Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

Research Article

Cancer Epidemiology, Biomarkers & Prevention

MCF-7 as a Model for Functional Analysis of Breast Cancer Risk Variants

Alix Booms, Gerhard A. Coetzee, and Steven E. Pierce

Abstract

Background: Breast cancer genetic predisposition is governed by more than 142 loci as revealed by genome-wide association studies (GWAS). The functional contribution of these risk loci to breast cancer remains unclear, and additional post-GWAS analyses are required.

Methods: We identified active regulatory elements (enhancers, promoters, and chromatin organizing elements) by histone H3K27 acetylation and CTCF occupancy and determined the enrichment of risk variants at these sites. We compared these results with previously published data and for other cell lines, including human mammary epithelial cells, and related these data to gene expression.

Results: In terms of mapping accuracy and resolution, our data augment previous annotations of the MCF-7 epigenome. After intersection with GWAS risk variants, we found 39 enhancers and 15 CTCF occupancy sites that, between them, overlapped 96 breast cancer credible risk variants at 42 loci. These risk enhancers likely regulate the expression of dozens of genes, which are enriched for GO categories, including estrogen and prolactin signaling.

Conclusions: Ten (of 142) breast cancer risk loci likely function via enhancers that are active in MCF-7 and are well suited to targeted manipulation in this system. In contrast, risk loci cannot be mapped to specific CTCF-binding sites, and the genes linked to risk CTCF sites did not show functional enrichment. The identity of risk enhancers and their associated genes suggests that some risk may function during later processes in cancer progression.

Impact: Here, we report how the ER$^+$ cell line MCF-7 can be used to dissect risk mechanisms for breast cancer.

Introduction

According to the traditional model of carcinogenesis, normal tissue undergoes rounds of oncogenic mutations during proliferation that eventually leads to metastatic tumor growth. An individual's inherited genetic predisposition for cancer, which can be measured by linkage analysis or genome-wide association studies (GWAS), influences the rate of those oncogenic somatic mutations and the environment in which the tumors develop (1). In breast cancer, rare but high-penetrance inherited mutations to genes, such as *BRCA1* and *BRCA2*, contribute about 30% to the familial risk of developing breast cancer and, in general, have well-understood biological consequences (2). However, only about 5% to 10% of breast cancer cases are actually associated with this type of germline mutation. In contrast, GWAS show that low-penetrance but common genetic variants explain up to 50% of disease heritability and contribute significant risk to the development of both familial and sporadic breast cancer (3). Unlike high-penetrance rare mutations, in most cases, these other risk genotypes are poorly understood and do not alter protein coding. Elucidating the functional basis of these common risk variants is therefore of great importance.

A recent GWAS uncovered 65 new breast cancer risk loci contributing to a total of roughly 142 reproducible breast cancer risk loci containing over 38,000 statistically significant (combined *P* value $< 5 \times 10^{-8}$) or near-significant (combined *P* value $< 1 \times 10^{-5}$) risk variants (4). These variants primarily consist of SNPs (we use "variant" and "SNP" interchangeably in this text). Due to the large number of risk-associated SNPs as well as to their presence in noncoding DNA, it can be difficult to pinpoint which SNP or combination of SNPs are causal for a disease, let alone explain the biological mechanisms/genes involved. This will only become more challenging as the number of identified risk SNPs is expected to increase as the sizes of case–control studies grow larger in the near future. Our goal is therefore to prioritize SNPs based on their potential effects on cell-type–specific genomic activity.

Risk SNPs are nonrandomly distributed throughout the genome and have been shown to be enriched in tissue-specific noncoding regulatory elements (RE), mainly in enhancers (5–7). REs are defined as regions of noncoding DNA that regulate the transcription of genes. Enhancers are a class of RE that influence cell fate and development through coordinated interaction between transcription factors (TF) and their target promoters to alter the transcription of genes (8). Enhancers are marked by surrounding histone modifications and nucleosome depletion. The histone modification most often used is H3K27 acetylation, because it has been shown that it marks active (engaged) enhancers (9). Enhancers are also relatively transient, and their activity is dictated by extracellular signals that activate complex enhancer networks to promote cell-type– and condition-specific gene expression. Common breast cancer risk variants present in gene

Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, Michigan.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

G.A. Coetzee and S.E. Pierce codirected this work.

Corresponding Author: Gerhard A. Coetzee, Van Andel Research Institute, 333 Bostwick Avenue N.E., Grand Rapids, MI 49503. Phone: 646-696-1560; E-mail: gerry.coetzee@vai.org

Cancer Epidemiol Biomarkers Prev 2019;28:1735–45

doi: 10.1158/1055-9965.EPI-19-0066

REs are likely to subtly alter gene expression patterns, compared with protective alleles, in ways that predispose some individuals to cancer. For example, studies of the noncoding region 8q24 upstream of *MYC* have shown enrichment of variants associated with different tumor types, including breast, in multiple enhancers at this location (10–12).

Another important feature shown to coincide with risk SNPs is CTCF (CCCTC binding-factor). CTCF is a highly conserved 11-zinc finger protein and is the only known insulator protein in vertebrates (13). It is most often enriched at loop and topologically associated domain (TAD) boundaries that separate transcriptionally active and repressed genes (14). This implicates its importance in the organization of chromatin compartments to prevent aberrant gene–enhancer interactions. Lupianez and colleagues (15) showed that removing the CTCF-associated boundary elements at the Epha4 TAD causes abnormal interactions between adjacent TADs and expression of genes within those TADs. Others have shown that CTCF is important for spatial organization of the genome for proper induction and silencing of transcription (14, 16). Although the relationship between SNPs and CTCF is less clear than that of SNPs within enhancers, SNP-mediated disruptions of CTCF consensus sequences have been shown to change the binding affinity of the CTCF protein (17). This could result in gene spatial network rewiring that has a strong influence on gene expression changes that indirectly lead to breast cancer. Such rewiring was demonstrated by a CRISPR-mediated deletion of prostate cancer risk–associated CTCF sites, which identified repressive chromatin loops [17].

In the present study, we used the MCF-7 cell line as a model of breast cancer. It is one of the most utilized breast cancer cell lines in cancer research due to its long history and expression of the estrogen receptor (ER; ref. 18). The majority of patients (72.7%) with a known HR/HER status (ER + progesterone receptor/human epidermal growth factor receptor) are HR$^+$/HER2$^-$ (19). MCF-7 cells are one of the few breast cancer cell lines that is also HR$^+$/HER2$^-$, making it an extremely relevant model for the study of invasive breast cancer. Even, so there is currently a lack in the literature of post-GWAS analyses based on experimental manipulation in ER$^+$ systems, including MCF-7 (20). Although a detailed meta-analysis of breast cancer risk loci was recently conducted and credible risk variants (CRV; see Materials and Methods for a detailed definition) genome-wide were defined (4), potential functionality was only alluded to in broad terms. Here, we analyze epidemiologically defined genetic risk loci within the MCF-7 cancer cell line. In so doing, we identified functional targets that can further be tested by genetic manipulation. We report a priority list of breast cancer risk enhancers and risk CTCF-binding sites tailored to the MCF-7 cell line.

Materials and Methods

Cell culture

MCF-7 cells were obtained from the ATCC and cultured in DMEM (ATCC, cat # 30-2003) supplemented with 10% FBS and 0.01 mg/mL human recombinant insulin (Gibco, ref # 12585-014). They were incubated in a humidified 37°C, 5% CO_2 incubator. For routine passaging, cells were grown in T25 and T75 culture flasks and passaged using 0.25% Trypsin/EDTA.

