Supplemental Figures and Legends

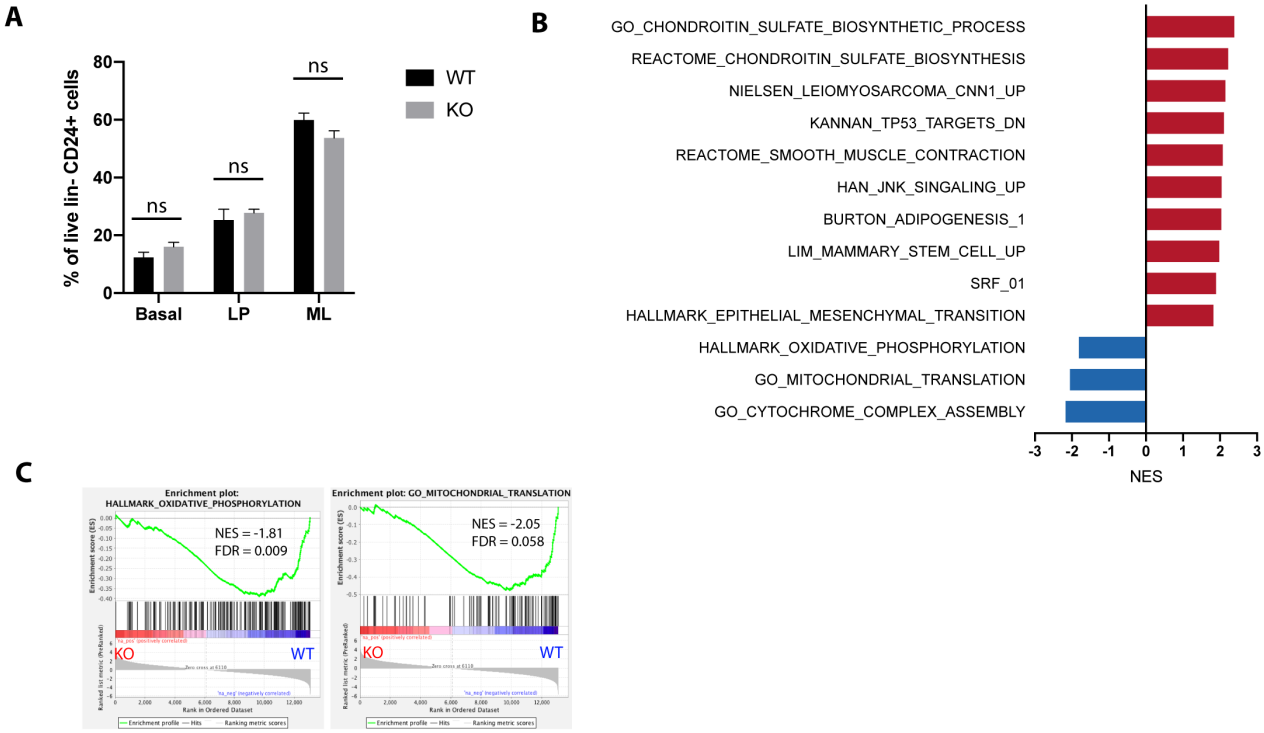

Figure S1. GSEA of ID4 knockout sorted basal cells, Related to Figure 1.

A) Proportions of basal, luminal progenitor (LP), and mature luminal (ML) subpopulations between ID4 WT and KO mammary epithelial cells. Unpaired two-tailed students t-test. ns = not significant. Error bars represent SEM. N=4. **B)** The top 10 positive and negatively enriched pathways with an FDR<0.1 are displayed. Only 3 pathways were negatively enriched with an FDR<0.1. **C)** Representative GSEA enrichment plots displaying the profile of the running Enrichment Score (green) and positions of gene set members on the rank ordered list for pathways related to cell growth**.** NES and FDR are indicated on the plots.

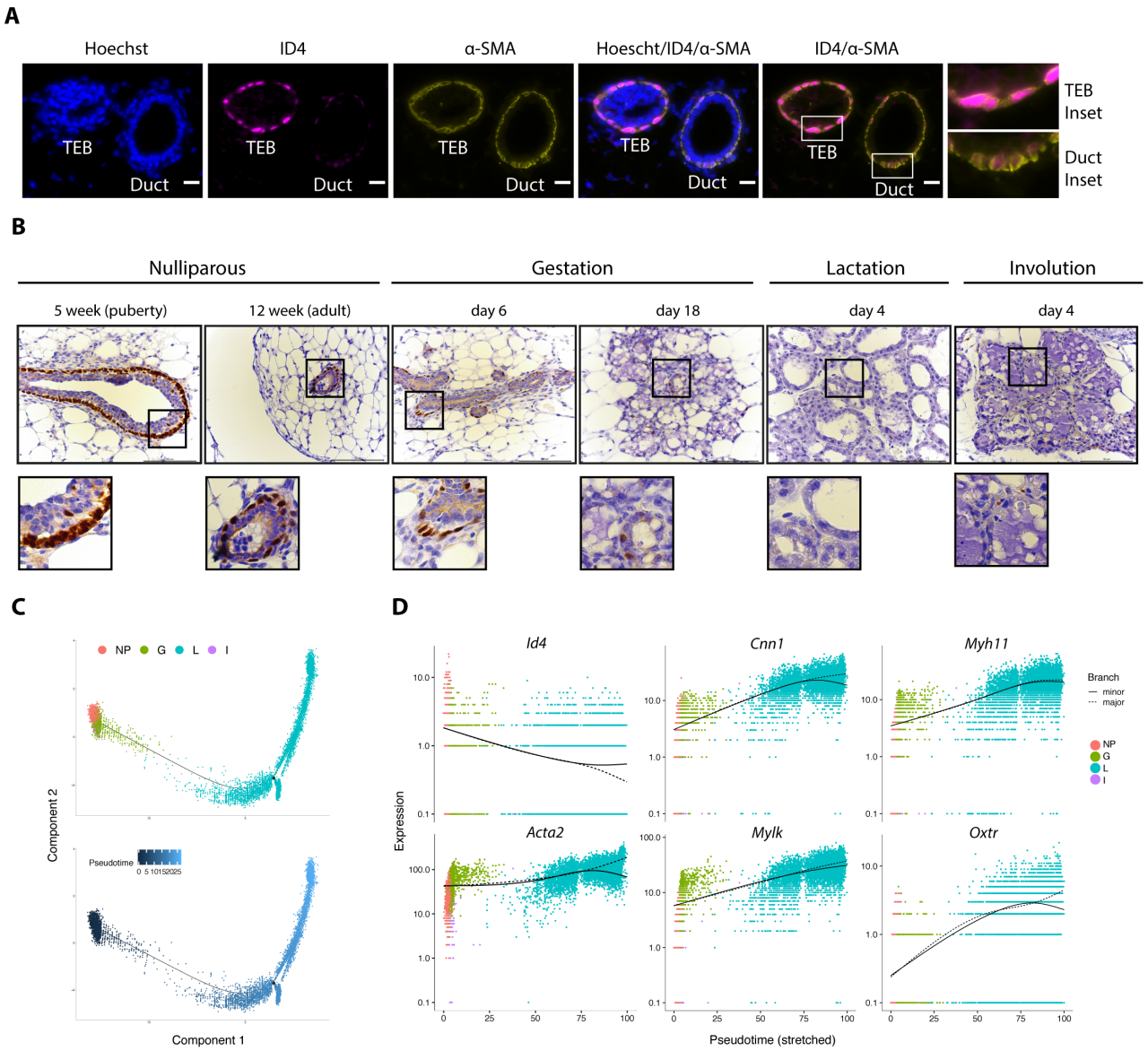

Figure S2. ID4 expression decreases in terminally differentiated myoepithelial cells, Related to Figure 2.

