Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars

Bastiaan Spanjaard1,3, Bo Hu1,3, Nina Mitic1, Pedro Olivares-Chauvet1, Sharan Janjuha2, Nikolay Ninov2 & Jan Philipp Junker1

A key goal of developmental biology is to understand how a single cell is transformed into a full-grown organism comprising many different cell types. Single-cell RNA-sequencing (scRNA-seq) is commonly used to identify cell types in a tissue or organ1. However, organizing the resulting taxonomy of cell types into lineage trees to understand the developmental origin of cells remains challenging. Here we present LINNAEUS (lineage tracing by nuclease-activated editing of ubiquitous sequences)—a strategy for simultaneous lineage tracing and transcriptome profiling in thousands of single cells. By combining scRNA-seq with computational analysis of lineage barcodes, generated by genome editing of transgenic reporter genes, we reconstruct developmental lineage trees in zebrafish larvae, and in heart, liver, pancreas, and telencephalon of adult fish. LINNAEUS provides a systematic approach for tracing the origin of novel cell types, or known cell types under different conditions.

Measuring lineage relationships between cell types is important for understanding fundamental mechanisms of cell differentiation in development and disease2,3. In early development and in adult systems with a constant turnover of cells, short-term lineage predictions can be computed directly on scRNA-seq data by ordering cells along pseudotemporal trajectories according to transcriptome similarity4–6. However, the developmental origin of cells in the adult body cannot be identified using these approaches alone. Several approaches for lineage tracing exist. Genetically encoded fluorescent proteins are widely used as lineage markers5,8, but due to limited spectral resolution, optical lineage tracing methods have mostly been restricted to relatively small numbers of cells. Pioneering studies based on viral barcoding6,10, transposon integration sites11, microsatellite repeats12, somatic mutations13,14, Cre-mediated recombination15, and genome editing of reporter constructs16,17 have used sequence information to increase the diversity of lineage labels. However, these methods have not been coupled with single-cell transcriptome sequencing and therefore do not provide any information on cell type.

Here we present LINNAEUS for simultaneous measurement of single-cell transcriptomes and lineage markers in vivo. The approach is based on the observation that, in the absence of a template for homologous repair, Cas9 produces short insertions or deletions at its target sites, which are variable in their length and position16,18,19. We reasoned that these insertions or deletions (hereafter referred to as genetic ‘scars’) constitute heritable cellular barcodes that can be used for lineage analysis and read-out by scRNA-seq (Fig. 1a). To ensure that genetic scarring does not interfere with normal development, we targeted an RFP transgene in the existing zebrafish line zebrafish line M, which has 16–32 independent integrations of the transgenic construct20. Since these integrations are in different genomic loci (as opposed to being in tandem), we could make sure that scars cannot be removed or overwritten by Cas9-mediated excision. We injected Cas9 and an sgRNA for RFP into one-cell-stage embryos to mark individual cells with genetic scars at an early time point in development (Fig. 1b). Loss of RFP fluorescence in injected embryos served as a direct visual confirmation of efficient scar formation (Supplementary Fig. 1). At a later stage, we dissociated the animals into a single-cell suspension and analyzed the scars by targeted sequencing of RFP transcripts (Online Methods). Simultaneously, we sequenced the transcriptome of the same cells by conventional scRNA-seq, using droplet microfluidics21 (Fig. 1c and Supplementary Figs. 2 and 3).

We analyzed single-cell transcriptomes of >70,000 cells from dissociated larvae at 5 days post-fertilization (dpf). On average, we detected ~3,000 unique transcripts from ~700 genes detected per cell (Supplementary Data set 1). Unsupervised clustering of single-cell transcriptomes22 revealed 70 groups of cells with distinct gene expression programs (Fig. 1d and Supplementary Fig. 4). We assigned these clusters to cell types based on differentially expressed genes (Supplementary Fig. 5 and Supplementary Data set 2). We found that Cas9 generated hundreds of unique scars per animal when targeting a single site in RFP (Fig. 1e,f and Supplementary Fig. 6), suggesting that analysis of genetic scars constitutes a useful approach for whole-organism lineage analysis. Bulk analysis of 32 individual larvae revealed that some scar sequences are more likely to be

1 Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany. 2DFG-Center for Regenerative Therapies Dresden, Technische Universität Dresden, Dresden, Germany. 3These authors contributed equally to this work. Correspondence should be addressed to J.P.J. (janphilipp.junker@mdc-berlin.de).

Received 3 October 2017; accepted 15 March 2018; published online 9 April 2018; doi:10.1038/nbt.4124
created than others, probably through mechanisms like microhomology-mediated repair\(^2\) (Fig. 1c). The scars with the highest intrinsic probabilities may be created multiple times per embryo and are therefore uninformative for lineage reconstruction. We therefore excluded the most frequent scars (\(P > 0.01\)) from further analysis. We found that scarring continued until around 10 h after fertilization, a stage at which zebrafish already have thousands of cells (Fig. 1g). Thus, our injection-based approach for Cas9 induction allowed us to label cells in an important developmental period during which the germ layers are formed and precursor cells for most organs are specified.

We detected variable numbers of scars in single cells, with the average number of scars per cell ranging from \(-2\) for erythrocytes to \(-5\) for epidermal cells. (Supplementary Data set 3). This indicated that some lineage information was lost due to the sparsity of scRNA-seq data. To investigate this issue in more detail, we analyzed single cells from the offspring of fish injected with Cas9 (Supplementary Fig. 7).