Chromatin immunoprecipitation sequencing

For chromatin immunoprecipitation (ChIP), we followed previously published protocols from Rhie and colleagues (21) with slight modifications. Roughly 30 to 40 $\times$ 10^6 cells were used per ChIP. Upon reaching 70% to 80% confluency, cells were directly fixed in T75 flasks by adding 16% formaldehyde to the culture medium to a final concentration of 1%. The reaction was quenched for 5 minutes at room temperature with 10X (1.15 mol/L) glycine. Using a Bioruptor Pico (Diagenode, Cat # B01060001), the isolated chromatin was sonicated for 30-second on and 30-second off cycles to yield DNA fragments between 200 and 500 base pairs. Note that 100 μg of sonicated chromatin was used for immunoprecipitation, and 1 μg (1%) was used for the input control. To probe for CTCF or H3K27ac, samples were incubated at 4°C overnight with a primary antibody [CTCF: Cell Signaling monoclonal (D31H2), Cat# 3418; H3K27ac: Active Motif, Cat #39133] or an IgG control (Sigma, Cat # R9133). A/G magnetic beads (Pierce, Cat # 88802) were then incubated with the samples for 2 hours at 4°C. Following this incubation, the beads were washed with a series of buffers of varying salt concentrations before overnight elution at 67°C. The ChIP, IgG, and Input samples were all purified using a QIAprep Spin Miniprep Kit (Qiagen, Cat # 27104).

Construction and sequencing of ChIP-Seq libraries

Libraries for input and IP samples were prepared by the Van Andel Genomics Core from 10 ng of input material and all available IP material using the KAPA Hyper Prep Kit (v5.16; Kapa Biosystems). Prior to PCR amplification, end-repaired and polyadenylated DNA fragments were ligated to Bio Scientific NEXTflex Adapters (Bio Scientific). The quality and quantity of the finished libraries were assessed using a combination of Agilent DNA High Sensitivity chip (Agilent Technologies, Inc.), QuantiFluor dsDNA System (Promega Corp.), and Kapa Illumina Library Quantification qPCR assays (Kapa Biosystems). Sequencing (75 bp, single end) was performed on an Illumina NextSeq 500 sequencer using a 75-bp sequencing kit (v2; Illumina Inc.). Base calling used Illumina NextSeq Control Software (NCS) v2.0, and the output of NCS was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v1.9.0.

Identification of ChIP-Seq peaks

Two biological replicates of MCF-7 were used for input and ChIP for H3K27Ac and separately for CTCF. Following sequencing, fastq files were aligned to the HG19 genome assembly using default setting for BWA v0.7.15 (22). Mapped sequencing depth was 40 to 50 million reads ChIP, approximately 110 million reads input for CTCF and 60 to 70 million reads ChIP, approximately 150 million reads input for H3K27ac. Aligned reads were called using MACS2 v2.1 at a liberal FDR cutoff of 0.1 (23). For comparison of our CTCF data with the ENCODE CTCF data [ENCODE1 = ENCSR000DMV, ENCODE2 = ENCSR560BUE, primary antibody: Millipore polyclonal (07-729); ENCODE3 = ENCSR000DWH primary antibody: Cell Signaling (2899)], analysis started with fastq files and were aligned with BWA as above, except: reads less than 50 bp in our data and reads less than 20 bp in the ENCODE data were removed using Trim Galore prior to peak calling. Peaks were filtered by irreproducible discovery rate (IDR; V2.0.2) according to default settings. For H3K27Ac, Trim Galore was not used, and ENCODE comparison was done starting with peaks files (also generated using BWA and MACS2). Existing

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

data for MCF-7 H3K27ac ChIP-Seq ENCODE1 (ENCSR000EWR, read length: 32 bp, ChIP depth = ~20 million reads, input = ~10 million reads), ENCODE2 (ENCSR752UOD, read length: 36 bp, ChIP depth: rep1, 60 million and rep2, 20 million reads, input: 70–80 million reads), and for human mammary epithelial cell (HMEC; ENCSR000ALW) were used for comparison. NarrowPeak calls (i.e., ENCFF187RUK, ENCFF37ORFF, ENCFF537JMI, and ENCFF208IPB) were filtered by IDR to generate plots in Fig. 1 and for peak coordinates. All H3K27Ac peak data were filtered by IDR (V2.0.2) with the following parameters for peak-merge: –rank P value –soft-idr-threshold 0.01 –peak-merge-method max (24). FRiP was calculated using deepTools (V.3.1.3; ref. 25). Intersected datasets are defined by coverage, not peaks, i.e., every bp annotated in all datasets is included; overlapping peaks are not merged. Prediction for allele-dependent CTCF binding was made using MotifBreakR (V.1.8; ref. 26).

Breast cancer risk SNPs

All SNPs used here were derived from the recent breast cancer GWAS meta-analysis by Michailidou and colleagues (4). Their report provided a set of 11.8 million 1000 Genomes SNPs which were associated with breast cancer and 20,989 SNPs with significant or near-significant (combined $P < 1 \times 10^{-5}$) association values. Across 142 regions, they selected the most significant SNP or SNPs and identified those nearby SNPs within 500 Kb and 2 orders of magnitude significance. These they called CRVs. The risk loci correspond roughly to sets of significant risk SNPs (combined P value $< 5 \times 10^{-8}$) with at least 400 to 500 Kb intervening distance. We combined the sets of CRVs and associated annotations from their Supplementary Tables S2, S6, S8, S11, S13, and S14 to produce a list of 4,453 breast cancer CRVs corresponding to 65 new risk loci and 77 reconfirmed previously published risk loci. Across these regions, we also selected the most significant SNPs per locus and used RaggR (27) to obtain the set of 8,687 phase 3 1000 genomes in LD $R^2 > 0.8$ based on European linkage maps.

Overlap and enrichment of breast cancer risk SNPs in REs

SNP enrichment was obtained using Bedtools v2.26.0 to overlap risk SNP genomic locations (hg19) with RE locations (28). The ratios of overlapping risk SNP/total risk SNPs were compared with the background ratio: either based on the overlap of all 11.8 million background SNPs with association values by Michailidou and colleagues (4) or based on the ratio of only those 11.8 million background SNPs within 1 Mb of each risk locus (1.25 million SNPs). The hypergeometric distribution was used to obtain significance values of overall overlap frequencies (Supplementary Table S2).

These enrichment ratios are depicted in Fig. 2B–D and are based on the overlap of all breast cancer risk variants with all enhancers, compared with a single background set. This is useful for determining in general whether the entire set of risk SNPs is related to a specific type of genomic element or to tissue-specific activity. The metric quantifies how relevant any particular model is likely to be for examining risk. Unfortunately, when examining each locus individually, this sort of enrichment calculation can be misleading. This is because, although the background set of SNPs are independent at both the level of genome and single locus, due to genetic linkage, risk SNPs are not independent of each other at the level of a locus. Although multiple significant risk SNPs can be present at a single locus, they should not be treated as independent, but instead represent a single risk signal. As such, an enrichment score for a single locus can be a function of the local LD structure, which we do not expect to be a good indicator of disease relevance. For this reason, the background rate of overlap should be adjusted separately for each locus.

To compare the significance of risk span/RE overlap at each locus separately, the most distant CRVs for each locus were used to define a CRV span, and the proportion of that span, in bp, which overlapped RE peak coverage, was calculated using Bedtools. That value was then ranked against a background distribution of overlap proportions for each of the 142 loci. The background distributions were calculated by permutation testing using R, wherein background SNPs within 1 MB were randomly drawn 10,000 times and used to form the center of a span equal in length to the risk span, and then overlapped with REs. This comparison generates a probability value corresponding to the proportion of the background distribution with an equal or greater amount of overlap as the risk span. The same set of random spans for each locus was used to compare overlap with different tissue or types of REs.