A) Co-immunofluorescent staining of ID4 and α-SMA in a section containing both a TEB and a duct from a pubertal (6 week) mammary gland. Scale bar = 20 µm**.** High power insets are shown. **B)** Mouse mammary glands at different developmental stages were stained by IHC for ID4. Scale bar = 100 µm. Representative images from 3 animals per stage shown. High power insets are displayed below images. **C)** Differentiation trajectory of basal cells from (Bach et al., 2017) coloured by developmental stage (upper) and pseudotime (lower). Low values (dark blue) represent undifferentiated cells. NP = Nulliparous (8 week), G= Gestation (Day 14.5), L = Lactation (Day 6), I = Involution (Day 11). **D)** Expression of *Id4, Cnn1, Myh11, Acta2, Mylk* and *Oxtr* as a function of pseudotime**.**

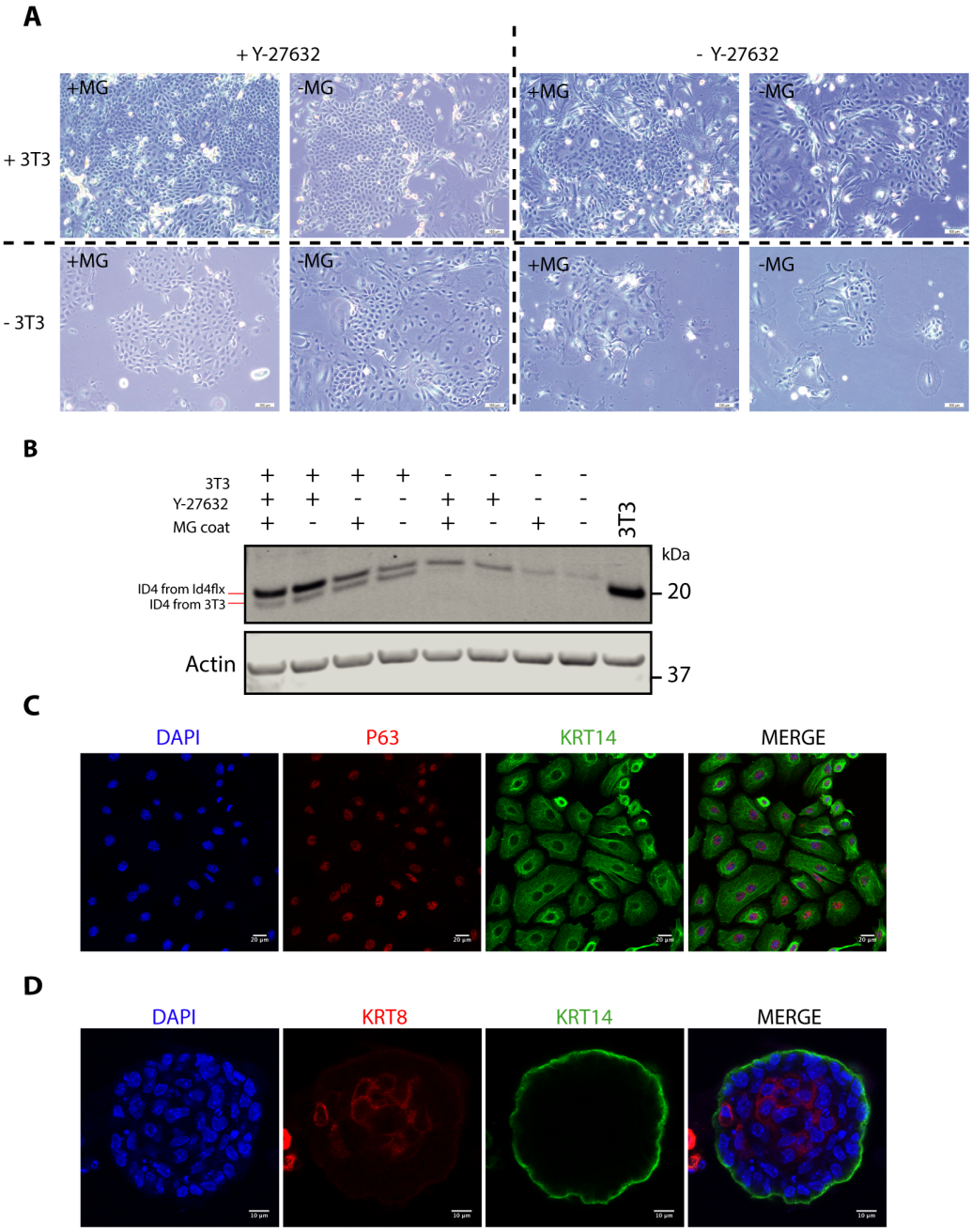

Figure S3. Establishment of 2D cultures and 3D organoids from primary basal cells, Related to Figure 3.

A) Phase contrast images of passage 11 conditionally reprogrammed cells grown in the presence and absence of ROCK inhibitor Y-27632, irradiated NIH-3T3 feeder cells, and Matrigel (MG) coated tissue culture flasks. Scale bar = 100 µm. **B)** ID4 western blot in cell lysates collected from cells shown in panel A. ID4 from basal cells runs at a higher molecular weight to ID4 expressed by NIH-3T3 cells. **C)** Co-immunofluorescent staining of primary basal cells with basal markers P63 and KRT14. Scale bar = 20 µm. **D)** Basal cells grown in 3D as organoids on top a plug of Matrigel. Confocal image of an organoid stained for luminal marker KRT8 and basal marker KRT14. Scale bar = 10 µm

