Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA-sequencing

Karsten Bach[1,2,5], Sara Pensa[1,5], Marta Grzelak[2,5], James Hadfield[2,5], David J. Adams[3,5], John C. Marioni[2,3,4], Walid T. Khaled[1,5]

1. Department of Pharmacology, University of Cambridge, Cambridge, UK
2. Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
3. Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
4. European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, UK
5. Cancer Research UK Cambridge Cancer Centre, Cambridge, UK

Correspondence to: Walid T. Khaled (wtk22@cam.ac.uk) or John C. Marioni (marioni@ebi.ac.uk)

Keywords: Single-cell RNAseq, Development, Mammary gland

Abstract:

The mammary gland is a unique organ as it undergoes most of its development during puberty and adulthood. Characterising the hierarchy of the various mammary epithelial cells (MECs) and how they are regulated in response to gestation, lactation and involution is important for understanding breast cancer. Recent studies have used only a handful of markers to isolate and characterise the different MEC compartments within the adult gland. There is however, a need for a comprehensive description of MECs at different developmental stages. To this end we used single cell RNA-sequencing (scRNAseq) to determine the gene expression profiles of individual MECs across four adult developmental stages; nulliparous, mid gestation, lactation and post weaning (full natural involution). Our data from 3,471 individual cells identifies 9 distinct MEC clusters and allows their hierarchical structure across development to be charted. We show that only few clusters could be fully characterized by a single marker gene. We argue instead that the epithelial cells – especially in the luminal compartment – should rather be conceptualized as being part of a continuous spectrum of differentiation. This view highlights the plasticity of the tissue and might help to explain some of the conflicting results from lineage tracing studies.

Main

The purpose of the mammary gland is to provide nourishment and passive immunity to the young until they are capable of feeding themselves. From a developmental biology perspective, the mammary gland is a unique organ as it undergoes most of its development during puberty and adulthood[1-4]. In the pre-pubertal mouse the mammary gland consists of a rudimentary epithelial ductal structure embedded within a mammary fatpad, which is connected to the nipple[5, 6]. In response to the hormonal changes during puberty, the rudimentary ductal structure will proliferate and migrate to fill the entire mammary fatpad leaving a developed network of ductal structures that later serve as channels for milk transport during lactation. At the onset of pregnancy a highly proliferative stage is initiated characterised by further ductal side-branching and widespread lobuloalveolar development[1]. Differentiation of the epithelial cells within alveoli prepares the gland for milk production and secretion. Towards the end of pregnancy the gland is extremely dense and primarily occupied by epithelial cells and only few adipocytes. This morphology is largely maintained throughout lactation. However, in response to cessation of suckling the gland undergoes involution, which is characterised by extensive cell death and tissue remodelling[4, 7]. Towards the end of involution the gland reaches a morphology resembling that prior to pregnancy and subsequent pregnancies will trigger the same chain of events.

Recent efforts have focused on the identification and characterisation of the various MEC lineages within the gland that contribute to this developmental homeostasis. Pioneering fat-pad transplantation studies nearly 70 years ago were the first to demonstrate the regenerative and differentiation capacity of small numbers of cells[8-

[10]. More recently the use of cell surface markers coupled with flow cytometry has been used to enrich for various progenitor and stem cell compartments[11-14] and shown that the imbalance of such cell populations results in cellular transformation and subsequently breast cancer[15, 16]. Other studies, inspired by breast cancer transcriptomic profiling, have identified transcriptional regulators of MEC types such as *GATA3* in the luminal cells[14, 17]. More recently, elegant lineage-tracing studies used key markers to address the contribution of each lineage to adult MEC homeostasis[4, 18-20]. However, in all of these studies only a handful of markers and genes were used to define the cellular hierarchy of the MECs, with a principal focus on the nulliparous developmental stage. Therefore, to properly understand its changing role throughout life, there is a need for an unbiased and comprehensive characterisation of MEC compartments at different developmental stages.

To this end we herein use single-cell RNA sequencing (scRNAseq) to map the cellular dynamics of MECs across four adult developmental stages; nulliparous, mid gestation, lactation and post weaning (full natural involution). Our data from 3,471 individual cells identifies 9 distinct cell populations within the gland and allows their hierarchical structure across developmental time points to be charted.

Results

Single-cell RNA sequencing identifies 9 clusters of MECs

First, we isolated MECs from four developmental time points; nulliparous (NP), day 14.5 gestation (G), day 6 lactation (L) and two weeks post natural involution (PI). For each time point we sorted MECs based on the Epcam cell surface marker from two independent mice (Fig. 1a). All samples were then prepared for single-cell RNA-

sequencing using the 10x Chromium platform[21]. Following quality control (materials and methods), this yielded an average of 11,175 unique transcripts and 3,113 genes detected from 3,471 cells (1,681 from NP, 609 from G, 604 from L and 605 from PI) (Supplementary Fig. 1a).

Visual inspection of the data using t-distributed stochastic neighbour embedding (t-SNE) suggested that although there is some grouping of cells by time point, there are also other factors that underlie structure within the dataset (Fig. 1b and Supplementary Fig. 2a). We used unsupervised hierarchical clustering based on the expression of 704 highly variable genes to identify 9 clusters of epithelial cells (C1-C9). Importantly, these clusters were consistent with the groups observed in the t-SNE (Fig. 1c). Based on the expression of *Krt18, Krt8, Krt5, Krt14 and Acta2* we grouped the clusters into 6 luminal (C1, C2, C3, C4, C5 & C8) and 3 basal clusters (C6, C7 & C9) (Fig. 1D, Supplementary Fig. 2b and 2c). Differential gene expression analysis allowed us to further characterise the clusters and infer putative identities (Fig. 1d, Table 1 and Supplementary Table 1). For example, C2 expresses high levels of beta-casein, *Csn2,* and is composed exclusively from cells from the gestation and lactation time points, suggesting that these cells are secretory alveolar cells (Fig. 1d). C1 and C3 on the other hand represent the clusters with the highest expression of hormone receptors. C5 expresses high levels of the luminal progenitor marker *Aldh1a3*[22] suggesting progenitor function, while C7 shows characteristics of myoepithelial cells such as high levels of *Acta2*, *Oxtr* and *Krt15* (Fig. 1d).

Reconstruction of the luminal differentiation hierarchy

We focused on cells from the NP and G time points to investigate mammary epithelial differentiation states of the gland. These differentiation states and the transitions between them can be computationally reconstructed using diffusion maps. Briefly, the method embeds the data in a low-dimensional space, where distances between cells represent the progression through a gradual but stochastic process such as differentiation. In diffusion maps constructed from all epithelial cells, we observed a clear segregation between the luminal and basal clusters, with virtually no transition states between the two (Fig. 2a). This supports the hypothesis that during normal tissue homoeostasis the two lineages are largely self-maintained, which is in agreement with the majority of the lineage tracing studies[18, 23, 24]. In contrast, the luminal compartment showed a distinct structure, with gradual transitions between different clusters and cells originating from a common origin (Fig. 2b). We confirmed the robustness of this bifurcation by verifying that it was present when different methods of feature selection, algorithms for trajectory inference and down-sampling were employed (Supplementary Fig. 3). The expression of the progenitor marker *Aldh1a3* gradually decreased as cells progressed away from their common origin (Fig. 2c), which was largely composed of C5. We further noted that the left arm of the differentiation trajectory terminates at C2 and shows increasing expression of *Csn2 and Glycam1* (Fig. 2c), consistent with a secretory phenotype. Between C5 and C2 we found cells from C8 (Fig. 2b). C8 is the second most rare cell population after C9 and is found in all developmental stages in approximately equal numbers (6 cells in NP, 5 cells in G, 8 cells in L and 7 cells in PI). C8 cells express *Aldh3a1, Elf5, Cd14 and Kit* similar to C5 but in addition express high levels of prolactin induced protein *(Pip)* and apolipoprotein D *(Apod)* (Fig. 1d). On the right arm of the differentiation trajectory, cells in C5 transitioned to C3 and C1, during

which the expression of *Esr1* and *Pgr* increase suggesting that this branch represents differentiation towards hormone-sensing luminal cells (Fig. 2c).

Being confident that the diffusion map recapitulates the luminal differentiation process, we then computationally inferred the two branches and ordered the cells according to their progression through "pseudotime"[25] (see materials and methods for further explanation) (Fig. 3a). We found 828 genes that showed a pseudotime-dependent expression with the same directionality along both differentiation trajectories (Fig. 3b, top panel of heatmap, Supplementary Table 2, Supplementary Fig. 4a). These included genes associated with known progenitor characteristics such as *Aldh1a3*, *Kit*, *Cd14* and *Tspan8* as well as transcription factors that have not previously been associated with luminal differentiation such as *Creb5*, *Hey1* and *Fosl1* (Fig. 3c). In addition, we identified 1459 genes with branch-specific expression patterns (Fig. 3b, bottom panel of heatmap, Supplementary Table 3, Supplementary Fig. 4b), including milk related genes (*Wap*, *Csn1s1*, *Csn3*) as well as genes involved in hormone-receptor signalling (*Pgr, Prlr, S100a6, Cited1, Areg*). Amongst the branch-specific transcription factors we found for example *Runx1, Tox2* and *Bhlhe41* (also known as *Sharp1*) to be transcribed during differentiation towards the hormone-sensing lineage (Fig. 3d) and *Elf5, Foxs1* and *Ehf* in the secretory lineage (Fig. 3e). Interestingly, *Runx1* is a known repressor of *Elf5* and its deletion has been shown to be deleterious for ductal morphogenesis[26]. The expression of the known progenitor master regulator *Elf5* is maintained and further increased during secretory differentiation suggesting that its transcriptional level is fine-tuned in luminal progenitors (Fig. 3e).

Cluster 4 represents a post involution specific progenitor cell

From the initial clustering analysis we noticed that C4 appeared only after the animals had undergone a full pregnancy and was predominantly made up of cells from PI (0 cells from NP, 0 cells from G, 83 cells from L and 339 cells from PI) (Fig. 1b). Moreover, multiple progenitor marker genes from C5 were also present in C4 (Fig. 1c, Supplementary Fig. 5). The similarity between the two clusters also held true on a transcriptome-wide level (Fig. 4a, Supplementary Fig. 10c). We thus hypothesised that C4 represents the luminal progenitor population in the post-parous gland, which is supported by the observation that C5 virtually disappears after parity (Table 1). To further support this hypothesis, we identified genes that distinguish C5 from the rest of the luminal compartment to see if they are also characteristic for the proposed post-parity progenitor population C4. Indeed, we find genes that are differentially expressed between C5 and the rest of the luminal compartment to show the same trend between C4 and the luminal compartment (Fig. 4b). In a similar manner we can distinguish C4 from the rest of the PI gland in a principal component analysis (PCA) using the identified progenitor genes (Supplementary Fig. 6a and b). Given this high degree of similarity we were not surprised that only 120 genes were differentially expressed between C4 and C5 (Supplementary Table 4). Interestingly, genes that were up-regulated in C4 were significantly enriched for pathways that are involved in the immune response and lactation (Fig. 4c, Supplementary Fig. 6c). Furthermore, the expression of milk genes was surprisingly maintained at high levels in C4 even after lactation (Fig. 4d). Of note, the genes of the casein locus (*Csn2*, *Csn1s1*, *Csn1s2a*, *Csn3*) have previously been reported to be up-regulated in the parous gland, most likely due to changes in chromatin accessibility[27, 28]. However, it has not been shown before that this effect is confined to the progenitor population of

188 the luminal compartment (Supplementary Fig. 7). Together, the data suggest that
189 luminal progenitor cells maintain memory of having undergone gestation and
190 involution. This memory could potentially prime progenitor cells towards the alveolar
191 fate to facilitate alveologenesis in subsequent pregnancies.
192
193 **Discussion**
194 We have reported here the use of single-cell RNA sequencing to comprehensively
195 map the transcriptomes of thousands of MECs across four developmental time
196 points. Our analysis identified 9 clusters of epithelial cells, some of which are only
197 present during specific developmental stage (e.g. C2 and C4). This study provides a
198 rich dataset that can be mined online (see link in materials and methods) to identify
199 marker genes and lineage specific factors that can be used to trace populations of
200 cells *in vivo*. We note, however, that only some of the clusters can be fully
201 characterized by a single marker gene. Instead we argue that the epithelial cells –
202 especially in the luminal compartment – should rather be conceptualized as being
203 part of a continuous spectrum of differentiation as visualised in Fig. 2. This view
204 highlights the plasticity of the tissue and might help to explain some of the conflicting
205 results from lineage tracing studies[4]. In this study we could not provide any evidence
206 for contribution of a putative multipotent stem cell to the day-to-day homoeostasis of
207 the gland. We do note, however, that the cluster C9 is placed between the luminal
208 and basal cells in Fig. 2a. C9 represents the cluster with the fewest numbers of cells
209 captured and expresses low levels of both luminal and basal markers, high levels of
210 the stem cell marker *Procr*[19] and high levels of the luminal progenitor marker
211 *Notch3*[29] (Fig. 1d). Yet, we did not capture cells in transition between C9 and any
212 other clusters suggesting that if these cells are indeed, multipotent progenitors their

involvement in tissue homoeostasis is negligible. Alternatively, we cannot exclude the possibility that C9 is enriched for doublets.

Based on the gene expression data presented here, the luminal compartment appears to have one common progenitor population (C5). We also identified a previously described progenitor population with expression of hormone-receptor related genes (C3) that appears to represent a meta-stable state on the differentiation path from luminal progenitors towards fully differentiated hormone-sensing cells (C1). It is worth noting that the expression of *Ly6a (Sca-1)* a marker recently described to differentiate between two sets of luminal progenitors[22, 30] is absent from C5 and is present in C3 (Fig. 1d). In the same study they describe that Sca-1^{+ve} cells are also ER^{+ve} and Sca1^{-ve} progenitors are ER^{-ve} this is in agreement with the expressions of both genes in C5 and C3. Furthermore, we characterised gene expression patterns along the differentiation hierarchy, thus enabling to disentangle the cellular hierarchy in the mammary gland. We were able to identify new transcription factors, which mark the progenitor populations and the two differentiated lineages (Fig. 3 and Supplementary Table 2 and 3). Investigating how these novel transcription factors are involved in cell fate decisions and cancer development should be investigated in the future.

By analysing the mammary gland at various stages of development we were also able to describe the molecular effects of parity at cellular resolution. We found that the luminal progenitor compartment undergoes lasting changes at the transcriptional level. This is especially interesting in light of the protective effect of pregnancies against breast cancer[31, 32] and the role of luminal progenitors as cell of origin[15]. The

majority of the changes were related to pathways involved in immunity and lactation, suggesting that, in particular, the luminal progenitors maintain a memory of gestation and involution. It is reasonable to assume that C4 overlaps with the previously described parity-induced MECs (PI-MECs)[33].

In summary, this study provides a novel view of mammary gland development. Our unbiased approach helps support some previously formed hypotheses in the mammary gland field and describes differentiation processes at a high cellular resolution. The dataset will be a useful resource for future studies that aim to understand the relationship of the different cell types in the gland and how breast cancer develops and progresses.

Figure legends

Figure 1. Single-cell RNA sequencing identifies 9 clusters of MECs

(a) Schematic diagram highlighting the experimental setup for isolating and sequencing the RNA of single cells using the 10X Chromium system. **(b)** t-SNE plot of 3,471 cells visualizes general structure in the data. Cells are coloured by the four developmental time points Pink=NP, Dark Green=G, Light Green=L, Purple=PI. **(c)** Same as **(b)** but colouring cells by clusters. **(d)** Heatmap highlighting some key marker genes. Colour scale represents log-transformed and normalized UMI counts scaled to a maximum of 1 per row. The upper panel shows genes that were used to distinguish between luminal and basal cells. Upper bars represent the cluster assignment and stages for the individual cells. For visualization purposes only 100 randomly selected cells were shown for each of the large clusters. The list of differentially expressed genes between clusters can be found in Supplementary Table 1.

