Supplemental Information

Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity

Simon Joost, Amit Zeisel, Tina Jacob, Xiaoyan Sun, Gioele La Manno, Peter Lönnérberg, Sten Linnarsson, and Maria Kasper
Figure S1
Figure S1. Cell isolation and technical performance. Related to STAR Methods.

(A) Representative pictures of Ki67 immunostainings used to stage the dorsal epidermis of all experimental mice. Mice in mid or late telogen (1st and 2nd panel) were included in the study, while mice in anagen (3rd panel) were excluded.

(B) Number of captured cells from experimental mice in mid (green bar) and late telogen (yellow bar) that were included in the dataset.

(C) Separation of SCA-1+ and SCA-1− cells using microbeads. Left side: illustration of murine telogen epidermis including HF and SG showing the expected expression pattern of SCA-1. Right side: flow cytometry and immunomicroscopy images of epidermal cell suspensions prior to (‘all cells’) and after (‘SCA-1+ fraction’, ‘SCA-1− fraction’) cell separation with SCA-1 microbeads.

(D) Number of SCA-1+ (red bar) and SCA-1− cells (blue bar) included in the dataset.

(E) Number of cells passing the quality control per C1 chip. 34 C1 chips were sequenced for this study, and in total 1,422 single-cell transcriptomes passed the quality control. Red and blue bars signify chips loaded with either SCA-1+ or SCA-1− cells. Black line: mean over all C1 chips.

(F) Size distribution of captured cells included in the dataset (green bars) compared to input cell suspension (red line). Size of captured cells was determined based on the cell area in the microphotographs of cells in the C1 chip. Size distribution of input cell suspensions was measured using a Millipore Scepter Cell Counter and averaged over all experiments. While the single-cell capturing exhibits a minor bias towards larger cells, single cells in the dataset represent the whole size range of the cell input.

(G–H) Total mapped reads (G) and total mapped reads per cell (H) for cells passing quality criteria in each sequenced C1 chip. Red and blue bars signify chips loaded with either SCA-1+ or SCA-1− cells. Black lines denote the mean over all C1 chips; error bars in (H) show the standard deviation between individual cells.

(I) Census of all reads included in the data based on their alignment to the genome (left side) and census of all unique mRNA molecules based on their class (right side).

(J) Number of sequenced reads per molecule (unique molecular identifiers [UMI]). Red line: mean value.

(K) Number of unique mRNA molecules (RNA spike-ins and repeats excluded) sequenced from every cell of the initial dataset. Cells were ordered according to number of unique mRNA molecules from highest (38,000 unique molecules) to lowest and cells with less than 2,000 unique molecules were excluded (leaving 1,422 cells in the final dataset). Inset: histogram of cells according to mRNA yield. Red lines represent the average number of unique mRNA molecules over all cells.

(L) Efficiency of RNA spike-in detection. For each ERCC spike-in species, the number of molecules added to the reaction is plotted against the average number of molecules detected over all 1,422 cells included in the dataset. The diagonal lines demarcate the efficiency boundaries. Inset: histogram aggregating the detection efficiency of each ERCC spike-in species. Red line: median value.

(M) RNA spike-in detection limit. For each ERCC spike-in species, the number of molecules added to the reaction is plotted against the fraction of cells in the dataset in which the molecule was detected at least once. The red line represents a logistic curve fitted to the data. The dotted line marks the inferred minimal number of spike-in molecules necessary for detection in 50% of cases.

(N) Uniformity of single-cell cDNA synthesis reaction environment. All cells included in the dataset (ordered based on 1st level clustering) were correlated according to their ERCC spike-in data.
Figure S2
**Figure S2. First-level clustering of epidermal cells. Related to Figure 1.**

(A) Selection of genes with high variance for unsupervised clustering using a negative binomial noise model. Scatter represents average number of transcripts (log$_2$) versus coefficient of variation (CV; log$_2$) for all genes with average number of transcripts > 0.05 and at least five highly correlated neighbors (7,123 genes). The red line denotes the expected CV as a function of transcript mean according to our fitted noise model log$_2$(CV) = log$_2$(mean$^{alpha}$ + k) with alpha = -0.52 and k = 0.63. The 2,500 genes with the largest difference between expected and observed CV are colored in green and were used for 1st level clustering.

(B–C) Unsupervised clustering of cells (B) and genes (C). Pearson correlation of cells or genes was used as input for clustering with affinity propagation (AP) and Ward’s linkage was subsequently used to define in-cluster order (see STAR Methods). Ordered Pearson correlation matrices of 1,422 cells (B) and 2,500 genes (C) with different color scale cut-offs (right panels) are shown. Group membership for cells in (B) is marked in left- and bottom-panels.

(D) Heatmap showing the expression of 2,500 genes (rows) in 1,422 single cells (columns). Cells and genes are clustered as in (B) and (C), respectively. Group membership of cells is color-coded in the bottom panel and explained in Figure 1C. The four top panels show cell-specific metadata: number of unique mRNA molecules and size of every cell are visualized in the first and second panel while telogen stage and SCA-1 expression are categorized in the third and fourth panel.

