Identification of quiescent and spatially-restricted mammary stem cells that are hormone responsive

Nai Yang Fu,[1,2,8,9] Anne C. Rios,[1,2,8] Bhupinder Pal,[1,2] Charity W. Law,[2,3] Paul Jamieson[1], Ruijie Liu,[3] François Vaillant,[1,2] Felicity Jackling,[1] Kevin He Liu,[1] Gordon K. Smyth,[4,5] Geoffrey J. Lindeman,[1,6,7] Matthew E. Ritchie,[3,5] and Jane E. Visvader*,[1,2]

[1]ACRF Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia. [3]Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [4]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [5]School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia. [6]Parkville Familial Cancer Centre, The Royal Melbourne Hospital and Peter MacCallum Cancer Centre, Parkville, Victoria 3050, Australia. [7]Department of Medicine, The University of Melbourne, Parkville, Victoria 3010, Australia. [9]Present address: Cancer and Stem Cell Biology Program, Duke-NUS Medical School, Singapore 169857, Singapore.

[8] These authors contributed equally

*Correspondence: visvader@wehi.edu.au (J.E.V.)

SUMMARY

Despite accumulating evidence for a mammary differentiation hierarchy, the basal compartment comprising stem cells remains poorly characterized. Through gene expression profiling of $Lgr5^{+}$ basal epithelial cells, we identify a new marker Tetraspanin8 (Tspan8). Fractionation based on Tspan8 and Lgr5 expression uncovered three distinct mammary stem cell (MaSC) subsets in the adult mammary gland. These exist in a largely quiescent state but differ in their reconstituting ability, spatial localisation, and their molecular and epigenetic signatures. Interestingly, the deeply quiescent MaSC subset ($Lgr5^{+}Tspan8^{hi}$) resides within the proximal region throughout life, and has a transcriptome strikingly similar to that of claudin-low tumours. $Lgr5^{+}Tspan8^{hi}$ cells appear to originate from the embryonic mammary primordia before switching to a quiescent state post-natally but can be activated by ovarian hormones. Our findings reveal an unexpected degree of complexity within the adult MaSC compartment and identify a dormant subset poised for activation in response to physiological stimuli.

INTRODUCTION

The isolation and characterization of tissue-specific stem cells is fundamental to understanding organ development and homeostasis, as well as the perturbation of cellular architecture during oncogenesis. Analogous to the paradigm established by the hematopoietic and other systems, the mammary epithelium appears to be organized in a differentiation hierarchy[1]. The purification of MaSCs from the basal compartment (comprising stem, progenitor and myoepithelial cells), however, has been an immense challenge given the molecular similarities between its constituent cells[2].

Lgr5 is of considerable interest as it marks actively cycling stem cells in several epithelial tissues[3] but its relevance in the mammary gland remains highly controversial. In one study, only the $Lgr5^+$ population demonstrated potent regenerative activity[4], while another study reported the converse[5]. Instead, a subset of $Lgr5^-$ cells that expressed the Protein C receptor (*Procr*), was shown to comprise cycling, multipotent MaSCs. Other reports have indicated that both $Lgr5^+$ and $Lgr5^-$ cells harbour repopulating capacity[6,7]. Lineage tracing studies have also yielded discrepant data, through the findings that *Lgr5* can mark unipotent[6,8] or bipotent MaSCs[7]. The discrepancies can in part be explained by the requirement for performing transplantation studies at limiting dilution[6,7], the dosage of tamoxifen utilised, and the low frequency of mammary epithelial labelling evident in the Lgr5-GFP-IRES-creERT2 model[6,7].

Emerging evidence suggests that there may be more than one type of MaSC in the adult mammary gland. Label-retention studies have indicated the presence of slow-cycling MaSCs in the basal cell population based on PKH26-labelling of mammary epithelial cells in mammosphere cultures[9] or analysis of H2B-GFP mice[10]. DNA nucleotide analogue-retaining studies have also suggested that mammary epithelial cells may undergo asymmetric division

and retain their template-strand[11]. More recently, a subset of embryonic cells was shown to contribute to long-lived basal cells in the mammary gland but whether or not these correspond to MaSCs has not yet been established[12]. Thus, fundamental questions on the degree of intrinsic heterogeneity within the adult MaSC compartment and the existence of quiescent MaSCs remain.

Based on expression of *Lgr5* and the tetraspanin family member Tspan8, here we uncover distinct MaSC subsets that co-exist in a largely quiescent state in the adult mammary gland. Remarkably, these subsets reside in distinct locations along the mammary ductal tree, with deeply quiescent MaSCs restricted to the proximal region. Furthermore, this dormant population emanates from embryonic mammary cells, exhibits unique molecular properties and appears to serve as a stem cell reservoir that remains highly responsive to hormonal stimuli.

RESULTS

Identification of a cell surface marker Tspan8 expressed on Lgr5$^+$ basal cells

To further explore the utility of Lgr5 as a marker of mammary stem or progenitor cells, we performed 3D confocal imaging[7,13] of extensive portions of the epithelial tree from Lgr5-GFP-IRES-creER^{T2} mice[7,13]. During puberty, Lgr5-GFP$^+$ cells were visible in the nipple region and along the ducts but not within the proliferative TEBs that drive ductal elongation and branching (Fig. 1a). The majority of Lgr5-GFP$^+$ cells corresponded to non-proliferating myoepithelial cells. In adult tissue, *Lgr5* was exclusively expressed by myoepithelial cells scattered throughout the entire epithelial tree, based on whole-mount 3D confocal imaging and flow cytometric analyses (Fig. 1b-d; Supplementary Fig. 1a), consistent with previous findings[4,14].

To interrogate the molecular characteristics of $Lgr5^+$ cells in the adult mammary gland, we determined the gene expression profiles of $Lgr5^+$ versus $Lgr5^-$ cells sorted from the basal compartment ($Lin^-CD29^{hi}CD24^+$) of Lgr5-GFP-IRES-$creER^{T2}$ mammary glands (Fig. 1e). Substantial changes in the gene expression profiles were evident, with 215 upregulated genes and 206 downregulated genes in $Lgr5^+$ relative to $Lgr5^-$ basal cells (Supplementary Fig. 1b, c). Gene ontology-analysis of differentially expressed (DE) genes in $Lgr5^+$ cells revealed significant enrichment of pathways associated with stem cells, embryonic development and the negative regulation of non-canonical Wnt signaling (Supplementary Fig. 1d). Conversely, $Lgr5^-$ basal cells were enriched for genes involved with cell division, an extensive range of metabolic processes, nucleoside biosynthesis and RNA splicing, suggesting an active cycling status.

Tspan8 was identified as the top upregulated gene between $Lgr5^+$ versus $Lgr5^-$ basal cells (Fig. 1e). This gene encodes an integral plasma membrane protein comprising four transmembrane domains. As modulators of larger molecular complexes, the tetraspanin family regulates integrin compartmentalisation and recycling, and also influences cellular invasion and metastasis[15-17]. Indeed, flow cytometric analysis of mammary glands from adult Lgr5-GFP-IRES-$creER^{T2}$ females (9 weeks of age) using antibodies against Tspan8, CD29 and CD24 revealed that approx. 10% of cells in the basal compartment expressed Tspan8 (Fig.2a, b). Interestingly, Lgr5 and Tspan8 are not always co-expressed and four subsets could be distinguished: three small subpopulations corresponding to $Lgr5^+Tspan8^{hi}$, $Lgr5^-Tspan8^{hi}$ and $Lgr5^+Tspan8^-$, each of which constituted 4 - 6% of the total basal compartment, and a large population containing $Lgr5^-Tspan8^-$ cells. Of note, the stroma was devoid of Tspan8 expression. A subset of luminal cells (30-50%) also expressed Tspan8 but the level

was 10-fold lower than that seen in the basal compartment (Supplementary Fig. 2a, b). Further analysis of Tspan8 expression in the luminal population from *Elf5*-GFP reporter mice[7] showed that almost all Tspan8$^+$ luminal cells were Elf5-GFP$^+$, suggesting that Tspan8 marks committed luminal progenitor cells (Supplementary Fig. 2c).

Lgr5$^+$Tspan8^{hi} cells have potent repopulating activity *in vivo* and may lie at the apex of the MaSC hierarchy

Functional analysis of the four basal epithelial subsets defined by Lgr5 and Tspan8 expression revealed substantially different properties in clonogenic assays *in vitro* and repopulating assays *in vivo*. While all subsets generated comparable numbers of colonies in 3D Matrigel assays, the size and nature of these colonies differed markedly. Both Lgr5$^+$Tspan8^{hi} and Lgr5$^-$Tspan8^{hi} cells yielded larger colonies than the other two subsets, but the Lgr5$^+$Tspan8^{hi} subset had the most potent clonogenic potential since it produced heterotypic and semi-branched basal colonies (Fig. 2c, d). Similarly, upon transplantation into cleared fat pads, the Lgr5$^+$Tspan8^{hi} subset exhibited superior repopulating capacity, with a MRU frequency of approximately 1 in 16 for single-sorted cells (Table 1). The mammary reconstituting frequencies of the Lgr5$^-$Tspan8^{hi} and Lgr5$^+$Tspan8$^-$ subsets were comparable, but 4- to 5-fold lower than that of the Lgr5$^+$Tspan8^{hi} subpopulation. In contrast, Lgr5$^-$ Tspan8$^-$ cells had low regenerative potential, despite harbouring substantial *in vitro* clonogenic activity, consistent with enrichment for progenitor (and not stem) cells. At the lowest cell dose (≤50 cells), fat pad filling by primary outgrowths derived from Lgr5$^+$Tspan8^{hi} cells was ≥75%, compared to 25-50% for Lgr5$^-$Tspan8^{hi} and Lgr5$^+$Tspan8$^-$ cells and <25% for Lgr5$^-$Tspan8$^-$ cells (Supplementary Fig. 2d). Thus, the extent of fat pad filling was highest for the double-positive subset at the low cell dose, but transplantation of large numbers of cells yielded extensive outgrowths for the four subsets. Moreover, these

could undergo differentiation to milk-producing cells when subject to pregnancy (Supplementary Fig. 2e), and secondary transplantation of primary outgrowths from limiting numbers of cells demonstrated that all subpopulations could generate outgrowths in multiple recipients (Supplementary Fig. 2f).

Despite their *in vivo* repopulating activity, the ductal outgrowths generated by the different subsets were not identical (Fig. 2e). When a relatively large number of cells was transplanted, only the double-positive subpopulation yielded the normal repertoire of epithelial cells based on FACS analysis of primary outgrowths for expression of Lgr5-GFP and Tspan8 in the basal population ($Lin^{-}CD29^{hi}CD24^{+}$). Interestingly, the $Tspan8^{-}$ subpopulations generated dramatically fewer $Tspan8^{hi}$ cells in primary outgrowths, compared to those derived from the two $Tspan8^{hi}$ subsets (Fig. 2e). Conversely, $Lgr5^{-}$ cells could generate $Lgr5^{+}$ cells, indicating that expression of this gene does not occur in a uni-linear fashion. Taken together, these data reveal an unexpected degree of heterogeneity in the MaSC compartment and suggest that $Lgr5^{+}Tspan8^{hi}$ stem cells may lie at the top of the stem cell hierarchy since only $Tspan8^{hi}$ cells generated the entire repertoire of basal cells.

Identification of a highly quiescent MaSC population

We next investigated the cell cycle status of the four basal cell populations defined by Lgr5 and Tspan8 expression by performing FACS analysis for PyroninY (RNA content) and 7-AAD (DNA content) (Fig. 3a,b). Different profiles were observed: $Lgr5^{+}Tspan8^{hi}$ cells, $Lgr5^{-}Tspan8^{hi}$ and $Lgr5^{+}Tspan8^{hi}$ subsets comprised a high proportion of cells in the G0 phase (average of 86%, 70% and 65%, respectively), whereas the $Lgr5^{-}Tspan8^{-}$ subpopulation was the least quiescent. Analysis of luminal subsets revealed that more than 63% of cells were cycling in both the $Tspan8^{+}$ and $Tspan8^{-}$ subsets (Supplementary Fig. 2h), in contrast to the

basal subpopulations[18]. Together these data reveal the presence of three quiescent MaSC subsets in the steady-state adult mammary gland. Notably, the cell cycle status of the most quiescent Lgr5^{+}Tspan8^{hi} pool closely parallels that of other quiescent stem cells such as hematopoietic stem cells[19].

To probe molecular pathways that distinguish the different MaSC subpopulations, RNA-seq analysis was performed on freshly sorted cellular subsets. Each of the basal subsets exhibited strong basal character based on the expression of core myoepithelial genes (Supplementary Fig. 3a). The heat map of DE genes for the comparison of Lgr5^{+}Tspan8^{hi} cells versus all other subpopulations (Fig. 3c) highlights two important features: (1) the Lgr5^{+}Tspan8^{hi} subset differs profoundly from the other two basal subpopulations that also harbour repopulating potential, and (2) the Lgr5^{-}Tspan8^{hi} subpopulation has a molecular profile intermediate between that of Lgr5^{+}Tspan8^{hi} and Lgr5^{+}Tspan8^{-} cells. As anticipated, Lgr5^{-}Tspan8^{-} cells bear a gene expression signature distinct from the other three subsets.

Comparison of the gene expression profiles of Lgr5^{+}Tspan8^{hi} cells versus the average of the other three subsets showed 728 upregulated and 103 downregulated genes (Supplementary Fig. 3b,c). Gene ontology enrichment analysis of DE genes between the Lgr5^{+}Tspan8^{hi} and Lgr5^{+}Tspan8^{-} subpopulations revealed that downregulated genes in Lgr5^{+}Tspan8^{hi} cells primarily belong to functional groups for cell division, DNA replication and the DNA damage response (Supplementary Fig. 3d), consistent with findings for other quiescent stem cells[20]. For example, the cell cycle genes *cdk1*, *cdc25*, *cyclinB1* and *cyclinA2* were downregulated. Upregulated genes were predominantly associated with cell surface receptor signalling, cell communication and migration pathways (Supplementary Fig. 3d). Moreover, several inhibitors of Wnt signalling (including *Dkk2*, *Sfrp1*, *Sfrp2*, *Sfrp4*, *Sfrp5*, *Sox17*) were

markedly upregulated in the Lgr5^{+}Tspan8^{hi} (Fig. 3d), consistent with their non-cycling cell status and the established role of Wnt in promoting MaSC expansion[21].

To explore similarities between the molecular portraits of quiescent MaSCs and other tissue-specific stem cells, we interrogated two independent data-sets for quiescent muscle stem cells (MuSCs)[22,23], as well as those for hematopoietic stem cells (HSCs)[24] and quiescent hair follicle stem cells (HFSCs)[25]. Significant molecular similarities were observed between the signatures of these stem cells and the highly quiescent Lgr5^{+}Tspan8^{hi} subset in all comparisons performed (Fig. 3e). Closer scrutiny of the expression profiles of MuSCs versus quiescent MaSCs uncovered 104 and 51 genes shared between quiescent MaSCs and the Fukada *et al* and Liu *et al* datasets, which utilized different markers for cell sorting and therefore enriched for slightly different populations of MuSC-enriched cells (Fig. 3f). Interestingly, a common signature encompassing 26 genes included *Bmp4, Bmp6* and *Gli2,* all of which have been implicated in regulating stem cell quiescence[26,27].

Finally, the intrinsic subtypes of breast cancer[28] were interrogated with the gene expression signatures of the four basal subpopulations. A striking molecular correlation was evident, by signature expression scores ($p=0.00005$) and gene-set testing, between Lgr5^{+}Tspan8^{hi} cells and claudin-low cancers, which have been presumed to arise from MaSCs[29] (Supplementary Fig. 3e, f). None of the other basal/MaSC subsets showed significant similarity to this breast cancer subtype. Thus, claudin-low cancers retain the molecular signature of highly quiescent MaSCs, but the clinical relevance of this correlation remains to be determined.

Highly quiescent MaSCs are epigenetically distinct

To explore the relevance of histone modification to the regulation of gene expression in the different MaSC subsets, chromatin immunoprecipitation sequencing (ChIP–seq) for H3K4me3 and H3K27me3 modifications was performed on the four basal populations. In each subset, H3K4me3 occupancy typically peaked sharply around the transcriptional start site (TSS) of each gene and correlated significantly with gene expression[30], as shown for the Lgr5^{+}Tspan8^{hi} cellular subset (Fig. 4a, b). Conversely, repressive H3K27me3 marks were more evenly spread over the gene body and inversely correlated with gene expression. Interestingly, the highly quiescent subset showed a marked enrichment for H3K27me3 modifications relative to the other three subsets, compatible with widespread repression of gene expression (Fig. 4a, b).

The heatmap (Fig. 4c) shows the overall pattern of H3K4me3 and H3K27me3 histone modifications for the unique signature genes of Lgr5^{+}Tspan8^{hi} cells (versus all other populations), and reveals differing profiles amongst the four subsets. When the roast gene-set testing method[31] was applied to address modifications on DE genes, down-regulated genes were found to be preferentially bound by H3K27me3 mark, whereas up-regulated genes tended to be bound less by K27me3 modifications (roast p-value=0.003). The opposite was found to be true of H3K4me3 marks on DE genes (roast p-value=0.001). The Lgr5^{+}Tspan8^{hi} subset differed drastically from Lgr5^{-}Tspan8^{-} cells for both epigenetic marks, suggesting that histone methylation is a key mediator of gene expression changes. Illustrative read coverage graphs of H3K4me3 and H3K27me3 patterns across candidate genes (*Tspan8, Lgr5, Dclk1, Bmp6*) expressed in the unique gene expression signature of the most dormant subset are shown (Fig. 4d). These genes exhibit enrichment for H3K4me3 at the TSS and diminution of H3K27me3 marks across the wider gene body. Of note, Dclk1 was recently shown to define quiescent precursor cells in the pancreas[32].

Localisation of highly quiescent stem cells to the proximal ductal tree

The localisation of stem cells within the mammary ductal tree remains a key question. The mammary gland can be divided into distinct anatomical areas that include the proximal region (between the nipple and lymph node of the fourth gland) and a more distal region that extends beyond the lymph node to the fat pad edge (Fig. 5a). To quantify the distribution of the different MaSC subsets at different stages, we turned to FACS analysis of the Lin^- $CD29^{hi}CD24^+$ basal compartment in dissected portions of mammary glands from Lgr5-GFP-IRES-creER^{T2} mice. At 5 weeks (puberty), when the ductal tree and TEBs have substantially penetrated the distal portion of the fat pad, $Tspan8^{hi}$ basal cells were only detectable in the proximal region, with negligible $Lgr5^+Tspan8^{hi}$ or $Lgr5^-Tspan8^{hi}$ cells in the distal area (Fig. 5b,c). By contrast, $Tspan8^+$ luminal cells were clearly resident in both the proximal and distal regions of the ductal tree (Supplementary Fig. 4a). Moreover, $Tspan8^{hi}$ basal cells remain confined to this region in older adult females (9 months) and over three cycles of pregnancy (Supplementary Figure 4b). Notably, dissection of the mammary gland into multiple pieces confirmed that $Tspan8^{hi}$ cells span the entire proximal region, with a higher percentage in the nipple area (Supplementary Fig. 4c).

Immunofluorescence staining for Tspan8 expression using newly generated rat anti-Tspan8 monoclonal antibodies, confirmed that Tspan8 was restricted to basal cells lining the ducts in the proximal but not the distal region, nor was it expressed within the TEBs of pubertal glands (Fig. 5d,e). Not all $p63^+$ basal cells in the proximal region expressed Tspan8, consistent with FACS data (Fig. 2a). Collectively, these results indicate that quiescent $Tspan8^{hi}$ MaSCs reside in the proximal area of the gland throughout post-natal development whereas a distinct pool of $Lgr5^+Tspan8^-$ cells localise to the distal area. Of note, two pools of

Lgr5^{+} cells are apparent in the adult mammary gland, one largely restricted to the proximal region and a second pool that lies along the distal branches.

