eQTL colocalization analyses identify *NTN4* as a candidate breast cancer risk gene

Jonathan Beesley[1*#], Haran Sivakumaran[1*], Mahdi Moradi Marjaneh[1,2*], Wei Shi[1], Kristine M Hillman[1], Susanne Kaufmann[1], Nehal Hussein[1,3], Siddhartha Kar[4], Luize G Lima[1], Sunyoung Ham[1], Andreas Möller[1,3], Georgia Chenevix-Trench[1*], Stacey L Edwards[1*#], Juliet D French[1*]

[1] Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, 4029, Australia.

[2] Present address: UK Dementia Research Institute, Imperial College London, W12 0BZ, UK.

[3] Faculty of Medicine, The University of Queensland, Brisbane, 4006, Australia.

[4] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.

* These authors jointly contributed.

Correspondence: Jonathan.Beesley@qimrberghofer.edu.au

Stacey.Edwards@qimrberghofer.edu.au

ABSTRACT

Breast cancer genome-wide association studies (GWAS) have identified 150 genomic risk regions containing more than 13,000 credible causal variants (CCVs). The CCVs are predominantly noncoding and enriched in regulatory elements. However, the genes underlying breast cancer risk associations are largely unknown. Here, we used genetic colocalization analysis to identify loci at which gene expression could potentially explain breast cancer risk phenotypes. Using data from the Breast Cancer Association Consortium (BCAC) and quantitative trait loci (QTL) from the Genotype-Tissue Expression (GTEx) project and The Cancer Genome Project (TCGA), we identify shared genetic relationships and reveal novel associations between cancer phenotypes and effector genes. Seventeen genes, including *NTN4,* were identified as potential mediators of breast cancer risk. For *NTN4,* we showed the rs61938093 CCV at this region was located within an enhancer element that physically interacts with the *NTN4* promoter, and the risk allele reduced *NTN4* promoter activity. Furthermore, knockdown of *NTN4* in breast cells increased cell proliferation *in vitro* and tumor growth *in vivo*. These data provide evidence linking risk-associated variation to genes that may contribute to breast cancer predisposition.

The influence of common genetic variation on gene expression underlies a considerable proportion of the heritability associated with complex traits. Mapping of expression QTL (eQTL), where genetic variants are tested for association with gene expression levels, is widely used to identify genes that are regulated by trait-associated variants. Several studies have shown that eQTLs are enriched in cell types relevant to the trait of interest[1; 2]. For example, T cell-specific eQTLs are over-represented for autoimmune risk alleles and monocyte-specific eQTLs for Alzheimer's [MIM: 104300] and Parkinson's [MIM: 168600] disease alleles[2]. For breast cancer [MIM: 114480], several studies have used eQTL data from tumor and normal tissues datasets to identify candidate target genes[3-6]. Recent studies have also showed that breast cancer risk variants could regulate genes in cells of the tumor microenvironment, such as immune cells and fibroblasts[7; 8]. Because eQTLs are widespread, overlap between GWAS and eQTL signals is likely to occur by chance when using nominal significance

levels. To mitigate false positive findings it is therefore important to show that the same genetic signal underlies gene expression and disease susceptibility.

Several statistical colocalization approaches have been developed to determine whether molecular traits (e.g. gene expression) and a disease trait share common causal variants. The simplest Bayesian model used in tools such as QTLMatch[9] and COLOC[10] tests for colocalization for two traits and determines whether they are driven by distinct variants or share a single causal signal. For example, Parker et al used COLOC to identify 32 emphysema-associated [MIM: 130700] regions where it is likely that colocalized GWAS and eQTL signals arise from the same causal variant[11]. Additional functional studies then showed that the emphysema-associated variant rs1690789 regulates *TGFB2* (encoding transforming growth factor beta 2 [MIM: 190220]) expression in human lung fibroblasts. A recent implementation of COLOC, called HyPrColoc (Hypothesis Prioritization in multi-trait Colocalization), identifies colocalized association signals using summary statistics on large number of traits[12]. This method has been used to identify regulatory loci underlying quantitative haematopoietic traits[13].

