






ARTICLE

DOI: 10.1038/s41467-017-01560-x

OPEN

# Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling

Bhupinder Pal<sup>1,2</sup>, Yunshun Chen<sup>2,3</sup>, François Vaillant <sup>1,2</sup>, Paul Jamieson<sup>1</sup>, Lavinia Gordon <sup>4</sup>, Anne C. Rios<sup>1,2</sup>, Stephen Wilcox<sup>5</sup>, Naiyang Fu<sup>1,2</sup>, Kevin He Liu<sup>1</sup>, Felicity C. Jackling<sup>1</sup>, Melissa J. Davis<sup>3,6</sup>, Geoffrey J. Lindeman <sup>1,7,8</sup>, Gordon K. Smyth <sup>3,9</sup> & Jane E. Visvader <sup>1,2</sup>

The mammary epithelium comprises two primary cellular lineages, but the degree of heterogeneity within these compartments and their lineage relationships during development remain an open question. Here we report single-cell RNA profiling of mouse mammary epithelial cells spanning four developmental stages in the post-natal gland. Notably, the epithelium undergoes a large-scale shift in gene expression from a relatively homogeneous basal-like program in pre-puberty to distinct lineage-restricted programs in puberty. Interrogation of single-cell transcriptomes reveals different levels of diversity within the luminal and basal compartments, and identifies an early progenitor subset marked by CD55. Moreover, we uncover a luminal transit population and a rare mixed-lineage cluster amongst basal cells in the adult mammary gland. Together these findings point to a developmental hierarchy in which a basal-like gene expression program prevails in the early post-natal gland prior to the specification of distinct lineage signatures, and the presence of cellular intermediates that may serve as transit or lineage-primed cells.

<sup>1</sup>ACRF Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia. <sup>2</sup>Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia. <sup>3</sup>Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia. <sup>4</sup>Australian Genome Research Facility, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia. <sup>5</sup>Systems Biology & Personalised Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia. <sup>6</sup>Department of Biochemistry and Molecular Biology, The University of Melbourne, Parkville, VIC 3010, Australia. <sup>7</sup>Department of Medicine, The University of Melbourne, Parkville, VIC 3010, Australia. <sup>8</sup>Parkville Familial Cancer Centre and Department of Medical Oncology, The Royal Melbourne Hospital and Peter MacCallum Cancer Centre, Parkville, VIC 3050, Australia. <sup>9</sup>School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia. Bhupinder Pal, Yunshun Chen and François Vaillant contributed equally to this work. Gordon K. Smyth and Jane E. Visvader jointly supervised this work. Correspondence and requests for materials should be addressed to J.E.V. (email: [visvader@wehi.edu.au](mailto:visvader@wehi.edu.au))

The mammary gland is a remarkably dynamic organ whose epithelium undergoes dramatic changes during morphogenesis and the reproductive cycle. Architecturally, the epithelium comprises two primary cellular lineages: an inner layer of luminal cells that surround the lumen and an outer layer of myoepithelial cells that lie in a basal position adjacent to the basement membrane. Cumulative evidence based on transplantation, colony-forming assays, and lineage tracing studies in mouse models indicates the presence of stem and committed progenitor cells that lie upstream of the mature epithelial cell types (myoepithelial, ductal luminal, and alveolar luminal) resident in the ductal tree<sup>1, 2</sup>. However, little is known about the spatio-temporal regulation of molecular pathways important for lineage specification in the mammary gland, thus highlighting the need for more refined transcriptional mapping studies.

Morphogenesis of the mammary gland occurs through distinct stages, with the majority of development taking place in the post-natal animal<sup>3</sup>. At birth, a rudimentary ductal tree exists and extends by allometric growth until puberty. During this stage, the epithelium undergoes massive expansion to form a highly elaborate and branched ductal tree that characterizes the adult gland. Ductal elongation and branching during puberty is largely driven by terminal endbuds (TEBs) located at the termini of the growing ducts. The gene expression portraits of different mammary epithelial cell types have been described at a population level<sup>4–8</sup> but not at the single-cell level. Hence, a comprehensive understanding of heterogeneity within the different epithelial populations is lacking.

The global analysis of transcriptomes at the single-cell level has emerged as a powerful tool to understand cellular heterogeneity and genomic states. Such studies have provided valuable insights into lineage relationships, rare cellular subsets, and novel biomarkers for diverse organs. For example, single-cell RNA-seq (scRNA-seq) analysis of cerebral cortex cells from the developing brain<sup>9</sup>, developing heart<sup>10</sup>, the adult mouse forebrain<sup>11</sup>, lung epithelium<sup>12</sup>, intestinal cells<sup>13</sup>, olfactory neurons<sup>14</sup>, and pancreatic cells<sup>15</sup> has revealed novel cellular subsets based on transcriptional and/or signaling pathways. Moreover, this methodology has been utilized to follow the induction of mouse embryonic fibroblasts to neuronal cells, identifying distinct intermediate stages during reprogramming<sup>16</sup>. The identification of lineage-primed or multipotent cells through single-cell analysis of haematopoietic<sup>17, 18</sup>, pancreatic<sup>19</sup> and intestinal cells<sup>20</sup> has provided important insights into rare cellular states.

Here we present comprehensive single-cell transcriptomes of epithelial cells in the post-natal mouse mammary gland at different developmental stages spanning pre-puberty, puberty, adulthood and pregnancy, as well as at different points of the estrus cycle. Transcript profiling was performed on two different platforms: the 10X Genomics Chromium System<sup>21</sup> for large-scale analyses and the Fluidigm C1 platform for high-resolution sequencing. Determination and compilation of the transcriptomes of individual cells across distinct developmental stages revealed that a major transcriptional switch occurs at the onset of puberty from a relatively homogeneous to heterogeneous landscape. In the adult mammary gland, the luminal compartment was more stratified than the basal population, but rare basal subsets could be delineated. Interestingly, mixed-lineage intermediates poised towards a luminal fate were identified in purified basal cells of the adult as well as in pubertal and pregnant mammary glands. Collectively, these single-cell datasets spanning different developmental stages provide a valuable resource to decipher regulatory decisions in the mammary epithelium executed at the single-cell level.

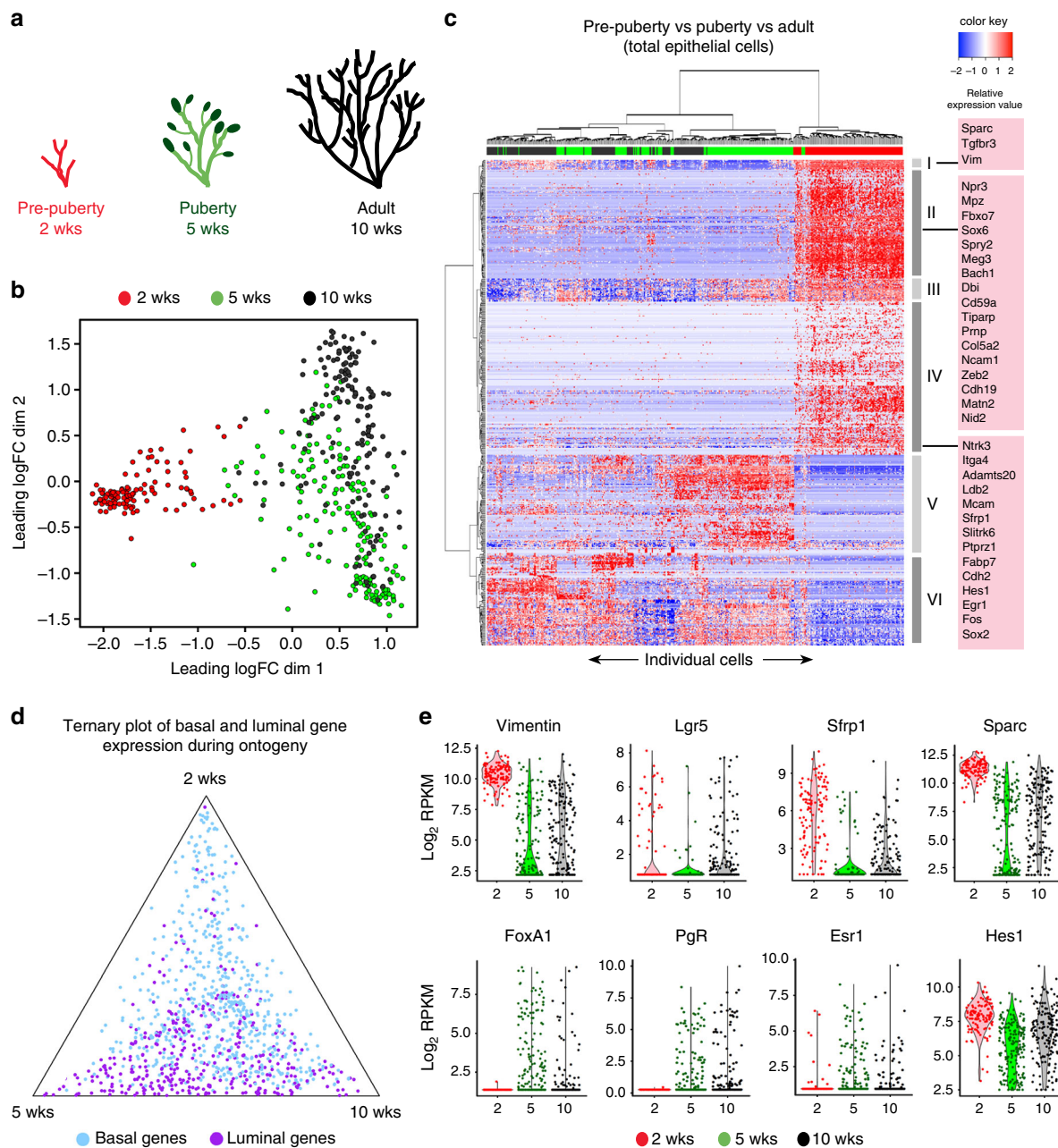
## Results

**Reprogramming of the epithelial transcriptional landscape at puberty.** To explore changes in cell diversity in the mammary epithelium during post-natal gland development (Fig. 1a), we performed scRNA-seq. We employed the Fluidigm C1 platform for single-cell capture to achieve in-depth expression profiling of individual mammary epithelial cells through different developmental stages (Supplementary Table 1). Combined with Illumina sequencing, this platform allowed us to assign up to 2 million read-pairs and to detect up to 8000 genes per cell (Supplementary Fig. 1a). The presence of a single cell per capture site was confirmed by direct visualization prior to sequencing.

We began by examining molecular heterogeneity in the mammary epithelium as the gland transitions from a rudimentary state in the early post-natal gland (2 weeks i.e. 14 days) to a developing tree in mid-puberty (5 weeks) and a highly branched ductal network in the adult (10 weeks) (Fig. 1a). An unsupervised plot of distances between the single-cell transcriptomes of 460 cells spanning the three developmental stages revealed a clear separation between pre-pubertal epithelial cells and those from pubertal or adult mice (Fig. 1b). A leading fold-change of more than 4-fold separates most pre-pubertal cells from the 5 or 10 week clusters. Pre-pubertal cells are relatively homogeneous whereas pubertal and adult cells are spread over a wider range and show more overlap (Fig. 1b). To understand the major genes driving these dramatic transcriptional differences, we clustered the cells using the most differentially expressed (DE) genes across the three developmental stages (Fig. 1c). Genes that were highly expressed in the pre-pubertal gland included the known basal genes *Vimentin* (*Vim*), *Ncam1*, the secreted matrix protein *Sparc*, the Wnt pathway modulator *Sfrp1*, and the transcriptional regulators, *Zeb2*, *Sox6*, *Egr1*, *Fos*, *Ldb2*, and *Bach1* (Gene clusters I, II, and IV in Fig. 1c; Supplementary Table 2). Gene cluster III included *Nras*, *Pten*, *Nf1*, and *Gsk3b* and is interesting in that it is expressed in both pre-pubertal cells and most pubertal cells but not in adult cells. Genes restricted to pubertal and adult mammary glands are discussed below.

