Cancer Cell, Volume *30*

Supplemental Information

SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin

Giustina Ferone, Ji-Ying Song, Kate D. Sutherland, Rajith Bhaskaran, Kim Monkhorst, Jan-Paul Lambooij, Natalie Proost, Gaetano Gargiulo, and Anton Berns

Supplemental Data

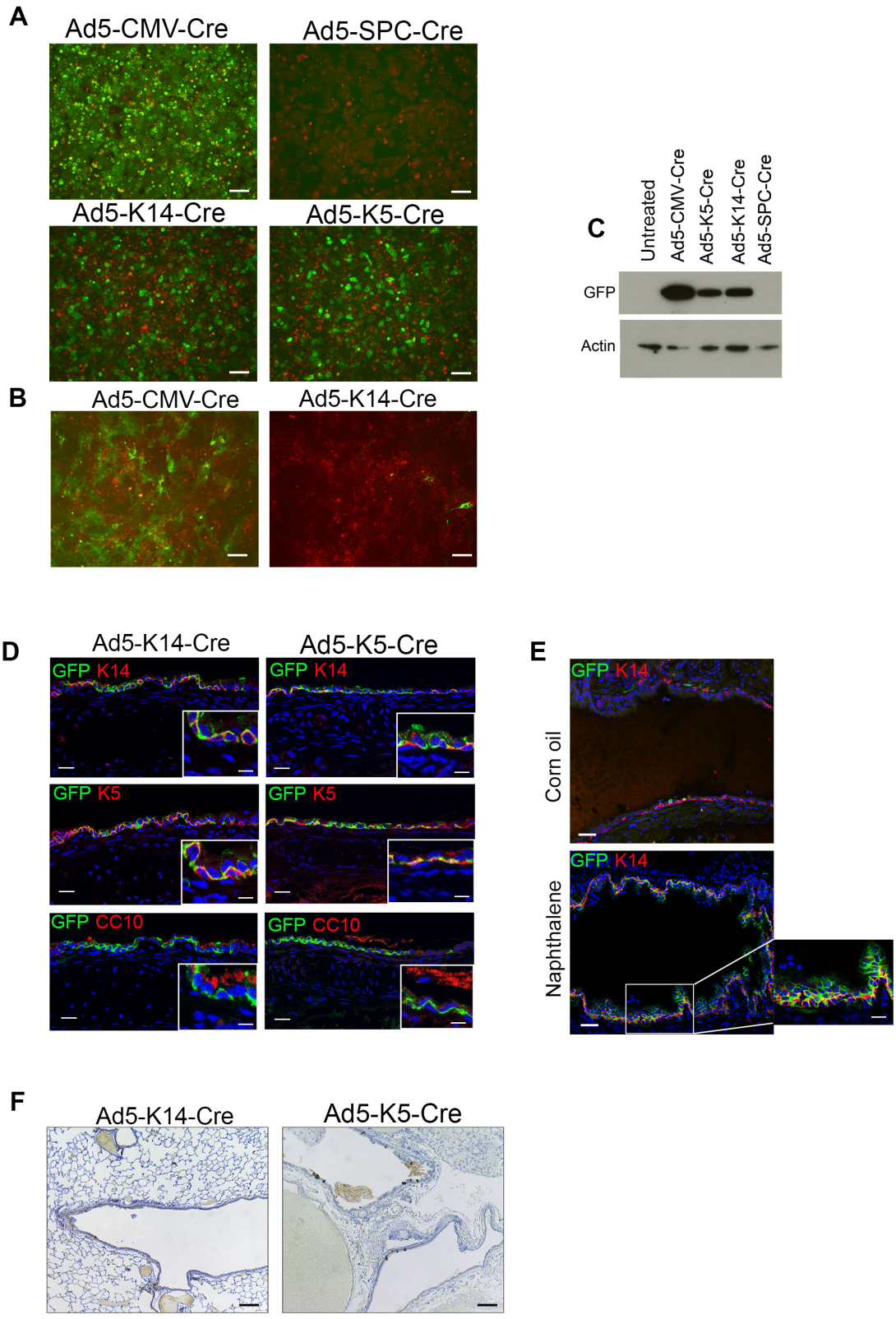

Figure S1 (related to Figure 1). Characterization of Ad5-K5-Cre and Ad5-K14-Cre both in vitro and in vivo.

(A) Fluorescent images of mouse keratinocytes isolated from *mT/mG* reporter mice at P0 and collected 48 hr upon infection with 100 MOI of the indicated Cre Adenoviruses: Ad5-CMV-Cre is used as positive control of infection and Ad5-SPC-Cre is used as negative control of infection; both Ad5-K14-Cre and Ad5-K5-Cre enable keratinocytes to switch from Tomato to GFP. (B) MEFs isolated from *mTmG* mice and collected 48 hr upon infection with either Ad5-CMV-Cre or Ad5-K14-Cre. (C) Immunoblotting analysis of protein extracts isolated from keratinocytes 48 hr upon infection with the indicated adenoviruses and incubated with GFP antibody. Actin is used to normalize the protein levels. (D) Dual IF with GFP/K14, GFP/K5, GFP/CC10, performed on tracheas isolated from *mTmG* mice 7 days following Ad5-K14-Cre and Ad5-K5-Cre infection; mice were administered naphthalene (250 mg/kg) 3 days prior to adenovirus infection. (E) Dual IF with GFP/K14, performed on tracheas isolated from *mTmG* mice 21 days following Ad5-K14-Cre infection; mice were administered naphthalene (250 mg/kg) or vehicle (corn oil control) 3 days prior to adenovirus infection. (F) GFP IHC staining showing positive cells in the bronchial lining of *mT/mG* mice 15 months upon naphthalene treatment and Ad5-K14-Cre and Ad5-K5-Cre injection. Scale bars, A, B, E, F:100 μm; D:20 μm. Inset in D:10 μm. Inset in E: 20 μm.

