LETTER

doi:10.1038/nature13173

Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq

Barbara Treutlein[1]*, Doug G. Brownfield[2]*, Angela R. Wu[1], Norma F. Neff[1], Gary L. Mantalas[1], F. Hernan Espinoza[2], Tushar J. Desai[3], Mark A. Krasnow[2] & Stephen R. Quake[1]

The mammalian lung is a highly branched network in which the distal regions of the bronchial tree transform during development into a densely packed honeycomb of alveolar air sacs that mediate gas exchange. Although this transformation has been studied by marker expression analysis and fate-mapping, the mechanisms that control the progression of lung progenitors along distinct lineages into mature alveolar cell types are still incompletely known, in part because of the limited number of lineage markers[1–3] and the effects of ensemble averaging in conventional transcriptome analysis experiments on cell populations[1–5]. Here we show that single-cell transcriptome analysis circumvents these problems and enables direct measurement of the various cell types and hierarchies in the developing lung. We used microfluidic single-cell RNA sequencing (RNA-seq) on 198 individual cells at four different stages encompassing alveolar differentiation to measure the transcriptional states which define the developmental and cellular hierarchy of the distal mouse lung epithelium. We empirically classified cells into distinct groups by using an unbiased genome-wide approach that did not require a priori knowledge of the underlying cell types or the previous purification of cell populations. The results confirmed the basic outlines of the classical model of epithelial cell-type diversity in the distal lung and led to the discovery of many previously unknown cell-type markers, including transcriptional regulators that discriminate between the different populations. We reconstructed the molecular steps during maturation of bipotential progenitors along both alveolar lineages and elucidated the full life cycle of the alveolar type 2 cell lineage. This single-cell genomics approach is applicable to any developing or mature tissue to robustly delineate molecularly distinct cell types, define progenitors and lineage hierarchies, and identify lineage-specific regulatory factors.

In mice, alveolar epithelial cells differentiate between embryonic days (E)16.5 and 18.5: distal airway tips expand into sac-like configurations ('sacculation') as a morphologically uniform population of columnar progenitors proceeds towards the fate of either flat alveolar type 1 (AT1) cells specialized for gas exchange or surfactant-secreting cuboidal alveolar type 2 (AT2) cells (Extended Data Fig. 1). At each time point during sacculation, progenitors, intermediates and recently differentiated cells coexist (Fig. 1a)[6]. To resolve the cellular composition of the developing bronchio-alveolar epithelium, we initially sequenced transcriptomes of 80 individual live cells of the developing mouse lung epithelium late in sacculation (E18.5; three biological replicates). Single-cell suspensions of micro-dissected distal lung regions were purified by magnetic-activated cell sorting (MACS) to deplete leukocytes and alveolar macrophages and enrich for epithelial cells ($CD45^-$ / $EpCAM^+$) (Extended Data Fig. 2). An automated microfluidic platform was used to capture and lyse individual epithelial cells, reverse transcribe RNA and amplify complementary DNA.

RNA-seq libraries from the amplification products of single cells as well as bulk control samples were sequenced to a depth of (2–5) × 10^6 reads per library (Methods). Saturation analysis confirmed that this sequencing depth is sufficient to detect most genes expressed by single cells (Extended Data Fig. 3a). Technical noise and dynamic range were

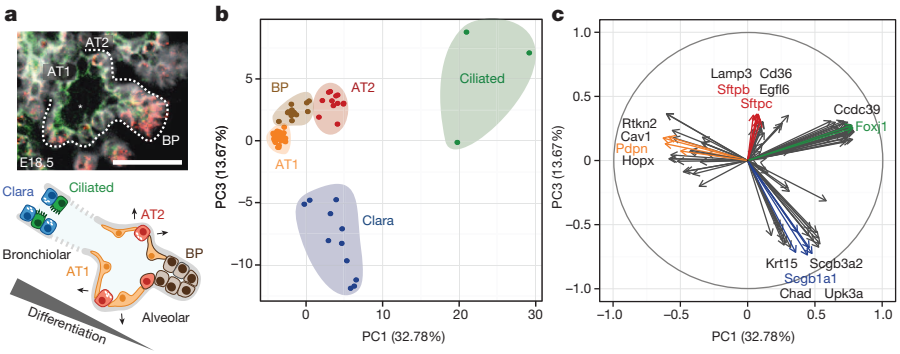


Figure 1 | Single-cell RNA-seq of 80 embryonic (E18.5) mouse lung epithelial cells enables unbiased identification of alveolar, bronchiolar and progenitor cell populations. **a**, Spatially heterogeneous differentiation of distal lung epithelium. The micrograph of a newly forming alveolar sac (asterisk) and the diagram below illustrate cell types and the gradient of developmental intermediates comprising the distal lung epithelium during sacculation (E18.5). Micrograph: green, Pdpn, alveolar type 1 (AT1) marker; red, Sftpc, AT2 marker; white, E-cadherin, pan-epithelial marker. BPs are characterized by co-expression of some AT1 and AT2 markers. In the diagram, BPs (brown) persist at the tip, and nascent AT2 (red) and AT1 (orange) cells are located more proximally. Ciliated (green) and Clara (blue) cells are located in the bronchiolar epithelium (not labelled in the micrograph). Scale bar, 75 μm. **b**, PCA of 80 single-cell transcriptomes (three biological replicates) at E18.5 distinguishes between major bronchiolar and alveolar cell lineages. PC, principal component. **c**, Distinct gene groups characterize each cell population on the basis of differential correlation with PC1 and PC3. The arrow tip denotes the correlation coefficient of the respective gene with each principal component.

[1]Departments of Bioengineering and Applied Physics, Stanford University School of Medicine and Howard Hughes Medical Institute, Stanford, California 94305, USA. [2]Department of Biochemistry, Stanford University School of Medicine and Howard Hughes Medical Institute, Stanford, California 94305, USA. [3]Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, Stanford University School of Medicine, Stanford, California 94305, USA.
*These authors contributed equally to this work.

assessed by using RNA control spike-in standards and by comparing single cells with the bulk samples (Extended Data Fig. 3b–e). The results are consistent with previous data from our group[7] and others[8–20]; we obtained single-transcript sensitivity and high ($\sim 10^5$) dynamic range. Comparison of three biological replicate experiments showed that median expression of all genes across single cells was strongly correlated ($r = 0.91$ and $r = 0.92$; Extended Data Fig. 3f, g).

We performed principal component analysis (PCA) on all 80 single-cell transcriptomes by using genes expressed in more than two cells and with a non-zero variance (8,578 genes). Genes with highest loadings in the first four principal components were analysed by unsupervised hierarchical clustering as well as PCA (Fig. 1b, c, Fig. 2a and Supplementary Data). This unbiased approach detected five different cell populations and four different gene families, which permutation analysis showed to be highly significant (Methods). Using known marker genes within the different clusters, we were able to associate cells with four previously reported epithelial cell types (Clara (*Scgb1a1*), ciliated (*Foxj1*), AT1 (*Pdpn*, *Ager*) and AT2 (*Sftpc*, *Sftpb*) cells). The fifth group was characterized by co-expression of AT1 and AT2 marker genes and was located on the PCA plot between the populations of AT1 and AT2 cells, suggesting either an intermediate population undergoing a transition between the two alveolar lineages or a population of bipotential alveolar progenitors. As discussed below, transcriptional profiles of distal lung epithelial cells at E16.5 implicate this fifth population as alveolar bipotential progenitor (BP) cells[6]. We validated these findings in two biological replicates of pooled E18.5 lungs by microfluidic single-cell quantitative PCR (qPCR) experiments: hierarchical clustering of ten known alveolar and bronchiolar marker genes identified the same five populations (Extended Data Fig. 5a–d). Together, these results show that single-cell RNA-seq enables the identification and molecular characterization of cell types and developmental intermediates retrospectively without the need to first purify populations of interest.

In addition to classifying the epithelial cell populations in the distal lung at E18.5, our analysis identified sets of genes specific to each population, providing a battery of previously unknown markers that can be used to distinguish cells from each alveolar and bronchiolar lineage. We used Guilt-by-Association and correlation analysis to assess the significance of co-expression of genes in all cells belonging to a specific cell type (Methods, Fig. 2b and Supplementary Data). The large number of lineage-specific genes allowed us to annotate functions of individual cell types by gene ontology and pathway enrichment analysis[21] (Extended Data Fig. 4a and Supplementary Data): AT1 cells were enriched in pathways associated with extracellular matrix-receptor interaction, focal adhesion, tight and adherens junctions and regulation of the actin cytoskeleton; AT2 cells were enriched for adipocytokine and PPAR signalling and for lysosome pathways; the Clara cell lineage was enriched for metabolism of xenobiotics by cytochrome P450, drug metabolism and glutathione metabolism; and ciliated cells showed enrichment for progesterone-mediated oocyte maturation and cell cycle pathways. Furthermore, we identified transcription factors, receptors and ligands whose expression profile across all single cells was strongly correlated with the individual cell types (Extended Data Fig. 4b, c).

Among the numerous newly identified putative cell-type markers, several are of particular interest. *Hopx* transcription factor was previously reported to regulate alveolar maturation by suppressing surfactant protein production in AT2 cells[22]; our data show that *Hopx* is expressed in BPs, turns off in maturing AT2 cells and is maintained in AT1 cells. We validated the AT1-specific expression of *Hopx* by transgenic labelling and co-localization with two AT1 markers, *Pdpn* and *Ager* (Fig. 2c and Extended Data Fig. 4e). We also found that *Vegfa* endothelial growth factor is specifically expressed in the AT1 lineage, presumably serving as a signal to activate nearby capillary endothelial cells; AT1-specific expression was validated by single-cell qPCR (Extended Data Fig. 4d). *Egfl6*, encoding a protein implicated in cell adhesion and cell differentiation, is specifically expressed in AT2 cells; AT2-specific expression was confirmed by multiplex *in situ* hybridization with the canonical AT2