Figure 2. Computational reconstruction of differentiation processes in the mammary gland

(a) Diffusion map of epithelial cells from the NP and G time points, showing the first three diffusion components. **(b)** Differentiation trajectory of the luminal compartment based on the first two diffusion components. **(c)** This panel depicts the log-transformed normalized expression levels of *Aldh1a3, Csn2, Glycam1, Pgr and Esr1* in all cells along the differentiation trajectory shown in **(b)**. Colour scale represents log-transformed and normalized UMI counts scaled to a maximum of 1.

Figure 3. Pseudotime ordering identifies genes associated with luminal differentiation

(a) Definition of the hormone-sensing and secretory differentiation branch. Cells are coloured by their progression through pseudotime, where low values represent undifferentiated cells. **(b)** Heatmap of the top 50 pseudotime-dependent genes (rows) with the same overall trend along both branches (top panel) or branch-specific trends (bottom panel, see materials and methods for definition). The black vertical line represents the common origin from the progenitor population. Arrows highlight differentiation direction. Pseudotime and the cluster assignment are annotated above the heatmap. The values in the heatmap represent z-scaled, spline-smoothed expression values. **(c-e)** Examples of transcription factors with pseudotime-dependent expression with the same overall trend on both branches **(c)** or branch specific trends **(d,e)**.

Figure 4. The effect of parity on the transcriptomic landscape of the luminal progenitor compartment

(a) Colour-coded similarity matrix between the log-transformed mean expression values of the 9 epithelial cell clusters. Similarities are based on Euclidean distances. **(b)** Comparison of fold changes from C5 versus luminal compartment and fold changes from C4 versus the luminal cells. The genes represent the top 500 differentially expressed genes between C5 and luminal cells. **(c)** Volcano plot illustrates differential expression between C5 and C4, coloured dots represent significant genes with known function in lactation and immunity, and dashed lines highlight the P value threshold of 0.01 and a log fold change of 1. P values are adjusted for multiple testing using Benjamini-Hochberg. **(d)** Visualization of

298 expression difference for some lactation related genes from **(c)**, where C4 is split
299 according to the developmental time point.
300
301 **Table 1. Summary of MEC clusters**
302 Overview of the different clusters including number of cells captured for each time
303 point and key genes that were used to infer their identities.
304
305 **Supplementary Figure 1. Quality control of sequencing data**
306 **(a)** Table summarizing quality control criteria per sample. Number of unique
307 molecules, genes detected and number of reads represent the median value for each
308 sample. **(b-c)** Histograms for the four conditions showing the distributions of the
309 number of genes detected **(b)** or total number of molecules **(c)** metric with indicated
310 threshold (dashed line). **(d)** Scatterplot of number of genes detected versus
311 percentage of mitochondrial RNA molecules, the threshold at 5% is indicated. For
312 choice of thresholds see materials and methods.
313
314 **Supplementary Figure 2. Differences between replicates and expression of**
315 **luminal and basal markers**
316 **(a)** t-SNE plot coloured by the eight different samples. Some minor structures (e.g.
317 two groups PI1 and PI2 at the top) are mainly replicate dependent, but all of the
318 clusters identified in the analysis are represented by multiple replicates. **(b-c)** t-SNE
319 plots highlighting the expression of luminal **(b)** and basal **(c)** genes in all 3,471 cells.
320
321 **Supplementary Figure 3. Luminal bifurcation is robust to various alterations of**
322 **the trajectory inference**

(a-b) Eigenvalues for the diffusion map used in Fig. 2a. **(a)** and Fig. 2b **(b)**. **(c)** The differentiation trajectory as determined by Monocle (see materials and methods) coloured by cluster assignment. **(d)** The diffusion map is robust towards down-sampling of cells (100, 50 or 25% of all cells were used in the left, middle or right panel, respectively) as well as the method of feature selection (all= all genes with mean expression level above 0.1, HVG= highly variable genes, PCA= first 50 components from PCA, selected= a manually selected choice of genes that are known to be involved in luminal cell differentiation). The gene list included: *Csn2, Gata3, Prlr, Elf5, Esr1, Pgr, Aldh1a3, Wap, Tspan8, Krt18, Krt8, Areg, Fgfr1, Fgfr2, Notch1, Notch3, Foxc1* and *Zeb2*.

Supplementary Figure 4. Full heatmaps of pseudotime dependent expression

Heatmaps as in Fig. 3, showing all genes that are pseudotime dependent in their expression along the trajectories with the same overall trend **(a)** or opposing trend **(b)**. Full gene list can be found in Supplementary Tables 2 and 3.

Supplementary Figure 5. Expression of progenitors markers in C4

Panel of t-SNE plots (compare to Fig. 1) with coloured normalized log-expression values of progenitor markers.

Supplementary Figure 6. C4 is a post-parity progenitor population

(a) Principal component analysis (PCA) on all luminal cells from the PI time point. The PCA was computed on the top 500 differentially expressed genes between C5 and the rest of the luminal NP gland. PC1 separates the progenitors and the differentiated cells, with the progenitors showing negative PC1 values. **(b)** Genes that

are higher expressed in C5 compared to the rest of the NP gland also have negative PC1 loadings. **(c)** Top 20 GO-terms (biological processes) that are significantly enriched in upregulated genes from Fig. 4c. Dashed line indicates P value threshold at 0.001.

Supplementary Figure 7. Expression of milk genes after parity is restricted to the progenitor compartment

Panel of t-SNE plots (compare to Fig. 1) with coloured normalized log-expression values of milk genes from Fig. 4d.

Supplementary Figure 8. Quality control of cells from the Lactation sample

t-SNE plot computed only on the lactation sample. Cells that were flagged as low quality are highlighted in grey and cells that passed the QC are shown in black. Although a majority of the sample was removed, the cells that passed represent at least all major groups from this sample (see materials and methods).

Supplementary Figure 9. Outline of the clustering strategy

(a) Dendrogram of the hierarchical clustering using Euclidean distances and average linkage. Colour bars indicate the animal (binary, only comparable within one condition), the condition and cluster assignment. **(b)** Cluster stability as determined by the overlap between the original cluster and the most similar cluster in the bootstrap samples. The distributions of jaccard indeces are shown as a boxplot for 100 bootstrap samples. Stability is shown for three values of the deepSplit parameter (0,1,2). Deep split 1 was chosen as this maintained overall a high cluster stability while identifying relevant cell types (see materials and methods).

Supplementary Figure 10. Clustering Quality Control and removal of C10

(a) Total molecule count grouped by different clusters. **(b)** Markers that were used to identify C10 as lymphocyte cluster[34] **(c)** Separation of C10 from all other clusters as shown by a PCA on the cluster average gene expression.

Table S1. Differentially expressed genes between clusters

Table S2. List of genes with pseudotime dependent gene expression with same overall trend in the two branches

Table S3. List of genes with pseudotime dependent gene expression with different trends in the two branches

Table S4. Genes differentially expressed between C4 and C5

Materials and methods

Animals

All experimental animal work was performed in accordance to the Animals (Scientific Procedures) Act 1986, UK and approved by the Ethics Committee at the Sanger Institute. C57BL/6N mice were housed in individually ventilated cages under a 12:12 h light-dark cycle, with water and food available *ad libitum*. The experiment was set up to allow for all of the developmental time points to be collected and tissues to be processed at the same time. Mice were euthanized by terminal anaesthesia. Females were mated with studs and allowed to litter. Tissues were then harvested at gestation day 14.5 (G), lactation day 6 (L) and day 11 post natural weaning of the pups (PI). Tissue from NP females was harvested at 8 weeks of age. Two individual mice per developmental time point were used in the study.

Mammary gland dissociation into single-cell suspension

Lymph node divested mammary glands (excluding the cervical pair and the proximal region of one of the fourth glands) were dissected from the mice and mechanically dissociated. The finely minced tissue was transferred to a digestion mix consisting of DMEM/F12 (Gibco) + 10 mM HEPES (Gibco) + 1 mg ml^{-1} collagenase (Roche) + 100 U ml^{-1} hyaluronidase (Sigma) + 50 µg ml^{-1} gentamicin (Gibco) for 2.5 hours at 37 °C and vortexed every 30 minutes. After the lysis of red blood cells in NH_4Cl, cells were briefly digested with warm 0.05% Trypsin-EDTA (Gibco), 5 mg ml^{-1} dispase (Sigma) and 1 mg ml^{-1} DNase (Sigma), and filtered through a 40 µm cell strainer (BD Biosciences).

Cell labelling, flow cytometry and sorting

Single cell suspensions were incubated in HF medium (Hank's balanced salt solution (Gibco) + 1% foetal bovine serum, Gibco) + 10% normal rat serum (Sigma) for 20 min on ice to pre-block before antibody staining. All antibody incubations were performed for 10 min on ice in HF media. Mammary cells were stained with the following primary antibodies: 1 μg ml^{-1} CD31–biotin (eBioscience); 1 μg ml^{-1} CD45–biotin (eBioscience); 1 μg ml^{-1} Ter119–biotin (eBioscience) and 0.5 μg ml^{-1} EpCAM–PE (Biolegend). Cells were then stained with 0.4 μg ml^{-1} streptavidin–PE-CF594 (BD-Biosciences). Propidium Iodide (PI, 1 μg ml^{-1}; Sigma) was used to detect dead cells. Cells were filtered through a 30 μm cell strainer (Partec) before sorting. Sorting of cells was done using a SH800 sorter (SONY). Single-stained control cells were used to perform compensation manually and unstained cells were used to set gates. Chip alignment and sorting calibration was performed with automatic setup beads (SONY) immediately prior to sorting. Doublets, dead cells and contaminating haematopoietic, endothelial and stromal cells were gated out and EpCAM positive cells were sorted in FACS tubes (BD Biosciences) containing 1 mL of HF. After sorting, cells were transferred to a 1.5 ml tube through a 30 μm cell strainer, spun down and resuspended in 15-50 μl of HF according to the number of cells sorted per sample. Representative samples were manually counted using an improved Neubauer chamber to estimate sample loss from centrifugation and filtering. Equal numbers of cells per sample were processed for scRNA library preparation. Samples were processed for scRNA library preparation within 10 hours from tissue isolation.

Library preparation and sequencing

Library preparation was performed according to instruction in the 10X chromium single cell kit.

RNA-seq data processing

Read processing was performed as previously reported[21]. Briefly, the Cell Ranger Single-Cell Software Suite was used for demultiplexing, barcode assignment and UMI quantification (http://software.10xgenomics.com/single-cell/overview/welcome). The reads were aligned to the mm10 reference genome using a pre-built annotation package obtained from the 10X Genomics website. All lanes per sample were processed using the "cellranger count" function. The outputs from different lanes were then aggregated using "cellranger aggr" with –normalize set to "none".

Quality control and preprocessing

In total the Cell Ranger software identified 5,598 barcodes that contained enough unique molecules to be considered as cells (1,707 in NP, 623 in G, 2,648 in L and 620 in PI). Libraries prepared from the NP, G and P time points all showed high quality that was reproducible between the two biological replicates (Supplementary Fig. 1a). We used the following metrics to flag poor quality cells: number of genes detected, total number of unique molecules (UMIs) and percentage of molecules mapped to mitochondrial genes. For these samples we then identified poor quality cells by setting a threshold on the number of genes and number of UMIs that was defined as four median absolute deviations (MAD) below the median for each sample (Supplementary Fig. 1b and c). The samples from the lactation sample, however, showed on average nearly one order of magnitude fewer genes detected and total number of UMIs. As this was specific to both lactation samples, we inferred that this was driven by the biological phenotype of the sample rather than technical artefacts. Nonetheless, we flagged a substantial amount of the lactation sample as poor quality cells by setting a fixed threshold of 500 detected genes and 1000 total molecules

detected (Supplementary Fig. 1b and c). This was necessary in order to allow for normalisation of all samples; otherwise, keeping more cells of lower quality would inflate the range of normalisation factors amplifying technical noise due to cell-specific biases and adversely impact downstream analyses. We ensured that we still sampled cells from all major structures present in the lactation sample (Supplementary Fig. 8). Finally, all cells with 5% or more of UMIs mapping to mitochondrial genes were defined as non-viable or apoptotic and removed form the analysis (Supplementary Fig. 1d). This left us with a total number of 3,499 cells (1,681 in NP, 609 in G, 604 in L and 605 in P). The cells were then normalised by size factors as previously described[35]. The log-transformed (log2(counts+1)) counts of highly variable genes (HVGs) were used as features for dimensionality reduction and clustering. HVGs were defined using the method from Brennecke et al.[36]. Genes with a Benjamini-Hochberg adjusted P value smaller than 0.1 were considered to be significant. The t-SNE embedding in Fig. 1 was computed using the “Rtsne” package on scaled, log-transformed expression values with default settings and perplexity set to 25 (https://github.com/jkrijthe/Rtsne).

Clustering

As we expected the epithelial compartment to show a hierarchical organisation, we used agglomerative hierarchical clustering to identify cell types. We first computed the pairwise Euclidean distances between cells based on the log-transformed, normalized counts of HVGs. The dissimilarity matrix was used to perform hierarchical clustering (Supplementary Fig. 9a) with the 'hclust' function in R using average linkage, as this resulted in the highest cophenetic correlation coefficient between the tree and the dissimilarity matrix (average:0.91 single: 0.73, complete: 0.83, ward.D2:

485 0.73). Clusters were then defined with the 'cutreeDynamic' function from the
486 'dynamicTreeCut' package (minClusterSize= 15, deepSplit=1, method="hybrid")[37].
487 The deepSplit parameter was set to 1 based on evaluating the robustness of the
488 clustering. At a deepSplit of 0, all clusters were highly robust as determined by
489 bootstrapping[38]. Increasing the parameter to 1, led to splitting C1 from C3 as well as
490 C9 from C6. C1, C3 and C6 were highly robust in the bootstrap (see Supplementary
491 Fig. 9b). C9 showed a below average robustness. However, due to the expression of
492 marker genes for a previously reported cell type[19] we decided to keep it as a
493 separate cluster. Interestingly, the putative cell type of C9 is known to have an
494 expression profile that is in between luminal and basal cells (see main text), which
495 might explain its assignment to other clusters during the bootstrap. An alternative
496 explanation that cannot be excluded at this point is that C9 is enriched for doublets
497 (see discussion). Further increasing the deepSplit parameter led to unstable clusters
498 (Supplementary Fig. 9b). With this method, we identified ten clusters. In addition, the
499 dynamic tree cut method allows for a "noise component" for cells that do not match to
500 any cluster. 1 cell from the lactation sample was flagged as noise by the algorithm
501 and removed from further analysis. C10 (27 cells) expressed lymphocyte related
502 genes was also removed from all downstream analyses (as previously reported[34],
503 Supplementary Fig. 10).

504 **Differential expression analysis**

505 Differential gene expression analysis was performed using "edgeR"[39]. Genes with a
506 mean expression level below 0.1 counts were removed from the analysis. A negative
507 binomial generalized log-linear model was fitted to the remaining genes with the

cluster assignments as covariate(s). The 'glmTreat' function was used to identify genes that have a significantly higher log fold change than 1 at an FDR of 0.01.

