LETTER

doi:10.1038/nature14966

Single-cell messenger RNA sequencing reveals rare intestinal cell types

Dominic Grün[1,2]*, Anna Lyubimova[1,2]*, Lennart Kester[1,2], Kay Wiebrands[1,2], Onur Basak[1,2], Nobuo Sasaki[1,2], Hans Clevers[1,2] & Alexander van Oudenaarden[1,2]

Understanding the development and function of an organ requires the characterization of all of its cell types. Traditional methods for visualizing and isolating subpopulations of cells are based on messenger RNA or protein expression of only a few known marker genes. The unequivocal identification of a specific marker gene, however, poses a major challenge, particularly if this cell type is rare. Identifying rare cell types, such as stem cells, short-lived progenitors, cancer stem cells, or circulating tumour cells, is crucial to acquire a better understanding of normal or diseased tissue biology. To address this challenge we first sequenced the transcriptome of hundreds of randomly selected cells from mouse intestinal organoids[1], cultured self-organizing epithelial structures that contain all cell lineages of the mammalian intestine. Organoid buds, like intestinal crypts, harbour stem cells that continuously differentiate into a variety of cell types, occurring at widely different abundances[2]. Since available computational methods can only resolve more abundant cell types, we developed RaceID, an algorithm for rare cell type identification in complex populations of single cells. We demonstrate that this algorithm can resolve cell types represented by only a single cell in a population of randomly sampled organoid cells. We use this algorithm to identify *Reg4* as a novel marker for enteroendocrine cells, a rare population of hormone-producing intestinal cells[3]. Next, we use *Reg4* expression to enrich for these rare cells and investigate the heterogeneity within this population. RaceID confirmed the existence of known enteroendocrine lineages, and moreover discovered novel subtypes, which we subsequently validated *in vivo*. Having validated RaceID we then applied the algorithm to *ex vivo*-isolated *Lgr5*-positive stem cells and their direct progeny. We find that *Lgr5*-positive cells represent a homogenous abundant population of stem cells mixed with a rare population of *Lgr5*-positive secretory cells. We envision broad applicability of our method for discovering rare cell types and the corresponding marker genes in healthy and diseased organs.

Single-cell mRNA sequencing has emerged as a powerful method to simultaneously measure cell-to-cell expression variability of thousands of genes[4]. Recently, it was demonstrated that sequencing of randomly selected cells from spleen[5] and lung tissue[6] permits the identification of known cell types within these organs. The approaches used in these and other recently published studies[7–10] show good performance in discovering abundant cell types but cannot detect rare cell types.

To profile cell types of widely varying abundance within a complex mixture we introduce a method for rare cell type identification (RaceID) and apply it to investigate rare cell types in the mouse small intestine[11].

The continuously self-renewing intestinal epithelium is arranged in crypts and villi. A small number of intestinal stem cells reside near the crypt bottom and give rise to rapidly proliferating transit amplifying (TA) cells. While migrating upward along the crypt–villus axis TA cells develop into the terminally differentiated cell types[2,12]. Absorptive enterocytes constitute the most abundant cell type, while all other mature cell types contribute only a few percent or less. The secretory lineage comprises mucus producing goblet, hormone secreting enteroendocrine, and Paneth cells, which provide a niche for the stem cell and secrete bactericidal products. In addition, tuft cells are believed to sense the luminal content.

To obtain clean random mixtures of intestinal cells without contamination of non-epithelial cell types, we use intestinal organoids, small epithelial structures containing all major cell types found in the intestinal epithelium[1]. Using a modified version of the cell expression by linear amplification and sequencing (CEL-seq) method[13] incorporating unique molecular identifiers to count transcripts[14] (Fig. 1a and Extended Data Fig. 1), we sequenced 238 randomly selected organoid cells with more than 3,000 transcripts in total each, and quantified 3,777 genes with more than five transcripts in at least one cell.

Hierarchical clustering of the transcriptome correlation matrix suggested the presence of three major groups of cells (Extended Data Fig. 2a). To screen for abundant cell types more systematically we employed *k*-means clustering of the correlation matrix with six clusters as inferred by the gap statistic[15] (see Methods, Fig. 1b and Extended Data Fig. 2b). We visualized these clusters in two dimensions (Fig. 1c) using *t*-distributed stochastic neighbour embedding (t-SNE)[16] and examined if expression of known intestinal marker genes was restricted to specific clusters (Extended Data Fig. 3). The intestinal alkaline phosphatase (*Alpi*) is a known enterocyte marker and showed a gradual expression increase across clusters 1, 4 and 5 (Extended Data Fig. 3a). Cluster 3 comprises distinct cells with non-overlapping expression of marker genes for diverse secretory cell types, such as the enteroendocrine marker *Chga*, the goblet cell marker *Muc2*, or the Paneth cell marker *Lyz1* (Extended Data Fig. 3b–d). The central cluster 2 does not express specific marker genes, but shows pronounced expression of genes encoding ribosomal proteins (Extended Data Fig. 3e), indicating the presence of transit amplifying cells. The bottom part of this cluster contains cells expressing low levels of the stem cell marker *Lgr5* (Extended Data Fig. 3f).

To detect rare cell types, we screened for outliers that could not be explained by a background model accounting for technical and biological gene expression noise (see Methods, Fig. 2a and Extended Data Fig. 4). Distinct outliers were grouped into 10 novel clusters based on transcriptome correlation (see Methods, Fig. 2b). Differential gene expression analysis revealed the presence of rare cell types among these clusters, comprising goblet, tuft, Paneth and enteroendocrine cells (Fig. 2c and Supplementary Table 1). Moreover, RaceID detected three secretory precursor clusters, co-expressing *Neurog3* with *Krt7*, *Pax4* or *Ang4* (Fig. 2c). Available methods for cell type identification[5] were clearly out-performed by RaceID (Extended Data Fig. 5a and Supplementary Note). Extensive experimental validation proved the

[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), 3584 CT Utrecht, The Netherlands. [2]University Medical Center Utrecht, Cancer Genomics Netherlands, 3584 CG Utrecht, The Netherlands.

*These authors contributed equally to this work.

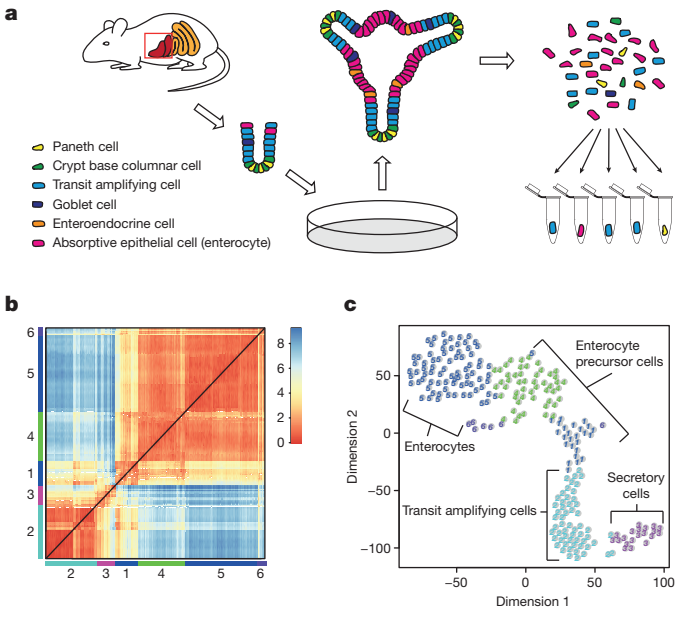


Figure 1 | Profiling cell composition of mouse intestinal organoids with single-cell sequencing. a, Intestinal crypts were isolated from mice and grown into intestinal organoids as described previously[1]. Organoids were dissociated and single cells, collected by fluorescence-activated cell sorting (FACS), were sequenced by a modified version of the CEL-seq method[13,14] (see Methods). **b**, Heat map indicating similarities between 238 single cells measured by Euclidean distances of the transcriptome correlation matrix (unitless; see Methods). *k*-means clustering identified six major groups of cells colour coded along the axes. **c**, t-SNE map representation of transcriptome similarities between individual cells. Clusters identified in **b** were highlighted with different colours and corresponding intestinal cell types identified on the basis of known marker genes are indicated.

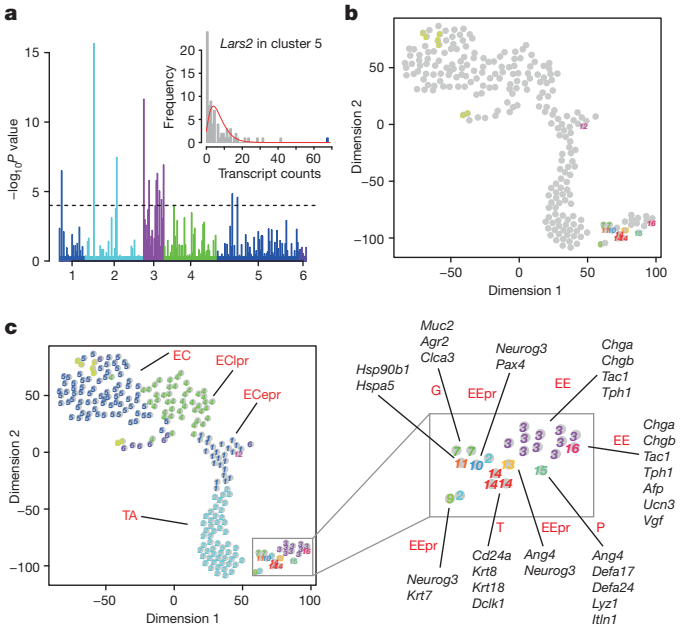


Figure 2 | RaceID algorithm identifies rare cell types among hundreds of sequenced cells. a, Histogram showing the negative logarithm of the probability that transcript levels in a particular cell are not explained by a background model accounting for the expected variability. Clusters are highlighted as in Fig. 1. The probability threshold for outlier identification (10^{-4}) was included (black broken line). The inset explains the derivation of outlier probabilities (see Methods) for the gene *Lars2* in cluster 5. The transcript count histogram (grey) is compared to the background model (red) and an outlier event is highlighted (purple). **b**, t-SNE map with additional clusters highlighted that were inferred from outlier cells (see Methods). **c**, t-SNE map of all clusters obtained by the RaceID algorithm (left) and close-up of all clusters in the secretory lineage (cluster 3 in Fig. 1c). Genes corresponding to known markers specifically upregulated in each cluster of the secretory lineage are indicated. EC, enterocytes; EClpr, late enterocyte precursors; ECepr, early enterocyte precursors; TA, transit amplifying cells; G, goblet cells; EE, enteroendocrine cells; EEpr, enteroendocrine precursors; P, Paneth cells; T, tuft cells.

specificity and sensitivity of RaceID (Extended Data Fig. 4c–e and Supplementary Note). We further showed that cell cycle related genes are unlikely to affect the results of RaceID for our data set (Extended Data Fig. 5f).

Enteroendocrine cells control metabolism by secreting at least ten different hormones[3,17] and individual cells produce subsets of those[18,19]. To profile heterogeneity of enteroendocrine cells, we aimed at purifying a random population of mature enteroendocrine cells. We identified enteroendocrine markers by a *z*-score analysis (Fig. 3a). Among the top scoring genes were novel markers such as the proteinase *Pappa2* and the largely uncharacterized gene *Reg4*. We focused on the latter, since it was highly expressed in enteroendocrine cells with hundreds of sequenced mRNAs. We validated *Reg4* as an *in vivo* marker for enteroendocrine cells by single-molecule fluorescent *in situ* hybridization[20] (smFISH) in the mouse intestine. Co-staining of *Chga* and *Reg4* revealed high levels of *Reg4* in enteroendocrine cells and substantially lower levels in Paneth cells at the crypt bottom (Fig. 3b). We then purified *Reg4*-positive organoid cells derived from a *Reg4*–red fluorescent protein (dsRed) reporter mouse (see Methods and Extended Data Fig. 6). RaceID predicted three major groups among the 161 cells surviving our filtering criteria (Fig. 3c, d). Upregulation of defensins suggested that one of these groups comprises maturation stages of the Paneth cell lineage (Extended Data Fig. 6e). Within the second group of cells we identified a contamination with TA cells (Extended Data Fig. 6f, g). The remaining 60 cells (37%) arise from the enteroendocrine lineage proving a pronounced enrichment (~eightfold) of this rare cell. We observed two major subgroups with low and high levels of *Chga*, respectively (Fig. 3d and Extended Data Fig. 6h). Hormones expressed in cells with low levels of *Chga* comprise *Cck*, *Ghrl*, *Sct*, *Nts* and *Gcg* (Extended Data Fig. 7), identifying this group as an intestinal specific branch of enteroendocrine cells[3,18,19].

The other sub-group co-expressed *Tph1* and *Tac1*, indicating hormone production of serotonin and substance P, respectively, and therefore comprises enterochromaffin cells (Fig. 3d and Extended Data Fig. 8a, b).

