PLOS | GENETICS

RESEARCH ARTICLE

Identification of breast cancer associated variants that modulate transcription factor binding

Yunxian Liu[1], Ninad M. Walavalkar[1], Mikhail G. Dozmorov[2], Stephen S. Rich[3], Mete Civelek[3,4], Michael J. Guertin[1,3]*

1 Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America, **2** Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, United States of America, **3** Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, United States of America, **4** Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, United Statess of America

* guertin@virginia.edu

Citation: Liu Y, Walavalkar NM, Dozmorov MG, Rich SS, Civelek M, Guertin MJ (2017) Identification of breast cancer associated variants that modulate transcription factor binding. PLoS Genet 13(9): e1006761. https://doi.org/10.1371/journal.pgen.1006761

Editor: Scott M. Williams, Case Western Reserve University School of Medicine, UNITED STATES

Received: November 16, 2016

Accepted: April 12, 2017

Published: September 28, 2017

Data Availability Statement: All data were deposited in the Gene Expression Omnibus (GEO) database (accession number GSE89013).

Funding: This research was supported by the Farrow Fellowship to YL, an NCI Cancer Center Support Grant P30 CA44579 subaward to MJG and YL, and an American Cancer Society #IRG 81-001-26 subaward to MJG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Genome-wide association studies (GWAS) have discovered thousands loci associated with disease risk and quantitative traits, yet most of the variants responsible for risk remain uncharacterized. The majority of GWAS-identified loci are enriched for non-coding single-nucleotide polymorphisms (SNPs) and defining the molecular mechanism of risk is challenging. Many non-coding causal SNPs are hypothesized to alter transcription factor (TF) binding sites as the mechanism by which they affect organismal phenotypes. We employed an integrative genomics approach to identify candidate TF binding motifs that confer breast cancer-specific phenotypes identified by GWAS. We performed *de novo* motif analysis of regulatory elements, analyzed evolutionary conservation of identified motifs, and assayed TF footprinting data to identify sequence elements that recruit TFs and maintain chromatin landscape in breast cancer-relevant tissue and cell lines. We identified candidate causal SNPs that are predicted to alter TF binding within breast cancer-relevant regulatory regions that are in strong linkage disequilibrium with significantly associated GWAS SNPs. We confirm that the TFs bind with predicted allele-specific preferences using CTCF ChIP-seq data. We used The Cancer Genome Atlas breast cancer patient data to identify ANKLE1 and ZNF404 as the target genes of candidate TF binding site SNPs in the 19p13.11 and 19q13.31 GWAS-identified loci. These SNPs are associated with the expression of ZNF404 and ANKLE1 in breast tissue. This integrative analysis pipeline is a general framework to identify candidate causal variants within regulatory regions and TF binding sites that confer phenotypic variation and disease risk.

Author summary

The promise of effective personalized medicine is dependent upon the ability to identify genetic variants in the population that influence disease risk and then use this

Competing interests: The authors have declared that no competing interests exist.

information to accurately predict the likelihood of disease incidence for individual patients. High-risk individuals may be entered into clinical trails, pre-clinical intervention strategies, or increased frequency of screening to detect early disease onset. However, the contribution of any one genetic variant to increase disease susceptibility is typically small, with many potential causal variants in the genomic region associated with risk. Therefore, it is important to understand the biological mechanisms by which the variants within a genetic region influence disease susceptibility by refining the set of all variants to those that are highly plausible to be causal. Herein, we describe a method to integrate molecular genomics data with genetic epidemiological data to inform on the underlying molecular mechanisms that influence breast cancer risk. This approach identifies the important transcription factors that directly regulate gene expression to modulate disease susceptibility.

Introduction

Genome-wide association studies (GWAS) have identified more than 90 genomic loci and common genetic variants associated with breast cancer [1–5]. The single nucleotide polymorphisms (SNPs) associated with breast cancer have been shown to be enriched in DNA regulatory regions [6, 7], with few residing in coding regions of genes. The mechanisms by which most of these variants contribute to breast cancer biology remain unknown [8–11]. The effects of putative causal non-coding SNPs are challenging to interpret as they may alter transcription factor (TF) binding sites [12], lncRNA structure [13], splicing [14], transcription start or termination signals, or DNA shape [15]. Non-coding SNPs that alter TF binding sites are the most easily interpreted because they have the potential to modulate gene expression to mediate their effects on disease risk [16]. Therefore, it is possible to identify putative causal SNPs by focusing on those that alter TF binding sites in breast tissue.

TF dysregulation is a hallmark of many cancers [17, 18]. Genes encoding TFs in tumor cells are often amplified, deleted, rearranged via chromosomal translocation, or subjected to point mutations that result in a gain- or loss-of-function [18]. For example, transcriptional amplification of c-Myc reduces rate-limiting constraints for tumor cell growth and proliferation; high c-Myc expression correlates with tumor aggression and poor clinical outcome [19]. Estrogen receptor (ER) is a TF that regulates cell proliferation, which is the defining feature of luminal breast cancers [20]. Identifying the full set of TFs that function within a cell type remains a challenge.

Enzymatic accessibility assays identify open chromatin in the genome, which is an indirect measure of regulatory element activity and TF binding events [21–25]. TFs that directly or indirectly recruit cofactors, such as histone modifiers and nucleosome remodelers, recognize sequence motifs that are enriched in regions of open chromatin characterized by active histone marks and enzymatic hypersensitivity peaks [26, 27]. These cofactors maintain the chromatin structure at regulatory elements. Therefore, one strategy for identifying the functional TFs in a cell is to query the sequence underlying accessible regions for over-represented motifs [28–31]. Alternatively, depletion of signal within hypersensitive regions (footprints) [30, 31] coupled to motif analysis can be used to infer TF binding. The reliance on hypersensitivity footprinting to define a near-comprehensive set of TF motifs is limiting because many TFs do not have footprints for biological reasons [31] and the enzymes exhibit sequence-specificity that can be misinterpreted as footprint signatures [31–33]. These methods strictly identify motifs that are over-represented within regions of open chromatin, but families of TFs often contain paralogous DNA binding domains and thus recognize indistinguishable sequence motifs. One can directly measure the expression of TFs to identify putative functional TFs among related TFs [34]; however, expression of a TF is an imperfect proxy for TF function. For example, many nuclear receptors are not transcriptionally functional in the absence of ligand regardless of expression levels. Here we propose an approach that integrates open chromatin genomic data and gene expression data to identify candidate TFs that are functional in breast cancer-relevant cells.

Expression quantitative trait loci (eQTL) analysis identifies genetic variants that correlate with gene expression differences in a population. eQTL analysis complements genetic association data by predicting causal genes whose expression differences dictate organismal phenotypes [35, 36]. The gene that is responsible for a trait may be located relatively far from the GWAS associated SNPs, as the causal SNPs may modulate TF binding and TFs can act distally to regulate gene transcription. In these cases, preferential binding of a TF to one allele causes differential regulation of gene expression to confer the phenotype [37]. Several studies have provided evidence of causal relationships for gene expression mediating the association between GWAS SNPs and traits [16, 38]. Phenotype-associated SNPs are enriched for eQTLs, suggesting that eQTL analysis can enhance discovery of causal genes associated with complex phenotypes [39, 40].

The goal of this study is to gain mechanistic insight into how breast cancer susceptibility alleles confer disease risk. First we identify sequence motifs in breast-relevant tissue and cell lines that are enriched within enzymatic hypersensitive sites and support their functional role with evolutionary conservation analysis and gene expression data. Next we identify breast cancer risk alleles that are predicted to modulate TF binding and we perform eQTL analysis to find candidate causal risk genes.

Results

Identification of functional TFs within breast-derived cells and tissues

Regulatory regions are bound by a milieu of protein complexes and regulatory nucleic acids, such as transcription factors, histone modifiers, nucleosome remodelers, and lncRNA. Sequence-specific TFs are directly or indirectly responsible for the recruitment of downstream transcriptional modifiers in a sequence-dependent manner. Therefore, we sought to identify all TFs that bind within open chromatin in breast cancer-relevant cells and tissues. We performed ATAC-seq in a mammary epithelial cell line (MCF10A) to complement publicly available DNase-seq data from Encyclopedia of DNA Elements (ENCODE) [41] and the Roadmap Epigenomics project [42]. We quantified open chromatin and identified regulatory elements genome-wide in five breast cancer-relevant cell lines and breast tissue: MCF7 cells, MCF10A cells, T47D cells, cultured human mammary epithelial cells (HMEC), and primary breast variant HMECs (vHMEC). Many regions of chromatin accessibility are shared between the cell lines, although the degree of accessibility for a region can vary between cell types (Fig 1). Enzymatic accessibility coverage (peaks) [28, 29] or depletions of signal in hypersensitive region (footprints) [30, 31] coupled to motif analysis are routinely used to identify TFs that maintain open chromatin structure.