In this study, we extracted eQTL association effect estimates and standard errors for all variants at the 150 breast cancer risk loci previously analysed by BCAC[14] (mean region size = 1.09 Mb). GWAS summary data were available for overall breast cancer risk from 122,977 cases and 105,974 controls[3]; and for estrogen receptor negative (ER-) breast cancer risk from 21,468 cases and 100,594 controls, combined with 18,908 *BRCA1* mutation carriers (9414 with breast cancer)[15] all of European ancestries. Variant IDs were converted to GRCH38 build co-ordinates and harmonized with GTEx data (0.86% failed conversion and were dropped from the analysis). The GTEx version 8 release includes data from normal breast tissue from 396 individuals. GTEx eQTL association data for variants within ±1 Mb windows of transcription start sites were extracted based on the variants present in the breast cancer risk data. Colocalization of the GWAS and eQTL signals were calculated using the HyPrColoc R package[12]. Breast cancer risk phenotypes and each proximal gene were analyzed separately with default parameters. Signals were considered to be plausibly colocalizing if posterior probability for colocalization (PPFC) > 0.7, resulting in a false discovery rate of 5%[12].

We identified 17 genes at 14 loci where the GTEx eQTL association *P* values are < 10^{-6} (**Table 1**). For every locus, all candidate SNPs met the GWAS significance *P* value threshold ($5x10^{-8}$) for overall or ER- breast cancer risk (**Table 1**). For 11 loci (*NTN4, PIDD1* [MIM: 605247]*, CBX8* [MIM: 617354]*, L3MBTL3* [MIM: 618844], *RCCD1* [MIM: 617997], *PRC1-AS1*, *SSBP4* [MIM: 607391], *MARCH11* [MIM: 613338], *ZNF596*, *RP5-855D21.3 and RP11-53O19.1*), the candidate colocalized SNPs have been previously nominated as strong candidate causal signals using multivariate logistic regression[14] (**Table 1** and **Figure 1**). However, at six loci (*ATG10* [MIM: 610800]*, CCDC88C* [MIM: 611204]*, PPM1K* [MIM: 611065]*, RP11-250B2.3, RP1-265C24.5* and *RP11-250B2.5*), the colocalized signals are independent secondary signals based on stepwise multinomial logistic regression analysis ($10^{-6} < P < 10^{-4}$)[14]. While this does not rule out causality, larger GWAS would be required to confirm genome-wide significance (**Figure S1**)[14].

Published computational predictions of target genes at breast cancer risk loci using the INQUISIT pipeline (which interrogates data including ChIA-PET, Hi-C, ChIP-seq and eQTL data independent of GTEx) provide further support for ten colocalized genes (**Table S1**)[3; 14]. Of these, *NTN4, PIDD1, L3MBTL3* and *RCCD1* have the strongest evidence from functional genomics data. Transcriptome-wide association studies also suggest that 13 of the 17 genes are regulated by breast cancer risk variants[5-7; 16; 17] (**Table S1**). Moreover, previous eQTL analysis based on TCGA breast tumor data have identified three of these candidate genes[3; 14] (**Figure S2**). For three genes (*PIDD1, L3MBTL3* and *SSBP4*), CCVs are located in the promoter regions, and for *PIDD1* previous reporter assays indicate the risk haplotype increases promoter activity[3]. Our recent capture Hi-C data showed also chromatin looping occurs between putative regulatory regions containing CCVs and the promoters of four genes (*NTN4, PRC1-AS1, ATG10* and *RP1-265C24.5*) in breast cell lines[18]. For the remaining loci, multiple CCVs were located in the introns of target genes and/or intergenic regions, but lacked demonstrable CCV-gene interactions. It is possible that some *cis*-regulatory interactions are only detected in specific breast cell subpopulations, or that CCVs are acting through other mechanisms such as perturbation of pre-messenger RNA splicing or altered noncoding RNA stability, structure and/or function. Of note, three genes (*PIDD1*, *CBX8*, and *L3MBTL3*) contain breast cancer CCVs in

their exons which are predicted to change the amino acid sequence, thus we cannot rule out that these are functional variants that affect the protein product.

One high probability colocalization signal, targeting *NTN4,* was detected at a locus at 12q22 (**Table 1**, **Figures 1** and **2**). Genetic fine-mapping studies have identified one risk signal at 12q22 that contains two CCVs (rs61938093 and rs17356907; odds ratio = 1.094, $r_2 = 1$)[14]. Both CCVs fall within putative regulatory elements (PREs) marked by open chromatin in B80T5 and MCF10A non-tumorigenic breast cell lines (**Figure 3A**). The PREs map to a large intergenic region between *USP44* [MIM: 610993] (encoding ubiquitin-specific protease 44) and *NTN4* (encoding Netrin 4; **Figure 3A**). Using promoter capture HiC data[18], we observed that the PREs frequently participate in long-range chromatin interactions with the *NTN4* promoter in non-tumorigenic and tumorigenic breast cell lines (**Figures 3A** and **S2A**). Notably, no other eQTLs or chromatin interactions from the PRE to promoter regions were detected in the breast cell lines we examined (**Figures 3A** and **S2A)**[18], suggesting *NTN4* is the likely target gene at this signal.

