nature genetics

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

Richard Cowper-Sal·lari[1,2,7], Xiaoyang Zhang[1,2,7], Jason B Wright[1], Swneke D Bailey[3,4], Michael D Cole[1], Jerome Eeckhoute[5,6], Jason H Moore[1,2] & Mathieu Lupien[3,4]

Genome-wide association studies (GWAS) have identified thousands of SNPs that are associated with human traits and diseases. But, because the vast majority of these SNPs are located in non-coding regions of the genome, the mechanisms by which they promote disease risk have remained elusive. Employing a new methodology that combines cistromics, epigenomics and genotype imputation, we annotate the non-coding regions of the genome in breast cancer cells and systematically identify the functional nature of SNPs associated with breast cancer risk. Our results show that breast cancer risk–associated SNPs are enriched in the cistromes of FOXA1 and ESR1 and the epigenome of histone H3 lysine 4 monomethylation (H3K4me1) in a cancer- and cell type–specific manner. Furthermore, the majority of the risk-associated SNPs modulate the affinity of chromatin for FOXA1 at distal regulatory elements, thereby resulting in allele-specific gene expression, which is exemplified by the effect of the rs4784227 SNP on the *TOX3* gene within the 16q12.1 risk locus.

SNPs that associate with disease risk map predominantly to non-coding regions of the genome[1]. Over 70% of the risk-association loci in the National Human Genome Research Institute (NHGRI) GWAS catalog lack variants that map to exons within their haplotype block, indicating that they do not disrupt coding sequences[2]. A 'cistrome' has been defined as the complete set of binding sites for a transcription factor, assessed across the genome for a given cell type under a specific treatment. Similarly, an 'epigenome' is the complete set of DNA elements that carry the same epigenomic mark for a specific cell type and treatment. Cistromes and epigenomes, defined by chromatin immunoprecipitation combined with sequencing (ChIP-seq), identify the positions of regulatory elements and novel transcripts in both coding and non-coding regions across the whole genome[3–8]. Notably, these approaches reveal 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 α; also known as ERα) have unique distributions in breast cancer cells compared to other cell types[4,11–13]. Similarly, comparison of the epigenomes in diverse cell types of specific histone modifications found at regulatory elements, such as mono- or dimethylation of lysine 4 on histone H3 (H3K4me1 or H3K4me2), shows that functional regulatory elements are also lineage specific[4,5]. The interplay between cistromes and epigenomes directs the transcriptional activity of the critical FOXA1 and ESR1 factors in breast cancer development to specific genomic regions, driving a unique gene expression profile[11,12,14]. Notably, 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 relationship between breast cancer–associated SNPs and cistromes and epigenomes in breast cancer using a novel integrative functional genomics approach.

RESULTS

Risk-associated SNP enrichment

We first gathered the coordinates for the SNPs associated with breast cancer risk from the GWAS catalog available through the NHGRI (see URLs). Of these SNPs, only 1 in 44 mapped to coding exons, with 25 mapping to introns, 1 to a FOXA1-binding site and 3 to ESR1-binding 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[15]. Statistically, it is therefore more likely for risk-associated SNPs to be in linkage disequilibrium (LD) with causal variants than to be causal themselves[16]. Therefore, we generated a list of all SNPs that were in strong LD with each breast cancer–associated SNP using data from the HapMap project (**Supplementary Table 1**). We used a highly stringent LD threshold of logarithm of odds (LOD) of >2 and *D*′ of 0.99. Through this imputation approach, we extended the coverage of risk-associated SNPs to include a more comprehensive list of putative functional SNPs that are in LD with these SNPs, thereby defining clusters that contain directly associated and linked SNPs (**Fig. 1b**). We refer to the comprehensive collection of

[1] Department of Genetics, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, New Hampshire, USA. [2] Institute for Quantitative Biomedical Sciences, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, New Hampshire, USA. [3] Ontario Cancer Institute, Princess Margaret Hospital–University Health Network, Toronto, Ontario, Canada. [4] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [5] Institut National de la Santé et de la Recherche Médicale (INSERM), Unité Mixte de Recherche (UMR) 1011, Université de Lille Nord de France, Lille, France. [6] Institut Pasteur de Lille, Lille, France. [7] These authors contributed equally to this work. Correspondence should be addressed to M.L. (mlupien@uhnres.utoronto.ca) or J.H.M. (jason.h.moore@dartmouth.edu).

Received 19 March; accepted 29 August; published online 23 September 2012; doi:10.1038/ng.2416

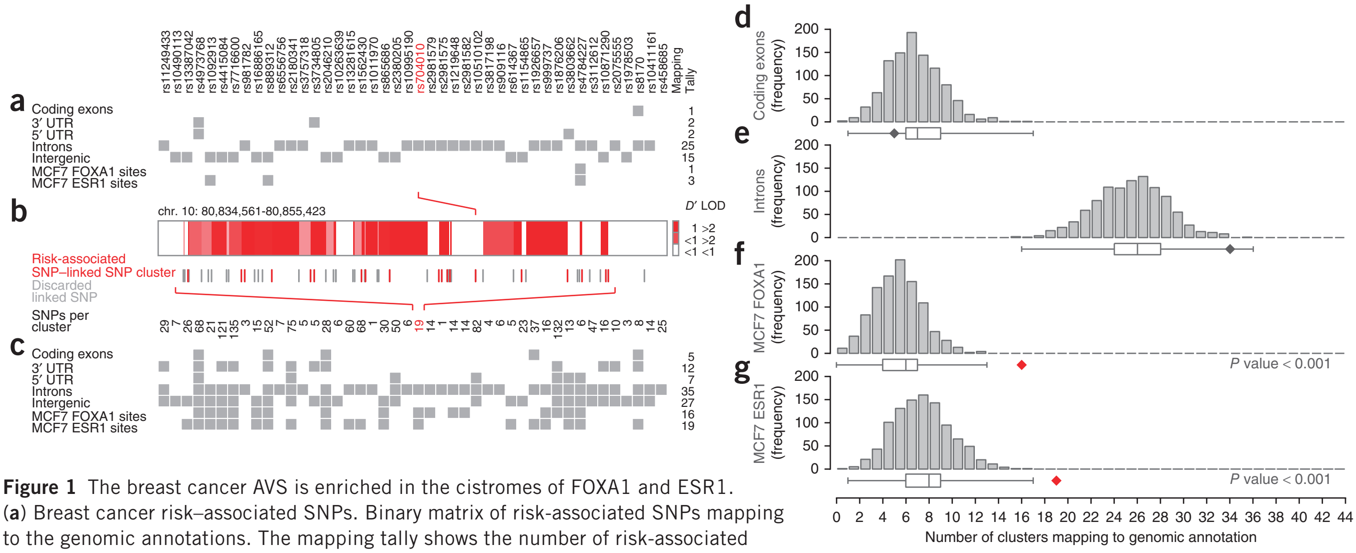


Figure 1 The breast cancer AVS is enriched in the cistromes of FOXA1 and ESR1. (**a**) Breast cancer risk–associated SNPs. Binary matrix of risk-associated SNPs mapping to the genomic annotations. The mapping tally shows the number of risk-associated SNPs per annotation. (**b**) Haplotype block for rs704010. The bottom row of numbers indicates the number of linked SNPs in each cluster of risk-associated and linked SNPs. (**c**) Clusters of breast cancer–associated and linked SNPs. Binary matrix of 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 the coding exon (**d**), intron (**e**), MCF7 FOXA1-binding site (**f**) and MCF7 ESR1-binding site (**g**) annotations. Null distributions are based on 1,000 matched random variant sets (MRVSs). Diamonds show mapping tallies for the breast cancer–associated clusters. Red diamonds highlight mapping tallies for genomic annotations that fall outside of the null distribution ($P < 0.001$).

