

1 **Whole organism lineage tracing by combinatorial and cumulative genome editing**

2

3 Aaron McKenna^{1†}, Gregory M. Findlay^{1†}, James A. Gagnon^{2†}, Marshall S. Horwitz^{1,3},
4 Alexander F. Schier^{2,4,5,6*}, Jay Shendure^{1,7*}

5

6 **Affiliations:**

7 ¹Department of Genome Sciences, University of Washington, Seattle WA, USA

8 ²Department of Molecular and Cellular Biology, Harvard University, Cambridge MA, USA

9 ³Department of Pathology, University of Washington, Seattle WA, USA

10 ⁴Center for Brain Science, Harvard University, Cambridge MA, USA

11 ⁵The Broad Institute of Harvard and MIT, Cambridge MA, USA

12 ⁶FAS Center for Systems Biology, Harvard University, Cambridge MA, USA

13 ⁷Howard Hughes Medical Institute, Seattle WA, USA

14

15 †These authors contributed equally to this work

16 *Correspondence to shendure@uw.edu and schier@fas.harvard.edu

17

18 **Abstract**

19

20 Multicellular systems develop from single cells through a lineage, but current lineage tracing
21 approaches scale poorly to whole organisms. Here we use genome editing to progressively
22 introduce and accumulate diverse mutations in a DNA barcode over multiple rounds of cell
23 division. The barcode, an array of CRISPR/Cas9 target sites, records lineage relationships in the
24 patterns of mutations shared between cells. In cell culture and zebrafish, we show that rates and
25 patterns of editing are tunable, and that thousands of lineage-informative barcode alleles can be
26 generated. We find that most cells in adult zebrafish organs derive from relatively few
27 embryonic progenitors. Genome editing of synthetic target arrays for lineage tracing
28 (GESTALT) will help generate large-scale maps of cell lineage in multicellular systems.

29

30 **Introduction**

31

32 The tracing of cell lineages was pioneered in nematodes by Charles Whitman in the 1870s, at a
33 time of controversy surrounding Ernst Haeckel's theory of recapitulation (1). This line of work
34 culminated a century later in the complete description of mitotic divisions in the roundworm *C.*
35 *elegans* - a tour de force facilitated by its visual transparency as well as the modest size and
36 invariant nature of its cell lineage (2).

37

38 Over the past century, a variety of creative methods have been developed for tracing cell lineage
39 in developmentally complex organisms (3). In general, subsets of cells are marked and their
40 descendants followed as development progresses. The ways in which cell marking has been
41 achieved include dyes and enzymes (4-6), cross-species transplantation (7), recombinase-
42 mediated activation of reporter gene expression (8, 9), insertion of foreign DNA (10-12), and
43 naturally occurring somatic mutations (13-15). However, despite many powerful applications,
44 these methods have limitations for the large-scale reconstruction of cell lineages in multicellular
45 systems. For example, dye and reporter gene-based cell marking are uninformative with respect
46 to the lineage relationships *between* descendent cells. Furthermore, when two or more cells are
47 independently but equivalently marked, the resulting multitude of clades cannot be readily
48 distinguished from one another. Although these limitations can be overcome in part with
49 combinatorial labeling systems (16, 17) or through the introduction of diverse DNA barcodes
50 (10-12), these strategies fall short of a system for inferring lineage relationships throughout an
51 organism and across developmental time. In contrast, methods based on somatic mutations have
52 this potential, as they can identify lineages and sub-lineages within single organisms (13, 18).
53 However, somatic mutations are distributed throughout the genome, necessitating whole
54 genome sequencing, (14, 15), which is expensive to scale beyond small numbers of cells and
55 not readily compatible with *in situ* readouts (19, 20).

56

57 What are the requirements for a system for comprehensively tracing cell lineages in a complex
58 multicellular system? First, it must uniquely and incrementally mark cells and their descendants

59 over many divisions and in a way that does not interfere with normal development. Second, these
60 unique marks must accumulate irreversibly over time, allowing the reconstruction of lineage
61 trees. Finally, the full set of marks must be easily read out in each of many single cells.

