ARTICLE

# Two distinct mechanisms underlie estrogen-receptor-negative breast cancer susceptibility at the 2p23.2 locus

Gustavo Mendoza-Fandiño[1,2], Paulo Cilas M. LyraJr[3], Thales C. Nepomuceno[4], Carly M. Harro[1,5,6], Nicholas T. Woods[7], Xueli Li[1], Leticia B. Rangel[8], Marcelo A. Carvalho[4,9], Fergus J. Couch[10] and Alvaro N. A. Monteiro[1]✉



Genome wide-association studies (GWAS) have established over 400 breast cancer risk loci defined by common single nucleotide polymorphisms (SNPs), including several associated with estrogen-receptor (ER)-negative disease. Most of these loci have not been studied systematically and the mechanistic underpinnings of risk are largely unknown. Here we explored the landscape of genomic features at an ER-negative breast cancer susceptibility locus at chromosome 2p23.2 and assessed the functionality of 81 SNPs with strong evidence of association from previous fine mapping. Five candidate regulatory regions containing risk-associated SNPs were identified. Regulatory Region 1 in the first intron of *WDR43* contains SNP rs4407214, which showed allele-specific interaction with the transcription factor USF1 in in vitro assays. CRISPR-mediated disruption of Regulatory Region 1 led to expression changes in the neighboring *PLB1* gene, suggesting that the region acts as a distal enhancer. Regulatory Regions 2, 4, and 5 did not provide sufficient evidence for functionality in *in silico* and experimental analyses. Two SNPs (rs11680458 and rs1131880) in Regulatory Region 3, mapping to the seed region for miRNA-recognition sites in the 3′ untranslated region of *WDR43*, showed allele-specific effects of ectopic expression of miR-376 on *WDR43* expression levels. Taken together, our data suggest that risk of ER-negative breast cancer associated with the 2p23.2 locus is likely driven by a combinatorial effect on the regulation of *WDR43* and *PLB1*.

*European Journal of Human Genetics* (2022) 30:465–473; https://doi.org/10.1038/s41431-021-01005-6

## INTRODUCTION

Breast cancer (BC) is the most commonly diagnosed cancer among women in the world [1]. Estrogen receptor (ER)-negative BCs are characterized by the lack of estrogen receptor expression. They account for 20–30% of all BC, and are more common in premenopausal women and women of African ancestry [2]. Genetic factors, such as pathogenic variants in *BRCA1*, have been shown to contribute to ER-negative tumors [3]. Triple negative BCs (TNBC) are a subset of ER-negative tumors and lack expression of ER, the progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). TNBC is particularly challenging to treat since tumors are not responsive to routine endocrine therapy or HER2-targeted therapies such as Trastuzumab and Lapatinib [4].

Currently, over 700 SNPs ($P \leq 5 \times 10^{-8}$) defining more than 400 loci associated with risk for BC have been identified through Genome-wide association studies (GWAS) (reviewed in [5]). Over 60 SNPs have been shown to be associated with ER-negative disease [6–14]. Although there has been progress in assessing the biological mechanisms at BC risk loci, those conferring risk to ER-negative disease risk have been underexplored [15–17].

In 2016, Couch et al. reported four loci associated with susceptibility to ER-negative BC ($P < 5 \times 0^{-8}$) with three SNPs (rs67073037, rs6734079, and rs4577244) at the 2p23.2 locus representing novel associations [18]. Interestingly, no association with ER-positive BC was detected for 2p23.2, suggesting that the association is specific to ER-negative disease [18]. The SNP most significantly associated at the locus was rs67073037 ($P = 4.76 \times 10^{-9}$) [18]. Preliminary analysis, including eQTL, chromatin marks, luciferase-reporter assays, and electrophoretic mobility shift assays suggested WD repeat domain 43 (*WDR43*) and tRNA methyltransferase 61B (*TRMT61B*) as possible target genes [18]. In 2020, fine mapping of the region conducted by Fachal et al. [19]. identified 81 SNPs with strong evidence of association with extensive overlap with the previously identified set proposed by Couch et al. [18]. None are located in coding regions, but 14 SNPs overlap with transcriptional-regulatory elements in BC cell lines [18]. Here we assess the functional contributions of the set of 81 risk-associated SNPs to identify SNPs, regulatory elements, and target genes likely to be the underlying risk at the locus.

## MATERIALS AND METHODS

### Bioinformatics-analysis pipeline and datasets

SNPs significantly associated with ER-negative BC at the locus after fine mapping (conditional $P < 10^{-6}$) were retrieved from [19]. Then, SNPs were functionally annotated by examination of the overlap with

[1]Cancer Epidemiology Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. [2]Corporación Universitaria Remington, Medellin, Colombia. [3]Faculdade de Ensino e Meio Ambiente, Ariquemes, RO, Brazil. [4]Instituto Nacional de Câncer, Programa de Pesquisa Clínica, Rio de Janeiro 20231-050, Brazil. [5]Department of Cell Biology, Microbiology, and Molecular Biology, College of Arts and Sciences, University of South Florida, Tampa, FL, USA. [6]Cancer Biology PhD Program, University of South Florida Tampa, Tampa, FL 33612, USA. [7]Eppley Institute for Research in Cancer, University of Nebraska Medical Center, Omaha, NE 68198, USA. [8]Biotechnology/RENORBIO Program, Federal University of Espírito Santo, Vitória, ES, Brazil. [9]Instituto Federal do Rio de Janeiro - IFRJ, Rio de Janeiro 20270-021, Brazil. [10]Mayo Clinic, Rochester, MN 55905, USA. ✉email: alvaro.monteiro@moffitt.org

Received: 27 June 2021 Revised: 24 October 2021 Accepted: 8 November 2021
Published online: 22 November 2021

histone-modification data for normal human mammary epithelial cell line (HMEC) obtained by publicly available (ENCODE/Broad Institute) chromatin immunoprecipitation (ChIP)-seq data visualized via the UCSC Genome Browser https://genome.ucsc.edu/. These features are indicative of putative transcriptional regulatory regions [20, 21] and therefore SNPs that are located in these regions were retained for further analysis. SNPs that did not overlap with these features were not analyzed further for transcriptional-regulatory activity. Next, to evaluate the potential of each SNP to contribute to the regulatory activity in the region, we used RegulomeDB ranks retrieved from http://www.regulomedb.org/ [22]. RegulomeDB guides interpretation of regulatory variants integrating high-throughput, experimental data sets from ENCODE and other sources, computational predictions, and manual annotations. Variants with lower ranks (predicted high regulatory activity) were prioritized for analysis. Because genes likely to be regulated by the candidate regions are expected to be contained in the same topologically associating domains (TAD), we used the EpiTAD tool [23] and the Yue Lab 3 C browser (http://promoter.bx.psu.edu/hi-c/view.php) with the Lierberman-raw 5 kb resolution data from HMEC (GRCh37/hg19) to identify putative target genes.