These comparisons were used to generate the heatmap shown in Fig. 3. This was done, in addition to using more common enrichment scores, primarily because we think that the location of individual risk variants within a locus is affected by how risk association is propagated during imputation according to LD related to some single risk element. Thus, these multiple variants represent a single signal. Enrichment calculations or statistical tests that require multiple independent tests are not valid. Randomly drawing individual SNPs from a large set of background SNPs for comparison is likewise not appropriate. Instead we asked: given a span of DNA associated with breast cancer risk, what percentage overlaps an MCF-7 RE and how does that compare with a background of equal-sized spans nearby? We believe that the identification of likely risk region is probably more accurate for small, well-defined REs in inactive regions, than for large REs in regions of high activity. For instance, Supplementary Fig. S2 shows 2 loci with high-risk SNP enrichment but with widely spaced CRVs so that they are characterized by a low span overlap significance and correspond to ambiguity in risk RE identification.

Construction and sequencing of directional mRNA-Seq libraries

Libraries were prepared by the Van Andel Research Institute Genomics Core from 1 μg of material using the KAPA Stranded mRNAseq Kit (v4.16) (Kapa Biosystems). RNA was sheared to 250–300 bp. Prior to PCR amplification, cDNA fragments were ligated to Bio Scientific NEXTflex Adapters (Bioo Scientific). The quality and quantity of the finished libraries were assessed using a combination of Agilent DNA High Sensitivity chip (Agilent Technologies, Inc.), QuantiFluor dsDNA System (Promega Corp.), and Kapa Illumina Library Quantification qPCR assays (Kapa Biosystems).

Differential gene expression analysis

Existing RNA-seq data from ENCODE for MCF-7 (ENCSR000CPT) and HMEC (ENCFF000GDZ) and from NCBI for HMEC (GSE47933) were used for comparison, without alteration. To these, we compared newly generated expression data from 8 WT biological replicates of MCF-7, which were sequenced in two separate experiments. For our experiments, 8 different

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

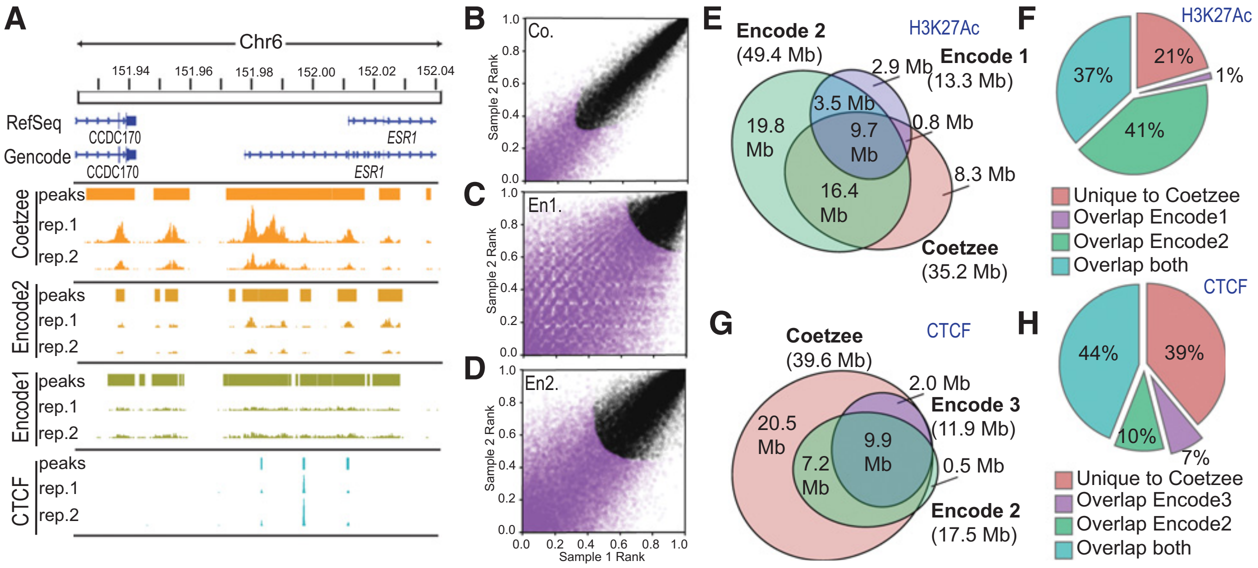
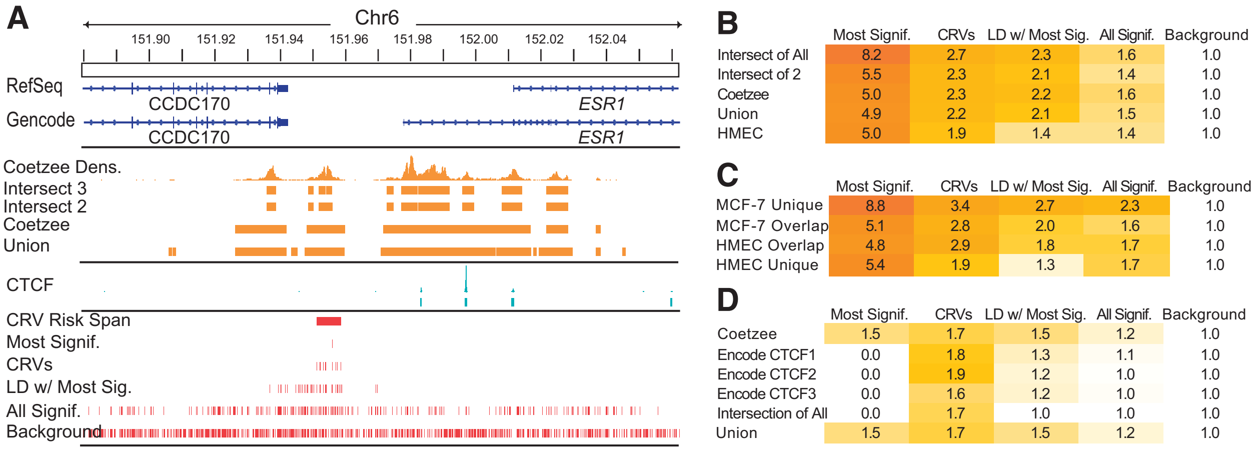


Figure 1.

Comparison of MCF-7 ChIP-Seq data. **A,** Genome browser view of representative locus near *ESR1*. **B,** IDR plot showing the correspondence between replicates 1 and 2 and the threshold for filtering (in black < 0.01) for our H3K27Ac ChIP-Seq (Coetzee; **C**) for ENCODE dataset 2, and (**D**) for ENCODE dataset 1. **E,** Comparison of H3K27Ac ChIP-Seq annotation coverage. **F,** Proportion of IDR-defined H3K27Ac peaks from our (Coetzee) data that overlap peaks in ENCODE1, ENCODE2, both, or neither; by at least 1 bp. **G,** Comparison of CTCF ChIP-Seq annotation coverage. **H,** Proportion of IDR-defined CTCF peaks from our (Coetzee) data that overlap peaks in ENCODE CTCF 2, ENCODE CTCF 3, both, or neither; by at least 1 bp.