(E) Robustness of 1st level clustering was evaluated by resampling (100 iterations) of the dataset and randomly excluding 25% of all cells per iteration. Each subset was reclustered, and the percentage of cells from each cell population that were assigned to the same group was determined (blue dots). The red dots represent the percentage of cells from each group that end up together by pure chance after permutation of cell labels. The black lines show the group means.

(F) Comparison of affinity propagation (AP) clustering and unsupervised clustering with backSPIN. Shown is the relative distribution of cells within each AP cluster over all clusters defined by backSPIN.

(G) Identification of genes that are most highly expressed over Baseline in each population based on negative binomial regression of 1st level clustering data. For each population, the ten genes whose population-specific expression coefficient exceeds the Baseline coefficient with 99.9% posterior probability and who show the largest gap to the Baseline (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline) are reported. The gray and colored violin plots show the posterior probability distribution of the Baseline and population-specific coefficients respectively (scale in molecules).

(H) Identification of genes that are most highly uniquely expressed in each population based on negative binomial regression of 1st level clustering data. For each population, the ten genes whose population-specific expression coefficient exceeds the Baseline and all other populations-specific coefficients with 99.9% posterior probability and who show the largest gap to the second highest coefficient (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the second highest coefficient) are reported. The gray and colored violin plots show the posterior probability distribution of the second highest and population-specific coefficients respectively (scale in molecules).

(I) Barplots showing the absolute expression of selected marker genes in each cell. Cells are ordered into groups according to the clustering in (D). Group membership is color-coded in the upper panel and explained in Figure 1C. Black lines show the average expression over each group.

(J) Remapping of cell populations according to marker gene expression. For every group, either one or two marker genes were selected and antibodies and single molecule FISH probes (gene names in italics) were used to determine their spatial expression pattern in telogen skin. Counterstaining displayed in gray: DAPI (nuclei) or WGA (cell membranes). Scale bars, 10µm.
Population-specific additional expression (colors) compared to Baseline (grey) [molecules]
Population-specific additional expression (colors) compared to second highest population (grey) [molecules]

Select all keratinocyte populations
Select cells with inner bulge signature
Select cells with outer bulge signature
Select cells with upper HF signature
Select cells with IFE basal signature
Select non-infundibular IFE cells

Number of unique molecules
Cell size [µm]

Select cells with upper HF (uHF) signature
Select cells with IFE basal (IFE B) signature
Select cells with inner bulge (IB) signature
Select cells with outer bulge (OB) signature
Select non-infundibular IFE cells

Affinity propagation cluster 1
Affinity propagation cluster 2
Affinity propagation cluster 3
Affinity propagation cluster 4
Affinity propagation cluster 5
Affinity propagation cluster 6
Affinity propagation cluster 7

Figure S3
Figure S3. Second-level clustering of epidermal cells. Related to Figure 2.

(A) Summary of the cell reselection approach for 2nd level clustering. All cells from the dataset, excluding the sebaceous gland and immune cells, were selected and cells with inner bulge, outer bulge, upper HF and IFE basal signatures were chosen in order of primacy. For pseudotemporal ordering, IFE basal cells, excluding infundibular cells, were combined with the remaining IFE cells (intermediate, mature and terminally differentiated).

(B) Heatmaps showing the subclustering of IFE basal, upper HF, outer, and inner bulge cells. Group membership of cells is color-coded in the bottom panel and explained in Figures 2F–2G. The four top panels show cell-specific metadata: number of unique mRNA molecules and size of every cell are visualized in the first and second panel while telogen stage and SCA-1 expression are categorized in the third and fourth panel.

(C–F) Identification of genes that are most highly expressed over Baseline and most highly uniquely expressed in each IFE basal (C), upper HF (D), outer bulge (E) and inner bulge (F) subpopulation based on negative binomial regression of 2nd level clustering data. Left panels: for each subpopulation, the ten genes whose population-specific expression coefficient exceeds the Baseline coefficient with 95% posterior probability and who show the largest gap to the Baseline (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline) are reported. The gray and colored violin plots show the posterior probability distribution of the Baseline and population-specific coefficients respectively (scale in molecules). Right panels: for each subpopulation, the ten genes whose subpopulation-specific expression coefficient exceeds the Baseline and all other subpopulations-specific coefficients (limited to either the subpopulations of the IFE basal, uHF, OB or IB) with 95% posterior probability and who show the largest gap to the second highest coefficient (difference between the 25th percentile of the subpopulation-specific coefficient and the 75th percentile of the second highest coefficient) are reported. The gray and colored violin plots show the posterior probability distribution of the second highest and subpopulation-specific coefficients respectively (scale in molecules).

(G–J) Robustness of IFE basal (G), upper HF (H), outer bulge (I) and inner bulge (J) clustering was evaluated by resampling (100 iterations) of the dataset and randomly excluding 25% of all cells per iteration. Each subset was reclustered, and the percentage of cells from each cell population that were assigned to the same group was determined (blue dots). The red dots represent the percentage of cells from each group that end up together by pure chance after permutation of cell labels. The black lines show the group means.