To further interpret the lineage switches that occur upon ductal morphogenesis, we examined the behavior of genes identified in previous studies<sup>22</sup> (Supplementary Table 3) to distinguish the basal/myoepithelial (denoted basal from hereon) and luminal lineages in the mammary gland. This revealed that basal gene expression occurs throughout all three developmental stages whereas expression of almost all luminal genes is confined to puberty and adulthood (Fig. 1d). Indeed, gene cluster V in the heatmap highlights the expression of definitive luminal genes (e.g., *Epcam*, *Keratin* (*Krt*)8/18/19) from puberty onwards (Fig. 1c; Supplementary Table 2). Violin plots of gene expression (Fig. 1e) illustrate that most pre-pubertal cells express abundant *Vim*, *Lgr5*, *Sfrp1*, and *Sparc* transcripts, whereas luminal lineage determinant genes such as the estrogen and progesterone receptors (*Esr1*, *PgR*) and *FoxA1* were expressed at low levels. By contrast, the Notch pathway effector gene *Hes1* was found in the majority of cells in 2-week-old mammary glands. Thus, a basal gene expression program prevails in pre-pubertal epithelial cells.

Interestingly, different sets of basal genes were expressed through pre-puberty, puberty, and adulthood, and a substantial subset was specific to early post-natal glands (Fig. 1c). To explore this further, we clustered individual cells according to their expression of basal and luminal lineage genes<sup>22</sup> (Supplementary Fig. 1b). This confirmed that early post-natal cells were highly enriched for basal-specific genes, yet harbored a signature distinct from that of basal cells isolated from pubertal or adult glands (Supplementary Fig. 1b, c). Many cells in pre-puberty expressed elevated levels of basal-specific genes (e.g., *Sparc*, *Ngfr*, and *ErbB3*), whereas other basal genes were more highly expressed in pubertal and adult

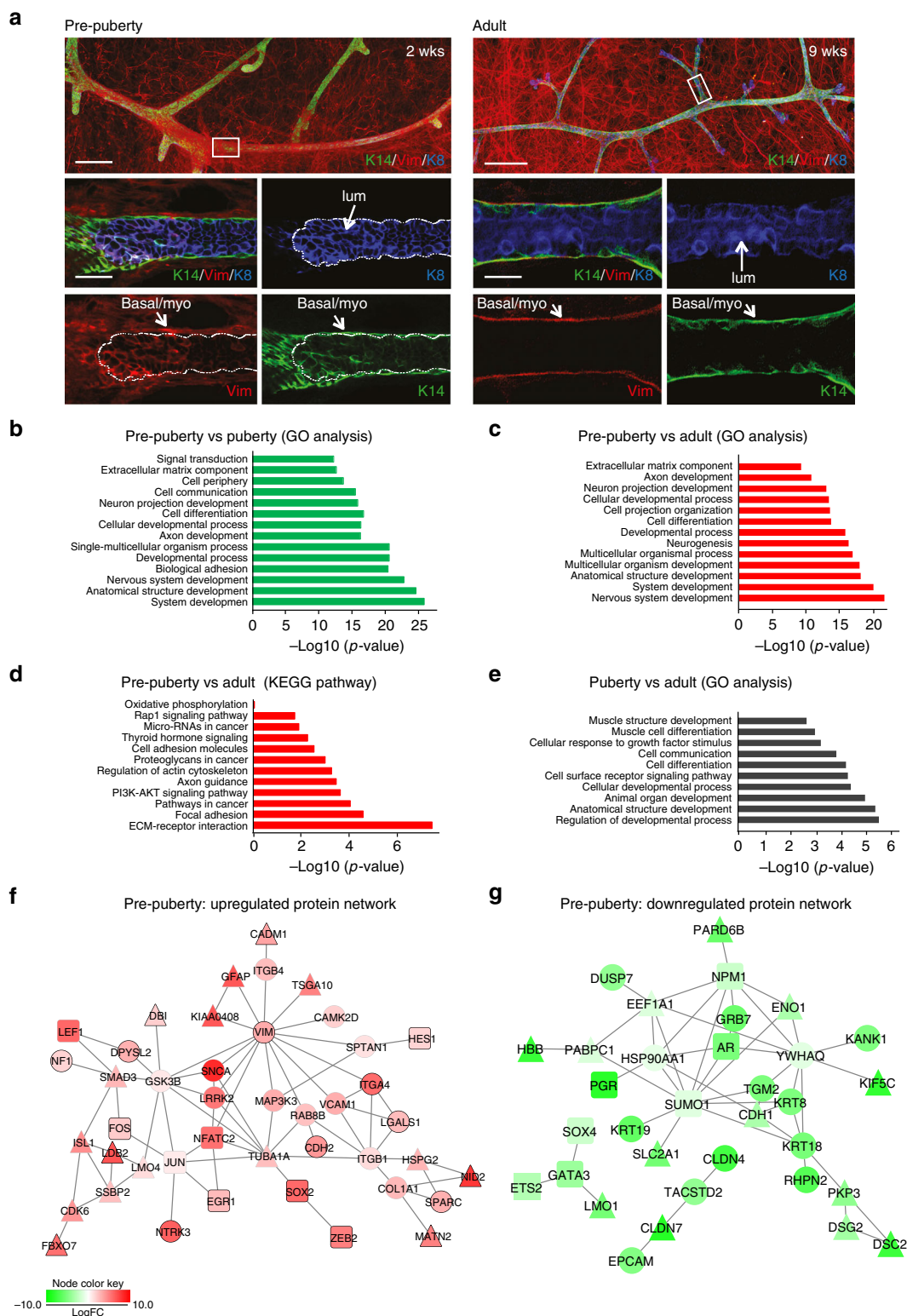


**Fig. 1** Acquisition of molecular heterogeneity in the mammary epithelium during puberty. **a** Schematic diagram of the mammary gland at pre-puberty (2 weeks; red), puberty (5 weeks; green, with TEBS in dark green) and adulthood (10 weeks; black). **b** Multidimensional scaling plot of scRNA-seq profiles. Total epithelial cells were collected at pre-puberty, puberty, and adulthood and profiled using the Fluidigm C1 platform. Distances on the plot correspond to leading log<sub>2</sub>-fold changes between cells. The first plot dimension (x-axis) clearly separates pre-pubertal cells from those in puberty or adult glands. Pre-puberty, 12 mice; puberty, 6 mice; adult, 4 mice per sort; three independent experiments. **c** Gene expression heatmap of the same 460 single cells shown in **b**. Rows correspond to the top 500 DE genes across the three developmental stages. Columns correspond to cells colored by stage as in **a** and **b**, indicated by bars at the top. Major gene clusters I to VI are highlighted in the side column; Supplementary Table 2 lists genes in these clusters. Selected genes are shown for clusters I, II, and IV, comprising genes specifically expressed at 2 weeks. Heatmap values are log expression scaled by row (red = high expression; blue = low expression). **d** Ternary plot showing relative expression of selected genes in the three developmental stages. For each gene, expression was averaged over all cells at each stage. Most genes known to be luminal-specific (purple) are strongly upregulated in cells from 5 and 10-week-old glands while different basal-specific genes (sky blue) are expressed at different developmental stages. Genes correspond to the top 1000 DE genes from Fu et al. 2015<sup>22</sup>. **e** Violin plots showing gene expression of selected genes for each cell by developmental stage. Vertical axis shows expression as log<sub>2</sub>-RPKM

glands, such as the transcription factors *Trp63*, *Id4*, and *Snai2*, the Wnt inhibitory factor *Dkk3*, and myoepithelial contractile/cytoskeletal proteins including *Myh11* and *Acta2* (smooth muscle actin). These data are consistent with the presence of a unique basal-like gene expression program in the early post-natal gland (14 day old)

prior to the establishment of distinct luminal and basal lineage-specific differentiation programs near the onset of puberty.

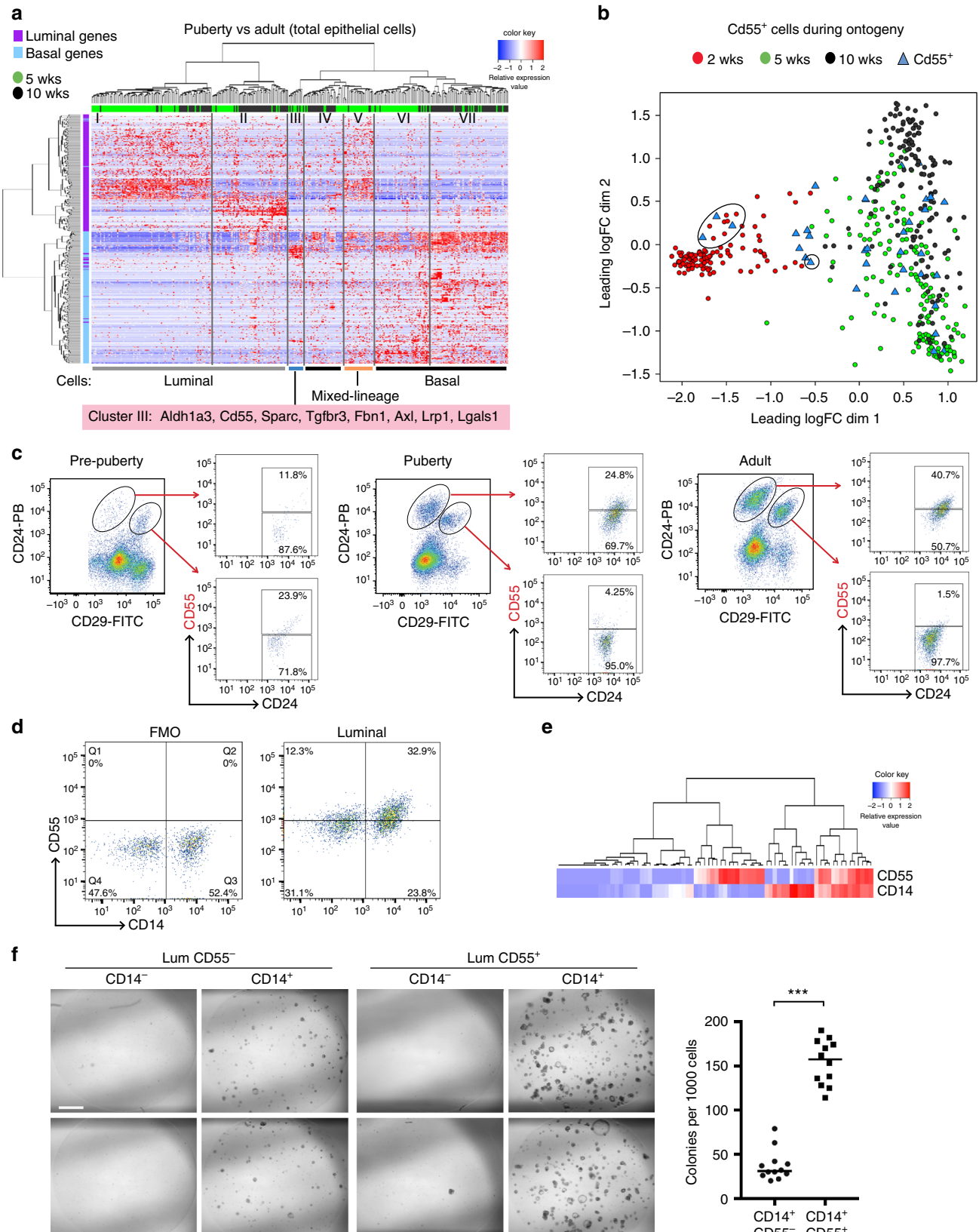
Although the ductal tree is bilayered at 2 weeks, with distinct inner cuboidal cells and outer elongated cells, the largely homogeneous gene expression pattern of epithelial cells at the



