TITLE PAGE

Full Title: Epigenetic reprogramming 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[1,8]

[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

[8]Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada

24 Corresponding authors:

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

26 Current address:

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

ABSTRACT

Enhancer reprogramming has been proposed as a key source of transcriptional dysregulation during tumorigenesis, but the molecular mechanisms underlying this process remain unclear. Here, we identify an enhancer cluster required for normal development that is aberrantly activated in breast and lung carcinoma. Deletion of the SRR124–134 cluster disrupts transcription of the *SOX2* oncogene and dysregulates genome-wide chromatin accessibility in cancer cells. Analysis of primary tumors reveals a correlation between chromatin accessibility at this cluster and *SOX2* overexpression in breast and lung cancer patients. We demonstrate that FOXA1 is an activator and NFIB is a repressor of SRR124–134 activity and *SOX2* transcription in cancer cells, revealing a co-opting of the regulatory mechanisms involved in early development. Notably, we show that the conserved SRR124 and SRR134 regions are essential during mouse development, where homozygous deletion results in the lethal failure of esophageal-tracheal separation. These findings provide insights into how developmental enhancers can be reprogrammed during tumorigenesis and underscore the importance of understanding enhancer dynamics during development and disease.

INTRODUCTION

Developmental enhancers are commissioned during early embryogenesis, as transcription factors progressively restrict the epigenome through the repression of regulatory regions associated with pluripotency (Zhu et al. 2013; Hawkins et al. 2010) and the activation of enhancers that control the expression of lineage-specific developmental genes (Rada-Iglesias et al. 2011; Creyghton et al. 2010; Rubin et al. 2017). This establishes a cell type-specific epigenetic regulatory "memory" that maintains cell lineage commitment and reinforces transcriptional programs (Stergachis et al. 2013). As cells mature and development ends, developmental-associated enhancers are decommissioned, and the enhancer landscape becomes highly restrictive and developmentally stable (Stergachis et al. 2013). This landscape, however, becomes profoundly disturbed during tumorigenesis, as cancer cells aberrantly acquire euchromatin features at regions near oncogenes (Hnisz et al. 2013; Lovén et al. 2013) that are often associated with earlier stages of cell lineage specification (Stergachis et al. 2013). This "enhancer reprogramming" has been proposed to result in a dysfunctional state that causes widespread abnormal gene expression and cellular plasticity (Fu et al. 2019; Roe et al. 2017; Bi et al. 2020; Pomerantz et al. 2015; Lupien et al. 2008). Although the misactivation of enhancers has been suggested as a major source of transcriptional dysregulation (reviewed in Richart et al. 2021; Okabe and Kaneda 2021), it remains largely unclear how this mechanism unfolds during the progression of cancer. To study this process, we evaluated cis-regulatory elements involved in driving transcription during normal development and disease.

66 SRY-box transcription factor 2 (SOX2) is a pioneer transcription factor required for
67 pluripotency maintenance in embryonic stem cells (Avilion 2003; Chew et al. 2005),
68 involved in reprogramming differentiated cells to induced pluripotent stem cells in
69 mammals (Takahashi and Yamanaka 2006; Takahashi et al. 2007; Yu et al. 2007), and
70 acts as an oncogene in several different types of cancer (reviewed in Wuebben and
71 Rizzino 2017; Novak et al. 2019). During later development, SOX2 is also required for
72 tissue morphogenesis and homeostasis of the brain (Graham et al. 2003), eyes
73 (Taranova et al. 2006), esophagus (Que et al. 2007), inner ear (Kiernan et al. 2005), lungs
74 (Gontan et al. 2008), skin (Driskell et al. 2009), stomach (Francis et al. 2019), taste buds
75 (Okubo et al. 2006) and trachea (Que et al. 2009) in both human and mouse. In these
76 tissues, *SOX2* expression is regulated precisely in space and time at critical stages of
77 development, although in most cases the cis-regulatory regions that mediate this
78 precision remain unknown. For example, proper levels of *Sox2* expression are required
79 for the complete separation of the anterior foregut into the esophagus and trachea in mice
80 (Que et al. 2007; Teramoto et al. 2020; Chakraborty et al. 2023) and in humans (Zenteno
81 et al. 2006; Williamson et al. 2006; Chassaing et al. 2007), as the disruption of *Sox2*
82 expression leads to an abnormal developmental condition known as esophageal atresia
83 with distal tracheoesophageal fistula (EA/TEF) (reviewed in Brunner and van Bokhoven
84 2005; Que et al. 2006). After the anterior foregut is properly separated, *Sox2* expression
85 ranges from the esophagus to the stomach in the gut (Que et al. 2007; Francis et al.
86 2019), and throughout the trachea, bronchi, and upper portion of the lungs in the
87 developing airways (Que et al. 2009). Proper branching morphogenesis at the tip of the
88 lungs, however, requires temporary downregulation of *Sox2*, followed by reactivation after

89 lung bud establishment (Gontan et al. 2008). *Sox2* also retains an essential function in
90 multiple mature epithelial tissues, where it is highly expressed in proliferative and self-
91 renewing adult stem cells necessary for maintaining and replacing terminally
92 differentiated cells within the epithelium of the brain, bronchi, esophagus, stomach, and
93 trachea (Arnold et al. 2011; Que et al. 2009; Tompkins et al. 2009; Francis et al. 2019).
94 However, the expression of *SOX2* becomes repressed within terminally differentiated
95 epithelial cells in these tissues (Arnold et al. 2011).

96 As an oncogene, overexpression of *SOX2* is linked to increased cellular replication
97 rates, aggressive tumor grades, and poor patient outcomes in breast carcinoma (BRCA)
98 (Chen et al. 2008; Leis et al. 2012; Liu et al. 2018b; Meng et al. 2020; Piva et al. 2014);
99 colon adenocarcinoma (COAD) (Takeda et al. 2018; Talebi et al. 2015; Zhang et al.
100 2020b; Zhu et al. 2021); glioblastoma (GBM) (Alonso et al. 2011; Cox et al. 2012;
101 Gangemi et al. 2009; Hägerstrand et al. 2011); liver hepatocellular carcinoma (LIHC) (Sun
102 et al. 2013); lung adenocarcinoma (LUAD) (Chou et al. 2013; Sholl et al. 2010;
103 Nakatsugawa et al. 2011); and lung squamous cell carcinoma (LUSC) (Bass et al. 2009;
104 Hussenet et al. 2010). These clinical and molecular characteristics arise from the
105 participation of *SOX2* in the formation and maintenance of tumor-initiating cells that
106 resemble tissue progenitor cells, as evidenced by BRCA (Domenici et al. 2019; Piva et
107 al. 2014; Simões et al. 2011), GBM (Bulstrode et al. 2017; Gangemi et al. 2009; Jeon et
108 al. 2011; Zhang et al. 2020a, 24), LUAD (Singh et al. 2012), and LUSC (Boumahdi et al.
109 2014) studies. *SOX2* knockdown, on the other hand, often results in diminished levels of
110 cell replication, invasion, and treatment resistance in these tumor types (Bass et al. 2009;
111 Berezovsky et al. 2014; Chen et al. 2008; Chou et al. 2013; Fang et al. 2011;

112 Nakatsugawa et al. 2011; Piva et al. 2014; Stolzenburg et al. 2012). Despite the
113 involvement of *SOX2* in the progression of multiple types of cancer, little is known about
114 the mechanisms that cause *SOX2* overexpression during tumorigenesis. Two proximal
115 enhancers were once deemed crucial for driving *Sox2* expression during early
116 development: *Sox2* Regulatory Region 1 (SRR1) and SRR2 (Ferri et al. 2004; Tomioka
117 et al. 2002; Zappone et al. 2000). Deletion of SRR1 and SRR2, however, has no effect
118 on *Sox2* expression in mouse embryonic stem cells (Zhou et al. 2014a). In contrast,
119 deletion of a distal *Sox2* Control Region (SCR), 106 kb downstream of the *Sox2* promoter,
120 causes a profound loss of *Sox2* expression in mouse embryonic stem cells (Li et al. 2014;
121 Zhou et al. 2014a) and in blastocysts, where SCR deletion causes peri-implantation
122 lethality (Chakraborty et al. 2023). The contribution of these regulatory regions in driving
123 *SOX2* expression during tumorigenesis, however, remains poorly defined.

124 Here, we investigated the mechanisms underlying *SOX2* overexpression in cancer.
125 We found that, in breast and lung cancer, *SOX2* is driven by a novel developmental
126 enhancer cluster we termed SRR124–134, rather than the previously identified SRR1,
127 SRR2, or the SCR. This novel distal cluster contains two regions located 124 and 134 kb
128 downstream of the *SOX2* promoter that drive transcription in breast and lung cancer cells.
129 Deletion of this cluster results in significant *SOX2* downregulation, leading to genome-
130 wide changes in chromatin accessibility and a globally disrupted transcriptome. The
131 SRR124–134 cluster is highly accessible in most breast and lung patient tumors, where
132 chromatin accessibility at these regions is correlated with *SOX2* overexpression and is
133 regulated positively by FOXA1 and negatively by NFIB. Finally, we found that both
134 SRR124 and SRR134 are highly conserved in the mouse and are essential for postnatal

135 survival, as homozygous deletion of their homologous regions results in lethal EA/TEF.
136 These findings serve as a prime example of how different types of cancer cells reprogram
137 enhancers that were decommissioned during development to drive the expression of
138 oncogenes during tumorigenesis.

139 **RESULTS**

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

141 *SOX2* overexpression occurs in multiple types of cancer (reviewed in Wuebben and
142 Rizzino 2017; Novak et al. 2019). To examine which cancer types have the highest levels
143 of *SOX2* upregulation, we performed differential expression analysis by calculating the
144 $\log_2$ fold change ($\log_2$ FC) of *SOX2* transcription from 21 TCGA primary solid tumors (see
145 Supplementary Table S1 for cancer type abbreviations) compared to normal tissue
146 samples (Hoadley et al. 2014). We found that BRCA ($\log_2$ FC = 3.31), COAD ($\log_2$ FC =
147 1.38), GBM ($\log_2$ FC = 2.05), LIHC ($\log_2$ FC = 3.22), LUAD ($\log_2$ FC = 1.36), and LUSC
148 ($\log_2$ FC = 4.91) tumors had the greatest *SOX2* upregulation ($\log_2$ FC > 1; FDR-adjusted
149 $Q < 0.01$; Figure 1A, Supplementary Table S2). As a negative control, we ran this same
150 analysis using the housekeeping gene *PUM1* (Krasnov et al. 2019) and found no cancer
151 types with significant upregulation of this gene (Supplementary Figure S1A,
152 Supplementary Table S3).

153 Next, we divided BRCA, COAD, GBM, LIHC, LUAD, and LUSC patients (n = 3,064)
154 into four groups according to their *SOX2* expression. Gene expression levels were
155 measured by RNA-seq counts normalized by library size and transformed to a $\log_2$ scale,
156 hereinafter referred to as $\log_2$ counts. Cancer patients within the top group (25% highest
157 *SOX2* expression; $\log_2$ counts > 10.06) have a significantly ($P = 1.27\times10^{-23}$, log-rank test)
158 lower overall probability of survival compared to cancer patients within the bottom group
159 (25% lowest *SOX2* expression; $\log_2$ counts < 1.68) (Supplementary Figure S1B,
160 Supplementary Table S4). We also examined the relationship between *SOX2* copy
161 number and *SOX2* overexpression within these six tumor types. Although previous

162 studies have shown that *SOX2* is frequently amplified in squamous cell carcinoma (Bass
163 et al. 2009; Hussenet et al. 2010; Maier et al. 2011; Liu et al. 2021), we found that most
164 BRCA (88%), COAD (98%), GBM (91%), LIHC (94%), and LUAD (92%) tumors were
165 diploid for *SOX2*. In addition, BRCA (P = 0.011, Holm-adjusted Dunn's test), GBM (P =
166 1.18×10^{-3}), LIHC (P = 0.016), LUAD (P = 0.012), and LUSC (P = 2.72×10^{-11}) diploid
167 tumors significantly overexpressed *SOX2* compared to normal tissue (Figure 1B,
168 Supplementary Table S5). This indicates that gene amplification is dispensable for driving
169 *SOX2* overexpression in most cancer types.

170 We investigated whether the *SOX2* locus gains epigenetic features associated with
171 active enhancers in cancer cells. Enhancer features commonly include accessible
172 chromatin determined by either Assay for Transposase Accessible Chromatin with high-
173 throughput sequencing (ATAC-seq) (Buenrostro et al. 2013) or DNase I hypersensitive
174 sites sequencing (DNase-seq) (Boyle et al. 2008), and histone modifications including
175 histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation
176 (H3K27ac) (Heintzman et al. 2007, 2009). To study gains in enhancer features within the
177 *SOX2* locus, we initially focused our analyses on luminal A breast cancer, the most
178 common subtype of BRCA to significantly (P = 0.021, Tukey's test) overexpress *SOX2*
179 (Supplementary Figure S1C) (Hoadley et al. 2014; Berger et al. 2018). MCF-7 cells are a
180 widely used $ER^+/PR^+/HER2^-$ luminal A breast adenocarcinoma model (Soule et al. 1973),
181 which have been previously described to overexpress *SOX2* (Chen et al. 2008; Liang et
182 al. 2013; Ling et al. 2012; Stolzenburg et al. 2012). After confirming that *SOX2* is one of
183 the most upregulated genes in MCF-7 cells ($\log_2$ FC = 10.75; FDR-adjusted Q = $2.20\times10^{-}$
184 ^{36}; Supplementary Figure S1D, Supplementary Table S6) compared to normal breast

185 epithelium (Roadmap Epigenomics Consortium et al. 2015), we contrasted their
186 chromatin accessibility and histone modifications (ENCODE Project Consortium 2012).
187 By intersecting 1,500 bp regions that contain at least 500 bp overlap between H3K27ac
188 and ATAC-seq peaks, we found that 19 putative enhancers gained ($\log_2$ FC > 1) both
189 these features within ± 1 Mb from the *SOX2* transcription start site (TSS) in MCF-7 cells
190 (Figure 1C, Supplementary Table S7). Besides the *SOX2* promoter (pSOX2), we
191 identified a downstream cluster containing two regions that have gained the highest
192 ATAC-seq and H3K27ac signal in MCF-7 cells: SRR124 (124 kb downstream of pSOX2)
193 and SRR134 (134 kb downstream of pSOX2). The previously described SRR1, SRR2
194 (Ferri et al. 2004; Tomioka et al. 2002; Zappone et al. 2000), and the human ortholog of
195 the mouse SCR (hSCR) (Li et al. 2014; Zhou et al. 2014a), however, lacked substantial
196 gains in enhancer features within MCF-7 cells.

197 Alongside gains in chromatin features, another characteristic of active enhancers is
198 the binding of numerous (> 10) transcription factors (Chen et al. 2012; Mitchell et al. 2012;
199 Singh et al. 2021). Chromatin Immunoprecipitation Sequencing (ChIP-seq) data from
200 ENCODE (ENCODE Project Consortium 2012) on 117 transcription factors revealed 48
201 different factors present at the SRR124–134 cluster in MCF-7 cells, with the majority (47)
202 of these factors present at SRR134 (Figure 1D). Transcription factors bound at both
203 SRR124 and SRR134 include CEBPB, CREB1, FOXA1, FOXM1, NFIB, NR2F2, TCF12,
204 and ZNF217. An additional feature of distal enhancers is that they contact their target
205 genes through long-range chromatin interactions (Tolhuis et al. 2002; Carter et al. 2002).
206 Chromatin Interaction Analysis by Paired-End-Tag sequencing (ChIA-PET) (Fullwood et
207 al. 2009) showed two interesting RNA Polymerase II (RNAPII)-mediated chromatin

208 interactions in MCF-7 cells: one between the *SOX2* gene and SRR134, and one between
209 SRR124 and SRR134 (Figure 1E). Beyond MCF-7 cells, we found that H520 (LUSC),
210 PC-9 (LUAD), and T47D (luminal A BRCA) cancer cell lines, which display varying levels
211 of *SOX2* expression (Supplementary Figure S1E), also gained substantial enhancer
212 features at SRR124 and SRR134 when compared to normal tissue (Figure 1E) (ENCODE
213 Project Consortium 2012; Chan et al. 2018; Sato et al. 2019; Gopi and Kidder 2021).
214 Together, these data suggest SRR124 and SRR134 could be active enhancers driving
215 *SOX2* transcription in BRCA, LUAD, and LUSC.
216
217 **Figure 1: A cluster 124–134 kilobases downstream of *SOX2* gains enhancer**
218 **features in cancer cells. (A)** Super-logarithmic RNA-seq volcano plot of *SOX2*
219 expression from 21 cancer types compared to normal tissue (Hoadley et al. 2014). Cancer
220 types with $\log_2$ FC > 1 and FDR-adjusted $Q < 0.01$ were considered to significantly
221 overexpress *SOX2*. Error bars: standard deviation. **(B)** *SOX2* $\log_2$-normalized expression
222 ($\log_2$ counts) associated with the *SOX2* copy number from BRCA (n = 1174), COAD (n =
223 483), GBM (n = 155), LIHC (n = 414), LUAD (n = 552), and LUSC (n = 546) patient tumors
224 (Hoadley et al. 2014). RNA-seq reads were normalized to library size using DESeq2
225 (Love et al. 2014). Error bars: standard deviation. Significance analysis by Dunn's test
226 (Dunn 1964) with Holm correction (Holm 1979). **(C)** 1,500 bp genomic regions within ± 1
227 Mb from the *SOX2* transcription start site (TSS) that gained enhancer features in MCF-7
228 cells (ENCODE Project Consortium 2012) compared to normal breast epithelium
229 (Roadmap Epigenomics Consortium et al. 2015). Regions that gained both ATAC-seq
230 and H3K27ac ChIP-seq signal above our threshold ($\log_2$ FC > 1, dashed line) are

231 highlighted in pink. Each region was labelled according to their distance in kilobases (kb)
232 to the *SOX2* promoter (pSOX2, bolded). **(D)** ChIP-seq signal for H3K4me1 and H3K27ac,
233 ATAC-seq signal, and transcription factor ChIP-seq peaks at the SRR124–134 cluster in
234 MCF-7 cells. Datasets from ENCODE (ENCODE Project Consortium 2012)**. (E)** UCSC
235 Genome Browser (Kent et al. 2002) display of H3K4me1 and H3K27ac ChIP-seq signal,
236 DNAse-seq and ATAC-seq chromatin accessibility signal, and ChIA-PET RNA
237 Polymerase II (RNAPII) interactions around the *SOX2* gene within breast (normal tissue
238 and 2 BRCA cancer cell lines) and lung (normal tissue, one LUAD, and one LUSC cancer
239 cell lines) samples (ENCODE Project Consortium 2012; Chan et al. 2018; Sato et al.
240 2019; Gopi and Kidder 2021). Relevant RNAPII interactions (between SRR124 and
241 SRR134, and between SRR134 and pSOX2) are highlighted in maroon.
242

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

244 To assess SRR124 and SRR134 enhancer activity alongside the embryonic-
245 associated SRR1, SRR2, and hSCR regions, we used a reporter vector containing the
246 firefly luciferase gene under the control of a minimal promoter (minP, pGL4.23). We
247 transfected each enhancer construct into the BRCA (MCF-7, T47D), LUAD (PC-9), and
248 LUSC (H520) cell lines and measured luciferase activity as a relative fold change (FC)
249 compared to the empty minP vector. SRR134 demonstrated the strongest enhancer
250 activity, with the MCF-7 (FC = 6.42; $P < 2\times10^{-16}$, Dunnett's test), T47D (FC = 3.36; $P =$
251 9.34×10^{-10}), H520 (FC = 2.37; $P = 1.22\times10^{-6}$), and PC-9 (FC = 2.03; $P = 9.79\times10^{-5}$) cell
252 lines displaying a significant increase in luciferase activity compared to minP (Figure 2A).
253 SRR124 also showed a modest, significant increase in luciferase activity compared to

254 minP in the MCF-7 (FC = 1.53; $P = 4.27\times10^{-2}$), T47D (FC = 1.80; $P = 4.57\times10^{-2}$), and PC-
255 9 (FC = 1.60; $P = 4.27\times10^{-2}$) cell lines. The SRR1, SRR2, and hSCR enhancers, however,
256 showed no significant enhancer activity ($P > 0.05$) in all four cell lines.