Within this sub-group RaceID identified three novel subtypes (Fig. 3d, e). In 7 out of 16 (41%) cells within cluster 3 we detected co-expression of *Tac1* and *Cck* (Extended Data Fig. 7a), previously considered to be markers of separate subtypes[3]. Expression of urocortin 3 (*Ucn3*), a ligand of the corticotropin-releasing hormone (*Crh*) receptor type 2 was significantly elevated in cluster 7 ($P < 4.1 \times 10^{-4}$, see Methods and Extended Data Fig. 8c). Although colonic expression of *Ucn3* has been described[21], it was not known to be expressed in the small intestine. Finally, we observed strong upregulation of Albumin (*Alb*, $P < 4.3 \times 10^{-6}$, see Methods) and the related alphafeto-protein (*Afp*, $P \sim 0$, see Methods) in cluster 2 (Extended Data Fig. 8d). Albumin can bind lipophilic hormones and could regulate their accessibility[22]. In the same cluster, we measured upregulation of *VGF* nerve growth factor (Extended Data Fig. 8e). See Extended Data Fig. 8f and Supplementary Table 2 for additional marker genes.

We validated the existence of the novel enterochromaffin subtypes *in vivo* at the mRNA and protein level by conducting smFISH and immunofluorescence experiments in the mouse intestinal epithelium (Fig. 4, Extended Data Fig. 9, Methods and Supplementary Table 4).

After having shown that RaceID can discriminate cell types we wanted to test our method on stem cells marked by *Lgr5* expression. Heterogeneity of the intestinal stem cell pool is still controversial[23–27] and single-cell sequencing could help to better characterize this

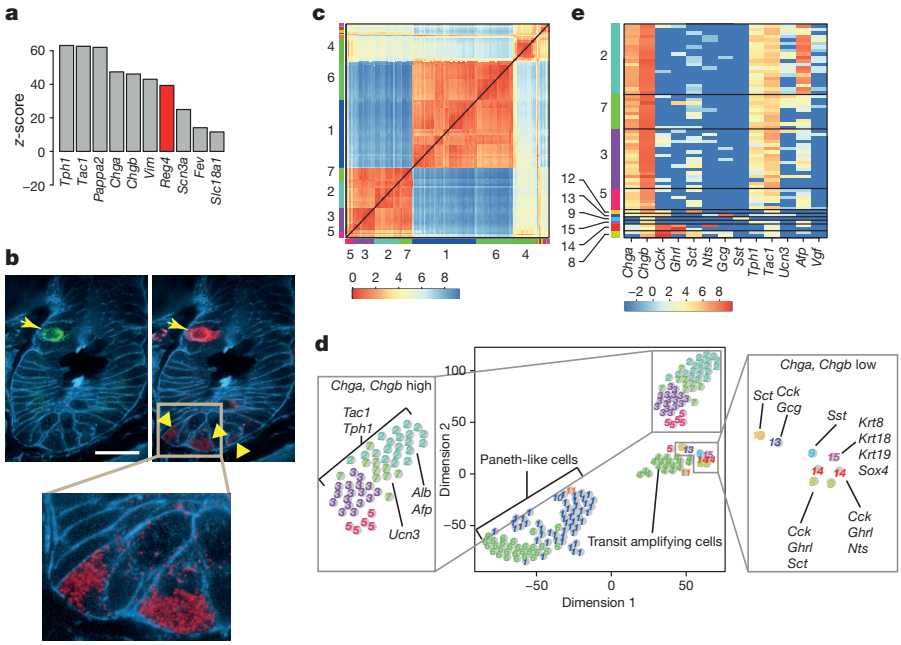


Figure 3 | *Reg4* is a novel marker of differentiated enteroendocrine cells. **a**, Histogram of top ten *z*-scores for upregulation in differentiated enteroendocrine cells. For each gene, we extracted the average level observed in mature enteroendocrine cells (cluster 3 and 16), subtracted the average level observed across all remaining cells, and divided by the standard deviation of transcript levels in these cells. **b**, Validation of *Reg4* as an *in vivo* marker of mature enteroendocrine cells by single-molecule fluorescent *in situ* hybridization (smFISH). Cryosections of mouse small intestine were hybridized with smFISH probes against *Chga*, conjugated to tetramethylrhodamine (TMR, green) and *Reg4*, conjugated to cyanine 5 (Cy5, red). Cell borders were visualized with AlexaFluor 488-conjugated phalloidin (blue). The enteroendocrine cell (arrow) expresses a high level of *Chga* and co-expresses *Reg4*. Paneth cells at the crypt bottom (arrowheads and inset) express strongly reduced levels of *Reg4*. Scale bar, 20 µm. **c**, Heat map representing the transcriptome similarities measured by the Euclidean distance of the transcriptome correlation matrix (see Methods) for *Reg4*-positive cells. RaceID clusters are colour-coded along the axes. Cluster numbers are shown for the bigger clusters. **d**, t-SNE map showing all clusters identified by RaceID for *Reg4*-positive cells. Different colours and numbers highlight distinct clusters. Selected upregulated genes are shown for individual clusters ($P < 10^{-3}$, see Methods). Close-ups are shown for clusters of enteroendocrine cells. **e**, Heat map of hormone expression ($\log_2$ scale) in subtypes of enteroendocrine cells identified by RaceID. The two groups of cells with low and high expression of *Chga*, respectively, display distinct patterns.

population. We first used intestinal organoids derived from an Lgr5–green fluorescent protein (GFP) reporter mouse[23] to purify and sequence 96 Lgr5–GFP$^+$ cells. RaceID detected only a single large cluster and few outliers that were mostly Paneth cells (Extended Data Fig. 10), suggesting that intestinal stem cells represent a uniform population. Since a distinct reserve pool of quiescent *Lgr5*-positive cells has been suggested[25–27], we next tried to characterize a population of *ex vivo* isolated *Lgr5* expressing cells. For this experiment, we isolated 192 Lgr5–enhanced GFP$^+$ (EGFP$^+$) cells from the small intestine of an Lgr5–EGFP reporter mouse[24] (Supplementary Table 5). RaceID classified these cells into a single large homogenous cluster and a few outliers (Fig. 5a, b).

As a complementary approach, we traced the progeny of *Lgr5*-positive cells *in vivo*, using a reporter mouse that expresses CreERT2 from an *Lgr5* promoter and YFP from a Rosa26 promoter with a loxP-flanked transcriptional roadblock. Administration of tamoxifen leads

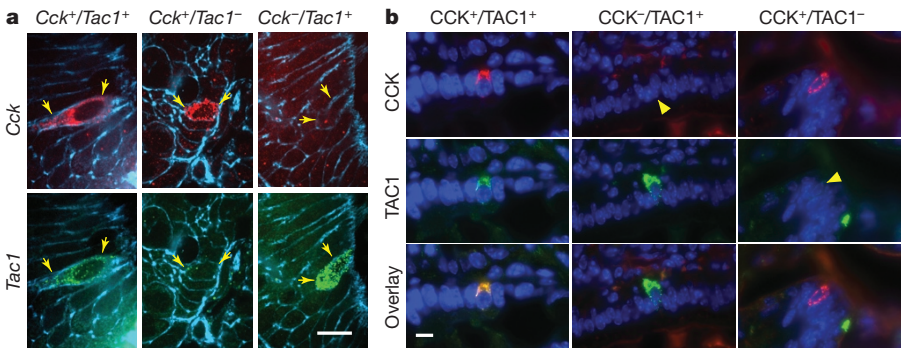


Figure 4 | Single-molecule FISH and immunofluorescence experiments confirm expression of markers for enteroendocrine cell sub-populations in the mouse small intestine. **a**, Small intestine cryosections were hybridized with smFISH probe libraries. Scale bar, 10 µm. *Cck* and *Tac1* are expressed by a subset of enteroendocrine cells. Probes against *Cck*, conjugated to Cy5 (upper panel, red) and against *Tac1*, conjugated to TMR (lower panel, green) were used for hybridization. Cell borders were visualized by staining with phalloidin, conjugated with AlexaFluor 488 (blue). Enteroendocrine cells co-expressing the two markers were observed (left), as well as cells expressing only *Cck* (middle) or *Tac1* (right). Arrows point at cell borders. **b**, Immunostaining was performed on cryosections of mouse small intestinal tissue. Scale bar, 20 µm. Expression of CCK and TAC1 was observed in a subset of enteroendocrine cells. CCK and TAC1 were visualized by indirect immunostaining with antibodies against CCK (upper panel, red) and against TAC1 (middle panel, green). Nuclei were counterstained with DAPI (blue). Rare cells, co-expressing the two markers, were observed (left), as well as cells, expressing only TAC1 (middle), or only CCK (right). Arrowheads point at CCK or TAC1-negative cells.

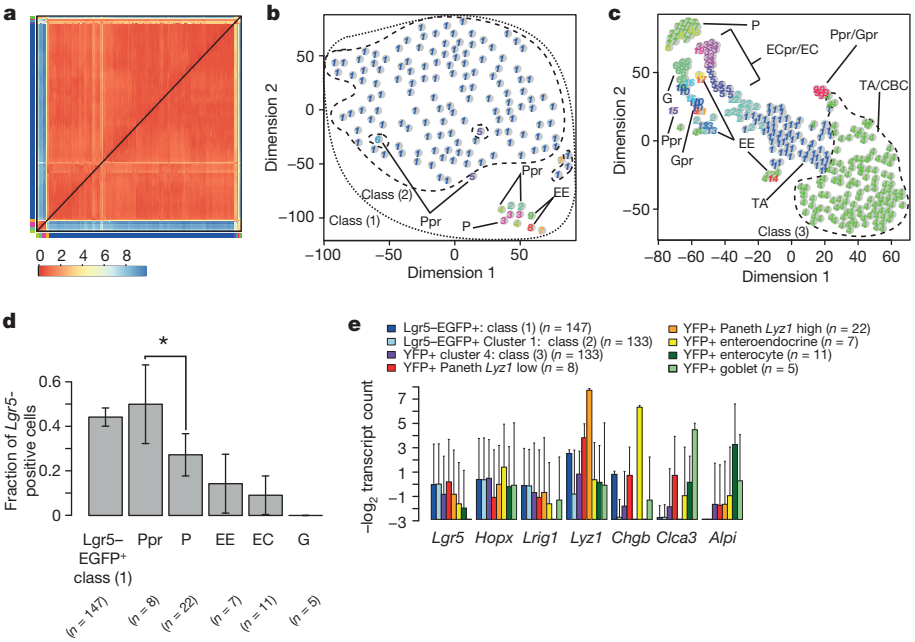


Figure 5 | Characterization of *ex vivo* isolated Lgr5–EGFP$^+$ cells. **a**, Heat map of transcriptome similarities measured by the Euclidean distance of the transcriptome correlation matrix (see Methods) for Lgr5–EGFP$^+$ cells purified using an Lgr5–EGFP reporter mouse[24]. RaceID clusters are colour coded along the axes. Colours correspond to panel **b**. **b**, t-SNE map of RaceID clusters for Lgr5–EGFP$^+$ cells. Cells of cluster 7 express non-coding RNAs (*Malat1*, *Kcnq1ot1*) and could not be characterized. The other cell types were assigned based on marker gene expression. The stem cell classes (1) (dotted line) and (2) (dashed line) are outlined (see text for details). **c**, t-SNE map of RaceID clusters for 5-day lineage tracing progeny of Lgr5-positive (YFP$^+$) cells. Cell types identified based on marker genes are indicated. The stem cell class (3) is outlined (dashed line, see text for details). **d**, Fraction of *Lgr5*-positive cells for the major intestinal lineages. See main text for details. Error bars and *P* value were inferred from binomial statistics reflecting uncertainty due to sampling. The number of cells (*n*) is indicated below the plot. *$P < 0.05$. **e**, Mean expression of marker genes in different sets of *Lgr5*-positive cells and in cells of the major intestinal lineages. Class (1) corresponds to all Lgr5–EGFP$^+$ cells shown in **b**, while class (2) and (3) correspond to the sub-populations indicated by the dashed line in **b** and **c**, respectively. Error bars indicate the standard deviation across cells. The number of cells (*n*) is given in the legend. EC, enterocytes; ECpr, enterocyte precursors; TA, transit amplifying cells; CBC, crypt base columnar cells; G, goblet cells; Gpr, goblet cell precursors; EE, enteroendocrine cells; P, Paneth cells; Ppr, Paneth cell precursors; Ppr/Gpr, early precursors of the Paneth and goblet cell lineage.

to yellow fluorescence protein (YFP) protein production in *Lgr5*-positive cells and their progeny. We sequenced 432 YFP-positive cells collected five days after label induction (Supplementary Table 5). As expected, RaceID detected differentiated cells of all major lineages, with a relatively large proportion of Paneth cells (Fig. 5c). A possible explanation for the over-representation of Paneth cells is label induction in mature Paneth cells or their precursors[27]. We then quantified the fraction of *Lgr5*-positive cells in all major lineages. To extract cells of a lineage independent of the maturation state, we only required >5 transcripts of a lineage marker (*Lyz1* for Paneth cells, *Chgb* for enteroendocrine cells, *Alpi* for enterocytes, and *Clca3* for goblet cells). Paneth cells were split into early and late stages, with *Lyz1* expression lower or higher than the median, respectively. While we detected *Lgr5* transcripts in ~45% of the *Lgr5*–EGFP$^+$ cells, this fraction was lower than 15% for most of the other major cell types (Fig. 5d). Only for Paneth cells we observed an elevated proportion of *Lgr5*-positive cells, which was significantly higher in early (~50%) compared to late Paneth cells (~28%) (Fig. 5d). This could be due to the *Lgr5* RNA half-life exceeding the rapid transition time of stem cells into Paneth cells, leading to a transient state where stem and Paneth cell genes are co-expressed.

