Published in final edited form as:

Nat Genet. 2012 November ; 44(11): 1191–1198. doi:10.1038/ng.2416.

Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression

Richard Cowper-Sal·lari[1,2,#], **Xiaoyang Zhang**[1,2,#], **Jason B. Wright**[1], **Swneke D. Bailey**[5,6], **Michael D. Cole**[1], **Jerome Eeckhoute**[3,4], **Jason H. Moore**[1,2,*], and **Mathieu Lupien**[5,6,*]

[1]Department of Genetics, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH, 03756

[2]Institute for Quantitative Biomedical Sciences, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH, 03756

[3]Université de Lille Nord de France, INSERM UMR1011, UDSL, F-59000, Lille, France

[4]Institut Pasteur de Lille, F-59019, Lille, France

[5]Ontario Cancer Institute, Princess Margaret Hospital-University Health Network, Toronto, ON, Canada

[6]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

Abstract

Genome-wide association studies (GWASs) have identified thousands of single nucleotide polymorphisms (SNPs) associated with human traits and diseases. But because the vast majority of these SNPs are located in the noncoding regions of the genome their risk promoting mechanisms are elusive. Employing a new methodology combining cistromics, epigenomics and genotype imputation we annotate the noncoding regions of the genome in breast cancer cells and systematically identify the functional nature of SNPs associated with breast cancer risk. Our results demonstrate that breast cancer risk-associated SNPs are enriched in the cistromes of FOXA1 and ESR1 and the epigenome of H3K4me1 in a cancer and cell-type-specific manner. Furthermore, the majority of these risk-associated SNPs modulate the affinity of chromatin for FOXA1 at distal regulatory elements, which results in allele-specific gene expression, exemplified by the effect of the rs4784227 SNP on the *TOX3* gene found within the 16q12.1 risk locus.

Disease risk-associated SNPs (raSNPs) map predominantly to noncoding regions[1]. Over 70% of the risk-associations in the National Human Genome Research Institute (NHGRI)

[*]Correspondence: M.L (mlupien@uhnres.utoronto.ca) or J.H.M (Jason.H.Moore@Dartmouth.edu)..
[#]R.C-S. and X.Z. contributed equally to this work.

Accession numbers ChIP-seq files generated for this manuscript are available on GEO: accession number: GSE31151

Author Contributions All experiments were conceived by R.C-S., X.Z., J.E., M.D.C., J.H.M. and M.L. Experiments were conducted by X.Z. and J.B.W. Computational methods were conceived and implemented by R.C-S with help from J.H.M. and M.L. Computational analysis was conducted by R.C-S. and S.D.B. The manuscript was written by R.C-S., X.Z. and M.L. with help from J.E., M.D.C. and J.H.M. Figures were designed and prepared by R.C-S and X.Z.

GWAS catalog lack variants mapping to exons within their haplotype block, absolving them from coding-sequence disruption[2]. A `cistrome' has been defined as the complete set of binding sites for a transcription factor assessed genome-wide for a given cell type under a specific treatment. Similarly, an `epigenome' is the complete set of elements carrying the same epigenomic mark for a specific cell type and treatment. Cistromes and epigenomes obtained through ChIP-seq identify the positions of regulatory elements and novel transcripts in both coding and noncoding regions across the whole genome[3–8]. Importantly, this is revealing that cistromes and epigenomes are cell-type-specific[4,5,9,10]. For instance, the cistromes for the pioneer factor FOXA1 (also known as HNF3A) and the transcription factor ESR1 (estrogen receptor alpha; also known as ERα) have unique distributions in breast cancer (BCa) cells compared to other cell types[4,11–13]. Similarly, the epigenomes of specific histone modifications found in regulatory elements, such as the mono or dimethylation of lysine 4 on histone H3 (H3K4me1 or me2), in diverse cell types demonstrate that functional regulatory elements are also lineage-specific[4,5]. The interplay between cistromes and epigenomes guides the transcriptional activity of these two critical factors in BCa development to specific regions driving the proper gene expression profile[11,12,14]. Importantly, differential recruitment of FOXA1 at enhancers drives cell-type-specific transcriptional programs[4]. Because cistromes and epigenomes lie at the source of cell identity we have investigated the functional relation between BCa raSNPs and BCa cistromes and epigenomes using a novel integrative functional genomics approach.

Risk SNP enrichment

We first gathered the coordinates for the raSNPs for BCa from the GWAS catalog available through the NHGRI[15]. Of these, only one in forty-four mapped to coding exons, twenty-five to introns, one to a FOXA1 site, and three to ESR1 sites (Fig. 1a). However, the number of SNPs assessed in GWAS genotyping microarrays provides low genomic coverage; less than 10% of all annotated SNPs (dbSNP build 135) are covered by any given platform[16]. Statistically, it is therefore more likely for raSNPs to be in linkage disequilibrium (LD) with causal variants than to be causal themselves[17]. Therefore, we identified for each BCa raSNP the list of all SNPs in strong LD (ldSNPs) using the HapMap project data (Supplementary Table 1). We used a highly stringent LD threshold of LOD > 2 and D' > 0.99. Through this imputation approach we extended the coverage of raSNPs to a more comprehensive list of putative functional ldSNPs, defining raSNP/ldSNP clusters (Fig. 1b). We refer to the comprehensive collection of all raSNP/ldSNP clusters associated with a specific disease or trait, in this case BCa, as its associated variant set (AVS). The BCa AVSs is then composed of forty-four raSNPs and 1315 ldSNPs (Supplementary Table 1). The genome wide distribution of raSNP/ldSNP clusters is reminiscent of the recruitment profiles of many transcription factors such as FOXA1 and ESR1 as they predominantly map to intronic and intergenic regions[4,11,12,14]. By considering ldSNPs, the number of raSNP/ldSNP clusters mapping to coding exons increased from one to five and the number of clusters mapping to FOXA1 and ESR1 binding sites increased from one to sixteen and three to nineteen, respectively (Fig. 1c). Significance of these mapping tallies was calculated through a novel computational method; Variant Set Enrichment (VSE). VSE is a method based on permutation testing that generates null distributions based on randomized variant sets while

taking into account the heterogeneous LD structure of the BCa AVS (Fig. 1d–g; a detailed description is provided in the methods section). It is an extension of the work of Hindorff and colleagues[18]. VSE reveals that the BCa AVS is not enriched in genes (Fig. 1d–e) but strongly enriched for the FOXA1 and ESR1 cistromes (Fig. 1f–g).

We extended the VSE analysis to include a total of 72 cistromes from sixteen different transcription factors and 27 epigenomes from eight different histone modifications in BCa cells. These were assessed in BCa cell lines MCF7, T47D and ZR75. The primary treatment of the samples is estradiol (E2), which is intended to stimulate ESR1. Cistromes and epigenomes were either generated by our lab or obtained from the literature and the Nuclear Receptor Cistrome project. Only significant peaks for cistromes and epigenomes were used in the analysis ($P<10e^{-5}$)[4,19]. Only enrichment scores satisfying a stringent Bonferroni corrected threshold for significance are reported subsequently (103 tests; $P<10^{-3.3}$).

Among the 24 distinct BCa epigenomes and exon and intron annotations assessed, only H3K4me1 (untreated and treated) shows enrichment (Fig 2a–c and Supplementary Tables 2 and 3), a histone modification associated with regulatory elements, mainly enhancers. Among the 72 distinct BCa cistromes, only the FOXA1 (three samples, untreated and treated) and ESR1 (one sample, treated) transcription factors show enrichment (Fig. 2d,e and Supplementary Tables 4 and 5). H3K4me1 enrichment is independent of FOXA1 and ESR1. After subtracting FOXA1 and ESR1 binding sites from the H3K4me1 elements, the epigenome under estrogen treatment remains significant ($P=10^{-3.532}$, Supplementary Table 6). FOXA1, ESR1 and H3K4me1 remain significant when using r^2 as a measure of LD instead of D' and LOD for thresholds of $r^2 = 0.8$ or greater (Supplementary Tables 7, 8 and 9). These six enriched cistromes and epigenomes in BCa cells span 70% (31/44) of the BCa raSNP/ldSNP clusters (Fig. 2a–e, dark heatmap rows). Five out of these six enrichment events are also significantly enriched with the BCa AVS when a null distribution based on DNase accessibility is used in a subsequent VSE run (Supplementary Tables 10 and 11).