To determine how the PRE alters *NTN4* transcriptional activity, we targeted a nuclease-defective dCas9 fused to the Kruppel-associated box (lentiviral vector pHR-SFFV-dCas9-BFP-KRAB; a gift from Stanley Qi & Jonathan Weissman, Addgene plasmid #46911) to the PRE. Two independent single-guide RNAs (sgRNAs) targeting the PRE were designed (**Table S2**) and cloned into the lentiviral vector pgRNA-humanized (a gift from Stanley Qi, Addgene plasmid #44248). Lentiviral particles were produced from HEK293 cells transfected with accessory plasmids pCMV-dR8.91 and pCMV-VSV-G (gifts from David Harrich, QIMR Berghofer), and with dCas9-KRAB or pgRNA constructs using Lipofectamine 2000 (Life Technologies). Supernatants from dCas9-KRAB and pgRNA cultures were mixed and transduced into Bre80-TERT1 breast cells. Cells expressing both dCas9-KRAB (co-expressing blue fluorescent protein) and pgRNA (co-expressing mCherry) were enriched by FACS on the Aria IIIu platform (Becton-Dickinson). Notably, silencing of the PRE significantly reduced *NTN4* expression in Bre80-TERT1 cells, suggesting that the PRE acts as an transcriptional enhancer (**Figure 3B**).

The regulatory capability of the PRE, combined with the effects of the CCVs, was further examined in reporter assays. An *NTN4* promoter-driven luciferase reporter construct was generated by the insertion of a PCR amplified genomic fragment into the KpnI/HindIII sites of pGL3-basic (Promega). A 1010 base pairs (bp) fragment containing a PRE1, with either the risk or protective alleles of rs61938093, or a 983 bp fragment containing a PRE2, with either the risk or protective alleles of rs17356907, were synthesized as gBlocks (Integrated DNA Technologies) and then cloned into the BamHI/SalI sites of the *NTN4*-promoter construct (genomic coordinates and primers are listed in **Table S2**). MCF10A and Bre80-TERT1 breast cells were transfected with the reporter constructs and luciferase activity was measured 24 h post-transfection using the Dual-Glo Luciferase System (Promega). To correct for any differences in transfection efficiency, *Firefly* luciferase activity was normalized to *Renilla*. Reporter assays confirmed strong enhancer activity of the PRE1 on the *NTN4* promoter in MCF10A and Bre80-TERT1 cells and inclusion of the rs6198093 risk allele significantly reduced *NTN4* promoter activity (**Figure 3C**). In contrast, inclusion of PRE2 with the protective or risk alleles of rs17356907 had no significant effects on the NTN4 promoter activity. pGL3-promoter luciferase results ???? These results suggest that rs61938093 alters transactivation of NTN4, but it is possible that the CCVs also influence recruitment of key proteins required for chromatin looping between the enhancer and *NTN4* promoter, which would not be observed in a reporter assay.

To assess the potential impact of the CCVs on chromatin looping, quantitative allele-specific 3C was performed in heterozygous MCF10A and T47D breast cell lines. 3C libraries were generated using HindIII as previously described[19]. Three independent 3C libraries or genomic input DNA were amplified for 15 cycles with two separate xxxxx 3C specific or genomic DNA PCR primers (primers are listed in **Table S2**) and purified by QIAGEN columns. Allele-specific PCR products were then quantified using a custom TaqMan SNP genotyping assay for rs61938093 (Life Technologies) on the Rotor-Gene 6000 platform. Purified PCR products were Sanger sequenced by the Australian Genome Research Facility (AGRF). The results showed a preference for the protective *t*-allele (**Figures 3D, 3E** and **S2B-S2E**), indicating that risk alleles may abrogate looping between the enhancer and *NTN4* which in turn may reduce *NTN4* expression.

Electrophoretic mobility shift assays (EMSAs) then assessed transcription factor (TF) binding for the protective and risk alleles of the CCVs. Nuclear lysates were prepared from Bre80-TERT1 and MCF10A breast cells using the NE-PER nuclear and cytoplasmic protein extraction kit (ThermoFisher). Biotinylated oligonucleotides representing either the risk or protective allele were synthesized (Integrated DNA Technologies; **Table S2**) and annealed to form double-stranded duplexes. Duplex-bound complexes were resolved by electrophoresis in 10% (w/v) Tris-borate-EDTA polyacrylamide (Lonza) and transferred to positively-charged nylon membranes by semi-dry transfer (Bio-Rad). Membranes were processed using the LightShift Chemiluminescent EMSA kit (ThermoFisher) and visualized with the C-DiGit blot scanner. The EMSAs showed that rs61938093 and rs17356907 altered protein binding *in vitro* in Bre80-TERT1 and MCF10A cell lysates (**Figures 3F, S3A** and **S3B**). *In silico* prediction tools including HaploReg[20] and Alibaba2[21] predicted both CCVs to alter TF binding. However, EMSAs using competitor DNA against predicted and other breast-relevant TFs were unable to identify the specific protein(s) binding to the alleles (**Figures S3C** and **S3D**).

