nature genetics

Regulators of genetic risk of breast cancer identified by integrative network analysis

Mauro A A Castro[1], Ines de Santiago[2,3], Thomas M Campbell[2,3], Courtney Vaughn[2,3], Theresa E Hickey[4], Edith Ross[2], Wayne D Tilley[4], Florian Markowetz[2], Bruce A J Ponder[2,3] & Kerstin B Meyer[2,3]

Genetic risk for breast cancer is conferred by a combination of multiple variants of small effect. To better understand how risk loci might combine, we examined whether risk-associated genes share regulatory mechanisms. We created a breast cancer gene regulatory network comprising transcription factors and groups of putative target genes (regulons) and asked whether specific regulons are enriched for genes associated with risk loci via expression quantitative trait loci (eQTLs). We identified 36 overlapping regulons that were enriched for risk loci and formed a distinct cluster within the network, suggesting shared biology. The risk transcription factors driving these regulons are frequently mutated in cancer and lie in two opposing subgroups, which relate to estrogen receptor (ER)$^+$ luminal A or luminal B and ER$^-$ basal-like cancers and to different luminal epithelial cell populations in the adult mammary gland. Our network approach provides a foundation for determining the regulatory circuits governing breast cancer, to identify targets for intervention, and is transferable to other disease settings.

Polygenic disease susceptibility results in a distribution of risk within the population. Given the large number of known risk loci in such diseases, there are a huge number of possible combinations of genotypes associated with high risk. Therefore, in parallel with the ongoing analysis of individual loci, a framework is needed to understand how multiple risk variants can combine at the cellular level and to indicate whether these variants work through many different mechanisms or converge on just a few. The latter scenario would be more tractable for understanding disease biology and for developing intervention strategies. Germline variants will interact not only with each other but with exposures and with acquired somatic events. Ideally, the framework should be able to capture these interactions.

Systems biology approaches may be able to provide such a framework[1]. Protein-protein interaction networks have been derived in attempts to shed light on the pathways underlying risk[2], but most of these networks remain sparse and have only yielded limited insight into cancer risk. Most germline risk variants are thought to affect gene expression. Therefore, regulatory networks may be an appropriate starting point to understand the combinatorial effect of risk variants.

Here we model breast cancer as such a gene regulatory network[3], onto which the loci relating to risk can be mapped to identify key regulators[4]. We extend our previous analysis[4] to map onto the network all genes that are associated with the known breast cancer loci derived from genome-wide association studies (GWAS)[5]. We found that transcription factors regulating the genes linked to risk loci cluster within the network, suggesting potential commonality of mechanisms. We also show that the same transcription factors are frequently mutated in breast cancer. Our analysis provides insight into the gene regulatory circuits operating in breast cancer and has implications for treatment and for the identification of novel therapeutic targets. The approach can be applied in any other settings where data from GWAS, large-scale genotyping and gene expression analyses are available.

RESULTS

Mapping of breast cancer risk loci to regulatory networks

Briefly, our analysis builds a regulatory network and then asks for each regulon in the network whether the genes within it are linked to more risk loci than would be expected by chance. In a subsequent step, we examine whether the risk-associated regulons, and the transcription factors driving them, cluster in the overall network.

First, we created a regulatory network for breast cancer using ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks)[3,4], which defines regulons (possible target genes) for a set of curated transcription factors. The regulon for each transcription factor is composed of all the genes whose expression data display significant mutual information with that of a given transcription factor and are therefore likely to be regulated by that transcription factor. We previously validated the functional importance of these regulons using chromatin immunoprecipitation and sequencing (ChIP-seq) data and transcription factor knockdown studies[4]. Regulatory networks were inferred using separate analyses on gene expression data from METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohort I ($n = 997$) and cohort II ($n = 995$)[6]. Within each network, regulons overlap because many genes are regulated by more than one transcription factor. We confirmed that copy

[1]Bioinformatics and Systems Biology Laboratory, Federal University of Paraná (UFPR), Polytechnic Center, Curitiba, Brazil. [2]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK. [3]Department of Oncology, University of Cambridge, Hutchison/Medical Research Council (MRC) Research Centre, Cambridge, UK. [4]Dame Roma Mitchell Cancer Research Laboratories, School of Medicine, University of Adelaide, Adelaide, South Australia, Australia. Correspondence should be addressed to K.B.M. (kerstin.meyer@cruk.cam.ac.uk).

Received 23 July; accepted 6 November; published online 30 November 2015; doi:10.1038/ng.3458

number variation does not substantially influence the network structure (**Supplementary Fig. 1** and **Supplementary Note**).

Second, we identified regulons enriched for genes associated with risk loci using EVSE (**e**QTL-conditioned **v**ariant **s**et **e**nrichment)[4]. GWAS identify risk loci, marked by tagging SNPs that may themselves not be causative. Therefore, each tagging SNP was expanded into an associated variant set (AVS)[7] that includes all SNPs in strong linkage disequilibrium (Online Methods). We then used variation in gene expression to determine which risk loci can be assigned to a given regulon using eQTLs[4] (SNPs where allelic differences determine expression of a target gene). We used a multivariate eQTL analysis to test the association between the genotypes of the SNPs in each AVS and, for each regulon separately, the expression of all the genes that lay within a ±250-kb window around the AVS. If such an association was found, the locus was counted toward a mapping tally of the number of GWAS loci associated with genes in the regulon. Finally, the statistical significance of the mapping tally was assessed by permutation analysis (Online Methods and **Supplementary Fig. 2**). We refer to transcription factors whose regulons were significantly enriched for risk loci as 'risk TFs'.

We carried out the EVSE analysis independently for cohorts I and II of the METABRIC cancer data set and identified 63 and 61 transcription factors, respectively, with significant enrichment scores, but we identified none using the much smaller data set from normal tissue (**Supplementary Fig. 3**). Frequently, a single risk locus contributes to the mapping tally of many regulons. This overlap can be driven by a single gene that is part of many regulons or by multiple distinct genes present at that locus contributing to the association with different regulons (**Supplementary Fig. 4** and **Supplementary Note**). The regulons for 36 transcription factors were significant in both cohorts (**Fig. 1a**,**b**).

Validation of the risk TFs

To gain confidence in the identification of the 36 risk TFs, we tested the effects of changing the input GWAS data or regulons on the resultant enrichment scores. The red box plots in **Figure 1c** show the average enrichment score for each of the 36 risk TFs using eQTLs and regulons from METABRIC. When replacing the breast cancer GWAS data, we found that GWAS hits for bone mineral density (BMD) and chronic lymphocytic leukemia (CLL) and random SNPs did not give significant enrichment scores (**Fig. 1c**, blue box plots). For prostate cancer GWAS loci, the scores obtained were lower than with the breast cancer data but still significant, probably reflecting similarities in these two hormone-driven cancers[8]. When we replaced the regulons calculated from METABRIC data with random regulons of similar size (**Fig. 1c**, gray box plots), none of the associations were significant. These results support the specificity and validity of the EVSE analysis. Our results were not confounded by population stratification (**Supplementary Fig. 5** and **Supplementary Note**). We did not find enrichment when using normal breast samples from METABRIC to calculate eQTLs (**Fig. 1c**, white box plots). This is possibly surprising, as one might expect inherited risk to be expressed in normal tissue. However, eQTL discovery is dependent on sample size[9], and only 144 normal tissue samples were available in this data set.

Comparison of ARACNe-EVSE analysis to other methods

We compared our analysis to alternative methods for derivation of network structure and expansion of tagging SNPs into AVSs and obtained very similar results (**Supplementary Figs. 6–8** and **Supplementary Note**). We also compared our EVSE algorithm to analyses in which the multivariate eQTL step was replaced by distance-based gene selection or by the use of 'predefined' eQTLs[10] from the same sample set (**Supplementary Figs. 9–12** and **Supplementary Note**). EVSE identified more risk TFs and showed better reproducibility than the other methods tested.

Risk TFs are frequently mutated in breast cancers

To ask whether somatic and germline variation are associated with the same regions of the network, we examined the frequency of mutations and/or copy number changes affecting transcription factor genes in data from The Cancer Genome Atlas (TCGA)[11]. Collectively, our 36 risk TFs have a significantly (empirical $P < 0.0001$) increased frequency of alterations in comparison to random genes (**Fig. 1d** and **Supplementary Table 1**) and are mutated at a similar frequency as annotated cancer-related genes for which mutations have been causally implicated in cancer[12].

Confirmation of risk association using ChIP-seq data

To validate that our risk TFs are indeed associated with the regulation of GWAS loci, we examined ChIP-seq data[13] that were generated for transcription factor–eGFP fusion proteins whose expression was driven from endogenous sequences in MCF-7 breast cancer cells. We used these data in a variant set enrichment (VSE) analysis[7] to test whether risk TF binding sites are enriched at risk SNPs. Our analysis correlated the position of transcription factor binding sites with risk AVSs. ChIP-seq data were available for nine of our 36 risk TFs and were compared to data for nine low-risk TFs chosen from the EVSE analysis. Five of the nine high-risk transcription factors but none of the low-risk transcription factors (**Fig. 2a**,**b**) yielded a significant enrichment score for risk association. The signal in this analysis is likely to be relatively low, as fusion proteins rather than the native transcription factors were assayed. When we used ChIP-seq data obtained with antibodies to FOXA1 and ESR1, much higher enrichment scores were obtained (**Fig. 2c**), corroborating previous results[7]. CEBPβ binding was also enriched at breast cancer risk loci (**Fig. 2c**). Some of the transcription factors, such as androgen receptor (AR) and PPARδ, are expressed at very low levels in MCF-7 cells. We therefore tested whether AR-binding sites were significantly enriched for GWAS hits in the MDA-MB-453 cell line, which belongs to the molecular apocrine subclass[14] and expresses high levels of AR. After AR activation, AR targets yield significant enrichment scores in this cell line (**Fig. 2d**). Collectively, the ChIP-seq experiments strongly support our conclusion that the risk TFs have a role in regulating transcription at risk-associated SNPs.

