HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2014 November 15.

Published in final edited form as:

Nature. 2014 May 15; 509(7500): 371–375. 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,+], and **Stephen R. Quake**[1,+]

[1]Departments of Bioengineering and Applied Physics, Stanford University and Howard Hughes Medical Institute, Stanford, CA 94305

[2]Department of Biochemistry, Stanford University and Howard Hughes Medical Institute, Stanford, CA 94305

[3]Department of Medicine, Stanford University, Stanford, CA 94305

Abstract

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 remain obscure, in part due to the limited number of lineage markers[1-3] and the effects of ensemble averaging in conventional transcriptome analysis experiments on cell populations[1–5]. We used microfluidic single cell RNA sequencing (RNA-seq) on 198 individual cells at 4 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 using an unbiased genome-wide approach that did not require *a priori* knowledge of the underlying cell types or prior 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 novel cell type markers and 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 lifecycle of the alveolar type 2 cell lineage. This single cell genomics approach is applicable to any developing or mature tissue to robustly

[+]To whom correspondence should be addressed. quake@stanford.edu, krasnow@stanford.edu, tdesai@stanford.edu.
[*]These authors contributed equally to this work

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. analyzed 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.

Conflict of Interest: S.R. Quake is a founder and consultant for Fluidigm Corporation.

Accession codes

Gene Expression Omnibus: GSE52583.

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 low 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 Figure 1). At each time point during sacculation, progenitors, intermediates, and recently differentiated cells coexist (Figure 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 (embryonic day E18.5, 3 biological replicates). Single cell suspensions of micro-dissected distal lung regions were purified using magnetic-activated cell sorting (MACS) to deplete leukocytes and alveolar macrophages and enrich for epithelial cells ($CD45^{-}/EpCAM^{+}$) (Extended Data Figure 2). An automated microfluidic platform was used to capture and lyse individual epithelial cells, reverse transcribe RNA, and amplify cDNA.

RNA-seq libraries from the amplification products of single cells as well as bulk control samples were sequenced to a depth of 2-5 million reads per library (Methods). Saturation analysis confirmed that this sequencing depth is sufficient to detect most genes expressed by single cells (Extended Data Figure 3a). Technical noise and dynamic range were assessed using RNA control spike-in standards and by comparing single cells with the bulk samples (Extended Data Figure 3b-e). The results are consistent with previous data from our group[7] and others[8–20]; we obtained single transcript sensitivity and high (~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 Figure 3f-g).

We performed principal component analysis (PCA) on all 80 single cell transcriptomes using genes expressed in more than two cells and with a non-zero variance (8578 genes). Genes with highest loadings in the first four principal components were analyzed by unsupervised hierarchical clustering as well as PCA (Figure 1b-c, Figure 2a, Supplementary Data 3). 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 transitioning 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 using microfluidic single cell qPCR experiments: hierarchical clustering of 10 known alveolar and bronchiolar marker genes identified the

same five populations (Extended Data Figure 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 novel 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 (Figure 2b, Supplementary Data 4, Methods). 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 Figure 4a, Supplementary Data 5): AT1 cells were enriched in pathways associated with ECM-receptor interaction, focal adhesion, tight and adherens junction as well as regulation of actin cytoskeleton; AT2 cells were enriched for adipocytokine and PPAR signaling as well as lysosome pathways; the Clara cell lineage was enriched for metabolism of xenobiotics by cytochrome P450, drug metabolism as well as 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 strongly correlated with the individual cell types (Extended Data Figure 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 AT1 specific expression of *Hopx* by transgenic labeling and colocalization with two AT1 markers, *Pdpn* and *Ager* (Figure 2c, Extended Data Figure 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 Figure 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 marker *Sftpc* (Figure 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 (Figure 2e). Finally, we used single cell multiplexed qPCR to validate 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 Figure 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 or low expressed (Extended Data Figure 4f). In summary, 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]. Therefore, we measured transcript levels of 10 known marker genes in 107 single cells of the distal lung epithelium at both E16.5 (33 cells) and E18.5 (74 cells) using multiplexed single cell qPCR (Extended Data Figure 5a-d). 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, as marker genes for both subpopulations were commonly expressed by all cells, while 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 based on partial coexpression of AT1 and AT2 marker genes. We used the newly identified genes specific for each mature alveolar cell type to sub-classify 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 (Figure 3). We confirmed the presence of developmental intermediates showing heterogeneity in marker gene expression by immunofluorescence (Extended Data Figure 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 Figure 5e). The transcription factors *Sox9* and *Cited2* were expressed in BP and AT2 cells, whereas *Hes1* was expressed in BP and AT1 cells. Interestingly, we did not detect any transcription factors that initiated expression exclusively in either of the maturing alveolar lineages, suggesting that lineage commitment involves down-regulation of factors active in alveolar progenitor cells rather than *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 either a BP (*Fzd2*), BP/AT2 (*Fgfr2*) or BP/AT1 (*Gprc5a*) pattern. These results show that our approach can be utilized 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 prior to (E14.5, 45 progenitor cells), early in (E16.5, 27 progenitor cells), and long after (adult, 46 transgenically labeled AT2 cells) sacculation. We performed unsupervised hierarchical clustering analysis of *Sftpc*-positive cells (124 cells) using genes with the highest PC loadings in a PCA analysis (Figure 4a,b, Supplementary Data 6). Cells clustered in groups that highly correlated with developmental stage of cell isolation, in sequence from early progenitors (EP), via BP and nascent AT1 and AT2 cells, to mature AT2 cells. Thus, AT2 cell maturation occurs in a progressive manner via transcriptionally distinct intermediates that can be robustly discriminated by expression profile throughout embryonic and adult life. The population of EP cells co-expresses AT2 marker *Sftpc* and AT1 marker *Pdpn*, indicating that these cells are located at the tips of the branching

epithelial tree (Extended Data Figure 6a). EP cells segregate into two sub-groups, 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, Figure 4a,c), which are downregulated during the transition of EPs to BPs. The down-regulated EP-specific genes include transcription factor *Sox11*, which is expressed in the developing airway epithelium and causes an alveolar defect when mutant[23], as well as *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 reduced expression of AT1 markers (nascent AT2) or AT2 markers (nascent AT1). Mature AT2 cells are characterized by expression of genes involved in respiratory gas exchange and immune response (gene group IV) and were only detected at adult stages (Figure 4a). Interestingly, 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 6000 genes, whereas mature AT2 cells expressed about 4-6 times fewer genes (Figure 4a and Extended Data Figure 7a). In summary, we followed the full lifecycle of *Sftpc*$^{+}$ cells and identified seven gene sets that robustly distinguish multipotential, bipotential, nascent, and mature AT2 cell states (Extended Data Figure 6b).

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

Online-only Methods

Mouse strains

Timed-pregnant C57BL/6J females (JAX) were used for all embryonic time points reported with gestation age verified by crown-rump length prior to use. For adult mice, a transgenic-labeling 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) and heterozygous for an inserted transgene, which drives the expression of a GFP tagged human histone 1 in a tetracycline-dependent manner (tetO-HIST1H2BJ/GFP). To validate 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) and heterozygous for a transgenic insertion into the Rosa26 locus encoding a two color, membrane tethered fluorophore reporter that switches expression from a red to green fluorophore upon 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 ethanol precipitation. Mice were housed in filtered cages and all experiments were performed in accordance with approved IACUC protocols.

Isolation and disaggregation of lung tissue

Single cell experiments were performed on embryonic mouse lung at E14.5, 16.5, and 18.5 as well as on adult mouse lung. In general, embryonic experiments were performed on pooled sibling lungs of one litter (5-7 lungs per pool). One of three replicate experiments at E18.5 (cells referred to as "E18_2_Cxx" in Supplementary Data 1 and 3) was performed on a single embryonic lung.

Adult mice were euthanized by CO_2 administration. For time points E14.5, 16.5, and 18.5, embryos were removed and lungs isolated *en bloc* without perfusion and pooled by litter (5-7 embryos) for further processing. Lungs from E14.5 and 16.5 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; Worthington Biochemical Corporation) and DNase I (0.33 U/mL; Roche) in DMEM/F12, incubated with frequent agitation at 37 °C for 45 minutes, 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, Thermo Scientific) was added to single cell suspensions prior to passing the suspension through a 100 μm mesh filter (Fisher) and centrifugation at 400×g for 10 minutes. Pelleted cells were resuspended in red blood cell lysis buffer (BD Biosciences), incubated for 2 minutes, passed through a 40 μm mesh filter (Fisher), centrifuged at 400×g for 10 minutes and then resuspended in sorting buffer (PBS supplemented with 0.05% BSA and 2 mM EDTA).

Purification of embryonic distal lung epithelial cells by magnetic-activated cell sorting (MACS)

Lung epithelial cells for embryonic time points (E14.5, E16.5, E18.5) were purified by magnetic-activated cell sorting (MACS) using MS columns (Miltenyi Biotec) in MACS buffer (2mM EDTA, 0.5% BSA in PBS, filtered and degassed) according to the protocol provided by the vendor. Prior to 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 (eBioscience, clone G8.8) followed by a secondary antibody against biotin conjugated to magnetic beads (Miltenyi Biotec).