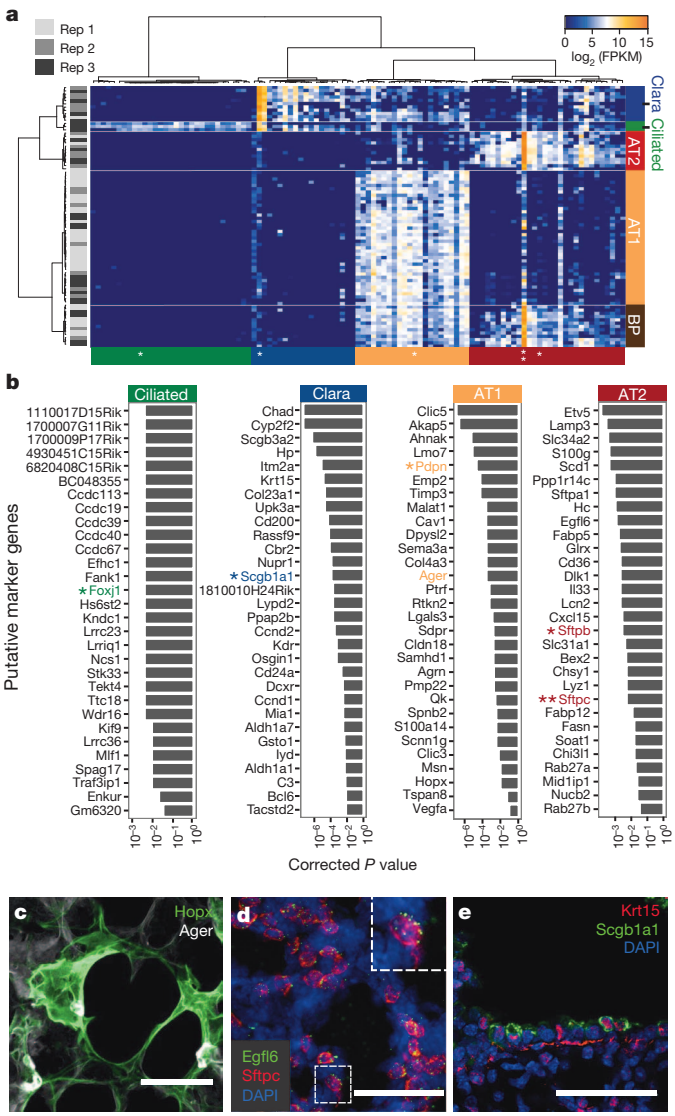


Figure 2 | Single-cell transcriptome analysis discovers previously unknown markers. **a**, Hierarchical clustering of RNA-seq data from 80 single distal lung epithelial cells (E18.5, three biological replicates) identifies five molecularly distinct populations, assigned to alveolar and bronchiolar lineages based on the presence of canonical marker genes (asterisks) within the respective gene clusters (AT2 (red), *Sftpb* and *Sftpc*; AT1 (orange). *Pdpn*; ciliated (green), *Foxj1*; Clara (blue), *Scgb1a1*). BPs (brown) co-express AT1 and AT2 markers. Each row represents a single cell, each column a gene (104 genes in total; Supplementary Data). Permutation analysis supports the significance of the presented clustering ($P = 2.89 \times 10^{-122}$; Methods). FPKM, fragments per kilobase of transcript per 10^6 mapped reads. **b**, Bar graphs showing the top 30 putative marker genes for each cell lineage inferred from the E18.5 single-cell transcriptomes as a function of the multiple testing corrected *P* value for each gene (Guilt-by-Association; Methods). Canonical markers are bold and coloured. **c**, Validation of *Hopx* expression in AT1 cells. A lung section from a transgenic *Hopx-Cre-ERT2*$^{+/-}$;*mTmG*$^{+/tg}$ adult mouse was co-stained for AT1 marker Ager. Maximum intensity projections of confocal *z* stacks show that AT1 cells expressing membrane-localized green fluorescent protein (GFP; green) also express Ager (white). Scale bar, 50 μm. **d**, Validation of *Egfl6* expression in AT2 cells. Multiplexed *in situ* hybridization of E18.5 lungs shows co-localization of probes targeting *Egfl6* (green) and AT2 marker *Sftpc* (red) messenger RNA. Inset, close-up of boxed region. Blue, 4′,6-diamidino-2-phenylindole (DAPI)-stained nuclei. Scale bar, 50 μm. **e**, Validation of Krt15 expression in Clara cells. Immunofluorescent staining of E18.5 lungs with the use of antibodies against Krt15 (red) and Clara cell marker Scgb1a1 (green). Blue, DAPI-stained nuclei. Krt15 is also expressed outside the epithelium. Scale bar, 50 μm.

marker *Sftpc* (Fig. 2d). Krt15, a component of intermediate filaments, was specifically expressed in the Clara cell lineage, which we validated by co-staining with the canonical Clara cell marker Scgb1a1 (Fig. 2e). Finally, we used single-cell multiplexed qPCR to validate the lineage-specific expression of six additional genes including *Itgb4* and *Top2a* for ciliated cells, *Cftr*, *Cebpa*, *Sftpd* and *Id2* for the AT2 lineage, and *Vegfa* for the AT1 lineage (Extended Data Fig. 4d). Most genes specifically expressed by the AT2 lineage at E18.5 were also detected by single-cell RNA-seq in mature AT2 cells of an adult mouse lung, whereas genes specific to AT1, Clara or ciliated cells were not expressed or were expressed only at a low level (Extended Data Fig. 4f). Thus, we identified a large number of new, and potentially more specific, markers for various biological processes and stages relevant to alveolar and bronchiolar maturation.

Identification of the progenitor and differentiated cell types at E18.5 prompted further investigation of developmental intermediates in the alveolar maturation pathway. Sacculation of distal airway tubules commences at E16.5, and the distal epithelium is dominated by alveolar progenitor cells at this time[6]. We therefore measured transcript levels of ten known marker genes in 107 single cells of the distal lung epithelium at both E16.5 (33 cells) and E18.5 (74 cells) with multiplexed single-cell qPCR (Extended Data Fig. 5a–d). The marker gene expression profile and PCA identified Clara and ciliated cells distinct from alveolar lineages at both E16.5 and E18.5, corroborating the earlier separation of bronchiolar from alveolar maturation pathways. However, gene expression of alveolar cells showed no segregation into AT1 and AT2 lineages at E16.5, because marker genes for both subpopulations were commonly expressed by all cells, whereas by E18.5 they had clearly separated. This is consistent with a recent temporo-spatial marker study suggesting that AT1 and AT2 lineages emerge from a common BP[6]. In addition to BPs and mature alveolar cells at E18.5, we observed cells in intermediate maturation stages on the basis of partial co-expression of AT1 and AT2 marker genes. We used the newly identified genes specific for each mature alveolar cell type to subclassify these intermediates and thereby reconstruct the molecular pathway of differentiation of BPs into AT1 and AT2 lineages, grouping the genes into early and late markers of either lineage (Fig. 3). We confirmed the presence of developmental intermediates showing heterogeneity in marker gene expression by immunofluorescence (Extended Data Fig. 5f–i).

The constructed hierarchy identified transcription factors, receptors and ligands showing expression changes that correlated with specific transitions in the maturation states of alveolar cells (Extended Data Fig. 5e). The transcription factors *Sox9* and *Cited2* were expressed in BP and AT2 cells, whereas *Hes1* was expressed in BP and AT1 cells. We did not detect any transcription factors that initiated expression exclusively in either of the maturing alveolar lineages, suggesting that lineage commitment involves the downregulation of factors active in alveolar progenitor cells rather than the *de novo* expression of a lineage-specific transcription factor. Ligands were expressed in either BP and AT2 (*Cxcl15*, *Cmtm8*) or BP and AT1 cells (*Sema3a*, *Tgfb*, *Vegfa*), and receptors were expressed in a BP (*Fzd2*), BP/AT2 (*Fgfr2*) or BP/AT1 (*Gprc5a*) pattern. These results show that our approach can be used to characterize transcriptional profiles of transient cellular intermediates during a dynamic maturation process within a complex tissue.

Finally, we explored temporal changes within the distal lung by sequencing additional single-cell transcriptomes before sacculation (E14.5; 45 progenitor cells), early in sacculation (E16.5; 27 progenitor cells) and long after sacculation (adult; 46 transgenically labelled AT2 cells). We performed unsupervised hierarchical clustering analysis of *Sftpc*-positive cells (124 cells), using genes with the highest principal-component loadings in a PCA analysis (Fig. 4a, b and Supplementary Data). Cells clustered in groups that were highly correlated with developmental stage of cell isolation, in sequence from early progenitors (EPs), through BP and nascent AT1 and AT2 cells, to mature AT2 cells. Thus, AT2 cell maturation occurs in a progressive manner through transcriptionally distinct intermediates that can be robustly discriminated by expression

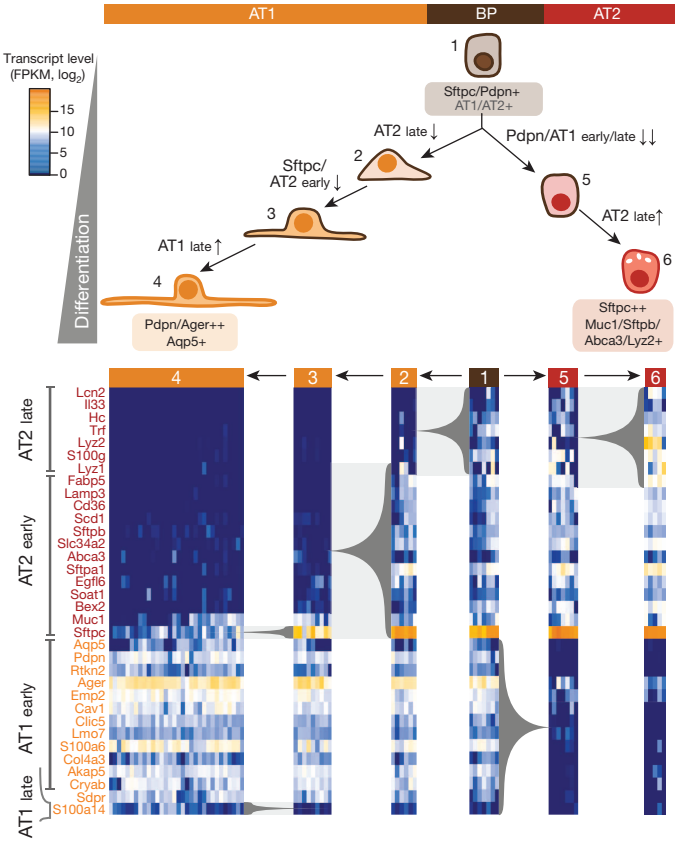


Figure 3 | Molecular profiles distinguish between developmental intermediates during the differentiation of AT1 and AT2 cells from a common BP. Developmental sequence of AT1 (orange) and AT2 (red) specification from a common BP (brown). Two and three maturation intermediates were identified in the specification processes of AT2 and AT1 cell types, respectively, on the basis of the expression of known and previously unknown marker genes for both alveolar lineages measured by single-cell RNA-seq. Genes were grouped into early and late markers of each lineage. Arrows, differentiation pathway; grey braces, change in transcript level of respective genes, tip pointing towards lower expression.

profile throughout embryonic and adult life. The population of EP cells co-expresses the AT2 marker *Sftpc* and the AT1 marker *Pdpn*, indicating that these cells are located at the tips of the branching epithelial tree (Extended Data Fig. 6a). EP cells segregate into two subgroups, one exclusive to E14.5 (early EPs; EP-A) and the other present at both E14.5 and E16.5 (late EPs; EP-B), indicating that cellular differentiation is not fully synchronous throughout the lung. Both EP populations show high expression of genes involved in cell cycle progression and chromosome dynamics (gene groups IIIa, IIb and IIIa; Fig. 4a, c), which are downregulated during the transition of EPs to BPs. The downregulated EP-specific genes include transcription factor *Sox11*, which is expressed in the developing airway epithelium and causes an alveolar defect when knocked out[23], and also *Tuba1a*, a putative target of *Sox11*[24]; this suggests that *Sox11* could be involved in maintaining the proliferative competence of EP cells. At E18.5, BP cells expressing both AT1 and AT2 markers appear in conjunction with intermediate populations with a decreased expression of AT1 markers (nascent AT2) or AT2 markers (nascent AT1). Mature AT2 cells are characterized by the expression of genes involved in respiratory gas exchange and immune response (gene group IV) and were only detected at adult stages (Fig. 4a). The overall number of genes as well as the total number of transcripts expressed in each cell were strongly correlated with its differentiation state: early progenitor cells at E14.5 expressed up to 6,000 genes, whereas mature AT2 cells expressed about fourfold to sixfold fewer genes (Fig. 4a and Extended Data Fig. 7a). Thus, we followed the full life cycle of *Sftpc*$^+$ cells and identified seven gene sets that robustly distinguish between