(K) Transcriptomic similarity of 2nd level subclusters visualized by Ward’s linkage hierarchical clustering of single-cell gene expression data averaged over each group.

(L) Remapping of subpopulations to their spatial location in the epidermis by immunostaining or single molecule FISH (gene symbols in italics). A summary of the populations’ spatial localization can be found in Figure 2G. Arrowheads highlight the positions of the populations. uHF IV (empty arrowhead) / uHF V (filled arrowhead), OB II (arrowhead marks Lgr5(dim) cells in Lgr5-EGFP-Ires-CreERT2 mice using anti-EGFP staining). HS, hair shaft. SG, sebaceous gland. CH, club hair. Scale bars, 10μm.
Figure S4

| Pseudotime | Network | TFs | Mol. signatures |
|------------|---------|-----|-----------------|
| I          |         |     |                 |
| II         |         |     |                 |
| III        |         |     |                 |
| IV         |         |     |                 |
| V          |         |     |                 |
| VI         |         |     |                 |
| VII        |         |     |                 |
| VIII       |         |     |                 |

- Extracellular matrix
- Focal adhesion
- ECM-receptor interaction
- Tissue development
- Cellular component assembly
- Protein transport
- Chloride channel activity
- Lipid metabolic process
- Ribosome
- Cell development
- Apoptosis
- Steroid / cholesterol biosynthesis
- Immune system
- Adherens junction
- Tight junction
- Nectin pathway
- Protein transport
- Secretory pathway
- Wnt signaling
- Biopolymer metabolic process
- Immune system
- Proteolysis
- Endocytosis
- MAPK signaling pathway
- Epidermis development
- Met pathway
- Spingolipid metabolism
- MAPK signaling pathway
- Cytokine / cytokine receptor interaction

Figure S4 (A) Pseudotime Network TFs Mol. signatures

(A) Pseudotime Network TFs Mol. signatures

(B) Pseudotime Network TFs Mol. signatures

(C) Pseudotime Network TFs Mol. signatures

(D) Pseudotime Network TFs Mol. signatures

(E) Pseudotime Network TFs Mol. signatures

(F) Pseudotime Network TFs Mol. signatures

(G) Pseudotime Network TFs Mol. signatures
**Figure S4. Modeling of the epidermal differentiation process. Related to Figure 3.**

**(A)** Robustness of pseudotemporal ordering. Far left panel: comparison of the pseudotemporal ordering selected for Figures 3 and S4 (x-axis) to a pseudotemporal ordering acquired without dimensional reduction through t-SNE (y-axis). Center left panel: comparison of the selected pseudotemporal ordering (x-axis) to one hundred randomly acquired alternative orderings based on different initial t-SNE plots (y-axis). Center right panel: comparison of the selected pseudotemporal ordering (x-axis) to one hundred alternative orderings after randomly removing 25% of cells (y-axis). Far right panel: comparison of the selected pseudotemporal ordering (x-axis) to one hundred alternative orderings after shuffling cell labels (y-axis). The black line in the last three plots shows the median position of each cell over all one hundred iterations, while the blue areas cover the range between the 5th and 95th percentile.

**(B)** Center: average expression of 7,345 genes expressed during IFE differentiation plotted against pseudotime-dependency. Pseudotime-dependency was tested against a pseudotime-independent restricted model using an approximate likelihood ratio test (see STAR Methods) and the p-values are reported. The 1,627 genes (green) with p-values below a Bonferroni-corrected significance threshold of 0.001 (red line) were used for further analysis. Upper/lower panels: examples of low and high expressed genes with or without pseudotime-dependency, respectively.

**(C)** Shared nearest neighbor network of 1,627 pseudotime-dependent genes. Genes are colored according to group membership as established in Figure 3C.

**(D)** Characteristics of different subgroups of differentiation-related genes as defined in Figure 3C. Pseudotime: averaged expression of subgroup-specific genes over pseudotime. The red line shows the median while the gray lines demarcate the 25th and 75th percentile. Network: position of subgroup-specific genes in the shared nearest neighbor network established in (C). TFs: transcription factors included in each subgroup of genes. Mol. signatures: molecular and functional signatures linked to each subgroup of genes.

**(E)** P-values corresponding to the best correlation (highest correlation coefficient) of each cell to the pseudotime model as shown in Figure 3F. Cells are grouped according to (sub) population membership as defined by 2nd level clustering. Black lines denote the median over each group.

**(F)** Robustness of each cell’s correlation to the pseudotime model. To measure robustness of correlation, cells were re-correlated to the pseudotime model for one hundred times after randomly removing 75% of pseudotime-dependent genes. Shown is the average distance between a cell’s pseudotime position in the full model and its position in the re-correlations. It is assumed that a small average distance is indicative of a more robust link to a particular stage in the pseudotime model.

