nature methods

ARTICLES

https://doi.org/10.1038/s41592-018-0009-z

Mapping the physical network of cellular interactions

Jean-Charles Boisset[1,2], Judith Vivié[1,2], Dominic Grün[1,2,3], Mauro J. Muraro[1,2], Anna Lyubimova[1,2] and Alexander van Oudenaarden[1,2*]

A cell's function is influenced by the environment, or niche, in which it resides. Studies of niches usually require assumptions about the cell types present, which impedes the discovery of new cell types or interactions. Here we describe ProximID, an approach for building a cellular network based on physical cell interaction and single-cell mRNA sequencing, and show that it can be used to discover new preferential cellular interactions without prior knowledge of component cell types. ProximID found specific interactions between megakaryocytes and mature neutrophils and between plasma cells and myeloblasts and/or promyelocytes (precursors of neutrophils) in mouse bone marrow, and it identified a Tac1$^+$ enteroendocrine cell–Lgr5$^+$ stem cell interaction in small intestine crypts. This strategy can be used to discover new niches or preferential interactions in a variety of organs.

Networks are studied in many fields and yield powerful information, as network topology and features can help scientists predict the behavior of a system[1–3]. The numerous cell types present in multicellular organisms interact and cooperate during embryonic development and adult life, forming a dynamic cellular network. Cellular functions are often lost or perturbed when neighboring cells are absent or malfunctioning[4,5]. Examples of cellular interactions that lead to specific functions include plasma cell–eosinophil[6] and erythroblast–macrophage (erythroblastic island)[7] interactions, as well as the hematopoietic stem cell niche in bone marrow (BM)[8]. Lgr5$^+$ stem cells in intestinal crypts also interact directly with Paneth cells to maintain stemness[9], and tumor cells interact with the local microenvironment to promote growth and survival[10]. These interactions are important but are typically studied one at a time, as existing tools generally do not allow this type of analysis to be carried out systematically.

The field of spatial transcriptomics has evolved rapidly in recent years owing to major improvements to in situ methods[11–15], transcriptomics of tissue sections[16], single-cell transcriptome mapping of in situ references[17,18], and spatially arrayed barcoding methods[19]. However, precise and systematic determination of the relative spatial positions of cells and their transcriptomes has not been achieved. We reasoned that we could create a cellular interaction network by microdissecting small interacting cell structures (e.g., cell doublets or triplets) and inferring the cell types present in the dissected entity via single-cell mRNA sequencing (scRNA-seq). Indeed, scRNA-seq has been used to discover new cell types, refine lineages, and study cell-to-cell heterogeneity in different organs[20–24]. In the present study, we used mouse BM as a model, as it contains the full hierarchy of hematopoietic cells, but their relative spatial cellular distribution is poorly known. Creating a network of physical interactions in mouse BM allowed us to find two new preferential interactions: promyelocyte/myeloblast–plasma cell interactions and megakaryocyte–mature neutrophil interactions. Next, we built a network of cell interactions in small intestinal crypts using a modified approach that does not require extensive microdissection. This allowed us to find a new preferential interaction between Tac1$^+$ enteroendocrine cells and Lgr5$^+$ stem cells. We call this strategy ProximID, and we expect that it will be useful for identifying new niche interactions in different organs.

Results

Cell types in small BM interacting structures. To create a cellular interaction network, we manually dissected 727 small interacting BM structures for a total of 1,728 cells across 18 independent experiments (Methods). We inferred the cell types present in the microdissected units by scRNA-seq (Fig. 1a and Supplementary Video 1). We removed cells with a low total number of transcripts from the dataset and applied the RaceID2 algorithm[20,25] to the remaining 911 cells, which returned 45 potential cell types/states (Fig. 1b,c). Different experimental replicates did not cluster together, which suggests that batch effects are negligible in this approach (Supplementary Fig. 1a). On the basis of the genes identified by RaceID2 as significantly upregulated in each cluster ($P < 0.01$), we were able to distinguish progenitor cells that express mainly ribosomal protein genes but no differentiation marker genes (Supplementary Fig. 2a); neutrophils (expressing *S100a8* and *S100a9*; Supplementary Fig. 2b); erythroid cells (*Beta-s* (*Hbb-bs*); Supplementary Fig. 2c); eosinophils (*Ear1* and *Ear2*; Supplementary Fig. 2d); megakaryocytes (*Ppbp*; Supplementary Fig. 2e); plasma cells (*Igj* (*Jchain*); Supplementary Fig. 2f); macrophages (*Hmox1*, *C1qa*, and *C1qb*; Supplementary Fig. 2g,h), and mast cells/basophils (*Mcpt8*; Supplementary Fig. 2i).

Next, we analyzed differentiation states present within the neutrophil lineage. To become fully mature, neutrophils sequentially build up granules[26–28]: *Elane* is expressed in azurophil granules at the myeloblast and promyelocyte stages (Supplementary Fig. 3a), *Ltf* (lactoferrin) and *Ngp* are expressed in specific granules in myelocytes (Supplementary Fig. 3b,c), and *Fpr1* and *Slc11a1* (Nramp1) are expressed in gelatinase and secretory granules in band cells and mature neutrophils (Supplementary Fig. 3d,e). We found that *Retnlg* expression best defined these later states (Supplementary Fig. 3f). Therefore, as a whole, the data acquired from microdissected

[1]Oncode Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, the Netherlands. [2]University Medical Center Utrecht, Cancer Genomics Netherlands, Utrecht, the Netherlands. [3]Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany.
*e-mail: a.vanoudenaarden@hubrecht.eu

