Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars

Jan Philipp Junker[1,2,3,4], Bastiaan Spanjaard[1,3], Josi Peterson-Maduro[1], Anna Alemany[1], Bo Hu[2], Maria Florescu[1], Alexander van Oudenaarden[1,4]

[1] Hubrecht Institute–KNAW and University Medical Center Utrecht, 3584 CT Utrecht, the Netherlands

[2] Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, 13092 Berlin-Buch, Germany

[3] equal contribution

[4] correspondence: janphilipp.junker@mdc-berlin.de (J.P.J.), a.vanoudenaarden@hubrecht.eu (A.v.O.)

A key goal of developmental biology is to understand how a single cell transforms into a full-grown organism consisting of many cells. Although impressive progress has been made in lineage tracing using imaging approaches, analysis of vertebrate lineage trees has mostly been limited to relatively small subsets of cells. Here we present scartrace, a strategy for massively parallel whole-organism lineage tracing based on Cas9 induced genetic scars in the zebrafish.

The timing of each cell division and the fate of the progeny define the lineage of an organism. Analysis of the lineage history of cell populations can reveal the developmental origin and the clonality of cell populations[1]. Genetically encoded fluorescent proteins are widely used as lineage markers[2,3], but due to limited spectral resolution this approach has mostly been restricted to tracking the lineage of a relatively small subset of cells. Recent progress in live imaging has allowed for following many individual cells over time in optically transparent samples such as early fly and zebrafish embryos[4,5]. Nevertheless, direct observation of all cell divisions is generally only possible at the earliest developmental stages. RNA sequencing has emerged as a powerful method for systematic expression profiling of single cells and for computational inference of differentiation dynamics[6-8]. However, our ability to harness the enormous multiplexing capacity of high-throughput sequencing for lineage tracing has so far been lagging behind, despite pioneering studies in the hematopoietic system based on viral barcoding[9,10] or transposon

tagging[11]. Here we present scartrace, a method for massively parallel whole-organism lineage tracing based on CRISPR/Cas9 genome editing.

Scartrace is based on the observation that, in the absence of a template for homologous repair, Cas9 produces short insertions or deletions (indels) at its target sites, which are variable in their length and position[12,13]. We reasoned that these indels (hereafter referred to as genetic "scars") constitute a permanent, heritable cellular barcode that can be used for lineage tracing. To ensure that genetic scarring does not interfere with the viability of the cells, we targeted GFP in a zebrafish line with 4-5 integrations of a histone-GFP transgene[14] (Supplementary Fig. 1). We injected sgRNA for GFP and Cas9 mRNA or protein into the yolk of 1-cell stage embryos in order to mark individual cells with genetic scars at an early time point in development (Fig. 1a). Loss of GFP fluorescence in injected embryos served as a direct visual confirmation of efficient scar formation. Scars were then analyzed at a later time by targeted sequencing of GFP (see Methods).

In order to determine how many cell lineages can be distinguished with scartrace, we analyzed the complexity of scar sequences in whole embryos at 24 hours post fertilization (hpf). We found that Cas9 generated hundreds of unique scars when targeting a single site in GFP (Fig. 1b and Supplementary Table 1), suggesting that analysis of genetic scars constitutes a powerful approach for whole-organism lineage tracing. Scar abundances spanned several orders of magnitude. While many scars were unique to a particular embryo, we also observed that others, in particular the most abundant ones, appeared in multiple embryos and displayed correlated abundances between different samples. This finding indicated that some scar sequences are more likely to be created than others, possibly through mechanisms like microhomology-mediated repair[15]. Consequently, the scars with the highest intrinsic probabilities will be created multiple times in different embryos. Scarring continued until around 8 hpf, a stage at which zebrafish already have thousands of cells (Fig. 1c). Embryos consistently exhibited a higher percentage of scarred GFP when injecting Cas9 protein compared to Cas9 mRNA, suggesting that protein may act earlier than mRNA (Fig. 1d). The dynamics of scar formation can potentially be adjusted further by injecting variants of Cas9. To illustrate this, we constructed an unstable variant of Cas9 (uCas9). Embryos injected with uCas9 mRNA had lower wildtype levels than embryos injected with Cas9 mRNA (Fig. 1d). Thus, our simple 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[16].