Confirmation of risk association by master regulator analysis

Estrogen and fibroblast growth factor receptor 2 (FGFR2) signaling pathways are known to be associated with breast cancer risk. We examined differential gene expression in response to estrogen and FGFR2 signaling in three ER^+ breast cancer cell lines: MCF-7, T-47D and ZR-75-1 cells. Using master regulator analysis (MRA)[15] (Online Methods), we identified master regulators consistently associated with these responses (**Supplementary Table 2** and **Supplementary Note**) and found a high prevalence of risk TFs among the master regulators, providing further support that our risk TFs are indeed functionally related to breast cancer risk.

Clustering of risk TFs and clues to function

To examine whether the different risk TFs converge on common mechanisms, we used ARACNe to calculate the breast cancer regulatory network and mapped onto this network the *P* values for risk association (shown in orange to red in **Fig. 3**) using METABRIC cohort I data. The network was visualized by the degree of overlap of regulons (**Fig. 3**

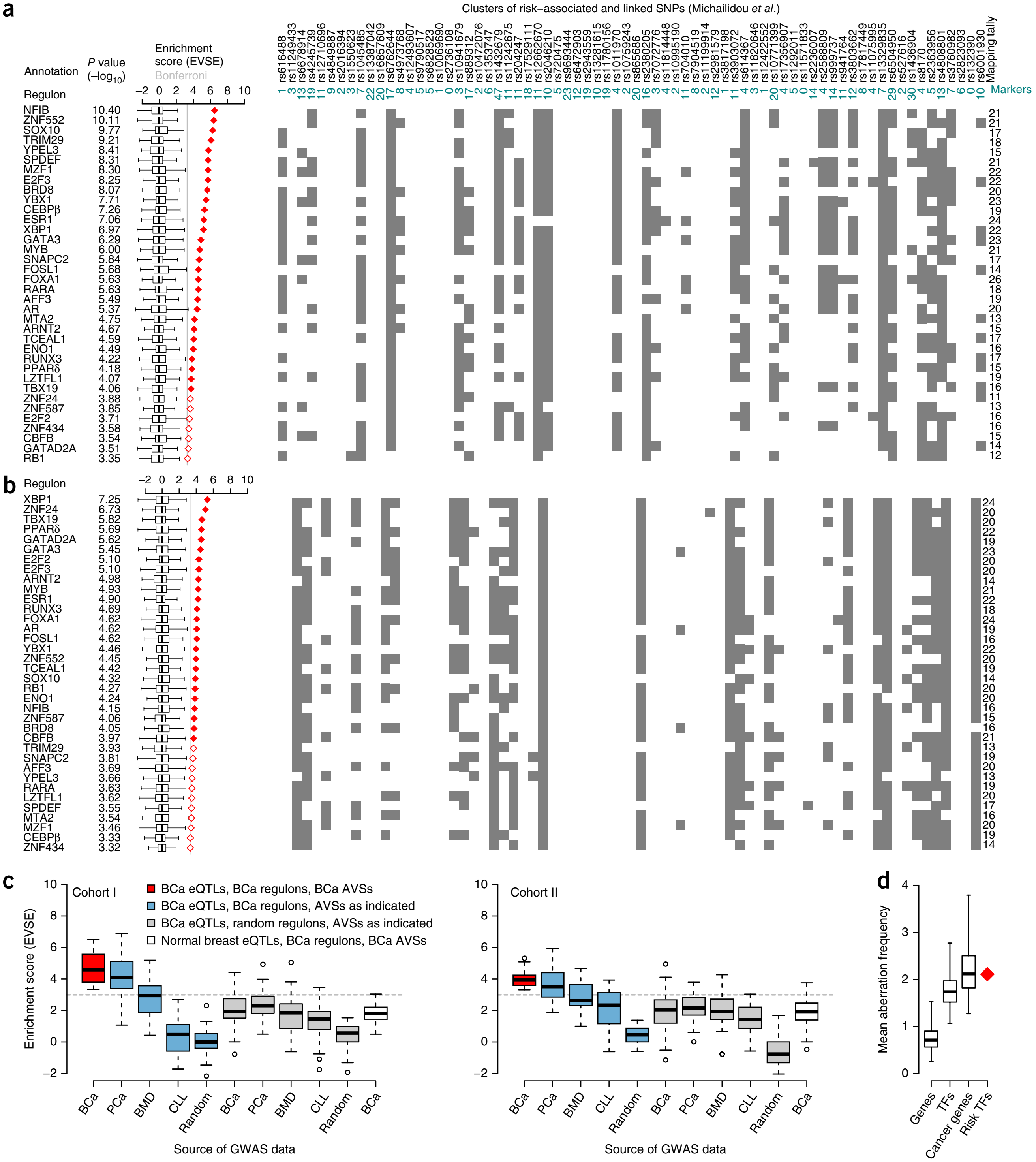


Figure 1 EVSE-based identification of 36 risk TFs. (**a**,**b**) Lists of 36 transcription factor regulons identified in the EVSE analysis, showing the enrichment score– and *P* value–based rank order of risk TFs for METABRIC cohort I (**a**) and cohort II (**b**). The tagging SNP for each breast cancer GWAS hit[5] is listed above, together with the number of markers (SNPs in the AVS for which genotypes were available in METABRIC) for each locus. The matrix shows each multivariate eQTL test with a significant result as a gray box. Mapping tallies are summed on the right of the matrix. Box plots show the normalized null distributions of the enrichment scores (box, first-to-third quartiles; bars, extremes). Solid and open red diamonds highlight enrichment scores that satisfy a Bonferroni-corrected threshold for significance of $P < 0.01$ and $P < 0.05$, respectively. *P* values are based on null distributions from 1,000 random AVSs. (**c**) Computational validation of the EVSE analysis for cohorts I and II. Averages of enrichment scores obtained in the EVSE analysis using different GWAS data sets (breast cancer (BCa), prostate cancer (PCa), bone mineral density (BMD) and chronic lymphocytic leukemia (CLL)) or random SNPs are shown along the *x* axis, using different regulons and eQTLs (origin indicated by color). The gray dashed line highlights the Bonferroni-corrected significance threshold ($P < 0.05$). (**d**) Mean aberration frequency of the 36 risk TFs in comparison to sets of 36 random genes (empirical $P < 0.001$; box-plot whiskers extend to the 1st and 99th percentiles of the random distribution with 10,000 random sets). Aberration frequencies for sets of random transcription factors (TFs) and cancer genes[12] are also shown.