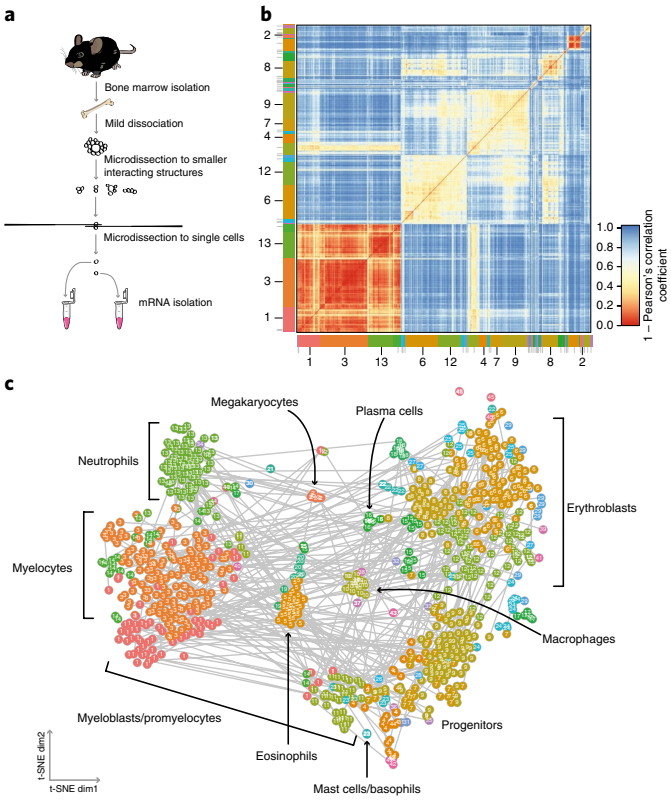

Fig. 1 | Unbiased dissection of small interacting structures and scRNA-seq resolve key cell types in the BM. a, Experimental scheme. BM is mildly dissociated and interacting structures are subdissected into smaller units, which are single-cell dissected and placed in different tubes for further analysis. **b**, Transcriptome similarities (1 – Pearson's correlation coefficient) between all 911 cell pairs in the final analysis ($n = 18$ animals). RaceID2 clustering is indicated along the axes by random colors. **c**, *t*-distributed stochastic neighbor embedding (t-SNE) map of transcriptome similarities. Colors represent RaceID2 clusters; gray edges connecting cells represent the actual physical interaction between cells. The putative cell types for the different clusters are indicated.

small interacting structures recapitulate the main cell types present in vivo in the BM. To verify that our dissection method does not introduce cell type biases, we checked the clusters represented within the different units of BM (≥4 cells) used to further subdissect doublets, triplets, and so on (Supplementary Fig. 1b,c). We did not observe any clonal units, which indicated high cell-type mixing within small BM regions.

To determine whether the dissected interacting structures reflected the normal composition of the BM, we sequenced 768 BM cells (after filtering) randomly sorted by flow cytometry. We merged both datasets and clustered cells on the basis of their transcriptomes using RaceID2 (Supplementary Fig. 4a,b). When we compared the frequency of cells per RaceID2 clusters to label permutation simulations, we found that the sorted cell population contained fewer erythroblasts and eosinophils, whereas handpicked cells contained fewer neutrophils, progenitors, and putative macrophage progenitors (expressing *S100a4*) (Supplementary Fig. 4c). Despite these few differences in frequency, the cell types present in the two datasets were similar.

Creation of a cellular interaction network. Next, we used physical connectivity to detect preferential or unfavorable interactions (Methods and Supplementary Software). We used label permutation to create a statistical distribution of chance interactions under the null hypothesis; this is an assumption-free procedure that keeps the original data structure. By comparing the actual number of interactions between cell type pairs to this distribution, we detected interactions that were significantly ($P < 0.05$) enriched or depleted compared with numbers in the background model (Fig. 2a). Notably, we observed enriched interactions between cells of the same type, and depleted interactions between cells of divergent lineages (for example, neutrophil versus erythroid lineages). However, we also observed enriched interactions among different cell types: macrophages and erythroblasts (clusters 10 and 12) (Fig. 2b), plasma cells and myeloblasts/promyelocytes (clusters 1 and 16) (Fig. 2c), and megakaryocytes and neutrophils (clusters 2 and 13) (Fig. 2d). The first interaction was already identified as the erythroblastic island, an important niche for the proper maturation of red blood cells[7,29]. We also observed expression of *Beta-s* in cells that clustered with macrophages, which supports previous evidence that erythroblast nuclei are phagocytized by central macrophages[29]. We achieved optimal detection of these three types of physical interactions (macrophage–erythroblast, plasma cell–myeloblast/promyelocyte, and megakaryocyte–neutrophil) when we used a threshold of 800–1,200 transcripts per cell (Supplementary Fig. 5c), and we note that the frequency of detected interactions exhibited a linear dependence on the number of picked structures (Supplementary Fig. 5d).

To test the reliability of ProximID, we also dissected interacting structures from the fetal liver (FL), the main hematopoietic organ during embryonic development. We carried out four independent experiments, acquiring 356 cells (after filtering) from FL from mice at embryonic day 13.5 (E13.5) to E15.5 (Supplementary Fig. 6a,b). RaceID2 identified hepatocytes (*Afp* and *Alb*), erythroblasts expressing adult (*Beta-s*; Supplementary Fig. 6c) or embryonic-type hemoglobin (*Hbb-y* and *Hba-x*), macrophages (*Hmox1*; Supplementary Fig. 6d), and putative progenitors. Compared with the random model, we observed enrichment for a specific interaction between macrophages and a subtype of adult erythroblasts (Supplementary Fig. 6e,f). This again corresponds to the erythroblastic island, and further confirms the robustness of the method for the discovery of cellular interactions without prior knowledge.

To assess whether interactions can potentially be created in vitro during the period of cell collection, we microdissected small interacting structures from an equal mixture of mildly dissociated C57BL/6 and CAST/EiJ BM (two genetically distinct strains of mice; $n = 2$ each). Although we observed that the majority of physical interactions took place between cells of the same strain, we also observed interstrain interactions, meaning that some interactions happened after dissociation (Supplementary Fig. 7a,b). In simulations where genotype labels for each cell were permutated, we observed that the frequency of interstrain interactions was significantly lower ($P = 0.0036$) than what would be expected if cellular interactions occurred only randomly after dissociation (Supplementary Fig. 7c). Nevertheless, the fact that we observed some interstrain interactions prompted us to verify the newly found interactions in situ.

In situ validation of new preferential interactions. To validate the previously unknown preferential interactions in situ, we carried out single-molecule fluorescence in situ hybridization (smFISH)[30] on BM sections. To test the interaction between plasma cells and myeloblasts/promyelocytes, we designed probes for *Igj* and *Elane*, respectively, and imaged and analyzed a large surface of the section (Fig. 3a, Supplementary Fig. 8a, and Supplementary Software). We calculated the distance from the center coordinates of plasma cells to the closest manually segmented myeloblast/promyelocyte pixel. Out of 43 plasma cells, 17 were located within interacting distance of myeloblasts/promyelocytes (Fig. 3b). The number of interactions differed significantly from that in a background

