# TITLE PAGE

Full Title: Epigenetic misactivation of a distal developmental enhancer cluster drives *SOX2* overexpression in breast and lung cancer

Author List and Affiliations

Luis E. Abatti[1], Patricia Lado-Fernández[2,3], Linh Huynh[4], Manuel Collado[2], Michael M. Hoffman[4,5,6,7], and Jennifer A. Mitchell[1]

[1]Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada

[2]Laboratory of Cell Senescence, Cancer and Aging, Health Research Institute of Santiago de Compostela (IDIS), Xerencia de Xestión Integrada de Santiago (XXIS/SERGAS), Santiago de Compostela, Spain

[3]Department of Physiology and Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain

[4]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

[5]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

[6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

[7]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada

Corresponding authors:

ja.mitchell@utoronto.ca, luis.abatti@mail.utoronto.ca

Current address:

Department of Cell and Systems Biology, University of Toronto, Toronto, Canada

## ABSTRACT

Enhancer reprogramming has been proposed as a major source of gene expression dysregulation during tumorigenesis. Here, we identify *SOX2* developmental enhancers that are misactivated in breast and lung carcinoma. Deletion of the SRR124–134 enhancer cluster disrupts *SOX2* transcription and genome-wide chromatin accessibility in cancer cells. ATAC- and RNA-seq analysis of primary tumors shows that chromatin accessibility at this cluster is correlated with *SOX2* overexpression in breast and lung cancer. We further identify FOXA1 as an activator and NFIB as a repressor of SRR124–134 activity and *SOX2* transcription. Notably, the conserved SRR124 and SRR134 regions are essential during mouse development, where homozygous deletion results in the lethal failure of esophageal-tracheal separation. Our findings indicate that the SRR124–134 enhancer cluster drives *SOX2* expression during development. In breast and lung cancer, FOXA1-induced aberrant activity of the SRR124–134 cluster drives *SOX2* overexpression, demonstrating how developmental enhancers can be recommissioned during tumorigenesis. These results highlight the importance of understanding enhancer dynamics during development and disease while also providing new opportunities for therapeutic intervention by targeting aberrantly activated developmental enhancers.

## INTRODUCTION

Multicellular organisms have a deeply organized hierarchy of distinct cell lineages originally derived from the same embryonic progenitor. The differentiation of early embryonic cells involves the activation of specific transcriptional programs that drive their commitment toward distinct cell types. This process is mediated by key developmental-associated transcription factors [reviewed in 1], which interact genome-wide with cis-regulatory regions and are responsible for progressively restricting the epigenome, repressing regulatory regions associated with pluripotency [2,3] and activating enhancers that control the expression of lineage-specific genes [4–6]. This establishes an epigenetic regulatory "memory" that maintains cells in their own lineage compartment, reinforcing their transcriptional programs and repressing previous uncommitted states [7]. This epigenetic landscape, however, becomes profoundly disturbed during tumorigenesis [8–10], causing cells to lose their identity and assume a dysfunctional state that combines regulatory features from other cell lineages in a mechanism known as "enhancer reprogramming" [7–9,11]. Although this has been proposed as one of the sources of transcriptional dysregulation, it remains unclear if early developmental enhancers that were decommissioned during lineage differentiation are reactivated in the disease state.

SRY-box transcription factor 2 (SOX2) is a pioneer transcription factor required for pluripotency maintenance in embryonic stem cells [12,13] and reprogramming to induced pluripotent stem cells in mammals [14–16]. Two proximal enhancers were once deemed crucial for driving *Sox2* expression during early development: *Sox2* Regulatory Region 1 (SRR1) and SRR2 [17–19]. Deletion of SRR1 and SRR2, however, has no effect on *Sox2*

expression in mouse embryonic stem cells [20]. In contrast, deletion of a distal *Sox2* Control Region (SCR), 106 kb downstream of the *Sox2* promoter, causes a profound loss of *Sox2* expression in mouse embryonic stem cells [20,21] and in blastocysts, where SCR deletion causes peri-implantation lethality [22]. However, the contribution of these regulatory regions in driving *SOX2* expression in other contexts remains poorly understood.

SOX2 is also involved in tissue morphogenesis and homeostasis of the brain [23], eyes [24], esophagus [25], inner ear [26], lungs [27], skin [28], stomach [29], taste buds [30] and trachea [31] in both human and mouse. In these tissues, *SOX2* expression is regulated precisely in space and time at critical stages of development. For example, proper levels of *Sox2* expression are required for the complete separation of the anterior foregut into the esophagus and trachea in mice [22,25,32] and in humans [33,34], as the disruption of *Sox2* expression leads to an abnormal developmental condition known as esophageal atresia with distal tracheoesophageal fistula (EA/TEF) reviewed in [35,36]. After the anterior foregut is properly separated, *Sox2* expression ranges from the esophagus to the stomach in the gut [25,29], and throughout the trachea, bronchi, and upper portion of the lungs in the developing airways [31]. Proper branching morphogenesis at the tip of the lungs, however, requires temporary downregulation of *Sox2*, followed by reactivation after lung bud establishment [27]. *Sox2* also retains an essential function in multiple mature epithelial tissues, where it is highly expressed in proliferative and self-renewing adult stem cells necessary for maintaining and replacing terminally differentiated cells within the epithelium of the brain, bronchi, esophagus, stomach, and trachea [29,31,37,38].

Tumorigenesis of at least 25 different cancer types involves *SOX2* overexpression reviewed in [39]. This overexpression is linked to increased cellular replication rates, more aggressive tumor grades, and poor patient outcomes in breast carcinoma (BRCA) [40–44]; colon adenocarcinoma (COAD) [45–48]; glioblastoma (GBM) [49–52]; liver hepatocellular carcinoma (LIHC) [53]; lung adenocarcinoma (LUAD) [54–56]; and lung squamous cell carcinoma (LUSC) [57,58]. These clinical and molecular characteristics arise from the participation of *SOX2* in the formation and maintenance of tumor-initiating cells that resemble tissue progenitor cells, as evidenced by BRCA [44,59,60], GBM [51,61–63], LUAD [64], and LUSC [65] studies. *SOX2* knockdown, on the other hand, often results in diminished levels of cell replication, invasion, and treatment resistance in these cancer types [40,44,54,56,57,66–68]. Despite the involvement of *SOX2* in the progression of multiple cancer types, little is known about the mechanisms that cause *SOX2* overexpression in cancer.

Here, we identify a novel enhancer cluster misactivated in breast and lung cancer. This cluster contains two regions, located 124 and 134 kb downstream of the *SOX2* promoter and referred to as SRR124–134, that drive transcription in BRCA, LUAD, and LUSC. Deletion of this cluster results in significant *SOX2* downregulation, leading to genome-wide changes in chromatin accessibility and a globally disrupted transcriptome. The SRR124–134 cluster is highly accessible in most breast and lung patient tumors, where chromatin accessibility at these regions is correlated with *SOX2* overexpression and is regulated positively by FOXA1 and negatively by NFIB. Finally, we found that both SRR124 and SRR134 are highly conserved in the mouse and are essential for postnatal survival, as homozygous deletion of their homologous regions results in lethal EA/TEF. These findings serve as a prime example of how cancer cells activate

113 enhancers that were decommissioned during development to drive the expression of
114 developmentally associated transcription factors during tumorigenesis.

## RESULTS

### Two regions downstream of *SOX2* gain enhancer features in cancer cells

*SOX2* overexpression occurs in multiple types of cancer [reviewed in 39]. To examine which cancer types have the highest levels of *SOX2* upregulation, we performed differential expression analysis by calculating the $\log_2$ fold change ($\log_2$ FC) of *SOX2* transcription from 21 TCGA primary solid tumors (see Supplementary Table S1 for cancer type abbreviations) compared to normal tissue samples [69]. We found that BRCA ($\log_2$ FC = 3.31), COAD ($\log_2$ FC = 1.38), GBM ($\log_2$ FC = 2.05), LIHC ($\log_2$ FC = 3.22), LUAD ($\log_2$ FC = 1.36), and LUSC ($\log_2$ FC = 4.91) tumors had the greatest *SOX2* upregulation ($\log_2$ FC > 1; FDR-adjusted $Q < 0.01$; Figure 1A, Supplementary Table S2). As a negative control, we ran this same analysis using the housekeeping gene *PUM1* [70] and found no cancer types with significant upregulation of this gene (Supplementary Figure S1A, Supplementary Table S3).

Next, we divided BRCA, COAD, GBM, LIHC, LUAD, and LUSC patients (n = 3,064) into four groups according to their *SOX2* expression. Gene expression levels were measured by RNA-seq counts normalized by library size and transformed to a $\log_2$ scale, hereinafter referred to as $\log_2$ counts. Cancer patients within the top group (25% highest *SOX2* expression; $\log_2$ counts > 10.06) have a significantly ($P = 1.27\times10^{-23}$, log-rank test) lower overall probability of survival compared to cancer patients within the bottom group (25% lowest *SOX2* expression; $\log_2$ counts < 1.68) (Supplementary Figure S1B, Supplementary Table S4). We also examined the relationship between *SOX2* copy number and *SOX2* overexpression within these six tumor types. Although previous studies have shown that *SOX2* is frequently amplified in squamous cell

carcinoma [57,58,71,72], we found that most BRCA (88%), COAD (98%), GBM (91%), LIHC (94%), and LUAD (92%) tumors were diploid for *SOX2*. In addition, BRCA ($P$ = 0.011, Holm-adjusted Dunn's test), GBM ($P = 1.18\times10^{-3}$), LIHC ($P$ = 0.016), LUAD ($P$ = 0.012), and LUSC ($P = 2.72\times10^{-11}$) diploid tumors significantly overexpressed *SOX2* compared to normal tissue (Figure 1B, Supplementary Table S5). This indicates that gene amplification is dispensable for driving *SOX2* overexpression in most cancer types.

We investigated whether the *SOX2* locus gains epigenetic features associated with active enhancers in cancer. Enhancer features commonly include accessible chromatin determined by either Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq) [73] or DNase I hypersensitive sites sequencing (DNase-seq) [74], and histone modifications including histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac) [75,76]. To study gains in enhancer features within the *SOX2* locus, we initially focused our analyses on luminal A breast cancer, the most common subtype of BRCA to significantly ($P$ = 0.021, Tukey's test) overexpress *SOX2* (Supplementary Figure S1C) [69]. MCF-7 cells are a widely used $ER^+/PR^+/HER2^-$ luminal A breast adenocarcinoma model [77], which have been previously described to overexpress *SOX2* [40,68,78,79]. After confirming that *SOX2* is one of the most upregulated genes in MCF-7 cells ($\log_2$ FC = 10.75; FDR-adjusted $Q = 2.20\times10^{-36}$; Supplementary Figure S1D, Supplementary Table S6) compared to healthy breast epithelium [80], we contrasted their chromatin accessibility and histone modifications [81]. By intersecting 1,500 bp regions that contain at least 500 bp overlap between H3K27ac and ATAC-seq peaks, we found that 19 putative enhancers gained ($\log_2$ FC > 1) both these features within ± 1 Mb from the

*SOX2* transcription start site (TSS) in MCF-7 cells (Figure 1C, Supplementary Table S7). Besides the *SOX2* promoter (pSOX2), we identified a downstream cluster containing two regions that have gained the highest ATAC-seq and H3K27ac signal in MCF-7 cells: SRR124 (124 kb downstream of pSOX2) and SRR134 (134 kb downstream of pSOX2). The previously described SRR1, SRR2 [17–19], and the human ortholog of the mouse SCR (hSCR) [20,21], however, lacked substantial gains in enhancer features in MCF-7 cells.

Alongside gains in chromatin features, another characteristic of active enhancers is the binding of numerous (> 10) transcription factors [82–84]. Chromatin Immunoprecipitation Sequencing (ChIP-seq) data from ENCODE [81] on 117 transcription factors revealed 48 different factors present at the SRR124–134 cluster in MCF-7 cells, with the majority (47) of these factors present at SRR134 (Figure 1D). Transcription factors bound at both SRR124 and SRR134 include CEBPB, CREB1, FOXA1, FOXM1, NFIB, NR2F2, TCF12, and ZNF217. An additional feature of distal enhancers is that they contact their target genes through long-range chromatin interactions [85,86]. Chromatin Interaction Analysis by Paired-End-Tag sequencing (ChIA-PET) [87] showed two interesting RNA Polymerase II (RNAPII)-mediated chromatin interactions in MCF-7 cells: one between the *SOX2* gene and SRR134, and one between SRR124 and SRR134 (Figure 1E). Beyond MCF-7 cells, we found that H520 (LUSC), PC-9 (LUAD), and T47D (luminal A BRCA) cancer cell lines, which display varying levels of *SOX2* expression (Supplementary Figure S1E), also gained substantial enhancer features at SRR124 and SRR134 when compared to healthy tissue (Figure 1E) [88–90]. Together,

183 these data suggest SRR124 and SRR134 could be active enhancers driving *SOX2*
184 transcription in BRCA, LUAD, and LUSC.
185
186 **Figure 1: A cluster 124–134 kilobases downstream of *SOX2* gains enhancer**
187 **features in cancer cells. (A)** Super-logarithmic RNA-seq volcano plot of *SOX2*
188 expression from 21 cancer types compared to normal tissue [69]. Cancer types with $\log_2$
189 FC > 1 and FDR-adjusted $Q < 0.01$ were considered to significantly overexpress *SOX2*.
190 Error bars: standard error. **(B)** *SOX2* $\log_2$-normalized expression ($\log_2$ counts)
191 associated with the *SOX2* copy number from BRCA (n = 1174), COAD (n = 483), GBM
192 (n = 155), LIHC (n = 414), LUAD (n = 552), and LUSC (n = 546) patient tumors [69]. RNA-
193 seq reads were normalized to library size using DESeq2 [91]. Error bars: standard
194 deviation. Significance analysis by Dunn's test [92] with Holm correction [93]. **(C)** 1,500 bp
195 genomic regions within ± 1 Mb from the *SOX2* transcription start site (TSS) that gained
196 enhancer features in MCF-7 cells [81] compared to healthy breast epithelium [80]. Regions
197 that gained both ATAC-seq and H3K27ac ChIP-seq signal above our threshold ($\log_2$ FC
198 > 1, dashed line) are highlighted in pink. Each region was labelled according to their
199 distance in kilobases (kb) to the *SOX2* promoter (pSOX2, bolded). **(D)** ChIP-seq signal
200 for H3K4me1 and H3K27ac, ATAC-seq signal, and transcription factor ChIP-seq peaks
201 at the SRR124–134 cluster in MCF-7 cells. Datasets from ENCODE [81]**. (E)** UCSC
202 Genome Browser [94] display of H3K4me1 and H3K27ac ChIP-seq signal, DNAse-seq
203 and ATAC-seq chromatin accessibility signal, and ChIA-PET RNA Polymerase II
204 (RNAPII) interactions around the *SOX2* gene within breast (normal tissue and 2 BRCA
205 cancer cell lines) and lung (normal tissue, one LUAD, and one LUSC cancer cell lines)

samples [81,88–90]. Relevant RNAPII interactions (between SRR124 and SRR134, and between SRR134 and pSOX2) are highlighted in maroon.

## The SRR124–134 cluster is essential for *SOX2* expression in BRCA and LUAD cells

To assess SRR124 and SRR134 enhancer activity alongside the embryonic-associated SRR1, SRR2, and hSCR regions, we used a reporter vector containing the firefly luciferase gene under the control of a minimal promoter (minP, pGL4.23). We transfected each enhancer construct into the BRCA (MCF-7, T47D), LUAD (PC-9), and LUSC (H520) cell lines and measured luciferase activity as a relative fold change (FC) compared to the empty minP vector. SRR134 demonstrated the strongest enhancer activity, with the MCF-7 (FC = 6.42; $P < 2\times10^{-16}$, Dunnett's test), T47D (FC = 3.36; $P = 9.34\times10^{-10}$), H520 (FC = 2.37; $P = 1.22\times10^{-6}$), and PC-9 (FC = 2.03; $P = 9.79\times10^{-5}$) cell lines displaying a significant increase in luciferase activity compared to minP (Figure 2A). SRR124 also showed a modest, significant increase in luciferase activity compared to minP in the MCF-7 (FC = 1.53; $P = 4.27\times10^{-2}$), T47D (FC = 1.80; $P = 4.57\times10^{-2}$), and PC-9 (FC = 1.60; $P = 4.27\times10^{-2}$) cell lines. The embryonic-associated enhancers SRR1, SRR2, and hSCR, however, showed no significant enhancer activity ($P > 0.05$) in all four cell lines.

Although reporter assays can be used to assess enhancer activity, enhancer knockout approaches remain the current gold standard method for enhancer validation [95,96]. To investigate whether the SRR124–134 cluster drives *SOX2* expression in cancer cells, we used CRISPR/Cas9 to delete this cluster from the H520, MCF-7, PC-9, and

T47D cell lines. RT-qPCR showed that homozygous SRR124–134 deletion ($\Delta ENH^{-/-}$) causes a profound (> 99.5%) and significant ($P < 0.001$, Dunnett's test) loss of *SOX2* expression in both the MCF-7 and PC-9 cell lines (Figure 2B). Immunoblot analysis confirmed the depletion of the SOX2 protein in $\Delta ENH^{-/-}$ MCF-7 cells (Supplementary Figure S2A). Heterozygous SRR124–134 deletion ($\Delta ENH^{+/-}$) also significantly ($P < 0.001$) reduced *SOX2* expression by ~60% in both MCF-7 and PC-9 cells (Figure 2B). Although we were unable to isolate a homozygous deletion clone from T47D cells, multiple independent heterozygous $\Delta ENH^{+/-}$ T47D clonal isolates showed a significant downregulation (>50%; $P < 0.001$) in *SOX2* expression (Figure 2C). Interestingly, we did not find a significant ($P > 0.05$) impact on *SOX2* expression in $\Delta ENH^{+/-}$ or $\Delta ENH^{-/-}$ H520 cells (Supplementary Figure S2B), which indicates that *SOX2* transcription is sustained by a different mechanism in these cells. To assess the impact of the loss of *SOX2* expression in the tumor initiation capacity of enhancer-deleted cells, we performed a colony formation assay with MCF-7 and PC-9 $\Delta ENH^{-/-}$ cells. We found that both MCF-7 ($P = 3.53\times10^{-4}$, t-test) and PC-9 ($P = 1.26\times10^{-5}$) $\Delta ENH^{-/-}$ cells showed a significant decrease (> 50%) in their ability to form colonies compared to WT cells (Figure 2D), further suggesting that SRR124–134-driven *SOX2* overexpression is required to sustain high tumor initiation capacity in BRCA and LUAD.

