nature medicine

ARTICLES

https://doi.org/10.1038/s41591-018-0091-x

Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer

Darren K. Patten[1,13], Giacomo Corleone[1,13], Balázs Győrffy[2,3], Ylenia Perone[1], Neil Slaven[1], Iros Barozzi[1], Edina Erdős[4], Alina Saiakhova[5], Kate Goddard[6], Andrea Vingiani[7], Sami Shousha[8], Lőrinc Sándor Pongor[2], Dimitri J. Hadjiminas[8], Gaia Schiavon[9], Peter Barry[10], Carlo Palmieri[11], Raul C. Coombes[1], Peter Scacheri[5], Giancarlo Pruneri[12] and Luca Magnani[1*]

The degree of intrinsic and interpatient phenotypic heterogeneity and its role in tumor evolution is poorly understood. Phenotypic drifts can be transmitted via inheritable transcriptional programs. Cell-type specific transcription is maintained through the activation of epigenetically defined regulatory regions including promoters and enhancers. Here we have annotated the epigenome of 47 primary and metastatic estrogen-receptor (ERα)-positive breast cancer clinical specimens and inferred phenotypic heterogeneity from the regulatory landscape, identifying key regulatory elements commonly shared across patients. Shared regions contain a unique set of regulatory information including the motif for transcription factor YY1. We identify YY1 as a critical determinant of ERα transcriptional activity promoting tumor growth in most luminal patients. YY1 also contributes to the expression of genes mediating resistance to endocrine treatment. Finally, we used H3K27ac levels at active enhancer elements as a surrogate of intra-tumor phenotypic heterogeneity to track the expansion and contraction of phenotypic subpopulations throughout breast cancer progression. By tracking the clonality of SLC9A3R1-positive cells, a bona fide YY1-ERα-regulated gene, we show that endocrine therapies select for phenotypic clones under-represented at diagnosis. Collectively, our data show that epigenetic mechanisms significantly contribute to phenotypic heterogeneity and evolution in systemically treated breast cancer patients.

Breast cancer (BC) is the most common cancer type and the second most frequent cause of cancer-related death in women[1]. Among all BC cases, 70% contain variable amounts of estrogen-receptor (ERα)-positive cells. ERα is central to BC pathogenesis and serves as the target of endocrine therapies (ETs)[2]. ERα-positive BC is subdivided into 'intrinsic' subtypes (luminal A and luminal B[3]) characterized by distinct prognoses, highlighting functional heterogeneity. Recent analyses demonstrate that interpatient heterogeneity is more pervasive (reflected by histological[4], genetic architecture[5] and transcriptional differences[6]), ultimately influencing the long-term response to endocrine treatment[7]. Indeed, 30–40% of ERα BC patients relapse during or after completion of adjuvant ETs. At the time of relapse, ET resistance is commonplace, partly achieved via treatment-specific genetic evolutionary trajectories[8]. Yet, recent studies have shown that driver coding mutations do not significantly change between primary and metastatic luminal BC, with the notable exception of *ESR1* mutations[9], suggesting that alternative non-genetic mechanisms might contribute to BC progression and drug resistance. Parallel to genetic evolution, phenotypic/functional changes driven by epigenetic mechanisms can also contribute to BC progression and ET resistance in cell lines[10]. Epigenetic modifications of histone proteins have been successfully used to map regulatory regions and to annotate non-coding DNA[11,12]. Acetylation of lysine 27 on histone 3 (H3K27ac) is strongly associated with promoters and enhancers of transcriptionally active genes[13–15]. Increasing evidence suggests that epigenetic information can actively transfer gene transcription states across cell division[16–19]. Epigenetic modifications also modulate ERα binding to enhancers by interacting with ERα-associated pioneer factors[20,21]. Nevertheless, little is known about the epigenome of BC patients, its influence on intratumour phenotypic heterogeneity, and its role in BC progression. Here, we show the results of the first systematic investigation of the epigenetic landscape of ERα-positive primary and metastatic BC from 47 individuals. Using H3K27ac chromatin immunoprecipitation coupled with next generation sequencing (ChIP–seq) and ad hoc bioinformatics analyses, we have characterized inter- and intrapatient epigenetic heterogeneity and identified transcription factor (TF) YY1 as a novel key player in ERα-positive BC. Finally, we

[1]Department of Surgery and Cancer, The Imperial Centre for Translational and Experimental Medicine, Imperial College London, London, UK. [2]MTA TTK Lendület Cancer Biomarker Research Group, Institute of Enzymology, Hungarian Academy of Sciences, Budapest, Hungary. [3]Semmelweis University, 2nd Deptartment of Pediatrics, Budapest, Hungary. [4]Department of Biochemistry and Molecular Biology, Genomic Medicine and Bioinformatic Core Facility, University of Debrecen, Debrecen, Hungary. [5]Department of Genetics and Genome Sciences, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA. [6]Department of Breast and General Surgery, Charing Cross Hospital, Imperial College Healthcare NHS Trust, London, UK. [7]Department of Pathology, European Institute of Oncology, Milan, Italy. [8]Centre for Pathology, Department of Medicine, Imperial College London, London, UK. [9]IMED Biotech Unit, AstraZeneca, Cambridge, UK. [10]Department of Breast Surgery, The Royal Marsden NHS Foundation Trust, Sutton, UK. [11]Institute of Translational Medicine University of Liverpool, Clatterbridge Cancer Centre, NHS Foundation Trust, and Royal Liverpool University Hospital, Liverpool, Merseyside, UK. [12]Pathology Department, Fondazione IRCCS Istituto Nazionale Tumori and University of Milan, School of Medicine, Milan, Italy. [13]These authors contributed equally: Darren K. Patten, Giacomo Corleone. *e-mail: l.magnani@imperial.ac.uk

demonstrate that epigenetic mapping can efficiently estimate phenotypic heterogeneity changes throughout BC progression.

Results

Mapping of regulatory regions in primary and metastatic ERα-positive BC. We profiled 55 ERα-positive BC samples with H3K27ac ChIP–seq to build a comprehensive compendium of clinically relevant active regulatory regions (Fig. 1a; primary $n=39$ and metastatic $n=16$) (Fig. 1a, Supplementary Tables 1 and 2 and Supplementary Data 1). H3K27ac-enriched regions were classified into 23,976 proximal promoters and 326,719 enhancers. Of the promoters, 80% were identified by profiling four patients, whereas nearly 40 were needed to reach the same coverage for enhancers, reflecting the 10:1 ratio between captured-enhancers and promoters (Supplementary Fig. 1c). These data are in agreement with enhancers being the main determinants of cell-type-specific transcriptional differences[13,14,22,23]. To gain insights into the penetrance of each regulatory region, we developed a sharing index (SI) (Supplementary Computational Methods) by annotating all enhancers and promoters as a function of the number of patients sharing the H3K27ac signal at each specific location (Supplementary Fig. 1d). This analysis showed that a vast proportion of enhancers are patient-specific ($SI=1$), whereas active promoters typically show higher values of SI (Supplementary Fig. 1d). Collectively, these data demonstrate that enhancers account for the majority of potential epigenetic heterogeneity in ERα-positive BC.

Assessment of phenotypic heterogeneity by enhancer ranking. Genetic heterogeneity is a hallmark of most solid tumors[24] but its impact on phenotypic heterogeneity is characteristically hard to resolve. In agreement, despite extensive inter- and intratumoral genetic heterogeneity[25], the majority of ERα-positive patients benefit from systemic ET[7]. Furthermore, de novo metastatic patients initially respond well to ET, suggesting that genetic heterogeneity on its own cannot explain treatment resistance and response. Of note, phenotypic hierarchies can override genetic hierarchies in brain cancers[26,27], suggesting that inheritable epigenetic programs might contribute to phenotypic heterogeneity and treatment outcome. Phenotypic heterogeneity in breast cancer has been known for decades. For example, immunohistochemistry (IHC) assessment of the proportion of ERα-positive cells in single biopsies varies on a continuum from less than 1% to nearly 100% (ref. [28]). However, IHC can test only a few targets in each sample, and deconvolution from bulk transcriptional data is technically unfeasible (Fig. 1b). For instance, cells with focal gene amplification have higher bulk gene expression, but individual cells contribute stochastic discrete amounts, as shown by single-molecule single-cell RNA fluorescence in situ hybridization (FISH)[8]. Conversely, recent evidence has shown that the signal captured by one-way reaction chromatin assays such as the assay for transposase-accessible chromatin using sequencing (ATAC–seq) appears to be linearly proportional to the cells contributing to it[29]. Histone modifications can also be thought of as digital information, with each single nucleosome being on (K27ac) or off at any given time (Fig. 1b). Notably, even within genetically clonal cell lines, the H3K27ac signal varies considerably between different regulatory regions. Regulatory regions labelled as super enhancers, for example, have 10 to 100 times more H3K27ac signal than typical enhancers[14]. What accounts for the variation in signal is not known, but one possibility is that heterogeneity within the cell population (either clonal or subclonal) contributes to the signal intensity. Although other factors might partially contribute to variation in the signal (local antibody affinity, histone dynamic, cell cycle, sonication efficiency, dinucleotide content, mappability and copy number aberrations; see Supplementary Computational Methods and Supplementary Figs. 2–4), we propose that the ChIP–seq signal is robustly positively correlated with the number of contributing cells with a logistic relationship. Super-enhancers might represent regulatory regions active across most cells within a population at any given time (clonal, C peaks), while 'typical' enhancers with lower H3K27ac signal may represent subclones (S peaks, Fig. 1b). This interpretation is conceptually similar to using variant allele frequencies to infer genetic heterogeneity.

Phenotypical heterogeneity might be the consequence of heterogeneous cell populations (i.e., stromal, immune and cancer cells) or actual cancer-specific epigenetic subclones. As our ChIP–seq data are derived from samples with high tumor burden, we hypothesized that the H3K27ac signal could allow for a qualitative assessment of phenotypic heterogeneity (Fig. 1b). To test the relationship between clonality and ChIP signal we performed spike-in experiments in which known numbers of cells with well characterized enhancer activity (MCF7:on, MCF7-F:off) and similar genetic background[10] were admixed in incremental proportions before H3K27ac ChIP–qPCR. The data show that H3K27ac enrichment is positively correlated to the number of cells in the absence of genomic differences (Fig. 1c). These findings were corroborated by an independent analysis using a different antibody (ERα) (Supplementary Fig. 5). As the signal between different patients is not directly comparable, we quantile-normalized the data, assigning to each H3K27ac signal a rank index (RI: 1–100, strongest to weakest; Supplementary Computational Methods and Supplementary Fig. 6a). The signal from a low RI (C peaks) is then associated with clonal regulatory regions active in almost all cells. Conversely, a high RI (S peaks) mark more heterogeneous/subclonal enhancer activity. On investigating the relationship between RI and SI (Supplementary Computational Methods) we found an extremely robust correlation between the two parameters (Fig. 1d and Supplementary Fig. 6b), suggesting that clonal regulatory regions are more common between patients (low RI/high SI) whereas subclonal regulatory elements are more patient-specific (high RI/low SI). For follow-up analysis we split the enhancer elements into two main subgroups ($SI<21$ and $SI\geq21$) based on the hypothesis that $SI\geq21$ might more strongly contribute to the population phenotype.

Enhancers are associated with BC risk and SNP and control gene transcription. Previous analyses from ERα BC cell lines have shown that a genetic predisposition to BC might occur through single nucleotide polymorphisms (SNPs) that modulate TFs binding at enhancers (FOXA1 and ERα[30]). We tested the relationship between regulatory regions captured in patients and DNA risk variants specifically associated with BC through a genome-wide association study (GWAS)[30–32]. Almost the totality of known BC risk variants from two independent data sets overlapped with our H3K27ac database. This overlap is highly significant specifically for enhancers, and not for other annotations (Fig. 1f,g). Notably, this association is not replicated for colorectal cancer risk variants, suggesting that these enhancers might play a specific role in BC development (Fig. 1f). Currently, our patient-derived enhancer data set represents the most enriched annotation for GWAS variants in BC. Next, we assessed the relationship between estimated enhancer clonality and transcriptional output. As the average expression is a function of the number of cells engaged in active transcription and the number of RNA molecule within each cell[33], assuming a stochastic single-cell contribution, bulk mRNA levels should positively correlate with the number of transcribing cells. We could then test if clonal enhancers active in the majority of cells correlate with higher RNA levels. We thus linked enhancers to their potential target genes using CTCF insulated boundaries[34], and analysed three independent BC expression data sets[5,6,35] as a function of RI/SI indices. Our analyses support the hypothesis that genes associated with clonal enhancers have higher bulk RNA levels (Supplementary Fig. 7a). We observed more modest associations when analysing the transcriptome from normal breast tissue (Supplementary Fig. 7a, small insets),