We used sequences retrieved from the human genome browser for the 41 bp surrounding rs4407214 as the input for JASPAR (http://jaspar.genereg.net/) [24] for both alleles to retrieve the predicted human-transcription factors binding to the site with a relative profile-score threshold of 70%. MirSNP (http://cmbi.bjmu.edu.cn/mirsnp) was used to identify whether the SNPs overlapped with miRNA "seed regions." ER status and mRNA expression data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) were analyzed using cBioPortal [25, 26]. Samples were divided into two groups based on mRNA levels: >2 standard deviations (SD) above the mean = 'high'; <2 SD below the mean = 'low'. For TNBC and non-TNBC cell lines, we used RNA-seq data from the Cancer Cell Line Encyclopedia (CCLE) [27].

## Cell lines

We used immortalized cell lines representative of basal ER-negative breast tissue: normal human mammary epithelial cell (MCF10A) (ATCC; CRL-10317) and the breast carcinoma cell line CAL-51 (Dynamics; CSC-C0382). All experiments were conducted in MCF10A cells and for EMSA and enhancer scanning experiments, CAL-51 was also used. MCF10A cells were grown in DMEM/F12 (Invitrogen) with 5% donor horse serum (Invitrogen), 20 ng/ml EGF (Invitrogen), 10 µg/ml insulin (Sigma), 0.5 µg/ml hydrocortisone, 100 ng/ml cholera toxin (Sigma), and 1X Penicillin/Streptomycin (Invitrogen). CAL-51 cells were grown in DMEM (4.5 g/L glucose) with 10% fetal bovine serum. Cells were grown at 37 °C in a 5% $CO_2$-humidified incubator.

## Electrophoretic mobility shift assay (EMSA)

To determine allele-specific transcription-factor binding, log-growing MCF10A or CAL-51 cells were used to prepare nuclear extracts for EMSA as previously described [28]. EMSA probes covered each SNP ± 20 base pairs. Probe sequences are shown in Supplementary Table 1. Experiments were performed in at least two replicates using the same lysate (technical replicates) and two independent experiments (biological replicates).

To determine allele-specific transcription-factor binding to rs4407214, EMSA probes were designed to cover 41 bases around rs4407214 (20 bases on either side covering the region chr2:29,118,239-29,118,279), for both the reference and the alternative alleles. Probe for the candidate transcription factor USF1 was designed as reported previously [29, 30]. The USF probe was synthesized to contain the human Cathepsin D proximal promoter position +124/+104 for EMSA experiments [29] (Supplementary Table 1). The sequence is homologous to the adenovirus major late promoter element, which contains a half-estrogen response element (5′-GTACC-3′) and E box (5′-CACGTG-3′) as USF1-binding site [31] (Supplementary Table 1). Probes were suspended in TE buffer and annealed at a concentration of 10 µM, and heated at 99 °C for 5 min. Probes were allowed to cool until RT and stored in aliquots at −80 °C. DNA probes were labeled with ATP [γ-$^{32}$P] (Perkin Elmer) using T4 polynucleotide kinase (NEB) and cleaned using the QIAquick Nucleotide Removal Kit (Qiagen). Labeled probes were then incubated with nuclear extracts (10 µg), LightShift Poly (dI-dC) (Thermo), binding buffer (10 mM Tris, 50 mM KCl, and 1 mM DTT, pH 7.4) and unlabeled competing probes. The reactions were separated by electrophoresis on 6% nondenaturing polyacrylamide gel, 84 V overnight. Gels were vacuum-dried at 60 °C for 1 h, and high-performance films were exposed for 4–24 h. EMSA experiments were performed in at least two technical replicates and two independent experiments.

## RNA-guided site-specific DNA cleavage (CRISPR)

To cleave the genomic region around the SNP rs4407214 (RR1), five guide sequences were identified (20-nucleotide DNA sequences followed by a 5′-NGG or 5′-NAG PAM) (Supplementary Table 1). The *Streptococcus pyogenes* Cas9 targets 20 nucleotide DNA sequences followed by a 5′-NGG or 5′-NAG PAM sequence [32]. These sequences facilitate RNA-guided site-specific DNA cleavage by CRISPR (clustered regularly interspaced short palindromic repeats)/Cas system. The five guide sequences were cloned into the gRNA Cloning Vector (Addgene ID # 41824) according to the Prashant Mali protocol option B (http://www.addgene.org/static/cms/files/hCRISPR_gRNA_Synthesis.pdf) and were confirmed by Sanger sequencing using the M13-Rev primer (Supplementary Table 3).

For RNA-guided site-specific DNA cleavage, 70–80% confluent MCF10A cells, which are diploid, were used for transfection. The five gRNA-vectors were cotransfected with hCas9 (Addgene ID# 41815), and pBABE-puro (Addgene ID# 1764) using Lipofectamine 2000 (Invitrogen) at a 3:1 ratio of Lipofectamine 2000 volume (µL) to DNA (µg). After 24 h, cell cultures were trypsinized, plated at different densities (1:1, 1:9, and 1:20), and selected with 1 µg/mL puromycin (Invitrogen). Clones were isolated with plastic cloning rings and expanded. Excision of the rs4407214 region was confirmed by sequencing (Supplementary Table 3). Eight clones were isolated and characterized. Clones CCΔ1, CCΔ2, CCΔ3, and CCΔ4 displayed a deletion at the targeted locus. Clones CC5, CC6, and CC7 had no deletion at the targeted locus detected in the sequenced amplicon. It is possible that they may harbor genomic changes outside of the examined amplicon. Clone CC8 had a deletion on the 5′ region of the amplicon that did not remove the targeted locus. Clones CC5–8 were chosen to serve as negative controls because they represent cells that were transduced, selected, cloned, and expanded (in parallel with the clones with deletion).

## Quantitative RT-PCR

For MCF10A CRISPR clones CCΔ1–4 and CC5–8, RNA was extracted using a Qiagen RNeasy Mini-Prep Kit. Reverse transcriptase reaction was performed with 1 µg of total RNA using the Qiagen QuantiTect Reverse Transcription Kit with genomic DNA removal. Quantitative PCR reaction was performed using TaqMan gene expression assays for *RBKS* (Hs00223231_m1), *BRE* (Hs01046283_m1), *PLB1* (Hs00290809_m1), *FOSL2* (Hs01050117_m1), *PPP1CB* (Hs01027793_m1), *SPDYA* (Hs00736925_m1), *TRMT61B* (Hs00372418_m1), *WDR43* (Hs01064086_m1), *FAM179A* (Hs00416668_m1), *C2orf71* (Hs01079277_s1), *CLIP4* (Hs00372786_m1), and *ALK* (Hs01058318_m1). Eukaryotic 18 S assay was used as an internal control (ABI catalog # 4319413E), with TaqMan Universal PCR Master Mix (Life Tech), on 7900HT Fast Real-Time PCR System (Applied Biosystems). RT-PCR was performed with three technical replicates and two biological replicates. The results were analyzed by comparing ΔCT values using Student's t-test. No clone displayed detectable expression of *C2orf71*.

