



ARTICLE

DOI: 10.1038/s41467-018-04406-2

OPEN

# Genome-scale identification of transcription factors that mediate an inflammatory network during breast cellular transformation

Zhe Ji <sup>1,2,4</sup>, Lizhi He<sup>1</sup>, Asaf Rotem<sup>1,2,5</sup>, Andreas Janzer<sup>1,6</sup>, Christine S. Cheng<sup>2,7</sup>, Aviv Regev<sup>2,3</sup> & Kevin Struhl <sup>1</sup>

Transient activation of Src oncoprotein in non-transformed, breast epithelial cells can initiate an epigenetic switch to the stably transformed state via a positive feedback loop that involves the inflammatory transcription factors STAT3 and NF- $\kappa$ B. Here, we develop an experimental and computational pipeline that includes 1) a Bayesian network model (AccessTF) that accurately predicts protein-bound DNA sequence motifs based on chromatin accessibility, and 2) a scoring system (TFScore) that rank-orders transcription factors as candidates for being important for a biological process. Genetic experiments validate TFScore and suggest that more than 40 transcription factors contribute to the oncogenic state in this model. Interestingly, individual depletion of several of these factors results in similar transcriptional profiles, indicating that a complex and interconnected transcriptional network promotes a stable oncogenic state. The combined experimental and computational pipeline represents a general approach to comprehensively identify transcriptional regulators important for a biological process.

<sup>1</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA. <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. <sup>3</sup>Department of Biology, Howard Hughes Medical Institute and David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 20140, USA. <sup>4</sup>Present address: Department of Pharmacology and Biomedical Engineering, Northwestern University, Evanston 60611 IL, USA. <sup>5</sup>Present address: Department of Medical Oncology and Center for Cancer Precision Medicine, Dana-Farber Cancer Institute, Boston 02215 MA, USA. <sup>6</sup>Present address: Bayer Pharma, Berlin 13353, Germany. <sup>7</sup>Present address: Department Biology, Boston University, Boston 02215 MA, USA. These authors contributed equally: Zhe Ji, Lizhi He. Correspondence and requests for materials should be addressed to K.S. (email: [kevin@hms.harvard.edu](mailto:kevin@hms.harvard.edu))

Transcriptional regulatory proteins that bind specific DNA sequences are the major determinants for regulating gene expression programs that determine cell state and behavior<sup>1–3</sup>. It is therefore important to comprehensively identify the transcription factors and transcriptional regulatory circuits involved in dynamic biological processes and maintaining stable cell states. Individual experimental approaches provide specific types of information, but integration of the various datasets is necessary for comprehensive understanding. There are only a few examples in which transcription factors important for a biological process and transcriptional regulatory connections have been identified on a comprehensive basis<sup>4–8</sup>. None of these have been performed in the context of cellular transformation or cancer.

Transcriptional activator or repressor proteins recruit co-activator or co-repressor complexes to their target sites via protein-protein interactions, thereby altering the level of transcription by the general RNA polymerase II machinery<sup>9</sup>. Some co-activator and co-repressor complexes are enzymes that locally modify chromatin structure either by altering nucleosome positions, removing nucleosomes to generate accessible DNA, or modifying histones at specific residues through acetylation and methylation. Chromatin-modifying activities are important for transcriptional regulation, but they are not the major determinants of gene expression patterns due to their limited specificity for genomic DNA sequences and their widespread presence in different cell types. Nevertheless, locally altered chromatin structures represent genomic regions of transcription factor activity *in vivo* under the physiological conditions tested.

Accessible chromatin regions can be mapped on a genomic scale by DNase I hypersensitivity<sup>10</sup> or transposon-based ATAC-seq<sup>11</sup>. Accessible regions are typically several hundred bp in length, and they are generated by nucleosome-remodeling complexes that are recruited by the combined action of multiple DNA-binding (and associated) proteins that bind to motifs within the accessible region. This combinatorial recruitment is critical for biological specificity, because individual sequence motifs are short and hence occur very frequently throughout large mammalian genomes simply by chance<sup>12,13</sup>. Many DNase I hypersensitive regions are promoters or enhancers; these are distinguished by virtue of their proximity to the transcriptional initiation site and by histone modifications (e.g., tri-methylated H3-K4)<sup>14</sup>.

The above considerations make it possible to use genome-scale chromatin accessibility maps to identify transcription factors that regulate gene expression programs during biological progresses. One approach is to search for sequence motifs enriched in differentially accessible chromatin regions<sup>15,16</sup>. However, dynamic chromatin accessibility and differential gene expression are not always correlated, and many functionally important transcription factors play a constitutive (i.e., non-regulated) role and will not be identified by this approach. More directly, genome-scale DNase I footprinting<sup>17,18</sup> and transposon-based ATAC-seq<sup>11</sup> can identify genomic regions protected by bound proteins. However, these footprinting maps require ~10 times more sequencing reads than hypersensitivity maps, and hence are considerably more expensive.

Transcription factors important in various biological contexts have been identified by integrating chromatin accessibility or footprinting analyses with gene expression profiles<sup>19–21</sup>. However, these previous integrative analyses did not comprehensively evaluate transcription factors for their role in the biological process of interest. Here, we develop an experimental and computational pipeline to comprehensively identify transcription factors and transcriptional regulatory circuits involved in a biological process. We apply this approach to an inducible model of cellular transformation in which transient activation of v-Src

oncoprotein converts a non-transformed breast epithelial cell line (MCF-10A) into a stably transformed state within 24 h<sup>22,23</sup>. This epigenetic switch between stable non-transformed and transformed states is mediated by an inflammatory positive feedback loop involving the transcription factors NF- $\kappa$ B and STAT3<sup>23,24</sup>. A few transcriptional regulatory circuits involved in this transformation model have been identified, and these are important in some other cancer cell types and human cancers<sup>24–27</sup>. However, a comprehensive analysis of transcriptional circuitry involved in this or any other model of cellular transformation has not been described.

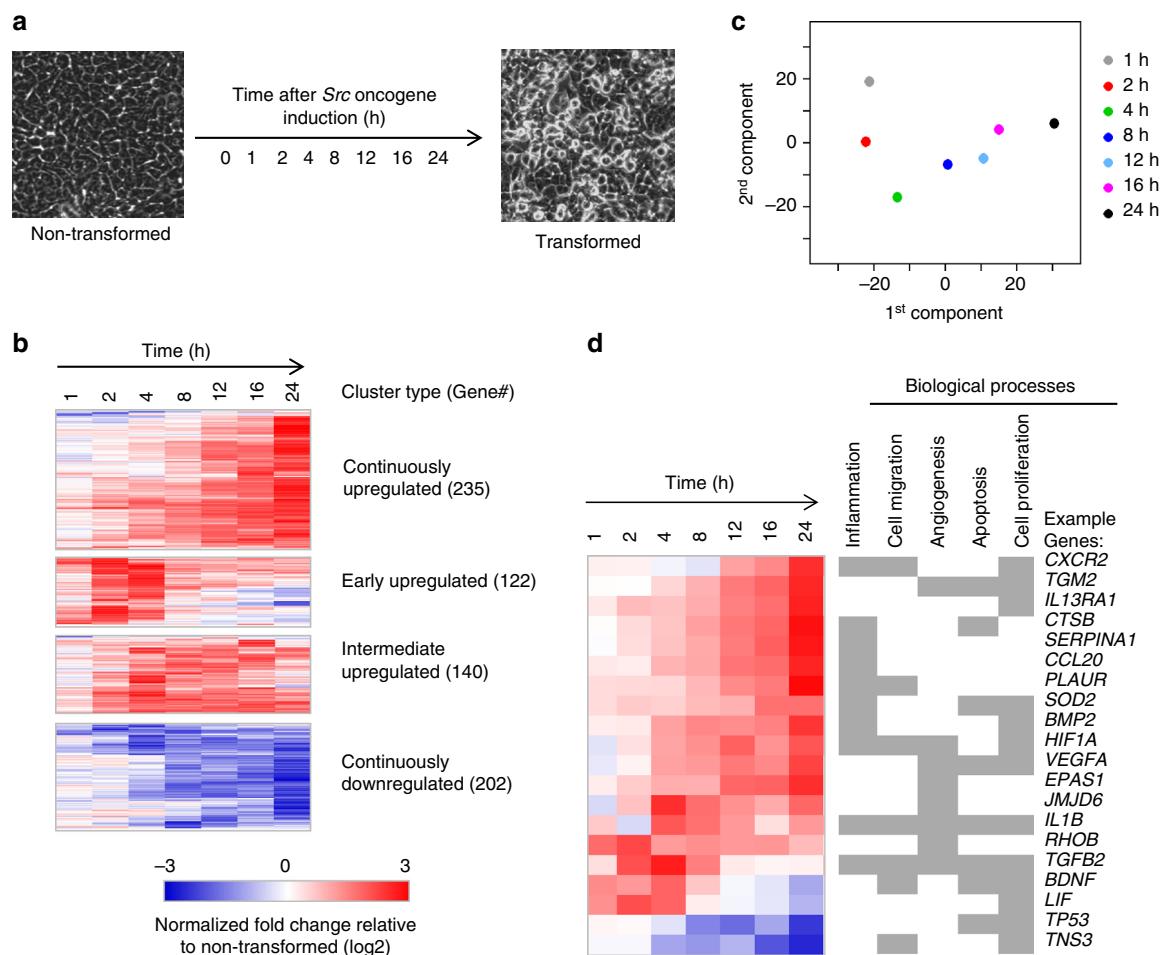
Using this approach, we show that >40 transcription factors are important for transformation in this model system. Furthermore, although these factors have different DNA-binding specificities, they can affect the expression of a common set of genes. This suggests that cellular transformation is mediated by a highly interconnected transcriptional regulatory circuit that depends on the combined inputs of many transcription factors.