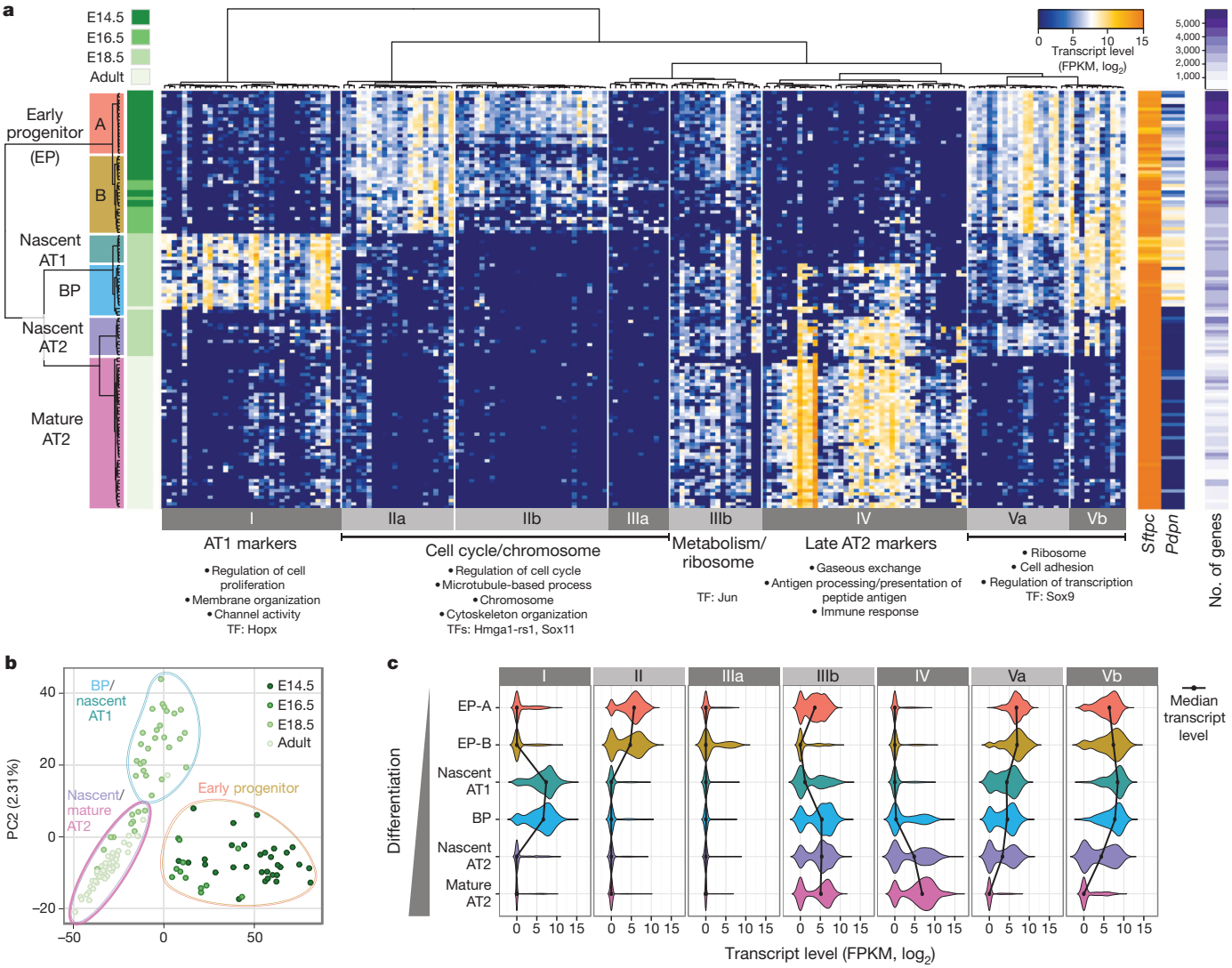


Figure 4 | Single-cell RNA-seq of *Sftpc*$^+$ cells at E14.5, E16.5, E18.5 and in the adult mouse lung explains progressive transcriptional states of the AT2 cell lineage throughout its life cycle. **a**, Hierarchical clustering of 124 *Sftpc*$^+$ cells from distal mouse lung epithelium of embryonic (E14.5, E16.5 and E18.5) and adult mice based on genes with highest principal-component loadings (Supplementary Data) in an unbiased PCA analysis (shown in **b**) of all cells and genes. Single cells are shown in rows; genes are shown in columns. Bars at the right show *Sftpc* and *Pdpn* expression, as well as the number of genes expressed by each single cell (see also Extended Data Fig. 7). Functional gene ontology enrichments and transcription factors (TFs) specific to each gene group (bottom grey-shaded bars) are shown (Supplementary Data). A similar analysis following the Clara cell lineage throughout development is shown in Extended Data Fig. 8. **b**, PCA of single-cell transcriptomes based on genes detected in more than two cells. Cells cluster into three major populations on the basis of different scores along the first two principal components. **c**, Violin plots depicting the course of expression of each of seven distinct gene groups across the six cell populations. Each violin plot shows the frequency distribution of the mean transcript level ($\log_2$-transformed FPKM) of all genes per cell.

multipotential, bipotential, nascent and mature AT2 cell states (Extended Data Fig. 6b).

We expect that a similar strategy to that pursued here can be applied to almost any tissue to empirically classify and characterize the full set of developing and mature cell types, explain the molecular regulation of these distinct populations, and explore how they are disrupted in disease.

METHODS SUMMARY

Single-cell suspensions were prepared from micro-dissected distal regions of embryonic mouse lung at E14.5, E16.5 and E18.5 and also from adult mouse lung. Epithelial cells were purified by either magnetic bead-activated cell sorting (MACS; Miltenyi Biotec) using CD45 depletion and EpCAM enrichment or by fluorescence-activated cell sorting (FACS) of transgenically labelled cells. Single cells were captured on a microfluidic chip on the C1 system (Fluidigm) and whole-transcriptome amplified cDNA was prepared on chip using the SMARTer Ultra Low RNA kit for Illumina (Clontech). Single-cell libraries were constructed as described previously[7] with the use of the Illumina Nextera XT DNA Sample Preparation kit and sequenced to a depth of (2–5) × 10^6 read pairs (HiSeq 2000; Illumina). Expression levels of transcripts were quantified with TopHat/Cufflinks[25]. Single-cell RNA-seq data was validated by the preparation of single-cell qPCR amplicons of 96 genes on a C1 microfluidic chip (Fluidigm) followed by multiplexed qPCR with BioMark (Fluidigm) as described previously[26]. Single-cell gene expression data was analysed with custom R scripts[27] (Supplementary Data). Gene ontology enrichment analysis was performed with DAVID informatics resources 6.7 (ref. 21). Additional experiments were performed with immunofluorescence and *in situ* hybridization (ACD's RNAscope *In situ* Hybridization Technology).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 July 2013; accepted 19 February 2014.
Published online 13 April 2014.

1. Kim, C. F. B. *et al.* Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* **121,** 823–835 (2005).

2. Zemke, A. C. *et al.* Molecular staging of epithelial maturation using secretory cell-specific genes as markers. *Am. J. Respir. Cell Mol. Biol.* **40,** 340–348 (2009).
3. Guha, A. *et al.* Neuroepithelial body microenvironment is a niche for a distinct subset of Clara-like precursors in the developing airways. *Proc. Natl Acad. Sci. USA* **109,** 12592–12597 (2012).
4. Gonzalez, R. *et al.* Freshly isolated rat alveolar type I cells, type II cells, and cultured type II cells have distinct molecular phenotypes. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **288,** L179–L189 (2005).
5. Xu, Y. *et al.* Transcriptional programs controlling perinatal lung maturation. *PLoS ONE* **7,** e37046 (2012).
6. Desai, T. J., Brownfield, D. G. & Krasnow, M. A. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507,** 190–194 (2014).
7. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* **11,** 41–46 (2013).
8. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21,** 1160–1167 (2011).
9. Islam, S. *et al.* Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nature Protocols* **7,** 813–828 (2012).
10. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality inexpression and splicing in immune cells. *Nature* **498,** 236–240 (2013).
11. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol.* **14,** R31 (2013).
12. Liu, C. L., Bernstein, B. E. & Schreiber, S. L. Whole genome amplification by T7-based linear amplification of DNA (TLAD). II. Second-strand synthesis and in vitro transcription. *CSH Protocols*, http://dx.doi.org/10.1101/pdb.prot5003 (2008).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
14. Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnol.* **30,** 777–782 (2012).
15. Tariq, M. A., Kim, H. J., Jejelowo, O. & Pourmand, N. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res.* **39,** e120 (2011).
16. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10,** 1096–1098 (2013).
17. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10,** 1093–1095 (2013).
18. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382 (2009).
19. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature Protocols* **5,** 516–535 (2010).
20. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6,** 468–478 (2010).
21. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
22. Yin, Z. *et al.* Hop functions downstream of Nkx2.1 and GATA6 to mediate HDAC-dependent negative regulation of pulmonary gene expression. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **291,** L191–L199 (2006).
23. Sock, E. *et al.* Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Mol. Cell. Biol.* **24,** 6635–6644 (2004).
24. Wang, X. *et al.* Gene expression profiling and chromatin immunoprecipitation identify DBN1, SETMAR and HIG2 as direct targets of SOX11 in mantle cell lymphoma. *PLoS ONE* **5,** e14085 (2010).
25. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
26. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnol.* **29,** 1120–1127 (2011).
27. R core team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing* http://www.R-project.org/.

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank W. Koh and B. Passarelli for help and discussions regarding bioinformatic pipelines and statistical analysis, S. I. Gonzalez for help with immunofluorescence, and J. G. Camp and members of the Krasnow laboratory for critical discussion and reading of the manuscript. This work was supported by National Heart, Lung, and Blood Institute (NHLBI) U01HL099995 Progenitor Cell Biology Consortium Grant (B.T., M.A.K., S.R.Q.), by National Institutes of Health (NIH) T32HD007249 (D.G.B.), by a Parker B. Francis Foundation Fellowship and NIH 5K08HL084095 Award (T.J.D.), and by NIH grant U01HL099999 (A.R.W., N.F.N.). M.A.K. and S.R.Q. are investigators of the Howard Hughes Medical Institute.

Author Contributions B.T., D.G.B., T.D., M.A.K. and S.R.Q. conceived the study and designed the experiments. B.T., D.G.B., F.H.E., A.R.W., N.F.N., G.L.M. and T.D. performed the experiments. B.T., D.G.B., A.R.W., F.H.E., T.D., M.A.K. and S.R.Q. analysed the data and/or provided intellectual guidance in their interpretation. B.T., D.G.B., F.H.E., T.D., M.K. and S.R.Q. wrote the paper.

Author Information The transcriptome sequencing data for all single cells has been deposited in the Gene Expression Omnibus database under accession number GSE52583. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R.Q. (quake@stanford.edu), M.A.K. (krasnow@stanford.edu), T.J.D. (tdesai@stanford.edu).

METHODS

Mouse strains. Timed-pregnant C57BL/6J females (JAX) were used for all embryonic time points reported; gestation age was verified by crown–rump length before use. For adult mice, a transgenic-labelling approach was employed to enrich for AT2 cells. Mice were bred to be homozygous for a knock-in allele into *Sftpc* encoding for a reverse tetracycline transactivator (SftpC-Cre-ERT2-rtTA)[28] and heterozygous for an inserted transgene that drives the expression of a GFP tagged human histone 1 in a tetracycline-dependent manner (tetO-HIST1H2BJ/GFP). To validate the expression of *Hopx*, mice were bred to be hetereozygous for the knock-in of a tamoxifen inducible Cre recombinase (Cre-ERT2) construct into the *Hopx* gene (Hopx-Cre-ERT2)[29] and heterozygous for a transgenic insertion into the *Rosa26* locus encoding a two-colour, membrane-tethered fluorophore reporter that switches expression from a red to green fluorophore on Cre-mediated recombination (mTmG) (cross of B6;129S-*Hopx*^{tm1Eno}/J and B6.129(Cg)-*Gt(ROSA)26Sor*$^{tm4(ACTB-tdTomato,-EGFP)Luo}$/J). Genotyping was performed using PCR with published primer sets from genomic DNA extracted from tails by Proteinase K (Sigma) digestion and precipitation with ethanol. Mice were housed in filtered cages and all experiments were performed in accordance with approved Institutional Animal Care and Use Committee protocols.

Isolation and disaggregation of lung tissue. Single-cell experiments were performed on embryonic mouse lung at E14.5, E16.5 and E18.5 and also on adult mouse lung. In general, embryonic experiments were performed on pooled sibling lungs of one litter (five to seven lungs per pool). One of three replicate experiments at E18.5 (cells referred to as 'E18_2_Cxx' in Supplementary Data) was performed on a single embryonic lung.