## Luciferase reporter assays (enhancer scanning)

Genomic tiles (~2 Mb) A–E spanning regions RR2–RR5 containing risk-associated SNPs were generated as previously described [33]. Briefly, forward and reverse primers contained attB1 and attB2 sequences, respectively (Supplementary Table 1). Tiles were cloned in forward and reverse orientations upstream of the SV40 promoter in pGL3-Pro-attb vector to test for enhancer regions. Each clone containing a tile was cotransfected in eight replicates using LipoFectamine 2000 (Life Technologies) into MCF10A or CAL51 cells with pRL-CMV (Promega), an internal control expressing *Renilla* luciferase, per well of 96-well plates. Luciferase was measured 24 h after transfection by Dual Glo Luciferase Assay (Promega) and two independent experiments were conducted.

To generate the luciferase reporter for the *WDR43* 3′UTR, the region was synthesized with restriction sites for *Spe*I, and *Hind*III (NEB) added to the end of the fragment (Supplementary Table 1). The fragments were synthesized with the reference [GG] or alternative/effect [TC] haplotypes for the rs11680458 and rs1131880, respectively. The fragment was cloned in using *Spe*I and *Hind*III (NEB) into the pMIR REPORT miRNA expression reporter vector (Ambion) according to the manufacturer's protocol. Positive clones were selected by restriction digestion using *Bam*HI (NEB) and confirmed by Sanger sequencing. For the luciferase reporter assay, MCF10A was plated at 10,000 cells per well in 96–well plates 24 h prior to

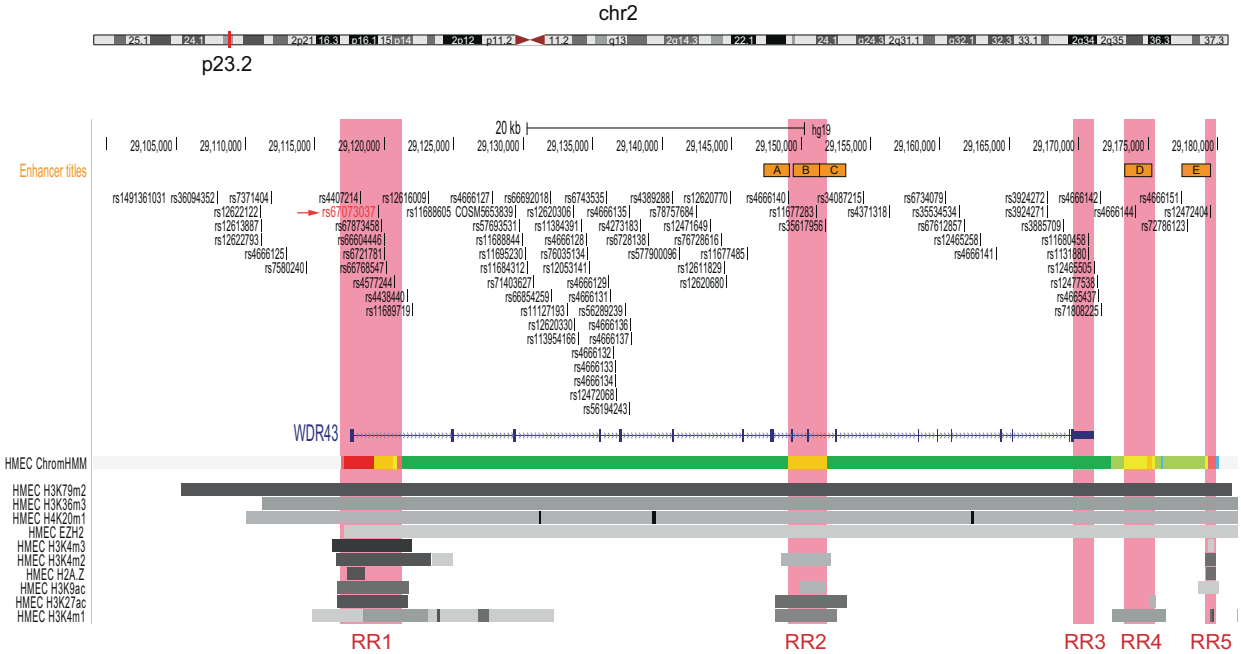


**Fig. 1 Chromatin landscape at the 2p23.2 locus.** Human Genome browser (hg19). SNPs significantly associated with ER-negative breast cancer at the 2p23.2 locus after fine-mapping (conditional $P < 10^{-6}$) previously reported by Fachal et al. [19]. The 2p23.2 locus region contains risk-associated SNPs spanning the *WDR43* gene at the 2p23.2 locus. Several risk-associated SNPs mapped to ENCODE features in human mammary epithelial cell line (HMEC). The index (most significantly associated in the meta-analysis) SNP, rs67073037, is shown in red. Tiles for luciferase-reporter assays to identify enhancers are shown as orange boxes A–E.

transfection. Allele-specific *WDR43* 3′UTR pMIR REPORT construct and pMIR-REPORT empty vector (EV) were cotransfected with pRL-TK (*Renilla* luciferase driven by TK promoter) as an internal control, mimic-hsa-miR-141-3p (Sigma), mimic-hsa-miR-376c (Sigma), mimic-hsa-miR-548ar-3p (Sigma), mimic-hsa-miR-200a-3p (Sigma), and mimic-has-miR-155-5p as a negative control that does not target the 3′ UTR *WDR43* region. Transfections were performed using lipofectamine 2000 (Invitrogen) at a ratio of 3:1 lipofectamine 2000 volume (μL) to DNA (μg). After 24 h, luciferase activity was measured using the Dual Glo Luciferase Assay Kit from Promega. Luciferase values were normalized to the internal control from *Renilla* luciferase values and compared with the mimic miR-155-5p-negative nontargeting control. Luciferase experiments were performed with eight technical replicates and two independent experiments. The results were analyzed using Student's *t*-test.

## RESULTS

### Characterizing the chromatin landscape at the 2p23.2 risk locus

We mapped the 81 SNPs that considered the credible set of SNPs (those most likely to contribute to the phenotype, conditional $P < 10^{-6}$) at the 2p23.2 locus by Fachal et al. [19] to the human genome and determined their overlap with genes obtained from ENCODE RefSeq to assess their functional relevance. None of the 81 risk-associated SNPs (Supplementary Table 2) were located in coding regions, consistent with the hypothesis that they might impact risk through allelic-specific differences in gene regulation rather than altering protein structure or function [34]. These risk-associated SNPs, spanning a region of approximately 72 kb, either overlapped or were proximal to *WDR43*, a protein-coding gene related to rRNA processing and ribosomal biogenesis (Fig. 1) [18].

