



Contents lists available at ScienceDirect

## Journal of Genetics and Genomics

Journal homepage: [www.journals.elsevier.com/journal-of-genetics-and-genomics/](http://www.journals.elsevier.com/journal-of-genetics-and-genomics/)

## Original Research

## Achieving single-cell-resolution lineage tracing in zebrafish by continuous barcoding mutations during embryogenesis

Zhan Liu, Hui Zeng, Huimin Xiang, Shanjun Deng, Xionglei He\*

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China

## ARTICLE INFO

## Article history:

Received 3 March 2024

Received in revised form

3 April 2024

Accepted 7 April 2024

## Keywords:

Lineage tracing

Cellular barcoding

Zebrafish cell phylogeny

Fin regeneration

Germ cell lineage

## ABSTRACT

Unraveling the lineage relationships of all descendants from a zygote is fundamental to advancing our understanding of developmental and stem cell biology. However, existing cell barcoding technologies in zebrafish lack the resolution to capture the majority of cell divisions during embryogenesis. A recently developed method, a substitution mutation-aided lineage-tracing system (SMALT), successfully reconstructed high-resolution cell phylogenetic trees for *Drosophila melanogaster*. Here, we implement the SMALT system in zebrafish, recording a median of 14 substitution mutations on a one-kilobase-pair barcoding sequence for one-day post-fertilization embryos. Leveraging this system, we reconstruct four cell lineage trees for zebrafish fin cells, encompassing both original and regenerated fin. Each tree consists of hundreds of internal nodes with a median bootstrap support of 99%. Analysis of the obtained cell lineage trees reveals that regenerated fin cells mainly originate from cells in the same part of the fins. Through multiple times sampling germ cells from the same individual, we confirm the stability of the germ cell pool and the early separation of germ cell and somatic cell progenitors. Our system offers the potential for reconstructing high-quality cell phylogenies across diverse tissues, providing valuable insights into development and disease in zebrafish.

Copyright © 2024, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, and Genetics Society of China. Published by Elsevier Limited and Science Press. All rights reserved.

## Introduction

Mapping lineage relationships between cells in a multicellular organism is a fundamental question in developmental biology (Kretzschmar and Watt, 2012). Completed in the 1980s, the description of the entire cell lineage of the nematode *Caenorhabditis elegans* by direct observation established a paradigm for cell lineage tracing (Sulston and Horvitz, 1977; Sulston et al., 1983). However, direct observation may not be suitable for complex organisms without invariant lineages and containing millions to billions of cells (VanHorn and Morris, 2021). Advancements in cell marking technologies encompass various approaches such as viral barcoding, Cre-mediated recombination, integration of random barcodes, and the utilization of natural somatic mutations (Salipante and Horvitz, 2006; Barker et al., 2007; Lu et al., 2011; Pei et al., 2017; Rodriguez-Fraticelli et al., 2018; Wagner et al., 2018). In recent years, progression in high-throughput sequencing and genome editing technology has made it feasible to trace cell lineages in complex

multicellular organisms comprehensively. Among these methods, the application of Cas9 genome editing on target transgenic cassettes has been widely adopted (Alemany et al., 2018; Raj et al., 2018; Spanjaard et al., 2018; Bowling et al., 2020; Li et al., 2023). Although Cas9 creates diverse editing on barcodes, these methods rely on indels as genetic scars, where large deletions could remove previously recorded lineage information (VanHorn and Morris, 2021). Indeed, in zebrafish, the recorded mutations per barcode are approximately three, significantly fewer than the tens of mitotic divisions occurring from a zygote to a fully developed complex organism, indicating that these systems miss recording lots of mitotic divisions.

Recently, SMALT, a substitution mutation-aided lineage-tracing system, was used in mapping cell phylogeny for *Drosophila* (Liu et al., 2021). SMALT utilized AID10-iScel to induce C-to-T transition mutations on a 3-kb synthetic cassette, successfully recording average over 20 mutations in early-adult fly cells. Here, we applied this system to zebrafish, utilizing a more condensed 1-kb barcoding sequence. In the 1-day post-fertilization (dpf) embryo, we documented a median of 14 substitution mutations on this 1-kb barcode. This mutation count is comparable to the estimated number of mitotic cell divisions from

\* Corresponding author.

E-mail address: [hexiong1@mail.sysu.edu.cn](mailto:hexiong1@mail.sysu.edu.cn) (X. He).

zygote to 1dpf embryo, providing a detailed recording of cell lineage history. Thus, our barcoding system outperforms the state-of-the-art Cas9-based lineage tracing method in zebrafish by recording a much higher number of mutations.

To demonstrate the capabilities of our system in zebrafish lineage reconstruction, we applied it to fin regeneration. Despite decades of detailed research by scientists on cell types and molecular pathways involved in zebrafish fin development and regeneration (Poss et al., 2000; Whitehead et al., 2005; Petrie et al., 2014; Pfefferli and Jazwińska, 2015; Sehring and Weidinger, 2020), there is still a lack of quantitative descriptions of original and regenerated fin cell relationships based on cell phylogenetic trees. Besides, previous studies on fin generation mainly focus on the caudal fin (Lee et al., 2005; Blum and Begemann, 2015). By utilizing the mutated 1-kb barcode as alleles, we reconstructed four nearly single-cell-resolution lineage trees using zebrafish fin cells from 8 parts of fins in four individuals. Each tree comprised hundreds of internal nodes with 99% median bootstrap values supported. Using these trees, we inferred that regenerated fin cells mainly originate from the cells of the same part of the fins.

Furthermore, with this system, we sampled germ cells from separate spawning events of the same individual. Based on the cell lineage trees, we observed the mixing of germ cells from separate spawning and an early separation between the progenitor of germ cells and somatic cells. Besides, by analyzing the monophyletic clades formed by germ cells and performing a targeting coalescent analysis (TarCA) (Deng et al., 2024), we revealed a relatively small germ cell progenitor number.

## Results

### High mutation recording and barcoding diversity in the zebrafish embryo

Building upon the previous SMALT mutation data collected from *Drosophila* and yeast, we re-designed and synthesized a more compact 1-kb barcode. Compared to the original 3-kb barcode, this 1-kb barcode still retains 16 iScel binding motifs but with only 80 bp between the downstream of the two adjacent iScel binding sites facing each other (Supplementary Note 1). Hence, we anticipated that it would facilitate PCR amplification more effectively than the original 3-kb readout sequence while still preserving a considerably large mutation space. We used the Tol2 transposon system to generate a stable single-copy transgenic line carrying the 1-kb barcode (Kawakami, 2007; Kwan et al., 2007; Mosimann et al., 2011). We have bred this transgenic line to F7, and these fish exhibit normal morphology and fertility. We injected in-vitro transcribed AID10-iScel-UGI mRNA, generated from our SMALT system, into 1-cell stage heterozygous transgenic embryos (Fig. 1A). To assess the overall mutation efficiency and diversity of our SMALT system in zebrafish, we performed PCR amplification of the 1-kb barcode sequence at 1dpf using the genomic DNA from a pool of 10 AID10-iScel-UGI mRNA-injected individuals. In the injected embryos, 18 out of 20 sequenced clones contained mutations, with a median of 14 substitution mutations per 1-kb barcode (Fig. S1A). Conversely, in the uninjected 1dpf control embryos, we sequenced 13 clones, and only 5 clones had one mutation each, while the remaining clones showed no mutation, confirming the robust barcoding mutations in our system (Fig. S1A). Besides, all 18 mutated 1-kb barcode clone sequences were distinct, indicating the high barcoding diversity of our system (Fig. S1C). Cell lysis count data from a previous study showed that embryos at 1dpf have around 50,000 cells (Wagner et al., 2018). Assuming no cell deaths and each cell divides to produce two daughter cells, there would be approximately 16 mitotic divisions from zygote to 1dpf embryos, as

calculated by  $\log_2(50000)$ . Thus, it is estimated that, on average, 0.875 mutations are recorded on the readout sequence per cell generation (calculated as 14 mutations over 16 cell divisions), which is close to the goal of one mutation per division. However, it is essential to note that these mutation per generation estimations are rough and likely to be overestimated due to assumptions such as no cell deaths and all symmetric cell divisions (Cole and Ross, 2001).

