Functional annotation of the 2q35 breast cancer risk locus implicates a structural variant in influencing activity of a long-range enhancer element

Joseph S. Baxter,1,* Nichola Johnson,1 Katarzyna Tomczyk,1 Andrea Gillespie,1 Sarah Maguire,2 Rachel Brough,1,3 Laura Fachal,4 Kyriaki Michailidou,5,6,7 Manjeet K. Bolla,7 Qin Wang,7 Joe Dennis,7 Thomas U. Ahearn,8 Irene L. Andrulis,9,10 Hoda Anton-Culver,11 Natalia N. Antonenkova,12 Volker Arndt,13 Kristan J. Aronson,14 Annelie Augustinsson,15 Heiko Becher,16 Matthias W. Beckmann,17 Sabine Behrens,18 Javier Benitez,19,20 Marina Bermisheva,21 Natalia V. Bogdanova,12,22,23 Stig E. Bojesen,24,25,26 Hermann Brenner,13,27,28 Sara Y. Brucker,29 Qiyun Cai,30 Daniele Campa,18,31 Federico Canzian,32 Jose E. Castelao,33 Tsun L. Chan,34,35 Jenny Chang-Claude,18,36 Stephen J. Chanock,3 Georgia Chenevix-Trench,37 Ji-Yeob Choi,38,39,40 Christine L. Clarke,41 NBCS Collaborators,42,43,44,45,46,47,48,49,50,51,52 Sarah Colonna,53 Don M. Conroy,4 Fergus J. Couch,54 Angela Cox,55 Simon S. Cross,56 Kamila Czene,57 Mary B. Daly,58 Peter Devilee,59,60 Thilo Dörk,23 Laure Dossus,61 Miriam Dwek,62 Diana M. Eccles,63 Arif B. Ekici,64 A. Heather Eliassen,65,66

Abstract

A combination of genetic and functional approaches has identified three independent breast cancer risk loci at 2q35. A recent fine-scale mapping analysis to refine these associations resulted in 1 (signal 1), 5 (signal 2), and 42 (signal 3) credible causal variants at these loci. We used publicly available in silico DNase I and ChIP-seq data with in vitro reporter gene and CRISPR assays to annotate signals 2 and 3. We identified putative regulatory elements that enhanced cell-type-specific transcription from the IGFBP5 promoter at both signals (30- to 40-fold increased expression by the putative regulatory element at signal 2, 2- to 3-fold by the putative regulatory element at signal 3). We further identified one of the five credible causal variants at signal 2, a 1.4 kb deletion (esv3594306), as the likely causal variant; the deletion allele of this variant was associated with an average additional increase in IGFBP5 expression of 1.3-fold (MCF-7) and 2.2-fold (T-47D). We propose a model in which the deletion allele of esv3594306 juxtaposes two transcription factor binding regions (annotated by estrogen receptor alpha ChIP-seq peaks) to generate a single extended regulatory element. This regulatory element increases cell-type-specific expression of the tumor suppressor gene IGFBP5 and, thereby, reduces risk of estrogen receptor-positive breast cancer (odds ratio = 0.77, 95% CI 0.74–0.81, p = 3.1 × 10−31).

Introduction

Over the last 15 years, genome-wide association studies have transformed our ability to map genetic variation underlying complex traits.1 The vast majority of variants identified in genome-wide association studies are non-coding and are thought to influence transcriptional regulation,2,3 a process which can be highly cell type and tissue

© 2021 The Authors. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
In this analysis we report fine-scale mapping of the 2q35 region in European and Asian individuals with breast cancer and control subjects from the Breast Cancer Association Consortium. We confirm three independent, high-confidence signals at 2q35 annotated by rs13387042 (signal 1), rs138522813 (signal 2), and rs16857609 (signal 3). We carry out functional annotation of credible variants at signals 2 and 3 and implicate the deletion variant (esv3594306) at signal 2 as causally associated with increased IGFBP5 expression and reduced breast cancer risk.

Material and methods

Fine-scale mapping of the 2q35 breast cancer risk locus

Fine-scale mapping of the 2q35 breast cancer risk locus was carried out as part of a large collaborative project; full details have been published. Briefly, for the current analysis we accessed data from 94,391 individuals with invasive breast cancer and 83,477 individuals of European ancestry and 12,481 individuals with invasive breast cancer and 12,758 control subjects of Asian ancestry from 87 studies participating in the Breast Cancer Association Consortium. All participating studies were approved by their appropriate ethics review board and all subjects provided informed consent.

Directly genotyped or imputed (info score > 0.8) calls for 10,314 SNPs mapping to a 1.4 Mb region at 2q35 (chr2:217,405,832–218,796,508; GRCh37/hg19) were available for analysis. At this threshold, the proportions of common variants (MAF ≥ 0.05), low-frequency variants (0.01 ≤ MAF < 0.05), and rare variants (0.001 ≤ MAF < 0.01) that could be analyzed were 89.7%, 68.5%, and 3.6%, respectively, for OncoArray and 64.2%, 40.5%, and 0.8%, respectively, for iCOGS. Analysis of the association between each SNP and risk of breast cancer was performed using unconditional logistic regression assuming a log-additive genetic model, adjusted for study and up to 15 ancestry-inflammatory principal components. p values were calculated using Wald tests. Forward stepwise logistic regression was used to explore whether additional loci in the fine-mapping region were independently associated with breast cancer risk. We carried out stratified analyses to determine whether each of the independent associations...
differed according to estrogen receptor (ER) status; heterogeneity between stratum-specific estimates was assessed using Cochran’s Q-test. All statistical analyses were carried out using R version 3.6.1.