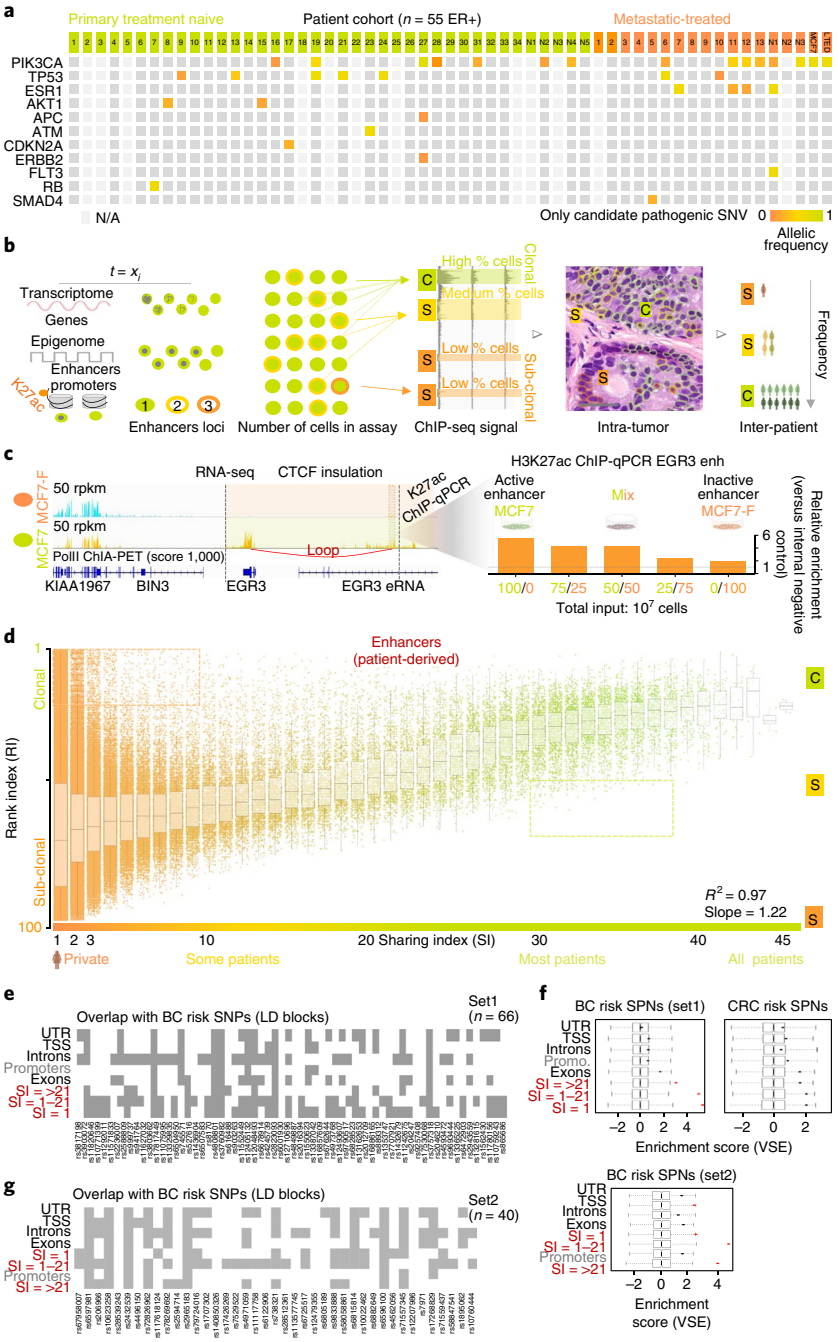

Fig. 1 | Assessment of inter- and intratumour epigenetic heterogeneity. a, Mutational analysis for common cancer driver genes in the patient cohort selected for the study. **b**, Main hypothesis of the study. RNA is ultimately an analogue signal in which each individual cell, at any given time, can contribute a stochastic amount of RNA, while transcriptional data from bulk tissue represent an average over a million cells. For chromatin data, at any given time ($t = X_i$), each cell can only contribute a deterministic value to the bulk signal, generally from two alleles. Therefore, the relative strength of ChIP–seq data is dependent on the number of cells carrying an epigenetic signal at discrete loci. C and S represent strong and medium/weak signal, respectively. Scale bars, 50 μm. Clonal regulatory regions are commonly shared by BC patients, whereas weak enhancers are more patient-specific. **c**, EGR3 mRNA is expressed in MCF7 but not derivative MCF7-F cells. eRNA and Pol-II ChIA-PET show enhancer activity in MC7 but not MCF7-F. CTCF insulated perimeter is shown in yellow. Predicted looping from ChIA-PET is shown in red. The observed ChIP–qPCR signal for H3K27ac at EGR3 enhancers increases with increasing number of MCF7 cells mixed in the sample. Similar results were obtained from three independent experiments. **d**, Linear regression shows that clonal enhancers are commonly shared between BC patients. Of note, a small but discrete proportion of promoters/enhancers escape this general trend of having extremely low RI despite being patient-specific or higher RI while being shared (dotted areas). *y* axis = RI, *x* axis = SI. SI indicates the number of patients sharing the regulatory region. Each dot represents the median RI (all patients) for each single enhancer. Boxplots show the median RI value and interquartile ranges for regulatory regions with the same SI. **e**, Overlap between BC risk variants and annotated DNA elements. **f**, Variant set enrichment analysis, indicating that BC-specific but not CRC-specific GWAS risk variants occur more frequently than expected within the enhancer elements identified in our study. **g**, Overlap with annotated DNA elements and variant set enrichment analysis for the most recent independent set of BC risk variants. SNV, somatic nucleotide variant; RPKM, reads per kilobase millions of sequenced reads; LD, linkage disequilibrium; CRC, colorectal cancer; UTR, untranslated region; TSS, transcriptional start site.

suggesting that our analysis has identified a subset of regulatory regions associated with malignant outgrowth. These data indicate that transcripts identified as disregulated in BC might reflect changes in the size of phenotypic subpopulations between the heterogeneous normal tissue and a cancer population dominated by epithelial features. Collectively, our data show that enhancer activity strongly tracks transcriptional changes in BC patients.

Imputed TFs landscape of ERα BC patients. Enhancers store regulatory information in the form of TF binding motifs[36]. The vast majority of TFs require accessible chromatin to bind their cognate DNA sequences[37]. To extrapolate the TFs landscape from our data we integrated the DNaseI signal (DHS) from 129 cell lines with inferred nucleosome patterns obtained from the H3K27ac signal (Fig. 2a, Supplementary Computational Methods and Supplementary Fig. 7b). As expected, this analysis could identify well-known BC TFs according to their promoter–enhancer bias (Supplementary Fig. 7c). Applying TF motif analysis to regulatory regions defined by the same SI followed by unsupervised clustering identified two major clades (Supplementary Fig. 8). Remarkably, high and low SI clustered together, suggesting that putative clonal and subclonal enhancers contain distinct regulatory information (Supplementary Fig. 8). Functional TF binding is often associated with TF leaving a footprint within chromatin accessible regions[36,38]. Analysing footprints as a function of RI in ERα-positive MCF7 BC cells revealed that enhancers with RI < 20 accumulate more footprints than expected (Fig. 2b). These data show that clonal enhancers might recruit TFs with longer residence time[38]. Unexpectedly, we find estrogen-response element (ERE) motifs significantly enriched only in low SI subclonal enhancers (Fig. 2e and Supplementary Fig. 8). By integrating in vivo ERα binding[39] with our data set we find that the proportion of binding sites increases with SI for enhancers (Fig. 2c) but not for promoters (Fig. 2c), consistent with ERα preferential binding at enhancer elements[40]. These data imply that shared enhancers have a strong propensity for ERα binding, despite being generally under-represented in EREs. Interestingly, although the majority of ERα binding events appear to be patient-specific (ERα SI = 1), 0.003% of ERα are shared across most primary and metastatic patients (484 core-ERα)[39] (Fig. 2d). Together, these data support TF imaging data indicating that only a small fraction of ERα-binding events with longer residency time are functional[38]. We therefore conclude that the largest portion of ERα binding identified in patients occurs at patient-specific, subclonal enhancers and might reflect transient ERα–DNA interactions occurring while the receptor scans the genome[38]. The discrepancy between the small amount of highly shared ERα core binding and the observation of ERE-poor clonal enhancers led us to hypothesize that other TFs might collaborate with ERα to increase its transcriptional efficiency at clonal enhancers. Examining the bias of TF motifs towards high SI enhancers we identified YY1 as the top candidate (Fig. 2e). YY1 is also the top ranked motif within the footprints of clonal MCF7 enhancers (Fig. 2b). It has recently been implicated in the de novo formation of enhancer promoter looping during neural development[41,42] and the MYC-like ability to potentiate gene expression[43], indicating a potential role in modulating the enhancer landscape in ERα-positive BC.

YY1 enhancer activity marks a dominant phenotypic clone in BC. YY1 is an ubiquitously expressed TF (Supplementary Fig. 9a,b) that can act as an activator or repressor by binding DNA, RNA and chromatin modifiers[44,45]. Interestingly, the YY1 Drosophila homologue PhoRC is involved in epigenetic memory by recruiting the Polycomb repressor complex to sequence specific regions[46], but the role of YY1 in mammals is only partially understood. Collectively, our analyses predict that most BC cells should be YY1-positive, so the enhancer driving YY1 should be clonal. To test this, we identified three bona fide enhancers looping at the YY1 promoter using 3D chromatin data[47] (Supplementary Fig. 10a). Enhancer A (SI = 41) directly interacts with enhancer B–C, suggesting a multi-enhancers interaction with the YY1 promoter. Enhancer A consistently ranks among the most clonal enhancers in our data set (Fig. 3a). By comparison, YY1 enhancer A activity is more variable in most normal tissues profiled by H3K27ac within the Epigenome Roadmap consortium[11], implying that some tissues might harbour YY1 subclonal subpopulations (Fig. 3b). Consistent with these predictions, immunocytochemistry (IHC) meta-analysis (Fig. 3b) shows subclonal YY1-positive populations in tissue with high RI (Fig. 3b and Supplementary Fig. 10b). To directly test the regulatory potential of enhancer A, we used CRISPR–Cas9-mediated deletion to generate enhancer-KO (knock-out) ERα positive MCF7 cells (eKO cells, Fig. 3c). Deletion of 2/5 alleles directly reduces the YY1 mRNA level by 30–35% (Fig. 3d). Collectively, these data show that enhancer ranking can capture qualitative changes in intratumoral heterogeneity, and that YY1 enhancer activity marks a dominant phenotypic clone in ERα-positive BC.

Tumor tissues generally have a significantly higher expression level for YY1 compared to normal tissues (Supplementary Fig. 11a). This observation was replicated in an independent BC data set (Fig. 3e and Supplementary Fig. 11b). These data suggest that BC lesions might contain a larger fraction of YY1-positive cells than normal breast tissue (Fig. 3b). Meta-analysis of the METABRIC[5] data sets showed that ERα-positive patients with higher bulk YY1 mRNA at diagnosis have significantly worse outcomes, but this does not hold true for ERα-negative patients (Fig. 3e). The prognostic value of YY1 in ERα-positive patients is maintained when adjusting for other clinical features (Fig. 3e). To test if increased YY1 mRNA levels could be driven by an expansion of YY1-positive cells from a more heterogeneous population, we stained normal breast tissue sections for IHC. Our data show that lobules and ducts contain distinct YY1-positive subclonal populations, whereas nearby tumor tissue is overwhelmingly YY1-positive (Fig. 3f,g). Interestingly, YY1 staining was absent or limited in specimens from patients with a different subtype of BC, specifically triple negative breast cancer (Supplementary Fig. 11c).

YY1 modulates functional ERα binding at enhancer regions. To gain mechanistic insights into the role of YY1 we performed ChIP–Seq in estrogen-deprived and estrogen-stimulated luminal BC MCF7 cells. In the absence of estrogen, YY1 occupies a small set of enhancers and promoters near housekeeping genes (Fig. 4a). Strikingly, estrogen stimulation induced a 23-fold expansion of the YY1 binding repertoire, mostly at enhancer regions associated with ERα-BC signatures (Fig. 4a). Orthogonal analyses showed that induced YY1 binding involves almost all MCF7 active regulatory regions and is strongly associated with H3K27ac marks (Fig. 4b). Conversely, YY1 binding is absent from silenced genes (Supplementary Fig. 12a), demonstrating that YY1 does not associate with PRC2-mediated repression in BC cells. Our in vivo analyses suggest that YY1-motif-enriched enhancers are generally deprived of EREs (Fig. 2b). In agreement, our in vitro data show only a marginal overlap between YY1 and ERα or its pioneer factor FOXA1 (Fig. 4b,c). Nevertheless, YY1, ERα and FOXA1 co-localization becomes significant at core-ERα loci in MCF7 cells (Fig. 4c). Similar observations were made by comparing YY1 overlap with patient-derived ERα binding site analyses (Fig. 4d). In addition, we found that genes defining the luminal subtype in The Cancer Genome Atlas (TCGA) patients are significantly associated with YY1-ERα core binding but not patient-unique ERα (Fig. 4e). Overall, these data further suggest that YY1 might contribute to ERα binding transcriptional output at a small subset of enhancers captured in most tumor cells and most patients. We further show that YY1 depletion is sufficient to abrogate transcription from an ERα-driven reporter (Fig. 4f).