We performed iterative rounds of *de novo* motif analysis using the sequence underlying enzymatic hypersensitivity peaks to identify overrepresented TF recognition sites in each data set (S1 File). All but one of the motifs we identified were previously characterized and described in databases [43–47]. These motifs found in hypersensitive peaks are, on average, evolutionarily conserved (Fig 2). Additionally, we identified a potential TF recognition sequence that has no known cognate TF binding partner; we refer to this sequence element as

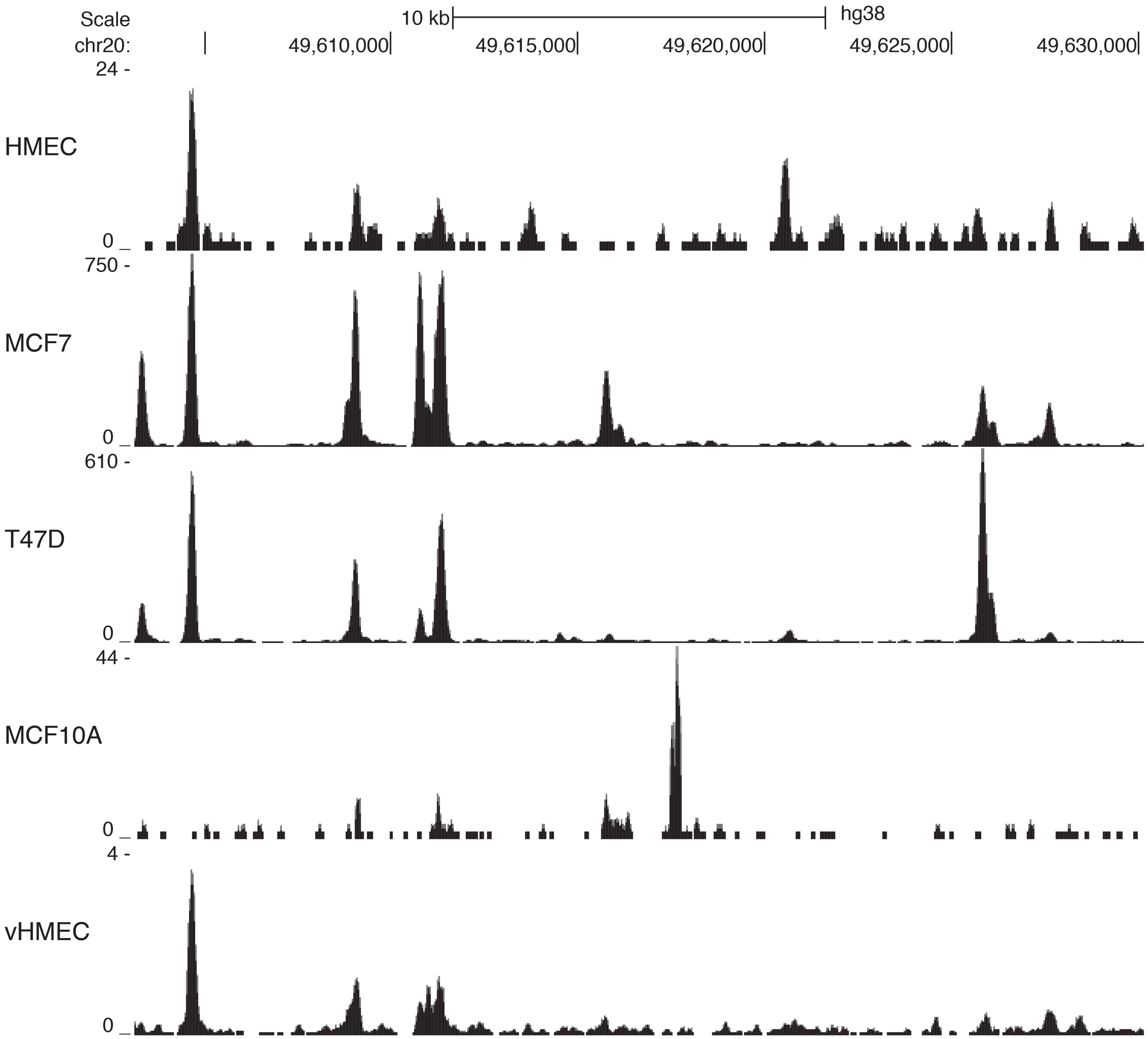

Fig 1. DNase-seq and ATAC-seq quantify differential open chromatin in breast cancer-relevant tissue and cell lines. We show smoothed DNase-seq and ATAC-seq tracks at a locus on chromosome 20. MCF7, T47D, HMEC and vHMEC DNase-seq data share several chromatin accessibility regions and MCF10A ATAC-seq data identifies a peak of differential chromatin accessibility that is distinct from the other data sets.

https://doi.org/10.1371/journal.pgen.1006761.g001

an *orphan motif* (Fig 2B). This orphan motif is evolutionarily conserved, as measured by phastCons [48] and phyloP [49] scores, in hypersensitivity peaks. Hypersensitivity footprints result from protection of the DNA by a bound TF [50]; however, approximately half of all TF-bound motifs do not exhibit composite footprints [31, 51] because the TF dissociates during the nuclei isolation procedure. We corrected the DNase data for intrinsic sequence bias [33], but we do not observe a composite footprint for this orphan motif (Fig 3B). However, the hypersensitivity pattern surrounding the motif is not uniform—the region downstream of this orphan motif is more hypersensitive than upstream. This directional pattern of enzyme accessibility is common with many TFs [29], including CTCF (Fig 3A). We hypothesize that this

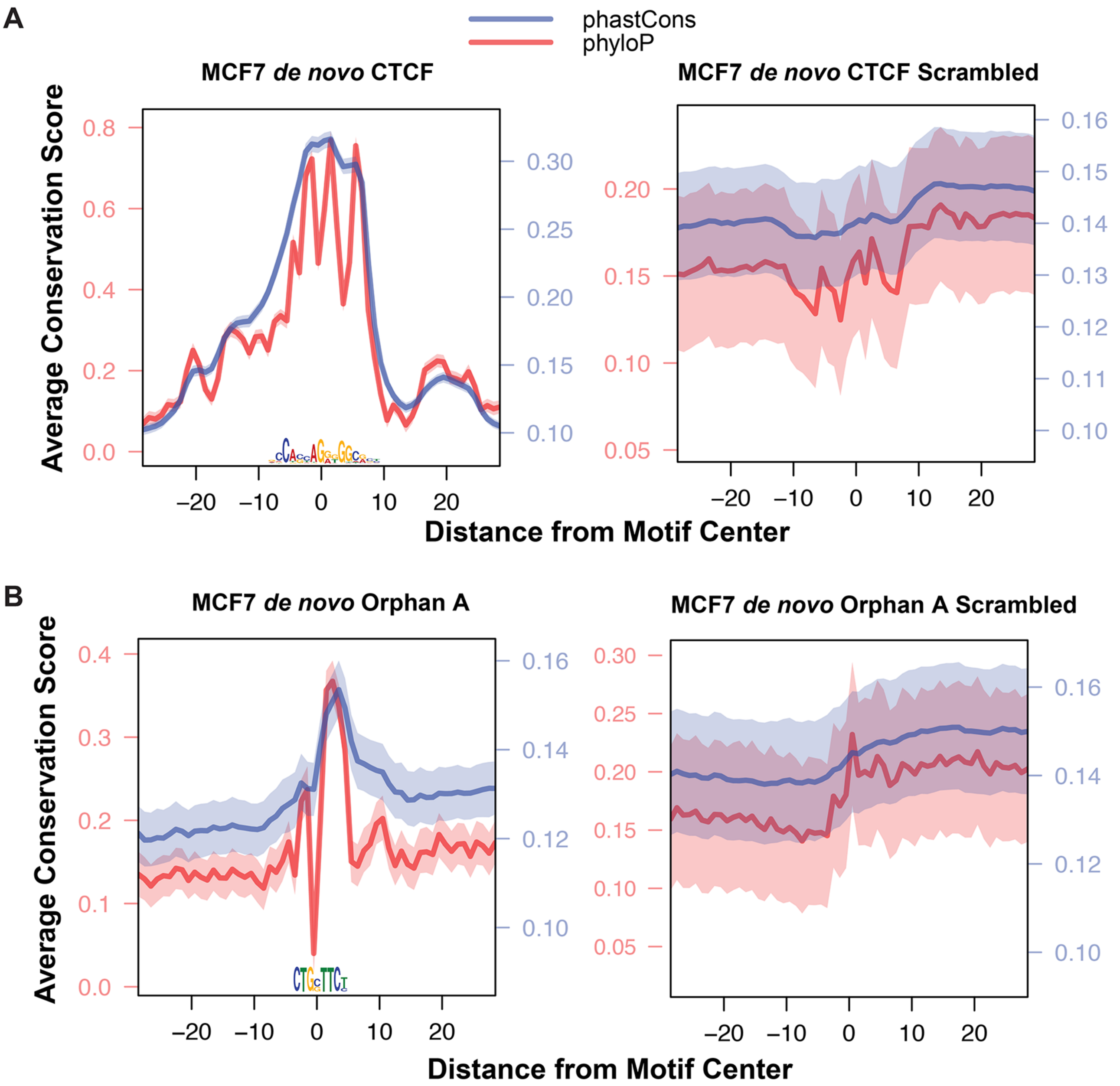

Fig 2. *De novo* identified motifs within open chromatin are evolutionarily conserved. (A) CTCF motifs within hypersensitive sites are, on average, conserved; note the peak of phastCons (blue trace) and pyhloP (red trace) intensity at the motif compared to the flanking region. The right panel is an average of 20 scrambled CTCF weight matrices. We do not observe any conservation peak after scrambling. (B) A novel orphan motif, which does not have a known cognate TF partner, is also evolutionarily conserved within open chromatin regions.

https://doi.org/10.1371/journal.pgen.1006761.g002

orphan motif is functional and directs chromatin accessibility by serving as a recognition site for an uncharacterized TF.