**In silico annotation of credible variants**
Credible variants at each of the three independent signals were aligned with DNase I and ChIP-seq data (P300 [EP300 (MIM: 602700)], H3K27Ac, H3K4me1, FOXA1, GATA3 [MIM: 131320], ERα [ESR1 (MIM: 133430)]) generated in T-47D and MCF-7 breast cancer cells.15-17 (Table S1).

**Cloning of reporter assay constructs**
All reporter assay plasmids were derived using the pGL4 reporter vector (Promega). Reporter vectors were constructed using a restriction digest-based cloning approach. The IGFBP5 promoter and putative regulatory element regions (containing WT alleles) were synthesized as gblocks (Integrated DNA Technologies, full details in Table S2). Double restriction digests of plasmid or gBlock were performed using BglII and XhoI (for IGFBP5) or SalI and BamHI (for putative regulatory element regions) according to the manufacturer’s instructions (New England Biolabs [NEB]). Ligations were performed in a 3:1 insert:vector ratio using T4 DNA ligase (NEB), according to manufacturer’s instructions. Correct cloning was validated by Sanger sequencing using a commercially available service (Eurofins Genomics). Alternative (ALT) alleles of each variant were introduced into reporter vectors using QuikChange Lightning Site-directed Mutagenesis kit (Agilent Technologies), according to the manufacturer’s instructions. Accurate mutagenesis was confirmed by Sanger sequencing (Eurofins Genomics). All reporter gene constructs are shown in Figure S1.

**Cell Culture**
T-47D cells were grown in RPMI (GIBCO) supplemented with 10% FBS (GIBCO), 10 μg/mL human insulin (Sigma), and 100 μL penicillin with 100 μg/mL streptomycin (Sigma). HCT116 cells were grown in RPMI supplemented with 10% FBS, 100 μL penicillin, and 100 μg/mL streptomycin. HepG2 cells were grown in EMEM (LGC Standards-ATCC) supplemented with 10% FBS and 100 μL penicillin with 100 μg/mL streptomycin. MCF-7 cells (including derivative Cas9-expressing cell lines) and 293T cells were grown in DMEM (GIBCO) supplemented with 10% FBS and 100 μL penicillin with 100 μg/mL streptomycin. All cell lines were routinely short tandem repeat (STR)-typed and tested for mycoplasma contamination.

**Reporter assays**
Reporter assays were performed in T-47D, MCF-7, 293T, HCT116, and HepG2 cell lines. Antibiotics were removed from standard growth media 24 h before transfection to improve viability. For assays performed under standard conditions, approximately 16,000 cells were seeded per well of a 96-well plate for T-47D, MCF-7, and HepG2, and approximately 8,000 cells were seeded per well of a 96-well plate for 293T and HCT116. Transfection was performed upon reaching 70% confluency (~24 h after cell seeding). For assays performed after 17β-estradiol treatment, cells were first hormone starved for 48 h. Approximately 10,000 cells (T-47D) and 8,000 cells (MCF-7) were seeded, per well of a 96-well plate, in standard growth media and cultured for 24 h. The media was then replaced with phenol red-free media (GIBCO) supplemented with 10% charcoal-stripped FBS (GIBCO), 100 μL penicillin with 100 μg/mL streptomycin, 10 nM fulvestrant (I4409, Sigma), and 10 μg/mL human insulin (T-47D only). After 48 h, growth media was replaced with phenol red-free media supplemented with 10% charcoal-stripped FBS, 10 μg/mL human insulin (T-47D only), with the addition of either (1) 10 nM 17β-estradiol (E2758, Sigma) or (2) vehicle (ethanol). Transfection was performed upon reaching 80% confluency (6 h after 17β-estradiol or vehicle treatment).

Transfection was performed using X-tremeGENE HP DNA transfection reagent (Roche). Equimolar amounts of the test pGL4-based firefly luciferase vector and pRL-TK renilla luciferase control (Promega) were combined in a 3:1 reagent:DNA ratio in OptiMEM (Fisher Scientific). After a 30 min incubation at room temperature, 10 μL transfection mixture was added per well. Each biological replicate was performed in technical triplicates with non-transfected, mock-transfected, and pEGFP-transfected controls (Takara Bio Inc.). Cells were screened for luciferase activity 48 h after transfection using the Dual-Glo Luciferase Assay System (Promega) according to the manufacturer’s instructions.

**Confirmatory genotyping and sequencing of putative regulatory element 2 (PRE2)**
Four of the five variants mapping to PRE2 (rs72951831, rs199804270, rs138522813, and esv3594306) are highly correlated based on 1000 Genomes data (1KGp), with the ALT alleles of rs72951831, rs199804270, and rs138522813 all predicted to occur in combination with the ALT (deletion) allele of esv3594306 (esv3594306: rs72951831 $r^2 = 1$, $D' = 1$; esv3594306: rs199804270 $r^2 = 0.95$, $D' = 1$; esv3594306: rs138522813 $r^2 = 1$, $D' = 1$). However, rs572022984 (hg19, chr2:217955897) theoretically maps within the esv3594306 deleted region (chr2:217,955,891–217,957,273) casting doubt on whether the (imputed) rs572022984-del allele could occur in combination with the esv3594306 deletion allele. To clarify this, we genotyped all five variants in 300 randomly selected women participating in the Generations Study using MassARRAY (Agena Bioscience; full details of primers available on request). The number of carriers of the alternative (A-) allele at rs572022984 (MAF = 0.035) was 0 (expected number = 21; p = 0.00002). To confirm our genotyping, we carried out Sanger sequencing (Eurofins) of a 2.4 kb region spanning (chr2:217,955,866–217,958,000) in two individuals who were heterozygous at the linked PRE2 SNP rs138522813. Primers were: forward 5’-CGCTCCCTCCCTACATCGT-3’ and reverse 5’-TCTCTAGGCAAGCTCAG-3’. Sequencing confirms the presence of REF and ALT alleles of esv3594306, rs72951831, and rs199804270 (rs138522813 maps just outside the amplified region) but only REF alleles at rs572022984; on this basis we excluded rs572022984 from further analyses.