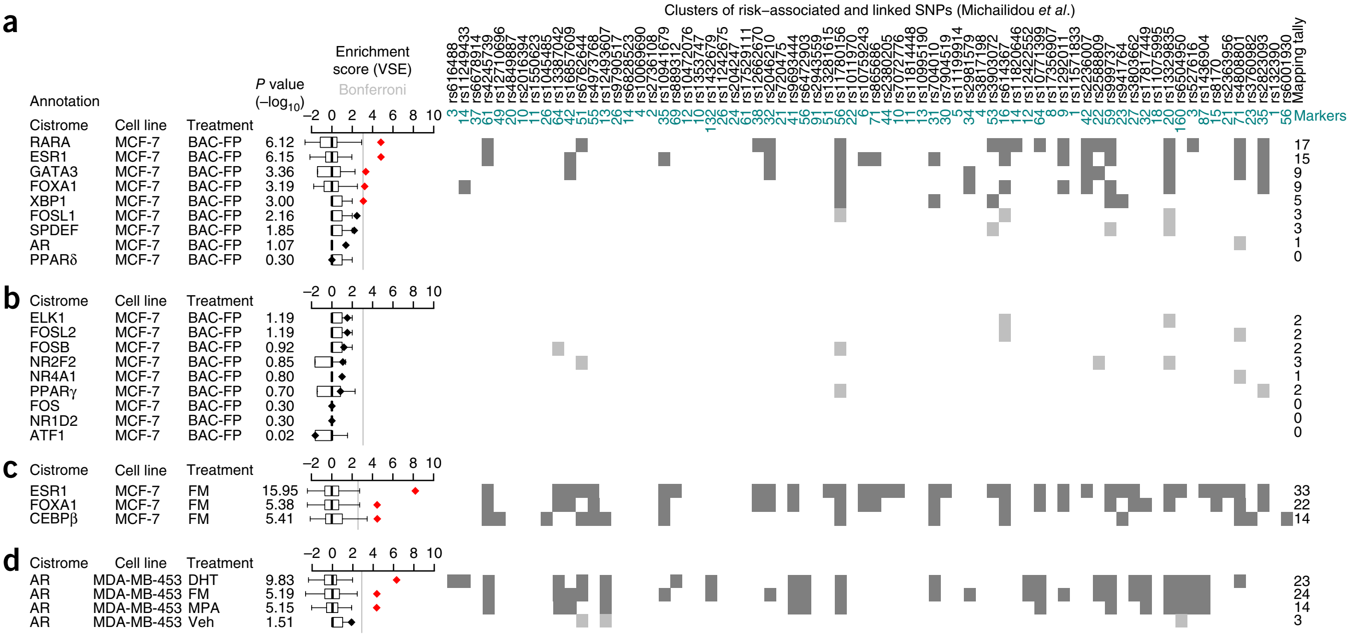


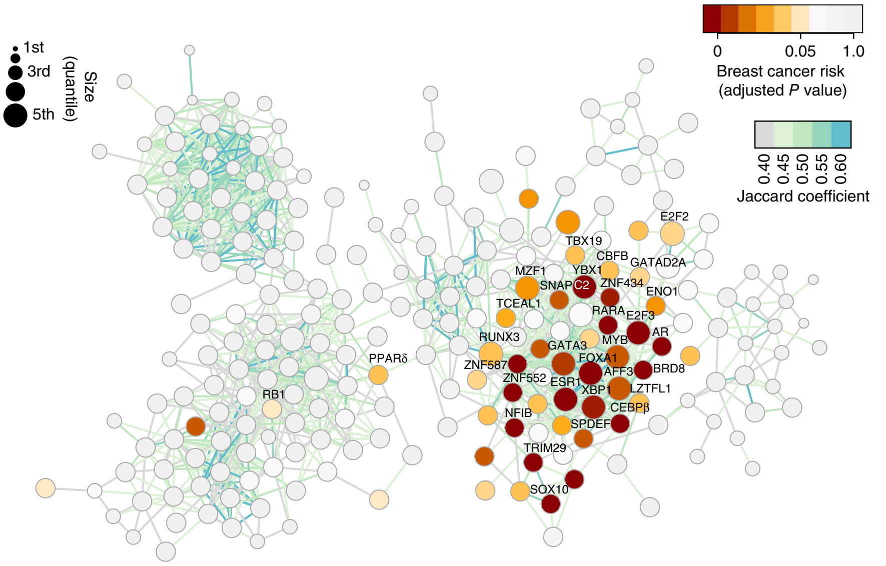
Figure 2 Enrichment of risk TF binding sites at breast cancer GWAS loci. (**a**,**b**) VSE analysis[7] of the cistromes of nine risk TFs (**a**) and nine non-risk TFs (**b**), defined in the EVSE analysis, for which ChIP-seq data were available[13] in MCF-7 cells. Cells were transfected to express BAC fusion proteins (BAC-FP) of the relevant transcription factor and eGFP and grown in full medium. Antibody against eGFP was used in the ChIP experiments. VSE tallies that yielded a significant enrichment score are shown in dark gray, and those that did not are shown in light gray. (**c**) VSE analysis of ChIP-seq experiments using antibodies to ESR1, FOXA1 and CEBPβ in MCF-7 cells. Cells were grown in full medium (FM). (**d**) VSE analysis of ChIP-seq data for AR using the molecular apocrine cell line MDA-MB-453 stimulated as indicated with DHT (5α–dihydrotestosterone), MPA (medroxyprogesterone acetate) or vehicle (Veh) or grown in full medium. Box plots show the normalized null distributions (box, first-to-third quartiles; bars, extremes). Diamonds show the corresponding VSE scores, either in black or in red, for mapping tallies that satisfy a Bonferroni-corrected threshold for significance ($P < 0.01$). *P* values are based on null distributions from 1,000 random AVSs.

and **Supplementary Fig. 13**). The enriched regulons mostly cluster together, suggesting that the risk TFs share biological function.

To refine the clustering analysis and look for clues to biological function, we extended the RTN package (Online Methods) to include the direction of association for any pair of transcription factors and their target genes, using Pearson correlation. For all pairs of transcription factors with a target gene in common, the correlation values were used to assess whether the transcription factors regulate the shared target gene in the same direction (up- or downregulation) or in different (opposite) directions (**Fig. 4a**–**c**). This analysis was carried out for all transcription factors in our regulatory network, and the correlation heat map was used in unsupervised clustering to generate the dendrogram depicted above the matrix in **Figure 4d**. The positions of the 36 risk TFs are highlighted by the black bars below the dendrogram.

An enlargement of the analysis for just the 36 risk TFs is shown in **Figure 4e**. These risk TFs fall into two distinct groups, with high correlation within each group: gene targets shared by two transcription factors in the same group are regulated in the same direction by both transcription factors, whereas gene targets shared by a transcription

Figure 3 Regulatory network for breast cancer, showing clustering of breast cancer risk. The network is depicted on the basis of the overlap of regulons, with risk association shown in yellow to red (based on data from cohort I of METABRIC). The 36 consensus risk TFs identified in both cohorts are labeled. The coloring of the edges (light green to blue) indicates the overlap as measured by Jaccard coefficient, and the size of the circles corresponds to the size of each regulon. Only regulons with a Jaccard coefficient ≥0.4 are shown in the diagram. All regulons and a heat map depicting the overlap of the regulons of risk TFs are shown in **Supplementary Figure 13**. Breast cancer risk is represented by Bonferroni-adjusted *P* values obtained for each regulon when calculating the enrichment with breast cancer GWAS loci in the EVSE analysis, using cohort I to calculate the network and eQTLs. *P* values are based on null distributions from 1,000 random AVSs.



factor in the first group and a transcription factor in the second group are regulated in opposite directions, suggesting the existence of two distinct regulatory groups of transcription factors each able to oppose the effects of the other. Transcription factors in group 1 are highly expressed in ER^+ tumors, whereas those in group 2 are highly expressed in ER^- tumors (**Fig. 4f**). Bootstrap analysis demonstrated that the split into two distinct groups is extremely stable (**Supplementary Fig. 14**). The behavior of shared gene targets was mirrored in the correlation between the expression of the transcription factors themselves, but with much weaker signals (**Supplementary Fig. 15a–c**). The weaker signals may reflect the difference between the regulatory activity of a transcription factor, which is influenced by post-translational regulation and the presence of interacting factors, and the level of transcription factor expression.

With respect to the five intrinsic breast cancer subtypes, group 1 transcription factors are highly expressed in luminal A and luminal B subclasses, whereas group 2 transcription factors are highly expressed in basal tumors. The remaining two intrinsic subtypes, Her2 and normal-like tumors, showed more heterogeneous gene expression patterns (**Supplementary Fig. 16**). Given this distribution, we tested the enrichment of each regulon for genes upregulated in ER^+ or ER^- tumors using MRA (Online Methods). We split each regulon into activated and repressed targets and found that group 1 positive targets were enriched in the ER^+ gene signature, whereas the negative targets were enriched in the ER^- signature (**Fig. 4e**, bar above the matrix). Group 2 generated the opposite pattern, demonstrating that each group of transcription factors is associated with gene expression changes in both tumor subtypes, but with opposite effects.

Identification of clusters associated with known breast cancer subtypes

The dendrogram generated in **Figure 4d** was used to draw a tree-and-leaf diagram (**Fig. 5a**) representing the 555 transcription factors whose regulons in cohort I were of sufficient size to be analyzed in the EVSE pipeline. Coloring of regulons indicates *P* values for breast cancer risk association. Although some risk TFs occur scattered throughout the diagram, two distinct clusters emerge. Cluster 1 (enlarged in **Fig. 5b**) corresponds closely to group 1 in the previous analysis. These transcription factors include those important for the FGFR2 and estrogen responses and also correlate with the transcription factors highly expressed in the luminal A and luminal B subtypes. Group 2 transcription factors are somewhat more dispersed throughout the tree, but there is clear clustering around the YBX1, CBFB, NFIB, TRIM29 and SOX10 transcription factors, labeled as cluster 2 (**Fig. 5c**). Another branch in this node contains the risk TFs CEBPβ and TBX19.