Purification of adult AT2 cells by fluorescence-activated cell sorting (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) and sacked 3 days later. After incubation of the single cell suspension with a viability stain (Sytox Blue, Invitrogen) for 15 minutes, 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 all distal lung epithelial cells, an initial experiment was performed using a microfluidic chip with a 17-25μm 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 Figure 2b). Cells were loaded onto the chip at a concentration of 300-500 cells/μl, stained for viability (LIVE/DEAD cell viability assay, Molecular Probes, Life Technologies) and imaged by phase-contrast and fluorescence microscopy to assess 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 mRNA. For qPCR experiments, amplicons were prepared using 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 analyzer (Advanced Analytical). Illumina libraries were constructed in 96 well plates using the Illumina Nextera XT DNA Sample Preparation kit according to 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 quantitated by Agilent Bioanalyzer using High Sensitivity DNA analysis kit as well as 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 bp paired-end on Illumina HiSeq 2000 to a depth of 2-6 million reads (3 replicate experiments of distal mouse lung epithelial cells at embryonic day 18.5 (E18.5), 1 experiment at E14.5, 1 experiment on adult AT2 cells) or 150 bp paired end on Illumina MiSeq (1 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 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[1] 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,

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 Ct values to approximate numbers of transcripts in a subset of 90 genes (Extended Data Figure 7).

Processing, analysis and graphic display of single cell RNA-seq data

Raw reads were pre-processed with sequence grooming tools FASTQC[2], cutadapt[3], and PRINSEQ[4] followed by sequence alignment using the Tuxedo suite (Bowtie[5] Bowtie2[6] TopHat[7] and SAMtools[8] using default settings. See Supplementary Data 1 for information about number of total reads and 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 randomly select raw reads 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 analyzing the correlation between concentration of exogenous ERCC spike-in mRNA sequences and their respective mean FPKM values as measured by RNA-seq (Extended Data Figure 3c). The spike-in sequences reflect a diverse range of sequence content and length, have low homology with eukaryotic transcripts as they are from microbial sources, and they span a large range of concentrations to allow empirical determination of the limit of detection[9]–[11]. 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 in $\log_2$ scale). Therefore, transcripts with an FPKM value below or equal to 1 were considered not expressed. Cells not expressing either of two housekeeping genes *Actb* and *Gapdh*, or expressing them below 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 3 replicate experiments at E18.5 (2x pooled sibling lungs (20 and 26 cells), 1x 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 1017-4998 expressed genes per single cell, 10,946 in the union of all single cells and 8653 in the 200 cell control bulk sample, indicating the heterogeneity of the analyzed single cells. 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 approximate number of transcripts using the correlation between number of transcripts of exogenous spike-in mRNA sequences and their respective measured mean FPKM values (Extended Data Figure 3c). The number of spike-in transcripts per single cell lysis reaction was calculated using the concentration of each spike-in provided by the vendor (Ambion, Life Technologies), the approximate volume of the lysis chamber (10 nl) as well as the dilution of spike-in transcripts in the lysis reaction mix (40,000x). Transcript levels were converted to the log-space by taking the $\log_2$(FPKM/ number of transcripts). When calculating the “average” single cell expression (Extended

Data Figure 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 Figure 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 Figure 3a and 7a) were generated as previously described[11]. Briefly, 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 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[12] were used to perform principal component analysis (PCA), hierarchical clustering, Guild-by-Association and permutation analysis as well as to construct violin plots, correlation plots and histograms. The scripts can be found as Supplementary Data 2. 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 PC loadings (highest absolute correlation coefficient with one of the first three to 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 Figure 2a identifying five distinct cell populations was assessed using a permutation analysis approach. Therefore, the sum of squares within groups (SSW) was calculated for the cell grouping presented in Figure 2a as well as for 50,000 random permutations thereof, keeping the size of cell groups and the total number of groups constant. With $x_{i,j}^k$ being the transcript level of gene j in cell i belonging to group k, the SSW can be calculated as

$$WSS = \sum_{k=1}^{n} \sum_{j=1}^{m} \left(x_{i,j}^k - {}_i\overline{x}_j^k \right)^2,$$

with ${}_i\overline{x}_j^k = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}^k$ 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 all other permutations (p-value = 2.89×10^{-122}).

When *Sftpc*$^+$ cells were isolated from all single cell RNA-seq data sets (Figure 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 novel 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 no expression (FPKM = 0) in all other cells. We then determined the pair-wise 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 was screened for cell type specificity was obtained from the online animal transcription factor database http://www.bioguo.org/AnimalTFDB/[13]. 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 4 together with information about the top 30 genes per cell type regarding previous detection in cell types in the lung, available literature or known mouse knock-out phenotypes.

Guild-By-Association analysis[14] was used to calculate the probability to observe a given coexpression of two genes by chance. Therefore, gene expression values were scaled gene-by-gene by mean-centering 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. Pair-wise comparisons were performed between the perfect marker gene for each of the 4 mature cell types (AT1, AT2, Clara, Ciliated) and all other genes expressed in at least one cell (10,946 genes in total). P-values were calculated using the hypergeometric distribution as described by Walker et al.[14] and multiple testing was accounted for using the Benjamini-Hochberg method (Figure 2b, Supplementary Data 4).

Gene ontology and KEGG pathway enrichment analyses were performed using DAVID informatics Resources 6.7 of the National Institute of Allergy and Infectious Diseases, NIH[15,16] (Supplementary Data 5 and 6).

Analysis and graphic display of microfluidic single cell multiplexed qPCR data

Single-cell multiplexed qPCR data was analyzed and displayed using custom R scripts[12]. qPCR experiments were performed for E16.5 (2 biological replicates), E18.5 (2 biological replicates) distal lung epithelial cells and for adult AT2 cells (1 replicate). The limit of detection of multiplexed qPCR values was determined as 22 threshold cycles (Ct) by a calibration experiment with 16-fold serial dilutions of total lung cDNA and 6 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 Ct values to transform Ct values to $\log_2$ expression values ($\log2Ex = Ct_{LoD} - Ct$, $Ct_{LoD} = 22$). Cells not expressing either of two housekeeping genes *Actb* and *Gapdh*, or expressing them below 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 (5-9 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-centering 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 10 canonical marker genes for bronchiolar and alveolar cells (*Abca3, Sftpb, Muc1, Sftpc, Lyz2, Aqp5, Pdpn, Ager, Foxj1, Scgb1a1*).

Immunofluorescence

E18.5 lungs were removed *en bloc* and for whole mount staining, the tip of the accessory lobe was excised. Lungs and tips were immersion fixed in PFA (4% in PBS) overnight at 4°C, and then dehydrated and stored in methanol at −20°C until staining. Lungs of adult mice were collected as above except that following 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 (Sigma, 2% in PBS). Ice-cold PBS was dripped into the thorax to solidify the agarose, inflated lungs were removed *en bloc* and processed as above. E18.5 lungs were rehyrdated, cryoprotected in 30% sucrose overnight at 4°C, submerged in OCT (Tissue Tek) in an embedding mold, frozen on dry ice, then stored at −80°C. Sections of 10 μm thickness were obtained using a cryostat (Leica CM3050S) were collected on chambered glass slides and stored at 4°C prior to staining.

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 minutes, PBS + 0.3% Triton X-100), washed (3X 5 minutes, PBS + 0.1% Tween 20), and blocked (1 hour, PBS+10% donkey serum) before overnight primary antibody incubation. 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, R&D, AF4584), S100a6 (rat, dendritics, DDX0192) and Krt15 (mouse, SCBT LHK15) directly conjugated to a fluorophore following manufacturer's instructions (Alexa Fluor Antibody Labeling 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) for 1 hour, followed by washing and mounting in Vectashield (Vector). Lung tissues were imaged using a laser scanning confocal microscope (LSM 780, Zeiss).