**Cloning of guides for CRISPR-based enhancer perturbation**
Guides were designed using the online design tool CHOPCHOP (http://chopchop.cbu.ubc.ca). Guides were selected based on their proximity to variants of interest and specificity scores. Full details are provided in Table S3. Cloning was performed essentially as described in Ran et al. Briefly, guides were produced as two complementary oligonucleotides with overhangs to facilitate cloning. Oligos were annealed with T4 Polynucleotide Kinase (NEB). The expression vector pKLV-U6gRNA(BbsI)-PGKpuro2ABFP (Addgene #50946) was digested using BbsI (NEB), and ligation performed using T4 DNA ligase (NEB). Cloning was validated by sequencing (Eurofins Genomics).
CRISPR-based enhancer perturbation

All CRISPR cell lines were derived from a parental MCF-7 cell line. Expression of each dCas9 construct was introduced by transduction with a specific Cas9-expressing lentivirus: pGH125_dCas9-Blast (Addgene #85417) for dCas9; pHHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene #60954) for dCas9-KRAB; Lent-hEFl-BLAST-dCas9-VPR (Dharmacon, CAS11916) for dCas9-VPR. Successfully transduced cells were then selected for by mCherry expression (dCas9-KRAB) or treatment with 10 μg/ml blasticidin (dCas9 and dCas9-VPR; Gibco). Cells were then seeded into 24-well plates at a density of 50,000 cells per well. 100 μL of sgRNA lentivirus was added. After 24 h, media was replaced and after 48 h cells were lysed using the Cells-to-Ct kit (Life Technologies) for subsequent gene expression analysis by RT-PCR.

Real-time PCR

Real-time PCR analysis of gene expression in cDNA samples was performed using Taqman probes (Life Technologies) for IGFBP2 (MM: 146731), IGFBP5, and RPL37A (MM: 613314) normalized to the housekeeping gene GAPDH (ThermoFisher; IGFBP2: Hs01040719_m1, IGFBP5: Hs00181213_m1, RPL37A: Hs011023 45_m1, GAPDH: Hs03929097_g1). Reactions of 5 μL were established using Taqman Universal Master Mix II, without UNG (Applied Biosystems) according to the manufacturer's instructions.

Statistical analysis of reporter gene assays and CRISPR-based enhancer perturbation

Firefly luciferase activity was internally normalized to renilla luciferase activity, and each condition normalized to the “IGFBP2 promoter-alone” (IGFBP2-PROM) construct. Setting IGFBP2-PROM to 1.0, for each putative enhancer-containing reporter gene construct we used t tests to test (1) $H_0$: the mean dual luciferase ratio does not differ from 1.0 and (2) $H_0$: the ALT construct does not differ from the REF construct. To compare mean dual luciferase ratios for each combination of SNP and SV at PRE2, we used three-way analysis of variance adjusting each variant for all other variants. To account for multiple testing, we used a Bonferroni corrected p value of 0.0056 (individual constructs, Figure 2; 9 tests) and 0.017 (PRE2 combinations, Figure 3; 3 tests).

Relative gene expression was calculated using the $\Delta \Delta C_T$ method. For the negative control sgRNAs (TAG-1 and TAG-2), we used t tests to test $H_0$: the relative gene expression does not differ from 1.0. To maximize the power of subsequent analyses, we then combined the negative control data and for each of the other sgRNAs we tested $H_0$: relative gene expression does not differ from the combined negative control relative gene expression. To account for multiple testing, we used a Bonferroni corrected p value of 0.017 (PROM sgRNAs Figures 4A; 3 tests per gene) and 0.0056 (PRE2 sgRNAs, Figures 4B and 4C; 9 tests per gene).

Ethics approval and consent to participate

All participating studies were approved by their appropriate ethics review board and all subjects provided informed consent.

Results

Fine-scale mapping of a 1.4 Mb region at 2q35 (chr2:217,407,297–218,770,424; GRCh37/hg19; Figure 1A) in combined data from up to 109,900 individuals with breast cancer and 88,937 control subjects of European Ancestry from the Breast Cancer Association Consortium confirmed the presence of three independent signals ($p < 5 \times 10^{-8}$; Figure S2) at this region. After conditioning on the top SNP at each of these three signals (signal 1, rs4442975; signal 2, rs138522813; signal 3, rs5838651), there were no additional high-confidence signals (defined as signals for which $p < 1 \times 10^{-6}$). Defining credible causal variants at each signal as variants with conditional p values within two orders of magnitude of the index variant there were 1, 5, and 42 credible causal variants at PRE1, PRE2, and PRE3, respectively (Table S4). Fine-scale mapping of this region in women of Asian Ancestry (12,481 affected individuals and 12,758 control subjects) did not identify any population-specific signals (all associations $p > 5 \times 10^{-8}$; Figure S3). None of the credible causal variants at signal 2 was present in women of Asian ancestry. The published causal variant at signal 1 (rs4442975) and all of the signal 3 credible causal variants (Table S5) were nominally associated with breast cancer risk in Asian women ($p < 0.05$). At signal 3, the index variants differ between Europeans and Asians (rs5838651 and 2:218265091:G:<INS:AME:ALU>:218265367, respectively) but none of the European credible causal variants could be excluded on the basis of the Asian data.