**(G)** Comparison of pseudotemporal ordering of cells in the IFE and the uHF. Left panel: comparison of differentiation-dependency of genes involved in IFE and uHF differentiation. Right panel: comparison of the pseudotime-positions of epidermal cells derived from an IFE- and uHF-based model of differentiation.
### Figure S5

| Pseudospace | Network | TFs | Mol. signatures |
|-------------|---------|-----|-----------------|
| I           |         | Nfkbiz, Klf6 | Ribosome / Translation |
| II          |         | Klf2, Tfp2c, Zlp362, Ets2, N4a, Gata3, Tsc22d1 | Extracellular matrix, Hematopoetic cell lineage, Focal adhesion, Axon guidance, Cell proliferation |
| III         |         | Gata6, Hes1 | Epidermis development, Immune system, Apoptosis, Cell development |
| IV          |         | Hoxb8, Gli1, Gli2, Runx1 | Extracellular matrix, Hedgehog pathway, Axon guidance, (Nervous) System development, Bone remodeling |
| V           |         | Inr5, Lhx2, Tbx1, Sox, Id3, Id6, Setbp1, Smarca2 | Extracellular matrix, Focal adhesion, Ncam signaling, Syndecan 1 pathway, System development |
| VI          |         | Vdr, Nfatc1, Tcf12, Hopx, Mitf, Nfla, Nnat, Sox9, Tfp2b, Nt3c1, Id4, Bnc2, Mitf, Foxp1, Casz1, Lmnap1, Mxi1, Foxc1, Hox13 | Melanogenesis, Wnt pathway, Focal adhesion, Adherens junctions, Transcription |
| VII         |         |              | System development, Fgf signaling, Epidermis development |

### Diagrams

- A: Pseudospace network graphs
- B: Graphs showing expression patterns
- C: 3D visualization of gene expression
- D: Table of TFs and molecular signatures
- E: Heatmap of gene expression
- F: Bar chart showing gene expression changes
- G: Graph showing gene expression changes over time
- H: Scatter plot of gene expression
- I: Bar chart showing gene expression summary
Figure S5. Modeling of spatial gene expression signatures. Related to Figure 4.

(A) Robustness of pseudospatial ordering. Upper left panel: comparison of the pseudospatial ordering selected for Figures 4 and S5 (x-axis) to a pseudospatial ordering acquired without dimensional reduction through t-SNE (y-axis). Upper right panel: comparison of the selected pseudospatial ordering (x-axis) to one hundred randomly acquired alternative orderings based on different initial t-SNE plots (y-axis). Lower left panel: comparison of the selected pseudospatial ordering (x-axis) to one hundred alternative orderings after randomly removing 25% of cells (y-axis). Lower right panel: comparison of the selected pseudospatial ordering (x-axis) to one hundred alternative orderings after shuffling cell labels (y-axis). The black line in the last three plots shows the median position of each cell over all one hundred iterations, while the green areas cover the range between the 5th and 95th percentile.

(B) Center: average expression of 6,788 genes expressed in basal cells plotted against pseudospace-dependency. Pseudospace-dependency was tested against a restricted model using an approximate likelihood ratio test (see STAR Methods) and the p-values are reported. The 547 genes (green) with p-values below a Bonferroni-corrected significance threshold of 0.001 (red line) were used for further analysis. Upper/lower panels: examples of low and high expressed genes with or without pseudospace-dependency, respectively.

(C) Shared nearest neighbor network of 547 pseudospace-dependent genes. Genes are colored according to group membership as established in Figure 4C.

(D) Characteristics of different subgroups of spatial genes as defined in Figure 4C. Pseudospace: averaged expression of subgroup-specific genes over pseudospace. The blue line shows the median while the gray lines demarcate the 25th and 75th percentile. Network: position of subgroup-specific genes in the shared nearest-neighbor network established in (C). TFs: transcription factors included in each subgroup of genes. Mol. signatures: molecular and functional signatures linked to each subgroup of genes.

(E) Overlap of genes expressed over Baseline in basal-cell populations (and IB I). Genes were called from the negative binomial regression model of 2nd level clustering if the population-specific regression coefficient exceeded Baseline with 95% posterior probability.

(F) Expression of spatial genes in all epidermal (sub) populations defined by either 1st or 2nd level clustering. A gene was considered expressed in a population if its population-specific coefficient in the negative binomial regression model exceeded Baseline with 95% posterior probability. Genes are ordered according to group membership introduced in Figure 4C. The shaded populations exhibit gene expression inconsistent with any distinct spatial signature.

(G) Position of epidermal cells from each population on the spatial axis as determined by highest Pearson correlation. The shaded cells belong to populations that show gene expression inconsistent with any spatial signature.

(H) P-values corresponding to the best correlation of each cell to the pseudospace model as shown in (G). Black lines denote the median over each group.

(I) Robustness of each cell’s correlation to the pseudospace model. To measure robustness of correlation, cells were re-correlated to the spatial axis for one hundred times after randomly removing 75% of pseudospace-dependent genes. Shown is the average distance between a cell’s pseudospace position in the full model and its position in the re-correlations. It is assumed that a small average distance is indicative of a more robust link to a particular stage in the pseudospace model.
Figure S6
Figure S6. Explaining cellular heterogeneity using differentiation and spatial signatures. Related to Figure 5.

