1 Enhancers mapping uncovers phenotypic heterogeneity and evolution in 2 patients with luminal breast cancer

3

4 Darren K. Patten^{*1}, Giacomo Corleone^{*1}, Balázs Györfy^{2,3}, Edina Erdős⁴, Alina
5 Saiakhova⁵, Kate Goddard⁶, Andrea Vingiani⁷, Sami Shousha⁸, Lőrinc Sándor
6 Pongor², Dimitri J. Hadjiminis⁸, Gaia Schiavon⁹, Peter Barry¹⁰, Carlo Palmieri¹¹, Raul
7 C. Coombes¹, Peter Scacheri⁵, Giancarlo Pruner¹², Luca Magnani^{#1}.

8

9 * Equal contribution

10 #To whom correspondence should be addressed: l.magnani@imperial.ac.uk

11

12 ¹Department of Surgery and Cancer, The Imperial Centre for Translational and
13 Experimental Medicine, Imperial College London, Hammersmith Campus, London,
14 U.K.

15 ²MTA TTK Lendület Cancer Biomarker Research Group, Institute of Enzymology,
16 Hungarian Academy of Sciences, 1117, Budapest, Hungary

17 ³Semmelweis University 2nd Dept. of Pediatrics, 1094, Budapest, Hungary

18 ⁴Department of Biochemistry and Molecular Biology, Genomic Medicine and
19 Bioinformatic Core Facility, University of Debrecen, Debrecen 4032, Hungary

20 ⁵Department of Genetics and Genome Sciences, Case Comprehensive Cancer
21 Center, Case Western Reserve University, Cleveland, OH 44106

22 ⁶Department of Breast and General Surgery, Charing Cross Hospital, Imperial
23 College Healthcare NHS Trust, London, U.K.

24 ⁷Department of Pathology, European Institute of Oncology, Milan

25 ⁸Centre for Pathology, Department of Medicine, Imperial, College London, Charing
26 Cross, London, UK

27 ⁹Translational Science, IMED Oncology, AstraZeneca, Cambridge, UK.

1 ¹⁰Department of Breast Surgery, The Royal Marsden NHS Foundation Trust,
2 Orchard House, Downs Road, Sutton, SM2 5PT, UK.

3 ¹¹Institute of Translational Medicine University of Liverpool, Clatterbridge Cancer
4 Centre, NHS Foundation Trust, and Royal Liverpool University Hospital, Liverpool,
5 Merseyside, UK

6 ¹²Pathology Department, Fondazione IRCCS Istituto Nazionale Tumori and
7 University of Milan, School of Medicine.

8

9

10

1 **Abstract**

2 The degree of intrinsic and interpatient phenotypic heterogeneity and its role in
3 tumour evolution is poorly understood. Phenotypic divergence can be achieved via
4 the inheritance of alternative transcriptional programs^{1,2}. Cell-type specific
5 transcription is maintained through the activation of epigenetically-defined regulatory
6 regions including promoters and enhancers^{1,3,4}. In this work, we annotated the
7 epigenome of 47 primary and metastatic oestrogen-receptor (ER α)-positive breast
8 cancer specimens from clinical samples, and developed strategies to deduce
9 phenotypic heterogeneity from the regulatory landscape, identifying key regulatory
10 elements commonly shared across patients. Highly shared regions contain a unique
11 set of regulatory information including the motif for the transcription factor YY1. *In*
12 *vitro* work shows that YY1 is essential for ER α transcriptional activity and defines the
13 critical subset of functional ER α binding sites driving tumor growth in most luminal
14 patients. YY1 also control the expression of genes that mediate resistance to
15 endocrine treatment. Finally, we show that H3K27ac levels at active enhancer
16 elements can be used as a surrogate of intra-tumor phenotypic heterogeneity, and to
17 track expansion and contraction of phenotypic subpopulations throughout breast
18 cancer progression. Tracking YY1 and SLC9A3R1 positive clones in primary and
19 metastatic lesions, we show that endocrine therapies drive the expansion of
20 phenotypic clones originally underrepresented at diagnosis. Collectively, our data
21 show that epigenetic mechanisms significantly contribute to phenotypic
22 heterogeneity and evolution in systemically treated breast cancer patients.

23

24

25

26

27

28

29

1 Introduction

2 Breast cancer (BC) is the most common cancer type and the second most frequent
 3 cause of cancer related death in women⁵. 70% of all BC cases contain variable
 4 amounts of oestrogen receptor-alpha (ER α) positive cells. ER α is central to BC
 5 pathogenesis and serves as the target of endocrine therapies (ET)^{6,7}. ER α -positive
 6 BC is typically subdivided in two 'intrinsic' molecular subtypes (luminal A and luminal
 7 B⁸) characterized by distinct prognosis, highlighting functional inter-patient
 8 heterogeneity. Recent analyses demonstrate that patient-to-patient heterogeneity is
 9 more pervasive (reflected by histological⁹, genetic architecture¹⁰ and transcriptional¹¹
 10 differences) ultimately influencing long-term response to endocrine treatment¹².
 11 Indeed, 30-40% of ER α BC patients relapse during or after completion of adjuvant
 12 endocrine therapies. At the time of relapse, almost all patients will have developed
 13 resistance to ET, partly through treatment-specific genetic evolutionary trajectories¹³.
 14 Additionally, the presence of genetic intra-tumor heterogeneity has also now been
 15 extensively documented in several cancer types, demonstrating the role of clonal
 16 evolution in cancer¹⁴. Yet, recent studies have shown that driver coding-mutations
 17 do not significantly change between primary and metastatic luminal breast cancer,
 18 with the notable exception of *ESR1* mutations¹⁵, suggesting that alternative
 19 mechanisms might contribute to BC progression and drug-resistance. Parallel to
 20 genetic evolution, phenotypic/functional changes driven by epigenetic mechanisms
 21 can also contribute to breast cancer progression and ET resistance in cell lines^{16,17}.
 22 Nevertheless, little is known about the epigenome of BC patients, its influence on
 23 intra-tumour phenotypic heterogeneity and its role in breast cancer progression.

24 Epigenetic modifications consist of chemical modifications targeting the
 25 DNA/RNA (e.g. DNA methylation) and the chromatin (histone modifications). Histone
 26 modifications have been successfully used to map regulatory regions and to
 27 annotate the non-coding DNA^{1,3}. Acetylation of lysine 27 on histone 3 (H3K27ac) is
 28 strongly associated with promoters and enhancers of transcriptionally active genes
 29 ^{4,18,19}. Increasing evidence suggests that epigenetic information can actively transfer
 30 gene transcription states across cell division²⁰⁻²³. Epigenetic modifications play also
 31 a central role in modulating ER α binding to the DNA possibly by interacting with
 32 ER α -associated pioneer factors^{24,25}. Finally, *in vitro* studies have shown that

epigenetic reprogramming might play a central role in ER α BC cells that adapt to endocrine therapy^{16,17}.

Here we show the results of a systematic investigation of the epigenetic landscape of ER α positive primary and metastatic breast cancer from 47 individuals. Our results represent the first large scale topographic mapping of the active regulatory landscape of longitudinal ER α -positive BC. Using H3K27ac we mapped active promoters and enhancers across treatment naïve primary and endocrine treated metastatic patients. We used bioinformatic approaches to deconvolute the complex regulatory landscape and identified inter- and intra-patient epigenetic heterogeneity. We mined promoters and enhancers from clinically relevant breast cancer samples for potential regulatory drivers identifying YY1 as a novel key player in ER α -positive BC. Finally, we demonstrate that epigenetic mapping can efficiently estimate phenotypic heterogeneity changes throughout BC progression.

Results

Mapping enhancers and promoters of primary and metastatic ER α positive breast cancer

To build a comprehensive compendium of all the clinically relevant active regulatory regions of luminal BC we profiled fifty-five ER α positive BC samples (primary n=39, and metastatic n=16) with H3K27ac ChIP-seq (Supplementary Table S1). To minimize the introduction of noise from non-tumor tissues we used samples with high tumor burden (>70%, Supplementary Figures S1). 85% of samples yielded satisfactory results (47/55, Supplementary Figure 2A and Table S2). H3K27ac-enriched regions were classified into 23,976 gene-proximal (1kb upstream of transcription start site (TSS), promoters) and 326,719 gene-distal (enhancers). Considering the ten-fold difference in H3K27ac signal, it was not surprising to observe that 80% of promoters can be captured by profiling 4 individual, while nearly 40 are needed to reach the same coverage for enhancers, as indicated by saturation plots (Supplementary Figure 2B). These data are in agreement with enhancers being the main determinants of cell-type specific transcriptional differences^{4,18,26,27}. To gain insights on the penetrance of each regulatory region, we developed a Sharing Index (SI) by annotating all enhancers and promoters in function of the number of

patients sharing the H3K27ac signal at each specific location (Supplementary Figure 2C). In agreement with saturation analyses, we find that a large portion of enhancers are patient-specific (SI=1) while active promoters are more commonly shared between patients (Supplementary Figure 2C). Collectively, these data demonstrate that enhancers account for the majority of potential epigenetic heterogeneity in ER α -positive BC.

Enhancer activity allows the qualitative assessment of phenotypic heterogeneity

Genetic heterogeneity is an hallmark of most solid tumours²⁸. Nonetheless, genetic intra-tumoral heterogeneity does not often directly translates to phenotypic heterogeneity. In agreement, despite extensive inter- and intratumoral clonal genetic diversity²⁹, the majority of ER α -positive tumors benefit from systemic ET¹². Likewise, treatment-naïve metastatic patients generally respond well to ET, at least initially, suggesting that genetic heterogeneity on its own cannot explain treatment resistance/response. On the other hand, phenotypic hierarchies can override genetic hierarchies in brain cancers^{2,30}, suggesting that inheritable epigenetic program might ultimately contribute to phenotypic heterogeneity and treatment outcome.

The existence of intra-tumoral phenotypic heterogeneity in breast cancer patients has been known to pathologists for decades, at least for a small number of biomarkers. For example, immunohistochemistry (IHC) assessment of the proportion of ER α -positive within luminal cancer patients varies on a *continuum* from less than 1% to nearly 100%³¹. Unfortunately, IHC is low-throughput method, and typically only a few proteins can be studied before the sample is consumed. In contrast, assessing phenotypical heterogeneity from patient bulk transcriptional data is unfeasible as transcription is ultimately an analogue signal in which each individual cell can contribute a stochastic amount of RNA, making data deconvolution impractical (Fig 1A). For instance, cells with focal gene amplification have higher bulk gene expression but individual cells can contribute radically different amounts as shown by single-molecule single-cell RNA FISH¹³. On the other hand, recent evidence show that the signal captured by chromatin assays such as ATAC-seq appears to be directly proportional to the cells contributing to it³². Similarly, ChIP-seq signal can be

1 thought of as digital information with each single nucleosome being ON (K27ac) or
 2 OFF at any given time (Fig. 1A). Notably, even within genetically clonal cell lines, the
 3 H3K27ac signal varies considerably between different regulatory regions. Regulatory
 4 regions labelled as super enhancers, for example, have 10-100-times more
 5 H3K27ac signal than typical enhancers¹⁸. What accounts for the variation in signal is
 6 not known, but one possibility is that heterogeneity within the cell population (either
 7 clonal or sub-clonal) contribute to signal intensity. For example, super-enhancers
 8 might represent regulatory regions active across most or all cells within a population
 9 at any given time (clonal, H-peaks), while “typical” enhancers with lower H3K27ac
 10 signal may represent sub-clones (M and L peaks, Fig. 1A). This concept is similar to
 11 measuring variant allele frequencies (VAF) to infer genetic heterogeneity.
 12 Phenotypical heterogeneity might be the consequence of heterogeneous cell
 13 populations (i.e. tumor, stroma and immune infiltrate) or actual cancer-specific
 14 epigenetic subclones. As our ChIP-seq data are derived from high tumor burden
 15 samples, we hypothesized that H3K27ac signal could allow for a qualitative
 16 assessment of phenotypic heterogeneity. We further theorised that direct correlation
 17 might exist between clonal prevalence (intra-tumour) and population prevalence
 18 (inter-patients) (Fig. 1A).

