frontiers in Genetics

TECHNOLOGY AND CODE
published: 20 July 2020
doi: 10.3389/fgene.2020.00730

ABC-GWAS: Functional Annotation of Estrogen Receptor-Positive Breast Cancer Genetic Variants

*Mohith Manjunath[1,2], Yi Zhang[2,3], Shilu Zhang[4], Sushmita Roy[4,5], Pablo Perez-Pinera[2,3,6,7] and Jun S. Song[1,2,7]**

[1] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [3] Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [4] Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, WI, United States, [5] Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, United States, [6] The Carle Illinois College of Medicine, Champaign, IL, United States, [7] Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Urbana, IL, United States

OPEN ACCESS

Edited by:
George Bebis,
University of Nevada, Reno,
United States

Reviewed by:
Dylan Glubb,
QIMR Berghofer Medical Research Institute, The University of Queensland, Australia
Jun Zhong,
National Cancer Institute (NCI),
United States

***Correspondence:**
Jun S. Song
songj@illinois.edu

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Genetics

Received: 18 February 2020
Accepted: 16 June 2020
Published: 20 July 2020

Citation:
Manjunath M, Zhang Y, Zhang S, Roy S, Perez-Pinera P and Song JS (2020) ABC-GWAS: Functional Annotation of Estrogen Receptor-Positive Breast Cancer Genetic Variants. Front. Genet. 11:730. doi: 10.3389/fgene.2020.00730

Over the past decade, hundreds of genome-wide association studies (GWAS) have implicated genetic variants in various diseases, including cancer. However, only a few of these variants have been functionally characterized to date, mainly because the majority of the variants reside in non-coding regions of the human genome with unknown function. A comprehensive functional annotation of the candidate variants is thus necessary to fill the gap between the correlative findings of GWAS and the development of therapeutic strategies. By integrating large-scale multi-omics datasets such as the Cancer Genome Atlas (TCGA) and the Encyclopedia of DNA Elements (ENCODE), we performed multivariate linear regression analysis of expression quantitative trait loci, sequence permutation test of transcription factor binding perturbation, and modeling of three-dimensional chromatin interactions to analyze the potential molecular functions of 2,813 single nucleotide variants in 93 genomic loci associated with estrogen receptor-positive breast cancer. To facilitate rapid progress in functional genomics of breast cancer, we have created "Analysis of Breast Cancer GWAS" (ABC-GWAS), an interactive database of functional annotation of estrogen receptor-positive breast cancer GWAS variants. Our resource includes expression quantitative trait loci, long-range chromatin interaction predictions, and transcription factor binding motif analyses to prioritize putative target genes, causal variants, and transcription factors. An embedded genome browser also facilitates convenient visualization of the GWAS loci in genomic and epigenomic context. ABC-GWAS provides an interactive visual summary of comprehensive functional characterization of estrogen receptor-positive breast cancer variants. The web resource will be useful to both computational and experimental biologists who wish to generate and test their hypotheses regarding the genetic susceptibility, etiology, and carcinogenesis of breast cancer. ABC-GWAS can also be used as a user-friendly educational resource for teaching functional genomics. ABC-GWAS is available at http://education.knoweng.org/abc-gwas/.

Keywords: GWAS, breast cancer, functional characterization, variant annotation, web resource

INTRODUCTION

Genome-wide association studies (GWAS) have implicated thousands of genetic variants in various complex traits, including diseases (MacArthur et al., 2017). However, only a few studies to date have been successful in characterizing the underlying molecular mechanisms that govern how genetic variations affect molecular interactions (Musunuru et al., 2010; Cowper-Sal lari et al., 2012; Bauer et al., 2013; Huang et al., 2014; Smemo et al., 2014; Gallagher et al., 2017; Zhang et al., 2018b). Studying the molecular function of a typical GWAS locus presents several key challenges (Gallagher and Chen-Plotkin, 2018). First, most of the variants found through GWAS are located in non-coding regions of the human genome; as a result, the precise link between a non-coding variant and some target protein's function is not immediately clear. Second, GWAS variants may indirectly correlate with a phenotype through a complex gene regulatory network involving multiple target genes, unknown causal variants, and transcription factors (TFs). For example, a reported GWAS variant may simply be genetically linked to another proximal variant that itself directly perturbs the binding affinity of a TF and changes the expression of a distal target oncogene or tumor suppressor forming a chromatin loop with the causal variant. In such cases, there is the additional complexity of having to dissect how different components of a gene regulatory network are altered and function together to modulate a trait. Finally, functional characterization of GWAS loci must be carried out in the right cell type representing the phenotype in question; however, one often lacks a complete set of data in genomic, epigenomic, and transcriptomic contexts in the cell type of interest or even faces a difficulty in determining the right cell type. Therefore, there is an urgent need for comprehensive and easily accessible resources that integrate information from heterogeneous large-scale datasets to facilitate rapid functional characterization of GWAS findings and ultimately contribute toward the development of therapeutic preventions and interventions.

Building on the public catalog of GWAS variants (MacArthur et al., 2017), there are currently a few databases providing functional annotation of disease variants. The GRASP database annotates GWAS results by summarizing millions of single nucleotide variant-phenotype associations from 1,390 GWAS studies through correlations such as expression quantitative trait loci (eQTLs), metabolite QTLs, and methylation QTLs (Leslie et al., 2014). Similarly, GWASdb curates trait-associated single nucleotide polymorphisms (SNPs) with detailed functional annotations including eQTL and disease ontology terms (Li et al., 2016). Phenoscanner is a curated database containing variant-phenotype associations of several types such as disease, methylation, gene expression, and protein levels (Staley et al., 2016). More recently, Qtlizer provides associations of variants with gene expression levels and protein abundance using published QTLs (Munz et al., 2019). In the context of cancer, PancanQTL provides a comprehensive list of cis- and trans-eQTLs, including GWAS-related eQTLs, in 33 cancer types (Gong et al., 2018). These web resources have specific advantages, such as having a detailed annotation of GWAS SNPs and/or a list of potential target genes found through eQTL analysis. However, these resources do not perform an in-depth integrative analysis of a specific cancer type using state-of-the-art information about cell type-specific epigenetic landscape, chromatin contact interactions, and TF binding affinity, required for a complete functional characterization of GWAS loci.