### Demonstrating cell lineage reconstruction in embryos

To demonstrate the principle of our system in zebrafish cell phylogenetic tree reconstruction, we performed PCR on the 1-kb barcode of three 1dpf individuals (hereafter named individuals A, B, and C) separately. For each individual, we sequenced 15 clones using the Sanger sequencing. The median substitution mutations per readout for individuals A, B, and C were 17, 17.5, and 14.5, respectively (Fig. 1B). Moreover, when we combined all of the Sanger-sequenced clones from 10 mixed individuals and individuals A, B, and C, all 1-kb barcode sequences were distinct (Fig. S1C), providing further evidence for the extensive mutation barcoding diversity within our SMALT system. Among the 514 G/C sites on our 1-kb barcode, we identified mutations at 223 sites in at least one clone, which indicates that mutable sites are widely distributed in our 1-kb barcode (Table S1). The barcoding mutation peaks occur around 20bp downstream of the center of iScel binding motifs (Fig. S1B). Around 99% of recorded mutations are CG-to-TA, consistent with the cytidine deaminase activity of AID (Fig. 1C). Since our 1-kb barcode exists as a single copy in the zygotic genome, a 1-kb barcode from a single clone can roughly represent an individual cell. Employing the 1-kb barcode as alleles, we reconstructed maximum-likelihood cell phylogenetic trees for each of the three fishes (Figs. 1D and S2). The wild-type reference sequence served as the root, with each allele representing an individual cell in these trees. All three reconstructed trees exhibit a dichotomous structure, which is optimal for phylogenetic inference. Despite the limited sample size for each tree, we still observed overall high bootstrap values supporting our internal nodes. The median SH-aLRT bootstrap values for the three trees were 77%, 78%, and 75%, respectively. Taken together, these proof-of-principle results suggest that the SMALT system can be employed for lineage tracing in zebrafish during embryonic development.

### Testing SMALT barcoding performance in adult fin

To investigate if the barcoding mutation edited during embryogenesis could be recovered in adult zebrafish, we raised the injected transgenic fish to adulthood (Fig. 2A). We amputated all eight parts of the zebrafish fins, including the unpaired caudal fin (ventral and dorsal side separately), anal fin, dorsal fin, as well as the paired fins pectoral and pelvic (left and right separately). For individuals ZF1 and ZF4, we amputated the original fins and the fully regenerated fins two times (at time intervals of 20 days) (Fig. 2A). For individual ZF2, we amputated the original fins and the fins regenerated once. Individual ZF3 had only the original fins amputated. After genome extraction, we PCR amplified the 1-kb barcode using primers containing unique molecular identifiers (UMIs) to distinguish different DNA molecules (Fig. 2A). We used single-molecule real-time (SMRT) sequencing on the PacBio Sequel II system to sequence the PCR amplicon. Using the high-quality HiFi reads, we recovered 4028 and 4030 1-kb barcode with distinct UMIs for ZF1 and ZF4. For ZF2 and ZF3, we obtained 1903 and 772 1-kb barcode, due to the reduced sampling times (Fig. S3A). The median substitution mutations per readout for ZF1 to ZF4 were 17, 15, 16, and 16, respectively (Fig. S3A). The median mutation number per 1-kb barcode varies from 9 to 26 across different fins (Fig. 2B). Moreover, all the recovered 1-kb barcodes