Diffusion maps and pseudotime inference

For inferring the differentiation trajectory we used diffusion maps. First, we selected all cells from the NP and G time point (Fig. 2a) and detected the HVGs as described above. The log-transformed (log2(count+1)) gene counts were then used to compute the first twenty diffusion components using the 'DiffusionMap' function (k=50, otherwise the default parameters as in 'destiny'[40]. After inspecting the fraction of explained variation for the first twenty components, we decided to retain the first 3. In Fig. 2b we then focused on the luminal compartment and recomputed the diffusion map based only on the luminal cells, using the aforementioned procedure. Here we retained only the first two components (Supplementary Fig. 3a). Notably, the structure inferred by the diffusion map algorithm was robust to the choice of features and down-sampling of cells (Supplementary Fig. 3d). The structure of a common origin and the two branches could also be inferred using Monocle with standard settings[41] (Supplementary Fig. 3c). For inferring the branches and pseudotime ordering, we defined the following three tips, the cell with the smallest value for the second eigenvector (which was set as root) and the cells with the largest and smallest values for the first eigenvector (compare Fig. 2b).

Pseudotime-dependent expression

To identify genes whose expression was significantly associated with the pseudotime we first fitted a natural cubic spline with three degrees of freedom to the log-transformed (log2(counts+1)) expression data in each branch. A likelihood ratio test

was then used to assess statistical significance of the fit compared to a null (pseudotime-independent) model. Genes with a Benjamini-Hochberg corrected P value below 0.001 were considered to be significantly pseudotime-dependent. We then used a heuristic definition of branch-specific expression instead of modelling the branch assignment explicitly. This was motivated as follows. We were interested in general expression trends of genes, i.e. increase or decrease along the differentiation towards one of the two cell types, rather than comparing the exact timing of gene activation/inactivation between the branches. Any approach trying to do the latter would have been complicated by the different cell densities along the branches and by the difficulty of verifying any such hypotheses *in vivo*. Hence, we defined genes to be branch-specific when they were pseudotime-dependent in their expression in at least one of the two branches and when the gradient differed in signs between two branches. The gradient was determined as the coefficient of a linear model fit to the spline-smoothed expression values, which was set to 0 if the coefficient was not significantly different from 0 at alpha=0.01. Consequently, the gradient of a gene could either be -1 (decreasing), 0 (flat) or 1 (increasing).

Gene set enrichment analysis

A gene set enrichment analysis based on gene-ontology (GO) terms was conducted to characterize genes that were up-regulated in C4 compared to C5. Genes with positive log fold changes were compared to all genes that were tested for differential expression using topGO with default settings[42].

Code availability

All computational analyses were performed in R (Version 3.3.3) using standard functions unless otherwise indicated. Code is available online at https://github.com/MarioniLab/MammaryGland. The processed data can be browsed and accessed at https://karstenbach.shinyapps.io/webapp/

Author contribution

K.B. performed most of the experiments and all the computational analysis. S.P. and K.B. setup and collected the MECs. M.G. and J.H performed the 10X library production and sequencing. J.C.M, D.A., K.B and W.T.K conceptualised the study and wrote the manuscript.

Acknowledgements

We would like to thank the staff at Sanger Institute, Research Service Facility (RSF) for their assistance. We would like to thank Dr. Aaron T. Lun (CRUK CI) for helpful discussions and comments on the manuscript. K.B. is funded by a Cambridge Cancer Centre studentship. D.A. is funded by CRUK and Wellcome Trust. S.P. is funded by CRUK. J.C.M is funded by CRUK and EMBL. W.T.K is funded by a CRUK career establishment award (C47525/A17348), University of Cambridge and Magdalene College, Cambridge.

References

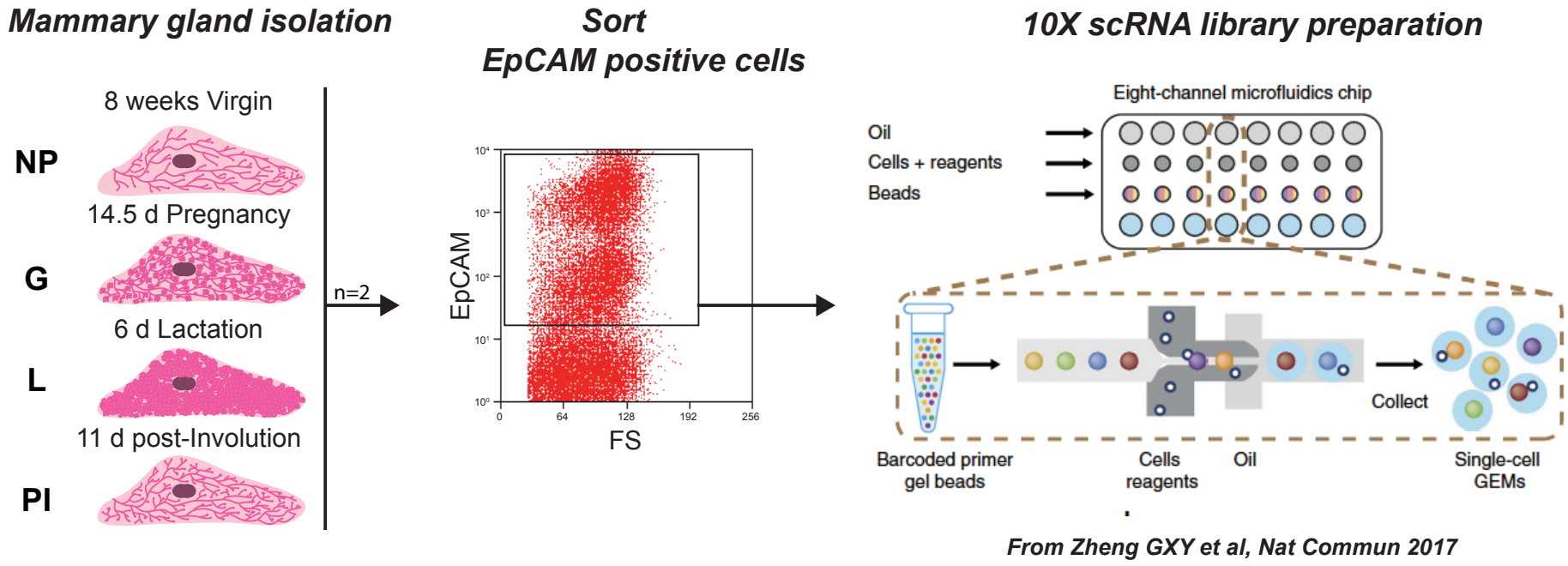
1. Watson, C.J. & Khaled, W.T. Mammary development in the embryo and adult: a journey of morphogenesis and commitment. *Development (Cambridge, England)* **135**, 995-1003 (2008).
2. Hennighausen, L. & Robinson, G.W. Information networks in the mammary gland. *Nat Rev Mol Cell Biol* **6**, 715-725 (2005).
3. Hennighausen, L. & Robinson, G.W. Think globally, act locally: the making of a mouse mammary gland. *Genes & development* **12**, 449-455 (1998).

4. Inman, J.L., Robertson, C., Mott, J.D. & Bissell, M.J. Mammary gland development: cell fate specification, stem cells and the microenvironment. *Development (Cambridge, England)* **142**, 1028-1042 (2015).
5. Mikkola, M.L. & Millar, S.E. The mammary bud as a skin appendage: unique and shared aspects of development. *Journal of mammary gland biology and neoplasia* **11**, 187-203 (2006).
6. Hens, J.R. & Wysolmerski, J.J. Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland. *Breast Cancer Res* **7**, 220-224 (2005).
7. Watson, C.J. Post-lactational mammary gland regression: molecular basis and implications for breast cancer. *Expert Reviews in Molecular Medicine* **8**, 1 (2006).
8. Kordon, E.C. & Smith, G.H. An entire functional mammary gland may comprise the progeny from a single cell. *Development (Cambridge, England)* **125**, 1921-1930 (1998).
9. Faulkin, L.J., Jr. & Deome, K.B. Regulation of growth and spacing of gland elements in the mammary fat pad of the C3H mouse. *J Natl Cancer Inst* **24**, 953-969 (1960).
10. Daniel, C.W. Regulation of cell division in aging mouse mammary epithelium. *Adv Exp Med Biol* **61**, 1-19 (1975).
11. Smalley, M.J., Titley, J. & O'Hare, M.J. Clonal characterization of mouse mammary luminal epithelial and myoepithelial cells separated by fluorescence-activated cell sorting. *In Vitro Cell Dev Biol Anim* **34**, 711-721 (1998).
12. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993-997 (2006).
13. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84-88 (2006).
14. Asselin-Labat, M.L. *et al.* Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nature cell biology* **9**, 201-209 (2007).
15. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature medicine* **15**, 907-913 (2009).
16. Molyneux, G. *et al.* BRCA1 Basal-like Breast Cancers Originate from Luminal Epithelial Progenitors and Not from Basal Stem Cells. *Cell Stem Cell* **7**, 403-417 (2010).
17. Kouros-Mehr, H., Slorach, E.M., Sternlicht, M.D. & Werb, Z. GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell* **127**, 1041-1055 (2006).
18. Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189-193 (2011).
19. Wang, D. *et al.* Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* **517**, 81-84 (2015).
20. 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).
21. Zheng, G.X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
22. Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res* **14**, R134 (2012).

23. 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).
24. Davis, F.M. *et al.* Single-cell lineage tracing in the mammary gland reveals stochastic clonal dispersion of stem/progenitor cell progeny. *Nature communications* **7**, 13053 (2016).
25. Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845-848 (2016).
26. van Bragt, M.P., Hu, X., Xie, Y. & Li, Z. RUNX1, a transcription factor mutated in breast cancer, controls the fate of ER-positive mammary luminal cells. *Elife* **3**, e03881 (2014).
27. Dos Santos, C.O., Dolzhenko, E., Hodges, E., Smith, A.D. & Hannon, G.J. An epigenetic memory of pregnancy in the mouse mammary gland. *Cell Rep* **11**, 1102-1109 (2015).
28. Rijnkels, M. *et al.* Epigenetic modifications unlock the milk protein gene loci during mouse mammary gland development and differentiation. *PloS one* **8**, e53270 (2013).
29. Lafkas, D. *et al.* Notch3 marks clonogenic mammary luminal progenitor cells in vivo. *The Journal of cell biology* **203**, 47-56 (2013).
30. Giraddi, R.R. *et al.* Stem and progenitor cell division kinetics during postnatal mouse mammary gland development. *Nature communications* **6**, 8487 (2015).
31. Russo, I.H. & Russo, J. Pregnancy-induced changes in breast cancer risk. *Journal of mammary gland biology and neoplasia* **16**, 221-233 (2011).
32. Lyons, T.R., Schedin, P.J. & Borges, V.F. Pregnancy and breast cancer: when they collide. *Journal of mammary gland biology and neoplasia* **14**, 87-98 (2009).
33. Wagner, K.U. *et al.* An adjunct mammary epithelial cell population in parous females: its role in functional adaptation and tissue renewal. *Development (Cambridge, England)* **129**, 1377-1386 (2002).
34. Scheele, C.L. *et al.* Identity and dynamics of mammary stem cells during branching morphogenesis. *Nature* **542**, 313-317 (2017).
35. Lun, A.T., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome biology* **17**, 75 (2016).
36. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**, 1093-1095 (2013).
37. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)* **24**, 719-720 (2008).
38. Hennig, C. fpc: Flexible Procedures for Clustering. R package version 2.1-10. https://CRAN.R-project.org/package=fpc ((2015)).
39. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-140 (2010).
40. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics (Oxford, England)* **32**, 1241-1243 (2016).
41. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).

681 42. J, A.A.a.R. topGO: Enrichment Analysis for Gene Ontology. *R package*
682 *version 2.26.0.* (2016).
683

