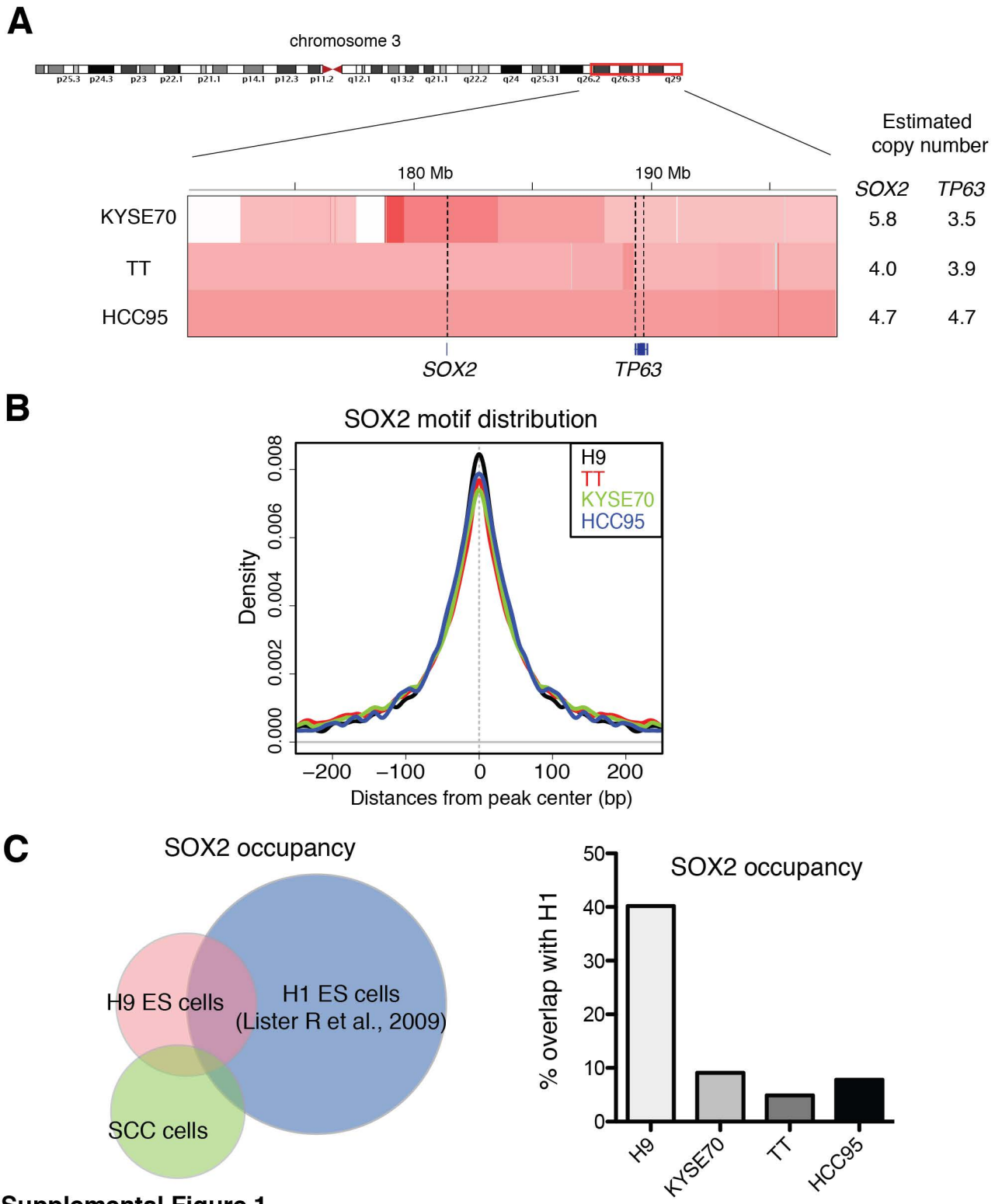
Supplementary Information for

SOX2-p63 interaction and genomic co-localization in squamous cell carcinomas

Hideo Watanabe, Qiuping Ma, Shouyong Peng, Guillaume Adelmant, Danielle Swain, Wenyu Song, Cameron Fox, Joshua M. Francis, Chandra Sekhar Pedamallu, David S. DeLuca, Angela N. Brooks, Su Wang, Jianwen Que, Anil K. Rustgi, Kwok-kin Wong, Keith L. Ligon, X. Shirley Liu, Jarrod A. Marto, Matthew Meyerson, Adam J. Bass

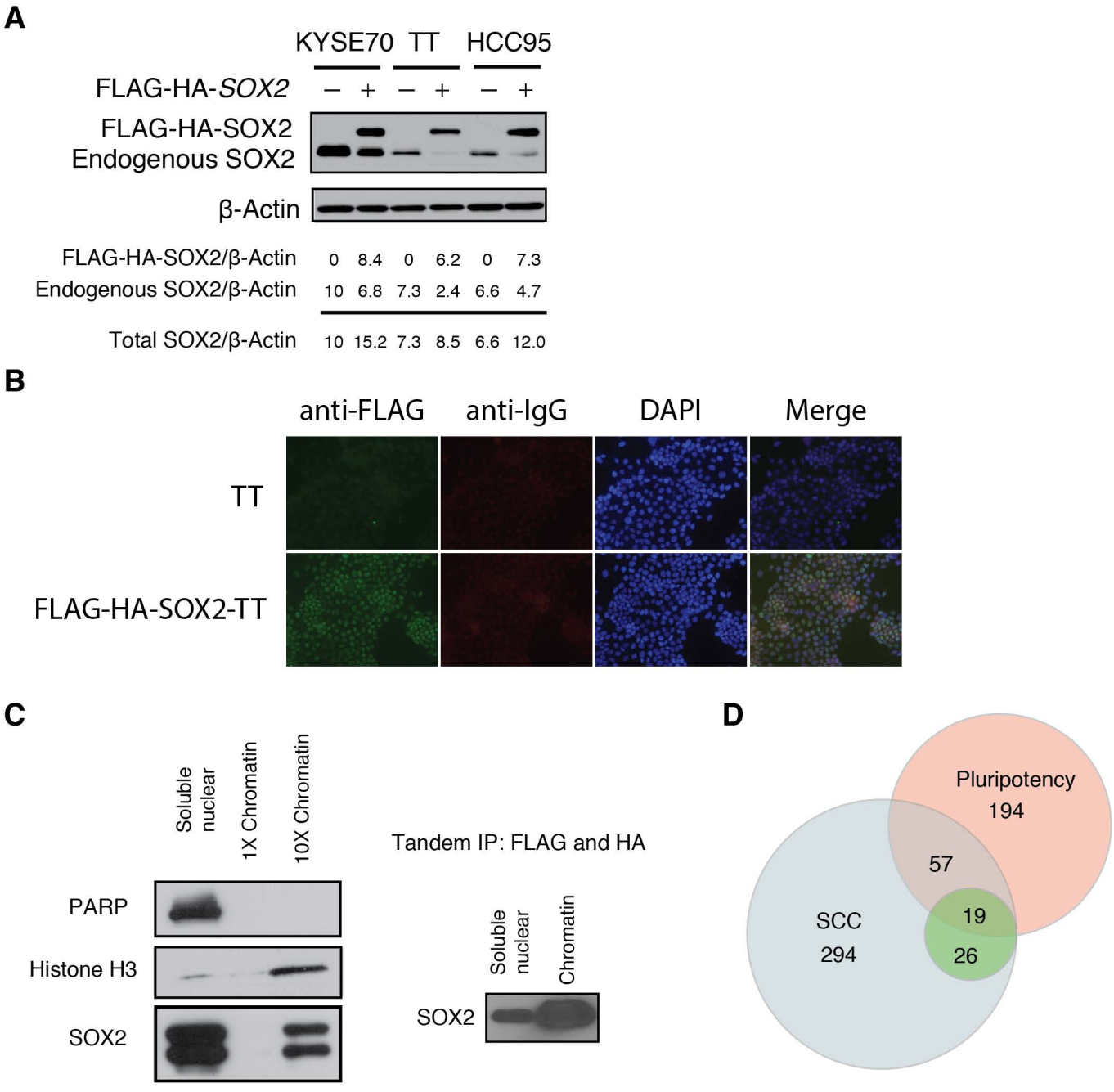
*To whom correspondence should be addressed: Adam_Bass@dfci.harvard.edu

