

# **Long noncoding RNAs, *CUPID1* and *CUPID2*, mediate breast cancer risk at 11q13 by modulating response to DNA damage**

Joshua A Betts<sup>1,2,9</sup>, Mahdi Moradi Marjaneh<sup>1,9</sup>, Fares Al-Ejeh<sup>1,9</sup>, Yi Chieh Lim<sup>1</sup>, Wei Shi<sup>1</sup>, Haran Sivakumaran<sup>1</sup>, Romain Tropée<sup>3</sup>, Ann-Marie Patch<sup>1</sup>, Michael B Clark<sup>4,5</sup>, Nenad Bartonicek<sup>5,6</sup>, Adrian P Wiegman<sup>1</sup>, Kristine M Hillman<sup>1</sup>, Susanne Kaufmann<sup>1</sup>, Amanda L Bain<sup>1</sup>, Brian S Gloss<sup>5,6</sup>, Joanna Crawford<sup>7</sup>, Stephen Kazakoff<sup>1</sup>, Shivangi Wani<sup>1</sup>, Shu W Wen<sup>1</sup>, Bryan Day<sup>1</sup>, Andreas Möller<sup>1</sup>, Nicole Cloonan<sup>1</sup>, John Pearson<sup>1</sup>, Melissa A Brown<sup>2</sup>, Timothy R Mercer<sup>5,6</sup>, Nicola Waddell<sup>1</sup>, Kum Kum Khanna<sup>1</sup>, Eloise Dray<sup>3,8</sup>, Marcel E Dinger<sup>5,6</sup>, Stacey L Edwards<sup>1,10,\*</sup>, Juliet D French<sup>1,10,\*</sup>

<sup>1</sup> Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, 4029, Australia.

<sup>2</sup> School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, 4006, Australia.

<sup>3</sup> Queensland University of Technology at the Translational Research Institute, Brisbane, 4102, Australia.

<sup>4</sup> Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, OX1 2JD, UK.

<sup>5</sup> Garvan Institute of Medical Research, Sydney, 2010, Australia.

<sup>6</sup> St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, 2052, Australia.

<sup>7</sup> Institute for Molecular Bioscience, University of Queensland, Brisbane, 4072, Australia.

<sup>8</sup> Queensland University of Technology, Institute of Health and Biomedical Innovation, Brisbane, 4059, Australia.

<sup>9</sup> These authors contributed equally to this study

<sup>10</sup> These authors contributed equally to this study

\*Correspondence: [stacey.edwards@qimrberghofer.edu.au](mailto:stacey.edwards@qimrberghofer.edu.au) (S.L.E.),  
[juliet.french@qimrberghofer.edu.au](mailto:juliet.french@qimrberghofer.edu.au) (J.D.F.)

## ABSTRACT

Breast cancer risk is strongly associated with an intergenic region on 11q13. We have previously shown that the strongest risk-associated SNPs fall within a distal enhancer that regulates *CCND1*. Here, we report that, in addition to *CCND1*, this enhancer regulates two estrogen-regulated long noncoding RNAs, *CUPID1* and *CUPID2*. We provide evidence that the risk-associated SNPs are associated with reduced chromatin looping between the enhancer and the *CUPID1/2* bidirectional promoter. We further show that *CUPID1* and *CUPID2* are predominantly expressed in hormone receptor-positive breast tumors and play a role in modulating double strand break (DSB) repair pathway choice. These data reveal a mechanism for the involvement of this region in breast cancer.

## KEYWORDS

Long noncoding RNAs, breast cancer, GWAS, enhancer, DNA repair.

## INTRODUCTION

Genome wide association studies (GWAS) provide an excellent means of identifying single nucleotide polymorphisms (SNPs) that are associated with breast cancer [MIM: 114480] risk.<sup>1,2</sup> As the majority of identified SNPs fall outside of protein coding regions,<sup>3</sup> a major challenge is determining how genetic variation in these regions contributes to disease. Several studies have shown that GWAS SNPs are enriched in regulatory regions and can influence the expression of nearby protein coding genes.<sup>3-6</sup> However, the contribution of noncoding RNAs to disease risk is still relatively unexplored, in part due to the fact that noncoding RNAs are poorly annotated.

One of the strongest breast cancer risk associations identified is with rs614367 located in an intergenic region at 11q13.<sup>7</sup> This association is restricted to risk of estrogen receptor positive (ER<sup>+</sup>) tumors. We previously fine-mapped this locus and found that the strongest genetic signal contained four SNPs that were located within a distal transcriptional enhancer (called PRE1) of *CCND1* [MIM: 168461].<sup>4</sup> Here, we identify two lncRNAs transcribed from this region and we name them *CUPID1* and *CUPID2* (*CCND1*-*Upstream Intergenic DNA Repair 1/2*). We find that PRE1 acts as a strong enhancer on the *CUPID1/2* bidirectional promoter. We additionally demonstrate, through functional analyses, the potential relevance of these lncRNAs to breast cancer.

## **MATERIALS AND METHODS**

### ***Cell lines***

Breast cancer cell lines MCF7, BT474 and T47D were grown in RPMI medium supplemented with 10% fetal calf serum (FCS), sodium pyruvate, 10 µg/ml insulin and antibiotics. ZR751, CAL51, Hs578T and MDAMB231 were grown in DMEM medium supplemented with 10% FCS and antibiotics. Normal breast epithelial cell lines, MCF10A and Bre-80 (a gift from Roger Reddel, CMRI, Sydney) were grown in DMEM/F12 medium with 5% horse serum, 10 µg/ml insulin, 0.5 µg/ml hydrocortisone, 20 ng/ml epidermal growth factor, 100 ng/ml cholera toxin and antibiotics. HMECs were purchased from Lonza and grown in basal medium (MEBM, Lonza) supplemented with SingleQuots (MEGM BulletKit, Lonza). Cell lines were maintained routinely tested for *Mycoplasma* and short tandem repeat (STR) profiled.



### ***Estrogen induction***

MCF7 cells were induced with 17 $\beta$ -Estradiol (Sigma-Aldrich) by first incubating cells with 10nM Fulvestrant (ICI 182780, Sigma) for 48 h and then treating the cells with either 10 nM 17 $\beta$ -Estradiol or DMSO (vehicle control) for 24 h.

### ***Cell fractionation***

MCF7 cells were fractionated as previously described.<sup>8</sup>

### ***Quantitative real-time PCR***

cDNA was synthesised using SuperScript III (Invitrogen). qPCR was performed using Taqman assays (Life Technologies) or intron-spanning primers to determine gene expression (**Table S1**).

### ***RNA CaptureSeq***

RNA CaptureSeq was performed on MCF7 cells with and without estrogen stimulation and on five breast cell lines; HMEC, T47D, BT474, Hs578T and MDAMB231. Oligonucleotide probes were designed to capture a 390kb intergenic region on 11q13 (chr11:69064954-69455655) as part of a larger study using the Roche NimbleGen Sequence Capture design algorithm. Locations of the capture probes are provided in **Table S2**. Sequencing libraries were prepared and capture hybridisations performed as described in.<sup>9</sup> De novo transcript assembly was conducted as previously described.<sup>9,10</sup>

### ***Source data files***

Raw and processed files for RNA CaptureSeq are available under E-MTAB-4519 at ArrayExpress.

### ***Chromatin conformation capture (3C)***

3C libraries were generated using *NcoI* or *EcoRI* as described previously.<sup>5</sup> 3C interactions were quantitated by qPCR using primers designed within restriction fragments (**Table S1**).

### ***Luciferase assays***

Promoter-driven reporter constructs were generated by inserting a *CUPID1/2* promoter sequence (chr11:69,294,069-69,296,583) into the *KpnI-XhoI* sites of pGL3-Basic in the 5'-3' (*CUPID1*) or 3'-5' (*CUPID2*) direction (**Figure S1**). A 3,334 bp fragment (chr11: 69,329,695-69,333,028) containing PRE1 with or without the risk SNPs was then inserted into the *BamHI-SalI* site of promoter-driven reporter constructs. Primers used to generate the promoter constructs are listed in **Table S1**. Luciferase assays were performed as described previously.<sup>5</sup>

### ***CUPID1 and CUPID2 overexpression***

Sequences corresponding to the main isoforms of *CUPID1* and *CUPID2* (**Figure S2**) were synthesised by GenScript and inserted into the *Nhe1-Not1* or *Nhe1-BamHI* sites of pCDNA3.1. To re-establish *CUPID1/2* expression; (i) For RAD51 foci formation, MCF7 *CUPID1/2*-CRISPRi cells were transfected with 2 µg of *CUPID1/2* expression plasmids using the Nucleofector kit V (Amaxa). (ii) For the HR assay, MCF7 DR-GFP cells were co-transfected with I-SceI and 2 µg of *CUPID1/2* expression plasmids using Lipofectamine 3000 (Invitrogen).

### ***siRNA silencing***

Two independent siRNAs were designed against *CUPID1* or *CUPID2* using the Dharmacon siDesign tool. A pool of three siRNAs were used against RAD51. The sequences are listed in **Table S1**. 10 nM of siRNA was transfected into MCF7 cells using RNAiMax (Invitrogen) and cells were harvested 48h later.

### ***CRISPR interference (CRISPRi)***

Single-guide RNA sequences targeting the 11q13 PRE1 or *CUPID1/2* promoter are provided in **Table S1**. The sgRNA target, Cas9 binding handle and terminator sequences were synthesized (Integrated DNA Technologies) and cloned into the lentiviral vector pgRNA-humanized. Virus-like particles (VLPs) containing either dCas9-KRAB or a targeting sgRNA were generated by transfection of HEK293 cells with Lipofectamine 2000 (Thermo Fisher). Cells were cotransfected with the packaging plasmid pCMV-dR8.91, the VSV-G expression plasmid pCMV-VSV-G,<sup>11</sup> and with either pHR-SFFV-dCas9-BFP-KRAB or pgRNA-humanized. VLPs were collected, mixed in equal volume used to transduce MCF7 DR-GFP cells. Cells expressing mCherry (pgRNA) and blue fluorescent protein (dCas9-KRAB) were FACS sorted on an ARIA IIIu (Becton-Dickinson).

### ***RNAseq***

RNAseq analysis was performed on three independent replicates for each siRNA (nontargeting control, *CUPID1* or *CUPID2*). Total RNA was extracted 48h after transfection with siRNAs using Trizol and 1 µg was rRNA depleted using an Illumina Ribo-Zero rRNA Removal Kit. RNAseq libraries were prepared using the NEBNext

Ultra Directional RNA Library Prep Kit for Illumina (NEB), and sequenced at 1 x 75bp on an Illumina NextSeq 500. The sequencing reads were assessed for data quality using fastqc v0.11.3. The RNAseq reads were then mapped to the human transcriptome (GRCh38.primary.assembly.genome.fa and annotation.gtf files from GENCODE Release 23) using RNA STAR v2.4.0k.<sup>12</sup> The expression levels of the transcripts were quantified using rsem-1.2.22.<sup>13</sup> The resulted FPKM values were transferred to Partek Genomics Suite 6.6 package. Principal components analysis (PCA) was performed with all FPKM values for the experiment and control samples to assess the variability between the replicates. Differentially expressed genes between the *CUPID1/2* siRNA experiments and controls were then identified using ANOVA ( $p < 0.05$ ). The significance of overlap between *CUPID1/2* regulated genes was tested using hypergeometric distribution carried out by R *phyper* function. Ingenuity Pathway Analysis (IPA®) on the differentially expressed genes was conducted as described previously.<sup>14</sup>

### ***Expression of CUPID1/2 in the TCGA cohort***

RNAseq data from The Cancer Genome Atlas (TCGA) breast invasive carcinoma cohort (1226 samples) was obtained from the UCSC Cancer Genomics Hub (30/09/2015). Reads were trimmed for adapter sequences using Cutadapt (version 1.11) and aligned using STAR (version 2.5.2a) to the GRCH37 assembly with the gene, transcript, and exon features of *CUPID1* and *CUPID2* appended to the Ensembl (release 75) gene model. Quality control metrics were computed using RNA-SeQC (version 1.1.8) and expression was estimated using RSEM (version 1.2.30).

### ***Analysis of TCGA whole genome sequencing (WGS) data***

WGS data for 118 tumor and matched normal breast samples was accessed from TCGA data portal (**Table S3**). All TCGA data underwent reanalysis in order to standardise the results used in the downstream analyses across all samples. Initially fastq files were recreated from the BAM files. Sequence reads were then trimmed using Cutadapt (version 1.9) prior to alignment to GRCh37 with BWA-MEM (version 0.7.13-r1126).<sup>15</sup> Duplicate aligned reads were marked with Picard (version 1.141) and aligned data was coordinate sorted using Samtools (version 1.3). Somatic substitutions were detected using a dual calling strategy; with substitutions detected using qSNP<sup>16</sup> and the Haplotype caller module of GATK<sup>17</sup>. High confidence somatic substitutions used in downstream analysis passed variant characteristic filtering: minimum coverage depth of 8 reads for control data and 12 reads for tumor data; at least 5 variant supporting reads where the variant is not within the first or last 5 bases and have individual start positions; and the variant supported by reads in both sequencing directions and more than 5 base pairs from a mono-nucleotide run of 7 or more bases in length; somatic variants had less than 3% variant evidence identified in the control sample.

Defects in homologous recombination were assessed using the presence of mutational signatures and the distribution pattern of structural rearrangements. The flanking nucleotide context of high confidence whole genome somatic substitution mutations was used to generate mutational signatures for each sample using a published framework.<sup>18</sup> Identified signatures were compared to other validated signatures (i.e. COSMIC) and the frequency of each signature per Mb was determined. Somatic structural variants were identified using the qSV tool and for

genomes with >200 SV events distinct patterns or clusters of breakpoint distributions were determined as described previously.<sup>19-21</sup> For the characterization of genomic breakpoint distributions chromosomes bearing a significantly clustered distribution of breakpoints (as described,<sup>22</sup> goodness of fit threshold  $P < 0.00001$ ) and those containing outlying high numbers of rearrangement events were identified (outliers had a breakpoint per megabase rate exceeding five times the length of the inter-quartile range from the seventy-fifth percentile for each sample with a minimum threshold of 30 breakpoints per chromosome). Genomes with less than 8 chromosomes with significantly clustered breakpoints and at least 1 containing an extreme outlying density of breakpoints were classified as containing focal rearrangement events. Genomes with eight or more chromosomes with clustered breakpoints and a maximum of one extreme density events were classified as scattered. Manual review of breakpoint distribution confirmed the classifications and characterized borderline cases.

### ***H2AX / FACS method***

Flow cytometric assays using biotinylated anti- $\gamma$ H2AX (JBW301, Millipore) and DNA content staining with 7AAD (Invitrogen) were carried out as described previously<sup>23</sup> and samples were acquired and analyzed using FACSCanto II (Becton Dickinson).

### ***Immunofluorescence***

Cells were seeded on coverslips and exposed to 5 Gy of ionising-radiation (IR) in a Gammacel. Cells were harvested post-IR at different time-points and pre-extracted with 0.5% NP-40 in cytoskeletal (CSK) buffer (100 mM NaCl, 300 mM sucrose, 3 mM MgCl<sub>2</sub>, 10 mM PIPES [pH 6.8]). Immunostaining for anti-RAD51 (sc-8349, Santa

Cruz), anti-phospho-RPA (S4S8, Bethyl), anti-γH2AX (JBW301, Millipore) and anti-53BP1 (4937s, Cell Signaling) was performed as described previously.<sup>24</sup>

### ***Homologous Recombination (HR) assay***

MCF7 DR-GFP cell line was a gift from Dr. Maria Jasin and contains the DR–GFP cassette in MCF7 cells. The HR assay was performed as described previously.<sup>25</sup>

### ***Non-homologous end joining (NHEJ) assay***

The NHEJ reporter substrate has been described previously.<sup>26</sup>

## **RESULTS**

### **LncRNAs expressed from the 11q13 breast cancer risk locus**

To identify lncRNAs transcribed from the 11q13 breast cancer risk locus, we performed RNA CaptureSeq on breast cancer cell lines using Capture probes designed to a 390kb intergenic region on 11q13 flanked by *MYEOV* and *CCND1* (**Table S2**). We identified two lncRNAs, which share a bidirectional promoter, transcribed ~20 kb away from PRE1 which we named *CUPID1* (transcribed from the positive strand) and *CUPID2* (transcribed from the negative strand; **Figures 1A, 1B, and S2**). Multiple splice isoforms were detected for *CUPID1*, one of which corresponds to the RefSeq lncRNA gene LINC01488. A predicted lncRNA transcript corresponding to *CUPID2* (AP000439.3) was also found in GENCODE (Version 19; **Figures 1A and 1B**).

### **Effect of risk SNPs on the *CUPID1* and *CUPID2* promoter**

PRE1 is a distal transcriptional enhancer within the 11q13 breast cancer risk region

that we have previously shown to regulate the expression of *CCND1*.<sup>4</sup> Using chromosome conformation capture (3C) we showed that PRE1 also frequently interacted with the predicted *CUPID1/2* promoter in normal breast and cancer cell lines (**Figures 1C** and **S3**). We also observed an additional interaction of unknown significance ~5 kb from the *CUPID1/2* promoter. Reporter assays indicated that PRE1 acts as a strong enhancer on the *CUPID1/2* promoter, and inclusion of the minor (risk) allele of two of the four 11q13 breast cancer risk SNPs (rs661204 and rs78540526) significantly reduced the PRE1 enhancer activity only on the *CUPID1* promoter (**Figures 1D, S4A** and **S4B**). Furthermore, in CAL51 cells, a breast cancer cell line heterozygous for the risk SNPs, allele-specific 3C showed only the protective allele participates in chromatin looping suggesting that the risk allele abrogates looping between PRE1 and the *CUPID1/2* promoter (**Figures 1E** and **S4C**). We also silenced PRE1 by targeting a nuclease-inactive dCas9 fused to the Krüppel-associated box (KRAB) repressor (dCas9-KRAB)<sup>27</sup> to the PRE1 enhancer and confirmed that *CUPID1*, *CUPID2* and *CCND1* levels are significantly reduced (**Figure 1F**).