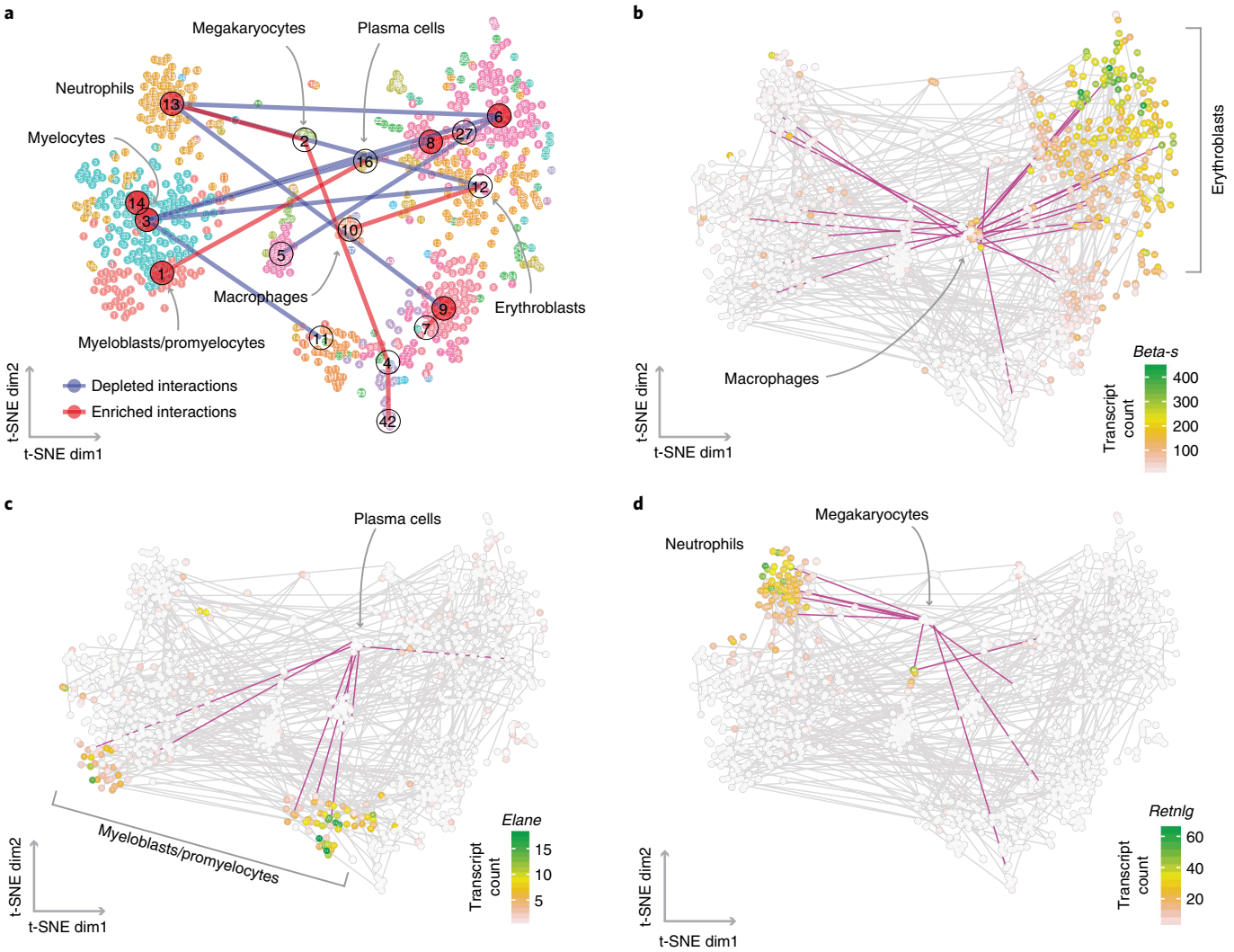

Fig. 2 | Identification of enriched and depleted interactions in the BM. a, t-SNE map of transcriptome similarities with enriched and depleted interactions. Nodes represent cluster centers, edges represent intercluster interactions, and solid nodes represent intracluster interactions. **b**–**d**, t-SNE map of transcriptome similarities, with color-coded representation of transcript counts for *Beta-s* (**b**), *Elane* (**c**), and *Retnlg* (**d**). Pink edges represent all interactions stemming from cluster 10 (macrophages; **b**), cluster 16 (plasma cells; **c**), and cluster 2 (megakaryocytes; **d**).

model where an equal number of cells were placed randomly on the sampling field ($P=0.0001$; Fig. 3c). Additionally, the mean distance of all plasma cells to the closest myeloblasts/promyelocytes was significantly lower than would be expected to occur by chance ($P<0.0001$; Fig. 3d). Therefore, plasma cells and myeloblasts/promyelocytes are neighbors and tend to physically interact in situ in the BM. Furthermore, we found that plasma cells express Slpi, an inhibitor of Elane, which indicates a probable function in protecting their environment from the deleterious effect of elastases (expressed by myeloblasts and promyelocytes) or in promoting neutrophil differentiation[31,32] (Fig. 3e).

To test the neutrophil–megakaryocyte interaction, we carried out smFISH using probes for *Retnlg* (neutrophils) and *Ngp* (myelocytes; negative control), as the network data (Fig. 2a) indicated that myelocytes do not specifically interact with megakaryocytes despite being the direct precursor of neutrophils. We added an antibody to CD41 to localize megakaryocytes (Fig. 4a and Supplementary Fig. 8b). We detected a total of 655 *Ngp*-positive cells and 603 *Retnlg*-positive cells, among which 39 and 67, respectively, were within interacting distance of megakaryocytes (Fig. 4b). Neutrophils interacted with megakaryocytes significantly more often than myelocytes did (Fisher's exact test; $P=0.00018$). To further exclude the possibility that arbitrary positioning of neutrophils in the BM could explain their interaction with megakaryocytes, we compared observed interactions to a background model of an equal number of cells placed randomly on the sampling field. The number of *Ngp*-positive cells within interacting distance of megakaryocytes did not differ from that in the background model ($P=0.3625$; Fig. 4c), but *Retnlg*-positive cells were within interacting distance significantly more often than would be expected to happen by chance ($P<0.0001$; Fig. 4d). We conclude that mature neutrophils and megakaryocytes favorably interact with each other in the BM.

An interaction network in intestinal crypts. To increase throughput, we designed an exploratory approach for generating a cellular interaction network without separating doublet or triplet structures, which we demonstrated on intestinal crypts (Fig. 5a and Methods). We first sequenced the transcriptome of 2,688 single cells sorted from intestinal crypts, and determined cell types with the RaceID3 algorithm[33] ($n=2$ independent experiments; 1,172 cells after filtering). We devised four different methods to infer cell types present in sequenced structures from mildly dissociated tissue ($n=3$

