine Dugue,"\textsuperscript{66} Christopher A. Haiman,"\textsuperscript{59} Gertraud Maskarinec,"\textsuperscript{68} Christy Woolcott,"\textsuperscript{69} Brian E. Henderson,"\textsuperscript{59} Mark S. Goldberg,"\textsuperscript{70,71} Jacques Simard,"\textsuperscript{72} Soo H. Teo,"\textsuperscript{73,74} Shivaani Mariapun,"\textsuperscript{73,74} Aslaug Helland,"\textsuperscript{75} Vilde Haakensen,"\textsuperscript{76} Wei Zheng,"\textsuperscript{77} Alicia Beeghly-Faiel,"\textsuperscript{78} Rulli Tamimi,"\textsuperscript{78,79,80} Arja Jukkola-Vuorinen,"\textsuperscript{81} Robert Winqvist,"\textsuperscript{82,83} Irene L. Andrulis,"\textsuperscript{84,85} Julia A. Knight,"\textsuperscript{86,87} Peter Devilee,"\textsuperscript{88,89} Robert A.E.M. Tollenaar,"\textsuperscript{90} Jonine Figueroa,"\textsuperscript{91} Montserrat Garcia-Closas,"\textsuperscript{92,93} Kamila Czene,"\textsuperscript{1} Maartje J. Hooning,"\textsuperscript{94} Madeleine Tilianus-Linthurst,"\textsuperscript{95}

(Affiliations continued on next page)
One of the identified SNPs, rs10995190 (combined p = \(5.1 \times 10^{-15}\)), lies within intron 4 of ZNF365 (MIM: 607818) at 10q21.2. Analysis by the Breast Cancer Association Consortium (BCAC) confirmed this association (p = 1.3 \times 10^{-16}) in those of European ancestry, with a non-significant association among Asians where this variant is rare.\(^6\) Instead, Cai et al.\(^7\) identified a common SNP, rs10822013 located 26.7 kb upstream, associated with breast cancer risk among East Asians. Further analysis by the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) showed that rs10995190 is associated with breast cancer risk among BRCA2 carriers,\(^8\) particularly for estrogen-receptor-positive breast cancer. This 10q21.2 locus was the first to be associated with percent mammographic density (PD) (combined p = 9.6 \times 10^{-10})\(^9\) and also shows a possible association with breast size.\(^10\) Varghese et al.\(^11\) showed that rs10995190 and rs10509168 (\(r^2 = 0.13\)) are both associated with mammographic density, and through analysis of polygenic risk scores found that PD and breast cancer have a shared genetic basis.

These data indicate that 10q21.2 is an important susceptibility region for both breast cancer and mammographic density and must harbor one or more SNPs causally related to these phenotypes. In an attempt to identify the most likely causal SNPs underlying these associations, we assessed 428 SNPs across the 10q21.2 region, applying multiple analyses aimed at exploring the target genes and functional basis of the associations with breast cancer risk and mammographic density.
Mammographic Density

Mammographic density measurements were available for 6,886 women from nine BCAC studies that were also part of the Marker Of DEnsity (MODE) consortium\textsuperscript{6} and DENSNP consortium\textsuperscript{14} (Table S1). Covariate data were obtained through self-administered postal questionnaires, in-person interviews, or telephone interviews, and anthropometric variables were self-reported or measured by trained staff (Table S2 of Vachon et al.).\textsuperscript{14} Seven studies estimated density (absolute and percent dense area) with the CUMULUS program,\textsuperscript{15} and two of the studies, the BBCC and NBCS studies, used the Madena software.\textsuperscript{16}

Statistical Analysis

For each SNP, we estimated a per-allele log-odds ratio (OR) and standard error using logistic regression, including principal components and per-study fixed-effects to adjust for study-specific differences in allele frequency, as described previously.\textsuperscript{7} Analyses were carried out separately for Europeans and Asians. We estimated genetic main effects for ER-positive and ER-negative breast cancer using logistic regression and restricting the cases to a specific subtype. We evaluated heterogeneity of association by tumor subtype in a case-only analysis, treating subtype status as the dependent variable. We derived the p values for association regression analyses, utilizing the step function in R with a penalty term set to 10.\textsuperscript{17} To account for uncertainty in the data resulting from the imputation process, we conducted analysis by regressing on the allele dosage for each genotype. We estimated haplotype-specific ORs using an EM algorithm implemented in the haplo.stats package in R. For this analysis we used the most probable genotypes and included study and principal components as covariates. We grouped haplotypes with a frequency < 0.01 together into one subgroup of rare haplotypes.