Table S1. Related to Figure 2, Figure 3, Figure 4
Tumor incidence and tumor types in mice with different genotypes

Genotype	# Mice with GFP^+ staining / Total # Mice analyzed	Tumor latency (months)	# Mice with Lung Tumors / # Mice with GFP^+ staining	# Mice with Atypical Hyperplasia / # Mice with GFP^+ staining	Histopathology of Lung Tumors
PC	**19**/25	10 -15	10/19	3/19	Broncho-alveolar Adenoma Polyp-like spindle cell neoplasia Adenofibroma Osteoma Adenocarcinoma Spindle cell sarcoma Hemangiosarcoma Chondroid sarcoma Neoplasia with rhabdoid differentiation
Fgfr1PC	**21**/27	1.5-6.5	16/21*	2/21	Sporadic squamous cell differentiation Broncho-alveolar Adenoma Polyp-like adenoma Adenofibroma Osteoma Adenocarcinoma Papillary carcinoma Spindle cell carcinoma Chondroid sarcoma
Sox2PC	**15/21**	**7-9**	**11/15**	**4/15**	**Squamous cell carcinoma**

*Sporadic squamous cell differentiation is present in 3 cases out of 16

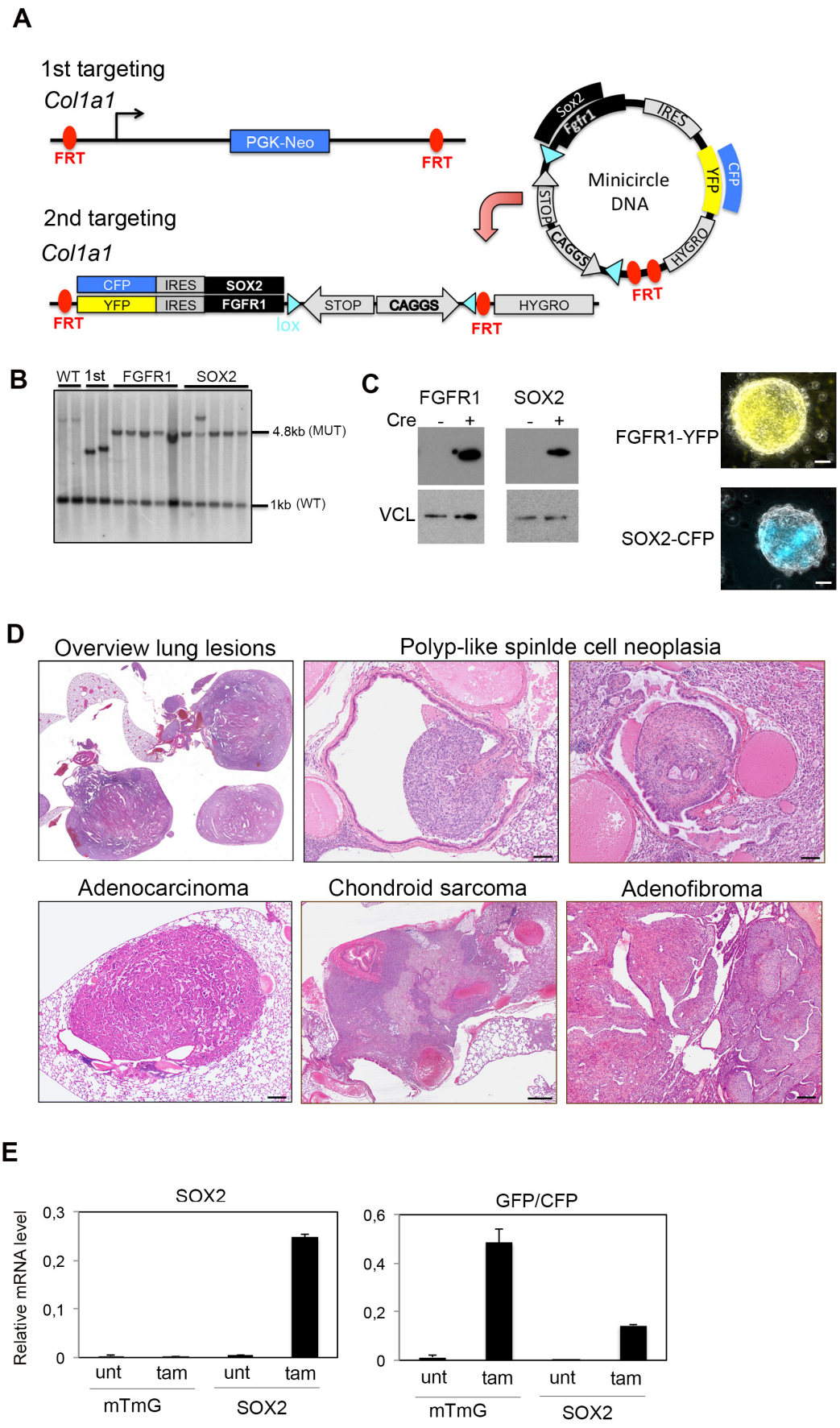

Figure S2 (related to Figure 3 and Figure 4). Generation of the conditional mice *LSL-Fgfr1*^{K656E} and *LSL-Sox2*