Supplementary Data

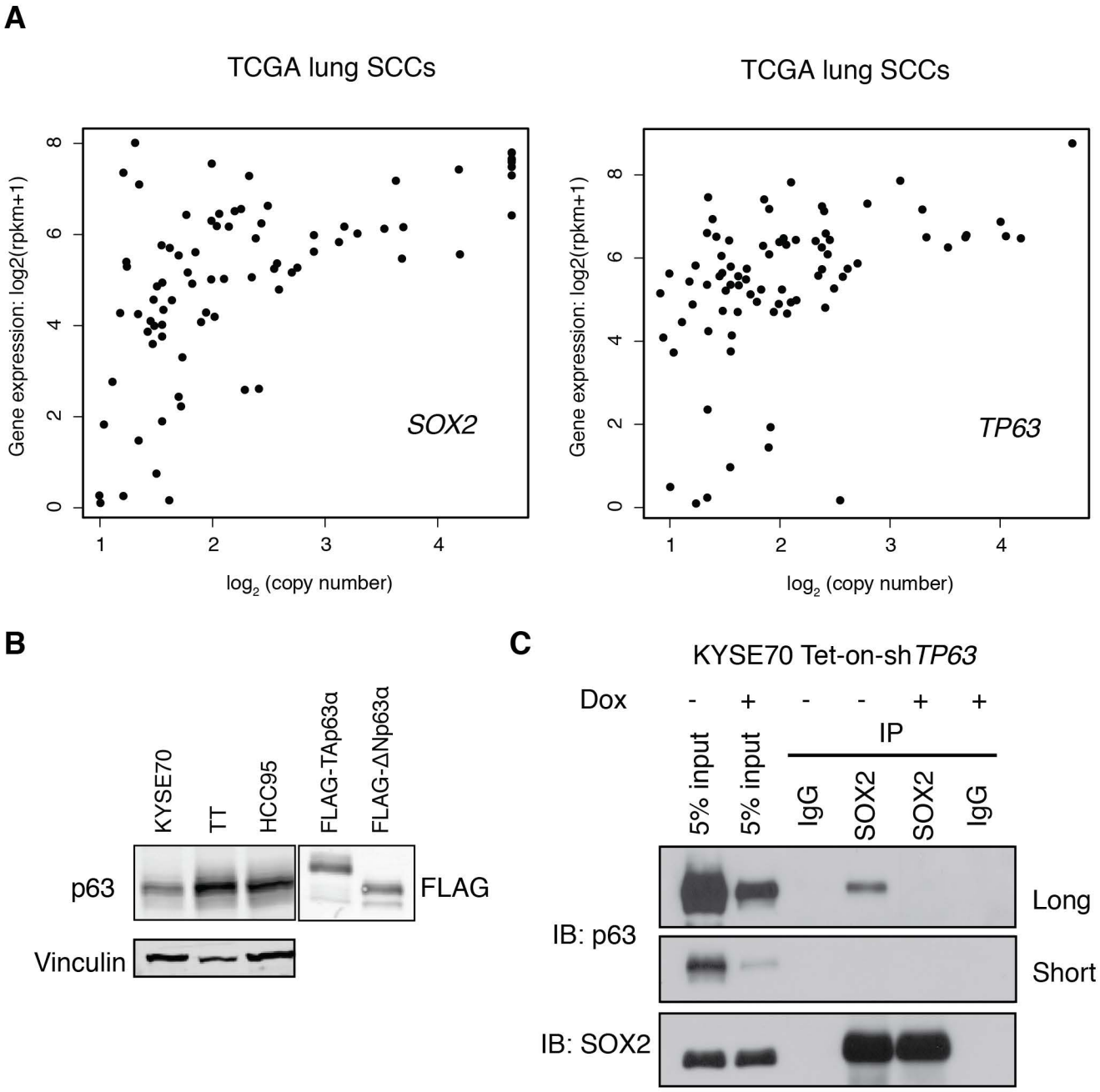


Supplemental Figure 1.

SOX2 binding sites in 3q-amplified SCC cells and in hES cells are enriched for SOX2 motifs and distinct from each other. (**A**) Estimated copy numbers at chromosome 3q segment containing the *SOX2* and *TP63* loci from SNP array 6.0 data for three SCC lines used in this study. (**B**) Appearance of SOX2 DNA binding motif is plotted around the summit of SOX2 binding peaks in the three SCC cell lines and the H9 ES cell line. The motif is highly enriched within 100bp from the SOX2 peak centers in all samples. (**C**) Overlap of SOX2 occupancy between SCC cells (composite data), H1 ES cell (analyzed with published data from Lister et al.) and H9 ES cell is depicted in Venn diagram (left). Overlaps of SOX2 occupancy in individual SCC cells and H9 cell with H1 ES cell (right).