In these fish, all cells (independent of the cell type) have the same scar profile, as they are derived from the same pair of germ cells. This analysis confirmed that scar detection efficiencies are dependent on cell type, which probably reflects differences in cell size or promoter strength. Furthermore, we observed some variance in scar-detection abilities may be created multiple times per embryo and are therefore uninformative for lineage reconstruction. Notably, we did not find any highly expressed scars that were undetectable in specific cell types, suggesting that developmental silencing of specific integrations is of no major concern.

To validate that genetic scars contain useful information about lineage relationships, we calculated enrichment or depletion of scar connections between pairs of cell types (Supplementary Fig. 8). Clustering cell types by scar connection strength revealed three groups, each of which contained either mostly ectodermal or mesendodermal cell types (Supplementary Fig. 9). We suggest this...
pattern was caused by a small number of scars that were created during the first cell divisions and then expanded locally. These groups of cell types formed contiguous domains on the zebrafish fate map, but did not strictly correspond to germ layers, since the domain boundaries of scar clones do not necessarily align with the boundaries between germ layers.

Next, we set out to analyze the data at higher resolution and reconstruct lineage trees on the level of single cells instead of cell types. As our previous filtering of frequent scars removed scars that may have been created multiple times, the lineage tree should fulfill the maximum parsimony principle, with every scar being created exactly once. Indeed, maximum parsimony approaches have previously been used for inferring trees from CRISPR–Cas9 lineage data. Earlier studies indicated that missing data do not need to be detrimental to maximum parsimony tree-building methods. However, such studies typically focused on a regime with an order of magnitude fewer taxa.

As our previous filtering of frequent scars removed scars that may have been created multiple times, the lineage tree should fulfill the maximum parsimony principle, with every scar being created exactly once. Indeed, maximum parsimony approaches have previously been used for inferring trees from CRISPR–Cas9 lineage data. Earlier studies indicated that missing data do not need to be detrimental to maximum parsimony tree-building methods. However, such studies typically focused on a regime with an order of magnitude fewer taxa.
than we have cells, and more characters than we have scars. Using two simulated data sets, we found that Camin–Sokal maximum parsimony failed to reconstruct the correct tree for our system (Supplementary Figs. 10 and 11). While it might be possible to solve this issue using modified versions of maximum parsimony or other established tree-reconstruction algorithms, we developed an algorithm that is custom-tailored to our experimental system. Our custom-built strategy also facilitated integration of a filtering step to remove spurious connections. We therefore developed a computational method that fulfills the maximum parsimony criterion and allows for reconstruction of the correct tree in our system even if not all scars are detected in every cell (Supplementary Fig. 10 and Supplementary Note 1). Our algorithm is based on the observation that there is a correspondence between the underlying lineage tree and the resulting scar network graph, a representation of all pairwise combinations of scars that are experimentally observed together in single cells (Fig. 2a). If all scar connections are detected, the scar that is created first has the most connections in the scar network graph, followed by scars that were created next, enabling lineage tree reconstruction in an iterative manner (Fig. 2b). To remove spurious connections, caused by cell doublets, for example, scar connections that do not occur in enough cells were not taken into consideration (Supplementary Note 1).

Using a simulated data set with realistic parameters, we found that our computational method correctly reconstructed lineage trees (Supplementary Fig. 11). Finally, we placed all single cells in the lineage tree based on the detected scars (Fig. 2c). Scar dropouts meant that we did not have full lineage information about every single cell. However, the reconstructed lineage tree allowed us to infer a large part of the missing scar information (Supplementary Fig. 12). The resulting single-cell lineage trees were then converted to a condensed representation for easier interpretation (Fig. 2d).

For the 5-dpf larvae we found that, as expected, the major developmental lineages shown in Figure 1d were separated at least partially from each other in the reconstructed lineage trees (Fig. 2e and Supplementary Figs. 13 and 14). These data can be explored at different levels of granularity, and we decided to focus next on the cell types of the lateral plate mesoderm (Fig. 2f). We found that the different blood cell types had a shared lineage, but we observed that the erythrocytes were also found in an additional branch that did not contain any immune cells. This observation probably reflects the transition from primitive to definitive hematopoiesis in early zebrafish development, as primitive hematopoiesis produces mostly erythrocytes, whereas definitive hematopoietic stem cells are capable of generating all blood cell types. The primitive and definitive hematopoietic stem cells are known to have different developmental origins. We found that the putative definitive hematopoietic cells have a shared lineage origin with endothelial cells (Fig. 2f and Supplementary Fig. 14), which is to be expected, as the definitive hematopoietic stem cells (but not the primitive ones) are derived from endothelial cells of the dorsal aorta. For endodermal and neuronal/neural crest cell types, we observed a similar structure of partially cell-type-specific lineage branches (Supplementary Fig. 15). Due to the stochastic nature of cell labeling in LINNAEUS, scar creation is not synchronized with mitosis. It is therefore important to note that reconstructed lineage trees did not necessarily contain all cell divisions (Supplementary Fig. 11). Furthermore, early zebrafish development is highly variable. We cannot, therefore, expect to find exact correspondence of early lineage trees for all cell types in different animals.