To examine if the population of *Lgr5* expressing cells show any kind of fate bias towards a particular lineage, we first distinguished three classes of stem cells: (1) all Lgr5–EGFP$^+$ cells (Fig. 5b), (2) the subset of Lgr5–EGFP$^+$ cells after removal of the outliers identified by RaceID (Cluster 1, Fig. 5b), and (3) the stem cell/early TA cluster from the lineage tracing data (Cluster 4, Fig. 5c). Based on the RaceID prediction we consider class (2) as a homogenous pool of stem cells, while class (1) contains a few additional *Lgr5* expressing cells of other lineages. Class (3) represents the homogenous stem cell population identified by RaceID in the lineage tracing data and is thus expected to resemble class (2). We then performed a marker gene analysis in all three classes and, for comparison, in cell populations of all major lineages (Fig. 5e). The three classes of stem cells showed similar expression of *Lgr5* and other stem cell markers (*Hopx* and *Lrig1*). Expression of secretory lineage markers (*Lyz1* and *Chgb*) was substantially lower in class (2) and (3) compared to class (1), and did not exceed the background level observed in any other lineage (Fig. 5e). This argues against additional secretory cells in class (2) and (3) that could have remained undetected by RaceID. Elevated expression of *Lyz1* and *Chgb* in class (1) is thus solely due to the few Lgr5–EGFP$^+$ secretory cells identified by RaceID (Fig. 5b). Interestingly, *Lgr5* transcript levels in early Paneth cells were similar to those in stem cells and reduced in the other cell types (Fig. 5e).

Taken together, we conclude that, both in organoids and *in vivo*, *Lgr5*-positive cells represent a homogenous population of cells mixed with a rare population of Paneth and enteroendocrine cells. In comparison to the other lineages, Paneth cells express the highest level of *Lgr5*, consistent with the observation that Paneth cell precursors can revert to the stem-cell state upon tissue damage[25,27]. It remains a possibility that high *Lgr5* expression in Paneth cells is an artefact due to sequencing doublets of Paneth and crypt bottom stem cells. However, we consider this unlikely, because expression of *Lgr5* is significantly elevated in early versus late Paneth cells (Fig. 5d). Finally, we would like to caution that heterogeneity among lowly expressed genes could still exist within the stem cell pool, which would be invisible owing to the limited sensitivity of current single cell sequencing protocols. Alternatively, stem cell heterogeneity could extend to *Lgr5*-low cells as described previously[26], which are not captured by our enrichment strategy.

In summary, we demonstrated here the ability of RaceID to correctly classify different cell types in a complex mixture and reveal heterogeneity

among rare cells. We believe that single-cell mRNA sequencing in combination with the RaceID algorithm is a powerful tool to unravel heterogeneity of rare cell types in both healthy and diseased organs.

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

Received 14 November 2014; accepted 23 July 2015.
Published online 19 August 2015.

1. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. *Nature* **459,** 262–265 (2009).
2. van der Flier, L. G. & Clevers, H. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annu. Rev. Physiol.* **71,** 241–260 (2009).
3. Engelstoft, M. S., Egerod, K. L., Lund, M. L. & Schwartz, T. W. Enteroendocrine cell types revisited. *Curr. Opin. Pharmacol.* **13,** 912–921 (2013).
4. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Rev. Genet.* **14,** 618–630 (2013).
5. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).
6. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).
7. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).
8. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (2014).
9. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnol.* **32,** 1053–1058 (2014).
10. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).
11. Clevers, H. The intestinal crypt, a prototype stem cell compartment. *Cell* **154,** 274–284 (2013).
12. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nature Rev. Mol. Cell Biol.* **15,** 19–33 (2014).
13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
14. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature Methods* **11,** 637–640 (2014).
15. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B.* **63,** 411–423 (2001).
16. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** 2570–2605 (2008).
17. Schonhoff, S. E., Giel-Moloney, M. & Leiter, A. B. Minireview: Development and differentiation of gut endocrine cells. *Endocrinology* **145,** 2639–2644 (2004).
18. Habib, A. M. *et al.* Overlap of endocrine hormone expression in the mouse intestine revealed by transcriptional profiling and flow cytometry. *Endocrinology* **153,** 3054–3065 (2012).
19. Egerod, K. L. *et al.* A major lineage of enteroendocrine cells coexpress CCK, secretin, GIP, GLP-1, PYY, and neurotensin but not somatostatin. *Endocrinology* **153,** 5782–5795 (2012).
20. Raj, A., Van Den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5,** 877–879 (2008).
21. Saruta, M. *et al.* Urocortin 3/stresscopin in human colon: possible modulators of gastrointestinal function during stressful conditions. *Peptides* **26,** 1196–1206 (2005).
22. Baker, M. E. Albumin, steroid hormones and the origin of vertebrates. *J. Endocrinol.* **175,** 121–127 (2002).
23. Tian, H. *et al.* A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. *Nature* **478,** 255–259 (2011).
24. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449,** 1003–1007 (2007).
25. Roth, S. *et al.* Paneth cells in intestinal homeostasis and tissue injury. *PLoS ONE* **7,** e38965 (2012).
26. Li, N. *et al.* Single-cell analysis of proxy reporter allele-marked epithelial cells establishes intestinal stem cell hierarchy. *Stem Cell Rep.* **3,** 876–891 (2014).
27. Buczacki, S. J. A. *et al.* Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature* **495,** 65–69 (2013).

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

Acknowledgements This work was supported by an European Research Council Advanced grant (ERC-AdG 294325-GeneNoiseControl) and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award.

Author Contributions D.G., A.L. and A.v.O. conceived the project. D.G. developed the algorithm, performed all computations and wrote the manuscript. A.L., L.K. and K.W. performed all sequencing experiments. A.L. performed the lineage tracing experiment and all imaging experiments. N.S. made the *Reg4*–dsRed mouse and was supervised by H.C.; O.B. contributed the Lgr5–GFP organoids. A.L., L.K., N.S. and H.C. edited the manuscript. A.v.O. guided experiments, data analysis and writing of the manuscript, and edited the manuscript.

Author Information RNA-seq data are deposited in Gene Expression Omnibus, accession number GSE62270. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu).

METHODS

No statistical methods were used to predetermine sample size, and the experiments were not randomized

Generation of the *Reg4*–dsRed mouse. *Reg4*–dsRed knock-in mice were generated by homologous recombination in embryonic stem cells by targeting a diphtheria toxin receptor-2A peptide–dsRed express2 cassette to the ATG start codon of *Reg4*. Generation of the knock-in mouse and experiments were performed according to guidelines and reviewed by the Dier Experimenten Commissie (DEC) of the KNAW.

***Lgr5*–GFP organoids.** Organoids from *Lgr5*–GFP-DTR reporter mice[23] were cultured under standard contitions[1]. Nine days after splitting, the organoids were dissociated into single cells using TrypLE (Invitrogen) for 15 min at 37 °C and mechanical disruption using a glass Pasteur pipette. Cells were washed twice in advanced DMEM/F12 (GIBCO) and resuspended in advanced DMEM/F12 + 4 μg μl^{-1} DNase I (Roche) and Propidium Iodide (Sigma). Among PI negative fraction, high level of GFP (top 10%) expressing cells were sorted directly in 96-wells containing 100 μl TRIzol.

Lineage tracing experiments. For lineage tracing experiments we injected 0.4 mg tamoxifen into 3-month-old Lgr5-CreERT2 C57Bl6/J mice crossed to a Rosa26LSL-YFP reporter mice.

Isolation of crypts from mouse small intestine. Crypts were isolated from Reg4–dsRed mouse as described previously[1]. Briefly, the whole of the small intestine was dissected, flushed with cold Ca^{2+}- and Mg^{2+}-free PBS and cut to 4–5 cm pieces for convenience. Intestines were cut open longitudinally and villi were scraped off with a glass slide. Intestine fragments were washed twice with cold Ca^{2+}- and Mg^{2+}-free PBS, then incubated with 5 mM EDTA in PBS at 4 °C for 30 min, with gentle agitation. Crypts were released by vigorous shaking of the tissue fragments, pelleted by centrifugation (200*g* at 4 °C for 5 min), washed once with cold PBS and once with Advanced DMEM/F12 medium (Life Technologies), pelleted by centrifugation and used to generate organoids.

Isolation of Lgr5–EGFP^{+} cells from the mouse intestine. Freshly isolated small intestines of Lgr5-EGFP-IRES-creERT2 mice[24] were incised along their length and villi were removed by gentle scraping. The tissue was then washed in ice-cold PBS0 and subsequently incubated in PBS0/EDTA (5 mM) for 5 min followed by gentle shaking to remove remaining villi. The intestine was then incubated in PBS/EDTA for 30 min at 4 °C. Vigorous shaking yielded free crypts that were incubated in MEM (Gibco) supplemented with trypsin (2 mg ml^{-1}; Sigma) and DNase I (2,000 U ml^{-1}; Sigma) for 30 min at 37 °C. Subsequently, cells were spun down, resuspended in MEM/DNase and DAPI (Life Technologies) and filtered through a 40-mm mesh. DAPI-negative, GFP-expressing cells were directly sorted in TRIzol (Life Technologies) using a BD FACSAria II cell sorter (BD Bioscience).

Intestinal organoid culture. Villin-Cre organoids were a gift from H. Farin. Organoid culture was carried out as described[1]. Briefly, organoids were grown in a drop of Matrigel (BD Biosciences), overlaid with the ENR medium (see below). Organoids were passaged weekly, with 1:4 dilution. Briefly, the old medium was aspirated, old Matrigel drop broken, organoids were washed and pelleted by centrifugation at 200–300*g*, then mixed with fresh Matrigel. Drops of Matrigel–organoid mix were placed on the bottom of tissue culture dish, let to solidify, and overlaid with ERM culture medium. Reg4–dsRed organoids were derived from the Reg4–dsRed mouse as described[1]. Briefly, small intestinal crypts were isolated, pelleted and mixed with Matrigel. Crypt–Matrigel mix was plated and cultured in the ENR medium. Newly generated organoids were expanded in culture for at least three weeks before harvesting for the experiment. The ENR medium is an Advanced DMEM/F12 medium, supplemented with penicillin/streptomycin, 1× GlutaMAX, 10 mmol l^{-1} HEPES, 1XB27 (Life Technologies), Noggin (conditioned medium, 10% volume), R-spondin 1 (conditioned medium, 10% volume), 1 mmol l^{-1} N-acetylcysteine (Sigma) and 50 ng ml^{-1} recombinant mouse EGF (Peprotech). The R-spol 1 and Noggin conditioned media were generated in HEK293T cells, stably expressing HA-mouse Rspo1-Fc (gift from Calvin Kuo, Stanford University), or transiently transfected with mouse Noggin-Fc plasmid. Advanced DMEM/F12 with penicillin/streptomycin, 10 mmol l^{-1} HEPES, and 1× GlutaMAX was conditioned for 1 week.

Single-cell suspension preparation, FACS sorting and RNA extraction. Organoids were dissociated to single cells as previously[28], with a few modifications. Briefly, medium was removed from organoid cultures, organoids were resuspended in TryplE, incubated at 37 °C for 5–15 min, with passaging through a glass pipette each 5 min and microscopic monitoring. Upon are disruption of most of the cell aggregates, cells were pelleted by centrifugation (5 min at 300–400*g*), washed twice with Advanced DMEM/F12 with 10% fetal calf serum, resuspended in the same medium. Cells were strained through 20 μm mesh filter and stained with either propidium iodide (Villin-Cre cells), or DAPI (Reg4–dsRed cells). Single cells were then sorted by flow cytometry (FACS Aria, BD). Each cell was sorted directly into single wells of 96-well PCR plates, each well containing 100 μl TRIzol reagent. Identical quantity of the ERCC Spike-in RNA (0.03 μl of 1:50,000 dilution) was added to each well. Total RNA was extracted from each cell, according to the TRIZOL manufacturer's protocol with a few alterations. To facilitate RNA precipitation, 0.2 μl of GlycoBlue reagent (Life Technologies) was added to each sample. Isopropanol precipitation was carried out overnight. RNA pellets were air dried for up to 15 min, then resuspended in the CEL-Seq first-strand primer solution.