## Results

**Transcriptional regulatory modules during transformation.** To improve our initial transcriptional profiling analysis<sup>22</sup>, we re-analyzed mRNA expression profiles during the process of transformation (0, 1, 2, 4, 8, 16, and 24 h time points in the presence of tamoxifen, which induces v-Src; Fig. 1a). Approximately 700 genes are differentially expressed with >1.5-fold change consistently in two biological replicates in at least one time-point (False Discovery Rate <0.007). These genes form four coherent clusters: continuously up-regulated; early up-regulated at 2 h; intermediate up-regulated at 12 h; continuously down-regulated (Fig. 1b). Principal component analysis indicates that the transcriptional program gradually evolves during the transformation process (Fig. 1c). As expected, differentially expressed genes are enriched in pathways strongly associated with cancer progression such as the inflammatory response, cell migration, angiogenesis, regulation of apoptosis, and cell proliferation (Fig. 1d and Supplementary Fig. 1).

**Genome-scale mapping of transcriptional regulatory regions.** Genome-scale mapping of DNase hypersensitive sites (DNase-seq)<sup>10</sup> of cells at 0, 6, and 24 h after tamoxifen treatment reveals ~212,423 accessible regions (an example region is shown in Fig. 2a). To further classify types of such regulatory regions<sup>14,28</sup>, we performed ChIP-seq for 6 histone modifications at 0, 2, 12, 24, and 36 h after tamoxifen treatment (Fig. 2a). These results indicate that 12% of the open chromatin regions are located in active promoters (H3-K27ac and H3-K4me3), 25% are in active enhancers (H3-K27ac but no H3-K4me3), 19% are in primed enhancers (H3-K4me1 but no H3-K27ac or H3-K4me3), 16% are in heterochromatin (H3-K9me3) or polycomb-repressed regions (H3-K27me3), and the remaining 28% uncharacterized based on our histone modification analysis (Fig. 2b).

Chromatin accessibility and H3-K27ac levels are dynamically regulated during the transformation process, while the levels of various types of histone methylation are largely unchanged (Fig. 2c). Open chromatin regions in enhancers and heterochromatin are more likely to be dynamically regulated than open regions in promoters, and 5 times as many genomic regions show increased accessibility upon transformation as opposed to decreased accessibility (Fig. 2d). Among open acetylated regions, those showing increased accessibility during transformation tend to be more acetylated (Fig. 2e). This suggests that many chromatin changes are due to increased function of transcriptional activator proteins bound at enhancers that recruit nucleosome remodeler and histone acetylase complexes.



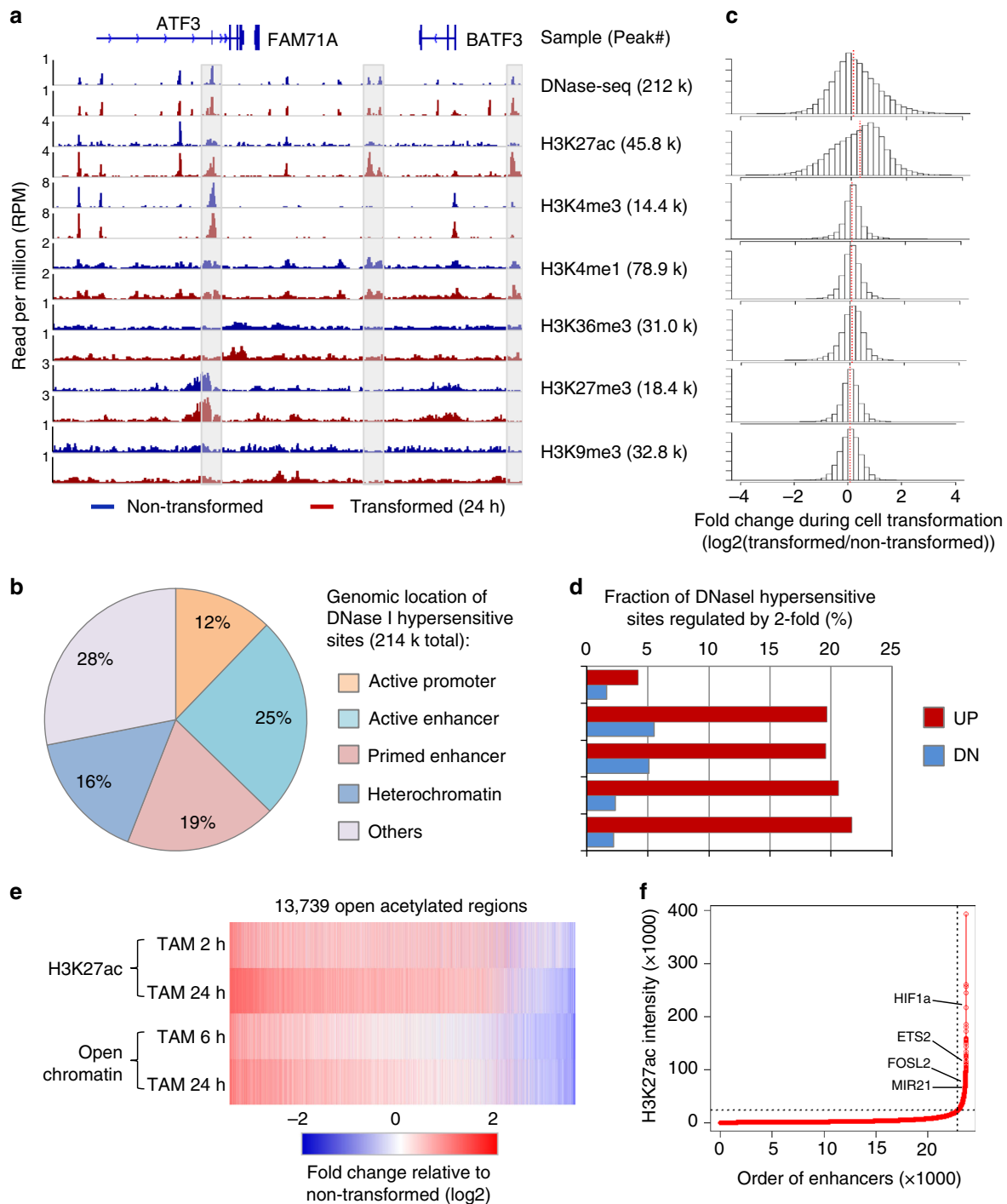
**Fig. 1** Differential gene expression during transformation. **a** Time-series of cells treated with tamoxifen for 0 h, 1 h, 2 h, 4 h, 8 h, 12 h, 16 h, and 24 h. **b** The K-mean clustering of 699 differentially expressed genes into four coherent clusters. **c** Principal component analyses of differential RNA expression profile. **d** Example genes in pathways enriched with differentially expressed genes. One gene may regulate several oncogenic processes as indicated in gray color

**Super-enhancer genes are preferentially activated.** Super-enhancers, previously termed dominant or local control regions<sup>29</sup>, are large clusters of individual enhancers that typically drive expression of genes defining cell identity<sup>30–32</sup>. Using the ROSE software<sup>30–32</sup> and ChIP-seq data for H3-K27ac, we identified 1050 super-enhancers in at least one time point at 0, 2, and 24 h after tamoxifen treatment (Fig. 2f and Supplementary Data 1), most of which (85%) pre-exist in non-transformed cells. 367 super-enhancers (35%) show increased acetylation levels >1.5 fold after 24 h of cell transformation, whereas only 18 showed >1.5 fold decreased acetylation (Supplementary Fig. 2a). Expression of the genes located in these activated super-enhancer regions tend to be up-regulated upon transformation (Supplementary Fig. 2b). Gene ontology analyses of genes located in the super-enhancer regions are enriched in the oncogenic pathways, such as cell migration, cell proliferation, intracellular signaling cascade, angiogenesis and gene transcription (Supplementary Fig. 3).

**Bayesian network model to predict TF binding sites in vivo.** As DNase I hypersensitive regions are generated ultimately by transcription factors bound (directly or indirectly) to DNA sequences, motif analysis of these accessible regions is a straightforward approach to identify the relevant transcription factors. However, this approach involves arbitrary cut-off choices

for the quality of sequence motifs, it does not distinguish between motifs at the center or edges of accessible regions, and it does not account for different levels of accessibility. Here, we describe a Bayesian Network model approach (AccessTF) that starts from all known sequence motifs in the human genome to predict protein-binding sites in vivo from DNase I hypersensitivity data. AccessTF integrates quantitative DNase I hypersensitive measurements with the following motif information: motif quality; the distance to the closest transcription start site; conservation among vertebrates (Fig. 3a, b). We define each motif to be in a bound or unbound state, with a motif more likely to be bound if located in a DNase hypersensitive region, has higher quality, higher conservation level, and is more proximal to a transcription start site (Fig. 3b). The Bayes algorithm is converged and calculates the probability that a motif is bound.