Relationship between embryonic label-retaining cells and adult MaSCs

Given the proximal localisation of double-positive cells within the adult ductal tree, we next explored the expression of Lgr5 and Tspan8 in the mammary primordia. The vast majority of embryonic mammary cells expressed *Lgr5* from E16.5 to E18.5 (Fig. 6a; Movie 1). At E18.5, co-expression of luminal (K8/K18) and myoepithelial (K5) markers was evident on many Lgr5-GFP^{+} cells within the mammary primordia, particularly within the developing branches where more than 70% of cells expressed Lgr5 and the keratin genes (Supplementary Fig. 5a-f). These data are compatible with the basal and luminal lineages not yet being specified at early stages of development. Comparative flow cytometric analysis of fetal skin versus mammary rudiments was performed, as it is technically impossible to dissect the fetal mammary rudiments without any contamination from skin. CD24^{hi} cells were identified as mammary epithelial cells and shown to be Lgr5-GFP^{+}. FACS analysis of mammary rudiments from Lgr5-GFP-IRES-creER^{T2} female embryos at E18.5 revealed a discrete population (~10%) of embryonic CD29^{+}CD24^{hi} cells that expressed Lgr5 and Tspan8 (Supplementary Fig. 5g, h). Short-term EdU-labelling during late embryogenesis further revealed that approximately 20% of Lgr5^{+}Tspan8^{+} cells were actively dividing at E16.5-17.5 (Fig. 6c), consistent with the dramatically higher clonogenic activity of fetal MaSCs *in vitro*[33,34]. Notably, *Lgr5* expression was significantly higher in embryonic versus adult mammary tissue (Supplementary Fig. 5i). Immunofluorescence staining for Tspan8 showed expression in a high proportion of cells in both the inner and outer (p63^{+}) layers of the primitive ducts (Supplementary Fig. 6a).

To examine the contribution of *Lgr5*-expressing primordial cells to the adult epithelial tree, we pulsed Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato dams with a single low dose of tamoxifen at E17.5 and analysed mammary glands 11 weeks later in adulthood. Tomato^{+} cells contributed substantially to both the luminal and basal lineages (Supplementary Fig. 6b, c), but it remains to be determined whether these derive from bipotent or unipotent cells. As embryonic cells appear largely uncommitted to a specific lineage, it is possible that many of these cells have bipotent capacity.

To determine the relationship between the adult MaSC subsets and embryonic mammary cells, we performed EdU label-retention studies. Lgr5-GFP-IRES-creER^{T2} dams were injected between E14.5-18.5 and then evaluated 6 weeks later. FACS analysis showed that the highest proportion of label-retaining cells was found in the Lgr5^{+}Tspan8^{hi} population (Fig. 6b,c). Furthermore, whole-mount confocal imaging in 3D demonstrated that virtually all EdU^{+} cells remained in the proximal region of Lgr5-GFP-IRES-creER^{T2} glands at 6-weeks post-EdU labelling in late embryogenesis (Supplementary Fig. 6d). Intriguingly, distal branches in the adult tree were devoid of EdU^{+} cells. Thus, embryonic label-retaining cells primarily seed the highly quiescent MaSC fraction in the adult mammary gland, suggesting that this fraction is enriched for the most primitive cells (Fig. 6e).

Proximal-restricted, dormant MaSCs can be activated by steroid hormones

The presence of spatially-restricted quiescent MaSCs in the adult gland prompted the question of whether they could be activated by ovarian hormones. Although the expression of *Tspan8* in both basal and luminal cells negates a lineage tracing strategy, we could track the fate of *Lgr5*-expressing cells in response to physiological stimuli using the Lgr5-GFP-IRES-creER^{T2} model. We first examined the hormonal milieu of pregnancy. Lineage tracing studies

in adult Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mice confirmed that Lgr5-expressing cells contribute to both alveolar luminal and basal cells in the mammary glands of pregnant females[7]. Interestingly, dissection of proximal versus distal regions of the ductal tree revealed substantial expansion of Tomato^{+} clones in the proximal region only (Fig.7a). The sparsely distributed Lgr5-GFP^{+} cells evident along the distal branches of the same glands indicated a lack of activation of Lgr5^{+} cells in this region.

Analysis of the basal subsets in mid-pregnant versus virgin mammary glands revealed a striking but anticipated decrease in the proportion of the highly quiescent subset (Lgr5^{+}Tspan8^{hi}) as well as the other quiescent subpopulations (Lgr5^{-}Tspan8^{hi} and Lgr5^{+}Tspan8^{-})(Fig.7b). Concomitantly, the Lgr5^{-}Tspan8^{-} subset underwent expansion, compatible with previous findings that a MaSC population with reduced self-renewal capacity is amplified in mid-pregnancy[35]. To directly assess whether Lgr5^{+}Tspan8^{hi} cells were responsive to hormonal stimuli, Lgr5-GFP-IRES-creER^{T2} mice were induced with tamoxifen in adulthood, subjected to pregnancy two weeks later, and then labelled with EdU before harvesting in mid-pregnancy. FACS analysis of EdU-labelled cells in the mammary basal population indicated a substantial increase in cycling cells in all basal subsets at day 14.5. The most profound response was observed in the case of Lgr5^{+}Tspan8^{hi} cells, which showed a 20-fold increase in EdU^{+} cells. Thus, the dormant Lgr5^{+}Tspan8^{hi} population is highly responsive to steroid hormones and contributes to alveologenesis in the proximal region.

Similar findings were made for mice treated with a mitogenic hormonal stimulus comprising estrogen (E) plus medroxyprogesterone acetate (MPA). Adult Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mice were induced with tamoxifen, and treated two weeks later

with the MPA+E for 7 days. Remarkable clonal expansion of Lgr5-derived cells occurred in the proximal region of the mammary gland, in contrast to the sparsely distributed tdTomato^{+} cells in the same area of control glands (Supplementary Fig. 7a-f). The expanded tdTomato^{+} domains included myoepithelial-only as well as bi-lineage clonal areas, as demonstrated by the presence of abutting E-cadherin^{+} luminal cells and elongated myoepithelial cells. The alveolar buds that give rise to lateral branches were notably enriched for cells of both lineages. A substantial increase in cycling cells was observed for the Lgr5^{+}Tspan8^{hi} basal subset as well as the luminal population (Supplementary Figure 7g). Upon exposure to MPA+E, very few labelled cells were apparent in the distal branches, recapitulating that observed for pregnancy-associated hormones. Intriguingly, prominent labelling of ducts in the proximal mammary tree was evident during remodelling of the mammary gland during involution in Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mice (Supplementary Fig. 8a-d), suggesting that quiescent MaSCs may be responsive to other stimuli acting in this phase.

DISCUSSION

In this study, Lgr5 together with Tspan8 has enabled the prospective isolation of three distinct, largely quiescent stem cell subsets in the adult mammary gland (Fig. 7d). Quiescence is a common property of diverse adult stem cells and likely evolved to protect the integrity of these long-lived cells, which are required for tissue maintenance throughout life. Highly quiescent stem cells demonstrate potent reconstituting potential *in vivo*, implying that intrinsic mechanisms link the quiescent state to regenerative potential[36,37]. Similarly, the 'dormant' MaSC subset (Lgr5^{+}Tspan8^{hi}) was highly enriched for cells in G0 and had a greater than 20-fold higher repopulating capacity than Lgr5^{-}Tspan8^{-} cells. Notably, the signature of the Lgr5^{+}Tspan8^{hi} population displays the hallmark features of other quiescent tissue-resident stem cells, including downregulation of genes necessary for cell cycle

progression and DNA replication[20].

The different MaSC subsets were found to be distributed in distinct spatial locations within the mammary fat pad. High levels of Tspan8 and Lgr5 were expressed on dormant MaSCs, which predominantly localise to the proximal region. The other quiescent Tspan8^{hi} subset (Lgr5^{-}) also resides within the proximal region, but is distinguished by its lower repopulating ability, lower proportion of cells in G0 and its molecular signature, thus raising the possibility that it represents an intermediate population. Although stem cells in the resting gland may exist in a continuum of states, these states proved to be physically isolatable and suggest an ordered rather than stochastic process. Recent studies in muscle have indicated that quiescent MuSCs exist in two physically distinct states: G0 and an 'alert' phase that enables more rapid cell cycle entry in response to muscle injury[38].

A separate pool of Lgr5^{+}Tspan8^{-} cells was identified in the distal portion of the adult mammary gland, distinguishing them from Tspan8^{+} cells. Curiously, the presence of quiescent Lgr5^{+} stem cells in the mammary gland contrasts with actively dividing Lgr5^{+} cells in several other tissues such as the small intestine, stomach and hair follicles[13,39,40]. Nevertheless, a small fraction of Lgr5^{+} cells has been reported to be slow cycling in the small intestine[41]. The lack of Lgr5^{+} cells within the TEBs of pubertal mammary glands is striking and indicates that *Lgr5* does not govern cell proliferation and ductal morphogenesis during this phase.

EdU label-retention studies provided direct evidence that embryonic cells contribute to the quiescent Lgr5^{+}Tspan8^{hi} population in the adult mammary gland. Both neural stem cells (NSCs)[42,43] and HSCs[44] in the adult have also been shown to originate from fetal stem cells.

Similar to NSCs and HSCs, proliferative fetal MaSCs become quiescent in the post-natal period, but the intrinsic and extrinsic determinants of this switch remain obscure.

Although MaSCs predominantly adopt a quiescent state in the adult, they can be activated by ovarian hormones and the hormonal milieu of pregnancy. Interestingly, activation was only evident in the proximal region of the mammary ductal tree, where Tspan8^{hi} cells reside throughout life. The role of sparsely distributed Lgr5^{+} cells in the distal branches remains to be determined. The recruitment of quiescent stem cells into the cell cycle and the expansion of their progeny is compatible with the observed activation of embryonic label-retaining cells in adult tissue[12]. Moreover, this is reminiscent of HSCs, which normally lie in a dormant state but are highly responsive to bone marrow injury or G-CSF stimulation[37]. Overall, these findings imply that quiescent MaSCs constitute an essential reserve of cells required to maintain a functional compartment of activated MaSC/progenitor cells throughout life. The longevity of these cells positions them as prime candidates for the accumulation of genetic errors and ensuing cellular transformation, either in these cells or their descendants.

METHODS

Mice

R26R-tdTomato and Lgr5-GFP-IRES-creER^{T2} (C57BL/6) mice were obtained from the Jackson Laboratory. Elf5-rtTA-GFP mice were generated as previously described[7]. Wild-type C57BL/6 and FVB/N mice were provided by the animal facility of the Walter and Eliza Hall Institute (WEHI). For timed pregnancies, adult female mice were mated with FVB/N or C57BL/6 males, and scored for the presence of vaginal plugs. Mice were considered as P0.5 on the day of the observed plug. For EdU (5-ethynyl-2-deoxyuridine) label retention studies, pregnant mice were injected intraperitoneally (IP) with 0.2 mg of EdU (200 µl, 1 mg/ml in

PBS) twice daily from E14.5 to E18.5. For cell proliferation analysis, mice were injected IP with 0.15 mg of EdU three times within 24 hours before collection. For acute hormonal treatment, the synthetic progesterone analogue medroxyprogesterone acetate (MPA) pellet, (50 mg, 90 day release; Innovative Research of America) was implanted subcutaneously into 8 week-old FVB/N mice. Mice were also injected subcutaneously with 10 μg estradiol (Sigma) in sunflower oil: ethanol (9:1) daily for 7 days, and injected IP with 0.2 mg of EdU twice daily for the final three days before harvest. No activation of creERT2 was observed by estrogen (plus MPA), consistent with the designed tamoxifen-specificity of creERT2. For pregnancy, dams were IP injected with 0.2 mg of EdU daily from E8.5 to E14.5. For lineage tracing of fetal mammary stem cells, Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mice were induced with 1.5 mg of tamoxifen (50 μl of 30 mg/ml diluted in sunflower seed oil, Sigma) at E17.5, in puberty (28 days) or adulthood (9 weeks). When the dams (about 40% of pregnant females injected with tamoxifen at E17.5) had difficulty in the delivery of the pups, C-section was performed and pups were nursed by a foster mum. For lineage tracing of adult MaSCs, females at the age of 9-10 weeks were injected 1.5 mg of tamoxifen twice with 48 hours interval and set up for plugging two weeks after the last tamoxifen injection. The mums were harvested for analysis at E14.5 or 2 weeks after involution. All mice were bred and maintained in the WEHI animal facility according to institutional guidelines. All experiments were approved by the WEHI Animal Ethics Committee.

Confocal analysis on whole-mounts and histology sections of mammary glands

Tissues were dissected, fixed in 4% paraformaldehyde and incubated overnight at 4 ˚C with primary antibodies. The following day, tissues were incubated with secondary antibodies. For EdU labelling, the tissues were incubated overnight using the Click-it Kit Imaging 647 from Invitrogen, after the secondary antibody step. Tissues were subsequently incubated in 80%

glycerol before dissection for three-dimensional imaging, as previously described[7]. The following primary antibodies were used: Keratin5 (rabbit, Covance; 1:500 dilution), K8/K18 (TROMA-I, rat, DSHB; 1:500 dilution), GFP (chicken, Abcam; 1:500 dilution) and E-cadherin (rat, clone ECCD-2, Invitrogen; 1:200 dilution). Rat monoclonal antibodies against peptides (NGAADWGNNF and NETLYENAKLLS) derived from two non-overlapping regions of the mouse Tspan8 protein were generated in-house (1:400 dilution). Several hybridomas for each peptide were tested and antibodies against the independent peptides yielded identical staining results. One monoclonal antibody against each peptide was selected for further analysis. All secondary antibodies were Alexa Fluor-conjugated: anti-rabbit Alexa Fluor 555 (Invitrogen; 1:500 dilution), anti-chicken Alexa Fluor 488 (Invitrogen; 1:500 dilution), anti-rat Alexa Fluor 555 (Invitrogen; 1:500 dilution).

Imaging analysis for K5/K8 colocalisation in E18.5 embryonic mammary glands

We performed colocalisation measurements using the colocalisation module in Imaris. The threshold was set at 1000 (as recommended) and a colocalisation channel was built and extracted from the colocalisation signal between K5 (basal marker) and K8/K18 (luminal markers). Lgr5-GFP$^+$ cells, K5/K8/K18$^+$, K5$^+$ or K8/K18$^+$ were then counted manually using the orthoslice module in Imaris. Approximately 200 GFP$^+$ cells per developing branch were counted and 100 GFP$^+$ cells per trunk. Three mice were used for image analysis. Statistical analysis was performed using GraphPad Prism software using the Student's t-test.

Mammary cell preparation from adult or embryonic mammary tissue, cell sorting and FACS analysis

Mammary glands from adult female mice or female embryos from Lgr5-GFP-IRES-creER^{T2} dams at E18 were collected. Mammary rudiments (E18) from embryos confirmed to be Lgr5-

GFP$^+$ were dissected under a fluorescence dissection microscope. Single-cell suspensions were prepared essentially as previously described[45]. The following antibodies were used: APC/cy7 anti-mouse/rat CD29 (rat, clone HM1-1, 102226, 1:200 dilution), Pacific Blue anti-mouse CD24 (rat, clone M1/69, Cat#101820, 1:200 dilution), APC anti-mouse CD31 Antibody (Cat#102410, 1:50 dilution), APC anti-mouse CD45 Antibody (Cat#103112, 1:100 dilution), APC anti-mouse TER-119/Erythroid Cell Antibody (Cat#116212, 1:100 dilution) from BioLegend; PE anti-mouse Tspan8 (Rat IgG2b Clone #657909, Cat# FAB6524P, 1: 75 dilution) and APC anti-mouse Tspan8 (Rat IgG2b Clone #657909, Cat# FAB6524A, 1:75 dilution) from R&D Systems. To exclude dead cells, cells were re-suspended in 0.2 μg/ml 7-AAD (Sigma) prior to analysis. For EdU labelling analysis, sorted cells were fixed and stained with the Click-it APC Kit (Invitrogen) according to the manufacturer's instructions. For cell cycle analysis, sorted cells were fixed, then stained with pyronin Y and 7-AAD following standard protocols. FACS analysis and cell sorting were performed on a FACS Aria (Becton Dickinson). The Lin$^-$ population was defined as Ter119$^-$CD31$^-$CD45$^-$ [45]. FACS data were analyzed using FlowJo software (Tree Star).

Transplantation and in vitro colony-forming assays

Freshly sorted basal populations defined by Lgr5 and Tspan8 expression from 9 week-old Lgr5-GFP-IRES-creER^{T2} female mice were implanted (in 25% growth-factor-reduced Matrigel (BD PharMingen)) into the cleared 4th glands of 3-week-old C57BL/6 females, as described previously[45]. Primary outgrowths were collected at 10 weeks after transplantation; basal cells were then sorted and transplanted into 3-week-old C57BL/6 females for secondary outgrowths. Limiting dilution analysis for the primary transplantation was performed as described (Extreme Limiting Dilution Analysis: http://bioinf.wehi.edu.au/software/elda/index.html).

For colony forming assays, freshly sorted cells were embedded in Matrigel (BD PharMingen) and cultured in DMEM/F12 medium supplemented with 1 mM glutamine, 5 μg/ml insulin, 500 ng/ml hydrocortisone, 10 ng/ml epidermal growth factor, 20 ng/ml cholera toxin, and 5% FCS, in a low-oxygen incubator at 37 °C for 7-8 days. Images were captured and colony number and size were analysed by ImageJ. The cutoff for colony size was 2,000 μm^2.

Microarray analysis

Total RNA was purified from sorted cell populations from 9 week-old Lgr5-GFP-IRES-creER^{T2} female mice using the RNeasy Micro kit (Qiagen). The quality of RNA from three biological replicates was assessed with the Agilent Bioanalyzer 2100 (Agilent Technologies) by using the Agilent RNA 6000 Nanokit (Agilent Technologies) according to the manufacturer's protocol. Up to 120 ng of RNA was amplified with the standard Total Prep RNA amplification kit (Ambion), and complementary RNA (1.5 μg) was labelled and hybridized to Illumina MouseWG-6 v2.0 BeadChips at the Australian Genome Research Facility (AGRF), Melbourne. After washing, the chips were scanned using an Illumina BeadArray Reader and summary probe profiles were output by GenomeStudio for each experiment separately. Subsequent analysis was carried out in R[46] using the *limma* package[47]. Intensities were background corrected and quantile normalized using the neqc method[48]. These data have been deposited in the Gene Expression Omnibus (Accession number GSE73045). Probes with consistently low expression (Detection p-value<0.95 on fewer than 3 arrays) or poor annotation (according to[49] were removed from further analysis. Expression levels for the different cell populations were estimated using linear models[50] with array quality weights[51] and correlations between samples collected at different times[52]. Pair-wise contrasts between the various populations were calculated and differential expression was

assessed using moderated *t*-statistics. Probes with false discovery rate (FDR) <0.05 and fold-change ≥2 were considered DE. An over-representation analysis of Gene Ontology terms was performed using the *goana* function in *limma*. In all heat maps shown, expression values are on a $\log_2$ scale and have been row-scaled.

RNA-seq analysis

Total RNA was extracted from sorted luminal or basal populations defined by Lgr5 and Tspan8 expression from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} female mice Total RNA (50 ng) for each of the two biological replicates was used to generate libraries for whole-transcriptome analysis following Illumina's TruSeq RNA v2 sample preparation protocol. Libraries were sequenced on an Illumina HiSeq 2000 at the Australian Genome Research Facility (AGRF), Melbourne. At least 30 million 100 bp single-end reads were obtained for each sample. Reads were aligned to the mouse genome (mm10) using the subread algorithm available from the *Rsubread* package[53]. The number of reads overlapping each Entrez gene was counted using the RefSeq gene annotation by the featureCounts function[54]. Filtering and normalization used the *edgeR* package[55]. Genes with low expression (defined as having a count per million (CPM) of less than 0.5 in fewer than 2 samples) were removed from further analysis. Compositional differences between libraries were normalized using the trimmed mean of *M*-values (TMM) method[56]. Subsequent differential expression analysis was performed using the *limma* package[47]. Counts were transformed to $\log_2$-CPM values (with an offset of 0.5) with associated precision weights using voom[57]. A linear model with effects for cell population and a blocking effect for experimental batch were fitted. Various contrasts between the 4 populations were estimated and differential expression was assessed using moderated *t*-statistics. Genes with FDR<0.05 and fold-change≥2 were considered DE. Gene ontology analysis used the *goana* function, which includes a correction

for gene length bias as per *goseq*[58]. Gene set testing using the *roast* method[31] for various published expression signatures were carried out. Genes were matched between experiments using Gene symbols. These data have been deposited in the Gene Expression Omnibus (Accession number GSE73111). In all heat maps shown, expression values are $\log_2$-CPM scale and have been row-scaled.