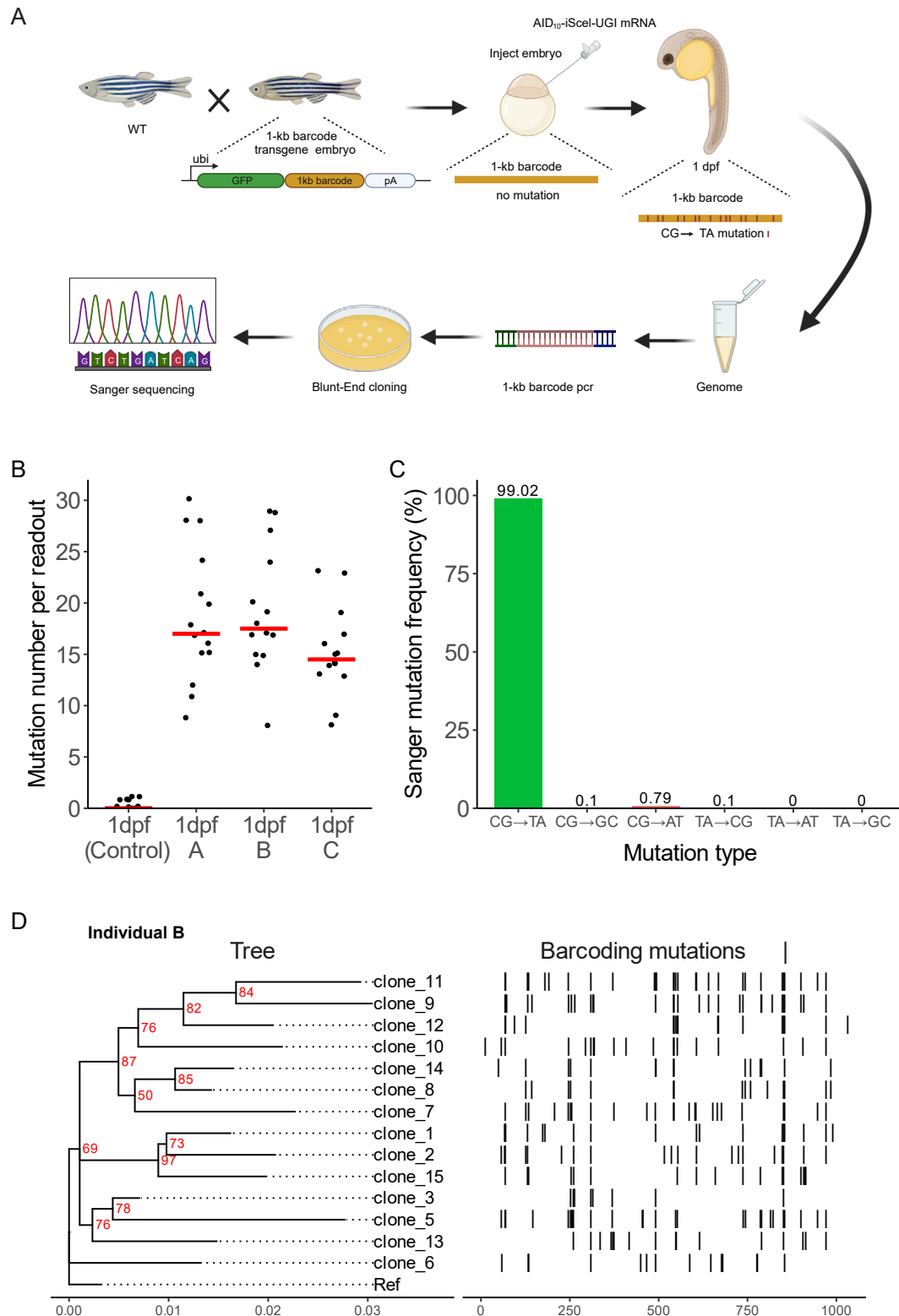


**Fig. 1.** Applying SMALT barcoding system in zebrafish embryos. **A:** A schematic diagram showing the generation of transgenic zebrafish with 1-kb barcode for SMALT barcoding system. The stable 1-kb barcode transgenic line was established using the Tol2 transposase system. Mutations were introduced by injection SMALT AID10–iScel–UGI mRNA to the single copy 1-kb heterozygotes embryo. The mutations were detected through PCR using the extracted embryo genome as a template, followed by blunt-end cloning and Sanger sequencing. **B:** The number of mutations per 1-kb barcode. Left are the 1 dpf uninjected control embryos, and the right three are AID10–iScel–UGI mRNA-injected 1 dpf embryos. Each dot represents a 1-kb barcode readout mutation number, and red horizontal line indicates the median mutation number for that group. **C:** The bar plots showing the relative frequency of different types of substitution mutations detected by Sanger sequencing. The number on the top of each bar represents the frequency of this particular type of mutation. **D:** Left, a maximum-likelihood cell phylogenetic tree for AID10–iScel–UGI mRNA-injected individual B using the Sanger-sequenced clones as representative of cells. The tips of the trees are

exhibited mutations, with ZF1 recording a minimum of one mutation, while ZF2 to ZF4 had at least three recorded mutations (Fig. S3A). Besides, CG-to-TA mutations constitute 99.1% of all mutations, closely aligning with the Sanger sequencing result of 99% (Fig. 2C). The barcoding mutations are pervasive across the 1-kb sequence, with peaks occurring around 20bp downstream of the center of iScel binding motifs, which is similar to our Sanger clone data (Fig. S3B). For each individual, out of the 514 G/C sites on our 1-kb barcode, we detected mutations on 321, 262, 242, and 309 sites for ZF1 to ZF4. When combining mutations from all four fish, we detected mutations on 437 sites, with 374 sites exhibiting at least two mutations (Table S2). The per-site mutation rate was highly consistent between our Sanger and PacBio data with a Pearson correlation coefficient of 0.95 (Fig. S3C), supporting the high quality of our PacBio data and UMI mutation calling strategy. The per-site mutation frequency has a wide range from  $2^{-12}$  to  $2^0$ . The per-site mutation frequency observed in our study exhibits a wide range, spanning from  $2^{-12}$  to  $2^0$ . This frequency is decided by the number of cell divisions during which the mutations are generated. For instance, mutations arising from the first cell division would be inherited by half of the subsequent cells, thus having a mutation frequency of  $2^{-1}$ . Similarly, mutations originating from subsequent cell divisions would be inherited by progressively smaller fractions of cells and smaller mutation frequency. Consequently, our analysis demonstrates that the barcoding mutations cover all the mitotic divisions from the zygote to the number of recovered cells (Figs. 2D, S3D–S3F).

### Cell phylogeny in adult fin

Utilizing the aforementioned PacBio sequencing data, we reconstructed maximum-likelihood phylogenetic trees using IQ-TREE (Nguyen et al., 2015; Minh et al., 2020) for each of the four fish, using the wild-type sequence as the root (Fig. 3A). Considering the proximity of the median mutation numbers in the adult fin and embryo and the degradative nature of the injected mRNA, we anticipated that the majority of mutation accumulation on our 1-kb barcode occurs during embryonic development. Indeed, the edge length near the terminal cells is quite small, which indicates less mutation recorded. Based on the edge length count distribution (Fig. S4A), we merged near-zero internal branches (branch length less than  $1e-4$ ), which indicates polytomy. After merging short edges, the median UFBoot support values for ZF1 to ZF4 were all 99% (Fig. 3B). From the documentation of IQ-TREE, typically, one would start to rely on the clade if its UFBoot  $\geq 95\%$ . Thus, we obtained four phylogenetic trees with high bootstrap support. The tree for ZF1 to ZF4 each comprised 234, 160, 138, and 258 internal nodes (Fig. 3B), indicating the recording of hundreds of mitotic cell divisions. A detailed examination of the four cell phylogenetic trees revealed that most polytomy nodes contained cells from the same part of the fin in origin and regeneration (Fig. 3A). This suggests that, after amputation, regenerated cells mainly originate from the cells of the same part of the fins.

By utilizing the distance from the most recent common ancestor (MRCA) internal node to the root as node heights, representing the division history of the cells under this node, we calculated all pairwise distances between every cell in the phylogenetic trees. Phylogenetically more related cells shared a longer lineage history from zygote to their MRCA internal node, thus longer node heights than distant cells. After grouping the node heights based on the cells sampled from each fin part, we selected the top 20 percent quantile to represent closely related cells on the phylogeny. We obtained a matrix representing all fins by combining the distance matrices from

four individuals. The heatmap of this matrix shows that for all eight parts of the fin, the regenerated fin and the original fin of the same part are phylogenetic close to each other (Fig. 3C). The heatmap from different sampling proportions showed consistent results (Fig. S4B and S4C). The box plots illustrating node heights from separate sampling parts further emphasize that cells from the same part, whether original or regenerated, share a longer lineage history than cells from different parts (Fig. S5). In conclusion, our findings indicate that for all parts of the fin, regenerated cells mainly originate from cells within the same part of the fin.

### Mutation recording and clone diversity in germ cells

To test the performance of our barcoding system in germ cells, we conducted PCR on the progeny of both ZF1 and ZF4. Given that adult ZF1 and ZF4 were raised from injected embryos, the AID10–iScel–UGI mRNA would undergo degradation before sexual maturity. We anticipated no additional mutation accumulation on the 1-kb barcode in the subsequent generation of embryonic development. Consequently, we speculated that the mutations observed in the 1-kb barcode of their offspring could be indicative of germ cell mutations. We conducted PCR amplification on a total of 94 and 92 offsprings separately obtained from ZF1 and ZF4, respectively, spanning three distinct spawnings on different days. The median mutation numbers were 10 and 14.5 for ZF1 and ZF4 (Fig. 4A). The germ cell mutation number is lower than the fin cell mutation number from the same individual (Fig. 4A). Indeed, previous studies using *vasa* RNA expression as a molecular marker showed that the primordial germ cells (PGCs) number remained constant during cleavage stage and did not change greatly after gastrulation to the first day of development (Yoon et al., 1997; Knaut et al., 2000; Raz, 2003; Tzung et al., 2015). The lower mitotic division of PGC during embryogenesis may contribute to the lower number of barcoding mutations.

Next, we analyzed the clonality of the germ cells from separate spawnings to figure out whether gametes in each spawning originated from the same group of germline stem cells. We detected five different clones, each having at least two cells with the same 1-kb readout in ZF1. Five out of five clones were used in the first spawning, with the second and third spawning having four clones each (Fig. 4B). For ZF4, six clones were detected, with five, six, and six clones in each spawning, respectively. This observation suggests that different sets of gametogenesis originate from the same pool of germline stem cells. This is coherent with the previous study using golden progeny fraction from mutated founder fish (Walker and Streisinger, 1983).