The T-allele of rs4442975 was associated with reduced breast cancer risk (per allele OR $= 0.88$, 95% CI 0.87–0.89, $p = 1.3 \times 10^{-75}$ and OR $= 0.94$, 95% CI 0.89–1.00, $p = 0.04$ in European and Asian women, respectively) and the delG-allele of rs5838651 was associated with increased risk (per allele OR $= 1.07$, 95% CI 1.05–1.08, $p = 1.5 \times 10^{-16}$ and OR $= 1.07$, 95% CI 1.03–1.11, $p = 0.0008$ in European and Asian women, respectively; Table 1). The delT-allele of rs138522813 was associated with reduced risk (carrier OR $= 0.80$, 95% CI 0.77–0.83, $p = 5.5 \times 10^{-32}$). Stratifying by ER status, the signal 1 (rs4442975) and signal 2 (rs138522813) SNPs were more strongly associated with ER+ disease; for the signal 3 SNP (rs5838651), there was no evidence that the ORs differed by ER status (Table S6).

Prioritization of credible variants for functional follow up

Fachal and colleagues used a Bayesian approach (PAINTOR) that combines genetic association, linkage disequilibrium, and enriched genomic features to determine variants with high posterior probabilities of being causal (Table S4). rs4442975, the only credible causal variant at signal 1 (posterior probability $= 0.84$), has previously been proposed to have a functional effect on breast cancer risk. Four of the five variants at signal 2 had posterior probabilities $\geq 0.20$ (combined posterior probability 0.997); none of the variants at signal 3 had posterior probabilities $> 0.15$. To further prioritize putative causal variants at signals 2 and 3, we aligned the 47 credible variants at these signals with markers of open chromatin (DNase I), active transcription (F300), active
enhancers (H3K27Ac, H3K4me1), and breast-relevant TFs (FOXA1, GATA3, ERα) generated in T-47D and MCF-7 breast cancer cells. Consistent with the PAINTOR posterior probabilities, four variants at signal 2 colocalized with at least one of these features. In addition, we identified two variants at signal 3 that colocalized with one of these features. These six variants were prioritized for further functional annotation.

Reporter gene assays of prioritized variants

For SNPs, we generated reference (REF) and alternative (ALT) constructs in which the putative regulatory element, defined in the first instance as a 500 to 700 bp region centered on the SNP or SNP pair (PRE2A rs572022984; PRE2B rs199804270 and rs72951831; PRE3 rs12694417 and rs12988242) and a structural variant esv3594306, which is defined by the presence (REF) or absence (ALT) of a 1.4 kb region (chr2:217,955,891–217,957,273; GRCh37/hg19), we generated separate REF constructs for PRE2A and PRE2B and a single ALT construct in which the centromeric sequences at PRE2A were juxtaposed to the telomeric sequences at PRE2B with the intervening 1.4 kb deleted.

Comparing the REF construct at each region with the IGFBP5 promoter construct (IGFBP5-PROM), there was evidence that two of the putative regulatory elements (PRE2B and PRE3) enhanced transcription from the IGFBP5 promoter.

Figure 1. 2q35 breast cancer risk locus

(A) Fine-scale mapping at 2q35 identified three high-confidence (p < 1 x 10^-6) signals annotated by rs4442975 (signal 1), rs138522813 (signal 2), and rs5838651 (signal 3). The putative target gene (IGFBP5) maps 360 kb, 399 kb, and 703 kb from signals 1, 2, and 3, respectively. All coordinates are based on GRCh37/hg19.

(B) Putative regulatory element 2 (PRE2; chr2:217,955,458–217,957,767) at signal 2 colocalizes with four highly correlated variants: three single-nucleotide polymorphisms (SNPs; rs572022984, rs199804270, and rs72951831) and a 1.4 kb insertion/deletion variant (esv3594306; indicated by a black bar). A fourth SNP (rs138522813) maps outside the proposed boundaries of PRE2. Regions of open chromatin (DNase I) and ChIP-seq binding peaks for transcription factors are shown as gray bars where the shade of gray indicates the strength of the ChIP-seq peak (light gray, weak binding; dark gray, strong binding). Also shown (yellow bars) are the coordinates of three reporter gene constructs (PRE2A, PRE2B, and PRE2DEL) and the locations of sequences targeted by nine small guide (sg)RNAs.

