Inhibitor of Differentiation 4 (ID4) represses myoepithelial differentiation of mammary stem cells through its interaction with HEB

Holly Holliday[1,2], Daniel Roden[1,2], Simon Junankar[1,2], Sunny Z. Wu[1,2], Laura A. Baker[1,2], Christoph Krisp[3,4], Chia-Ling Chan[1], Andrea McFarland[1], Joanna N. Skhinas[1], Thomas R. Cox[1,2], Bhupinder Pal[5], Nicholas Huntington[6], Christopher J. Ormandy[1,2], Jason S. Carroll[7], Jane Visvader[5], Mark P. Molloy[3], Alexander Swarbrick[1,2]

1. The Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
2. St Vincent's Clinical School, Faculty of Medicine, UNSW Sydney, Sydney, NSW 2010, Australia
3. Australian Proteome Analysis Facility, Macquarie University, Sydney, NSW 2109, Australia
4. Institute of Clinical Chemistry and Laboratory Medicine, Mass Spectrometric Proteomics, University Medical Center Hamburg-Eppendorf, Hamburg 20251, Germany
5. ACRF Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
6. Biomedicine Discovery Institute, Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3168, Australia
7. Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK

Keywords

Differentiation / Epithelium / Mammary gland / Myoepithelial / Transcription factor

Abstract

Differentiation of stem cells embedded within the mammary epithelium is orchestrated by lineage-specifying transcription factors. Unlike the well-defined luminal hierarchy, dissection of the basal lineage has been hindered by a lack of specific markers. Inhibitor of Differentiation 4 (ID4) is a basally-restricted helix-loop-helix (HLH) transcription factor essential for mammary development. Here we show that ID4 is highly expressed in basal stem cells and decreases during myoepithelial differentiation. By integrating transcriptomic, proteomic, and ChIP-sequencing data, we reveal that ID4 is required to suppress myoepithelial gene expression and cell fate. We identify the bHLH protein HEB as a direct binding partner of ID4, and describe a previously-unknown role for this regulator in mammary development. HEB binds to E-boxes in regulatory elements of developmental genes, negatively regulated by ID4, involved in extracellular matrix synthesis and contraction. Together our findings support a model whereby HEB promotes myoepithelial specialisation by *cis*-transcriptional regulation, which is opposed by biochemical interaction with ID4. These new insights expand our current understanding into control of myoepithelial differentiation and mammary gland morphogenesis.

Introduction

The mammary gland undergoes marked tissue remodelling throughout life. During murine pubertal development, terminal end buds (TEBS) located at the tips of the

ducts invade into the surrounding stromal fat pad (Macias & Hinck, 2012, Williams & Daniel, 1983). This process is driven by collective migration and rapid proliferation of outer cap cells, which surround multiple layers of inner body cells (Williams & Daniel, 1983). As the ducts elongate, the cap cells differentiate into the basal cell layer and the body cells adjacent to the basal cells give rise to the luminal cell layer (Williams & Daniel, 1983), while the innermost body cells undergo apoptosis to sculpt the bilayered ductal tree present in the adult gland (Paine & Lewis, 2017). During pregnancy the gland undergoes alveolar morphogenesis in preparation for production and secretion of milk at lactation. The luminal cells differentiate into milk-producing alveolar cells and the milk is ejected from the gland by the contractile action of specialised myoepithelial cells, smooth muscle-like epithelial cells which differentiate from basal cells (Macias & Hinck, 2012). From hereon 'basal' refers to the basal lineage, encompassing cap cells, duct basal cells, and myoepithelial cells.

Initial transplantation experiments demonstrated that the basal lineage contains bipotent mammary stem cells (MaSCs), capable of forming both myoepithelial and luminal populations, while luminal cells are restricted in their differentiation potential (Shackleton, Vaillant et al., 2006, Stingl, Eirew et al., 2006). Subsequent lineage tracing experiments support the existence of a small population of basally-restricted bipotent stem cells (Rios, Fu et al., 2014, Wang, Cai et al., 2015), with a majority of unipotent basal and luminal mammary stem cells (Davis, Lloyd-Lewis et al., 2016, Lilja, Rodilla et al., 2018, Lloyd-Lewis, Davis et al., 2018, Prater, Petit et al., 2014, Scheele, Hannezo et al., 2017, van Amerongen, Bowman et al., 2012, Van Keymeulen, Rocha et al., 2011, Wuidart, Ousset et al., 2016, Wuidart, Sifrim et al., 2018) that replenish the myoepithelial and luminal lineages throughout life.

Lineage specifying transcription factors are responsible for directing luminal and myoepithelial differentiation, and also for maintaining the self-renewal of uncommitted stem cells upstream in the mammary epithelial hierarchy. Transcriptomic profiling of sorted epithelial subpopulations has been instrumental in identifying such lineage specifying transcription factors that regulate each step of luminal-alveolar differentiation (Asselin-Labat, Sutherland et al., 2007, Bouras, Pal et al., 2008, Buchwalter, Hickey et al., 2013, Carr, Kiefer et al., 2012, Chakrabarti, Wei et al., 2012, Kouros-Mehr, Slorach et al., 2006, Liu, Ginestier et al., 2008, Oakes, Naylor et al., 2008, Yamaji, Na et al., 2009). However, due to lack of specific cell markers that can resolve stem and myoepithelial populations, it has been challenging to dissect molecular regulators of basal differentiation. While a number of basal-specific transcription factors have been identified, such as P63, SLUG, SOX9, SRF and MRTFA (Guo, Keckesova et al., 2012, Li, Chang et al., 2006, Mills, Zheng et al., 1999, Sun, Boyd et al., 2006, Yang, Schweitzer et al., 1999), their role in the basal compartment and myoepithelial specialisation is poorly understood. ID proteins (ID1-4) are helix-loop-helix (HLH) transcriptional regulators that lack a DNA binding domain. They function by dimerizing with basic HLH (bHLH) transcription factors and preventing them from binding to E-box DNA motifs and regulating transcription (Benezra, Davis et al., 1990). E-box motifs are found in regulatory regions of genes involved in lineage specification and as such ID proteins and bHLH transcription factors are critical regulators of stemness and differentiation across diverse cellular linages (Massari & Murre, 2000). The expression of ID4 in mouse and human mammary epithelium is exclusive to the basal population (Lim, Wu et al., 2010). We and others have demonstrated that ID4 is a key regulator of mammary stem cells,

93 required for ductal elongation during puberty (Best, Hutt et al., 2014, Dong, Huang
94 et al., 2011, Junankar, Baker et al., 2015). ID4 was also shown to have a role in
95 blocking luminal differentiation (Best et al., 2014, Junankar et al., 2015).The precise
96 molecular mechanism by which ID4 functions in the mammary gland, including its
97 full repertoire of transcriptional targets and interacting partners, have yet to be
98 determined. Here we show that ID4 marks immature basal cells and demonstrate
99 that ID4 inhibits myoepithelial differentiation. Moreover, using unbiased interaction
100 proteomics, we identify a novel player in the mammary differentiation hierarchy –
101 the bHLH transcription factor HEB. By mapping the genome-wide binding sites of
102 HEB we show that it directly binds to regulatory elements of ID4 target genes
103 involved in myoepithelial functions such as contraction and extracellular matrix
104 (ECM) synthesis.

105 Results

106 **Loss of ID4 causes upregulation of myoepithelial genes in basal cells**

107 In order to determine the genes regulated by ID4 in mammary epithelial cells, we
108 FACS-enriched basal, luminal progenitor and mature luminal cells from wild type
109 (WT) and ID4 knockout (KO) mice (Yun, Mantani et al., 2004) and performed RNA-
110 sequencing (RNA-seq) (Fig. 1a). No significant differences were observed in the
111 proportion of mammary epithelial subpopulations between WT and KO mice (Fig.
112 S1a). 96 genes were significantly (FDR<0.05) upregulated and 141 genes
113 downregulated in ID4 KO basal cells. In contrast, only 2 genes were differentially
114 expressed in the two luminal subpopulations (Fig. 1b-c and Table S1) suggesting that
115 ID4 predominantly regulates gene expression within basal cells *in vivo*.

In line with the known role of ID4 in promoting proliferation of mammary epithelial cells (Dong et al., 2011, Junankar et al., 2015), the genes downregulated in ID4 KO basal cells were enriched for pathways involved in cell growth such as translation and metabolism (Fig. 1d-e and Fig. S1b). Conversely, the genes upregulated in KO basal cells were enriched for pathways related to basal cells and smooth muscle function. These included a muscle-derived leiomyosarcoma signature (Nielsen, West et al., 2002), smooth muscle contraction genes, epithelial-mesenchymal transition (EMT) and Serum Response Factor (SRF) targets (Fig. 1d and Fig. S1c), which shared common genes as indicated by the connecting edges in the enrichment map network (Fig. 1e). SRF target genes are of relevance as SRF is a master regulator of cytoskeletal contraction, and is one of the only transcription factors implicated in myoepithelial differentiation (Li et al., 2006, Miano, Long et al., 2007, Sun et al., 2006). Taken together, loss of ID4 causes basal cells to decrease proliferation and adopt a more differentiated myoepithelial and mesenchymal gene expression program, implicating ID4 in the repression of basal cell specialisation.

ID4 marks basal stem cells and is dynamically regulated during myoepithelial differentiation

We previously showed that ID4 is heterogeneously expressed in basal cells and that ID4-positive cells have enhanced mammary reconstitution activity (Junankar et al., 2015). To further characterise the phenotype of ID4-positive basal cells, we used an ID4-GFP reporter mouse in which the ID4 promoter drives GFP expression (Best et al., 2014). Basal cells with high expression of the stem cell marker CD49f (also known as ITGA6) and the epithelial marker EPCAM have been shown to be enriched for

140 MaSC activity (Prater et al., 2014, Stingl et al., 2006). ID4-GFP expression was
141 maximal in this EPCAM^{hi} CD49f^{hi} subset (Fig. 2a). ID4-GFP expression within the basal
142 gate was binned into 3 groups: bright (top 10%), intermediate (middle 80%) and dim
143 (bottom 10%) and the median fluorescent intensity (MFI) of EPCAM and CD49f was
144 analysed within these groups (Fig. 2b-c). ID4-bright cells had significantly higher
145 EPCAM and CD49f MFI than ID4-dim and intermediate cells, suggesting ID4 marks
146 basal stem cells.
147 We next sought to locate ID4-high and ID4-low populations in a tissue context and to
148 further investigate the association between ID4 and markers of myoepithelial
149 differentiation. During ductal elongation, the cap cells differentiate at the neck of
150 the TEBs and mature into myoepithelial cells which form the outer basal layer of the
151 ducts (Paine & Lewis, 2017). Mammary gland sections from pubertal mice, in which
152 TEBs and ducts are both present, were stained for ID4 and the myoepithelial marker
153 alpha-smooth muscle actin (α-SMA). ID4 expression was highest in the nuclei of cap
154 cells at the extremity of the TEBs and expressed at lower levels in basal cells of ducts
155 (Fig. 2d and Fig. S2). Conversely, α-SMA expression was higher in ductal basal cells
156 and lower in cap cells. ID4-high cap cells had a compact cuboidal epithelial
157 appearance compared to the more separated elongated morphology of the ID4-low
158 duct cells (Fig. 2d and Fig. S2). Quantification of fluorescence demonstrated that ID4
159 was more highly expressed in cap cells of TEBs than in basal cells of ducts, while the
160 opposite was true for α-SMA expression (Fig. 2e-f). A negative correlation between
161 α-SMA and ID4 was observed, with clear separation between cap (dark blue Fig. 2g)
162 and duct basal cells (light blue Fig. 2g). As a negative control, α-SMA fluorescence
163 was compared with nuclear stain Hoechst and no correlation or separation based on

region was observed (Fig. 2h). Thus, based on marker expression, morphology and spatial localization, ID4 expression is high in epithelial-like cap cells and is lower in myoepithelial cells.

Terminal differentiation of basal cells into contractile myoepithelial cells occurs during lactation. We interrogated a published single cell RNA-seq (scRNA-seq) data-set (Bach, Pensa et al., 2017) to examine *Id4* expression dynamics over postnatal murine mammary gland development. In this study, individual EPCAM+ mammary epithelial cells from four developmental stages: nulliparous (8 week), gestation (day 14.5), lactation (day 6), and involution (day 11) were captured and profiled. We limited our analysis to basal cell clusters (9,663 cells), defined by expression of both *Krt5* and *Krt14*. Basal cells broadly clustered by developmental time point (Fig. 2i). Increased differentiation of myoepithelial cells appears to proceed from nulliparous to gestation to lactation, associated with decreased expression of epithelial marker *Epcam* and increased expression of *Acta2* (encoding α-SMA) (Fig. 2j-k), consistent with gradual acquisition of a smooth muscle phenotype and loss of adherent epithelial features (Deugnier, Moiseyeva et al., 1995). Like *Epcam*, *Id4* expression was highest in basal cells from nulliparous mice, and decreased in basal cells of pregnant and lactating mice (Fig. 2j-k). This result was validated on the protein level by immunohistochemical staining for ID4 on mammary gland sections at different developmental time points (Fig. S2b). Using the Monocle 2 package (Trapnell, Cacchiarelli et al., 2014), we performed pseudotemporal ordering of all basal cells to form a myoepithelial differentiation trajectory (Fig. S2c). As expected, the nulliparous and involution cells clustered together in pseudotime space in the least differentiated part of the trajectory. Basal cells from gestating mice were dispersed

188 between the nulliparous and lactation stages, while basal cells at lactation were the
189 most differentiated. *Id4* expression decreased, while several myoepithelial markers
190 (*Acta2, Cnn1, Mylk, Myh11* and *Oxtr*) increased over pseudotime (Fig. S2d).
191 We next sought to identify transcriptional signatures associated with high and low
192 *Id4* expressing cells in the mammary basal epithelium across all developmental time
193 points (Fig 2l and Table S2). Basal cells with high *Id4* expression were enriched for
194 genes involved in RNA binding, metabolic processes, translation and ribosome
195 biogenesis (Fig. 2m). Conversely, genes enriched in the *Id4*-low basal cells were
196 involved in muscle contraction and response to cytokine and hormone stimuli, and
197 circulatory system development (Fig. 2n). Notably, myoepithelial genes such as *Oxtr,*
198 encoding the oxytocin receptor, and contractile genes *Mylk, Cnn1, Cav1* and *Myh11,*
199 were among the top differentially expressed genes in the cells with low *Id4*
200 expression (Fig. 2l and Table S2). Thus, ID4 downregulation during pregnancy and
201 lactation is associated with terminal differentiation into functionally mature
202 contractile myoepithelial cells.
203
204 **Loss of ID4 results in myoepithelial differentiation of mammary organoids**
205 To functionally validate the role of ID4 in suppressing the myoepithelial
206 differentiation of basal cells we generated a primary basal cell organoid model (Fig.
207 3a). The organoid model system complements and expands upon the findings from
208 the KO mouse, as the acute consequence of ID4 loss on basal cell phenotype can be
209 determined. Furthermore, organoids are less complex cellular systems than tissue,
210 thus cell-autonomous effects can be isolated more precisely. ID4-positive basal cells
211 from mice in which exons 1 and 2 of *Id4* are floxed (*Id4*$^{fl/fl}$) (Best et al., 2014) were

212 isolated using cell sorting. To overcome culture-induced senescence, basal cells were
213 conditionally reprogrammed into a proliferative stem/progenitor state using an
214 established protocol utilizing irradiated 3T3 fibroblast feeders and Rho Kinase
215 (ROCK) inhibition (Liu, Ory et al., 2012, Prater et al., 2014) (Fig. 3a). Conditionally
216 reprogrammed basal cells adopted an epithelial cobblestone morphology and
217 expressed high levels of ID4 (Fig. S3a-b). Cells cultured in the absence of ROCK
218 inhibitor or feeders adopted a flattened differentiated/senescent cell morphology
219 and had reduced ID4 expression (Fig. S3a-b). Reprogrammed cells maintained basal
220 marker expression of P63 and KRT14 (Fig. S3c) and could be grown as 3D organoids
221 with basal marker KRT14 on the outer cell layer and luminal marker KRT8 on the
222 inner cells of the organoids (Fig. S3d).
223 ID4 expression was limited to the outer basal cells (Fig. 3b), recapitulating expression
224 of ID4 in cap cells of TEBs (Fig. 2d). To test whether ID4 regulates differentiation of
225 basal cells we deleted ID4 with Adenoviral-Cre and compared these to cells treated
226 with control Adenoviral-GFP. ID4 protein was markedly downregulated in organoids
227 infected with Cre adenovirus confirming successful gene deletion (Fig. 3b-c). Loss of
228 ID4 resulted in increased α-SMA fluorescent intensity (Fig. 3d-e). This finding
229 demonstrates that as well as marking undifferentiated basal cells, ID4 has a cell
230 autonomous role in impeding maturation of basal cells into myoepithelial cells.
231
232 **ID4 inhibits expression of contractile and ECM genes *in vitro***
233 Given the negative association between ID4 and the differentiated myoepithelial
234 phenotype, we next sought to determine direct ID4 target genes by performing RNA-
235 seq following siRNA-mediated ID4 depletion in the spontaneously immortalized

236 mouse mammary epithelial cell line Comma-Dβ (Danielson, Oborn et al., 1984). This
237 normal-like cell line expresses basal markers and is commonly used as a model of
238 mammary stem/progenitors cells as they retain the capacity of multi-lineage
239 differentiation (Best et al., 2014, Danielson et al., 1984, Deugnier, Faraldo et al.,
240 2006, Idoux-Gillet, Nassour et al., 2018, Junankar et al., 2015). They can also be
241 readily expanded *in vitro* to produce a large amount of material for investigation of
242 mechanisms by molecular biology and biochemistry. Western blotting confirmed 70-
243 80% knockdown (KD) of ID4 protein 48 hr after siRNA transfection (Fig. 4a). RNA-seq
244 analysis of ID4 KD cells resulted in 859 and 838 (FDR<0.05) down and upregulated
245 genes, respectively, compared to the non-targeting siRNA control (Table S3).
246 Downregulated genes were predominantly involved in cell proliferation and growth
247 pathways (Fig. 4b-c) including hallmark gene sets such as MYC targets, E2F targets
248 and G2M checkpoint (Fig. S4a-b). Driving enrichment were several key cell cycle
249 genes such as *Aurka, Mki67, Cdk1, Cdk2, Cdk6, Ccnb1* and *Ccnd1* (Table S3). This
250 result is consistent with the loss of cell growth gene expression programs in ID4 KO
251 basal cells *in vivo* (Fig. 1d-e).
252 Genes acutely upregulated upon ID4 depletion were involved in development,
253 morphogenesis, ECM remodelling, and immune signalling (Fig. 4b). Consistent with
254 the gene expression profiling of sorted basal cells (Fig. 1d-e), loss of ID4 resulted in
255 upregulation of SRF targets, EMT and myogenesis gene signatures (Fig. 4d and Fig.
256 S4a-b). Several myoepithelial genes involved in contraction were significantly
257 increased in ID4 depleted cells including *Acta2, Cnn1, Cnn2, Tagln, Lmod1, Mylk* and
258 *Oxtr* (Table S3). Many of these contractile genes were also inversely correlated with
259 *Id4* expression in the scRNA-seq analysis (Fig. 2l and Table S2).