### Cell phylogeny confirms the early separation of germ cells from somatic cells

To further investigate the separation between progenitors of somatic cells and germ cells, we incorporated these germ cells along with those first sampled original fin cells to build two lineage trees. After merging near-zero internal branches, in the case of ZF1, our reconstructed tree comprises 769 cells and 139 internal nodes. For ZF4, the tree encompasses 1126 cells and 168 internal nodes (Fig. 4C). Both trees have a median UFBoot support of 99% (Fig. 4D). In ZF1, all 94 sampled germ cells were grouped under one internal node with 100% UFBoot support values, forming a large monophyletic clade (Fig. S6A). In ZF4, 88 out of the 92 sampled germ cells form a large monophyletic clade, while the remaining 4 cells constitute a separate, smaller monophyletic clade. These two monophyletic clades in ZF4 both have 99% UFBoot supporting the separation of germ cells and fin cells (Fig. S6A). These data imply a

different Sanger-sequenced clones, and the red number represents the SH-aLRT bootstrap value support for the internal node. Right, the corresponding barcoding mutations on the 1-kb barcode for each clone. Mutations are represented as short vertical lines along the 1-kb barcode to indicate their positions.

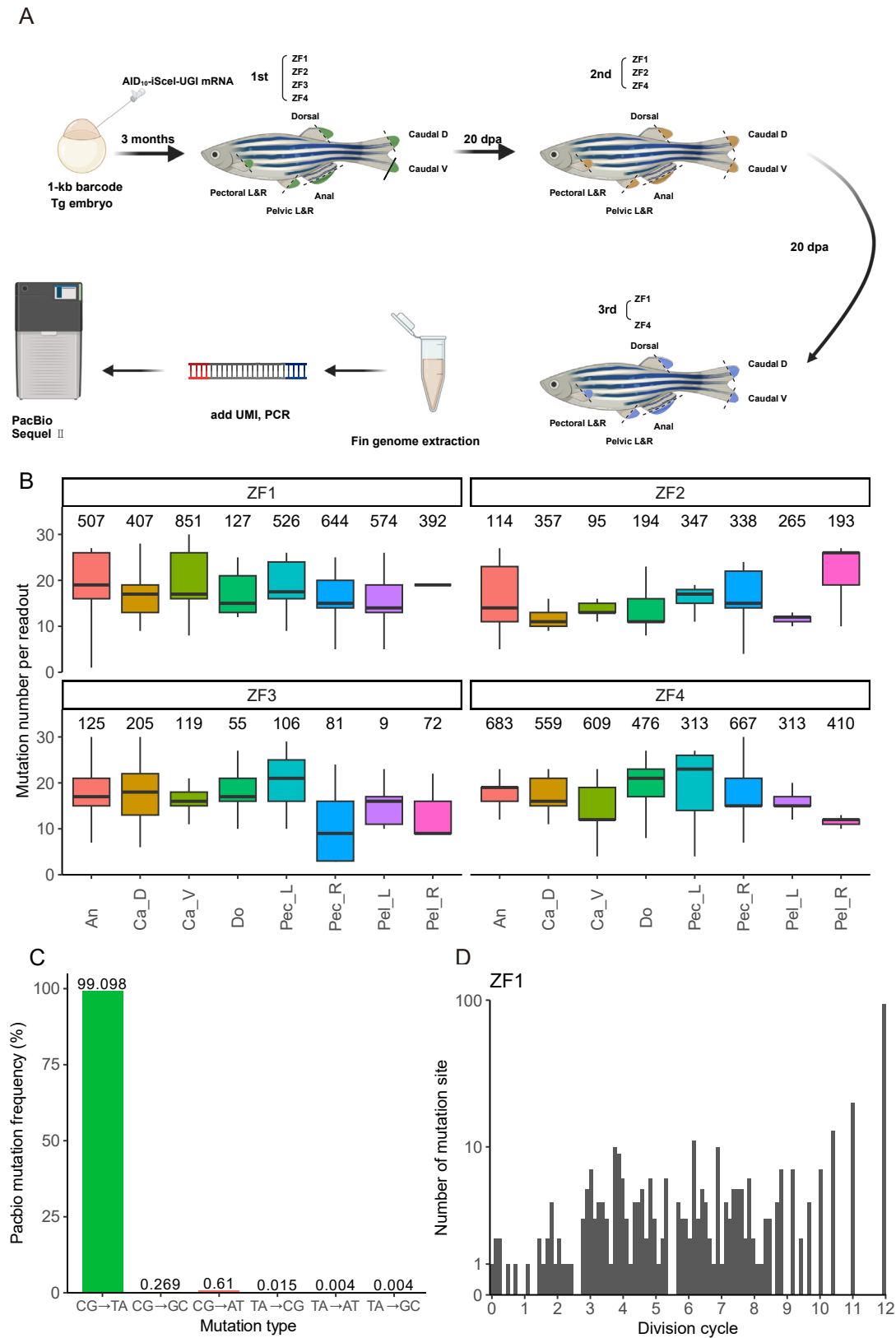


**Fig. 2.** SMALT barcoding performance in adult fin. **A:** A schematic diagram demonstrating the sampling of zebrafish fin cells from eight anatomic parts, each in four individuals, and the generation of UMI containing a 1-kb barcode library for SMRT sequencing. The genome of each fin part from every sampling was individually extracted, with green color representing the first, brown color representing the second, and purple color representing the third sampling. The extracted fin genome was used as PCR template and UMI was added on the 1-kb barcode. The UMI added barcode was further amplified to introduce sample and batch index. The cleaned 1-kb PCR library was sequenced on the PacBio Sequel II system. **B:** The box plots illustrating the per-allele mutation numbers for different fin parts in four zebrafish individuals. Each allele represents a high-quality readout sequence with a UMI, characterized by SMRT sequencing. Each color represents a different anatomical part of the fin. The number on the top of each box plot represents the number of alleles recovered for

relatively small progenitor cell number for germ cells and an early separation between the progenitor of germ cells and somatic cells in fins. In the monophyletic clades of germ cells in ZF1 and ZF4, germ cells from separate spawnings were mixed under each of the internal nodes, which suggests germ cells from separate spawnings shared the same group of progenitor cells (Fig. S6B).

To calculate the number of progenitors, we employed a recently developed targeting coalescent analysis (TarCA), which calculates the probability that two randomly sampled cells coalesce within a monophyletic clade for a given tissue (Deng et al., 2024). The inverse of the probability serves as a measure of the progenitor number of the tissue. We specifically focused on the monophyletic clades composed of germ cells. We utilized these fin somatic cells as background and used TarCAa to calculate the progenitor number of germ cells. The outcomes revealed a relatively small number of germ cell progenitors, with 1 and 1.09 for ZF1 and ZF4, respectively (see Materials and methods). These fate-determined progenitor cells for germ cells are characterized by their commitment exclusively to generating germ cells, with none of their progeny capable of differentiating into somatic cells. These germ cell progenitor numbers are estimations when the progenitor of germ cells and fin somatic cells are set apart. Previous studies demonstrated that *vasa* RNA is initially observed along the first two cleavage planes and subsequently condenses into four subcellular clumps by the 32-cell stage (Knauf et al., 2000; Yoon et al., 1997). The difference between the number of *vasa* RNA-containing cells and our phylogenetic progenitor estimation could be explained in two ways. Firstly, although the fin cells comprised a variety of cell types, including epithelial, mesenchymal, and hematopoietic populations (Hou et al., 2020), using fin cells as a proxy for all types of somatic cells is still underrepresented, which might lower the estimation of germ cell progenitors. Secondly, the marker gene-defined progenitors may be divided from the same ancestor in the phylogeny, suggesting the delayed cell state divergence during cell fate commitment of germ cells (Wagner and Klein, 2020).