### ***CUPID1* and *CUPID2* expression**

RNA CaptureSeq data showed that *CUPID2* is widely expressed in multiple tissues, however *CUPID1* was predominantly expressed in ER<sup>+</sup> breast cancer cell lines (**Figure S5**). Using qPCR in breast cell lines we showed that *CUPID1/2* transcripts are more highly expressed in ER<sup>+</sup> breast cancer cell lines. Higher expression was observed in BT474 and MCF7 cells, which are amplified at the 11q13 region (**Figure 2A**). We determined the subcellular localisation of *CUPID1* and *CUPID2* relative to a number of well-characterized controls. As shown in **Figure 2B**, *CUPID1* was



enriched in the nucleus, whereas *CUPID2* is nuclear and cytoplasmic. *CUPID1/2* expression was significantly induced in MCF7 cells treated with estrogen (**Figure 2C**). Notably, estrogen induction of the *CUPID1/2* promoter depended on the presence of PRE1, however inclusion of the risk-associated SNPs did not attenuate this effect (**Figure S6**). We also examined the expression of *CUPID1/2* in human breast tumors using TCGA RNAseq data. *CUPID1/2* were expressed in 60.1% and 78.7% of all breast tumors (FPKM $\geq$ 0.1), respectively. Their expression was only weakly associated with 11q13 amplification, indicating that copy number variation is not the sole mechanism underlying the observed overexpression (**Figure S7**). *CUPID1/2* expression was highly correlated with each other (correlation=0.796) and similar to *CCND1* both were more highly expressed in hormone receptor positive (HR<sup>+</sup>) tumors, specifically in luminal A and B subtypes (**Figures 2D, 2E and S8**).

### **Effect of *CUPID1* and *CUPID2* on gene expression**

To gain insight into the potential functions of *CUPID1/2*, we silenced both lncRNAs using independent siRNAs in MCF7 cells and performed cell cycle analyses (**Figure S9**). In contrast to *CCND1* knockdown, *CUPID1* and *CUPID2* siRNA did not significantly alter cell cycle progression (**Figure S10**). Of note, *CUPID1* and *CUPID2* likely regulate each other as shown by reduced expression of both lncRNAs after either *CUPID1* or *CUPID2* siRNA treatment. We then performed RNAseq and identified 1847 and 1835 genes that were differentially regulated upon repression of *CUPID1* and *CUPID2*, respectively, compared to a non-targeted control siRNA (**Figure 2F and Table S4**). Notably, we observed a significant overlap (362 genes,  $P<1e-100$ ) of the genes regulated by both *CUPID1/2* (**Figure 2F and Table S5**). Consistent with this, Ingenuity Pathway Analysis (IPA) of the *CUPID1/2* regulated

genes showed an overlap in biological functions (**Figure S11A** and **Table S6**), the majority of which are commonly disrupted in cancer. IPA analysis of the overlapping genes identified five networks with the top networks being Molecular Transport, Cellular Assembly and Organisation, DNA replication, Recombination and Repair (**Figure S11B**, **Tables S7** and **S8**). Of the 362 genes regulated by both *CUPID1/2* identified by silencing, approximately one quarter (91/362, 25.1%) correlated with *CUPID1/2* expression in the TCGA cohort (**Figure 2G** and **Table S9**). Notably, we showed that *CUPID1/2*-induced genes shared a similar pattern of high expression in HR<sup>+</sup>, luminal A and B breast cancer subtypes. Conversely, *CUPID1/2*-suppressed genes were inversely correlated with *CUPID1/2* expression across the same breast cancer subtypes (**Figure 2G**). These data suggest that *CUPID1/2* drives the expression of these genes in breast tumors.

### ***CUPID1* and *CUPID2* silencing inhibits HR mediated DNA repair**

We have previously shown that the strongest risk-associated SNPs at 11q13 act to reduce *CCND1* expression, which conflicts with its demonstrated oncogenic function.<sup>4</sup> However, Cyclin D1 also plays a role in homologous recombination-mediated DNA repair (HRR), a function independent of its role in cell cycle control.<sup>28</sup> Given this and the fact that pathway analyses indicated that *CUPID1/2*-regulated genes affect DNA repair and recombination we hypothesized that PRE1 may act as an enhancer on other genes in the HRR pathway, including *CUPID1/2*. In support of this, we show that breast tumors with low *CUPID1/2* or *CCND1* expression frequently have somatic mutation signatures<sup>18</sup> that are consistent with defective HRR (**Figures 3A, S12, S13A** and **Table S3**). The same signature has previously been shown to be associated with mutations in two other HRR genes, *BRCA1* [MIM:

113705] and *BRCA2* [MIM: 600185].<sup>29</sup> Consistent with impaired HRR, we also show that low *CUPID1/2* or *CCND1* expression is associated with a large number of structural variants distributed through the genome<sup>19</sup> (**Figures 3B, S13B, S14 and S15**).

We then used a stable MCF7 cell line with two non-functional GFP alleles (MCF7 DR-GFP), in which GFP is activated only after HR-mediated repair and showed reduced HR repair efficiency upon *CUPID1/2* knockdown using two independent siRNAs (**Figures 3C and S16A**). Consistent with this, we show that depletion of *CUPID1/2* by targeting dCAS9-KRAB to the *CUPID1/2* bidirectional promoter also impaired HRR (**Figures 3D and S17**). Importantly, re-expression of *CUPID1* and/or *CUPID2* in these cells almost completely restores HR repair efficiency (**Figures 3D, and S18A**). We also noted that re-expression of *CUPID2* also restored *CUPID1* levels (**Figure S18A**). We then reduced *CUPID1/2* levels using either siRNA or dCas9-KRAB targeted to the *CUPID1/2* promoter and observed that recruitment of RAD51 to ionizing radiation (IR)-induced sites of DNA damage was impaired (**Figures 3E, 3F, 3G, S16B, S17 and S19**). Again, we show that re-expression of *CUPID1* and/or *CUPID2* in the *CUPID1/2* silenced cells restores RAD51 foci formation after IR treatment (**Figures 3G and S18B**). There was no significant change in *CCND1* expression following knockdown or re-expression of either *CUPID1* or *CUPID2* (**Figure S18C**).

### ***CUPID1/2* regulates NHEJ/HR pathway choice.**

End resection is a key step in the initiation of HRR that involves the generation of a single stranded DNA (ssDNA) stretch. This ssDNA stretch is then rapidly coated by

the phosphorylated version of Replication Protein A (RPA) at residues S4/S8 (pRPA), until it is subsequently replaced by RAD51. Given RAD51 recruitment was impaired in *CUPID1/2* silenced cells, we performed immunofluorescence assays to interrogate pRPA foci formation after irradiation (**Figure 4A**). In contrast to control cells where pRPA peaked at 2h, we observed that cells depleted for *CUPID1/2* had significantly lower pRPA at 2h that persisted until 4h before resolution (**Figures 4A, 4C and S20**). In addition, while *CUPID1/2* silenced cells had similar levels of  $\gamma$ H2AX, a marker of DSBs, at 0.5h post irradiation, *CUPID1/2* depleted cells had significantly higher levels of  $\gamma$ H2AX foci 2h post irradiation suggesting a mild delay in DSB repair (**Figures 4B and 4C**). Despite this delay,  $\gamma$ H2AX were fully resolved after 16h. Consistent with this, we observed an increase in 53BP1 foci 2h post-irradiation (**Figures 4B, 4C and 4D**) indicating that DSB repair by non-homologous end joining (NHEJ) may be compensating for reduced HRR efficiency. We also calculated the percentage of cells with more than five  $\gamma$ H2AX/53BP1 foci/cell (**Figure 4E**) as an indicator for DSBs primed for NHEJ repair marked by their co-localization. Together, the inverse correlation of the low number of pRPA foci (**Figure 4D**) compared to the high number of 53BP1/ $\gamma$ H2AX foci (**Figure 4E**) at 2h indicates a preference for NHEJ for DSB repair. To confirm whether the DSBs formed could be repaired by NHEJ in this HRR deficient context, we used a reporter assay for NHEJ, whereby successful repair of NHEJ after DNA digestion will restore expression of GFP. We observed a mild increase in NHEJ repair in MCF7 cells transfected with either *CUPID1* or *CUPID2* siRNA suggesting that NHEJ may compensate for the loss of HRR (**Figures 4F and S16C**). Finally, the lack of a noticeable difference in the delay of the cell cycle progression between *CUPID1/2* silenced cells and control cells

(Figure 4G) supports the notion that the DNA DSBs in these mutant cells were eventually repaired.

## DISCUSSION

A key challenge for post-GWAS analysis is interpreting the mechanisms of action of risk-associated SNPs, as the majority lie in noncoding regions of the genome. Although it is clear that *cis*-regulatory variation is a common mechanism underlying many associations, the role of lncRNAs in influencing disease susceptibility is only now being realized. Recent studies indicate lncRNAs are transcribed from cancer risk loci and that these transcripts can play important roles in tumorigenesis.<sup>30-32</sup> Here, we describe two estrogen-regulated lncRNAs (*CUPID1* and *CUPID2*) identified by targeted RNA sequencing that may contribute to the risk of developing breast cancer by modulating pathway choice of DSB repair.

We showed by chromosome conformation capture, reporter assays and CRISPR interference that PRE1 acts as an enhancer on the *CUPID1/2* bidirectional promoter. In addition, allele-specific 3C between PRE1 and the *CUPID1/2* promoter performed in a cell line heterozygous for the risk SNPs showed preferential looping of the protective alleles. These findings suggest that risk-associated SNPs may abrogate chromatin looping, resulting in reduced promoter activity and subsequent transcription.<sup>33</sup> In addition, we also showed that the risk SNPs reduced PRE1 activity on the *CUPID1* promoter. However, in the absence of chromatin looping between PRE1 and the *CUPID1/2* promoter, the significance of this observation is unclear. These data would clearly be strengthened by analysis of isogenic lines with and

without incorporation of the risk alleles. However, due to amplification of this region in ER<sup>+</sup> cell lines, it has not been possible to generate such cell lines.

We showed that both lncRNAs are induced following estrogen stimulation. Interestingly, the estrogen induction of *CUPID1* and *CUPID2* requires the presence of PRE1, as estrogen stimulation of a construct containing the promoter alone had no effect. In reporter assays, the risk SNPs did not affect the estrogen induction of the *CUPID1/2* promoters. This was not unexpected as the risk SNPs are not predicted to affect estrogen receptor binding. However, in cells heterozygous for the risk SNPs we would expect that estrogen induction of *CUPID1/2* would only occur on the protective allele as chromatin looping was abrogated in the presence of the risk allele. Unfortunately, we did not identify any ER<sup>+</sup> cell lines that are heterozygous for the risk SNPs therefore we could not confirm this by allele-specific expression analyses.

We provide evidence that *CUPID1* and *CUPID2* are important for HR mediated repair (HRR). Silencing of *CUPID1/2* delayed the formation of phosphorylated RPA foci and inhibited RAD51 recruitment to double strand breaks (DSBs) post irradiation suggesting that reduced *CUPID1/2* expression may lead to a defect in end resection, a critical step required for the initiation of the HRR pathway. Notably, we show that *CUPID1/2* silencing does not lead to increased  $\gamma$ H2AX foci suggesting that *CUPID1/2* does not affect overall DSB repair but likely affects DNA repair pathway choice. Consistent with this, we show that the recruitment of 53BP1 to DSBs is promoted and NHEJ is still functional while HRR is decreased. Importantly, these data are consistent with our observations in breast tumors where low *CUPID1/2*

expression is associated with a HR mutation signature and high levels of scattered structural variants.

We previously reported that the breast cancer risk SNPs at 11q13 act to reduce *CCND1* expression, whose protein product, Cyclin D1, is required for G1 to S phase cell cycle transition. Cyclin D1 is recruited to DNA damage sites and reduction of cyclin D1 impairs the recruitment of RAD51 to DSBs and impedes HR DNA repair.<sup>28</sup> Our study herein on the same locus that affects *CUPID1/2* expression has revealed that these lncRNAs also play a role in DSBs repair similar to cyclin D1. As depicted in our proposed model (**Figure S21**), the loss of *CUPID1/2* expression results in reduced DNA end resection, defect in pRPA and RAD51 recruitment leading to reduced HR DNA repair and favors 53BP1 recruitment and the choice of NHEJ as the DNA repair pathway. The role of BRCA1 and 53BP1 and associated partners in the choice of DNA repair pathway is an area of ongoing research.<sup>34</sup> Our results implicate *CUPID1/2* in this field and future work should determine whether the lncRNAs directly affect DNA end resection, the relationship between 53BP1/RIF1/PTIP and BRCA1/CtIP at the DSB sites and/or chromatin remodelling particularly through histone modification around the DNA breaks. Alternatively, *CUPID1* may be acting as an RNA scaffold for DNA repair complexes in a similar way as LINP1, which has recently been shown to promote NHEJ by acting as a RNA scaffold for Ku70-Ku80 and DNA-PKcs.<sup>35</sup>

In summary, we have shown that breast cancer risk variants at 11q13 may regulate two lncRNAs, in addition to *CCND1*, and provide evidence that reduced expression of the lncRNAs mediates risk by favoring a switch from HRR to NHEJ DSB repair.

Importantly, these data are consistent with observations in breast tumors whereby a HR mutation signature is associated with increased structural variants presumably caused by DSB repair through error prone pathways. This study highlights the importance of annotating all noncoding RNAs expressed from GWAS loci and further consolidates the premise that aberrant lncRNA expression contributes to the etiology of many human cancers. Given that the majority of noncoding RNAs have no assigned function, it is likely that lncRNAs will provide a wealth of opportunities for uncovering novel pathways that could potentially be targeted for cancer therapy.

## **ACKNOWLEDGMENTS**

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC; 1021731, 1058421). JDF, SLE, APW were supported by Fellowships from the National Breast Cancer Foundation (NBCF) Australia, JAB was supported by an Australian Postgraduate Scholarship, FA was supported by a Future Fellowship from the Australia Research Council (ARC - FT130101417), MBC was supported by an NHMRC Early Career Fellowship (APP1072662) and EMBO Long Term Fellowship (ALTF 864-2013), NW was supported by an NHMRC Career Development Fellowship (APP1112113) and KKK was supported by an NHMRC Senior Principal Research Fellowship. The results published in this study are in part based upon data generated by the TCGA Research Network. The contents of the published material are solely the responsibility of the administering institution, a participating institution or individual authors and do not reflect the views of NHMRC. The procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and that proper informed consent was obtained. TRM is the recipient of a Roche



Discovery Agreement (2014). MBC has received research support from Roche/Nimblegen for an unrelated research project. No potential conflicts of interest were disclosed by the other authors.

## WEB RESOURCES

ArrayExpress, <https://www.ebi.ac.uk/arrayexpress/>

COSMIC, <http://cancer.sanger.ac.uk/cosmic/signatures/>

ENCODE, <https://www.encodeproject.org/>

Fastqc v0.11.3, <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>

OMIM, <https://www.omim.org/>

Partek, <http://www.partek.com/>

qSV tool, <http://sourceforge.net/projects/adamajava/>

The Cancer Genome Atlas, <http://cancergenome.nih.gov/>

## REFERENCES

1. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353-361.
2. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373-380.
3. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195.
4. French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.* **92**, 489-503.
5. 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.

6. Glubb, D.M., Maranian, M.J., Michailidou, K., Pooley, K.A., Meyer, K.B., Kar, S., Carlebur, S., O'Reilly, M., Betts, J.A., Hillman, K.M., et al. (2015). Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am. J. Hum. Genet.* **96**, 5-20.
7. Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C.S., et al. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504-507.
8. Vance, K.W., and Ponting, C.P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet.* **30**, 348-355.
9. Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9**, 989-1009.
10. Clark, M.B., Mercer, T.R., Bussotti, G., Leonardi, T., Haynes, K.R., Crawford, J., Brunck, M.E., Cao, K.A., Thomas, G.P., Chen, W.Y., et al. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat. Methods* **12**, 339-342.
11. Zufferey, R., Nagy, D., Mandel, R.J., Naldini, L., and Trono, D. (1997). Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat. Biotechnol.* **15**, 871-875.
12. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21.
13. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323.
14. Al-Ejeh, F., Miranda, M., Shi, W., Simpson, P.T., Song, S., Vargas, A.C., Saunus, J.M., Smart, C.E., Mariasegaram, M., Wiegman, A.P., et al. (2014). Kinome profiling reveals breast cancer heterogeneity and identifies targeted therapeutic opportunities for triple negative breast cancer. *Oncotarget* **5**, 3145-3158.
15. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595.
16. Kassahn, K.S., Holmes, O., Nones, K., Patch, A.M., Miller, D.K., Christ, A.N., Harliwong, I., Bruxner, T.J., Xu, Q., Anderson, M., et al. (2013). Somatic point mutation calling in low cellularity tumors. *PLoS One* **8**, e74380.
17. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303.
18. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. *Nature* **500**, 415-421.
19. Waddell, N., Pajic, M., Patch, A.M., Chang, D.K., Kassahn, K.S., Bailey, P., Johns, A.L., Miller, D., Nones, K., Quek, K., et al. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501.
20. Patch, A.M., Christie, E.L., Etemadmoghadam, D., Garsed, D.W., George, J., Fereday, S., Nones, K., Cowin, P., Alsop, K., Bailey, P.J., et al. (2015). Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-494.

21. Nones, K., Waddell, N., Wayte, N., Patch, A.M., Bailey, P., Newell, F., Holmes, O., Fink, J.L., Quinn, M.C., Tang, Y.H., et al. (2014). Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nat. Commun.* 5, 5224.
22. Korbel, J.O., and Campbell, P.J. (2013). Criteria for inference of chromothripsis in cancer genomes. *Cell* 152, 1226-1236.
23. Al-Ejeh, F., Staudacher, A.H., Smyth, D.R., Darby, J.M., Denoyer, D., Tsopelas, C., Hicks, R.J., and Brown, M.P. (2014). Postchemotherapy and tumor-selective targeting with the La-specific DAB4 monoclonal antibody relates to apoptotic cell clearance. *J. Nucl. Med.* 55, 772-779.
24. Wiese, C., Dray, E., Groesser, T., San Filippo, J., Shi, I., Collins, D.W., Tsai, M.S., Williams, G.J., Rydberg, B., Sung, P., et al. (2007). Promotion of homologous recombination and genomic stability by RAD51AP1 via RAD51 recombinase enhancement. *Mol. Cell* 28, 482-490.
25. Parpys, A.C., Zhao, W., Sharma, N., Groesser, T., Liang, F., Maranon, D.G., Leung, S.G., Grundt, K., Dray, E., Idate, R., et al. (2015). NUCKS1 is a novel RAD51AP1 paralog important for homologous recombination and genome stability. *Nucl. acids Res.* 43, 9817-9834.
26. Lim, Y.C., Roberts, T.L., Day, B.W., Stringer, B.W., Kozlov, S., Fazry, S., Bruce, Z.C., Ensbey, K.S., Walker, D.G., Boyd, A.W., et al. (2014). Increased sensitivity to ionizing radiation by targeting the homologous recombination pathway in glioma initiating cells. *Mol. Oncol.* 8, 1603-1615.
27. Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* 154, 442-451.
28. Jirawatnotai, S., Hu, Y., Michowski, W., Elias, J.E., Becks, L., Bienvenu, F., Zagazdzon, A., Goswami, T., Wang, Y.E., Clark, A.B., et al. (2011). A function for cyclin D1 in DNA repair uncovered by protein interactome analyses in human cancers. *Nature* 474, 230-234.
29. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell* 149, 979-993.
30. Jendrzewski, J., He, H., Radomska, H.S., Li, W., Tomsic, J., Liyanarachchi, S., Davuluri, R.V., Nagy, R., and de la Chapelle, A. (2012). The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. U S A* 109, 8646-8651.
31. Pasmant, E., Sabbagh, A., Vidaud, M., and Bieche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444-448.
32. Chung, S., Nakagawa, H., Uemura, M., Piao, L., Ashikawa, K., Hosono, N., Takata, R., Akamatsu, S., Kawaguchi, T., Morizono, T., et al. (2011). Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.* 102, 245-252.
33. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233-1244.