a



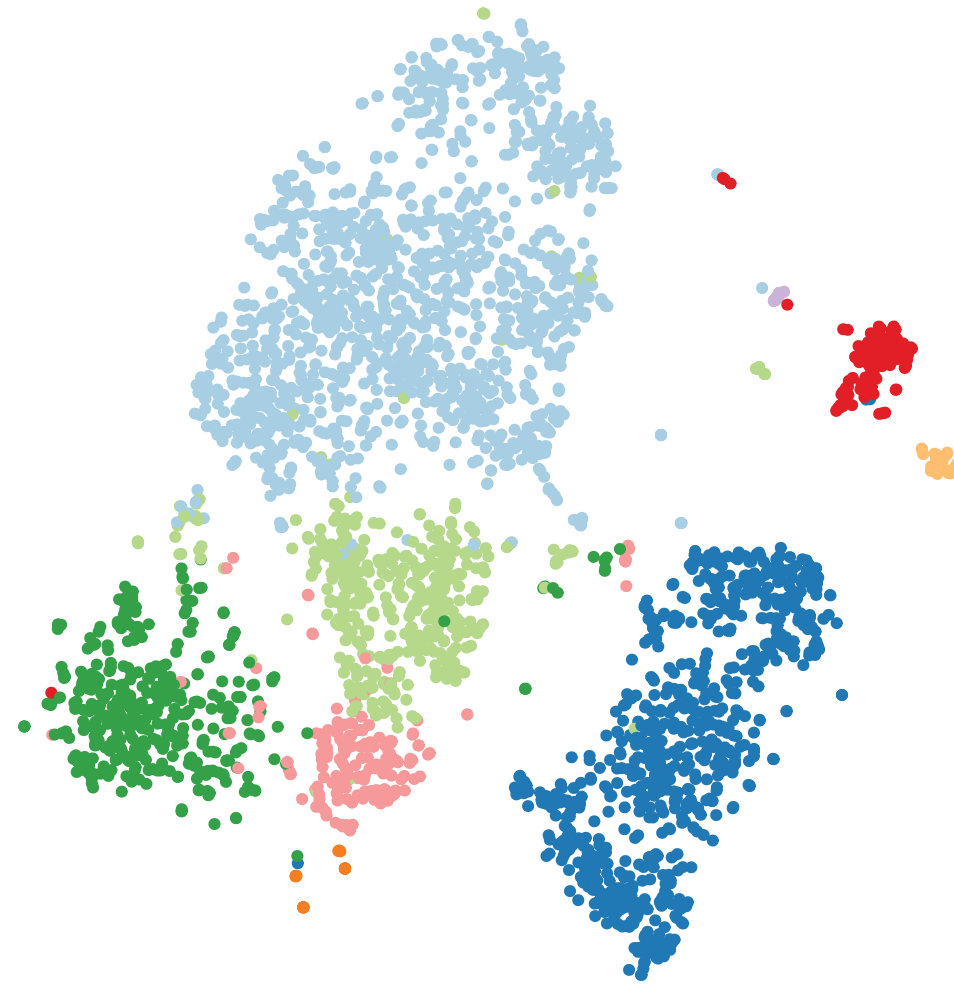
From Zheng GXY et al, Nat Commun 2017

b

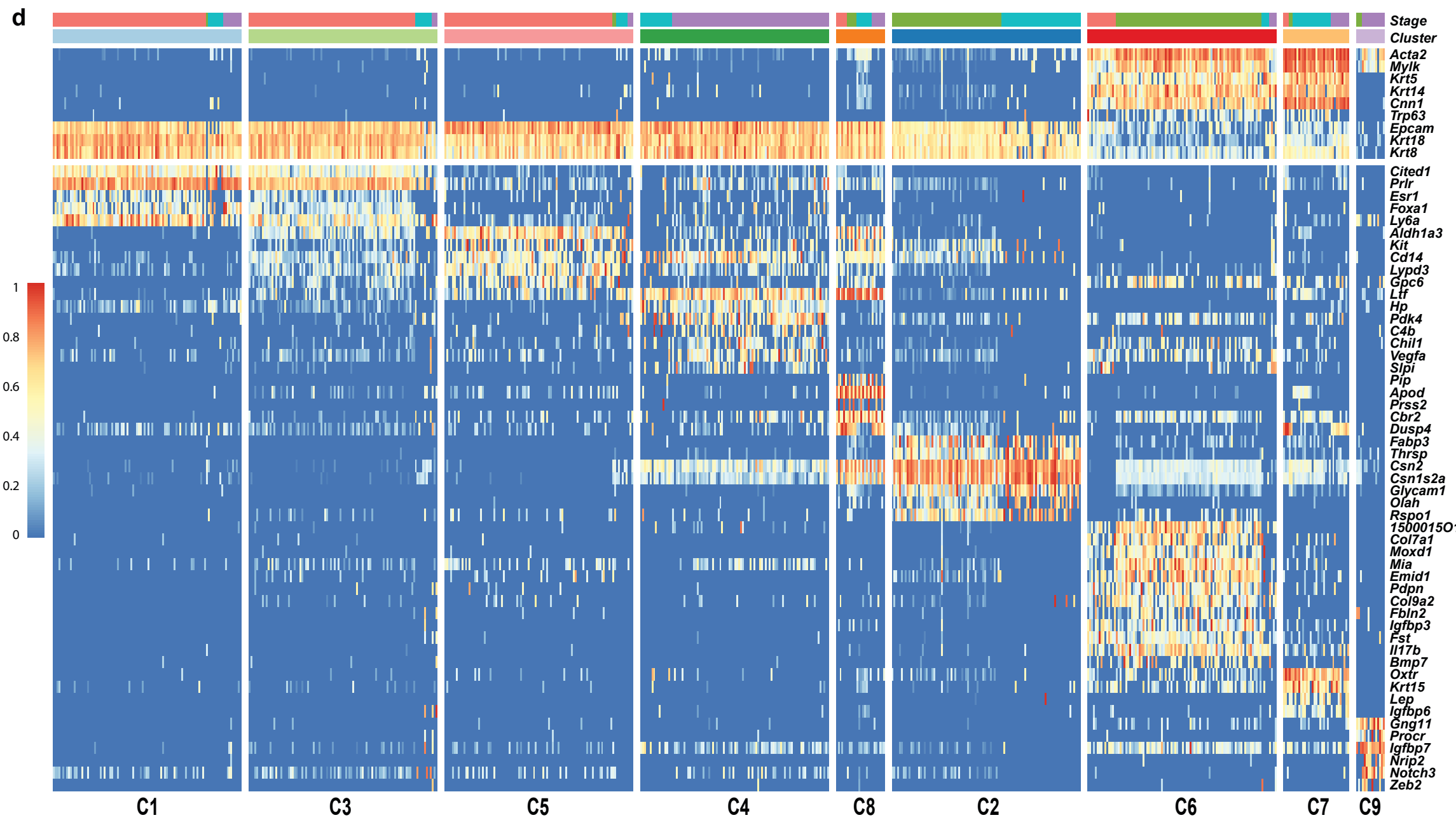
t-SNE coloured by timepoint

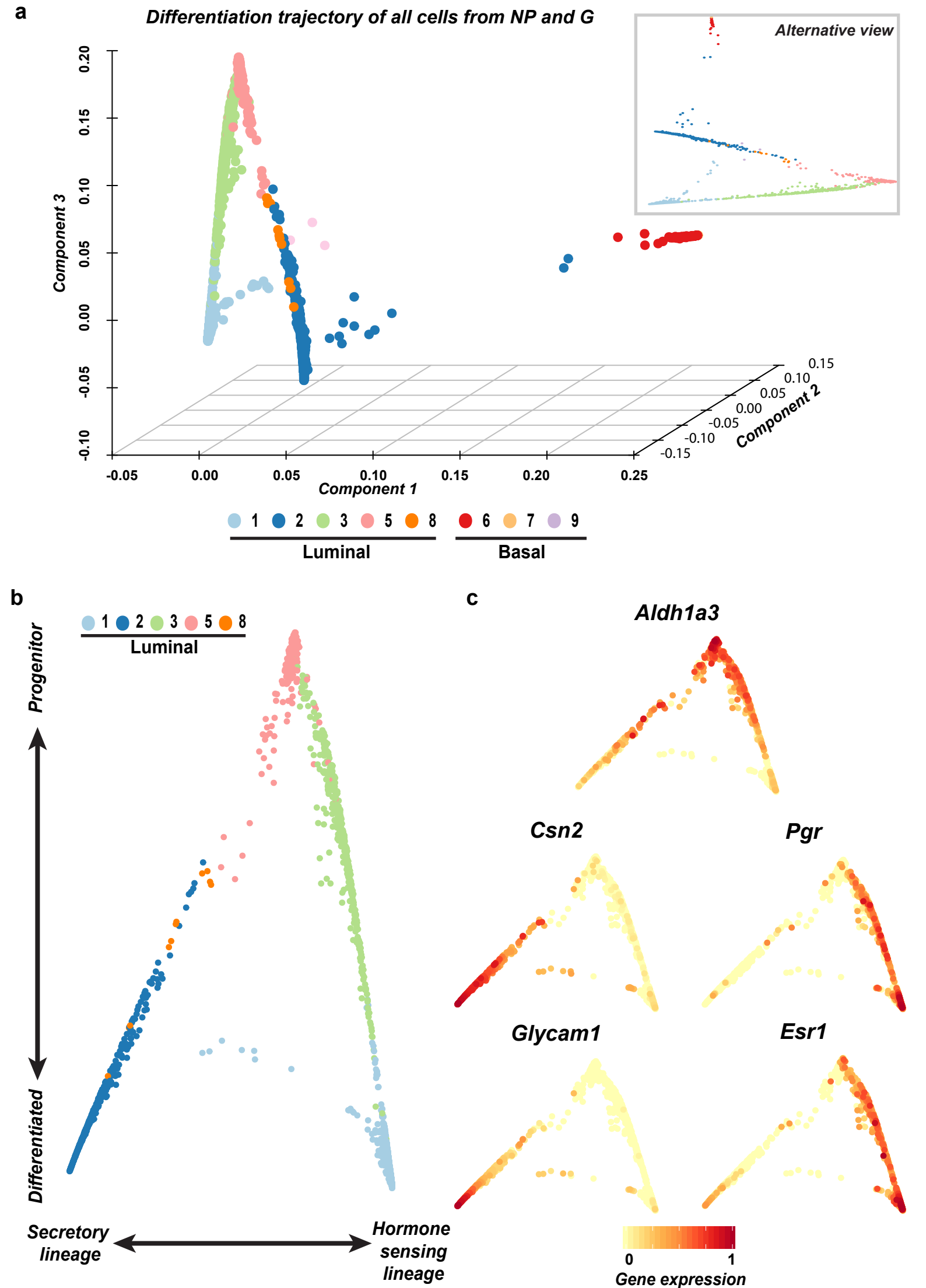
c

t-SNE coloured by cluster

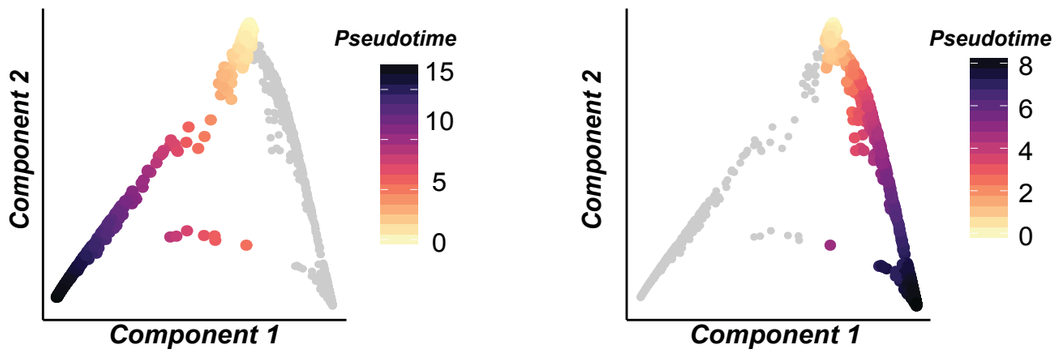


d

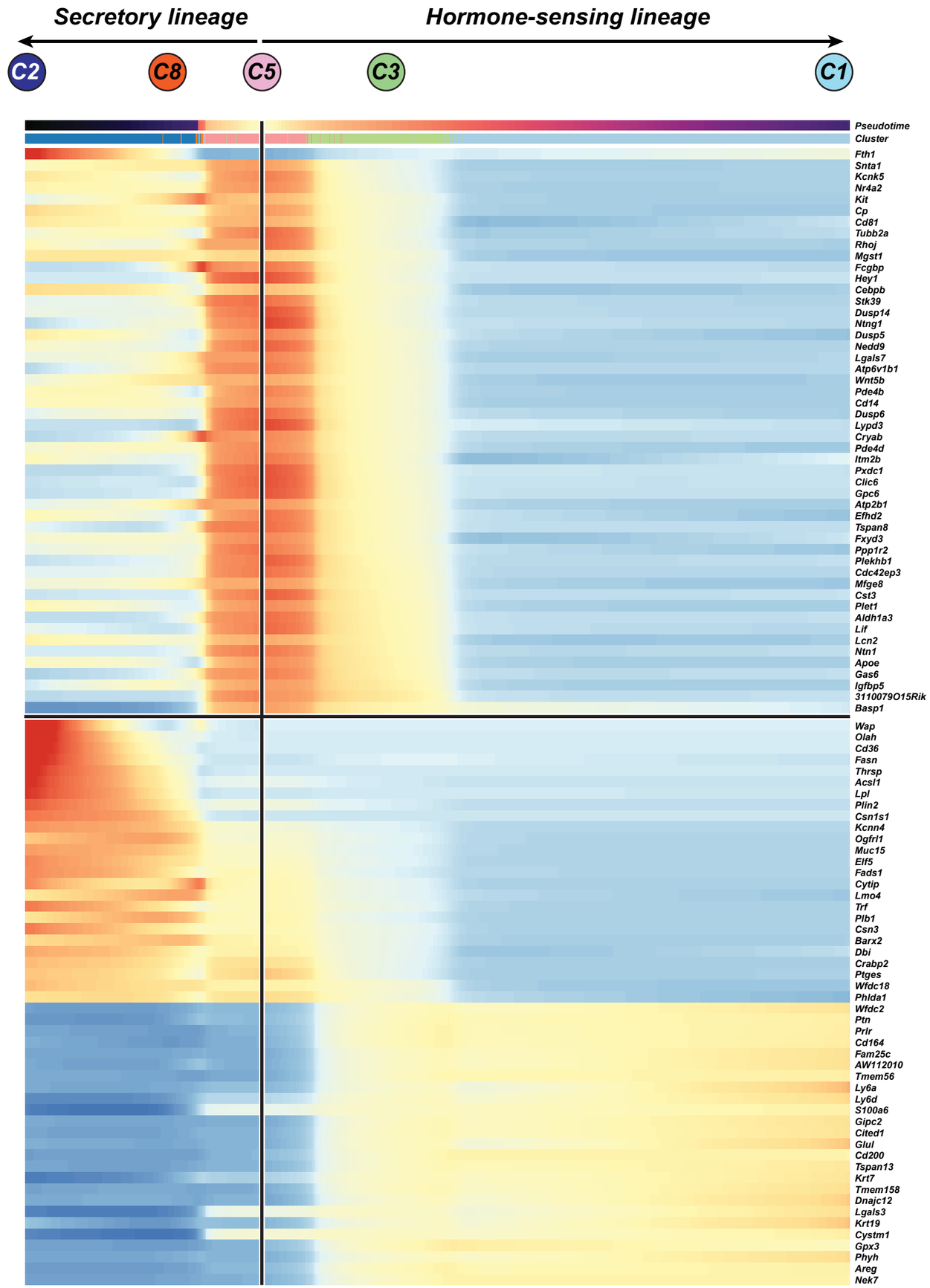



a
Differentiation trajectory of all cells from NP and G
Alternative view
Component 3
Component 1
Component 2
1 2 3 5 8
Luminal
6 7 9
Basal
b
Progenitor
Differentiated
Secretory lineage
Hormone sensing lineage
c
Aldh1a3
Csn2
Pgr
Glycam1
Esr1
0
1
Gene expression