We examined expression of *NTN4* in matched normal and cancerous breast tissues using TCGA RNA-seq data. *NTN4* was more highly expressed in normal tissue, a mixture of cell types, compared to adjacent tumor samples (**Figure 4A**) and is expressed across the histological subtypes, albeit with lower expression in the basal subtype (**Figure 4B**). To explore the effect of reduced *NTN4* on breast cancer cell proliferation, MCF7 cells were transfected with ON-TARGETplus negative control or *NTN4* siRNA smartpools (Dharmacon) using RNAiMAX (Life Technologies). *NTN4* silencing was confirmed by TaqMan qPCR gene expression assay 72 h post-transfection (**Figures S3E** and **S3F**). Notably, *NTN4* depletion promoted anchorage-dependent and -independent cell growth in MCF7 cells (**Figures 4C** and **4D**). To assess the effect of reduced *NTN4* on tumor growth, we stably depleted *NTN4* in MCF7 cells by targeting dCAS9-KRAB to the promoter of *NTN4,* and injected the cells in the mammary fat pad of nude mice. Female BALB/c-Foxn1^{nu}/Arc mice were first subcutaneously implanted with 17β-estradiol (0.72 mg/pellet, 90 day release; Innovative Research of America) at 8 weeks of age. MCF7 control-CRISPRi or *NTN4*-CRISPRi cells were orthotopically injected into mammary fatpads 3 days later at 10^7 cells per mouse (6-7 mice per cell line). Tumor

volumes were measured every 2 days until experimental end, at which point mice were euthanized and their tumors excised and weighed. All animal procedures were conducted in accordance with Australian National Health and Medical Research regulations on the use and care of experimental animals, and approved by the QIMR Berghofer Medical Research Institute Animal Ethics Committee (P1499). Compared to control MCF7 cells containing non-targeting sgRNA, *NTN4* depletion led to a marked increase in tumor growth (**Figures 4E, 4F** and **S3G**), which was reflected in increased tumor weight (**Figure 4G**).

NTN4 encodes the Netrin-4 secreted protein which has been implicated in various developmental processes including axon guidance, angiogenesis, mammary and lung morphogenesis[22]. Several studies show that NTN4 is involved in cancer, but the exact role of NTN4 appears to be dependent on the cancer type. For example, NTN4 knockdown reduces cell proliferation and motility in gastric cancer [MIM: 613659] and melanoma [MIM: 155600][23; 24], but promotes cell migration and invasion in colorectal cancer [MIM: 114500] and breast cancer[25; 26]. *NTN4* has also been implicated in breast cancer progression. For example, reduced NTN4 is reported to promote migration and invasion of breast cancer cells through epithelial to mesenchymal transition[26]. In addition, NTN4 has been shown to be an independent biomarker for prognosis of survival in breast cancer[27; 28]. We and others have demonstrated that SNPs can alter chromatin loop formation between promoters and enhancers[29; 30]. Here, we provide evidence that the same mechanism may explain how breast cancer CCVs alter *NTN4* expression and that suppressed *NTN4* increases cancer-related processes including cell proliferation and tumor growth. However, we acknowledge that further functional studies will be required to clarify how NTN4 contributes to breast tumor initiation.

Seven additional colocalized target genes have prior evidence for a functional role in cancer. For example, PIDD1 (p53-induced death domain protein 1) is implicated in DNA-damage-induced apoptosis and tumorigenesis[31]. CBX8 is overexpressed in breast cancer and correlates with poor survival[32]. CBX8 functions by interacting with the H3K4 methyltransferase complex component WDR5 to activate genes involved in Notch signaling and promote breast tumorigenesis[32]. Furthermore, Guo et al showed ectopic ATG10 (autophagy related 10) overexpression decreases

colony formation of MDA-MB-231 breast cancer cells, suggesting it may play a role in breast tumorigenesis[5]. Notably, ten genes have no reported involvement in breast tumorigenesis and may represent new genes that influence the susceptibility to breast cancer. This list includes five lncRNAs which are arguably more challenging to investigate as they can have multiple functions. However, there is increasing evidence that dysregulated lncRNAs contribute to breast cancer etiology[29; 33; 34].