34. Daley, J.M., and Sung, P. (2014). 53BP1, BRCA1, and the choice between recombination and end joining at DNA double-strand breaks. *Mol. Cell. Biol.* **34**, 1380-1388.
35. Zhang, Y., He, Q., Hu, Z., Feng, Y., Fan, L., Tang, Z., Yuan, J., Shan, W., Li, C., Hu, X., et al. (2016). Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat. Struct. Mol. Biol.* **23**, 522-530.

## FIGURE LEGENDS

**Figure 1. Breast cancer risk SNPs alter the promoter activity of novel lncRNAs identified by RNA CaptureSeq. (A)** Gene structures are depicted with exons (blue boxes) joined by introns (lines). The location of *CUPID1* and *CUPID2* are shown by red arrows, PRE1 is shown as a black box and known risk-associated SNPs (rs661204, rs78540526, rs554219, rs657686) are shown as red vertical lines. Data from the UCSC genome browser including epigenetic marks for histone H3 lysine 27 acetylation (H3K27ac) and transcription factor binding from ENCODE are shown. **(B)** Intron-exon structures of *CUPID1* and *CUPID2* transcripts identified by RNA CaptureSeq in MCF7 cells with or without estrogen stimulation. RefSeq lncRNA gene LINC01488 (*CUPID1*) and predicted lncRNA AP000439.3 (*CUPID2*) are shown in green, and *CUPID1/2* transcripts verified by PCR in MCF7 cells are indicated by red arrows (**Figure S2**). Bottom panel shows ChIP-seq data from ENCODE for DNaseI hypersensitivity (indicative of open chromatin), transcription factor binding sites (TFBS) and chromatin state (ChromHMM). **(C)** 3C interaction profiles between PRE1 and the *CUPID1/2* bidirectional promoter in MCF7 and CAL51 cells. 3C libraries were generated with *NcoI*, with the anchor point set at the PRE1. A physical map of the region interrogated by 3C is shown above, with the gray shading representing the position of the *CUPID1/2* promoter. Error bars, SD (n=3). **(D)** Luciferase reporter assays following transient transfection of MCF7 cells. PRE1 (PRE) was cloned downstream of a *CUPID1* promoter-driven (Pr) luciferase (Luc)

construct with and without the 11q13 risk-associated SNPs (rs ID). Error bars, 95% confidence intervals (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (\*\**P*<0.01, \*\*\**P*<0.001). **(E)** 3C followed by sequencing of PRE1 in heterozygous CAL51 breast cancer cells. Chromatograms represent one of two independent 3C libraries generated and sequenced. **(F)** PRE1 was epigenetically silenced in MCF7 cells using two different single guide RNAs (SgPRE1 or SgPRE2). SgCON contains a non-targeting control guide RNA. Gene expression was measured by qPCR and normalized to *TBP* [MIM: 600075] and *GUSB* [MIM: 611499]. Error bars, SEM (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (\*\**P*<0.01, \*\*\**P*<0.001).

**Figure 2. *CUPID1* and *CUPID2* expression and function in breast cell lines and tumors.** **(A)** qPCR for *CUPID1/2* expression normalized against *TBP* in breast normal and cancer cell lines. Error bars, SD (n=2). **(B)** Relative enrichment of RNA in the nucleus relative to the cytoplasm for *CUPID1*, *CUPID2*, *MALAT1* [MIM: 607924], *HOTAIR* [MIM: 611400], *BACTIN* [MIM: 102630] and *GAPDH* [MIM: 138400]. Expression is shown as a ratio of nuclear to cytoplasmic RNA abundance for each replicate with a value of 1 indicating equal distribution across the compartments. Boxes correspond to mean  $\pm$  SEM (n=4). **(C)** qPCR for *CUPID1/2* expression following estrogen stimulation of MCF7 cells. Error bars, SEM (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (\*\*\**P*<0.001). **(D)** *CUPID1/2* expression in breast tumors using TCGA RNAseq data across the histological subtypes (n=708; based on ER, PR and HER2 IHC status). *P*-values were determined by comparing against HR+ groups using Kruskal-Wallis H Test (\*\*\*\**P*<0.0001). **(E)** *CUPID1/2* expression in TCGA

breast tumors stratified by the PAM50 intrinsic molecular subtypes (n=812). *P*-values were determined by one-way ANOVA of the box plots with Tukey outlier test ( $***P<0.001$ ,  $****P<0.0001$ ). **(F)** Overlap of deregulated genes ( $P<0.05$ ) after siRNA-mediated silencing of *CUPID1* or *CUPID2* in MCF7 cells (**Table S5**). **(G)** 91/362 *CUPID* regulated genes showed significant correlation with expression in TCGA breast tumors (n=812) consistent with the direction observed by *CUPID1/2* silencing (**Table S9**). Heatmaps show the average expression of each of the 91 genes in tumour samples from the TCGA dataset with the lowest *CUPID1/2* expression (quartile 1; Q1: bottom 25%), highest expression (quartile 4, Q4: top 25%) and the intermediate quartiles (quartile 2; Q2: 25-50% or quartile 3, Q3: 50-75%). Box plots show the average expression of *CUPID1/2*-induced (top graph) or *CUPID1/2*-suppressed genes (bottom graph) in TCGA tumors stratified by the PAM50 intrinsic molecular subtypes. *P*-values were determined by one-way ANOVA of the box plots with Tukey outlier test ( $****P<0.0001$ ).

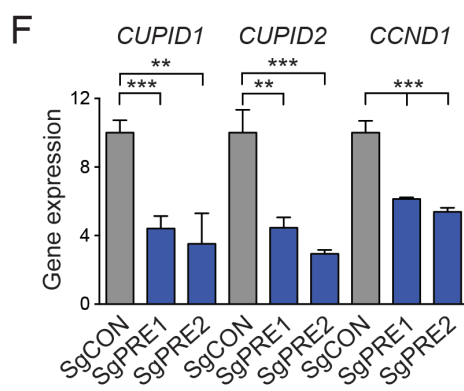
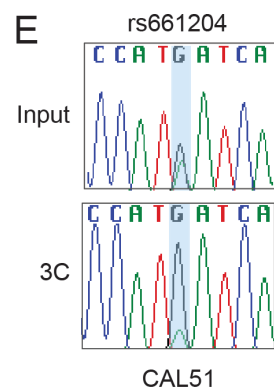
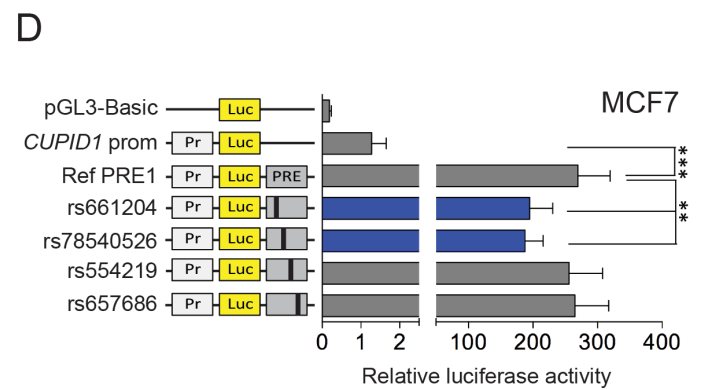
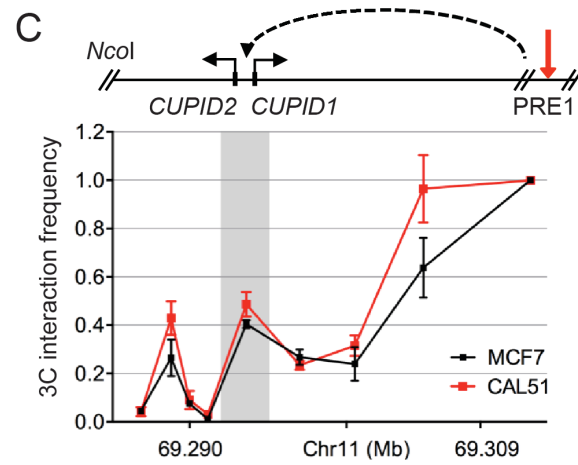
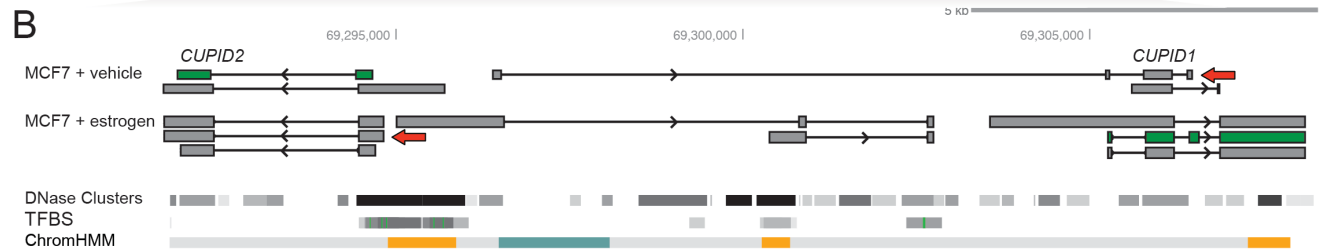
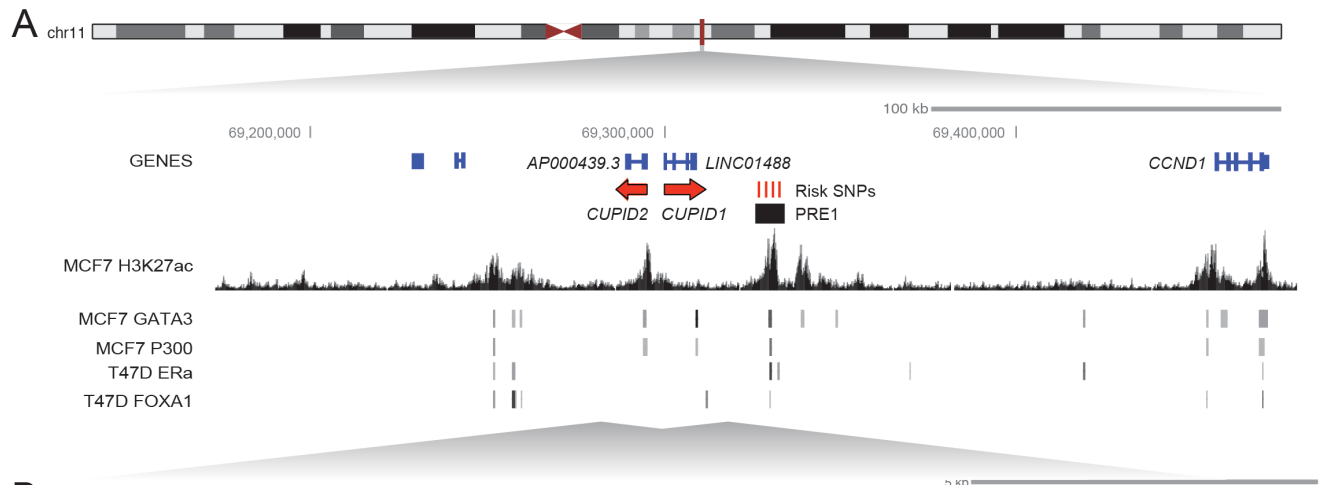
**Figure 3. *CUPID1* and *CUPID2* silencing impairs HRR. (A)** Low *CUPID1/2* expression is associated with a HR mutation signature. TCGA tumors were ranked based on BRCA signature mutations per Mb and those with high and low mutation rates (MR), defined as the top and bottom quartiles, were compared for *CUPID1/2* expression using a two-sided *t*-test. **(B)** TCGA tumors were classified based on the observed patterns of SV distribution (focal, scattered, and mixed). The focal and scattered groups were compared for *CUPID1/2* expression using a two-sided *t*-test. **(C)** MCF7 DR-GFP or **(E, F)** MCF7 cells were transiently transfected with two independent siRNA constructs (A and B) against *CUPID1* or *CUPID2* for 48h. siCON denotes a non-targeting siRNA control, siRAD51 was used as a positive control. **(D)**

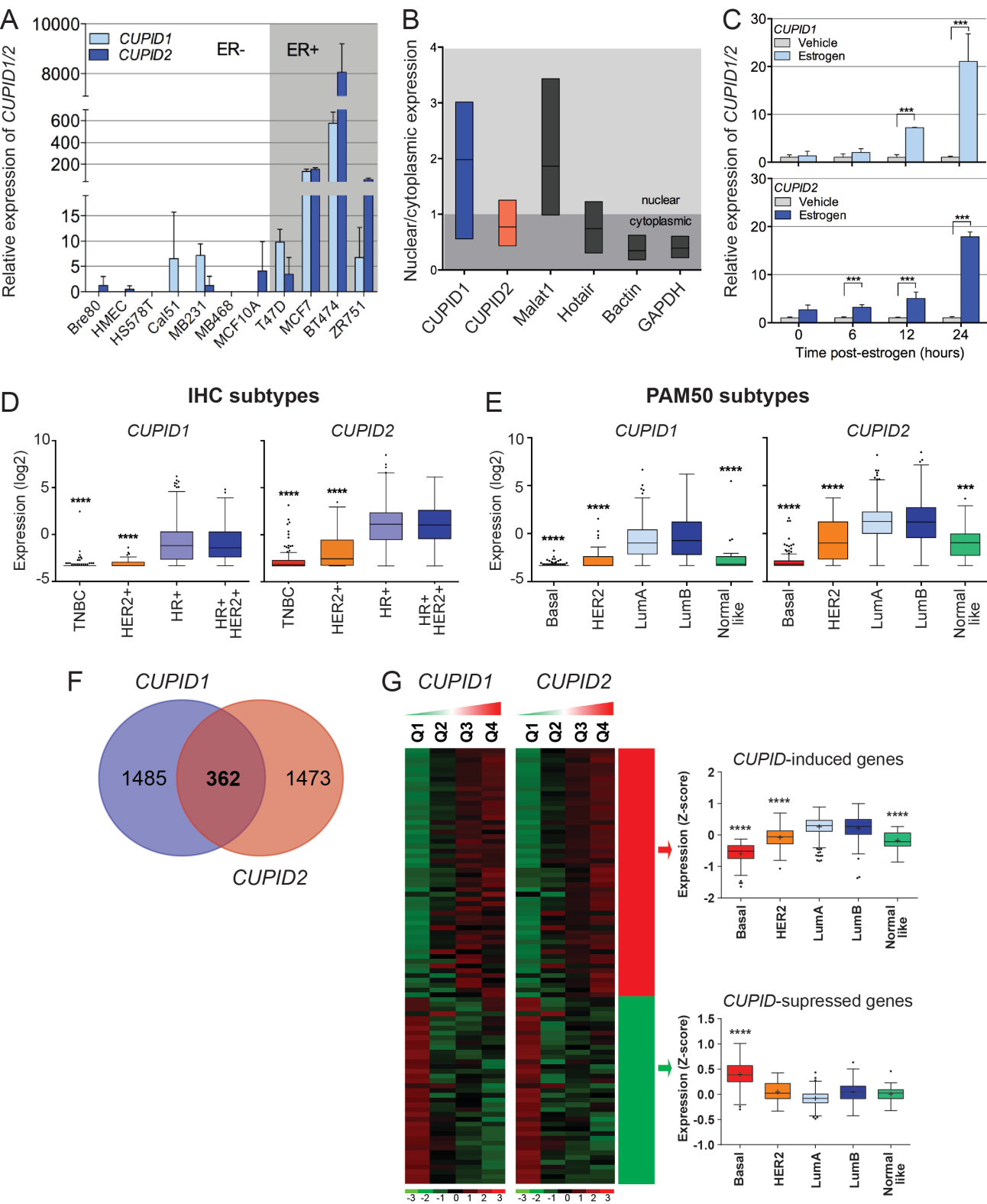
MCF7 DR-GFP or **(G)** MCF7 cells were depleted of *CUPID1/2* by targeting dCAS9-KRAB to the bidirectional promoter (*CUPID1/2*-CRISPRi). *CUPID1* and/or *CUPID2* expression was re-established by cDNA transient transfection. PgRNA denotes a CRISPRi non-targeting control. **(C, D)** HR/DR-GRP reporter assay. For each condition, the percentage of GFP-positive cells in the siCON or vector control was subtracted from the percentage of pCMV-3xNLS-transfected cells. HR% was calculated as the percentage of GFP positive cells in pCMV-ISceI-3xNLS transfected cells divided by GFP positive cells in the corresponding cells transfected with GFP plasmid. Error bars, SEM (n=3). *P*-values were determined by one-way ANOVA followed by Dunnett's multiple comparisons test (\*\* $P<0.001$ , \*\*\*\* $P<0.0001$ ). **(E-G)** IR-induced RAD51 foci formation. Cells were exposed to a single dose of 5Gy IR and RAD51 positive cells quantified by counting nuclei with >5 foci (Scale bar in panel e; 10  $\mu$ m). Error bars, SEM (n=3). *P*-values were determined by one-way ANOVA followed by Dunnett's multiple comparisons test (\* $P<0.05$ ).