Next, we performed total RNA sequencing (RNA-seq) to measure changes in the transcriptome of $\Delta ENH^{-/-}$ MCF-7 cells compared to WT MCF-7 cells. As expected, all three replicates of each genotype clustered together (Supplementary Figure S2C). In addition to *SOX2* downregulation (Figure 2E), differential expression analysis showed a total of 529 genes differentially ($|\log_2 FC| > 1$; FDR-adjusted $Q < 0.01$) expressed in

ΔENH$^{-/-}$ MCF-7 cells (Figure 2F, Supplementary Table S8). From these, 312 genes significantly lost expression (59%), whereas 217 (41%) genes significantly gained expression in ΔENH$^{-/-}$ MCF-7 cells compared to WT MCF-7 cells (Supplementary Figure S2D). *SOX2* was the gene with the highest loss in expression ($log_2$ FC = -10.24; $Q = 1.23\times10^{-43}$) in ΔENH$^{-/-}$ MCF-7 cells, followed by *CT83* ($log_2$ FC = -8.43; $Q = 1.07\times10^{-8}$), and *GUCY1A1* ($log_2$ FC = -6.96; $Q = 5.09\times10^{-15}$). On the other hand, genes with the most significant gain in expression within ΔENH$^{-/-}$ MCF-7 cells included the protocadherins *PCDH7* ($log_2$ FC = 5.34; $Q < 1\times10^{-200}$), *PCDH10* ($log_2$ FC = 5.29; $Q < 1\times10^{-200}$), and *PCDH11X* ($log_2$ FC = 4.73; Q = $9.29\times10^{-110}$). In addition, deletion of the SRR124–134 cluster reduced *SOX2* expression back to the levels found in healthy breast epithelium (*P* = 0.48, Tukey's test) [80,81] (Figure 2G). Together, these data confirm that the SRR124–134 cluster drives *SOX2* overexpression in BRCA and LUAD.

**Figure 2: The SRR124–134 cluster drives *SOX2* overexpression in MCF-7, T47D, and PC-9 cells. (A)** Enhancer reporter assay comparing luciferase activity driven by the SRR1, SRR2, SRR124, SRR134, and hSCR regions to an empty vector containing only a minimal promoter (minP). Enhancer constructs were assayed in the BRCA (MCF-7, T47D), LUAD (PC-9), and LUSC (H520) cell lines. Dashed line: average activity of minP. Error bars: standard deviation. Significance analysis by Dunnett's test (n = 5; * *P* < 0.05, *** *P* < 0.001, ns: not significant) [97]. **(B)** RT-qPCR analysis of *SOX2* transcript levels in SRR124–134 heterozygous- (ΔENH$^{+/-}$) and homozygous- (ΔENH$^{-/-}$) deleted MCF-7 (BRCA) and PC-9 (LUAD) clones compared to WT cells. Error bars: standard deviation. Significance analysis by Dunnett's test (n = 3; *** *P* < 0.001). **(C)** RT-qPCR

analysis of *SOX2* transcript levels in three independent SRR124–134 heterozygous-deleted ($\Delta ENH^{+/-}$) T47D clonal isolates compared to WT cells. Error bars: standard deviation. Significance analysis by Dunnett's test (n = 4; *** $P < 0.001$). **(D)** Crystal violet absorbance (570 nm) from a colony formation assay with WT and $\Delta ENH^{-/-}$ MCF-7 and PC-9 cells. Total absorbance was normalized to the average absorbance from WT cells within each cell line. Significance analysis by t-test with Holm correction (n = 5; *** $P < 0.001$). **(E)** UCSC Genome Browser [94] view of the SRR124–134 cluster deletion in $\Delta ENH^{-/-}$ MCF-7 cells with RNA-seq tracks from normal breast epithelium [80], WT and $\Delta ENH^{-/-}$ MCF-7 cells. Arrow: reduction in RNA-seq signal at the *SOX2* gene in $\Delta ENH^{-/-}$ MCF-7 cells. **(F)** Volcano plot with DESeq2 [91] differential expression analysis between $\Delta ENH^{-/-}$ and WT MCF-7 cells. Blue: 312 genes that significantly lost expression ($\log_2$ FC < -1; FDR-adjusted $Q < 0.01$) in $\Delta ENH^{-/-}$ MCF-7 cells. Pink: 217 genes that significantly gained expression ($\log_2$ FC > 1; $Q < 0.01$) in $\Delta ENH^{-/-}$ MCF-7 cells. Grey: 35,891 genes that maintained similar ($-1 \leq \log_2 FC \leq 1$) expression between $\Delta ENH^{-/-}$ and WT MCF-7 cells. **(G)** Comparison of *SOX2* transcript levels between WT MCF-7 and either $\Delta ENH^{-/-}$ MCF-7 or healthy breast epithelium cells [80], and between $\Delta ENH^{-/-}$ MCF-7 and healthy breast epithelium cells. RNA-seq reads were normalized to library size using DESeq2 [91]. Error bars: standard deviation. Significance analysis by Tukey's test (*** $P < 0.001$, ns: not significant) [98].

## SOX2 regulates pathways associated with epithelium development in luminal A BRCA

297 Because SOX2 regulates cell proliferation and differentiation pathways in other
298 epithelial cells [38,99], we decided to further investigate the molecular function of SOX2 in
299 luminal A BRCA cells by utilizing our *SOX2*-depleted ΔENH$^{-/-}$ MCF-7 cell model. Gene
300 Set Enrichment Analysis (GSEA) showed a significant (FDR-adjusted $Q$ < 0.05)
301 depletion of multiple epithelium-associated processes within the transcriptome of
302 ΔENH$^{-/-}$ MCF-7 cells, as indicated by normalized enrichment score [NES] < 1
303 (Supplementary Table S9). These processes included epidermis development (NES = -
304 1.93; $Q$ = 0.001; Figure 3A), epithelial cell differentiation (NES = -1.67; $Q$ = 0.007;
305 Figure 3B), and cornification (NES = -2.11; $Q$ = 0.006; Figure 3C). This suggests that
306 SOX2 regulates epithelial development and differentiation in luminal A BRCA cells.
307 SOX2 is a pioneer transcription factor that associates with its motif in
308 heterochromatin [100] and recruits chromatin-modifying complexes [101] in embryonic and
309 reprogrammed stem cells. We performed ATAC-seq in ΔENH$^{-/-}$ MCF-7 cells and
310 compared chromatin accessibility to WT MCF-7 cells to identify genome-wide loci that
311 are dependent on SOX2 to remain accessible in luminal A BRCA. As expected, the
312 ATAC-seq signal from all replicates was highly enriched around gene TSS
313 (Supplementary Figure S3A), with both WT and ΔENH$^{-/-}$ samples having higher
314 chromatin accessibility at the TSS of highly expressed genes (Supplementary Figure
315 S3B and Supplementary Figure S3C). Correlation analysis also confirmed the clustering
316 of all three replicates from each genotype (Supplementary Figure S3D). Together with
317 the SRR124–134 cluster and pSOX2 (Figure 3D), a total of 3,076 500-bp regions had
318 significant ($|\log_2 FC| > 1$; FDR-adjusted $Q$ < 0.01) changes in chromatin accessibility in
319 ΔENH$^{-/-}$ compared to WT MCF-7 cells (Figure 3E, Supplementary Table S10). Most

regions (86%, 2,636 regions) significantly lost chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells and 76% (2,024 regions) of these regions also gained chromatin accessibility in WT MCF-7 compared to healthy breast epithelium [80] (Supplementary Table S11). Together, this indicates that SOX2 has an important role in regulating the chromatin accessibility changes acquired in luminal A BRCA.

We used TOBIAS [102] to analyze changes in transcription factor footprints within ATAC-seq peaks in ΔENH$^{-/-}$ compared to WT MCF-7 cells. From 841 vertebrate motifs [103], we found a total of 281 motifs with a significant ($|\log_2 FC| > 0.1$; FDR-adjusted $Q < 0.01$) differential binding score (Figure 3F, Supplementary Table S12). Most of these motifs (97%, 272 motifs) were underrepresented within ATAC-seq peaks in ΔENH$^{-/-}$ compared to WT MCF-7 cells, indicating that reduced *SOX2* expression affects the binding of multiple other transcription factors. Among them, the GRHL1 ($\log_2 FC = -0.519$; $Q = 3\times10^{-179}$), TFCP2 ($\log_2 FC = -0.462$; $Q = 1.03\times10^{-172}$), RUNX2 ($\log_2 FC = -0.352$; $Q = 8.02\times10^{-164}$), GRHL2 ($\log_2 FC = -0.343$; $Q = 4.43\times10^{-174}$), TEAD3 ($\log_2 FC = -0.235$; $Q = 9.74\times10^{-155}$), and SOX4 ($\log_2 FC = -0.232$; $Q = 5.33\times10^{-167}$) motifs (Figure 3G) had the most reduced binding score in ΔENH$^{-/-}$ MCF-7 cells compared to WT MCF-7 cells. These factors belong to three main motif clusters: GRHL/TFCP (cluster 33; aaAACAGGTTtcAgtt), RUNX (cluster 60; ttctTGtGGTTttt), TEAD (cluster 2; tccAcATTCCAggcCTTta), and SOX (cluster 8; acggaACAATGgaagTGTT) [103]. The SOX cluster also included the SOX2 ($\log_2 FC = -0.175$; $Q = 6.61\times10^{-139}$) motif.

Next, we aimed to analyze ChIP-seq data from transcription factors within these motif clusters in MCF-7 cells. We utilized two published datasets: GRHL2 [89] and RUNX2 [104]. Regions that lost ($\log_2 FC < -1$; $Q < 0.01$) chromatin accessibility in ΔENH$^{-/-}$ MCF-7

cells significantly overlapped with regions with binding of either of these transcription factors ($P < 2\times10^{-16}$, hypergeometric test). Among the 2,636 regions that lost chromatin accessibility, 40% (750 regions) also show GRHL2 binding (Supplementary Figure S3E), whereas 21% (552 regions) share RUNX2 binding (Supplementary Figure S3F). We found multiple SOX motifs significantly (FDR-adjusted $Q < 0.001$) enriched within peaks from both GRHL2 (Supplementary Table S13) and RUNX2 (Supplementary Table S14) ChIP-seq data, further suggesting that SOX2 collaborates with GRHL2 and RUNX2 to maintain chromatin accessibility in luminal A BRCA. Expression levels of either *GRHL2* or *RUNX2*, however, were not significantly affected by *SOX2* downregulation in ΔENH$^{-/-}$ MCF-7 cells ($-1 \leq \log_2 FC \leq 1$; Supplementary Table S8), indicating that they are not directly regulated by SOX2 at the transcriptional level but may interact at the protein level.

**Figure 3: *SOX2* downregulation impacts chromatin accessibility in luminal A BRCA. (A – C)** Gene Set Enrichment Analysis (GSEA) in the transcriptome of ΔENH$^{-/-}$ compared to WT MCF-7 cells. Genes were ranked according to their change in expression ($\log_2$ FC). A subset of GO terms significantly enriched among downregulated genes in ΔENH$^{-/-}$ MCF-7 cells are displayed, indicated by the normalized enrichment score (NES) < 1: **(A)** epidermis development, **(B)** epithelial cell differentiation, and **(C)** cornification. GSEA was performed using clusterProfiler [105] with an FDR-adjusted $Q < 0.05$ threshold. Green line: running enrichment score. **(D)** UCSC Genome Browser [94] view of the SRR124–134 deletion in ΔENH$^{-/-}$ MCF-7 cells with ATAC-seq tracks from breast epithelium [80], WT, and ΔENH$^{-/-}$ MCF-7 cells. **(E)** Volcano

plot with differential ATAC-seq analysis of ΔENH$^{-/-}$ MCF-7 cells compared to WT. Blue: 2,638 regions that lost ($log_2$ FC < -1; FDR-adjusted $Q$ < 0.01) chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells. Pink: 440 regions that gained ($log_2$ FC > 1; $Q$ < 0.01) chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells. Grey: 132,726 regions that retained chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells (-1 ≤ $log_2$ FC ≤ 1). Regions were labelled with their closest gene within a ± 1 Mb distance threshold. Differential chromatin accessibility analysis was performed using diffBind [106]. **(F)** Volcano plot with ATAC-seq footprint analysis of differential transcription factor binding in ΔENH$^{-/-}$ MCF-7 cells compared to WT. Blue: 272 underrepresented ($log_2$ FC < -0.1; FDR-adjusted $Q$ < 0.01) motifs in ATAC-seq peaks from ΔENH$^{-/-}$ MCF-7 cells. Pink: 9 overrepresented ($log_2$ FC > 0.1; $Q$ < 0.01) motifs in ATAC-seq peaks from ΔENH$^{-/-}$ MCF-7 cells. Grey: 560 motifs with no representative change (-0.1 ≤ $log_2$ FC ≤ 0.1) within ATAC-seq peaks from ΔENH$^{-/-}$ MCF-7 cells. **(G)** Sequence motifs of the top 6 transcription factors with the lowest binding score in ΔENH$^{-/-}$ compared to WT MCF-7 cells: GRHL1, TFCP2, RUNX2, GRHL2, TEAD3, SOX4. Footprint analysis was performed using TOBIAS [102] utilizing the JASPAR 2022 motif database [103].

## The SRR124–134 cluster is associated with *SOX2* overexpression in primary tumors

With the confirmation that the SRR124–134 cluster drives *SOX2* overexpression in the BRCA and LUAD cell lines, we investigated chromatin accessibility at this enhancer cluster within primary tumors isolated from cancer patients. By analyzing the pan-cancer ATAC-seq dataset from TCGA [107], we found that SRR124 and SRR134 are most

accessible within LUSC, LUAD, BRCA, bladder carcinoma (BLCA), stomach adenocarcinoma (STAD), and uterine endometrial carcinoma (UCEC) patient tumors (Figure 4A). We also quantified the ATAC-seq signal at six other regions (genomic coordinates in Supplementary Table S15): the *SOX2* embryonic-associated enhancers (SRR1, SRR2, hSCR), pSOX2, a gene regulatory desert with no enhancer features located between the *SOX2* gene and the SRR124–134 cluster (desert), and the promoter of the housekeeping gene *RAB7A* (pRAB7A, positive control). We then compared the chromatin accessibility levels at each of these regions to the promoter of the repressed olfactory gene *OR5K1* (pOR5K1, negative control). Both SRR124 and SRR134 showed significantly increased ($P < 0.05$, Holm-adjusted Dunn's test) chromatin accessibility when compared to pOR5K1 in BLCA (SRR124 $P = 0.014$; SRR134 $P = 1.52\times10^{-3}$; Holm-adjusted Dunn's test), BRCA (SRR124 $P = 1.70\times10^{-20}$; SRR134 $P = 1.03\times10^{-16}$), LUAD (SRR124 $P = 6.76\times10^{-7}$; SRR134 $P = 3.26\times10^{-6}$), LUSC (SRR124 $P = 1.62\times10^{-6}$; SRR134 $P = 7.08\times10^{-4}$), STAD (SRR124 $P = 1.15\times10^{-4}$; SRR134 $P = 1.96\times10^{-7}$), and UCEC (SRR124 $P = 3.15\times10^{-5}$; SRR134 $P = 0.025$) patient tumors (Figure 4B).

One explanation for increased chromatin accessibility is locus amplification. While LUSC had high levels of chromatin accessibility likely related to previously described *SOX2* amplifications [57,58,71,72], most patient tumors showed no evidence of locus amplifications extending to the SRR124–134 cluster, as evidenced by the lack of significant ($P > 0.05$) accessibility at the intermediate desert region. In contrast, the SRR124–134 cluster displayed a consistent pattern of accessible chromatin across multiple cancer types: BLCA, BRCA, LUAD, LUSC, STAD, and UCEC (Figure 4C).

GBM and LGG tumors lacked accessible chromatin at this cluster but displayed increased chromatin accessibility at the SRR1 and SRR2 enhancers (Supplementary Figure S4A, Supplementary Table S16), which is consistent with the evidence that SRR1 and SRR2 drive *SOX2* expression in the neural lineage [17,19,108].

Next, we reasoned that an accessible SRR124–134 cluster drives subsequent *SOX2* transcription within patient tumors. If this is the case, we expect to find positive and significantly correlated chromatin accessibility between this enhancer cluster and pSOX2. Indeed, we found that the majority of BRCA (58%), LUAD (82%), and LUSC (69%) tumors have concurrent accessibility ($\log_2$ RPM > 0) at pSOX2, SRR124 and SRR134. Patient tumors also showed a significant correlation (Pearson, *R*) between accessible chromatin signal at pSOX2 and at both SRR124 and SRR134 in BRCA and LUAD (Figure 4D). LUSC tumors showed a significant correlation between accessible chromatin at pSOX2 and SRR124, but not at SRR134 (Figure 4D). As a negative control, we measured the correlation between chromatin accessibility at pSOX2 and at the *SOX2* desert region and found no significant ($P > 0.05$) correlation in any of these cancer types (Supplementary Figure S4B). We also conducted a similar analysis after segregating BRCA tumors into luminal A, luminal B, $HER2^+$, and basal-like subtypes. Interestingly, we found that both luminal A and luminal B tumors possess a significant ($P < 0.05$) correlation between enhancer accessibility and pSOX2 accessibility, whereas for $HER2^+$ tumors the correlation was weaker (Supplementary Figure S4C). Basal-like tumors, on the other hand, display no accessible chromatin at either SRR124 or SRR134. In summary, a luminal-like BRCA phenotype correlates with increased accessibility at the SRR124–134 cluster.

Finally, by separating BRCA, LUAD, and LUSC patient tumors according to their chromatin accessibility at SRR124 and SRR134, we found that tumors with the most accessible chromatin at each of these regions also significantly ($P < 0.05$, t-test) overexpress *SOX2* compared to tumors with low chromatin accessibility at these regions (Figure 4E, Supplementary Table S17). Together, these data are consistent with a model in which increased chromatin accessibility at the SRR124–134 cluster drives *SOX2* overexpression in BRCA, LUAD, and LUSC patient tumors.