(A) Schematic representation of the Flp-recombinase mediated cassette exchange technology: in the first step a cassette containing PGK-Neomycin (PGK-Neo) and flanked by FRT sites was targeted to the *Col1a1* locus by homologous recombination; in the second step, a minicircle DNA containing the cDNA of either SOX2 or FGFR1 ^{K656E} followed by a reporter protein (respectively CFP and YFP) was transfected in ES cell clones positive to the first targeting, together with Flip recombinase, which mediated the cassette exchange.
(B) Southern blotting of BglII digested genomic DNA, isolated from ES cells and hybridized to the "*Col1a1* 3' probe", which is a 842 bp genomic fragment that anneals to a fragment of 1kb in wild-type (WT) mice (Line 1 and 2) and to a fragment of 4.9 kb in mutant (MUT) mice. Line 2 and 3 are the positive clones derived from the 1st targeting and used for the second targeting (1st). 5 out of 5 FGFR1 ES clones (Line 5 to 9) and 4 out of 5 SOX2 ES clones (Line 10 to 14) were correctly targeted. (C) ES cells positive to the Flp recombinase mediated cassette exchange and transfected with permeable Cre or left untreated. Left panel: Immunoblotting analysis with FGFR1 and SOX2 antibodies showing their expression upon Cre mediated cassette switch; Right panel: Confocal images of Cre transfected cells showing the expression of CFP and YFP. VCL (Vinculin) is used to normalize the protein levels.
(D) HE staining of lung sections showing that *Fgfr1PC* mice develop large, invasive and heterogeneous tumor lesions, as indicated. (E) Real time RT-PCR on RNA isolated from *mTmG;CreERT2* (mTmG) and *LSL-Sox2;CreERT2* (SOX2) MEFs and treated with 2.5mM of Tamoxifen (tam) for 48 hr or left untreated (unt). Data are represented as means ±SD. Samples are normalized by using Actin RNA level. Scale bars, C: 20 μm; D: 200 μm.

Table S2. Related to Figure 4
Scoring of IHC analysis of histopathological markers performed on human and mouse LSCC

Species	Case number	SOX2	P63	TTF-1
Human	1 (P[1]and C[2])	++/+++	+/++	_
	2 (P and C)	60 % +/+++	++	_
	3 (P)	++/+++	+++	_
	4 (C)	_	+/++	_
	5 (C)	++/+++	+++	_
	6 (P)	+++	+++	_
	7 (P and C)	+++	+/+++	_
	8 (C)	+++	+++	_
	9 (C)	++/+++	++/+++	_
	10 (P and C)	+/++ (also cytoplasm)	+++	_
	11 (P)	10% +	++/+++	_
Mouse	14GFE121 (Ad5-**K14**-Cre)	++	+++	_
	14GFE121 (Ad5-**K14**-Cre)	70% +/+++	40% +	_
	15GFE019 (Ad5-**K14**-Cre)	90 % +/+++	70% +	_
	15GFE021 (Ad5-**K14**-Cre)	70% +/+++	65% +	_
	15GFE024 (Ad5-**K14**-Cre)	90% ++/+++	80% +	_
	15GFE028 (Ad5-**K14**-Cre)	70% +/+++	60% +	_
	16GFE001 (Ad5-**SPC**-Cre)	90% +/+++	80% +	_
	16GFE003 (Ad5-**SPC**-Cre)	90% +/+++	65% +	_
	16GFE011 (Ad5-**CC10**-Cre)	99% +/+++	80% +	_
	16GFE017 (Ad5-**SPC**-Cre)	95% +/+++	80% +	_
	16GFE018 (Ad5-**SPC**-Cre)	90% +/+++	50% +	_
	16GFE019 (Ad5-**CC10**-Cre)	90% +/+++	70% +	_
	16GFE020 (Ad5-**CC10**-Cre)	99% +/+++	80%	-

