Functional maps of a genomic locus reveal confinement of an enhancer by its target gene

Mathias Eder[1,2,3]†, Christina J.I. Moene[1,2,3]†, Lise Dauban[1,2,3], Christ Leemans[1,3], Bas van Steensel[1,2,3]*

Divisions of [1]Gene Regulation and [2]Molecular Genetics, Netherlands Cancer Institute, Amsterdam, the Netherlands.
[3]Oncode Institute, the Netherlands

† These authors contributed equally to this work
* Corresponding author. Email: b.v.steensel@nki.nl

Abstract

Genes are often activated by enhancers located at large genomic distances. The importance of this positioning is poorly understood. By relocating promoter-reporter constructs into >1,000 alternative positions within a single locus, we dissected the positional relationship between the mouse *Sox2* gene and its distal enhancer. This revealed an intricate, sharply confined activation landscape, in which the native *Sox2* gene occupies an optimal position for its activation. Deletion of the gene relaxes this confinement and broadly increases reporter activity. Surprisingly, the confining effect of the *Sox2* gene is partially conferred by its ~1 kb coding region. Our local relocation approach provides high-resolution functional maps of a genomic locus and reveals that a gene can strongly constrain the realm of influence of its enhancer.

MAIN TEXT

Mammalian enhancers can activate genes across large genomic distances, up to hundreds of kilobases (*1, 2*), presumably by contacting their target promoter through folding of the chromatin fiber. Indeed, enhancers tend to have increased contact with the promoters they regulate (*3*), but it is unclear whether this is a cause or effect of their function. In addition, enhancer – promoter pairs often reside within the same topologically associated domain (TAD), a region of preferential self-interaction in the genome that is often demarcated by CTCF binding sites (*4, 5*). These TADs are thought to form functional domains (*6, 7*), but it is largely unclear whether enhancers can activate promoters anywhere within a TAD, or whether some positions within a TAD are more favorable than others. Recent studies in two loci with little three-dimensional (3D) structure indicated that promoter activity increases with decreasing genomic distance to its enhancer (*8, 9*). However, in more structured genomic loci the effect of position on enhancer – promoter communication has remained elusive. Here, we addressed these questions by constructing high-resolution maps of promoter activity as function of position in a native locus with pronounced 3D folding.

High throughput relocation by Sleeping Beauty transposition

To unravel the logic that underlies the linear arrangement of regulatory elements, it is necessary to alter the genomic position of such elements systematically and probe the functional consequences. Here we provide such a systematic transplantation of regulatory elements by adapting a methodology based on the cut-and-paste mechanism of transposable elements (*9-11*). When mobilized from the genome by its transposase, the Sleeping Beauty (SB) transposon excises and re-integrates, usually within ~1-2 Mb from the original location (**Fig. 1A**) (*10, 12*). Thus, any sequence inserted between the two SB inverted terminal repeats (ITRs) can be randomly relocated ("hopped") throughout a genomic locus. We developed a workflow to generate and map thousands of integrations in a single locus, and link the locations to the expression of a transcribed reporter, yielding high-resolution functional maps of the local regulatory landscape.

For this study we chose the *Sox2* locus in mouse embryonic stem cells (mESCs). In these cells, *Sox2* expression is primarily determined by the *Sox2* control region (SCR), which is a cluster of enhancers located 110 kb downstream of the gene in the same TAD (*13-17*). Within this TAD, the *Sox2* gene has been found to preferentially contact the SCR (*18*) (**Fig. 1B**). The 1.5 Mb region around *Sox2* contains no other protein-coding genes (*13*). To ease genetic engineering, we used F1-hybrid (*129/Sv:CAST/EiJ*) mESCs, in which the *CAST* and the *129S1* alleles can be distinguished by a high density of sequence polymorphisms. We tagged both alleles of *Sox2* with either *eGFP* (*CAST*) or *mCherry* (*129S1*) to be able to quantify allelic *Sox2* expression by flow cytometry (**Fig. S1A**). To confirm the regulation by the SCR, we transfected the

35 dual-tagged cell line with Cas9 and two gRNAs that cause deletion of the SCR. In line with the reported
36 key role of the SCR (*13, 14, 16*), a proportion of cells lost more than 90% of *eGFP* or *mCherry* expression,
37 most likely corresponding to a deletion of the SCR on the corresponding allele (**Fig. S1B**).

38 Next, we used Cas9 editing to create a "launch pad" by integrating a SB transposon cassette 6
39 kb upstream of the *Sox2::mCherry* allele (-116 kb relative to the SCR, **Fig. 1B**). The cassette contained a
40 double selection marker (mPGK-HyTK) flanked by a pair of heterotypic Flp-recombinase recognition sites
41 (FRT/F3), enabling us to easily change the DNA sequence within the transposon by recombination-
42 mediated cassette exchange (RMCE) (**Fig. S1C**).

43 In a proof-of-principle experiment, we induced hopping of the SB cassette containing an
44 arbitrary 282 bp sequence (see methods). After culturing the cells for one week to dilute the transposase-
45 containing plasmid and prevent further re-hopping events, we isolated and expanded 200,000 cells and
46 mapped the integrations using a Tn5-based sequencing approach (*19*). In some of the cells, the transposon
47 was never mobilized, which is reflected by 33% of the mapping reads originating from the launch site. Of
48 all the hopped integrations, 72% were located on the same chromosome as the *Sox2* locus (chr3) (**Fig.**
49 **S1D**). Allele-specific mapping indicated that most integrations happened in *cis* (**Fig. S1E**). Within -/+ 2 Mb
50 of the launch pad we mapped 2699 unique integrations (**Fig. 1C**), which were equally distributed between
51 the forward (1342) and reverse (1357) orientations (**Fig. S1F**). We note that this mapping is not exhaustive
52 and the actual pool likely contains more integrations, since more than 70% of the integrations were only
53 supported by one unique mapping read (**Fig S1G**). This experiment demonstrates the effectiveness of SB
54 hopping to generate thousands of genomic integrations.

55

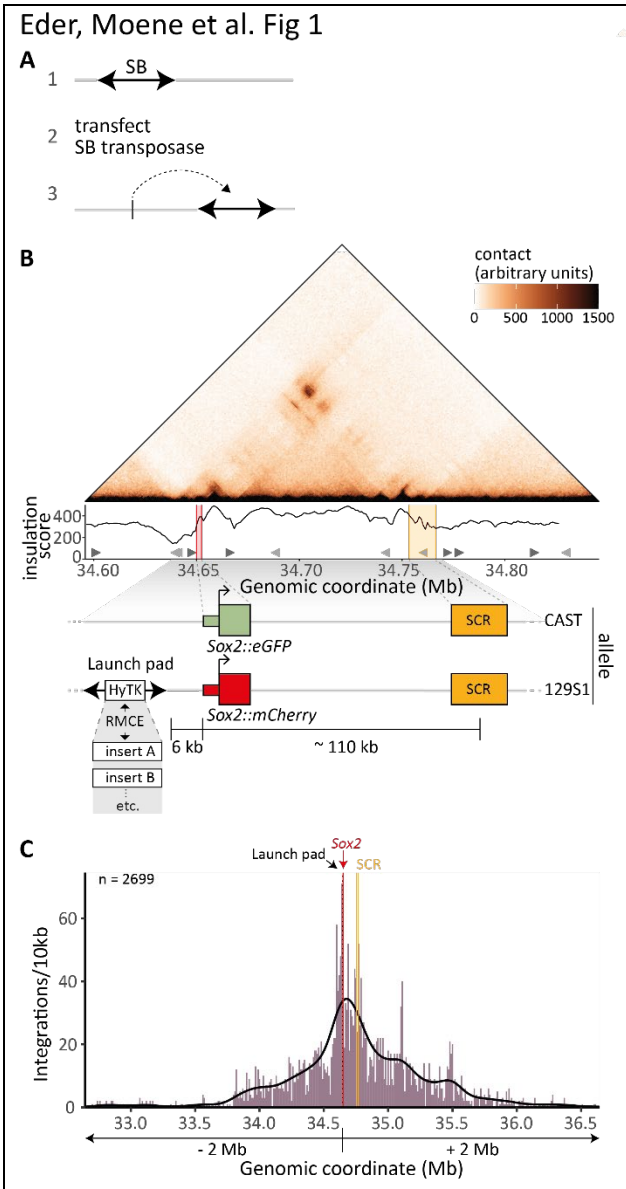


***Fig. 1* Sleeping Beauty mobilization creates integrations covering the *Sox2* locus.**

(A) Cut-and-paste local relocation of a Sleeping Beauty (SB) transposon by transient expression of SB transposase. **(B)** Overview of the *Sox2* locus. Top to bottom: contact map in mESCs (region-capture micro-C data [(*18*)], 1000-bp resolution) and insulation scores based on the same data; positions and orientations of CTCF binding sites (triangles); and positions of SCR and eGFP/mCherry tagged *Sox2* alleles, the SCR, and a launch pad (129S1 allele) containing SB with a recombination cassette. **(C)** Distribution of hopped SB carrying a 282 bp inert DNA sequence, mobilized from the launch pad; n, total number of integrations mapped in the plotted region.

66 **A highly detailed activation landscape of the *Sox2* locus**

67 Next, we set out to generate a detailed map of the expression potential of the *Sox2* locus, *i.e.*, to
68 determine quantitatively how the location throughout the locus affects the ability of the *Sox2* promoter
69 to be activated. For this we engineered a reporter comprising 1.9 kb of the *Sox2* promoter (excluding a
70 promoter-proximal CTCF binding site) driving the expression of mTurquoise2 (*mTurq*), a blue fluorescent
71 protein (Fig. 2A). We integrated this reporter into the launch pad using RMCE. This resulted in low but
72 detectable reporter expression, without affecting endogenous *Sox2::mCherry* expression (Fig. S2A,
73 reporter -116 kb).

74 We then induced SB hopping and monitored reporter fluorescence by flow cytometry. Compared
75 to the non-hopped control, the resulting cell pool showed a broadened distribution of reporter
76 fluorescence, indicating that the relocation at least in part affected reporter expression (Fig. 2B). We then
77 split the full range of reporter expression values into six gates (P1-P6) and sorted pools of cells from each
78 gate (Fig. 2B), and subsequently mapped the locations of the reporters in each pool. As a control, we also
79 mapped the locations in unsorted cells. Since the integration patterns were reproducible between
80 biological replicates (Fig. S2B-C), we combined the replicates for further downstream analysis.

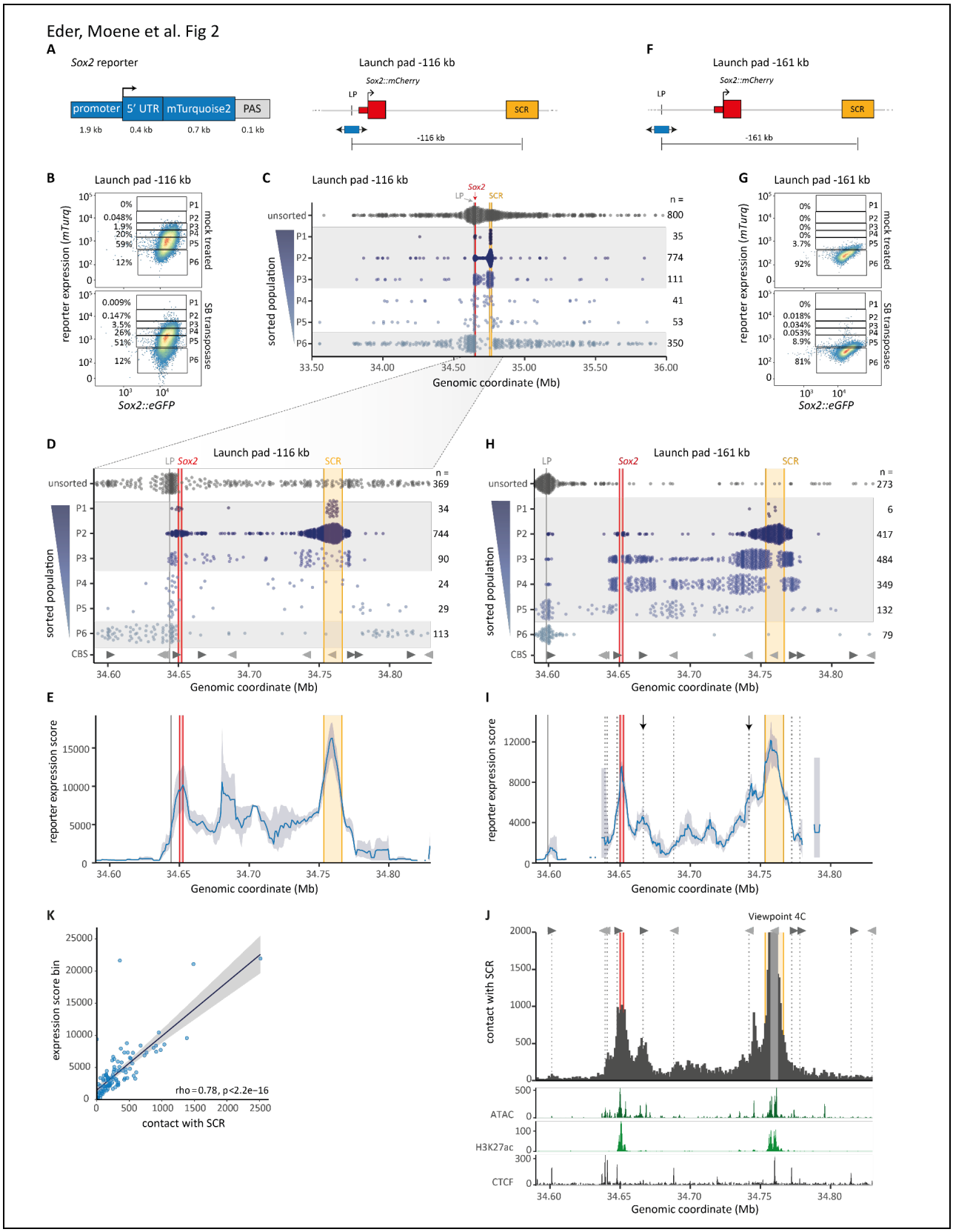
81 As we observed previously with the arbitrary insert sequence (Fig. 1C), reporter integrations in
82 the control pool are mostly concentrated in a 2 Mb region surrounding the launch site (Fig. 2C, top track).
83 In contrast, the cell pools with different expression levels yielded clearly distinct distributions (Fig. 2C-D,
84 P1 – P6 tracks). Reporters with the highest activity (P1) were concentrated almost exclusively in the center
85 of the SCR or close to the endogenous *Sox2* gene, while the second-most active reporters (P2) were more
86 dispersed around these regions. Intermediate activity (P3 – P5) occurred predominantly near the launch
87 site and between *Sox2* and SCR. Finally, silent (P6) reporters were almost exclusively located outside the
88 *Sox2*-SCR TAD.

