


RESEARCH

Open Access



# Chromatin interactome mapping at 139 independent breast cancer risk signals

Jonathan Beesley<sup>1†</sup>, Haran Sivakumaran<sup>1†</sup>, Mahdi Moradi Marjaneh<sup>1,2†</sup>, Luize G. Lima<sup>1</sup>, Kristine M. Hillman<sup>1</sup>, Susanne Kaufmann<sup>1</sup>, Natasha Tuano<sup>3</sup>, Nehal Hussein<sup>1,4</sup>, Sunyoung Ham<sup>1</sup>, Pamela Mukhopadhyay<sup>1</sup>, Stephen Kazakoff<sup>1</sup>, Jason S. Lee<sup>1</sup>, Kyriaki Michailidou<sup>5,6</sup>, Daniel R. Barnes<sup>5</sup>, Antonis C. Antoniou<sup>5</sup>, Laura Fachal<sup>7</sup>, Alison M. Dunning<sup>7</sup>, Douglas F. Easton<sup>5,7</sup>, Nicola Waddell<sup>1</sup>, Joseph Rosenbluh<sup>3</sup>, Andreas Möller<sup>1</sup>, Georgia Chenevix-Trench<sup>1</sup>, Juliet D. French<sup>1\*†</sup> and Stacey L. Edwards<sup>1\*†</sup> 

## Abstract

**Background:** Genome-wide association studies have identified 196 high confidence independent signals associated with breast cancer susceptibility. Variants within these signals frequently fall in distal regulatory DNA elements that control gene expression.

**Results:** We designed a Capture Hi-C array to enrich for chromatin interactions between the credible causal variants and target genes in six human mammary epithelial and breast cancer cell lines. We show that interacting regions are enriched for open chromatin, histone marks for active enhancers, and transcription factors relevant to breast biology. We exploit this comprehensive resource to identify candidate target genes at 139 independent breast cancer risk signals and explore the functional mechanism underlying altered risk at the 12q24 risk region.

**Conclusions:** Our results demonstrate the power of combining genetics, computational genomics, and molecular studies to rationalize the identification of key variants and candidate target genes at breast cancer GWAS signals.

## Introduction

Breast cancer is known to have an important inherited component. While rare coding mutations in susceptibility genes such as *BRCA1*, *BRCA2*, and *PALB2* confer a high risk of breast cancer, these account for less than one quarter of the familial risk [1]. Much of the remaining heritability is due to the combination of a large number of common, low-penetrance variants [2, 3]. Genome-wide association studies (GWAS) have been a powerful tool to identify disease-associated genetic variants, but these studies do not directly address the underlying biological mechanisms. A combination of fine-scale mapping and bioinformatic and functional studies are required to establish this link [4]. The Breast Cancer Association Consortium (BCAC) and the

Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) have recently performed large-scale genetic fine-mapping of 150 breast cancer susceptibility regions in ~217,000 breast cancer cases and controls of European ancestry [5]. Step-wise multinomial logistic regression analysis identified 196 high confidence independent risk signals, defined as having association  $p$  values  $< 10^{-6}$  after adjusting for other variants. Fachal et al. [5] used these data to define sets of credible causal variants (CCVs) for each signal, defined as variants with  $p$  values within 2 orders of magnitude of the top variant.

The majority of CCVs mapped to non-protein-coding regions of the genome and are enriched at regulatory DNA elements such as enhancers, silencers, and insulators [2, 5]. It is established that many regulatory elements are located long distances from their target gene promoters and that regulation of transcription involves direct physical interactions brought about by chromatin looping [6]. Importantly, individual enhancers often loop to and regulate multiple genes, including protein-coding and non-coding RNA genes. Adding to the complexity,

\* Correspondence: [Juliet.French@qimrberghofer.edu.au](mailto:Juliet.French@qimrberghofer.edu.au); [Stacey.Edwards@qimrberghofer.edu.au](mailto:Stacey.Edwards@qimrberghofer.edu.au)

<sup>†</sup>Jonathan Beesley, Haran Sivakumaran, and Mahdi Moradi Marjaneh contributed equally.

Juliet D. French and Stacey L. Edwards jointly directed this work.

<sup>1</sup>Cancer Program, QIMR Berghofer Medical Research Institute, Brisbane, Australia

Full list of author information is available at the end of the article



enhancers do not necessarily act on the closest promoter but can bypass neighboring genes to regulate genes located more distally. There is also considerable evidence that most enhancer-promoter interactions occur in *cis* and within chromatin structures called topologically associating domains (TADs) [7]. TADs are typically several hundred kilobases to a few megabases in size and are relatively stable between cell types and in response to extracellular signals [8, 9].

Various chromatin conformation capture (3C)-based methods have been developed to map chromatin contacts at a genome-wide level. The basic principle of 3C involves chromatin fragmentation of formaldehyde-fixed nuclei (usually by restriction digestion), followed by ligation of linked DNA fragments, then detection and quantification of ligation products [10]. One of these methods, Hi-C, is an unbiased but relatively low-resolution approach that quantifies interactions between all possible DNA fragment pairs in the genome [11]. Hi-C has been used extensively to analyze the three-dimensional organization of genomes, including the compartmentalization of chromatin and the position of TADs [12, 13]. To increase Hi-C resolution, several groups have developed sequence capture to enrich for chromosomal interactions involving targeted regions of interest [14–17]. There are several capture methodologies, but typically, RNA or DNA oligonucleotide baits are directed to the ends of targeted DNA fragments to enrich for ligation events prior to next-generation sequencing [18, 19]. Promoter Capture Hi-C (PCHi-C) is the most widely used approach where baits are designed to annotated promoters, resulting in a strong enrichment for promoter-anchored interactions [15–17, 20]. A few post-GWAS studies have also used Region Capture Hi-C, in which baits target linkage disequilibrium blocks or restriction fragments containing genetic variants associated with the disease of interest [21, 22].

Here, we applied Variant Capture Hi-C (VChi-C) and PCHi-C to normal breast and breast cancer cell lines to generate a catalog of interactomes. We report several hundred candidate target genes in breast cancer risk regions including some known cancer driver genes but also many molecular targets not previously implicated in breast cancer etiology.

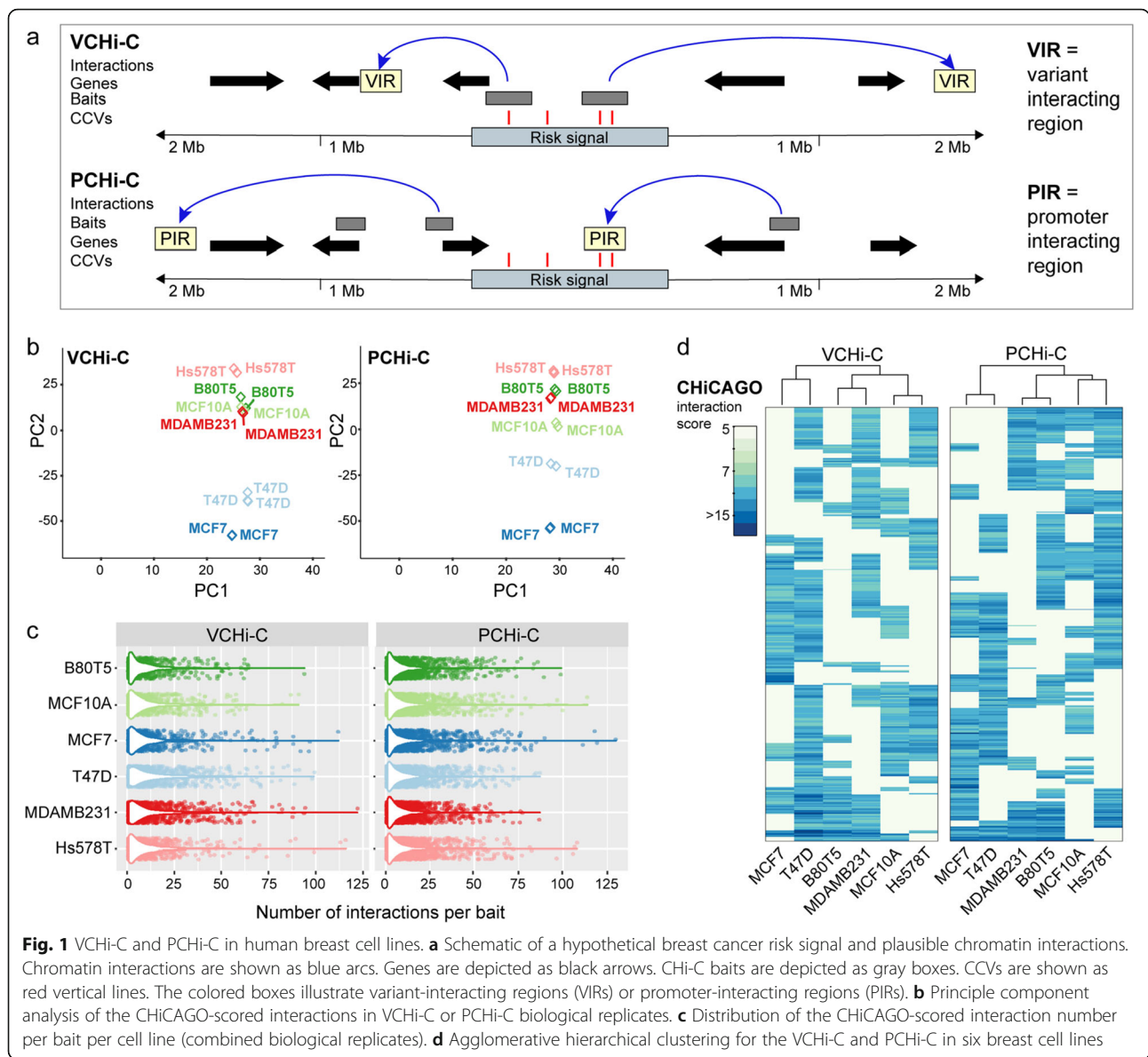
## Results

### VChi-C and PCHi-C interaction profiling

To enrich for chromatin interactions relevant to breast cancer risk, we designed two capture arrays, Variant Capture (VC) and Promoter Capture (PC). The VChi-C baits were designed to *Hind*III fragments that contained at least 1 CCV, regardless of the CCV regulatory potential (Fig. 1a [5]);. We could design baits to 1432 *Hind*III fragments encompassing 6044/7394 CCVs. The PCHi-C

baits were designed to 4045 *Hind*III fragments containing 8216 annotated GENCODE v19 promoters within 1 Mb of CCVs at breast cancer risk signals (Fig. 1a). This dual-capture approach ensured comprehensive coverage of each risk signal and provided independent validation of interactions. We performed in situ VChi-C and PCHi-C [16, 18] in 2 non-tumorigenic breast cell lines (B80T5, MCF10A), 2 estrogen receptor-positive (ER+; MCF7, T47D) breast cancer cell lines, and 2 ER– (MDAMB231, Hs578T) breast cancer cell lines. Sequencing of both captures produced over 1 billion unique di-tags involving CCV-containing fragments and annotated promoters (Additional file 2: Table S1). To assess the robustness of the approach, each CHi-C experiment was conducted in 2 biological replicates per cell type. We observed a strong correlation between the replicates, particularly when captured interaction pairs were within 0.5 Mb (Additional file 1: Figure S1a).