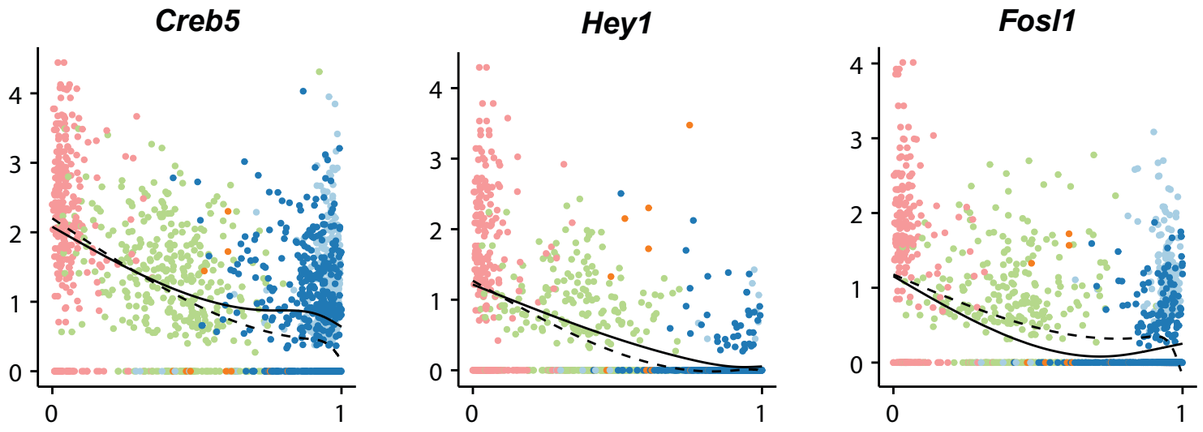
a



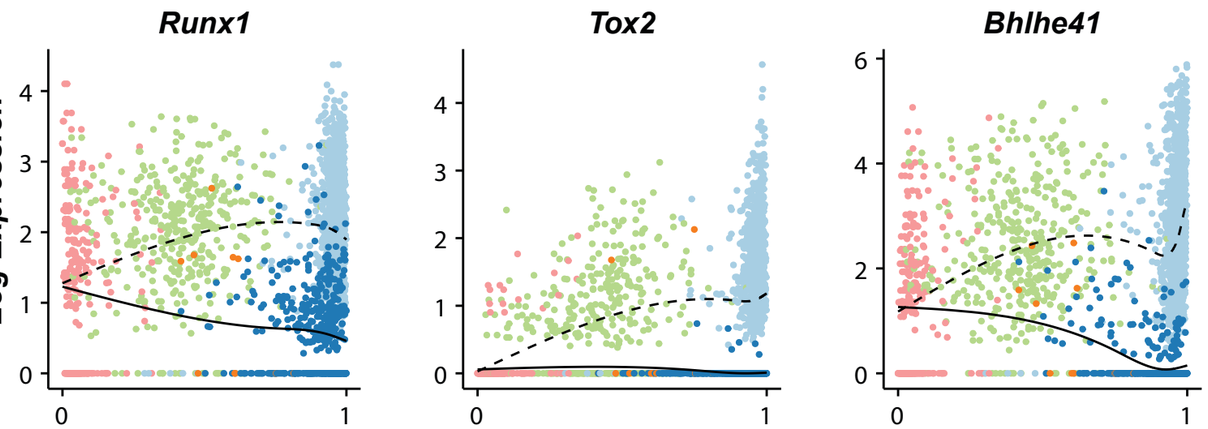
b



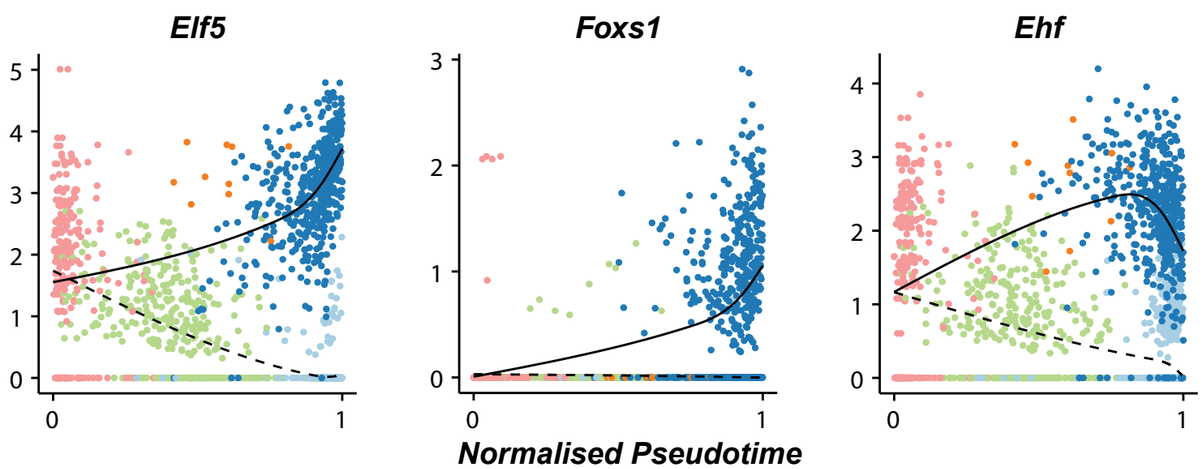
c



d

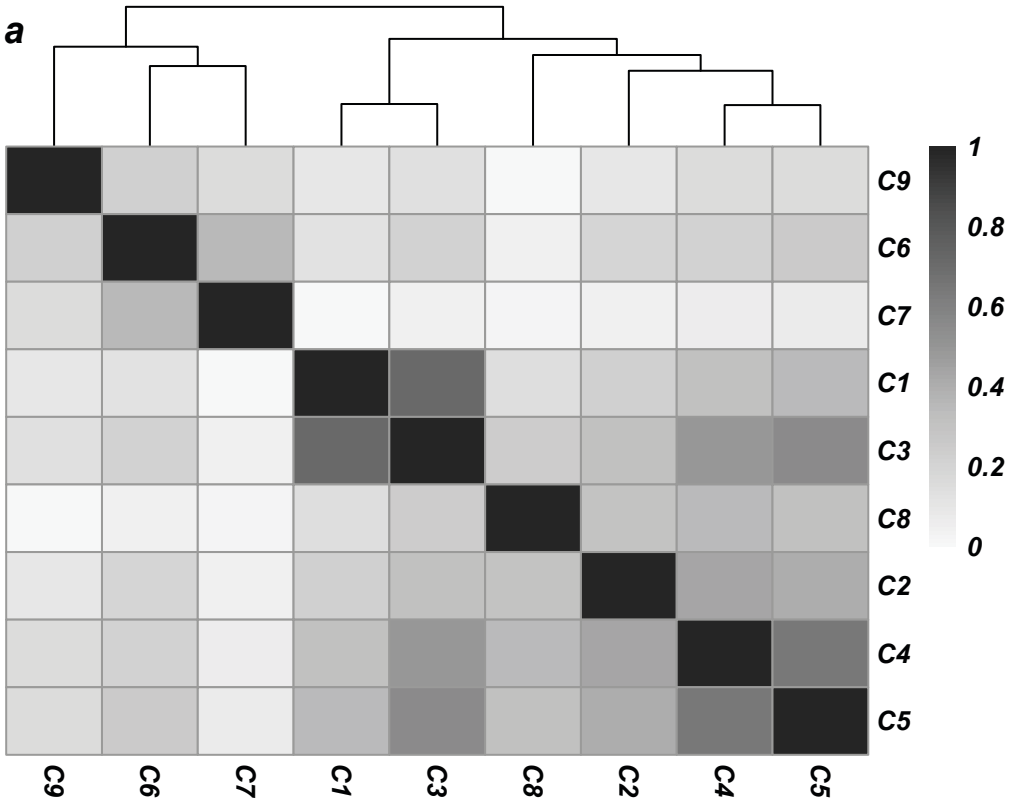


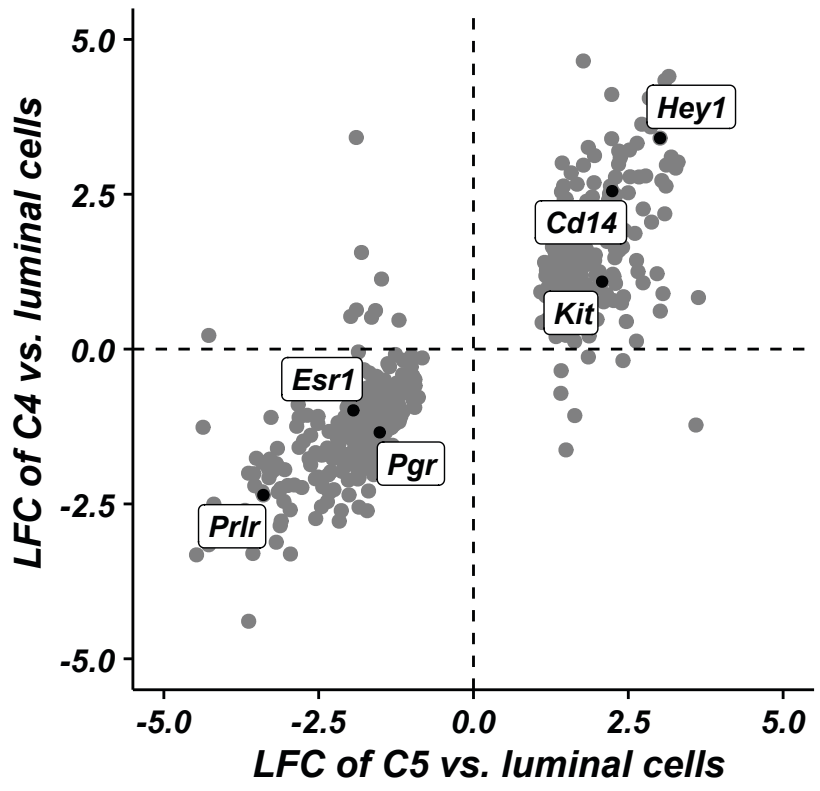
e

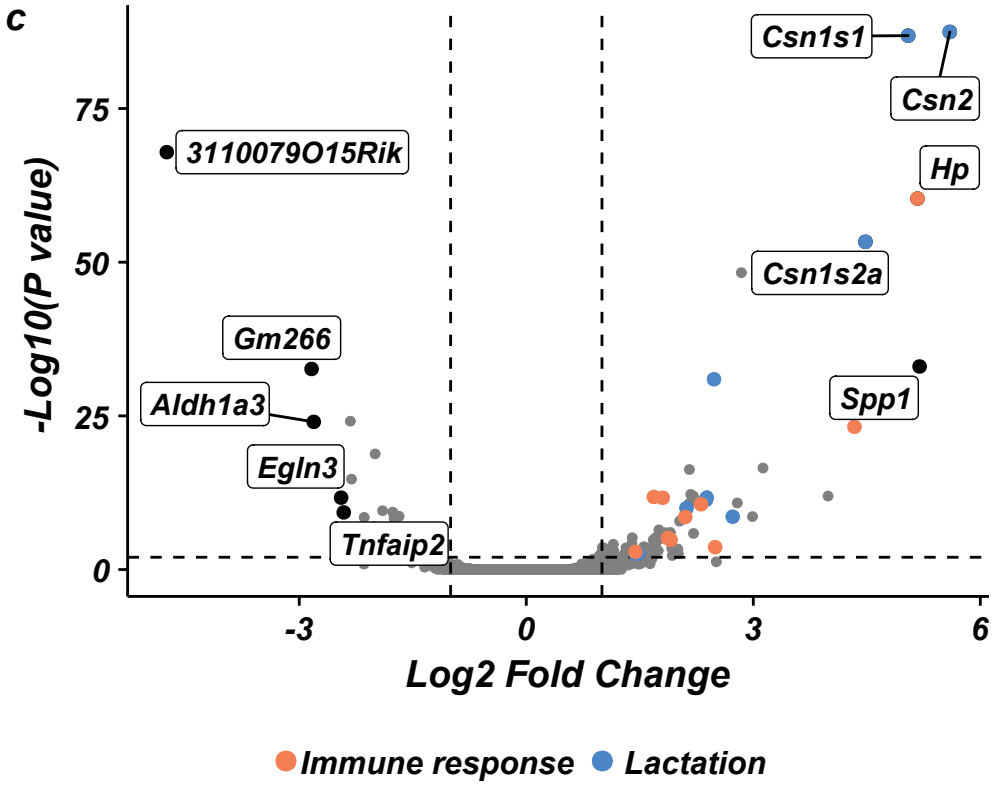


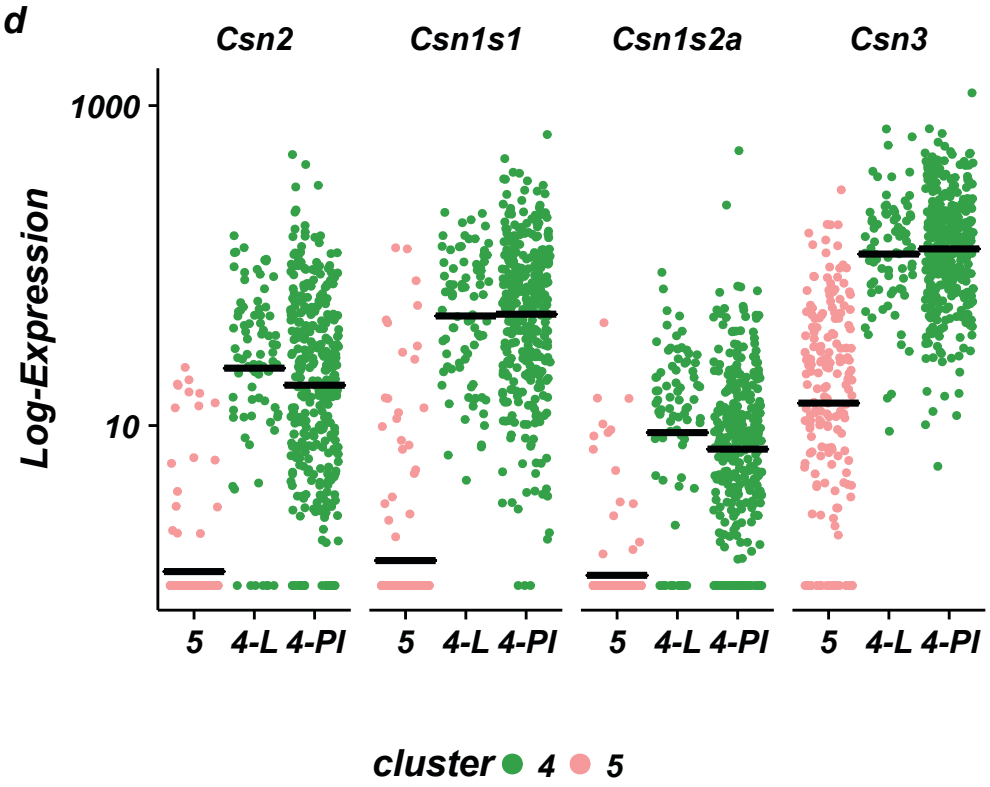
Hormone-sensing lineage — Secretory lineage

1 2 3 5 8


a
C9
C6
C7
C1
C3
C8
C2
C4
C5
1
0.8
0.6
0.4
0.2
0


b
LFC of C4 vs. luminal cells
LFC of C5 vs. luminal cells
5.0
2.5
0.0
-2.5
-5.0
Hey1
Cd14
Kit
Esr1
Pgr
Prlr

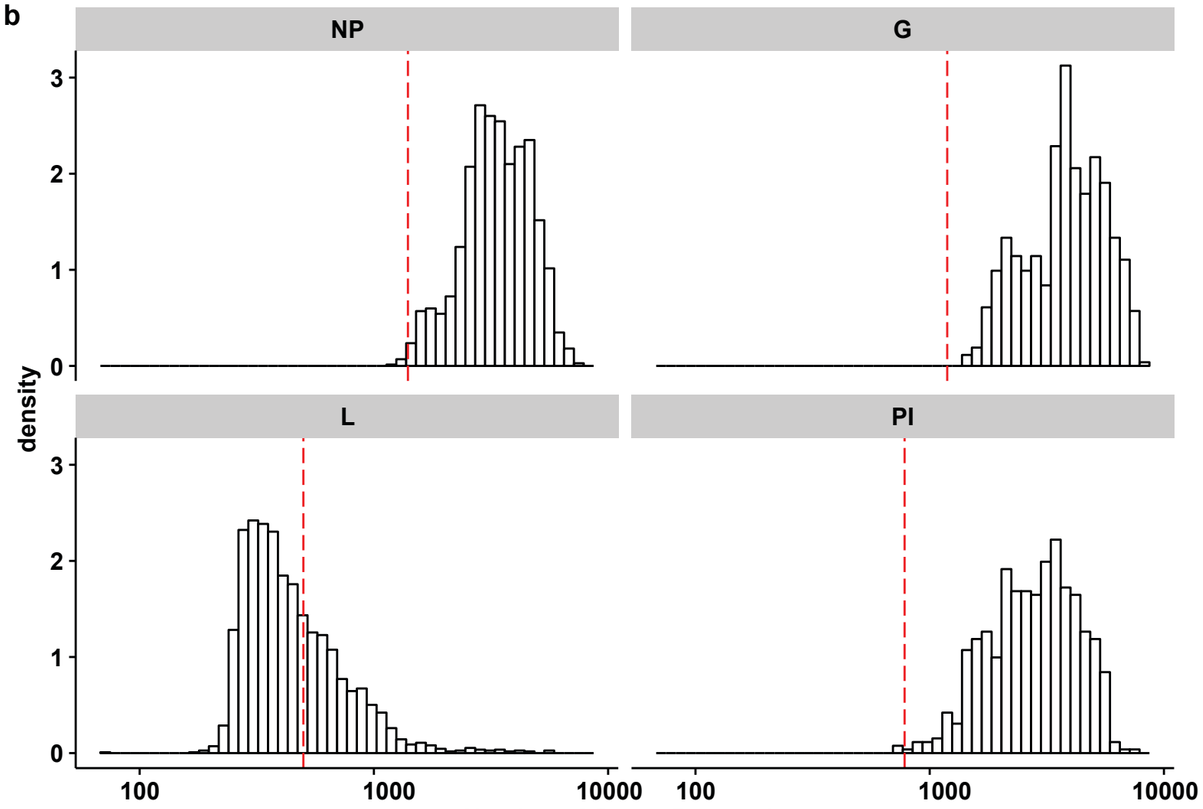

c
-Log10(P value)
Log2 Fold Change
75
50
25
0
-3
0
3
6
Csn1s1
Csn2
Hp
Csn1s2a
Spp1
3110079O15Rik
Gm266
Aldh1a3
Egln3
Tnfaip2
Immune response
Lactation