MCF-7 WT clones were expanded until reaching 80% confluency in a T25 flask. RNA was isolated using a Qiagen RNeasy mini kit (Cat # 74104). Paired-end mRNA libraries were then prepared by the Van Andel Research Institute Genomics Core as described in the previous section. Following sequencing, fastq files were aligned to HG19 using STAR v2.5 (29). Alignments (bam files) were converted to feature counts using HTSeq v0.6.0 referenced against the ENSEMBLE annotation of HG19: Homo sapiens. GRCh37.87.gtf counting against the feature "exon," grouped by "gene_id," and using the strand parameter "reverse." This set included exon locations for 57,905 genomic entities including pseudogenes, lncRNAs, and 20,356 protein coding genes. The resulting gene_id map counts were normalized using edgeR (TMM) and tested for significant differential expression with Limma and Voom, in R (v3.3.1; refs. 30–32). The normalized count data for all 8 datasets revealed, through principle



B

	Most Signif.	CRVs	LD w/ Most Sig.	All Signif.	Background
Intersect of All	8.2	2.7	2.3	1.6	1.0
Intersect of 2	5.5	2.3	2.1	1.4	1.0
Coetzee	5.0	2.3	2.2	1.6	1.0
Union	4.9	2.2	2.1	1.5	1.0
HMEC	5.0	1.9	1.4	1.4	1.0

C

	Most Signif.	CRVs	LD w/ Most Sig.	All Signif.	Background
MCF-7 Unique	8.8	3.4	2.7	2.3	1.0
MCF-7 Overlap	5.1	2.8	2.0	1.6	1.0
HMEC Overlap	4.8	2.9	1.8	1.7	1.0
HMEC Unique	5.4	1.9	1.3	1.7	1.0

D

	Most Signif.	CRVs	LD w/ Most Sig.	All Signif.	Background
Coetzee	1.5	1.7	1.5	1.2	1.0
Encode CTCF1	0.0	1.8	1.3	1.1	1.0
Encode CTCF2	0.0	1.9	1.2	1.0	1.0
Encode CTCF3	0.0	1.6	1.2	1.0	1.0
Intersection of All	0.0	1.7	1.0	1.0	1.0
Union	1.5	1.7	1.5	1.2	1.0

Figure 2.

Enrichment of breast cancer risk variants in MCF-7 regulatory elements. **A,** Genome browser view of representative locus near *ESR1*, showing in red: the location of risk SNPs and background SNPs, and in orange: the location of H3K27ac peaks, based on our data (Coetzee), our data intersected with ENCODE 2 (Intersect 2), our data intersected with both ENCODE data sets (Intersect 3), or the union of the 3 datasets. **B,** Heat map showing risk the SNP enrichment ratio in each of the 4 MCF-7 datasets of MCF-7 H3K27ac peaks, or in HMEC (which has no H3K27ac at the *ESR1* locus) for each of the sets of breast cancer risk SNPs. **C,** Heat map showing the SNP enrichment ratio for enhancers that are unique to MCF-7, for MCF-7 enhancers that overlap HMEC enhancers, HMEC enhancers that overlap MCF-7 enhancers, or enhancers unique to HMEC. **D,** Heat map showing risk the SNP enrichment ratio for CTCF peaks.

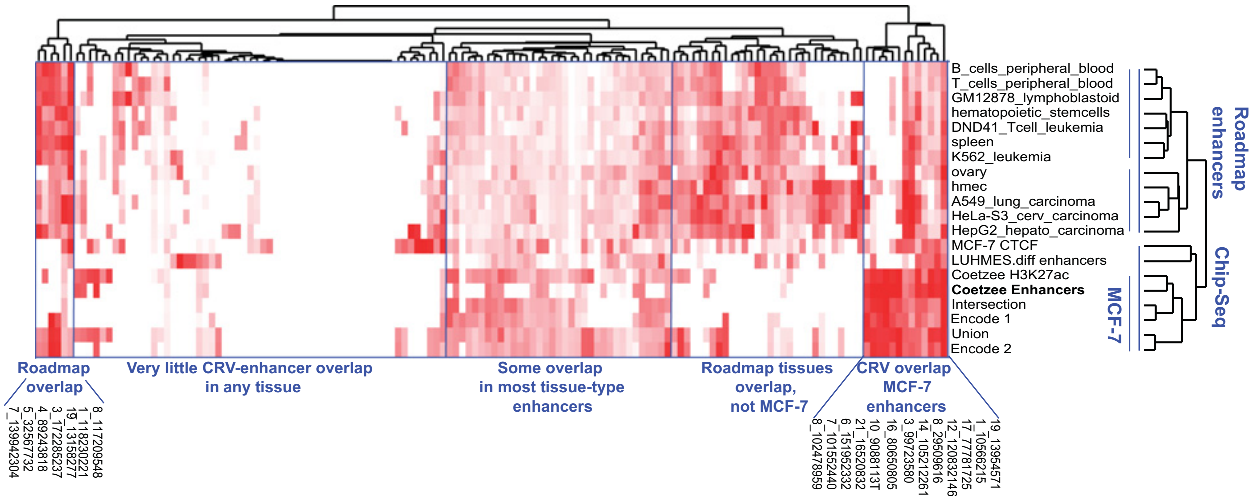


Figure 3.

Analysis of individual risk loci and tissue-specific regulatory elements. ChIP-Seq data and roadmap segmentations (subset for enhancers) were used to examine the significance of overlap with CRV risk spans for each of the 142 breast cancer risk loci (columns). Significance levels are depicted in red with darker red indicating a greater risk-span/RE overlap proportion compared with background levels for that locus. Dendrograms are based on complete linkage using city-block distance measurements of -log(*P* value). The individual loci showing the most significant overlap are identified by chromosome and position underneath. These are the loci with the most specific correspondence between tissue-specific regulatory elements and breast cancer CRVs.

component analysis, that there was very high similarity between our WT MCF-7 which was distinct from the ENCODE data (Supplementary Fig. S3). In total, 57,905 gene transcripts were mapped; however, existing MCF-7 and HMEC RNA-seq datasets only reported expression levels for a subset of these. After merging the expression datasets by gene_id, expression values for 15,810 genes remained.

Gene ontology analysis

Gene ontology enrichment analysis was done using String v.10.5 (33).

Availability of data and material

Publicly available MCF-7 H3K27ac and CTCF ChIP-Seq, and HMEC RNA-Seq datasets used during the current study are available from the encyclopedia of DNA elements (ENCODE; https://www.encodeproject.org/.

All data generated during this study are included in this published article (and its Supplementary information files) or have been deposited in NCBI's Gene Expression Omnibus (34) and are accessible through GEO Series accession number GSE130852.

Ethics approval

Human cell lines used in this study were obtained from the ATCC and do not require additional ethical approval. All other data used were obtained from public sources.

Results

MCF-7 ChIP-Seq and RNA-seq analyses

With the goal of identifying which genetic breast cancer risk variants are functional in the MCF-7 cell line, we first attempted to locate REs that are active in MCF-7 cells using ChIP followed by high-throughput sequencing (ChIP-Seq). ChIP-Seq has become widely used in many cell types, including MCF-7, to map genome-wide histone modifications and TF-binding sites. Whereas this method has been integral in locating gene REs, it can be highly variable (35). Antibody quality, sample preparation, sequencing depth, and MCF-7 cell line heterogeneity are the main sources of technical variability. Furthermore, following successful immunoprecipitation and sequencing, various algorithms then must be employed to define peaks: regions showing a high number of mapped reads relative to background. In particular, sufficiently deep sequencing is important for this part of the process wherein active REs are defined. Deeper sequencing allows for a lower relative detection threshold and so can capture low-expression transcripts, but also generates greater discrimination among peaks and more signal relative to control and so can map features with higher resolution and more reproducibly. This is important for the detection of variable regions that may not have robust enrichment but are still highly active and functionally important (35, 36). An example at one locus showing 3 different H3K27ac CHIP-Seq experiments, including our own reported here, is shown in Fig. 1A.