We assessed individual SNP associations with percentage density (PD), dense area (DA), and non-dense area (nDA) via linear regression. Because the distributions of estimated PD, DA, and nDA were positively skewed (skewness = 1.31, 2.32, and 1.02, respectively), each phenotype was square-root transformed. This transformation reduces skewness and has previously been shown to generate variables that are approximately normally distributed.\textsuperscript{18} In addition to study and principal components, we included age, body mass index (BMI), postmenopausal hormone replacement therapy (HRT), mammographic view, menopausal status, and case-control status as covariates in each model. Age and BMI were treated as continuous variables, and use of postmenopausal hormones at time of mammogram (0, never; 1, stopped prior to mammogram date; 2, current use at date of mammogram; 9, unknown), menopausal status at time of mammogram (0, postmenopausal; 1, premenopausal; 2, perimenopausal; 9, unknown), and case-control status were treated as categorical variables in the regression model. We also included mammographic view (1, medio-lateral oblique [MLO] view; 2, craniocaudal [CC] view) and treated it as categorical variable because it has been shown that the percent density measurements from the MLO view are systematically lower than those from the CC view.\textsuperscript{9} The mean percentage density across all studies was 17 from the MLO view as compared to 25 from the CC view.

Expression Quantitative Trait Locus Analysis

We examined the associations of germline-genotyped and imputed SNPs within 1 Mb of the risk region with the expression levels of all genes within 1 Mb, up- and downstream, of the SNP in question (including ADH [MIM: 611392], ARID5B [MIM: 608538], c10orf107, EGR2 [MIM: 129010], JMJD1C [MIM: 604503], JMJD1C-AS1, REEP3 [MIM: 603948], RTNK2 [MIM: 113708], and ZNF365) in normal breast, adjacent normal, and breast cancerous tissue from the following four cohorts. Normal breast 1 (NBI; n = 116) is comprised of women of European descent ascertained through multiple Norwegian hospitals. Gene expression data for the majority of women were derived from normal breast tissue in women who had not been affected with breast cancer; data for ten women were derived from normal tissue adjacent to a tumor. Genotyping was performed with the iCOGS SNP array, and gene expression levels were measured with the Agilent 44K array.\textsuperscript{18,19} NBI (n = 93) is the European subset of the TCGA study, for whom expression data were available from normal tissue adjacent to a tumor. Germline genotype data from Affymetrix SNP 6 array, processed through Birdseed, were obtained from TCGA dbGAP data portal.\textsuperscript{20} Gene expression levels were assayed by RNA sequencing, RSEM (RNaseq by Expectation-Maximization)\textsuperscript{21} normalized per gene, as obtained from the TCGA consortium.\textsuperscript{20} The data were log\textsubscript{2} transformed, and unexpressed genes were excluded prior to eQTL analysis. Breast carcinomas 1 (BC1; n = 241) is a series of women of European ancestry diagnosed with breast cancer and recruited through multiple Norwegian hospitals. Genotypes were obtained with the iCOGS SNP array, and mRNA expression data were from the Agilent 44K array.\textsuperscript{22} BCII (n = 765) is the TCGA breast cancer cohort; all non-European samples (as determined by clustering and PCA) were excluded from this analysis.\textsuperscript{20} The genotyping platform was Affymetrix SNP 6, and gene expression data, for the breast tumors, was derived from RNA sequencing analysis, in a similar manner to NBII. There is no overlap between women recruited to each of these studies.

In addition, we examined all the genotyped or imputed SNPs used in the risk analysis for association with expression of nine genes (ADO, ARID5B, c10orf107, EGR2, JMJD1C, NRBF2, REEP3, RTNK2, and ZNF365, represented by 14 expression probes) in the 1 Mb region on either side of the fine-mapping interval, using data from normal tissue in patients from METABRIC. METABRIC (n = 135) comprises normal tissues adjacent to the tumors from breast cancer patients of genetically confirmed European ancestry from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study.\textsuperscript{23} The samples were assayed with the Illumina HT-12 v3 microarray. Matched germline SNP genotypes were available from the Affymetrix SNP 6.0 platform.

For all cohorts, the genotyping data were processed as follows: SNPs with call rates < 0.95 or minor allele frequencies < 0.05 were excluded, as were SNPs out of Hardy-Weinberg equilibrium with p < 10\textsuperscript{-6}. All samples with a call rate below 80% were excluded. Identity by state was computed with the R GenABEL package,\textsuperscript{24} and closely related samples with IBS > 0.95 were removed. The SNP and sample filtration criteria were applied iteratively until all samples and SNPs met the stated thresholds. In total, 489 samples and 662,521 SNPs passed were kept in the analysis. Imputation was run on both the iCOGS and Affymetrix6 germline genotype data using the 1000 Genomes Project March 2012 v.3 release as the reference dataset.\textsuperscript{25} A two-stage imputation procedure, using SHAPEIT to derive phased genotypes and IMPUTEv2 to perform the imputation on the phased data, has been found to significantly reduce the computational burden.\textsuperscript{26}
The influence of germline genetic variations on gene expression was assessed using a linear regression model, as implemented in the R library eMAP. An additive effect was assumed by modeling copy number of the rare allele, i.e., 0, 1, or 2, for a given genotype. Only relationships in cis, i.e., where the SNP resided less than 1 Mb up or down from the center of the transcript, were investigated. Correction for multiple testing was performed using the false discovery rate (FDR) as implemented in the p.adjust function in R. Only FDR-adjusted, significant p values are reported in the Results section of the paper. Genotyping quality control and imputation for the METABRIC data are described in Guo et al.27 For the METABRIC data, association between genotype and expression was tested by linear regression with FDR control as implemented in the MatrixEQTL package in R. We also conducted eQTL haplotype analysis in NBI and BCII which had been genotyped on iCOGS. The correlations with expression for the haplotypes residing in the four iCHAVs were estimated using the haplo.score function implemented in the R library haplo.stats, assuming a Gaussian distribution for the expression data.