VSE analysis of the BCa AVS across all cistromes and epigenomes in BCa cells demonstrates that the enrichment for FOXA1 and ESR1 cistromes and H3K4me1 epigenome is highly factor-specific (Fig. 2a–e). However, because not all FOXA1 or ESR1 cistromes show significant VSE scores, enrichment for the other transcription factors assayed in the study cannot be completely discarded. The Nuclear Receptor Cistrome project data sets are highly heterogeneous; the number of binding sites for FOXA1 in MCF7 cells varies over an order of magnitude depending on the laboratory of origin. Given that we are intersecting 44 raSNPs with tens of thousands of binding sites, a ten-fold decrease in the number of binding sites dramatically reduces our power to detect enrichment.

Enrichment of the BCa AVS for FOXA1, ESR1 and H3K4me1 is also cancer and cell-type-specific. We used the AVS of prostate cancer (PCa), bone mineral density (BMD) and colorectal cancer (CRC) to control for the enrichments of FOXA1, ESR1 and H3K4me1, respectively (Supplementary Tables 12, 13 and 14). Prostate cancer development relies on FOXA1, bone mineral density is affected by estrogen signaling and the colorectal cancer AVS is enriched for the H3K4me1 epigenome in tumor samples[20]. None of these AVS were enriched in their target cistromes or epigenomes in breast cancer cells, showing that the

enrichment of the BCa AVS for these cistromes and epigenomes is cancer-type-specific (Fig. 2g–i, and Supplementary Tables 15, 16 and 17).

Conversely, the breast cancer AVS was tested for enrichment in prostate, bone and colon cell lines (LNCaP, VCaP, human metastatic cancer tissue, U2OS and Caco-2). Cistromes and epigenomes assessed in these cell lines for factors and modifications including FOXA1, ESR2, AR and H3K4me1 did not show enrichment for the BCa AVS, demonstrating that the enrichment of the BCa AVS is cell-type-specific (Fig. 2f and Supplementary Table 18).

FOXA1 modulation

Having obtained these results, we then tested the ability of the BCa raSNP/ldSNP clusters to disrupt the normal recruitment of FOXA1 to chromatin. We refer to the mechanism through which base pair changes alter the recognition of a binding motif as affinity modulation. Using a novel computational method, named Intra-Genomic Replicates (IGR), we are able to accurately predict the affinity modulation of a SNP. The approach delineates the affinity of a given transcription factor for both the reference and variant alleles of a SNP while accounting for its contextual sequence. The affinity of a transcription factor for a particular DNA sequence of length K (K-mer) can be obtained by averaging binding data across a genome-wide ChIP-seq dataset for that transcription factor. We refer to this measurement as the affinity model for that K-mer. If a canonical binding motif is altered by a variant allele, the binding factor can find a higher affinity configuration by moving a few bases or reversing its orientation. IGR accounts for displacement effects by computing affinity models over a sliding window of K-mers around the SNP of interest (Fig. 3a). Through this process, the collection of affinity models for increasing values of K are placed in a lattice structure that connects K-mers that are one base pair apart. Two lattices are constructed, one for each of the variants alleles (Fig. 3b). The affinity models within each lattice are then filtered based on their signal-to-noise ratios (a detailed description is provided in the methods section). The maxima among the remaining affinity models in the lattices are used to calculate the IGR score and p-value (Fig. 3c). This comparison between maxima represents the highest affinity for each allele given the genomic context of the SNP. Affinity model scores show little correlation with Position Weighted Matrix (PWM) scores for the same set of 256 K-mers (Fig. 3d). The models derived for a given transcription factor are consistent across laboratories and cell lines (Fig. 3e–f). Because the chromatin landscape is a known barrier to the interaction between transcription factors and DNA we restrict our K-mer search to chromatin regions favorable to binding[21]. When K-mer instances are not filtered based on a favorable chromatin landscape, the affinity of the highly ranked K-mers is underestimated (Fig. 3g).

Recruitment of FOXA1 is dependent on the recognition of the forkhead (FKH) motif (Fig. 4a). We explored the affinity landscape of the FKH motif by generating a collection of thirty 10-mers; one for each possible variation at each of the ten positions in the FKH motif (Fig. 4b–d). PWMs are useful in the identification of transcription factor binding sites, but their scores correlate poorly with binding affinity (Fig. 4b, d–e). The IGR method can accurately predict the affinity modulating properties of the thirty 10-mers (Fig. 4c–e). Furthermore, the physical association between a transcription factor and DNA cannot be fully captured by

PWMs that consider each base of the motif in isolation. Changes at one side of the motif can aggravate or compensate for changes at the other side. We refer to this phenomenon as interaction effects between the bases of a motif. These effects can be dramatic, to the extent of reversing the affinity modulation properties of an allele if the contextual sequence is ignored (Supplementary Fig. 1).

SNP rs4784227 is associated with BCa and is predicted to disrupt the binding of an undetermined transcription factor[22]. It is also a raSNP that directly maps to a FOXA1 binding site (Fig. 1a). This raSNP falls on the eighth position of the FKH motif recognized by FOXA1 (Fig. 4a). The context sequence of this SNP differs from the canonical FKH motif; it has a G in the sixth position as opposed to an A, and ends in GA instead of TT. The PWM for the FKH motif predicts a 9% affinity increase for the variant T risk allele at rs4784227 compared to the reference C allele (6.24 T allele v. 5.734 C allele, Fig. 4b). Taking into account the interaction effects between motif positions the IGR method predicts a 63% increase in affinity for the variant T allele compared to the reference C allele (fold change 1.63, $P<10e^{-10}$, Fig. 4f and Supplementary Fig. 1). Previous reports have shown the allele-specific TCF4 recruitment at the risk-associated rs6983267 SNP in colon cancer cells[23–25]. IGR validates the affinity modulation of rs6983267 for TCF4 (fold change=1.59, $P<10^{-10}$, Fig. 4f). The comparison between the IGR scores of the BCa raSNP rs4784227 assessed for TCF4 and the colon cancer raSNP rs6983267 assessed for FOXA1 shows that the affinity predictions are factor-specific (fold changes ~1; P>0.05, Fig. 4f).

Having probed a single raSNP, we then applied the IGR method to all raSNP/ldSNP clusters and found that the BCa AVS is enriched in affinity modulating SNPs for FOXA1 (Supplementary Table 19). Of the 44 raSNP/ldSNP clusters in the BCa AVS, 33 contain SNPs mapping to the FOXA1 cistrome or H3K4me2 epigenome in BCa cells. These clusters harbor the regions where FOXA1 binding sites can be created or destroyed; these were the only clusters considered in this analysis[4]. For each cluster, we computed IGR scores for all SNPs in FOXA1 or H3K4me2 regions and tallied the number of affinity modulating SNPs in each cluster. We then used the VSE method to generate a null distribution for the number of affinity modulating variants in randomized sets of clusters. The significance threshold for affinity modulation is based on a Bonferroni correction that takes into account all tests across all permutations ($P<10e^{-6}$). 24 out of the 33 clusters mapping to FOXA1 or H3K4me2 regions (73%) contained affinity-modulating SNPs (P = 0.009, Fig. 5a). This means that more than half of all SNPs associated with breast cancer modulate the affinity for FOXA1. Genotyping the SNPs predicted to be disruptive of FOXA1 binding in four breast cancer cell lines (MCF7, BT474, T47D and ZR75-1) identified 13 SNPs (including the rs4784227 variant) heterozygous in at least one cell line. 77% (10/13) of these predicted SNPs demonstrate a significant and concordant allele-specific binding preference for FOXA1 determined by *in vivo* allele-specific ChIP assays (Fig. 5b). This comprehensive approach validates the power of IGR to identify affinity-modulating SNPs.