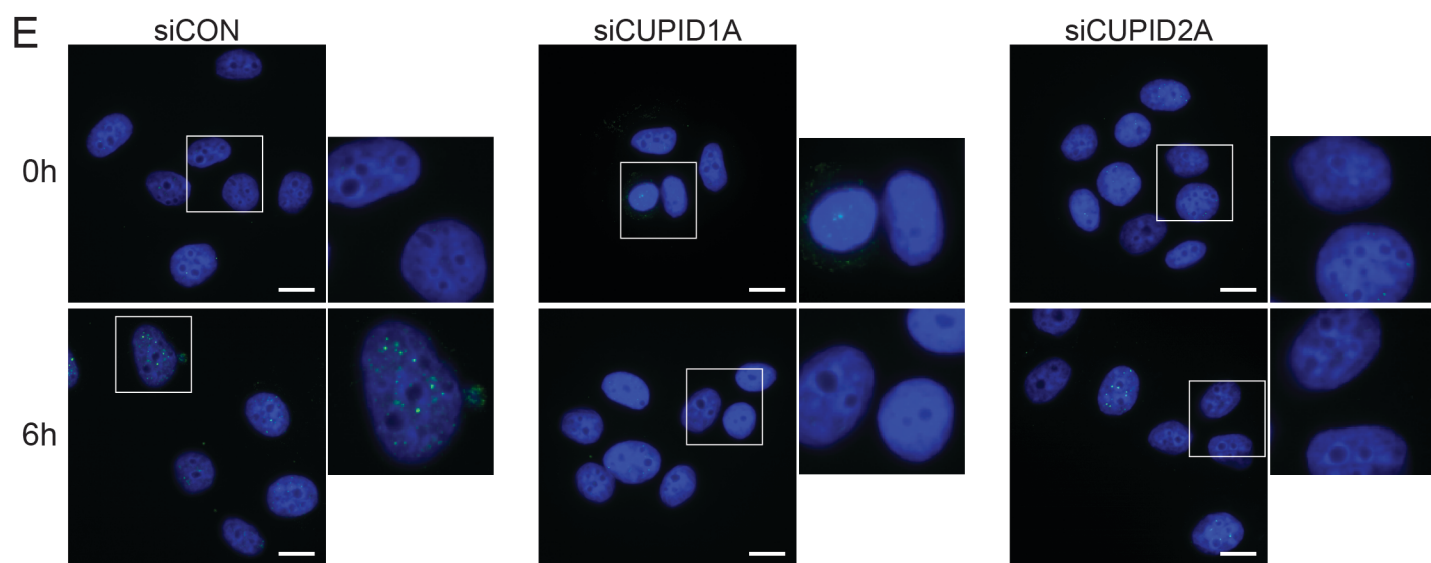
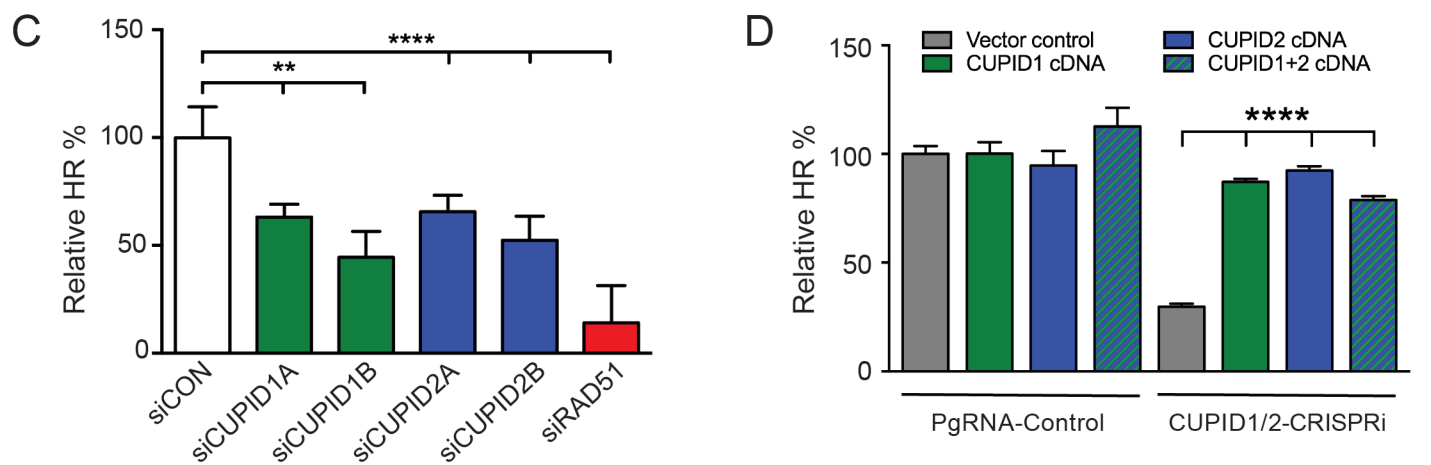
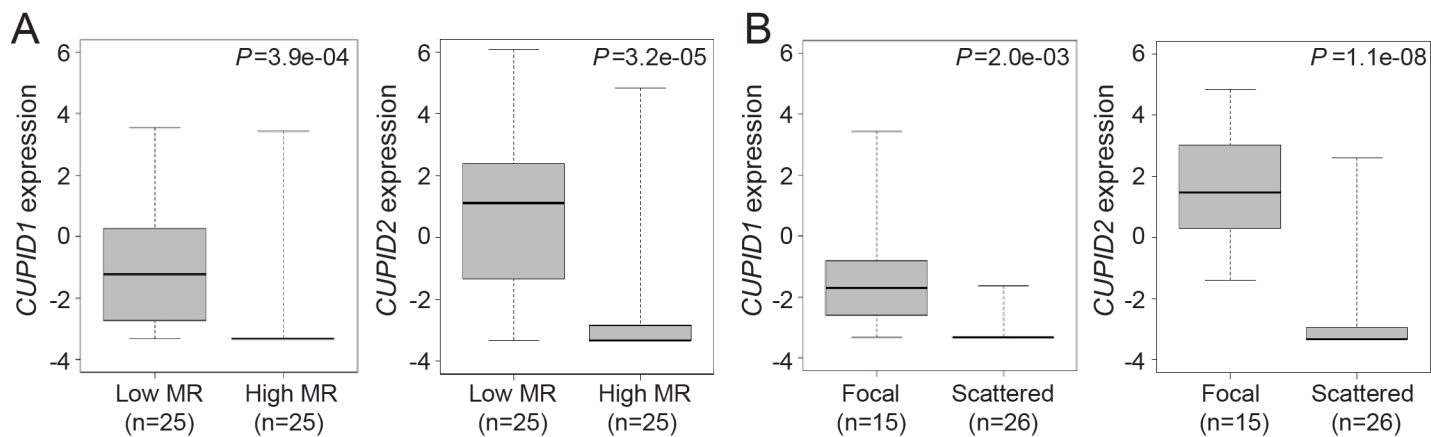
**Figure 4. Knockdown of *CUPID1* and *CUPID2* mediates compensation of DNA DSB repair through NHEJ. (A-D)** Immunofluorescence assays comparing pRPA,  $\gamma$ H2AX and 53BP1 foci formation in control and *CUPID1/2* silenced-MCF7 cells (*CUPID1/2*-CRISPRi) after irradiation (5 Gy). **(A, B)** Representative images of pRPA,  $\gamma$ H2AX and 53BP1 foci at 2h post irradiation (scale bar 20  $\mu$ m). **(C)** Quantification of the number of  $\gamma$ H2AX or 53BP1 foci in 120 nuclei (cells) and the number of pRPA foci in 220 nuclei (cells). Error bars, SEM. *P*-values were determined with one-way ANOVA (\* $P<0.0001$ ). **(D)** Number of foci/cell of pRPA,  $\gamma$ H2AX or 53BP1 after irradiation. Error bars, SEM. **(E)** Percentage of cells with >5 foci/cell with co-localization of  $\gamma$ H2AX and 53BP1 after irradiation. Error bars, SEM. **(F)** *CUPID1/2*

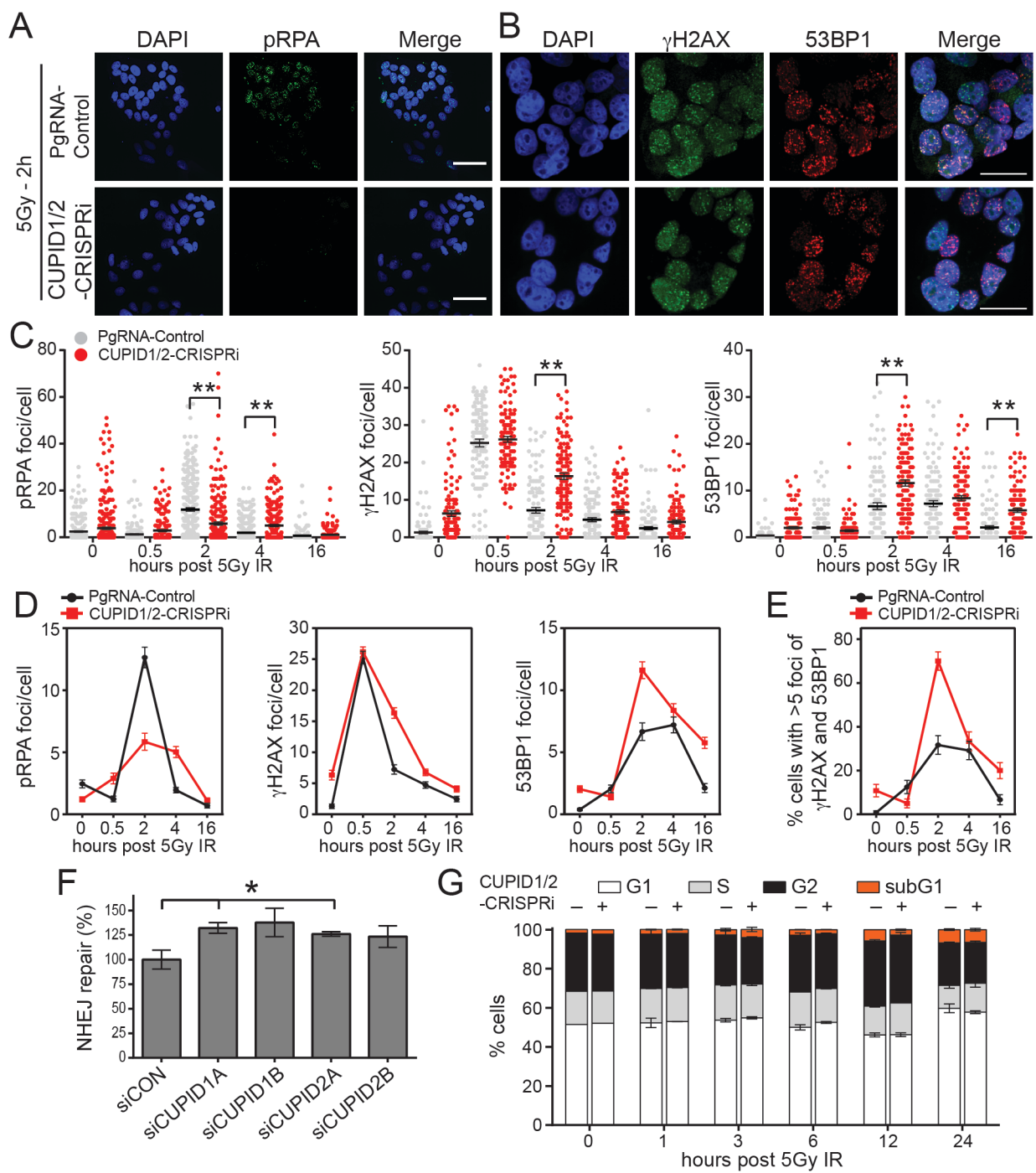
silenced-MCF7 cells were transfected with a linearized plasmid (DNA DSB template) and quantified 48h later for NHEJ activity by the total number of GFP fluorescence cells. Error bars, SD (n=2). *P*-values were determined with a two-way ANOVA (\**P*<0.05). **(G)** Control and *CUPID1/2* silenced-MCF7 cells (CUPID1/2-CRISPRi) were irradiated (5 Gy) then assayed over the specified time point by flow cytometry for 7AAD staining for cell cycle. Error bars, SEM (n=2).

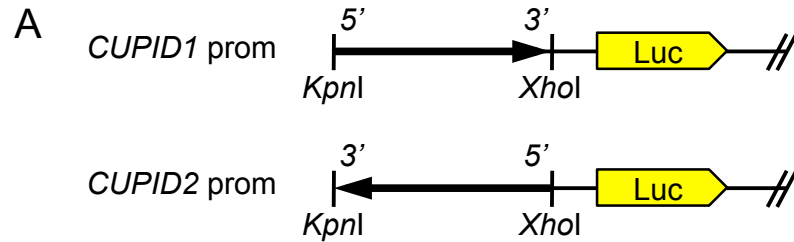










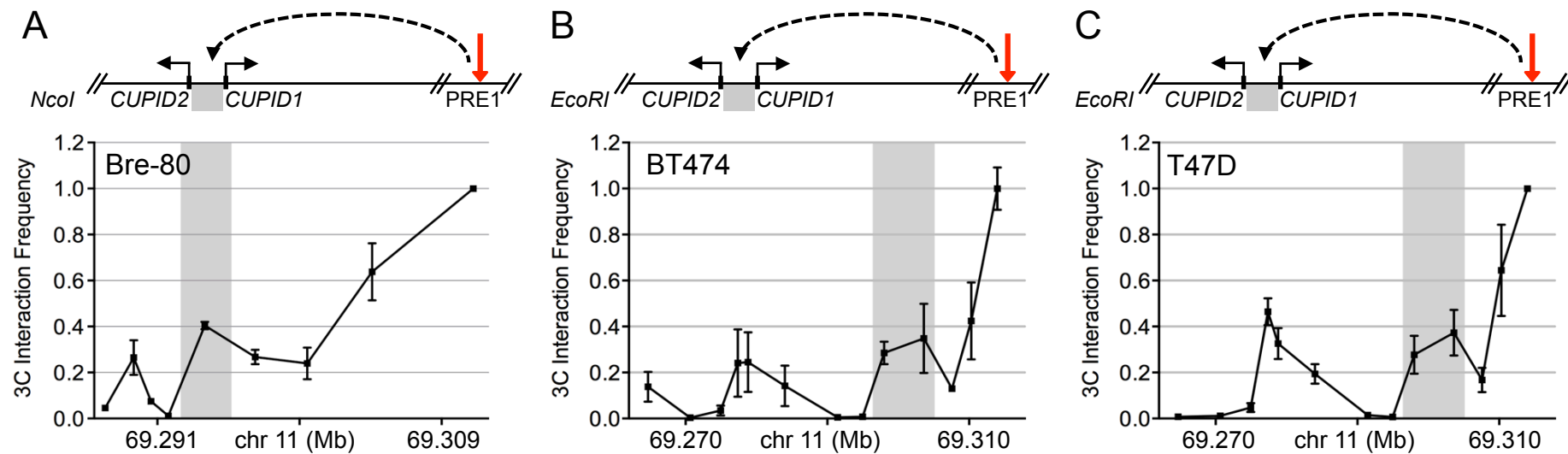


**B** GAGAAAAGAAAAACAAGTGGGTGGTATATATGTGGAAGAACATAAAAAAGACCAGAAGGAACCACATCTCAGTGTGAGCTCTGATGGTCTCTGCTG  
 AGGTTGCCTGCGGTGCCCTAACTCTCTCTGTTTCTGATTGCTTTTGCAATCAACAGGTATTATTCTCATGAGCAGAAAAACGCCCTGTTGTGCACATTTGGGA  
 TGAGAGAAGGGTGCTGCAGGTGAGGAGAAGACAAAGGGGTGGAGTGGGCCAGCTTCTCTAGTGTGGCATCTTCTCCTCTCCTCCTTGCCTGGCTGCCTGGAGCCT  
 GAGGGTCACAGCCTCCAGCCTCTTGGCTTGCACCTCTCGGCCTCTGCACGCAGCCTTATCTTCCCCAGAACGCACACTCACCTTCCGCTTCCCTCAGGACCTGCCA  
 CCTCCAAACCTCTGCCCATGCTGGGTCTCCACTGGGCCACCTCTCCTGCCCAGCTCGGCAGGTGCACATCTTCTAGCCTGGGGGGCCTGCTCAGAAGGCAGC  
 TCTTTCAGGAAGCCCTGGCTGGAGAGACCCACCTTCTGCGCCTCTCTGGCCTCTCTGCTGAGCTGCCCACTCTCTTCCCTTGGTCAAGTTCAGAGTTCAAGGTGTG  
 GCTGTCTTGGCTGTGAGGGCAAAGTCTGCAGCCGACTCGTCTCCCTGGCTGTGAGCTCCACAAGGCGGGCGGGTGCTCAGTAAACATTACTAAACTCATCAGCCC  
 CCTTTGGGCAAACGGGACAGTGGGGCGAGGCCTGGCGTGGCTGGAAAAGCAGTCTGGATGCCCCAGTTGCGGCTGGCTCAACGGCTTTATCTGACAGCCCTGGGC  
 ACCAAGACAAAGAGGCGTTCCCTGGCTGCCTTGGCCTCTCTCTCCAGGGGAAGGCAGGTGACATGGCGAGCCTGGGCCACACAGTGAGCCAGGCCGGAGCC  
 AGCAGCCGGTTCCCAACTGGCTAGGGCCCTTCTGCAGGCCTGACCCTGAGCACATCCAGGACCCCGTCAGCTTCCAGGGAGAGTTCCAGGAGTCTGTGCT  
 GGCGTGGAGAGGAGCTGGGGGCCCTGGGTGTGCGGCTGCAGACGTACGGGCACCCAGAACCTGACCAGAGTCAATGTTGTCGTCCCTTGGCATGCTGGAAA  
 AAAAAATTGCTCAGACAGGGCTGGAGGGCTCTCAGACCCCTTGTATCCTCTTCTCTGAGTCTTGGGAAATTGAGAAGTCTCTGAGAGCTTCCCTTCACCTGATAG  
 TCCACCCCTCAGGATGGAGCTGAGGCAGAGTCTGCACCTCTCAACCTCTTCTCTGTGAAACAGTTAGGAAATGTCTGCAGACATCATCATCCGGACTCAACTCC  
 TCTTCCCTGCTATGAATCGTGAACAAAGTGTTCAGGTCACCACAGTGAAGTCTGAGGACTCTGGCACTGAGCCGGCACTGGGCGGGATCAGTAGCTTGGATGGAATGCCGG  
 CAGGGTCCCTGCTGACCTCATCATGTGGAAGAGTTTCGAGAGAGTTCAGGAGTTGCCAAGGGCCTGCAGACAACGCCGTGTGAGGAGGACAGTGTGAGGTAGCACA  
 GGCCTCTTCGATTCCCAAGGTTCCCTGGAATCCTCAGGATTCCTCAAGGAACCTTGGATCACCCAGCACCACCTAGCCCAGGGACAGAGCCAGGCCTGAGTGCA  
 CAGCCCGCTAGCAACTGTGTAACAGGTGCCAGGCACCTACAGTGTGCTGGGCATGGGGTGAGGTGGGGCTGTGCTGGCAATGCTGCAGACAGCATGTTTACACC  
 AAGCACAGCTACAAGCCCAGGACCAGCTTCCCTCCTTCCGTCTCAGAGCAGCCACTGGGGCAGCCAGTGTACTAGATCCCCACTTGACACATGAGGAGGCACCAGG  
 AGGCTGTGCCACTCAGCCGATCTAGCGGTCCACACATCAGAGTCCATACTCTGCCACCATTAGAATGGGGCTTGTGATGGCTGCCCTCACTCAGTTCAGGGGTCC  
 CCACCAGAATCTCTCAGGCCACACCTGGGACCTCAGATACACTCTGGGCAAGTGGCAGGGGCCCTGAAAGGTCCAGAGTCCAGCCAACCTTCAGGTCTCCCAGGC  
 AAAAGGAGCCAGACCTCCCACTCACCCCTCGAACCAGCAGTCACTCCAACAGGGCTTCTCTGACAGTCCAGGCTGATGCAGCTCACTGTGGCAAAAGACCTTC  
 CCTGAATCAGGTCTCTACATAGAGGCCCATGCAGCCCTCTGGGAATGCGGAGGCTGGGGGCAGCCATAGACACTGGAGCTCCTTCCCTGTGACAAAGTGCCATCCC  
 ATGGGGAGGAGCCAAGCCCAACGAACACCCCAAGGCAAGAGGGAAGAGTCTGTGGGGCAGAGGCAGCCTGAGCATCTGCAAAACAGAGTGGCTCACAGAGAGGA  
 CACCTGCCCTGAGGACACCTGCAGGAGATCCCTACAGAGGAAGGGCTAGACCAGGCTCCATCCAGCAAGTAAGGTCTCTAACAGCTTTGAGATCGTAAGC