ChIP Sample Preparation and Sequencing

Freshly sorted cells (100,000 primary cells) of four basal sub-populations defined by the Lgr5 and Tspan8 expression were crosslinked with 1% paraformaldehyde. The ChIP assay was carried out as per the manufacturer's protocol (Millipore #17-371). Briefly, cells were lysed and the chromatin was sheared to a size range of 200 to 400 bp using Covaris M220 sonicator. Sheared chromatin was diluted and incubated at 4°C overnight with antibodies against Histone H3 trimethyl Lys4 (K4, Millipore #07-473, 1:100 dilution), Histone H3 trimethyl Lys27 (K27, Millipore #07-449, 1:100 dilution) or mouse isotype control (Millipore #12-371, 1:100 dilution). Immune complexes were handled according to the manufacturer's protocol.

The ChIP DNA samples were prepared and indexed for Ilumina sequencing using the TruSeq DNA sample Prep Kit (Ilumina) as per manufacturer's instruction. The library was quantified using the Agilent Tapestation and the Qubit™ RNA assay kit for Qubit 2.0® Fluorometer (Life Technologies). The indexed libraries were then prepared for paired-end 75 bp sequencing on a NextSeq500 instrument using the 150 cycle kit v2 chemistry (Ilumina) as per manufacturer's instructions. Reads were aligned to the mouse genome (mm10) using the subread algorithm[53] and counted into bins associated with Entrez gene identifiers (27080 genes in total) using featureCounts[54] in two different ways. Gene-body counts summarized the number of reads overlapping anywhere between the first and last base of a given gene

while promoter counts summarized the number of reads overlapping a region 3000 bases upstream to 2000 bases downstream of the TSS of each gene. For downstream analysis, gene-body counts were used for the H3K27me3 marks and promoter counts were used for the H3K4me3 marks. For each mark, $\log_2$-fold changes (logFC) and average $\log_2$-CPM values for each population were calculated for the 831 DE genes between Lgr5^{+}Tspan8^{hi} and the average of the remaining populations in the RNA-seq analysis. Boxplots were used to display logFC values of ChIP-seq versus RNA-seq for those DE genes, separating upregulated genes from those that were downregulated in the RNA-seq comparison. Average $\log_2$-CPM values were displayed in heatmaps for DE genes that were enriched for either the H3K4me3 or H3K27me3 marks. For H3K4me3, genes were classified as enriched if at least 2 samples have CPM values greater than the median CPM value. For H3K27me3 the 80th percentile of the CPM values was used in place of the median value. A total of 325 genes were plotted in the heatmap of the H3K4me3 data, and 469 genes for the heatmap of the H3K27me3 data. The average $\log_2$-CPM values were mean-adjusted by subtracting the gene-wise average $\log_2$-CPM values. Heatmaps show relative rather than absolute enrichment of each group and are ordered from high to low enrichment in the Lgr5^{+}Tspan8^{hi} subset.

In a separate analysis, peaks were called on aligned reads using MACS2[59] using the relevant input control with a *q*-value 0.01 cut-off. Results from replicate samples were combined and peak profiles were plotted using the ChIPseeker package[60] for the region 10,000 bases upstream and downstream of the TSS of genes defined using annotation from the TxDb.Mmusculus.UCSC.mm10.knownGene package (M Carlson, 2016, version 3.2.2.). These data have been deposited in the Gene Expression Omnibus (Accession number GSE89450).

Comparison of MaSC signatures with human breast cancer subtypes

Gene expression profiles from human breast cancers were obtained from[28] under Gene Expression Omnibus accession number GSE18229. Expression values measured on the Agilent Human 1A Oligo UNC custom microarray platform (GPL1390) from 5 sub-types (Basal, Luminal A, Luminal B, Her2 and Claudin-low) were used in the quiescent MaSCs signature score calculations. Mouse gene symbols were converted to human an ortholog table downloaded from the Mouse Genome Informatics resource (http://www.informatics.jax.org).

For a given mouse signature obtained from the RNA-seq experiment (genes with the FDR<0.05 and absolute fold-change>2), a per patient signature score was calculated as the average expression value for genes that could be matched on the Agilent platform, weighted by the respective RNA-seq $\log_2$ fold-change. Boxplots of the various signature scores per subtype were then made (Supp. Fig. 3d). These results were summarized using the barcode plot function from the *limma* package (Supp. Fig. 3e). First, linear models were fitted to the human data to compare gene expression levels between different sub-types. Genes were ranked by moderated *t*-statistics in the Claudin-low versus Basal-like subtype comparison and plotted as a rectangular block in the center of the figure. Next genes either up- (red bars) or down- (blue bars) regulated in the signature comparing $Lgr5^{+}Tspan8^{hi}$ expression versus the average of all other populations were overlayed. Finally, enrichment of these up and down genes was summarized by plotting a moving average calculated using a tri-cube weight function. Gene set testing of the mouse signatures in the human data was performed using the *roast* method[31].

Statistics and reproducibility

No statistical method was applied to predetermine sample size. The experiments were not randomized. No samples or animals were excluded from the analyses. The investigators were

not blinded to allocation during experiments and outcome assessment. Most of the experiments were repeated at least three times and the exact *n* is stated in the corresponding figure legend. Error bars are only shown when the data are derived from more than 3 independent experiments. Data are shown as mean ± standard error of the mean (SEM). The Student's t-test was used where applicable; a $p<0.05$ was considered significant, the exact p value is indicated in the figure legends.

Data availability

Microarray and RNA–seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under accession codes GSE73045 and GSE73111. ChIP-seq data have been deposited under accession number GSE89450. Previously published microarray data that were re-analysed here are available under accession codes GSE18229 and GPL1390. All other data supporting the findings of this study are available from the corresponding author upon reasonable request.

45. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84-88 (2006).
46. R Development Core Team *R: A language and environment for statistical computing.* . (R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. 2012).
47. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
48. Shi, W., Oshlack, A. & Smyth, G.K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res* **38**, e204 (2010).

49. Barbosa-Morais, N.L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* **38**, e17 (2010).
50. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).
51. Ritchie, M.E. *et al.* Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics* **7**, 261 (2006).
52. Smyth, G.K., Michaud, J. & Scott, H.S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067-2075 (2005).
53. Liao, Y., Smyth, G.K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**, e108 (2013).
54. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
55. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
56. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).
57. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
58. Young, M.D., Wakefield, M.J., Smyth, G.K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14 (2010).
59. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
60. Yu, G., Wang, L.G. & He, Q.Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383 (2015).

AUTHOR CONTRIBUTIONS

N.Y.F, A.C.R. designed and performed experiments and contributed to manuscript writing; F.V., B.P., P.J., F.J. and K.L. performed experiments; C.L., R.L., G.K.S. and M.E.R. performed bioinformatics analysis; G.J.L. contributed to interpretation of data. J.E.V. conceived the study and carried out manuscript writing.

ACKNOWLEDGMENTS

We are grateful to the Animal, FACS, Monoclonal Antibody (Bundoora), Imaging and Histology facilities at WEHI and to M. Milevskiy for discussions. This work was supported by the Australian National Health and Medical Research Council (NHMRC) grants #1016701, #1024852, #1054618, #1059622, #1085191, #1086727, #1100807; NHMRC IRIISS; the Victorian State Government through Victorian Cancer Agency (VCA) funding and Operational Infrastructure Support; and the Australian Cancer Research Foundation. N.Y.F. and A.C.R. were supported by a National Breast Cancer Foundation (NBCF)/Cure Cancer Australia Fellowship; B.P. by a NHMRC Fellowship #1016571 and a VCA Fellowship; G.K.S., G.J.L., M.E.R., and J.E.V. by NHMRC Fellowships #1058892, #1078730, #1104924; #1102742.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Visvader, J.E. & Clevers, H. Tissue-specific designs of stem cell hierarchies. *Nat Cell Biol* **18**, 349-355 (2016).
2. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993-997 (2006).
3. Barker, N., Tan, S. & Clevers, H. Lgr proteins in epithelial stem cell biology. *Development* **140**, 2484-2494 (2013).
4. Plaks, V. *et al.* Lgr5-expressing cells are sufficient and necessary for postnatal mammary gland organogenesis. *Cell Rep* **3**, 70-78 (2013).
5. Wang, D. *et al.* Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* **517**, 81-84 (2015).
6. de Visser, K.E. *et al.* Developmental stage-specific contribution of LGR5(+) cells to basal and luminal epithelial lineages in the postnatal mammary gland. *J Pathol* **228**, 300-309 (2012).
7. Rios, A.C., Fu, N.Y., Lindeman, G.J. & Visvader, J.E. In situ identification of bipotent stem cells in the mammary gland. *Nature* **506**, 322-327 (2014).
8. Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189-193 (2011).
9. Cicalese, A. *et al.* The tumor suppressor p53 regulates polarity of self-renewing divisions in mammary stem cells. *Cell* **138**, 1083-1095 (2009).
10. Dos Santos, C.O. *et al.* Molecular hierarchy of mammary differentiation yields refined markers of mammary stem cells. *Proc Natl Acad Sci U S A* **110**, 7123-7130 (2013).
11. Smith, G.H. Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* **132**, 681-687 (2005).

12. Boras-Granic, K., Dann, P. & Wysolmerski, J.J. Embryonic cells contribute directly to the quiescent stem cell population in the adult mouse mammary gland. *Breast Cancer Res* **16**, 487 (2014).
13. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003-1007 (2007).
14. Zhang, L. *et al.* Establishing estrogen-responsive mouse mammary organoids from single Lgr5+ cells. *Cell Signal* (2016).
15. Greco, C. *et al.* E-cadherin/p120-catenin and tetraspanin Co-029 cooperate for cell motility control in human colon carcinoma. *Cancer Res* **70**, 7674-7683 (2010).
16. Hemler, M.E. Tetraspanin functions and associated microdomains. *Nat Rev Mol Cell Biol* **6**, 801-811 (2005).
17. Zoller, M. Tetraspanins: push and pull in suppressing and promoting metastasis. *Nat Rev Cancer* **9**, 40-55 (2009).
18. Giraddi, R.R. *et al.* Stem and progenitor cell division kinetics during postnatal mouse mammary gland development. *Nat Commun* **6**, 8487 (2015).
19. Arai, F. & Suda, T. Quiescent stem cells in the niche, in *StemBook* (Cambridge, MA; 2008).
20. Cheung, T.H. & Rando, T.A. Molecular regulation of stem cell quiescence. *Nat Rev Mol Cell Biol* **14**, 329-340 (2013).
21. Zeng, Y.A. & Nusse, R. Wnt proteins are self-renewal factors for mammary stem cells and promote their long-term expansion in culture. *Cell Stem Cell* **6**, 568-577 (2010).
22. Fukada, S. *et al.* Molecular signature of quiescent satellite cells in adult skeletal muscle. *Stem Cells* **25**, 2448-2459 (2007).
23. Liu, L. *et al.* Chromatin modifications as determinants of muscle stem cell quiescence and chronological aging. *Cell Rep* **4**, 189-204 (2013).

24. Chambers, S.M. *et al.* Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* **1**, 578-591 (2007).
25. Lien, W.H. *et al.* Genome-wide maps of histone modifications unwind in vivo chromatin states of the hair follicle lineage. *Cell Stem Cell* **9**, 219-232 (2011).
26. Genander, M. *et al.* BMP signaling and its pSMAD1/5 target genes differentially regulate hair follicle stem cell lineages. *Cell Stem Cell* **15**, 619-633 (2014).
27. He, X.C. *et al.* BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt-beta-catenin signaling. *Nat Genet* **36**, 1117-1121 (2004).
28. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).
29. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* **15**, 907-913 (2009).
30. Pal, B. *et al.* Global changes in the mammary epigenome are induced by hormonal cues and coordinated by Ezh2. *Cell Rep* **3**, 411-426 (2013).
31. Wu, D. *et al.* ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176-2182 (2010).
32. Westphalen, C.B. *et al.* Dclk1 Defines Quiescent Pancreatic Progenitors that Promote Injury-Induced Regeneration and Tumorigenesis. *Cell Stem Cell* **18**, 441-455 (2016).
33. Makarem, M. *et al.* Developmental changes in the in vitro activated regenerative activity of primitive mammary epithelial cells. *PLoS Biol* **11**, e1001630 (2013).
34. Spike, B.T. *et al.* A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. *Cell Stem Cell* **10**, 183-197 (2012).
35. Asselin-Labat, M.L. *et al.* Control of mammary stem cell function by steroid hormone signalling. *Nature* **465**, 798-802 (2010).

36. Orford, K.W. & Scadden, D.T. Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation. *Nat Rev Genet* **9**, 115-128 (2008).
37. Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118-1129 (2008).
38. Rodgers, J.T. *et al.* mTORC1 controls the adaptive transition of quiescent stem cells from G0 to G(Alert). *Nature* **510**, 393-396 (2014).
39. Barker, N. *et al.* Lgr5(+ve) stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. *Cell Stem Cell* **6**, 25-36 (2010).
40. Jaks, V. *et al.* Lgr5 marks cycling, yet long-lived, hair follicle stem cells. *Nat Genet* **40**, 1291-1299 (2008).
41. Buczacki, S.J. *et al.* Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature* **495**, 65-69 (2013).
42. Fuentealba, L.C. *et al.* Embryonic Origin of Postnatal Neural Stem Cells. *Cell* **161**, 1644-1655 (2015).
43. Furutachi, S. *et al.* Slowly dividing neural progenitors are an embryonic origin of adult neural stem cells. *Nat Neurosci* **18**, 657-665 (2015).
44. Bowie, M.B. *et al.* Identification of a new intrinsically timed developmental checkpoint that reprograms key hematopoietic stem cell properties. *Proc Natl Acad Sci U S A* **104**, 5878-5882 (2007).

FIGURE LEGENDS

Table 1. Repopulating frequency of basal subsets defined by Lgr5 and Tspan8 expression isolated from the mammary glands of adult mice.

Limiting dilution analysis of the repopulating frequency of fractionated subsets of Lin^- $CD29^{hi}CD24^+$ cells from the mammary glands of virgin 9 week-old Lgr5-GFP-IRES-creER^{T2} females. Cells were injected into the cleared mammary fat pads of 3 week-old syngeneic recipients. Data are pooled from three independent experiments. *Shown as the number of outgrowths per number of injected fat pads. p-values for pairwise tests of differences in repopulating frequencies for the $Lgr5^+Tspan8^{hi}$ subset versus the $Lgr5Tspan8^+$, $Lgr5^+Tspan8$ and Lgr5Tspan8 subsets are 6.08e-07, 1.15e-05 and 1.8e-19, respectively.

Figure 1. Basal-restricted localisation and gene profiling of $Lgr5^+$ cells

(a) Whole-mount 3D confocal image and optical sections of a duct (I) and terminal endbud (TEB, II) enlarged from a whole-mounted mammary ductal portion isolated from a Lgr5-GFP-IRES-creER^{T2} mouse at 5 weeks of age (puberty) after 3 consecutive days of EdU administration. The glands were immunolabelled for GFP (green), EdU (red), and Keratin-5 (K5) (blue). The red arrow depicts the leading edge of EdU^+ cells and the white arrow indicates where Lgr5-GFP^+ cells begin to arise. No $Lgr5^+$ were observed in the TEB structures, shown in the optical section of a TEB, and no $Lgr5^+$ cells were EdU^+, as evident in the optical section of the duct (representative of 3 mice, 3 independent experiments). Scale bars, 300 μm (whole-mount); 50 μm (optical sections). (**b**) Whole-mount 3D confocal image of a ductal portion located in the distal part of the mammary gland from an adult (9 week-old) Lgr5-GFP-IRES-creER^{T2} female stained for E-cadherin (blue) (representative of 4 mice). $Lgr5^+$ cells were often apparent as small clusters and only appeared in the basal layer of the

adult mammary gland (representative of 3 mice, 3 independent experiments). Scale bars, 300 μm (whole-mount); 50 μm (optical section). (**c**) Representative FACS plots (180 mice were analysed in a total of 15 independent experiments) for detection of Lgr5-GFP^{+} cells in the indicated sub-populations from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females. (**d**) Bar graph showing the percentage of the Lgr5-GFP^{+} cells in the luminal (Lin^{-} CD29^{lo}CD24^{+}) and basal (Lin^{-}CD29^{hi}CD24^{+}) populations from the mammary glands of 9 week-old Lgr5CreER^{T2} females. Error bars represent mean ± SEM (120 mice were analysed in a total of n=15 independent experiments). (**e**) Heat map showing the top 100 upregulated genes in Lgr5^{+} versus Lgr5^{-} basal cells. Three biological replicates of luminal, Lgr5^{+} and Lgr5^{-} basal cells from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females were sorted and their expression profiles determined by microarray analysis. The cut-off for DE (Differential Expression) genes is FDR<0.05 and absolute fold-change ≥ 2 (70 mice were analysed in a total of n=3 independent experiments).

Figure 2. Prospective isolation of different subsets of mammary stem cells based on expression of Lgr5 and Tspan8

(**a**) The mammary basal compartment (Lin^{-}CD29^{hi}CD24^{+}) can be further separated into four distinct subpopulations based on Lgr5 and Tspan8 expression. Representative FACS plots showing the distribution of Lgr5^{+} and Tspan8^{+} cells in the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females (160 mice were analysed in a total of 15 independent experiments). (**b**) Bar graph showing the percentage of Lgr5-GFP^{+} and Tspan8^{hi} cells in the basal population isolated from adult Lgr5-GFP-IRES-creER^{T2} female mice (9 weeks). Error bars represent mean ± SEM (160 mice were analysed in a total of n=15 independent experiments). (**c**) Representative images showing the colony forming capacity of the different basal subpopulations defined by Lgr5 and Tspan8 expression. Cells were freshly isolated

from the mammary glands of 10-week-old Lgr5-GFP-IRES-creER^{T2} female and cultured in Matrigel for 7 days. Scale bar, 100 μm (3 independent experiments). (**d**) Bar graph showing colony size or percentage of branched colonies derived from the indicated basal subpopulations. Error bars represent mean ± SEM (n=3 independent experiments). **p<0.01, Student's t-test. (**e**) Representative FACS analysis of primary outgrowths for Lgr5-GFP and Tspan8 expression. Outgrowths from virgin recipient mice, transplanted with 400 cells, were pooled to procure sufficient cells for flow cytometry. The outgrowths were analysed by FACS at 10 weeks post-transplantation (36 recipient mice for each population, 3 independent experiments).

Figure 3. Cells marked by Lgr5 and Tspan8 are largely quiescent in the steady-state gland and share homology with other quiescent tissue-specific cells

(**a**) Representative FACS plots showing the DNA and RNA content of different basal sub-populations defined by Lgr5 and Tspan8 expression (Lin$^-$CD29^{hi}CD24$^+$). Sorted cells from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females were fixed and stained with 7-AAD and pyroninY (pY). G0 was defined as G1 cells with low RNA content (30 mice, 3 independent experiments). (**b**) Bar graph showing the percentage of cells at different cell cycle stages for the indicated basal subpopulations. * p < 0.05; **p < 0.01, Student's t-test (30 mice, n=3 independent experiments). (**c**) Heat map showing DE genes for Lgr5-GFP$^+$Tspan8^{hi} cells versus ALL other subpopulations (based on cutoffs of FDR<0.05 and absolute fold-change ≥ 2). RNA-seq analysis was performed on two biological replicates of the indicated subpopulations from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females (90 mice were analysed in a total of n=2 independent experiments). (**d**) Heat map of Wnt pathway genes (GO: 0030509) DE in (c). (**e**) Bar graph depicting *p*-values for roast analysis performed on comparison of the gene expression signatures of quiescent

MaSCs with those of quiescent muscle stem cells (MuSCs), quiescent hair follicle stem cells (HFSCs) and hematopoietic stem cells (HSCs). (**f**) Venn diagram showing common genes expressed between quiescent MaSCs and MuSCs from two different datasets [22,23]. A heat map of the 26 genes shared between all three signatures is shown. * $p < 0.05$; **$p < 0.005$; ***$p < 0.0005$; ****$p < 0.00005$, Roast analysis.