I*n situ* hybridisation

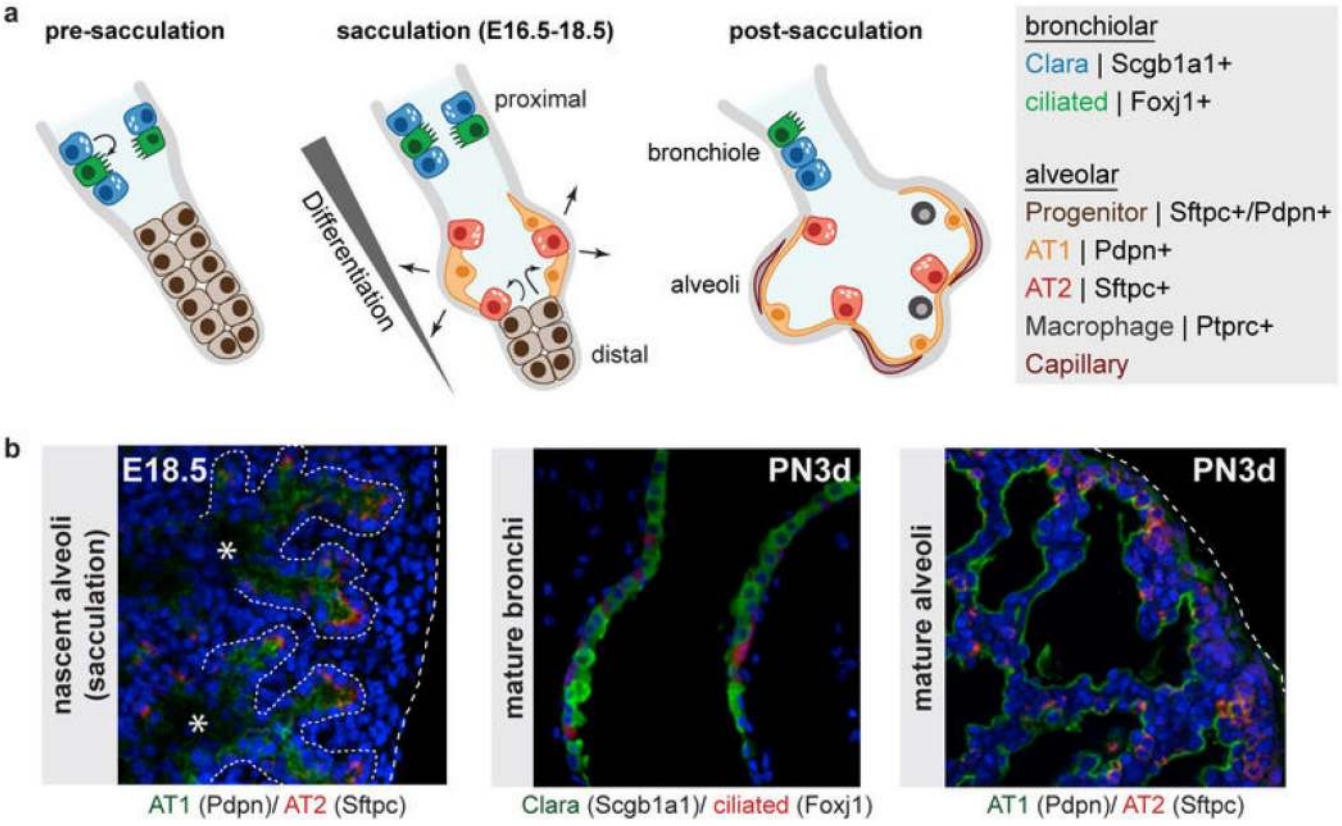
For *in situ* hybridisations, embryonic lungs were collected as for immunostaining (see above), washed briefly in PBS (autoclaved, DEPC-treated) and snap frozen in OCT for sectioning. Sections of 10 μm were generated on the cryostat and stored at -20°C before further processing. To validate AT2-specific 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, following the company's reported protocol (ACD's RNAscope® In Situ Hybridization Technology). To validate 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[17].

Validation of *Hopx* as novel AT1 marker gene by transgenic labeling

Cells actively transcribing Hopx in the adult mouse lung were labeled by injecting 2 mg of tamoxifen (Sigma) in corn oil at 20 mg/ml concentration intraperitoneally into postnatal day 28 Hopx-Cre-ERT2$^{+/-}$ mTmG$^{+/-}$ mice. 3 days later, lungs were collected as described above, fixed in PFA (4% in PBS) overnight at 4°C and stored in 80% glycerol at 4°C prior to imaging using a laser scanning confocal microscope (LSM 780, Zeiss) with a 0.8 NA, 25x oil-immersion objective and confocal z-sections with a thickness of 2.3 μm.

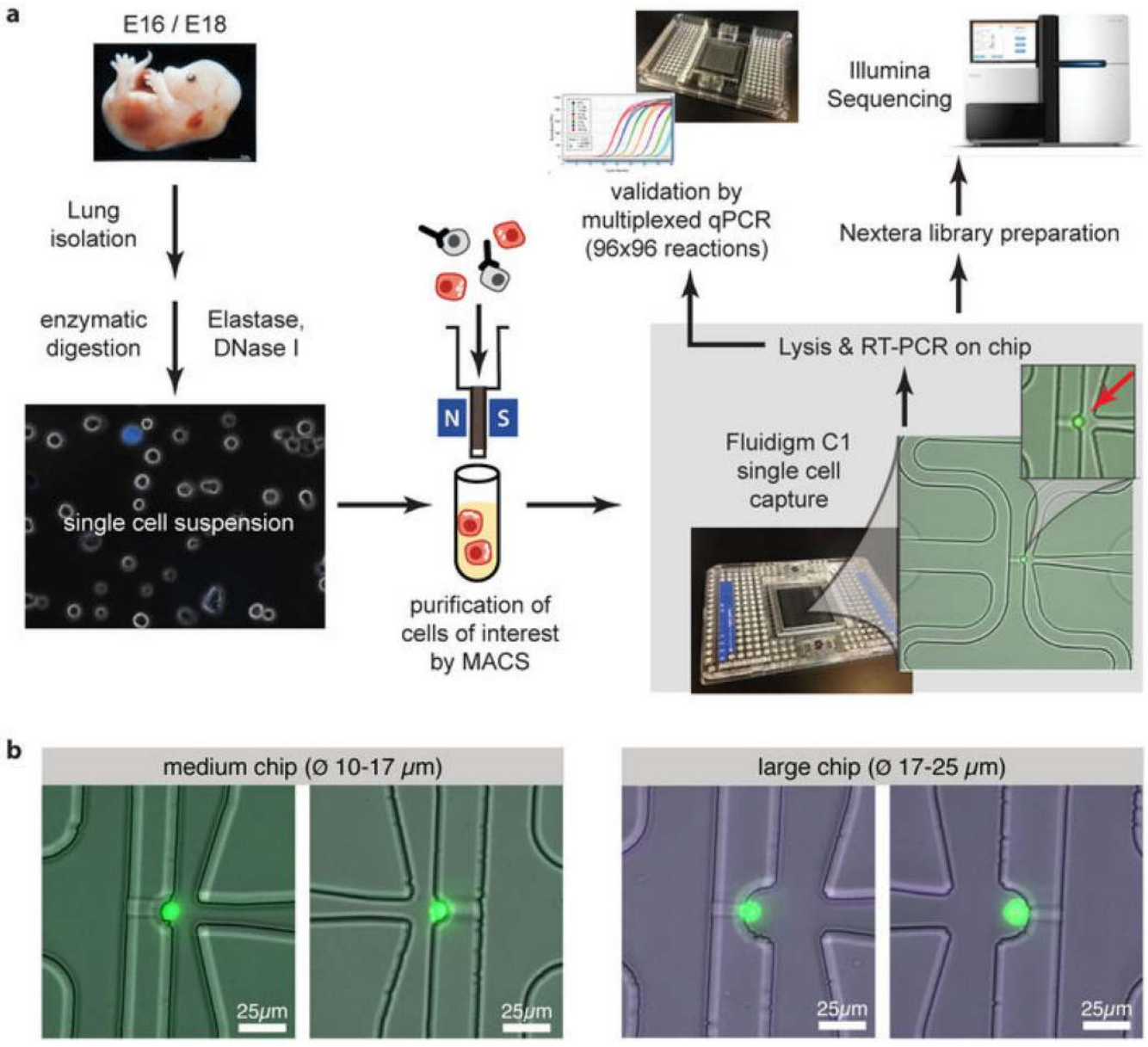
Extended Data



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, and therefore progenitor cells persist the longest at the tips. Ciliated (green) and Clara (blue) cells mature first, followed by differentiation of flat alveolar type 1 (AT1, orange) and cuboidal type 2 (AT2, red) cells from cuboidal alveolar progenitors during sacculation (embryonic day (E) 16-18.5), when distal airway tubules widen as nascent AT1 cells flatten to form the gas exchange surface.

(b) Micrographs of alveolar (E18.5, post-natal 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.

