



Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development

Kehui Liu^{1,2}, Shanjun Deng^{1,2}, Chang Ye^{1,2}, Zeqi Yao¹, Jianguo Wang¹, Han Gong¹, Li Liu¹ and Xionglei He¹

Mapping the cell phylogeny of a complex multicellular organism relies on somatic mutations accumulated from zygote to adult. Available cell barcoding methods can record about three mutations per barcode, enabling only low-resolution mapping of the cell phylogeny of complex organisms. Here we developed SMALT, a substitution mutation-aided lineage-tracing system that outperforms the available cell barcoding methods in mapping cell phylogeny. We applied SMALT to *Drosophila melanogaster* and obtained on average more than 20 mutations on a three-kilobase-pair barcoding sequence in early-adult cells. Using the barcoding mutations, we obtained high-quality cell phylogenetic trees, each comprising several thousand internal nodes with 84–93% median bootstrap support. The obtained cell phylogenies enabled a population genetic analysis that estimates the longitudinal dynamics of the number of actively dividing parental cells (Np) in each organ through development. The Np dynamics revealed the trajectory of cell births and provided insight into the balance of symmetric and asymmetric cell division.

Just as all species on earth are connected by a single phylogenetic tree, all cells in a multicellular organism can form a phylogenetic tree with leaves representing the cells, a root representing the zygote and internal nodes representing mitotic cell divisions¹. While the practice of mapping species phylogeny has become routine², the field of studying the cell phylogeny of complex organisms is in its infancy^{3,4}, because the millions to billions of cells in a typical complex organism pose a daunting technical challenge^{5,6}.

So far, the various methods developed to track the cell phylogeny of a complex organism, in which direct observation is often not feasible, have been based on the somatic mutations accumulated during mitotic cell divisions. Because of the extremely low rate of naturally occurring somatic mutations^{7,8}, a widely adopted strategy in model organisms is to engineer a genetic cassette that is designed to mutate at a very high rate^{9–21}. The cassette then serves as a barcode that records the mitotic division history of the cells. Despite rapid progress in the cell barcoding field in recent years^{22,23}, the achieved recording capacity of each barcode is on average about three mutations per barcode, and it therefore tracks at most three mitotic divisions. Because there are often tens of mitotic divisions for each cell in a fully developed complex organism during development, the cell phylogenies reconstructed by the available barcoding methods contain only limited information^{9,15,17,19}.

From a phylogenetic perspective, an informative cell phylogeny should comprise as many internal nodes as possible, because only internal nodes provide information about mitotic divisions and the related genealogy. Equally importantly, the informative (internal) nodes should be statistically supported (for example, with large bootstrap value)²⁴. To achieve this goal, the key parameter of a cell barcoding method is the number of recorded mutations per barcode, which should be comparable to the number of mitotic cell generations from zygote to adult (that is, the number of mitotic divisions an average cell in a fully developed adult has

experienced since the zygote stage). Although the number of mutations may vary among organisms and cell types, a minimum of 20 seems reasonable²⁵. In addition, the number of possible mutation profiles (or the barcoding space) should be larger than the number of cells in a complex organism (for example, roughly 10¹⁰ for a mouse²⁶). This is important for avoiding parallel acquisitions of the same mutation profiles.

Motivated by these considerations, we developed SMALT, a substitution mutation-aided lineage-tracing system. By applying SMALT to the fruit fly *D. melanogaster*, we observed on average roughly 21 substitution mutations per barcode in early-adult cells, with an effective barcoding space larger than 10³⁰. As a comparison, for the state-of-the-art cell barcoding systems there are typically around three mutations per barcode (Supplementary Table 1). Using the barcoding mutations, we obtained nearly single-cell-resolution cell phylogenetic trees, each comprising several thousand informative (internal) nodes supported by 84–93% median bootstrap values in two fly individuals. We then designed a phylodynamic analysis²⁷ and successfully reconstructed the developmental cell population dynamics in more than 12 fly organs.

Inventing SMALT in yeast

We started by searching for a protein that can mutate DNA. The cytidine deaminase gene family is known for the ability to induce C-to-T mutations²⁸. Specifically, the enzymes can remove the amino group of a cytosine to produce a uracil that will be recognized as a thymine in the next round of DNA replication, resulting in a C-to-T transition. We tested eight representative members of the enzyme family from different species by overexpressing them in the yeast *Saccharomyces cerevisiae* using conventional yeast genetic approaches (Fig. 1a)^{29,30}. We assessed the occurrences of L-canavanine-resistant yeast clones, which represent inactivation mutations on the yeast *CAN1* gene (Methods). The results indicate

¹MOE Key Laboratory of Gene Function and Regulation, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China. ²These authors contributed equally: Kehui Liu, Shanjun Deng, Chang Ye. [✉]e-mail: liul47@mail.sysu.edu.cn; hexiongli@mail.sysu.edu.cn

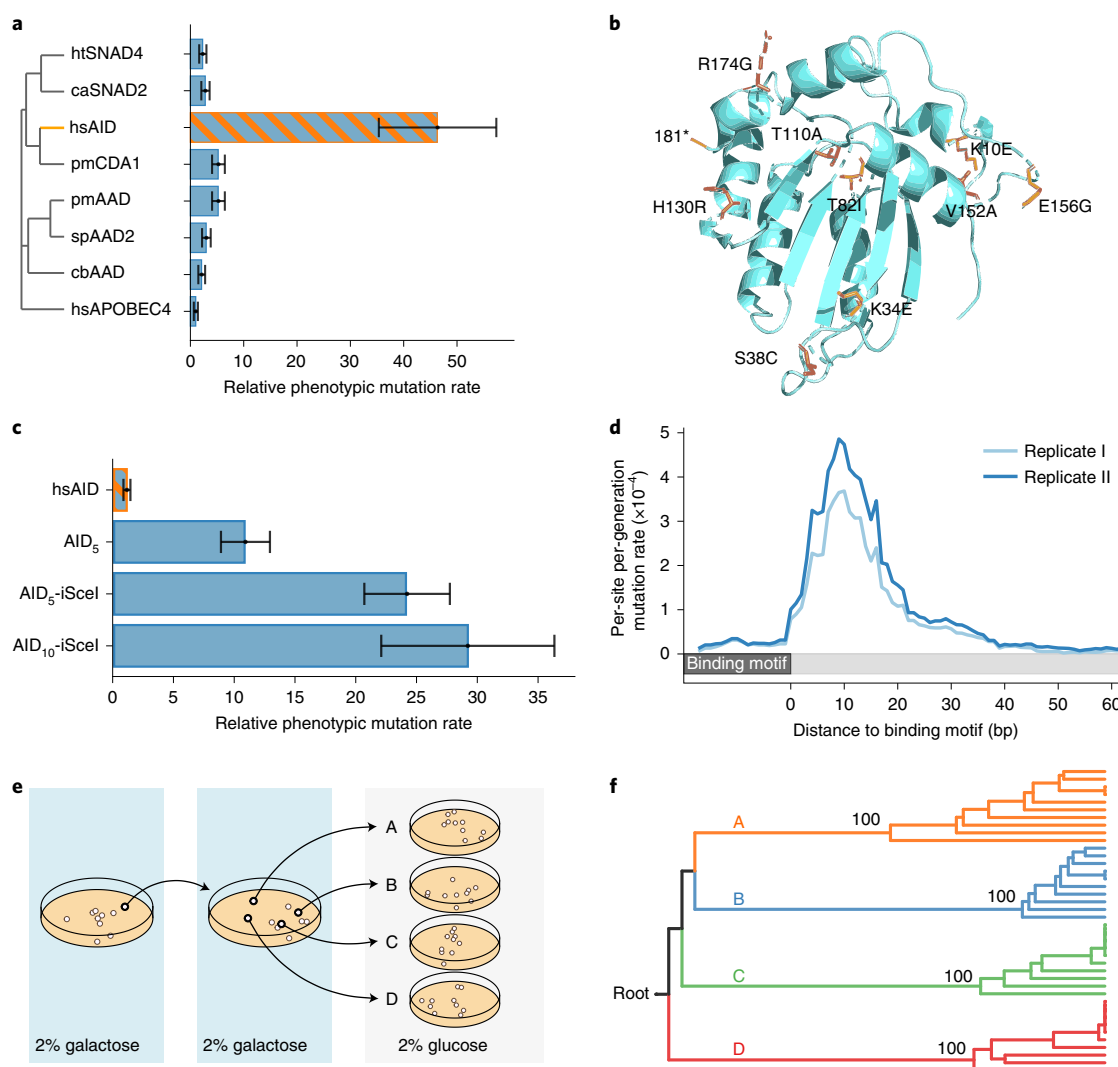


Fig. 1 | Invention of the SMALT cell barcoding system in yeast. a, Test of eight representative genes of the cytidine deaminase family for mutation rate. The rate of phenotypic mutations (L-canavanine resistance) depicted is relative to that induced by hsAPOBEC4. Error bars indicate 95% confidence intervals. **b**, Three-dimensional structure of the hsAID protein with the ten modifications used in the study. **c**, The phenotypic mutation rate relative to that induced by hsAID is shown, with error bars showing the 95% confidence intervals. **d**, Distribution of the per-CG site per-generation genetic mutation rate induced by AID₅-iSceI. For each site, the rate is the number of CG-to-TA and CG-to-GC mutations divided by the number of CG sites within a 15-bp window that covers the focal site and its downstream 14 sites. Two biological replicates are shown. **e**, A schematic diagram of the yeast evolution experiment. The SMALT system is active in plates supplied with 2% galactose but not in plates with glucose. **f**, A proof-of-concept demonstration in yeast that cell genealogy can be reconstructed using the barcoding mutations produced by SMALT. The unweighted pair group method with arithmetic mean (UPGMA) phylogenetic tree based on the barcoding mutations accurately represents the known genealogy of 40 yeast clones selected on the glucose plates, with 100% bootstrap supports on the four major clades.

that the *Homo sapiens* activation-induced cytidine deaminase (hsAID) is the most active enzyme, being able to boost the phenotypic mutation rate by at least roughly 45-fold (Fig. 1a).