Cell Lines
Normal breast epithelial cell lines MCF10A and Bre80 were grown in DMEM/F12 media supplemented with 5% horse serum, 10 μg/ml insulin, 0.5 μg/ml hydrocortisone, 20 ng/ml EGF, 100 ng/ml cholera toxin, and antibiotics. The MCF7 breast cancer cell line was grown in DMEM media supplemented with 10% fetal bovine serum, 10 μg/ml insulin, sodium pyruvate, and antibiotics. All cell lines were routinely tested for Mycoplasma and profiled by gel electrophoresis, gel purified, and sequenced to verify the NRBF2 10q21 locus and BAC clones (RP11-629KB and RP11-1021P21) that spanned the 10q21.2 locus. Interactions were quantified in triplicate. The results for all SNPs in the interval with an association p value < 10^{-7} to 10^{-29}; Figure 1. The results for all SNPs in the interval with an association p value < 10^{-3} for overall breast cancer risk (382 SNPs) are presented in Table S3.

Results
We successfully genotyped 428 SNPs across the 560-kb fine-mapping region in 46,450 case subjects and 42,600 control subjects from 41 case-control studies in populations of European ancestry, and in 6,269 case subjects and 6,624 control subjects from 9 case-control studies of Asian ancestry.5 We imputed genotypes for 3,409 SNPs in the interval in the European studies, at imputation r^2 > 0.3 and MAF ≥ 1%, using known genotypes in combination with data from the 1000 Genomes Project reference panel. Based on data from the European studies, 87 genotyped or imputed SNPs were convincingly associated with overall risk of breast cancer (p values 10^{-7} to 10^{-29}, Figure 1). The results for all SNPs in the interval with an association p value < 10^{-3} for overall breast cancer risk (382 SNPs) are presented in Table S3.

Among the genotyped SNPs, the strongest evidence of association with overall breast cancer risk among Europeans was for rs10995194 (OR [95% CI] = 0.86 [0.84–0.88], p = 3.77 × 10^{-29}). This SNP lies within intron 4 of ZNF365 and is strongly correlated with the original GWAS hit rs10995190 (r^2 = 0.99). Analysis of the imputed SNPs identified two markers with slightly stronger associations: the most significantly associated marker was rs10995201 (OR = 0.85 [0.83–0.88], p = 1.05 × 10^{-29}), which is also strongly correlated with rs10995194 (r^2 = 0.94).

Multiple Independent Signals at 10q21.2
To determine whether there were additional independent signals of association at the locus, we included in a forward stepwise regression model 230 SNPs with MAF ≥ 2% that displayed evidence of association with overall breast cancer risk at p < 10^{-4}. The most parsimonious model included four independent SNPs that mark four iCHAVs (Figure 1; Table 1). These were as follows: iCHAV1, rs10995201 (OR = 0.85 [0.83–0.88], p = 1.05 × 10^{-29};
These four SNPs were all imputed, with \( \text{conditional} \ p = 4.92 \times 10^{-16} \); iCHAV2, chr10: 64,258,684:D (OR = 0.93 [0.91–0.95], \( p = 4.24 \times 10^{-14} \); conditional \( p = 3.24 \times 10^{-3} \)); iCHAV3, rs7922449 (OR = 1.06 [1.04–1.08], \( p = 1.68 \times 10^{-7} \); conditional \( p = 5.78 \times 10^{-5} \)); iCHAV4, rs9971363 (OR = 0.94 [0.92–0.97], \( p = 6.54 \times 10^{-5} \); conditional \( p = 3.95 \times 10^{-4} \)). These four SNPs were all imputed, with IMPUTE2 info-score above 0.85. The strongest pairwise LD \( r^2 \) value among the four markers was 0.12 (between rs10995201 and chr10: 64,258,684:D). Among the genotyped SNPs, the top SNP correlated with the lead iCHAV1 SNP was rs10995194 (OR = 0.86 [0.84–0.88], \( p = 3.77 \times 10^{-29} \), \( r^2 = 0.94 \)), with the lead iCHAV2 SNP was rs2393886 (OR = 0.93 [0.91–0.95], \( p = 4.50 \times 10^{-14} \), \( r^2 = 0.99 \)), with the lead iCHAV3 SNP was rs4746428 (OR = 1.06 [1.04–1.08], \( p = 3.24 \times 10^{-8} \), \( r^2 = 0.68 \)), and with the lead iCHAV4 SNP was rs10995312 (OR = 0.95 [0.92–0.97], \( p = 1.38 \times 10^{-4} \), \( r^2 = 0.97 \)).