## Discussion

In this study, we applied the SMALT system to zebrafish, achieving a highly efficient barcoding capacity for lineage tracing. Specifically, for the 1dpf embryo, we recorded a median of 14 substitution mutations per 1-kb barcode. This closely aligns with the estimated number of cell divisions and surpasses the mutation count achieved by the current Cas9-based lineage tracing method in zebrafish, which typically records around three mutations per barcode. Applying this system to zebrafish fin and germ cells, we reconstructed high-quality cell phylogenetic trees comprising hundreds of internal nodes with high bootstrap value support. These proof-of-principle experiments illustrated that our system can effectively record barcoding mutations in diverse tissues, rendering it applicable to various fields in developmental biology. The high mutation recording capacity will enable the generation of detailed cell lineage relationships, enhancing a comprehensive understanding of development and disease. Notably, this barcoding system uses injected AID10–iScel–UGI mRNA to create mutations, and its efficacy could be further enhanced by incorporating state-of-the-art mRNA modification and circular RNA techniques.

The zebrafish cell lineage trees obtained in our study provided clear evidence for the early separation of germ cells and somatic cells progenitor. While prior investigations using molecular marker

expression have identified several germ-cell progenitors during early embryogenesis, our approach provides a detailed illustration from a phylogenetic perspective. The estimation of germ cell progenitors using targeting coalescent analysis introduces a novel method, complementing the conventional observation of *vasa* RNA-containing cells. To improve the precision of estimating the number of germ cell progenitors, it may be necessary to include additional cell types beyond those found in fin cells, thereby creating a broadly sampled population of somatic cells as background.

One of the main limitations of our experimental design is that we relied solely on anatomical position information for annotating our lineage tree. This approach overlooks the presence of diverse cell types within the same fin region, each characterized by unique molecular states. Recent advancements in single-cell genomics have enabled the detection of mutations from barcode mRNA, providing a means to simultaneously acquire the molecular state and lineage history of individual cells (Raj et al., 2018; Spanjaard et al., 2018; Bowling et al., 2020; Li et al., 2023). Our transgenic line, integrated as the 3' UTR of GFP, is expected to transcribe 1-kb barcode mRNA with the poly-A structure. Consequently, it could be incorporated into current scRNA-seq workflows that utilize oligo-dT for capturing mRNA. This adaptation will allow the incorporation of cell type annotations into the cell lineage tree, providing a comprehensive understanding of cell fate commitments during the development.

## Materials and methods

### Plasmids

The 1-kb DNA barcode HMF10 was designed based on previous mutation readout data from flies and yeast. The 1-kb represents a 1-kb DNA barcode containing 16 *iScel* binding motifs surrounded by AID-preferred motifs. Unlike the original 3-kb readout used in fly, the new 1-kb DNA barcode has only 80 bp of DNA sequence between the heads of two opposite *iScel* binding motifs and 4 bp of DNA sequence between the backs of *iScel* binding motifs. The 1-kb DNA barcode was synthesized commercially through GENEWIZ with BamHI and NotI restriction enzyme sites on two ends. This 1-kb DNA barcode was cloned into Tol2 vectors as 3' UTR sequences of *ubi::GFP*. The plasmid is available under request.

### Zebrafish transgenic line

All zebrafish were maintained at 28.5°C under a 14-h light/10-h dark cycle. Tol2 transposase mRNA was in vitro transcribed by the NEB E2060s kit and was purified by the NEB T2040S kit. Tol2 transposase mRNA and the Tol2\_ubi:GFP\_1-kb\_pA vector were mixed at a concentration of 12.5 ng/μL each. Then, 2 nL of the DNA/RNA mixture was injected into 1-cell stage AB wild-type embryos. F0 embryos with broad GFP fluorescence were selected and raised to adulthood. F0 adult was crossed with WT, and fluorescent F1 embryos with ubiquitous GFP expression were collected. F1 adult fin clipping is performed for genotyping. We used the fin genome for qPCR to calculate the relative ratio of GFP to ultra-conserved genomic sequence to determine the copy number of transgene insertion. Only the qPCR-identified single copy F1 that has approximately 50% GFP-positive progeny when outcrossed with WT were selected. F2 from single copy F1 was raised to adulthood as a stable transgenic line. All animal experiments were approved by the Animal

this part of the fin in this individual. An, anal fin; Ca\_D, caudal dorsal fin; Ca\_V, caudal ventral fin; Do, dorsal fin; Pec\_L, pectoral left fin; Pec\_R, pectoral right fin; Pel\_L, pelvic left fin; Pel\_R, pelvic right fin. **C:** The relative frequency of different types of substitution mutations detected by SMRT sequencing. The number on the top of each bar represents the frequency of this particular type of mutation. **D:** The estimated number of mutation sites created during each of the division cycles in ZF1. The y axis was log<sub>10</sub> normalized. The division cycle of the mutated site was calculated as the log<sub>2</sub> of the site mutation frequency. The site mutation frequency was calculated by dividing the number of cells with mutations on this particular site by the total number of cells detected at that site.