89 We converted these raw data into a reporter expression score, which is the average expression
90 level estimated in a sliding window across the locus (Fig. 2E, see Methods). This graph highlights that the
91 SCR and the region immediately surrounding the endogenous *Sox2* gene are the most optimal locations
92 for transcriptional activity. Precise fluorescence measurements in a panel of clonal cell lines with 36
93 unique integrations confirms this peak in activity at the center of the SCR and validates the pattern of
94 expression scores inferred from the sorted pools (Fig. S2D). Reporter integrations in the center of the SCR
95 display a 10-fold upregulation in reporter expression compared to the -116 kb launch pad position. A
96 remarkable lack of reporter expression occurs outside of the *Sox2*-SCR TAD, with a very sharp drop-off
97 (within about 10 kb) on either side (Fig. 2E). A plateau of intermediate expression is found between *Sox2*
98 and SCR. Together, this detailed landscape points to a strong confinement of the realm over which the
99 SCR is able to activate transcription.

100 The relatively large confidence intervals of the expression scores limited interpretation of the fine-
101 scale pattern between *Sox2* and SCR (**Fig. 2E**). This is due to a relative paucity of reporter integrations in
102 this range. The distribution of reporter expression in cells where the reporter was not mobilized, overlaps
103 the medium-level expression gates (P4 – P5, **Fig. 2C**), leading to a low number of sorted integrations for
104 those gates. To overcome this, we repeated the hopping experiments starting from a launch pad 51 kb
105 upstream of *Sox2::mCherry* (-161 kb relative to the SCR) (**Fig. 2F**). At this position the reporter showed no
106 measurable expression (**Fig. S2A**), and hence from this launch pad it should be easier to detect hopping
107 events in the medium-level expression bins. Indeed, this was the case (**Fig. 2G-H**). The resulting expression
108 score map (**Fig. 2I**) has narrow confidence intervals within the *Sox2*-SCR TAD (at a resolution of 5 kb), at
109 the cost of a paucity of data outside this range.

110 This high-resolution map confirmed that the highest reporter expression occurs at the SCR and
111 around the endogenous *Sox2* gene (**Fig. 2I**). Reporter activity throughout the region is independent of the
112 orientation of the reporter (**Fig. S2E**). Within the *Sox2*-SCR TAD, the map reveals an intricate activity
113 landscape with several local peaks of reporter expression, two of which overlap CTCF binding sites (**Fig.
114 2I**, black arrows). Strikingly, this landscape closely matches a map of SCR contact frequencies that was
115 obtained in the absence of reporter integrations (*18*) (**Fig. 2J**, top panel; **Fig. 2K**). Maps of open chromatin
116 (ATAC-seq (*20*)) and the histone mark H3K27ac (*21*), both thought to be features of regulatory elements
117 (*22*), showed only partial overlap with the fine pattern of reporter activity (**Fig. 2J**, bottom panel). These
118 results indicate that reporter expression primarily follows contact frequency with the SCR, suggesting that
119 these contacts are a major determinant of quantitative expression levels.

120


Eder, Moene et al. Fig 2
A
Sox2 reporter
promoter
5' UTR
mTurquoise2
PAS
1.9 kb
0.4 kb
0.7 kb
0.1 kb
Launch pad -116 kb
Sox2::mCherry
LP
SCR
-116 kb
F
Launch pad -161 kb
-161 kb
B
Launch pad -116 kb
reporter expression (mTurq)
mock treated
SB transposase
Sox2::eGFP
C
Launch pad -116 kb
Sox2
unsorted
sorted population
n =
Genomic coordinate (Mb)
G
Launch pad -161 kb
D
Launch pad -116 kb
CBS
H
Launch pad -161 kb
E
reporter expression score
I
K
expression score bin
contact with SCR
rho = 0.78, p<2.2e-16
J
Viewpoint 4C
contact with SCR
ATAC
H3K27ac
CTCF

121

122 ***Fig. 2*** **A highly detailed activation landscape of the Sox2 locus. (A)** Left panel: design of mTurquoise2 reporter driven
123 by *Sox2* promoter and 5'UTR, followed by SV40 polyadenylation signal (PAS). Right panel: this reporter was inserted
124 in the -116 kb launch pad (LP), upstream of the mCherry-tagged endogenous *Sox2* allele. **(B)** Expression levels of
125 reporter (mTurq) and *Sox2::eGFP* (control) in mock-transfected cells and in cells transfected with SB transposase,
126 measured by FACS. Boxes mark sorting gates used to obtain cell populations P1-P6 in **C-D**. The percentage of cells in
127 each gate is indicated. Representative result of three biological replicates. **(C)** Mapped integrations from a large pool
128 of unsorted cells (ctrl) and the six sorted populations (P1-P6), mobilized from the -116 kb launch pad. Each dot
129 indicates one mapped integration, number of plotted integrations in each population is indicated on the right.
130 Combined data from three biological replicates. Grey shading highlights populations that are most different from
131 the starting expression of the unhopped reporter. LP, launch pad position. **(D)** Zoom-in of C. Grey triangles indicate
132 CTCF binding sites (CBS) and their orientation. **(E)** Expression score derived from data shown in **D**, smoothened using
133 a 10 kb running window shifted in 1 kb steps and only plotted when the window contains 3 or more SB integrations.
134 Shaded region indicates 95% confidence interval. **(F-I)** Same as **A-B, D-E**, except that the reporter was mobilized from
135 a launch pad at position -161 kb. Data are from 2 biological replicates combined. Expression scores in **I** are calculated
136 using a 5 kb running window shifted in 500 bp steps. Arrows mark two peaks near CBSs. **(J)** Annotation of the locus.
137 Top: virtual 4C contact profile (from RCMC data, (*18*)) with the center of the SCR (indicated in lightgrey) as viewpoint,
138 at 1 kb resolution. Dashed lines indicate CBSs and grey triangles indicate their orientation. Bottom: ATAC-seq signal
139 (*20*), H3K27ac ChIP-seq p-value signal (*21*) and CTCF ChIP-seq coverage (own data), smoothened using a 500bp
140 running window. **(K)** Correlation between expression score and virtual 4C contact with the core SCR, for the datasets
141 from -161 kb launch pad, per 1 kb bin on the genomic region plotted in **H-J**. Only bins with 3 or more SB integrations
142 are included. Rho is Spearman's correlation coefficient.

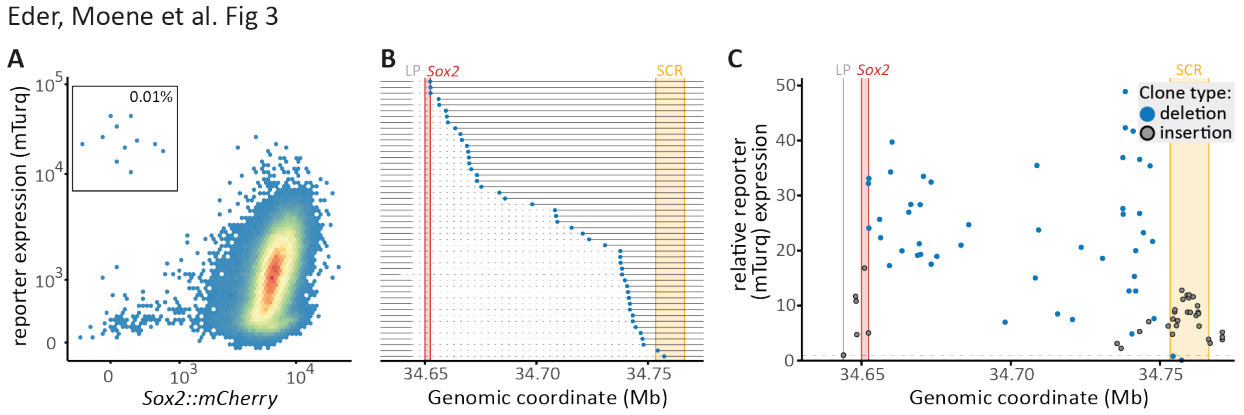
143

144 **Random deletion of the endogenous *Sox2* gene boosts reporter expression**

145 Close examination of the flow cytometry data of the transposase-transfected reporter cells (-
146 116 kb launch pad) revealed a small but distinct population of cells with high reporter expression but no
147 *Sox2*::*mCherry* expression (Fig. 3A). We isolated a panel of clones with these unusual characteristics and
148 mapped the SB location. Surprisingly, in these clones the left ITRs were still located at the launch site
149 (upstream of *Sox2::mCherry*), while the right ITRs were linked to sequences distributed throughout the
150 *Sox2* locus (Fig. 3B, S3A). This suggests that the genomic sequence between the launch site and
151 integration site has been lost. Sanger sequencing confirmed the loss of the 129S1 allele in this region (Fig.
152 S3B). We hypothesize that in rare instances, errors in the SB excision-reinsertion process can cause such
153 a deletion, which can lead to the loss of the *Sox2::mCherry* gene.

154 Strikingly, clones with such hopping-induced deletions exhibited very high reporter expression,
155 up to 49-fold higher than the -116 kb reporter clone (Fig. 3C). Reporters close to the SCR (within 20 kb)
156 were expressed about 5-fold higher when a deletion was present, compared to regular insertions in the
157 same area. Even the smallest deletions of 8.5 kb, removing only the endogenous *Sox2::mCherry* gene,
158 resulted in dramatically increased reporter expression. These data suggest that the endogenous *Sox2* gene

159 strongly limits reporter expression when present. Other sequences within the *Sox2*-SCR range may further
160 modulate this effect, indicated by the variable increases in expression across the deletions, which show
161 no correlation with distance to the SCR (Spearman's rho = 0.19, p = 0.19). In summary, these fortuitous
162 rare deletions suggested that the endogenous *Sox2* gene strongly reduces activity of the reporter gene
163 throughout the *Sox2*-SCR range.



164
165
166 ***Fig. 3* Hopping-induced deletions of the endogenous Sox2 gene increase reporter expression. (A)** Expression of
167 *Sox2::mCherry* and *Sox2* reporter (mTurq) after mobilization of SB from the -116 kb launch pad, first biological
168 replicate (also shown in Fig. 2B). Box indicates rare *Sox2::mCherry*-negative cells with high reporter expression. **(B)**
169 Unique deletions inferred from mapping in *Sox2::mCherry*-negative sorted clones (Fig. S3A), sorted by deletion size.
170 Horizontal dashed lines indicate individual deletions. Dots indicate the locations of the re-insertions. Almost all
171 deletions start at the launch pad position. **(C)** Relative reporter expression of clones with a hopping-induced deletion
172 (blue dots, location corresponds to the end of the deletion) or regular insertion (grey, same data as in Fig. S2D,
173 bottom).
174

175 ***Sox2* gene strongly controls the regulatory landscape of the enhancer**

176 To verify this interpretation, we used CRISPR/Cas9 to delete the entire *Sox2* gene, including its
177 promoter. Furthermore, to test whether the CTCF binding site (CBS) just upstream of the *Sox2* promoter
178 (which was also lost in the hopping-induced deletions) might play a role, we deleted either the *Sox2*
179 promoter and gene body (Δ*Sox2*), or the same region including this CBS (ΔCBS-*Sox2*). The guide RNAs used
180 for these deletions were not allele-specific, and thus could independently delete either the *Sox2::eGFP* or
181 *Sox2::mCherry* allele (homozygous *Sox2* deletion causes mESC differentiation (*23*)). By FACS analysis we
182 identified cells with either heterozygous deletion and measured the respective impacts on mTurq reporter
183 activity (Fig. 4A). We performed these experiments in three cell lines, with the reporter located at
184 positions -161 kb (no detectable activity), -116 kb (low activity), or -39 kb (between *Sox2* and SCR, medium
185 activity) (Fig. S4A) relative to the SCR.

186 Compared to cells that retained both *Sox2* alleles, cells which only lost *Sox2::mCherry* showed
187 strong upregulation of the reporter (on the same allele) in all three locations (**Fig. 4B**). This upregulation
188 did not occur in cells lacking the *Sox2::eGFP* allele, indicating that the effect is strictly *in cis*. The increase
189 in reporter expression was similar between the Δ*Sox2* and ΔCBS-*Sox2*. The upregulation of the reporter
190 at position -161 kb is remarkable, as it indicates that the endogenous *Sox2* gene normally prevents the
191 SCR from reaching sequences upstream, outside of the TAD. For reporters in the other two positions only
192 a sub-population of cells showed pronounced upregulation (**Fig. 4B**). The basis of this bimodality is
193 unclear.

194 To gain better understanding of the relationship between reporter location and activation, we
195 sorted out pools and clones of *Sox2::mCherry* negative cells (*cis*-deletion) with high reporter activity. After
196 cell expansion, the reporters at -161 kb and -39 kb displayed stable and uniform reporter expression (**Fig.**
197 **S4B-D**). In contrast, the -116 kb reporters again exhibited a bimodal expression pattern: a fraction of cells
198 lost reporter expression after sorting the pools (**Fig. S4B**) and clones (**S4E-F**). Comparison of the highest
199 reporter expression peaks of the sorted clones shows that the highest expression was not reached by the
200 reporter closest to the SCR (-39 kb), but rather by the reporter closest to the native location of *Sox2* (-116
201 kb) (**Fig. 4C**). Overall, these results underscore that the region around the endogenous *Sox2* gene is
202 optimal for high expression, even when this gene including its promoter is deleted.