In summary, we used eQTL colocalization to link breast cancer risk variants to seventeen target genes, including some potential cancer driver genes, but many with no reported role in breast cancer etiology. However, even with demonstration of shared genetic signals, it is as yet unknown how genes implicated by statistical colocalization analyses reflect true molecular mechanisms. It is therefore important to perform functional assays, as we have done for *NTN4,* to provide evidence that the variant affects gene expression and the gene plays a role in the disease etiology. Future work confirming the role of these genes or associated pathways in breast cancer development could ultimately lead to new avenues for breast cancer prevention or therapy.

SUPPLEMENTAL DATA

Supplemental Data include three figures and two tables.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Health and Medical Research Council of Australia (NHMRC; 1120563). S.L.E. is an NHMRC Senior Research Fellow (1135932). G.C.T is an NHMRC Senior Principal Research Fellow (1117073). J.D.F was supported by a Fellowship from the National Breast Cancer Foundation of Australia. N.A. was co-funded by a QIMR Berghofer International PhD Scholarship and a University of Queensland Research Training Scholarship. The results published here are in part based upon data generated by the TCGA Research Network. The breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec and grant PSR-SIIRI-701, The National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research UK

(C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). All studies and funders are listed in 3.

DECLARATION OF INTERESTS

The authors declare no conflict of interests.

WEB RESOURCES

Breast Cancer Association Consortium, http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/

GTEx Portal, https://gtexportal.org/home/

LDlink, https://ldlink.nci.nih.gov/

Bioconductor, https://www.bioconductor.org/

REFERENCES

1. Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., Consortium, G.T., and Dermitzakis, E.T. (2017). Estimating the causal tissues for complex traits and diseases. Nat Genet 49, 1676-1683.

2. Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science 344, 519-523.

3. Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemacon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94.

4. Li, Q., Seo, J.H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., Brown, M., Tyekucheva, S., and Freedman, M.L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152, 633-641.

5. Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M., Yuan, Y., Shi, J., Sun, Y., Han, M.R., et al. (2018). A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer

Susceptibility Loci Identified in Genome-wide Association Studies. Am J Hum Genet 102, 890-903.

6. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.O., Lu, Y., Cai, Q., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet 50, 968-978.

7. Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F., Aittomaki, K., Andrulis, I.L., Anton-Culver, H., Arason, A., Arndt, V., Aronson, K.J., Arun, B.K., et al. (2019). Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat Commun 10, 1741.

8. Geeleher, P., Nath, A., Wang, F., Zhang, Z., Barbeira, A.N., Fessler, J., Grossman, R.L., Seoighe, C., and Stephanie Huang, R. (2018). Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. Genome Biol 19, 130.

9. Plagnol, V., Smyth, D.J., Todd, J.A., and Clayton, D.G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics 10, 327-334.

10. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 10, e1004383.

11. Parker, M.M., Hao, Y., Guo, F., Pham, B., Chase, R., Platig, J., Cho, M.H., Hersh, C.P., Thannickal, V.J., Crapo, J., et al. (2019). Identification of an emphysema-associated genetic variant near TGFB2 with regulatory effects in lung fibroblasts. Elife 8.

12. Foley, C.N., Staley, J.R., Breen, P.G., Sun, B.B., Kirk, P.D.W., Burgess, S., and Howson, J.M.M. (2019). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. bioRxiv doi: https://doi.org/10.1101/592238.

13. Thom, C.S., and Voight, B.F. (2020). Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes. bioRxiv doi: https://doi.org/10.1101/787333.

14. Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S., Pooley, K.A., Dennis, J., Michailidou, K., Turman, C., et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet 52, 56-73.

15. Milne, R.L., Kuchenbaecker, K.B., Michailidou, K., Beesley, J., Kar, S., Lindstrom, S., Hui, S., Lemacon, A., Soucy, P., Dennis, J., et al. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 49, 1767-1778.

16. Barfield, R., Feng, H., Gusev, A., Wu, L., Zheng, W., Pasaniuc, B., and Kraft, P. (2018). Transcriptome-wide association studies accounting for colocalization using Egger regression. Genet Epidemiol 42, 418-433.

17. Feng, H., Gusev, A., Pasaniuc, B., Wu, L., Long, J., Abu-Full, Z., Aittomaki, K., Andrulis, I.L., Anton-Culver, H., Antoniou, A.C., et al. (2020). Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. Genet Epidemiol.