Figure 4. The four basal subpopulations defined by Lgr5 and Tspan8 exhibit distinct epigenetic profiles for H3K4me3 and H3K27me3.

(**a**) Global distribution of H3K4me3 and H3K27me3 modifications across genes in the different subpopulations: ChIP-seq analysis was performed on approx. 100,000 cells per subset. Average peak intensity 10 kb upstream and downstream of the TSS (Transcription Start Site sequence) across the genome is shown. **(b)** Boxplots display logFC for ChIP-seq and RNA-seq analyses of DE genes (by RNA-seq) for $Lgr5^{+}Tspan8^{hi}$ cells versus the other subpopulations. $p = 1.1x10^{-19}$ for H3K4me3 and $p = 6.3x10^{-26}$ for H3K27me3 (Wilcoxon rank sum test). Boxes show 25%, 50% and 75% percentiles. Whiskers extend to minimum and maximum values. **(c)** Mean-adjusted average log_2-CPM values are displayed in heatmaps for the DE genes that were bound by H3K4me3 or H3K27me3 in any of the four populations. **(d)** Read coverage graphs for H3K4me3 and H3K27me3 in $Lgr5^{+}Tspan8^{hi}$ (red) and $Lgr5^{-}Tspan8^{-}$ (green) subpopulations across the gene-body of four representative genes enriched in $Lgr5^{+}Tspan8^{hi}$ cells (80 mice were analysed in a total of 2 independent experiments).

Figure 5. Quiescent mammary stem cells in the adult are restricted to the proximal area and may derive from the embryo

(**a**) Schematic diagram demarcating the proximal and distal areas of mammary glands used for FACS analysis in (b). (**b**) Representative FACS plots showing $Lgr5^{+}$ and $Tspan8^{hi}$ cells in the proximal area of mammary glands from Lgr5-GFP-IRES-creER^{T2} females during puberty (5 weeks) or adulthood (9 weeks or 6 months). It is noteworthy that the proximal region contains the highest number of epithelial cells owing to the density of primary ducts in this area (15 mice were analysed for each age in a total of 3 independent experiments). (**c**) Bar graph showing the percentage of cells defined by Lgr5 and Tspan8 expression in the basal compartment ($Lin^{-}CD29^{hi}CD24^{+}$) isolated from either the proximal (prox) or distal (dist) area of Lgr5-GFP-IRES-creER^{T2} mammary glands at the indicated age (15 mice were analysed in a total of n=3 independent experiments). (**d, e**) Representative confocal images of Tspan8 expression in a proximal duct (d) or TEB (e). Proximal or distal areas of 5 week-old glands were stained with Tspan8 (red), p63 (green) and DAPI (blue). White and yellow arrows in (d) depict examples of cells that co-express Tspan8 and p63 or cells that only express p63, respectively (5 mice were analysed in a total of 3 independent experiments. Scale bars, 20 μm.

Figure 6. Characterisation of Lgr5 and Tspan8 expression in fetal mammary gland

(**a**) Whole-mount 3D confocal images (a1, b1, c1) and optical sections (a2, b2, c2, c3) of entire mammary ductal trees from Lgr5-GFP-IRES-creER^{T2} embryos. Lgr5-GFP-IRES-creER^{T2} mothers were injected with EdU (three times in 24 hours) prior to collection of the mammary primordia at E14.5 (a1, a2), E16.5 (b1, b2), or E18.5 (c1, c2, c3), stained for GFP (green), EdU (red) and K5 (blue). Three embryos for E14.5, 3 for E16.5 and 5 for E18.5. Although K5 is expressed in virtually all embryonic mammary epithelial cells, higher levels are restricted to the outer layer at E18.5. Scale bars, 100 μm (whole-mounts); 40 μm (optical sections) (10 embryos were analysed in a total of 3 independent experiments). (**b**)

Representative FACS plot showing Lgr5 and Tspan8 expression in fetal mammary glands. Embryos were harvested at E18.5 from pregnant Lgr5-GFP-IRES-creER^{T2} females. Mammary glands from female embryos were dissected under a fluorescence microscope and GFP^{+} mammary rudiments from 8-10 embryos were pooled for the preparation of single cell suspensions for FACS analysis (2 experiments). **(c)** Representative FACS plots showing EdU^{+} cells in the indicated populations of fetal mammary primordia defined by Lgr5 and Tspan8 expression (8 embryos, 2 independent experiments). (**d**) Embryonic origin of Lgr5^{+}Tspan8^{hi} cells in the adult mammary gland. Pregnant Lgr5CreER^{T2} mothers were injected with EdU from 14.5 to 18.5 days of pregnancy. Mammary glands were harvested at 6 weeks of age, sorted, fixed and analysed for EdU retention. Bar graph shows the percentage of EdU^{+} cells in each of the four basal subsets (8 embryos, n=2 independent experiments). (**e**) Schematic diagram summarising the distribution of the different basal subpopulations defined by Lgr5 and Tspan8 expression at different developmental stages.

Figure 7. Quiescent mammary stem cells can be activated by hormonal cues (**a**) Whole-mount 3D confocal image of the proximal and distal portions of a Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mammary gland at 14.5 days of pregnancy after tamoxifen injection at 9 weeks and immunostained for E-cadherin (blue). The enlargement in the proximal region shows an alveolus containing labelled luminal and myoepithelial cells. The arrowhead indicates an E-cadherin^{+} luminal cell and the arrow depicts an E-cadherin^{-} myoepithelial cell. Scale bars: 200 μm (whole-mounts), 15 μm (optical sections). Representative of 3 mice. (**b**) Representative FACS plots showing the distribution of Lgr5^{+} and Tspan8^{+} cells in the mammary glands from virgin and pregnant (14.5 days) Lgr5-GFP-IRES-creER^{T2} females. Bar chart depicting the fold-decrease in Lgr5^{+}Tspan8^{hi}, Lgr5^{-}Tspan8^{+} and Lgr5^{+}Tspan8^{-} subpopulations during pregnancy compared to virgin mice. Error bars represent mean ± SEM

(4 mice, n=4 independent experiments). (**c**) FACS plots showing EdU^+ cells in the four basal subpopulations isolated from Lgr5-GFP-IRES-creER^{T2} virgin and mid-pregnant mice. Bar chart showing the percentage of EdU^+ cells in the four basal subpopulations defined by Lgr5 and Tspan8 (25 mice were analysed in a total of n=3 independent experiments). Error bars represent mean ± SEM; * $p < 0.05$; **$p < 0.01$; ***$p < 0.001$, Student's t-test. (**d**) Proposed model of the MaSC differentiation hierarchy. The stem cell compartment comprises distinct subsets that exist in a largely quiescent state. Fetal MaSCs seed the post-natal MaSC pool, which then adopt a quiescent state, reminiscent of HSCs and NSCs. $Lgr5^+Tspan8^{hi}$ cells represent a dormant pool of MaSCs that are more quiescent than the $Lgr5^-Tspan8^+$ and $Lgr5^+Tspan8^-$ subsets and may lie upstream based on transplantation studies. $Lgr5^-Tspan8^-$ cells have not been included in the model but are likely to encompass myoepithelial progenitor and mature cells.

Table 1. Repopulating frequency of basal subsets defined by Lgr5 and Tspan8 expression isolated from the mammary glands of adult mice.

Sub-population	Number of cells transplanted	Number of outgrowths	Repopulating frequency (CI 95%)
$Lgr5^{+}Tspan8^{hi}$	10	10/12	16.2 (10.1-26.2)
	25	5/6	
	50	10/11	
	75	5/6	
	100	6/6	
	400	6/6	
$Lgr5^{-}Tspan8^{hi}$	10	3/12	73.1 (45.4-117.9)
	25	3/6	
	50	9/11	
	75	4/6	
	100	3/6	
	400	5/6	
$Lgr5^{+}Tspan8^{-}$	10	4/12	63.0 (39.6-100.4)
	25	3/6	
	50	6/11	
	75	3/6	
	100	4/6	
	400	6/6	
$Lgr5^{-}Tspan8^{-}$	10	2/12	361.2 (183.6-711.3)
	25	1/6	
	50	3/11	
	75	1/6	
	100	1/6	
	400	2/6	

Limiting dilution analysis of the repopulating frequency of fractionated subsets of Lin^{-} $CD29^{hi}CD24^{+}$ cells from the mammary glands of virgin 9 week-old Lgr5-GFP-IRES-creER^{T2} females. Cells were injected into the cleared mammary fat pads of 3 week-old syngeneic recipients. Data are pooled from three independent experiments. *Shown as the number of outgrowths per number of injected fat pads. p-values for pairwise tests of differences in repopulating frequencies for the $Lgr5^{+}Tspan8^{hi}$ subset versus the $Lgr5Tspan8^{+}$, $Lgr5^{+}Tspan8$ and Lgr5Tspan8 subsets are 6.08e-07, 1.15e-05 and 1.8e-19, respectively.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Gene expression profiling of Lgr5^{+} cells in the adult mammary gland

(**a**) Whole-mount 3D confocal image and optical section of a ductal portion located in the distal part of the mammary gland from an adult (9 week-old) Lgr5-GFP-IRES-creER^{T2} female (representative of n = 4 mice, three independent experiments) immunostained for K5 (blue). Lgr5^{+} cells only appear in the basal population of the adult mammary glands. Scale bars, 50 μm (whole-mount); 10 μm (optical sections). (**b**) Heat map showing all the DE genes in any pairwise comparison amongst the luminal, Lgr5-GFP^{+} (Lgr5^{+}) and Lgr5-GFP^{-} (Lgr5^{-}) basal populations. Expression values are on a log2 scale and are mean-corrected for each gene. Three biological replicates were sorted from 9 week-old Lgr5-GFP-IRES-CreER^{T2} females and their transcriptomes analysed by microarray. The cutoff for DE was FDR<0.05 and absolute fold-change ≥ 2 (n>50 mice, five independent experiments). (**c**) Heat map for the top 100 upregulated genes for Lgr5^{+} vs Lgr5^{-} basal cells in (a). (**d**) GO enrichment analysis of DE genes between Lgr5^{+} versus Lgr5^{-} basal cells in (b).

Supplementary Figure 2. FACS analysis of Tspan8-expressing cells in the adult mammary gland and representative outgrowth generated by different basal subpopulations

(**a**) Representative FACS plots showing Tspan8-expressing cells in the mammary glands of 9 week-old C57BL/6 females (n=15 mice, three independent experiments). (**b**) Representative FACS plots showing Tspan8-expressing cells in the mammary glands of 9 week-old FVB/N females (n=10 mice, three independent experiments). (**c**) Representative FACS plots showing overlap of Elf-GFP^{+} and Tspan8^{+} cells in the luminal population from the mammary glands of 9 week-old Elf5-rtTA-GFP females (n = 2 mice, two independent experiments). (**d**)

Transplantation of the different subpopulations (50 cells) into the cleared fat pads of 3-week old recipient female mice. Glands were collected 10 weeks post-transplantation. Representative whole-mount images are shown (n=6 mice for each population). Scale bar, 2 mm. **(e)** Transplantation of the different subpopulations (400 cells) into the cleared fat pads of recipient females. Females were mated 9-10 weeks after transplantation and mammary glands were collected at 18.5 days of pregnancy. Representative whole-mounts of outgrowths are shown (n=6 mice per population). Scale bar, 2 mm. **(f)** Representative whole-mount images of secondary outgrowths. Basal cells were sorted from primary outgrowths generated from the indicated subsets defined by Lgr5 and Tspan8 expression, and transplanted into the cleared fat pads of secondary recipient females. The secondary outgrowths were collected 10 weeks post-transplantation. Scale bar, 2 mm (n>5 mice for each population for each experiment, three independent experiments). (**g**) Representative FACS plots showing the DNA and RNA content of the luminal compartment isolated on the basis of Tspan8 expression (all $Lin^{-}CD29^{lo}CD24^{+}$). Sorted cells from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females were fixed and stained with 7-AAD and pyroninY (pY). G0 is defined as G1 cells with low RNA content (n=30 mice, three independent experiments).

Supplementary Figure 3. Gene expression signatures of quiescent mammary stem cells and comparison with the major subtypes of human breast cancer

(**a**) Relative expression (log2 cpm values) of known basal markers across the four different myoepithelial/basal subpopulations as determined by RNA-seq analysis. (**b, c**) Heat maps showing expression of the top 100 DE genes on comparison of $Lgr5^{+}Tspan8^{hi}$ cells versus the average of the other three subsets, either upregulated (b) or downregulated (c). Two biological replicates were sorted by flow cytometry from the mammary glands of 9 week-old Lgr5-GFP-IRES-creER^{T2} females and their transcriptomes were determined by RNA-seq

(n>50 mice, five independent experiments) (**d**) GO enrichment analysis of DE genes for $Lgr5^{+}Tspan8^{hi}$ versus $Lgr5^{+}Tspan8^{-}$ cells (n>50 mice, five independent experiments). (**e**) Box plots of quiescent MaSC signature scores by tumour subtype. The signature scores for the $Lgr5^{+}Tspan8^{hi}$ sub-population compared to any other population are strongly correlated with the claudin-low subtype of breast cancer (n>50 mice, five independent experiments). (**f**) Barcode plot depicting the strongly associated gene expression signatures of quiescent MaSCs and claudin-low tumours compared to the basal-like subtype. Genes are ordered from right to left as most upregulated to most downregulated in claudin-low cancer. The red lines designate upregulated genes in quiescent MaSCs ($Lgr5^{+}Tspan8^{hi}$ vs All Others), whereas blue lines designate downregulated genes. The cutoff for DE is FDR<0.05 and absolute fold-change ≥ 2 (n>50 mice, five independent experiments). *P* values measuring the overall correlation were derived from the 'roast' function of the limma software package.

Supplementary Figure 4. Localisation of $Tspan8^{hi}$ basal cells in the adult mammary gland

(**a**) Representative FACS plots showing $Tspan8^{+}$ luminal cells in the proximal and distal areas of mammary glands isolated from Lgr5-GFP-IRES-$creER^{T2}$ females at 5 weeks, 9 weeks or 6 months of age (n>50 mice, n>5 independent experiments). (**b**) Representative FACS plots showing $Tspan8^{hi}$ basal cells in the proximal and distal areas of mammary glands from C57BL/6 females after completion of three pregnancy cycles (n=3 mice, three independent experiments). (**c**) Representative FACS plots showing the basal subpopulations defined by expression of Lgr5 and Tspan8 in the nipple, middle and distal areas of mammary glands from the Lgr5-GFP-IRES-$creER^{T2}$ females at the age of 9 months (n=3 mice, three independent experiments).

Supplementary Figure 5. Characterisation of Lgr5- and Tspan8-positive cells in fetal mammary cells

(**a**) Whole-mount 3D confocal image of an entire mammary rudiment from a Lgr5-GFP-IRES-creER^{T2} embryo at E18.5 immunostained for GFP (green), Keratin 5 (red) and K8/K18 (blue) (representative of n=3). (**b**) Colocalisation channel (white) built in the Imaris software showing the colocalisation pattern for K5 (basal) and K8/K18 (luminal). Note that colocalisation mainly occurs in the growing ducts. (**c**) Enlargement from (a) showing the trunk portion of the embryonic mammary gland, where the epithelial layers are more defined and there are rare double-positive cells (for K5 and K8/K18). (**d**, **e**) Enlargement from (a) showing developing branches. (e) shows the colocalisation channel (white). Most cells are double-positive for K5 and K8/K18 (n=3 embryos, three independent experiments). Scale bars, 100 μm (whole-mounts); 20 μm (optical sections). (**f**) Bar chart showing the percentage of Lgr5-GFP^{+} K5/K8^{+}, K5^{+} and K8^{+} cells in the growing ducts (buds) and the trunk (n = 3 embryos). K8 in a-f designates the combined K8/K18 Troma antibody (n=3 embryos, three independent experiments). Error bars represent mean ± SEM; * $p < 0.05$; ****$p < 0.0001$. (**g**) Representative image of a mammary rudiment from a Lgr5-GFP-IRES-creER^{T2} female embryo at E18.5 (n=40). Scale bars, 500 μm. (**h**) Representative FACS plot showing Lgr5 and Tspan8 expression in fetal mammary glands and skin. Embryos were harvested at E18.5 from pregnant Lgr5-GFP-IRES-creER^{T2} females. Mammary glands from female embryos were dissected under a fluorescence microscope and GFP^{+} mammary rudiments from 8-10 embryos were pooled for the preparation of single cell suspensions for FACS analysis (n=8-10 female embryos for each experiment, two independent experiments). **(i)** Comparison of the expression levels of Lgr5-GFP in the fetal and adult MaSC-enriched populations (n=3 mice or 16 female embryos, three independent experiments).

Supplementary Figure 6. Contribution of fetal Lgr5^{+} cells and embryonic origin of Lgr5^{+}Tspan8^{hi} cells in the adult mammary gland

(**a**) Representative confocal images of an E18.5 female mammary ductal tree stained with Tspan8 (red), p63 (green) and DAPI (blue) (n=3 embryos, three independent experiments). (**b**) Embryonic *Lgr5*-expressing cells contribute to the luminal and basal lineages in the adult mammary gland. b1, Whole-mount image of an entire mammary gland from Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mice 11 weeks after tamoxifen injection at 17.5 days of pregnancy (representative of n = 5 mice, three independent experiments). Scale bars: 1 cm. a2, Whole-mount 3D image of a ductal portion labelled for E-cadherin (blue). Inset, optical section from the enlargement showing Tomato^{+} luminal and myoepithelial cells labelled in the duct. Scale bars, 100 μm (whole-mount); 50 μm (optical section). (**c**) Representative FACS plots demonstrating the mammary repopulating activity of embryonic *Lgr5*-expressing cells. Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mice were analysed at 11 weeks after tamoxifen injection at 17.5 days of pregnancy (n=4 mice, four independent experiments). (**d**) Whole-mount 3D confocal image of the proximal portion of a ductal tree from a Lgr5-GFP-IRES-creER^{T2} mouse. EdU was IP injected twice per day from E14.5 to E18.5 and then chased for 11 weeks. The whole-mount was labelled for GFP (green), EdU (red) and E-cadherin (blue). I, Optical section from the whole-mount image showing a ductal portion emanating from the nipple area, with no EdU retention. II, Optical section from the whole-mount image showing a ductal portion in the nipple area displaying EdU retention (representative of n=2 mice, two independent experiments). Scale bars, 300 μm (whole-mount); 50 μm (optical sections).

Supplementary Figure 7. Quiescent mammary stem cells can be activated by synthetic hormonal cues

(**a-b**) Whole-mount 3D confocal images and optical sections of the proximal (a) and distal (b) parts of a mammary gland from a Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mouse two weeks after tamoxifen injection at 9 weeks and immunolabelled for E-cadherin (blue). Mice were treated with vehicle. (I, II) Enlarged areas showing sparsely distributed dtTomato^{+} myoepithelial and luminal cells in vehicle-treated mice (a). Representative images are shown (n=4 mice, two independent experiments). (**c-f**) Whole-mount 3D confocal images and optical sections of the proximal (c, d) and distal (e, f) parts of a mammary gland from a Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mouse two weeks after tamoxifen injection at 9 weeks, during adulthood and immunolabelled for E-cadherin (blue). Mice were treated with the synthetic progestin medroxyprogesterone acetate (MPA) plus estrogen (E + MPA). (d) Enlargement from (c) depicting a branching bud with luminal and basal Tomato^{+} cells derived from Lgr5^{+} cells in mice treated with E + MPA. Right hand panel, optical section of the alveolar bud. (f**)** Enlargement from (e) showing a ductal branch in the distal area comprising only Tomato^{+} basal cells. Representative of n= 4 mice, two independent experiments. Scale bars: 300 μm (whole-mounts), 20 μm (optical sections). (**g**) FACS plots showing EdU^{+} cells in the luminal and basal populations isolated from Lgr5-GFP-IRES-creER^{T2} mice treated with vehicle or E + MPA. Representative plots are shown (n=10 mice, two independent experiments).