257 Although reporter assays can be used to assess enhancer activity, enhancer
258 knockout approaches remain the current gold standard method for enhancer validation
259 (Moorthy et al. 2017; Tobias et al. 2021). To investigate whether the SRR124–134 cluster
260 drives *SOX2* expression in cancer cells, we used CRISPR/Cas9 to delete this cluster from
261 breast (MCF-7, T47D) and lung (H520, PC-9) cancer cell lines. RT-qPCR showed that
262 homozygous SRR124–134 deletion ($\Delta ENH^{-/-}$) causes a profound (> 99.5%) and
263 significant ($P < 0.001$, Dunnett's test) loss of *SOX2* expression in both the MCF-7 and
264 PC-9 cell lines (Figure 2B). Immunoblot analysis confirmed the depletion of the SOX2
265 protein in $\Delta ENH^{-/-}$ MCF-7 cells (Supplementary Figure S2A). Heterozygous SRR124–
266 134 deletion ($\Delta ENH^{+/-}$) also significantly ($P < 0.001$) reduced *SOX2* expression by ~60%
267 in both MCF-7 and PC-9 cells (Figure 2B). Although we were unable to isolate a
268 homozygous deletion clone from T47D cells, multiple independent heterozygous $\Delta ENH^{+/-}$
269 T47D clonal isolates also showed a significant downregulation (>50%; $P < 0.001$) in *SOX2*
270 expression (Figure 2C). Interestingly, we did not find a significant ($P > 0.05$) impact on
271 *SOX2* expression in $\Delta ENH^{+/-}$ or $\Delta ENH^{-/-}$ H520 cells (Supplementary Figure S2B), which
272 indicates that *SOX2* transcription is sustained by a different mechanism in these cells. To
273 assess the impact of the loss of *SOX2* expression in the tumor initiation capacity of
274 enhancer-deleted cells, we performed a colony formation assay with MCF-7 and PC-9
275 $\Delta ENH^{-/-}$ cells. We found that both MCF-7 ($P = 3.53\times10^{-4}$, t-test) and PC-9 ($P = 1.26\times10^{-}$
276 ^{5}) $\Delta ENH^{-/-}$ cells showed a significant decrease (> 50%) in their ability to form colonies

277 compared to WT cells (Figure 2D), further indicating that SRR124–134-driven *SOX2*
278 overexpression is required to sustain high tumor initiation capacity in BRCA and LUAD.
279 Next, we performed total RNA sequencing (RNA-seq) to measure changes in the
280 transcriptome of $\Delta ENH^{-/-}$ MCF-7 cells compared to WT MCF-7 cells. As expected, all
281 three replicates of each genotype clustered together (Supplementary Figure S2C). In
282 addition to *SOX2* downregulation (Figure 2E), differential expression analysis showed a
283 total of 529 genes differentially ($|\log_2 FC| > 1$; FDR-adjusted $Q < 0.01$) expressed in
284 $\Delta ENH^{-/-}$ MCF-7 cells (Figure 2F, Supplementary Table S8). From these, 312 genes
285 significantly lost expression (59%), whereas 217 (41%) genes significantly gained
286 expression in $\Delta ENH^{-/-}$ MCF-7 cells compared to WT MCF-7 cells (Supplementary Figure
287 S2D). *SOX2* was the gene with the highest loss in expression ($\log_2 FC = -10.24$; $Q =$
288 1.23×10^{-43}) in $\Delta ENH^{-/-}$ MCF-7 cells, followed by *CT83* ($\log_2 FC = -8.43$; $Q = 1.07\times10^{-8}$),
289 and *GUCY1A1* ($\log_2 FC = -6.96$; $Q = 5.09\times10^{-15}$). On the other hand, genes with the most
290 significant gain in expression within $\Delta ENH^{-/-}$ MCF-7 cells included the protocadherins
291 *PCDH7* ($\log_2 FC = 5.34$; $Q < 1\times10^{-200}$), *PCDH10* ($\log_2 FC = 5.29$; $Q < 1\times10^{-200}$), and
292 *PCDH11X* ($\log_2 FC = 4.73$; $Q = 9.29\times10^{-110}$). In addition, deletion of the SRR124–134
293 cluster reduced *SOX2* expression back to the levels found in normal breast epithelium (P
294 $= 0.48$, Tukey's test) (ENCODE Project Consortium 2012; Roadmap Epigenomics
295 Consortium et al. 2015) (Figure 2G). Together, these data confirm that the SRR124–134
296 cluster drives *SOX2* overexpression in BRCA and LUAD.
297
298 **Figure 2: The SRR124–134 cluster drives *SOX2* overexpression in MCF-7, PC-9, and**
299 **T47D cells. (A)** Enhancer reporter assay comparing luciferase activity driven by the

300 SRR1, SRR2, SRR124, SRR134, and hSCR regions to an empty vector containing only
301 a minimal promoter (minP). Enhancer constructs were assayed in the BRCA (MCF-7,
302 T47D), LUAD (PC-9), and LUSC (H520) cell lines. Dashed line: average activity of minP.
303 Error bars: standard deviation. Significance analysis by Dunnett's test (n = 5; * $P < 0.05$,
304 *** $P < 0.001$, ns: not significant) (Dunnett 1955). **(B)** RT-qPCR analysis of *SOX2*
305 transcript levels in SRR124–134 heterozygous- ($\Delta ENH^{+/-}$) and homozygous- ($\Delta ENH^{-/-}$)
306 deleted MCF-7 (BRCA) and PC-9 (LUAD) clones compared to WT cells. Error bars:
307 standard deviation. Significance analysis by Dunnett's test (n = 3; *** $P < 0.001$). **(C)** RT-
308 qPCR analysis of *SOX2* transcript levels in three independent SRR124–134
309 heterozygous-deleted ($\Delta ENH^{+/-}$) T47D clonal isolates compared to WT cells. Error bars:
310 standard deviation. Significance analysis by Dunnett's test (n = 4; *** $P < 0.001$). **(D)**
311 Crystal violet absorbance (570 nm) from a colony formation assay with WT and $\Delta ENH^{-/-}$
312 MCF-7 and PC-9 cells. Total absorbance was normalized to the average absorbance from
313 WT cells within each cell line. Significance analysis by t-test with Holm correction (n = 5;
314 *** $P < 0.001$). **(E)** UCSC Genome Browser (Kent et al. 2002) view of the SRR124–134
315 cluster deletion in $\Delta ENH^{-/-}$ MCF-7 cells with RNA-seq tracks from normal breast
316 epithelium (Roadmap Epigenomics Consortium et al. 2015), WT and $\Delta ENH^{-/-}$ MCF-7
317 cells. Arrow: reduction in RNA-seq signal at the *SOX2* gene in $\Delta ENH^{-/-}$ MCF-7 cells. **(F)**
318 Volcano plot with DESeq2 (Love et al. 2014, 2) differential expression analysis between
319 $\Delta ENH^{-/-}$ and WT MCF-7 cells. Blue: 312 genes that significantly lost expression ($\log_2$ FC
320 < -1; FDR-adjusted $Q < 0.01$) in $\Delta ENH^{-/-}$ MCF-7 cells. Pink: 217 genes that significantly
321 gained expression ($\log_2$ FC > 1; $Q < 0.01$) in $\Delta ENH^{-/-}$ MCF-7 cells. Grey: 35,891 genes
322 that maintained similar ($-1 \leq \log_2 FC \leq 1$) expression between $\Delta ENH^{-/-}$ and WT MCF-7

323 cells. **(G)** Comparison of *SOX2* transcript levels between WT MCF-7 and either ΔENH$^{-/-}$
324 MCF-7 or normal breast epithelium cells (Roadmap Epigenomics Consortium et al. 2015),
325 and between ΔENH$^{-/-}$ MCF-7 and normal breast epithelium cells. RNA-seq reads were
326 normalized to library size using DESeq2 (Love et al. 2014). Error bars: standard deviation.
327 Significance analysis by Tukey's test (*** $P < 0.001$, ns: not significant) (Tukey 1949).
328

329 **SOX2 regulates pathways associated with epithelium development in luminal A**
330 **BRCA**

331 Because SOX2 regulates cell proliferation and differentiation pathways in other
332 epithelial cells (Tompkins et al. 2009, 2011), we decided to further investigate the
333 molecular function of SOX2 in luminal A BRCA cells by utilizing our *SOX2*-depleted
334 ΔENH$^{-/-}$ MCF-7 cell model. Gene Set Enrichment Analysis (GSEA) showed a significant
335 (FDR-adjusted $Q < 0.05$) depletion of multiple epithelium-associated processes within the
336 transcriptome of ΔENH$^{-/-}$ MCF-7 cells, as indicated by the normalized enrichment score
337 [NES] < 1 (Supplementary Table S9). These processes included epidermis development
338 (NES = -1.93; $Q = 0.001$; Figure 3A), epithelial cell differentiation (NES = -1.67; $Q = 0.007$;
339 Figure 3B), and cornification (NES = -2.11; $Q = 0.006$; Figure 3C). This suggests that
340 SOX2 regulates epithelial development and differentiation in luminal A BRCA cells.

341 SOX2 is a pioneer transcription factor that associates with its motif in
342 heterochromatin (Soufi et al. 2012) and recruits chromatin-modifying complexes (Soufi et
343 al. 2015) in embryonic and reprogrammed stem cells. We performed ATAC-seq in ΔENH^{-}
344 $^{/-}$ MCF-7 cells and compared chromatin accessibility to WT MCF-7 cells to identify
345 genome-wide loci that are dependent on SOX2 to remain accessible in luminal A BRCA.

346 As expected, the ATAC-seq signal from all replicates was highly enriched around gene
347 TSS (Supplementary Figure S3A), with both WT and ΔENH$^{-/-}$ samples having higher
348 chromatin accessibility at the TSS of highly expressed genes (Supplementary Figure S3B
349 and Supplementary Figure S3C). Correlation analysis also confirmed the clustering of all
350 three replicates from each genotype (Supplementary Figure S3D). Including the
351 SRR124–134 cluster and pSOX2 (Figure 3D), a total of 3,076 500-bp regions had
352 significant ($|\log_2 FC| > 1$; FDR-adjusted $Q < 0.01$) changes in chromatin accessibility in
353 ΔENH$^{-/-}$ compared to WT MCF-7 cells (Figure 3E, Supplementary Table S10). Most
354 regions (86%, 2,636 regions) significantly lost chromatin accessibility in ΔENH$^{-/-}$ MCF-7
355 cells and 76% (2,024 regions) of these regions also gained chromatin accessibility in WT
356 MCF-7 compared to normal breast epithelium (Roadmap Epigenomics Consortium et al.
357 2015) (Supplementary Table S11). Together, this indicates that SOX2 has an important
358 role in regulating the chromatin accessibility changes acquired in luminal A BRCA.
359 We used TOBIAS (Bentsen et al. 2020) to analyze changes in transcription factor
360 footprints within ATAC-seq peaks in ΔENH$^{-/-}$ compared to WT MCF-7 cells. From 841
361 vertebrate motifs (Castro-Mondragon et al. 2022), we found a total of 281 motifs with a
362 significant ($|\log_2 FC| > 0.1$; FDR-adjusted $Q < 0.01$) differential binding score (Figure 3F,
363 Supplementary Table S12). Most of these motifs (97%, 272 motifs) were
364 underrepresented within ATAC-seq peaks in ΔENH$^{-/-}$ compared to WT MCF-7 cells,
365 indicating that reduced *SOX2* expression affects the binding of multiple other transcription
366 factors. Among them, the GRHL1 ($\log_2 FC = -0.519$; $Q = 3\times10^{-179}$), TFCP2 ($\log_2 FC =$ -
367 0.462; $Q = 1.03\times10^{-172}$), RUNX2 ($\log_2 FC = -0.352$; $Q = 8.02\times10^{-164}$), GRHL2 ($\log_2 FC =$ -
368 0.343; $Q = 4.43\times10^{-174}$), TEAD3 ($\log_2 FC = -0.235$; $Q = 9.74\times10^{-155}$), and SOX4 ($\log_2 FC$

369 = -0.232; Q = 5.33×10^{-167}) motifs (Figure 3G) had the most reduced binding score in
370 ΔENH$^{-/-}$ MCF-7 cells compared to WT MCF-7 cells. These factors belong to three main
371 JASPAR (Castro-Mondragon et al. 2022) motif clusters: GRHL/TFCP (cluster 33;
372 aaAACAGGTTtcAgtt), RUNX (cluster 60; ttctTGtGGTTttt), TEAD (cluster 2;
373 tccAcATTCCAggcCTTta), and SOX (cluster 8; acggaACAATGgaagTGTT). The SOX
374 cluster also included the SOX2 ($\log_2$ FC = -0.175; Q = 6.61×10^{-139}) motif.
375 Next, we aimed to analyze ChIP-seq data from transcription factors within these
376 motif clusters in MCF-7 cells. We utilized two published datasets: GRHL2 (Cocce et al.
377 2019) and RUNX2 (Jeselsohn et al. 2017). Regions that lost ($\log_2$ FC < -1; Q < 0.01)
378 chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells significantly (P < 2×10^{-16}, hypergeometric
379 test) overlapped regions with binding of either of these transcription factors. Among the
380 2,636 regions that lost chromatin accessibility, 40% (750 regions) also show GRHL2
381 binding (Supplementary Figure S3E), whereas 21% (552 regions) share RUNX2 binding
382 (Supplementary Figure S3F). In addition, we found multiple SOX motifs significantly
383 (FDR-adjusted Q < 0.001) enriched within peaks from both GRHL2 (Supplementary Table
384 S13) and RUNX2 (Supplementary Table S14) ChIP-seq datasets, further suggesting that
385 SOX2 collaborates with GRHL2 and RUNX2 to maintain chromatin accessibility in luminal
386 A BRCA. Expression levels of either *GRHL2* or *RUNX2*, however, were not significantly
387 affected by *SOX2* downregulation in ΔENH$^{-/-}$ MCF-7 cells (-1 ≤ $\log_2$ FC ≤ 1;
388 Supplementary Table S8), indicating that they are not directly regulated by SOX2 at the
389 transcriptional level but may interact at the protein level.
390

391 **Figure 3: *SOX2* downregulation impacts chromatin accessibility in luminal A BRCA.**
392 **(A – C)** Gene Set Enrichment Analysis (GSEA) in the transcriptome of ΔENH$^{-/-}$ compared
393 to WT MCF-7 cells. Genes were ranked according to their change in expression ($\log_2$ FC).
394 A subset of GO terms significantly enriched among downregulated genes in ΔENH$^{-/-}$
395 MCF-7 cells are displayed, indicated by the normalized enrichment score (NES) < 1: **(A)**
396 epidermis development, **(B)** epithelial cell differentiation, and **(C)** cornification. GSEA was
397 performed using clusterProfiler (Yu et al. 2012) with an FDR-adjusted $Q < 0.05$ threshold.
398 Green line: running enrichment score. **(D)** UCSC Genome Browser (Kent et al. 2002) view
399 of the SRR124–134 deletion in ΔENH$^{-/-}$ MCF-7 cells with ATAC-seq tracks from breast
400 epithelium (Roadmap Epigenomics Consortium et al. 2015), WT, and ΔENH$^{-/-}$ MCF-7
401 cells. **(E)** Volcano plot with differential ATAC-seq analysis of ΔENH$^{-/-}$ MCF-7 cells
402 compared to WT. Blue: 2,638 regions that lost ($\log_2$ FC < -1; FDR-adjusted $Q < 0.01$)
403 chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells. Pink: 440 regions that gained ($\log_2$ FC >
404 1; $Q < 0.01$) chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells. Grey: 132,726 regions that
405 retained chromatin accessibility in ΔENH$^{-/-}$ MCF-7 cells ($-1 \leq \log_2$ FC ≤ 1). Regions were
406 labelled with their closest gene within a ± 1 Mb distance threshold. Differential chromatin
407 accessibility analysis was performed using diffBind (Stark R, Brown G 2011). **(F)** Volcano
408 plot with ATAC-seq footprint analysis of differential transcription factor binding in ΔENH^{-}
409 $^{/-}$ MCF-7 cells compared to WT. Blue: 272 underrepresented ($\log_2$ FC < -0.1; FDR-
410 adjusted $Q < 0.01$) motifs in ATAC-seq peaks from ΔENH$^{-/-}$ MCF-7 cells. Pink: 9
411 overrepresented ($\log_2$ FC > 0.1; $Q < 0.01$) motifs in ATAC-seq peaks from ΔENH$^{-/-}$ MCF-
412 7 cells. Grey: 560 motifs with no representative change ($-0.1 \leq \log_2$ FC ≤ 0.1) within ATAC-
413 seq peaks from ΔENH$^{-/-}$ MCF-7 cells. **(G)** Sequence motifs of the top 6 transcription

414 factors with the lowest binding score in ΔENH$^{-/-}$ compared to WT MCF-7 cells: GRHL1,
415 TFCP2, RUNX2, GRHL2, TEAD3, SOX4. Footprint analysis was performed using
416 TOBIAS (Bentsen et al. 2020) utilizing the JASPAR 2022 motif database (Castro-
417 Mondragon et al. 2022).
418

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

421 With the confirmation that the SRR124–134 cluster drives *SOX2* overexpression in
422 the BRCA and LUAD cell lines, we investigated chromatin accessibility at this enhancer
423 cluster within primary tumors isolated from cancer patients. By analyzing the pan-cancer
424 ATAC-seq dataset from TCGA (Corces et al. 2018), we found that SRR124 and SRR134
425 are most accessible within LUSC, LUAD, BRCA, bladder carcinoma (BLCA), stomach
426 adenocarcinoma (STAD), and uterine endometrial carcinoma (UCEC) patient tumors
427 (Figure 4A). We also quantified the ATAC-seq signal at six other regions (genomic
428 coordinates in Supplementary Table S15): the *SOX2* embryonic-associated enhancers
429 (SRR1, SRR2, hSCR), the *SOX2* promoter (pSOX2), a gene regulatory desert with no
430 enhancer features located between the *SOX2* gene and the SRR124–134 cluster
431 (desert), and the promoter of the housekeeping gene *RAB7A* (pRAB7A, positive control).
432 We then compared the chromatin accessibility levels at each of these regions to the
433 promoter of the repressed olfactory gene *OR5K1* (pOR5K1, negative control). Both
434 SRR124 and SRR134 showed significantly increased ($P < 0.05$, Holm-adjusted Dunn's
435 test) chromatin accessibility when compared to pOR5K1 in BLCA (SRR124 $P = 0.014$;
436 SRR134 $P = 1.52\times10^{-3}$; Holm-adjusted Dunn's test), BRCA (SRR124 $P = 1.70\times10^{-20}$;

437 SRR134 $P = 1.03\times10^{-16}$), LUAD (SRR124 $P = 6.76\times10^{-7}$; SRR134 $P = 3.26\times10^{-6}$), LUSC
438 (SRR124 $P = 1.62\times10^{-6}$; SRR134 $P = 7.08\times10^{-4}$), STAD (SRR124 $P = 1.15\times10^{-4}$; SRR134
439 $P = 1.96\times10^{-7}$), and UCEC (SRR124 $P = 3.15\times10^{-5}$; SRR134 $P = 0.025$) patient tumors
440 (Figure 4B).

441 One explanation for increased chromatin accessibility is locus amplification. While
442 LUSC had high levels of chromatin accessibility likely related to previously described
443 *SOX2* amplifications (Bass et al. 2009; Hussenet et al. 2010; Maier et al. 2011; Liu et al.
444 2021), most patient tumors showed no evidence of locus amplifications extending to the
445 SRR124–134 cluster, as evidenced by the lack of significant ($P > 0.05$) accessibility at
446 the intermediate desert region. In contrast, the SRR124–134 cluster displayed a
447 consistent pattern of accessible chromatin across multiple cancer types: BLCA, BRCA,
448 LUAD, LUSC, STAD, and UCEC (Figure 4C). GBM and LGG tumors lacked accessible
449 chromatin at this cluster but displayed increased chromatin accessibility at the SRR1 and
450 SRR2 enhancers (Supplementary Figure S4A, Supplementary Table S16), which is
451 consistent with the evidence that SRR1 and SRR2 drive *SOX2* expression in the neural
452 lineage (Ferri et al. 2004; Miyagi et al. 2006; Zappone et al. 2000).

453 Next, we reasoned that an accessible SRR124–134 cluster drives subsequent
454 *SOX2* transcription within patient tumors. If this is the case, we expect to find positive and
455 significantly correlated chromatin accessibility between this enhancer cluster and pSOX2.
456 Indeed, we found that the majority of BRCA (58%), LUAD (82%), and LUSC (69%) tumors
457 have concurrent accessibility ($\log_2$ RPM > 0) at pSOX2, SRR124 and SRR134. Patient
458 tumors also showed a significant ($P < 0.05$) correlation (Pearson, R) between accessible
459 chromatin signal at pSOX2 and at both SRR124 and SRR134 in BRCA and LUAD (Figure

460 4D). LUSC tumors showed a significant correlation between accessible chromatin at
461 pSOX2 and SRR124, but not at SRR134 (Figure 4D). As a negative control, we measured
462 the correlation between chromatin accessibility at pSOX2 and at the *SOX2* desert region
463 and found no significant ($P > 0.05$) correlation in any of these cancer types
464 (Supplementary Figure S4B). We also conducted a similar analysis after segregating
465 BRCA tumors into luminal A, luminal B, HER2$^+$, and basal-like subtypes (Berger et al.
466 2018; Corces et al. 2018). Interestingly, we found that both luminal A and luminal B tumors
467 possess a significant ($P < 0.05$) correlation between enhancer accessibility and pSOX2
468 accessibility, whereas for HER2$^+$ tumors the correlation was weaker (Supplementary
469 Figure S4C). Basal-like tumors, on the other hand, display no accessible chromatin at
470 either SRR124 or SRR134. In summary, a luminal-like BRCA phenotype correlates with
471 increased accessibility at the SRR124–134 cluster.

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

479

480 **Figure 4: The SRR124–134 cluster is associated with *SOX2* overexpression in**
481 **cancer patient tumors. (A)** ATAC-seq signal ($\log_2$ RPM) at SRR124 and SRR134 for
482 294 patient tumors from 14 cancer types (Corces et al. 2018). Cancer types are sorted in

483 descending order by the median signal between all three regions. Dashed line: regions
484 with a sum of reads above our threshold (log_2 RPM > 0) were considered "accessible".
485 Error bars: standard deviation. Underscore: top 6 cancer types with the highest ATAC-
486 seq median signal. **(B)** ATAC-seq signal (log_2 RPM) at the *RAB7A* promoter (pRAB7A),
487 *SOX2* promoter (pSOX2), SRR1, SRR2, SRR124, SRR134, hSCR, and a desert region
488 within the *SOX2* locus (desert) compared to the background signal at the repressed
489 *OR5K1* promoter (pOR5K1) in BLCA (n = 10), BRCA (n = 74), LUAD (n = 22), LUSC (n
490 = 16), STAD (n = 21), and UCEC (n = 13) patient tumors. Dashed line: regions with a sum
491 of reads above our threshold (log_2 RPM > 0) were considered "accessible". Error bars:
492 standard deviation. Significance analysis by Dunn's test with Holm correction (* $P < 0.05$,
493 ** $P < 0.01$, *** $P < 0.001$, ns: not significant). **(C)** UCSC Genome Browser (Kent et al.
494 2002) visualization of the *SOX2* region with ATAC-seq data from BLCA, BRCA, LUAD,
495 LUSC, STAD, and UCEC patient tumors (n = 5 in each cancer type) (Corces et al. 2018).
496 ATAC-seq reads were normalized by library size (RPM). Scale: 0 – 250 RPM. **(D)** ATAC-
497 seq signal at SRR124 and SRR134 regions against ATAC-seq signal for the *SOX2*
498 promoter (pSOX2) from 74 BRCA, 22 LUAD, and 16 LUSC patient tumors. Correlation is
499 shown for accessible chromatin (log_2 RPM > 0). Grey: tumors with closed chromatin (log_2
500 RPM < 0) at either region, not included in the correlation analysis. Significance analysis
501 by Pearson correlation. Bolded line: fitted linear regression model. Shaded area: 95%
502 confidence region for the regression fit. **(E)** Comparison of log_2-normalized *SOX2*
503 transcript levels (log_2 counts) between BRCA, LUAD, and LUSC patient tumors according
504 to the chromatin accessibility at SRR124 and SRR134 regions. Chromatin accessibility
505 at each region was considered "low" if log_2 RPM < -1, or "high" if log_2 RPM > 1. RNA-seq

506 reads were normalized to library size using DESeq2 (Love et al. 2014). Error bars:
507 standard deviation. Significance analysis by a two-sided t-test with Holm correction.
508

509 ***FOXA1* and *NFIB* are upstream regulators of the SRR124–134 cluster**

510 With the indication that the SRR124–134 cluster is driving *SOX2* overexpression in
511 cancer patient tumors, we investigated which transcription factors regulate this cluster in
512 BRCA, LUAD, and LUSC tumors from TCGA (Corces et al. 2018; Hoadley et al. 2014).
513 From a list of 1622 human transcription factors (Lambert et al. 2018), we found that the
514 expression of 115 transcription factors was significantly (FDR-adjusted $Q < 0.05$)
515 associated with chromatin accessibility levels at SRR124, whereas accessibility at
516 SRR134 was associated with the expression of 90 transcription factors (Figure 5A,
517 Supplementary Table S18). From this list, we focused our investigation on FOXA1 and
518 NFIB, which show binding at both SRR124 and SRR134 in ChIP-seq data from MCF-7
519 cells (ENCODE Project Consortium 2012).