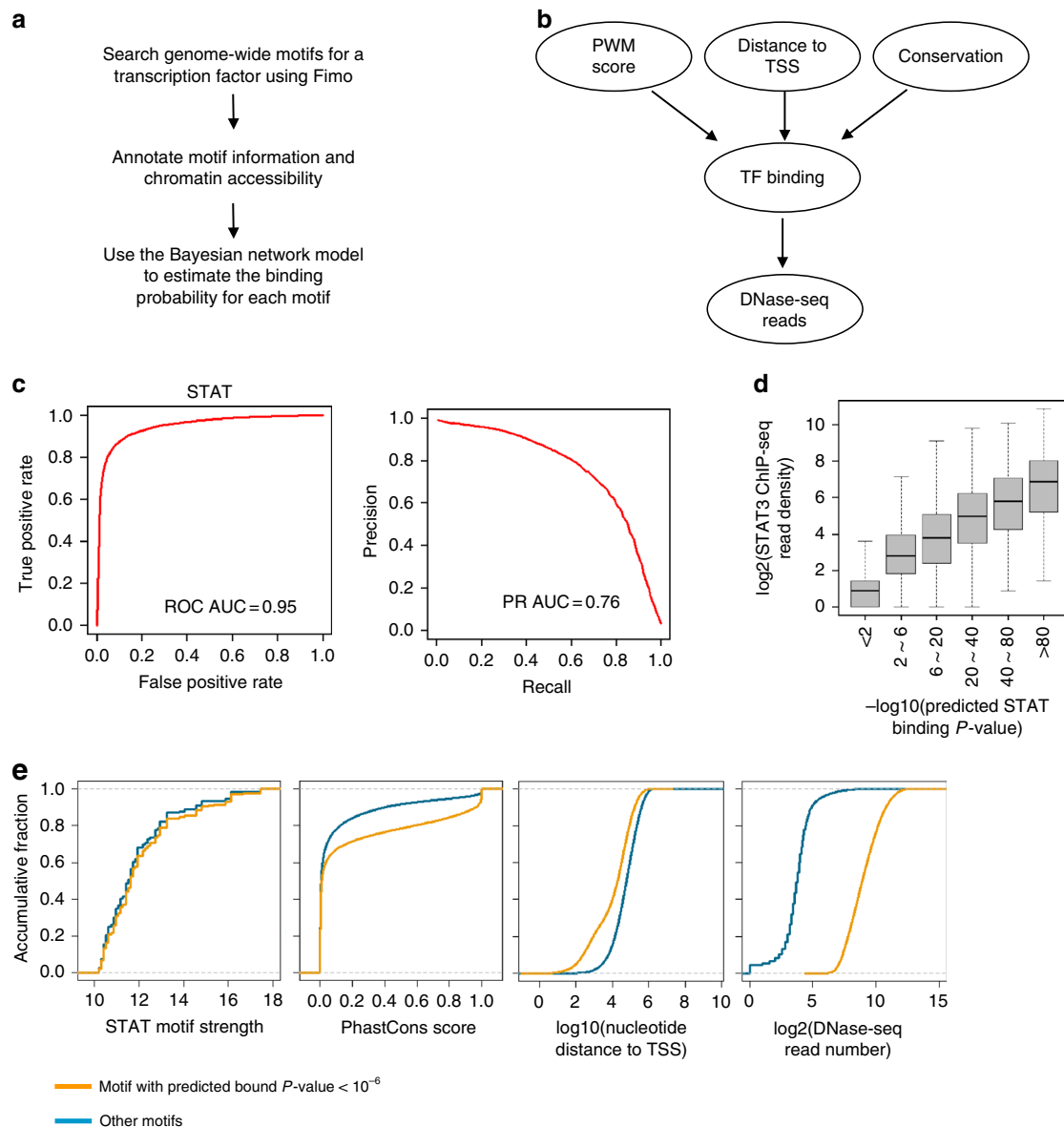
We tested the performance of the algorithm on binding sites for AP-1 and STAT3, factors for which we have ChIP-seq data in the same cell line<sup>33</sup>. The Area Under ROC Curve (ROC AUC) is >0.95 for both factors, the Area Under Precision-Recall Curve (PR AUC) is 0.76 for STAT3 and 0.86 for AP-1 (Fig. 3c and Supplementary Fig. 4c), and the predicted motif binding probability increases in accord with the factor binding level (Fig. 3d and Supplementary Fig. 4b). As expected, DNase I hypersensitivity around the motif is the major parameter for distinguishing bound vs. unbound motifs, although motif information adds some power to the prediction (Fig. 3e and



**Fig. 2** Differential chromatin accessibility and histone modification during transformation. **a** Data for the indicated chromatin features before (blue) and after (red) transformation. Total peak numbers are shown. **b** Distribution of genomic locations of DNase I hypersensitive sites (DNase-seq peaks) classified by histone modifications as active promoters (H3K4me3), active enhancers (H3K27ac), primed enhancers (H3K4me1), and heterochromatin (H3K27me3/H3K9me3). Regions without any modification were grouped as “Others”. **c** Fold change of chromatin accessibility and histone modification levels, calculated as the  $\log_2$  fold change of read density in cells after 24 h of tamoxifen treatment and non-transformed cells. **d** Fraction of DNase I hypersensitive sites differentially accessible ( $>2$ -fold change) during transformation. **e** Heat map showing differential chromatin accessibility and H3K27ac in 13,739 open acetylated regions during transformation. Pearson Correlation Coefficient of fold change values of accessibility vs. acetylation = 0.36 at 24 h after tamoxifen treatment;  $P < 10^{-100}$ . **f** Super-enhancers are identified by the sorted rank order based on H3K27ac levels at 24 h upon tamoxifen treatment. The analyses were based on the ROSE software<sup>30–32</sup>

Supplementary Fig. 4c). For the factors tested, the quality of the sequence motif makes a minimal contribution. As further validation of the algorithm, we performed a similar analysis in K562 cells using a DNase-seq dataset (for predictions) and ChIP-

seq datasets for many transcription factors (for testing) obtained by the ENCODE consortium<sup>34</sup> (Supplementary Fig. 5). For many factors, the AUC were  $>0.9$  and the PR AUC were  $>0.75$ , indicative of high performance. Some factors had lower AUC



**Fig. 3** AccessTF, a Bayesian network model to identify TF binding sites. **a** Steps of computational analyses for identifying transcription factor binding sites using DNase-seq data. **b** AccessTF integrates motif strength, distance to the closest transcription start site, Phastcon conservation score and surrounding DNase-seq reads. See Methods for a detailed description. **c** Area under ROC curves (ROC AUC) and Area under Precisions-Recall Curves (PR AUC) for measuring the performance of AccessTF predicting the STAT binding status. **d** STAT motifs grouped by predicted binding probabilities and plotted against STAT3 binding levels in 400 nt region around the motifs (estimated by ChIP-seq data) after transformation. For the box plot, the bounds of the box represent the first and third quartiles and the center line represents the median. **e** Comparing features of STAT motifs with predicted bound  $P$ -value  $< 10^{-6}$  vs. others

values (0.75–0.85), probably because these factors can bind sites that are not in open chromatin regions. Thus, the algorithm performs well to identify experimentally determined binding sites.

We then applied AccessTF to identify putative *in vivo* binding sites for all factors with PWM annotated by MotifDB<sup>35</sup>. Given the accuracy of the predictions, we expect that the vast majority of predicted sites are bound by their cognate factors *in vivo* under the conditions tested. However, this motif analysis does not distinguish among individual members of multi-protein families that recognize a common sequence motif (e.g., AP-1).

**Predicting TFs that regulate chromatin and expression.** To identify transcription factors important for the cellular

transformation, we examined the relative contribution of factor binding motifs, identified above, to differential chromatin accessibility and to their enrichment in promoters/enhancers of differentially expressed gene clusters. Open chromatin regions containing AP-1, NRF/MAF, STAT, and CEBP motifs are more likely to have increased accessibility during cell transformation, as compared to other open chromatin regions (Wilcoxon Rank Sum Test  $p$ -value  $< 10^{-40}$ ) (Supplementary Fig. 6a). Those motifs are also enriched in promoters/enhancers of differentially expressed gene clusters (Supplementary Fig. 6b). Interestingly, these motifs are associated both with genes that are continuously up-regulated and continuously down-regulated during transformation. Such locus-specific effects on transcriptional factor function are

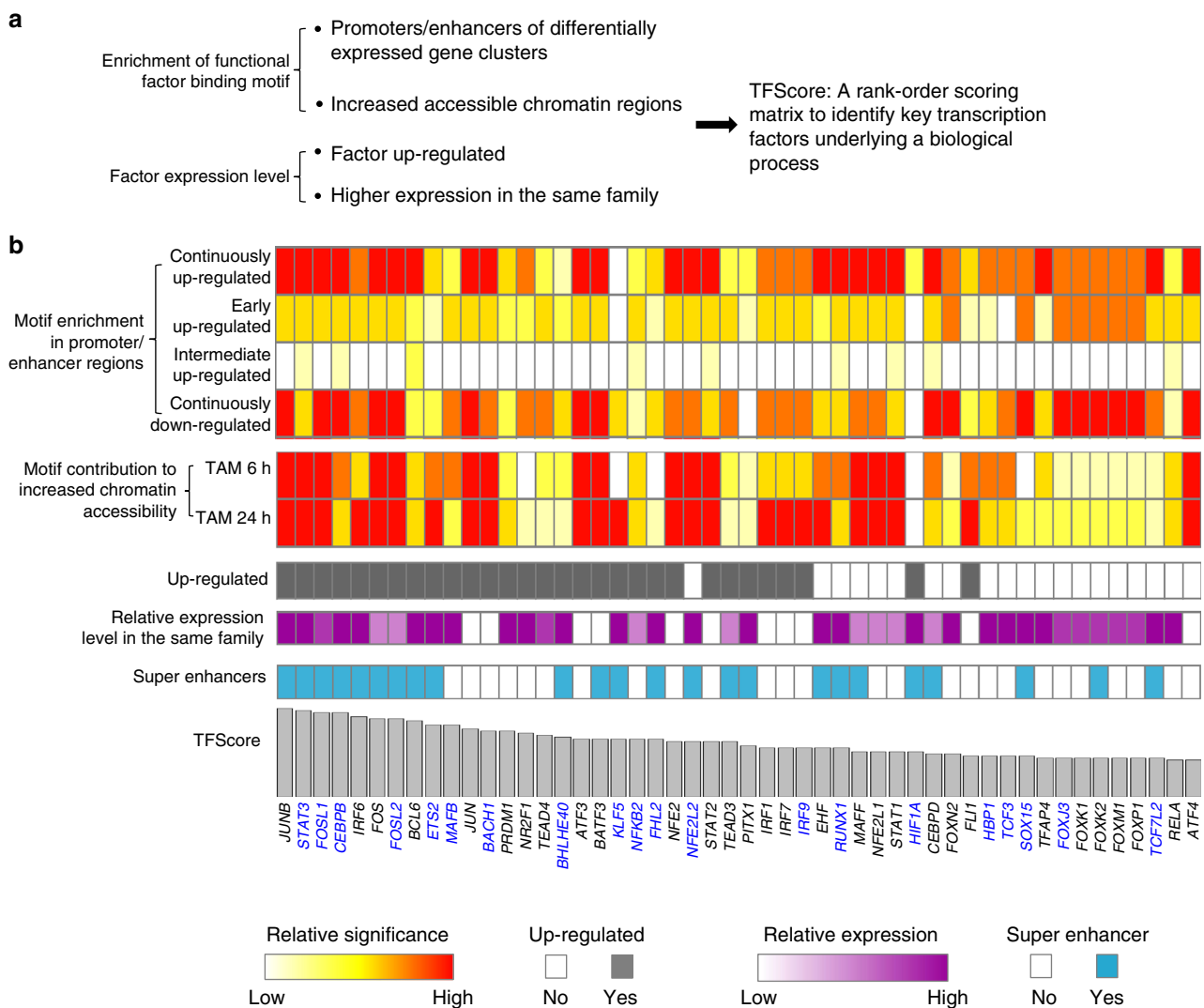
typically due to functional interactions with other factors that differentially bind the relevant genomic regions.