Control library with mouse embryonic stem cells and fibroblasts. Irradiated mouse embryonic fibroblasts were cultured in DMEM containing 10% FBS (Gibco), 2 mM GlutaMAX (Gibco), 0.1 mM MEM nonessential amino acids and 1% Pen/Strep (Gibco). Wild-type mouse embryonic stem cells were derived from C57BL6 mice and cultured in DMEM containing 10% FBS (Gibco), 2 mM GlutaMAX (Gibco), 0.1 mM MEM nonessential amino acids, 1% Pen/Strep (Gibco) and 1,000 U LIF ml^{-1} (ESGRO). The control library contained 5 ESCs with barcodes 1–5, 5 MEFs with barcodes 6–10, 75 mouse small intestinal organoids cells with barcodes 11–85, 6 controls without template but with reverse transcription primer, barcodes 87–92 and 5 empty controls.

Pool-and-split control sample preparation. Organoids, derived from the *Reg4*–dsRed knock-in mouse, were dissociated to a single-cell suspension and dsRed-positive cells were sorted using FACS, as described above. Pools of 100 cells were collected into single tubes containing 100 μl TRIzol reagent and processed for the total RNA extraction. Total RNA from 100 cells was resuspended in nuclease-free water and ERCC Spike-in RNA was added to the RNA solution (3 μl of 1:50,000 dilution, per 100 pooled cells). Then each of the pooled samples was split into 100 separate portions. CEL-Seq was performed on 200 resulting split samples.

Tissue preparation and immunofluorescence. Freshly dissected mouse small intestines were flushed and fixed in cold 4% paraformaldehyde in PBS for 3 h. After fixation, the intestines were incubated in cryoprotective solution (30% sucrose in PBS) at 4 °C overnight, then frozen blocks were prepared in Tissue-Tek OC. compound (VWR) and stored at −80 °C. Five μm thick cryosections were cut and mounted on poly-L-lysine (Sigma)-coated cover glass. Sections were fixed in 4% paraformaldehyde for 15 min, permeabilized with 0.25% Triton X100 (Sigma) on PBS for 5 min and blocked for 1 h in PBS containing 0.2% Triton X100, 1% BSA (Sigma), and 2% each normal donkey (Jackson) and goat (Monosan) serum. Next, the sections were incubated for 1 h with primary antibody solution, (in PBS with 1% BSA), washed, and incubated for 1 h with secondary antibody (in PBS). Nuclei counterstain was done with DAPI (100 ng ml^{-1} in PBS) for 10 min. The sections were then washed and mounted in Fluoromount-G (Electron Microscopy Sciences). Imaging was done on Leica fluorescence microscope with a 100× oil immersion objective, using MetaMorph imaging software. Images were processed and combined using ImageJ and Photoshop programs.

Antibodies. For indirect immunofluorescence the following antibodies were used: anti-mouse Urocortin 3 rabbit polyclonal antibody (Yanaihara Institute, Y364); anti-mouse CCK rabbit polyclonal antibody (LifeSpan BioSciences, aa26-33, LS-C190673); anti-mouse VGF rabbit polyclonal antibody (Abcam, ab69989); anti-mouse Tac1 guinea pig polyclonal antibody (Abcam, ab10353). The following secondary antibodies were used: Goat anti-rabbit IgG, Cy5-conjugated (Life Technologies, A10523); Donkey-anti guinea pig IgG, TRITC-conjugated (Jackson Labs, 706-025-148). For direct immunofluorescence we used the mouse monoclonal antibody against AFP, conjugated to Alexa Fluor 594 (Cell Signaling, 7877).

CEL-seq library preparation. Single cells were processed using the previously described CEL-seq technique[13], with several modifications. A 4-bp random barcode as unique molecular identifier (UMI) was added to the primer in between the cell specific barcode and the poly T stretch (Supplementary Table 3). Dried RNA, prepared from single cells by TRIzol extraction method, was resuspended in primer solution, denatured at 70 °C for 2 min and quickly chilled, after which the first strand synthesis mix was added. The rest of the protocol was carried out as published[13], with no substantial alterations. Libraries were sequenced on an Illumina HighSeq 2500 using 50 bp paired end sequencing.

Single-molecule FISH. Probe libraries were designed and fluorescently labelled as previously described[20]. All probe libraries consist of 20 to 39 oligonucleotides of 20-bp length (see Supplementary Table 3 for probe sequences) complementary to the coding sequence of the genes. Cells were hybridized overnight with probes at 30 °C, as previously described[20]. DAPI and phalloidin-AlexaFluor488 staining was done after washes. Images were acquired on a Perking-Elmer Spinning Disc confocal microscope with a 100× oil-immersion objective (numerical aperture 1.4) using Perking Elmer Volocity software. Images were recorded as stacks with a *z* spacing of 0.3 μm. Diffraction-limited dots corresponding to single mRNA molecules were automatically detected using custom Matlab software, based on previously described algorithms[20]. Briefly, the images were first filtered using a three-dimensional Laplacian of Gaussian filter, followed by selection of the intensity

threshold at which the number of connected components was least sensitive to the threshold.

Quantification of transcript abundance. Paired end reads obtained by CEL-seq were aligned to the transcriptome using bwa[29] (version 0.6.2-r126) with default parameters. The transcriptome contained all RefSeq gene models based on the mouse genome release mm10 downloaded from the UCSC genome browser[30] and contained 31,109 isoforms derived from 23,480 gene loci. All isoforms of the same gene were merged to a single gene locus. The right mate of each read pair was mapped to the ensemble of all gene loci and to the set of 92 ERCC spike-ins[31] in sense direction. Reads mapping to multiple loci were discarded. The left read contains the barcode information: the first eight bases correspond to the cell specific barcode followed by 4 bases representing the unique molecular identifier (Supplementary Table 3). The remainder of the left read contains a polyT stretch followed by few (<15) transcript-derived bases. The left read was not used for quantification. For each cell barcode we counted the number of unique molecular identifiers for every transcript and aggregated this number across all transcripts derived from the same gene locus. Based on binomial statistics we converted the number of observed unique molecular identifiers into transcript counts[14].

Rare cell type identification algorithm RaceID. *Data preparation.* The clustering algorithm takes as input a matrix with transcript counts for all genes in each cell. As a first preprocessing step cells with low overall transcript counts are removed. We require at least 3,000 transcripts per cell for the whole organoid data and 1,000 transcripts per cell for the *Reg4*-positive cells. For the latter we observed overall lower transcript numbers. Next, the total transcript count within each cell is normalized to the median transcript number across cells. Alternatively, downsampling of the transcript pool to the required minimal total transcript count can be applied. Hereafter, we add a pseudocount of 0.1 to the expression value of each gene to avoid divergences when computing fold changes. In the next step lowly expressed genes are filtered out. Genes that are not expressed with a minimum of five transcripts for the whole organoids data and three transcripts for the *Reg4*-positive cells in at least a single cell are discarded. Furthermore, highly expressed genes that saturate the pool of available UMIs (>500 transcripts after normalization for the whole organoid data and >2,000 transcripts after normalization for the *Reg4*-positive cells) are discarded, since these genes potentially introduce artefacts in the clustering. For the *Reg4*-positive cells we amended this last filtering step, since hormones crucial for the cell type determination saturated the UMIs in only very few cells. We tested the robustness of the RaceID predictions using the more relaxed setting also for the random organoid cells (Extended Data Fig. 5e) and the more stringent settings for the *Reg4*-positive cells (Extended Data Fig. 6d), respectively. In each case, we observed the same rare cell types for the cells that survive the filtering criteria of both settings. We also analysed *Lgr5*-positive intestinal cells. Here we applied the same filtering criteria as used for the whole organoid data, since the single cell sequencing yielded a high number of transcripts per cell. For all *ex vivo* isolated cells we also applied these settings. However, the data were downsampled to the same transcript number in all cells (that is, from all cells subsets of transcripts are sampled with a size corresponding to the minimal total transcript count across all cells surviving the filtering step). This approach was applied since batch effects due to combining different libraries were more pronounced and downsampling reduces technical noise caused by variation in library complexity.

k-*means clustering.* The clustering step of RaceID identifies larger clusters of different cells by *k*-means clustering. First, a similarity matrix is computed that contains Pearson's correlation coefficients for all pairs of cells. Subtracting the coefficients from one yields a distance matrix, which serves as input for *k*-means clustering. *k*-means clustering is applied to the similarity matrix using the Euclidean metric. In comparison to direct clustering of the expression matrix this approach yielded improved cluster separation. The number of clusters used for *k*-means clustering is determined from the gap statistic[15], that is, from the difference of the average within cluster dispersion in uniformly distributed and in the actual data. By default, the cluster number is determined as the first local maximum of the gap statistic where the maximum exceeds its neighbours by >25% of their standard deviation. If the gap statistic does not exhibit a clear maximum, the cluster number demarcating the point where the gap statistic starts to saturate should be used as input for the *k*-means clustering. Given the number of clusters, *k*-means clustering of the distance matrix is performed and cluster reproducibility is assessed by bootstrapping using the clusterboot function of the R package fpc. The algorithm computes Jaccard's similarity to quantify cluster reproducibility. If more than a single cluster has a Jaccard's similarity lower than 0.5, the clustering should be repeated with fewer clusters. Importantly, the outlier identification step of the algorithm will correct for an underestimation of the actual cluster number and it is thus recommended to start with a conservative estimate.

Identification of outlier cells. To identify outlier cells within each cluster the algorithm evaluates transcript count variability of every gene across all cells in this cluster. The expected baseline level of expression variability, quantified by the transcript count variance, is inferred from the ensemble of all cells. A second order polynomial is fitted to the transcript count variance as a function of the average transcript count in logarithmic space (Extended Data Fig. 4a). In comparison to a linear regression the polynomial leads to a significant improvement of the regression as was tested by ANOVA model comparison for the data sets presented here (ANOVA $P < 2.2*10^{-16}$ for both data sets). For the random organoid cells the residual sum of squares was reduced by 20% when using the second order polynomial instead of the linear regression. This polynomial serves as an estimate of the expected variance-mean dependence under the assumption, that the majority of genes do not exhibit cluster (or cell type) specific expression. Next, each cluster is screened for outlier cells by computing the transcript count probability in each cell for a given gene from a negative binomial distribution defined by the average transcript count of this gene across all cells in the cluster and the expected variance computed by the second order polynomial. Assuming a lower limit of Poissonian noise, values of the expected variance lower than the mean are replaced by the mean. In practice, this does not happen and the regression yields noise estimates well above the Poissonian limit for all data sets analysed so far. If the multiple testing (Benjamini-Hochberg) corrected transcript count probability of a specified number of genes (two for our data) is below a defined probability threshold ($<10^{-4}$ for our data) in a given cell, this cell is considered an outlier. The total number of outliers can be plotted as a function of the probability threshold (Extended Data Fig. 4b), which should be chosen such that it separates the tail of this distribution from the bulk behaviour.