We were determined to construct an hsAID-based mutagenesis system. Because in DNA the uracil resulting from cytosine deamination would be corrected back to cytosine by DNA repair systems³¹, the C-to-U/U-to-C reversible processes must produce an equilibrium of uracils in the genome as the source of C-to-T mutations. As equilibrium is not sensitive to the duration of the cell cycle, and the realization of C-to-T mutations from the C-to-U lesions relies on DNA replication, the hsAID-based mutagenesis system would be particularly suitable for tracking cell division events. To improve the performance of hsAID, we tested mutations in a variety of positions in the protein that may affect the enzyme activity based

on the three-dimensional structure according to previous studies³² (Fig. 1b). After extensive screenings we obtained a much more efficient variant of hsAID (AID₅) that differs from the wild type by five changes (K10E, K34E, T82I, E156G and a truncating mutation that removes the C-terminal 18 residues) (Fig. 1c).

To achieve targeting mutagenesis, the enzyme can be fused with a DNA-binding domain. We chose an inactive version of iSceI, a homing endonuclease that can specifically bind a DNA motif of 18 base pairs (bp)³³. AID₅ and iSceI were fused via a linker of 15 amino acids (3×GGGGS). An iSceI binding motif was also introduced into the 3' end of the *CAN1* gene in the yeast genome so that the AID₅-iSceI fusion protein expressed in the yeast could specifically mutate *CAN1*. We observed a roughly 25-fold increase in the phenotypic mutation rate compared to hsAID (Fig. 1c). We noted

that this number underestimates the power of AID₅-iSceI because it targets only the 3'-end of *CAN1*.

To find the full power of AID₅-iSceI we inserted a synthesized DNA sequence of roughly 400 bp with two divergent iSceI binding motifs at the middle into a functionally neutral locus of the yeast genome (Methods). We expressed AID₅-iSceI in the yeast for ten generations to accumulate mutations, amplified the roughly 400-bp sequence by polymerase chain reaction (PCR) and sequenced the amplicons by Illumina MiSeq sequencing. Mutations were primarily found within a window of roughly 50 bp away from the binding motifs (Fig. 1d). The rate of CG-to-TA and CG-to-GC mutations was roughly 3×10^{-4} per CG site per generation, about six orders of magnitude higher than the background mutation rate (2.2×10^{-10} per CG site per generation) of the yeast³⁴. We conducted further random mutagenesis on AID₅ using error-prone PCR to improve its efficiency. We obtained a few variants with slight improvements and chose one with five differences from AID₅ (S38C, T110A, H130R, V152A, R174G), which is termed AID₁₀ (Fig. 1c). We also attempted to optimize the binding domain but failed to identify one better than iSceI (Supplementary Fig. 1). Hence, AID₁₀-iSceI (hereafter referred to as AI) was used for further study.

We tested whether mutations induced by AI could be used to track cell genealogy in a yeast population. We used a 3-kb DNA cassette sequence designed to have 16 iSceI binding motifs surrounded by AID preferred motifs as the mutation readout (Methods). We engineered AI with the 3-kb readout sequence into a functionally neutral locus of the yeast genome. In an experimental evolution, a yeast colony starting from a single cell was redistributed to agar plates to form new single-cell-derived colonies, four of which, designated as lineages A, B, C and D, were selected and each redistributed again to new agar plates (Fig. 1e). The resulting single-cell clones were then analyzed. We found that mutations accumulated on the 3-kb readout sequence could reconstruct the known genealogy of 40 analyzed clones with no errors (Fig. 1f). We also found that the mutations depend heavily on the number of cell divisions (Extended Data Fig. 1). We termed the designed system (AI plus the readout sequence) SMALT.

Mapping the cell phylogeny of fruit fly with SMALT

To test the performance of SMALT in a complex organism, we engineered it for *D. melanogaster* (Fig. 2a and Supplementary Fig. 2). We used the (tubulin)-GAL4/UAS system to drive the expression of AI in the fly to ensure its ubiquitous activity across the whole body since early embryogenesis³⁵ (Extended Data Fig. 2). The SMALT transgenic flies are morphologically normal and fertile, and the effects of potential off-target mutations on the fly genome do not appear strong (Extended Data Fig. 3). We amplified the 3-kb readout sequence by PCR and sequenced the amplicons, after subcloning, by Sanger sequencing to identify mutations (Methods). We observed a continuous increase in substitution mutations on the 3-kb readout sequence from stages of embryo, larva and pupa, to adult (Fig. 2b), justifying the use of the mutations to track fly development. There are on average 21.3 substitution mutations per readout in early-adult cells, which is much higher than the typically around three mutations per barcode or readout for the state-of-the-art cell barcoding methods (Supplementary Table 1). Over 96% of the identified mutations are CG-to-TA or CG-to-GC, echoing the cytidine deaminase activity of AID. By considering the roughly 21 average mutations per readout and roughly 1,200 mutable sites on the 3-kb readout sequence, we estimated that the mutation space of SMALT achieved in fly is over 10^{35} (Supplementary Note 1), a number far larger than the cell or allele number of an individual fly³⁶, which is believed to be smaller than 10^8 . Assuming no cell deaths and that every cell divides to produce two daughter cells, there would be $\log_2 10^8$ equal to roughly 26 cell mitotic generations from zygote to adult. This is a lower-bound estimation, but the bias should not be too large since

the cell generation number is believed to be roughly 28–67, depending on tissue types, for a 10-week-old mouse²⁵. It is thus estimated that on average 0.8 ($=21/26$) mutation is recorded on the readout sequence per cell generation; the number would be roughly 1.3 ($=21/16$) if the initial ten ultrafast nuclear cleavage cycles (completed within roughly an hour after egg laying at 25°C) in the fly embryogenesis are not counted³⁷. Hence, the per-generation mutation rate should be somewhere between around 0.8 and 1.3. However, we noted that the estimations are very rough (likely to be overestimated by assuming no cell deaths and all symmetric cell divisions) and should be used with caution. The mutations are present pervasively across the 3-kb sequence despite its near absence on the iSceI binding motifs (Supplementary Fig. 3a,b). We also computed the mutual information between the mutated sites and found overall low interdependence (Supplementary Fig. 3c).

We conducted simulations to assess the possibility of obtaining a statistically reliable cell phylogenetic tree based on the barcoding information (Supplementary Note II). Unlike CRISPR-Cas9-based barcoding methods, which rely on indels²², SMALT produces substitution mutations on the readout sequence such that the data structure is the same between the reconstruction of cell phylogeny in this study and the conventional reconstructions of species phylogeny. Thus, SMALT is compatible with most sophisticated phylogeny reconstruction methods². We used a time-consuming but most reliable method, the maximum-likelihood method³⁸, in our phylogeny reconstruction. One advantage of using the likelihood method is that the various previous mutation rates of different sites can be automatically modeled in the calculation of likelihood. In addition, we used bootstrapping to model the effects of errors and insufficient barcoding mutations²⁴. The simulation results indicate that SMALT would enable high-quality mapping of cell phylogeny (Supplementary Fig. 4).

We thoroughly dissected two late-third instar larvae (wandering larvae) in which the SMALT system was active. We obtained more than 12 organs from each fly, with 16 shared between the two individuals (Fig. 2c). For each organ, we amplified the 3-kb readout sequence using primers containing unique molecular identifiers (UMIs), and sequenced the amplicons by single molecule real-time (SMRT) sequencing to identify mutations (Supplementary Figs. 5 and 6 and Methods). As a control, we examined the readout sequence of a fly sample in which SMALT was not activated, to benchmark the mutation identification process. We obtained over 5,000 high-quality readouts with distinct UMIs (hereafter referred to as alleles) in each fly (Supplementary Figs. 7 and 8 and Methods). Because the 3-kb readout sequence is present as a heterozygous single copy in the zygotic genome, each allele can be treated as a single cell for organs whose cells are diploid. For some organs in which the cells contain polytene chromosomes³⁹, an allele represents a set of chromatids that are equivalent to the genome a diploid cell has. The barcoding mutations are found pervasively across the 3-kb sequence with peaks near the iSceI binding motifs (Supplementary Fig. 9). The median mutation number per allele ranges from 6 to 28 among different organs, with largely consistent patterns between the two individuals (Fig. 2d). Up to roughly 99% of the mutations were CG-to-TA or CG-to-GC. The observed average mutation frequency is 1.0 and 0.002%, respectively, per CG site and AT site on the alleles of the two flies with active SMALT, while the numbers are 0.007 and 0.002% in the control sample (Fig. 2e). This indicates a negligible error rate of the identified barcoding mutations. In addition, over 80% of the alleles have a unique mutation profile (Fig. 2f).