On the other hand, chromatin regions with CTCF and NFI motifs tend to show less accessibility (Wilcoxon Rank Sum Test  $p$ -value  $< 10^{-40}$ ) (Supplementary Fig. 6c). Open chromatin regions with CTCF or NFI motifs and without AP-1, NRF/MAF, STAT, or CEBP motifs are even more likely to have decreased accessibility (Supplementary Fig. 6c). It is unclear whether such decreased accessibility reflects decreased activator function and hence decreased recruitment of the nucleosome remodelers or increased recruitment of transcriptional co-repressors (e.g., histone deacetylases) that inhibit the association and/or function of the remodelers.

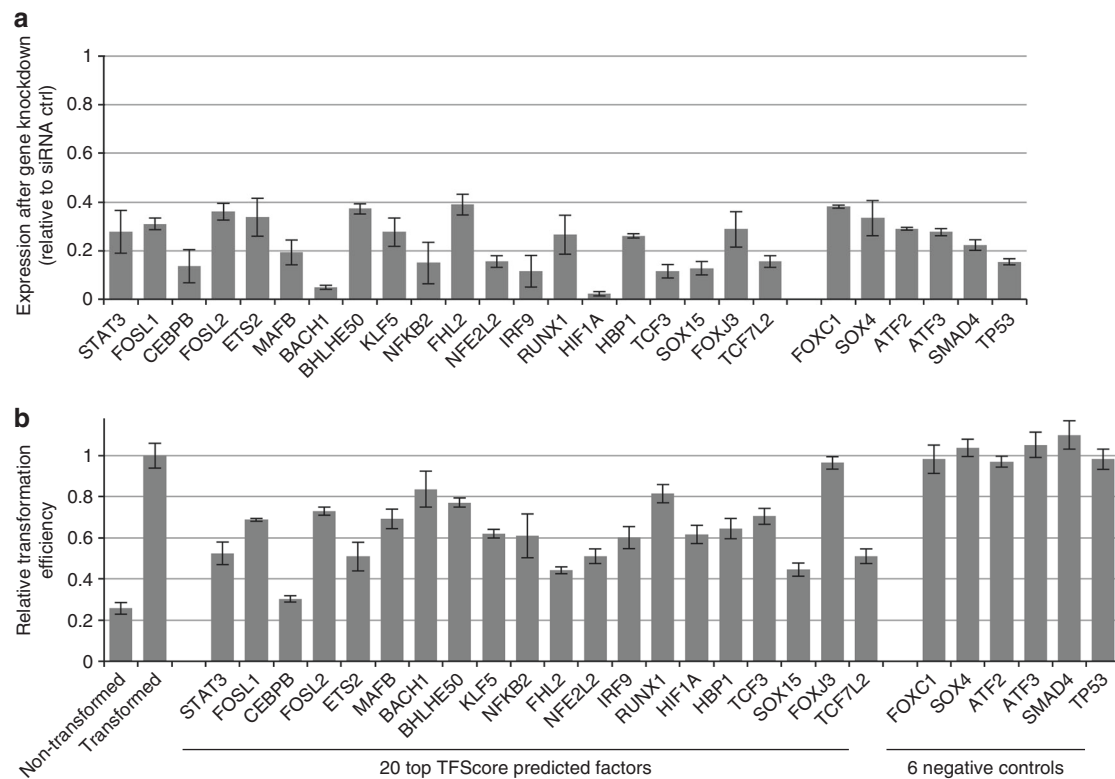
**TFScore to predict TFs important for transformation.** We developed a score schema (TFScore) to rank transcription factors in terms of their likelihood of being important for transformation. TFScore is based on 4 criteria (Fig. 4a): 1) higher motif enrichment in promoters/enhancers of the differentially expressed gene clusters (Fig. 1b); 2) higher motif occurrences in chromatin regions showing increased accessibility at 6 and 24 h after

tamoxifen treatment (Fig. 2e); 3) up-regulation of the factor during cell transformation; 4) higher relative expression level of an individual factor of a given transcription factor family that recognize a common sequence motif. The latter two criteria were used to distinguish the contributions of the various factors with similar DNA-binding specificities, based on the idea that factors expressed at higher levels and/or up-regulated are more likely to be important for transformation. The resulting rank-ordered list of transcription factors (Fig. 4b and Supplementary Data 2) reveals known regulators STAT3 and NF- $\kappa$ B near the top of the list and an unexpectedly large cohort of transcriptional regulators as potentially being important for the oncogenic transformation. Although only ~10% of human protein-coding genes are present in super-enhancer regions, 24 out of the top 50 TFScore-predicted transcriptional factors are located in super-enhancer regions (Fig. 4b and Supplementary Data 2), which is highly significant (Binomial Test  $P$ -value  $< 10^{-8}$ ) and suggestive of their functional importance.

**Many TFScore-predicted TFs are important for transformation.** To validate the functional importance of the predicted



**Fig. 4** Identifying transcription factors important for transformation. **a** Criteria for TFscore. See Methods for a detailed description. **b** Rank order of transcription factors potentially playing important regulatory roles during cell transformation, based on TFscores (top 50 are shown). For each transcription factor, the relative contribution of 4 criteria to TFscore, and annotated super-enhancer genes are shown. The 20 genes used for validation by siRNA knockdowns (Fig. 5) are shown in blue



**Fig. 5** TFscore-predicted transcription factors are important for transformation. **a** siRNA knockdown efficiencies of the indicated transcription factors as compared to siRNA controls. **b** Relative transformation efficiency of the indicated transcription factors as determined by their ability to grown under low attachment conditions