Most studies investigating breast cancer GWAS variants have so far focused only on eQTL analysis to find genes correlated with a variant genotype, while only few have pursued a systematic analysis of causal variants and target genes through chromatin structure and TFs (Cowper-Sal lari et al., 2012; French et al., 2013; Li et al., 2013; Ghoussaini et al., 2014, 2016; Darabi et al., 2015; Dunning et al., 2016; Michailidou et al., 2017; Zhang et al., 2018b; Zhang Y. et al., 2019). This paper presents ABC-GWAS, an interactive database containing our comprehensive analysis of 70 manually curated estrogen receptor-positive (ER+) breast cancer GWAS loci and 23 additional ER+ breast cancer loci from a recent fine mapping study (Fachal et al., 2020). The set of 70 loci was obtained from the literature on breast cancer GWAS (Turnbull et al., 2010; Michailidou et al., 2013, 2015). Utilizing large-scale multi-omics datasets such as the Cancer Genome Atlas (TCGA) and the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium, 2012), our analysis pipeline includes eQTL analyses for identifying putative target genes, causal variant prioritization utilizing relevant epigenomic datasets, motif and expression correlation analyses for identifying putative TFs, and three-dimensional chromatin contact predictions for assessing long-distance enhancer-gene interactions. ABC-GWAS aggregates and organizes these results, not readily available in other existing databases, via a user-friendly web interface, making them easily accessible to researchers for additional analysis or experimental validation. It features an embedded genome browser that includes histone modification, chromatin interaction, and TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) tracks for further exploration of the GWAS locus and linked non-coding variants of interest. ABC-GWAS also shows the average DNA copy number information in TCGA breast cancer samples at each GWAS locus. Our resource thus provides useful practical results and conceptual approaches to the functional genomics community in general and breast cancer researchers in particular.

MATERIALS AND METHODS

TCGA Data and Genotype Imputation

The processed RNA-seq expression data in RSEM (RNA-Seq by Expectation-Maximization) units for 794 ER+ breast cancer patients were obtained from the TCGA Genomic Data Commons (GDC) Legacy Archive (Grossman et al., 2016). The germline genotypes of 788 patients in birdseed format for TCGA-BRCA (Breast Invasive Carcinoma) patients were also obtained from the TCGA Data Portal. The copy number segmentation data for 693 patients in hg19 coordinates were retrieved from the GDC Legacy Archive (Grossman et al., 2016). For genotype imputation of the raw genotypes in birdseed format, confidence score greater than 0.1 was used to mark the probed genotypes as missing,

which was then imputed along with the non-probed SNPs. We used the Michigan Imputation Server for imputation (Das et al., 2016), choosing the Haplotype Reference Consortium (HRC) r1.1 2016 as a reference panel (Loh et al., 2016a), Eagle v2.3 for phasing (Loh et al., 2016b), and EUR population as the quality control option. Imputed genotypes were retained if the minor allele frequency (MAF) exceeded 0.005 and estimated imputation accuracy (R^2) exceeded 0.4.

Credible Causal Variants in 23 Additional GWAS Loci

We obtained the full list of credible causal variants (CCVs) from Fachal et al. (2020) and then selected the variants that are single-nucleotide variants, associated with ER+ breast cancer (column ERpos = 1), and have posterior probability of being causal greater than zero (column PP_ERpos > 0). We further removed SNPs that did not pass the quality control tests in the Michigan Imputation Server or for which genotypes could not be imputed confidently in the TCGA data. Finally, excluding 227 CCV SNPs already present in the list of 2,510 SNPs that were in high linkage disequilibrium ($r^2 > 0.8$, 1000 Genomes Phase 3, EUR population) with the reported GWAS SNPs in the 70 manually curated regions yielded 303 CCVs with non-zero posterior probability of being causal in ER+ breast cancers. The 303 CCVs resided in 32 GWAS regions, and 23 of these regions differed from the 70 manually curated regions. ABC-GWAS thus contains the analysis of 530 CCVs out of the 1,238 CCVs reported for ER+ breast cancer.

Genome Browser

The WashU EpiGenome Browser source code was obtained from their GitHub repository (Li et al., 2019; WashU, 2019). The browser uses hg19 coordinates. The JavaScript files from the source code were used to generate the tracks in the embedded browser of ABC-GWAS. The tracks included TF ChIP-seq peaks publicly available in ReMap 2018 database (Cheneby et al., 2018), ENCODE DNase-seq signals, and ESR1, GATA3, and FOXA1 ChIP-seq signals in MCF-7 and T-47D cell lines, POLR2A, CTCF, and ESR1 ChIA-PET interactions, and chromatin interaction predictions in MCF-7 cell line. CTCF is known to play an important role in defining the activity of ESR1 in ER+ breast cancer (Carroll et al., 2006; Chan and Song, 2008). The above datasets were downloaded from the corresponding sources and integrated into our server (**Supplementary Table 1**).

Chromatin Interaction Predictions

To predict SNP-associated interactions, we applied HiC-Reg (Zhang S. et al., 2019), a tool for predicting Hi-C contact counts between pairs of genomic loci from their one-dimensional regulatory signals such as histone modification data, TF ChIP-seq, and chromatin accessibility. We obtained ChIP-seq datasets for 10 histone marks and TFs, and DNase-seq datasets in five cell lines from ENCODE (**Supplementary Material**). HiC-Reg can be trained using cell-line-specific datasets for a cell line with available high-resolution (5 kb) Hi-C data, e.g., the five human cell lines available from Rao et al. (2014). Once trained, HiC-Reg takes as input the genomic features associated with a pair of regions and predicts the chromatin contact count for that pair. We used the method to make predictions in the MCF-7 cell line by training eight different models at 5 kb resolution (**Supplementary Material**). To interpret our results, we averaged the predictions across eight models and displayed the resulting contact count profile associated with each SNP on ABC-GWAS.

eQTL Analysis

To identify candidate target genes for each GWAS SNP, we scanned all genes within 4 Mb centered at the SNP by constructing a multivariate linear regression model with the expression level of each gene as the response variable and the genotype of the GWAS SNP and the copy number (CN) of the gene as predictors (Zhang et al., 2018b; Zhang Y. et al., 2019). The processed gene expression levels in RSEM units were transformed as $\log_2(RSEM + 1)$. The patients with ER+ breast cancer based on TCGA clinical information were retained for subsequent analysis. The genotypes of each GWAS SNP were encoded as the number of risk alleles based on the risk allele information from the NHGRI GWAS catalog (MacArthur et al., 2017). The tumor copy number segmentation values were transformed into gene copy number by taking gene length-weighted average and using $CN = 2 \times 2^{\{segmentation\}}$. We then performed multivariate linear regression and selected genes with mean RSEM larger than 1 and genotype p-value less than 0.05 as candidate target genes for each breast cancer GWAS SNP. On the website, a violin plot using plotly.js is displayed to show the distribution of a candidate target gene's mRNA expression as a function of the GWAS SNP's genotype status (Plotly, 2018).

