nature research

Corresponding author(s): Luca Magnani

Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

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

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

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

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

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

Data analysis

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

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

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

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

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

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

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

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

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

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

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

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

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

Data

Policy information about availability of data

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

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

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

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences

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

Life sciences

Study design

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

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

Materials & experimental systems

Policy information about availability of materials

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

Antibodies

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

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

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

Validation

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

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

Authentication

Authentication Karyotyping was perfomed for all cell lines

Mycoplasma contamination

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

Commonly misidentified lines
(See ICLAC register)

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

Human research participants

Policy information about studies involving human research participants

Population characteristics

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

Method-specific reporting

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

ChIP-seq

Data deposition

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

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

Data access links
May remain private before publication.

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

Files in database submission

Raw reads and Peak files

Genome browser session
(e.g. UCSC)

NA

Methodology

Replicates

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

Sequencing depth

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

Antibodies

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

Peak calling parameters

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

Data quality

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

Software

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