520 The expression of *FOXA1* is positively (Pearson correlation $R > 0$) and significantly
521 correlated to accessible chromatin at both SRR124 ($R = 0.39$; FDR-adjusted $Q = 1.97\times10^{-}$
522 ^{3}) and SRR134 ($R = 0.46$; $Q = 1.41\times10^{-4}$) (Figure 5B). By separating BRCA, LUAD, and
523 LUSC patient tumors according to the chromatin accessibility levels at each region, we
524 found that tumors with the most accessible chromatin within SRR124 ($P = 2.38\times10^{-4}$, t-
525 test) and SRR134 ($P = 1.53\times10^{-4}$) also significantly overexpress *FOXA1* compared to
526 tumors with low accessibility at these regions (Figure 5C, Supplementary Table S19). On
527 the other hand, we found the expression of *NFIB* to be negatively ($R < 0$) and significantly
528 correlated with chromatin accessibility at both SRR124 ($R = -0.49$; $Q = 4.12\times10^{-5}$) and

529 SRR134 (R = -0.51; Q = 1.32×10^{-5}) (Figure 5D). Patient tumors with highly accessible
530 chromatin within SRR124 (P = 1.46×10^{-6}) and SRR134 (P = 1.24×10^{-5}) also display
531 significantly downregulated *NFIB* expression (Figure 5E, Supplementary Table S20).
532 These data suggest that whereas FOXA1 could be inducing increased accessibility at the
533 SRR124–134 cluster, NFIB expression could counteract FOXA1 by acting as a repressor.

534 To assess the contribution of these transcription factors to enhancer activity, we
535 overexpressed either *FOXA1* or *NFIB* in H520, MCF-7, PC-9, and T47D cells and
536 compared SRR124 and SRR134 enhancer activity to cells transfected with an empty
537 vector (mock) containing only a fluorescent marker. Although endogenous *FOXA1* and
538 *NFIB* expression levels are already high in both MCF-7 and T47D cells (Supplementary
539 Figure S5A), we found that overexpression of *FOXA1* significantly increased ($\log_2$ FC >
540 1; $P < 0.05$, Tukey's test) the enhancer activity of both SRR124 and SRR134 in the H520,
541 MCF-7, PC-9, and T47D cell lines, whereas *NFIB* overexpression led to a significant
542 decrease ($\log_2$ FC < 1; $P < 0.05$) in SRR124 and SRR134 enhancer activity in the H520,
543 MCF-7, and T47D cell lines (Figure 5F). This further indicates that *FOXA1* overexpression
544 increases SRR124–134 activity, whereas NFIB represses the enhancer activity of this
545 cluster.

546 To assess the importance of FOXA1 and NFIB motifs in modulating enhancer
547 activity, we analyzed the SRR134 sequence using the JASPAR2022 motif database
548 (Castro-Mondragon et al. 2022) and mutated FOXA1 (GTAAACA) or NFIB
549 (TGGCAnnnnGCCAA) motifs (mutated SRR134 sequences in Supplementary Table
550 S21). We found that mutation of the FOXA1 motif abolished SRR134 enhancer activity
551 compared to the WT SRR134 sequence within MCF-7 (P = 1.53×10^{-5}, Tukey's test), PC-

552 9 ($P = 1\times10^{-2}$), and T47D ($P = 4.48\times10^{-6}$) cells, whereas no significant change ($P > 0.05$)
553 in enhancer activity was found for the NFIB-mutated construct (Figure 5G). These data
554 indicate that the FOXA1 motif is crucial for sustaining SRR134 activity, whereas the NFIB
555 motif is dispensable in this context, as would be expected for a negative regulator under
556 conditions where the activity of the target is high.

557 With the evidence that these transcription factors are modulating SRR124–134
558 activity, we investigated their transcriptional effects over *SOX2* expression. We used
559 CRISPR homology-directed repair (HDR) to create an MCF-7 cell line in which the *SOX2*
560 gene is tagged with a 2A self-cleaving peptide (P2A) followed by a blue fluorescent protein
561 (tagBFP). This cell line, MCF-7 *SOX2*-P2A-tagBFP, allows rapid visualization of *SOX2*
562 transcriptional changes by measuring tagBFP signal through fluorescence-activated cell
563 sorting (FACS). To validate this model, we sorted cells within the top 10% (BFP^{+ve}) and
564 bottom 10% (BFP^{-ve}) tagBFP signal (Supplementary Figure S5B). We found that BFP^{+ve}
565 cells showed a significant ($P = 4.25\times10^{-5}$, paired t-test) increase in *SOX2* expression, and
566 displayed significantly upregulated transcription of enhancer RNA (eRNA) at SRR124 (P
567 $= 1.54\times10^{-4}$) and SRR134 ($P = 5.13\times10^{-5}$) compared to BFP^{-ve} cells (Figure 5H). This
568 confirms that the tagBFP signal is directly correlated to *SOX2* transcription levels and
569 enhancer output in MCF-7 *SOX2*-P2A-tagBFP cells.

570 Finally, we overexpressed *FOXA1* or *NFIB* in MCF-7 SOX2-P2A-tagBFP to assess
571 changes in *SOX2* transcription. Although overexpression of *FOXA1* did not significantly
572 (chi-squared T(x)=63.70) change tagBFP signal, we found that overexpression of *NFIB*
573 significantly (chi-squared T(x)=1168.88) reduced tagBFP signal compared to transfection
574 of an empty vector (mock) (Figure 5I). This confirms the repressive effect of NFIB over

575 *SOX2* expression and illustrates a potential mechanism upstream of *SOX2* that
576 modulates chromatin accessibility at the SRR124–134 cluster and subsequent control of
577 *SOX2* transcription in cancer cells.
578
579 **Figure 5: FOXA1 and NFIB are upstream regulators of SRR124 and SRR134. (A)**
580 Heatmap of the Pearson correlation between transcription factor expression (Hoadley et
581 al. 2014) and chromatin accessibility (Corces et al. 2018) at SRR124 and SRR134 in
582 BRCA, LUAD, and LUSC patient tumors (n = 111). Transcription factors are ordered
583 according to their correlation to chromatin accessibility at each region. Red: transcription
584 factors with a positive correlation ($R > 0$; FDR-adjusted $Q < 0.05$) to chromatin
585 accessibility. Blue: transcription factors with a negative correlation ($R < 0$; $Q < 0.05$) to
586 chromatin accessibility. Asterisk: transcription factors that show binding at SRR124 or
587 SRR134 by ChIP-seq (ENCODE Project Consortium 2012). **(B)** Correlation analysis
588 between *FOXA1* expression ($\log_2$ counts) and chromatin accessibility ($\log_2$ RPM) at
589 SRR124 and SRR134 regions in BRCA (n = 74), LUAD (n = 21), and LUSC (n = 16)
590 tumors. RNA-seq reads were normalized to library size using DESeq2 (Love et al. 2014).
591 Significance analysis by Pearson correlation (n = 111). Bolded line: fitted linear regression
592 model. Shaded area: 95% confidence region for the regression fit. **(C)** Comparison of
593 *FOXA1* expression ($\log_2$ counts) from BRCA, LUAD, and LUSC patient tumors according
594 to their chromatin accessibility at the SRR124 and SRR134 regions. Chromatin
595 accessibility at each region was considered “low” if $\log_2$ RPM < 1, or “high” if $\log_2$ RPM >
596 1. RNA-seq reads were normalized to library size using DESeq2 (Love et al. 2014). Error
597 bars: standard deviation. Significance analysis by a two-sided t-test with Holm correction.

598 **(D)** Correlation analysis between *NFIB* expression (log_2 counts) and chromatin
599 accessibility (log_2 RPM) at SRR124 and SRR134 regions in BRCA (n = 74), LUAD (n =
600 21), and LUSC (n = 16) tumors. RNA-seq reads were normalized to library size using
601 DESeq2 (Love et al. 2014). Significance analysis by Pearson correlation (n = 111). Bolded
602 line: fitted linear regression model. Shaded area: 95% confidence region for the
603 regression fit. **(E)** Comparison of *NFIB* expression (log_2 counts) from BRCA, LUAD, and
604 LUSC patient tumors according to their chromatin accessibility at the SRR124 and
605 SRR134 regions. Chromatin accessibility at each region was considered "low" if log_2 RPM
606 < 1, or "high" if log_2 RPM > 1. RNA-seq reads were normalized to library size using
607 DESeq2 (Love et al. 2014). Error bars: standard deviation. Significance analysis by a two-
608 sided t-test with Holm correction. **(F)** Relative fold change (log_2 FC) in luciferase activity
609 driven by SRR124 and SRR134 after overexpression of either *FOXA1* or *NFIB* compared
610 to an empty vector containing (mock negative control, miRFP670). Dashed line: average
611 activity of the mock control. Error bars: standard deviation. Significance analysis by
612 Tukey's test (n = 5; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not significant). **(G)** Relative
613 luciferase activity driven by WT, FOXA1-mutated, and NFIB-mutated SRR134 constructs
614 compared to a minimal promoter (minP) vector in the MCF-7, PC-9, and T47D cell lines.
615 Dashed line: average activity of minP. Error bars: standard deviation. Significance
616 analysis by Tukey's test (n = 5; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not significant).
617 **(H)** RT-qPCR comparison of transcripts at *SOX2*, SRR124, and SRR134 between sorted
618 BFP^{-ve} and BFP^{+ve} MCF-7 cells normalized to unsorted population. Error bars: standard
619 deviation. Significance analysis by paired t-test with Holm correction (n = 6; *** $P < 0.001$).
620 **(I)** FACS density plot comparing tagBFP signal between *SOX2*-P2A-tagBFP MCF-7 cells

621 transfected with an empty vector (mock negative control, miRFP670), FOXA1-T2A-
622 miRFP670, or NFIB-T2A-miRFP670. tagBFP signal was acquired from successfully
623 transfected live cells (miRFP^{+}/PI^{-}) after 5 days post-transfection. Significance analysis by
624 FlowJo's chi-squared T(x) test. T(x) scores above 1000 were considered "strongly
625 significant" (*** $P < 0.001$), whereas T(x) scores under 100 were considered "non-
626 significant".
627
628 **SRR124 and SRR134 are conserved enhancers across mammals and are required**
629 **for the separation of the anterior foregut**
630 *SOX2* is required for the proper development of multiple tissues (Arnold et al. 2011),
631 including the digestive and respiratory systems in the mouse (Francis et al. 2019; Gontan
632 et al. 2008; Que et al. 2007, 2009; Teramoto et al. 2020; Tompkins et al. 2009) and in
633 humans (Zenteno et al. 2006; Williamson et al. 2006; Chassaing et al. 2007). Therefore,
634 we questioned whether the SRR124–134 cluster drives *SOX2* expression in additional
635 contexts other than cancer. A compilation of chromatin accessibility data from cardiac,
636 digestive, embryonic, lymphoid, musculoskeletal, myeloid, neural, placental, pulmonary,
637 renal, skin, and vascular tissues (ENCODE Project Consortium 2012; Meuleman et al.
638 2020; Roadmap Epigenomics Consortium et al. 2015) showed that both SRR124 and
639 SRR134 display increased chromatin accessibility in digestive and respiratory tissues
640 alongside cancer samples (Figure 6A). By comparing DNase-seq signal from fetal lung
641 and stomach tissues (ENCODE Project Consortium 2012), we found that both SRR124
642 (lung $P = 1.25\times10^{-6}$; stomach $P = 9.64\times10^{-4}$; Holm-adjusted Dunn's test) and SRR134
643 (lung $P = 1.14\times10^{-3}$; stomach $P = 0.045$), together with SRR2 (lung $P = 1.55\times10^{-3}$;

644 stomach $P = 5.74\times10^{-5}$), are significantly more accessible than pOR5K1 (Figure 6B,
645 Supplementary Table S22). This suggests that SRR124 and SRR134 are contributing to
646 *SOX2* expression during the development of the digestive and respiratory systems.

647 Since critical developmental genes are often controlled by highly conserved
648 enhancers across species (Pennacchio et al. 2006; Woolfe et al. 2005), we hypothesized
649 that the SRR124–134 cluster might regulate *SOX2* expression during the development of
650 other species. By analyzing PhyloP conservation scores (Pollard et al. 2010; Kent et al.
651 2002), we discovered that both SRR124 and SRR134 contain a highly conserved core
652 sequence that is preserved across mammals, birds, reptiles, and amphibians (Figure 6C).
653 After aligning and comparing enhancer sequences between humans and mice, we found
654 the core sequence at both SRR124 and SRR134 are highly conserved (> 80%) in the
655 mouse genome (Supplementary Figure S6A). We termed these homologous regions as
656 mSRR96 (96 kb downstream of the mouse *Sox2* promoter; homologous to the human
657 SRR124) and mSRR102 (102 kb downstream of the mouse *Sox2* promoter; homologous
658 to the human SRR134). Enhancer feature analysis in the developing lung and stomach
659 tissues in the mouse (ENCODE Project Consortium 2012; Liu et al. 2019) showed that
660 both mSRR96 and mSRR102 display increased chromatin accessibility and H3K27ac
661 signal throughout developmental days E14.5 to the 8th post-natal week (Figure 6D).
662 Interestingly, mSRR96 and mSRR102 display higher ATAC-seq and H3K27ac signal
663 towards the later stages of development in the lungs, but at early stages of development
664 in the stomach. This suggests a distinct spatiotemporal contribution of this homologous
665 cluster to *Sox2* expression during the development of these tissues in the mouse. ATAC-
666 seq quantification (genomic coordinates in Supplementary Table S23) showed that both

667 mSRR96 (lung $P = 5.54\times10^{-5}$; stomach $P = 2.37\times10^{-4}$; Holm-adjusted Dunn's test) and
668 mSRR102 (lung $P = 1.27\times10^{-3}$; stomach $P = 0.046$) are significantly more accessible than
669 the repressed promoter of the olfactory gene *Olfr266* (pOlfr266, negative control) during
670 the development of the lungs and stomach in the mouse (Supplementary Figure S6B,
671 Supplementary Table S24). Together, these results suggest a conserved *SOX2*
672 regulatory mechanism across multiple species and support a model in which the SRR124
673 and SRR134 enhancers and their homologs regulate *SOX2* expression during the
674 development of the digestive and respiratory systems.

675 To assess the contribution of the mSRR96 and mSRR102 regions to the
676 development of the mouse, we generated a knockout containing a deletion spanning the
677 mSRR96–102 enhancer cluster (ΔmENH) (Figure 6E). We crossed animals carrying a
678 heterozygous mSRR96–102 deletion ($\Delta mENH^{+/-}$) and determined the number of pups
679 alive at weaning (P21) from each genotype. We found a significant ($P = 1.13\times10^{-4}$, Chi-
680 squared test) deviation from the expected mendelian ratio with no homozygous mice
681 ($\Delta mENH^{-/-}$) alive at weaning (Figure 6F), demonstrating that the mSRR96–102 enhancer
682 cluster is crucial for survival in the mouse. To investigate the resulting phenotype in a
683 homozygous mSRR96–102 enhancer deletion, we collected E18.5 embryos and
684 prepared cross-sections at the thymus level from five animals of each genotype (WT,
685 $\Delta mENH^{+/-}$, and $\Delta mENH^{-/-}$) (Figure 6G). Similar to other studies that interfered with *Sox2*
686 expression during development (Que et al. 2007; Teramoto et al. 2020; Chakraborty et
687 al. 2023), we found that all five $\Delta mENH^{-/-}$ embryos developed EA/TEF, where the
688 esophagus and trachea fail to separate during embryonic development (Figure 6H). WT
689 and $\Delta mENH^{+/-}$ embryos, on the other hand, showed normal development of the

690 esophageal and tracheal tissues. Immunohistochemistry showed the complete absence
691 of the SOX2 protein within the EA/TEF tissue in ΔmENH$^{-/-}$ embryos, whereas WT and
692 ΔmENH$^{+/-}$ embryos showed high levels of SOX2 protein within both the esophagus and
693 tracheal tubes (Figure 6I). Finally, immunofluorescence staining for NKX2.1, a
694 transcription factor associated with the inner epithelium of the respiratory tract (Minoo et
695 al. 1999), showed high protein levels within the inner layer of the EA/TEF tissue in
696 ΔmENH$^{-/-}$ embryos, indicating that this aberrant tissue resembles a tracheal-like structure
697 lacking SOX2 (Supplementary Figure S6C). Together, these results demonstrate that
698 mSRR96 and mSRR102 are required to drive *Sox2* expression during the development
699 and separation of the esophagus and trachea.
700
701 **Figure 6: The SRR124 and SRR134 enhancers are conserved across species and**
702 **are required for the separation of the esophagus and trachea in the mouse. (A)**
703 UCSC Genome Browser (Kent et al. 2002) view of the *SOX2* region containing a
704 compilation of chromatin accessibility tracks of multiple human tissues (ENCODE Project
705 Consortium 2012; Meuleman et al. 2020; Roadmap Epigenomics Consortium et al. 2015).
706 Arrow: increased chromatin accessibility at the SRR124–134 cluster in cancer, digestive,
707 and respiratory tissues. **(B)** DNAse-seq quantification ($\log_2$ RPM) at the *RAB7A* promoter
708 (pRAB7A), *SOX2* promoter (pSOX2), SRR1, SRR2, SRR124, SRR134, human SCR
709 (hSCR), and a desert region within the *SOX2* locus (desert) compared to the background
710 signal at the repressed *OR5K1* promoter (pOR5K1) in lung and stomach embryonic
711 tissues (ENCODE Project Consortium 2012). Dashed line: Regions with a sum of reads
712 above our threshold ($\log_2$ RPM > 0) were considered “accessible”. Error bars: standard

713 deviation. Significance analysis by Dunn's test with Holm correction (* $P < 0.05$, ** $P <$
714 0.01, *** $P < 0.001$, ns: not significant). **(C)** UCSC Genome Browser (Kent et al. 2002)
715 with PhyloP conservation scores (Pollard et al. 2010) at the SRR124 and SRR134
716 enhancers across mammals, birds, reptiles, and amphibians species. Black lines: highly
717 conserved sequences. Empty lines: variant sequences. **(D)** UCSC Genome Browser
718 (Kent et al. 2002) view of the *Sox2* region in the mouse. ATAC-seq and H3K27ac ChIP-
719 seq data from lung and stomach tissues throughout developmental days E14.5 to the 8th
720 post-natal week (ENCODE Project Consortium 2012; Liu et al. 2019). mSRR96:
721 homologous to SRR124. mSRR102: homologous to SRR134. Reads were normalized to
722 library size (RPM). **(E)** Illustration demonstrating the mSRR96–102 enhancer cluster
723 CRISPR deletion (ΔmENH) in C57BL/6J mouse embryos. **(F)** Quantification and
724 genotype of the C57BL/6J progeny from mSRR96–102-deleted crossings (ΔmENH$^{+/-}$).
725 Pups were counted and genotyped at weaning (P21). Significance analysis by chi-
726 squared test to measure the deviation in the number of obtained pups from the expected
727 mendelian ratio of 1:2:1 (WT : ΔmENH$^{+/-}$: ΔmENH$^{-/-}$). **(G)** Transverse cross-section of
728 fixed E18.5 embryos at the start of the thymus. **(H)** Embryo sections stained with
729 Hematoxylin and Eosin (H&E). Scale bar: 500µm. Es: esophagus; Tr: trachea; EA/TEF:
730 esophageal atresia with distal tracheoesophageal fistula. **(I)** Embryo cross-sections
731 stained for SOX2. Scale bar: 500µm. Es: esophagus; Tr: trachea; EA/TEF: esophageal
732 atresia with distal tracheoesophageal fistula.

733 **DISCUSSION**

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

746 The observation that enhancers involved in the development of the digestive and
747 respiratory systems are reprogrammed to support *SOX2* upregulation during
748 tumorigenesis is in line with previous observations that tumor-initiating cells acquire a less
749 differentiated phenotype (Bonnet and Dick 1997; Chaffer et al. 2011; Lapidot et al. 1994,
750 199; Gupta et al. 2011). It is more surprising, however, that the *SOX2* gene is regulated
751 by common enhancers in both breast and lung cancer cells as enhancers are usually
752 highly tissue-specific (Pennacchio et al. 2006; Woolfe et al. 2005; Thurman et al. 2012;
753 Stergachis et al. 2013). Our observation that *FOXA1* expression is significantly correlated
754 to chromatin accessibility at the SRR124–134 cluster and increases the transcriptional
755 output of the SRR124 and SRR134 enhancers provides a mechanistic link between

756 breast and lung developmental programs and cancer progression. FOXA1 is directly
757 involved in the branching morphogenesis of the epithelium in breast (Bernardo et al. 2010;
758 Liu et al. 2016) and lung (Besnard et al. 2005, 1; Paranjapye et al. 2020) tissues, where
759 SOX2 also plays an important role (Domenici et al. 2019; Gontan et al. 2008).
760 Overexpression of both *FOXA1* (Camolotto et al. 2018; Fu et al. 2016, 2019; Lupien et
761 al. 2008; Orstad et al. 2022; Roe et al. 2017; Stergachis et al. 2013) and *SOX2* (Boumahdi
762 et al. 2014; Chou et al. 2013, 2; Liu et al. 2013) have also been individually linked to the
763 activation of transcriptional programs associated with multiple types of cancer. Therefore,
764 we propose that FOXA1 is one of the key players responsible for the reprogramming of
765 the SRR124–134 cluster in cancer, which then drives *SOX2* overexpression in breast and
766 lung tumors. As mutation of the FOXA1 motif disrupted SRR134 enhancer activity, and
767 this motif is shared among other members of the forkhead box (FOX) transcription factor
768 family (Pierrou et al. 1994), it remains unclear if FOXA1 alone activates the SRR124–134
769 cluster, or whether other FOX proteins are involved in this process. For example, *FOXM1*
770 overexpression, which also showed binding at both SRR124 and SRR134 in MCF-7 cells,
771 has similarly been associated with poor patient outcomes in multiple types of cancer (Li
772 et al. 2017).