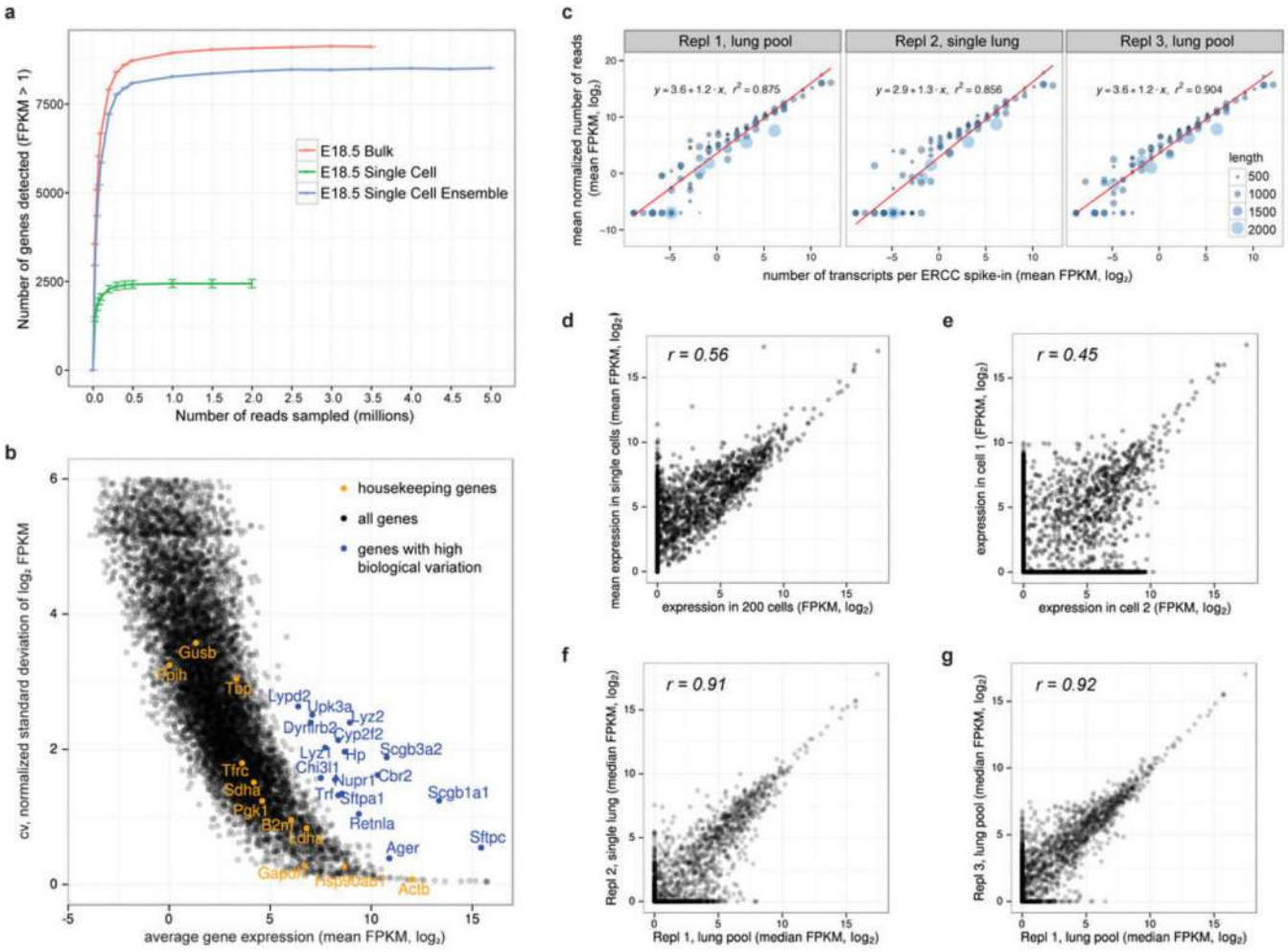


Extended Data Figure 2. Single cell transcriptomics analysis workflow

(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 10-17 μm (medium, left) or 17-25 μm (large, right) diameter. Cells are stained

for viability using 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

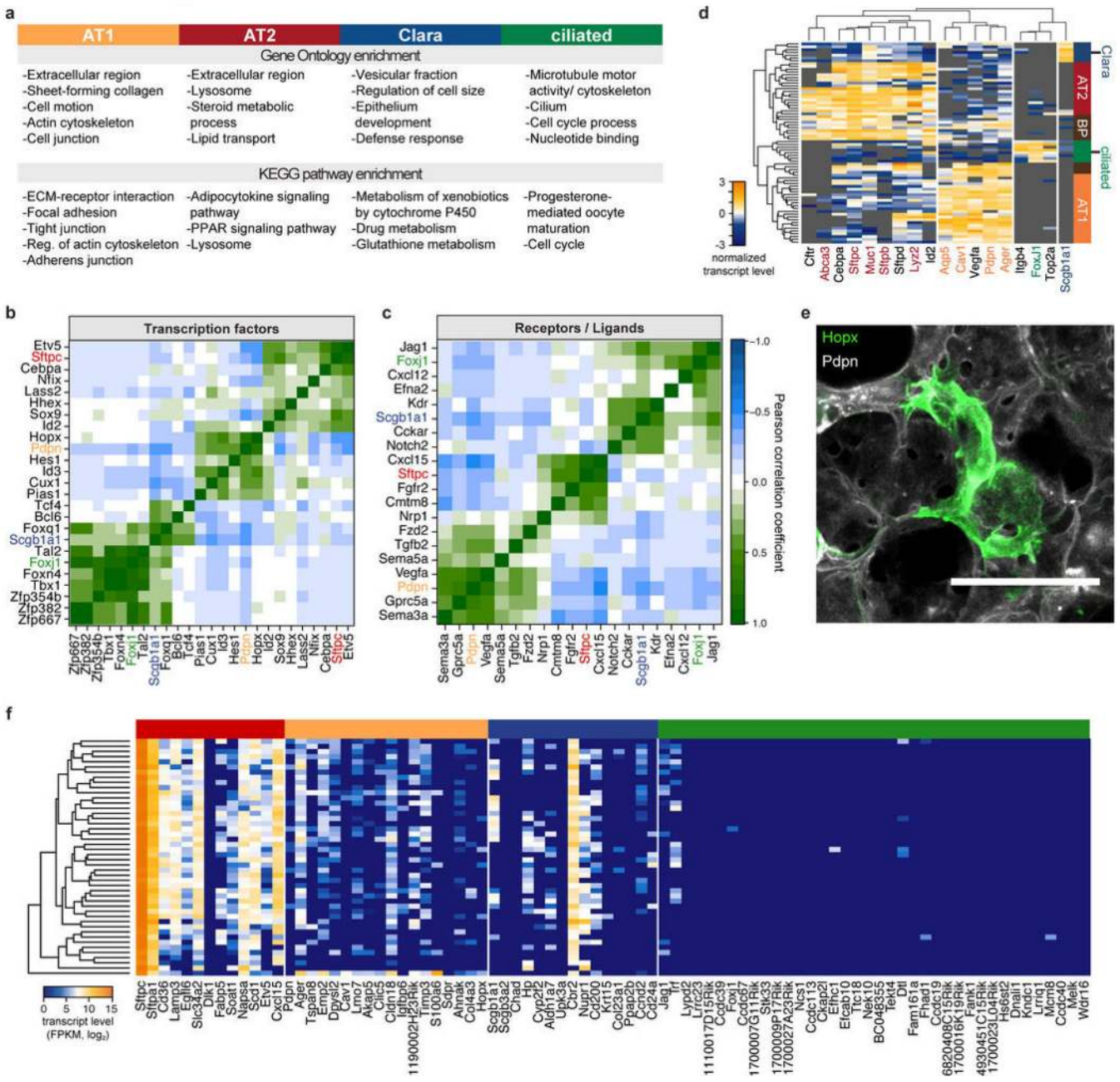
(a) Saturation analysis reveals the sequencing depth required for the detection of most genes expressed by single cells. To detect most expressed genes, single cell RNA-seq libraries have to be sequenced only to a depth of about 1 million reads, whereas libraries of bulk samples have to be sequenced deeper. 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 mean FPKM >1. Each point represents four replicate sub-samplings, error bars represent standard errors.

(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 exhibit strong biological variation (blue), as they exhibit 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 color 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 6 orders of magnitude, in agreement with a previous study evaluating microfluidic single cell RNA-seq[7].

(d,e) Correlation between (d) transcript levels of a 200-cell population and median transcript levels of single cells of the same pool of embryonic lungs, and (e) transcript levels of two single AT2 cells. *r*, Pearson correlation coefficients.

(f,g) Correlation between (f) transcript levels of all genes detected in the single lung and the pooled lung experiment and between (g) transcript levels of all genes detected in the two independent experiments on pooled embryonic lungs. 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

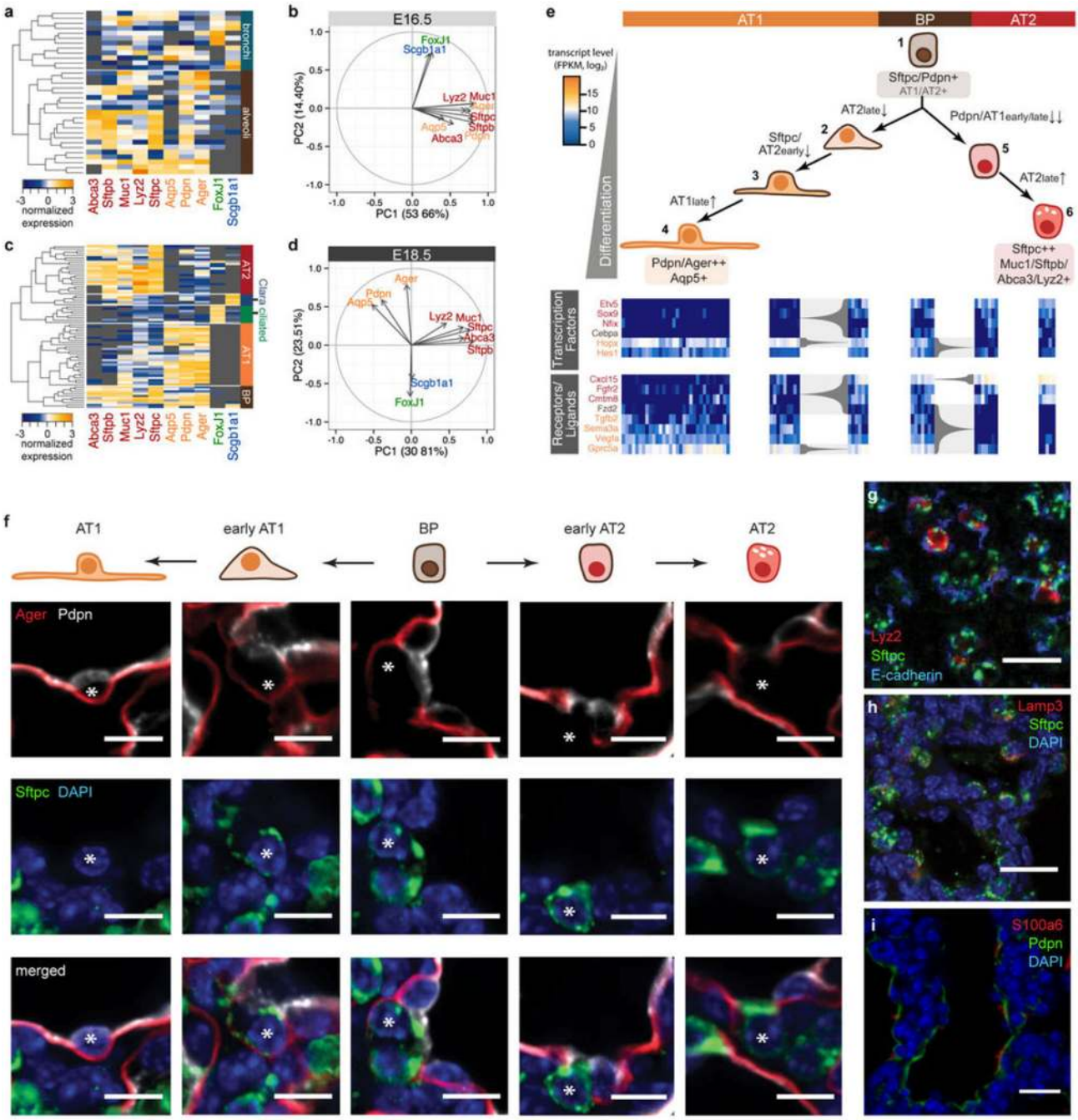
(a) Results of gene ontology (GO) and KEGG pathway enrichment analyses for distal lung epithelial cell types based on lineage-specific genes identified by single cell RNA-seq of 80 E18.5 distal lung epithelial cells (Supplementary Data 5).