all such clusters associated with a specific disease or trait, in this case, breast cancer, as its associated variant set (AVS). The breast cancer AVS is thus composed of 44 risk-associated SNPs and 1,315 linked SNPs (**Supplementary Table 1**). The genome-wide distribution of the clusters containing these SNPs is reminiscent of the binding profiles of many transcription factors, such as FOXA1 and ESR1, as the clusters predominantly map to intronic and intergenic regions[4,11,12,14]. By considering SNPs in LD with the risk-associated SNPs, the number of clusters mapping to coding exons increased from 1 to 5, and the number of clusters mapping to FOXA1- and ESR1-binding sites increased from 1 to 16 and 3 to 19, respectively (**Fig. 1c**). The significance of these mapping tallies was calculated through a novel computational method that we call variant set enrichment (VSE). VSE is a method based on permutation testing that generates null distributions from randomized variant sets while taking into account the heterogeneous LD structure of the breast cancer AVS (**Fig. 1d**–**g** and Online Methods). This method is an extension of the work of Hindorff and colleagues[17]. By using VSE, we determined that the breast cancer AVS is not enriched in genes (**Fig. 1d**,**e**) but is strongly enriched for the FOXA1 and ESR1 cistromes (**Fig. 1f**,**g**).

We extended the VSE analysis to include a total of 72 cistromes from 16 different transcription factors and 27 epigenomes from 8 different histone modifications in breast cancer cells, including MCF7, T47D and ZR-75-1 cells (**Fig. 2**). Samples were primarily treated with estradiol (E2), which was intended to stimulate the transcriptional activity of ESR1. Cistromes and epigenomes were either generated by our laboratory or were obtained from the literature and the Nuclear Receptor Cistrome project (see URLs). Only significant peaks ($P < 1 \times 10^{-5}$) for cistromes and epigenomes were used in the analysis[4,18]. Only enrichment scores satisfying a stringent Bonferroni-corrected threshold for significance (103 tests; $P < 5.01 \times 10^{-4}$) are reported subsequently.

Among the 24 distinct breast cancer epigenomes and exon and intron annotations assessed, only H3K4me1—a histone modification that is associated with regulatory elements, mainly enhancers—showed enrichment (in samples untreated and treated with estrogen) (**Fig. 2a**–**c** and **Supplementary Tables 2** and **3**). Among the 72 distinct breast cancer cistromes, only the FOXA1 (three samples, untreated and treated) and ESR1 (one sample, treated) transcription factors showed enrichment (**Fig. 2d**,**e** and **Supplementary Tables 4** and **5**). H3K4me1 enrichment was independent of that for FOXA1- and ESR1-binding sites. After subtracting FOXA1- and ESR1-binding sites from the list of H3K4me1-modified elements, the epigenome for this mark remained significant ($P = 2.94 \times 10^{-4}$) under estrogen treatment (**Supplementary Table 6**). Enrichment of FOXA1- and ESR1-binding sites and of H3K4me1-modified elements remained significant when using r^2 as a measure of LD instead of D' and LOD for thresholds of r^2 of 0.8 or greater (**Supplementary Tables 7**–**9**). These six enriched cistromes and epigenomes in breast cancer cells span 70% (31 of 44) of the clusters of risk-associated and linked SNPs for breast cancer (**Fig. 2a**–**e**, dark heatmap rows). Five out of these six enrichment events were also significantly enriched with the breast cancer AVS when a null distribution based on DNase accessibility was used in a subsequent VSE run (**Supplementary Tables 10** and **11**).

VSE analysis of the breast cancer AVS across all cistromes and epigenomes in breast cancer cells showed that the enrichment for FOXA1 and ESR1 cistromes and the H3K4me1 epigenome was highly factor specific (**Fig. 2a**–**e**). However, because not all FOXA1 or ESR1 cistromes showed significant VSE scores, enrichment for the cistromes of the other transcription factors assayed in the study cannot be completely ruled out. 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 in which data were generated. Given that we were intersecting 44 risk-associated SNPs with tens of thousands of binding sites, a tenfold decrease in the number of binding sites would be expected to markedly reduce our power to detect enrichment.

Enrichment of the breast cancer AVS for the FOXA1 and ESR1 cistromes and the H3K4me1 epigenome was also dependent on cancer