203 To further explore the impact of the native *Sox2* gene on the regulatory landscape throughout the
204 locus, we repeated the reporter hopping experiment from the -161 kb launch site in a ΔCBS-*Sox2::mCherry*
205 clone (**Fig. S4C**). Since the founder cell line has intermediate reporter expression (**Fig. S4G**), hopped
206 integrations were expected to be most enriched in the highest (P1) and lowest (P6) expression gates. In
207 the presence of the endogenous *Sox2::mCherry* (*Sox2*$^{+/+}$), P1-level integrations were rare and only found
208 at the SCR and endogenous gene (**Fig. 4D**, bottom). In contrast, when the endogenous *Sox2* gene is
209 deleted, such highly active reporter integrations are found throughout the *Sox2*-SCR range and even
210 extend beyond the cluster of CTCF binding sites ~7 kb upstream of the deleted *Sox2* promoter. Most silent
211 (P6) reporters are only found at least 55 kb upstream of the deleted gene and 38 kb downstream of the
212 SCR, indicating that the region permissive for activity expanded by roughly 50%, extending beyond the
213 original TAD-boundaries (**Fig. 4D**, top). Together, these data demonstrate that the *Sox2* gene strongly
214 constrains both the level of activation of a competing promoter and the realm of influence of the SCR.

215

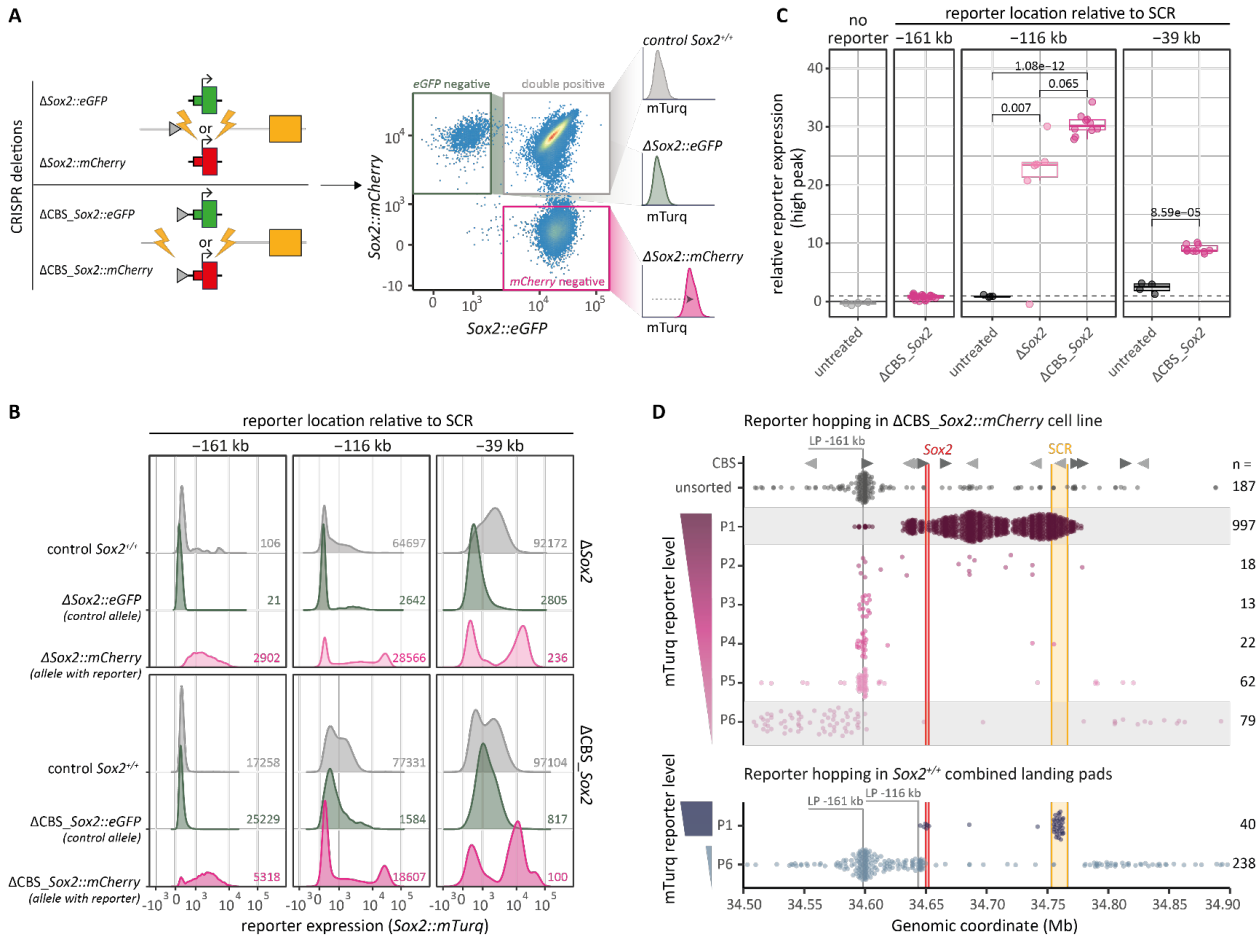


Fig. 4 **Sox2 gene strongly confines the activation realm of the SCR. (A)** Left panel: Cas9-mediated deletion of tagged *Sox2* alleles, with or without CBS (grey triangle); right panel: gating and analysis strategy. **(B)** Distribution of mTurq reporter expression levels in three cell lines with reporter inserted in positions -161 kb, -116 kb, and -39 kb, each treated and analyzed as in **A**. Numbers indicate the number of cells measured. **(C)** Maximum relative reporter expression of clonal cell lines after *Sox2::mCherry* or CBS_*Sox2::mCherry* deletion, at reporter locations as indicated. The maximum relative reporter expression is the maximum expression peak indicated in Fig. S4C-F (vertical lines), which is then autofluorescence subtracted, normalized to *Sox2::GFP*, and then normalized to the median of the standard -116 kb reporter cell line measured on the same day. Dashed line indicates 1 (standard cell line). Each dot represents one clone (CRISPR edited clones) or one measurement day (untreated clones). **(D)** Top: hopping of reporter (Sox2P) from the -161 kb launch pad in a ΔCBS_*Sox2::mCherry* cell line. Mapped integrations from a large pool of unsorted cells (ctrl) and the six sorted populations P1-P6 (same as in Fig. 2B). Data are from 2 biological replicates combined. Grey shading indicates the populations with reporter expression levels that are most different from the starting expression of the unhopped reporter. Bottom: the P1 and P6 sorted integrations combined from hopping in the original -161 kb (2 replicates) and -116 kb (3 replicates) reporter cell lines with *Sox2::mCherry* intact (same data as in Fig 2D and 2H). Grey triangles indicate CBSs and their orientation.

234 **Unbalanced competition is partially driven by the *Sox2* gene body**

235 While the endogenous *Sox2* gene significantly restricts reporter expression, we found that the
236 reciprocal impact of the reporter on the *Sox2* gene is very minor. First, the reporter inserted in the launch
237 site at -116 kb did not detectably affect *Sox2::mCherry* expression (**Fig. S2A**). Second, among all clonal
238 lines with a hopped reporter, *Sox2::mCherry* expression is remarkably robust. Even in clones with a
239 reporter in the center of the SCR (with ~10-fold higher expression than at -116 kb) *Sox2::mCherry*
240 expression is reduced by only ~25% (**Fig. 5A**). This is extremely modest compared to the 3-30 fold
241 upregulation of the reporter upon loss of *Sox2::mCherry* (**Fig. 4C**). Hence, the endogenous gene affects
242 the reporter considerably more than *vice versa*.

243 We wondered what might give the endogenous *Sox2* gene this competitive edge. The reporter
244 contained the exact promoter and 5' UTR sequence from the *Sox2* gene, but it lacked the 1.0 kb coding
245 sequence (CDS) and the 1.1 kb 3' UTR and polyadenylation signal (PAS) (*Sox2* has no introns). Re-analysis
246 of high-resolution region capture micro-C data (*18*) indicates that the SCR not only contacts the promoter
247 of the endogenous *Sox2* gene, but also the gene body (**Fig. S5A**), indicating that the latter sequence might
248 play a role. We therefore generated two new reporters in which we added the *Sox2* CDS alone or in
249 combination with the 3' UTR/PAS (**Fig. 5B**). We kept the *mTurquoise2* sequence separated from the CDS
250 by a P2A cleavage signal (*24*), so the stability of the fluorescent protein would not be affected by a fusion
251 with the SOX2 protein.

252 We first tested these new reporters in positions -161 kb and -116 kb after deletion of
253 *Sox2::mCherry*. Strikingly, across multiple independent clonal cell lines, the CDS and CDS/3'UTR additions
254 boosted reporter activity in the -161 kb location by 6-8 fold (**Fig. 5C, Fig. S5B**). Reporter insertion in the -
255 116 kb location, just 6 kb upstream of the deleted *Sox2* gene, is a close approximation of replacing the
256 gene with a reporter. Here, the promoter-only reporters showed very high but bimodal expression (**Fig**
257 **5C, S5B, S5C**), as observed before (**Fig S4B, E, F**). Surprisingly, the CDS-containing reporters had lower but
258 stable expression (**Fig. 5C, S5C**). We speculate that this lower expression is the result of a negative
259 feedback mechanism on high SOX2 protein levels, which can be triggered by the CDS-containing reporter
260 but not the original *Sox2P* reporter (*25, 26*). Overall, these results suggest that the regulation of the CDS-
261 containing reporter more closely resembles the native *Sox2* gene.

262 Next, we tested the effect of the CDS when the competing *Sox2::mCherry* gene is also present in
263 the locus. In both the -116 kb and -39 kb position, clonal cell lines with the CDS-containing reporters had
264 3-7 fold higher expression than those with the original reporter (**Fig. 5D** top, **Fig. S5D**). We then asked
265 whether the presence of the CDS also shifts the competitive balance with the endogenous *Sox2* gene.
266 Indeed, in these cells the addition of the CDS to the reporter caused a 13-18% reduction in *Sox2::mCherry*
267 expression, while additional inclusion of the 3' UTR in the reporter even caused a 21-31% reduction. (**Fig.**
268 **5D** bottom, **Fig. S5D**).

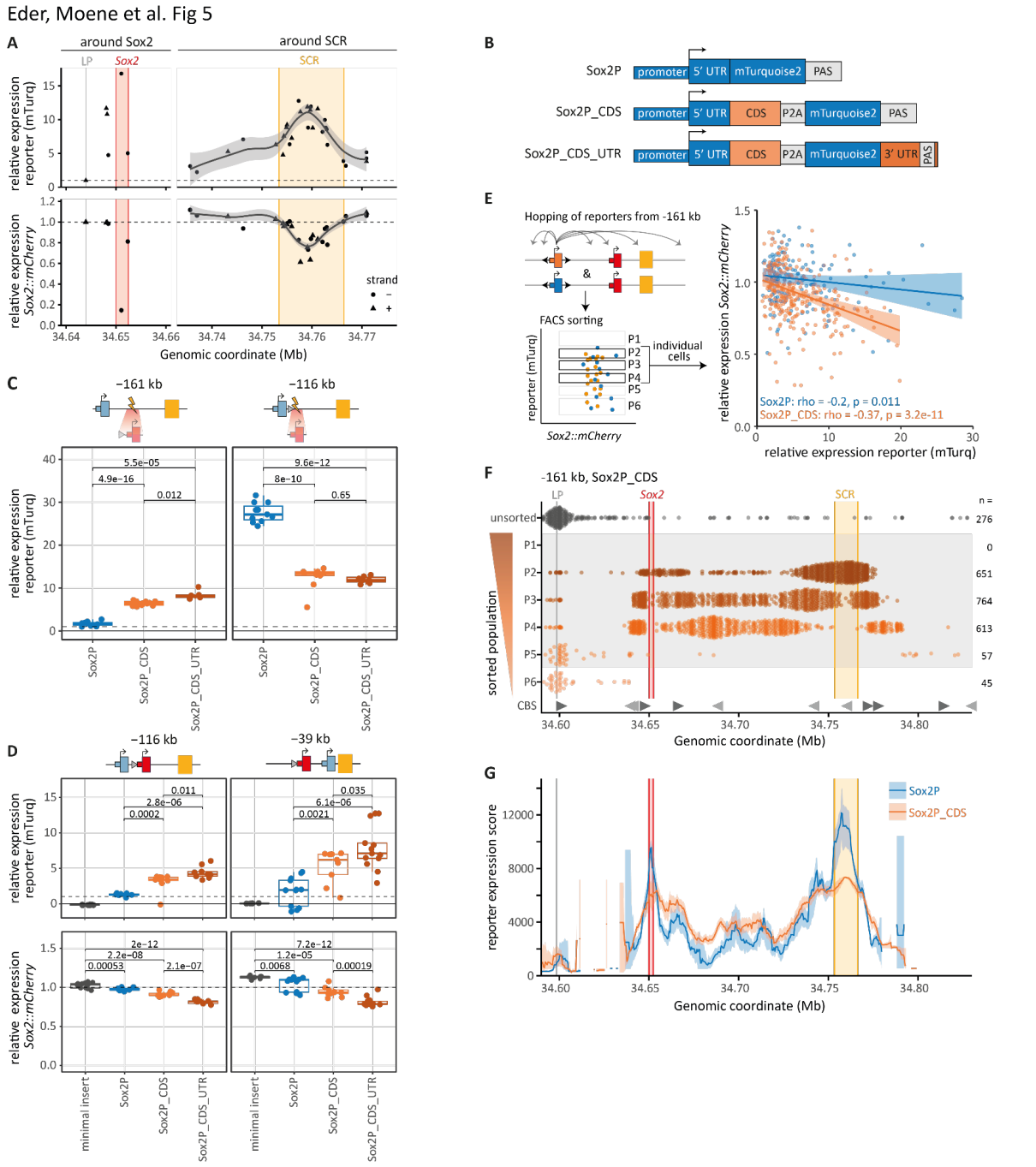


Fig. 5 **Unbalanced competition is partially driven by Sox2 gene body. (A)** Relative reporter (mTurq) and *Sox2::mCherry* expression in clones around *Sox2* and the SCR (median expression, estimated autofluorescence subtracted, normalized to GFP, and then normalized to the unhopped reporters at -116 kb). Loess curve shows the trend in expression. Same clones and mTurq data as in Fig. S2D, bottom. **(B)** Design of original *Sox2* promoter reporter (Sox2P) and two coding sequence (CDS)-containing reporters. PAS, poly-adenylation signal. **(C)** Relative expression levels of the three reporters at -161 kb and -116 kb, in a ΔCBS_*Sox2::mCherry* cell line. Each dot represents an individual clone. Relative expression was calculated as in **A**. Dashed line indicates 1 (control cell line). Fig. S5B shows the same expression values, not normalized to *eGFP*. P-values according to Welch's unequal variances *t*-test.