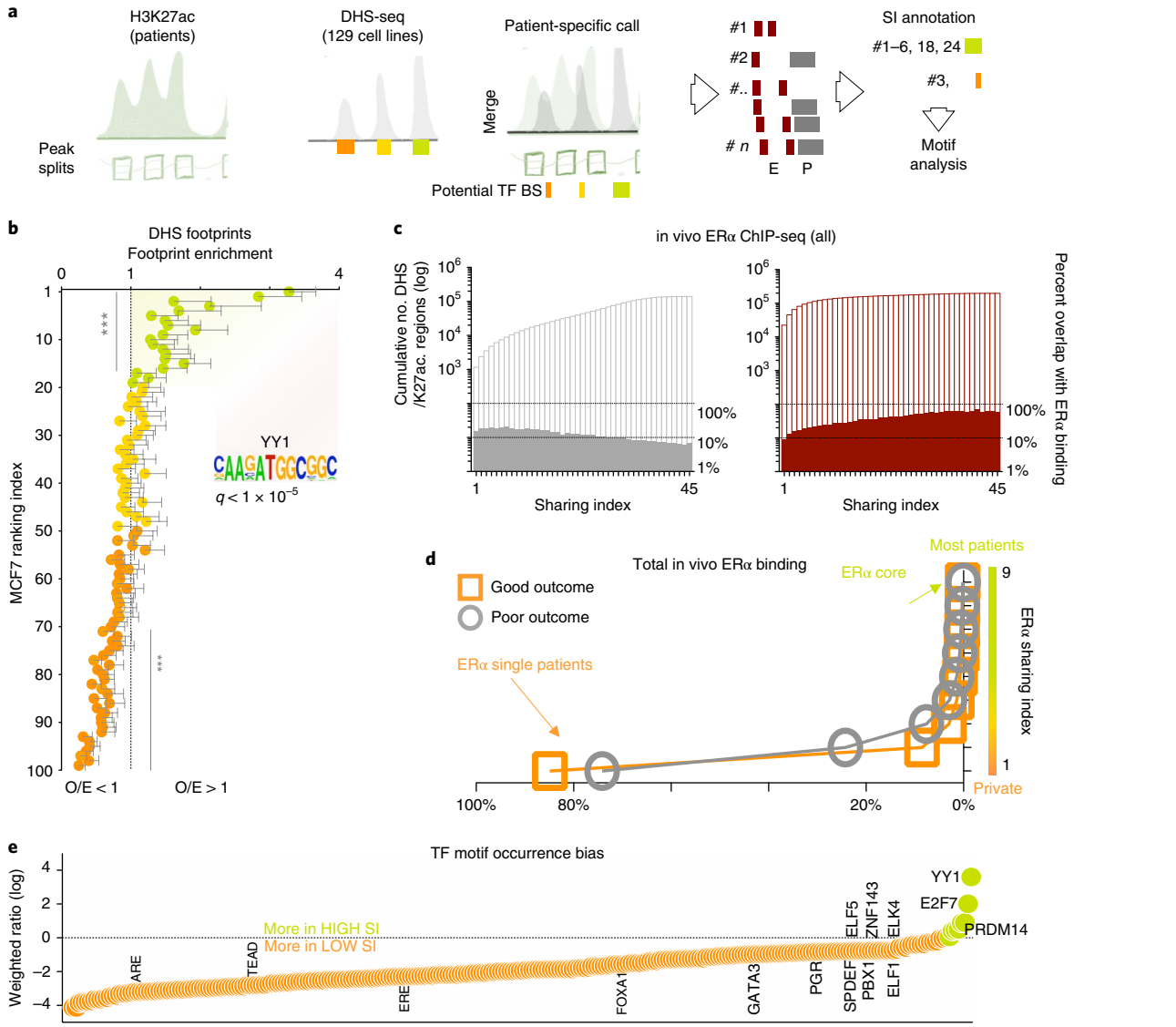

Fig. 2 | Clonal and subclonal regulatory regions contain distinct regulatory information. a, Bioinformatic framework of the analyses. H3K27ac calls were split to identify approximate nucleosome-level enrichment (subpeaks). Subpeak data were integrated with ENCODE-derived DHS-seq calls to identify potential sites of TF binding (BS, binding sites). Individual imputed DHS regions were assigned SI values based on the number of patients sharing the region. **b**, Clonal enhancers in MCF7 cells (RI < 20) are characterized by a higher number of TF footprints, while subclonal enhancers (RI > 70) have less footprint than expected. Each bin contains 34 nucleosome-free regions. The number of footprints (from Wellington) was normalized for enhancer size. Observed (O)/expected (E) values were calculated by dividing the number of normalized footprints in each enhancer by the overall average (2.7 footprints per enhancer). Each value and error bar represent average footprint and 95% CI. Asterisks represent a *P* values of <0.001 in a Wilcoxon signed rank test. **c**, Overlap of imputed DHS regions with in vivo derived ERα binding sites. Left *y* axis: cumulative DHS regions. Right *y* axes: percentage of overlap based on total DHS in each SI bin. **d**, Distribution plot of in vivo derived ERα binding sites versus the number of patients in which they were observed. **e**, YY1 motif is enriched in putatively clonal enhancers in luminal BC patients. TF motifs within imputed DHS are plotted based on their bias towards highly shared enhancers (green) or more private enhancers (orange).

YY1 depletion also abrogates cell proliferation in response to estrogen stimulation in MCF7 (Fig. 4g), suggesting that YY1 is a direct driver of the clonal proliferation observed in BC (Fig. 3d,e). These observations were replicated in independent luminal BC cell models (ZR75 and T47D, Supplementary Fig. 12b,c). YY1 depletion leads to significant downregulation of core-ERα target genes in luminal BC cell line models (Supplementary Fig. 12d). Finally, monitoring cell proliferation at the single cell level using eKO cell lines, we show that deletion of YY1 enhancer A is sufficient to reduce MCF7 growth in estrogen-supplemented conditions (Fig. 4h). Collectively these data identify YY1 as a novel essential TF significantly contributing to ERα regulatory network transcriptional activity.

YY1 contributes to endocrine resistance in luminal BC. YY1-positive cells appear to dominate both primary and metastatic lesions in luminal patients, suggesting that this might remain important even after ET (Fig. 3a). YY1 depletion is indeed sufficient to abrogate proliferation in long-term estrogen-deprived (LTED) cells, an MCF7-derivative mimicking AI-treated BC cells[10](Fig. 4i). Interestingly, LTED cells have an expanded repertoire of ERα binding compared to MCF7, fuelled by endogenous ligands[8,10]. Nonetheless, YY1 and ERα overlap remains restricted to a minority of sites (Supplementary Fig. 13a). Intriguingly, the set of enhancers engaged by ERα and YY1 in LTED cells is radically different from MCF7, with the majority of ERα-YY1 being specific to

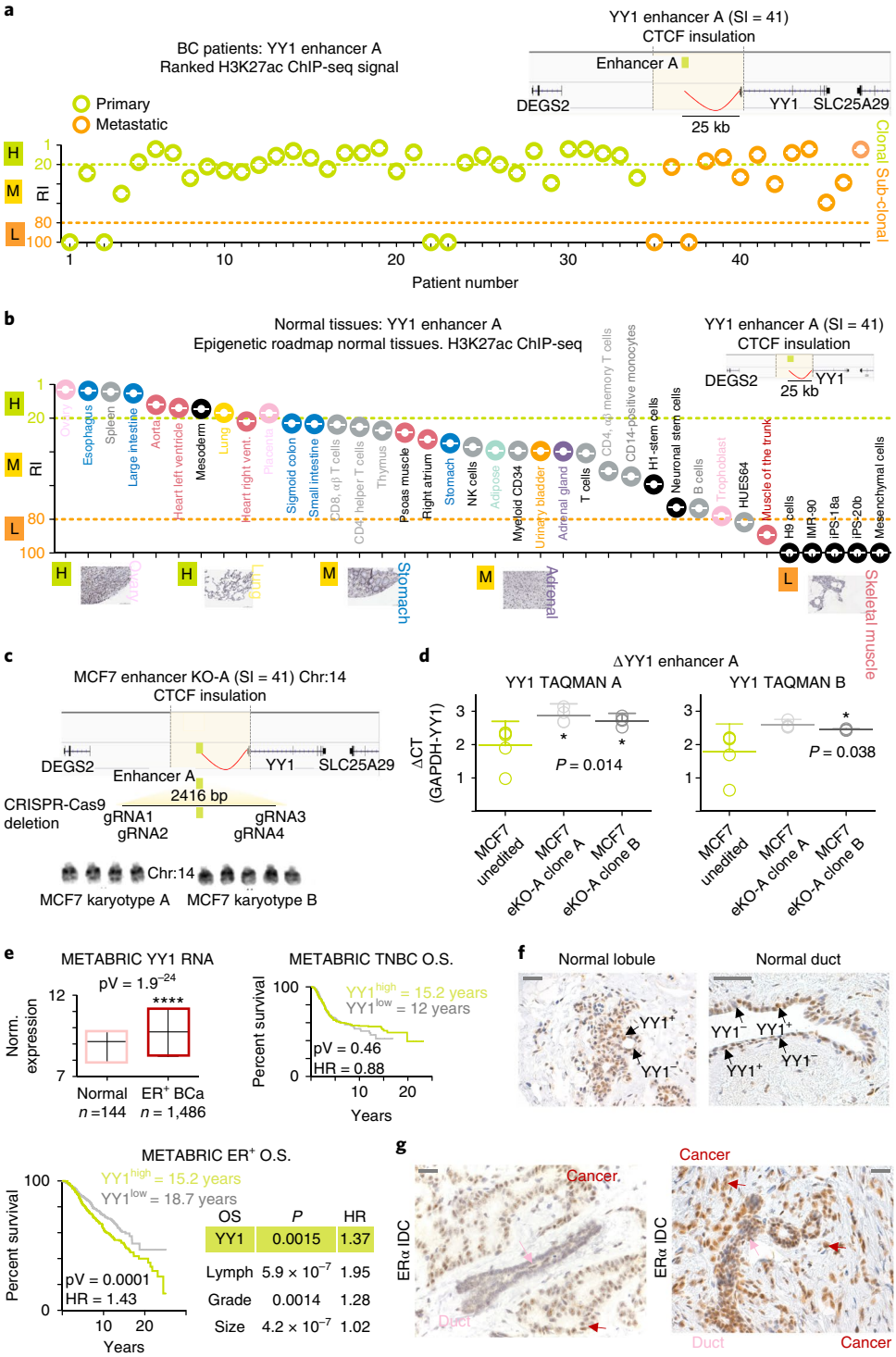

Fig. 3 | YY1 identifies a dominant phenotypic clone in ERα BC. a, RIs for the YY1 enhancer within all individual patients included in the current study. YY1 enhancer locations with 3D interactions are shown in the top right inset. **b**, YY1 enhancer ranking analysis of available Epigenome Roadmap H3K27ac data sets. Tissues are displayed from the strongest to weakest YY1 enhancer activity (based on RI). Representative IHC analysis of normal tissues stained with a YY1 antibody are also shown. **c**, eKO cell lines were generated by deleting 2.4 kb containing YY1-A enhancers in MCF7 cells. Actual karyotyping (bottom) was performed on 10 individual cells. **d**, YY1 expression in control and eKO cell lines was measured using RT–qPCR. Lines and error bars represent average and 95% CI of five independent experiments. Significance was calculated with a one-way ANOVA followed by Tukey's test. **e**, Top left, YY1 expression in ERα-positive BC compared to normal breast tissue. Median, lowest and highest values are reported. Top right, YY1 prognostic value in triple negative BCs. Bottom left, YY1 prognostic value in luminal BCs (CI: 1.19–1.76). Bottom right, Multivariate correction for the luminal BC data set. Analyses included 1,476 ERα-positive and 432 ERα-negative patients. Comparison of survival curves was performed using a log-rank (Mantel-Cox) test. **f**, IHC analysis of normal breast tissues highlights YY1 functional subclones in normal breast. Similar results were observed in 10 independent clinical specimens from independent individuals. **g**, IHC analysis of ERα-positive invasive ductal carcinomas identifies YY1-positive clones as the dominant clonal population. Scale bars, 50 μm. HR, hazard ratio; Pv, *P* value; OS, overall survival; IDC, invasive ductal carcinoma.