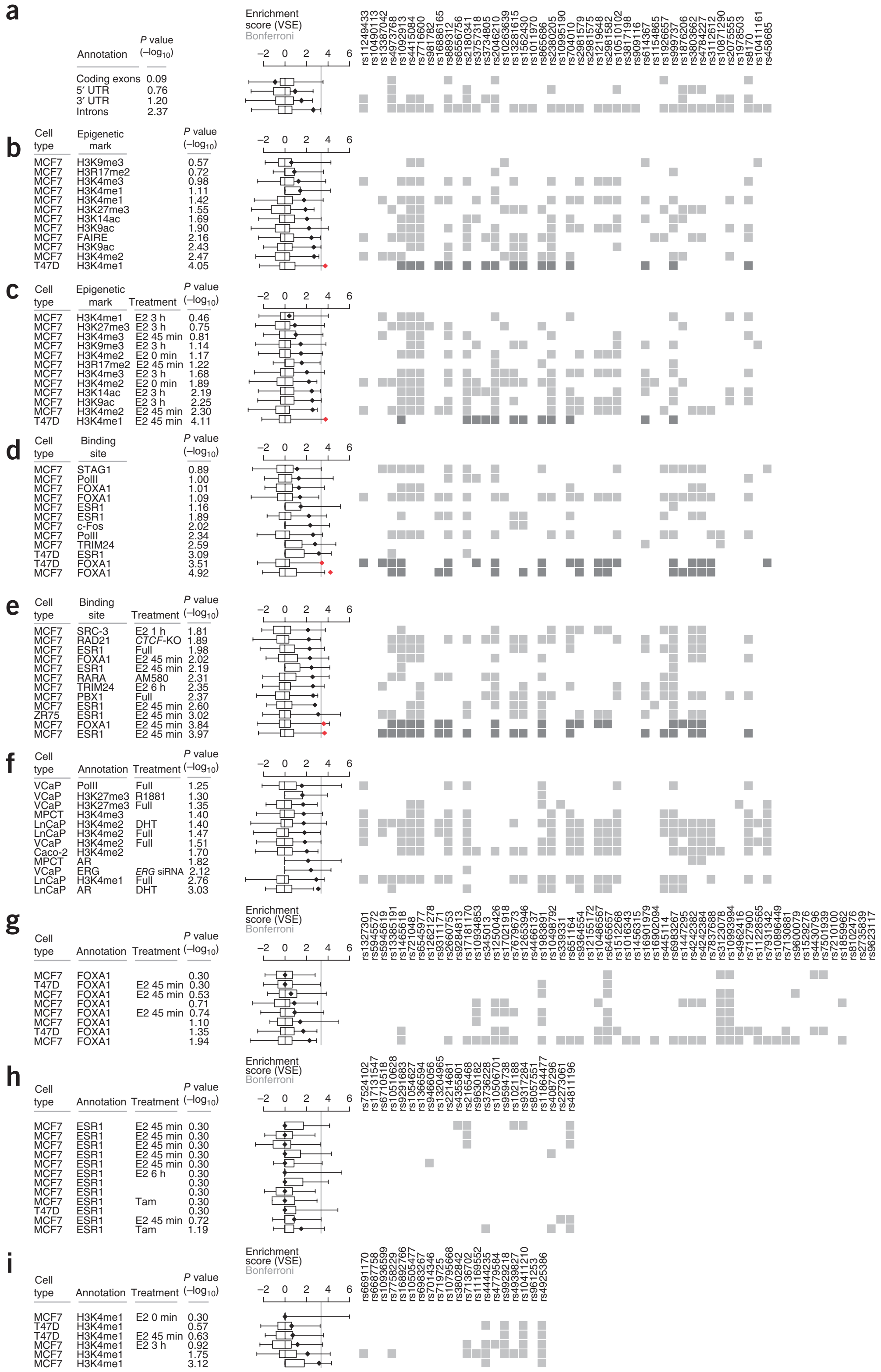


Figure 2 Enrichment of the breast cancer AVS is factor, cell type and cancer type specific. (**a**–**f**) VSE plots for the breast cancer AVS. (**a**) Core gene annotations. (**b**) Epigenomes in untreated breast cancer cells. (**c**) Epigenomes in treated breast cancer cells. (**d**) Cistromes in untreated breast cancer cells. (**e**) Cistromes in treated breast cancer 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 BMD AVS against ESR1 cistromes. (**i**) VSE plot for the CRC AVS against H3K4me1 epigenomes. 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 ($P < 5.01 \times 10^{-4}$). P values are based on null distributions from 1,000 MRVSs. Binary matrices show clusters of risk-associated and linked SNPs with at least one SNP mapping to the genomic annotations. Rows highlighted in dark gray show statistically significant enrichment for genomic annotations. KO, knockout; tam, tamoxifen; DHT, dihydrotestosterone; R1881, AR agonist; AM580, RARα agonist; E2, estradiol; full, full medium.

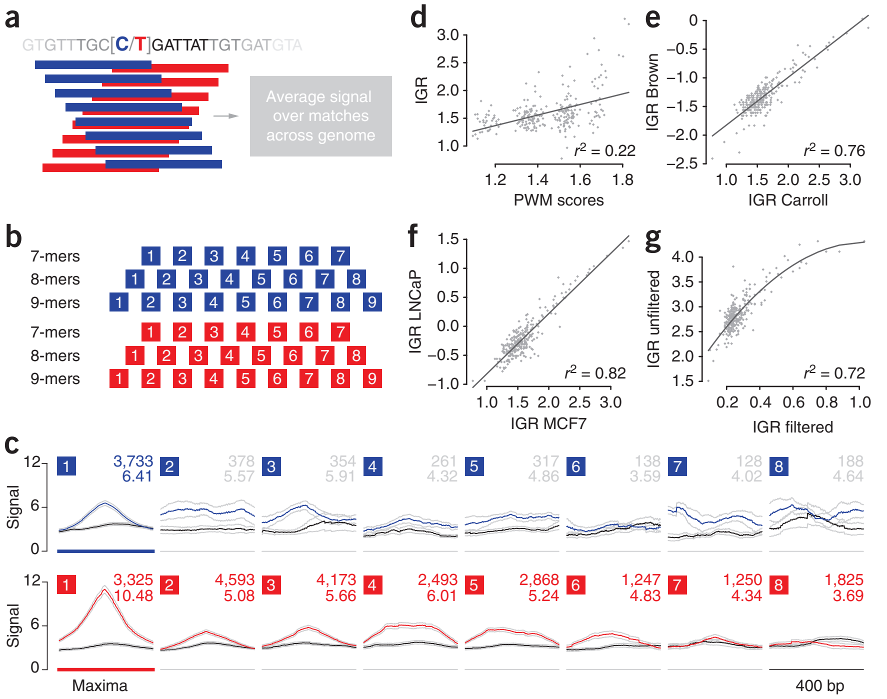


Figure 3 The intragenomic replicates (IGR) method. (**a**) Affinity models are calculated over a sliding window of *k*-mers around the SNP of interest for both reference (C, blue) and variant (T, red) alleles. (**b**) Lattice structures of connected *k*-mer affinity models. (**c**) Lattice rows corresponding 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. Gray 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**) IGR affinity scores for FOXA1 in MCF7 cells for a set of 256 8-mers against PWM scores based on the FKH motif for the same set of 256 8-mers. (**e**) Comparison of IGR scores for FOXA1 in MCF7 cells obtained in two separate laboratories. (**f**) Comparison of IGR scores for FOXA1 obtained in MCF7 breast cancer cells and LNCaP prostate cancer cells. (**g**) Comparison of IGR scores for FOXA1 in MCF7 cells obtained with or without the H3K4me2 accessibility filter.

and cell type. We used the AVSs of prostate cancer, bone mineral density (BMD) and colorectal cancer (CRC) to control for enrichment of FOXA1 and ESR1 cistromes and the H3K4me1 epigenome, respectively (**Supplementary Tables 12–14**). Prostate cancer development relies on FOXA1, BMD is affected by estrogen signaling and the CRC AVS is enriched for the H3K4me1 epigenome in tumor samples[19]. None of these variant sets were enriched in their target cistromes or epigenomes in breast cancer cells, showing that the enrichment of the breast cancer AVS for these cistromes and epigenomes is specific to this cancer type (**Fig. 2g–i** and **Supplementary Tables 15–17**).

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). Assessment of these cell lines for enriched transcription factor–binding sites and epigenetic modifications, including FOXA1, ESR2 and androgen receptor sites and H3K4me1 elements, respectively, did not show enrichment for the breast cancer AVS,

