Fine scale mapping of the 17q22 breast locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGS)

Hatef Darabi1, Jonathan Beesley2,3*, Arnaud Droit4,5*, Siddhartha Kar4,6, Silje Nord5,7, Mahdi Moradi Marjaneh4,8, Penny Soucy6, Kyríaki Michailidou7,8, Maya Ghoussaini4, Hanna Fues Wahl9, Manjeet K. Bolla9, Qin Wang9, Joe Dennis9, M. Rosario Alonso9, Irene L. Andrulis10,11, Hoda Anton-Culver12, Volker Arndt13, Matthias W. Beckmann16, Javier Benitez15,16, Natalia V. Bogdanova17, Stig E. Bojesen18,19,20, Hiltlrun Brauch21,22,23, Hermann Brenner24,25,26,27,28, Annegien Broeks25,26, Thomas Bruning26, Barbara Burwinkel27,28, Jenny Chang-Claude29,30, Ji-Yeob Choi31,32, Don M. Conroy33, Fergus J. Couch33, Angela Cox34, Simon S. Cross35, Kamila Czene3, Peter Devilee36,37, Thilo Dörk38, Douglas F. Easton39,40, Peter A. Fasching41,42, Jonine Figueroa43,44, Henrik Flyger44,45, Eva Galle45,46, Montserrat García-Closas47,48, Graham G. Giles49,50, Mark S. Goldberg49,50, Anna González-Neira15, Pascal Guénet51, Christopher A. Haiman52, Emily Hallberg53, Ute Hamann54, Mikael Hartman55,56, Antoinette Hollestelle57, John L. Hopper48, Hidemi Ito58, Anna Jakubowska59, Nichola Johnson50,51, Dahee Kang51,52,60, Sofia Khan51, Veli-Matti Kosma52,61,62,63,64, Mieke Kriege57, Vessela Kristensen55,56,65,66, Diether Lambrechts55,67, Loïc Le Marchand67, Soo Chinn Lee68,69, Annika Lindblom70, Arpitaya Lophatanaanon71, Jan Lubinski72,73, Arto Mannnermaa74,75, Siranoush Manoukian76,77,78, Sara Margolin73,79, Keitaro Matsuo74,75, Rebecca Mayes80, James McKay75, Alfons Meindl76, Roger L. Milne77,78, Kenneth Muir77,78, Susan L. Neuhausen78, Heli Nevalinniemi51, Curtis Olswold79, Nick O’R5, Paolo Peterlongo79, Guillermo Pita80,81, Anja Rudolph29, Suleeporn Sangrajrang82, Elinor J. Sawyer83, Marjanka K. Schmidt84, Rita K. Schmutzler85,86, Caroline Seynaeve57, Mitul Shah7, Chen-Yang Shen77,87,88, Xiao-Ou Shu89, Melissa C. Southey90, Daniel O. Stram92,93, Harald Surowy27,28, Anthony Swerdlow93,94, Soo H. Teo92,93, Daniel C. Tessler94, Ian Tomlinson85,86, Diana Torres45,67, Thérèse Truong81, Celine M. Vachon53, Daniel Vincent94, Robert Winqvist85,86, Anna H. Wu92, Pei-Ei Wu92, Cheng Har Yip93, Wei Zheng80, Paul D. P. Pharaoah97, Per Hall98, Stacey L. Edwards9, Jacques Simard9, Juliet D. French2, Georgia Chenevix-Trench2 & Alison M. Dunning4

Genome-wide association studies have found SNPs at 17q22 to be associated with breast cancer risk. To identify potential causal variants related to breast cancer risk, we performed a high resolution fine-mapping analysis that involved genotyping 517 SNPs using a custom Illumina iSelect array (iCOGS) followed by imputation of genotypes for 3,134 SNPs in more than 89,000 participants of European ancestry from the Breast Cancer Association Consortium (BCAC). We identified 28 highly correlated common variants, in a 53 Kb region spanning two introns of the STXBP4 gene, that are strong candidates for driving breast cancer risk (lead SNP rs2787486 (OR = 0.92; CI 0.90–0.94; $P = 8.96 \times 10^{-15}$)) and are correlated with two previously reported risk-associated variants at this locus, SNPs rs6504950 (OR = 0.94, $P = 2.04 \times 10^{-09}$, $r^2 = 0.73$ with lead SNP) and rs1156287 (OR = 0.93, $P = 3.41 \times 10^{-11}$, $r^2 = 0.83$ with lead SNP). Analyses indicate only one causal SNP in the region and several enhancer elements targeting STXBP4 are located within the 53 kb association signal. Expression
studies in breast tumor tissues found SNP rs2787486 to be associated with increased STXBP4 expression, suggesting this may be a target gene of this locus.
Breast cancer is one of the most common epithelial malignancies in women. Ahmed et al. carried out a multi-stage genome wide association study (GWAS) for breast cancer susceptibility involving studies from the Cancer Genetic Markers of Susceptibility (CGEMS) and Breast Cancer Association Consortium (BCAC) and reported strong evidence for a susceptibility locus at 17q22 with single nucleotide polymorphism (SNP) rs6504950, OR = 0.95, 95% confidence interval (CI) 0.92–0.97, P = 1.4 × 10^{-8}. Turnbull et al. found confirmatory evidence for association with SNPs at the same locus; they reported a breast cancer risk association with SNP rs1156287 (OR = 0.91; 95% CI 0.85–0.97; P = 5.8 × 10^{-3}), which links 20 kb from originally reported SNP rs6504950 (r^2 = 0.91). Using data from the National Cancer Institute’s Breast and Prostate Cancer Cohort Consortium (BPC3), Campa et al. also confirmed the association with rs6504950 (OR = 0.92; 95% CI 0.88–0.97; P = 5.8 × 10^{-4}). Broeks et al. further investigated this association with respect to tumor estrogen receptor (ER) status, and reported that rs6504950 had stronger association with ER positive (ER+) status; OR = 0.93; 95% CI 0.90–0.95; P = 7.2 × 10^{-4}) than ER negative disease (ER−; OR = 1.00; 95% CI 0.95–1.05; P = 0.94). Tang et al. conducted a meta-analysis which further confirmed the association with SNP rs6504950 (OR = 0.93; 95% CI 0.87–0.99). SNP rs6504950 lies in an intron of STXB4 (Syntax binding protein 4) and two other genes are found within 200 kb including COX11 (cytochrome C assembly protein 11) and TOMM1 (target of myb1-like). As part of the Collaborative Oncological Gene-Environment Study (COGS), we conducted a comprehensive fine-scale mapping of this 17q22 breast cancer susceptibility locus using 517 SNPs chosen to give dense coverage across this locus. These were genotyped on a custom-designed Illumina Select genotyping array (iCOGS) in 50 studies participating in BCAC. We used these data to define the variants most strongly associated with risk, and combined these data with additional in-silico and functional data in an attempt to determine the most likely causal variants.