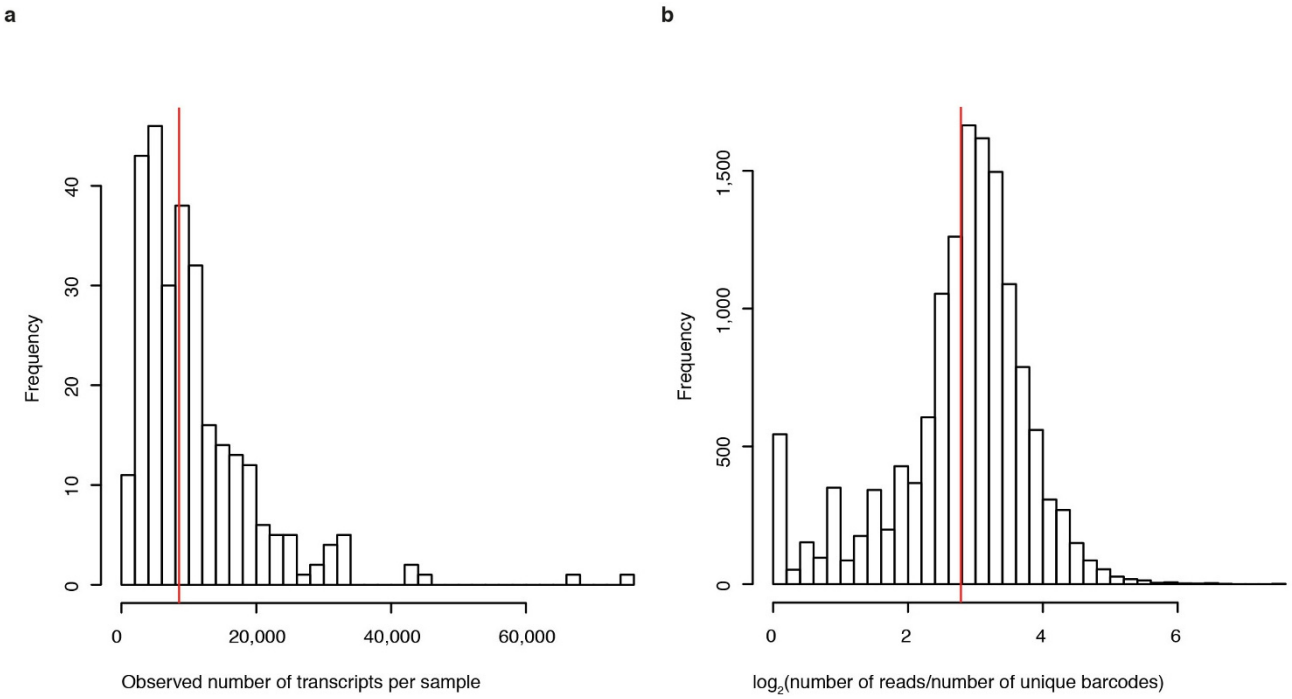
Inference of final clusters. Given the set of outlier cells the final clusters that should largely correspond to different cell-types or –states are inferred. To this end, outlier cells are first merged to outlier clusters if their transcriptome correlation exceeds the 75%-quantile of the distribution of cell-to-cell correlation within the original clusters after outlier removal. Subsequently, new cluster centres are computed for the remaining original and the new outlier clusters by averaging transcript counts within these clusters and each cell is reassigned to the most highly correlated cluster centre.

Two-dimensional representation of cell type maps. For visual inspection of cell clusters and associated cell types we apply a dimensional reduction of cell-to-cell distances as computed by the distance matrix (see above) using a machine learning algorithm termed t-distributed stochastic neighbour embedding (t-SNE)[16]. Briefly, this algorithm converts the original point-to-point distance distribution to a lower dimensional Student's *t*-distribution. The location of all points in the map is determined by a stochastic minimization of the Kullback–Leibler divergence of the original distances with respect to the mapped distances.

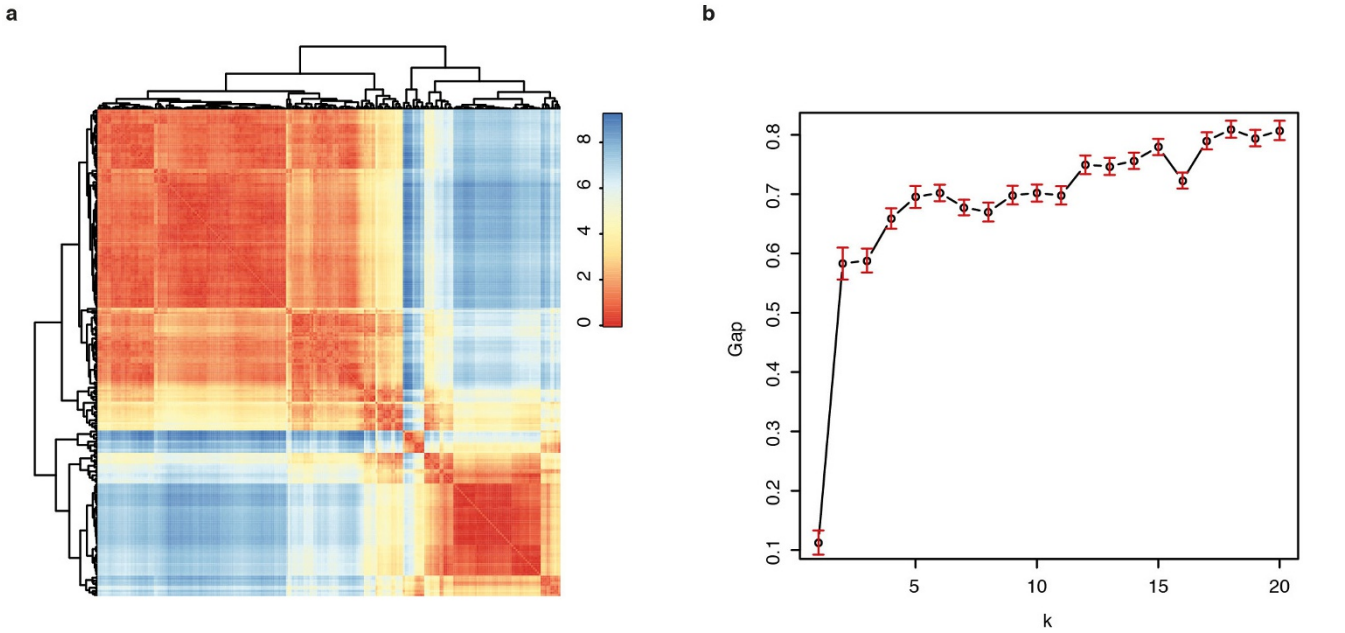
Identification of differentially expressed genes. To identify genes that were on average up- or downregulated within a cluster compared to the ensemble of all cells, the fold change in absolute transcript counts was computed after normalizing the total transcript count in a cell to the median transcript count within the cluster under consideration. As shown previously[14] for single cell sequencing data, median normalized transcript levels exhibit Poissonian noise for most genes with small deviations at high expression. A *P* value for significant up- or downregulation was therefore computed based on Poissonian statistics and multiple testing corrected by the Benjamini–Hochberg method.

Code availability. The RaceID algorithm is supplied as an R script (Supplementary Data 1) along with sample code (Supplementary Data 2), an extensive reference manual (Supplementary Data 3) and sample data (Supplementary Table 6), corresponding to the random organoid cell transcriptome data analysed in this paper. Bug fixes and updates of RaceID can be downloaded from https://github.com/dgrun/RaceID.

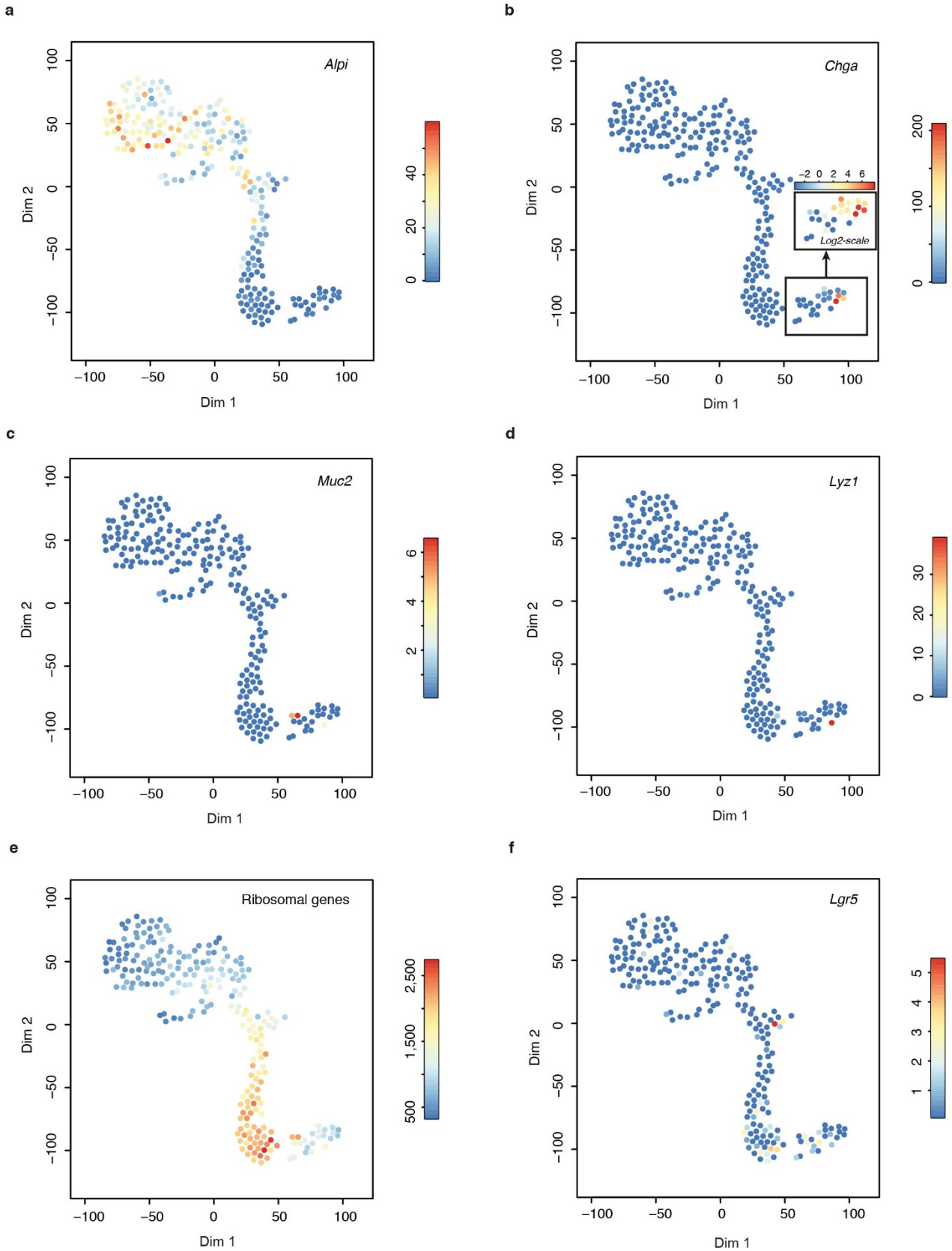
28. Yin, X. *et al.* Niche-independent high-purity cultures of Lgr5$^+$ intestinal stem cells and their progeny. *Nature Methods* **11,** 106–112 (2014).
29. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
30. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* **41,** D64–D69 (2013).
31. The External RNA Controls Consortium. a progress report. *Nature Methods* **2,** 731–734 (2005).



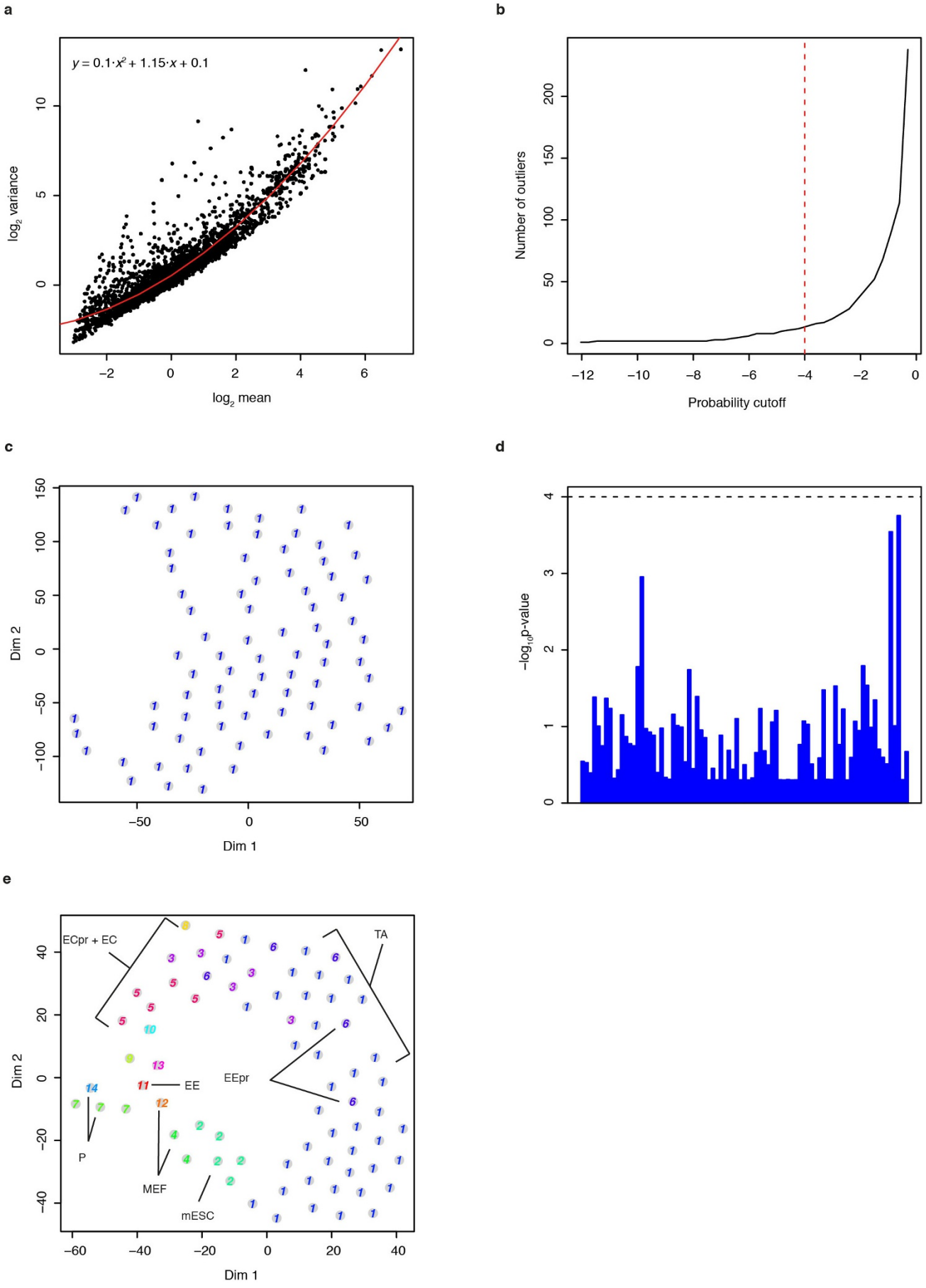
Extended Data Figure 1 | Sequencing statistics. **a**, Histogram of the number of sequenced transcripts per cell. The median (red line) is 8,559. The 288 cells were sequenced on two lanes to a total depth of 106,950,038 reads. Of those, 32,694,069 (31%) were mapped with a valid cell-specific barcode. **b**, Histogram of the total number of reads per cell divided by the total number of sequenced transcripts as counted with unique molecular identifiers. The average level of oversequencing across all genes is 6.9-fold (red line).