**Fig. 2** The majority of pre-pubertal epithelial cells exhibit a unique basal-like signature. **a** Whole-mount confocal 3D images (top panels) and optical sections (middle and bottom panels) of intact mammary ductal trees immunolabeled for Keratin 14 (green), Vimentin (red), and Keratin 8/18 (blue). Mammary glands were collected at 2 (left panels) and 9 weeks of age (right panels). Scale bars, 200  $\mu\text{m}$  (top panels); 20  $\mu\text{m}$  (bottom panels) ( $n = 3$  mice per time point). Basal/myo indicates basal/myoepithelial cells in the outer layer. **b** Gene ontology terms upregulated in 2 vs. 5 week epithelial cells, based on differential expression analysis of the scRNA-seq profiles. **c** Gene ontology terms upregulated in 2 vs. 10 week epithelial cells. **d** KEGG pathways upregulated in 2 vs. 10 week epithelial cells. **e** Gene ontology terms downregulated at 5 vs. 10 weeks. **f** Protein interaction subnetwork of selected genes upregulated (red) in pre-pubertal cells relative to puberty. Node shape represents protein function (circle = signal transduction; square = transcription factor; rounded square = dual function, and triangle = none of the above). **g** Protein interaction subnetwork of selected genes downregulated (green) in pre-pubertal cells relative to puberty

pre-pubertal stage suggests that they lie in an immature state. Indeed, confocal whole-mount 3D image analysis of pre-pubertal glands for expression of the canonical basal (Krt14 and Vim) and luminal markers (Krt8/18) revealed sporadic expression of basal proteins within the luminal layer of the distal ductal branches (Fig. 2a). Vimentin levels were examined as this is one of the most

highly expressed genes in 2-week-old glands. The expression of basal markers in luminal cells is compatible with the scRNA-seq data that shows widespread expression of these genes across individual cells in the pre-pubertal gland. As anticipated, in the adult gland, Krt14 and Vimentin were exclusively expressed in the outer myoepithelial layer (Fig. 2a). Although E18.5<sup>23, 24</sup> and



newborn mammary glands<sup>25</sup> frequently co-express Krt8 and Krt14, the ductal tree is not well-defined at these stages and does not comprise elongated myoepithelial cells<sup>26</sup>, which are evident by 2 weeks of age. Western blot analysis of individual cells<sup>27</sup> from pubertal and adult mammary glands using the Milo system confirmed strong co-expression of Krt8/18 and Krt14 in ~13% of pre-pubertal epithelial cells but not in adult basal cells (Supplementary Fig. 1d).

**Predicted changes in protein interaction networks during ontogeny.** Gene ontology and KEGG analyses of gene expression signatures derived for the different developmental stages showed that pre-pubertal cells were enriched for signaling pathways involved in developmental processes, anatomical structure, axon development, and cell projection organization relative to pubertal and adult epithelial cells (Fig. 2b–d). Interestingly, axon guidance molecules have been implicated in maintaining cell proliferation and adhesion during normal mammary gland development<sup>28, 29</sup>. Conversely, comparative analysis of 5 vs. 10 weeks revealed significant downregulation of developmental and anatomical processes and muscle differentiation pathways in pubertal cells (Fig. 2e).

We next integrated DE genes between the different developmental stages with protein–protein interaction data for the corresponding human orthologues (Methods section) to identify potential biological interactions that regulate the transition of epithelial cells through ontogeny. Our analysis focused on 1796 mouse genes (with 1673 human orthologues) that were DE (FDR < 0.001) between 2 and 5 weeks. We identified direct interactions among the proteins and annotated the resulting network with functional associations drawn from the gene ontology to identify parts of the network involved in signal transduction and transcriptional regulation, using Cytoscape analysis. The networks of transcription factors (TFs) and signal transducers (STs) that are DE between pre-puberty and puberty reveal potential mechanisms involved in driving ontogeny and are consistent with our previous studies<sup>5</sup> (Fig. 2f, g; Supplementary Fig. 2a, b). Notably, the basal marker Vimentin, which is upregulated in pre-pubertal epithelial cells vs. those in puberty, forms a major hub in both the main network and the network of STs and TFs (Fig. 2f, Supplementary Fig. 2a). Interestingly, Vimentin interacts directly with other basal proteins highly expressed in pre-pubertal epithelium such as VCAM1, ITGB4/A4, CAMK2d, MAP3K3, and GSK3B, and the TF NFATC2. Upregulated networks implicated in indirect interactions via the vimentin node included several extracellular components (Collagen type 1, SPARC, Matrilins (MATN2)), the tyrosine kinase receptor NTRK3, as well as the TFs SOX2, ZEB2, FOS, EGR1, and SMAD3 (Fig. 2f). SMAD3 is also predicted to have direct links with a plethora of

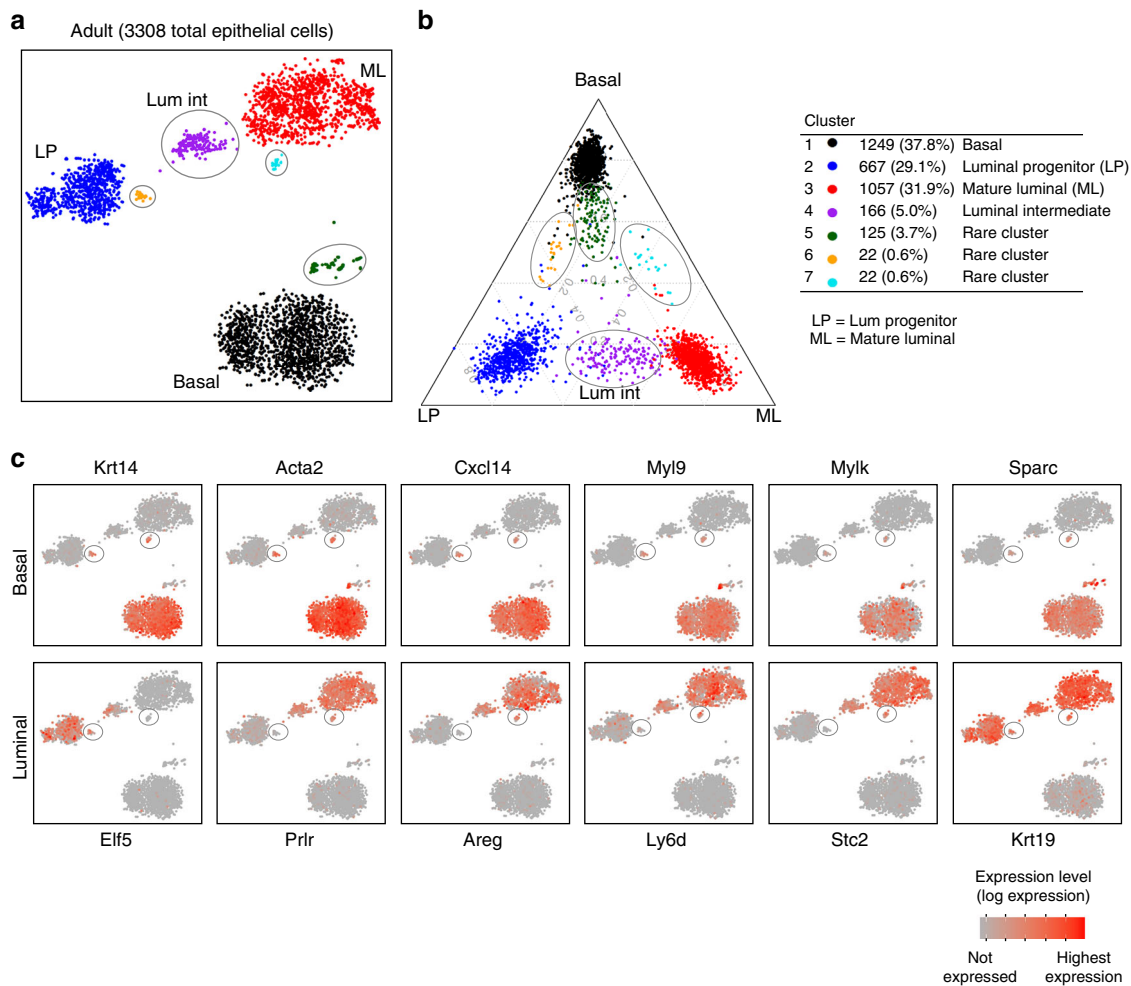
other transcriptional regulators that include ISL1, LEF1, LDB2, and LMO4. Taken together, the network analysis has highlighted the upregulation of multiple interacting proteins important for synthesis of extracellular matrix components and/or interactions with the matrix, thus pointing to a key role for the microenvironment in ductal development within the early post-natal gland.

Relative to puberty, the pre-pubertal gland was marked by downregulation of a protein interaction network interconnected by the shuttling nucleoprotein NPM1 and SUMO1 (Fig. 2g). SUMO1, which catalyzes the post-translational modification of proteins, including the activity of many TFs<sup>30</sup>, has been shown to bind Keratins and alter filament dynamics during cellular development<sup>31</sup>. Intriguingly, in this downregulated network, SUMO1 interacts with the luminal-specific androgen receptor, the luminal structural proteins KRT8, KRT18, and KRT19, and indirectly with E-CAD, the expression of which are very low until the onset of puberty (Fig. 2g). NPM1 is predicted to interact directly with SUMO, GRB7, the glycolytic enzyme ENO1, and the polarity complex protein PARD6B. Interrogation of a larger cohort of DE TFs and STs revealed additional information on interactions between known lineage-specific molecular regulators including the upregulated molecules ID2, MYOG, LEF1, ARNT, NGFR, and RARB, and the downregulated molecules GATA3, PGR, EHF, FGFR2, Klf5, AREG, ELF3, and FOXA1 (Supplementary Fig. 2a).

Comparison of puberty to adulthood showed relatively few genes that were DE at the very strict threshold we applied for network construction, and resulted in a smaller network than that observed for pre-puberty vs. puberty. Notable upregulated networks included NRAS/GRB10/GRB14/INHBA and GSK3B/CHD3/PLP1 nodes, while TF networks including FOS, MAFF, and the steroid hormone receptor PPARα were downregulated (Supplementary Fig. 2b).