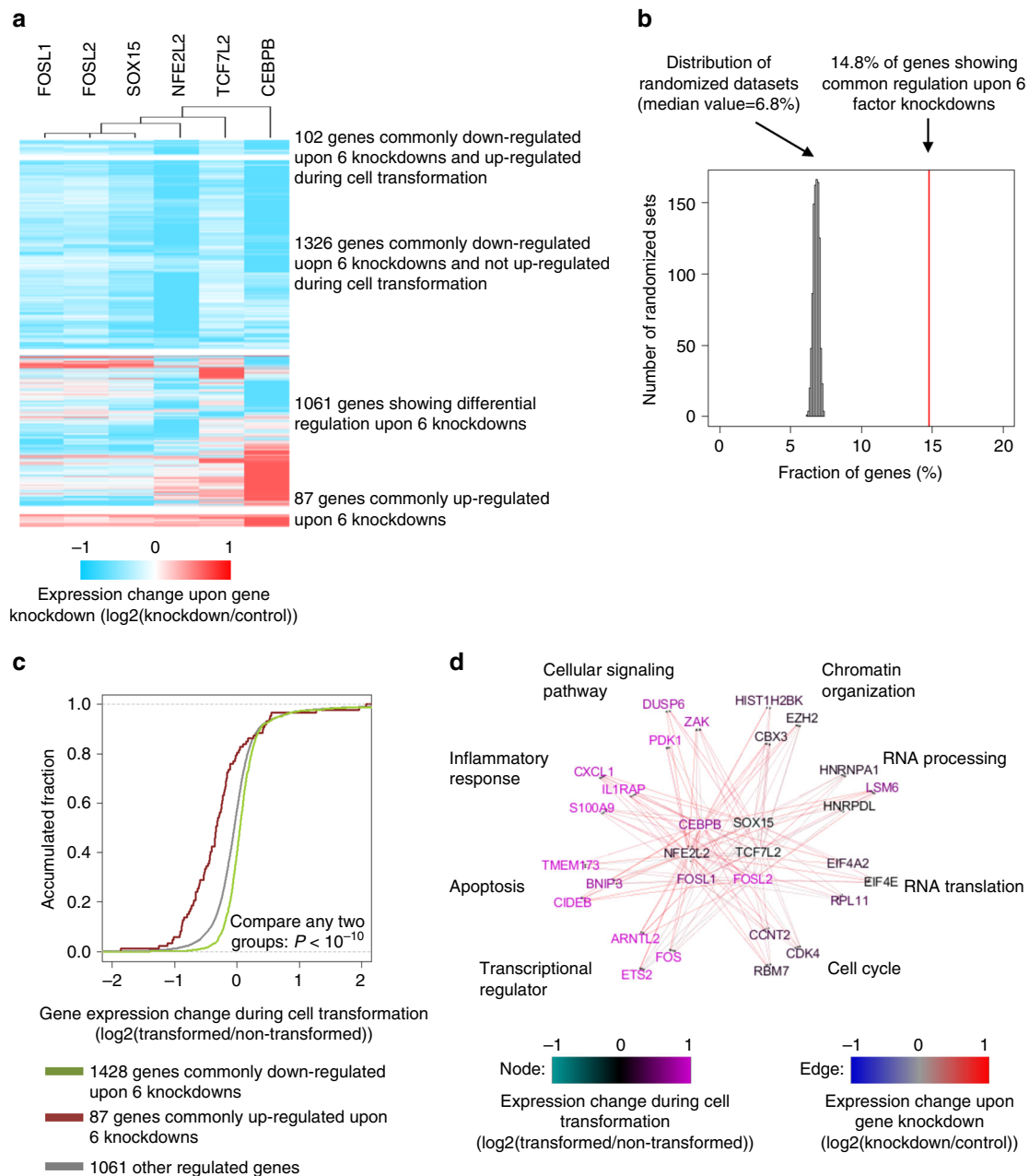
transcription factors, we genetically inhibited expression of their genes via siRNAs and tested the resulting cells for their ability to grow under conditions of low attachment<sup>36</sup>, a property of transformed cells. We randomly selected 20 of the top 50 candidate transcription factors predicted by TFscore, and as determined by mRNA levels, we obtained knockdown efficiencies of >60% (Fig. 5a). Of the 20 knockdowns tested, 17 resulted in >20% inhibition of cellular transformation and 19 resulted in >15% inhibition (Fig. 5b), indicating that the corresponding transcription factors are important for transformation. These factors include the expected STAT3 and NFKB2, but they also include FOSL1, FOSL2, CEBPB, ETS2, MAFB, BHLHE50, KLF5, FHL2, NFE2L2, IRF9, HIF1A, HBP1, TCF3, SOX15, and TCF7L2. As negative controls, knockdowns of 6 factors (FOXO1, SOX4, ATF2, ATF3, SMAD4, and TP53) with middle or low TFscores did not significantly affect cell transformation efficiencies. Thus, TFscore predicts transcription factors that are important for cellular transformation with very good accuracy with  $P$ -value < 0.0005 (Fisher's Exact Test comparing top candidate factors vs. negative controls). Some of these (TCF7L2, HIF1A, NFKB2, SOX15, FHL2, and BHLHE40) do not show high enrichment of binding sites in accessible chromatin regions that are dynamically regulated. These results indicate a surprisingly large number of factors are important for the process of cellular transformation in our model.

#### TFs important for transformation co-regulate common genes.

To examine the effects of individual factors on gene expression, we performed transcriptional profiling (RNA-seq) by using siRNAs to individually knockdown expression of 6 transcription factors (CEBPB, NFE2L2, FOSL1, FOSL2, SOX15, and TCF7L2). When normalized to a control knockdown experiment, 2576

genes show over 2-fold differential expression upon at least one factor knockdown. Remarkably, the transcriptome-scale gene expression patterns in these 6 knockdowns are quite similar, even though the factors have different DNA-binding specificities (Fig. 6a). Knockdowns of FOSL1 and FOSL2 show the most similar gene transcriptional response as compared to other knockdowns, indicating redundancy of transcription factors in the same family (Fig. 6a). For the 6 factors tested, 1428 genes (14% of the total expressed genes) are commonly down-regulated, whereas 87 (0.8% of the total expressed genes) are commonly up-regulated. The similarities in the gene expression profiles for these 6 knockdown experiments are far above random expectation (Fig. 6b). In accord with the relevance of these transcription factors to transformation, the 1428 genes commonly down-regulated upon knockdown are more likely to be up-regulated during transformation, whereas the 87 genes commonly up-regulated upon knockdown are more likely to be down-regulated during transformation (Fig. 6c).

Only 102 (7.1%) of the 1428 genes that are commonly down-regulated upon these factor knockdowns are induced >1.5 fold during cell transformation (Fig. 6a). Those genes encode important regulators of inflammatory response, cellular signaling pathways, and apoptosis (Fig. 6d). For each of these 102 inducible genes, the AccessTF-predicted binding sites in the corresponding promoter/enhancer regions provide strong evidence for which of the 6 factors directly interact with the DNA and affect transcription of the gene (on average, 4.6 AccessTF-predicted binding sites per gene). These enhancers and promoters typically lack one or more motifs, and hence direct binding sites, for transcription factors that nevertheless influence transcription of the gene. Such “non-directly-bound” transcription factors could associate with the promoter/



**Fig. 6** Gene expression network mediated by transcriptional regulators. **a** Genes showing significant differential gene expression upon factor knockdowns, with 2-fold differential expression in at least one knockdown. **b** Fraction of expressed genes showing consistent up- or down-regulation upon 6 factor knockdowns compared to those obtained by randomizing each column of the datasets. **c** Dynamic regulation of gene expression of gene groups in **a** during transformation. The cumulative distribution function fold changes values ( $\log_2(\text{transformed/non-transformed})$ ) is plotted. **d** The gene transcription network mediated by the transcription factors. The edges represent direct factor binding sites in promoters/enhancers of targeted genes. For each pathway, 3 randomly picked genes are shown

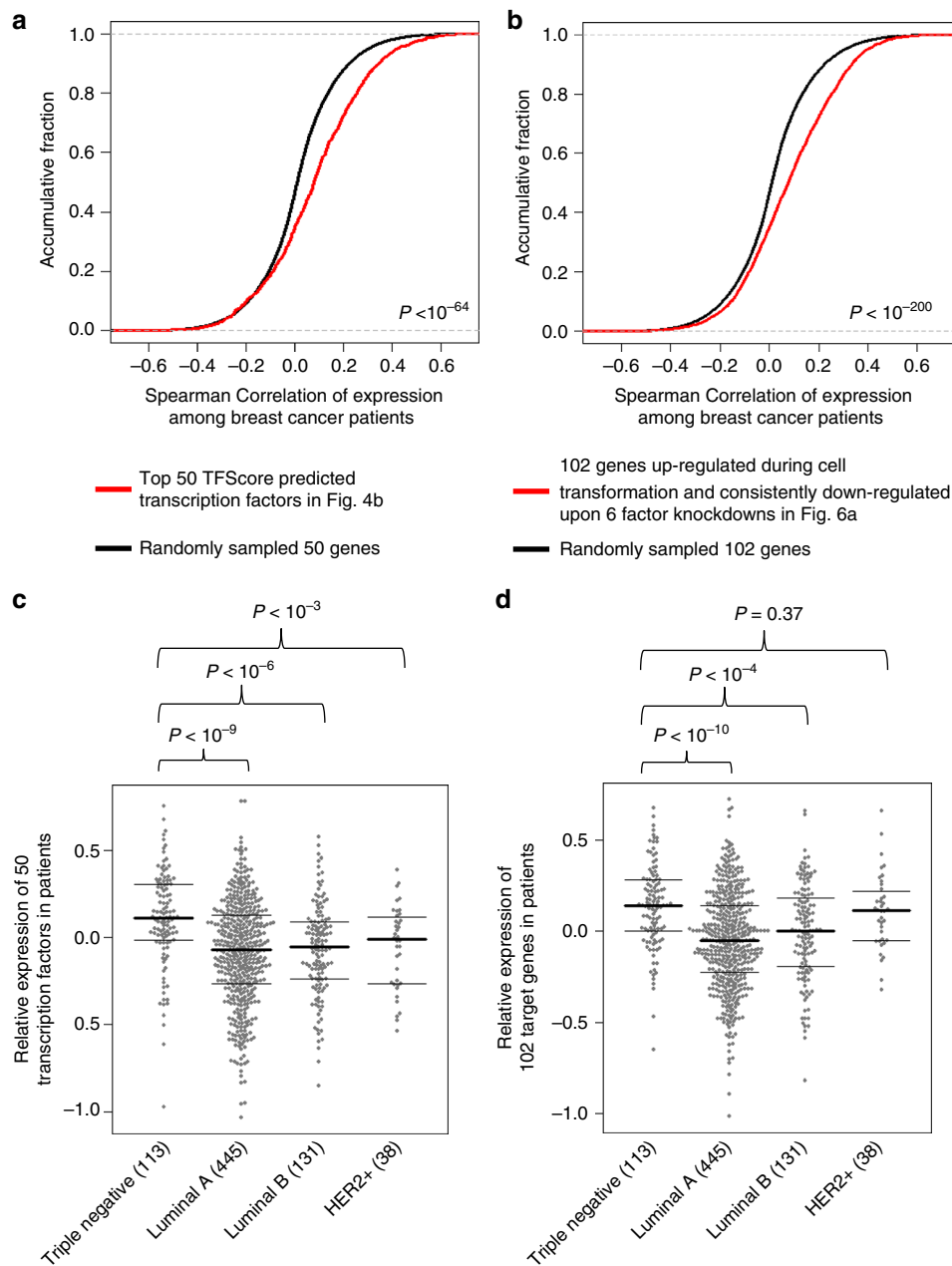
enhancer via protein-protein interactions with other factors directly bound to a different motif. Alternatively, they may indirectly affect transcription of a given gene via effects on other genes that contribute to the cell state.