In addition to technical variation, MCF-7 is prone to instability, and major genetic and expression changes can be present between subclones (18). Therefore, prior to any genetic manipulation of MCF-7 for disease modeling, recent and accurate characterization is recommended. We reviewed existing ENCODE MCF-7 CHIP-Seq H3K27ac and CTCF datasets and compared these studies with our own (Fig. 1B–D). We mapped a total of 14,722 H3K27ac peaks, corresponding to a coverage of roughly 35.2 Mb of DNA (Supplementary Table S1). In comparison, the ENCODE 1 and ENCODE 2 datasets generate 11,304 and 20,102 reproducible peaks covering about 13.3 and 49.4 Mb, respectively (Fig. 1). However, peak size can vary, even if the same peaks have been called in multiple datasets. By comparing our peak locations (labeled "Coetzee") with the ENCODE datasets, we found that only 37% of our peaks overlapped H3K27ac regions from both ENCODE datasets (by at least 1 bp). In addition, 21% are completely unique to our data, 1% overlap peaks in ENCODE

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

dataset 1 (but not 2), and 41% overlap peaks present in ENCODE dataset 2 (but not 1). Examining unique DNA coverage across the three datasets yields a total of 61.4 Mb, which are annotated as H3K27ac in at least 1 experiment and 9.7 Mb (15%), which are annotated in all three datasets (Fig. 1E). Based upon an analysis of the proportion of reads in peaks (FRiP), which indicates the specificity of library following immunoprecipitation, our data are similar to that of ENCODE 2, both studies of which used the same primary antibody for H3K27Ac (Supplementary Fig. S1). The ENCODE 1 dataset in contrast (the oldest study) used a different antibody and shows a lower proportion of reads in peaks. Analyzing the reproducibility of peak calling between the three studies showed larger differences, as seen by IDR analysis. This analysis filters out peaks which are not similar across two biological replicates. In this case, our data retained 62% of defined peaks, ENCODE 2 retained 44%, and the oldest data, ENCODE 1 retained only 23% of peaks. ENCODE 1 used both a different primary antibody and was sequenced to a much lower depth for both input and control.

We used the same approach in comparing our MCF-7 CTCF data with three separate experiments from ENCODE (labeled 1–3 and unrelated to the H3K27Ac ENCODE data). We detected almost double the number of peaks in our data set (65,997 peaks) compared with the other three ENCODE datasets with 38,553, 33,312, and 36,057 peaks, respectively, that pass an IDR cutoff of $q < 0.01$ (IDR score > 830; Supplementary Table S1). When combining all three datasets, they only share 20,601 common peaks. After removing CTCF ENCODE1, the least reproducible data, we further evaluated CTCF ENCODE datasets 2 and 3 in comparison with our own. With our CTCF ChIP-Seq data, we detected a substantially greater amount of unique CTCF-bound DNA (20.5 Mb) across the genome in comparison with CTCF ENCODE 2 and 3 (0.5 and 0.03 MB; Fig. 1G). The majority of CTCF sites detected by the CTCF ENCODE 2 and 3 datasets were also detected in our data. Out of the 65,997 CTCF peaks found in our experiment, 39% were unique to our dataset, 10% were shared with CTCF ENCODE 3 but not CTCF ENCODE 2, 7% were shared with CTCF ENCODE 2 but not CTCF ENCODE 3, and 44% were shared between all three datasets (Fig. 1H). These data illustrate that CTCF peaks are largely conserved across datasets, but deeper sequencing may improve the detection of weaker CTCF-binding signals.

Coincidence of breast cancer risk variants with regulatory regions

Although most risk variants do not alter protein coding, they must still show allelic differences in some functional aspect of genome biology if they are causally related to the disease. Identifying active regulatory regions that coincide with risk variant locations can point to such mechanisms. In order to identify breast cancer risk processes which are active in MCF-7, we intersected the locations of risk variants with both our own and the ENCODE MCF-7 ChIP-Seq H3K27ac peaks and CTCF-binding peaks. To maximize both sensitivity and specificity, multiple sets of SNPs were used in this analysis based on different definitions of which variants are most likely to confer breast cancer risk, all deriving from Michailidou and colleagues (4) and based on different thresholds for significance (see Materials and Methods for details, Supplementary Table S2). Overall, we examined GWAS breast cancer risk variants corresponding to 142 separate loci. These loci each had one or small number of equally most significant variants per locus, with 210 in total. In addition, Michailidou and colleagues (4) reported a set of CRVs for each locus, with an average of 31 (stdev. 50) CRVs per locus that spanned an average of 114 Kb (stdev. 179 Kb). CRVs are most likely to be functional and are based on a locus-specific significance threshold (see Materials and Methods).

By examining the location of genomic regulatory activity with respect to risk variants, we can most easily infer which risk loci are *not* functional within a particular cell line. In MCF-7, 76 (of 142) breast cancer risk loci, and 93% of total reported CRVs, do not overlap any active REs. It is therefore very unlikely that these 76 loci confer risk via processes which are active in MCF-7. In contrast, the remaining 66 risk loci *may* function via processes that are intrinsic to, and active in, some breast cancer cells, making them logical targets for experimental manipulated in MCF-7. However, the possibility of spurious coincidence remains, as some rSNPs will overlap active REs simply by chance. Out of the 4,453 reported breast cancer CRVs, only 70 (1%) CRVs overlap (31) H3K27ac peak locations present in all three datasets, whereas 191 CRVs overlap (80) H3k27ac peaks based on our data alone (Supplementary Table S2). Similarly, only 13 CRVs overlap a CTCF-binding peak present in all of the MCF-7 CTCF ChIP-seq datasets, and 96 CRVs overlap (70) peaks based on our data alone. For CTCF, of the 13, just 5 (rs3008455, rs8103622, rs1800437, rs10231350, and rs10116233), and of the 96, 16 (Supplementary Table S4) are predicted to also cause allele-dependent disruption of a known CTCF-binding motif.

Coincidence of breast cancer risk variants with REs in MCF-7

If GWAS-measured breast cancer risk is functional in MCF-7 cells, then we predict that the location of breast cancer risk SNPs is correlated with the locations of active MCF-7 enhancers and/or CTCF-binding sites. That is, we expect that the character of MCF-7 as breast-derived cancer cells can be captured uniquely by the location of REs, and this unique activity pattern will predict the location of breast cancer CRVs. To test this hypothesis, we compared the location of breast cancer risk SNPs with entire set of imputed SNPs generated by Michailidou and colleagues, the vast majority of which are not statistically associated with breast cancer. Polymorphisms are distributed nonrandomly throughout the genome and tend to occur near active regions of transcription and so some breast cancer–unrelated SNP set is required to define a background rate of coincidence. The simplest calculation for SNP enrichment is to compare two ratios: the proportion of risk variants overlapping enhancers or CTCF sites and the proportion background SNPs overlapping the same. We made these enrichment calculations using slightly different definitions of MCF-7 enhancer or CTCF peaks, based on merging our data with the ENCODE data, and for different definitions of risk SNPs, as described above.

To calculate enrichment of breast cancer risk in MCF-7 then, we first measured the ratios of CTCF peaks overlapped by both risk variants and with the entire set of imputed SNPs. By using this comparison, we found that overall enrichment of SNPs coinciding with CTCF-binding sites is very modest and is close to the overlap expected by chance (Fig. 2D). Though the difference in enrichment scores between datasets was not large, CRVs showed the strongest enrichment for all ChIP datasets, indicating that this definition of risk SNPs may be the most relevant to breast cancer (see Supplementary Table S2 for CTCF sites that overlap CRVs). Out of the 210 most significant risk variants dataset, our ChIP data

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

were the only ones to coincide with any of them (rs4971059, rs56069439, rs12449271, and rs35383942). The effects of a CTCF motif disruption may apply across a broader range of cell types as CTCF occupancy is highly consistent (37, 38). To demonstrate this in the scope of our work with breast cancer, we compared the location of HMEC CTCF peaks (from ENCODE) with our CTCF peaks and found that 99% (15,095/15,138) of HMEC peaks were shared (at least 1 base pair) with our MCF-7 CTCF peaks.