Using the above alleles, we then reconstructed a maximum-likelihood phylogenetic tree for each of the two flies with the wild-type readout sequence as the root (Fig. 3a–c and Methods). We used bootstrapping to evaluate the statistical reliability of the trees. Consistent with the simulation results, the internal nodes were overall supported by large bootstrap values, with medians of

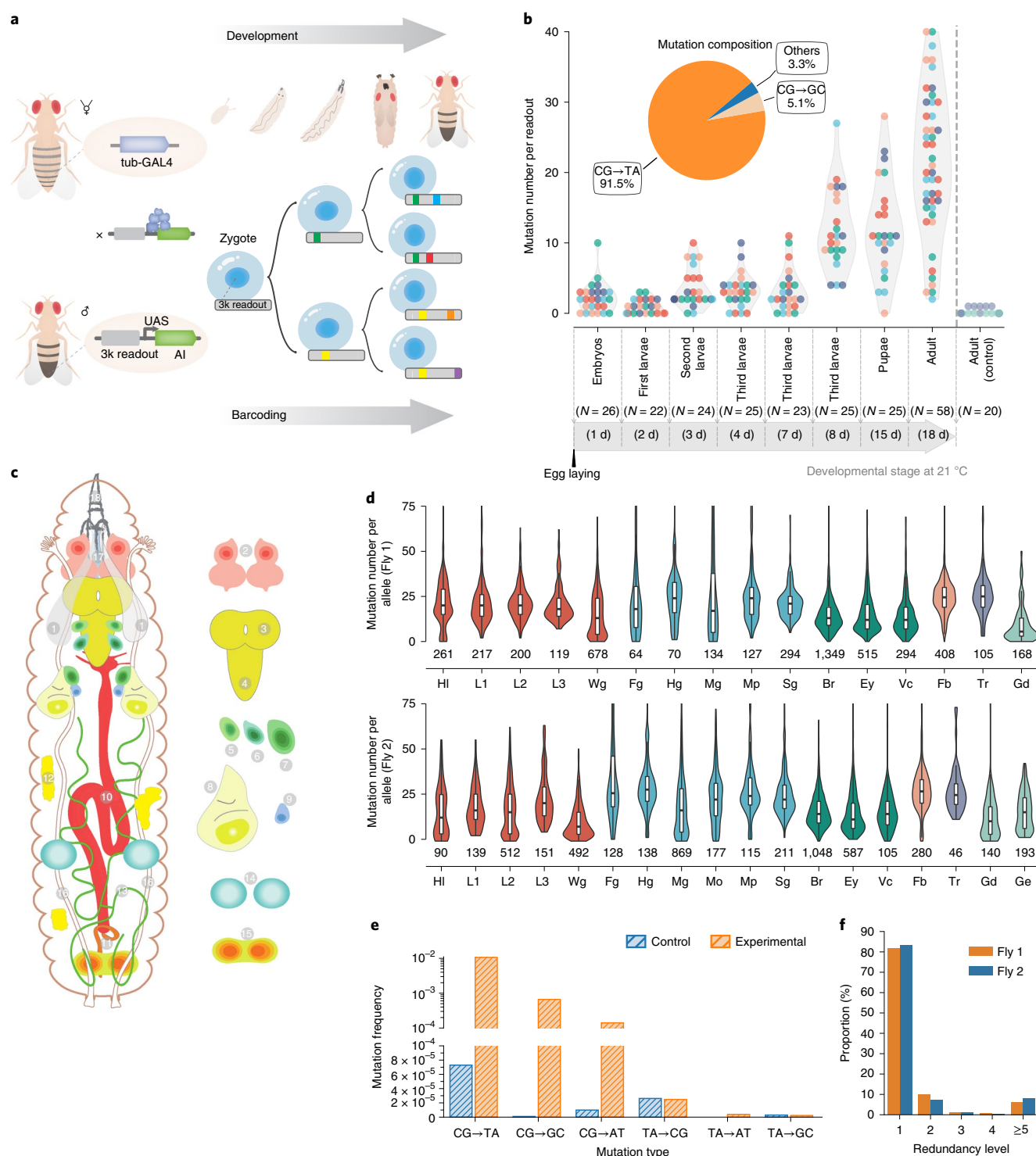


Fig. 2 | Performance of SMALT in fruit fly. a, A schematic diagram showing how SMALT could barcode fly development. **b**, The number of mutations per readout sequence observed at various development stages specified by the time windows (shown in parentheses) within which whole-body samples are collected. Each dot represents a readout sequence, with colors showing different individuals. The pie chart shows the relative frequencies of the different types of substitution mutation examined here. The control represents adult individuals without the GAL4 activator. **c**, A schematic diagram showing the different organs obtained from larva. Numbers relate to organs and abbreviations as follows: 1, salivary glands (Sg); 2, eye-antennal discs (Ey); 3, brain disc (Br); 4, VNC disc (Vc); 5, leg discs vT1 (L1); 6, leg discs vT2 (L2); 7, leg discs vT3 (L3); 8, wing discs (Wg); 9, haltere discs (Hl); 10, midgut (Mg); 11, hindgut (Hg); 12, fat body (Fb); 13, malpighian tubule (Mp); 14, gonad (Gd); 15, genitalia disc (Ge); 16, tracheae (Tr); 17, foregut (Fg) and 18, mouthparts (Mo). **d**, Violin plots of the per-allele mutation number for each of the organs of two fly individuals (an allele is a high-quality readout sequence with UMI characterized by SMRT sequencing). A standard box plot is also provided within each violin, with the center showing the median and the lower and upper hinges showing the first and third quartiles of the data. The number below each violin shows the number of alleles included in the organ. **e**, Estimation of the overall error rate of the mutations identified on the roughly 10,000 alleles by SMRT sequencing by referring to a control sample in which SMALT is not activated. **f**, Proportion of alleles with unique mutation profiles in both flies.

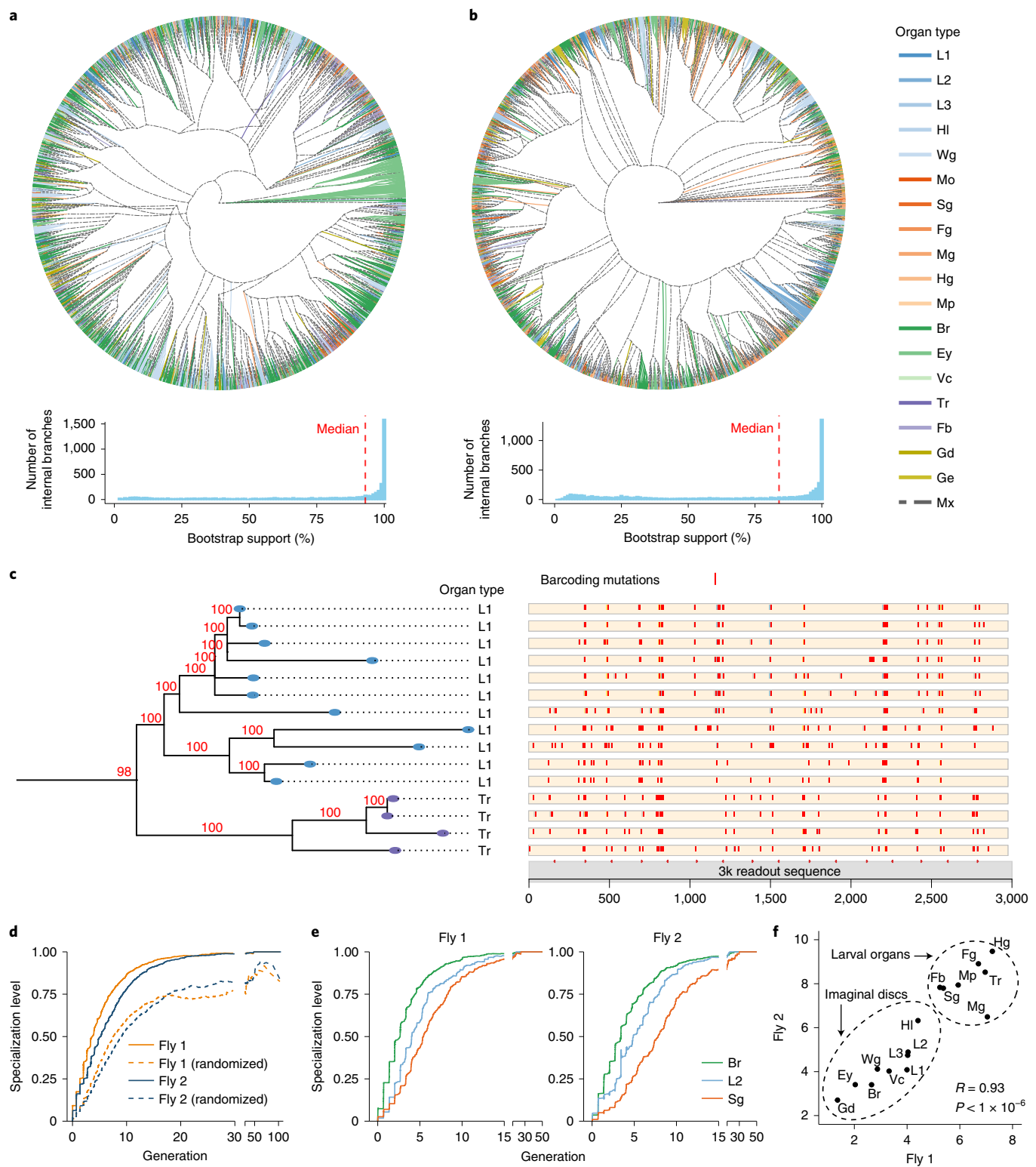


Fig. 3 | Building high-quality cell phylogenetic trees of two fly individuals. a, A dendrogram of the cell phylogeny comprising over 5,000 alleles from fly 1, with the original readout sequence as the root and nodes of different organs in different colors. The accompanying density plot shows the frequency distribution of the bootstrap values on the internal branches of the tree. The median bootstrap value is 93%. Mx signifies mixed organs (for clades containing more than one organ). **b**, The cell tree of fly 2, with a median bootstrap value 84%. **c**, Details of an exemplar clade comprising 15 terminal nodes trimmed from the tree of fly 1, with the bootstrap values shown on the branches and the corresponding barcoding mutations on the right. **d**, The trajectory of specialization level measured by the divergence of organ composition among subclades of a given tree height (generation). The 'randomized' curve was obtained by swapping the organ identities of all alleles. **e**, The specialization trajectory of a specific organ relative to the other organs. Three example organs are shown for both flies. **f**, The tree height (generation) corresponding to a specialization level of 0.5 is shown for the 16 organs available in both flies. The specialization rate is generally higher for imaginal discs than for organs that function specifically in larvae, with highly correlated patterns between the two individuals.