We next aimed to use scartrace to determine the clonality of specific cell populations. We chose zebrafish germ cells for this proof of concept experiment, as previous studies have established that the entire germ cell pool is derived from 4 primordial germ cells specified at around the 32 cell stage. These 4 founder germ cells start to proliferate at around the 4,000 cell stage, and their total number increases to 25-50 by the end of the first day of development[17]. We hence raised selected histone-GFP zebrafish that were injected with Cas9 protein and sgRNA to adulthood. We then bred a heterozygous female with a wildtype male and sequenced scars in the separate resulting embryos (Fig. 2a). This approach allowed us to sequence the clonal complexity of the maternal germ cell pool on a single cell level. As expected, we detected multiple scars in the F1 generation, with fractions that are consistent with the number of integrations we observed in whole-genome sequencing. In two crosses of the same pair of fish we observed the same 3 dominant clones of germ cells (Fig. 2b), suggesting that one of the four founder germ cell clones was lost almost completely during subsequent development. In summary, these experiments validated scartrace as a reliable and reproducible approach for whole-organism lineage tracing and systematic quantification of cell population clonality.

After this first proof of principle, we decided to apply scartrace to a more complex, yet well-studied biological system. We chose to focus on the zebrafish caudal fin, a structure that consists of about a dozen different cell types. Zebrafish have the remarkable capacity to regenerate fins upon amputation, with most cell types in the regenerated organ derived from cell-type specific precursors[18]. We dissected amputated fins into pieces consisting of proximal, central, and distal positions of individual rays and interrays (Fig. 2c). This procedure was repeated twice after the fin had fully regenerated. We found that the scartrace data segregated into four clusters, with each cluster consisting of pieces from the original as well as both regenerated fins (Fig. 2d). These clusters were spatially organized along the dorsoventral and anteroposterior axis (Fig. 2e). Interestingly, the spatial position of the individual clusters remained largely constant between the original and the regenerated fins. Correlation was particularly strong between pieces belonging to the same ray or interray, suggesting that growth patterns predominantly followed the

proximal-distal axis (Supplementary Fig. 2). These findings indicate, in agreement with previous reports[18,19], that formation of the caudal fin proceeds similarly during development and during regeneration. Interestingly, we found that the number of detected scars decreased only mildly in the regenerated fin (Supplementary Fig. 3). This observation suggests that most clones that gave rise to the original fin survived until adulthood and were reactivated upon amputation.

Here we introduced scartrace, a simple approach for systematic lineage tracing of whole organisms and for quantification of the clonality of tissues and organs. A key advantage of our approach is that cell labeling as well as detection of lineage markers can be performed in a high-throughput manner. For instance, previous studies of fin regeneration required manual analysis of hundreds or thousands of fish, in which the progeny of only one or a few cells was traced[18,19]. Our approach allows lineage tracing of many cells in the same organism, facilitating for instance quantitative studies of clonality changes during life and after perturbation. Importantly, scartrace does not require prior knowledge of cell type markers. Scartrace will hence be an ideal method for elucidating the origin of anatomical structures whose developmental lineage is unknown. Injection of Cas9 and sgRNA into the zygote of histone-GFP zebrafish enabled us to label most cells in the organism at important developmental stages such as mid-blastula transition and gastrulation. Scartrace is based on a fish line that is already present in most zebrafish facilities and on easily available reagents, and can hence be adopted immediately by other labs.

Our strategy is similar to a recently published method using concatemerized Cas9 target sites[20]. Concatemerized target sites have the advantage that multiple scars can be sequenced on the same read. However, a potential downside is that target sites may be excised, leading to loss of previously established scars. Cumulative acquisition of scars can be used for reconstructing lineage trees. It is however important to take into account that scars have different intrinsic probabilities, and that the probability for repeated generation of the same scar in a single embryo is sequence-dependent. We expect that there will be variants of Cas9-mediated lineage tracing in the future, using for instance self-targeting sgRNAs[21,22] or inducible systems. Combination of scartrace with single-cell RNA-seq will ultimately enable unbiased cell type identification and simultaneous lineage reconstruction of heterogeneous cell populations.