18. Beesley, J., Sivakumaran, H., Moradi Marjaneh, M., Lima, L.G., Hillman, K.M., Kaufmann, S., Tuano, N., Hussein, N., Ham, S., Mukhopadhyay, P., et al. (2020). Chromatin interactome mapping at 139 independent breast cancer risk signals. Genome Biol 21, 8.

19. Ghoussaini, M., Edwards, S.L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., Kar, S., Hillman, K.M., Kaufmann, S., Glubb, D.M., et al. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nat Commun 4, 4999.

20. Ward, L.D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 44, D877-881.

21. Grabe, N. (2002). AliBaba2: context specific identification of transcription factor binding sites. In Silico Biol 2, S1-15.

22. Wilson, B.D., Ii, M., Park, K.W., Suli, A., Sorensen, L.K., Larrieu-Lahargue, F., Urness, L.D., Suh, W., Asai, J., Kock, G.A., et al. (2006). Netrins promote developmental and therapeutic angiogenesis. Science 313, 640-644.

23. Lv, B., Song, C., Wu, L., Zhang, Q., Hou, D., Chen, P., Yu, S., Wang, Z., Chu, Y., Zhang, J., et al. (2015). Netrin-4 as a biomarker promotes cell proliferation and invasion in gastric cancer. Oncotarget 6, 9794-9806.

24. Jayachandran, A., Prithviraj, P., Lo, P.H., Walkiewicz, M., Anaka, M., Woods, B.L., Tan, B., Behren, A., Cebon, J., and McKeown, S.J. (2016). Identifying and targeting determinants of melanoma cellular invasion. Oncotarget 7, 41186-41202.

25. Eveno, C., Broqueres-You, D., Feron, J.G., Rampanou, A., Tijeras-Raballand, A., Ropert, S., Leconte, L., Levy, B.I., and Pocard, M. (2011). Netrin-4 delays colorectal cancer carcinomatosis by inhibiting tumor angiogenesis. Am J Pathol 178, 1861-1869.

26. Xu, X., Yan, Q., Wang, Y., and Dong, X. (2017). NTN4 is associated with breast cancer metastasis via regulation of EMT-related biomarkers. Oncol Rep 37, 449-457.

27. Esseghir, S., Reis-Filho, J.S., Kennedy, A., James, M., O'Hare, M.J., Jeffery, R., Poulsom, R., and Isacke, C.M. (2006). Identification of transmembrane proteins as potential prognostic markers and therapeutic targets in breast cancer by a screen for signal sequence encoding transcripts. J Pathol 210, 420-430.

28. Esseghir, S., Kennedy, A., Seedhar, P., Nerurkar, A., Poulsom, R., Reis-Filho, J.S., and Isacke, C.M. (2007). Identification of NTN4, TRA1, and STC2 as prognostic markers in breast cancer in a screen for signal sequence encoding proteins. Clin Cancer Res 13, 3164-3173.

29. Betts, J.A., Moradi Marjaneh, M., Al-Ejeh, F., Lim, Y.C., Shi, W., Sivakumaran, H., Tropee, R., Patch, A.M., Clark, M.B., Bartonicek, N., et al. (2017). Long Noncoding RNAs CUPID1 and CUPID2 Mediate Breast Cancer Risk at 11q13 by Modulating the Response to DNA Damage. Am J Hum Genet 101, 255-266.

30. Visser, M., Kayser, M., and Palstra, R.J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res 22, 446-455.

31. Lin, Y., Ma, W., and Benchimol, S. (2000). Pidd, a new death-domain-containing protein, is induced by p53 and promotes apoptosis. Nat Genet 26, 122-127.

32. Chung, C.Y., Sun, Z., Mullokandov, G., Bosch, A., Qadeer, Z.A., Cihan, E., Rapp, Z., Parsons, R., Aguirre-Ghiso, J.A., Farias, E.F., et al. (2016). Cbx8 Acts Non-canonically with Wdr5 to Promote Mammary Tumorigenesis. Cell Rep 16, 472-486.

33. Moradi Marjaneh, M., Beesley, J., O'Mara, T.A., Mukhopadhyay, P., Koufariotis, L.T., Kazakoff, S., Hussein, N., Fachal, L., Bartonicek, N., Hillman, K.M., et al. (2020). Non-coding RNAs underlie genetic predisposition to breast cancer. Genome Biol 21, 7.
34. Xu, S., Kong, D., Chen, Q., Ping, Y., and Pang, D. (2017). Oncogenic long noncoding RNA landscape in breast cancer. Mol Cancer 16, 129.
35. Liu, B., Gloudemans, M.J., Rao, A.S., Ingelsson, E., and Montgomery, S.B. (2019). Abundant associations with gene expression complicate GWAS follow-up. Nat Genet 51, 768-769.
36. Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31, 3555-3557.