19 We tested the initial assumption by performing spike-in experiments in which
 20 known numbers of cells with well-characterized regulatory region activity (and similar
 21 genetic background)¹⁶ were admixed in incremental proportions prior to H3K27ac
 22 ChIP-qPCR. The data shows that H3K27ac enrichment is proportional to the number
 23 of cells with the active enhancer (Fig. 1B). These findings are corroborated by an
 24 independent analysis using a different antibody (ER α) (Supplementary Figure S3A).
 25 As the signal between different patients is not directly comparable, we normalized
 26 the data using a ranking approach, assigning to each H3K27ac signal a Rank Index
 27 (RI, 1 to 100, strongest to weakest) (Fig. 1C). Signal from low RI (H peaks) might be
 28 associated with clonal regulatory regions active in almost all cells. Conversely, high
 29 RI (M-L peaks) mark more heterogeneous/sub-clonal enhancer activity. By
 30 investigating the relationship between RI and SI we found extremely high correlation
 31 between these two parameters (Fig. 1D), suggesting that clonal regulatory regions
 32 are more common between patients while sub-clonal regulatory elements are more
 33 patient-specific. We defined clonal low-RI/high-SI loci as regulatory drivers (RD,
 34 SI>21) and high-RI/low-SI that might originate from sub-clonal populations as

regulatory noise (RN, SI<21). Of note, a small but discrete proportion of promoters/enhancers escape this general trend having extremely low RI despite being patient-specific or higher RI while being shared (dot-boxes, Fig. 1D).

Enhancers are associated with BC risk-SNP and control gene transcription

We next investigated the extent to which regulatory regions identified in our cohort associate to BC. Previous analyses from ER α -BC cell lines, have shown that genetic predisposition might occur through SNPs that modulate transcription factors binding at enhancers (FOXA1 and ER α ³³). We then tested the relationship between DNA risk variants specifically associated with BC through GWAS^{33,34} and regulatory regions captured in patients. Strikingly, almost the totality of validated BC risk variants is contained within our H3K27ac database. Currently, this dataset represents the most enriched annotation for GWAS variants in breast cancer (Fig 1E). This overlap is highly significant specifically for enhancers but not for other annotations (Fig. 1F). Notably, this association is not replicated using colorectal cancer risk variants suggesting that these enhancers might play a specific role in BC development (Fig. 1F).

Next, we assessed the relationship between estimated enhancers clonality and transcriptional output. Transcriptional data obtained using microarray and RNA-seq estimate the average expression level within a population. The average expression is function of the number of cells engaged in active transcription and the number of RNA molecule within each cell³⁵. Interestingly, several lines of evidence suggest that RNA transcription is stochastic^{13,36}, thus implying that the total number of cells with active transcription significantly contribute to changes in average RNA levels in bulk populations. As our analysis allows for qualitatively prediction of clonality in enhancer activity, it allows to test if clonal enhancers active in the majority of cells correlate with higher RNA levels. To do so we linked enhancers to their potential target genes using CTCF insulated boundaries³⁷. We then analysed three independent BC expression datasets (METABRIC¹⁰, TCGA³⁸ and Affymetrix³⁹) in function of RI/SI indexes. Our analyses show a predictable increase for mRNA levels with parallel increases in the associated SI, further suggesting that RDs drive RNA expression in a progressively increasing number of cells (Supplementary Figure 3B). These results were more modest when analysing the transcriptome from normal

breast tissue (Supplementary Figure 3B, small insets) suggesting that our analysis has identified a subset of regulatory regions strongly associated with malignant outgrowth. These data indicate that transcripts identified as dis-regulated in BC might reflect changes in the size of phenotypic subpopulations. This could be in part driven by a selection process, as normal breast transcriptional data reflect the heterogeneous composition of the tissue (adipocyte, myoepithelial and epithelial cells), while BC is normally dominated by epithelial morphology. Collectively, our data show that enhancer activity strongly tracks transcriptional changes in breast cancer patients.

Imputed transcription factors landscape of ER α breast cancer patients

Enhancers stores regulatory information in the form of transcription factors (TFs) binding motifs⁴⁰. The vast majority of TFs require accessible chromatin in order to bind their cognate DNA sequences⁴¹. We reasoned that a systematic investigation of the predicted TFs landscape in function of enhancer activity could reveal potential transcriptional BC drivers. To narrow down to the accessible DNA within active enhancers and promoters we integrated the DNaseI signal from 129 cell lines with the inferred nucleosome pattern obtained from the H3K27ac data (Fig. 2A). Initial analyses collapsing all imputed DHS in relationship with their enhancers and promoter location identified correctly well-known BC-TFs according to their promoter–enhancer bias (Supplementary Figure 4). We then stratified the complete set of enhancers and promoter regions based on the associated SI and repeated TF motif analysis focusing within each SI-defined bin followed by unsupervised clustering. This analysis generated two major clades (Fig. 2B), indicating the presence of different classes of regulatory regions. Strikingly, we find that RD and RN enhancers and promoters cluster specifically into the two major clades, suggesting that putative clonal and sub-clonal enhancers contain distinct regulatory information (Fig. 2B). Functional TF binding is associated with TF leaving a footprint within chromatin accessible regions^{40,42}. Interestingly, clonal enhancers in ER α -positive MCF7 breast cancer cells are significantly enriched in TF footprints¹⁶, while sub-clonal enhancers are significantly deprived of footprints suggesting that TFs might bind clonal enhancers with longer residence time⁴² (Fig. 2C).

1 We then focused on estrogen-response elements (ERE) as they constitute the
2 canonical DNA sequence to which ER α binds. Unexpectedly, ERE motifs are
3 enriched only in RN enhancers, suggesting that a significant amount of ER α binding
4 occurs in sub-clonal/less functional enhancers (Fig. 2B). To gain further insights on
5 ER α dynamics we turned to a recently published ER α dataset obtained from patient
6 material (n=15)⁴³. Generally, the proportion of ER α binding sites overlapping
7 enhancers increase with the SI (9% vs. 70%, SI-1 vs. SI-39, Fig. 2C). This was not
8 observed for promoters (15% vs. 6%, SI-1 vs. SI-39, Fig. 2C), and is consistent with
9 previous studies demonstrating a significant bias for ER α binding at active enhancer
10 elements^{44,45}. These data imply that shared enhancers have a strong propensity for
11 ER α binding despite being generally under-enriched in EREs (Fig. 2B). More
12 importantly, the bulk of ER α binding were captured only once in fifteen patients (ER α
13 SI=1), with less than 0.003% of ER α being shared across 75% of the patients (484
14 core ER α)⁴³ (Fig. 2D). Together, these data support biochemical evidence that
15 suggest that only a small fraction of ER α binding events with longer-residency time is
16 functional⁴². We therefore concluded that the largest portion of ER α binding identified
17 in patients occur at patient specific, sub-clonal enhancers and might be the
18 consequence of the transient ER α -DNA interactions occurring while the receptor
19 scans the genome⁴². The discrepancy between the small number functional ER α
20 core binding and the observation of ERE-poor RD enhancers led us to hypothesize
21 that other TFs might collaborate with ER α to increase its transcriptional efficiency at
22 clonal enhancer. Most TF motifs enriched in RD enhancers are also largely observed
23 in RN regions, with the notable exception of the motif for the transcription factor YY1
24 (Fig. 2B). Interestingly, TF analysis of the footprints within MCF7 clonal enhancer
25 (Fig. 2C, RI<20) similarly identifies YY1 as the top hit (qVal=0.001). YY1 has been
26 recently implied in *de novo* formation of enhancer promoter looping during neural
27 development⁴⁶ and MYC-like ability to potentiate gene expression⁴⁷ indicating a
28 potential role in modulating the enhancer landscape in ER α -positive BC.

29

30 **YY1 enhancer activity mark a dominant phenotypic clone in BC**

31 YY1 is a ubiquitously expressed TF (Supplementary Figure 5A-B) that can act as an
32 activator or repressor by binding DNA, RNA and chromatin modifiers^{48,49}. YY1

1 function in breast cancer is poorly understood, as it has been linked to different
2 outcomes depending on breast cancer cells subtypes. In luminal BC, YY1 appears to
3 be positively correlated with AP2 to promote HER2 activity⁵⁰, while in triple negative
4 BC models appears to be a tumor suppressor by controlling BRCA1 expression⁵¹.
5 Interestingly, YY1 drosophila homolog PhoRC is involved in epigenetic memory by
6 recruiting of Polycomb repressor complex to sequence specific regions⁵², but YY1's
7 role in mammals is not entirely understood^{46,47,53}. Our TF analysis shows that YY1
8 might actually operate as a global reader of active clonal enhancers common to the
9 majority of ER α -positive patients. These data therefore predict that most luminal
10 breast cancers should contain a dominant YY1-positive clone. To assess the size of
11 YY1 phenotypic clone in our patient's dataset we identified the *bona fide* enhancers
12 looping at YY1 promoter using 3D chromatin maps⁵⁴ (Supplementary Fig. 6A). We
13 found three potential enhancers with high SI within a CTCF-insulated region with
14 YY1 promoter (SI A=41, B=33 and C=26, Supplementary Fig. 6A). Interestingly
15 Enhancer A also directly interacts with Enhancer B-C, suggesting a multi-enhancers
16 interaction with YY1 promoter. Enhancer A consistently ranks among the most clonal
17 enhancer in nearly all patients, suggesting that YY1 is transcribed in almost all cells
18 (Fig. 3A). By comparison, in most normal tissues profiled by H3K27ac within the
19 Epigenome Roadmap consortium⁴⁵, YY1 Enhancer A activity is more variable while
20 remaining relatively dominant, implying that some tissues may harbour YY1-
21 subclonal subpopulations (Fig 3B). Consistent with these predictions,
22 immunocytochemistry (IHC) meta-analysis (Fig 3B) showed a decreasing number of
23 YY1 positive cells in correspondence to increasing RI scores (Insets, Fig 3B and
24 Supplementary Figure 6B). Collectively, these data suggest that enhancer ranking
25 can capture qualitative changes in intra-tumoral heterogeneity, and that YY1-
26 enhancer activity marks a dominant phenotypic clone in ER α -positive BC.

27 Next, we looked at the significance of YY1 mRNA expression in a pan-cancer
28 analysis and found that tumor tissue generally have significantly higher expression
29 level for YY1 as compared to normal tissues (Supplementary Figure 7A). This does
30 not appear to reflect simply the proliferation status as we found no correlation
31 between YY1 and Ki67 expression in 2509 Breast Cancer patients⁵⁵. This
32 observation was replicated in an independent large BC dataset as well (Fig. 3C and
33 Supplementary Figure 7B). Of note, YY1 is not subject to recurrent genomic

1 aberrations (data not show). These data suggest that cancer lesions might contain a
 2 larger fraction of YY1-expressing cells as compared to more heterogeneous tissues
 3 (Fig. 3B). Meta-analysis of the METABRIC¹⁰ datasets shows that patients with higher
 4 YY1 mRNA level at diagnosis have significantly worse outcome (Fig. 3C). We
 5 replicated this finding using TCGA RNA-seq data with significant stratification
 6 occurring in Luminal A patients, a subtype typically associated with good prognosis
 7 (Fig. 3C). To test if YY1 increased mRNA expression could be driven by an
 8 expansion of YY1-positive cells from a more heterogeneous population, we stained
 9 normal breast section with IHC. Our data show that lobules and ducts contain distinct
 10 YY1 positive sub-clonal populations within the luminal and basal compartments (Fig
 11 3D). On the other hand, nearby tumour tissue is overwhelmingly YY1 positive,
 12 demonstrating the existence of a clonal YY1 population. The data demonstrate that
 13 the expansion of a YY1 phenotypic clone drive the changes in bulk RNA levels
 14 between normal and tumor samples and reinforce the notion that YY1 might play a
 15 central role in ER α -positive BC.