Adult mice were euthanized by administration of CO_2. For time points E14.5, E16.5 and E18.5, embryos were removed and lungs were isolated *en bloc* without perfusion and pooled by litter (five to seven embryos) for further processing. Lungs from E14.5 and E16.5 time points were dissociated in Dispase (BD Biosciences) and triturated with glass Pasteur pipettes until a single-cell suspension was attained. For E18.5 and adult time points, either total lung (adult) or peripheral lobe edges (E18.5) were minced with a razor blade into 1 mm^3 fragments, suspended in 5 ml of digestion buffer consisting of elastase (3 U ml^{-1}; Worthington Biochemical Corporation) and DNase I (0.33 U ml^{-1}; Roche) in DMEM/F12 medium, incubated with frequent agitation at 37 °C for 45 min, and triturated briefly with a 5-ml pipette. For all time points, an equal volume of DMEM/F12 supplemented with 10% FBS and penicillin–streptomycin (1 U ml^{-1}; Thermo Scientific) was added to single-cell suspensions before passing the suspension through a 100-μm mesh filter (Fisher) and centrifugation at 400*g* for 10 min. Pelleted cells were resuspended in red blood cell lysis buffer (BD Biosciences), incubated for 2 min, passed through a 40-μm mesh filter (Fisher), centrifuged at 400*g* for 10 min and then resuspended in sorting buffer (PBS supplemented with 0.05% BSA and 2 mM EDTA).

Purification of embryonic distal lung epithelial cells by MACS. Lung epithelial cells for embryonic time points (E14.5, E16.5 and E18.5) were purified by MACS using MS columns (Miltenyi Biotec) in MACS buffer (2 mM EDTA, 0.5% BSA in PBS, filtered and degassed) in accordance with the protocol provided by the vendor. Before loading, the single-cell suspension was passed through a 35-μm cell strainer (BD Biosciences). Leukocytes and alveolar macrophages were removed by depletion with an antibody against the surface antigen CD45 conjugated to magnetic beads (Miltenyi Biotec) followed by enrichment for epithelial cells, incubating first with a biotinylated primary antibody for EpCAM (clone G8.8; eBioscience) followed by a secondary antibody against biotin conjugated to magnetic beads (Miltenyi Biotec).

Purification of adult AT2 cells by FACS. For AT2 cells from the adult lung, an adult Sftpc-Cre-ERT2-rtta$^{-/-}$ tetO-HIST1H2BJ-GFP$^{+/-}$ mouse was injected with 2 mg of doxycycline (Sigma) three days prior to the single cell experiments. After isolation and disaggregation of the lung, the single-cell suspension was incubated with a viability stain (Sytox Blue; Invitrogen) for 15 min and viable GFP^{+} cells were sorted by FACS (Aria II; BD Biosciences) into DMEM containing 10% FBS.

Capturing of single cells and preparation of cDNA. Single embryonic lung epithelial cells were captured on a medium-sized (10–17 μm cell diameter) microfluidic RNA-seq or STA chip (Fluidigm) using the Fluidigm C1 system. To ensure unbiased and comprehensive profiling of all distal lung epithelial cells, an initial experiment was performed with a microfluidic chip with a 17–25-μm capture range; however, no cells with diameter greater than ~15 μm were captured, indicating that no major cell populations were missed by using the smaller capture range (Extended Data Fig. 2b). Cells were loaded onto the chip at a concentration of 300–500 cells μl^{-1}, stained for viability (LIVE/DEAD cell viability assay; Molecular Probes, Life Technologies) and imaged by phase-contrast and fluorescence microscopy to assess the number and viability of cells per capture site. Only single, live cells were included in the analysis. For RNaseq experiments, cDNAs were prepared on chip using the SMARTer Ultra Low RNA kit for Illumina (Clontech). ERCC (External RNA Controls Consortium) RNA spike-in Mix (Ambion, Life Technologies) was added to the lysis reaction and processed in parallel to cellular messenger RNA. For qPCR experiments, amplicons were prepared with pooled DELTAgene assays (Fluidigm) and Ambion (Life Technologies) Cells to CT lysis and pre-amplification kit, using the protocol provided by Fluidigm.

RNA-seq library construction. Single-cell cDNA size distribution and concentration was assessed on a capillary electrophoresis-based fragment analyser (Advanced Analytical). Illumina libraries were constructed in 96-well plates using the Illumina Nextera XT DNA Sample Preparation kit as described previously[7] using the protocol supplied by Fluidigm. For each C1 experiment, a bulk RNA control (about 200 cells) and a no-cell negative control were processed in parallel in PCR tubes, using the same reagent mixes as used on chip. Libraries were quantified by Agilent Bioanalyzer, using High Sensitivity DNA analysis kit, and also fluorometrically, using Qubit dsDNA HS Assay kits and a Qubit 2.0 Fluorometer (Invitrogen, Life Technologies).

DNA sequencing. Single-cell Nextera XT (Illumina) libraries of one experiment were pooled and sequenced 100 base pairs (bp) paired-end on Illumina HiSeq 2000 to a depth of (2–6) × 10^6 reads (three replicate experiments of distal mouse lung epithelial cells at E18.5, one experiment at E14.5 and one experiment on adult AT2 cells) or 150 bp paired-end on Illumina MiSeq (one experiment at E16.5) to a depth of 100,000–550,000 reads with v3 chemistry. CASAVA 1.8.2 was used to separate out the data for each single cell by using unique barcode combinations from the Nextera XT preparation and to generate *.fastq files.

Microfluidic single-cell multiplexed qPCR. Single-cell multiplexed qPCR was performed in a M96 quantitative PCR DynamicArray on the Fluidigm Biomark instrument as described previously[26], using a panel of 96 DELTAgene assays (Fluidigm; Supplementary Table 2). In three of five single-cell qPCR experiments, ERCC spike-in transcripts (Ambion Live Technologies) were added to each single-cell lysis reaction on chip. Primer pairs for 6 of the 92 exogenous RNA spike-ins (ERCC spike-ins ERCC-00033, ERCC-00136, ERCC-00044, ERCC-00164, ERCC-00054 and ERCC-00074) were added to the preamplification reaction on chip and were subsequently used in the multiplexed qPCR experiment to detect the transcript level of each RNA spike-in. qPCR detection of the spike-in transcripts was later used to convert measured C_t values to approximate numbers of transcripts in a subset of 90 genes (Extended Data Fig. 7).

Processing, analysis and graphic display of single-cell RNA-seq data. Raw reads were pre-processed with the sequence-grooming tools FASTQC[30], cutadapt[31] and PRINSEQ[32] followed by sequence alignment with the Tuxedo suite (Bowtie[33], Bowtie2 (ref. 34), TopHat[25]) and SAMtools[35], using default settings. See Supplementary Data for information about the number of total reads and the percentage of mapped reads for each single cell. Transcript levels were quantified as fragments per kilobase of transcript per million mapped reads (FPKM) generated by TopHat/Cufflinks. Where depth matching was done, Seqtk (H. Li, https://github.com/lh3/seqtk/) was used to select raw reads randomly from each library, and the same pre-processing and alignment pipelines were used to obtain FPKM values for the depth-matched samples. Limit of detection of microfluidic single-cell RNA-seq was determined by analysing the correlation between the concentration of exogenous ERCC spike-in mRNA sequences and their respective mean FPKM values as measured by RNA-seq (Extended Data Fig. 3c). The spike-in sequences reflect a diverse range of sequence content and length, they have low homology with eukaryotic transcripts because they are from microbial sources, and they span a large range of concentrations to allow an empirical determination of the limit of detection[7,36,37]. The limit of detection was on the order of 0.5 molecules per reaction chamber, which is reflected as an FPKM value of ~1 (or 0 on a $\log_2$ scale). Transcripts with an FPKM value below or equal to 1 were therefore considered not expressed. Cells not expressing either of two housekeeping genes *Actb* and *Gapdh* (encoding β-actin and glyceraldehyde-3-phosphate dehydrogenase, respectively), or expressing them at less than three standard deviations below the mean, were scored as unhealthy and removed from the analysis. After applying this filter, a total of 80 cells remained for three replicate experiments at E18.5 (2 × pooled sibling lungs (20 and 26 cells), 1 × single lung (34 cells)), 45 cells remained for one experiment at E14.5, 27 cells remained for one experiment at E16.5 and 46 cells remained for an experiment of adult AT2 cells yielding 198 single cells in total.

For the lung epithelial cells at E18.5, we detected between 1,017 and 4,998 expressed genes per single cell, 10,946 in the union of all single cells and 8,653 in the 200 cell control bulk sample, indicating the heterogeneity of the analysed single cells. A total of 81 genes were commonly expressed in all single cells (Supplementary Table 1), which were mainly enriched for general processes such as translation.

FPKM values were converted to an approximate number of transcripts by using the correlation between the number of transcripts of exogenous spike-in mRNA sequences and their respective measured mean FPKM values (Extended Data Fig. 3c). The number of spike-in transcripts per single-cell lysis reaction was calculated from the concentration of each spike-in provided by the vendor (Ambion, Life Technologies), the approximate volume of the lysis chamber (10 nl) and the dilution of spike-in transcripts in the lysis reaction mix (×40,000). Transcript levels were converted to logarithmic space by taking $\log_2$(FPKM/number of transcripts). When calculating the characteristic single-cell expression (Extended Data Fig. 3b, d, f, g),

we used either the mean FPKM or the median FPKM value of each gene across all single cells transformed to the $\log_2$ space. To calculate the coefficient of variation of each gene across single cells (Extended Data Fig. 3b), the standard deviation of the $\log_2$-transformed FPKM values of a gene across all single cells was divided by the mean $\log_2$ FPKM value of the same gene.

Saturation plots (Extended Data Fig. 3a and 7a) were generated as described previously[7]. In brief, a corresponding number of millions of raw reads were randomly selected from each sample library and then, using the same alignment pipeline, FPKM values were called for each gene. This random subsampling was repeated, for each sample replicate, a total of four subsampled data sets per point, and the mean number of genes with an FPKM greater than 1 was plotted. For generating the 'single-cell ensemble' data set, raw reads from all the single-cell RNA-seq libraries were bioinformatically pooled to mimic a bulk RNA-seq experiment.

Custom R scripts[27] were used to perform principal component analysis (PCA), hierarchical clustering, Guilt-by-Association and permutation analysis as well as to construct violin plots, correlation plots and histograms. The scripts can be found in Supplementary Data. PCA analysis was performed on all cells using all genes expressed in more than two cells and with a variance in transcript level ($\log_2$(FPKM)) across all single cells greater than 0.5. Subsequently, genes with the highest principal-component loadings (highest absolute correlation coefficient with one of the first three or four principal components) were identified. Hierarchical clustering was performed on cells and on the genes identified by PCA, using Euclidean or correlation distance metric.

The specificity of the hierarchical clustering in Fig. 2a identifying five distinct cell populations was assessed with a permutation analysis approach. The sum of squares within groups (SSW) was therefore calculated for the cell grouping presented in Fig. 2a as well as for 50,000 random permutations thereof, keeping the size of cell groups and the total number of groups constant. With $x^k_{i,j}$ being the transcript level of gene j in cell i belonging to group k, the SSW can be calculated as

$$\mathrm{SSW} = \sum_{k=1}^{n}\sum_{j=1}^{m}(x^k_{i,j} - {}_i\overline{x}^k_j)^2$$

with ${}_i\overline{x}^k_j = \frac{1}{n}\sum_{i=1}^{n} x^k_{i,j}$ being the mean transcript level of gene j in all cells i = 1, 2, …, n belonging to group k. The SSWs for all 50,001 permutations were normally distributed and the SSW for our chosen clustering was significantly lower than that for all other permutations ($P = 2.89 \times 10^{-122}$).

When *Sftpc*$^+$ cells were isolated from all single-cell RNA-seq data sets (Fig. 4), a *Sftpc* transcript level of $\log_2$(FPKM) = 10 was chosen as threshold to separate cells with background *Sftpc* expression from cells with high *Sftpc* expression (referred to as *Sftpc*$^+$ cells).