FIGURE LEGENDS

Figure 1. Comparison of association signals from BCAC and GTEx v8 for breast tissue. LocusCompare plots[35] for eleven high-probability colocalized signals**.** Gene names and the relevant breast cancer phenotypes are shown in the plot headings. Points are coloured based on linkage disequilibrium (LD) bins relative to the candidate SNP prioritized by HyPrColoc (purple diamond labelled with rsID; red: ≥ 0.8, orange: 0.6 – 0.8, green: 0.4 – 0.6, light blue: 0.2 – 0.4, and dark blue: < 0.2). LD data from 1000 Genomes phase 3, version 5 were retrieved from the LDlink portal[36]. Strong candidate causal variants for breast cancer risk are annotated as small diamonds and weaker secondary risk variants as squares.

Figure 2. Regional association plots at the 12q22 breast cancer risk locus. Single variant associations with overall breast cancer risk (top panel) and with *NTN4* expression in normal breast tissue from GTEx v8 (lower panel). Variants are represented by points colored relative to linkage disequilibrium (LD) with the candidate variant detected by HyPrColoc (rs17356907; red: ≥ 0.8, orange: 0.6 – 0.8, green: 0.4 – 0.6, light blue: 0.2 – 0.4, and dark blue: < 0.2).

Figure 3. Breast cancer CCVs distally regulate *NTN4*. (**A**) WashU genome browser showing topologically associating domains (TADs) as horizontal gray bars above GENCODE annotated

coding genes (blue). The promoter capture Hi-C (PCHi-C) baits are depicted as black boxes. The putative regulatory element (PRE) containing the CCVs is shown as red colored vertical lines. The ATAC-seq tracks for B80T5 and MCF10A breast cells are shown as blue histograms. PCHi-C chromatin interactions are shown as black arcs. Red arcs depict chromatin looping between CCVs and the *NTN4* promoter region. (**B**) dCAS9-KRAB was targeted to the PRE using two different sgRNAs (sgPRE1 and sgPRE2) in Bre80-TERT1 breast cells. SgCON contains a non-targeting control guide RNA. Gene expression was measured by qPCR and normalized to *beta-glucuronidase* (*GUSB*) expression. Error bars, SEM (n=3). *P*-values were determined by one-way ANOVA followed by Dunnett's multiple comparisons test (**p < 0.01). (**C**) Luciferase reporter assays following transient transfection of MCF10A and Bre80-TERT1 breast cells. A PRE1 containing the protective (Prot.) or risk alleles of rs61938093 and a PRE2 containing the protective (Prot.) or risk alleles of rs1735907 were cloned into *NTN4*-promoter driven luciferase constructs. Error bars, SEM (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (**p < 0.01, ****p < 0.0001). (**D,E;** *left panels*) 3C interaction profiles between the *NTN4* promoter and the genomic region containing the PRE in MCF10A (**D**) and T47D (**E**) 3C libraries generated with HindIII. A physical map of the region interrogated by 3C is shown above; the blue shading represents the position of the PRE and the anchor point set at the *NTN4* promoter. Representative 3C profiles are shown. Error bars, SD (n=3). (**D,E;** *right panels*) Allele-specific qPCR to quantify the allelic ratio at CCV rs61938093. Error bars, SEM (n=3). *P*-values were determined using a Student's t-test (***p < 0.001). (**F**) EMSA for oligonucleotide duplexes containing CCVs rs61938093 or rs17356907 with either the risk allele (R) or protective allele (P) as indicated, assayed using Bre80-TERT1 nuclear extracts. Competitor oligonucleotides are listed above each panel and were used at 100-fold molar excess: (-) no competitor; (Neg) a non-specific competitor; (Self) an identical oligonucleotide with no biotin label. Red arrowheads indicate band mobility differences between alleles.

