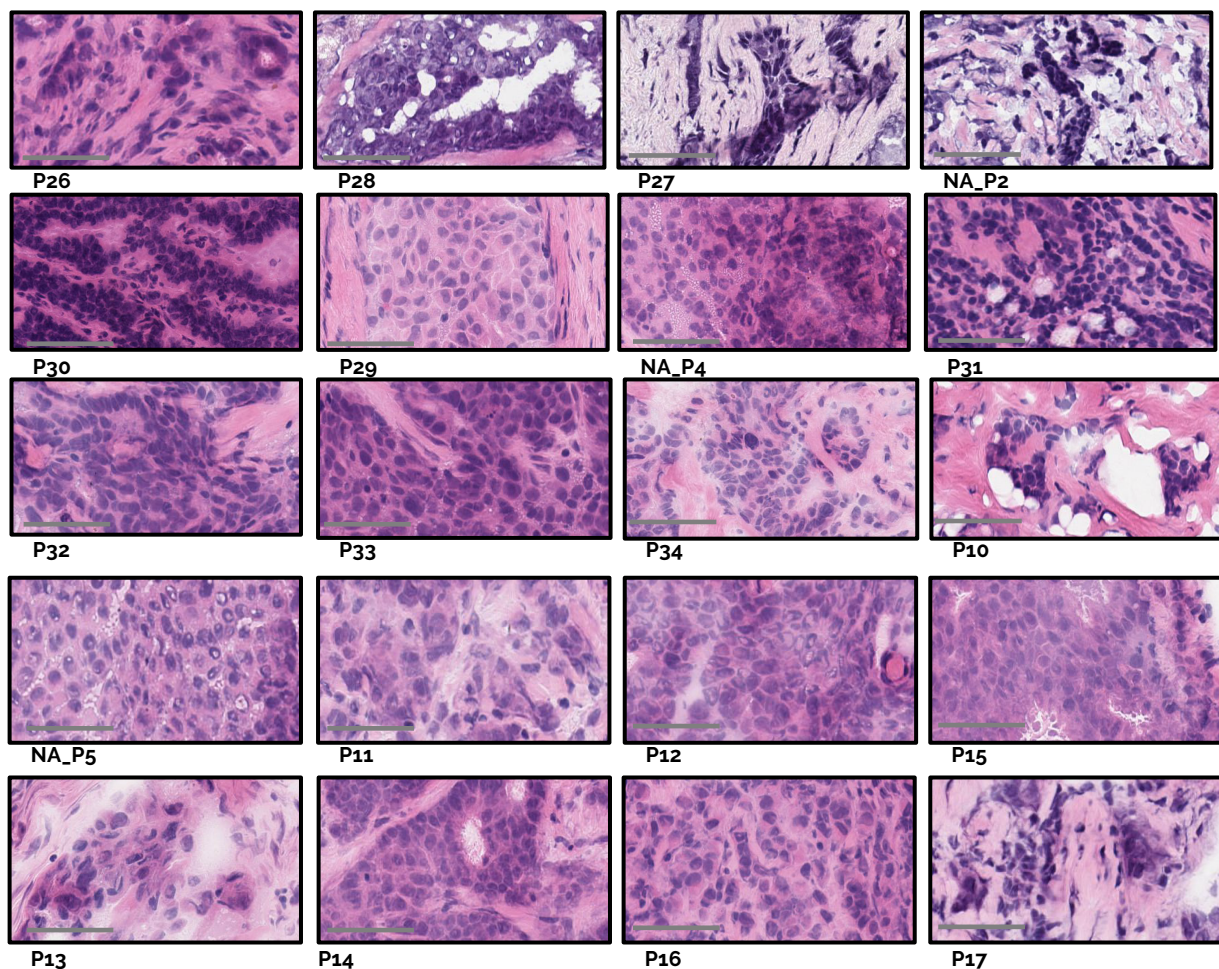
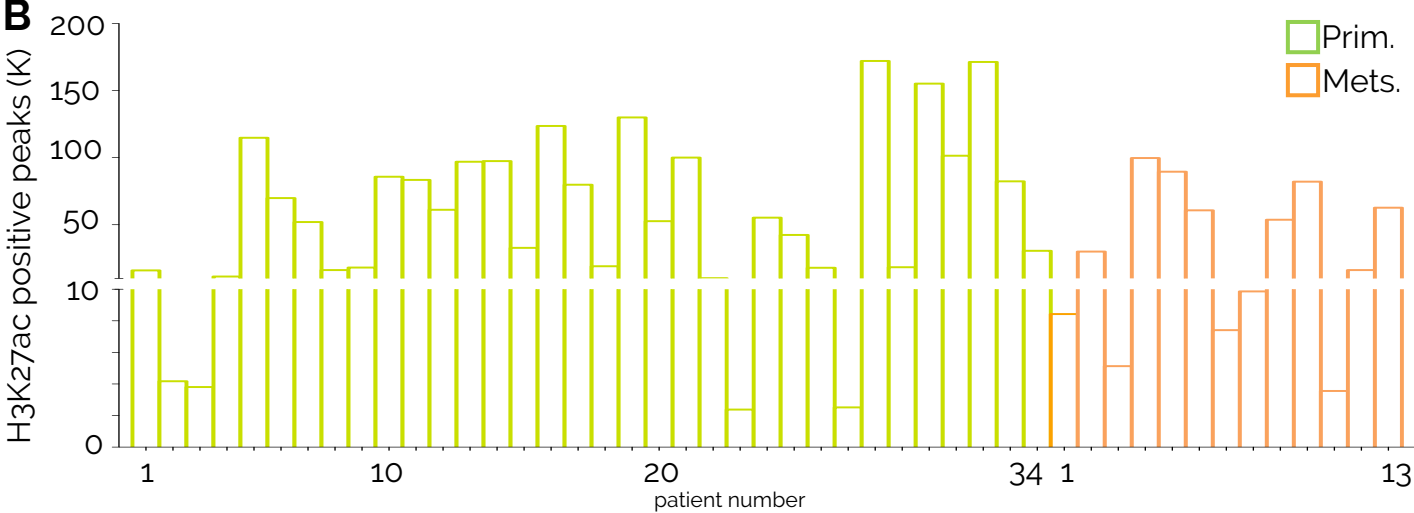
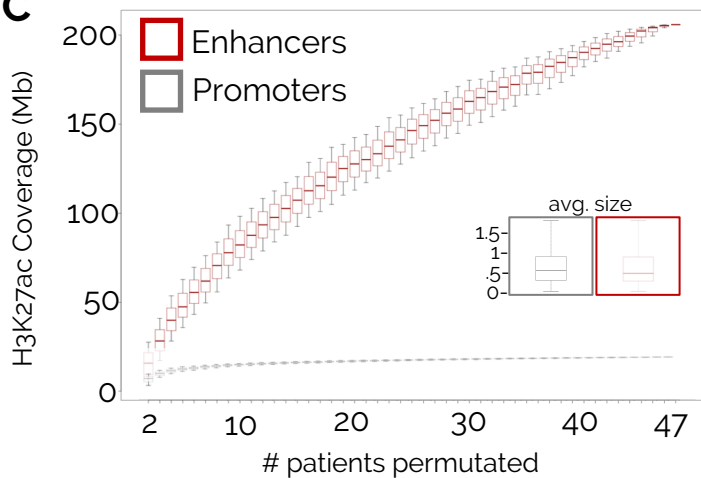
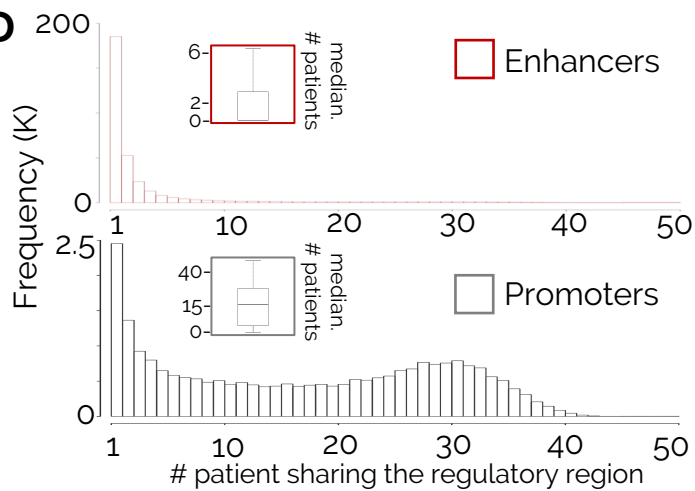
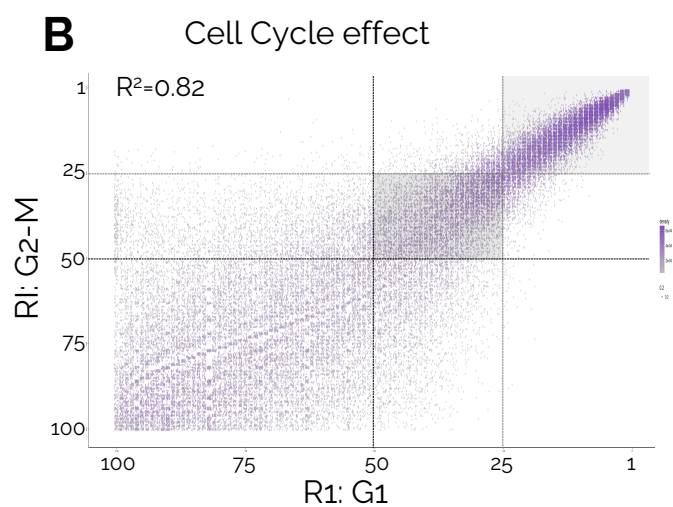
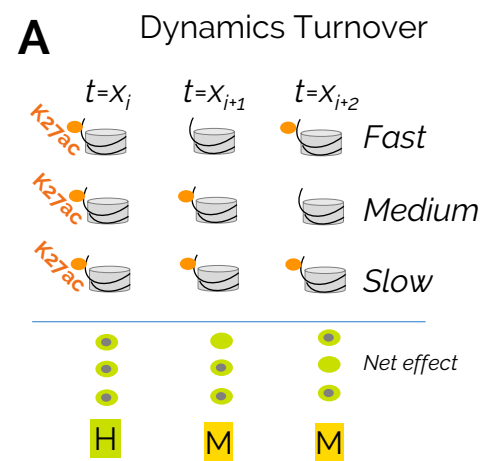
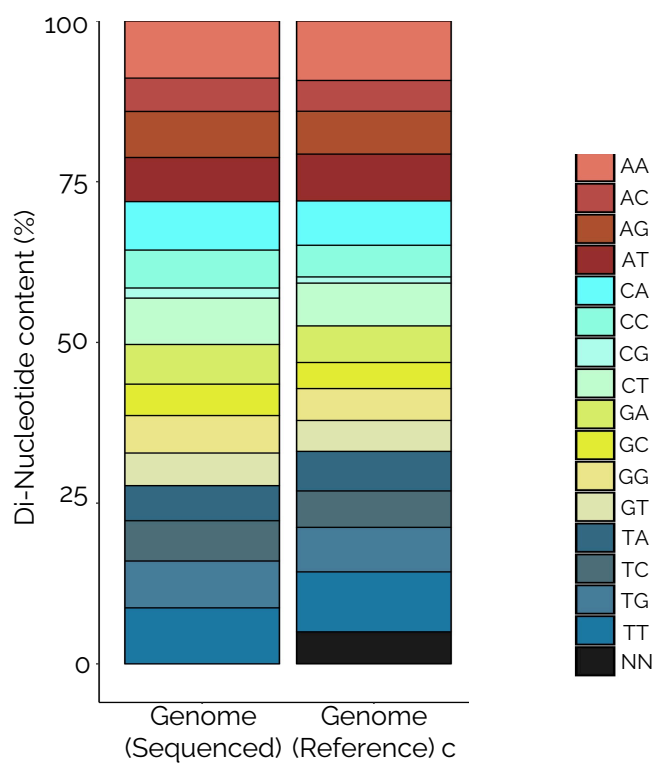
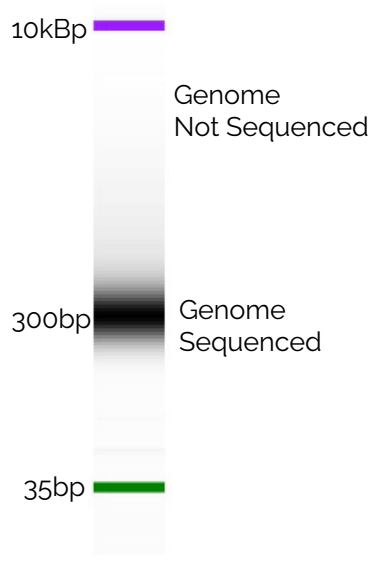