We performed similar analyses for the remaining 1326 genes that are commonly down-regulated upon factor knockdowns, but are not differentially expressed during cell transformation. Gene ontology analyses show that these genes encode proteins enriched in functions such as RNA processing, cell cycle, RNA translation, and chromatin modification (Fig. 6d and Supplementary Fig. 7). Compared to inducible genes during transformation, these non-

regulated genes are less likely to be direct targets of six factors with 3.2 AccessTF-predicted binding sites per gene (Wilcoxon Rank Sum Test  $P$ -value  $< 10^{-12}$ ). Interestingly, some transcription factors functionally important for cell transformation (HIF1A, ETS2, and FOS) are also common targets.

**Functional TFs and target genes are co-expressed in patients.** Using RNA-seq for the Human Cancer Cell Atlas (TCGA) database, we examined the expression of functional transcription factors and target genes learned from our cellular transformation model in breast cancer patients<sup>37</sup>. Expression of the top 50





**Fig. 7** Expression of TFs and targets in 1182 breast cancer patient samples. **a** The Spearman's correlation of expression levels of top 50 TFScore predicted transcription factors in Fig. 4b. We randomly picked 50 expressed transcription factors and calculated their correlation values as the background distribution. The Wilcoxon Rank Sum test  $P$ -value is shown. **b** The Spearman's Correlation of expression levels of 102 genes up-regulated during cell transformation and consistently down-regulated upon 6 factor knockdowns in Fig. 5a. We randomly picked 102 expressed genes and calculated their correlation values as the background distribution. The Wilcoxon Rank Sum test  $P$ -value is shown. **c** Relative expression of top 50 TFScore predicted transcription factor in indicated breast cancer subtypes. The numbers of patient samples were shown in parentheses, and the Wilcoxon Rank Sum test  $P$ -values comparing different groups are shown. **d** Relative expression of 102 target genes in indicated breast cancer subtypes. The numbers of patient samples were shown in parentheses, and the Wilcoxon Rank Sum test  $P$ -values comparing different groups are shown

TFScore-predicted functional transcription factors in Fig. 4b are positively correlated among 1182 patient samples compared to randomly sampled genes (Fig. 7a). Similarly, the 102 genes that are up-regulated during cell transformation and are consistently down-regulated upon 6 factor knockdowns, show significant positive expression correlation in patient samples as compared to randomly sampled genes (Fig. 7b). In addition, the transcription factors (Fig. 7c) and target genes (Fig. 7d) identified here tend to have higher expression levels in triple negative breast cancers, as

compared to ER-positive breast cancers. Consistent with previous analyses of individual regulatory circuits in our transformation model<sup>24–27</sup>, these data indicate that the functional transcriptional network identified in our cellular transformation model is relevant in human cancer.

## Discussion

We describe a combined experimental and computational approach to comprehensively identify transcription factors that

are important for mediating dynamic changes in gene expression between two physiological states. Experimentally, this approach combines genome-scale mapping of accessible chromatin regions via DNase I hypersensitivity, histone modifications, and transcriptional profiling. Although DNase I hypersensitive regions represent a functional assay for transcription factors bound to these regions, they do not directly identify motifs or other sequences bound by proteins *in vivo*. Instead, bound motifs, and hence the cognate proteins, are inferred. In contrast, genome-scale, DNase I footprinting directly identifies sequences bound by proteins *in vivo*<sup>38–41</sup>. However, while individual sequence reads contribute directly to the identification of DNase hypersensitive regions, numerous sequence reads are necessary to identify underrepresented regions of DNase I cleavage that define DNase footprints. As such, footprinting methods require much higher (~10 times more) sequencing depth and hence are considerably more expensive, especially for experiments involving multiple samples.

Computationally, we first developed a Bayesian network model, AccessTF, to predict protein-binding sites *in vivo* in which information about all known DNA sequence motifs is combined with quantitative measurements of DNase I hypersensitivity centered on the motifs. This approach is advantageous over standard motif analyses of accessible regions. It does not involve arbitrary cut-offs for quality of sequence motifs, it accounts for where the motif is located within the accessible region, it accounts for the level of chromatin accessibility, and it yields binding probabilities for each motif. Most importantly, validation using *in vivo* binding for multiple transcription factors (ChIP-seq data) yields high ROC and PR AUC values for predictions of AccessTF. Of course, any motif- or footprint-based approach cannot identify transcription factors that directly associate with target sites but instead are recruited to such sites via interactions with factors directly bound to the motif.

Secondly, we develop a novel scoring system, TFScore, to identify key transcriptional regulators by integrating the AccessTF-predicted binding sites with four layers of functional information. This integrative approach provides more powerful predictions and identifies more functional regulators, and it doesn't require a factor to meet all four criteria. TFScore yields a rank-ordered list of transcription factors that are predicted to be important for the process of interest. Experimental validation using siRNA-knockdowns indicates that most of the top 50 factors on the list are important for transformation, in contrast to all 6 factors tested with low TFScores. This high validation rate suggests that the computational pipeline should be generally applicable to identify key transcriptional regulators in other biological processes.

In principle, our integrated computational analysis is comprehensive and loosely analogous to a genetic screen, because it gives each known transcription factor a score that predicts its relative importance in transformation. Transcription factors important for cellular transformation have also been identified on a genomic scale by screening shRNA or CRISPR libraries<sup>42,43</sup>. This genetic approach does not identify physiological target sites or transcriptional regulatory circuits, but the identified genes are not restricted to DNA-binding transcription factors. As such, these approaches are complementary.

Our results indicate that numerous transcription factors play a functional role in transformation in a single experimental model. Extrapolation of the result that 85% of the top 50 factors (17 out of 20 tested) affect transformation suggests the involvement of at least 40 transcription factors in over 20 protein families in this oncogenic model. Moreover, it is likely that additional factors further down the list will also be important, although we have not

experimentally determined false discovery rates throughout the list. Some factors identified (e.g., NF- $\kappa$ B, STAT3, FOS) are known to be involved for transformation in our model, others (e.g., CEBPB, HIF1a, ETS2, FHL2, TCF7L2, and NFE2L2) have been described as oncogenes in other settings, and some proteins (BHLHE40 and MAFB) have not been well linked to cancer. Transformation in other cellular models also involves many transcription factors, although not necessarily the same set identified here<sup>42,43</sup>. Similarly, we hypothesize that many transcription factors will play a functional role in individual human cancers, even if only a small number of them are oncogenic drivers. The involvement of numerous transcription factors in a dynamic gene expression program has been observed in dendritic cells responding to pathogens<sup>4,5</sup>, differentiation of Th17 cells<sup>6,7</sup>, and hematopoiesis<sup>8</sup>.

An important observation arising from the siRNA knockdown experiments is that the 6 factors tested affect the expression of a common group of genes. It seems likely that many of the 11 other factors validated to affect transformation will behave in a similar manner. And more broadly, as the 17 factors shown to be important for transformation were selected from and distributed among the top 50 factors, it seems likely that many of them will affect the common group of genes. This would seem to be surprising because the factors recognize different motifs, and different genes within the common group have different combinations and organizations of motifs. However, similar results have been observed in other biological processes<sup>44–46</sup>.

Previously, we described the transformation process in our model as an epigenetic switch from a stable non-transformed state to a stable transformed state mediated by an inflammatory feedback loop<sup>23,24</sup>. This epigenetic switch between stable cell states is analogous to what occurs in cellular differentiation and formation of distinct and stable cell types from a common progenitor<sup>1–3</sup>. A similar epigenetic switch involving an inflammatory feedback loop occurs in a liver cell model of transformation<sup>47</sup>. STAT3 is a critical player in both epigenetic switches, but otherwise the described pathways involved different genes. In both cases, a molecular pathway involving a small number of genes was described.

The comprehensive analysis presented here suggests that this feedback loop is much more extensive, involving numerous transcription factors that control a large and common set of genes. These genes not only include those induced during the cellular transformation process, but also many genes that are constitutively expressed yet are affected in a common fashion by these factors. In this regard, the set of genes induced by a given environmental stress is not well conserved across yeast species, whereas the overall category of genes is highly conserved<sup>48</sup>. By analogy with stable developmental states, we suggest that the critical transcription factors form a stable regulatory loop for each other's expression, thereby leading to a common set of target genes. In this view, transient induction of Src leads to changes in transcription factor activity or levels, and the altered state of transcription factors is self-reinforcing, leading to a new and stable state of gene expression.

Cancer occurs primarily as a consequence of somatic mutations and DNA methylation of tumor suppressor genes, and every cancer is genetically and epigenetically distinct. As such, an epigenetically stable cancer state is presumably not derived from evolutionary selection, but rather reflects a natural state of the wild-type organism. The simplest view is that this natural state represents a de-differentiated state in early development, where rapid growth is important. Thus, we suggest that the regulatory loop that is critical to maintain the stable transformed state in our model is not generated *de novo*, but rather reflects the induction

of a natural de-differentiated, rapid-growth state by a transient inflammatory stimulus.

## Methods

**Cell culture and cell transformation assays.** MCF-10A-ER-*Src* cells were cultured in DMEM/F12 medium with the supplements as previously described<sup>22,23</sup>. Tamoxifen (TAM, Sigma, H7904) 0.4 mM for 24 h was used to transform this inducible cell line when the cells were grown to 30% confluence. The transformation assay that measures growth under low attachment conditions has been described previously<sup>36</sup>.