To search for further previously unknown cell-type markers and cell-type specific transcription factors or receptors/ligands beside the genes identified by PCA, we defined a 'perfect marker gene' for each cell type with a high transcript level ($\log_2$(FPKM) = 10) in all cells of the respective cell type, and with no expression (FPKM = 0) in all other cells. We then determined the pairwise Pearson correlation between the single-cell expression profile of each perfect marker gene and every other transcribed gene. The list of murine transcription factors that were screened for cell-type specificity was obtained from the online animal transcription factor database (http://www.bioguo.org/AnimalTFDB/) (ref. 38). All genes identified as cell-type-specific by PCA analysis and hierarchical clustering (see above) also had a high Pearson correlation coefficient with the corresponding perfect marker gene. The Pearson correlation coefficients for the most strongly correlating genes are shown in Supplementary Data together with information about the top 30 genes per cell type regarding previous detection in cell types in the lung and regarding available literature or known mouse knockout phenotypes.

Guilt-by-Association analysis[39] was used to calculate the probability of observing a given co-expression of two genes by chance. Gene expression values were therefore scaled gene-by-gene by mean-centring and dividing by the standard deviation of the respective gene across all single cells, and a binary expression matrix was constructed by defining a gene as expressed in a given cell if the scaled expression level was greater than or equal to 0, and as not expressed if it was smaller than 0. Pairwise comparisons were performed between the perfect marker gene for each of the four mature cell types (AT1, AT2, Clara and ciliated) and all other genes expressed in at least one cell (10,946 genes in total). P values were calculated with the hypergeometric distribution as described in ref. 37, and multiple testing was accounted for by the Benjamini–Hochberg method (Fig. 2b and Supplementary Data).

Gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed with DAVID informatics resources 6.7 of the National Institute of Allergy and Infectious Diseases of the NIH[21,40] (Supplementary Data).

Analysis and graphic display of microfluidic single-cell multiplexed qPCR data. Single-cell multiplexed qPCR data were analysed and displayed by using custom R scripts[27]. qPCR experiments were performed for E16.5 (two biological replicates), E18.5 (two biological replicates) distal lung epithelial cells and for adult AT2 cells (one replicate). The limit of detection of multiplexed qPCR values was determined as 22 threshold cycles (C_t) by a calibration experiment with 16-fold serial dilutions of total lung cDNA and six replicates for each concentration. Genes that were not expressed were given a value higher than the limit of detection and the limit of detection was subtracted from all C_t values to transform C_t values to $\log_2$ expression values ($\log_2\mathrm{Ex} = C_{t,\mathrm{LoD}} - C_t$, $C_{t,\mathrm{LoD}} = 22$). Cells not expressing either of two housekeeping genes *Actb* and *Gapdh*, or expressing them at less than three standard deviations below the mean, were scored as unhealthy and removed from the analysis. After applying this filter, 74 single cells remained for two experiments at E18.5, 33 cells for two experiments at E16.5, and 48 cells for the experiment with adult AT2 cells. In all experiments, cells were isolated from pooled lungs from one litter (five to nine lungs). To combine experiments from different chips for the same embryonic time point, the expression value of each gene for a given cell was normalized to the median gene expression value of that cell. Normalized gene expression values were further scaled gene by gene by mean-centring and dividing by the standard deviation of expressing cells. PCA and hierarchical clustering using Euclidean distance metric were performed in R for all cells, using ten canonical marker genes for bronchiolar and alveolar cells (*Abca3*, *Sftpb*, *Muc1*, *Sftpc*, *Lyz2*, *Aqp5*, *Pdpn*, *Ager*, *Foxj1* and *Scgb1a1*).

Immunofluorescence. E18.5 lungs were removed *en bloc*; for whole-mount staining, the tip of the accessory lobe was excised. Lungs and tips were immersion-fixed overnight in paraformaldehyde (4% in PBS) at 4 °C, and then dehydrated and stored in methanol at −20 °C until being stained. Lungs of adult mice were collected as above except that after clearance of the pulmonary vasculature, the ventral trachea was incised and cannulated, and lungs were gently inflated to full capacity with molten low-melting-point agarose (2% in PBS; Sigma). Ice-cold PBS was dripped into the thorax to solidify the agarose, and inflated lungs were removed *en bloc* and processed as above. E18.5 lungs were rehydrated, cryoprotected overnight in 30% sucrose at 4 °C, submerged in OCT (Tissue Tek) in an embedding mould, frozen on solid CO_2, then stored at −80 °C. Sections of thickness 10 μm obtained with a cryostat (Leica CM3050S) were collected on chambered glass slides and stored at 4 °C before being stained.

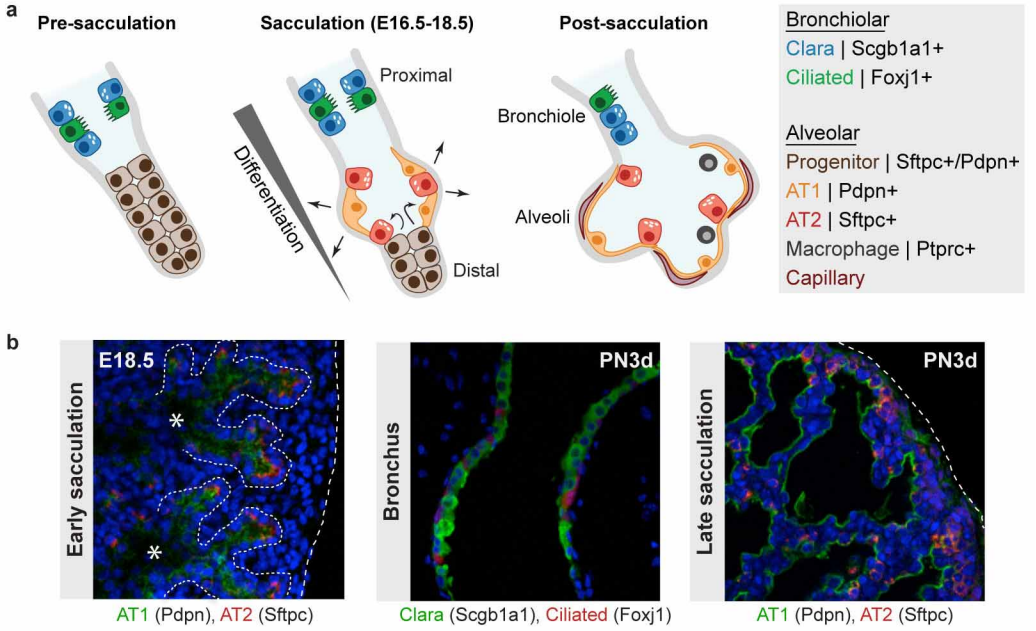
Similar immunofluorescence protocols were used on whole mounts as on sections, except that incubation times were increased to compensate for tissue thickness. Lung tissue was permeabilized (10 min, PBS containing 0.3% Triton X-100), washed (three 5-min washes in PBS containing 0.1% Tween 20) and blocked (1 h, PBS containing 10% donkey serum) before incubation overnight with primary antibody. Adult lungs did not require further permeabilization. Primary antibodies against the following antigens (used at 1:200 dilution unless otherwise noted) were pro-SftpC (rabbit; Chemicon AB3786), Pdpn (hamster; DSHB 8.1.1), E-cadherin (rat; Zymed ECCD-2), Rage (rat; R&D), Scgb1a1 (rabbit; Upstate), Lamp3 (sheep, AF4584; R&D), S100a6 (rat, DDX0192; Dendritics) and Krt15 (mouse, LHK15; SCBT) directly conjugated to a fluorophore in accordance with the manufacturer's instructions (Alexa Fluor Antibody Labelling Kit; Invitrogen). After further washing, sections were incubated with appropriate secondary antibodies conjugated to an Alexa fluorophore (donkey A488, A555 or A633; Invitrogen) as well as DAPI (5 ng ml^{-1}) for 1 h, followed by washing and mounting in Vectashield (Vector). Lung tissues were imaged with a laser scanning confocal microscope (LSM 780; Zeiss).

***In situ* hybridization.** For *in situ* hybridizations, embryonic lungs were collected as for immunostaining (see above), washed briefly in PBS (autoclaved, diethyl pyrocarbonate-treated) and snap-frozen in OCT for sectioning. Sections of thickness 10 μm were generated on the cryostat and stored at −20 °C before further processing. To validate AT2-specific expression of *Egfl6* RNA by *in situ* hybridization, sections were transported on dry ice to a company specializing in processing and imaging dual *in situ* hybridized samples, in accordance with the company's reported protocol (ACD's RNAscope *In situ* Hybridization Technology). To validate the expression of *Sftpc* and explore its spatial expression pattern in the embryonic mouse lung (E11.5, E13.5 and E14.5), *in situ* hybridizations were performed on whole-mount mouse lungs as described previously[41].

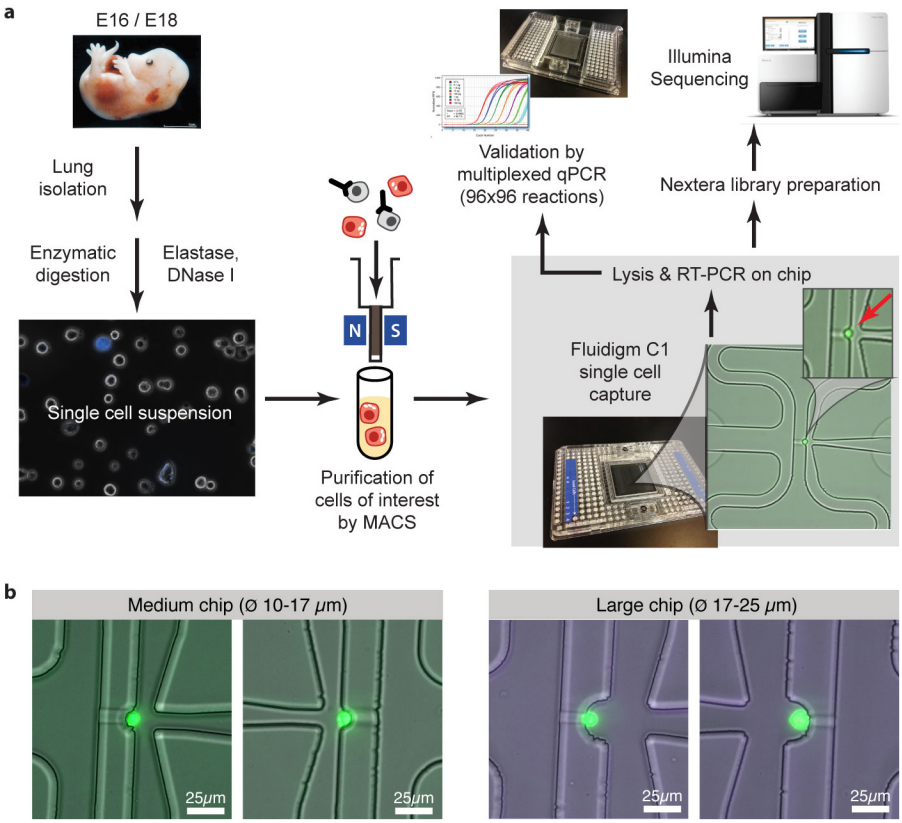
Validation of *Hopx* as AT1 marker gene by transgenic labelling. Cells actively transcribing Hopx in the adult mouse lung were labelled by injecting 2 mg of tamoxifen (Sigma) in corn oil at 20 mg ml^{-1} concentration intraperitoneally into postnatal day-28 Hopx-Cre-ERT2$^{+/-}$ mTmG$^{+/-}$ mice. Three days later, lungs were collected as described above, fixed overnight in paraformaldehyde (4% in PBS) at 4 °C and stored in 80% glycerol at 4 °C before imaging with a laser scanning confocal microscope (LSM 780; Zeiss) with a 0.8 numerical aperture, 25× oil-immersion objective and confocal z-sections with a thickness of 2.3 μm.

28. Chapman, H. A. *et al.* Integrin α6β4 identifies an adult distal lung epithelial population with regenerative potential in mice. *J. Clin. Invest.* **121,** 2855–2862 (2011).