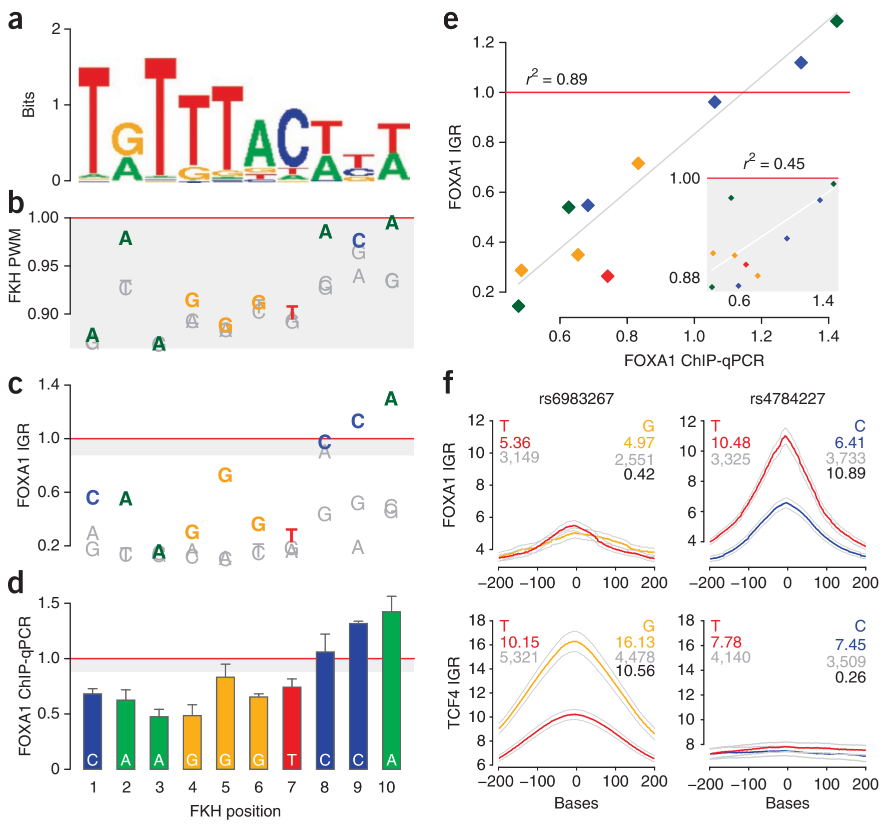


Figure 4 The breast cancer–associated SNP rs4784227 modulates FOXA1 affinity by altering the FKH motif. (**a**) Sequence logo of the FKH motif. (**b**,**c**) FKH motif PWM (**b**) and FOXA1 IGR (**c**) scores for each single-base variation to the FKH motif. Values are normalized to their respective scores for the canonical FKH motif (red line). Gray areas show the range of PWM scores. (**d**) FOXA1 ChIP assays combined with quantitative PCR (ChIP-qPCR) in MCF7 cells for each of the highest scoring variants across the 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 comparison of the FOXA1 IGR scores and FOXA1 ChIP-seq profiles in MCF7 cells. Inset, comparison of FKH PWM scores and ChIP-seq profiles. (**f**) IGR profiles for rs6983267 and rs4784227. Colored numbers, average binding across instances; gray numbers, number of *k*-mer instances in the genome; black numbers, $-\log_{10}$ *P* values obtained by IGR.

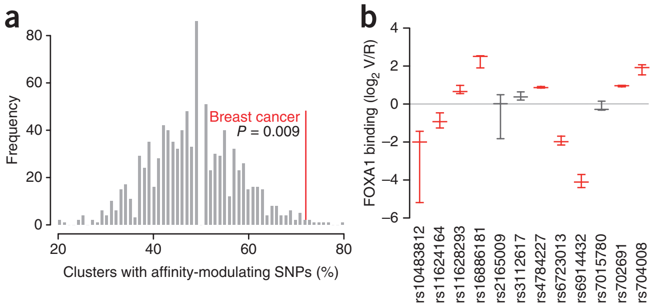


Figure 5 The breast cancer AVS is enriched for affinity-modulating SNPs. (**a**) The percentage of FOXA1 affinity-modulating clusters of risk-associated and linked SNPs over all clusters with SNPs mapping to FOXA1-binding or H3K4me2 modification sites in breast cancer–associated SNPs. The histogram shows the null distribution of this percentage over 1,000 MRVSs. (**b**) Allele-specific ChIP-qPCR result for the 13 heterozygous SNPs found in breast cancer cell lines. The *y* axis represents the $\log_2$-transformed fold change in FOXA1 binding between variant and reference alleles for each SNP. Red indicates the SNPs that show a significant and concordant allele-specific binding preference for FOXA1 as predicted by IGR. Error bars represent maximum and minimum values. Three replicates were used.

showing that the enrichment of the breast cancer AVS is cell type specific (**Fig. 2f** and **Supplementary Table 18**).

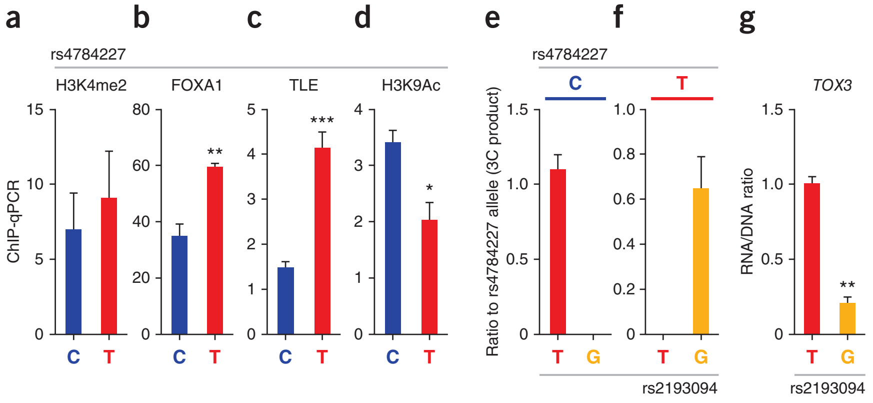
FOXA1 modulation

Having obtained these results, we then tested the ability of the clusters of breast cancer–associated and linked SNPs 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 intragenomic replicates (IGR), we are able to accurately predict affinity modulation by 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 the genomic context around the SNP. The affinity of a transcription factor for a particular DNA sequence of length *k* (a *k*-mer) can be obtained by averaging binding data across a genome-wide ChIP-seq data set 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 variant alleles (**Fig. 3b**). The affinity models within each lattice are then filtered on the basis of their signal-to-noise ratios (Online Methods). 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. We found that affinity model scores showed 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 were 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 restricted our *k*-mer search to chromatin regions that are favorable to binding[20]. When *k*-mer instances are not filtered on the basis of a favorable chromatin landscape, the affinity of the highly ranked *k*-mers is underestimated (**Fig. 3g**).

Recruitment of FOXA1 is dependent on recognition of the forkhead (FKH) motif (**Fig. 4a**). We explored the affinity landscape of the FKH motif by generating a collection of 30 10-mers, 1 for each possible variation at each of the 10 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 30 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 marked, even reversing the affinity-modulating properties of an allele if the contextual sequence is ignored (**Supplementary Fig. 1**).