TOX3 disruption

To demonstrate *in vivo* that disruptive SNPs identified through the VSE and IGR methods can promote BCa development we focused on the BCa raSNP rs4784227 heterozygous in

MCF7 BCa cells. Heterozygosity allows us to determine the allele-specific function of a variant within its natural genomic context. The SNP rs4784227 directly overlaps a FKH motif within a FOXA1 binding site that lies in a regulatory element (marked by H3K4me2) 18 kb away from the *TOX3* gene (also known as *TNRC9* and *CAGF9*; Fig. 6a,b). The IGR method predicts that the variant T allele will favor FOXA1 binding over the reference C allele (Fig. 4f and Supplementary Fig. 1). Accordingly, allele-specific directed ChIP assays *in vivo* confirm that FOXA1 is preferentially recruited to the T risk allele (Fig. 6b). This occurs independently of changes to H3K4me2, the epigenetic signature favorable to FOXA1 binding, as H3K4me2 levels are equivalent between the T and C alleles at rs4784227 in MCF7 BCa cells[4] (Fig. 6a). While FOXA1 commonly promotes gene expression, co-binding to chromatin with Groucho/TLE proteins leads to local chromatin condensation and transcriptional repression[26]. The risk variant T allele of rs4784227 associated with increased FOXA1 binding is also significantly bound by Groucho/TLE compared to the C allele (Fig. 6c). Furthermore, H3K9Ac (a chromatin signature for active enhancers) is significantly less enriched at the T allele versus the C allele[7] (Fig. 6d).

Chromatin conformation capture assays (3C) using BglII or MspI restriction enzyme revealed that the rs4784227 region physically interacts with the promoter of the *TOX3* gene (Fig. 6e,f). Consequently, *TOX3* is a gene target of the regulatory region harboring rs4784227. In order to assess impact of the rs4784227 alleles on *TOX3* gene expression, we found a heterozygous SNP (rs2193094) in the first intron of *TOX3* in MCF7 cells. Sequencing the 3C product revealed that the reference C allele and the variant T allele of rs4784227 are physically linked to the T and G alleles of rs2193094, respectively (Fig. 6e,f and Supplementary Fig. 2). Allele-specific expression assays revealed that the T allele of rs2193094 is preferentially expressed compared to the G allele, suggesting a repression effect of the variant T allele at rs4784227 (Fig. 6g). This agrees with results from previous *in vitro* luciferase reporter assays[22].

RNA-Seq data from MCF7 cells revealed six genes (*TOX3, CHD9, RBL2, AKTIP, RPGRIP1L and FTO*) to be expressed in a ~3.5 Mb genomic window centered on the rs4784227 SNP (Supplementary Fig. 3). In addition to the rs2193094 SNP within the first intron of *TOX3*, we identified heterozygous SNPs within the intronic regions of *CHD9* and *FTO* (rs35925303 and rs11646260, respectively) by genotyping MCF7 cells. While the variant G allele of rs2193094 is associated with a decrease in *TOX3* RNA levels compared to the reference T allele, no significant difference in RNA levels was detected between the variant and reference alleles for either the rs35925303 (*CHD9*) or rs11646260 (*FTO*) SNPs (Fig. 7a). Furthermore, we genotyped the rs4784227 SNP in eleven ESR1+ and FOXA1+ breast cancer cell lines and extracted the expression data for *TOX3* from the Neve breast cancer cell lines database[1]. Only one cell line, T47D, was homozygous for the T allele of rs4784227. Correspondingly, T47D had the lowest expression of *TOX3*. Comparing all heterozygous (N=3) cell lines with those homozygous for the C allele of rs4784227 (N=7) we observed a significant association between the T allele of rs4784227 and a decrease in the expression of *TOX3* (one-sided $p = 0.046$, Fig. 7b). This association agrees with a previous study that identified an association between the rs3803662 SNP and *TOX3* gene expression using 1,401 breast cancer patients of European ancestry[27]. The rs3803662 SNP is

in significant LD with the rs4784227 SNP in individuals of European ancestry (r2=0.864, D'=1, HapMap CEU) (Supplementary Table 20). In fact, the T allele of rs4784227 only segregates with the A allele of the rs3803662 which is also associated with a decrease in TOX3 gene expression supporting our conclusion that the functional T allele of rs4784227 is associated with a decrease in the expression of *TOX3*. In addition, we included the remaining five genes (*CHD9, RBL2, AKTIP, RPGRIP1L and FTO*) that are expressed within a ~3.5 Mb window centered on rs4784227 to the e-QTL assay (Fig. 7b and Supplementary Fig. 3). The rs4784227 SNP was not significantly associated with the expression of these genes. Overall, this suggests that the allele-specific expression is unique to *TOX3* at this locus and dependent on the rs4784227 SNP.

In order to assess the role of TOX3 in breast cancer cells, we performed a cell proliferation assay after silencing *TOX3*. We used the ZR75-1 breast cancer cell line, because of its high expression level of *TOX3*. Silencing of *TOX3* in ZR75-1 breast cancer cells significantly increased cell proliferation compared to the control treatment (siRNA against Luciferase), suggesting that TOX3 functions as a tumor suppressor in breast cancer cells (Fig.7c,d).

Discussion

The mechanisms underlying BCa raSNPs are unknown. As with most other complex traits these raSNPs map to the noncoding regions of the genome[2]. Here, we demonstrate that BCa raSNPs are enriched for FOXA1 and ESR1 transcription factor binding sites and the H3K4me1 histone modification. Furthermore, we reveal that this enrichment is factor-specific, cell-type-specific and cancer-type-specific. The body of evidence supporting regulatory mechanisms for GWAS raSNPs is steadily growing[28–31]. Heterozygous sites with differential allelic occupancy within 100 bp of transcription start sites have been shown to have a strong association with differential gene expression and to be enriched for GWAS SNPs[32]. Binding of the FOXA1 pioneer factor is central for chromatin opening and nucleosome positioning favorable to transcription factor recruitment[4,11,12,33,34]. In addition, FOXA1 is central to the establishment of transcriptional programs responding to estrogen stimulation in ESR1-positive BCa cells[4,11,12,33,35,36].

Among the elusive BCa raSNPs, rs4784227 is the most studied. The 16q12.1 locus, which harbors rs4784227 and the *TOX3* gene, has been associated with breast cancer through GWAS in populations of European, Asian and African ancestry[37–40]. Fine scale mapping of the 16q12.1 locus has revealed that rs4784227 is the most strongly associated SNP in European and Asian populations[22,41]. *In vitro* experiments have shown that the risk allele T reduces luciferase activity and alters DNA-protein binding patterns, suggesting a regulatory role for rs4784227[22]. Functionally, TOX3 belongs to the HMG-box family of chromatin structure modifying proteins[42]. It is expressed mainly in epithelial cells and targets both anti-apoptotic and pro-apoptotic transcripts, protecting against cell death[43]. Furthermore, it stimulates estrogen-response element (ERE)-dependent transcriptional programs[43]. Statistically, *TOX3* is differentially expressed in breast cancers that relapse to bone[44]. SNPs contained within the gene interact multiplicatively with mutations in BRCA1 or BRCA2 increasing breast cancer risk[45]. Here, we show that the risk allele T at rs4784227 produces a five-fold decrease in *TOX3* gene expression. This reduction is due to an increase in affinity

for FOXA1 at the enhancer where rs4784227 is located. FOXA1 binding is coupled to the recruitment of the TLE repressor, which diminishes the strength of the enhancer (H3K9Ac reduction). Mediated by a chromatin loop to the *TOX3* promoter, the weakened rs4784227 enhancer downregulates the gene's transcriptional output (Fig. 8). This mechanism of action for a cancer risk-associated SNP, i.e., the disruption of transcription factor binding at a distal enhancer and the subsequent change in gene expression, such as reported in colorectal cancer, appears to be a common mechanism of action for risk-associated SNPs[23–25].