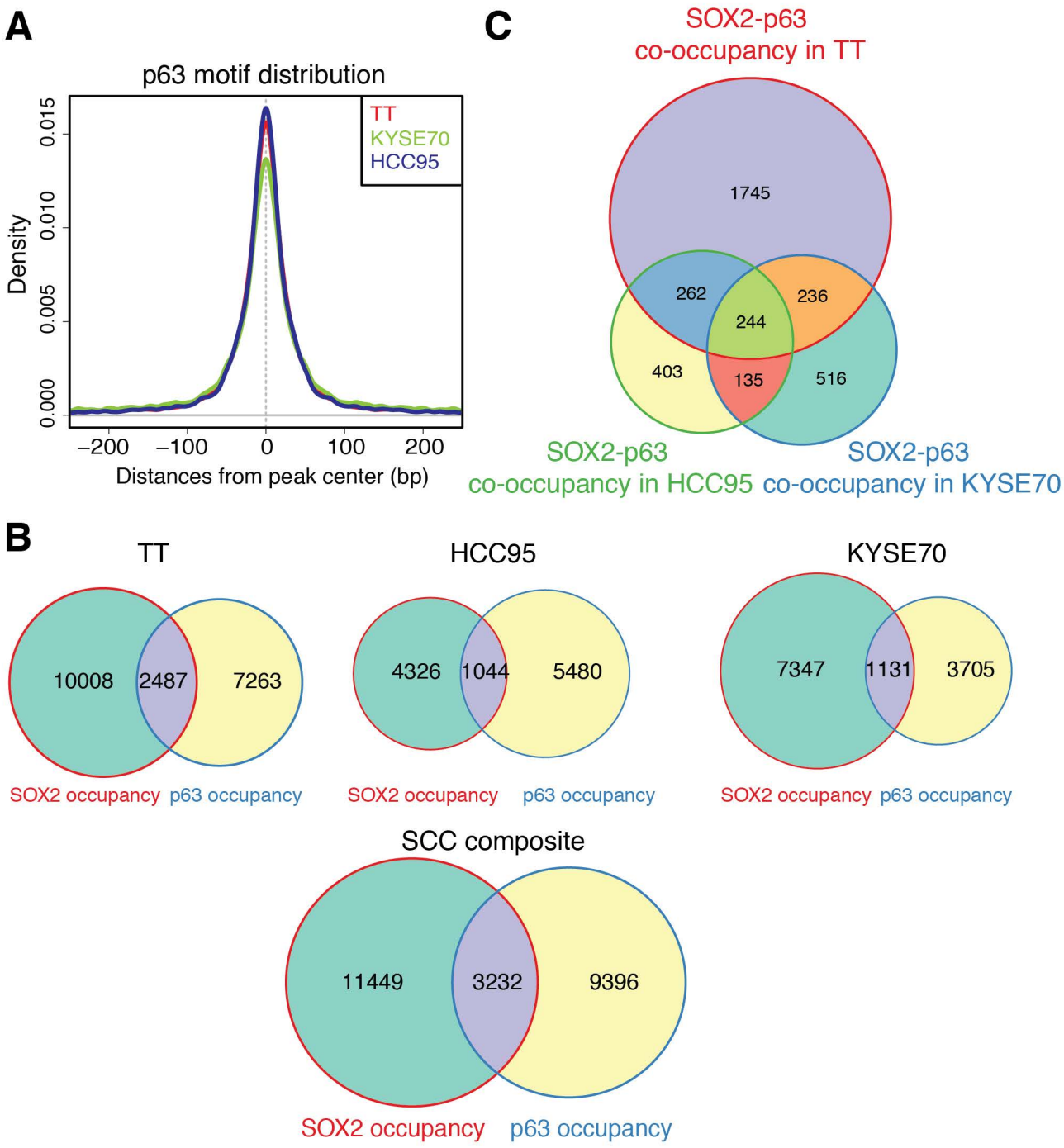


Supplemental Figure 2.
(**A**) Expressions of endogenous SOX2 and ectopic FLAG-HA-tagged SOX2 in the whole cell lysate of parental and stable three SCC lines are determined by immunoblotting with anti-SOX2 antibody and quantitation for each signal is shown below the immunoblots. (**B**) IgG control staining of endogenous SOX2 in parental and stable TT cells. Expression of ectopic FLAG-HA-tagged SOX2 and endogenous SOX2 are determined by immunofluorescence with anti-FLAG and anti-IgG antibodies, nuclei (DAPI) and merge figures are also shown. (**C**) (*left*) Separation of soluble and insoluble nuclear extracts were confirmed by immunoblotting with antibodies against known soluble nuclear (PARP), insoluble chromatin Histone H3 markers as well as SOX2 in FLAG-HA-SOX2-TT cells. Proteins from soluble nuclear and chromatin extracts were loaded to represent the same amount of cells or 10 times for chromatin. (*right*) Representative relative expression of SOX2 in nuclear fraction and chromatin fraction in TT cells following IP determined by immunoblotting with anti-SOX2 antibody. (**D**) Venn diagram depicting the overlaps of proteins identified in any of the three SCC cell lines (blue), the subset of 45 common SOX2 binding partners found across all three cell lines (green) and proteins previously identified as pluripotent interactome (pink).

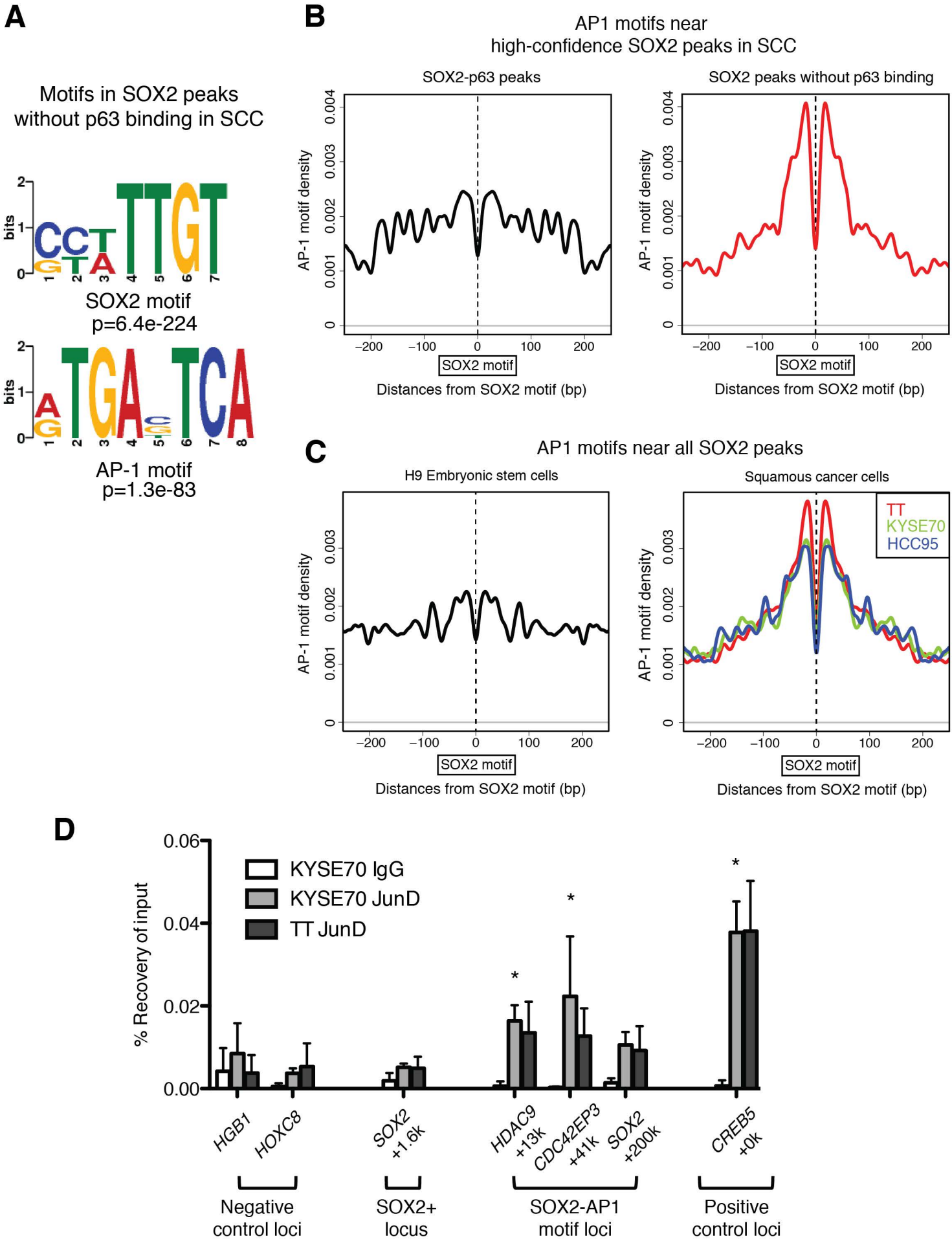


Supplemental Figure 3.

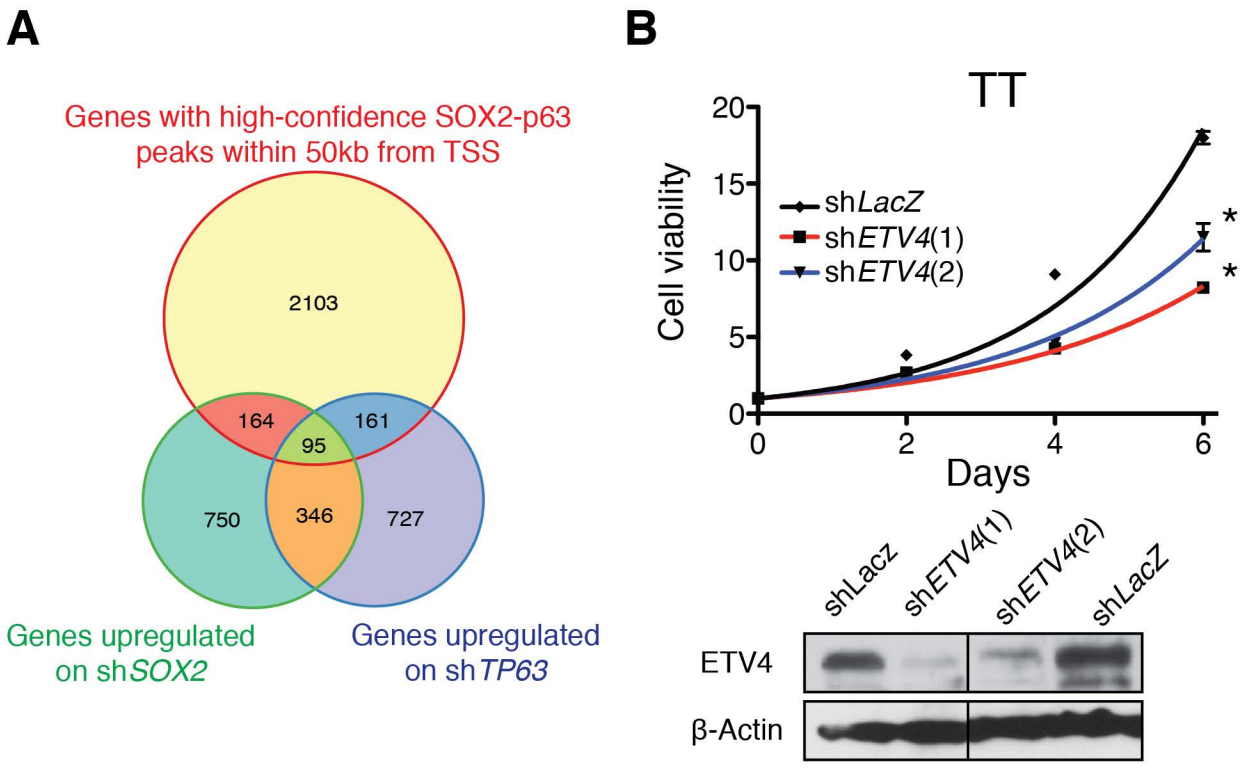
(**A**) Scatter plot of copy number (X-axis) and gene expression (Y-axis) of the *SOX2* gene (*left*) and the *TP63* gene (*right*) in lung SCCs from TCGA project. (**B**) Isoforms of p63 expressed in the three SCC lines were determined by immunoblotting with an anti-p63 antibody that detect all isoforms of p63 and compared to the migration of ectopically expressed two isoforms (TAp63α and ΔNp63α) with FLAG-tag in 293T cells immunoblotted with anti-FLAG antibody. All samples were run on the same gel, transferred to a membrane and probed with each indicated antibody. Images were then merged to retain the same migrated position. (**C**) Antibody against endogenous SOX2 co-immunopreciptates p63 in KYSE70 cells with inducible sh*TP63* and the corresponding p63 band was reduced upon suppression of *TP63*. Longer and shorter exposure for p63 immunoblots are shown.