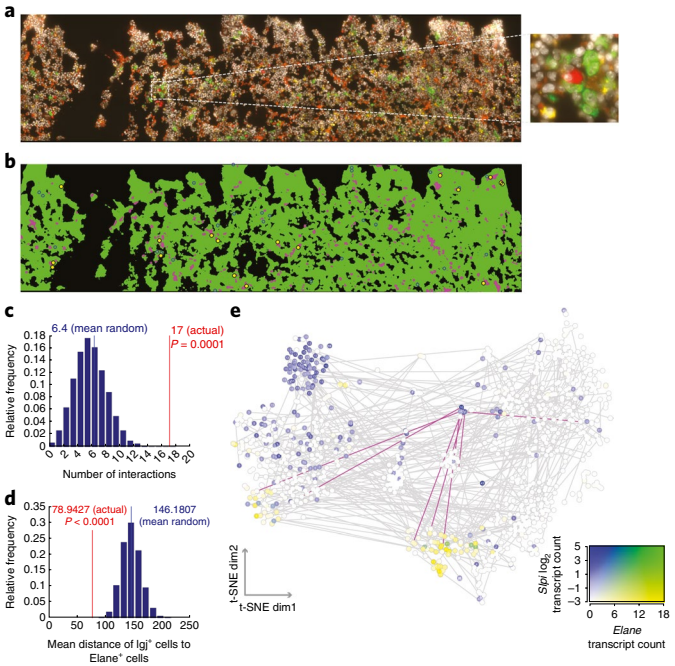

Fig. 3 | Plasma cells specifically interact with myeloblasts and promyelocytes. a, Tile scan of a BM cryosection stained with smFISH probes for *Elane* (green) and *Igj* (red) mRNAs, conjugated with Cy5 and TAMRA, respectively ($n=1$). Nuclei were stained with DAPI (gray). The magnified image (magnified five times relative to the primary image) on the right shows one plasma cell (red) interacting with myeloblasts/promyelocytes (green). Scale bar, 20 µm. **b**, Same region as in **a**, showing occupation by cells (green) and myeloblasts/promyelocytes (pink). Plasma cells (central coordinates shown for each cell) are indicated as noninteracting (blue circles) or interacting with myeloblasts/promyelocytes (yellow circles). **c**, Probability distribution of plasma cell interactions with myeloblasts/promyelocytes in 10,000 randomly positioned simulations (blue bars). The red bar represents the number of interactions observed in situ. **d**, Probability distribution of mean distances between plasma cells and the closest myeloblast/promyelocyte in the simulations from **c** (blue bars). The red bar represents the mean distance observed in situ. **e**, t-SNE map of transcriptome similarities, with a two-dimensional color-coded representation of transcript counts for *Slpi* and *Elane*. Magenta edges represent all interactions stemming from cluster 16 (plasma cells).

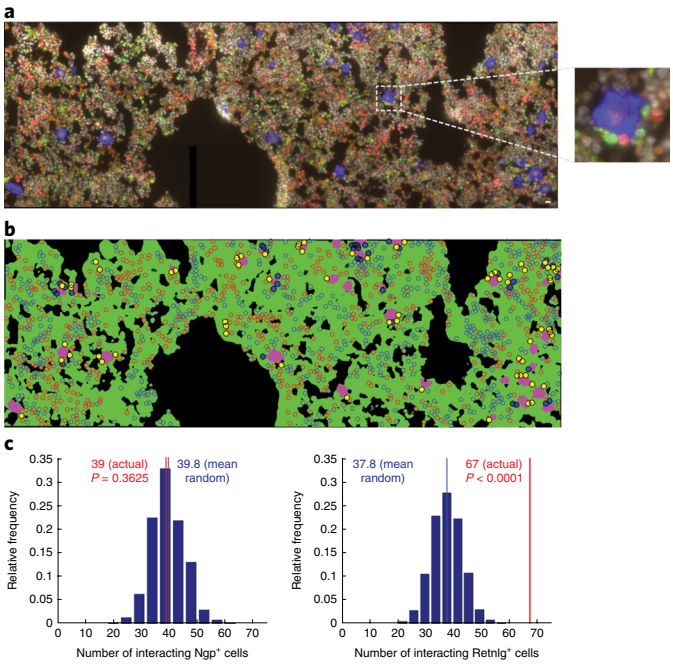

Fig. 4 | Megakaryocytes specifically interact with neutrophils. a, Tile scan of a BM cryosection stained with biotin-conjugated anti-CD41 and Alexa Fluor 488–streptavidin conjugates (blue), as well as smFISH probes for *Retnlg* (green) and *Ngp* (red) mRNAs conjugated with Cy5 and TAMRA, respectively ($n=1$). Nuclei were stained with DAPI (gray). The magnified image (magnified four times relative to the primary image) on the right shows neutrophils (green) and myelocytes (red) interacting with a megakaryocyte (blue). Scale bar, 20 µm. **b**, Same region as in **a**, showing occupation by cells (green) and megakaryocytes (pink). Neutrophil and myelocyte central coordinates are indicated as noninteracting (open blue and red circles, respectively) or interacting (solid yellow and blue circles, respectively) with megakaryocytes. **c**,**d**, Probability distribution of the number of myelocytes (**c**) and neutrophils (**d**) interacting with megakaryocytes in 10,000 randomly positioned simulations (blue bars). Red bars indicate the number of interactions observed in situ.

independent experiments), and compared their performance by using a test dataset (Supplementary Software).

We first trained a random forest classifier to recognize virtual structures of doublets or triplets, for all combinations of cell types, based on the single-cell transcriptomes. As a second approach, we trained 20 random forest classifiers (one per cell type) to recognize whether a particular cell type was present in a virtual structure. For each classifier, we determined a subjective probability threshold to limit false positive detection and called doublets of the same cell type for cases in which only a single classifier probability was above the threshold. Otherwise, doublets were assigned the cell types with the two highest probabilities. We applied a similar principle for the triplets. For the third strategy, we integrated probabilities from the 20 independent classifiers by training a second layer of random forests to recognize all possible combinations of probabilities in doublets or triplets. For the final strategy, we used a Monte Carlo approach in which we randomly chose two or three cells from the single-cell training set and used them to create virtual structures by summing their transcriptomes. We calculated the correlation of the virtual random structure to the test structure iteratively, with one cell in the random structure exchanged in each iteration, until the highest correlation was reached.

The four methods yielded diverse accuracies for cell doublets and triplets, from which we inferred interactions (Fig. 5b). We compared the pairwise interactions with label permutation simulations, as described for BM, to detect enriched and depleted interactions (Fig. 5c–f). In agreement with previously published results[9], we observed enriched interaction between Paneth cells and Lgr5$^+$ stem cells in the first method (Fig. 5c). Although detected interactions varied by method, some were identified with multiple approaches. For example, we repeatedly detected a Paneth cell–Nts$^+$ enteroendocrine cell interaction (Fig. 5c,e) and an Lgr5$^+$ stem cell–Tac1$^+$ enteroendocrine cell interaction (Fig. 5d,f). We verified the latter interaction by smFISH with probes for *Lgr5* and *Tac1*. Among ten *Tac1*-expressing cells found across 33 imaged crypts, three cells did not interact in 2 crypts, and seven cells were in direct contact with *Lgr5*-expressing cells at the bottom of 6 crypts (a representative example is shown in Fig. 5g), thus validating ProximID's prediction. We conclude that it is feasible to construct a network of cellular interaction without physically separating cells, using a single-cell transcriptome reference. Nevertheless, the detection of enriched and depleted interactions is dependent on cell-type identification and therefore requires the validation of prospective interactions in situ.