**Dissection of heterogeneity in puberty and adulthood.** To interrogate heterogeneity in pubertal and adult glands in more detail, we excluded pre-pubertal cells and re-clustered epithelial cells from these two stages according to their expression of defined basal and luminal lineage genes<sup>22</sup> (Supplementary Table 3). Hierarchical clustering revealed seven major cell clusters (Fig. 3a; Supplementary Table 4). As expected, both luminal cells (clusters I and II) and basal cells (clusters VI and VII) could be clearly identified at these stages, in contrast to the homogeneous basal-like cells evident in pre-pubertal glands (Supplementary Fig. 1b). Clusters I and II were enriched for definitive markers of mature luminal (ML) and luminal progenitor cells, respectively. Conversely, cluster III was enriched for basal genes (*Axl*, *Sparc*, *Tgfb3*, and *Lrp1*), *Cd55* and the progenitor gene *Aldh1a3* (Fig. 3a), and was more prevalent in pubertal mammary glands.

**Fig. 3** Identification of multiple subsets in pubertal and adult mammary glands. **a** Heatmap showing hierarchical clustering of 343 epithelial cells from pubertal (green) and adult (black) mammary glands based on their expression of lineage-specific genes (200 basal and 200 luminal genes; see Supplementary Table 3). Columns correspond to cells colored by stage, indicated by bars at the top. The cells shown are the same 5 and 10-week-old cells as in Fig. 1b, c and Supplementary Fig. 1b, c. Seven major cell clusters are indicated (gene lists shown in Supplementary Table 4). Basal genes = blue; luminal genes = purple; red = high expression; blue = low expression. **b** Multidimensional scaling plot of scRNA-seq profiles. These are the same as in Fig. 1b but cells expressing *Cd55* (RPKM > 40) are highlighted as blue triangles. *Cd55*<sup>+</sup> cells are present at all three stages but pre-pubertal glands harbor only rare cells (circled in the plot). **c** *CD55* expression within epithelial subsets isolated from mammary glands in pre-puberty (2 wks) (left), puberty (5 weeks) (center), and adult (10 weeks) (right). Representative dot plots of *CD29* vs. *CD24* expression in Lin<sup>-</sup> cells (see gating strategy in Supplementary Fig. 3) are shown on the left, while *CD29* vs. *CD55* expression in either the luminal (top) or basal (bottom) compartment is shown on the right (*n* = 3 independent experiments per time-point: 12 mice per experiment for pre-puberty, 6 mice per experiment for puberty, and 4 mice per experiment for adult). **d** The *CD14*/*CD55* flow cytometric plots of freshly sorted Lin<sup>-</sup>*CD29*<sup>lo</sup>*CD24*<sup>+</sup> (luminal) mammary epithelial cells isolated from 10-week-old FVB/N mice (representative of 5 independent experiments, 4 mice per experiment). **e** Hierarchical clustering of 76 purified luminal cells based on *Cd55* and *Cd14* gene expression (Fluidigm C1 platform; see Supplementary Fig. 4c). Red = high expression; blue = low expression. **f** Clonogenic activity of *CD14*<sup>-</sup>*CD55*<sup>+</sup>, *CD14*<sup>+</sup>*CD55*<sup>-</sup>, and *CD14*<sup>+</sup>*CD55*<sup>+</sup> luminal (*CD29*<sup>lo</sup>*CD24*<sup>+</sup>) subsets from 10-week-old FVB/N mice, plated in Matrigel (1000 cells per population). Representative of three independent experiments, 4 mice per experiment. *P*-value (unpaired *t*-test) < 0.0001. Scale bar; 1000 μm



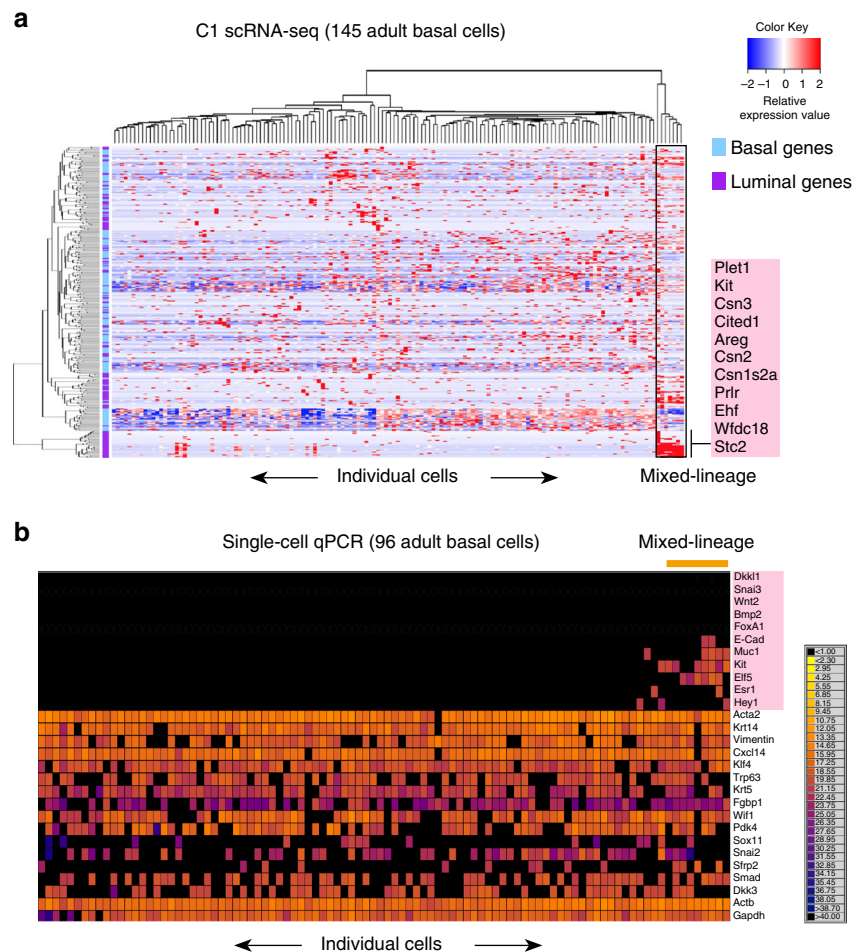
**Fig. 4** Transcriptome analysis of 3308 adult total epithelial cells reveals a luminal intermediate cluster. **a** t-SNE plot of the transcriptomes of total epithelial cells from adult mammary glands ( $n = 4$  mice) generated using the 10X Genomics Chromium platform. The major clusters could be identified as basal (cluster #1, 1249 cells), luminal progenitor (LP; cluster #2, 667 cells), mature luminal (ML; cluster #3, 1057 cells), luminal intermediate (Lum Int; cluster #4, 166 cells), and cluster #5 (green, 125 cells), as indicated. **b** The ternary plot classifies total single epithelial cells into basal, LP, and ML cell clusters based on known lineage-specific signatures for these cell types. Cells are plotted according to the relative proportion of genes they express that are specific to the three cell types. This analysis also showed the presence of the Lum Int cluster (purple cells) between LP and ML cells. **c** The epithelial cell clusters were interrogated for the expression of typical basal and luminal genes to determine their patterning amongst the cell clusters. Color indicates expression level in log<sub>2</sub>-CPM

Cluster V was also mainly found amongst pubertal cells and expressed a subset of both basal and luminal genes, indicative of a mixed-lineage subpopulation. Cluster IV consisted mostly of adult cells and expressed a subset of basal genes. In a recent report, scRNA-seq of 5-week-old pubertal glands indicated that TEB-enriched cells largely cluster into basal and luminal groups despite their heterogeneous expression profile<sup>32</sup>, consistent with findings presented here. Interestingly, close examination of the basal cell clusters VI and VII revealed that cluster VI mainly comprised pubertal cells, whereas cluster VII was primarily found amongst adult basal cells. Moreover, a higher proportion of pubertal cells expressed a gene subset that included *Snai2*, *Moxd1*, *Nrg1*, *Wif1*, and *Bmp7* (Supplementary Table 4; bottom of cluster VI). It is plausible that these cells correspond to the cap cells of TEBs, which characterize the pubertal gland.

**CD55 marks distinct populations through development.** To further explore features of cluster III (Fig. 3a), we selected *Cd55* since this gene encodes a cell surface marker that enables flow cytometry to track protein expression during development.

*CD55* has been shown to be a discriminating marker of one of the earliest branch-points in the hematopoietic system<sup>18</sup>, so we returned to pre-pubertal cells to determine whether *Cd55*-expressing cells are already present at this stage. Intriguingly, a small number (4 out of 117) of cells from 2-week-old glands expressed *Cd55* transcripts (Fig. 3b). Flow cytometry (gating strategy shown in Supplementary Fig. 3) showed *CD55* expression in a minority of cells at the pre-pubertal stage but in larger numbers of pubertal and adult mammary epithelial cells (Fig. 3c), compatible with the scRNA-seq data. In the pre-pubertal gland, *CD55*<sup>+</sup> cells were primarily localized to the basal-like population, but during development, *CD55* expression shifted towards the luminal lineage. Indeed, *CD55*<sup>+</sup> cells were exclusively found in the luminal compartment of the adult gland and comprised approximately half of the Lin<sup>-</sup>CD29<sup>lo</sup>CD24<sup>+</sup> population (Fig. 3d). *CD55*<sup>+</sup> luminal cells from pubertal or adult tissue harbored 3.0- or 3.3-fold higher colony-forming capacity than *CD55*<sup>-</sup> luminal cells, respectively, suggesting that *CD55* enriches for progenitor cells.

Comparison of *CD55* levels relative to that of another adult luminal progenitor marker CD14<sup>4, 33</sup> revealed that adult mammary glands comprised ~33% double-positive cells in the



**Fig. 5** Purified basal epithelial cells in the adult gland comprise a lineage-primed subset. **a** Heatmap showing expression of the top 200 DE basal (blue) and top 200 DE luminal (purple) lineage genes in 145 sorted basal epithelial cells ( $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ ) from adult mammary glands that were captured on the Fluidigm C1 platform (red = high expression, blue = low expression) ( $n = 4$  mice). Mixed-lineage intermediate cells are marked in the box. **b** Heatmap showing qRT-PCR expression for 96 sorted adult basal cells ( $n = 6$  mice; 4 independent experiments). Taqman assays for 13 basal and 11 luminal lineage genes were applied using the Fluidigm Biomark multiplexed qPCR platform. The color key shows Ct values from black (Ct > 40, no expression) to yellow (highest expression). Mixed-lineage cells expressing both basal and luminal genes are marked by the orange bar

luminal fraction (Fig. 3d), compatible with the overlap observed at the messenger RNA level (Fig. 3e). Colony-forming assays in Matrigel revealed that  $\text{CD}14^+ \text{CD}55^+$  cells were significantly enriched for clonogenic activity (Fig. 3f), thus reflecting a highly purified population of committed progenitor cells. The large acinar colonies are consistent with the presence of luminal progenitor cells. Collectively, these observations suggest that  $\text{CD}55$  expression is first acquired by rare basal-like cells prior to the onset of puberty and that it subsequently demarcates a potent progenitor cell that becomes restricted to a luminal cell fate.