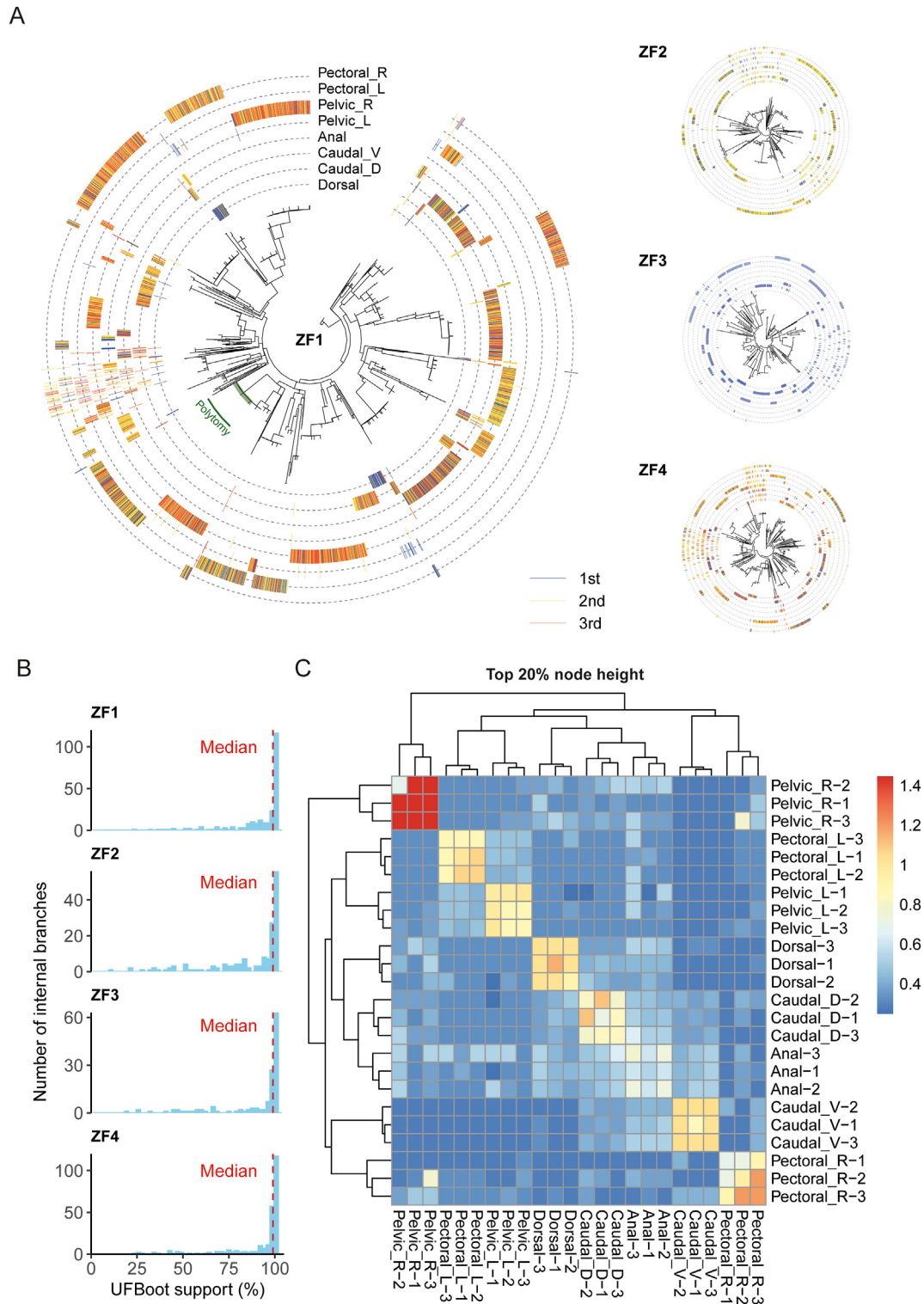


**Fig. 3.** Building high-quality cell lineage trees for zebrafish fin. **A:** The dendrogram representing four zebrafish fin cell lineage trees. The trees include 4028 alleles for ZF1, 1903 alleles for ZF2, 772 alleles for ZF3, and 4030 alleles for ZF4. The left tree represents ZF1, while the three trees from top to bottom on the right side represent ZF2, ZF3, and ZF4, respectively. The outer eight layers represent eight different anatomical parts of the fin, color-coded to distinguish between original and regenerated sampled cells. The original sampled fin cells are denoted in blue, the second-time sampled cells in yellow, and the third-time sampled cells in red. In the ZF1 tree, a representative polytomy containing cells from the same part of the fin in both the original and regenerated states is highlighted in dark green. **B:** The accompanying density plot showing the frequency distribution of UFBoot support values on the internal branches of the four trees. The four zebrafish cell phylogenetic trees all have 99% median UFBoot support, denoted by the red vertical dashed line. **C:** The heatmap showing the similarity of original cells and regenerated cells for each part of the fin based on shared phylogenetic distance. A higher value denoted as red represents cells from this part are phylogenetically closer to each other, and a lower value denoted as blue represents cells from this part are phylogenetically farther.

Research and Ethics Committees of Sun Yat-Sen University and performed under the National Guideline for the Care and Use of Laboratory Animals (China).

### Mutation creation

The AID10–iScel–UGI from the SMALT system was PCR cloned into the pCS2 vector. AID10–iScel–UGI mRNA was in vitro transcribed. The single copy insertion homozygotes *Tg(ubi:GFP\_1-kb\_pA)* transgenic fish were outcrossed with WT, and 300–pg mRNA was injected into 1-cell stage embryos. Transgenic larval with normal development and ubiquitous GFP expression were raised to adult. The 1-kb was PCR amplified from fin genomes and cloned by blunt end cloning (Clone Smarter C5851). The mutations on the 1-kb were calculated from the Sanger sequencing data.

### Fin 1-kb barcode library preparation for SMRT sequencing

Four adult zebrafish were anesthetized by a standard dose (164 mg/L) of tricaine. We amputated eight parts of the fins for each fish as illustrated in Fig. 2A. After amputation, the adult fish were immediately transferred to the fish tank. Each part of the fin was collected in a 1.5-mL microtube filled with 180  $\mu$ L of genome extraction buffer ATL from QIAamp DNA Micro Kit (QIAGEN). The genomic DNA was extracted following the ‘isolation of genomic DNA from tissues’ protocol and quantified with Qubit v.3.0 Fluorometer (ThermoFisher).

We used 30-ng genome from each part of the fin from each sampling, about 6500 copies of the zebrafish genome, as a PCR template. We implemented a three-step strategy utilizing five primers (P1–P5). All PCR reactions used KOD One™ PCR Master Mix (TOYOBO).

In the initial step, we performed a single cycle of linear amplification utilizing the P1 primer, which incorporates 16-nucleotide UMIs to label the initial readout molecules. Then, we cleaned up the product using 0.6 $\times$  SPRIselect beads’ left side size selection. Next, we performed 15 cycles of nested PCR utilizing primers P2 (which includes a 10-nucleotide fin sample index) and P3 to amplify and enrich the readout molecules containing UMIs. Subsequently, the resulting products were purified using a cycle-pure kit (Omega Bio-tek d6492-02). Finally, purified products from the Step 2 were used as templates to be amplified using P4 (containing a 10-nucleotide batch sample index) and P5. Detailed PCR primer sequences are listed in Table S3. The detailed thermal conditions are as follows:

Step 1: 0.3  $\mu$ M P1 in 100  $\mu$ L of solution; thermal condition: 94°C/3 min (98°C/10 s, 62°C/1 min, 68°C/1 min)  $\times$  1 cycle, 68°C/10 min followed by 0.6 $\times$  bead purification;

Step 2: 0.3  $\mu$ M P2 and P3 each in 100  $\mu$ L of solution; thermal condition: 94°C/3 min (98°C/10 s, 57°C/5 s, 68°C/20 s)  $\times$  15 cycles, 68°C/2 min and mixed samples from the same sampling batches and cycle pure the DNA;

Step 3: 0.3  $\mu$ M P4 and P5 each in 100  $\mu$ L of solution; thermal condition: 94°C/3 min (98°C/10 s, 57°C/5 s, 68°C/20 s)  $\times$  20 cycles, 68°C/2 min and mixed samples from the same sampling batches and then 0.6 $\times$  bead purification.

### Germ cell 1-kb barcode library preparation for SMRT sequencing

Female adults ZF1 and ZF4 were individually mated with wild-type AB adults at intervals of 1–2 weeks. For each pair of mating, all zygotes were collected, raised to 3 dpf, and all larvae displaying GFP fluorescence were selected. These larvae were individually lysed in a 96-well PCR plate using 50  $\mu$ L of 50 mM NaOH at 95°C for 5 min for genome extraction. Subsequently, 5  $\mu$ L of 1 M Tris-HCl, pH 8.0, was

added to neutralize the sample, and 2  $\mu$ L of the mixture was used as the PCR template. A two-step PCR strategy was employed using four primers (P6–P9). In the first step, we performed a 15-cycle PCR with dual index (P6: Partial read1-l5\_cell\_index-1-kb F; P7: Partial read2-l7\_cell\_index-1-kb R). All PCR reactions on larvae from the same individual’s mating on the same day were pooled and purified by cycle-pure. In the second step, PCR products from the previous step were used as templates and the batch index was added by a 15-cycle PCR using P8 and P9 primers. Detailed PCR primer sequences are listed in Table S3.