**Chromatin immunoprecipitation sequence.** ChIP was performed as described previously<sup>33</sup> with some modifications. Cells were treated with ethanol or tamoxifen (1  $\mu$ M) for 24 h and then cross-linked using ethylene glycol bis (succinimidyl succinate) (EGS), disuccinimidyl glutarate (DSG) and disuccinimidyl suberate (DSS) mixture (2.5  $\mu$ M each) for 45 min at room temperature. After this initial crosslinking, cells were further fixed using 1% formaldehyde for 20 min at room temperature and then quenched by glycine (0.125 M). Chromatin in sonication buffer (50 mM HEPES, pH7.5, 140 mM NaCl, 2 mM EDTA, 2 mM EGTA, 1% Triton-100, and 0.4% SDS) was sheared using Branson Microtip Sonifier 450 (12 cycles of 15 s at a sonication setting of output 4 and duty cycle 60%) to a size mostly between 100–150 bp. The sonicated chromatin solution was diluted to 0.085% SDS and immunoprecipitated with antibodies against H3K4me3 (ab8580), H3K27ac (ab4729), H3K4me1 (ab8895), H3K36me3 (ab9050), H3K27me3 (ab6002), and H3K9me3 (ab8898). Immunoprecipitated chromatin was decross-linked using RNase Cocktail (Ambion, AM2286) and Pronase (Roche, 10165921001). ChIP DNA was end repaired, addition of “A” and adapters ligation and PCR amplification to produce ChIP-seq libraries. The DNA concentration was measured by Bioanalyzer before sequencing using HiSeq 2000 at the Bauer Core Facility, Harvard.

**DNase-seq.** The procedures for DNase treatment of chromatin and library preparation have been described previously<sup>10</sup>.

**siRNA transfection and qPCR.** Cells were seeded for 24 h and then transfected with siRNAs, 50 nM (Dharmacon) and Lipofectamine RNAiMax (Life Technologies). siRNA sequences were in Supplementary Data 3. After 24 h, cells were split and treated with either ethanol or Tamoxifen (0.4  $\mu$ M, Sigma-Aldrich, H7904) plus AZD0530 (0.4 nM, Selleck Chemicals, S1006) for 24 h. Total RNA was isolated using mRNeasy Mini Kit (Qiagen, No. 217004). Two microgram RNA was used for SYBR Green based. Primers were listed in Supplementary Data 4.

**RNA-seq library preparation.** Briefly, RNA was extracted using mRNeasy Mini Kit following the manufacturer’s instruction. RNA-seq libraries were prepared using TruSeq Ribo Profile Mammalian Kit (Illumina, RPHMR12126) as per manufacturer’s instruction. RNA-seq libraries were sequenced by Bauer Core Facility using HiSeq 2000.

**Analyses of time-series of mRNA expression data.** We profiled mRNA expression profiles using Affymetrix Human U133 2.0 A expression arrays, at 0 h, 1 h, 2 h, 4 h, 8 h, 12 h, 16 h, and 24 h upon *Src* oncogene induction with two biological replicates at each time point (GSE17941)<sup>22</sup>. The gene expression values were calculated by the RMA approach using Affymetrix Expression Console Software. We used MAS5 algorithm to estimate whether a gene is expressed in a sample and required the genes should be expressed in two biological replicates. Differently expressed genes were selected using the cutoff >1.5 fold change consistently in two biological replicates as compared to control in at least one time point during cell transformation. To test the validity of the cutoff, we randomized the fold-change values across genes 100,000 times, and applied the cutoff to estimate the false discovery rate (FDR) for differentially expressed genes as  $<7 \times 10^{-3}$ . The expression values were then mean-normalized and standardized. We used K-mean clustering to group differentially expressed genes into four coherent clusters, with median Pearson Correlation values >0.7 of genes in each cluster.

**DNase-seq and ChIP-seq data analyses.** Raw Fastq reads were aligned to human reference genome (hg19) using Bowtie<sup>49</sup> allowing up to 2 mismatches. Only the uniquely mappable reads were used for subsequent analyses. For DNase-seq data, we used MACS<sup>50</sup> to call peaks with the cutoff *P*-value  $< 10^{-11}$  in at least one sample and using the following parameters “macs2 callpeak --llocal 1000000 -g 2.7e9”. For ChIP-seq data for H3K27ac, H3K4me3, and H3K4me1, we used MACS<sup>50</sup> to call peaks with the cutoff *P*-value  $< 10^{-8}$  in at least one sample and using the following parameters “macs2 callpeak --llocal 1000000 -g 2.7e9”. For ChIP-seq for H3K27me3, H3K9me3 and H3K36me3, we used SICER<sup>51</sup> to call peaks with the cutoff *E*-value >40, window size 200 bp and gap size 600 bp, which is better for identifying broad read peaks. Then for each data type, we merged overlapping significant peaks from samples in different time points. For each

**Table 1 Bayesian Network Model Parameters**

	Normalized value	Type	Range
PWM score ( $f_i$ )	(Fimo_Score <sub><i>i</i></sub> -10)/10	Continuous	(0, 1)
Distance to TSS ( $d_i$ )	$1/(1+TSS\_dist_i/1000)$	Continuous	(0, 1)
Conservation ( $c_i$ )	PhastCon Score in placenta	Continuous	(0, 1)
TF binding ( $b_i$ )	Hidden	Binary	0, 1
DNase I tag ( $n_i$ )	DNase-seq read # in 200 bp upstream or downstream from the motif	Continuous	(0, +)

merged peak, its expression level in a sample was measured by the Reads per Million (RPM) value.

**Bayesian network model to identify potential functional TFBS.** To identify potential functional TFBS, we considered transcription factors with annotated Position weight matrix (PWM) in human, mouse and rat defined by MotifDB<sup>35</sup>. Based on those PWM, we used FIMO<sup>52</sup> to search potential TFBS in human genome (hg19) with default cutoff *E*-value  $< 10^{-4}$ .

For each potential TFBS *i* in the genome, its binding status is a hidden variable, either bound ( $b_i = 1$ ) or unbound ( $b_i = 0$ ). To estimate the binding probability, we hypothesized that a motif is more likely to be functional and bound by the factor, if it is closer to a transcription start site (TSS), show higher conservation during evolution, is more similar to the consensus sequence (higher PWM score) and is located more accessible chromatin regions. For each TFBS, we calculated its distance to the closest TSS defined by refSeq (TSS\_dist<sub>*i*</sub>), normalized as  $d_i = 1/(1 + TSS\_dist_i/1000)$ . The PWM score is calculated based on Fimo Score, as  $f_i = (\text{Fimo\_Score}_i - 10)/10$ . We used the averaged PhastCons<sup>53</sup> score across 44 placental mammals to measure its conservation level as  $c_i$ . During our analyses, we found motifs located at edges of or close to DNase-seq peaks are less likely to be bound (as determined by ChIP-seq), and can cause false positives in prediction. So, we calculated the number of DNase-seq reads 200 bp upstream and downstream a motif, respectively, and picked the lower number to represent the chromatin accessibility. We used the Bayesian Network Model (Table 1) to estimate the contribution of PWM score ( $f_i$ ), distance to TSS ( $d_i$ ), conservation level ( $c_i$ ), and DNase I tag ( $n_i$ ) to the probability of motif binding ( $P(b_i = 1)$ ), as shown in Fig. 3b.

The contribution of PWM score ( $f_i$ ), distance to TSS ( $d_i$ ) and conservation levels ( $c_i$ ) to TF binding probability ( $y_i = P(b_i = 1)$ ) is modeled by logistic regression:

$$\log\left(\frac{y_i}{1-y_i}\right) = \beta_0 + \beta_1 \times f_i + \beta_2 \times d_i + \beta_3 \times c_i$$

The TF binding probability ( $y_i = P(b_i = 1)$ ) is correlated with the chromatin accessibility, which is measured as the number of DNase-seq tags around the motif ( $n_i$ ). The distribution is modeled by the negative binomial distribution:

$$P(n_i | b_i = 0) = \text{Negative Binomial}(n_i | K_0, r_0) \\ = \frac{(n_i + K_0 - 1)!}{n_i! (K_0 - 1)!} (1 - r_0)^{K_0} r_0^{n_i}$$

$$P(n_i | b_i = 1) = \text{Negative Binomial}(n_i | K_1, r_1) \\ = \frac{(n_i + K_1 - 1)!}{n_i! (K_1 - 1)!} (1 - r_1)^{K_1} r_1^{n_i}$$

The expectation-maximization (EM) algorithm was used to find maximum likelihood estimates of parameters in the model, including  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ,  $K_0$ ,  $r_0$ ,  $K_1$ ,  $r_1$ . We randomly picked 10,000 motifs for training to learn the parameters of AccessTF, and applied the parameters to predict the binding status of another randomly picked 10,000 motifs for testing to evaluate the algorithm performance. We set the prior binding status based on the number of DNase I tags around the motif. Motifs with top 5% number of tags were set as bound and others were set as unbound. We tried different prior probabilities and obtained similar predicted posterior binding probabilities after converge. We picked the selected one as it converges quickest to get the parameter of AccessTF on the training set and the prediction performed the best to predict the motif binding status on the testing set.

To examine the performance of the algorithm in predicting the motif status for AP1 and STAT in ER-*Src* cells, we analyzed ChIP-seq data for AP1 (FOS, JUN and JUNB) and STAT3 factors, respectively. We used MACS<sup>50</sup> to call ChIP-seq peaks using the following parameters “macs2 callpeak --llocal 1000000 -g 2.7e9”, with the cutoff *P*-value  $< 10^{-8}$ . The motifs with defined ChIP-seq peaks were considered as true positive, and those not overlapping with ChIP-seq peaks were considered as true negative.

Using the same analyses procedure, we applied AccessTF to predict transcription factor binding sites in K562 cells using DNase-seq data in ENCODE project<sup>34</sup>. ChIP-seq data for transcription factors in K562 cells were used to measure the algorithm performance.