[1]P: peripheral location
[2]C: central location

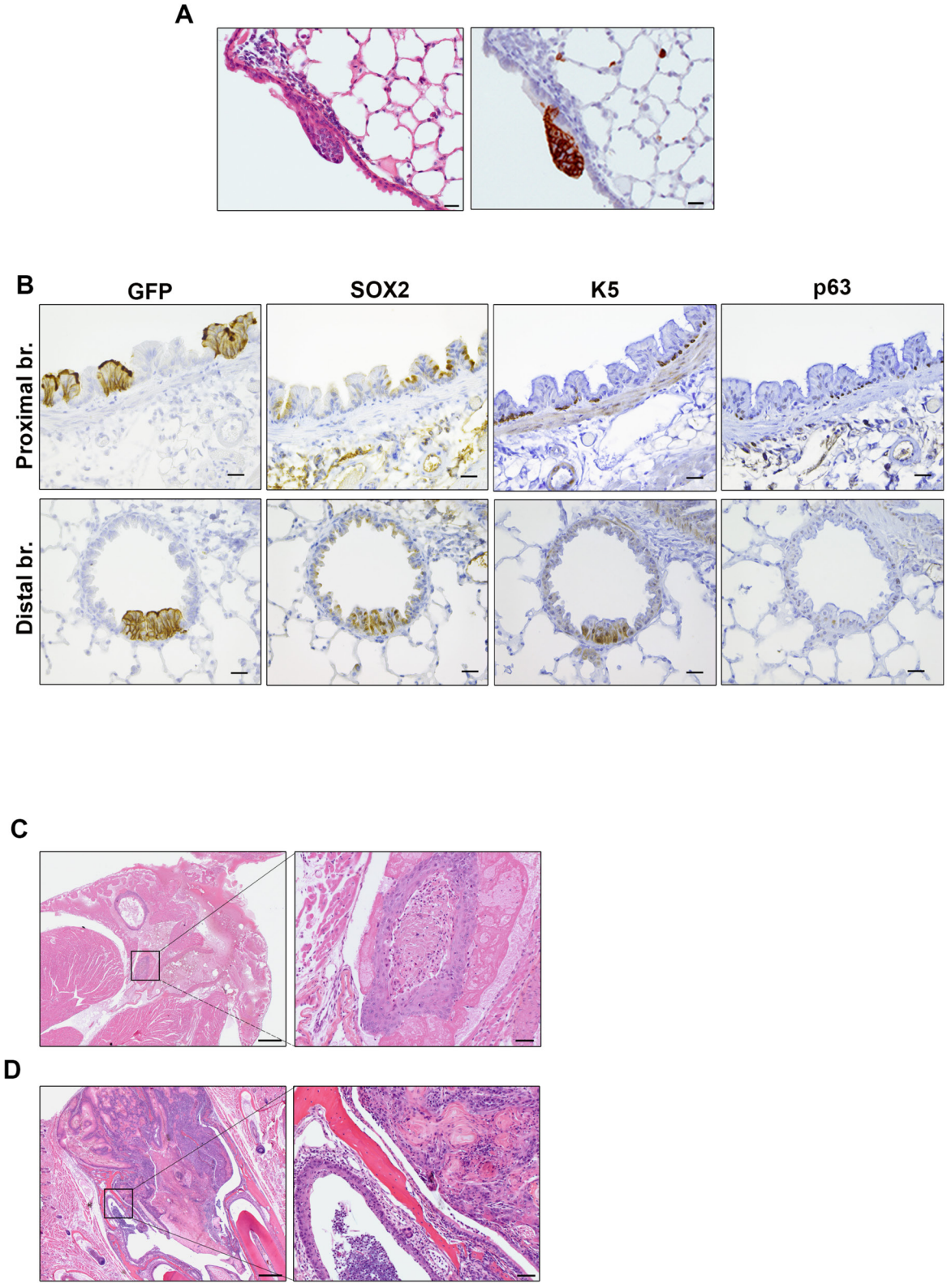

Figure S3 (related to Figure 4) Histological characterization of *Sox2PC* mice
(A) Atypical hyperplasia of bronchial cells showed by HE staining (left panel) and monitored by expression of GFP (right panel). (B) IHC analysis performed on lung tissues using antibodies against the indicated markers of LSCC in proximal bronchioles (Proximal br.) and distal bronchioles (Distal br.). (C) HE staining showing a large lesion of SCC in the lumen of the left atrium of the heart (left panel) and its higher magnification (right panel). (D) HE staining of sections of nasal cavity, showing a large lesion of SCC (right panel) and its higher magnification (right panel). Scale bars, A-B: 20 μm; C-D left panel: 500 μm; C-D right panel: 50 μm.

Table S3. Related to Figure 5
Human RNA seq samples (LSCC and LADC) downloaded from TCGA

Human LSCC	Human LADC
TCGA-39-5019-01A-01R-1820-07	TCGA-44-5645-01A-01R-1628-07
TCGA-34-5929-01A-11R-1820-07	TCGA-44-6146-01A-11R-1755-07
TCGA-22-5471-01A-01R-1635-07	TCGA-44-6147-11A-01R-1858-07
TCGA-60-2712-01A-01R-0851-07	TCGA-44-6148-11A-01R-1858-07
TCGA-22-5489-11A-01R-1635-07	TCGA-44-6776-01A-11R-1858-07
TCGA-60-2707-01A-01R-0851-07	TCGA-44-6778-11A-01R-1858-07
TCGA-66-2758-01A-02R-0851-07	TCGA-50-5931-01A-11R-1755-07
TCGA-66-2781-01A-01R-0851-07	TCGA-44-6777-11A-01R-1858-07
TCGA-66-2742-01A-01R-0980-07	TCGA-49-4512-11A-01R-1858-07
TCGA-18-3415-01A-01R-0980-07	TCGA-49-6742-11A-01R-1858-07
TCGA-66-2794-01A-01R-1201-07	TCGA-49-6743-11A-01R-1858-07
TCGA-66-2800-01A-01R-1201-07	TCGA-49-6744-01A-11R-1858-07
TCGA-21-1071-01A-01R-0692-07	TCGA-49-6745-11A-01R-1858-07
TCGA-66-2770-01A-01R-0851-07	TCGA-50-5931-01A-11R-1755-07
TCGA-22-5474-01A-01R-1635-07	TCGA-50-5932-11A-01R-1755-07
TCGA-22-5473-01A-01R-1635-07	TCGA-50-5935-01A-11R-1755-07
TCGA-46-3768-01A-01R-0980-07	TCGA-55-6968-01A-11R-1949-07
TCGA-18-4083-01A-01R-1100-07	