93 and 84% in the two trees. The bootstrap values on the internal nodes near the root are relatively small (Extended Data Figs. 4 and 5), reflecting the lack of barcoding mutations at the early nuclear cleavage cycles of fly embryogenesis³⁷. In each tree there are over 3,000 internal nodes (3,392 in fly 1 and 3,577 in fly 2), suggesting that thousands of mitotic cell divisions are tracked. By considering the number of internal nodes as well as their statistical supports, the two cell phylogenetic trees obtained here are unprecedentedly informative compared to the available cell trees (Supplementary Table 2).

A close examination of the two cell trees revealed rather small monophyletic clades formed by a specific organ, highlighting the heterogeneity of the genealogical origins of each organ. To track the process of cell fate specialization we moved from root to leaves and found, in both trees, rapid divergence of organ composition between subclades (Fig. 3d and Methods). We then mapped the specialization trajectory of a specific organ relative to the other organs (Fig. 3e and Supplementary Fig. 10). We observed a highly consistent pattern between the two cell trees (Fig. 3f), with gonad (Gd) being the first organ to separate from the others, which is in line with the fact that gonad lineages are the first to be fate determined⁴⁰. As for the remaining organs, the discs (eye-antennal, brain, wing, VNC, haltere and legs) overall have an earlier specialization time than the larva-specific organs (fat body, salivary gland, midgut, hindgut, foregut, malpighian tubule and tracheae)³⁶. This finding suggests that the larval organs in general have more complex genealogical origins than the discs^{36,41,42}. Notably, for the larva-specific organs in which the cells often undergo endoreplication to generate polytene chromosomes³⁹, the new lineages generated by the endoreplication are in the form of chromosomes but not cells.

Cell phylogeny informs cell population dynamics

In addition to recording the genealogy, cell phylogeny should also inform the dynamics of the cell population⁴³. By adapting the methods used in conventional population genetic analyses^{44–46}, we designed an instantaneous coalescent method⁴⁷ to reconstruct the dynamics of the number of actively dividing parental cells (N_p) during development (Fig. 4a, Extended Data Fig. 6 and Methods). Here, parental cells refer to those that have divided before the moment of sampling. To validate the method, we simulated an organ development process and found that the N_p dynamics can be reliably reconstructed from the phylogeny of sampled cells, which is not sensitive to cell deaths (Fig. 4b, Extended Data Fig. 6 and Methods). Notably, this coalescent method operates in a way that is insensitive to population structure, although more samples seem necessary for structured populations than their panmictic counterparts (Extended Data Fig. 7). This property is important because organs are often highly structured cell populations that would heavily confound conventional coalescent analyses.

We focused on ten organs that were well sampled in both flies (Supplementary Fig. 11). For each of them we obtained the longitudinal dynamics of N_p under the assumption of a strict molecular clock (with one mutation per cell generation) by using the above coalescent method (Methods). We skipped the first two cell generations of each cell tree because of the poor bootstrap supports on the tree branches within the time range (Extended Data Figs. 4 and 5). The estimated N_p at early cell generations (starting from the third cell generation) is generally small and increases gradually to reach the highest level of a few thousand, followed in some cases by a gradual reduction (Fig. 4c). Further subsampling analyses show that the dynamics is reasonably robust despite apparent variations in less sampled organs (Supplementary Fig. 12). We also used Phylodyn⁴⁸ to infer the dynamics and obtained similar curves (Supplementary Fig. 13). Notably, the N_p dynamics of the whole body (ten organs) by adding up the ten dynamics is similar to that derived from all alleles of the ten organs (Fig. 4d). The result highlights the ability of the coalescent method to deal with structured cell population.

The slight differences are explained in part by shared populations among some organs, which inflates the sum of the N_p values (Supplementary Fig. 14).

Since the time interval used in the N_p estimation is one cell generation, there should be one net cell gain from each parental cell within the time interval. Hence, N_p here is equivalent to the number of cell births (net gains) of each generation, and the final cumulative N_p (Cum_ N_p) would measure the total number of cell births in an organ (Methods). Notably, for the three organs that function specifically in larva (Fb, Sg and Mp), their cells do not divide after the embryonic stage but undergo endoreplication to generate polytene chromosomes³⁹. Hence, their N_p measured alleles rather than cells although for simplicity we do not emphasize the distinction. We found that the final Cum_ N_p was highly similar between the two flies (Pearson's $R=0.97$, $n=10$, $P<0.0001$, one-tailed t -test; Fig. 5a). Furthermore, the estimated total cell births, or Cum_ N_p , explains the actual number of cells in the organs (Fig. 5b and Supplementary Table 3). The pattern can be ascribed to neither the sampling size of the alleles nor the average number of mutations on the alleles in each of the organs. These findings support the quality of the cell phylogenies as well as the coalescent method.

The N_p dynamics illustrate the three possible cell division modes a parental cell could have (Fig. 5c). In mode I, which we called symmetric cell division, the two descendants are both a parental cell that can further divide. In mode II, which we called asymmetrical cell division, one descendant is parental and the other one is a quiescent cell that does not further divide. In mode III, which we called terminal cell division, the two descendants are both quiescent. Notably, in this study quiescent cells refer to those that have not divided at the moment of sampling. When cell deaths are not considered, the relative frequency of the three cell division modes determines the ratio of $N_{p,g+1}$ to $N_{p,g}$, where $N_{p,g+1}$ and $N_{p,g}$ are the estimated N_p of generation g and generation $g+1$, respectively. The $N_{p,g+1}/N_{p,g}$ ratio would be two if all cell divisions of the generation g are mode I, or 1 if all are mode II, or 0 if all are mode III. By assuming that mode III is rare in early development^{37,49}, the $N_{p,g+1}/N_{p,g}$ ratio should be in the range of roughly 1–2 for early cell generations. This provides an additional test for the N_p estimations. We plotted $N_{p,g+1}/N_{p,g}$ for $g=3–10$ in each of the organs and found them well in the range of roughly 1–2, with an overall signal in each fly much stronger than expected ($P<0.001$ in both individuals, one-sided permutation test; Fig. 5d). Those far away from the expected range were found mainly in the less sampled organs, suggesting sampling noise rather than underlying errors. We repeated the above analyses by assuming molecular clocks with various per-generation mutation rates and observed qualitatively the same patterns (Supplementary Fig. 15). The derived $N_{p,g+1}/N_{p,g}$ ratio would be a useful parameter for gauging the expansion (mode I) and differentiation (modes II and III) of a cell population during development.

Discussion

In this study we developed SMALT, a highly efficient cell barcoding method for mapping the cell phylogeny of complex organisms. As shown in Supplementary Table 1, SMALT has three main strengths compared to the state-of-the-art cell barcoding methods. (1) A short barcode sequence at a single genomic locus is used to represent a cell. This outperforms the design in many other methods that use a set of barcode sequences (barcode set) present in separate genomic loci to represent a cell. The drawback of using a barcode set to represent a cell is the difficulty of combining the information from the separate barcodes in each cell. Without an efficient strategy for combining per-cell information one could either work on a small number of isolated single cells^{18,43} or track the relationship of bulk tissues by examining the mutation frequency spectrum¹⁴. One attractive solution is to make the separate barcodes express and obtain the combined information by single-cell RNA-sequencing