(C) PRE3 (chr2:218,305,944–218,306,443) indicated by a blue bar colocalizes with two SNPs (rs12694417 and rs12988242). Regions of open chromatin and ChIP-seq binding peaks are as in (B).
strong enhancer activity (PRE2B-REF/REF: fold change $\text{FC} = 27.9$, $p = 0.004$ and $\text{FC} = 28.7$, $p = 0.005$; PRE2-DEL-ALT/ALT: $\text{FC} = 50.5$, $p = 0.004$ and $\text{FC} = 44.9$, $p = 0.03$ in MCF-7 and T-47D, respectively). For PRE3 the activity was more modest and only significant ($p < 0.0056$; Material and methods) for the ALT allele in T-47D (PRE3-REF/REF: $\text{FC} = 1.8$, $p = 0.03$ and $\text{FC} = 2.9$, $p = 0.006$; PRE3-ALT/ALT $\text{FC} = 2.2$, $p = 0.006$ and $\text{FC} = 2.8$, $p = 0.003$ in MCF-7 and T-47D, respectively; Figure 2). To test these constructs for cell type specificity, we used HepG2 (hepatocyte carcinoma), 293T (embryonic kidney), and HCT116 (colorectal carcinoma) cells. The only construct that influenced transcription from the $\text{IGFBP5}$ promoter in these non-breast cells was PRE2DEL-ALT/ALT in 293T cells and with an effect size that was an order of magnitude lower ($\text{FC} = 1.9$, $p = 0.002$; Figure S4) compared to the breast cancer cell lines ($\text{FC} > 40$; Figure 2). Comparing PRE2 and PRE3 responsive to low-dose estradiol (Figure S5A and S5B) but only PRE2 showed a difference between alleles, with the (protective) PRE2-DEL-ALT/ALT allele once again being associated with significantly greater activity than the

**Table 1.** Association of rs4442975, rs138522813 and rs5838651 among women of European and Asian ancestry

<table>
<thead>
<tr>
<th>SNP</th>
<th>MAF</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
<th>MAF</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
<th>MAF</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4442975</td>
<td>0.49</td>
<td>36,471</td>
<td>37,251</td>
<td>0.88</td>
<td>0.86-0.89</td>
<td>$4.9 \times 10^{-35}$</td>
<td>0.48</td>
<td>57,920</td>
<td>46,226</td>
<td>0.88</td>
<td>0.87-0.90</td>
<td>1.7 $\times 10^{-42}$</td>
<td>0.88</td>
<td>87.08-0.89</td>
<td>$1.3 \times 10^{-75}$</td>
</tr>
<tr>
<td>rs138522813</td>
<td>0.035</td>
<td>-</td>
<td>-</td>
<td>0.81</td>
<td>0.76-0.86</td>
<td>$2.2 \times 10^{-12}$</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
<td>0.79</td>
<td>0.75-0.83</td>
<td>3.0 $\times 10^{-21}$</td>
<td>-</td>
<td>0.77-0.83</td>
<td>$5.5 \times 10^{-12}$</td>
</tr>
<tr>
<td>rs5838651</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>0.10</td>
<td>1.05-1.10</td>
<td>$4.2 \times 10^{-9}$</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>1.06</td>
<td>1.04-1.08</td>
<td>4.6 $\times 10^{-9}$</td>
<td>-</td>
<td>1.07-1.08</td>
<td>$1.5 \times 10^{-16}$</td>
</tr>
</tbody>
</table>

**Europeans**

<table>
<thead>
<tr>
<th></th>
<th>MAF</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4442975</td>
<td>0.49</td>
<td>36,471</td>
<td>37,251</td>
<td>0.88</td>
<td>0.86-0.89</td>
<td>$4.9 \times 10^{-35}$</td>
<td>0.48</td>
<td>57,920</td>
<td>46,226</td>
<td>0.88</td>
<td>0.87-0.90</td>
<td>1.7 $\times 10^{-42}$</td>
<td>0.88</td>
</tr>
<tr>
<td>rs138522813</td>
<td>0.035</td>
<td>-</td>
<td>-</td>
<td>0.81</td>
<td>0.76-0.86</td>
<td>$2.2 \times 10^{-12}$</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
<td>0.79</td>
<td>0.75-0.83</td>
<td>3.0 $\times 10^{-21}$</td>
<td>-</td>
</tr>
<tr>
<td>rs5838651</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>0.10</td>
<td>1.05-1.10</td>
<td>$4.2 \times 10^{-9}$</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>1.06</td>
<td>1.04-1.08</td>
<td>4.6 $\times 10^{-9}$</td>
<td>-</td>
</tr>
</tbody>
</table>

**Asians**

<table>
<thead>
<tr>
<th></th>
<th>MAF</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
<th>Cases Controls</th>
<th>OR</th>
<th>95% CI</th>
<th>$P_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4442975</td>
<td>0.87</td>
<td>4,994</td>
<td>5,866</td>
<td>0.96</td>
<td>0.88-1.04</td>
<td>0.29</td>
<td>0.88</td>
<td>7,487</td>
<td>6,892</td>
<td>0.93</td>
<td>0.87-1.01</td>
<td>0.07</td>
<td>12,481</td>
</tr>
<tr>
<td>rs138522813</td>
<td>0.035</td>
<td>-</td>
<td>-</td>
<td>0.97</td>
<td>0.97-1.10</td>
<td>0.29</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
<td>1.09</td>
<td>1.04-1.14</td>
<td>0.0005</td>
<td>-</td>
</tr>
<tr>
<td>rs5838651</td>
<td>0.61</td>
<td>-</td>
<td>-</td>
<td>1.03</td>
<td>0.97-1.10</td>
<td>0.29</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
<td>1.09</td>
<td>1.04-1.14</td>
<td>0.0005</td>
<td>-</td>
</tr>
</tbody>
</table>

*MAF, minor allele frequency
bOR, per allele odds ratio
c$P_1$, test of H0 no association between SNP and breast cancer risk
d$P_{\text{Het1}}$, test of H0 no difference between iCOGS and OncoArray data
e$P_{\text{Het2}}$, test of H0 no difference between European and Asian data

rs138522813 DEL allele is extremely rare in Asians (MAF ~0.05%) and was not analyzed in Asian data