Supplementary Figure 8. Quiescent mammary stem cells can be activated during involution

(**a-d**) Whole-mount 3D confocal images and optical sections from enlargements of the proximal (a, b) and distal (c, d) parts of a mammary gland from a Lgr5-GFP-IRES-creER^{T2}/R26R-tdTomato mouse two weeks during involution after tamoxifen injection at 9 weeks of age. Glands were immunolabelled with E-cadherin (blue). Expansion can be

visualised in both regions of the gland. Scale bars: 200 μm (whole-mounts), 20 μm (optical sections). Representative images are shown (n = 3 mice, three independent experiments).

Supplementary Video 1. Movie depicting 3D reconstruction of a mammary sprout at E18.5 (represented in Figure 6Ac1). This movie shows $Lgr5^{+}$ (GFP, green) and $K5^{+}$ (blue) cells. Representative video of n=3 embryos, three independent experiments.

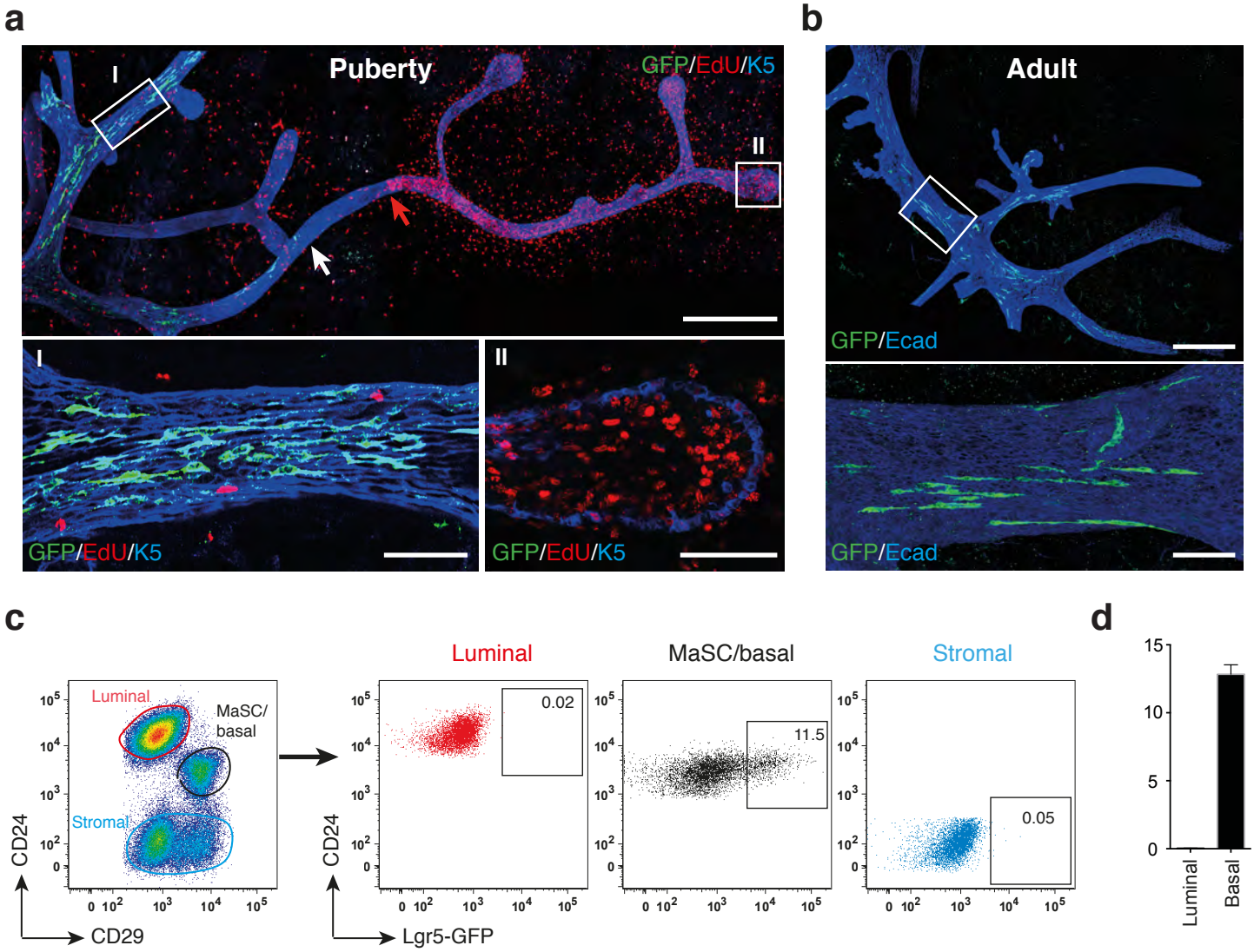

a
Puberty
GFP/EdU/K5
I
II
GFP/EdU/K5
GFP/EdU/K5
b
Adult
GFP/Ecad
GFP/Ecad
c
Luminal
MaSC/basal
Stromal
MaSC/ basal
Stromal
CD24
CD29
0.02
11.5
0.05
CD24
Lgr5-GFP
d
15
10
5
0
Luminal
Basal
e
-2 -1 0 1 2
luminal
Lgr5⁻ basal
Lgr5⁺ basal
Tspan8
Lgr5

Figure 1

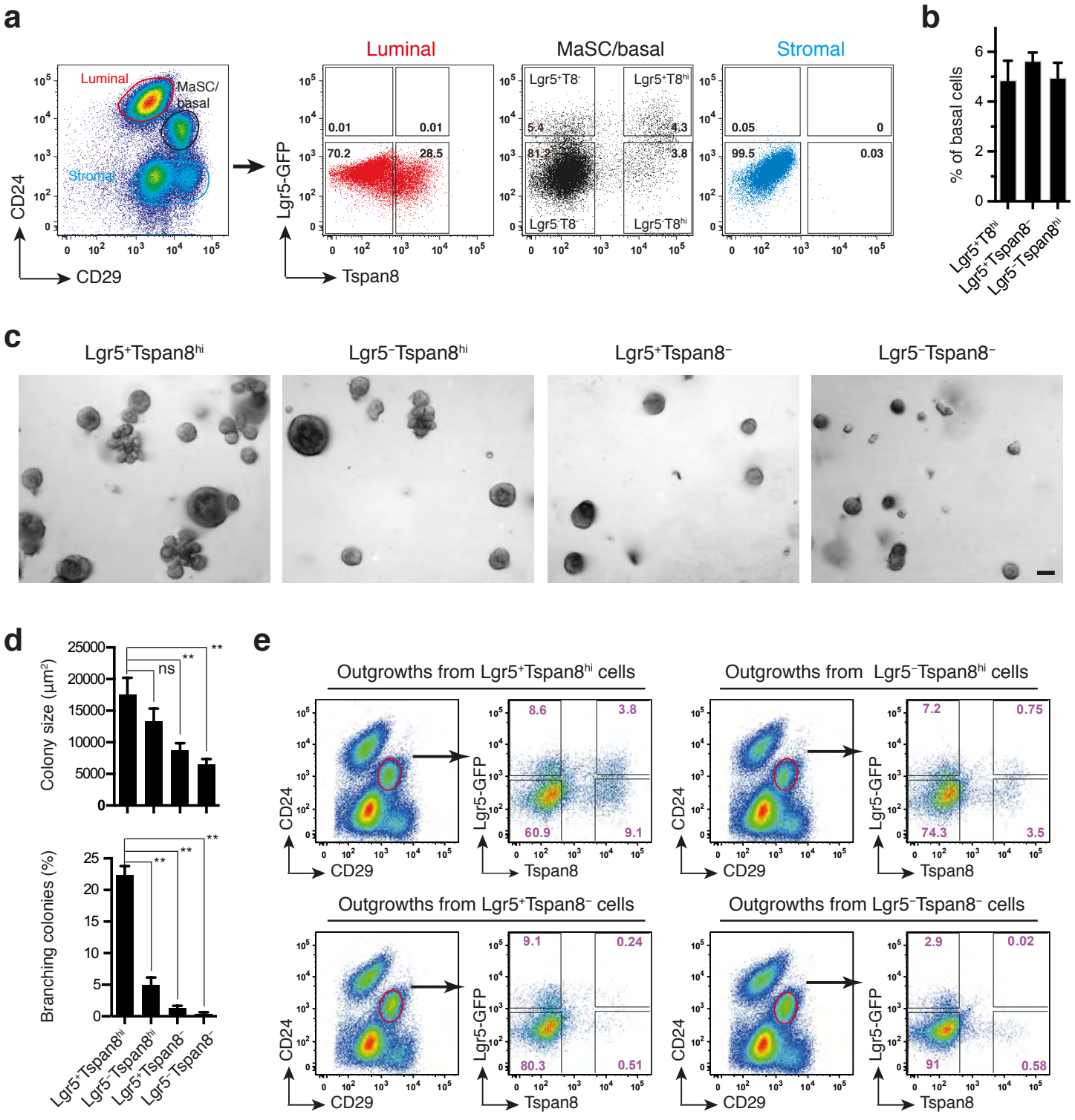

a
Luminal
MaSC/basal
Stromal
CD24
CD29
Lgr5-GFP
Tspan8
Lgr5⁺T8⁻
Lgr5⁺T8ʰⁱ
Lgr5⁻T8⁻
Lgr5⁻T8ʰⁱ
0.01
0.01
70.2
28.5
5.4
4.3
81.2
3.8
0.05
0
99.5
0.03
b
% of basal cells
Lgr5⁺T8ʰⁱ
Lgr5⁺Tspan8⁻
Lgr5⁻Tspan8ʰⁱ
c
Lgr5⁺Tspan8ʰⁱ
Lgr5⁻Tspan8ʰⁱ
Lgr5⁺Tspan8⁻
Lgr5⁻Tspan8⁻
d
Colony size (µm²)
ns
**
Branching colonies (%)
Lgr5⁺Tspan8ʰⁱ
Lgr5⁻Tspan8ʰⁱ
Lgr5⁺Tspan8⁻
Lgr5⁻Tspan8⁻
e
Outgrowths from Lgr5⁺Tspan8ʰⁱ cells
Outgrowths from Lgr5⁻Tspan8ʰⁱ cells
Outgrowths from Lgr5⁺Tspan8⁻ cells
Outgrowths from Lgr5⁻Tspan8⁻ cells
8.6
3.8
60.9
9.1
7.2
0.75
74.3
3.5
9.1
0.24
80.3
0.51
2.9
0.02
91
0.58

Figure 2

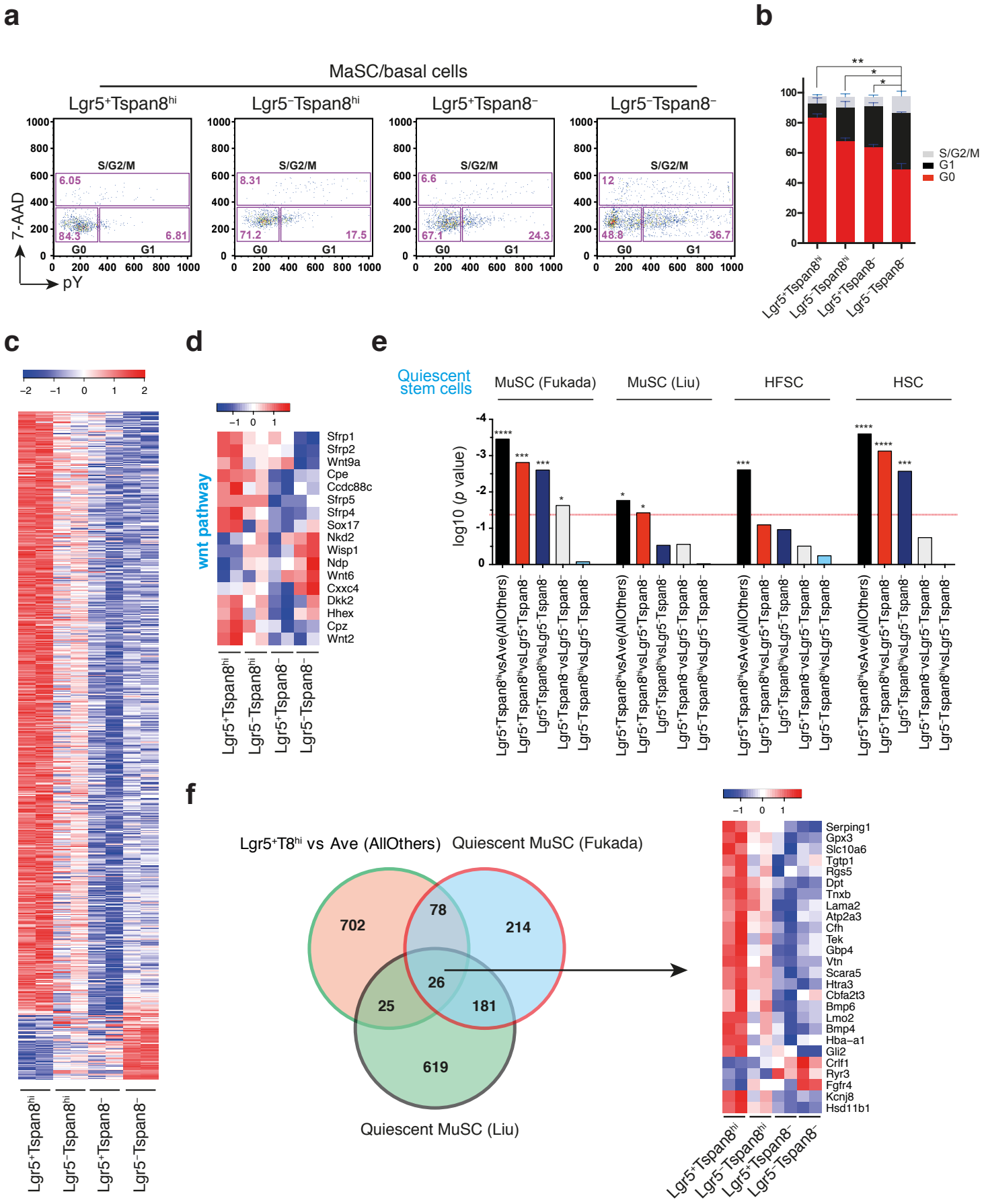

a
MaSC/basal cells
Lgr5+Tspan8hi
Lgr5−Tspan8hi
Lgr5+Tspan8−
Lgr5−Tspan8−
S/G2/M
6.05
84.3
6.81
8.31
71.2
17.5
6.6
67.1
24.3
12
48.8
36.7
G0
G1
7-AAD
pY
b
S/G2/M
G1
G0
c
d
wnt pathway
Sfrp1
Sfrp2
Wnt9a
Cpe
Ccdc88c
Sfrp5
Sfrp4
Sox17
Nkd2
Wisp1
Ndp
Wnt6
Cxxc4
Dkk2
Hhex
Cpz
Wnt2
e
Quiescent stem cells
MuSC (Fukada)
MuSC (Liu)
HFSC
HSC
log10 (p value)
f
Lgr5+T8hi vs Ave (AllOthers)
Quiescent MuSC (Fukada)
Quiescent MuSC (Liu)
702
78
214
26
25
181
619
Serping1
Gpx3
Slc10a6
Tgtp1
Rgs5
Dpt
Tnxb
Lama2
Atp2a3
Cfh
Tek
Gbp4
Vtn
Scara5
Htra3
Cbfa2t3
Bmp6
Lmo2
Bmp4
Hba−a1
Gli2
Crlf1
Ryr3
Fgfr4
Kcnj8
Hsd11b1

Figure 3

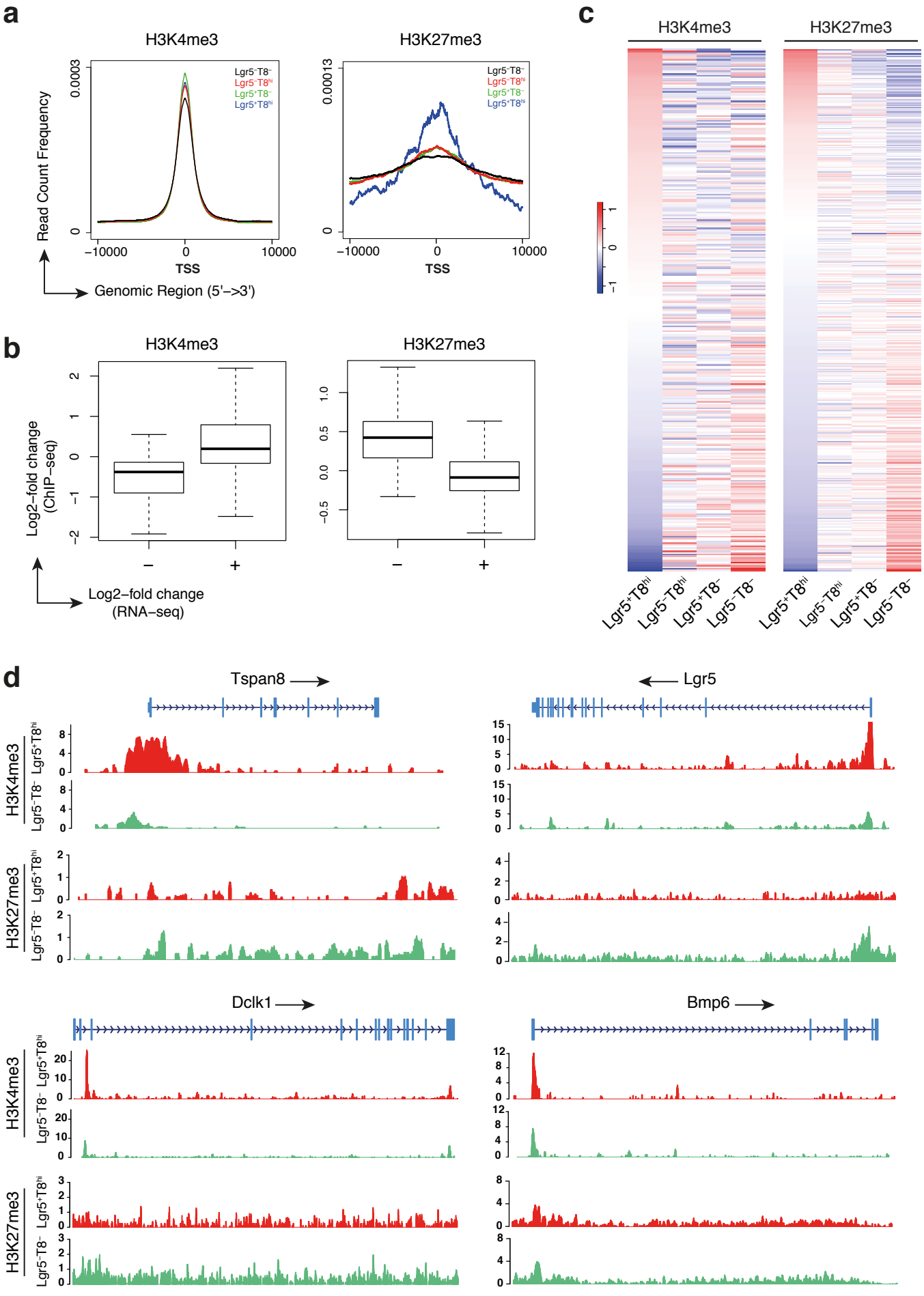

a
H3K4me3
H3K27me3
Lgr5⁻T8⁻
Lgr5⁻T8ʰⁱ
Lgr5⁺T8⁻
Lgr5⁺T8ʰⁱ
Read Count Frequency
0.0003
0.00013
0
−10000
0
10000
TSS
Genomic Region (5'−>3')
b
H3K4me3
H3K27me3
Log2−fold change (ChIP−seq)
Log2−fold change (RNA−seq)
c
H3K4me3
H3K27me3
Lgr5⁺T8ʰⁱ
Lgr5⁻T8ʰⁱ
Lgr5⁺T8⁻
Lgr5⁻T8⁻
d
Tspan8
Lgr5
Dclk1
Bmp6
H3K4me3
H3K27me3