16

17 **YY1 is a global enhancers modulator marking functional ER α binding**

18 The TF analysis in our epigenomic patient dataset revealed YY1 uniquely
 19 associated with clonal regulatory regions and potentially collaborating with ER α at
 20 critical enhancers. To gain mechanistic insight on the role of YY1 we performed YY1
 21 ChIP-Seq in quiescent (estrogen-deprived) and estrogen-stimulated luminal BC
 22 MCF7 cells. Cells were stimulated with estrogen for 45 minutes, upon which
 23 maximum ER α -binding to chromatin occurs⁴⁵. ChIP-seq biological replicates show
 24 very high correlation ($R^2=0.98$), thus we kept consensus loci for further analyses. In
 25 quiescent cells, YY1 occupies a very small set of enhancers and promoters near
 26 housekeeping genes⁵⁶ (Fig. 4A). Strikingly, estrogen stimulation induce a 23-fold
 27 expansion of the YY1 binding repertoire, mostly at enhancer regions (Fig. 4A). Newly
 28 occupied loci are associated with ER α -BC signatures and epigenetic editors (Fig.
 29 4A). Interestingly, only ~10% of all binding is characterized by a high affinity YY1
 30 motif suggesting that induced YY1 could also bind directly to modified nucleosomes
 31 through its chromatin remodelling partner INO80⁵⁷. Orthogonal analyses show that
 32 induced-YY1 binding involves almost all MCF7 active regulatory regions and is

strongly associated with H3K27ac marks (Fig. 4B). Nonetheless, it is unlikely that YY1 binding directly promote or requires H3K27ac as we did not find any difference between quiescent (estrogen-deprived) and estrogen-stimulated H3K27ac epigenomes (data not shown). Conversely, YY1 binding is absent from silenced genes (Supplementary Figure 7C), demonstrating that YY1 does not associate with PRC2 mediated repression in BC cancer cells.

Our *in vivo* analysis suggest that clonal YY1-bound enhancers are generally not enriched for EREs or ER α , with the exception of the atypical core-ER α (Fig 2B). In agreement, our *in vitro* data show only marginal overlap between YY1 and ER α or its pioneer factor FOXA1 (Fig. 4B-D) indicating that generally YY1 recruitment is independent from ER α . On the other hand, YY1, ER α and FOXA1 co-localize at increased frequencies at core-ER α loci in MCF7 cells (Fig. 4C). Similar observations were made by comparing YY1 overlap with *in vivo* derived ER α binding (60% overlap with core ER α vs. 18% overlap with patient-unique ER α). In addition, we find that genes defining the luminal subtype in TCGA patients are significantly enriched for ER α core binding with YY1 but not patient-unique ER α (Fig. 4D). Overall, these data further suggest that YY1 might stabilize ER α binding⁴² at a small subset of transcriptionally productive enhancers (core-ER α) captured in most tumor cells and most patients⁴³. To test if YY1 can contribute to ER α driven transcription, we measured luciferase activity from a promoter driven by an array of estrogen response elements (EREs) in MCF7 cells in presence or absence of YY1 (Fig. 4E) and show absolute dependencies on YY1. Furthermore, YY1 depletion also abrogates cell proliferation in response to estrogen stimulation in MCF7 (Fig. 4F) suggesting that YY1 is a direct driver of the clonal proliferation observed in the BCa (Fig. 3D-E). These observations were replicated in other independent luminal BC cell models (ZR75 and T47D, Supplementary Figure 8A-B). Finally, we show that YY1 depletion leads to significant downregulation of core-ER α target genes in luminal BC cell line models (Supplementary Figure 8C). Collectively these data identify YY1 as a novel essential transcription factor significantly contributing to ER α regulatory network transcriptional activity.

YY1 contributes to drug-resistance in luminal BC

YY1 motif is highly enriched in clonal enhancers identified in primary and metastatic luminal patients (Fig 2B). All metastatic patients included in this study relapsed following adjuvant endocrine therapies suggesting that YY1 might also play role in this setting. In agreement, primary and metastatic samples show clonal YY1 enhancer activity, indicating that YY1 positive cells are not effectively cleared by the therapy (Fig. 3A). Therefore, we investigated the role of YY1 in LTED cells, an MCF7-derivative that develop estrogen-independent growth partly through constitutive activation of ER α signalling¹⁶. YY1 depletion leads to complete abrogation of LTED growth demonstrating that YY1 is still required at this stage (Fig. 4G). Interestingly, LTED cells have an expanded repertoire of ER α binding compared to MCF7, fuelled by endogenous ligands^{13,16}. The set of enhancers engaged by ER α and YY1 in LTED cells is radically different compared to MCF7, with the majority of ER α -YY1 being specific to each cell type (Fig 4I, LTED only: 3598/5037). ER α -YY1 bound enhancers in LTED strongly associates with the transcription of genes involved with acquired endocrine therapy, suggesting that during epigenetic reprogramming¹⁶, YY1 might stabilize ER α to LTED specific enhancers (Fig. 4J). To further examine the relationship between YY1 and endocrine resistance we analyzed a set of estrogen responsive genes whose transcription cannot be antagonized by Tamoxifen treatment in MCF7 cells⁵⁸. These genes were not enriched for patient-private ER α , but we saw an ever-increasing association with ER α -YY1 bound enhancers, especially with core ER α -YY1 (Fig. 4K). For example, ER α -YY1 is found near CXXC5 and SLC9A3R1, ranked respectively first and second as the most strongly estrogen-induced gene that cannot be antagonized by Tamoxifen⁵⁸. Collectively, these data strongly support the role of YY1 in ER α BC growth and progression.

26

27 **YY1-ER α promote SCL9A3R1 expression despite endocrine treatment**

28 SLC9A3R1 (NHERF1/EBP50)⁵⁹ SLC9A3R1 encodes a Na/H exchanger regulatory
29 cofactor. SLC9A3R1 null mice have disrupted protein-kinase-A-dependent cAMP-
30 mediated phosphorylation⁶⁰. In agreement with SLC9A3R1 potential role in
31 endocrine resistance, meta-analysis of patient-derived data using all available genes
32 (n=22,277) reveals that SLC9A3R1 expression is amongst the top 1% of genes with

1 the strongest prognostic association with relapse in a cohort of 724 ET-treated ER α -
2 positive patients ³⁹ (Fig. 5A). High expression of SLC9A3R1 also significantly
3 correlates with poor survival in additional independent ER α -BC datasets
4 (Supplementary Figure 9A). In addition to Tamoxifen treatment, SLC9A3R1 remains
5 transcriptionally active in most endocrine therapy resistant BC cell lines that retain
6 ER α expression (Supplementary Figure 9B) but genetic or pharmacological
7 (Fulvestrant) suppression of ER α is sufficient to block SLC9A3R1 transcription
8 (Supplementary Figure 9B-C). Specifically, SLC9A3R1 expression is not
9 antagonized by estrogen deprivation⁶¹ (LTED models, Supplementary Figure 9B-D),
10 nor Raloxifene *in vitro* (Supplementary Figure 9E) or neo-adjuvant AI treatment in
11 clinically treated patients *in vivo* (Fig. 5B). Overall, these data demonstrate that
12 SLC9A3R1 is a direct ER α target whose expression cannot be antagonized by first-
13 line endocrine therapies.

14 Bulk RNA-seq data from a panel of cancer cell lines demonstrate that ER α -
15 positive BC cells have the highest levels of SLC9A3R1 mRNA (Supplementary
16 Figure 10A). More importantly, TCGA RNA-seq analysis shows that SLC9A3R1
17 expression is higher specifically in ER α BC patients compared to normal tissue or
18 other subtypes (Supplementary Figure 10B). Chromatin analyses of MCF7 and
19 LTED cells identify ER-bound enhancers at 3 independent loci within the insulated
20 SLC9A3R1 locus (E1-E3), a RD region directly looping to the YY1-bound SLC9A3R1
21 promoter within a CTCF insulated perimeter (Supplementary Figure 10C). Strikingly,
22 E1 and E2 contain core ER α bindings in addition to YY1, while E3 contains a patient
23 unique ER α binding with no YY1 (Supplementary Figure 10C). *In vivo* transcriptional
24 analysis demonstrates that SLC9A3R1 is the only gene near the E1-E2 enhancers
25 that shows a significant increase in bulk-RNA level when comparing normal breast
26 tissue with ER α -positive BC (Supplementary Figure 10D). Interestingly, enhancer-
27 activity appears to be immune to endocrine therapy (MCF7 vs. MCF7 tamoxifen
28 resistant and LTED, Supplementary Figure 10C). Collectively, these data strongly
29 support the notion that SLC9A3R1 expression is driven by a breast cancer specific
30 enhancer within the expanding ER α -YY1 clone during tumor initiation. Nonetheless,
31 SLC9A3R1 expression is dependent on YY1 (Supplementary Figure 11A),
32 demonstrating that both ER α and YY1 are essential for full enhancer activity.
33 Silencing SLC9A3R1 is sufficient to abrogate oestrogen-induced growth in ER α -

positive cells (Fig 5C). Intriguingly, SLC9A3R1 is not essential for a second in ER α -positive model (T47D) but appears to be a critical gene for both AI-resistant cells models (Fig. 5C and Supplementary Figure S11B). Collectively, these data demonstrate that ER α -YY1 regulate SLC9A3R1 via enhancer binding and identify SLC9A3R1 as a novel player involved in ET resistance.

Mapping phenotypic heterogeneity using YY1 and SLC9AR1 enhancer activity

Both SLC9A3R1 and YY1 enhancers are commonly activated in our patient's dataset (SI=34 and SI=41 respectively). Yet, YY1 enhancer identifies YY1-positive cells as a dominant clone in almost all patients (RI $\leq$ 20, Fig 3A). Conversely, SLC9A3R1 enhancer activity indicates that SLC9A3R1 marks a potentially dynamic sub-clonal population in most primary patients (RI $\geq$ 20, Fig 5D). Our *in vitro* data suggest that SLC9A3R1 transcription cannot be antagonized by endocrine therapies while SLC9A3R1 is important for resistant BC cell lines. This predicts that the SLC9A3R1-positive population should increase under adjuvant treatment. Interestingly, the only evidence of a clonal SLC9A3R1 population was found in samples from three metastatic, endocrine-resistant patients (Fig. 5D).