**Figure 2.** Luciferase reporter assays following transient transfection of PRE2 and PRE3, REF and ALT constructs, into MCF-7, T-47D, and HepG2 cells

The PRE containing the reference (REF) allele at each SNP was cloned downstream of the $\text{IGFBP5}$ promoter to generate reference (REF) luciferase constructs. Alternative (ALT) alleles were generated by site-directed mutagenesis. Coordinates of the PREs are given in Table S2, diagrams are in Figure S1. Error bars denote standard deviations based on three independent experiments each done in triplicate. $p$ values were determined by t tests and a Bonferroni correction was applied to account for multiple testing. Comparing each PRE containing construct to $\text{IGFBP5-PROM}$, $*p < 0.0056$, $**p < 0.0005$, ***$p < 0.00005$.
PRE2B-REF/REF allele, this time in T-47D cells (MCF-7 FC = 1.5, p = 0.15; T-47D FC = 2.7, p = 0.002; Figure S5A).

The PRE2DEL-ALT/ALT construct comprises a haplotype of three tightly linked variants: the ALT alleles of the two SNPs (rs199804270:GA;G, rs72951831:G:T) with the ALT (deletion) allele of the structural variant (esv3594306) that brings two separate ERα, FOXA1, GATA3, and P300 ChIP-seq peaks into juxtaposition (Figure 1B). To differentiate individual effects, each allele of each SNP was introduced onto esv3594306 insertion and deletion backgrounds separately using site-directed mutagenesis. The PRE2A SNP (rs572022984) was not considered further due to technical issues (Material and methods). In a combined analysis, adjusting each variant for the other two variants, there was evidence that deletion constructs consistently showed greater activity than insertion constructs (MCF-7: DEL FC = 43.4, INS FC = 34.4, i.e., average additional FC for DEL = 1.3, p_{het} = 0.01; T-47D: DEL FC = 47.3, INS FC = 21.6, i.e., average additional FC for DEL = 2.2, p_{het} = 1.7 \times 10^{-8}; Figure 3).

CRISPR-based perturbation of PRE2

Reporter gene assays do not reflect the “normal” genomic context of a regulatory element. Specifically, the assay tests whether the putative regulatory element can influence expression in an episomal context21 and from a distance of a few kilobases; in vivo, PRE2 maps approximately 400 kb from the IGFBP5 promoter. To determine whether PRE2 acts as an enhancer element in a cellular context, we used a systematic CRISPR-based enhancer perturbation approach. We hypothesized that if PRE2 acts as an enhancer in vivo, targeting a catalytically inactive Cas9 (dCas9) fused to a repressive (KRAB) domain to regions within PRE2 would result in lower levels of expression of IGFBP5 (CRISPR interference; CRISPRi); by contrast, targeting dCas9 fused to an activating VPR domain would result in higher levels of expression of IGFBP5 (CRISPR activation; CRISPRa).22,23 We designed CRISPR single-guide (sg) RNAs to the ERα ChIP-seq peak at the centromeric breakpoint of the deletion (guides PRE2-1 and -2), within the esv3594306 deletion region (guides PRE2-3 to -6) and to the ERα ChIP-seq peak at the telomeric breakpoint of the deletion (guides PRE2-7 to -9; Figure 1B). As positive controls we designed sgRNAs to target the IGFBP5 promoter (guides PROM-1 to -3; Figure S6A) and the previously characterized causal variant (rs4442975, guide PRE1-1; Figure S6B). As negative controls we designed sgRNAs to the published genome-wide association study signal 1 tag SNP (rs13387042, guides TAG-1 and -2; Figure S6B). We used MCF-7 cell lines engineered to stably express (1) dCas9 with a repressive KRAB domain and (2) dCas9 with an activating VPR domain; as an additional control we used MCF-7 cells that expressed dCas9 without the KRAB or VPR domains.

In the dCas9 cell line, there was just one sgRNA (PROM-2) that influenced IGFBP5 expression; this sgRNA targets the IGFBP5 promoter, colocalizing with the transcription start site (TSS) and likely reduces expression of IGFBP5 by steric hindrance (60% reduction, p = 0.004; Figure S7A). In the CRISPRi setting, all three sgRNAs targeting the IGFBP5 promoter repressed IGFBP5 expression significantly to 8%–15% of levels in the negative controls (p = 0.001, p = 0.001, and p = 0.0008 for guides PROM-1, -2, and -3, respectively; Figure S8A). No sgRNA targeting non-promoter sequences influenced IGFBP5 expression (Figures S8A and S8B). In the CRISPRa setting, the sgRNA S′ to the IGFBP5 promoter (PROM-3; Figure 4A) enhanced IGFBP5 expression more than 60-fold (p = 0.00008) and the PRE-1-positive control sgRNA (PRE1-1) targeting rs4442975 also enhanced IGFBP5 expression (FC = 3.7, p = 0.006; Figure 4A). In addition, four of the nine sgRNAs targeting sequences at PRE2 enhanced IGFBP5 expression; specifically PRE2-1 and -2 targeting the ERα ChIP-seq peak at the centromeric deletion breakpoint (PRE2-1: FC = 3.7, p = 0.0005; PRE2-2: FC = 3.1, p = 0.001), PRE2-5 at the distal end of the deletion region (PRE2-5: FC = 3.2, p = 0.002), and PRE2-8 targeting the ERα ChIP-seq peak immediately telomeric to the deletion region (PRE2-8: FC = 5.3, p = 0.002; Figures 4B and 5A). None of the sgRNAs influenced expression of two genes mapping immediately 3′ to IGFBP5 (IGFBP2 and RPL37A; Figure 4C).
**Discussion**