d
Csn2
Csn1s1
Csn1s2a
Csn3
Log-Expression
1000
10
5
4-L
4-Pl
cluster
4
5

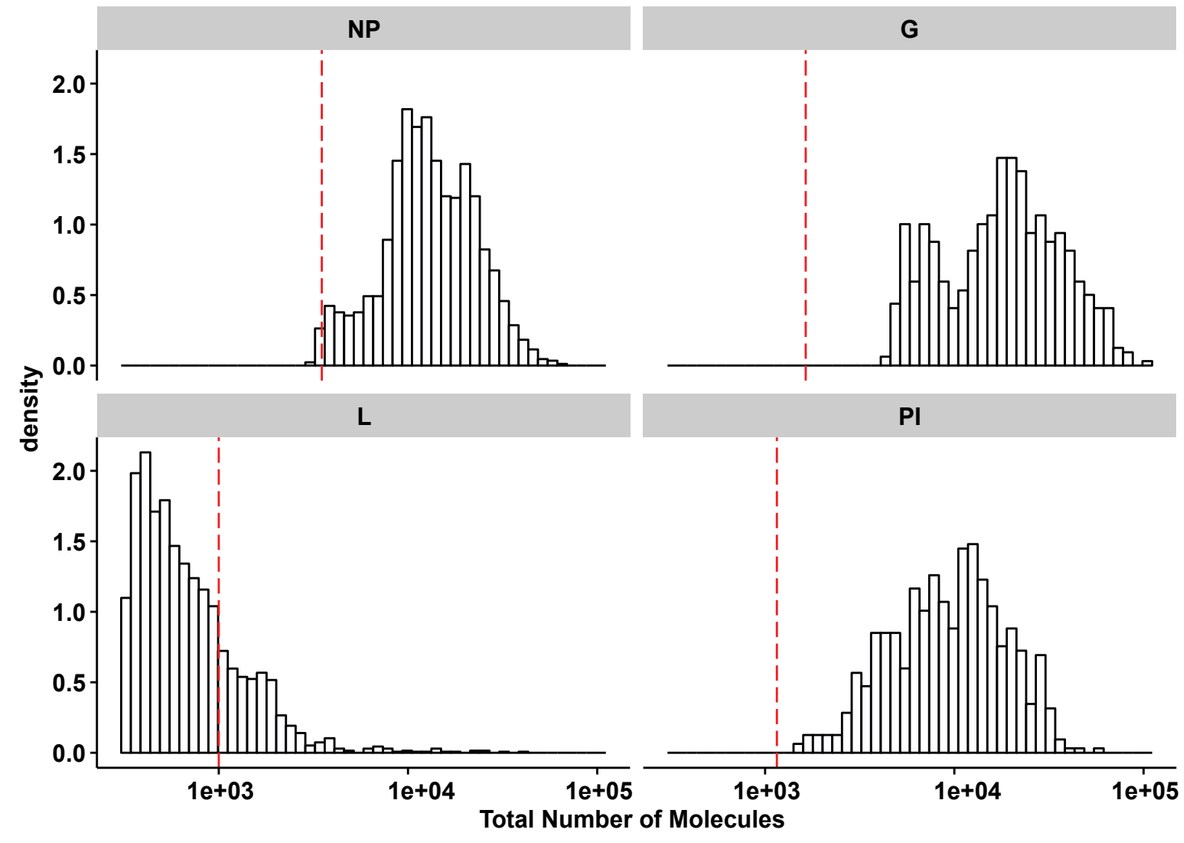
a

Sample	Number of cells	Unique molecules	Genes Detected	Number of reads	Sequencing Saturation
NP1	596	12778.5	3326.0	257639	86.3
NP2	1111	12640.0	3358.0	124686	75.3
G1	248	20043.5	4106.5	574902	91.4
G2	375	18068.0	3754.0	372905	88.3
L1	1312	620.5	414.0	106789	97.7
L2	1336	571.5	389.5	111926	98.0
PI1	469	9329.0	2773.0	264971	90.2
PI2	151	10473.0	3061.0	914353	97.2