There is a striking disparity between the number of SNPs associated to disease and the number of SNPs (rs6983267, rs4784227 and rs1859962) that have been functionally characterized[46,47]. Therefore, it has been impossible to extrapolate the general mechanisms of genetic risk promotion for any particular disease. Here, we show that for BCa, the majority of risk-associated SNPs modulate FOXA1 binding. First, they are in complete linkage disequilibrium with SNPs contained in sites of FOXA1 binding. And second, these linked SNPs are capable of changing the recruitment of FOXA1 in a significant manner. Pioneer factors such as FOXA1 and lineage-specific factors such as ESR1 underlie the transcriptional programs that establish cell identity[4]. Accordingly, we have shown that the majority of SNPs that can disrupt normal breast cell identity modulate the binding of the FOXA1 pioneer factor.

It is time to shift gears towards the post-GWAS phase of human genetics[48]. Freedman et al. have proposed a set of principles for the post-GWAS functional characterization of cancer risk loci[46]. But there is still one major hurdle to overcome before this transition can be made. The number of GWAS raSNPs is large and growing, and each raSNP has hundreds or thousands of ldSNPs. Furthermore, because the majority of ldSNPs lie in regulatory regions, each ldSNP has an array of potential gene targets. This branching morass makes the exhaustive functional characterization of all GWAS results unfeasible. A mechanism is required with which to prioritize this over-abundance of putative causal SNPs. The methods developed and applied in this study provide such a mechanism (Supplementary Fig. 4). First, they determine the most promising transcription factors, histone modification or other genomic annotations to interrogate (VSE). And second, they identify candidate SNPs among the thousands of ldSNP for experimental follow-up studies (IGR). This integration of functional genomic data will allow the post-GWAS characterization of risk loci to gain traction and advance.

URLs. NHGRI GWAS Catalog, www.genome.gov/gwastudies; Nuclear Receptor Cistrome project, cistrome.dfci.harvard.edu/NR_Cistrome/.

Online Methods

Computational methods

Variant Set Enrichment—Variant Set Enrichment (VSE) is a computational method that calculates a score and a p-value for the enrichment or depletion of a set of variants in a genomic annotation. Genomic annotations are lists of chromosomal coordinates to which a particular property or function has been attributed. Single nucleotide polymorphisms (SNPs) are used in genome-wide association studies (GWAS) in the search of variation underlying

genetic traits and diseases. Disease risk-associated SNP (raSNP) in particular are the focus of this study and are the constituents of the variant sets tested for enrichment. SNPs contained within the haplotype block of an raSNP are referred to as its ldSNPs (linkage disequilibrium SNPs). Each individual raSNP and its collection of ldSNPs are referred to as an raSNP/ldSNP cluster. The set of clusters for all SNPs associated with a single trait or disease is in turn referred to as the Associated Variant Set (AVS) for that trait or disease.

The first step is to obtain all raSNPs for a disease of interest from the GWAS catalog (http://www.genome.gov/gwastudies/). We then identify the list of all SNPs in strong LD (ldSNPs) with each raSNP by using HapMap project data (UCSC Genome Browser's CEPH HapMap Linkage Disequilibrium Phase II table, hg18, hapmapLdPhCeu). We used a high-stringency LD threshold based on LOD > 2 and D' > 0.99. Haplotype blocks are heterogeneous in size and porous, i.e. strong LD values are not contiguous across the block and are interspersed with variants with no linkage (Fig. 1a). The first measure that is calculated is the mapping tally between the AVS and the genomic annotation. The mapping tally is the number of raSNP/ldSNP clusters in the AVS that have at least one ldSNP that overlaps the genomic annotation. The intersectBed program from the BEDTools suite is used to compute the overlap between chromosomal coordinates. Each raSNP/ldSNP cluster is intersected independently. The mapping tally is a preliminary measure of the functional relationship between the AVS and the genomic annotation. The mapping tally is also a function of confounding factors such as the size and structure of the haplotype blocks and the abundance and distribution of the genomic annotation. In order to correct for these factors, we build a null distribution for the mapping tally based on randomly permuting the AVS.

In order to account for the size and structure of the haplotype blocks in the AVS, each permuted variant set is matched to the original AVS, cluster by cluster. In order to account for the genomic heterogeneity of the annotation, sets of variants are sampled at random from a comprehensive list of tag variants used in GWAS (Illumina HumanOmniExpress). For each randomly sampled tag SNP (pseudo-raSNPs) its set of ldSNPs is imputed from HapMap data. Each set is built so as to be composed of a collection of pseudo-raSNPs/ldSNP clusters that match the size and structure of the AVS. We refer to these as Matched Random Variant Sets (MRVSs). Through this process, all variants with a given number of LD variants in the sampling list are binned together. For each raSNP in the AVS, a pseudo-raSNP is chosen at random from the bin that matches its number of LD variants. pseudo-raSNP are sampled with replacement within each bin and their distribution across bins is checked to ensure that no bin is depleted. All raSNP contained in the AVS are removed from the sampling list. Finally, the chromosomal coordinates for all pseudo-raSNP and ldSNP variants belonging to the MRVS are obtained.

We then derive the null distribution of the mapping tally between an AVS and a genomic annotation by intersecting collections of MRVSs with the annotation (Fig. 1d–g, gray histograms). Specifically, we estimate the probability of finding a set of variants with a mapping tally equal to that of the observed variants by chance alone. In order to obtain an enrichment score that is comparable across genomic annotations each AVS mapping tally is centered to the median and scaled to the standard deviation of its null distribution. The

enrichment score is therefore the number of standard deviation that the mapping tally deviates from the null mapping tally mean.

An approximate p-value can be obtained directly from the null distribution if the AVS mapping tally falls within the null's range. This is calculated as the ratio between the number of MRVSs that have a more extreme value than the AVS and the total number of MRVSs. Depending on which tail end of the distribution is used we obtain a p-value for enrichment or depletion. When the mapping tally falls near the end of the null's range, it becomes increasingly imprecise. When the mapping tally falls beyond the null's range, the p-value is smaller than the inverse of the number of MRVSs in the null. An exact p-value requires fitting a density function to the null distribution derived from the MRVSs. The Kolmogorov-Smirnov test is used to check that the distribution does not deviate significantly from normality. If the null distribution fails the Kolmogorov-Smirnov test, the Box-Cox procedure is used to find a transformation of the null that approaches normality. If the Box-Cox procedure fails to normalize the null, the test is considered non-significant. Exact p-values are then calculated from the density function. All methods are implemented in the R statistical environment. Functions and packages used: box.cox(car), ks.test (stats), pnorm(stats).

VSE for non-disjoint raSNP/ldSNP clusters—Most of the AVS considered in this study have one or more raSNP with ldSNPs in common. When this is the case, a special scoring procedure must be used for VSE to ensure that raSNPs are not counted more than once when only one ldSNP maps to a functional genomic element. This would result in inflated mapping tallies and could lead to false positives. A program was created, LDXI (linkage disequilibrium extension with intersections), which computes all intersections within the set of raSNPs. The ldSNPs that are shared by one or more raSNP become separate raSNP/ldSNP cluster. We call this the `intersection' class of clusters and it is treated separately when calculating the mapping tally. The rest of the ldSNPs are grouped into raSNP/ldSNP clusters as in the standard version of VSE and are added to the `unique' class. Therefore, the list of both intersection and unique clusters is disjoint. The non-disjoint VSE scoring algorithm works as follows. Tally all unique clusters as in the standard VSE. Look up all raSNPs in each intersection cluster in the unique tally. If all raSNPs in the cluster are accounted for, discard cluster. If at least one of the raSNPs in the cluster is not present in the unique tally, count the cluster towards the mapping tally once. This counting mechanism ensures that each raSNP that is counted toward the tally has at least one putatively functional variant to account for its association that it does not share with any other raSNP.

