nature research

Corresponding author(s): Cédric BLANPAIN

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The confocal imaging data-sets generated were merged and displayed with the ZEN 2 software. Flow cytometry data analysis and cell sorting was performed on a FACSAria sorter using the FACS DiVa software (BD Biosciences).
Data analysis	The confocal imaging data-sets generated were analysed with the ZEN 2 software. Flow cytometry data analysis was performed on a FACSAria sorter using the FACS DiVa software (BD Biosciences). All microarray results were normalized using the RMA normalization algorithm using R-bioconductor affy package with standard parameters. Cross experiment normalization was further performed to eliminate the batch effect using non-parametric empirical Bayes frameworks for adjusting data implemented in ComBat function of the Surrogate Variable Analysis package (SVA) in R-bioconductor. Venn diagrams were generated using Venny 2.0. GSEA analysis was performed using ranked fold change of probe expression values between Lgr5 and BCs or LCs and genes upregulated in LCs or BCs for the displayed dataset. Genes up-regulated in each subset of the Venn Diagrams were tested for enrichment in each Gene Ontology class using the DAVID web server. Sequencing reads were trimmed for adapter sequences using cutadapt (version 1.13) and reads were aligned to the GRCm38 reference genome including ERCC sequences using STAR with default parameters (version 2.5.2b)79. The expression count matrix was generated using HTSeq (version 0.6.0)80 on GENCODE M12 transcript annotations and counts for each protein coding gene were collapsed. Quality control was performed using the scater R package (version 1.2.0)81. Cells that complied with one of the following conditions were excluded: had fewer than 105 counts, showed expression of fewer than 2000 unique genes, had more than 20% counts belonging to

ERCC sequences, had more than 8% counts belonging to mitochondrial sequences. BCs and EMPs that showed no expression of neither K5 nor K14, and LCs that did not express K8 were further excluded. Out of the 377 samples which passed sequencing, 281 passed quality control. Genes for which less than 20 counts were observed across the complete dataset were excluded from further analysis. Read counts were normalized using scran with default parameters (version 1.2.2) 82. Clustering using the SC3 R package (version 1.3.18)49 and PCA was performed using the prcomp function in R, plots were generated using the ggplot2 R package (version 2.2.1). We chose k=4 (all cells), k=2 (for EMP-only and LC vs BC clustering) for SC3 as this best represented the heterogeneity in our dataset and recapitulated the studied cell lineages. For cluster marker gene discovery we set the thresholds to all genes with an AUC higher than 0.8 and p-value lower than 0.01. Heatmaps were generated using a modified version of the gplots R package (3.0.1). Cell-cycle phase was automatically assigned using the scran package.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Microarray, RNAseq and single cell RNA sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE109711. Previously published microarray data that were re-analysed here are available under accession code GSE69290.
Source data for Figure 1d, 1l, 1q, 6d and 7c have been provided as Supplementary Table 1.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. All experiments were repeated at least three times with similar results, except for microarray and RNAseq for which experiments were repeated twice.
Data exclusions	For the single cell RNA sequencing, after initial quality control 24 samples were discarded with cDNA yields of less than 21ng. These samples were replaced with 24 cells from another 384 well plate from the same cell sort by plate reformatting with an acoustic dispenser (LabCyte Echo 525). Cells coming from a row F of the 384-well plate which showed systematic mixing of LC and BC markers were excluded from further analysis due to a likely pipetting error. Out of the 377 samples which passed sequencing, 221 passed quality control. Genes for which less than 20 counts were observed across the complete dataset were excluded from further analysis.
Replication	All attempts at replication were successfull and are shown, n is described in legends.
Randomization	The experiments were not randomized. Transgenic mice of mixed background were used and compared to control mice of mixed background.
Blinding	The investigators were not blinded to allocation during experiments and outcome assessment. Blinding was not possible as the same investigator processed the animals and analysed the data.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☐	☒ Research animals
☒	☐ Human research participants

Antibodies

Antibodies used

The following primary antibodies were used: anti-GFP (chicken, 1:1000, ab13970, Abcam), anti-K8 (rat, 1:1000, Troma-I, Developmental Studies Hybridoma Bank, University of Iowa), anti-K14 (rabbit or chicken, 1:1000, Thermo), anti-K5 (rabbit, 1:1000, PRB-160P-0100, Covance), anti-CD49f-PE (rat, 1:100, clone GoH3, eBiosciences), anti-p63 (rabbit, 1:500, clone EPR5701, Abcam), anti-SMA-Cy3 (mouse, 1:500, clone 1A4, Sigma), anti-smMHC (rabbit, 1:100, BT562, Biomedical Technologies), anti-Sox9 (rabbit, 1:5000, AB5535, Millipore), anti-FoxA1 (rabbit, 1:100, clone EPR10881, Abcam), anti-ER (rabbit, 1:300, sc542, Santa Cruz). The following secondary antibodies, diluted 1:400, were used: anti-rabbit (A21206), anti-rat (A21208), anti-chicken (A11039) conjugated to AlexaFluor488 (Molecular Probes), anti-rabbit (711-295-152), anti-rat (712-295-153), anti-chicken (703-295-155) Rhodamine Red-X or anti-rabbit (711-605-152), anti-rat (712-605-153), anti-chicken (703-605-155) Cy5 (JacksonImmunoResearch).

Validation

Antibodies specificity was controlled by using sections of tissues known to express high levels of the antigen, as well as sections of tissues known to lack expression of the antigen.
Antibodies used in this study were described previously in Van Keymeulen et al, Nature 2011, Van Keymeulen et al, Nature 2015, Wuidart et al, Genes and dev 2016, Van Keymeulen et al, Cell Reports 2017.