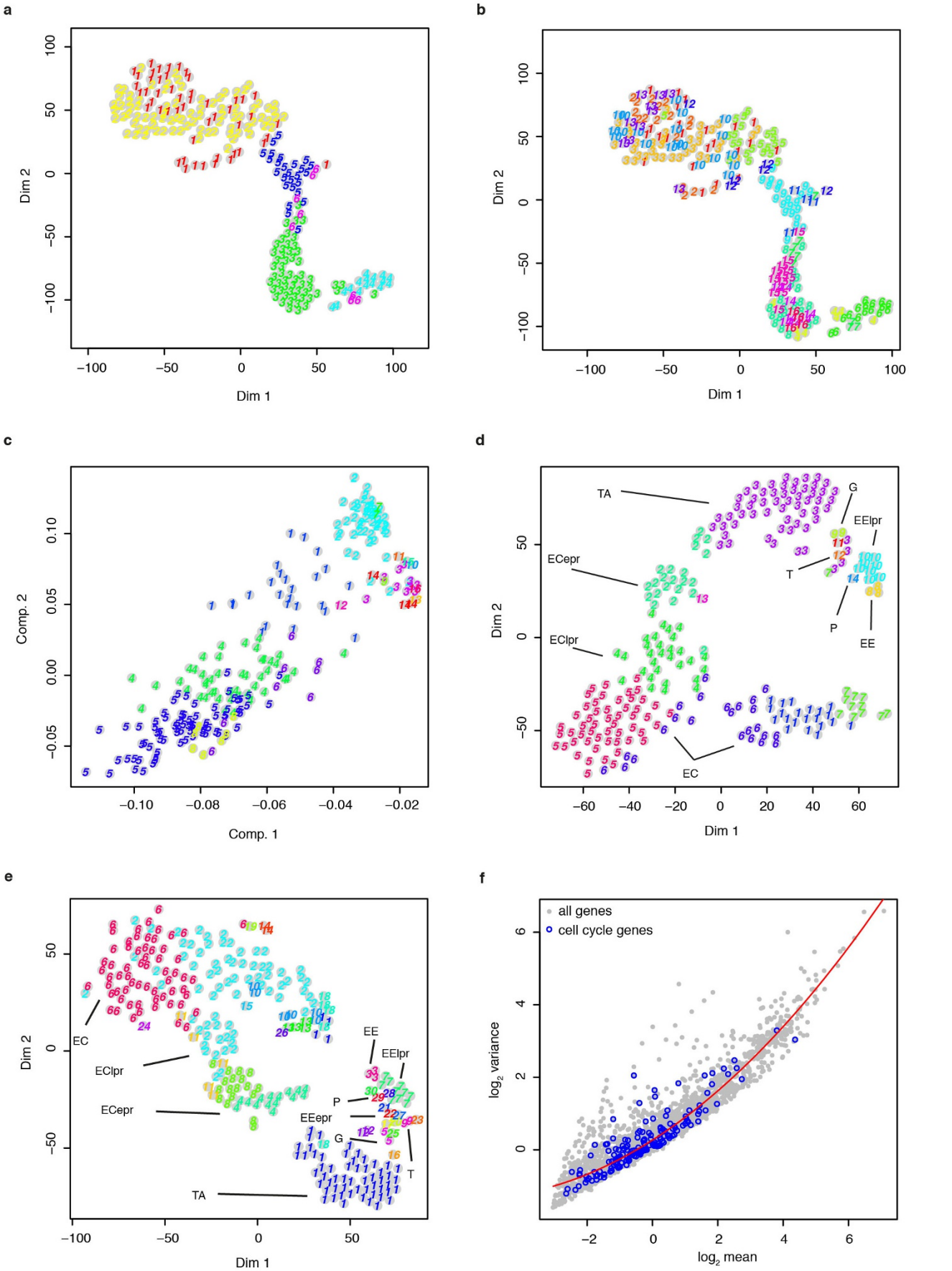
Extended Data Figure 2 | Clustering reveals major transcriptome differences between intestinal cells. **a**, Dendrogram obtained by hierarchical clustering of the transcriptome correlation matrix of 238 intestinal cells that survived all filtering steps using Euclidean distance metric. At least three distinct groups of cells can be recognized. **b**, Gap statistic of *k*-means clustering of the correlation matrix as a function of the cluster number. The gap statistic reflects the difference of the average within cluster distance of points in uniformly distributed data and the actual data. The first local maximum provides a good estimate for the number of clusters that achieves optimal separation of the data into clusters[15]. Data points and error bars represent mean and standard deviation across 50 bootstrap samples. For the intestinal cells a number of six clusters was predicted on the basis of the gap statistic.



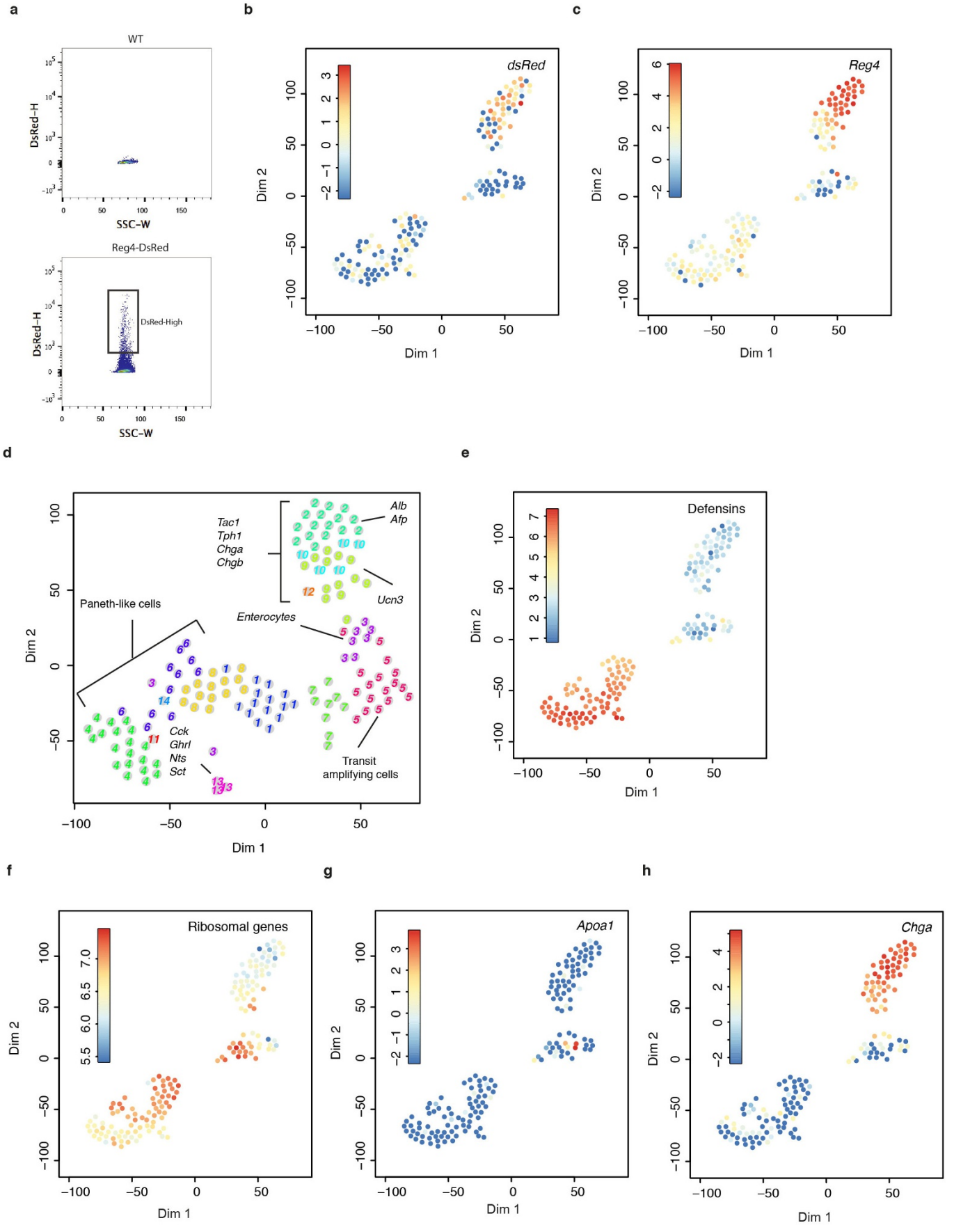
Extended Data Figure 3 | Marker gene expression reveals intestinal cell types. **a**–**f**, In the t-SNE maps the transcript count of known marker genes is colour-coded. Shown are maps for the enterocyte specific gene *Alpi* (**a**), the enteroendocrine marker *Chga* (**b**), the goblet cell marker *Muc2* (**c**), the Paneth cell marker *Lyz1* (**d**), for transcript counts aggregated across all ribosomal genes (**e**) and for the stem cell marker *Lgr5* (**f**). The latter two are upregulated in cells of cluster 2, for which no other specific markers could be identified. These cells most likely correspond to transit-amplifying cells. The inset shown in **b** depicts transcript counts of *Chga* on a logarithmic scale, since the dynamic range of this gene was very large. Dim, dimension.