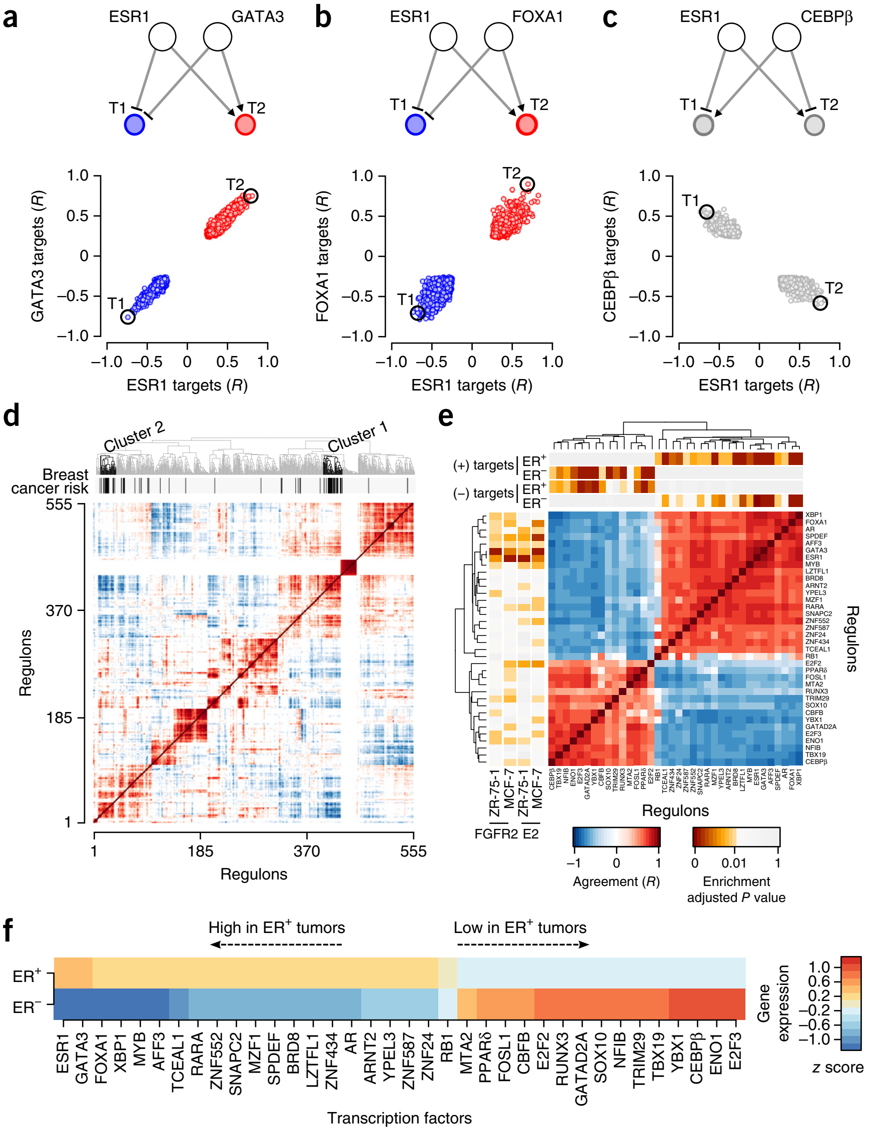


Figure 4 Correlation of expression of targets shared by transcription factor pairs in breast tumors. (**a**–**c**) Correlation (*R*) of gene expression between a given transcription factor and its targets is plotted for three different transcription factor–pairs, including ESR1-GATA3 (**a**), ESR1-FOXA1 (**b**) and ESR1-CEBPβ (**c**). Top, schematics depict the observed interactions. Red circles indicate co-activation, and blue circles indicate co-repression; targets are shown in gray if the two transcription factors have opposing effects on the target. (**d**) Heat map of the correlation of gene expression for targets shared by any pair of the 555 transcription factors (cohort I, METABRIC) whose regulons were of sufficient size to be analyzed in the EVSE pipeline. Unsupervised clustering was applied to this correlation heat map, resulting in the dendrogram shown at the top. The black bars depict the 36 risk TFs, which fall into two distinct clusters. (**e**) Enlargement of the correlation heat map for the risk TFs only. Above the matrix, a bar with yellow-to-red coloring depicts the results (Bonferroni-Hochberg–adjusted *P* values) of MRA for enrichment within each regulon of positive and negative targets that are upregulated in ER^+ and ER^- tumors, respectively, in cohort I of the METABRIC samples. The panel to the left of the matrix shows the master regulators identified for the FGFR2 and E2 responses. (**f**) Relative gene expression levels of the risk TFs in ER^+ and ER^- tumors in samples from cohort I of METABRIC: expression levels were averaged in all ER^+ or all ER^- tumors and compared to expression levels averaged across all samples. The transcription factors are shown ranked by differential gene expression between ER^+ and ER^- tumors.

A literature survey confirmed that the transcription factors in cluster 2 are primarily associated with basal-like breast cancer. We therefore tested whether a gene signature for basal tumors[16] was linked to cluster 2 using MRA. Of the six consensus master regulators for basal-like cancers obtained from the METABRIC cohorts (**Supplementary Table 3**), the two most strongly associated transcription factors map to this cluster

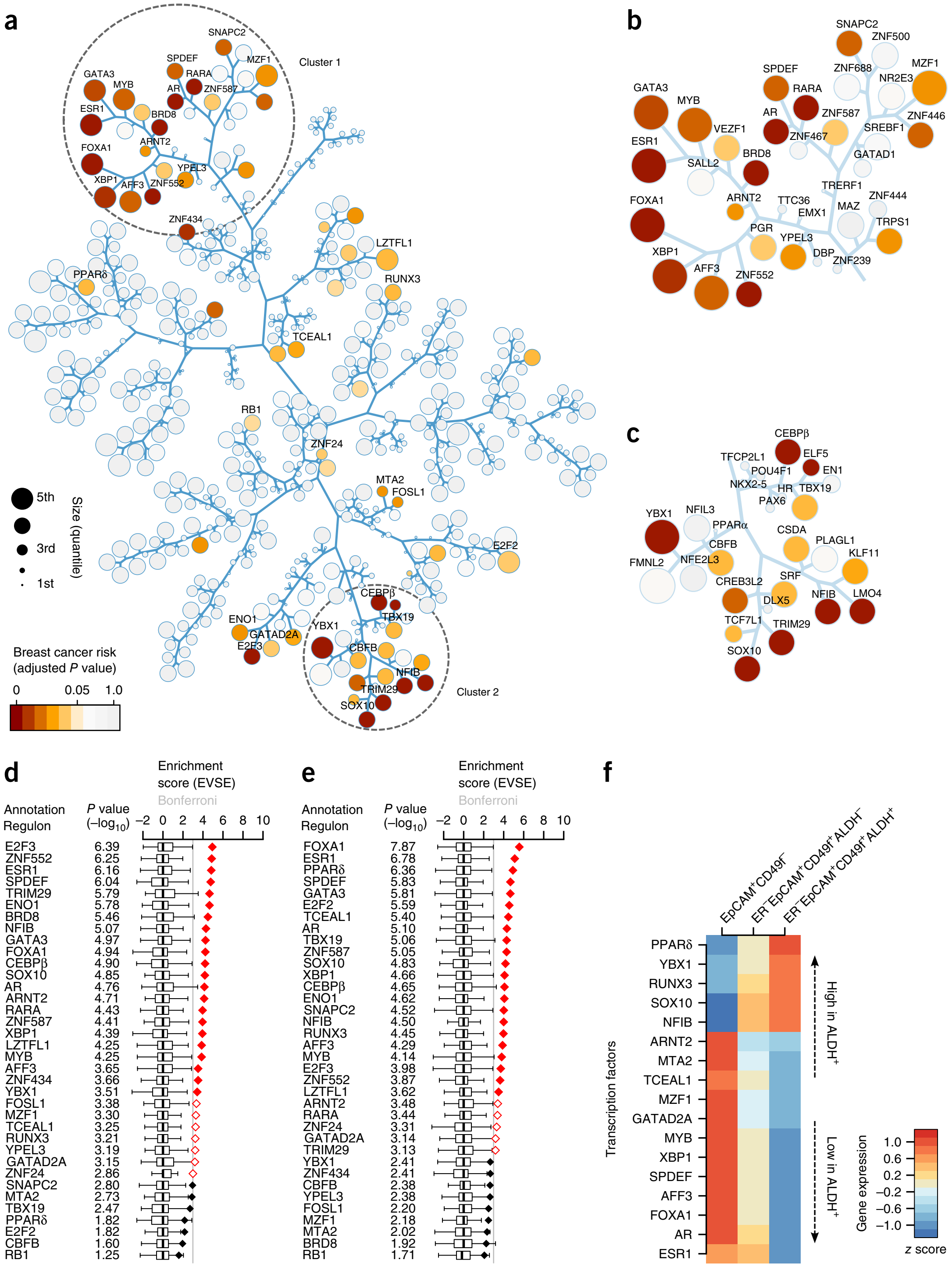


Figure 5 A tree-and-leaf representation of the correlation matrix shows two clusters of risk TFs. (**a**) Tree-and-leaf representation of the dendrogram depicted in **Figure 4d**, where branches represent the arms in the dendrogram. The size of regulons is represented by circle size, and Bonferroni-adjusted *P* values for EVSE enrichment of regulons for breast cancer GWAS loci in cohort I are represented by color. Only consensus risk TFs are labeled. (**b**,**c**) Enlargement of cluster 1 (**b**) and cluster 2 (**c**) of the correlation heat map. All transcription factors present in these clusters are labeled, independently of risk association. (**d**,**e**) EVSE analysis showing enrichment score– and *P* value–based rank order of the 36 risk TFs for cohort I (**d**) and cohort II (**e**) using only ER$^+$ tumors from the METABRIC data set. The mapping tallies are shown in **Supplementary Figure 17**. (**f**) Relative gene expression levels of the risk TFs that were differentially expressed in a comparison of three primary human luminal mammary cell populations[17] ($P < 0.05$, Bonferroni-Hochberg adjusted from limma comparisons; Online Methods). Expression (z score) in each subpopulation is calculated relative to the average in the three populations analyzed and ranked by differential expression between the ALDH$^+$ and ALDH$^-$ cell populations.

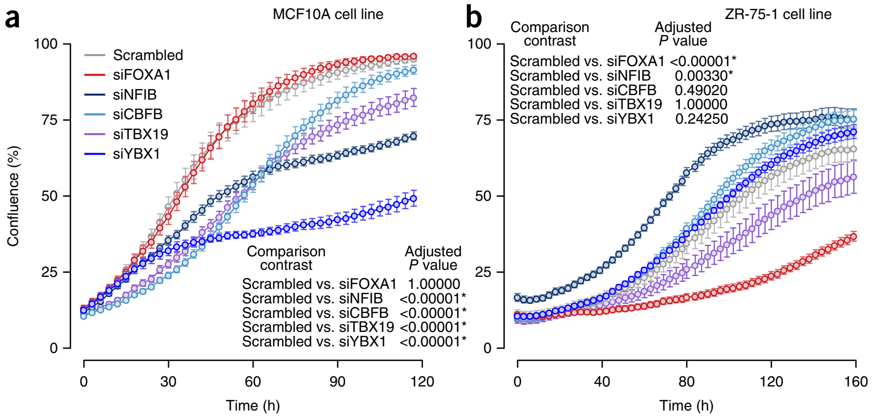


Figure 6 Effects of risk TF knockdown on cell proliferation. (**a**,**b**) Growth curves for the ER$^-$MCF10A cell line (**a**) and the ER$^+$ ZR-75-1 cell line (**b**) after transient transfection with the indicated siRNAs. Cells transfected with a scrambled siRNA were included as a control. Error bars, s.e.m. of eight wells each in a minimum of two independent experiments (Online Methods). The statistical analysis (insets) compares the growth curves using 100,000 simulations, with *P* values adjusted by the Benjamini-Yekutieli correction method (**P* < 0.05).

(SOX10 and TRIM29). PLAGL1 also maps to cluster 2, whereas none of the basal-like cancer master regulators fall within cluster 1.