Supplemental Figure 4.
SOX2 and p63 share common genomic binding loci in SCC cell lines. (**A**) Appearance of p63 DNA binding motif is plotted around the summit of p63 binding peaks in the three SCC lines. The p63 motif is highly enriched within 50bp from the p63 peaks. (**B**) Venn diagrams depict overlaps between SOX2 and p63 binding sites by ChIP-seq in each of the three SCC lines (top) as well as in a composite analysis using ChIP-seq data from all three SCC lines (bottom). All peaks were determined with the p-value threshold < $1x10^{-5}$. The overlaps of the peaks (each approximately 500-800 bp in size) for SOX2 and p63 in the three cell lines were 13.3%-25.5%. (**C**) Venn diagram depicts overlaps of the SOX2-p63 overlapping binding sites in each SCC line.


A
Motifs in SOX2 peaks
without p63 binding in SCC
bits
SOX2 motif
p=6.4e-224
AP-1 motif
p=1.3e-83
B
AP1 motifs near
high-confidence SOX2 peaks in SCC
SOX2-p63 peaks
SOX2 peaks without p63 binding
AP-1 motif density
SOX2 motif
Distances from SOX2 motif (bp)
C
AP1 motifs near all SOX2 peaks
H9 Embryonic stem cells
Squamous cancer cells
TT
KYSE70
HCC95
D
% Recovery of input
KYSE70 IgG
KYSE70 JunD
TT JunD
HGB1
HOXC8
SOX2 +1.6k
HDAC9 +13k
CDC42EP3 +41k
SOX2 +200k
CREB5 +0k
Negative control loci
SOX2+ locus
SOX2-AP1 motif loci
Positive control loci

Supplemental Figure 5.
SOX2 binding peaks without p63 occupancy in SCC are enriched for AP-1 binding motif. (**A**) Motif analysis of recurrent SOX2 peaks without p63 binding in *SOX2*-amplified SCC lines reveals significant enrichment of SOX2 motif ($p=6.4e^{-224}$) as well as AP-1 transcriptional response element ($p=1.3e^{-83}$). (**B**) Appearance of AP-1 DNA binding motifs is plotted around the SOX2 motif near the summit of high-confidence SOX2-p63 co-occupied peaks in the composite SCC ChIP data (left) and recurrent SOX2 peaks without p63 occupancy (right) showing enrichment of the AP-1 motif within 50 bp from the SOX2 motif only in SOX2 peaks without p63 co-occupancy. (**C**) Appearance of AP-1 DNA binding motif is plotted around the SOX2 motif near the summit of all SOX2 binding peaks in the H9 ES cell line (left) and the three SCC lines (right) showing enrichment of the AP-1 motif within 50 bp from the SOX2 motif in SOX2 peaks in all three SCC lines. (**D**) ChIP enrichment of JunD in KYSE70 and TT cells at three candidate SOX2-occupied sites with adjacent AP-1 motif (SOX2-AP1 motif loci) as well as negative control regions (*HGB1* promoter, *HOXC8* promoter, and a SOX2-occupied SOX2 enhancer without adjacent AP-1 motif (SOX2+)) and a positive control locus for AP-1 binding regions (*CREB5* promoter) by qPCR. Percent recovery of input for ChIP was calculated based on 10% non-IP DNA sample for each experiment. Mean % recovery ± s.d. of triplicates are shown. * denotes $P < 0.05$ in two-way ANOVA-test with Bonferroni's post-test correction.



Supplemental Figure 6.
SOX2 and p63 cooperatively regulate gene expression and *ETV4* is essential for *SOX2*-amplified SCC cells. (**A**) Venn diagram depicts the overlap between the 2,523 genes with high-confidence SOX2-p63 co-occupied ChIP peaks within 50kb from their TSS, genes whose expression is upregulated (>1.5 fold) upon *SOX2* suppression, and genes whose expression is upregulated (>1.5 fold) upon *TP63* suppression in KYSE70 cells RNA-seq data. See Supplemental Table S6 for 95 genes at the intersection. (**B**) Cell growth curve of TT cells after seeding following infection of two independent shRNAs against *ETV4* or one shRNA against *LacZ* as control. The plots represent mean cell viabilities ± s.d. of cell plated in six wells and* denotes $P < 0.0001$ for sum-of-squares F-test of exponential growth model compared to sh*LacZ* control. Immunoblots for ETV4 protein and β-actin in the lysates from corresponding TT cells with shRNAs are shown.

Supplemental Table S2.
Gene expression correlation with SOX2 of SOX2-associated proteins

gene	correlation	p.value
TP63	0.487	1.02E-09
TFAP4	0.36	1.25E-05
ZNF384	0.317	1.35E-04
CUX1	0.299	3.35E-04
TMPO	0.15	7.77E-02
RUNX1	0.137	1.07E-01
CDC5L	0.112	1.89E-01
MECP2	0.046	5.92E-01
PSIP1	0.029	7.35E-01
CBX3	0.028	7.42E-01
TFAM	-0.031	7.13E-01
PBX1	-0.072	3.95E-01
NFIB	-0.095	2.62E-01
CEBPB	-0.348	2.51E-05

Supplemental Table S3. SOX2-p63 target genes

		Co-oocupied by SOX2 and p63		Fisher's exact test
		Yes	No	
Down-regulated by shSOX2	Yes	201	1066	
	No	2322	17933	p=6.35e-6
Down-regulated by shTP63	Yes	218	1137	
	No	2305	17862	p=7.66e-7
Down-regulated by both shSOX2 and shTP63	Yes	93	2430	
	No	454	18545	p=2.04e-4

Supplemental Table S5:
93 SOX2-p63 target genes downregulated by sh*SOX2* and sh*TP63*

Gene	Fold change sh*SOX2*	Fold change sh*TP63*	Correlation (r) with *SOX2*
ADAMTS15	0.75	0.78	0.05
ADRM1	4.34	4.43	0.088
BCAS2	2.14	2.08	0.056
C16orf74	1.94	2.20	-0.003
CCDC85B	3.01	2.58	-0.083
CCT6A	4.57	4.81	0.244
CD109	2.61	2.23	0.306
CDCA5	3.73	2.72	0.309
CEP55	2.24	1.82	0.238
CSE1L	3.22	3.10	0.249
DDX56	3.81	3.40	0.275
DEK	3.50	3.23	0.166
DHX9	3.75	3.48	0.409
DOHH	2.41	2.80	0.156
DUSP7	1.54	1.74	0.038
EFNB2	3.00	2.78	0.154
EIF4H	5.15	5.11	0.431
EIF5A	7.26	6.71	0.125
EIF6	4.68	4.69	-0.019
ETV4	4.06	2.35	0.396
FAM104A	2.17	2.63	0.223
FOSL1	2.70	1.47	0.036
FRMD8	3.06	3.13	0.005
GLRX3	2.76	3.07	0.135
GPS2	3.71	3.52	0.131
HAUS1	1.50	1.39	0.166
HCFC1	3.86	3.97	0.273
HIST1H1D	2.59	2.27	-0.02
HIST1H1E	4.52	4.72	0.01
HIST1H2AE	3.59	3.89	0.029
HIST1H2BE	3.61	3.65	0.046
HIST1H2BF	2.11	2.41	-0.056
HIST1H2BG	3.24	3.71	-0.036
HIST1H3D	2.72	3.29	-0.096
HIST1H4D	2.02	2.37	0.017
HIST1H4E	3.74	3.89	0.088
HNRNPU	4.82	4.96	0.225
HNRPDL	3.90	3.83	0.074