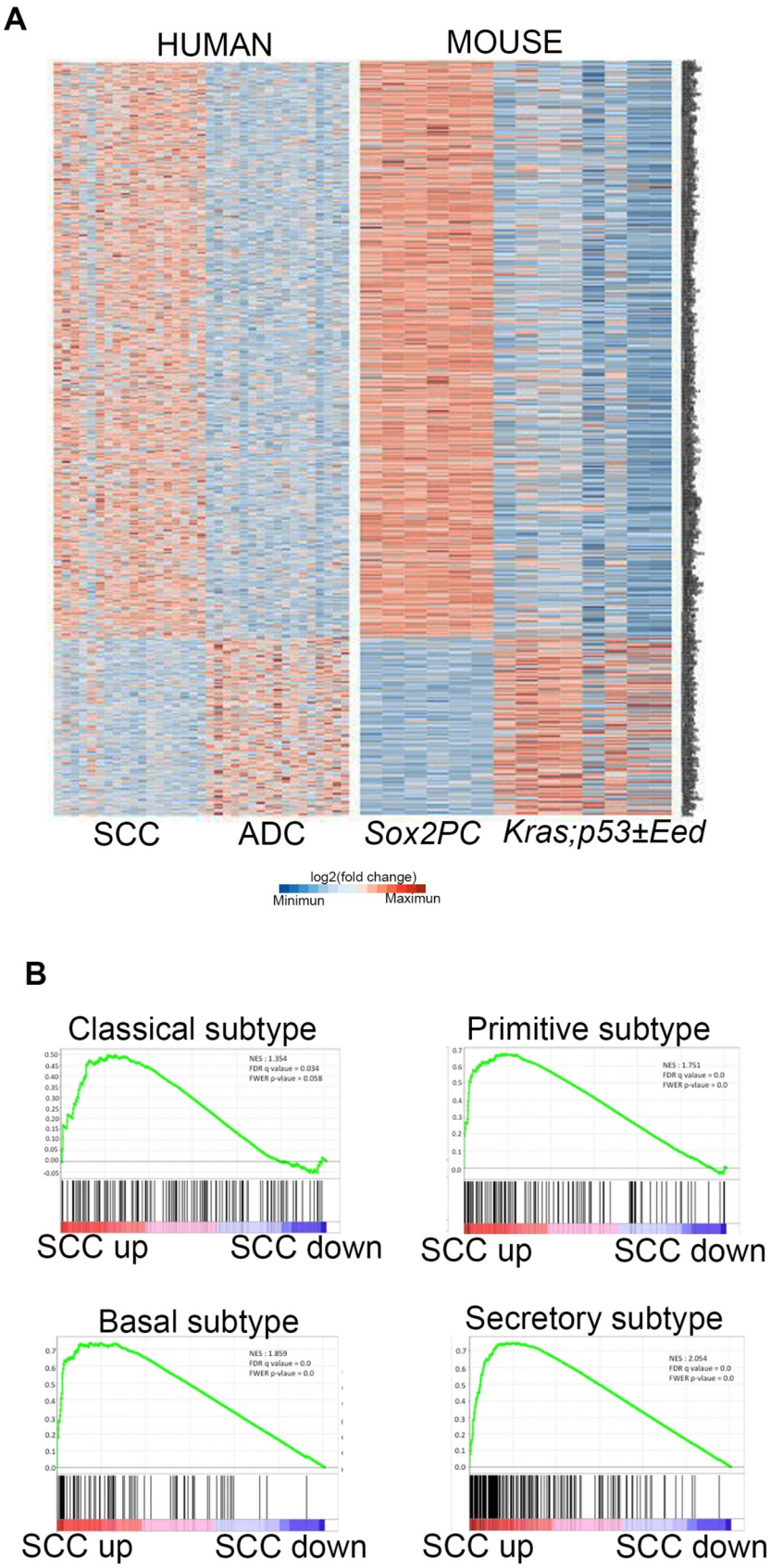

Figure S4 (related to Figure 5) Transcriptional profile of *Sox2PC* mice
(A) Unsupervised clustering heatmap for genes differentially expressed between LSCC and LADC of human (SCC, ADC on the left) or mouse (*Sox2PC* and *Kras;p53±Eed* on the right) origin. (B) GSEA plot for genes upregulated (SCC up versus SCC down) in the four different LSCC subtypes. (Classical, Primitive, Basal, Secretory).

Table S4. Related to Figure 6
Scoring of IHC analysis of immune cell markers performed on human and mouse LSCC

Species	Case number	PD-L1	PD-1[2]	HIF-1α	CD4[2]	CD8[2]	Neutrophils[3]
Human	1 (P[1]and C[2])	0.5% +	++ /+	Focally weak	+ / -	++/+	Sporadic
	2 (P and C)	20% +	+ /-	Partially weak	++ /-	+/-	+++
	3 (P)	_	++/+	Partially weak	++ /-	++/-	+/++
	4 (C)	2% +	++/++	Partially weak	**++/++**	++/+	+++
	5 (C)	0.5% +	_	Weak to strong	+ /-	+/-	Sporadic
	6 (P)	_	+/-	Partially weak	+ / -	+/-	Sporadic (eosinophils++)
	7 (P and C)	20% +	++/ +	Cluster weak	++ / -	++/-	Sporadic
	8 (C)	One cluster +	+/-	Partially weak	+/-	+/-	Sporadic
	9 (C)	80% +	+/+	Focally weak	+/-	+/++	Sporadic
	10 (P and C)	2% +	++/ +	_	++/++	++/-	Sporadic
	11 (P)	2% +	++/++	Partially weak	++/+	++/+	++
	12 (P and C)	20% +	+/+	Focally weak	+/-	+/-	Sporadic
Mouse	14GFE121 (Ad5-**K14**-Cre)	NA[1]	NA	NA	NA	NA	+
	15GFE019 (Ad5-**K14**-Cre)	50% +	Sporadic	Partially +	++/+	+/+	+
	15GFE021 (Ad5-**K14**-Cre)	60% +	Sporadic	Partially +	++/+	+/+	+/++
	15GFE024 (Ad5-**K14**-Cre)	40% +	Sporadic	Partially +	++/+	+/-	++
	15GFE028 (Ad5-**K14**-Cre)	50% +	Sporadic	Partially +/++	++/+	+/+	++
	16GFE001 (Ad5-**SPC**-Cre)	40% +	Sporadic	Partially +/++	++/+	+/+	+/++
	16GFE003 (Ad5-**SPC**-Cre)	60% +	Sporadic	Partially +	+/-	+/+	+
	16GFE011 (Ad5-**CC10**-Cre)	50% +	Sporadic	Partially +/++	+/+	+/+	+/++
	16GFE017 (Ad5-**SPC**-Cre)	40% +	Sporadic	Partially +	+/+	_ /+	+
	16GFE018 (Ad5-**SPC**-Cre)	50% +	Sporadic	Partially +	+/+	+/+	+
	16GFE019 (Ad5-**CC10**-Cre)	40% +	Sporadic	Partially +	+/+	-/+	+/++
	16GFE020 (Ad5-**CC10**-Cre)	40% +	Sporadic	Partially +	+/+	+/+	+/++

[1]NA: not available
[2]Infiltrating cells in the tumor stroma/among tumor cells
[3]Human, based on HE staining; mouse, based on LY6G.