Intra-Genomic Replicates—Intra-Genomic Replicates (IGR) is a computational method that calculates the modulation in affinity produced by a SNP at a transcription factor binding site. The method requires transcription factor ChIP-chip or ChIP-seq data and the human genome sequence. The affinity between a K-mer and a transcription factor can be obtained from ChIP-seq data in the following manner. First, all instance of the K-mer are searched genome-wide. Second, the binding information at each of the instances of the K-mer is retrieved from the ChIP-seq data. Finally, the signal from the ChIP-seq data set is averaged

over all instances to obtain an affinity model for that K-mer. Each model is derived from a different number of instances based on the frequency of the K-mer.

A set of scrambled K-mers is used to obtain a baseline affinity for each model and to separate the import of sequence and composition on affinity (Supplementary Fig. 1, black profiles). The variable number of instances for each K-mer over different values of K confounds the comparison between sliding windows. We picked a stringent cutoff of one thousand instances to discard unreliable observations based on their signal to noise ratio within each node in the lattice (Fig. 3c). Furthermore, among possible values of K, eight provided a good tradeoff between number of instances and sequence specificity. The two collections of 8-mer affinity models are filtered based on number of instances. Among the remaining windows for both alleles, the affinities are compared across the two allele lattices. The number of cells in the lattice that satisfy the instance threshold is recorded and results can be filtered based on completeness of the assessment. This comparison between maxima represents the highest possible affinity for each allele given its genomic context. The default parameters include a restriction of our K-mer search to chromatin regions favorable to binding; for instance only K-mer matches that fall within H3K4me2 or DNaseI hypersensitive regions are considered. To further increase the contrast of IGR, open chromatin elements are only allowed to contain a single copy of the K-mer in order to avoid cooperative binding to chromatin. T-tests are used to assess the statistical significance of the affinity modulation between the two K-mers with the maximum affinities.

Enrichment of affinity modulating SNPs—The VSE and IGR methods can be combined to calculate a score and p-value for the enrichment or depletion of a set of variants in affinity modulators. We first determine how many of the SNPs in the AVS map to transcription factor binding sites or histone modifications favorable to binding. These represent the regulatory regions where binding sites can be created or destroyed; we refer to these as regSNPs, which are processed in a similar manner to ldSNPs. We then calculate the modulator quotient. The numerator of this quotient is the number of raSNP/regSNP clusters that have at least one significant affinity modulating regSNP. The denominator is the number of raSNP/regSNPs clusters in the AVS. raSNPs that do not contain regSNPs are not considered in the modulator quotient. The modulator quotient null distribution over the set of MRVSs is then used as in VSE to calculate a p-value for the enrichment of the AVS in affinity modulating SNPs.

Experimental procedures

Nucleosome-resolution ChIP-seq assays

The MCF7 breast cancer cells were treated with MNase (Sigma N3755) and ChIP was done with H3K4me2 antibody (Millipore CMA303) as previously published[34]. Libraries were prepared for Illumina Genome Analyzer according to manufactures instructions. H3K4me2 significantly ($P<10e^{-5}$) enriched regions were detected using the MACS software using default parameters[49]. H3K4me2 enrichment was validated by ChIP-qPCR. Files are available on GEO, accession number: GSE31151.

In vivo allele-specific binding assays—ChIP assays were performed as previously described[4]. Antibodies used in ChIP assays include: FoxA1 (abcam: ab5089), H3K4me2 (Millipore 07-030), TLE (santacruz biotechnology: sc-13373X) and H3K9Ac (Abcam: ab4441). Allele-specific MAMA PCR-based genotyping technique was applied to assess differential ChIP enrichment on heterozygous alleles as described previously[47,50].

3C-qPCR assays—Chromosome conformation capture (3C) technology coupled to qPCR was performed according to published protocol[51]. Briefly, 5 million MCF7 cells were processed using formaldehyde crosslinking (1%, 10 minutes at room temperature) of interacting chromatin segments. This was followed by BglII (or MspI for a higher resolution) digestion (400 units, overnight at 37°C) and ligation (T4 DNA ligase 4000 units, 4 hours at 16°C). DNA fragments were cleaned by phenol-chloroform, followed by ethanol extraction. qPCR were performed to quantify ligated DNA fragments. Bacterial artificial chromosome (BAC) clones were used to verify primer efficiency and normalize the 3C interaction frequency.

In vivo allele-specific gene expression—A primer outside of the rs2193094 SNP and its closest MspI restriction enzyme site and a primer outside of the rs4784227 SNP and its closest MspI site were used to PCR amplify the MspI 3C product from MCF7 cells. PCR amplified products were first cloned into an empty vector and then sequenced using the Sanger sequencing approach, which revealed the linkage between the two alleles of the rs2193094 and rs4784227 SNPs. MCF7 genomic DNA was extracted using QIAGEN Dneasy blood and tissue kit. MCF7 nuclear total RNA was extracted using Trizol (Invitrogen), followed by RT reaction to convert RNA into cDNA. Primers surrounding the rs2193094, rs35925303 and rs11646260 SNPs were used to generate cDNA and for PCR assays. Sanger sequencing was performed to measure the DNA and RNA level of each allele of the rs2193094, rs35925303 and rs11646260 SNPs.

Cell line-based Expression Quantitative Trait Loci assay (e-QTL)—The rs4784227 SNP was genotyped in eleven ERα+ and FOXA1+ breast cancer cell lines (T47D, HCC1428, MCF7, MDA-MB-415, UACC-812, ZR-75-1, BT-474, BT-483, MDA-MB-175VII, 600MPE and ZR-75-30). The gene expression data for *TOX3*, *AKTIP*, *CHD9*, *FTO*, *RBL2* and *RPGRIP1* (probes: 214774_x_at, 218373_at, 212615_at, 209702_at, 212331_at and 206608_at, respectively) in these cell lines was extracted from the Neve breast cancer cell lines database[52]. SNP genotype data was then correlated with expression level of each of the above genes.

Cell proliferation assay after si*TOX3*—*TOX3* was silenced by small-interfering RNA duplexes (siRNAs). Two sets of siRNAs against *TOX3* were designed through Dharmacon (Supplementary table 21). siRNA against Luciferase was used as a negative control (siLuc). RNA was isolated from ZR75-1 breast cancer cells using RNeasy mini kit (QIA-GEN) according to manufacturer's recommendation. Real-time reverse transcription-PCR (RT-qPCR) was done as previously described[4] to confirm the silencing effect of si*TOX3*. RNA level of *TOX3* was normalized to 28S ribosomal RNA (*R28S*), and further normalized to

siLuc. The number of ZR75-1 breast cancer cells was counted 2 days after *TOX3* silencing. ZR75-1 cells treated with siLuc were used as control.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Drs. Myles Brown and Matthew L. Freedman (Dana-Farber Cancer Institute), and Mathieu Lemaire (Yale University) for discussions as well as Drs. Benjamin G. Neel and Richard Marcotte for technical assistance. We acknowledge support from the NCI (2P30CA023108-32), the ACS (IRG-82-003-27 to M.L.) the NIH (R01LM009012 to J.H.M and R01CA155004 to M.L.) and the Princess Margaret Hospital Foundation (to M.L.).