Given the highly differential expression of these clusters of transcription factors in ER$^+$ and ER$^-$ tumors, we carried out the EVSE analysis separately in ER$^+$ and ER$^-$ tumors. Risk TFs for ER$^+$ tumors map to both cluster 1 and cluster 2 (**Fig. 5d**,**e** and **Supplementary Fig. 17**), reinforcing our previous observation that both groups of risk TFs can have a role in ER$^+$ and ER$^-$ tumors, most likely with opposite effects. Both clusters were also marked by VSE analysis using predefined eQTLs for ER$^+$ tumors or using different network construction tools (**Supplementary Fig. 18** and **Supplementary Table 4**). EVSE analysis with ER$^-$ tumors found very few, non-reproducible risk TFs (data not shown).

Activity of cluster 1 and 2 transcription factors in primary cells

Next, we examined the expression patterns of our risk TFs in primary cell populations isolated from the normal human mammary gland. Gene expression patterns for three luminal cell populations have previously been described[17], including an EpCAM$^+$CD49f$^-$ population highly enriched in ER$^+$ cells that express high levels of luminal cell differentiation markers, ER$^-$EpCAM$^+$CD49f$^+$ALDH$^+$ cells that function as alveolar precursor cells and ER$^-$EpCAM$^+$CD49f$^+$ALDH$^-$ luminal cells that have a phenotype intermediate between those of the EpCAM$^+$CD49f$^-$ and ER$^-$EpCAM$^+$CD49f$^+$ALDH$^+$ subpopulations. The risk TFs that showed differential gene expression across these three populations (adjusted *P* value < 0.05) are listed in **Figure 5f**. Eight cluster 1 transcription factors were overexpressed in the population enriched for ER$^+$ cells, whereas several cluster 2 transcription factors (**Fig. 5f** and **Supplementary Fig. 19a**) were overexpressed in ER$^-$EpCAM$^+$CD49f$^+$ALDH$^+$ alveolar progenitors. The ALDH$^-$ population showed an intermediate pattern. Myoepithelial and stromal cells showed no clear expression pattern for the clusters (**Supplementary Fig. 19b**). The gene expression patterns seen in the ALDH$^+$ population versus the population enriched for primary ER$^+$ cells are reminiscent of those seen in basal-like versus luminal cancers.

Functional analysis of cluster 2 transcription factors

We examined the effect of small interfering RNA (siRNA)-mediated knockdown of cluster 2 risk TFs (NFIB, YBX1, CBFB and TBX19) in ER$^-$ (MCF10A) and ER$^+$ (ZR-75-1) cell lines. In MCF10A cells, siRNA targeting YBX1 strongly reduced proliferation (**Fig. 6a**), and siRNAs targeting CBFB, NFIB, TBX19 and LMO4 (**Fig. 6a** and **Supplementary Fig. 20**) all had a significant antiproliferative effect. In contrast, repression of the cluster 2 transcription factors in ZR-75-1 cells had either no or little effect on proliferation, whereas repression of FOXA1 strongly inhibited growth (**Fig. 6b**). Interestingly, in ZR-75-1 cells, siRNA targeting of NFIB led to a slight but significant increase in proliferation, in keeping with the hypothesis that members of the two clusters have opposing effects.

Although a group of ESR1-cooperating factors is already well defined, our analysis has extended the ESR1 cluster and identified a group of transcription factors opposing ESR1 function, which are likely to be important in regulating basal-like cancers and their precursors.

Regulon activity as prognostic readout

The ESR1 regulon consists of estrogen-induced and estrogen-repressed genes in approximately equal proportions[4]. Our current analysis suggests that the relative activity of these two groups of genes may be important for determining the phenotype of the cell. We therefore devised a two-tailed GSEA (**Fig. 7a**,**b** and Online Methods) in which positive and negative targets of the ESR1 regulon are considered separately to generate a differential enrichment score (dES; Online Methods) representing the activity of the regulon. We used this score in stratified survival analysis of the METABRIC data (**Fig. 7c**,**d**). We found a continuous spectrum of differential enrichment scores across the tumors, except near the transition between the active and repressed states of the ESR1 regulon, which was characterized by an abrupt change. There was a strong trend for better survival with a high differential enrichment score. Interestingly, we identified a set of patients with histochemically ER$^+$ tumors who had a repressed ESR1 regulon and significantly worse outcome than patients with tumors with an active ESR1 regulon (**Fig. 7e**,**f** and **Supplementary Fig. 21**). This difference in survival is not apparent when stratifying by *ESR1* gene expression alone (**Supplementary Fig. 22**). We also tested the effect of tamoxifen treatment on the activity of the ESR1 regulon in MCF-7 cells using two-tailed GSEA. As expected, we found that estrogen induction of steroid-starved MCF-7 cells led to strong activation of the ESR1 regulon (**Fig. 7g**). However, with combined estrogen and tamoxifen treatment, the ESR1 regulon was shifted toward a more repressed state than with estrogen alone (**Fig. 7h**). This finding suggests that tamoxifen, although inhibiting proliferation, might also push luminal tumors to a more basal-like state[18].

DISCUSSION

Our goal was to develop a network-based approach to understand how the effects of multiple GWAS loci combine to influence susceptibility. We derived a transcription factor–centric regulatory network for breast cancer and asked by eQTL analysis which regulons were enriched for association with confirmed breast cancer GWAS loci. We identified 36 regulons that were enriched in both of two separate analyses. The transcription factors controlling these regulons are frequently mutated in breast cancer, implying a convergence of

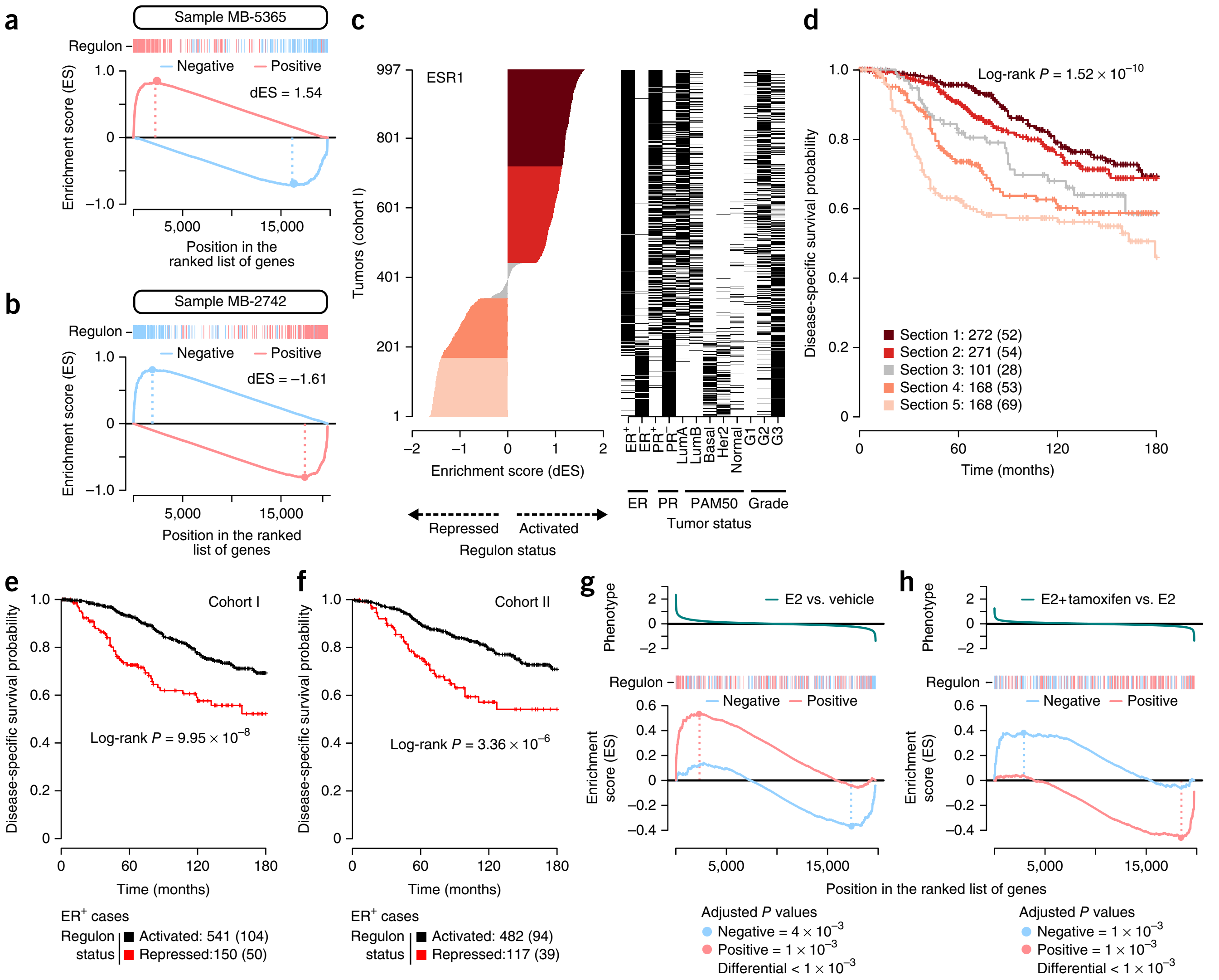


Figure 7 The ESR1 regulon as readout of cell state. (**a**,**b**) Examples of two tumors for which two-tailed GSEA was carried out. The ESR1 regulon is split into targets activated by ESR1 (red bars) and targets repressed by ESR1 (blue bars). GSEA is carried out for each group. Running enrichment scores are shown. Differential enrichment scores (dESs) are obtained by subtracting the maximal deviation from zero for the running enrichment score for repressed targets from that obtained for activated targets. The MB-5365 sample (**a**) represents a tumor whose ESR1 regulon is in an activated state relative to all others of the same cohorts, whereas the MB-2742 sample (**b**) is in a repressed state. (**c**) Differential enrichment scores calculated for all tumors in METABRIC cohort I. Black bars indicate ER status, PAM50 subclass and tumor grade for each of the tumors analyzed. (**d**) Kaplan-Meier survival curves for disease-specific survival for each of the tumor subgroups highlighted in **c**. The number of patients in each section is listed, with the number of patients who died in parentheses. (**e**,**f**) Kaplan-Meier survival curves for immunohistochemically ER$^+$ tumors in cohort I (**e**) and cohort II (**f**) of METABRIC, comparing patients for whom the ESR1 regulon is in an activated state to those with a repressed ESR1 regulon (Online Methods). (**g**,**h**) Two-tailed GSEA in MCF-7 cells activated by estrogen (E2) or estrogen plus tamoxifen. Phenotypes were defined as differential gene expression between estrogen- and vehicle-treated cells (**g**) or between cells treated with estrogen plus tamoxifen and estrogen alone (**h**).