In contrast to CTCF, we found highly significant enrichment of breast cancer risk SNPs in H3k27ac peaks (enhancers and promoters) for all sets of MCF-7 enhancers and all definitions of breast cancer risk variants (Fig. 2B). For example, we found that breast cancer CRVs overlapped 81 MCF-7 H3K27ac peaks (based on our data alone) and were 2.3-fold more likely to coincide with peaks than were the background SNPs overall. This enrichment means that the location of MCF-7 enhancer activity is predictive of risk SNP location. Therefore, as expected, at least some gene regulation that is active in MCF-7 also increases the risk of developing breast cancer.

To begin to define risk element functionality, we first categorized MCF-7 H3K27ac peaks as enhancer or promoter based on gene transcription start site locations. As expected, promoter peaks were uniformly less likely than enhancer peaks to show enrichment for risk variants compared with background. Focusing on enhancers, we then compared MCF-7–active enhancer locations with normal HMEC active enhancer locations, based on ENCODE H3K27ac CHIP-Seq data (Supplementary Table S1). We reasoned that HMEC exhibits a noncancerous phenotype and so represents the active-enhancer profile of normal tissue, whereas MCF-7 cells represent the active enhancer profile of ER^+ cancerous breast epithelial tissue. Therefore, they can be used to categorize risk enhancers active in two distinct cell types representing two extremes of breast epithelial cells. Although not yet demonstrated, we speculate that risk enhancers that are present only in HMEC point to risk imposed through precancer cell functions (such as apoptotic checkpoints) that must be altered or bypassed for carcinogenic initiation. Risk enhancers present only in MCF-7 could affect mechanisms involved in tumor progression and malignancy (such as cell adhesion and immune avoidance). Finally, risk enhancers present in both cell types may regulate processes involved in breast cancer at all stages of oncogenic promotion (such as DNA synthesis and cell proliferation). The data presented in Fig. 2C demonstrate that although all enhancer subtypes are enriched for breast cancer–associated SNPs compared with background, MCF-7–only enhancers are most correlated with risk SNPs. This indicates that a greater proportion of MCF-7–specific processes are involved in conferring breast cancer risk compared with HMEC. It also suggests that some of GWAS measured risk is involved with later stage processes of cancer development.

Next, we measured the significance of overlap between risk SNPs and MCF-7 enhancers at each locus individually. The total number and relative locations of risk SNPs in a locus are influenced by both the underlying disease biology and the LD structure. Therefore, we sought to control for LD and simultaneously treat the rSNPs at each locus as fundamentally connected (see Materials and Methods for more details). As a very simple partial solution to this problem, we chose to merge the set of risk SNPs within a locus into a single peak or span. We then used permutation testing to compare that risk span (the distance between the outer most significant CRVs at a locus) against a background set of randomly sampled equal-sized spans across the same locus (e.g., Fig. 2A). We measured the proportion of the risk span that overlaps active REs and generated a percentile score based on the number of background comparisons with greater overlap (Fig. 3, Supplementary Fig. S2, and Supplementary Table S3). A low value indicates greater confidence in the identification of a specific risk RE within a locus. This span overlap metric is helpful, in addition to SNP enrichment, and *P* value ranking, in order to determine which REs are most suitable for follow-up testing.

We calculated the risk span overlap at each locus for enhancers active in 11 related roadmap tissues as well as HMEC and MCF-7 (Fig. 3). Notably, the largest group of loci shows low overlap in any tissue. The remaining loci are easier to link to likely REs. For instance, at locus 6:151952332 near *ESR1*, the actual breast cancer CRV risk span overlaps a greater proportion of MCF-7 enhancers than more than 98% of other possible cases within 1 Mb. In contrast, no other cell types examined showed significant overlap. It is likely that the enhancer(s) at this locus, which completely overlaps breast cancer credible risk SNPs, is enhancing *ESR1* expression. In the other cell types, no enhancer is present, and *ESR1* expression is probably off, such as it is in HMEC cells. Breast cancer risk may be imposed through altered function of this enhancer, and this can be experimentally verified in MCF-7. Thirteen other loci show a greater overlap of risk SNPs with MCF-7 enhancers than expected by chance. Unlike *ESR1*, the risk locus, 19:13954571, near *NANOS3*, shows greater than expected overlap with MCF-7 enhancers as well as those of multiple other tissues. *NANOS3* is expressed in many tissues and is likely involved in general proliferation. The function of this locus, then, may be related to and can be queried in MCF-7 in addition to the other cell types.

The risk span overlap metric may also be misleading in some cases. For example, widely spaced CRVs can give spans that overlap large amounts of RE even when no individual risk SNPs do. However, using this measure in conjunction with a simple enrichment calculation and also with GWAS significance-based SNP ranking is an improvement over any one method alone. Table 1 shows 10 breast cancer loci that are highly appropriate for study in MCF-7 based on all three metrics.

Putative MCF-7 breast cancer risk genes

In order to begin to characterize the processes active in MCF-7 that confer risk for breast cancer via GWAS-identified risk variants, we sought to link risk enhancers to the genes which they may regulate. Therefore, we performed RNA-seq on MCF-7 cells and also compared that expression data with existing RNA-Seq data for both MCF-7 and HMEC cells (Supplementary Fig. S3). Our data corresponded well to published MCF-7 data (Supplementary Fig. S3A). Combining the MCF-7 datasets, we found 11,309 protein-coding genes expressed at an average of at least 1 count per million, whereas 12,294 genes were expressed in HMEC (Supplementary Table S5). Comparing the two cell lines, 10,657 genes (82.3%) were minimally expressed in both, and 8,087 genes were differentially expressed (adjusted P value < 0.05 and fold change > 2) with 4,899 more highly expressed in HMEC and 3,188 more in MCF-7 cells. The expression changes between the cell lines were not unexpected. MCF-7 genes were, roughly, overenriched for differentiated cell anatomical functions, migration, and intercellular interactions, and also enriched for KEGG cancer pathways. The HMEC genes were, roughly, overenriched for proliferation, metabolism, and nuclear organization. For

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

Table 1. MCF-7 breast cancer risk enhancers; listed are the 10 best breast cancer risk loci in MCF-7 and the corresponding 10 MCF-7 risk enhancers that are most suitable for follow-up testing