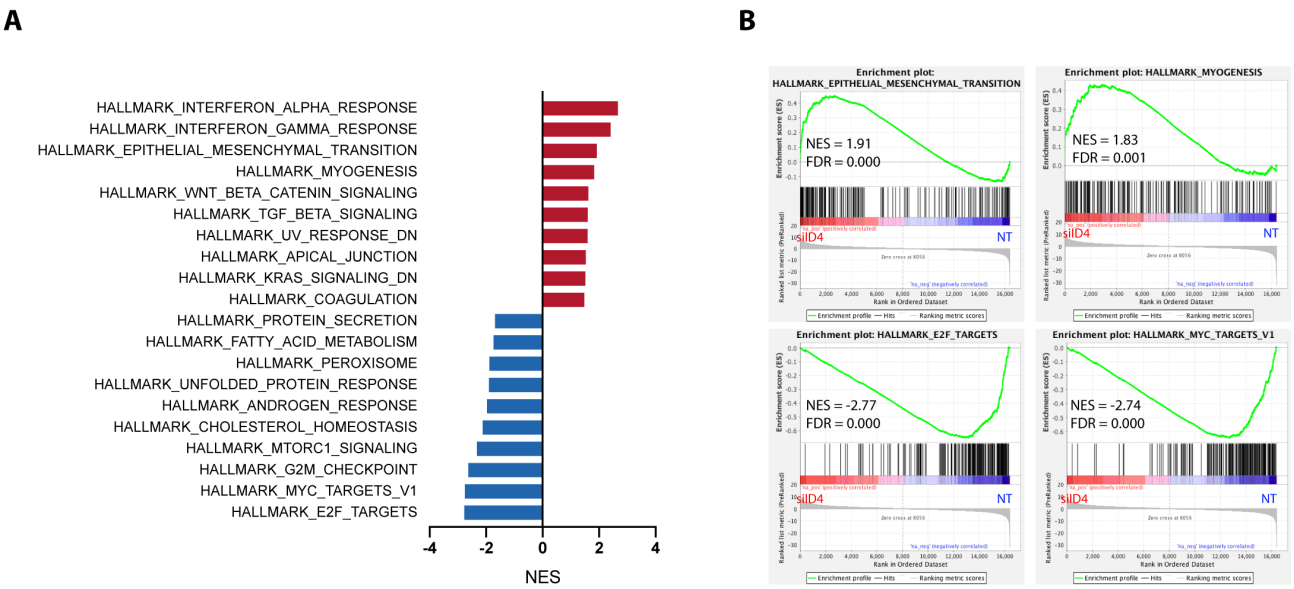

Figure S4. GSEA of ID4 knockdown Comma-Dβ cells, Related to Figure 4.

A) Genes were ranked based on the limma t-statistic comparing NT and siID4 cells and GSEA was carried out using the Hallmark gene sets. The top 10 positively (red) and negatively (blue) enriched pathways are displayed. **B)** Representative Hallmark GSEA enrichment plots displaying the profile of the running Enrichment Score (green) and positions of gene set members on the rank ordered list.

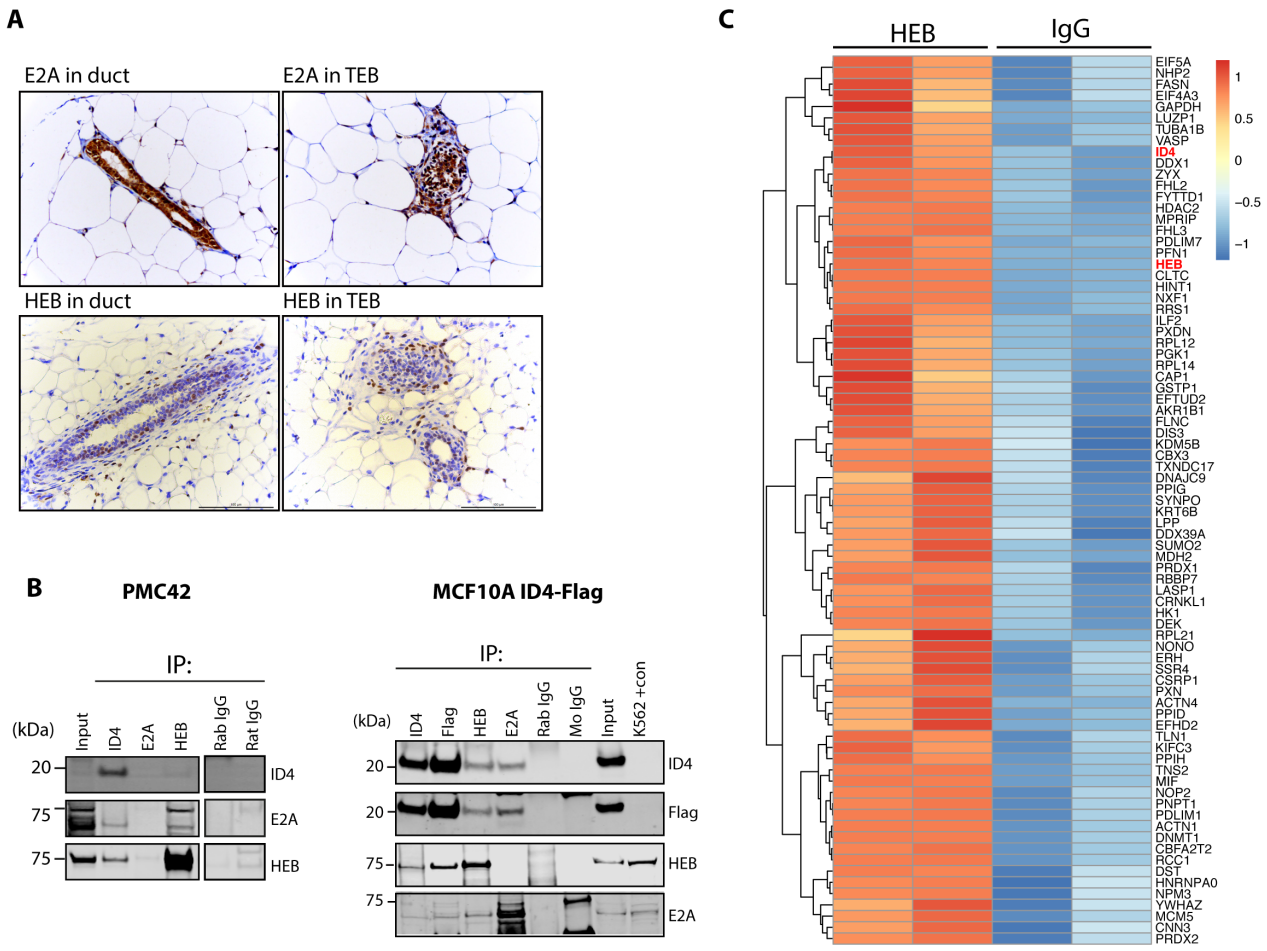

Figure S5. ID4 interacts with E-proteins E2A and HEB, Related to Figure 5.

A) IHC staining for E2A and HEB in murine mammary gland sections. **B)** Co-immunoprecipitation and western blotting of ID4, E2A and HEB from human normal-like mammary epithelial cell lines PMC42 (left) and MCF10A cells overexpressing Flag-tagged ID4 protein (right). **C)** Unsupervised hierarchical clustering heat map of SWATH RIME data from Comma-Dβ cells. Proteins with significantly higher abundance (p-value<0.05) in the HEB IPs compared to IgG IPs are displayed. Log2 protein area was used to generate the heatmap.