In another set of experiments, we applied LINNAEUS to dissected organs of adult fish (Supplementary Data sets 4 and 5). Analysis of >40,000 cells from the telencephalon, heart, liver, and the primary pancreatic islet by scRNA-seq allowed us to identify many different cell types in these organs (Fig. 3a and Supplementary Figs. 16 and 17). We first analyzed the resulting lineage trees at low granularity, which revealed a strong separation of the individual organs (Fig. 3b and Supplementary Fig. 18). However, we also detected several cell types, mostly from the immune system, that were present in multiple organs. We found that, as expected, the immune cells from different organs were grouped together in the lineage tree, even though the sequencing libraries for the different organs were prepared separately. This observation is an additional important validation of the scar-filtering pipeline, since it shows that even small cell populations such as these immune cells do not acquire scars from other cell types in their organ of origin by mechanisms such as cell doublets or sequencing errors.

![Figure 3](image.png)

**Figure 3** Single-cell lineage analysis of adult organs reveals hierarchies of cell fate decisions. (a) t-SNE representations of scRNA-seq data for dissociated organs from adult zebrafish (n = 3 animals). (b) Lineage tree for organs from one adult. Pie charts are plotted small for n < 50, medium for n ≥ 50, and large for n ≥ 1,000. Scars with creation probability 20.01 were excluded from the analysis. Color code as in a. (c) Lineage tree zoomed into immune cell types from same adult as in b (see color code). As expected, immune cells from different organs cluster together in the lineage tree, even though the sequencing libraries for the different organs were prepared separately. This observation is an additional important validation of the scar-filtering pipeline, since it shows that even small cell populations such as these immune cells do not acquire scars from other cell types in their organ of origin by mechanisms such as cell doublets or sequencing errors.
our computational approach for tree reconstruction that is robust to dropout events under realistic experimental conditions, and by our experimental strategy that uses independent scarring sites whose scars, once created, cannot be changed again. Within a single experiment, data analysis can be performed at different levels of granularity, from germ layers to organs and cell types. Our combined experimental and computational platform thus provides a strategy for dissecting the lineage origin of uncharacterized cell types and for measuring the capacity of lineage trees to adapt to genetic or environmental perturbations. Our approach is based on an existing transgenic animal with multiple integrations of a transgenic construct, which should facilitate adaptation of the method to other model systems.

The observation that Camin–Sokal maximum parsimony failed to reconstruct the correct tree for our system (Supplementary Figs. 10 and 11) serves as a cautionary note regarding computational analysis of CRISPR–Cas9 lineage data. However, additional studies would be necessary to systematically compare our algorithm to existing methods for tree reconstruction under different parameter regimes. Developing a general statistical framework for disentangling biological and technological variability of CRISPR–Cas9 lineage tracing remains another important open challenge for the future. We anticipate that future modifications of the experimental platform, such as, for instance, inducible systems, will enable longer periods of lineage tracing and molecular recording of cellular signaling events during cell fate decisions.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank R. Opitz, M. Guedes Simoes, D. Panakova, T. Durovic, and J. Ninkovic for help with zebrafish experiments. Work in J.P.J.’s laboratory was funded by a European Research Council Starting Grant (ERC-StG 715361 J. Richter for help with cell-type identification. We also acknowledge support by MDC/SPACEVAR, a Fondation Leducq Transatlantic Networks Grant (16CVD03), and a CFI grant (SPC216). B.H. was supported by a PhD fellowship from Studienstiftung des deutschen Volkes.

AUTHOR CONTRIBUTIONS
J.P.J., B.S. and B.H. conceived and designed the project. B.H., N.M. and S.I. optimized dissection and dissociation protocols. B.H. developed the experimental approach for combined scar and transcriptome detection, and B.H. and N.M. performed experiments. B.S. developed computational methods and analyzed the data, with support by P.O.-C. J.P.J. and N.N. guided experiments, and J.P.J. guided analysis. J.P.J. and B.S. wrote the manuscript, with input from all other authors. All authors discussed and interpreted results.

COMPETING INTERESTS
The authors declare no competing interests.

1. Grün, D. & van Oudenaarden, A. Design and analysis of single-cell sequencing experiments. Cell 163, 799–810 (2015).
2. Woodworth, M.B., Ginskis, K.M. & Walsh, C.A. Building a lineage from single cells: genetic techniques for cell lineage tracking. Nat. Rev. Genet. 18, 230–244 (2017).
3. Spanjaard, B. & Junker, J.P. Methods for lineage tracing on the organism-wide level. Curr. Opin. Cell Biol. 49, 16–21 (2017).
4. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386 (2014).
5. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat. Biotechnol. 34, 637–643 (2016).
6. Haghverdi, L., Bittner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods 13, 845–848 (2016).
7. Barker, N. et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449, 1003–1007 (2007).
8. Livet, J. et al. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature 450, 56–62 (2007).
9. Lu, R., Nett, N.F., Quake, S.R. & Weissman, I.L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat. Biotechnol. 29, 928–933 (2011).
10. Naik, S.H. et al. Diverse and heritable lineage imprinting of early hematopoietic progenitors. Nature 496, 229–232 (2013).
11. Sun, J. et al. Clonal dynamics of native hematopoiesis. Nature 514, 322–327 (2014).
12. Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. PLoS Comput. Biol. 1, e50 (2005).
13. Ledoigt, M.A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 350, 94–98 (2015).
14. Ju, Y.S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature 543, 714–718 (2017).
15. Pei, W. et al. Polypox barcoding reveals haematopoietic stem cell fate realized in vivo. Nature 548, 456–460 (2017).
16. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science 353, aaf7907 (2016).
17. Frieda, K.L. et al. Synthetic recording and in situ readout of lineage information in single cells. Nature 541, 107–111 (2017).
18. Junker, J.P. et al. Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. Preprint at bioRxiv https://doi.org/10.1101/056499 (2017).
19. Schmidt, S.T., Zimmerman, S.M., Wang, J., Kim, S.K. & Quake, S.R. Quantitative analysis of synthetic cell lineage tracing using nucleic acid barcoding. ACS Synth. Biol. 6, 936–942 (2017).
20. Pan, Y.A. et al. Zebrabow: multispectral cell labeling for cell tracking and lineage analysis in zebrafish. Development 140, 2835–2846 (2013).
21. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015).
22. Saliba, R., Farrell, J.A., Gerinert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495–502 (2015).
23. Villarreal, D.D. et al. Microhomology directs diverse DNA break repair pathways and chromosomal translocations. PLoS Genet. 8, e1003026 (2012).
24. Schier, A.F. & Talbot, W.S. Molecular genetics of axis formation in zebrafish. Annu. Rev. Genet. 39, 561–613 (2005).
25. Wiens, J.J. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39, 34–42 (2006).
26. Jagnnathan-Bogdan, M. & Zon, L.I. Hematopoiesis. Development 140, 2463–2467 (2013).
27. Kimmel, C.B. & Warga, R.M. Indeterminate cell lineage of the zebrafish embryo. Dev. Biol. 124, 269–280 (1987).
28. Keegan, B.R., Meyer, D. & Yelon, D. Organization of cardiac chamber progenitors in the zebrafish blastula. Development 131, 3081–3091 (2004).
29. Alemany, A. et al. Whole-organism clone tracing using single-cell sequencing. Nature doi:10.1038/nature29969 (2018).
30. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. doi:10.1038/nbt.4103 (2018).
**ONLINE METHODS**