Several ECM genes encoding collagens (e.g. *Col1a1, Col1a2, Col4a3* and *Col5a1*), basement membrane laminins (e.g. *Lamc1* and *Lama4*), matricellular proteins (e.g. *Sparc*), and secreted factors (e.g. *Wnt5b* and *Igf1*) were also upregulated upon ID4 depletion (Fig. 4c-d and Table S3). The ECM provides physical support to the mammary gland and is a source of biochemical signals important for coordinating morphogenesis (Muschler & Streuli, 2010). ECM gene expression changes are associated with EMT and cellular contractility (Kiemer, Takeuchi et al., 2001, Liu, Eischeid et al., 2012). To functionally validate the role of ID4 in regulating ECM proteins we examined whether ID4 represses ECM deposition *in vivo* using picrosirius red staining of mammary glands from ID4 WT and KO mice to visualise fibrillar collagen (Fig. 4e). ID4 KO TEBs in pubertal mammary glands were surrounded by a thickened collagen-dense ECM when compared to ID4 WT TEBs (Fig. 4e i-ii), indicating that ID4 normally restrains collagen expression by cap cells during puberty. In addition to total abundance, the thickness of bundled collagen fibres can be further distinguished using polarized light. Analysis of birefringence signal revealed an increase in thick fibres and a decrease in thin fibres in the ECM surrounding ID4 KO TEBs (Fig. 4e iii) signifying a redistribution of collagen composition, as well as an overall increase in collagen abundance, in the absence of ID4.

Hence, ID4 positively regulates proliferative genes and negatively regulates genes involved in myoepithelial functions such as contraction and ECM synthesis in mammary epithelial cells. These ID4-regulated functions are likely to be important for morphogenesis of the ductal tree during pubertal development.

284 **ID4 interacts with E-proteins in mammary epithelial cells**

285 As ID4 lacks a DNA binding domain, it inevitably influences transcription through its
286 interaction with other DNA-binding proteins. We used Rapid Immunoprecipitation
287 Mass spectrometry of Endogenous proteins (RIME) to discover the binding partners
288 of ID4 in Comma-Dβ cells (Mohammed, D'Santos et al., 2013). We identified 48
289 proteins that were significantly ($p<0.05$) more abundant in the ID4 IPs compared to
290 the IgG negative control IPs in three independent RIME experiments (Fig. 5a and
291 Table S4). ID4 was consistently identified in all replicates and was the top hit (Table
292 S4), verifying the validity of the technique. Among the putative ID4 binding partners
293 were many DNA and RNA interacting proteins (Fig. 5a).

294 The E-proteins E2A and HEB were identified as ID4 interactors in our RIME analysis.
295 There are three members of the E-protein family (E2A, HEB, and ITF-2), that dimerise
296 with other E-proteins or tissue-specific bHLH transcription factors (e.g. MyoD and
297 NeuroD) to regulate expression of lineage commitment genes (Wang & Baker, 2015).
298 E2A has been implicated in branching morphogenesis of mammary organoids (Lee,
299 Gjorevski et al., 2011), however, there are no previous studies implicating HEB in
300 mammary gland development. Given this, and that E-proteins are the canonical
301 binding partners of ID proteins in other lineages, we chose to pursue the ID4-HEB
302 interaction further.

303 Reciprocal co-immunoprecipitation followed by western blotting (co-IP WB)
304 experiments were performed to validate the ID4-E-protein interactions in the same
305 Comma-Dβ cell line, as well as in normal human mammary epithelial cell lines
306 PMC42 and MCF10A. IP of ID4 resulted in co-IP of E2A and HEB and correspondingly,
307 IP of E2A and HEB resulted in co-IP of ID4 (Fig. 5b and S5a-b). E2A and HEB did not

form heterodimers with each other in Comma-Dβ cells (Fig. 5b), implying that E-proteins either function as homodimers, or heterodimers with other bHLH proteins in this context. To identify mammary specific bHLH factors binding to HEB, we performed another RIME experiment by immunoprecipitating HEB protein. HEB (HTF4_MOUSE in Table S4) was identified as the top hit, and ID4 was also identified (Fig. S5c). However, we did not identify any other bHLH transcription factors in this experiment (Fig. S5c and Table S4). This suggests that HEB either binds DNA as a homodimer or binds another factor that was not detected by this assay. To independently validate the interaction between ID4 and HEB, the Proximity Ligation Assay (PLA) was used to visualize protein-protein interactions *in situ*. Multiple PLA foci were detected in the nuclei of Comma-Dβ cells co-stained with ID4 and HEB antibodies (Fig. 5c-d).

To test if ID4 and HEB interacted *in vivo*, we engineered a tagged ID4 mouse model in which a FlagV5 tag, a very efficient and specific target for immunoprecipitation, was integrated into the *Id4* locus downstream of the open reading frame. The result was FlagV5-tagged ID4 protein under the control of endogenous regulatory elements. We first ensured that ID4 and HEB were co-expressed by performing co-immunofluorescent staining for V5 and HEB. ID4-FlagV5 expression was tightly restricted to the cap and ductal basal cells, while HEB was more ubiquitous in its expression; detected in basal and luminal epithelial cells and surrounding stromal cells (Fig. 5e). ID4 and HEB co-localized in the nuclei of cap and duct basal cells (insets; Fig. 5e). Reciprocal co-IP was carried out from digested mammary glands of ID4-FlagV5 mice using antibodies raised against ID4, Flag, V5 (two different antibodies) and HEB (Fig. 5f). Each antibody tested was able to precipitate both ID4

332 and HEB, confirming interaction between the two transcription factors *in vivo*.

333

334 **HEB binds to regulatory elements of a subset of ID4 differentially expressed genes**

335 To establish if ID4 regulates gene expression through HEB, we sought to determine if
336 HEB directly binds to genes regulated by ID4 by performing ChIP-seq for HEB in
337 Comma-Dβ cells. Across four biological replicates there were a total of 2752 HEB
338 peaks identified (FDR<0.05). This was narrowed down to 956 consensus peaks which
339 were present in at least 2 replicates. Transcription factor motif enrichment was
340 carried out using MEME-ChIP (Machanick & Bailey, 2011) and the top enriched
341 motifs were canonical E-box motifs (CANNTG), which are the binding sites for E-
342 proteins (Fig. 6a). The majority of peaks were mapped to intergenic and intronic
343 regions, and approximately 5% of the peaks occurred at gene promoters (Fig. S6a).
344 We used the Genomics Regions Enrichment of Annotations Tool (GREAT) to analyse
345 the functional significance of the regions bound by HEB (McLean, Bristor et al.,
346 2010). In this unbiased analysis, top enriched pathways were related to actin
347 cytoskeleton and ECM organization (Fig. 6b), resembling the pathways negatively
348 regulated by ID4 in the gene expression profiling (Fig. 4c-d and Fig. S4).
349 The consensus peaks were annotated to 1320 genes, using the default GREAT basal
350 plus extension gene annotation rule (McLean et al., 2010). We overlapped these
351 genes with those regulated following ID4 KD to determine the genes directly
352 regulated by HEB (Fig. 6c). Approximately 10% of these genes had an associated HEB
353 peak, which is more than expected by chance (p<1.595E-27; hypergeometric test).
354 The remainder of the genes regulated by ID4 KD are likely due to secondary effects
355 of ID4 KD or HEB independent mechanisms. The fact that there was a similar number

356 of genes overlapping in both upregulated (85) and downregulated (92) genes,
357 suggests that HEB can bind to sites both positively and negatively regulated by ID4.
358 In line with this, E-proteins have previously been demonstrated to act as both
359 activators and repressors of gene transcription by recruiting different co-factors
360 (Bayly, Chuen et al., 2004, Zhang, Kalkum et al., 2004). Genes negatively regulated by
361 ID4 (up in ID4 KD), and associated with a HEB peak were involved in ECM-related
362 pathways (Fig. S6b), however no pathways were enriched in the HEB-bound genes
363 positively regulated by ID4 (down in ID4 KD).
364 In parallel we also performed ChIP-seq for three histone modifications; H3K4Me3
365 (active promoter mark), H3K27Ac (active enhancer mark) and H3K27Me3 (repressive
366 chromatin mark) to elucidate the chromatin context of HEB-bound regions. A
367 number of HEB peaks associated with ID4 regulated genes demonstrated a bimodal
368 distribution of H3K27Ac signal (Fig. 6d and Fig. S6c) suggesting that HEB binds to
369 enhancers. Some of the peaks were also localized to active promoters (Fig. 6d and
370 Fig. S6c). HEB peaks were observed at enhancer-marked chromatin upstream of ID4-
371 repressed ECM genes expressed in myoepithelial cells (Barsky & Karlin, 2005) such as
372 *Sparc, Col1a1, Col1a2, Col3a1* and *Col5a1* (Fig. 6e and Fig. S6d). HEB binding was also
373 observed near the promoter of contractile gene *Cnn2* (Fig. 6e). Of relevance *Cnn2*,
374 encoding Calponin 2, was recently discovered to be regulated by a super enhancer
375 specifically accessible in myoepithelial cells (Pervolarakis, Nguyen et al., 2019). This
376 suggests that ID4 inhibits HEB's ability to activate transcription of genes that define
377 the myoepithelial fate.
378 We next performed HEB ChIP-seq on Comma-Dβ cells in which ID4 had been
379 depleted by siRNA in order to determine whether HEB DNA binding is augmented

380 when released of inhibition by ID4. Western blotting revealed that ID4 protein was
381 reduced to approximately 20% of control levels, while HEB expression was
382 unchanged (Fig. S6e). E-box motifs were again enriched in both conditions (Fig. S6f).
383 Differential binding analysis revealed a total of 290 regions changing upon ID4
384 depletion ($p<0.05$) (Table S5). More peaks were increased than decreased (263
385 compared to 27) suggesting that, as expected, depletion of ID4 increased HEB's DNA
386 binding activity (Fig. 6f and Fig. S6g). GREAT analysis revealed that the peaks that
387 increased were involved in processes such as gland morphogenesis, skeletal
388 development, and branching morphogenesis (Fig. S6h). No pathways were enriched
389 in the regions that were decreased when ID4 was knocked down. Finally, upon ID4
390 depletion, HEB binding was increased at genes regulated by ID4 (Fig. 6g).
391 Together, our ChIP-seq analysis suggests that HEB binds to E-box motifs in regulatory
392 elements of basal developmental genes involved in ECM and the contractile
393 cytoskeleton, and this is antagonised by its interaction with ID4 (Fig. 6h).

394 Discussion

395 The mammary gland is the defining morphological feature of mammals and is also
396 the source of one of the most common human cancers. Beyond their role in milk
397 ejection at lactation, myoepithelial cells can also sense, synthesize and remodel the
398 ECM, and control the polarity of neighbouring luminal cells (Barsky & Karlin, 2005,
399 Gudjonsson, Rønnov-Jessen et al., 2002, Ingthorsson, Hilmarsdottir et al., 2015).
400 Furthermore, myoepithelial cells are "natural tumour suppressors"; acting as
401 physical barriers to prevent cancer cell invasion, secreting factors that limit growth
402 and rarely undergoing malignant transformation (Gudjonsson, Adriance et al., 2005,

Sternlicht, Kedeshian et al., 1997). Thus, understanding the regulation of the basal lineage has implications for fundamental mammary gland biology as well as breast cancer.

Both bipotent and unipotent stem cells have been identified within the mammary epithelium (reviewed in (Visvader & Clevers, 2016)). ID4 has previously been demonstrated to block luminal commitment of basal cells via inhibition of key luminal driver genes including *Elf5, Notch, Brca1, Esr1, PR* and *FoxA1* (Best et al., 2014, Junankar et al., 2015). We now show that ID4 can also block genes associated with myoepithelial differentiation, in part through the inhibition of the E-protein HEB. The dual inhibition of both luminal and myoepithelial differentiation by ID4 may protect the stem-cell phenotype of uncommitted basal cells during development. Subsequent downregulation of ID4, through a currently unknown mechanism, may then allow basal cells to adopt a luminal or myoepithelial fate depending on the cellular context.

HEB has been implicated in the specification of lymphocyte (Braunstein & Anderson, 2012), haematopoietic (Li, Brauer et al., 2017), mesodermal (Yoon, Foley et al., 2015), neuronal (Mesman & Smidt, 2017), and skeletal muscle (Conway, Pin et al., 2004) lineages. It however has not been associated with lineage commitment of epithelial tissues. Our proteogenomic analyses support the model outlined in Fig. 6h. When ID4 is highly expressed, such as in cap cells, it sequesters HEB off chromatin, preventing expression of differentiation genes, thus determining a stem-like state (Fig. 6h; left). When ID4 expression is low, HEB is able to bind to E-box DNA motifs at

427 promoters/enhancers to activate transcription of developmental genes that specify
428 functional myoepithelial cells (Fig. 6h; right). Further functional studies are needed
429 to demonstrate HEB's requirement in promoting myoepithelial differentiation.
430 Compared with the luminal lineage, the molecular regulators controlling the basal
431 lineage remain poorly understood. One of the few transcription factors known to
432 promote myoepithelial differentiation is MADS-box protein SRF and its associated
433 co-activator MRTFA (Li et al., 2006, Sun et al., 2006). Interestingly, HEB and SRF
434 cooperatively activate transcription of *Acta2* in cultured fibroblasts (Kumar, Hendrix
435 et al., 2003). This occurred in an E-box dependent manner and was inhibited by ID1
436 and ID2 overexpression (Kumar et al., 2003). While we show that ID4 suppresses
437 *Acta2*, we did not observe HEB binding to the *Acta2* promoter. However, it is
438 possible that in the absence of ID4, HEB and SRF cooperate to drive expression of
439 other myoepithelial genes. This is supported by the positive enrichment of SRF
440 targets upon ID4 depletion. Subsequent studies should test whether HEB and SRF
441 cooperate in mammary epithelial cells.
442
443 The newly elucidated roles of ID4 in regulating myoepithelial commitment and ECM
444 deposition expand upon why ID4 is required for pubertal mammary gland
445 morphogenesis (Best et al., 2014, Dong et al., 2011, Junankar et al., 2015). TEBs
446 undergo collective migration, enabling the coordinated movement of adherent cells
447 into the stromal fat pad (Ewald, Brenot et al., 2008). We hypothesise that the high
448 levels of ID4 in cap cells prevent epithelial cells from acquiring a
449 mesenchymal/myoepithelial phenotype. Similar mechanisms have been observed
450 for the transcriptional repressor OVOL2, which inhibits EMT to allow for collective

451 migration (Watanabe, Villarreal-Ponce et al., 2014), and for C/EBPα, which maintains
452 epithelial homeostasis of human mammary epithelial cells (Lourenço, Roukens et al.,
453 2020). During ductal elongation, collagenous stromal ECM is absent directly in front
454 of the invading TEBs (Paine & Lewis, 2017, Silberstein, Strickland et al., 1990,
455 Sternlicht, 2006). We show that ID4 suppresses collagen synthesis and deposition
456 around TEBs, which may otherwise act as a physical barrier to impede invasion. In
457 support of this idea, ectopic deposition of collagen by mammary epithelial cells
458 induced by exogeneous TGF-β, or forced expression of recombinant type I collagen
459 which is resistant to collagenase attack, causes ensheathment of TEBs by collagen,
460 and retardation of ductal elongation (Feinberg, Zheng et al., 2018, Silberstein et al.,
461 1990).
462
463 Developmental transcription factors are often dysregulated in cancer. ID4 is highly
464 expressed in ~50% of basal-like breast cancer (BLBC) cases and associates with poor
465 prognosis (Baker, Holliday et al., 2016, Junankar et al., 2015). Given the cell-intrinsic
466 role of ID4 in promoting growth/proliferation and inhibiting differentiation in the
467 mammary gland, it is easy to envision how overexpression of ID4 could lead to an
468 aggressive breast cancer phenotype. Additionally, the suppression of ECM synthesis
469 may allow tumour cells to easily invade into the surrounding stroma, akin to the
470 invasion of cap cells during ductal elongation. Future work should test whether the
471 mechanisms discovered here are conserved in cancer. If so, targeting ID4 may
472 represent an attractive therapeutic strategy for ID4-positive BLBCs. Excitingly, a
473 small-molecule Pan-ID inhibitor has recently been identified by the Benezra
474 laboratory, which works by inhibiting the interaction between ID and E-proteins

475 (Wojnarowicz, Lima E Silva et al., 2019).