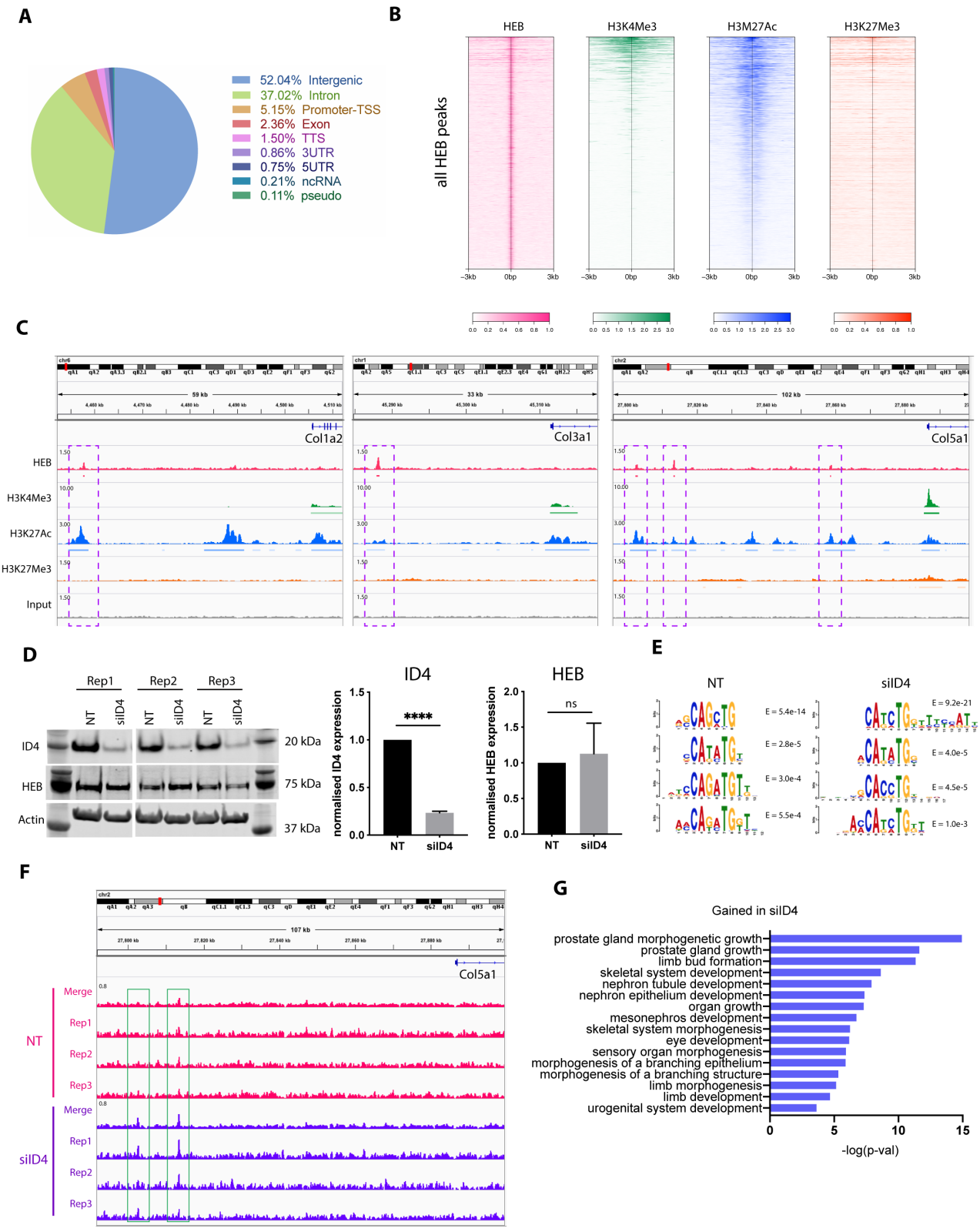

Figure S6. HEB directly binds to a subset of ID4 regulated genes and binding increases upon ID4 knockdown, Related to Figure 6.

A) Genomic distribution of HEB consensus peaks. TSS = transcriptional start site. TTS = transcriptional termination site. **B)** Heatmaps of HEB, H3K4Me3, H3K27Ac, and H3K27Me3 ChIP-seq signal at HEB-

bound regions. **C)** Examples of HEB and histone mark peaks occurring upstream of ID4 repressed genes *Col1a2, Col3a1*, and *Col5a1* (multiple HEB peaks) from the Integrative Genomics Viewer (IGV). Bars beneath peaks represent consensus MACS call (FDR<0.05) in at least 2 of 4 biological replicates. Input was used as a negative control. Purple boxes highlight HEB binding regions. Refseq genes shown in blue. Data scales for each track are indicated. **D)** Western blot analysis of ID4 and HEB expression in Comma-Dβ cells treated with NT or ID4-targeting siRNA. Irrelevant lanes were digitally removed indicated by the gap. Densitometry quantification of ID4 and HEB bands. Band intensity was normalised to β-Actin and expressed as fold change relative to NT control. N=3. Unpaired two-tailed students t-test. Error bars represent SEM. **** p<0.0001. ns = not significant. **E)** Top 4 enriched transcription factor binding motifs determined using MEME-ChIP for consensus HEB ChIP-seq peaks in Comma-Dβ cells treated with NT or ID4-targeting siRNA. E-values are displayed**. F)** Example of HEB peaks in control NT cells and siID4 cells upstream of *Col5a1* from the Integrative Genomics Viewer (IGV). Green boxes highlight HEB bound regions. Individual replicates and merged tracks are displayed. Refseq genes shown in blue. Data scales indicated. **G)** GREAT pathway analysis of 263 peaks increased in siID4 compared to NT from Fig. 6F. Top 16 Gene Ontologies (Biological process, cellular component, molecular function) are displayed.

Transparent Methods

Mice

All mice experiments were performed in accordance with the ethical regulations of the Garvan Institute Animal Experimentation Committee.

ID4 KO mice were generated as previously described (Yun et al. 2004).

ID4-FlagV5 mice were produced by the Mouse Engineering Garvan/ABR (MEGA) Facility using CRISPR/Cas9 gene targeting in C57BL/6J mouse embryos following established molecular and animal husbandry techniques (Yang et al. 2014). A single guide RNA (sgRNA) was produced based on a Cas9 target site that contained the TGA stop codon in exon 2 of *Id4* (TCTCTGCCGC**TGA**GCTGCGA*TGG*) (stop codon, bold underlined; protospacer-associated motif = PAM, italics underlined). sgRNA was microinjected into the nucleus and cytoplasm of C57BL/6J zygotes together with polyadenylated *S.pyogenes* Cas9 mRNA and a 150 base single-stranded, sense, deoxy-oligonucleotide homologous recombination substrate encoding for the insertion of a FLAG-V5 tag (DYKDDDDKGKPIPNPLLGLDST) immediately prior to the stop codon. A founder mouse heterozygous for the desired 66 bp insertion was identified by PCR amplification and Sanger DNA sequencing and the line maintained by backcrossing with inbred C57BL/6J mice.

The ID4floxGFP mice were generated as previously described (Best et al. 2014). All mice used were on the FVB/N background. For the gene expression profiling experiment, mice were synchronised in estrus to reduce hormone-induced gene expression variation (Dalal et al. 2001), and checked by vaginal swab cytology.

Mammary epithelial cell preparations

Mammary epithelial cells were prepared from freshly harvested 3rd and 4th mammary glands pooled from 4-8 female mice at the indicated ages. Glands were mechanically disrupted using a McIlwain tissue chopper then digested with 15,000 U collagenase (Sigma-Aldrich C9891) and 500 U hyaluronidase (Sigma-Aldrich H3506) in FV media (DMEM/F12 (Gibco), 5% (v/v) FBS (HyClone) 10 mM HEPES (Gibco), 0.14 IU/mL Insulin (Novo Nordisk), 500 ng/mL Hydrocortisone (Sigma-Aldrich), 20 ng/mL Cholera Toxin (Sigma-Aldrich), 2 ng/mL mEGF (Life Technologies)). Digestion was carried out in a shaking incubator for 1 hr at 37°C 225 rpm. The resulting organoids were further digested with warm 0.25% trypsin for 1 min with constant pipetting followed by treatment with 5 mg/mL Dispase (Roche 165859) for 5 min in at 37°C. Cells were incubated with 1X red blood cell lysis buffer (BD Biosciences) for 5 min at room temperature. Cells were passed through a 70 µm strainer followed by a 40 µm strainer to remove clumps and debris. Live cells were counted using trypan blue (Thermo Fisher Scientific) and a haemocytometer prior to use in downstream applications.