(b,c) Correlograms visualizing correlation of single cell gene expression profiles between (b) transcription factors or (c) receptors/ligands and the major canonical marker genes for bronchiolar and alveolar lineages (AT1: *Pdpn*, AT2: *Sftpc*, Clara: *Scgb1a1*, ciliated: *Foxj1*). The color bar denotes the Pearson correlation coefficient from −1 (blue, anticorrelated genes) to 1 (green, positively correlated genes).

(d) Validation of novel 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, novel: *Cftr*, *Cebpa*, *Sftpd*, *Id2*), (AT1, orange, novel: *Vegfa*), (ciliated, green, novel: *Itgb4*, *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 labeled 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.

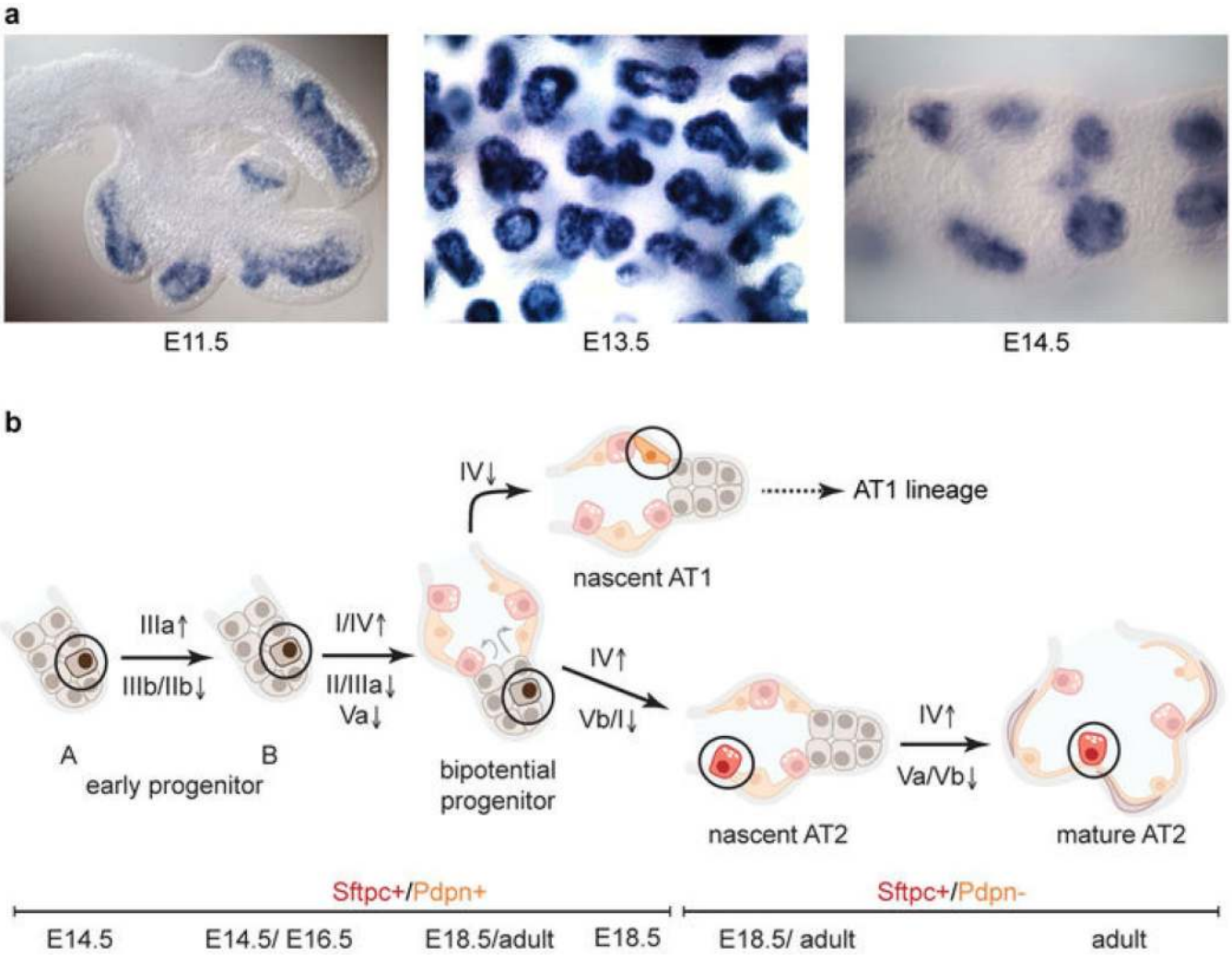


Extended Data Figure 5. Molecular profiles distinguish developmental intermediates during the differentiation of AT1 and AT2 cells from a common bipotential progenitor

(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 (figure 1c). AT1 and AT2 specific marker genes do not segregate into distinct populations at E16.5.

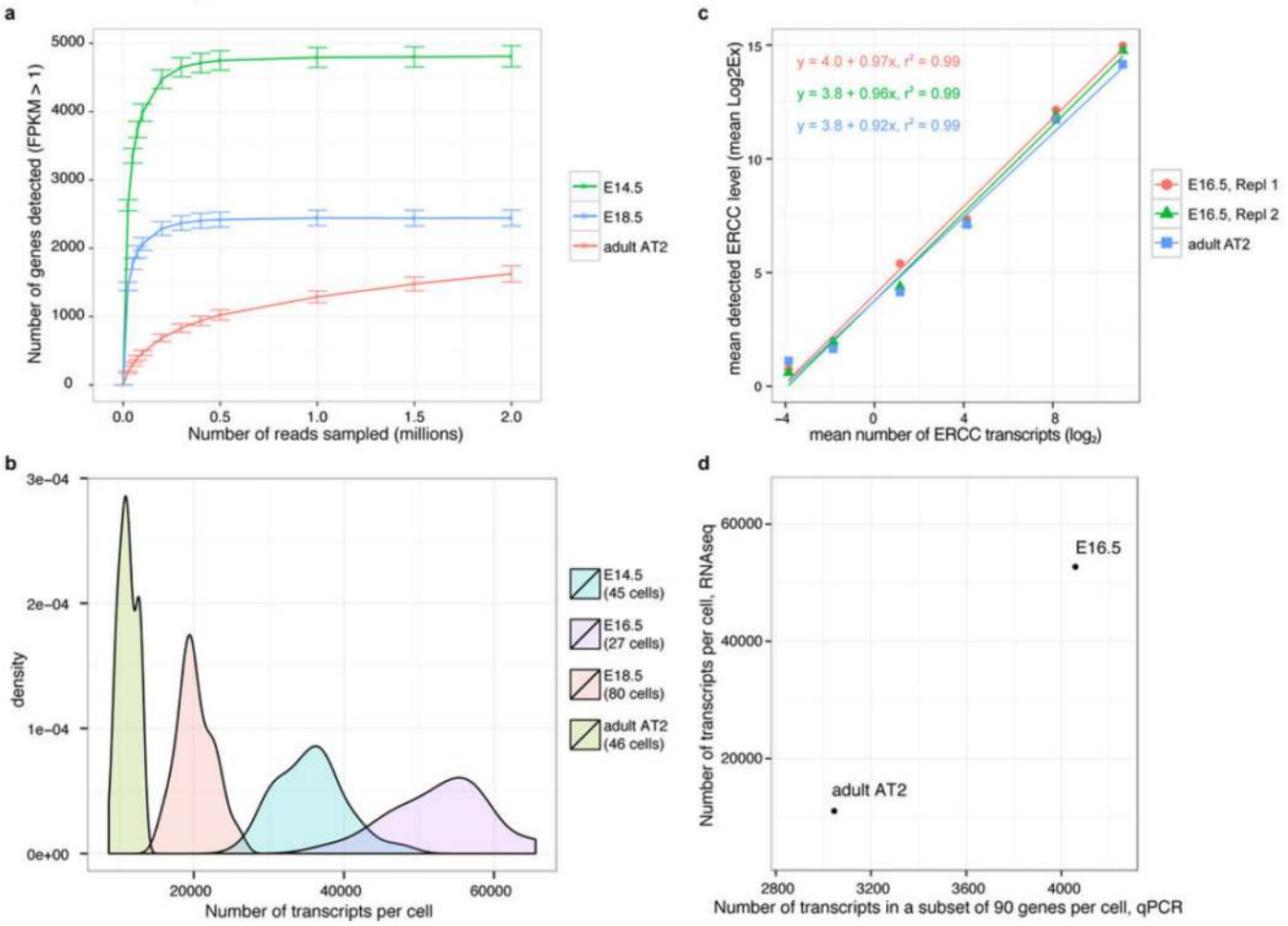
(c) Hierarchical clustering of multiplexed qPCR gene expression data for 74 single embryonic lung epithelial cells ($CD45^{-}/EpCAM^{+}$) at E18.5 shows multiple distinct cell populations consistent with RNA-sequencing data at this time point: BP, AT1, AT2, Clara and ciliated cells. Each row represents a single cell and each column a gene. Cells are clustered based on 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 based on 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, based on the expression of known and novel marker genes for both alveolar lineages measured by single cell RNA-seq (Figure 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; gray 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, 10μm. (g) Markers of late AT2 cells are heterogeneously expressed at E18.5. Immunofluorescent micrograph of a lung from a Lyz2-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-positve 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 lifecycle