476

477 To conclude, these new insights into ID4 and HEB function have helped unravel
478 regulation within the elusive basal differentiation hierarchy, with broader
479 implications for the regulation of epithelial stem cells in general, and also in tumour
480 progression.

481 Materials and methods

482 **Mice**

483 All mice experiments were performed in accordance with the ethical regulations of
484 the Garvan Institute Animal Experimentation Committee. ID4 KO mice were
485 generated as previously described (Yun et al., 2004). The founder ID4-FlagV5 mouse
486 was engineered using Crispr/Cas9 gene editing by the Mouse Engineering at
487 Garvan/ABR (MEGA) facility. The ID4floxGFP mice were generated as previously
488 described (Best et al., 2014). All mice used were on the FVB/N background. For the
489 gene expression profiling experiment, mice were synchronised in estrus to reduce
490 hormone-induced gene expression variation (Dalal, Estep et al., 2001), and checked
491 by vaginal swab cytology.

492

493 **Mammary epithelial cell preparations**

494 Mammary epithelial cells were prepared from freshly harvested 3rd and 4th
495 mammary glands pooled from 4-8 female mice at the indicated ages. Glands were
496 mechanically disrupted using a McIlwain tissue chopper then digested with 15,000 U
497 collagenase (Sigma-Aldrich) and 500 U hyaluronidase (Sigma-Aldrich) in FV media

498 (DMEM/F12 (Gibco), 5% (v/v) FBS (HyClone) 10 mM HEPES (Gibco), 0.14 IU/mL
499 Insulin (Novo Nordisk), 500 ng/mL Hydrocortisone (Sigma-Aldrich), 20 ng/mL Cholera
500 Toxin (Sigma-Aldrich), 2 ng/mL mEGF (Life Technologies)). Digestion was carried out
501 in a shaking incubator for 1 hr at 37°C 225 rpm. The resulting organoids were further
502 digested with warm 0.25% trypsin for 1 min with constant pipetting followed by
503 treatment with 5 mg/mL Dispase (Roche) for 5 min in at 37°C. Cells were incubated
504 with 1X red blood cell lysis buffer (BD Biosciences) for 5 min at room temperature.
505 Cells were passed through a 70 µm strainer followed by a 40 µm strainer to remove
506 clumps and debris. Live cells were counted using trypan blue (Thermo Fisher
507 Scientific) and a haemocytometer prior to use in downstream applications.
508
509 **Flow cytometry and FACS**
510 Single-cell suspensions of mouse mammary epithelial cells were blocked in Fc block
511 cocktail (FACS buffer, 6.25 µg/mL Mouse BD Fc block (BD Bioscience) and 200 µg/mL
512 Rat Gamma Globulin (Jackson ImmunoResearch)) for 10 min on ice. Fluorophore- or
513 biotin-conjugated antibodies were diluted in Fc block cocktail and incubated on the
514 cells on ice for 20 min in the dark. Antibodies used were CD24-PE (1:200, BD
515 Biosciences, Clone M1/69), EPCAM-PerCP/Cy5.5 (1:200, Biolegend, Clone G8.8),
516 CD29-APC/Cy7 (BioLegend, Clone HMβ1-1), CD61-APC (1:50, Thermo Fisher
517 Scientific, Clone HMβ3-1), CD49f-APC (1:100, BioLegend, Clone GoH3), Ter119-biotin
518 (1:80, BD Biosciences, Clone TER119), CD45-biotin (1:100, BD Biosciences, Clone 30-
519 F11), CD31-biotin (1:40, Biolegend, Clone 390) and BP1-biotin (1:500, Thermo Fisher
520 Scientific, Clone 6C3). Cells were washed and stained with streptavidin-BV421
521 (1:100, Biolegend). Following washing, cells were resuspended in FACS buffer at a

522 density of $1x10^7$ cells/mL and DAPI was added (1:1000, Invitrogen). FACS was
523 performed on a FACS Aria III 4 laser 15 colour sorter with BD FACS DIVA software.
524 For analytical flow cytometry, a CytoFLEX 3 laser 13 colour flow cytometer (Beckman
525 Coulter) was used. All flow cytometry data was analysed using FlowJo software
526 version 10 (Tree Star Inc).
527
528 **RNA-seq**
529 RNA was extracted using the Qiagen miRNease mini kit (Qiagen) for Comma-Dβ cells
530 or the Qiagen miRNease micro kit (Qiagen) for FACS-sorted mammary epithelial
531 cells. Cells were sorted directly into 700 µL Qiazol lysis reagent to minimise cell loss,
532 with a maximum of $5x10^4$ cells per 1.5 mL Eppendorf tube. RNA was eluted in RNAse
533 free water. The Qubit RNA BR Assay kit (Thermo Fisher Scientific) was used to
534 measure RNA concentration and RNA integrity was determined using the Agilent
535 Bioanalyser 2100 with the 6000 Nano Assay (Agilent Technologies).
536 For low input RNA extracted from FACS sorted mammary glands, the Ovation RNA-
537 seq System V2 kit (NuGEN) was used to synthesise cDNA with RNA inputs ranging
538 from 0.5-2 ng as per the manufacturers protocol. The Ovation Ultralow System V2 kit
539 was then used to prepare libraries from the cDNA. cDNA libraries were sequenced
540 on a HiSeq 2500 system (high output mode) (Illumina), with 125 bp paired-end
541 reads. For RNA extracted from Comma-Dβ cells, The Illumina TruSeq Stranded mRNA
542 Library Prep Kit (Illumina) was used with 1 µg of input RNA following the
543 manufacturer's instructions. cDNA libraries were sequenced on the NextSeq system
544 (Illumina), with 75 bp paired-end reads.
545 Quality control was checked using FastQC to remove poor quality reads and to trim

546 Illumina sequencing adapters (Andrews, 2010). Reads were then aligned to the
547 mouse reference genome mm10 using STAR ultrafast universal RNA-seq aligner
548 (Dobin, Davis et al., 2013). RSEM accurate transcript quantification for RNA-seq data
549 was used to generate a gene read count table and to filter genes with read counts of
550 0 (Li & Dewey, 2011). Differential gene expression analysis was performed using
551 edgeR (McCarthy, Chen et al., 2012, Robinson, McCarthy et al., 2010) and
552 voom/limma (Ritchie, Phipson et al., 2015) R packages. Genes were ranked based on
553 the limma moderated t-statistic and this was used as input for the GSEA desktop
554 application (Broad Institute) using Molecular Signature Database (MSigDB)
555 collections (v6.1). The EnrichmentMap Cytoscape plugin (Merico, Isserlin et al.,
556 2010) was used to visualise the GSEA results.
557
558 **Analysis of public single-cell RNA-seq data**
559 Raw Unique Molecular Identifier (UMI) expression data was taken from Bach *et al.*
560 and re-processed using Seurat v2 using default parameters, as recommended by the
561 developers (Satija, Farrell et al., 2015). All basal cell clusters, defined by *Krt5* and
562 *Krt14* expression, were extracted for subsequent analysis. Developmental
563 trajectories were inferred using Monocle 2 (Qiu, Mao et al., 2017). Clustering was
564 first performed using the 'densityPeak' and 'DDRTree' methods, with a delta local
565 distance threshold of 2, and a rho local density threshold of 3. For inferring
566 pseudotime, we first selected all genes with a mean expression greater than 0.5, and
567 an empirical dispersion greater than 1, and proceeded with the top 1000 significant
568 genes for ordering. Differential gene expression between the two branch states were
569 performed using the MAST algorithm through the Seurat package (Finak, McDavid 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. ID4 Primary antibody (1:400, Biocheck BCH9/82-12) 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

594 diluent and incubated overnight at 4°C. Antibodies used were ID4 (1:400, Biocheck
595 BCH9/82-12), α-SMA (1:200, Sigma-Aldrich A5228), HEB (1:200, Protein Tech 14419-
596 1-AP), V5 (1:200, Santa Cruz sc-58052). Following washing, slides were incubated
597 with fluorescent secondary antibody (1:500, Jackson ImmunoResearch) for 1 hr at
598 room temperature. Nuclei were stained using Hoechst 33342 (Sigma-Aldrich). Slides
599 were mounted using Prolong Diamond mounting media (Thermo Fisher Scientific).
600 Fluorescent images were captured using a Leica DM 5500 microscope.
601 For quantification of fluorescence of ID4 and α-SMA in cap and ductal basal cells, 3
602 representative ducts and TEBs were imaged from each mouse (n=9). Using FIJI
603 software, a region of interest was drawn around 10 randomly selected basal cells
604 from each image. The mean fluorescence of ID4 and α-SMA was determined within
605 each cell.
606
607 **Picrosirius red staining**
608 Picrosirius red staining was performed following as per (Vennin, Chin et al., 2017,
609 Vennin, Mélénec et al., 2019). Briefly, dewaxed FFPE sections were stained with
610 haematoxylin. Sections were then treated with 0.2% Phosphomolybdic acid (Sigma-
611 Aldrich) followed by 0.1% Sirius red (Sigma-Aldrich) Picric Acid-Saturated Solution
612 (Sigma-Aldrich). Slides were then rinsed with acidified water (90 mM glacial acetic
613 acid) then with 70% ethanol. A Leica DM 4000 microscope was used to image total
614 collagen staining in the tissue sections by phase contrast. To image birefringence of
615 collagen fibres, 2 polarised filters were used. Matched phase-contrast and polarised
616 images were taken for each region of interest. Image analysis was performed using
617 FIJI image analysis software.

618

619 **Proximity Ligation Assay**

620 Comma-Dβ cells were grown on glass coverslips in 6 well plates until ~80% confluent.
621 Cells were fixed in 4% Paraformaldehyde (PFA) (ProSciTech) for 15 min then
622 permeablised in 0.2% Triton-X-100 for 15 min. The Duolink Proximity Ligation Assay
623 (PLA) (Sigma-Aldrich) was performed as per the manufacturer's protocol and using
624 the following antibodies: ID4 (1:50, Santa Cruz sc-365656) and HEB (1:200, Protein
625 Tech 14419-1-AP), and equivalent concentrations of IgG negative controls (sc-2027
626 and sc-2025). DAPI (Life Technologies) and Phalloidin (Life Technologies) were added
627 to the final wash step. Coverslips were mounted onto glass slides with Prolong
628 Diamond mountant (Thermo Fisher Scientific).
629 Cells were imaged using a Leica DMI Sp8 confocal microscope (63X oil objective). Six
630 random fields of view per image were captured, each with approximately 50 cells,
631 and images were quantified using Andy's algorithms FIJI package (Law, Yin et al.,
632 2017) to enumerate the number of PLA foci per nuclei.

633

634 **Cell lines**

635 The mouse mammary epithelial cell line Comma-Dβ was a gift was a gift from Joseph
636 Jeffery (University of Massachusetts, Amherst, MA, USA). Comma-Dβ cells were
637 maintained in DMEM/F12 media (Gibco) supplemented with 2% FBS (HyClone), 10
638 mM HEPES (Gibco), 0.125 IU/mL Insulin (Novo Nordisk) and 5 ng/mL mEGF (Life
639 Technologies). The human mammary epithelial cell line PMC42 was a gift from
640 Professor Leigh Ackland (Deakin University, Melbourne, Victoria, Australia). PMC42
641 cells were maintained in RPMI 1640 (Gibco) supplemented with 10% FBS (HyClone).

642 The MCF10A cell line was obtained from the American Type Culture Collection and
643 were maintained in DMEM/F12 (Gibco), 5% Horse Serum (Thermo Fisher Scientific),
644 20 ng/mL hEGF (In Vitro Technologies), 0.5 mg/mL Hydrocortison (Sigma-Aldrich),
645 100 ng/mL Cholera Toxin (Sigma-Aldrich) and 0.125 IU/mL Insulin (Novo Nordisk).
646

647 **Conditional reprogramming of primary mouse basal cells**

648 Viable basal cells were purified by FACS from 10-12 week old mice as described
649 above. Cells were collected into FAD media (DMEM/F12 3:1 (Gibco) supplemented
650 with 10% FBS, 0.18 mM Adenine (Sigma-Aldrich), 500 ng/mL Hydrocortisone (Sigma-
651 Aldrich), 8.5 ng/mL Cholera Toxin (Sigma-Aldrich), 10 ng/mL mEGF (Life
652 Technologies), 0.14 IU/mL Insulin (Novo Nordisk), 5 µM Y-27632 (Seleckchem), 1X
653 AB/AM (Gibco) and 50 µg/mL Gentamicin (Life Technologies). Basal cells were
654 maintained in culture using as per (Prater et al., 2014). Tissue culture flasks coated
655 with Growth Factor-reduced Matrigel (Corning) diluted 1:60 in PBS for 30 min to 1 hr
656 at 37°C. Excess Matrigel/PBS solution was aspirated from the flasks and basal cells
657 were seeded at a density of ~5000 cells/cm^2. Basal cells were co-cultured with
658 irradiated (50 Gy) NIH-3T3 feeder cells at a density of 10,000 cells/cm^2. Cells were
659 maintained in a 37°C low 5% oxygen 5% CO_2 incubator. Differential trypsinisation
660 was used to first remove the less-adherent 3T3 cells when passaging.
661

662 **Organoid culture**

663 A 40 µL plug of Matrigel (Corning) was added to each well of 8-well chamber slides
664 (Corning) and allowed to set for 30 min in a 37°C incubator. Primary cells were
665 resuspended in Epicult-B (Stemcell Technologies) media containing 2% Matrigel and

666 12,000 cells were seeded into each chamber in a total volume of 400 µL. Chambers
667 were observed using a light microscope to ensure that cells were not aggregated.
668 Organoids were allowed to form over 1 week and media was changed every 3 days.
669

670 **Immunofluorescent staining of organoids**

671 Organoids were stained within the chamber slides. Media was aspirated and
672 organoids were fixed with 2% PFA diluted in PBS for 20 min at room temperature.
673 Organoids rinsed with PBS for 5 min following permeabilisation with 0.5% Triton X-
674 100 in PBS for 10 min at 4°C then rinsed 3 times with 100 mM Glycine (Astral
675 Scientific) in PBS for 10 min each. Organoids were blocked for 1 hr in IF buffer with
676 10% goat serum (Vector Laboratories). Primary antibody made up in blocking buffer
677 was incubated in the chambers overnight at 4°C in a humidified chamber. Antibodies
678 used were ID4 (1:200, Biocheck BCH9/82-12), α-SMA (1:100, Abcam ab5694), KRT14
679 (1:1000, Covance PRB-155P), KRT8 (1:500, DSHB TROMA1) and P63 (1:100, Novus
680 NB100-691). The following day, slides were equilibrated to room temperature for 1-2
681 hr. Organoids were rinsed 2 times with IF buffer (0.1% BSA, 0.2% Triton-X-100, 0.05%
682 Tween-20 in PBS) for 20 min each. Fluorescent secondary antibody (Jackson
683 ImmunoResearch) diluted in blocking buffer (1:500) was added to the chambers and
684 incubated for 45 min in the dark. Secondary antibody was rinsed off for 20 min in IF
685 buffer, then 2 times with PBS for 10 min each. DAPI (Life Technologies) was
686 incubated in the chambers for 15 min followed by a 10 min PBS rinse. The walls of
687 the chamber slides were removed and one drop of Prolong Diamond mounting
688 media (Thermo Fisher Scientific) was added to each well. Slides were coverslipped
689 and edges sealed with clear nail varnish. Slides were allowed to dry for 24 hr at room

690 temperature protected from light.