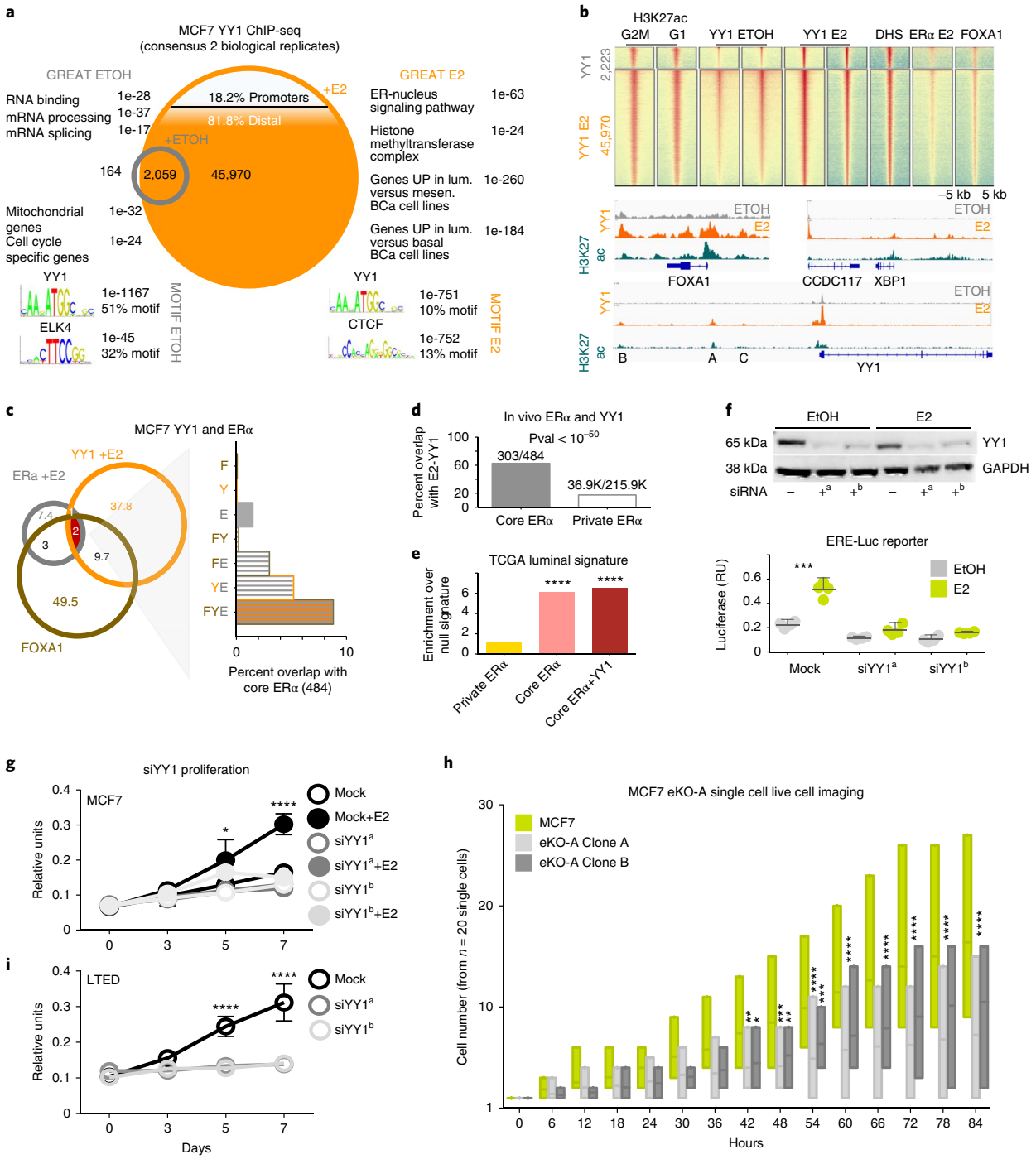

Fig. 4 | YY1 marks critical enhancers in BC cells. a, ChIP–seq data from ERα-positive MCF7 for YY1 in quiescent or 17β-estradiol (E2) stimulated cells. **b**, Heatmaps showing global enrichment profiles of several chromatin markers associated with active regulatory regions in MCF7 cells. **c**, Overlap between ERα, YY1 and FOXA1 in MCF7 cells. Right, Potential overlap with in vivo derived core ERα binding sites. **d**, ERα core binding sites are strongly enriched for YY1 binding in MCF7 cells, while patient-specific ERα bindings are generally YY1-free. Proportions were compared using Fisher's exact test. **e**, Genes used to classify luminal BC patients are strongly enriched for ERα-YY1 binding sites. Asterisks represent $P<1\times10^{-5}$ in a Fisher's exact test versus private ERα . **f**, YY1 depletion leads to transcriptional shutdown of an ERE-driven luciferase reporter. The blot has been cropped (for full blot see Supplementary Fig. 12b). Bottom, bars and error bars represent the average and 95% CI of four independent experiments. Asterisks represent significance at $P<0.001$ after ANOVA with Dunnet's correction. RU, relative unit. **g**, Silencing YY1 blocks estrogen-induced growth in MCF7 cells. Proliferation assays were conducted in three independent biological replicates. Symbols and error bars indicate average and 95% CIs. Significance was calculated using a two-way ANOVA with Bonferroni's correction. **h**, YY1-A enhancer deletion directly leads to reduced proliferation in MCF7 cells. Bars represent highest, lowest and median count for cell numbers from individual colonies ($n=20$) monitored individually using single-cell live imaging. Significance was calculated using a two-way ANOVA with Bonferroni's correction Asterisks represent significant differences after ANOVA followed by Dunnet's test $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$, $^{****}P<0.0001$. **i**, Silencing YY1 blocks growth in LTED cells. Proliferation assays were conducted in three independent biological replicates. Symbols and error bars indicate average and 95% CIs. Significance was calculated using a two-way ANOVA with Bonferroni's correction. Letters [a] and [b] represent different siRNA molecules used in the experiments

each cell type (Supplementary Fig. 13a). ERα-YY1 bound enhancers in LTED strongly associate with the transcription of genes involved with acquired ET, suggesting that during epigenetic reprogramming, YY1 might stabilize ERα to LTED-specific enhancers (Supplementary Fig. 13b). Previous studies have shown that the transcription of a small set of estrogen-activated genes is not antagonized by current ETs[48]. Examining the regulatory landscape near these genes we found an ever increasing association with ERα-YY1 bound enhancers, especially with core ERα-YY1 (Supplementary Fig. 13c). Collectively, these data strongly support the role of YY1 in ERα BC growth and progression.

YY1-ERα promotes SCL9A3R1 expression despite endocrine treatment. By ranking the set of endocrine unresponsive genes bound by YY1-ERα for gene-specific prognostic power calculated in patients treated with ETs[35], we identified SLC9A3R1 as a potential driver of ET resistance (Fig. 5a). *SLC9A3R1* (*NHERF1/EBP50*) encodes a Na/H exchanger regulatory cofactor with a potential role in metastatic invasion[49]. High expression of SLC9A3R1 independently correlates with poor survival in other ERα-BC data sets (Supplementary Fig. 14a). Despite being an ERα target, SLC9A3R1 expression is not suppressed by tamoxifen in MCF7 cells[48]. Additionally, SLC9A3R1 remains transcriptionally active in most ET-resistant BC cell lines that retain ERα expression (Supplementary Fig. 14b–e), demonstrating that ERα activity remains critical for SLC9A3R1 expression. In vivo SLC9A3R1 expression is also unaffected by neo-adjuvant AI treatment (Fig. 5b). Notably, bulk RNA–seq data from a panel of cancer cell lines demonstrate that ERα-positive BC cells have the highest levels of SLC9A3R1 mRNA (Supplementary Fig. 15a). More importantly, TCGA RNA–seq analysis shows that SLC9A3R1 expression is higher specifically in ERα-positive BC patients compared to normal tissue or other subtypes (Supplementary Fig. 15b). Chromatin analyses of MCF7 and LTED cells identify three potential enhancers within the insulated SLC9A3R1 locus (E1–E3). Interestingly, E1–E2 enhancers loop to the SLC9A3R1 promoter and are characterized by high SI, YY1/core-ERα binding sites (Supplementary Fig. 15c). In vivo transcriptional analysis demonstrates that SLC9A3R1 is the only gene near the E1–E2 enhancers that shows a significant increase in bulk RNA level when comparing normal breast tissue with ERα–positive BC (Supplementary Fig. 15d). Remarkably, enhancer activity appears to be resistant to ETs (Supplementary Fig. 15c). Furthermore, SLC9A3R1 expression is dependent on YY1 (Supplementary Fig. 16a), demonstrating that both ERα and YY1 are essential for full enhancer activity. Collectively, these data demonstrate that SLC9A3R1 expression is driven by a BC-specific YY1-ERα bound enhancer. Silencing SLC9A3R1 is sufficient to reduce estrogen-induced growth in ERα-positive cells (Fig. 5c). Intriguingly, SLC9A3R1 is not essential for a second ERα-positive model (T47D) but appears to be a critical gene for both AI-resistant cells models (Fig. 5c and Supplementary Fig. 16b). Overall, these data identify SLC9A3R1 as a novel player involved in ET resistance, the function of which remains to be elucidated.

Mapping phenotypic heterogeneity using YY1 and SLC9AR1 enhancer activity. SLC9A3R1 enhancer activity (E1–E2, SI = 34, RI ≥ 20) indicates that SLC9A3R1 marks subclonal populations in most primary patients (Fig. 5d). Meta-analysis of SLC9A3R1 enhancer activity (RI) within the ENCODE H3K27ac data sets indicates that MCF7 cells are the only cancer cells containing a clonal SLC9A3R1 population (Supplementary Fig. 16c). Of note, the size of the subclonal population correlates with total RNA content for cells contained in both assays, suggesting that the decreasing bulk RNA signal is driven by a progressively smaller subpopulation (Supplementary Fig. 16c). Similar analyses of YY1 enhancers indicate that cancer cell lines are prevalently clonal for YY1 expression (Supplementary Fig. 16d) whereas both YY1 and SLC9A3R1 RIs in mammary epithelial cells predict smaller subclonal populations. These observations fit extremely well with experimental data from IHC profiles from normal and malignant breast (Fig. 3d and Supplementary Fig. 11c). Meta-analyses from the Epigenome Roadmap predict mainly SLC9A3R1-positive subclonal populations, with the exception of gastrointestinal tissues, and these data fit well with RNA–seq measurements from independent cohorts (Fig. 5e and Supplementary Fig. 17a). Analogous to YY1 analysis, SLC9A3R1 IHC data identifies decreasing SLC9A3R1-positive cells in specimens characterized by increasing RI scores (Fig. 5e and Supplementary Fig. 17b). To validate that the RI index can estimate phenotypic clones, we retrospectively collected available biopsies for the BC patients profiled with H3K27ac ChIP–seq ($n = 19$). IHC analysis of YY1 (Fig. 5f) showed that, with the exception of one metastatic sample (M3), YY1 staining robustly correlates with RI, confirming large clonal YY1-positive populations in all examined tissues (Fig. 5f). In parallel, SLC9A3R1 enhancer activity correctly estimated the size of the subclonal subpopulations in individual patients (Fig. 5g). Additional meta-analyses on Protein Atlas data support these findings by identifying YY1 clonal populations and SLC9A3R1 subclonal populations in most ERα BC samples (Supplementary Fig. 18). Overall, these data show that enhancer activity can be used to qualitatively deconvolute heterogeneous populations into phenotypical subclones.

Phenotypic evolution during BC progression is shaped by endocrine treatment. Tumor evolution studies have primarily focused on treatment-naive patients, taking advantage of multiregional sampling to monitor changes in clonality[50,51]. Clonal tracking is dependent in part on passenger mutations, and the effect of therapy has rarely been taken into account[8,52]. More importantly, clonality has been traced using genetic variants, with the intrinsic limitation of correlating genetic changes to phenotypic ones. For example, genetic subclones might be phenotypically equivalent, while a recent study using barcoded glioblastoma cells shows that phenotypic clones might evolve independently from genetic clones[26]. The few studies that looked at driver coding mutation changes in BC show relatively similar mutational landscapes[9] (Fig. 1a), suggesting a potential role for epigenetically driven phenotypic evolution. We thus leveraged our ability to infer phenotypic clones through enhancer activity to interrogate our patient's data set, focusing on events occurring between treatment-naive primaries and treatment-resistant metastatic BC (Fig. 6a). We hypothesized that phenotypic clonal evolution might be driven by a coordinated activation/selection of groups of enhancers during BC progression, and this could be influenced by treatment. Our previous results suggest that YY1-positive cells remain clonal during progression (Fig. 3a). Conversely, we show that SLC9A3R1 expression is not antagonized by endocrine treatment, suggesting that SLC9A3R1-positive clones could expand during progression. We then calculated changes in RI (ΔRI) for all enhancers captured in at least three patients (SI > 3, $n = 88{,}935$) between primary and metastatic samples (Fig. 6b). SLC9A3R1 ranks among the enhancers with the strongest increase in predicted clonality going from primary to metastatic samples (Fig. 6b,c). Conversely, YY1 enhancer activity remains relatively unchanged (Fig. 6b,c). To substantiate these data, we mapped the size of YY1 and SLC9A3R1-positive phenotypic clones in an independent cohort of 20 primary tumor and metastasis-matched longitudinal biopsies. We found that YY1-positive cells remain clonal in both settings, while SLC9A3R1-positive subclones significantly expand during metastatic progression (Fig. 6d). Interestingly, the only metastatic case in which we have observed a contraction of the SLC9A3R1-positive clone also showed a concomitant loss of ERα and PR positivity, demonstrating that SLC9A3R1 remains an ERα-dependent target despite being ET-insensitive in vivo (Fig. 6d). Overall, these data demonstrate that