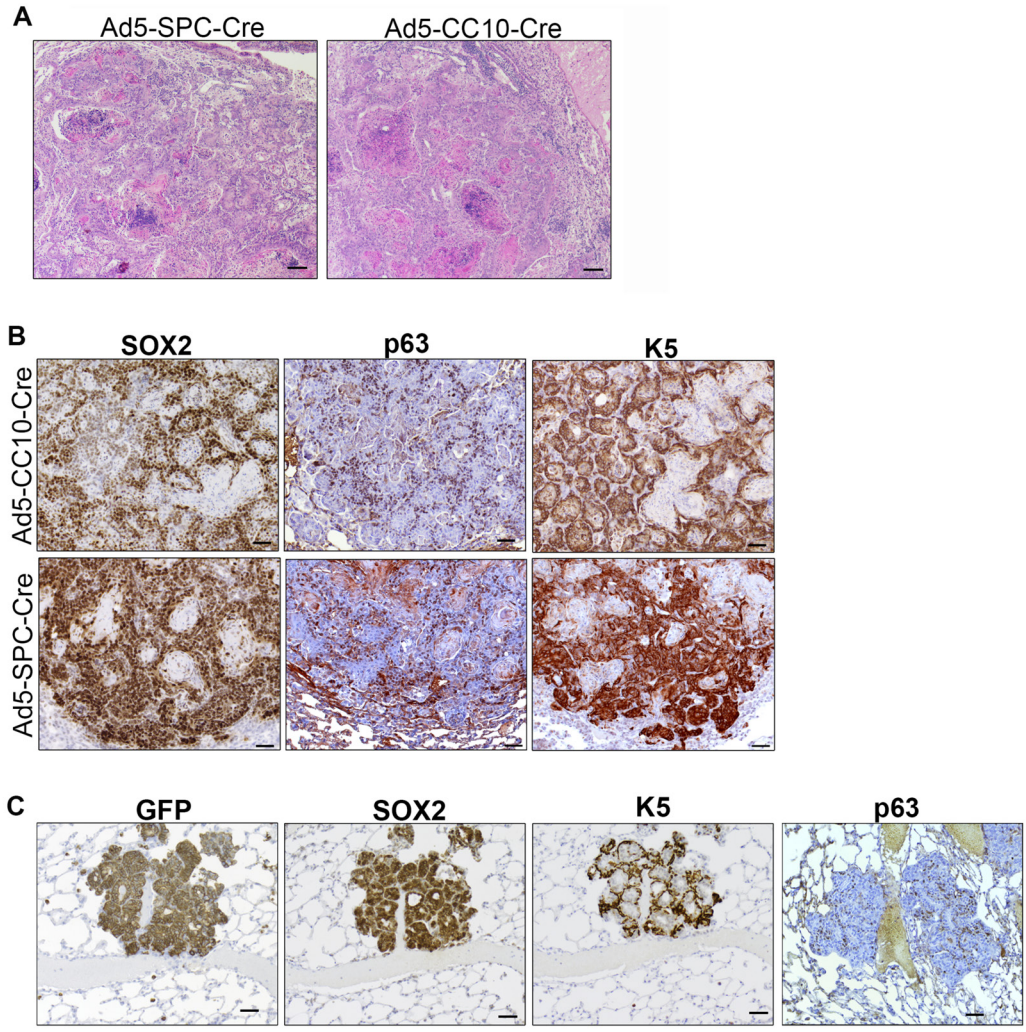

Figure S5 (related to Figure 7) Morphology and biomarkers of LSCC isolated from Ad5-SPC/CC10-Cre injected mice

(A) Histopathology of large lesions of LSCC of *Sox2PC* mice injected with Ad5-SPC-Cre or Ad5-CC10-Cre, as indicated. (B) IHC analysis of squamous markers performed on sections of LSCC isolated from *Sox2PC* mice injected with the indicated adenoviruses. (C) A representative early lesion of Ad5-SPC-Cre injected mice stained with the indicated squamous markers. Scale bars, A 100 μm; B, C: 50 μm.

Table S5. Relative to Figure 7
Significant differences between Ad5-SPC-Cre and Ad5-CC10-Cre injected *Sox2PC* mice

Adenovirus	# Mice with GFP+ staining / Total # Mice analyzed	# Mice with Lung Tumors / # Mice with GFP+ staining	Tumor latency (months)	#Lung SCC/ # Lung Tumors	Average of Carcinoma vs Early lesions	Biomarker staining
Ad5-SPC-Cre	**6**/8	6/6	7 -8	28/30	5 vs 29.5	**Single cell/small clusters**: $SPC^{pos}GFP^{pos}SOX2^{pos}$ or $SPC^{pos}GFP^{pos}CC10^{pos}SOX2^{pos}$ **Early lesions**: $SPC^{neg}GFP^{pos}CC10^{pos}$ $SOX2^{pos}K5^{pos}p63^{pos}$ **Carcinoma:** $SPC^{neg}GFP^{pos}CC10^{neg}$ $SOX2^{pos}$ $K5^{pos}p63^{pos}$
Ad5-CC10-Cre	**4**/6	4/4	7-8	57/59	14.25 vs 6.25	**Single cell/small clusters**: $CC10^{pos}GFP^{pos}SOX2^{pos}$ **Early lesions**: $GFP^{pos}CC10^{pos}$ $SOX2^{pos}$ $K5^{pos}p63^{pos}$ **Carcinoma:** $GFP^{pos}CC10^{neg}$ $SOX2^{pos}$ $K5^{pos}p63^{pos}$

Supplemental Experimental Procedures

Mouse generation

We targeted FVB ES cells with a vector containing a Flp-in module just after the 3′UTR of the *Col1a1* locus (Beard et al., 2006). This module, named Col1A1-frt, serves as a docking site for introduction of transgene-coding plasmids by Flp recombinase-mediated integration. The vector contained also a PGK-Neomycin cassette for positive selection of ES clones. The Col1A1-frt targeted GEMM-ESC clones were subsequently injected into morulae or blastocysts to produce chimeric mice to assess their transmittability. Both injected ES clones produced germline competent chimeras.

In the second step, a transgenic construct was introduced in the Col1A1-frt locus of germline competent ES clones, using the Flp recombinase.

The transgenic construct was carried by a minicircle DNA, devoid of bacterial elements (S2A). It has been previously demonstrated that bacterial DNA linked to a mammalian expression cassette resulted in transcriptional silencing of the transgene (Chen et al., 2004). The two transgenic constructs we used to generate *LSL-Sox2* and *LSL-Fgfr1*^{K656E}, were called Sox2-frt-invCAG and Fgfr1- frt-invCAG, respectively. They contained the cDNA of either SOX2 or FGFR1, which, following Cre mediated inversion is expressed from a constitutive CAG promoter (Figure S2A). Before the inversion the transgene is flanked by a Lox-stop-Lox cassette. For identification of cells in which recombination had occurred, we inserted an IRES followed by a fluorescent protein: YFP to track FGFR1 expression and CFP to track SOX2 expression (Figure S2C). We introduced a mutation in the cDNA of FGFR1 (K656E) which results in its activation in the absence of ligand (Jin et al., 2003). These vectors were introduced into a Col1A1-frt targeted FVB ES cells with 100% efficiency. Colonies were screened by PCR and correctly targeted clones were confirmed by Southern blotting (Figure S2B). The *Col1a1* locus has been shown to allow for ubiquitous expression of transgenes when combined with the CAGGS promoter (Figure S2A) (Huijbers et al., 2014).