**Figure S1****A****B****C****D**

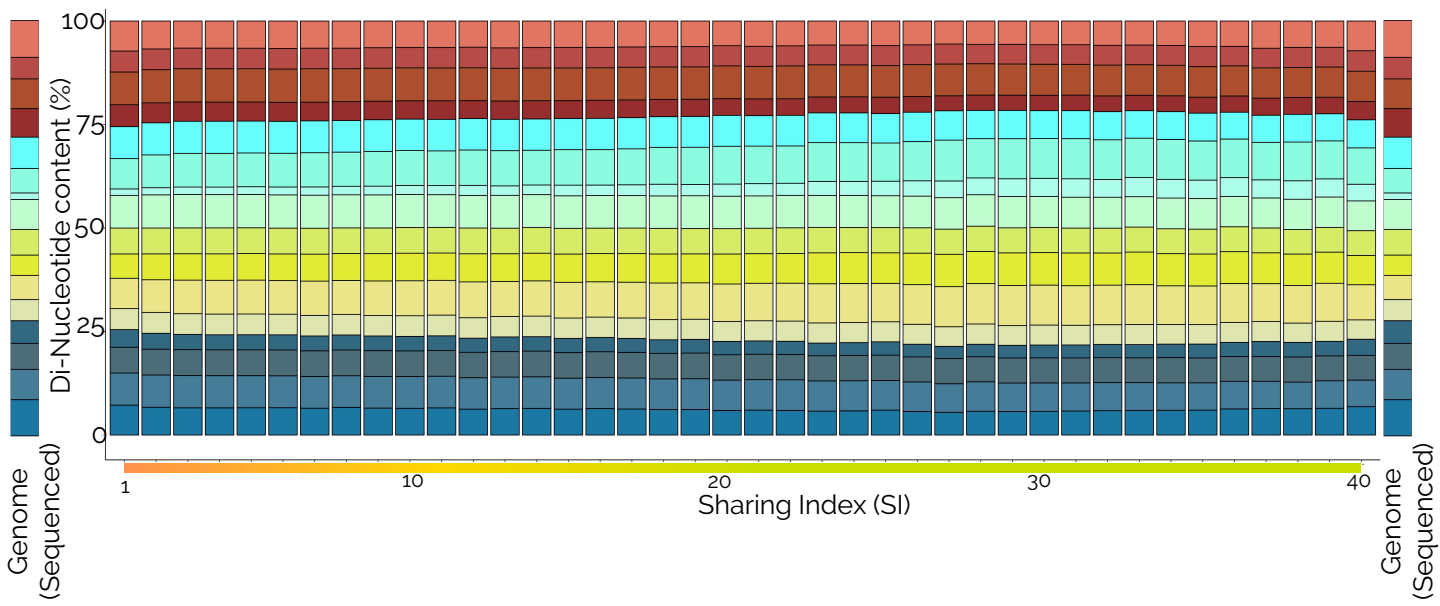
# Figure S2



**C** Sonication effect

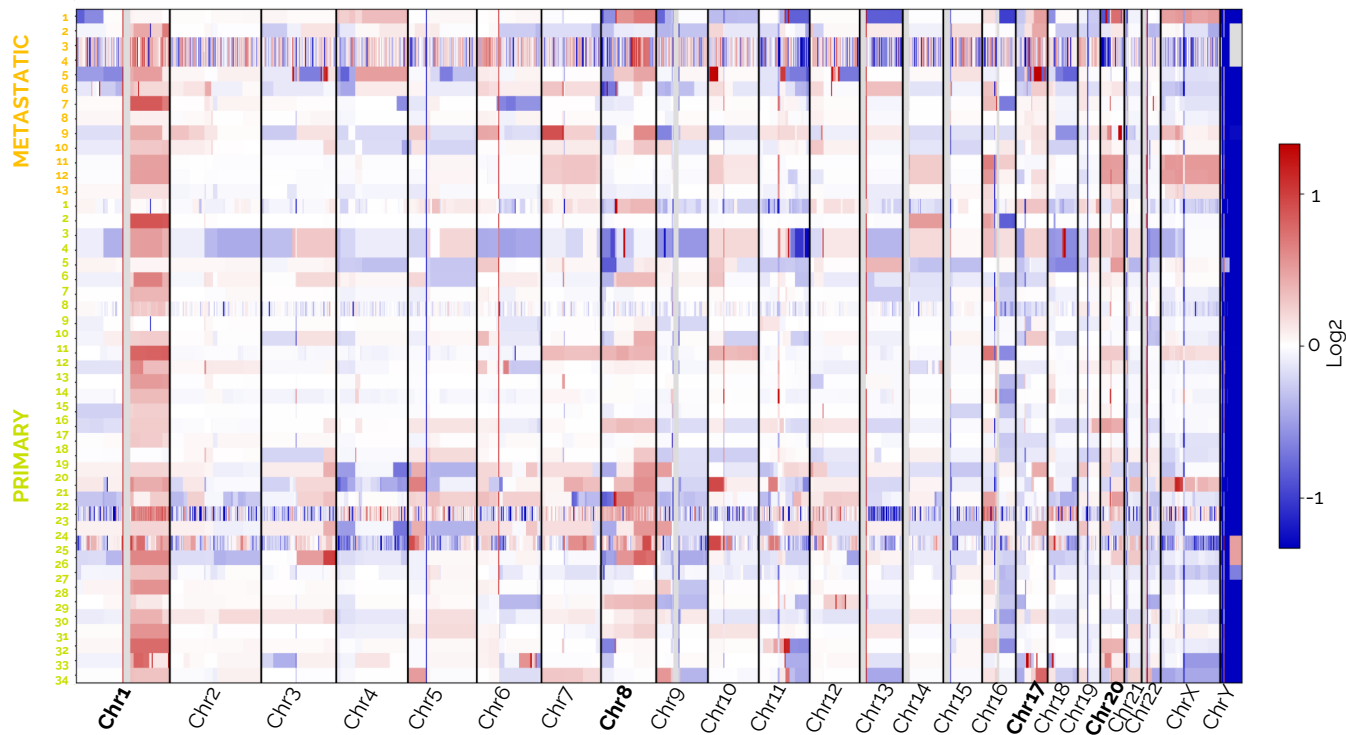


**D** Di-Nucleotide content

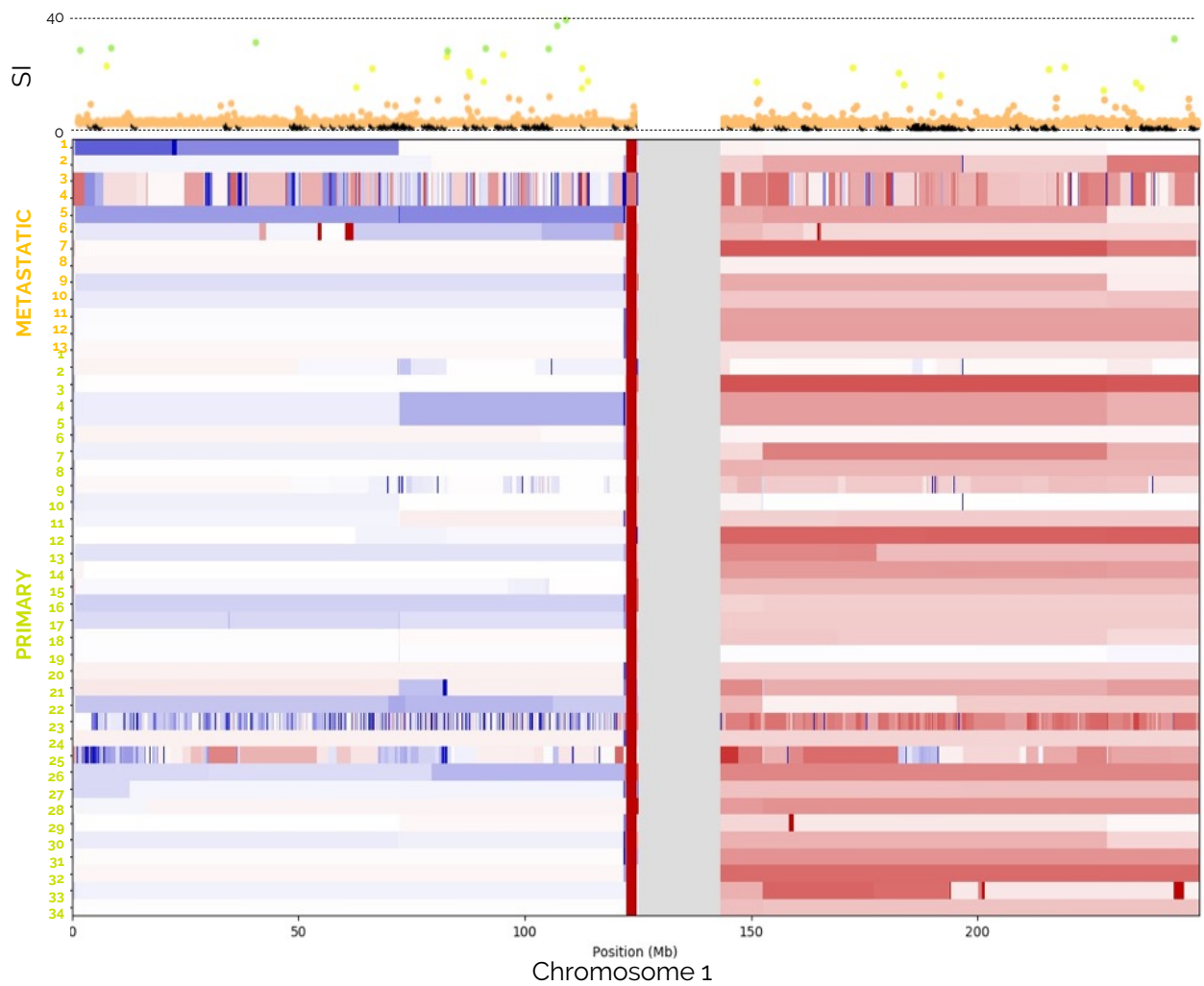


# Figure S3

## A Copy Number from patient dataset (Input Shallow-Seq)

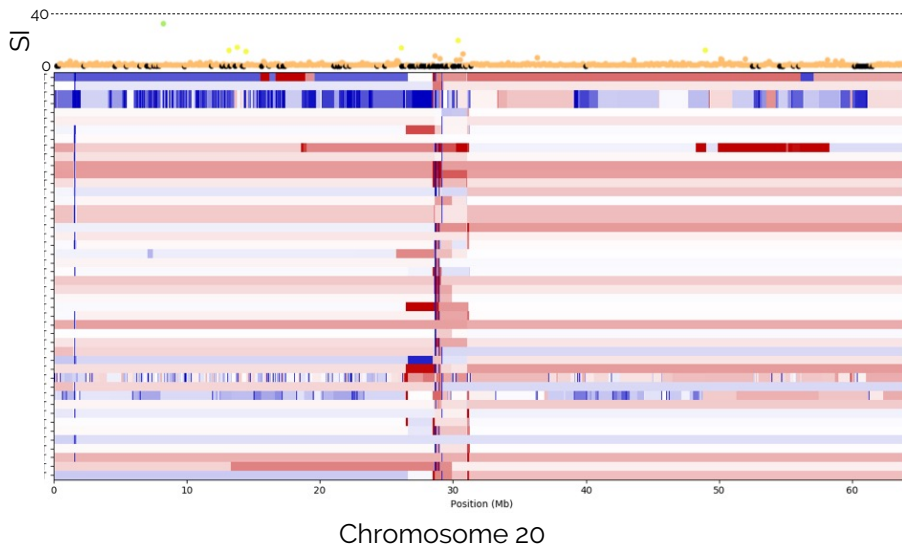
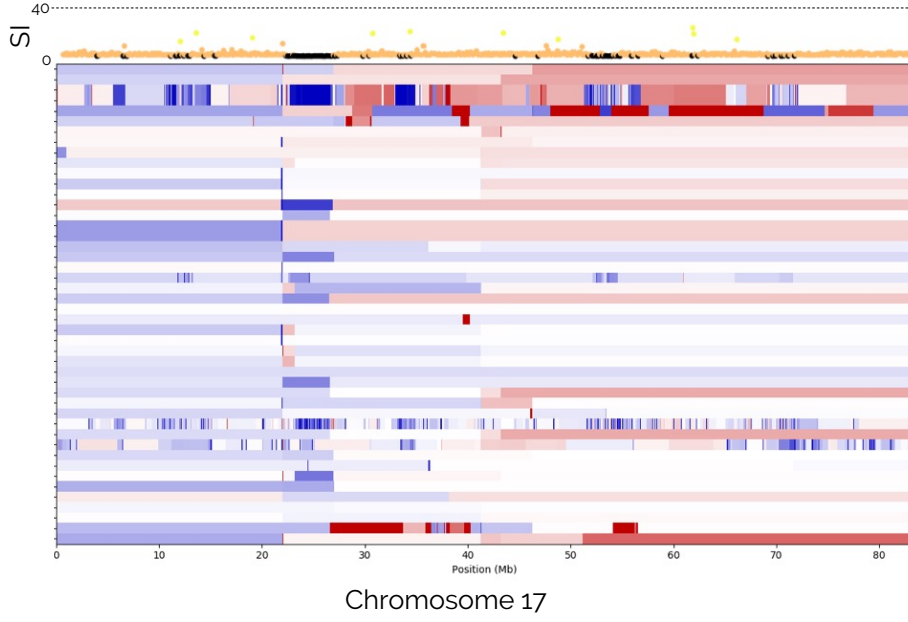
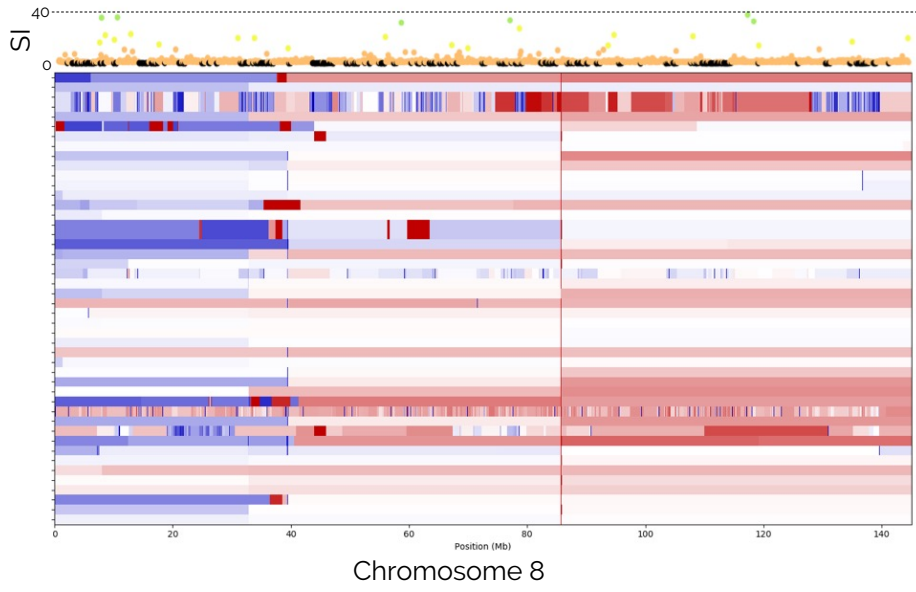