**Large-scale scRNA-seq identifies luminal intermediates in the adult.** To investigate molecular heterogeneity more fully within the mammary epithelial compartment, we used the 10X Genomics Chromium platform to profile gene expression in thousands of individual cells (at lower sequencing depth than provided by the Fluidigm C1 technology) (Supplementary Table 1). 10X profiling of 3308 individual epithelial cells ( $\text{Lin}^- \text{CD}24^+$ ) from adult mice revealed three major clusters as well as several satellite clusters (clusters 4–7) (Fig. 4a) based on t-distributed stochastic neighbor embedding (t-SNE)<sup>34</sup>. To uncover the lineage identity of each cluster, we overlaid the t-SNE clusters onto a ternary plot with each cell positioned according to its expression of known markers for mammary stem cell (MaSC)/basal (basal), luminal progenitor

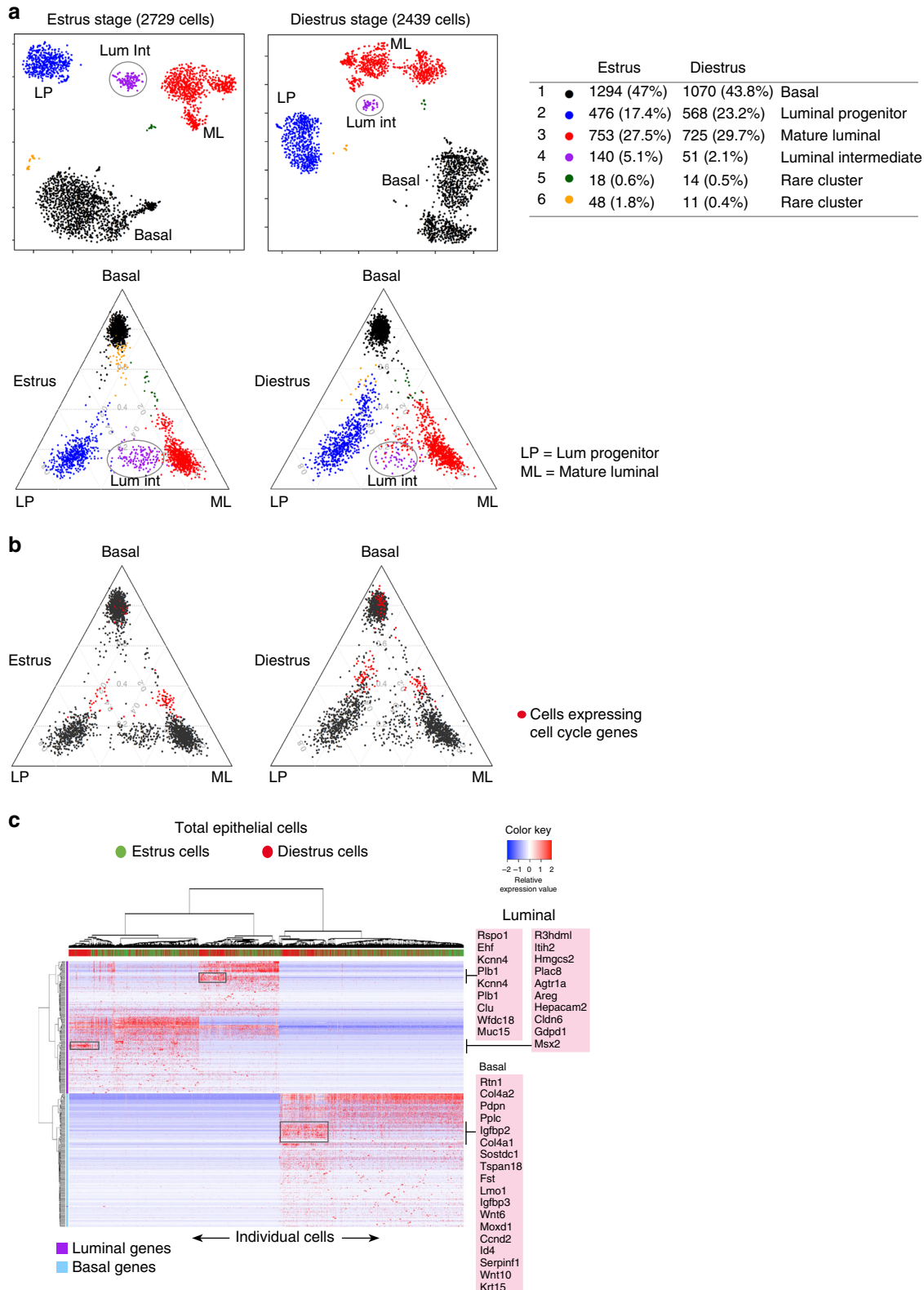
(LP), and ML cell types<sup>8</sup> (Fig. 4b). The ternary plot shows the proportion of basal, LP, and ML-specific genes expressed by each cell and recapitulates the t-SNE clusters to a striking degree. Expression analysis of the cell clusters for basal genes (*Krt14*, *Acta2*, *Cxcl14*, *Sparc*, *Mylk*, and *Myl9*), and luminal genes (*Krt19*, *Areg*, the prolactin receptor (*Prlr*), the estrogen-regulated secreted glycoprotein *Stc2*, the antigen *Ly6d*, and the progenitor marker *Elf5*) confirmed the identity of the major clusters (Fig. 4c). These data indicated three luminal clusters and a largely homogeneous basal population associated with a small cluster (#5) that is closely aligned with the basal population at the molecular level (Fig. 4a, b). The genes most highly expressed in this cluster relative to basal cells are shown in Supplementary Fig. 4a. Other very rare clusters (#6 and 7) apparent in Fig. 4a may represent different regulatory states, although this remains to be determined. These clusters were noted to express basal as well as selective luminal genes including the epidermal growth factor (EGF) family member amphiregulin (*Areg*), *Prlr* and *Krt19* (discussed further below). Analysis of sorted luminal cells on the 10X platform confirmed the presence of three luminal populations and did not reveal mixed-lineage clusters (data not shown).

In the context of the luminal lineage, a novel cluster (#4) was identified to lie between the two major luminal populations in both the t-SNE and ternary plots (Fig. 4a, b). Interestingly, this



was found to express the majority of genes at an intermediate level between that in progenitor (LP) and mature (ML) cells (Supplementary Fig. 4b). We therefore denoted this cluster Luminal Intermediate or Lum Int. Of 3711 genes that were DE between the LP, Lum Int, and ML cell clusters (FDR < 0.01), more than 88% showed steadily increasing or decreasing expression from LP to Lum Int to ML, further supporting the

interpretation of the Lum Int population as an intermediate state between LP and ML (Supplementary Fig. 4b). In addition, a small number of genes (e.g., *Jund*, *Irx5*, *Sox4*, *Ncor1*, and *Igfbp2/6*) showed at least 20% higher average expression in the Lum Int population relative to either LP or ML. Collectively, these findings suggest that Lum Int cells represent a transit population between the progenitor and differentiated states, rather than a distinct type



of luminal cell. As anticipated, intermediate luminal cells were detected amongst purified luminal cells (Lin<sup>-</sup>CD29<sup>lo</sup>CD24<sup>+</sup>) from adult mammary glands upon interrogation with the gene expression profiles of LP and ML cells<sup>5,8</sup> (Supplementary Fig. 4c). Finally, the Lum Int cluster was readily demarcated by scRNA-seq of total mammary epithelial cells (5387) isolated from pubertal glands (Supplementary Fig. 5a, b). Therefore, this intermediate population appears to arise in puberty, when definitive commitment to the luminal lineage is first observed.

#### Deep scRNA-seq of basal cells uncovers mixed-lineage cells.

High-resolution transcriptome mapping (Fluidigm C1) of 145 purified basal cells (Lin<sup>-</sup>CD29<sup>hi</sup>CD24<sup>+</sup>) from the adult mammary gland confirmed a largely dominant basal population, as also observed by 10X Chromium scRNA-seq analysis of 4754 sorted basal cells (Supplementary Fig. 5c). Interestingly, clustering of the C1 platform-based single-cell transcriptomes of basal cells according to basal and luminal lineage genes indicated the presence of mixed-lineage cells (Fig. 5a). These cells comprised ~5% of the basal compartment, with 7 out of 145 cells expressing luminal genes such as *Kit*, *Areg*, *Prlr*, the TFs *Ehf* and *Cited1*, and the milk protein genes *Csn1s2a*, *Csn2*, and *Csn3*, in addition to core basal genes. Such cells, albeit rare, were also evident amongst total epithelial cells (Fig. 3a; Supplementary Fig. 1b). Additional evidence for the existence of a lineage-primed population derived from the analysis of 96 sorted basal cells using the Fluidigm Biomark Real-Time PCR system<sup>35</sup>. These basal cells were found to express the luminal genes *Elf5*, *Muc1*, *Kit*, *Esrl*, and/or *E-cad* (Fig. 5b), implying the presence of a transient subpopulation that may be primed for luminal lineage commitment. Conversely, interrogation of purified adult LP and mature cells for basal gene expression did not reveal basal lineage-priming in these cells by multiplex quantitative reverse transcription PCR (Supplementary Fig. 6a, b). Together, these data suggest that the basal state may precede commitment to a luminal cell fate in the post-natal mammary gland.

**Mixed-lineage cells exist at other developmental stages.** A larger mixed-lineage population was evident in pubertal compared to adult mammary glands. This was first apparent in the developmental analysis across three stages, which revealed cells with dual-lineage features in puberty and adulthood but not in the pre-pubertal phase (Fig. 3a; Supplementary Fig. 1b, c). The variable expression patterns evident in this mixed-lineage population (Cluster V in Fig. 3a) indicate that individual cells express distinct combinations of basal and luminal genes, compatible with the properties of lineage-primed cells. Mixed-lineage clusters (#6 and 7) were also observed amongst pubertal mammary epithelial cells captured on the 10X Chromium platform (Supplementary Fig. 5a). These cells expressed core basal genes including *Krt14*, *Acta2*, *Myl9*, *Sparc*, *Mylk*, and/or *Cxcl14* and luminal genes such as *Areg*, *Prlr*, *Krt19*, *Cited1*, and/or *Stc2*.

To investigate whether mammary glands during pregnancy comprised an intermediate subset, total epithelial cells isolated at

mid-pregnancy (12.5 days) were subjected to high-resolution scRNA-seq analysis. While the multidimensional scaling (MDS) plot (Supplementary Fig. 6c) showed dispersed cells, hierarchical clustering of cells according to lineage-specific genes revealed the presence of an expanded mixed-lineage intermediate that expressed canonical basal genes together with numerous luminal genes (*Elf5*, *Krt8*, *Cyp24a1*, *Wap*, *Csn1s2b*, *Csn2*, and *Lalba*) (Supplementary Fig. 6d). These findings suggest that basal cells are primed towards the alveolar lineage and milk protein production in pregnancy to enable expansion and differentiation in preparation for lactation. A small cluster highly enriched for ML cell-specific genes such as *Prlr*, *Cited1*, *Esrrb*, and *Cxcl15* was also noted.