773 In addition to the activating role of FOXA1, we identified NFIB as a negative
774 regulator of *SOX2* expression through inhibition of SRR124–134 activity. NFIB is normally
775 required for the development of multiple tissues (reviewed in Harris et al. 2015), including
776 the brain and lungs (Steele-Perkins et al. 2005; Gründer et al. 2002; Hsu et al. 2011),
777 tissues in which *SOX2* expression is also tightly regulated (Favaro et al. 2009; Gontan et
778 al. 2008). In the lungs, NFIB is essential for promoting the maturation and differentiation

779 of progenitor cells (Gründer et al. 2002; Steele-Perkins et al. 2005). This is in stark
780 contrast to SOX2, which inhibits the differentiation of lung cells (Gontan et al. 2008).
781 Interestingly, NFIB seems to have paradoxical roles in cancer, acting both as a tumor
782 suppressor and as an oncogene in different tissues (Becker-Santos et al. 2017). Among
783 its tumor suppressor activity, NFIB acts as a barrier to skin carcinoma progression (Zhou
784 et al. 2014b), and its downregulation is associated with dedifferentiation and
785 aggressiveness in LUAD (Becker-Santos et al. 2016). On the other hand, SOX2 promotes
786 skin (Boumahdi et al. 2014) and lung (Ferone et al. 2016) cancer progression. As an
787 oncogene, *NFIB* promotes cell proliferation and metastasis in STAD (Wu et al. 2018),
788 where *SOX2* downregulation is associated with poor patient outcomes (Otsubo et al.
789 2008; Wang et al. 2015; Zhang et al. 2010). With this contrasting relationship between
790 *SOX2* and *NFIB* across multiple tissues, we propose that NFIB normally acts as a
791 suppressor of SRR124–134 activity and *SOX2* expression during the differentiation of
792 progenitor cells; downregulation of *NFIB* expression then results in *SOX2* overexpression
793 during tumorigenesis of the breast and lung.

794 We initially hypothesized that SRR1 and SRR2 (Miyagi et al. 2004; Zappone et al.
795 2000; Tomioka et al. 2002), and/or the SCR (Li et al. 2014; Zhou et al. 2014a) might be
796 recommissioned during cancer progression, as stem cell-related enhancers have been
797 shown to acquire enhancer features in tumorigenic cells (Aran et al. 2016). Although other
798 studies have also proposed the activation of either SRR1 (Leis et al. 2012; Stolzenburg
799 et al. 2012) or SRR2 (Iglesias et al. 2014; Jung et al. 2015) as the main drivers of *SOX2*
800 overexpression in BRCA, we found no evidence of this mechanism and instead identified
801 the SRR124–134 cluster as the main driver of *SOX2* expression in BRCA and LUAD. Our

802 patient tumor analysis did show that GBM and LGG were the only cancer types that
803 display a unique and consistent pattern of accessible chromatin at SRR1 and SRR2,
804 which is likely related to glioma cells assuming a neural stem cell-like identity to sustain
805 high levels of cell proliferation in the brain (Bulstrode et al. 2017). In fact, SRR2 deletion
806 was shown to downregulate *SOX2* and reduce cell proliferation in GBM cells (Saenz-
807 Antoñanzas et al. 2021), highlighting enhancer specificity to different tumor types. In line
808 with these findings, our observation that PC-9 LUAD cells are dependent on SRR124–
809 134 for *SOX2* transcription, whereas in H520 LUSC cells SRR124–134 is dispensable,
810 again highlights these tumor-type specific regulatory mechanisms. LUSC tumors
811 frequently amplify the *SOX2* locus (Bass et al. 2009; Hussenet et al. 2010; Liu et al. 2021;
812 Maier et al. 2011), whereas LUAD tumors do not (Björkqvist et al. 1998), indicating that
813 different mechanisms are involved in genome dysregulation in these two subtypes of lung
814 cancer. Interestingly, a further downstream enhancer cluster located ~55 kb away from
815 SRR124–134 is co-amplified with *SOX2* in LUSC cell lines (Liu et al. 2021), revealing an
816 additional mechanism that could sustain *SOX2* overexpression in the absence of the
817 SRR124–134 cluster in certain types of LUSC but not in LUAD.

818 Deletion of mSRR96–102, homolog of the human SRR124–134 cluster, resulted in
819 EA/TEF which is also observed in human cases with *SOX2* heterozygous mutations
820 (Zenteno et al. 2006; Williamson et al. 2006; Chassaing et al. 2007). Interestingly, a recent
821 study showed that insertion of a CTCF insulation cluster downstream of the *Sox2* gene,
822 but upstream of mSRR96–102, disrupts *Sox2* expression, impairs separation of the
823 esophagus and trachea, and results in perinatal lethality due to EA/TEF in the mouse
824 (Chakraborty et al. 2023). This was of particular interest for understanding enhancer

825 functional nuances during development since the SCR, which is required for *Sox2*
826 transcription at implantation, can partially overcome the insulator effect of this insertion.
827 The authors proposed that enhancer density might explain the EA/TEF phenotype, as
828 chromatin features suggested that enhancers in the developing lung and stomach tissues
829 might be spread over a 400 kb domain (Chakraborty et al. 2023). The 6 kb deletion that
830 removes the mSRR96–102 cluster causing EA/TEF suggests this is not the case. Instead,
831 we propose that the sensitivity of each cell type to gene dosage is behind the differing
832 ability of CTCF to block distal enhancers. This is based on two observations: in humans,
833 heterozygous *SOX2* mutations are linked with the anophthalmia-esophageal-genital
834 syndrome (Zenteno et al. 2006; Williamson et al. 2006; Chassaing et al. 2007); in mice,
835 hypomorphic *Sox2* alleles display similar phenotypes in the eye (Taranova et al. 2006)
836 and EA/TEF (Que et al. 2007; Teramoto et al. 2020). This suggests that cells from the
837 peri-implantation phase are less sensitive to lower *Sox2* dosages compared to cells from
838 the developing airways and digestive systems in both species and explains the aberrant
839 phenotypes observed at term.

840 Overall, our findings illustrate how cis-regulatory regions can similarly drive gene
841 expression in both normal and diseased contexts and serve as a prime example of how
842 decommissioned developmental enhancers may be reprogrammed during tumorigenesis.
843 The fact that we have found a digestive/respiratory-associated enhancer cluster driving
844 gene expression in a non-native context such as BRCA remains intriguing and reinforces
845 a model in which tumorigenic cells often revert to a progenitor-like state that combines
846 cis-regulatory features of progenitor cells from multiple developing lineages (Stergachis
847 et al. 2013). This “dys-differentiation” mechanism seems to be centered around the

848 overexpression of a few key development-associated pioneer transcription factors such
849 as FOXA1 and SOX2. Identifying additional mechanisms that regulate the reprogramming
850 of these enhancers could lead to new approaches to target tumor-initiating cells that
851 depend on *SOX2* overexpression.
852

853 **MATERIALS AND METHODS**

854 *Cell Culture*

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

866

867 *Genome editing*

868 Pairs of gRNA plasmids were constructed by inserting a 20 bp target sequence
869 (Supplementary Table S25) into an empty gRNA cloning vector (a gift from George
870 Church; Addgene plasmid # 41824; http://n2t.net/addgene:41824;
871 RRID:Addgene_41824) (Mali et al. 2013, 9) containing either miRFP670 (Addgene
872 plasmid #163748) or tagBFP (Addgene plasmid #163747) fluorescent markers. Plasmids
873 were sequenced to confirm correct insertion. Both gRNA (1 µg each) vectors were co-
874 transfected with 3 µg of pCas9_GFP (a gift from Kiran Musunuru; Addgene plasmid
875 #44719; http://n2t.net/addgene:44719; RRID:Addgene_44719) (Ding et al. 2013) using

876 Neon electroporation (Life Technologies). After 72 hours of transfection, cells were FACS
877 sorted to select clones that contained all three plasmids. Sorted
878 tagBFP$^+$/GFP$^+$/miRFP670$^+$ cells were grown in a bulk population and serially diluted into
879 individual wells to generate isogenic populations. Once fully grown, each well was
880 screened by PCR to confirm the deletion.
881
882 *Gene tagging*
883 *SOX2* was tagged with a P2A-tagBFP sequence in both alleles using CRISPR-mediated
884 homology-directed repair (HDR) (Ran et al. 2013). This strategy results in the expression
885 of a single transcript that is further translated into two separate proteins due to ribosomal
886 skipping (Ahier and Jarriault 2014). In summary, we designed a gRNA that targets the 3'
887 end of the *SOX2* stop codon (Supplementary Table S25, Addgene plasmid #163752). We
888 then amplified ~800 bp homology arms upstream and downstream of the gRNA target
889 sequence using high-fidelity Phusion Polymerase. We purposely avoided amplification of
890 the *SOX2* promoter sequence to reduce the likelihood of random integrations in the
891 genome. Both homology arms were then joined at each end of a P2A-tagBFP sequence
892 using Gibson assembly. Flanking primers containing the gRNA target sequence were
893 used to reamplify *SOX2*-P2A-tagBFP and add gRNA targets at both ends of the fragment;
894 this approach allows excision of the HDR sequence from the backbone plasmid once
895 inside the cell (Zhang et al. 2017). Finally, the full HDR sequence was inserted into a
896 pJET1.2 (Thermo Scientific) backbone, midiprepped, and sequenced (Addgene
897 #163751). 3µg of HDR template was then co-transfected with 1µg of hCas9 (a gift from
898 George Church; Addgene plasmid #41815; http://n2t.net/addgene:41815 ;

899 RRID:Addgene_41815) (Mali et al. 2013) and 1µg of gRNA plasmid using Neon
900 electroporation (Life Technologies). A week after transfection, tagBFP$^+$ cells were FACS
901 sorted as a bulk population. Sorted cells were further grown for two more weeks, and
902 single tagBFP$^+$ cells were isolated to generate isogenic populations. Once fully grown,
903 each clone was screened by PCR and sequenced to confirm homozygous integration of
904 P2A-tagBFP into the *SOX2* locus.
905
906 *Luciferase assay*
907 Luciferase activity was measured using the dual-luciferase reporter assay (Promega
908 #E1960) that relies on the co-transfection of two plasmids: pGL4.23 (Firefly Luciferase,
909 *luc2*) and pGL4.75 (Renilla Luciferase). Assayed plasmids were constructed by
910 subcloning the empty pGL4.23 vector containing a minimal promoter (minP). SRR124,
911 SRR134, SRR1, SRR2, and hSCR were PCR-amplified (primers in Supplementary Table
912 S26) from MCF-7 genomic DNA using high-fidelity Phusion Polymerase and inserted in
913 the forward position downstream of the *luc2* gene at the NotI restriction site. Constructs
914 were sequenced to confirm correct insertions.
915 JASPAR2022 (Castro-Mondragon et al. 2022) was used to find FOXA1 (GTAAACA) and
916 NFIB (TGGCAnnnnGCCAA) motifs in the SRR134 sequence. Only motifs with a score of
917 80% or higher were further analyzed. Bases within each motif sequence were mutated
918 until the score was reduced below 80% without affecting co-occurring motifs or creating
919 novel binding sites. In total, four FOXA1 motifs and two NFIB motifs were mutated
920 (Supplementary Table S21). Engineered sequences were ordered as gene blocks

921 (Eurofins) and inserted into pGL4.23 in the forward position. Constructs were sequenced
922 to confirm correct insertions.
923 Cells were plated in 96-well plates with 4 technical replicates at 2.10^4 cells per well. After
924 24 hours, a 200ng 50:1 mixture of enhancer vector and pGL4.75 was transfected using
925 Lipofectamine 3000 (0.05µl Lipofectamine:1µl Opti-mem). For transcription factor
926 overexpression analysis, a 200ng 50:10:1 mixture of enhancer vector, expression
927 plasmid, and pGL4.75 was transfected. After 48 hours of transfection, cells were lysed in
928 1x Passive Lysis Buffer and stored at -80°C until all 5 biological replicates were
929 completed. Luciferase activity was measured in the Fluoroskan Ascent FL plate reader.
930 Enhancer activity was calculated by normalizing the firefly signal from pGL4.23 to the
931 *Renilla* signal from pGL4.75.
932
933 *Colony formation assay*
934 MCF-7 and PC-9 cells were seeded at low density (2,000 cells/well) into 6-well plates in
935 triplicate for each cell line. Culture media was renewed every 3 days. After 12 days, cells
936 were fixed with 3.7% paraformaldehyde for 10 minutes and stained with 0.5% crystal
937 violet for 20 minutes to quantify the number of colonies formed. Crystal violet staining was
938 then eluted with 10% acetic acid and absorbance was measured at 570 nm to evaluate
939 cell proliferation. Each 6-well plate was considered one biological replicate and the
940 experiment was repeated five times for each cell line (n = 5).
941
942 *FACS analysis*

943 For analyzing the effects of *FOXA1* and *NFIB* overexpression, 2.10^6 SOX2-P2A-tagBFP
944 cells were transfected with 50nM of plasmid expressing either miRFP670 (a gift from
945 Vladislav Verkhusha; Addgene plasmid #79987; http://n2t.net/addgene:79987;
946 RRID:Addgene_79987), FOXA1-T2A-miRFP670 (Addgene plasmid #182335), or NFIB-
947 T2A-miRFP670 (Addgene plasmid #187222) in 5 replicates. Five days after transfection,
948 miRFP670, tagBFP, and propidium iodide (PI) (live/dead stain) signals were acquired; the
949 amount of tagBFP signal from miRFP670$^+$/PI$^-$ cells was compared between each
950 treatment across all replicates.
951 FlowJo's chi-squared T(x) test was used to contrast the effects of each treatment over
952 tagBFP expression; T(x) scores above 1000 were considered "strongly significant" (***),
953 whereas T(x) scores under 100 were considered "non-significant".
954
955 *Transcriptome analysis*
956 Total RNA was isolated from WT and enhancer-deleted (ΔENH) cell lines using the
957 RNeasy kit. Genomic DNA was digested by Turbo DNAse. 500-2,000ng of total RNA was
958 used in a reverse transcription reaction with random primers. cDNA was diluted in H_2O
959 and amplified in a qPCR reaction using SYBR Select Mix (primers in Supplementary
960 Table S27). Amplicons were sequenced to confirm primer specificity. Gene expression
961 was normalized to *PUM1* (Kılıç et al. 2014; Krasnov et al. 2019; Lyng et al. 2008).
962 Total RNA was sent to The Centre for Applied Genomics (TCAG) for paired-end rRNA-
963 depleted total RNA-seq (Illumina 2500, 125 bp). Read quality was checked by fastQC,
964 trimmed using fastP (Chen et al. 2018) and mapped to the human genome
965 (GRCh38/hg38) using STAR 2.7 (Dobin et al. 2013). Normal breast epithelium RNA-seq

966 was obtained from ENCODE (Supplementary Table S28) (ENCODE Project Consortium
967 2012; Roadmap Epigenomics Consortium et al. 2015). Mapped reads were quantified
968 using featureCounts (Liao et al. 2014) and imported into DESeq2 (Love et al. 2014) for
969 normalization and differential expression analysis. Genes with a $|\log_2 FC| > 1$ and FDR-
970 adjusted $Q < 0.01$ were considered significantly changing. Differential gene expression
971 was plotted using the EnhancedVolcano package. Correlation and clustering heatmaps
972 were plotted using the pheatmap R package (https://cran.r-
973 project.org/web/packages/pheatmap/index.html). Signal enrichment plot was prepared
974 using NGS.plot (Shen et al. 2014).
975 Cancer patient transcriptome data were obtained from TCGA (Hoadley et al. 2014) using
976 the TCGAbiolinks package (Colaprico et al. 2016). The overall survival KM-plot was
977 calculated using clinical information from TCGA (Liu et al. 2018a). Tumor transcriptome
978 data were compared to normal tissue using DESeq2. RNA-seq reads were normalized to
979 library size using DESeq2 (Love et al. 2014) and transformed to a $\log_2$ scale [$\log_2$ counts].
980 Differential gene expression was considered significant if $|\log_2 FC| > 1$ and $Q < 0.01$.
981 Gene Set Enrichment Analysis (GSEA) was performed by ranking genes according to
982 their $\log_2$ FC in $\Delta ENH^{-/-}$ versus WT MCF-7 cells. The ranking was then analyzed using
983 the GSEA function from the clusterProfiler package (Yu et al. 2012) with a threshold of
984 FDR-adjusted $Q < 0.05$ using the MSigDB GO term database (C5).
985
986 *Chromatin accessibility analysis*
987 Cells were grown in three separate wells (n = 3) and 50,000 cells were sent to Princess
988 Margaret Genomics Centre for ATAC-seq library preparation using the Omni-ATAC

989 protocol (Corces et al. 2017). ATAC-seq libraries were sequenced using 50 bp paired-
990 ended parameters in the Illumina Novaseq 6000 platform. Read quality was checked by
991 fastQC, trimmed using fastP and mapped to the human genome (GRCh38/hg38) using
992 STAR 2.7. narrowPeaks were called using Genrich (https://github.com/jsh58/Genrich).
993 Differential chromatin accessibility analysis was performed using diffBind (Stark R, Brown
994 G 2011). ATAC-seq peaks with a $|\log_2 FC| > 1$ and FDR-adjusted $Q < 0.01$ were
995 considered significantly changing. Correlation heatmaps were generated using diffBind.
996 Signal enrichment plot was prepared using NGS.plot (Shen et al. 2014). Genes were
997 separated into three categories according to their expression levels in our WT MCF-7
998 RNA-seq data.
999 Transcription factor footprint analysis was performed using TOBIAS (Bentsen et al. 2020)
1000 with standard settings. Motifs with a $|\log_2 FC| > 0.1$ and FDR-adjusted $Q < 0.01$ were
1001 considered significantly enriched in each condition. Replicates (n = 3) were merged into
1002 a single BAM file for each treatment. Motif enrichment at differential ATAC-seq peaks was
1003 performed using HOMER (Heinz et al. 2010). ATAC-seq peaks were assigned to their
1004 closest gene within ± 1 Mb distance from their promoter using ChIPpeakAnno (Zhu et al.
1005 2010).
1006 Cancer patient ATAC-seq data was obtained from TCGA (Corces et al. 2018). DNAse-
1007 seq from human developing tissues were obtained from ENCODE (Supplementary Table
1008 S28) (ENCODE Project Consortium 2012; Roadmap Epigenomics Consortium et al.
1009 2015). Read quantification was calculated at the *RAB7a* (pRAB7a), *OR5K1* (pOR5K1),
1010 and *SOX2* (pSOX2) promoters, together with SRR1, SRR2, SRR124, SRR134, hSCR,
1011 and desert regions with a 1,500 bp window centered at the core of each region (genomic

1012 coordinates of each region in Supplementary Table S15). Reads were normalized to
1013 library size (RPM) and transformed to a $\log_2$ scale ($\log_2$ RPM) using a custom script
1014 (https://github.com/luisabatti/BAMquantify). Each region's average $\log_2$ RPM was
1015 compared to the *OR5K1* promoter for differential analysis using Dunn's test with Holm
1016 correction. Correlations were calculated using Pearson's correlation test and considered
1017 significant if FDR-adjusted $Q < 0.05$. Chromatin accessibility at SRR124 and SRR134
1018 regions was considered low if $\log_2$ RPM < -1, medium if -1 ≤ $\log_2$ RPM ≤ 1, or high if $\log_2$
1019 RPM > 1.
1020 ATAC-seq from developing mouse lung and stomach tissues were obtained from
1021 ENCODE (Supplementary Table S28) (ENCODE Project Consortium 2012) and others
1022 (Liu et al. 2019). Conserved mouse regulatory regions were lifted from the human build
1023 (GRCh38/hg38) to the mouse build (GRCm38/mm10) using UCSC liftOver (Kent et al.
1024 2002, 200). The number of mapped reads was calculated at the *Egf* (pEgf), *Olfr266*
1025 (pOlfr266), and *Sox2* (pSox2) promoters, together with the mouse mSRR1, mSRR2,
1026 mSRR96, mSRR102, mSCR and desert regions with a 1,500 bp window at each location
1027 (genomic coordinates in Supplementary Table S23). Each $\log_2$-transformed region's
1028 reads per million ($\log_2$ RPM) was compared to the negative *Olfr266* promoter control for
1029 differential analysis using Dunn's test with Holm correction.
1030
1031 *Conservation analysis*
1032 Cross-species evolutionary conservation was obtained using phyloP (Pollard et al. 2010).
1033 Pairwise comparisons between human SRR124 and SRR134 (GRCh38/hg38) and

1034 mouse mSRR96 and mSRR102 (GRCm38/mm10) sequences were plotted using
1035 FlexiDot (Seibt et al. 2018) with an 80% conservation threshold.
1036
1037 *ChIP-seq analysis*
1038 Transcription factor and histone modifications ChIP-seq were obtained from ENCODE
1039 (ENCODE Project Consortium 2012) (Supplementary Table S28) and others (Chan et al.
1040 2018; Cocce et al. 2019; Sato et al. 2019) (Supplementary Table S29). H3K4me1 and
1041 H3K27ac tracks were normalized to input and library size ($\log_2$ RPM). ATAC-seq reads
1042 were normalized to library size (RPM). Histone modification ChIP-seq tracks and
1043 transcription factor ChIP-seq peaks were uploaded to the UCSC browser (Kent et al.
1044 2002, 20) for visualization. Normalized H3K4me1, H3K27ac and ATAC-seq reads were
1045 quantified and the difference in normalized signal was calculated using diffBind. Peaks
1046 with a $|\log_2 FC| > 1$ and $Q < 0.01$ were considered significantly changing.
1047 Overlapping ChIP-seq and ATAC-seq peaks were analyzed using ChIPpeakAnno (Zhu
1048 et al. 2010). The hypergeometric test was performed by comparing the number of
1049 overlapping peaks to the total size of the genome divided by the median peak size.
1050
1051 *Mouse line construction*
1052 Our mSRR96–102 knockout mouse line (C57BL/6J-Del(3)1Tcp/Jmit) was ordered from
1053 and generated by The Centre for Phenogenomics (TCP) in Toronto, ON. The protocol for
1054 the generation of the mouse line has been previously described (Gertsenstein and Nutter
1055 2021). Briefly, C57BL/6J zygotes were collected from superovulated, mated, and plugged
1056 female mice at 0.5-day post coitum. Zygotes were electroporated with Cas9 RNPs

1057 complexes (gRNA sequences in Supplementary Table S25) and transferred into
1058 pseudopregnant female recipients within 3-4 hours of electroporation. Born pups
1059 (founders) were screened by end-point PCR and sequenced to confirm allelic mSRR96–
1060 102 deletions. Heterozygous mSRR96–102 founders (ΔmENH$^{+/-}$) were then backcrossed
1061 to the parental strain to confirm germline transmission. Once the mouse line was
1062 established and the mSRR96–102 deletion was fully confirmed and sequenced in the N1
1063 offspring, ΔmENH$^{+/-}$ mice were crossed and the number of live pups from each genotype
1064 (WT, ΔmENH$^{+/-}$, ΔmENH$^{-/-}$) was assessed at weaning (P21). The obtained number of
1065 live pups from each genotype was then compared to the expected mendelian ratio of
1066 1:2:1 (WT : ΔmENH$^{+/-}$: ΔmENH$^{-/-}$) using a chi-squared test. Once the lethality of the
1067 homozygous deletion was confirmed at weaning, E18.5 embryos generated from new
1068 ΔmENH$^{+/-}$ crosses were collected for further histological analyses. All procedures
1069 involving animals were performed in compliance with the Animals for Research Act of
1070 Ontario and the Guidelines of the Canadian Council on Animal Care. The TCP Animal
1071 Care Committee reviewed and approved all procedures conducted on animals at the
1072 facility.
1073
1074 *Histological analyses*
1075 A total of 46 embryos were collected at E18.5 and fixed in 4% paraformaldehyde. Each
1076 embryo was genotyped, and a total of 15 embryos, 5 of each genotype (WT, ΔmENH$^{+/-}$,
1077 ΔmENH$^{-/-}$), were randomly selected, processed, and embedded in paraffin for sectioning
1078 and further analysis. Tissue sections were collected at 4μm thickness roughly at the start
1079 of the thymus. Sections were prepared by the Pathology Core at TCP.