ENCODE Data

ChIP-seq files for 715 TFs and histone marks in 231 cell lines and tissues were obtained from the ENCODE website (Davis et al., 2018). The locations of the breast cancer risk variants along with the high LD SNPs were then intersected with the peaks of each TF or histone mark in every cell line using bedtools (Quinlan and Hall, 2010). A list of TFs, relevant cell lines, and distance of the SNP from peak center were then tabulated for display.

Motif Analysis

Position weight matrices (PWM) for TFs were obtained from several public databases included in the MotifDb and motifbreakR packages on Bioconductor (Coetzee et al., 2015; Shannon, 2017). The public databases included Jaspar 2018 (Khan et al., 2018), HOCOMOCO (Kulakovskiy et al., 2018), hPDI (Xie et al., 2010), Jolma (Jolma et al., 2013), cisbp (Weirauch et al., 2014), UniPROBE (Hume et al., 2015), Swiss Regulon (Pachkov et al., 2013), HOMER (Heinz et al., 2010), ENCODE motifs (Kheradpour and Kellis, 2014), and FactorBook (Wang et al., 2012). TRANSFAC matrices were also added to the above list (Wingender, 2008). In the first step, the motifbreakR package was used to get possible motif disruptions by candidate SNPs with a p-value threshold of 10^{-3}. We then used our previously developed random mutation model to test the significance of difference in motif scores for the two sequences carrying reference and alternative alleles (Zhang et al., 2018b). The motif disruptions that passed the permutation test p-value threshold of

0.05 were denoted as significant and subsequently included in the ABC-GWAS database.

Correlation Analysis

The list of putative TFs from motif analysis was filtered by removing TFs whose log-transformed mean expression levels across TCGA ER+ breast cancer patients were less than 1 (mean $\log_2(RSEM + 1) < 1$). For each putative target gene from eQTL analysis and TFs passing the expression cut-off threshold, we computed the Pearson correlation coefficient between the expression levels of the target gene and TF across TCGA ER+ breast cancer primary tumor samples, stratifying the patients into three genotype groups: homozygous-risk, heterozygous, and homozygous-alternative. We reasoned that for a good candidate TF, the correlation should be strongest in the homozygous genotype group preserving the TF motif and weakest in the homozygous genotype group disrupting the motif.

RESULTS

Analysis Pipeline for Prioritization of Functional Candidates

We applied the analysis pipeline from our previous work (Zhang et al., 2018b), summarized in **Figure 1**, on a list of manually curated ER+ breast cancer GWAS variants and all SNPs in high linkage disequilibrium (LD) with the GWAS variants, as well as an additional 303 credible causal variants (CCVs) with non-zero posterior probability of being causal in ER+ breast cancers (Fachal et al., 2020; section "Materials and Methods"). The basic framework performs various genomic analyses outlined below to infer how a GWAS variant or a linked SNP changes the binding affinity of a TF in a regulatory region, which in turn alters the transcription of a target gene. In our analysis, linked SNPs residing in accessible open chromatin sites with activating histone modifications (H3K4me1 and H3K27ac) are prioritized as candidate causative SNPs. The genotypes and gene expression data were obtained from TCGA, where the non-probed SNPs' genotypes were imputed using the Michigan imputation server (Das et al., 2016; section "Materials and Methods"). We gathered various heterogeneous datasets from high-throughput experimental techniques such as DNase I hypersensitive sites sequencing (DNase-seq) for prioritization of candidate causal variants, ChIP-seq for TF binding evidence, and chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) and RNA-seq for target gene prioritization in breast cancer samples or cell lines (section "Materials and Methods"; **Supplementary Table 1**). In order to assess how a SNP may perturb a TF's binding affinity and consequently modulate a target gene's expression, we performed eQTL analysis, motif analysis, and TF vs. target gene expression correlation analysis to determine a list of candidate (SNP, target gene, TF) triplets (section "Materials and Methods").

ABC-GWAS User Interface

ABC-GWAS is divided into several modules for interactive data exploration. In the query module, the user first selects a GWAS SNP of interest from the list of 70 SNPs which represent the best reported variants in the manually curated implicated loci, after which a list of high LD ($r^2 > 0.8$, 1000 Genomes Phase 3, EUR population) SNPs of the queried GWAS SNP is populated (**Figure 2**). A table containing the list of GWAS studies implicating the selected SNP in breast cancer is shown on the