**Zebrafish lines and animal husbandry.** We used the transgenic zebrafish line zebrabow M\(^{32}\) for LINNAEUS. This line has multiple integrations of a transgenic construct that expresses RFP from the ubi promoter, which is constitutively active in all cell types. Fish were maintained according to standard laboratory conditions. All animal procedures were conducted as approved by the local authorities (LAGeSo, Berlin, Germany) under license number G0211/16. We set up crosses between zebrabow M adults with high RFP fluorescence, and we injected the embryos at the one-cell stage with 2 nl Cas9 protein (NEB, final concentration 350 ng/μl) in combination with an sgRNA targeting RFP (final concentration 50 ng/μl, sequence: GGGTGTCGACGTTAGTAGAGCGTTTATGAAATAGCTAAGTAAATATTAGGTGTTGCTGTCTTTC). Since injection efficiencies may vary (Supplementary Fig. 1), we selected embryos with low RFP fluorescence for single-cell analysis. For control experiments in Supplementary Figures 2 and 7 we set up crosses between pairs of adult Cas9-injected fish.

The sgRNA was transcribed in vitro from a template using the MEGAscript T7 Transcription Kit (Thermo Scientific). The sgRNA template was synthesized with T4 DNA polymerase (New England Biolabs) by partially annealing two single-stranded DNA oligonucleotides containing the T7 promoter and the RFP binding sequence, and the tracrRNA sequence, respectively. In the experiments described here, we did not use the ability of the line zebrabow M to switch from RFP to YFP or CFP expression upon addition of Cre\(^{33}\).

**Preparation of single-cell suspensions.** Single larvae at 5 dpf were transferred into 50 μl HBSS containing 1× TrypLE (Thermo Fisher Scientific) and incubated at 33 °C for ~20 min with intermittent mixing with a pipette (every 5 min) until the larva was no longer visible. 500 μl cold HBSS (Thermo Fisher Scientific) supplemented with 1% BSA was then added to the suspension, and the cells were pelleted in a table-top centrifuge at 4 °C for 300g for 5 min. The pellet was washed with 500 μl cold HBSS supplemented with 0.05% BSA and centrifuged down again. The resulting pellet was resuspended in the same buffer and filtered through a cell strainer of 35-μm diameter.

Adult zebrafish were euthanized by an overdose of tricaine in combination with low water temperature. Afterwards, heart, brain, pancreas islets, and liver were isolated from the fish. Single-cell suspensions of the organs were obtained using different protocols:

- **Heart.** The zebrafish heart including atrium, ventricle and bulbus arteriosus was transferred into cold HBSS and opened carefully with forceps, allowing most of the erythrocytes to be washed away. Afterwards, the heart tissue was transferred into 500 μl HBSS containing Liberase enzyme mix (Sigma-Aldrich, 0.26 U/ml final concentration) and Pluronic F-68 (Thermo Fisher Scientific, 0.1%). The reaction was incubated at 37 °C for 30 min while shaking at 750 r.p.m. with intermittent pipette mixing. Afterwards, most of the tissue was dissociated. The reaction was stopped by adding 500 μl cold HBSS supplemented with 1% BSA. The cells were pelleted by centrifuging at 200g in a table-top centrifuge at 4 °C, then washed and filtered following the procedure described above for 5-dpf larvae.

- **Brain.** The telencephalon without olfactory bulbs was isolated in cold HBSS and immediately transferred to a solution of HBSS with 0.81% d-glucose and 15 mM HEPES. Dissociation was initiated by adding 0.1× TrypLE and 0.1% Pluronic F-68 (final concentrations). The tissue was incubated for 30 min at 37 °C while shaking at 750 r.p.m., with occasional gentle mixing. The dissociation reaction was stopped by addition of equal volume of EBS solution containing 4% BSA and 20 mM HEPES. The sample was filtered using a 70-μm filter and centrifuged at 300g for 5 min, after which the pellet was washed once with PBS and resuspended in HBSS with 0.04% BSA. Finally, the suspension was filtered with a 35-μm filter.