The rs4784227 SNP is associated with breast cancer and is predicted to disrupt the binding of an undetermined transcription factor[21]. It is also a risk-associated SNP that directly maps to a FOXA1-binding site (**Fig. 1a**). This SNP localizes to position 8 of the FKH motif that is recognized by FOXA1 (**Fig. 4a**). The sequence surrounding this SNP differs from that encountered in the canonical FKH motif in that position 6 consists of G instead of A and the motif ends in GA instead of TT. The PWM for the FKH motif predicts a 9% affinity increase for the variant rs4784227[T] risk allele compared to the reference rs4784227[C] allele (6.24 versus 5.734, respectively; **Fig. 4b**). Taking into account the interaction effects between motif positions, the IGR method predicts a 63% increase in affinity for the rs4784227[T] variant allele compared to the rs4784227[C] reference allele (fold change of 1.63, $P < 1 \times 10^{-10}$; **Fig. 4f** and **Supplementary Fig. 1**). Previous reports have shown allele-specific TCF4 recruitment at the risk-associated rs6983267 SNP in colon cancer cells[22–24]. IGR validates the affinity modulation by rs6983267 of TCF4 (fold change of 1.59,

Figure 6 The breast cancer–associated SNP rs4784227 disrupts enhancer function through FOXA1 affinity modulation. (**a**–**d**) ChIP-qPCR for H3K4me2 histone modification (enhancer) (**a**) FOXA1 (**b**) and TLE (**c**) binding and H3K9Ac (active enhancer) modification (**d**) at rs4784227. rs4784227[C] is the reference allele, and rs4784227[T] is the risk variant. (**e**,**f**) 3C sequencing results showing the physical interactions between heterozygous SNPs at the enhancer, rs4784227[C] (**e**) and rs4784227[T] (**f**), and different alleles at the rs2193094 SNP in the *TOX3* intron. (**g**) Ratio between RNA and DNA levels for both alleles in the *TOX3* intron. $^{*}P \le 0.05$; $^{**}P \le 0.01$; $^{***}P < 0.0001$. Error bars, s.e.m. from three replicate measures.



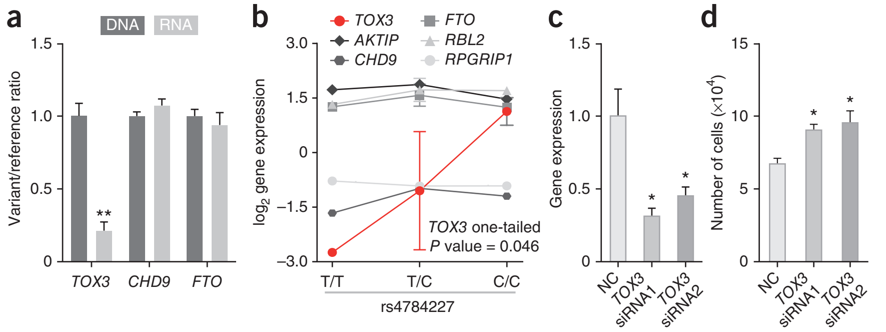


Figure 7 The breast cancer–associated SNP rs4784227 affects cell proliferation by disrupting *TOX3* gene expression. (**a**) Ratio between variant and reference alleles for DNA and RNA levels of each SNP (rs2193094 in *TOX3*; rs35925303 in *CHD9*; rs11646260 in *FTO*). (**b**) Cell line–based eQTL results for rs4784227 and expression of the *TOX3* gene (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 ZR-75-1 cells. NC, siRNA targeting luciferase. (**d**) The number of cells counted 2 d after *TOX3* silencing. $^*P \leq 0.05$; $^{**}P \leq 0.01$. Error bars, s.e.m. from three replicate measures.

$P < 1 \times 10^{-10}$; **Fig. 4f**). Comparison of the IGR scores of the breast cancer risk–associated SNP rs4784227 assessed for TCF4 affinity and the colon cancer risk–associated SNP rs6983267 assessed for FOXA1 affinity showed that the affinity predictions are factor specific (fold changes of ~1, $P > 0.05$; **Fig. 4f**).

Having probed a single risk-associated SNP, we then applied the IGR method to all risk-associated and linked SNP clusters and found that the breast cancer AVS is enriched in affinity-modulating SNPs for FOXA1 (**Supplementary Table 19**). Of the 44 SNP clusters in the breast cancer AVS, 33 contained SNPs mapping to the FOXA1 cistrome or the H3K4me2 epigenome in breast cancer cells. These clusters harbored 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-binding or H3K4me2-modified regions, thereby excluding linked SNPs lacking the ability to disrupt FOXA1 function in breast cancer cells, 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 was based on a Bonferroni correction that took into account all tests across all permutations ($P < 1 \times 10^{-6}$). Out of the 33 clusters mapping to FOXA1-binding or H3K4me2-modified regions, 24 (73%) contained affinity-modulating SNPs ($P = 0.009$; **Fig. 5a**). This finding signifies that more than half of all SNPs that are associated with breast cancer modulate the affinity of chromatin for FOXA1. Genotyping of the SNPs that were predicted to be disruptive of FOXA1 binding in 4 breast cancer cell lines (MCF7, BT474, T47D and ZR-75-1) identified 13 SNPs (including the rs4784227 variant) that were heterozygous in at least 1 cell line. Of these predicted SNPs, 77% (10 of 13) showed a significant and concordant allele-specific binding preference for FOXA1, as 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 show *in vivo* that the disruptive SNPs that were identified through the VSE and IGR methods can promote breast cancer development, we initially focused on the breast cancer–associated SNP rs4784227, which is heterozygous in MCF7 breast cancer cells. This heterozygosity allowed us to determine the allele-specific function of a variant within its natural genomic context. The rs4784227 SNP directly overlaps an FKH motif in a FOXA1-binding site that lies in a regulatory element (that is subject to H3K4me2 modification) 18 kb away from the *TOX3* gene (also known as *TNRC9* and *CAGF9*; **Fig. 6a**,**b**). The IGR method predicted that the rs4784227[T] variant allele would favor FOXA1 binding over the rs4784227[C] reference allele (**Fig. 4f** and **Supplementary Fig. 1**). Accordingly, allele-specific directed ChIP assays *in vivo* confirmed that FOXA1 is preferentially recruited to the rs4784227[T] risk allele (**Fig. 6b**). This change in affinity is independent from changes in H3K4me2 modification—the epigenetic signature favorable to FOXA1 binding—as H3K4me2 levels were equivalent for the two rs4784227 alleles in MCF7 breast cancer cells[4] (**Fig. 6a**). Whereas FOXA1 commonly promotes gene expression, co-binding to chromatin with Groucho/TLE proteins leads to local chromatin condensation and transcriptional repression[25]. The variant rs4784227[T] risk allele associated with increased FOXA1 binding was also significantly bound by Groucho/TLE compared to the rs4784227[C] allele (**Fig. 6c**). Furthermore, acetylation of histone H3 at lysine 9 (H3K9ac, a chromatin signature for active enhancers) was significantly less enriched at the rs4784227[T] allele versus the rs4784227[C] allele[7] (**Fig. 6d**).

Chromatin conformation capture assays (3C) using BglII or MspI restriction enzyme digestion revealed that the region containing rs4784227 physically interacts with the promoter of the *TOX3* gene ($P < 0.04$). Consequently, *TOX3* is a gene target of the regulatory region harboring rs4784227. To assess the impact of the rs4784227 alleles on *TOX3* gene expression, we identified a SNP (rs2193094) in the first intron of *TOX3* that is heterozygous in MCF7 cells. Sequencing the 3C product revealed that the rs4784227[C] reference allele and the rs4784227[T] variant allele are physically linked to the rs2193094[T] and rs2193094[G] alleles, respectively (**Fig. 6e**,**f** and **Supplementary Fig. 2**). Allele-specific expression assays showed that the rs2193094[T] allele is preferentially expressed compared to the rs2193094[G] allele, suggesting that the rs4784227[T] variant allele has a repressive effect on *TOX3* expression (**Fig. 6g**). This finding agrees with the results of previous *in vitro* luciferase reporter assays[21].