(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 Figure 4a).

(b) Schematic 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, E18.5) and adult mice. The cell transitions from an early (A) and late (B) early progenitor state into a bipotential progenitor state before either taking the AT1 fate (nascent AT1), or following along 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 (Figure 4a and Supplementary Data 6). 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 putative locations are shown.



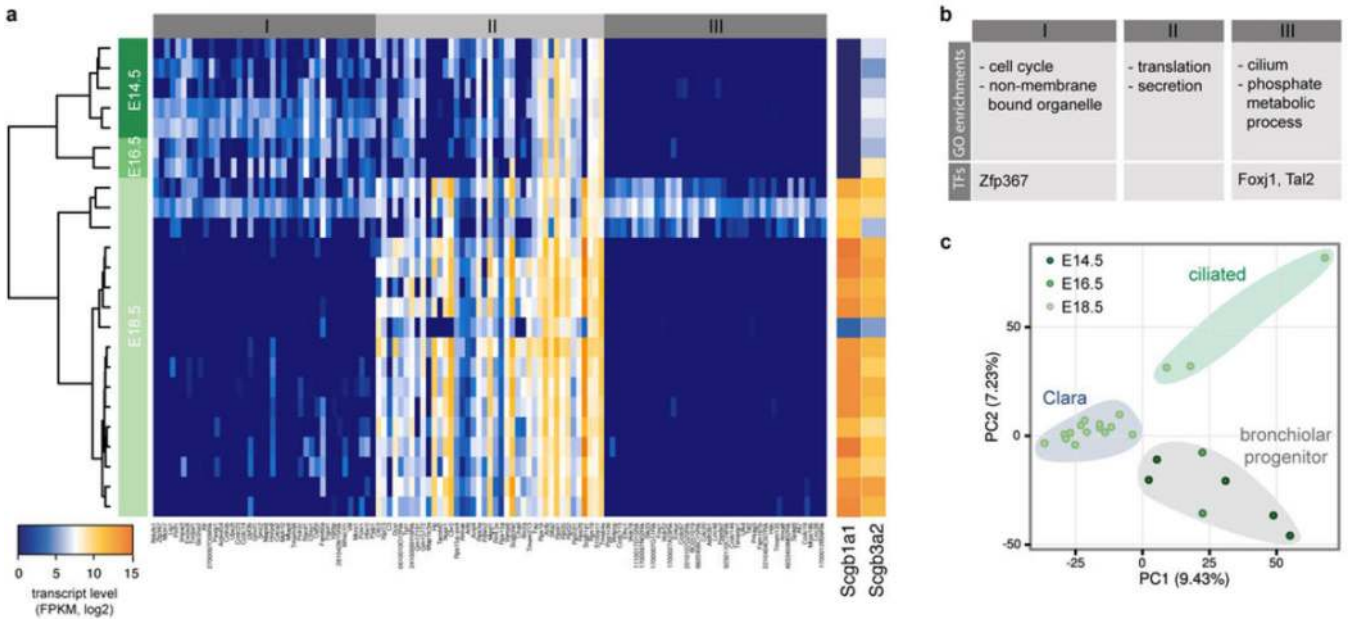
Extended Data Figure 7. The number of unique genes and the total number of transcripts expressed by a single cell strongly correlates with its differentiation state

(a) Saturation analysis of single cell RNA-seq data of lung epithelial cells at different embryonic and adult time points (E14.5, E18.5, 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 over 6000 genes, whereas cells at E18.5 express approximately 3000 and mature AT2 cells only around 2000 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 mean FPKM >1. Each point represents four replicate sub-samplings. Error bars represent standard errors. All libraries were sequenced to a depth of at least 2 million 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. Number of transcripts per cell were 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 (exemplary calibration curves are shown in Extended Data Figure 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), 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 million reads, whereas the libraries for cells at E16.5 were sequenced to a lower depth of 100,000-550,000 reads.

(c) Calibration of Ct values measured by single cell qPCR to number of molecules. Average detected transcript levels (log2Ex = Ct_{LoD} − Ct, Ct_{LoD} = 22) for 6 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, 2 replicates, red and green) and adult mouse lung (adult AT2, 1 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 as compared to 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, E18.5) based on the genes with highest PC loadings in an unbiased PCA analysis of all cells and all genes (panel c). Cells are shown in rows and genes are shown in columns. Cells cluster into 3 major groups. *Scgb3a2* and *Scgb1a1* transcript levels are shown in side-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 to be an early Clara cell marker.

(b) Gene Ontology (GO) enrichments of the three different gene clusters as well as transcription factors (TF) 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to acknowledge W. Koh and B. Passarelli for help and discussions regarding bioinformatic pipelines and statistical analysis, S.I. Gonzalez for help with immunofluorescence as well as J.G. Camp and members of the Krasnow lab for critical discussion and reading of the manuscript.

This work was supported by an NHLBI U01 Progenitor Cell Biology Consortium grant (B.T., M.K.), by NIH T32HD007249 (D.G.B.), a Parker B. Francis Foundation Fellowship and NIH 5KO8HL084095 Award (T.D.), and by NIH U01HL099999-01 (A.R.W., N.F.N). M.K. and S.Q.R. are investigators of the Howard Hughes Medical Institute.

References

1. Kim CFB, et al. Identification of bronchioalveolar stem cells in normal lung and lung cancer. Cell. 2005; 121:823–835. [PubMed: 15960971]
2. Zemke AC, et al. Molecular staging of epithelial maturation using secretory cell-specific genes as markers. Am J Respir Cell Mol Biol. 2009; 40:340–348. [PubMed: 18757308]
3. Guha A, et al. Neuroepithelial body microenvironment is a niche for a distinct subset of Clara-like precursors in the developing airways. Proceedings of the National Academy of Sciences. 2012; 109:12592–12597.
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. 2005; 288:L179–L189. [PubMed: 15447939]
5. Xu Y, et al. Transcriptional Programs Controlling Perinatal Lung Maturation. PLoS ONE. 2012; 7:e37046. [PubMed: 22916088]
6. Desai TJ, Brownfield DG, Krasnow MA. Alveolar progenitor and stem cells in lung development, maintenance and cancer. Nature. 2014:1–16.
7. Wu AR, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nature Methods. 2013
8. Islam S, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011; 21:1160–1167. [PubMed: 21543516]
9. Islam S, et al. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. Nature Protocols. 2012; 7:813–828. [PubMed: 22481528]
10. Shalek AK, et al. Single-cell transcriptomics reveals bimodality inexpression and splicing in immune cells. Nature. 2013; 498:236–240. [PubMed: 23685454]
11. Sasagawa Y, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. Genome Biol. 2013; 14:R31. [PubMed: 23594475]
12. Liu CL, Bernstein BE, Schreiber SL. Whole Genome Amplification by T7-Based Linear Amplification of DNA (TLAD): II. Second-Strand Synthesis and In Vitro Transcription. CSH Protoc. 2008; 2008 pdb prot5003.
13. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012; 2:666–673. [PubMed: 22939981]
14. Ramskold D, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 2012; 30:777–782. [PubMed: 22820318]