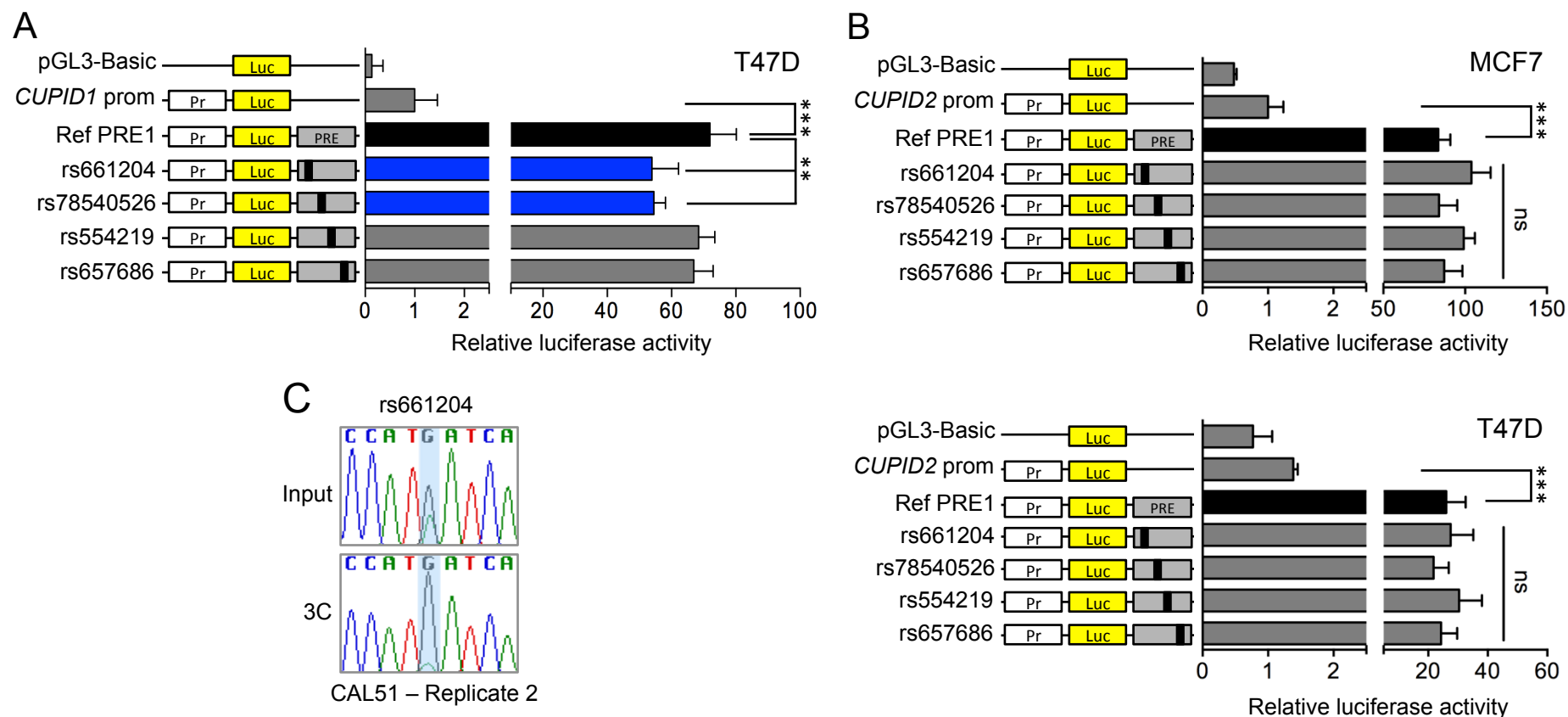
**Figure S1. Schematic and sequence of the *CUPID1/2* bidirectional promoter.** (A) Promoter-driven reporter constructs were generated by inserting a 2515 bp DNA sequence (hg19; chr11:69,294,069-69,296,583) into the *KpnI*-*XhoI* sites of pGL3-Basic the 5'-3' (*CUPID1*) or 3'-5' (*CUPID2*) direction. (B) Sanger sequencing of the *CUPID1/2* bidirectional promoter (5'-3'). Colored text indicates the primers used for PCR amplification.



**Figure S2. PCR validation of *CUPID1* and *CUPID2* transcripts.** (A) PCR amplification of *CUPID1* and *CUPID2* transcripts from MCF7 cDNA. (B) Sanger sequencing of the major transcripts (red arrows). Colored text denotes exon-intron boundaries.



**Figure S3. Chromatin interactions at 11q13 in breast cell lines.** 3C interaction profiles between PRE1 and the *CUPID1/2* bidirectional promoter in **(A)** Bre-80 normal breast cells, **(B)** BT474 or **(c)** T47D breast cancer cells. 3C libraries were generated with *NcoI* or *EcoRI*, with the anchor point set at the PRE1. A physical map of the region interrogated by 3C is shown above, with the grey shading representing the position of the *CUPID1* bi-directional promoter. Error bars, SEM (n=3).



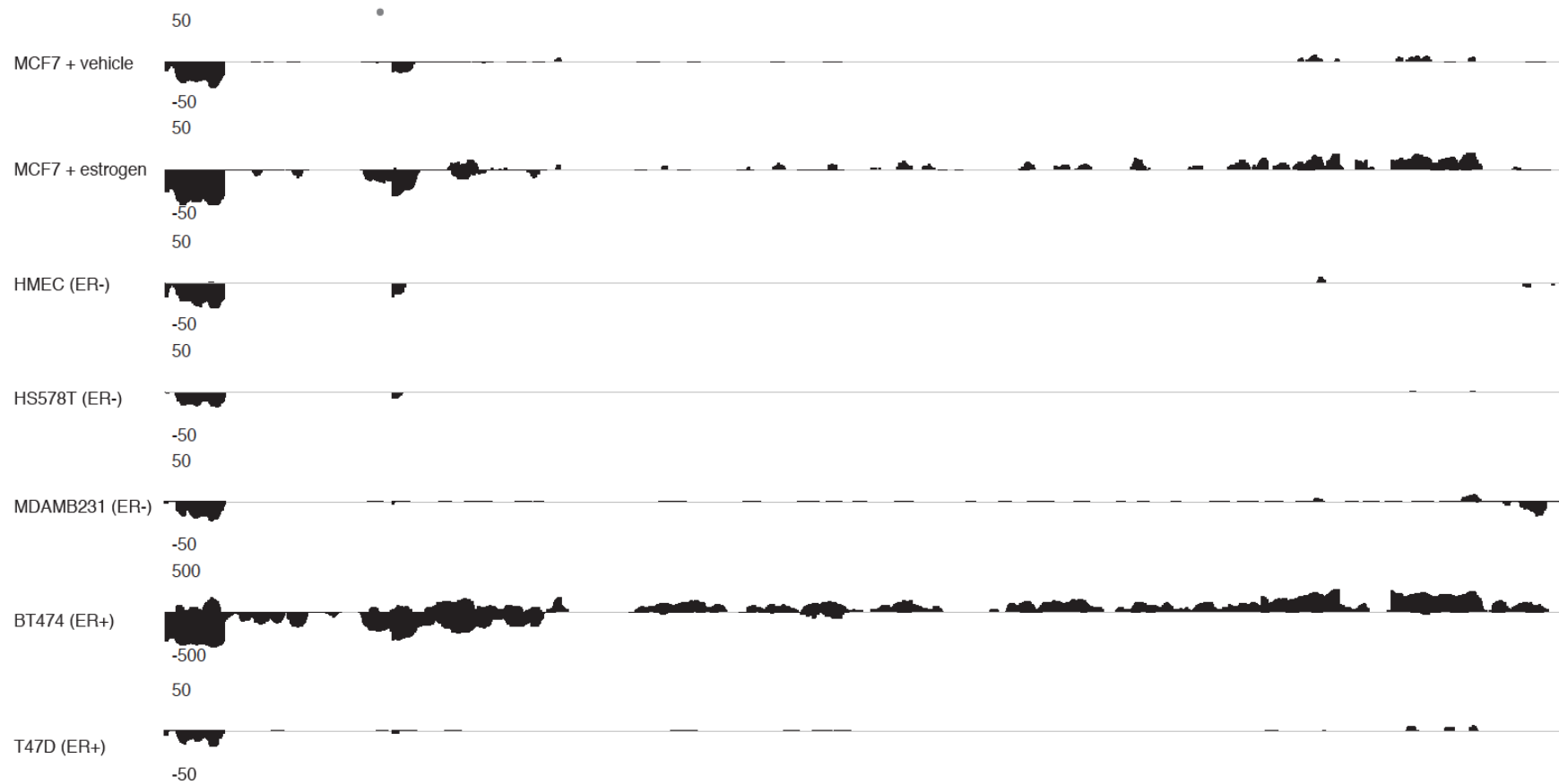
**Figure S4. Luciferase reporter assays and allele-specific 3C in breast cancer cell lines.** PRE1 (PRE) was cloned downstream of (A) *CUPID1* or (B) *CUPID2* promoter-driven (Pr) luciferase (Luc) constructs with and without the 11q13 risk-associated SNPs (rs ID). Error bars, 95% confidence intervals (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (\*\* $P < 0.01$ , \*\*\* $P < 0.001$ , ns=not significant). (C) 3C followed by sequencing for PRE1 in heterozygous CAL51 breast cancer cells. Chromatograms represent one of two independent 3C libraries generated and sequenced.