278 **(D)** As C, but for the relative reporter (top) and endogenous *Sox2*::*mCherry* (bottom) expression of the three
279 reporters and a minimal insert (not containing any promoter or mTurq) inserted at -116 kb and -39 kb launch pads
280 with *Sox2*::mCherry intact. Fig. S5D shows the same expression values, not normalized to *eGFP*. **(E)** Left panels:
281 schematic of hopping (from launch pad -161 kb) and sorting Sox2P and Sox2P_CDS reporters based on reporter
282 expression (P1-P6). Right panel: correlation between relative *Sox2::mCherry* and reporter (mTurq) expression from
283 the Sox2P and Sox2P_CDS reporter, from cells with reporter expression in the gates P2-P4 (see Fig. 2G and S5H).
284 Relative expression was calculated as in **A**, but per individual cell. Rho, Spearman's correlation coefficient. Data are
285 from two biological replicates combined. For visualization, the y-axis is cut at 1.5, excluding 8 outlier points (they are
286 included in the correlation calculation); Fig. S5F shows the same data with uncut axes. **(F)** Mapping of the
287 integrations in sorted populations P1-P6 (Fig. S5E) and unsorted control cells (ctrl). Visualization as in Fig. 2D; data
288 from 2 biological replicates combined. **(G)** Expression score of the Sox2P and Sox2P_CDS reporter across the *Sox2*
289 locus (-161 kb launch pad); 5 kb running window shifted in 500 bp steps. Expression score of the Sox2P reporter is
290 same as in Fig. 2I. Shaded region indicates the 95% confidence interval. Expression score is only plotted when the
291 window contains 3 or more SB integrations.
292

293 To test whether this competitive characteristic of the CDS reporters is a general feature
294 throughout the locus, we mobilized the CDS-containing reporter from the -161 kb location – where it is
295 not detectably expressed (Fig. S5E) – in cells carrying the intact *Sox2::mCherry* (Fig. 5E). We compared the
296 relative *Sox2::mCherry* and reporter expression of individual cells with active reporters, as measured by
297 flow cytometry. If the reporter competes with *Sox2::mCherry*, this should be reflected in reduced
298 expression of the latter when the reporter is more active. With the reporter lacking the CDS, this effect is
299 detectable but weak (Fig. 5E, S5F). However, when the CDS is present, this effect is about three times
300 stronger ($p = 0.008$), as indicated by the respective slopes of a fitted regression model (Fig. 5E, S5G). Thus,
301 across the random active locations sampled in this hopping experiment, the CDS strengthens the reporter
302 in its competition with the *Sox2* gene.

303 To further assess if the presence of the CDS in the reporter can change its activation landscape
304 throughout the locus, we mapped the integrations in cells sorted for different reporter expression levels
305 from the CDS hopping experiment (Fig. S5H). The sorted integrations of CDS-containing reporter follow a
306 similar pattern as the original reporter (Fig. 5F), indicating that the two reporters are stimulated by the
307 SCR in a similar way. However, we did not detect any cells with the highest reporter expression (P1),
308 causing flattening of the activity landscape near the SCR and the endogenous *Sox2* location. This is possibly
309 explained by the same negative feedback that we proposed above. Despite this, the CDS boosted the
310 reporter activity throughout the *Sox2*-SCR range (excluding *Sox2::mCherry* and SCR regions) (Fig. 5G).
311 Together, these data indicate that the CDS has two effects on promoter activity: it can boost expression
312 independent of competition; and it increases the competitive edge of a promoter.

313

314 **Discussion**

315 Here we developed a novel high-throughput “hopping” approach to create highly detailed quantitative
316 maps of the activation potential of a native genomic locus, based on thousands of reporter integrations.
317 We expect that this approach may also be used to systematically relocate other regulatory elements, such
318 as enhancers, CBSs and other structural elements, and a wide range of synthetic elements.

319 Earlier probing at low resolution in mouse embryos suggested that promoters respond to an
320 enhancer whenever they are in the same TAD (*11, 27*). Systematic insertion of enhancers and promoters
321 in heterologous genomic loci supported this notion, and pointed to a monotonous decay of promoter
322 activity with increasing genomic distance to the enhancer (*8, 9, 28*). Our high-resolution functional maps
323 of the native *Sox2* locus confirm that the influence realm of the SCR coincides with the TAD, but also reveal
324 an intricate landscape in which pre-existing SCR contact frequency rather than genomic distance is a key
325 determinant of promoter activity. Moreover, the region around the endogenous *Sox2* gene is a "sweet
326 spot" for activation, even in the absence of the gene itself. Possibly, synergistic interactions between the
327 SCR and enhancer-like elements surrounding the gene (*16*) contribute to this.

328 Another striking result is that the *Sox2* gene is not only passively controlled by the SCR, but also
329 actively constrains the SCR, as indicated by its strong dominance over the reporter. Various studies of
330 native and artificial loci, both in *Drosophila* (*29, 30*) and mammals (*31-33*), showed competition between
331 two promoters for activation by one enhancer. In each case, the enhancer preferentially activated the
332 closest promoter. In contrast, our results show that the SCR strongly favors the native *Sox2* gene
333 regardless of the location of the reporter. Surprisingly, this preference is in part mediated by the *Sox2*
334 coding sequence. Possibly the CDS DNA sequence binds one or more transcription factors that boost SCR
335 interactions; or the transcribed CDS RNA facilitates SCR interactions, as has been proposed for other
336 transcripts (*34*). Our functional maps of SCR – gene interplay lay the foundation for further mechanistic
337 dissection.

338 **References**

339 1. J. H. Yang, A. S. Hansen, Enhancer selectivity in space and time: from enhancer-promoter
340 interactions to promoter activation. *Nat Rev Mol Cell Biol* **25**, 574-591 (2024).

341 2. C. Paliou *et al.*, Preformed chromatin topology assists transcriptional robustness of Shh during limb
342 development. *Proc Natl Acad Sci U S A* **116**, 12390-12399 (2019).

343 3. C. P. Fulco *et al.*, Activity-by-contact model of enhancer-promoter regulation from thousands of
344 CRISPR perturbations. *Nat Genet* **51**, 1664-1669 (2019).

345 4. J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin
346 interactions. *Nature* **485**, 376-380 (2012).

347 5. E. P. Nora *et al.*, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*
348 **485**, 381-385 (2012).

349 6. G. R. Cavalheiro, T. Pollex, E. E. Furlong, To loop or not to loop: what is the role of TADs in
350 enhancer function and gene regulation? *Curr Opin Genet Dev* **67**, 119-129 (2021).

351 7. J. Dekker, E. Heard, Structural and functional diversity of Topologically Associating Domains. *FEBS*
352 *Lett* **589**, 2877-2884 (2015).

353 8. N. J. Rinzema *et al.*, Building regulatory landscapes reveals that an enhancer can recruit cohesin
354 to create contact domains, engage CTCF sites and activate distant genes. *Nat Struct Mol Biol* **29**,
355 563-574 (2022).

356 9. J. Zuin *et al.*, Nonlinear control of transcription through enhancer-promoter interactions. *Nature*
357 **604**, 571-577 (2022).

358 10. V. W. Keng *et al.*, Region-specific saturation germline mutagenesis in mice using the Sleeping
359 Beauty transposon system. *Nat Methods* **2**, 763-769 (2005).

360 11. O. Symmons *et al.*, Functional and topological characteristics of mammalian regulatory domains.
361 *Genome Res* **24**, 390-400 (2014).

362 12. L. Dauban *et al.*, Genome - nuclear lamina interactions are multivalent and cooperative. *bioRxiv*,
363 2024.2001.2010.574825 (2024).

364 13. Y. Li *et al.*, CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse
365 embryonic stem cells. *PLoS One* **9**, e114485 (2014).

366 14. H. Y. Zhou *et al.*, A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation
367 potential. *Genes Dev* **28**, 2699-2711 (2014).

368 15. T. Taylor *et al.*, Transcriptional regulation and chromatin architecture maintenance are decoupled
369 functions at the Sox2 locus. *Genes Dev* **36**, 699-717 (2022).

370 16. R. Brosh *et al.*, Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2
371 locus. *Mol Cell* **83**, 1140-1152 e1147 (2023).

372 17. B. Bonev *et al.*, Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-
373 572 e524 (2017).

374 18. V. Y. Goel, M. K. Huseyin, A. S. Hansen, Region Capture Micro-C reveals coalescence of
375 enhancers and promoters into nested microcompartments. *Nat Genet* **55**, 1048-1056 (2023).

376 19. D. L. Stern, Tagmentation-Based Mapping (TagMap) of Mobile DNA Genomic Insertion Sites.
377 *bioRxiv*, 037762 (2017).

378 20. H. W. King, N. A. Fursova, N. P. Blackledge, R. J. Klose, Polycomb repressive complex 1 shapes
379 the nucleosome landscape but not accessibility at target genes. *Genome Res* **28**, 1494-1507
380 (2018).

381 21. A. Sethi *et al.*, Supervised enhancer prediction with epigenetic pattern recognition and targeted
382 validation. *Nat Methods* **17**, 807-814 (2020).

383 22. M. Gasperini, J. M. Tome, J. Shendure, Towards a comprehensive catalogue of validated and
384 target-linked human enhancers. *Nat Rev Genet* **21**, 292-310 (2020).

385 23. S. Masui *et al.*, Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse
386 embryonic stem cells. *Nat Cell Biol* **9**, 625-635 (2007).

387 24. A. L. Szymczak, D. A. Vignali, Development of 2A peptide-based strategies in the design of
388 multicistronic vectors. *Expert Opin Biol Ther* **5**, 627-638 (2005).

389 25. J. L. Kopp, B. D. Ormsbee, M. Desler, A. Rizzino, Small increases in the level of Sox2 trigger the
390 differentiation of mouse embryonic stem cells. *Stem Cells* **26**, 903-911 (2008).

391 26. B. D. Ormsbee Golden, E. L. Wuebben, A. Rizzino, Sox2 expression is regulated by a negative
392 feedback loop in embryonic stem cells that involves AKT signaling and FoxO1. *PLoS One* **8**,
393 e76345 (2013).

394 27. O. Symmons *et al.*, The Shh Topological Domain Facilitates the Action of Remote Enhancers by
395 Reducing the Effects of Genomic Distances. *Dev Cell* **39**, 529-543 (2016).

396 28. H. Thomas *et al.*, Enhancer cooperativity can compensate for loss of activity over large genomic
397 distances. *bioRxiv*, 2023.2012.2006.570399 (2023).

398 29. T. Fukaya, B. Lim, M. Levine, Enhancer Control of Transcriptional Bursting. *Cell* **166**, 358-368
399 (2016).

400 30. J. R. Bateman, J. E. Johnson, Altering enhancer-promoter linear distance impacts promoter
401 competition in cis and in trans. *Genetics* **222**, (2022).

402 31. N. Dillon, T. Trimborn, J. Strouboulis, P. Fraser, F. Grosveld, The effect of distance on long-range
403 chromatin interactions. *Mol Cell* **1**, 131-139 (1997).

404 32. S. Oh *et al.*, Enhancer release and retargeting activates disease-susceptibility genes. *Nature* **595**,
405 735-740 (2021).

406 33. S. K. Topfer *et al.*, Disrupting the adult globin promoter alleviates promoter competition and
407 reactivates fetal globin gene expression. *Blood* **139**, 2107-2118 (2022).

408 34. L. Liang *et al.*, Complementary Alu sequences mediate enhancer-promoter selectivity. *Nature* **619**,
409 868-875 (2023).

410 35. J. C. Rivera-Mulia *et al.*, Allele-specific control of replication timing and genome organization during
411 development. *Genome Res* **28**, 800-811 (2018).

412 36. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif.
413 *Bioinformatics* **27**, 1017-1018 (2011).

414 37. R. Schep *et al.*, Impact of chromatin context on Cas9-induced DNA double-strand break repair
415 pathway balance. *Mol Cell* **81**, 2216-2230 e2210 (2021).

416 38. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823
417 (2013).

418 39. F. Hahne *et al.*, flowCore: a Bioconductor package for high throughput flow cytometry. *BMC*
419 *Bioinformatics* **10**, 106 (2009).

420 40. A. G. Doran *et al.*, Deep genome sequencing and variation analysis of 13 inbred mouse strains
421 defines candidate phenotypic alleles, private variation and homozygous truncating mutations.
422 *Genome Biol* **17**, 167 (2016).

423 41. T. M. Keane *et al.*, Mouse genomic variation and its effect on phenotypes and gene regulation.
424 *Nature* **477**, 289-294 (2011).

425 42. G. Finak, W. Jiang, R. Gottardo, CytoML for cross-platform cytometry data sharing. *Cytometry A*
426 **93**, 1189-1196 (2018).

427 43. M. Malek *et al.*, flowDensity: reproducing manual gating of flow cytometry data by automated
428 density-based cell population identification. *Bioinformatics* **31**, 606-607 (2015).

429 44. S. Chen, Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using
430 fastp. *Imeta* **2**, e107 (2023).

431 45. R. H. van der Weide *et al.*, Hi-C analyses with GENOVA: a case study with cohesin variants. *NAR*
432 *Genom Bioinform* **3**, lqab040 (2021).

433

Acknowledgments

We thank the Giorgetti lab for sharing the *Sox2* fosmid; the NKI Genomics, Flow Cytometry, Protein Production and Research High Performance Computing core facilities for excellent support; Marantha Kaagman for valuable experimental contributions; and Elphege Nora, Luca Giorgetti, Elzo de Wit, Tineke Lenstra and members from our lab and the divisions of Gene Regulation and Molecular Genetics for inspiring and helpful discussions.