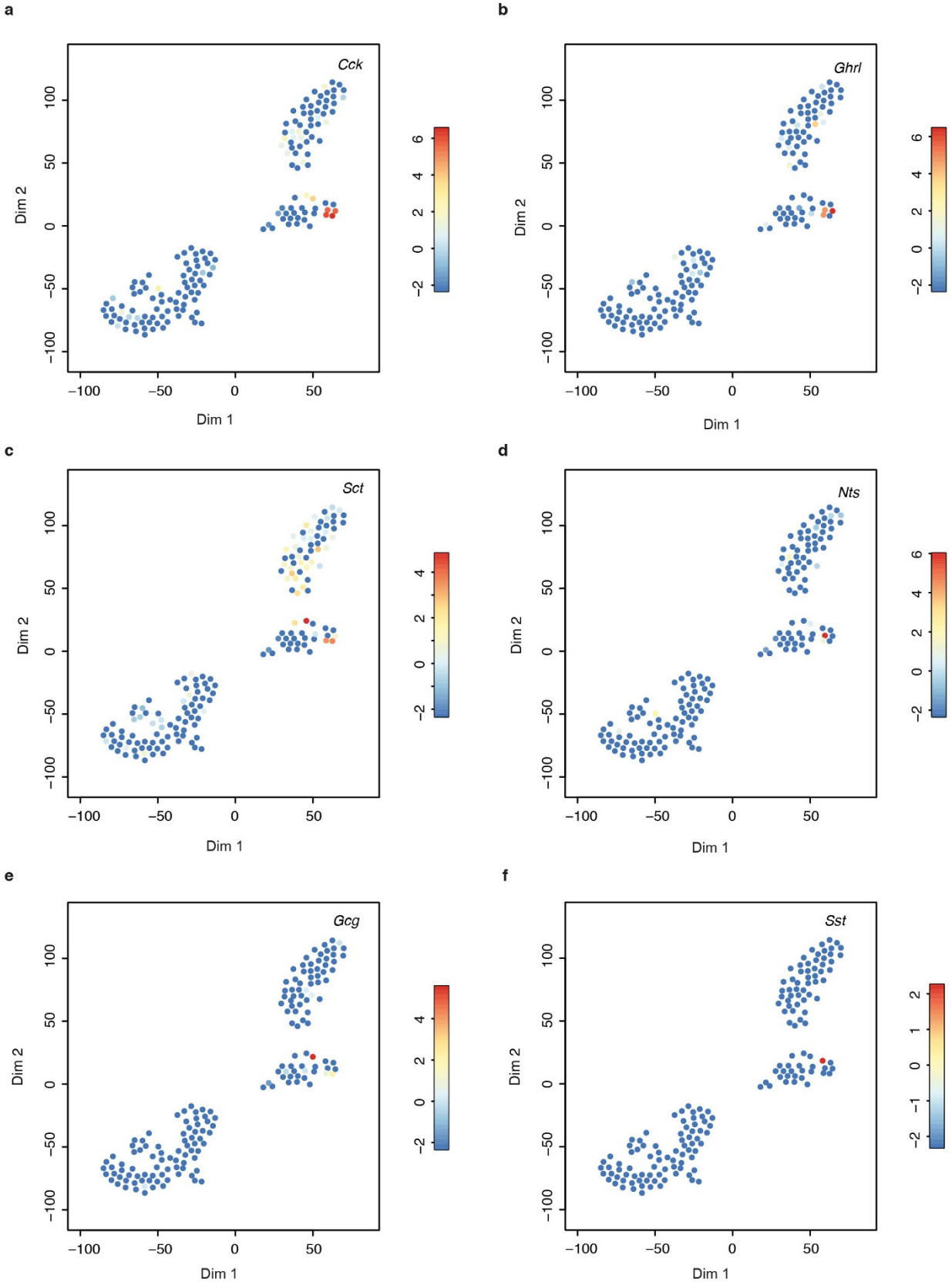
Extended Data Figure 4 | Identification of rare cell types with RaceID. **a**, Variance of transcript count across the entire ensemble of sequenced intestinal cells as a function of mean expression. The red line represents a second order polynomial (upper left corner) that was fitted to the data. Assuming that a large number of genes follows a similar transcript count distribution across different cell types this function can be used to estimate the parameters for a negative binomial that represents a background model for the expected transcript count variability at a given mean expression. The probability of observing a given transcript count in a particular cell of a cluster can be computed using this distribution with the average expression within this cluster as input. If the expression of at least two genes has a probability $<10^{-4}$ after multiple testing correction a cell is considered an outlier. **b**, Number of outlier cells as a function of the probability threshold. The threshold used in this study (10^{-4}) is indicated (red broken line). **c**, **d**, RaceID of pool-and-split controls reflects a low false-positive rate (see Supplementary Note). **c**, t-SNE map of 93 pool-and-split controls. RaceID identifies only a single large cluster with no outliers. **d**, Outlier probability for all pool-and-split controls. The p-value of all cells is higher than the default threshold for outlier identification (10^{-4}). **e**, RaceID on a defined mixture of cells demonstrates high specificity (see Supplementary Note). RaceID clusters for a mixture of cells comprising 75 random organoid cells, 5 mouse embryonic stem cells (mESCs) and 5 mouse embryonic fibroblasts (MEFs). Two out of five MEFs did not pass the filtering criteria (>3,000 transcripts per cell). EC, enterocytes; ECpr, enterocyte precursors; TA, transit amplifying cells; EE, enteroendocrine cells; EEpr, enteroendocrine precursors; P, Paneth cells. Dim, dimension.



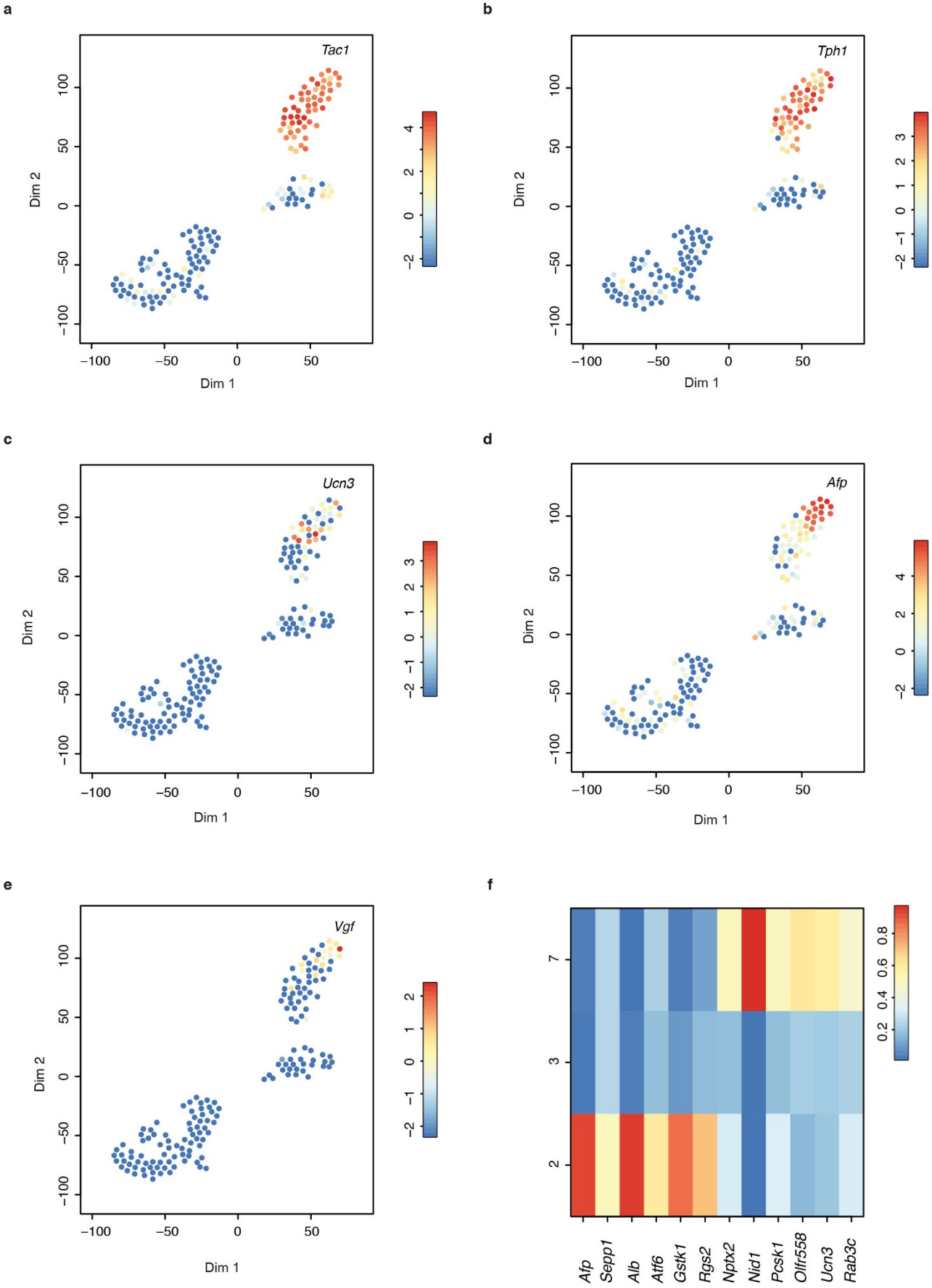
Extended Data Figure 5 | Benchmarking RaceID. **a**, **b**, To benchmark the RaceID algorithm we compared to a previously published method developed to distinguish cell types from the spleen[5] in thousands of sequenced single cells. We implemented the method according to the specification provided in the original paper. A shortcoming of the method is that it has to be initialized with an expected number K of cell types. Running the algorithm with $K = 6$ (**a**) yields results very similar to k-means clustering with $K = 6$ (Fig. 1c). However, when the algorithm is run with a larger number of cell types, e. g. $K = 20$ (**b**), rare cells of the secretory lineage can still not be separated while clusters corresponding to relatively uniform cell types fall apart. We conclude that this algorithm performs well for more abundant cell types but cannot identify rare cell types. **c**, Principal component analysis (PCA) of the transcriptome similarities. The cell types identified by RaceID are highlighted. The first two principal components can only classify major groups of enterocytes, transit amplifying cells, and secretory cells. **d**, To demonstrate that our method is not affected by technical noise due to varying detection efficiency across individual cells, we downsampled the transcriptome of all cells with >3,000 transcripts to the same size, given by the minimum total transcript counts across all cells that passed the filtering, and repeated the outlier identification. The t-SNE map shows all cell types identified by this strategy and the results are highly consistent with the normalization-based approach. **e**, The t-SNE map shows the results of RaceID run with relaxed filtering constraints (>1,000 transcripts per cell and only genes with more than three transcripts in at least one cell) as used for the *Reg4*-positive organoid cells. All the rare secretory cell types identified with the original settings were recovered. The different stages of enterocyte differentiation are also apparent. EC, enterocytes; EClpr, late enterocyte precursors; ECepr, early enterocyte precursors; TA, transit amplifying cells; G, goblet cells; EE, enteroendocrine cells; EElpr, late enteroendocrine precursors; EEepr, early enteroendocrine precursors; P, Paneth cells; T, tuft cells. **f**, Same as Extended Data Fig. 4a, but cell cycle related genes are highlighted as blue circles. This set of genes comprises all genes containing "cell cycle" within their associated "biological process" Gene Ontology terms. Cell cycle related genes do not show increased variability and are thus unlikely to lead to false positives in the outlier detection by RaceID. Dim, dimension.



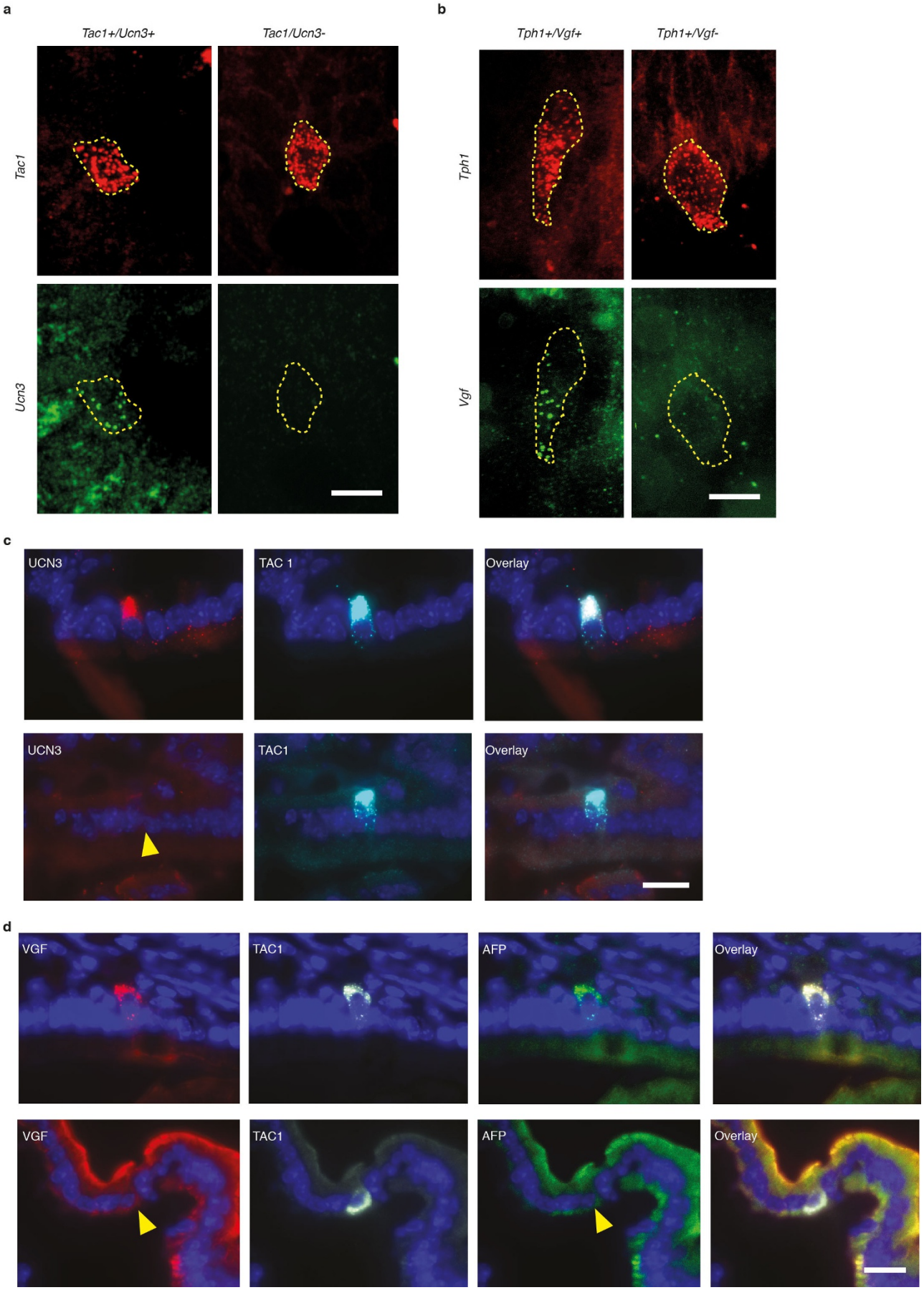
Extended Data Figure 6 | Purification of *Reg4*-positive cells from reporter mouse organoids. In total, 288 cells were sequenced. Ninety-six and 192 cells were analysed from two separate sorting experiments. **a**, Single small intestine cells derived from the wild-type (WT, upper panel) and Reg4–dsRed (lower panel) mice were sorted by FACS. Live cells were gated as DsRed-positive (lower panel, gate denoted by black rectangle, DsRed-High). SSC-W, side scatter width, DsRed-H, DsRed height. A median number of 2,813 transcripts per cell were quantified. After filtering out cells with $<$1,000 transcripts and genes with $<$3 transcripts in all cells or $>$2,000 transcripts in a single cells, 161 cells remained for analysis. **b**, **c**, In the t-SNE maps of *Reg4*-positive cells the transcript count of dsRed, driven by a *Reg4* promoter (**b**) and endogenous *Reg4* (**c**) are colour-coded on a logarithmic ($\log_2$) scale. Assuming a previously estimated sensitivity of our sequencing protocol[14], we measure $\sim$10% of all expressed transcripts. Reporter expression is about eightfold reduced in comparison to endogenous *Reg4*, but expression of both the reporter gene and *Reg4* is strongest in the *Tac1*/*Tph1* expressing enteroendocrine cells, while expression in Paneth-like cells is reduced. Noticeably, expression of *Reg4* and the reporter gene is also reduced in the *Cck*-positive enteroendocrine cells, similar to *Chga*. **d**, The t-SNE map shows the results of RaceID with more stringent filtering constraints ($>$3,000 transcripts per cell and only genes with a minimum of five transcripts in at least one cell) as used for the random organoid cells. Although the total number of cells is reduced to 135, most subtypes and rare cells identified with the relaxed settings are still observed, including the *Afp* and *Alb* expressing sub-types, the *Ucn3*-positive cells, the *Cck*-positive cells, the contamination by enterocytes and transit amplifying cells as well as the different Paneth cell states. **e**–**h**, Marker gene expression reveals intestinal cell types among *Reg4*-positive cells. In the t-SNE maps of *Reg4*-positive cells the transcript count of known marker genes is colour-coded on a logarithmic ($\log_2$) scale. Shown are maps, for transcript counts aggregated across all defensin genes which are highly expressed in Paneth cells (**e**), for transcript counts aggregated across all ribosomal genes (**f**), for the enterocyte marker *Apoa1* (**g**) and for the enteroendocrine markers *Chga* (**h**). Dim, dimension.