## B Copy Number vs. Sharing Index



# Figure S4

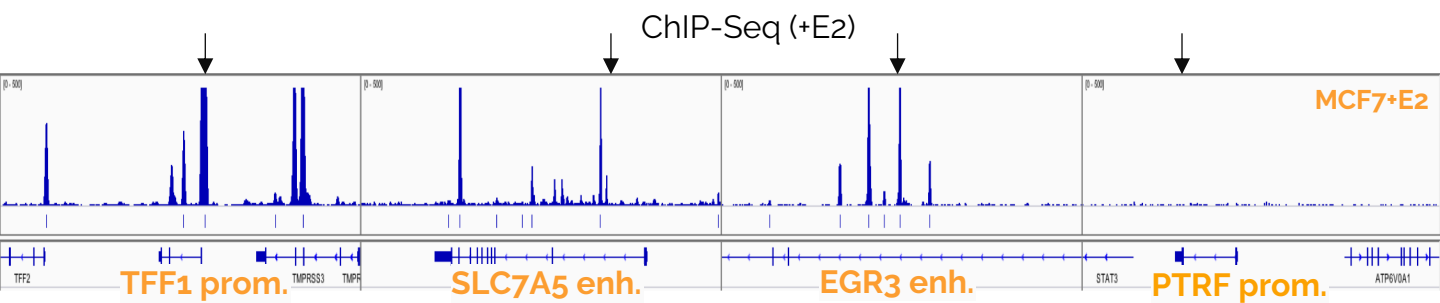
Copy Number vs. Sharing Index



# Figure S5

## A

Enhancer Selection: **MCF7+E2: ON**, MCF7-E2: OFF



## B

MCF7+E2

MCF7-E2



ChIP- ERα

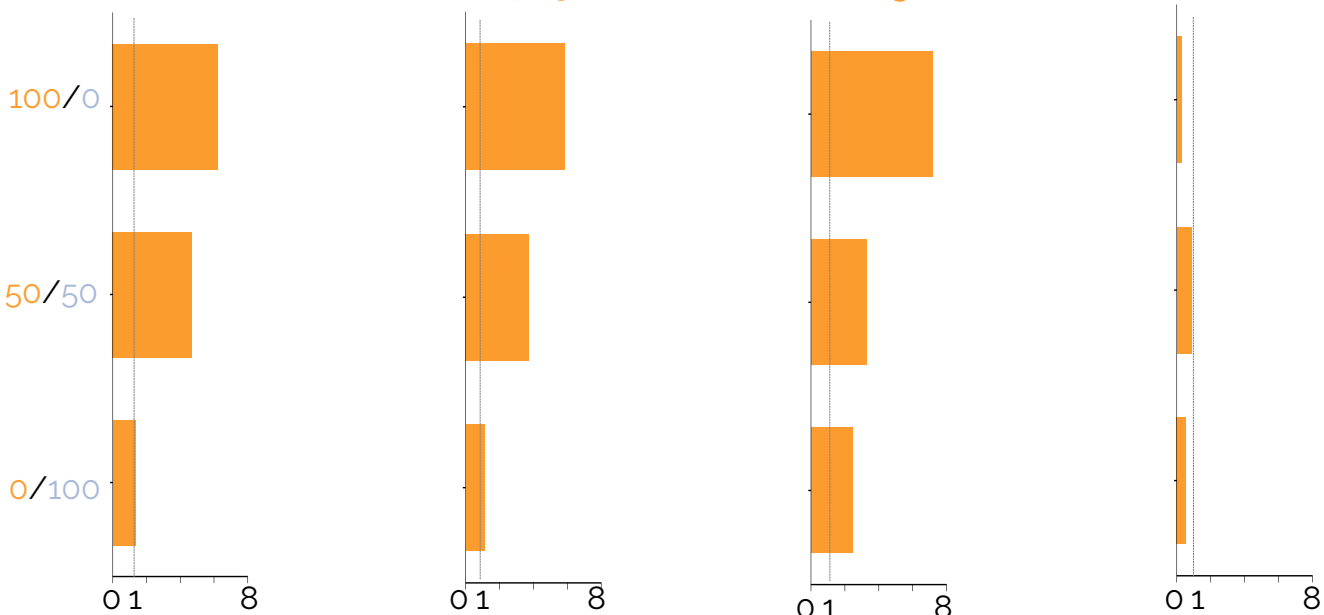


TFF1

SLC7A5

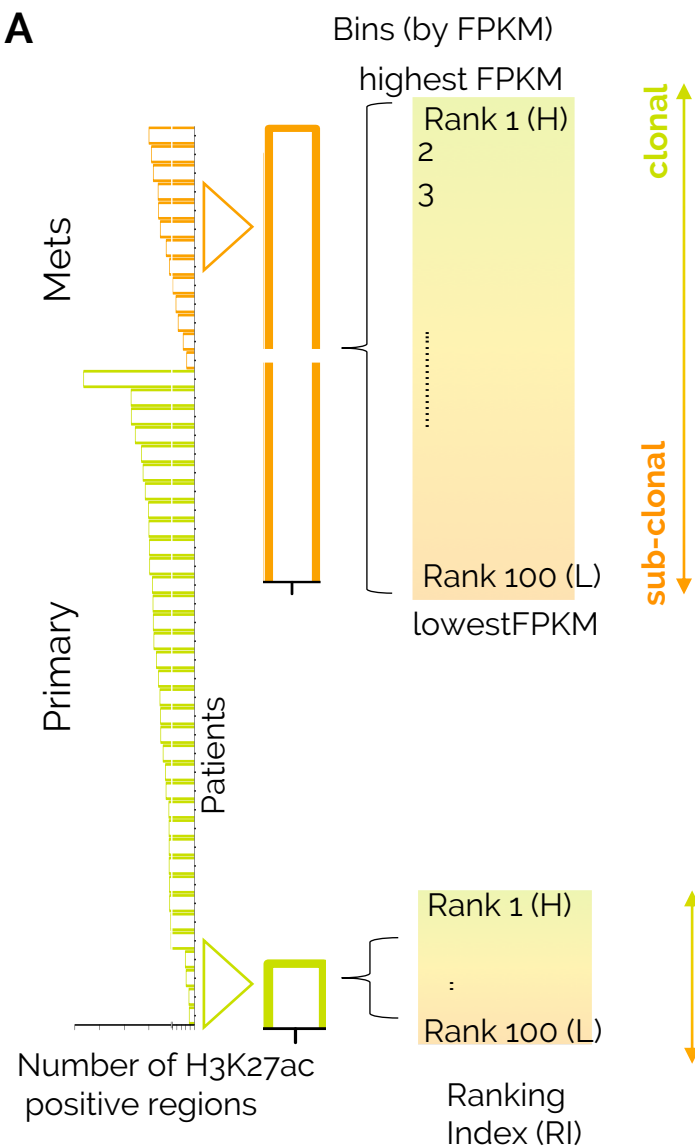
EGR3

PTRF

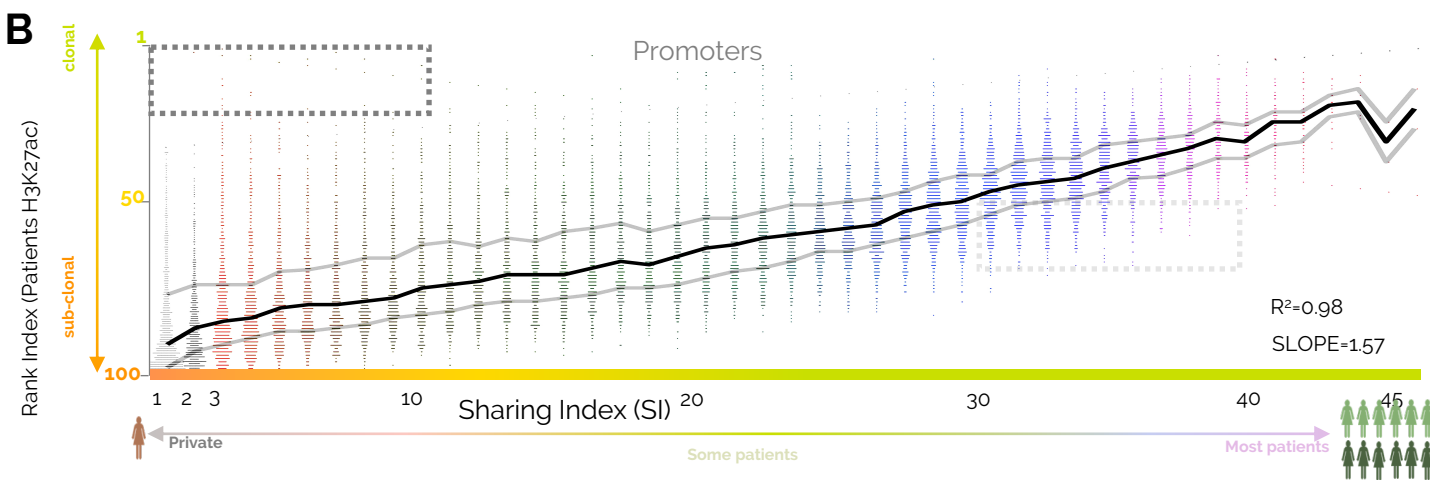


# Figure S6

## A

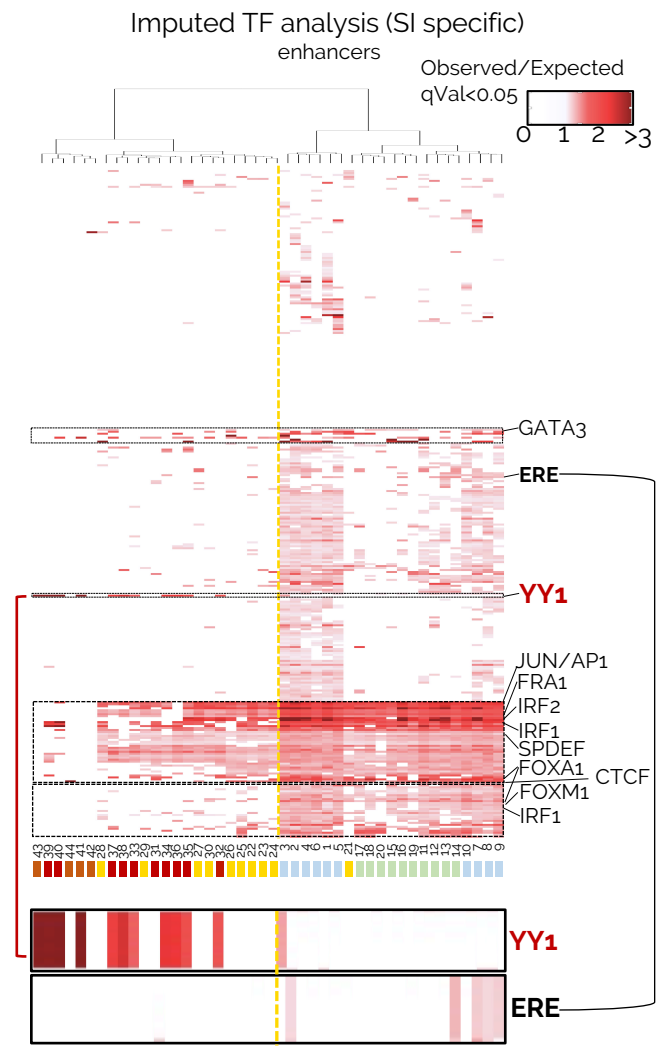
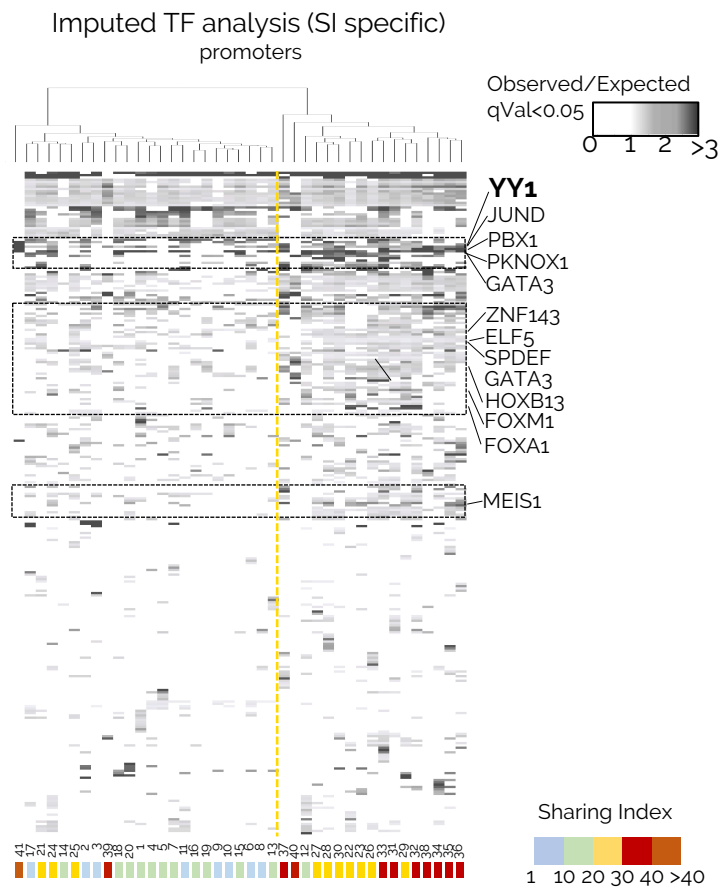