#### Global changes in gene expression during the estrus cycle.

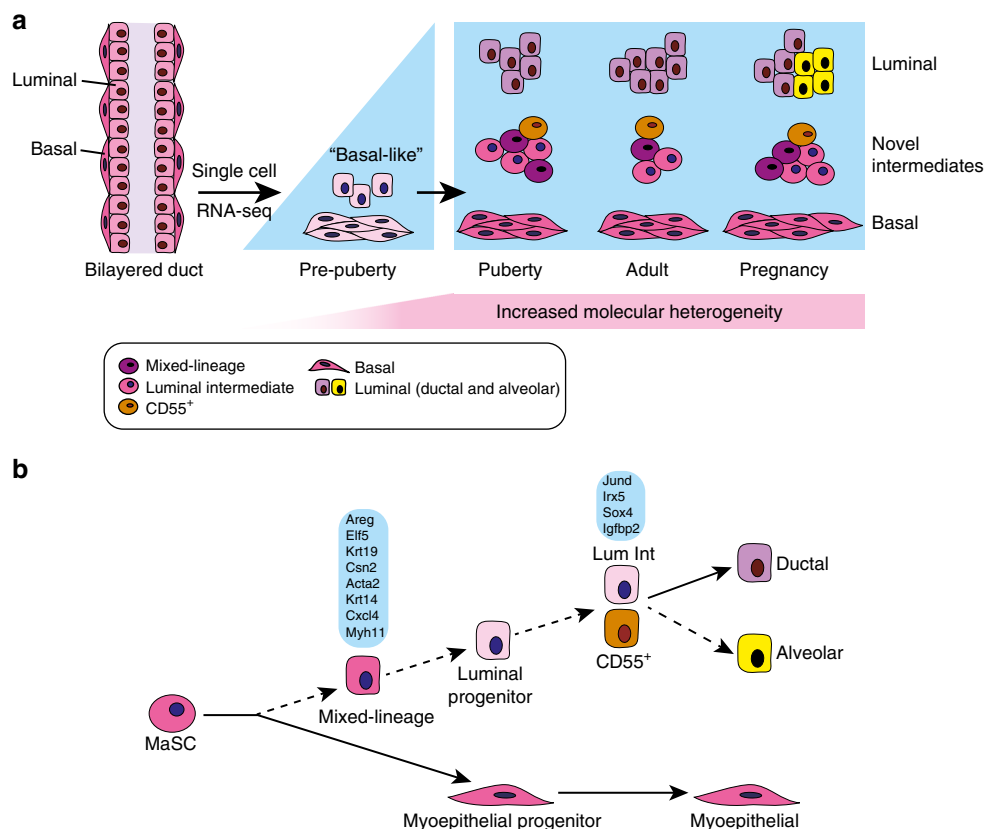
The mouse mammary gland undergoes hormonal cycling every 3–5 days<sup>36</sup>. To evaluate gene expression changes that accompany cycling at the single-cell level, we collected mammary glands at different phases of the estrus cycle. Estrus and diestrus represent the luteal and follicular phases of the hormonal cycle, respectively, whereby the diestrus phase is characterized by maximal progesterone levels, and estrus by a surge of circulating estrogen<sup>37</sup>. Individual adult mice were staged by vaginal cytology at the time of mammary gland collection and glands from two mice subsequently pooled for estrus and diestrus. scRNA-seq of 2729 total epithelial cells in estrus and 2439 cells in diestrus using the 10X Chromium platform indicated the presence of the three major populations (basal, LP, and ML) (Fig. 6a). However, the Lum Int population (cluster 4) was substantially diminished in diestrus whereas the LP population was unaltered, suggesting a rapid transit to a more differentiated state. Moreover, the ML population appeared to segregate into two subsets in diestrus, and one of the clusters tightly associated with ML signature genes such as *PgR*<sup>38,39</sup>. Given that progesterone is a potent mitogen and is elevated in diestrus, we examined the expression of cell cycle genes including *Ki67*. A notable increase in the number of cycling/proliferative cells was observed in the basal population in diestrus (Fig. 6b; Supplementary Fig. 7a, b), compatible with previous findings that the MaSC-enriched/basal population is expanded during diestrus or upon treatment with progesterone<sup>40</sup>.

Hierarchical clustering of the scRNA-seq profiles of combined estrus and diestrus cells with respect to the top DE basal and luminal genes further highlighted that changes across the stages were not widespread but more localized (Fig. 6c). Three specific cell clusters of upregulated genes could be identified in diestrus. The upregulated basal genes included *cyclin D2*, the ligands *Wnt6*, *Wnt10* and basal transcriptional regulator *Id4*, while luminal genes included *Areg* and *Rspodin1* (*Rspo1*), all of which are compatible with the increase in proliferation that occurs in this phase (Fig. 6c).

#### Discussion

Unmasking cellular heterogeneity within a given organ is critical for understanding the biology of normal and diseased tissues.

**Fig. 6** Transcriptome analysis of estrus and diestrus adult epithelial cells reveals specific changes in molecular heterogeneity. **a** t-SNE plots of the transcriptomes of total epithelial cells (Lin<sup>-</sup>CD24<sup>+</sup>; *n* = 2 mice for each phase) from adult mammary glands staged at estrus (2729 cells) and diestrus (2439 cells) were generated using the 10X Genomics Chromium platform (top panels). The major epithelial clusters (1–3; see table on right) representing basal, mature luminal (ML) and luminal progenitor (LP) cell populations could be identified by ternary plots based on known lineage-specific signatures for these cell types (bottom panels). Cells are plotted according to the relative proportion of genes they express that are specific to the three cell types. Cluster 4 depicts the Lum Int (luminal intermediate) population. **b** Ternary plots as for **a** but with cells marked by expression of cell cycle genes. Cells marked red belong to the cell clusters with cell cycle upregulation identified in Supplementary Fig. 7. **c** Heatmap showing hierarchical clustering of all estrus (green) and diestrus (red) epithelial cells together based on their expression of lineage-specific signature genes (200 DE basal and 200 DE luminal genes). The luminal and basal gene clusters specifically upregulated in diestrus stage are marked by boxes in the side columns (basal genes = blue; luminal genes = purple; red = high expression; blue = low expression)



**Fig. 7** Schematic model depicting lineage relationships and an increase in molecular and cellular heterogeneity during mammary gland ontogeny. **a** A profound shift in gene expression occurs between pre-puberty and the onset of puberty. New intermediate clusters identified include a luminal intermediate (Lum Int) population and a mixed-lineage or ‘lineage-primed’ subset. Lum Int expresses luminal genes at levels intermediate between LP and ML, but a small number of genes (four indicated) showed at least 20% higher average expression. CD55<sup>+</sup> cells with intermediate properties were readily detectable in puberty and become restricted to the luminal lineage during development. **b** Proposed position of the newly identified clusters along the epithelial hierarchy. The mixed-lineage subset within the basal compartment may represent a transient population that precedes commitment to the luminal lineage. Molecular evidence places the Lum Int cluster between the luminal progenitor and mature ductal/alveolar cells

Here we provide high-resolution single-cell transcriptomes of mammary epithelial cells at different stages of ontogeny as well as large-scale analyses of the adult mammary epithelium, which have enabled the construction of a temporal lineage specification map. The developmental phase in the post-natal animal was found to profoundly affect cellular and genetic heterogeneity (Fig. 7). Strikingly, epithelial cells in the early post-natal gland exist in a unique molecular state that is reminiscent of basal cells. These cells exhibit a relatively homogeneous transcriptional landscape, which is distinct from that of adult basal cells and show little luminal gene expression. Curiously, luminal cells that line the bilayered ducts in 2-week-old glands display the molecular characteristics of basal-like cells, suggestive of an immature state. Near the onset of puberty, a major transcriptional reprogramming event occurs, with the emergence of cells that harbor distinct lineage-specific gene signatures. Thus, the overall patterning suggests a branched trajectory of development with progression from a basal-like population in the early post-natal gland towards a more diverse array of lineage-specific epithelial cells from puberty into adulthood.

The basal compartment of the mammary gland appears to be dynamically regulated through the different stages of development in the post-natal animal. In the pre-pubertal gland, the basal-like gene expression signature differs substantially from that of basal cells in pubertal and adult mammary glands, although a few genes were common across the three stages. Moreover, the composition of the basal compartment was similar but not identical between

pubertal and adult glands. High-resolution single-cell transcriptome analysis of purified basal cells from adult mice revealed an apparent continuum of cells that exhibited strong basal character. Nevertheless, hierarchical clustering uncovered a rare basal subset that displayed ‘mixed-lineage’ features, namely the expression of canonical basal genes together with numerous luminal genes such as *Elf5*, *Csn2*, *Areg*, *Cd14*, and *Prlr*.

The mixed-lineage or ‘lineage-primed’ cluster in the basal compartment was observed at different developmental stages including puberty, adulthood, and pregnancy. These cells, however, were absent in 2-week-old mammary glands and thus appear to coincide with the increased epithelial diversity that emerges at the onset of puberty. Although scarce amongst basal cells in the steady-state adult mammary gland, mixed-lineage cells were more prevalent in puberty and pregnancy, commensurate with the massive epithelial expansion that accompanies these stages. Their presence in mid-pregnancy is consistent with microarray analysis of the bulk basal cell population at day 12.5<sup>41</sup>. The detection of mixed-lineage or lineage-primed intermediates across different developmental stages suggests that transcriptional priming may not be stochastic. It is presumed that these cells represent a transient population that is poised for commitment to the luminal lineage, reminiscent of ‘lineage-primed’ stem and progenitor cells initially reported in the haematopoietic system<sup>18, 42–44</sup>. Indeed, only luminal and not myoepithelial/basal cells express milk protein genes at the protein level, consistent with the existence of transcriptionally-primed cells. Given that highly purified basal

cells appear to comprise lineage-primed cells, it is presumed that these cells are descendants of basal epithelial cells, although the possibility remains that luminal cells could transition to a basal-like state under certain conditions. The demonstration that a single basal cell has the capacity to reconstitute an entire ductal tree upon transplantation into the fat pad<sup>45, 46</sup> is consistent with the notion of 'primed' multipotent cells. Notably, single-cell RNA analyses have identified lineage-primed cells in the haematopoietic system<sup>17, 18</sup>, intestine<sup>20</sup>, and pancreas<sup>19</sup>, as well as heterogeneity in the haematopoietic stem cell pool<sup>47</sup>.

In the luminal compartment of the adult mammary gland, three distinct luminal clusters were visualized through interrogation of single-cell transcriptomes. These corresponded to a novel luminal intermediate cluster (Lum Int) as well as the expected LP and ML populations, which largely reflect differentiation status. Moreover, in contrast to the ML subset, LP cells are largely hormone receptor-negative. The novel cluster forms a prominent population that exhibits properties intermediate between that of the LP and ML subsets, and likely represents a transit population. Interestingly, this cluster was reduced in size in diestrus, perhaps indicating a faster transit time between the progenitor and mature states in response to higher serum progesterone levels. Indeed, the ML population, which expresses high levels of *Esr1* and *Pgr*<sup>7, 8, 48</sup>, was more pronounced in diestrus than estrus.

CD55 was identified as a novel marker of early progenitor cells. Rare CD55<sup>+</sup> cells could be discerned in the pre-pubertal mammary gland, where it may mark the earliest prospective luminal precursor cells. Pertinently, CD55 marks the functional separation of the megakaryocytic-erythroid and lympho-myeloid branches in the haematopoietic system<sup>18</sup>. In pubertal glands, a tight cluster of cells that co-express Cd55 with *Aldh1a3*, a marker of mammary progenitor cells<sup>4, 7, 49</sup> and the receptor tyrosine kinase *Axl* that has been recently implicated in the epithelial mesenchymal transition<sup>50</sup>, could be distinguished. In adult tissue, CD55 was found to mark progenitor cells exclusively restricted to the luminal lineage. Moreover, the CD55<sup>+</sup> subset falls within the CD14<sup>+</sup> (and CD61<sup>+</sup>) luminal fractions, where it substantially enriches for highly clonogenic progenitor cells. Interestingly, *Axin-2* cell-fate mapping studies have indicated that the output of cells induced at different stages can differ markedly<sup>51</sup>. Similar observations may apply to CD55<sup>+</sup> cells during mammary gland development. Although CD55 has been implicated in breast tumorigenesis<sup>52</sup>, its precise role remains to be determined.