In order to identify the candidate causal SNPs in each iCHAV for subsequent functional analysis, we calculated the likelihood ratio of each SNP relative to the best independent signal with which it was correlated (\( r^2 > 0.6 \)), after adjusting for the lead SNP of preceding iCHAV(s). SNPs with a relative likelihood ratio of <1:100 compared with the most significant SNP for each iCHAV were excluded from consideration as being potentially causative.16,30 Eleven SNPs had a relative likelihood ratio of >1:100 compared with the most significant SNP (rs10995201) and hence could not be excluded as causative for the lead signal—these SNPs were all strongly correlated with rs10995201 and span an interval of 31.2 kb (iCHAV1, Figure 1, Table S4). A 12th SNP, a single base insertion chr10: 64,291,099, was excluded at this threshold, but not strongly so (likelihood ratio ~1:600); all other SNPs could be clearly excluded (likelihood ratios < 1:10\(^{12} \)). After adjustment for lead SNP of iCHAV1, we identified 16 strongly correlated SNPs and with a likelihood ratio >1:100 relative to the lead SNP in iCHAV2 (chr10: 64,258,684:D); these SNPs span an interval of 25 kb (iCHAV2, Figure 1, Table S5). SNP rs16917302 (OR = 0.96 [0.93–0.99], \( p = 1.03 \times 10^{-5} \), reported by Couch et al.,31 lies within the region that iCHAV2 spans. However, rs16917302 is only weakly correlated with the lead SNP (\( r^2 = 0.12 \)) in iCHAV2. After adjusting for the effects of lead SNPs of iCHAV1 and iCHAV2, 17 highly correlated SNPs had a relative likelihood >1:100 compared to rs7922449 (iCHAV3, Figure 1, Table S6) and span an interval of 29.1 kb. Finally, only one other SNP, rs7090365, had a relative likelihood of >1:100 with respect to rs9971363 the lead SNP of iCHAV4, after adjusting for the lead SNPs in iCHAVs 1–3 (Figure 1, Table S7). iCHAVs 1, 2, and 3 all lie within ZNF365. The SNPs in iCHAV1 and iCHAV3 span regions that physically overlap, whereas the SNPs in iCHAV2 lie telomeric to iCHAV1 and iCHAV3. iCHAV4 lies \( S' \) of \( ADO \), which encodes 2-aminoethanethiol dioxygenase (Figure 1).

**Association with Breast Cancer Subtypes**

Based on data from European studies, 31 genotyped SNPs and 14 imputed SNPs were associated with risk of ER-positive breast cancer (\( p \) values \( 10^{-7} \) to \( 10^{-23} \)). The most strongly associated SNP for overall breast cancer (rs10995201) was also the most strongly associated SNP
Table 1. Four iCHAVs at 10q21.2 Associated with Breast Cancer Risk or Mammographic Density Phenotype in Europeans

<table>
<thead>
<tr>
<th>iCHAV1</th>
<th>rs10995201&lt;sup&gt;a&lt;/sup&gt;</th>
<th>p Value&lt;sup&gt;b&lt;/sup&gt;</th>
<th>OR (95% CI)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>SNP&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p Value&lt;sup&gt;b&lt;/sup&gt;</th>
<th>OR (95% CI)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>SNP&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p Value&lt;sup&gt;b&lt;/sup&gt;</th>
<th>OR (95% CI)&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>1.05 × 10&lt;sup&gt;-29&lt;/sup&gt;</td>
<td>0.85 (0.83–0.88)</td>
<td>4.24 × 10&lt;sup&gt;-14&lt;/sup&gt;</td>
<td>0.93 (0.91–0.95)</td>
<td>1.68 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>1.06 (1.04–1.08)</td>
<td>6.54 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>0.94 (0.92–0.97)</td>
<td></td>
</tr>
<tr>
<td>ER&lt;sup&gt;+&lt;/sup&gt;</td>
<td>2.51 × 10&lt;sup&gt;-23&lt;/sup&gt;</td>
<td>0.85 (0.82–0.88)</td>
<td>8.01 × 10&lt;sup&gt;-11&lt;/sup&gt;</td>
<td>0.93 (0.91–0.95)</td>
<td>6.70 × 10&lt;sup&gt;-6&lt;/sup&gt;</td>
<td>1.06 (1.03–1.09)</td>
<td>1.26 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>0.93 (0.90–0.96)</td>
<td></td>
</tr>
<tr>
<td>ER&lt;sup&gt;-&lt;/sup&gt;</td>
<td>9.60 × 10&lt;sup&gt;-8&lt;/sup&gt;</td>
<td>0.87 (0.82–0.91)</td>
<td>0.0055</td>
<td>0.95 (0.91–0.98)</td>
<td>0.000268</td>
<td>1.08 (1.04–1.13)</td>
<td>0.776</td>
<td>0.99 (0.94–1.05)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p Value&lt;sup&gt;c&lt;/sup&gt;</th>
<th>β (SE)&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p Value&lt;sup&gt;c&lt;/sup&gt;</th>
<th>β (SE)&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p Value&lt;sup&gt;c&lt;/sup&gt;</th>
<th>β (SE)&lt;sup&gt;c&lt;/sup&gt;</th>
<th>p Value&lt;sup&gt;c&lt;/sup&gt;</th>
<th>β (SE)&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density area</td>
<td>1.45 × 10&lt;sup&gt;-7&lt;/sup&gt;</td>
<td>−0.25 (0.05)</td>
<td>0.0055</td>
<td>−0.09 (0.03)</td>
<td>0.308</td>
<td>0.04 (0.04)</td>
<td>0.634</td>
</tr>
<tr>
<td>Percent density</td>
<td>1.32 × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
<td>−0.15 (0.04)</td>
<td>0.00155</td>
<td>−0.08 (0.02)</td>
<td>0.634</td>
<td>0.01 (0.03)</td>
<td>0.629</td>
</tr>
</tbody>
</table>