Bulk transcriptional data show that average SLC9A3R1 expression is significantly higher in ER α positive BC cells but do not inform about potential subpopulations (Supplementary Figure 10A-B). Conversely, meta-analysis of SLC9A3R1 enhancer activity (RI) within the ENCODE H3K27ac datasets indicates that MCF7 cells contain a clonal SLC9A3R1 population while all other cell lines appear to have decreasing sub-clonal populations (Supplementary Figure 11C). Of note, the size of the sub-clonal population correlates with total RNA content for the cells contained in both assays, suggesting that the decreasing bulk RNA signal is driven by a progressively smaller subpopulation (Supplementary Figure 11C). Similar analysis of YY1 enhancer indicate that cancer cell lines are vastly clonal for YY1 expression (Supplementary Figure 11D) in agreement with clinical samples. Notably, both YY1 and SLC9A3R1 RIs in mammary epithelial cells suggest the presence of sub-clonal populations. These observations fit well with experimental data from IHC profiles from normal breast (Fig. 3D and Supplementary Figure 12B). Meta-analyses of H3K27ac from the Epigenome Roadmap database predict that most tissues

1 potentially have only sub-clonal populations, as determined by SLC9A3R1 enhancer
2 activity (Fig. 5E). In agreement, the size of the sub-clonal population tracks the
3 mRNA signature of each tissue with SLC9A3R1 potentially clonal tissues
4 accumulating the highest amount of RNA-seq tags (Small and Large Intestine,
5 Supplementary Figure 12). Analogously to YY1, meta-analysis of IHC data identifies
6 decreasing SLC9A3R1-positive with increasing RI scores (Fig. 5E and
7 Supplementary Figure 12B). Finally, to further validate that RI index can estimate
8 phenotypic clones, we retrospectively collected available FFPE biopsies for the BC
9 patients profiled with H3K27ac ChIP-seq (n=19). We then performed IHC using YY1
10 (Fig. 5F) and SLC9A3R1 (Fig. 5G) antibodies and compared the predicted enhancer
11 activity (RI) with the actual size of YY1 and SLC9A3R1-positive clones within each
12 patient. With the exception of one metastatic sample (M3), YY1 staining robustly
13 correlate with RI, confirming large clonal YY1 positive populations in all examined
14 tissues (Fig. 5F). In parallel, SLC9A3R1 enhancer activity correctly estimated the
15 size of the sub-clonal subpopulations in individual patients (Fig 5D). Further meta-
16 analyses on Protein Atlas data support these findings, by identifying YY1 clonal
17 populations and SLC9A3R1 sub-clonal populations in most ER α BC samples
18 (Supplementary Figure 13A-C). Overall, these data strongly support the notion that
19 enhancer activity can be used to qualitatively deconvolute heterogeneous
20 populations into phenotypical subclones.

21

22 **Phenotypic evolution during BC progression is shaped by endocrine treatment**

23 Tumor evolution studies have primarily focused on treatment naïve patients, taking
24 advantage of multi-regional sampling to retrospectively monitor changes in
25 clonality^{14,62}. Clonal tracking is dependent in part on passenger mutations, and the
26 effect of therapy has been rarely accounted for^{13,63}. More importantly, clonality has
27 been traced uniquely using genetic variants, with the intrinsic limitation of correlating
28 genetic changes to phenotypic ones. For example, sub-clones defined by passenger
29 mutations might be phenotypically equivalent, while a recent study using barcoded
30 glioblastoma cells shows that phenotypic clones might evolve independently from
31 mutational signatures². In addition, the few studies that looked at driver mutations in
32 coding regions of primary and metastatic BC disease found relatively similar

1 mutational landscapes¹⁵, suggesting that mapping phenotypic clones through BC
2 progression might reveal new targets. Our ability to acquire qualitative estimates of
3 phenotypic clones using enhancer ranking provides for a potential approach for
4 tracking changes in tumor heterogeneity with the additional advantage of predicting
5 for potentially functional changes. We interrogated our patient's dataset focusing on
6 events occurring between treatment-naïve primaries and treatment-resistant
7 metastatic BC (Fig. 6A). We hypothesized that phenotypic clonal evolution might be
8 driven by a coordinated activation/selection of groups of enhancers during BC
9 progression and this could be influenced by treatment. Our previous results suggest
10 that YY1+ cells remain clonal during progression (Fig 3A). Conversely, we show that
11 SLC9A3R1 expression is not antagonized by endocrine treatment suggesting that
12 SLC9A3R1-positive clones could expand during progression. We then calculated
13 changes in RI (Δ RI) for all enhancers captured in at least three patients ($SI > 3$,
14 $n=88935$) between primary and metastatic samples (Fig. 6B). SLC9A3R1 ranks
15 amongst the enhancers with the strongest increase in predicted clonality going from
16 primary to metastatic samples (Fig. 6B-C, 3.86σ from median Δ RI). Conversely, YY1
17 enhancer activity remains relatively unchanged (Fig. 6B-C). These data support our
18 initial hypothesis and suggest that SLC9A3R1-positive clones might expand in
19 response to treatment. To substantiate these data, we mapped the size of YY1 and
20 SLC9A3R1 phenotypic clones using IHC in an independent series of 20 matched
21 longitudinal biopsies. All surgical biopsies were obtained from treatment naïve
22 patients, while all the metastatic biopsies were taken at first relapse after endocrine
23 treatment¹³. We found YY1+ cells clonal both in primary and metastatic biopsies
24 (Fig. 6D). Conversely, SLC9A3R1+ subclones significantly expand during metastatic
25 progression to become completely clonal (100% staining) in 13/20 patients.
26 Interestingly, the only metastatic case in which we have observed a contraction of
27 the SLC9A3R1+ clone also showed a concomitant loss of ER α and PR positivity,
28 confirming our *in vitro* analysis and demonstrating that SLC9A3R1 remains an ER α
29 dependent-target despite being ET insensitive *in vivo* (Fig. 6D, red line and
30 Supplementary Figure 9B-E). Overall, these data demonstrate that changes in
31 enhancer ranking can estimate functional evolution during breast cancer
32 progression.

1 To gain more insight on functional evolution, we systematically annotated all
2 regulatory regions based on bias in detection between primary and metastatic
3 patients (Fig 6E). As expected, the bulk of enhancers and promoters do not show
4 bias toward primary and metastatic BC patients (common enhancers, CE). However,
5 we could successfully identify two distinct sets of regulatory regions that are
6 preferentially associated with primary (primary enhancers, PE) or metastatic
7 (metastatic enhancers, ME) patients. Remarkably, while CE do not show stage-
8 specific changes in RI, PE underlie larger sub-clonal populations in primary cancers
9 (statistically higher RIs in primary compared to metastatic, Fig. 6E). Likewise, ME
10 have lower RI in metastatic samples suggesting that the number of cells carrying
11 these enhancers have increased during progression (Fig. 6E). We next explored the
12 potential causes and functional consequences driving these coordinated epigenetic
13 changes. We thus identified the potential transcriptional targets of our enhancers
14 taking in account CTCF boundaries³⁷. Strikingly, we find that PE-associated gene-
15 transcription is associated with significantly better outcome while ME-associated
16 gene-transcription in primary samples is associated with poor prognosis (Fig 6D).
17 These data imply that primary samples containing larger subpopulations of
18 phenotypic clones with metastatic features relapse earlier.

19 We then mined PE and ME regulatory regions to identify the associated
20 biological features⁵⁶. PE appear to promote abnormal proliferation and
21 vascularization, two key events in early tumorigenesis. Remarkably, metastatic
22 samples switch to functional clones characterized by genes associated with BC
23 progression (FOXA1⁴³) or endocrine therapy resistance^{64,65} (Fig. 6E). Altogether,
24 these data suggest that endocrine therapies play a central role in shaping phenotypic
25 clonal evolution. Additional in-depth studies are needed to dissect the temporal
26 events triggered during phenotypic clonal evolution. Phenotypic subclones could
27 evolve by early coordinated activation and decommissioning of epigenetically
28 defined regulatory regions (*acquired*), selection of the fittest pre-existent epigenomic
29 landscape (*de novo*) or a combination of both.

30

31 Discussion

1 Our work describes the first systematic epigenetic profiling of primary and metastatic
 2 luminal breast cancer and reveals several critical principles underlying phenotypic-
 3 functional heterogeneity and its role in breast cancer progression. By mapping
 4 H3K27ac in untreated and treated patient samples we have also identified YY1 and
 5 SLC9A3R1, two new key players contributing to BC. While genomic profiling of
 6 breast cancer patients has revealed extensive clonal heterogeneity and retrospective
 7 tumour evolution^{28,66}, the vast majority of the mutational burden can be considered
 8 composed of passenger mutations^{29,67} making difficult to extrapolate actual
 9 phenotypes. Most RNA-based analysis, which may better reflect the phenotypic state
 10 of cancer cells, is generally obtained from bulk tissue and cannot inform on the
 11 existence of distinct subpopulations. Finally, molecular pathology can inform on the
 12 relative amount of protein abundance at the single-cell level but is laborious and not
 13 suitable for testing multiple targets simultaneously. In this work, we used epigenomic
 14 analyses to extrapolate phenotypic heterogeneity in solid tumour samples. Our
 15 analysis reveals that histone-based ChIP-seq signals, similarly to ATAC-seq³²,
 16 generally correlates with the number of cells in a population carrying the specific
 17 epigenetic information. Our predictions using YY1 and SLC9A3R1 enhancer fit
 18 extremely well with experimental data derived from normal tissues or BC patients.
 19 The findings that clonal regulatory regions dominating the landscape of individual
 20 tumor samples are shared across many patients, parallel recent genomic evidences
 21 showing that truncal (high allele frequency) mutations are also the most common
 22 mutations within cancer cohorts.

23 The results described here have several practical implications for BC. First, by
 24 comparing samples from drug-resistant metastatic patients with drug-naïve primary
 25 samples, we uncovered a set of enhancers marking phenotypic clones that
 26 significantly expand during breast cancer progression. Notably, these enhancers are
 27 strongly associated with genes specifically transcribed in cells that acquire endocrine
 28 therapy resistance (Fig. 6H). Conversely, enhancers progressively lost during tumour
 29 progression are linked to processes that often occur early in tumorigenesis. A set of
 30 enhancers expanding in metastatic samples point at progressive activation of
 31 FOXA1 and its network. It was recently reported that FOXA1 levels are increased in
 32 metastatic samples^{43,68}. Our data then predict that, similarly to SLC9A3R1, FOXA1
 33 positivity increases as a consequence of the expansion of a phenotypic clone

1 marked by an active FOXA1 enhancer and not via increased transcription of the
2 FOXA1 gene within single cells. It is tempting to speculate that this paradigm might
3 be valid for other genes. If correct it might signify that during cancer evolution, the
4 proportion of cells activating transcription is more important than the absolute
5 changes in transcription at the single cell levels. Interestingly, a set of enhancers
6 deactivated during progression involve IL-2 signalling (Fig. 6H). Reduction in IL-2
7 signalling was identified as a potential marker of relapse⁶⁹. Whether the IL-2 signal
8 source is the BC cells⁷⁰ themselves or it is due to a small contamination of immune
9 cells, needs to be defined. Equally, it will be important to measure real-time
10 activation/selection of enhancers in appropriate systems to ultimately establish if
11 phenotypic cancer evolution can be driven by Lamarckian events.