HRAS	3.66	3.21	0.277
HSPD1	5.33	5.08	0.391
HSPE1	5.60	5.63	0.298
HSPH1	4.11	3.59	0.231
IMP4	4.25	4.25	0.104
INPP5F	0.76	0.65	-0.067
JUN	2.91	3.09	-0.103
KLF6	1.77	1.88	-0.344
LMNA	5.40	5.77	0.038
MCM7	5.33	4.56	0.385
MORN2	0.67	1.25	-0.199
MPV17L2	2.38	2.49	0.181
MRPL1	1.14	1.35	-0.092
MYOF	4.28	4.87	-0.216
NCAPG	1.79	1.31	0.253
NOP16	3.06	2.57	0.105
NUP188	3.29	3.35	0.259
ORC1	1.87	1.15	0.302
PADI1	0.76	0.23	0.018
PCNA	4.72	4.80	0.121
PDCL3	1.78	1.99	0.35
PFKFB3	2.80	2.72	-0.16
PRPF38A	1.49	1.96	-0.103
PSMC3IP	2.50	1.93	0.387
PYGB	3.93	4.51	-0.128
RAB15	2.50	2.30	0.125
RBM4	0.65	0.51	0.039
RRM1	2.67	2.49	0.469
RRP12	2.97	2.60	0.068
RUVBL2	4.25	4.17	0.024
S100A2	6.02	5.69	-0.108
SEH1L	1.70	2.01	0.22
SEPT11	1.43	1.48	NA
SERPINE1	2.04	1.51	-0.172
SNRPE	2.83	2.98	0.113
SRSF2	4.69	4.68	0.233
SYNCRIP	2.73	2.85	0.157
TAGLN2	5.02	5.39	-0.067
TBL1XR1	2.01	2.11	0.503
TCHP	2.13	1.83	0.178
TGFA	2.24	2.42	-0.109
TK1	4.37	3.77	0.348
TNS4	2.71	2.78	0.064

TOX2	1.59	1.52	-0.158
TRMT61A	2.51	2.21	0.024
TUBA4A	2.01	2.63	-0.123
TUBB	6.80	6.62	0.071
TUBG1	4.17	4.09	0.374
TXNRD1	4.73	5.51	0.443
UBE2S	6.30	5.70	0.134
UBXN8	1.53	1.02	0.032
UTP15	1.17	1.10	0.325
VGF	2.16	0.47	-0.084
VRK1	1.28	1.17	0.149
ZBED2	1.30	0.63	-0.073

Supplemental Table S6:
95 SOX2-p63 target genes upregulated by sh*SOX2* and sh*TP63*

Gene	Log_2 FC on sh*SOX2*	Log_2 FC on sh*TP63*	Correlation (r) with *SOX2*
ABCC5	4.19	4.10	0.619
ACBD4	2.21	2.49	-0.179
ADAM23	3.81	3.24	0.581
AGRN	6.54	5.75	-0.228
ANKRD9	4.18	3.55	0.052
ARHGAP23	2.59	2.90	0.026
C1S	3.70	3.57	-0.128
CD82	4.56	4.30	-0.321
CHD6	3.30	2.76	0.293
CLDN1	2.87	2.76	-0.063
CNFN	1.45	2.47	0.125
COL9A2	1.58	1.77	-0.055
CRIM1	4.47	3.90	-0.139
CTDSP2	4.96	5.55	-0.078
CYP4F12	1.66	2.05	-0.101
CYR61	4.90	4.54	-0.251
DAPL1	1.06	1.50	0.212
DDIT4	5.39	4.85	-0.184
DDR1	5.61	5.18	0.387
EGR1	2.75	4.39	0.037
EMID2	0.99	1.13	-0.053
EXT1	3.68	2.98	0.253
FGFR2	3.02	3.07	0.61
FJX1	4.15	3.92	-0.048
GAL3ST4	2.71	2.22	0.099
GAS6	3.79	3.47	-0.314
GGT1	1.82	2.29	-0.14
GOLIM4	3.39	3.29	0.506
HADHB	3.29	3.41	0.022
HCK	3.17	3.62	0.114
HERPUD1	2.97	2.94	-0.115
HES4	4.05	3.32	-0.091
HRCT1	2.81	3.17	-0.106
HS6ST1	5.09	4.57	0.208
ICAM4	4.83	5.35	-0.171
ICAM5	5.40	4.75	-0.138
IFITM1	4.44	4.42	-0.24
IFITM3	6.14	7.07	-0.334

IGFBP5	2.33	4.55	0.172
IL20RB	4.79	3.32	-0.124
KCNJ2	1.99	2.16	0.205
KREMEN1	3.23	3.04	0.177
LAMP1	6.56	6.23	-0.026
LBH	2.26	2.35	0.027
LEPREL4	3.81	3.39	0.237
LRP1	3.00	3.02	-0.218
LRP4	3.34	3.17	0.399
MEGF8	2.67	2.55	-0.038
MMP11	3.56	2.38	0.094
MMP15	3.32	2.97	-0.139
NEAT1	4.74	4.78	0.028
NISCH	3.89	3.47	-0.196
NRBP1	4.41	4.42	0.273
NUPR1	4.46	5.52	-0.218
PBXIP1	4.04	4.10	-0.191
PCYT1B	2.24	2.15	0.568
PGD	7.10	7.27	0.381
PIK3IP1	1.75	2.06	-0.025
PLEKHA4	2.33	3.31	-0.234
PLLP	3.39	3.17	-0.231
PLXDC2	2.20	1.33	-0.158
PNRC1	2.50	3.92	-0.201
PTCH1	2.63	2.07	0.226
PTGES	1.31	0.87	-0.248
PTMS	7.23	7.86	0.236
RBMS2	1.90	2.22	-0.279
REEP2	2.86	3.66	0.031
SCD	7.52	6.38	0.189
SCPEP1	4.46	4.13	0.194
SEL1L3	3.54	3.39	0.187
SEMA4B	3.71	2.98	-0.08
SERPINB13	2.02	1.30	0.122
SGPP1	2.31	2.31	-0.256
SIL1	3.46	2.88	-0.083
SLC25A28	4.82	4.49	0.097
SLC25A29	2.24	2.49	-0.021
SMO	4.80	4.29	0.662
SNCG	4.18	5.81	-0.255
SOST	5.13	3.64	0.266
SPARCL1	2.02	1.95	0.16
SRCRB4D	4.03	4.51	0.061

ST3GAL3	2.57	2.60	0.571
ST6GALNAC	2.64	2.65	0.499
STOM	3.48	3.03	-0.092
STON2	3.73	3.67	0.445
TACC2	1.71	2.05	0.324
TIMP3	4.42	4.25	-0.347
TMEM187	2.48	1.51	0.104
TMEM59L	1.93	1.17	0.006
TMEM92	1.22	1.01	-0.33
TNKS1BP1	4.99	5.33	-0.046
TSPAN33	3.48	3.52	0.608
VAMP1	2.62	2.80	0.049
VMAC	1.87	1.54	0.283
ZNF362	2.79	3.17	-0.096