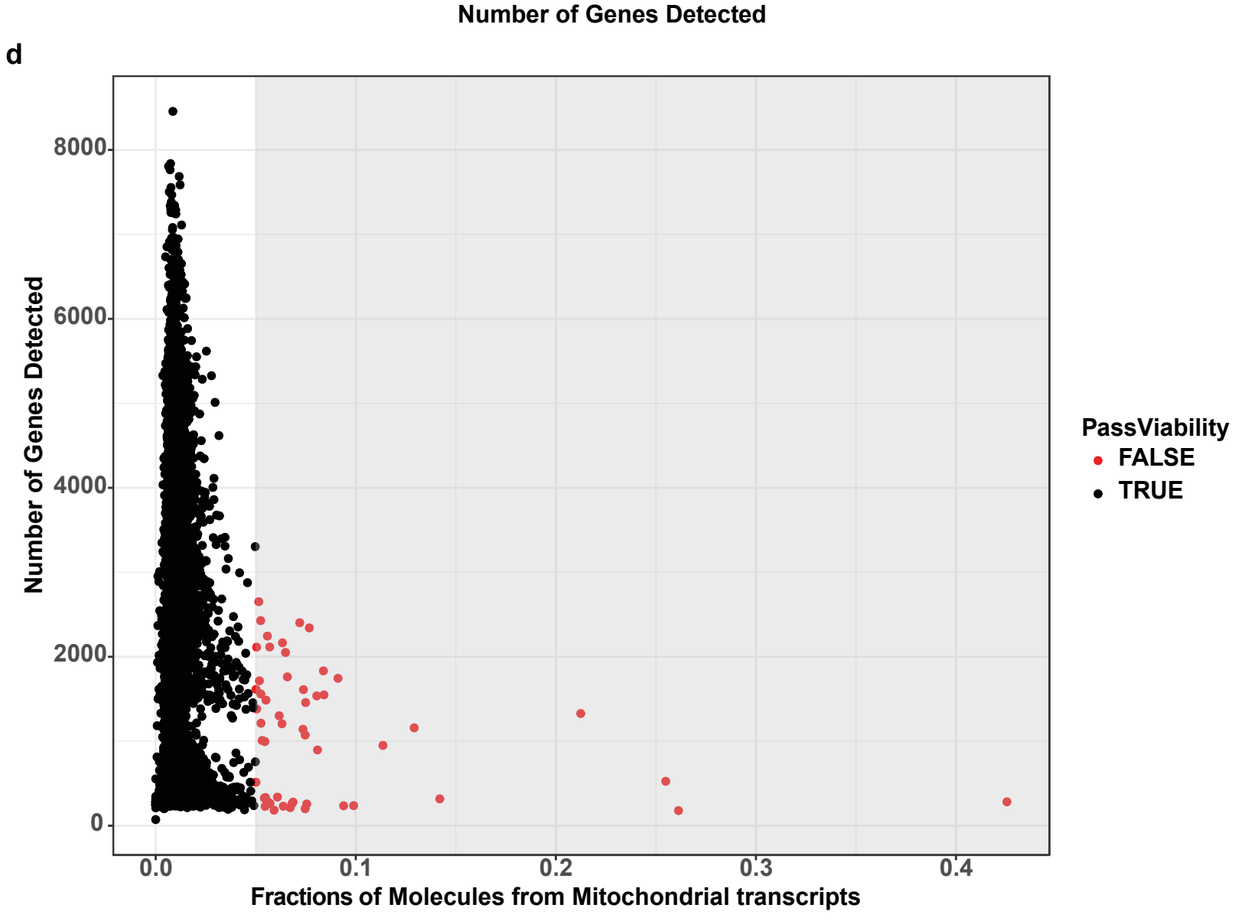
b

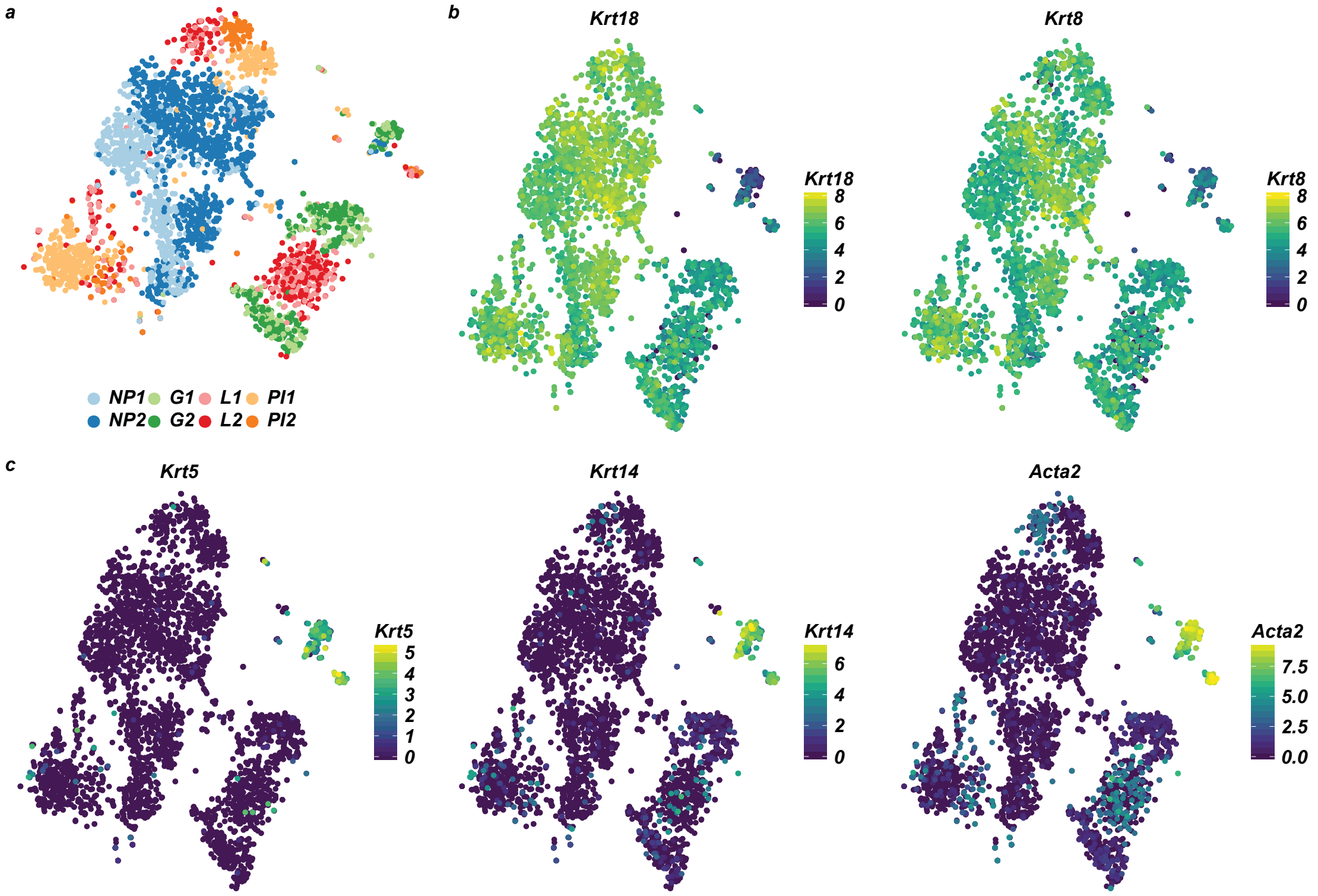


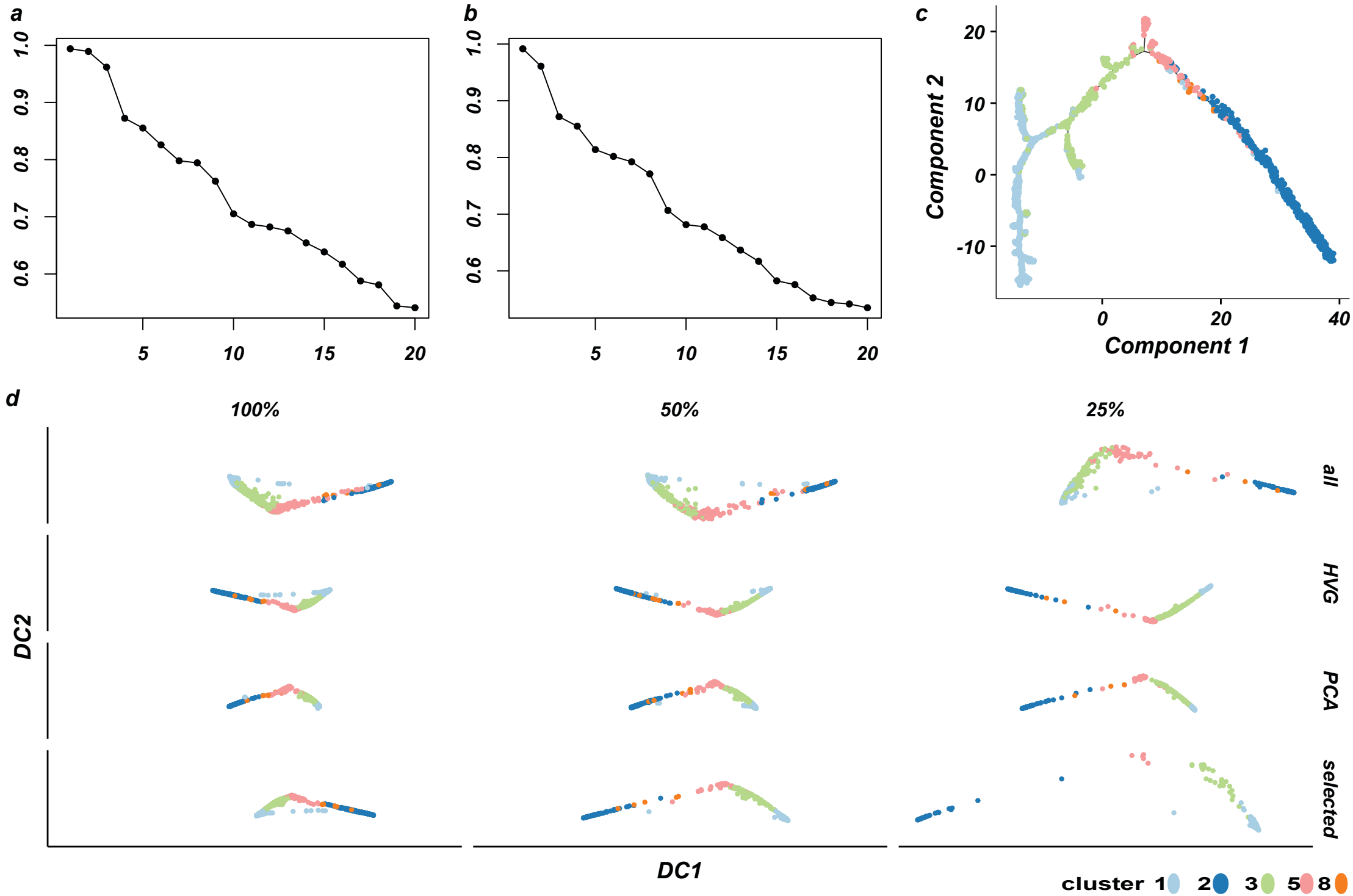
c



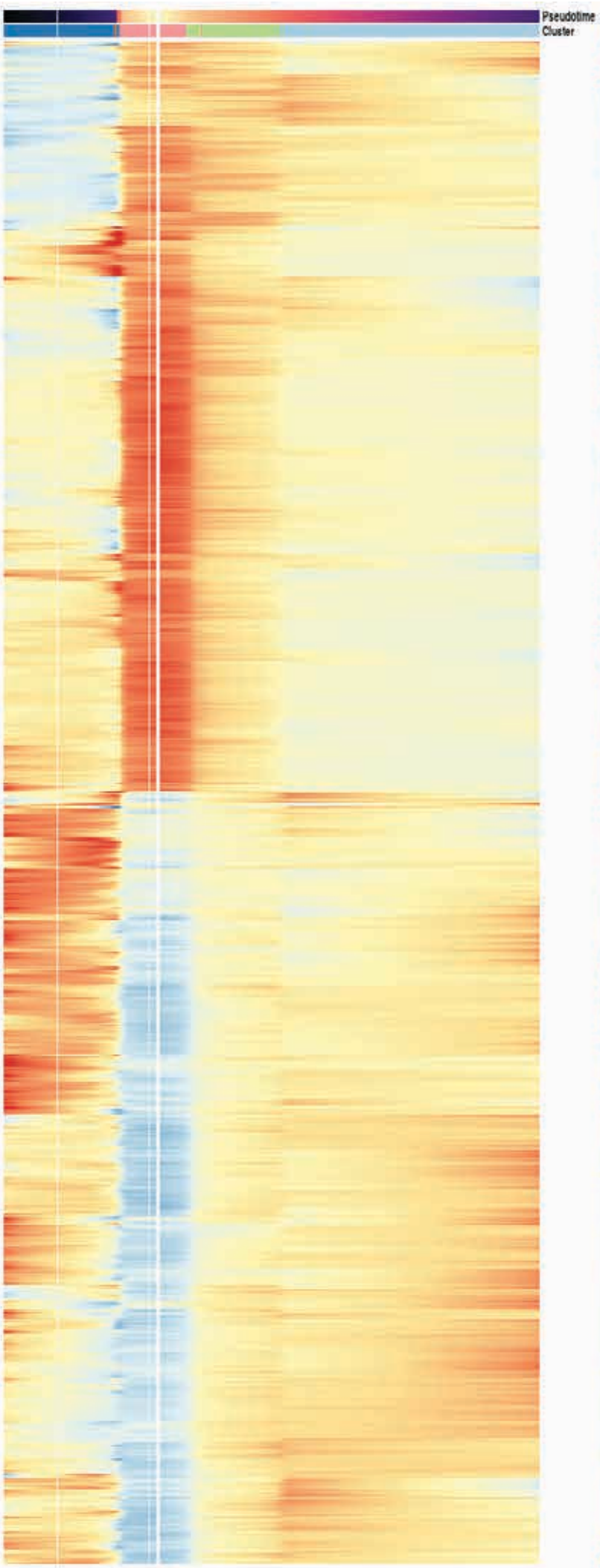
d



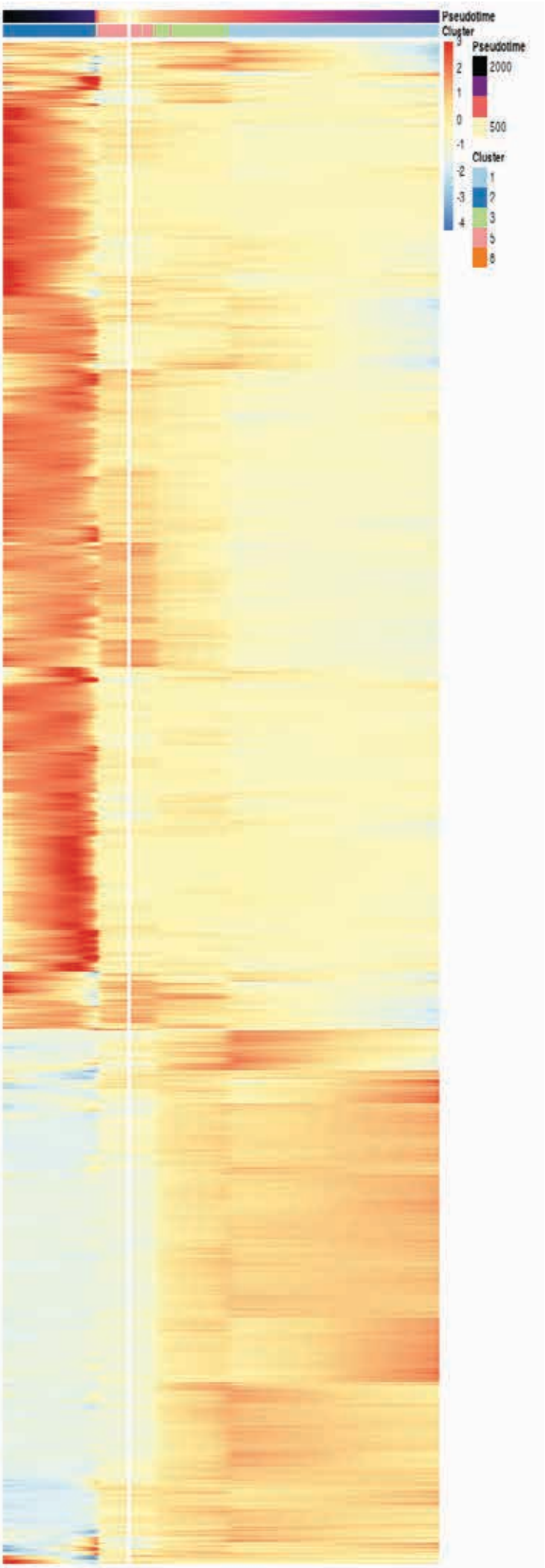

a
NP1
G1
L1
Pl1
NP2
G2
L2
Pl2
b
Krt18
Krt18
8
6
4
2
0
Krt8
Krt8
8
6
4
2
0
c
Krt5
Krt5
5
4
3
2
1
0
Krt14
Krt14
6
4
2
0
Acta2
Acta2
7.5
5.0
2.5
0.0

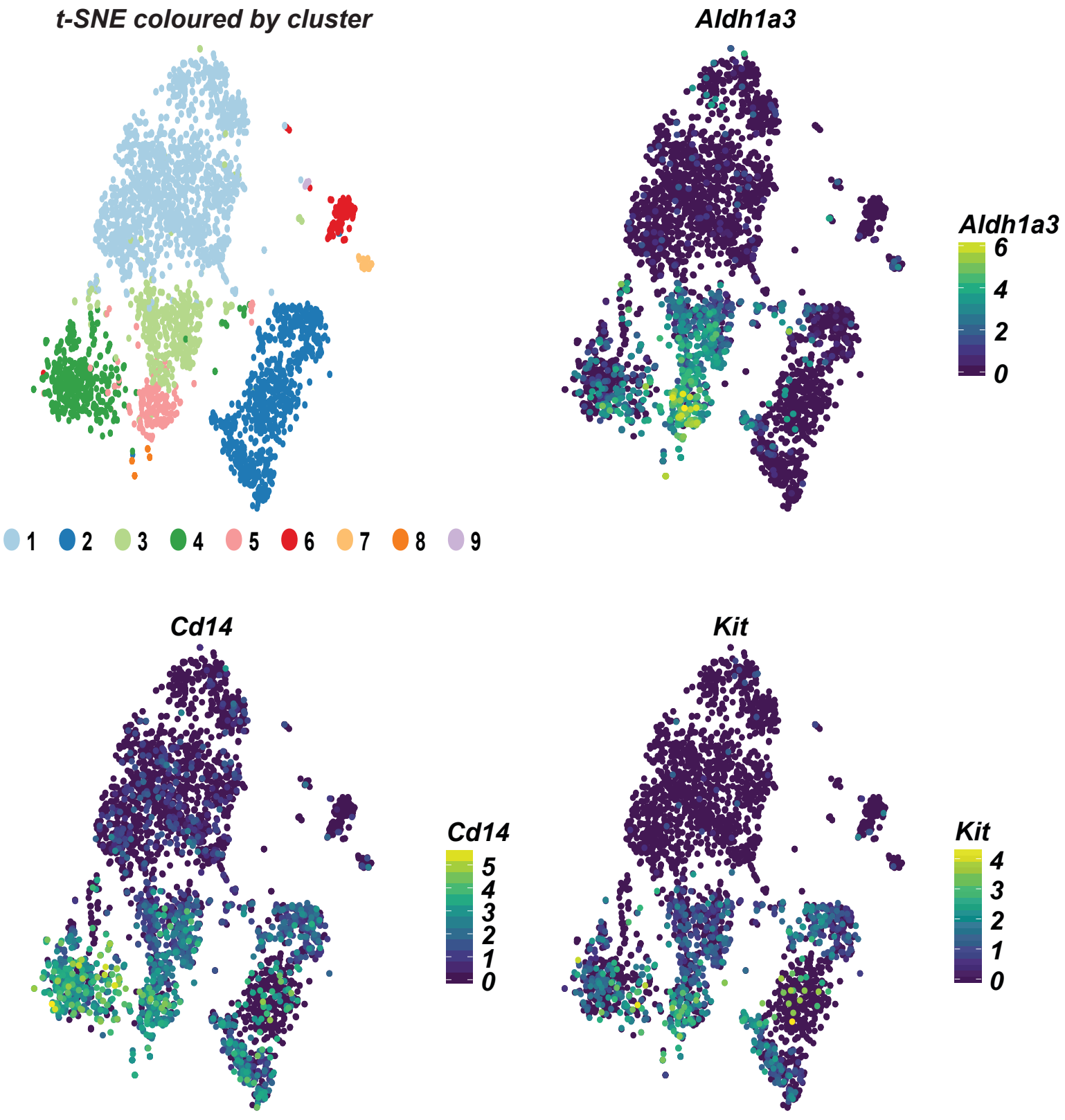

a
b
c
d
1.0
0.9
0.8
0.7
0.6
5
10
15
20
Component 2
Component 1
100%
50%
25%
all
HVG
PCA
selected
DC2
DC1
cluster 1 2 3 5 8