691 Organoids were imaged using a Leica DMI Sp8 confocal microscope using the 40X oil
692 objective and 3X optical zoom. The FIJI software package was used to quantify the
693 fluorescence within the organoids. For α-SMA, a circle was drawn around the entire
694 organoid and fluorescence measured. For ID4, a mask was made using the DAPI
695 channel and signal was measured within the nuclear mask. Multiple representative
696 regions of background staining were quantified. Corrected fluorescence (CF) was
697 calculated as described in (McCloy, Rogers et al., 2014).

698 $$CF =$$

699 $$integrated\ density - (area\ of\ selected\ organoid *$$

700 $$mean\ fluorescence\ of\ all\ backgrond$$ readings)

701 **Cre mediated deletion of ID4 from primary cells**

702 Following differential trypsinisation to remove irradiated 3T3 feeder cells, 1.0-
703 1.5x10^5 ID4floxGFP primary basal cells were seeded into a T75 with 7.5x10^5 fresh
704 irradiated 3T3 cells in 20 mL FAD media. Adenovirus (Vector Biolabs) was added to
705 the media at a multiplicity of infection (MOI) of 100. Adenoviruses' used were codon
706 optimised Cre (iCre) and GFP adenovirus (#1772) and control eGFP adenovirus
707 (#1060). Infection efficiency was assessed the following day using a fluorescence
708 microscope to check GFP expression within the cells. Cells were harvested after 72 hr
709 for downstream analysis.

710

711 **siRNA transfections**

712 Comma-Dβ cells were seeded at a density of 1.5x10^4 cells/cm^2 into 6 well plates
713 (Corning) for RNA-seq or 100 mm dishes (Corning) for ChIP-seq in antibiotic-free

714 media. The following day siRNA constructs (Dharmacon) were transfected into the
715 cells using Dharmafect-4 transfection reagent (Dharmacon) as per the
716 manufacturer's instructions at 20 nM. siRNA used were siGENOME Mouse Id4
717 SMARTpool (M-043687) and ON-TARGETplus Non-targeting Control Pool (D-001810).
718 Media was changed the following day and cells were harvested 48 hr post-
719 transfection.
720
721 **Western blotting**
722 Cells were lysed in RIPA buffer (50 mM Tris-HCl pH 7.4, 1% NP-40, 0.5% Sodium
723 Deoxycholate, 0.1% SDS, 1% Glycerol, 137.5 mM NaCl, 100 µM Sodium
724 Orthovanadate, 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche))
725 and protein concentration was quantified using the Pierce BCA Protein Assay Kit
726 (Thermo Fisher Scientific) according to the manufacturer's instructions. Protein
727 lysates (15-30 µg protein) were mixed with 1X NuPage loading buffer (Life
728 Technologies) and 1X NuPage reducing agent (Life Technologies) and denatured by
729 heating at 85°C for 5 min. Samples were run on a 4-12% Bis/Tris gels (Life
730 Technologies) in MES or MOPS running buffer (Life Technologies). Protein was
731 transferred to a 0.45 µm PVDF membrane (Merck Millipore) using BioRad transfer
732 modules in transfer buffer (25 mM Tris, 192 mM Glycine, 20% Methanol).
733 Membranes were blocked in Odyssey blocking buffer (LiCOR) for 1 hr at room
734 temperature then incubated in primary antibody diluted in 5% BSA/TBS overnight at
735 4°C. The following antibodies were used for western blotting: ID4 (1:20,000,
736 Biocheck BCH9/82-12), E2A (1:1000, from Dr Nicolas Huntington and Prof Lynn
737 Corcoran from the Walter Eliza Hall Institute (WEHI)), HEB (1:1000, from Dr Nicolas

738 Huntington and Prof Lynn Corcoran, WEHI or 1:1000 ProteinTech 144191-1-AP), Flag
739 (1:5000, Sigma-Aldrich F1804), V5 (1:200, sc-58052), β-Actin (1:1000, Sigma-Aldrich
740 A5441) and α-Tubulin (1:1000, Santa Cruz sc-5286). Fluorescent secondary antibody
741 conjugated to IRDye680 or IRDye800 (LiCOR) diluted in Odyssey blocking buffer
742 (1:15,000 - 1:20,000) were used for detection. An Odyssey CLx Infrared Imaging
743 System (LiCOR) was used to image and quantify the western blots.
744
745 **Co-immunoprecipitation**
746 Pierce Protein A/G magnetic breads (Thermo Fisher Scientific) were incubated with
747 antibody for a minimum of 4 hr at 4°C on a rotating platform. For each IP 15 µL of
748 beads and 2.5 µg antibody were used. Immunoprecipitating antibodies used were
749 ID4 (pool of Santa Cruz sc-491 and sc-13047), E2A (from Dr Nicolas Huntington and
750 Prof Lynn Corcoran, WEHI), HEB (Santa Cruz sc-357 and antibody from Dr Nicolas
751 Huntington and Prof Lynn Corcoran, WEHI), Flag (Sigma-Aldrich F1804), V5 (Santa
752 Cruz sc-58052 and Thermo Fisher R960-25), species matched IgG negative controls
753 (sc-2027, sc-2025 and BioLegend 400602). Protein was extracted in IP lysis buffer
754 (10% Glycerol, 0.03% $MgCl_2$, 1.2% HEPES, 1% Sodium acid pyrophosphate, 1% Triton-
755 X, 0.8% NaCl, 0.4% NaF, 0.04% EGTA, 1x cOmplete ULTRA Tablet (Roche), 100 µM
756 Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and cells were passed 5 times
757 through a 23 g needle to aid in lysis. Protein lysates (0.5-2 mg per IP) were added to
758 washed beads and incubated overnight at 4°C on a rotating platform. Following
759 washing, beads were resuspended in 2X NuPage loading buffer (Life Technologies)
760 and 2X reducing agent (Life Technologies) in IP lysis buffer. Samples were incubated
761 at 85°C for 5 min. Beads were separated on a magnetic rack and supernatant loaded

762 onto NuPage gel for SDS-PAGE and western blotting. For detection of protein,
763 fluorescent TrueBlot secondary antibodies were used (Jomar Life Research).
764
765 **RIME**
766 The RIME protocol was adapted from the protocol developed by Mohammed *et al.*
767 (Mohammed et al., 2013). Comma-Dβ cells were grown in 150 mm tissue culture
768 dishes (Corning) until 80-90% confluent. A total of 8x 150 mm dishes were used per
769 RIME sample. Cells were fixed with 1% PFA (ProSciTech) in DMEM/F12 (Gibco) for 7
770 min at room temperature on a rocking platform. Cross-linking was quenched by
771 addition of molecular grade glycine (Astral Scientific) to a final concentration of 125
772 mM for 2 min on a rocking platform. Cells were washed twice with ice-cold PBS
773 (Gibco) and scraped in 1 mL of ice-cold PBS containing magnesium and calcium salts
774 (Gibco). Cells were centrifuged at 4°C for 5 min at 1200 rpm. Cell nuclei were
775 enriched through a series of lysis buffers. Cells were first resuspended in 10 mL lysis
776 buffer 1 (50 mM HEPES-KOH pH 7.5, 140 mM Sodium, 1 mM EDTA, 10% Glycerol,
777 0.5% NP-40 or Igepal CA-630, 0.25% Triton X-100, 100 µM Sodium Orthovanadate,
778 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche) and incubated for
779 30 min at 4°C on a rotating platform. Cells were then centrifuged at 3000 g for 3 min
780 at 4°C and supernatant removed. The pellet was then resuspended in 10 mL lysis
781 buffer 2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 100 µM
782 Sodium Orthovanadate, 20 µM MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet
783 (Roche)) and incubated for a further 30 min at 4°C on a rotating platform. Following
784 centrifugation, the resulting nuclei pellet was lysed in 2.5 mL lysis buffer 3 (10 mM
785 Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Sodium

786 Deoxycholate, 0.5% N-lauroylsarcosine, 100 µM Sodium Orthovanadate, 20 µM
787 MG132, 1 mM DTT and 1x cOmplete ULTRA Tablet (Roche)). The nuclear lysate was
788 sonicated using a Bioruptor sonicator (Diagenode) for 20 cycles of 30 sec on/30 sec
789 off. Triton-X-100 was added to a final concentration of 1% and the sample was then
790 centrifuged at 4700 g for 5 min at 4°C to remove cellular debris.
791 Immunoprecipitation was carried out from the sheared nuclear supernatant
792 described above with 20 µg of antibody and 100 µL of magnetic beads per sample.
793 The following day, the beads were washed 10 times with RIPA buffer then 5 times
794 with 100 mM ammonium hydrogen carbonate (AMBIC) (Sigma-Aldrich) solution to
795 remove salts and detergents, resuspended in 50 µL AMBIC solution and transferred
796 into clean tubes.
797
798 **Mass Spectrometry**
799 Samples were processed as described in (Huang, Yang et al., 2015) for liquid
800 chromatography coupled mass spectrometry (LC-MS/MS) using Sequential
801 Windowed Acquisition of all THeoretical fragment-ion spectra (SWATH) acquisition.
802 Briefly, samples were denatured in 100 mM triethylammonium bicarbonate and 1%
803 sodium deoxycholate, disulfide bonds were reduced in 10 mM DTT, alkylated in 20
804 mM iodo acetamide, and proteins digested on the magnetic beads using trypsin.
805 After C18 reversed phase (RP) StageTip sample clean up, peptides were analysed
806 using a TripleToF 6600 mass spectrometer (SCIEX, MA, USA) coupled to a nanoLC
807 Ultra 2D HPLC system (SCIEX). Peptides were separated for 60 min using a 15 cm chip
808 column (ChromXP C18, 3 µm, 120 Å) (SCIEX) with an acetonitrile gradient from 3-
809 35%. The MS was operated in positive ion mode using either a data dependent

810 acquisition method (DDA) or SWATH acquisition mode. DDA was performed of the
811 top 20 most intense precursors with charge stages from 2+ - 4+ with a dynamic
812 exclusion of 30 s. SWATH-MS was acquired using 100 variable size precursor
813 windows. DDA files were searched using ProteinPilot 5.0 (SCIEX) against the
814 reviewed UniProt *Mus musculus* protein database (release February 2016) using an
815 unused score of 0.05 with decoy search strategy enabled. These search outputs were
816 used to generate a spectral library for targeted information extraction from SWATH-
817 MS data files using PeakView v2.1 with SWATH MicroApp v2.0 (SCIEX) importing only
818 peptides with < 1% FDR. Protein areas, summed chromatographic area under the
819 curve of peptides with extraction FDR ≤ 1%, were calculated and used to compare
820 protein abundances between bait and control IPs.
821
822 **ChIP-seq**
823 Our ChIP-seq protocol was adapted from (Khoury, Achinger-Kawecka et al., 2020).
824 Comma-Dβ cells were grown in 150 mm or 100 mm dishes (Corning) until 80-90%
825 confluent. For the ChIP-seq experiment on unperturbed cells 2x 150 mm dishes were
826 used for HEB ChIP and 1x 150 mm dishes were used for histone marks. For the ID4
827 siRNA experiment 3x 100 mm dishes per condition were used. Cells were scraped in
828 ice-cold PBS (Gibco) and resuspended by passing 10 times through a 19 g syringe.
829 Cells were fixed in 1% PFA (ProSciTech) in PBS (Gibco) for 15 min at room
830 temperature followed by quenching with glycine to a final concentration of 125 mM
831 for 5 minutes followed by two PBS washes. Nuclei was extracted by resuspending
832 cells in nuclei extraction buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$,
833 0.1 mM EDTA, 0.5% IGEPAL, 1x cOmplete ULTRA Tablet (Roche)) and incubated on

834 ice for 10 min followed by 20 passes on a tight Dounce homogeniser, or until nuclei
835 were extracted. Nuclei were visually inspected using trypan blue and a
836 haemocytometer. Nuclei were pelleted by centrifugation and resuspended in
837 sonication buffer (50 mM Tris-HCl pH 8, 1% SDS, 10 mM EDTA, 1x cOmplete ULTRA
838 Tablet (Roche)) and sonicated using a Bioruptor sonicator (Diagenode) to achieve
839 DNA fragments between 100-500 bp with a mean fragment size of approximately
840 300 bp. In the siRNA experiment, protein was quantified using the Pierce BCA assay
841 to load equal amounts of input material for NT and ID4 siRNA for
842 immunoprecipitation. Sonicated sample was diluted with IP dilution buffer to 1 mL
843 (16.7 mM Tris-HCl pH 8, 0.01% SDS, 1% Triton X-100, 167 mM NaCl, 1.2 mM EDTA).
844 Samples were cleared with protein A/G magnetic beads for 1.5 hr at 4°C. 1% of the
845 cleared nuclear lysate was removed for the input control and immunoprecipitation
846 was performed as described previously. For HEB ChIP 100 µL of beads and 20 µg of
847 HEB antibody (sc-58052) was used with chromatin from approximately 20-30x10^6
848 cells. For histone mark ChIP, 50 µL beads with 10 µg of the following antibodies –
849 H3K4Me3 (Active Motif 28431), H3K27Me3 (Merck Millipore 07-449), was used with
850 chromatin from approximately 10-15x10^6 cells. The following day, beads were
851 washed for 5 min each with 1 mL of the following buffers: Low salt buffer (0.1% SDS,
852 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8, 150 mM NaCl), High salt buffer
853 (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8, 500 mM NaCl), LiCl
854 buffer (0.25 M LiCl, 1% IGEPAL, 1% deoxycholic acid (sodium salt), 1 mM EDTA, 10
855 mM Tris, pH 8) followed by 2 washes with TE buffer (10 mM Tris-HCl, 1 mM EDTA,
856 pH 8). DNA was eluted twice in 100 uL ChIP elution buffer (1% SDS, 0.1 M Sodium
857 Biocarbonate) for 15 minutes at room temperature each. Cross-linking was reversed

858 overnight by treating samples with 200 mM NaCl and 250 µg/mL Proteinase K (New
859 England Biolabs) and incubating overnight at 65°C. The following day samples were
860 treated with 100 µg/mL RNAse A (Qiagen) for 1 hr at 37°C. DNA was purified using
861 Phase Lock Gel Light tubes (Quantabio 5Prime) according to the manufacturer's
862 instructions. DNA was eluted in 20 uL nuclease-free TE buffer, pH8 (Qiagen).
863 DNA concentration was measured using the Qubit HS Assay Kit (Thermo Fisher
864 Scientific). Libraries were prepared using the Illumina TruSeq ChIP library prep kit
865 (Illumina) following the manufacturer's instructions except the gel purification step
866 was replaced with a two-sided AMPure XP bead (Beckman Coulter) size selection to
867 obtain libraries between 200-500 bp. Library sizes were verified using the 4200
868 TapeStation System (Agilent) with a D1000 ScreenTape (Agilent) and concentration
869 was determined using the Qubit HS Assay Kit (Thermo Fisher Scientific). ChIP
870 libraries were sequenced on the NextSeq system (Illumina), with 75 bp paired-end
871 reads.
872 Quality control was checked using FastQC to remove poor quality reads and to trim
873 Illumina sequencing adapters (Andrews, 2010). Reads were aligned using Bowtie for
874 Illumina (Langmead & Salzberg, 2012) followed by Sam-to-Bam conversion and
875 alignment using Bwa (Li, Handsaker et al., 2009). Alignment statistics were generated
876 using Samtools Flagstat (Li et al., 2009). ChIP-seq peaks were called using the peak
877 calling algorithm MACS (Zhang, Liu et al., 2008) and ENCODE blacklist regions were
878 removed. For unperturbed cells, consensus peaks present in at least 2 of 4 replicates
879 were used for downstream analysis. Due to lower amount of input material available
880 from the siRNA experiment, less consensus peaks were called and for this reason
881 analysis was conducted on merged peaks from the 3 replicates. Motif enrichment

882 analysis was performed using MEME-ChIP (Machanick & Bailey, 2011). Peaks were
883 annotated to genomic features using HOMER (Heinz, Benner et al., 2010). GREAT
884 was used for functional enrichment analysis (v4.0.4) and gene annotation using the
885 default parameters (McLean et al., 2010). SeqPlots (Stempor & Ahringer, 2016) and
886 IGV (Robinson, Thorvaldsdóttir et al., 2011) software were used for data
887 visualisation. Differential binding analysis was performed using the DiffBind package
888 (Stark, 2011). For integration of RNA-seq and ChIP-seq data a hypergeometric test
889 was used to determine if overlap was significant, assuming 30,000 genes in the
890 mouse genome.