(A) Pseudotime- and pseudospace-dependency of genes. Black lines mark the Bonferroni-corrected significance threshold of 0.001. Of 7,893 genes, 1,409 were uniquely pseudotime-, 329 uniquely pseudospace- and 218 both pseudotime- and pseudospace-dependent. Only the uniquely pseudotime- or pseudospace-dependent genes were considered in the pseudospacetime model.
(B) Percentage of molecules explained (green), underexplained (red) or overexplained (blue) by the pseudospacetime model. Molecules were pooled across the cells per population.
(C) Percentage of molecules explained (green), underexplained (red) or overexplained (blue) by the pseudospacetime model in each single cell per population.
(D) Accuracy of pseudospacetime, 1st level clustering, 2nd level clustering and shuffled pseudospacetime model stratified according to cell populations defined in either 1st or 2nd level clustering.
(E) Fraction of explained molecules contributed by Baseline, differentiation axis, spatial axis and other (sebaceous gland, immune) signatures for each cell population.
Figure S7
Figure S7. Cellular heterogeneity of stem cell populations. Related to Figure 6.

(A) Immunostaining and single molecule FISH (gene symbols in italics) of SCMs in epidermis. Note that most markers are expressed in several epidermal compartments. Scale bars, 20µm.

(B) Cells projected onto the t-SNE map of basal cells (see Figure 6B), colored according to 1st level, 2nd level and selective clustering of basal cells.

(C) Matrix showing the overlap in SCM expression. Percentage of cells expressing each SCM Lgr5, Cd34, Gli1, Lgr6, Lrig1, or Krt14 (rows) co-expressing additional SCM genes (columns).

(D–F) Analyses of gene expression signatures for SCM-expressing populations. (D) Identification of the top ten genes that were most highly expressed over Baseline in each stem cell population based on negative binomial regression. For each population, the ten genes whose population-specific expression coefficient exceeds the Baseline coefficient with 95% posterior probability and which show the largest gap to the Baseline (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline) are reported. The gray and colored violin plots show the posterior probability distribution of the Baseline and population-specific coefficients, respectively (scale in molecules). (E) Identification of genes that are most highly and uniquely expressed in each stem cell population based on negative binomial regression. For each population, the ten genes whose population-specific expression coefficient exceeds the Baseline and all other populations-specific coefficients with 95% posterior probability and that show the largest gap to the second highest coefficient (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the second highest coefficient) are reported. The gray and colored violin plots show the posterior probability distribution of the second highest and population-specific coefficients, respectively (scale in molecules). (F) Characteristics of different stem cell populations. Genes: number of genes expressed over Baseline with 95% posterior probability. TFs: transcription factors included in each set of genes. Mol. signatures: molecular and functional signatures linked to each subgroup of genes.

(G–I) Analyses of shared gene expression signatures in SCM+ cells. (G) Identification of the top ten genes that were most highly expressed over Baseline among all SCM+ cells. In contrast to (D), a 90% posterior probability cut-off was chosen. (H) Identification of the top ten genes expressed in all SCM+ cells that were most highly expressed compared to SCM− cells. A 90% posterior probability cut-off was chosen and only genes whose SCM+ cell-specific expression exceeded 0.25 molecules were chosen. The black and red violin plots show the posterior probability distribution of the SCM+ and SCM− specific coefficients respectively. (I) Characteristics of the SCM+ population. Genes: number of genes with higher expression than in the SCM− population based on a 90% posterior probability cut-off. TFs: transcription factors. Mol. signatures: molecular and functional signatures.

(J) Percentage of shared genes between the specific signatures of Lgr5+, Cd34+, Gli1+, Lgr6+, Lrig1+, Krt14+, and SCM+ cells. The specific signatures were defined as specified in (D), (F), (G) and (I).

(K) Pseudotime vs. pseudospace-dependency of stem cell-specific genes. Plotted is the difference between the −log10 transformed p-value of pseudotime- and pseudospace-dependency. “Stem cell population associated genes” include all genes, which are expressed over Baseline in at least one stem/progenitor population. “Genes linked to SCM+ or SCM− cells” are the 44 genes that are expressed differently between SCM+ and SCM− cells as specified in Figures 6G and S7G–S7I. The black lines denote the median.
A. Description of cell populations from 1st defined in this study.

Identification of genes is based on negative binomial regression of 1st level clustering data. For each population, genes whose population-specific expression coefficient exceeded the Baseline coefficient (left column) or all other populations-specific coefficients (right column) with 99.9% posterior probability and which show the largest gap to the Baseline / the second highest population (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline / the second highest population) are listed.

(Supplied as Excel file: Joost Table S1.xls)

Table S2. Marker genes – 2nd level clustering. Related to Figure 2.

Lists of genes that are most highly expressed over Baseline (vs. Baseline, left column) or the second highest population (vs. other groups, right column). Identification of genes is based on negative binomial regression of 2nd level clustering data. For each population, genes whose population-specific expression coefficient exceeded the Baseline coefficient (left column) or all other populations-specific coefficients (right column) with 95% posterior probability and which show the largest gap to the Baseline / the second highest population (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline / the second highest population) are listed.