Notably, MaSCs were not readily definable through scRNA-seq analyses. Conversely, functional studies based on cell surface marker expression and flow cytometry have led to further fractionation of stem cells from the basal compartment. For example, EpCAM and Procr enrich for MaSCs in combination with other markers<sup>7, 53</sup>. More recently, *Lgr5* and *Tspan8* were utilized to resolve three basal subsets with stem cell activity, one of which was highly enriched for dormant MaSCs<sup>26</sup>. Whilst molecular profiling represents a powerful tool for unraveling heterogeneity on a global scale, an understanding of the molecular and functional diversity amongst stem cells in the basal compartment may only be possible through fractionation based on cell surface markers. Interestingly, although hematopoietic stem cells (HSCs) are well-defined, scRNA-seq analysis of human hematopoietic cells indicated a continuum of molecular states downstream of HSCs, implying that lineage restriction can occur in multiple directions<sup>54</sup>.

In summary, the deconstruction of heterogeneity at different developmental time-points by ordering individual cells on the basis of transcriptome similarity has led to unexpected findings on the transcriptional landscape of mammary epithelial cells. The gene expression programs of the definitive mammary lineages

appear to be defined near the onset of puberty, with the majority of epithelial cells existing in a unique molecular state prior to this period. In the adult mammary gland, at least two novel intermediate subsets were delineated. The lineage-primed subset evident amongst purified adult basal cells by hierarchical clustering suggests that a basal expression state precedes commitment to the luminal lineage. Collectively, these data on the molecular identity of individual mammary epithelial cells should provide a useful resource for future studies on understanding the molecular networks that drive specification and differentiation.

## Methods

**Mouse strains.** The FVB/NJ mouse strain was used for preparation of mammary epithelial cell suspensions for all scRNA-seq and RT-PCR work. All animal experiments conformed to regulatory standards and were approved by the Walter and Eliza Hall Institute Animal Ethics Committee.

**Mammary cell preparation and cell sorting.** Mammary glands from 2-week-old (precisely 14 days old), 5-week-old (between 4.7–5 weeks and designated 5 weeks herein) and 10-week-old female mice were collected and single-cell suspensions were prepared and stained<sup>45</sup>. The following antibodies were used: FITC anti-mouse CD29 (rat, clone HM $\beta$ 1-1, BioLegend Cat#102206, 1/200 dilution), Pacific Blue anti-mouse CD24 (Armenian Hamster, clone M1/69, BioLegend Cat#101820, 1/200), APC anti-mouse CD31 (rat, clone 390, BioLegend Cat#102410, 1/40 dilution), APC anti-mouse CD45 (rat, clone 30-F-11, BioLegend Cat#103112, 1/100 dilution), APC anti-mouse TER-119/erythroid cell (rat, clone TER-119, BioLegend Cat#116212, 1/100 dilution), PE anti-mouse CD55 (Armenian hamster, clone RIKO-3, BioLegend #131803, 1/50 dilution), and biotin anti-mouse CD14 (rat, clone Sa2-8, eBioscience #13-0141-85, 1/100 dilution). To exclude dead cells, cells were re-suspended in 0.5  $\mu$ g/ml propidium iodide prior to analysis. FACS analysis and cell sorting were performed on a FACS Aria (Becton Dickinson). The Lin<sup>-</sup> population was defined as Ter119<sup>-</sup>CD31<sup>-</sup>CD45<sup>-</sup>. FACS data were analyzed using FlowJo software (v 10.1r7, Tree Star).

**Colony-forming assays.** 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  $\mu$ g/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 was evaluated by ImageJ.

**Single cell capture and library preparation for sequencing.** Freshly sorted epithelial cells were submitted to the Fluidigm C1 machine for single-cell capture and complementary DNA (cDNA) preparation according to the manufacturer's specifications. The 10–15  $\mu$ m C1 integrated fluidics circuits (IFCs) were used to capture cells, convert polyA<sup>+</sup> RNA into cDNA, and perform cDNA amplification. Cells were visualized under the microscope to ensure integrity of the captured single cells prior to cDNA preparation. For two experiments, single-cell capture was confirmed by examining 3D images of 100's of individual cells through acquiring Z-stacks of mammary epithelial cells captured on C1 IFCs. The 3D analysis corroborated observations made by regular microscopy. The libraries were prepared using the Nextera XT kit and sequencing was carried out on either an Illumina HiSeq 2000 to achieve 100 bp paired-end reads on a NextSeq 500 to achieve paired-end 75 bp reads.

For high through-put single-cell studies, a 10X Genomics Chromium machine was used for 3000–5000 single-cell capture and cDNA preparation according to the Single Cell 3' Protocol recommended by the manufacturer. The cells are first washed and prepared in an ideal concentration. The machine partitions thousands of cells into nanoliter-scale Gel Bead-In-EMulsions (GEMs), where all generated cDNA share a common 10x Genomics Barcode but uses a pool of ~750,000 barcodes to separately index each cell's transcriptome. The silane magnetic beads and Solid Phase Reversible Immobilization beads were used to clean up the GEM reaction mixture and the barcoded cDNA was then amplified in a PCR step. Optimal cDNA amplicon size was achieved using Covaris machine prior to library construction. The P7 and R2 primers were added during the GEM incubation and the P5, and R1 during library construction via end repair, A-tailing, adapter ligation and PCR. The final libraries contain the P5 and P7 primers used in Illumina bridge amplification. Sequencing was carried out on an Illumina NextSeq 500 to achieve 75 bp reads.

**Multiplex single-cell RT-PCR.** Single-cell RT-PCR were performed using 96:96 IFC microfluidic chips (Fluidigm). Single cells were flow sorted directly into the individual wells of 96-well PCR plates containing 5  $\mu$ l/well of CellsDirect PCR mix and RNase Inhibitor mix. Following single-cell sorting, each well was supplemented with 1  $\mu$ l SuperScript III RT/Platinum Taq (Invitrogen), 1.5  $\mu$ l Tris-EDTA (TE) buffer and 2.5  $\mu$ l of 96 pooled TaqMan assays (Applied Biosystems). Single-

cell lysates were directly reverse transcribed into cDNA (50 °C for 15 min, 95 °C for 2 min) and pre-amplified for 14 PCR cycles (60 °C for 4 min, 95 °C for 15 s). A mixture containing 2.25 µl amplified cDNA (diluted 1:5), 2.5 µl TaqMan quantitative PCR (qPCR) mix (Applied Biosystems) and 0.25 µl Fluidigm sample loading agent was prepared and loaded onto the IFC chip. A 2.5 µl aliquot of TaqMan assay was mixed with 2.5 µl Fluidigm assay loading agent and loaded onto the IFC chip. The 96:96 chip was then submitted to HX IFC Controller (Fluidigm) before setting up qPCR run on Biomark real-time PCR machine (Fluidigm) following the manufacturer's instructions; the data were analyzed using the Fluidigm program. Cells not expressing ACTB ( $\beta$ -actin) and GAPDH (Glyceraldehyde 3-phosphate dehydrogenase) were removed from the analysis.

**scRNA-seq bioinformatic analysis.** Read pairs from the Fluidigm C1 scRNA-sequencing were mapped to the mm10 mouse genome using the Subread aligner<sup>55</sup> and assigned to genes using featureCounts<sup>56</sup> and Rsubread's built-in refSeq annotation. Mitochondrial genes, ribosomal genes, and Entrez gene IDs without current annotation were removed. To ensure good coverage per cell, cells were dropped from the analysis if fewer than 100,000 read pairs were assigned to the remaining genes or fewer than 1500 genes were detected. For epithelial cells from pregnant mice and basal cells from adult mice, these thresholds were relaxed to 50,000 read pairs and 1000 genes. Finally, genes that failed to achieve 1 count-per-million (CPM) in at least 3 cells were filtered from the analyses. Illumina output from 10X Genomic Chromium sequencing reads was processed using Cell Ranger v2.0. Genewise read counts were exported to Matrix Market format files and read into R.

Statistical analyses of the C1 data were conducted using the edgeR<sup>57</sup>, limma<sup>58</sup>, and Rtsne software packages for R. MDS plots were generated with edgeR's plotMDS function. Distances on the MDS plots represent leading log fold change, which is the root mean square average of the largest 500 expression log<sub>2</sub>-fold changes between each pair of cells. Ternary plots were generated using the vcd package with quantile normalized log<sub>2</sub>-CPM values. Log<sub>2</sub>-RPKM values calculated using edgeR's rpkm function with a prior count of 5. Heatmaps were generated using the heatmap.2 function of the gplots package. Log<sub>2</sub>-RPKM values were standardized to have mean 0 and standard deviation 1 for each gene before producing the heatmaps, after which genes and cells were clustered by Euclidean distance. Differential expression analyses used likelihood ratio tests with negative binomial dispersions estimated by the estimateDisp function<sup>57, 59</sup>. Gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed using the goana and kegg functions. t-SNE plots were generated using the Rtsne package.

**Bulk RNA-seq bioinformatic analysis.** Signature genes for the basal and luminal cell lineages were obtained from previously published RNA-seq data<sup>22</sup>. The differential expression analysis was performed using the glmTreat function in edgeR with a fold change threshold of 3. Only protein-coding genes were considered in the analysis. Signature genes for basal, LP, and ML cell types were obtained from other previously published RNA-seq data<sup>8</sup>. Genes were considered cell type specific if they were upregulated in one cell type vs. both other types. The comparisons were performed with a limma TREAT fold change threshold<sup>60</sup> of 2 for basal vs LP and basal vs ML and 1.3 for LP vs ML. A false discovery rate (FDR) cut-off of 0.05 was applied for each comparison.

**Protein network analysis.** The mouse interactome is sparse compared with that available for human<sup>61</sup>, so mouse genes were mapped to human orthologs using the mouse-human ortholog mapping available through the Mouse Genome Informatics resource<sup>62</sup>. Protein interactions were retrieved, from a large collection of manually curated experimentally identified protein interactions<sup>63</sup>, for the orthologs of genes found to be DE between pre-puberty, puberty and adulthood from the scRNA-seq with FDR < 0.001. Only direct interactions between genes were used. Gene Ontology terms "signal transduction" (GO:0007165) and "TF activity" (GO:0003700) were used to identify signal transduction and transcript factor genes, respectively. Networks were merged with differential expression data and functional categories, and imported to Cytoscape for analysis and visualization. Networks were generated for the 2 vs. 5 week comparison, and for the 5 vs. 10 week comparison. Small subnetworks corresponding to sets of genes of interest were sampled from these background networks.