**TFScore.** We rank-ordered the candidate factors based on 4 following criteria:

(1) Relative contribution of the motif to general increased chromatin accessibility during the cell transformation. If the motif occurrence is higher, the corresponding TF is more likely to be important. For each accessible motif identified in AccessTF, we calculated the sum of reads in 200 nt upstream and downstream, normalized it to total number of reads and obtained the read per million (RPM) value to represent its surrounding chromatin accessibility in a sample. For each PWM, we calculated the sum of fold changes at 6 h and 24 h after Tamoxifen treatment relative to 0 h to represent its contribution to increased chromatin accessibility, using the following scoring definition. The regulation at 6 h: +0 ( $\leq 600$ ), +1 (601–1000); +2 (1001–1500); +3 (1501–3000); +4 (3001–5000); +5 ( $>5001$ ). The regulation at 24 h: +0 ( $\leq 2000$ ), +1 (2001–4000); +2 (4001–15000); +3 (1501–3000); +4 (3001–5000); +5 ( $>5001$ ).

(2) Relative enrichment of motifs in promoter/enhancer regions of differentially expressed gene clusters. A factor is more likely to be important, if the motif enrichment is higher. For each PWM, we assigned the accessible motifs to the nearest closest expressed gene with the distance between the motif and TSS smaller than 100 kb. We also associated the motif and gene if the distance is within 20 kb. We used the Fisher Exact test to check the enrichment of the motifs in differentially expressed gene clusters (Fig. 1b) as compared to expressed genes which do not show differential expression. The  $-\log_{10}$  ( $P$ -value) was used to indicate the relative enrichment, using the following scoring definition: +0 ( $\leq 8$ ); +1 (8–11); +2 (11–14); +3 (14–17); +4 (17–20); +5 ( $>20$ ).

(3) If a transcription factor is significantly up-regulated over 1.5 fold, we added scoring +15. An up-regulated factor is likely to be more important

(4) For transcription factors in the same family which have similar binding motifs, we picked a representative PWM and rank-ordered their relative importance based on their expression levels. A factor expressed at a higher levels if more likely to be important. Suppose the highest expression level of genes in a family is  $E$ , Following is the scoring definition: +5 ( $=E$ ); +3 ( $E/2-E$ ); +0 ( $E/4-E/2$ ); -5 ( $E/6-E/4$ ); -10 ( $\leq E/6$ ).

The final TFScore is the sum of the above four criteria.

**RNA-seq data analyses.** Raw reads were aligned to GENCODE<sup>41</sup> defined transcripts and then human reference genome (hg19) using Tophat<sup>49</sup> allowing up to 2 mismatches. Only the uniquely mappable reads were used for subsequent gene expression analyses. Gene expression levels were calculated as transcripts per million (TPM) value.

**Gene ontology analyses.** The Database for Annotation, Visualization and Integrated Discovery (DAVID)<sup>54</sup> was used for gene ontology analyses.

**TCGA data analyses.** RNA-seq gene expression and genetic annotation data for 1182 breast cancer patients were downloaded from TCGA database<sup>37</sup>. We calculated Spearman's rank correlation coefficient values between gene pairs among transcription factors and target genes. We also randomly selected expressed genes and calculated Spearman's correlation as the background distribution. We grouped breast cancer patients based on their genetic subtypes as following: Triple Negative (ER-, PR-, and HER2-), Luminal A (ER+, PR+, and HER2-), Luminal B (ER+, PR+, and HER2+), and HER2+ (ER-, PR-, and HER2+). To calculate the relative expression of a set of transcription factors and target genes, we first median-normalized the  $\log_2$  gene expression levels across patient samples. And then we took the median normalized values across genes in a gene set to indicate the relative expression level of the gene set in a sample.

**Data availability.** All sequencing data that support the findings of this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) and are accessible through the GEO series accession numbers GSE100259, GSE100255, GSE100257 and GSE100258. All computational codes are available from the authors upon request.

Received: 10 July 2017 Accepted: 26 April 2018

Published online: 25 May 2018

## References

- Johnson, A. D. et al. Lambda repressor and cro- components of an efficient molecular switch. *Nature* **294**, 217–223 (1981).
- Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
- Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).
- Amit, I. et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326**, 257–263 (2009).
- Garber, M. et al. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell* **47**, 810–822 (2012).
- Yosef, N. et al. Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461–468 (2013).
- Ciofani, M. et al. A validated regulatory network for Th17 cell specification. *Cell* **151**, 289–303 (2012).
- Novershtern, N. et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
- Struhl, K. Fundamentally different logic of gene expression in eukaryotes and prokaryotes. *Cell* **98**, 1–4 (1999).
- Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
- Buenostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin. DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
- Yang, A. et al. Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell* **24**, 593–602 (2006).
- Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl Acad. Sci. USA* **107**, 564–568 (2010).
- 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).
- Denny, S. K. et al. Nf1b promotes metastasis through a widespread increase in chromatin accessibility. *Cell* **166**, 328–342 (2016).
- He, H. H. et al. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* **22**, 1015–1025 (2012).
- Vierstra, J., Wang, H., John, S., Sandstrom, R. & Stamatoyannopoulos, J. A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat. Methods* **11**, 66–72 (2014).
- Vierstra, J. & Stamatoyannopoulos, J. A. Genomic footprinting. *Nat. Methods* **13**, 213–221 (2016).
- Blatti, C., Kazemian, M., Wolfe, S., Brodsky, M. & Sinha, S. Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res.* **43**, 3998–4012 (2015).
- Xu, J. et al. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev. Cell* **23**, 796–811 (2012).
- Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl Acad. Sci. USA* **114**, E4914–E4923 (2017).
- Hirsch, H. A. et al. A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. *Cancer Cell* **17**, 348–361 (2010).
- Iliopoulos, D., Hirsch, H. A. & Struhl, K. An epigenetic switch involving NF- $\kappa$ B, lin 28, let-7 microRNA, and IL6 links inflammation to cell transformation. *Cell* **139**, 693–706 (2009).
- Iliopoulos, D., Jaeger, S. A., Hirsch, H. A., Bulyk, M. L. & Struhl, K. STAT3 activation of miR-21 and miR-181b, via PTEN and CYLD, are part of the epigenetic switch linking inflammation to cancer. *Mol. Cell* **39**, 493–506 (2010).
- Iliopoulos, D. et al. Loss of miR-200 inhibition of Suz12 leads to polycomb-mediated repression required for the formation and maintenance of cancer stem cells. *Mol. Cell* **39**, 761–772 (2010).
- Iliopoulos, D., Rotem, A. & Struhl, K. Inhibition of miR-193a expression by Max and RXRa activates K-Ras and PLAU to mediate distinct aspects of cellular transformation. *Cancer Res.* **71**, 5144–5153 (2011).
- Polytarchou, C., Iliopoulos, D. & Struhl, K. An integrated transcriptional regulatory circuit that reinforces the breast cancer stem cell state. *Proc. Natl Acad. Sci. USA* **109**, 14470–14475 (2012).
- Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
- Talbot, D. et al. A dominant control region from the human beta-globin locus conferring integration site-independent gene expression. *Nature* **338**, 352–355 (1989).
- Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
- Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- Loven, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).

33. Fleming, J. D. et al. STAT3 acts through pre-existing nucleosome-depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer. *Epigenetics Chromatin* **8**, 7 (2015).
34. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
35. Shannon, P. MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs v1.12.1 (The R Foundation, 2015).
36. Rotem, A. et al. Alternative to the soft-agar assay that permits high-throughput drug and genetic screens for cellular transformation. *Proc. Natl Acad. Sci. USA* **112**, 5708–5713 (2015).
37. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
38. Raj, A., Shim, H., Gilad, Y., Pritchard, J. K. & Stephens, M. msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLoS ONE* **10**, e0138030 (2015).
39. Gusmao, E. G., Dieterich, C., Zenke, M. & Costa, I. G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* **30**, 3143–3151 (2014).
40. Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods* **13**, 303–309 (2016).
41. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
42. Marcotte, R. et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
43. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
44. Parnas, O. et al. A Genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. *Cell* **162**, 675–686 (2015).
45. Dixit, A. et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
46. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896 (2016).
47. Hatzia Apostolou, M. et al. An HNF1 $\alpha$ -miRNA inflammatory feedback circuit regulates hepatocellular oncogenesis. *Cell* **147**, 1233–1247 (2011).
48. Tirosh, I., Wong, K.-H., Barkai, N. & Struhl, K. Extensive divergence of the yeast stress response through transitions between induced and constitutive activation. *Proc. Natl Acad. Sci. USA* **108**, 16693–16698 (2011).
49. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
50. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
51. Zang, C. et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).
52. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
53. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
54. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

### Acknowledgements

We thank John Stamatoyannopoulos for generating the DNase-seq data and Chaolin Zhang (Columbia University) for generous sharing of the computational codes of the Bayesian Network model. This work was supported by a fellowship from the Postdoc Programme of the German Academic Exchange Service, DAAD, (A.J.), the NIH Ruth L. Kirschstein National Research Service Award for postdoctoral fellowship (C.C.), the Searle Leadership Fund in the Life Sciences from Northwestern University (Z.J.), and research grants CA 107486 (K.S.) and K99 CA 207865 (Z.J.) from the National Institutes of Health.

### Author contributions

Z.J. conceived and performed all the bioinformatic analyses, L.H. performed all the siRNA experiments and RNA-seq library preparation, L.H., A.R., A.J., and C.S.C. performed the ChIP-seq experiments, Z.J., A.R., and K.S. analyzed the experimental and bioinformatic analyses and wrote the paper, and K.S. conceived the experimental design.

### Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-018-04406-2>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018