Figure 4

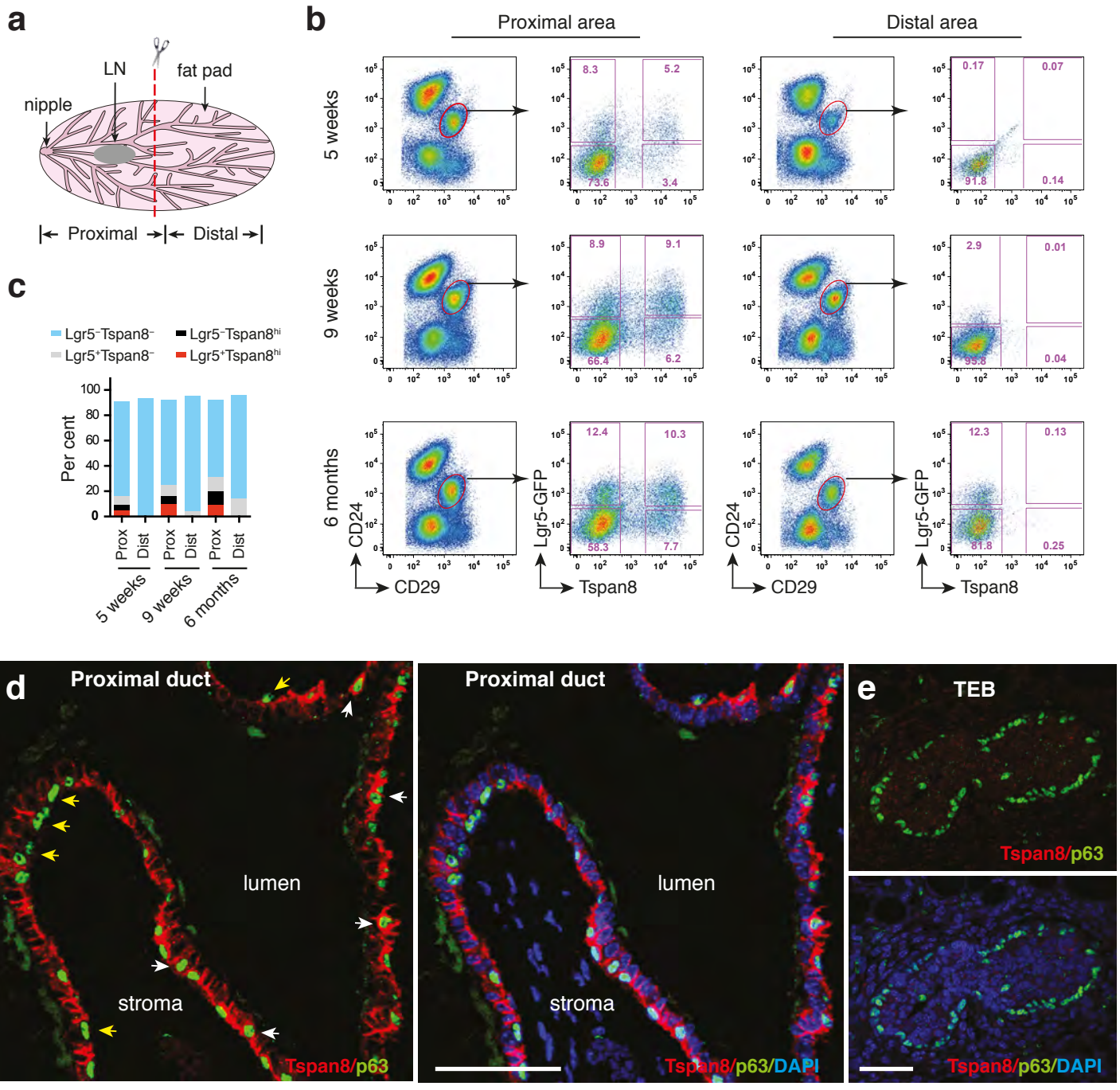
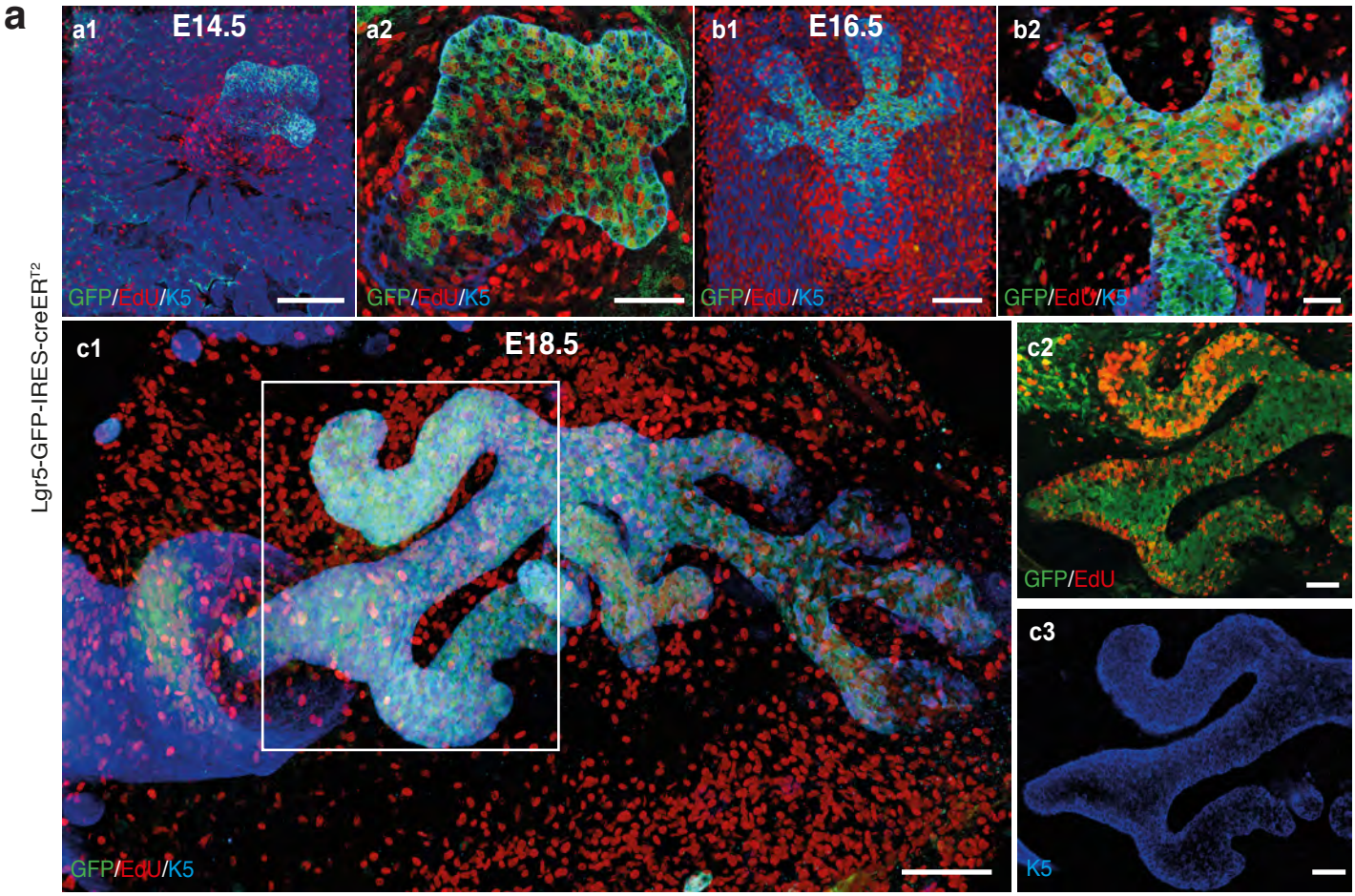
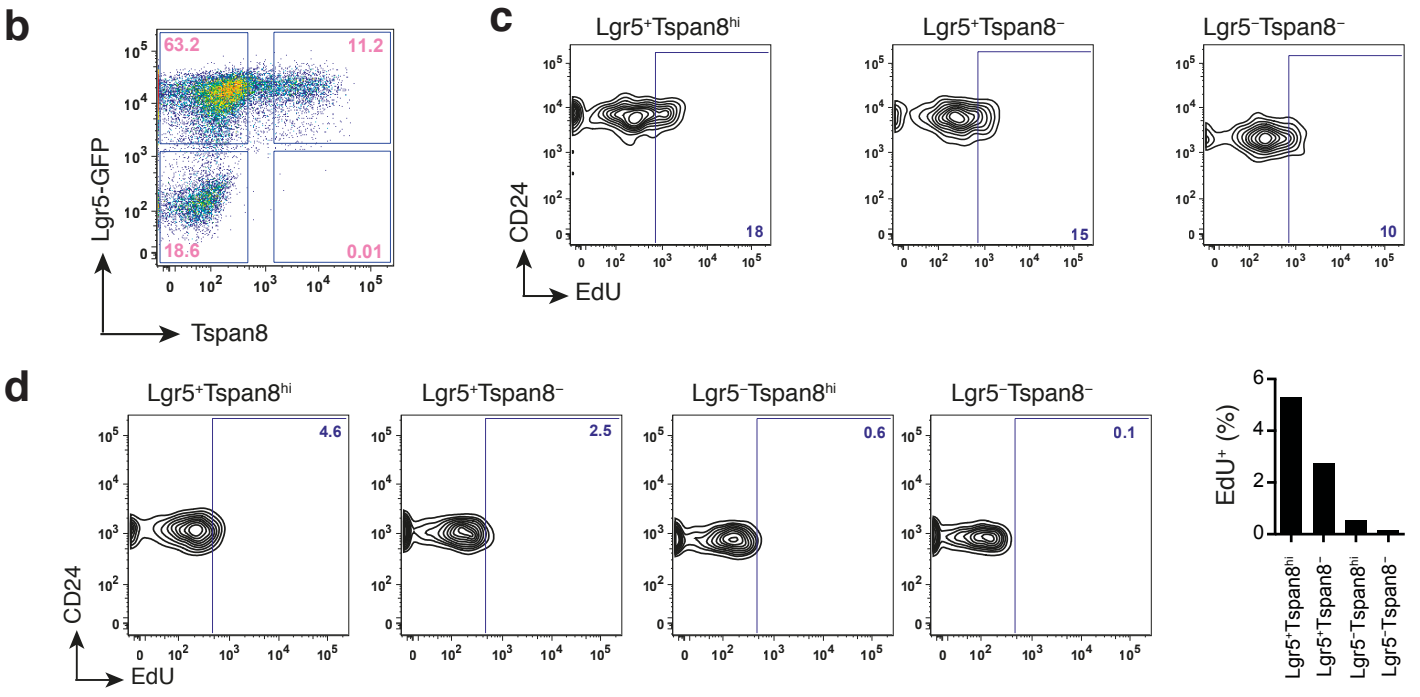

a
LN
fat pad
nipple
Proximal
Distal
b
Proximal area
Distal area
5 weeks
9 weeks
6 months
CD24
CD29
Lgr5-GFP
Tspan8
8.3
5.2
73.6
3.4
0.17
0.07
91.8
0.14
8.9
9.1
66.4
6.2
2.9
0.01
93.8
0.04
12.4
10.3
58.3
7.7
12.3
0.13
81.8
0.25
c
Lgr5⁻Tspan8⁻
Lgr5⁻Tspan8hi
Lgr5⁺Tspan8⁻
Lgr5⁺Tspan8hi
Per cent
100
80
60
40
20
0
Prox
Dist
5 weeks
9 weeks
6 months
d
Proximal duct
lumen
stroma
Tspan8/p63
Proximal duct
lumen
stroma
Tspan8/p63/DAPI
e
TEB
Tspan8/p63
Tspan8/p63/DAPI

Figure 5


a
a1
E14.5
a2
b1
E16.5
b2
Lgr5-GFP-IRES-creER^{T2}
GFP/EdU/K5
c1
E18.5
c2
GFP/EdU
c3
K5


b
63.2
11.2
18.6
0.01
Lgr5-GFP
Tspan8
c
Lgr5$^+$Tspan8^{hi}
Lgr5$^+$Tspan8$^-$
Lgr5$^-$Tspan8$^-$
18
15
10
CD24
EdU

d
Lgr5$^+$Tspan8^{hi}
Lgr5$^+$Tspan8$^-$
Lgr5$^-$Tspan8^{hi}
Lgr5$^-$Tspan8$^-$
4.6
2.5
0.6
0.1
CD24
EdU
EdU$^+$ (%)

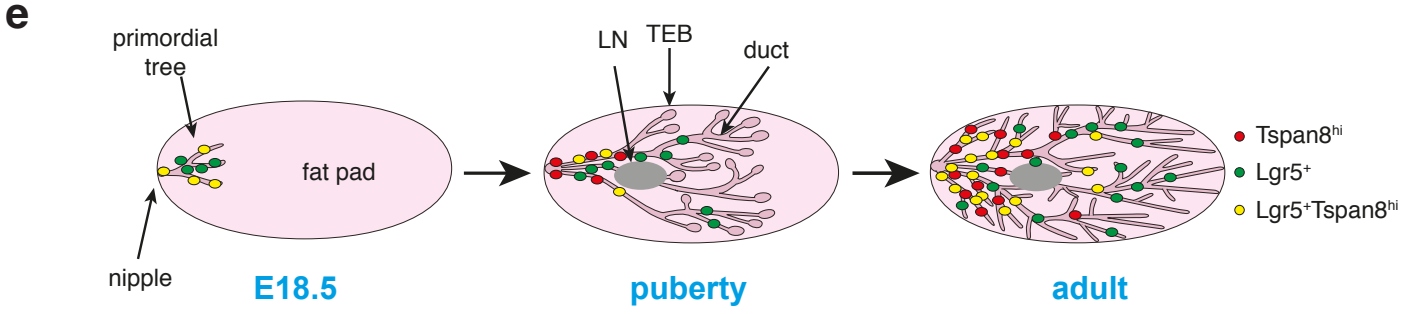

e
primordial tree
fat pad
nipple
E18.5
LN
TEB
duct
puberty
adult
Tspan8^{hi}
Lgr5$^+$
Lgr5$^+$Tspan8^{hi}

Figure 6

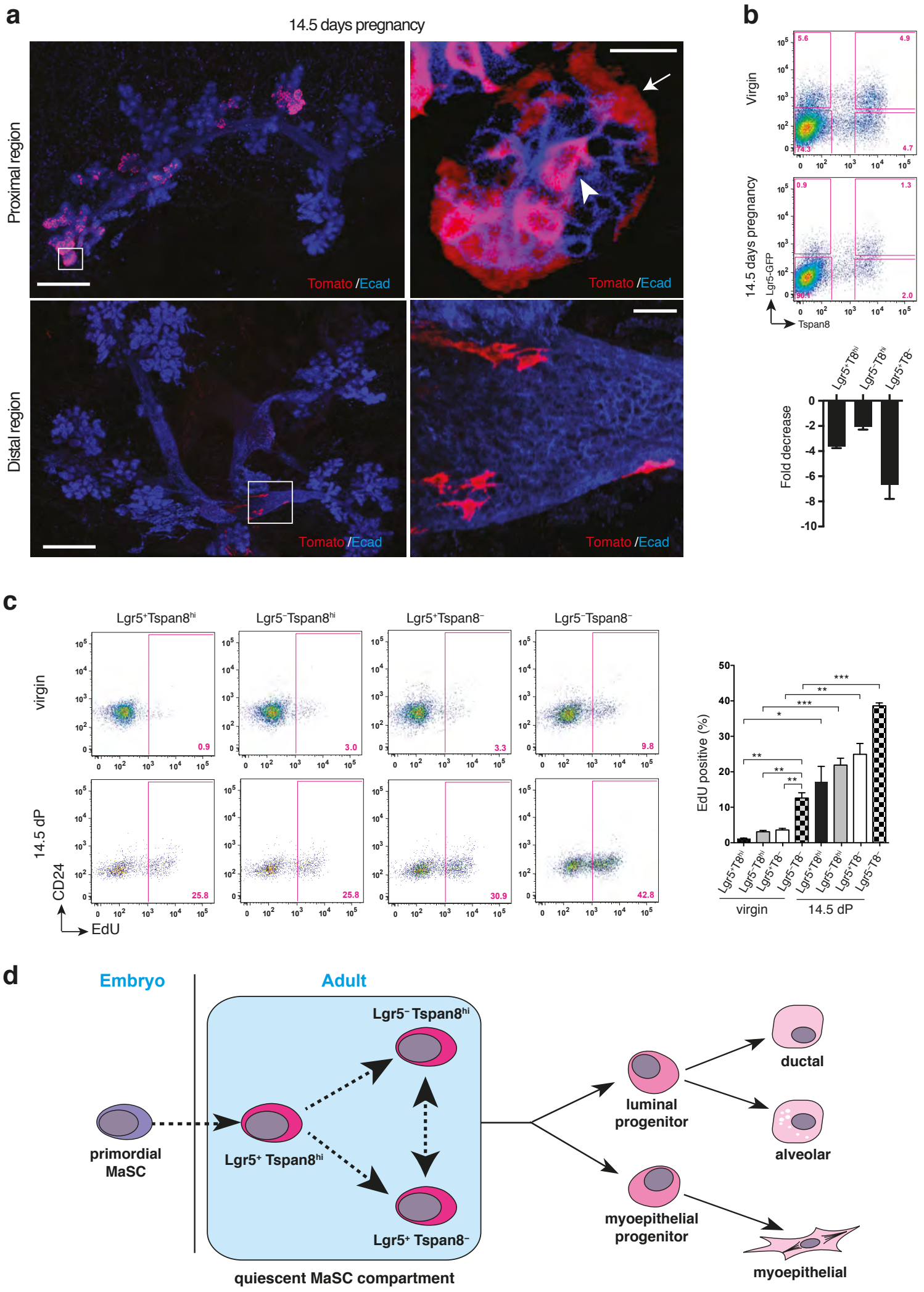
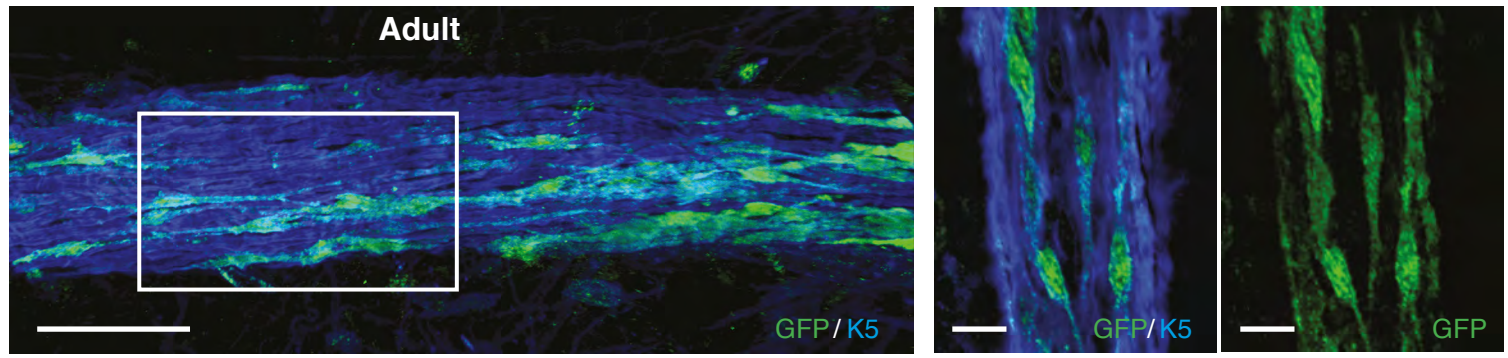
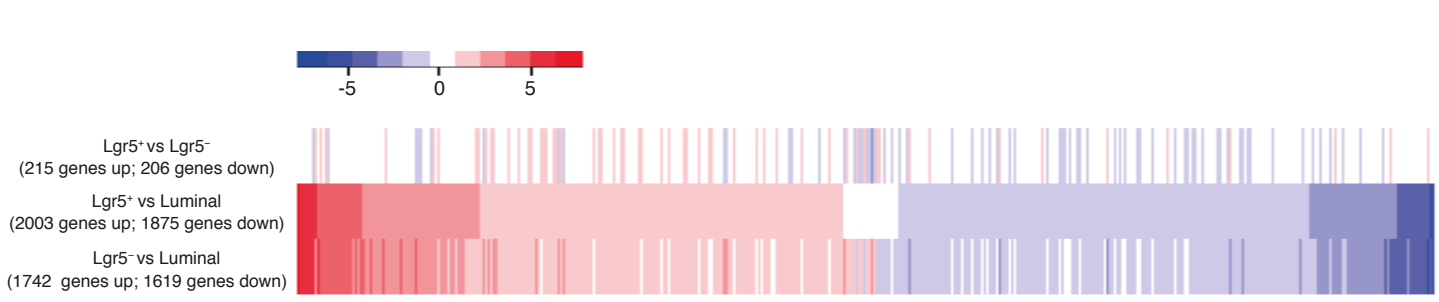
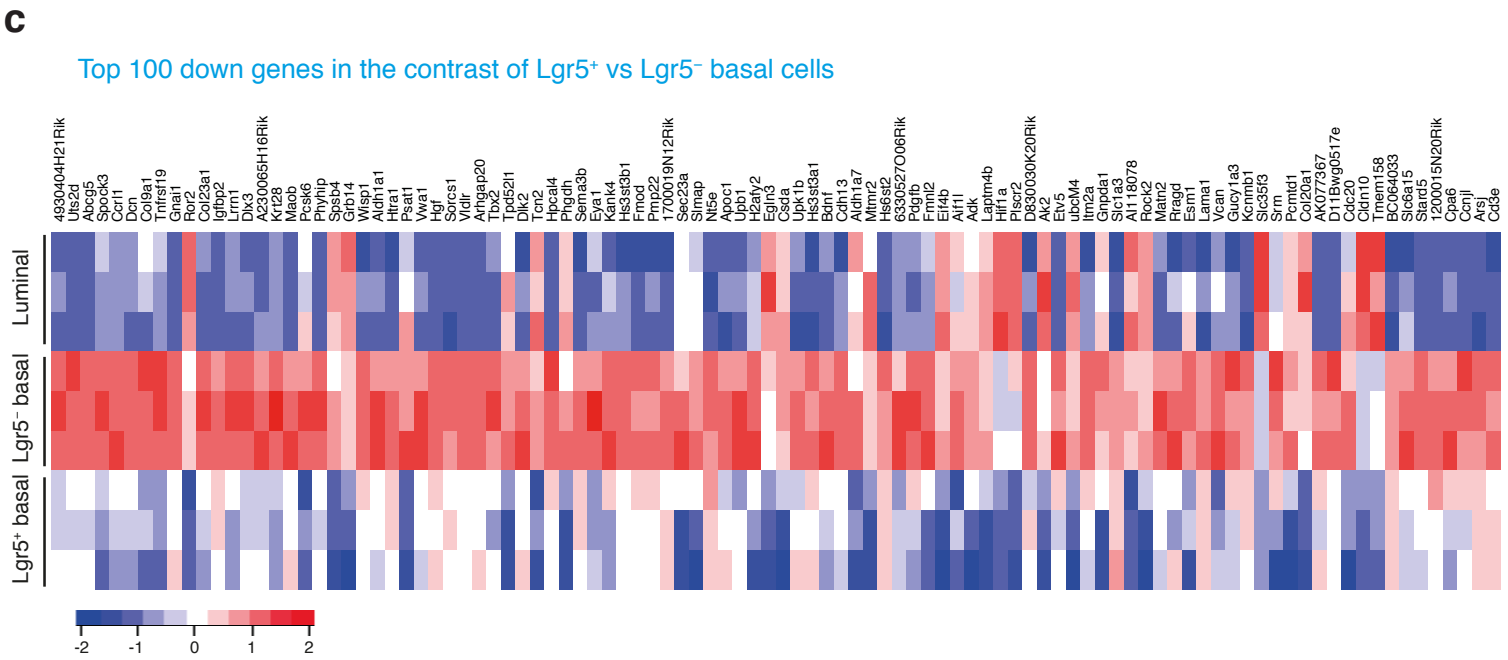
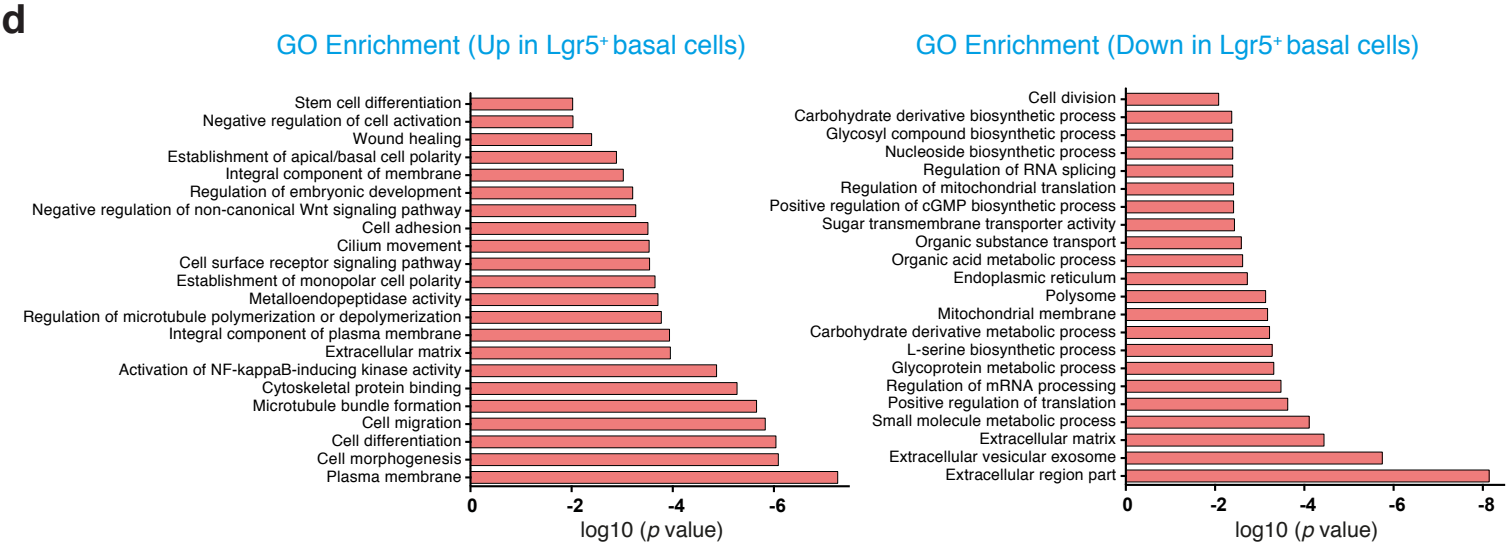