**Single-cell western analysis.** Freshly sorted single-cell suspensions (30,000 cells per ml) were loaded onto the scWest chip for 20 min. Optical visualization was performed to confirm and mark the single-cell capture sites followed by gentle washing in resuspension buffer according to the manufacturer's conditions. The scWest chip was then submitted to the Milo machine (Protein Simple) for lysis, electrophoresis and protein immobilization. Protein targets were probed on-chip with primary antibodies against Keratin 14 (rabbit polyclonal, Thermofisher, cat#: LBVRB-9020-P0, 1:80 dilution) and Keratin 8/18 (rat, clone: TROMA-1, DSHB, 1:80 dilution) overnight at 4 °C and then with secondary antibodies, either anti-rat Alexa-647 (Invitrogen; 1:100 dilution) or anti-rabbit Alexa-555 (Invitrogen; 1:100 dilution) for one hour at room temperature. The probed chip was washed, air-dried and analyzed using a fluorescence microarray scanner (Innoscan 710).

**Confocal analysis on whole-mounts and histological 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. Tissues were subsequently incubated in 80% glycerol before dissection for three-dimensional imaging<sup>64</sup>. The following primary antibodies were used: Keratin 14 (rabbit polyclonal, Thermofisher, cat#: LBVRB-9020-P0, 1:500 dilution), Keratin 8/K18 (rat, clone: TROMA-1, DSHB, 1:500 dilution), and Vimentin (mouse, clone LN-6, Sigma cat# V2258, 1:200 dilution). All secondary antibodies were Alexa Fluor-conjugated: anti-rabbit Alexa Fluor 488 (Invitrogen; 1:500 dilution), anti-mouse Alexa Fluor 555 (Invitrogen; 1:500 dilution), anti-rat Alexa Fluor 647 (Invitrogen; 1:500 dilution).

**Data availability.** Sequencing data have been deposited in the GEO database under accession codes: GSE98131 (C1 data) and GSE103275 (10X data). The authors declare that all data supporting the findings of this study are available within the article and its [Supplementary Information](#) files.

Received: 4 July 2017 Accepted: 27 September 2017

Published online: 20 November 2017

## References

- Inman, J. L., Robertson, C., Mott, J. D. & Bissell, M. J. Mammary gland development: cell fate specification, stem cells and the microenvironment. *Development* **142**, 1028–1042 (2015).
- Visvader, J. E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev.* **28**, 1143–1158 (2014).
- Macias, H. & Hinck, L. Mammary gland development. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 533–557 (2012).
- Kendrick, H. et al. Transcriptome analysis of mammary epithelial subpopulations identifies novel determinants of lineage commitment and cell fate. *BMC Genomics* **9**, 591 (2008).
- Lim, E. et al. Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).
- Raouf, A. et al. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell Stem Cell* **3**, 109–118 (2008).
- Shehata, M. et al. Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.* **14**, R134 (2012).
- Sheridan, J. M. et al. A pooled shRNA screen for regulators of primary mammary stem and progenitor cells identifies roles for Asap1 and Prox1. *BMC Cancer* **15**, 221 (2015).
- Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).
- DeLaughter, D. M. et al. Single-cell resolution of temporal gene expression during Heart development. *Dev. Cell* **39**, 480–490 (2016).
- Luo, Y. et al. Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. *Cell* **161**, 1175–1186 (2015).
- Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
- Grun, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
- Hanchate, N. K. et al. Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science* **350**, 1251–1255 (2015).
- Wollny, D. et al. Single-cell analysis uncovers clonal acinar cell heterogeneity in the adult pancreas. *Dev. Cell* **39**, 289–301 (2016).
- Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
- Olsson, A. et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702 (2016).
- Guo, G. et al. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* **13**, 492–505 (2013).
- Grun, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
- Kim, T. H. et al. Single-cell transcript profiles reveal multilineage priming in early progenitors derived from Lgr5(+) intestinal stem cells. *Cell Rep.* **16**, 2053–2060 (2016).
- Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
- Fu, N. Y. et al. EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nat. Cell Biol.* **17**, 365–375 (2015).
- Makarem, M. et al. Developmental changes in the in vitro activated regenerative activity of primitive mammary epithelial cells. *PLoS Biol.* **11**, e1001630 (2013).

24. 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).
25. 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).
26. Fu, N. Y. et al. Identification of quiescent and spatially restricted mammary stem cells that are hormone responsive. *Nat Cell Biol* **19**, 164–176 (2017).
27. Hughes, A. J. et al. Single-cell western blotting. *Nat. Methods* **11**, 749–755 (2014).
28. Srinivasan, K., Strickland, P., Valdes, A., Shin, G. C. & Hinck, L. Netrin-1/neogenin interaction stabilizes multipotent progenitor cap cells during mammary gland morphogenesis. *Dev. Cell* **4**, 371–382 (2003).
29. Strickland, P., Shin, G. C., Plump, A., Tessier-Lavigne, M. & Hinck, L. Slit2 and netrin 1 act synergistically as adhesive cues to generate tubular bi-layers during ductal morphogenesis. *Development* **133**, 823–832 (2006).
30. Gareau, J. R. & Lima, C. D. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nat. Rev. Mol. Cell Biol.* **11**, 861–871 (2010).
31. Snider, N. T., Weerasinghe, S. V., Iniguez-Lluhi, J. A., Herrmann, H. & Omary, M. B. Keratin hyperumoylation alters filament dynamics and is a marker for human liver disease and keratin mutation. *J. Biol. Chem.* **286**, 2273–2284 (2011).
32. Scheele, C. L. G. J. et al. Identity and dynamics of mammary stem cells during branching morphogenesis. *Nature* **542**, 313–317 (2017).
33. Asselin-Labat, M. L. et al. Gata-3 negatively regulates the tumor-initiating capacity of mammary luminal progenitor cells and targets the putative tumor suppressor caspase-14. *Mol. Cell Biol.* **31**, 4609–4622 (2011).
34. van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
35. Sanchez-Freire, V., Ebert, A. D., Kalisky, T., Quake, S. R. & Wu, J. C. Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat. Protoc.* **7**, 829–838 (2012).
36. Richert, M. M., Schwertfeger, K. L., Ryder, J. W. & Anderson, S. M. An atlas of mouse mammary gland development. *J. Mammary Gland Biol. Neoplasia* **5**, 227–241 (2000).
37. Fata, J. E., Chaudhary, V. & Khokha, R. Cellular turnover in the mammary gland is correlated with systemic levels of progesterone and not 17 $\beta$ -estradiol during the estrous cycle. *Biol. Reprod.* **65**, 680–688 (2001).
38. Briskin, C. Progesterone signalling in breast cancer: a neglected hormone coming into the limelight. *Nat. Rev. Cancer* **13**, 385–396 (2013).
39. 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).
40. Joshi, P. A. et al. Progesterone induces adult mammary stem cell expansion. *Nature* **465**, 803–807 (2010).
41. Asselin-Labat, M. L. et al. Control of mammary stem cell function by steroid hormone signalling. *Nature* **465**, 798–802 (2010).
42. Cabezas-Wallscheid, N. et al. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507–522 (2014).
43. Hu, M. et al. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.* **11**, 774–785 (1997).
44. Laslo, P. et al. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* **126**, 755–766 (2006).
45. Shackleton, M. et al. Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84–88 (2006).
46. Stingl, J. et al. Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993–997 (2006).
47. Wilson, N. K. et al. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* **16**, 712–724 (2015).
48. Pal, B. et al. Integration of microRNA signatures of distinct mammary epithelial cell types with their gene expression and epigenetic portraits. *Breast Cancer Res.* **17**, 85 (2015).
49. Eirew, P. et al. Aldehyde dehydrogenase activity is a biomarker of primitive normal human mammary luminal cells. *Stem Cells* **30**, 344–348 (2012).
50. Asiedu, M. K. et al. AXL induces epithelial-to-mesenchymal transition and regulates the function of breast cancer stem cells. *Oncogene* **33**, 1316–1324 (2014).
51. van Amerongen, R., Bowman, A. N. & Nusse, R. Developmental stage and time dictate the fate of Wnt/ $\beta$ -catenin-responsive stem cells in the mammary gland. *Cell Stem Cell* **11**, 387–400 (2012).
52. Ikeda, J. et al. Prognostic significance of CD55 expression in breast cancer. *Clin. Cancer Res.* **14**, 4780–4786 (2008).
53. Wang, D. et al. Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* **517**, 81–84 (2015).
54. Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
55. 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).
56. 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).
57. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
58. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
59. Chen, Y., Lun, A. T. L. & Smyth, G. K. in *Statistical Analysis of Next Generation Sequence Data.* (eds. Datta, S. & Nettleton, D. S.) Ch.3 (Springer, 2014).
60. McCarthy, D. J. & Smyth, G. K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25**, 765–771 (2009).
61. Shin, C. J., Davis, M. J. & Ragan, M. A. Towards the mammalian interactome: Inference of a core mammalian interaction set in mouse. *Proteomics* **9**, 5256–5266 (2009).
62. Blake, J. A. et al. Mouse genome database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45**, D723–D729 (2017).
63. Cursons, J. & Davis, M. J. Determining the significance of protein network features and attributes using permutation testing. *Methods Mol. Biol.* **1549**, 199–208 (2017).
64. 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).

## Acknowledgements

We are grateful to L. Whitehead for assistance with confocal analysis, G. Wahl for sharing the single-cell qPCR method, B. Capaldo for imaging assistance, K. Hogg, and the Imaging, Animal and FACS facilities at WEHI. This work was supported by the Australian National Health and Medical Research Council (NHMRC) grants #1016701, #1054618, #1100807, #1113133; NHMRC IRIISS; the Victorian State Government through VCA funding of the Victorian Breast Cancer Research Consortium and Operational Infrastructure Support; the Australian Cancer Research Foundation and the Ian Potter Foundation. M.J.D., N.F. and A.C.R. were supported by a National Breast Cancer Foundation (NBCF) Fellowship; G.J.L. and G.K.S. by NHMRC Fellowships #1078730, #1058892; and J.E.V. by NHMRC Fellowships #1037230 and 1102742.

## Author contributions

B.P., J.E.V., and G.J.L. designed the study; B.P., F.V., P.J., A.C.R., S.W., N.F., K.H.L., and F.C.J. performed experiments and analyzed data; Y.C., G.K.S., L.G., and M.J.D. performed bioinformatic analyses; B.P., Y.C., G.K.S., and J.E.V. carried out interpretation of data and manuscript writing.

## Additional information

**Supplementary Information** accompanies this paper at doi:10.1038/s41467-017-01560-x.

**Competing interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://ngp.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017



Minerva Access is the Institutional Repository of The University of Melbourne

**Author/s:**

Pal, B; Chen, Y; Vaillant, F; Jamieson, P; Gordon, L; Rios, AC; Wilcox, S; Fu, N; Liu, KH; Jackling, FC; Davis, MJ; Lindeman, GJ; Smyth, GK; Visvader, JE

**Title:**

Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling

**Date:**

2017-11-20

**Citation:**

Pal, B., Chen, Y., Vaillant, F., Jamieson, P., Gordon, L., Rios, A. C., Wilcox, S., Fu, N., Liu, K. H., Jackling, F. C., Davis, M. J., Lindeman, G. J., Smyth, G. K. & Visvader, J. E. (2017). Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. NATURE COMMUNICATIONS, 8 (1), <https://doi.org/10.1038/s41467-017-01560-x>.

**Persistent Link:**

<http://hdl.handle.net/11343/270600>

**File Description:**

Published version

**License:**

CC BY