We identified five noncoding regions, containing 14 risk-associated SNPs, hypothesized to constitute transcriptional regulatory regions (RR) [18] (Fig. 1). Candidate transcriptional regulatory regions were defined based on tissue-specific features and chromatin-state segmentation from ENCODE ChIP-seq data for H3K4m1, H3K4m2, H3K4m3, H3K9ac, H3K27ac, H3K36m3, H3K79m2, H4K20m1, EZH2, and H2AZ in normal human mammary epithelial cell line (HMEC). Previously, using two breast cancer datasets, we only found significant eQTL associations to *TRMT61B* association [18]. We searched for eQTL and sQTL associations in GTEx (V8) using the breast (mammary tissue, $n = 469$) and confirmed that only eQTL associations to *TRMT61B* are found at the locus (data not shown). Risk-associated SNPs in RR1 are located within the first intron of *WDR43*, and SNPs in RR2 overlap with introns 8–10 of *WDR43*. SNPs in RR3 map to a miRNA "seed region" at the 3′ untranslated region (UTR) of *WDR43* (Fig. 1). SNPs in RR4 and RR5 are in the intergenic region between *WDR43* and *FAM179A* (Fig. 1).

### Regulatory Region 1: transcription factors interacting with rs4407214

We previously reported rs4407214 allele-specific binding of nuclear proteins from MCF10A and CAL-51 cells using electrophoretic mobility shift assay (EMSA) [18]. Evidence from luciferase reporter assays, which showed allele-specific activity, and ENCODE features, is consistent with this region harboring a regulatory region that could act as an enhancer element [18]. To identify transcription factors binding to rs4407214 in an allele-specific manner, we combined bioinformatics prediction of transcription-factor binding with ChIP-Seq data available in the ENCODE project obtained from the University of California, Santa Cruz (UCSC) Human Genome browser repository site.

We conducted a transcription-factor search analysis using the 41-base sequence surrounding SNP rs4407214 using the JASPAR web tool [24]. JASPAR identified 1852 predicted binding sites for 266 unique human transcription factors binding to the region containing rs4407214 [G or T alleles] (Fig. 2A). The rs4407214 SNP was part of the binding site for 188 transcription factors (Fig. 2A) (Supplementary Table 3). Next, we used ChIP-Seq data from the ENCODE Factorbook [35] repository to identify 57 unique transcription factors that have been experimentally shown to bind to the region containing rs4407214 (Fig. 2A) (Supplementary Table 4). Of those, 13 had a ration of supporting/total experiments > 0.8 and only five had more than one experiment (Fig. 2A). Using

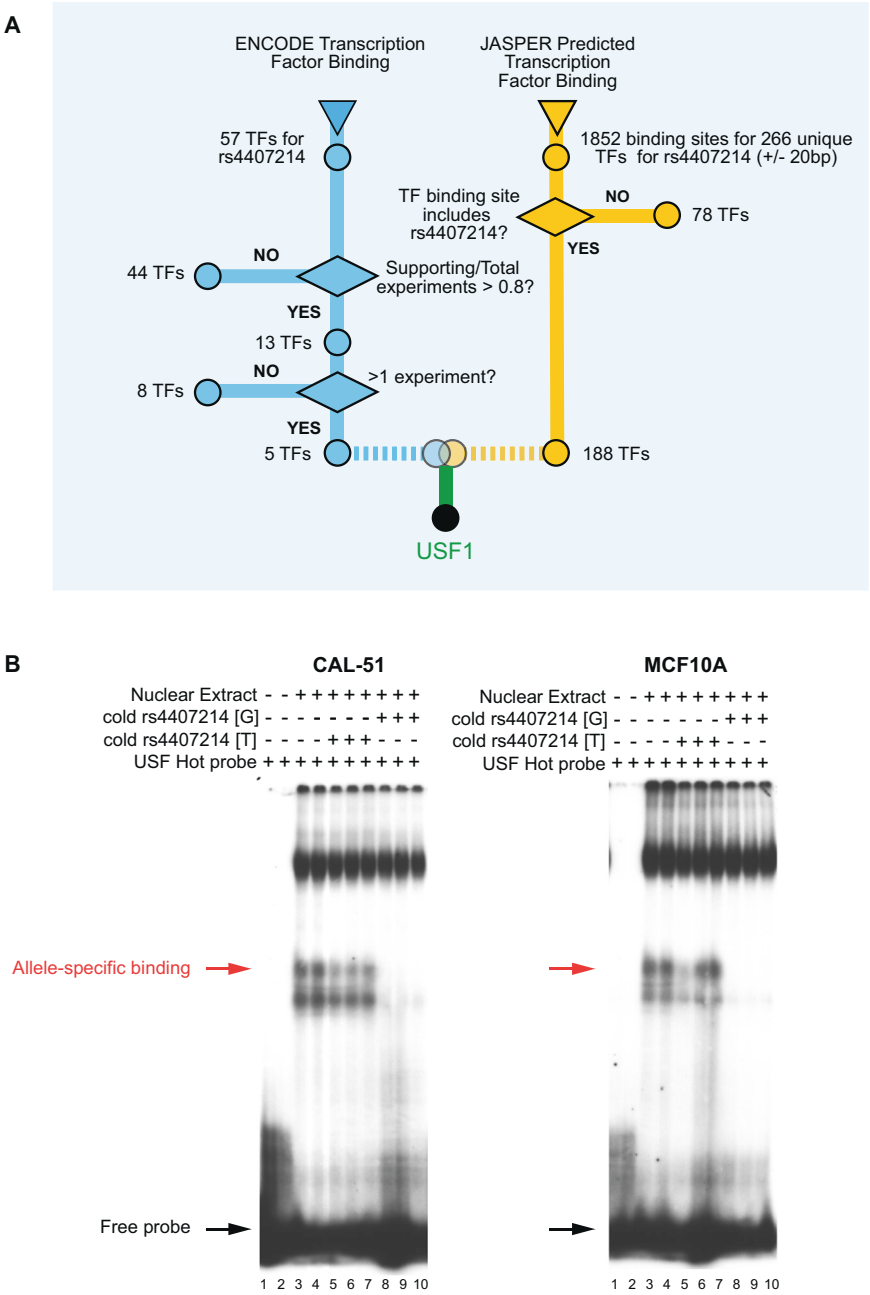


**Fig. 2 Allele-specific transcription factor binding. A** 'Subway chart' showing the *in silico* analysis pipeline for determining candidate transcription factors binding to rs4407214 using predictions from JASPAR database and ENCODE Factorbook data. **B**, Electrophoretic mobility shift assays using nuclear extracts from CAL-51 (left panel) and MCF10A cells (right panel), respectively, for competition between unlabeled [T] or [G] alleles with a labeled probe for USF1 and USF2. Red arrows indicate allele specific binding. Control lanes (hot probe only and nuclear extract plus hot probe) are loaded as two technical replicates side-by-side. Cold competitor lanes are loaded as three technical replicates side-by-side.

these stringent cutoffs, we identified USF1 as present in both sets **(**Fig. 2A).