**Notes:**
- <sup>a</sup>Lead SNP for each iCHAV.
- <sup>b</sup>p value, odds ratio (OR), and 95% confidence interval (CI) for association with overall breast cancer risk, estrogen-positive (ER<sup>+</sup>), and estrogen-negative (ER<sup>-</sup>) disease.
- <sup>c</sup>p value, beta, and standard error (SE) for association with percentage density, dense area.
- <sup>d</sup>SNPs correlated ($r^2 > 0.6$) with a likelihood ratio of $>1:100$ with respect to overall risk association relative to lead SNP of each iCHAV and corresponding p value for association with overall breast cancer risk.
- <sup>e</sup>IMPUTE2 info score.
- <sup>f</sup>Genotyped marker.
- <sup>g</sup>Imputed marker.

---

28 The American Journal of Human Genetics 97, 22-34, July 2, 2015
for ER-positive disease (OR = 0.85 [0.82–0.88], p = 2.51 × 10^{-23}) and had a similar association for ER-negative disease (OR = 0.87 [0.82–0.91], p = 9.60 × 10^{-08}, \text{P}_{\text{heterogeneity}} = 0.34) (Table 1, Figures S1 and S2). The most strongly associated, well-imputed SNP (IMPUTE2 info-score > 0.9) for ER-negative disease was rs10995182 (OR = 0.85 [0.8–0.9], p = 6.68 × 10^{-8}), another SNP within iCHAV1 strongly correlated with rs10995201. Thus, the results indicate that susceptibility SNPs in iCHAV1 confer similar relative risks for ER-positive and ER-negative disease. Similarly, the lead SNPs in iCHAVs 2 and 3 showed similar associations for ER-positive and ER-negative disease (ORER+ = 0.96 [0.94–0.98], \text{P}_{\text{heterogeneity}} = 0.86, ORER- = 0.98 [0.94–1.02], \text{P}_{\text{heterogeneity}} = 0.44 for chr10: 64,258,684:D adjusted for iCHAV1, ORER+ = 1.04 [1.02–1.07], \text{P}_{\text{heterogeneity}} = 0.15 [1.02–1.11], \text{P}_{\text{heterogeneity}} = 0.44 for rs7922449 adjusted for iCHAVs 1 and 2). However, the lead SNP in iCHAV4 showed no association with ER-negative disease (ORER- = 0.93 [0.90–0.97], \text{P}_{\text{heterogeneity}} = 7.72 × 10^{-05}, \text{OR}_{\text{ER}} = 1.00 [0.94–1.05], \text{P}_{\text{heterogeneity}} = 0.862, \text{P}_{\text{heterogeneity}} = 0.016 for rs9971363 adjusted for iCHAVs 1, 2, and 3).

To determine whether there were additional subtype-specific signals of association, we included all SNPs displaying evidence for association with ER-positive disease (222 SNPs, p < 10^{-4} and MAF ≥ 2%) and ER-negative disease (19 SNPs, p < 10^{-4} and MAF ≥ 2%) in separate forward stepwise regression models. For ER-positive disease, two iCHAVs were identified: iCHAV1-ER+ rs10995201 (OR = 0.85 [0.83–0.88], p = 2.51 × 10^{-23}, conditional p = 1.65 × 10^{-18} in iCHAV1 and iCHAV2-ER+ chr10: 64,258,684:D (OR = 0.93 [0.91–0.95], p = 8.01 × 10^{-11}, conditional p = 2.59 × 10^{-05}) in iCHAV2.