a
14.5 days pregnancy
Proximal region
Distal region
Tomato /Ecad
Tomato /Ecad
Tomato /Ecad
Tomato /Ecad
b
Virgin
14.5 days pregnancy
Lgr5-GFP
Tspan8
5.6
4.9
74.3
4.7
0.9
1.3
90.1
2.0
Lgr5⁺T8ʰⁱ
Lgr5⁻T8ʰⁱ
Lgr5⁺T8⁻
Fold decrease
0
-2
-4
-6
-8
-10
c
Lgr5⁺Tspan8ʰⁱ
Lgr5⁻Tspan8ʰⁱ
Lgr5⁺Tspan8⁻
Lgr5⁻Tspan8⁻
virgin
14.5 dP
CD24
EdU
0.9
3.0
3.3
9.8
25.8
25.8
30.9
42.8
EdU positive (%)
50
40
30
20
10
0
***
**
***
*
**
**
**
Lgr5⁺T8ʰⁱ
Lgr5⁻T8ʰⁱ
Lgr5⁺T8⁻
Lgr5⁻T8⁻
Lgr5⁺T8ʰⁱ
Lgr5⁻T8ʰⁱ
Lgr5⁺T8⁻
Lgr5⁻T8⁻
virgin
14.5 dP
d
Embryo
Adult
primordial MaSC
Lgr5⁻ Tspan8ʰⁱ
Lgr5⁺ Tspan8ʰⁱ
Lgr5⁺ Tspan8⁻
quiescent MaSC compartment
luminal progenitor
ductal
alveolar
myoepithelial progenitor
myoepithelial

Figure 7

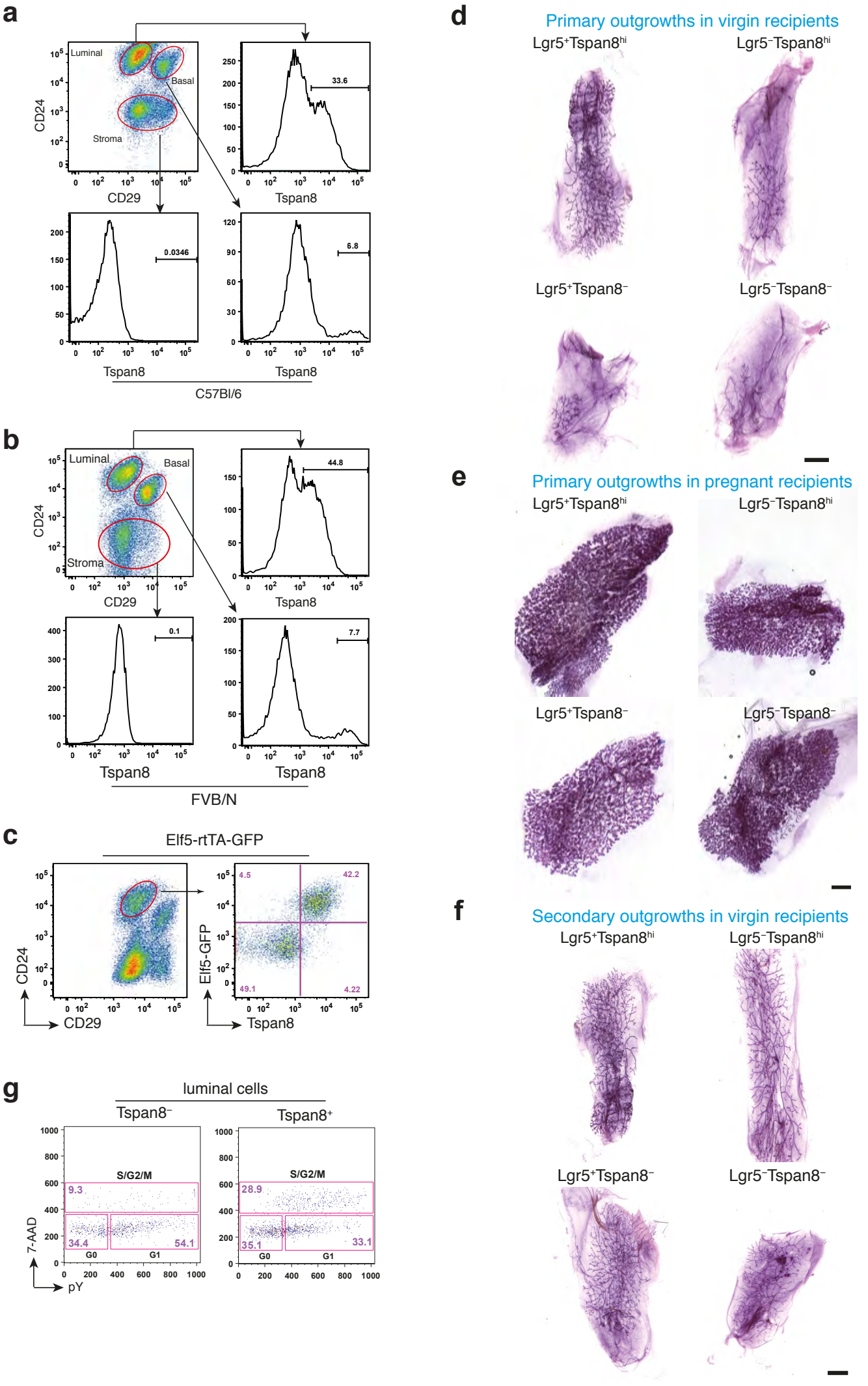








Supplementary Figure 1

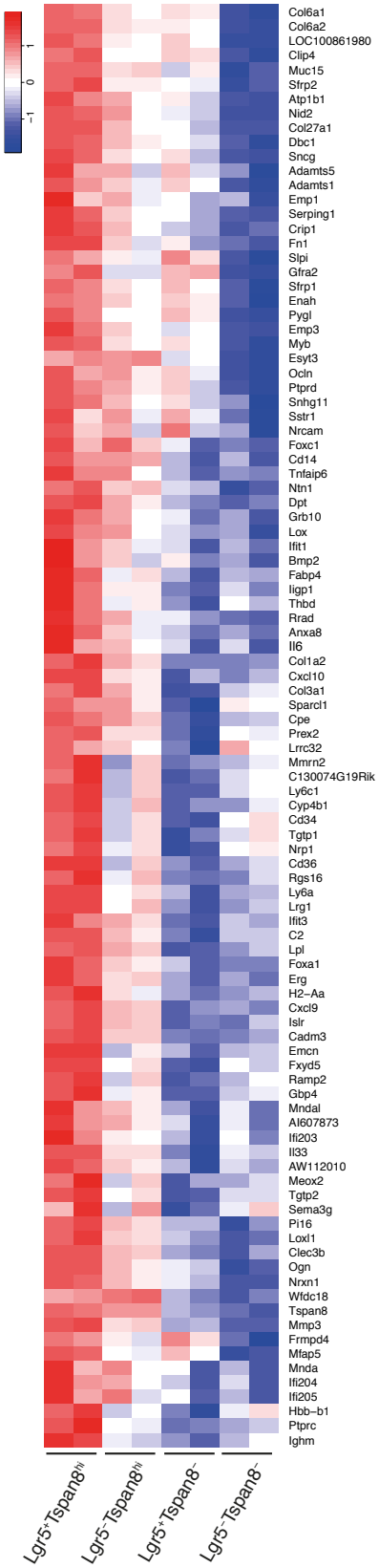


Supplementary Figure 2

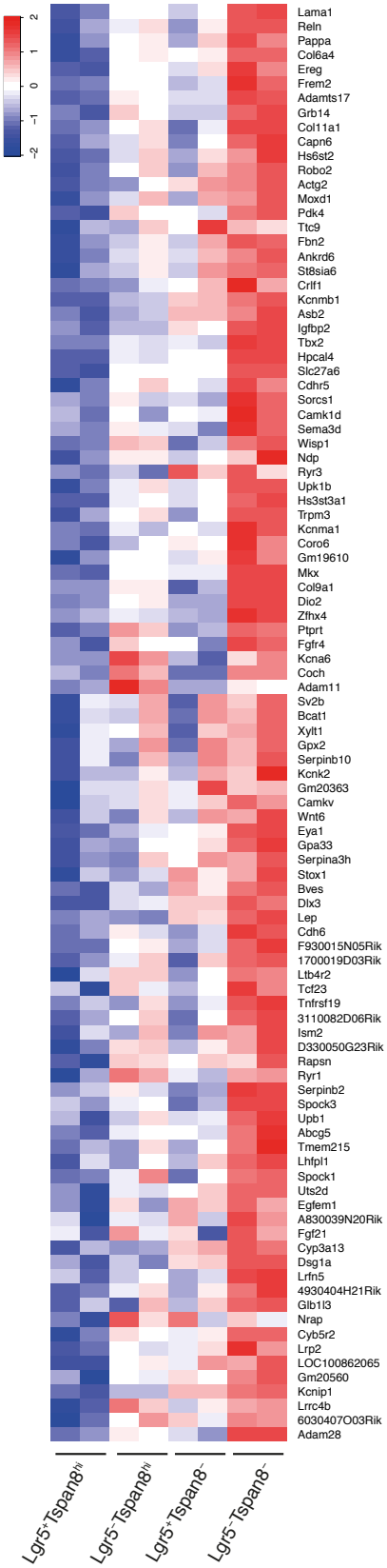
a

Gene	Lgr5⁻Tspan8⁻	Lgr5⁻Tspan8^{hi}	Lgr5⁺Tspan8⁻	Lgr5⁺Tspan8^{hi}
Acta2	12.76	12.03	12.45	11.85
Krt5	11.05	11.15	11.36	11.31
Krt14	12.23	12.37	12.57	12.44
Krt15	8.43	8.21	8.72	8.51
Krt17	11.34	11.41	11.92	11.55
Trp63	8.03	7.58	7.81	7.45
Col14a1	9.67	9.61	10.04	9.81
Actg2	8.45	7.2	8.0	6.64
Myh11	12.17	11.18	11.68	11.01
Snai2	6.49	6.18	6.32	6.18
Id4	6.65	6.6	6.51	6.37