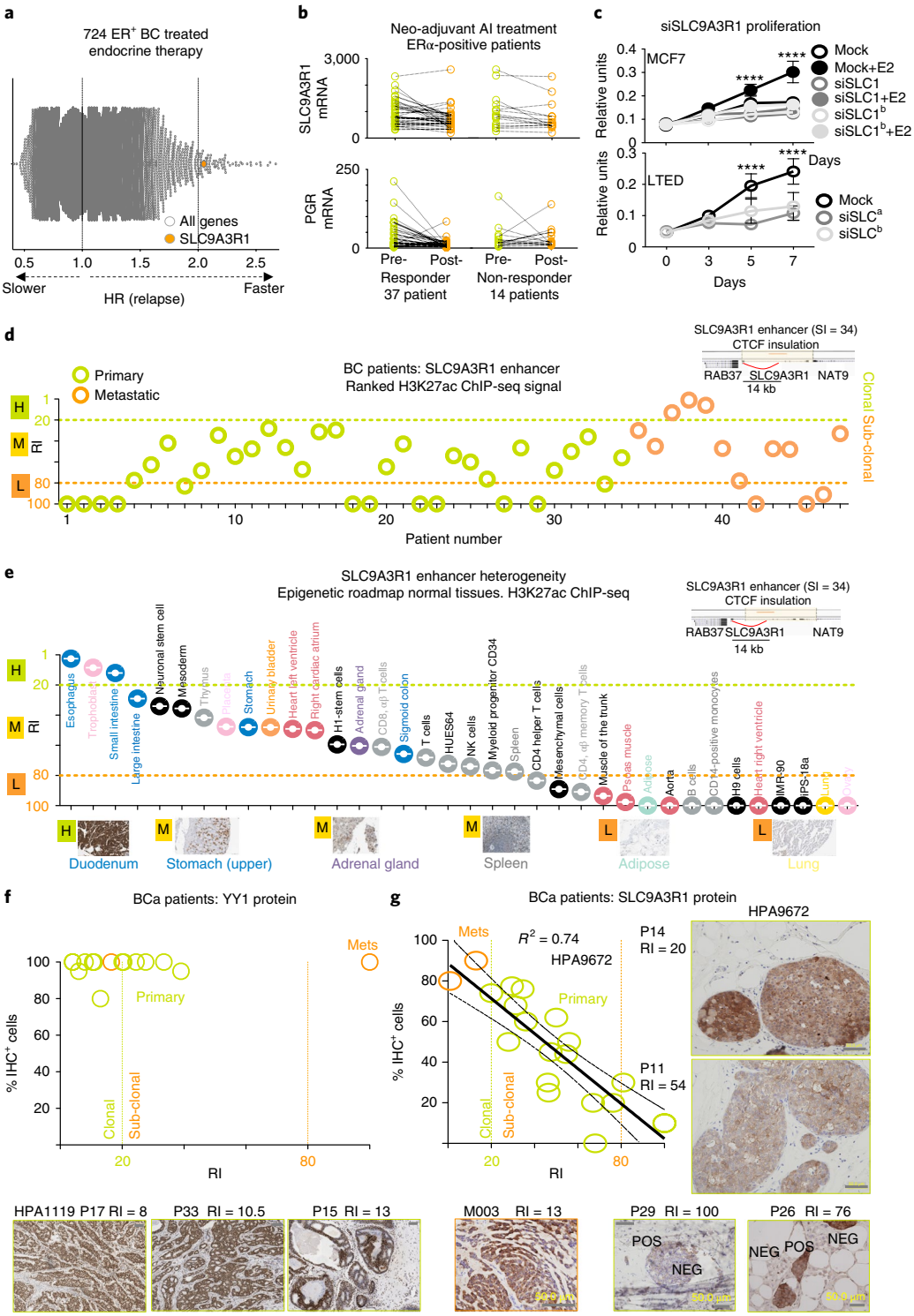

Fig. 5 | Epigenomic mapping predicts the size of phenotypic clones in patients. a, Global Kaplan–Meier analysis summarizes univariate analysis for 22,278 genes included in the Affymetrix microarray platform. Hazard ratios are plotted on the *x* axis. **b**, SLC9A3R1 RNA levels pre and post short-term aromatase inhibitor treatment in responder and non-responder patients[35]. Estrogen-dependent expression of progesterone receptor mRNA is shown for comparison. **c**, Silencing SLC9A3R1 leads to proliferation arrest in response to estrogen stimulation in MCF7 and estrogen-independent growth in LTED cells. Proliferation assays were conducted in biological triplicate. Symbols and error bars indicate average and 95% CIs. $^{****}P < 0.0001$; two-way ANOVA with Bonferroni's correction. **d**, RIs for the SLC9A3R1 enhancer within all the individual patients included in the current study. SLC9A3R1 enhancer location and its 3D interactions are shown in the top right inset. **e**, SLC9A3R1 enhancer ranking analysis of available Epigenome Roadmap H3K27ac data sets. Tissues are displayed from the strongest to weakest SLC9A3R1 enhancer activity (based on RI). Representative IHC analyses of normal tissues stained with a SLC9A3R1 antibody are shown. Scale bars, 50 µm. **f**,**g**, YY1 and SLC9A3R1 IHC analysis of BC patients profiled using H3K27ac ChIP-seq. Predicted activity (RI) of YY and SLC9A3R1 enhancers is shown on the *x* axis. The number of cells positively stained for YY1 and SLC9A3R1 protein is indicated on the *y* axis. Representative images are shown. One slide was stained for each patient. Linear regression R^2 values, CIs and representative staining are shown.

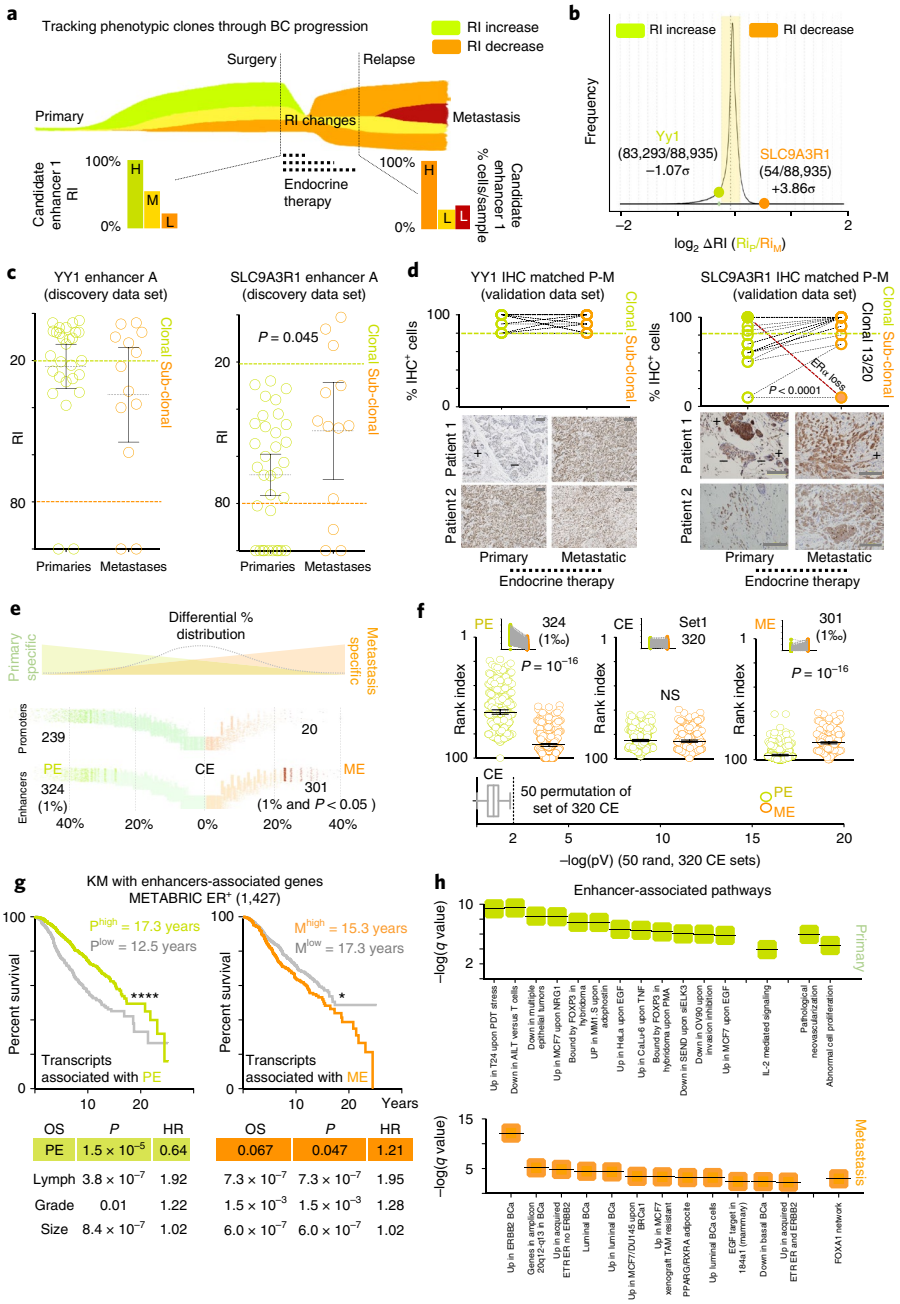

Fig. 6 | Endocrine treatment shapes phenotypic evolution. a, Theoretical framework of the analysis. The relative size of phenotypic clones can be tracked using enhancer ranking (RIs). Phenotypic clones can be positively or negatively selected during BC progression in response to ETs. **b**, Expanding or contracting phenotypic clones were defined based on the RI ratio of primary (P) and metastatic (M) samples (RI_P/RI_M). The distribution of RI ratios shows that YY1 enhancers' RI does not change significantly during progression compared to other enhancers, while SLC9A3R1 RI ranks among the enhancers with a stronger increase in activity during progression. Vertical bars represent 1σ (s.d.) increments from the population median. **c**, Scatterplot of YY1 and SLC9A3R1 enhancer ranking according to patient stage. Bars indicate mean and 95% CIs. Asterisks represent significance at $P < 0.05$ after Student's two-tailed *t*-test. **d**, IHC staining for YY1- and SLC9A3R1-positive cells in an independent matched longitudinal cohort of 22 ERα BC patients. Scale bars, 100 μm. All primary patients are treatment-naive. All metastatic patients have received ETs (tamoxifen or aromatase inhibitors). Statistical significance was calculated using a pair-wise, two-tailed *t*-test. Representative images are also shown. **e**, Enhancer and promoter stratification based on frequency of usage in primary and metastatic patients. Percentages were calculated for each regulatory region for each stage (primary and metastatic) and the differential was then derived and plotted on the *x* axis. All enhancers and promoters in Fig. 1 were used. PE and ME were called by taking the top 1/1,000 in the distribution that also satisfied a Fisher exact test $P < 0.05$. **f**, Dot-plot representing RI indices for all PE (324) and ME (301) are plotted. As a control, RI for common enhancers (CE = 320) are also plotted. Bottom, Permutation was used to assess changes in RI in 50 randomly selected sets of 320 CE. Box and whiskers represent median and 1–99th percentile for the *P* distribution. A Wilcoxon matched-pairs signed rank test was used to test for statistical significance. NS, not significant. **g**, Kaplan–Meier (KM) analysis using 1,427 ERα-positive patients and averaged RNA expression of genes associated with PE or ME regulatory regions. CI for PE: 0.39–0.61. CI for ME: 1.1–1.67. Comparison of survival curves was performed using a log-rank (Mantel–Cox) test. Genes were assigned considering CTCF insulated perimeters. Multivariate correction for the comparisons is also shown. P, primary; M, metastatic. **h**, Pathway analysis for genes associated with PE or ME regulatory regions. Pathways were identified using GREAT and are listed in order of significance (symbols indicate *q* value).

changes in enhancer ranking can estimate functional evolution during BC progression.

To gain insight into functional evolution, we systematically annotated all regulatory regions based on bias in detection between primary and metastatic patients (Fig. 6e). As expected, the bulk of enhancers and promoters do not show bias towards primary and metastatic BC patients (common enhancers, CEs). However, we could identify two distinct sets of regulatory region where activity is stronger in primary (primary enhancers, PEs) or metastatic (metastatic enhancers, MEs) patients (Fig. 6f). We next explored the potential causes and functional consequences driving these coordinated epigenetic changes by identifying the associated transcriptional targets of MEs and PEs[34]. Strikingly, we find that PE-driven gene transcription is associated with a significantly better outcome, while ME-associated gene transcription in primary samples is associated with poor prognosis (Fig. 6g). These data imply that primary samples containing larger subpopulations of phenotypic clones with metastatic features relapse earlier. PEs are associated with abnormal proliferation and vascularization, two key events in early tumorigenesis. Remarkably, MEs are associated with genes promoting BC progression (FOXA1[39]) or ET resistance (Fig. 6h). Altogether, these data suggest that ETs play a central role in shaping phenotypic clonal evolution. Additional in-depth studies are needed to dissect the temporal events triggered during phenotypic clonal evolution. Phenotypic subclones could evolve by early coordinated activation and decommissioning of epigenetically defined regulatory regions (acquired), selection of the fittest pre-existent epigenomic landscape (de novo), or a combination of both.

Discussion

Genomic profiling of BC patients has revealed extensive clonal heterogeneity and evolution[24,53], but it remains difficult to link genotype to actual phenotypes. Most RNA-based analyses, which may better reflect the phenotypic state of cancer cells, cannot inform on the existence of distinct subpopulations. Finally, molecular pathology can inform on the relative amount of protein abundance at the single-cell level, but is laborious and not suitable for testing multiple targets simultaneously. In this work, we have used epigenomic analyses to extrapolate phenotypic heterogeneity in solid tumor samples. Our analysis reveals that histone-based ChIP–seq signals, similarly to ATAC–seq[29], generally correlate with the number of cells in a population carrying the specific epigenetic information. Our predictions using YY1 and the SLC9A3R1 enhancer fit extremely well with experimental data derived from normal tissues or BC patients. The finding that clonal regulatory regions dominating the landscape of individual tumor samples are shared across many patients parallels recent genomic evidence showing that truncal (high allele frequency) mutations are also the most common mutations within cancer cohorts.

Our work reveals several critical principles underlying phenotypic–functional heterogeneity and its role in BC progression. First, by comparing samples from drug-resistant metastatic patients with drug-naive primary samples, we uncovered a set of enhancers marking phenotypic clones that significantly expand during BC progression. A set of enhancers expanding in metastatic samples point at progressive activation of FOXA1 and its network. It was recently reported that FOXA1 levels are increased in metastatic samples[39]. Our data predict that, similar to SLC9A3R1, FOXA1 positivity increases as a consequence of the expansion of a phenotypic clone marked by an active FOXA1 enhancer. It is tempting to speculate that this paradigm might be valid for other genes. If correct, it might signify that during cancer evolution, the proportion of cells activating transcription is more important than the absolute changes in transcription at single-cell levels. Interestingly, a set of enhancers deactivated during progression involve interleukin-2 (IL-2) signalling (Fig. 6h). Reduction in IL-2 signalling was identified as a potential marker of relapse[54]. Whether the IL-2 signal source is the BC cells themselves or is due to a small contamination of immune cells needs to be defined. Equally, it will be important to measure real-time activation/selection of enhancers in appropriate systems to ultimately establish if phenotypic cancer evolution can be driven by Lamarckian events.

Additionally, our analysis has identified two novel drivers of luminal BC. First, we identified YY1 as a key TF associated with clonal enhancers and promoters in BC patients. Our data strongly support the idea that YY1 acts as a global co-activator associated with the entire active epigenetic landscape in BC cells. Several lines of evidence indicate that YY1 might interact directly with modified nucleosomes, possibly through its partner INO80[55]. YY1 widespread association with a clonal enhancer suggests it might play a role in epigenetic memory. Intriguingly, a positive screen for factors that improve induced pluripotent cells formation (iPS) identified YY1 as the top hit, further supporting its potential role as an enhancer gatekeeper[56]. More specifically to ERα BC, we hypothesize that YY1 plays a critical role in stabilizing ERα binding at the transcriptionally productive core-ERα enhancers. Single-molecule imaging shows that estrogen-activated ERα increases its residency time on the chromatin[38], and recent evidence has shown that eRNA can trap YY1 on the chromatin[45]. Altogether, these data raise the intriguing hypothesis that YY1 might contribute to increased ERα residency at clonal enhancers (Supplementary Fig. 19). This could explain why some ERα occupancy is captured in most patients, as a longer residency time would increase the chances of being captured by ChIP–Seq[39]. Longer residency might also explain the increased transcriptional activity (Fig. 4d) and increased TF footprints (Fig. 2c) of these enhancers. Another possibility is that YY1 defines the set of ERα-bound enhancers with transcriptionally productive looping at target genes[41,42,57]. Further studies will investigate these hypotheses. Future studies are also required to investigate the exact mechanisms through which SLC9A3R1 contributes to BC and efficient strategies to antagonize its transcription. We recently demonstrated that individual ETs can drive parallel genetic evolution in vivo[8] and epigenetic reprogramming in vitro[10]. Our data now strongly support the notion that therapeutic interventions also play an essential role driving specific epigenetic evolution during BC progression in the clinic.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41591-018-0091-x.