### SMRT sequencing and data processing

One microgram of the PCR 1-kb readout sequence containing UMIs and fin-specific index was utilized for SMRTbell library preparation (PacBio Sequel II), following the standard SMRT sequencing protocol. Sequencing runs were conducted using one SMRT cell for each SMRTbell library (Xi’an Haorui Genomics Technologies Ltd.). For each subread group, circular consensus sequences (CCS) were generated by pbccs v.6.4.0 with the parameters: ccs –min-length = 800 –max-length = 1300. After this step, ccs reads were aligned to the 1-kb reference sequence using minimap2 under the parameter setting: -ax -am20. After alignment, reads that were correctly mapped were selected using samtools as samtools view -b -F 4. The sampling batch index was extracted using customized R scripts, which used the P7 PCR adaptor sequence as an anchor to extract the following 10 nucleotides. The CCS reads were then grouped by batch. Then, we extracted the fin sample index and cell UMI from CCS reads using read1 as the anchor sequence. Using these indexes, we assigned CCS reads to different parts of the amputated fin. The UMI was used to separate CCS reads to different cells. Each UMI that has less than 2 CCS reads was excluded for further analysis. Mutation calling was performed using bcftools, and only substitution mutations were kept. For each mutation site, we only kept the site that keeps the same for all CCS reads or at least a 3 to 1 ccs reads ratio of the mutation base to the reference base. After filtering, for each UMI, a consensus readout sequence was obtained as an allele. In the end, we got 4028, 1903, 772, and 4030 alleles for ZF1, ZF2, ZF3, and ZF4 fin. For germ cells, we got 94 and 92 alleles for ZF1 and ZF4.

### Reconstruction of fin and germ cell phylogeny

We utilized a maximum-likelihood-based software, IQ-TREE, to reconstruct the cell trees. We employed the -m MFP parameter in IQ-TREE, directing it to execute ModelFinder. The ModelFinder calculates log-likelihoods for an initial parsimony tree across multiple models and utilizes the Bayesian information criterion (BIC). Subsequently, it selects the model that minimizes the BIC score. For every reconstruction, we performed 1000 rounds of ultrafast bootstrap approximation (-bb 1000) and 1000 rounds of SH-like approximate likelihood ratio test (-alrt 1000) to assess the quality of consensus tree. We used the R package ggtree v3.10.0 for tree plotting (Xu et al., 2021; Yu et al., 2017, 2018).

### Fin distance computation based on phylogeny

We used the get\_mrca\_of\_set (tree, descendants) function from the R package castor v1.7.11 for each pair of terminal cells on the tree to get the most recent common ancestor internal node. For each internal node, the nodeheight (tree, node) function from R package phytools v2.0.3 was utilized to compute the height above the root for that internal node, and this height represents the distance of this pair of terminal cells. Then, the distance was normalized by dividing the median node height of all tips. We combined the distance from all

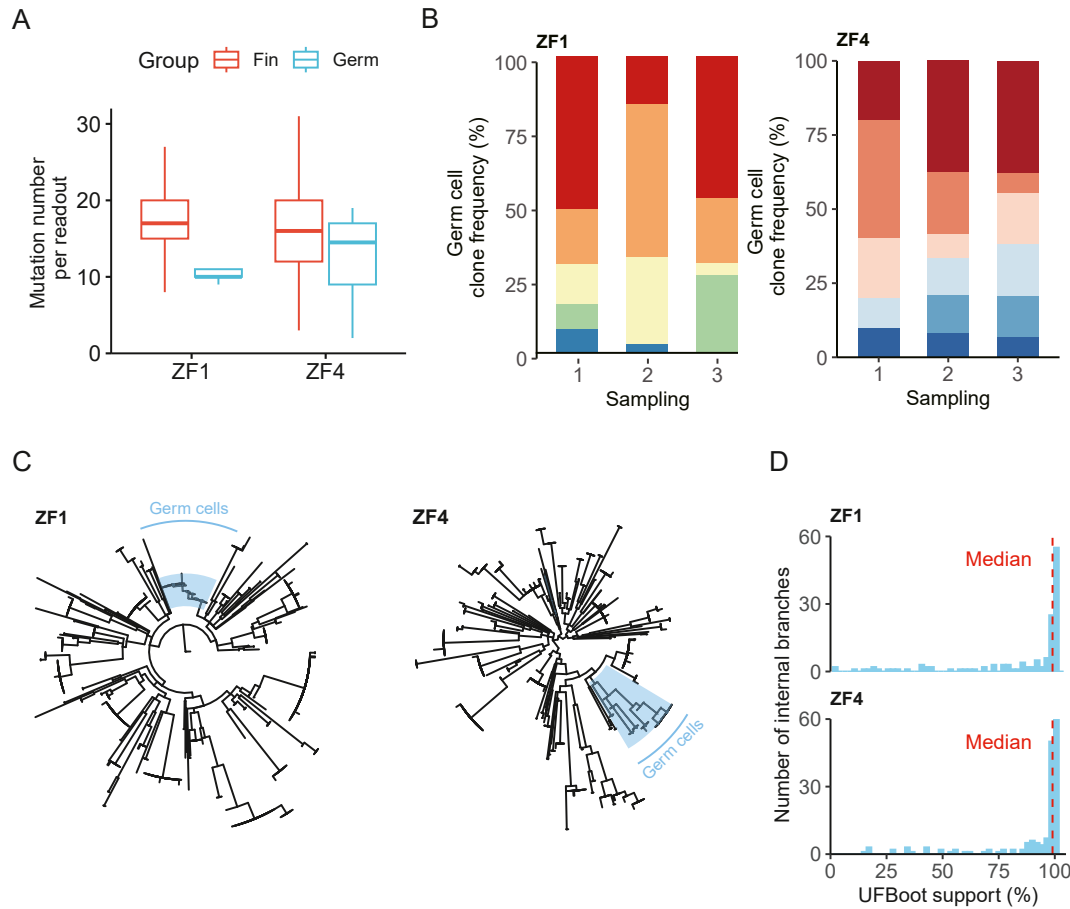


**Fig. 4.** Clonality and cell phylogeny of the germ cell. **A:** The box plots showing the comparison of germ cell and fin cell barcoding mutation numbers from the same individual. The red box represents the mutation number in fin cells, and the blue box represents the mutation number in germ cells. **B:** The clonal composition of germ cells from ZF1 and ZF4 across three separate spawning events. Each color represents a clone with a unique combination of barcoding mutations, and the height of the bar indicates the proportion of cells belonging to that specific clone. **C:** The cell phylogenetic tree of germ and fin cells for ZF1 and ZF4. The light blue clade highlight represents the clade full of germ cells, while other fin cells in these trees are annotated with black lines. **D:** The density plot showing the distribution of UFBoot support values on the internal branches of the ZF1 and ZF4 germ cell and fin cell trees. The red dotted line indicates the median UFBoot support of 99%.

four fish trees. Subsequently, we grouped all the pair-wise cell-to-cell distances based on the specific part and sampled times of the fin to which they belonged and computed the median distance for each part of the fin. We used the reciprocal of this distance matrix for hierarchical clustering using UPGMA. Finally, we used the R package gg dendro v0.1.23 to plot the dendrogram.

#### Germ cell progenitor number estimation

For ZF1 and ZF4, the trees that incorporated all the originated fin cells and germ cells were used. All three times of sampled germ cells were annotated as germ cells and all eight parts of the fin cells were annotated as fin cells. Then, we used the  $N_p$  Estimator function from TarCA and the tree and annotation table as input to estimate the number of progenitor cells of germ cells. The details of the calculation are as follows. For ZF1, the coalescent rate:  $Pr = C_{94}^2 / C_{94}^2 = 1$ , thus  $N_p = 1/Pr = 1$ . For ZF4, the coalescent rate:  $Pr = (C_{88}^2 + C_4^2) / C_{92}^2$ , thus  $N_p = 1/Pr = 1.09$ .