- **Pancreas and liver.** The pancreatic tissue containing preferentially the primary pancreatic islet was isolated under a stereomicroscope and transferred into 500 μl HBSS containing 1× TrypLE and 0.1% Pluronic F-68. The liver was isolated and dissected into small pieces, one of which was transferred into 500 μl HBSS containing 1× TrypLE and 0.1% Pluronic F-68. After 30 min of incubation at 37 °C with intermittent pipetting, the suspensions were pelleted, washed and filtered following the procedure described above for 5-dpf larvae.

All final single-cell suspensions were quantified and controlled for quality by microscopy using a hemocytometer.

**Scar detection in bulk samples.** DNA-based scar detection. DNA of single animals was extracted by heating the samples in 50 μl of 50 mM NaOH at 95 °C for 20 min. 1/10 volume of 1 M Tris-HCl, pH = 8.4 was then used to neutralize the mixture. We took 20 μl of the DNA for amplification of scar sequences using RFP-specific barcoded primers. The RFP primers were chosen such that the cut site of Cas9 was positioned approximately in the middle of the sequencing read. We then pooled the PCR products, performed a clean-up reaction using magnetic beads (AMPure Beads, Beckman Coulter), and added Illumina sequencing adapters in a second PCR reaction. Primer sequences are provided in Supplementary Table 1.

RNA-based scar detection. RNA of single or pooled animals was extracted with TRIzol Reagent (Thermo Fisher Scientific) according to the manufacturer’s protocol. The RNA was precipitated using isopropanol, and the pellet was washed twice with 75% ethanol, air-dried, and resuspended in 10 μl of reverse transcription mix (0.3 μM poly-T primer, 1× first strand buffer (Thermo Fisher Scientific), 10 μM DTT, 1 mM dNTPs, 0.5 μl RNaseOUT (Thermo Fisher, Cat. No. 10777019), 0.5 μl SuperScript II (Thermo Fisher, Cat. No. 1806-014)). The reaction was incubated at 42 °C for 2 h for reverse transcription, followed by scar-specific PCR amplification as described above for DNA-based scar detection.

**Transcriptome and scar detection in single cells.** Single cells were captured using Chromium (10X Genomics, PN-120233), a droplet-based scRNA-seq device according to the manufacturer’s recommendations. Briefly, the instrument encapsulates single cells with barcoded beads, followed by cell lysis and reverse transcription in droplets. Reverse transcription was performed with polyT primers containing cell-specific barcodes, Unique Molecular Identifiers\(^{31}\) (UMI), and adaptor sequences. After pooling and a first round of amplification, the library was split in half. The first half was fragmented and processed into a conventional scRNA-seq library using the manufacturer’s protocols. We used the second, unfragmented, half to amplify scar reads by two rounds of PCR, using two nested forward primers that are specific to RFP, and reverse primers binding to the adaptor site. The RFP primers were chosen such that the cut site of Cas9 was positioned approximately in the middle of the sequencing read, ensuring that a broad range of deletion lengths can be reliably detected. Primer sequences are provided in Supplementary Table 1. We confirmed successful library preparation by Bioanalyzer (DNA HS kit, Agilent). Samples were sequenced on Illumina NextSeq 500 2x 75 bp and Illumina HiSeq 2500 2x 100 bp.

**Mapping and extraction of single-cell mRNA transcript counts.** A zebrafish transcriptome was created with Cell Ranger 2.0.2 from GRCz10, release 90. Alignment and transcript counting of libraries was done using Cell Ranger. Cell numbers to be extracted were set at a minimum of 6,000 but were increased if there were substantially more cells with more than 500 unique transcripts. Exact numbers can be found in Supplementary Data set 1.

**Mapping and filtering of single-cell scar data.** Scar reads have the same structure as transcript reads: they consist of a barcode, a UMI and a scar. The scar sequences were aligned using bwa mem\(^{32}\) to a reference of RFP. Valid cell barcodes were identified based on the single-cell transcriptome data (see previous paragraph). We removed reads that were unmapped, had an incorrect barcode, or did not start with the exact PCR primer we used. We truncated all scar sequences to 75 nucleotides and filtered out shorter sequences.

To mitigate the effect of sequencing errors, we implemented several rounds of scar filtering (Supplementary Fig. 2). We started by counting the number of times each molecule was sequenced. Sequencing errors will typically have fewer reads than the actual scars they originate from. As a first filtering step, we therefore removed all molecules only seen once to reduce the complexity in the data set for consecutive filtering steps.

In the second filtering step, we aimed to remove easily recognizable sequencing errors and chimeric reads\(^{33}\). To this end, we consecutively considered scar sequences that have the same cellular barcode and UMI, UMIs that have the same cellular barcode and scar sequence, and cellular barcodes that have the

---

**doi:10.1038/nbt.4124**

© 2018 Nature America, Inc., part of Springer Nature. All rights reserved.
same UMI and scar sequence. In each step, we kept only the molecule with the highest number of reads. The rationale behind this is that it is very improbable to have two valid scar sequences in the same cell with the same UMI, or to have a scar sequence with the same UMI appear in two different cells. The observation of two different UMIs for the same scar in the same cell is much more likely and corresponds to detection of multiple transcripts from the same locus, but information about scar expression levels was not required in our downstream analysis.