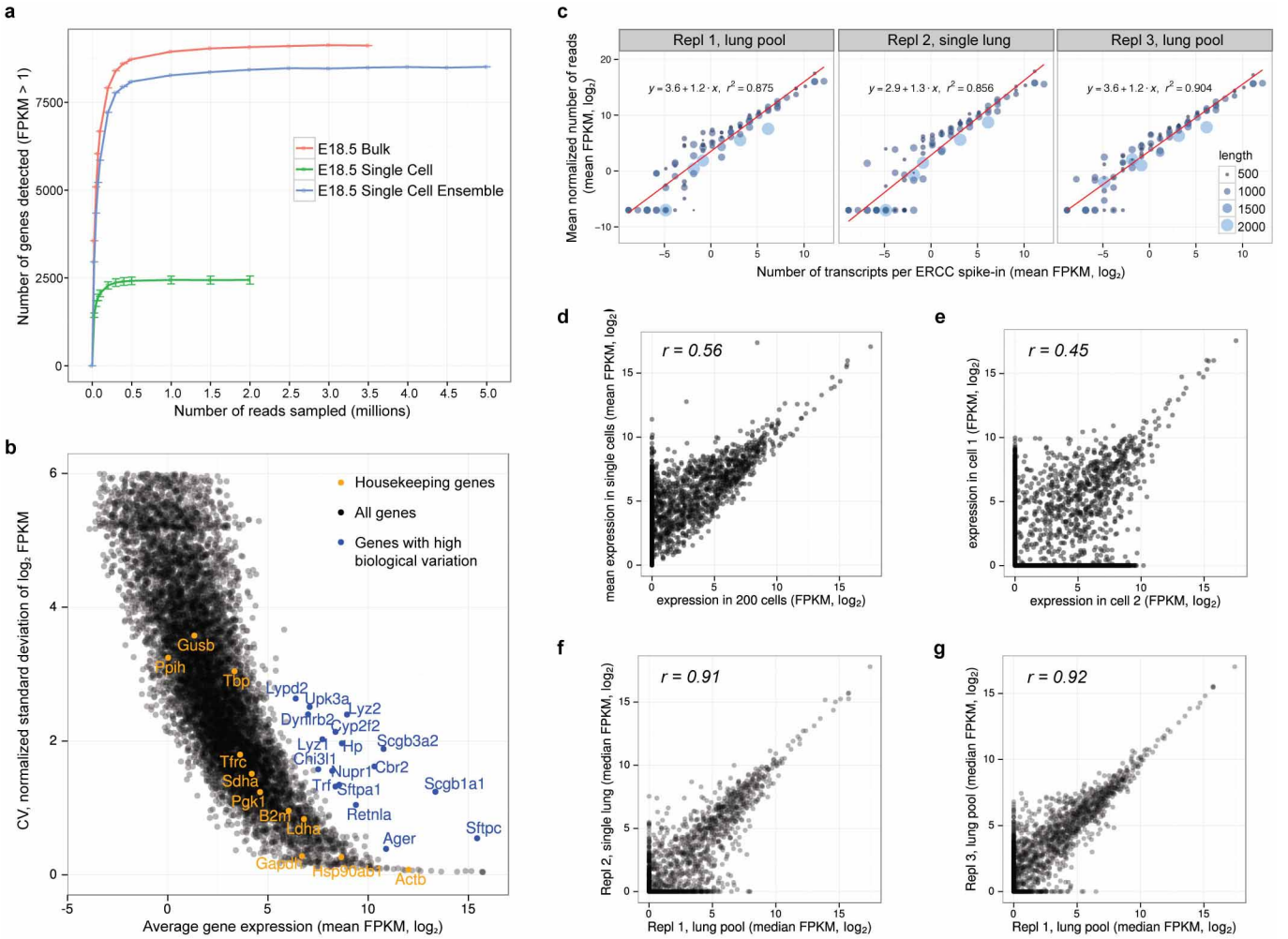
29. Takeda, N. *et al.* Interconversion between intestinal stem cell populations in distinct niches. *Science* **334,** 1420–1424 (2011).
30. Babraham Institute. Babraham Bioinformatics. FASTQC. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc.
31. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
32. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
33. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
35. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
36. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nature Methods* **2,** 731–734 (2005).
37. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21,** 1543–1551 (2011).
38. Zhang, H.-M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.* **40,** D144–D149 (2012).
39. Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D. & Klingler, T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.* **9,** 1198–1203 (1999).
40. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
41. Greif, D. M. *et al.* Radial construction of an arterial wall. *Dev. Cell* **23,** 482–493 (2012).



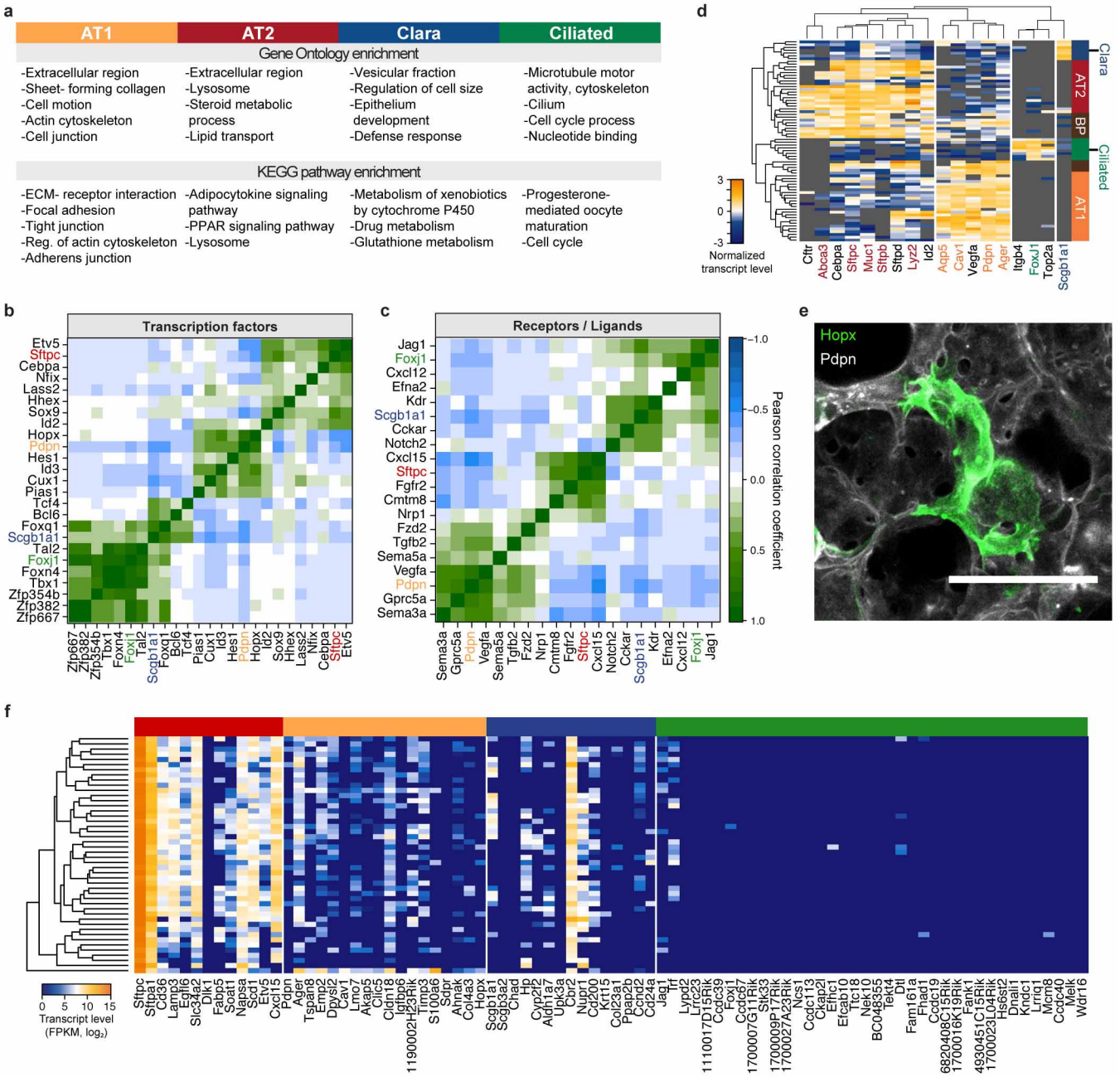
Extended Data Figure 1 | Schematic illustration of the process of sacculation. **a**, Schematic illustration of morphological and molecular changes of the distal airways during development. Cell differentiation progresses in a directional manner from the bronchio-alveolar junction (proximal) to the distal tip (distal) of each terminal airway; progenitor cells therefore persist the longest at the tips. Ciliated (green) and Clara (blue) cells mature first, followed by the differentiation of flat alveolar type 1 (AT1, orange) and cuboidal type 2 (AT2, red) cells from cuboidal alveolar progenitors during sacculation (E16–18.5), when distal airway tubules widen as nascent AT1 cells flatten to form the gas-exchange surface. **b**, Micrographs of alveolar (E18.5, postnatal 3 days (PN3d)) and bronchiolar (PN3d) sections of a mouse lung co-stained for Clara (Scgb1a1, green) and ciliated (Foxj1, red) cell markers as well as AT1 (Pdpn, green) and AT2 (Sftpc, red) specific markers. Progenitor cells at the tips of sacculating alveoli are detected by an overlap of AT1 and AT2 specific markers. Newly forming alveolar sacs are marked by asterisks.



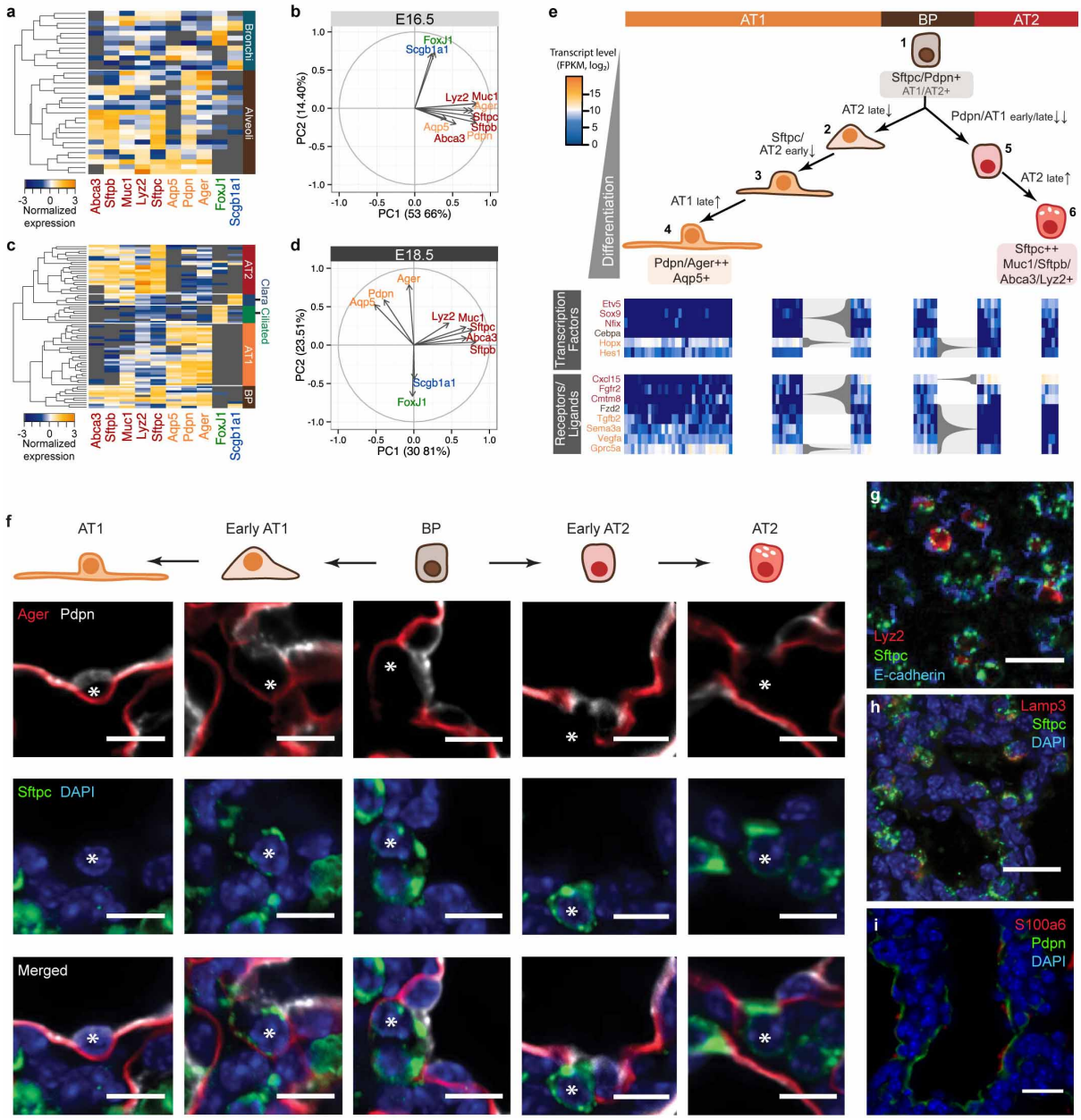
Extended Data Figure 2 | Single-cell transcriptomics analysis workflow. **a**, Workflow of single-cell transcriptomics analysis of mouse lung epithelial cells. A single captured lung epithelial cell stained with Alexa488 for EpCAM (green) is indicated by a red arrow. **b**, Single lung epithelial cells captured in microfluidic chips with capture sites designed to trap cells with a diameter of 10–17 μm (medium, left) or 17–25 μm (large, right). Cells were stained for viability with Calcein AM. Even cells captured by the large chip did not exceed a diameter of ~15 μm, indicating that the medium-sized chips are sufficient for comprehensively profiling distal mouse lung epithelial cells.