Flow cytometry and FACS

Single-cell suspensions of mouse mammary epithelial cells were blocked in Fc block cocktail (FACS buffer, 6.25 µg/mL Mouse BD Fc block (BD Bioscience) and 200 µg/mL Rat Gamma Globulin (Jackson ImmunoResearch)) for 10 min on ice. Fluorophore- or biotin-conjugated antibodies were diluted in Fc block cocktail and incubated on the cells on ice for 20 min in the dark. Antibodies used were CD24-PE (1:200, BD Biosciences, Clone M1/69), EPCAM-PerCP/Cy5.5 (1:200, Biolegend, Clone G8.8), CD29-APC/Cy7 (BioLegend, Clone HMβ1-1), CD61-APC (1:50, Thermo Fisher Scientific, Clone HMβ3-1), CD49f-APC (1:100, BioLegend, Clone GoH3), Ter119-biotin (1:80, BD Biosciences, Clone TER119), CD45-biotin (1:100, BD Biosciences, Clone 30-F11), CD31-biotin (1:40, Biolegend, Clone 390) and BP1-biotin (1:500, Thermo Fisher Scientific, Clone 6C3). Cells were washed and stained with streptavidin-BV421 (1:100, Biolegend). Following washing, cells were resuspended in FACS buffer at a density of

$1x10^7$ cells/mL and DAPI was added (1:1000, Invitrogen). FACS was performed on a FACS Aria III 4 laser 15 colour sorter with BD FACS DIVA software. For analytical flow cytometry, a CytoFLEX 3 laser 13 colour flow cytometer (Beckman Coulter) was used. All flow cytometry data was analysed using FlowJo software version 10 (Tree Star Inc).

RNA-sequencing

For the FACS-enriched mammary subpopulation RNA-seq experiment, cells were sorted directly into 700 Qiazol lysis reagent to minimise cell loss, with a maximum of $5x10^4$ cells per 1.5 mL tube and RNA was extracted using the Qiagen miRNeasy micro kit (Qiagen). Four biological replicates were performed. RNA was extracted from Comma-Dβ cells using the Qiagen miRNeasy mini kit (Qiagen) in biological triplicate. The Qubit RNA BR Assay kit (Thermo Fisher Scientific) was used to measure RNA concentration and RNA integrity was determined using the Agilent Bioanalyser 2100 with the 6000 Nano Assay (Agilent Technologies).

For low input RNA extracted from FACS sorted cells, the Ovation RNA-seq System V2 kit (NuGEN) was used to synthesise cDNA with RNA inputs ranging from 0.5-2 ng. The Ovation Ultralow System V2 kit was then used to prepare libraries from the cDNA. For the Comma-Dβ experiment the Illumina TruSeq Stranded mRNA Library Prep Kit (Illumina) was used with 1 μg of input RNA.

cDNA libraries from were sequenced on a HiSeq 2500 system (high output mode) (Illumina), with 125 bp paired-end reads for the FACS experiment or the NextSeq system (Illumina), with 75 bp paired-end reads for the Comma-Dβ experiment.

Quality control was checked using FastQC (Andrews 2010) to remove poor quality reads. Illumina sequencing adapters were then trimmed using Cutadapt (Martin 2011). Reads were then aligned to the mouse reference genome mm10 using STAR ultrafast universal RNA-seq aligner (Dobin et al. 2013). RSEM accurate transcript quantification for RNA-seq data was used to generate a gene read

count table and to filter genes with read counts of 0 (Li and Dewey 2011).

Differential gene expression analysis was performed using edgeR (Robinson et al. 2010; McCarthy et al. 2012) and voom/limma (Ritchie et al. 2015) R packages. Genes were ranked based on the limma moderated t-statistic and this was used as input GSEA pre-ranked (Subramanian et al. 2005) using Molecular Signature Database (MSigDB) collections (v6.2). The EnrichmentMap Cytoscape plugin (Merico et al. 2010) was used to visualise the GSEA results.

Analysis of public single-cell RNA-seq data

Raw Unique Molecular Identifier (UMI) expression data was taken from Bach *et al.* and re-processed using Seurat v2 using default parameters, as recommended by the developers (Satija et al. 2015). All basal cell clusters, defined by *Krt5* and *Krt14* expression, were extracted for subsequent analysis. Developmental trajectories were inferred using Monocle 2 (Qiu et al. 2017). Clustering was first performed using the 'densityPeak' and 'DDRTree' methods, with a delta local distance threshold of 2, and a rho local density threshold of 3. For inferring pseudotime, we first selected all genes with a mean expression greater than 0.5, and an empirical dispersion greater than 1, and proceeded with the top 1000 significant genes for ordering. Differential gene expression between the two branch states were performed using the MAST algorithm through the Seurat package (Finak et al. 2015; Satija et al. 2015). For comparisons between *Id4*-high and *Id4*-low states, we first filtered for cells with the detection of *Id4* to avoid technical biases from gene drop out. Differential gene expression was performed (as described above) between the top and bottom 200 cells grouped based on log normalised gene expression of *Id4*.

Immunohistochemistry and immunofluorescence on mammary tissue sections

IHC staining was performed on 4 μm sections of formalin-fixed paraffin-embedded (FFPE) tissue.

Slides were dewaxed with xylene and hydrated through graded alcohols. Antigen retrieval was performed for 1 min in a pressure cooker in DAKO target retrieval reagent s1699. A DAKO Autostainer was used for subsequent steps. Briefly, slides were incubated with DAKO peroxide block for 5 min and blocked with DAKO protein block for 30 min. Primary antibody (ID4 1:400 Biocheck BCH9/82-12, E2A 1:100 abcam 69999, HEB ProteinTech 21073-1-AP 1:250) was incubated on the slides for 60 min. Following washing, slides were incubated with Envision rabbit secondary antibody (Agilent Technologies, Santa Clara, CA, USA) for 30 min. Slides were washed then incubated with DAKO DAB+ (Agilent Technologies) reagent for 10 min. Slides were then rinsed and counterstained with haematoxylin for 20-30 sec and dehydrated through graded alcohols, cleared using xylene and mounted using Ultramount #4 (Fronine). Bright-field images were captured using a Leica DM 4000 microscope with high-resolution colour camera (DFC450).

Immunofluorescence was performed manually on antigen retrieved tissue sections prepared as described for IHC. Slides were blocked with Mouse on Mouse (MOM) blocking buffer (Vector Biolabs) for 1 hour. Primary antibody was diluted in MOM diluent and incubated overnight at 4°C. Antibodies used were ID4 (1:400, Biocheck BCH9/82-12), α-SMA (1:200, Sigma-Aldrich A5228), HEB (1:200, Protein Tech 14419-1-AP), V5 (1:200, Santa Cruz sc-58052). Following washing, slides were incubated with fluorescent secondary antibody (1:500, Jackson ImmunoResearch) for 1 hr at room temperature. Nuclei were stained using Hoechst 33342 (Sigma-Aldrich). Slides were mounted using Prolong Diamond mounting media (Thermo Fisher Scientific). Fluorescent images were captured using a Leica DM 5500 microscope.