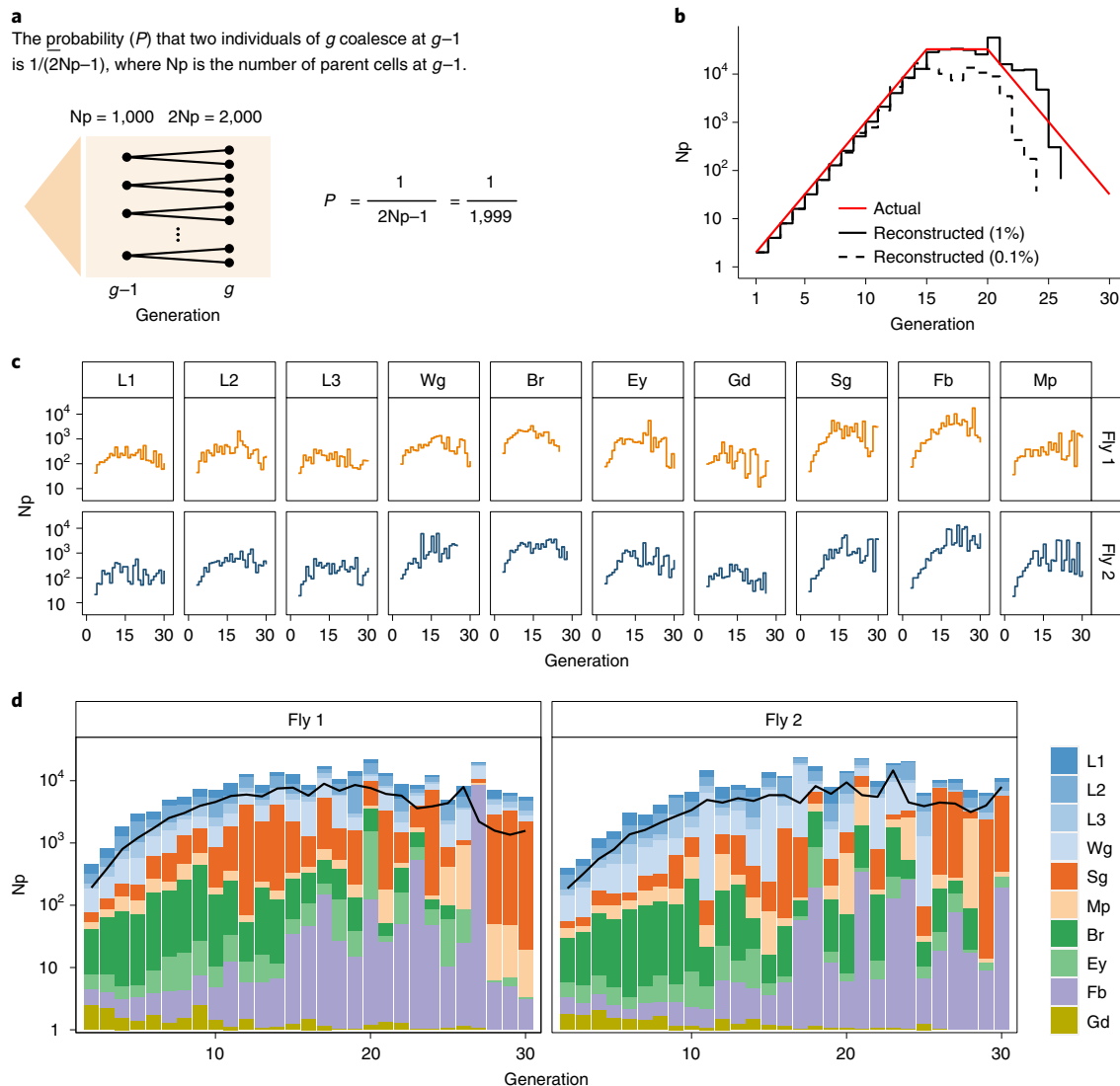


Fig. 4 | Population reconstruction of the demographic history of ten fly organs. a, A diagram explaining how coalescent rate can be used to estimate Np , the number of parental cells, at the previous generation. Each parental cell of generation $g-1$ produces two daughter cells of generation g . **b**, In a simulated organ development, the actual dynamics of Np can be reconstructed by the coalescent method from cell phylogeny. We tested sampling proportions of 1 and 0.1% of the cells of the simulated organ, and the 1% sampling proportion seemed to be required for a reliable reconstruction under a mutation rate of one mutation per cell generation. **c**, The Np dynamics of ten fly organs reconstructed from the two cell phylogenetic trees. Results of the first 30 cell generations are shown. The first two cell generations are not considered because of the low bootstrap values within the time period in the trees. **d**, The add-up of the ten Np dynamics is consistent with the Np dynamics (black line) derived from the coalescent probability of all alleles of the ten organs.

(scRNA-seq). Because both cell type and cell phylogeny information can be simultaneously obtained, this strategy has been adopted by ScarTrace (eight barcodes)¹³, Linneaus (32 barcodes)¹⁶ and Molecular Recorder (15 barcodes)¹⁷. It is, however, not ideal from the perspective of mapping cell phylogeny because of the high dropout rate of the barcodes in scRNA-seq data. Specifically, on average no more than half of the barcodes can be recovered from a cell in any of these barcoding systems. Hence, there would be less than one-quarter of the shared barcodes between two cells, and this number would drop exponentially as a function of the number of cells under examination. The prevalent missing data would confound the reconstruction of cell phylogeny². Another drawback of using barcode sets is the difficulty of obtaining a stable transgenic line because the multiple barcodes at different genomic loci segregate independently, which undermines reproducibility. (2) The average number of mutations per barcode reaches around 21,

comparable to the number of cell generations from zygote to an adult organism. Hence, the detailed history of mitotic cell divisions during development can be recorded. As a comparison, there are on average roughly three mutations per barcode in the state-of-the-art methods including GESTALT⁹, Polylox¹⁰, CARLIN¹⁹ and TRACE²¹. By considering a barcode set, the number of mutations for barcoding each cell would increase. However, the increase is moderate because, so far, the achieved highest mutation number per barcode set is roughly 6.6 using in Molecular Recorder¹⁷. (3) Because the goal of a cell barcoding method is to produce a reliable cell phylogeny, a good method should fit the goal. SMALT produces substitution mutations, the typical mutation type used in reconstructing species phylogeny². Hence, the many sophisticated phylogeny reconstruction algorithms that have been extensively tested in the past few decades can readily applied to data generated via SMALT. There are established strategies for evaluating the statistical confidence of

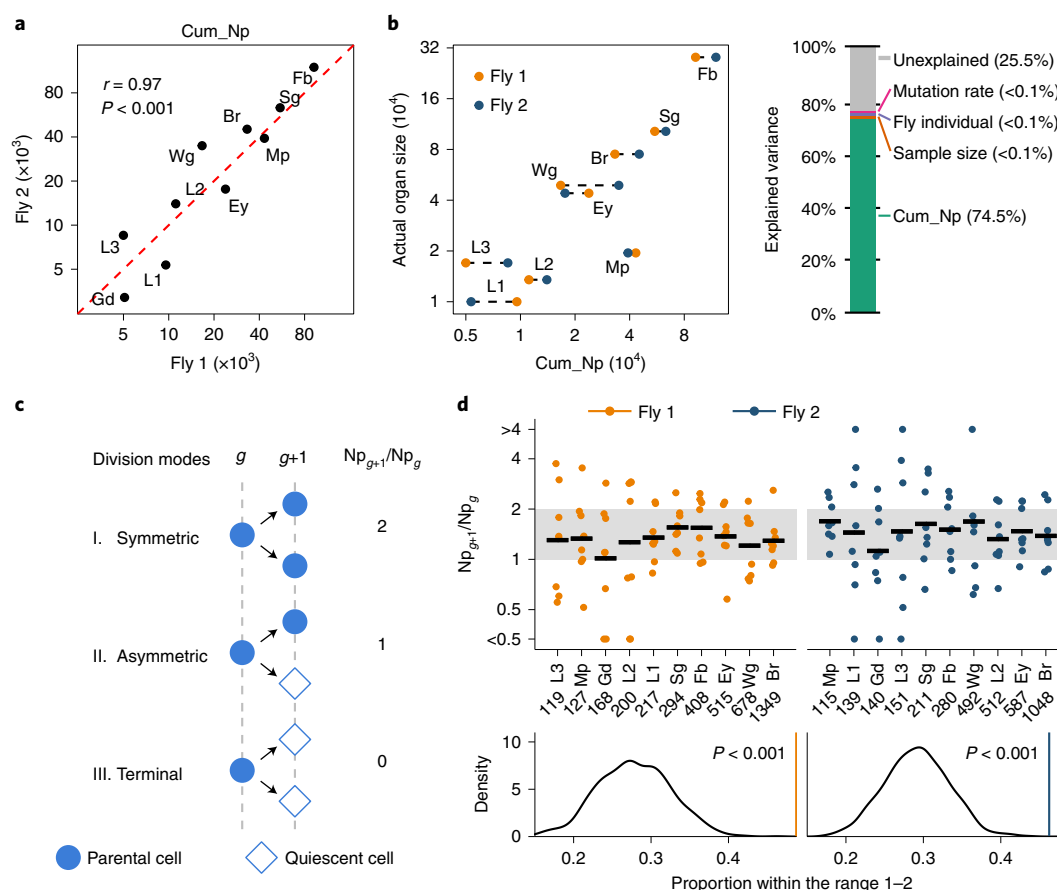


Fig. 5 | The Np dynamics informs the rate of cell births and cell division modes. **a**, The final cumulative Np (Cum_Np) of each organ is highly consistent between the two individuals (Pearson's $R = 0.97$, $n = 10$, $P < 0.0001$, one-tailed t -test). **b**, The final Cum_Np, which measures the total births of an organ, explains well the actual organ size. Only nine organs are considered because the gonad (Gd) has no reported cell number available. A multiple variable analysis that includes the average mutation rate on the alleles, the fly individual difference, the sampling size of the alleles and the estimated total births (that is, Cum_Np). **c**, The three modes of cell divisions, each resulting in a different Np_{g+1}/Np_g ratio. **d**, The Np_{g+1}/Np_g ratios for early cell generations ($g = 3-10$ roughly) in each of the ten organs of the two individuals. Within each organ a dot represents a cell generation, and the number beneath the organ name shows the number of alleles the organ has. The two density plots show the expected proportion of Np_{g+1}/Np_g ratios ($g = 3-10$) within the range of roughly 1–2, which is estimated by swapping the Np values of the same generation among the ten organs of an individual (with 1,000 replicates). The vertical line shows the observed proportion, which is 48.7 and 46.2%, respectively, in the two individuals.

the phylogeny reconstructed based on substitution mutations²⁴. By contrast, current CRISPR–Cas9-based barcoding methods rely on indels. Because deletion would easily remove mutation information already recorded on a barcode sequence⁹, the resulting barcodes have a lot of missing information. The difficulty of modeling the rate of the indel mutations and dealing with missing information has led to the general lack of statistical supports on the cell phylogenies obtained by these methods⁵⁰. While SMALT currently outperforms existing methods, a future updated version of SMALT in which the barcode is expressed to be recoverable by scRNA-seq would further expand its application scope, thus enabling the reconstruction of cell population dynamics for specific tissues or cell types.

The cell phylogenies obtained in this study enabled a population genetic reconstruction of the developmental demographic history of fruit fly organs. There are several related technical limitations worth discussing. First, the performance of the barcoding system at very early nuclear cleavage cycles in the fly was not satisfactory. As a consequence, we did not have the resolution to reconstruct the detailed population divergence history between organs and instead concentrated on intraorgan history in this study. Second, in the population reconstruction a strict molecular clock was applied to the cell trees⁵¹, which may not be ideal given the heterogeneity

among organs and developmental stages. Thus, a new and more flexible model that incorporates the related developmental conditions could lead to improved reconstruction. Third, despite generally strong bootstrap supports, there are still many errors in the cell trees. A more sophisticated method incorporating the errors would improve the population inferences. Despite these limitations, SMALT-reconstructed history revealed the trajectory of cell births during organ development and provided information about cell division modes. This resembles the reconstruction of the evolutionary demographic history of organisms, which can shed light on selection, migration, geographic catastrophes and so on⁵². With the emergence of new methods capable of producing accurate and high-quality cell phylogeny, we anticipate more new biological insights from such phylodynamic analyses²⁷.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-021-01325-x>.