In addition to this orphan motif, we identified hundreds of position-specific weight matrices (PSWM) from our exhaustive analyses, many were redundant between cell lines and found in multiple rounds of motif analysis in the same cells. To reduce the complexity of these data, we mapped the comprehensive set of motifs found in MCF7, MCF10A, T47D, HMEC and vHMEC into distinct non-redundant motif families. This operation resulted in identification

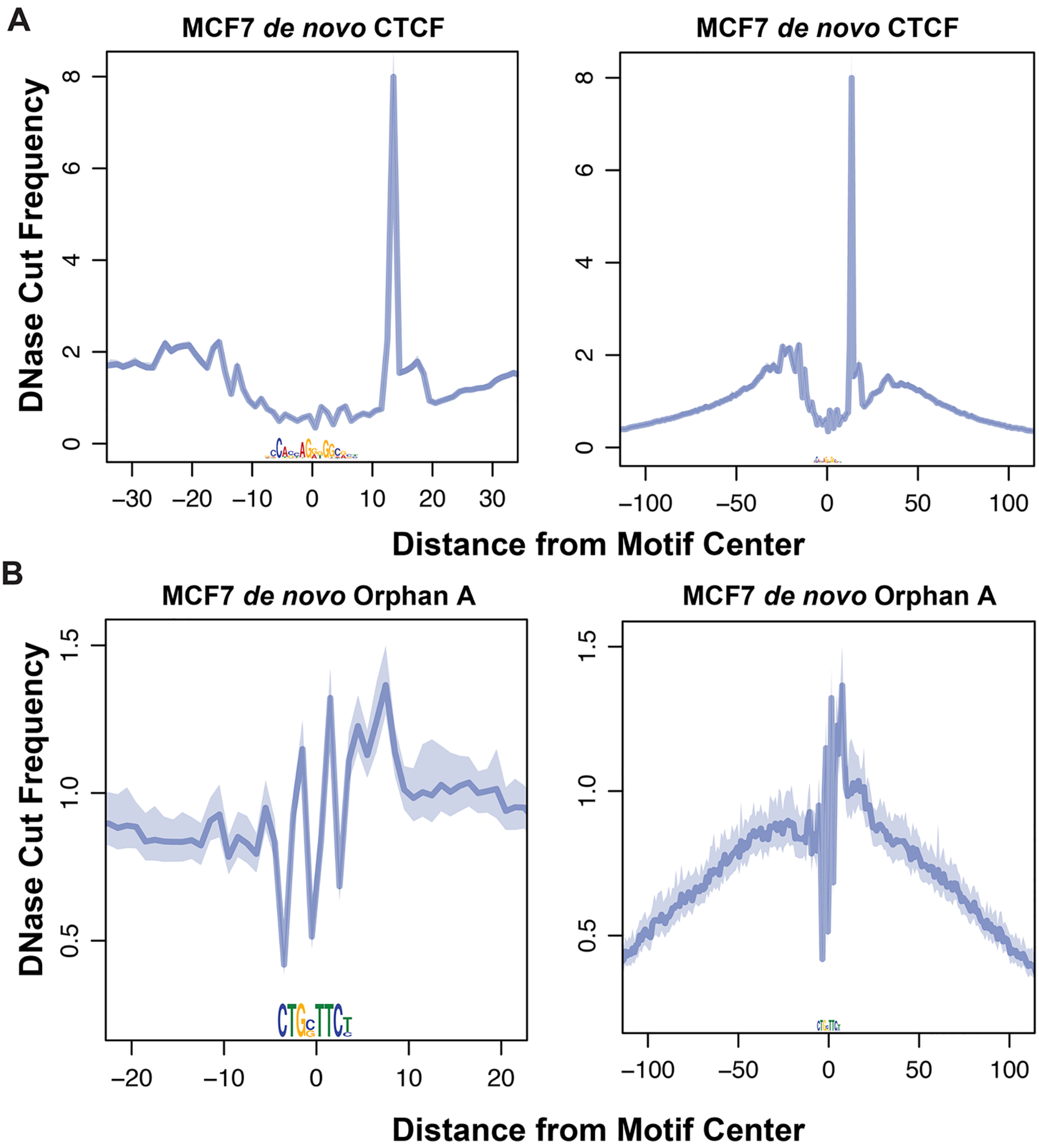

Fig 3. Composite footprints and directional patterns of enzyme accessibility indicate TF occupancy. (A) The CTCF motif exhibits one of the most striking composite footprints and directional patterns of accessibility among sequence-specific binding proteins. (B) We do not observe a composite footprint for this orphan motif; however, the enzyme accessibility pattern around this motif is directional, with higher degree of cleavage downstream from the motif.

https://doi.org/10.1371/journal.pgen.1006761.g003

of 37 sequence motif classes across the five breast cancer-relevant cell lines and tissue (S1 Fig). Many TFs share paralogous DNA binding domains and these TFs often recognize the same sequence motifs. We defined the full set of TFs that recognize each motif by using known TF/sequence interaction data from ChIP-seq [43, 45, 52], protein binding microarrays [53, 54], and SELEX [46, 47] data. We identified 23 TF families in at least two cell lines/breast tissue (Table 1). Fourteen TF families were uniquely found in one cell line or tissue (Table 1). These 37 TF families represent the binding motifs for 235 distinct TFs (Table 1). However, these TFs are not all expressed in breast tissue. We identified the TFs that are most likely candidates for maintaining open chromatin and the gene regulatory expression profiles in breast-relevant cells by examining the relative expression of all of the TFs in each family using TCGA expression data (Fig 4 and S2 Fig). For example, ESR1, ESR2, and PPAR-γ contain paralogous DNA binding domains and they recognize indistinguishable sequence elements (S1 Fig). We find that ESR1 is the most highly expressed TF that recognizes the motif (Fig 4). This result is consistent with the biological role of ESR1 in the etiology of breast cancer and breast biology compared to ESR2 and PPAR-γ. Similarly, FOXA1 is the most well-charactered TF within the Forkhead Box family of TFs in terms of estrogen signaling and interplay with ER [55]. As expected, we find that FOXA1 is the most highly expressed TF in the family of 22 Forkhead Box TFs (Fig 4). We find that many of these highly expressed TFs correlate with breast cancer survival time in a subtype-specific manner (S3 Fig). Silencing of IRF7 pathways in breast cancer cells promotes breast cancer metastasis, and high expression of the IRF7-regulated genes with breast cancer is associated with prolonged survival [56]. Similarly, we find that high expression of IRF7 is correlated ($P = 0.029$) with positive breast cancer patient outcome in Luminal A subtype (S3 Fig). We find that high expression of BATF ($P = 0.0035$) and TP73 ($P = 0.0077$) is correlated with breast cancer patient survival in HER2+ and Basal-like subtypes, respectively (S3 Fig). Taken together, these data support the notion that TF expression levels may serve as biomarkers of patient outcome.

Breast cancer-associated genetic variants affect TF binding

A major goal of this study was to identify a set of plausible causal SNPs that modulate TF binding from a list of SNPs associated with breast cancer in GWAS-defined loci. We identified a total of 463 SNPs in strong LD ($r^2 \geq 0.8$) with the most associated breast cancer GWAS SNPs defined in 93 distinct genomic loci; these SNPs are predicted to affect the binding of TFs belonging to at least 30 TF families. Six examples of candidate causal breast cancer-associated SNPs are shown in Table 2. Transcription factors from the following TF families are predicted to have their binding affected: CTCF, GABPA, RUNX, GRHL2, USF1, ZBTB33, and ZNF143. For example, SNP rs11540855 (3′ UTR of ABHD8 on chromosome 19p13.11) is within a DNase-defined regulatory element in human mammary epithelial cells and should affect the binding of CTCF (Fig 5). rs11540855 is in strong LD with rs8170 ($r^2 = 0.98$), which is associated with breast cancer risk [58–60]. Likewise, rs3760982 (1.1kb 5′ of KCNN4 on chromosome 19q13.31) is associated with breast cancer susceptibility [3] and we find that its A allele is predicted to enhance RUNX binding (Fig 5). More examples of candidate causal breast cancer-associated SNPs disrupting TF binding sites within breast cancer GWAS loci are shown in the supporting information (S4 Fig).

Two SNPs (rs4414128 and rs8103622) are predicted to strongly affect CTCF binding (Fig 6 and S5 Fig); therefore, we tested our predictions of how the alleles would affect CTCF binding by analyzing ENCODE ChIP-seq data. The SNPs rs4414128 and rs8103622 showed allele-specific binding that was consistent with the direction predicted based on tolerated degeneracy from the consensus binding site (Fig 6 and S5 Fig). For rs4414128 (Fig 6), 34 of 44 cell types/

Table 1. Iterative *de novo* motif analysis identified a set of 37 overrepresented motif families within the regulatory elements in MCF7, MCF10A, T47D, HMEC and vHMEC cells. TFs that recognize similar regulatory sequences are clustered into families. Each row contains one TF family and we denote the cell lines/tissues with each identified TF family by check mark.

TF family	MCF7	MCF10A	T47D	HMEC	vHMEC
HNF4A HNF4G NR1H2 NR2C2 NR2F1 NR2F6 PPARG RXRA RXRB RXRG	✓				
BCL6B E2F1 E2F3 E2F4 E2F6 EGR1 KLF1 KLF12 KLF13 KLF14 KLF16 KLF4 KLF5 KLF7 SP1 SP2 SP3 SP4 SP8 ZNF410	✓	✓	✓	✓	✓
ARNT ARNTL BHLHE40 BHLHE41 CLOCK CREB3L2 MAX MITF MITF MLX MLXIPL MYCN NPAS2 RBPJ SREBF1 SREBF2 TCFL5 TFE3 TFEB TFEC USF1 USF2	✓				
ELK1 ELK3 ELK4 ERF ERG ETS1 ETV1 ETV2 ETV3 ETV4 ETV5 ETV6 FEV FLI1 GABPA	✓	✓	✓	✓	✓
MAZ PAX4 RREB1 RUNX1 RUNX2 ZNF263 ZNF281 ZNF740	✓	✓	✓	✓	✓
ATF1 ATF2 ATF3 ATF7 BATF3 CREB1 CREB3 CREB5 JDP2 JUN JUND MAFB XBP1	✓	✓	✓	✓	
ESRRA ESRRB ESRRG NR1H3 NR2F1 NR2F2 NR4A1 NR4A2 NR5A2 RARA RARB RARG RXRA	✓		✓	✓	
SOX10 SOX2 SOX21 SOX3 SOX4 SOX6 SOX9 SRY		✓			
TFAP2A TFAP2B TFAP2C	✓		✓	✓	
NF1 NFIA NFIB NFIC NFIX TLX1	✓		✓	✓	
FOXA1 FOXA2 FOXB1 FOXC1 FOXC1 FOXC2 FOXD1 FOXD2 FOXD3 FOXF2 FOXG1 FOXG1 FOXI1 FOXJ1 FOXJ2 FOXJ2 FOXJ3 FOXJ3 FOXK1 FOXL1 FOXL1 FOXO1 FOXO3 FOXO4 FOXP1 FOXP2 FOXP3	✓		✓	✓	
EHF ELF1 ELF3 ELF4 ELF5 ETS1 SPDEF SPI1		✓			
RFX1 RFX2 RFX3 RFX4 RFX5 RFX7	✓		✓		✓
ZBTB33	✓		✓	✓	✓
CEBPA CEBPB	✓			✓	
IRF1 IRF2 IRF3 IRF4 IRF5 IRF6 IRF7 IRF8 IRF9 PRDM1 STAT1 STAT2			✓		
GRHL1 GRHL2 TFCP2 TFCP2L1	✓		✓	✓	
ZNF143	✓	✓	✓	✓	✓
ATF3 BATF FOS FOSL1 FOSL2 JDP2 JUN JUNB JUND NFAT5 NFATC1 NFATC3 NFE2	✓	✓			
ATF4 DDIT3	✓				
NFYA NFYB	✓	✓	✓	✓	✓
NRF1	✓	✓	✓	✓	✓
YY1		✓			
NFKB1 NFKB2 REL RELA	✓				✓
TEAD1 TEAD2 TEAD3 TEAD4	✓		✓		
POU1F1 POU2F1 POU2F2 POU2F3 POU3F1 POU3F2 POU3F3 POU3F4 POU5F1 POU5F1B SOX2	✓				
SOX1 SOX10 SOX11 SOX14 SOX15 SOX17 SOX18 SOX2 SOX21 SOX3 SOX4 SOX7 SOX8 SOX9 SRY				✓	
ESR1 ESR2 PPARG			✓		
CTCF	✓	✓	✓	✓	✓
REST	✓		✓	✓	✓
BACH1 BACH2 MAF MAFA MAFB MAFF MAFG MAFK MAFK NFE2 NFE2L2 NRL	✓	✓	✓	✓	
TFAP2A TFAP2B TFAP2C TFAP2E					✓
EGR1 EGR2 EGR3 EGR4	✓	✓	✓		
TP53 TP63 TP73	✓			✓	✓