Received: 25 September 2017; Accepted: 14 May 2018;
Published online: 23 July 2018

References

1. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. Ali, S., Buluwela, L. & Coombes, C. Antiestrogens and their therapeutic applications in breast cancer and other diseases. *Ann. Rev. Med.* **62**, 217–232 (2010).
3. Perou, C. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
4. Genestie, C. et al. Comparison of the prognostic value of Scarff–Bloom–Richardson and Nottingham histological grades in a series of 825 cases of breast cancer: major importance of the mitotic count as a component of both grading systems. *Anticancer Res.* **18**, 571–576 (1998).
5. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
6. Koboldt, D. et al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
7. EBCTCG. Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. *Lancet* **386**, 1341–1352 (2015).

8. Magnani et al. Acquired CYP19A1 amplification is an early specific mechanism of aromatase inhibitor resistance in ERα metastatic breast cancer. *Nat. Genet.* **49**, 444–450 (2017).
9. Yates, L. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184 (2017).
10. Nguyen, V. et al. Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. *Nat. Commun.* **6**, 10044 (2015).
11. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
12. ENCODE Project Consortium. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
13. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
14. Whyte, W. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
15. Heintzman, N. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
16. Falahi, F. et al. Towards sustained silencing of HER2/neu in cancer by epigenetic editing. *Mol. Cancer Res.* **11**, 1029–1039 (2013).
17. Laprell, F., Finkl, K. & Müller, J. Propagation of polycomb-repressed chromatin requires sequence-specific recruitment to DNA. *Science* **356**, 85–88 (2017).
18. Wang, X. & Moazed, D. DNA sequence-dependent epigenetic inheritance of gene silencing and histone H3K9 methylation. *Science* **356**, 88–91 (2017).
19. Coleman, R. T. & Struhl, G. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a *Drosophila* HOX gene. *Science* **356**, eaai8236 (2017).
20. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet.* **27**, 465–474 (2011).
21. Jozwik, K. M. & Carroll, J. S. Pioneer factors in hormone-dependent cancers. *Nat. Rev. Cancer* **12**, 381–385 (2012).
22. Hnisz, D. et al. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell* **58**, 362–370 (2015).
23. Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
24. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
25. Williams, M. J., Werner, B., Barnes, C., Graham, T. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
26. Lan, X. et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **549**, 227–232 (2017).
27. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
28. Harvey, J. M., Clark, G. M., Osborne, C. K. & Allred, D. C. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *J. Clin. Oncol.* **17**, 1474–1481 (1999).
29. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
30. Cowper-Sal Iari, 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).
31. Cohen, A. J. et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat. Commun.* **8**, 14400 (2017).
32. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
33. Levsky, J. M. & Singer, R. H. Gene expression and the myth of the average cell. *Trends Cell Biol.* **13**, 4–6 (2003).
34. Wang, S. et al. Target analysis by integration of transcriptome and ChIP–seq data with BETA. *Nat. Protoc.* **8**, 2502–2515 (2013).
35. Gyorffy, B. et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res.* **123**, 725–731 (2010).
36. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
37. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
38. Paakinaho, V. et al. Single-molecule analysis of steroid receptor and cofactor action in living cells. *Nat. Commun.* **8**, 15896 (2017).
39. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
40. Carroll, J. S. et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the Forkhead protein FoxA1. *Cell* **122**, 33–43 (2005).
41. Beagan, J. A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* **27**, 1139–1152 (2017).
42. Weintraub, A. et al. YY1 is a structural regulator of enhancer–promoter loops. *Cell* **171**, 1573–1588 (2017).
43. Vella, P., Barozzi, I., Cuomo, A., Bonaldi, T. & Pasini, D. Yin Yang 1 extends the Myc-related transcription factors network in embryonic stem cells. *Nucleic Acids Res.* **40**, 3403–3418 (2012).
44. Jeon, Y. & Lee, J. T. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146**, 119–133 (2011).
45. Sigova, A. A. et al. Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978–981 (2015).
46. Klymenko, T. et al. A polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. *Genes Dev.* **20**, 1110–1122 (2006).
47. Tang, Z. et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
48. Hurtado, A., Holmes, K., Ross-Innes, C., Schmidt, D. & Carroll, J. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet.* **43**, 27–33 (2011).
49. Cardone, R. A., Casavola, V. & Reshkin, S. J. The role of disturbed pH dynamics and the Na^+/H^+ exchanger in metastasis. *Nat. Rev. Cancer* **5**, 786–795 (2005).
50. Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
51. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27**, 15–26 (2015).
52. Juric, D. et al. Convergent loss of PTEN leads to clinical resistance to a PI(3) Kα inhibitor. *Nature* **518**, 240–244 (2015).
53. Shah, S. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
54. Arduino, S. et al. Reduced IL-2 level concentration in patients with breast cancer as a possible risk factor for relapse. *Eur. J. Gynaecol.* **17**, 535–537 (1996).
55. Cai, Y. et al. YY1 functions with INO80 to activate transcription. *Nat. Struct. Mol. Biol.* **14**, 872–874 (2007).
56. Onder, T. et al. Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **486**, 598–602 (2012).
57. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).

Acknowledgements

The authors acknowledge and thanks all patients and their families for support and for donating research samples. The authors thank the Breast Cancer Now Tissue Bank (project TR0121), Imperial Tissue Bank and the LEGACY study for contributing tissues. The authors acknowledge infrastructure support from the Cancer Research UK Imperial Centre, the Imperial Experimental Cancer Medicine Centre and the National Institute for Health Research Imperial Biomedical Research Centre. L.M. was supported by a CRUK fellowship (C46704/A23110) and an Imperial Junior Fellowship (G53019). D.P. was supported by a Wellcome Trust PhD studentship (103034/Z/13/Z). G.C. was supported by a Marie Skłodowska Curie Training Grant (642691, EpiPredict). G.P. was supported by AIRC IG 2016-18696. The authors thank J.A. Buendia, L. Watson and J. Carrol for constructive comments on the manuscript.

Author contributions

L.M. conceived the study. D.K.P., E.E., N.S. and Y.P. performed the experiments. L.M., G.C., B.G., A.S., L.S.P., I.B. and P.S. developed and performed bioinformatic analyses. K.G. organized tissue collection. D.J.H., G.S., P.B., C.P. and R.C.C. recruited patients and supplied tissues. S.S. performed pathology assessment of ChIP–seq processed samples. G.P. provided matched material. A.V. and G.P. performed IHC staining and scoring. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0091-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.M.

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

Methods

Tumor tissue processing. BC samples for ChIP–seq were collected by the Imperial Tissue Bank (project ethic approval R15021) and from Breast Cancer Now Tissue Bank (BCNTB- TR000053-MTA and TR000040). BC fresh frozen tissue samples each underwent aseptic macroscopic adipose tissue dissection. The dissected tumor tissue was sectioned into 2 mm × 2 mm fragments in a Petri dish placed over dry ice. Tumor fragments were then fixed using 1% formaldehyde solution for 10 min. Cold glycine (1 M) was added to the formaldehyde-fixed tissue for 10 min. The fragments were then pulverized using pestle and mortar and homogenized using liquid nitrogen. We used samples with high tumor burden to minimize the introduction of noise from non-tumor tissues (>70%, Supplementary Fig. 1a). Wherever possible, we profiled patients for known cancer drivers using targeted enrichment sequencing (Fig. 1a and Supplementary Data 1). A total of 85% of samples yielded satisfactory results (47/55, Supplementary Fig. 1b and Supplementary Table 2).

Cancer hotspot mutations. See Supplementary Computational Methods.

ChIP. The ChIP protocol was conducted as described in ref. [58] with few modifications. In summary, following fixation, the tumor tissue underwent chromatin extraction and sonication using a Bioruptor Pico sonication device (Diagenode, B01060001) using 20 cycles (30 s on and 30 s off) at maximum intensity. Purified chromatin was then separated for the following: (1) immunoprecipitation using 4 µg of H3k27ac antibodies (Abcam; ab4729) per ChIP experiment or using 4 µg of YY1 antibodies (Santa Cruz; sc-281X) (ChIP–seq experiments for YY1 were performed in biological duplicates, cells were stimulated with estrogen for 45 min, after which maximum ERα-binding to chromatin occurs; biological replicates showed very high correlation ($R^2 = 0.98$), so only consensus loci were kept for further analyses); (2) non-immunoprecipitated chromatin, used as input control; (3) assessment of sonication efficiencies using a 1% agarose gel. Before construction of ChIP–seq libraries (NEB Ultra II kit, Supplementary Methods), enrichment of the immunoprecipitated sample was ascertained using positive and negative controls for ChIP–qPCR. Library preparation was performed using 10–50 ng of immunoprecipitated and Input samples. Before sequencing, libraries were again retested to confirm enrichment using positive and negative controls.

ChIP–qPCR. Briefly, reactions were carried out in 10 µl volume containing 5 µl of SYBR-green mix (ABI; 4472918), 0.5 µl of primer (5 µM final concentration), 2.5 µl of genomic DNA and 2 µl of DNASE/RNASE-free water. A three-step cycle programme and a melting analysis were applied. The cycling steps were as follows: 10 s at 95 °C, 30 s at 60 °C and 30 s at 72 °C, repeated 40 times.

Ranking and sharing index. See Supplementary Computational Methods.

Variant set enrichment. See Supplementary Computational Methods.

DHS imputations and TF motif analyses. See Supplementary Computational Methods.

Imputed DHS with in vivo ERα binding overlap. Dataset of ERα binding derived from BC patients were obtained from ref. [39]. The ERα SI was calculated using the same algorithm used for the H3K27ac dataset (see Supplementary Computational Methods). Overlap with imputed DHS was calculated using BedTools by calculating the overlap (at least one base pair) with the Cistrome Pipeline Analysis Suite (http://cistrome.org/Cistrome/Cistrome_Project.html). Percentages of overlap were calculated using binned DHS as a variable first dataset and all the concatenated in vivo ERα as the second dataset.

Footprint analysis. See Supplementary Computational Methods.

Encode and epigenomic roadmap ranking. See Supplementary Computational Methods.

Immunocytochemistry. Haematoxylin and eosin staining of clinical samples was performed to calculate the tumor burden before ChIP–seq. Briefly, 4-µm-thick sections were obtained from formalin-fixed and paraffin-embedded specimens. After dewaxing in xylene and graded ethanol, sections were incubated in 3% H_2O_2 solution for 25 min to block endogenous peroxidase activities and then subjected to microwaving in EDTA buffer for antigen retrieval. For YY1 (Protein Atlas HPA001119, Atlas Antibodies cat. no. HPA001119, RRID:AB_1858930) flowing conditions were used: tissue sections were incubated with primary monoclonal antibody overnight at 4 °C, and chromogen development was performed using the Envision system (DAKO Corporation). A minimum of 500 tumor cells were scored, with the percentage of tumor cell nuclei in each category recorded. For SLC9A3R1 (HPA9672 and HPA27247, Atlas Antibodies cat. no. HPA009672, RRID:AB_1857215 and Atlas Antibodies cat. no. HPA027247, RRID:AB_10601162, respectively) the following conditions were used. HPA9672 was diluted 1:400 and HPA27247 was diluted 1:1,500. Staining was automatized with a Ventana Benchmark-Ultra using epitope retrieval ER2 for 20 min. ER and PgR immunoreactivity were assessed with the FDA-approved ER/PR PharmDX kit (Dako). The prevalence of ER/PgR-positive invasive cancer cells, independent of staining intensity, was quantitatively annotated in the original diagnostic reports. In accordance with ASCO/CAP guidelines, tumors with ≥1% of immunoreactivity were considered positive.

Cell culture. MCF7 was cultured using Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal calf serum (FCS) and 100 U penicillin/0.1 mg ml^{-1} streptomycin, 2 mM L-glutamine plus 10^{-8} 17-β-estradiol (SIGMA E8875). MCF7 long-term estrogen-deprived (MCF7-LTED) cells were grown in phenol-free DMEM with 10% charcoal-stripped FCS (DCFCS) and 100 U penicillin/0.1 mg ml^{-1} streptomycin and 2 mM L-glutamine. T47D and T47D-LTED cells were passaged using DMEM containing 10% FCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin, 2 mM L-glutamine and phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin and 2 mM L-glutamine, respectively. ZR75-1 cells were grown in DMEM containing 10% FCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin, 2 mM L-glutamine.

siRNA. siRNA against SLC9A3R1 (gene ID 9368: Ambion s17919, s17920), YY1 (gene ID 7528: Ambion s14958, s14959, s14960) and *Silencer* negative control (Ambion AM4611). 1.5×10^5 cells were seeded per well using a six-well plate. MCF7 cells were seeded in phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin and 2 mM L-glutamine. After 24 h, the cells were transfected with siRNA using Lipofectamine 3000 (Invitrogen L3000015). T47D and ZR75-1 cells were seeded in DMEM containing 10% FCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin, 2 mM L-glutamine. After 24 h, the cells were transfected with siRNA using Lipofectamine 3000 (Invitrogen L3000015). Cells were collected for analysis following at least 48 h of transfection.

CRISPR/–Cas9 enhancer knockout. See Supplementary Methods.