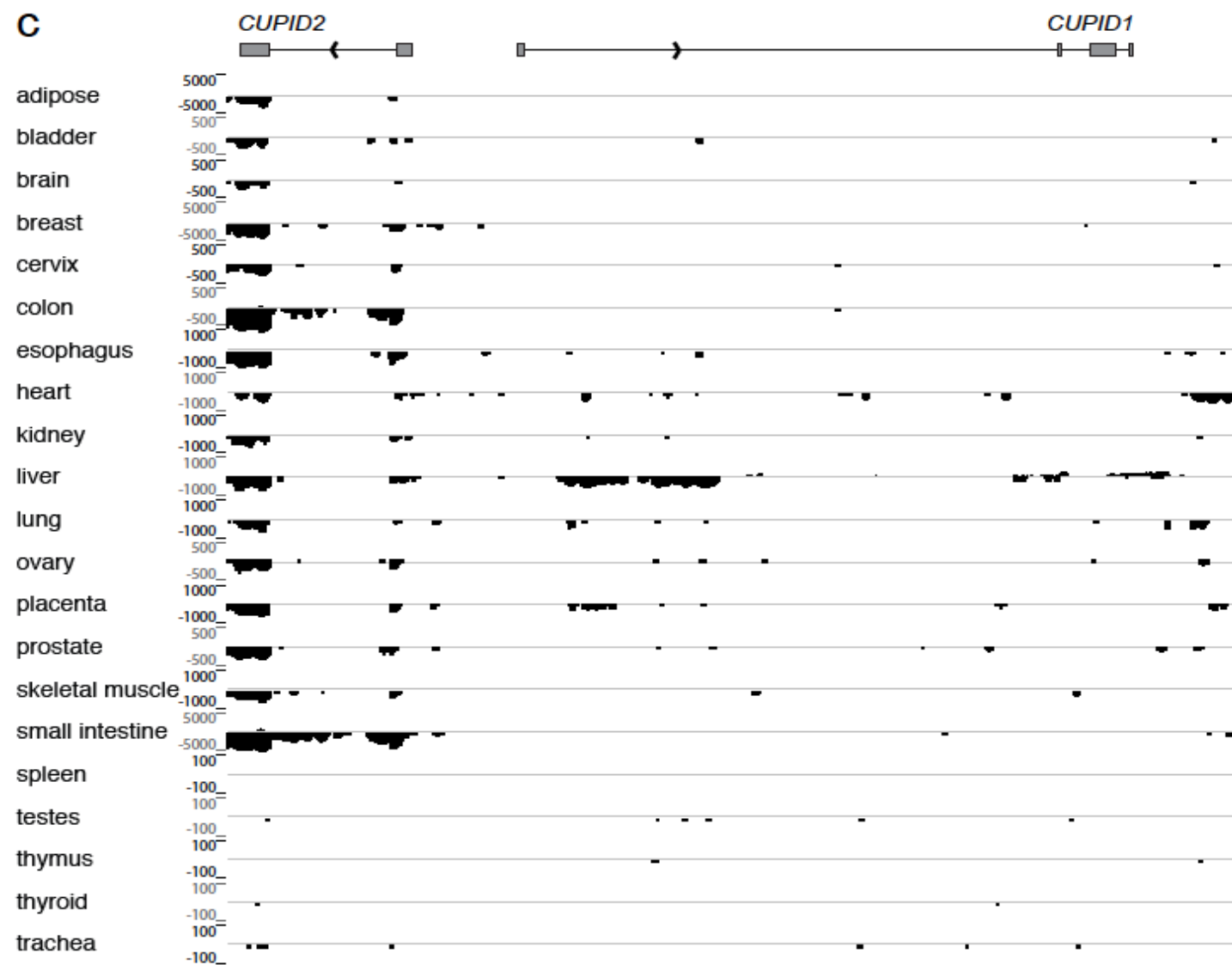


A

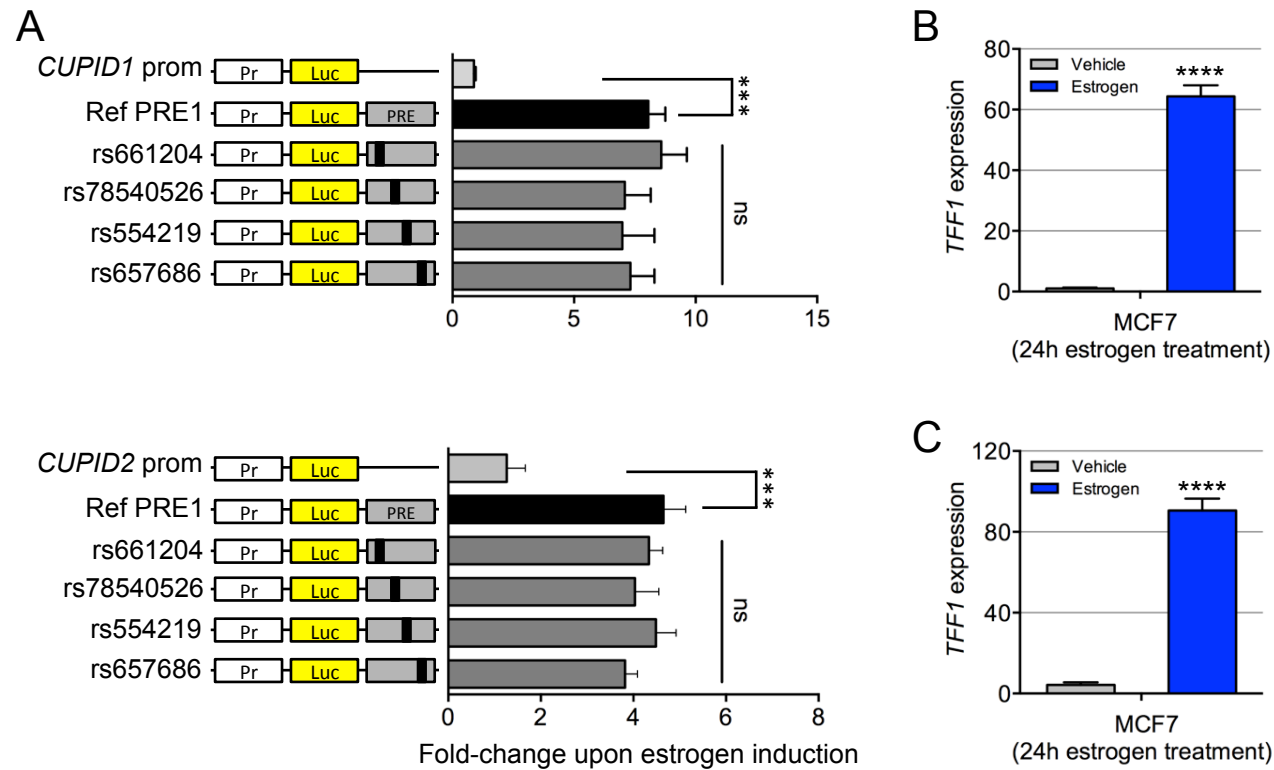


B

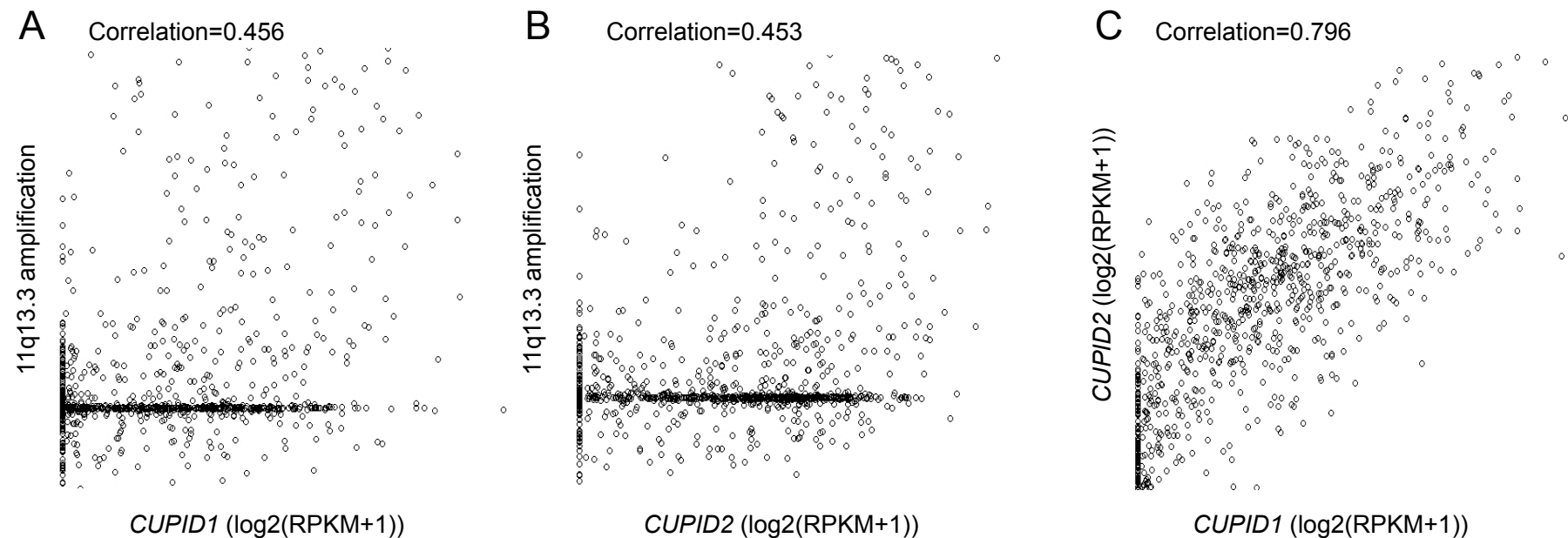




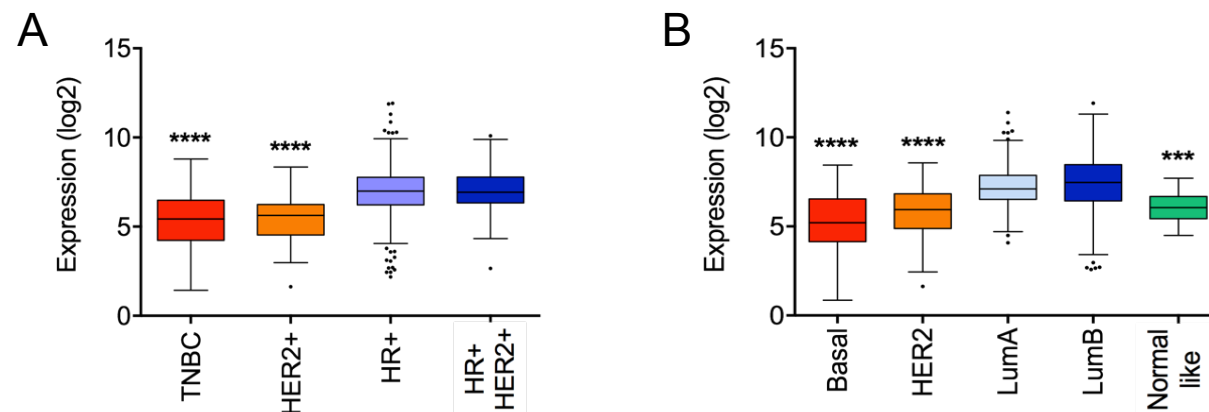
**Figure S5. Captured lncRNAs from the 11q13 breast cancer risk locus across multiple cell lines and tissues. (A)** Intron-exon structures of *CUPID1* and *CUPID2* transcripts identified by RNA Captureseq in MCF7 cells. Histogram of RNA Captureseq reads (log<sub>10</sub> scale) indicating expression of *CUPID1* and *CUPID2* transcripts in **(B)** breast normal and cancer cell lines and **(C)** different tissues.



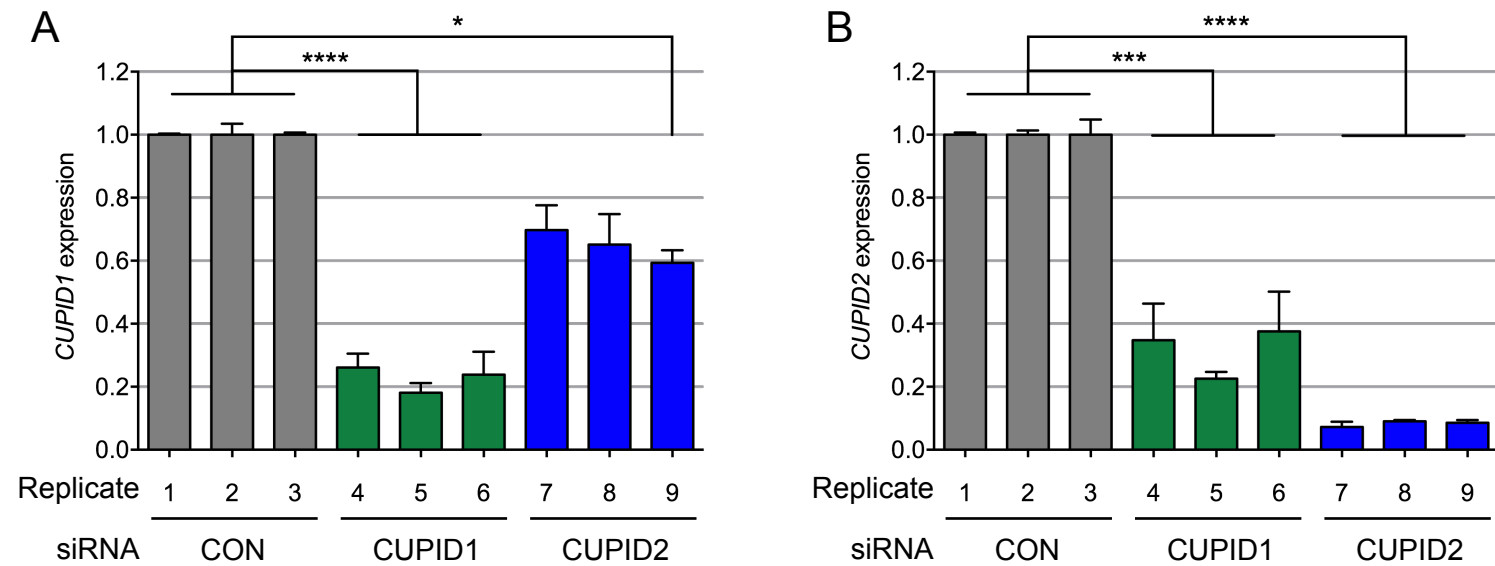
**Figure S6. Luciferase reporter assays and TaqMan qPCR following estrogen induction in MCF7 cells. (A)** PRE1 (PRE) was cloned downstream of a *CUPID1* or *CUPID2* promoter-driven (Pr) luciferase (Luc) constructs with and without the 11q13 risk-associated SNPs (rs ID). Cells were transiently transfected with constructs, pretreated with 10 nM ICI 182780 for 48 hr, then stimulated with estradiol (100 nM) or vehicle for 24 hr. Error bars, 95% confidence intervals (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (\*\*\**P*<0.001, ns=not significant). **(B)** Refers to **Figure S6A** and **(C)** refers to **Figure 2C**; *TFF1* expression was measured by qPCR and normalized to *GUSB*. Error bars, SEM (n=3). *P*-values were determined with a two-tailed *t* test (\*\*\*\**P*<0.0001).



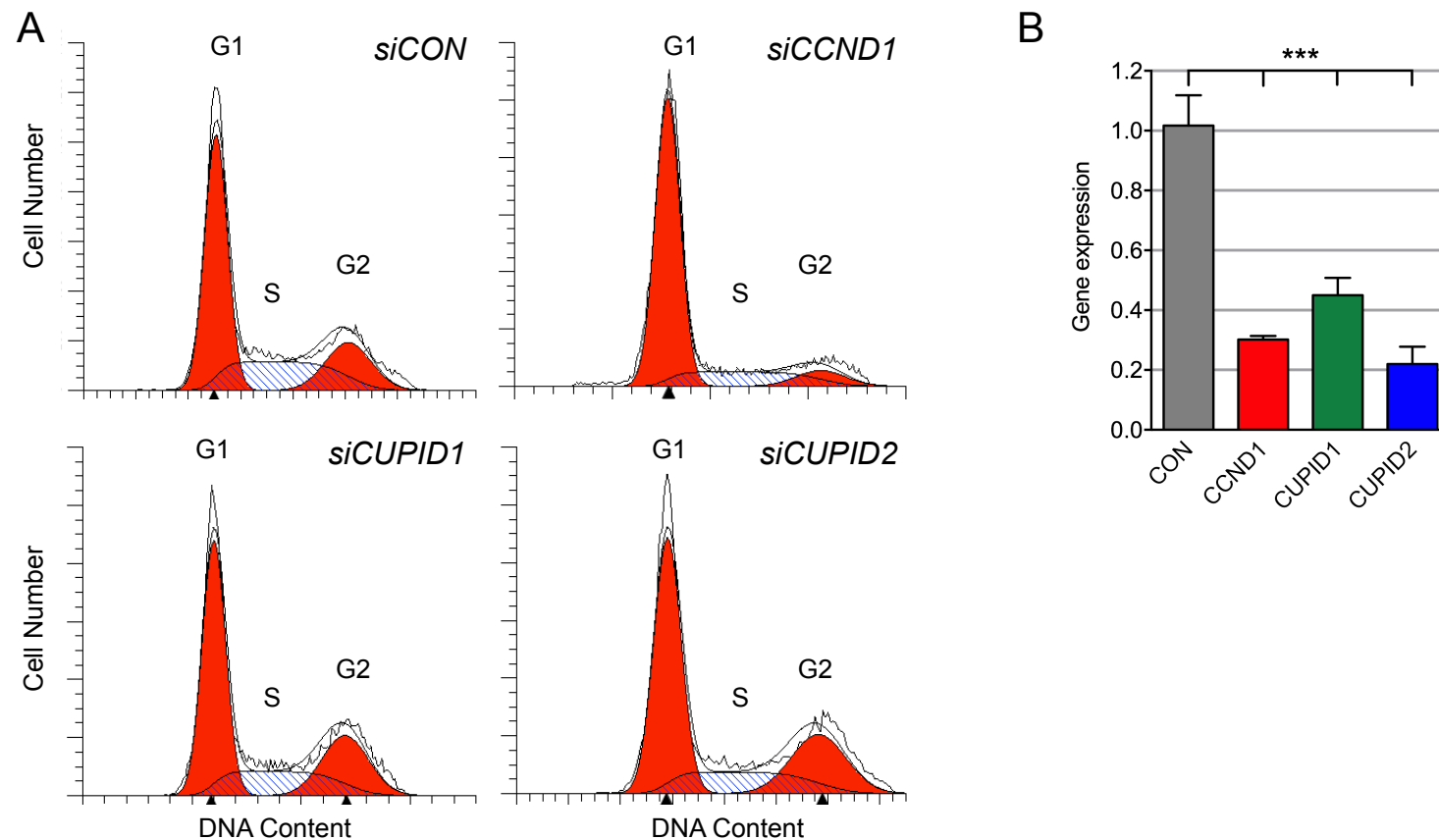
**Figure S7.** (A) Correlation between *CUPID1* expression and 11q13.3 amplification, (B) correlation between *CUPID2* expression and 11q13.3 amplification and (C) correlation between *CUPID1* and *CUPID2* expression in the TCGA cohort (n=1074). Each dot represents a breast cancer individual. Expression values are log transformed (base 2). To avoid infinite values the original RPKM values have been added by a pseudo-count of 1.



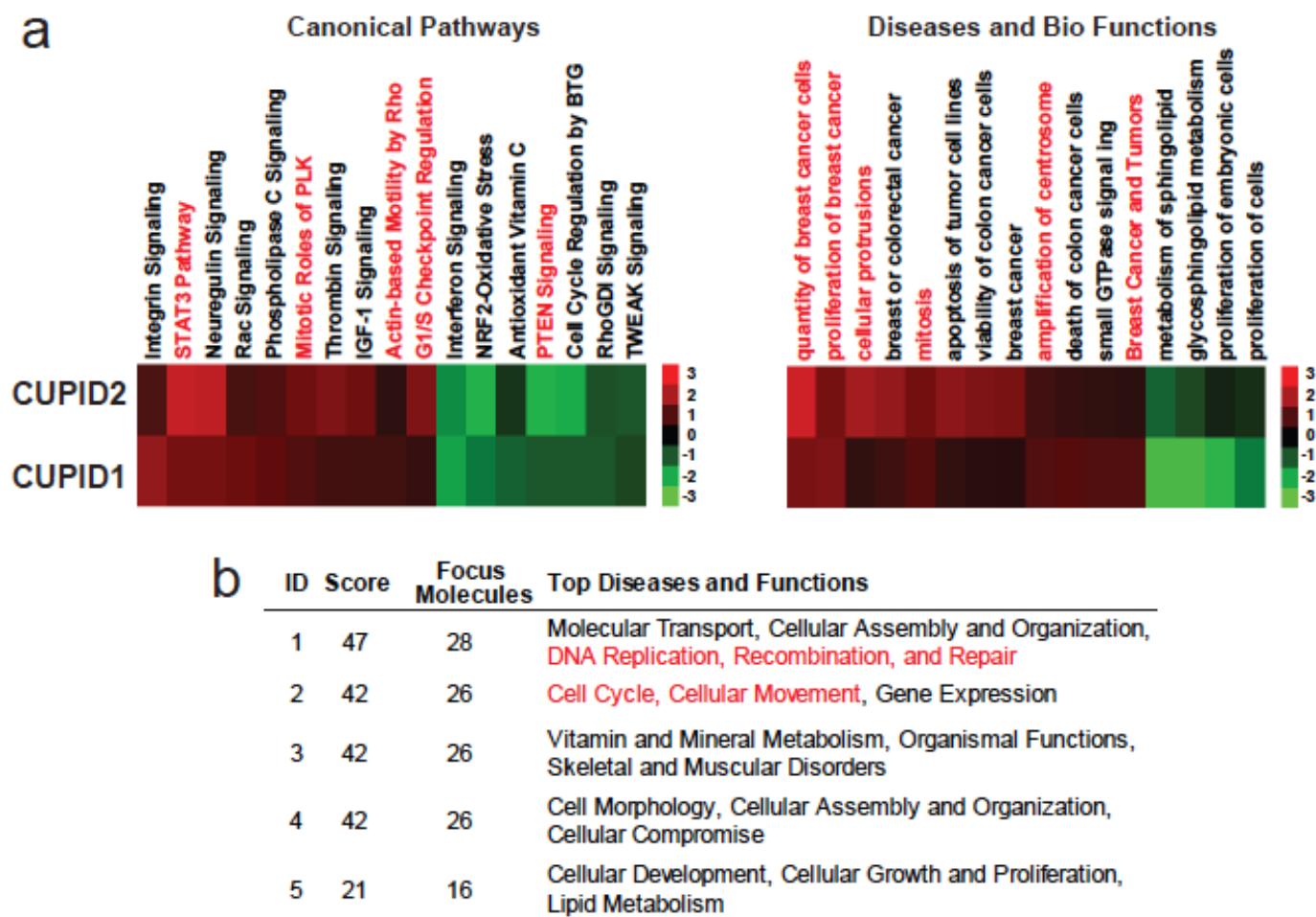
**Figure S8. *CCND1* expression in human breast tumors. (A)** *CCND1* expression in breast tumors using TCGA RNAseq data across the histological subtypes (n=708; based on ER, PR and HER2 IHC status). *P*-values were determined by comparing against HR+ groups using Kruskal-Wallis H Test (\*\*\*\**P*<0.0001). **(B)** *CCND1* expression in TCGA breast tumors stratified by the PAM50 intrinsic molecular subtypes (n=812). *P*-values were determined by one-way ANOVA of the box plots with Tukey outlier test (\*\*\**P*<0.001, \*\*\*\**P*<0.0001).



**Figure S9. RNAi-based knockdown of *CUPID1/2* for RNA-seq.** qPCR confirming expression of **(A) *CUPID1*** or **(B) *CUPID2*** following transient transfection of siRNAs into MCF7 cells. CON denotes transfection with a nontargeting siRNA negative control. Error bars, SEM (n=3). *P*-values were determined by one-way ANOVA followed by Dunnett's multiple comparisons test (\**P*<0.05, \*\*\**P*<0.001, \*\*\*\**P*<0.0001).