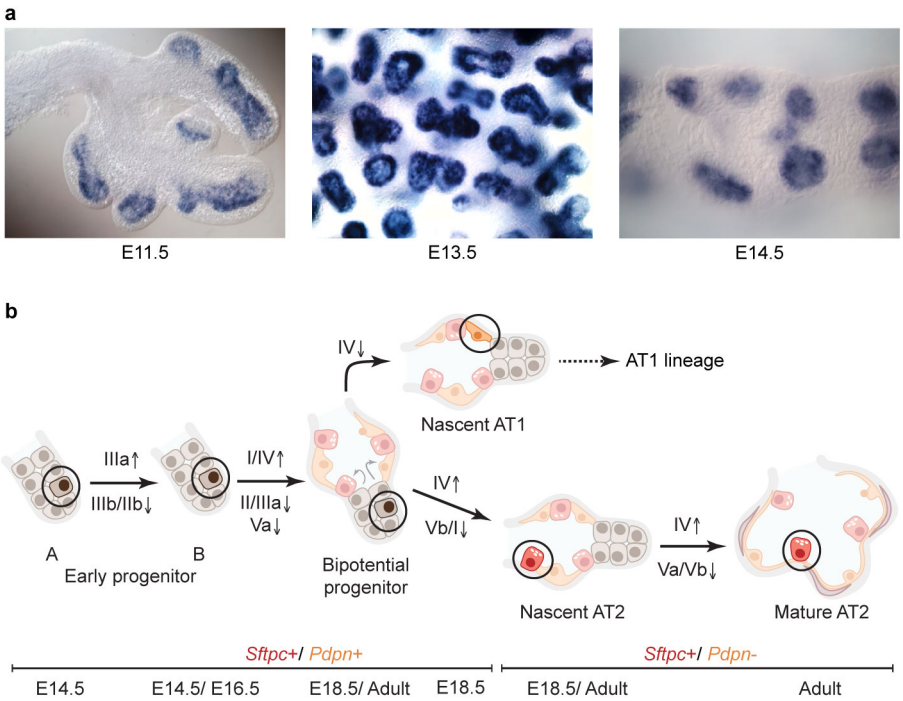
Extended Data Figure 3 | Assessment of required sequencing depth, technical and biological variation, dynamic range and reproducibility of single-cell RNA-seq data of 80 single distal lung epithelial cells at E18.5. **a**, Saturation analysis reveals the sequencing depth required for the detection of most genes expressed by single cells. To detect most expressed genes, single-cell RNA-seq libraries have to be sequenced only to a depth of about 10^6 reads, whereas libraries of bulk samples have to be sequenced more deeply. The number of genes detected in the ensemble of all single cells (synthetic bulk) is comparable to the number of genes detected in the true bulk experiment. Each point on the saturation curve was generated by randomly selecting a number of raw reads from each sample library (bulk, 200 cell bulk library; single cell, single-cell RNA-seq libraries of 80 lung epithelial cells; single-cell ensemble, bioinformatically pooled single-cell libraries) and then using the same alignment pipeline to call genes with a mean FPKM of more than 1. Each point represents four replicate subsamplings; error bars represent s.e.m. **b**, Technical noise and biological variation in single-cell RNA-seq data. Relationship between mean expression level and coefficient of variation for 10,946 genes in single embryonic lung epithelial cells. Several genes show strong biological variation (blue): they show higher variability than the average noise at a given average gene expression. Housekeeping genes are shown in yellow. **c**, Average detected transcript levels (mean FPKM, $\log_2$) for 92 ERCC RNA spike-ins as a function of provided number of molecules per lysis reaction for each of the three independent single-cell RNA-seq experiments performed at E18.5. Linear regression fits through data points are shown. The length of each ERCC RNA spike-in transcript is encoded in the size and colour of the data points. No particular bias towards the detection of shorter versus longer transcripts is observed. The method shows single transcript sensitivity as well as a dynamic range of approximately six orders of magnitude, in agreement with a previous study evaluating microfluidic single-cell RNA-seq[7]. **d**, **e**, Correlation between transcript levels of a 200-cell population and median transcript levels of single cells of the same pool of embryonic lungs (**d**), and transcript levels of two single AT2 cells (**e**). *r*, Pearson correlation coefficients. **f**, **g**, Correlation between transcript levels of all genes detected in the single lung and the pooled lung experiment (**f**) and between transcript levels of all genes detected in the two independent experiments on pooled embryonic lungs (**g**). Pearson correlation coefficients *r* are given.



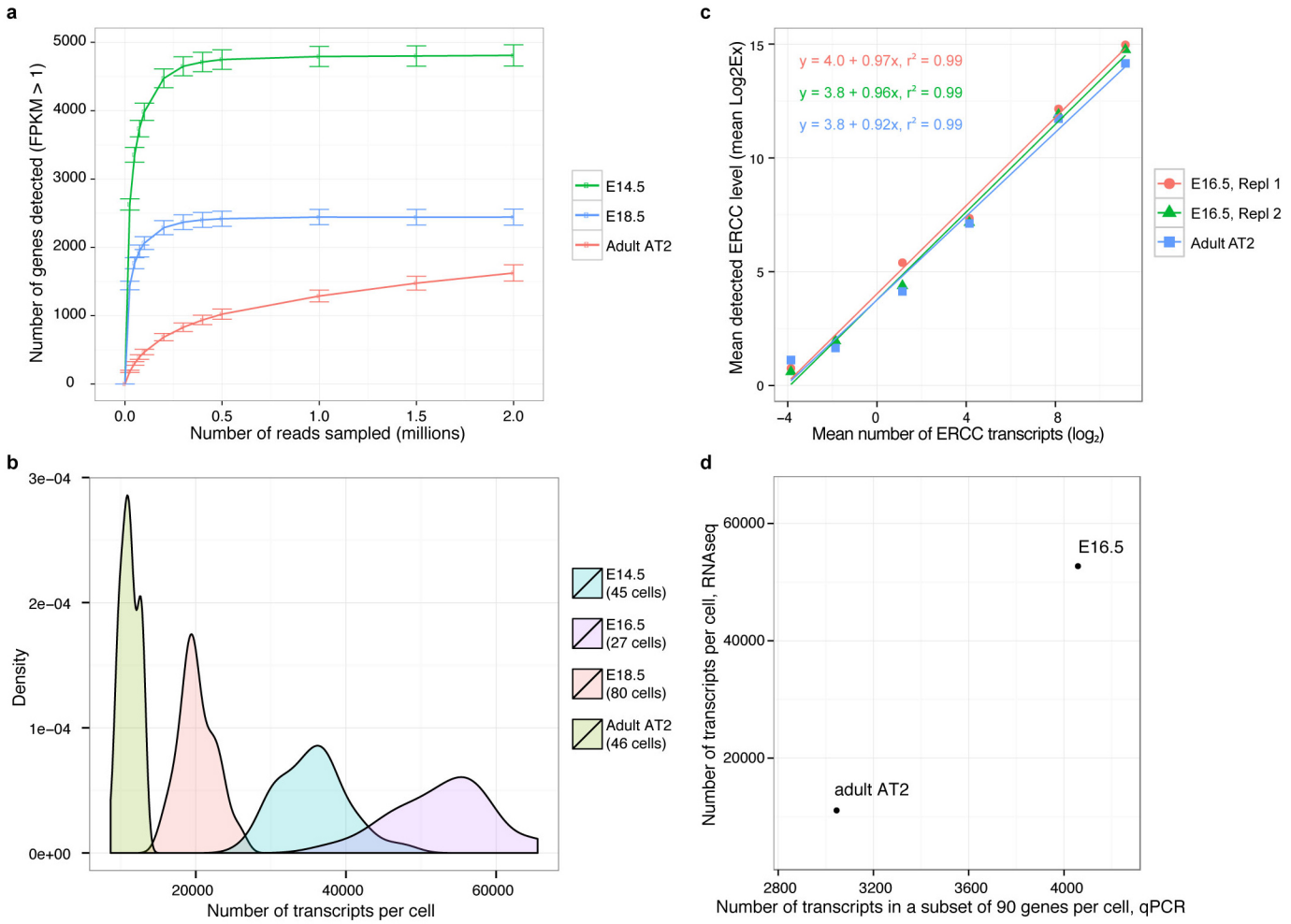
Extended Data Figure 4 | Lineage-specific genes identified by single-cell transcriptome analysis allow functional description of individual distal lung epithelial cell populations. **a**, Results of gene ontology (GO) and KEGG pathway enrichment analyses for distal lung epithelial cell types based on lineage-specific genes identified by single-cell RNA-seq of 80 E18.5 distal lung epithelial cells (Supplementary Data). **b**, **c**, Correlograms visualizing correlation of single-cell gene expression profiles between transcription factors (**b**) or receptors/ligands (**c**) and the major canonical marker genes for bronchiolar and alveolar lineages (AT1: *Pdpn*; AT2: *Sftpc*; Clara: *Scgb1a1*; ciliated: *Foxj1*). The colour bar denotes the Pearson correlation coefficient from −1 (blue, anticorrelated genes) to 1 (green, positively correlated genes). **d**, Validation of previously unknown marker genes by single-cell multiplexed qPCR on 74 single cells isolated from the distal mouse lung epithelium at E18.5. Lineage-specific expression of seven new marker genes is shown by clustering with known markers for respective lineages (AT2, red, previously unknown: *Cftr*, *Cebpa*, *Sftpd* and *Id2*; AT1, orange, previously unknown: *Vegfa*; ciliated, green, previously unknown: *Itgb4* and *Top2a*; Clara, blue). **e**, Validation of *Hopx* expression in AT1 cells. A lung section from a transgenic *Hopx>GFP* adult mouse (*Hopx-Cre-ERT2*$^{+/-}$;*mTmG*$^{+/tg}$) was co-stained for AT1 marker Pdpn. Maximum-intensity projections of confocal *z* stacks show that AT1 cells expressing the membrane-localized GFP reporter (green) also express Pdpn (white). Scale bar, 50 μm. **f**, Hierarchical clustering of 46 transgenically labelled mature *Sftpc*$^{+}$ AT2 cells, isolated by FACS from adult mouse lung. Most genes identified as AT2 lineage-specific from single-cell transcriptomes at E18.5 are transcribed also by mature AT2 cells. In contrast, no or low expression is observed in mature AT2 cells for the genes specific to the other alveolar or bronchiolar lineages as identified from single-cell RNA-seq data at E18.5.