Funding

This project was funded by the Boehringer Ingelheim Fonds (C.J.I.M.); the European Research Council: GoCADiSC, 694466 (B.v.S.); RE_LOCATE, 101054449 (B.v.S.) and the Oncode Institute. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Research at the Netherlands Cancer Institute is supported by an institutional grant of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport. The Oncode Institute is partially funded by the Dutch Cancer Society.

Author Contribution

Conceptualization: M.E., C.J.I.M., B.v.S.
Wet-lab experiments: M.E., C.J.I.M., L.D.
Data analysis and visualization: C.J.I.M., M.E.
Method development: M.E., L.D.
Code development: C.L., C.J.I.M.
Writing: M.E., C.J.I.M., B.v.S.
Funding Acquisition: B.v.S., C.J.I.M.
Supervision: B.v.S.

Competing interests:

The authors declare no conflict of interest

Data and materials availability:

Laboratory notebooks and supplementary data (including primer sequences, gRNAs, plasmids, etc.) are available on Zenodo (10.5281/zenodo.13350269). Raw sequencing data is available on GEO under accession numbers GSE275427.

Code and supplementary files used as input for the scripts are available on Zenodo (10.5281/zenodo.13362255).

All plasmids used will be made available to Addgene.

MATERIALS AND METHODS

A. EXPERIMENTAL PROCEDURE

Cell culture

F121/9 (CAST/EiJ x S129/Sv) (RRID:CVCL_VC42) female mouse embryonic stem cells (mECSs) F1 hybrid cell line (*35*) and derived clones were cultured in Serum+Lif+2i condition. Briefly, cells were cultured on gelatin-coated (0.1%) culture plates in Glasgow minimum essential medium (Sigma-Aldrich, G5154) supplemented with 15% fetal bovine serum (Thermo Fisher Scientific, 10270-106), 1% L-glutamine (Thermo Fisher Scientific, 25030024), 1% sodium pyruvate (Thermo Fisher Scientific, 11360-70), 1% MEM non-essential amino acids (Thermo Fisher Scientific, 11140-50), 1% penicillin-streptomycin (Thermo Fisher Scientific, 15070063), 100 µM ß-mercaptoethanol, 10 x 4 U leukemia inhibitory factor (LIF; esg1107, Millipore), 1 µM MEK inhibitor PD0325901 (Mirdametinib, MedChemExpress), 3 µM GSK-3β inhibitor CHIR99021 (Laduviglusib, MedChemExpress) in 5% CO_2 at 37 °C. Cells were passaged every 2 days. Mycoplasma contamination was ruled out by regular testing (#LT07-318; Lonza).

Genetic engineering of the *Sox2* locus

Tagging endogenous Sox2 alleles

The alleles of the *Sox2* gene were tagged by direct fusion with eGFP and mCherry using CRISPR/Cas9 mediated homologous recombination. We used a previously described plasmid for targeting *Sox2* (*Sox2* sgRNA, Addgene #175553) together with donor plasmids designed to include eGFP or mCherry in between two homology arms (200 bp homologies) for the *Sox2* gene. In brief, $1x10^6$ F121/9 mESCs were transfected with 1 µg of *Sox2* sgRNA and 2 µg of repair template (1 µg of eGFP + 1 µg of mCherry) using 9 µL Lipofectamine2000. Cells were plated in Serum+Lif+2i medium together with a DNAPKcs inhibitor (M3814, MCE, HY-101570) and cultured for two days. Subsequently, mCherry+/eGFP+ cells were sorted sparsely into a 24-well plate (cultured without inhibitor) and resulting colonies were picked and genotyped by PCR. Then they were re-measured by flow cytometry (LSRFortessaTM) using the following lasers and filters: eGFP 488nm BL[B] 530/30, mCherry 561nm YG[D] 610/20. Allele-specific genotyping of *Sox2* was performed by PCR amplification from the fluorophore to the genome across CAST/129S1 SNPs (primers: EMp324, EMp325, MEP68). A clone with the CAST allele *Sox2* gene fused with eGFP and the 129S1 allele *Sox2* gene fused with mCherry was selected as the parental clone.

Establishing -116 kb HyTK launch pad

We cloned the vector pME015 containing an mouse *PGK* promoter driving the expression of a double selection marker HyTK (HyTK amplified from plasmid #11684, Addgene), flanked by FRT and F3 heterotypic recombination sites embedded between Sleeping Beauty 5' and 3' inverted terminal repeats (ITRs). For establishing the -116 kb launch pad, we used the Alt-R™ CRISPR-Cas9 system from IDT and designed a crRNA that directs cutting 6 kb upstream (-116 kb from SCR) of the endogenous *Sox2*. The crRNA was assembled into an RNP together with the AltR CRISPR tracrRNA (1073189, IDT) and Cas9 nuclease

(1081058, IDT). We transfected $50x10^3$ parental cells (*Sox2* GFP+/mCherry+) with the RNP and the PCR product using lipofectamine CRISPRMAX Cas9 Transfection Reagent (CMAX00001, Thermo Fisher Scientific) according to the manufacturer's protocol. Two days after transfection, Hygromycin was added to the medium (200 µg/mL, 10687010, Invitrogen) to select for HyTK-expressing cells. We picked colonies 6 days later and screened clones for cassette integration by PCR and Sanger sequencing.

Establishing -39 kb HyTK launch pad

We created a launch pad in-between *Sox2* and SCR (-39 kb relative to SCR) by CRISPR/Cas9 mediated homology directed repair. For this, we added 1 kb homology arms of the target region upstream and downstream of the HyTK construct (pME015) to generate plasmid pME024, assembled the crRNA with the tracrRNA and Cas9 protein as explained above, and co-transfected the RNP complex together with the pME024 repair template into parental cells (*Sox2* GFP+/mCherry+). Similar as above, cells were selected with Hygromycin and colonies were picked and genotyped by PCR and Sanger sequencing.

RMCE donor plasmids

Minimal insert

We generated 2000 random 50 nucleotide sequences in silico with a GC content of 50% and scanned them by FIMO for transcription factor binding sites (*36*). We selected one sequence without any transcription factor binding sites, added FRT and F3 sites to the start and end of the sequence and ordered it as an DNA ultramer from IDT. We amplified this sequence by PCR with primers that have 40 bp overlap to the digested ends (BamH1+XhoI) of the target plasmid (pCR Zero, #120275, Addgene) (primers: MEP190, MEP191). Finally we used Gibson assembly to assemble the linearized pCR-Zero vector together with the amplified PCR product according to the manufacturer's protocol. Two µL of Gibson assembly mix were transformed into 50 µL DH5α (NEB, C2987H) chemically competent cells. Clones were screened by colony PCR and validated by Sanger sequencing.

Sox2 reporter construct

We amplified 2346 bp of the *Sox2* promoter (containing the 5'UTR) from a fosmid (Wl1-1017O07, position mm9: chr3:34535173-34570689, kindly provided by Luca Giorgetti's lab) by PCR using Phusion Polymerase (F-530L, Thermo Fisher Scientific) together with DMSO and GC rich buffer (primers: MEP145, MEP150). The resulting PCR product was cloned into pCR Zero vector (#120275, Addgene) digested with StuI (R0187S, NEB) and transformed into 50 µL DH5α (C2987H, NEB) chemically competent cells. Resulting colonies were picked and genotyped by PCR and validated by Sanger sequencing. To generate the *Sox2P*-mTurquoise2 plasmid, the *Sox2*-promoter-containing pCR Zero plasmid was digested with XbaI + SalI-HF (R3138S, NEB) and ligated (T4 Ligase 5 U/µL, 9015-85-4, Roche) with the XbaI+SalI-HF-digested mTurquoise2-plasmid (#118617, Addgene), resulting in a plasmid containing the *Sox2* promoter with the mTurquoise2 fluorophore. Finally, an SV40 polyA signal was cloned 3' of the mTurquoise2 (primers:

MEP171, MEP172) and F3 and FRT heterotypic Flp-recombinase sites were added by PCR (primers: MEP147, MEP177), resulting into the final pME034 plasmid. The plasmid was validated by Sanger sequencing.

Sox2-CDS reporter constructs

We generated two additional *Sox2* reporters containing the *Sox2* coding sequence (957 bp) and either the endogenous *Sox2* 3'UTR or the same polyA signal (SV40) as the original Sox2 reporter construct (pME034). We linearized the reporter plasmid (pME034) with SalI-HF & EcoRV-HF, and PCR amplified the mTurquoise2 gene ± polyA signal from pME034 (primers: MEP224, MEP224, MEP231) and *Sox2*-3'UTR (primers: MEP226, MEP230) and *Sox2*-CDS (primers: MEP222, MEP223) from the same fosmid as before (Wl1-1017O07). A P2A sequence with linker was added by the forward primer for the mTurquoise2 gene (MEP224). We used Gibson assembly to create either the *Sox2*-CDS-polyA(SV40) or the *Sox2*-CDS-3'UTR plasmids. Next, we checked the resulting plasmids by colony PCR and subsequent Sanger sequencing. The final plasmids were named pME040 (F3_*Sox2*_CDS_mTurq_3'UTR_FRT) and pME041 (F3_*Sox2*_CDS_mTurq_polyA_FRT).

Recombination-mediated cassette exchange (RMCE)

Inserts were loaded into the SB cassette by recombination-mediated cassette exchange (RMCE). 300,000 cells were transfected with 0.5 µg of Flp recombinase-encoding plasmid (Addgene #13787) and 2 µg of donor plasmid, using 7.5 µL lipofectamine 2000 (Invitrogen, 11668019). Two days after transfection, pools were seeded sparsely and ganciclovir (2.5 µg/mL) was added to the medium to kill HyTK-expressing cells. Resistant colonies were picked, expanded, and screened by PCR. To replace the SB insert in a Sox2P-containing clone derived by SB hopping or CRISPR editing, we first exchanged the Sox2P reporter for the original HyTK cassette by co-transfecting 400,000 to 500,000 cells with 0.5 µg of Flp recombinase-encoding plasmid and 2 µg of a HyTK-containing plasmid (pME015). Two days after transfection, pools of HyTK-expressing cells were selected by adding Hygromycin to the medium. Next, various constructs were loaded into these HyTK-containing pools using RMCE, as explained above.

Establishing Sox2-CDS clones

Clones containing the Sox2P, Sox2P-CDS or Sox2P-CDS-UTR construct in different launch pads (-39 kb, -116 kb, -161 kb) were generated by RMCE and single cell sorted in 96 well plates. Subsequently, clones were genotyped by PCR. Only clones containing the right insert and genomic background were measured by flow cytometry (*Sox2::mCherry, Sox2::eGFP*, reporters) and used for downstream analysis.

Establishing proof of principle cell line

For a different project we used a plasmid containing a stretch of 200 random nucleotides (free of any CTCF motifs), followed by a CTCF binding site flanked by LoxP sites (kindly provided by Luca Braccioli, Elzo de Wit group). We PCR amplified the insert with flanking F3/FRT sites (primers: MEP74, MEP75) and cloned the resulting PCR product into StuI blunt cut pCR-Zero vector, creating donor plasmid pME013. Next, we used RMCE to integrate the insert into the -116 kb launch pad of *Sox2* GFP+/mCherry+ F121/9 mESCs, resulting in a cell line called CBS16. To generate a ΔCTCF control cell line (CRE6), we transfected the cells with a CRE-Puro plasmid (pJK14-946-CRE-Puro), selected with Puromycin and seeded surviving cells as single cells in 96 well plates by FACS (BD FACSAriaTM Fusion Flow Cytometer). Resulting clones had an insert of 282 bp in between FRT/F3 sites and were genotyped for the floxed CTCF motif by PCR and validated by Sanger sequencing. Those cells were used for the proof of principle experiment. The sequence of the resulting insert (200 arbitrary nucleotides and one loxP site) can be found in the supplementary data sheet.

SB Hopping

SB hopped cell pools

For each replicate and cell line, SB hopping was induced by transfecting 2.5 to 5 million cells with the bicistronic plasmid (pME07) encoding SB transposase SB100x and the human nerve growth factor receptor (LNGFR), using 15 µL Lipofectamine 2000 (Invitrogen, 11668019) and 5 µg of plasmid per million cells. As a control, 0.5 million cells were transfected with a similar bicistronic plasmid encoding PiggyBac (PB) transposase and LNGFR (pLD042). After 30 hours, transfected cells were enriched using MACS MS columns (130-042-201, Miltenyi Biotec) and LNGFR MicroBeads (130-091-330, Miltenyi Biotec) according to the manufacturer's protocol. One week after transfection, 0.2 million SB-transposase-transfected cells were set apart as an unsorted pool and the remaining cells were sorted by FACS (BD FACSAriaTM Fusion Flow Cytometer) into pools for populations based on mTurq expression level, using the following lasers and filters: mTurq: 442nm V[F] 470/20, eGFP: 488nm BL[B] 530/30, mCherry: 561nm YG[D] 610/20 (see Fig 2B for the exact gates). PB transposase transfected cells served as mock-treated FACS control. In the first biological replicate of the -116 kb Sox2P-reporter (**Fig 2B, 3A**) cells were only selected based on mTurq level, in all later replicates we selected only *Sox2::mCherry*-positive cells for all *Sox2::mCherry* containing cell lines and only *Sox2::mCherry*-negative cells for the ΔCBS_*Sox2::mCherry* cell line. Cells were sorted into up to 5 pools of up to 1000 cells per pool. Sorted pools were expanded and crude lysates were obtained from a full 96 well plate by lysing with 50 µL DirectPCR Lysis Reagent (102-T, Viagen Biotec) supplemented with 100 µg/mL proteinase K, and incubating 2.5 h at 55 °C and 45 min at 85 °C. After expansion, gDNA was extracted from the unsorted control cell pools using the ISOLATE II Genomic DNA Kit (BIO-52067, Bioline).