## B





**Figure S8**





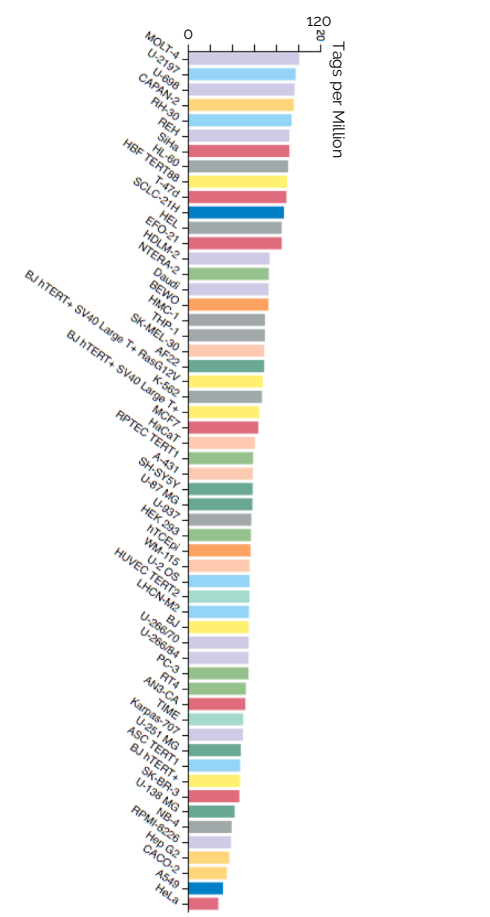
# Figure S9

## A

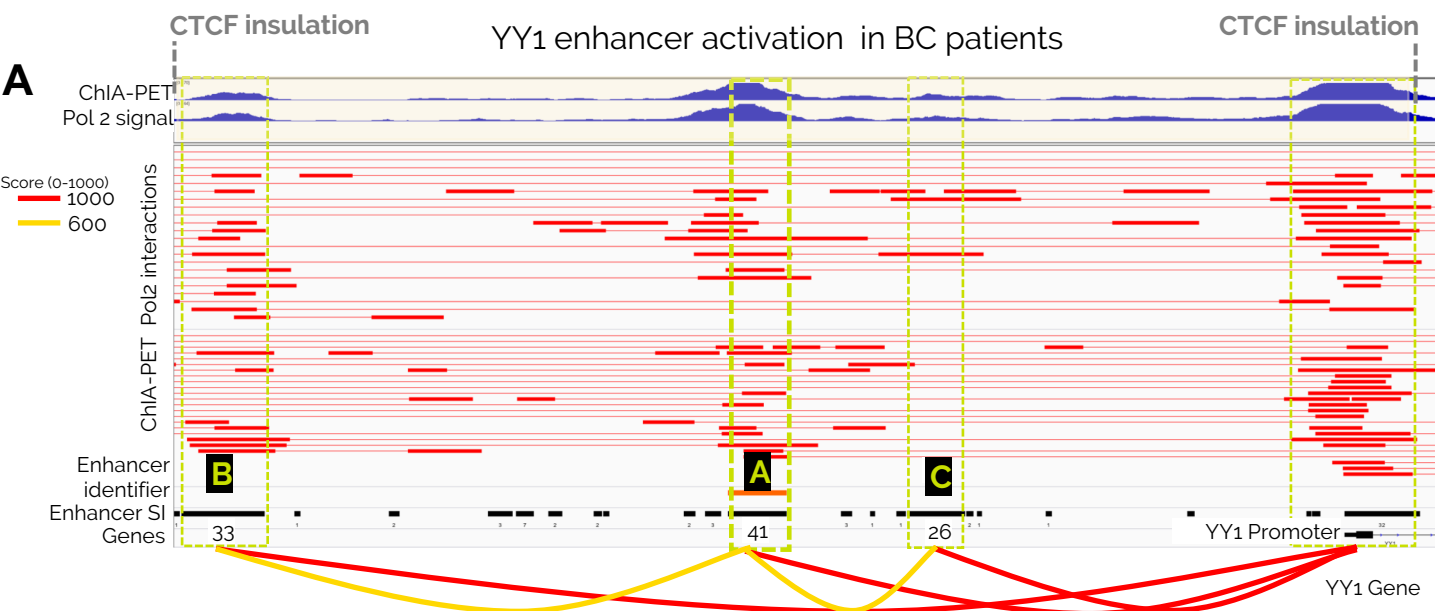


## B

### YY1 expression Cancer Cells



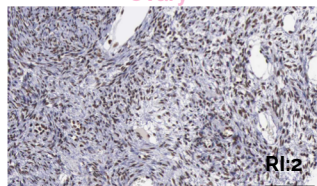
# Figure S10



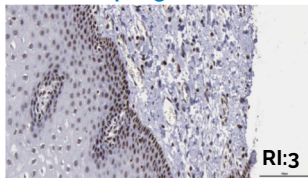
**B**

**H R.I. <20**

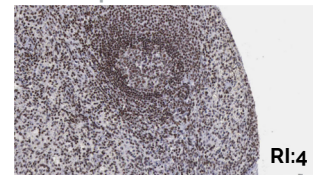
Ovary



Esophagus

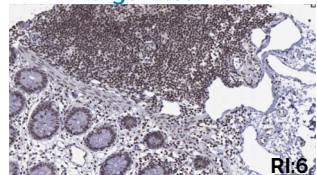


Spleen



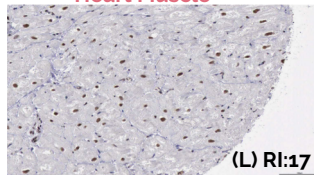
Cells in Red Pulp: >75%  
Intensity: High  
Cells in White Pulp: 25-75%

Large Intestine

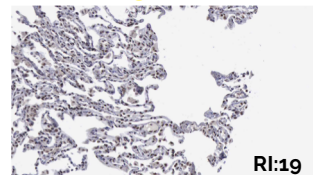


Glandular cells: >75%  
Peripheral Nerve: >75%  
Endothelial Cells: >75%

Heart Muscle



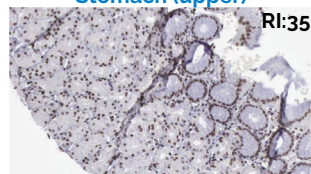
Lung



Macrophages: 25-75%  
Pneumocytes: 25-75%

**M R.I. 50-80**

Stomach (upper)