In the third filtering step, we specifically targeted sequencing errors within each cell. We compared the scar sequences found within a cell to each other. We filtered out sequences that had a Hamming distance of 2 or less to another scar sequence in the same cell that occurred in at least eight times as many reads. Scar sequences in the same cell that were one Hamming distance apart but had a read ratio less than 8 were tested on three criteria if both of them occurred at least twice in the scar library:

1. Do both scars have more than one transcript?
2. Do both scars occur in cells independently from each other?
3. Do the UMIs of both scars have Hamming distance of two or more?

If two of these criteria were true, the scars were kept and the sequences were placed on a list of validated scars that, if they occurred in the same cell in another library, did not have to be tested anymore. If one or zero criteria were true, the scar that had only one transcript, or the scar that did not occur independently, was filtered out.

In the fourth filtering step, we determined the distribution of reads for the scars we had kept so far. Based on this distribution we set a cut-off and filtered out the scars that did not have at least this number of reads. Finally, for each cell type we determined the distribution of different scars seen per cell and set a maximum number of scars a cell of that type can have. We filtered out cells in which we observed more than this maximum number as possible doublets.

While each scar is identified by its sequence, scars are labeled in the manuscript using their ranking in the bulk scar frequency distribution (e.g., “scar 77”) or their CIGAR code (e.g., “47M6D28M”) as a shorthand notation. Since scars cannot be modified once created, each scar is considered as a separate entity for lineage tracing independent of its sequence.

**Determination of scar probabilities.** We aligned reads from 32 single embryos (DNA-based bulk scar detection) to a reference of RFP. We filtered out unmapped reads and reads that did not start with the exact PCR primer, and truncated all reads to 100 nucleotides, removing shorter ones. To determine the creation probabilities of the different scars, we removed all unscarred RFP reads from each embryo. We normalized the scar content of each embryo to one and calculated scar probabilities as the average ratio with which each scar was observed.

To account for the slightly different sequencing read structure of single-cell and bulk scar detection (see above), we considered only the nucleotides that are shared between the two approaches, and we assigned the bulk scar probabilities to single-cell scars accordingly. Single-cell scars that were not detected in bulk had their probability set to the lowest probability value detected in bulk.

**Determination of scaring dynamics.** Embryos were injected with Cas9 and sgRNA at the one-cell stage. After 1, 2, 3, 4, 6, 8, 10, and 24 h, several embryos were collected and pooled (5–6 for earlier stages, 2–3 for later stages), followed by RNA and/or DNA extraction using TRIzol Reagent. Bulk scar libraries were produced as described above. For each sample, we calculated the percentage of unscarred RFP. We fit a negative exponential to these data, assuming that the fraction of unscarred RFP at t = 0 was 1.

**Identifying cell types.** We used the R package ‘Seurat’ version 2.1.0, for cell-type identification as described below. We removed genes that were not found in at least three cells, and removed cells that had less than 200 of those genes. We log-normalized the transcript counts and removed cells with more than 2,500 genes observed. For single cells from 5-dpf larvae and adult pancreas, we filtered out cells with a mitochondrial content of more than 7.5%, and for single cells from adult hearts and telencephalon we filtered out cells with a mitochondrial content of more than 15%; we expect the cardiomyocytes in these cases to have a mitochondrial content of more than 15%. For single cells from 5-dpf larvae, we filtered out cells with a mitochondrial content of more than 7.5%, and for 2,500 genes observed. For single cells from 5-dpf larvae and adult pancreas, we filtered out cells with a mitochondrial content of more than 15%; we expect the cardiomyocytes in these cases to have a mitochondrial content of more than 15%

**Connection enrichment analysis.** We used an analysis of the scars shared between cells to illuminate the overall structure of the sequencing results from 5-dpf larvae. We expect that cells in which we observe the same scar have a shared lineage. To understand the scarring process better, we aimed to find out which cell types share many scars—these cell types would have a strong lineage relationship, and which cell types do not share many scars—these cell types would not have many immediate shared precursors.

We call cells ‘connected’ if they share at least one scar that has a creation probability of less than 0.1% and is only present in one organism. To find out whether cell types have a higher number of connections between them than expected by chance, we developed the background model described below (see also Supplementary Fig. 8). The background model starts with the realization that a connection is defined by its endpoints, and that therefore the number of expected connections between two cell types is determined by the number of connection endpoints of the two cell types. More precisely, the chance of forming a connection between cell type A and B is given by \( p(A-B) = 2 \cdot CE(A)^{-1} \cdot CE(B) \cdot CE(tot)^2 \), and that of forming a connection within cell type A by \( p(A-A) = CE(A)^{-1} \cdot CE(tot)^2 \), with \( CE(A) \) the number of connection endpoints of cell type A, and \( CE(tot) \) the total number of connection endpoints. These probabilities define a binomial background model. Using this model, we calculate the enrichment z-score between cell types, that is, how many s.d. the observed number of connections between two cell types is away from the expected number of connections. A positive enrichment score indicates more connections than expected by chance, a negative enrichment score indicates fewer connections than expected by chance.

We define the distance between cell types based on their enrichment z-scores by the following equation: \( D(A, B) = 1 - (E(A, B) - E_{min} + E_{max}) \), with \( D(A, B) \) the distance between cell types A and B, \( E(A, B) \) the enrichment z-score between them, \( E_{min} \) the minimal enrichment z-score and \( E_{max} \) the maximum enrichment z-score. The term \( E - E_{min} \) can be understood as a translation of all enrichment scores to positive values. These values are then divided by the maximum value and subtracted from 1 to create distances scaled between 0 and 1. We performed hierarchical clustering on these distances, using average linkage as implemented by the hclust function in R. We performed this analysis for two larvae, cutting the dendrogram into three and four clusters, respectively (Supplementary Fig. 9).