**Figure 4: The SRR124–134 cluster is associated with *SOX2* overexpression in cancer patient tumors. (A)** ATAC-seq signal ($\log_2$ RPM) at SRR124 and SRR134 for 294 patient tumors from 14 cancer types [107]. Cancer types are sorted in descending order by the median signal between all three regions. Dashed line: regions with a sum of reads above our threshold ($\log_2$ RPM > 0) were considered "accessible". Error bars: standard deviation. Underscore: top 6 cancer types with the highest ATAC-seq median signal. **(B)** ATAC-seq signal ($\log_2$ RPM) at the *RAB7A* promoter (pRAB7A), *SOX2* promoter (pSOX2), SRR1, SRR2, SRR124, SRR134, hSCR, and a desert region within the *SOX2* locus (desert) compared to the background signal at the repressed *OR5K1* promoter (pOR5K1) in BLCA (n = 10), BRCA (n = 74), LUAD (n = 22), LUSC (n = 16), STAD (n = 21), and UCEC (n = 13) patient tumors. Dashed line: regions with a sum of reads above our threshold ($\log_2$ RPM > 0) were considered "accessible". Error bars: standard deviation. Significance analysis by Dunn's test with Holm correction (* $P <$ 0.05, ** $P < 0.01$, *** $P < 0.001$, ns: not significant). **(C)** UCSC Genome Browser [94] visualization of the *SOX2* region with ATAC-seq data from BLCA, BRCA, LUAD, LUSC,

458 STAD, and UCEC patient tumors (n = 5 in each cancer type) [107]. ATAC-seq reads were
459 normalized by library size (RPM). Scale: 0 – 250 RPM. **(D)** ATAC-seq signal at SRR124
460 and SRR134 regions against ATAC-seq signal for the *SOX2* promoter (pSOX2) from 74
461 BRCA, 22 LUAD, and 16 LUSC patient tumors. Correlation is shown for accessible
462 chromatin ($log_2$ RPM > 0). Grey: tumors with closed chromatin ($log_2$ RPM < 0) at either
463 region, not included in the correlation analysis. Significance analysis by Pearson
464 correlation. Bolded line: fitted linear regression model. Shaded area: 95% confidence
465 region for the regression fit. **(E)** Comparison of $log_2$-normalized *SOX2* transcript levels
466 ($log_2$ counts) between BRCA, LUAD, and LUSC patient tumors according to the
467 chromatin accessibility at SRR124 and SRR134 regions. Chromatin accessibility at
468 each region was considered "low" if $log_2$ RPM < -1, or "high" if $log_2$ RPM > 1. RNA-seq
469 reads were normalized to library size using DESeq2 [91]. Error bars: standard deviation.
470 Significance analysis by a two-sided t-test with Holm correction.
471
472 ***FOXA1* and *NFIB* are upstream regulators of the SRR124–134 cluster**
473 With the indication that the SRR124–134 cluster is driving *SOX2* overexpression in
474 patient tumors, we investigated which transcription factors regulate this cluster in BRCA,
475 LUAD, and LUSC. From a list of 1622 human transcription factors [109], we found that the
476 expression of 115 transcription factors was significantly (FDR-adjusted $Q$ < 0.05)
477 associated with chromatin accessibility levels at SRR124, whereas accessibility at
478 SRR134 was associated with the expression of 90 transcription factors (Figure 5A,
479 Supplementary Table S18). From this list, we focused our investigation on FOXA1 and
480 NFIB, which show binding at both SRR124 and SRR134 in MCF-7 cells [81].

The expression of *FOXA1* is positively (Pearson correlation $R > 0$) and significantly correlated to accessible chromatin at both SRR124 ($R$ = 0.39; FDR-adjusted $Q$ = $1.97\times10^{-3}$) and SRR134 ($R$ = 0.46; $Q$ = $1.41\times10^{-4}$) (Figure 5B). By separating BRCA, LUAD, and LUSC patient tumors according to the chromatin accessibility levels at each region, we found that tumors with the most accessible chromatin within SRR124 ($P$ = $2.38\times10^{-4}$, t-test) and SRR134 ($P$ = $1.53\times10^{-4}$) also significantly overexpress *FOXA1* compared to tumors with low accessibility at these regions (Figure 5C, Supplementary Table S19). On the other hand, we found the expression of *NFIB* to be negatively (Pearson correlation $R < 0$) and significantly correlated with chromatin accessibility at both SRR124 ($R$ = -0.49; $Q$ = $4.12\times10^{-5}$) and SRR134 ($R$ = -0.51; $Q$ = $1.32\times10^{-5}$) (Figure 5D). Patient tumors with highly accessible chromatin within SRR124 ($P$ = $1.46\times10^{-6}$) and SRR134 ($P$ = $1.24\times10^{-5}$) also display significantly downregulated *NFIB* expression (Figure 5E, Supplementary Table S20). These data suggest that whereas FOXA1 could be inducing increased accessibility at the SRR124–134 cluster, NFIB expression could counteract FOXA1 by acting as a repressor.

To assess the contribution of these transcription factors to enhancer activity, we overexpressed either *FOXA1* or *NFIB* in H520, MCF-7, PC-9, and T47D cells and compared SRR124 and SRR134 activity to cells transfected with an empty vector (mock) containing only a fluorescent marker. Although endogenous *FOXA1* and *NFIB* expression levels are already high in both MCF-7 and T47D cells (Supplementary Figure S5A), we found that overexpression of *FOXA1* significantly increased ($\log_2$ FC > 1; $P < 0.05$, Tukey's test) the enhancer activity of both SRR124 and SRR134 in the H520, MCF-7, PC-9, and T47D cell lines, whereas *NFIB* overexpression led to a

significant decrease ($\log_2$ FC < 1; $P$ < 0.05) in SRR124 and SRR134 enhancer activity in the H520, MCF-7, and T47D cell lines (Figure 5F). This further indicates that *FOXA1* overexpression increases SRR124–134 activity, whereas NFIB represses the activity of this cluster.

To assess the importance of FOXA1 and NFIB motifs in modulating enhancer activity, we analyzed the SRR134 sequence using the JASPAR2022 motif database [103] and mutated FOXA1 (GTAAACA) or NFIB (TGGCAnnnnGCCAA) motifs (mutated SRR134 sequences in Supplementary Table S21). We found that mutation of the FOXA1 motif abolished SRR134 enhancer activity compared to WT SRR134 within MCF-7 ($P = 1.53\times10^{-5}$, Tukey's test), PC-9 ($P = 1\times10^{-2}$), and T47D ($P = 4.48\times10^{-6}$) cells, whereas no significant change ($P$ > 0.05) in enhancer activity was found for the NFIB-mutated construct (Figure 5G). These data indicate that the FOXA1 motif is crucial for sustaining SRR134 activity, whereas the NFIB motif is dispensable in this context, as would be expected for a negative regulator under conditions where the activity of the target is high.

With the evidence that these transcription factors are modulating SRR124–134 activity, we investigated their transcriptional effects over *SOX2* expression. We used CRISPR homology-directed repair (HDR) to create an MCF-7 cell line in which the *SOX2* gene is tagged with a 2A self-cleaving peptide (P2A) followed by a blue fluorescent protein (tagBFP). This cell line, MCF-7 *SOX2*-P2A-tagBFP, allows rapid visualization of *SOX2* transcriptional changes by measuring tagBFP signal through fluorescence-activated cell sorting (FACS). To validate this model, we sorted cells within the top 10% ($BFP^{+ve}$) and bottom 10% ($BFP^{-ve}$) tagBFP signal (Supplementary Figure

S5B). We found that $BFP^{+ve}$ cells showed a significant ($P = 4.25\times10^{-5}$, paired t-test) increase in *SOX2* expression, and significantly upregulated transcription of enhancer RNA (eRNA) at SRR124 ($P = 1.54\times10^{-4}$) and SRR134 ($P = 5.13\times10^{-5}$) compared to $BFP^{-ve}$ cells (Figure 5H). This confirms that the tagBFP signal is directly correlated to *SOX2* transcription levels in MCF-7 *SOX2*-P2A-tagBFP cells.

Finally, we overexpressed *FOXA1* or *NFIB* in MCF-7 SOX2-P2A-tagBFP to assess changes in *SOX2* transcription. Although overexpression of *FOXA1* did not significantly (chi-squared T(x)=63.70) change tagBFP signal, we found that overexpression of *NFIB* significantly (chi-squared T(x)=1168.88) reduced tagBFP signal compared to transfection of an empty vector (mock) (Figure 5I). This confirms the repression effect of NFIB over *SOX2* expression and illustrates a potential mechanism upstream of *SOX2* that modulates chromatin accessibility at the SRR124–134 cluster and subsequent control of *SOX2* transcription in cancer cells.

**Figure 5: FOXA1 and NFIB are upstream regulators of SRR124 and SRR134. (A)** Heatmap of the Pearson correlation between transcription factor expression [69] and chromatin accessibility [107] at SRR124 and SRR134 in BRCA, LUAD, and LUSC patient tumors (n = 111). Transcription factors are ordered according to their correlation to chromatin accessibility at each region. Red: transcription factors with a positive correlation ($R > 0$; FDR-adjusted $Q < 0.05$) to chromatin accessibility. Blue: transcription factors with a negative correlation ($R < 0$; $Q < 0.05$) to chromatin accessibility. Asterisk: transcription factors that show binding at SRR124 or SRR134 by ChIP-seq [81]. **(B)** Correlation analysis between *FOXA1* expression ($\log_2$ counts) and chromatin

accessibility ($log_2$ RPM) at SRR124 and SRR134 regions in BRCA (n = 74), LUAD (n = 21), and LUSC (n = 16) tumors. RNA-seq reads were normalized to library size using DESeq2 [91]. Significance analysis by Pearson correlation (n = 111). Bolded line: fitted linear regression model. Shaded area: 95% confidence region for the regression fit. **(C)** Comparison of *FOXA1* expression ($log_2$ counts) from BRCA, LUAD, and LUSC patient tumors according to their chromatin accessibility at the SRR124 and SRR134 regions. Chromatin accessibility at each region was considered "low" if $log_2$ RPM < 1, or "high" if $log_2$ RPM > 1. RNA-seq reads were normalized to library size using DESeq2 [91]. Error bars: standard deviation. Significance analysis by a two-sided t-test with Holm correction. **(D)** Correlation analysis between *NFIB* expression ($log_2$ counts) and chromatin accessibility ($log_2$ RPM) at SRR124 and SRR134 regions in BRCA (n = 74), LUAD (n = 21), and LUSC (n = 16) tumors. RNA-seq reads were normalized to library size using DESeq2 [91]. Significance analysis by Pearson correlation (n = 111). Bolded line: fitted linear regression model. Shaded area: 95% confidence region for the regression fit. **(E)** Comparison of *NFIB* expression ($log_2$ counts) from BRCA, LUAD, and LUSC patient tumors according to their chromatin accessibility at the SRR124 and SRR134 regions. Chromatin accessibility at each region was considered "low" if $log_2$ RPM < 1, or "high" if $log_2$ RPM > 1. RNA-seq reads were normalized to library size using DESeq2 [91]. Error bars: standard deviation. Significance analysis by a two-sided t-test with Holm correction. **(F)** Relative fold change ($log_2$ FC) in luciferase activity driven by SRR124 and SRR134 after overexpression of either *FOXA1* or *NFIB* compared to an empty vector containing (mock negative control, miRFP670). Dashed line: average activity of the mock control. Error bars: standard deviation. Significance analysis by

Tukey's test (n = 5; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not significant). **(G)** Relative luciferase activity driven by WT, FOXA1-mutated, and NFIB-mutated SRR134 constructs compared to a minimal promoter (minP) vector in the MCF-7, PC-9, and T47D cell lines. Dashed line: average activity of minP. Error bars: standard deviation. Significance analysis by Tukey's test (n = 5; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not significant). **(H)** RT-qPCR comparison of transcripts at *SOX2*, SRR124, and SRR134 between sorted $BFP^{-ve}$ and $BFP^{+ve}$ MCF-7 cells normalized to unsorted population. Error bars: standard deviation. Significance analysis by paired t-test with Holm correction (n = 6; *** $P < 0.001$). **(I)** FACS density plot comparing tagBFP signal between *SOX2*-P2A-tagBFP MCF-7 cells transfected with an empty vector (mock negative control, miRFP670), FOXA1-T2A-miRFP670, or NFIB-T2A-miRFP670. tagBFP signal was acquired from successfully transfected live cells ($miRFP^{+}/PI^{-}$) after 5 days post-transfection. Significance analysis by FlowJo's chi-squared T(x) test. T(x) scores above 1000 were considered "strongly significant" (*** $P < 0.001$), whereas T(x) scores under 100 were considered "non-significant".

## SRR124 and SRR134 are conserved enhancers across mammals and are required for the separation of the anterior foregut

*SOX2* is required for the proper development of multiple tissues [37], including the digestive and respiratory systems in the mouse [25,27,29,31,32,38] and in humans [33,34]. Therefore, we questioned whether the SRR124–134 cluster drives *SOX2* expression in additional contexts other than cancer. A compilation of chromatin accessibility data from cardiac, digestive, embryonic, lymphoid, musculoskeletal, myeloid, neural, placental,

pulmonary, renal, skin, and vascular tissues [80,81,110] showed that both SRR124 and SRR134 display increased chromatin accessibility in digestive and respiratory tissues alongside cancer samples (Figure 6A). By comparing DNase-seq signal from fetal lung and stomach tissues [81], we found that both SRR124 (lung $P = 1.25\times10^{-6}$; stomach $P = 9.64\times10^{-4}$; holm-adjusted Dunn's test) and SRR134 (lung $P = 1.14\times10^{-3}$; stomach $P = 0.045$), together with SRR2 (lung $P = 1.55\times10^{-3}$; stomach $P = 5.74\times10^{-5}$), are significantly more accessible than pOR5K1 (Figure 6B, Supplementary Table S22). This suggests that SRR124 and SRR134 are contributing to *SOX2* expression during the development of the digestive and respiratory systems.

Since critical developmental genes are often controlled by highly conserved enhancers across species [111,112], we hypothesized that the SRR124–134 cluster might regulate *SOX2* expression during development in other species. By analyzing PhyloP conservation scores [94,113], we discovered that both SRR124 and SRR134 contain a highly conserved core sequence that is preserved across mammals, birds, reptiles, and amphibians (Figure 6C). After aligning and comparing enhancer sequences between humans and mice, we found the core sequence at both SRR124 and SRR134 are highly conserved (> 80%) in the mouse genome (Supplementary Figure S6A). We termed these homologous regions as mSRR96 (96 kb downstream of the mouse *Sox2* promoter; homologous to the human SRR124) and mSRR102 (102 kb downstream of the mouse *Sox2* promoter; homologous to the human SRR134). Enhancer feature analysis in the developing lung and stomach tissues in the mouse [81,114] showed that both mSRR96 and mSRR102 display increased chromatin accessibility and H3K27ac signal throughout developmental days E14.5 to the 8$^{th}$ post-natal week (Figure 6D).

Interestingly, mSRR96 and mSRR102 display higher ATAC-seq and H3K27ac signal towards the later stages of development in the lungs, but at early stages of development in the stomach. This suggests a distinct spatiotemporal contribution of this homologous cluster to *Sox2* expression during the development of these tissues in the mouse. ATAC-seq quantification (genomic coordinates in Supplementary Table S23) showed that both mSRR96 (lung $P = 5.54\times10^{-5}$; stomach $P = 2.37\times10^{-4}$; Holm-adjusted Dunn's test) and mSRR102 (lung $P = 1.27\times10^{-3}$; stomach $P = 0.046$) are significantly more accessible than the repressed promoter of the olfactory gene *Olfr266* (pOlfr266, negative control) during the development of the lungs and stomach in the mouse (Supplementary Figure S6B, Supplementary Table S24). Together, these results suggest a conserved *SOX2* regulatory mechanism across multiple species and support a model in which the SRR124 and SRR134 enhancers and their homologs regulate *SOX2* expression during the development of the digestive and respiratory systems.

To assess the contribution of the mSRR96 and mSRR102 regions to the development of the mouse, we generated a knockout containing a deletion spanning the mSRR96–102 enhancer cluster (ΔmENH) (Figure 6E). We crossed animals carrying a heterozygous mSRR96–102 deletion ($\Delta mENH^{+/-}$) and determined the number of pups alive at weaning (P21) from each genotype. We found a significant ($P = 3.92\times10^{-6}$, Chi-squared test) deviation from the expected mendelian ratio, with no homozygous mice ($\Delta mENH^{-/-}$) alive at weaning (Figure 6F), demonstrating that the mSRR96–102 enhancer cluster is crucial for survival in the mouse. To investigate the resulting phenotype in a homozygous mSRR96–102 enhancer deletion, we collected E18.5 embryos and prepared cross-sections at the thymus level from five animals of each

phenotype (WT, ΔmENH$^{+/-}$, and ΔmENH$^{-/-}$) (Figure 6G). Similar to other studies that interfered with *Sox2* expression during development [22,25,32], we found that all five ΔmENH$^{-/-}$ embryos developed EA/TEF, where the esophagus and trachea fail to separate during embryonic development (Figure 6H). WT and ΔmENH$^{+/-}$ embryos, on the other hand, showed normal development of the esophageal and tracheal tissues. Finally, immunohistochemistry showed the complete absence of the SOX2 protein within the EA/TEF tissue in ΔmENH$^{-/-}$ embryos, whereas WT and ΔmENH$^{+/-}$ embryos showed high levels of SOX2 protein within both the esophagus and tracheal tubes (Figure 6I). Together, these results demonstrate that mSRR96 and mSRR102 are imperative to drive *Sox2* expression during the development of the esophagus and trachea.

**Figure 6: The SRR124 and SRR134 enhancers are conserved across species and are required for the separation of the esophagus and trachea in the mouse. (A)** UCSC Genome Browser [94] view of the *SOX2* region containing a compilation of chromatin accessibility tracks of multiple human tissues [80,81,110]. Arrow: increased chromatin accessibility at the SRR124–134 cluster in cancer, digestive, and respiratory tissues. **(B)** DNAse-seq quantification ($\log_2$ RPM) at the *RAB7A* promoter (pRAB7A), *SOX2* promoter (pSOX2), SRR1, SRR2, SRR124, SRR134, human SCR (hSCR), and a desert region within the *SOX2* locus (desert) compared to the background signal at the repressed *OR5K1* promoter (pOR5K1) in lung and stomach embryonic tissues [81]. Dashed line: Regions with a sum of reads above our threshold ($\log_2$ RPM > 0) were considered “accessible”. Error bars: standard deviation. Significance analysis by Dunn’s

test with Holm correction (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not significant). **(C)** UCSC Genome Browser [94] with PhyloP conservation scores [113] at the SRR124 and SRR134 enhancers across mammals, birds, reptiles, and amphibians species. Black lines: highly conserved sequences. Empty lines: variant sequences. **(D)** UCSC Genome Browser [94] view of the *Sox2* region in the mouse. ATAC-seq and H3K27ac ChIP-seq data from lung and stomach tissues throughout developmental days E14.5 to the 8$^{th}$ post-natal week [81,114]. mSRR96: homologous to SRR124. mSRR102: homologous to SRR134. Reads were normalized to library size (RPM). **(E)** Illustration demonstrating the mSRR96–102 enhancer cluster CRISPR deletion (ΔmENH) in C57BL/6J mouse embryos. **(F)** Quantification and genotype of the C57BL/6J progeny from mSRR96–102-deleted crossings (ΔmENH$^{+/-}$). Pups were counted and genotyped at weaning (P21). The expected numbers of heterozygous and homozygous (ΔmENH$^{-/-}$) pups are twice and equal to, respectively, the number of obtained WT animals. Significance analysis by chi-squared test to measure the deviation in the number of obtained pups from the expected mendelian ratio of 1:2:1 (WT : ΔmENH$^{+/-}$ : ΔmENH$^{-/-}$). **(G)** Transverse cross-section of fixed E18.5 embryos at the start of the thymus. **(H)** Embryo sections stained with Hematoxylin and Eosin (H&E). The scale bar represents 500µm. Es: esophagus; Tr: trachea; EA/TEF: esophageal atresia with distal tracheoesophageal fistula. **(I)** Embryo sections stained with SOX2. The scale bar represents 500µm. Es: esophagus; Tr: trachea; EA/TEF: esophageal atresia with distal tracheoesophageal fistula.