62
63 We hypothesized that genome editing, which introduces diverse, irreversible edits in a highly
64 programmable fashion (21), could be repurposed for cell lineage tracing in a way that realizes
65 these requirements. To this end, we developed genome editing of synthetic target arrays for
66 lineage tracing (GESTALT), a method that uses CRISPR/Cas9 genome editing to accumulate
67 combinatorial sequence diversity to a compact, multi-target, densely informative barcode.
68 Importantly, edited barcodes can be efficiently queried by a single sequencing read from each of
69 many single cells (Fig. 1A). In both cell culture and in the zebrafish *Danio rerio*, we
70 demonstrate the generation of thousands of uniquely edited barcodes that can be related to one
71 another to reconstruct cell lineage relationships. In adult zebrafish, we observe that the majority
72 of cells of each organ are derived from a small number of progenitor cells. Furthermore,
73 ancestral progenitors, inferred on the basis of shared edits amongst subsets of derived alleles,
74 make highly non-uniform contributions to germ layers and organ systems.

75

76 **Results**

77

78 **Combinatorial and cumulative editing of a compact genomic barcode in cultured cells**

79

80 To test whether genome editing can be used to generate a combinatorial diversity of mutations
81 within a compact region, we synthesized a contiguous array of ten CRISPR/Cas9 targets
82 (protospacers plus PAM sequences) separated by 3 base-pair (bp) linkers (total length of 257
83 bp). The first target perfectly matched one single guide RNA (sgRNA), while the remainder
84 were off-target sites for the same sgRNA, ordered from highest to lowest activity (22). This
85 array of targets ('v1 barcode') was cloned downstream of an EGFP reporter in a lentiviral
86 construct (23). We then transduced HEK293T cells with lentivirus and used FACS to purify an
87 EGFP-v1 positive population. To edit the barcode, we co-transfected these cells with a plasmid
88 expressing Cas9 and the sgRNA and a vector expressing DsRed. Cells were sorted three days
89 post-transfection for high DsRed expression, and genomic DNA (gDNA) was harvested on day
90 7. The v1 barcode was PCR amplified and the resulting amplicons subjected to deep
91 sequencing.

92

93 To minimize confounding sequencing errors, which are primarily substitutions, we analyzed
94 edited barcodes for only insertion-deletion changes relative to the 'wild-type' v1 barcode. In
95 this first experiment, we observed 1,650 uniquely edited barcodes (each observed in ≥ 25 reads)
96 with diverse edits concentrated at the expected Cas9 cleavage sites, predominantly inter-target
97 deletions involving sites 1, 3 and 5, or focal edits of sites 1 and 3 (Fig. 1, B and C, and table

98 S1). These results show that combinatorial editing of the barcode can give rise to a large
99 number of unique sequences, *i.e.* “alleles”.

100

101 To evaluate reproducibility, we transfected the same editing reagents to cultures expanded from
102 three independent EGFP-v1 positive clones. Targeted RT-PCR and sequencing of EGFP-v1
103 RNA showed similar distributions of edits to the v1 barcode in the transcript pool, between
104 replicates as well as in comparison to the previous experiment (fig. S1). These results show that
105 the observed editing patterns are largely independent of the site of integration and that edited
106 barcodes can be queried from either RNA or DNA.

107

108 To evaluate how editing outcomes vary as a function of Cas9 expression, we co-transfected
109 EGFP-v1 positive cells with a plasmid expressing Cas9 and the sgRNA as well as an DsRed
110 vector, and after four days sorted cells into low, medium, and high DsRed bins and harvested
111 gDNA. Overall editing rates matched DsRed expression (frequency of non-wild-type barcodes:
112 low DsRed = 40%; medium DsRed = 69%; high DsRed = 91%). The profile of edits observed
113 remained similar, but there were fewer inter-target deletions in the lower DsRed bins (fig. S2).
114 These results show that adjusting expression levels of editing reagents can be used to modify
115 the rates and patterns of barcode editing.

116

117 We also synthesized and tested three barcodes (v2-v4) with nine or ten weaker off-target sites
118 for the same sgRNA as used for v1 (22). Genome editing resulted in derivative barcodes with
119 substantially fewer edits than seen with the v1 barcode, but a much greater proportion of these
120 edits were to a single target site, *i.e.* fewer inter-target deletions were observed (Fig. 1, D and E,
121 and fig. S3, A and B). As only a few targets were substantially edited in designs v1-v4, we
122 combined the most highly active targets to a new, twelve target barcode (v5). This barcode
123 exhibited more uniform usage of constituent targets, but with relative activities still ranging
124 over two orders of magnitude (fig. S3C and table S1). These results illustrate the potential value
125 of iterative barcode design.