Establishing clones

To generate a panel of clonal cell lines with active reporters, we FACS sorted single cells either from previously sorted P1 and P2 pools (biological replicate 1) or directly from the full pool of SB-transposase-transfected cells (replicates 2 and 3, only deletion clones). mCherry-positive and mCherry-negative cells were sorted to obtain clones with hopped insertions and hopping-induced deletions, respectively. When sorting from the full pool of SB-transfected cells we also gated for high (P1) mTurq expression. The cell line with a reporter integration at -161 kb was isolated in an earlier SB hopping experiment (not described here), by sorting out mTurq-negative cells from a pool of SB-transposase-transfected cells (from the -116 kb reporter cell line). Sorted clones were expanded and crude lysates were obtained from a full 96 well plate, as for the cell pools. Deletion clones presumably derived from one hopping event (from the same experiment with identical location and orientation) were averaged.

Genotyping hopping-induced deletion

A few hopping-induced deletion clones were genotyped to confirm the genomic deletions. We used PCR (primers: CMP116-121) to amplify three regions containing annotated SNPs on the CAST allele and Sanger sequencing to evaluate which allelic fragment(s) got amplified. By examining the Sanger sequencing traces, we could categorize if and which allele got deleted (**Fig. S3B**).

Tagmentation mapping

SB integrations were mapped by amplifying and sequencing the ITR-genome junction using a tagmentation-based approach, based on (*19*). The procedure was similar to (*37*), but using the following SB-ITR specific primers: MEP009 (5' ITR, reverse) or LD027 (3' ITR, forward) for the linear enrichment and MEP011 (5' ITR, reverse) or MEP034 (3' ITR, forward) for PCR1. Input for one tagmentation reaction was either 100ng gDNA or 3 µL crude lysate. The 72 °C extension times of all PCRs were increased to 1 min. In addition, the composition of the linear enrichment PCR was changed to 10 µL of tagmented DNA, 2 µL of 1 µM primer, 4 µL dNTPs (10mM), 8 µL 5x Phusion® HF Buffer (NEB), 0.5 µL Phusion® HS Flex polymerase (2 U/µL - NEB), in a final volume of 40 µL. For all clones and some experiments with sorted pools (proof-of-principle experiment, third replicate of the reporter hopping) we halved this reaction to a final volume of 20 µL. For the unsorted control pools, 10 separate tagmentation reactions and library preparations were performed, usually using one set of sequencing indices. Libraries were sequenced with 150 bp paired-end reads using either an Illumina MiSeq system (reagent kit v2) or Illumina NextSeq 550 system (mid-output kit) including 10% of PhiX spike-in.

CRISPR deletions

Deletion endogenous Sox2

To delete the endogenous *Sox2* gene, we designed two gRNAs targeting upstream of the CTCF site, one at the start of the *Sox2* promoter (not targeting the reporter) and three downstream of the *Sox2* gene (one in the 3' UTR, two after). gRNAs were cloned into an expression plasmid containing mCherry. 0.5 million cells were transfected with 1 µg Cas9 plasmid (PX330) and 1.5 µg gRNA plasmid pool (equimolar pool) using 7.5 µL lipofectamine 2000. Transfected cells were isolated after one day, by sorting cells with increased mCherry expression. More than three weeks later, *Sox2* and reporter expression were measured in these cells by flow cytometry (BD FACSAria™ Fusion) and *Sox2::mCherry* negative cells with high reporter expression were sorted as pools. *Sox2::mCherry* negative, reporter-high clones were either isolated at the same time (-116 kb ΔCBS_*Sox2*), or isolated later from the sorted pools by single cell dilution (-116 kb *ΔSox2*) or by FACS sorting *Sox2::mCherry*-negative mTurq-positive cells (-161 kb ΔCBS_*Sox2*, -39 kb ΔCBS_*Sox2*). We isolated gDNA or crude lysates (as before) for all clones and performed a PCR across the intended deletion (*ΔSox2*: CMP075+CMP081; ΔCBS_*Sox2*: CMP077+CMP081). In addition, we performed a PCR from the intact GFP or 3' UTR to the downstream region (CMP024+CMP081 or CMP080+CMP081) to confirm that the *Sox2::GFP* allele was not inverted. Clones where either PCR failed were excluded from all analyses.

SCR deletion

We cloned two previously published gRNA sequences (*13*) targeting 5' or 3' of the SCR into a Cas9 expressing px330 backbone (#158973, Addgene) according to the protocol published by the Zhang lab (*38*). Used gRNA sequences were reported to work in mESCs in an earlier publication (*13*). Resulting plasmids were verified by Sanger sequencing and transfected into *Sox2* GFP+/mCherry+ F121/9 mESCs. In brief, 0.5 million cells were transfected with 0.5 µg 5'SCR-sgRNA plasmid and 0.5 µg 3'SCR-sgRNA plasmid using Lipofectamine2000. Cells were analyzed for eGFP and mCherry expression by FACS (LSRFortessaTM) after 5 days of culturing using the following lasers and filters: eGFP BL[B] 530/30, mCherry YG[D] 610/20.

Flowcytometry

Allele-specific *Sox2* expression and reporter expression of clones and pools were measured on a BD LSRFortessa™, using the BD™ High Throughput Sampler (HTS) to directly sample from 96 well plates. The following lasers and filters were used: eGFP 488nm BL[B] 530/30, mCherry 561nm YG[D] 610/20, mTurquoise2 405nm V[F] 470/20. Single cells were gated using FlowJo and exported as fcs files which were loaded in R using the flowCore package (*39*) for further analysis.

CTCF ChIP

F121/9 (CAST/EiJ x S129/Sv) *Sox2* mCherry+/eGFP+ were cultured in Serum+Lif+2i condition. For chromatin preparation, 10 million cells (per condition) were cross-linked with a final concentration of 1%

formaldehyde for 10 min. The cross-linking reaction was quenched with 2.0 M glycine (0.2 M final concentration). The cross-linked cells were then lysed and sonicated to obtain approximately 300 bp chromatin fragments with a Bioruptor Plus sonication device (Diagenode). For ChIP assays, antibodies were first coupled with Protein G dynabeads (10003D, Thermo Fisher Scientific), and the sonicated chromatin was then incubated overnight at 4 °C. After incubation, the beads were washed, the captured chromatin was eluted and cross-linking was reversed. The released DNA fragments were purified with the MinElute PCR Purification kit (28004, Qiagen). The ChIP experiment was performed with the following antibody: CTCF (07-729, Merck Millipore; 5 µL per ChIP). The purified DNA fragments were prepared according to the protocol of the KAPA LTP Library Preparation kit (KR0453, Roche) before sequencing. All ChIP-seq libraries were sequenced with single-end 150 bp cycle mode on an Illumina NextSeq 550 (mid-output kit).

B. DATA PROCESSING AND COMPUTATIONAL ANALYSIS

Computational analysis of SB hopping

Tagmentation mapping

We used the pipeline published in (*12*) to map SB integrations. For experiments using the -161 kb launch pad we mapped to the mouse reference genome (Release M23 – mm10), for the -116 kb launch pad we mapped to mm10 with an *in silico* insertion of the SB cassette (including exact flanking indels from the CRISPR insertion) and lifted over the integration positions to the mm10 genome. In the pilot experiment (**Fig. 2C**), the pipeline was also used to assign integrations to the 129S1 or CAST allele based on SNPs in the mapping reads, using SNP annotation of CAST and 129S1 alleles from the Mouse Genomes Project (https://www.mousegenomes.org/publications/ (*40, 41*)). Reads were filtered for a minimal mapping quality of 10. For the sorted pools, integrations with at least two unique mapping reads were retained. To ensure that high reporter expression is due to location and not due to hopping-induced *Sox2::mCherry* deletions, the integrations in the in the P1 pools were further filtered for having at least one read mapping each from the 5' ITR and 3' ITR (reverse and forward read). For the large unsorted control pools we filtered integrations less stringently, including any integration with at least one mapping read. Note that the integrations in these large unsorted pools are not mapped exhaustively. In a few experiments, integrations exactly matching the -39 kb launch site (chr3:34721183-34721192) or an earlier derived hopped clone (chr3:25018987 and chr3:35232946) were found across many pools. This indicated a PCR contamination from an earlier experiment, therefore we filtered out all integrations exactly matching these coordinates.

Clones

To identify the true integrations in clonal cell lines, we filtered more stringently, requiring at least ten unique mapping reads. Next, we discarded any locations that were supported by <20% of the read count of the top location, since these are likely to be either index-swapped reads from other clones or off-target reads. To call an insertion with confidence, we required to find exactly one integration supported by 5' ITR and 3' ITR mapping. For confident deletion clones, we required to find exactly two locations: one

supported by the 5' ITR mapping, and one by the 3' ITR mapping. Any clones with more mapped locations were manually reviewed: any clones for which the extra reads could be explained by index-swapping were included, while clones with evidence of a secondary integration or an unexplained high number of mapping reads from a secondary location were discarded.

Expression score from sorted pools

To calculate expression scores from sorted pools, we first combined the integrations from each experiment per sorted population. Identical integrations found in multiple sorted pools from one experiment most likely originated from a single hopping event. Still, we treated them as separate data points because they were independently sorted into an expression gate. To calculate the expression score for any cell line, we first took the median fluorescence intensity ($Fluo_P$) of each sorted population P1 through P6 that was measured during the FACS sorting process. For -161 kb Sox2P P1 cells were sorted but no cells were recorded (frequency was less than 1 in 100.000 cells), so the mean $Fluo_P$ from -116 kb Sox2P P1 was used. Next, for any given window (W), the expression score was defined as the average of these population fluorescence values, weighted by the fraction of all integrations of each population falling within that window ($\frac{N_{W,P}}{N_P}$):

$$expression\ score_W = \sum_{P=1}^{6} (\frac{N_{W,P}}{N_P} * Fluo_P) \ / \sum_{P=1}^{6} \frac{N_{W,P}}{N_P}$$

Whenever replicate data were pooled, the median fluorescence per gate varied slightly between biological replicates so the fluorescence of each experiment (E) was weighted by the number of integrations from that experiment:

$$expression\ score_W = \sum_{E=1}^{n} \sum_{P=1}^{6} (\frac{N_{W,P,E}}{N_P} * Fluo_{P,E}) \ / \sum_{P=1}^{6} \frac{N_{W,P}}{N_P}$$

To obtain a running window expression score (e.g. 10 kb window), integrations were counted in smaller bins (e.g. 1 kb, this is the step size) and the expression score was calculated from a group of bins (e.g. 10 bins). Window and step size were selected based on the density of integrations in each experiment and region of interest. To estimate a 95% confidence interval, the genome-wide list of integrations for each cell line was bootstrapped 5000 times and the expression score was calculated from each bootstrap. Reported are the median of these bootstrapped expression scores as well as the 2.5th to 97.5th percentile (95% confidence interval), for each window with at least 3 sorted integrations. To obtain the strand-specific expression score, only integrations on a specific strand were included in integration count of a window ($N_{W,P}$), while integrations on both strands were included in the total number of integrations of the population (N_P).

Computational analysis flowcytometry

FACS color density plots

To show the exact FACS gating used to sort SB-transposase-transfected cells, we loaded the FACSDiva experiment file (.xml) and the raw fcs files in R using diva_to_gatingset from the CytoML package (*42*). Color density plots were created using ggcyto and geom_hex. The other color density plots (e.g. showing *Sox2::mCherry* and *Sox2::GFP* negative populations after CRISPR treatment) were created from gated single cells exported from FlowJo. The threshold for GFP or mCherry negative cells was determined by the 99th percentile of the WT control, while the threshold for GFP or mCherry positive cells was 1.5 times this value.

Normalization pools and clones

To compare the *Sox2* and reporter expression levels between clones (**Fig. 3C**), the fluorescent values were corrected and normalized. First, the median fluorescence intensity (MFI) was calculated for each fluorophore and each clone. Next, the MFIs of a mESC control cell line lacking any fluorescent protein were subtracted from the respective MFIs of the other samples, to correct for autofluorescence. Subsequently, the corrected mCherry and mTurquoise2 values were normalized by the corrected eGFP value of the same sample, to adjust for any global effects on *Sox2* expression. Because there was variation in the absolute MFIs between measurement days, we finally divided the normalized fluorescence values by the respective value of the original -116kb_Sox2P cell line, measured on the same day. In this way, we obtained relative *Sox2::mCherry* and reporter (mTurquoise2) expression levels that could be compared between measurement days and experiments (**Fig. 3C, 4C, 5A, 5C-D**). To compare the expression of all fluorophores without adjusting for global changes in SOX2 level, we divided the autofluorescence-corrected MFI values by those of the -116kb_Sox2P cell line ('relative expression, not normalized', **Fig. S5B, S5D**). The measurements of the clones derived by SB hopping (insertions and deletions, e.g. **Fig. 3C**) contained no wild-type control, so we used the wild-type control from a different day with closely matching fluorescent values of the original -116kb_Sox2P cell line.

Normalization single cells

To compute the relative reporter and *Sox2::mCherry* expression of individual SB-transposase-transfected cells (**Fig. 5E**), the MFIs of the wild-type mESC sample were subtracted from the fluorescent values of each cell. Next, the corrected mCherry and mTurquoise2 values were normalized by the eGFP value of the same cell. Finally, all individual cell measures were scaled to the corrected, normalized MFIs of the -116kb Sox2P cell line to be able to combine the data from two measurement days (biological replicates). Only cells in P2-P4 were used because P5 and P6 contain primarily many non-mobilized reporters (so individual cells do not represent a unique integration location) and no cells were recorded in P1.

Bimodal expression patterns

To determine the modes of the fluorescence density distribution in clones with a bimodal reporter expression pattern, we used the function getPeaks from the flowDensity package (*43*) on the bi-exponentially transformed data (as in the figures). We extracted at most two highest peaks and selected the peak with the highest expression value. This mTurquoise2 value was corrected and normalized to compare between experiments as previously explained for MFI values.

Computational analysis CTCF ChIP

Raw sequencing data were filtered (>=Q15) and adapters were trimmed using fastP (*44*). Sequences were mapped to mm10 reference genome using the BWA alignment tool (0.7.17-r1188) after indexing the mm10 reference genome. The coverage files (bigwig files) were generated with RPKM normalization using deepTools (version 3.0). Peak calling was performed with MACS2 at a q-value cutoff of 0.05 and CTCF motifs were called using Jaspar (CTCF:MA0139.1) database running MotifScan (v1.3.0). By manual inspection of the CTCF bigwig files we could see that one site was clearly bound but not annotated by the MotifScan. Therefore we added it manually (mm10; chr3: 34772210).