891 Data availability

892 The mass spectrometry proteomics data have been deposited to the
893 ProteomeXchange Consortium via the PRIDE partner repository with the dataset
894 identifier PXD017517

895 Acknowledgements

896 We are very grateful for Amanda Khoury for her extensive advice on the ChIP-seq
897 experiments, Samantha Oakes for providing the mammary gland FFPE blocks from
898 different developmental stages, David Gallego Ortega for advice on tissue
899 dissociation and sorting and William Hughes for help with microscopy image
900 analysis. We thankfully acknowledge the support of the Australia Postgraduate
901 Award and The National Health and Medical Research Council of Australia
902 (1107671). Aspects of this research were supported by access to the Australian
903 Proteome Analysis Facility, funded by the Australian Government's National

904 Collaborative Research Infrastructure Scheme. T.R.C and J.N.S are supported by the
905 Cancer Institute NSW (171105) and a Career Catalyst grant from Susan G. Komen.

906 Author Contributions

907 H.H designed and performed the experiments and wrote the manuscript. D.R
908 analysed the RNA-seq and ChIP-seq data. S.J co-supervised the project, contributed
909 to experiments and discussions and extensively revised the manuscript. S.Z.W
910 analysed public scRNA-seq data and revised the manuscript. L.A.B advised on RIME
911 experiments and contributed to discussions. C.K and M.P.M performed mass
912 spectrometry and analysis. C.H performed RNA-seq library preps and analysis. A.M
913 assisted with mouse experiments. J.N.S and T.R.C performed picrosirius red image
914 analysis. B.P assisted with ChIP-seq experiments. N.H contributed IP antibodies. C.J.O
915 co-supervised the project and revised the manuscript. J.S.C helped to conceptualise
916 the project and revised the manuscript. J.V contributed the reporter mice and
917 revised the manuscript. A.S conceived the study and was in charge of overall
918 direction and planning.
919

920 Conflicts of Interest

921 NH has ownership and stock options in oNKo-Innate Pty Ltd.

922 Figure legends

923 **Figure 1. Loss of ID4 results in upregulation of myoepithelial genes and**
924 **downregulation of cell growth genes in sorted mammary basal cells.**
925 **a)** Live lineage negative $CD24^{+}CD29^{hi}CD61^{+}$ basal, $CD24^{+}CD29^{lo}CD61^{+}$ luminal

926 progenitor and $CD24^{+}CD29^{lo}CD61^{-}$ mature luminal cells were isolated by FACS from
927 adult (10-12 weeks) ID4 wild type (WT) and knockout (KO) mice at estrus for RNA-
928 seq. Representative FACS plots shown from 4 experiments. **b)** Number of
929 significantly (FDR<0.05) differentially expressed genes (DEGs) upregulated (red) or
930 downregulated (blue) in ID4 KO epithelial subpopulations compared to ID4 WT. LP =
931 luminal progenitor, ML = mature luminal. **c)** Heat map displaying the 237 significant
932 differentially expressed genes between ID4 WT and ID4 KO basal cells. **d)** Genes
933 were ranked based on the limma t-statistic comparing ID4 WT and KO basal cells and
934 GSEA was carried out using the C2all, C3TF, C5 and Hallmark gene sets. The top 10
935 positive and negatively enriched pathways with an FDR<0.1 are displayed. Only 3
936 pathways were negatively enriched with an FDR<0.1. **e)** GSEA results were visualised
937 using the EnrichmentMap plugin on Cytoscape. Nodes represent gene sets and
938 edges represent overlap. Red nodes are positively enriched and blue nodes are
939 negatively enriched. Gene sets with an FDR<0.25 are shown.
940
941 **Figure 2. ID4 marks basal stem cells and decreases in terminally differentiated**
942 **myoepithelial cells.**
943 **a)** FACS analysis of EPCAM and CD49f in live lin- mammary cells from adult (10-14
944 week) Id4floxGFP reporter mice. ID4-GFP expression is indicated in the heatmap
945 scale. Representative plots from 5 experiments shown. Basal cells were binned into 3
946 groups based on ID4-GFP expression, ID4-bright (red), ID4-intermediate (orange) and
947 ID4-dim (green) and the median fluorescence intensity (MFI) of EPCAM **(b)** and
948 CD49f **(c)** were compared between the 3 gates. MFI expressed as a fold change
949 relative to the ID4-low basal cells. Ordinary one-way ANOVA test was used to test

950 significance. n=5. Error bars represent SEM. ** p<0.01, *** p<0.001, **** p<0.0001.
951 **d)** Representative co-immunofluorescent staining of ID4 and α-SMA in TEB and duct
952 from a pubertal (6 week) mammary gland. Scale bar = 20 µm**.** Comparison of ID4 **(e)**
953 and α-SMA **(f)** mean fluorescence between individual cap cells (dark blue) and basal
954 duct cells (light blue). Unpaired two-tailed students t-test. **** p<0.0001.
955 Correlation between α-SMA and ID4 **(g)** and Hoescht **(h)** mean fluorescence in
956 individual cap cells (dark blue) and basal duct cells (light blue)**.** R^2 and p values are
957 displayed. Data is pooled from 9 mice. Approximately 20 TEB cap cells and 20 ductal
958 basal cells were analysed per mouse. **i)** tSNE plot of 9663 Krt5+/Krt14+ basal cells
959 from (Bach et al., 2017). 2 mice were analysed per developmental stage. NP =
960 Nulliparous (8 week), G= Gestation (Day 14.5), L = Lactation (Day 6), I = Involution
961 (Day 11). Feature plots **(j)** and Violin plots **(k)** displaying expression of *Id4, Epcam*
962 and *Acta2* in single cells in the different developmental stages. **l)** Heatmap displaying
963 top and bottom 30 differentially expressed genes between the top and bottom 200
964 *Id4* high and *Id4* low basal cells across all stages. Top 10 GO terms enriched in the
965 top 50 genes upregulated in *Id4* high **(m)** and low **(n)** basal cells.
966
967 **Figure 3. ID4 inhibits myoepithelial differentiation of organoids.**
968 **a)** Schematic diagram of 3D Matrigel organoid assay. ID4-GFP+ basal cells were FACS
969 purified from adult (10-11 weeks) ID4-GFP reporter mice. Exon 1 and 2 of *Id4* are
970 floxed and a GFP reporter cassette introduced. Basal cells are reprogrammed in
971 culture using ROCK inhibitor Y-27632 and irradiated NIH-3T3 feeder cells. Adenoviral
972 Cre is used to knock out ID4 as shown by western blotting. Single cells are then
973 seeded on top of a Matrigel plug and grown for 6 days followed by

immunofluorescent staining and quantification. Organoids grown from conditionally reprogrammed basal cells were treated with control GFP adenovirus (ID4 WT) or with Cre Adeno virus (ID4 KO) then stained for ID4 **(b)** and α-SMA **(d).** Scale bar = 10 µm. Fluorescence was quantified for ID4 **(c)** and α-SMA **(e)** in approximately 10 organoids per experiment. n=4. Unpaired two-tailed students t-test. *** $p<0.001$, **** $p<0.0001$.

Figure 4. ID4 negatively regulates EMT and ECM production in mammary epithelial cells.

a) Western blot analysis of ID4 expression in Comma-Dβ cells treated with non-targeting (NT) or ID4-targeting siRNA. Representative results from 3 western blots shown **(i)**. Densitometry quantification of ID4 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 . **** $p<0.0001$. Error bars represent SEM. **(ii)**. **b)** Gene Ontology (GO) GSEA results were visualised using the EnrichmentMap plugin on Cytoscape. Nodes represent gene sets and edges represent overlap. Red nodes are positively enriched and blue nodes are negatively enriched in siID4 cells compared to control NT cells. Gene sets with an FDR<0.25 are shown. **c)** Top 10 positively (red) and negatively (blue) enriched GO pathways. **d)** Representative GSEA enrichment plots displaying the profile of the running Enrichment Score (green) and positions of gene set members on the rank ordered list. **e)** Collagen fibres were visualised by picrosirius red staining of TEBs from 6-week-old ID4 WT and KO mice. Scale bar = 50 µm **(i).** Total collagen staining **(ii)** and birefringence signal **(iii)** were quantified from approximately 3 TEBs from each mammary gland section. N=9 for WT and N=6 for

998 KO mice. Unpaired two-tailed students t-test. * $p<0.05$. *** $p<0.001$. Error bars
999 represent SEM.
1000
1001 **Figure 5. ID4 interacts with E-proteins E2A and HEB.**
1002 **a)** Unsupervised hierarchical clustering heat map of SWATH RIME data from Comma-
1003 Dβ cells. Proteins with significantly higher abundance (p-value<0.05) in the ID4 IPs
1004 compared to IgG IPs are displayed. Log2 protein area was used to generate the
1005 heatmap. Data from 3 independent experiments shown, each with 2-3 technical
1006 replicates. **b)** Co-immunoprecipitation (co-IP) and western blotting of ID4, E2A and
1007 HEB and IgG negative controls from uncrosslinked Comma-Dβ lysates. **c)** Proximity
1008 ligation assay (PLA) in Comma-Dβ cells for ID4 and HEB or corresponding negative
1009 control IgG. The cytoskeleton is stained with phalloidin and nucleus with DAPI. Scale
1010 bar = 20 µm. High power insets are shown below. **d)** Quantification of PLA foci from
1011 6 random fields of view for each condition, each with approximately 50 nuclei per
1012 image. Ordinary one-way ANOVA test was used to test significance. **** $p<0.001$.
1013 N=3. **e)** Co-immunofluorescent staining of V5 and HEB in TEB (upper) and duct
1014 (lower) from ID4-FlagV5 mouse mammary glands. High-power insets feature cells
1015 positive for both proteins. Scale bar = 20 µm. **f)** co-IP and western blotting from ID4-
1016 FlagV5 mammary gland protein extracts. ID4 was immunoprecipitated using
1017 antibodies raised against ID4 and Flag, and two independent V5 antibodies. ID4, V5
1018 and HEB were detected by western blotting.
1019
1020 **Figure 6. HEB directly binds to a subset of ID4 target genes**
1021 **a)** Top 4 enriched transcription factor binding motifs determined using MEME-ChIP

1022 for consensus HEB ChIP-seq peaks in Comma-Dβ cells. E-values are displayed. **b)**
1023 GREAT pathway analysis of consensus HEB peaks. Top 16 Gene Ontologies (Biological
1024 process, cellular component, molecular function) are displayed. **c)** Venn diagram
1025 showing overlap between genes associated with a HEB peak (GREAT basal plus
1026 extension annotation) and ID4 differentially expressed genes. **d)** Profile plots of
1027 average HEB, H3K4Me3, H3K27Ac, and H3K27Me3 signal intensity at regions
1028 associated with downregulated (left) and upregulated (right) RNA-seq differentially
1029 expressed genes in siID4 compared to NT. **e)** Examples of HEB and histone mark
1030 peaks occurring upstream of siID4 upregulated genes *Sparc*, *Col1a1* and *Cnn2* from
1031 the Integrative Genomics Viewer (IGV). Bars beneath peaks represent consensus
1032 MACS call (FDR<0.05) in at least 2 of 4 biological replicates. Input was used as a
1033 negative control. Purple boxes highlight HEB binding regions. Refseq genes shown in
1034 blue. Data scales for each track are indicated. **f)** Volcano plot of Differential Binding
1035 Analysis. Analysis using EdgeR of HEB binding in siID4 verses NT of 3 biological
1036 replicates. Regions with an FDR less than 0.05 are indicated in pink. **g)** Profile plots of
1037 average HEB signal intensity in NT control (pink) and siID4 (purple) conditions at
1038 regions associated with RNA-seq siID4 downregulated (left), upregulated, (middle)
1039 and not regulated (right) genes. **h)** Model of ID4 and HEB action in mammary
1040 epithelial cells. Left: when ID4 is expressed it interacts with HEB, antagonising its
1041 transcriptional activity. Right: when ID4 is depleted, HEB dimerises and binds to E-
1042 box motifs in the promoters and enhancers of developmental genes involved in
1043 contraction and ECM. Below are hypothetical ChIP signals for H3K27Ac (blue),
1044 H3K4Me3 (green), and HEB (pink).

1045 ## Supplementary Figure legends
1046

1047 **Figure S1 related to Figure 1.**

1048 **a)** Proportions of basal, luminal progenitor (LP), and mature luminal (ML)
1049 subpopulations between ID4 WT and KO mammary epithelial cells. Unpaired two-
1050 tailed students t-test. ns = not significant. Error bars represent SEM. N=4.
1051 Representative GSEA enrichment plots displaying the profile of the running
1052 Enrichment Score (green) and positions of gene set members on the rank ordered
1053 list for pathways related to growth **(b)** and myoepithelial function **(c).**

1054

1055 **Figure S2 related to Figure 2.**

1056 **a)** Co-immunofluorescent staining of ID4 and α-SMA in a section containing both a
1057 TEB and a duct from a pubertal (6 week) mammary gland. Scale bar = 20 µm**. b)**
1058 Mouse mammary glands at different developmental stages were stained by IHC for
1059 ID4. Scale bar = 100 µm. Representative images from 3 animals per stage shown.
1060 High power insets are displayed below images. **c)** Differentiation trajectory of basal
1061 cells from (Bach et al., 2017) coloured by developmental stage (upper) and
1062 pseudotime (lower). Low values (dark blue) represent undifferentiated cells. NP =
1063 Nulliparous (8 week), G= Gestation (Day 14.5), L = Lactation (Day 6), I = Involution
1064 (Day 11). **d)** Expression of *Id4, Cnn1, Myh11, Acta2, Mylk* and *Oxtr* over pseudotime**.**

1065

1066 **Figure S3 related to Figure 3.**

1067 **a)** Phase contrast images of passage 11 conditionally reprogrammed cells grown in
1068 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 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 positive and negatively 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 related to Figure 5.

Co-immunoprecipitation and western blotting of ID4, E2A and HEB from human normal-like mammary epithelial cell lines PMC42 **(a)** and MCF10A cells overexpressing Flag-tagged ID4 protein **(b)**. **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 related to Figure 6.

a) Genomic distribution of HEB consensus peaks. TSS = transcriptional start site. TTS = transcriptional termination site. **b)** GREAT pathway analysis of closest mapped gene to HEB consensus regions also repressed by ID4 from Fig. 6c. Top 16 Gene Ontologies (Biological process, cellular component, molecular function) are displayed. **c)** Heatmaps of HEB, H3K4Me3, H3K27Ac, and H3K27Me3 ChIP-seq signal at HEB-bound regions. **d)** 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. **e)** Western blot analysis of ID4 and HEB expression in Comma-Dβ cells treated with NT or ID4-targeting siRNA. 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. **** p<0.0001. ns = not significant Error bars represent SEM. **f)** 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**. g)** Example of HEB peaks in control NT cells and siID4 cells upstream of *Col5a1* from the Integrative Genomics Viewer (IGV). Green boxes highlight HEB binding regions. Individual replicates and merged tracks are displayed. Refseq genes shown in blue. Data scales indicated. **h)** GREAT pathway analysis of 263 peaks increased in siID4 compared to NT from Fig. 7f. Top 16 Gene Ontologies (Biological process, cellular component, molecular function) are displayed.

1117 ## Supplementary Tables

1118 **Table S1.** All differentially expressed genes between ID4 wild type and knockout
1119 sorted mammary populations (Basal, luminal progenitor and mature luminal).

1120 **Table S2.** Top 50 differentially expressed genes between *Id4*-high and *Id4*-lo basal
1121 cells from Bach et al. single cell RNA-seq analysis.

1122 **Table S3.** Top 50 differentially expressed genes between siID4 and Non-targeting
1123 control Comma-Dβ cells.

1124 **Table S4.** Proteins identified in ID4 and HEB RIME experiments.

1125 **Table S5.** Genomic regions of Differentially bound HEB ChIP-Seq peaks between siID4
1126 and NT control Comma-Dβ cells.