12 Finally, our analysis has identified two novel drivers of luminal BC. Firstly, we
13 identified YY1 as a key TF associated with clonal enhancers and promoters in BC
14 patients. Our data strongly support the idea that YY1 acts as a global co-activator in
15 cancer cells associating with the entire active epigenetic landscape. Several lines of
16 evidence indicate that YY1 might interact directly with modified nucleosomes,
17 possibly through its partner INO80⁵⁷. YY1 widespread association with clonal
18 enhancer suggests it might play a role in epigenetic memory. Intriguingly, a positive
19 screen for factors that improve induced pluripotent cells formation (iPS), identified
20 YY1 as the top hit, further supporting its potential role as enhancer gatekeeper⁷¹.
21 More specifically to ER α BC, we hypothesize that YY1 plays a critical role to stabilize
22 ER α binding at the transcriptionally productive core- ER α enhancers. Single-
23 molecule imaging shows that estrogen activated ER α increases its residency time on
24 the chromatin⁴² and recent evidence has shown that eRNA can trap YY1 on the
25 chromatin⁴⁹. More importantly, enhancer co-occupancy for YY1 and ER α occurs
26 almost exclusively at highly shared-highly functional core-ER α bound loci.
27 Altogether, these data raise the intriguing hypothesis that YY1 might contribute to
28 increased ER α residency at clonal enhancers. This could explain why some ER α are
29 captured in most patients, as longer residency time would increase chances of being
30 captured by ChIP-Seq⁴³. Longer residency might also explain the increased
31 transcriptional activity (Fig. 4D) and increased TF footprints (Fig. 2C) of these
32 enhancers.

1 YY1-ER α jointly control SLC9A3R1 enhancer activity, an event that cannot be
2 antagonized by conventional first-line endocrine therapies (Tamoxifen or AI) and that
3 drives SLC9A3R1 clonal expansion during breast cancer progression. Of note,
4 primary patients with high SLC9A3R1 expression might be viewed as containing
5 larger population of SLC9A3R1+ cells, thus resembling drug-resistant BC.
6 Intriguingly, SLC9A3R1 is amongst the strongest single prognostic genes for
7 relapse-free survival when considering endocrine treated patients (Fig. 5A). An
8 attractive possibility is that YY1 stabilizes ER α sufficiently at the SLC9A3R1
9 enhancer maintaining epigenetic memory in the presence of external antagonists.
10 Future studies are required to investigate the exact mechanisms through which
11 SLC9A3R1 contribute to BC and efficient strategies to antagonize its transcription,
12 possibly using CDK4/6 inhibitors⁷² to destabilize YY1. We recently demonstrated that
13 individual endocrine therapies can drive parallel genetic evolution *in vivo*¹³ and
14 epigenetic reprogramming *in vitro*¹⁶. Our data now strongly support the notion that
15 therapeutic interventions also play an essential role driving specific epigenetic
16 evolution during BC progression in the clinic. Metastatic re-biopsy at the time of
17 relapse, which is becoming commonplace in clinical practice should then examine
18 epigenetic changes in addition to newly acquired genomic ones, especially when no
19 new genetic drivers to guide further treatment are apparent¹⁵

20

1 **On Line Materials and Methods**

2

3 **Tumour tissue processing**

4 Breast cancer sample for ChIP-seq were collected by Imperial Tissue Bank (project
5 ethic approval R15021) and from Breast Cancer Now Tissue Bank (BCNTB-
6 TR000053-MTA & TR000040). Breast cancer fresh frozen tissue samples each
7 undergo aseptic macroscopic adipose tissue dissection. The dissected tumour tissue
8 is sectioned into 2mm x 2mm fragments in a petridish placed over dry ice. Tumour
9 fragments are then fixed using 1% formaldehyde solution for 10 minutes. Cold
10 glycine (1M) is added to the formaldehyde-fixed tissue for 10 minutes. The tumour
11 fragments are then pulverised using pestle and mortar and homogenised using liquid
12 nitrogen.

13

14 **Chromatin immunoprecipitation (ChIP)**

15 The ChIP protocol was conducted as described by Schmidt et al.⁷³ with few
16 modifications. In summary, following fixation, the tumour tissue undergoes chromatin
17 extraction and sonication using the Bioruptor Pico sonication device (Diagenode;
18 B01060001) using 20 cycles (30s on and 30s off) at maximum intensity. Purified
19 chromatin was then separated for 1. Immunoprecipitation using 4ug of H3k27ac
20 antibodies (Abcam; ab4729) per ChIP experiment or using 4ug of YY1 antibodies
21 (Santa Cruz; sc-281 X). ChIP-seq experiment for YY1 were performed in biological
22 duplicates. 2. Non-immunoprecipitated chromatin, used as Input control and 3.
23 Assessment of sonication efficiencies using a 1% agarose gel. Before construction of
24 ChIP-seq libraries (NEB Ultra II kit, see supplementary methods), enrichment of the
25 immunoprecipitated sample was ascertained using positive and negative controls for
26 ChIP-qPCR. Library preparation was performed using 10 – 50 ng of
27 immunoprecipitated and Input samples.

28

29 **ChIP-qPCR**

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

Ranking and Sharing Index

See Supplementary Computational Methods.

VSE

See Supplementary Computational Methods.

DHS imputations and TF motif analyses

See Supplementary Computational Methods.

Imputed DHS with vivo ER α binding Overlap

ER α binding from in breast cancer patients were obtained from ⁴³. ER α sharing index was calculated as before (see Supplementary Computational Methods). Overlap with imputed DHS was calculated using BedTools calculating the overlap (at least one base pair) via Cistrome Pipeline Analysis Suite (http://cistrome.org/Cistrome/Cistrome_Project.html). The percentage of overlap were calculated using binned DHS as variable first dataset and all the concatenated in vivo ER α as second dataset.

Footprint analysis

See Supplementary Computational Methods.

1 **Encode and Epigenomic Roadmap Ranking**

2 See Supplementary Computational Methods.

3 **Immunocytochemistry**

4 Hematoxylin and eosin staining of clinical samples was performed to calculate tumor
 5 burden prior to ChIP-seq. Briefly, 4- μ m-thick sections were obtained from formalin-
 6 fixed and paraffin-embedded specimens. After de-waxing in xylene and graded
 7 ethanol, sections were incubated in 3% H₂O₂ solution for 25 minutes to block
 8 endogenous peroxidase activities and then subjected to microwaving in EDTA buffer
 9 for antigen retrieval. For YY1 (Protein Atlas HPA001119, Atlas Antibodies
 10 Cat#HPA001119, RRID:AB 1858930) the following conditions were used: tissue
 11 sections were incubated with the primary monoclonal overnight at 4°C, and
 12 chromogen development was performed using the Envision system (DAKO
 13 Corporation, Glostrup, Denmark). A minimum of 500 tumor cells were scored with
 14 the percentage of tumor cell nuclei in each category recorded. For SLC9A3R1
 15 (HPA9672 and HPA27247, Atlas Antibodies Cat#HPA009672, RRID:AB 1857215
 16 and Atlas Antibodies Cat#HPA027247, RRID:AB 10601162 respectively) the
 17 following conditions were used. HPA9672 was diluted 1:400 and HPA27247 was
 18 diluted 1:1500. Staining was automatized with a Ventana Benchmark Ultra using
 19 epitope retrieval ER2 for 20 minutes. ER and PgR immunoreactivity was assessed
 20 by the FDA-approved ER/PR PharmDX kit (Dako). The prevalence of ER/PgR
 21 positive invasive cancer cells, independent of their staining intensity, was
 22 quantitatively annotated in the original reports. In accordance with ASCO/CAP
 23 guidelines, tumours with $\geq 1\%$ of immunoreactivity were considered positive

24

25 **Cell culture**

26 MCF7 was cultured using Dulbecco's modified Eagle's medium (DMEM) containing
 27 10% fetal calf serum (FCS) and 100 U penicillin/0.1 mg ml⁻¹ streptomycin, 2mM L-
 28 glutamine plus 10⁻⁸ 17- β -estradiol (SIGMA E8875). MCF7 long term oestrogen
 29 deprived (MCF7-LTED) cells were grown in phenol-free DMEM with 10% charcoal-
 30 stripped FCS (DCFCS) and 100 U penicillin/0.1 mg ml⁻¹ streptomycin and 2mM L-
 31 glutamine. T47D and T47D-LTED cells were passaged using DMEM containing 10%

1 FCS and 100 U penicillin/0.1 mg ml⁻¹ streptomycin, 2mM L-glutamine and phenol-
2 free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml⁻¹ streptomycin and
3 2mM L-glutamine, respectively. ZR75-1 cells were grown in DMEM containing 10%
4 FCS and 100 U penicillin/0.1 mg ml⁻¹ streptomycin, 2mM L-glutamine.

5

6 **siRNA**

7 Small interfering RNA (siRNA) against SLC9A3R1 (Gene ID; 9368: Ambion; s17919,
8 s17920), YY1 (Gene ID; 7528: Ambion; s14958, s14959, s14960) and *Silencer*
9 negative control (Ambion; AM4611). 1.5 x 10⁵ cells were seeded per well using a 6-
10 well plate. MCF7 cells were seeded in phenol-free DMEM with 10% DCFCS and 100
11 U penicillin/0.1 mg ml⁻¹ streptomycin and 2mM L-glutamine. Following 24 hours, the
12 cells were then transfected with siRNA using Lipofectamine 3000 (Invitrogen;
13 L3000015). T47D and ZR75-1 cells were seeded in DMEM containing 10% FCS and
14 100 U penicillin/0.1 mg ml⁻¹ streptomycin, 2mM L-glutamine. Following 24 hours, the
15 cells were then transfected with siRNA using Lipofectamine 3000 (Invitrogen;
16 L3000015). Cells were harvested for analysis following at least 48 hours of
17 transfection.

18

19 **Cell lysis and Western blot**

20 Cells were washed twice in ice-cold PBS and lysed in RIPA (Sigma-Aldrich; R02780)
21 buffer supplemented with protease (Roche 11697498001) and phosphatase
22 (Sigma-Aldrich 93482) inhibitors for 30 minutes with intermittent vortexing. Samples
23 were centrifuged at 4°C at maximum speed for 30 minutes after which, the
24 supernatant is transferred to a clean Eppendorf. Protein concentrations for each
25 sample was ascertained using the bicinchoninic acid (BCA) assay (ThermoFisher
26 Scientific; 23227). Equal amounts of lysates were loaded into BOLT 4-12% Bis-Tris
27 Plus Gel (Invitrogen; NW04120BOX). Proteins were transferred to a Biotrace
28 nitrocellulose membrane (VWR; PN66485) and incubated with primary antibodies
29 overnight. Proteins were then visualised using goat anti-mouse (ThermoFisher
30 Scientific; 31446) and anti-rabbit (ThermoFisher Scientific; 31462) HRP conjugated
31 secondary antibodies. Amersham ECL start Western Blotting Detection reagent (GE

1 Healthcare Life Sciences; RPN3243) was used for chemiluminescent imaging using
2 the Fusion solo (Vilber; Germany) imager.

3

4 **Transcriptional profiling**

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

9

10 **RNA extraction and real-time PCR**

11 Total RNA was extracted using RNeasy Mini Kit (Qiagen; 74106), and the cDNA was
12 reverse transcribed from 1ug of RNA using iScript cDNA synthesis kit (Bio-Rad;
13 #1708891). Real time-qPCR (RT-qPCR) reactions were carried out in 10 uL volume
14 containing 5 uL of sybergreen mix (ABI; 4472918), 0.5 ul of primer (2.5 uM final
15 concentration), 2.5 ul of genomic DNA and 2 ul of DNASE/RNASE-free water. A
16 three-step cycle programme and a melting analysis were applied. The cycling steps
17 were as follows: 10s at 95 °C, 30s at 60 °C and 30s at 72 °C, repeated 40 times¹⁹.