External data

Region capture micro C

Processed region capture micro-C (RCMC) data from (*18*), aligned to the mm39 genome, was visualized using the GENOVA package (*45*). For the contact matrices in **Fig. 1B** and **Fig. S5A**, all annotations of the *Sox2* locus were lifted over to the mm39 genome. Insulation score was calculated using the function insulation_score, on the 1 kb resolution data, with window = 20 (20 kb window size). Trans-interactions between the *Sox2* gene and SCR were visualized using the function trans_matrixplot, using the 200 bp resolution RCMC data. To correlate contact probability to expression score, we created a virtual 4C on the mm39 genome using as a viewpoint the region containing the two essential enhancer elements in the SCR (mm39: chr3:34811722-34816213, (*16*)) at 1 kb resolution. Then we lifted over the contact probabilities to the mm10 genome build and calculated the expression score on the same 1 kb bins. Only bins with at least 3 reporter integrations were included in the correlation.

Supplementary Figures

Eder, Moene et al. Fig S1

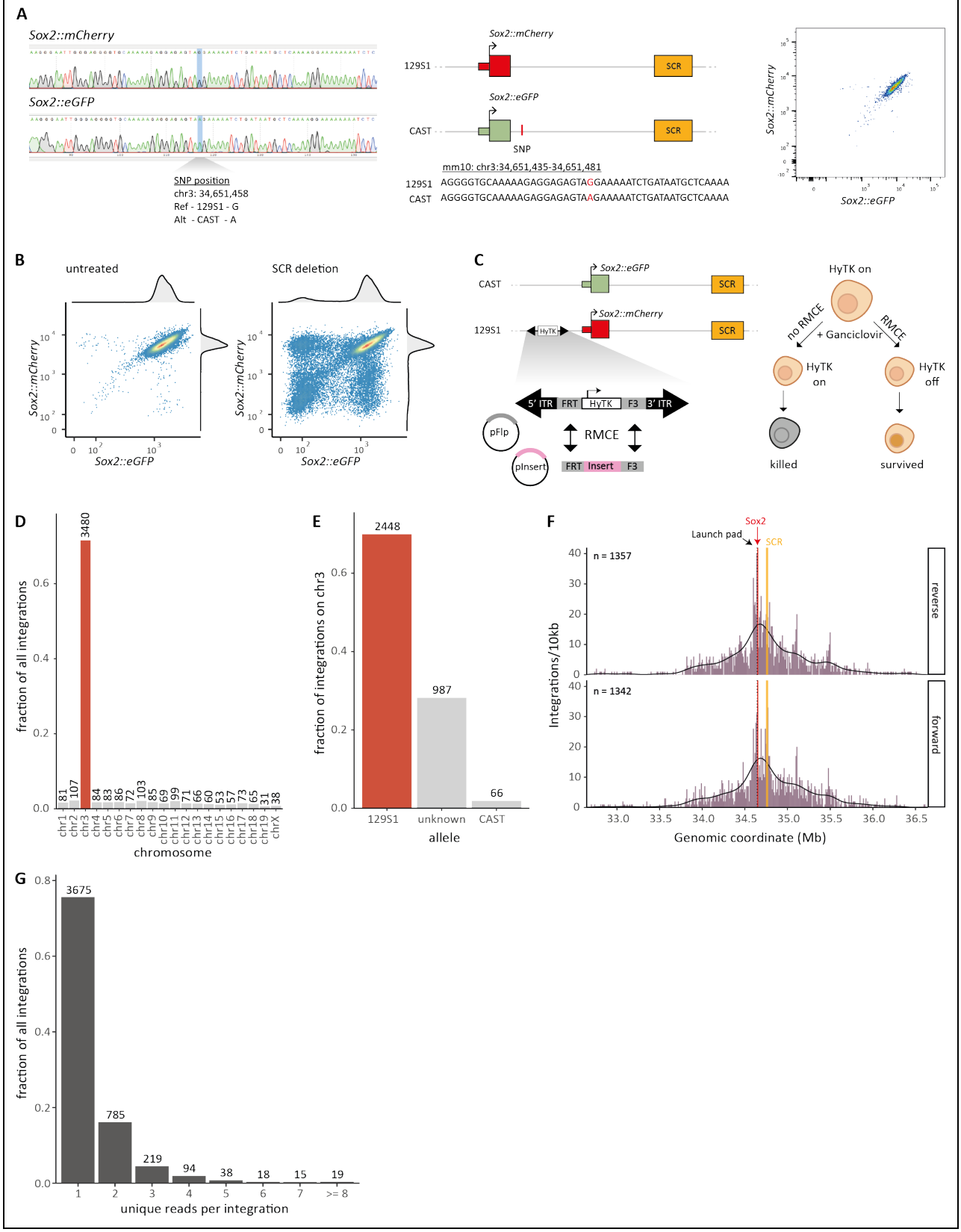


Fig. S1. **Cell line generation and proof of principle experiment**. **(A)** Left and middle: PCR and Sanger sequencing with mCherry and eGFP specific primers indicates on which allele each fluorophore is inserted. Right: expression level of Sox2::mCherry and Sox2::eGFP measured by flow cytometry. **(B)** Allele-specific Sox2 expression in the dual tagged cell line, in cells treated with CRISPR to delete the SCR and in untreated control cells. **(C)** Recombination-mediated cassette exchange: The hygromycin phosphotransferase-thymidine kinase fusion gene (HyTK) is flanked by heterotypic Flp recombinase recognition sites (FRT & F3) and embedded into a Sleeping Beauty transposon (SB-ITRs), integrated via CRISPR-Cas9-mediated knock-in 6 kb upstream of the Sox2::mCherry fusion gene. Co-transfection of the Flp recombinase expression plasmid (pFlp) and donor plasmid (pInsert) with subsequent ganciclovir selection for 7 days results in cells with the intended insert. **(D-G)** Statistics from the proof-of-principle hopping experiment, where a SB transposon carrying a 282 bp arbitrary DNA sequence was relocated. Number of integrations is indicated (above the bars or n). **(D)** Fraction of integrations mapped per chromosome. The launch pad is on chr 3. **(E)** Fraction of all integrations on chr 3 mapped to each allele, or not assigned to either allele. The launch pad is on the 129S1 allele. **(F)** Number of integrations per 10 kb bin, split by SB orientation. **(G)** Distribution of the number of unique mapping reads supporting each integration. **(B, D-G)** Data from one (proof-of-principle) biological replicate.

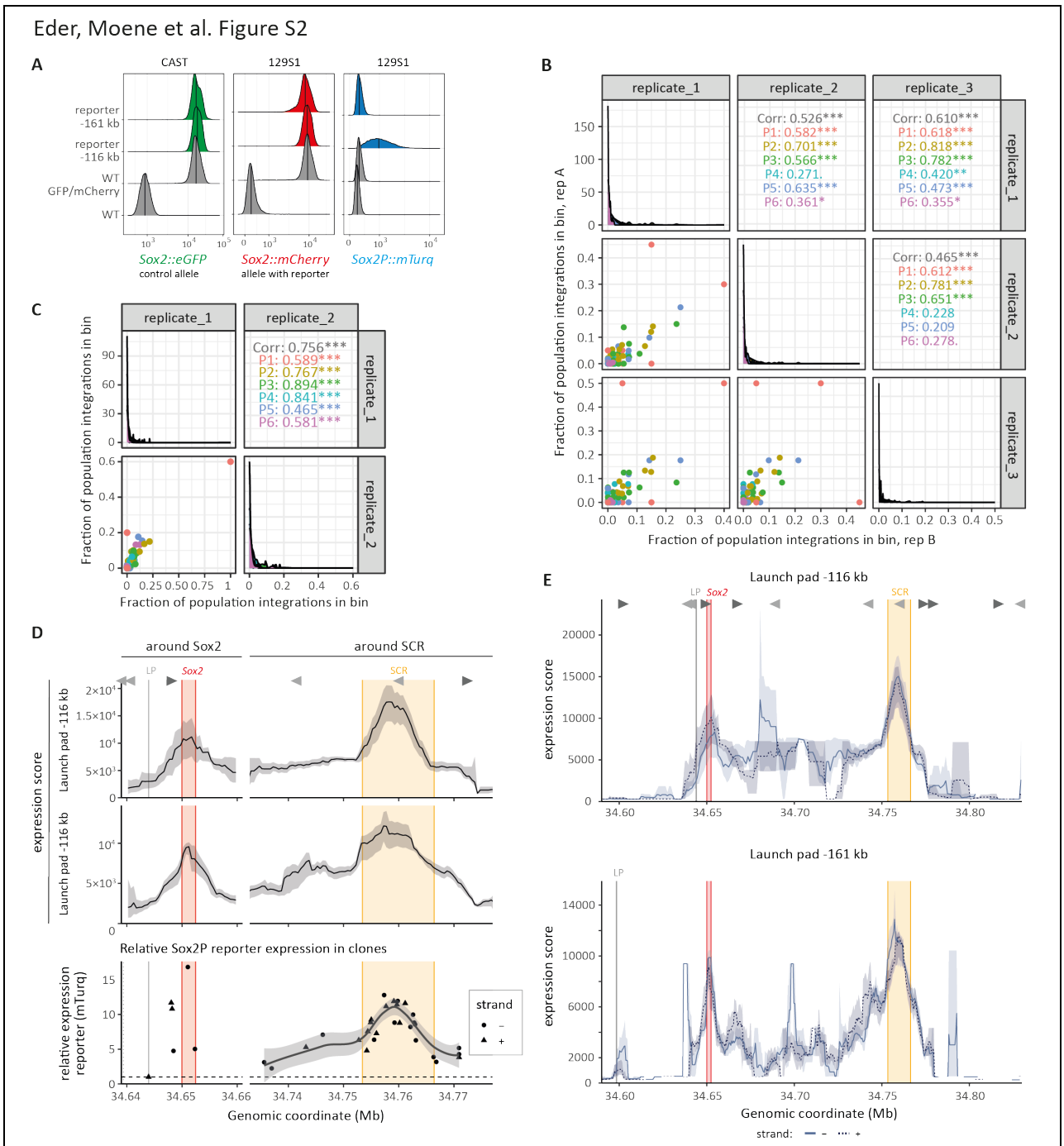


Fig. S2. **Reproducibility of reporter hopping and validation of the expression score. (A)** Distribution of expression of *Sox2::eGFP* (green), *Sox2::mCherry* (red) and the reporter (mTurq, blue), in the -161 kb and -116 kb reporter cell lines and the parental control cell lines measured by flow cytometry. **(B)** Reproducibility of the fraction of sorted integrations across the *Sox2* locus, from the -116 kb launch pad. Bottom left panels: each dot represents the fraction of all integrations of that population found in a 5 kb bin in the *Sox2* locus, compared between two biological replicates. Colors indicate the 6 sorted populations. Panels along diagonal: distribution of observed fractions in each replicate. Top right panels: Spearman correlation between two replicates. **(C)** As B, but for the -161 kb launch pad, with two biological replicates. **(D)** Top: the computed expression scores around *Sox2* gene and SCR based on the sorted integrations from the -116 kb and -161 kb launch pad (see Fig. 2D, H), smoothened using a 5 kb running

window shifted in 500 bp steps. Shaded region indicates 95% confidence interval. Grey triangles indicate CTCF binding sites (CBS) and their orientation. LP, launch pad position. Bottom: relative reporter expression (median mTurq, estimated autofluorescence subtracted, normalized to eGFP, and then normalized to the unhopped reporters at -116 kb) of clones around the *Sox2* gene and the SCR. Loess curve shows the trend in expression. **(E)** Expression score of the positive and negative strand integrations (forward and reverse oriented promoter, respectively), across the *Sox2* locus, based on the integrations from the -116 kb and -161 kb launch pad. Smoothened using a 10 kb running window, shifted in 1 kb steps. Shaded region indicates 95% confidence interval. Locus annotation as in **D**.