We initially used the CHiCAGO pipeline [23] to assign confidence scores to interactions derived from the VChi-C and PCHi-C (Additional file 2: Tables S2, S3). Principal component analysis (PCA) based on the CHiCAGO scores demonstrated concordance for individual replicates in the VChi-C and PCHi-C. Consistent with previous gene expression profiling [24], PCA separated ER+ breast cancer from normal-like breast or ER– breast cancer cell lines (Fig. 1b). Using a strict interaction threshold (CHiCAGO score  $\geq 5$ , intrachromosomal and interaction distance  $\leq 2$  Mb), we detected on average  $\sim 10,000$  VChi-C and  $\sim 27,000$  PCHi-C high-confidence interactions per cell type (Fig. 1c and Additional file 2: Tables S2, S3). The difference in the interaction number between the captures likely reflects the higher number of PCHi-C baits. In addition, VChi-C baits were designed to all possible CCV-containing *Hind*III fragments, but some CCVs will be correlated with passenger variants or function through alternative non-looping mechanisms, such as promoter variants. For the VChi-C, we detected a median of 5 variant-interacting regions (VIRs; Fig. 1a) per bait per cell type, of which 3.6–5.5% interacted with an annotated protein- or non-coding promoter. This CCV-promoter interaction number is consistent with the assumption that only a small subset of CCVs from each signal are functional regulatory variants. Similarly, for the PCHi-C, we detected a median of 5 promoter-interacting regions (PIRs; Fig. 1a) per bait per cell type, where 2.2–2.7% specifically interacted with a CCV-containing fragment (Additional file 1: Figure S1b and Additional file 2: Tables S2, S3). The median linear distance between interactions from either capture ranged from 192 to 405 kb (Additional file 1: Figure S1c), and  $\sim 70\%$  of the CHi-C interactions occurred within TAD boundaries. Hierarchical clustering based on the CHiCAGO scores separated the cell lines based



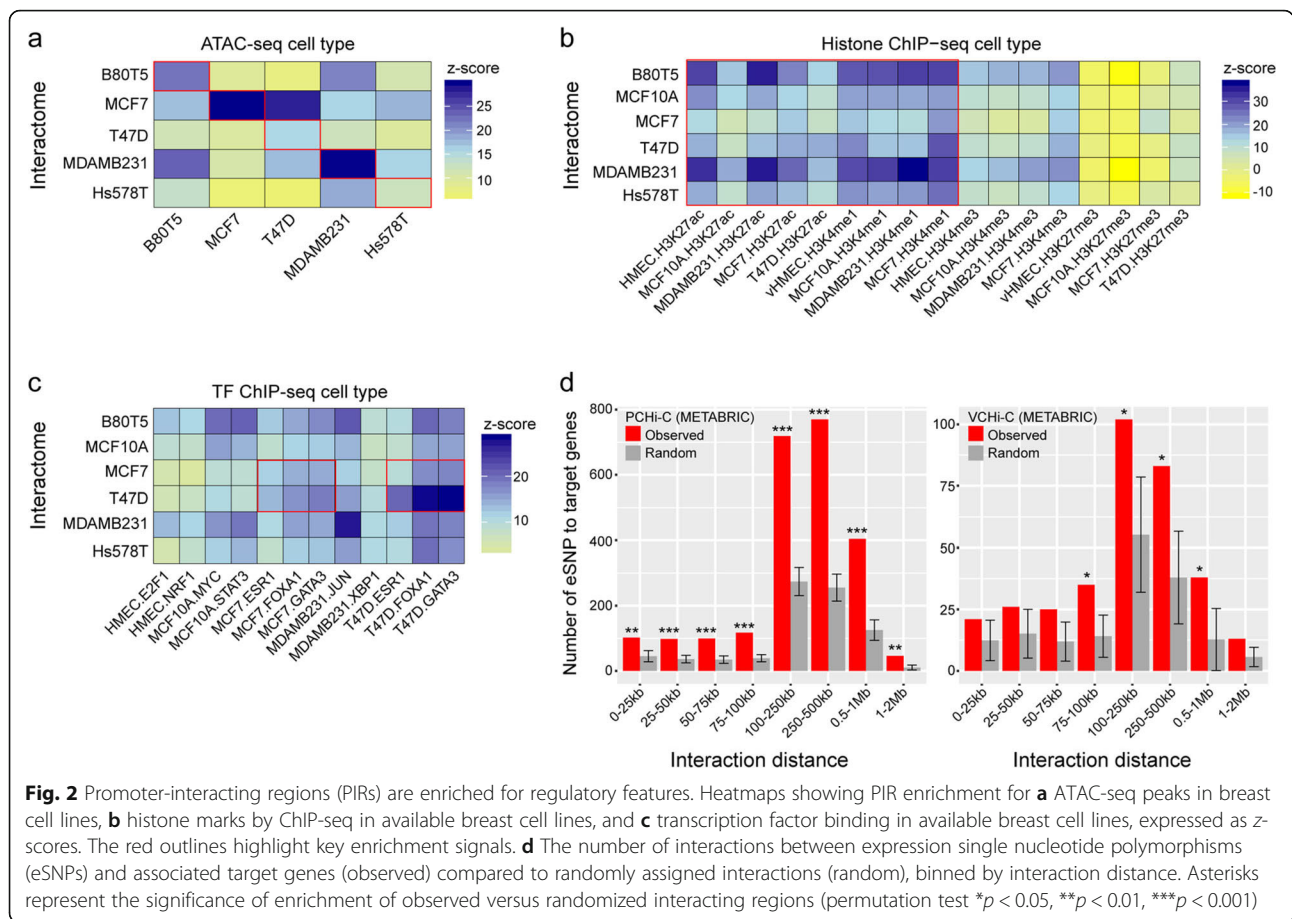
on ER status (Fig. 1d), which suggested that ER status mediates cell-type specificity of the interactomes. We also observed a positive correlation (Pearson's  $r = 0.60-0.84$ ) in the CHiCAGO scores for interactions detected in both the VChi-C and PChi-C (Additional file 1: Figure S1d), thus validating our approach.

**Interacting regions are enriched for regulatory features, eQTLs, and CCVs in breast cells**

We first annotated CHiCAGO-scored PIRs in each breast cell type with DNase-seq data derived from a diverse panel of cells and tissues as part of the Roadmap Epigenomics Project [25]. We found PIRs to be enriched for regions of accessible chromatin in human mammary epithelial cells (HMEC) (Additional file 1: Figure S2a and Additional file 2: Table S4). To explore this

observation in additional breast cells, we annotated PIRs with assay for transposase-accessible chromatin sequencing (ATAC-seq) peaks in five breast cell lines (Additional file 2: Table S5) and noted that the enrichment signals were stronger from PIRs detected in the matched cell line (Fig. 2a and Additional file 2: Table S4). We next investigated the epigenetic makeup of PIRs using ChIP-seq data for histone modifications and other DNA-binding proteins in human cell lines. PIRs were significantly enriched for histone marks associated with active enhancers (H3K27ac and H3K4me1) in the majority of cell lines as compared to inactive elements which are typically marked by the polycomb-associated mark H3K27me3 (Fig. 2b and Additional file 2: Table S4).

Binding sites for several structural proteins with established roles in chromatin looping were also enriched in



PIRs, including CTCF and the cohesin subunits RAD21 and STAG1 (Additional file 1: Figure S2b and Additional file 2: Table S4), consistent with the role of these factors in mediating long-range genomic interactions [9, 12]. Associations were also observed for the cistromes of established breast cancer transcription factors (TFs); ESR1, FOXA1, and GATA3 [26, 27] (Fig. 2c). Notably, ESR1 ChIP-seq peaks were observed at 18% of PIRs in MCF7 cells. Furthermore, the subset of interacting regions with ESR1 binding was significantly enriched for GATA3 binding ( $p < 2.2e-16$ ), consistent with the known role for GATA3 in modulating estrogen signaling [27]. This enrichment was stronger in the ER+ MCF7 and T47D cell lines as compared to available ER- breast cancer, normal breast, and other non-breast cell lines (Fig. 2c, Additional file 1: Figure S2b, and Additional file 2: Table S4), consistent with an additional layer of ER-mediated cell-type specificity [27]. Applying the same enrichment criteria, we also found VIRs to be enriched in ATAC-seq peaks in the matched cell associated with active enhancers (H3K27ac and H3K4me1) and also for H3K4me3, which marks active gene promoters (Additional file 1: Figure S2c, d and Additional file 2: Table S6), supporting the notion that

promoters and enhancers cooperatively communicate through transcriptionally active chromatin [28].

To demonstrate PIR and VIR gene regulatory function, we assessed the overlap of expression quantitative trait loci (eQTLs) in normal breast tissue from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort [2, 29]. We found 800 eQTL genes (eGenes) with eSNPs (false discovery rate (FDR) < 0.05) within PIRs in at least 1 analyzed breast cell line. Examination of the VIR data also revealed 184 eGenes interacting with eSNPs (Fig. 2d). To assess the specificity of eQTL localization to interacting regions, we maintained the interaction network by assigning baits to randomly selected promoters and compared the number of interactions supported by eQTL-target gene pairs. We found that eQTLs were significantly more likely to loop to their associated gene than expected by chance, across a broad range of linear distances from their target promoters (Fig. 2d).

#### Fine-mapping of VCHI-C and PCHI-C profiles

While the CHiCAGO pipeline is extremely useful for interaction detection in CHi-C data [23], many of the generated contact maps contain contiguous restriction



fragments linked with the same target. It is hypothesized that such collateral contacts might result from inaccuracy during the cross-linking process in CHi-C [30] or from bait migration via Brownian motion [31]. Therefore, as a complementary interaction scoring method, we also used a recently developed Bayesian sparse variable selection approach (“*Peaky*” [32]). The model proposes that for any given bait, the expected CHi-C signal at each prey fragment is expressed as a sum of the contributions from a set of fragments directly contacting that bait [32]. We applied *Peaky* to the ~1300 baits from the VChi-C and ~3200 baits from the PChi-C (Additional file 2: Table S7) to derive a measure of confidence in the location of a direct contact called the marginal posterior probability of a contact (MPPC) [32].

To facilitate a comparison with CHiCAGO-scored interactions, we applied an interaction threshold of MPPC  $\geq$  0.1. We filtered for intrachromosomal and interaction distance  $\leq$  2 Mb and detected ~3500 VChi-C and ~7400 PChi-C interactions per cell type (Additional file 1: Figure S3a and Additional file 2: Tables S8, S9). For the VChi-C, ~11% of CCV-containing fragments interacted with an annotated protein- or non-coding promoter, and for the PChi-C, ~2.5% of promoter fragments specifically interacted with a CCV-containing fragment (Additional file 1: Figure S3b and Additional file 2: Tables S8, S9). There were fewer interactions detected by *Peaky*, perhaps because *Peaky* can distinguish and rank a subset of direct contacts from long stretches of chromatin interactions [32]. The median linear distance between interactions from either capture was longer than the CHiCAGO-scored interactions (ranged from 294 to 489 kb; Additional file 1: Figure S3c). Similar to the CHiCAGO-scored interactions, hierarchical clustering based on MPPC scores also separated the cell lines based on ER status (Additional file 1: Figure S3d). We then compared the CHiCAGO and MPPC scores for each bait-prey pair. As reported by Eijbsbouts et al. [32], we noted that CHiCAGO and MPPC scores were positively correlated (Additional file 1: Figure S3e; Spearman’s  $\rho = 0.22$ – $0.37$ ). *Peaky* was able to refine the number of CHiCAGO-scored interactions by 12–17% in both captures; however, a proportion of interactions were identified by *Peaky* but not CHiCAGO (Additional file 1: Figure S3f). To provide a more stringent list of CCVs and candidate target genes, we combined inferences from the two approaches.