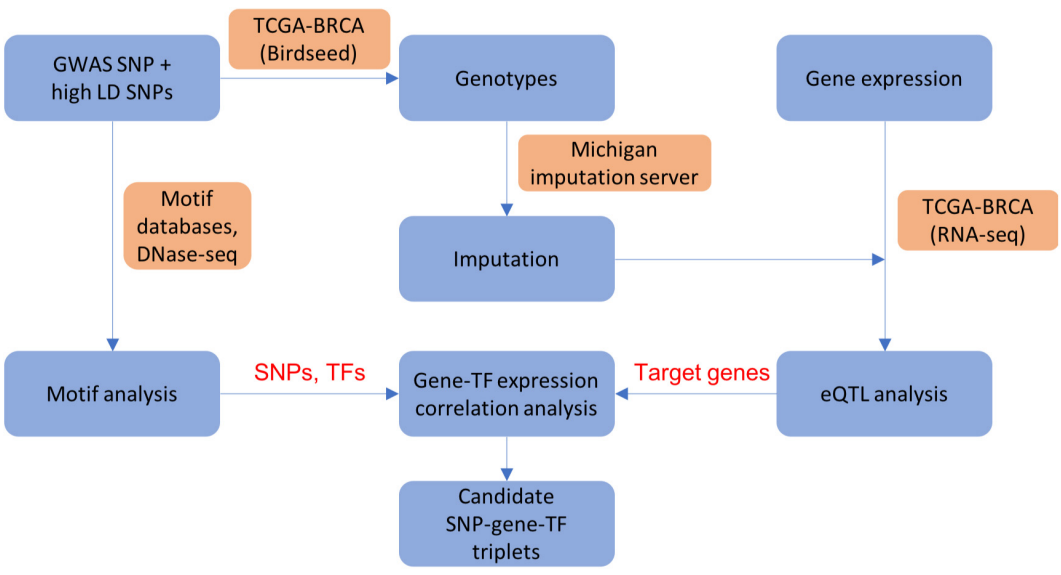

FIGURE 1 | A flowchart showing integrative analysis pipeline used in ABC-GWAS. For each GWAS locus, we perform eQTL analysis, motif analysis, and gene-TF expression correlation analysis to obtain candidate (SNP, gene, TF) triplets. Blue boxes show analysis steps, orange boxes indicate data and/or resources used in the analysis, and text in red shows intermediate results.

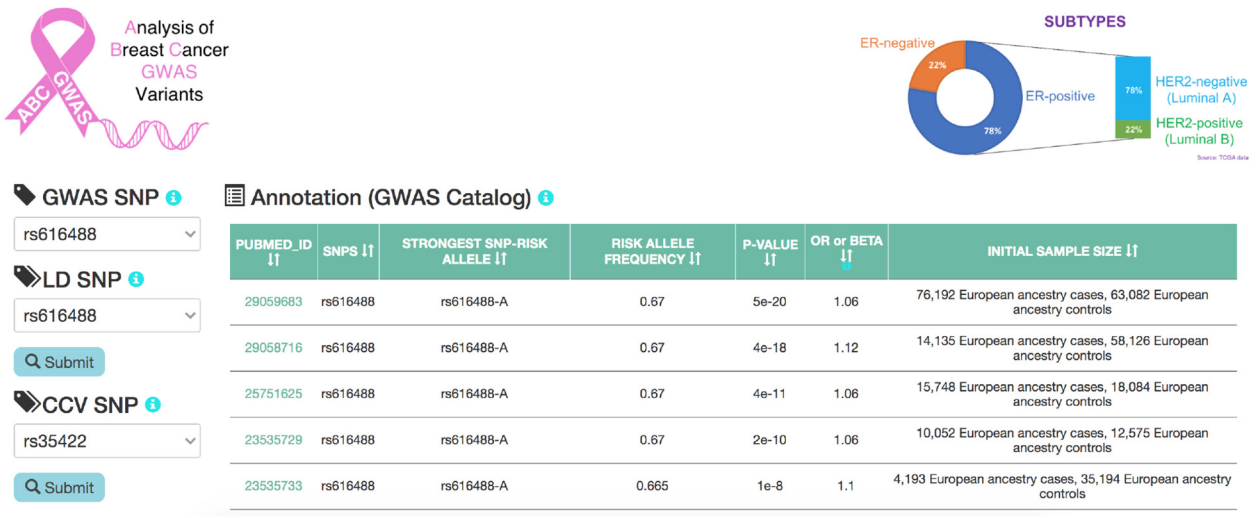

FIGURE 2 | A snapshot of the homepage of ABC-GWAS. Selecting a GWAS SNP using the left-hand-side drop-down menu populates the table on the right with relevant GWAS publications. Upon selecting an LD SNP or a CCV and clicking on the "Submit" button, various tabs on the bottom containing analysis modules are loaded.

right-hand side of the query module (**Figure 2**). Alternatively, the user may choose one of the additional 303 CCVs, not found in the list of all high LD SNPs. After submitting a high LD or CCV SNP as the query variant, all the analysis tabs below the query module get updated. The first tab contains an embedded genome browser showing ChIP-seq, DNase-seq, and ChIA-PET sequencing tracks around the queried SNP locus (**Figure 3A**; section "Materials and Methods"). The second tab displays predicted chromatin interactions in the MCF-7 breast cancer cell line, showing significant interactions between the queried LD SNP location and nearby gene promoters (**Figure 3B**; section "Materials and Methods", and **Supplementary Material**); this track is not available for the 303 CCVs. The third tab consists of two modules. One module shows the average DNA copy number around the queried GWAS SNP location using the TCGA copy number segmentation data for normal and tumor samples (**Figure 3C**; section "Materials and Methods"). The other module checks whether the queried SNP is a CCV (Fachal et al., 2020); when available, a list of likely target genes of the queried SNP obtained from the same study is also displayed. The fourth tab summarizes our eQTL analysis results for the selected GWAS SNP or CCV using the genotypes and RNA-seq data from TCGA breast cancer samples (section "Materials and Methods"). A table containing significant eQTL results and a violin plot of the target gene's expression stratified into genotype groups are displayed. The fifth tab shows a table of all ENCODE ChIP-seq peaks that intersect the queried SNP (section "Materials and Methods"). The peaks are categorized based on whether the experiment is for a TF or histone modification. The results can also be filtered to show peaks occurring only in breast tissue or breast cancer-related cell lines. The last tab contains two modules showing putative TFs, the binding activities of which are predicted to be affected by the given SNP, as assessed by motif analysis (section "Materials and Methods") and expression correlation analysis (section "Materials and Methods"). A motif logo with the nucleotide perturbed by the SNP is available for each of the putative TFs. In the "Expression correlation" tab, the putative TFs from motif analysis are further prioritized based on the expression correlation between each TF and eQTL target genes. Pearson correlation coefficients are displayed as a heatmap with the putative TFs along the rows and genotype groups along the columns.