germline and somatic events in the etiology of breast cancer. Many of the risk TFs are master regulators of pathways associated with breast cancer risk, such as estrogen and FGFR2 signaling. Within the regulatory network, almost all of the risk TFs clustered around a group of transcription factors already known to be central to breast cancer risk: ESR1, FOXA1, GATA3 and SPDEF[4,7]. This clustering supports the functional importance of the newly identified risk TFs and suggests that risk TFs share regulatory mechanisms.

The validity of the ARACNe-EVSE analysis was confirmed through extensive comparisons to other methods. The 36 risk TFs identified were specific for hormone-driven cancer and could be validated experimentally. EVSE analysis avoids the multiple-testing problems of unrestrained eQTL calling and was therefore able to identify more risk TFs than other methods. However, as in other analyses[10], we identified eQTLs for only a minority of GWAS loci. Our method used gene expression data from breast tumors. Yet, our hypothesis is that inherited variation exerts its effects on normal tissue and, indeed, on specific cell types within that tissue. To detect such effects, improved, context-specific methods for eQTL identification[19,20] are required. The EVSE analysis we have developed can provide a general approach to interpret GWAS data in the context of regulatory networks.

Consideration of the direction (up or down) of the response of shared target genes led to the identification of two distinct clusters of risk TFs: those in cluster 1, whose positive targets were overexpressed in ER$^+$ cancers, and those in cluster 2, whose positive targets were overexpressed in basal-like, ER$^-$ cancers. However, the inverse also

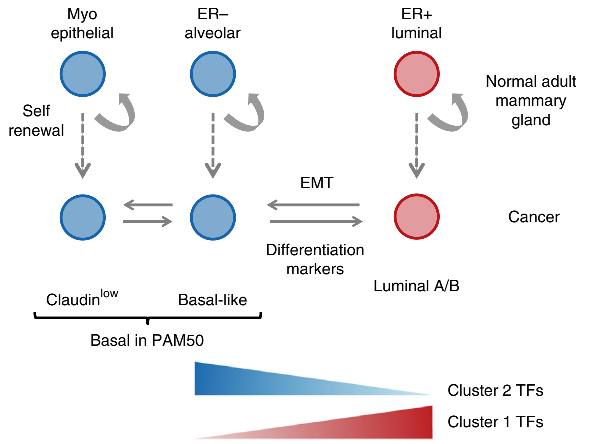


Figure 8 Schematic model of mammary gland cell populations. In this model, we show the predominant expression of cluster 1 versus cluster 2 risk TFs with respect to the cell populations found in the mammary gland and the cancer subtypes that arise from them. In normal mammary gland, all three populations have self-renewal capacity. Claudin^{low} tumors were originally classified as basal in the PAM50 signature but are likely to represent a separate lineage arising from myoepithelial cells[39]. Basal-like cancer is thought to arise from alveolar progenitor cells (the somewhat misleading term ‘basal-like’ reflects the fact that these tumors express not only epithelial but also mesenchymal cell surface markers that are also highly expressed in the myoepithelial lineage located near the basal membrane), and luminal A and luminal B cancer is thought to arise from ER$^+$ precursors.

holds true: cluster 2 transcription factors repress genes associated with ER$^+$ cancers, and cluster 1 transcription factors repress genes associated with ER$^-$ cancers. Therefore, both clusters of transcription factors are likely to be important for the establishment of ER$^+$ and ER$^-$ tumors, albeit with opposing effects. This hypothesis is supported by GWAS results, where the majority of loci confer risk for both ER$^+$ and ER$^-$ disease[21]. Furthermore, our EVSE analysis using only ER$^+$ tumors identified risk TFs from both clusters.

Some cluster 1 transcription factors have previously been reported as critical for ER$^+$ disease[22–24] (ESR1, FOXA1, GATA3 and SPDEF). We confirmed these and added more validated risk TFs: XBP1, RARA and AR. *XBP1* and *ESR1* gene expression is highly correlated in laser-microdissected breast tissue[25], and RARA cooperates with ESR1 to drive estrogen-induced transcription[26]. Recent data suggest that, in ER$^-$ apocrine tumors, expressing high levels of AR, this steroid receptor is able to replace the function of ESR1, leading to a luminal-like gene expression profile[27]. The identification of XBP1, RARA and AR as risk TFs fits the overall framework that estrogen-driven gene expression is the predominant determinant of luminal breast cancer risk.

Cluster 2 comprises YBX1, CBFB, NFIB, TRIM29, SOX10, CEBPβ and TBX19, which are all highly expressed in ER$^-$ tumors. Of these transcription factors, our functional assays identified YBX1, NFIB and CBFB as important for proliferation in ER$^-$ cells in culture. Existing literature links individual transcription factors in cluster 2 to basal-like breast cancer[28–31], which is associated with increased aggressiveness, metastasis and epithelial-to-mesenchymal transition (EMT). Here we suggest a network of cooperating transcription factors important in determining this cancer subtype. The link of cluster 2 transcription factors to basal-like breast cancer is further supported by increased binding at GWAS loci by CEBPβ, a transcription factor required for lobuloalveolar development[32] whose loss is associated with EMT[33].

The most striking aspect of cluster 1 and cluster 2 transcription factors is the opposing regulatory effects they exert on their target genes. We postulate that this mutually exclusive activity reflects the decision of a progenitor to commit to either an ER$^+$ ductal or an ER$^-$ alveolar cell fate. In line with this hypothesis, we find that, in primary human mammary cell populations[17], cells representative of ER$^-$ alveolar progenitors show differential upregulation of cluster 2 transcription factors, whereas ER$^+$ luminal cells display higher expression of cluster 1 transcription factors (**Fig. 8**). Recent genetic tracing experiments have shown that ER$^+$ ductal progenitors and ER$^-$ alveolar progenitors are self-renewing in the mouse mammary gland[34–36]. The differential expression of risk TFs in these two self-renewing populations may suggest that these are the populations where risk-associated genes are effective and cell transformation occurs. In line with this hypothesis, the transcriptional profiles of basal-like tumors most resemble that of ER$^-$ alveolar progenitors[37,38], whereas luminal A and luminal B tumors phenocopy ER$^+$ ductal cells[18,38–40]. Furthermore, the ER$^-$ alveolar progenitor population is expanded in *BRCA1* mutation carriers[38], which are predisposed to develop ER$^-$ breast cancer.

The opposing activity of two distinct networks of transcription factors has not previously been reported but is consistent with studies carried out for individual transcription factors. For example, ELF5, an important inducer of alveolar differentiation[41], can reduce estrogen sensitivity in ER$^+$ cell lines[42]. FOXA1 in combination with GATA3 and ESR1 can specify an estrogen-responsive phenotype[23] and, conversely, is able to repress the basal phenotype[43]. The concept of antagonism between transcription factors led us to two-tailed GSEA of the ESR1 regulon (**Fig. 7**). Of potential clinical relevance, the analysis identifies a subgroup of histochemically ER$^+$ patients in whom the ESR1 regulon is functionally in a repressed state and for whom anti-estrogen treatment might not be effective. Our results also highlight the possibility that repression of cluster 1 transcription factors may lead to a shift in cell state toward more basal-like cancer, which is potentially associated with a more aggressive tumor phenotype and resistance to therapy. Better understanding of the interplay of key regulators will be critical for optimal therapeutic strategies.

In summary, we have shown that EVSE analysis, together with gene regulatory networks, can identify key regulators that may influence disease risk. The analysis can be applied to any combination of GWAS loci for which eQTLs can be interrogated, not just to those for which causative SNPs and genes are already known. For breast cancer, the risk-enriched regulons include many driven by transcription factors already implicated in breast cancer but many others that are not. The mutual antagonism of the two identified clusters of risk TFs provides new insights into their interactions, with potential clinical implications.

URLs. National Human Genome Research Institute (NHGRI) GWAS Catalog, http://www.genome.gov/gwastudies/; Cancer Gene Census, http://cancer.sanger.ac.uk/census; RTN R package, http://bioconductor.org/packages/RTN/; RedeR R package, http://bioconductor.org/packages/RedeR/; statmod R package, http://CRAN.R-project.org/package=statmod.