For quantification of fluorescence of ID4 and α-SMA in cap and ductal basal cells, 3 representative ducts and TEBs were imaged from each mouse (n=9). Using FIJI software, a region of interest was drawn around 10 randomly selected basal cells from each image. The mean fluorescence of ID4 and α-SMA was determined within each cell.

Picrosirius red staining

Picrosirius red staining was performed following as per (Vennin et al. 2017; Vennin et al. 2019). Briefly, dewaxed FFPE sections were stained with haematoxylin. Sections were then treated with 0.2% Phosphomolybdic acid (Sigma-Aldrich) followed by 0.1% Sirius red (Sigma-Aldrich) Picric Acid-Saturated Solution (Sigma-Aldrich). Slides were then rinsed with acidified water (90 mM glacial acetic acid) then with 70% ethanol. A Leica DM 4000 microscope was used to image total collagen staining in the tissue sections by phase contrast. To image birefringence of collagen fibres, 2 polarised filters were used. Matched phase-contrast and polarised images were taken for each region of interest. Image analysis was performed using FIJI image analysis software.

Proximity Ligation Assay

Comma-Dβ cells were grown on glass coverslips in 6 well plates until ~80% confluent. Cells were fixed in 4% Paraformaldehyde (PFA) (ProSciTech) for 15 min then permeablised in 0.2% Triton-X-100 for 15 min. The Duolink Proximity Ligation Assay (PLA) (Sigma-Aldrich) was performed as per the manufacturer's protocol and using the following antibodies: ID4 (1:50, Santa Cruz sc-365656) and HEB (1:200, Protein Tech 14419-1-AP), and equivalent concentrations of IgG negative controls (sc-2027 and sc-2025). DAPI (Life Technologies) and Phalloidin (Life Technologies) were added to the final wash step. Coverslips were mounted onto glass slides with Prolong Diamond mountant (Thermo Fisher Scientific).

Cells were imaged using a Leica DMI Sp8 confocal microscope (63X oil objective). Six random fields of view per image were captured, each with approximately 50 cells, and images were quantified using Andy's algorithms FIJI package (Law et al. 2017) to enumerate the number of PLA foci per nuclei.

Cell lines

The mouse mammary epithelial cell line Comma-Dβ was a gift was a gift from Joseph Jeffery (University of Massachusetts, Amherst, MA, USA). Comma-Dβ cells were maintained in DMEM/F12 media (Gibco) supplemented with 2% FBS (HyClone), 10 mM HEPES (Gibco), 0.125 IU/mL Insulin (Novo Nordisk) and 5 ng/mL mEGF (Life Technologies). The human mammary epithelial cell line PMC42 was a gift from Professor Leigh Ackland (Deakin University, Melbourne, Victoria, Australia). PMC42 cells were maintained in RPMI 1640 (Gibco) supplemented with 10% FBS (HyClone). The MCF10A cell line was obtained from the American Type Culture Collection and were maintained in DMEM/F12 (Gibco), 5% Horse Serum (Thermo Fisher Scientific), 20 ng/mL hEGF (In Vitro Technologies), 0.5 mg/mL Hydrocortison (Sigma-Aldrich), 100 ng/mL Cholera Toxin (Sigma-Aldrich) and 0.125 IU/mL Insulin (Novo Nordisk).

Overexpression of ID4

Comma-Dβ cells ($1.1x10^5$) were seeded into a 6-well plate. 16-24 hours later the cells were infected with pMSCV-Id4-DSred or pMSCV-DSred retrovirus diluted 1:10 in Comma-D media with 8 µg/ml polybrene. 24 hours later the media was changed. DSred positive cells were then FACS enriched using the BD FACSAria fluorescence activated cell sorter and BD FACSDIVA software.

Conditional reprogramming of primary mouse basal cells

Viable basal cells were purified by FACS from 10-12 week-old female mice as described above. Cells were collected into FAD media (DMEM/F12 3:1 (Gibco) supplemented with 10% FBS, 0.18 mM Adenine (Sigma-Aldrich), 500 ng/mL Hydrocortisone (Sigma-Aldrich), 8.5 ng/mL Cholera Toxin (Sigma-Aldrich), 10 ng/mL mEGF (Life Technologies), 0.14 IU/mL Insulin (Novo Nordisk), 5 µM Y-27632 (Seleckchem), 1X AB/AM (Gibco) and 50 µg/mL Gentamicin (Life Technologies). Basal cells

were maintained in culture using as per (Prater et al. 2014). Tissue culture flasks coated with Growth Factor-reduced Matrigel (Corning) diluted 1:60 in PBS for 30 min to 1 hr at 37°C. Excess Matrigel/PBS solution was aspirated from the flasks and basal cells were seeded at a density of ~5000 cells/cm^2. Basal cells were co-cultured with irradiated (50 Gy) NIH-3T3 feeder cells at a density of 10,000 cells/cm^2. Cells were maintained in a 37°C low 5% oxygen 5% CO_2 incubator. Differential trypsinisation was used to first remove the less-adherent 3T3 cells when passaging.

Cre mediated deletion of ID4 from primary cells

1.0-1.5x10^5 ID4floxGFP primary basal cells were seeded into a T75 with 7.5x10^5 irradiated 3T3 cells in 20 mL FAD media. Adenovirus was added to the media at a multiplicity of infection (MOI) of 100. Adenoviruses' used were codon optimised Cre (iCre) and GFP adenovirus (Vector Biolabs 1772) and control eGFP adenovirus (Vector Biolabs 1060). Infection efficiency was assessed the following day using a fluorescence microscope to check GFP expression within the cells. Cells were harvested after 72 hr for downstream experiments and analysis.

Organoid culture

A 40 µL plug of Matrigel (Corning) was added to each well of 8-well chamber slides (Corning) and allowed to set for 30 min in a 37°C incubator. Primary cells were resuspended in Epicult-B (Stemcell Technologies) media containing 2% Matrigel and 12,000 cells were seeded into each chamber in a total volume of 400 µL. Chambers were observed using a light microscope to ensure that cells were not aggregated. Organoids were allowed to form over 1 week and media was changed every 3 days. Organoids from each chamber were photographed using an inverted epifluorescence microscope (4x objective) and the average organoid area per image was determined using Andy's algorithms FIJI package (Law et al. 2017).

Immunofluorescent staining of organoids

Organoids were stained within the chamber slides. Media was aspirated and organoids were fixed with 2% PFA diluted in PBS for 20 min at room temperature. Organoids rinsed with PBS for 5 min following permeabilisation with 0.5% Triton X-100 in PBS for 10 min at 4°C then rinsed 3 times with 100 mM Glycine (Astral Scientific) in PBS for 10 min each. Organoids were blocked for 1 hr in IF buffer with 10% goat serum (Vector Laboratories). Primary antibody made up in blocking buffer was incubated in the chambers overnight at 4°C in a humidified chamber. Antibodies used were ID4 (1:200, Biocheck BCH9/82-12), α-SMA (1:100, Abcam ab5694), KRT14 (1:1000, Covance PRB-155P), KRT8 (1:500, DSHB TROMA1) and P63 (1:100, Novus NB100-691). The following day, slides were equilibrated to room temperature for 1-2 hr. Organoids were rinsed 2 times with IF buffer (0.1% BSA, 0.2% Triton-X-100, 0.05% Tween-20 in PBS) for 20 min each. Fluorescent secondary antibody (Jackson ImmunoResearch) diluted in blocking buffer (1:500) was added to the chambers and incubated for 45 min in the dark. Secondary antibody was rinsed off for 20 min in IF buffer, then 2 times with PBS for 10 min each. DAPI (Life Technologies) was incubated in the chambers for 15 min followed by a 10 min PBS rinse. The walls of the chamber slides were removed and one drop of Prolong Diamond mounting media (Thermo Fisher Scientific) was added to each well. Slides were coverslipped and edges sealed with clear nail varnish. Slides were allowed to dry for 24 hr at room temperature protected from light.