Correctly recombined ES cell clones were treated with a permeable Cre-recombinase in order to validate the activation of FGFR1 or SOX2 expression (Figure S2C left panel). ES cells also expressed the fluorescent protein, as indicated by confocal images (Figure S2C right panel). To exclude that any rearrangement had occurred in vivo, MEFs were isolated from E13.5 embryos obtained by crossing *LSL-Sox2* mice with *Rosa26-CreERT2* mice, and treated with Tamoxifen, to switch on the expression of the transgene. Both SOX2 and CFP were expressed in MEFs carrying the transgene only upon Tamoxifen treatment, as assayed by real time RT PCR (Figure S2E). WT MEFs were obtained from *mT/mG* E13.5 embryos as control of expression of the fluorescent protein (Figure. S2E).

Generation of tissue specific Cre Adenoviruses

In order to target basal cells we utilized adenoviral vectors carrying a Cre-recombinase gene whose expression is driven either by a DNA fragment containing the 2 kb human K14 promoter, or the 5.2 kb of upstream 5' flanking and promoter sequences from the bovine K5 locus (pHR2 plasmid generously provided by Sabine Werner).

The Cre open reading frame with an N-terminal synthetic intron and C-terminal polyadenylation signal was isolated from pOG231 (O'Gorman et al., 1997) and inserted in pDONR^{TM}221 under the described K5 and K14 promoters, previously inserted in the same vector. Cloned pDONR^{TM}221 constructs were then recombined into promoter-less pAd-PL DEST vectors (Invitrogen) by Gateway LR recombination, to generate Ad5-K5-Cre and Ad5-K14-Cre adenoviral constructs. High titer adenoviruses were amplified and purified for use in vivo by the University of Iowa Gene Transfer Vector Core, supported in part by the NIH and the Roy J. Carver Foundation, for viral vector preparation.

Cell culture

Newborn mouse skin was isolated from *mT/mG* mice and then placed in dispase o/n at 4 °C. The day after keratinocytes were isolated from epidermis by enzymatic dissociation in trypsin, and cultured in defined CnT-Prime Epithelial cell culture medium (CnT07, CELLnTEC) as described previously (Strachan et al., 2008). Once attached, they were infected with 100 MOI and analyzed for GFP expression 48 hours upon infection. Fluorescent signals were monitored under a Leica CTR6000 image microscope. Protein extracts were also collected and samples were analyzed by immunoblotting with anti-GFP antibodies.

MEFs were isolated from 13.5 postcoitum (p.c.) mouse embryos of either *mT/mG;CreERT2* or *LSL-Sox2;CreERT2* and *LSL-Fgfr1*^{K656E}*;CreERT2* mice. The embryos, after removal of internal organs, were dissociated and then trypsinized to produce single-cell suspensions. Cells were treated with 4-Hydroxytamoxifen (0.2mm) to switch on the expression of the transgenes. Total RNA was collected 48 hr upon treatment. RNA was extracted in TRIzol reagent (Invitrogen). Complementary DNA (cDNA) synthesis was obtained using SuperScript™ III Reverse Transcriptase (Invitrogen).

Real-time RT PCR was performed using the SYBR Green PCR master mix (Applied Biosystems) in the Applied Biosystems® StepOnePlus™ Real-Time PCR System. Expression of target genes was normalized for Actinb.

Gene expression profile analysis

Data set

For mouse LADC, we used GSE61190 dataset, which contains 19 tumor samples isolated from *Kras;p53* mice with and without Eed. RNAseq profiles for human primary tumor (LSCC and LADC) were obtained from The Cancer Genome Atlas (TCGA) dataset. Human LSCC was selected based on *SOX2* amplification and deletion of *PTEN, CDKN2A, CDKN2B* (Table S4).

Read mapping, assembly, and expression analysis

After quality filtering according to the illumina pipeline, 51-bp single-end reads were mapped to the mouse genome (assembly NCBIM37.67), using TopHat (2.0.12) (Trapnell et al., 2009). TopHat was run with default. Reads with mapping quality less than 10 and non-primary alignments were discarded. Remaining reads were counted using HTSeq-cout (Anders et al., 2015). Statistical analysis of the differential expression of genes was performed using DESeq2 (Love et al., 2014). Genes with False Discovery Rate (FDR) for differential expression lower than 0.05 were considered significant. Batch effect with in tumor samples from different source was corrected using ComBat with default options through the Bioconductor package sva 3.10 (Johnson et al., 2007; Leek and Storey, 2007).

Determination of differentially expressed genes in mouse and human LSCC

To identify genes correlating with the phenotypic groups, fold changes of gene expression in mouse LSCC (6 samples) was compared to mouse LADC (8 samples) and human LSCC (18 samples) compared to human LADC (17 samples). Multiple hypothesis testing was corrected for using the Benjamini and Hochberg method (BH) (Benjamini et al., 2001), and significantly differentially expressed genes are reported.

To identify genes correlating with the phenotypic groups, we used DESeq2 to compute the variance stabilized expression values between three groups: Basal SCC, Club SCC and AT2 SCC. The expression heatmap of tumor subtypes are plotted using unsupervised consensus clustering of the top 100 most variable genes. The genes with padj < 0.01 and log2fold change > 1 are considered to be significant.

Gene set enrichment and functional set enrichment analysis

Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was used to investigate the correlation of gene set significantly overrepresented in the transcriptome of either mouse or human LSCC.