Live cell imaging. MCF7 and YY1-EKO clones cells were plated at a density of 3×10^3 in a 96-well plate in FluoroBrite DMEM medium (ThermoFisher) supplemented with 1×10^{-8} M estradiol. Cells were culture in an Incucyte Zoom (EssenBioscience) programmed to capture images every 6 h. Twenty single cells for each cell line were followed over the course of 84 h and their doubling time recorded and plotted.

Cell lysis and western blot. Cells were washed twice in ice-cold PBS and lysed in RIPA (Sigma-Aldrich R02780) buffer supplemented with protease (Roche 11697498001) and phosphastase (Sigma-Aldrich 93482) inhibitors for 30 min with intermittent vortexing. Samples were centrifuged at 4 °C at maximum speed for 30 min, then the supernatant was transferred to a clean Eppendorf tube. Protein concentrations for each sample were ascertained using a bicinchoninic acid (BCA) assay (ThermoFisher Scientific 23227). Equal amounts of lysates were loaded into BOLT 4–12% Bis-Tris Plus Gel (Invitrogen NW04120BOX). Proteins were transferred to a Biotrace nitrocellulose membrane (VWR; PN66485) and incubated with primary antibodies overnight. Proteins were then visualized using goat anti-mouse (ThermoFisher Scientific 31446) and anti-rabbit (ThermoFisher Scientific 31462) horseradish peroxidase (HRP) conjugated secondary antibodies. Amersham ECL Prime Western Blotting Detection reagent (GE Healthcare Life Sciences RPN3243) was used for chemiluminescent imaging using a Fusion solo (Vilber) imager. For SLC9A3R1 we used HPA027247 (Protein Atlas) at 1:1,000 dilution, and for YY we used Santa Cruz sc-281 at a 1:500 dilution. For GAPDH we used Abcam ab9385 at a 1:5,000 dilution.

Transcriptional profiling. Following 48 h of transfection, MCF7 cells were either treated with 10^{-8} 17-β-estradiol (SIGMA E8875) or control treatment for 6 h before RNA extraction. T47D and ZR75-1 cells lines were harvested for RNA following 48 h of transfection. No treatments were added.

RNA extraction and real-time PCR. Total RNA was extracted using an RNeasy Mini Kit (Qiagen 74106), and the cDNA was reverse transcribed from 1 µg of RNA using iScript cDNA synthesis kit (Bio-Rad 1708891). Real-time qPCR (RT–qPCR) reactions were carried out in a 10 µl volume containing 5 µl sybergreen mix (ABI 4472918), 0.5 µl primer (2.5 µM final concentration), 2.5 µl genomic DNA and 2 µl DNASE/RNASE-free water. A three-step cycle programme and a melting analysis were applied. The cycling steps were 10 s at 95 °C, 30 s at 60 °C and 30 s at 72 °C, repeated 40 times.

Luciferase reporter assay. MCF7 cells were seeded in a 24-well plate at 5×10^4 cells per well in phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin and 2 mM L-glutamine. After 24 h of incubation, transfection of plasmid DNA was performed using Lipofectamine 3000 (Invitrogen L3000015). Cells were transfected with 100 ng of ERE_Luciferase reporter, 10 ng of the renilla luciferase control plasmid (pRL-CMV), 10 ng of pSG5_ER-α, 15 nm of siRNA and 280 ng of Bluescribe DNA (BSM) per well, totalling 400 ng of DNA per well. After 12 h of transfection the medium was replaced with fresh phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml^{-1} streptomycin and 2 mM L-glutamine.

Treatment with 10^{-8} 17-β-estradiol (Sigma E8875) or control treatment was administered and the cells incubated for 24 h. Cell lysates were then obtained using Passive lysis 5× buffer (Promega E1941). Firefly and renilla luciferase activity were determined using a DualGlo luciferase assay kit (Promega E2920) according to the manufacturer's protocol. The renilla luciferase activity measurement was utilized as a control for transfection efficiency, so the ERE_Luciferase activity was normalized to the reading obtained for renilla luciferase activity.

Sulforhodamine B assay. Briefly, the sulforhodamine B (SRB) assay was used to monitor the effects of silencing either SLC9A3R1 or YY1, using siRNAs, on cell proliferation monolayer cultures. Cells were seeded in flat-bottomed 96-well plates (Costar CLS3585) at a density of 2×10^3. Cells were allowed to attach overnight, then the first plate (Day 0) was assayed once the cells had become adherent. Prospective plates were assayed sequentially after 3 days, 5 days and 7 days. The cells were fixed by adding 200 µl of cold 40% (wt/vol) trichloroacetic acid (TCA) to each well for at least 60 min. The plates were washed five times with distilled water, 100 µl of SRB reagent (0.4% wt/vol SRB in 1% wt/vol acetic acid) was added to each well, and the plates were allowed to incubate for 30 min. The plates were then washed five times in 1% (wt/vol) acetic acid and allowed to dry overnight. SRB solubilization was performed by adding 100 µl of 10 mM Tris HCl per well to the plates, followed by shaking for 30 min. Optical density was then measured using a Sunrise microplate reader (Tecan) at 492 nm. Cell proliferation was calculated over the 7 day period (with day 0 as baseline).

Enrichment scores. See Supplementary Computational Methods.

RI–IHC correlation. Formalin-fixed, paraffin-embedded sections for the patients used in the ChIP–seq section were retrieved from Imperial Tissue bank. Sections were stained with YY1 or SLC9A3R1 antibodies. Stained sections were divided into 20 sectors. Five sectors with high tumor burden were scored for the number of IHC-positive cells and the results averaged. The number of IHC-positive cells and the matched RI were analysed using linear regression using Prism 5 (GraphPad software).

ΔRI. See Supplementary Computational Methods.

YY1 and SLC9A3R1 Pan cancer expression analysis. YY1 and SLC9A3R1 expression profiles for matched normal versus cancer samples were obtained using the TIMER diff.exp option (https://cistrome.shinyapps.io/timer/). YY1 transcriptional analyses of BC subtypes were performed in the Metabric Dataset (Curtis Breast) using probe ILMN_1770892 or the TCGA dataset using Oncomine (https://www.oncomine.org/resource/login.html).

SLC9A3R1 meta-analyses. The SLC9A3R1 expression profile in drug-resistant cell lines was performed by analysis of RNA–seq data from ref. [10]. The SLC9A3R1 expression profile in MCF7 cells transfected with siRNA against ERα was performed by analysis of microarray data from GSE27473. The SLC9A3R1 expression profile in additional LTED models was performed by analysis of microarray data from E-GEOD-19639. All statistical analyses were performed using Prism 5 (GraphPad software). Kaplan–Meier analysis using SLC9A3R1 expression was performed by reanalysis of 23 independent microarray data sets (KMPLOT), TCGA RNA–seq data or the combined Metabric Dataset. Multivariate Cox proportional hazards survival analysis was performed using gene expression and clinical variables including nodal status, grade and size in the Metabric and Affymetrix data sets, using Winstat for Excel 2017. A multivariate analysis in the TCGA data set employing available clinical data including tumor node metastasis, histology, menopausal status and race did not deliver significant results for any of the included parameters, probably due to the short follow-up combined with limited number of events. The SLC9A3R1 transcriptional profile in BC cell lines was obtained from the HPA RNA–seq data set (http://www.proteinatlas.org/about/download). The SLC9A3R1 transcriptional profile from tissues was obtained from the HPA, GTEx and FANTOM5 RNA-seq data sets (http://www.proteinatlas.org/about/download).

Data availability. H3K27ac data for all patients' samples have been deposited at the ENA (http://www.ebi.ac.uk/ena) under project no. PRJEB22757.

References

58. Schmidt, D. et al. ChIP–seq: using high-throughput sequencing to discover protein–DNA interactions. *Methods* **48**, 240–248 (2009).

nature research

Corresponding author(s): Luca Magnani

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

Targeted capture was performed using NEB Cancer Hotspot panel modified to include ESR1 ligand binding domain (NEB E7000X). Sonicated Input material from ChIP-seq analysis (frozen tissues) was used as an input (minimum 50ng) as specified by the manufacturer. Sequencing was performed on a NextSeq Illumina machine by multiplexing 24 samples per lane in two lanes (Single End 75bp flow cell). Single-end 75-base pairs reads were aligned to the hg38 human reference genome using bwa1 version 0.7.15 (parameters: -q 0). Samtools (PMID: 19505943) version 1.3.1 was then used to obtain indexed bam files. Aligned reads from each captured sample were pre-processed using Picard (http://broadinstitute.github.io/picard) version 2.6.0, applying functions AddOrReplaceReadGroups (parameters: RGID=1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=1) and sortSam (parameters: SORT_ORDER=coordinate). GATK 2 version 3.6 was then used for variant identification. PCR duplicates were marked using the MarkDuplicates function from Picard (parameters REMOVE_DUPLICATES=False AS=True). Re-alignment around indels was performed using functions RealignerTargetCreator and IndelRealigner from GATK (known indels from the GATK bundle: Mills_and_1000G_gold_standard.indels.hg38.vcf). This step was followed by base quality score recalibration (GTAK BaseRecalibrator). Mutect2 (part of GATK v3.6) was finally run separately on each capture, without control samples. The identified variants were then annotated to known SNPs (1000G_phase1.snps.high_confidence.hg38.vcf in the GATK bundle) and to COSMIC 3 version 34 (hg38). Variants showing alternate allele frequency lower than 1% were excluded from further analyses. Those supported by evidence from both alleles and covered by ten or more reads were retained. Variants overlapping known SNPs were excluded. Among the remaining variants, only those previously reported in COSMIC were kept. As a final step, those protein-coding variants predicted as “Neutral” by FATHMM 4were filtered out.

Reads were quality controlled with FastQC v0.11.5 and aligned to the human hg38 reference using bowtie v1.1.2 5 with default parameters. The generated sequence alignments were converted into binary files (BAM), then sorted and indexed using the SAMtools

v1.3. H3K27ac peaks were called with MACS2 v2.1.16 (command-line parameters: -callpeak --format AUTO -B --SPMR --call-summits -q 0.01) using matched input DNA as a control. Samples showing either less than 2K or more than 200K H3K27ac peaks were not considered for further analysis.

We re-analysed ChIP-seq data of H3K27ac profile across 33 cell lines from ENCODE 10 and 37 tissues from the Epigenomic Roadmap11, for a total of 337 epigenomic profiles. We downloaded matching .bam and .bed profiles from ENCODE and matching raw reads of input and ChIP from Epigenomic Roadmap. The epigenomic profiles of ENCODE cell lines from human hg19 reference genome were lifted to the human hg38 assembly using CrossMap v0.2.312. Peaks from the Epigenomic Roadmap samples were called following the procedure above. The BC active promoter and enhancer sets were intersected with all the epigenomic profiles and the RI calculation of each peak was repeated as above.

We downloaded 1000 Genomes Project genotypes data (Phase 3 release 20130502) and excluded any genotype calls in individuals of non-European ancestry. We then ran PLINK (v1.90b3.46)14 on the filtered genotypes data and a list of 66 CEU BC risk variants to retrieve 1000 Genomes variants in LD with each BC variant. We defined LD variants as those within 500KB of a BC variant and having an allele count squared correlation >=0.8 with that variant. We also ran PLINK with the same settings on a list of 20 CEU CRC risk variants to obtain their LD information. The PLINK output files were then converted into BED format to be used in downstream analyses by VSE R library (v0.99).
We ran VSE separately for BC and CRC variant sets to assess the enrichment of those variants in the following list of genomic features on hg19: 5' and 3' UTR, Refseq gene TSS, Refseq gene introns, Refseq gene exons, active BC promoters, active BC enhancers with SI =1, active BC enhancers with SI between 1 and 21 exclusive, and active BC enhancers with SI >=21. Active BC promoters and enhancers were converted from hg38 to hg19 using liftOver prior to running VSE. During each VSE analysis, an associated variant set (AVS) was constructed using LD block information from PLINK-generated variant lists. 1000 matched random variant sets (MRVS) from 1000 Genome Project Phase III data were then generated. The final step was to compute the enrichment of AVS in the set of previously described genomic features compared to the null distribution (MRVS). Enrichment results are shown in Figure 1F with Bonferroni adjusted p-value < 0.05 marked in red. We also generated a heatmap (Figure 1E) showing the overlaps between BC risk variants as well as variants in LD and the genomic features of interest.

Data analysis

Functional characterization of the peaks. The identification of promoter and enhancer peaks was performed using an in-house pipeline based on BEDTOOLS v2.25.0 6and custom BASH scripts. A promoter annotation which classifies the promoter as the region 1kb upstream of the transcription-start site (TSS) was generated using UCSC table browser (PMID 27899642) (assembly: hg38; groups: Genes and Gene predictions; track: GENCODE v24) 7.
Peaks were then intersected using BEDTOOLS intersect (default parameters) to identify the promoter specific peaks. Annotated promoters which were not overlapping with the patient signal were considered inactive. In order to produce a master list of active core promoters, a multiple intersection between the promoter peaks was performed using BEDTOOLS multiinter to identify the common overlapping signal. The book-ended regions from the core signal file were merged using BEDTOOLS merge, then intersected with the original peak calls and sorted. All those peaks showing no overlap with the promoter annotation were considered enhancers. The procedure used to derive active core promoters (outlined in the previous paragraph) was applied to these signals to generate a master list of active enhancers.

Assessment of the level of heterogeneity. Active promoters and enhancers were further processed in order to reveal whether the available dataset achieves a high genomic coverage. The saturation analysis was performed with ACT SaturationPlotCreator8 with default parameters. The frequency distribution and the average peak size distribution of each regulatory region was calculated intersecting the peaks from each individual with the master lists of active promoters and enhancers and then plotted using BASH and R in-house scripts. The size of each peak was extracted from the MACS2 output files (_peaks.xls) and the peaks binned by sharing index.