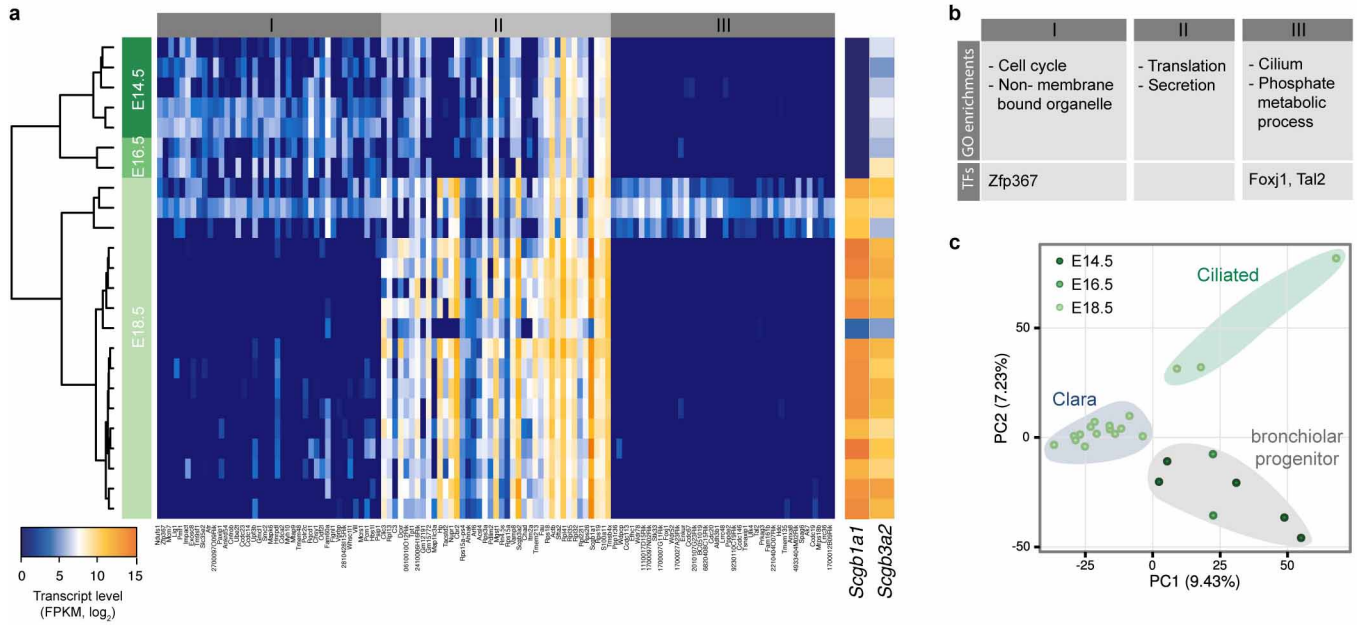
Extended Data Figure 5 | Molecular profiles distinguish developmental intermediates during the differentiation of AT1 and AT2 cells from a common BP. **a**, Hierarchical clustering of multiplexed qPCR gene expression data for 33 single cells from E16.5 lung epithelium ($CD45^-/EpCAM^+$) suggests the presence at this time point of two major cell lineages, bronchiolar (cyan) and alveolar (brown) progenitors. Note that alveolar progenitors express a subset of both AT1 and AT2 marker genes. **b**, PCA of multiplexed qPCR data of lung epithelial cells at E16.5 identifies two gene groups in contrast to three observed at E18.5 (Fig. 1c). AT1 and AT2 specific marker genes do not segregate into distinct populations at E16.5. **c**, Hierarchical clustering of multiplexed qPCR gene expression data for 74 single embryonic lung epithelial cells ($CD45^-/EpCAM^+$) at E18.5 shows multiple distinct cell populations consistent with RNA-sequencing data at this time point: BP, AT1, AT2, Clara and ciliated cells. Each row represents a single cell and each column a gene. Cells are clustered on the basis of expression of marker genes for alveolar and bronchiolar lineages (AT2: *Abca3*, *Sftpb*, *Muc1*, *Lyz2*, *Sftpc*; AT1: *Aqp5*, *Pdpn*, *Ager*; ciliated: *Foxj1*; Clara: *Scgb1a1*). **d**, PCA of multiplexed qPCR data replicates gene families found by single-cell RNA-seq at E18.5. Gene groups were characterized on the basis of differential correlation with the first two principal components. **e**, Developmental sequence of AT1 (orange) and AT2 (red) specification from a common BP (brown). Two and three maturation intermediates were identified in the specification process of AT2 and AT1 cell types, respectively, on the basis of the expression of known and previously unknown marker genes for both alveolar lineages measured by single-cell RNA-seq (Fig. 3). Transcription factors and receptors/ligands shown here were found to be expressed in BP cells and subsequently restricted to one of the alveolar lineages. Arrows, differentiation pathway; grey braces, change in transcript level of respective genes with tip pointing towards lower expression. **f–i**, Protein level heterogeneity of alveolar epithelial markers during sacculation. **f**, Immunofluorescent micrograph from an E19.5 lung with mature AT1 and AT2 cells stained for their respective markers (Pdpn (white) and Ager (red) for AT1; Sftpc (green) for AT2). BPs are positive for all three markers. Cells in intermediate states are observed, such as early AT1 (Pdpn and Ager positive, Sftpc low) and early AT2 cells (Sftpc positive, and either Pdpn positve/Ager low or Pdpn low/Ager negative). Scale bar, 10 μm. **g**, Markers of late AT2 cells are expressed heterogeneously at E18.5. Immunofluorescence micrograph of a lung from a Lyz2–enhanced green fluorescent protein (eGFP) transgenic mouse, in which within the epithelium (E-cadherin, blue) only a subset of Sftpc (green)-positive AT2 cells are Lyz2 (red)-positive. Scale bar, 20 μm. **h**, Immunofluorescent staining of E18.5 lung tissue for Lamp3 (red) shows heterogeneous expression of Lamp3 in Sftpc-positive cells (green): Proximal cells show higher Lamp3 expression than distal cells. Blue, DAPI-stained nuclei. Scale bar, 20 μm. **i**, Immunofluorescent staining of E18.5 lung tissue for S100a6 (red) shows heterogeneous expression of the secreted protein S100a6 in Pdpn-positve cells (green). Blue, DAPI-stained nuclei. Scale bar, 20 μm.



Extended Data Figure 6 | Following *Sftpc*-expressing cells throughout their life cycle. **a**, Whole-mount *in situ* hybridizations of embryonic mouse lungs at E11.5, E13.5 and E14.5 using probes against *Sftpc* mRNA show expression of *Sftpc* specific to the tips of the epithelial tree branches. Moreover, variations in signal intensity indicate heterogeneity in the level of *Sftpc* expression across cells, which is in agreement with our single-cell RNA-seq data of *Sftpc*$^{+}$ cells at E14.5 (see Fig. 4a). **b**, Diagram of the different transcriptional states in the specification of an AT2 cell as identified by single-cell RNA-seq of *Sftpc*$^{+}$ cells from distal mouse lung epithelium of embryonic (E14.5, E16.5 and E18.5) and adult mice. The cell undergoes a transition from an early (A) and late (B) early progenitor state into a BP state before either taking the AT1 fate (nascent AT1), or following the AT2 pathway to become a nascent and finally a mature AT2 cell. Groups of genes turning on/up or off/down during the individual transitions are shown above and below each arrow, respectively (Fig. 4a and Supplementary Data). Whereas EP and BP cells are double positive for *Sftpc* and *Pdpn*, nascent and mature AT2 cells express *Sftpc* but turn off expression of the AT1 marker *Pdpn*. The developmental time points at which the individual cell states were detected, and their putative locations, are shown.



Extended Data Figure 7 | The number of unique genes and the total number of transcripts expressed by a single cell strongly correlates with its differentiation state. **a**, Saturation analysis of single-cell RNA-seq data of lung epithelial cells at different embryonic and adult time points (E14.5, E18.5 and adult AT2) reveals that the number of unique genes expressed by single lung epithelial cells decreases with progressing differentiation state. Distal lung epithelial cells at E14.5 express more than 6,000 genes, whereas cells at E18.5 express about 3,000 genes, and mature AT2 cells only about 2,000 genes. Each point on the saturation curve was generated by randomly selecting a number of raw reads from each sample library and then using the same alignment pipeline to call genes with a mean FPKM of more than 1. Each point represents four replicate subsamplings. Error bars represent s.e.m. All libraries were sequenced to a depth of at least 2×10^6 reads. **b**, Single-cell RNA-seq reveals that the total number of transcripts expressed by single cells decreases with increasing differentiation state of the cell. The number of transcripts per cell was calculated from the FPKM values of all genes in each cell, using the correlation between number of transcripts of exogenous spike-in mRNA sequences and their respective measured mean FPKM values (example calibration curves are shown in Extended Data Fig. 3c for three replicates at E18.5). Area-normalized density distributions are shown for embryonic cells at E14.5 (45 cells), E16.5 (27 cells) and E18.5 (80 cells), and for 46 *Sftpc*$^+$ adult AT2 cells. The number of transcripts is highest in lung epithelial progenitor cells at E16.5 and E14.5 and decreases in cells at E18.5 and even further in mature AT2 cells. Note that single-cell RNA-seq libraries for E14.5, E18.5 and adult AT2 cells were sequenced to a depth of (2–6) $\times 10^6$ reads, whereas the libraries for cells at E16.5 were sequenced to a lower depth of 100,000–550,000 reads. **c**, Calibration of C_t values measured by single-cell qPCR to number of molecules. Average detected transcript levels ($\log_2 Ex = C_{t,LoD} - C_t$, $C_{t,LoD} = 22$) for six ERCC RNA spike-ins as a function of provided number of molecules per lysis reaction for each of three independent single-cell qPCR experiments performed on embryonic (E16.5, two replicates; red and green) and adult mouse lung (adult AT2, one replicate; blue). Linear regression fits through data points and corresponding equations are shown and were used to convert C_t values measured by qPCR into numbers of transcripts. **d**, Single-cell qPCR confirms the presence of a higher number of transcripts in lung epithelial progenitor cells in comparison with fully differentiated alveolar epithelial cells. The median number of transcripts per cell as detected by single-cell RNA-seq (*y* axis) and by single-cell multiplexed qPCR of 90 genes (*x* axis) is shown for distal lung epithelial cells at E16.5 (qPCR, 33 cells; RNA-seq, 27 cells) and mature AT2 cells (qPCR, 48 cells; RNA-seq, 46 cells).



Extended Data Figure 8 | Transcriptional states during the early lifetime of the Clara cell lineage identified by single-cell RNA-seq of *Scgb3a2*$^{+}$ cells at E14.5, E16.5 and E18.5. **a**, Hierarchical clustering of 24 *Scgb3a2*-positive cells from distal mouse lung epithelium at different embryonic time points (E14.5, E16.5 and E18.5) based on the genes with highest principal-component loadings in an unbiased PCA analysis of all cells and all genes (shown in **c**). Cells are shown in rows, genes in columns. Cells cluster into three major groups. *Scgb3a2* and *Scgb1a1* transcript levels are shown in bars on the right. Whereas canonical Clara cell marker *Scgb1a1* is first detected at E18.5, *Scgb3a2* is detected as early as E14.5, suggesting that it is an early Clara cell marker. **b**, Gene Ontology (GO) enrichments of the three different gene clusters as well as transcription factors (TFs) belonging to the different groups of genes. **c**, PCA analysis of all *Scgb3a2*-positive cells and all genes identifies three different cell populations that were identified as bronchiolar progenitor as well as Clara and ciliated cells.