18

19 **Luciferase reporter assay**

20 MCF7 cells were seeded in a 24-well plates at 5×10^4 cells per well in phenol-free
21 DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml⁻¹ streptomycin and 2mM L-
22 glutamine. After 24 hours of incubation, transfection of plasmid DNA was performed
23 using Lipofectamine 3000 (Invitrogen; L3000015). Cells were transfected with 100ng
24 of ERE_Luciferase reporter, 10ng of the renilla luciferase control plasmid (pRL-
25 CMV), 10ng of pSG5_ER- α , 15 nm of siRNA and 280ng of Bluescribe DNA (BSM)
26 per well; totalling 400ng of DNA/well. After 12 hours of transfection the media was
27 replaced with fresh phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg
28 ml⁻¹ streptomycin and 2mM L-glutamine. Treatment with 10^{-8} 17- β -estradiol (SIGMA
29 E8875) or control treatment was administered and the cells incubated for 24 hours.
30 Cell lysates are then obtained using Passive lysis 5X buffer (Promega; E1941). The

firefly and renilla luciferase activity was determined using DualGlo luciferase assay kit (Promega; E2920) according to the manufacturer protocol. The renilla luciferase activity measurement was utilised as control for transfection efficiency and therefore the ERE_Luciferase activity was normalised to the reading obtained for the renilla luciferase activity.

SRB assay

Briefly, the sulphorhodamine 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 after which, the first plate (Day 0) is assayed after the cells have become adherent. Prospective plates are assayed sequentially after 3 days, 5 days and 7 days. The cells are fixed by adding 200uL of cold 40% (weight/volume) of trichloroacetic acid (TCA) to each well for at least 60 minutes. The plates were washed five times with distilled water and then 100 uL/well of SRB (0.4% wt/vol SRB in 1% wt/vol acetic acid) reagent is added to each well and the plates are allowed to incubate for 30 minutes. The plates were then washed five times in 1% (wt/vol) acetic acid and allowed to dry overnight. SRB solubilisation was performed by adding 100 uL/well of 10 mM Tris HCl to the plates and allowed to shake for 30 minutes. Optical density was then measured using the Sunrise microplate reader (Tecan; Sunrise) at 492 nm. Cell proliferation is then calculated over the 7-day period using Day 0 as a baseline measurement.

Enrichment scores.

Overlap for ER α (*in vivo*) vs enhancers and promoters were calculated by betoold intersect were the percentage overlap is calculated over the total number of regulatory regions within each bin against the concatenate ER α binding set (all ER α in all patients). For YY1, FOXA1 and ER α in MCF7, intersections were calculated using Cistrome. YY1 BEDFILES were the consensus narrow peaks of two biological experiment, FOXA1 ChIP-seq data and ER α were obtained in house¹⁶. The core ER α was BEDFILE was obtained by converting the published dataset from ⁴³ to

1 HG19. The private ER α BEDFILE was obtained by iterative process to identify ER α
 2 binding unique to single patients prior to concatenation into a single file. Overlap
 3 represent the fraction of the original datasets (first dataset) overlapping with core
 4 ER α (second dataset). The TCGA luminal signature was obtained from ³⁸. Each
 5 gene was extended for 20Kb upstream keeping in consideration the direction of
 6 transcription. A null gene list was generated by subtracting the TCGA luminal
 7 signature from a genome-wide gene list. Genes from the null list were extended in a
 8 similar way and enrichment was calculated by comparing the fraction of TCGA gene
 9 list with nearby binding vs. the null list. A list of estrogen target genes that do not
 10 respond to Tamoxifen was obtained from ⁵⁸. Each gene was extended for 20Kb
 11 upstream keeping in consideration the direction of transcription. A null gene list was
 12 generated by subtracting the signature from a genome-wide gene list. Genes from
 13 the null list were extended in a similar way and enrichment was calculated by
 14 comparing the fraction of TAM resistant estrogen dependent gene list with nearby
 15 binding vs. the null list.

16

17 **RI-IHC correlation**

18 FFPE sections for the patients used in the ChIP-seq section were retrieved from
 19 Imperial Tissue bank. Sections were stained with YY1 or SLC9A3R1 antibodies.
 20 Stained sections were divided in 20 sectors. 5 sectors with high tumor burden were
 21 scored for the number of IHC+ cells and results averaged. The number of IHC+ cells
 22 and the matched RI was analyzed using linear regression using Prism 5 (GraphPad
 23 software Inc.).

24

25 **Δ RI**

26 See Supplementary Computational Methods.

27

28 **YY1 and SLC9A3R1 Pan cancer expression analysis**

29 YY1 and SLC9A3R1 expression profile for matched Normal vs. Cancer samples was
 30 obtained using TIMER diff.exp option (<https://cistrome.shinyapps.io/timer/>). YY1

transcriptional analyses of breast cancer subtypes was performed in the Metabric Dataset (Curtis Breast) using probe ILMN_1770892 or TCGA dataset using Oncomine (<https://www.oncomine.org/resource/login.html>).

SLC9A3R1 Meta-analyses

SLC9A3R1 expression profile in drug resistant cell lines was performed by analysis of RNA-seq data from ¹⁶. SLC9A3R1 expression profile in MCF7 cells transfected with siRNA against ER α was performed by analysis of microarray data from GSE27473. 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 Inc.). Kaplan-Meier analysis using SLC9A3R1 expression were performed by re-analysis of 23 independent microarray datasets (KMPlot), TCGA RNA-seq data or the combined Metabric Dataset. SLC9A3R1 transcriptional profile in breast cancer cell lines was obtained from the HPA RNA-seq dataset (<http://www.proteinatlas.org/about/download>). SLC9A3R1 transcriptional profile from tissues was obtained from the HPA, GTEx and FANTOM5 RNA-seq datasets (<http://www.proteinatlas.org/about/download>).

Author Contribution

L.M. conceived the study. D.P., E.E performed the experiments. L.M., G.C., B.G., A.S., L.S., and P.S., developed and performed bioinformatic analyses. K.G., organised tissue collection. D.H., G.S., P.B., C.P., 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.

Acknowledgments

We want to acknowledge and thanks all patients and their families for the support and for donating the research samples. We thank Breast Cancer Now Tissue Bank (project TR0121), Imperial Tissue Bank and the LEGACY study for contributing

tissues. The authors gratefully 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 (P64250) and 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). We acknowledge Lorna Watson, Iros Barozzi, Ylenia Perone and Jason Carrol for their constructive comments on the manuscript.

Data Accession

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

References

1. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
2. Lan, X. *et al.* Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **60**, 5954 (2017).
3. Consortium, T. E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**, 57–74 (2012).
4. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
5. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–86 (2015).
6. Carroll, J. S. Steroids, nuclear receptors and breast cancer. Preface. *Mol. Cell. Endocrinol.* **382**, 623 (2014).
7. Ali, S., Buluwela, L. & Coombes, R. C. Antiestrogens and Their Therapeutic Applications in Breast Cancer and Other Diseases. *Annu. Rev. Med.* **62**, 217–232 (2011).
8. Perou, C. M. *et al.* Molecular portraits of human breast tumours : Article : *Nature*. *Nature* **406**, 747–752 (2000).
9. 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).
10. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
11. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
12. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Aromatase

- 1 inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis
2 of the randomised trials. *Lancet* (2015). doi:10.1016/S0140-6736(15)61074-1
- 3 13. Magnani, L. *et al.* Acquired CYP19A1 amplification is an early specific
4 mechanism of aromatase inhibitor resistance in ER α metastatic breast cancer.
5 *Nature Genetics* **49**, 444–450 (2017).
- 6 14. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor
7 heterogeneity in cancer evolution. *Cancer Cell* **27**, 15–26 (2015).
- 8 15. Yates, L. R. *et al.* Genomic Evolution of Breast Cancer Metastasis and
9 Relapse. *Cancer Cell* **32**, 169–184.e7 (2017).
- 10 16. Nguyen, V. T. M. *et al.* Differential epigenetic reprogramming in response to
11 specific endocrine therapies promotes cholesterol biosynthesis and cellular
12 invasion. *Nature Communications* **6**, 10044 (2015).
- 13 17. Magnani, L. *et al.* Genome-wide reprogramming of the chromatin landscape
14 underlies endocrine therapy resistance in breast cancer. *Proceedings of the*
15 *National Academy of Sciences* **110**, E1490–E1499 (2013).
- 16 18. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-
17 enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- 18 19. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of
19 transcriptional promoters and enhancers in the human genome. *Nature*
20 *Genetics* **39**, 311–318 (2007).
- 21 20. Falahi, F. *et al.* Towards Sustained Silencing of HER2/neu in Cancer By
22 Epigenetic Editing. *Molecular Cancer Research* **11**, 1029–1039 (2013).
- 23 21. Laprell, F., Finkl, K. & Müller, J. Propagation of Polycomb-repressed chromatin
24 requires sequence-specific recruitment to DNA. *Science* eaai8266 (2017).
25 doi:10.1126/science.aai8266
- 26 22. Wang, X. & Moazed, D. DNA sequence-dependent epigenetic inheritance of
27 gene silencing and histone H3K9 methylation. *Science* eaaj2114 (2017).
28 doi:10.1126/science.aaj2114
- 29 23. Coleman, R. T. & Struhl, G. Causal role for inheritance of H3K27me3 in
30 maintaining the OFF state of a Drosophila HOX gene. *Science* eaai8236
31 (2017). doi:10.1126/science.aai8236
- 32 24. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing
33 transcriptional regulators within the chromatin environment. *Trends Genet.* **27**,
34 465–474 (2011).
- 35 25. Jozwik, K. M. & Carroll, J. S. Pioneer factors in hormone-dependent cancers.
36 1–5 (2012). doi:10.1038/nrc3263
- 37 26. Hnisz, D. *et al.* Convergence of developmental and oncogenic signaling
38 pathways at transcriptional super-enhancers. *Molecular Cell* **58**, 362–370
39 (2015).
- 40 27. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect
41 global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
- 42 28. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed
43 by multiregion sequencing. *Nature Medicine* **21**, 751–759 (2015).
- 44 29. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A.
45 Identification of neutral tumor evolution across cancer types. *Nat Genet* **48**,
46 238–244 (2016).
- 47 30. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in
48 human oligodendroglioma. *Nature* (2016). doi:10.1038/nature20123
- 49 31. Harvey, J. M., Clark, G. M., Osborne, C. K. & Allred, D. C. Estrogen receptor
50 status by immunohistochemistry is superior to the ligand-binding assay for