(*Continued*)

Table 1. (*Continued*)

TF family	MCF7	MCF10A	T47D	HMEC	vHMEC
CEBPA					✓
HBP1			✓		
SMAD3		✓			

TF family found by iterative *de novo* motif analysis in a specific cell line/tissue is denoted by check mark (✓).

https://doi.org/10.1371/journal.pgen.1006761.t001

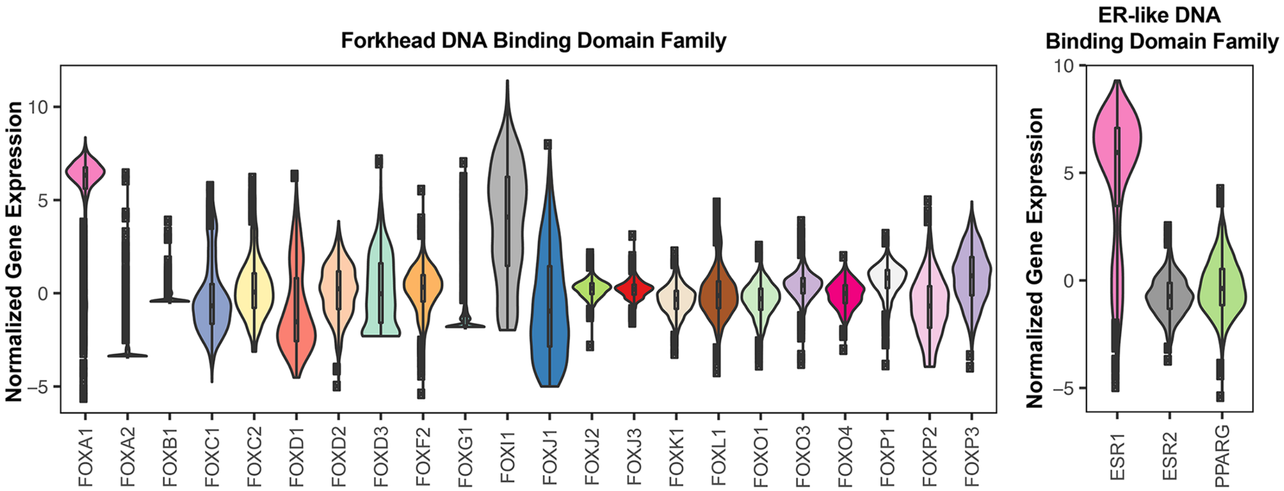

Fig 4. The most highly expressed TFs with paralogous DNA binding domains are most relevant to breast cancer. The relative expression of TFs that recognize the same sequence motif can identify the top candidate functional TFs. FOXA1 and ESR1 are the highest expressed TF in each of their TF families. We quantified gene expression using TCGA breast cancer patient solid tumor samples [57].

https://doi.org/10.1371/journal.pgen.1006761.g004

Table 2. Six examples of SNPs that are associated with breast cancer susceptibility and predicted to affect TF binding and gene expression regulation. Candidate SNPs are: 1) in strong LD ($r^2 \geq 0.08$) with the most associated breast cancer GWAS SNP; 2) within DNase/ATAC-seq defined regulatory region of MCF7, MCF10A, T47D, HMEC or vHMEC cell lines and tissue; 3) contain high information content (IC) in the TF binding PSWM ($IC \geq 0.5$); and 4) are eQTLs in breast cancer patient solid tumor sample and GTEx breast tissue.

Candidate Causal SNP	Breast Cancer GWAS SNP	Affect Gene Expression	TF Affected Binding	eQTL P-value in Breast Cancer Tumor	eQTL P-value in Breast Tissue
rs11540855	rs8170 $r^2 = 0.90$	ANKLE1	Affects CTCF binding	2.75E-10	9.40E-07
rs73509996	rs8170 $r^2 = 1.00$	ANKLE1	Affects GABPA binding	4.35E-10	7.40E-07
rs3760982	rs3760982	ZNF404	Affects RUNX binding	3.16E-07	1.50E-06
rs11665924	rs3760982 $r^2 = 1.00$	ZNF404	Affects GRHL2 binding	3.40E-07	1.50E-06
rs11669175	rs3760982 $r^2 = 1.00$	ZNF404	Affects USF1 binding Affects ZBTB33 binding	1.13E-06	1.50E-06
rs4802200	rs3760982 $r^2 = 1.00$	ZNF404	Affects ZNF143 binding	2.26E-06	3.00E-06

https://doi.org/10.1371/journal.pgen.1006761.t002

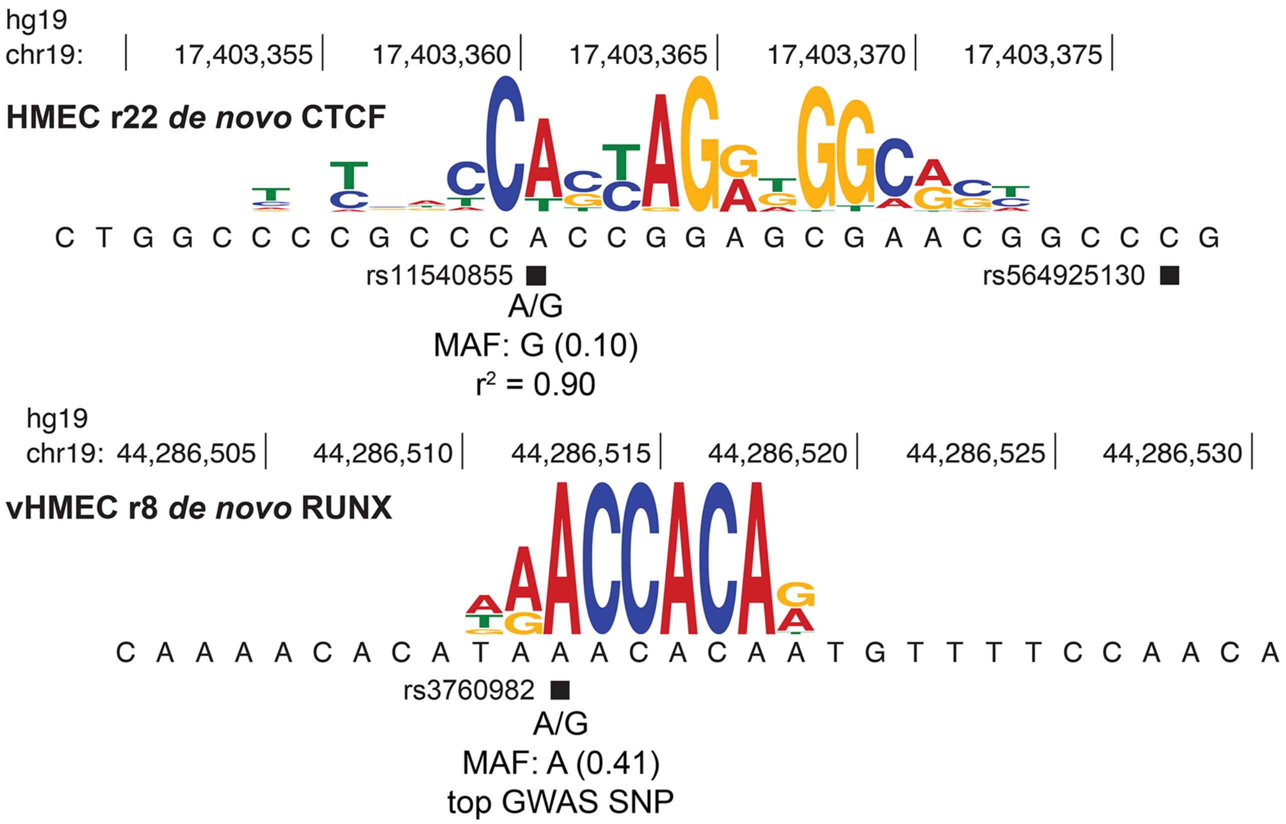
hg19
chr19: 17,403,355 17,403,360 17,403,365 17,403,370 17,403,375
HMEC r22 de novo CTCF
C T G G C C C C G C C C A C C G G A G C G A A C G G C C C G
rs11540855
rs564925130
A/G
MAF: G (0.10)
r2 = 0.90
hg19
chr19: 44,286,505 44,286,510 44,286,515 44,286,520 44,286,525 44,286,530
vHMEC r8 de novo RUNX
C A A A A C A C A T A A A C A C A A T G T T T T C C A A C A
rs3760982
A/G
MAF: A (0.41)
top GWAS SNP

Fig 5. Two examples of SNPs that are in strong LD ($r^2 \geq 0.8$) with the most associated breast cancer GWAS SNP and are predicted to affect the binding of TFs in breast cancer-relevant tissue and cell lines. Reference SNP rs11540855 is in strong LD ($r^2 = 0.90$) with the breast cancer GWAS SNP rs8170. The risk allele G of rs11540855 (MAF = 0.10) is predicted to compromise the binding of CTCF, which was identified in round 22 of *de novo* motif analysis from HMECs. Reference SNP rs3760982 is the most associated breast cancer GWAS SNP. Its risk allele A (MAF = 0.41) is predicted to enhance the binding of RUNX, which was identified in round 8 of *de novo* motif analysis from vHMEC.

https://doi.org/10.1371/journal.pgen.1006761.g005

replicates show the expected C preference with a range between 52% and 75%. Eight cell lines/replicates show modest allelic imbalance favoring the T allele (31–49% of the reads spanning the SNP); two experiments are balanced. Importantly, both replicates of human mammary epithelial cells (HMEC) show an allelic imbalance favoring C in 58% and 75% of the reads. Across 37 cell types (and replicates) that are heterozygous and normal karyotype, rs8103622 (S5 Fig) shows allele-specific preference of C as expected in 34 cell types/replicates with the range between 53% and 88%. The other three instances exhibit allelic balance, with allele frequencies between 48% and 52%.