1127 ## References

1128 Andrews S (2010) FastQC: A quality control tool for high throughput sequence data. .
1129 In
1130 Asselin-Labat ML, Sutherland KD, Barker H, Thomas R, Shackleton M, Forrest NC,
1131 Hartley L, Robb L, Grosveld FG, van der Wees J, Lindeman GJ, Visvader JE (2007)
1132 Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell
1133 differentiation. *Nature cell biology* 9: 201-9
1134 Bach K, Pensa S, Grzelak M, Hadfield J, Adams DJ, Marioni JC, Khaled WT (2017)
1135 Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA
1136 sequencing. *Nature communications* 8: 2128
1137 Baker LA, Holliday H, Swarbrick A (2016) ID4 Controls Mammary Lineage
1138 Commitment and Inhibits BRCA1 in Basal Breast Cancer. *Endocrine-related cancer*
1139 23: R381-92
1140 Barsky SH, Karlin NJ (2005) Myoepithelial cells: autocrine and paracrine suppressors
1141 of breast cancer progression. *Journal of mammary gland biology and neoplasia* 10:
1142 249-260
1143 Bayly R, Chuen L, Currie RA, Hyndman BD, Casselman R, Blobel GA, LeBrun DP (2004)
1144 E2A-PBX1 interacts directly with the KIX domain of CBP/p300 in the induction of
1145 proliferation in primary hematopoietic cells. *The Journal of biological chemistry* 279:
1146 55362-71
1147 Benezra R, Davis RL, Lockshon D, Turner DL, Weintraub H (1990) The protein Id: a
1148 negative regulator of helix-loop-helix DNA binding proteins. *Cell* 61: 49-59

1149 Best SA, Hutt KJ, Fu NY, Vaillant F, Liew SH, Hartley L, Scott CL, Lindeman GJ, Visvader
1150 JE (2014) Dual roles for Id4 in the regulation of estrogen signaling in the mammary
1151 gland and ovary. *Development (Cambridge, England)* 141: 3159-64
1152 Bouras T, Pal B, Vaillant F, Harburg G, Asselin-Labat ML, Oakes SR, Lindeman GJ,
1153 Visvader JE (2008) Notch signaling regulates mammary stem cell function and
1154 luminal cell-fate commitment. *Cell stem cell* 3: 429-41
1155 Braunstein M, Anderson MK (2012) HEB in the spotlight: Transcriptional regulation
1156 of T-cell specification, commitment, and developmental plasticity. *Clinical &*
1157 *developmental immunology* 2012: 678705
1158 Buchwalter G, Hickey MM, Cromer A, Selfors LM, Gunawardane RN, Frishman J,
1159 Jeselsohn R, Lim E, Chi D, Fu X, Schiff R, Brown M, Brugge JS (2013) PDEF promotes
1160 luminal differentiation and acts as a survival factor for ER-positive breast cancer
1161 cells. *Cancer cell* 23: 753-67
1162 Carr JR, Kiefer MM, Park HJ, Li J, Wang Z, Fontanarosa J, DeWaal D, Kopanja D,
1163 Benevolenskaya EV, Guzman G, Raychaudhuri P (2012) FoxM1 regulates mammary
1164 luminal cell fate. *Cell reports* 1: 715-29
1165 Chakrabarti R, Wei Y, Romano RA, DeCoste C, Kang Y, Sinha S (2012) Elf5 regulates
1166 mammary gland stem/progenitor cell fate by influencing notch signaling. *Stem cells*
1167 *(Dayton, Ohio)* 30: 1496-508
1168 Conway K, Pin C, Kiernan JA, Merrifield P (2004) The E protein HEB is preferentially
1169 expressed in developing muscle. *Differentiation; research in biological diversity* 72:
1170 327-340
1171 Dalal SJ, Estep JS, Valentin-Bon IE, Jerse AE (2001) Standardization of the Whitten
1172 Effect to induce susceptibility to Neisseria gonorrhoeae in female mice.
1173 *Contemporary topics in laboratory animal science / American Association for*
1174 *Laboratory Animal Science* 40: 13-7
1175 Danielson KG, Oborn CJ, Durban EM, Butel JS, Medina D (1984) Epithelial mouse
1176 mammary cell line exhibiting normal morphogenesis in vivo and functional
1177 differentiation in vitro. *Proceedings of the National Academy of Sciences of the*
1178 *United States of America* 81: 3756-60
1179 Davis FM, Lloyd-Lewis B, Harris OB, Kozar S, Winton DJ, Muresan L, Watson CJ (2016)
1180 Single-cell lineage tracing in the mammary gland reveals stochastic clonal dispersion
1181 of stem/progenitor cell progeny. *Nature communications* 7: 13053
1182 Deugnier MA, Faraldo MM, Teuliere J, Thiery JP, Medina D, Glukhova MA (2006)
1183 Isolation of mouse mammary epithelial progenitor cells with basal characteristics
1184 from the Comma-Dbeta cell line. *Developmental biology* 293: 414-25
1185 Deugnier MA, Moiseyeva EP, Thiery JP, Glukhova M (1995) Myoepithelial cell
1186 differentiation in the developing mammary gland: progressive acquisition of smooth
1187 muscle phenotype. *Developmental dynamics : an official publication of the American*
1188 *Association of Anatomists* 204: 107-17
1189 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
1190 Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford,*
1191 *England)* 29: 15-21
1192 Dong J, Huang S, Caikovski M, Ji S, McGrath A, Custorio MG, Creighton CJ, Maliakkal
1193 P, Bogoslovskaia E, Du Z, Zhang X, Lewis MT, Sablitzky F, Brisken C, Li Y (2011) ID4
1194 regulates mammary gland development by suppressing p38MAPK activity.
1195 *Development (Cambridge, England)* 138: 5247-56

Ewald AJ, Brenot A, Duong M, Chan BS, Werb Z (2008) Collective epithelial migration and cell rearrangements drive mammary branching morphogenesis. *Developmental cell* 14: 570-581

Feinberg TY, Zheng H, Liu R, Wicha MS, Yu SM, Weiss SJ (2018) Divergent Matrix-Remodeling Strategies Distinguish Developmental from Neoplastic Mammary Epithelial Cell Invasion Programs. *Developmental cell* 47: 145-160.e6

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* 16: 278

Gudjonsson T, Adriance MC, Sternlicht MD, Petersen OW, Bissell MJ (2005) Myoepithelial cells: their origin and function in breast morphogenesis and neoplasia. *Journal of mammary gland biology and neoplasia* 10: 261-72

Gudjonsson T, Rønnov-Jessen L, Villadsen R, Rank F, Bissell MJ, Petersen OW (2002) Normal and tumor-derived myoepithelial cells differ in their ability to interact with luminal breast epithelial cells for polarity and basement membrane deposition. *Journal of cell science* 115: 39-50

Guo W, Keckesova Z, Donaher JL, Shibue T, Tischler V, Reinhardt F, Itzkovitz S, Noske A, Zurrer-Hardi U, Bell G, Tam WL, Mani SA, van Oudenaarden A, Weinberg RA (2012) Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* 148: 1015-28

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

Huang Q, Yang L, Luo J, Guo L, Wang Z, Yang X, Jin W, Fang Y, Ye J, Shan B, Zhang Y (2015) SWATH enables precise label-free quantification on proteome scale. *Proteomics* 15: 1215-23

Idoux-Gillet Y, Nassour M, Lakis E, Bonini F, Theillet C, Du Manoir S, Savagner P (2018) Slug/Pcad pathway controls epithelial cell dynamics in mammary gland and breast carcinoma. *Oncogene* 37: 578-588

Ingthorsson S, Hilmarsdottir B, Kricker J, Magnusson MK, Gudjonsson T (2015) Context-Dependent Function of Myoepithelial Cells in Breast Morphogenesis and Neoplasia. *Current molecular biology reports* 1: 168-174

Junankar S, Baker LA, Roden DL, Nair R, Elsworth B, Gallego-Ortega D, Lacaze P, Cazet A, Nikolic I, Teo WS, Yang J, McFarland A, Harvey K, Naylor MJ, Lakhani SR, Simpson PT, Raghavendra A, Saunus J, Madore J, Kaplan W et al. (2015) ID4 controls mammary stem cells and marks breast cancers with a stem cell-like phenotype. *Nature communications* 6: 6548

Khoury A, Achinger-Kawecka J, Bert SA, Smith GC, French HJ, Luu P-L, Peters TJ, Du Q, Parry AJ, Valdes-Mora F, Taberlay PC, Stirzaker C, Statham AL, Clark SJ (2020) Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nature communications* 11: 54-54

Kiemer AK, Takeuchi K, Quinlan MP (2001) Identification of genes involved in epithelial-mesenchymal transition and tumor progression. *Oncogene* 20: 6679-6688

Kouros-Mehr H, Slorach EM, Sternlicht MD, Werb Z (2006) GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell* 127: 1041-55

Kumar MS, Hendrix JA, Johnson AD, Owens GK (2003) Smooth muscle alpha-actin gene requires two E-boxes for proper expression in vivo and is a target of class I basic helix-loop-helix proteins. *Circulation research* 92: 840-7

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9: 357-359

Law AMK, Yin JXM, Castillo L, Young AIJ, Piggin C, Rogers S, Caldon CE, Burgess A, Millar EKA, O'Toole SA, Gallego-Ortega D, Ormandy CJ, Oakes SR (2017) Andy's Algorithms: new automated digital image analysis pipelines for FIJI. *Scientific reports* 7: 15717

Lee K, Gjorevski N, Boghaert E, Radisky DC, Nelson CM (2011) Snail1, Snail2, and E47 promote mammary epithelial branching morphogenesis. *The EMBO journal* 30: 2662-74

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12: 323

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25: 2078-2079

Li S, Chang S, Qi X, Richardson JA, Olson EN (2006) Requirement of a myocardin-related transcription factor for development of mammary myoepithelial cells. *Molecular and cellular biology* 26: 5797-808

Li Y, Brauer PM, Singh J, Xhiku S, Yoganathan K, Zúñiga-Pflücker JC, Anderson MK (2017) Targeted Disruption of TCF12 Reveals HEB as Essential in Human Mesodermal Specification and Hematopoiesis. *Stem cell reports* 9: 779-795

Lilja AM, Rodilla V, Huyghe M, Hannezo E, Landragin C, Renaud O, Leroy O, Rulands S, Simons BD, Fre S (2018) Clonal analysis of Notch1-expressing cells reveals the existence of unipotent stem cells that retain long-term plasticity in the embryonic mammary gland. *Nature cell biology* 20: 677-687

Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE (2010) Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast cancer research : BCR* 12: R21

Liu J, Eischeid AN, Chen XM (2012) Col1A1 production and apoptotic resistance in TGF-β1-induced epithelial-to-mesenchymal transition-like phenotype of 603B cells. *PloS one* 7: e51371

Liu S, Ginestier C, Charafe-Jauffret E, Foco H, Kleer CG, Merajver SD, Dontu G, Wicha MS (2008) BRCA1 regulates human mammary stem/progenitor cell fate. *Proceedings of the National Academy of Sciences of the United States of America* 105: 1680-5

Liu X, Ory V, Chapman S, Yuan H, Albanese C, Kallakury B, Timofeeva OA, Nealon C, Dakic A, Simic V, Haddad BR, Rhim JS, Dritschilo A, Riegel A, McBride A, Schlegel R (2012) ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. *The American journal of pathology* 180: 599-607

Lloyd-Lewis B, Davis FM, Harris OB, Hitchcock JR, Watson CJ (2018) Neutral lineage tracing of proliferative embryonic and adult mammary stem/progenitor cells. *Development (Cambridge, England)* 145

Lourenço AR, Roukens MG, Seinstra D, Frederiks CL, Pals CE, Vervoort SJ, Margarido AS, van Rheenen J, Coffer PJ (2020) C/EBPα is crucial determinant of epithelial

maintenance by preventing epithelial-to-mesenchymal transition. *Nature communications* 11: 785
Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics (Oxford, England)* 27: 1696-1697
Macias H, Hinck L (2012) Mammary gland development. *Wiley interdisciplinary reviews Developmental biology* 1: 533-57
Massari ME, Murre C (2000) Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Molecular and cellular biology* 20: 429-40
McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research* 40: 4288-97
McCloy RA, Rogers S, Caldon CE, Lorca T, Castro A, Burgess A (2014) Partial inhibition of Cdk1 in G 2 phase overrides the SAC and decouples mitotic events. *Cell cycle (Georgetown, Tex)* 13: 1400-12
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* 28: 495-501
Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one* 5: e13984
Mesman S, Smidt MP (2017) Tcf12 Is Involved in Early Cell-Fate Determination and Subset Specification of Midbrain Dopamine Neurons. *Frontiers in Molecular Neuroscience* 10
Miano JM, Long X, Fujiwara K (2007) Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus. *American journal of physiology Cell physiology* 292: C70-C81
Mills AA, Zheng B, Wang XJ, Vogel H, Roop DR, Bradley A (1999) p63 is a p53 homologue required for limb and epidermal morphogenesis. *Nature* 398: 708-13
Mohammed H, D'Santos C, Serandour AA, Ali HR, Brown GD, Atkins A, Rueda OM, Holmes KA, Theodorou V, Robinson JL, Zwart W, Saadi A, Ross-Innes CS, Chin SF, Menon S, Stingl J, Palmieri C, Caldas C, Carroll JS (2013) Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. *Cell reports* 3: 342-9
Muschler J, Streuli CH (2010) Cell-matrix interactions in mammary gland development and breast cancer. *Cold Spring Harbor perspectives in biology* 2: a003202
Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van de Rijn M (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet (London, England)* 359: 1301-7
Oakes SR, Naylor MJ, Asselin-Labat ML, Blazek KD, Gardiner-Garden M, Hilton HN, Kazlauskas M, Pritchard MA, Chodosh LA, Pfeffer PL, Lindeman GJ, Visvader JE, Ormandy CJ (2008) The Ets transcription factor Elf5 specifies mammary alveolar cell fate. *Genes & development* 22: 581-6
Paine IS, Lewis MT (2017) The Terminal End Bud: the Little Engine that Could. *Journal of mammary gland biology and neoplasia* 22: 93-108
Pervolarakis N, Nguyen Q, Gutierrez G, Sun P, Jhutty D, Zheng GX, Nemec CM, Dai X, Watanabe K, Kessenbrock K (2019) Integrated single-cell transcriptomics and

chromatin accessibility analysis reveals novel regulators of mammary epithelial cell identity. *bioRxiv*: 740746

Prater MD, Petit V, Alasdair Russell I, Giraddi RR, Shehata M, Menon S, Schulte R, Kalajzic I, Rath N, Olson MF, Metzger D, Faraldo MM, Deugnier MA, Glukhova MA, Stingl J (2014) Mammary stem cells have myoepithelial cell properties. *Nature cell biology* 16: 942-50, 1-7

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* 14: 979-982

Rios AC, Fu NY, Lindeman GJ, Visvader JE (2014) In situ identification of bipotent stem cells in the mammary gland. *Nature* 506: 322-7

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43: e47

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. *Nature biotechnology* 29: 24-26

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26: 139-40

Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* 33: 495-502

Scheele CL, Hannezo E, Muraro MJ, Zomer A, Langedijk NS, van Oudenaarden A, Simons BD, van Rheenen J (2017) Identity and dynamics of mammary stem cells during branching morphogenesis. *Nature* 542: 313-317

Shackleton M, Vaillant F, Simpson KJ, Stingl J, Smyth GK, Asselin-Labat ML, Wu L, Lindeman GJ, Visvader JE (2006) Generation of a functional mammary gland from a single stem cell. *Nature* 439: 84-8

Silberstein GB, Strickland P, Coleman S, Daniel CW (1990) Epithelium-dependent extracellular matrix synthesis in transforming growth factor-beta 1-growth-inhibited mouse mammary gland. *The Journal of cell biology* 110: 2209-19

Stark RB, G. D. (2011) DiffBind: differential binding analysis of ChIP-seq peak data. In Bioconductor:

Stempor P, Ahringer J (2016) SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics [version 1; peer review: 2 approved, 1 approved with reservations]. *Wellcome Open Research* 1

Sternlicht MD (2006) Key stages in mammary gland development: the cues that regulate ductal branching morphogenesis. *Breast cancer research : BCR* 8: 201

Sternlicht MD, Kedeshian P, Shao ZM, Safarians S, Barsky SH (1997) The human myoepithelial cell is a natural tumor suppressor. *Clinical cancer research : an official journal of the American Association for Cancer Research* 3: 1949-58

Stingl J, Eirew P, Ricketson I, Shackleton M, Vaillant F, Choi D, Li HI, Eaves CJ (2006) Purification and unique properties of mammary epithelial stem cells. *Nature* 439: 993-7

Sun Y, Boyd K, Xu W, Ma J, Jackson CW, Fu A, Shillingford JM, Robinson GW, Hennighausen L, Hitzler JK, Ma Z, Morris SW (2006) Acute myeloid leukemia-associated Mkl1 (Mrtf-a) is a key regulator of mammary gland function. *Molecular and cellular biology* 26: 5809-26

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* 32: 381-386

van Amerongen R, Bowman AN, Nusse R (2012) Developmental stage and time dictate the fate of Wnt/beta-catenin-responsive stem cells in the mammary gland. *Cell stem cell* 11: 387-400

Van Keymeulen A, Rocha AS, Ousset M, Beck B, Bouvencourt G, Rock J, Sharma N, Dekoninck S, Blanpain C (2011) Distinct stem cells contribute to mammary gland development and maintenance. *Nature* 479: 189-93