126

127 To determine whether the means of editing reagent delivery influences patterns of barcode
128 editing, we introduced a lentiviral vector expressing Cas9 and the same sgRNA to cells
129 containing the v5 barcode (24). After two weeks of culturing a population bottlenecked to 200
130 cells by FACS, we observed diverse barcode alleles but with substantially fewer inter-target
131 deletions than with episomal delivery of editing reagents (fig. S3D). This finding demonstrates
132 that the allelic spectrum can also be modulated by the delivery mode of editing reagents.

133

134 Taken together, these results show that editing multiple target sites within a compact barcode
135 can generate a combinatorial diversity of alleles, and also that these alleles can be read out by
136 single sequencing reads derived from either DNA or RNA. Rates and patterns of barcode
137 editing are tunable by using targets with different activities and/or off-target sequences, by

138 iteratively recombining targets to new barcode designs, and by modulating the concentration
139 and means of delivery of editing reagents.

140

141 **Reconstruction of lineage relationships in cultured cells**

142

143 To determine whether GESTALT could be used to reconstruct lineage relationships, we applied
144 it to a designed lineage in cell culture (Fig. 2). A monoclonal population of EGFP-v1 positive
145 cells was transfected with editing reagents to induce a first round of mutations in the v1
146 barcode. Clones derived from single cells were expanded, sampled, split, and re-transfected
147 with editing reagents to induce a second round of mutations of the v1 barcode. For each clonal
148 population, two 100-cell samples of the re-edited populations were expanded and harvested for
149 gDNA. In these experiments, we began incorporating unique molecular identifiers (UMIs; 10
150 bp) during amplification of barcodes by a single round of polymerase extension (fig. S4A).
151 Each UMI tags the single barcode present within each single cell, thereby allowing for
152 correction of subsequent PCR amplification bias and enabling each UMI-barcode combination
153 to be interpreted as deriving from a single cell (25).

154

155 Seven of twelve clonal populations we isolated contained mutations in the v1 barcode that were
156 unambiguously introduced during the first round of editing (Fig. 2A). Additional edits
157 accumulated in re-edited cells but generally did not disrupt the early edits (Fig. 2B and fig. S5).
158 We next sought to reconstruct the lineage relationships between all alleles observed in the
159 experiment using a maximum parsimony approach (fig. S4B)(26). The resultant tree contained
160 major clades that were defined by the early edits present in each lineage (Fig. 2C). Four clonal
161 populations (#3, #5, #7 and #8) were cleanly separated upon lineage reconstruction, with
162 >99.7% of cells accurately placed into each lineage's major clade. Two lineages (#1 and #6)
163 were mixed because they shared identical mutations from the first round of editing. These most
164 likely represent the recurrence of the same editing event across multiple lineages, but could also
165 have been daughter cells subsequent to a single, early editing event prior to isolating clones.
166 Consequently, 99.9% of cells of these two lineages were assigned to a single clade (Fig. 2C,
167 blue). One clonal population (#4) appears to have derived from two independent cells, one of
168 which harbored an unedited barcode. Later editing of these barcodes confounded the assignment
169 of this lineage on the tree. Overall, however, these results demonstrate that GESTALT can be
170 used to capture and reconstruct cell lineage relationships in cultured cells.

171

172 **Combinatorial and cumulative editing of a compact genomic barcode in zebrafish**

173

174 To test the potential of GESTALT for *in vivo* lineage tracing in a complex multicellular
175 organism, we turned to the zebrafish *Danio rerio*. We designed two new barcodes, v6 and v7,
176 each with ten sgRNA target sites that are absent from the zebrafish genome and predicted to be
177 highly editable (methods). In contrast to v1-v5, in which the target sites are variably editable by

178 one sgRNA, the targets within v6 or v7 are designed to be edited by distinct sgRNAs. We
179 generated transgenic zebrafish that harbor each barcode in the 3' UTR of DsRed driven by the
180 ubiquitin promoter (27, 28) and a GFP marker that is expressed in the cardiomyocytes of the
181 heart (fig. S6) (29). To evaluate whether diverse alleles could be generated by *in vivo* genome
182 editing, we injected Cas9 and ten different sgRNAs with perfect complementarity to the barcode
183 target sites into single-cell v6 embryos (Fig. 3A). Editing of integrated barcodes had no
184 noticeable effects on development (fig. S7). To characterize barcode editing *in vivo*, we
185 extracted gDNA from a series of single 30 hours post fertilization (hpf) embryos, and UMI-
186 tagged, amplified and sequenced the v6 barcode. In control embryos (Cas9-; n = 2), all 4,488
187 captured barcodes were unedited. In contrast, in edited embryos (Cas9+; n = 8), fewer than 1%
188 of captured barcodes were unedited. We recovered barcodes from hundreds of cells per embryo
189 (median 943; range 257-2,832) and identified dozens to hundreds of alleles per embryo (median
190 225; range 86-1,323). 41% +/- 10% of alleles were observed recurrently within single embryos,
191 most likely reflecting alleles that were generated in a progenitor of two or more cells. Fewer
192 than 0.01% of alleles were shared in pairwise comparisons of embryos, revealing the highly
193 stochastic nature of editing in different embryos. These results demonstrate that GESTALT can
194 generate very high allelic diversity *in vivo*.