Received: 4 March 2021; Accepted: 18 October 2021;
Published online: 2 December 2021

References

1. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
2. Nei, M. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* **30**, 371–403 (1996).
3. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
4. Woodworth, M. B., Girsakis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).
5. Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, 382–394 (2005).
6. Salipante, S. J. & Horvitz, M. S. Phylogenetic fate mapping. *Proc. Natl Acad. Sci. USA* **103**, 5448–5453 (2006).
7. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
8. Luo, T., He, X. & Xing, K. Lineage analysis by microsatellite loci deep sequencing in mice. *Mol. Reprod. Dev.* **83**, 387–391 (2016).
9. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
10. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
11. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).
12. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
13. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
14. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
15. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
16. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
17. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–7 (2019).
18. Hwang, B. et al. Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nat. Commun.* **10**, 1234 (2019).
19. Bowling, S. et al. An engineered CRISPR–Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. *Cell* **181**, 1410–1422 (2020).
20. Ye, C., Chen, Z. X., Liu, Z., Wang, F. & He, X. L. Defining endogenous barcoding sites for CRISPR/Cas9-based cell lineage tracing in zebrafish. *J. Genet. Genomics* **47**, 85–91 (2020).
21. Chen, H. Q. et al. Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor. *Nat. Biotechnol.* **38**, 165–166 (2020).
22. Baron, C. S. & van Oudenaarden, A. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.* **20**, 753–765 (2019).
23. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
24. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
25. Wasserstrom, A. et al. Estimating cell depth from somatic mutations. *PLoS Comput. Biol.* **4**, e1000058 (2008).
26. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* <https://doi.org/10.1371/journal.pbio.1002533> (2016).
27. Stadler, T., Pybus, O. G. & Stumpf, M. P. H. Phylodynamics for cell biologists. *Science* **371**, <https://doi.org/10.1126/science.aah6266> (2021).
28. Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* **10**, 1247–1253 (2002).
29. Chen, X. et al. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* **335**, 1235–1238 (2012).
30. Chen, P., Wang, D., Chen, H., Zhou, Z. & He, X. The nonessentiality of essential genes in yeast provides therapeutic insights into a human disease. *Genome Res* **26**, 1355–1362 (2016).
31. Prorok, P. et al. Uracil in duplex DNA is a substrate for the nucleotide incision repair pathway in human cells. *Proc. Natl Acad. Sci. USA* **110**, E3695–E3703 (2013).
32. Wang, M., Yang, Z. Z., Rada, C. & Neuberger, M. S. AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. *Nat. Struct. Mol. Biol.* **16**, 769–776 (2009).
33. Fonfara, I., Curth, U., Pingoud, A. & Wende, W. Creating highly specific nucleases by fusion of active restriction endonucleases and catalytically inactive homing endonucleases. *Nucleic Acids Res.* **40**, 847–860 (2012).
34. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl Acad. Sci. USA* **111**, E2310–E2318 (2014).
35. Brand, A. H. & Perrimon, N. J. D. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
36. Bate, M., Martinez Arias, A. & Hartenstein, V. *The Development of Drosophila melanogaster* (Cold Spring Harbor Laboratory Press, 1993).
37. Farrell, J. A. & O'Farrell, P. H. From egg to gastrula: how the cell cycle is remodeled during the *Drosophila* mid-blastula transition. *Annu. Rev. Genet.* **48**, 269–294 (2014).
38. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
39. Edgar, B. A., Zielke, N. & Gutierrez, C. Endocycles: a recurrent evolutionary innovation for post-mitotic cell growth. *Nat. Rev. Mol. Cell Biol.* **15**, 197–210 (2014).
40. Underwood, E. M., Caulton, J. H., Allis, C. D. & Mahowald, A. P. Developmental fate of pole cells in *Drosophila melanogaster*. *Dev. Biol.* **77**, 303–314 (1980).
41. Buchon, N. et al. Morphological and molecular characterization of adult midgut compartmentalization in *Drosophila*. *Cell Rep.* **3**, 1725–1738 (2013).
42. Miguel-Aliaga, I., Jasper, H. & Lemaitre, B. Anatomy and physiology of the digestive tract of *Drosophila melanogaster*. *Genetics* **210**, 357–396 (2018).
43. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
44. Fu, Y. X. A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**, 685–692 (1994).
45. Pybus, O. G., Rambaut, A. & Harvey, P. H. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437 (2000).
46. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
47. Yao, Z., Liu, K., Deng, S. & He, X. An instantaneous coalescent method insensitive to population structure. *J. Genet. Genomics* **48**, 219–224 (2021).
48. Karcher, M. D., Palacios, J. A., Lan, S. & Minin, V. N. phylodyn: an R package for phylodynamic simulation and inference. *Mol. Ecol. Resour.* **17**, 96–100 (2017).
49. Hu, Z., Fu, Y. X., Greenberg, A. J., Wu, C. I. & Zhai, W. W. Age-dependent transition from cell-level to population-level control in murine intestinal homeostasis revealed by coalescence analysis. *PLoS Genet.* **9**, e1003326 (2013).
50. Salvador-Martinez, I., Grillo, M., Averof, M. & Telford, M. J. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife* **8**, e40292 (2019).
51. Ho, S. Y. W. & Duchene, S. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* **23**, 5947–5965 (2014).
52. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Invention of SMALT in yeast. *Yeast strain and plasmid.* The yeast strain GIL104 (*MATa*, *URA3*, *leu2*, *trp1*, *CAN1*, *ade2*, *his3*, *bar1Δ::ADE2*), which is derived from the W303 background, was used to test the mutagenesis performance of the cytidine deaminases. The strain was regularly maintained using YPD (1% yeast extract, 2% peptone, 2% glucose) or synthetic complete (SC) medium with 2% glucose as the carbon source. Plasmid *pGAL1*, bearing the *LEU2* selectable marker and the galactose-inducible *GAL1* promoter, was used as backbone in the yeast induction expression experiment. Both the tested gene fragments and the vector were digested by HindIII and XbaI (NEB), and then ligated with T4 DNA ligase (NEB). Transformed yeast was selected in synthetic medium deprived of leucine (SC-Leu) to maintain the vector. SC-Leu+Glu media (SC-Leu with 2% glucose) was used for expression suppression, and SC-Leu+Gal media (SC-Leu with 2% galactose and 1% raffinose) for expression induction. The standard LiAc transformation method was used to transform DNA into the yeast cells⁵³, and gene replacements on chromosomes were achieved by homologous recombination.

Cytidine deaminases. The genes of the cytidine deaminase family were commercially synthesized (GENEWIZ), and all PCR primers were also obtained from the same vendor. Unless otherwise stated, PCR was performed using the high-fidelity polymerase KOD FX Neo (TOYOBO). Site-directed mutagenesis for the human *AID* sequence were introduced via overlapping PCR with mutagenic primers. Deaminase gene was cloned into *pGAL1* vector for testing. Sequences of the genes and the map of vectors are available in Supplementary Dataset I.

Phenotypic mutation rate estimation. Mutation assays were performed similar to our previous studies, as in refs.^{29,30}. In brief, a single yeast clone was first cultured in SC-Leu+Glu to saturation (roughly 10^7 cells per ml), and then transferred by 1:10,000 dilution to SC-Leu+Gal for induction of the cytidine deaminase gene expression. The cultures were maintained for 2 d at 30 °C without shaking, and then resuspended using a pipet before plating. They were plated onto overdried SC-Leu-Arg agar plates supplied with 2% glucose and $1\times$ canavanine (60 mg l^{-1} L-canavanine, Sigma-Aldrich). The number of mutants per spot was counted after 2 d of growth and the phenotypic mutation rate was calculated using Ma–Sandri–Sarkar maximum-likelihood estimation equations⁵⁴.

Genetic mutation rate estimation. A 454-bp sequence with two divergent iSceI binding motifs was commercially synthesized (Life Technologies). The sequence was integrated into the *URA3* locus of the yeast strain GIL104, which was subsequently transformed with the vector *pGAL1_AID₅-iSceI*. There were two biological replicates with galactose induced expression of *AID₅-iSceI* and one control experiment with no induction of *AID₅-iSceI*. All three yeast samples started from 10,000 cells. After ten generations of growth in SC-Leu liquid media (2% galactose added only to the two induced samples and 2% glucose to the control sample), the target sequence was amplified with a pair of primers each containing a 3-bp sample index. An Illumina sequencing library was prepared by following the standard protocol, and PE300 sequencing was performed on the Illumina MiSeq platform for the amplicons. Reads with an average Phred quality score less than 30 were dropped. The remaining paired-end reads were merged based on their overlapping sequence (140 bp) using *fastq_mergepairs*. We obtained 350,763 reads for the control sample, and 217,398 and 299,168 reads for two induced replicate samples. The average Phred score per site was greater than 38 for each sample. These reads were mapped to the reference sequence using the *bwa-mem* aligner with default settings. The alignment file was piled up by site, and the substitution mutations were called from the pileup data only for sites with a Phred score greater than 35. The rate of mutation was calculated for each CG site of the sequence. The two divergent fragments were merged as a single one according to the relative distance of the sites to the corresponding binding motif. The average mutation rate was calculated for each position that is away from the iSceI binding motif. In the end the net mutation rate in the two induced replicate samples was estimated by subtracting the corresponding background mutation rate estimated in the control sample.