Methods

Zebrafish

Embryos of the transgenic zebrafish line Tg((H2Af/va:H2Af/va-GFP)^{kca66} [14] were injected at the 1-cell stage with 1 nl Cas9 protein (NEB, final concentration 1590 ng/µl) or mRNA (final concentration 300 ng/µl) or uCas9 (final concentration 300 ng/µl) in combination with an sgRNA targeting GFP (final concentration 25 ng/µl, sequence: GGTGTTCTGCTGGTAGTGGT). The uCas9 was constructed by introducing an 8 amino acid long destruction box used in the zFucci system[23] into the N-terminus of the pCS2-nCas9n vector. Cas9 mRNA and uCas9 mRNA were in vitro transcribed from linearized pCS2-nCas9n vector (Addgene plasmid # 47929)[12] using the mMESSAGE mMACHINE SP6 Transcription Kit (Thermo Scientific). The sgRNA was in vitro transcribed 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 promotor and the GFP binding sequence, and the tracrRNA sequence, respectively[18]. All studies involving vertebrate animals were performed with institutional approval of the Hubrecht Institute, and were reviewed by the dierexperimentencommissie (DEC) of the KNAW.

Scartrace protocol

The protocol was performed using either reverse transcribed RNA or genomic DNA as starting material. RNA was extracted from homogenized zebrafish samples with TRIzol reagent (Ambion) according to the manufacturer's instructions. For DNA extraction we used either TRIzol extraction or SEL buffer + Proteinase K (50 mM KCl, 2.5 mM MgCl2, 10 mM Tris pH 8.3, 0.045% Igepal, 0.045% Tween-20, 0.05% Gelatine and 0.1 mg/ml Proteinase K). For extraction with SEL buffer + Proteinase K, samples were incubated for 1 h at 60 °C and 15 min at 95 °C. Scar detection for the experiments shown in Figure 1b was done on mRNA, all other experiments shown here were performed on genomic DNA. For scarring dynamics experiments in Figure 1c, embryos were collected regardless of GFP expression. For all other experiments, individual fish with successful injections were selected based on disappearance of GFP fluorescence.

GFP sequences were amplified by PCR with primers complementary to GFP, including an 8 bp barcode sequence and adapter sequences for Illumina sequencing. Samples were then pooled and subjected to magnetic bead cleanup (AMPure XP beads – Beckman Coulter). Finally, sequenceable libraries were generated by a second round of PCR with indexed primers from Illumina's TruSeq Small RNA Sample Prep Kit. We confirmed successful library preparation by Bioanalyzer (DNA HS kit, Agilent). Samples were sequenced on Illumina NextSeq 500 2x75bp.

Determination of GFP-integrations by whole-genome sequencing

Genomic DNA from a homozygotic histone-GFP fish was extracted using DNeasy Blood & Tissue Kit (Qiagen). Following extraction, the DNA was fragmented using an S2 Focused-ultrasonicator (Covaris) with a target peak of 500 bp. For library preparation we used the NEBNext Ultra library preparation kit for Illumina (E7370S) and NEB Multiplex Oligos for Illumina (E7500L). We modified the genome sequence danRer10[24] by adding a sequence for the histone-GFP fusion gene, and removing chromosome five, the location of the h2afv histone to minimize the number of reads both mapping to the fusion gene and the fifth chromosome. We used bwa mem 0.7.10[25] to align the reads to this modified genome, selected reads for which both mates align, and removed duplicate reads.

We binned these reads into bins of 1,000 bases; one of these bins contained 91% of the 717 bases that make up the GFP-sequence. To determine the copy number of this bin we first performed GC-correction[26]: we calculated the mean bin reads for each integer GC-percentage for which we had at least two thousand bins. We then divided the number of reads for each bin by the mean bin reads corresponding to its GC-percentage. A histogram of all corrected bin counts showed a clear peak around 1, as expected (Supplementary Fig. 1). The corrected bin count for the GFP-bin is 4.6, indicated by a red line in the figure. We concluded that whole-genome sequencing is consistent with eight or ten GFP-integrations for a homozygote, and four or five GFP-integrations for a heterozygote.

Determination of scar abundance

Sequencing data were mapped to GFP using bwa mem 0.7.10. We only considered

reads for which the left mate was mapped in the forward and the right mate in the reverse direction, for which the left mate contained a correct barcode, and for which the right mate started with the primer sequence and had a length of 76 nucleotides. The PCR primer locations were chosen such that the scar can be found in the right mate. To identify different scars, we first classified right mate reads using the CIGAR string, which describes length and position of insertions and deletions in the alignment, combined with the 3' location of the right mate. Within each CIGAR string we performed a subclassification based on which sequences it contained. We retained those CIGAR strings for which we saw at least 20 reads in a library, and those sequences that made up at least 5% of all reads in its CIGAR-class.