Breast Cancer Risk Associations in Asian Studies

The top associated marker among Asians was the genotyped SNP rs7914770 (OR = 0.93 [0.89–0.98], p = 0.006). This SNP, which lies between iCHAV1 and iCHAV4, as defined in analysis of the European population, was not associated with overall breast cancer risk among Europeans (OR = 1.01 [0.98–1.03], p = 0.534; Table S8). The most strongly associated genotyped SNP in Europeans, rs10995194 in iCHAV1, showed a borderline association with breast cancer risk in Asians, but in the opposite direction (Asians: OR = 1.18 [1.00–1.39], p = 0.04; Europeans: OR = 0.86 [0.84–0.88], p = 3.77 × 10^{-29}). The minor (C) allele of rs10995194 (MAF = 0.15) and the other iCHAV1 SNPs were much rarer in Asians (MAF = 0.02), but the ORs estimates nevertheless differed significantly (p = 10^{-4}). In iCHAV2, the most strongly associated genotyped SNP in Europeans, rs2393886 (OR = 0.93, MAF = 0.47 for the A allele), was also associated with risk in Asians; however, in contrast to iCHAV1, the effects were in the same direction and of comparable magnitude (OR = 0.95 [0.90–1.00], p = 0.04, allele frequency 0.51 for the A allele in Asians). In this iCHAV the most strongly associated SNP in Asians was rs4746409 (OR = 1.06 [1.00–1.11], p = 0.03), but this is strongly correlated with the lead SNP rs2393886 (r^2 = 0.88). SNP rs10822013 within iCHAV2, recently identified by Cai et al. in an Asian GWAS, was more weakly associated with risk in Asians in our data (OR = 0.96 [0.90–1.10], p = 0.07). None of the genotyped markers in iCHAV3 or iCHAV4 that were significant in Europeans showed association with overall breast cancer risk in the Asian population.

Haplotype Analysis

We conducted haplotype analysis using the most probable genotype for the imputed lead SNPs in each of the iCHAVs. We grouped haplotypes with a frequency < 0.01 into one (rare) group. The haplo.score procedure estimated 11 haplotypes with non-zero frequencies for all three phenotypes (overall, ER\(^+\), and ER\(^-\) risk) (Table S9). The most significant association was observed for haplotype H6, carrying the rare allele of the lead SNPs of iCHAV1 and iCHAV2: OR = 0.86 (0.82–0.90), p = 8.52 × 10^{-10}. Consistent with the regression analyses, the three haplotypes (H6, H14, H5) carrying the rare allele of the iCHAV1 SNP rs10995201 were all associated with a similar, reduced breast cancer risk. Three other haplotypes, all carrying the rare allele of the iCHAV2 SNP, chr10: 64,258,684:D, were also associated with a reduced risk, relative to the baseline haplotype, consistent with an independent effect of iCHAV2.

Association with Mammographic Density Phenotypes

Multiple linear regression was used to investigate the association between mammographic density phenotypes (PD, percent density; DA, dense area; nDA, non-dense area) and genotypes after adjustment for other covariates. Among the three density phenotypes, the strongest associations were observed with DA. The strongest association with DA and PD was seen with the lead SNP of iCHAV1, rs10995201 (β (SE) = -0.25 [0.05], p = 1.45 × 10^{-07} for DA, β (SE) = -0.15 [0.04], p = 1.32 × 10^{-05} for PD; Tables S10, S11; Figures S3 and S4). After adjusting for the lead SNP in iCHAV1, no SNPs in iCHAV2 were associated with any of the density phenotypes at p < 0.01 (Table S12); similarly, none of the iCHAV3 or iCHAV4 SNPs were associated with density phenotypes (Tables S13 and S14). The strongest association with nDA was seen with imputed marker rs224303 about 33 kb away from iCHAV4 (Figure S5), which was not associated with overall risk of breast cancer (p = 0.745).

To assess the extent to which the observed association with breast cancer risk might be mediated through a mammographic density phenotype, we estimated the association with breast cancer risk before and after adjustment for PD, DA, or nDA in a pooled analysis of the 2,379 breast cancer case subjects and 4,507 control subjects for whom density measurements were available. The association between lead iCHAV1 SNP rs10995201 (i.e., the only iCHAV for which an association with mammographic density had been identified) and breast cancer in this subset (OR = 0.87...
[0.77–0.98], p = 0.02) was similar to that in the complete BCAC set and was only slightly attenuated after adjustment for PD (OR = 0.88 [0.78–1.00], p = 0.05) or DA (OR = 0.89 [0.79–1.01], p = 0.06) (Table S15).

