Is it possible to reconstruct an accurate cell lineage using CRISPR recorders?

Irepan Salvador-Martínez[1†], Marco Grillo[2,3†], Michalis Averof[2,3*], and Maximilian J. Telford[1*]

[1]Centre for Life's Origins and Evolution, Department of Genetics Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, UK
[2]Institut de Génomique Fonctionnelle de Lyon (IGFL), École Normale Supérieure de Lyon, 32 avenue Tony Garnier, 69007 Lyon, France
[3]Centre National de la Recherche Scientifique (CNRS), France
[†]These authors contributed equally to this work
[*]For correspondence

Cell lineages provide the framework for understanding how multicellular organisms are built and how cell fates are decided during development. Describing cell lineages in most organisms is challenging, given the number of cells involved; even a fruit fly larva has ∼50,000 cells and a small mammal has more than 1 billion cells. Recently, the idea of using CRISPR to induce mutations during development as heritable markers for lineage reconstruction has been proposed and trialled by several groups. While an attractive idea, its practical value depends on the accuracy of the cell lineages that can be generated by this method. Here, we use computer simulations to estimate the performance of this approach under different conditions. Our simulations incorporate empirical data on CRISPR-induced mutation frequencies in *Drosophila*. We show significant impacts from multiple biological and technical parameters - variable cell division rates, skewed mutational outcomes, target dropouts and different mutation sequencing strategies. Our approach reveals the limitations of recently published CRISPR recorders, and indicates how future implementations can be optimised to produce accurate cell lineages.

Correspondence:
michalis.averof@ens-lyon.fr
m.telford@ucl.ac.uk

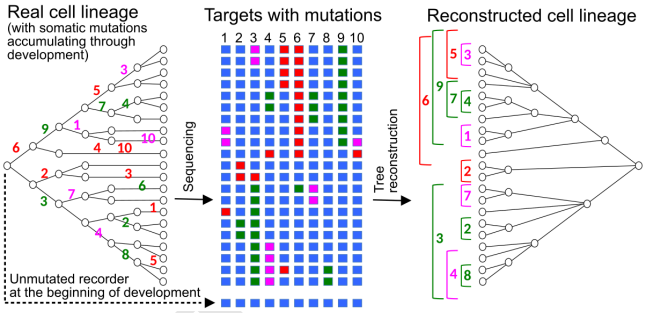

Fig. 1. Reconstructing cell lineages using CRISPR-induced somatic mutations. Left: Development begins with a zygote carrying in its genome a lineage recorder composed of a series of CRISPR targets (blue boxes). During subsequent cell divisions, any target of the recorder can be cleaved by Cas9 in any cell, leaving a specific mutational signature on the target which will be inherited by all the descendants of the cell. Numbers represent the the cleaved target in the recorder and its mutational signature is represented with a colour. Middle: At the end of development, the recorder of every cell is sequenced, recovering the pattern of accumulated mutations in each of the targets (coloured boxes). Right: The pattern of mutations is used to reconstruct the cell lineage, in a similar way to how a phylogenetic tree is inferred from the sequences of homologous genes.

Introduction

Starting from a single cell - the fertilised egg - multicellular organisms undergo repeated rounds of cell division to produce the adult form. The divisions that generate these adult cells constitute a genealogical tree with the fertilised egg at its root and each differentiated cell as a terminal branch. Knowing the cell lineage that produces a fully developed organism from a single cell provides the framework for understanding when, where and how cell fate decisions are made.

Obtaining high resolution (single-cell level) lineages is a challenging task that has been solved only in simple cases, such as the nematode *Caenorhabditis elegans*: its complete lineage (∼1000 cells) was deduced by painstaking observation of each cell division under the microscope. This approach is impossible in larger animals, in which most cells are inaccessible to microscopy and their number becomes quickly unmanageable. The 16 rounds of cell division required to produce a hatched *Drosophila* larva, for example, result in about 50,000 cells (1) and further rounds of division produce an adult with approximately 10^6 cells. The bodies of mice and humans consist of 10^{10} to 10^{14} cells respectively (2).

Recently it was proposed that naturally occurring somatic mutations, which accumulate in cells during the lifetime of an organism, could be used as lineage markers to reconstruct its entire cell lineage (3, 4). This is directly analogous to the use of heritable mutations, accumulating through time, to reconstruct a species phylogeny. While this approach is theoretically possible (3), it is nevertheless limited by the enormous challenge of detecting these rare mutations within the genomes of individual cells.

As a solution to the problem of reading the mutations, several recent papers have explored the idea of using CRISPR-induced somatic mutations, targeted to artificial sequences inserted as transgenes into the genome (termed "CRISPR recorders") (5–14). The recorders consist of arrays of CRISPR target sites, targeted by their cognate sgRNAs and Cas9 during development. Starting in early embryogenesis, CRISPR-induced mutations occur stochastically at those target sites, in each cell of the body, and these mutations are stably inherited by the progeny of these cells. In most cases, the mutation destroys the match between target and sgRNA meaning a mutated target is immune to further change. At the end of development only the recorder sequence has to be read rather than the whole genome; the accumulated mutations can then be used as phylogenetic characters allowing

the reconstruction of a tree of relationships between all cells (Figure 1).

The basic principle of recorder-based lineage tree reconstruction is easy to grasp. What is far less clear is whether the lineages produced by these methods are accurate enough for us to draw meaningful conclusions from them. Of course the required accuracy will depend on the intended use of the lineage, but to date there seems to have been minimal consideration of how accurate the lineages produced might be.

The ideal way of assessing the accuracy of these techniques would be to compare the real cell lineage of an organism against the lineage inferred by the recorder (11). This is difficult to implement in practice, however, because in most cases the real cell lineages are unknown.

We have taken the alternative approach of computationally simulating the processes of cell division and accumulation of mutations in a recorder and then comparing the lineage inferred from the recorder to the known *in silico* reference tree. We have used this approach to estimate the accuracy of lineage reconstruction in different situations (type and complexity of recorder, mutation rates, cell lineage depth, etc.), taking into account empirical measures of mutation rates and frequencies of different mutational outcomes derived from *in vivo* experimental data from *Drosophila melanogaster*. While some previous studies used simulations to evaluate the reconstruction of small cell lineages, no study has attempted this on cell lineages of tens of thousands of cells (6, 11).

Different designs of CRISPR recorders have been implemented, including recorders that register point mutations on arrays of barcoded targets (GESTALT; McKenna et al. 5, Raj et al. 12), ones that rely on "collapsing" target arrays through deletions (MEMOIR; Frieda et al. 6), recorders that target identical target sites located on separate transgenes (ScarTrace and LINNAEUS; Junker et al. 7, Alemany et al. 10, Schmidt et al. 11, Attardi et al. 13, Spanjaard et al. 14) and ones that target the CRISPR gRNA itself (8, 9). In this work we have simulated the behaviour of the first two types of recorders, but the insights that we have gained should apply to all types of recorders. Ultimately, these simulations will allow us to establish a set of criteria for the optimal design of CRISPR-based lineage recorders, as well as to understand the limitations of these techniques when addressing real biological questions.

To assess the power of CRISPR-based lineage recorders in cell lineage reconstruction, we focus on the conditions required to reconstruct a cell lineage of ∼65,000 cells. This roughly corresponds to the size of the cell lineage of a *Drosophila* first instar larva, of a pharyngula stage zebrafish embryo, or a stage E8.0 mouse embryo (Lehner et al. 1, Kane 15, Kojima et al. 16, respectively).

Results

General description of the simulations. In our simulations, a cell is implemented as a vector of m targets. We begin each simulation with one cell, representing the fertilised egg, that has all its targets in an unmutated state. The initial cell then undergoes a series of cell divisions (d), growing into a population of N cells, where $N = 2^d$. Following each cell division, each unmutated target can mutate (with a given probability μ_d) to one of several possible mutated states. Once a target is mutated, it can no longer change, either to revert to the unmutated state or to transit to a new state (Figure 2A).

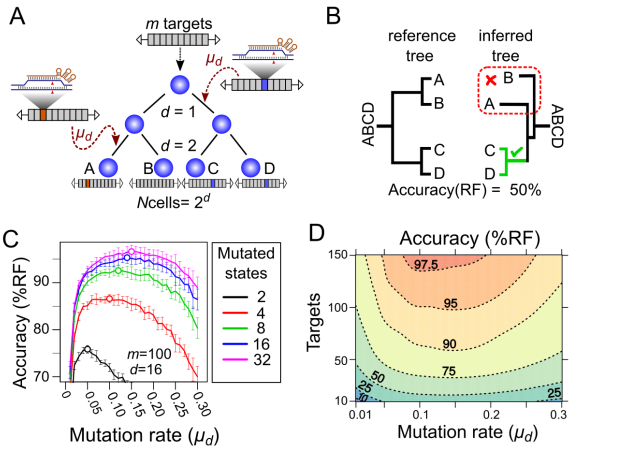

Fig. 2. Computational simulation of CRISPR recorders A. In our simulations, an initial cell with multiple CRISPR targets (m) yields N cells after a given number of cell divisions (d). The recorder accumulates independent CRISPR-induced mutations with a probability, in each target, of μ_d per cell division, which are inherited in subsequent cell divisions. The pattern of mutations accumulated in each cell is used to infer the lineage tree. **B.** The accuracy of lineage reconstruction was determined by comparing the inferred tree with the reference tree using the Robinson Foulds algorithm. The unmutated state of the recorder was used to root the tree. **C.** Accuracy of lineage reconstruction with a recorder of 100 CRISPR targets after 16 cell divisions (yielding 65,536 cells) over a range of mutation rates. Each line represents the mean accuracy (10 simulations) for simulations resulting in different numbers of equiprobable mutated states. The optimal mutation rate for each number of mutated states is indicated with an open circle. Vertical lines represent 95% confidence intervals. **D.** Accuracy of lineage reconstruction for different mutation rates and numbers of CRISPR targets. Mutations were set to result in 16 equiprobable mutated states. Dashed lines represent different accuracy thresholds (levelplot) after a LOESS regression. For each parameter combination, we plot the mean accuracy of 10 simulations after 16 cell divisions.

A unique label was given to each cell during the simulation. The sequence of simulated cell divisions were recorded in the form of a tree, whose topology describes the lineage relationships between all cells (the "reference tree"). At the end of the simulation, we randomly sampled a number of cells and we used the pattern of mutations accumulated in those cells to infer their cell lineage (the "inferred tree") using the Neighbor-Joining method (17) (for a comparison with parsimony see Figure Suppl. 6).

The accuracy of lineage reconstruction of each simulation was determined by comparing the inferred tree with the reference tree using the Robinson-Foulds algorithm (18), which calculates the fraction of splits in the reference tree that are precisely recovered in the inferred tree (Figure 2B). If the inferred tree is identical to the reference tree, the Robinson-Foulds accuracy is 100%. This provides a strict measure of the global accuracy of the inferred lineage tree. The accuracy of each lineage reconstruction was estimated as the mean accuracy of 10 subsamples of 1,000 cells.