SNPs that modulate TF binding correlate with ZNF404 and ANKLE1 expression

We identified rs11540855 as the most significant (P-value: 2.75E-10) eQTL SNP in the GWAS locus that is in strong LD ($r^2 = 0.90$) with the most associated breast cancer GWAS SNP rs8170 [58] (Fig 7C). The G/G genotype at rs11540855 is predicted to compromise CTCF binding (Fig 5) and G/G individuals have, on average, higher expression levels of ANKLE1 (Fig 7A and 7B). The most significantly associated GWAS SNP in this locus (rs8170-T) increases the risk of ER-negative breast cancer with an odds ratio of 1.10 [60] and the T allele is in LD with the G allele

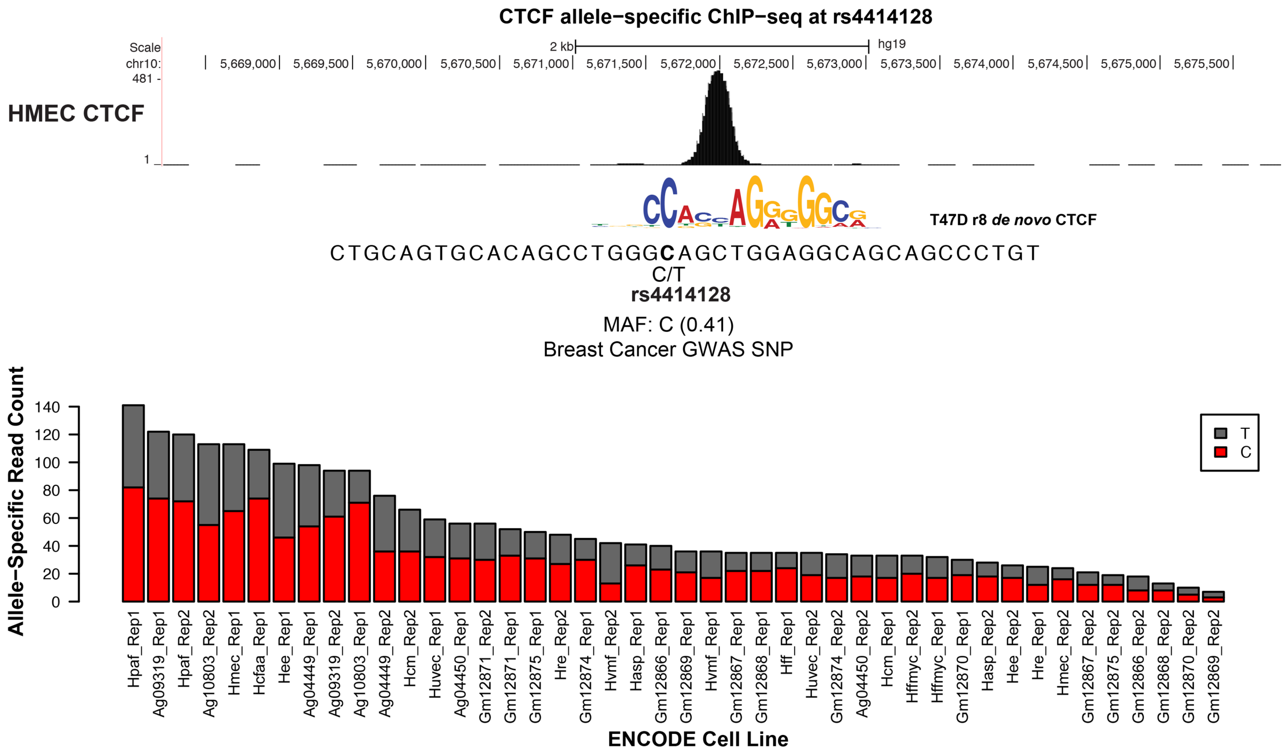

Fig 6. Reference SNP rs4414128 affects CTCF binding as measured by allele-specific ChIP-seq among many diploid, heterozygous cell lines. We analyzed allele-specific binding of all ENCODE cell lines with reported normal karyotype that are heterozygous at rs4414128. CTCF binding is unbalanced in favor of the C allele, which conforms more strongly to the consensus sequence.

https://doi.org/10.1371/journal.pgen.1006761.g006

of rs11540855. Therefore, higher expression of ANKLE1 (Fig 7A and 7B) is associated with increased breast cancer risk. We also identified rs3760982 as one of the most associated eQTL SNPs (P-value: 3.16E-07) and rs3760982 is correlated with ZNF404 expression (Fig 8C). The A/A genotype at rs3760982 is predicted to increase RUNX binding (Fig 5) and is correlated with higher expression of ZNF404 in breast cancer tumor samples (Fig 8A) and breast tissue (Fig 8B). The rs3760982-A allele is associated with an increased risk (odds ratio of 1.06) of breast cancer [3], thus higher expression of ZNF404 correlates with increased breast cancer risk. Therefore, we prioritized SNPs within GWAS loci that are predicted to affect transcription factor binding and module expression of ANKLE1 and ZNF404 to confer breast cancer risk.

Discussion

GWAS have discovered more than 90 genetic loci and common genetic variants associated with breast cancer susceptibility [1–5], and the majority of SNPs in these loci are enriched in non-coding regions. Non-coding genetic variants can contribute to complex traits and diseases through many molecular mechanisms [12–14, 61, 62] including having an effect on TF binding affinities, which can result in differential gene expression. Herein, we describe an integrative genomics methodology to identify a near-comprehensive set of TFs that are actively maintaining open chromatin in a cell type. We identify which GWAS-relevant SNPs are predicted to modulate TF binding intensity and use ChIP-seq data to confirm our predictions. Lastly, we use eQTL data to identify the likely target genes of SNPs that affect TF binding affinity. Taken together, this approach can identify likely causal SNPs associated with breast cancer risk.