Next, we determined whether USF1 displayed allele-specific binding activity using electrophoretic mobility shift assays (EMSA). Nuclear protein extracts from CAL-51 and MCF10A were incubated with DNA probes containing consensus sequence for USF 1 (Supplementary Table 1). To evaluate the binding-competition capability of the SNP rs4407214 [T] or [G] alleles, probes for each allele were combined with nuclear extracts in the binding reaction. The SNP rs4407214 [T] allele did not compete for the binding with the USF factor (Fig. 2B, lanes 5–7) unlike the [G] allele (Fig. 2B, lanes 8–10) in both MCF10A and CAL-51 cell lines. These results suggest that USF1 modulates the activity of the enhancer/promoter element that includes rs4407214 and support the notion that the SNP rs4407214 contributes mechanistically to ER-negative BC risk.

### Deletion of putative regulatory region 1 containing rs4407214 via CRISPR–Cas9

To evaluate the role of RR1 as an enhancer on genes at the 2p23.2 locus, we removed the genomic region around SNP rs4407214 via the CRISPR/Cas-9 system. Our screening revealed four MCF10A clones (CCΔ1–4) in which the region around rs4407214 was deleted, and four clones (CC5–8) that retained the region and served as negative controls (Fig. 3A).

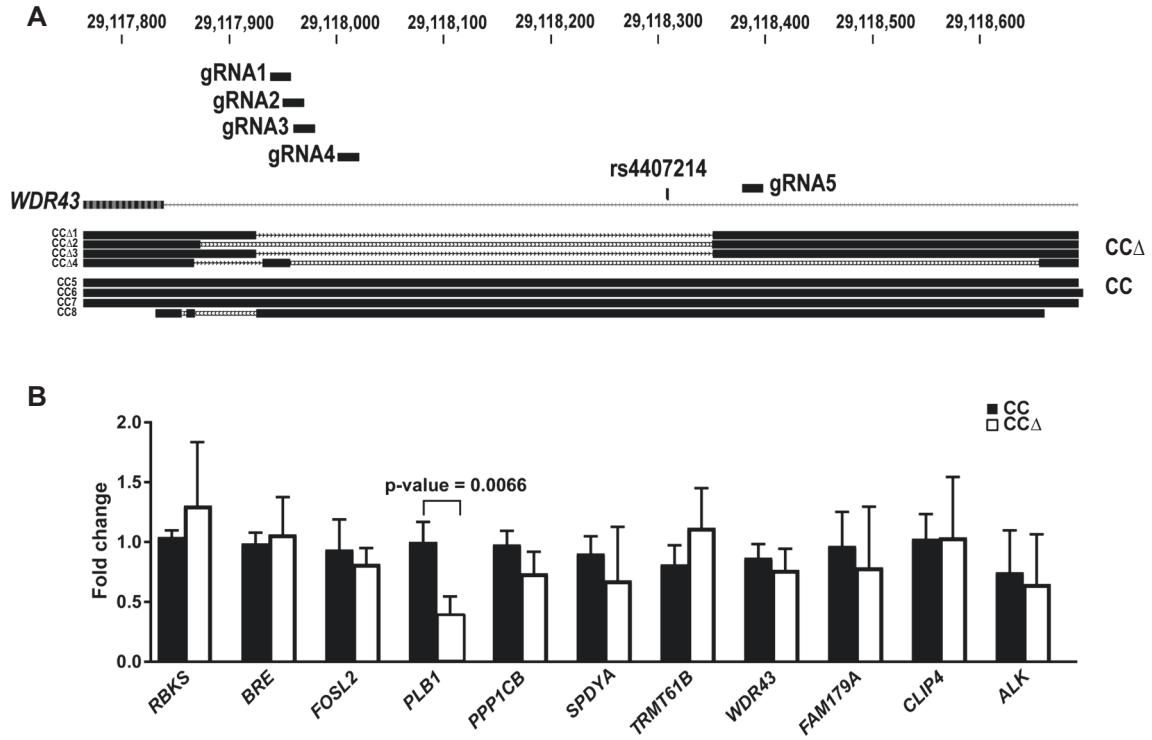


**Fig. 3 Deletion of regulatory region 1. A** The region around rs4407214 was removed using CRISPR/Cas-9 system with the indicated guide RNAs (gRNAs) in the first intron of the *WDR43* gene. Clones were screened using Sanger sequencing and aligned to the region to ensure the removal of rs4407214. **B** RT-qPCR using TaqMan assays for the expression of 11 genes spanning 2 Mb. Values for CRISPR clones containing the removed region (CCΔ 1-4) were compared with the negative controls, or clones with intact regions (CC 5-8). Significant change ($p \leq 0.05$) in gene expression between the CCΔ and CC clones is indicated.

Next, we compared expression levels of genes in a 2-Mb region at the locus (Chr2: 28,000,000 to 30,000,000; *RBKS, BRE, FOSL2, PLB1, PPP1CB, SPDYA, TRMT61B, WDR43, FAM179A, C2ORF71, CLIP4*, and *ALK*) in clones CCΔ1–4 compared with our CC5–8 with 18 S acting as a control-housekeeping gene. No clone showed detectable expression of *C2ORF71*. Quantitative TaqMan assays revealed a significant change in expression of phospholipase B1 (*PLB1*) ($P = 6.6 \times 10^{-3}$), which was consistently downregulated in all CCΔ clones (Fig. 3B). This indicates that RR1 acts as an enhancer to *PLB1*, which is located 400 kb telomeric to *WDR43*. We did not detect any significant change in expression of *WDR43*. We also did not observe any significant changes in *TRMT61B*, a methyltransferase for N(1)-methyl adenine on tRNAs, expression between CC clones and CCΔ clones as we anticipated based on the previous eQTL analysis [18]. These data indicate that the regulatory region around SNP rs4407214 targets the expression of *PLB1* in a mammary gland epithelial cell line.

### Contribution of the Regulatory Regions 2–5 to the regulation at the 2p23.2 locus

Next, we examined the functional roles of the SNPs in RR2-RR5. We identified RR2, RR4, and RR5 as regions with evidence of enhancer activity using ChromHMM [36], a multivariate hidden Markov model that models chromatin marks to identify different chromatin states (Fig. 1). We also used RegulomeDB which assigns functions based on a collection of regulatory information on a set of variants, with lower ranks associating with higher regulatory function. Most candidate SNPs in RR2–RR5 had weak evidence for functionality defined by high (≥4) RegulomeDB ranks, except rs11677283 and rs4666144 (Table 1).