Impact of mutation rate on the accuracy of cell lineage reconstruction. We simulated a lineage with a depth of 16 cell divisions ($d = 16$), yielding 65,536 cells (2^{16}). To deter-

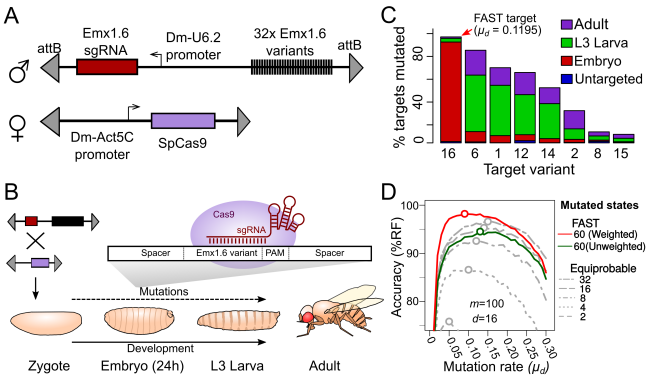

Fig. 3. Tuning the mutation rate of a CRISPR recorder *in vivo*. **A.** CRISPR recorder designed to test the mutation rates of 32 variants of the Emx1.6 target in *Drosophila*. The recorder consists of two transgenic constructs brought together by genetic crosses. The first construct carries an array of the 32 target variants and a transgene expressing the Emx1.6 sgRNA under the constitutive *Drosophila* U6.2 promoter (19). The second construct expresses the *Streptococcus pyogenes* Cas9 gene under the constitutive *Drosophila Act5C* promoter (19). **B.**Double heterozygotes carrying both constructs were collected at embryonic, late larval (L3) and adult stages and analysed for mutations in the target array by PCR amplification and sequencing of the recorder. **C.** Proportion of targets mutated at different stages, for the 8 most efficient target variants. "Untargeted" represents background mutations or sequencing errors observed in the absence of the Cas9 transgene. **D.** Estimates of cell lineage accuracy from computer simulations (as in Figure 2C) using the mutational outcomes observed *in vivo* on the FAST target.

mine the effect of varying the mutation rate on the accuracy of lineage reconstruction, we performed simulations with a recorder carrying 100 targets ($m = 100$) and a rate of mutation μ_d varying from 0.01 to 0.3 mutations per cell division per target. In parallel, we tested how the diversity of mutational outcomes (number of distinct mutated states) at each target could influence the accuracy of lineage reconstruction, by varying the number of possible mutational outcomes at each target between 2 and 32. In each case the different mutational outcomes were considered to be equiprobable.

As expected, these simulations show that the mutation rate and the diversity of mutations have a strong effect on the accuracy of cell lineage reconstruction (Figure 2C). A low diversity of possible mutational outcomes gives poorer results than a higher diversity. Mutation rates show a broad optimum between between 0.05 and 0.2 mutations per cell division per target; under these rates, 56-97% of target sites are mutated after 16 cell divisions. Lower mutation rates lead to more targets having no mutations, thus contributing no information for reconstructing the cell lineage. Higher mutation rates lead to most targets being mutated during the early cell divisions, leaving few targets available for recording later events.

In practice, CRISPR activity generates a range of mutations (mostly small deletions or insertions) at varying frequencies. We have measured the actual rates and diversity of CRISPR-induced mutations *in vivo* (see below) and used these empirical data in our subsequent simulations.

Tuning the mutation rate of a CRISPR recorder *in vivo*.

Our simulations show that specific mutation rates must be achieved experimentally in order to optimise cell lineage reconstruction. There are several ways to vary CRISPR mutation rates *in vivo*, including the use of different sgRNA:target pairs, varying the expression levels of sgRNA and Cas9, and using variants of sgRNA or Cas9 that influence their stability or activity. We chose to adjust the mutation rate by altering the target sequence in order to introduce mismatches in the sgRNA:target pairing; this is known to reduce the targeting efficiency (20, 21). We have measured the mutation rate of a series of different variants of a CRISPR target to find those with the optimum rates for cell lineage reconstruction of the *Drosophila* embryo.

We took advantage of a previous study by (20) who analysed the effects of sgRNA:target pairing mismatches on the efficiency of targeting a section of the human *EMX1* gene. Based on this study, we selected 32 variants of the Emx1.6 target (20), including the wild-type sequence and 31 variants with single- or double-nucleotide changes at different positions within the target sequence (Suppl. Table 1). To compare the mutation rates of the 32 targets, we designed and synthesised a single construct that carries all 32 variants in tandem. In the same construct, we incorporated a transgene constitutively expressing the Emx1.6 sgRNA under the *Drosophila* U6.2 promoter (19) (Figure 3A). We generated transgenic *Drosophila* lines carrying a single copy of this construct at the 37B7 locus, using ϕC31-mediated integration (22).

Males carrying the Emx1.6 sgRNA and the target array were crossed with virgin females carrying a constitutively expressed Cas9 transgene (Actin::Cas9, Port et al. 19) to generate progeny carrying a single copy of the CRISPR target array, the sgRNA and the Cas9 transgene. We collected these progeny at different developmental stages (end of embryogenesis, late L3 larvae, newly eclosed adults) to assess the number of mutations that had accumulated in each of the 32 target variants at these three stages (Figure 3B).

We pooled individuals collected at each of the three chosen developmental stages, performed PCR on genomic DNA and used high throughput sequencing to characterise the mutated targets. In individual animals, mutational frequencies are influenced both by the probability of each mutational outcome of CRISPR and by the clonal expansion of cells that carry each mutation. Since we have analysed populations of individuals and expect no systematic clonal biases associated with specific mutations in CRISPR recorders, we expect that our estimates of mutational frequencies largely reflect the frequencies of CRISPR-induced mutations on our target.

Our results confirm that, by employing different target variants, we can achieve widely different rates of mutation. As expected, the target that has perfect complementarity with the Emx1.6 sgRNA (target 16, named the "FAST" target) showed the highest mutation rate; having corrected for sequencing errors we observed that 87% of the targets carried a mutation at the end of embryogenesis (Figure 3C). This corresponds to a mutation rate of μ_d= 0.1195 per cell division, assuming a constant mutation rate per cell division (see later for consideration of uneven rates per cell division). The other targets showed lower mutation rates: in the six variants with the highest rates, μ_d ranged from $4\text{x}10^{-4}$ (target 15) to $6\text{x}10^{-2}$ (target 6) mutations per cell division (Suppl. Table 2).

The mutation rate of the FAST target (μ_d= 0.1195) falls within the optimal range we had estimated for reconstruct-

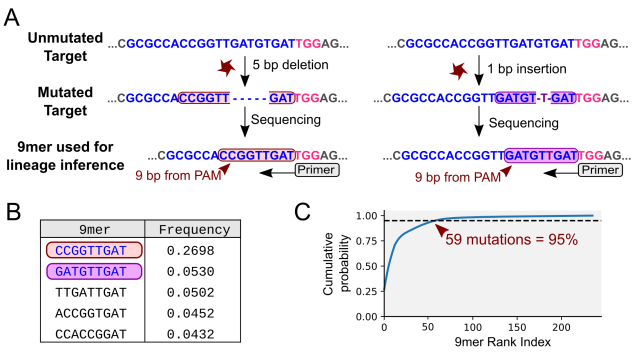

Fig. 4. Mutational outcomes of CRISPR *in vivo*. A. Examples of two mutational outcomes after CRISPR cleavage. The 9-nucleotide sequence located immediately upstream of the PAM (coloured box) captures most of the variation resulting from CRISPR-induced mutations. The target sequence is shown in blue, PAM sequence in pink, flanking sequence in grey. **B.** Relative frequencies of the five most common mutational outcomes in the FAST target. **C.** Cumulative probability of the mutational outcomes. 59 mutations account for 95% of the total number of mutations.

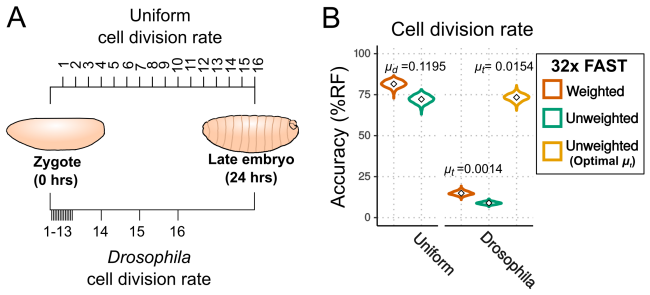

Fig. 5. Effects of cell division rate on lineage reconstruction. A. Scheme of the comparison between uniform and actual cell division rates in *Drosophila* embryos. **B.** Accuracy of lineage reconstruction under a uniform cell division rate (left) compared to rates that approximate those actually observed during *Drosophila* development (right) (25, 26), using mutation rates calculated from real experiments for the FAST target (μ_t= 0.0014), or optimised for increased accuracy of reconstruction (μ_t= 0.0154). Violin plots represent the distribution of reconstruction accuracies of 1000 simulations after 16 cell divisions. The accuracy of reconstruction using 32 FAST targets, with or without weighting of mutations, is represented in orange and green respectively. In yellow is the accuracy of 32 targets with an optimal μ_t (with no weighting).

ing the lineage of 65,536 cell embryos, assuming a uniform rate of cell division (see Figure 2C). Targets with slower mutation rates would be suited for lineaging past the embryonic stages. Conversely, faster mutation rates would be optimal for lineaging embryos at earlier stages, following fewer cell divisions. Instances of rapid or unequal rates of cell division would also require faster mutation rates (see below).

Simulating cell lineage reconstruction based on experimentally observed mutational outcomes. Thus far, our simulations assumed that the targets can mutate to a certain number of character states with equal probability. This assumption does not reflect the complexity of CRISPR mutagenesis observed *in vivo*. Our sequencing data for the 32 variants of the Emx1.6 target in *Drosophila* provide empirical measurements not only of the rate of mutation but also of the diversity of different mutational outcomes and their relative frequencies in a CRISPR recorder *in vivo*. Using these data we refined our simulations using the real set of mutational outcomes and their observed relative frequencies.

We focused on the complexity of mutational outcomes affecting the FAST target. As reported in previous studies (23, 24), we found that most of the mutations were located close to the Cas9 editing site. This suggests that most of the mutational information can be extracted by reading the nucleotides surrounding the editing site. Focusing on the 9 bp adjacent to the PAM sequence (Figure 4B) we observed >200 mutated states. The frequencies of mutations follow an exponential curve, with a few variants occurring at high frequency (Figure 4C), in contrast to a naive assumption of equiprobable mutational outcomes.

Using the observed distribution of these 9mers and the estimated overall μ_d, we carried out 1,000 simulations of the mutational process in a hypothetical construct carrying 32 identical FAST targets. We used 32 targets because we have shown that synthesising and generating transgenic flies with such a construct is feasible. For convenience, we considered 61 out of the ∼200 observed states: 59 states representing the 59 most common mutations (which account for 95% of the total observed mutations; see Figure 4C), a 60th state with a frequency of 5% that accounts for all other outcomes combined, and the 61st state representing the unmutated target.

Using the experimentally measured distribution of mutational outcomes, the accuracy of cell lineage reconstruction is 72% (see Figure 5B). Rarely occurring mutations are less likely to appear independently in the same target in more than one branch of the lineage tree (an instance of homoplasy), suggesting that rare mutations are better lineage markers than more frequent mutations. To take advantage of this we introduced a weighting scheme whereby mutations are weighted in inverse proportion to their frequency of occurrence (see Methods). Following this approach, the accuracy of lineage reconstruction using the same 61 states improved from 72% to 82% (Figure 5B).

Impact of uneven rates of cell division on the accuracy of cell lineage reconstruction. So far we have assumed that the probability of mutation per available target (μ_d) is the same in every cell division. This would be a reasonable assumption if all cells have a similar rate of cell division and if that rate remains constant during the course of development. In many species, however, the rate of cell division in embryogenesis varies among cells and through time. Early *Drosophila* embryos, for example, initially go through a series of 13 rapid and near-synchronous nuclear divisions to generate a uniform syncytial blastoderm (25); during this phase the duration of each mitotic cycle is very short (∼10 minutes). After cellularisation at cell cycle 14, the rate of cell division in the embryo slows considerably and becomes non-uniform (26, 27).