#### Prioritizing CCVs by *Peaky* fine-mapping of the PChi-C data

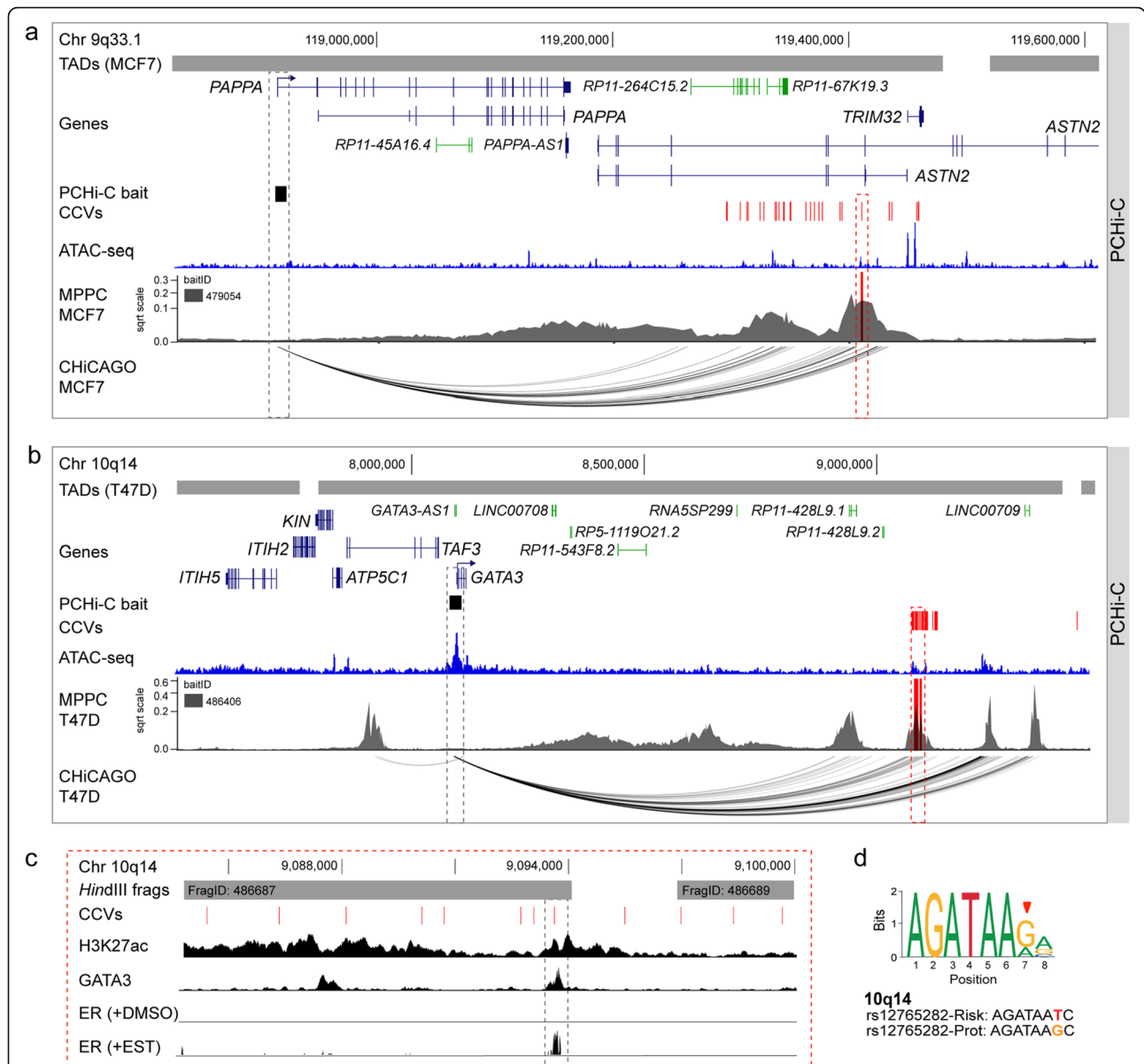
At many signals, we noted that CHiCAGO identified long stretches of PIRs, some of which contained CCVs. We therefore used *Peaky* to fine-map the CHiCAGO-identified interactions to identify the likely driver contacts within these stretches. This approach proved particularly

useful at 9q33.1, where CHiCAGO-identified 24 PIRs starting at ~340 kb from a pregnancy-associated plasma protein A (*PAPPA*) promoter (Fig. 3a). *Peaky* fine-mapping using a *PAPPA* promoter bait indicated this stretch of interactions might be explained by a subset of contacts (MPPC  $\geq$  0.1), which spanned 1 (rs811688) out of 29 CCVs in MCF7 cells (Fig. 3a). 3C provided further support that the *HindIII* fragment containing rs811688 was the most frequently interacting fragment with the *PAPPA* promoter (Additional file 1: Figure S4a). *PAPPA* encodes a secreted zinc metalloproteinase and is an important regulatory component of the insulin-like growth factor system [33]. Recent studies indicate *PAPPA* is frequently overexpressed in luminal B breast tumors [34] and identify *PAPPA* as a pregnancy-dependent oncogene that promotes the formation of pregnancy-associated breast cancer [35]. Another example is 10q14, where CHiCAGO-identified 59 PIRs located ~1 Mb from the GATA binding protein 3 (*GATA3*) promoter. Interactions between *GATA3* and CCVs were restricted to the ER+ (T47D and MCF7) breast cell lines and spanned 49 CCVs (Fig. 3b). *Peaky* fine-mapping indicated this stretch of interactions might be explained by a subset of four contacts, one of which spanned a region containing CCVs. Two *HindIII* fragments within the CCV-containing peak surpassed the 0.1 MPPC interaction threshold and contained 11 out of the 49 CCVs (Fig. 3b). 3C provided further support that the *HindIII* fragment containing 8 CCVs (FragID: 486687) was the most frequently interacting fragment with the *GATA3* promoter (Additional file 1: Figure S4b). Notably, 1 CCV (rs12765282) within the 3C-identified peak mapped to a putative regulatory element as defined by H3K27ac marks and TF binding in T47D cells (Fig. 3c). This CCV is predicted to alter a *GATA3*-binding motif, with the risk allele likely acting to decrease *GATA3* binding. ChIPseq data showed that *GATA3* and ER bound to the CCV site in T47D cells, which are homozygous for the protective *g*-allele (Fig. 3c, d). *GATA3* is important in mediating enhancer accessibility for ER [27], raising the possibility of a *GATA3*-mediated regulatory loop underlying risk at this region.

Taken together, at 77 signals where we could detect at least 1 promoter-CCV interaction, we could prioritize 839 out of 4208 CCVs using the combined CHiCAGO (score  $\geq$  5) and *Peaky* (MPPC  $\geq$  0.1) fine-mapping approach. This included 33 signals where the number of prioritized genetically indistinguishable CCVs could potentially be reduced to less than 5 at each signal (Additional file 2: Table S10).

#### Prioritizing target genes by sequential CHiCAGO and *Peaky* fine-mapping

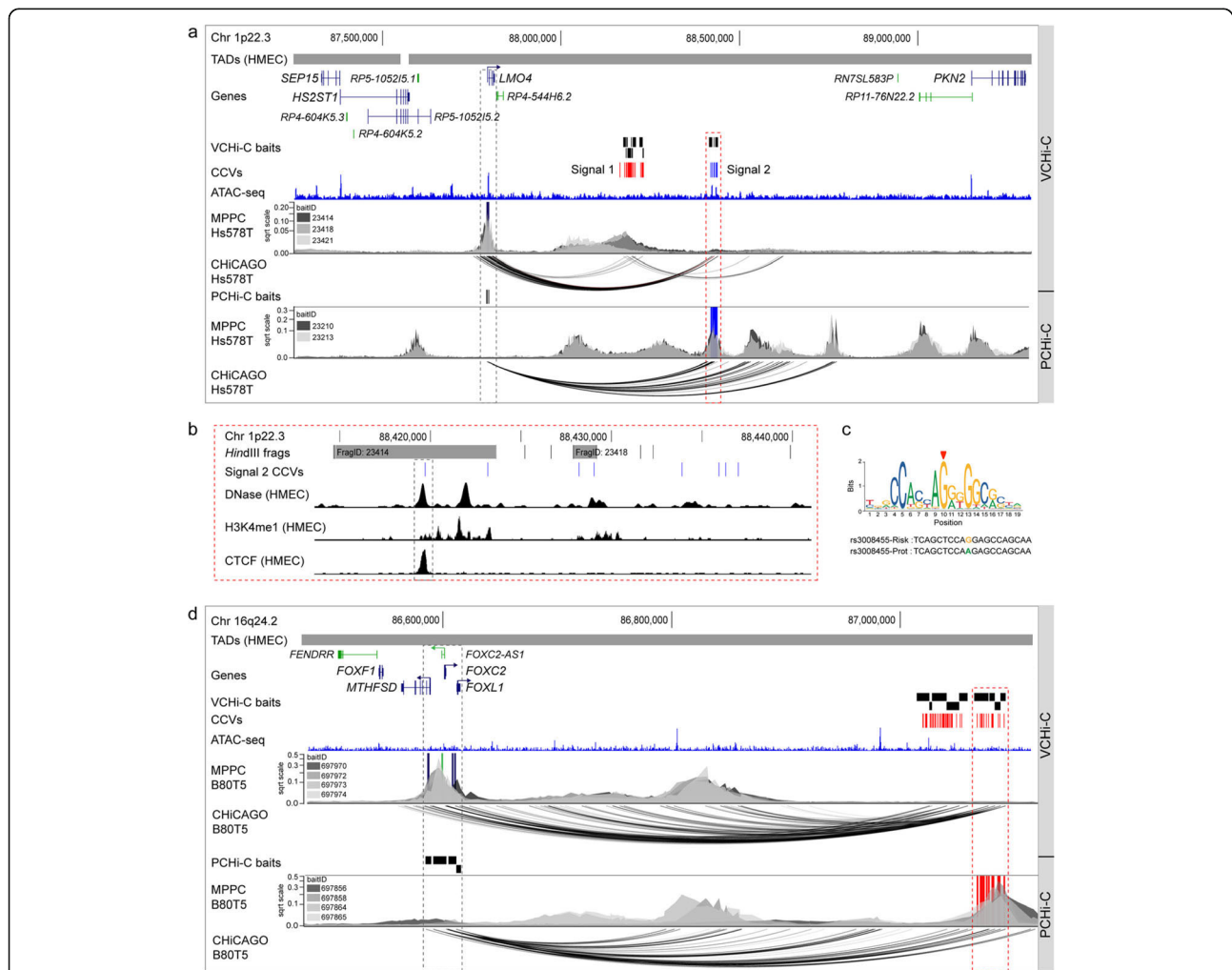
The combined analyses can be extended to integrate, where possible, the VChi-C and PChi-C data.



**Fig. 3** PCHI-C Peaky fine-mapping prioritizes CCVs at 10q14 and 6p22.3. **a** Chromatin interactions at 9q33.1 in MCF7 breast cancer cells. Topologically associating domains (TADs; Additional file 2: Table S16) are shown as horizontal gray bars above GENCODE-annotated coding (blue) and non-coding (green) genes. The PCHI-C bait is depicted as a black box. CCVs are shown as red vertical lines. The ATAC-seq track is shown as a dark blue histogram. Peaky-defined MPPC values (from PCHI-C BaitID: 479054) are plotted with the prioritized CCV overlaid as a red vertical line. CHiCAGO-scored interactions are shown as black arcs. The dashed red outline highlights the prioritized CCV rs811688 and the dashed gray outline the target gene (*PAPPA*). **b** Chromatin interactions at 10q14 in T47D breast cancer cells. Topologically associating domains (TADs) are shown as horizontal gray bars above GENCODE-annotated coding (blue) and non-coding (green) genes. The PCHI-C bait is depicted as a black box. CCVs are shown as red vertical lines. The ATAC-seq track is shown as a dark blue histogram. Peaky-defined MPPC values (from PCHI-C BaitID: 486406) are plotted with the prioritized CCVs overlaid as red vertical lines. CHiCAGO-scored interactions are shown as black arcs. The dashed red outline highlights the prioritized CCVs and the dashed gray outline the target gene (*GATA3*). **c** Zoomed in view of prioritized CCVs at 10q14. *Hind*III fragments are shown as gray bars with their fragment IDs. CCVs are shown as red vertical lines. Black histograms denote ChIP-seq data from T47D cells for H3K27ac, GATA3, and estrogen receptor (ER; cells treated with DMSO or 17 beta-estradiol (EST)). The dashed gray outline highlights CCV rs12765282. **d** Position weight matrix of the GATA3 binding site from JASPAR (red arrowhead indicates the CCV position in the motif), with homology to the risk (*t*) and protective (*g*) alleles of rs12765282 colored below

One example is 1p22.3, where CHiCAGO-detected interactions in the VCHi-C data between two independent signals and the LIM-only protein 4 (*LMO4*) promoter in Hs578T breast cancer cells (Fig. 4a). Peaky fine-mapping using signal 2 VCHi-C baits then provided further support that *LMO4* was the likely target gene (Fig. 4a). Peaky was also applied to signal 1 VCHi-C baits, but the

resulting contact peaks did not reach the 0.1 MPPC interaction threshold (Additional file 1: Figure S4c). We then interrogated the PCHi-C data using two *LMO4* promoter baits in Hs578T cells. CHiCAGO identified 84 PIRs starting at ~612 kb from the *LMO4* promoter (Fig. 4a). Peaky fine-mapping using the same promoter baits indicated this stretch of interactions might be