Vennin C, Chin VT, Warren SC, Lucas MC, Herrmann D, Magenau A, Melenec P, Walters SN, Del Monte-Nieto G, Conway JRW, Nobis M, Allam AH, McCloy RA, Currey N, Pinese M, Boulghourjian A, Zaratzian A, Adam AAS, Heu C, Nagrial AM et al. (2017) Transient tissue priming via ROCK inhibition uncouples pancreatic cancer progression, sensitivity to chemotherapy, and metastasis. *Science translational medicine* 9: eaai8504

Vennin C, Mélénec P, Rouet R, Nobis M, Cazet AS, Murphy KJ, Herrmann D, Reed DA, Lucas MC, Warren SC, Elgundi Z, Pinese M, Kalna G, Roden D, Samuel M, Zaratzian A, Grey ST, Da Silva A, Leung W, Australian Pancreatic Genome I et al. (2019) CAF hierarchy driven by pancreatic cancer cell p53-status creates a pro-metastatic and chemoresistant environment via perlecan. *Nature communications* 10: 3637-3637

Visvader JE, Clevers H (2016) Tissue-specific designs of stem cell hierarchies. *Nature cell biology* 18: 349-55

Wang D, Cai C, Dong X, Yu QC, Zhang XO, Yang L, Zeng YA (2015) Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* 517: 81-4

Wang LH, Baker NE (2015) E Proteins and ID Proteins: Helix-Loop-Helix Partners in Development and Disease. *Developmental cell* 35: 269-80

Watanabe K, Villarreal-Ponce A, Sun P, Salmans ML, Fallahi M, Andersen B, Dai X (2014) Mammary morphogenesis and regeneration require the inhibition of EMT at terminal end buds by Ovol2 transcriptional repressor. *Developmental cell* 29: 59-74

Williams JM, Daniel CW (1983) Mammary ductal elongation: differentiation of myoepithelium and basal lamina during branching morphogenesis. *Developmental biology* 97: 274-90

Wojnarowicz PM, Lima E Silva R, Ohnaka M, Lee SB, Chin Y, Kulukian A, Chang S-H, Desai B, Garcia Escolano M, Shah R, Garcia-Cao M, Xu S, Kadam R, Goldgur Y, Miller MA, Ouerfelli O, Yang G, Arakawa T, Albanese SK, Garland WA et al. (2019) A Small-Molecule Pan-Id Antagonist Inhibits Pathologic Ocular Neovascularization. *Cell reports* 29: 62-75.e7

Wuidart A, Ousset M, Rulands S, Simons BD, Van Keymeulen A, Blanpain C (2016) Quantitative lineage tracing strategies to resolve multipotency in tissue-specific stem cells. *Genes & development* 30: 1261-77

Wuidart A, Sifrim A, Fioramonti M, Matsumura S, Brisebarre A, Brown D, Centonze A, Dannau A, Dubois C, Van Keymeulen A, Voet T, Blanpain C (2018) Early lineage segregation of multipotent embryonic mammary gland progenitors. *Nature cell biology* 20: 666-676

Yamaji D, Na R, Feuermann Y, Pechhold S, Chen W, Robinson GW, Hennighausen L (2009) Development of mammary luminal progenitor cells is controlled by the transcription factor STAT5A. *Genes & development* 23: 2382-7

Yang A, Schweitzer R, Sun D, Kaghad M, Walker N, Bronson RT, Tabin C, Sharpe A, Caput D, Crum C, McKeon F (1999) p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. *Nature* 398: 714-8

Yoon SJ, Foley JW, Baker JC (2015) HEB associates with PRC2 and SMAD2/3 to regulate developmental fates. *Nature communications* 6: 6546

Yun K, Mantani A, Garel S, Rubenstein J, Israel MA (2004) Id4 regulates neural progenitor proliferation and differentiation in vivo. *Development (Cambridge, England)* 131: 5441-8

Zhang J, Kalkum M, Yamamura S, Chait BT, Roeder RG (2004) E protein silencing by the leukemogenic AML1-ETO fusion protein. *Science (New York, NY)* 305: 1286-9

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9: R137-R137

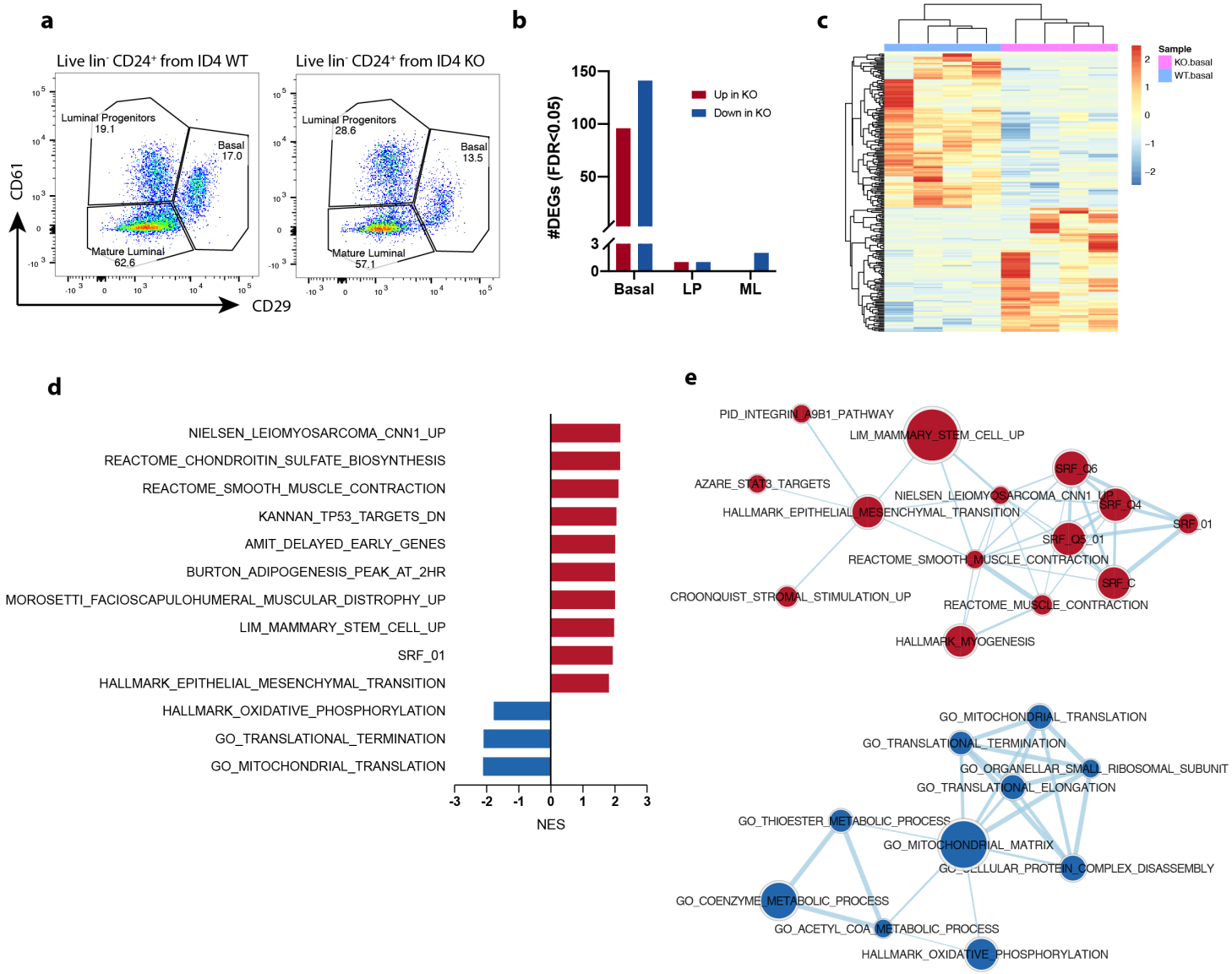

Figure 1. Loss of ID4 results in upregulation of myoepithelial genes and downregulation of cell growth genes in sorted mammary basal cells.
a) Live lineage negative CD24+CD29hiCD61+ basal, CD24+CD29loCD61+ luminal progenitor and CD24+CD29loCD61- mature luminal cells were isolated by FACS from adult (10-12 weeks) ID4 wild type (WT) and knockout (KO) mice at estrus for RNA-seq. Representative FACS plots shown from 4 experiments. b) Number of significantly (FDR<0.05) differentially expressed genes (DEGs) upregulated (red) or downregulated (blue) in ID4 KO epithelial subpopulations compared to ID4 WT. LP = luminal progenitor, ML = mature luminal. c) Heat map displaying the 237 significant differentially expressed genes between ID4 WT and ID4 KO basal cells. d) Genes were ranked based on the limma t-statistic comparing ID4 WT and KO basal cells and GSEA was carried out using the C2all, C3TF, C5 and Hallmark gene sets. 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. e) GSEA results were visualised using the EnrichmentMap plugin on Cytoscape. Nodes represent gene sets and edges represent overlap. Red nodes are positively enriched and blue nodes are negatively enriched. Gene sets with an FDR<0.25 are shown.

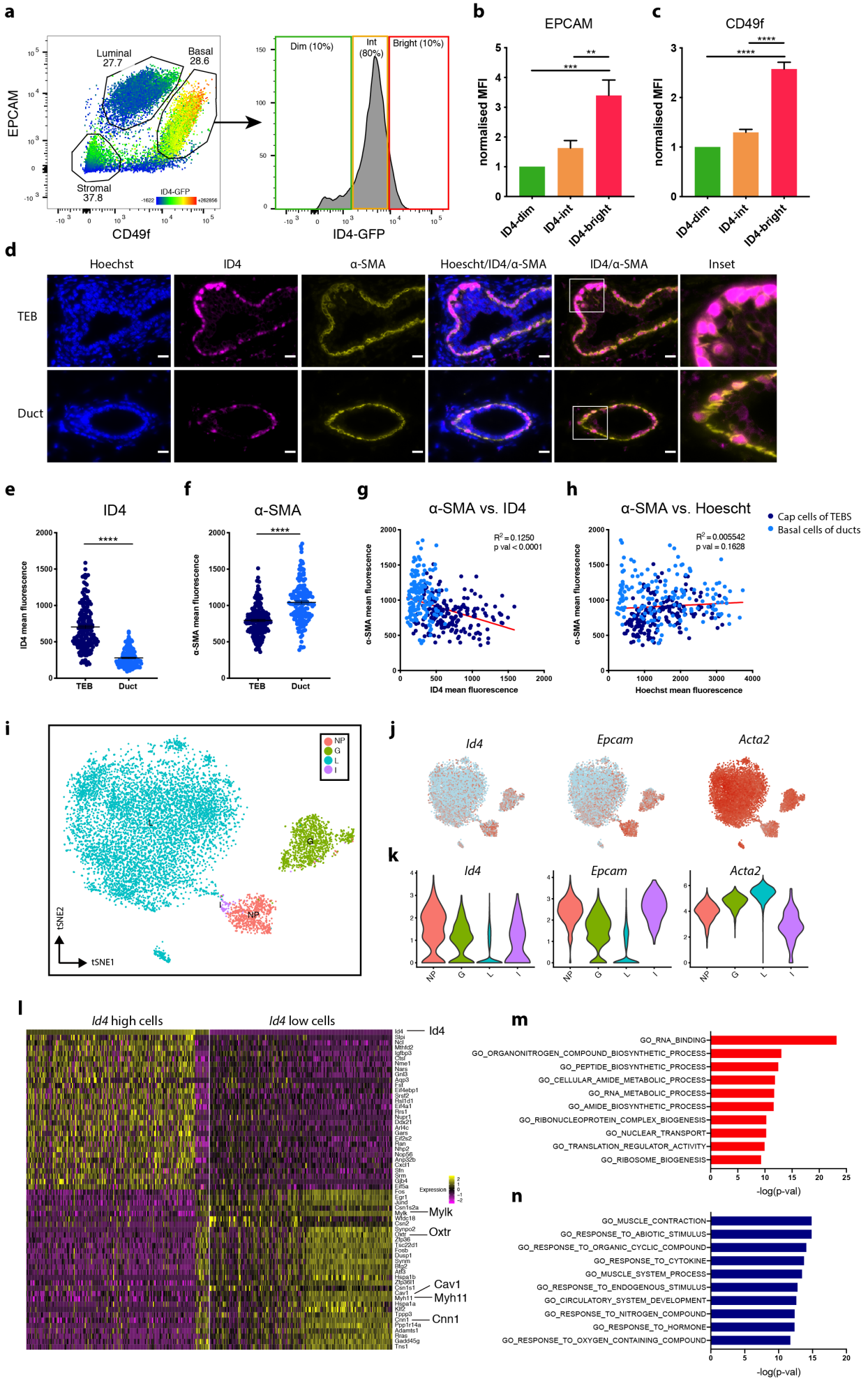

Figure 2. ID4 marks basal stem cells and decreases in terminally differentiated myoepithelial cells.

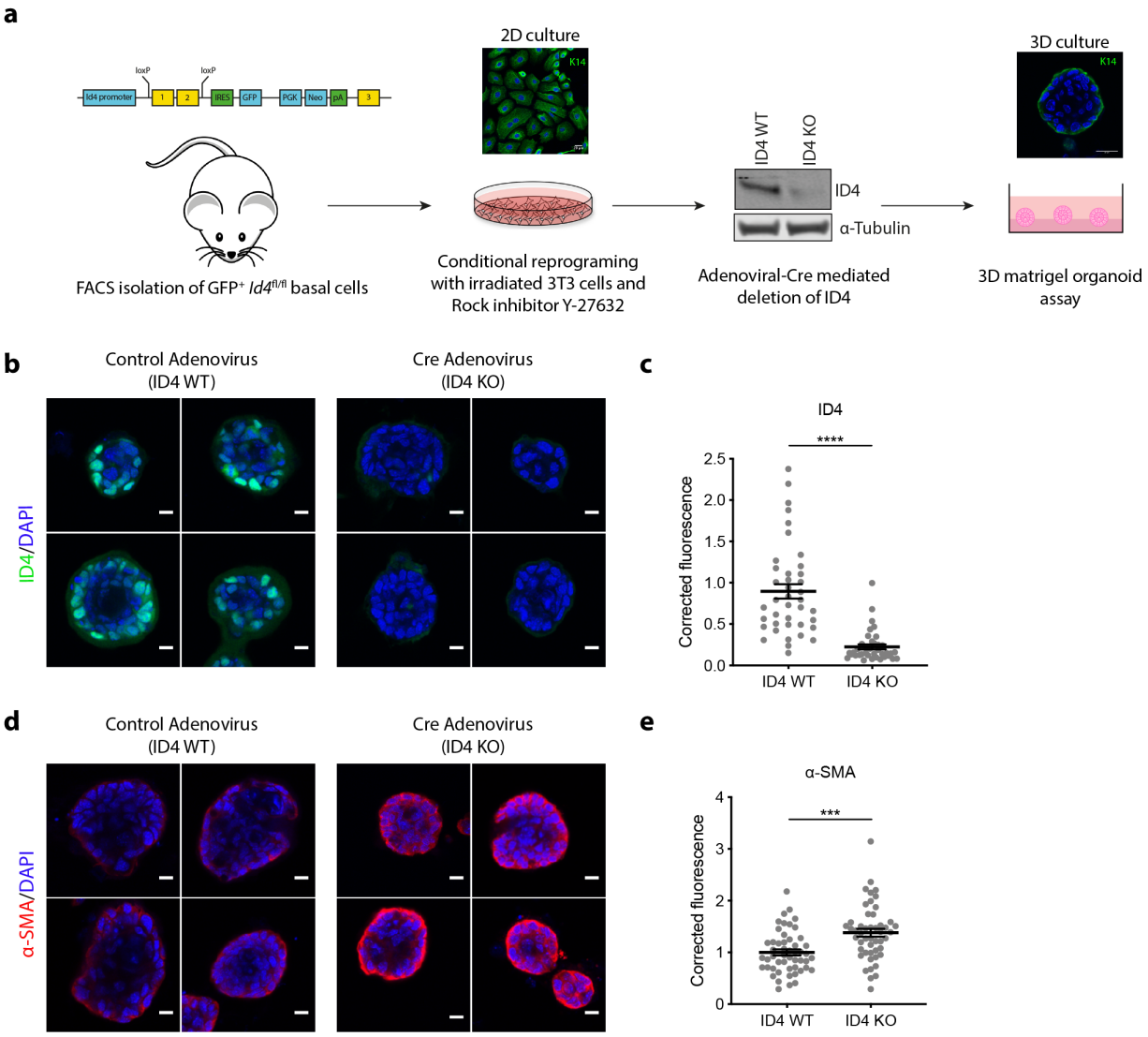

Figure 3. ID4 inhibits myoepithelial differentiation of organoids.
a) Schematic diagram of 3D Matrigel organoid assay. ID4-GFP+ basal cells were FACS purified from adult (10-11 weeks) ID4-GFP reporter mice. Exon 1 and 2 of Id4 are floxed and a GFP reporter cassette introduced. Basal cells are reprogrammed in culture using ROCK inhibitor Y-27632 and irradiated NIH-3T3 feeder cells. Adenoviral Cre is used to knock out ID4 as shown by western blotting. Single cells are then seeded on top of a Matrigel plug and grown for 6 days followed by immunofluorescent staining and quantification. Organoids grown from conditionally reprogrammed basal cells were treated with control GFP adenovirus (ID4 WT) or with Cre Adeno virus (ID4 KO) then stained for ID4 (b) and α-SMA (d). Scale bar = 10 µm. Fluorescence was quantified for ID4 (c) and α-SMA (e) in approximately 10 organoids per experiment. n=4. Unpaired two-tailed students t-test. *** $p<0.001$, **** $p<0.0001$.