## DISCUSSION

Our findings reveal that the SRR124–134 enhancer cluster is essential for *Sox2* expression in the developing airway and digestive systems and is required for the separation of the esophagus and trachea during mouse development. When embryogenesis is complete, *Sox2* expression is downregulated in most cell types as its developmental enhancers are decommissioned. We propose that aberrant upregulation of the pioneer factor *FOXA1* recommissions both SRR124 and SRR134 in tumor cells, driving *SOX2* overexpression in breast and lung cancer. As SOX2 is also a pioneer transcription factor, increased levels of this protein further reprogram the chromatin landscape of cancer cells, binding at multiple downstream regulatory regions, increasing chromatin accessibility, and driving subsequent upregulation of genes associated with epithelium development, ultimately supporting a tumor-initiating phenotype.

The observation that enhancers involved in the development of the airway and digestive systems are recommissioned to support *SOX2* upregulation during tumorigenesis is in line with observations that tumor-initiating cells acquire a less differentiated phenotype [115–118]. It is more surprising, however, that the *SOX2* gene is regulated by common enhancers in both breast and lung cancer cells as enhancers are usually highly tissue-specific [7,111,112,119]. Our observation that *FOXA1* expression is significantly correlated to chromatin accessibility at the SRR124–134 cluster and increases the transcriptional output of the SRR124 and SRR134 enhancers provides a mechanistic link between breast and lung developmental programs and cancer progression. FOXA1 is directly involved in the branching morphogenesis of the epithelium in breast [120,121] and lung [122,123] tissues, where SOX2 also plays an important

role [27,59]. Overexpression of both *FOXA1* [7–9,11,124–126] and *SOX2* [54,65,127] have also been individually linked to the activation of transcriptional programs associated with multiple types of cancer. Therefore, we propose that FOXA1 is one of the key players responsible for misactivation of the SRR124–134 cluster in cancer, which then drives *SOX2* overexpression in breast and lung tumors. As mutation of the FOXA1 motif disrupted SRR134 enhancer activity, and this motif is shared among other members of the forkhead box (FOX) transcription factor family [128], it remains unclear if FOXA1 alone activates the SRR124–134 cluster, or whether other FOX proteins are involved in this process. For example, *FOXM1* overexpression, which also showed binding at both SRR124 and SRR134 in MCF-7 cells, has similarly been associated with poor patient outcomes in multiple types of cancer [129].

In addition to the activating role of FOXA1, we identified NFIB as a negative regulator of *SOX2* expression through inhibition of SRR124–134 activity. NFIB is normally required for the development of multiple tissues [reviewed in 130], including the brain and lungs [131–133], tissues in which *SOX2* expression is also tightly regulated [27,134]. In the lungs, *NFIB* is essential for promoting the maturation and differentiation of progenitor cells [131,132]. This is in stark contrast to *SOX2*, which inhibits the differentiation of lung cells [27]. Interestingly, NFIB seems to have paradoxical roles in cancer, acting both as a tumor suppressor and as an oncogene in different tissues [135]. Among its tumor suppressor activity, NFIB acts as a barrier to skin carcinoma progression [136], and its downregulation is associated with dedifferentiation and aggressiveness in LUAD [137]. On the other hand, *SOX2* promotes skin [65] and lung [138] cancer progression. As an oncogene, NFIB promotes cell proliferation and metastasis in STAD [139], where *SOX2*

731 downregulation is associated with poor patient outcomes [140–142]. With this contrasting
732 relationship between *SOX2* and *NFIB* across multiple tissues, we propose that NFIB
733 normally acts as a suppressor of SRR124–134 activity and *SOX2* expression during the
734 differentiation of progenitor cells; downregulation of *NFIB* expression then results in
735 *SOX2* overexpression during tumorigenesis of the breast and lung.

736 We initially hypothesized that the neural enhancers SRR1 and SRR2 [18,19,143],
737 and/or the pluripotency-associated SCR [20,21] might be recommissioned during cancer
738 progression, as stem cell-related enhancers have been shown to acquire enhancer
739 features in tumorigenic cells [144]. Although other studies have also proposed the
740 activation of either SRR1 [41,68] or SRR2 [145,146] as the main drivers of *SOX2*
741 overexpression in BRCA, we found no evidence of this mechanism and instead
742 identified the SRR124–134 cluster as the main driver of *SOX2* expression in BRCA and
743 LUAD. Our patient tumor analysis did show that GBM and LGG were the only cancer
744 types that display a unique and consistent pattern of accessible chromatin at SRR1 and
745 SRR2, which is likely related to glioma cells assuming a neural stem cell-like identity to
746 sustain high levels of cell proliferation in the brain [61]. In fact, SRR2 deletion was shown
747 to downregulate *SOX2* and reduce cell proliferation in GBM cells [147], highlighting
748 enhancer specificity to different tumor types. In line with these findings, our observation
749 that PC-9 LUAD cells are dependent on SRR124–134 for *SOX2* transcription, whereas
750 in H520 LUSC cells SRR124–134 is dispensable, again highlights these tumor-type
751 specific regulatory mechanisms. LUSC tumors frequently amplify the *SOX2* locus
752 [57,58,71,72], whereas LUAD tumors do not [148], indicating that different mechanisms are
753 involved in genome dysregulation in these two lung cancer subtypes. Interestingly, a

754 further downstream enhancer cluster located ~55 kb away from SRR124–134 is co-
755 amplified with *SOX2* in LUSC cell lines [72], revealing an additional mechanism that could
756 sustain *SOX2* overexpression in the absence of the SRR124–134 cluster in certain
757 types of LUSC but not in LUAD.

758 Deletion of mSRR96–102, homolog of the human SRR124–134 cluster, resulted in
759 EA/TEF which is also observed in human cases with *SOX2* heterozygous mutations
760 [33,34]. Interestingly, a recent study showed that insertion of a CTCF insulation cluster
761 downstream of the *Sox2* gene, but upstream of mSRR96–102, disrupts *Sox2*
762 expression, impairs separation of the esophagus and trachea, and results in perinatal
763 lethality due to EA/TEF in the mouse [22]. This was of particular interest for understanding
764 enhancer functional nuances since the SCR, which is required for *Sox2* transcription at
765 implantation, can overcome the insulator effect of this insertion. The authors proposed
766 that enhancer density might explain the EA/TEF phenotype, as chromatin features
767 suggested that enhancers in the developing lung and stomach tissues might be spread
768 over a 400 kb domain [22]. The 6 kb deletion that removes the mSRR96–102 cluster
769 causing EA/TEF suggests this is not the case. Instead, we propose that the sensitivity of
770 each cell type to gene dosage is behind the differing ability of CTCF to block distal
771 enhancers. This is based on two observations: in humans, heterozygous *SOX2*
772 mutations are linked with the anophthalmia-esophageal-genital syndrome; in mice,
773 hypomorphic *Sox2* alleles display similar phenotypes in the eye [24] and EA/TEF [25,32].
774 This suggests that cells from the peri-implantation phase are less sensitive to lower
775 *Sox2* dosages compared to cells from the developing airways and digestive systems in
776 both species and explains the aberrant phenotypes observed at term.

Our findings illustrate how cis-regulatory regions can similarly drive gene expression in both healthy and diseased contexts and serve as a prime example of how developmental-associated enhancers may become misactivated in cancer. The fact that we have found a digestive/respiratory-associated enhancer cluster driving gene expression in a non-native context such as BRCA remains intriguing and reinforces a model in which tumorigenic cells often revert to a progenitor-like state that combines cis-regulatory features of progenitor cells from their own tissue compartments with those of other developing lineages [7]. This “dys-differentiation” mechanism seems to be centered around the overexpression of a few key development-associated pioneer transcription factors such as FOXA1. Identifying additional mechanisms that regulate this enhancer recommissioning could lead to new approaches to target tumor-initiating cells that depend on *SOX2* overexpression.

## MATERIALS AND METHODS

*Cell Culture*

MCF-7 cells were obtained from Eldad Zacksenhaus (Toronto General Hospital Research Institute, Toronto, CA). H520 (HTB-182) and T47D (HTB-133) cells were acquired from ATCC. PC-9 (90071810) cells were obtained from Sigma. Cell line identities were confirmed by short tandem repeat profiling. MCF-7 and T47D cells were grown in phenol red-free DMEM high glucose (Gibco), 10% FBS (Gibco), 1x Glutamax (Gibco), 1x Sodium Pyruvate (Gibco), 1x Penicillin-Streptomycin (Gibco), 1x Non-essential amino acids (Gibco), 25 mM HEPES (Gibco) and 0.01 mg/ml insulin (Sigma). H520 and PC-9 cells were grown in phenol red-free RPMI-1640 (Gibco), 10% FBS (Gibco), 1x Glutamax (Gibco), 1x Sodium Pyruvate (Gibco), 1x Penicillin-Streptomycin (Gibco), 1x Non-essential amino acids (Gibco), and 25 mM HEPES (Gibco). Cells were either passaged or had their medium replenished every three days.

*Genome editing*

Pairs of gRNA plasmids were constructed by inserting a 20 bp target sequence (Supplementary Table S25) into an empty gRNA cloning vector (a gift from George Church; Addgene plasmid # 41824; http://n2t.net/addgene:41824; RRID:Addgene_41824) [149] containing either miRFP670 (Addgene plasmid #163748) or tagBFP (Addgene plasmid #163747) fluorescent markers. Plasmids were sequenced to confirm correct insertion. Both gRNA (1 µg each) vectors were co-transfected with 3 µg of pCas9_GFP (a gift from Kiran Musunuru; Addgene plasmid #44719; http://n2t.net/addgene:44719; RRID:Addgene_44719) [150] using Neon electroporation

(Life Technologies). After 72 hours of transfection, cells were FACS sorted to select clones that contained all three plasmids. Sorted tagBFP$^{+}$/GFP$^{+}$/miRFP670$^{+}$ cells were grown in a bulk population and serially diluted into individual wells to generate isogenic populations. Once fully grown, each well was screened by PCR to confirm the deletion.

*Gene tagging*

*SOX2* was tagged with a P2A-tagBFP sequence in both alleles using CRISPR-mediated homology-directed repair (HDR) [151]. This strategy results in the expression of a single transcript that is further translated into two separate proteins due to ribosomal skipping [152]. In summary, we designed a gRNA that targets the 3' end of the *SOX2* stop codon (Supplementary Table S25, Addgene plasmid #163752). We then amplified ~800 bp homology arms upstream and downstream of the gRNA target sequence using high-fidelity Phusion Polymerase. We purposely avoided amplification of the *SOX2* promoter sequence to reduce the likelihood of random integrations in the genome. Both homology arms were then joined at each end of a P2A-tagBFP sequence using Gibson assembly. Flanking primers containing the gRNA target sequence were used to reamplify *SOX2*-P2A-tagBFP and add gRNA targets at both ends of the fragment; this approach allows excision of the HDR sequence from the backbone plasmid once inside the cell [153]. Finally, the full HDR sequence was inserted into a pJET1.2 (Thermo Scientific) backbone, midiprepped, and sequenced (Addgene #163751). 3µg of HDR template was then co-transfected with 1µg of hCas9 (a gift from George Church; Addgene plasmid #41815; http://n2t.net/addgene:41815 ; RRID:Addgene_41815) [149] and 1µg of gRNA plasmid using Neon electroporation (Life Technologies). A week after transfection,

tagBFP$^{+}$ cells were FACS sorted as a bulk population. Sorted cells were further grown for two more weeks, and single tagBFP$^{+}$ cells were isolated to generate isogenic populations. Once fully grown, each clone was screened by PCR and sequenced to confirm homozygous integration of P2A-tagBFP into the *SOX2* locus.

*Luciferase assay*

Luciferase activity was measured using the dual-luciferase reporter assay (Promega #E1960) that relies on the co-transfection of two plasmids: pGL4.23 (Firefly Luciferase, *luc2*) and pGL4.75 (Renilla Luciferase). Assayed plasmids were constructed by subcloning the empty pGL4.23 vector containing a minimal promoter (minP). SRR124, SRR134, SRR1, SRR2, and hSCR were PCR-amplified (primers in Supplementary Table S26) from MCF-7 genomic DNA using high-fidelity Phusion Polymerase and inserted in the forward position downstream of the *luc2* gene at the NotI restriction site. Constructs were sequenced to confirm correct insertions.

JASPAR2022 [103] was used to find FOXA1 (GTAAACA) and NFIB (TGGCAnnnnGCCAA) motifs in the SRR134 sequence. Only motifs with a score of 80% or higher were further analyzed. Bases within each motif sequence were mutated until the score was reduced below 80% without affecting co-occurring motifs or creating novel binding sites. In total, four FOXA1 motifs and two NFIB motifs were mutated (Supplementary Table S21). Engineered sequences were ordered as gene blocks (Eurofins) and inserted into pGL4.23 in the forward position. Constructs were sequenced to confirm correct insertions.

Cells were plated in 96-well plates with 4 technical replicates at $2.10^4$ cells per well. After 24 hours, a 200ng 50:1 mixture of enhancer vector and pGL4.75 was transfected using Lipofectamine 3000 (0.05µl Lipofectamine:1µl Opti-mem). For transcription factor overexpression analysis, a 200ng 50:10:1 mixture of enhancer vector, expression plasmid, and pGL4.75 was transfected. After 48 hours of transfection, cells were lysed in 1x Passive Lysis Buffer and stored at -80$^o$C until all 5 biological replicates were completed. Luciferase activity was measured in the Fluoroskan Ascent FL plate reader. Enhancer activity was calculated by normalizing the firefly signal from pGL4.23 to the *Renilla* signal from pGL4.75.

*Colony formation assay*

MCF-7 and PC-9 cells were seeded at low density (2,000 cells/well) into 6-well plates in triplicate for each cell line. Culture media was renewed every 3 days. After 12 days, cells were fixed with 3.7% paraformaldehyde for 10 minutes and stained with 0.5% crystal violet for 20 minutes to quantify the number of colonies formed. Crystal violet staining was then eluted with 10% acetic acid and absorbance was measured at 570 nm to evaluate cell proliferation. Each 6-well plate was considered one biological replicate and the experiment was repeated five times for each cell line (n = 5).

*FACS analysis*

For analyzing the effects of *FOXA1* and *NFIB* overexpression, $2.10^6$ SOX2-P2A-tagBFP cells were transfected with 50nM of plasmid expressing either miRFP670 (a gift from Vladislav Verkhusha; Addgene plasmid #79987; http://n2t.net/addgene:79987;

RRID:Addgene_79987), FOXA1-T2A-miRFP670 (Addgene plasmid #182335), or NFIB-T2A-miRFP670 (Addgene plasmid #187222) in 5 replicates. Five days after transfection, miRFP670, tagBFP, and propidium iodide (PI) (live/dead stain) signals were acquired; the amount of tagBFP signal from $miRFP670^{+}/PI^{-}$ cells was compared between each treatment across all replicates.

FlowJo's chi-squared T(x) test was used to contrast the effects of each treatment over tagBFP expression; T(x) scores above 1000 were considered "strongly significant" (***), whereas T(x) scores under 100 were considered "non-significant".

*Transcriptome analysis*

Total RNA was isolated from WT and enhancer-deleted (ΔENH) cell lines using the RNeasy kit. Genomic DNA was digested by Turbo DNAse. 500-2,000ng of total RNA was used in a reverse transcription reaction with random primers. cDNA was diluted in $H_2O$ and amplified in a qPCR reaction using SYBR Select Mix (primers in Supplementary Table S27). Amplicons were sequenced to confirm primer specificity. Gene expression was normalized to *PUM1* [70,154,155].

Total RNA was sent to The Centre for Applied Genomics (TCAG) for paired-end rRNA-depleted total RNA-seq (Illumina 2500, 125 bp). Read quality was checked by fastQC, trimmed using fastP [156] and mapped to the human genome (GRCh38/hg38) using STAR 2.7 [157]. Healthy breast epithelium RNA-seq was obtained from ENCODE (Supplementary Table S28) [80,81]. Mapped reads were quantified using featureCounts [158] and imported into DESeq2 [91] for normalization and differential expression analysis. Genes with a $|\log_2 FC| > 1$ and FDR-adjusted $Q < 0.01$ were considered significantly

changing. Differential gene expression was plotted using the EnhancedVolcano package. Correlation and clustering heatmaps were plotted using the pheatmap R package (https://cran.r-project.org/web/packages/pheatmap/index.html). Signal enrichment plot was prepared using NGS.plot [159].

Cancer patient transcriptome data were obtained from TCGA [69] using the TCGAbiolinks package [160]. The overall survival KM-plot was calculated using clinical information from TCGA [161]. Tumor transcriptome data were compared to healthy tissue using DESeq2. RNA-seq reads were normalized to library size using DESeq2 [91] and transformed to a $\log_2$ scale [$\log_2$ counts]. Differential gene expression was considered significant if $|\log_2 FC| > 1$ and $Q < 0.01$.

Gene Set Enrichment Analysis (GSEA) was performed by ranking genes according to their $\log_2$ FC in $\Delta ENH^{-/-}$ versus WT MCF-7 cells. The ranking was then analyzed using the GSEA function from the clusterProfiler package [105] with a threshold of FDR-adjusted $Q < 0.05$ using the MSigDB GO term database (C5).

*Chromatin accessibility analysis*

Cells were grown in three separate wells (n = 3) and 50,000 cells were sent to Princess Margaret Genomics Centre for ATAC-seq library preparation using the Omni-ATAC protocol [162]. ATAC-seq libraries were sequenced using 50 bp paired-ended parameters in the Illumina Novaseq 6000 platform. Read quality was checked by fastQC, trimmed using fastP and mapped to the human genome (GRCh38/hg38) using STAR 2.7. narrowPeaks were called using Genrich (https://github.com/jsh58/Genrich). Differential chromatin accessibility analysis was performed using diffBind [106]. ATAC-seq peaks with

a $|\log_2 FC| > 1$ and FDR-adjusted $Q < 0.01$ were considered significantly changing. Correlation heatmaps were generated using diffBind. Signal enrichment plot was prepared using NGS.plot [159]. Genes were separated into three categories according to their expression levels in our WT MCF-7 RNA-seq data.

Transcription factor footprint analysis was performed using TOBIAS [102] with standard settings. Motifs with a $|\log_2 FC| > 0.1$ and FDR-adjusted $Q < 0.01$ were considered significantly enriched in each condition. Replicates (n = 3) were merged into a single BAM file for each treatment. Motif enrichment at differential ATAC-seq peaks was performed using HOMER [163]. ATAC-seq peaks were assigned to their closest gene within ± 1 Mb distance from their promoter using ChIPpeakAnno [164].