To estimate the impact of uneven rates of cell division on cell lineage reconstruction we modelled the mutation events as a Poisson process dependent on time rather than on cell divisions. A Poisson process assumes that a given event (in this case a CRISPR-induced mutation) occurs stochastically at a given rate μ_t. We estimated that setting μ_t at 0.0014 mutations per site per minute would produce the observed proportion of mutated FAST targets (87%) after 24 hours of

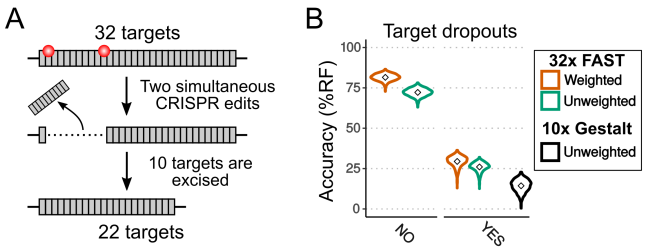

Fig. 6. Effects of dropouts on lineage reconstruction. A. Schematic representing how targets can be dropped out by simultaneous CRISPR edits. **B.** Accuracy of lineage reconstruction without dropouts (left) or with the presence of dropouts (right), using a μ_d= 0.1195. Violin plots represent the distribution of reconstruction accuracies of 1000 simulations after 16 cell divisions. The accuracy of reconstruction using 32 FAST targets, with or without weighting of mutations, is represented in orange and green respectively. In black is the accuracy when simulating the 10 targets of the Gestalt v7 construct (with no weighting).

embryonic development. We set the cell division intervals to approximate those known from *Drosophila* development (see Methods) and we modelled the frequency and diversity of mutational outcomes on those observed in the FAST target (see previous section). Under these conditions, we would expect the accuracy of lineage reconstruction to be considerably worse, as there will be many fewer mutations accumulated in the rapid early cell cycles, and indeed the accuracy fell to just 9% (without using the weighting scheme; see Figure 5B).

We hypothesized that the optimal value of μ_t would be different in this scenario of unequal cell divisions: that a higher μ_t should improve the accuracy of the reconstructed lineage because it would help to lineage the rapid early cell cycles. To test this hypothesis we performed simulations using different values of μ_t. The results show that the accuracy did indeed improve with increasing rates of mutation (Figure Suppl. 1), with a maximum accuracy of 73% at μ_t= 0.0154 (11 times higher than the optimal rate for embryos with a uniform rate of cell division; Figure 5B). These results suggest that higher mutation rates are needed for high lineaging accuracy when the rates of cell division are uneven.

Modelling the effects of target dropouts. Given the number of targets needed to reconstruct a cell lineage accurately (Figure 2D), lineage recorders must include arrays of tens or hundreds of targets. CRISPR activity affecting multiple targets simultaneously, in the same cell, can result in deletions of the DNA between these targets (see Figure 6). Such deletions could remove multiple targets, hampering accurate cell lineage reconstruction. We modelled the potential impact of these "dropouts" on the accuracy of lineage reconstruction, by conducting simulations (with uniform cell divisions, μ_d = 0.1195 and m=32, as before) under a scenario in which every time two or more targets were mutated in a given cell at a given cell cycle, we removed all the targets located between them (see Methods). We find that dropouts have a major impact on the accuracy of lineage reconstruction (Figure 6B); after 16 cell divisions, the accuracy dropped from 72% to 26% (or from 82% to 29% with weighted mutational outcomes).

Optimising cell lineage reconstruction for in situ sequencing with 2, 4 or 16 character states. Besides the biological constraints that influence our ability to reconstruct the cell lineage based on CRISPR recorders (mutation rates, diversity of CRISPR mutations, rates of cell division, target dropouts), there are technical constraints that currently limit our ability to read the information contained in these recorders. Thus far, our simulations have assumed that we can reliably read up to 9 nucleotides surrounding each target site over tens of targets, from individual cells. This can be achieved in dissociated single cells using modern high-throughput sequencing technologies (10, 12, 14).

Ideally, CRISPR-based lineage recorders could also be used in combination with spatially resolved sequencing (*in situ* sequencing), so that lineage information of single cells could be recorded together with their exact position in the developed embryo. Achieving accurate sequencing of multiple nucleotides in tens of targets in cells *in situ* is currently impractical, however, less ambitious *in situ* approaches have been proposed. The MEMOIR approach (6) has addressed this by employing single molecule *in situ* hybridization to distinguish mutated from unmutated targets.

In MEMOIR, only two character states can be detected per target ("scratchpad"), mutated versus unmutated. Moreover, successive rounds of *in situ* hybridization are needed to interrogate many distinct targets, which places a constraint on the number of targets that can be read. (6) have shown that 3 targets can be read per hybridization and up to 9 rounds of hybridization are feasible (6); thus, reading 2 character states per target over ∼30 targets seems to be achievable by the MEMOIR approach.

We carried out simulations to test how MEMOIR would perform using 32 targets, a mutation rate μ_d=0.1195 and a readout of 2 character states ("mutated" or "unmutated"). We find that the accuracy is only 4%. Even with an optimal mutation rate resulting in 50% target saturation (6) the accuracy of lineage reconstruction would be only ∼15% (data not shown).

In the future, *in situ* sequencing methods could be developed to interrogate the sequence of each target. These methods would be subject to different technical constraints than MEMOIR. Thus far, *in situ* sequencing efforts have mostly been based on sequencing by ligation and used the SOLiD sequencing technology (28, 29), which uses consecutive ligations of fluorescent oligonucleotides to interrogate pairs of dinucleotides in the target sequence sequentially (30)). The SOLiD colour code is degenerate, as 4 colours are used to represent all 16 possible DNA dinucleotides.

As a first step we wanted to explore the SOLiD parameter space extensively, to determine how the number of targets (m) and mutation rates (μ_d) affect the accuracy of lineage reconstruction when reading each target with one SOLiD ligation/detection cycle (only 4 character states). For this, we performed 10 simulations over a range of values for μ_d (from 0.01 to 0.3 mutations per cell division) and m (from 10 to 300 targets). In these simulations we assumed that the 4 possible mutated states were equiprobable and used the complete inferred tree to estimate the accuracy. Our results show that

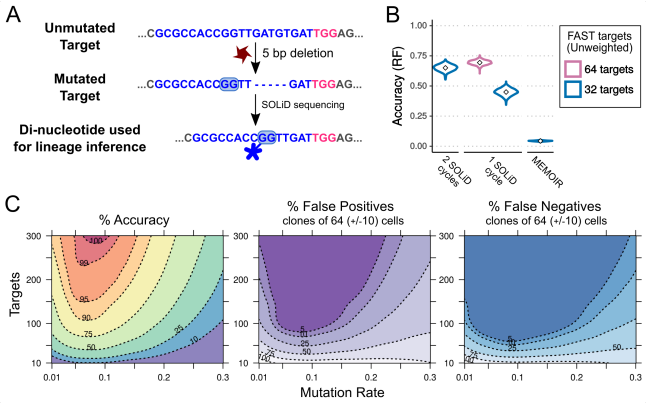

Fig. 7. Combining CRISPR lineaging with in situ sequencing. A. The most common mutational outcome of the FAST target is shown. The coloured box highlights the most informative dinucleotide position to read by SOLiD sequencing (6-7bp from PAM after CRISPR cleavage) and its SOLiD colour code (see Figure Suppl. 2). Sequence colours as in Figure 4. **B.** Accuracy of lineage reconstruction after sequencing with 2 SOLiD sequencing reads (left), 1 SOLiD read (center) and as in MEMOIR (right) using a μ_d= 0.1195. In blue and pink are the accuracy of a construct with 32 and 64 FAST targets, respectively. **C.** Accuracy of lineage reconstruction using in situ SOLiD sequencing, for different mutation rates and numbers of CRISPR targets, using a μ_d= 0.1195 and assuming equiprobable colour frequencies after 1 SOLiD read; Robinson Foulds global accuracy (left), false positives (center) and false negatives (right). Dashed lines represent different accuracy thresholds (levelplot) after a LOESS regression. For each parameter combination, we used the mean accuracy of 10 simulations after 16 cell divisions.

the optimal mutation rate for lineage reconstruction by this approach lies between 0.05 and 0.12 mutations per cell division, and that is possible to get up to 99% accuracy with 260 targets or more.

In SOLiD sequencing, the number of ligation/detection cycles that can be performed is limited by photodamage of the target amplicons and by the time required to perform this type of sequencing (10 days for 30 ligation cycles, Lee et al. 31). The practical upper limit on the number of SOLiD cycles that can be performed is therefore currently in the order of 30-60 cycles. Given these constraints, it is important to optimise the sequencing strategy, so as to maximise the amount of sequence information obtained for a given number of SOLiD sequencing cycles. We can ask, for example, whether it would be preferable to perform a single ligation/detection cycle on 64 targets rather than two ligation/detection cycles on 32 targets Given the experimentally measured spectrum of CRISPR-induced mutations on the targets, we can also determine which nucleotides of the target we should interrogate in order to extract the most information.

We determined that positions 6-7 bp 5' from the PAM sequence yield the most equiprobable colour frequencies for the FAST target (Figure 7A), minimising homoplasy in the observed character states (see Figure 5). The frequency of each mutated state was determined by the real frequency of mutations observed (see above) and the overall frequency of mutation was set to μ_d= 0.1195 per cell division. We note that the unmutated state (red) is indistinguishable from one of the four mutated states. With 4-character states, homoplasy will arise frequently from convergent appearance of the same colour (even arising from different mutated states) in independent cells. Our results show that, with a single SOLiD read, using a recorder with 32 targets, the mean accuracy of reconstructed cell lineages is 45% (Figure 7B).

Clearly, increasing the number of targets will improve performance, but we wanted to know whether it would be better instead to double the number of reads per target, which represents the same sequencing effort. We found that the reconstruction accuracy obtained by performing 1 SOLiD sequencing cycle on 64 FAST targets is higher (69%) than performing 2 SOLiD cycles on 32 FAST targets (65%) (Figure 7C). For the second SOLiD cycle we used the positions 11-12 bp 5' from the PAM sequence, as in SOLiD the sequentially interrogated dinucleotide pairs are typically separated by 5 nucleotides (31).

Interpreting accuracy in terms of correctly assigning cells to clones. As an alternative measure of tree accuracy, which could be more useful when thinking about the clonal composition of tissues, we also estimated the proportion of false positive and false negative assignments of cells to clones in the reconstructed cell lineage (Figure 7C). False positives were defined as the proportion of cells that are erroneously assigned to a given cell clone. Conversely, false negatives were defined as the proportion of cells that are not assigned to a cell clone to which they belong. Our measurements of false positives and false negatives were performed on clones of 64($\pm$10) cells, as described in the Methods section.

Using 60 targets, a single SOLiD read per target and a mutation rate of $\mu_d = 0.08$, we find that a global (Robinson-Foulds) accuracy of 70% corresponds to 13% false positives and to 3% false negatives in $\sim$64-cell clones.

Assessing the accuracy of an existing recorder. Recently, a number of lineaging approaches using CRISPR recorders have been tested in the nematode *Caenorhabditis elegans* and the zebrafish *Danio rerio*, as well as in cultured human cells (5, 6, 11). We have used our simulation approach to assess the accuracy of GESTALT, one of the first and most ambitious approaches, which aimed to reconstruct the cell lineage of the tens of thousands of cells of the zebrafish embryo (5). It is important to note that in GESTALT the cell lineage is reconstructed at a coarse-grained level, with clones (instead of cells) as nodes in the tree, whereas our measure of success assesses the ability to reconstruct the complete cell lineage at a single-cell level.

GESTALT uses arrays of 10 different CRISPR targets, mutated by injecting fertilised eggs with 10 corresponding sgRNAs and Cas9. The mutated targets are then sequenced at different developmental stages. We based our simulations on the mutational outcomes derived from the GESTALT recorder v7 at 30 hours post-fertilisation (downloaded from the Dryad repository). At this stage the zebrafish embryo consists of approximately 25,000 cells, resulting from $\sim$15 rounds of cell division.

In our simulations, we assumed a constant mutation rate (per cell division), which, as we have shown, will probably overestimate of the accuracy of the inferred lineage. For each of the 10 CRISPR targets, we estimated the mutation rate (μ_d) necessary to obtain the fraction of mutated targets observed

after 15 cell divisions (Figure Suppl. 3). The estimated mutation rate ranges from ∼0.01 (for target 10) to ∼0.23 (for target 7) per cell division.

The v7 GESTALT construct shows a high incidence of target dropouts which were modelled as previously described. The mutational process was modelled as a gamma distribution of 60 possible mutated states (Figure Suppl. 3), with frequencies closely approximating the observed distribution of mutations reported in the GESTALT publication (see Methods for detail). We compared the number of different alleles (i.e., unique combinations of mutated targets) obtained in the simulated and the experimental results; the mutational complexity used in our simulations generated a number of alleles that closely approximates the experimentally observed number (see Methods for more details).