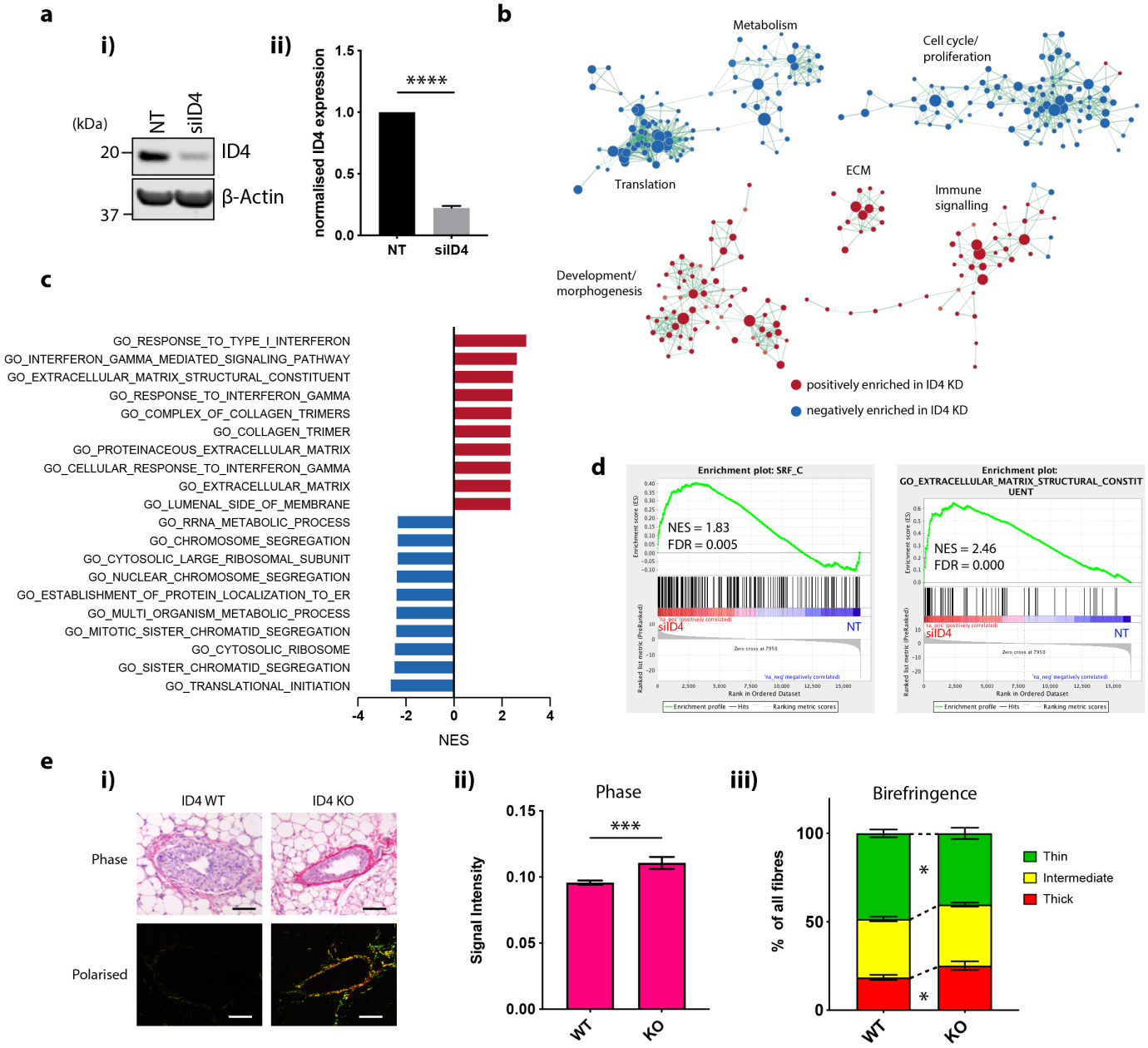

Figure 4. ID4 negatively regulates EMT and ECM production in mammary epithelial cells.
a) Western blot analysis of ID4 expression in Comma-Dβ cells treated with non-targeting (NT) or ID4-targeting siRNA. Representative results from 3 western blots shown (i). Densitometry quantification of ID4 bands. Band intensity was normalised to β-Actin and expressed as fold change relative to NT control. N=3. Unpaired two-tailed students t-test . **** $p<0.0001$. Error bars represent SEM. (ii). b) Gene Ontology (GO) GSEA results were visualised using the EnrichmentMap plugin on Cytoscape. Nodes represent gene sets and edges represent overlap. Red nodes are positively enriched and blue nodes are negatively enriched in siID4 cells compared to control NT cells. Gene sets with an FDR<0.25 are shown. c) Top 10 positively (red) and negatively (blue) enriched GO pathways. d) Representative GSEA enrichment plots displaying the profile of the running Enrichment Score (green) and positions of gene set members on the rank ordered list. e) Collagen fibres were visualised by picrosirius red staining of TEBs from 6-week-old ID4 WT and KO mice. Scale bar = 50 µm (i). Total collagen staining (ii) and birefringence signal (iii) were quantified from approximately 3 TEBs from each mammary gland section. N=9 for WT and 6 for KO mice. Unpaired two-tailed students t-test. * $p<0.05$. *** $p<0.001$. Error bars represent SEM.

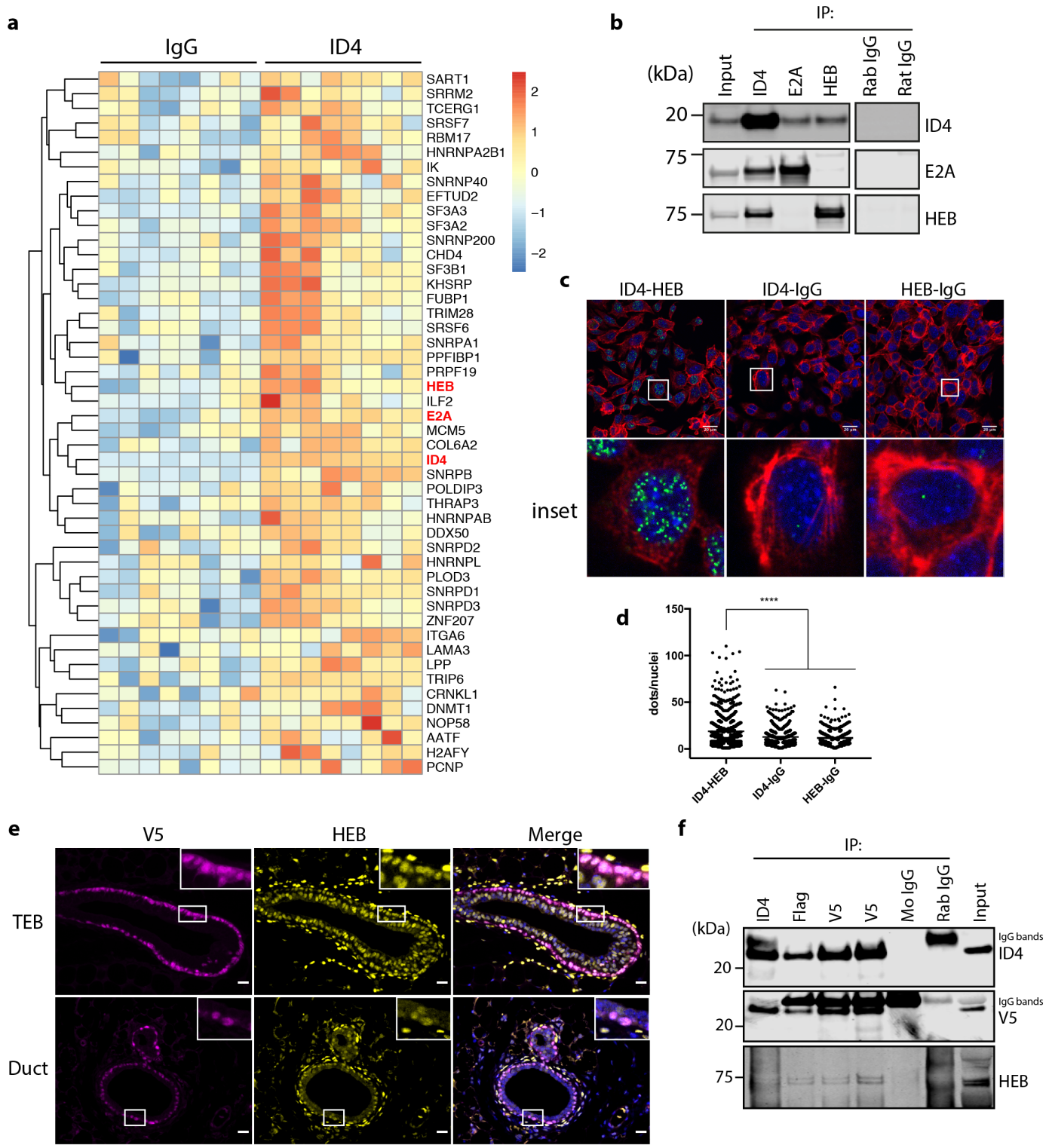

Figure 5. ID4 interacts with E-proteins E2A and HEB.
a) Unsupervised hierarchical clustering heat map of SWATH RIME data from Comma-Dβ cells. Proteins with significantly higher abundance (p-value<0.05) in the ID4 IPs compared to IgG IPs are displayed. Log2 protein area was used to generate the heatmap. Data from 3 independent experiments shown, each with 2-3 technical replicates. b) Co-immunoprecipitation (co-IP) and western blotting of ID4, E2A and HEB and IgG negative controls from uncrosslinked Comma-Dβ lysates. c) Proximity ligation assay (PLA) in Comma-Dβ cells for ID4 and HEB or corresponding negative control IgG. The cytoskeleton is stained with phalloidin and nucleus with DAPI. Scale bar = 20 μm. High power insets are shown below. d) Quantification of PLA foci from 6 random fields of view for each condition, each with approximately 50 nuclei per image. Ordinary one-way ANOVA test was used to test significance. **** p<0.001. N=3. e) Co-immunofluorescent staining of V5 and HEB in TEB (upper) and duct (lower) from ID4-FlagV5 mouse mammary glands. High-power insets feature cells positive for both proteins. Scale bar = 20 μm. f) co-IP and western blotting from ID4-FlagV5 mammary gland protein extracts. ID4 was immunoprecipitated using antibodies raised against ID4 and Flag, and two independent V5 antibodies. ID4, V5 and HEB were detected by western blotting.

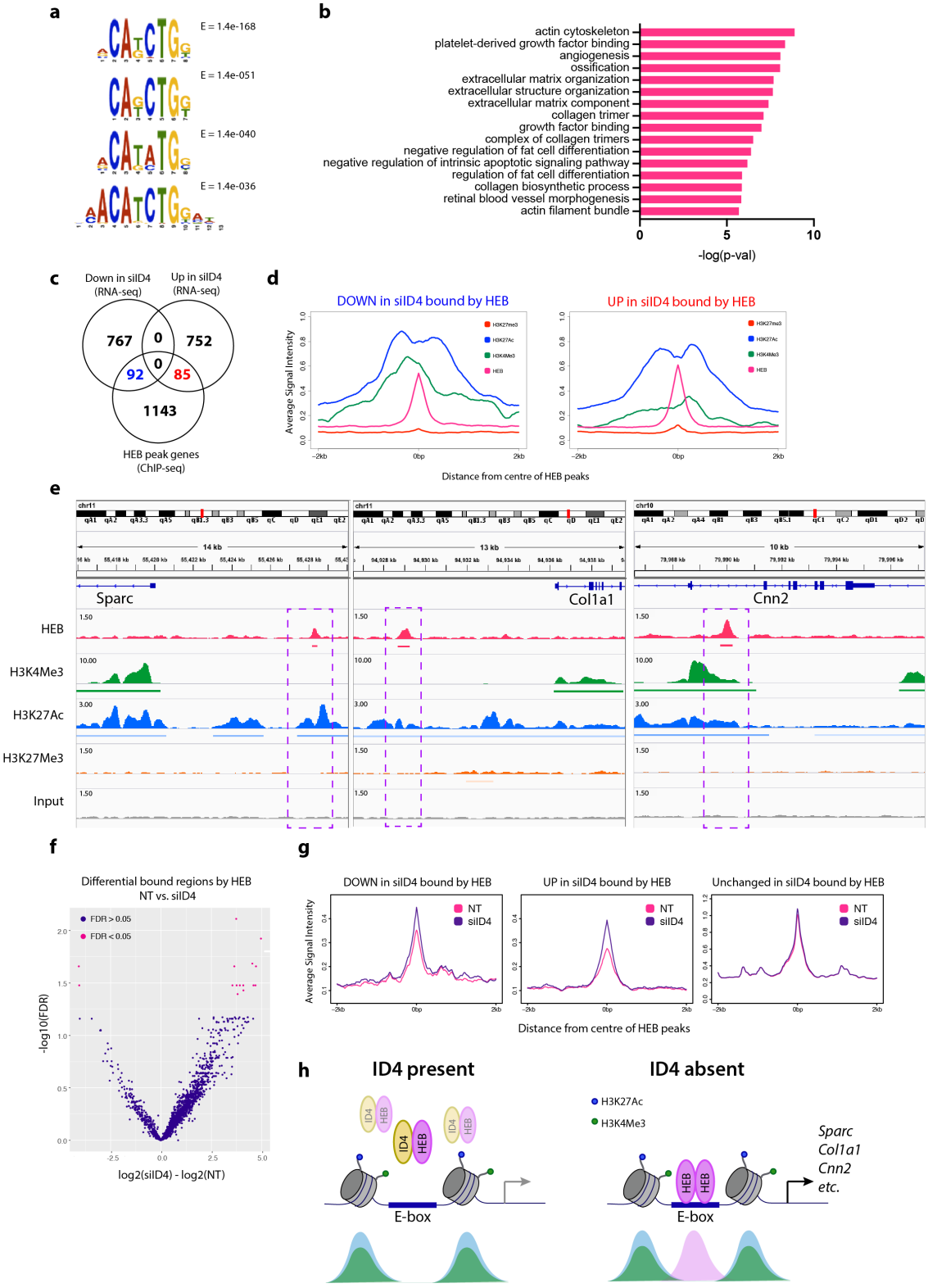

Figure 6. HEB directly binds to a subset of ID4 target genes

a) Top 4 enriched transcription factor binding motifs determined using MEME-ChIP for consensus HEB ChIP-seq peaks in Comma-Dβ cells. E-values are displayed. b) GREAT pathway analysis of consensus HEB peaks. Top 16 Gene Ontologies (Biological process, cellular component, molecular function) are displayed. c) Venn diagram showing overlap between genes associated with a HEB peak (GREAT basal plus extension annotation) and ID4 differentially expressed genes. d) Profile plots of average HEB, H3K4Me3, H3K27Ac, and H3K27Me3 signal intensity at regions associated with downregulated (left) and upregulated (right) RNA-seq differentially expressed genes in siID4 compared to NT. e) Examples of HEB and histone mark peaks occurring upstream of siID4 upregulated genes Sparc, Col1a1 and Cnn2 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. f) Volcano plot of Differential Binding Analysis. Analysis using EdgeR of HEB binding in siID4 verses NT of 3 biological replicates. Regions with an FDR less than 0.05 are indicated in pink. g) Profile plots of average HEB signal intensity in NT control (pink) and siID4 (purple) conditions at regions associated with RNA-seq siID4 downregulated (left), upregulated, (middle) and not regulated (right) genes. h) Model of ID4 and HEB action in mammary epithelial cells. Left: when ID4 is expressed it interacts with HEB, antagonising its transcriptional activity. Right: when ID4 is depleted, HEB dimerises and binds to E-box motifs in the promoters and enhancers of developmental genes involved in contraction and ECM. Below are hypothetical ChIP signals for H3K27Ac (blue), H3K4Me3 (green), and HEB (pink).