Cancer patient ATAC-seq data was obtained from TCGA [107]. DNAse-seq from human developing tissues were obtained from ENCODE (Supplementary Table S28) [80,81]. Read quantification was calculated at the *RAB7a* (pRAB7a), *OR5K1* (pOR5K1), and *SOX2* (pSOX2) promoters, together with SRR1, SRR2, SRR124, SRR134, hSCR, and desert regions with a 1,500 bp window centered at the core of each region (genomic coordinates of each region in Supplementary Table S15). Reads were normalized to library size (RPM) and transformed to a $\log_2$ scale ($\log_2$ RPM) using a custom script (https://github.com/luisabatti/BAMquantify). Each region's average $\log_2$ RPM was compared to the *OR5K1* promoter for differential analysis using Dunn's test with Holm correction. Correlations were calculated using Pearson's correlation test and considered significant if FDR-adjusted $Q < 0.05$. Chromatin accessibility at SRR124 and SRR134 regions was considered low if $\log_2$ RPM < -1, medium if $-1 \leq \log_2$ RPM $\leq 1$, or high if $\log_2$ RPM > 1.

ATAC-seq from developing mouse lung and stomach tissues were obtained from ENCODE (Supplementary Table S28) [81] and others [114]. Conserved mouse regulatory regions were lifted from the human build (GRCh38/hg38) to the mouse build (GRCm38/mm10) using UCSC liftOver [94]. The number of mapped reads was calculated at the *Egf* (pEgf), *Olfr266* (pOlfr266), and *Sox2* (pSox2) promoters, together with the mouse mSRR1, mSRR2, mSRR96, mSRR102, mSCR and desert regions with a 1,500 bp window at each location (genomic coordinates in Supplementary Table S23). Each $\log_2$-transformed region's reads per million ($\log_2$ RPM) was compared to the negative *Olfr266* promoter control for differential analysis using Dunn's test with Holm correction.

*Conservation analysis*

Cross-species evolutionary conservation was obtained using phyloP [113]. Pairwise comparisons between human SRR124 and SRR134 (GRCh38/hg38) and mouse mSRR96 and mSRR102 (GRCm38/mm10) sequences were plotted using FlexiDot [165] with an 80% conservation threshold.

*ChIP-seq analysis*

Transcription factor and histone modifications ChIP-seq were obtained from ENCODE [81] (Supplementary Table S28) and others [88–90] (Supplementary Table S29). H3K4me1 and H3K27ac tracks were normalized to input and library size ($\log_2$ RPM). ATAC-seq reads were normalized to library size (RPM). Histone modification ChIP-seq tracks and transcription factor ChIP-seq peaks were uploaded to the UCSC browser [94] for visualization. Normalized H3K4me1, H3K27ac and ATAC-seq reads were quantified

and the difference in normalized signal was calculated using diffBind. Peaks with a $|\log_2 FC| > 1$ and $Q < 0.01$ were considered significantly changing.

Overlapping ChIP-seq and ATAC-seq peaks were analyzed using ChIPpeakAnno [164]. The hypergeometric test was performed by comparing the number of overlapping peaks to the total size of the genome divided by the median peak size.

*Mouse line construction*

Our mSRR96–102 knockout mouse line (ΔmENH) was ordered from and generated by The Centre for Phenogenomics (TCP) in Toronto, ON. The protocol for the generation of the mouse line has been previously described [166]. Briefly, C57BL/6J zygotes were collected from superovulated, mated, and plugged female mice at 0.5-day post coitum. Zygotes were electroporated with Cas9 RNPs complexes (gRNA sequences in Supplementary Table S25) and transferred into pseudopregnant female recipients within 3-4 hours of electroporation. Born pups (founders) were screened by end-point PCR and sequenced to confirm allelic mSRR96–102 deletions. Heterozygous mSRR96–102 founders ($\Delta mENH^{+/-}$) were then backcrossed to the parental strain to confirm germline transmission. Once the mouse line was established and the mSRR96–102 deletion was fully confirmed and sequenced in the N1 offspring, $\Delta mENH^{+/-}$ mice were crossed and the number of live pups from each genotype (WT, $\Delta mENH^{+/-}$, $\Delta mENH^{-/-}$) was assessed at weaning (P21). The obtained number of live pups from each genotype was then compared to the expected mendelian ratio of 1:2:1 (WT : $\Delta mENH^{+/-}$ : $\Delta mENH^{-/-}$) using a chi-squared test. Once the lethality of the homozygous deletion was confirmed at weaning, E18.5 embryos generated from new $\Delta mENH^{+/-}$

crosses were collected for further histological analyses. All procedures involving animals were performed in compliance with the Animals for Research Act of Ontario and the Guidelines of the Canadian Council on Animal Care. The TCP Animal Care Committee reviewed and approved all procedures conducted on animals at the facility.

*Histological analyses*

A total of 46 embryos were collected at E18.5 and fixed in 4% paraformaldehyde. Each embryo was genotyped, and a total of 15 embryos, 5 of each genotype (WT, $\Delta mENH^{+/-}$, $\Delta mENH^{-/-}$), were randomly selected, processed, and embedded in paraffin for sectioning and further analysis. Tissue sections were collected at 4µm thickness roughly at the start of the thymus. Sections were prepared by the Pathology Core at TCP.

Tissue sections were stained with Hematoxylin and Eosin (H&E) using an auto-stainer to ensure batch consistency. Slides were scanned using a Hamamatsu Nanozoomer slide scanner at 20X magnification. Images were then cropped and centered around the esophageal (Es) and tracheal (Tr) tissues.

Embryo sections were submitted to heat-induced epitope retrieval with TRIS-EDTA (pH 9.0) for 10 minutes, followed by quenching of endogenous peroxidase with Bloxall reagent (Vector). Non-specific antibody binding was blocked with 2.5 % normal horse serum (Vector), followed by incubation for 1 hour in Rabbit anti-SOX2 (Abcam, ab92494, 1:500). After washes, sections were incubated for 30 minutes with ImmPRESS Anti-Rabbit HRP (Vector) followed by DAB reagent, and counterstained in Mayer's hematoxylin.

## SUPPORTING INFORMATION

**Supplementary Figure S1: (A)** Super-logarithmic volcano plot of *PUM1* expression from RNA-seq of 21 cancer types compared to normal tissue [69]. Cancer types with $\log_2$ FC > 1 and FDR-adjusted $Q < 0.01$ were considered to significantly overexpress *PUM1*. Error bars: standard error. **(B)** Kaplan-Meier plot [167] of overall survival against time since diagnosis for 3,064 patients with BRCA (n = 1089), COAD (n = 453), GBM (n = 153), LIHC (n = 370), LUAD (n = 504), and LUSC (n = 495) tumors [161]. We divided patients into four equal groups and compared two groups: high *SOX2* expression (range: 10.06–16.36 $\log_2$ counts) and low *SOX2* expression (range: 0–1.67 $\log_2$ counts). RNA-seq reads were normalized to library size using DESeq2 [91]. Significance analysis by logrank test. The shadowed area represents the 95% confidence interval. **(C)** Comparison of *SOX2* expression ($\log_2$ counts) between luminal A (n = 560), luminal B (n = 207), HER2+ (n = 82), basal-like (n = 190) breast cancer subtypes and normal mammary tissue (n = 152) [69]. RNA-seq reads were normalized to library size using DESeq2 [91]. Error bars: standard deviation. Significance analysis by Tukey's test (*** $P < 0.001$, * $P < 0.05$, ns: not significant). **(D)** Volcano plot with DESeq2 [91] differential expression analysis between WT MCF-7 cells and breast epithelium [80]. Blue: 7,937 genes that significantly lost expression ($\log_2$ FC < -1; FDR-adjusted $Q < 0.01$) in WT MCF-7 cells. Pink: 5,335 genes that significantly gained expression ($\log_2$ FC > 1; $Q < 0.01$) in WT MCF-7 cells. Grey: 25,342 genes that maintained similar ($-1 \leq \log_2 \text{FC} \leq 1$) expression between WT MCF-7 and breast epithelium cells. **(E)** RT-qPCR analysis of *SOX2* transcript levels in the MCF-7, T47D, PC-9 and H520 cell lines. Error bars: standard deviation. Significance analysis by Tukey's test (n = 3; *** $P < 0.001$).

**Supplementary Figure S2: (A)** SOX2 protein levels in mouse embryonic stem cells (mESC, positive control), WT, ΔENH$^{+/-}$, and ΔENH$^{-/-}$ MCF-7 clones. Cyclophilin A (CypA) was used as a loading control across all samples. **(B)** RT-qPCR analysis of *SOX2* transcript levels in SRR124–134 heterozygous- (ΔENH$^{+/-}$) and homozygous- (ΔENH$^{-/-}$) deleted H520 (LUSC) clones compared to WT cells. Error bars: standard deviation. Significance analysis by Dunnett's test (n = 3, ns: not significant). **(C)** Euclidean distance pairwise comparison between WT and ΔENH$^{-/-}$ MCF-7 replicates (n = 3) using variance stabilizing transformed RNA-seq reads from DESeq2 [91]. Darker colors indicate a higher correlation. **(D)** Euclidean hierarchical clustering of 529 differentially expressed genes ($|\log_2 FC| > 1$; FDR-adjusted $Q < 0.01$) based on RNA-seq analysis between WT and ΔENH$^{-/-}$ MCF-7 replicates (n = 3). Reads were normalized for each gene across treatments (Z-score). Blue color indicates downregulated genes (Z-score < 0). Red color indicates upregulated genes (Z-score > 0).

**Supplementary Figure S3: (A)** ATAC-seq metagene enrichment plot ± 2 kb around the transcription start site (TSS) across all genes from WT and ΔENH$^{-/-}$ MCF-7 cells (n = 3). Reads were normalized by library size (RPM). Grey: TSS. Shaded area: standard deviation. **(B)** ATAC-seq metagene enrichment plot ± 2 kb around the transcription start site (TSS) across 12,167 high (average $\log_2$ counts = 9.12), 12,167 medium (average $\log_2$ counts = 2.94), and 12,167 low (average $\log_2$ counts = 0.43) expressed genes in WT MCF-7 cells. Genes were split into each group according to RNA-seq data. RNA-

seq reads were normalized to library size using DESeq2 [91]. Grey: TSS. (**C)** ATAC-seq metagene enrichment plot ± 2 kb around the transcription start site (TSS) across 12,167 high (average $\log_2$ counts = 9.10), 12,167 medium (average $\log_2$ counts = 2.97), and 12,167 low (average $\log_2$ counts = 0.47) expressed genes in ΔENH$^{-/-}$ MCF-7 cells. Genes were split into each group according to RNA-seq data. RNA-seq reads were normalized to library size using DESeq2 [91]. Grey: TSS. **(D)** Pairwise Pearson correlation comparison between WT and ΔENH$^{-/-}$ MCF-7 replicates (n = 3) using ATAC-seq normalized signal from diffBind [106]. Darker colors indicate a higher correlation. **(E)** Overlap between GRHL2 ChIP-seq peaks [89] and ATAC-seq peaks that significantly ($\log_2$ FC < -1; $P$ < 0.01) lost chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells. Significance analysis by the hypergeometric test [164]. **(F)** Overlap between RUNX2 ChIP-seq peaks [104] and ATAC-seq peaks that significantly ($\log_2$ FC < -1; $P$ < 0.01) lost chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells. Significance analysis by the hypergeometric test [164].

**Supplementary Figure S4: (A)** ATAC-seq signal at the *RAB7A* promoter (pRAB7A), *SOX2* promoter (pSOX2), SRR1, SRR2, SRR124, SRR134, human SCR (hSCR) and desert region versus the background signal at the repressed *OR5K1* promoter (pOR5K1) in BLCA (n = 10), BRCA (n = 74), COAD (n = 38), ESCA (n = 18), GBM (n = 9), HNSC (n = 9), LGG (n = 13), LIHC (n = 16), LUAD (n = 22), LUSC (n = 16), PRAD (n = 26), STAD (n = 21), TGCT (n = 9), and UCEC (n = 13) patient tumors. Dashed line: regions with $\log_2$ RPM > 0 were considered “accessible”. Error bars: standard deviation. Significance analysis by Dunn’s test with Holm correction (* $P$ < 0.05, ** $P$ < 0.01, *** $P$

< 0.001, ns: not significant). **(B)** ATAC-seq signal at the *SOX2* desert region (desert) against ATAC-seq signal for the *SOX2* promoter (pSOX2) from 74 BRCA, 22 LUAD, and 16 LUSC patient tumors. Dashed line: regions with $\log_2$ RPM > 0 were considered "accessible". Significance analysis by Pearson correlation. Bolded line: fitted linear regression model. Shaded area: 95% confidence region for the regression fit. **(C)** ATAC-seq signal at SRR124 and SRR134 regions against ATAC-seq signal for the *SOX2* promoter (pSOX2) from BRCA patient tumors separated into luminal A (n = 31), luminal B (n = 16), $HER2^+$ (n = 10), and basal-like (n = 14) subtypes. Correlation is shown for accessible chromatin ($\log_2$ RPM > 0). Grey: tumors with closed chromatin ($\log_2$ RPM < 0) at either region, not included in the correlation analysis. Significance analysis by Pearson correlation. Bolded line: fitted linear regression model. Shaded area: 95% confidence region for the regression fit.

**Supplementary Figure S5: (A)** RT-qPCR analysis of *FOXA1 and NFIB* expression in the H520, MCF-7, PC-9, and T47D cell lines. Error bars: standard deviation. Significance analysis by Tukey's test (n = 3; ** $P < 0.01$, *** $P < 0.001$). **(B)** FACS plot of tagBFP signal (450 nm) over side scatter (SSC) in WT and *SOX2*-P2A-tagBFP MCF-7 cells. Cell populations within the top 10% tagBFP signal were considered "tagBFP positive" ($BFP^{+ve}$), whereas populations within the bottom 10% BFP signal were considered "tagBFP negative" ($BFP^{-ve}$).

**Supplementary Figure S6: (A)** Dot-plot alignment of human (GRCh38/hg38, y-axis) SRR124 and SRR134, and mouse (GRCm38/mm10, x-axis) mSRR96 and mSRR102

homologous sequences (1,500 bp). Lines indicate high conservation scores (> 80%) across both species. Sequence alignment using Clustal Omega [168]. **(B)** ATAC-seq quantification ($log_2$ RPM) at the promoter of the housekeeping gene *Egf* (pEgf, positive control), *Sox2* promoter (pSox2), mSRR1, mSRR2, mSRR96, mSRR102, mSCR, and a mouse desert (mdesert) region compared to the background signal at the repressed *Olfr266* promoter (pOlfr266) in lung and stomach embryonic tissues from the mouse [81]. mSRR96: homologous to SRR124. mSRR102: homologous to SRR134. Dashed line: regions with a sum of reads above our threshold ($log_2$ RPM > 0) were considered "accessible". Error bars: standard deviation. Significance analysis by Dunn's test with Holm correction (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not significant).

**Supplementary Table S1:** List of TCGA tumor type abbreviations.

**Supplementary Table S2:** *SOX2* differential expression analysis between primary tumor vs. normal tissue across TCGA cancer types.

**Supplementary Table S3:** *PUM1* differential expression analysis between primary tumor vs. normal tissue across TCGA cancer types.

**Supplementary Table S4:** TCGA cancer patient overall survival analysis relative to *SOX2* expression levels.

**Supplementary Table S5:** TCGA copy number variation (CNV) and *SOX2* expression analysis.

**Supplementary Table S6:** RNA-seq differential expression analysis between WT MCF-7 vs. breast epithelium (ENCODE).

**Supplementary Table S7:** Differential ATAC-seq, H3K4me1, and H3K27ac analysis within ± 1 Mb of the *SOX2* gene in WT MCF-7 vs. Breast epithelium (ENCODE).

**Supplementary Table S8:** RNA-seq differential expression analysis comparing ΔENH$^{-/-}$ versus WT MCF-7 cells.

**Supplementary Table S9:** Gene set enrichment analysis (GSEA) in WT versus ΔENH$^{-/-}$ MCF-7 cells.

**Supplementary Table S10:** Significantly changing ATAC-seq peaks in ΔENH$^{-/-}$ versus WT MCF-7 cells.

**Supplementary Table S11:** ATAC-seq peaks that commonly gained signal in WT MCF-7 vs. breast epithelium and lost signal in ΔENH$^{-/-}$ MCF-7 cells.

**Supplementary Table S12:** ATAC-seq footprint analysis in ΔENH$^{-/-}$ vs. WT MCF-7 cells.

**Supplementary Table S13:** ChIP-seq motif analysis of GRHL2 peaks in WT MCF-7 cells.

**Supplementary Table S14:** ChIP-seq motif analysis of RUNX2 peaks in WT MCF-7 cells.

**Supplementary Table S15:** Coordinates of regions used in genome-wide analysis in humans (GRCh38/hg38).

**Supplementary Table S16:** Chromatin accessibility analysis across TCGA cancer types.

**Supplementary Table S17:** ATAC-seq quantification used to separate patient tumors into expression groups and their *SOX2* expression levels.

**Supplementary Table S18:** Significantly correlated transcription factors to accessible chromatin at the SRR124–134 cluster in BRCA, LUAD, and LUSC tumors.

**Supplementary Table S19:** *FOXA1* transcript levels and chromatin accessibility at the SRR124–134 cluster in BRCA, LUAD, and LUSC patient tumors.

**Supplementary Table S20:** *NFIB* transcript levels and chromatin accessibility at the SRR124–134 cluster in BRCA, LUAD, and LUSC patient tumors.

**Supplementary Table S21:** WT, FOXA1, and NFIB mutated SRR134 sequences.

**Supplementary Table S22:** Chromatin accessibility analysis in human (GRCh38/hg38) lung and stomach embryonic tissues.

**Supplementary Table S23:** Coordinates of regions used in genome-wide analysis in the mouse (GRCm38/mm10).

**Supplementary Table S24:** Chromatin accessibility analysis in mouse (GRCm38/mm10) embryonic lung and stomach tissues

**Supplementary Table S25:** List of gRNA sequences used for CRISPR/Cas9.

**Supplementary Table S26:** List of primers used for enhancer cloning.

**Supplementary Table S27:** List of primers used in RT-qPCR experiments.

**Supplementary Table S28:** ENCODE datasets used in this paper.

**Supplementary Table S29:** GEO datasets used in this paper.

## DATA AVAILABILITY

Sequencing and processed data files were submitted to the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) repository (GSE132344).

## ACKNOWLEDGMENTS

We thank all the members of the Mitchell laboratory for helpful discussions and support. We also thank the ENCODE Consortium and the TCGA project for generating and releasing data to the scientific community. Finally, we thank the contribution of the staff at TCP (The Centre for Phenogenomics), including Kyle Roberton, who handled the embedding, cutting, and H&E staining of E18.5 mouse embryos, and Vivian Bradaschia, responsible for the IHC staining. This work was supported by the Canadian Institutes of Health Research (FRN PJT153186 and PJT180312), the Canada Foundation for Innovation, and the Ontario Ministry of Research and Innovation (operating and infrastructure grants held by J.A.M.). BioRender.com was used to create parts of Figures 6E and 6G.