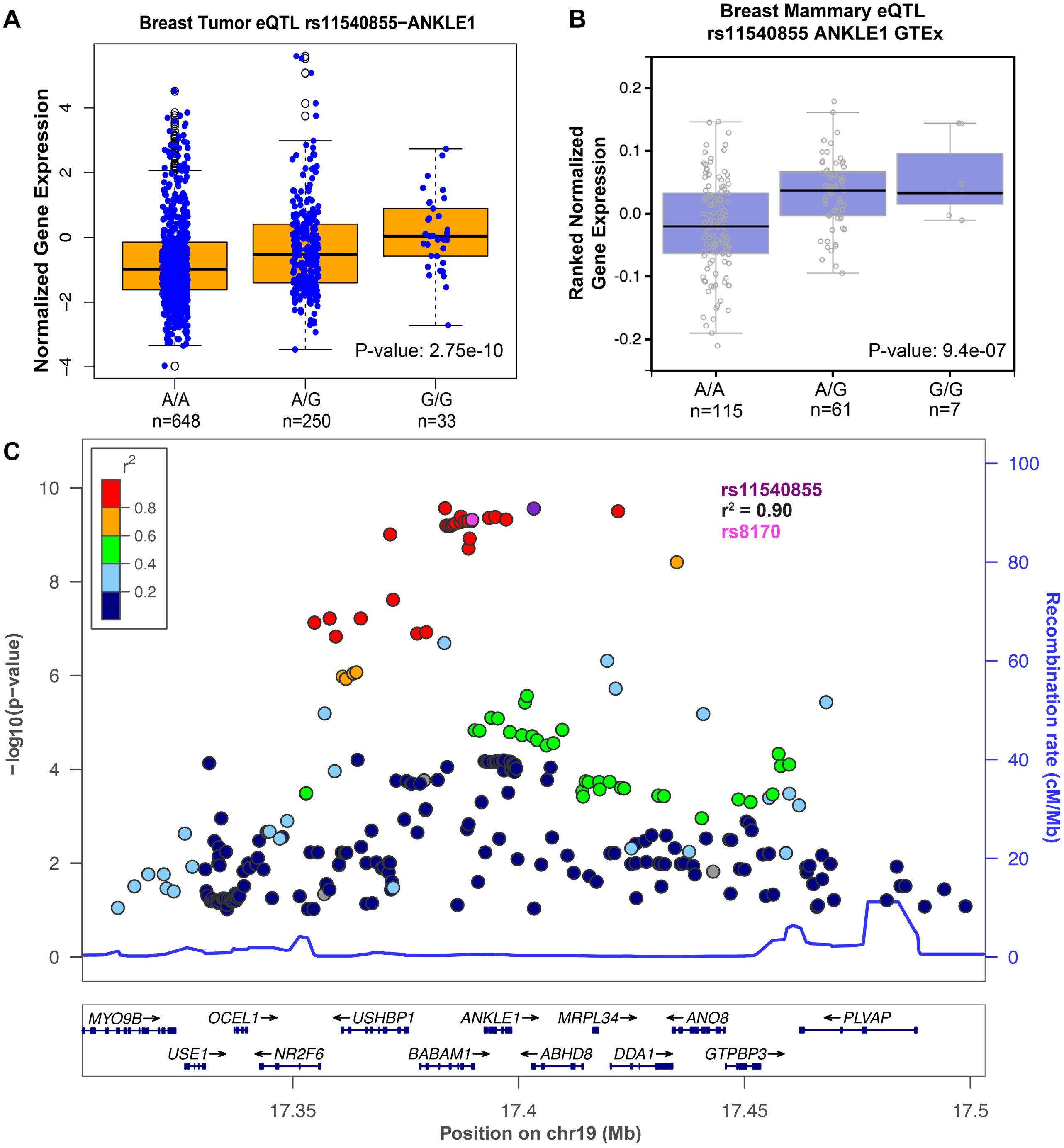

Fig 7. Candidate causal SNP rs11540855 is the strongest eQTL for ANKLE1. (A) rs11540855 genotypes are correlated with ANKLE1 expression. Genotype data was imputed from TCGA breast cancer patient blood sample and gene expression is from TCGA breast cancer patient solid tumor samples [57]. (B) GTEx data confirm the association of rs11540855 genotype with ANKLE1 expression using breast cancer tissue expression data. (C) SNP rs11540855 is the top eSNP in the same haloptype region. The SNP rs8170 (highlighted in pink) is the most associated GWAS hit for breast cancer susceptibility and rs8170 is in strong LD ($r^2 = 0.90$) with rs11540855.

https://doi.org/10.1371/journal.pgen.1006761.g007

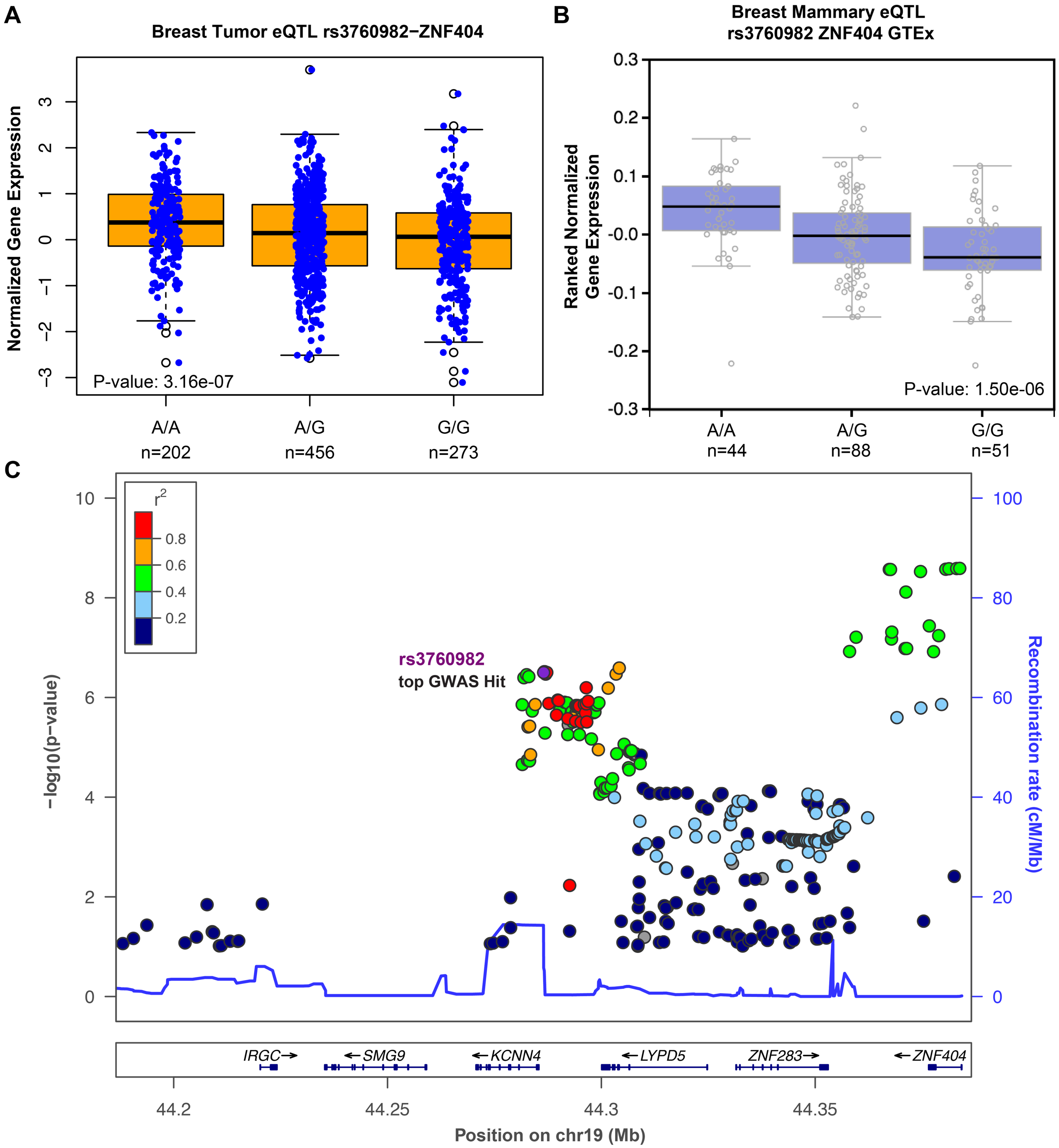

Fig 8. Candidate causal SNP rs3760982 is one of the top eQTLs for gene ZNF404. (A) The A/A genotype at rs3760982 is correlated with higher expression of ZNF404 in breast cancer patient solid tumor samples. (B) GTEx data confirm that the A/A genotype at rs3760982 is correlated with higher expression of ZNF404 in breast tissue. (C) rs3760982 is one of the top eQTL SNPs of ZNF404 and rs3760982 is the most associated GWAS hit for breast cancer susceptibility.

https://doi.org/10.1371/journal.pgen.1006761.g008

We found that rs3760982 variants are predicted to modulate RUNX binding (Fig 5) and we confirm previous work showing that rs3760982 is an eQTL for ZNF404 (Fig 8) [3, 63]. The A allele of rs3760982 conforms more stringently to the RUNX consensus sequence and this allele is predicted to enhance RUNX binding; allele-specific ChIP-seq in breast tissue, would test whether RUNX family TFs prefer binding the A allele *in vivo*. The RUNX family of TFs are canonical transcriptional activators [64–67], so we hypothesize that increased RUNX binding is a mechanism by which the ZNF404 is regulated. Testing this hypothesis would necessitate specific gene editing of rs3760982 and subsequent measurements of ZNF404 expression. While CRISPR-mediated [68] deletions of genetic elements is routine, precise changes of specific alleles remains a challenge. Deletion of the rs3760982 variant by CRISPR, followed by measuring ZNF404 expression would confirm or refute the role of rs3760982 variants in ZNF404 expression. One could also test allele-specific expression of alleles within transcription units that are phased with rs3760982 variants. We propose using precision global run-on sequencing [69] to measure nascent RNA expression to capture informative intronic SNPs. While genomic approaches are a means to develop novel hypotheses, the advent of genetic editing approaches permits hypothesis testing to define mechanisms by which genes and genetic variants contribute to disease risk.

We found that rs11540855 is an eQTL for ANKLE1 (Fig 7) and rs11540855 variants are predicted to affect CTCF binding (Fig 5). The rs11540855 SNP is in high LD ($r^2 = 0.90$) with the breast cancer-associated GWAS SNP rs8170, which was first found as a modifier of breast cancer risk in BRCA1 mutation carriers [58]. Subsequently, this SNP was found to be associated with breast cancer susceptibility in ER-negative breast cancer [58–60]. ANKLE1 is an evolutionarily conserved non-membrane-bound LEM protein that harbors endonuclease activity, but its cellular functions remain uncharacterized [70, 71]. Future work will need to determine the allele-specificity of CTCF binding at rs11540855 and test the role that this site has upon ANKLE1 expression. These approaches will be able to define the relationship between TF binding and gene expression, but it is challenging to develop a physiologically relevant model of breast cancer risk that incorporates human genetic variation. GWAS-identified SNPs are common and typically confer relatively small differences in risk. Further, the cumulative affects of differential gene expression over the lifetime of an individual cannot be easily recapitulated in a controlled environment.

Our research identified a previously uncharacterized DNA sequence motif that is enriched in open chromatin, evolutionarily conserved, and is associated with directional hypersensitivity profiles (Figs 2 and 3). We hypothesize that this orphan motif is the recognition site for a previously uncharacterized transcription factor. Future work, such as DNA affinity chromatography [72], will be needed to identify this candidate TF.

Genomic approaches are ideally suited to address fundamental biological questions in a relatively unbiased manner. Integrative genomic approaches and analyses can clarify the null-hypothesis and permit the development of novel hypotheses that were previously inconceivable. These approaches serve as a first-step in understanding the biology of breast cancer risk and targeted experimental follow-up is necessary to define the mechanistic roles of genes and genetic variants in breast cancer susceptibility and disease progression.

Materials and methods

ATAC-seq library preparation

We cultured MCF10A cells in Dulbecco's modified Eagle's medium (Invitrogen) with 5% horse serum (Invitrogen), 1% penicillin/streptomycin (Invitrogen), 20 ng/ml EGF (Peprotech), 0.5 μg/ml hydrocortisone (Sigma), 100 ng/ml cholera toxin (Sigma) and 10 μg/ml insulin

(Sigma) in a humidified incubator at 37˚C with 5% CO_2. The ATAC-seq library was prepared as previously described [73] with several modifications: 1) IGEPAL CA-630 was omitted from the lysis buffer; 2) we performed two additional wash steps with lysis buffer; and 3) we performed PCR-clean up using AMPure XP beads to select DNA <600 bp. The MCF10A ATAC-seq data were deposited in the Gene Expression Omnibus (GEO) database, with accession number GSE89013. We mapped reads to the hg38 human reference genome using Bowtie2 [74] and merged replicate aligned files. We used the merged data for all subsequent analysis; refer to S1 File for ATAC-seq data analysis details.