**Fig. 4** Sequential CHiCAGO and Peaky fine-mapping prioritizes CCVs and target genes. **a** Chromatin interactions at 1p22.3 in Hs578T breast cancer cells. Topologically associating domains (TADs) are shown as horizontal gray bars above GENCODE-annotated coding (blue) and non-coding (green) genes. The VCHi-C or PCHi-C baits are depicted as black boxes. Risk signals 1 and 2 are numbered, and the CCVs within each signal are shown as colored vertical lines. The ATAC-seq track is shown as a dark blue histogram. Peaky-defined MPPC values (from specified BaitIDs) are plotted with the prioritized gene overlaid as a dark blue vertical line or prioritized CCVs overlaid as royal blue vertical lines. CHiCAGO-scored interactions for specified BaitIDs are shown as black arcs. The dashed red outline highlights the prioritized CCVs and the dashed gray outline the target gene (*LMO4*). **b** Zoomed-in view of prioritized signal 2 CCVs at 1p22.3. VCHi-C baits are shown as gray bars with their fragment IDs. CCVs are shown as blue vertical lines. Black histograms denote DNase I hypersensitivity sites or ChIP-seq data for H3K4me1 and CTCF binding from HMEC cells. The dashed gray outline highlights CCV rs3008455. **c** Position weight matrix of the CTCF binding site from JASPAR (red arrowhead indicates the CCV position in the motif), with homology to the risk (*g*) and protective (*a*) alleles of rs3008455 colored below. **d** Chromatin interactions at 16q24.2 in B80T5 normal breast cells. Topologically associating domains (TADs) are shown as horizontal gray bars above GENCODE-annotated coding (blue) and non-coding (green) genes. The VCHi-C or PCHi-C baits are depicted as black boxes. CCVs are shown as red vertical lines. The ATAC-seq track is shown as a dark blue histogram. Peaky-defined MPPC values (from specified BaitIDs) are plotted with the prioritized genes overlaid as dark blue or green vertical lines and prioritized CCVs overlaid as red vertical lines. CHiCAGO-scored interactions for specified BaitIDs are shown as black arcs. The dashed red outline highlights the prioritized CCVs and the dashed gray outline the prioritized target genes (*MTHFSD*, *FOXC2*, *FOXL1*, and *FOXC2-AS1*)

explained by a subset of three direct contacts (MPPC  $\geq$  0.1). One contact spanned two *HindIII* fragments within signal 2 and potentially prioritized four out of eight CCVs at this signal (Fig. 4b). Of these, one CCV (rs3008455) mapped to a putative regulatory element as defined by open chromatin and TF binding in normal breast cells (Fig. 4b). This CCV is predicted to alter a CTCF-binding motif, with the risk allele promoting increased CTCF binding (Fig. 4c). LMO4 is a transcriptional modulator that is overexpressed in > 50% of breast tumors [36]. Overexpression of LMO4 promotes cell proliferation, invasion, and tumor formation and induces mammary hyperplasia in transgenic mice [37].

A more complex example is 16q24.2, where CHiCAGO detected 62 VIRs spanning a ~320-Kb genomic region derived from 9 separate VChi-C baits (Fig. 4d). Peaky fine-mapping of this VChi-C data then prioritized *FOXC2*, *FOXC2-ASI*, *FOXL1*, and *MTHFSD* as the likely target genes in B80T5 breast cells (Fig. 4d). We interrogated the PChi-C data using the 4 target gene promoter baits in B80T5 cells. CHiCAGO identified 40 PIRs spanning a ~500-Kb genomic region. Peaky fine-mapping using the same promoter baits indicated this stretch of interactions might be explained by a subset of 2 direct contacts (MPPC  $\geq$  0.1). One contact spanned 5 *HindIII* fragments and potentially prioritized 21 out of the possible 85 CCVs at this signal (Fig. 4d). Preliminary in silico analyses revealed many of the 21 prioritized CCVs display regulatory activity, and therefore, additional studies would be required to determine which are the likely functional variants. *FOXC2* and *FOXL1* are members of the Forkhead family of transcription factors with important functions in biological processes such as cell cycle control, proliferation, and development [38]. *FOXC2* has been implicated in triple-negative breast cancer progression and therapy resistance [39], while *FOXL1* is reported to inhibit breast cancer cell proliferation, invasion, and migration [40]. Little is known about methenyltetrahydrofolate synthetase domain-containing (*MTHFSD*), but a recent report suggests the gene encodes a stress granule-associated RNA-binding protein [41].

#### Identification of 651 candidate target genes at 139 breast cancer risk signals

We defined candidate target genes of breast cancer risk signals by CHiCAGO- and/or Peaky-scored CCV-gene promoter interactions in VChi-C or PChi-C in at least 2 cell lines. This combined analysis resulted in 651 candidate target genes at 139 breast cancer risk signals, including 419 protein-coding genes (Additional file 2: Table S11). Most of the identified target genes are expressed in normal breast tissue, ductal carcinoma in situ (DCIS), or breast tumors, with 66% of genes differentially expressed between normal breast and breast

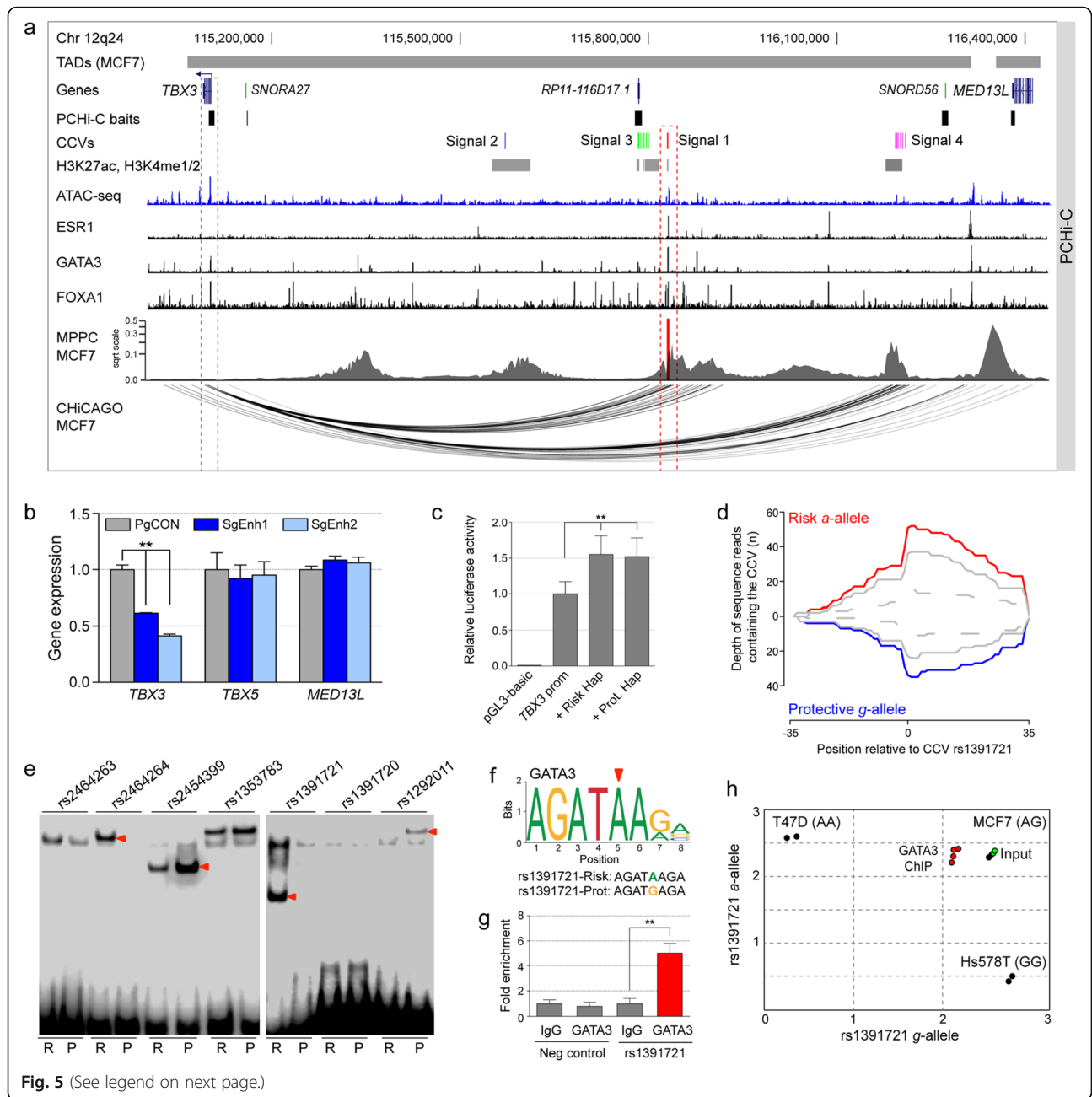
tumor (Additional file 2: Table S12) or 33% between breast organoid and DCIS samples (Additional file 2: Table S13). The majority of candidate target genes interacted with 1 signal, but ~13% interacted with 2 or more independent signals (Additional file 1: Figure S4d). The 6q25 region is one of the more extreme examples, where 5 out of 6 independent signals all loop to and potentially regulate *ESR1* [42, 43] (Additional file 1: Figure S4e). More than 80% of signal-target gene interactions skipped at least 1 annotated gene promoter, and ~75% of signals interacted with at least 2 promoter-containing fragments (Additional file 1: Figure S4f). One example that demonstrates both characteristics is 8q24.13, where signal 1 CCVs interact with 6 candidate target genes (*WDYHVI*, *FBXO32*, *CTD-2552K11.2*, *ANXA13*, *FAM91A1*, and *TRMT12*) including skipping 3 annotated genes to contact the *TRMT12* promoter (Additional file 1: Figure S4g). Notably, 181 candidate target genes were identified by both CHiCAGO and Peaky (Additional file 1: Figure S4h), which may further prioritize these genes for functional validation. This priority list includes established breast cancer driver genes such as *MYC* and *GATA3* [44] but also includes many genes with no reported role in breast cancer (Additional file 2: Table S11).

#### Chi-C identifies *TBX3* as the target of multiple risk signals

To further illustrate the power of combining genetic fine-mapping, CHi-C, and functional studies, we examined in detail the 12q24 susceptibility region. Genetic fine-mapping of 12q24 identified at least four independent signals [2, 5] (listed in Additional file 2: Table S14); signal 1 (seven CCVs), signal 2 (one CCV), and signal 4 (six CCVs) were more strongly associated with ER+ tumors, whereas signal 3 (eight CCVs) was associated with both ER+ and ER- breast cancer (Additional file 2: Table S14). The CCVs in all four signals are located in a large intergenic region on 12q24 between *TBX3* and *MEDI3L* (Fig. 5a). We used ATAC-seq and available ChIP-seq datasets [45, 46] to map CCVs relative to transcriptional regulatory elements. These analyses showed evidence of putative regulatory elements overlapping the CCVs at each signal, indicating that one or more CCVs likely have high regulatory potential (Fig. 5a). CHi-C and 3C identified T-Box 3 (*TBX3*) as the most likely target gene (Figs. 5a, Additional file 1: Figure S5a, and Additional file 2: Table S15). Notably, we detected interactions between *TBX3* and each of the four independent signals in a cell type-specific manner (Fig. 5a and Additional file 1: Figure S5b).

Our functional studies focused on the strongest signal 1 CCVs. CRISPRi silencing of the signal 1 element in ER+ MCF7 cells showed that *TBX3*, but not *TBX5* and *MEDI3L*, levels were significantly reduced (Fig. 5b). Reporter assays then confirmed that the element acts as an enhancer on the *TBX3* promoter in the presence of





(See figure on previous page.)