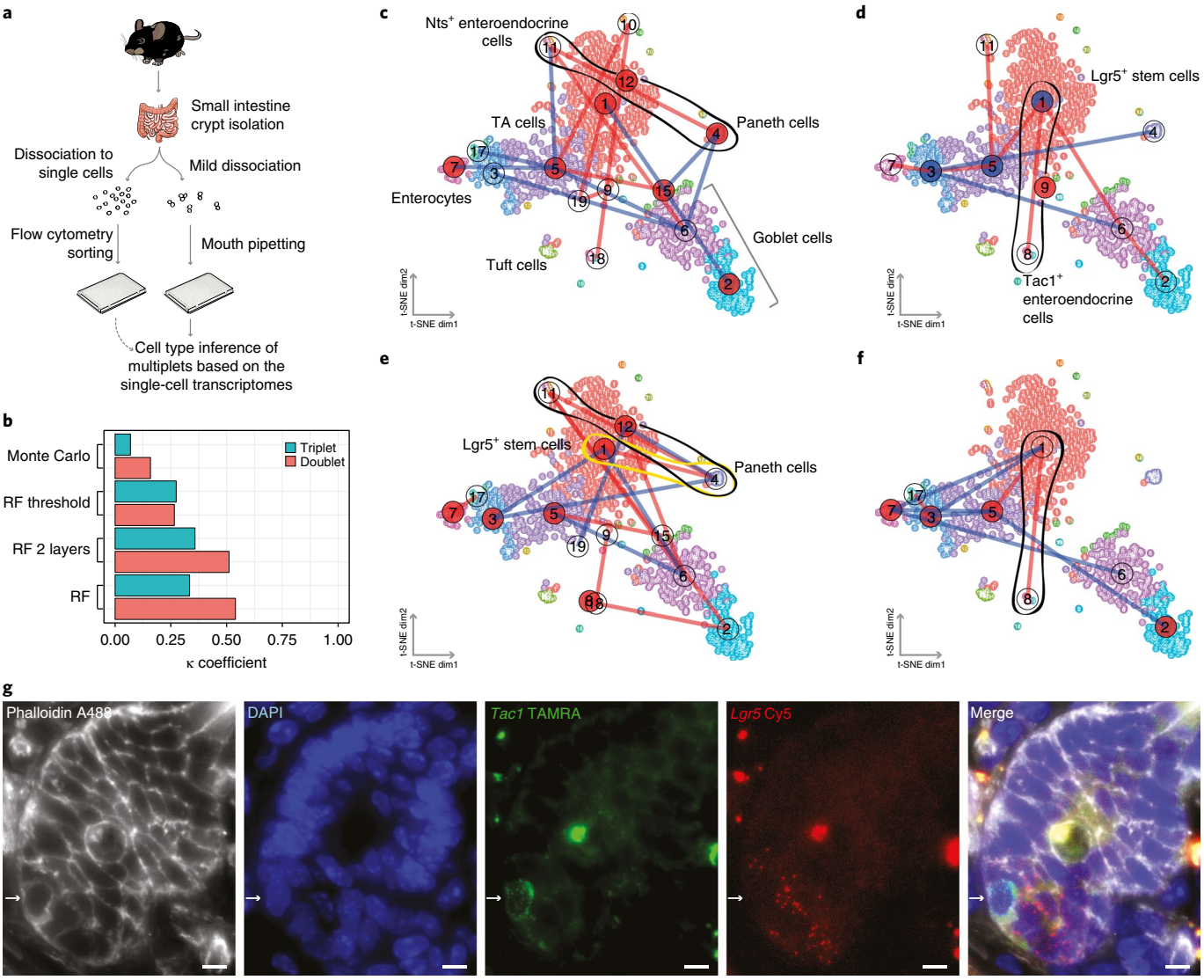

Fig. 5 | Identification of a Tac1$^+$ enteroendocrine cell–Lgr5$^+$ stem cell interaction in the intestinal crypt without microdissection. a, Experimental scheme. Small intestine crypts were either dissociated to single cells to create a transcriptome reference or mildly dissociated to collect doublets or triplets that were computationally inferred by training on the single-cell references. **b**, Overall performance, as measured by Cohen's κ coefficient for a test set, of four methods (based on Monte Carlo simulations or random forests (RF)) for inference of cell types present in doublets or triplets, based on a single-cell transcriptome reference ($n=1$ training). **c**–**f**, t-SNE maps of transcriptome similarities with enriched (red edges) and depleted (blue edges) interactions. Nodes represent cluster centers, solid nodes indicate intracluster interactions, and edges indicate intercluster interactions. The yellow line in **e** highlights the Paneth cell–Lgr5$^+$ cell interaction, and the black lines highlight two new interactions found with two methods of cell-type identification. Cell-type identification of the small structures was based on an RF trained on all possible combinations of doublets and triplets (**c**) or on the presence or absence of a particular cell type in the structure, with the decision based on thresholds for probabilities (**d**) or on a second layer of training by an RF (**e**). **f**, Cell-type identification based on Monte Carlo simulations. **g**, Epifluorescence images of an intestinal crypt stained with phalloidin–Alexa Fluor 488 (gray), DAPI (blue), and smFISH probes for *Tac1* (TAMRA; green) and *Lgr5* (Cy5; red), showing a representative example of a *Tac1*-expressing cell interacting with an *Lgr5*-expressing cell ($n=1$ animal). The arrow points to a *Tac1*-positive cell at the bottom of an intestinal crypt. The image is a maximum-intensity projection of nine *z*-stacks. Scale bars, 10 μm.

Discussion

ProximID builds a physical cellular interaction network to predict preferential associations between cells on the sole basis of the requirement of physical attachment. We discovered two new preferential interactions and confirmed them in situ in BM. We also captured, in both BM and FL, a previously identified functional interacting structure, the erythroblastic island[7,29], which further highlights the robustness of ProximID. We did not detect interactions such as eosinophil–plasma cell[6] or megakaryocyte–hematopoietic stem cell[34] interactions, probably because these do not require physical attachment or are infrequent. Concerning the megakaryocyte–neutrophil interaction, it is interesting to note that in some pathological conditions, neutrophils are often seen living inside megakaryocytes (emperipolesis), which further reinforces their close connection[35].

As a proof of principle, we also showed that it is possible to create a network of cellular interaction by sampling doublets or triplets as such, without physical separation of the cells. The cell types present in the small structures are inferred on the basis of a reference set of single-cell transcriptomes. This alternative increased throughput while diminishing the experimental load. However, cell type identification is not perfect, and we therefore recommend that novel

interactions always be verified in situ. The new finding of a Tac1$^+$ enteroendocrine cell–Lgr5$^+$ stem cell interaction could help scientists better understand the role of a subset of Tac1$^+$ enteroendocrine cells that remain in the crypt instead of migrating toward the villus[36].