Case Study (Validated Result From the Literature): (rs4784227, *TOX3*, FOXA1)

Cowper-Sal lari et al. (2012) analyzed the functional mechanism of the GWAS SNP rs4784227 and proposed it to be a causal regulatory SNP targeting the gene *TOX3*. Furthermore, the risk allele rs4784227-T was shown to increase the binding affinity of the pioneer factor FOXA1, resulting in a fivefold decrease in *TOX3* gene expression. Here, we sought to verify the reported mechanism at the rs4784227 locus using the results from our database. **Figure 4A** shows a snapshot of the genomic region around rs4784227 from the embedded genome browser. The MCF-7 DNase tracks in **Figure 4A** clearly indicate that the GWAS SNP is located within open chromatin region. Furthermore, the "ReMap 2018 Peaks" track, which represent TF binding peak locations collected from ENCODE and Gene Expression Omnibus (GEO) datasets (Barrett et al., 2013; Cheneby et al., 2018), showed several TF binding sites, supporting that this SNP is likely a causal SNP. The eQTL results showed a negative correlation between the risk allele rs4784227-T and the mRNA level of *TOX3* in TCGA breast cancer samples (**Figure 4B**). Our motif analysis results further suggested FOXJ3 as one of the top candidate TFs (**Figure 4C**); given the similarity of FOXJ3 and FOXA1 motifs (q-value = 0.0098), as predicted by

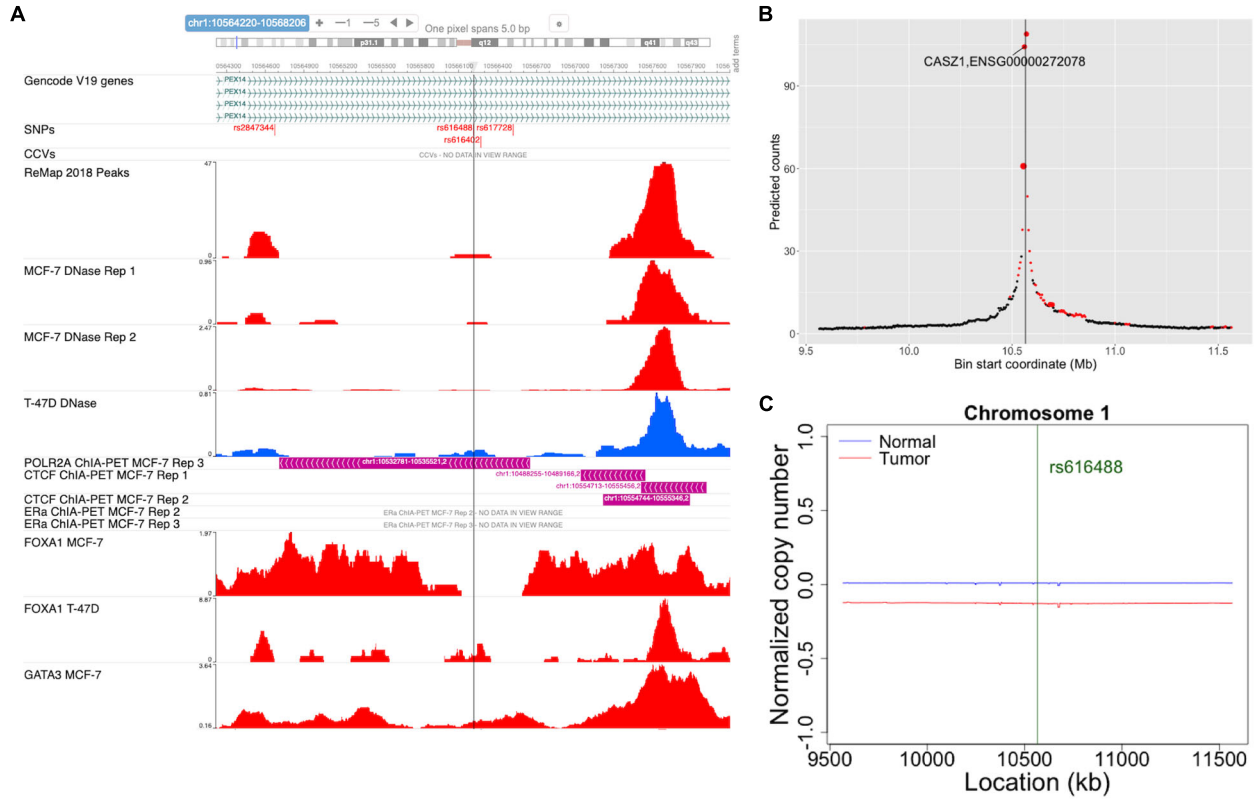

FIGURE 3 | Snapshots of some of the ABC-GWAS analysis modules. **(A)** Embedded genome browser showing the queried LD SNP location (vertical line) and ChIP-seq tracks in MCF-7 and T-47D cell lines. The "SNPs" track shows the locations of the GWAS SNPs and high LD SNPs, while the "CCVs" track shows the locations of the credible causal variants and their GWAS lead SNPs. **(B)** The plot shows average predicted chromatin contact counts across several HiC-Reg models in MCF-7 as a function of genomic location centered at the queried LD SNP (vertical line). The predictions for which at least one model shows significance ($q < 0.05$) is filled in red. The size of the markers is proportional to the number of models showing significance. **(C)** Normalized DNA copy number for normal (blue) and tumor (red) samples from TCGA in a 2 Mb window centered at the GWAS SNP location.

the Tomtom motif comparison tool from MEME web resource (Gupta et al., 2007; Bailey et al., 2009), our overall results were thus consistent with the findings of Cowper-Sal lari et al. (2012).