Fine-scale mapping at the 2q35 breast cancer locus in women of European ancestry\(^3\) confirmed rs4442975 as the probable causal variant at signal 1 and reduced the number of credible causal variants at signal 2 from 14 to 5;\(^3,14\) at signal 3, however, there remained 42 credible causal variants that could not be excluded as causal on statistical grounds alone in either the European or the Asian data. Low-throughput functional approaches that are used to investigate putative causal variants, including reporter gene assays and CRISPR screens, become prohibitive with large numbers of credible causal variants and most single locus\(^11,14,24–38\) and global\(^1,6\) annotation studies have used co-localization of credible causal variants with markers of open chromatin, active histone modifications, and transcription factor binding in relevant cell types to prioritize credible causal variants for functional follow up. Of the 811 annotation tracks that were examined in a recent global fine-scale mapping analysis,\(^3\) credible causal variants were enriched at three types of genomic features that are relevant to long-range regulatory elements: (1) open chromatin in ER\(^+\) cell lines and normal breast, (2) the active histone marks H3K4me1 and H3K27ac in MCF-7 cells, and (3) ESR1, FOXA1, GATA3, and P300 TF binding sites. By aligning the five credible causal variants at PRE2 and the 42 credible causal variants at PRE3 with these marks (Table S4), we were able to prioritize 4 of the 5 credible causal variants at PRE2 and 2 of the 42 credible causal variants at PRE3 for follow-up studies. By taking this approach there is, inevitably, the possibility that we have excluded one or more causal variants from our follow-up analyses. For PRE2 this seems unlikely as we selected four out of the five credible causal variants for further follow-up studies. For PRE3 it is entirely possible, or even probable, that we failed to prioritize one or more causal variant(s); improving our ability to discriminate more accurately between potentially functional variants and large numbers of correlated variants will require genome-wide datasets with functional outputs\(^21,39,40\) generated in more relevant cellular disease models and taking advantage of single-cell technologies.\(^1\)

Using reporter gene assays, we have demonstrated that both the distal region of PRE2 (PRE2B) and the entire PRE3 region can enhance transcription from the IGFBP5 promoter in a cell-type-specific manner. Despite co-localizing with multiple markers, we found no evidence that
the proximal region of PRE2 (PRE2A) acts as an independent enhancer element. The ChIP-seq peaks at this region are, however, relatively weak (Figure 1B); combining data from both PRE2A alleles, in both breast cancer cell lines to increase our power (i.e., using 12 replicates rather than 3) the overall mean fold change for PRE2A was 1.14 (1.03–1.26, \( p = 0.01 \)), consistent with the presence of a very modest enhancer element. Comparing REF constructs with ALT constructs, we found no evidence that either of the credible causal variants at PRE3 (rs12694417, rs12988242) altered the activity of the PRE. This does not exclude these SNPs as functional; as above, modest effects on enhancer activity may be difficult to detect and variants that, for example, influence chromatin accessibility may not be detectable in transient assays.\(^1\) However, without preliminary in vitro evidence to suggest that one of these variants alters cell-type-specific transcription from the IGFBP5 promoter, pursuing further functional studies that are predicated on this very assumption seems unlikely to be fruitful. By contrast, one comparison that was consistent and significant between constructs and across the two breast cancer cell lines was that PRE2 deletion alleles had stronger enhancer activity than PRE2 insertion alleles.

The purpose of our CRISPR-based enhancer perturbation was 2-fold: specifically, to interrogate the PRE2 region within its normal genomic context and more generally to evaluate CRISPRi and CRISPRa approaches for interrogating long-range regulatory elements that harbor credible causal variants. As none of our PRE2 sgRNAs impacted IGFBP5 expression significantly in the CRISPRi setting, our analysis raises questions as to the utility of this approach for characterizing long-range regulatory elements (PRE2 maps approximately 400 kb telomeric to the IGFBP5 promoter). This is at odds with results of a systematic CRISPRi screen to identify enhancer elements in K562 cells, which demonstrated CRISPRi-mediated repression of c-MYC expression by sgRNAs targeting sequences mapping up to 1.9 Mb downstream of c-MYC.\(^2\) In this analysis, however, CRISPRi-mediated repression by these distal elements was modest compared to CRISPRi-mediated repression by more proximal elements and, even based on 12 biological replicates, of borderline statistical significance.\(^2\) By contrast, using CRISPRa we were able to confirm that one or more elements within PRE2 can act as a long-range regulatory element that specifically targets IGFBP5 (rather than IGFBP2 or RPL37A). Four of the nine guide RNAs targeting dCas9-VPR to sequences at PRE2 increased expression of IGFBP5; three of these colocalized with ERz, FOXA1, and GATA3 ChIP-seq peaks (PRE2-1, -2, -6, -8).
and -8) and a fourth (PRE2-5) mapped within the esv3594306 deleted region (Figure 5A). There were also two guides which targeted dCas9-VPR to sequences that map close to the distal ERx, FOXA1, and GATA3 ChiP-seq peak (PRE2-6 and -7) but did not increase IGFBP5 expression; this may reflect the very variable efficiency of different guide RNAs. We present a theoretical model in which we hypothesize that all of the PRE2 guides that increased expression of IGFBP5 increased the local density of activating TF domains by bringing a VPR domain into the proximity of a cluster of TF ChiP-seq peaks; one implication of the increase in IGFBP5 expression we observed with PRE2-5, which maps approximately 450 bp from the center of the nearest cluster of ChiP-seq peaks (Figure 5A), is that these regulatory elements may extend over relatively large (>1 kb) regions. This should not, perhaps, be surprising: at a subset of strongly activated E2-responsive enhancers, it has previously been shown that ERα recruits DNA-binding transcription factors in trans, to form a large (1–2 MDa) complex.