Material and Methods

Genetic Mapping. Tagging strategy for the fine-scale mapping. We defined the region for fine-mapping by identifying the flanking SNPs with minor allele frequency (MAF) > 2% and detectable correlation (r^2 > 0.1) with rs6504950, based on the 1000 genomes project European population (March 2010 Pilot version 60 CEU project data). From this 468 kb interval we selected all SNPs correlated with rs6504950 at r^2 > 0.1, plus a set of SNPs designed to tag all remaining SNPs with r^2 > 0.9. We thus aimed to genotype 525 SNPs, between chromosome 17 positions 52,816,899 and 53,284,506 (NCBI build 37 assembly), that had an Illumina designability score (DS) > 0.9. Of these, 517 were successfully genotyped on the array and passed QC filters.

iCOGS genotyping and imputation. Case and control samples were drawn from studies participating in the BCAC, of which 41 (total: 46,450 cases/42,600 controls) were predominantly of European ancestry and nine (6269 cases/6624 controls) of Asian ancestry. We performed iCOGS genotyping in four centres, as part of the Collaborative Oncological Gene-Environment Study (COGS). All BCAC studies had local human ethical approvals as described previously. We then used the genotype data from 517 SNPs that passed quality control to impute genotypes, among European subjects, at all additional known variants in the interval, using IMPUTE version 2.0 (IMPUTE2, without pre-phasing) and the 1000 genome project multi-population data (March 2012 version) as a reference panel. IMPUTE2 was run with default parameters and ‘effective size’ of the population Ne = 20,000. Using an imputation--r^2 > 0.3 in Europeans, we successfully imputed 3,134 SNPs (MAF ≥ 1%).

Statistical Analysis. For each SNP, we estimated the per-allele log-odds ratio (OR) and standard error using logistic regression, including principal components and per-study fixed-effects to capture study-specific
differences as previously described. For the analyses of European subjects, we included the first six principal components as covariates, together with a seventh component derived specifically for one study (LMBC) for which there was substantial inflation not accounted for by the components derived from the analysis of all studies (this component was set to zero for all other studies). For the analysis of Asian subjects, we included two principal components. We estimated per allele ORs under the assumption of a log-additive mode of inheritance, i.e. SNPs were coded according to the number of minor alleles 0, 1 or 2. We estimated main effects by subtype specific status (ER +/−) using case-control logistic regression and restricting the case sample to a specific subtype. We evaluated heterogeneity of association across tumour subtypes in a case-only analysis, treating subtype status as a dependent variable. We derived the P values by means of a likelihood-ratio test (one degree of freedom). Tests were two-sided. We carried out analyses separately among women of European and of Asian ancestry, defined by multiple dimensional scaling as previously described. We performed multiple logistic regression analyses to identify SNPs independently associated with each phenotype. To identify the most parsimonious model, we included all SNPs with a P < 10−4 and MAF ≥ 2% in the single SNP analysis in forward selection regression analyses, utilizing the step function in R with penalty term set to 20. We also used a joint analysis of all SNPs using a Bayesian-inspired penalised maximum likelihood approach (HyperLasso). To correctly account for uncertainty in the data resulting from the imputation process, we conducted analysis by regressing on the allele dosage for each genotype. For HyperLasso we utilized the most probable genotype as input, based on the posterior probability from the imputation algorithm (set to missing if all posterior probabilities were < 0.9).

Bioinformatic analyses. We combined multiple sources of in silico annotation from public databases to help identify potential functional SNPs. To investigate functional elements enriched across the previously defined fine-mapped region, more specifically in the region encompassing the strongest candidate causal SNPs, we analysed chromatin biofeatures data from the Encyclopedia of DNA Elements (ENCODE) Project named: Chromatin State Segmentation by Hidden Markov Models (chromHMM), DNase I hypersensitivity sites (DHS) and histone modifications of epigenetic markers H3K4, H3K9, and H3K27 in Human Mammary Epithelial Cells (HMEC) and MCF7 breast cancer cells. To identify putative target genes, we examined chromatin interactions between distal and proximal regulatory transcription-factor binding sites and gene promoters, using Chromatin Interaction Analysis by Paired End Tag (Chi-PET) in MCF7 cells. This detects genome-wide interactions associated with CCCTC-binding factor (CTCF) and DNA polymerase II (Pol2) – both involved in transcriptional regulation. Putative regulatory elements were determined using data from ENCODE, Roadmap Epigenomics, the “Predicting Specific Tissue Interactions of Genes and Enhancers” (PreSTIGE) algorithm, Hnisz and FANTOM. Intersections between candidate causal variants and regulatory elements were identified using Galaxy, and visualised in the UCSC Genome Browser. We used the ENCODE RNA-seq data to evaluate the expression of exons across the 17q22 locus in HMEC and MCF7 cell lines. The alignment files for HMEC (4 biological replicates) and MCF7 (19 biological replicates) were downloaded from ENCODE and the read count in the defined region was extracted and normalized in reads per million (RPM).

Allele specific expression (ASE) analysis. ASE analysis was performed using The Cancer Genome Atlas (TCGA) breast cancer data as described previously. SNP rs2787481, genotyped on the Affymetrix SNP Array 6.0 was used as a representative SNP for the candidate causal variants (r2 = 0.90 with rs2787486). SNP rs2787481 genotype calls and the corresponding confidence scores were retrieved using level 2 TCGA SNP array Birdseed data downloaded from TCGA portal. Genotypes with confidence scores equal to or above 0.1 were excluded.

We utilised RNA-sequencing data from 742 breast cancer samples from women of Caucasian ancestry. The corresponding RNA-sequencing BAM files and metadata are available from the Cancer Genomics Hub (CGHub). Markers used to assess relative allelic expression were exonic SNPs located in KIF2B, TOM1L1, COX11, STXB4, HLE, MMD, TMEM100, PCTP, and ANKFN1 extracted from dbSNP human Build 142. Homozygote marker SNPs, those with low coverage (less than 15x) and those within overlapping regions of the target genes, were removed. RNA-sequencing read counts on SNP sites for reference and alternative alleles were computed. The major allele fraction (μ), representing allelic imbalance for each marker SNP, was computed and an average of allelic imbalance for each gene was calculated for individual tumour samples. Marker SNPs with extreme μ values (μ > 0.75) were not included in the analysis. Level 3 SNP array data were downloaded from TCGA portal and GISTIC version 2.0.16 was used to identify copy number variations (CNVs) for each sample. Samples with low or high CNV levels, as presented in the gene-based GISTIC module report, were excluded from the analysis of the corresponding gene.