RNA sequencing (RNA-seq) data from MCF7 cells revealed that six genes (*TOX3*, *CHD9*, *RBL2*, *AKTIP*, *RPGRIP1L* and *FTO*) are expressed from the ~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. Whereas the rs2193094[G] variant allele was associated with a decrease in *TOX3* RNA levels compared to the rs2193094[T] reference allele, no significant difference in RNA levels was detected for the variant and reference alleles rs35925303 (*CHD9*) or rs11646260 (*FTO*) (**Fig. 7a**). Furthermore, we genotyped the rs4784227 SNP in 11 breast cancer cell lines expressing ESR1 and FOXA1 and extracted the expression data for *TOX3* from the Neve breast cancer cell lines database[1]. Only one cell line, T47D, is homozygous for the rs4784227[T] allele. Correspondingly, T47D had the lowest expression of *TOX3*. Comparing all heterozygous ($n = 3$) cell lines with those that are homozygous for the rs4784227[C] allele ($n = 7$), we observed a significant association between the

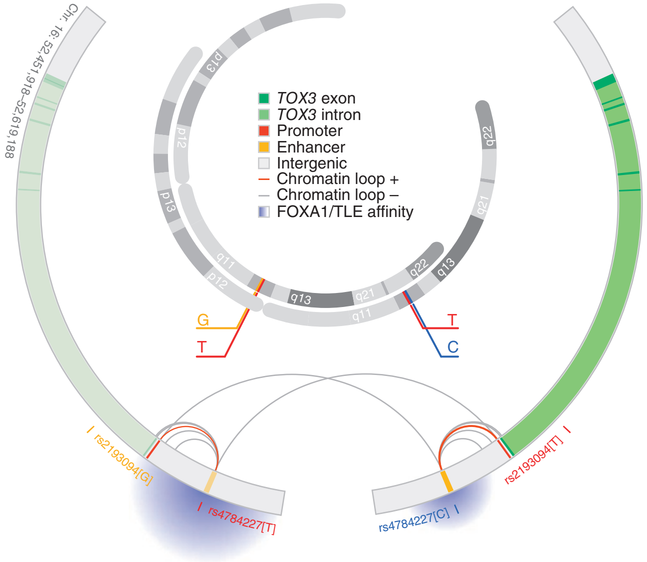


Figure 8 Diagram of the *TOX3* locus and the physical interactions between and within chromosome 16 homologs.

rs4784227[T] allele 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 in a study of 1,401 individuals of European ancestry with breast cancer[26]. rs3803662 is in significant LD with rs4784227 in individuals of European ancestry ($r^2 = 0.864$, $D' = 1$; HapMap Utah residents of Northern and Western European ancestry (CEU)) (**Supplementary Table 20**). In fact, rs4784227[T] only segregates with rs3803662[A], which is also associated with a decrease in *TOX3* gene expression, supporting our conclusion that the functional rs4784227[T] allele is associated with a decrease in the expression of *TOX3*. In addition, we included the remaining five genes (*CHD9*, *RBL2*, *AKTIP*, *RPGRIP1L* and *FTO*) expressed from the ~3.5-Mb window centered on rs4784227 in expression quantitative trait locus (eQTL) assays (**Fig. 7b** and **Supplementary Fig. 3**). rs4784227 was not significantly associated with the expression of these genes. Overall, these data suggest that allele-specific expression is unique to *TOX3* at this locus and is dependent on the genotype at rs4784227.

To assess the role of TOX3 in breast cancer cells, we performed a cell proliferation assay after silencing *TOX3*. We used the ZR-75-1 breast cancer cell line for this analysis because of its high expression of *TOX3*. Silencing of *TOX3* in ZR-75-1 breast cancer cells significantly increased cell proliferation compared to the control treatment (small interfering RNA (siRNA) against luciferase), suggesting that TOX3 functions as a tumor suppressor in breast cancer cells (**Fig. 7c,d**).

DISCUSSION

The mechanisms underlying breast cancer risk–associated SNPs are unknown. As with most other complex traits, these risk-associated SNPs map to the non-coding regions of the genome[2]. Here, we show that breast cancer–associated SNPs are enriched for FOXA1 and ESR1 transcription factor–binding sites and H3K4me1 histone modification. Furthermore, we reveal that enrichment is dependent on factor, cell type and cancer type. The body of evidence supporting regulatory mechanisms for GWAS-identified risk-associated SNPs is steadily growing[27–30]. 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-identified SNPs[31]. Binding of the FOXA1 pioneer factor is central for chromatin opening and nucleosome positioning favorable to transcription factor recruitment[4,11,12,32,33]. In addition, FOXA1 is central to the establishment of the transcriptional programs that respond to estrogen stimulation in ESR1-positive breast cancer cells[4,11,12,32,34,35].

Among the elusive breast cancer–associated SNPs, 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[36–39]. Fine-scale mapping of the 16q12.1 locus has revealed that rs4784227 is the most strongly associated SNP in European and Asian populations[21,40]. In *in vitro* experiments, the rs4784227[T] risk allele reduces luciferase reporter activity and alters DNA-protein binding patterns, suggesting a regulatory role for the rs4784227 SNP[21]. Functionally, TOX3 belongs to the high-mobility-group (HMG)-box family of proteins that modify chromatin structure[41]. It is expressed mainly in epithelial cells and targets both antiapoptotic and proapoptotic transcripts, protecting against cell death[42]. Furthermore, it stimulates estrogen-response element (ERE)-dependent transcriptional programs[42]. Statistically, *TOX3* is differentially expressed in breast cancers that relapse to bone[43]. SNPs contained within the gene interact multiplicatively with mutations in *BRCA1* or *BRCA2*, increasing breast cancer risk[44]. Here, we show that the rs4784227[T] risk allele produces a fivefold 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 (through diminished H3K9ac modification). Mediated through a chromatin loop that interacts with 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, that is, the disruption of transcription factor binding at a distal enhancer and the subsequent change in gene expression, such as reported in CRC, seems to be a common mechanism of action for risk-associated SNPs[22–24].

There is a marked disparity between the number of SNPs associated with all cancers and the number of SNPs (rs6983267, rs4784227 and rs1859962) that have been functionally characterized[45,46]. Therefore, it has been impossible to extrapolate the general mechanisms of genetic risk promotion for any particular disease. Here, we show that, for breast cancer, the majority of risk-associated SNPs modulate FOXA1 binding. First, they are in complete LD with SNPs localized to 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 toward the post-GWAS phase of human genetics[47]. A set of principles has been proposed for the post-GWAS functional characterization of cancer risk loci[45]. But there is still one major hurdle to overcome before this transition can be made. The number of GWAS-identified risk-associated SNPs is large and growing, and each associated SNP has hundreds or thousands of linked SNPs. Furthermore, because the majority of linked SNPs lie in regulatory regions, each of these SNPs has an array of potential gene targets through multiple interactions mediated by chromatin loops. This branching morass makes the exhaustive functional characterization of all GWAS results unfeasible. A mechanism is required that can prioritize this overabundance 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 modifications or other genomic annotations to interrogate (VSE). Second, they identify candidate SNPs among the thousands of linked SNPs for experimental follow-up studies (IGR). This integration of functional genomics data will allow the post-GWAS characterization of risk loci to gain traction and advance.