eQTL Analysis
We analyzed 2,238, 5,122, 2,250, 5,211, and 3,814 SNPs in NBI, NBII, BCI, BCII, and METABRIC, respectively, for association with the expression levels of all genes within 1 Mb, up- and downstream, of the SNP in question. Significant eQTL associations were observed for both normal breast and tumors. Multiple SNPs within the fine mapping region associated with expression of c10orf107 in the NBII cohort (strongest association, chr10: 63,427,159:D, which is not a candidate causal risk SNP; FDR-adjusted p = 6.1 × 10−5 and r² = 0.325). For the BCI cohort, multiple eQTLs were found for both c10orf107 and RTKN2 (strongest associations chr10: 63,427,159:D, FDR-adjusted p = 4.3 × 10−6 and r² = 0.15, and rs870988, FDR-adjusted p = 1.4 × 10−5 and r² = 0.13, respectively). In BCII we observed multiple eQTLs for three different genes, c10orf107, EGR2, and ADO (rs12781009, FDR-adjusted p = 5.6 × 10−9 and r² = 0.066; rs34632941, FDR-adjusted p = 0.02 and r² = 0.02; and rs224045, FDR-adjusted p = 0.02 and r² = 0.02, respectively). No significant associations with expression were found for NBI or METABRIC, and there were no significant associations between the putative causal SNPs in iCHAVs 1–4 and expression of any of the genes analyzed in the region in any study (Tables S16 and S17). Significant associations (without correction for multiple testing) were found between haplotypes of iCHAV1 and NRBF2 in NBI and BCI (p = 0.005 and p = 0.042, respectively) and between haplotypes of iCHAV2 and NRBF2 in NBI (p = 0.011). In addition, we found associations between haplotypes of iCHAV1 and REEP3, EGR2, and RTKN2 (p = 0.011–0.049), haplotypes of iCHAV2 and RTKN2 (p = 0.012), haplotypes of iCHAV3 and REEP3, JMJD1C, ARID5B, ADO, RTKN2, and TMEM26 (p = 0.001–0.036), and haplotypes of iCHAV4 and ARID5B (p = 0.012) in either NBI or BCI, but not both.

Chromosome Conformation Capture Analyses Identify ZNF365 and NRBF2 as Target Genes
We performed 3C experiments to determine whether there were any chromatin interactions between the ZNF365 and NRBF2 promoters and iCHAVs 1, 2, and 3. We identified significant interactions between iCHAV2 at the 10q21.2 locus and both promoter regions in the breast cancer cell line, MCF7, and in two normal breast cell lines, MCF10A and Bre80 (Figures 2 and S6). No reproducible interactions were detected between iCHAVs 1 or 3 and ZNF365 and NRBF2 in the cell lines analyzed. Although we did observe some possible interactions in MCF7 and MCF10A between different parts of iCHAV 1/3 and the promoters of both ZNF365 and NRBF2, they were not reproducible (Figures 2 and S6).
Identification of a Putative Regulatory Region within iCHAV2 Regulating both ZNF365 and NRBF2

We used available ENCODE ChIP-seq data to identify putative regulatory elements (PREs) within iCHAV2 (Figure 3A). We identified two PREs (PRE1 and PRE2), as defined by DNaseI hypersensitivity sites (indicative of regions of open chromatin), in several normal and cancer breast cell lines, and H3K4me1 and H3K4me2 histone modifications in HMECs (Figure 3B). PRE1 lies in a complex repetitive region and we were unable to clone and analyze this region. However, 6 (chr10: 64,258,684:D, rs2393886, rs10995176, c10_pos64258017, rs10509168, and chr10: 64,258,692:D) of the 17 candidate causal SNPs lie within PRE2 (Figure 3B). We examined the regulatory capability of PRE2, combined with the effect of the protective haplotype, using luciferase constructs containing the ZNF365 or NRBF2 promoters. Inclusion of the reference haplotype of iCHAV2 did not significantly alter the effect of the NRBF2 promoter in Bre80 cells. However, the construct containing the protective haplotype acted as a silencer relative to the iCHAV2 reference allele (p = 0.003; Figures 4A and S7). A similar trend was seen in MCF7 cells (p = 0.029; Figure 4C). Although constructs containing the ZNF365 promoter showed a trend reducing its activity with the reference haplotype of iCHAV2, this was significant only in MCF7 cells (Figures 4B and 4D). Similarly, inclusion of the protective haplotype reduced the ZNF365 promoter activity but had no effect relative to the iCHAV2 reference haplotype in either Bre80 or MCF7 cells (Figures 4B and 4D).

Discussion

Our large combined dataset provides clear confirmation of a susceptibility locus for breast cancer at 10q21.2 for breast cancer as originally reported by Turnbull et al. and for mammographic density as reported by Lindström et al. Multiple regression and haplotype analyses showed clear evidence of at least two, and potentially four, independent susceptibility loci in this region in Europeans. The most strongly associated SNP, rs10995201 in iCHAV1, showed clear evidence of association with both ER-positive and ER-negative disease, with the ORs being similar, and was also the SNP most strongly associated with mammographic density (DA, and hence PD). iCHAV2 (lead SNP, chr10: 64,258,684) and iCHAV3 (lead SNP, rs7922449) also appeared to be associated with both ER-positive and ER-negative disease. Evidence of a fourth iCHAV, associated only with ER-positive breast cancer, was weaker. In contrast to the results in Europeans, there was less evidence of association between SNPs at this locus and breast cancer risk in Asian women. The top associated genotyped SNP among Asians was rs7914770, which was not associated with overall breast cancer risk among Europeans. Furthermore, the lead genotyped SNP in Europeans, rs10995194, showed borderline evidence of association with breast cancer in Asians in the opposite direction, such that the effect sizes were clearly different. There are several possible explanations for the difference in the effect of iCHAV1 by ethnicity. The difference might reflect the differential effect of another SNP, or SNPs, in the
10q21.2 region. Alternatively, it might reflect trans-interactions with SNPs elsewhere in the genome. Larger studies in Asian populations, both for breast cancer risk and mammographic density, might help to resolve this paradox. The SNPs in iCHAV2, however, showed effects that were consistent between the two populations.