Glandular cells: 25-75%

Adipose

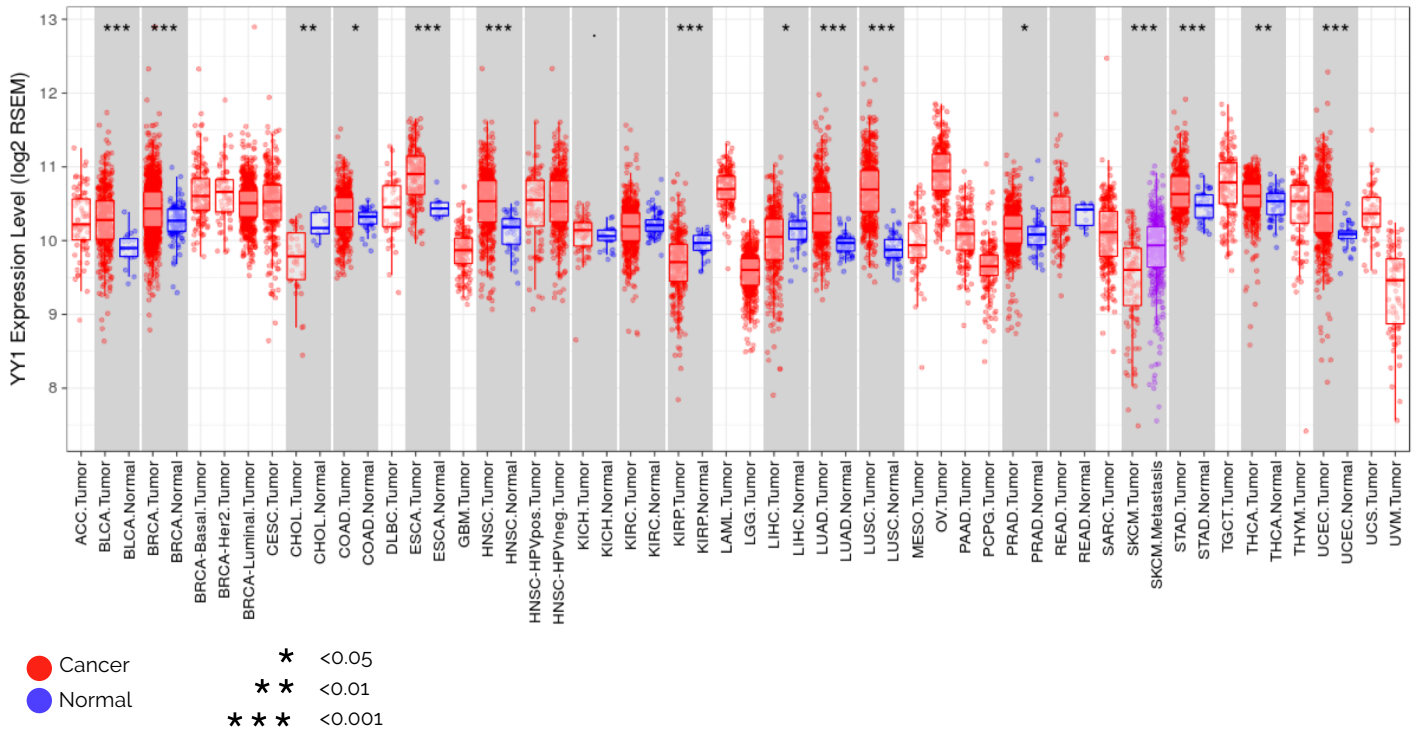


Adipocytes: >75%  
Fibroblast: not detected

# Figure S11

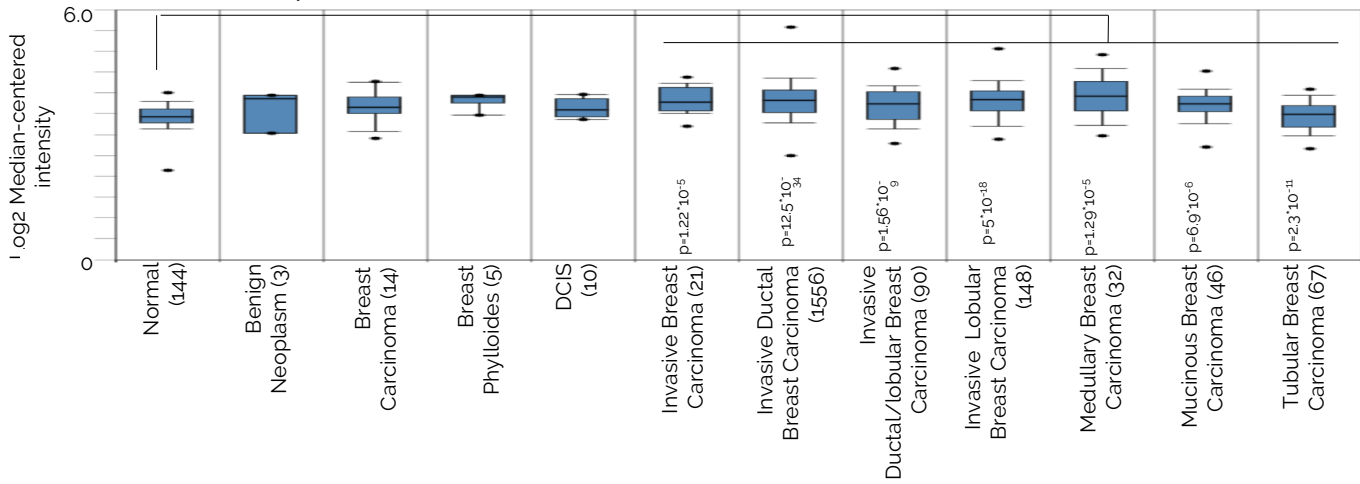
**A**

## YY expression PAN CANCER TCGA



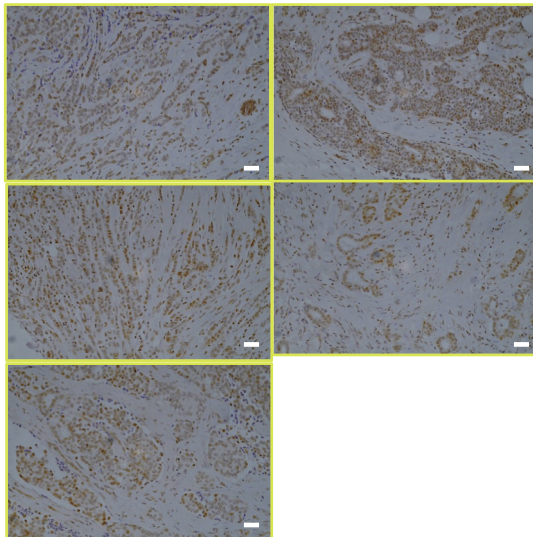
**B**

## YY1 Expression METABRIC BCa

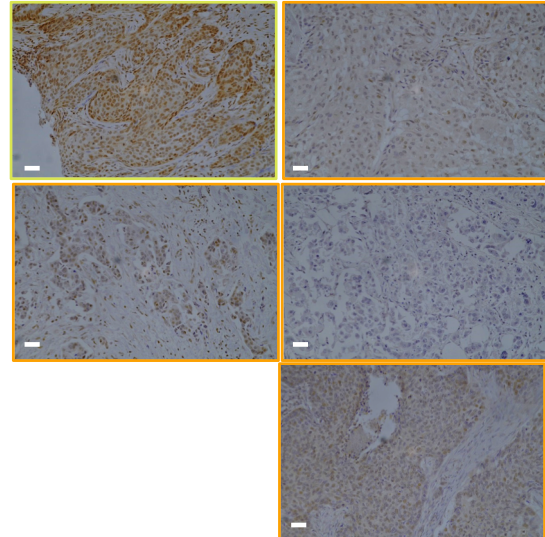


**C**

## YY1 IHC ER+ BC

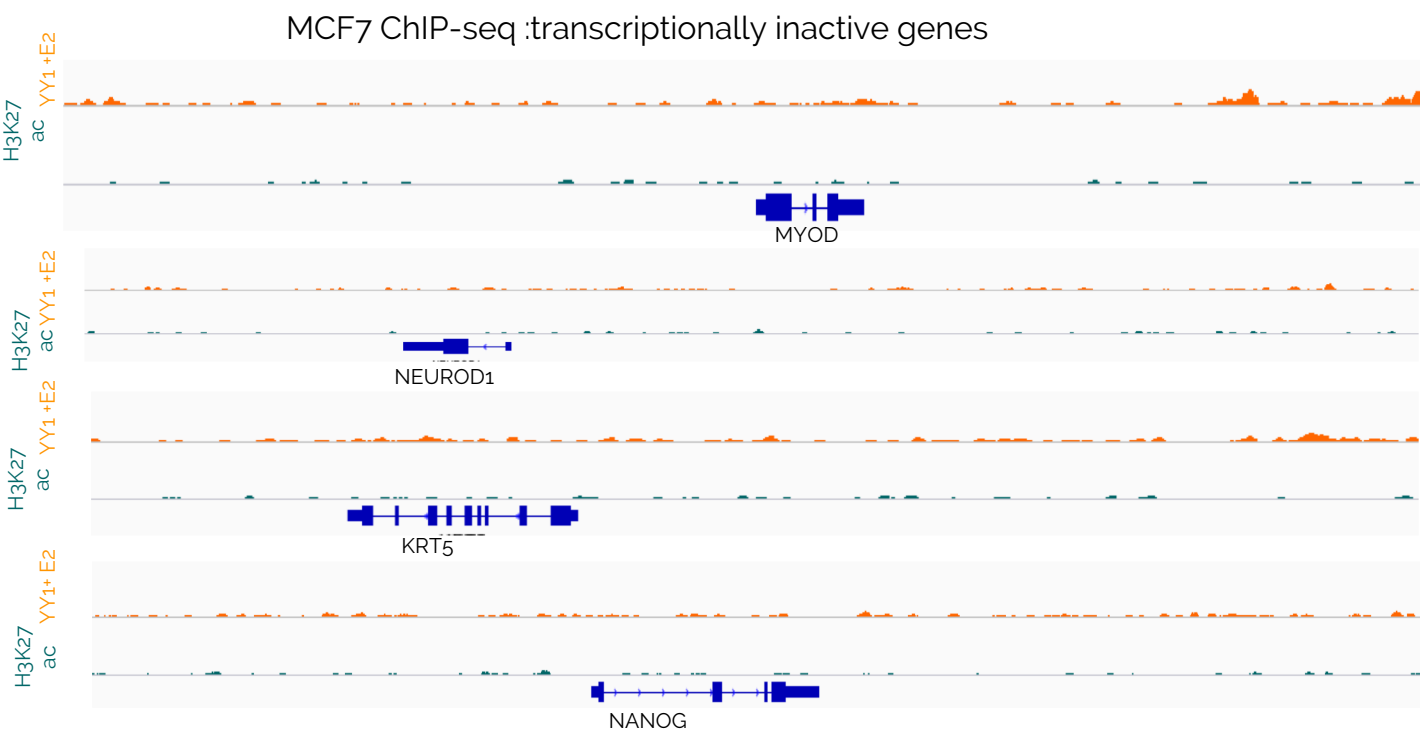


## YY1 IHC ER- BC

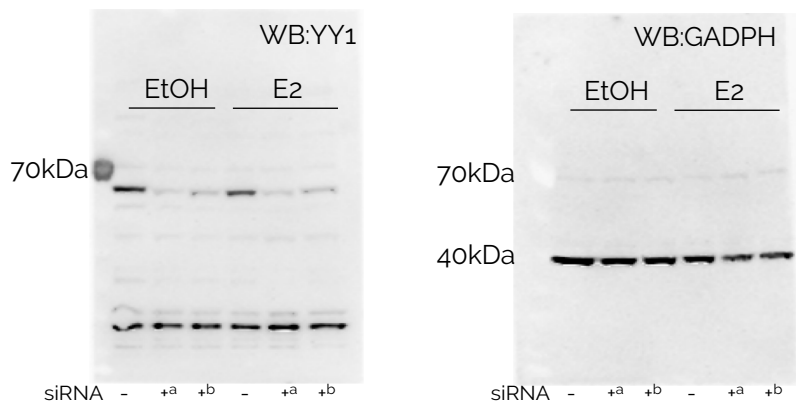


# Figure S12

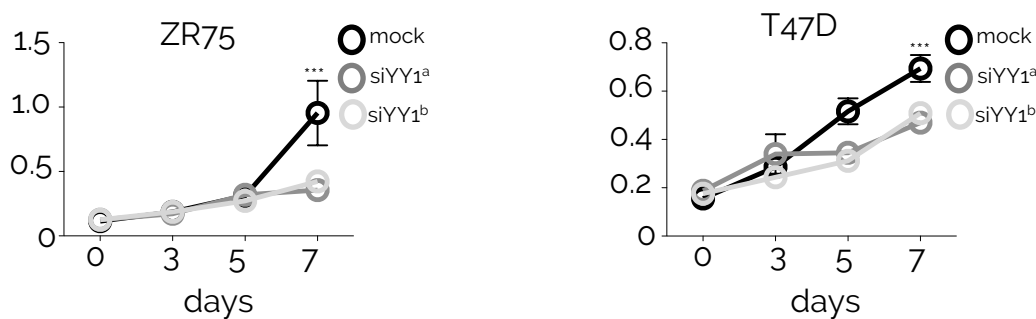
## A



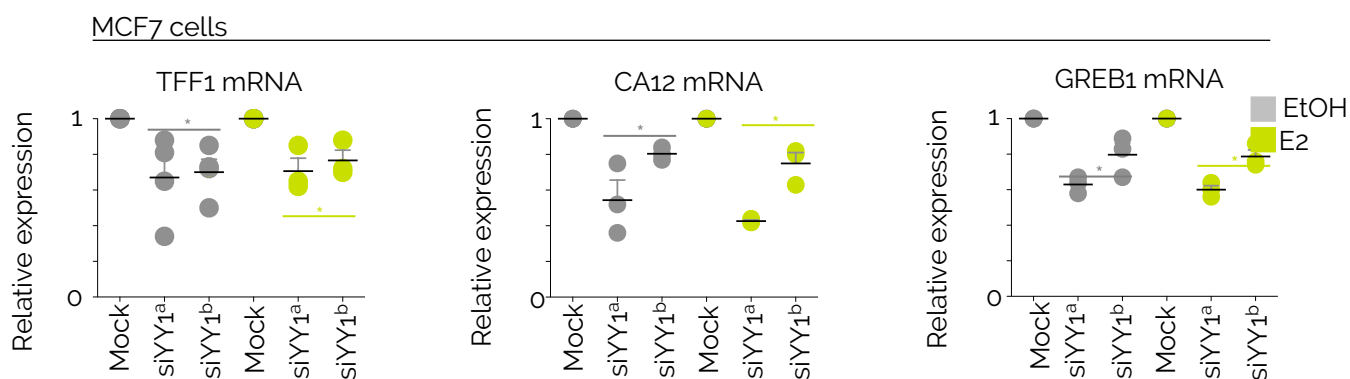
## B



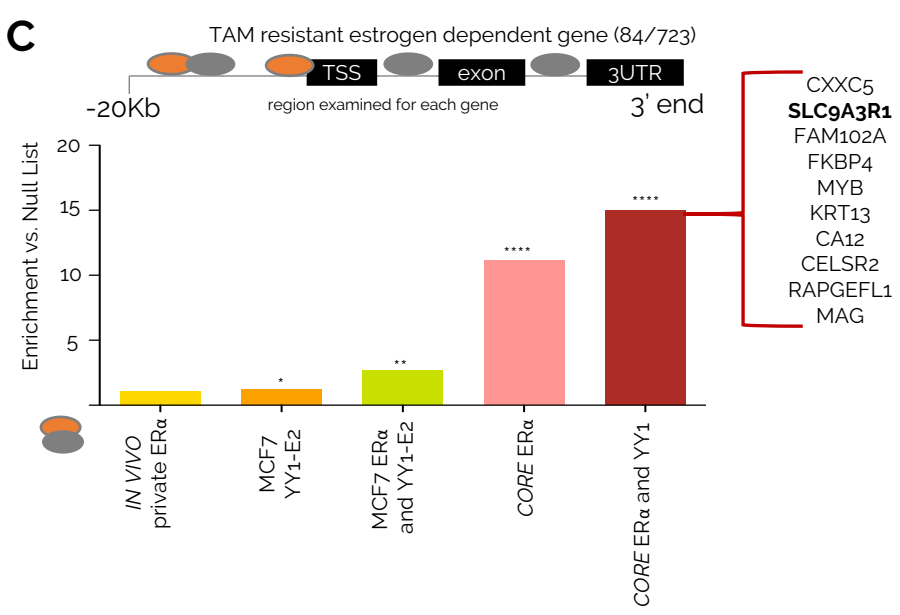
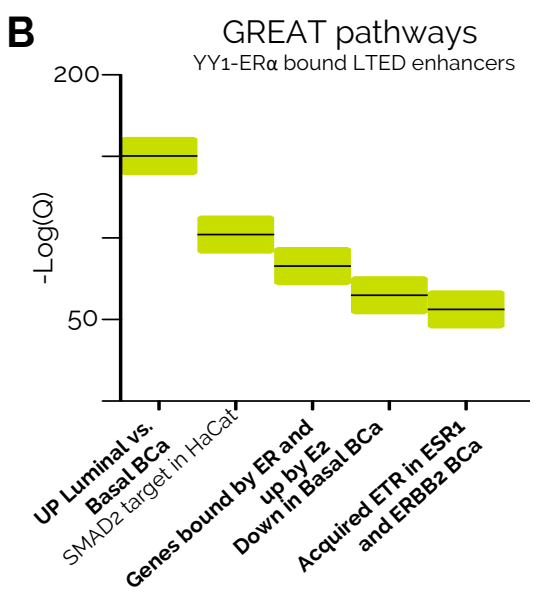
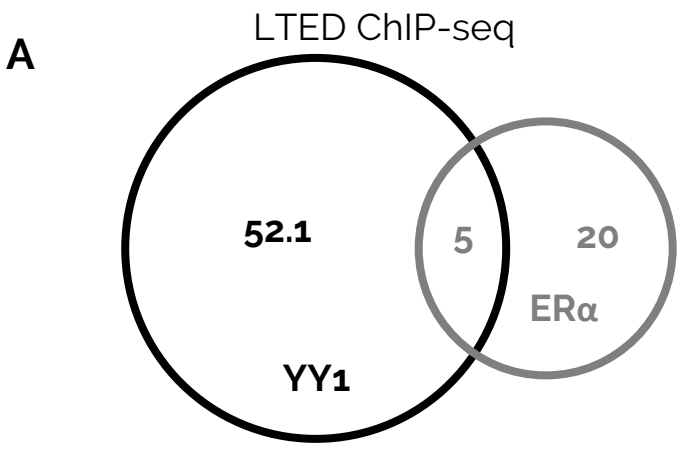
## C