Iterative *de novo* motif analysis from regulatory regions

We performed iterative rounds of *de novo* motif analysis using a 120-base pair window centered on the summit of hypersensitivity as defined by ATAC-seq or DNase-seq (S1 File). In each cell type we found hundreds of over-represented position specific weight matrices (PSWMs). We identified all instances of each PSWM within breast-specific regulatory elements, while accounting for the possibility that the reference genome contains variants that will conform more or less strictly to the PSWM. This approach allowed us to identify potential binding sites that contain SNPs, even if the reference allele does not match the queried PSWM.

Clustering of *de novo* found motif PSWMs

Although we identified hundreds of distinct PSWMs, many PSWMs are similar to one another and are likely to represent redundant specificity of a single TF or TF family. To consolidate similar PSWMs into known TF families, we systematically classified several public PSWM repositories [43–47] into families with distinct features. PSWMs were first divided into clusters based on connectivity; connectivity between motif nodes was measured by negative $\log_{10}$ E-value as calculated by TOMTOM [75]. An edge was inferred between two motif nodes if their similarity exceeded a negative $\log_{10}$ E-value of 10. We defined a motif cluster as a connected set of nodes; connectivity is defined by the existence of a path between every pair of nodes. A fast greedy modularity algorithm [76] further divided each motif cluster into families.

Identification of SNPs that affect binding of TFs

We downloaded a curated list of breast cancer associated SNPs from the GWAS catalog [77]. The SNP that exhibits the most statistically significant association with breast cancer in any locus may not be causal due to linkage disequilibrium (LD) and the sampling variation that interrogated the specific SNP. To better define the list of likely causal variants for each locus, we identified all SNPs satisfying the following three criteria: 1) SNPs that are in strong LD ($r^2 \geq 0.8$) with the most significant reported GWAS SNP; 2) SNPs that are located within putative TF binding sites identified by hypersensitivity assays; and 3) SNPs that are within critically important positions that affect TF binding affinity (Information Content $\geq$ 0.5).

TF allele-specific binding preference analysis

We analyzed ENCODE CTCF ChIP-seq data for allele-specific preference of SNPs that are predicted to modulate CTCF binding affinity. All CTCF ChIP-seq data are provided within S1 File. We analyzed the highest intensity CTCF sites to ensure that sequencing reads would span the query SNP. We exclusively queried normal karyotype cell lines that had SNPs that were heterozygous within each locus to reduce the chances that copy number variations (i.e., aneuploidy) of alleles would skew our analyses.

eQTL analysis

To identify putative causal genes whose expression may be affected by polymorphisms, we performed eQTL analysis using TCGA breast cancer data [57] with fastQTL [78]. We imputed the genotypes from TCGA SNP6 arrays that were hybridized with DNA extracted from the blood of patients with breast cancer. We retrieved genotype data from dbGaP (phs000178.v9.p8) and imputed genotypes using the Michigan Imputation Server [79] with the following parameters: 1000G Phase 1 v3 Shapeit2 Reference Panel, ShapeIT Phasing, Mixed Population, and Quality Control/Imputation Mode. Following imputation, we removed SNPs with the following features: Hardy-Weinberg Equilibrium $p < 1 \times 10^{-6}$ and minor allele frequency (MAF) < 5%. We used UCSC-curated TCGA RNA-seq data [57] from breast cancer patient solid tumor samples as the gene expression data to identify eQTLs. TCGA clinical data were incorporated as the covariates for eQTL analysis such as sample RNA concentration, RIN value, sex, and ethnicity. We performed Principal Component Analysis (PCA) on the quantitative variables from clinical data and used the first three principal components as covariates. We retained other qualitative variables as categorical covariates.

Supporting information

S1 File. A step-by-step guide to reproducing this publication's analyses and results. In this vignette, we provide data sources for DNase-seq data, ChIP-seq data, TCGA RNA-seq data, TCGA genotype data, TCGA phenotype data, breast cancer GWAS catalog information, and PSWM databases used in our study. We provide detailed documentation of the computational analyses we performed.
(PDF)

S2 File. A bundle of all the individual executable scripts we used in this publication. This zipped file contains the full set of R, Python, and Shell scripts used in this publication.
(ZIP)

S1 Fig. All the motifs found by iterative *de novo* motif analysis in the five data sets mapped to 37 distinct motif families. (A) We clustered several publicly available PSWMs into families based on their similarities. In this illustration each node is a PSWM and the size of each node is proportional its number of edges. Contiguously connected nodes are colored uniquely and each family of PSWM is outlined by a distinct background color. We inferred edges using TOMTOM [75] to match PSWMs. The width of edges denote the similarity between two PSWMs and this width is proportional to the $-\log_{10}$ E-value of the match. (B) TFs are organized into families that recognize similar sequences and a representative motif for each regulatory sequence family is presented as a seqLogo. For example, three known TFs (ESR1, ESR2, and PPARG-γ) recognize the estrogen response element (ERE).
(TIF)

S2 Fig. TF members in the same family have different levels of gene expression in the solid tumor sample of breast cancer patients. Each TF is numbered in accordance with its row number in Table 1. Normalized RNA-seq gene expression data are from TCGA breast cancer patient solid tumor samples [57].
(TIF)

S3 Fig. TF expression correlates with patient outcome in a subtype-specific manner. High expression of IRF7, BATF, and TP73 is correlated with better breast cancer patient outcome in

Luminal A, HER2+, and Basal-like subtype, respectively. TF expression groups are classified as either high or low expression and with the corresponding patient number is noted in the parentheses. We use FDR to correct the P-values for multiple testing. Kaplan-Meier analysis is performed using TCGA breast cancer patient solid tumor sample RNA-seq expression and overall breast cancer patient survival data [57].
(TIF)

S4 Fig. Candidate SNPs are predicted to modulate the binding of breast cancer-relevant TFs. Each SNP is in strong LD ($r^2 \geq 0.8$) with the most associated breast cancer GWAS SNP and affects TF binding affinity. We report the minor allele frequency (MAF) and LD association (r^2) of each SNP with the most associated breast cancer GWAS SNP.
(TIF)

S5 Fig. Genotypic variants of rs8103622 affect CTCF binding in many cell types. ENCODE ChIP-seq data (top) indicates that CTCF is strongly bound in HMEC at rs8103622. The risk allele C of rs8103622 is predicted to increase CTCF binding, which was identified in round 33 of our *de novo* motif analysis from MCF7 (middle). By analyzing ENCODE ChIP-seq count data, we show that there is an allelic imbalance favoring C (denoted in red) versus T (denoted in black) in 34 out of 37 cell types/replicates that are diploid and heterozygous for the C/T allele at rs8103622.
(TIF)

Acknowledgments

We thank all members of the Guertin lab for discussion and comments.

Author Contributions

Conceptualization: YL MJG.

Data curation: YL MJG.

Formal analysis: YL MJG.

Funding acquisition: YL MJG.

Investigation: YL NMW MJG.

Methodology: YL MJG.

Project administration: YL MGD SSR MC MJG.

Resources: MJG.

Software: YL MJG.

Supervision: MJG.

Validation: YL.

Visualization: YL MJG.

Writing – original draft: YL MJG.

Writing – review & editing: YL NMW MGD SSR MC MJG.

References

1. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nature genetics. 2013; 45(4):392–398. https://doi.org/10.1038/ng.2561 PMID: 23535733
2. Han MR, Long J, Choi JY, Low SK, Kweon SS, Zheng Y, et al. Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. Human Molecular Genetics. 2016;p. ddw164.
3. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013; 45(4):353–361. https://doi.org/10.1038/ng.2563
4. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nature genetics. 2015; 47(4):373–380. https://doi.org/10.1038/ng.3242
5. Song C, Chen GK, Millikan RC, Ambrosone CB, John EM, Bernstein L, et al. A genome-wide scan for breast cancer risk haplotypes among African American women. PloS one. 2013; 8(2):e57298. https://doi.org/10.1371/journal.pone.0057298 PMID: 23468962
6. Cowper-Sal R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nature genetics. 2012; 44(11):1191–1198. https://doi.org/10.1038/ng.2416
7. Bailey SD, Desai K, Kron KJ, Mazrooei P, Sinnott-Armstrong NA, Treloar AE, et al. Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. Nature Genetics. 2016; 48(10):1260–1266. https://doi.org/10.1038/ng.3650 PMID: 27571262
8. Donnelly P. Progress and challenges in genome-wide association studies in humans. Nature. 2008; 456(7223):728–731. https://doi.org/10.1038/nature07631 PMID: 19079049
9. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nature genetics. 2011; 43(6):513–518. https://doi.org/10.1038/ng.840 PMID: 21614091
10. Paul DS, Soranzo N, Beck S. Functional interpretation of non-coding sequence variation: Concepts and challenges. Bioessays. 2014; 36(2):191–199. https://doi.org/10.1002/bies.201300126 PMID: 24311363
11. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics. 2012; 90(1):7–24. https://doi.org/10.1016/j.ajhg.2011.11.029 PMID: 22243964
12. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337(6099):1190–1195. https://doi.org/10.1126/science.1222794 PMID: 22955828
13. Castellanos-Rubio A, Fernandez-Jimenez N, Kratchmarov R, Luo X, Bhagat G, Green PH, et al. A long noncoding RNA associated with susceptibility to celiac disease. Science. 2016; 352(6281):91–95. https://doi.org/10.1126/science.aad0467 PMID: 27034373
14. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016; 352(6285):600–604. https://doi.org/10.1126/science.aad9417 PMID: 27126046
15. Lazarovici A, Zhou T, Shafer A, Machado ACD, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proceedings of the National Academy of Sciences. 2013; 110(16):6376–6381. https://doi.org/10.1073/pnas.1216822110
16. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466(7307):714–719. https://doi.org/10.1038/nature09266
17. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013; 152(6):1237–1251. https://doi.org/10.1016/j.cell.2013.02.014 PMID: 23498934
18. Bhagwat AS, Vakoc CR. Targeting Transcription Factors in Cancer. Trends in cancer. 2015; 1(1):53–65. https://doi.org/10.1016/j.trecan.2015.07.001 PMID: 26645049
19. Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012; 151(1):56–67. https://doi.org/10.1016/j.cell.2012.08.026 PMID: 23021215
20. Pike MC, Spicer DV, Dahmoush L, Press MF. Estrogens progestogens normal breast cell proliferation and breast cancer risk. Epidemiologic reviews. 1993; 15(1):17–35. https://doi.org/10.1093/oxfordjournals.epirev.a036102 PMID: 8405201

21. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132(2):311–322. https://doi.org/10.1016/j.cell.2007.12.014 PMID: 18243105

22. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature methods. 2009; 6(4):283–289. https://doi.org/10.1038/nmeth.1313 PMID: 19305407

23. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods. 2013; 10(12):1213–1218. https://doi.org/10.1038/nmeth.2688 PMID: 24097267

24. Grøntved L, Bandle R, John S, Baek S, Chung HJ, Liu Y, et al. Rapid genome-scale mapping of chromatin accessibility in tissue. Epigenetics & chromatin. 2012; 5(1):1.

25. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome research. 2011; 21 (10):1757–1767. https://doi.org/10.1101/gr.121541.111 PMID: 21750106

26. Guertin MJ, Lis JT. Chromatin landscape dictates HSF binding to target DNA elements. PLoS Genet. 2010; 6(9):e1001114. https://doi.org/10.1371/journal.pgen.1001114 PMID: 20844575

27. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, et al. Chromatin accessibility predetermines glucocorticoid receptor binding patterns. Nature genetics. 2011; 43(3):264–268. https://doi.org/10.1038/ng.759 PMID: 21258342

28. Guertin MJ, Martins AL, Siepel A, Lis JT. Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet. 2012; 8(3):e1002610. https://doi.org/10.1371/journal.pgen.1002610 PMID: 22479205

29. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature biotechnology. 2014; 32(2):171–178. https://doi.org/10.1038/nbt.2798 PMID: 24441470

30. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome research. 2011; 21 (3):447–455. https://doi.org/10.1101/gr.112623.110 PMID: 21106904

31. Sung MH, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Molecular cell. 2014; 56(2):275–285. https://doi.org/10.1016/j.molcel.2014.08.016 PMID: 25242143

32. He HH, Meyer CA, Chen MW, Zang C, Liu Y, Rao PK, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nature methods. 2014; 11(1):73–78. https://doi.org/10.1038/nmeth.2762

33. Martins AL, Walavalkar NM, Anderson WD, Zang C, Guertin MJ. Universal correction of enzymatic sequence bias. bioRxiv. 2017;p. 104364.

34. Wang Z, Martins AL, Danko CG. RTFBSDB: an integrated framework for transcription factor binding site analysis. Bioinformatics. 2016;p. btw338.

35. Brænne I, Civelek M, Vilne B, Di Narzo A, Johnson AD, Zhao Y, et al. Prediction of causal candidate genes in coronary artery disease loci. Arteriosclerosis, thrombosis, and vascular biology. 2015; 35 (10):2207–2217. https://doi.org/10.1161/ATVBAHA.115.306108 PMID: 26293461

36. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics. 2015; 16(4):197–212. https://doi.org/10.1038/nrg3891 PMID: 25707927

37. Cavalli M, Pan G, Nord H, Arzt EW, Wallerman O, Wadelius C. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. Genomics. 2016; 107(6):248–254. https://doi.org/10.1016/j.ygeno.2016.04.006

38. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, et al. Variations in DNA elucidate molecular networks that cause disease. Nature. 2008; 452(7186):429–435. https://doi.org/10.1038/nature06757 PMID: 18344982

39. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6(4):e1000888. https://doi.org/10.1371/journal.pgen.1000888 PMID: 20369019

40. Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501(7468):506–511. https://doi.org/10.1038/nature12531 PMID: 24037378

41. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. https://doi.org/10.1038/nature11247

42. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nature biotechnology. 2010; 28(10):1045–1048. https://doi.org/10.1038/nbt1010-1045 PMID: 20944595

43. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research. 2004; 32(suppl 1):D91–D94. https://doi.org/10.1093/nar/gkh012 PMID: 14681366

44. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009;p. gkp335.

45. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell. 2010; 38(4):576–589. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

46. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome research. 2010; 20 (6):861–873. https://doi.org/10.1101/gr.100552.109 PMID: 20378718

47. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013; 152(1):327–339. https://doi.org/10.1016/j.cell.2012.12.009 PMID: 23332764

48. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research. 2005; 15(8):1034–1050. https://doi.org/10.1101/gr.3715005 PMID: 16024819

49. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research. 2010; 20(1):110–121. https://doi.org/10.1101/gr.097857.109 PMID: 19858363

50. Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell. 1979; 16(4):797–806. https://doi.org/10.1016/0092-8674(79)90095-3 PMID: 455449

51. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012; 489(7414):83–90. https://doi.org/10.1038/nature11212 PMID: 22955618

52. Mathelier A, Fornes O, Arenillas DJ, Chen Cy, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic acids research. 2015;p. gkv1176.

53. Berger MF, Bulyk ML. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. Gene Mapping, Discovery, and Expression: Methods and Protocols. 2006;p. 245–260. https://doi.org/10.1385/1-59745-097-9:245

54. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein —DNA interactions. Nucleic acids research. 2009; 37(suppl 1):D77–D82. https://doi.org/10.1093/nar/gkn660 PMID: 18842628

55. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, et al. Genome-wide analysis of estrogen receptor binding sites. Nature genetics. 2006; 38(11):1289–1297. https://doi.org/10.1038/ng1901

56. Bidwell BN, Slaney CY, Withana NP, Forster S, Cao Y, Loi S, et al. Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape. Nature medicine. 2012; 18(8):1224–1231. https://doi.org/10.1038/nm.2830 PMID: 22820642

57. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, et al. The UCSC cancer genomics browser: update 2015. Nucleic acids research. 2014;p. gku1073.

58. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nature genetics. 2010; 42(10):885–892. https://doi.org/10.1038/ng.669

59. Stevens KN, Vachon CM, Lee AM, Slager S, Lesnick T, Olswold C, et al. Common breast cancer susceptibility loci are associated with triple-negative breast cancer. Cancer research. 2011; 71(19):6240–6249. https://doi.org/10.1158/0008-5472.CAN-11-1266 PMID: 21844186

60. Stevens KN, Fredericksen Z, Vachon CM, Wang X, Margolin S, Lindblom A, et al. 19p13.1 Is a Triple-Negative—Specific Breast Cancer Susceptibility Locus. Cancer research. 2012; 72(7):1795–1803. https://doi.org/10.1158/0008-5472.CAN-11-3364

61. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Human molecular genetics. 2015; 24(R1):R102–R110. https://doi.org/10.1093/hmg/ddv259

62. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. Nature Reviews Genetics. 2016; 17(2):93–108. https://doi.org/10.1038/nrg.2015.17 PMID: 26781813

63. Haiman CA, Han Y, Feng Y, Xia L, Hsu C, Sheng X, et al. Genome-wide testing of putative functional exonic variants in relationship with breast and prostate cancer risk in a multiethnic population. PLoS Genet. 2013; 9(3):e1003419. https://doi.org/10.1371/journal.pgen.1003419 PMID: 23555315

64. Ogawa E, Inuzuka M, Maruyama M, Satake M, Naito-Fujimoto M, Ito Y, et al. Molecular cloning and characterization of PEBP2*β*, the heterodimeric partner of a novel Drosophila runt-related DNA binding protein PEBP2*α*. Virology. 1993; 194(1):314–331. https://doi.org/10.1006/viro.1993.1262 PMID: 8386878

65. Bruhn L, Munnerlyn A, Grosschedl R. ALY, a context-dependent coactivator of LEF-1 and AML-1, is required for TCRalpha enhancer function. Genes & development. 1997; 11(5):640–653. https://doi.org/10.1101/gad.11.5.640

66. Zhang DE, Hetherington CJ, Meyers S, Rhoades KL, Larson CJ, Chen HM, et al. CCAAT enhancer-binding protein (C/EBP) and AML1 (CBF alpha2) synergistically activate the macrophage colony-stimulating factor receptor promoter. Molecular and Cellular Biology. 1996; 16(3):1231–1240. https://doi.org/10.1128/MCB.16.3.1231 PMID: 8622667

67. Pelletier N, Champagne N, Stifani S, Yang XJ. MOZ and MORF histone acetyltransferases interact with the Runt-domain transcription factor Runx2. Oncogene. 2002; 21(17):2729–2740. https://doi.org/10.1038/sj.onc.1205367

68. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA—guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337(6096):816–821. https://doi.org/10.1126/science.1225829 PMID: 22745249

69. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013; 339(6122):950–953. https://doi.org/10.1126/science.1229386 PMID: 23430654

70. 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(5):e63925. https://doi.org/10.1371/journal.pone.0063925 PMID: 23717510

71. Brachner A, Braun J, Ghodgaonkar M, Castor D, Zlopasa L, Ehrlich V, et al. The endonuclease Ankle1 requires its LEM and GIY-YIG motifs for DNA cleavage in vivo. J Cell Sci. 2012; 125(4):1048–1057. https://doi.org/10.1242/jcs.098392 PMID: 22399800

72. Gadgil H, Jurado LA, Jarrett HW. DNA affinity chromatography of transcription factors. Analytical biochemistry. 2001; 290(2):147–178. https://doi.org/10.1006/abio.2000.4912 PMID: 11237318

73. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology. 2015;p. 21–29.

74. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4):357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

75. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome biology. 2007; 8(2):R24. https://doi.org/10.1186/gb-2007-8-2-r24 PMID: 17324271

76. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. Physical review E. 2004; 70(6):066111. https://doi.org/10.1103/PhysRevE.70.066111

77. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014; 42(D1):D1001–D1006. https://doi.org/10.1093/nar/gkt1229 PMID: 24316577

78. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2015;p. btv722.

79. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016; 48(10):1284–1287. https://doi.org/10.1038/ng.3656 PMID: 27571263