F1 embryo data analysis

After determining scar abundances as described above, we selected samples with a sufficiently high read count. We first calculated the average number of reads for samples in a given library, and then filtered out all samples that did not have at least 10% of this average.

After this, we singled out scars that made up at least 5% of all reads of at least one embryo. The abundances of all other scars were summed up and displayed as 'other'.

Fin data analysis

We selected fin pieces with sufficiently high read count similarly to F1 embryos as described above. We calculated the pairwise Pearson correlations after removing non-scarred GFP. The hierarchical clustering in Figure 2d was done on pairwise correlations between the pieces, using Ward's method for agglomeration. We separated the resulting tree into four clusters.

Confidence intervals

The 95% confidence intervals shown in Figures 1c, 1d and Supplementary Figure 2 were generated by sampling the populations with replacement, determining the mean of the population, repeating this thousand times and taking the upper boundary of the

lowest 2.5%-quantile and the lower boundary of the highest 2.5%-quantile of all means generated. Sample size was 100 for Figure 1c and 40 for Figure 1d.

References

1. Kretzschmar, K. & Watt, F. M. Lineage Tracing. *Cell* **148,** 33–45 (2012).
2. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449,** 1003–1007 (2007).
3. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450,** 56–62 (2007).
4. Keller, P. J., Schmidt, A. D., Wittbrodt, J. & Stelzer, E. H. K. Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. *Science* **322,** 1065–1069 (2008).
5. Amat, F. *et al.* Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat Methods* **11,** 951–958 (2014).
6. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25,** 1491–1498 (2015).
7. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14,** 618–630 (2013).
8. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145 (2015).
9. Lu, R., Neff, 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. *Nature Biotechnology* **29,** 928–933 (2011).
10. Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496,** 229–232 (2013).
11. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514,** 322–327 (2014).
12. Jao, L.-E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc Natl Acad Sci USA* **110,** 13904–13909 (2013).
13. Varshney, G. K. *et al.* High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res* **25,** 1030–1042 (2015).
14. Pauls, S., Geldmacher-Voss, B. & Campos-Ortega, J. A. A zebrafish histone variant H2A.F/Z and a transgenic H2A.F/Z:GFP fusion protein for in vivo studies of embryonic development. *Dev Genes Evol* **211,** 603–610 (2001).
15. Villarreal, D. D. *et al.* Microhomology Directs Diverse DNA Break Repair Pathways and Chromosomal Translocations. *PLoS Genet* **8,** e1003026–12 (2012).
16. Woo, K. & Fraser, S. E. Order and coherence in the fate map of the zebrafish nervous system. *Development* **121,** 2595–2609 (1995).
17. Raz, E. Primordial germ-cell development: the zebrafish perspective. *Nature Reviews Genetics* **4,** 690–700 (2003).
18. Tu, S. & Johnson, S. L. Fate Restriction in the Growing and Regenerating

Zebrafish Fin. *Dev Cell* **20,** 725–732 (2011).

19. Stewart, S. & Stankunas, K. Limited dedifferentiation provides replacement tissue during zebrafish fin regeneration. *Dev Biol* **365,** 339–349 (2012).
20. McKenna, A. *et al.* Whole organism lineage tracing by combinatorial and cumulative genome editing. *Science* aaf7907 (2016). doi:10.1126/science.aaf7907
21. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *bioRxiv* 1–19. doi:10.1101/055863
22. Perli, S., Cui, C. & Lu, T. K. Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells. *bioRxiv* 1–42. doi:10.1101/053058
23. Sugiyama, M. *et al.* Illuminating cell-cycle progression in the developing zebrafish embryo. *Proc Natl Acad Sci USA* **106,** 20812–20817 (2009).
24. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** 498–503 (2014).
25. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997v2 (2013).
26. Zhang, C. *et al.* A Single Cell Level Based Method for Copy Number Variation Analysis by Low Coverage Massively Parallel Sequencing. *PLoS ONE* **8,** e54236–9 (2013).