## D

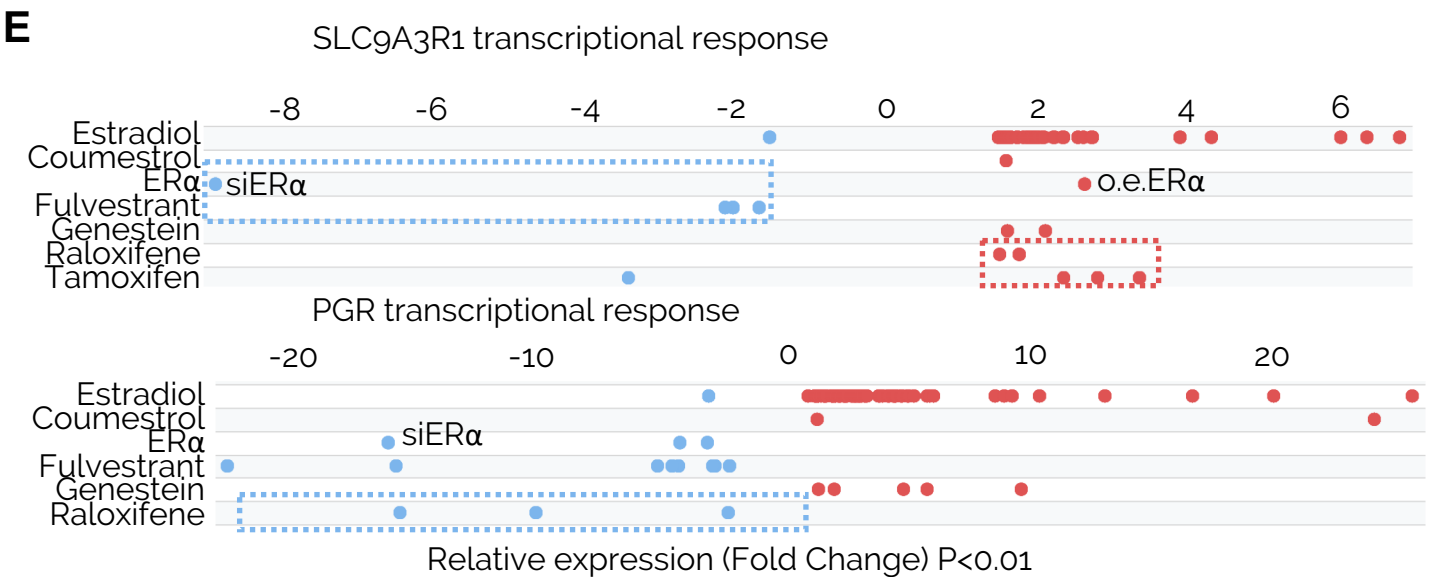
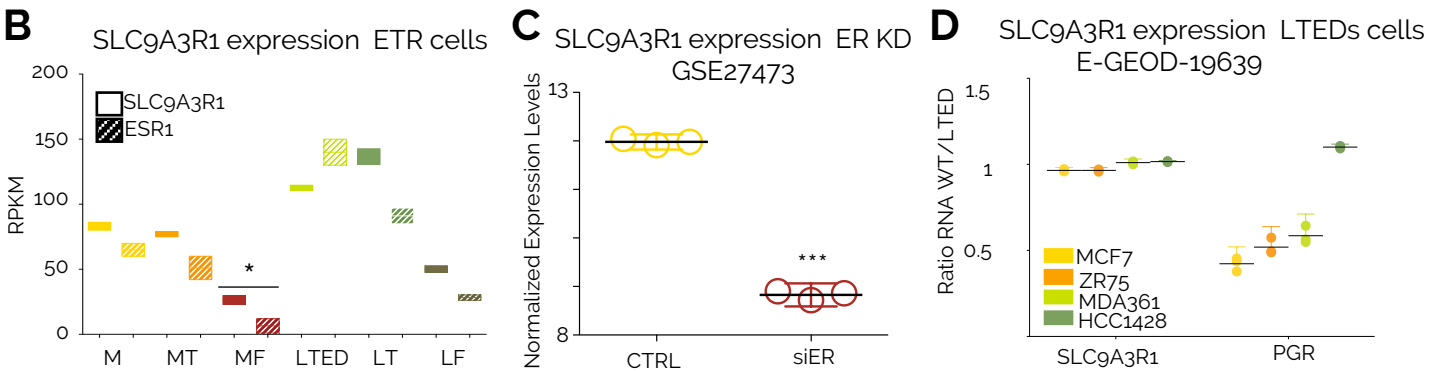
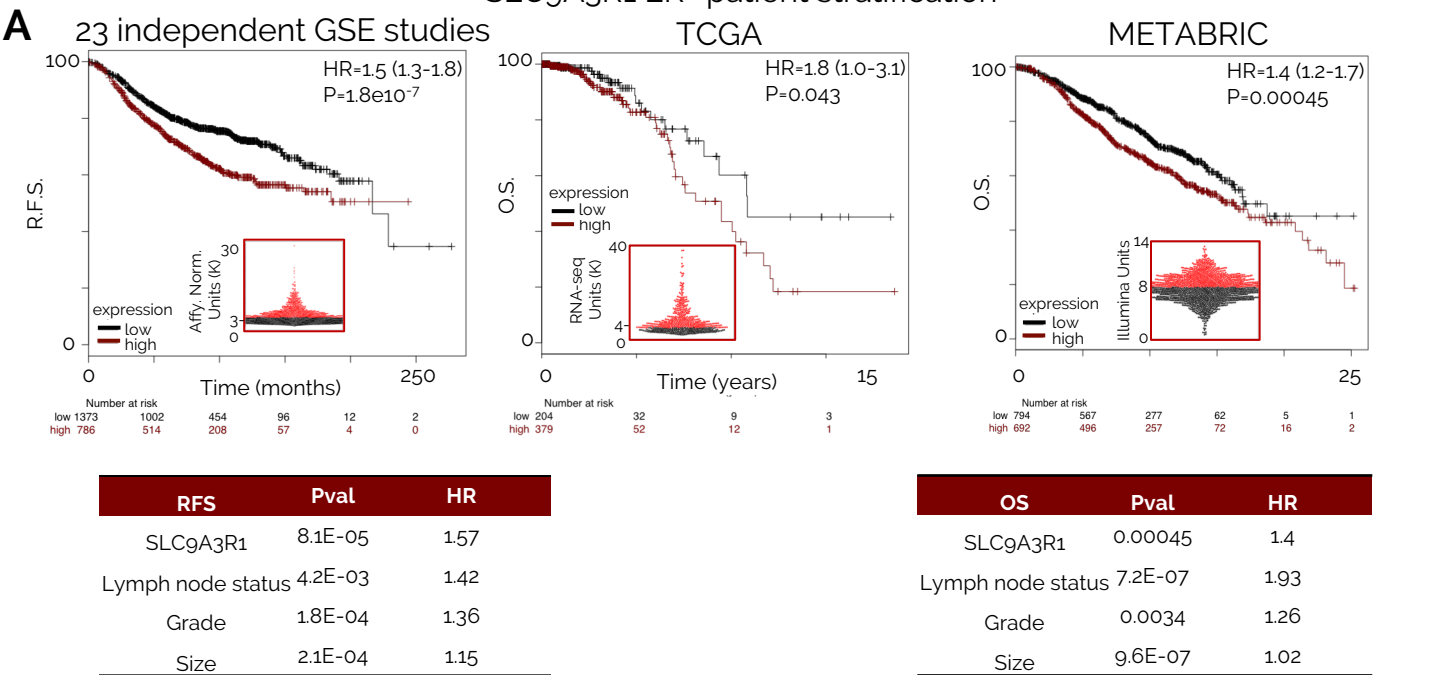


# Figure S13



# Figure S14

## SLC9A3R1 ER+ patient stratification

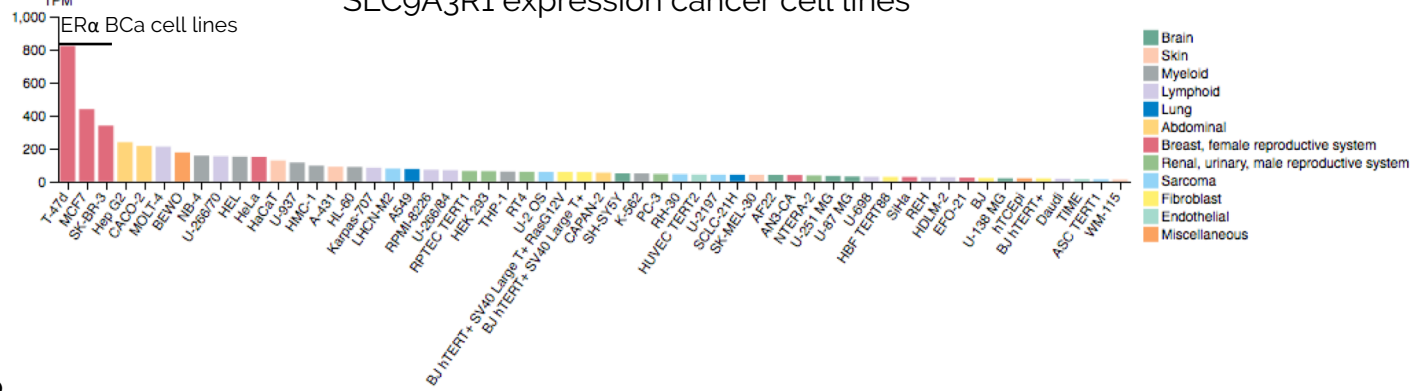


# Figure S15

A

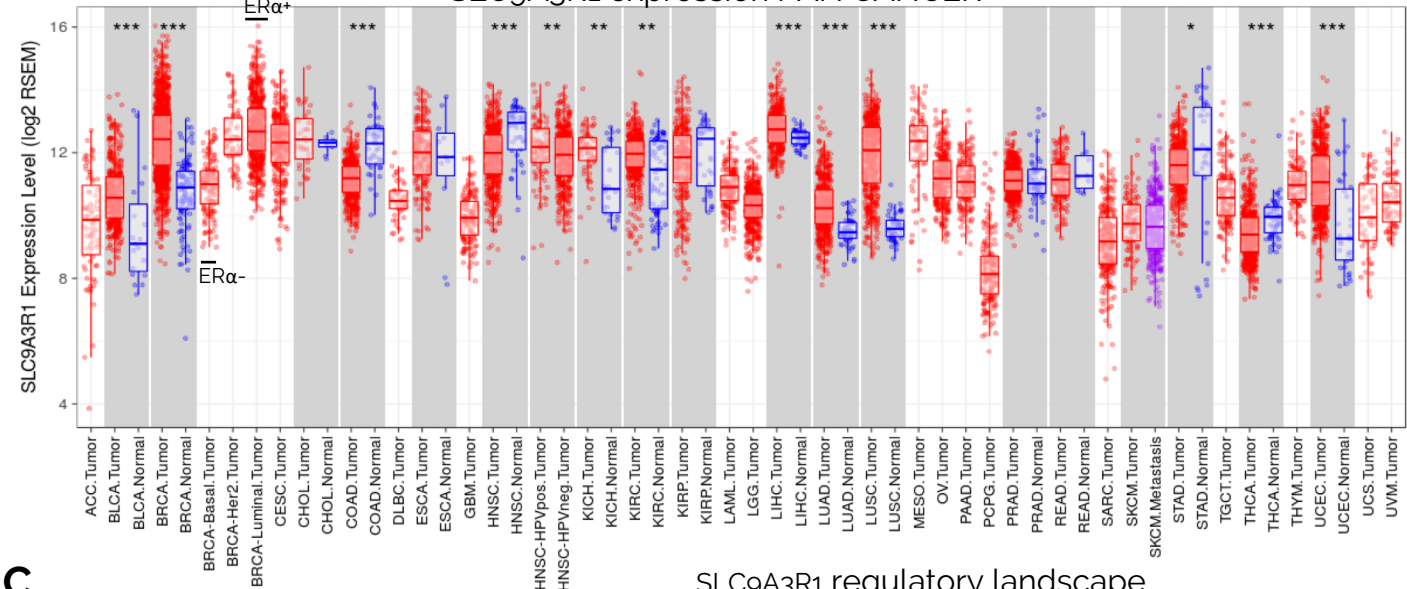
## SLC9A3R1 expression cancer cell lines