Allelic imbalance for the target transcripts was compared between rs2787481 heterozygote (CT) and homozygote (CC and TT) samples using Levene’s Test for equality of variances. KIF2B, TMEM100, and ANKFN1 were excluded from the statistical analyses as they did not have enough informative marker SNPs left after applying the filtering criteria.

Local gene expression by SNP (eQTL) association analysis. We examined the association of all genotyped or imputed SNPs with expression of nine genes (KIF2B, TOM1L1, COX11, STXB4, HLE, MMD, TMEM100, PCTP, and ANKFN1) in the 1 Mb region on either side of the fine-mapping interval, using data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study. METABRIC comprises normal tissues adjacent to tumours from breast cancer patients genetically confirmed to be of European ancestry. The samples (n = 135) were assayed for expression with the Illumina HT-12 v3 microarray. Matched germline SNP genotypes were derived using the Affymetrix SNP 6.0 array. Genotyping quality control and imputation for the METABRIC data are described in Guo et al. Association between genotype and expression was tested by linear regression with FDR control as implemented in the MatrixEQTL package in R.
Figure 1. Association Results for Overall Breast Cancer Risk. Directly genotyped SNPs are shown as filled black circles, and imputed SNPs ($r^2 > 0.3$, MAF $> 0.02$) are shown as open red circles, plotted as the negative log of the $P$ value against relative position across the locus. A schematic of the gene structures is shown. Signals, encompassing all SNPs with a likelihood ratio of $<1:100$ compared with the most significant SNP, are labelled and are shown as grey regions. The dashed purple line represents genome-wide significance ($P < 5 \times 10^{-8}$).

Four additional SNP-expression data sets were available and analysed separately: (1) NB116 consists of 116 Caucasian normal breast samples (the majority of Norwegian descent) with $n = 10$ tumour-adjacent normal biopsies. (2) BC241 consists of 241 Caucasian tumor (all stages) samples (the majority of Norwegian origin). These were both genotyped on the “iCOGS” SNP array, and gene expression levels were measured with Agilent 44 K23. (3) NB93 consists of 93 Caucasian adjacent normal breast samples from TCGA. Birdseed processed germline genotype data from the Affy6 SNP array were obtained from the TCGA dbGaP data portal24. (4) BC765 consists of 765 Caucasian breast tumour samples from TCGA24. Gene expression levels were assayed by RNA sequencing, RSEM (RNAseq by Expectation-Maximization25) normalized per gene or isoform, as obtained from the TCGA consortium24. Unexpressed and minimally expressed genes/isoforms whose sum in expression level was less than ten were excluded, and the data log2 transformed prior to analysis. The influence of SNPs on local gene expression (transcripts within 1 MB from the most strongly associated SNP) was assessed using a linear regression model, as implemented in the R library eMAP26. An additive effect was assumed by modelling the patient's number of copies of the rare allele, i.e. 0, 1 or 2 for a given genotype. Correction for multiple testing was performed using the false discovery rate (FDR) as implemented in the p.adjust function in R.

eQTL data from the Genotype-Tissue Expression (GTEx) project27 were downloaded from the v6 release.

Results

A total of 517 SNPs at chromosome 17 positions 52,816,899 to 53,284,506 (NCI build 37 assembly) were successfully genotyped using the iCOGs chip. Genotypes of other common variants across the region were imputed in the European studies using known genotypes in combination with a reference panel from the 1000 Genomes Project. 3,134 SNPs and insertion/deletion (indel) polymorphisms were reliably imputed (imputation $r^2$ score $> 0.3$, MAF $> 0.01$) and included in further analysis together with the 517 genotyped SNPs. In the European studies 139 genotyped or imputed SNPs were associated with overall risk of breast cancer ($P$ values $< 10^{-7}$) (Fig. 1). This set included SNPs rs6504950 and rs1156287 ($r^2 = 0.84$), both previously reported3,4 to be associated with breast cancer risk among Europeans (Supplementary Table 1).

Among the European ancestry studies the strongest association detected was with imputed SNP rs2787486 (OR [minor/major allele] = 0.92 [C/A]; 95% CI 0.90–0.94; $P = 8.96 \times 10^{-15}$), located in an intron of STXBP4 and strongly correlated with both previously reported GWAS hits ($r^2 = 0.83$ with rs1156287; $r^2 = 0.73$ with rs6504950). The strongest genotyped SNP association was rs244353 (OR = 0.92; 95% CI 0.90–0.94; $P = 5.75 \times 10^{-14}$) which lies $\sim 15$ kb from rs2787486 and is correlated with it ($r^2 = 0.99$). A regression model suggests that both rs2787486 ($P = 0.02$) and rs244353 ($P = 0.11$) are detecting the same risk association. To dissect further the observed associations all SNPs displaying evidence for association ($P < 10^{-4}$ and MAF $\geq 0.02$) with overall breast cancer risk (228 SNPs, Supplementary Table 1) in European studies were included in a forward stepwise regression model. This analysis identified a single association signal marked by top imputed variant rs2787486 (that is, no further SNPs were associated after adjustment for rs2787486). We also utilized penalized logistic regression models (based on the normal exponential gamma probability density) implemented in HyperLasso13, including all typed and imputed variants with an specified lambda of 0.05 and a penalty of 491 for overall risk (based on the sample size and a type I error of 0.001)26.

In this analysis, the best fitting model also included just one SNP, rs2787486.

On the assumption of a single causal variant, we calculated the likelihood ratio of each SNP relative to rs2787486 with respect to overall risk and SNPs with a relative likelihood ratio of $<1:100$ were excluded from further consideration26. After this exclusion process 28 SNPs (17 genotyped and 11 imputed), spanning 52.3 Kb (positions 53,176,211 to 53,228,543), remained as candidate causal variants (Table 1, Supplementary Table 2). These SNPs have very similar allele frequencies and are strongly correlated with SNP rs2787486. The two SNPs first reported to be associated with breast cancer were both excluded from this set of 28 candidate causal variants by likelihood ratio tests relative to SNP rs2787486 (likelihood ratios: 1:177439 for rs6504950 and 1:3271 for rs1156287). Caswell et al.30 subsequently identified marker rs11658717 as a potential causal candidate, but this variant is ranked 49th and has a likelihood ratio of 1:3875 relative to lead SNP rs2787486 - hence this has also been ruled out as potential causal candidate by our analysis.
Association with breast cancer subtypes. Based on data from European studies, 66 genotyped SNPs and 72 imputed SNPs were associated with risk of ER+ breast cancer (P values $10^{-7}$ to $10^{-14}$). The most strongly associated SNP for overall breast cancer (rs2787486) was also the most strongly associated for ER+ disease (OR = 0.91 (0.88–0.93), P = 1.39 × 10^{-14}), but was more weakly associated with ER− disease (OR = 0.95 (0.91–0.99), P = 1.77 × 10^{-13}, $P_{\text{heterogeneity}} = 0.017$). The most strongly associated SNP for ER− disease was c17_ pos53079506 (OR = 1.19 (1.07–1.33), P = 0.0017, $P_{\text{heterogeneity}} = 0.015$) located ~130 kb from rs2787486.