## AUTHOR CONTRIBUTIONS

L.E.A. designed and performed bioinformatic analyses, cell culture work, CRISPR deletions, data curation, gene expression quantification and led the conceptualization and writing of the manuscript; P.L.F. assessed cellular phenotypes, including the colony formation assay; L.H. acquired and processed TCGA ATAC-seq data and assisted in writing review & editing; M.C. assisted in the writing review & editing; M.M.H. provided TCGA data access and assisted in writing review & editing; J.A.M. was involved in

supervision, funding acquisition, data interpretation, experimental design and writing review & editing. All authors have participated in the editing and approval of the manuscript.

## REFERENCES


1. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
2. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
3. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
4. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
5. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010).
6. Rubin, A. J. *et al.* Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet* **49**, 1522–1528 (2017).
7. Stergachis, A. B. *et al.* Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell* **154**, 888–903 (2013).
8. Fu, X. *et al.* FOXA1 upregulation promotes enhancer and transcriptional reprogramming in endocrine-resistant breast cancer. *Proc Natl Acad Sci USA* **116**, 26823–26834 (2019).
9. Roe, J.-S. *et al.* Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. *Cell* **170**, 875-888.e20 (2017).

10. Bi, M. *et al.* Enhancer reprogramming driven by high-order assemblies of transcription factors promotes phenotypic plasticity and breast cancer endocrine resistance. *Nat Cell Biol* **22**, 701–715 (2020).
11. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958–970 (2008).
12. Avilion, A. A. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & Development* **17**, 126–140 (2003).
13. Chew, J.-L. *et al.* Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4/Sox2 Complex in Embryonic Stem Cells. *Mol Cell Biol* **25**, 6031–6046 (2005).
14. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
15. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131**, 861–872 (2007).
16. Yu, J. *et al.* Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. *Science* **318**, 1917–1920 (2007).
17. Ferri, A. L. M. *et al.* Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development* **131**, 3805–3819 (2004).
18. Tomioka, M. *et al.* Identification of Sox-2 regulatory region which is under the control of Oct-3/4–Sox-2 complex. *Nucleic Acids Res* **30**, 3202–3213 (2002).
19. Zappone, M. V. *et al.* Sox2 regulatory sequences direct expression of a (beta)-geo transgene to telencephalic neural stem cells and precursors of the mouse embryo,

revealing regionalization of gene expression in CNS stem cells. *Development* **127**, 2367–2382 (2000).

20. Zhou, H. Y. *et al.* A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes & Development* **28**, 2699–2711 (2014).
21. Li, Y. *et al.* CRISPR Reveals a Distal Super-Enhancer Required for Sox2 Expression in Mouse Embryonic Stem Cells. *PLoS ONE* **9**, e114485 (2014).
22. Chakraborty, S. *et al.* Enhancer–promoter interactions can bypass CTCF-mediated boundaries and contribute to phenotypic robustness. *Nat Genet* (2023) doi:10.1038/s41588-022-01295-6.
23. Graham, V., Khudyakov, J., Ellis, P. & Pevny, L. SOX2 Functions to Maintain Neural Progenitor Identity. *Neuron* **39**, 749–765 (2003).
24. Taranova, O. V. *et al.* SOX2 is a dose-dependent regulator of retinal neural progenitor competence. *Genes Dev* **20**, 1187–1202 (2006).
25. Que, J. *et al.* Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm. *Development* **134**, 2521–2531 (2007).
26. Kiernan, A. E. *et al.* Sox2 is required for sensory organ development in the mammalian inner ear. *Nature* **434**, 1031–1035 (2005).
27. Gontan, C. *et al.* Sox2 is important for two crucial processes in lung development: Branching morphogenesis and epithelial cell differentiation. *Developmental Biology* **317**, 296–309 (2008).
28. Driskell, R. R., Giangreco, A., Jensen, K. B., Mulder, K. W. & Watt, F. M. Sox2-positive dermal papilla cells specify hair follicle type in mammalian epidermis. *Development* **136**, 2815–2823 (2009).

29. Francis, R. *et al.* Gastrointestinal transcription factors drive lineage-specific developmental programs in organ specification and cancer. *Sci. Adv.* **5**, eaax8898 (2019).

30. Okubo, T., Pevny, L. H. & Hogan, B. L. M. Sox2 is required for development of taste bud sensory cells. *Genes & Development* **20**, 2654–2659 (2006).

31. Que, J., Luo, X., Schwartz, R. J. & Hogan, B. L. M. Multiple roles for Sox2 in the developing and adult mouse trachea. *Development* **136**, 1899–1907 (2009).

32. Teramoto, M. *et al.* The absence of SOX2 in the anterior foregut alters the esophagus into trachea and bronchi in both epithelial and mesenchymal components. *Biology Open* **9**, bio048728 (2020).

33. Zenteno, J. C., Perez-Cano, H. J. & Aguinaga, M. Anophthalmia-esophageal atresia syndrome caused by an SOX2 gene deletion in monozygotic twin brothers with markedly discordant phenotypes. *Am J Med Genet A* **140**, 1899–1903 (2006).

34. Williamson, K. A. *et al.* Mutations in SOX2 cause anophthalmia-esophageal-genital (AEG) syndrome. *Hum Mol Genet* **15**, 1413–1422 (2006).

35. Brunner, H. G. & van Bokhoven, H. Genetic players in esophageal atresia and tracheoesophageal fistula. *Curr Opin Genet Dev* **15**, 341–347 (2005).

36. Que, J., Choi, M., Ziel, J. W., Klingensmith, J. & Hogan, B. L. M. Morphogenesis of the trachea and esophagus: current players and new roles for noggin and Bmps. *Differentiation* **74**, 422–437 (2006).

37. Arnold, K. *et al.* Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell* **9**, 317–329 (2011).

1288 38. Tompkins, D. H. *et al.* Sox2 Is Required for Maintenance and Differentiation of
1289 Bronchiolar Clara, Ciliated, and Goblet Cells. *PLoS ONE* **4**, e8248 (2009).
1290 39. Wuebben, E. L. & Rizzino, A. The dark side of SOX2: cancer-a comprehensive
1291 overview. *Oncotarget* **8**, 44917–44943 (2017).
1292 40. Chen, Y. *et al.* The Molecular Mechanism Governing the Oncogenic Potential of
1293 SOX2 in Breast Cancer. *Journal of Biological Chemistry* **283**, 17969–17978 (2008).
1294 41. Leis, O. *et al.* Sox2 expression in breast tumours and activation in breast cancer
1295 stem cells. *Oncogene* **31**, 1354–1365 (2012).
1296 42. Liu, P. *et al.* SOX2 Promotes Cell Proliferation and Metastasis in Triple Negative
1297 Breast Cancer. *Front Pharmacol* **9**, (2018).
1298 43. Meng, Y., Xu, Q., Chen, L., Wang, L. & Hu, X. The function of SOX2 in breast
1299 cancer and relevant signaling pathway. *Pathology - Research and Practice* **216**,
1300 153023 (2020).
1301 44. Piva, M. *et al.* Sox2 promotes tamoxifen resistance in breast cancer cells. *EMBO*
1302 *Molecular Medicine* **6**, 66–79 (2014).
1303 45. Takeda, K. *et al.* Sox2 is associated with cancer stem-like properties in colorectal
1304 cancer. *Scientific Reports* **8**, 17639 (2018).
1305 46. Talebi, A., Kianersi, K. & Beiraghdar, M. Comparison of gene expression of SOX2
1306 and OCT4 in normal tissue, polyps, and colon adenocarcinoma using
1307 immunohistochemical staining. *Adv Biomed Res* **4**, 234 (2015).
1308 47. Zhang, X.-H., Wang, W., Wang, Y.-Q., Zhu, L. & Ma, L. The association of SOX2
1309 with clinical features and prognosis in colorectal cancer: A meta-analysis. *Pathol*
1310 *Res Pract* **216**, (2020).

48. Zhu, Y. *et al.* SOX2 promotes chemoresistance, cancer stem cells properties, and epithelial-mesenchymal transition by β-catenin and Beclin1/autophagy signaling in colorectal cancer. *Cell Death Dis* **12**, 449 (2021).

49. Alonso, M. M. *et al.* Genetic and Epigenetic Modifications of Sox2 Contribute to the Invasive Phenotype of Malignant Gliomas. *PLoS ONE* **6**, e26740 (2011).

50. Cox, J. L., Wilder, P. J., Desler, M. & Rizzino, A. Elevating SOX2 levels deleteriously affects the growth of medulloblastoma and glioblastoma cells. *PLoS One* **7**, e44087 (2012).

51. Gangemi, R. M. R. *et al.* SOX2 silencing in glioblastoma tumor-initiating cells causes stop of proliferation and loss of tumorigenicity. *Stem Cells* **27**, 40–48 (2009).

52. Hägerstrand, D. *et al.* Identification of a SOX2-dependent subset of tumor- and sphere-forming glioblastoma cells with a distinct tyrosine kinase inhibitor sensitivity profile. *Neuro Oncol* **13**, 1178–1191 (2011).

53. Sun, C. *et al.* Sox2 expression predicts poor survival of hepatocellular carcinoma patients and it promotes liver cancer cell invasion by activating Slug. *Med Oncol* **30**, 503 (2013).

54. Chou, Y.-T. *et al.* The Emerging Role of SOX2 in Cell Proliferation and Survival and Its Crosstalk with Oncogenic Signaling in Lung Cancer. *STEM CELLS* **31**, 2607–2619 (2013).

55. Sholl, L. M., Barletta, J. A., Yeap, B. Y., Chirieac, L. R. & Hornick, J. L. Sox2 protein expression is an independent poor prognostic indicator in stage I lung adenocarcinoma. *Am J Surg Pathol* **34**, 1193–1198 (2010).

56. Nakatsugawa, M. *et al.* SOX2 is overexpressed in stem-like cells of human lung adenocarcinoma and augments the tumorigenicity. *Lab Invest* **91**, 1796–1804 (2011).

57. Bass, A. J. *et al.* SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nature Genetics* **41**, 1238–1242 (2009).

58. Hussenet, T. *et al.* SOX2 Is an Oncogene Activated by Recurrent 3q26.3 Amplifications in Human Lung Squamous Cell Carcinomas. *PLOS ONE* **5**, e8960 (2010).

59. Domenici, G. *et al.* A Sox2–Sox9 signalling axis maintains human breast luminal progenitor and breast cancer stem cells. *Oncogene* (2019) doi:10.1038/s41388-018-0656-7.

60. Simões, B. M. *et al.* Effects of estrogen on the proportion of stem cells in the breast. *Breast Cancer Research and Treatment* **129**, 23–35 (2011).

61. Bulstrode, H. *et al.* Elevated FOXG1 and SOX2 in glioblastoma enforces neural stem cell identity through transcriptional control of cell cycle and epigenetic regulators. *Genes Dev* **31**, 757–773 (2017).

62. Jeon, H.-M. *et al.* ID4 imparts chemoresistance and cancer stemness to glioma cells by derepressing miR-9*-mediated suppression of SOX2. *Cancer Res* **71**, 3410–3421 (2011).

63. Zhang, L.-H. *et al.* TRIM24 promotes stemness and invasiveness of glioblastoma cells via activating Sox2 expression. *Neuro Oncol* **22**, (2020).

64. Singh, S. *et al.* EGFR/Src/Akt signaling modulates Sox2 expression and self-renewal of stem-like side-population cells in non-small cell lung cancer. *Molecular Cancer* **11**, 73 (2012).

65. Boumahdi, S. *et al.* SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature* **511**, 246–250 (2014).

66. Berezovsky, A. D. *et al.* Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* **16**, 193–206, 206.e19–25 (2014).

67. Fang, X. *et al.* The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. *BMC Genomics* **12**, 11 (2011).

68. Stolzenburg, S. *et al.* Targeted silencing of the oncogenic transcription factor SOX2 in breast cancer. *Nucleic Acids Research* **40**, 6725–6740 (2012).

69. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

70. Krasnov, G. S. *et al.* Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization. *Frontiers in Genetics* **10**, (2019).

71. Maier, S. *et al.* SOX2 amplification is a common event in squamous cell carcinomas of different organ sites. *Human Pathology* **42**, 1078–1088 (2011).

72. Liu, Y. *et al.* A predominant enhancer co-amplified with the SOX2 oncogene is necessary and sufficient for its expression in squamous cancer. *Nat Commun* **12**, 7139 (2021).

73. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat Methods* **10**, 1213–1218 (2013).

74. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).

75. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).

76. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

77. Soule, H. D., Vazquez, J., Long, A., Albert, S. & Brennan, M. A human cell line from a pleural effusion derived from a breast carcinoma. *Journal of the National Cancer Institute* **51**, 1409–1416 (1973).

78. Liang, S. *et al.* Isolation and characterization of human breast cancer cells with SOX2 promoter activity. *Biochemical and Biophysical Research Communications* **437**, 205–211 (2013).

79. Ling, G.-Q., Chen, Dong-bo, Wang, Bao-Qing & Zhang, Lan-Sheng. Expression of the pluripotency markers Oct3/4, Nanog and Sox2 in human breast cancer cell lines. *Oncology Letters* (2012) doi:10.3892/ol.2012.916.

80. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

81. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

82. Chen, C., Morris, Q. & Mitchell, J. A. Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. *BMC Genomics* **13**, 152 (2012).

83. Mitchell, J. A. *et al.* Nuclear RNA sequencing of the mouse erythroid cell transcriptome. *PLoS ONE* **7**, e49274 (2012).

84. Singh, G. *et al.* A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome Res.* **31**, 564–575 (2021).

85. Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. & de Laat, W. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell* **10**, 1453–1465 (2002).

86. Carter, D., Chakalova, L., Osborne, C. S., Dai, Y. & Fraser, P. Long-range chromatin regulatory interactions in vivo. *Nature Genetics* **32**, 623 (2002).

87. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).

88. Chan, H. L. *et al.* Polycomb complexes associate with enhancers and promote oncogenic transcriptional programs in cancer through multiple mechanisms. *Nat Commun* **9**, 3377 (2018).

89. Cocce, K. J. *et al.* The Lineage Determining Factor GRHL2 Collaborates with FOXA1 to Establish a Targetable Pathway in Endocrine Therapy-Resistant Breast Cancer. *Cell Rep* **29**, 889-903.e10 (2019).

90. Sato, T. *et al.* Epigenomic Profiling Discovers Trans-lineage SOX2 Partnerships Driving Tumor Heterogeneity in Lung Squamous Cell Carcinoma. *Cancer Research* **79**, 6084–6100 (2019).

91. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, (2014).

92. Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics* **6**, 241–252 (1964).

93. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).

94. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

95. Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* **27**, 246–258 (2017).

96. Tobias, I. C. *et al.* Transcriptional enhancers: from prediction to functional assessment on a genome-wide scale. *Genome* **64**, 426–448 (2021).

97. Dunnett, C. W. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association* **50**, 1096–1121 (1955).

98. Tukey, J. W. Comparing Individual Means in the Analysis of Variance. *Biometrics* **5**, 99 (1949).

99. Tompkins, D. H. *et al.* Sox2 Activates Cell Proliferation and Differentiation in the Respiratory Epithelium. *Am J Respir Cell Mol Biol* **45**, 101–110 (2011).

100. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell* **151**, 994–1004 (2012).

101. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).

102. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**, 4267 (2020).

103. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165–D173 (2022).

104. Jeselsohn, R. *et al.* Embryonic transcription factor SOX9 drives breast cancer endocrine resistance. *Proceedings of the National Academy of Sciences* **114**, E4482–E4491 (2017).

105. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

106. Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf (2011).

107. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).

108. Miyagi, S. *et al.* The Sox2 Regulatory Region 2 Functions as a Neural Stem Cell-specific Enhancer in the Telencephalon. *Journal of Biological Chemistry* **281**, 13374–13381 (2006).

109. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

110. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

111. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).

112. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).

113. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–121 (2010).

114. Liu, C. *et al.* An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci Data* **6**, 65 (2019).

115. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* **3**, 730–737 (1997).

116. Chaffer, C. L. *et al.* Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *PNAS* **108**, 7950–7955 (2011).

117. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).

118. Gupta, P. B. *et al.* Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633–644 (2011).

119. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

120. Bernardo, G. M. *et al.* FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. *Development* **137**, 2045–2054 (2010).

121. Liu, Y. *et al.* Foxa1 is essential for mammary duct formation. *Genesis* **54**, 277–285 (2016).

122. Besnard, V., Wert, S. E., Kaestner, K. H. & Whitsett, J. A. Stage-specific regulation of respiratory epithelial cell differentiation by Foxa1. *Am J Physiol Lung Cell Mol Physiol* **289**, L750-759 (2005).

123. Paranjapye, A., Mutolo, M. J., Ebron, J. S., Leir, S.-H. & Harris, A. The FOXA1 transcriptional network coordinates key functions of primary human airway epithelial cells. *Am J Physiol Lung Cell Mol Physiol* **319**, L126–L136 (2020).

124. Camolotto, S. A. *et al.* FoxA1 and FoxA2 drive gastric differentiation and suppress squamous identity in NKX2-1-negative lung cancer. *Elife* **7**, e38579 (2018).

125. Fu, X. *et al.* FOXA1 overexpression mediates endocrine resistance by altering the ER transcriptome and IL-8 expression in ER-positive breast cancer. *PNAS* **113**, E6600–E6609 (2016).

126. Orstad, G. *et al.* FoxA1 and FoxA2 control growth and cellular identity in NKX2-1-positive lung adenocarcinoma. *Dev Cell* **57**, 1866-1882.e10 (2022).

127. Liu, K.-C. *et al.* The multiple roles for Sox2 in stem cell maintenance and tumorigenesis. *Cell Signal* **25**, (2013).

128. Pierrou, S., Hellqvist, M., Samuelsson, L., Enerbäck, S. & Carlsson, P. Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. *The EMBO Journal* **13**, 5002–5012 (1994).

129. Li, L., Wu, D., Yu, Q., Li, L. & Wu, P. Prognostic value of FOXM1 in solid tumors: a systematic review and meta-analysis. *Oncotarget* **8**, 32298–32308 (2017).

130. Harris, L., Genovesi, L. A., Gronostajski, R. M., Wainwright, B. J. & Piper, M. Nuclear factor one transcription factors: Divergent functions in developmental versus adult stem cell populations. *Dev Dyn* **244**, 227–238 (2015).

131. Steele-Perkins, G. *et al.* The transcription factor gene Nfib is essential for both lung maturation and brain development. *Mol Cell Biol* **25**, 685–698 (2005).

132. Gründer, A. *et al.* Nuclear factor I-B (Nfib) deficient mice have severe lung hypoplasia. *Mech Dev* **112**, 69–77 (2002).

133. Hsu, Y.-C. *et al.* Mesenchymal nuclear factor I B regulates cell proliferation and epithelial differentiation during lung maturation. *Dev Biol* **354**, 242–252 (2011).