(Supplied as Excel file: Joost Table S2.xls)

Table S3. Description of cell populations and comparison to literature. Related to Figures 1, 2 and 6.

(A) Description of cell populations defined during 1st and 2nd level clustering. Described are the molecular and spatial characteristics of each population defined in this study and their relation to previous work.

(B) Previously defined murine epidermal (stem) cell populations and their relation to populations defined in this study.

### SUPPLEMENTAL TABLES

**Table S1. Marker genes – 1st level clustering. Related to Figure 1.**

| Populations | Molecular and spatial description | Previous descriptions |
|-------------|-----------------------------------|-----------------------|
| Interfollicular basal I (IFE B I) | IFE basal cell population marked by TH17/ Foxp3 expression and higher than average expression of IFE basal genes such as Krt14, 6RT and 79. IFE B I cells are interspersed with IFE B II cells in the IFE basal layer. | A TH17 (+) subpopulation of the IFE has not been previously described. This population is not congruent with the IVE population or Lgr5+ population described by Mascré et al., 2012, and Füllgrabe et al., 2015. |
| Interfollicular basal II (IFE B II) | IFE basal cell population marked by absence of 6RT and shows that average expression of IFE basal genes such as Krt14, 6RT, 79. IFE B II cells are interspersed with IFE B I cells in the IFE basal layer. | A HIF (basal) population of the IFE has not been previously described. This population has not been previously described. This population is not congruent with the IVE population or Lgr5+ population described by Mascré et al., 2012, and Füllgrabe et al., 2015. |
| Infundibular basal (INFU B) | A population of cells dominated by typical IFE basal markers such as Krt14, 6RT and 79, which additionally expresses pan and upper HF markers such as Sostdc1, Cyst6, Fts and Krt15. Located in the infundibular region close to the HF opening. The specific spatial position of this population is clearly identifiable by Posh expression. While it has been shown that pan HF markers such as Sostdc1/reach to the infundibular region (Collette et al., 2013), the particularly IFE basal character of this population in combination with the displayed gradual expression of HF markers has not been clarified before. |
| IFE differentiated cells I (IFE D I) | Transient population of cells marked by low level expression of both (IFE basal (Krt14, 6RT) and (IFE) suprabasal (Krt10, 79) markers. Expresses aHIF as one of a few specific markers. Available expression of HF marker levels clustering data. For each population, genes whose population | The existence of the IVE basal-to-suprabasal transient population is expected according to the accepted model of epidermal differentiation, it has never been resolved at a transcriptional level. |
| IFE differentiated cells II (IFE D II) | Mature, suprabasal population expressing high levels of well-established spinous layer markers such as Krt10 and Sbsn. The cells of the spinous layer are well described and characterized. See for instance Fuchs, 1990. |
| IFE keratinized layer I (IFE K I) | A transient population of cells which is marked by decreasing levels of spinous layer markers such as Krt10 and Sbsn (when compared to IFE D II) and increasing levels of granular layer markers including Lgr6 and Flg2. While it has been shown that pan HF markers such as Sostdc1/reach to the infundibular region (Collette et al., 2013), the particularly IFE basal character of this population in combination with the displayed gradual expression of HF markers has not been clarified before. |
| IFE keratinized layer II (IFE K II) | A population of flat, keratinized cells expressing high levels of well-established granular layer markers such as Lgr6 and Flg2. The cells of the granular layer are well described and characterized. See for instance Fuchs, 1990. |
| Upper HIF (UHF I) | A population of cells marked by a typical uHF signature (Krt14, 1977, Gm64) in combination with expression of a gene module distinguished by high K5/ K14, Deft8 and Cnfl expression. Additionally, epidermis- and compartment-unique markers such as Krt10 and Cry1b are expressed. Could be mapped to two rings of suprabasal cells above and below the SG opening based on highest Cst6 expression. Accordingly, KLK10 is mostly found secreted into the hair canal at the corresponding positions. | Has not been previously described in molecular detail. Although Zeewen et al., 2002 and Veniaminova et al., 2013 described Cst6 expression in the upper HF, they were unable to differentiate a set of uHF populations with strong Cst6 expression and a set of uHF populations with weak or absent Cst6 expression or subdivide those populations further. It is not clear whether this population contains cells of BLIMP1 / Prdm1 population described by Horney et al., 2006. |
| Upper HIF (UHF II) | A population of cells marked by a typical uHF signature (Krt14, 1977, Gm64) in combination with high expression of a gene module distinguished by high K5/ K14, Deft8 and Cnfl expression. This population is additionally distinguished by a K5/ K14 / Log2 signature while the K5/ K14 module is absent. Can be mapped to the SG opening based on the location of highest Deft8 expression in the epidermis. Additionally, the absence of Krt14 / Prdm1 expression in the population does not allow its location in the basal layer of the upper HF / junctional zone while the presence of Krt14 argues against suprabasal localization (e.g. adjacent to uHF I). | Has not been previously described. Although it has been shown previously that Log2+ cells are located in different compartments of the skin including the uHF and the SG (Paget et al., 2013), and Krt79 and Cst6 are expressed in the uHF (Veniaminova et al., 2013; Zeewen et al., 2002), these populations were never dissected on a molecular / transcriptional level. |
B. Previously defined murine epidermal (stem) cell populations