- 1 predicting response to adjuvant endocrine therapy in breast cancer. *Journal of*
2 *Clinical Oncology* **17**, 1474–1481 (1999).
- 3 32. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of
4 regulatory variation. *Nature* **523**, 486–490 (2015).
- 5 33. Cowper-Salari, R. *et al.* Breast cancer risk-associated SNPs modulate the
6 affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics* **44**,
7 1191–1198 (2012).
- 8 34. Cohen, A. J. *et al.* Hotspots of aberrant enhancer activity punctuate the
9 colorectal cancer epigenome. *Nature Communications* **8**, 14400 (2017).
- 10 35. Levisky, J. M. & Singer, R. H. Gene expression and the myth of the average
11 cell. *Trends in Cell Biology* **13**, 4–6 (2003).
- 12 36. Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time
13 observation of transcription initiation and elongation on an endogenous yeast
14 gene. *Science* **332**, 475–478 (2011).
- 15 37. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq
16 data with BETA. *Nature Protocols* **8**, 2502–2515 (2013).
- 17 38. Cancer Genome Atlas Network. Comprehensive molecular portraits of human
18 breast tumours. *Nature* **490**, 61–70 (2012).
- 19 39. Györfy, B. *et al.* An online survival analysis tool to rapidly assess the effect of
20 22,277 genes on breast cancer prognosis using microarray data of 1,809
21 patients. *Breast Cancer Res Treat* **123**, 725–731 (2009).
- 22 40. Neph, S. *et al.* An expansive human regulatory lexicon encoded in
23 transcription factor footprints. *Nature* **489**, 83–90 (2012).
- 24 41. Thurman, R. E. *et al.* The accessible chromatin landscape of the human
25 genome. *Nature* **489**, 75–82 (2012).
- 26 42. Paikinaho, V. *et al.* Single-molecule analysis of steroid receptor and cofactor
27 action in living cells. *Nature Communications* **8**, 15896 (2017).
- 28 43. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated
29 with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
- 30 44. Kittler, R. *et al.* A comprehensive nuclear receptor network for breast cancer
31 cells. *Cell Reports* **3**, 538–551 (2013).
- 32 45. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding
33 reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122**,
34 33–43 (2005).
- 35 46. Beagan, J. A. *et al.* YY1 and CTCF orchestrate a 3D chromatin looping switch
36 during early neural lineage commitment. *Genome Research* **27**, 1139–1152
37 (2017).
- 38 47. Vella, P., Barozzi, I., Cuomo, A., Bonaldi, T. & Pasini, D. Yin Yang 1 extends
39 the Myc-related transcription factors network in embryonic stem cells. *Nucleic*
40 *Acids Res.* **40**, 3403–3418 (2012).
- 41 48. Jeon, Y. & Lee, J. T. YY1 tethers Xist RNA to the inactive X nucleation center.
42 *Cell* **146**, 119–133 (2011).
- 43 49. Sigova, A. A. *et al.* Transcription factor trapping by RNA in gene regulatory
44 elements. *Science* **350**, 978–981 (2015).
- 45 50. Powe, D. G. *et al.* Investigating AP-2 and YY1 protein expression as a cause
46 of high HER2 gene transcription in breast cancers with discordant HER2 gene
47 amplification. *Breast Cancer Res.* **11**, R90 (2009).
- 48 51. Lee, M. H. *et al.* Yin Yang 1 positively regulates BRCA1 and inhibits mammary
49 cancer formation. *Oncogene* **31**, 116–127 (2011).
- 50 52. Klymenko, T. *et al.* A Polycomb group protein complex with sequence-specific

- 1 DNA-binding and selective methyl-lysine-binding activities. *Genes &*
2 *Development* **20**, 1110–1122 (2006).
- 3 53. Schwalie, P. C. *et al.* Co-binding by YY1 identifies the transcriptionally active,
4 highly conserved set of CTCF-bound regions in primate genomes. *Genome*
5 *Biology* **14**, R148 (2013).
- 6 54. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals
7 Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
- 8 55. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines
9 their genomic and transcriptomic landscapes. *Nature Communications* **7**,
10 11479 (2016).
- 11 56. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory
12 regions. *Nature Biotechnology* **28**, 495–501 (2010).
- 13 57. Cai, Y. *et al.* YY1 functions with INO80 to activate transcription. *Nature*
14 *Structural & Molecular Biology* **14**, 872–874 (2007).
- 15 58. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S.
16 FOXA1 is a key determinant of estrogen receptor function and endocrine
17 response. *Nat Genet* **43**, 27–33 (2010).
- 18 59. Vaquero, J., Nguyen Ho-Bouloires, T. H., Clapéron, A. & Fouassier, L. Role
19 of the PDZ-scaffold protein NHERF1/EBP50 in cancer biology: from signaling
20 regulation to clinical relevance. *Oncogene* (2017). doi:10.1038/onc.2016.462
- 21 60. Murtazina, R. *et al.* Tissue-specific regulation of sodium/proton exchanger
22 isoform 3 activity in Na(+)/H(+) exchanger regulatory factor 1 (NHERF1) null
23 mice. cAMP inhibition is differentially dependent on NHERF1 and exchange
24 protein directly activated by cAMP in ileum versus proximal tubule. *Journal of*
25 *Biological Chemistry* **282**, 25141–25151 (2007).
- 26 61. Miller, W. R. *et al.* Gene Expression Profiles Differentiating Between Breast
27 Cancers Clinically Responsive or Resistant to Letrozole. *Journal of Clinical*
28 *Oncology* **27**, 1382–1387 (2009).
- 29 62. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell
30 carcinomas defined by multiregion sequencing. *Nature Genetics* **46**, 225–233
31 (2014).
- 32 63. Juric, D. *et al.* Convergent loss of PTEN leads to clinical resistance to a
33 PI(3)Ka inhibitor. *Nature* **518**, 240–244 (2015).
- 34 64. Creighton, C. J. *et al.* Development of Resistance to Targeted Therapies
35 Transforms the Clinically Associated Molecular Profile Subtype of Breast
36 Tumor Xenografts. *Cancer Res.* **68**, 7493–7501 (2008).
- 37 65. Massarweh, S. *et al.* Tamoxifen Resistance in Breast Tumors Is Driven by
38 Growth Factor Receptor Signaling with Repression of Classic Estrogen
39 Receptor Genomic Function. *Cancer Res.* **68**, 826–833 (2008).
- 40 66. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at
41 single nucleotide resolution. *Nature* **461**, 809–813 (2009).
- 42 67. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal
43 Passenger Mutations in Cancer Evolution. *PLoS Comput Biol* **12**, e1004731
44 (2016).
- 45 68. Mohammed, H. *et al.* Progesterone receptor modulates ER α action in breast
46 cancer. *Nature* **523**, 313–317 (2015).
- 47 69. Arduino, S. *et al.* Reduced IL-2 level concentration in patients with breast
48 cancer as a possible risk factor for relapse. *Eur. J. Gynaecol. Oncol.* **17**, 535–
49 537 (1996).
- 50 70. García-Tuñón, I. *et al.* Interleukin-2 and its receptor complex (α , β and γ

1 chains) in in situ and infiltrative human breast cancer: an immunohistochemical
2 comparative study. *Breast Cancer Res.* **6**, R1 (2003).
3 71. Onder, T. T. *et al.* Chromatin-modifying enzymes as modulators of
4 reprogramming. *Nature* **483**, 598–602 (2012).
5 72. Turner, N. C. *et al.* Palbociclib in Hormone-Receptor–Positive Advanced
6 Breast Cancer. *N Engl J Med* **373**, 209–219 (2015).
7 73. Schmidt, D. *et al.* ChIP-seq: Using high-throughput sequencing to discover
8 protein–DNA interactions. *Methods* **48**, 240–248 (2009).
9
10

1 **Figure Legends**

2 **Figure 1: Assessment of inter- and intra-tumor epigenetic heterogeneity** A)

3 Main hypothesis of the study. Transcriptional data from bulk tissue represent the
 4 average over million cells. Each cell contributes a value from a continuous
 5 distribution of potential mRNA molecules. For chromatin data, each cell can only
 6 contribute a deterministic value to the bulk signal, generally from two alleles.
 7 Therefore, the relative strength of ChIP-seq data is dependent on the number of cells
 8 carrying epigenetic signal at discrete loci. H, M and L represent strong, medium and
 9 weak signal, respectively. Clonal regulatory regions are commonly shared by BC
 10 patients while weak enhancers are more patient specific B) EGR3 mRNA is
 11 expressed in MCF7 but not MCF7-F cells. eRNA and Pol-II ChIA-PET show
 12 enhancer activity in MC7 but not MCF7-F¹⁶. CTCF insulated perimeter is shown in
 13 yellow. Predicted looping from ChIA-PTE is shown in red. The observed ChIP-qPCR
 14 signal for H3K27ac at EGR3 enhancers decrease with increasing number of MCF7-F
 15 cells mixed in the sample C) Ranking strategy: H3K27ac signal is normalized at each
 16 locus and assigned a ranking index based on relative strength within each single
 17 ChIP-seq experiment (1=strongest, 100=weakest, binning on RPKM signal). Binning
 18 is repeated for each patient. D) Linear regression shows that clonal enhancers are
 19 commonly shared between breast cancer patients. Y axis=Ranking Index, X
 20 axis=Sharing index. Sharing index indicate the number of patients sharing the
 21 regulatory region. Each dot represents the median RI (all patients) for a single
 22 regulatory region. The interpolating lines represent the median RI value and
 23 interquartile ranges for regulatory regions with the same SI E) Overlap between BC
 24 risk variants and annotated DNA elements F) Variant Set Enrichment analysis
 25 indicates that BC-specific but not CRC-specific GWAS risk variants occur more
 26 frequently than expected within the enhancers elements identified in our study.

27

28 **Figure 2: Clonal and sub-clonal regulatory regions contain distinct regulatory**

29 **information.** A) Bioinformatic framework of the analyses. H3K27ac calls were split
 30 to identify approximate nucleosome-level enrichment (sub-peaks). Sub-peaks data
 31 were integrated with ENCODE-derived DHS-seq calls to identify potential sites of TF
 32 binding. Individual imputed DHS regions were assigned SI values based on the

1 number of patient sharing the region B) Transcription factor motif analysis of
 2 individual bins (SI) followed by unsupervised clustering. RD and RN regions cluster
 3 separately in two distinct clades. ERE and YY1 motif are blown up at the bottom C)
 4 Clonal enhancers in MCF7 cells (RI<20) are characterized by a higher number of TF
 5 footprints, while sub-clonal enhancers (RI>70) have less footprint than expected
 6 (O/E=1). Asterisks represent a pValue of <0.001 in a Wilcoxon Signed Rank Test D)
 7 Overlap of imputed DHS regions with *in vivo* derived ER α binding sites. The left Y
 8 axis indicates cumulative DHS regions. The right Y axes indicate the percentage of
 9 overlap based on total DHS in each SI bin E) Distribution plot of *in vivo* derived ER α
 10 binding sites versus the number of patients in which they were observed⁴³.

11

12 **Figure 3: YY1 identify a dominant phenotypic clone in ER α BC** A) RIs for the
 13 YY1 enhancer within all the individual patients included in the current study. YY1
 14 enhancer location with its 3D interactions are shown in the top right inset B) YY1
 15 enhancer ranking analysis of available Epigenome Roadmap H3K27ac datasets.
 16 Tissues are displayed from the strongest to the weakest YY1 enhancer activity
 17 (based on RI). Representative IHC analysis of normal tissues stained with a YY1
 18 antibody are shown C) Top left: YY1 expression in ER α -positive breast cancer
 19 compared to normal breast tissue. Bottom left: Kaplan-Meier analysis of patient
 20 outcome using YY1 expression to stratify patients. Right: Kaplan-Meier analysis of
 21 patient outcome using YY1 expression. All BC subtypes were analysed separately
 22 D) IHC analysis of normal breast tissues highlights YY1 functional subclones in
 23 normal breast E) IHC analysis of ER α positive invasive ductal carcinomas identify
 24 YY1 positive clones as the dominant clonal population.

25

26 **Figure 4: YY1 marks critical enhancers in breast cancer cells** A) ChIP-seq data
 27 from ER α -positive MCF7 for YY1 in quiescent or 17 β -estradiol (E2) stimulated cells
 28 B) Heatmaps showing global enrichment profiles of several chromatin markers
 29 associated with active regulatory regions in MCF7 cells C) Overlap between ER α ,
 30 YY1 and FOXA1 in MCF7 cells. The right panel shows the potential overlap with *in*
 31 *vivo*- derived core ER α binding sites D) ER α core binding sites are strongly enriched

1 for YY1 binding in MCF7 cells while patient-specific ER α bindings are generally YY1-
2 free. E) Genes used to classify luminal breast cancer patients are strongly enriched
3 for ER α -YY1 binding sites. Asterisks represent $p < 10^{-5}$ in a Fisher's Exact test vs.
4 private ER α F) YY1 depletion leads to transcriptional shut-down of an ERE-driven
5 luciferase reporter. Bars and error bars represent the average of 5 independent
6 experiments with SE. Asterisks represent significance at $P < 0.001$ after ANOVA with
7 Dunnet's correction. G) Silencing YY1 blocks estrogen-induced growth in MCF7 cells
8 H) YY1 depletion leads to growth arrest in AI resistant LTED cells. Proliferation
9 assays were conducted in biological triplicate. Error bars indicate 95% confidence
10 intervals. Asterisks represent significance at $P < 0.05$, 0.01, 0.001 and 0.0001 after 2-
11 way ANOVA with Tukey's post-test I) Overlap of YY1 and ER α binding sites in LTED
12 cell lines J) ER α -YY1 bound enhancers in LTED cells underlie the transcription of
13 genes associated with luminal breast cancer and acquired endocrine therapy
14 resistance K) core ER α -YY1 bound enhancers are strongly enriched near estrogen
15 responsive genes that are not suppressed by Tamoxifen co-treatment.