1080 Tissue sections were stained with Hematoxylin and Eosin (H&E) using an auto-stainer to
1081 ensure batch consistency. Slides were scanned using a Hamamatsu Nanozoomer slide
1082 scanner at 20X magnification. For immunohistochemistry staining, E18.5 embryo cross-
1083 sections were submitted to heat-induced epitope retrieval with TRIS-EDTA (pH 9.0) for
1084 10 minutes, followed by quenching of endogenous peroxidase with Bloxall reagent
1085 (Vector). Non-specific antibody binding was blocked with 2.5 % normal horse serum
1086 (Vector), followed by incubation for 1 hour in Rabbit anti-SOX2 (Abcam, ab92494, 1:500).
1087 After washes, sections were incubated for 30 minutes with ImmPRESS Anti-Rabbit HRP
1088 (Vector), followed by DAB reagent and counterstained in Mayer's hematoxylin.
1089 For immunofluorescence staining, E18.5 embryo cross-sections were collected onto
1090 charged slides and then baked at 60 ºC for 30 minutes. Tissue sections were submitted
1091 to heat-induced epitope retrieval with Citrate Buffer pH 6.0 for 10 minutes. Non-specific
1092 antibody binding was blocked with Protein Block Serum-Free (Dako) for 10 minutes,
1093 followed by overnight incubation at 4 ºC in a primary antibody cocktail (Rabbit anti-
1094 NKX2.1, Abcam ab76013 at 1:200; Rat anti-SOX2, Thermo Fisher Scientific 14-9811-80
1095 at 1:100). After washes with TBS-T, sections were incubated for 1 hour with a cocktail of
1096 Alexa Fluor-conjugated secondary antibodies at 1:200 (Goat anti-Rabbit IgG AF488 –
1097 Thermo Fisher Scientific A32731; Goat anti-Rat IgG AF647 – Thermo Fisher Scientific
1098 A21247), followed by counterstaining with DAPI. Scanning was performed using an
1099 Olympus VS-120 slide scanner and imaged using a Hamamatsu ORCA-R2 C10600
1100 digital camera for all dark-field and fluorescent images.

1101 **SUPPORTING INFORMATION**

1102 **Supplementary Figure S1: (A)** Super-logarithmic volcano plot of *PUM1* expression from
1103 RNA-seq of 21 cancer types compared to normal tissue (Hoadley et al. 2014). Cancer
1104 types with $\log_2$ FC > 1 and FDR-adjusted Q < 0.01 were considered to significantly
1105 overexpress *PUM1*. Error bars: standard deviation. **(B)** Kaplan-Meier plot (Kaplan and
1106 Meier 1958) of overall survival against time since diagnosis for 3,064 patients with BRCA
1107 (n = 1089), COAD (n = 453), GBM (n = 153), LIHC (n = 370), LUAD (n = 504), and LUSC
1108 (n = 495) tumors (Liu et al. 2018a). We divided patients into four equal groups and
1109 compared two groups: high *SOX2* expression (range: 10.06–16.36 $\log_2$ counts) and low
1110 *SOX2* expression (range: 0–1.67 $\log_2$ counts). RNA-seq reads were normalized to library
1111 size using DESeq2 (Love et al. 2014). Significance analysis by logrank test. The
1112 shadowed area represents the 95% confidence interval. **(C)** Comparison of *SOX2*
1113 expression ($\log_2$ counts) between luminal A (n = 560), luminal B (n = 207), HER2+ (n =
1114 82), basal-like (n = 190) breast cancer subtypes (Berger et al. 2018; Hoadley et al. 2014)
1115 and normal mammary tissue (n = 152) (Hoadley et al. 2014). RNA-seq reads were
1116 normalized to library size using DESeq2 (Love et al. 2014). Error bars: standard deviation.
1117 Significance analysis by Tukey's test (*** P < 0.001, * P < 0.05, ns: not significant). **(D)**
1118 Volcano plot with DESeq2 (Love et al. 2014, 2) differential expression analysis between
1119 WT MCF-7 cells and breast epithelium (Roadmap Epigenomics Consortium et al. 2015).
1120 Blue: 7,937 genes that significantly lost expression ($\log_2$ FC < -1; FDR-adjusted Q < 0.01)
1121 in WT MCF-7 cells. Pink: 5,335 genes that significantly gained expression ($\log_2$ FC > 1;
1122 Q < 0.01) in WT MCF-7 cells. Grey: 25,342 genes that maintained similar (-1 ≤ $\log_2$ FC ≤
1123 1) expression between WT MCF-7 and breast epithelium cells. **(E)** RT-qPCR analysis of

1124 *SOX2* transcript levels in the MCF-7, T47D, PC-9 and H520 cell lines. Error bars: standard
1125 deviation. Significance analysis by Tukey's test (n = 3; *** $P < 0.001$).
1126
1127 **Supplementary Figure S2: (A)** SOX2 protein levels in mouse embryonic stem cells
1128 (mESC, positive control), WT, $\Delta ENH^{+/-}$, and $\Delta ENH^{-/-}$ MCF-7 clones. Cyclophilin A (CypA)
1129 was used as a loading control across all samples. **(B)** RT-qPCR analysis of *SOX2*
1130 transcript levels in SRR124–134 heterozygous- ($\Delta ENH^{+/-}$) and homozygous- ($\Delta ENH^{-/-}$)
1131 deleted H520 (LUSC) clones compared to WT cells. Error bars: standard deviation.
1132 Significance analysis by Dunnett's test (n = 3, ns: not significant). **(C)** Euclidean distance
1133 pairwise comparison between WT and $\Delta ENH^{-/-}$ MCF-7 replicates (n = 3) using variance
1134 stabilizing transformed RNA-seq reads from DESeq2 (Love et al. 2014). Darker colors
1135 indicate a higher correlation. **(D)** Euclidean hierarchical clustering of 529 differentially
1136 expressed genes ($|\log_2 FC| > 1$; FDR-adjusted $Q < 0.01$) based on RNA-seq analysis
1137 between WT and $\Delta ENH^{-/-}$ MCF-7 replicates (n = 3). Reads were normalized for each
1138 gene across treatments (Z-score). Blue color indicates downregulated genes (Z-score <
1139 0). Red color indicates upregulated genes (Z-score > 0).
1140
1141 **Supplementary Figure S3: (A)** ATAC-seq metagene enrichment plot ± 2 kb around the
1142 transcription start site (TSS) across all genes from WT and $\Delta ENH^{-/-}$ MCF-7 cells (n = 3).
1143 Reads were normalized by library size (RPM). Grey: TSS. Shaded area: standard
1144 deviation. **(B)** ATAC-seq metagene enrichment plot ± 2 kb around the transcription start
1145 site (TSS) across 12,167 high (average $\log_2$ counts = 9.12), 12,167 medium (average $\log_2$
1146 counts = 2.94), and 12,167 low (average $\log_2$ counts = 0.43) expressed genes in WT

1147 MCF-7 cells. Genes were split into each group according to RNA-seq data. RNA-seq
1148 reads were normalized to library size using DESeq2 (Love et al. 2014). Grey: TSS. (**C)**
1149 ATAC-seq metagene enrichment plot ± 2 kb around the transcription start site (TSS)
1150 across 12,167 high (average $\log_2$ counts = 9.10), 12,167 medium (average $\log_2$ counts =
1151 2.97), and 12,167 low (average $\log_2$ counts = 0.47) expressed genes in ΔENH$^{-/-}$ MCF-7
1152 cells. Genes were split into each group according to RNA-seq data. RNA-seq reads were
1153 normalized to library size using DESeq2 (Love et al. 2014). Grey: TSS. **(D)** Pairwise
1154 Pearson correlation comparison between WT and ΔENH$^{-/-}$ MCF-7 replicates (n = 3) using
1155 ATAC-seq normalized signal from diffBind (Stark R, Brown G 2011). Darker colors
1156 indicate a higher correlation. **(E)** Overlap between GRHL2 ChIP-seq peaks (Cocce et al.
1157 2019) and ATAC-seq peaks that significantly ($\log_2$ FC < -1; P < 0.01) lost chromatin
1158 accessibility in ΔENH$^{-/-}$ MCF-7 cells. Significance analysis by the hypergeometric test
1159 (Zhu et al. 2010). **(F)** Overlap between RUNX2 ChIP-seq peaks (Jeselsohn et al. 2017)
1160 and ATAC-seq peaks that significantly ($\log_2$ FC < -1; P < 0.01) lost chromatin accessibility
1161 in ΔENH$^{-/-}$ MCF-7 cells. Significance analysis by the hypergeometric test (Zhu et al.
1162 2010).
1163
1164 **Supplementary Figure S4: (A)** ATAC-seq signal at the *RAB7A* promoter (pRAB7A),
1165 *SOX2* promoter (pSOX2), SRR1, SRR2, SRR124, SRR134, human SCR (hSCR) and
1166 desert region versus the background signal at the repressed *OR5K1* promoter (pOR5K1)
1167 in BLCA (n = 10), BRCA (n = 74), COAD (n = 38), ESCA (n = 18), GBM (n = 9), HNSC (n
1168 = 9), LGG (n = 13), LIHC (n = 16), LUAD (n = 22), LUSC (n = 16), PRAD (n = 26), STAD
1169 (n = 21), TGCT (n = 9), and UCEC (n = 13) patient tumors. Dashed line: regions with $\log_2$

1170 RPM > 0 were considered "accessible". Error bars: standard deviation. Significance
1171 analysis by Dunn's test with Holm correction (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns:
1172 not significant). **(B)** ATAC-seq signal at the *SOX2* desert region (desert) against ATAC-
1173 seq signal for the *SOX2* promoter (pSOX2) from 74 BRCA, 22 LUAD, and 16 LUSC
1174 patient tumors. Dashed line: regions with $\log_2$ RPM > 0 were considered "accessible".
1175 Significance analysis by Pearson correlation. Bolded line: fitted linear regression model.
1176 Shaded area: 95% confidence region for the regression fit. **(C)** ATAC-seq signal at
1177 SRR124 and SRR134 regions against ATAC-seq signal for the *SOX2* promoter (pSOX2)
1178 from BRCA patient tumors separated into luminal A (n = 31), luminal B (n = 16), $HER2^+$
1179 (n = 10), and basal-like (n = 14) subtypes (Berger et al. 2018; Corces et al. 2018).
1180 Correlation is shown for accessible chromatin ($\log_2$ RPM > 0). Grey: tumors with closed
1181 chromatin ($\log_2$ RPM < 0) at either region, not included in the correlation analysis.
1182 Significance analysis by Pearson correlation. Bolded line: fitted linear regression model.
1183 Shaded area: 95% confidence region for the regression fit.
1184
1185 **Supplementary Figure S5: (A)** RT-qPCR analysis of *FOXA1 and NFIB* expression in the
1186 H520, MCF-7, PC-9, and T47D cell lines. Error bars: standard deviation. Significance
1187 analysis by Tukey's test (n = 3; ** $P < 0.01$, *** $P < 0.001$). **(B)** FACS plot of tagBFP signal
1188 (450 nm) over side scatter (SSC) in WT and *SOX2*-P2A-tagBFP MCF-7 cells. Cell
1189 populations within the top 10% tagBFP signal were considered "tagBFP positive"
1190 (BFP^{+ve}), whereas populations within the bottom 10% BFP signal were considered
1191 "tagBFP negative" (BFP^{-ve}).
1192

1193 **Supplementary Figure S6: (A)** Dot-plot alignment of human (GRCh38/hg38, y-axis)
1194 SRR124 and SRR134, and mouse (GRCm38/mm10, x-axis) mSRR96 and mSRR102
1195 homologous sequences (1,500 bp). Lines indicate high conservation scores (> 80%)
1196 across both species. Sequence alignment using Clustal Omega (Sievers et al. 2011). **(B)**
1197 ATAC-seq quantification ($\log_2$ RPM) at the promoter of the housekeeping gene *Egf* (pEgf,
1198 positive control), *Sox2* promoter (pSox2), mSRR1, mSRR2, mSRR96, mSRR102, mSCR,
1199 and a mouse desert (mdesert) region compared to the background signal at the repressed
1200 *Olfr266* promoter (pOlfr266) in lung and stomach embryonic tissues from the mouse
1201 (ENCODE Project Consortium 2012). mSRR96: homologous to SRR124. mSRR102:
1202 homologous to SRR134. Dashed line: regions with a sum of reads above our threshold
1203 ($\log_2$ RPM > 0) were considered "accessible". Error bars: standard deviation. Significance
1204 analysis by Dunn's test with Holm correction (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns:
1205 not significant). **(C)** Immunofluorescence staining of WT, $\Delta ENH^{+/-}$, and $\Delta ENH^{-/-}$ E18.5
1206 embryos with DAPI (blue), NKX2.1 (green), and SOX2 (white). Cross-sections were
1207 prepared at the thymus levels. Scale bar: 200µm. Es: esophagus; Tr: trachea; EA/TEF:
1208 esophageal atresia with distal tracheoesophageal fistula.
1209

1210 **Supplementary Table S1:** List of TCGA tumor type abbreviations.

1211 **Supplementary Table S2:** *SOX2* differential expression analysis between primary tumor
1212 vs. normal tissue across TCGA cancer types.

1213 **Supplementary Table S3:** *PUM1* differential expression analysis between primary tumor
1214 vs. normal tissue across TCGA cancer types.

1215 **Supplementary Table S4:** TCGA cancer patient overall survival analysis relative to
1216 *SOX2* expression levels.

1217 **Supplementary Table S5:** TCGA copy number variation (CNV) and *SOX2* expression
1218 analysis.

1219 **Supplementary Table S6:** RNA-seq differential expression analysis between WT MCF-
1220 7 vs. breast epithelium (ENCODE).

1221 **Supplementary Table S7:** Differential ATAC-seq, H3K4me1, and H3K27ac analysis
1222 within ± 1 Mb of the *SOX2* gene in WT MCF-7 vs. Breast epithelium (ENCODE).

1223 **Supplementary Table S8:** RNA-seq differential expression analysis comparing $\Delta ENH^{-/-}$
1224 versus WT MCF-7 cells.

1225 **Supplementary Table S9:** Gene set enrichment analysis (GSEA) in WT versus ΔENH^{-}
1226 $^{/-}$ MCF-7 cells.

1227 **Supplementary Table S10:** Significantly changing ATAC-seq peaks in $\Delta ENH^{-/-}$ versus
1228 WT MCF-7 cells.

1229 **Supplementary Table S11:** ATAC-seq peaks that commonly gained signal in WT MCF-
1230 7 vs. breast epithelium and lost signal in $\Delta ENH^{-/-}$ MCF-7 cells.

1231 **Supplementary Table S12:** ATAC-seq footprint analysis in $\Delta ENH^{-/-}$ vs. WT MCF-7 cells.

1232 **Supplementary Table S13:** ChIP-seq motif analysis of GRHL2 peaks in WT MCF-7 cells.

1233 **Supplementary Table S14:** ChIP-seq motif analysis of RUNX2 peaks in WT MCF-7 cells.

1234 **Supplementary Table S15:** Coordinates of regions used in genome-wide analysis in
1235 humans (GRCh38/hg38).

1236 **Supplementary Table S16:** Chromatin accessibility analysis across TCGA cancer types.

1237 **Supplementary Table S17:** ATAC-seq quantification used to separate patient tumors
1238 into expression groups and their *SOX2* expression levels.

1239 **Supplementary Table S18:** Significantly correlated transcription factors to accessible
1240 chromatin at the SRR124–134 cluster in BRCA, LUAD, and LUSC tumors.

1241 **Supplementary Table S19:** *FOXA1* transcript levels and chromatin accessibility at the
1242 SRR124–134 cluster in BRCA, LUAD, and LUSC patient tumors.

1243 **Supplementary Table S20:** *NFIB* transcript levels and chromatin accessibility at the
1244 SRR124–134 cluster in BRCA, LUAD, and LUSC patient tumors.

1245 **Supplementary Table S21:** WT, FOXA1, and NFIB mutated SRR134 sequences.

1246 **Supplementary Table S22:** Chromatin accessibility analysis in human (GRCh38/hg38)
1247 lung and stomach embryonic tissues.

1248 **Supplementary Table S23:** Coordinates of regions used in genome-wide analysis in the
1249 mouse (GRCm38/mm10).

1250 **Supplementary Table S24:** Chromatin accessibility analysis in mouse (GRCm38/mm10)
1251 embryonic lung and stomach tissues

1252 **Supplementary Table S25:** List of gRNA sequences used for CRISPR/Cas9.

1253 **Supplementary Table S26:** List of primers used for enhancer cloning.

1254 **Supplementary Table S27:** List of primers used in RT-qPCR experiments.

1255 **Supplementary Table S28:** ENCODE datasets used in this paper.

1256 **Supplementary Table S29:** GEO datasets used in this paper.

1257

1258 **DATA AVAILABILITY**

1259 Sequencing and processed data files were submitted to the Gene Expression Omnibus
1260 (GEO; https://www.ncbi.nlm.nih.gov/geo/) repository (GSE132344).

1261

1262 **ACKNOWLEDGMENTS**

1263 We thank all the members of the Mitchell laboratory for helpful discussions and Mathieu
1264 Lupien for manuscript review. We also thank the ENCODE Consortium and the TCGA
1265 project for generating and releasing data to the scientific community. Finally, we thank
1266 the contribution of the staff at TCP (The Centre for Phenogenomics), including Kyle
1267 Roberton, who handled the embedding, cutting, and H&E staining of E18.5 mouse
1268 embryos, and Vivian Bradaschia, responsible for the IHC staining. This work was
1269 supported by the Canadian Institutes of Health Research (FRN PJT153186 and
1270 PJT180312), the Canada Foundation for Innovation, and the Ontario Ministry of Research
1271 and Innovation (operating and infrastructure grants held by J.A.M.). BioRender.com was
1272 used to create parts of Figures 6E and 6G.

1273

1274 **AUTHOR CONTRIBUTIONS**

1275 L.E.A. designed and performed bioinformatic analyses, cell culture work, CRISPR
1276 deletions, data curation, gene expression quantification, and led the conceptualization
1277 and writing of the manuscript; P.L.F. assessed cellular phenotypes, including the colony
1278 formation assay; L.H. acquired and processed TCGA ATAC-seq data, and assisted in
1279 writing review & editing; M.C. assisted in the writing review & editing; M.M.H. provided
1280 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.

1284 **REFERENCES**

1285 Ahier A, Jarriault S. 2014. Simultaneous Expression of Multiple Proteins Under a Single
1286 Promoter in *Caenorhabditis elegans* via a Versatile 2A-Based Toolkit. *Genetics*
1287 **196**: 605–613.

1288 Alonso MM, Diez-Valle R, Manterola L, Rubio A, Liu D, Cortes-Santiago N, Urquiza L,
1289 Jauregi P, de Munain AL, Sampron N, et al. 2011. Genetic and Epigenetic
1290 Modifications of Sox2 Contribute to the Invasive Phenotype of Malignant Gliomas
1291 ed. B.W. Futscher. *PLoS ONE* **6**: e26740.

1292 Aran D, Abu-Remaileh M, Levy R, Meron N, Toperoff G, Edrei Y, Bergman Y, Hellman A.
1293 2016. Embryonic Stem Cell (ES)-Specific Enhancers Specify the Expression
1294 Potential of ES Genes in Cancer. *PLoS Genet* **12**: e1005840.

1295 Arnold K, Sarkar A, Yram MA, Polo JM, Bronson R, Sengupta S, Seandel M, Geijsen N,
1296 Hochedlinger K. 2011. Sox2(+) adult stem and progenitor cells are important for
1297 tissue regeneration and survival of mice. *Cell Stem Cell* **9**: 317–329.

1298 Avilion AA. 2003. Multipotent cell lineages in early mouse development depend on SOX2
1299 function. *Genes & Development* **17**: 126–140.

1300 Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, Kim SY, Wardwell L,
1301 Tamayo P, Gat-Viks I, et al. 2009. SOX2 is an amplified lineage-survival oncogene
1302 in lung and esophageal squamous cell carcinomas. *Nature Genetics* **41**: 1238–
1303 1242.

1304 Becker-Santos DD, Lonergan KM, Gronostajski RM, Lam WL. 2017. Nuclear Factor I/B:
1305 A Master Regulator of Cell Differentiation with Paradoxical Roles in Cancer.
1306 *EBioMedicine* **22**: 2–9.

1307 Becker-Santos DD, Thu KL, English JC, Pikor LA, Martinez VD, Zhang M, Vucic EA, Luk
1308 MT, Carraro A, Korbelik J, et al. 2016. Developmental transcription factor NFIB is
1309 a putative target of oncofetal miRNAs and is associated with tumour
1310 aggressiveness in lung adenocarcinoma. *J Pathol* **240**: 161–172.