195

196 **Reconstruction of lineage relationships in embryos**

197

198 To evaluate whether lineage relationships can be reconstructed using edited barcodes, we
199 focused on the v6 embryo with the lowest rates of inter-target deletions and edited target sites
200 (Fig. 3B; avg. 58% +/- 27% of target sites no longer a perfect match to the unedited target,
201 compared to 87% +/- 21% for all other 30 hpf v6 embryos). Application of our parsimony
202 approach (fig. S4B) to the 1,961 cells in which we observed 1,323 distinct alleles generated the
203 large tree shown in Fig. 4. 1,307 of the 1,323 (98%) alleles could be related to at least one other
204 allele by one or more shared edits, 85% by two or more shared edits, and 56% by three or more
205 shared edits. These results illustrate the principle of using patterns of shared edits between
206 distinct barcode alleles to reconstruct their lineage relationships *in vivo*.

207

208 **Developmental timing of barcode editing**

209

210 To determine the developmental timing of barcode editing, we injected Cas9 and ten sgRNAs
211 into one-cell stage v7 transgenic embryos and harvested genomic DNA before gastrulation
212 (dome stage, 4.3 hpf; n = 10 animals), after gastrulation (90% epiboly / bud stage, 9 hpf; n = 11
213 animals), at pharyngula stage (30 hpf; n = 12 animals), and from early larvae (72 hpf; n = 12
214 animals) (Fig. 3A). We recovered barcode sequences from a median of 8,785 cells per embryo
215 (range 461-31,640; total of 45 embryos), comprising a median of 1,223 alleles per embryo
216 (range 15-4,195) (Fig. 3C). Within single embryos, 65% +/- 6% of alleles were observed
217 recurrently, whereas in pairwise comparisons of embryos only 2% +/- 5% of alleles were

218 observed recurrently. The abundances of alleles were well-correlated between technical
219 replicates for each of two 72 hpf embryos (fig. S8A and B), and alleles containing many edits
220 were more likely to be unique to an embryo than those with few edits (fig. S8C). To assess
221 when editing begins, we analyzed the proportions of the most common editing events across all
222 barcodes sequenced in a given embryo, reasoning that the earliest edits would be the most
223 frequent. Across eight v6 and 45 v7 embryos, we never observed an edit that was present in
224 100% of cells. This observation indicates that no permanent edits were introduced at the one-
225 cell stage. In nearly all embryos, we observe that the most common edit is present in >10% of
226 cells, and in some cases in ~50% of cells (Fig. 3D and fig. S9). This observation also holds in
227 ~4,000-cell dome stage embryos, which result from approximately 12 rounds of largely
228 synchronous division unaccompanied by cell death. Most of these edits are rare or absent in
229 other embryos, suggesting they are unlikely to have arisen recurrently within each lineage.
230 These results suggest that the edits present in ~50% of cells were introduced at the two-cell
231 stage and that the edits present in >10% of cells were introduced before the 16-cell stage.

232
233 How long does barcode editing persist? Two aspects of the data suggest that it tapers relatively
234 early in development. First, in dome stage embryos (4.3 hpf), we captured barcodes from a
235 median of 2,086 cells, in which a median of 4.8 targets were edited. Although the number of
236 cells and alleles that we were able to sample increased at the later developmental stages, the
237 proportion of edited sites appeared relatively stable (Fig. 3C). If editing were occurring
238 throughout this time course, we would instead expect the proportion of edited sites to increase
239 substantially. Second, the number of unique alleles appears to saturate early, never exceeding
240 4,200 (Fig. 3E). For example, only 4,195 alleles were observed in a 72 hpf embryo in which we
241 sampled the highest number of cells ($n = 31,639$). These results suggest that the majority of
242 editing events occurred before dome stage.