Figure 1

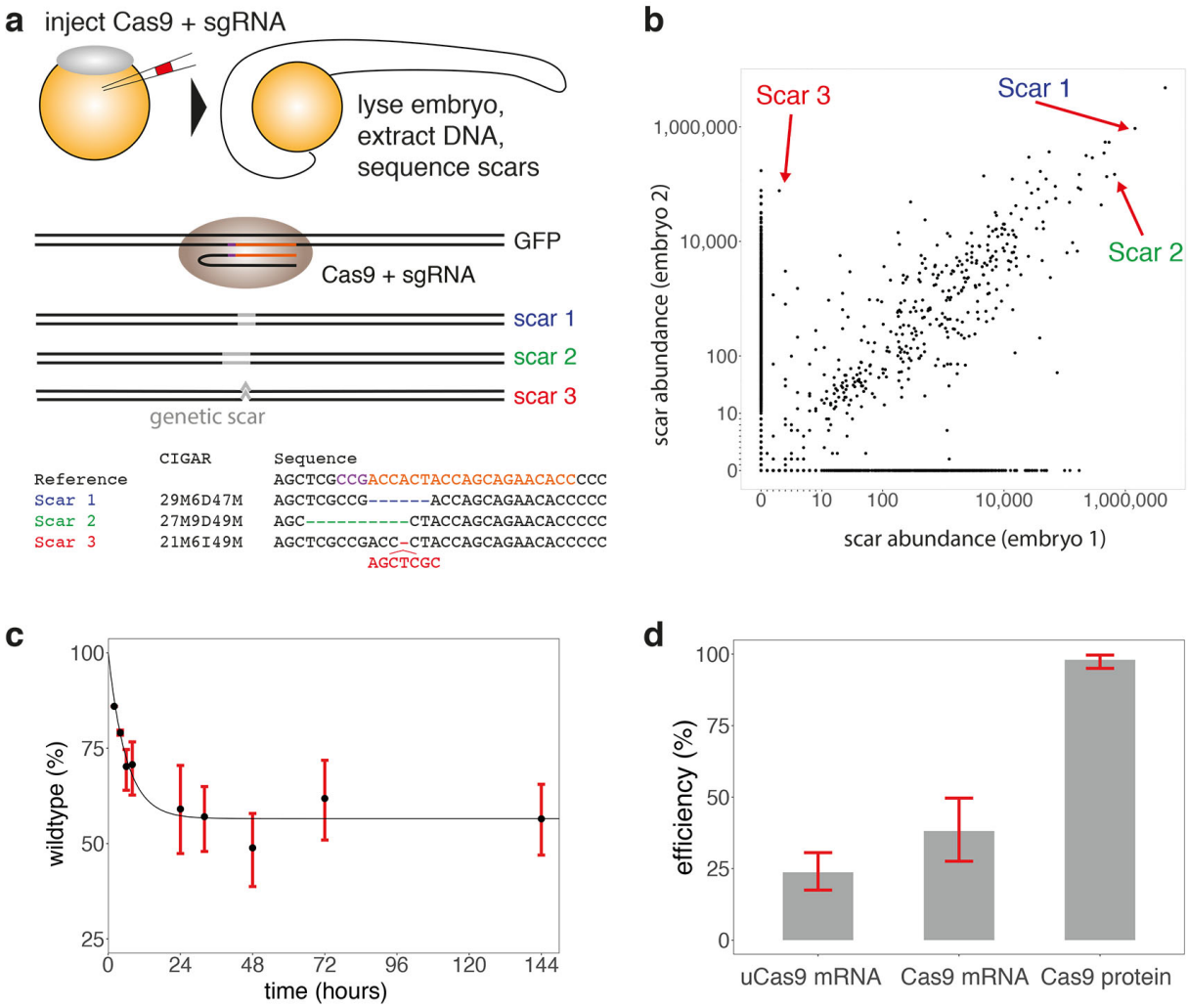

Figure 1. Cas9 generates large diversity of single-cell barcodes. a. Sketch of the experimental protocol. Injection of Cas9 and sgRNA into the zygote marks cells with genetic scars at an early developmental stage. Scars are analyzed by using the CIGAR code, which describes length and position of insertions and deletions. b. Correlation of scar abundances between two different 24 hpf embryos. Each black dot represents a different unique scar. Axes are linear between zero and ten, and logarithmic for higher abundances. c. Dynamics of scar formation. Fraction of wildtype reads, averaged over multiple embryos, as a function of time. We fitted a negative exponential as guide to the eye. d. Scarring efficiency after injection of uCas9 mRNA, Cas9 mRNA or protein.