Five genomic tiles (Tiles A–E) covering these regions and containing risk-associated SNPS were assessed for enhancer activity in both forward and reverse orientation. Only tile D (RR4, rs4666144) displayed activity in at least one orientation in both cell lines (Fig. 4A) (Table 1). Tile D displayed allele-specific activity depending on the rs4666144 allele [C] or [T] but only in one cell line (Fig. 4B). We then tested whether any of the five risk-associated SNPs in these regions displayed allelic-specific binding of EMSA. SNPs rs11672283 and rs4666144 showed evidence of allele-specific binding in CAL51 nuclear extracts (Supplementary Fig. 1). Overall, evidence for allele-specific enhancer activity was present for rs4666144, albeit weakly. Next, we investigated other potential mechanisms of regulation within the locus.

### Regulatory region 3: allele-specific miRNA-mediated regulation

SNPs rs11680458, rs1131880, and rs12465505 mapped to the 3′ untranslated region (UTR) of *WDR43* (Fig. 1). Thus, we explored it as a possible mechanism of regulation, which we refer to as Regulatory Region 3 (RR3). SNPs rs11680458 and rs1131880 were predicted by MirSNP database [37] to occur within miRNA "seed regions" (Fig. 4C). The rs11680458 reference allele [G] was predicted to be recognized by the miR-141 and 200, while the effect allele [T] was predicted to be recognized by miR-548. The rs1131880 effect allele [C] was predicted to be recognized by the miR-376 and miR-577 (Supplementary Table 5). Our working hypothesis was that one or both SNPs might alter miRNA-binding sites affecting the expression of *WDR43*.

To assess the effects of rs11680458 and rs113188 alleles, we compared luciferase expression controlled by allele-specific sequences of the *WDR43* 3′-UTR. We designed the insert with both alleles to match the most frequent haplotype present in CEU populations, retrieved by the LDHap tool [38]. The reference haplotype for both rs11680458 and rs1131880 [GG] was present in the same construct, while the effect haplotype [TC] was present on a different construct (Supplementary Table 1). We cotransfected MCF10A cells with expression vectors for miRNA mimics

**Table 1.** Candidate functional SNPs in regulatory regions RR2-RR5.

| rsID | Position (Chr2) | A1 | A2 | P value[a] | RegdB[b] | LD (r2)[c] | Elements[d] | Region[e] (tile) | EMSA[f] | ES[g] | mir[h] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs4666140 | 29149051 | T | C | $8.44\times10^{-9}$ | 6 | 0.97 | E | R2 (tile A) | No | No | – |
| rs11677283 | 29151035 | T | C | $7.74\times10^{-9}$ | 2b | 1 | E | R2 (tile B) | Yes | No | – |
| rs35617956 | 29151714 | A | AT | $1.93\times10^{-8}$ | 5 | 1 | E | R2 (tile C) | No | No | – |
| rs11680458 | 29170623 | G | T | $1.11\times10^{-8}$ | 7 | 1 | – | R3 | – | – | No |
| rs1131880 | 29170676 | G | C | $1.43\times10^{-8}$ | 7 | 1 | – | R3 | – | – | Yes |
| rs4666144 | 29174105 | C | T | $1.49\times10^{-8}$ | 3a | 1 | E | R4 (tile D) | Yes | Yes | – |
| rs12472404 | 29179452 | G | C | $1.46\times10^{-8}$ | 4 | 0.97 | E | R5 (tile E) |  | No | – |

[a]*P* value for the meta-analysis [18].
[b]RegulomeDB ranks.
[c]Linkage disequilibrium ($r^2$) with tag SNP rs67073037 in CEU.
[d]Regulatory elements such as enhancers (E) and promoters (P).
[e]Regions defined in the current manuscript as in Fig. 1.
[f]Allele-specific activity in Electrophoretic mobility shift assays.
[g]Allele-specific activity in Enhancer Scanning.
[h]miRNA binding to predicted binding site.

miR-141, miR548, miR-376, or nontargeting negative control miR-155, a pMIR-REPORT Luciferase vector (Ambion) containing the *WDR43* 3′UTR region with either haplotype set or an empty vector, and a *Renilla* sp. luciferase driven by a TK promoter as an internal control (Fig. 4C). The pMIR-REPORT empty vector did not show a statistically significant change in luciferase expression when miR-141, miR-548, or miR-376 were cotransfected in MCF10A cells (Fig. 4D, first three samples; Fig. 4E, first two samples). We detected statistically significant increases in cells transfected with the pMIR-REPORT vector containing the *WDR43* 3′UTR with the reference haplotype [GG] and the mimic miR-548 or with the risk haplotype [TC] and the mimic miR-141 or miR-376 (Fig. 4D–E). These results suggest that variation in the risk-associated SNPs rs1131880 and rs11680458 can control the expression of *WDR43* by modulating transcript stability.

### Role of target gene in ER-negative disease

Our analysis shows that *PLB1* and *WDR43* are regulated by risk-associated SNPs at the locus. BC data from METABRIC [25] show that ER-negative status is more frequent in patients with high *WDR43* expression (79.82%) when compared with low *WDR43* expression (3.23%). The same is not true for *PLB1* (27.9% versus 19.1%) (Supplementary Fig. 2A). Both *PLB1* and *WDR43* expressions are higher in TNBC cell lines from the Cancer Cell Line Encyclopedia (Supplementary Fig. 2B) [27]. Further studies are warranted to determine how changes in expression of *PLB1* and *WDR43* contribute to the cancer phenotype.

In summary, we explored five candidate-regulatory regions, RR1–RR5, at the 2p23.2 defined by the presence of SNPs associated with risk to ER-negative BC ($P < 5\times10^{-8}$) at the 2p23.2 locus. Two regions, RR1 and RR3, risk-associated SNPs (part of the credible set of 81 SNPs) displayed allele-specific activity and are therefore likely to contribute to risk. Our data suggest that *PLB1* and WDR43 are targets for RR1 and RR3, respectively.

## DISCUSSION

Although over two hundred loci have been identified through GWAS for BC susceptibility, few have been thoroughly dissected via post-GWAS functional analysis. Among these loci, ~20 have been associated with risk for ER-negative breast tumors [10, 18]. These loci associated with ER-negative BC remain relatively unexplored and little is known about the molecular mechanism driving risk. Here, our objective was to identify likely causal SNPs, those that contribute to increased risk for the ER-negative BC subtype at the locus via allele-specific effects, and illuminate biological mechanisms driving risk at the locus.