We performed 1,000 simulations and inferred the cell lineage of 1,000 randomly sampled cells from each simulation. We find that the mean accuracy of the GESTALT approach is just 14% after 16 cell divisions (Figure 6B). This means that this implementation of GESTALT is not suited for reconstructing a complete, accurate cell lineage.

Discussion

The use of CRISPR-induced somatic mutations is emerging as an attractive approach for reconstructing complex cell lineages. A variety of CRISPR-based lineage recorders has been developed to test this approach (5–14). If the results of these methods are to be useful for gaining biological insights, however, it is essential that the inferred lineage trees are sufficiently reliable, i.e. that they accurately reconstruct the real cell lineages of the organism. The potential accuracy of the trees inferred using these methods has not yet been established.

We have used simulations of the process of cell division and the accumulation of mutations across a lineage tree covering tens of thousands of cells, to examine the effects of different factors on the accuracy of a reconstructed tree. Our simulations allowed us to look at the influence of different rates of mutation on the CRISPR targets, of different designs of lineage recorders and of how mutations could be read experimentally. We have also investigated the effects of irregular cell divisions, target deletions following simultaneous double-stranded cuts and the variable mutational outcomes of the CRISPR process itself.

Unsurprisingly, the accuracy of lineage reconstruction largely rests on the quantity and quality of lineage information carried by the recorders, which is influenced by several factors. Although it is obvious that the accuracy of the lineage tree depends on the number of CRISPR targets in the recorder, our results serve to place strict upper limits on the level of accuracy that we can expect from CRISPR recorders. Under ideal conditions (optimized mutation rates, uniform cell divisions, fully sequenced targets), 30 targets are sufficient to reach an overall tree accuracy of ∼70% for a lineage of ∼65,000 cells; 100 targets would yield trees that have an accuracy above 90% (Figure 2D). If we were only able to take a single 4-colour SOLiD read per target, more than 200 targets would be required to get a highly accurate (>95%) tree (Figure 7C).

A second important requirement is to match the mutation rate to the rate of cell division; mutation rates that are too low will leave many cell divisions unmarked, while mutations that accumulate too rapidly will quickly saturate the targets and leave very few available to record later cell divisions. The range of mutation rates that can produce accurate lineage reconstruction fortunately proves to be quite broad for a given tree size; 0.05 to 0.25 mutations per cell division can yield reasonably high levels of accuracy for trees of ∼65,000 cells, if the division rates are relatively even (Fig. 2C). Alongside the number of targets, mutation rates are an attribute of the experiment that can be adjusted. Rates can potentially be increased by increasing the expression levels of the CRISPR effectors, or decreased by introducing mismatches between the sgRNA and the CRISPR targets, as we have shown experimentally (Figure 3C).

The information carried by CRISPR recorders is also influenced by the diversity of the experimentally observed mutations accumulating at each CRISPR target. Our observations of CRISPR mutations in *Drosophila* show that these are biased towards a small number of frequently observed outcomes. Simulation shows how targets that accumulate a broad set of more equiprobable mutations generate more reliable trees (Figure 2C). If, as expected, the diversity of mutations and their relative frequencies vary depending on the target sequence (23, 24) sampling different targets to approach this optimum would be worthwhile.

Some factors affecting tree reconstruction accuracy are outside of experimental control, but simulating their effects can nevertheless show which responses can successfully mitigate them. We have shown, for example, that uneven rates of cell division across the tree require faster mutation rates and/or larger numbers of targets, to provide sufficient coverage during the fastest divisions. In an extreme case, such as the *Drosophila* embryonic lineage where 13 of the 16 cell divisions take place at a very high rate (1 cell division every ∼10 minutes), the optimum mutation rate proves to be >10 times higher than in an equivalent tree with uniform division rates. Even with this optimised mutation rate, the potential accuracy of lineage reconstruction with a given number of targets is much lower (Figure 5A).

Besides the intrinsic limitations imposed by CRISPR mutagenesis, the information that we can obtain from each CRISPR target is further constrained by our ability to read and to discriminate between the mutational outcomes. As an obvious goal would be to sequence the mutated targets in individual cells *in situ*, we have explored the specific case of obtaining a single 4-colour SOLiD sequencing read per target. It is encouraging to find that accurate lineage reconstruction is still possible given a sufficient number of targets (Figure 7C).

Finally, we show the degree to which the accuracy of lineage reconstruction is sensitive to loss of information caused by the loss of targets through deletion resulting from simultaneous cleavage at two sites (Figure 6B). While we have used

the most pessimistic estimate of the frequency of dropouts - assuming that every pair of targets cleaved in the same cell would lead to a deletion of the intervening targets in the array - data from GESTALT suggest that target dropouts are frequent when mutation rates are high (5). The strong deleterious effect of dropouts that we observe in simulations highlights the need to address this issue. The problem of dropouts could be reduced by opting for the lower end of the optimal range of mutation rates; or eliminated by targeting separate loci in the genome rather than arrays of targets.

Available implementations of CRISPR type recorders are based on different conceptual designs: barcoded arrays recording point mutations (5, 12), "collapsing" arrays (6), targets distributed in different genomic locations (7, 10, 11, 13, 14) and mutations induced by self-targeting guide RNAs (8, 9). Here we have simulated the first two types of recorders, but we expect that the insights that we have gained on the importance of optimising mutation rates, target numbers and the complexity of character states will apply to all types of recorders.

Our analysis suggests that most of the CRISPR recorders published to date, which rely on at most 10 CRISPR targets (5–7, 10–12), yield trees of very low overall accuracy and lineage resolution. While these recorders must, nevertheless, carry lineage information of lower resolution, it is sensible to interpret the results from these recorders in the light of this expected low level of accuracy.

A simulation-guided design of lineage recorders, taking into account the specific parameters of each experimental system, is essential. We hope our study will encourage the general use of simulations of lineage recorders, with the aim of testing their limits, adjusting their design and improving their performance. This approach should stimulate the development of a new generation of CRISPR recorders that could finally allow the reconstruction of accurate cell lineages of complex multicellular organisms at the level of a single cell.

Material and Methods

CRISPR recorder and sequencing.

Design and synthesis. We designed a DNA construct containing an array of 32 targets of the human Emx1.6 sgRNA (20), including the wild-type Emx1.6 target sequence and 31 variants carrying 1 or 2 mismatches and/or an alternative PAM sequence (see Suppl. Table 1). To facilitate synthesis of this construct, between each pair of targets we introduced 80 bp spacers, harbouring unique sequences which would be recognised by specific primers (Suppl. Table 3). We optimised these spacer sequences *in silico* to minimise the presence of repetitive sequences. Unique KpnI and NotI cloning sites were included at either end of the array to help with subsequent cloning steps.

We designed a second plasmid carrying the KpnI and NotI restriction sites and the *Drosophila* U6.2 promoter driving expression of the Emx1.6 sgRNA (20) using a standard sgRNA scaffold: GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAG (19, 20). This DNA sequence was flanked by two attB sites. The *Drosophila* U6.2 promoter has been shown by previous studies to produce lower levels of CRISPR activity when compared to the U6.1 and U6.3 promoters (19).

Both constructs were synthesised by Biomatik (Ontario, Canada) using standard gene synthesis techniques. The CRISPR target array was excised from the first construct by KpnI-NotI digestion and subcloned into the KpnI and NotI sites of the second plasmid.

Fly transgenesis, genetics and strains. The construct carrying the CRISPR target array and U6.2::Emx1.6 sgRNA was inserted via recombinase-mediated cassette exchange (22) into the 2nd chromosome of *Drosophila melanogaster* (acceptor strain # 27387) using a commercially available service (BestGene Inc., U.S.A.).

Homozygous Act-5C-Cas9 females (Bloomington stock # 54590) were crossed with homozygous males carrying the CRISPR target array (Bloomington stock # 54590), set to lay eggs over 30 minute intervals in order to obtain synchronised egg collections, and the progeny were collected at different developmental stages (24h embryos, third instar larvae, recently hatched adults). As negative controls, to account for sequencing errors, we used adults carrying the CRISPR target array (in heterozygous condition), but lacking the Cas9 transgene.

DNA extraction, generation of libraries and sequencing. For DNA extraction and sequencing, we pooled approximately 100 embryos, 10 larvae or 20 adults (10 males and 10 females). We extracted genomic DNA by phenol chloroform extraction followed by alcohol precipitation, and generated libraries by PCR using primers with extended adapter sequences ("fusion PCR") barcoded by condition (see Suppl. Table 3) for sequencing on Ion Torrent Personal Genome Machine (PGM, Life Technologies). As the maximum PGM read length is 400 bp and each target repeat in our construct is 100 bp long, we amplified the repeats in 10 groups of 3 units (amplicons 1-10), plus a group of 2 units (amplicon 11). Amplicons were mixed in equimolar amounts and the final pooled mix was sequenced on the PGM sequencer with a 318 v2 chip, as well as a calibration standard to enhance the read quality.

Filtering of sequencing data. The 7,347,400 reads obtained were de-multiplexed by condition and trimmed to meet quality standards using the Phred software included in the seqtk_trimfq package of the Galaxy software (32). We next eliminated sequencing reads that were shorter than 100 bp, lacked the 5' primer sequence, or lacked a target-specific sequence of 11-20 bp downstream of the target (including the PAM) using a custom Python script. In each sequencing read, we used the 9bp adjacent to the PAM sequence (9mer) to determine whether a target was mutated (the results are shown in Suppl. Table 1).

We quantified sequencing errors (with a custom Python script) by analysing the target sequences of adult flies car-

rying the CRISPR recorder and the sgRNA but not carrying Act-Cas9 ("untargeted" condition): in these animals we expect any differences from the unmutated state to reflect sequencing errors. In target 16 (FAST target) we found two frequent sequencing errors (single nucleotide deletions) downstream of the target; we decided to include the reads carrying these errors. Targets 17 and 18 did not yield a sufficient number of good quality reads, and targets 13, 21, and 23 showed a high proportion of sequencing errors (Suppl. Table 1).

Estimating mutation rate and mutational complexity.

FAST target. We estimated the mutation rate of the FAST target based on the proportion of targets that were mutated at the end of embryonic development (86.95%) using a custom Python script based on the geometric cumulative distribution function. The mutation rate of μ_d=0.1195 mutations per cell division produces the observed saturation of 86.95% after 16 cell divisions.

We modelled the mutational outcomes of the FAST target based on the mutational outcomes observed at the end of embryonic development (>200 distinct 9mers with frequencies following an exponential curve; see Figure 4C). We considered that a mutation would result in a change to one of 59 states with a probability reflecting the observed occurrence of the 59 most frequent real mutations (95% of the total; see Figure 4C) or to a 60th state with a probability of 0.05.

GESTALT. To analyse the accuracy of GESTALT, we used data from the v7 construct at the 30 hours post-fertilisation stage (available at https://datadryad.org/resource/doi:10.5061/dryad.478t9). These consist of six biological replicates. The v7 construct contains 10 different CRISPR targets that were targeted with 10 different sgRNAs.

For each biological replicate we quantified the frequency of mutations and dropouts in each target (Figure Suppl. 3) using a custom Perl script. We considered any deletion greater than 26 bp to be a dropout, as this would affect more than one target (each target is 23 bp). For each target, we quantified saturation as the proportion of reads of the target that were mutated. For each target, we estimated the mutation rate per cell division (μ_d) necessary to produce the level of saturation (proportion of mutated targets) observed after 15 cell divisions, assuming that mutations follow a geometric distribution. The estimated mutation rate ranges from ∼0.01 (target 10) to ∼0.23 (target 7) (Figure Suppl. 3).