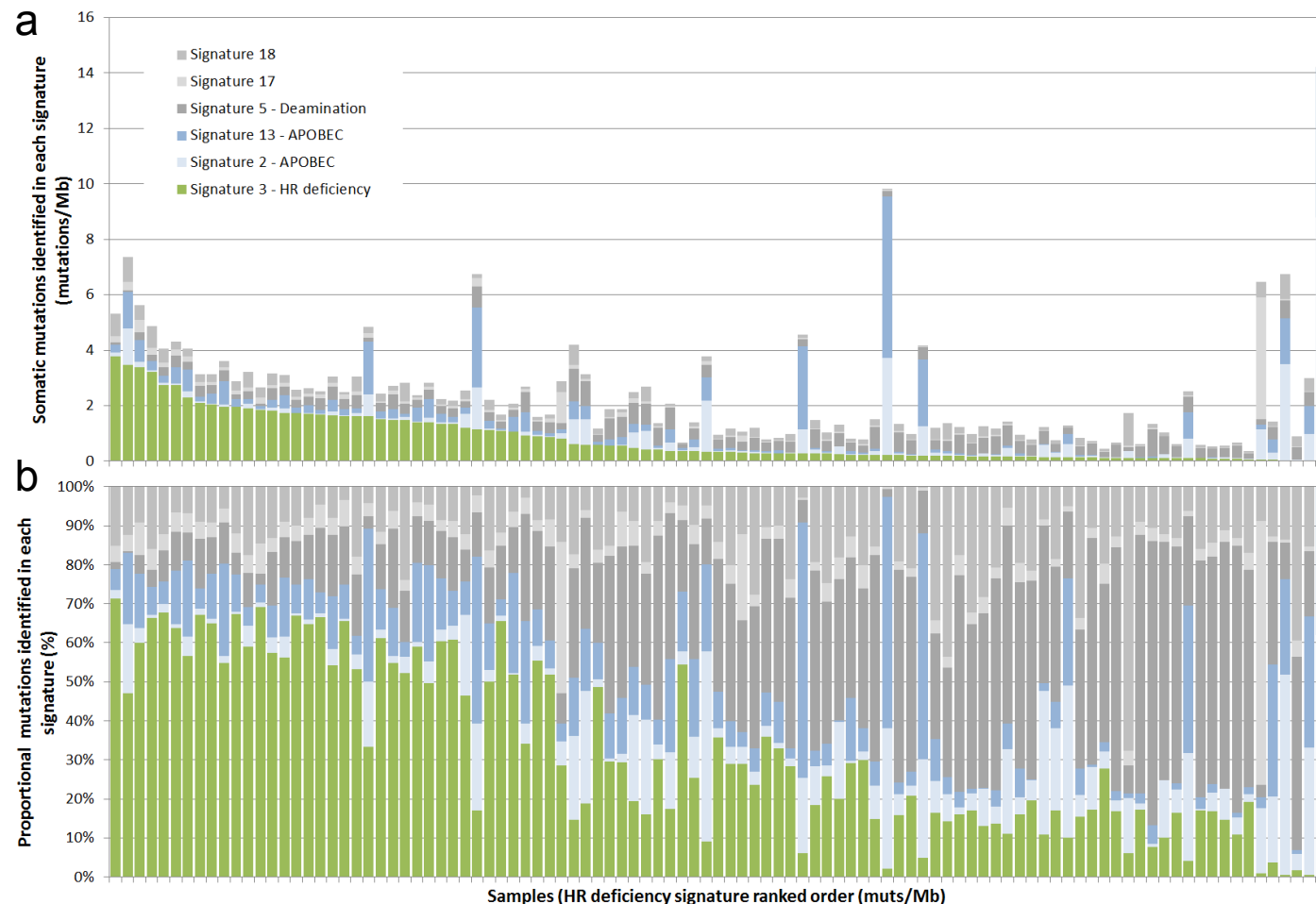


**Figure S10. Effect of RNAi-based knockdown of *CCND1* or *CUPID1/2* on cell cycle progression.** (A) Flow cytometry profiles showing percentage of cells remaining in the G1-phase following siRNA knockdown of *CCND1*, *CUPID1* or *CUPID2* in MCF7 cells. *siCON* denotes transfection with a nontargeting siRNA negative control. (B) qPCR confirming knockdown of *CCND1*, *CUPID1* or *CUPID2*. Error bars, SD (n=2). *P*-values were determined with a two-tailed t test (\*\*\**P*<0.001).

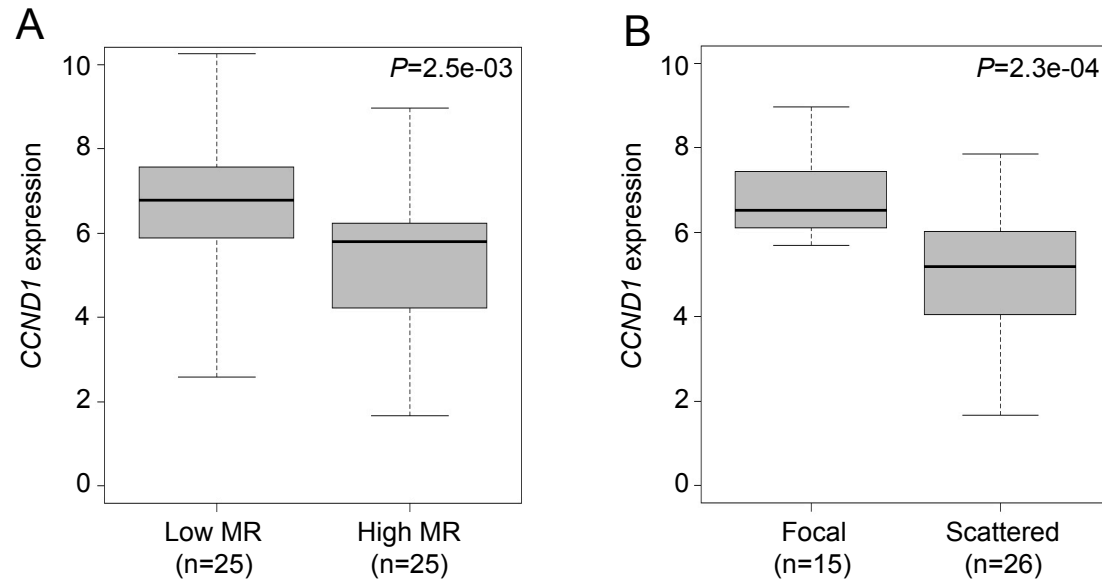


**Figure S11. *CUPID1* and *CUPID2* function.** (A) Deregulated genes after *CUPID1* (1847 annotated genes) or *CUPID2* (1835 genes) siRNA-mediated silencing were subjected to Ingenuity Pathway Analysis (IPA®) and then compared. Heatmaps summarize the top Canonical Pathways (left) and the top Disease/Biological Functions (right). Red denotes high z-score for activation while green denotes low z-score of activation. The complete lists and z-scores are summarized in **Table S6**. (B) The top five networks from IPA® based on the overlapping 362 genes deregulated after *CUPID1* or *CUPID2* silencing (**Table S7** shows the complete list of networks and **Table S8** shows the canonical pathways).

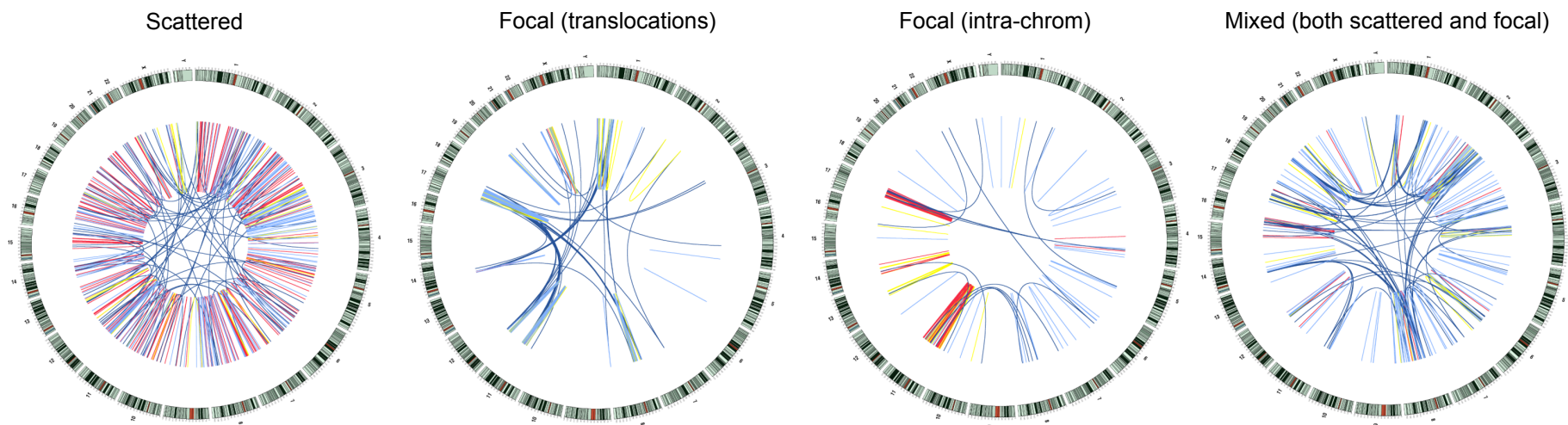




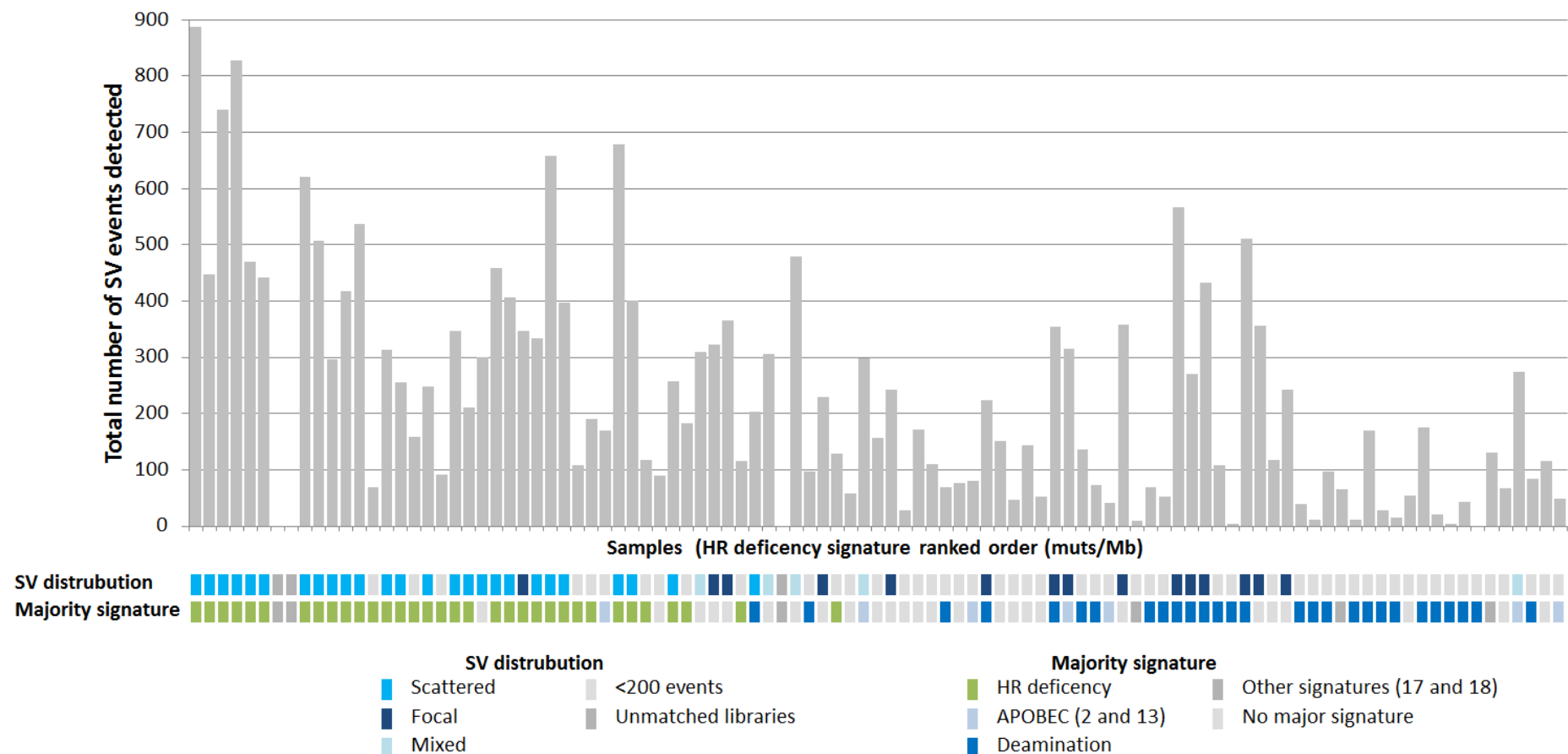
**Figure S12. Mutational signatures in TCGA breast whole genome data.** The tri-nucleotide context of somatic substitution variants were used to identify the mutational signatures from TCGA whole genome sequencing data of 118 breast cancer samples. Six mutation signatures were identified: HR repair deficiency signature (COSMIC: Signature 3); APOBEC (COSMIC: Signatures 2 and 13); Deamination associated with age of onset (COSMIC: Signature 5); and two further signatures of unknown aetiology (COSMIC: Signatures 17 and 18). **(A)** The number of mutations per Mb that contribute to the signatures in each sample and **(B)** the proportion of each signature within the samples is shown. Samples are ordered in both charts from left to right with the number of somatic mutations contributing to the HR deficiency signature descending.



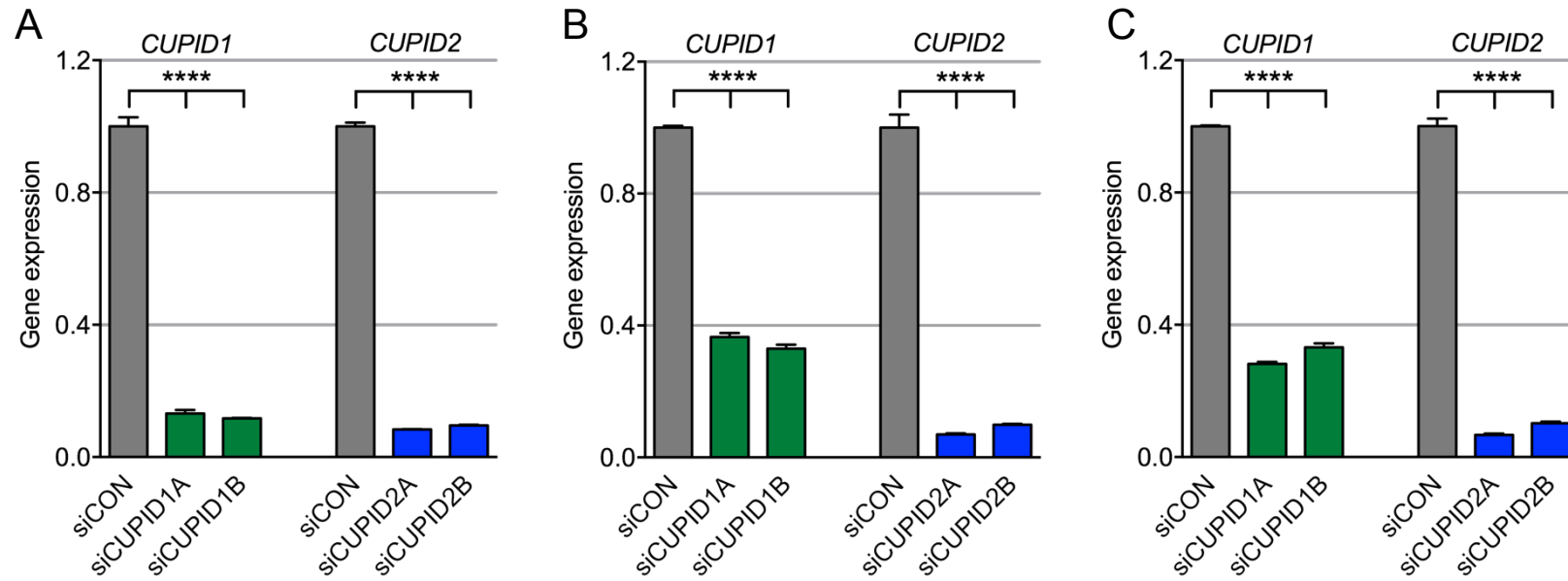
**Figure S13. Low *CCND1* expression is associated with a HR mutation signature. (A)** TCGA tumors were ranked based on BRCA signature mutation per Mb and those with high and low mutation rate (MR) defined as the top and bottom quartiles were compared for gene expression ( $\log_2$  (FPKM+0.1)) using a two-sided t-test. **(B)** TCGA tumors were classified based on the observed patterns of SV distribution (focal, scattered, and mixed). The focal and scattered groups were compared for *CCND1* expression using a two-sided t-test.



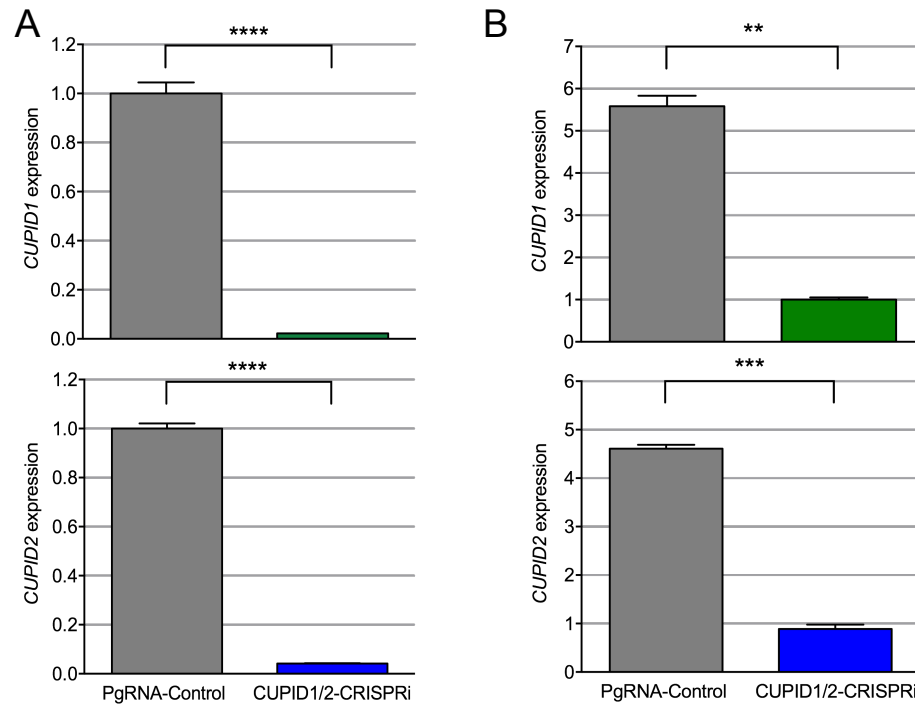
**Figure S14. Distribution patterns of structural variant (SV) breakpoints in TCGA breast cancer whole genome data.** The genomic distribution patterns of the SV breakpoints were used to classify highly rearranged tumors with more than 200 identified SV events as: Scattered, where events are distributed throughout the genome on  $\geq 8$  chromosomes; Focal, where there is a concentration of events on less than 8 chromosomes that commonly involve high numbers of translocations and/or intra-chromosomal events in patterns of complex rearrangement such as chromothripsis or breakage-fusion-bridge; and Mixed, tumors contain characteristics of both scattered and focal events.