1311 Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner
1312 J, Kuenne C, Braun T, et al. 2020. ATAC-seq footprinting unravels kinetics of
1313 transcription factor binding during zygotic genome activation. *Nat Commun* **11**:
1314 4267.

1315 Berezovsky AD, Poisson LM, Cherba D, Webb CP, Transou AD, Lemke NW, Hong X,
1316 Hasselbach LA, Irtenkauf SM, Mikkelsen T, et al. 2014. Sox2 promotes malignancy
1317 in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* **16**:
1318 193–206, 206.e19–25.

1319 Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H,
1320 Ravikumar V, et al. 2018. A Comprehensive Pan-Cancer Molecular Study of
1321 Gynecologic and Breast Cancers. *Cancer Cell* **33**: 690-705.e9.

1322 Bernardo GM, Lozada KL, Miedler JD, Harburg G, Hewitt SC, Mosley JD, Godwin AK,
1323 Korach KS, Visvader JE, Kaestner KH, et al. 2010. FOXA1 is an essential
1324 determinant of ERalpha expression and mammary ductal morphogenesis.
1325 *Development* **137**: 2045–2054.

1326 Besnard V, Wert SE, Kaestner KH, Whitsett JA. 2005. Stage-specific regulation of
1327 respiratory epithelial cell differentiation by Foxa1. *Am J Physiol Lung Cell Mol*
1328 *Physiol* **289**: L750-759.

1329 Bi M, Zhang Z, Jiang Y-Z, Xue P, Wang H, Lai Z, Fu X, De Angelis C, Gong Y, Gao Z, et
1330 al. 2020. Enhancer reprogramming driven by high-order assemblies of
1331 transcription factors promotes phenotypic plasticity and breast cancer endocrine
1332 resistance. *Nat Cell Biol* **22**: 701–715.

1333 Björkqvist AM, Husgafvel-Pursiainen K, Anttila S, Karjalainen A, Tammilehto L, Mattson
1334 K, Vainio H, Knuutila S. 1998. DNA gains in 3q occur frequently in squamous cell
1335 carcinoma of the lung, but not in adenocarcinoma. *Genes Chromosomes Cancer*
1336 **22**: 79–82.

1337 Bonnet D, Dick JE. 1997. Human acute myeloid leukemia is organized as a hierarchy that
1338 originates from a primitive hematopoietic cell. *Nat Med* **3**: 730–737.

1339 Boumahdi S, Driessens G, Lapouge G, Rorive S, Nassar D, Le Mercier M, Delatte B,
1340 Caauwe A, Lenglez S, Nkusi E, et al. 2014. SOX2 controls tumour initiation and
1341 cancer stem-cell functions in squamous-cell carcinoma. *Nature* **511**: 246–250.

1342 Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford
1343 GE. 2008. High-Resolution Mapping and Characterization of Open Chromatin
1344 across the Genome. *Cell* **132**: 311–322.

1345 Brunner HG, van Bokhoven H. 2005. Genetic players in esophageal atresia and
1346 tracheoesophageal fistula. *Curr Opin Genet Dev* **15**: 341–347.

1347 Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of
1348 native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat*
1349 *Methods* **10**: 1213–1218.

1350 Bulstrode H, Johnstone E, Marques-Torrejon MA, Ferguson KM, Bressan RB, Blin C,
1351 Grant V, Gogolok S, Gangoso E, Gagrica S, et al. 2017. Elevated FOXG1 and
1352 SOX2 in glioblastoma enforces neural stem cell identity through transcriptional
1353 control of cell cycle and epigenetic regulators. *Genes Dev* **31**: 757–773.

1354 Camolotto SA, Pattabiraman S, Mosbruger TL, Jones A, Belova VK, Orstad G, Streiff M,
1355 Salmond L, Stubben C, Kaestner KH, et al. 2018. FoxA1 and FoxA2 drive gastric

1356 differentiation and suppress squamous identity in NKX2-1-negative lung cancer.
1357 *Elife* **7**: e38579.

1358 Carter D, Chakalova L, Osborne CS, Dai Y, Fraser P. 2002. Long-range chromatin
1359 regulatory interactions in vivo. *Nature Genetics* **32**: 623.

1360 Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L,
1361 Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022.
1362 JASPAR 2022: the 9th release of the open-access database of transcription factor
1363 binding profiles. *Nucleic Acids Research* **50**: D165–D173.

1364 Chaffer CL, Brueckmann I, Scheel C, Kaestli AJ, Wiggins PA, Rodrigues LO, Brooks M,
1365 Reinhardt F, Su Y, Polyak K, et al. 2011. Normal and neoplastic nonstem cells can
1366 spontaneously convert to a stem-like state. *PNAS* **108**: 7950–7955.

1367 Chakraborty S, Kopitchinski N, Zuo Z, Eraso A, Awasthi P, Chari R, Mitra A, Tobias IC,
1368 Moorthy SD, Dale RK, et al. 2023. Enhancer–promoter interactions can bypass
1369 CTCF-mediated boundaries and contribute to phenotypic robustness. *Nat Genet*.
1370 https://www.nature.com/articles/s41588-022-01295-6 (Accessed February 1,
1371 2023).

1372 Chan HL, Beckedorff F, Zhang Y, Garcia-Huidobro J, Jiang H, Colaprico A, Bilbao D,
1373 Figueroa ME, LaCava J, Shiekhattar R, et al. 2018. Polycomb complexes
1374 associate with enhancers and promote oncogenic transcriptional programs in
1375 cancer through multiple mechanisms. *Nat Commun* **9**: 3377.

1376 Chassaing N, Gilbert-Dussardier B, Nicot F, Fermeaux V, Encha-Razavi F, Fiorenza M,
1377 Toutain A, Calvas P. 2007. Germinal mosaicism and familial recurrence of a SOX2
1378 mutation with highly variable phenotypic expression extending from AEG
1379 syndrome to absence of ocular involvement. *Am J Med Genet A* **143A**: 289–291.

1380 Chen C, Morris Q, Mitchell JA. 2012. Enhancer identification in mouse embryonic stem
1381 cells using integrative modeling of chromatin and genomic features. *BMC*
1382 *Genomics* **13**: 152.

1383 Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
1384 *Bioinformatics* **34**: i884–i890.

1385 Chen Y, Shi L, Zhang L, Li R, Liang J, Yu W, Sun L, Yang X, Wang Y, Zhang Y, et al.
1386 2008. The Molecular Mechanism Governing the Oncogenic Potential of SOX2 in
1387 Breast Cancer. *Journal of Biological Chemistry* **283**: 17969–17978.

1388 Chew J-L, Loh Y-H, Zhang W, Chen X, Tam W-L, Yeap L-S, Li P, Ang Y-S, Lim B, Robson
1389 P, et al. 2005. Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the
1390 Oct4/Sox2 Complex in Embryonic Stem Cells. *Mol Cell Biol* **25**: 6031–6046.

1391 Chou Y-T, Lee C-C, Hsiao S-H, Lin S-E, Lin S-C, Chung C-H, Chung C-H, Kao Y-R,
1392 Wang Y-H, Chen C-T, et al. 2013. The Emerging Role of SOX2 in Cell Proliferation

1393 and Survival and Its Crosstalk with Oncogenic Signaling in Lung Cancer. *STEM*
1394 *CELLS* **31**: 2607–2619.

1395 Cocce KJ, Jasper JS, Desautels TK, Everett L, Wardell S, Westerling T, Baldi R, Wright
1396 TM, Tavares K, Yllanes A, et al. 2019. The Lineage Determining Factor GRHL2
1397 Collaborates with FOXA1 to Establish a Targetable Pathway in Endocrine
1398 Therapy-Resistant Breast Cancer. *Cell Rep* **29**: 889-903.e10.

1399 Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM,
1400 Pagnotta SM, Castiglioni I, et al. 2016. TCGAbiolinks: an R/Bioconductor package
1401 for integrative analysis of TCGA data. *Nucleic Acids Research* **44**: e71–e71.

1402 Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld
1403 C, Wong CK, Cho SW, et al. 2018. The chromatin accessibility landscape of
1404 primary human cancers. *Science* **362**: eaav1898.

1405 Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S,
1406 Satpathy AT, Rubin AJ, Montine KS, Wu B, et al. 2017. An improved ATAC-seq
1407 protocol reduces background and enables interrogation of frozen tissues. *Nature*
1408 *Methods* **14**: 959–962.

1409 Cox JL, Wilder PJ, Desler M, Rizzino A. 2012. Elevating SOX2 levels deleteriously affects
1410 the growth of medulloblastoma and glioblastoma cells. *PLoS One* **7**: e44087.

1411 Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J,
1412 Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates
1413 active from poised enhancers and predicts developmental state. *Proc Natl Acad*
1414 *Sci USA* **107**: 21931–21936.

1415 Ding Q, Regan SN, Xia Y, Oostrom LA, Cowan CA, Musunuru K. 2013. Enhanced
1416 efficiency of human pluripotent stem cell genome editing through replacing
1417 TALENs with CRISPRs. *Cell Stem Cell* **12**: 393–394.

1418 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
1419 Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:
1420 15–21.

1421 Domenici G, Aurrekoetxea-Rodríguez I, Simões BM, Rábano M, Lee SY, Millán JS,
1422 Comaills V, Oliemuller E, López-Ruiz JA, Zabalza I, et al. 2019. A Sox2–Sox9
1423 signalling axis maintains human breast luminal progenitor and breast cancer stem
1424 cells. *Oncogene*. http://www.nature.com/articles/s41388-018-0656-7 (Accessed
1425 January 14, 2019).

1426 Driskell RR, Giangreco A, Jensen KB, Mulder KW, Watt FM. 2009. Sox2-positive dermal
1427 papilla cells specify hair follicle type in mammalian epidermis. *Development* **136**:
1428 2815–2823.

1429 Dunn OJ. 1964. Multiple Comparisons Using Rank Sums. *Technometrics* **6**: 241–252.

1430 Dunnett CW. 1955. A Multiple Comparison Procedure for Comparing Several Treatments
1431 with a Control. *Journal of the American Statistical Association* **50**: 1096–1121.

1432 ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the
1433 human genome. *Nature* **489**: 57–74.

1434 Fang X, Yoon J-G, Li L, Yu W, Shao J, Hua D, Zheng S, Hood L, Goodlett DR, Foltz G,
1435 et al. 2011. The SOX2 response program in glioblastoma multiforme: an integrated
1436 ChIP-seq, expression microarray, and microRNA analysis. *BMC Genomics* **12**: 11.

1437 Favaro R, Valotta M, Ferri ALM, Latorre E, Mariani J, Giachino C, Lancini C, Tosetti V,
1438 Ottolenghi S, Taylor V, et al. 2009. Hippocampal development and neural stem
1439 cell maintenance require Sox2-dependent regulation of Shh. *Nature Neuroscience*
1440 **12**: 1248–1256.

1441 Ferone G, Song J-Y, Sutherland KD, Bhaskaran R, Monkhorst K, Lambooij J-P, Proost
1442 N, Gargiulo G, Berns A. 2016. SOX2 Is the Determining Oncogenic Switch in
1443 Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. *Cancer*
1444 *Cell* **30**: 519–532.

1445 Ferri ALM, Cavallaro M, Braida D, Di Cristofano A, Canta A, Vezzani A, Ottolenghi S,
1446 Pandolfi PP, Sala M, DeBiasi S, et al. 2004. Sox2 deficiency causes
1447 neurodegeneration and impaired neurogenesis in the adult mouse brain.
1448 *Development* **131**: 3805–3819.

1449 Francis R, Guo H, Streutker C, Ahmed M, Yung T, Dirks PB, He HH, Kim T-H. 2019.
1450 Gastrointestinal transcription factors drive lineage-specific developmental
1451 programs in organ specification and cancer. *Sci Adv* **5**: eaax8898.

1452 Fu X, Jeselsohn R, Pereira R, Hollingsworth EF, Creighton CJ, Li F, Shea M, Nardone A,
1453 Angelis CD, Heiser LM, et al. 2016. FOXA1 overexpression mediates endocrine
1454 resistance by altering the ER transcriptome and IL-8 expression in ER-positive
1455 breast cancer. *PNAS* **113**: E6600–E6609.

1456 Fu X, Pereira R, De Angelis C, Veeraraghavan J, Nanda S, Qin L, Cataldo ML, Sethunath
1457 V, Mehravaran S, Gutierrez C, et al. 2019. FOXA1 upregulation promotes
1458 enhancer and transcriptional reprogramming in endocrine-resistant breast cancer.
1459 *Proc Natl Acad Sci USA* **116**: 26823–26834.

1460 Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei
1461 PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin
1462 interactome. *Nature* **462**: 58–64.

1463 Gangemi RMR, Griffero F, Marubbi D, Perera M, Capra MC, Malatesta P, Ravetti GL,
1464 Zona GL, Daga A, Corte G. 2009. SOX2 silencing in glioblastoma tumor-initiating
1465 cells causes stop of proliferation and loss of tumorigenicity. *Stem Cells* **27**: 40–48.

1466 Gertsenstein M, Nutter LMJ. 2021. Production of knockout mouse lines with Cas9.
1467 *Methods* **191**: 32–43.

1468 Gontan C, de Munck A, Vermeij M, Grosveld F, Tibboel D, Rottier R. 2008. Sox2 is
1469 important for two crucial processes in lung development: Branching
1470 morphogenesis and epithelial cell differentiation. *Developmental Biology* **317**:
1471 296–309.

1472 Gopi LK, Kidder BL. 2021. Integrative pan cancer analysis reveals epigenomic variation
1473 in cancer type and cell specific chromatin domains. *Nat Commun* **12**: 1419.

1474 Graham V, Khudyakov J, Ellis P, Pevny L. 2003. SOX2 Functions to Maintain Neural
1475 Progenitor Identity. *Neuron* **39**: 749–765.

1476 Gründer A, Ebel TT, Mallo M, Schwarzkopf G, Shimizu T, Sippel AE, Schrewe H. 2002.
1477 Nuclear factor I-B (Nfib) deficient mice have severe lung hypoplasia. *Mech Dev*
1478 **112**: 69–77.

1479 Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, Lander ES. 2011.
1480 Stochastic state transitions give rise to phenotypic equilibrium in populations of
1481 cancer cells. *Cell* **146**: 633–644.

1482 Hägerstrand D, He X, Bradic Lindh M, Hoefs S, Hesselager G, Ostman A, Nistér M. 2011.
1483 Identification of a SOX2-dependent subset of tumor- and sphere-forming
1484 glioblastoma cells with a distinct tyrosine kinase inhibitor sensitivity profile. *Neuro*
1485 *Oncol* **13**: 1178–1191.

1486 Harris L, Genovesi LA, Gronostajski RM, Wainwright BJ, Piper M. 2015. Nuclear factor
1487 one transcription factors: Divergent functions in developmental versus adult stem
1488 cell populations. *Dev Dyn* **244**: 227–238.

1489 Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y,
1490 Klugman S, et al. 2010. Distinct epigenomic landscapes of pluripotent and lineage-
1491 committed human cells. *Cell Stem Cell* **6**: 479–491.

1492 Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK,
1493 Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers
1494 reflect global cell-type-specific gene expression. *Nature* **459**: 108–112.

1495 Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar
1496 S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of
1497 transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**:
1498 311–318.

1499 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H,
1500 Glass CK. 2010. Simple combinations of lineage-determining transcription factors
1501 prime cis-regulatory elements required for macrophage and B cell identities. *Mol*
1502 *Cell* **38**: 576–589.

1503 Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA.
1504 2013. Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**: 934–
1505 947.

1506 Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu
1507 B, McLellan MD, Uzunangelov V, et al. 2014. Multiplatform analysis of 12 cancer
1508 types reveals molecular classification within and across tissues of origin. *Cell* **158**:
1509 929–944.

1510 Holm S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian*
1511 *Journal of Statistics* **6**: 65–70.

1512 Hsu Y-C, Osinski J, Campbell CE, Litwack ED, Wang D, Liu S, Bachurski CJ, Gronostajski
1513 RM. 2011. Mesenchymal nuclear factor I B regulates cell proliferation and epithelial
1514 differentiation during lung maturation. *Dev Biol* **354**: 242–252.

1515 Hussenet T, Dali S, Exinger J, Monga B, Jost B, Dembelé D, Martinet N, Thibault C,
1516 Huelsken J, Brambilla E, et al. 2010. SOX2 Is an Oncogene Activated by Recurrent
1517 3q26.3 Amplifications in Human Lung Squamous Cell Carcinomas. *PLOS ONE* **5**:
1518 e8960.

1519 Iglesias JM, Leis O, Pérez Ruiz E, Gumuzio Barrie J, Garcia-Garcia F, Aduriz A, Beloqui
1520 I, Hernandez-Garcia S, Lopez-Mato MP, Dopazo J, et al. 2014. The Activation of
1521 the Sox2 RR2 Pluripotency Transcriptional Reporter in Human Breast Cancer Cell
1522 Lines is Dynamic and Labels Cells with Higher Tumorigenic Potential. *Front Oncol*
1523 **4**: 308.

1524 Jeon H-M, Sohn Y-W, Oh S-Y, Oh S-Y, Kim S-H, Beck S, Kim S, Kim H. 2011. ID4 imparts
1525 chemoresistance and cancer stemness to glioma cells by derepressing miR-9*-
1526 mediated suppression of SOX2. *Cancer Res* **71**: 3410–3421.

1527 Jeselsohn R, Cornwell M, Pun M, Buchwalter G, Nguyen M, Bango C, Huang Y, Kuang
1528 Y, Paweletz C, Fu X, et al. 2017. Embryonic transcription factor SOX9 drives
1529 breast cancer endocrine resistance. *Proceedings of the National Academy of*
1530 *Sciences* **114**: E4482–E4491.

1531 Jung K, Gupta N, Wang P, Lewis JT, Gopal K, Wu F, Ye X, Alshareef A, Abdulkarim BS,
1532 Douglas DN, et al. 2015. Triple negative breast cancers comprise a highly
1533 tumorigenic cell subpopulation detectable by its high responsiveness to a Sox2
1534 regulatory region 2 (SRR2) reporter. *Oncotarget* **6**: 10366–10373.

1535 Kaplan EL, Meier P. 1958. Nonparametric Estimation from Incomplete Observations.
1536 *Journal of the American Statistical Association* **53**: 457–481.

1537 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D.
1538 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

1539 Kiernan AE, Pelling AL, Leung KKH, Tang ASP, Bell DM, Tease C, Lovell-Badge R, Steel
1540 KP, Cheah KSE. 2005. Sox2 is required for sensory organ development in the
1541 mammalian inner ear. *Nature* **434**: 1031–1035.

1542 Kılıç Y, Çelebiler AÇ, Sakızlı M. 2014. Selecting housekeeping genes as references for
1543 the normalization of quantitative PCR data in breast cancer. *Clinical and*
1544 *Translational Oncology* **16**: 184–190.

1545 Krasnov GS, Kudryavtseva AV, Snezhkina AV, Lakunina VA, Beniaminov AD, Melnikova
1546 NV, Dmitriev AA. 2019. Pan-Cancer Analysis of TCGA Data Revealed Promising
1547 Reference Genes for qPCR Normalization. *Frontiers in Genetics* **10**.
1548 https://www.frontiersin.org/article/10.3389/fgene.2019.00097/full (Accessed
1549 March 10, 2020).

1550 Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes
1551 TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.

1552 Lapidot T, Sirard C, Vormoor J, Murdoch B, Hoang T, Caceres-Cortes J, Minden M,
1553 Paterson B, Caligiuri MA, Dick JE. 1994. A cell initiating human acute myeloid
1554 leukaemia after transplantation into SCID mice. *Nature* **367**: 645–648.

1555 Leis O, Eguiara A, Lopez-Arribillaga E, Alberdi MJ, Hernandez-Garcia S, Elorriaga K,
1556 Pandiella A, Rezola R, Martin AG. 2012. Sox2 expression in breast tumours and
1557 activation in breast cancer stem cells. *Oncogene* **31**: 1354–1365.

1558 Li L, Wu D, Yu Q, Li L, Wu P. 2017. Prognostic value of FOXM1 in solid tumors: a
1559 systematic review and meta-analysis. *Oncotarget* **8**: 32298–32308.

1560 Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B. 2014. CRISPR
1561 Reveals a Distal Super-Enhancer Required for Sox2 Expression in Mouse
1562 Embryonic Stem Cells ed. Q. Wu. *PLoS ONE* **9**: e114485.

1563 Liang S, Furuhashi M, Nakane R, Nakazawa S, Goudarzi H, Hamada J, Iizasa H. 2013.
1564 Isolation and characterization of human breast cancer cells with SOX2 promoter
1565 activity. *Biochemical and Biophysical Research Communications* **437**: 205–211.

1566 Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for
1567 assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

1568 Ling G-Q, Chen, Dong-bo, Wang, Bao-Qing, Zhang, Lan-Sheng. 2012. Expression of the
1569 pluripotency markers Oct3/4, Nanog and Sox2 in human breast cancer cell lines.
1570 *Oncology Letters*. http://www.spandidos-publications.com/10.3892/ol.2012.916
1571 (Accessed November 9, 2016).

1572 Liu C, Wang M, Wei X, Wu L, Xu J, Dai X, Xia J, Cheng M, Yuan Y, Zhang P, et al. 2019.
1573 An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci Data* **6**: 65.

1574 Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ,
1575 Benz CC, Levine DA, Lee AV, et al. 2018a. An Integrated TCGA Pan-Cancer
1576 Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**:
1577 400-416.e11.

1578 Liu K-C, Lin B-S, Zhao M, Yang X, Chen M, Gao A, Que J, Lan X-P. 2013. The multiple
1579 roles for Sox2 in stem cell maintenance and tumorigenesis. *Cell Signal* **25**.
1580 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3871517/ (Accessed April 5, 2017).

1581 Liu P, Tang H, Song C, Wang J, Chen B, Huang X, Pei X, Liu L. 2018b. SOX2 Promotes
1582 Cell Proliferation and Metastasis in Triple Negative Breast Cancer. *Front*
1583 *Pharmacol* **9**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6110877/ (Accessed
1584 March 21, 2019).

1585 Liu Y, Wu Z, Zhou J, Ramadurai DKA, Mortenson KL, Aguilera-Jimenez E, Yan Y, Yang
1586 X, Taylor AM, Varley KE, et al. 2021. A predominant enhancer co-amplified with
1587 the SOX2 oncogene is necessary and sufficient for its expression in squamous
1588 cancer. *Nat Commun* **12**: 7139.

1589 Liu Y, Zhao Y, Skerry B, Wang X, Colin-Cassin C, Radisky DC, Kaestner KH, Li Z. 2016.
1590 Foxa1 is essential for mammary duct formation. *Genesis* **54**: 277–285.