Figure 2

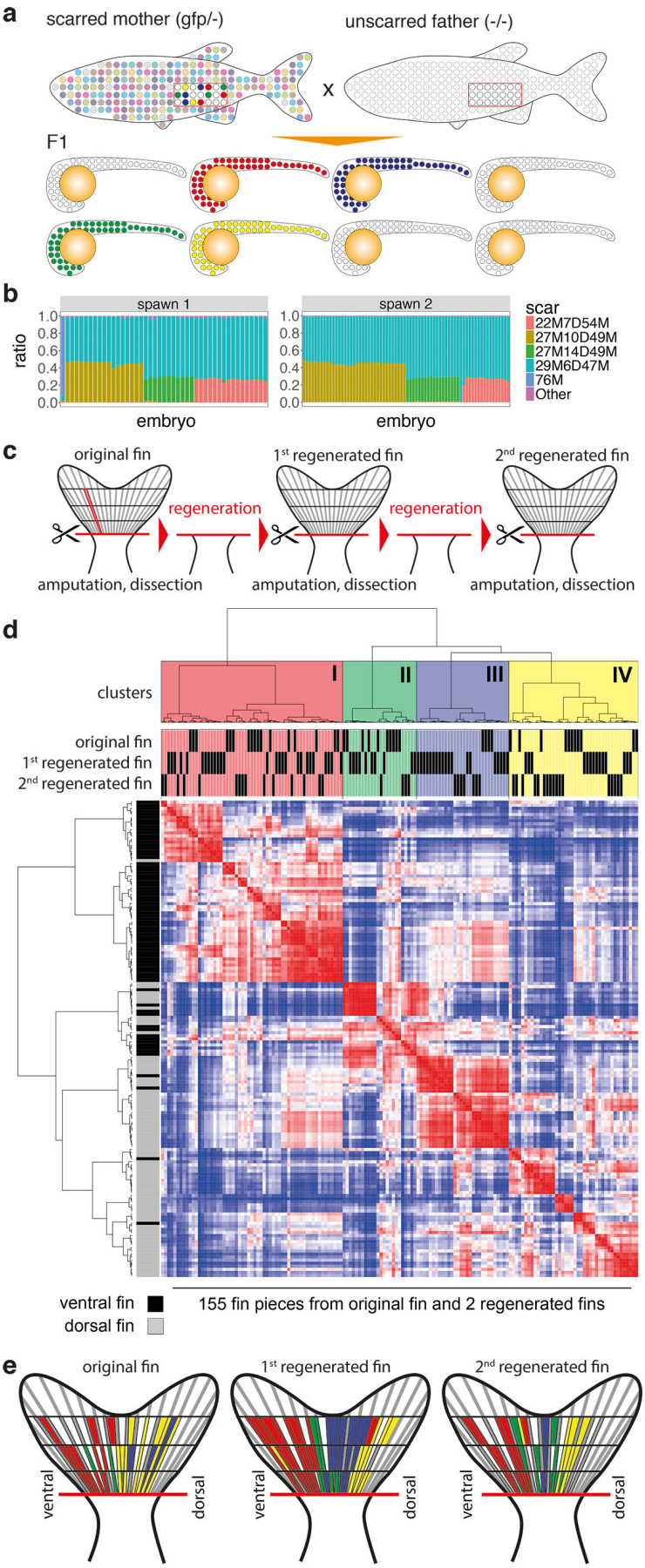

Figure 2. Clonality of the germ cell pool and the regenerating caudal fin. a. Sketch of the experimental protocol. An adult female with Cas9-induced scars was crossed with a wildtype male. Since we used a heterozygous fish only half of the germ cells (red box) carried scars. Cells with different scar profiles are indicated by circles filled with different colors. To analyze the clonality of the mother's germ cell pool, we sequenced the scars of the F1 generation. b. Results for 2 independent crosses of the same fish. We detected 3 dominant clones of germ cells with identical scar profiles. c. Sketch of the experimental protocol. After amputation of the caudal fin, we dissected rays and interrays into proximal, central, and distal pieces. The experiment was repeated twice after regeneration of the fin. d. Clustering of scar abundances. Clusters were determined for a combined dataset including all data from the original and regenerated fins. e. Spatial profile of scar clusters. Color code as in Figure 2d.