To determine whether there were additional subtype-specific association signals, we included all SNPs displaying evidence for association with ER+ disease (345 SNPs, P < 10^{-4} and MAF ≥ 2%) in a separate forward stepwise regression model. The top associated SNP was rs2787486 (OR = 0.91 (0.88–0.93), P = 1.39 × 10^{-14}) - the same SNP best-associated with overall risk and the same signal was localized by the HyperLasso search with penalty term set to 424. We calculated the likelihood ratio of each ER+ associated SNP ($r^2 > 0.6$) relative to rs2787486 and retained a list of 37 markers (17 genotyped, 20 imputed) with a likelihood ratio of $>1:100$. This list included all 28 candidate causal SNPs for overall risk, except for imputed variant rs187242. No stepwise selection for ER− risk was performed as none of the markers fulfilled the inclusion criteria.

Overall breast cancer and subtype risk association in Asian studies. Among Asian studies, the strongest association with overall breast cancer risk was observed for genotyped SNP rs244353 (OR = 0.91 (0.85–0.93), P = 2.57 × 10^{-3}). This SNP was one of the candidate causal SNPs in Europeans, and conferred a similar OR among Asians (rs244353, P = 0.91 (0.89–0.93), P = 1.77 × 10^{-13}, $P_{\text{heterogeneity}} = 0.017$). The most strongly associated SNP was rs244358 (OR = 0.92 (0.90–0.94), P = 1.77 × 10^{-13}) located ~130 kb from rs244353. The most strongly associated SNP for ER+ disease was c17_pos53079506 (OR = 1.19 (1.07–1.33), P = 0.0017, $P_{\text{heterogeneity}} = 0.015$) located ~130 kb from rs2787486.
analyses of overlap between candidate causal variants and regulatory sites. The 28 candidate causal variants (Table 1, Supplementary Table 5) fall in a 53.2 kb region spanning two introns of STXBP4 (Fig. 1). We mapped these to regulatory annotations from ENCODE. Analysis of DNase hypersensitivity clusters indicates that SNP rs244353 overlaps with a DHS in 23 cell lines, while rs2787481 and rs244317 show overlap in one and
three cell lines, respectively. However, none of these overlaps were observed in mammary cells. None of the candidate causal SNPs overlapped with histone modification marks (H3K4me1, H3K4me3, H3K9ac, H3K27ac) in the mammary cells line HMEC and MCF7 breast cancer cells (Fig. 2A). We analysed enhancer-promoter interactions using Chromatin Interaction Analysis by Paired End Tag (ChiA-PET) data for CCCTC-binding factor (CTCF) and DNA polymerase II (Pol2) in MCF7 breast tumour derived cells. Although multiple chromosomal interactions were observed across the locus for both Pol2 and CTCF in MCF7 cells there was a notable dearth of such interactions in the region encompassing the strongest candidate causal variants (Fig. 2B). No interactions were observed in Hi-C data from HMEC cells in this region (data not shown).

Data from Hnisz et al. 18 indicates the existence of several enhancers across the region, including a small one predicted to target the STXBP4 gene (observed in both HUVEC and CD4 memory cells) that includes the candidate causal variant SNP rs244353 (Fig. 3). However, PreSTIGE 17 indicates that an overlapping enhancer element (also containing rs244353) active in HepG2 cells may target the HLF gene. Another PreSTIGE element containing rs244336 and rs244337 is predicted to target HLF in colon crypt cells (Fig. 3).

Local Gene Expression analyses. ENCODE RNA-seq data show that COX11 and TOM1L1 are highly expressed in both MCF7 and HMEC cells lines while STXBP4 shows much lower expression levels (Fig. 2B). We performed allele specific expression (ASE) analysis using RNAseq and SNP array genotype data from TCGA 19. Allelic imbalance at marginal statistical significance ($P = 0.032$) in COX11 expression was detected with the alleles of candidate causal SNP rs2787481 ($r^2 = 0.90$ with rs2787486 ~1.3 Kb away) but not with any other genes within 1 Mb (TOM1L1, STXBP4, HLF, MMD, PCTP) using the same SNP (Supplementary Figure 3, Supplementary Table 6).

We also examined the associations of SNPs with the expression levels of the same local genes. In the normal tissue samples ($n = 135$) from the METABRIC study the top breast cancer candidate causal variant was also associated with COX11 expression levels (Supplementary Table 7). The most significant breast cancer associated SNP, rs2787486, was associated with differential expression of COX11 ($P = 0.00019$, FDR corrected $P = 0.05$) but not significantly associated with expression of any other genes after FDR correction. However, other SNPs across this region were more significantly associated with COX11 expression in normal breast tissue is unlikely to be the main driver of breast cancer risk. By contrast, no associations with COX11 expression were observed in the TCGA breast tumour samples with the top breast cancer risk SNPs (Supplementary Figure 5). However, in TCGA multiple SNPs associate with expression of the shortest isoform of STXBP4 (uc010dcc) with the top breast cancer risk SNP, rs2787486 having a FDR corrected $P = 4.0 \times 10^{-8}$ ($r^2 = 0.06$, Supplementary Figure 4). Other SNPs, including rs244317 and rs11658717 displayed more significant associations with expression of this isoform (FDR corrected $P = 3.8 \times 10^{-9}$, $r^2 = 0.07$, Supplementary Figure 4) than the top breast cancer risk SNP. The minor alleles are associated with increased expression of isoform uc010dcc, and explain 7% of the variation in its expression levels. Of note Caswell et al. 30 reported that the A-G base change of SNP rs11658717 mediates the use of different splice junction between exons 5 and 6 of the STXBP4 gene and thus generates the shorter uc010dcc isoform. Our expression data thus support this report but our
association evidence (likelihood ratio 1:3875 relative to lead SNP rs2787486) indicates that this SNP is unlikely to be a causal variant driving breast cancer risk.