Error-prone PCR. The mutant library of *AID₅* was generated by error-prone PCR using the Diversify PCR Random Mutagenesis Kit (Clontech). As suggested by the protocol, we used 1 ng of template DNA with $1\text{ }\mu\text{M}$ of primers, chose condition 8 as the aimed level of mutagenesis and obtained in the end on average roughly four substitution mutations per molecule. The PCR products underwent agarose gel electrophoresis and purification using the QIAquick PCR Purification Kit (Qiagen). Similarly, the library of *AID₅*-derived mutants was ligated into *pGAL1* for activity test.

Selection of homing endonuclease. Ten homing endonucleases without the DNA cutting activity were commercially synthesized (GENEWIZ). To achieve target-specific mutagenesis, deaminases and a homing endonuclease were fused with a linker of 30 amino acids ($6\times$ GGGGS), and cloned into *pGAL1* for testing the targeting mutagenesis performance in yeast. Sequences of the homing endonucleases are available in Supplementary Dataset I.

Using SMALT to track the genealogy of yeast cells. The 3-kb DNA cassette was integrated into the *HO* locus of the yeast strain BY4741, and the AI sequence was integrated into the same locus of the BY4742 strain under the control of the *GAL1* promoter. After crossing the two strains a diploid yeast clone with both the AI and the 3-kb readout sequence was obtained. In the experimental evolution, the AI was induced by 2% galactose on SC-Ura agar plate for the first two rounds of barcoding. Four single-cell-derived clones after the second round of barcoding were picked and distributed onto 2% glucose SC-Ura agar plate, from which 40 single-cell-derived clones were chosen for further analysis. Mutations on the 3-kb readout sequence of the 40 clones were identified by Sanger sequencing, and an unweighted pair group method with arithmetic mean tree was constructed for the 40 sequences.

SMALT mutations depend on cell divisions. A newly designed 120-bp DNA barcode (yBC120) was integrated into the *HO* locus of the BY4741 strain, and the AI sequence was integrated into the same locus of the BY4742 strain under the control of Tet-On promoter. After crossing the two strains a diploid yeast clone with both the AI and the yBC120 barcode sequence was obtained. It is worth noting that the leaky expression of Tet-On system still exists, although we have done several modifications according to a recent paper⁵⁵. Because of the heterogeneity of the diploid clones and the leaky expression, we tested nearly 100 independent clones by examining the induction sensitivity and uniformity after adding Doxycycline (Sigma D9891-25G-9). We obtained three clones with excellent performance in both sensitivity and uniformity for the follow-up experiment. In brief, we controlled the yeast cell division by changing the culture medium, and then tested the accumulation of SMALT mutations on the 120-bp barcode sequence. We used YPDA (1% yeast extract + 2% peptone + 0.05% adenine + 2% glucose) as the cell dividing medium, and YEKAC (0.1% yeast extract + 2% potassium acetate) and PBS as two nondividing media for growing the yeast cells for roughly 18 h (roughly 12 generations in YPDA). The yeast cells were then collected and the 120 bp barcode sequence was PCR amplified (the amplicon is 227 bp), followed by Illumina sequencing (PE150). Over 1 million reads were obtained for each sample and processed to call mutations. After cleaning with *trim_galore*, processed reads were mapped to the 120 barcode using *bwa-mem*, followed by *bcftools mpileup* for mutation calling.

Applying SMALT to fly. For the design of the barcoding system in the fly, to suppress the potential repair of C-to-U mutations caused by AID we fused *AID₁₀-iSceI* with UGI, a protein that can inhibit the associated repair system⁵⁶. The 3-kb readout sequence was designed by considering the mutation range and the preferred targets of AID, with 16 iSceI binding motifs present with equal distance throughout the sequence. The readout sequence was commercially synthesized (Life Technologies). Details of the fusion protein and the 3 kb readout sequence are available in Supplementary Dataset II.

Preparing vectors for microinjection. Unless otherwise noted, PCR was performed using the KOD FX Neo (TOYOBO), and cloning was performed using the ClonExpress Kit (Vazyme Biotech). The sequences of all constructs were verified by Sanger sequencing before microinjection. We started with the vector *pUASTattB-eGFP*. Two vectors were prepared. In the first vector, the *eGFP* of *pUASTattB-eGFP* was replaced by the 3-kb readout sequence via the EcoRI–XbaI window. In the second vector, the *eGFP* was replaced by a codon-optimized AI with accompanying translational enhancers via the same window. For each vector, a transgenic fly line was obtained, and a further cross of the two lines was necessary to obtain the final line used in this study.

Obtaining transgenic flies. Unless otherwise noted, flies were normally maintained at 25 °C with 50% humidity and a 12-h light/12-h dark cycle. Embryo microinjections were conducted and transgenic lines were obtained by following standard procedures. All transgenic flies in this study were integrated as a single copy into the defined genomic locus using the ϕ C31/attP/attB system⁵⁷. All fly lines and vectors used in this study are available in Supplementary Dataset II. Imaging data were collected using the software of AMG EVOS FL Digital Inverted Microscope (Life Technologies, EVOS software revision 25566).

Off-target analysis in fly. Approximated 20 tubulin-GAL4 virgins and ten males (transgenic SMALT) were mated together at 21 °C. Offspring were collected after roughly 20 d for total genomic DNA extraction. For single fly genomic DNA extraction, each fly was homogenized in a microcentrifuge tube with a plastic pestle, and DNA was extracted using the Magen HiPure Insect DNA Kit (D3129-02) according to the manufacturer's instructions. A Heffl NG5 OnePot II DNA Library Prep Kit (Yeasen Biotech) was used for library preparation. The final libraries were quantified by $1\times$ double-stranded DNA High Sensitivity Kit in Qubit 3.0 Fluorometer (both from ThermoFisher Science) and Standard Quantitative Cartridge Kit in Qsep1 Bio-Fragment Analyzer (both from BiOptic, Inc.). Normalized samples were sent to the GENEWIZ Company for Illumina HiSeq platform sequencing (PE150). In total we obtained roughly 0.8 billion reads after trimming with *trim_galore*, leaving on average roughly 108 sequencing coverages for each individual (12 individuals). The processed sequences were mapped to the

D. melanogaster genome (BDGP6) using bwa-mem and duplicates were removed, followed by variants calling with GATK HaplotypeCaller.

Tracking the development-coupled barcoding mutations. We crossed a single male founder carrying the AI system (homozygous) with roughly ten tubulin-GAL4 virgins. Samples were collected at typical stages of the fly life cycle as specified. The 3-kb readout sequence was amplified from the genomic DNA of a whole individual and the products were cloned into the *pUC19* followed by Sanger sequencing.

Dissection of larval organs. We picked out a wandering male larva from the vial with the aid of a dissecting needle, and then transferred the larva to a silicon-coated dish filled with 1× PBS. Under a dissecting microscope, we carefully exposed the fat body, digestive system and imaginal discs in order, and obtained the anatomical structures as specified in Fig. 2c. These anatomical structures are loosely referred to as organs. The dissociated organs were transferred to a 1.5-ml microtube filled with 190 µl of genome extraction buffer ATL from the HiPure Tissue DNA Micro Kit (Magen). The genomic DNA of the organs was extracted using the same kit and quantified with Qubit v.3.0 Fluorometer (ThermoFisher).

Library preparation for SMRT sequencing. For each tissue approximately 10,000 copies of the fly genome were used as the template of a PCR reaction using KOD FX Neo (TOYOBO). We adopted a four-step strategy with six primers (P1–P6) to ensure an efficient PCR amplification with UMIs of the 3-kb readout sequences (Supplementary Fig. 6). Specifically, we first conducted one cycle of linear amplification using P1 primer that contains 14-nucleotide UMIs to tag the initial readout molecules, followed by ten cycles of linear amplification by adding sixfold more P2 to neutralize the remaining P1 primer. And then clean up the product with 0.7× AMPure XP beads (Beckman Coulter). Next, we conducted ten cycles of nested PCR using P3 and P4 to enrich the UMI-containing readout molecules, and the products were purified with 0.7× beads. In the last step, products of beads purification previous step were used as template to be further amplified using P5 and P6, both primers contain 16-nucleotide symmetric multiplexing organ-specific barcodes as well as the same padding sequence for the concern of unequal ligation efficiencies. Below are the details of the thermal conditions:

Step 1: 0.3 µM P1 in 50 µl of solution; thermal condition: 94°C/3 min (98°C/10 s, 58°C/1 min, 68°C/5 min) × 1 cycle, 68°C/10 min;
 Step 2: 1.8 µM P2 in 50 µl of solution; thermal condition: (55°C/30 s, 68°C/1 min) × 10 cycles, 68°C/30 s and then 0.7× bead purification;
 Step 3: 0.2 µM P3 and P4 each in 50 µl of solution; thermal condition: 94°C/1 min (98°C/10 s, 60°C/30 s, 68°C/5 min) × 10 cycles, 68°C/10 min and then 0.7× bead purification;
 Step 4: 0.3 µM P5 and P6 each in 50 µl of solution; thermal condition: 94°C/1 min (98°C/10 s, 55°C/30 s, 68°C/5 min) × 5 cycle, (98°C/10 s, 60°C/30 s, 68°C/5 min) × 20 cycles, 68°C/10 min and then 0.7× bead purification.

SMRT sequencing and data processing. One microgram of PCR products of the 3-kb readout sequence containing UMIs and organ-specific barcodes were used for SMRTbell library preparation (Sequel platform), which follows the standard protocol of SMRT sequencing. Sequencing runs were performed on two SMRT cells for each SMRTbell library (GENEWIZ). In total, roughly 1.2 M effective zero-mode waveguides were obtained for each individual.