The mutational complexity varied between targets and replicates, from ∼25 to ∼200 different mutations per target. In all cases, however, their frequencies followed an exponential curve, with one mutation usually accounting for 20-30% of the total reads and with the majority of the mutations observed only rarely. For each target we modelled the mutational outcome as 60 different mutations with frequencies sampled from a random gamma distribution, with shape parameter κ=0.1 and scale parameter θ=2, which approximate the observed distribution (see Figure Suppl. 3D).

Computer simulations. Computer simulations were performed using Matlab v2017a (Mathworks, 2017) and are available at (https://github.com/irepansalvador/CRISPR_recorders_sims). CRISPR mutations were simulated following a geometric or a poisson distribution.

Simulating mutation events using a geometric distribution. To simulate mutations using a geometric distribution, the probability of mutation was the same for all targets per cell division. Given a mutation rate μ_d (per cell division), the probability that a site remains unmutated after d cell divisions is $(1-\mu_d)^d$. Thus, we can determine the mutation rate μ_d from the proportion of targets that are mutated after a given number of cell divisions.

Simulating mutation events using a Poisson distribution. Under the Poisson model, given a mutation rate μ_t (per minute), the probability that a site remains unmutated after t minutes is: $e^{-(\mu_t t)}$). Thus, we can determine the mutation rate μ_t from the proportion of targets that are mutated after a given amount of time. The time interval for each cell division was set to approximate the rates of cell division in early *Drosophila* embryos: for the first 13 cell divisions the interval was set to 10 minutes, and to 130 minutes per division for the last 3 divisions (25, 26).

Simulation of target dropouts. For the dropouts simulations, if any two targets were hit during a given cell division, all the targets between them were removed. When three or more targets were hit during the same cell division, two were selected randomly and the intervening targets were removed. In subsequent phylogenetic analyses, dropouts were treated as missing data.

Simulations of GESTALT. We performed 1,000 simulations with the estimated μ_d for each target over 16 cell divisions. We accounted for dropouts as described previously. To test whether our simulations match the experimental results in terms of mutational complexity, we compared the number of "alleles" (unique combinations of mutated targets) found in the experimental and in the simulated data.

Our simulations encompassed 15 cell divisions, yielding 32,768 cells, which approximates the 30 hpf zebrafish embryo (∼25,000 cells). For each simulation, we took 100 random samples of 10,000 cells and counted the number of alleles in each sample. Our simulated samples produced an average of 3,409 alleles (s.d.= 952 alleles; see Figure Suppl. 3), compared to the 1,000-2,500 alleles found in the experimental data (5).

Analysis of simulated targets. The main outcome of each simulation was a T matrix of size $N \times m$, for N cells and m targets. This matrix is equivalent to a DNA alignment with sequences as rows and DNA positions as columns. For most simulations, 10 random samples of 1,000 cells were chosen for lineage reconstruction and for assessing the accuracy of the reconstructed cell lineage. A "root" taxon with unmutated character states was added to the alignment prior to the

lineage inference. For some simulations, we inferred cell lineages using all cells after 16 cell divisions (N =65,536) and found that their global accuracy was similar to that when subsampling 1,000 cells (Figure Suppl. 4).

For the target dropouts simulations we added to the T matrix an extra character for each distinct dropout of one or more targets that was shared between ≥32 cells (character state "1" if present, "0" if absent). This was done to take advantage of the information coming from shared target dropouts.

Cell lineage inference.

Reconstructing lineage trees using Neighbor Joining (PAUP*). Most cell lineages were inferred using the Neighbor-Joining method (NJ). We used the Neighbor joining algorithm as implemented in the PAUP* software (version 4.0a build 158; Swofford 33). In PAUP*, up to 64 character states can be specified, with the possibility of giving different weights to the occurrence of specific mutations. We used a substitution matrix based on the frequency of each mutation. The matrix has size $s \times s$ for s number of states, where the distance from state i to state j is weighted according to the natural logarithm of the inverse of their frequencies (34) with the equation:

$$d(i,j) = \begin{cases} 0, & \text{if } i = j \\ log(\frac{1}{Freq_j}), & \text{if } i = unmutated \\ log(\frac{1}{Freq_i}), & \text{if } j = unmutated \\ log(\frac{1}{Freq_i}) + log(\frac{1}{Freq_j}), & \text{otherwise} \end{cases}$$

where $Freq_i$ and $Freq_j$ are the frequencies of states i and j respectively.

In simulations where we modelled dropouts, an extra character state was assigned to each cell containing a dropout that was shared by ≥32 cells. For these simulations, the distance matrix was applied to the m original targets and for the extra characters the following distance was applied:

$$d(i,j) = \begin{cases} 0, & \text{if } i = j \\ 100, & \text{otherwise} \end{cases}$$

Reconstructing lineage trees using FastTree. When inferring complete cell lineage trees in the simulations of SOLiD sequencing data (N =65,536 cells), we used an heuristic method that approximates the Maximum Likelihood approach, implemented by the FastTree software (35). FastTree was chosen for its ability to infer trees from large alignments, consisting of tens of thousands of sequences, and for doing so very efficiently.

Reconstructing lineage trees using Maximum Parsimony (PAUP*). The use of parsimony for the cell lineage reconstruction of our simulations was not practical for the thousands of cells/taxa we consider. Nevertheless, to assess the relative performance of NJ and Maximum Parsimony in the context of lineage data, we compared the two methods using trees of 100 randomly sampled cells (Figure Suppl. 6). We calculated the accuracy of lineage reconstruction with NJ and parsimony methods on 10 separate simulations, based on the mutational frequencies of FAST targets, 4 character states and a mutation rate of 0.1195 over 16 cell divisions. For the parsimony analysis, we used the Camin-Sokal model (i.e., irreversible mutated states) and a substitution matrix based on character states frequencies (as used in GESTALT). For the NJ analysis we used a weighting scheme based on character states frequencies.

Tree accuracy estimation.

Robinson-Foulds algorithm. The accuracy of each cell-lineage reconstruction was determined by calculating the Robinson-Foulds distance (RF) between the reference and the inferred trees. For this task we used the CompareTree software (CompareTree.pl is available at http://www.microbesonline.org/fasttree/treecmp.html). RF is 1 when the inferred and reference trees are identical however for easier comprehension we report RF as a percentage of identical splits, instead of a fraction.

Calculating false positives and false negatives. False positives (FP) and false negatives (FN) were calculated by comparing the reference tree (R) with the inferred tree (I) using the newick-tools software (36). False positives were measured by counting the proportion of cells that need to be pruned from a branch of the inferred tree to match a given branch the reference tree. More formally, the false positives were estimated as follows (see Figure Suppl. 5):

1) We extracted the x number of subtrees in R that contained 64 (±10) cells (subtrees $R'_{(1-x)}$)

2) Then for each R' subtree we find the subtree from I (I' subtree) that includes all the cells present in R'. The FP is then calculated with the following equation:

$$FP_{(R,I)} = \frac{1}{x}\sum_{i=1}^{x}\frac{[I'_i] - [R'_i]}{[I'_i]}$$

where $[R'_i]$ and $[I'_i]$ are the number of cells in trees R'_i and I'_i respectively.

False negatives were measured by counting the proportion of cells that need to be pruned from a branch of the reference tree to match a given branch of the inferred lineage tree. More formally, the false negatives were estimated as follows (see Figure Suppl. 5):

1) We extracted the x number of subtrees from I that contained 64 (±10) cells (subtrees $I'_{(1-x)}$).

2) Then for each I' subtree we extracted the subtree from R (R' subtree) that included all cells from the I' tree. The FN is calculated then with the following equation:

$$FN_{(R,I)} = \frac{1}{x}\sum_{i=1}^{x}\frac{[R'_i] - [I'_i]}{[R'_i]}$$

where $[R'_i]$ and $[I'_i]$ are the number of cells in trees R'_i and I'_i respectively.

ACKNOWLEDGEMENTS

We thank Sandrine Hughes and Benjamin Gillet for expert assistance in sequencing the CRISPR target arrays, Tomas Flouri for helping implementing the false positives/negatives analysis, Je Hyuk Lee, Jessica Svedlund and Mats Nilsson for discussions on *in situ* sequencing, Nikos Konstantinides, Pierre Martinez, James Cotterell, Rosa Barrio, Michael Akam, Isaac Salazar-Ciudad and Richard Copley for critical comments on the manuscript, and the High Performance Computing platform at the UCL Department of Computer Science. This research was supported by a grant from the Human Frontier Science Programme (HFSP RGP0002/2016).