We also interrogated candidate variants in the v6 data release from the Gene-Tissue Expression (GTEX) project. We found a significant association between the minor allele of SNP rs244353 and decreased expression of their measured STXBP4 (full length) isoform in multiple tissues including breast (n = 183; P = 1.3 × 10^-6; Supplementary Figure 6, Supplementary Table 8). These different METABRIC, TCGA and GTEX findings appear contradictory of each other: SNPs rs244353 and rs244317 are highly correlated (r^2 = 0.90 and yet their minor alleles are significantly associated with decreased STXBP4 expression in GTEX but increased expression of isoform (uc010dcc) in TCGA. One possible explanation is that the STXBP4 full length transcript (measured in GTEX) and the short transcript (uc010dcc, measured in TCGA) are regulated by different mechanisms.

**Discussion**

In this study, using more than 100,000 cases and controls of European and Asian ancestry participating in BCAC, we have confirmed previous reports of associations of SNPs in the 17q22 region with risk of breast cancer. Moreover, we identified a set of 28 strong candidate causal variants, of which one or more is the likely driver of their measured association. Of these, SNP rs2787486, which is correlated with previously reported candidates: rs6504950 (r^2 = 0.73), rs1156287 (r^2 = 0.83) and rs11658717 (r^2 = 0.84)\(^5\); was the most strongly associated variant with overall risk (OR = 0.92 (95% CI: 0.90–0.94), P = 8.96 × 10^-11). A similar magnitude of association was observed in both European and Asian women, consistent with the same causal variant mediating risk in both populations. The association was stronger for ER+ than ER- breast cancer.

All the remaining candidate causal variants lie in a 53 Kb region (positions 52,176,211 to 53,228,543) spanning two introns of the STXBP4 gene. None are predicted to alter the coding sequence of this gene and so it is most likely that the association is mediated through altering the regulation of one or more nearby genes. CHIA-Pet studies in the breast cancer MCF7 cell line reveal many chromatin interactions across the wider region (Fig. 2); however, there is a dearth of such interactions in the region encompassing the strongest candidate causal variants. Furthermore, in MCF7 or HMEC mammary cell lines there was no evidence of histone modification or open chromatin, indicative of the existence of regulatory regions, overlapping the best candidate causal variants, although such regions do exist in the wider region studied (Fig. 3). In this respect, this association signal differs from other breast cancer association signals in which strong evidence of regulatory elements in mammary cell lines has been observed. An enhancer is predicted by FANTOM in many cell types while data from Hnisz et al.\(^5\) (Fig. 3) indicates the existence of a small enhancer region, targeting STXBP4 (observed in both HUVEC and CD4 memory cells) that overlaps with candidate causal variant rs244353. However, PreSTIGE data indicate that nearby enhancer elements may target HLF in HepG2 and colonic crypt cells (Fig. 3).

Of the candidate genes in the region, both COX11 and TOM1L1 are highly expressed in both the HMEC and MCF7 breast cancer cell lines, while STXBP4 shows much lower expression (detected by RNAseq in TCGA). In support of COX11 and TOM1L1 being the targets of this breast cancer susceptibility locus, eQTL analyses in normal breast tissue showed borderline significant associations of the risk alleles of top candidate causal SNP rs2787486 with increased expression levels of both TOM1L1 and COX11; candidate SNP rs2787481 also showed evidence of allelic imbalance in COX11 expression. COX11 encodes a cytochrome c oxidase copper chaperone – a nuclear-encoded protein component of a mitochondrial-membrane-embedded respiratory complex and TOM1L1 encodes a Target of myb1-like1 membrane trafficking protein. Both genes are expressed in the majority of tissues examined in the Human Protein Atlas\(^5\); eQTL analysis in breast tumour tissues in TCGA find the risk allele of top candidate breast cancer risk SNP rs2787486, to be significantly associated with increased STXBP4 expression, but not with COX11 (Supplementary Figures 4 and 5).

Furthermore, Hnisz et al.\(^5\) indicates the presence of an enhancer element that overlaps with candidate causal SNP rs244353 and potentially targets STXBP4 (observed in both HUVEC and CD4 memory cells). Consistent with this, TCGA eQTL studies in breast tumour tissues find the risk allele of top candidate breast cancer risk SNP, rs2787486, to be significantly associated with increased STXBP4 mRNA expression. The STXBP4 gene encodes Syntaxin binding protein 4, a scaffold protein, which has been shown to stabilise and prevent degradation of an isoform of p63. P63 is, in turn, a member of the p53 tumour suppressor protein family and thus possibly a biologically more plausible candidate cancer gene than COX11 or TOM1L1.

We conclude that one or more of the 28 variants we identified is causally related to breast cancer risk, most likely through regulation of STXBP4, COX11 and TOM1L1, with the balance of the evidence favouring STXBP4 as the most important target. It remains possible, however, that the target gene(s) is more distant (>1 Mb) from the associated variants and so have not yet been considered. Further functional analyses will be required to determine the mechanism underlying this association and the downstream targets.

**References**

1. Ferlay, J. et al. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 127, 2893–2917, doi: 10.1002/ijc.25516 (2010).
2. Ferlay, J. et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136, E359–386, doi: 10.1002/ijc.29210 (2015).
3. Ahmed, S. et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41, 585–590, doi: 10.1038/ng.354 (2009).
4. Turnbull, C. et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 42, 504–507, doi: 10.1038/ng.586 (2010).
5. Campo, D. et al. Interactions between genetic variants and breast cancer risk factors in the breast and prostate cancer cohort consortium. J Natl Cancer Inst 103, 1252–1263, doi: 10.1093/jnci/djr265 (2011).
6. Brooks, A. et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet 20, 3289–3303, doi: 10.1093/hmg/ddr228 (2011).