URLs. NHGRI GWAS Catalog (accessed January 2012), http://www.genome.gov/gwastudies/; Nuclear Receptor Cistrome project, http://cistrome.dfci.harvard.edu/NR_Cistrome/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. ChIP-seq files generated for this manuscript are available at the Gene Expression Omnibus (GEO) under accession GSE31151.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Brown, M.L. Freedman and M. Lemaire for discussions, as well as B.G. Neel and R. Marcotte for technical assistance. We acknowledge support from the US National Cancer Institute (NCI; 2P30CA023108-32), the American Cancer Society (ACS; IRG-82-003-27 to M.L.), the US National Institutes of Health (NIH; R01LM009012 to J.H.M. and R01CA155004 to M.L.) and the Princess Margaret Hospital Foundation (M.L.).

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.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2416.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
2. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
4. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
5. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell type–specific gene expression. *Nature* **459**, 108–112 (2009).
6. Khalil, A.M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
7. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
8. Zentner, G.E., Tesar, P.J. & Scacheri, P.C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21**, 1273–1283 (2011).
9. Lupien, M. & Brown, M. Cistromics of hormone-dependent cancer. *Endocr. Relat. Cancer* **16**, 381–389 (2009).
10. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20**, 578–588 (2010).
11. Carroll, J.S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122**, 33–43 (2005).
12. Carroll, J.S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289–1297 (2006).
13. Krum, S.A. *et al.* Unique ERα cistromes control cell type–specific gene regulation. *Mol. Endocrinol.* **22**, 2393–2406 (2008).
14. Wang, Q. *et al.* Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138**, 245–256 (2009).
15. Grant, S.F. & Hakonarson, H. Microarray technology and applications in the arena of genome-wide association. *Clin. Chem.* **54**, 1116–1124 (2008).
16. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
17. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
18. Johnson, W.E. *et al.* Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* **103**, 12457–12462 (2006).
19. Akhtar-Zaidi, B. *et al.* Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736–739 (2012).
20. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet.* **27**, 465–474 (2011).
21. 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 Genet.* **6**, e1001002 (2010).
22. Pomerantz, M.M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with *MYC* in colorectal cancer. *Nat. Genet.* **41**, 882–884 (2009).
23. Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* **41**, 885–890 (2009).
24. Wright, J.B., Brown, S.J. & Cole, M.D. Upregulation of c-*MYC* in *cis* through a large chromatin loop linked to a cancer risk–associated single-nucleotide polymorphism in colorectal cancer cells. *Mol. Cell. Biol.* **30**, 1411–1420 (2010).
25. Sekiya, T. & Zaret, K. Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates compacted chromatin *in vitro* and impairs activator binding *in vivo*. *Mol. Cell* **28**, 291–303 (2007).
26. 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.* **133**, 843–851 (2012).
27. De Gobbi, M. *et al.* A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**, 1215–1217 (2006).
28. Gaulton, K.J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
29. Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.* **5**, e1000597 (2009).
30. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
31. Reddy, T.E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22**, 860–869 (2012).
32. Eeckhoute, J., Carroll, J.S., Geistlinger, T.R., Torres-Arzayus, M.I. & Brown, M. A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes Dev.* **20**, 2513–2526 (2006).
33. He, H.H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* **42**, 343–347 (2010).
34. Laganière, J. *et al.* Location analysis of estrogen receptor α target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proc. Natl. Acad. Sci. USA* **102**, 11651–11656 (2005).
35. Hurtado, A., Holmes, K.A., Ross-Innes, C.S., Schmidt, D. & Carroll, J.S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet.* **43**, 27–33 (2011).
36. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
37. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).
38. Ruiz-Narváez, E.A. *et al.* Polymorphisms in the *TOX3*/*LOC643714* locus and risk of breast cancer in African-American women. *Cancer Epidemiol. Biomarkers Prev.* **19**, 1320–1327 (2010).
39. Hutter, C.M. *et al.* Replication of breast cancer GWAS susceptibility loci in the Women's Health Initiative African American SHARe Study. *Cancer Epidemiol. Biomarkers Prev.* **20**, 1950–1959 (2011).
40. Udler, M.S. *et al.* Fine scale mapping of the breast cancer 16q12 locus. *Hum. Mol. Genet.* **19**, 2507–2515 (2010).
41. O'Flaherty, E. & Kaye, J. TOX defines a conserved subfamily of HMG-box proteins. *BMC Genomics* **4**, 13 (2003).
42. 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. *J. Cell Sci.* **124**, 252–260 (2011).
43. Smid, M. *et al.* Genes associated with breast cancer metastatic to bone. *J. Clin. Oncol.* **24**, 2261–2267 (2006).
44. Antoniou, A.C. *et al.* Common breast cancer–predisposition alleles are associated with breast cancer risk in *BRCA1* and *BRCA2* mutation carriers. *Am. J. Hum. Genet.* **82**, 937–948 (2008).
45. Freedman, M.L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.* **43**, 513–518 (2011).
46. 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 Res.* **22**, 1437–1446 (2012).
47. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).

ONLINE METHODS

Computational methods. *Variant set enrichment (VSE).* 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. SNPs are the genetic markers used in GWAS in the search for the variation underlying genetic traits and diseases. GWAS-identified risk-associated SNPs are the focus of this study and were the constituents of the variant sets tested for enrichment. SNPs contained within the haplotype block of a risk-associated SNP are referred to as its linked SNPs (in LD). Each individual risk-associated SNP and its collection of linked SNPs make up a 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 in the VSE pipeline is to obtain all risk-associated SNPs for a disease of interest from the GWAS catalog (see URLs). We then identify the list of all SNPs in strong LD with each risk-associated SNP by using HapMap project data (UCSC Genome Browser's CEPH HapMap Linkage Disequilibrium Phase 2 table, hg18, HapMapLdPhCeu). We used a high-stringency LD threshold based on LOD >2 and D' of >0.99. Haplotype blocks are heterogeneous in size and porous, meaning that strong LD values are not contiguous across the block but 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 clusters of directly associated and linked SNPs in the AVS that have at least one linked SNP that overlaps the genomic annotation. The intersectBed program from the BEDTools suite is used to compute the overlap between chromosomal coordinates, with each cluster intersected independently. However, 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. To correct for these factors, we build a null distribution for the mapping tally that is based on random permutations of the AVS.

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. Sets of tag SNPs are sampled at random from a comprehensive list of marker variants used in GWAS (Illumina Human OmniExpress). For each randomly sampled tag SNP (matched tag SNP), its set of linked SNPs is imputed from HapMap data. Each set is built so as to be composed of a collection of clusters of matched tag SNPs and their linked SNPs that match the size and structure of the AVS. We refer to these as matched random variant sets (MRVS). For each permutated variant risk-associated SNP in the AVS, a matched tag SNP is chosen at random from a bin of tag SNPs that matches its number of linked SNPs. Matched tag SNPs are sampled with replacement within each bin, and their distribution across bins is checked to ensure that no bin is depleted. All risk-associated SNPs contained in the AVS are removed from the sampling list.