Eder, Moene et al. Fig S3

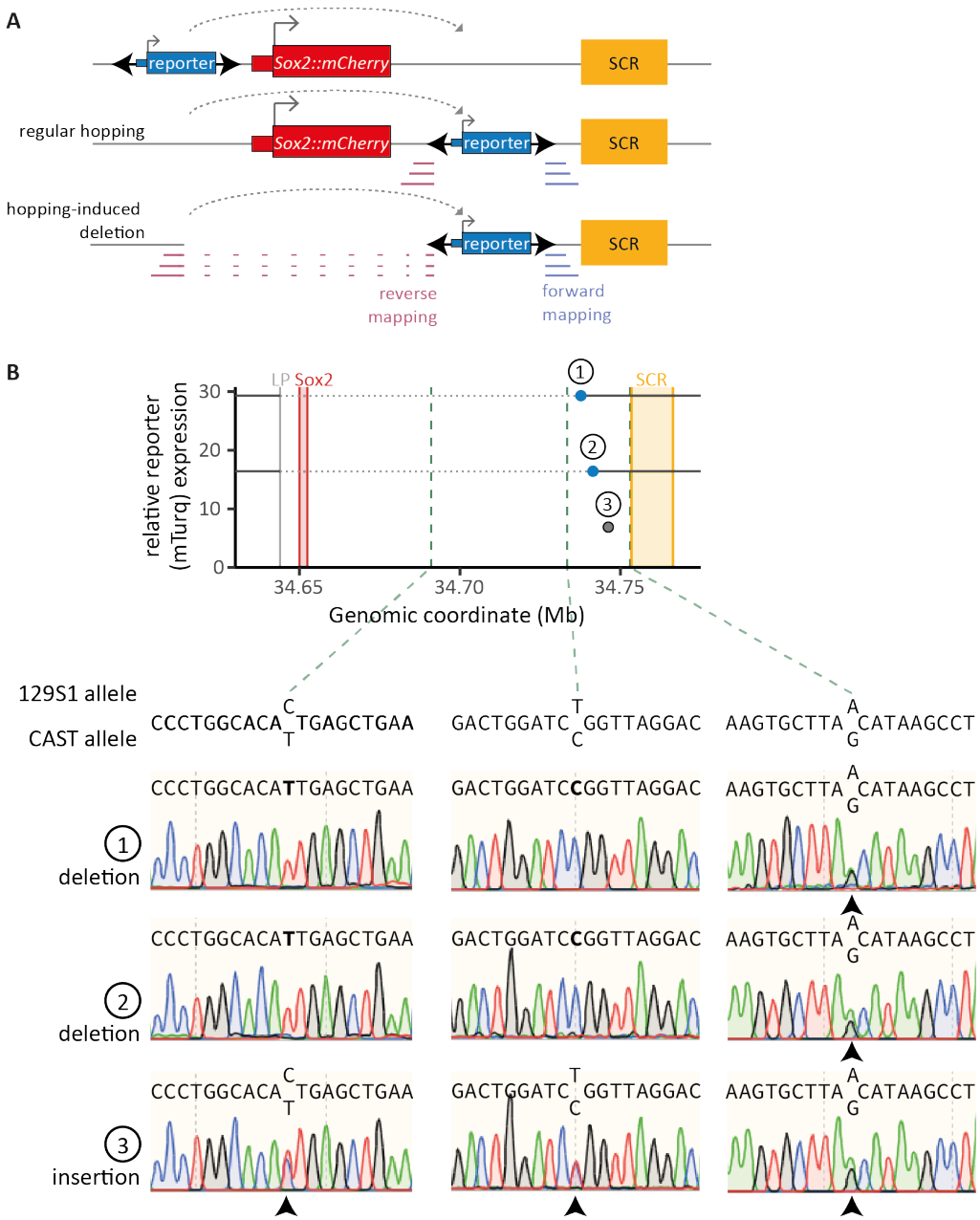


Fig. S3. **Mapping of hopping induced deletions and insertions. (A)** Schematic of Tn5-based mapping in cells with a regular hopping event or a hopping-induced deletion. In case of a hopping-induced deletion, one SB ITR will be mapped to the launch site while the other is mapped to the insertion location. **(B)** Confirmation of the loss of the 129S1 allele in the deleted region of deletion clones. Top: genomic location and expression of two hopping-induced deletion clones (1, 2, blue) and one regular insertion clone (3, grey). Horizontal dashed grey lines indicate expected deletions, vertical dashed green lines indicate the locations of three known SNPs between the 129S1 and CAST alleles. Bottom: Sanger sequencing across the three SNPs in the three cell lines. Black arrowheads indicate double peaks at the SNPs, indicating that both alleles are present. When only the CAST allele is present, the SNP is indicated in bold in the sequence above the Sanger trace.

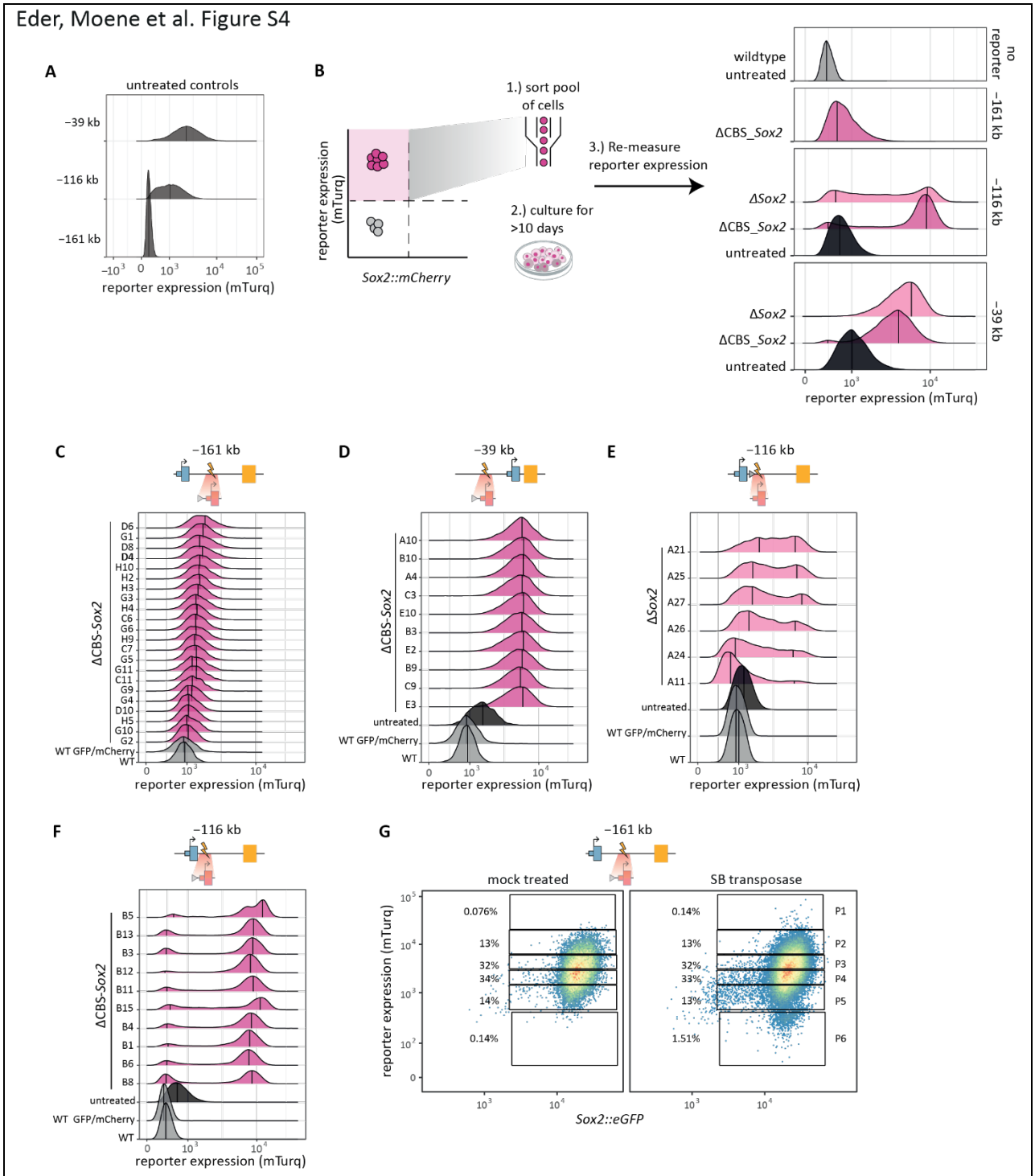


***Fig. S4*. Sox2P-reporter expression in Sox2::mCherry deleted pools and clones. (A)** Reporter expression in untreated control cell lines (-39 kb, -116 kb, -161 kb). **(B)** Distribution of reporter expression in *Sox2::mCherry*-negative, reporter-high sorted pools. **(C)** Distribution of reporter expression in CBS_*Sox2::mCherry* deleted, reporter-high sorted clones from the -161 kb reporter cell line. Clone D4 (in bold) was used for hopping (see Fig. 4D). **(D)** As in C, but for the -39 kb reporter cell line. Untreated reporter cell line is indicated in black. **(E)** As D, but for *Sox2::mCherry* deletion in the -116 kb reporter cell line. **(F)** As D, but for CBS_*Sox2::mCherry* deletion in the -116 kb reporter cell line. **(C-F)** Vertical lines indicate the (at most) two highest peaks in the density distribution, the peak with the highest

mTurq level is used in Fig. 4B. **(G)** Expression of *Sox2::eGFP* (control) and the *Sox2* reporter (mTurq) in the untreated cells and cells transfected with the SB transposase plasmid, for the -161 kb ΔCBS_*Sox2::mCherry* reporter cell line. The percentage of cells in each sorted gate is indicated. Cells are gated for *Sox2::mCherry*-negative single cells. Representative image of 2 biological replicates.

Eder, Moene et al. Fig S5

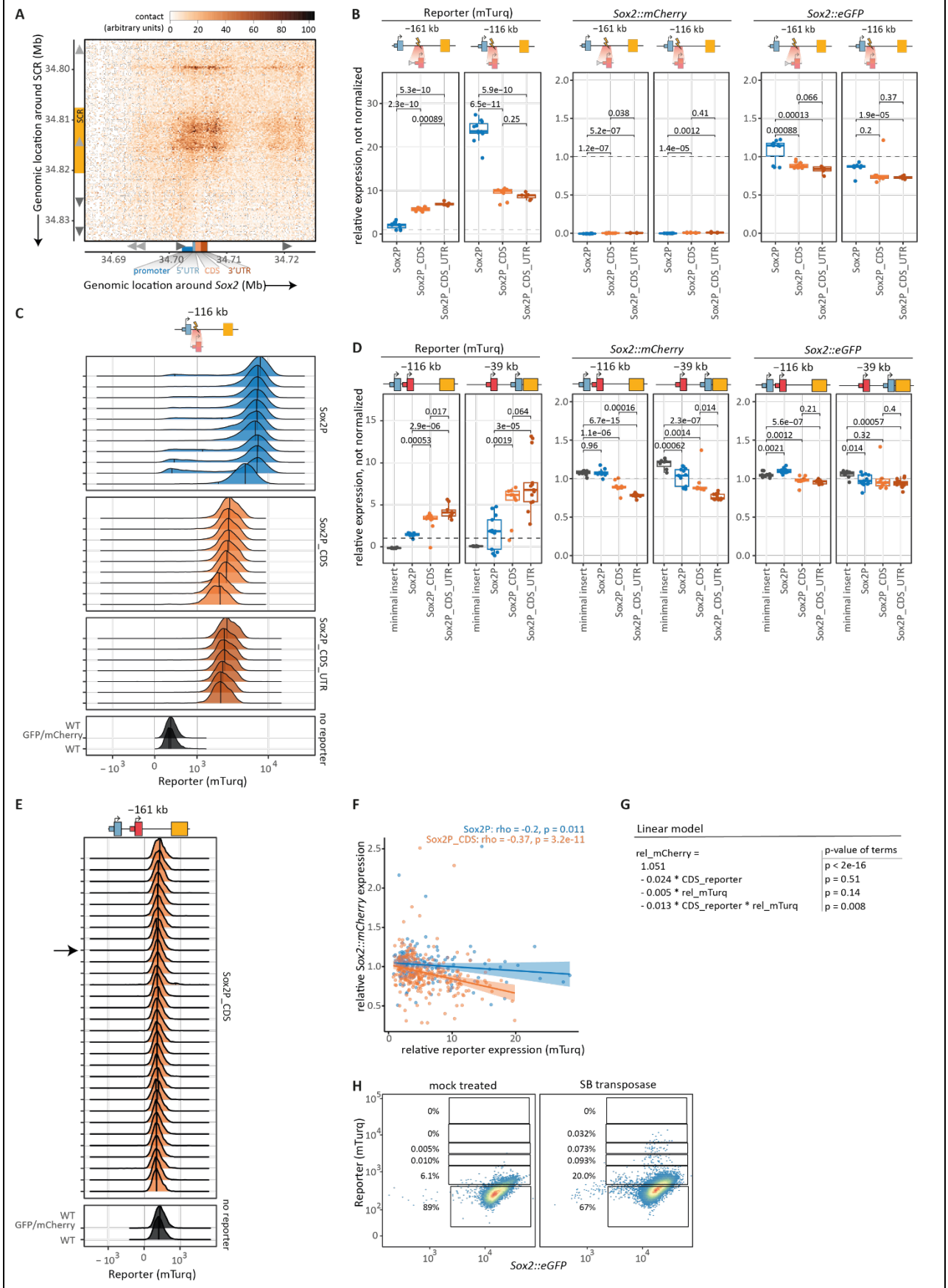


Fig. S5. **Impact of the Sox2 coding sequence on reporter and *Sox2* expression. (A)** Contact between *Sox2* and the SCR (region-capture micro-C data [(*18*)], 200-bp resolution). Positions of the components of the *Sox2* gene (blue, orange, red; x-axis) and SCR (yellow; y-axis) are indicated. Grey triangles indicate CTCF binding sites and their orientation. Arrows indicates the genomic direction. **(B)** Relative expression of the reporter (mTurq), *Sox2::mCherry* and *Sox2::eGFP*, not normalized to *Sox2::eGFP*, for the three reporters inserted at -161 kb and -116 kb. Each dot represents one clone. Expression values are autofluorescence subtracted, and then normalized to the control reporter cell line (*Sox2* reporter at -116 kb) measured on the same day. Dashed line indicates 1 (level of the control cell line). P-values according to Welch's unequal variances *t*-test. **(C)** Distribution of reporter expression of reporters integrated at -116 kb in a ΔCBS_*Sox2::mCherry* cell line. Each density plot is one clone, vertical lines indicate the (at most) two highest peaks in the density distribution. **(D)** As **B**, for the three reporters and a minimal insert (not containing any promoter or mTurq) inserted at -116 kb and -39 kb launch pads with *Sox2::mCherry* intact. **(E)** Reporter expression in a panel of clones with the Sox2P_CDS reporter integrated at the -161 kb launch pad with *Sox2::mCherry* intact. The arrow indicates the clone used for hopping (see **Fig. 5F**). **(F)** Correlation between relative *Sox2::mCherry* and reporter (mTurq) expression from the Sox2P and Sox2P_CDS reporter, from cells with reporter expression in the gates P2-P4 after induction of hopping. As in **Fig. 5E**, but also showing the outlier points. **(G)** Linear model of the data in **F**, describing how the relative *Sox2::mCherry* expression (rel_mCherry) depends on the identity of the reporter (CDS_reporter, is 0 for the Sox2P reporter and 1 for the Sox2P_CDS reporter), and the relative reporter expression (rel_mTurq). Right side indicates the p-values of the coefficients of the model, testing whether they are significantly different from 0. **(H)** Expression of *Sox2::eGFP* (control) and the *Sox2* reporter (mTurq) in the untreated cells and cells transfected with the SB transposase plasmid, for the -161 kb Sox2P_CDS reporter cell line. The percentage of cells in each sorted gate is indicated. Cells are gated for *Sox2::mCherry*-positive single cells. Representative image of 2 biological replicates.