Acknowledgements
The authors thank all the individuals who took part in these studies and all the researchers, study staff, clinicians, and other healthcare providers, technicians, and administrative staff who have enabled this work to be carried out. In particular, they thank: COGS: Andrew Berchuck (OCAC), Rosalind A. Eales, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Antonis Antoniou, Lesley McGuinness and Ken Offit (CIMBA), Joe Dennis, Andrew Lee, and Ed Dicks, Craig Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory and the staff of the CNIO genotyping unit, Francois Bacot, Sylvie Lapoissiere and Frederic Robidoux and the staff of the McGill University and Genome Quebec Innovation Centre, Sune F. Nielsen, Borge G. Nordestgaard, and the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility; ABCFS: Maggie Angelakos, Judi Maskiell, Gillian Dite; ABCS: Sten Corneliussen, Richard van Hien, Linde Braaf, Frans Hogervorst, Senno Verhoef, Laura van’t Veer, Emiel Rutgers, C Ellen van der Schoot, Femke Atsma; ACP: The ACP study wishes to thank the participants in the Thai Breast Cancer study. Special Thanks also go to the Thai Ministry of Public Health (MOPH), doctors and nurses who helped with the data collection process. Finally, the study would like to thank Dr Prat Boonyawongviroj, the former Permanent Secretary of MOPH and Dr Pornthep Sirivaranusong, the Department Director-General of Disease Control who have supported the study throughout; BBCS: Eileen Williams, Elaine Young-Mills, Kara Sargus; BIGGS: Niall McInerney, Gabrielle Colleran, Andrew Rowan, Angela Alonso; BSUCH: Peter Bugert, Medical Faculty Mannheim; CGPS: Staff and participants of the Copenhagen General Population Study. For the excellent technical assistance; Dorte Ultdall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthie Kjeldgård Hansen. The Danish Cancer Biobank is acknowledged for providing infrastructure for the collection of blood samples for the cases; CNIO-BCS: Charo Alonso, Daniel
Herrero, Nuria elvarez, Pilar Zamora, Primitiva Menendez, the Human Genotyping-CEGEN Unit (CNIO); CTS: The CTS Steering Committee includes Leslie Bernstein, James Lacey, Sophia Wang, Huiyan Ma, Yani Lu, and Jessica Clague DeHart at the Beckman Research Institute of City of Hope, Dennis Deapen, Rich Fender, Eun Jung Lee, and Fred Schumacher at the University of Southern California, Pam Horn-Ross, Peggy Reynolds, Christina Clarke Dur and David Nelson at the Cancer Prevention Institute of California, Argyrios Zogas and Hannah Park at the University of California Irvine; ESTHER: Hartwig Ziegler, Sonja Wolf, Volker Hermann, Christa Stempmaier, Katja Butterbach; GC-HBCC: Stefanie Engert, Heide Hellebrand, Sandra Kröber; GENICA: The GENICA Network: Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany [HB, Wing-Yee Lo, Christina Justenhooven], German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ) [HB], Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany [Yon-Dschun Ko, Christian Baish], Institute of Pathology, University of Bonn, Germany [Hans-Peter Fischer], Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany, Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany [TB, Beate Pesch, Sylvia Rabstein, Anne Lotz]; and Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany [Volker Harth]; HEBCS: Kirsimari Aaltonen, Karl von Smitten, Tuomas Heikkinen, Irja Erkkilä; HMBCS: Peter Hilleman, Hans Christiansen and Johann H. Karstens; KBCP: Eija Mäyliäinen, Helena Kemiäinen; kConFab/AOCS: We wish to thank Heather Thorne, Eveline Niedermayer, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab; LAABC: We thank all the data collection team, especially Annie Fung and June Yashiki; LMB: Gillian Peuteman, Dominiek Smeets, Thomas Van Brussel and Kathleen Corthouts; MARIE: Petra Seibold, Dieter Fleisch-Jans, Judith Heinz, Nadia Obi, Alina Vrieling, Sabine Behrens, Ursula Elber, Muhubbett Celik, Til Ochters and Stefan Nickels; NBCS: Bernard Peissel and Jacopo Azzollini Daniela Zaffaroni of the Fondazione IRCCS Istituto Nazionale dei Tumori (INT); Bernardo Bonanni, Monica Barile and Irene Ferone of the Istituto Europeo di Oncologia (IEO) and the personnel of the Cogentech Cancer Genetic Test Laboratory; MTLGBCS: We would like to thank Martine Tranchant (CHU de Québec Research Center), Marie-France Valois, Annie Turgeon and Lea Heguy (McGill University Health Center, Royal Victoria Hospital; McGill University) for DNA extraction, sample management and skillful technical assistance. J.S. is Chairholder of the Canada Research Chair in Oncogenetics; MYBRCA: Phuah Sze Yee, Peter Kang, Kang In Nee, Kavitta Sivanandan, Shivaani Mariapun, Yoon SooK-Yee, Daphne Lee, Teh Yueh Ching and Nur Aishah Mohd Taib for DNA Extraction and patient recruitment; NBCS: The following are NBCS Collaborators: Dr. Kristine K. Sahilberg, PhD (Department of Research, Vestre Viken Hospital, Drammen, Norway and Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalte, Oslo, Norway), Dr. Lars Ottestad, MD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalte, Oslo, Norway), Prof. Em. Rolf Käresen, MD (Institute of Clinical Medicine, University of Oslo, Oslo, Norway and Department of Breast- and Endocrine Surgery, Division of Surgery, Cancer and Transplantation, Oslo University Hospital, Oslo, Norway), Dr. Anita Langerod, PhD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalte, Oslo, Norway), Dr. Ellen Schlichting, MD (Section for Breast- and Endocrine Surgery, Department of Cancer, Division of Surgery, Cancer and Transplantation Medicine, Oslo University Hospital, Oslo, Norway), Dr. Marit Muri Holmen, MD (Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway), Prof. Toril Sauer, MD (Department of Pathology at Akershus University hospital, Lørenskog, Norway and Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway), Dr. Vilde Haakensen, MD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalte, Oslo, Norway), Dr. Olav Engebråten, MD (Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, Department of Oncology, Division of Surgery and Cancer and Transplantation Medicine, Oslo University Hospital, Oslo, Norway and Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway), Prof. Bjørn Naume, MD (Department of Oncology, Division of Surgery and Cancer and Transplantation Medicine, Oslo University Hospital-Radiumhospitalte, Oslo, Norway and K.G. Jebsen Centre for Breast Cancer, Institute for Clinical Medicine, University of Oslo, Oslo, Norway), Dr. Cecile E. Kiserud, MD (National Advisory Unit on Late Effects after Cancer Treatment, Department of Oncology, Oslo University Hospital, Oslo, Norway and Department of Oncology, Oslo University Hospital, Oslo, Norway), Dr. Kristin V. Reinertsen, MD (National Advisory Unit on Late Effects after Cancer Treatment, Department of Oncology, Oslo University Hospital, Oslo, Norway and Department of Oncology, Oslo University Hospital, Oslo, Norway), Assoc. Prof. Åslaug Helland, MD (Department of Genetics, Institute for Cancer Research and Department of Oncology, Oslo University Hospital-Radiumhospitalte, Oslo, Norway), Dr. Margit Riis, MD (Dept of Breast- and Endocrine Surgery, Oslo University Hospital, Ullevål, Oslo, Norway), Dr. Ida Bukholm, MD (Department of Breast-Endocrine Surgery, Akershus University Hospital, Oslo, Norway and Department of Oncology, Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital, Oslo, Norway), Prof. Per Eystein Lønning, MD (Section of Oncology, Institute of Medicine, University of Bergen and Department of Oncology, Haukeland University Hospital, Bergen, Norway), OSBREAC (Oslo Breast Cancer Research Consortium), Prof. Anne-Lise Borresen-Dale, PhD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalte, Oslo, Norway and Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Norway) and Grethe I. Grenaker Alnæs, M.Sc. (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalte, Oslo, Norway); NBHIS: We thank study participants and research staff for their contributions and commitment to this study; OBCS: We thank Arja Jukkola-Vuorinen, Mervi Grip, Saita Kauppila, Meeri Otsuikka and Kari Mononen for their contributions to this study; OFBCR:
Teresa Selander, Nayana Weerasooriya; ORIGO: We thank E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The LUMC survival data were retrieved from the Leiden hospital-based cancer registry system (ONCDOC) with the help of Dr. J. Molenaar; PBCS: Louise Brinhton, Mark Sherman, Neomila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner; pKARMA: The Swedish Medical Research Counsel; RBCS: Petra Bos, Jannet Bram, Ellen Crepin, Elisabeth Huijskens, Annette Heemskerk, the Erasmus MC Family Cancer Clinic; SASBAC: The Swedish Medical Research Counsel; SBCGS: We thank study participants and research staff for their contributions and commitment to this study; SBCS: Sue Higham, Helen Cramp, Ian Brock, Sabapathy Balasubramanian and Dan Connley; SEARCH: The SEARCH and EPIC teams; SGBCC: We thank the participants and research coordinator Kimberley Chua; SKKDKFZS: We thank all study participants, clinicians, family doctors, researchers and technicians for their contributions and commitment to this study; TNBCC: Robert Pilarski and Charles Shapiro were instrumental in the formation of the OSU Breast Cancer Tissue Bank. We thank the Human Genetics Sample Bank for processing of samples and providing OSU Columbus area control samples. UKBKS: We thank Breast Cancer Now and the Institute of Cancer Research for support and funding of the Breakthrough Generations Study, and the study participants, study staff, and the doctors, nurses and other health care providers and health information sources who have contributed to the study. We acknowledge NHS funding to the Royal Marsden/ICR NIHR Biomedical Research Centre. The authors would also like to acknowledge Dr Katherine A. Hoadley for normalization and sharing of all of TCGA BRCA RNAseq gene expression data. The work conducted for this project is supported by BCAC: BCAC is funded by Cancer Research UK [CI1287/A10118, CI1287/A12014]; and by the European Community’s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175); COGS: Funding for the COGS infrastructure came from: the European Community’s Seventh Framework Programme, grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (CI1287/A10118, CI1287/A10710, CI12292/A11174, CI1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund; ABCFS: The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow; ABCS: The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]; BMBRI-NL, which is a Research Infrastructure financed by the Dutch government (NWO 184.021.007); and the Dutch National Genomics Initiative; ACP: The ACP study is funded by the Breast Cancer Research Trust, UK; BBCC: The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen; BBCS: The BBCS is funded by Cancer Research UK and Breast Cancer Now and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN); BIGGS: ES is supported by NIHR Comprehensive Biomedical Research Centre, Guy’s & St. Thomas’ NHS Foundation Trust in partnership with King’s College London, United Kingdom. It is supported by the Oxford Biomedical Research Centre; BSUCH: The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholz Society and the German Cancer Research Center (DKFZ); CECCILE: The CECCILE study was funded by Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Ligue contre le Cancer Grand Ouest, Agence Nationale de Sécurité Sanitaire (ANSES), Agence Nationale de la Recherche (ANR); CGPS: The CGPS was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council and Herlev Hospital; CNIO-BCS: he CNIO-BCS was supported by the Instituto de Salud Carlos III, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI12/00070); CTS: The CTS was initially supported by the California Breast Cancer Act of 1993 and the California Breast Cancer Research Fund (contract 97-10500) and is currently funded through the National Institutes of Health (R01 CA77398). Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885. HAC receives support from the Lom V Smith Foundation (LV539420); ESTHER: The ESTHER study was supported by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe); GC-HBOC: The GC-HBOC (German Consortium of Hereditary Breast and Ovarian Cancer) is supported by the German Cancer Aid (grant no 110837, coordinator: Rita K. Schmutzler); GENICA: The GENICA was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, as well as the Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany; HEBCS: The HEBCS was financially supported by the Helsinki University Central Hospital Research Foundation, Academy of Finland (266528), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation; HERPACC: The HERPACC was supported by MEXT Kakenhi (No. 17015018 and 26253041) from the Ministry of Education, Science, Sports, Culture, and Technology of Japan, by a Grant-in-Aid
for the Third Term Comprehensive 10-Year Strategy for Cancer Control from Ministry Health, Labour and Welfare of Japan, by Health and Labour Sciences Research Grants for Research on Applying Health Technology from Ministry Health, Labour and Welfare of Japan, by National Cancer Center Research and Development Fund, and "Practical Research for Innovative Cancer Control (15ck0106177h0001)" from Japan Agency for Medical Research and development, AMED, and Cancer Bio Bank Aichi; HMBCS: The HMBCS was supported by a grant from the Friends of Hannover Medical School and by the Rudolf Bartling Foundation; KARBAC: Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantzows foundation; KBCP: The KBCP was financially supported by the special Government Funding (EVO) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations, and by the strategic funding of the University of Eastern Finland; kConFab/ACOS: kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. Financial support for the AOCs was provided by the United States Army Medical Research and Materiel Command [DAMD17-01-1-0729], Cancer Council Victoria, Queensland Cancer Fund, Cancer Council New South Wales, Cancer Council South Australia, The Cancer Foundation of Western Australia, Cancer Council Tasmania and the National Health and Medical Research Council of Australia (NHMRC; 400413, 400281, 199600). G.C.T. and P.W. are supported by the NHMRC. RB was a Cancer Institute NSW Clinical Research Fellow; LAABC: LAABC is supported by grants (IR8-0287, 3PB-0102, 5PB-0018, 10PB-0098) from the California Breast Cancer Research Program. Incident breast cancer cases were collected by the USC Cancer Surveillance Program (CSP) which is supported under subcontract of Health, The California Department of Health, the National Cancer Institute's Division of Cancer Prevention and Control Surveillance, Epidemiology, and End Results Program, under contract number N01CN25403; LMB: LMB is supported by the 'Stichting tegen Kanker' (232–2008 and 196–2010). Diether Lambrechts is supported by the FWO and the KULP2F/10/016-SymBioSysII; MARIE: The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR 1, 106332, 108253, 108419], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]; MBCSG: MBCSG is supported by grants from the Italian Association for Cancer Research (AIRC) and by funds from the Italian citizens who allocated the 5/1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale Tumori, according to Italian laws (INT-Institutional strategic projects “5 × 1000”); MCBCS: The MCBCS was supported by the NIH grants CA128978, CA111617, CA176785 an NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation and a generous gift from the David F. and Margaret T. Grohne Foundation and the Ting Tsung and Wei Fong Chao Foundation; MCCS: MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR); the Cancer Council South Australia, Cancer Council Tasmania and the National Health and Medical Research Council of Australia (NHMRC; 400413, 400281, 199600). G.C.T. and P.W. are supported by the NHMRC. RB was a Cancer Institute NSW Clinical Research Fellow; MEC: The MEC was supported by NIH grants CA63464, CA54281, CA098738 and CA132839; MTLGBCS: The work of MTLGBCS was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701; MYBRCA: MYBRCA is funded by research grants from the Malaysian Ministry of Science, Technology and Innovation (MOSTI), Ministry of Higher Education (UM.C/HIR/2008/016-SymBioSysII) and Research Initiatives Foundation (CARIF). Additional controls were recruited by the Singapore Eye Research Institute, which was supported by a grant from the Biomedical Research Council (BMRC08/1/35/19/550), Singapore and the National Medical Research Council, Singapore (NMRC/CG/SERI/2010); NBCS: The NBCS has received funding from the K.G. Jebsen Centre for Breast Cancer Research; the Research Council of Norway grant 193387/V50 (to A-L Børresen-Dale and V.N. Kristensen) and grant 193387/H10 (to A-L Børresen-Dale and V.N. Kristensen), South Eastern Norway Health Authority (grant 39346 to A-L Børresen-Dale) and the Norwegian Cancer Society (to A-L Børresen-Dale and V.N. Kristensen); NBHS: The NBHS was supported by NI grant R01CA100374. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485; OBCC: The OBCC was supported by research grants from the Finnish Cancer Foundation, the Academy of Finland (grant number 250083, 122715 and Center of Excellence grant number 251314), the Finnish Cancer Foundation, the Sigrid Juselius Foundation, the University of Oulu, the University of Oulu Support Foundation and the special Governmental EVO funds for Oulu University Hospital-based research activities; OFBCR: The Ontario Familial Breast Cancer Registry (OFBCR) was supported by grant UMI CA164901 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR; ORIGO: The ORIGO study was supported by the Dutch Cancer Society (RUL 1997–1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16); PBCS: The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA; pKARMA: The pKARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer; RBCS: The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009–4318); SASBAC: The SASBAC study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation; SBCGS: The SBCGS was supported primarily by NIH grants R01CA64277, R01CA148867, and R37CA70867. Biological sample preparation was conducted the Survey and
N.J., D.K., S.Khan, V.M.K., M.K., V.K., D.L., L.L.M., S.C.L., A. Lindblom, A.Lophatananon, J.L., A. Mannermaa, P.A.F., J.F., O.F., H.F., E.G., M.G.C., G.G.G., M.S.G., A.G.N., P.G., C.A.H., E.H., U.H., M.H., A.H., J.L.H., H.I., A.J., S.N., M.M.M., P.S., K. Michailidou, M.G., H.F.W., M.K.B., Q.W., J.D., M.R.A., I.L.A., H.A.C., V.A., M.W.B., J. Benitez, G.C.T. and A.M.D. Approved the final version of the manuscript and were involved in sample and data collection.