References

Krt14+/iv- and Krt14+/iv+ IFE basal cells (Mascre et al., 2012)

While the infundibulum (INFU) and upper HF (uHF IV-V) basal cells shared a higher uHF-like baseline expression, krt14+ was rarely found in the IFE basal populations (ICE VIII-IX). Instead, krt14+ expression was predominantly detected in the differentiating and terminally differentiated basal populations. It is possible that Mcaseo et al., 2012 targeted cells, which are in the process of transition from basal to suprabasal and show both basal and suprabasal characteristics.

Spinoius layer cells (Miscellaneous)

Cells that show a spinous layer signature could be found in the IFE (ICE I), the upper HF (uHF VI-VII) and the outer bulge (OB V) and the inner bulge compartment (IB III).

Granular layer cells (Miscellaneous)

Cells distinguished by a granular layer signature are present in the IFE (ICE I and II) and the upper HF compartment (uHF VI). This population is most likely to be located in the outer bulge region.

Lgr4+ cells located in the isthmus (Nijhoff, 2008)

Expression of fter4 (H20000280RA) could be detected in all populations in the upper HF and in terminally differentiated cells of the IFE. In the upper HF, fter4 expression peaked in the terminally differentiated populations (uHF VI and uHF VII) and in the basal compartment of uHF I-VII. A distinct Lgr4+ isthmus (adult) population could not be observed. It was neither possible to clearly demarcate Lgr4+ and Lgr4- populations in the IFE basal layer.

GI1+ cells in the upper bulge (Browne et al., 2011)

Could be identified (OB IV) as cells with an outer bulge signature and a unique set of co-expressed genes (Dnmt3, Apln, Krt5).

Krt15+ / CD34+ mid bulge cells (Cosentino et al., 1990, Morris et al., 2004)

CD34 is the most prominent marker of the bulge and in our dataset we confirmed CD34 as a pan outer bulge and in all outer bulge populations. Cells of the mid bulge (CD34+/Lgr5-) were most likely included in OB I and OB II. However, it was not possible to resolve those cells as a distinct population.

Krt15+ / CD34+ / Lgr5+ lower bulge cells (Jaks et al., 2008)

Although some p-cadherin (Cdhr5) expressing cells could be found in OB I, OB II, and OB III, it was not possible to resolve those as a distinct population.

Suprabasal CD34- + Igfr2 cells (Blanpain et al., 2004)

Most likely represented by the population of cells with both an outer bulge and a spinous layer signature (OB V).

Eg86+ bulge population which provides attachment to the arrector pili muscle (Fujisawa et al., 2011)

Egfl6+ expressed consistently over all outer bulge populations and sporadically in the IFE. No distinct Egfl6+ population could be resolved.

Inner bulge cells formed a highly distinct 1st level cluster (Krt14+) which could be further divided into three subpopulations (IB I–III).
## Table S4. Differentiation-related genes. Related to Figure 3.

List of significant pseudotime-dependent genes. Genes are grouped according to clustering shown in Figure 3C. Within each cluster, genes are ordered from lowest to highest p-value.
Gene expression coefficients for each population were used to identify marker genes. In the case of Lgr5, Cd34, Gli1, Lgr6, and Krt14 expression, each population was compared to either a shared Baseline or all other populations. For each population, genes whose population-specific expression coefficient exceeded the Baseline coefficient (left column) or all other populations-specific coefficients (right column) with 95% posterior probability and which show the largest gap to the Baseline / the second highest population (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline / the second highest population) are listed. The SCM + basal cells were compared to SCM− basal cells or a Baseline shared by both populations. Genes which exceed the Baseline coefficient (left column) or the SCM− basal cell coefficient (right column) with 95% posterior probability and which show the largest gap to the Baseline / the SCM− basal cells (difference between the 25th percentile of the population-specific coefficient and the 75th percentile of the Baseline / the SCM− basal cells) are listed.
Table S7. Immunohistochemistry and single molecule FISH stainings. Related to STAR Methods.