**Tree building.** Our computational method for lineage tree reconstruction consists of two phases. First, we derive the order of scarring events. To do so, we make use of cell network graphs, a representation of all pairwise combinations of scars that are experimentally observed together in single cells (Supplementary Fig. 2). If all scar connections are detected, the scar that is created first has
the highest degree of connections in the scar network graph, followed by scars that were created next, enabling lineage tree reconstruction in an iterative manner (Fig. 2b). In the second phase, we place all cells in the lineage tree according to their scar profile (Fig. 2c). Cells are placed as low in the tree as their scars allow. Due to incomplete scar detection, we do not have full lineage information about every single cell. However, the structure of the scar network graph is robust toward scar dropouts, since it is based on the collective information of thousands of single cells (see simulations in Supplementary Fig. 10). To ensure that lineage tree reconstruction is not affected by known experimental biases, we also included the following measures:

**Double scarring.** Some scars have a higher intrinsic probability than others (Fig. 1e). To minimize the chance of considering scars that may have been created twice or more in the same fish, we excluded all scars that have a probability higher than 0.1%. With this threshold, most scars were unique to a single fish among the replicates studied (Supplementary Fig. 6). Any remaining scars that were not unique to a single replicate were also excluded from the subsequent analysis.

**Cell doublets.** Co-encapsulation of two cells in one droplet is a known limitation of scRNA-seq techniques that are based on droplet microfluidics. Cell doublets can lead to spurious connections between scars in the network graph. Incomplete tissue dissociation, limited barcode diversity, barcode sequencing errors, and free-floating RNA from cells burst in the microfluidic system may potentially have similar consequences. As a protection against this effect, we only accept connections in the scar network graph that are more highly detected than expected by chance given a library-specific doublet rate (typically around 10%, depending on the experimental cell loading rate). See Supplementary Note 1 for details.

**Missing connections.** In case of very low cell numbers or scar detection efficiencies, it is possible that a connection is missed in the scar network graph. To address this issue, we performed a statistical test for each scar to check whether the number of observed connections is compatible with the scar being on top of the current sub-branch, given the numbers of cells and the observed scar dropout rates (Supplementary Fig. 20, Supplementary Note 1). In each iteration, we tested only those scars whose inferred detection rate (if placed on top of the corresponding sub-branch) was higher than 0.1, a threshold derived experimentally in Supplementary Fig. 7.

**Pruning the tree.** Especially for later, smaller branches, it is possible that not enough connections are observed to accurately place them in the lineage tree, resulting in positioning of the branch too high up in the tree (Supplementary Fig. 20 and Supplementary Note 1). We prune the lineage tree for such branches by removing branches that have less than 25% of the cells their siblings have.

Using simulated data under realistic conditions (including cell doublets, as well as cell-type-and integration-site-dependent scar detection efficiencies), we demonstrate that this approach reconstructs the correct lineage tree (Supplementary Fig. 11). Lineage trees were visualized by expanding the R-package `collapsibleTree` (https://github.com/AdeelK93/collapsibleTree/) with previous authorization from the author. This package relies on the D3 javascript libraries.

**Simulations.** We simulated the scarring process during embryo development (Supplementary Figs. 10 and 11). To do this, we used a simple model that starts with one cell, and in which all cells present undergo synchronized mitosis. Every cell cycle, the RFP integrations of the cells can acquire a unique scar. The chance of creating a scar is fixed for every integration for every cell division, and all scars are transmitted to the progeny of the cells.

After simulating the scarring events during development, we also simulate a sequencing experiment that produces data for tree building. To this end, the cells at the bottom of the tree are clonally expanded, generating many copies that all have the same scar profile. The experimental data consist of a sample of these cells, with a scar detection rate determining the chance of seeing a scar that is present in a cell.

We simulated two distinct trees. The first is a simple tree of three generations, where all cell divisions are marked by acquisition of new scars (Supplementary Fig. 10). From this tree we sampled 125 cells with a scar detection rate of 0.3, yielding 99 cells in which at least one scar was detected. This data set was then used to compare LINNAEUS tree building with maximum parsimony tree building.

The second simulated tree was a more realistic tree in which six generations of cells can potentially receive a scar on ten target sites (Supplementary Fig. 11). Here, we used a cell division rate of four per hour, as measured by microscopy. A scaring rate of 0.4 per hour reproduced the scaring dynamics of the exponential fit during the first three hours (Supplementary Fig. 11a). We can use this simulation to estimate the number of new scars per cell division (Supplementary Fig. 11b). In this simulation, we assumed three cell types (fraction 15%, 25%, 60%) with different detection rates (70%, 30%, 10%, respectively). We furthermore assumed that two of the ten target sites are much harder to detect (by a factor 20, that is, detection rates 3.5%, 1.5%, 0.5%). The resulting developmental tree is shown in Supplementary Figure 11c. Due to the stochasticity of scar creation, scars are not created in all precursor cells, and in Supplementary Figure 11d we show the maximal lineage tree that can be measured by scars. We expand all final branches (not shown) and sample 2,000 cells from the resulting pool with a cell doublet rate of 5%, yielding 1,716 cells (including doublet cells) with at least one scar.