Supplemental Table S7. shRNA sequences

Name	Target sequence	TRC clone ID or Reference
sh*LacZ*	CGCTAAATACTGGCAGGCGTT	TRCN0000072231
sh*ETV4*(1)	CCAGGATCTAAGTCACTTCCA	TRCN0000013937
sh*ETV4*(2)	CCCAACAAATGCCCATTTCAT	TRCN0000013936
sh*SOX2*(1)	CGCTCATGAAGAAGGATAAGT	TRCN0000355694
sh*SOX2*(2)	CTGCCGAGAATCCATGTATAT	TRCN0000003253
sh*SOX2*(3)	AGGAGCACCCGGATTATAAAT	TRCN0000231640
sh*ΔNP63*	GGACAGCAGCATTGATCAA	Sabbisetti, V; PLoS One;2009
sh*TP63*(1)	CCGTTTCGTCAGAACACACAT	TRCN0000006506
sh*TP63*(2)	CATCTGACCTGGCATCTAATT	TRCN0000420642

Supplemental Table S8. Primers for PCR

Primers for quantitative RT-PCR	
18S Forward	GTAACCCGTTGAACCCCATT
18S Reverse	CCATCCAATCGGTAGTAGCG
ETV4 Forward	TGGAAATCAGGAACAAACTGC
ETV4 R	GCCCCTCGACTCTGAAGAT
Primers for ChIP-qPCR	
HGB1 F -435	ACACTAATCTATTACTGCGCTG
HGB1 R -637	CCAGGATTTTTGACGGGAC
HOXC8 F -506	TGCTGGCTCAGGTATCTCTCT
HOXC8 R -391	GCTAATGGCTGGGGCCAGTT
SOX2 F -1587	GGATAACATTGTACTGGGAAGGGACA
SOX2 R -1787	CAAAGTTTCTTTTATTCGTATGTGTGAGCA
CDC42EP3 F "-4k"	TCGTTCTGTTCTTGCTGGGT
CDC42EP3 R "-4k"	AAGCCTTCTACTTTGCCTCAA
SOX2 F "+200k"	TTCCTGTTTGGCAGGACCTT
SOX2 R "+200k"	AGCAATAATCGGCCAGCCTT
HDAC9 F "+13k"	ATGCCTCATGGGGTTTTCCC
HDAC9 R "+13k"	GATAAAGGCTTCCCCACCCC
CREB5 F -47	ATCGCTTGGCCTGGAATCAA
CREB5 R +70	ATTCCCTCCTGAGGTGCAGA
ETV4 P1 F	CGCTAGTAGATTGGGGGCTG
ETV4 P1 R	AAGCCACAGCCTGCCTTTAT
ETV4 P2 F	ACACAGCTTCGTGGACACAT
ETV4 P2 R	AGAGCTCAGCCCCCAATCTA
CD55 F -17	CCCAGAGGGTGTTGCCAGCG
CD55 R +172	GTGCTTTGGGCGGGGTCTGG
CCND1 F +76	GAGCAGCAGAGTCCGCACGCTC
CCND1 R +216	TGTTCCATGGCTGGGGCTCTTC
HDAC9 F -604	TGTGCCTGTGGGCCATGCTG
HDAC9 R -471	TCCAGCCGGTGCCCTTCTCC
JUN F +100	TCGCTGCTTCAGCCACACTCA
JUN R -37	CTGAGAGCGACGCGAGCCAATG

Supplemental Table S9. Motifs used for scanning across the genome

Transcription Factor	Motif PSSM from database	P value	Number of motif occurrence in genome	Reference
SOX2	MA0143.1 (1st half)	5x10^-4	4,780,848	Jasper Core (Bryne JC et. al. 2008)
TP63	M01656	3x10^-4	756,286	Transfac database (Wingender et.al 2000)
		(5x10^-4)	1,378,815	
OCT4	MA0143.1 (2nd half)	1x10^-4	1,747,599	Jasper Core (Bryne JC et. al. 2008)
		(1x10^-5)	169,779	
AP1	MA0099.2	3x10^-4	461,659	Jasper Core (Bryne JC et. al. 2008)
		(5x10^-4)	3,424,688	

Supplemental Methods

Antibodies, and plasmids

Antibodies used are anti-SOX2 (Cell Signaling Technology, D6D9 or R&D, AF2018), anti-p63 (Santa Cruz, sc-8431 or sc-8344), anti-OCT4 (Santa Cruz, sc-5279), anti-ETV4 (Santa Cruz, sc-113), anti-JunD (Santa Cruz, sc-74), anti-vinculin (Sigma), anti-β-actin (Sigma), anti-FLAG (Sigma), anti-PARP (Cell Signaling Technology), anti-Histone H3 (Cell Signaling Technology) and goat IgG antibody (R&D). *SOX2* or *GFP* open reading frame (ORF) was cloned into pMSCV-FLAG-HA gateway vector. pcDNA3-TAp63α-FLAG and pcDNA3-ΔNp63α-FLAG were obtained through Addgene and a site encoding Glycine at amino acid position 59 in pcDNA3-ΔNp63α-FLAG was corrected to Serine. GST-ΔNp63α was generated using pcDNA3-ΔNp63α-FLAG and pGEX4T3 (GE Healthcare). pcDNA3-SOX2 and GST-SOX2 were generated by cloning *SOX2* ORF into pcDNA3.1/Hygro (Invitrogen) or pGEX4T3 vector.

Tandem affinity purification

Stable SCC cells expressing pMSCV-FLAG-HA-SOX2 or pMSCV-FLAG-HA-GFP were established by infecting with retrovirus using standard viral transfer protocol followed by puromycin selection. Nuclear pellets were prepared from these stables cells by the methods described previously (1) and tandem affinity purification were performed as recently described (2) with a little modifications. Briefly, the nuclear pellets were thoroughly resuspended in 5 volumes of FLAG-IP buffer (150 mM NaCl, 50 mM Tris (pH 7.5), 1 mM EDTA, 0.5% NP40, 10% Glycerol, Protease inhibitors (Roche Applied Sciences)) and continued the lysis for 30 min. Soluble nuclear fraction and insoluble chromatin pellet were separated after centrifuging at 14,000 rpm for 15 min. The insoluble pellet was washed in 5 pellet volumes of MNase buffer (20 mM Tris (pH 7.5), 100 mM KCl, 2 mM $MgCl_2$, 1 mM $CaCl_2$, Protease inhibitors) and re-pelleted for 5 min at 10,000 rpm. The pellet then was thoroughly resuspended in one pellet-volume of MNase buffer containing 3 units of Micrococcal nuclease (Nuclease S7, Roche Applied Sciences) per µl and incubated with constant vortexing until complete digestion and

solubilization of the chromatin. Solubilized chromatin was then diluted and vortexed with an equal volume of 2x HA-IP buffer (300 mM NaCl, 100 mM Tris (pH 7.5), 2 mM EDTA, 0.1% NP40, 20% Glycerol, Protease inhibitors) for 10 min followed by centrifugation for 10 min at 14,000 rpm. The supernatant containing the solubilized chromatin extract was then incubated with anti-FLAG-agarose (Sigma) for 2 h, followed by elution with FLAG peptide (5 mg/ml) (Sigma) in HA-IP buffer after 3 washes with FLAG-IP buffer. The eluate was filtered through Ultrafree-MC Centrifuge filters (Millipore) and further incubated with anti-HA-agarose conjugate (Santa Cruz) for 3 h, followed by elution with HA peptide (4 mg/ml) in HA-IP buffer after 3 washes with HA-IP buffer. The final HA eluate from CE fractions were further processed for LC–MS/MS sequencing and data analysis.

Protein digestion, LC-MS/MS data acquisition, and database searching

Protein digestion: Proteins from the TAP samples were directly processed in solution: Cysteine residues were reduced with 10 mM DTT for 30 min at 56°C and alkylated with 22.5 mM iodoacetamide for 20 min at room temperature in the dark. Proteins were digested overnight at 37ºC using 5 µg of trypsin after adjusting the pH to 8.0 with 1M Tris. The resulting tryptic peptide solutions were acidified by adding TFA to a final concentration of 1% and desalted by C18 solid phase extraction followed by Strong Cation Exchange (SCX), both performed in batch-mode format. Eluted peptides were concentrated in a vacuum concentrator, and reconstituted with 20 µl of 0.1% TFA.