#### Data availability

The authors confirm that all experimental data are available and accessible via the main text and the Extended Data. Plasmids and

transgenic zebrafish line used in this study are available upon request. The raw PacBio sequence data reported in this paper have been deposited in the Genome Sequence Archive in the National Genomics Data Center under accession 1082510.

#### CRedit authorship contribution statement

**Zhan Liu:** Conceptualization, Data curation, Formal analysis, Writing - Original draft, Writing - Review & Editing. **Hui Zeng:** Data curation. **Huimin Xiang:** Data curation. **Shanjuan Deng:** Conceptualization, Formal analysis, Funding acquisition. **Xionglei He:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing - Review & Editing.

#### Conflict of interest

The authors declare that they have no conflict of interest.

#### Acknowledgments

This work was supported by the National Key R&D Program of China (2021YFA1302500 and 2021YFA1302501) and the National Natural Science Foundation of China (32293190, 32293191, 31970570, and 32200492).

## Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jgg.2024.04.004>.

## References

- Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J., Van Oudenaarden, A., 2018. Whole-organism clone tracing using single-cell sequencing. *Nature* 556, 108–112.
- Barker, N., Van Es, J.H., Kuipers, J., Kujala, P., Van Den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P.J., et al., 2007. Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* 449, 1003–1007.
- Blum, N., Begemann, G., 2015. Osteoblast de- and redifferentiation are controlled by a dynamic response to retinoic acid during zebrafish fin regeneration. *Development* 142, 2894–2903.
- Bowling, S., Sritharan, D., Osorio, F.G., Nguyen, M., Cheung, P., Rodriguez-Fraticelli, A., Patel, S., Yuan, W.C., Fujiwara, Y., Li, B.E., et al., 2020. An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. *Cell* 181, 1410–1422.
- Cole, L.K., Ross, L.S., 2001. Apoptosis in the developing zebrafish embryo. *Dev. Biol.* 240, 123–142.
- Deng, S., Gong, H., Zhang, D., Zhang, M., He, X., 2024. A statistical method for quantifying progenitor cells reveals incipient cell fate commitments. *Nat. Methods* 21, 597–608.
- Hou, Y., Lee, H.J., Chen, Y., Ge, J., Osman, F.O.I., McAdow, A.R., Mokalled, M.H., Johnson, S.L., Zhao, G., Wang, T., 2020. Cellular diversity of the regenerating caudal fin. *Sci. Adv.* 6, eaba2084.
- Kawakami, K., 2007. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol.* 8, 1–10.
- Knaut, H., Pelegri, F., Bohmann, K., Schwarz, H., Nüsslein-Volhard, C., 2000. Zebrafish *vasa* RNA but not its protein is a component of the germ plasm and segregates asymmetrically before germline specification. *J. Cell Biol.* 149, 875–888.
- Kretzschmar, K., Watt, F.M., 2012. Lineage tracing. *Cell* 148, 33–45.
- Kwan, K.M., Fujimoto, E., Grabher, C., Mangum, B.D., Hardy, M.E., Campbell, D.S., Parant, J.M., Yost, H.J., Kanki, J.P., Chien, C.B., 2007. The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs. *Dev. Dyn.* 236, 3088–3099.
- Lee, Y., Grill, S., Sanchez, A., Murphy-Ryan, M., Poss, K.D., 2005. Fgf signaling instructs position-dependent growth rate during zebrafish fin regeneration. *Development* 132, 5173–5183.
- Li, L., Bowling, S., McGeary, S.E., Yu, Q., Lemke, B., Alcedo, K., Jia, Y., Liu, X., Ferreira, M., Klein, A.M., et al., 2023. A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* 186, 5183–5199.
- Liu, K., Deng, S., Ye, C., Yao, Z., Wang, J., Gong, H., Liu, L., He, X., 2021. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat. Methods* 18, 1506–1514.
- Lu, R., Neff, N.F., Quake, S.R., Weissman, I.L., 2011. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* 29, 928–933.
- Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., Teeling, E., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534.
- Mosimann, C., Kaufman, C.K., Li, P., Pugach, E.K., Tamplin, O.J., Zon, L.I., 2011. Ubiquitous transgene expression and Cre-based recombination driven by the *ubiquitin* promoter in zebrafish. *Development* 138, 169–177.
- Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- Pei, W., Feyerabend, T.B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., et al., 2017. *Polylox* barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* 548, 456–460.
- Petrie, T.A., Strand, N.S., Tsung-Yang, C., Rabinowitz, J.S., Moon, R.T., 2014. Macrophages modulate adult zebrafish tail fin regeneration. *Development* 141, 2581–2591.
- Pfefferli, C., Jazwińska, A., 2015. The art of fin regeneration in zebrafish. *Regeneration* 2, 72–83.
- Poss, K.D., Shen, J., Nechiporuk, A., McMahon, G., Thisse, B., Thisse, C., Keating, M.T., 2000. Roles for Fgf signaling during zebrafish fin regeneration. *Dev. Biol.* 222, 347–358.
- Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., Schier, A.F., 2018. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36, 442–450.
- Raz, E., 2003. Primordial germ-cell development: the zebrafish perspective. *Nat. Rev. Genet.* 4, 690–700.
- Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., Camargo, F.D., 2018. Clonal analysis of lineage fate in native haematopoiesis. *Nature* 553, 212–216.
- Salipante, S.J., Horwitz, M.S., 2006. Phylogenetic fate mapping. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5448–5453.
- Sehring, I.M., Weidinger, G., 2020. Recent advancements in understanding fin regeneration in zebrafish. *Wiley Interdiscip. Rev. Dev. Biol.* 9, e367.
- Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjua, S., Ninov, N., Junker, J.P., 2018. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* 36, 469–473.
- Sulston, J.E., Horvitz, H.R., 1977. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* 56, 110–156.
- Sulston, J.E., Schierenberg, E., White, J.G., Thomson, J.N., 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* 100, 64–119.
- Tzung, K.W., Goto, R., Saju, J.M., Sreenivasan, R., Saito, T., Arai, K., Yamahara, E., Hossain, M.S., Calvert, M.E.K., Orbán, L., 2015. Early depletion of primordial germ cells in zebrafish promotes testis formation. *Stem Cell Rep.* 4, 61–73.
- VanHorn, S., Morris, S.A., 2021. Next-generation lineage tracing and fate mapping to interrogate development. *Dev. Cell* 56, 7–21.
- Wagner, D.E., Klein, A.M., 2020. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* 21, 410–427.
- Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., Klein, A.M., 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981–987.
- Walker, C., Streisinger, G., 1983. Induction of mutations by  $\gamma$ -rays in pregonial germ cells of zebrafish embryos. *Genetics* 103, 125–136.
- Whitehead, G.G., Makino, S., Lien, C.L., Keating, M.T., 2005. *fgf20* is essential for initiating zebrafish fin regeneration. *Science* 310, 1957–1960.
- Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., Tang, W., Feng, T., Chen, M., Zhan, L., et al., 2021. ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* 38, 4039–4042.
- Yoon, C., Kawakami, K., Hopkins, N., 1997. Zebrafish *vasa* homologue RNA is localized to the cleavage planes of 2- and 4-cell-stage embryos and is expressed in the primordial germ cells. *Development* 124, 3157–3165.
- Yu, G., Lam, T.T.Y., Zhu, H., Guan, Y., 2018. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Mol. Biol. Evol.* 35, 3041–3043.
- Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.Y., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36.