Organoids were imaged using a Leica DMI Sp8 confocal microscope using the 40X oil objective and 3X optical zoom. The FIJI software package was used to quantify the fluorescence within the organoids. For α-SMA, a circle was drawn around the entire organoid and fluorescence measured. For ID4, a mask was made using the DAPI channel and signal was measured within the nuclear mask. Multiple representative regions of background staining were quantified. Corrected fluorescence (CF)

was calculated as described in (McCloy et al. 2014). CF = integrated density – (area of selected organoid * mean fluorescence of all background readings)

siRNA transfections

Comma-Dβ cells were seeded at a density of $1.5x10^4$ cells/cm^2 into 6 well plates (Corning) for RNA-seq or 100 mm dishes (Corning) for ChIP-seq in antibiotic-free media. The following day siRNA constructs (Dharmacon) were transfected into the cells using Dharmafect-4 transfection reagent (Dharmacon) as per the manufacturer's instructions at 20 nM. siRNA used were siGENOME Mouse Id4 SMARTpool (M-043687) and ON-TARGETplus Non-targeting Control Pool (D-001810). Media was changed the following day and cells were harvested 48 hr post-transfection.

Western blotting

Cells were lysed in RIPA buffer (50 mM Tris-HCl pH 7.4, 1% NP-40, 0.5% Sodium Deoxycholate, 0.1% SDS, 1% Glycerol, 137.5 mM NaCl, 100 µM Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche)) and protein concentration was quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. Protein lysates (15-30 µg protein) were mixed with 1X NuPage loading buffer (Life Technologies) and 1X NuPage reducing agent (Life Technologies) and denatured by heating at 85°C for 5 min. Samples were run on a 4-12% Bis/Tris gels (Life Technologies) in MES or MOPS running buffer (Life Technologies). Protein was transferred to a 0.45 µm PVDF membrane (Merck Millipore) using BioRad transfer modules in transfer buffer (25 mM Tris, 192 mM Glycine, 20% Methanol). Membranes were blocked in Odyssey blocking buffer (LiCOR) for 1 hr at room temperature then incubated in primary antibody diluted in 5% BSA/TBS overnight at 4°C. The following antibodies were used for western blotting: ID4 (1:20,000, Biocheck BCH9/82-12), E2A (1:1000, from Dr Nicolas Huntington from the Walter Eliza Hall

Institute (WEHI)), HEB (1:1000, from Dr Nicolas Huntington, WEHI or 1:1000 ProteinTech 144191-1-AP), Flag (1:5000, Sigma-Aldrich F1804), V5 (1:200, sc-58052), α-SMA (1:1000, Abcam ab5694), SNAIL (Cell Signalling Technology 3879), SLUG (Cell Signalling Technology 9585), CNN2 (ProteinTech 21073-1-AP), ZEB1 (Cell Signalling Technology 3396), β-Actin (1:1000, Sigma-Aldrich A5441) and α-Tubulin (1:1000, Santa Cruz sc-5286). Fluorescent secondary antibody conjugated to IRDye680 or IRDye800 (LiCOR) diluted in Odyssey blocking buffer (1:15,000 - 1:20,000) were used for detection. An Odyssey CLx Infrared Imaging System (LiCOR) was used to image and quantify the western blots.

Co-immunoprecipitation

Pierce Protein A/G magnetic breads (Thermo Fisher Scientific) were incubated with antibody for a minimum of 4 hr at 4°C on a rotating platform. For each IP 15 µL of beads and 2.5 µg antibody were used. Immunoprecipitating antibodies used were ID4 (pool of sc-491 and sc-13047), E2A (from Dr Nicolas Huntington, WEHI), HEB (Santa Cruz sc-357 and antibody from Dr Nicolas Huntington, WEHI), Flag (Sigma-Aldrich F1804), V5 (Santa Cruz sc-58052 and Thermo Fisher R960-25), species matched IgG negative controls (sc-2027, sc-2025 and BioLegend 400602). Protein was extracted in IP lysis buffer (10% Glycerol, 0.03% $MgCl_2$, 1.2% HEPES, 1% Sodium acid pyrophosphate, 1% Triton-X, 0.8% NaCl, 0.4% NaF, 0.04% EGTA, 1x cOmplete ULTRA Tablet (Roche), 100 µM Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and cells were passed 5 times through a 23 g needle to aid in lysis. Protein lysates (0.5-2 mg per IP) were added to washed beads and incubated overnight at 4°C on a rotating platform. Following washing, beads were resuspended in 2X NuPage loading buffer (Life Technologies) and 2X reducing agent (Life Technologies) in IP lysis buffer. Samples were incubated at 85°C for 5 min. Beads were separated on a magnetic rack and supernatant loaded onto NuPage gel for SDS-PAGE and western blotting. For detection of protein, fluorescent TrueBlot secondary antibodies were used (Jomar Life Research).

RIME

The RIME protocol was adapted from the protocol developed by Mohammed *et al.* (Mohammed et al. 2013). Comma-Dβ cells were grown in 150 mm tissue culture dishes (Corning) until 80-90% confluent. A total of 8x 150 mm dishes were used per RIME sample. Cells were fixed with 1% PFA (ProSciTech) in DMEM/F12 (Gibco) for 7 min at room temperature on a rocking platform. Cross-linking was quenched by addition of molecular grade glycine (Astral Scientific) to a final concentration of 125 mM for 2 min on a rocking platform. Cells were washed twice with ice-cold PBS (Gibco) and scraped in 1 mL of ice-cold PBS containing magnesium and calcium salts (Gibco). Cells were centrifuged at 4°C for 5 min at 1200 rpm. Cell nuclei were enriched through a series of lysis buffers. Cells were first resuspended in 10 mL lysis buffer 1 (50 mM HEPES-KOH pH 7.5, 140 mM Sodium, 1 mM EDTA, 10% Glycerol, 0.5% NP-40 or Igepal CA-630, 0.25% Triton X-100, 100 µM Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche)) and incubated for 30 min at 4°C on a rotating platform. Cells were then centrifuged at 3000 g for 3 min at 4°C and supernatant removed. The pellet was then resuspended in 10 mL lysis buffer 2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 100 µM Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche)) and incubated for a further 30 min at 4°C on a rotating platform. Following centrifugation, the resulting nuclei pellet was lysed in 2.5 mL lysis buffer 3 (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Sodium Deoxycholate, 0.5% N-lauroylsarcosine, 100 µM Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche)). The nuclear lysate was sonicated using a Bioruptor sonicator (Diagenode) for 20 cycles of 30 sec on/30 sec off. Triton-X-100 was added to a final concentration of 1% and the sample was then centrifuged at maximum speed for 5 min at 4°C to remove cellular debris. Immunoprecipitation was carried out from the sheared nuclear supernatant described above with

20 µg of antibody (ID4 sc-491 and sc-13047 pool, HEB sc-357) and 100 µL of magnetic beads per sample. The following day, the beads were washed 10 times with RIPA buffer then 5 times with 100 mM ammonium hydrogen carbonate (AMBIC) (Sigma-Aldrich) solution to remove salts and detergents, resuspended in 50 µL AMBIC solution and transferred into clean tubes.