Case Study (Novel): (rs1250003, *ZMIZ1*, GATA)

The SNP rs704010, residing within an intron of the gene *ZMIZ1*, was reported to be associated with increased breast cancer risk in Turnbull et al. (2010), and this association was subsequently verified in later studies (Michailidou et al., 2013, 2015, 2017; Zhang et al., 2018a). **Figure 5A** shows a snapshot of the locus from the embedded genome browser. Among the 12 high LD SNPs shown in the first track, we identified rs1250003 to be the only SNP residing within an open chromatin region in MCF-7 and also to a lesser extent in T-47D, as shown by the DNase tracks. This candidate SNP rs1250003 was located about 5 kb from the GWAS SNP and in high LD with the GWAS SNP (r^2 = 0.99, 1000 Genomes Phase 3, EUR population). We also found that in the European population (1000 Genomes, Phase 3), rs1250003 was in perfect LD with two SNPs (rs1250008, rs1250009) previously reported to be CCVs (Fachal et al., 2020). Several TFs relevant to breast cancer – such as ESR1, FOXA1, and GATA3 – were found to bind near the SNP, as shown by the corresponding ChIP-seq tracks, indicating an important regulatory role of the SNP. The genotype status of rs704010 significantly correlated with the mRNA level of *ZMIZ1* ($p = 7.7 \times 10^{-4}$) (**Figure 5B**). POLR2A ChIA-PET track in MCF-7 further showed a chromatin-looping interaction between the SNP location and the promoter of *ZMIZ1* (**Figure 5A**). A significant interaction was also computationally predicted between the two loci in MCF-7 (**Figure 5C**). Our integrative analysis thus implicated *ZMIZ1* to be the top candidate target gene for the locus. Finally, we found GATA family binding motifs to be significantly disrupted by the SNP (**Figure 5D**), consistent with the ChIP-seq data. Thus, a quick analysis based on ABC-GWAS found the triplet (rs1250003, *ZMIZ1*, GATA) to be a novel putative functional mechanism behind the GWAS SNP rs704010 for increasing risk for breast cancer.

DISCUSSION

We demonstrated the capability of ABC-GWAS to find known, as well as novel, functional mechanisms of breast cancer GWAS

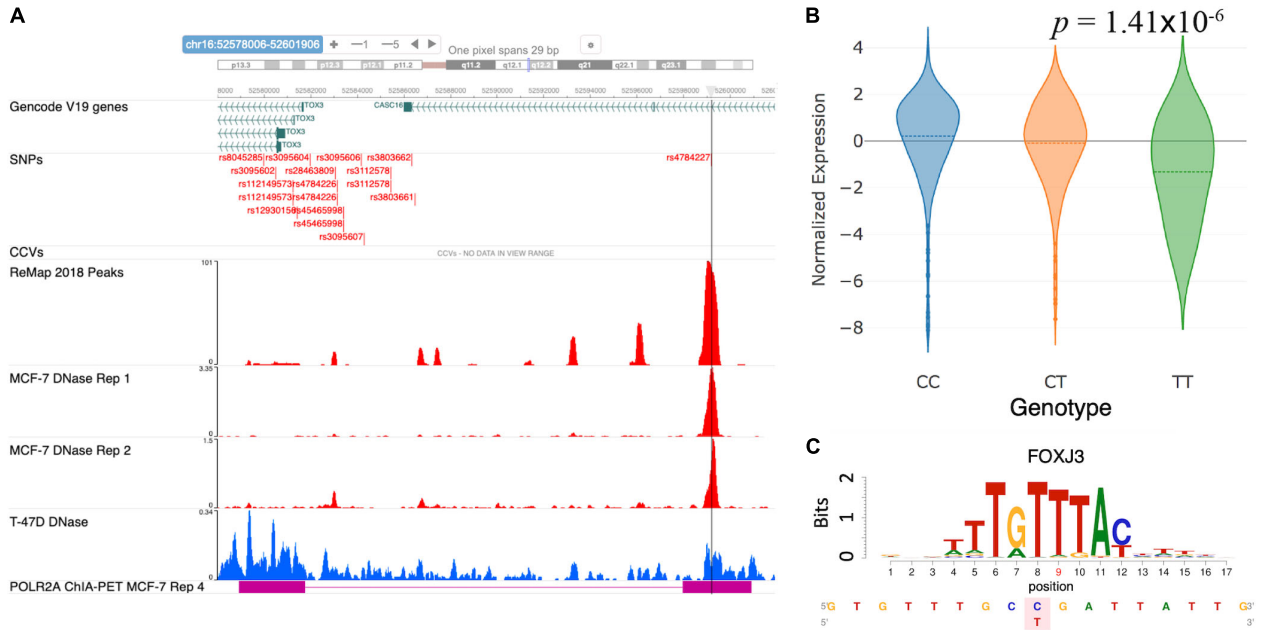

FIGURE 4 | Case study of GWAS locus at *TOX3*: (rs4784227, *TOX3*, FOXA1). **(A)** Embedded genome browser showing the SNP rs4784227 within an open chromatin region (MCF-7 DNase tracks) and several TF peaks ("ReMap 2018 Peaks"). **(B)** eQTL plot showing significant correlation between *TOX3* expression and rs4784227 genotypes. **(C)** Predicted TF candidate FOXJ3 motif from Kheradpour and Kellis (2014).