1591 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion
1592 for RNA-seq data with DESeq2. *Genome Biol* **15**.
1593 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/ (Accessed November
1594 18, 2017).

1595 Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young
1596 RA. 2013. Selective Inhibition of Tumor Oncogenes by Disruption of Super-
1597 Enhancers. *Cell* **153**: 320–334.

1598 Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown
1599 M. 2008. FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-
1600 Specific Transcription. *Cell* **132**: 958–970.

1601 Lyng MB, Lænkholm A-V, Pallisgaard N, Ditzel HJ. 2008. Identification of genes for
1602 normalization of real-time RT-PCR data in breast carcinomas. *BMC Cancer* **8**.
1603 http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-8-20 (Accessed
1604 June 17, 2020).

1605 Maier S, Wilbertz T, Braun M, Scheble V, Reischl M, Mikut R, Menon R, Nikolov P,
1606 Petersen K, Beschorner C, et al. 2011. SOX2 amplification is a common event in
1607 squamous cell carcinomas of different organ sites. *Human Pathology* **42**: 1078–
1608 1088.

1609 Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013.
1610 RNA-Guided Human Genome Engineering via Cas9. *Science* **339**: 823–826.

1611 Meng Y, Xu Q, Chen L, Wang L, Hu X. 2020. The function of SOX2 in breast cancer and
1612 relevant signaling pathway. *Pathology - Research and Practice* **216**: 153023.

1613 Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F,
1614 Teodosiadis A, et al. 2020. Index and biological spectrum of human DNase I
1615 hypersensitive sites. *Nature* **584**: 244–251.

1616 Minoo P, Su G, Drum H, Bringas P, Kimura S. 1999. Defects in tracheoesophageal and
1617 lung morphogenesis in Nkx2.1(-/-) mouse embryos. *Dev Biol* **209**: 60–71.

1618 Mitchell JA, Clay I, Umlauf D, Chen C-Y, Moir CA, Eskiw CH, Schoenfelder S, Chakalova
1619 L, Nagano T, Fraser P. 2012. Nuclear RNA sequencing of the mouse erythroid cell
1620 transcriptome. *PLoS ONE* **7**: e49274.

1621 Miyagi S, Nishimoto M, Saito T, Ninomiya M, Sawamoto K, Okano H, Muramatsu M,
1622 Oguro H, Iwama A, Okuda A. 2006. The Sox2 Regulatory Region 2 Functions as
1623 a Neural Stem Cell-specific Enhancer in the Telencephalon. *Journal of Biological*
1624 *Chemistry* **281**: 13374–13381.

1625 Miyagi S, Saito T, Mizutani K, Masuyama N, Gotoh Y, Iwama A, Nakauchi H, Masui S,
1626 Niwa H, Nishimoto M, et al. 2004. The Sox-2 regulatory regions display their
1627 activities in two distinct types of multipotent stem cells. *Mol Cell Biol* **24**: 4207–
1628 4220.

1629 Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L,
1630 Martchenko A, So V, Macpherson NN, Mitchell JA. 2017. Enhancers and super-
1631 enhancers have an equivalent regulatory role in embryonic stem cells through
1632 regulation of single or multiple genes. *Genome Res* **27**: 246–258.

1633 Nakatsugawa M, Takahashi A, Hirohashi Y, Torigoe T, Inoda S, Murase M, Asanuma H,
1634 Tamura Y, Morita R, Michifuri Y, et al. 2011. SOX2 is overexpressed in stem-like
1635 cells of human lung adenocarcinoma and augments the tumorigenicity. *Lab Invest*
1636 **91**: 1796–1804.

1637 Novak D, Hüser L, Elton JJ, Umansky V, Altevogt P, Utikal J. 2019. SOX2 in development
1638 and cancer biology. *Seminars in Cancer Biology*.
1639 http://www.sciencedirect.com/science/article/pii/S1044579X18301858 (Accessed
1640 August 19, 2019).

1641 Okabe A, Kaneda A. 2021. Transcriptional dysregulation by aberrant enhancer activation
1642 and rewiring in cancer. *Cancer Sci* **112**: 2081–2088.

1643 Okubo T, Pevny LH, Hogan BLM. 2006. Sox2 is required for development of taste bud
1644 sensory cells. *Genes & Development* **20**: 2654–2659.

1645 Orstad G, Fort G, Parnell TJ, Jones A, Stubben C, Lohman B, Gillis KL, Orellana W, Tariq
1646 R, Klingbeil O, et al. 2022. FoxA1 and FoxA2 control growth and cellular identity in
1647 NKX2-1-positive lung adenocarcinoma. *Dev Cell* **57**: 1866-1882.e10.

1648 Otsubo T, Akiyama Y, Yanagihara K, Yuasa Y. 2008. SOX2 is frequently downregulated
1649 in gastric cancers and inhibits cell growth through cell-cycle arrest and apoptosis.
1650 *Br J Cancer* **98**: 824–831.

1651 Paranjapye A, Mutolo MJ, Ebron JS, Leir S-H, Harris A. 2020. The FOXA1 transcriptional
1652 network coordinates key functions of primary human airway epithelial cells. *Am J*
1653 *Physiol Lung Cell Mol Physiol* **319**: L126–L136.

1654 Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky
1655 S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human
1656 conserved non-coding sequences. *Nature* **444**: 499–502.

1657 Pierrou S, Hellqvist M, Samuelsson L, Enerbäck S, Carlsson P. 1994. Cloning and
1658 characterization of seven human forkhead proteins: binding site specificity and
1659 DNA bending. *The EMBO Journal* **13**: 5002–5012.

1660 Piva M, Domenici G, Iriondo O, Rábano M, Simões BM, Comaills V, Barredo I, López-
1661 Ruiz JA, Zabalza I, Kypta R, et al. 2014. Sox2 promotes tamoxifen resistance in
1662 breast cancer cells. *EMBO Molecular Medicine* **6**: 66–79.

1663 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral
1664 substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

1665 Pomerantz MM, Li F, Takeda DY, Lenci R, Chonkar A, Chabot M, Cejas P, Vazquez F,
1666 Cook J, Shivdasani RA, et al. 2015. The androgen receptor cistrome is extensively
1667 reprogrammed in human prostate tumorigenesis. *Nat Genet* **47**: 1346–1351.

1668 Que J, Choi M, Ziel JW, Klingensmith J, Hogan BLM. 2006. Morphogenesis of the trachea
1669 and esophagus: current players and new roles for noggin and Bmps. *Differentiation*
1670 **74**: 422–437.

1671 Que J, Luo X, Schwartz RJ, Hogan BLM. 2009. Multiple roles for Sox2 in the developing
1672 and adult mouse trachea. *Development* **136**: 1899–1907.

1673 Que J, Okubo T, Goldenring JR, Nam K-T, Kurotani R, Morrisey EE, Taranova O, Pevny
1674 LH, Hogan BLM. 2007. Multiple dose-dependent roles for Sox2 in the patterning
1675 and differentiation of anterior foregut endoderm. *Development* **134**: 2521–2531.

1676 Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique
1677 chromatin signature uncovers early developmental enhancers in humans. *Nature*
1678 **470**: 279–283.

1679 Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering
1680 using the CRISPR-Cas9 system. *Nat Protoc* **8**: 2281–2308.

1681 Richart L, Bidard F-C, Margueron R. 2021. Enhancer rewiring in tumors: an opportunity
1682 for therapeutic intervention. *Oncogene* **40**: 3475–3491.

1683 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen
1684 A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative
1685 analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

1686 Roe J-S, Hwang C-I, Somerville TDD, Milazzo JP, Lee EJ, Da Silva B, Maiorino L, Tiriac
1687 H, Young CM, Miyabayashi K, et al. 2017. Enhancer Reprogramming Promotes
1688 Pancreatic Cancer Metastasis. *Cell* **170**: 875-888.e20.

1689 Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, Kim
1690 DS, Boxer LD, Cairns J, Spivakov M, et al. 2017. Lineage-specific dynamic and
1691 pre-established enhancer-promoter contacts cooperate in terminal differentiation.
1692 *Nat Genet* **49**: 1522–1528.

1693 Saenz-Antoñanzas A, Moncho-Amor V, Auzmendi-Iriarte J, Elua-Pinin A, Rizzoti K,
1694 Lovell-Badge R, Matheu A. 2021. CRISPR/Cas9 Deletion of SOX2 Regulatory
1695 Region 2 (SRR2) Decreases SOX2 Malignant Activity in Glioblastoma. *Cancers*
1696 *(Basel)* **13**: 1574.

1697 Sato T, Yoo S, Kong R, Sinha A, Chandramani-Shivalingappa P, Patel A, Fridrikh M,
1698 Nagano O, Masuko T, Beasley MB, et al. 2019. Epigenomic Profiling Discovers
1699 Trans-lineage SOX2 Partnerships Driving Tumor Heterogeneity in Lung
1700 Squamous Cell Carcinoma. *Cancer Research* **79**: 6084–6100.

1701 Seibt KM, Schmidt T, Heitkam T. 2018. FlexiDot: highly customizable, ambiguity-aware
1702 dotplots for visual sequence analyses ed. J. Hancock. *Bioinformatics* **34**: 3575–
1703 3577.

1704 Shen L, Shao N, Liu X, Nestler E. 2014. ngs.plot: Quick mining and visualization of next-
1705 generation sequencing data by integrating genomic databases. *BMC Genomics*
1706 **15**: 284.

1707 Sholl LM, Barletta JA, Yeap BY, Chirieac LR, Hornick JL. 2010. Sox2 protein expression
1708 is an independent poor prognostic indicator in stage I lung adenocarcinoma. *Am J*
1709 *Surg Pathol* **34**: 1193–1198.

1710 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H,
1711 Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein
1712 multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

1713 Simões BM, Piva M, Iriondo O, Comaills V, López-Ruiz JA, Zabalza I, Mieza JA, Acinas
1714 O, Vivanco M d.M. 2011. Effects of estrogen on the proportion of stem cells in the
1715 breast. *Breast Cancer Research and Treatment* **129**: 23–35.

1716 Singh G, Mullany S, Moorthy SD, Zhang R, Mehdi T, Tian R, Duncan AG, Moses AM,
1717 Mitchell JA. 2021. A flexible repertoire of transcription factor binding sites and a
1718 diversity threshold determines enhancer activity in embryonic stem cells. *Genome*
1719 *Res* **31**: 564–575.

1720 Singh S, Trevino J, Bora-Singhal N, Coppola D, Haura E, Altiok S, Chellappan SP. 2012.
1721 EGFR/Src/Akt signaling modulates Sox2 expression and self-renewal of stem-like
1722 side-population cells in non-small cell lung cancer. *Molecular Cancer* **11**: 73.

1723 Soufi A, Donahue G, Zaret KS. 2012. Facilitators and Impediments of the Pluripotency
1724 Reprogramming Factors' Initial Engagement with the Genome. *Cell* **151**: 994–
1725 1004.

1726 Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. 2015. Pioneer
1727 transcription factors target partial DNA motifs on nucleosomes to initiate
1728 reprogramming. *Cell* **161**: 555–568.

1729 Soule HD, Vazquez J, Long A, Albert S, Brennan M. 1973. A human cell line from a pleural
1730 effusion derived from a breast carcinoma. *Journal of the National Cancer Institute*
1731 **51**: 1409–1416.

1732 Stark R, Brown G. 2011. DiffBind: differential binding analysis of ChIP-Seq peak data.
1733 http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind
1734 .pdf.

1735 Steele-Perkins G, Plachez C, Butz KG, Yang G, Bachurski CJ, Kinsman SL, Litwack ED,
1736 Richards LJ, Gronostajski RM. 2005. The transcription factor gene Nfib is essential
1737 for both lung maturation and brain development. *Mol Cell Biol* **25**: 685–698.

1738 Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB,
1739 Thurman RE, Sandstrom R, et al. 2013. Developmental Fate and Cellular Maturity
1740 Encoded in Human Regulatory DNA Landscapes. *Cell* **154**: 888–903.

1741 Stolzenburg S, Rots MG, Beltran AS, Rivenbark AG, Yuan X, Qian H, Strahl BD,
1742 Blancafort P. 2012. Targeted silencing of the oncogenic transcription factor SOX2
1743 in breast cancer. *Nucleic Acids Research* **40**: 6725–6740.

1744 Sun C, Sun L, Li Y, Kang X, Zhang S, Liu Y. 2013. Sox2 expression predicts poor survival
1745 of hepatocellular carcinoma patients and it promotes liver cancer cell invasion by
1746 activating Slug. *Med Oncol* **30**: 503.

1747 Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007.
1748 Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined
1749 Factors. *Cell* **131**: 861–872.

1750 Takahashi K, Yamanaka S. 2006. Induction of Pluripotent Stem Cells from Mouse
1751 Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**: 663–676.

1752 Takeda K, Mizushima T, Yokoyama Y, Hirose H, Wu X, Qian Y, Ikehata K, Miyoshi N,
1753 Takahashi H, Haraguchi N, et al. 2018. Sox2 is associated with cancer stem-like
1754 properties in colorectal cancer. *Scientific Reports* **8**: 17639.

1755 Talebi A, Kianersi K, Beiraghdar M. 2015. Comparison of gene expression of SOX2 and
1756 OCT4 in normal tissue, polyps, and colon adenocarcinoma using
1757 immunohistochemical staining. *Adv Biomed Res* **4**: 234.

1758 Taranova OV, Magness ST, Fagan BM, Wu Y, Surzenko N, Hutton SR, Pevny LH. 2006.
1759 SOX2 is a dose-dependent regulator of retinal neural progenitor competence.
1760 *Genes Dev* **20**: 1187–1202.

1761 Teramoto M, Sugawara R, Minegishi K, Uchikawa M, Takemoto T, Kuroiwa A, Ishii Y,
1762 Kondoh H. 2020. The absence of SOX2 in the anterior foregut alters the
1763 esophagus into trachea and bronchi in both epithelial and mesenchymal
1764 components. *Biology Open* **9**: bio048728.

1765 Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC,
1766 Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin
1767 landscape of the human genome. *Nature* **489**: 75–82.

1768 Tobias IC, Abatti LE, Moorthy SD, Mullany S, Taylor T, Khader N, Filice MA, Mitchell JA.
1769 2021. Transcriptional enhancers: from prediction to functional assessment on a
1770 genome-wide scale. *Genome* **64**: 426–448.

1771 Tolhuis B, Palstra R-J, Splinter E, Grosveld F, de Laat W. 2002. Looping and Interaction
1772 between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell* **10**:
1773 1453–1465.

1774 Tomioka M, Nishimoto M, Miyagi S, Katayanagi T, Fukui N, Niwa H, Muramatsu M, Okuda
1775 A. 2002. Identification of Sox-2 regulatory region which is under the control of Oct-
1776 3/4–Sox-2 complex. *Nucleic Acids Res* **30**: 3202–3213.

1777 Tompkins DH, Besnard V, Lange AW, Keiser AR, Wert SE, Bruno MD, Whitsett JA. 2011.
1778 Sox2 Activates Cell Proliferation and Differentiation in the Respiratory Epithelium.
1779 *Am J Respir Cell Mol Biol* **45**: 101–110.

1780 Tompkins DH, Besnard V, Lange AW, Wert SE, Keiser AR, Smith AN, Lang R, Whitsett
1781 JA. 2009. Sox2 Is Required for Maintenance and Differentiation of Bronchiolar
1782 Clara, Ciliated, and Goblet Cells ed. C.M. Verfaillie. *PLoS ONE* **4**: e8248.

1783 Tukey JW. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* **5**:
1784 99.

1785 Wang S, Tie J, Wang R, Hu F, Gao L, Wang W, Wang L, Li Z, Hu S, Tang S, et al. 2015.
1786 SOX2, a predictor of survival in gastric cancer, inhibits cell proliferation and
1787 metastasis by regulating PTEN. *Cancer Lett* **358**: 210–219.

1788 Williamson KA, Hever AM, Rainger J, Rogers RC, Magee A, Fiedler Z, Keng WT, Sharkey
1789 FH, McGill N, Hill CJ, et al. 2006. Mutations in SOX2 cause anophthalmia-
1790 esophageal-genital (AEG) syndrome. *Hum Mol Genet* **15**: 1413–1422.

1791 Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P,
1792 Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are
1793 associated with vertebrate development. *PLoS Biol* **3**: e7.

1794 Wu C, Zhu X, Liu W, Ruan T, Wan W, Tao K. 2018. NFIB promotes cell growth,
1795 aggressiveness, metastasis and EMT of gastric cancer through the Akt/Stat3
1796 signaling pathway. *Oncol Rep* **40**: 1565–1573.

1797 Wuebben EL, Rizzino A. 2017. The dark side of SOX2: cancer-a comprehensive
1798 overview. *Oncotarget* **8**: 44917–44943.

1799 Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R Package for Comparing
1800 Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*
1801 **16**: 284–287.

1802 Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J,
1803 Jonsdottir GA, Ruotti V, Stewart R, et al. 2007. Induced Pluripotent Stem Cell Lines
1804 Derived from Human Somatic Cells. *Science* **318**: 1917–1920.

1805 Zappone MV, Galli R, Catena R, Meani N, De Biasi S, Mattei E, Tiveron C, Vescovi AL,
1806 Lovell-Badge R, Ottolenghi S, et al. 2000. Sox2 regulatory sequences direct
1807 expression of a (beta)-geo transgene to telencephalic neural stem cells and
1808 precursors of the mouse embryo, revealing regionalization of gene expression in
1809 CNS stem cells. *Development* **127**: 2367–2382.

1810 Zenteno JC, Perez-Cano HJ, Aguinaga M. 2006. Anophthalmia-esophageal atresia
1811 syndrome caused by an SOX2 gene deletion in monozygotic twin brothers with
1812 markedly discordant phenotypes. *Am J Med Genet A* **140**: 1899–1903.

1813 Zhang J-P, Li X-L, Li G-H, Chen W, Arakaki C, Botimer GD, Baylink D, Zhang L, Wen W,
1814 Fu Y-W, et al. 2017. Efficient precise knockin with a double cut HDR donor after
1815 CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biology* **18**: 35.

1816 Zhang L-H, Yin Y-H, Chen H-Z, Feng S-Y, Liu J-L, Chen L, Fu W-L, Sun G-C, Yu X-G,
1817 Xu D-G. 2020a. TRIM24 promotes stemness and invasiveness of glioblastoma
1818 cells via activating Sox2 expression. *Neuro Oncol* **22**.

1819 Zhang X, Yu H, Yang Y, Zhu R, Bai J, Peng Z, He Y, Chen L, Chen W, Fang D, et al.
1820 2010. SOX2 in gastric carcinoma, but not Hath1, is related to patients'
1821 clinicopathological features and prognosis. *J Gastrointest Surg* **14**: 1220–1226.

1822 Zhang X-H, Wang W, Wang Y-Q, Zhu L, Ma L. 2020b. The association of SOX2 with
1823 clinical features and prognosis in colorectal cancer: A meta-analysis. *Pathol Res*
1824 *Pract* **216**.

1825 Zhou HY, Katsman Y, Dhaliwal NK, Davidson S, Macpherson NN, Sakthidevi M, Collura
1826 F, Mitchell JA. 2014a. A Sox2 distal enhancer cluster regulates embryonic stem
1827 cell differentiation potential. *Genes & Development* **28**: 2699–2711.

1828 Zhou M, Zhou L, Zheng L, Guo L, Wang Y, Liu H, Ou C, Ding Z. 2014b. miR-365 promotes
1829 cutaneous squamous cell carcinoma (CSCC) through targeting nuclear factor I/B
1830 (NFIB). *PLoS One* **9**: e100620.

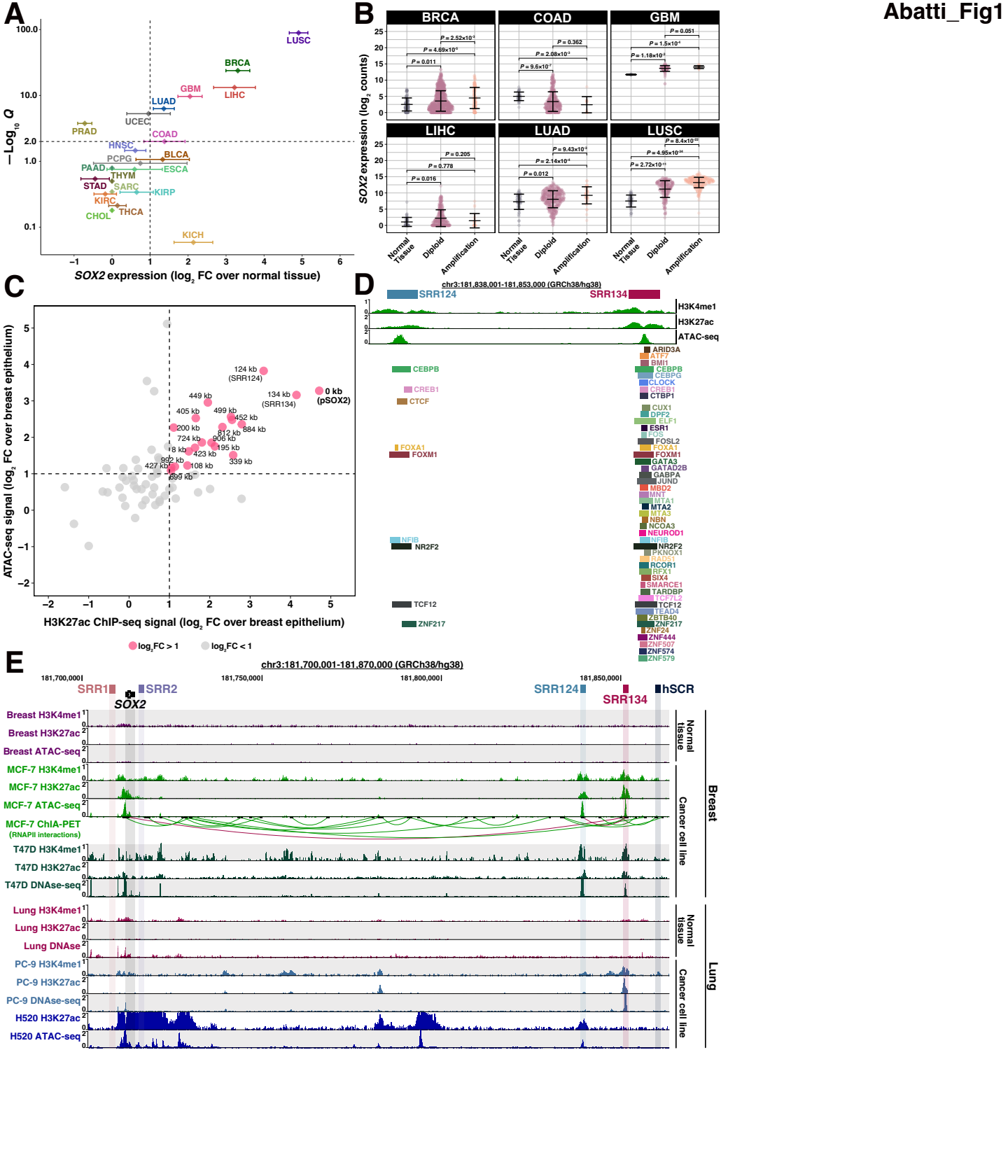
1831 Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande
1832 V, De Jager PL, et al. 2013. Genome-wide chromatin state transitions associated
1833 with developmental and environmental cues. *Cell* **152**: 642–654.