**Fig. 5** Molecular analysis of signal 1 CCVs at 12q24. **a** Chromatin interactions in MCF7 cells. Topologically associating domains (TADs) are shown as horizontal gray bars above GENCODE-annotated coding (blue) and non-coding (green) genes. The PCHi-C baits are depicted as black boxes. Risk signals 1–4 are numbered, and the CCVs within each signal are shown as colored vertical lines. ENCODE ChIP-seq data for available histone marks are depicted as gray boxes. The ATAC-seq track is shown as a dark blue histogram. ESR1, GATA3, and FOXA1 binding are shown as black histograms. Peak defined MPPC values (from PCHi-C BaitID: 596031) are plotted with the prioritized CCVs overlaid as red vertical lines. CHiCAGO-scored interactions are shown as black arcs. The dashed red outline highlights signal 1 CCVs and the dashed gray outline the target gene (*TBX3*). **b** The 12q24 enhancer was repressed by targeting dCas9-KRAB to the enhancer in MCF7 cells with two different CRISPRi single-guide (sg) RNAs (SgEnh1 and SgEnh2). PgCON contains a non-targeting control sgRNA. Gene expression of *TBX3*, *TBX5*, and *MED13L* was measured by qPCR and normalized to *GUSB*. Error bars represent the SEM ( $n = 3$ ).  $p$  values were determined by two-way ANOVA followed by Dunnett's multiple-comparison test (\*\* $p < 0.01$ ). **c** Luciferase reporter assays following transient transfection of MCF7 cells. The 12q24 enhancer containing either the risk or protective (Prot.) haplotype was cloned into *TBX3* promoter-driven luciferase constructs (*TBX3* prom). Error bars represent the SEM ( $n = 3$ ).  $p$  values were determined by two-way ANOVA followed by Dunnett's multiple-comparison test (\*\* $p < 0.01$ ). **d** Allele-specific DNase I hypersensitivity at CCV rs1391721 in heterozygous MCF7 cells. The depth of reads containing the risk (red) and protective (blue) alleles are shown. **e** EMSAs for signal 1 CCVs to detect allele-specific binding of nuclear proteins. Labeled oligonucleotide duplexes were incubated with MCF7 nuclear extract. Red arrowheads show the bands of different mobility detected between risk (R) and protective (P) alleles. **f** Position weight matrix of the GATA3 binding site from JASPAR, with homology to the risk (*a*) and protective (*g*) alleles of rs1391721 colored below. **g** Allele-specific GATA3 ChIP-PCR results assessed at CCV rs1391721 in heterozygous MCF7 cells. Error bars represent the SEM ( $n = 3$ ).  $p$  values were determined by a two-tailed Student's  $t$  test (\*\* $p < 0.01$ ). **h** Allelic discrimination plot of the GATA3 ChIP in MCF7 cells. Genomic DNA extracted from homozygous T47D and Hs578T breast cancer cells were used as controls

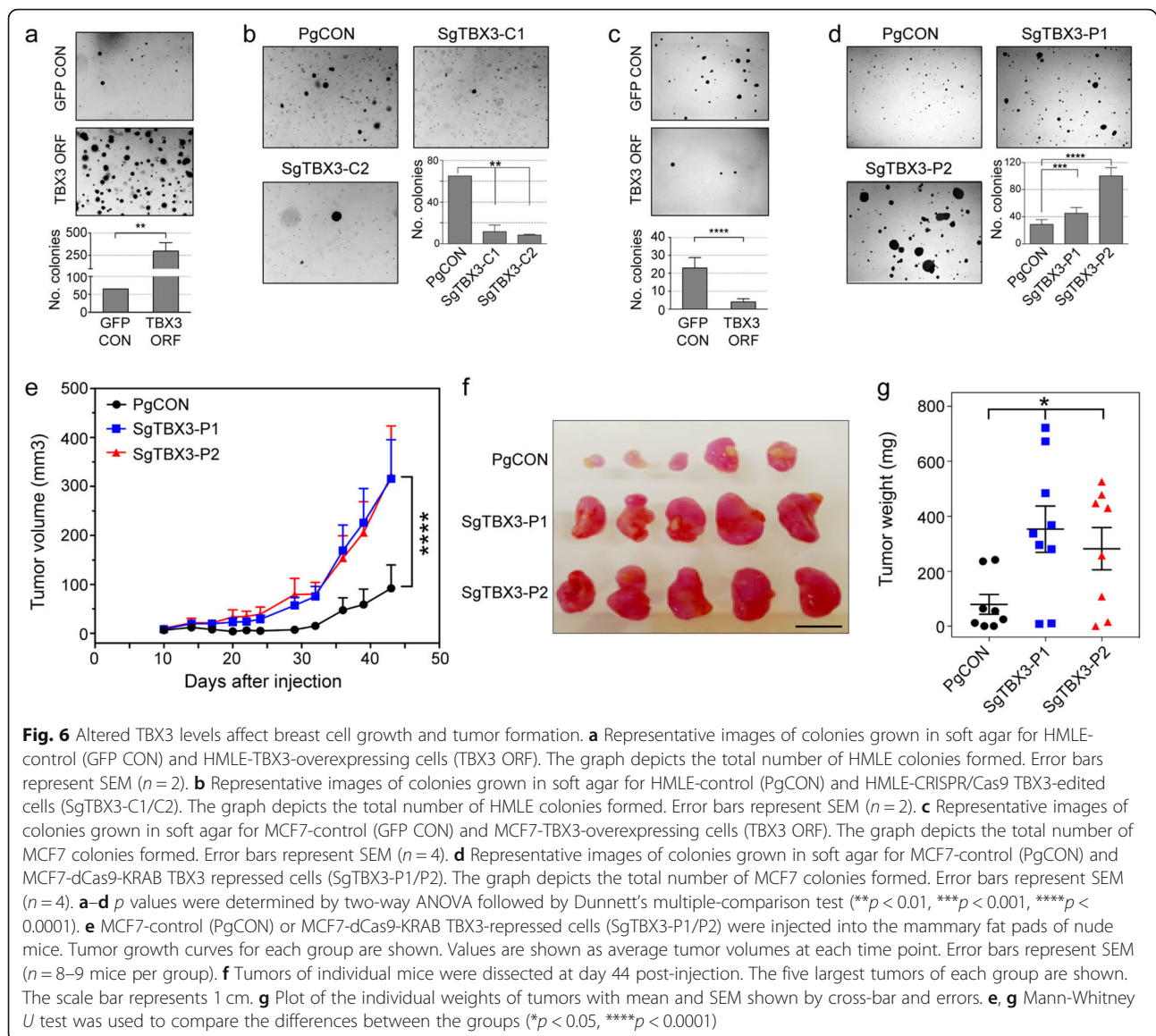
either the risk or protective haplotypes (Fig. 5c). We used available DNase-seq data derived from heterozygous MCF7 cells to show that the risk *a*-allele of CCV rs1391721 may promote allele-specific open chromatin (Fig. 5d). Electrophoretic mobility shift assays (EMSAs) then assessed TF binding for each of the signal 1 CCVs. Allele-specific binding by nuclear proteins was observed for CCVs rs2464264, rs2454399, rs1391721, and rs1292011 in MCF7 and BT474 extracts (Fig. 5e and Additional file 1: Figure S6a). Further EMSAs using competitor DNA against predicted TFs suggested GATA3 bound to the rs1391721 site (Additional file 1: Figure S6b). Similar to the 10q14 CCV, rs1391721 is also predicted to lie in a GATA3 binding site. Here, the risk *a*-allele promoted increased GATA3 binding compared to the protective *g*-allele (Fig. 5f), as evident in the GATA3 ChIP-seq data derived from heterozygous MCF7 cells (Additional file 1: Figure S6c). Two other TFs involved in estrogen signaling, ESR1 and FOXA1, also colocalize at this enhancer (Additional file 1: Figure S6d). To assess occupancy of GATA3 in vivo, we performed ChIP followed by allele-specific qPCR in MCF7 cells and found that GATA3 was preferentially recruited to the *a*-allele of rs1391721 (Fig. 5g, h). As further support, we investigated the correlation between *GATA3* and *TBX3* expression in the TCGA cohort. A stronger correlation was observed between *GATA3* and *TBX3* expression in normal breasts as compared with the breast tumor samples (Additional file 1: Figure S6e).

*TBX3* is a T-Box TF that has been linked to tumorigenesis by impacting senescence and apoptosis as well as promoting proliferation and tumor formation [47]. To determine whether *TBX3* can promote a tumorigenic phenotype in breast cells, we stably overexpressed or repressed *TBX3* in the human mammary epithelial

(HMLE) cell line and the MCF7 breast cancer cell line. HMLEs have been engineered to express *hTERT* and the SV40 large T antigen and can grow in soft agar and form tumors in immune-deficient mice only upon the introduction of an additional oncogenic insult [48]. Overexpression of *TBX3* in HMLE cells resulted in a significant increase in cell colony growth in soft agar, suggesting that overexpression promotes anchorage-independent growth (Fig. 6a and Additional file 1: Figure S6f), while CRISPR/Cas9-mediated *TBX3* silencing showed a reciprocal effect (Fig. 6b). These results are consistent with our in vitro data which indicated breast cancer risk was likely associated with increased *TBX3* expression. The HMLE-*TBX3*-overexpressing cells were also injected into the mammary fat pads of nude mice, but no tumors were observed, suggesting elevated levels of *TBX3* alone is not enough to promote tumor development from these cells. In contrast, overexpression of *TBX3* in MCF7 cells decreased cell colony growth in soft agar (Fig. 6c and Additional file 1: Figure S6g), while depletion of *TBX3* by targeting dCas9-KRAB to the *TBX3* promoter resulted in a significant increase in growth (Fig. 6d and Additional file 1: Figure S6h). To further investigate *TBX3* in tumor growth, *TBX3*-depleted MCF7 cells were injected into the mammary fat pads of nude mice. Compared to control cells, reduced *TBX3* levels resulted in a marked increase in tumor growth in vivo (Fig. 6e, f), which was reflected in increased tumor weights (Fig. 6g). As reported previously [49], these data suggest that *TBX3* can be oncogenic or tumor suppressive depending on the cellular context.

## Discussion

The field of 3D chromatin interaction mapping is rapidly changing how we view the genome and is revealing



important insights into disease biology. Interpretation of findings from GWAS has particularly benefited from the influx of chromatin data, allowing more accurate mapping and redefining of candidate causal genes. In this study, we generated high-resolution chromatin maps in human breast cells to delineate gene-regulatory interactions between breast cancer CCVs and target genes. We used two independent algorithms to score chromatin interactions. Peaky assisted the identification of the probable direct contacts from long stretches of CHiCAGO-identified interactions. This proved useful when examining PIRs as we were able to further prioritize the list of CCVs, which will be valuable in future in-depth functional studies. The de-prioritized variants may simply represent those in linkage disequilibrium with the true causal variant(s). Similarly, we observed an overlap

between the CHiCAGO- and Peaky-detected target genes but noted that a proportion was detected by only one method. This was not unexpected given the different statistical models, and further studies will be required to establish parameters for improved resolution of direct interactions. Collectively, we could identify 651 candidate target genes at 139 independent breast cancer risk signals. Of particular interest for post-GWAS functional studies, 65 signals could be prioritized to one or two candidate target genes (Table 1). Some of the listed genes already have functional data linking breast cancer CCVs and somatic point mutations to altered target gene expression, including *ESR1* [42, 43], *FGFR2* [50, 51], and *MAP3K1* [44, 52].

A recent study used CHi-C to identify 110 putative target genes at 33 breast cancer risk loci [53].