METHODS

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

Accession codes. ChIP-seq data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) under accession GSE74069; all microarray gene expression data have been deposited under accession GSE70759.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Califano for helpful discussions and J. Stingl for advice and critical reading of the manuscript. We are grateful to the genomics and bioinformatics cores at the Cancer Research UK Cambridge Institute for their support. This work was funded by Cancer Research UK and the Breast Cancer Research Foundation. M.A.A.C. is funded by the National Research Council (CNPq) of Brazil. T.E.H. held a fellowship from the US Department of Defense Breast Cancer Research Program (W81XWH-11-1-0592) and is currently supported by a Royal Adelaide Hospital Career Development Fellowship (Australia). T.E.H. and W.D.T. are funded by the National Health and Medical Research Council (NHMRC) of Australia (1008349 to W.D.T. and 1084416 to W.D.T. and T.E.H.) and Cancer Australia/National Breast Cancer Foundation (627229 to W.D.T. and T.E.H.; ID: PS-15-041-W.D.T.). B.A.J.P. is a Gibb Fellow of Cancer Research UK and a National Institute for Health Research (NIHR) Senior Investigator. We would like to acknowledge the support of the University of Cambridge, Cancer Research UK and Hutchison Whampoa, Ltd.

AUTHOR CONTRIBUTIONS

M.A.A.C. and K.B.M. designed the experiments, and M.A.A.C. and I.d.S. carried out the computational analysis. T.M.C. carried out the microarray experiments, and C.V. performed the siRNA transfection and proliferation analysis. T.E.H. and W.D.T. performed AR ChIP-seq experiments. E.R. carried out copy number normalization and eQTL calling. F.M. provided computational expertise. K.B.M., M.A.A.C. and B.A.J.P. developed the ideas and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Califano, A., Butte, A.J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).
2. Leiserson, M.D., Eldridge, J.V., Ramachandran, S. & Raphael, B.J. Network analysis of GWAS data. *Curr. Opin. Genet. Dev.* **23**, 602–610 (2013).
3. Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37**, 382–390 (2005).
4. Fletcher, M.N. *et al.* Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* **4**, 2464 (2013).
5. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
6. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
7. Cowper-Sal lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
8. Risbridger, G.P., Davis, I.D., Birrell, S.N. & Tilley, W.D. Breast and prostate cancer: more similar than different. *Nat. Rev. Cancer* **10**, 205–212 (2010).
9. Schliekelman, P. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* **178**, 2201–2216 (2008).
10. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
11. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
12. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
13. Kittler, R. *et al.* A comprehensive nuclear receptor network for breast cancer cells. *Cell Rep.* **3**, 538–551 (2013).
14. Hickey, T.E., Robinson, J.L., Carroll, J.S. & Tilley, W.D. Minireview: the androgen receptor in breast tissues: growth inhibitor, tumor suppressor, oncogene? *Mol. Endocrinol.* **26**, 1252–1267 (2012).
15. Carro, M.S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).
16. Bertucci, F. *et al.* Gene expression profiling shows medullary breast cancer is a subgroup of basal breast cancers. *Cancer Res.* **66**, 4636–4644 (2006).
17. Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.* **14**, R134 (2012).
18. Haughian, J.M. *et al.* Maintenance of hormone responsiveness in luminal breast cancers by suppression of Notch. *Proc. Natl. Acad. Sci. USA* **109**, 2742–2747 (2012).
19. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8**, e1002431 (2012).
20. Montgomery, S.B. & Dermitzakis, E.T. From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.* **12**, 277–282 (2011).
21. Fachal, L. & Dunning, A.M. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr. Opin. Genet. Dev.* **30**, 32–41 (2015).
22. 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).
23. Kong, S.L., Li, G., Loh, S.L., Sung, W.K. & Liu, E.T. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. *Mol. Syst. Biol.* **7**, 526 (2011).
24. Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
25. Andres, S.A. & Wittliff, J.L. Relationships of ESR1 and XBP1 expression in human breast carcinoma and stromal cells isolated by laser capture microdissection compared to intact breast cancer tissue. *Endocrine* **40**, 212–221 (2011).
26. Ross-Innes, C.S. *et al.* Cooperative interaction between retinoic acid receptor-α and estrogen receptor in breast cancer. *Genes Dev.* **24**, 171–182 (2010).
27. Robinson, J.L. *et al.* Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. *EMBO J.* **30**, 3019–3027 (2011).
28. Davies, A.H. *et al.* YB-1 transforms human mammary epithelial cells through chromatin remodeling leading to the development of basal-like breast cancer. *Stem Cells* **32**, 1437–1450 (2014).
29. Cimino-Mathews, A. *et al.* Neural crest transcription factor Sox10 is preferentially expressed in triple-negative and metaplastic breast carcinomas. *Hum. Pathol.* **44**, 959–965 (2013).
30. Moon, H.G. *et al.* NFIB is a potential target for estrogen receptor–negative breast cancers. *Mol. Oncol.* **5**, 538–544 (2011).
31. Ai, L. *et al.* TRIM29 suppresses TWIST1 and invasive breast cancer behavior. *Cancer Res.* **74**, 4875–4887 (2014).
32. Seagroves, T.N. *et al.* C/EBPβ, but not C/EBPα, is essential for ductal morphogenesis, lobuloalveolar proliferation, and functional differentiation in the mouse mammary gland. *Genes Dev.* **12**, 1917–1928 (1998).
33. Johansson, J. *et al.* MiR-155–mediated loss of C/EBPβ shifts the TGF-β response from growth inhibition to epithelial-mesenchymal transition, invasion and metastasis in breast cancer. *Oncogene* **32**, 5614–5624 (2013).
34. Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189–193 (2011).
35. Lafkas, D. *et al.* Notch3 marks clonogenic mammary luminal progenitor cells *in vivo*. *J. Cell Biol.* **203**, 47–56 (2013).
36. Rodilla, V. *et al.* Luminal progenitors restrict their lineage potential during mammary gland development. *PLoS Biol.* **13**, e1002069 (2015).
37. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in *BRCA1* mutation carriers. *Nat. Med.* **15**, 907–913 (2009).
38. Molyneux, G. *et al.* *BRCA1* basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell Stem Cell* **7**, 403–417 (2010).
39. Perou, C.M. & Borresen-Dale, A.L. Systems biology and genomics of breast cancer. *Cold Spring Harb. Perspect. Biol.* **3**, a003293 (2011).
40. Lim, E. *et al.* Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).
41. Oakes, S.R. *et al.* The Ets transcription factor Elf5 specifies mammary alveolar cell fate. *Genes Dev.* **22**, 581–586 (2008).
42. Kalyuga, M. *et al.* ELF5 suppresses estrogen sensitivity and underpins the acquisition of antiestrogen resistance in luminal breast cancer. *PLoS Biol.* **10**, e1001461 (2012).
43. Bernardo, G.M. *et al.* FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene* **32**, 554–563 (2013).

ONLINE METHODS

Computational analysis. *ARACNe-EVSE analysis.* Regulons were calculated on the basis of mutual information using the ARACNe algorithm[3]. Of the 809 transcription factors[3] tested, we were able to assign regulons to 555 in cohort I and 635 in cohort II of the METABRIC data set. The EVSE analysis has been described before[4], and here we extended our previous computational pipeline (RTN) to allow the testing of all regulons defined in the network. The steps and data sets used in this analysis are illustrated in **Supplementary Figure 2**. In more detail, EVSE was carried out using the 72 breast cancer risk SNPs identified by Michailidou *et al.*[5]. For most of these GWAS loci, neither the causative SNP nor the potential target genes are known. To deal with the former, the top hit at each locus (tagging SNP) was expanded into an AVS including all SNPs with linkage (D') >0.99 and logarithm of odds (LOD) >3.0 (**Supplementary Fig. 2a**), following the previously published VSE method[7]. This approach gave similar results to those obtained using the r^2 linkage metric to expand the tagging SNP into an AVS (**Supplementary Fig. 8**). To identify potential target genes at each GWAS locus, we used gene expression and genotyping data in a multivariate eQTL analysis[4]. When considering multiple GWAS loci in a single analysis, the number of potential target genes may vary strongly for each GWAS locus to be analyzed, making statistical comparisons between the loci difficult. For this reason, we carried out a single multivariate eQTL analysis at each GWAS locus, asking whether there is an association of any of the SNPs in the AVS with the expression of any of the genes in a given regulon in a ±250-kb window around the AVS (**Supplementary Fig. 2b**,**c**). (For each AVS, only the SNPs for which genotyping data were available in METABRIC were considered in the analysis.) If a positive association was found, the locus was counted toward a mapping tally (**Supplementary Fig. 2d**), as described by Cowper SalIari *et al.*[7]. In a subsequent step, statistical significance was assessed (**Supplementary Fig. 2e**). To reduce the cost of the computational analysis when interrogating many regulons, we ran a low-resolution analysis to remove obviously non-significant regulons (RTN package, Reconstruction of Transcriptional Networks and analysis of master regulators). For all remaining regulons, the EVSE analysis using breast cancer GWAS hits was tested against a null distribution based on random permutations of the AVS (matched random variant sets). These distributions were normalized and centered on the null to obtain an enrichment score, which is the number of standard deviations by which the observed mapping tally deviates from the mapping mean for the null distribution. From these null distributions, *P* values were calculated. To gain confidence in our results, we used cohorts I and II of the METABRIC data set separately and only considered regulons that were significant in both cohorts. Where different GWAS results were tested (BMD, prostate cancer and CLL), each GWAS set was controlled with the appropriate number of random SNPs. As a threshold for significance, a Bonferroni correction was applied.

eQTL analysis. We performed a *cis*-eQTL analysis for cohort I and cohort II breast cancer samples generated by the METABRIC study[6]. The analysis largely followed that by Li *et al.*[10]. We required probes to map to one of the RefSeq genes according to the annotation data obtained from the R package illuminaHumanv3.db. Probes mapping to genes in the highly polymorphic human leukocyte antigen (HLA) region were excluded from the analysis. Genes with low expression levels (the lowest 10% quantile of all expression values) were removed. Probes mapping to the same gene were treated independently in the eQTL analysis.

Copy number values for each gene in each sample were estimated from segmented copy numbers by averaging the copy number of all segments that fell in the region of the gene while using the length of the copy number segments as weights.