Locus	Alt. locus ID	CRVs Enh. overlap	CRVs in locus	Hyper. *P* val. (-log)	Ratio enrich.	Span overlap signif. (-log)	Lead SNP ID	Comb. GWAS *P* val.	Comb. GWAS *P* val., ER Pos.	Risk enhancer start	Risk enhancer stop	Gene	MCF-7 DE Genes 0–500 Kb	MCF-7 DE Genes 500 Kb–1 Mb
1p36.22	1_10566215	8	50	4.29	5.79	1.72	rs2506889	2.38E-20	1.10E-07	10595877	10599012	*PEX14*	*APITD1, TARDBP, UBE4B, PEX14*	*SRM, SLC25A33, PIK3CD, LZIC, PTCHD2, CLSTN1, EXOSC10, MTOR*
3p26.1	3_4742276	1	22	0.65	3.31	1.71	rs6787391	9.07E-19	1.20E-15	4727872	4732087	*EGOT, ITPR1*	*BHLHE40, ARL8B, ITPR1, SUMF1*	*EDEM1*
6q25	6_151952332	10	10	12.50	17.64	2.19	rs60954078	2.84E-54	1.70E-22	151948050	151959756	*ESR1*	*ZBTB2, RMND1, AKAP12, C6orf211, SYNE1, ESR1*	*MTHFD1L, PLEKHG1*
8q22.3	8_102478959	2	8	1.50	6.62	1.66	rs514192	5.61E-09	9.90E-09	102478133	102481107	*YWHAZ, GRHL2, KLF10*	*ZNF706, GRHL2*	*ANKRD46, YWHAZ, RRM2B, UBR5, PABPC1*
10p14	10_9088113	30	42	23.50	8.93	2.15	rs67958007	1.73E-10	6.00E-08	9074392	9089320	*GATA3*	N/A	*GATA3*
12q24.31	12_120832146	2	26	0.80	2.32	1.54	rs206966	3.79E-08	1.40E-06	120831824	120833249	*DYNLL1, GATC*	*TRIAP1, SRSF9, POP5, PXN, MSI1, DYNLL1, COQ5, UNC119B, CCDC64, RNF10*	*PRKAB1, P2RX4, CIT, ANAPC5*
16q23.2	16_80650805	8	17	9.69	26.12	2.66	rs7500067	4.06E-27	2.90E-21	80647147	80651103	*CDYL2*	*ATMIN, CENPN, CDYL2*	*GAN, CMIP*
17q25.3	17_77781725	3	13	2.56	9.98	1.90	rs8082452	1.14E-10	2.40E-06	77769769	77772332	N/A	*CBX2, CBX8, CBX4, CARD14, SGSH, SLC26A11, TBC1D16, EIF4A3*	*TIMP2, LGALS3BP, ENDOV, USP36, RPTOR*
19p13.12	19_13954571	19	43	13.35	8.74	2.74	rs2594714	1.08E-08	1.70E-05	13948902	13955856	*JUNB, NANOS3*	*MIR24-2, PALM3, SAMD1, CCDC130, ZSWIM4, C19orf57, PODNL1, ASF1B, IL27RA, LPHN1*	*CALR, SYCE2, TRMT1, CD97, GIPC1, DNAJB1, HOOK2, MAST1, NFIX, PKN1, ZNF333*
21q21.1	21_16520832	2	2	4.32	143.22	2.52	rs2403907	1.87E-32	4.20E-23	16570278	16574820	*NRIP1*	*NRIP1*	*HSPA13, USP25*

NOTE: Each enhancer overlaps the most significant GWAS SNP at the locus, and the locus shows both an enrichment greater than 2 (frequency of breast cancer CRVs overlapping enhancers compared with background overlap frequency), as well as a proportion of the CRV risk span that overlaps enhancers greater than more than 95% background overlap proportions. The enrichment and overlap values may be modified by overlap of risk SNPs with other enhancers at the same locus. The overall GWAS significance, ER+ GWAS significance, and reported genes from Michailidou and colleagues are listed. In addition, all but the enhancer at locus 19 are unique to MCF-7 and do not overlap HMEC enhancers locations, so the differentially expressed genes (relative to HMEC) within 500 Kb or 1 MB of each lead SNP are identified as putative risk genes in MCF-7.
Abbreviation: N/A, not applicable.

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

instance, the top 1,000 genes most significantly upregulated in MCF-7 were statistically enriched for 382 GO biological processes including "vasculature development," "anatomical structure morphogenesis," "regulation of cellular component movement," and "extracellular matrix organization" (FDRs = 5.2E-12, 5.2E-12, 2.0E-11, 2.2E-11; Supplementary Table S5). The top 1,000 most significant HMEC genes were enriched for 63 processes including "RNA metabolic process," "nucleic acid-templated transcription," "nucleobase-containing compound metabolic process," and "RNA biosynthetic process" (FDRs = 7.0E-8, 7.0E-8, 2.1E-7, and 2.2E-7; Supplementary Table S5).

Multiple methods exist that attempt to pair enhancers with target genes (39–41). The simplest method, which we use here, is to identify nearby genes which are coexpressed with active enhancers specifically in a particular cell type, such as MCF-7. For this purpose, we used only one definition of MCF-7 enhancers based on our own H3K27ac data (see Fig. 1) following the removal of promoter peaks. The majority of previously reported enhancer-gene regulatory pairing occur at distances less than 1 Mb, though far cis and even interchromosomal interactions do exist (42, 43). Based on our previous work, we used here a distance threshold of 500 kb distance to identify putative risk genes expressed in MCF-7. By this proximity cutoff, 522 genes are near-risk enhancers in MCF-7 cells. These potential risk-associated genes expressed in MCF-7 are statistically enriched for 102 GO biological process terms including "nucleosome assembly," "DNA methylation," "chromatin silencing," and "positive regulation of DNA repair" (FDR = 1.4E-5, 7.9E-4, 9.6E-3, and 1.5E-2; Supplementary Table S6). Likewise, there are 386 putative risk genes expressed in HMEC and 260 genes common to both the HMEC and MCF-7 sets. HMEC were statistically enriched for 42 biological processes. However, almost no significantly enriched HMEC processes were unique, 80% were also significant for the set of MCF-7 risk genes. The exception being enrichment for GO pathways related to Notch signaling or apoptosis (Supplementary Table S6). This suggests that few breast cancer risk processes are unique to HMEC, compared with MCF-7.

A potentially more accurate method to link REs with target genes is by using expression quantitative trait loci (eQTL) data, from sources such as GTEx. These results are biased by tissue sample type and population origin and cannot identify change is high variable or lowly expressed genes. However, an eQTL describes a direct association between the allelic variation at an SNP and an expression change for a gene within 1 Mb. In order to integrate eQTL data, we identified all SNPs (not just breast cancer risk SNPs) located within MCF-7 risk enhancers and queried the GTEx database for all significant eQTL genes, associated with those SNPs in breast tissue. We then removed from this set those genes which are not expressed in MCF-7. Doing so produced 48 genes (Supplementary Table S7). These genes were not enriched for any GO annotations and may be too stringently filtered to include likely risk genes. For instance, *ESR1*, a gene known to be important in breast cancer, and near a risk enhancer (Fig. 1), has no identified significant eQTLs in the GTEx dataset. Moreover, the CRVs located at that risk locus, one of the most significant of all loci, have also not been measured to be eQTLs for any gene. Thus, this clear test case fails reidentification by GTEx. So, although eQTL can provide strong support for linking a risk locus to a gene, we think it is currently too restrictive for further use here.

Because enhancer regions active in MCF-7, and not active in HMEC, were most highly correlated with the locations of risk variants, we sought to associate this subset of enhancers with putative genes. Enhancers alter the expression of nearby genes so we can assume that most genes that are regulated by this subset of risk enhancers will appear differentially expressed relative to HMEC expression levels, in which the risk enhancers are not active. Although most enhancers upregulate nearby genes, H3K27ac ChIP-Seq will also identify regulatory regions that can downregulate nearby genes. For this reason, we considered DE genes that were both up- and downregulated with respect to HMEC. Based on 500 KB proximity, 138 DE genes are associated with MCF-7–only risk enhancers (Supplementary Table S7). These are enriched for multiple broad GO categories, but also including the KEGG pathways for estrogen and prolactin signaling, identified via 6 and 5 genes, respectively (Supplementary Table S7).

Finally, in Table 1 we list risk MCF-7 enhancers at 10 breast cancer risk loci, which are highly enriched for SNPs, coincide with the span of DNA-containing breast cancer CRVs above background levels, and show overlap for the most significant risk SNP. We linked MCF-7 genes to these 10 enhancers based on expression in MCF-7, significant difference in expression relative to HMEC, and a distance closer than 1 Mb. These represent the most promising enhancer targets for further functional analysis.