**Table 1** Independent breast cancer risk signals with  $\leq 2$  candidate protein-coding genes

Cytoband	Locus	Signal	Target gene/s	Cytoband	Locus	Signal	Target gene/s
1p22.3	chr1:87656923_88656923	2	<i>LMO4</i>	8q24.21	chr8:127424659_130041931	3	<i>MYC</i>
1q32.1	chr1:200937832_201937832	1	<i>IPO9</i>	9q31.2	chr9:109803808_111395353	1	<i>KLF4</i>
2p23.3	chr2:28670676_29670676	1	<i>ALK,SPDYA</i>	9q33.1	chr9:118813486_119813486	1	<i>PAPPA</i>
2p24.1	chr2:18815791_19820803	1	<i>OSR1</i>	10p14	chr10:8588113_9588113	1	<i>GATA3</i>
2q35	chr2:217405832_218796508	1	<i>IGFBP5</i>	10q25.2	chr10:114273927_115286154	1	<i>TCF7L2</i>
2q35	chr2:217405832_218796508	3	<i>IGFBP5</i>	10q26.12	chr10:122593901_123849324	2	<i>FGFR2</i>
3p24.1	chr3:26827965_28285247	2	<i>AZ12,CMC1</i>	11p15.5	chr11:1398664_2442575	1	<i>LSP1</i>
3q23	chr3:140612859_141612859	1	<i>ZBTB38</i>	11q13.3	chr11:68831418_69879161	1	<i>MYEOV</i>
4p14	chr4:38312876_39312876	1	<i>TBC1D1,TLR10</i>	11q13.3	chr11:68831418_69879161	2	<i>MYEOV</i>
4q24	chr4:105569013_106856761	1	<i>GSTCD,PPA2</i>	11q13.3	chr11:68831418_69879161	3	<i>MYEOV</i>
5p13.3	chr5:32067732_33067732	1	<i>ZFR</i>	11q24.3	chr11:128952507_129961171	1	<i>BARX2</i>
5p15.33	chr5:779790_1797488	1	<i>SLC6A18</i>	12p11.22	chr12:27639846_29034415	1	<i>CCDC91</i>
5p15.33	chr5:779790_1797488	2	<i>SLC6A18</i>	12p13.1	chr12:13913931_14913931	1	<i>ATF7IP</i>
5q11.1	chr5:49141645_50695093	2	<i>CTD-2203A3.1,ISL1</i>	12q22	chr12:95527759_96527759	1	<i>NTN4,RP11-536G4.1</i>
5q11.2	chr5:55531884_56587883	1	<i>MAP3K1</i>	12q24.21	chr12:115336522_116336522	1	<i>TBX3</i>
5q11.2	chr5:55531884_56587883	4	<i>MAP3K1</i>	12q24.21	chr12:115336522_116336522	2	<i>TBX3</i>
5q11.2	chr5:55531884_56587883	5	<i>MAP3K1</i>	12q24.21	chr12:115336522_116336522	3	<i>TBX3</i>
5q11.2	chr5:57684061_58865569	1	<i>GAPT</i>	13q13.1	chr13:32468810_33472626	1	<i>FRY</i>
5q11.2	chr5:57684061_58865569	2	<i>PDE4D</i>	13q22.1	chr13:73464519_74464519	1	<i>KLF5</i>
5q33.3	chr5:157730013_158744083	1	<i>EBF1</i>	14q13.3	chr14:36632769_37635752	1	<i>SLC25A21,SLC25A21-AS1</i>
6p22.3	chr6:15899557_16899557	1	<i>ATXN1</i>	14q24.1	chr14:68117194_69534682	1	<i>ZFP36L1</i>
6q14.1	chr6:81628386_82795951	1	<i>AL359693.1</i>	14q24.1	chr14:68117194_69534682	2	<i>ZFP36L1</i>
6q23.1	chr6:129849119_130849119	1	<i>AKAP7,TMEM244</i>	16q12.2	chr16:53300954_54355291	2	<i>IRX5,LPCAT2</i>
6q25	chr6:151418856_152937016	2	<i>ESR1</i>	16q23.2	chr16:80148327_81150805	1	<i>CDYL2</i>
6q25	chr6:151418856_152937016	3	<i>SYNE1</i>	18q11.2	chr18:23832476_25075396	1	<i>KCTD1</i>
6q25	chr6:151418856_152937016	5	<i>ESR1</i>	19q12	chr19:29777729_30777729	1	<i>CCNE1</i>
6q25.1	chr6:149086328_150086328	1	<i>TAB2</i>	19q13.31	chr19:43783447_44786513	1	<i>KCNN4</i>
7q22.1	chr7:101054599_102054599	1	<i>COL26A1</i>	20p12.3	chr20:5448227_6448227	1	<i>GPCPD1</i>
7q34	chr7:139442304_140442304	1	<i>SLC37A3</i>	21q21.1	chr21:16073983_17073983	1	<i>HSPA13,NRIP1</i>
8p12	chr8:29009616_30009616	1	<i>DUSP4</i>	21q21.1	chr21:16073983_17073983	2	<i>HSPA13,NRIP1</i>
8q21.11	chr8:75730301_76917937	2	<i>CRISPLD1</i>	22q13.31	chr22:45783297_46783297	1	<i>ATXN10</i>
8q23.3	chr8:116709548_117709548	1	<i>TRPS1</i>	22q13.31	chr22:45783297_46783297	2	<i>ATXN10,WNT7B</i>
8q24.21	chr8:127424659_130041931	2	<i>FAM84B,MYC</i>				

Surprisingly, only 30 of the 110 genes were also identified in our study. The lack of concordance may firstly result from a fundamental difference in capture design; Baxter et al. were based on SNPs correlated with the published SNP ( $r^2 \geq 0.2$ ), whereas the present study captures only those fragments containing CCVs based on fine-mapping analysis of a very large association dataset. In addition, the design used by Baxter and colleagues included many examples where oligonucleotide probes were tiled across large genomic regions rather than restricted to individual *HindIII* fragments. Baxter et al. also reported multiple genes as putative targets at some risk

signals, while our analysis of the same signals prioritized only 1 or 2 genes. For example, at 11p15.5, Baxter et al. identified 9 target genes, whereas our combined statistical analyses reduced this number to just 2 candidates, *LSP1* and *MIR4298*.

We acknowledge that some CCV-target gene interactions may have been missed due to intrinsic biases in the capture. False negatives may result from the lack of suitable baits for some CCV- and promoter-containing fragments, short-range contact constraints, or due to the transient and cell type-specific nature of regulatory chromatin interactions. It is also possible that the proportion



of our observed interactions may be cell culture condition dependent. It is also important to keep in mind that interactions between a CCV and gene promoter do not infer causality. It is likely that correlated CCVs within some signals have no effect on TF binding or enhancer activity, or they may act via alternate mechanisms. Consistent with other GWAS follow-up studies [2, 26, 54], our results support the hypothesis that *cis*-acting regulatory variation is a predominant molecular mechanism at breast cancer risk signals. However, we saw no CCV-target gene looping interactions at 57 (out of 196) risk signals. Twelve signals contained promoter or coding CCVs, suggesting that direct gene alteration is a probable mechanism underlying these risk associations. The remaining signals ( $n = 45$ ) contained baited variant or promoter fragments, but the lack of detected CCV-gene interactions suggests mechanisms other than distal regulation. A recent study has incorporated some of the proposed alternate CCV mechanisms together with the distally regulated genes from this study to generate a complete catalog of candidate target genes and biological pathways [5].

We provided functional evidence that breast cancer risk at 12q24 is driven by the TF, TBX3. TBX3 is over-expressed in many cancers including breast cancer and contributes to oncogenesis at multiple levels including the promotion of proliferation, tumor formation, and metastasis [47]. Consistent with previous findings, our *in vitro* data indicate that the signal 1 CCVs likely act to increase TBX3 expression through recruitment of GATA3 to the CCV site, resulting in an increased looping of the risk CCV-containing enhancer to the *TBX3* promoter. Recent studies have suggested that TBX3 may also function as a tumor suppressor depending on the cellular context [49]. Indeed, in MCF7 breast cancer cells, we showed that TBX3 repression promoted colony formation and *in vivo* tumor formation. Furthermore, somatic *TBX3* mutations in primary breast tumors are predominantly loss-of-function through impaired transcriptional repression [55]. Interestingly, a recent report showed that many of these “double-agent” genes are TFs and that breast cancer is the second most common cancer type associated with dual-function genes [56]. The molecular mechanisms underlying this duality are largely unknown, but differing mutation spectrums, interaction partners, and cellular contexts have been implicated. Dual-function genes likely contribute to the heterogeneity of cancer cells, and some are already considered promising targets for breast cancer therapy. It will therefore be important to refine therapeutic strategies to selectively block one function without compromising the other.

In summary, we report the most comprehensive study linking regulatory CCVs to candidate breast cancer

genes. This forms an important resource for the breast cancer research community that will facilitate the generation of hypotheses, functional experimentation, and insights into breast cancer biology. We anticipate that many of the candidate target genes may represent drug repositioning opportunities or be suitable for future drug targeting.

## Methods

### Availability of data and materials

Raw sequencing data has been deposited at EBI: PRJEB29716. Processed Capture Hi-C data is available from <https://osf.io/2cnw7/>. Processed chromatin interaction data can be visualized at the Washington Epigenome Browser via <http://epigenomegateway.wustl.edu/browser/live/Tei2kygBF>. The custom scripts used in the study are available from [https://github.com/jmbeesley/Beesley\\_GenomeBiol2019](https://github.com/jmbeesley/Beesley_GenomeBiol2019). All datasets and software used are listed in Additional file 2: Table S16 [5, 12, 16, 23, 25, 32, 45, 46, 57–68].

### Cell lines

Estrogen receptor-positive (ER+) breast cancer cell lines MCF7 and T47D were grown in RPMI medium with 10% (vol/vol) fetal bovine serum (FBS), 1 mM sodium pyruvate, 10  $\mu$ g/ml insulin, and 1% (vol/vol) antibiotics. ER- breast cancer cell lines MDAMB231 and Hs578T were grown in DMEM medium with 10% (vol/vol) FBS and 1% (vol/vol) antibiotics. The B80T5 mammary epithelial cell line [69] (provided by Roger Reddel, CMRI, Australia) was grown in RPMI medium with 10% (vol/vol) FBS and 1% (vol/vol) antibiotics. The MCF10A mammary epithelial cell line was grown in DMEM/F12 medium with 5% (vol/vol) horse serum, 10  $\mu$ g/ml insulin, 0.5  $\mu$ g/ml hydrocortisone, 20 ng/ml epidermal growth factor, 100 ng/ml cholera toxin, and 1% (vol/vol) antibiotics. Cell lines were maintained under standard conditions (37 °C, 5% CO<sub>2</sub>), tested for *Mycoplasma*, and profiled for short tandem repeats.

### Hi-C library preparation

Hi-C libraries were prepared from 4 to 8  $\times 10^7$  cells per library (2 biological replicates per cell line; 3 replicates for the T47D VCHi-C) as described previously [11], but using *in-nucleus* ligation as described in [70]. The immobilized Hi-C libraries were amplified using the SureSelect<sup>XT</sup> ILM Indexing pre-capture primers (Agilent Technologies) with 8 PCR amplification cycles. Each Hi-C library (750 ng) was hybridized and captured individually using the SureSelect<sup>XT</sup> Target Enrichment System reagents and protocol (Agilent Technologies). After library enrichment, a post-capture PCR amplification step was carried out using SureSelect<sup>XT</sup> ILM Indexing post-

capture primers (Agilent Technologies) with 14–16 PCR amplification cycles.

#### Biotinylated RNA bait library design

The SureSelect<sup>XT</sup> Custom Target Enrichment Arrays were designed using the eARRAY software (Agilent Technologies). For the VChi-C, biotinylated 120-mer RNA baits were designed to both ends of *Hind*III restriction fragments that contained at least 1 CCV [5]. A total of 1448 *Hind*III fragments were captured, covering 6044/7394 CCVs. For the PChi-C, biotinylated 120-mer RNA baits were designed to both ends of *Hind*III restriction fragments that overlapped annotated promoters within 1 Mb of CCVs [5]. A total of 4049 *Hind*III fragments were captured, overlapping 2298 Ensembl-annotated promoters (GRCh38) [16]. A bait sequence was accepted if its GC content was between 25 and 65%, the sequence contained no more than 2 consecutive nucleotides of the same identity, and was within 330 bp of the *Hind*III restriction fragment end. Repetitive elements were masked using SureDesign masking tools with the highest level of stringency.

#### Sequencing of Chi-C libraries

PChi-C and VChi-C libraries were sequenced on the Illumina HiSeq 2500 platform (Kinghorn Centre for Clinical Genomics, Australia). Two PChi-C or three VChi-C libraries were multiplexed per sequencing lane.

#### PChi-C and VChi-C sequence alignment and data processing

Raw sequencing reads were truncated, mapped to the hg19 reference genome, and filtered using the HiCUP pipeline [62]. Individual library statistics are presented in Additional file 2: Table S1. Significant interactions were identified using the CHiCAGO pipeline [23]. For both captures, replicate libraries for each cell line were analyzed separately to learn weights which were then used to merge replicates into a single dataset per cell type. Interactions with CHiCAGO scores  $\geq 5$  in at least one cell type were considered high-confidence interactions.