1834 Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. 2010.
1835 ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip
1836 data. *BMC Bioinformatics* **11**.
1837 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-237
1838 (Accessed June 17, 2020).

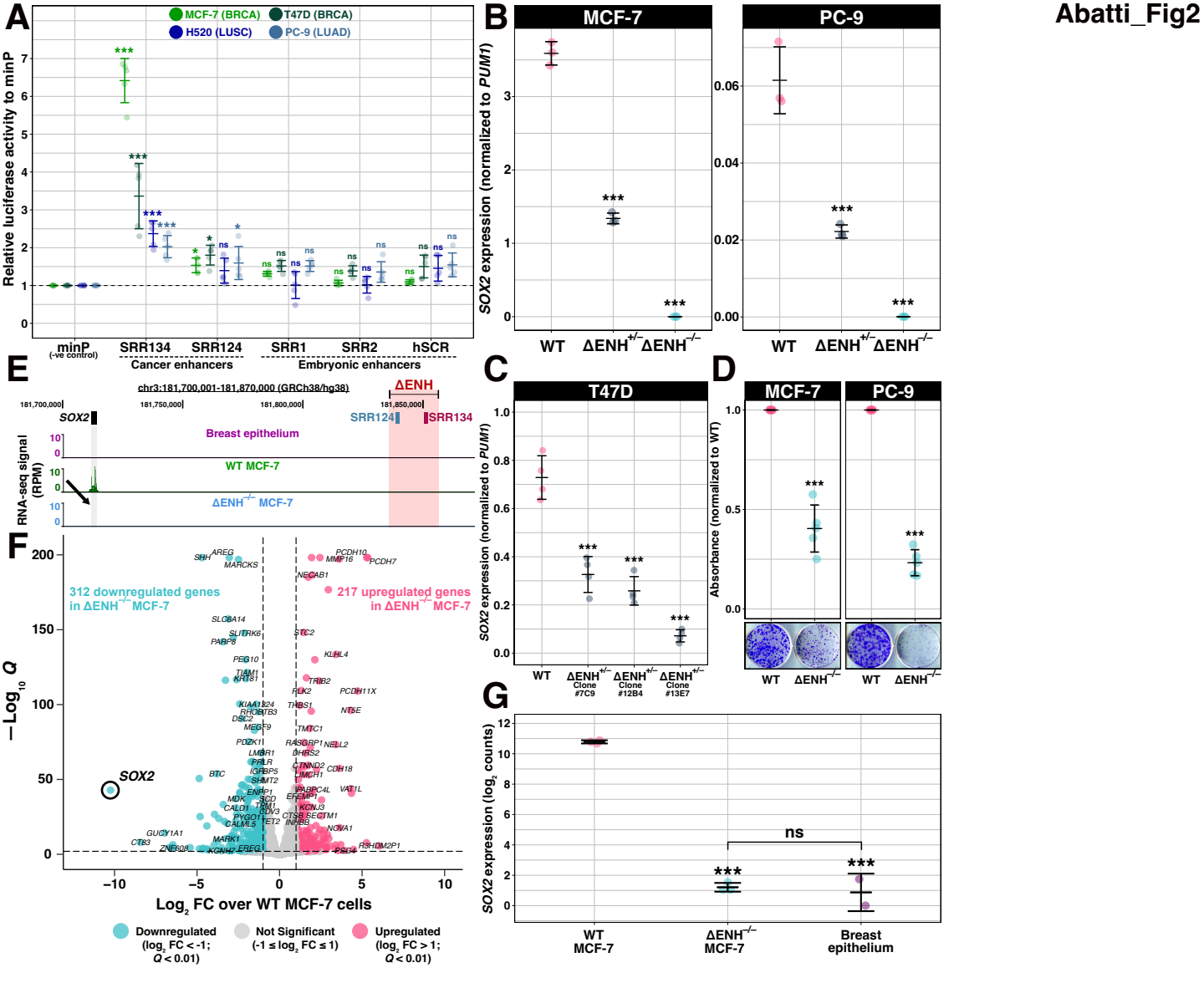
1839 Zhu Y, Huang S, Chen S, Chen J, Wang Z, Wang Y, Zheng H. 2021. SOX2 promotes
1840 chemoresistance, cancer stem cells properties, and epithelial-mesenchymal
1841 transition by β-catenin and Beclin1/autophagy signaling in colorectal cancer. *Cell*
1842 *Death Dis* **12**: 449.

1843

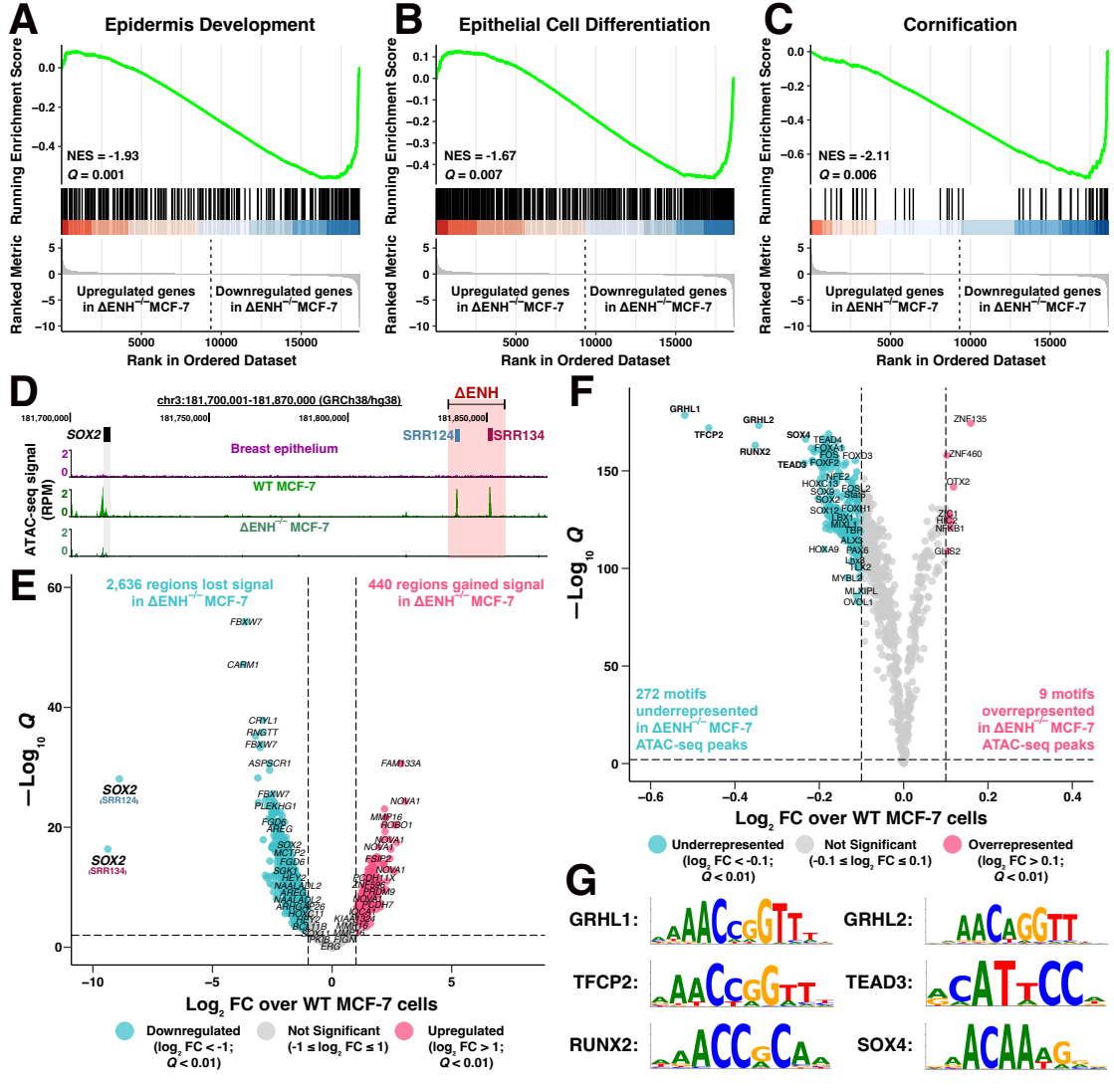
Abatti_Fig1


A
B
C
D
E
SOX2 expression (log2 FC over normal tissue)
−Log10 Q
LUSC
BRCA
GBM
LIHC
LUAD
UCEC
PRAD
COAD
HNSC
BLCA
PCPG
PAAD
ESCA
THYM
STAD
SARC
KIRP
KIRC
THCA
CHOL
KICH
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)
0 kb (pSOX2)
124 kb (SRR124)
134 kb (SRR134)
log2FC > 1
log2FC < 1
chr3:181,838,001-181,853,000 (GRCh38/hg38)
SRR124
SRR134
H3K4me1
H3K27ac
ATAC-seq
CEBPB
CREB1
CTCF
FOXA1
FOXM1
NFIB
NR2F2
TCF12
ZNF217
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

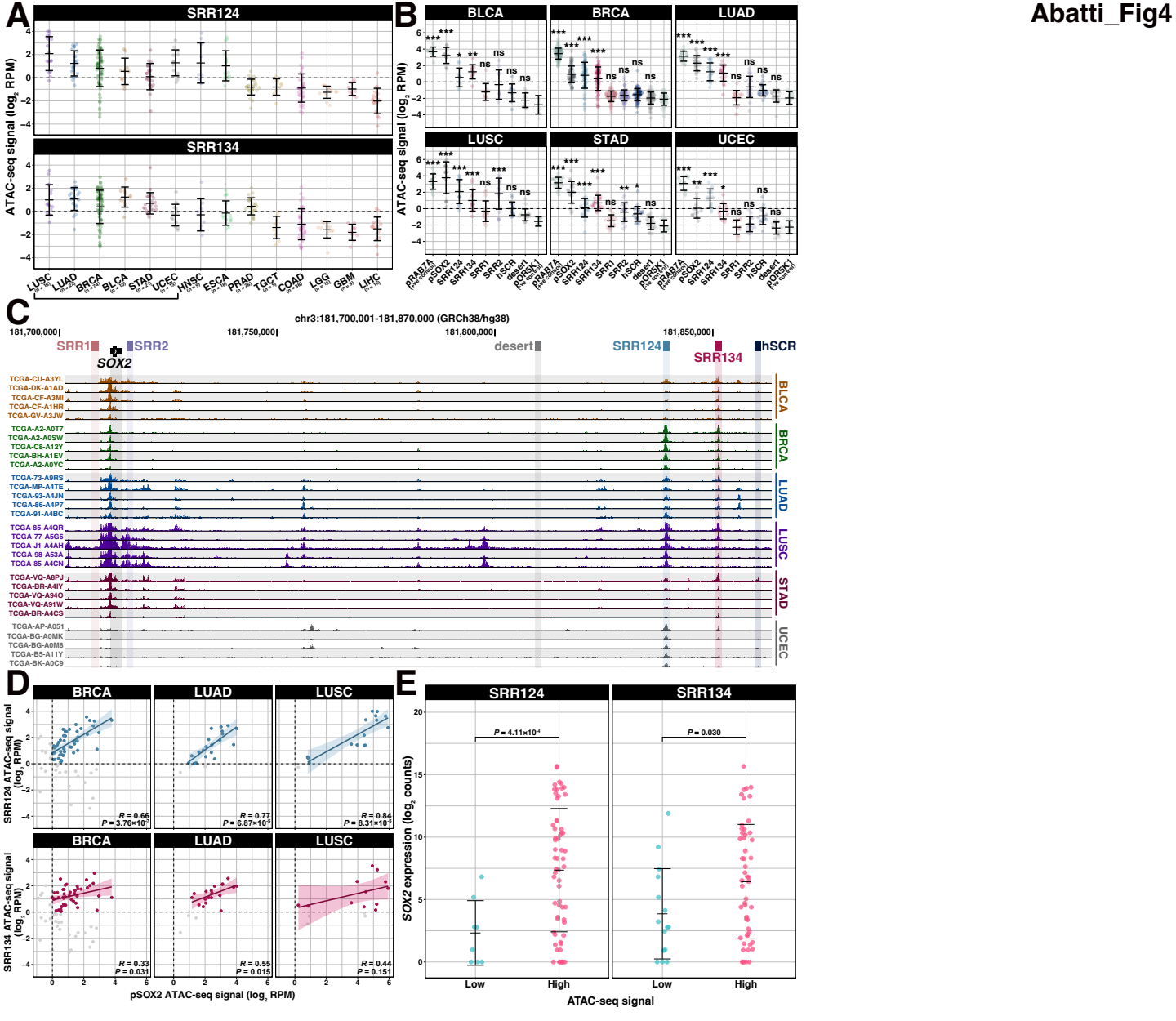
Abatti_Fig2


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
Clone #7C9
Clone #12B4
Clone #13E7
D
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

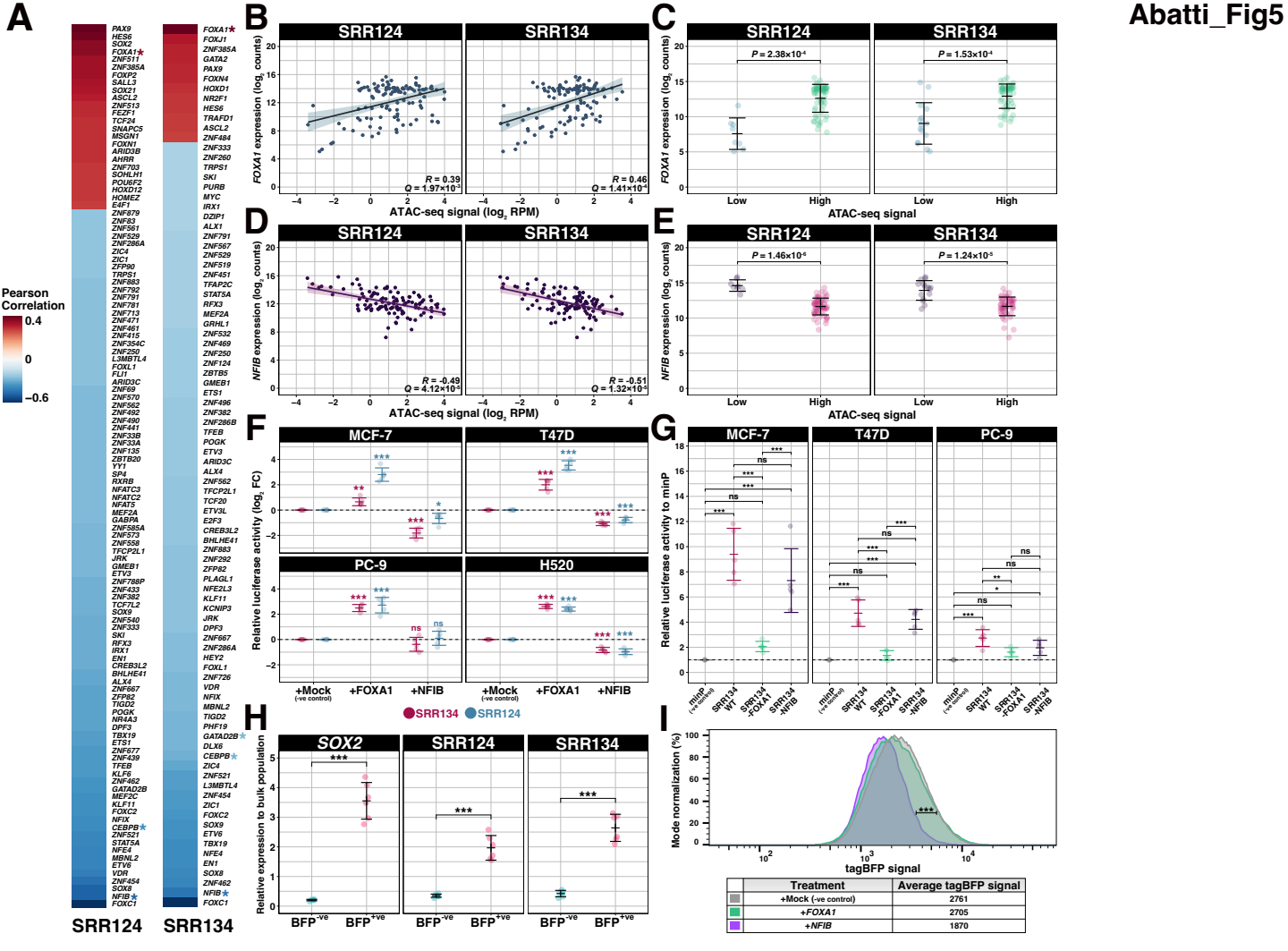
Abatti_Fig3


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
SRR124
SRR134
Breast epithelium
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
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:

Abatti_Fig4

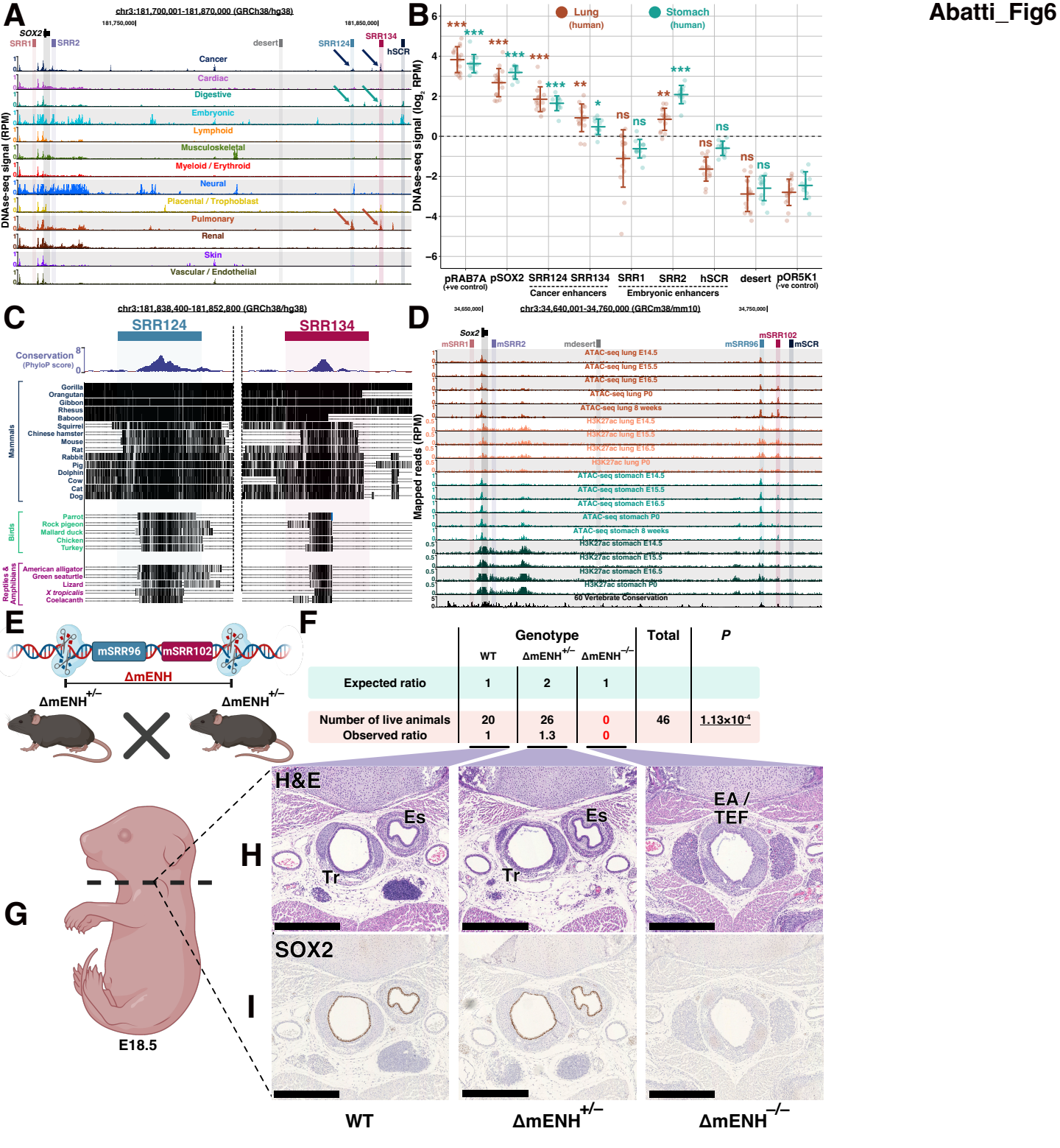

A
SRR124
SRR134
ATAC-seq signal (log2 RPM)
LUSC
LUAD
BRCA
BLCA
STAD
UCEC
HNSC
ESCA
PRAD
TGCT
COAD
LGG
GBM
LIHC
B
BLCA
BRCA
LUAD
LUSC
STAD
UCEC
pRABL7A
pSOX2
SRR124
SRR134
SRR1
SRR2
hSCR
desert
pOR5K1
C
chr3:181,700,001-181,870,000 (GRCh38/hg38)
181,700,000
181,750,000
181,800,000
181,850,000
SRR1
SRR2
SOX2
desert
SRR124
SRR134
hSCR
BLCA
BRCA
LUAD
LUSC
STAD
UCEC
D
BRCA
LUAD
LUSC
SRR124 ATAC-seq signal (log2 RPM)
SRR134 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
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

Abatti_Fig5



	Treatment	Average tagBFP signal
	+Mock (-ve control)	2761
	+FOXA1	2705
	+NFIB	1870

Abatti_Fig6



	Genotype			Total	*P*
	WT	ΔmENH$^{+/-}$	ΔmENH$^{-/-}$		
Expected ratio	1	2	1		
Number of live animals	20	26	0	46	1.13×10^{-4}
Observed ratio	1	1.3	0		