Bibliography

1. C F Lehner, H W Jacobs, K Sauer, and C A Meyer. Regulation of the embryonic cell proliferation by *Drosophila* cyclin D and cyclin E complexes. In Gregory R. Bock, Gail Cardew, and Jamie A. Goode, editors, *The Cell Cycle and Development*, volume 237, pages 43–54. John Wiley and Sons, Wiley, Chichester, 2001. ISBN 0471496626.
2. Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, 14(8):e1002533, aug 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002533.
3. Dan Frumkin, Adam Wasserstrom, Shai Kaplan, Uriel Feige, and Ehud Shapiro. Genomic variability within an organism exposes its cell lineage tree. *PLoS Computational Biology*, 1 (5):e50, 2005. ISSN 1553-734X. doi: 10.1371/journal.pcbi.0010050.
4. Stephen J Salipante and Marshall S Horwitz. Phylogenetic fate mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(14):5448–53, apr 2006. ISSN 0027-8424. doi: 10.1073/pnas.0601265103.
5. Aaron McKenna, Gregory M Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907–aaf7907, jul 2016. ISSN 0036-8075. doi: 10.1126/science.aaf7907.
6. Kirsten L. Frieda, James M. Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K. Chow, Zakary S. Singer, Mark W. Budde, Michael B. Elowitz, and Long Cai. Synthetic recording and *in situ* readout of lineage information in single cells. *Nature*, 541(7635):107–111, nov 2016. ISSN 0028-0836. doi: 10.1038/nature20777.
7. Jan Philipp Junker, Bastiaan Spanjaard, Josi Peterson-Maduro, Anna Alemany, Bo Hu, Maria Florescu, and Alexander van Oudenaarden. Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars. Technical report, 2016.
8. Reza Kalhor, Kian Kalhor, Kathleen Leeper, Amanda Graveline, Prashant Mali, and George M Church. A homing CRISPR mouse resource for barcoding and lineage tracing. *bioRxiv*, page 280289, mar 2018. doi: 10.1101/280289.
9. Samuel D. Perli, Cheryl H. Cui, and Timothy K. Lu. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*, 353(6304):aag0511–aag0511, 2016. ISSN 0036-8075. doi: 10.1126/science.aag0511.
10. Anna Alemany, Maria Florescu, Chloé S. Baron, Josi Peterson-Maduro, and Alexander Van Oudenaarden. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556 (7699):108–112, mar 2018. ISSN 14764687. doi: 10.1038/nature25969.
11. Stephanie Tzouanas Schmidt, Stephanie M. Zimmerman, Jianbin Wang, Stuart K. Kim, and Stephen R. Quake. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synthetic Biology*, mar 2017. ISSN 2161-5063. doi: 10.1021/acssynbio.6b00309.
12. Bushra Raj, Daniel E Wagner, Aaron McKenna, Shristi Pandey, Allon M Klein, Jay Shendure, James A Gagnon, and Alexander F Schier. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36(5):442–450, mar 2018. ISSN 15461696. doi: 10.1038/nbt.4103.
13. Andrea Attardi, Tim Fulton, Maria Florescu, Gopi Shah, Leila Muresan, Jan Huisken, Alexander van Oudenaarden, and Benjamin Steventon. Neuromesodermal progenitors are a conserved source of spinal cord with divergent growth dynamics. *bioRxiv*, page 304543, apr 2018. doi: 10.1101/304543.
14. Bastiaan Spanjaard, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature Biotechnology 2018*, apr 2018. ISSN 1546-1696. doi: 10.1038/nbt.4124.
15. Donald A. Kane. *Cell cycles and development in the embryonic zebrafish.* Methods in Cell Biology Series. Academic Press, 1999. ISBN 9780125441612.
16. Yoji Kojima, Oliver H. Tam, and Patrick P.L. Tam. Timing of developmental events in the early mouse embryo. *Seminars in Cell & Developmental Biology*, 34:65–75, oct 2014. ISSN 1084-9521. doi: 10.1016/J.SEMCDB.2014.06.010.
17. N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, jul 1987. ISSN 1537-1719. doi: 10.1093/oxfordjournals.molbev.a040454.
18. D.F. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, feb 1981. ISSN 0025-5564. doi: 10.1016/0025-5564(81)90043-2.
19. Fillip Port, Hui-Min Chen, Tzumin Lee, and Simon L Bullock. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in *Drosophila*. *Proceedings of the National Academy of Sciences*, 111(29):E2967—-E2976, 2014.
20. Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao, and Feng Zhang. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, 31(9):827–832, jul 2013. ISSN 1087-0156. doi: 10.1038/nbt.2647.
21. Yanfang Fu, Jeffry D Sander, Deepak Reyon, Vincent M Cascio, and J Keith Joung. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnology*, 32(3):279–284, mar 2014. ISSN 15461696. doi: 10.1038/nbt.2808.
22. Jack R Bateman, Anne M Lee, and Others. Site-specific transformation of *Drosophila* via ϕC31 integrase-mediated cassette exchange. *Genetics*, 173(2):769–777, 2006.
23. Megan van Overbeek, Daniel Capurso, Matthew M. Carter, Matthew S. Thompson, Elizabeth Frias, Carsten Russ, John S. Reece-Hoyes, Christopher Nye, Scott Gradia, Bastien Vidal, Jiashun Zheng, Gregory R. Hoffman, Christopher K. Fuller, and Andrew P. May. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Molecular Cell*, 63 (4):633–646, aug 2016. ISSN 10974164. doi: 10.1016/j.molcel.2016.06.037.
24. Giang T.H. Vu, Hieu X. Cao, Friedrich Fauser, Bernd Reiss, Holger Puchta, and Ingo Schubert. Endogenous sequence patterns predispose the repair modes of CRISPR/Cas9-induced DNA double-stranded breaks in *Arabidopsis thaliana*. *Plant Journal*, 92(1):57–67, oct 2017. ISSN 1365313X. doi: 10.1111/tpj.13634.
25. M Zalokar and I Erk. Division and migration of nuclei during early embryogenesis of *Drosophila melanogaster*. *Journal de Microscopie et de Biologie Cellulaire*, 25(2):97–&, 1976.
26. V E Foe. Mitotic domains reveal early commitment of cells in *Drosophila* embryos. *Development (Cambridge, England)*, 107(1):1–22, sep 1989. ISSN 0950-1991.
27. Volker Hartenstein. *Atlas of Drosophila development.* Cold Spring Harbor Laboratory Press, 1993.
28. J. H. Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L Yang, Thomas C Ferrante, Richard Terry, Sauveur S F Jeanty, C. Li, Ryoji Amamoto, D. T. Peters, Brian M Turczyk, Adam H Marblestone, Samuel A Inverso, A. Bernard, Prashant Mali, Xavier Rios, J. Aach, and G. M. Church. Highly multiplexed subcellular RNA sequencing *in situ*. *Science*, 343(6177):1360–1363, mar 2014. ISSN 0036-8075. doi: 10.1126/science.1250212.
29. Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. *In situ* sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, jul 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2563.
30. Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, Joel A Malek, Gina Costa, Kevin McKernan, Arend Sidow, Andrew Fire, and Steven M Johnson. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7):1051–63, jul 2008. ISSN 1088-9051. doi: 10.1101/gr.076463.108.
31. Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, 10(3):442–458, feb 2015. ISSN 1754-2189. doi: 10.1038/nprot.2014.191.
32. Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Greg Von Kuster, Eric Rasche, Nicola Soranzo, Nitesh Turaga, James Taylor, Anton Nekrutenko, and Jeremy Goecks. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, 44(W1):W3–W10, jul 2016. doi: 10.1093/nar/gkw343.
33. D.L. Swofford. *PAUP* Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0a.* Sinauer Associates, Sunderland, Massachusetts, 2017.
34. Joseph Felsenstein. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, 16(3):183–196, nov 1981. ISSN 10958312. doi: 10.1111/j.1095-8312.1981.tb01847.x.
35. Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, mar 2010. ISSN 19326203. doi: 10.1371/journal.pone.0009490.
36. Tomas Flouri, Alexandros Stamatakis, and Paschalia Kapli. newick-tools: a novel software for simulating and processing phylogenetic trees, 2018.

Supplementary Figures and Tables

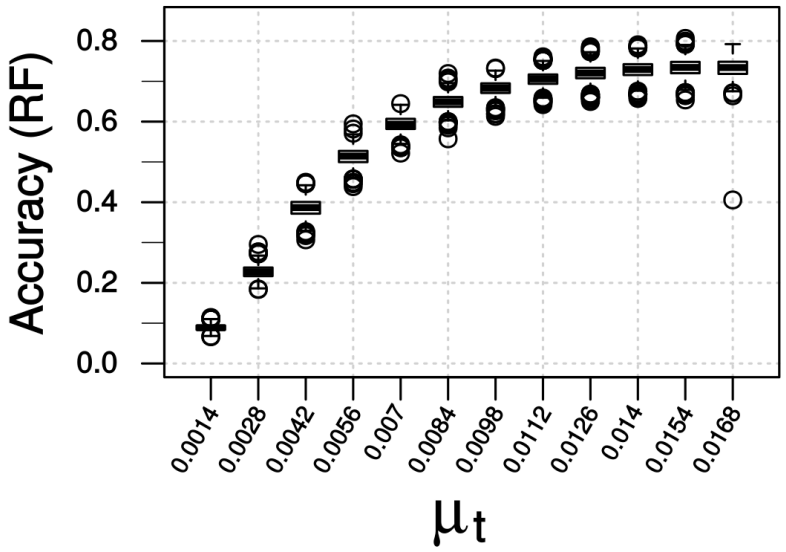

Fig. Suppl. 1. Finding the optimal mutation rate for the real rates of cell division in *Drosophila* embryos
Accuracy of lineage reconstruction is given for different mutation rates (μ_t). Simulations were performed to approximate *Drosophila*'s known cell division rate over 16 cell divisions, under a Poisson model. Boxplots represent the distribution of 1,000 simulations.

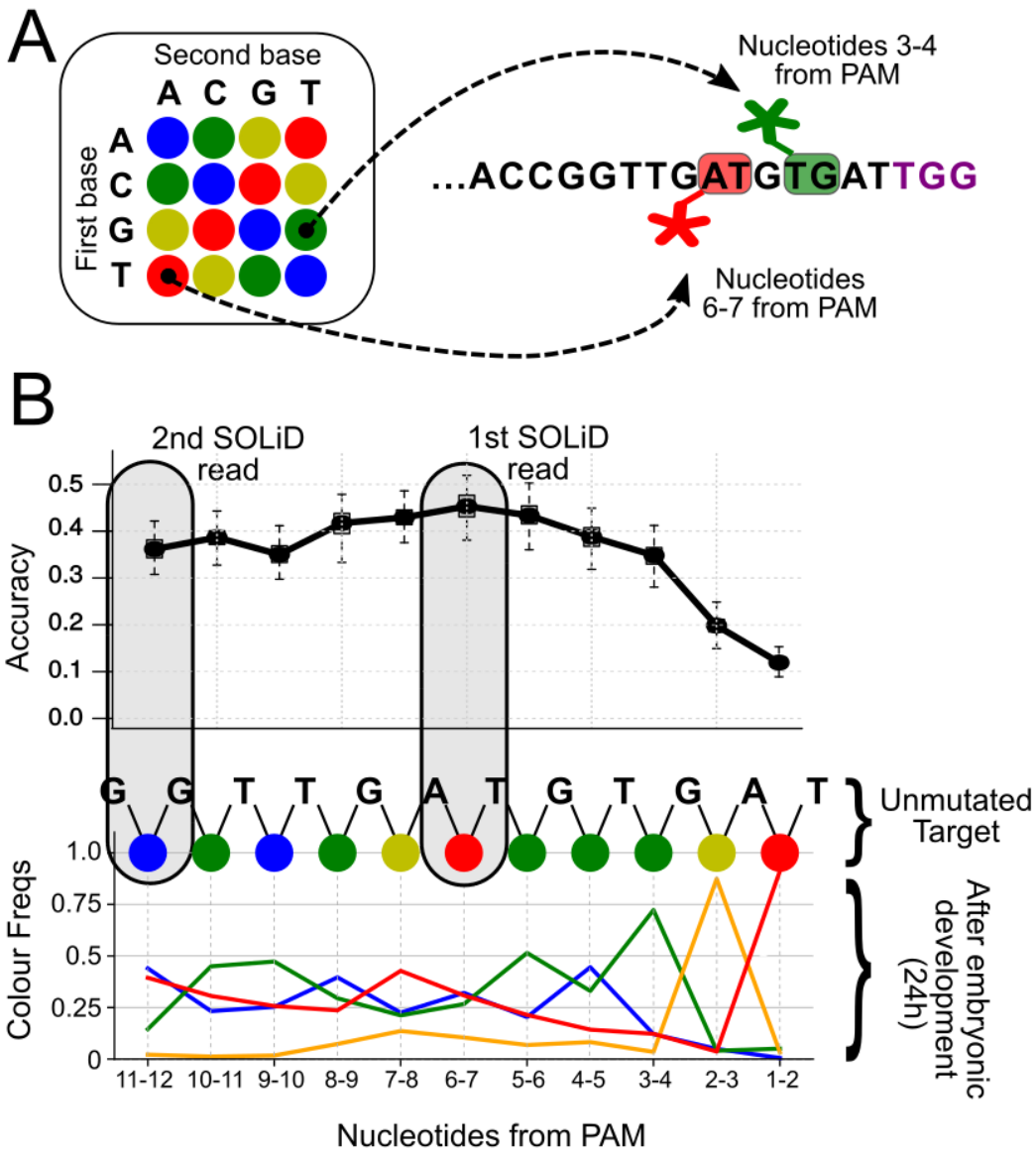

Fig. Suppl. 2. Distribution of SOLiD sequencing outcomes on the FAST target, to identify the most informative sites
A. SOLiD colour-space coding. Each dinucleotide-specific probe is labelled with 1 of 4 fluorescent markers. The colour code for the 16 possible dinucleotides is shown on the left. The outcome of interrogating two different dinucleotides in the unmutated target is shown in the right; positions 3-4 (green) and 6-7 (red) from the PAM sequence (in purple). **B.** The frequencies of experimentally observed mutational outcomes on the FAST target are shown, using the SOLiD colour code. For reference, the sequence of colours for the unmutated target is shown at the top. The dinucleotide located 6-7 nucleotides upstream of the PAM is the most informative for lineage reconstruction. The grey boxes highlight the dinucleotides used for the SOLiD simulations.