Our strategy (Supplementary Fig. 3) started from the assumption that SNPs mechanistically involved in driving risk at the locus would be represented in a set containing 81 credible SNPs from fine mapping analysis of the locus [19]. Integration with publicly available datasets of genomic features (e.g., coding regions, open-chromatin regions, and histone marks of enhancers or promoters) revealed five candidate-functional regions narrowing the set of credible causal SNPs to 12 SNPs located in regions with evidence for promoter/enhancer activity and 2 additional SNPs in a microRNA-binding sequence (Supplementary Fig. 3). We then rigorously evaluated these regions and their risk-associated SNPs in experiments designed to identify allele-specific effects. Experimental analysis identified three transcriptional-regulatory regions, RR1, RR3, and RR4 mediating risk at the locus, although the latter region only had weak evidence of activity.

RR1 contains one risk-associated SNP (rs4407214, $P_{meta} = 7.63\times10^{-09}$) with allele-specific effect on binding to the upstream stimulator factor (USF) family of proteins. *USF1* and *USF2* encode 43–44 kDa proteins that function as both homo- and heterodimers to regulate transcription of target genes [31, 39]. Removal of the sequence surrounding rs4407214 led to a downregulation of *PLB1*, a phospholipase involved in the metabolism of choline. Surprisingly, we did not observe changes in gene expression from the predicted eQTL target *TRMT61* [18]. Interestingly, breast carcinomas have increased concentrations of choline containing compounds and choline kinase inhibitors are selectively cytotoxic to tumor cells [40], which identify *PLB1* as a promising candidate biomarker for triple-negative BC. Our working hypothesis is that rs4407214 is critical to the activity of an enhancer region and the effect allele [G] increases recruitment of transcription factor USF1 to the enhancer, leading to changes in expression levels of *PLB1* with a predicted protective effect on BC risk.

Risk-associated SNPs in RR3 operate through mechanisms different from RR1. In this region we uncovered allele-specific effects of risk-associated haplotypes on the expression of *WDR43* by modulating binding of microRNA (miRNA) to the 3′ UTR of the *WDR43* transcript. *WDR43* is involved in rRNA processing [18]; however, its role in BC is not defined.

Taken together, our data suggest that risk to ER-negative BC associated with the 2p23.2 locus is driven, at least in part, by two common variants, rs4407214 and rs1131880, operating through distinct mechanisms acting on *PLB1* and *WDR43*. While the risk allele in RR3 is expected to lead to degradation of *WDR43* transcript, ER-negative tumors from METABRIC are associated with

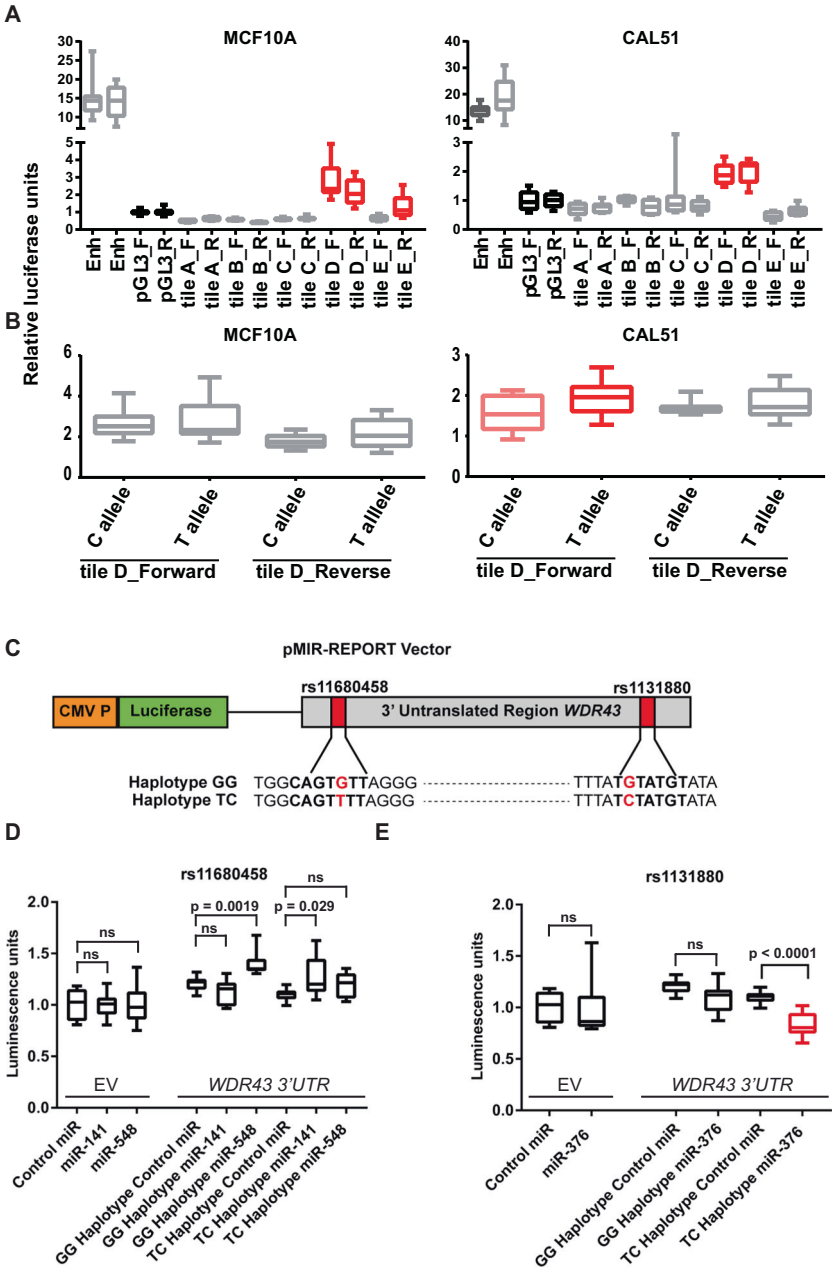


**Fig. 4 Reporter assays. A** Enhancer scanning activity detected by luciferase assays using tiles **A**, **B**, and **C** in RR2, tile **D** in RR4, and tile **E** in RR5 in MCF10A and CAL-51 cells. Red boxes indicate significant ($p < 0.05$) difference in luciferase activity in relation to the no-enhancer control pGL3. **B** Allele-specific enhancer activity detected by luciferase assays using tile D in RR4 in MCF10A and CAL-51 cells. Red boxes indicate significant ($p < 0.05$) difference in luciferase activity between the two alleles. **C** Luciferase reporter assay for allele-specific miRNA regulation. Overview of 3′ UTR region of *WDR43* construct containing "seed regions" (underlined) for miRNA binding to SNPs rs11680458 and rs1131880. Luciferase assays transfected MCF10A cells with either an empty vector (EV) or construct containing either the [GG] or [TC] alleles for rs11680458 (**D**) and rs1131880 (**E**), respectively (*WDR43* 3′UTR). MCF10A cells were cotransfected with miR-141 and miR-548 against rs11680458, and miR-376 against rs1131880 along with miRNA control (miR-155) and a *Renilla* sp. luciferase internal control. Significant differences in a two-tailed test ($p < 0.05$) are shown (ns, nonsignificant).

high *WDR43* expression, which may reflect a compensatory mechanism. Further studies on the role of these genes in BC are needed to determine their biological relevance.