References

1. Frazer K, Murray S, Schork N, Topol E. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics. 2009; 10:241–251.
2. Durbin R, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
3. Heintzman N, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics. 2007; 39:311–318. [PubMed: 17277777]
4. Lupien M, et al. Foxa1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell. 2008; 132:958–970. [PubMed: 18358809]
5. Heintzman N, et al. Histone modifications at human enhancers reflect global cell type-specific gene expression. Nature. 2009; 459:108. [PubMed: 19295514]
6. Khalil A, et al. Many human large intergenic noncoding rnas associate with chromatin-modifying complexes and affect gene expression. Proceedings of the National Academy of Sciences. 2009; 106:11667.
7. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. [PubMed: 21441907]
8. Zentner G, Tesar P, Scacheri P. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Research. 2011; 21:1273–1283. [PubMed: 21632746]
9. Lupien M, Brown M. Cistromics of hormone-dependent cancer. Endocrine-related cancer. 2009; 16:381. [PubMed: 19369485]
10. Schmidt D, et al. A ctcf-independent role for cohesin in tissue-specific transcription. Genome research. 2010; 20:578. [PubMed: 20219941]
11. Carroll J, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein foxa1. Cell. 2005; 122:33–43. [PubMed: 16009131]
12. Carroll J, et al. Genome-wide analysis of estrogen receptor binding sites. Nature genetics. 2006; 38:1289–1297. [PubMed: 17013392]
13. Krum S, et al. Unique erα cistromes control cell type-specific gene regulation. Molecular Endocrinology. 2008; 22:2393–2406. [PubMed: 18818283]
14. Wang Q, et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell. 2009; 138:245–256. [PubMed: 19632176]
15. Hindorff, L.; Junkins, H.; Mehta, J.; Manolio, T., et al. A catalog of published genome-wide association studies. [Accessed January 2012]
16. Grant S, Hakonarson H. Microarray technology and applications in the arena of genome-wide association. Clinical chemistry. 2008; 54:1116. [PubMed: 18499899]
17. McClellan J, King M. Genetic heterogeneity in human disease. Cell. 2010; 141:210–217. [PubMed: 20403315]
18. Hindorff L, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009; 106:9362.

19. Johnson W, et al. Model-based analysis of tiling-arrays for chip-chip. Proceedings of the National Academy of Sciences. 2006; 103:12457.
20. Akhtar-Zaidi B, et al. Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. Science. 2012; 336:736–739. [PubMed: 22499810]
21. Magnani L, Eeckhoute J, Lupien M. Pioneer factors: directing transcriptional regulators within the chromatin environment. Trends in Genetics. 2011
22. Long J, et al. Identification of a functional genetic variant at 16q12. 1 for breast cancer risk: results from the asia breast cancer consortium. PLoS Genetics. 2010; 6:e1001002. [PubMed: 20585626]
23. Pomerantz M, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with myc in colorectal cancer. Nature genetics. 2009; 41:882–884. [PubMed: 19561607]
24. Tuupanen S, et al. The common colorectal cancer predisposition snp rs6983267 at chromosome 8q24 confers potential to enhanced wnt signaling. Nature genetics. 2009; 41:885–890. [PubMed: 19561604]
25. Wright J, Brown S, Cole M. Upregulation of c-myc in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Molecular and cellular biology. 2010; 30:1411. [PubMed: 20065031]
26. Sekiya T, Zaret K. Repression by groucho/tle/grg proteins: genomic site recruitment generates compacted chromatin in vitro and impairs activator binding in vivo. Molecular cell. 2007; 28:291–303. [PubMed: 17964267]
27. Riaz M, et al. Correlation of breast cancer susceptibility loci with patient characteristics, metastasis-free survival, and mRNA expression of the nearest genes. Breast Cancer Res Treat. 2011
28. De Gobbi M, et al. A regulatory snp causes a human genetic disease by creating a new transcriptional promoter. Science. 2006; 312:1215. [PubMed: 16728641]
29. Gaulton K, et al. A map of open chromatin in human pancreatic islets. Nature genetics. 2010; 42:255–259. [PubMed: 20118932]
30. Jia L, et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genetics. 2009; 5:e1000597. [PubMed: 19680443]
31. McDaniell R, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010; 328:235. [PubMed: 20299549]
32. Reddy T, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Research. 2012
33. Eeckhoute J, Carroll J, Geistlinger T, Torres-Arzayus M, Brown M. A cell-type-specific transcriptional network required for estrogen regulation of cyclin d1 and cell cycle progression in breast cancer. Genes & development. 2006; 20:2513. [PubMed: 16980581]
34. He H, et al. Nucleosome dynamics define transcriptional enhancers. Nature genetics. 2010; 42:343–347. [PubMed: 20208536]
35. Laganière J, et al. Location analysis of estrogen receptor α target promoters reveals that foxa1 defines a domain of the estrogen response. Proceedings of the National Academy of Sciences. 2005; 102:11651.
36. Hurtado A, Holmes K, Ross-Innes C, Schmidt D, Carroll J. Foxa1 is a key determinant of estrogen receptor function and endocrine response. Nature genetics. 2010; 43:27–33. [PubMed: 21151129]
37. Easton D, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007; 447:1087–1093. [PubMed: 17529967]
38. Stacey S, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics. 2007; 39:865–869. [PubMed: 17529974]
39. Ruiz-Narváez E, et al. Polymorphisms in the tox3/loc643714 locus and risk of breast cancer in african-american women. Cancer Epidemiology Biomarkers & Prevention. 2010; 19:1320.
40. Hutter C, et al. Replication of breast cancer gwas susceptibility loci in the women's health initiative african american share study. Cancer Epidemiology Biomarkers & Prevention. 2011; 20:1950–1959.
41. Udler M, et al. Fine scale mapping of the breast cancer 16q12 locus. Human molecular genetics. 2010; 19:2507–2515. [PubMed: 20332101]

42. O'Flaherty E, Kaye J. Tox defines a conserved subfamily of hmg-box proteins. BMC genomics. 2003; 4:13. [PubMed: 12697058]
43. Dittmer S, et al. Tox3 is a neuronal survival factor that induces transcription depending on the presence of cited1 or phosphorylated creb in the transcriptionally active complex. Journal of Cell Science. 2011; 124:252–260. [PubMed: 21172805]
44. Smid M, et al. Genes associated with breast cancer metastatic to bone. Journal of Clinical Oncology. 2006; 24:2261–2267. [PubMed: 16636340]
45. Antoniou A, et al. Common breast cancer-predisposition alleles are associated with breast cancer risk in brca1 and brca2 mutation carriers. The American Journal of Human Genetics. 2008; 82:937–948. [PubMed: 18355772]
46. Freedman M, et al. Principles for the post-gwas functional characterization of cancer risk loci. Nature genetics. 2011; 43:513–518. [PubMed: 21614091]
47. Zhang X, et al. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Research. 2012
48. Park J, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature genetics. 2010; 42:570–575. [PubMed: 20562874]
49. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
50. Li B, Kadura I, Fu DJ, Watson DE. Genotyping with TaqMAMA. Genomics. 2004; 83:311–20. [PubMed: 14706460]
51. Hagege H, et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc. 2007; 2:1722–33. [PubMed: 17641637]
52. Neve RM, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell. 2006; 10:515–27. [PubMed: 17157791]

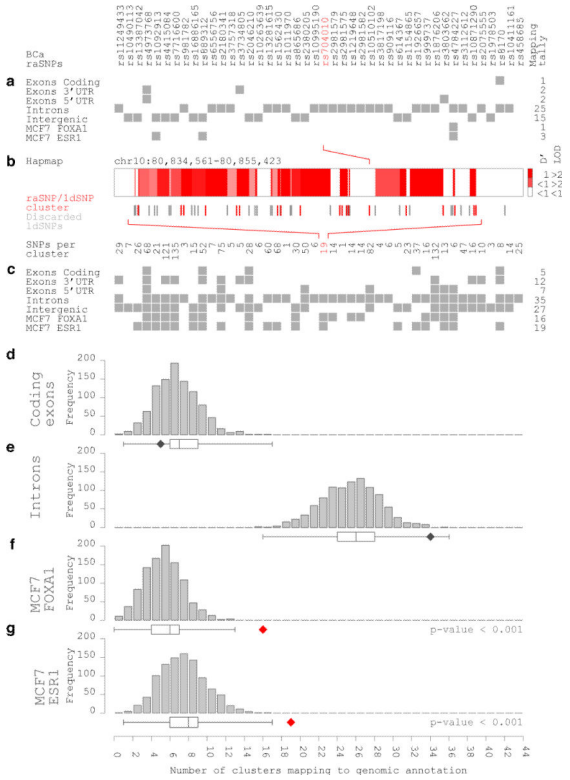

Figure 1.