Transcripts were ranked by the difference of classes (metric for gene ranking) and using the following settings: number of permutations = 1,000, permutation type = gene set, chip platform = Null, enrichment statistic = weighted, gene list sorting mode = real, gene list ordering mode = descending, max gene set size = 500, min gene set size = 15. The gene set were manually created specific for unregulated and downregualted genes in LSCC over LADC in both mouse and human tumor model. A gene set was identified as significantly enriched when associated with p value scores ≤ 0.05.

Functional enrichment analyses were generated with the DAVID tool (Huang da et al., 2009). The GO enrichment analysis was carried out in the "two lists mode", using the lists of DEGs and as background the corresponding list of expressed genes.

Significant GO terms (p value<0.05) were mapped with the REViGO online tool (http://revigo.irb.hr) with default parameters except for the resulting list that was setting as small size, which removes redundant GO terms and visualizes the semantic similarity of remaining terms (Supek et al., 2011). The results were visualized as bar charts.

Antibodies

We performed IHC for anti-GFP (goat polyclonal, 1:500, Abcam), anti–K5 (rabbit polyclonal, 1:2000, Chemicon), anti-Sox2 (mouse monoclonal, 1:1000, Cell signaling), anti-p63 (mouse monoclonal, 1:200, SantaCruz); anti-Fgfr1 (rabbit polyclonal, 1:1000, Cell signaling), anti-K14 (rabbit polyclonal, 1:10000, Covance), anti-TTF1 (mouse monoclonal, 1:1000, DAKO), anti-Ly6G (monoclonal mouse, 1:500 BD Bioscences), anti-F4-80 (clone CI:A3, 1:1000 AbD Serotec), MPO (rabbit polyclonal, 1:300, DakoCytomation), anti-CC10 (goat polyclonal, 1:200, Santa Cruz), anti-pro SPC (rabbit polyclonal, 1:1000, Millipore), anti-CD4 (rat polyclonal, 1:2000 eBioscience), anti-CD8 (rat polyclonal, 1:2000 eBioscience), anti-HIF-1α (rabbit polyclonal, NovusBio, 1:6000), anti-PD-1 (rabbit polyclonal, Protein Tech, 1:200), anti-PD-L1 (rabbit polyclonal, Protein Tech, 1:200). For IHC performed on human samples we used: anti-TTF1 (Monosan, MONX10584), anti-p63 (Immunologic, 4A4), anti-CD4 (Cell Marque, SP35); anti-CD8 (Dako, C8144B), anti-

PD-1 (AbCam, NAT), anti-PD-L1 (Cell Signalling, E1L3N). Streptavidin-peroxidase (DAKO) or Powervision Poly-HRP (Leica Microsystems) was used for visualization and diaminobenzidine as a chromagen (DAKO). We performed immunofluorescence analysis for anti-GFP and anti-K14 using as secondary antibodies Alexa Fluor 488 donkey anti-goat and Alexa Fluor 594 donkey anti-rabbit respectively.

Real time RT-PCR oligonucloetides

Oligonucleotide Name	Oligonucleotide Sequence
mSox2-total RT For	ctggactgcgaactggagaag
mSox2-total RT rev	tttgcacccctcccaattc
GFP-RT For	aagttcatctgcaccaccg
GFP-RT Rev	tgctcaggtagtggttgtcg
mSox2 endogenous RT For	ggcagagaagagagtgtttgc
mSox2 endogenous RT Rev	tcttctttctcccagcccta
mSox2 exogenous RT For	tggctctcctcaagcgtatt
mSox2 exogenous RT Rev	cccatacaatggggtaccttc
mKLK10-RT For	gcaagagtgtcaggtctcagg
mKLK10-RT Rev	ggaacagctcaggctcctatt
mDsg3-RT For	gatgaggacacgggtaaagc
mDsg3-RT Rev	accatcattacgacccagga
mTMPRSS11D-RT For	cagcagctcattgcttcaaa
mTMPRSS11D-RT Rev	tctcagcctagggctcattg
mADAM17-RT For	tgtggttatttaaatgcagatagtga
mADAM17-RT Rev	tctcttcactcgacgaacaaac

Supplemental Reference

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Beard, C., Hochedlinger, K., Plath, K., Wutz, A., and Jaenisch, R. (2006). Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. Genesis *44*, 23-28.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. Behav Brain Res *125*, 279-284.

Chen, Z. Y., He, C. Y., Meuse, L., and Kay, M. A. (2004). Silencing of episomal transgene expression by plasmid bacterial DNA elements in vivo. Gene Ther *11*, 856-864.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huijbers, I. J., Bin Ali, R., Pritchard, C., Cozijnsen, M., Kwon, M. C., Proost, N., Song, J. Y., de Vries, H., Badhai, J., Sutherland, K.*, et al.* (2014). Rapid target gene validation in complex cancer mouse models using re-derived embryonic stem cells. EMBO Mol Med *6*, 212-225.

Jin, C., McKeehan, K., Guo, W., Jauma, S., Ittmann, M. M., Foster, B., Greenberg, N. M., McKeehan, W. L., and Wang, F. (2003). Cooperation between ectopic FGFR1 and depression of FGFR2 in induction of prostatic intraepithelial neoplasia in the mouse prostate. Cancer Res *63*, 8784-8790.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Leek, J. T., and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet *3*, 1724-1735.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

O'Gorman, S., Dagenais, N. A., Qian, M., and Marchuk, Y. (1997). Protamine-Cre recombinase transgenes efficiently recombine target sequences in the male germ line of mice, but not in embryonic stem cells. Proc Natl Acad Sci U S A *94*, 14602-14607.

Strachan, L. R., Scalapino, K. J., Lawrence, H. J., and Ghadially, R. (2008). Rapid adhesion to collagen isolates murine keratinocytes with limited long-term repopulating ability in vivo despite high clonogenicity in vitro. Stem Cells *26*, 235-243.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800.
Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.