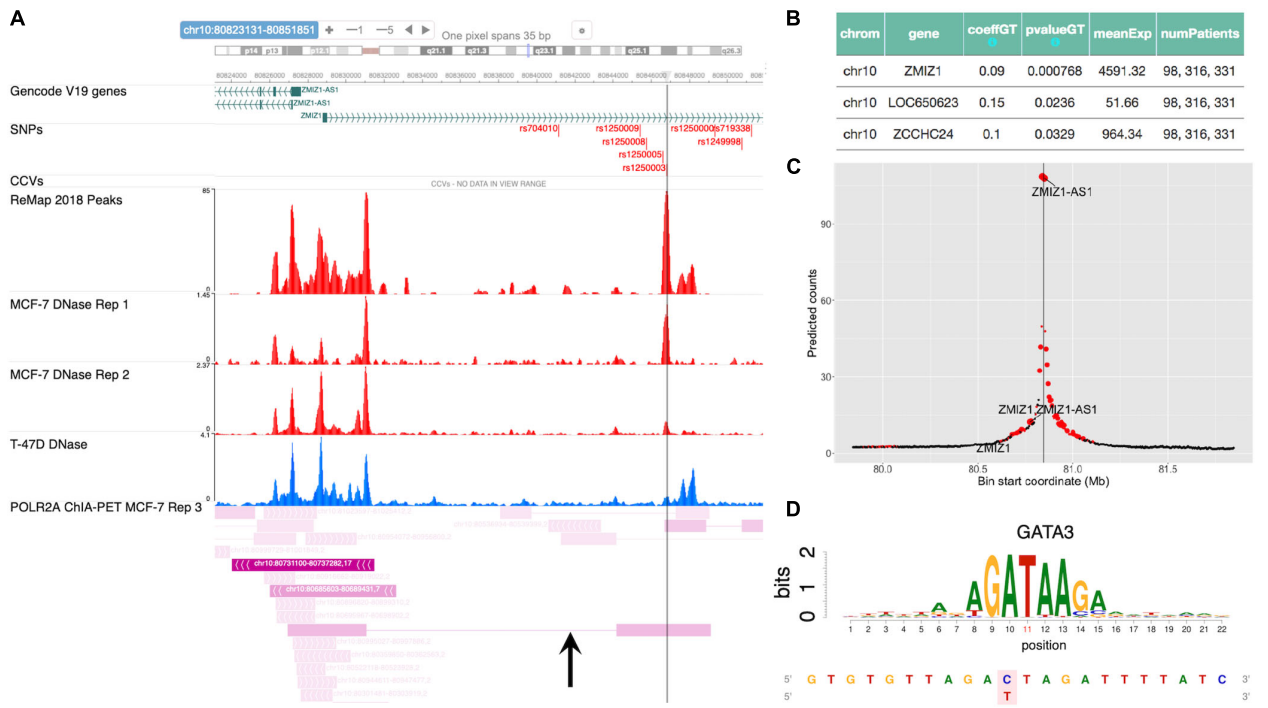

FIGURE 5 | Case study of GWAS locus at *ZMIZ1*: (rs1250003, *ZMIZ1*, GATA). **(A)** Embedded genome browser showing the LD SNP rs1250003 within an open chromatin region (MCF-7 DNase track) and several TF peaks ("ReMap 2018 Peaks"). Black arrow indicates a chromatin looping interaction between the SNP locus and the *ZMIZ1* promoter. **(B)** eQTL results showing the top significant target gene of the GWAS SNP to be *ZMIZ1*. **(C)** Average predicted chromatin contact counts across several HiC-Reg models in MCF-7, showing the *ZMIZ1* promoter as one of the loci significantly interacting with the LD SNP. **(D)** Predicted TF candidate GATA motif from Kheradpour and Kellis (2014).

loci. The computational and organizational framework of ABC-GWAS can be readily extended to other cancers. Once a (SNP, target gene, TF) triplet is identified through ABC-GWAS, several molecular experiments can be performed to validate the prediction. For example, the genotype of the predicted causative SNP could be modified through CRISPR-Cas9 base editors to study its effect on target gene expression (Komor et al., 2016). ChIP-quantitative polymerase chain reaction (ChIP-qPCR) is one way to measure how the SNP's genotype status modulates the binding affinity of the predicted TF. ABC-GWAS thus provides a valuable resource, currently not available in other databases, for functional characterization of GWAS results. ABC-GWAS currently contains analysis results for only a predetermined set of SNPs, and a useful future extension could allow our integrative analysis pipeline to be performed on any genetic variant of interest chosen by the user. Another informative feature could be to provide a pathway analysis of candidate target genes and transcription factors in the context of breast cancer biology.

ABC-GWAS is an interactive web resource containing results from an integrative functional analysis of ER+ breast cancer variants. We combined data from TCGA, ENCODE, and several motif databases to create a comprehensive resource that includes an embedded genome browser with relevant tracks in breast cancer cell lines and several modules describing results from eQTL, motif, and expression correlation analyses. Using our resource, we have verified the known functional mechanism of a genetic variant regulating the gene *TOX3* and also proposed a novel mechanism targeting the *ZMIZ1* locus. ABC-GWAS aims to take GWAS discoveries to the next level by providing a one-stop resource for in-depth functional analyses critical for interpreting and prioritizing GWAS variants. We thus hope that our resource will help both experimental and computational researchers accelerate breast cancer research.

SOFTWARE AVAILABILITY

1. Project name: ABC-GWAS.
2. Project home page: http://education.knoweng.org/abc-gwas/.
3. Operating system(s): Platform independent.
4. Programming language: HTML, JavaScript, R, Python.
5. Other requirements: JavaScript supporting web browser.
6. License: GNU GPL v3.0.

DATA AVAILABILITY STATEMENT

The datasets analyzed in the current study are available in the ENCODE project (https://www.encodeproject.org/) and the TCGA repository (http://cancergenome.nih.gov/) through GDC (https://portal.gdc.cancer.gov/projects) and dbGaP (https://www.ncbi.nlm.nih.gov/gap).

ETHICS STATEMENT

The usage of NIH controlled-access datasets was approved by the NCBI dbGaP.

AUTHOR CONTRIBUTIONS

JS contributed to the conception and design of the project. MM, YZ, and SZ contributed to the data curation and analysis. MM, YZ, SZ, SR, and JS contributed to the methodology. MM and YZ developed the resource and visualization tools. SR, PP-P, and JS supervised the research and secured the funding. MM and JS wrote the original draft. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by funds from the National Institutes of Health (NIH) R01CA163336, NIH U54GM114838, and the Grainger Engineering Breakthroughs Initiative to JS, Planning Grant from the Cancer Center at Illinois to PP-P and JS, and NIH R01-HG010045-01, NIH U54 AI117924, and UW Madison Vilas fellowship to SR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00730/full#supplementary-material

REFERENCES

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37(Suppl._2), W202–W208.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bauer, D. E., Kamran, S. C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* 342, 253–257. doi: 10.1126/science.1242088

Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., et al. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* 38, 1289–1297. doi: 10.1038/ng1901

Chan, C. S., and Song, J. S. (2008). CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Res.* 68, 9041–9049. doi: 10.1158/0008-5472.CAN-08-2632

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 46, D267–D275. doi: 10.1093/nar/gkx1092

Coetzee, S. G., Coetzee, G. A., and Hazelett, D. J. (2015). motifbreakR: an R/bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* 31, 3847–3849. doi: 10.1093/bioinformatics/btv470

Cowper-Sal lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., et al. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* 44, 1191–1198. doi: 10.1038/ng.2416