The BCa AVS is enriched in the cistromes of FOXA1 and ESR1. (**a**) Breast cancer (BCa) raSNPs. Binary matrix encodes raSNPs mapping to the genomic annotations. The mapping tally shows the number of raSNPs per annotation. (**b**) Haplotype block for SNP rs704010. SNPs included in raSNP/ldSNP clusters are highlighted in red. The bottom row of numbers indicates the number of ldSNPs per raSNP/ldSNP cluster. (**c**) BCa raSNP/ldSNP clusters. Binary matrix encodes clusters with at least one SNP mapping to the genomic annotations. The mapping tally shows the number of clusters per annotation. (**d–g**) Histograms and box plots showing the null distributions of the mapping tallies for each annotation. Nulls are based on 1000 Matched Random Variant Sets (MRVSs). Diamonds show mapping tallies for the BCa clusters. Red diamonds highlight mapping tallies for genomic annotations that fall outside of the null (P<0.001)

Figure 2.

The enrichment of the BCa AVS is factor, cell-type and cancer-type specific. (**a–f**) Variant Set Enrichment (VSE) plots for the breast cancer AVS. Box plots in each panel show the normalized null distributions. Diamonds show the corresponding VSE scores. Colored diamonds highlight mapping tallies for genomic annotations that satisfy a Bonferroni corrected threshold for significance ($P<10^{-3.3}$). P-values are based on null distributions based on 1000 MRVSs. Binary matrices encode raSNP/ldSNP clusters with at least one SNP mapping to the genomic annotations. Rows highlighted in dark gray show statistically significant enrichment for genomic annotations. (**a**) Core gene annotations. (**b**) Untreated epigenomes in BCa cells. (**c**) Treated epigenomes in BCa cells. (**d**) Untreated cistromes in BCa cells. (**e**) Treated cistromes in BCa cells. (**f**) Cistromes and epigenomes for control cell lines. (**g**) VSE plot for the prostate cancer AVS against FOXA1 cistromes. (**h**) VSE plot for the bone mineral density AVS against ESR1 cistromes. (**i**) VSE plot for the colorectal cancer AVS against H3K4me1 epigenomes.

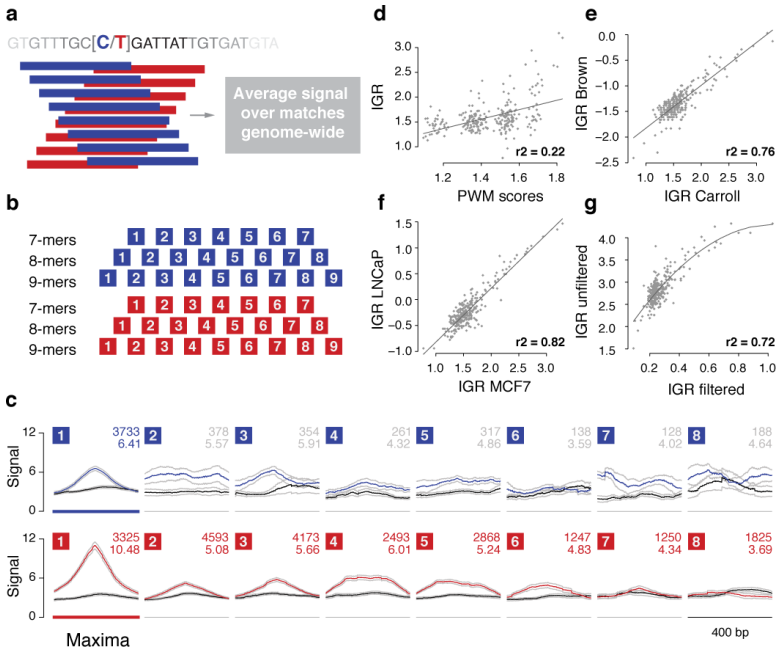

Figure 3.

The Intra-Genomic Replicates method. (**a**) Affinity models are calculated over a sliding window of K-mers around the SNP of interest for both reference (C, blue) and variant alleles (T, red). (**b**) Lattice structures of connected K-mer affinity models. (**c**) Lattice rows corresponding to 8-mers. Each cell, numbered 1 to 8, corresponds to an affinity model. The top numbers on the right of each cell indicate the number of genomic matches for each 8-mer. The bottom numbers indicate the average binding signal across all 8-mer matches. Greyed numbers indicate discarded cells within the lattice due to an insufficient number of matches. Colored lines show the averaged binding profiles over a 400 bp window centered on each 8-mer match. Black lines show the averaged binding profile for the scrambled 8-mer sequence. (**d**) Intra-Genomic Replicates (IGR) affinity scores for FOXA1 in MCF7 cells for a set of 256 8-mers against the Position Weighted Matrix (PWM) scores based on the forkhead (FKH) motif for the same set of 256 8-mers. (**e**) Comparison between IGR scores for FOXA1 in MCF7 cells obtained in two separate laboratories. (**f**) Comparison between IGR scores for FOXA1 obtained in MCF7 BCa cells and LNCaP PCa cells. (**g**) Comparison between IGR scores for FOXA1 in MCF7 cells obtained with or without the H3K4me2 accessibility filter.

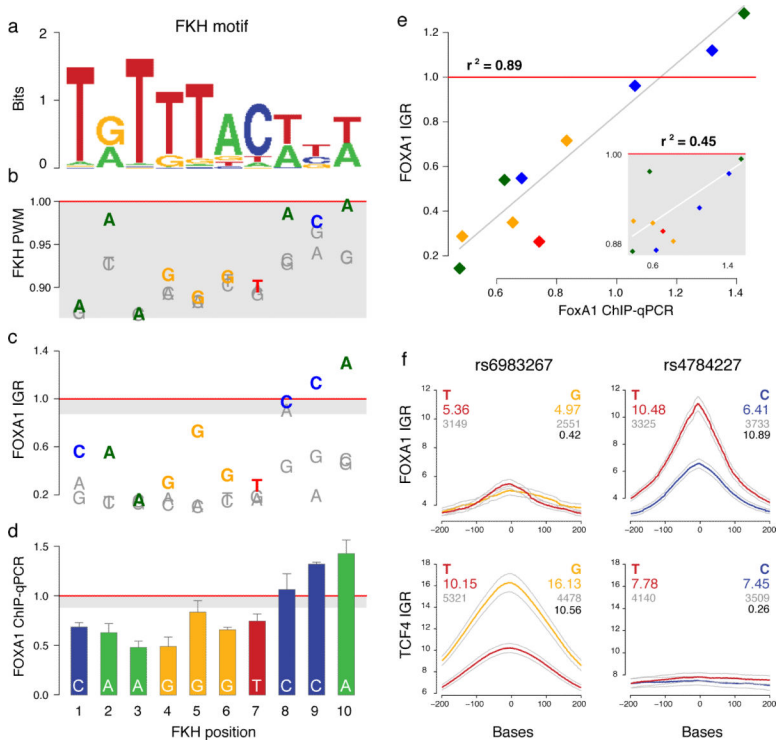

Figure 4.