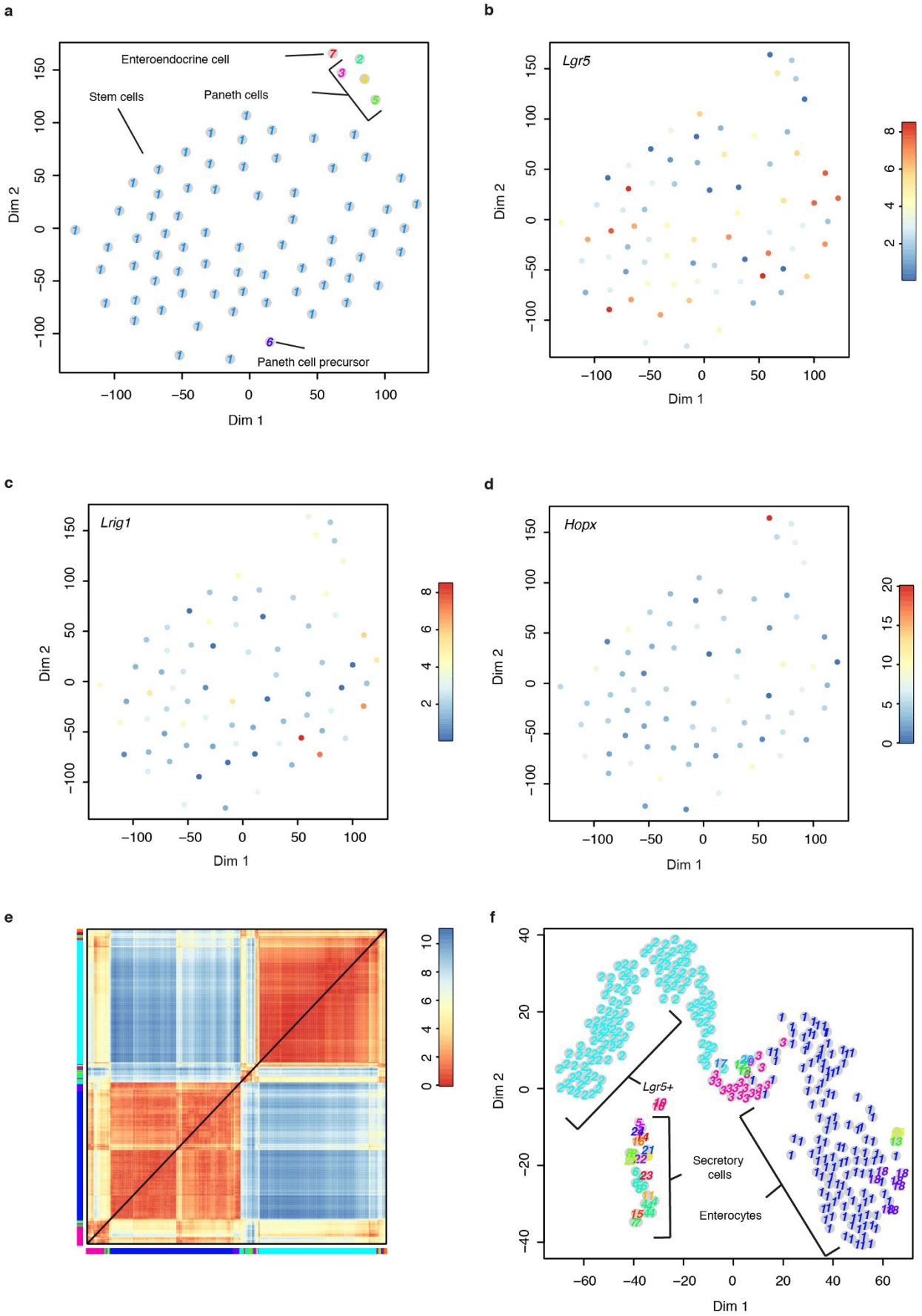
Extended Data Figure 7 | Expression of intestinal hormones in *Reg4*-positive cells. In the t-SNE maps of *Reg4*-positive cells the transcript count of hormone encoding genes is colour-coded on a logarithmic ($\log_2$) scale. **a**–**f**, Shown are maps for cholecystokinin (*Cck*) (**a**), ghrelin (*Ghrl*) (**b**), secretein (*Sct*) (**c**), neurotensin (*Nts*) (**d**), proglucagon (*Gcg*) (**e**), and somatostatin (*Sst*) (**f**). Dim, dimension.



Extended Data Figure 8 | Heterogeneity of substance P producing enteroendocrine cells. In the t-SNE maps of *Reg4*-positive cells the transcript count of genes specifically expressed in subtypes of enteroendocrine cells is colour-coded on a logarithmic (log_2) scale. Shown are maps for tachykinin (*Tac1*), which encodes substance P, (**a**), tryptophan hydroxylase 1 (*Tph1*) (**b**), urocortin 3 (*Ucn3*) (**c**), alpha-fetoprotein (*Afp*) (**d**), and VGF nerve growth factor inducible (*Vgf*) (**e**). **f**, The heat map shows the average expression of inferred marker genes for the enterochromaffin subtypes (cluster 2, 3 and 7). To show all genes on the same scale, the sum of average expression levels in each of the three clusters was normalized to one. Expression levels are shown for six cluster 2 markers and for six cluster 7 markers. Cluster 3 is distinct by the downregulation of both sets. Cluster 5 (not shown here) does not have specific markers and differs from the other clusters by lower expression of mature enterochromaffin markers (*Chga*, *Chgb*, *Tac1*, *Tph1*). This cluster likely comprises non-mature enterochromaffin cells. Dim, dimension.



Extended Data Figure 9 | Single-molecule FISH and immunofluorescence experiments confirm expression of markers for enteroendocrine cell subpopulations in the mouse small intestine. **a**, **b**, Small intestine cryosections were hybridized with smFISH probe libraries. Scale bar, 10 μm. **a**, The novel marker *Ucn3* was found in a small number of *Tac1*-positive cells. Probes against *Tac1*, conjugated with TMR (upper panel, red), and against *Ucn3*, conjugated with Cy5 (lower panel, green), were hybridized to the tissue sections. Dashed lines indicate cell borders. A cell co-expressing the two markers (*Tac1*$^+$/*Ucn3*$^+$) is shown in the left column. A cell expressing *Tac1*, but not *Ucn3* (*Tac1*$^+$/*Ucn3*$^-$), is shown in the right column. **b**, The novel marker *Vgf* is expressed by a subpopulation of *Tph1*-positive cells. Probes against *Tph1*, conjugated to TMR (upper panel, red), and against *Vgf*, conjugated to *Cy5* (lower panel, green), were used for hybridization. Cell borders were stained by phalloidin-AlexaFluor 488 (not shown). Dashed lines demarcate cell borders. A *Tph1*-positive cell, expressing *Vgf* (*Tph1*$^+$/*Vgf*$^+$) is shown in the left column. An example of a *Tph1*-positive cell, expressing no *Vgf* (*Tph*$^+$/*Vgf*$^-$) is shown in the right column. **c**, **d**, Immunostaining was performed on cryosections of mouse small intestinal tissue. Scale bar, 20 μm. **c**, Expression of UCN3 was observed in a few TAC1-positive cells within the jejunum. Frozen tissue sections were indirectly stained with anti-UCN3 (left panel, red), and anti-TAC1 (middle panel, light blue) antibodies. Nuclei were visualized with DAPI (dark blue). A cell, expressing both markers (TAC1$^+$/UCN3$^+$), is shown in the upper row. A cell, positive for TAC1, but not UCN3 (TAC1$^+$/UCN3$^-$), is shown in the lower row. The arrowhead points at the UCN3-negative cell. **d**, VGF is expressed by a subpopulation of TAC1-positive jejunal and ileal cells. VGF (left panel, red) and TAC1 (second panel from the left, grey) expression was visualized with indirect immunostaining. Expression of AFP was detected using a directly conjugated antibody (second panel from the right, green). Nuclei were counterstained with DAPI (blue). A TAC1-positive cell, expressing VGF and AFP (TAC1$^+$/VGF$^+$/AFP$^+$) is shown in the upper panel. An example of a TAC1-positive cell, expressing no VGF or AFP (TAC1$^+$/VGF$^-$/AFP$^-$) is shown in the lower panel. Arrowheads point at the VGF- and AFP-negative cell.



Extended Data Figure 10 | Purification of Lgr5–GFP^{+} cells from reporter mouse[23] organoids. Single small intestinal organoid cells, derived from *Lgr5*–GFP mice were sorted by FACS. In total, 96 cells were sequenced from a single experiment on four lanes with 31,065,854 reads in total of which 33% could be mapped to the transcriptome. Every UMI-derived transcript was sequenced on average 6.4 times. A median number of 9,626 transcripts per cell were quantified. After filtering out cells with $<$3,000 transcripts and genes with $<$5 transcripts in all cells or $>$500 transcripts in a single cell, 74 cells remained for analysis. **a**, The t-SNE map shows the cell types identified by RaceID. Only a single predominant cell type and few outliers were observed. Cluster 1 comprises intestinal stem cells while the few outliers represent Paneth and enteroendocrine cells. **b**–**d**, The t-SNE maps show expression of the stem cell marker *Lgr5* (**b**), the stem cell marker *Lrig1* (**c**), and the +4 niche marker *Hopx* (**d**). All markers are homogenously expressed across all cells at low transcript counts. We only observed marginal expression of the stem cell marker *Bmi1* in few cells and we did not observe expression of *Tert* in any of the cells, which is likely owing to the limited sensitivity of the method. The RaceID results indicate that *Lgr5*-positive intestinal stem cells represent a uniform population of cells. **e**, **f**, Combined analysis of random organoid and *Lgr5*-positive cells using RaceID. The initial clusters of the random organoid cells are conserved. The *Lgr5*-positive cells give rise to a uniform group that merges with the CBC/TA cluster from the random organoid cells (cluster 2). Shown is a heat map representation (**e**) and a t-SNE map (**f**) of the cell-to-cell transcriptome distances. Clusters are indicated by the same colours along the axes of the heat map (**e**) and for individual data points in the t-SNE map (**f**). Dim, dimension.