Recent developments in spatial transcriptomics aim at probing the transcriptomes of single cells as well as their exact spatial position. Importantly, these methods do not discern whether cells are actually physically interacting or just in close proximity. Because we manually microdissected our BM samples and considered only interactions in which cells are physically attached to each other, we were able to reconstruct the physical interaction network. However, the nature and strength of the interactions is unknown and could influence the sampling of the cells involved in these interactions. Indeed, the frequency of certain cell types was biased when structures were handpicked compared with observations after random single-cell sorting. Also, in the intestinal crypts, critical adhesion molecules such as β-catenin and E-cadherin are differentially expressed, which could influence dissociation and sampling[37]. These parameters are mostly unknown, and thus performance is likely to change depending on factors such as the tissue of interest or dissociation conditions.

This study lays a foundation for further exploration of the function of the newly found interactions, which should help to improve understanding of the production of mature neutrophils, the retention of antibody-producing plasma cells in the BM, and the cross-talk between hormone production and stem cell self-renewal in the intestinal crypts. Niches are usually analyzed from a single point of view (for example, the influence of the stroma on stem cells). Here we highlight the possible mutual benefits of preferential cell association, as demonstrated by the promyelocyte/myeloblast–plasma cell interaction. Indeed, promyelocytes and myeloblasts are likely to be a niche for plasma cells in the BM, but plasma cells could also influence the differentiation process toward neutrophils.

The ProximID workflow is straightforward and scalable, as small structures do not necessarily need to be physically separated in order for interacting cells to be inferred, and it should be applicable to many tissues and organs. We hope that ProximID will be a potent tool for the discovery of new prospective niches, especially when cell types and relative spatial positions are unknown.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41592-018-0009-z.

Received: 19 April 2017; Accepted: 20 March 2018;
Published online: 21 May 2018

References

1. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
2. Albert, R., & Barabasi, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
3. Newman, M. E. J. *Networks: An Introduction* (Oxford Univ. Press, New York, NY, 2010).
4. Raaijmakers, M. H. et al. Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature* **464**, 852–857 (2010).
5. Scadden, D. T. Nice neighborhood: emerging concepts of the stem cell niche. *Cell* **157**, 41–50 (2014).
6. Chu, V. T. et al. Eosinophils are required for the maintenance of plasma cells in the bone marrow. *Nat. Immunol.* **12**, 151–159 (2011).
7. Chow, A. CD169$^+$ macrophages provide a niche promoting erythropoiesis under homeostasis and stress. *Nat. Med.* **19**, 429–436 (2013).
8. Mendelson, A., & Frenette, P. S. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. *Nat. Med.* **20**, 833–846 (2014).
9. Sato, T. et al. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469**, 415–418 (2011).
10. Becker, A. et al. Extracellular vesicles in cancer: cell-to-cell mediators of metastasis. *Cancer Cell* **30**, 836–848 (2016).
11. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
12. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
13. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
14. Lovatt, D. et al. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–196 (2014).
15. Shah, S., Lubeck, E., Zhou, W.., & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
16. Junker, J. P. et al. Genome-wide RNA tomography in the zebrafish embryo. *Cell* **159**, 662–675 (2014).
17. Achim, K. et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
18. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
19. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
20. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
21. 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).
22. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
23. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
24. Zeisel, A. et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
25. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
26. Borregaard, N. & Cowland, J. B. Granules of the human neutrophilic polymorphonuclear leukocyte. *Blood* **89**, 3503–3521 (1997).
27. Fouret, P. et al. Expression of the neutrophil elastase gene during human bone marrow cell differentiation. *J. Exp. Med.* **169**, 833–845 (1989).
28. Pham, C. T. et al. Neutrophil serine proteases: specific regulators of inflammation. *Nat. Rev. Immunol.* **6**, 541–550 (2006).
29. Chasis, J. A. & Mohandas, N. Erythroblastic islands: niches for erythropoiesis. *Blood* **112**, 470–478 (2008).
30. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
31. Janoff, A. Elastase in tissue injury. *Annu. Rev. Med.* **36**, 207–216 (1985).
32. Klimenkova, O. et al. A lack of secretory leukocyte protease inhibitor (SLPI) causes defects in granulocytic differentiation. *Blood* **123**, 1239–1249 (2014).
33. Herman, J. S., & Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
34. Bruns, I. et al. Megakaryocytes regulate hematopoietic stem cell quiescence through CXCL4 secretion. *Nat. Med.* **20**, 1315–1320 (2014).
35. Centurione, L. et al. Increased and pathologic emperipolesis of neutrophils within megakaryocytes associated with marrow fibrosis in GATA-1(low) mice. *Blood* **104**, 3573–3580 (2004).
36. Aiken, K. D., & Roth, K. A. Temporal differentiation and migration of substance P, serotonin, and secretin immunoreactive enteroendocrine cells in the mouse proximal small intestine. *Dev. Dyn.* **194**, 303–310 (1992).
37. Tan, C. W., Hirokawa, Y., Gardiner, B. S., Smith, D. W., & Burgess, A. W. Colon cryptogenesis: asymmetric budding. *PLoS One* **8**, e78519 (2013).
38. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst* **3**, 385–394 (2016).
39. Lyubimova, A. et al. Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.* **8**, 1743–1758 (2013).
40. Preibisch, S., Saalfeld, S., & Tomancak, P. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics* **25**, 1463–1465 (2009).

Acknowledgements

We thank J. Korving for help with the microdissection microscope; the animal facility; A. de Graaf and the microscope facility; the sequencing facility; K. Wiebrands for suggesting a name for the method and assistance in intestinal crypt dissociation; N. Battich and B. De Barbanson for help with machine learning; L. Kester for help with the Monte Carlo simulations; and R. van der Linden for help with the sorting experiments. This work was supported by the European Research Council (advanced grant ERC-AdG 742225-IntScOmics to A.v.O.) and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) OPEN award (NWO-ALWOP189 to A.v.O.). This work is part of the Oncode Institute, which is partly financed by the Dutch Cancer Society. In addition, we thank the Hubrecht Sorting Facility and the Utrecht Sequencing Facility, subsidized by the University Medical Center Utrecht, Hubrecht Institute, and Utrecht University.

Author contributions

J.-C.B. and A.v.O. conceived the project. J.-C.B. performed experiments. J.V. performed single-cell RNA preparation for sorted BM and FL cells. M.J.M. performed single-cell sequencing of intestinal crypt single cells. A.L. prepared intestinal crypt cells. J.-C.B. and D.G. analyzed the data. J.-C.B. wrote the manuscript. A.v.O. guided experiments and data analysis and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0009-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.v.O.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Animals. We used C57BL/6 and CAST/EiJ female or male mice aged 23 weeks to 9 months, bred in our facility. Experimental procedures were approved by the Dier Experimenten Commissie (DEC) of the KNAW and performed according to the guidelines.