#### Principal component and cluster analyses

Principal component analysis (PCA) of the CHiCAGO interaction scores was performed for both variant and promoter capture arrays for each individual biological replicate. Interaction length  $< 2$  Mb and CHiCAGO score  $> 0$  were included. PCA was performed using the R utility *prcomp* with unit variance scaling. Hierarchical clustering with average linkage based on Euclidian distances was performed on the 1000 interactions with most variance using R's *heatmap.2* function. Cell types were clustered based on profiles including interactions with CHiCAGO score  $\geq 5$  and length  $< 2$  Mb.

Interactions with score  $\geq 5$  in at least one cell line were considered.

#### PChi-C and VChi-C concordance

To examine the overall concordance between promoter and variant captures, we identified interactions common to both experiments from the full range of CHiCAGO scores ( $> 0$ ) for each cell type. The Pearson correlation between the CHiCAGO scores for interactions from each of the captures was computed. Interaction scores for each capture were plotted after inverse hyperbolic sine (*asinh*) transformation with Loess-smoothed regression lines.

#### Enrichment of genomic features within interacting regions

Positions of genomic features including DNase-seq peak, histone modification ChIP-seq peaks, transcription factor ChIP-seq peaks (web links provided in Additional file 2: Table S16), and ATAC-seq peaks were intersected with PIRs from each cell line. Enrichment was estimated by comparing to a set of background PIRs generated by maintaining the distribution of interaction distances and interaction counts relative to promoter baits for each cell type. Interactions were grouped in 50-kb distance bins, and 100 sets of random PIR sets were built for each cell line. We removed the baited fragments from the pool of possible PIRs. *Z* scores were calculated for each genomic annotation (Additional file 2: Tables S4, S5).

#### Fine-mapping of chromatin contacts

PChi-C and VChi-C contact mapping was performed using the *Peakpy* Bioconductor package [32]. We first pooled the aligned reads from replicate Chi-C libraries. Probable interaction-driving contacts were then modeled for each bait from each cell line independently. We maintained the default  $\Omega$  value (5) for each bait. Two parallel chains were run, and correlation between MPPC values for interacting prey fragments was tested until  $r > 0.75$  (typically after  $20^6$  iterations). We achieved successful convergence for  $> 93\%$  tested baits. Distributions derived from parallel chains were then merged to generate cell type- and bait-specific contact maps. An arbitrary MPPC threshold of 0.1 was used for downstream analysis.

#### Expression quantitative trait loci analysis

To determine whether eSNP-target gene pairs were overrepresented within captured interactions, we assigned interactions to random promoters within the same chromosome. This randomization procedure was repeated 10,000 times. The frequency of eSNP-gene occurrences within interactions was then tallied in the

observed interaction set and compared to random expectation.

#### ATAC-seq library preparation and data analysis

ATAC-seq was performed as previously described [71]. Briefly,  $5 \times 10^4$  cells were resuspended in lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl<sub>2</sub>, and 0.1% (vol/vol) IGEPAL CA-630), then centrifuged at 5000×g for 10 min at 4 °C. Pellets were resuspended in TD buffer (10 mM Tris (pH 7.6), 5 mM MgCl<sub>2</sub>, 10% (vol/vol) dimethylformamide), and 2.5 µl of TDE1 enzyme (Illumina). Transposed fragments were purified using a MinElute PCR purification kit (QIAGEN), then amplified and indexed with unique library indices using NEBNext High-Fidelity 2× PCR Master Mix (New England BioLabs). PCR products were purified with AMPure XP beads (Beckman-Coulter) and quantified with a Qubit dsDNA High-Sensitivity Assay kit (Thermo Fisher Scientific) and BioAnalyzer High Sensitivity DNA Kit (Agilent Technologies). Pools of six libraries were sequenced per lane on an Illumina HiSeq 2500 (Kinghorn Centre for Clinical Genomics). Raw sequencing reads were trimmed for adapter sequences using Cutadapt (version 1.9 [64]); and aligned using BWA-MEM (version 0.7.12 [72]); to the GRCh37 assembly. The aligned reads were coordinate sorted using Samtools (version 1.1 [65]); and duplicate alignments were marked with Picard (version 1.129). Qprofiler assessed the sequence quality and provide fragment length distribution. Peaks were called for each sample using MACS2 [66]. Peak annotation was performed using HOMER [67].

#### 3C validation

3C libraries were generated using *HindIII* as described previously [73]. 3C interactions were quantified by real-time PCR (qPCR) using primers designed within restriction fragments (Additional file 2: Table S15). qPCR was performed on a RotorGene 6000 using MyTaq HS DNA polymerase (Bioline) with the addition of 25 µM Syto9, annealing temperature of 66 °C, and extension time of 30 s. 3C analyses were performed in two independent 3C libraries from each cell line quantified in duplicate. BAC clones covering each region were used to create artificial libraries of ligation products to normalize for PCR efficiency. Data were normalized to the signal from the BAC clone library and, between cell lines, by reference to a region within *GAPDH*.

#### CRISPR/Cas9 interference and cutting

For CRISPR interference (CRISPRi), the sgRNA targets (listed in Additional file 2: Table S15), Cas9 binding handle and terminator sequences were synthesized (Integrated DNA Technologies, IDT) 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 Scientific). Cells were cotransfected with the packaging plasmid pCMV-dR8.91, the VSV-G envelope expression plasmid pCMV-VSV-G, and with either pHR-SFFV-dCas9-BFP-KRAB or pgRNA-humanized. VLPs were collected from culture supernatants, mixed in equal volume, and transduced into MCF7 cells. Cells expressing both mCherry (via pgRNA-humanized) and blue fluorescent protein (via dCas9-KRAB) were isolated by FACS on an ARIA IIIu (Becton-Dickinson). For CRISPR cutting (CRISPRc), the GFP control and sgRNA targets (listed in Additional file 2: Table S15) were synthesized (IDT) and cloned into the pXPR\_011 lentiviral vector. Virus-like particles (VLPs) containing the GFP control or targeting sgRNAs were generated by the transfection of HEK293 cells with FuGene (Promega). VLPs were collected from culture supernatant, transduced into HMLE-Cas9 cells, and selected using puromycin for at least 48 h.

#### Quantitative real-time PCR

Complementary DNA (cDNA) was synthesized from RNA samples using SuperScript III (Invitrogen). qPCR was performed using TaqMan assays (Thermo Fisher Scientific; listed in Additional file 2: Table S15).

#### Plasmid construction and reporter assays

The *TBX3* promoter-driven luciferase construct was generated by insertion of a PCR-amplified promoter fragment into the *NheI* and *HindIII* sites of the pGL3-basic vector (primers are listed in Additional file 2: Table S15). The 12q24 signal 1 enhancer, containing either the risk or protective CCV alleles, was synthesized as gBlocks (IDT) and cloned into the *BamHI* and *Sall* sites of the *TBX3* promoter construct (coordinates are listed in Additional file 2: Table S15). Sanger sequencing of all constructs confirmed variant incorporation. MCF7 cells were transfected with equimolar amounts of luciferase reporter plasmids and pRL-TK transfection control plasmid with Lipofectamine 3000 (Thermo Fisher Scientific). Luciferase activity was measured 24 h post-transfection by the Dual-Glo Luciferase Assay System (Promega). To correct for any differences in transfection efficiency or cell lysate preparation, *Firefly* luciferase activity was normalized to *Renilla* luciferase activity, and the activity of each construct was expressed relative to the reference promoter constructs, which were defined to have an activity of 1.

#### Electromobility shift assays

Gel shift assays were performed with MCF7 or BT474 nuclear lysates and biotinylated oligonucleotide duplexes (listed in Additional file 2: Table S15). Nuclear lysates

were prepared using the NE-PER nuclear and cytoplasmic extraction reagents (Thermo Fisher Scientific) as per the manufacturer's instructions. Total protein concentrations in nuclear lysates were determined by Bradford's method. Duplexes were prepared by combining sense and antisense oligonucleotides in NEBuffer2 (New England Biolabs) and heat annealing at 80 °C for 10 min followed by slow cooling to 25 °C for 1 h. Binding reactions were performed in binding buffer (10% (vol/vol) glycerol, 20 mM HEPES (pH 7.4), 1 mM DTT, protease inhibitor cocktail (Roche), and 0.75 µg poly(dI:dC) (Sigma-Aldrich)) with 7.5 µg of nuclear lysate. For competition assays, binding reactions were pre-incubated with 1 pmol of competitor duplex (competitor sequences are listed in Additional file 2: Table S15) at 25 °C for 10 min before the addition of 10 fmol of biotinylated duplex and incubation at 25 °C for 15 min. Reactions were separated on 10% (wt/vol) Tris-Borate-EDTA (TBE) polyacrylamide gels (Bio-Rad) in TBE buffer at 160 V for 40 min. Duplex-bound complexes were transferred onto Zeta-Probe positively charged nylon membranes (Bio-Rad) by semi-dry transfer at 25 V for 20 min, then cross-linked onto the membranes under 254 nm ultra-violet light for 10 min. Membranes were processed with the LightShift Chemiluminescent EMSA kit (Thermo Fisher Scientific) as per the manufacturer's instructions. Chemiluminescent signals were visualized with the C-DiGit blot scanner (LI-COR).

#### Chromatin immunoprecipitation

MCF7 cells were cross-linked with 1% (wt/vol) formaldehyde at 37 °C for 10 min, rinsed once with ice-cold PBS containing 5% (wt/vol) BSA and once with PBS, and harvested in PBS containing protease inhibitor cocktail (Roche). Harvested cells were centrifuged for 2 min at 3000 rpm. Cell pellets were resuspended in 0.35 ml of lysis buffer (1% (wt/vol) SDS, 10 mM EDTA, and 50 mM Tris-HCl (pH 8.1)), protease inhibitor cocktail and sonicated three times for 15 s at 70% duty cycle (Branson SLPt) followed by centrifugation at 13000 rpm for 15 min. Supernatants were collected and diluted in dilution buffer (1% (wt/vol) Triton X-100, 2 mM EDTA, 150 mM NaCl, and 20 mM Tris-HCl (pH 8.1)). Two micrograms of anti-GATA3 antibody (Santa Cruz) or control IgG (Santa Cruz) was prebound for 6 h to protein G Dynabeads (Thermo Fisher Scientific) and then added to the diluted chromatin for overnight immunoprecipitation. The magnetic bead-chromatin complexes were collected and washed six times in RIPA buffer (50 mM HEPES (pH 7.6), 1 mM EDTA, 0.7% (vol/vol) sodium deoxycholate, 1% (vol/vol) NP-40, 0.5 M LiCl), then twice with TE buffer. To reverse cross-linking, the magnetic bead complexes were incubated overnight at 65 °C in elution buffer (1% (wt/vol) SDS, 0.1 M NaHCO<sub>3</sub>). DNA fragments

were purified using the QIAquick Spin Kit (QIAGEN). For qPCR (primers are listed in Additional file 2: Table S15), 2 µl from a 100-µl immunoprecipitated chromatin extraction was amplified for 40 cycles. All PCR products were sequenced by Sanger sequencing.

#### TBX3 overexpression

The TBX3 overexpression construct (pLX307/TBX3) was generated by Gateway cloning from pDONR201 containing the full-length TBX3 cDNA into the pLEX\_307 lentiviral destination vector (Thermo Fisher Scientific). A negative control construct (pLX307/CON) was generated by excising TBX3 via *NheI* and *SpeI* restriction enzyme digestion and self-ligating the vector backbone. VLPs were generated from HEK293 cells transfected with pLX307/CON or pLX307/TBX3 as described above and transduced into HMLE or MCF7 cells. Transductants were selected with puromycin for at least 48 h.

#### Western blotting

Cell pellets were lysed in RIPA buffer (50 mM Tris-HCl (pH 8.0), 150 mM NaCl; 1% (vol/vol) IGEPAL CA-630, 0.5% (vol/vol) sodium deoxycholate, 0.1% (wt/vol) SDS, 1 mM DTT, protease inhibitor cocktail) and clarified by centrifugation to remove cell debris. Forty micrograms of lysate supernatants was separated by SDS-polyacrylamide gel electrophoresis, electroblotted onto PVDF membranes by semi-dry transfer (Bio-Rad), and blocked in blocking buffer (1% (wt/vol) casein, 0.1% (vol/vol) Tween 20, PBS). TBX3 was detected with 1 µg/ml rabbit anti-TBX3 antibody (Thermo Fisher Scientific) and actin with 400 ng/ml of rabbit anti-actin antibody (Sigma-Aldrich). Primary antibodies were detected with horseradish peroxidase-conjugated goat anti-rabbit IgG (Cell Signaling). Detected proteins were visualized with enhanced chemiluminescence substrate (Bio-Rad) and the G:BOX Chemi XX6 gel documentation system (Syngene).