134. Favaro, R. *et al.* Hippocampal development and neural stem cell maintenance require Sox2-dependent regulation of Shh. *Nature Neuroscience* **12**, 1248–1256 (2009).

135. Becker-Santos, D. D., Lonergan, K. M., Gronostajski, R. M. & Lam, W. L. Nuclear Factor I/B: A Master Regulator of Cell Differentiation with Paradoxical Roles in Cancer. *EBioMedicine* **22**, 2–9 (2017).

136. Zhou, M. *et al.* miR-365 promotes cutaneous squamous cell carcinoma (CSCC) through targeting nuclear factor I/B (NFIB). *PLoS One* **9**, e100620 (2014).

137. Becker-Santos, D. D. *et al.* Developmental transcription factor NFIB is a putative target of oncofetal miRNAs and is associated with tumour aggressiveness in lung adenocarcinoma. *J Pathol* **240**, 161–172 (2016).

138. Ferone, G. *et al.* SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. *Cancer Cell* **30**, 519–532 (2016).

139. Wu, C. *et al.* NFIB promotes cell growth, aggressiveness, metastasis and EMT of gastric cancer through the Akt/Stat3 signaling pathway. *Oncol Rep* **40**, 1565–1573 (2018).

140. Otsubo, T., Akiyama, Y., Yanagihara, K. & Yuasa, Y. SOX2 is frequently downregulated in gastric cancers and inhibits cell growth through cell-cycle arrest and apoptosis. *Br J Cancer* **98**, 824–831 (2008).

141. Wang, S. *et al.* SOX2, a predictor of survival in gastric cancer, inhibits cell proliferation and metastasis by regulating PTEN. *Cancer Lett* **358**, 210–219 (2015).

142. Zhang, X. *et al.* SOX2 in gastric carcinoma, but not Hath1, is related to patients' clinicopathological features and prognosis. *J Gastrointest Surg* **14**, 1220–1226 (2010).

143. Miyagi, S. *et al.* The Sox-2 regulatory regions display their activities in two distinct types of multipotent stem cells. *Mol. Cell. Biol.* **24**, 4207–4220 (2004).

144. Aran, D. *et al.* Embryonic Stem Cell (ES)-Specific Enhancers Specify the Expression Potential of ES Genes in Cancer. *PLoS Genet* **12**, e1005840 (2016).

145. Iglesias, J. M. *et al.* The Activation of the Sox2 RR2 Pluripotency Transcriptional Reporter in Human Breast Cancer Cell Lines is Dynamic and Labels Cells with Higher Tumorigenic Potential. *Front Oncol* **4**, 308 (2014).

146. Jung, K. *et al.* Triple negative breast cancers comprise a highly tumorigenic cell subpopulation detectable by its high responsiveness to a Sox2 regulatory region 2 (SRR2) reporter. *Oncotarget* **6**, 10366–10373 (2015).

147. Saenz-Antoñanzas, A. *et al.* CRISPR/Cas9 Deletion of SOX2 Regulatory Region 2 (SRR2) Decreases SOX2 Malignant Activity in Glioblastoma. *Cancers (Basel)* **13**, 1574 (2021).

148. Björkqvist, A. M. *et al.* DNA gains in 3q occur frequently in squamous cell carcinoma of the lung, but not in adenocarcinoma. *Genes Chromosomes Cancer* **22**, 79–82 (1998).

149. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823–826 (2013).

150. Ding, Q. *et al.* Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell* **12**, 393–394 (2013).

151. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281–2308 (2013).

152. Ahier, A. & Jarriault, S. Simultaneous Expression of Multiple Proteins Under a Single Promoter in *Caenorhabditis elegans* via a Versatile 2A-Based Toolkit. *Genetics* **196**, 605–613 (2014).

153. Zhang, J.-P. *et al.* Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biology* **18**, 35 (2017).

154. Kılıç, Y., Çelebiler, A. Ç. & Sakızlı, M. Selecting housekeeping genes as references for the normalization of quantitative PCR data in breast cancer. *Clinical and Translational Oncology* **16**, 184–190 (2014).

155. Lyng, M. B., Lænkholm, A.-V., Pallisgaard, N. & Ditzel, H. J. Identification of genes for normalization of real-time RT-PCR data in breast carcinomas. *BMC Cancer* **8**, (2008).

156. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

157. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

158. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

159. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284 (2014).

160. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71–e71 (2016).

161. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

162. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* **14**, 959–962 (2017).

163. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

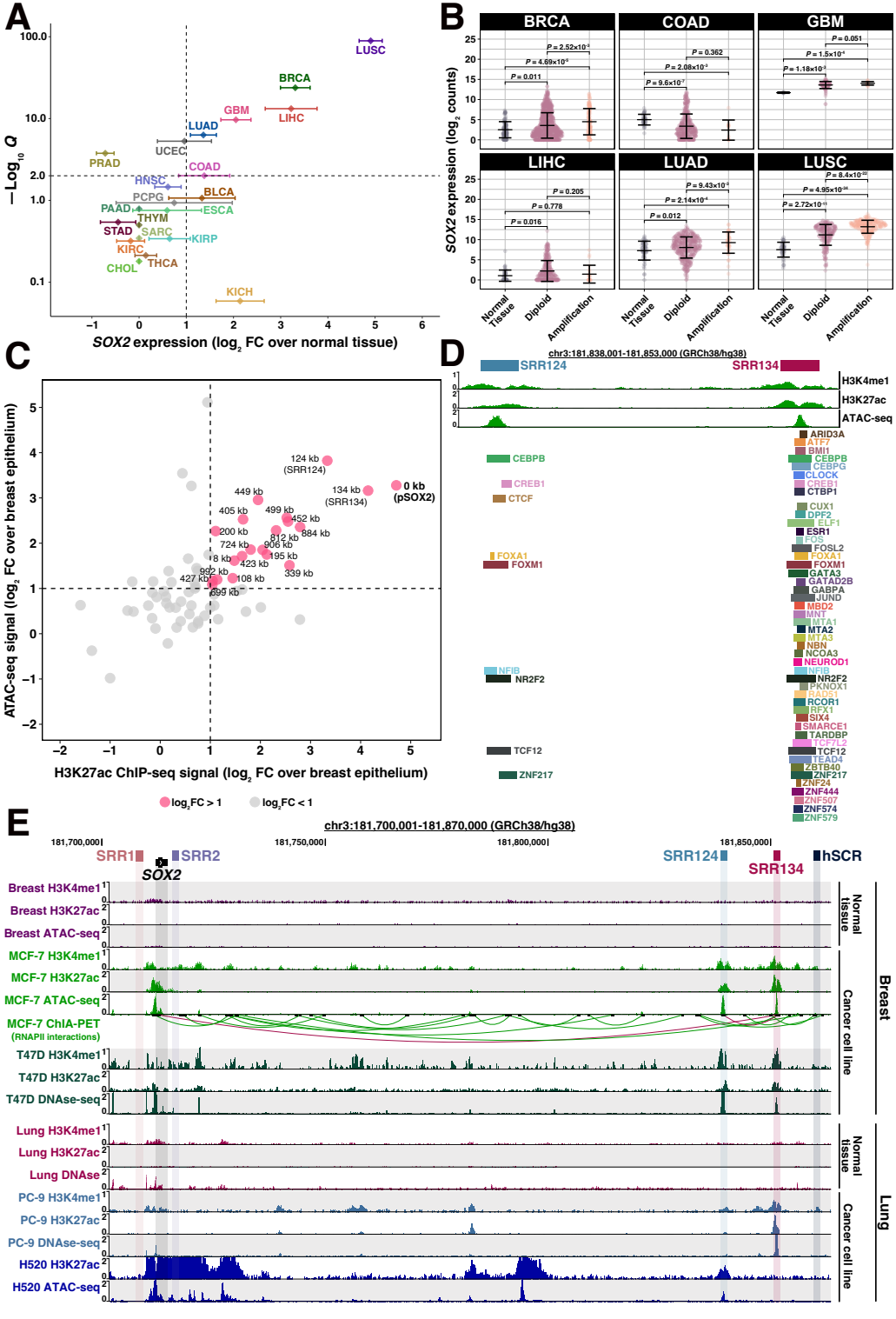
164. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, (2010).

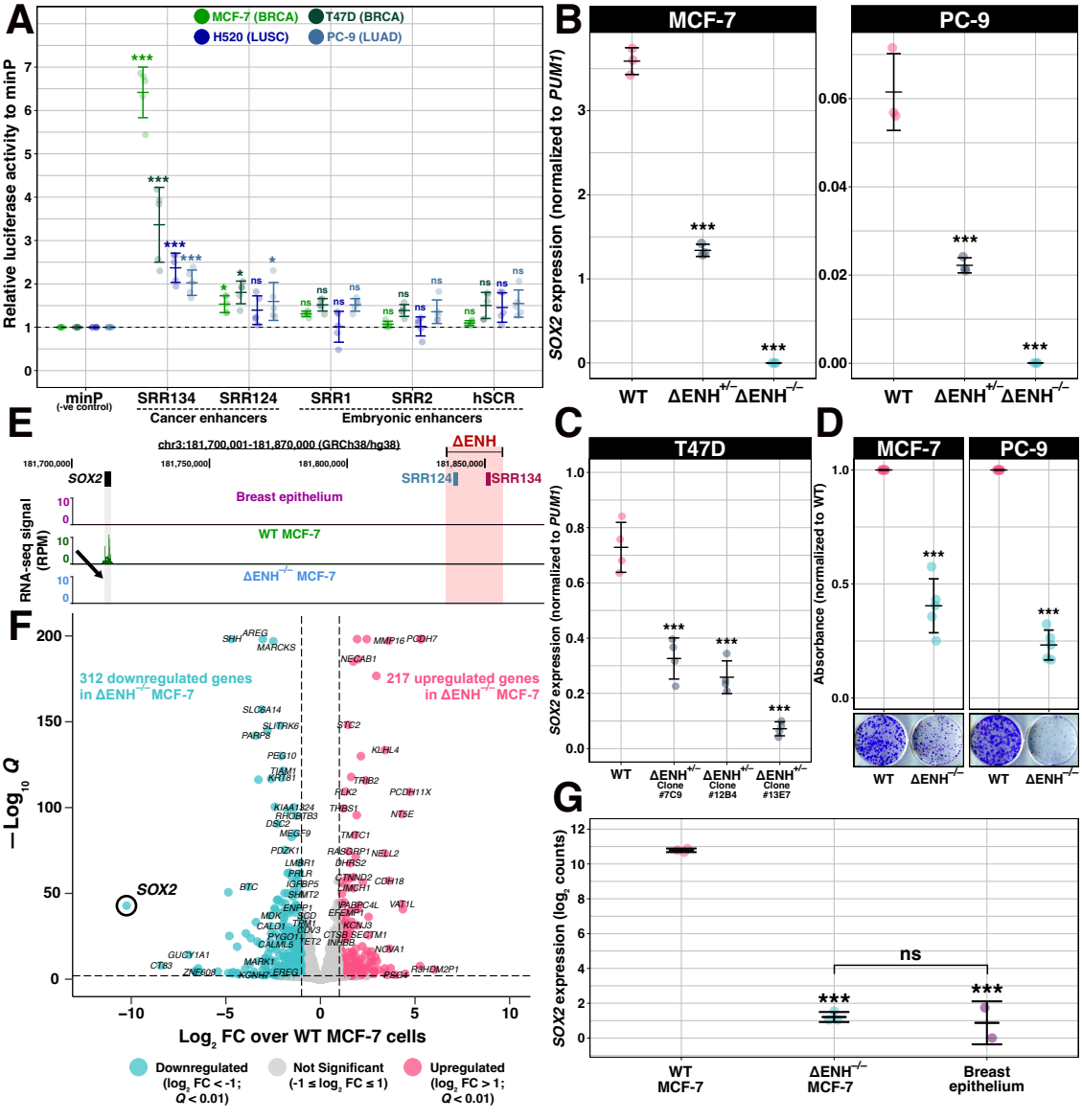
165. Seibt, K. M., Schmidt, T. & Heitkam, T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics* **34**, 3575–3577 (2018).

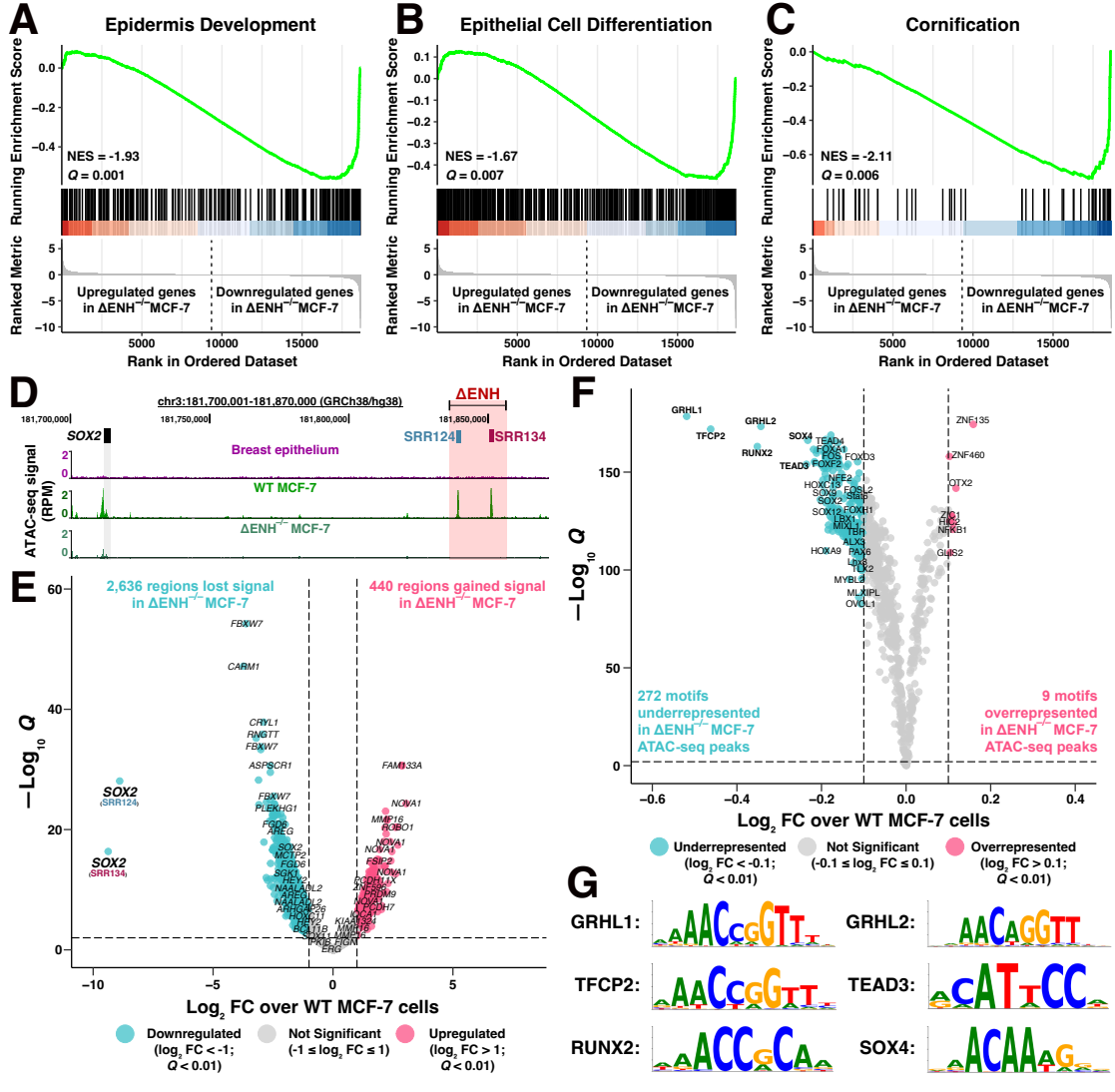
166. Gertsenstein, M. & Nutter, L. M. J. Production of knockout mouse lines with Cas9. *Methods* **191**, 32–43 (2021).

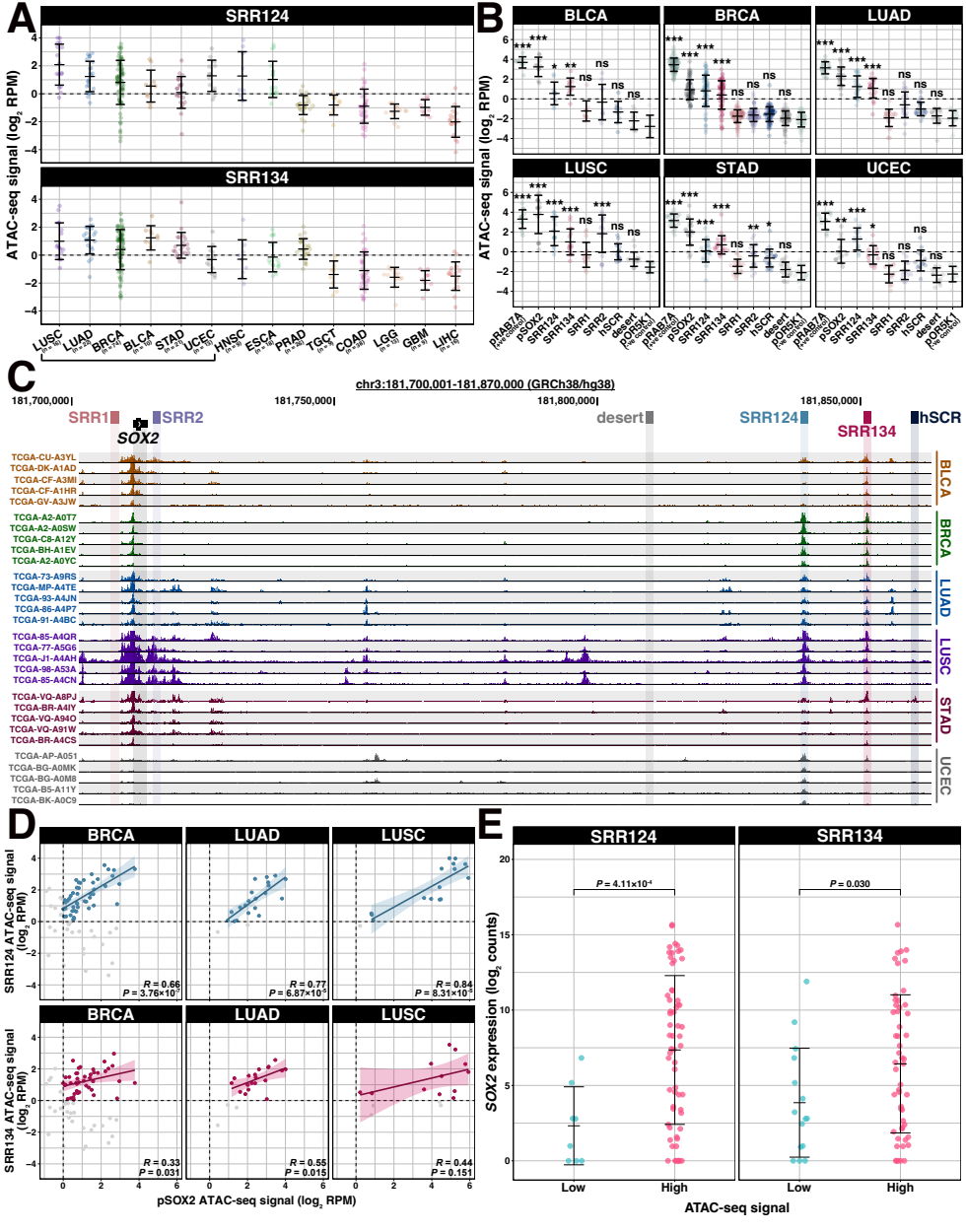
167. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457–481 (1958).