Darabi, H., McCue, K., Beesley, J., Michailidou, K., Nord, S., Kar, S., et al. (2015). Polymorphisms in a putative enhancer at the 10q21.2 breast cancer risk locus regulate NRBF2 expression. *Am. J. Hum. Genet.* 97, 22–34. doi: 10.1016/j.ajhg.2015.05.002

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Dunning, A. M., Michailidou, K., Kuchenbaecker, K. B., Thompson, D., French, J. D., Beesley, J., et al. (2016). Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat. Genet.* 48, 374–386. doi: 10.1038/ng.3521

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Fachal, L., Aschard, H., Beesley, J., Barnes, D. R., Allen, J., Kar, S., et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat. Genet.* 52, 56–73. doi: 10.1038/s41588-019-0537-1

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., et al. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.* 92, 489–503. doi: 10.1016/j.ajhg.2013.01.002

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS Era: from association to function. *Am. J. Hum. Genet.* 102, 717–730. doi: 10.1016/j.ajhg.2018.04.002

Gallagher, M. D., Posavi, M., Huang, P., Unger, T. L., Berlyand, Y., Gruenewald, A. L., et al. (2017). A dementia-associated risk variant near TMEM106B alters chromatin architecture and gene expression. *Am. J. Hum. Genet.* 101, 643–663. doi: 10.1016/j.ajhg.2017.09.004

Ghoussaini, M., Edwards, S. L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., et al. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat. Commun.* 4:4999. doi: 10.1038/ncomms5999

Ghoussaini, M., French, J. D., Michailidou, K., Nord, S., Beesley, J., Canisus, S., et al. (2016). Evidence that the 5p12 variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through FGF10 and MRPS30 regulation. *Am. J. Hum. Genet.* 99, 903–911. doi: 10.1016/j.ajhg.2016.07.017

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., et al. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res.* 46, D971–D976. doi: 10.1093/nar/gkx861

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375, 1109–1112. doi: 10.1056/NEJMp1607591

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8:R24.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Huang, Q., Whitington, T., Gao, P., Lindberg, J. F., Yang, Y., Sun, J., et al. (2014). A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat. Genet.* 46, 126–135. doi: 10.1038/ng.2862

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., and Bulyk, M. L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 43, D117–D122. doi: 10.1093/nar/gku1045

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-binding specificities of human transcription factors. *Cell* 152, 327–339. doi: 10.1016/j.cell.2012.12.009

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 46:D1284. doi: 10.1093/nar/gkx1188

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* 42, 2976–2987. doi: 10.1093/nar/gkt1249

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424. doi: 10.1038/nature17946

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 46, D252–D259. doi: 10.1093/nar/gkx1106

Leslie, R., O'Donnell, C. J., and Johnson, A. D. (2014). GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* 30, i185–i194. doi: 10.1093/bioinformatics/btu273

Li, D., Hsu, S., Purushotham, D., Sears, R. L., and Wang, T. (2019). WashU epigenome browser update 2019. *Nucleic Acids Res.* 47, W158–W165. doi: 10.1093/nar/gkz348

Li, M. J., Liu, Z., Wang, P., Wong, M. P., Nelson, M. R., Kocher, J. P., et al. (2016). GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* 44, D869–D876. doi: 10.1093/nar/gkv1317

Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., et al. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 152, 633–641. doi: 10.1016/j.cell.2012.12.034

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, A. Y., Finucane, H. K., et al. (2016a). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443–1448. doi: 10.1038/ng.3679

Loh, P. R., Palamara, P. F., and Price, A. L. (2016b). Fast and accurate long-range phasing in a UK biobank cohort. *Nat. Genet.* 48, 811–816. doi: 10.1038/ng.3571

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* 47, 373–380. doi: 10.1038/ng.3242

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* 45, 353–361. doi: 10.1038/ng.2563

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94. doi: 10.1038/nature24284

Munz, M., Wohlers, I., Simon, E., Busch, H., Schaefer, A. S., and Erdmann, J. (2019). Qtlizer: comprehensive QTL annotation of GWAS results. *bioRxiv* [Preprint]. doi: 10.1101/495903

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466, 714–719. doi: 10.1038/nature09266

Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E., and van Nimwegen, E. (2013). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.* 41, D214–D220. doi: 10.1093/nar/gks1145

Plotly (2018). *Plotly JavaScript Graphing Library*. Available online at: https://plot.ly/javascript/ (accessed September 11, 2018).

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Shannon, P. (2017). *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. R Package Version 1.18.0*.

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gomez-Marin, C., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375. doi: 10.1038/nature13138

Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* 32, 3207–3209. doi: 10.1093/bioinformatics/btw373

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., et al. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* 42, 504–507. doi: 10.1038/ng.586

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22, 1798–1812. doi: 10.1101/gr.139105.112

WashU (2019). *EpiGenome Gateway - WashU EpiGenome Browser*. Available online at: https://github.com/epgg/eg (accessed March 19, 2019).

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* 9, 326–332. doi: 10.1093/bib/bbn016

Xie, Z., Hu, S., Blackshaw, S., Zhu, H., and Qian, J. (2010). hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics* 26, 287–289. doi: 10.1093/bioinformatics/btp631

Zhang, S., Chasman, D., Knaack, S., and Roy, S. (2019). In silico prediction of high-resolution Hi-C interaction matrices. *Nat. Commun.* 10:5449. doi: 10.1038/s41467-019-13423-8

Zhang, Y., Manjunath, M., Yan, J., Baur, B. A., Zhang, S., Roy, S., et al. (2019). The cancer-associated genetic variant Rs3903072 modulates immune cells in the tumor microenvironment. *Front. Genet.* 10:754. doi: 10.3389/fgene.2019.00754

Zhang, Y., Manjunath, M., Zhang, S., Chasman, D., Roy, S., and Song, J. S. (2018a). Abstract 1220: integrative genomic analysis discovers the causative regulatory mechanisms of a breast cancer-associated genetic variant. *Cancer Res.* 78, 1220–1220.

Zhang, Y., Manjunath, M., Zhang, S., Chasman, D., Roy, S., and Song, J. S. (2018b). Integrative genomic analysis predicts causative Cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084. *Cancer Res.* 78, 1579–1591. doi: 10.1158/0008-5472.CAN-17-3486

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.