TPM



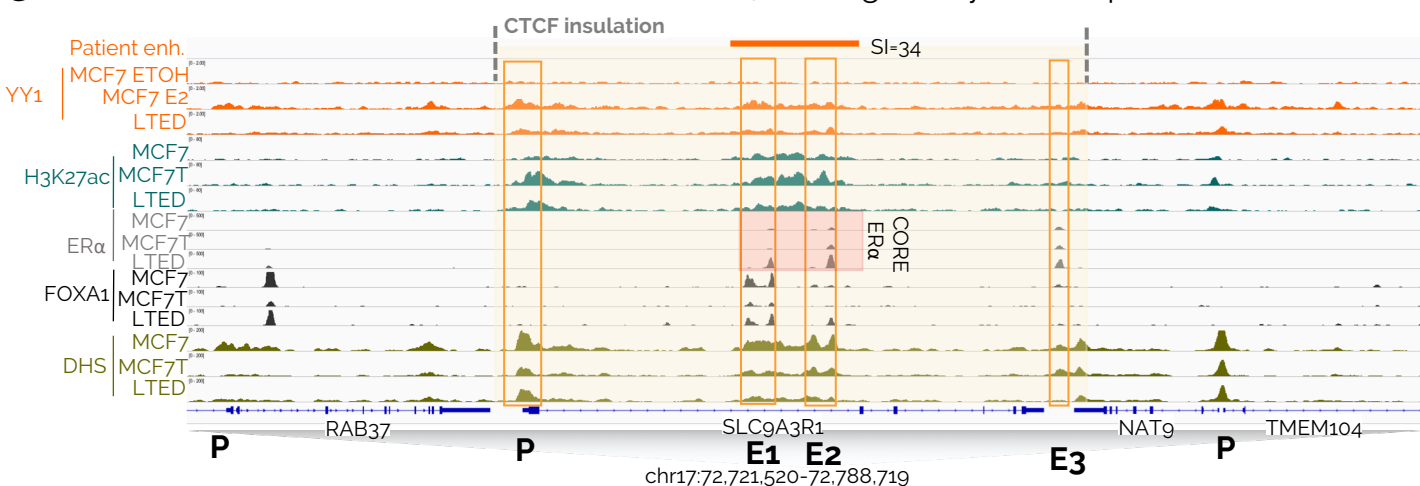
B

## SLC9A3R1 expression PAN CANCER

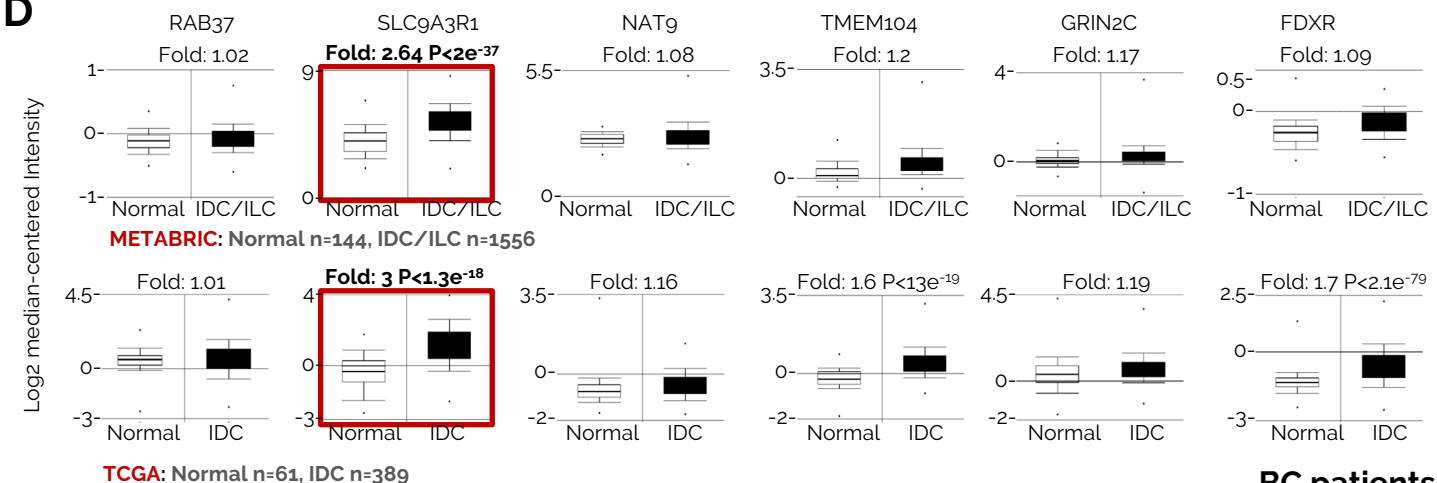


C

## SLC9A3R1 regulatory landscape

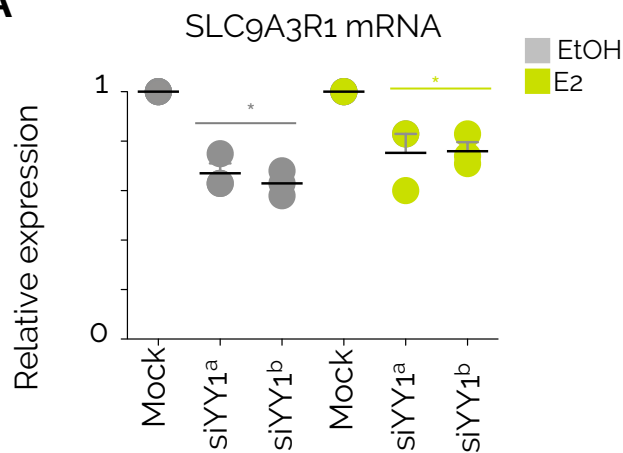


D

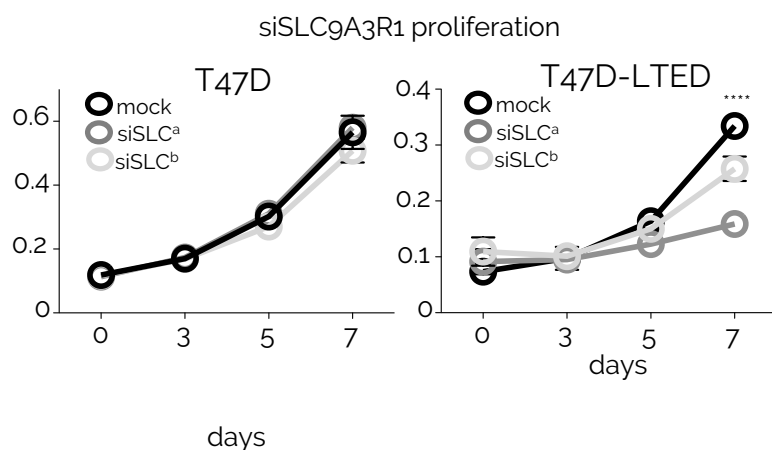
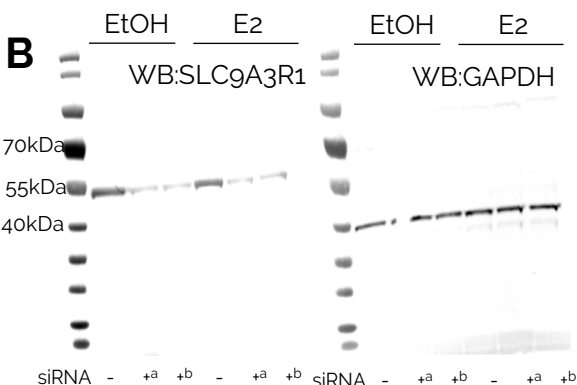


# Figure S16

**A**

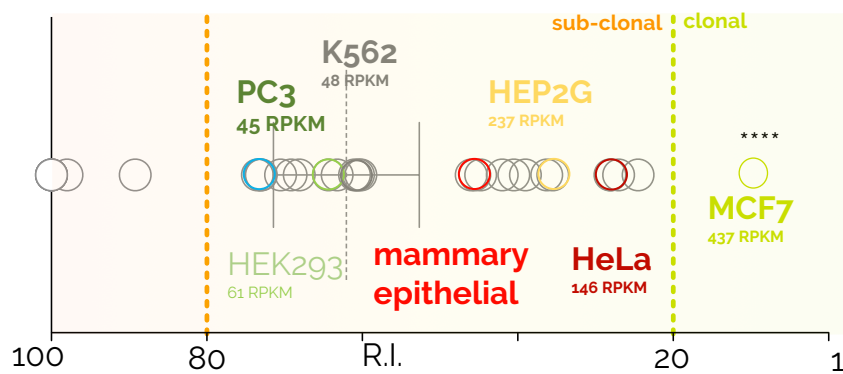


**B**



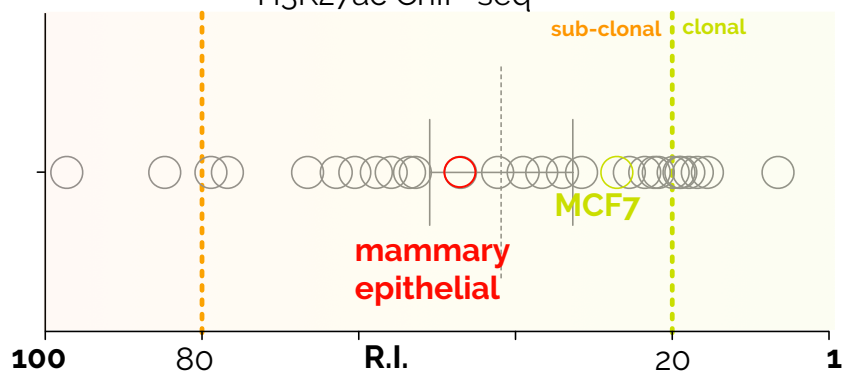
**C**

SLC9A3R1 Enhancer activity ENCODE  
H3K27ac CHIP-seq



**D**

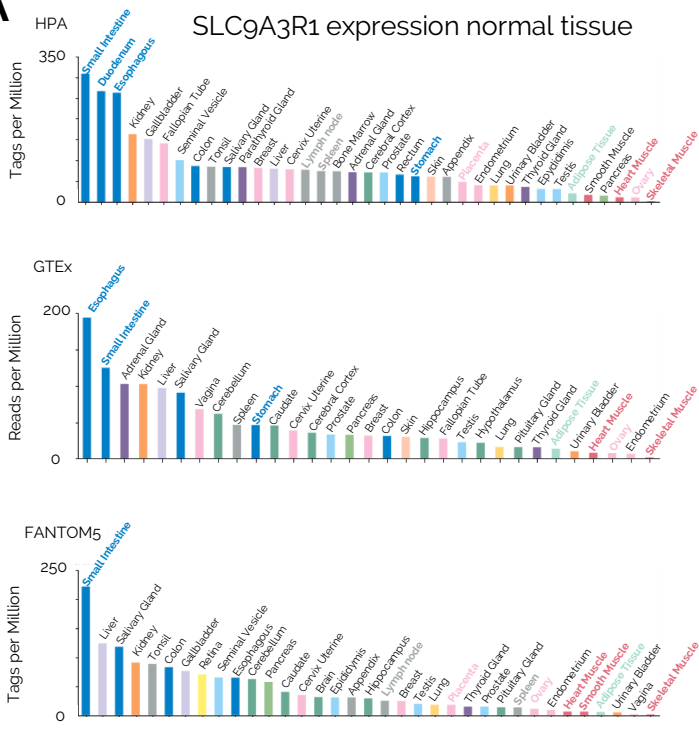
YY1Enhancer activity ENCODE  
H3K27ac CHIP-seq





# Figure S17

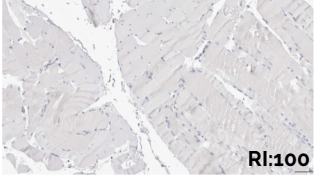
A



B

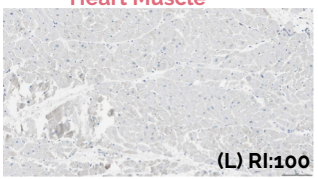
**L R.I. => 80**

**Skeletal Muscle**



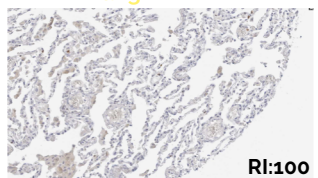
RI:100

**Heart Muscle**



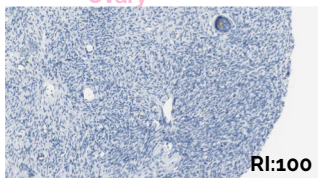
(L) RI:100

**Lung**



RI:100

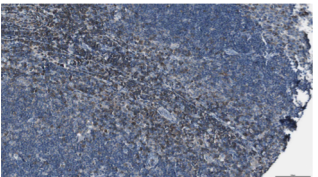
**Ovary**



RI:100

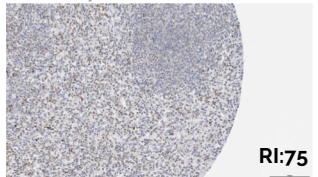
**M R.I. 50-80**

**Lymph node**



Germinal center cells: <25%  
Intensity: Moderate  
Non-Germinal center cells: 25-75%  
Intensity: Moderate

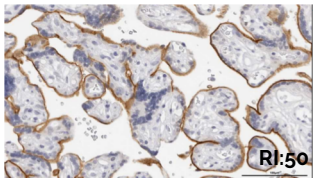
**Spleen**



RI:75

Cells in Red Pulp: 25-75%  
Intensity: Moderate  
Cells in White Pulp: Not Detected

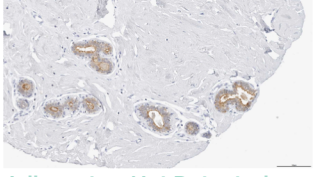
**Placenta**



RI:50

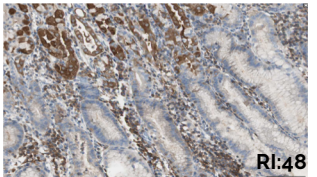
Trophoblastic cell: 75%  
Decidua cells: Not Detected

**Breast**



Adipocytes: Not Detected  
Myoepithelial: Not Detected  
Glandular: 25-75% cells  
Intensity: Moderate

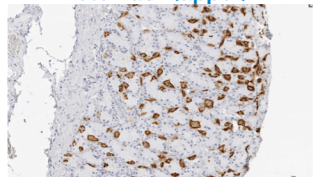
**Stomach (lower)**



RI:48

Glandular cells: <25%  
Intensity: Strong

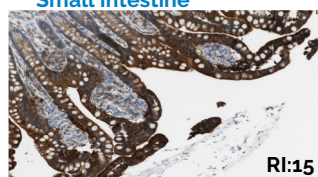
**Stomach (upper)**



Glandular cells: 25-75%  
Intensity: Strong

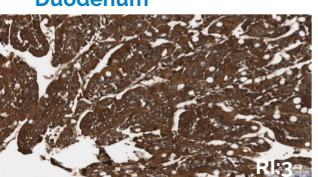
**H R.I. <20**

**Small Intestine**



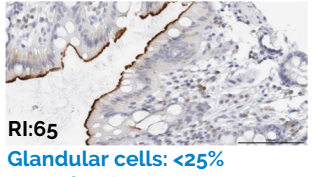
RI:15

**Duodenum**



RI:13

**Rectum**

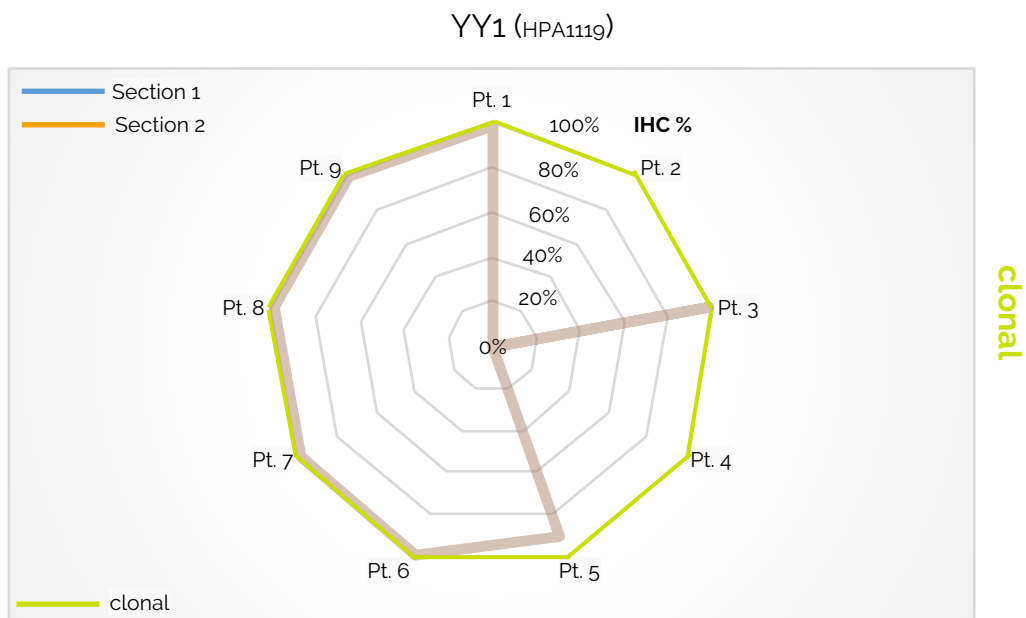


RI:65

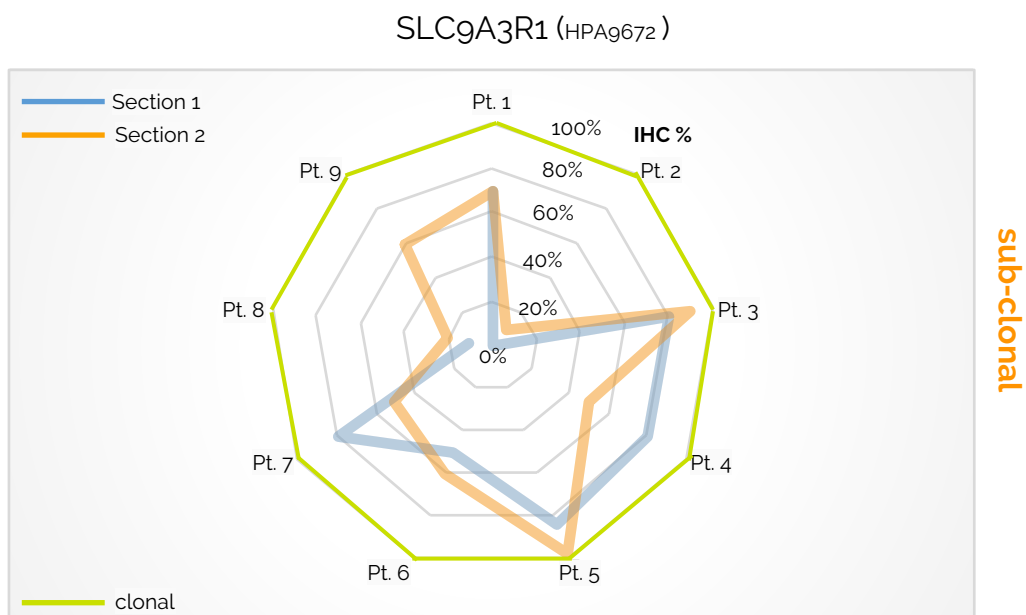
Glandular cells: <25%  
Intensity: Strong