LC-MS/MS analysis: Purified tryptic peptides were analyzed by LC-MS/MS (3) on an LTQ-Orbitrap-XL mass spectrometer (Thermo) equipped with a Digital PicoView electrospray source platform (New Objective). The spectrometer was operated in data dependent mode where the top 9 most abundant ions in each MS scan were subjected to CAD (35% normalized collision energy, isolation width = 2.8 Da, threshold = 20,000). Dynamic exclusion was enabled with a repeat count of 1 and exclusion duration of 30 seconds. ESI voltage was set to 2.2 kV.

Data processing: MS spectra were recalibrated using the background ion $(Si(CH_3)_2O)_6$ at *m/z* 445.12 +/- 0.03 and converted into a Mascot generic file format (.mgf) using multiplierz scripts (4, 5). Spectra were searched using Mascot (version 2.3) against three appended databases consisting of: i) human protein sequences (downloaded from RefSeq on 07/11/2011); ii) common lab contaminants and iii) a decoy database generated by reversing the sequences from these two databases. For Mascot searches, precursor tolerance was set to 1 Da and product ion tolerance to 0.6 Da. Search parameters included trypsin specificity, up to 2 missed cleavages, fixed carbamidomethylation (C, +57 Da) and variable oxidation (M, +16 Da). Precursors with a recalibrated mass error greater than 10 ppm were discarded. Spectra matching to peptides from the reverse database were used to calculate a global false discovery rate, and were discarded. Data were further processed to remove peptide spectral matches (PSMs) to the forward database with an FDR greater than 1.0%. Protein abundance was estimated by calculating the sum of the extracted ion chromatograms area of the three most intense peptides(6). Any protein identified in more than 1% of 108 negative TAP controls was removed from the sets of SOX2 interactors (7).

SOX2 binding partner analysis

The list of proteins associated with pluripotent factors (pluripotency network) (8-10) was compiled from lists downloaded from BIOGRID. Official HGNC (HUGO Gene Nomenclature Committee) symbols for the human orthologues of all pluripotency genes were retrieved. The significance of the overlap observed between the 291 pluripotency genes (a full list of genes is available in Table S1) and the 45 SOX2 interactors common to all SCCs was calculated using a Fisher's exact test, with a background set consisting of 2818 human genes with "Nuclear lumen" as cellular component annotation (GO: 0031981 and children). This set was extended to include 61 genes from the two test sets. Transcriptional regulators were defined as genes with "Regulation of transcription, DNA-dependent" as biological process annotation (GO: 0006355 and children). Using protein abundance data from the three most intense peptides for each protein, we

generated a relative protein abundance score with color intensity using the R package heatmap.2.

RNA interference and quantitative RT-PCR assays

sh*ΔNP63* sequence was published previously (11). Other shRNA sequences were obtained from the RNAi Consortium (TRC) (http://www.broadinstitute.org/rnai/trc; Table S7). Oligonucleotides were cloned into pLKO.1 or Dox-inducible pLKO-Tet-On lentiviral vector at AgeI and EcoRI sites. Viral transduction and generation of cell lines stably expressing shRNAs were performed as previously described (12). Optimization for suppression of targeted genes was performed via immunoblotting with antibodies against SOX2 or p63 after treating with Dox at different concentration for different time-points. Quantitative RT-PCR assays were also performed as previously described (12) with results normalized to *18SrRNA* expression. Primers are listed in Table S8.

Cell proliferation assays, anchorage-independent growth assays

Cell proliferation assays were performed as previously described (12) with a little modification. Specifically, for Dox-inducible shRNA, cell medium were changed every 48 h with fresh Dox at 25 ng/ml in indicated cells. Cell proliferation was assayed at day 1, 3, 5, and 7. Cell numbers at day 3, 5 and 7 were corrected for the ratio of cells from the day 1 reading to correct for plating unevenness. Cell viability determined by CellTiter-Glo (Promega) for Dox-treated KYSE or TT cell lines with inducible sh*LacZ*, sh*SOX2* or sh*TP63* was normalized to Dox-untreated control cell lines and presented as fraction over time-course. For sh*ETV4*, measurements of cell viability were plotted as cell growth curve after seeding. A representative experiment is shown with viability ± 1 standard deviation (s.d.) of the reading from the six wells shown. Effects of RNAi on cellular proliferation were analyzed with the use of GraphPad Prism software. Curves for cell viability fraction to the non-Dox treated cell lines for sh*LacZ*, sh*SOX2* or sh*TP63* expressing cell lines were modeled to fit exponential decay curve and compared to the non-decay model (k=0). For sh*ETV4*, growth curve for each shRNA-expressing cell line

was modeled to fit exponential growth curve and compared to that of the corresponding sh*LacZ*-expressing control cell line. *F* tests were used to determine *P* values. *P* values < 0.05 were considered significant. Anchorage-independent growth assays were performed as previously described (12).

Immunoblotting, immunoprecipitation, and GST pull-down assays

To determine protein expression, cells were lysed in modified RIPA (mRIPA) buffer (50 mM Tris (pH. 7.4), 150 mM NaCl, 1 mM EDTA, 1% NP40, 1 mM Na_3VO_4, 1 mM NaF, and Protease inhibitors) and immunoblotting was performed as previously described(13). For immunoprecipitation, whole cell lysate of KYSE70 cells or 293T cells transfected with pcDNA3-ΔNp63α-FLAG and pcDNA3-SOX2 were prepared in mRIPA buffer followed by incubation with SOX2 antibody or goat IgG antibody and Dynabeads Protein G (Invitrogen) or anti-FLAG-agarose on a rotator overnight at 4°C. After washing with mRIPA buffer, the beads were boiled for 10 min in 2x SDS sample buffer. The immunoprecipitates were then processed for immunoblotting. GST pull-down assay was performed as previously described(13) with modification. Briefly, GST, GST-ΔNp63α and GST-SOX2 proteins were purified. SOX2 protein was produced by cleavage of purified GST-SOX2 protein with Thrombin (GE Healthcare) and then incubated with equal amount of GST-tagged proteins in binding buffer (1x PBS, 0.5% Triton X-100) for 1 h at room temperature followed by four washes with binding buffer. Bound proteins were resolved in SDS-PAGE gel and probed with anti-SOX2 antibody.

Chromatin immunoprecipitation, quantitative PCR (qPCR), and Ilumina sequencing

SOX2, p63 and JunD ChIP and qPCR were performed as previously described (14), except that fragmented chromatin with lengths of 200-to-1500bp after sonication was incubated overnight at 4ºC with pre-incubated Dynabeads Protein G with either anti-SOX2 (R&D, AF2018), anti-p63 (Santa Cruz, sc-8431) or with control goat IgG antibodies. Primers used for qPCR are listed in Table S8. For ChIP-sequencing, immunoprecipitated DNA with indicated antibody and input DNA without

immunoprecipitation were treated with RNase A and Proteinase K, followed by sonication to enrich DNA fragment lengths between 100-300bp. Ten to 50ng of DNA were used for the library construction. DNA libraries for Illumina cluster generation were prepared according to the manufacturer's protocol. Cluster generation and sequencing was done with Illumina HiSeq 2000.

ChIP-seq data analysis and motif analysis

Peak calling: Sequenced reads were mapped to human reference genome (build 37, hg18) using bowtie aligner (15). Uniquely aligned reads were subsequently normalized for copy number variation to average DNA copies on each cell line used as previously described (14). The summits of binding sites for SOX2 or p63 were detected with the use of MACS (16), with a *p*-value cutoff of 10^{-5} and default values for other parameters. Binding profiles were generated using MACS with the resolution of 10 bp, and were used to generate representative images with IGV (17). Significantly enriched peaks in composite data from the three SCC lines were also identified using MACS to increase the power of binding sites detection. The composite data for SOX2 or p63 were direct combinations of aligned reads used for each cell line, which has approximately equal number of reads. Data are publicly available at GEO (accession no. GSE46837).

Correlation analysis: For correlation analysis between SOX2 ChIP-seq datasets, a complete set of peaks composed of all identified peaks from SOX2 ChIP-seq from H9 and three SCC cell lines were used. Peaks from different dataset whose summits are within 600 bp from each other are regarded as overlap and replaced by the genomic locus at the midpoint until the complete peak set contains no "duplicated" peaks. Correlation matrix of ChIP-seq datasets is constructed from Pearson correlations between the peak occurrence profiles of ChIP-seq datasets.