Figure 4. *NTN4* depletion promotes breast cell proliferation and tumor formation. (**A**) Boxplot showing *NTN4* expression in normal breast and paired tumor tissue samples from TCGA. Boxplots indicate median (centre line), interquartile range (box limits) and range (whiskers). P-value was determined using a two-tailed t-test. (**B**) Boxplot showing *NTN4* expression in breast tumors from

TCGA stratified by PAM50 molecular subtypes (n=841). Boxplots indicate median (centre line), interquartile range (box limits) and range (whiskers). (**C**) Proliferation of MCF7 cells transfected with a non-targeting control (siCON) or *NTN4* (siNTN4) ON-TARGETplus siRNAs. Cells were grown in 24-well plates and confluency of the wells was measured by the IncuCyte live-cell imaging system. Results represent relative cell growth rates. Error bars, SD (n=2). P-value was determined by Student's t-test comparing confluency at the last time point measured (***p < 0.001). (**D**) MCF7 cells were transfected with the siCON or siNTN4 and grown over 7 days in ultra low-attachment conditions. Cell growth was assessed using the CellTiter-Glo luminescent cell viability assay. Graph shows fold change in luminescence of siNTN4 treated cells relative to siCON treated cells. Error bar, SEM (n=3). P-value was determined by Student's t-test test (**p < 0.01). (**E**) MCF7-control (PgCON) or MCF7-dCas9-KRAB NTN4 repressed cells (SgNTN4-P1/P2) were orthotopically injected into the mammary fat pads of nude mice. Tumor growth curves for each group are shown. Values are shown as average tumor volumes at each time point. Error bars, SEM (n=6-7 mice per group). (**F**) Tumors of individual mice were dissected at day 38 post-injection. The scale bars represent 1 cm. (**G**) Plot of the individual weights of tumors with mean and SEM shown by cross-bar and error bars. (**E, G**) Mann-Whitney *U* test was used to compare differences between groups (*p < 0.05, **p < 0.01, ***p < 0.001).

Table 1. Candidate breast cancer risk genes identified by eQTL colocalization analyses.

Ensembl ID	Gene symbol	Breast cancer risk association	Genomic coordinates (hg19)	Posterior probability	Candidate SNP	Posterior explained by SNP	GTEx eQTL *P* value	Breast cancer risk *P* value
ENSG00000074527.11	*NTN4*	Overall risk	chr12:95527759-96527759	0.9466	rs17356907	0.97	8.01E-09	1.02E-39
ENSG00000141570.10	*CBX8*	Overall risk	chr17:77281387-78281725	0.9178	rs9905914	0.49	7.92E-23	4.00E-09
ENSG00000198945.7	*L3MBTL3*	Overall risk	chr6:129849119-130849119	0.7998	rs7740107	1.00	5.88E-40	2.90E-11
ENSG00000183654.8	*MARCH11*	Overall risk	chr5:15687358-16687528	0.8369	rs1013018	0.16	3.05E-09	1.65E-11
ENSG00000177595.17	*PIDD1*	Overall risk	chr11:303017-1303017	0.9695	rs6597981	0.22	6.53E-27	1.35E-12
ENSG00000166965.12	*RCCD1*	Overall risk	chr15:91009215-92009215	0.9633	rs113343095	0.60	2.44E-24	3.37E-15
ENSG00000130511.15	*SSBP4*	Overall risk	chr19:18050434-19071141	0.7800	rs7258465	0.09	7.87E-08	2.79E-28
ENSG00000172748.13	*ZNF596*	ER- risk	chr8:0-670692	0.9059	rs35346588	0.79	2.17E-08	1.39E-08
ENSG00000258725.1	*PRC1-AS1**	Overall risk	chr15:91009215-92009215	0.9302	rs2290202	0.22	5.89E-10	1.87E-15
ENSG00000251141.5	*RP11-53O19.1**	Overall risk	chr5:44013304-45206498	0.9347	rs10941679	1.00	4.41E-07	5.61E-73
ENSG00000272812.1	*RP5-855D21.3**	ER- risk	chr8:0-670692	0.9769	rs3008281	0.81	6.11E-08	6.23E-09
ENSG00000152348.15	*ATG10*#	Overall risk	chr5:80928261-82038046	0.7904	rs144580806	0.36	2.56E-40	8.07E-12
ENSG00000015133.18	*CCDC88C*#	Overall risk	chr14:91341069-92368623	0.9465	rs8018155	0.50	9.15E-11	4.03E-12
ENSG00000163644.14	*PPM1K*#	Overall risk	chr4:88743818-89743818	0.9935	rs10022462	0.58	1.60E-08	1.55E-09
ENSG00000233967.6	*RP11-250B2.3**#	Overall risk	chr6:80594287-81594287	0.8473	rs9448940	0.22	4.65E-11	9.85E-09
ENSG00000260645.1	*RP11-250B2.5**#	Overall risk	chr6:80594287-81594287	0.8227	rs1436864	0.08	1.97E-08	3.89E-09
ENSG00000219392.1	*RP1-265C24.5**#	Overall risk	chr6:26180698-27180698	0.9901	rs35768595	0.38	5.95E-10	3.16E-09

* Noncoding RNAs

Weaker secondary signals[14]