Cell preparation. We isolated BM from femurs and tibias by flushing bones with Hank's Balanced Salt Solution (HBSS; Invitrogen) without calcium or magnesium, supplemented with 1% heat-inactivated FCS (Sigma). We then mildly dissociated the BM by pipetting it up and down a few times. Small interacting structures were selected by visual inspection under a dissection stereomicroscope (Leica) and transferred by mouth pipetting to a microscope (Zeiss) equipped with micromanipulators (Narishige). These structures could be combinations of cells such as doublets or triplets, or slightly bigger units composed of around 10–20 cells. In the case of small structures, the cells were manually pulled apart, without enzymatic dissociation, with the help of two pulled needles. For the bigger units, small structures were first sequentially trimmed off the unit with dissection needles, and then single-cell dissected as described previously. The same strategy was used for FL, with some variations: after dissection, FLs were cut into pieces, and borders were scraped with insulin needles to release small interacting structures. The microdissection was then performed in TrypLE (Thermo Fisher Scientific). The intestinal crypt single cells were obtained as described previously[20]. To obtain doublets or triplets, we checked the trypsin incubation every 5 min under a microscope and stopped it by washing the sample before complete dissociation of the crypts was achieved.

For BM and FLs, single cells were mouth-pipetted directly into Eppendorf tubes containing 100 µL of TRIzol (Life Technologies), 0.02 µL of 1:50,000 ERCC spike-in RNA (Ambion), and 0.2 µL of GlycoBlue (Ambion). Tubes were immediately frozen on dry ice. The pipette used for mouth pipetting was always washed between pipetting rounds with HBSS with 1% FCS.

For flow cytometry sorting, BM was dissociated to single cells, filtered on a 40-µm strainer, and lysed with IOTest 3 lysing solution (Beckman Coulter). Before sorting, DAPI was added to exclude dead cells. Live singlet cells were sorted on a FACSAria II directly into 100 µL of TRIzol, as done for the picked cells.

For the small intestine crypts, DAPI was added to exclude dead cells, and live singlet cells were sorted into 384-well plates as described previously[38]. Doublets or triplets were mouth-pipetted into 384-well plates containing 0.02 µL of 1:50,000 ERCC spike-in RNA (Ambion), 0.015 µL of 1% IGEPAL (Sigma), 0.035 µL of RNase inhibitor (Clonetech), and 0.020 µL of 10 mM dNTPs and water to a total volume of 100 nL, which was then covered with 5 µL of mineral oil (Sigma).

CEL-Seq library preparation, clustering, and differential gene expression. Total RNA was prepared according to the manufacturer's instructions, with overnight precipitation. CEL-Seq was performed directly on the dried pellet after the 70% ethanol wash as previously described[20]. Alternatively, 384-well plates were further processed as previously described[38]. The libraries were sequenced on an Illumina HiSeq 2500 and NextSeq 500. Mapping of the sequenced reads (mm10 transcriptome) and quantification of the number of transcripts were performed as previously described[20]. Cells containing fewer than 900 total transcripts were discarded as a compromise to remove low-quality cells but retain cell types with inherently low numbers of transcripts. Indeed, some cell types, including neutrophils (cluster 13), exhibit inherently low levels of transcripts (Supplementary Fig. 5a,b) and get discarded when higher thresholds are used. The remaining cells were downsampled to 900 transcripts for normalization. For small intestine crypt single cells, cells expressing high levels of *Kcnq1ot1*, *Malat1*, and *Rn45s* were discarded, as we often observed potential artifactual expression of these genes in low-quality sample cells. Then cells with fewer than 2,000 transcripts were discarded, and the remaining cells were normalized to the median transcript count of all cells. Clustering was done with the RaceID2 algorithm[25], or RaceID3[33] for intestinal crypt cells.

For attribution of strain identity for each cell isolated from mixed C57BL/6 and CAST/EiJ bone marrow, we first created a CAST/EiJ transcriptome reference, where all CAST/EiJ single-nucleotide polymorphisms were introduced into the C57BL/6 RefSeq transcriptome gene model based on mouse genome release mm10. Transcripts were mapped to both C57BL/6 and CAST/EiJ transcriptomes. Transcripts that mapped uniquely to only one of the transcriptomes were kept. We then determined strain identity by calculating the ratio of the total number of CAST/EiJ transcripts to that of C57BL/6 transcripts.

Cell-type inference for intact doublets and triplets. We used four different methods to infer the cell types present in small intestine crypt doublets and triplets. We distributed 70% of the single-cell transcriptomes for training, and the remaining 30% of the cells for testing. RaceID clusters composed of only one cell were resampled. For the first method, we selected the top ten variable genes between clusters, for each cluster, with the RaceID 'clustdiffgenes' function. We used the training cells to create doublets or triplets with all possible combinations of cell types, with 100 random cells sampled for each class of doublets and 10 cells sampled for each class of triplets. Transcript counts were normalized for all structures to the same number (here the lowest transcript count sum for all training structures). We used the randomForest package in R to compute the random forests based on all different classes (8,000 trees for doublets and 800 trees for triplets). The test dataset was prepared like the training set via the combination of all possible cell types, with 100 examples for each class. The accuracy of the predictions and the κ performance metric was computed with the 'confusionMatrix' function from the 'caret' R package.

For the second method, we used the training dataset to create 100 examples for each possible combination of doublets or triplets. The classes were defined on the basis of the presence or absence of a particular cell type in the virtual structure. To avoid class imbalance, we downsampled the number of structures not containing the cell type of interest to the number of structures containing the cell type of interest. Random forests were computed for each cell type independently (500 trees for each). We calculated the false discovery rate per different thresholds of probabilities given by the random forests when testing with the test dataset. We opted for doublets to contain the same cell type twice when only one of the cell type probabilities was higher than a threshold set at an arbitrary $1e^{-5}$ probability of false positive discovery; otherwise, the two cell types with the highest probabilities were chosen to compose the structure. The same principle was applied for triplets: if only one cell type probability passed the threshold, the triplet was composed of only that cell type. If two cell type probabilities passed, the triplet was composed of two cells of the cell type with the highest probability and a third cell with the second highest probability. Otherwise, the cell types with the three highest probabilities were chosen.

For the third method, instead of using a threshold based on the false positive discovery rate, we used a second layer of random forests to integrate the probabilities of the different cell-type-specific random forests. Because we needed two different test datasets, we randomly distributed the single cells into three parts. We used one for the training of the individual random forests, and one to test the models. This first test set was then used to train random forests (1,000 trees) to assign all possible classes of doublets or triplets on the basis of the probabilities given by the individual random forest models. The third set was used to test the two layers of random forests.