It has previously been suggested that sequences mapping to PRE2 act as a repressor element which, in the presence of low-dose estradiol, acts to reduce IGFBP5 expression. By contrast, our data support PRE2 acting as a powerful enhancer element with the deletion allele increasing expression of IGFBP5 over and above that of the insertion allele with or without estradiol stimulation. Overall, our data are consistent with a hypothetical model in which the juxtaposition of the two ERx, FOXA1, GATA3 binding sites at PRE2 by deletion of approximately 1.4 kb of intervening sequence generates a single extended binding region (Figure 5B) that is causally associated with increased enhancer activity, higher levels of expression of the putative tumor suppressor gene IGFBP5, and a reduction in breast cancer risk (OR = 0.77, p = 2.2 × 10⁻²⁹) that is largely restricted to ER⁻ disease.

In conclusion, we have identified putative enhancer elements at two additional 2q35 breast cancer risk loci. One of these, mapping approximately 400 kb telomeric to IGFBP5, enhances transcription from the IGFBP5 promoter by a factor of 30- to 40-fold. For this element we provide evidence that a deletion of 1.4 kb is causally associated with increased enhancer activity and suggest a mechanism for this increased activity.

Data and code availability
Summary results for all variants genotyped by the Breast Cancer Association Consortium BCAC (including rs45446698) are available at http://bcac.ccge.medschl.cam.ac.uk/. Requests for data can be made to the corresponding author or the Data Access Coordination Committee (DACC) of the Breast Cancer Association Consortium via email to BCAC@medschl.cam.ac.uk.

Supplemental information
Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.05.013.

Consortia
The NBCS Collaborators are Anne-Lise Børresen-Dale, Grethe I. Grenaker Alnaes, Kristine K. Sahlberg, Lars Ottestad, Rolf Kåresen, Ellen Schlichting, Marit Muri Holmen, Toril Sauer, Vilde Haakens, Olav Engebråten, Bjørn Naume, Alexander Fossá, Cecile E. Kiservud, Kristin V. Reinertsen, Åslaug Helland, Margit Riis, Jürgen Geisler, and OSBREAC. The kConFab Investigators are David Amor, Lesley Andrews, Yoland Antill, Rosemary Balleine, Jonathan Beesley, Ian Bennett, Michael Bogwitz, Leon Botes, Meagan Brennan, Melissa Brown, Michael Buckley, Jo Burke, Phyllis Butow, Liz Caldon, Ian Campbell, Deepa Chauhan, Manisha Chauhan, Georgia Chenevix-Trench, Alice Christian, Paul Cohen, Alison Colley, Ashley Crook, James Cui, Margaret Cummings, Sarah Jane Dawson, Anna deFazio, Martin Delatycki, Rebecca Dickson, Joanne Dixon, Ted Edkins, Stacey Edwards, Gelareh Farshid, Andrew Fellows, Georgina Fenton, Michael Field, James Flanagan, Peter Fong, Laura Forrest, Stephen Fox, Juliet French, Michael Friedlander, Clara Gaff, Mike Gattas, Peter George, Sian Greening, Marion Harris, Stewart Hart, Nick Hayward, John Hopper, Cass Hoskins, Clare Hunt, Paul James, Mark Jenkins, Alexa Kidd, Judy Kirk, Jessica Koehler, James Kollias, Sunil Lakhan, Mitchell Lawrence, Geoff Lindeman, Lara Lipton, Liz Lobb, Graham Mann, Deborah Marsh, Sue Anne McLachlan, Bettina Meiser, Roger Milne, Sophie Nightingale, Shona O’Connell, Sarah O’Sullivan, David Gallego Ortega, Nick Pachter, Briony Patterson, Amy Pearn, Kelly Phillips, Ellen Pieper, Edwina Rickard, Bridget Robinson, Mona Saleh, Elizabeth Salisbury, Christobel Saunders, Jodi Saunus, Rodney Scott, Clare Scott, Adrienne Sexton, Andrew Shelling, Peter Simpson, Melissa Southey, Amanda Spurdle, Jessica Taylor, Renea Taylor, Heather Thorne, Alison Trainer, Kathy Tucker, Jane Visvader, Logan Walker, Rachael Williams, Ingrid Winship, and Mary Ann Young. The ABCTB Investigators are Christine Clarke, Deborah Marsh, Rodney Scott, Robert Baxter, Desmond Yip, Jane Carpenter, Alison Davis, Nirmala Pathmanathan, Peter Simpson, Dinny Graham, and Mythily Sachchithananthan.

Declaration of interests
M.W.B. conducts research funded by Amgen, Novartis, and Pfizer. P.A.F. conducts research funded by Amgen, Novartis, and Pfizer and received honoraria from Roche, Novartis, and Pfizer. A.W.K. received research funding to her institution from Myriad Genetics for an unrelated project (funding dates 2017-2019). U.M. has stockownership in Abcodia Ltd. All other authors declare no conflict of interest.

Received: October 20, 2020
Accepted: May 25, 2021
Published: June 18, 2021

Web resources
1000 Genomes Project, https://www.internationalgenome.org/

References