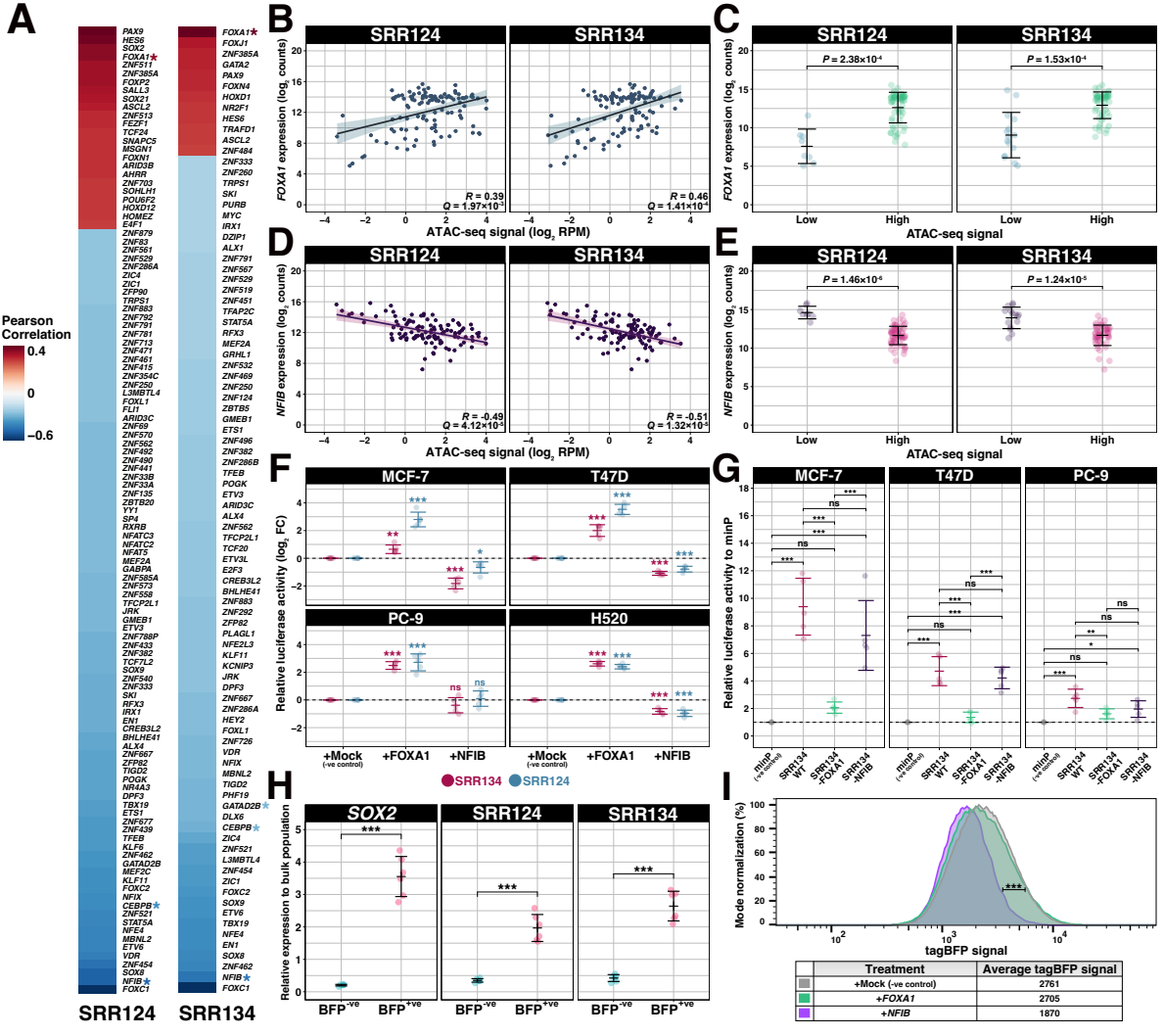
168. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).

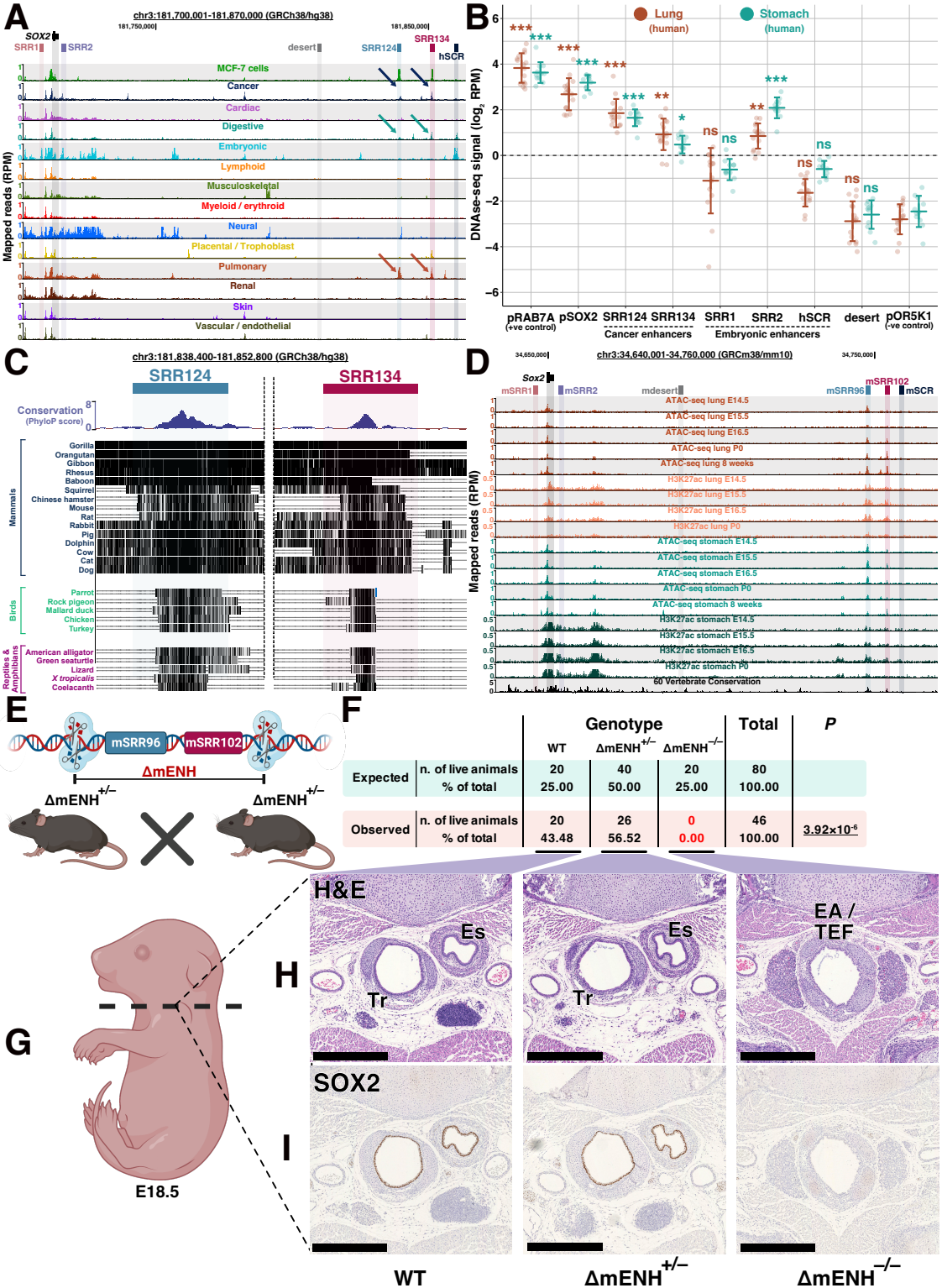

A
B
C
D
E
SOX2 expression (log2 FC over normal tissue)
−Log10 Q
BRCA
COAD
GBM
LIHC
LUAD
LUSC
SOX2 expression (log2 counts)
Normal Tissue
Diploid
Amplification
ATAC-seq signal (log2 FC over breast epithelium)
H3K27ac ChIP-seq signal (log2 FC over breast epithelium)
log2FC > 1
log2FC < 1
chr3:181,838,001-181,853,000 (GRCh38/hg38)
SRR124
SRR134
H3K4me1
H3K27ac
ATAC-seq
chr3:181,700,001-181,870,000 (GRCh38/hg38)
SRR1
SRR2
SOX2
hSCR
Breast H3K4me1
Breast H3K27ac
Breast ATAC-seq
MCF-7 H3K4me1
MCF-7 H3K27ac
MCF-7 ATAC-seq
MCF-7 ChIA-PET
(RNAPII interactions)
T47D H3K4me1
T47D H3K27ac
T47D DNAse-seq
Lung H3K4me1
Lung H3K27ac
Lung DNAse
PC-9 H3K4me1
PC-9 H3K27ac
PC-9 DNAse-seq
H520 H3K27ac
H520 ATAC-seq
Normal tissue
Cancer cell line
Breast
Lung


A
MCF-7 (BRCA)
T47D (BRCA)
H520 (LUSC)
PC-9 (LUAD)
Relative luciferase activity to minP
minP
(-ve control)
SRR134
SRR124
SRR1
SRR2
hSCR
Cancer enhancers
Embryonic enhancers
B
MCF-7
PC-9
SOX2 expression (normalized to PUM1)
WT
ΔENH+/−
ΔENH−/−
C
T47D
SOX2 expression (normalized to PUM1)
Clone #7C9
Clone #12B4
Clone #13E7
D
MCF-7
PC-9
Absorbance (normalized to WT)
E
chr3:181,700,001-181,870,000 (GRCh38/hg38)
ΔENH
SOX2
SRR124
SRR134
Breast epithelium
WT MCF-7
ΔENH−/− MCF-7
RNA-seq signal (RPM)
F
312 downregulated genes in ΔENH−/− MCF-7
217 upregulated genes in ΔENH−/− MCF-7
SOX2
− Log10 Q
Log2 FC over WT MCF-7 cells
Downregulated (log2 FC < -1; Q < 0.01)
Not Significant (-1 ≤ log2 FC ≤ 1)
Upregulated (log2 FC > 1; Q < 0.01)
G
SOX2 expression (log2 counts)
ns
WT MCF-7
ΔENH−/− MCF-7
Breast epithelium


A
Epidermis Development
NES = -1.93
Q = 0.001
B
Epithelial Cell Differentiation
NES = -1.67
Q = 0.007
C
Cornification
NES = -2.11
Q = 0.006
Running Enrichment Score
Ranked Metric
Upregulated genes in ΔENH−/− MCF-7
Downregulated genes in ΔENH−/− MCF-7
Rank in Ordered Dataset
D
chr3:181,700,001-181,870,000 (GRCh38/hg38)
ΔENH
SOX2
Breast epithelium
SRR124
SRR134
WT MCF-7
ΔENH−/− MCF-7
ATAC-seq signal (RPM)
E
2,636 regions lost signal in ΔENH−/− MCF-7
440 regions gained signal in ΔENH−/− MCF-7
−Log10 Q
Log2 FC over WT MCF-7 cells
Downregulated (log2 FC < -1; Q < 0.01)
Not Significant (-1 ≤ log2 FC ≤ 1)
Upregulated (log2 FC > 1; Q < 0.01)
F
272 motifs underrepresented in ΔENH−/− MCF-7 ATAC-seq peaks
9 motifs overrepresented in ΔENH−/− MCF-7 ATAC-seq peaks
Log2 FC over WT MCF-7 cells
Underrepresented (log2 FC < -0.1; Q < 0.01)
Not Significant (-0.1 ≤ log2 FC ≤ 0.1)
Overrepresented (log2 FC > 0.1; Q < 0.01)
G
GRHL1:
GRHL2:
TFCP2:
TEAD3:
RUNX2:
SOX4:


A
SRR124
SRR134
ATAC-seq signal (log2 RPM)
LUSC
LUAD
BRCA
BLCA
STAD
UCEC
HNSC
ESCA
PRAD
TGCT
COAD
LGG
GBM
LIHC
B
BLCA
BRCA
LUAD
LUSC
STAD
UCEC
ATAC-seq signal (log2 RPM)
pRAB7A (ATAC-seq)
pSOX2
SRR124
SRR134
SRR1
SRR2
hSCR
desert
pOR5K1 (ATAC-seq)
C
chr3:181,700,001-181,870,000 (GRCh38/hg38)
181,700,000
181,750,000
181,800,000
181,850,000
SRR1
SRR2
SOX2
desert
SRR124
SRR134
hSCR
TCGA-CU-A3YL
TCGA-DK-A1AD
TCGA-CF-A3MI
TCGA-CF-A1HR
TCGA-GV-A3JW
TCGA-A2-A0T7
TCGA-A2-A0SW
TCGA-C8-A12Y
TCGA-BH-A1EV
TCGA-A2-A0YC
TCGA-73-A9RS
TCGA-MP-A4TE
TCGA-93-A4JN
TCGA-86-A4P7
TCGA-91-A4BC
TCGA-85-A4QR
TCGA-77-A5G6
TCGA-J1-A4AH
TCGA-98-A53A
TCGA-85-A4CN
TCGA-VQ-A8PJ
TCGA-BR-A4IY
TCGA-VQ-A94O
TCGA-VQ-A91W
TCGA-BR-A4CS
TCGA-AP-A051
TCGA-BG-A0MK
TCGA-BG-A0M8
TCGA-B5-A11Y
TCGA-BK-A0C9
BLCA
BRCA
LUAD
LUSC
STAD
UCEC
D
BRCA
LUAD
LUSC
SRR124 ATAC-seq signal (log2 RPM)
R = 0.66
P = 3.76×10-7
R = 0.77
P = 6.87×10-5
R = 0.84
P = 8.31×10-5
SRR134 ATAC-seq signal (log2 RPM)
R = 0.33
P = 0.031
R = 0.55
P = 0.015
R = 0.44
P = 0.151
pSOX2 ATAC-seq signal (log2 RPM)
E
SRR124
SRR134
P = 4.11×10-4
P = 0.030
SOX2 expression (log2 counts)
Low
High
ATAC-seq signal


A
Pearson Correlation
0.4
0
−0.6
SRR124
SRR134
B
SRR124
SRR134
FOXA1 expression (log2 counts)
R = 0.39
Q = 1.97×10−3
R = 0.46
Q = 1.41×10−4
ATAC-seq signal (log2 RPM)
C
SRR124
SRR134
P = 2.38×10−4
P = 1.53×10−4
FOXA1 expression (log2 counts)
Low
High
ATAC-seq signal
D
SRR124
SRR134
NFIB expression (log2 counts)
R = -0.49
Q = 4.12×10−3
R = -0.51
Q = 1.32×10−3
ATAC-seq signal (log2 RPM)
E
SRR124
SRR134
P = 1.46×10−6
P = 1.24×10−5
NFIB expression (log2 counts)
Low
High
ATAC-seq signal
F
MCF-7
T47D
PC-9
H520
Relative luciferase activity (log2 FC)
+Mock (-ve control)
+FOXA1
+NFIB
SRR134
SRR124
G
MCF-7
T47D
PC-9
Relative luciferase activity to minP
minP (-ve control)
SRR134 WT
SRR134 +FOXA1
SRR134 +NFIB
H
SOX2
SRR124
SRR134
Relative expression to bulk population
BFP-ve
BFP+ve
I
Mode normalization (%)
tagBFP signal
Treatment
Average tagBFP signal
+Mock (-ve control)
2761
+FOXA1
2705
+NFIB
1870


A
chr3:181,700,001-181,870,000 (GRCh38/hg38)
SOX2
SRR1
SRR2
desert
SRR124
SRR134
hSCR
MCF-7 cells
Cancer
Cardiac
Digestive
Embryonic
Lymphoid
Musculoskeletal
Myeloid / erythroid
Neural
Placental / Trophoblast
Pulmonary
Renal
Skin
Vascular / endothelial
Mapped reads (RPM)
B
Lung (human)
Stomach (human)
DNAse-seq signal (log2 RPM)
pRAB7A (+ve control)
pSOX2
SRR124
SRR134
Cancer enhancers
SRR1
SRR2
hSCR
Embryonic enhancers
desert
pOR5K1 (-ve control)
C
chr3:181,838,400-181,852,800 (GRCh38/hg38)
SRR124
SRR134
Conservation (PhyloP score)
Mammals
Gorilla
Orangutan
Gibbon
Rhesus
Baboon
Squirrel
Chinese hamster
Mouse
Rat
Rabbit
Pig
Dolphin
Cow
Cat
Dog
Birds
Parrot
Rock pigeon
Mallard duck
Chicken
Turkey
Reptiles & Amphibians
American alligator
Green seaturtle
Lizard
X tropicalis
Coelacanth
D
chr3:34,640,001-34,760,000 (GRCm38/mm10)
Sox2
mSRR1
mSRR2
mdesert
mSRR96
mSRR102
mSCR
ATAC-seq lung E14.5
ATAC-seq lung E15.5
ATAC-seq lung E16.5
ATAC-seq lung P0
ATAC-seq lung 8 weeks
H3K27ac lung E14.5
H3K27ac lung E15.5
H3K27ac lung E16.5
H3K27ac lung P0
ATAC-seq stomach E14.5
ATAC-seq stomach E15.5
ATAC-seq stomach E16.5
ATAC-seq stomach P0
ATAC-seq stomach 8 weeks
H3K27ac stomach E14.5
H3K27ac stomach E15.5
H3K27ac stomach E16.5
H3K27ac stomach P0
60 Vertebrate Conservation
Mapped reads (RPM)
E
mSRR96
mSRR102
ΔmENH
ΔmENH+/–
ΔmENH+/–
F
Genotype
WT
ΔmENH+/–
ΔmENH–/–
Total
P
Expected
n. of live animals 20 40 20 80
% of total 25.00 50.00 25.00 100.00
Observed
n. of live animals 20 26 0 46
% of total 43.48 56.52 0.00 100.00
3.92×10-6
G
E18.5
H
I
H&E
Es
Tr
EA / TEF
SOX2
WT
ΔmENH+/–
ΔmENH–/–