For the fourth method, we used a Monte Carlo simulation approach in which we chose a starting set of doublets or triplets by randomly picking from the training set two or three single-cell transcriptomes, respectively, which were summed. The Spearman correlation to the test structure was calculated, and we iteratively (10,000 times) replaced one of the cells from the training structure in order to reach a higher correlation. For optimization, we also introduced a temperature that was lowered stepwise every 1,000 iterations. This ensured that a virtual doublet or triplet, with a lower correlation to the test, could sometimes be chosen for the next starting pair.

Single-molecule FISH on bone marrow sections. We designed probes consisting of multiple 20-bp oligos, complementary to the coding sequence of the selected genes, using the Stellaris Probe Designer tool (Biosearch Technologies). Femurs were fixed in cold 4% paraformaldehyde in PBS for 3 h and then incubated in cryoprotective solution (30% sucrose in PBS) at 4 °C overnight. Femurs were then embedded in OCT compound (Leica), in frozen blocks, and stored at –80 °C. After cryosectioning with a cryostat (Thermo), BM sections were prepared as previously described[39], with some modifications: in some cases we added to the probes and hybridization buffer an anti-mouse CD41–biotin antibody (clone MWReg30; eBioscience), and during the first wash, the sections were incubated with streptavidin–Alexa Fluor 488 conjugate (Life Technologies). For intestinal crypt sections, we added phalloidin–Alexa Fluor 488 to the last washing step. Images were acquired as tile scans of *z*-stacks (25 steps with a 0.3-µm interval) on a Leica microscope equipped with an Andor Ikon camera and a 100× oil-immersion objective. Stacks were stitched back using the Grid/Collection stitching plugging included in Fiji[40], and five to nine planes were projected with maximum intensity into a single image.

Network analysis. Interacting structures were simplified to the four most commonly observed configurations: doublet, triplet (three different cells interacting with each other), central connector (*n* cells interacting with one common cell), and in-line (*n* cells interacting in successive order). We used the igraph package in R to construct a list of pairwise interactions and quantified the number of interactions occurring between the different clusters. We performed a permutation test to construct a background model. To this end, we randomly sampled cells from the pool of interacting cells and simulated the same number of doublet, triplet, central connector, and in-line configurations as present in the original data. We repeated this step 10,000 times to obtain a distribution for each type of interaction to which we could compare the experimental number of interactions. We rejected the null hypothesis when the experimental value was equal to or less than 0.05% of the rest of the random distribution. Significantly enriched interactions based on only one interaction were removed from the network visualization. The same strategy was used for the inferred interactions from small intestine crypt cells, but in that case only doublets and triplets were used.

Image analysis. To delimitate the sampling field, we converted RGB images first into a gray image in MATLAB, and then into a binary image by applying

a threshold, where null values indicated areas without cells. To segment the megakaryocytes and the myeloblasts/promyelocytes, we drew the cell boundaries manually in Photoshop and then converted the image into a binary image. These pixels were removed from the sampling field. The center coordinates of the neutrophils, myelocytes, and plasma cells were spotted manually. To estimate interacting distances, we calculated the Euclidian distance between 5–10 visually interacting cells and their closest megakaryocyte or myeloblast/promyelocyte boundary, and selected the highest value for the respective cell type. We determined the number of interacting cells over the whole area by identifying cells with a Euclidian distance between their center coordinates and all pixels corresponding to megakaryocytes or myeloblasts/promyelocytes that was less than or equal to the predetermined interacting distance. To create a background model, we randomly selected a number of coordinates, equal to the number of cells in the experimental data, from the sampling field (area normally occupied by the cell type of interest) to represent virtual cells; this process was repeated 10,000 times. The average radius of the different cell types was taken into account to model physical hindrance: pairs of cells for which the distance between them was less than the sum of both cells' radiuses were excluded and resampled. The *P* value was calculated as the fraction of simulations that had an equal or greater number of interacting cells compared with the actual number. For the distance, the *P* value was calculated as the fraction of simulations with a value equal to or less than the actual measured distance.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. All custom code used in this work is available in the Supplementary Software.

Data availability. Raw and processed sequencing data are available at the Gene Expression Omnibus (GEO) database under accession code GSE89379.

nature research

Corresponding author(s): Alexander van Oudenaarden

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

To build a network of cell interactions in the bone marrow we picked 727 cellular structures for a total of 1728 cells in 18 experiments (so 18 different mice). For the murine small intestine crypts, we picked 576 structures for a total of 1275 cells in 5 independent experiments. For the embryonic fetal liver, we picked 356 cells in 4 independent experiments. Sample size was not pre-determined: as observed in Supplementary figure 5d, the more we sampled, the higher the chance was to discover new enriched interactions. This is a consequence of acquiring more cells of rare cell types. In other words, sampling less leads to lower chances of finding the three new interactions we described, and sampling more would likely lead to discovering additional new enriched interactions.

2. Data exclusions

Describe any data exclusions.

We excluded cells with a total transcript count lower than a threshold. The threshold was chosen as a compromise to retain a maximum of cells (some cell types have an inherent low number of transcripts), while discarding lower quality cells. For the murine small intestine, we discarded cells expressing Kcnq1ot1, Malat1, or Rn45s, as we often observed potential artefactual expression of these genes in low-quality samples.

3. Replication

Describe whether the experimental findings were reliably reproduced.

The main conclusions of the manuscript are based on 18 independent replicates that converges, by adding all data, toward the conclusions presented in the manuscript. We did not observe major batch effects in between experiments. In addition, we quantified the probability of finding the same results with different parameters (threshold and downsampling) as shown in Supplementary figure 5c.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

NA. There are no different experimental groups.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Sampling small interacting structures (doublets, triplets, etc...) was done by mouth pipetting. This cannot be done blinded, but we tried to pick structures as randomly as possible, not selecting them based on arbitrary parameters such as cell size.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

R (3.4.3) on RStudio (1.1.423) with packages:
stringr (1.2.0)
ggplot2 (2.2.1)
gtools (3.5.0)
doMC (1.3.5)
foreach (1.4.4)
iterators (1.0.9)
parallel (3.4.3)
randomForest (4.6-12)
caret (6.0-78)
tidyr (0.8.0)
RaceID2 (https://github.com/dgrun/StemID)
RaceID3 (https://github.com/dgrun/RaceID3_StemID2)
MATLAB (R2012b, 8.0.0.783)
Fiji (Version 1.0) with package:
Stitching (1.2)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All material used are readily available from standard commercial sources, which are stated in the method section.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

We used an anti-mouse CD41-biotin antibody (clone: eBioMWReg30 (MWReg30), ref: 13-0411-81, lot: 4272600, eBioscience, concentration: 0.5 mg/mL used at 1:100). This monoclonal antibody has already been extensively used and described in the literature.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	We used C57Bl/6 and CAST/EiJ female or male mice, from 23 weeks to 9 months, bred in our facility. Experimental procedures were approved by the Dier Experimenten Commissie (DEC) of the KNAW, and performed according to the guidelines.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The study did not involve human research participants.