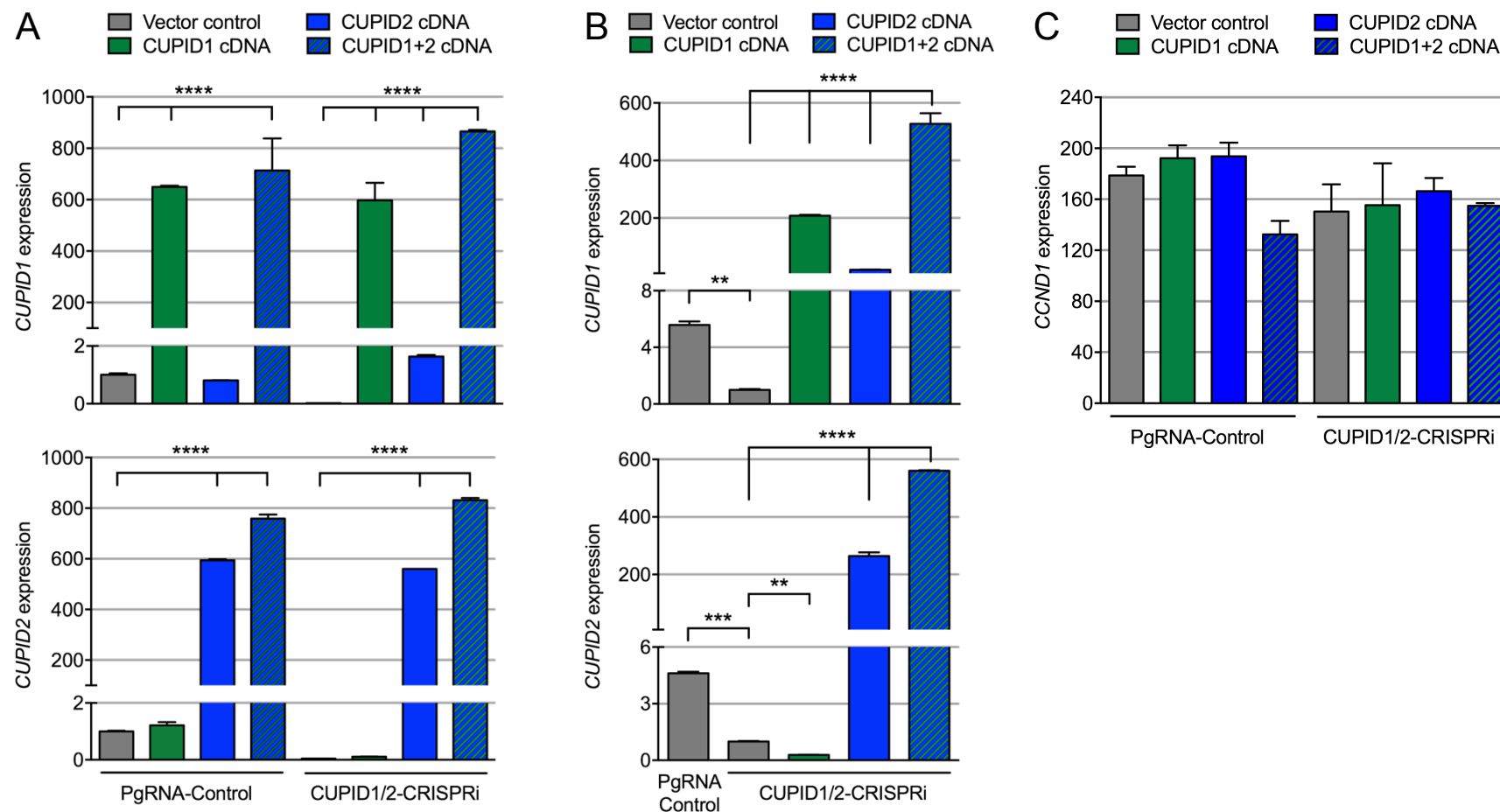
**Figure S15. Summary of structural variant (SV) genomic distribution pattern and the majority mutational signature identified for TCGA breast cancer whole genome data.** Total number of SV events identified is shown in the histogram. For 115 out of 118 TCGA breast cancer samples there was a well-matched library sizes between tumor and normal control samples. The genomic distribution pattern of SV breakpoints was categorized as scattered, focal or mixed for 46 highly rearranged tumors that contained a minimum of 200 SV events. The majority mutational signature (defined as that containing >40% of somatic mutations for that case) is indicated. A majority homologous recombination signature is most commonly identified together with a scattered SV distribution pattern that has previously been reported for HR deficient pancreatic ductal adenocarcinoma and ovarian high-grade serous carcinoma data<sup>19,20</sup>. These corroborative data suggest that there is a defective homologous recombination repair pathway in these samples. Samples are ordered from left to right with the number of somatic mutations contributing to the HR deficiency signature descending.



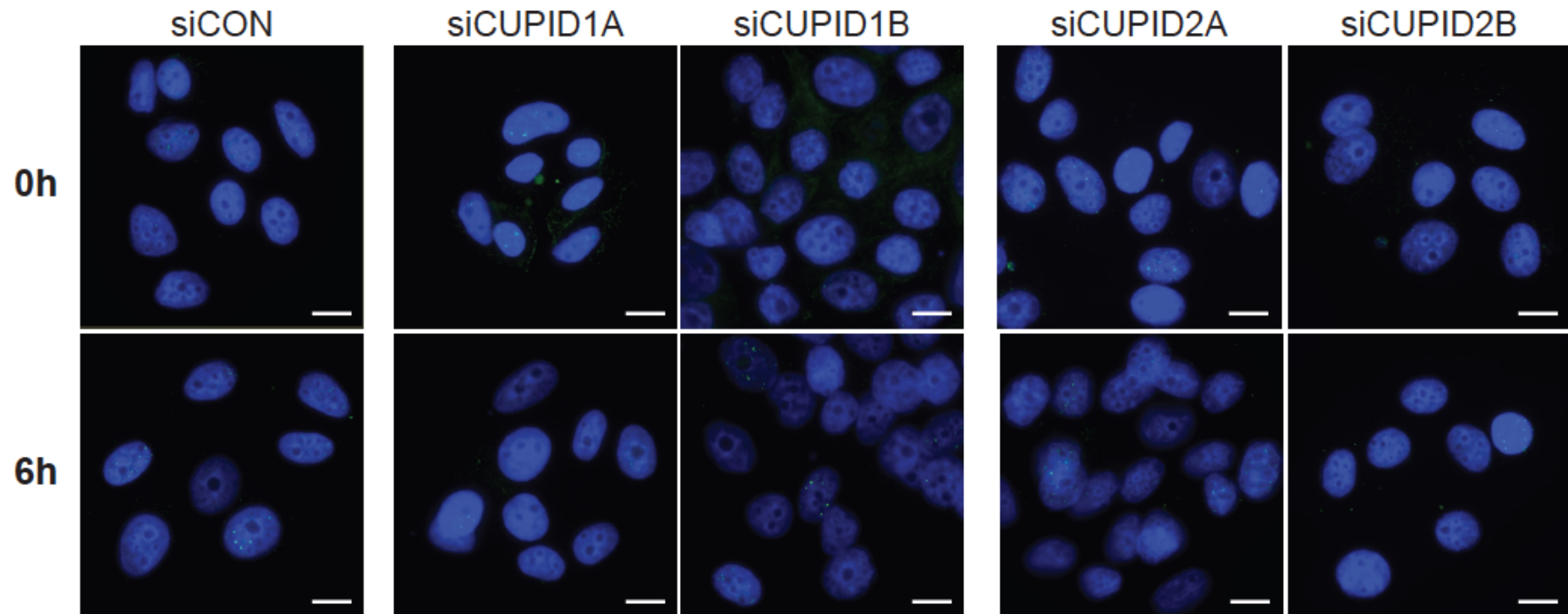
**Figure S16. RNAi-based knockdown of *CUPID1* or *CUPID2* for DNA repair assays.** qPCR confirming knockdown of *CUPID1* or *CUPID2* in MCF7 cells for (A) MCF7 DR-GFP assays (Refers to **Figure 3C**), (B) RAD51 foci assays (Refers to **Figures 3E, 3F** and, **S15**), or (C) cells for NHEJ repair (Refers to **Figure 4F**). siCON denotes transfection with a nontargeting siRNA control. Error bars, SD (n=2). *P*-values were determined by a two-tailed t-test (\*\*\*\**P*<0.0001).



**Figure S17. dCAS9-KRAB repression of *CUPID1* or *CUPID2* for DNA repair assays.** qPCR of *CUPID1/2* in MCF7 cells for **(A)** MCF7 DR-GFP assays (Refers to **Figure 3D**) or **(B)** RAD51 foci assays (Refers to **Figure 3G**). Error bars, SD (n=2). *P*-values were determined by a two-tailed t-test (\*\* $P < 0.01$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ ).

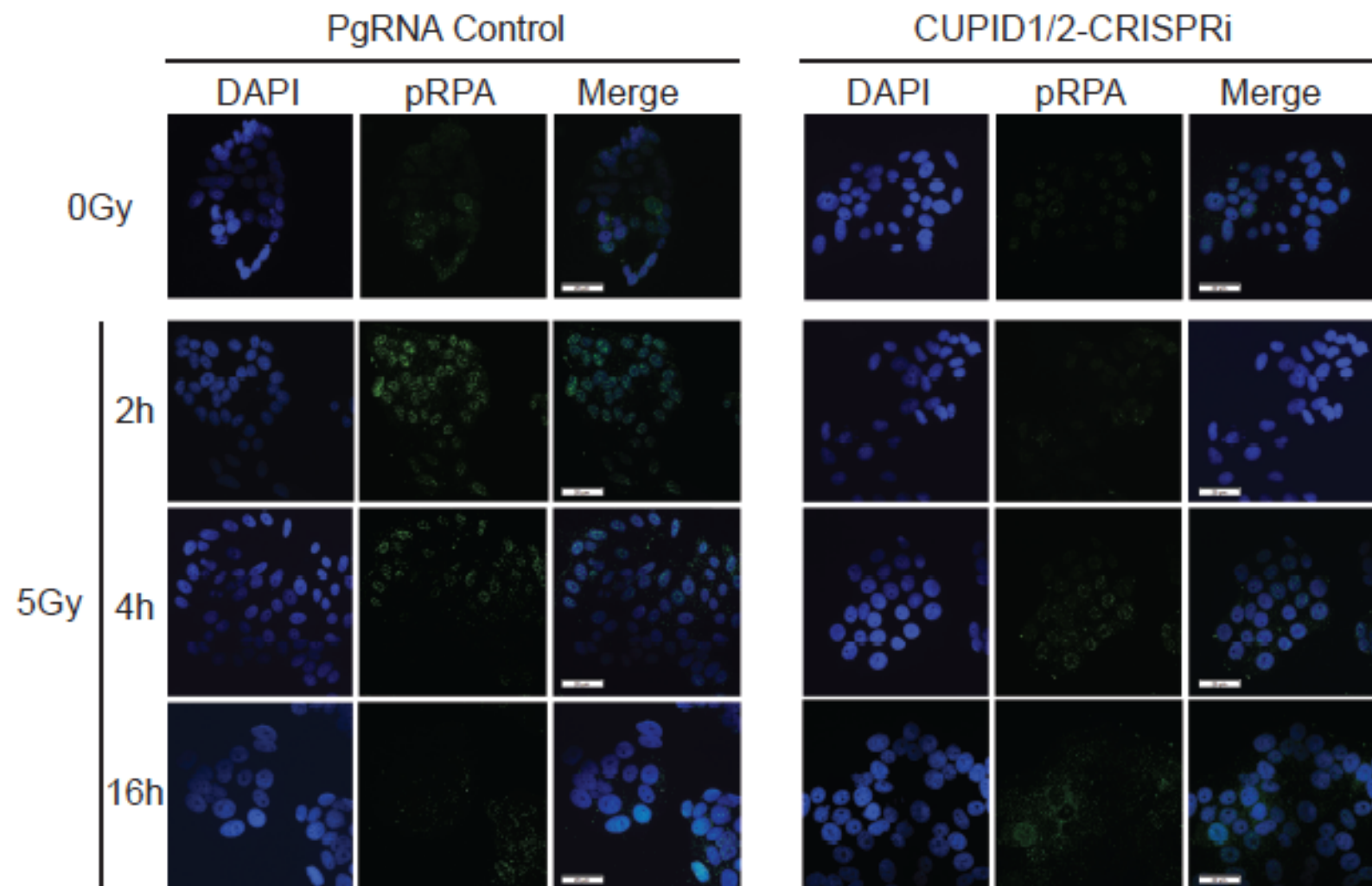


**Figure S18. dCAS9-KRAB repression and re-expression of *CUPID1* or *CUPID2* for DNA repair assays.** qPCR of *CUPID1/2* in MCF7 cells for **(A)** MCF7 DR-GFP assays (Refers to **Figure 3D**) or **(B)** RAD51 foci assays (Refers to **Figure 3G**). **(C)** qPCR of *CCND1* in MCF7 DR-GFP assays. Error bars, SD (n=2). *P*-values were determined by a two-tailed t-test (\*\**P*<0.01, \*\*\**P*<0.001, \*\*\*\**P*<0.0001).

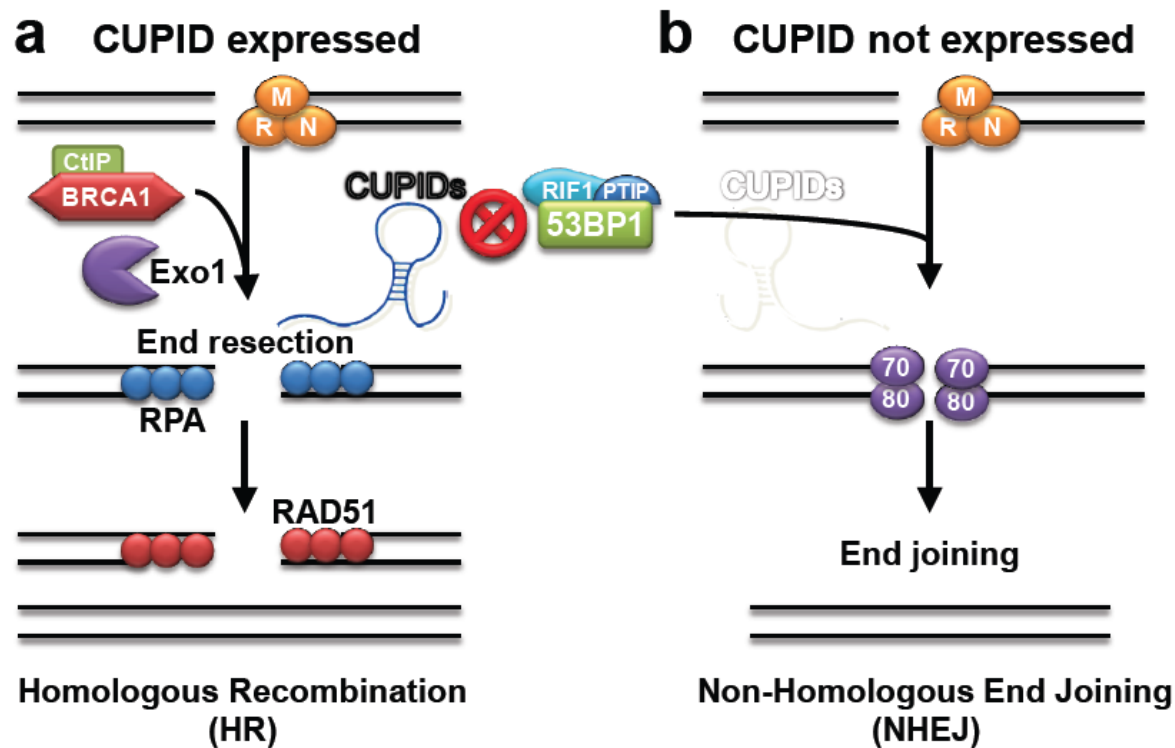


**Figure S19. *CUPID1* and *CUPID2* attenuate *RAD51* activation in HRR.** Refers to **Figures 3E** and **3F**; MCF7 cells were transiently transfected with two independent siRNA constructs (A and B) for *CUPID1* or *CUPID2* for 48h. siCON denotes transfection with a non-targeting siRNA control. Cells were exposed to a single dose of 5Gy IR and *RAD51* positive cells quantified by counting nuclei with >5 foci. Representative images of *RAD51* foci (green) and Hoechst nuclear stain (blue). Scale bar 10  $\mu$ m.





**Figure S20. Repression of *CUPID1/2* mediates compensation of DNA DSB repair through NHEJ.** Immunofluorescence assays for pRPA foci formation in MCF7 cells depleted of *CUPID1/2* by targeting dCAS9-KRAB to the bidirectional promoter (CUPID1/2-CRISPRi) after irradiation (5 Gy). PgRNA denotes a lenti CRISPR non-targeting control. Representative images of pRPA foci for 2, 4 and 16 hours post irradiation (scale bar 20  $\mu$ m).



**Figure S21. Proposed model for the role of *CUPIDs* in DSBs repair.** DNA double strand breaks (DSBs) coated by the MRE11, RAD50, NBS1 (MRN) complex initiates downstream DNA repair signaling and pathways. **(A)** The homologous recombination (HR) DNA repair pathway is mediated by BRCA/CtIP to start DNA end resection by Exo1 and generate ssDNA that is coated by RPA and finally followed by RAD51 before completion of homologous recombination after strand invasion. **(B)** In the absence of *CUPID* expression, RPA coating was significantly reduced leading to inhibition of RAD51 recruitment to DSBs and reduction in HR DNA repair. Instead, 53BP1 recruitment was significantly increased and NHEJ remained active, or mildly elevated, in order to repair the DSBs.