15. Tariq MA, Kim HJ, Jejelowo O, Pourmand N. Whole-transcriptome RNAseq analysis from minute amount of total RNA. Nucleic Acids Research. 2011; 39:e120. [PubMed: 21737426]
16. Picelli S, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods. 2013
17. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods. 2013
18. Tang F, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods. 2009; 6:377–382. [PubMed: 19349980]
19. Tang F, et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nature Protocols. 2010; 5:516–535. [PubMed: 20203668]
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. 2010; 6:468–478. [PubMed: 20452321]
21. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009; 4:44–57. [PubMed: 19131956]
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. 2006; 291:L191–L199. [PubMed: 16510470]
23. Sock E, et al. Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. Molecular and cellular biology. 2004; 24:6635–6644. [PubMed: 15254231]
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. 2010; 5:e14085. [PubMed: 21124928]
25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
26. Dalerba P, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol. 2011; 29:1120–1127. [PubMed: 22081019]
27. R core team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. at <http://www.R-project.org/>

Methods References

1. Dalerba P, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol. 2011; 29:1120–1127. [PubMed: 22081019]
2. Babraham Institute, Babraham Bioinformatics. FASTQC. bioinformatics.bbsrc.ac.uk. at <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>
3. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011; 17
4. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011; 27:863–864. [PubMed: 21278185]
5. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]
7. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
8. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
9. Baker SC, et al. The External RNA Controls Consortium: a progress report. Nat Methods. 2005; 2:731–734. [PubMed: 16179916]
10. Jiang L, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011; 21:1543–1551. [PubMed: 21816910]

11. Wu AR, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nature Methods. 2013
12. R core team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. at <http://www.R-project.org/>
13. Zhang H-M, et al. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Research. 2012; 40:D144–D149. [PubMed: 22080564]
14. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res. 1999; 9:1198–1203. [PubMed: 10613842]
15. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009; 37:1–13. [PubMed: 19033363]
16. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009; 4:44–57. [PubMed: 19131956]
17. Greif DM, et al. Radial construction of an arterial wall. Dev. Cell. 2012; 23:482–493. [PubMed: 22975322]

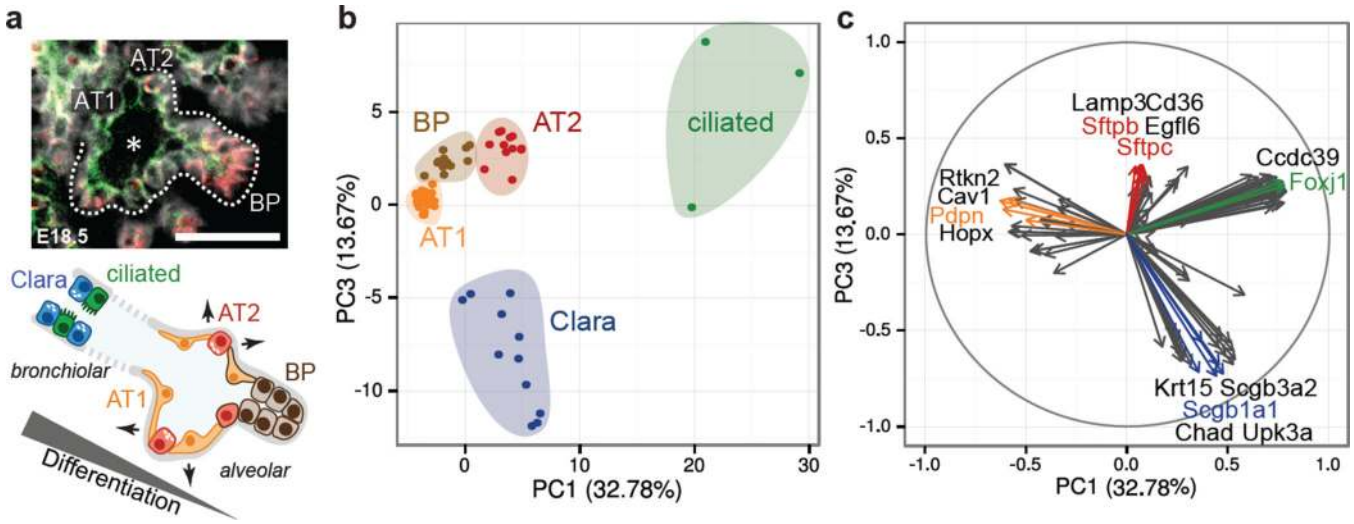


Figure 1. Single cell RNA-seq of 80 embryonic (E18.5) mouse lung epithelial cells enables unbiased identification of alveolar, bronchiolar and progenitor cell populations

(a) Spatially heterogeneous differentiation of distal lung epithelium. Micrograph of a newly forming alveolar sac (asterisk) and schematic below illustrate cell types and 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 (Ecad), pan-epithelial marker). Bipotential progenitor cells (BP) are characterized by co-expression of AT1 and AT2 markers. Schematic: BPs (brown) persist at the tip, nascent AT2 (red) and AT1 (orange) cells are located more proximally. Ciliated (green) and Clara (blue) cells are located in the bronchiolar epithelium (not labeled in micrograph). Scale bar 75 μm.

(b) Principal component analysis (PCA) of 80 single cell transcriptomes (3 biological replicates) at E18.5 distinguishes major bronchiolar and alveolar cell lineages.

(c) Distinct gene groups characterize each cell population based on differential correlation with PC1 and PC3. Arrow tip denotes correlation coefficient of the respective gene with each PC.

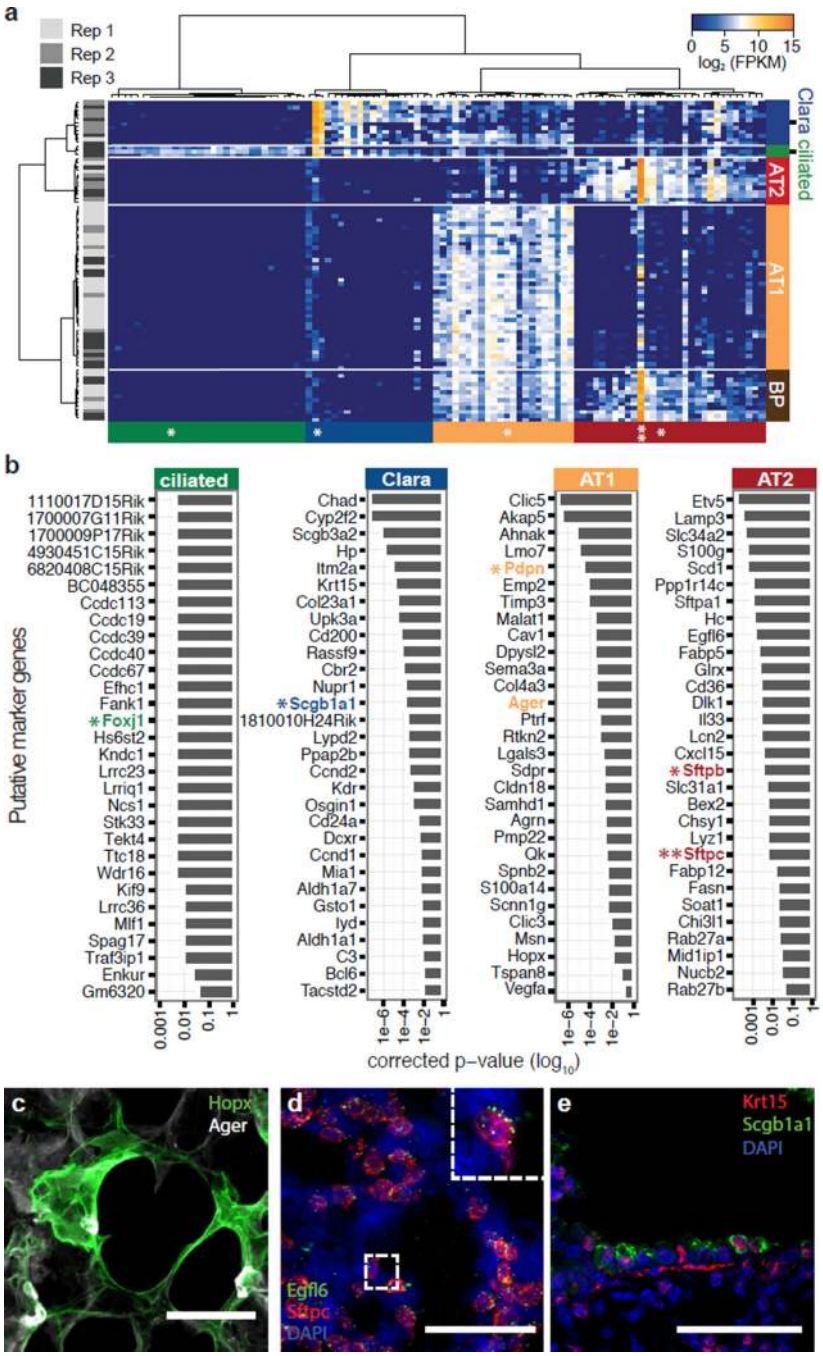


Figure 2. Single cell transcriptome analysis discovers novel markers

(a) Hierarchical clustering of RNA-seq data from 80 single distal lung epithelial cells (E18.5, 3 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*, *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 3). Permutation analysis supports the significance of the presented clustering (p-value = 2.89×10^{-122}, Methods).

(b) Bargraphs 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, bold and colored.

(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 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) mRNA. Inset, close up of boxed region. Blue, DAPI-stained nuclei. Scale bar 50 μm.

(e) Validation of Krt15 expression in Clara cells. Immunofluorescent staining of E18.5 lungs using antibodies against Krt15 (red) and Clara cell marker Scgb1a1 (green). Blue, DAPI-stained nuclei. Scale bar, 50μm.