#### Soft agar colony formation assay

Six-well plates were layered with 0.6% (wt/vol) noble agar (Becton-Dickinson) in RPMI or DMEM medium supplemented with 10% (vol/vol) FBS and antibiotics and allowed to set at 4 °C. Twenty-four hours later, the cells were trypsinized and  $8 \times 10^3$  MCF7 or  $5 \times 10^4$  HMLE cells were resuspended in 0.3% (wt/vol) noble agar and plated on top of bottom agar layers (three wells/cell line). Colonies were imaged after 3–4 weeks using a Leica MZ FLIII stereo microscope. Visible colonies > 1 mm in diameter were counted.



### Cell proliferation assay

Cell proliferation was measured using a label-free, non-invasive cellular confluence assay on the IncuCyte Live-Cell Imaging System (Essen Bioscience). MCF7 cells were seeded at 20,000 cells/well into 24-well plates and imaged on the IncuCyte using a  $\times 10$  objective lens every 3 h over 7 days. Imaging was performed in an incubator maintained at 37 °C under a 5% CO<sub>2</sub> atmosphere. Cell confluence in each well was measured using IncuCyte ZOOM 2016A software, and the data analyzed using GraphPad Prism.

### Mouse tumor xenograft model

A cholesterol-based pellet containing 17 $\beta$ -estradiol (0.72 mg, 90-day slow release, Innovative Research of America) was implanted subcutaneously in the interscapular region of 8-week-old female BALB/c-Foxn1<sup>tm</sup>/Arc mice. Three days later, MCF7 CRISPRi-suppressed cells ( $1 \times 10^7$  cells/mouse) were injected into the fourth right mammary fat pad (eight to nine mice per cell line). Tumor volumes were measured with a digital caliper every second day until the experimental end stage approved by the QIMR Berghofer animal ethics committee, 525 mm<sup>3</sup> according to the formula ( $\pi \times \text{length} \times \text{width}^2/6$ ).

### Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s13059-019-1877-y>.

**Additional file 1.** Supplementary figures.

**Additional file 2.** Supplementary tables.

### Acknowledgements

The results published here are in part based upon the data generated by the TCGA Research Network.

### Authors' contributions

JDF and SLE contributed to the conceptualization. JB, MMM, PM, SK, KM, DRB, and LF contributed to the bioinformatic and statistical analyses. HS, LGL, NT, KMH, SKaufmann, NH, SH, JSL, and KN contributed to the functional analyses. ACA, AMD, DFE, NW, JR, AM, and GCT contributed to the supervision. JB, JDF, and SLE contributed to the writing, with contributions from all authors. All authors read and approved the final manuscript.

### Funding

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC; 1058415 and 1120563), Cancer Council Queensland (1099810), and Perpetual IMPACT Program (IPAP2017/1497). SLE and NW are NHMRC Senior Research Fellows (1135932 and 1139071). GCT is an NHMRC Senior Principle Research Fellow (1117073). JDF was supported by a Fellowship from the National Breast Cancer Foundation of Australia. NA was co-funded by a QIMR Berghofer International PhD Scholarship and a University of Queensland Research Training Scholarship. This project has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie Individual Fellowships program under grant agreement no. MSCA-IF-2014-EF-656144.

### Ethics approval and consent to participate

All animal procedures were conducted in accordance with the Australian National Health and Medical Research regulations on the use and care of

experimental animals and approved by the QIMR Berghofer Medical Research Institute Animal Ethics Committee (A12617M, P1499).

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Cancer Program, QIMR Berghofer Medical Research Institute, Brisbane, Australia. <sup>2</sup>Current address: UK Dementia Research Institute, Imperial College London, London, UK. <sup>3</sup>Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia. <sup>4</sup>Faculty of Medicine, The University of Queensland, Brisbane, Australia. <sup>5</sup>Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. <sup>6</sup>Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. <sup>7</sup>Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.

Received: 11 March 2019 Accepted: 1 November 2019

### References

- Melchor L, Benitez J. The complex genetic landscape of familial breast cancer. *Hum Genet.* 2013;132:845–63.
- Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature.* 2017; 551:92–4.
- Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst.* 2015;107:1–15.
- Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWAS: illuminating the dark road from association to function. *Am J Hum Genet.* 2013;93:779–97.
- 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.* (in press).
- Cavalli G, Misteli T. Functional implications of genome topology. *Nat Struct Mol Biol.* 2013;20:290–9.
- Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet.* 2016;17:772.
- Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature.* 2012;485:381–5.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012;485:376–80.
- Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using chromosome conformation capture. *Methods.* 2012;58: 192–203.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozcy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009;326:289–93.
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 2014;159:1665–80.
- Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature.* 2013; 503:290–4.
- Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* 2014;24:1854–68.
- Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* 2015;25:582–97.
- Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell.* 2016;167:1369–84.
- Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet.* 2017;49: 1522–8.

18. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution Capture Hi-C. *Nat Genet.* 2015;47:598–606.
19. Davies JO, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods.* 2016;13:74–80.
20. Siersbaek R, Madsen JGS, Javierre BM, Nielsen R, Bagge EK, Cairns J, et al. Dynamic rewiring of promoter-anchored chromatin loops during adipocyte differentiation. *Mol Cell.* 2017;66:420–35.
21. McGovern A, Schoenfelder S, Martin P, Massey J, Duffus K, Plant D, et al. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* 2016;17:212.
22. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun.* 2015;6:10069.
23. Cairns J, Freire-Pritchett P, Wingett SW, Varnai C, Dimond A, Pagnon V, et al. CHICAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* 2016;17:127.
24. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell.* 2006;10:515–27.
25. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilienky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518:317–30.
26. Cowper-Sal Lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoutte J, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet.* 2012;44:1191–8.
27. Theodorou V, Stark R, Menon S, Carroll JS. GATA3 acts upstream of FOXA1 in mediating ESRI binding by shaping enhancer accessibility. *Genome Res.* 2013;23:12–22.
28. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature.* 2017; 544:59–64.
29. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature.* 2012;486:346–52.
30. Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, et al. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* 2014;28:2778–91.
31. Rosa A, Becker NB, Everaers R. Looping probabilities in model interphase chromosomes. *Biophys J.* 2010;98:2410–9.
32. Eijbsbouts CQ, Burren OS, Newcombe PJ, Wallace C. Fine mapping chromatin contacts in capture Hi-C data. *BMC Genomics.* 2019;20:77.
33. Conover CA. Key questions and answers about pregnancy-associated plasma protein-A. *Trends Endocrinol Metab.* 2012;23:242–9.
34. Mansfield AS, Visscher DW, Hart SN, Wang C, Goetz MP, Oxvig C, et al. Pregnancy-associated plasma protein-A expression in human breast cancer. *Growth Hormon IGF Res.* 2014;24:264–7.
35. Takabatake Y, Oxvig C, Nagi C, Adelson K, Jaffer S, Schmidt H, et al. Lactation opposes pappalysin-1-driven pregnancy-associated breast cancer. *EMBO Mol Med.* 2016;8:388–406.
36. Visvader JE, Venter D, Hahm K, Santamaria M, Sum EY, O'Reilly L, et al. The LIM domain gene LMO4 inhibits differentiation of mammary epithelial cells in vitro and is overexpressed in breast cancer. *Proc Natl Acad Sci U S A.* 2001;98:14452–7.
37. Sum EY, Segara D, Duscio B, Bath ML, Field AS, Sutherland RL, et al. Overexpression of LMO4 induces mammary hyperplasia, promotes cell invasion, and is a predictor of poor outcome in breast cancer. *Proc Natl Acad Sci U S A.* 2005;102:7659–64.
38. Hannenhalli S, Kaestner KH. The evolution of Fox genes and their role in development and disease. *Nat Rev Genet.* 2009;10:233–40.
39. Mani SA, Yang J, Brooks M, Schwanger G, Zhou A, Miura N, et al. Mesenchyme Forkhead 1 (FOXO2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers. *Proc Natl Acad Sci U S A.* 2007;104:10069–74.
40. Zhong J, Wang H, Yu J, Zhang J, Wang H. Overexpression of forkhead box L1 (FOXO1) inhibits the proliferation and invasion of breast cancer cells. *Oncol Res.* 2017;25:959–65.
41. MacNair L, Xiao S, Miletic D, Ghani M, Julien JP, Keith J, et al. MTHFS and DDX58 are novel RNA-binding proteins abnormally regulated in amyotrophic lateral sclerosis. *Brain.* 2016;139:86–100.
42. 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 ESRI, RMND1 and CCDC170. *Nat Genet.* 2016;48:374–86.
43. Bailey SD, Desai K, Kron KJ, Mazrooei P, Sinnott-Armstrong NA, Treloar AE, et al. Noncoding somatic and inherited single-nucleotide variants converge to promote ESRI expression in breast cancer. *Nat Genet.* 2016;48:1260–6.
44. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature.* 2016;534:47–54.
45. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57–74.
46. Guertin MJ, Zhang X, Coonrod SA, Hager GL. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Mol Endocrinol.* 2014;28:1522–33.
47. Willmer T, Cooper A, Peres J, Omar R, Prince S. The T-Box transcription factor 3 in development and cancer. *Biosci Trends.* 2017;11:254–66.
48. Elenbaas B, Spirio L, Koerner F, Fleming MD, Zimonjic DB, Donaher JL, et al. Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* 2001;15:50–65.
49. Willmer T, Cooper A, Sims D, Govender D, Prince S. The T-box transcription factor 3 is a promising biomarker and a key regulator of the oncogenic phenotype of a diverse range of sarcoma subtypes. *Oncogenesis.* 2016;5:e199.
50. Meyer KB, O'Reilly M, Michailidou K, Carlebur S, Edwards SL, French JD, et al. Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am J Hum Genet.* 2013;93:1046–60.
51. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature.* 2007;446: 153–8.
52. Glubb DM, Maranian MJ, Michailidou K, Pooley KA, Meyer KB, Kar S, et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am J Hum Genet.* 2015;96:5–20.
53. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat Commun.* 2018;9:1028.
54. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 2012;337:1190–5.
55. Fischer K, Pflugfelder GO. Putative breast cancer driver mutations in TBX3 cause impaired transcriptional repression. *Front Oncol.* 2015;5:244.
56. Shen L, Shi Q, Wang W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis.* 2018;7:25.
57. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res.* 2018;46:D794–801.
58. Cheneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 2018;46:D267–75.
59. Zhang G, Zhao Y, Liu Y, Kao LP, Wang X, Skerry B, et al. FOXA1 defines cancer cell specificity. *Sci Adv.* 2016;2:e1501473.
60. Rhie SK, Hazelett DJ, Coetzee SG, Yan C, Noushmehr H, Coetzee GA. Nucleosome positioning and histone modifications define relationships between regulatory elements and nearby gene expression in breast epithelial cells. *BMC Genomics.* 2014;15:331.
61. Barutcu AR, Lajoie BR, McCord RP, Tye CE, Hong D, Messier TL, et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.* 2015;16:214.
62. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* 2015;4:1310.
63. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29:24–6.
64. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal.* 2011;17:10–2.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25: 2078–9.

66. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9:R137.
67. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38:576–89.
68. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 2018;46:D1284.
69. Toouli CD, Huschtscha LI, Neumann AA, Noble JR, Colgin LM, Hukku B, et al. Comparison of human mammary epithelial cells immortalized by simian virus 40 T-antigen or by the telomerase catalytic subunit. *Oncogene.* 2002; 21:128–39.
70. Nagano T, Varnai C, Schoenfelder S, Javierre BM, Wingett SW, Fraser P. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.* 2015;16:175.
71. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods.* 2013;10:1213–8.
72. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26:589–95.
73. Ghousaini M, Edwards SL, Michailidou K, Nord S, Cowper-Sal Lari R, Desai K, et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat Commun.* 2014;4:4999.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