Gene expression levels were adjusted for copy number effects using the equation $T_i = \beta_i \mathrm{CN}_i + \varepsilon_I$, where T_i is the measured gene expression, CN_i is the copy number value, β_i is the regression coefficient and ε_i is the residual gene expression level of gene *i*.

The eQTL analysis was performed using MatrixEQTL in R (ref. 44) by correlating the genotypes of all remaining SNPs with the residual expression levels of proximal genes, that is, genes within 1 Mb of the SNP. In the case that multiple probes mapped to a gene, all probes for that gene were tested separately. Finally, significant associations were selected on the basis of a Benjamini-Hochberg false discovery rate (FDR) threshold of 0.1. Only SNPs with minor allele frequency (MAF) >0.05 were tested. This restriction is necessary because the effect of different genotypes on transcript levels cannot be evaluated if the genotypes at a given SNP locus are very homogeneous.

Master regulator analysis. MRA uses a hypergeometric test to assess whether a gene list is enriched in a given regulon[15]. If significant, the transcription factor controlling the regulon is likely to be involved in the regulation of the gene list. Our experimental design compares resting with cycling cells, and we therefore removed transcription factors that were also enriched with the Meta-PCNA signature[45] (**Supplementary Note**).

Variant set enrichment analysis. VSE analysis was carried out as previously described[7] using publically available data[4,13,22] (available under GEO accessions GSE48930, GSE41995 and GSM1010889 and ArrayExpress accessions E-MTAB-223 and E-MTAB-986). Briefly, VSE analysis tests enrichment of a chromosomal annotation, here transcription factor binding sites, at AVSs. An overlap between a ChIP-seq peak and a SNP in the AVS is counted toward a mapping tally that is tested against random SNPs, as in the EVSE analysis.

Differential gene expression. Differential gene expression was assessed using limma[46]. *z* scores were obtained by comparing the gene expression values averaged across all cell populations in the analysis against the averages for subgroups tested in each case. When determining significant differences in gene expression across primary cell populations, the following comparisons were considered: ALDH$^+$ versus ALDH$^-$ cells, ALDH$^+$ versus EpCAM$^+$CD49f$^-$ cells and ALDH$^-$ versus EpCAM$^+$CD49f$^-$ cells.

Two-tailed gene set enrichment analysis. GSEA[47] assesses the skewed distribution of a selected gene set (*S*), here the ESR1 regulon, in a list of genes (*L*) ranked by a particular phenotype, in this case the differential gene expression observed when comparing a given tumor with the average expression for all METABRIC tumors. An enrichment score (ES) was calculated by walking down list *L*, increasing by $1/|S|$ a running-sum statistic when encountering a gene in *S* and decreasing the statistic by $1/(|L| - |S|)$ when encountering a gene not in *S*. The enrichment score is the maximum deviation from zero. The two-tailed GSEA method is based on the Connectivity Map (CMAP) procedure[48]. The ESR1 regulon was derived by ARACNe from METABRIC cohort I data and filtered using genefilter in Bioconductor to remove uninformative genes, about 15% of the regulon, mostly of low variance. Feature selection was performed on cohort II and used to filter the regulon in cohort I and vice versa. The resultant regulon was split into two subgroups, positive targets (A) and negative targets (B), using Pearson's correlation to assign directionality. The distribution of A and B was then tested by the GSEA statistics in the ranked phenotype, producing independent enrichment scores for each subgroup. An additional step calculated the differential enrichment ($\mathrm{dES} = \mathrm{ES_A} - \mathrm{ES_B}$). Two-tailed GSEA was performed in R using the tni.gsea2 function in the RTN package[4] with 1,000 permutations.

Survival data[6] were used to plot Kaplan-Meier curves, and *P* values were calculated using log-rank statistics. On the basis of differential enrichment score values, the patients were divided into three groups: those with an active ESR1 regulon (dES >0, $\mathrm{ES_A}$ >0 and $\mathrm{ES_B}$ <0), those with a repressed ESR1 regulon (dES <0, $\mathrm{ES_A}$ <0 and $\mathrm{ES_B}$ >0) and a small group for whom the differential enrichment score values were around zero (inconclusive, with $\mathrm{ES_A}$ and $\mathrm{ES_B}$ distributions skewed to the same side). The two large groups were further subdivided in half.

We tested the response of the ESR1 regulon to estrogen or estrogen plus tamoxifen treatment by applying two-tailed GSEA to gene expression data from Hurtado *et al.*[22] using differential gene expression (estrogen versus vehicle and estrogen versus estrogen plus tamoxifen; GSE25316) as the phenotype to rank the gene list (*L*).

Cell culture. The human breast cancer cell lines MCF-7 and MDA-MB-453 (HTB 131; American Type Culture Collection (ATCC)) were cultured in DMEM (Invitrogen), and the ZR-75-1 and T-47D cell lines were cultured in RPMI (Invitrogen), all supplemented with 10% FBS and antibiotics. The MCF10A cell line was cultured in DMEM, 5% horse serum, 5 µg/ml insulin, 1 µg/ml hydrocortisone, 100 ng/ml cholera toxin, 20 ng/ml epidermal growth factor

doi:10.1038/ng.3458

(EGF) and 2 mM L-glutamine. Unless otherwise stated, all cells were from the Cancer Research UK Cambridge Institute biorepository and were maintained at 37 °C in 5% CO_2. All cell lines used were free of mycoplasma.

Chromatin immunoprecipitation. ChIP-seq experiments were carried out as previously described[49]. Cells were seeded at ~70% confluence in 15-cm tissue culture dishes (four per treatment). Following overnight attachment, cells were starved using base medium containing 5% steroid-stripped FBS. To ensure steroid depletion before treatment, medium was changed every day for 3 d; cells were then treated for 4 h with vehicle control (ethanol), DHT (10 nM) or MPA (10 nM). Cells were cross-linked, and ChIP-seq was performed using an antibody to AR (N20; sc-816, Santa Cruz Biotechnology; 10 μg per immunoprecipitation) with subsequent data processing as previously described[27]. Two independent experiments were performed in each cell line, and consensus AR chromatin-binding events were determined for each treatment condition.

Gene expression analysis for estrogen and FGF10 signaling. MCF-7 cells were plated at 5×10^5 cells/well in six-well dishes and left in complete medium overnight. Cell synchronization via estrogen starvation was then carried out for 3 d in estrogen-free medium (phenol red–free medium supplemented with 5% charcoal, dextran-treated FBS and 2 mM L-glutamine), with the medium changed every 24 h. Estrogen-deprived cells were stimulated with 1 nM β-estradiol (E2; Sigma) or with 100 ng/ml FGF10 (Invitrogen) in combination with 1 nM E2. Six hours after cell treatment, total RNA was isolated from three biological replicates, quality controlled, and used for cRNA amplification and labeling with the Illumina TotalPrep-96 kit (Ambion). cRNA was hybridized to HumanHT-12 v4 Expression BeadChips according to the manufacturer's protocol (Illumina, *WGGX DirectHyb Assay Guide 11286331 RevA*). Raw image files were processed and analyzed using the beadarray package from Bioconductor.

Transient transfection with siRNA. Cell lines were transfected with ON-TARGETplus SMARTpool siRNA (Dharmacon) directed against the risk TFs NFIB (L-008456-00), YBX1 (L-010213-00), CBFB (L-011602-00), LMO4 (L-012124-00), ELF5 (L-011265-02), TBX19 (L-011910-00) and SOX10 (L-017192-00). CEBPβ was not included in the analysis because multiple distinct isoforms with opposing functions may be present in the cell. A custom siRNA was used against FOXA1 (ref. 22). Knockdown of mRNA was confirmed for each cell line by RT-PCR of cDNA 48 h after transfection (**Supplementary Fig. 20c**) using the primer pairs listed in **Supplementary Table 5**. One microgram of total RNA was reverse transcribed using the High-Capacity cDNA Reverse Transcription kit (Applied Biosystems), and quantitative RT-PCR was performed using cDNA obtained from 10 ng of total RNA. Quantitative RT-PCR was performed using an ABI 9800HT Sequence Detection System (Applied Biosystems) with SDS software version 2.3. Amplification and detection were carried out in 384-well Optical Reaction Plates (Applied Biosystems) with Power SYBR Green Fast 2× qRT-PCR Mastermix (Applied Biosystems). All expression data were normalized to *DGUOK* expression. Primer specificity was confirmed at the end of each quantitative RT-PCR run through the generation of single peaks in melt-curve analysis. siRNA against SOX10 did not cause a reduction in mRNA levels and was not used for further analysis. A control, non-targeting pool of siRNAs (Dharmacon, D-001810-01-05) was included in each experiment. Transfections were carried out using Lipofectamine RNAiMAX Reagent (Invitrogen), according to the manufacturer's protocol. Growth was measured in 96-well plates using the IncuCyte (Essen BioScience) system every 3 h. Data from eight wells were averaged for each experiment, and at least two repeats were carried out for each cell line (MCF10A, $n = 3$; ZR-75-1, $n = 2$). The results of knockdown of transcription factors in cluster 2 that are not consensus risk TFs are shown in **Supplementary Figure 20**. Statistical analysis was carried out using the compareGrowthCurves command in the statmod package in R, generating Benjamini-Yekutieli[50] adjusted *P* values.

Code availability. The source code developed in this study is publicly available from Bioconductor[51] in the R packages RTN[4] and RedeR[52] (see URLs).

44. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
45. Venet, D., Dumont, J.E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7**, e1002240 (2011).
46. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
47. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
48. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
49. Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48**, 240–248 (2009).
50. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Satist.* **29**, 1165–1188 (2001).
51. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
52. Castro, M.A. *et al.* R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biol.* **13**, R29 (2012).