Provided critical review of the manuscript: H.D., J. Beesley, A.D., S. Kar, S.N., M.M.M., D.F.E., W.Z., S.L.E., J.D.F., P.A.E, J.F., O.F., H.E., E.G., M.G.C., G.G.G., M.S.G., A.G.N., P.G., C.A.H., E.H., U.H., M.H., A.H., J.L.H., H.I., A.J., N.J., D.K., S.Khan, V.M.K., M.K., V.K., D.L., L.L.M., S.C.L., A. Lindblom, A.Lophatananon, J.L., A. Mannermaa, S. Manoukian, S. Margolin, K. Matsuo, R.M., J.M., A.Meindl, R.L.M., K. Muir, S.L.N., H.N., C.O., N.O., P.P., P.G., K.P., A.R., S.S., E.J.S., M.K.S., R.K.S., C.S., M.S., C.Y.S., X.O.S., M.C.S., D.O.S., H.S., A.W., S.H.T., D.C.T., I.T., D.T., T.T., C.M.Y., D.V., R.W., A.H.W., P.E.W., C.H.Y., W.Z., P.D.P.P., P.H., S.L.E., J.S., J.D.F., G.C.T. and A.M.D.

Author Contributions
Analysed the data: H.D., J. Beesley, A.D., S. Kar, S.N., M.M.M. and A.M.D. J.Beesley prepared Figure 1,3 and Supplementary Figure 1, 2 and 6. A.D. prepared Figure 2. S.N. prepared Supplementary Figures 4 and 5. H.D., J. Beesley, A.D., S. Kar, S.N., M.M.M. and A.M.D. provided critical review of the manuscript: H.D., J. Beesley, A.D., S. Kar, S.N., M.M.M., D.F.E., W.Z., S.L.E., J.D.F., P.A.E., J.F., O.F., H.E., E.G., M.G.C., G.G.G., M.S.G., A.G.N., P.G., C.A.H., E.H., U.H., M.H., A.H., J.L.H., H.I., A.J., N.J., D.K., S.Khan, V.M.K., M.K., V.K., D.L., L.L.M., S.C.L., A. Lindblom, A.Lophatananon, J.L., A. Mannermaa, S. Manoukian, S. Margolin, K. Matsuo, R.M., J.M., A.Meindl, R.L.M., K. Muir, S.L.N., H.N., C.O., N.O., P.P., P.G., K.P., A.R., S.S., E.J.S., M.K.S., R.K.S., C.S., M.S., C.Y.S., X.O.S., M.C.S., D.O.S., H.S., A.W., S.H.T., D.C.T., I.T., D.T., T.T., C.M.Y., D.V., R.W., A.H.W., P.E.W., C.H.Y., W.Z., P.D.P.P., P.H., S.L.E., J.S., J.D.F., G.C.T. and A.M.D.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Darabi, H. et al. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Sci. Rep. 6, 32512; doi: 10.1038/srep32512 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016