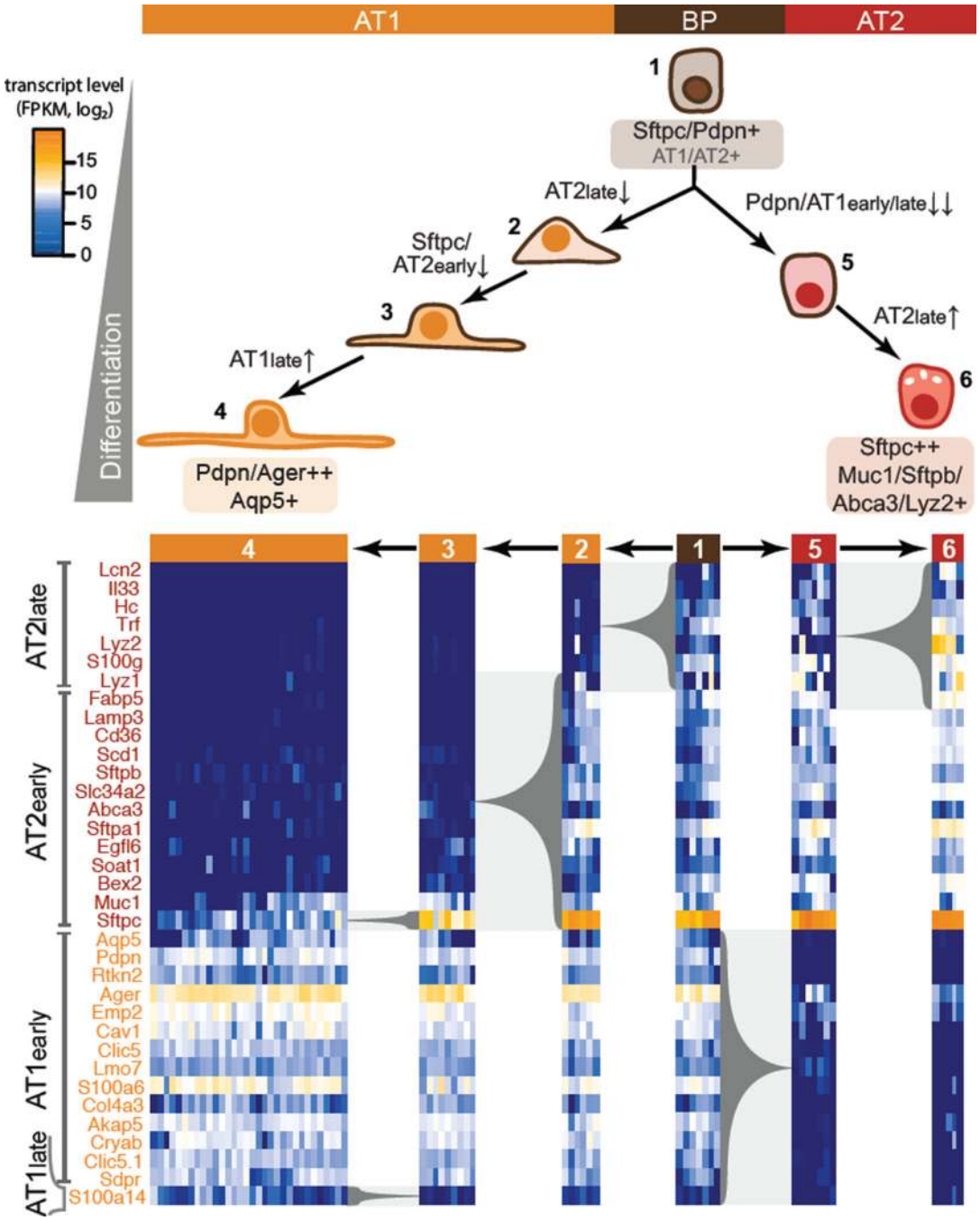


Figure 3. Molecular profiles distinguish developmental intermediates during the differentiation of AT1 and AT2 cells from a common bipotential progenitor

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, based on expression of known and novel marker genes for both alveolar lineages measured by single cell RNA-seq. Genes were grouped into early and late markers of either lineage. Arrows, differentiation pathway; gray braces, change in transcript level of respective genes, tip pointing towards lower expression.

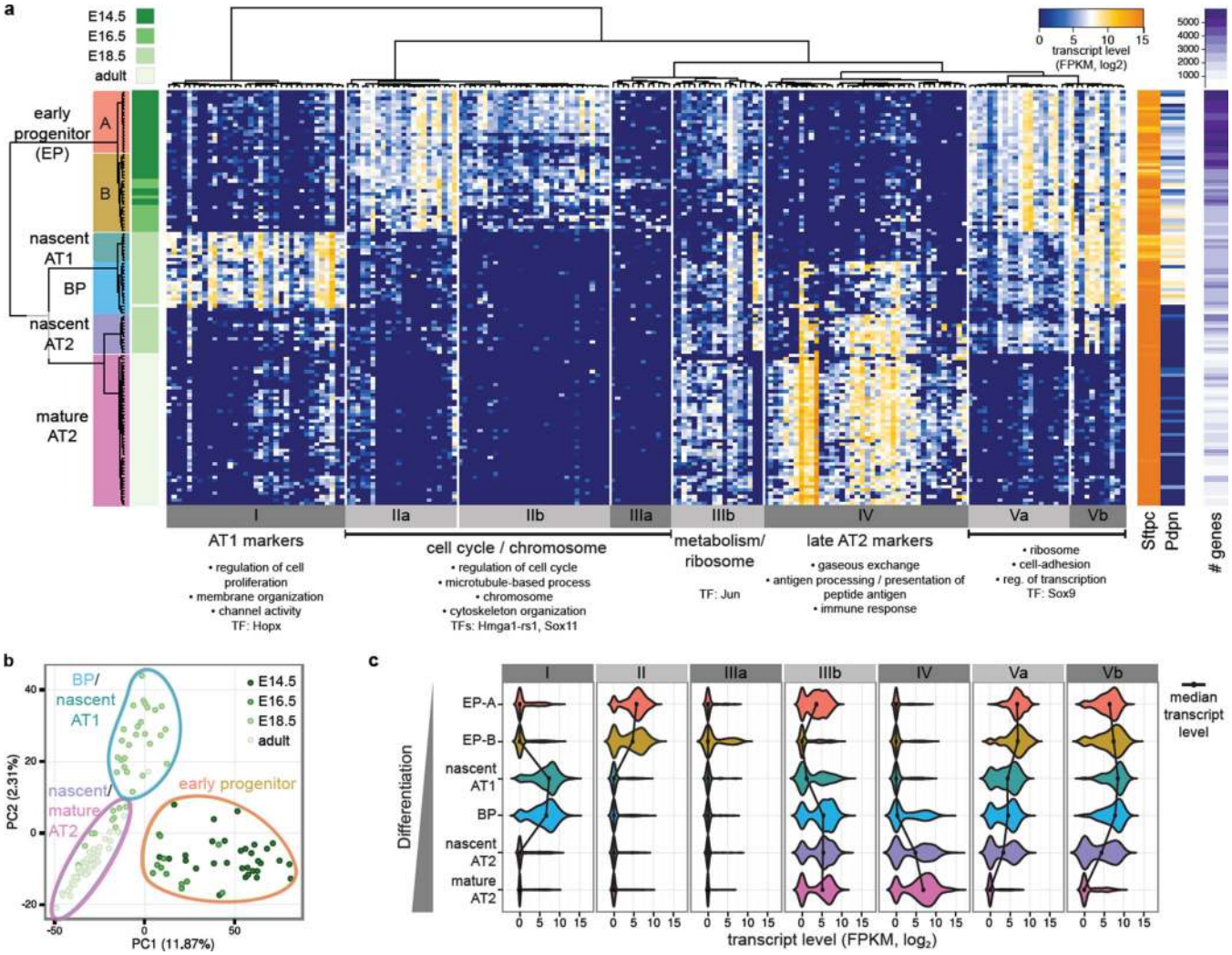


Figure 4. Single cell RNA-seq of *Sftpc*$^+$ cells at E14.5, E16.5, E18.5 and in the adult mouse lung elucidates progressive transcriptional states of the AT2 cell lineage throughout its lifecycle

(a) Hierarchical clustering of 124 *Sftpc*$^+$ cells from distal mouse lung epithelium of embryonic (E14.5, E16.5, E18.5) and adult mice based on genes with highest PC loadings (Supplementary Data 6) in an unbiased PCA analysis (panel b) of all cells and genes. Single cells are shown in rows, genes are shown in columns. Right side-bars show *Sftpc* and *Pdpn* expression, as well as the number of genes expressed by each single cell (see also Extended Data Figure 7). Functional gene ontology enrichments and transcription factors (TFs) specific to each gene group (bottom grey-shaded bars) are shown (Supplementary Data 6). A similar analysis following the Clara cell lineage throughout development is shown in Extended Data Figure 8.

(b) PCA of single cell transcriptomes based on genes detected in more than two cells. Cells cluster into three major populations based on 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 6 cell populations. Each violin plot shows the frequency distribution of the mean transcript level (log_2-transformed FPKM) of all genes per cell.