243 244 **Editing diversity in adult organs**

245
246 To evaluate whether barcodes edited during embryogenesis can be recovered in adults, we
247 dissected two edited 4-month old v7 transgenic zebrafish (ADR1 and ADR2) (Fig. 5A). We
248 collected organs representing all germ layers - the brain and both eyes (ectodermal), the
249 intestinal bulb and posterior intestine (endodermal), the heart and blood (mesodermal), and the
250 gills (neural crest, with contributions from other germ layers). We further divided the heart into
251 four samples - a piece of heart tissue, dissociated unsorted cells (DHCs), FACS-sorted GFP+
252 cardiomyocytes, and non-cardiomyocyte heart cells (NCs) (fig. S10). We isolated genomic
253 DNA from each sample, amplified and sequenced edited barcodes with high technical
254 reproducibility (fig. S11), and observed barcode editing rates akin to those in embryos (fig.
255 S12). For zebrafish ADR1, we captured barcodes from between 776 and 44,239 cells from each
256 tissue sample (median 17,335), corresponding to a total of 197,461 cells and 1,138 alleles. For
257 zebrafish ADR2, we captured barcodes from between 84 and 52,984 cells from each tissue

258 sample (median 20,973), corresponding to a total of 217,763 cells and 2,016 alleles. These
259 results show that edits introduced to the barcode during embryogenesis are inherited through
260 development and tissue homeostasis and can be detected in adult organs.

261

262 **Differential contribution of embryonic progenitors to adult organs**

263

264 To analyze the contribution of diverse alleles to different organs, we compared the frequency of
265 edited barcodes within and between organs. We first examined blood (of note, zebrafish
266 erythrocytes are nucleated (30)). Only 5 alleles defined over 98% of cells in the ADR1 blood
267 sample (Fig. 5B), suggesting highly clonal origins of the adult zebrafish blood system from a
268 few embryonic progenitors. Consistent with the presence of blood in all dissected organs, these
269 common blood alleles were also observed in all organs (10-40%; Fig. 5C) but largely absent
270 from cardiomyocytes isolated by flow sorting (0.5%). Furthermore, the relative proportions of
271 these five alleles remained constant in all dissected organs, suggesting that they primarily mark
272 the blood and do not substantially contribute to non-blood lineages (Fig. 5D). In performing
273 similar analyses of clonality across all organs (while excluding the five most common blood
274 alleles), we observed that a small subset of alleles dominates each organ (Fig. 5E). Indeed, for
275 all dissected organs, fewer than 7 alleles comprised >50% of cells (median 4, range 2-6), and,
276 with the exception of the brain, fewer than 25 alleles comprised >90% of cells (median 19,
277 range 4-38). Most of these dominant alleles were organ-specific, *i.e.* although they were found
278 rarely in other organs, they tended to be dominant in only one organ (Fig. 5F). For example, the
279 most frequent allele observed in the intestinal bulb comprised 13.6% of captured non-blood
280 cells observed in that organ, but <0.01% of cells observed in any other organ. There are
281 exceptions, however. For example, one allele is observed in 24.7% of sorted cardiomyocytes,
282 13.4% of the intestinal bulb, and at lower abundances in all other organs. Similar results were
283 observed in ADR2 (fig. S13). These results indicate that the majority of cells in diverse adult
284 organs are descended from a few differentially edited embryonic precursors.

285

286 **Reconstructing lineage relationships in adult organs**

287

288 To reconstruct the lineage relationships between cells both within and across organs on the basis
289 of shared edits, we again relied on maximum parsimony methods (fig. S4B). The resulting trees
290 for ADR1 and ADR2 are shown in Fig. 6 and fig. S14, respectively. We observed clades of
291 alleles that shared specific edits. For example, ADR1 had 8 major clades, each defined by
292 'ancestral' edits that are shared by all captured cells assigned to that clade (Fig. 7A; also
293 indicated by colors in the tree shown in Fig. 6). Collectively, these clades comprised 49% of
294 alleles and 90% of the 197,461 cells sampled from ADR1 (Fig. 7A). Blood was contributed to
295 by 3 major clades (#3, #6, #7) (Fig. 7B). After re-allocating the 5 dominant blood alleles from
296 the composition of individual organs back to blood (Fig. 5B and fig. S15), we observed that all
297 major clades made highly non-uniform contributions across organs. For example, clade #3

298 contributed almost exclusively to mesodermal and endodermal organs, while clade #5
299 contributed almost exclusively to ectodermal organs. These results reveal that GESTALT can
300 be used to infer the contributions of inferred ancestral progenitors to adult organs.