Listed are the markers used for the validation and localization of the defined cell populations, the respective number of mice and analyzed images, and the corresponding figures in the manuscript.

| Population | Clustering level | Staining to identify populations based on sequencing data | Number of mice | Number of images taken (HF / HF+IFE / IFE) | Number of images showing the respective population | Corresponding figure in manuscript |
|------------|------------------|--------------------------------------------------------|---------------|------------------------------------------|----------------------------------------------------|----------------------------------|
| IFE Bi     | 2                | KRT14(hi)/Thbs1(hi)                                    | 3             | 1 / 15 / 11                             | 25 out of 27                                       | 2E                                |
| IFE BII    | 2                | KRT14(dim)/Thbs1(lo)                                   | 3             | 1 / 15 / 11                             | 25 out of 27                                       | 2E                                |
| INFU B     | 2                | Postn(dim)                                             | 4             | 25 / 78 / 0                            | 61 out of 103                                      | 2E                                |
| IFE DI     | 1                | KRT10(dim)/PTGS1(dim)                                  | 2             | 0 / 7 / 0                              | 7 out of 7                                         | S2J                               |
| IFE DII    | 1                | KRT10(hi)/PTGS1(hi)                                    | 2             | 0 / 7 / 0                              | 7 out of 7                                         | S2J                               |
| IFE KI     | 1                | LOR(dim)/Flg2(dim)                                     | 3             | 0 / 11 / 5                             | 16 out of 16                                       | S2J                               |
| IFE KII    | 1                | LOR(hi)/Flg2(hi)                                       | 3             | 0 / 11 / 5                             | 16 out of 16                                       | S2J                               |
| uHF I      | 1                | KRT17(lo)/KRT79(lo)                                    | 3             | 0 / 17 / 0                             | 16 out of 17                                       | S2J                               |
| uHF II     | 1                | KRT17(dim)/KRT79(dim)                                  | 3             | 0 / 17 / 0                             | 16 out of 17                                       | S2J                               |
| uHF III    | 1                | KRT17(hi)/KRT79(hi)                                    | 3             | 0 / 17 / 0                             | 16 out of 17                                       | S2J                               |
| uHF I      | 2                | KRT14(lo)/Cst6(hi) and KLK10 to localize               | 3             | 0 / 16 / 0 for Krt14/Cst6; 3 for Klk10 | 16 out of 16 for Krt14/Cst6; 14 out of 15 for Klk10| 2E, S3L                           |
| uHF II     | 2                | KRT14(hi)/Cst6(hi/dim)                                 | 3             | 0 / 16 / 0 for Krt14/Cst6              | 16 out of 16 for Krt14/Cst6                         | 2E                                |
| uHF III    | 2                | can't be stained*                                      |               |                                        |                                                   | not applicable                    |
| uHF IV     | 2                | KRT14(dim)/Krt79(lo)                                   | 3             | 3 / 12 / 0                             | 15 out of 15                                       | S3L                               |
| uHF V      | 2                | KRT14(dim)/Krt79(hi)                                   | 3             | 3 / 12 / 0                             | 15 out of 15                                       | S3L                               |
| uHF VI     | 2                | KRT10(hi)/KRT79(hi)                                    | 3             | 0 / 14 / 0                             | 11 out of 14                                       | S3L                               |
| uHF VII    | 2                | LOR(hi)/KRT79(hi)                                      | 3             | 0 / 20 / 0                             | 17 out of 20                                       | S3L                               |
| SG         | 1                | MGST1(pos)                                             | 3             | 0 / 18 / 0                             | 18 out of 18                                       | 1F, S2J                           |
| OB         | 1                | CD34(hi)/Postn(hi)                                     | 2             | 7 / 7 / 0                              | 14 out of 14                                       | S2J                               |
| OB I       | 2                | Lgr5-EGFP(hi)/Postn(hi)                                | 4             | 25 / 78 / 0 for Postn; 2 for Lgr5**    | 95 out of 103 for Postn; 39 out of 43 for Lgr5     | S3L                               |
| OB II      | 2                | Lgr5-EGFP(dim)/Postn(hi)                               | 4             | 25 / 78 / 0 for Postn; 43 / 0 / 0 for Lgr5** | 95 out of 103 for Postn; 39 out of 43 for Lgr5     | S3L                               |
| OB III     | 2                | KRT15(lo)/Postn(hi)                                    | 3             | 0 / 16 / 0                             | 9 out of 16                                        | 2E                                |
| OB IV      | 2                | Postn(hi)/Krt79(dim)                                   | 4             | 5 / 20 / 0                             | 15 out of 25                                       | 2E                                |
| OB V       | 2                | KRT10(hi)/Postn(hi)                                    | 3             | 1 / 21 / 0                             | 18 out of 22                                       | 2E                                |
| IB I       | 2                | KRT6(hi)                                               | 4             | 6 / 24 / 0                             | 30 out of 30                                       | 1F, S2J                           |
| IB II      | 2                | KRT6(hi)/Postn(hi)                                     | 3             | 5 / 20 / 0                             | 19 out of 25                                       | 2E                                |
| IB III     | 2                | KRT6(hi)/Krt79(hi)                                     | 2             | 3 / 17 / 1                             | 14 out of 20                                       | 2E                                |
| TC         | 1                | CD3(pos)                                               | 3             | 0 / 25 / 0                             | 24 out of 25                                       | 1F, S2J                           |
| LH         | 1                | CD207(pos)                                             | 2             | 0 / 10 / 0                             | 10 out of 10                                       | 1F, S2J                           |

**Legend**

- hi = high; lo = low; pos = positive
- IHC / RNAscope
- Populations stained together

* was located via exclusion of positive stainings

** costaining was technically not possible