**Tree building on simulated data.** To validate our tree building method, we built trees from both simulated trees using the cells sampled as described in the section "Simulations." We compared our results to maximum parsimony tree building as done by the program "mix" in PHYLIP 3.695 (distributed by the author, J. Felsenstein, Univ. Washington, Seattle), using the Camin-Sokal algorithm with missing states encoded as "0." If multiple trees were tied for best tree, we took the first tree generated.

The simple developmental tree (Supplementary Fig. 10) was recreated flawlessly by the LINNAEUS tree building algorithm (Supplementary Fig. 10b). However, maximum parsimony was not able to resolve the tree correctly, creating unjustified complexity due to multiple creation events for the same scar (Supplementary Fig. 10c). The more realistic scar tree (Supplementary Fig. 11) was also recreated faithfully by LINNAEUS (Supplementary Fig. 11f). Maximum parsimony again created a large amount of unjustified complexity with a total of 265 scarring events for 46 scars, an average of over five times per scar (Supplementary Fig. 11g).

**Statistics.** We assessed RNA scar expression rates by comparing scar abundance in DNA to scar abundance in RNA in three animals 24 hours post fertilization. Using 70,251 data points, every one of which represents the RNA and DNA abundances of a sequence in one fish, we found a Pearson correlation of 0.97 between RNA and DNA abundances (Supplementary Fig. 3). We used Seurat to identify cell types in four data sets: 72,252 cells from 5-dpf larvae (n = 7 animals) and cells from three different organs in adult fish (n = 3 animals): heart (12,248 cells), telencephalon (7,045 cells), and pancreas/liver (20,777 cells). Distribution of cell numbers over identified clusters can be found in Supplementary Data sets 2 (larvae) and 4 (adults). We determined differential gene expression using Seurat’s "nhoodbin" test that includes a Benjamini–Hochberg correction of P-values.

To determine whether cell types had a statistically significant amount of connections (Supplementary Figs. 8 and 9), we first determined the theoretical connection probability of two cell types following the reasoning laid out above. We then used a two-tailed binomial test to assess whether the actual observed number of connections between the two cell types is different from the expected number of connections. The P-values were corrected for multiple testing using the Benjamini-Hochberg correction. Values for all 2,485 tests can be found in Supplementary Data 6.

**Data availability.** Sequencing data are deposited on Gene Expression Omnibus, accession number GSE106121. Interactive single-cell lineage trees are available at http://biostatsb.ics.uci.edu/junker/linnaeus/index.html.

**Code availability.** Custom code is provided at https://bitbucket.org/Bastiaanspanjaard/linnaeus. As part of the software package we provide a sample data set on which the code can be run. The function of all scripts is summarized in README.md.

**Life Sciences Reporting Summary.** Further information on experimental design is available in the Life Sciences Reporting Summary.
31. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9, 72–74 (2012).
32. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
33. Wang, G.C. & Wang, Y. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Appl. Environ. Microbiol. 63, 4645–4650 (1997).
34. Waltman, L. & van Eck, N.J. A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B 86, 471 (2013).
35. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).
36. Amir, A.D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. 31, 545–552 (2013).
37. McDavid, A. et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics 29, 461–467 (2013).
38. Howe, D.G. et al. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgensics. Nucleic Acids Res. 41, D854–D860 (2013).
39. Kane, D.A. & Kimmel, C.B. The zebrafish midblastula transition. Development 119, 447–456 (1993).
40. Kobitski, A.Y. et al. An ensemble-averaged, cell density-based digital model of zebrafish embryo development derived from light-sheet microscopy data with single-cell resolution. Sci. Rep. 5, 8601–8610 (2015).
Experimental design

1. Sample size

Describe how sample size was determined.

The number of cells to be analyzed was based on the maximum number attainable. Since the various detected cell types have vastly different abundances, we focused the analysis on those cell types that were detected at high enough numbers for statistical analysis.

2. Data exclusions

Describe any data exclusions.

No data was excluded. However, the following filtering steps were applied to the scar sequencing data in order to mitigate the effect of sequencing errors, PCR amplification errors and formation of chimeric reads: removal of all sequences that were detected only once; removal of molecules with the same BC+UMI but different scar, same BC+scar but different UMI, and same UMI+scar but different BC; removal of scars with a Hamming distance $\leq 2$ to a more highly detected scar; removal of the most lowly detected scars; and removal of cells that express more scars than expected (putative doublets).

For lineage analysis, only scars with creation probability $<0.01$ (as determined by bulk experiments) were used in order to minimize artifacts caused by repeated creation of the same scar sequence in one animal.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All attempts at replication were successful.

4. Randomization

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

Formal randomization of samples was not applicable, since no manual selection of single cells was performed during experiments and analysis.

5. Blinding

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

Formal blinding was not applicable, since no manual selection of single cells was performed during experiments and analysis.

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.

- Cell Ranger (version 2.0.2) was used to align and demultiplex single-cell transcriptomics data.
- bwa mem (version 0.7.12) was used to map scar data.
- All data analysis was done in R (version 3.3 and higher).
- The R package Seurat (version 2.1.0) was used to cluster single-cell transcriptomics data.
- We developed custom code to filter, analyze and visualize scar data. The custom code is provided at https://bitbucket.org/Bastiaanspanjaard/linnaeus. This repository contains an expansion of the R package collapsibleTree (version 0.1.6) with authorization of the author. The repository contains instructions on how to run the code.

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.

- No unique materials were used.

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

- No antibodies were used.

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.

Zebrafish (Danio rerio, strain TL), line zebrabow M (Pan et al., Development, 2013). Larvae were analyzed at 5 dpf without taking sex into account. Adult animals were males in the age range 6-9 months.

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.