We then derive the null distribution of the mapping tally between an AVS and a genomic annotation by intersecting collections of MRVS with the annotation (**Fig. 1d**–**g**, gray histograms). Specifically, we estimate the probability of finding a set of variants with a mapping tally greater or equal to that of the observed AVS by chance alone. 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 deviations 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 distribution's range. This is calculated as the ratio between the number of MRVS that have a value greater than or equal to the AVS and the total number of MRVS. 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 distribution's range, it becomes increasingly imprecise, and the *P* value is smaller than the inverse of the number of MRVS in the null. An exact *P* value requires fitting a density function to the null distribution derived from the MRVS. The Kolmogorov-Smirnov test is used to ensure 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 distribution, the test is considered nonsignificant. Exact *P* values are then calculated from the density function. All methods are implemented in the R statistical environment. Functions and packages used include box.cox(car), ks.test (stats) and pnorm(stats).

VSE for non-disjoint clusters of risk-associated and linked SNPs. Most of the AVSs considered in this study have one or more risk-associated SNPs with linked SNPs in common. When this is the case, we use a special scoring procedure for VSE to ensure that directly associated SNPs are not counted more than once when only one linked SNP maps to a functional genomic element, as this would result in inflated mapping tallies and could lead to false positives. We created the LDXI program (linkage disequilibrium extension with intersections) to compute all intersections within the set of risk-associated SNPs. The linked SNPs that are shared by one or more associated SNPs become separate clusters. We call this the 'intersection' class of clusters, and it is treated separately when calculating the mapping tally. The rest of the linked SNPs are grouped into clusters with risk-associated SNPs 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. All unique clusters are intersected and tallied as in the standard VSE. All risk-associated SNPs in each intersection cluster that map to the genomic annotation are searched for in the unique tally. If all associated SNPs in the intersection cluster are accounted for, the cluster is discarded, but, if at least one of the associated SNPs in the cluster is not present in the unique tally, the cluster is counted once toward the mapping tally. This counting mechanism ensures that each associated SNP 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 associated SNP.

Intragenomic replicates (IGR). IGR is a computational method that calculates the modulation in affinity produced by a SNP at a transcription factor–binding site or other quantitative changes for any continuous genomic annotation. 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 instances of the *k*-mer are searched across the genome. 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 depending 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 affect 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 1,000 instances to discard unreliable observations on the basis of their signal-to-noise ratios within each node in the lattice (**Fig. 3c**). Furthermore, among possible values of *k*, eight provided a good tradeoff between the number of instances and sequence specificity. The two collections of 8-mer affinity models are filtered on the basis of the 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 on the basis of the completeness of the assessment. The comparison between maxima represents the highest possible affinity for each allele given its genomic context. Furthermore, we 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 DNase I 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 inflating our results. *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 subset of linked SNPs contained in the regulatory regions where binding sites can be created or destroyed; we refer to these as regulatory SNPs. We then calculate the modulator quotient. The numerator of this quotient is the number of clusters of risk-associated and regulatory SNPs that have at least one significant affinity-modulating regulatory SNP. The denominator is the number of clusters of risk-associated and regulatory SNPs in the AVS. Risk-associated SNPs that do not contain regulatory SNPs are not considered in the modulator quotient. The modulator quotient null distribution over the set of MRVS 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*. MCF7 breast cancer cells were treated with MNase (Sigma, N3755), and ChIP assays were performed with antibody to H3K4me2 (Millipore, CMA303), as previously published[33]. Libraries were prepared for the Illumina Genome Analyzer according to the manufacturer's instructions. Regions that were significantly ($P < 1 \times 10^{-5}$) enriched for H3K4me2 were detected using MACS software with default parameters[48]. H3K4me2 enrichment was validated by ChIP-qPCR.

In vivo allele-specific binding assays. ChIP assays were performed as previously described[4]. ChIP assays were performed using antibodies to FoxA1 (Abcam, ab5089), H3K4me2 (Millipore, 07-030), TLE (Santa Cruz Biotechnology, sc-13373X) and H3K9Ac (Abcam, ab4441). The allele-specific MAMA PCR-based genotyping technique was applied to assess differential ChIP enrichment on heterozygous alleles, as described previously[46,49].

3C-qPCR assays. Chromosome conformation capture (3C) technology coupled to qPCR was performed according to a published protocol[50]. Briefly, 5×10^6 MCF7 cells were subjected to formaldehyde cross-linking (1% formaldehyde, 10 min at room temperature) of interacting chromatin segments. This step was followed by BglII (or MspI, for higher resolution) digestion (400 U, overnight at 37 °C) and ligation (T4 DNA ligase, 4,000 U, 4 h at 16 °C). DNA fragments were cleaned by phenol-chloroform and ethanol extraction. qPCR was performed to quantify ligated DNA fragments. BAC clones were used to verify primer efficiency and to normalize the 3C interaction frequency.

In vivo allele-specific gene expression. A primer binding to a sequence outside of the rs2193094 SNP and its closest MspI restriction enzyme site and a primer binding to a sequence 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 were then sequenced using the Sanger sequencing approach, thereby revealing the linkage between the two alleles of the rs2193094 and rs4784227 SNPs. MCF7 genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit. Total nuclear RNA from MCF7 cells was extracted using TRIzol (Invitrogen), and reverse transcription was performed 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 amount of DNA and RNA encoding each allele of the rs2193094, rs35925303 and rs11646260 SNPs.

Cell line–based eQTL assays. The rs4784227 SNP was genotyped in 11 ERα-positive and FOXA1-positive 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). Gene expression data for *TOX3*, *AKTIP*, *CHD9*, *FTO*, *RBL2* and *RPGRIP1* (generated using probes 214774_x_at, 218373_at, 212615_at, 209702_at, 212331_at and 206608_at, respectively) in these cell lines were extracted from the Neve breast cancer cell lines database[51]. SNP genotype data were then correlated with the expression levels of each of these genes.

Cell proliferation assays after TOX3 silencing. *TOX3* was silenced by siRNA duplexes. Two sets of siRNA against *TOX3* were designed through Dharmacon (**Supplementary Table 21**). siRNA against luciferase was used as a negative control. RNA was isolated from ZR-75-1 breast cancer cells using the RNeasy Mini kit (Qiagen) according to the manufacturer's recommendations. Quantitative RT-PCR was performed as previously described[4] to confirm the silencing effect of the siRNA against *TOX3*. *TOX3* RNA levels were normalized to those of 28S ribosomal RNA and were then normalized by *TOX3* RNA levels in cells that received negative control siRNA targeting luciferase. The number of ZR-75-1 breast cancer cells was counted 2 d after *TOX3* silencing. ZR-75-1 cells that received siRNA targeting luciferase were considered to be a control.

48. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
49. Li, B., Kadura, I., Fu, D.J. & Watson, D.E. Genotyping with TaqMAMA. *Genomics* **83**, 311–320 (2004).
50. Hagège, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.* **2**, 1722–1733 (2007).
51. Neve, R.M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).

doi:10.1038/ng.2146