The BCa raSNP rs4784227 modulates FOXA1 affinity by altering the forkhead (FKH) motif. (**a**) Sequence logo of the FKH motif. (**b,c**) FKH motif Position Weighted Matrix (PWM) and FOXA1 Intra Genomic Replicates (IGR) scores for each single base variation to the FKH motif. Values are normalized to their respective scores for the canonical FKH motif (red line). Gray areas shows the range of PWM scores. (**d**) FOXA1 ChIP-qPCR in MCF7 for each of the highest scoring variants across FKH positions. Each measurement is based on three replicates. FKH variants for each position assayed through ChIP-qPCR are highlighted in color and bold typeface in **b** and **c**. (**e**) Scatter plots and regressions for the comparison between the FOXA1 IGR scores and FOXA1 ChIP-Seq profiles in MCF7. The inset shows the comparison between FKH PWM scores and ChIP-Seq profiles. (**f**) IGR profiles for SNPs rs6983267 and rs4784227. Colored numbers: average binding across instances. Black: -log10 of the p-values obtained by IGR. Gray: numbers of K-mer instances in the genome.

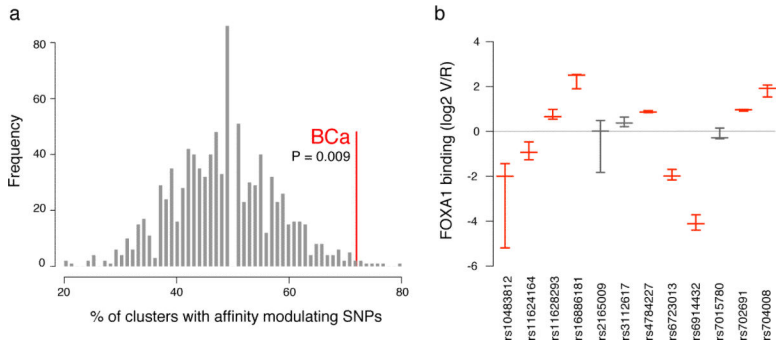

Figure 5.

The BCa AVS is enriched in affinity modulating SNPs. (**a**) The percentage of FOXA1 affinity modulating raSNP/ldSNP clusters over all clusters with SNPs mapping to FOXA1 or H3K4me2 sites in BCa raSNPs. The histogram shows the null distribution of this percentage over 1000 MRVSs. (**b**) Allele-specific ChIP-qPCR result for the 13 heterozygous SNPs found in BCa cell lines. *y* axis represents log2-transformed fold change of FOXA1 binding level between variant and reference alleles for each SNP. Red indicates the SNPs that demonstrate a significant and concordant allele-specific binding preference for FOXA1 predicted by IGR.

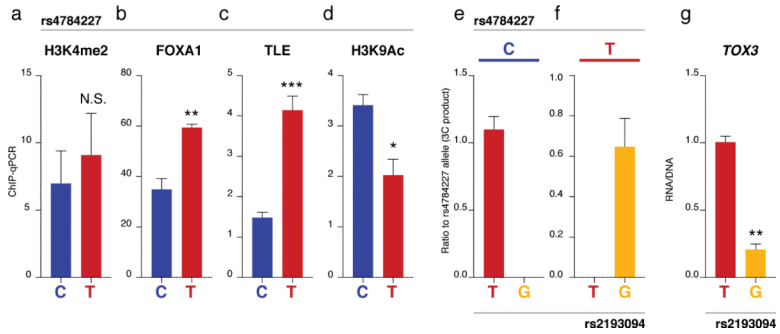

Figure 6.

The BCa raSNP rs4784227 disrupts enhancer function through FOXA1 affinity modulation. (**a–d**) ChIP-qPCR for H3K4me2 histone modification (enhancer), FOXA1 and TLE binding and H3K9Ac (active enhancer) at rs4784227. The C allele is the reference and the T allele is the risk variant. (**e,f**) 3C sequencing results showing the physical interactions between heterozygous SNPs at the enhancer (rs4784227) and *TOX3* intron (rs2193094). (**g**) Ratio between RNA and DNA levels for both alleles in the *TOX3* intron. Error bars, SEM from three replicate measures.

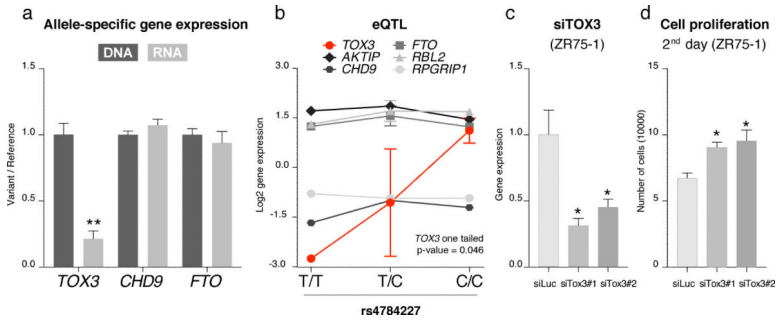

Figure 7.

The BCa raSNP rs4784227 affects cell proliferation by disrupting *TOX3* gene expression. (**a**) Ratio between variant and reference alleles for DNA and RNA levels of each SNP (rs2193094 in *TOX3*; rs35925303 in *CHD9*; rs11646260 in *FTO*). (**b**) Cell line-based e-QTL results for the SNP rs4784227 and expression level of the gene *TOX3* (rs4784227: T47D, T/T; HCC1428, C/T; MCF7, C/T; MDA-MB-415, C/T; UACC-812, C/C; ZR-75-1, C/C; BT-474, C/C; BT-483, C/C; MDA-MB-175VII, C/C; 600MPE, C/C; ZR-75-30, C/C). (**c**) TOX3 silencing in ZR75-1 cells. (**d**) The number of cells counted 2 days after *TOX3* silencing. N.S., not significant; *, p≤0.05; **, p≤0.01; ***, p≤0.001. Error bars, SEM from three replicate measures.

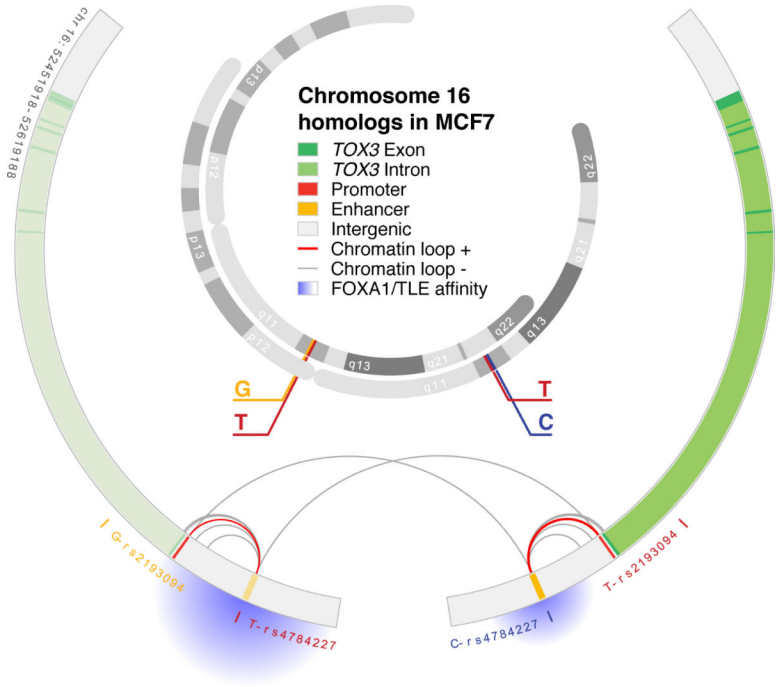

Figure 8.

Diagram of the *TOX3* locus and the physical interactions between and within chromosome 16 homologs.