Peak comparison: SOX2 ChIP-seq data for H1 cells (18) were downloaded from GEO (GSM456570), and sequenced reads were mapped to human reference genome (build 37, hg18) using bowtie aligner. Uniquely aligned reads were used for identifying SOX2

binding peaks using MACS with a p-value cutoff of $1x10^{-5}$ and default values for other parameters. Due to the lack of accompanying DNA input control data, ChIP-seq data could not be normalized for copy number variation or sequencing bias along the genome. Overlaps of SOX2 occupancy from our ChIP-seq data were calculated by comparing with the SOX2 occupancy identified from H1 cells.

Co-occupancy analysis: For comparison of SOX2 and p63 binding, numbers of genomic intervals shared by two or more samples were counted and represented in Venn diagrams. Binding sites whose summits are within 600 bp from each other are considered as overlap. To develop a list of high-confidence SOX2-occupied regions, peaks present in two or more of the three individual SOX2 ChIP-seq data from SCC cells were identified. These SOX2 peaks with and without evidence for significant p63 occupancy from the composite p63 ChIP-seq analysis were distinguished. Conversely, high-confidence p63-occupied regions, whose peaks found in two of three p63 ChIP-seq experiments, were queried for the presence of SOX2 occupancy in the composite SOX2 ChIP-seq data. Heatmap of binding profiles near binding sites was drawn using the heatmap module from cistrome project (http://cistrome.org/). Target genes by SOX2 or p63 are defined as genes whose TSS are within 50kb from SOX2 or p63 binding sites.

Motif analysis: SOX2, p63, OCT4 or AP1 consensus binding motif (19-21) occurrence along human reference genome sequence was identified using MAST module in MEME suite (22). Details about threshold and number of occurrences along the genome for each motif are listed in Table S9. Motifs closest to the center of SOX2 occupancy were used to profile the relative distance between motifs. SOX2-occupied regions without evidence of p63 binding were further scanned for sequence motifs with use of the MEME-ChIP module in MEME suite (23). Known motifs at the fixed distance from primary SOX2 motif identified from MEME algorithm were searched using SpaMo (24) module in MEME suite as described.

cDNA library construction for RNA-seq

KYSE70 stable cells expressing Dox-inducible sh*SOX2* or sh*TP63* were treated with or without Dox for 4 d. Total RNA were prepared and suppression of target genes were confirmed by quantitative RT-PCR. For each sample, polyA$^+$ RNA was isolated from 5 ug of total RNA using Dynabeads mRNA DIRECT Kit (Ambion). RNA was fragmented by incubation at 98 °C for 41 min in 1× RNA restoration buffer (Ambion). cDNA was synthesized with 20-40 ng of polyA$^+$ RNA with a peak length around 600-700 bases with the use of SuperScript III Reverse Transcriptase, SuperScript Double-Stranded cDNA Synthesis kit, and random primers (Invitrogen). Primers were annealed at room temperature for 10 min followed by first strand synthesis for 1 h at 50°C and second strand synthesis for 2.5 h at 16°C. Following cDNA size selection with 1.8× AMPure XP beads (Beckman Coulter), end repair, and A-tailing, adapter ligation was performed using the TruSeq DNA Sample Preparation kit (Illumina) according to manufacturer's instruction. Adaptor-ligated cDNAs were cleaned up with 0.8× AMPure XP beads followed by 8 cycles of PCR with primers complementary to the adaptor sequences with indexes. The PCR products were purified with 0.8× AMPure XP beads and then submitted to the Center for Cancer Genome Discovery at the Dana-Farber Cancer Institute for HiSeq 2000 sequencing.

RNA-seq analysis

RNA-seq reads for *SOX2* and *TP63* RNAi experiments were aligned to human reference genome (hg19) and exon-exon junctions (ensembl v64) with PRADA pipeline (25). Transcriptome was collapsed to gene level using the GENCODE v12 transcriptome model. RPKM values were generated by RNA-seQC (26) for each experiment, $\log_2$ transformed ($\log_2$ (RPKM+1)), and quantile-normalized. Gene expressions for each condition (sh*LacZ*, sh*SOX2*, and sh*TP63*) were represented by the mean of duplicates, and were used to calculate the $\log_2$ fold change between two experimental conditions. Gene expression changes (in $\log_2$ ratios) following suppression of either *SOX2* or *TP63* were plotted against each other. Genes whose $\log_2$ fold changes are more than $\log_2$ (1.5) were regarded as differentially expressed upon suppression via shRNA. Gene expression values in RPKM for TCGA lung SCC samples

were obtained through https://tcga-data.nci.nih.gov/docs/publications/lusc_2012/. While the values are log normally distributed, for comparison purpose $\log_2$ transformed ($\log_2$ (RPKM+1)) values were used. Of 144 samples with available RNA-seq data, *SOX2* high-expressing lung SCC samples at the top quartile were used for average expression of each annotated gene. Correlation to *SOX2* gene expression for expression of each gene was determined by Pearson's correlation (r value). For scatter plots of gene expression versus copy number for *SOX2* and *TP63,* inferred copy number from SNP 6.0 array data in TCGA lung SCC samples were obtained and plotted against the transformed RPKM value from corresponding RNA-seq data from TCGA.

Supplemental References

1. Tsukada Y-I, Zhang Y. Purification of histone demethylases from HeLa cells. *Methods (San Diego, Calif)*. 2006;40(4):318–326.

2. Adelmant G et al. DNA ends alter the molecular composition and localization of Ku multicomponent complexes. *Mol Cell Proteomics*. 2012;11(8):411–421.

3. Ficarro SB et al. Improved electrospray ionization efficiency compensates for diminished chromatographic resolution and enables proteomics analysis of tyrosine signaling in embryonic stem cells. *Anal Chem*. 2009;81(9):3440–3447.

4. Askenazi M, Parikh JR, Marto JA. mzAPI: a new strategy for efficiently sharing mass spectrometry data. *Nature methods*. 2009;6(4):240–241.

5. Parikh JR et al. multiplierz: an extensible API based desktop environment for proteomics data analysis. *BMC Bioinformatics*. 2009;10:364.

6. Silva JC, Gorenstein MV, Li G-Z, Vissers JPC, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics*. 2006;5(1):144–156.

7. Rozenblatt-Rosen O et al. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature*. 2012;487(7408):491–495.

8. Wang J et al. A protein interaction network for pluripotency of embryonic stem cells. *Nature*. 2006;444(7117):364–368.

9. Pardo M et al. An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell stem cell*. 2010;6(4):382–395.

10. van den Berg DLC et al. An Oct4-centered protein interaction network in embryonic stem cells. *Cell stem cell*. 2010;6(4):369–381.

11. Sabbisetti V et al. p63 promotes cell survival through fatty acid synthase. *PloS one*. 2009;4(6):e5877.

12. Bass AJ et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nature Genetics*. 2009;41(11):1238–1242.

13. Ma Q et al. FoxO1 mediates PTEN suppression of androgen receptor N- and C-terminal interactions and coactivator recruitment. *Mol Endocrinol*. 2009;23(2):213–225.

14. Watanabe H et al. Integrated cistromic and expression analysis of amplified NKX2-1 in lung adenocarcinoma identifies LMO3 as a functional transcriptional target. *Genes & development*. 2013;27(2):197–210.

15. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.

16. Zhang Y, Liu T, Meyer C, Eeckhoute J. Model-based analysis of ChIP-Seq (MACS). *Genome*. 2008;

17. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics*. 2013;14(2):178–192.

18. Lister R et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315–322.

19. Engelen E et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. *Nature Genetics*. 2011;43(6):607–611.

20. Kouwenhoven E et al. Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. *PLoS genetics*. 2010;6(8).

21. Wingender E et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic acids research*. 2000;28(1):316–319.

22. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. 1998;14(1):48–54.

23. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets.

Bioinformatics. 2011;27(12):1696–1697.

24. Whitington T, Frith MC, Johnson J, Bailey TL. Inferring transcription factor complexes from ChIP-seq data. *Nucleic acids research*. 2011;39(15):e98.

25. Berger MF et al. Integrative analysis of the melanoma transcriptome. *Genome research*. 2010;20(4):413–427.

26. DeLuca DS et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*. 2012;28(11):1530–1532.