Discussion

MCF-7 is a highly utilized breast cancer cell line. However, the range of its potential use for the dissection of breast cancer GWAS risk is not immediately obvious. Based on the classical theory of carcinogenesis, cancer arises from normal tissue and proceeds through the stages of initiation, promotion, and progression. Genetic factors affecting any of these stages may be picked up in GWAS studies, but only a subset of these risk mechanisms are likely to be active in MCF-7 itself. Only those processes active in MCF-7 can in turn be easily manipulated. By comparing MCF-7 with HMEC, we believe that risk arising from gene regulation involved in all three stages is active in MCF-7, but that MCF-7 is most suited for studying the risk mechanisms exacerbating tumor progression. In particular, estrogen and prolactin signaling gene networks are especially enriched in breast cancer GWAS risk biology in MCF-7. Although MCF-7 cells are classified as luminal A/ER$^+$, it is worth noting that this cell line may still be relevant for tumorigenic processes in other subtypes including ER$^-$ breast cancer. The developmental lineage of breast tumor subtypes is complex and not fully understood. Based on the hypothesis that different subtypes may be derived from the same cell type of origin, some of the active enhancer–driven processes leading to cancer in each subtype are most likely not mutually exclusive. Further studies need to be done to evaluate the overlapping risk in multiple breast cancer cell lines of different classifications.

We found that the work reported here is more reproducible in defining MCF-7–specific H3K27ac histone marks and CTCF occupancy than that in previously reported ENCODE datasets. The discrepancy in the total amount of H3K27ac or CTCF DNA and in continuously designated regions (peaks) is likely due both to differences in sample preparation and sequencing depth as well as to intrinsic differences between subclones.

A potential next step in dissecting breast cancer risk in MCF-7 is to perform allelic replacement or to delete the entire risk REs using CRISPR-CAS9 gene disruption. We found that multiple enhancers are well suited for follow-up experiments, and a smaller number of CTCF sites may also be suitable. Allele-dependent CTCF binding can potentially disrupt large TAD regions leading to large expression changes to nearby genes. For this reason, CTCF can be an ideal target. Unfortunately, we found very low enrichment of rSNPs, implying more breast cancer risk loci function via enhancers in MCF-7. It is possible that the enrichment scores maybe less informative for CTCF, though CTCF-binding sites/peaks are much more specific and narrower than that of H3K27ac peaks and are less likely to overlap multiple risk SNPs simply due to the close proximity of risk SNPs to each other. This and the greater number of CTCF peaks (which are largely tissue invariant) may contribute to lower overall enrichment. In total, we found 5 loci which are good targets for further CTCF examination.

In contrast, at least 10 loci point to specific enhancers as casual in breast cancer risk. For enhancers, we found the greatest degree of enrichment of risk variants in the MCF-7 cell line as compared with a normal precursor model (HMEC), indicating that MCF-7 cells are highly relevant for the study of processes leading to abnormal cell growth and tumor formation.

Overall, our results reported here bring into focus how MCF-7 can be used as a model to reveal breast cancer risk mechanisms in ER-positive genetic predisposition. For enhancers, we found the greatest degree of enrichment of risk variants in the MCF-7 cell line as compared with a normal precursor model (HMEC), indicating that MCF-7 cells are highly relevant for the study of processes leading to abnormal cell growth and tumor formation.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors' Contributions

Conception and design: G.A. Coetzee
Development of methodology: A. Booms
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A. Booms
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Booms, S.E. Pierce
Writing, review, and/or revision of the manuscript: A. Booms, G.A. Coetzee, S.E. Pierce
Study supervision: G.A. Coetzee, S.E. Pierce

Acknowledgments

The authors would like to thank everyone who contributed to this study. Van Andel Institute also provided the facilities and general support to conduct this research. Marie Adams and the rest of the Genomics Core conducted and advised on next-generation sequencing, and Lee Marshall provided analysis advice.

Research reported in this publication was supported by the NCI of the NIH under award number R01CA190182 (to G.A. Coetzee).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 4, 2019; revised May 17, 2019; accepted July 2, 2019; published first July 10, 2019.

References

1. Hirshfield KM, Rebbeck TR, Levine AJ. Germline mutations and polymorphisms in the origins of cancers in women. J Oncol 2010;2010: 297671.
2. Fachal L, Dunning AM. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr Opin Genet Dev 2015;30:32–41.
3. Skol AD, Sasaki MM, Onel K. The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. Breast Cancer Res 2016;18:99.
4. 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.
5. Coetzee SG, Pierce S, Brundin P, Brundin L, Hazelett DJ, Coetzee GA. Enrichment of risk SNPs in regulatory regions implicate diverse tissues in Parkinson's disease etiology. Sci Rep 2016;6:30509.
6. Pierce SE, Tyson T, Booms A, Prahl J, Coetzee GA. Parkinson's disease genetic risk in a midbrain neuronal cell line. Neurobiol Dis 2018;114: 53–64.
7. Rhie SK, Coetzee SG, Noushmehr H, Yan C, Kim JM, Haiman CA, et al. Comprehensive functional annotation of seventy-one breast cancer risk Loci. PLoS One 2013;8:e63925.
8. Plank JL, Dean A. Enhancer function: mechanistic and genome-wide insights come together. Mol Cell 2014;55:5–14.
9. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 2010;107:21931–6.
10. Sur I, Taipale J. The role of enhancers in cancer. Nat Rev Cancer 2016; 16:483–93.
11. Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet 2009;5:e1000597.
12. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 2009; 41:882–4.
13. Merkenschlager M, Nora EP.CTCF and cohesin in genome folding and transcriptional gene regulation. Annu Rev Genomics Hum Genet 2016;17: 17–43.
14. Nakamoto M, Ishihara K, Watanabe T, Hirosue A, Hino S, Shinohara M, et al. The glucocorticoid receptor regulates the ANGPTL4 gene in a CTCF-mediated chromatin context in human hepatic cells. PLoS One 2017;12: e0169225.
15. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 2015;161:1012–25.
16. Guo Y, Rhie SK, Hazelett DJ, Coetzee GA, Farnham PJ. CRISPR-mediated deletion of prostate cancer risk-associated CTCF sites identifies repressive chromatin loops. Genome Biol 2018;19:160.
17. Allen EK, Randolph AG, Bhangale T, Dogra P, Ohlson M, Oshansky CM, et al. SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans. Nat Med 2017;23: 975–83.
18. Lee AV, Oesterreich S, Davidson NE. MCF-7 cells–changing the course of breast cancer research and care for 45 years. J Natl Cancer Inst 2015;107. pii: djv073.
19. Howlader N, Altekruse SF, Li CI, Chen VW, Clarke CA, Ries LA, et al. US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. J Natl Cancer Inst 2014;106. pii: dju055.
20. Rivandi M, Martens JWM, Hollestelle A. Elucidating the underlying functional mechanisms of breast cancer susceptibility through post-GWAS analyses. Front Genet 2018;9:280.
21. 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.

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
23. 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.
24. Li QH, Brown JB, Huang HY, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann Appl Stat 2011;5:1752–79.
25. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 2016;44:W160–5.
26. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics 2015;31:3847–9.
27. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
29. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics 2015;51:11.14.1–19.
30. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:R29.
31. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
32. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25.
33. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43(Database issue):D447–52.
34. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
35. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform 2017;18: 279–90.
36. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22:1813–31.
37. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 2012;22:1680–8.
38. Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, Smith ST, et al. CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. EMBO Rep 2005;6:165–70.
39. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature 2012;488:116–20.
40. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009;459:108–12.
41. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43–9.
42. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res 2012;22:490–503.
43. Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R. Interchromosomal interactions and olfactory receptor choice. Cell 2006; 126:403–13.

Published OnlineFirst July 10, 2019; DOI: 10.1158/1055-9965.EPI-19-0066

MCF-7 as a Model for Functional Analysis of Breast Cancer Risk Variants

Alix Booms, Gerhard A. Coetzee and Steven E. Pierce

Cancer Epidemiol Biomarkers Prev 2019;28:1735-1745. Published OnlineFirst July 10, 2019.