# Figure S18

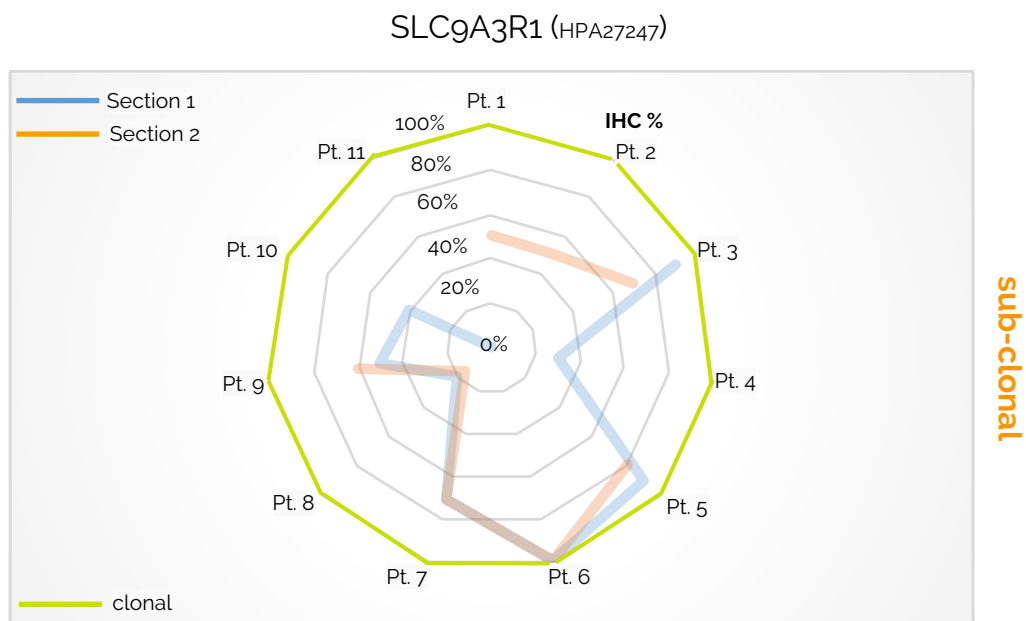
A

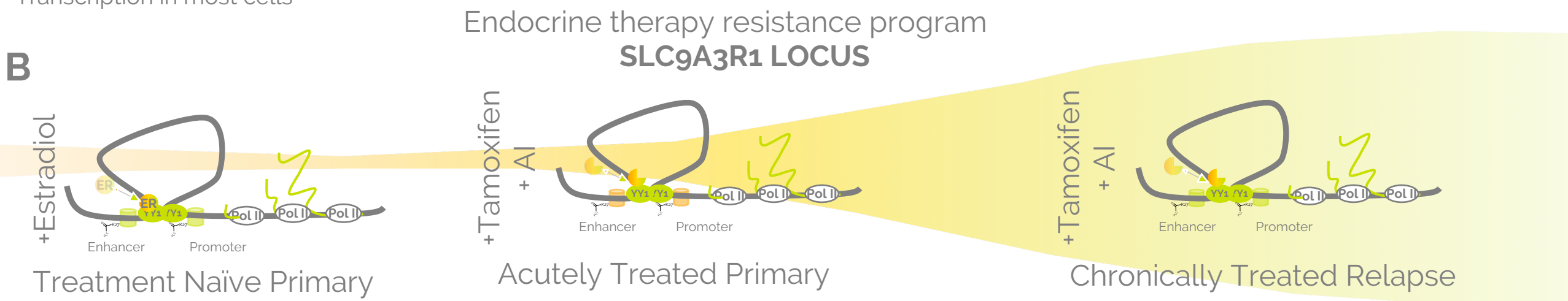
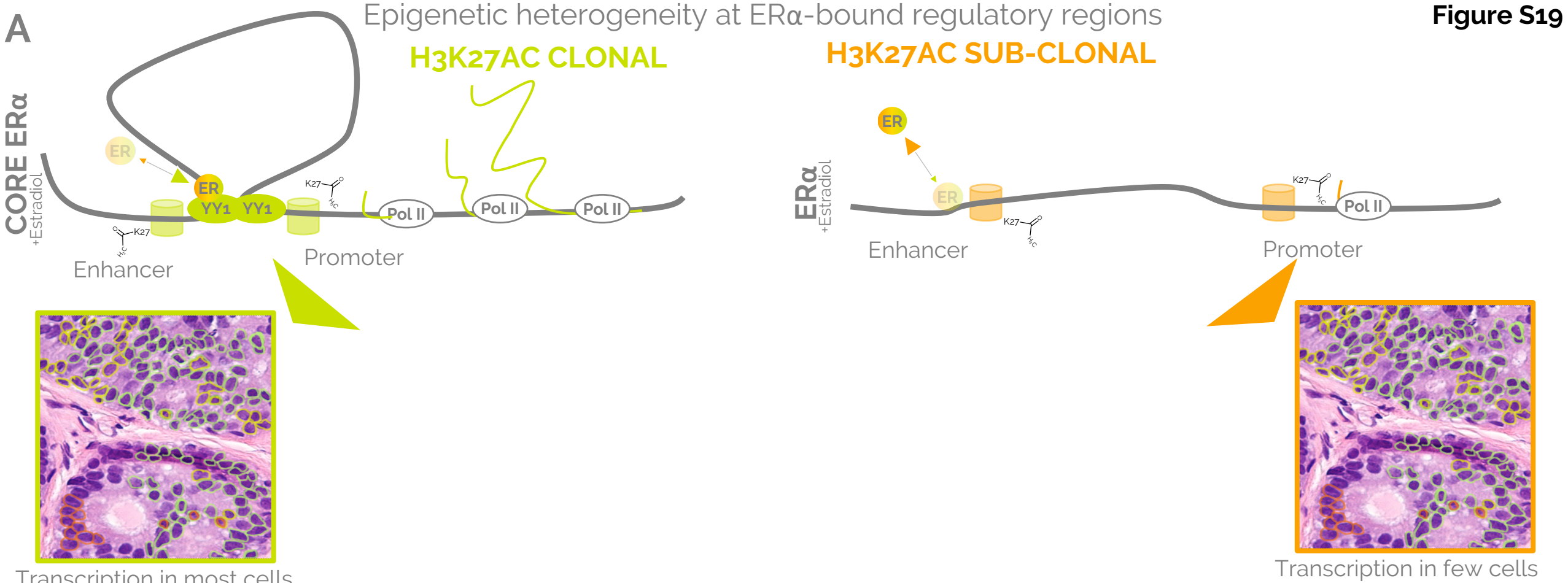


B



C





## Supplementary Computational Methods

### Targeted-Seq Cancer panel.

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 *bwa*<sup>1</sup> 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*<sup>2</sup> 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 (*GATK 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*<sup>3</sup> 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*<sup>4</sup> were filtered out.

**ChIP-Seq data processing.** Reads were quality controlled with *FastQC v0.11.5* and aligned to the human hg38 reference using *bowtie v1.1.2*<sup>5</sup> 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 v2.1.1<sup>6</sup> (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.

**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<sup>6</sup> and 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)<sup>7</sup>.

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 SaturationPlotCreator<sup>8</sup> 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<sup>9</sup> 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:  $Nscore = ((\text{peak read count} / \text{peak size}) \cdot 10^6) \cdot 10^3 / \text{total mapped reads (FPKM)}$ .

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<sup>10</sup> and 37 tissues from the Epigenomic Roadmap<sup>11</sup>, 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 v0.2.3<sup>12</sup>. 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  $\leq 0.05$ ) 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 library<sup>13</sup> 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)<sup>14</sup> 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  $\geq 0.8$  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  $\geq 21$ . 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 Wellington<sup>14,15</sup> 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 than the single narrow peaks defined by MACS, and allow the identification of all the NFRs. The number of footprints within each active regulatory region was calculated, and then normalized by the region size. The RI for each promoter and enhancer in MCF7 calls was calculated and plotted 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 QDNAseq<sup>16</sup> and CNVkit tools.<sup>17</sup> QDNAseq data processing involves 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<sup>18</sup> with 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 $\alpha$  (*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 $\alpha$  binding set (all ER $\alpha$  in all patients). For YY1, FOXA1 and ER $\alpha$  in MCF7, intersections were calculated using Cistrome<sup>19</sup>. YY1 BED files were defined as the consensus narrow peaks of two biological replicates. FOXA1 ChIP-seq data and ER $\alpha$  were obtained in house<sup>20</sup>. The core ER $\alpha$  BED file was obtained by lifting a published dataset<sup>21</sup> to hg19 coordinates. The private ER $\alpha$  BED file was obtained by iterative processing of the ER $\alpha$  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 $\alpha$  (second dataset). The TCGA luminal signature was obtained from<sup>22</sup>. 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<sup>23</sup>. 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.

**CRISPR/Cas9 Enhancer Knockout.** Four sgRNAs were designed using the CRISPR-DO software<sup>24</sup>, two at either end of the putative YY1 regulating Enhancer A and cloned into a gRNA expression vector (Church's lab, Addgene plasmid # 41824) using the Gibson Assembly Kit (NEB). Properly constructed plasmids were confirmed through Sanger sequencing. All gRNA vectors were simultaneously co-transfected with a pCas9-GFP plasmid (Musunuru's lab, Addgene plasmid # 44719) at a 1:1, gRNA to Cas9-GFP ratio into MCF7 cells using the 4D-Nucleofector system and Amaxa Cell Line Kit V (Lonza). 48 hours after transfection cells were sorted for GFP

expression using flow cytometry (Imperial Medical Research Council Flow Cytometry Facility). Sorted cells were plated at low density in 15 cm dishes to allow growth before full isolation using cloning discs (Sigma-Aldrich). Isolated clones were screened for successful enhancer knockout through PCR amplification and Sanger sequencing.

sgRNA and PCR primes used are shown in the table

YY1 Enhancer A gRNAs	Full primer	Target sequence	Loci
Upstream gRNA1	TTTCTTGGCTTTATATATCTTGTGGAAGGACGAAACA CCGctgctgcggggctcacgc	ctggtcgcggggctcacgccg	chr14:100680137-100680166
	GACTAGCCTTATTTAACTTGCTATTTCTAGCTCTAAA ACgctgagccccgcgaccagC		
Upstream gRNA2	TTTCTTGGCTTTATATATCTTGTGGAAGGACGAAACA CCGaaatagtggctggtcgcg	aaatagtggctggtcgcggg	chr14:100680147-100680176
	GACTAGCCTTATTTAACTTGCTATTTCTAGCTCTAAA ACcgcgaccagccaactatttC		
Downstream gRNA3	TTTCTTGGCTTTATATATCTTGTGGAAGGACGAAACA CCGgaccagaccctcaccgg	gaccagaccacctcaccggtg	chr14:100682121-100682150
	GACTAGCCTTATTTAACTTGCTATTTCTAGCTCTAAA ACccggtgaggtggtctggtcC		
Downstream gRNA4	TTTCTTGGCTTTATATATCTTGTGGAAGGACGAAACA CCGtgtatattaaactcacgg	tgtatattaaactcacggagg	chr14:100682225-100682254
	GACTAGCCTTATTTAACTTGCTATTTCTAGCTCTAAA ACccgtgagtttaataatacaC		
	<b>YY1 Enhancer PCR amplification primers</b>		
Forward	TTTTCTCTCTTTCCTTCTGCAA		
Reverse	CCTGAGAGAAACAGGCTTGA		
	<b>YY1 Enhancer sequencing primers</b>		
Forward	GCTCACTGCAGCCTTGACTT		
Reverse	TATCATTGCCTCACCGAACC		

## References

1. Li & Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 1754–1760 (2009). 25,
2. McKenna et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 1297–1303 (2010). 20,
3. Forbes et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research* D777–D783 (2017). 45,
4. Shihab et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* 57–65 (2013). 34,
5. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *R25* 10,
6. Quinlan & Hall. BEDTools: a flexible suite of utilities for comparing genomic features. 841–842 26,
7. Tyner et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research* D626–D634 (2017). 45,
8. Jee et al. ACT: aggregation and correlation toolbox for analyses of genome tracks. *Bioinformatics* 1152–1154 (2011). 27,
9. Ritchie et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* e47–e47 (2015). 43,
10. Consortium, T. et al. An integrated encyclopedia of DNA elements in the human genome. 57–74 488,
11. Bernstein, B. et al. The NIH Roadmap Epigenomics Mapping Consortium. 1045–1048 28,
12. Zhao et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 1006–1007 (2014). 30,
13. Valero-Mora. ggplot2:Elegant Graphics for Data Analysis. *Journal of Statistical Software* (2010). 35,
14. Ahmed et al. Variant Set Enrichment: an R package to identify disease-associated functional genomic regions. *BioData Mining* 9 (2017). 10,
15. Piper, J. et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. e201–e201 41,

16. Scheinin et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Research* 2022–2032 (2014). 24,
17. Talevich, Shain, Botton & Bastian. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology* e1004873 (2016). 12,
18. Rice, Longden & Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 276–277 (2000). 16,
19. Liu, T. et al. Cistrome: an integrative platform for transcriptional regulation studies. *R83* 12,
20. Nguyen, V. et al. Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. *Nature Communications* 10044 6,
21. Ross-Innes, C. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. 389–393 481,
22. Koboldt, D. et al. Comprehensive molecular portraits of human breast tumours. 61–70 490,
23. Hurtado, A., Holmes, K., Ross-Innes, C., Schmidt, D. & Carroll, J. FOXA1 is a key determinant of estrogen receptor function and endocrine response. 27–33 43,
24. Ma et al. CRISPR-DO for genome-wide CRISPR design and optimization. *Bioinformatics* 3336–3338 (2016). 32,