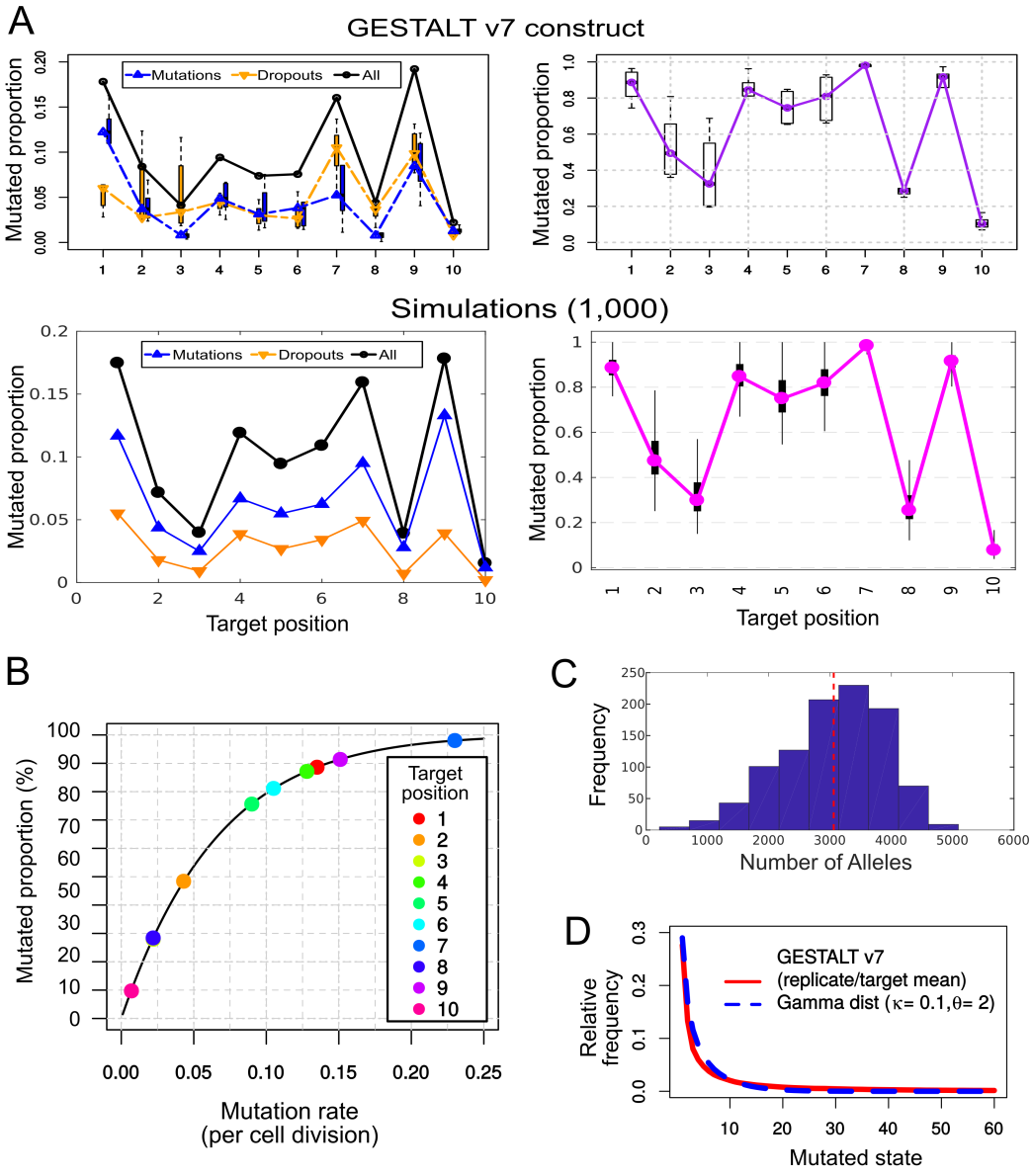

Fig. Suppl. 3. Simulating the mutational outcomes of the GESTALT v7 recorder.

A. Comparison between the observed target saturation of the GESTALT v7 recorder (top) and our simulations (bottom). Left: Relative frequency of mutations and dropouts affecting each target after 15 cell divisions. Right: Proportion of targets (remaining after dropouts) carrying a mutation.

B. Mutation rate (μ_d) necessary to produce the proportion of mutations observed in each target after 15 cell divisions, assuming a geometric distribution.

C. Histogram of the number of "alleles" found per simulation, in 1,000 GESTALT simulations. The red dashed line represents the mean number of alleles per simulation. 100 samples of 10,000 cells were analysed per simulation.

D. The relative frequencies of the 60 most common mutated states (mean values for all replicates and targets, in red) follow a gamma distribution with shape parameter κ= 0.1 and scale parameter θ= 2 (in blue).

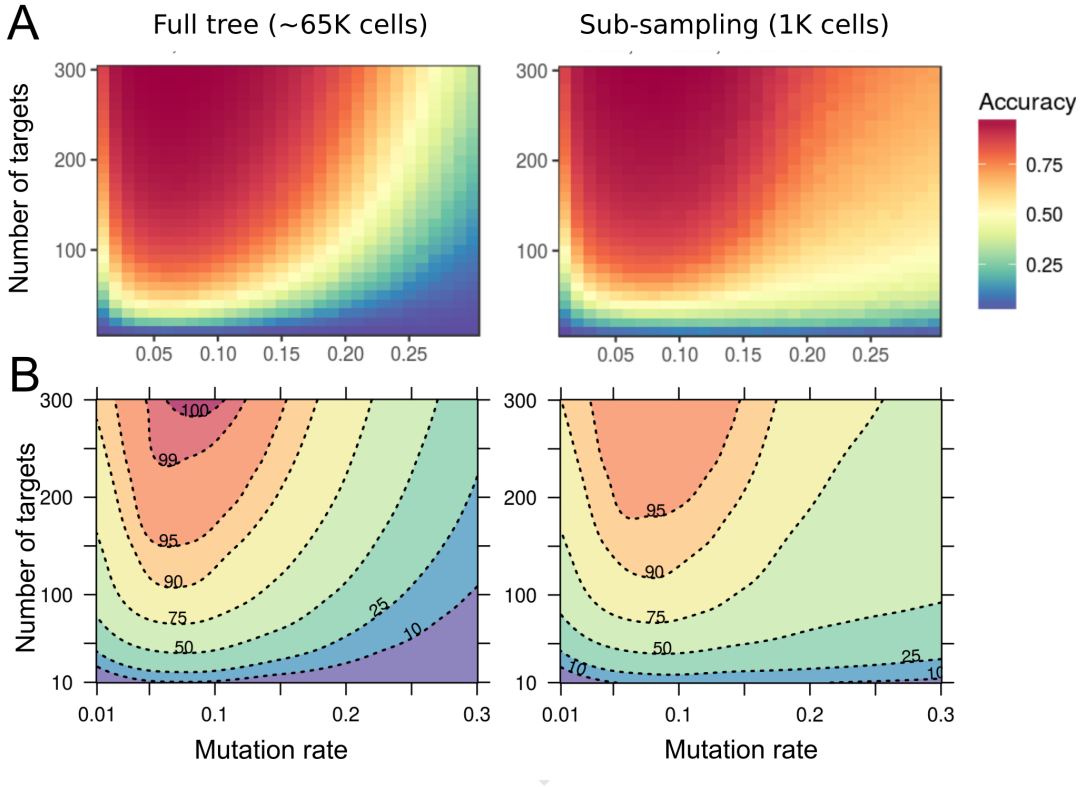

Fig. Suppl. 4. Accuracy of lineage reconstruction using a single-read of SOLiD sequencing.
A. Accuracy of lineage reconstruction, for different mutation rates and numbers of CRISPR targets, after a single read of SOLiD sequencing at positions 6-7 of the FAST target (see Suppl. Figure Suppl. 2). For each parameter combination, accuracy values represent the average of 10 simulations.
B.Accuracy thresholds after applying a Loess Regression on the same data.

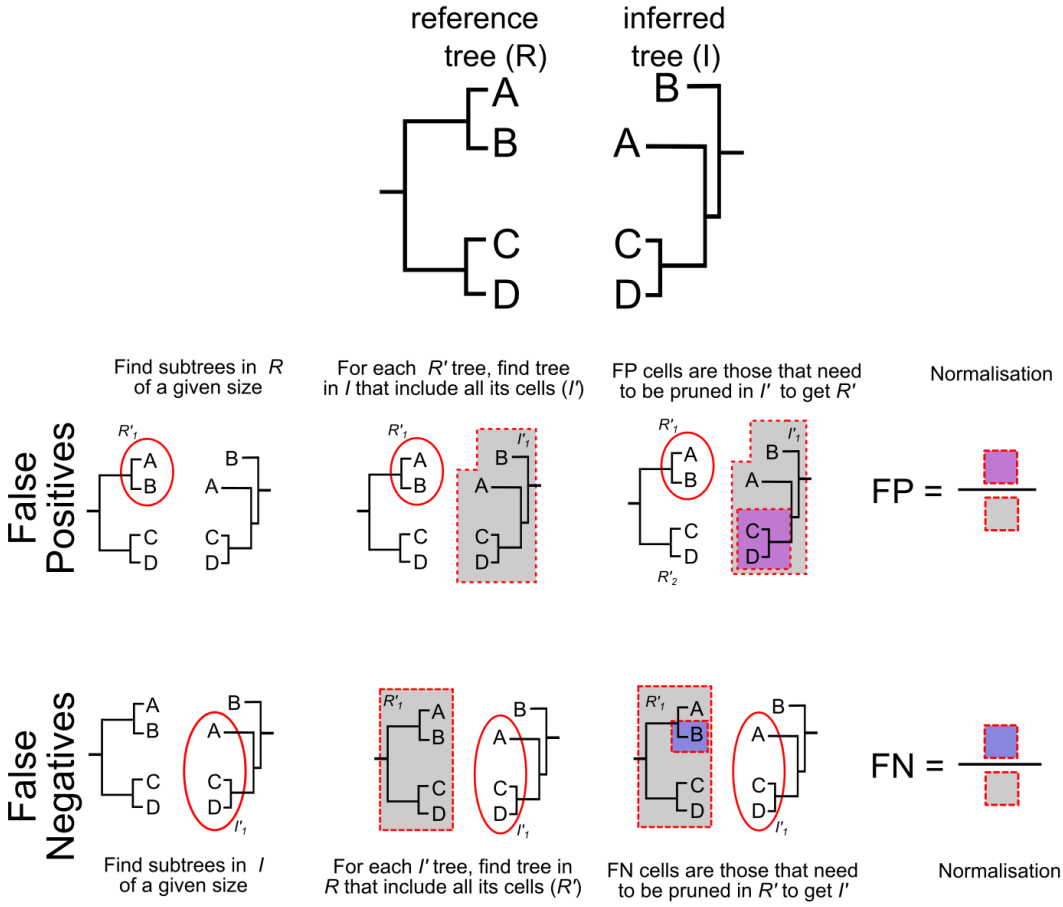

Fig. Suppl. 5. Method to estimate false positives and false negatives
False positives were measured by counting the proportion of cells that need to be pruned from a branch of the inferred tree to match a given branch in the reference tree. Similarly, false negatives were measured by counting the proportion of cells that need to be pruned from a branch of the reference tree to match a given branch of the inferred lineage tree. For details see Methods.

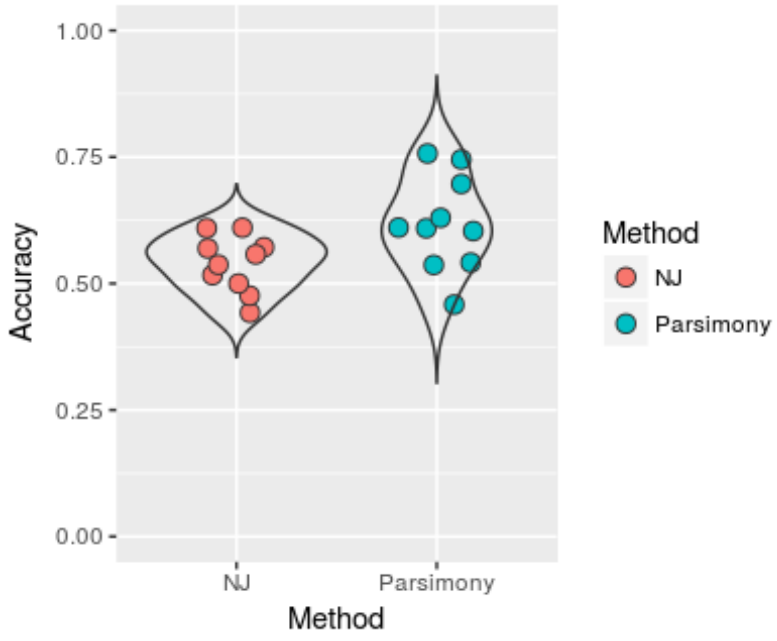

Fig. Suppl. 6. Comparing the performance of Neighbor Joining and Maximum Parsimony in lineage reconstruction
Violin plots show the distribution of accuracy of lineage reconstruction in 10 simulations, using single SOLiD reads on the FAST target (4 character states) across 100 subsampled cells. The simulations were performed with 32 targets and a mutation rate of μ_d=0.1195 per cell division, over 16 cell divisions. For the Neighbor Joining, we used a weighting scheme based on character state frequencies. For the Maximum Parsimony, we used the Camin-Sokal model (irreversible mutated states) and probabilities based on character state frequencies. The implementation of Maximum Parsimony is similar to the one used in the GESTALT approach (5)