We developed a workflow for mutation calling using single molecule sequencing data (Supplementary Fig. 8), which effectively controlled the influences of both low sequencing quality and strand-switching PCR amplifications that produce chimera molecules. In brief, the adapter sequence was first removed from each polymerase read by *baz2bam* with default settings to obtained subreads groups for each polymerase read. For each subread group, two strand-specific circular consensus sequences (CCS) were generated by *pbccs* v.4.0.0 with the parameters: --min-length=1000--num-threads=60--by-strand--min-passes=1--min-rq=0.66. After that, CCS reads were aligned to the reference 3-kb readout sequence by *minimap2* with slight modifications under the parameter setting: -A 4 -B 12 -O 10,15 -E 2,1 --score-C 20 --score-A 2 --score-D 4 --end-bonus 10 -a --MD -x map-pb. Then, the UMI, readout sequence and tissue barcode were separated from each CCS reads by the *pysam* package. The CCS reads were then grouped by their UMI and tissue barcode. Those CCS groups with <2 or >100 CCS reads were excluded from further analyses. For each remaining CCS group, the CCS reads were clustered according to their mutation profiles to identify the dominant subgroup that comprises >50% of the CCS reads. Then, for each dominant CCS subgroup the consensus readout sequence was obtained, with the posterior Phred score (Q) calculated for each site. In addition, we required a high-quality readout sequence to have an overall average quality of Q > 70, with <100 low-quality sites (Q < 50) and >1,000 high-quality sites (Q > 90). In the end, we obtained 5,003 high-quality readouts (or alleles) for fly 1 and 5,421 alleles for fly 2. The average Q score per site was larger than 90 in both individuals. We used the 3-kb readouts of a control fly sample mixed in SMRT sequencing library to estimate the error rate derived from the sequencing and data processing. Using the same computational pipeline, we obtained 728 high-quality readout sequences as benchmark. Scripts for processing the SMRT sequencing data are available on Github.

Reconstruction of cell phylogeny and fate specialization trajectory. We used a maximum-likelihood-based software IQ-TREE for the cell tree reconstructions. In the original readout sequence was assigned as the root. After evaluating all available mutation models, we used the model GTR2+FO+R10 for the tree reconstructions. Hence, the readout sequence was treated as a binary string, with '0' representing wild-type and '1' representing mutated. This resulted in misassignment of a very small proportion of the site information used in the tree reconstruction (0.042% in fly 1 and 0.04% in fly 2). The general time-reversible (GTR) model is the most complex model and considers the mutation rate of every site. In GTR2, the empirical state frequencies were optimized by maximum likelihood from the data (FO) and ten categorized rates (R10) generalized assumption of Gamma-distribution. For each reconstruction, we ran 1,000 rounds of ultrafast bootstrap approximation (-bb 1,000) and 1,000 rounds of SH-like approximate likelihood ratio test (-alrt 1,000) to evaluate the quality of the consensus tree. The likelihood rate of each site was estimated via --mlrate parameter. Ancestral sequences of the internal nodes were also reconstructed via the --asr parameter. To build a maximum-likelihood tree containing over 5,000 terminal nodes, the MPI version of IQ-TREE was used. The reconstructions were performed on high-performance computing clusters, taking roughly 5,000 CPU-hours for each reconstruction.

To map the fate specialization trajectory through a cell tree, we moved from root to leaves and obtained all subclades for each tree height (cell generation). The specialization level is

$$\text{Specialization level} = 1 - \frac{1 - \sum_i \omega_i \sum_k x_{ik}^2}{1 - \sum_k (\sum_i \omega_i x_{ik})^2}, \quad (1)$$

where x_{ik} is the frequency of the k th organ in the i th subclade, and ω_i is the weighted size of the i th subclade among all subclades of a given tree height. We swapped the organ identities of all alleles to obtain the null/randomized trajectory of the cell trees.

Population reconstruction of the Np dynamics of fly organs. In a cell population the probability (P) that two randomly sampled cells coalesce at the previous generation is

$$P = \frac{1}{2N_p - 1}, \quad (2)$$

where N_p is the number of parental cells that are dividing at the previous generation. With the time-resolved phylogeny of a set of sampled cells we can calculate the coalescent rate (CR) of each cell generation using

$$CR_g = \frac{2CE_g}{L_g \times (L_g - 1)}, \quad (3)$$

where L_g is the original number of lineages at generation g , and CE_g is the number of coalescent events for the L_g lineages within the time interval from g to $g-1$. According to equation (1), we have

$$N_{p_g} = 0.5 + \frac{1}{2CR_g}, \quad (4)$$

where CR_g represents the P of equation (1) at generation g . So long as the sampling is sufficiently dense and the mutation rate is high enough, the CR of each small time interval can be reliably estimated from the phylogeny of sampled cells to derive the dynamics of N_p .

To validate the coalescent method, we first simulated an organ development that spans 30 cell generations (as shown in Supplementary Fig. 14). It started with a single cell, followed by 15 rounds of symmetric cell divisions to generate $2^{15} = 32,768$ cells; in each of the next five generations only half of the cells divided (that is, all asymmetric cell divisions) such that the N_p remained constant (32,768); during the last ten generations the number of dividing cells (that is, N_p) reduced by a half every generation, resulting in 64 cells born in the 30th generation. Hence, by assuming no cell deaths, the whole process generated 262,110 cells that were present in the final simulated organ. Mutations were randomly assigned to the lineages at a specified per-generation rate, with no recurrent mutations allowed. Two lineages coalesce when they share a mutation. Using a similar procedure, we further simulated other development processes with designed N_p trajectories to test the performance of the coalescent method in dealing with population structure as well as cell deaths.

In the analysis of the two fly trees, we first dated the cell trees by assuming a strict molecular clock with one mutation per cell generation. We noted that the reconstruction quality and the shape of the N_p trajectory would largely remain if a different per-generation mutation rate (ranging from 0.5 to 1.5) was used, although the N_p values would be overall inflated or deflated. The trees were heterochronous given the various number of mutations of the alleles. For each organ, we extracted from the whole tree all lineages that seeded the organ's alleles to conduct organ-specific coalescent analyses. Using equation (3), we calculated the CR per cell generation to obtain the dynamics of N_p . If no coalescent events were observed

within the time interval from g to $g-1$ (that is, $CE_g=0$), we searched the previous generations further until we observed nonzero m coalescent events in the gk generation. The CE_g was set to be m to compute the CR_g , and the resulting CR_g divided by k was then used to derive the Np_g using equation (4). Observation of no coalescent events was mainly for late developmental stages when there are often a limited number of lineages available. To improve the confidence of the estimations, in each organ we ignored the latest lineages that contributed 5% of the alleles of an organ. Because of the relatively low bootstrap supports on the early branches of the cell trees, we did not consider the first two cell generations in all organs.

Statistics and reproducibility. All associations between paired samples were measured with a one-tailed Pearson's product-moment correlation coefficient, assuming the positive association between two groups. No statistical method was used to predetermine sample size. Sample sizes were determined by readout amplification and mixture in library preparation. No data were excluded from the analyses unless otherwise clearly stated. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment because our results are based on reproducible data processing instead of human estimations.

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

Data availability

Raw data have been deposited in the National Center for Biotechnology Information's Sequence Read Archive with accession numbers [PRJNA716791](#), [PRJNA761270](#) and [PRJNA761271](#). Source data supporting the findings of the present study are provided as online materials for this paper.

Code availability

Codes for processing the data are available at <https://github.com/CellLineage/SLOTH>.

References

53. Gietz, R. D. & Schiestl, R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 38–41 (2007).
54. Radchenko, E. A., McGinty, R. J., Aksenova, A. Y., Neil, A. J. & Mirkin, S. M. Quantitative analysis of the rates for repeat-mediated genome instability in a yeast experimental system. *Methods Mol. Biol.* **1672**, 421–438 (2018).
55. Roney, I. J., Rudner, A. D., Couture, J. F. & Kaern, M. Improvement of the reverse tetracycline transactivator by single amino acid substitutions that reduce leaky target gene expression to undetectable levels. *Sci. Rep.* **6**, 27697 (2016).
56. Mol, C. D. et al. Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell* **82**, 701–708 (1995).
57. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for *Drosophila* using germ-line-specific phi C31 integrases. *Proc. Natl Acad. Sci. USA* **104**, 3312–3317 (2007).

Acknowledgements

We are grateful to J. Zhang for inspiration, Y. Rong for information on the phiC31 system and for providing related fly strains, the Tsinghua Fly Center at the Tsinghua University for providing fly strains, Y. Zhao for suggestions and guidance with regard to the fly embryo microinjection, and T. Tang, C.-I. Wu, X. Shen and members of the Wu laboratory for help with fly work. We thank H. Chen, W. Qian, Y. Zhang, W. Zhai, C.-I. Wu, X. Huang, Z. Wang and J. Yang for discussion and comments on the paper. This work was supported by grants of the National Key R&D Program of China (grant no. 2017YFA0103504), the National Natural Science Foundation of China (grant nos. 31630042, 31970570, and 32070687), the Shanghai Municipal Science and Technology Major Project (grant no. 2017SHZDZX01) and the Guangdong Special Support Program (grant no. 2017TX04R395).

Author contributions

X.H., C.Y. and L.L. designed the study. C.Y., K.L. and L.L. conducted the yeast assays. K.L. and H.G. conducted the fly experiments. S.D., C.Y., Z.Y., K.L., J.W. and X.H. analyzed the data. X.H. supervised the study and wrote the paper with input from C.Y., S.D., K.L. and other coauthors.

Competing interests

A patent related to the developed technique has been filed.

Additional information

Extended data are available for this paper at <https://doi.org/10.1038/s41592-021-01325-x>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-021-01325-x>.

Correspondence and requests for materials should be addressed to Li Liu or Xiongfei He.

Peer review information *Nature Methods* thanks Hugo Bellen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Madhura Mukhopadhyay was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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