Mass Spectrometry

Samples were processed as described in (Huang et al. 2015) for liquid chromatography coupled mass spectrometry (LC-MS/MS) using Sequential Windowed Acquisition of all THeoretical fragment-ion spectra (SWATH) acquisition. Briefly, samples were denatured in 100 mM triethylammonium bicarbonate and 1% sodium deoxycholate, disulfide bonds were reduced in 10 mM DTT, alkylated in 20 mM iodo acetamide, and proteins digested on the magnetic beads using trypsin. After C18 reversed phase (RP) StageTip sample clean up, peptides were analysed using a TripleToF 6600 mass spectrometer (SCIEX, MA, USA) coupled to a nanoLC Ultra 2D HPLC system (SCIEX). Peptides were separated for 60 min using a 15 cm chip column (ChromXP C18, 3 µm, 120 Å) (SCIEX) with an acetonitrile gradient from 3-35%. The MS was operated in positive ion mode using either a data dependent acquisition method (DDA) or SWATH acquisition mode. DDA was performed of the top 20 most intense precursors with charge stages from 2+ - 4+ with a dynamic exclusion of 30 s. SWATH-MS was acquired using 100 variable size precursor windows. DDA files were searched using ProteinPilot 5.0 (SCIEX) against the reviewed UniProt *Mus musculus* protein database (release February 2016) using an unused score of 0.05 with decoy search strategy enabled. These search outputs were used to generate a spectral library for targeted information extraction from SWATH-MS data files using PeakView v2.1 with SWATH MicroApp v2.0 (SCIEX) importing only peptides with < 1% FDR. Protein areas, summed chromatographic area under the curve of peptides with extraction FDR ≤ 1%, were calculated and used to compare protein abundances between bait and control IPs.

ChIP-seq

ChIP-seq was adapted from (Khoury et al. 2020). Comma-Dβ cells were grown in 150 mm or 100 mm dishes (Corning) until 80-90% confluent. For the ChIP-seq experiment on unperturbed cells 2x 150 mm dishes were used for HEB ChIP and 1x 150 mm dishes were used for histone marks. Four independent replicates were conducted. For the ID4 siRNA experiment 3x 100 mm dishes per condition were used and the experiment was repeated 3 times. Cells were scraped in ice-cold PBS (Gibco) and resuspended by passing 10 times through a 19 g syringe. Cells were fixed in 1% PFA (ProSciTech) in PBS (Gibco) for 15 min at room temperature followed by quenching with glycine to a final concentration of 125 mM for 5 minutes followed by two PBS washes. Nuclei was extracted by resuspending cells in nuclei extraction buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 0.5% IGEPAL, 1x cOmplete ULTRA Tablet (Roche)) and incubated on ice for 10 min followed by 20 passes on a tight Dounce homogeniser, or until nuclei were extracted. Nuclei were visually inspected using trypan blue and a haemocytometer. Nuclei were pelleted by centrifugation and resuspended in sonication buffer (50 mM Tris-HCl pH 8, 1% SDS, 10 mM EDTA, 1x cOmplete ULTRA Tablet (Roche)) and sonicated using a Bioruptor sonicator (Diagenode) to achieve DNA fragments between 100-500 bp with a mean fragment size of approximately 300 bp. In the siRNA experiment, protein was quantified using the Pierce BCA assay to load equal amounts of input material for NT and ID4 siRNA for immunoprecipitation. Sonicated sample was diluted with IP dilution buffer to 1 mL (16.7 mM Tris-HCl pH 8, 0.01% SDS, 1% Triton X-100, 167 mM NaCl, 1.2 mM EDTA). Samples were cleared with protein A/G magnetic beads for 1.5 hr at 4°C. 1% of the cleared nuclear lysate was removed for the input control and immunoprecipitation was performed as described above. For HEB ChIP 100 µL of beads and 20 µg of HEB antibody (sc-58052) was used with chromatin from approximately 20-30x10^6 cells. For histone mark ChIP, 50 µL beads with 10 µg of the following

antibodies – H3K4Me3 (Active Motif 28431), H3K27Me3 (Merck Millipore 07-449), was used with chromatin from approximately 10-15x10^6 cells. The following day, beads were washed for 5 min each with 1 mL of the following buffers: Low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8, 150 mM NaCl), High salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8, 500 mM NaCl), LiCl buffer (0.25 M LiCl, 1% IGEPAL, 1% deoxycholic acid (sodium salt), 1 mM EDTA, 10 mM Tris, pH 8) followed by 2 washes with TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8). DNA was eluted twice in 100 uL ChIP elution buffer (1% SDS, 0.1 M Sodium Biocarbonate) for 15 minutes at room temperature each. Cross-linking was reversed overnight by treating samples with 200 mM NaCl and 250 µg/mL Proteinase K (New England Biolabs) and incubating overnight at 65°C. The following day samples were treated with 100 µg/mL RNAse A (Qiagen) for 1 hr at 37°C. DNA was purified using Phase Lock Gel Light tubes (Quantabio 5Prime) according to the manufacturer's instructions. DNA was eluted in 20 uL nuclease-free TE buffer, pH8 (Qiagen).

DNA concentration was measured using the Qubit HS Assay Kit (Thermo Fisher Scientific). Libraries were prepared using the Illumina TruSeq ChIP library prep kit (Illumina) following the manufacturer's instructions except the gel purification step was replaced with a two-sided AMPure XP bead (Beckman Coulter) size selection to obtain libraries between 200-500 bp. Library sizes were verified using the 4200 TapeStation System (Agilent) with a D1000 ScreenTape (Agilent) and concentration was determined using the Qubit HS Assay Kit (Thermo Fisher Scientific). ChIP libraries were sequenced on the NextSeq system (Illumina), with 75 bp paired-end reads.

Reads were aligned with BWA (Li and Durbin 2009) and all reads with a MAPQ<15 were removed. Alignment statistics were generated using Samtools Flagstat (Li et al. 2009). ChIP-seq peaks were called using the peak calling algorithm MACS (Zhang et al. 2008) and ENCODE blacklist regions were removed. For unperturbed cells, consensus peaks present in at least 2 of 4 replicates were used for downstream analysis. Due to lower amount of input material available from the siRNA experiment,

less consensus peaks were called and for this reason analysis was conducted on merged peaks from the 3 replicates. Motif enrichment analysis was performed using MEME-ChIP (Machanick and Bailey 2011). Peaks were annotated to genomic features using HOMER (Heinz et al. 2010). GREAT was used for functional enrichment analysis (v4.0.4) and gene annotation using the default parameters (McLean et al. 2010). SeqPlots (Stempor and Ahringer 2016) and IGV (Robinson et al. 2011) software were used for data visualisation. Differential binding analysis was performed using the DiffBind package (Stark 2011). For overlapping of siID4 RNA-seq and HEB ChIP-seq genes, a hypergeometric test was used to determine if overlap was significant, assuming 30,000 genes in the mouse genome.