The A allele of rs10995190 in iCHAV1 is associated with decreased ER-positive and ER-negative breast cancer risk, as well as with lower percent mammographic density and reduced dense area. These results are consistent with the hypothesis that the same causal SNP confers susceptibility to both mammographic density and breast cancer risk. However, adjustment for percent density or dense area caused only a minor attenuation of the association between breast cancer risk and the lead iCHAV1 SNP. The implications of this finding are not clear, but it is likely that percent and absolute mammographic density as captured by a two-dimensional mammogram, despite being strong predictors of breast cancer risk, are imperfect measures of the underlying mechanism that drives breast cancer risk. Alternatively, it might be that different SNPs in this iCHAV are acting independently on mammographic density and breast cancer risk, and through different mechanisms. We did not observe associations with mammographic density for breast cancer-associated SNPs in the other iCHAVs, but the smaller sample size, and hence lower statistical power, for the mammographic density analysis mean that more subtle effects in other iCHAVs would not have been detectable.

As we have reported for the breast cancer loci at 11q13 and 2q35, we did not find any evidence for single SNP eQTLs for any of the putatively causal SNPs in iCHAVs 1–4 in normal or cancerous breast samples. This might be because the power of the eQTL studies (n = 93–765) is limited, or because eQTLs are context dependent and might be expressed only in certain cell types or in response to certain stimuli. We have, however, identified a putative regulatory element in iCHAV2 at the 10q21.2 locus that interacts with both the NRBF2 and ZNF365 promoters in both normal breast and breast epithelial tumor cells. It is interesting that we detected haplotype associations with iCHAV1 and iCHAV2 and NRBF2 expression, but the significance of these is difficult to interpret given their modest p values and the fact that these iCHAVs are far too big to be cloned in their entirety for luciferase assays. We prioritized the 3C analysis to look for interactions between iCHAVs 1, 2, and 3 and the promoters of ZNF365 and NRBF2 because of their function. NRBF2 encodes nuclear receptor binding protein 2, an interaction partner of the peroxisome proliferator-activated receptor alpha (PPARα), and exhibits a gene activation function in mammalian cells. Furthermore, NRFB2 is thought to have a role in cell survival in neural progenitor cells and to suppress autophagy, a process that needs to be tightly controlled in breast cells during normal development, tissue differentiation, and response to stress. To our knowledge, NRFB2 has not previously been implicated in breast cancer tumorigenesis. Our results suggest that the haplotype in iCHAV2 associated with reduced risk of breast cancer is associated with silencing of the NRBF2 promoter. ZNF365 encodes the zinc finger protein 365, which plays a critical role in stabilizing fragile sites within the genome and telomeres and maintaining genome stability, and is therefore a good candidate for a breast cancer
susceptibility gene. Although we did not observe a differential effect of the protective haplotype on the ZNF365 promoter in luciferase assays, the protective SNPs might act through differential looping as we have previously observed at the 2q35 breast cancer risk locus. Alternatively, other candidate causal SNPs in PRE1 or elsewhere in iCHAV2 might have a differential effect on transactivation of the ZNF365 promoter. We did not find any convincing evidence for interactions between iCHAV1/3 and the promoters of ZNF365 or NRB2 in MCF7, MCF10A, or Bre80 cells. It is possible that other genes, such as ADO, which encodes 2-aminoethanethiol dioxygenase, or EGR2 encoding early growth response-2, are the targets of these iCHAVs, or that their interactions with ZNF365 or NRB2 are manifest in different cell types, such as of the immune system, or only in response to specific stimuli.

In conclusion, we have found evidence for four sets of correlated genetic variants (iCHAVs) at 10q21.2 independently associated with breast cancer risk, one of which is also associated with mammographic density. In one of these iCHAVs, we have identified candidate causal SNPs that affect expression of NRB2, which lies more than 600 kb away, suggesting that expression of NRB2 might play a role in transformation or progression of transformed breast cells.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments, 7 figures, and 17 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.05.002.

Received: February 8, 2015
Accepted: May 1, 2015
Published: June 11, 2015

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org
eMAP, http://www.bios.unc.edu/~weisun/software.htm
RegulomeDB, http://RegulomeDB.org/
UCSC Genome Browser, http://genome.ucsc.edu

References