The limitations of this work include the possibility that regulatory landscapes may differ significantly during development, the incomplete regulatory profiling information from normal human mammary-gland cells, and the possibility of not accounting for unknown very rare alleles or with low quality of imputation that contributes to the phenotype. Despite these limitations our strategy identified two SNPs operating to modulate ER-negative BC risk. Similar examples of risk loci containing multiple SNPs contributing functionally to risk have been uncovered in analysis of loci conferring risk to inflammatory bowel diseases [41]. Future studies will likely focus on dissecting the individual contributions from the two SNPs identified in this study.

## DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article and its supplementary information files.

## REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021;71:7–33.
2. Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J Clin Oncol. 2006;24:2137–50.
3. Foulkes WD, Metcalfe K, Sun P, Hanna WM, Lynch HT, Ghadirian P, et al. Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: influence of age, grade, and hitological type. Clin Cancer Res. 2004;10:2029–34.
4. Wahba HA, El-Hadaad HA. Current approaches in treatment of triple-negative breast cancer. Cancer Biol Med. 2015;12:106–16.
5. Fanfani V, Zatopkova M, Harris AL, Pezzella F, Stracquadanio G. Dissecting the heritable risk of breast cancer: From statistical methods to susceptibility genes. Semin Cancer Biol. 2021;72:175–84.
6. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet. 2013;45:392–8. 398e391-392.
7. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet.* 2011;43:1210–4.
8. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013;45:353–61. 361e351-352
9. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47:373–80.
10. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet. 2017;49:1767–78.
11. Purrington KS, Slager S, Eccles D, Yannoukakos D, Fasching PA, Miron P et al. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis* 2014;35:1012–9.
12. Huo D, Feng Y, Haddad S, Zheng Y, Yao S, Han YJ, et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. Hum Mol Genet. 2016;25:4835–46.
13. Fehringer G, Kraft P, Pharoah PD, Eeles RA, Chatterjee N, Schumacher FR, et al. Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. Cancer Res. 2016;76:5103–14.
14. Guo Q, Schmidt MK, Kraft P, Canisius S, Chen C, Khan S et al. Identification of novel genetic markers of breast cancer survival. *J Natl Cancer Inst.* 2015; 107: djv081.
15. Ghoussaini M, French JD, Michailidou K, Nord S, Beesley J, Canisus S, et al. Evidence that the 5p12 variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through FGF10 and MRPS30 regulation. Am J Hum Genet. 2016;99:903–11.
16. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat Genet.* 2016;48:374–386.
17. Wyszynski A, Hong CC, Lam K, Michailidou K, Lytle C, Yao S, et al. An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. Hum Mol Genet. 2016;25:3863–76.
18. Couch FJ, Kuchenbaecker KB, Michailidou K, Mendoza-Fandino GA, Nord S, Lilyquist J, et al. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun. 2016;7:11375.
19. Fachal L, Aschard H, Beesley J, Barnes DR, Allen J, Kar S, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet. 2020;52:56–73.
20. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
21. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43:513–8.
22. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.
23. Creed JH, Monteiro AN, Gerke TA epiTAD: a web application for visualizing high throughput chromosome conformation capture data in the context of genetic epidemiology. *bioRxiv* 2018.
24. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32:D91–4.
25. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.
26. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:pl1.
27. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8.
28. Buckley MA, Woods NT, Tyrer JP, Mendoza-Fandino G, Lawrenson K, Hazelett DJ, et al. Functional analysis and fine mapping of the 9p22.2 ovarian cancer susceptibility locus. Cancer Res. 2019;79:467–81.
29. Xing W, Archer TK. Upstream stimulatory factors mediate estrogen receptor activation of the cathepsin D promoter. Mol Endocrinol. 1998;12:1310–21.
30. Ongwijitwat S, Wong-Riley MT. Is nuclear respiratory factor 2 a master transcriptional coordinator for all ten nuclear-encoded cytochrome c oxidase subunits in neurons? Gene. 2005;360:65–77.
31. Dimova EY, Kietzmann T. Cell type-dependent regulation of the hypoxia-responsive plasminogen activator inhibitor-1 gene by upstream stimulatory factor-2. J Biol Chem. 2006;281:2999–3005.
32. Cong L, Ran FA, Cox D, Lin SL, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/cas systems. Science. 2013;339:819–23.
33. Buckley M, Gjyshi A, Mendoza-Fandino G, Baskin R, Carvalho RS, Carvalho MA, et al. Enhancer scanning to locate regulatory regions in genomic loci. Nat Protoc. 2016;11:46–60.
34. Monteiro AN, Freedman ML. Lessons from postgenome-wide association studies: functional analysis of cancer predisposition loci. J Intern Med. 2013;274:414–24.
35. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;22:1798–812.
36. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat methods. 2012;9:215–6.
37. Liu C, Zhang F, Li T, Lu M, Wang L, Yue W, et al. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics. 2012;13:661–661.
38. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31:3555–7.
39. Sirito M, Lin Q, Maity T, Sawadogo M. Ubiquitous expression of the 43- and 44-kDa forms of transcription factor USF in mammalian cells. Nucleic Acids Res. 1994;22:427–33.
40. Moestue SA, Borgan E, Huuse EM, Lindholm EM, Sitter B, Børresen-Dale A-L, et al. Distinct choline metabolic profiles are associated with differences in gene expression for basal-like and luminal-like breast cancer xenograft models. BMC Cancer. 2010;10:433–433.
41. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal Lari R, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014;24:1–13.

## AUTHOR CONTRIBUTIONS

G.M.F. and A.M. conceived the project and designed the experiments. G.M.F. and C.H. performed the experiments. G.M.F., P.L., T.N., C.H., and A.M. performed the analysis and interpreted the results. All authors contributed to the overall data interpretation, provided intellectual input, and approved the final paper.

## FUNDING

This work was funded by the Florida Breast Cancer Foundation, Moffitt Foundation, and by support from the Molecular Genomics Facilities at H. Lee Moffitt Cancer Center & Research Institute, an NCI designated Comprehensive Cancer Center (P30-CA076292).

## ETHICAL APPROVAL

This study does not include human subjects and therefore ethical approval is not applicable.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41431-021-01005-6.

**Correspondence** and requests for materials should be addressed to Alvaro N. A. Monteiro.