Research animals

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Animals/animal-derived materials

Lgr5-EGFP-IRES-CreER15 and Rosa-tdTomato55 mice were obtained from the Jackson Laboratory. Rosa-Confetti13 mice were provided by H. Clevers; K14rtTA transgenic mice56 were provided by Elaine Fuchs; TetO-Cre mice57 were provided by Andreas Nagy; Rosa26-ΔNp63-IRES-GFP mice51 were provided by Wim Declercq. The generation of K5CreER and of K8rtTA were previously described8, 10. Mice colonies were maintained in a certified animal facility in accordance with European guidelines. The experiments were approved by the local ethical committee (CEBEA) under protocols #477 and #527.The study is compliant with all relevant ethical regulations regarding animal research.
All experimental mice used in this study were females of mixed genetic backgrounds. Mice were analysed at embryonic stages E14, E15 and E17, during postnatal development at P1, P5, P21, P60, 7w and in adult mice (over 8w), as indicated in figure legends.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ Magnetic resonance imaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Mammary cell preparation – E14 Lgr5-GFP embryos were collected and the whole skin containing the mammary buds was dissected. Female tissues were placed in HBSS + 300U/ml collagenase (Sigma) + 300µg/ml hyaluronidase (Sigma) and digested for 1h30 at 37°C under shaking. EDTA at a final concentration of 5mM was added for 3min to the resultant organoid suspension, followed by a wash in 10% FBS/PBS and labelling in 2% FBS/PBS before filtration through a 40µm mesh. Adult mammary glands

were dissected and the lymph nodes removed. Tissues were briefly washed in HBSS, and chopped in 1mm3 pieces. Chopped tissues were placed in HBSS + 300 U/ml collagenase (Sigma) + 300µg/ml hyaluronidase (Sigma) and digested 2h at 37°C under agitation. Physical dissociation using a P1000 pipette was done every 15mins throughout the enzymatic digestion time. EDTA at a final concentration of 5mM was added for 5 min to the resultant organoid suspension, followed by 0,25% Trypsin-EGTA for 1 min before filtration through a 70-µm mesh, 2 successive washes in 2% FBS/PBS and labelling.

Cell labelling, flow cytometry and sorting – Samples were incubated in 250µl of 2% FBS/PBS with fluorochrome-conjugated primary antibodies for 30min, with shaking every 10min. Primary antibodies were washed with 2% FBS/PBS, and cells were resuspended in 2.5mg/ml DAPI (Invitrogen) before analysis. The following primary antibodies were used: APC-conjugated anti-CD45 (1:100, clone 30-F11, eBiosciences), APC-conjugated anti-CD31 (1:100, clone 390, eBiosciences), APC-conjugated anti-CD140a (1:100, clone APA5, eBiosciences) and PE-conjugated anti-CD49f (1:200, clone GoH3, eBiosciences) for embryos; PECy7-conjugated anti-CD24 (1:100, clone M1/69, BD Biosciences), APC-conjugated anti-CD29 (1:100, clone eBioHMb1-1, eBiosciences), PE-conjugated anti-CD45 (1:100, clone 30-F11, eBiosciences), PE-conjugated anti-CD31 (1:100, clone MEC 13.33, BD Biosciences), PE-conjugated anti-CD140a (1:100, clone APA5, eBiosciences) for adult MGs. Data analysis and cell sorting was performed on a FACSAria sorter using the FACS DiVa software (BD Biosciences). Dead cells were excluded with DAPI; CD45+, CD31+ and CD140a+ cells were excluded (Lin+) before analysis of the GFP+ cells.

Instrument

FACSAria 3 sorter (BD Biosciences)

Software

FACS DiVa software (BD Biosciences)

Cell population abundance

To control cell sorting quality, 5000 cells of the different populations of interest were sorted in PBS and their FACS profile was analyzed using exactly the same gating strategy (post-sort control). Minimum 90% of the sorted cells were located within the sorting gate. Cells usually showed a reduced fluorescent labelling, due to bleaching following the first sort, explaining why a fraction of the cells are not located within the sorting gate.

Gating strategy

Supplementary Figure 1. Gating strategy for flow cytometry analysis. A-G, Unicellular suspension of skin and mammary bud cells from Lgr5-IRES-GFP E14 embryos stained for Lin (CD31, CD45, CD140a) in APC and CD49f in PE were gated as shown in A to eliminate debris, doublets were discarded with gate shown in B followed by gate showed in C, the living cells were gated by DAPI dye exclusion as shown in D, the non-epithelial Lin positive cells were discarded in E. The CD49f Hi cells were gated as shown in F and the GFP+ cells were gated as shown in G. H-O, Unicellular suspension of mammary cells from adult K8rtTA/TetOCre/DNp63-IRES-GFP, induced at P30 and analysed at P45, stained for Lin (CD31, CD45, CD140a) in PE, CD24 in PECy7 and CD29 in APC, were gated as shown in H to eliminate debris, doublets were discarded with gates shown in I followed by gate shown in J, the living cells were gated by DAPI dye exclusion as shown in K, the non-epithelial Lin positive cells were discarded in L and the GFP+ cells were gated as shown in M. CD24 and CD29 expression was studied in Lin- cells (N) or in YFP+ cells (O). The CD24+CD29Lo gate corresponds to luminal cells (LC), while CD24+CD29Hi gate corresponds to basal cells (BC). The stromal population corresponds to the cells labelled due to the leakiness of the Tet-O-Cre, as described previously in reference 14.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.