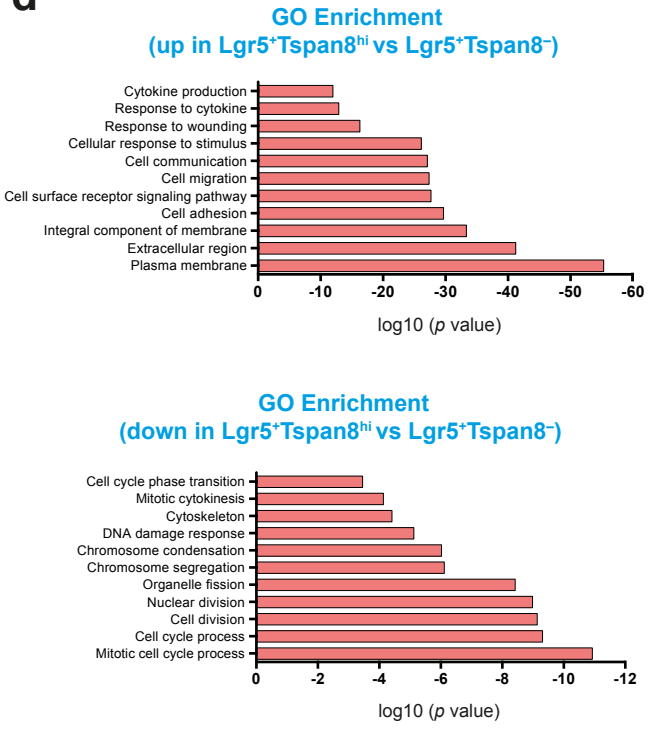
b



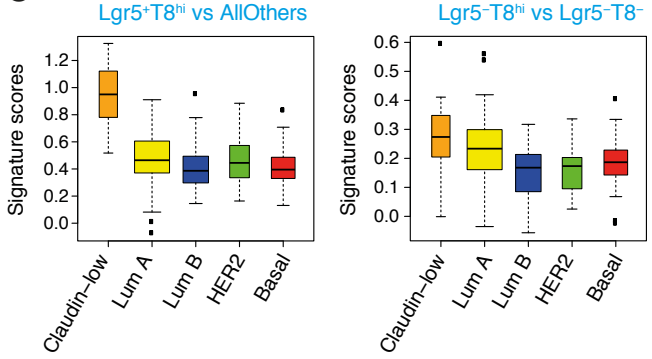
c



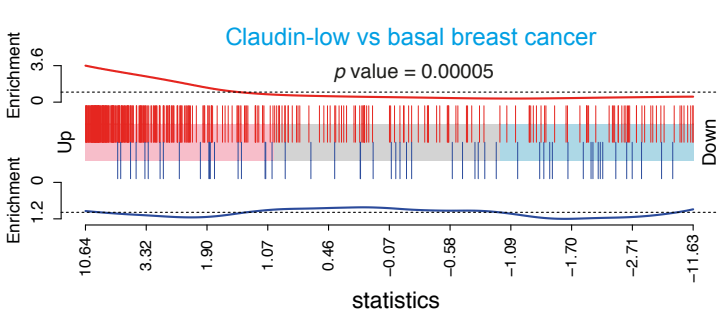
d



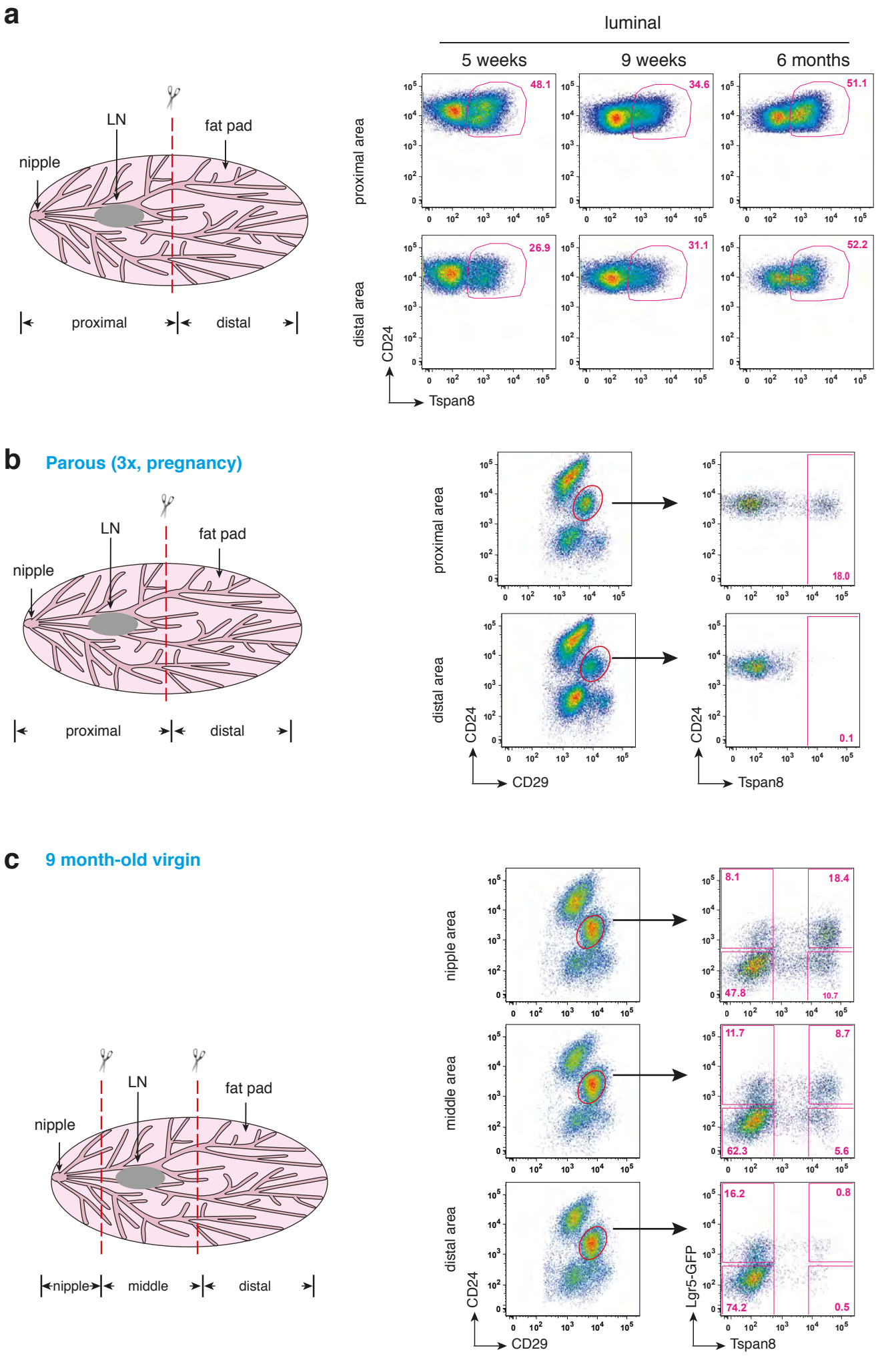
e



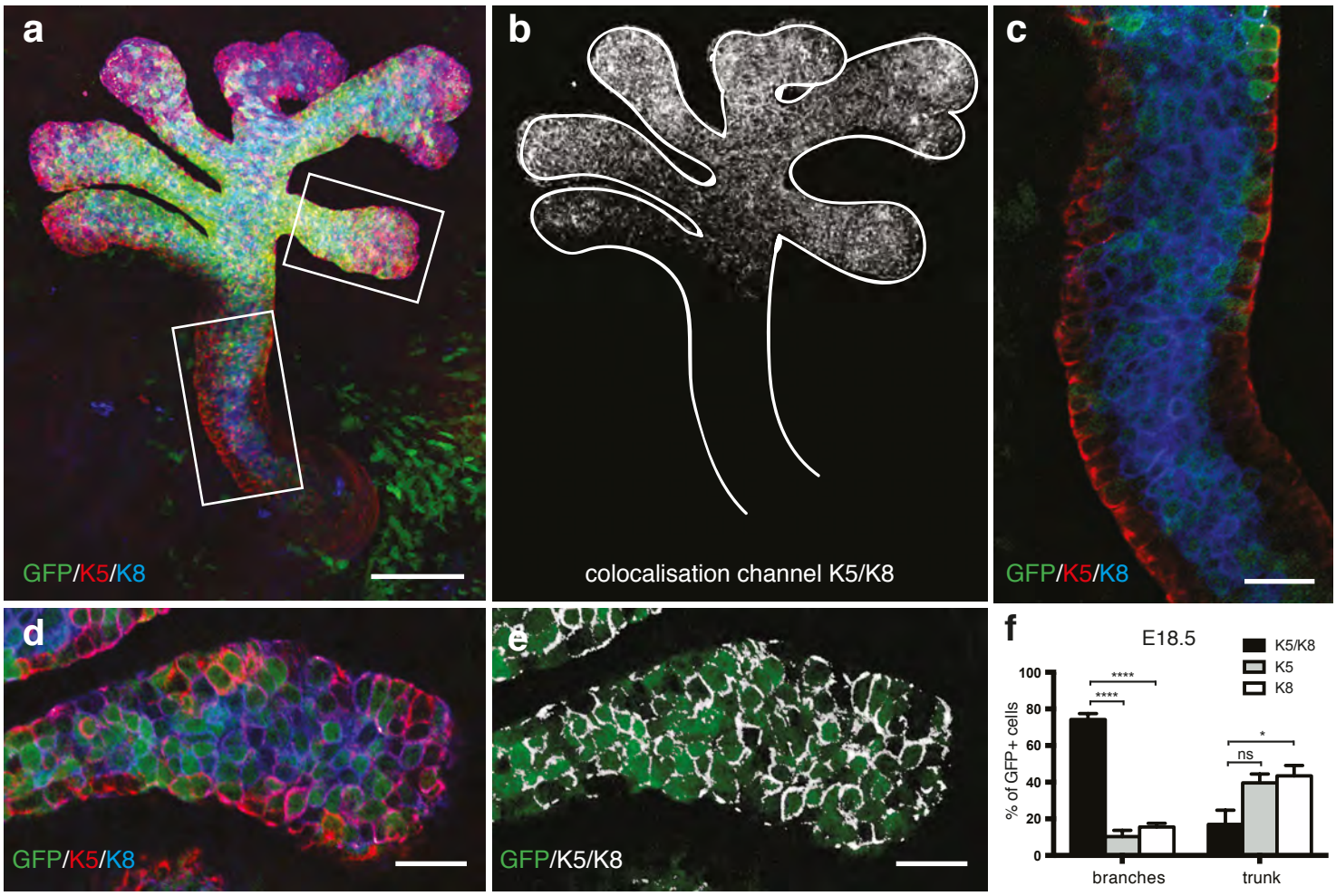
f



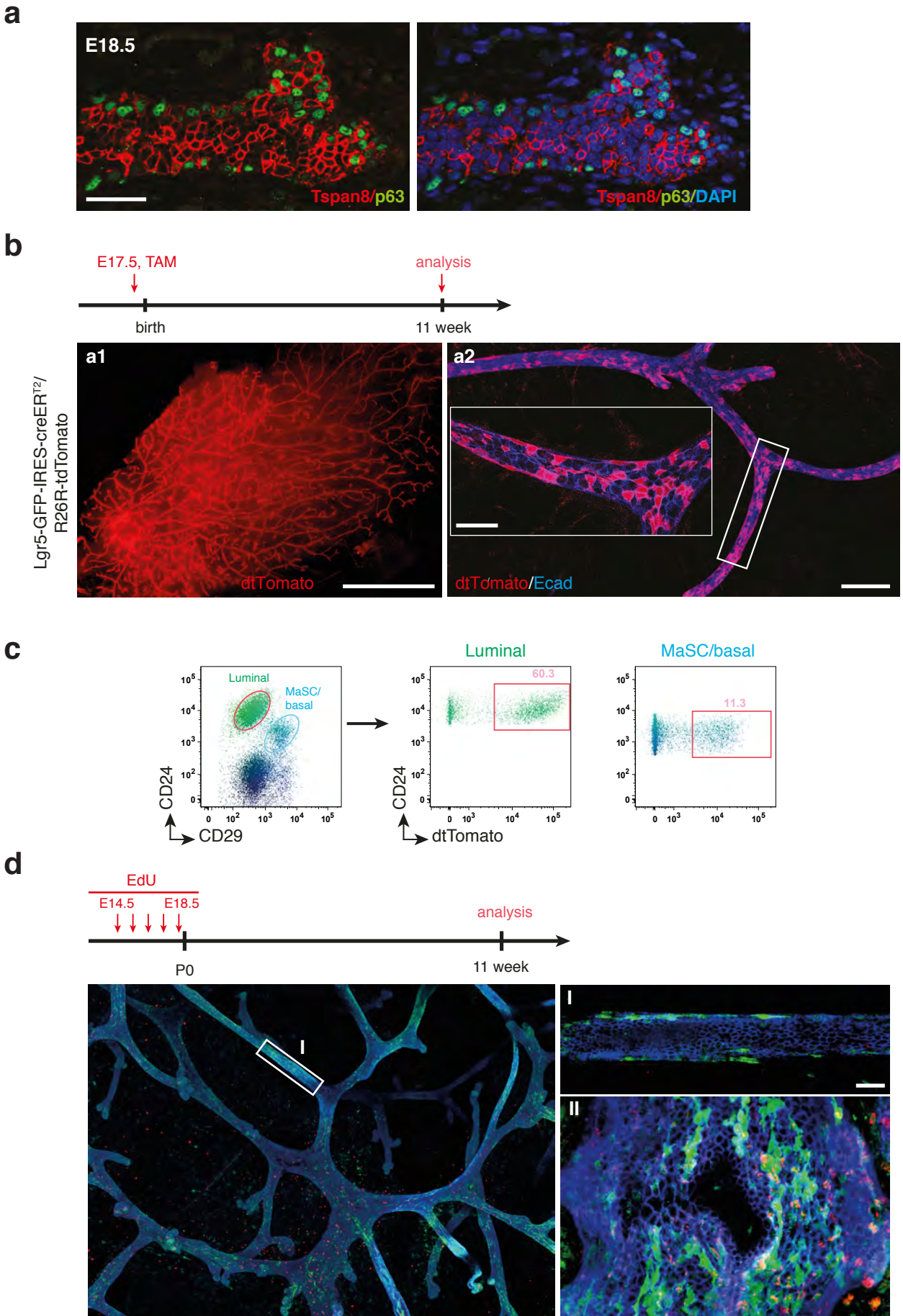
Supplementary Figure 3



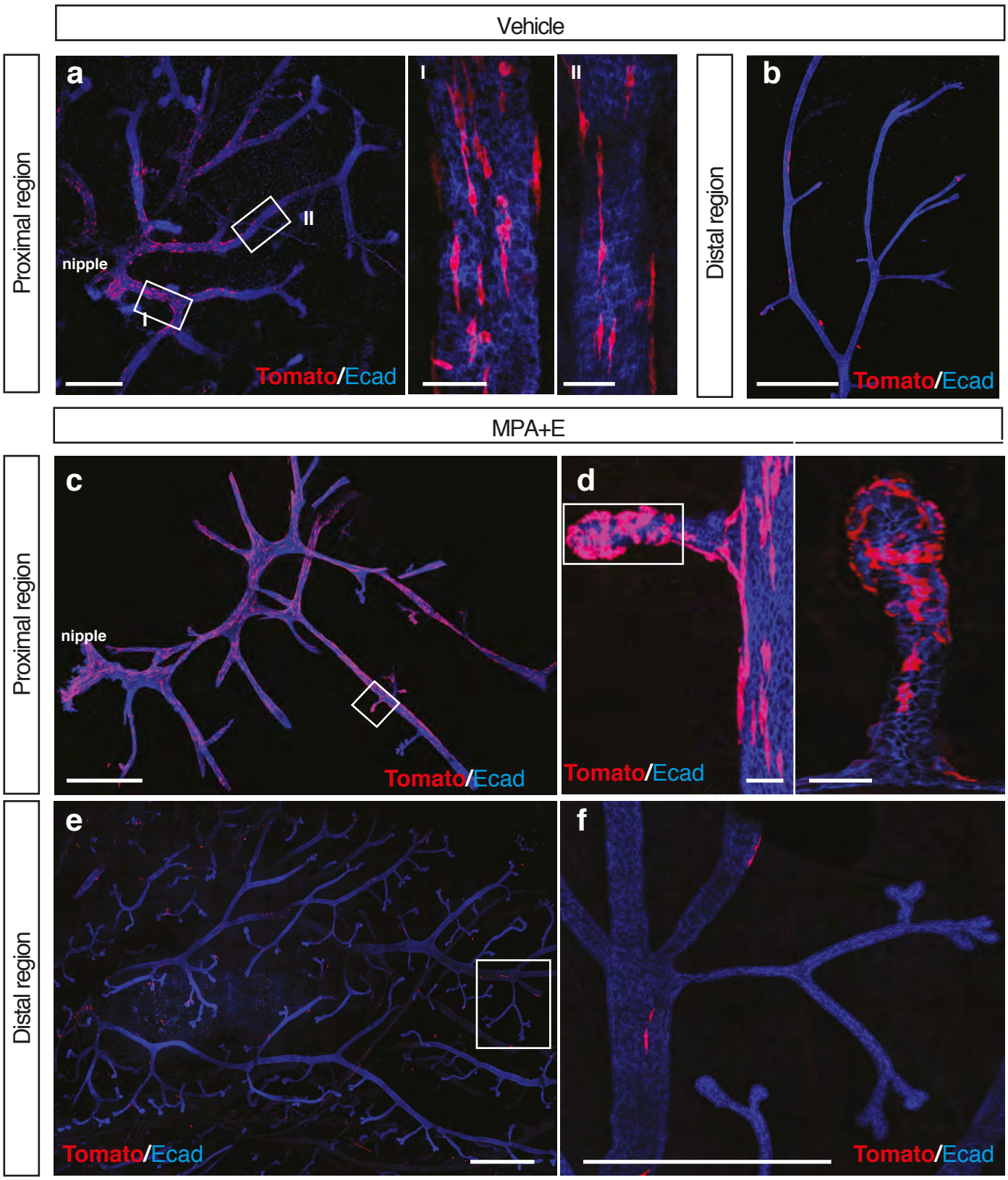
Supplementary Figure 4

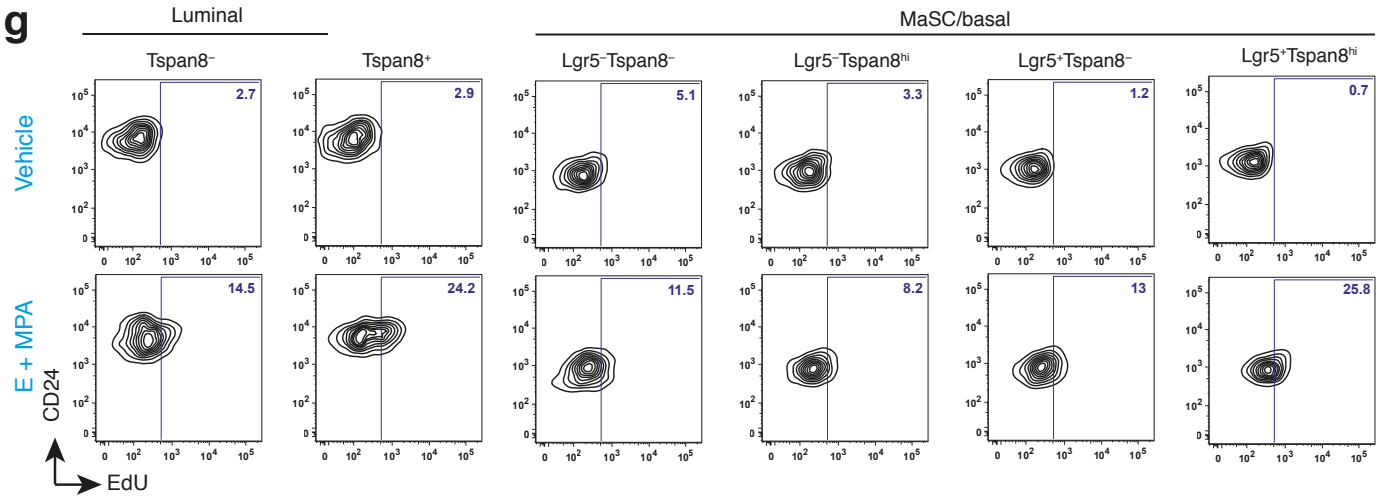


Supplementary Figure 5

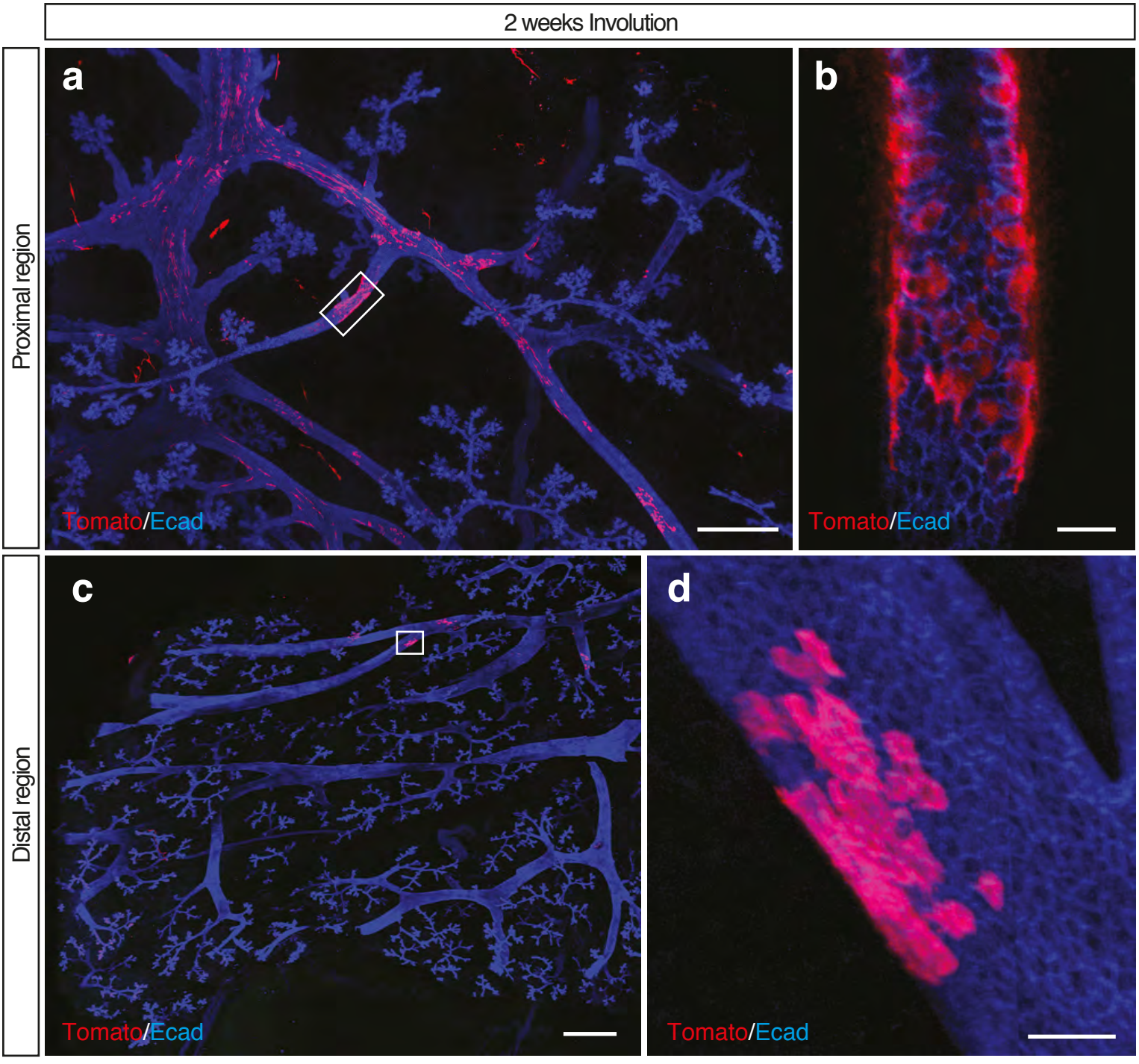


Supplementary Figure 6





Supplementary Figure 7



Supplementary Figure 8