Sharing Index. Sharing Index (SI) is a discrete metric introduced for measuring the usage of enhancer and promoter across the tumor samples. SI was calculated as the number of individual samples in which a regulatory region overlaps the master list with a coverage of at least 40% of its bases. This way, a discrete SI score was assigned to all promoters and enhancers in the master list. To add further significance to the accuracy of this metric, we compared it to a quantile normalized continuous equivalent of SI, calculated as follows. The number of deduplicated reads overlapping each regulatory region in the master list was calculated using BEDTOOLS Multicov with default parameters. A matrix showing the read count of each tumor sample across all the regions was derived and quantile normalized after Voom transformation (LIMMA 9 package available in Bioconductor). In addition, data were scaled (z-score) and compressed with (arcsinh) transformation.

Ranking Index. The level of enrichment of each regulatory region in the tumor sample dataset is scored using the Ranking Index (RI) metric. RIs were assigned to each called peak. Duplicated reads from the ChIP-Seq treatment files were filtered out using PICARD v2.1.1 MarkDuplicates (REMOVE_DUPLICATES=true) and only the uniquely mapped reads were retained for further analyses. Peak read count was obtained using BEDTOOLS Multicov function and this value was normalized using the following equation: Nscore= ((peak read count / peak size) · 106))* 103 /total mapped reads (FPKM).
Peak calls in each sample were categorized as promoter or enhancer as described in the previous paragraph, then sorted by their FPKM and assigned to their respective intra-sample percentile score where 1 is highest enrichment and 100 is the lowest. The peak calls were then intersected with the sets of active promoters and enhancers set and the average RI for each promoter and enhancer was calculated.

Ranking approach in cancer cell line and normal tissue epigenomes. We re-analysed ChIP-seq data of H3K27ac profile across 33 cell lines from ENCODE 10 and 37 tissues from the Epigenomic Roadmap11, for a total of 337 epigenomic profiles. We downloaded matching .bam and .bed profiles from ENCODE and matching raw reads of input and ChIP from Epigenomic Roadmap. The epigenomic profiles of ENCODE cell lines from human hg19 reference genome were lifted to the human hg38 assembly using CrossMap v0.2.312. Peaks from the Epigenomic Roadmap samples were called following the procedure above. The BC active promoter and enhancer sets were intersected with all the epigenomic profiles and the RI calculation of each peak was repeated as above.

Transcription factor profiling. The profile of the BC cistrome was imputed by taking all the potential accessible regions encoded in the active promoter and enhancer set. H3K27ac ChIP-Seq provides the location of the enriched histones while the transcription factors bind the accessible regions in the nucleosome-free region (NFR). NFRs were putatively characterized by the analysis of DNaseI-hypersensitivity

site (DHS) from 220 different ENCODE cell lines available at: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwDnase/ and http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromDnase/; DHS profiles were generated using MACS2 with the following parameters: --format AUTO --nomodel --shift -100 --extsize 200 -B --SPMR --call-summits -q 0.01 and lifted to the human hg38 assembly. After that, all the DHS peaks were concatenated into one sorted BED file. NFRs were identified as the regions between two sub-peaks at a distance of +- 71bps from the subpeak summit and the region between two broad-peaks distant at the most 500bps. DHS signals overlapping the NFRs were retained for the analysis. The retained DHS sites were sorted and elongated using BEDTOOLS merge to have a unique DHS signal for all the NFRs. Motif enrichment analysis was carried out separately on promoter and enhancer specific DHS signals in the BC datasets using the HOMER function findMotifsGenome.pl with parameters: -size given -preparse. The highest 50 ranked TFs in the two groups were selected and graphed in polar histograms with a custom R script.
We then binned promoters and enhancers by SI, overlapped the NFRs identified above and ran the motif enrichment analysis separately on each promoter and enhancer bin (in the same way described above). The motif enrichment results were filtered for statistical significance (q-value <= 0.05) and integrated with the observed/expected ratio (OEr) of each TF with a custom R script. Two heatmaps (one for promoters and one for enhancers) showing the OEr across the bins were generated using heatmap.2 from the ggplot2 R library13 In order to highlight the most significant results from the enhancer heatmap, we computed a differential analysis between the 2 clades of the heatmap (SI 1-21 and SI 22-44). We calculated the mean of OEr for each TF between the 2 clades and counted the number of significant enrichments in each clade. Then, we computed a weighted score specific to each TF multiplying the relative clade mean x number of significant clade enrichments. Furthermore, we calculated the log of the ratio, ranked and plot it. DHS regions imputed using the procedure outlined in this paragraph were compared to ENCODE Honey Badger DHS (https://personal.broadinstitute.org/meuleman/reg2map/) and found to be highly comparable.

Variant Set Enrichment VSE. We downloaded 1000 Genomes Project genotypes data (Phase 3 release 20130502) and excluded any genotype calls in individuals of non-European ancestry. We then ran PLINK (v1.90b3.46)14 on the filtered genotypes data and a list of 66 CEU BC risk variants to retrieve 1000 Genomes variants in LD with each BC variant. We defined LD variants as those within 500KB of a BC variant and having an allele count squared correlation >=0.8 with that variant. We also ran PLINK with the same settings on a list of 20 CEU CRC risk variants to obtain their LD information. The PLINK output files were then converted into BED format to be used in downstream analyses by VSE R library (v0.99).
We ran VSE separately for BC and CRC variant sets to assess the enrichment of those variants in the following list of genomic features on hg19: 5' and 3' UTR, Refseq gene TSS, Refseq gene introns, Refseq gene exons, active BC promoters, active BC enhancers with SI =1, active BC enhancers with SI between 1 and 21 exclusive, and active BC enhancers with SI >=21. Active BC promoters and enhancers were converted from hg38 to hg19 using liftOver prior to running VSE. During each VSE analysis, an associated variant set (AVS) was constructed using LD block information from PLINK-generated variant lists. 1000 matched random variant sets (MRVS) from 1000 Genome Project Phase III data were then generated. The final step was to compute the enrichment of AVS in the set of previously described genomic features compared to the null distribution (MRVS). Enrichment results are shown in Figure 1F with Bonferroni adjusted p-value < 0.05 marked in red. We also generated a heatmap (Figure 1E) showing the overlaps between BC risk variants as well as variants in LD and the genomic features of interest.

Footprint analysis. Footprints within the chromatin accessible regions in MCF7 were obtained using Wellington14,15 with parameters -fdr 0.01 -pv -5,-10,-20,-30,-50,-100. We identified the active regions in MCF7 and intersected them with the patients signals, which are broader then the single narrow peaks defined by MACS, and allow the identification of all the NFRs. The number footprints within each active regulatory region was calculated, and then normalized by the region size. The RI for eacg promoter and enhancer in MCF7 calls was calculated and plot in function of the number of footprints.

Estimation of somatic Copy Number Alterations (sCNA). Input BAM files from ChIP-seq experiment of tumor samples and cell lines were processed to estimate the chromosomal losses and gains in each tumor sample dataset. After removal of duplicated reads, the input BAM files were processed to detect sCNA using QDNAseq16 and CNVkit tools.17 QDNAseq data processing involve genome binning, correction for GC-content and mappability, and normalization. The hg38 genome was binned in 15kb and 100kb sized windows and copy numbers were inferred applying the standard procedure (https://cnvkit.readthedocs.io/en/stable/pipeline.html) (with default parameters. CNVkit was run with the default parameters of the batch command after creating a flat reference genome as suggested in the manual using the command reference.

Assessment of dinucleotide composition. The impact of possible sequence artifacts driving the SI scores has been assessed by a complete evaluation of the dinucleotide frequencies in each SI bin. We obtained the expected dinucleotide frequencies by processing the input BAM files of tumor samples in the dataset. Deduplicated Input BAM files from all patients were merged, sorted and indexed using SAMtools. The merged bam was then converted to FASTA. The frequencies of the 16 dinucleotides were computed using the compseq module of EMBOSS 18with parameter "-word 2". The frequencies of dinucleotides in the bins were obtained by coupling BEDTOOLS get fasta to convert the coordinates of regulatory regions in fasta format and EMBOSS compseq -word 2 to calculate the actual frequencies by bin.

Enrichment scores. Overlap for ER (in vivo) vs enhancers and promoters were calculated by BEDTOOLS intersect. The percentage overlap was calculated on the total number of regulatory regions within each bin against the concatenate ER binding set (all ER in all patients). For YY1, FOXA1 and ER in MCF7, intersections were calculated using Cistrome19. YY1 BED files were defined as the consensus narrow peaks of two biological replicates. FOXA1 ChIP-seq data and ER were obtained in house20. The core ER BED file was obtained by lifting a published dataset 21to hg19 coordinates. The private ER BED file was obtained by iterative processing of the ER binding sites unique to single patients prior to concatenation into a single file. Overlap represent the fraction of the original datasets (first dataset) overlapping with core ER (second dataset). The TCGA luminal signature was obtained from22. Each gene was extended for 20Kb upstream keeping in consideration the direction of transcription. A null gene list was generated by subtracting the TCGA luminal signature from a genome-wide gene list. Genes from the null list were extended in a similar way and enrichment was calculated by comparing the fraction of TCGA gene list with nearby binding vs. the null list. A list of estrogen target genes that do not respond to Tamoxifen was obtained from 23. Each gene was extended for 20Kb upstream keeping in consideration the direction of transcription. A null gene list was generated by subtracting the signature from a genome-wide gene list. Genes from the null list were extended in a similar way and enrichment was calculated by comparing the fraction of TAM resistant estrogen dependent gene list with nearby binding vs. the null list.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

H3K27ac data for all patients' samples have been deposited at the ENA (http://www.ebi.ac.uk/ena) under project number PRJEB22757.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size for the ChIP-seq cohort was not predetermined as this was a discovery-based project.
Data exclusions	We have excluded from the analysis samples that yielded less than 2000 calls or over 3000000 calls (as described in the manuscript and reported in the supplementary tables).
Replication	Each samples was exhausted after the analysis making replication of the in vivo part of the study impossible. Cell lines data were replicated (ChIP-seq n=2, other experiments n>5). Each replication was successful
Randomization	Randomization was not performed in the current study as this was a discovery based project and the goal was to compile a preliminary compendium of regulatory regions potentially involved in breast cancer. We did not design the study to compare between groups of patients or other clinical features.
Blinding	Pathological scoring was blinded. We only gave an anonymized set of slides for scoring to the two pathologists involved in the study. Data were married back after the scoring was finalized.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☐	☒ Human research participants

Antibodies

Antibodies used	WESTERN BLOT: For SLC9A3R1 we used HPA027247 (protein atlas) at 1:1000 dilution, for YY we used Santa Cruz; sc-281 at 1:500 dilution. For GAPDH we used Abcam #ab9385 at 1:5000 dilution. For IHC: For YY1 (Protein Atlas HPA001119, Atlas Antibodies Cat#HPA001119, RRID:AB_1858930) the flowing conditions were used: tissue sections were incubated with the primary monoclonal. overnight at 4°C, and chromogen development was performed using the Envision system (DAKO Corporation, Glostrup, Denmark). A minimum of 500 tumor cells were scored with the percentage of tumor cell nuclei in each category recorded. For SLC9A3R1 (HPA9672 and HPA27247, Atlas Antibodies Cat#HPA009672, RRID:AB_1857215 and Atlas Antibodies Cat#HPA027247, RRID:AB_10601162 respectively) the following conditions were used. HPA9672 was diluted 1:400 and HPA27247 was diluted 1:1500. Staining was automatized with a Ventana Benchmark-Ultra using epitope retrieval ER2 for 20 minutes. ER and PgR immunoreactivity was assessed by the FDA-approved

ER/PR PharmDX kit (Dako). The prevalence of ER/PgR positive invasive cancer cells, independent of their staining intensity, was quantitatively annotated in the original reports. In accordance with ASCO/CAP guidelines, tumors with ≥1% of immunoreactivity was considered positive

For ChIP: Immunoprecipitation using 4ug of H3k27ac antibodies (Abcam; ab4729) per ChIP experiment or using 4ug of YY1 antibodies (Santa Cruz; sc-281 X).

Validation

All the antibodies were commercially available and pre-validated using orthogonal methods (RNA-ICH correlation, siRNA, Protein/ peptide array and Mass Spec) . For IHC we used two independent antibodies to increase robustness.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Philippa Darbre, who received MCF7 cells on 21 October 1987 from Kent Osborne at passage 390 and called "MCF-7 McGrath". They were as described in his paper in detail of that year (Kent Osborne et al 1987 Biological differences among MCF-7 human breast cancer cell lines from different laboratories. Breast Cancer Res Treat 9: 111-121

Authentication

Authentication Karyotyping was perfomed for all cell lines

Mycoplasma contamination

Mycoplasma has been routinely tested throughout the study (once a week) and confirmed negative.

Commonly misidentified lines
(See ICLAC register)

None of the cell lines used are listed in the ICLAC database

Human research participants

Policy information about studies involving human research participants

Population characteristics

Participants were selected based on histo-pathological data (luminal invasive breast cancer, estrogen receptor positive). No selection was applied on grade, node, stage, size or age. All tissues were frozen. No covariate-relevant characteristics were collected excluded being ER-positive.

Method-specific reporting

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Data have been submitted to EBI and can be accessed using the PRJEB22757 code

Files in database submission

Raw reads and Peak files

Genome browser session
(e.g. UCSC)

NA

Methodology

Replicates

No Replicates are available for the in vivo part of the study. Two replicates were performed for YY1 ChIP-seq in cell lines

Sequencing depth

Sequencing depth At least 40M reads were used for each experiments.

Antibodies

H3K27ac was acquired from AbCam (ab4729). YY1 was bought from Santa Cruz (sc-281 X)

Peak calling parameters

All the details of the analysis are reported in the supplementary computational method file.

Data quality

All the details of the analysis are reported in the supplementary computational method file.

Software

All the details of the analysis are reported in the supplementary computational method file.