16

17 **Figure 5: Epigenomic mapping predicts the size of phenotypic clones in**
18 **patients** A) Global Kaplan-Meier analysis summarize univariate analysis for each
19 gene included in the Affymetrix microarray platform. Hazard Ratios are plotted in the
20 X axis B) SLC9A3R1 RNA levels pre- and post- short-term aromatase inhibitor
21 treatment in responder and non-responder patients⁶¹. Oestrogen-dependent
22 expression of progesterone receptor mRNA is shown as comparison C) Silencing
23 SLC9A3R1 leads to proliferation arrest in response to estrogen stimulation in MCF7
24 and estrogen independent growth in LTED cells. Proliferation assays were
25 conducted in biological triplicate. Error bars indicate 95% confidence intervals.
26 Asterisks represent significance at $P < 0.05$, 0.01, 0.001 and 0.0001 after 2-way
27 ANOVA with Tukey's post-test D) RIs for the SLC9A3R1 enhancer within all the
28 individual patients included in the current study. SLC9A3R1 enhancer location and
29 its 3D interactions are shown in the top right inset E) SLC9A3R1 enhancer ranking
30 analysis of available Epigenome Roadmap H3K27ac datasets. Tissues are displayed
31 from the strongest to the weakest SLC9A3R1 enhancer activity (based on RI).
32 Representative IHC analysis of normal tissues stained with a SLC9A3R1 antibody

are shown. 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. Linear regression R square, confidence intervals and representative staining are also shown.

6

Figure 6: Endocrine treatment shapes phenotypic evolution. A) Theoretical framework of the analysis. The relative size of phenotypic clones can be tracked using enhancer activity (RIs). Phenotypic clones can be positively or negatively selected during BC progression in response to endocrine therapies. B) Expanding or contracting phenotypic clones were defined based on the RI-ratio in primary and metastatic samples (RI_P/RI_M). Distribution of RI-ratio identified top candidate enhancers YY1 RI does not change significantly during progression, while SLC9A3R1 RI ranks among the enhancers with stronger increase in activity during progression. Vertical bars represent σ (Standard Deviation) increments from the population median C) Scatterplot of YY1 and SLC9A3R1 enhancer ranking according to patient stage. Bars indicate mean and 95% confidence intervals. Asterisks represent significance at $P < 0.05$ after students two-tail T-Test D) IHC staining for YY1 and SLC9A3R1 positive cells in an independent matched longitudinal cohort of ER α breast cancer patients. All normal and primaries are treatment naïve. All metastatic have received endocrine therapies (Tamoxifen or Aromatase inhibitors). Statistical significance was calculated using a pair-wise, two-tail 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 differential was then derived and plotted on the X-axis F) RI indexes for all PE and ME are plotted. As a control, RI for common enhancer (CE) are also plotted. Permutation was used to assess changes in RI in 50 randomly selected sets of CE G) Kaplan-Meier analysis using averaged RNA expression of genes associated with PE or ME regulatory regions. Genes were assigned considering CTCF insulated perimeters E) Pathway analysis for genes associated with PE or ME regulatory regions. Pathways were identified using GREAT and are listed in order of significance (qValue).

1 **Supplementary Figures**

2

3 Supplementary Figure 1. Hematoxylin-Eosin staining to evaluate tumor cellularity
4 was carried out for each sample profiled using ChIP-seq. Only tumors with cellularity
5 above 70% were analyzed.

6

7 Supplementary Figure 2. Summary statistics for ChIP-seq analyses. A) The number
8 of individual peaks called using MACS 2.0 are shown for each patient. A q value of
9 0.01 was used in the peak calling analysis B) Saturation plots. Patients were
10 permuted and the total number of region called after permutation is shown on the Y
11 axis. 80% of total promoters were covered by permutating 4 patients, while similar
12 saturation for enhancers was reached after permutating all 47 samples C)
13 Distribution plots show the frequencies in function of Sharing Index for each
14 regulatory region. Inset show median SI for promoters and enhancers.

15

16 Supplementary Figure 3. A) Spiking experiments show that relative enrichment for
17 ER α binding correlates with the number of cells carrying the binding event. MCF7
18 stimulated or not with estradiol to induce ER α at the specified enhancers were mixed
19 in different proportion with ER α -negative cells with similar genetic background before
20 ER α binding was measured using ChIP-qPCR. Arrows indicate the binding site
21 quantified using ChIP-qPCR B) Linear regression using patient-derived RNA levels
22 and patient-derived SI. The analysis was repeated in three independent cohorts.
23 Normal samples were analysed when available (small insets). The number in each
24 box summarize relative slope for RN and RD elements, the colour of the box
25 indicates the correlation coefficient R².

26

27 Supplementary Figure 4. Transcription factor motif analyses of the entire promoter
28 and enhancer imputed DHS landscapes. Motifs are ranked based on the ratio of
29 observed/expected. Motif were filtered for a q Value of 10⁻⁴

30

31 Supplementary Figure 5. YY1 RNA levels from three independent RNA-seq datasets

of normal tissues. Images were obtained using Protein Atlas Tools B) YY1 RNA levels from RNA-seq analysis of cancer cell lines. Images were obtained using Protein Atlas Tools.

Supplementary Figure 6. A) Chromatin landscape at the YY1 enhancer locus in breast cancer cells. The loops were obtained from Pol II ChIA-PET data (high score). The CTCF insulation perimeter was established from CTCF ChIA-PET data (ENCODE). Enhancers SI are shown at the bottom B) Meta-analysis of IHC data from Protein Atlas stained with YY1 antibody. RI for the individual tissues in indicated. Percentage of YY1 positive cells is also listed at the bottom of each image.

Supplementary Figure 7. A) YY1 expression comparing normal tissues and cancer tissues shows that YY1 median expression is significantly stronger in TCGA cancer tissues. Data were generated using TIMER (<https://cistrome.shinyapps.io/timer/>) (B) YY1 median expression is significantly higher in several breast cancers sub-classes compared to normal tissue. Data were obtained from Oncomine C) IGV snapshot of estradiol induced YY1 ChIP-seq and H3K27ac ChIP-seq from MCF7 cells near transcriptionally inactive genes.

Supplementary Figure 8. A-B) Silencing YY1 is sufficient to abrogate the growth of two independent cell line models of ER α breast cancer. Proliferation assays were conducted in biological triplicate. Error bars indicate 95% confidence intervals. Asterisks represent significance at P<0.05, 0.01, 0.001 and 0.0001 after 2-way ANOVA with Tukey's post-test C) YY1 depletion leads to reduced transcription of common ER α target genes. Each experiment was performed in biological triplicates. Column and bars represent the average and SEM of all experiment. Asterisks represent significance at P<0.05, 0.01, 0.001 and 0.0001 after 1-way ANOVA with Dunnet's post-test.

Supplementary Figure 9. A) SLC9A3R1 RNA expression in BC cell lines sensitive or resistant to endocrine therapies. RNA-seq data were obtained in house¹⁶ B) Meta-analysis of SLC9A3R1 expression in response to ER α depletion C) SLC9A3R1 RNA in additional oestrogen independent BC cell line models. Fold changes are

1 calculated as ratio compared to parental endocrine sensitive BC cells. Oestrogen-
 2 dependent expression of progesterone receptor mRNA is shown as comparison.
 3 Original microarray codes are shown. Microarray were downloaded from GEO and
 4 re-analysed D) SLC9A3R1 transcriptional response to several stimuli is shown. Data
 5 were analysed using NURSA Transcriptomime tool
 6 (<https://www.nursa.org/nursa/transcriptomine/index.jsf;jsessionid=J3hRuy3XjX6HeNr2aekh-rrTsXS-uVUXErsip0wY.nursa3>) E) Kaplan Meier survival plots were
 8 calculated using three independent large datasets using SLC9A3R1 expression in
 9 the primary cancer as a classifier.

10

11 Supplementary Figure10. A) Cell lines from the Protein Atlas initiative were ranked
 12 based on SLC9A3R1 expression as profiled by RNA-seq B) SLC9A3R1 expression
 13 comparing normal tissues and cancer tissues shows that SLC9A3R1 median
 14 expression is significantly stronger in ER α -positive Breast cancer patients from
 15 TCGA. Data were generated using TIMER (<https://cistrome.shinyapps.io/timer/>) C)
 16 Chromatin landscape at the SLC9A3R1 enhancer locus in breast cancer cells.
 17 Looping analysis was conducted using Pol II ChIA-PET data. CTCF insulation
 18 perimeter was established from CTCF ChIA-PET data (ENCODE). H3K27ac, YY1,
 19 ER α , FOXA1 and DHS-seq data were developed in house and deposited online.
 20 ER α binding identified as Core are highlighted in the red box D) Expression analysis
 21 comparing median RNA expression values for several genes localized near the
 22 active SLC9A3R1 putative enhancer. SLC9A3R1 expression is significantly
 23 increased in BC tissues compared to normal while the expression of other genes is
 24 not affected

25

26 Supplementary Figure11 A) YY1 silencing in MCF7 is sufficient to decrease
 27 SLC9A3R1 expression. Each experiment was performed in biological triplicates.
 28 Column and bars represent the average and SEM of all experiment. Asterisks
 29 represent significance at P<0.05, 0.01, 0.001 and 0.0001 after 1-way ANOVA with
 30 Dunnet's post-test B) Silencing SLC9A3R1 is sufficient to abrogate the growth of an
 31 endocrine therapy resistant cell line but not the parental, treatment naïve breast
 32 cancer cell line. Proliferation assays were conducted in biological triplicate. Error
 33 bars indicate 95% confidence intervals. Asterisks represent significance at P<0.05,

1 0.01, 0.001 and 0.0001 after 2-way ANOVA with Tukey's post-test C-D) YY1 and
 2 SLC9A3R1 enhancer RIs are shown for all available ENCODE cell lines. MCF7
 3 SLC9A3R1 RI is significantly different from the median value calculated on the entire
 4 population (without MCF7, green circle). RI and relative RNA levels (RPKM) are
 5 shown when available for the same cell type. Asterisks represent significance at
 6 $P < 0.05$, 0.01, 0.001 and 0.0001 after one sample T-Test.

7
 8 Supplementary Figure 12. A) SLC9A3R1 RNA levels from three independent RNA-
 9 seq datasets of normal tissues. Images were obtained using Protein Atlas Tools B)
 10 SLC9A3R1 enhancer activity was classified based on the relative RI index in each
 11 tissue. IHC meta-analysis from the Protein Atlas initiative supports the predicted
 12 heterogeneity based on enhancer activity.

13
 14 Supplementary Figure 13. A) Protein Atlas sections stained with the indicated
 15 antibody were scored and ICH+ positive cells were plotted on the radial axis. Data
 16 for each patient are plotted in each corner. Two sections were examined for most
 17 patients. Green lines indicate clonal staining.

18