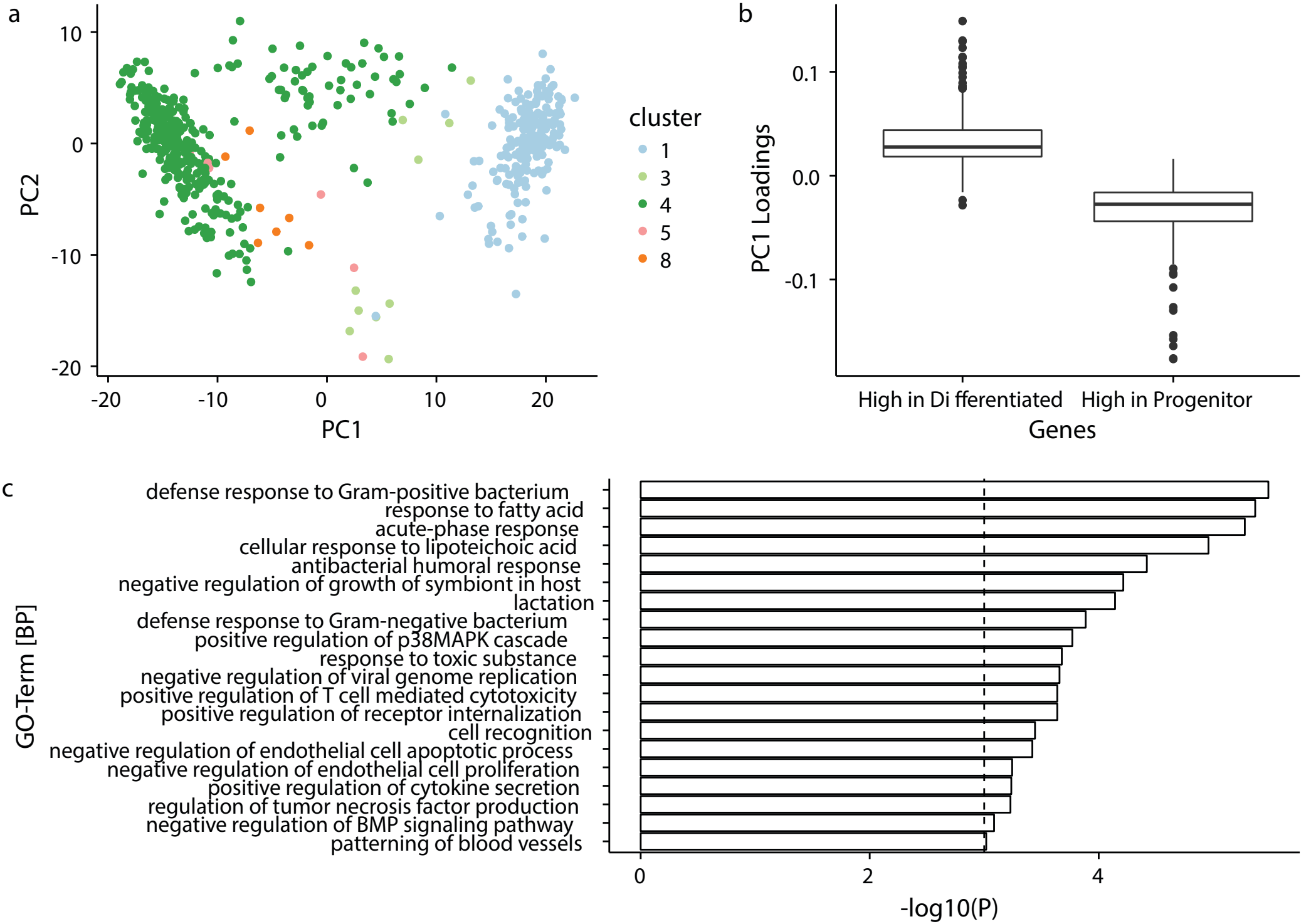
a

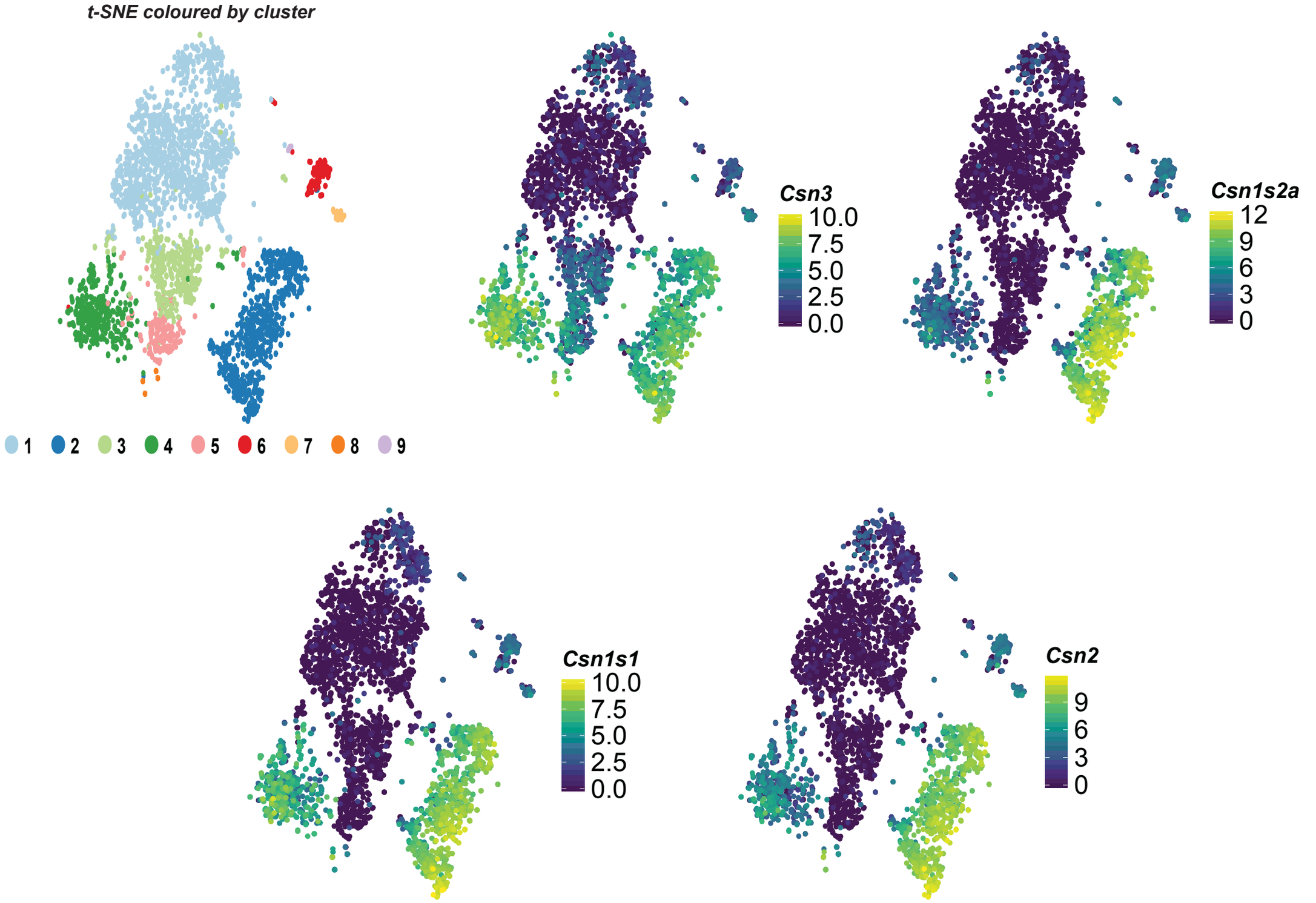

Pseudotime
Cluster

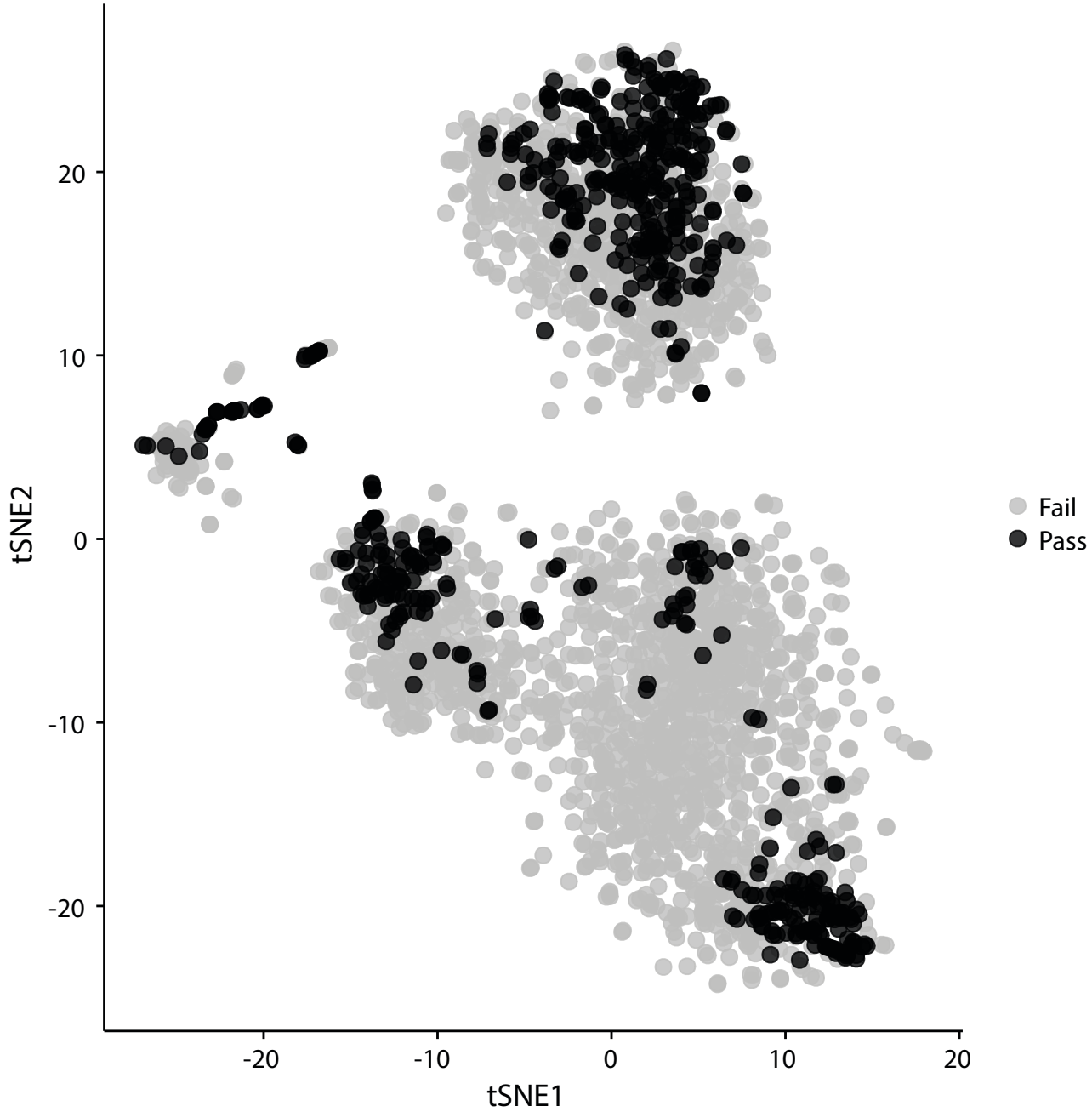
b

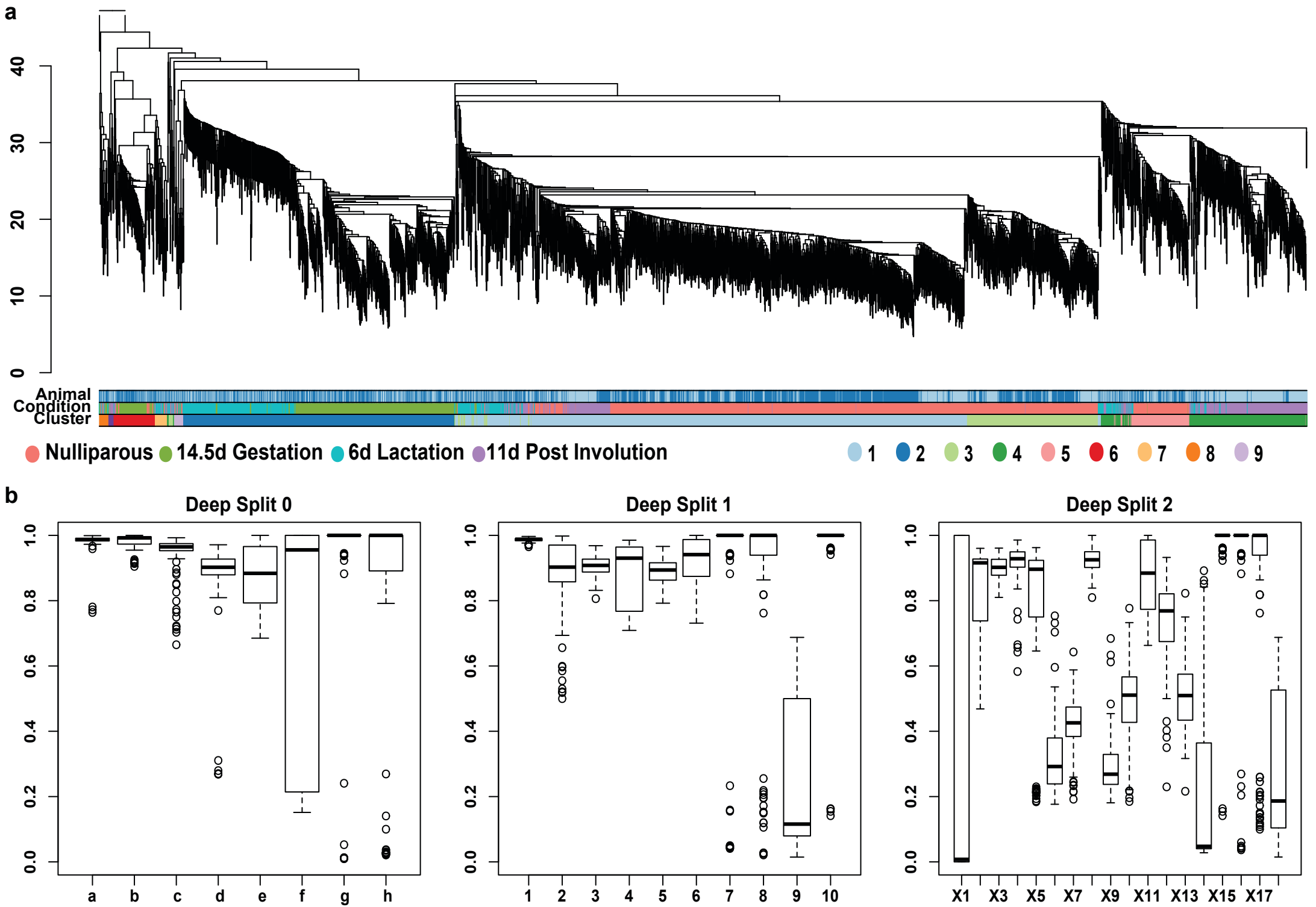

Pseudotime
Cluster
3
2
1
0
-1
-2
-3
-4
Pseudotime
2000
500
Cluster
1
2
3
5
6

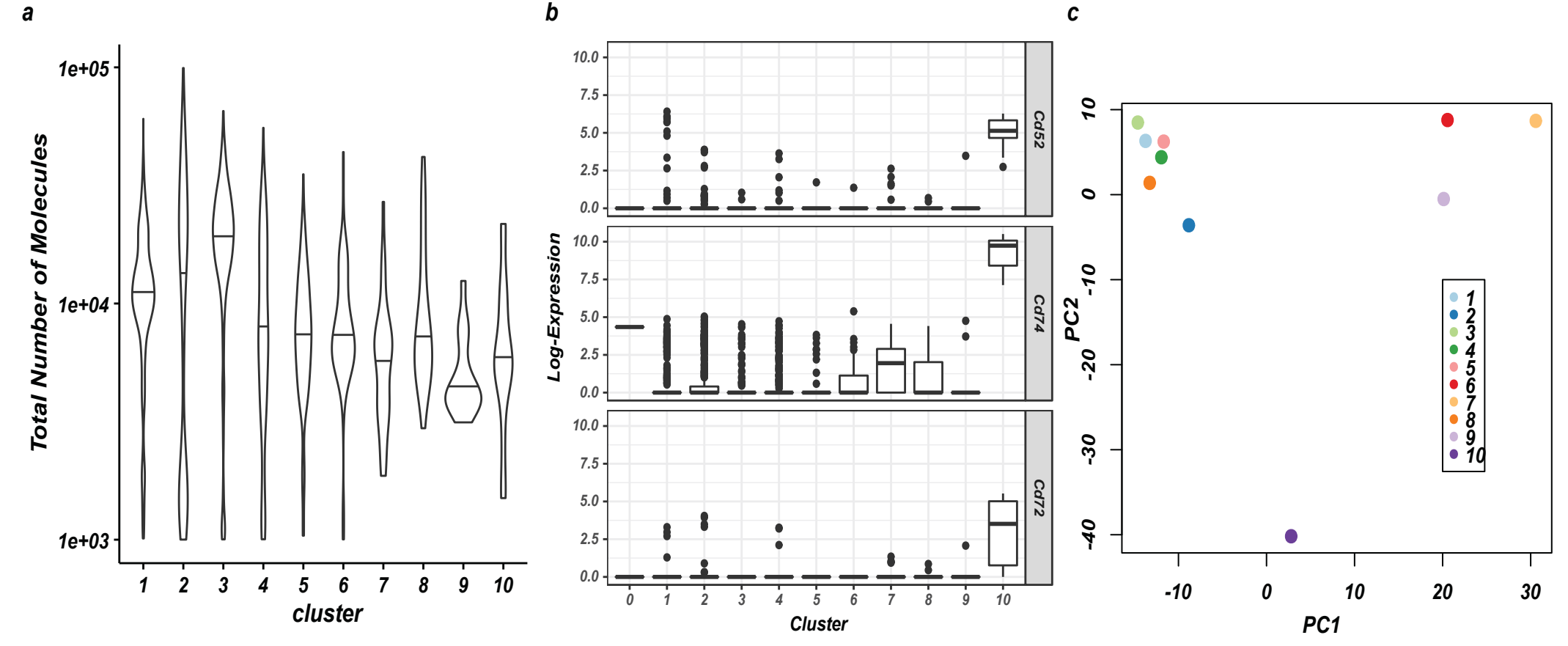

t-SNE coloured by cluster
1 2 3 4 5 6 7 8 9
Aldh1a3
Aldh1a3
6
4
2
0
Cd14
Cd14
5
4
3
2
1
0
Kit
Kit
4
3
2
1
0


a
PC2
PC1
cluster
1
3
4
5
8
b
PC1 Loadings
High in Di fferentiated
High in Progenitor
Genes
c
GO-Term [BP]
defense response to Gram-positive bacterium
response to fatty acid
acute-phase response
cellular response to lipoteichoic acid
antibacterial humoral response
negative regulation of growth of symbiont in host
lactation
defense response to Gram-negative bacterium
positive regulation of p38MAPK cascade
response to toxic substance
negative regulation of viral genome replication
positive regulation of T cell mediated cytotoxicity
positive regulation of receptor internalization
cell recognition
negative regulation of endothelial cell apoptotic process
negative regulation of endothelial cell proliferation
positive regulation of cytokine secretion
regulation of tumor necrosis factor production
negative regulation of BMP signaling pathway
patterning of blood vessels
-log10(P)


t-SNE coloured by cluster
1
2
3
4
5
6
7
8
9
Csn3
10.0
7.5
5.0
2.5
0.0
Csn1s2a
12
9
6
3
0
Csn1s1
10.0
7.5
5.0
2.5
0.0
Csn2
9
6
3
0


20
10
0
-10
-20
tSNE2
-20
-10
0
10
20
tSNE1
Fail
Pass


a
0
10
20
30
40
Animal
Condition
Cluster
Nulliparous
14.5d Gestation
6d Lactation
11d Post Involution
1
2
3
4
5
6
7
8
9
b
Deep Split 0
Deep Split 1
Deep Split 2
0.0
0.2
0.4
0.6
0.8
1.0
a
b
c
d
e
f
g
h
1
2
3
4
5
6
7
8
9
10
X1
X3
X5
X7
X9
X11
X13
X15
X17


a
Total Number of Molecules
1e+05
1e+04
1e+03
1 2 3 4 5 6 7 8 9 10
cluster
b
Log-Expression
Cd52
Cd74
Cd72
10.0
7.5
5.0
2.5
0.0
0 1 2 3 4 5 6 7 8 9 10
Cluster
c
PC2
PC1
10
0
-10
-20
-30
-40
-10 0 10 20 30
1
2
3
4
5
6
7
8
9
10

Cluster	Key genes expressed	Number of cells captured	Putative identity
1	*Esr1, Prlr, Pgr, S100a6, Cited1*	1103(NP), 12(G), 126(L), 211(PI)	*Differentiated hormone sensing luminal cells*
2	*Wap, Csn2, Glycam1*	480(G), 305(L)	*Secretory luminal cells*
3	*Esr1, Prlr, Pgr, Aldh1a3, Cd14, Kit, Ly6a (Sca-1)*	388(NP), 36(L), 12(PI)	*Hormone sensing luminal progenitor cells*
4	*Aldh1a3, Cd14, Kit*	83(L), 339(PI)	*Post-parity luminal progenitor cell*
5	*Aldh1a3, Cd14, Kit*	161(NP), 2(G), 13(L), 6(PI)	*Luminal progenitor cells*
6	*Krt4, Krt14, Pdpn, Etv5, Acta2*	18(NP), 94(G), 4(L), 4(PI)	*Basal progenitor cells*
7	*Oxtr, Acta2, Krt4, Krt14*	3(NP), 2(G), 20(L), 10(PI)	*Myoepithelial cells*
8	*Elf5, Aldh1a3, Kit*	6(NP), 5(G), 8 (L), 7 (PI)	*Luminal progenitor cell with bias towards the alveolar fate*
9	*Procr, Igfbp4, Gng11, and Zeb2*	3(G), 12(PI)	*Procr+ basal cells*