		Untargeted			Embryo			Larva			Adult		
Target	Sequence	Mutated	Total Reads	%Mutated	Mutated	Total Reads	%Mutated	Mutated	Total Reads	%Mutated	Mutated	Total Reads	%Mutated
1	gtgccaccggttgatatgattgg	121	99226	0.12	6121	96608	6.34	75720	149974	50.49	102245	155140	65.90
2	gcatcaccggttgatgtgattgg	346	87861	0.39	2498	81823	3.05	15940	129651	12.29	35264	124852	28.24
3	gcgctgccggttgatgtgattgg	420	32248	1.30	644	34267	1.88	1586	78040	2.03	4002	88117	4.54
4	gcgccattggttgatgtgattgg	1535	99480	1.54	2478	104618	2.37	4618	210606	2.19	5009	226852	2.21
5	gcgccaccaattgatgtgattgg	190	93798	0.20	200	96093	0.21	1195	191108	0.63	1607	200975	0.80
6	gcgccaccggccgatgtgattgg	86	10165	0.85	1177	11931	9.87	18926	31850	59.42	37834	46591	81.20
7	gcgccaccggttgatgtggctgg	61	18337	0.33	79	20498	0.39	93	42249	0.22	365	46009	0.79
8	gcgccaccggttgatgtgactgg	197	14782	1.33	416	17171	2.42	2161	35512	6.09	3494	36278	9.63
9	gcgccaccggctgatgtgactgg	586	16165	3.63	720	19268	3.74	1430	38407	3.72	1268	38662	3.28
10	gcgccgccggttgatgtgactgg	1533	64904	2.36	2199	72060	3.05	4717	149498	3.16	5104	194043	2.63
11	acgccaccggttgatgtgactgg	226	50987	0.44	312	57962	0.54	1075	113185	0.95	3274	141471	2.31
12	gcgccaccggttgatgtaattgg	854	44266	1.93	3259	45624	7.14	26607	62918	42.29	44721	72469	61.71
13	gcgtcaccggttgatgtaattgg	7658	63954	11.97	12712	71980	17.66	30353	151170	20.08	26080	102783	25.37
14	gcgccaccggttgatatgattgg	81	52995	0.15	2042	54820	3.72	31647	91934	34.42	33493	69298	48.33
15	gcgccatcggttgatatgattgg	236	52048	0.45	569	55779	1.02	4459	110828	4.02	6149	82185	7.48
16	gcgccaccggttgatgtgattgg	61	5300	1.15	9528	10777	88.41	10349	11293	91.64	20107	21644	92.90
17	gcgccaccggttgatgtgattag	0	3	0.00	0	4	0.00	0	2	0.00	0	2	0.00
18	gcatcaccggttgatgtgattag	0	0	0.00	1	15	6.67	0	3	0.00	1	47	2.13
19	gcgctgccggttgatgtgattag	194	79514	0.24	168	72777	0.23	438	183610	0.24	325	127976	0.25
20	gcgccattggttgatgtgattag	224	63655	0.35	231	56115	0.41	480	145963	0.33	388	98924	0.39
21	gcgccaccaattgatgtgattag	1081	14275	7.57	1012	11992	8.44	2621	32583	8.04	1961	23600	8.31
22	gcgccaccggttgatgtggctag	187	22862	0.82	170	19125	0.89	399	49154	0.81	447	55250	0.81
23	gcgccaccggttgatgtgactag	1308	10558	12.39	1209	9965	12.13	2788	24616	11.33	3328	28413	11.71
24	gcgccaccggctgatgtgactag	78	10639	0.73	49	9684	0.51	218	23365	0.93	212	22416	0.95
25	gcgccgccggttgatgtgactag	49	31929	0.15	80	29019	0.28	120	45388	0.26	134	72243	0.19
26	acgccaccggttgatgtgactag	61	29512	0.21	46	25542	0.18	76	39436	0.19	122	65409	0.19
27	gcgccaccggttgatgtaattag	281	33129	0.85	251	28662	0.88	379	44048	0.86	634	72444	0.88
28	gcgtcaccggttgatgtaattag	2423	94592	2.56	4050	81706	4.96	7880	192762	4.09	7081	213946	3.31
29	gcgccaccggttgatatgattag	75	60028	0.12	94	52281	0.18	168	124714	0.13	162	141312	0.11
30	gcgccatcggttgatatgattag	731	45879	1.59	607	39918	1.52	1526	96792	1.58	1687	105807	1.59
31	gtgccaccggttgatatgattag	75	90061	0.08	41	47984	0.09	114	123187	0.09	141	88167	0.16
32	gcgccaccggccgatgtgattag	560	73178	0.77	265	37469	0.71	715	100420	0.71	572	71699	0.80

Supplementary Table 1.
Proportion of mutated targets (target saturation) for each of the 32 Emx1.6 target variants, sampled at different developmental stages (embryos, L3 larvae, adults) and in the absence of Cas9 (untargeted). Targets 13, 17, 18, 21 and 23 were not analysed further because there were no good quality reads in the untargeted condition or because the targets showed a high proportion of sequencing errors.

Target	Embryo Saturation (%)	Mutation rate (per cell division)
16	87.26	0.1195
6	9.02	0.006
1	6.21	0.004
12	5.21	0.003
14	3.57	0.0025
2	2.66	0.002
8	1.09	0.0006
15	0.57	0.0004

Supplementary Table 2.
Proportion of mutated targets (target saturation) in the embryo after correcting for sequencing errors, and estimated mutation rates per cell division, for the target variants showing the highest mutation rates.

Primer	Sequence
BC1_1F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATactgcctgcctgaagattacgagac
BC1_2F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATgaccctaactagacgaacttgacga
BC1_3F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATgattgagtagggaggagtatcacga
BC1_4F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATaacccgataacgacgaaacgagctt
BC1_5F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATaggagggttggaagtacggatatag
BC1_6F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATttgagaagatagacagaatatgcgc
BC1_7F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATccgagacgaactgacgaacctgtgc
BC1_8F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATcagttaagagaaagccccagtagta
BC1_9F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATagagagagagcccaaaattccgaga
BC1_10F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATtaatagccgtagtaaacaagtcgta
BC1_11F	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTAAGGTAACGATccaccgccagagatagagttacgac
BC2_1F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATactgcctgcctgaagattacgagac
BC2_2F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATgaccctaactagacgaacttgacga
BC2_3F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATgattgagtagggaggagtatcacga
BC2_4F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATaacccgataacgacgaaacgagctt
BC2_5F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATaggagggttggaagtacggatatag
BC2_6F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATttgagaagatagacagaatatgcgc
BC2_7F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATccgagacgaactgacgaacctgtgc
BC2_8F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATcagttaagagaaagccccagtagta
BC2_9F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATagagagagagcccaaaattccgaga
BC2_10F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATtaatagccgtagtaaacaagtcgta
BC2_11F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGAACGATccaccgccagagatagagttacgac
BC3_1F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATactgcctgcctgaagattacgagac
BC3_2F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATgaccctaactagacgaacttgacga
BC3_3F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATgattgagtagggaggagtatcacga
BC3_4F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATaacccgataacgacgaaacgagctt
BC3_5F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATaggagggttggaagtacggatatag
BC3_6F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATttgagaagatagacagaatatgcgc
BC3_7F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATccgagacgaactgacgaacctgtgc
BC3_8F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATcagttaagagaaagccccagtagta
BC3_9F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATagagagagagcccaaaattccgaga
BC3_10F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATtaatagccgtagtaaacaagtcgta
BC3_11F	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGATTCGATccaccgccagagatagagttacgac
BC4_1F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATactgcctgcctgaagattacgagac
BC4_2F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATgaccctaactagacgaacttgacga
BC4_3F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATgattgagtagggaggagtatcacga
BC4_4F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATaacccgataacgacgaaacgagctt
BC4_5F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATaggagggttggaagtacggatatag
BC4_6F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATttgagaagatagacagaatatgcgc
BC4_7F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATccgagacgaactgacgaacctgtgc
BC4_8F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATcagttaagagaaagccccagtagta
BC4_9F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATagagagagagcccaaaattccgaga
BC4_10F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATtaatagccgtagtaaacaagtcgta
BC4_11F	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGATCGATccaccgccagagatagagttacgac

Supplementary Table 3.(Continues in the next page)

Primer	Sequence
BC5_1F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATactgcctgcctgaagattacgagac
BC5_2F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATgaccctaactagacgaacttgacga
BC5_3F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATgattgagtagggaggagtatcacga
BC5_4F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATaacccgataacgacgaaacgagctt
BC5_5F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATaggagggttggaagtacggatatag
BC5_6F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATttgagaagatagacagaatatgcgc
BC5_7F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATccgagacgaactgacgaacctgtgc
BC5_8F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATcagttaagagaaagccccagtagta
BC5_9F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATagagagagagcccaaaattccgaga
BC5_10F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATtaatagccgtagtaaacaagtcgta
BC5_11F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CAGAAGGAAC</u>GATccaccgccagagatagagttacgac
BC6_1F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATactgcctgcctgaagattacgagac
BC6_2F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATgaccctaactagacgaacttgacga
BC6_3F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATgattgagtagggaggagtatcacga
BC6_4F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATaacccgataacgacgaaacgagctt
BC6_5F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATaggagggttggaagtacggatatag
BC6_6F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATttgagaagatagacagaatatgcgc
BC6_7F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATccgagacgaactgacgaacctgtgc
BC6_8F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATcagttaagagaaagccccagtagta
BC6_9F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATagagagagagcccaaaattccgaga
BC6_10F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATtaatagccgtagtaaacaagtcgta
BC6_11F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<u>CTGCAAGTTC</u>GATccaccgccagagatagagttacgac
1R	CCTCTCTATGGGCAGTCGGTGATacatccctcctcatcctcctcctct
2R	CCTCTCTATGGGCAGTCGGTGATcattcatcttcgggcgggcagtttc
3R	CCTCTCTATGGGCAGTCGGTGATcgtctcagggtacatcaggtcggtt
4R	CCTCTCTATGGGCAGTCGGTGATacgggatctctggagggctctacta
5R	CCTCTCTATGGGCAGTCGGTGATctcaggtgggcttcgttcagacttc
6R	CCTCTCTATGGGCAGTCGGTGATgggcggtaattcgggctctcttcta
7R	CCTCTCTATGGGCAGTCGGTGATttcggtcttctagtcaggcatcggg
8R	CCTCTCTATGGGCAGTCGGTGATggtcgtatctctcgggtatcagggc
9R	CCTCTCTATGGGCAGTCGGTGATctctggaacttcttcggatcggagg
10R	CCTCTCTATGGGCAGTCGGTGATcgaacgttgctggtgtcggactctt
11R	CCTCTCTATGGGCAGTCGGTGATgctttcacttcagggagtcgtcggt

BC1=Unmutated
BC2=Embryos
BC3=Larva3Male*
BC4=Larva3Female*
BC5=AdultMaleº
BC6=AdultFemaleº
*BC3 and BC4 data were pooled in the analysis phase
ºBC5 and BC6 data were pooled in the analysis phase

Supplementary Table 3.
PCR primers used for preparation of the sequencing libraries. Forward primers (F) carry adapter sequences (uppercase), barcodes specific for each condition (underlined, BC1 to BC6), and sequences annealing to the spacers of the repeat construct (lowercase). Reverse primers (R) carry adapters (uppercase) and sequences annealing to the spacers of the repeat construct (lowercase); see Figure 3B and Materials and Methods.