301
302 Although some ancestral clades appear to contribute to all germ layers, we find that subclades,
303 defined by additional shared edits within a clade, exhibit greater specificity. For example, while
304 clade #1 contributes substantially to all organs except blood, additional edits divide clade #1
305 into three subclades with greater tissue restriction (Fig. 7C and D). The #1+A subclade
306 primarily contributes to mesendodermal organs (heart, both gastrointestinal organs) while the
307 #1+C subclade primarily contributes to neuroectodermal organs (brain, left eye, and gills).
308 Similar patterns are observed for clade #2 (Fig. 7E and F), where the #2+A subclade contributes
309 primarily to mesendodermal organs, the #2+B subclade to the heart, and the #2+C clade to
310 neuroectodermal organs. Additional edits divide these subclades into further tissue-specific sub-
311 subclades. For example, while the #2+A subclade is predominantly mesendoderm, additional
312 edits define #2+A+D (heart, primarily cardiomyocytes), #2+A+E (heart and posterior intestine),
313 and #2+A+F (intestinal bulb). All of the major clades exhibit similar patterns of increasing
314 restriction with additional edits (Fig. 7C-F and fig. S16). Similar observations were made in fish
315 ADR2 (fig. S17). These results indicate that GESTALT can record lineage relationships across
316 many cell divisions and capture information both before and during tissue restriction.

317

318 **Discussion**

319

320 We describe a new method, GESTALT, which uses combinatorial and cumulative genome
321 editing to record cell lineage information in a highly multiplexed fashion. We successfully
322 applied this method to both artificial lineages (cell culture) as well as to whole organisms
323 (zebrafish).

324

325 The strengths of GESTALT include: 1) the combinatorial diversity of mutations that can be
326 generated within a dense array of CRISPR/Cas9 target sites; 2) the potential for informative
327 mutations to accumulate across many cell divisions and throughout an organism's developmental
328 history; 3) the ability to scalably query lineage information from at least hundreds of thousands
329 of cells and with a single sequencing read per single cell; 4) the likely applicability of GESTALT
330 to any organism, from bacteria and plants to vertebrates, that allows genome editing, as well as
331 human cells (*e.g.* tumor xenografts). Even in organisms in which transgenesis is not established,
332 lineage tracing by genome editing may be feasible by expressing editing reagents to densely
333 mutate an endogenous, non-essential genomic sequence.

334

335 Our experiments also highlight several remaining technical challenges. Chief amongst these are:
336 1) the chance recurrence of identical edits or similar patterns of edits in distantly related cells can
337 confound lineage inference; 2) non-uniform editing efficiencies and inter-target deletions within

338 the barcode contribute to suboptimal sequence diversity and loss of information, respectively; 3)
339 the transient means by which Cas9 and sgRNAs are introduced likely restrict editing to early
340 embryogenesis; 4) the computational challenge of precisely defining the multiple editing events
341 that give rise to different alleles complicates the unequivocal reconstruction of lineage trees; and
342 5) the difficulty of isolating tissues without contamination by blood and other cells can hinder
343 the assignment of alleles to specific organs. A broader set of challenges includes the lack of
344 information about the precise anatomical location and exact cell type of each queried cell, the
345 fact that genome editing events are not directly coupled to the cell cycle, and the failure to
346 recover all cells. These challenges currently hinder the reconstruction of a lineage tree as
347 complete and precise as the one that Sulston and colleagues described for *C. elegans*. Despite
348 these limitations, our proof-of-principle study shows that GESTALT can inform developmental
349 biology by richly defining lineage relationships among vast numbers of cells recovered from an
350 organism.

351
352 The current challenges highlight the need for further optimization of the design of targets and
353 arrays, as well as the delivery of editing reagents. For example, an array containing twice as
354 many targets as used here could fit within a single read on contemporary sequencing platforms,
355 thus yielding more lineage information per cell without sacrificing throughput. Also, as we have
356 shown, adjustments to the target sequences and dosages of editing reagents can be used to fine-
357 tune mutation rates and to minimize undesirable inter-target deletions. Finally, sgRNA sequences
358 and lengths (31), Cas9 cleavage activity and target preferences (32, 33), and the means by which
359 Cas9 and sgRNA(s) are expressed (e.g. transient, constitutive (34), or induced (35, 36)), can be
360 altered to control the pace, temporal window and tissue(s) at which the barcodes are mutated. For
361 example, coupling editing to cell cycle progression might enable higher resolution reconstruction
362 of lineage relationships throughout development.

363
364 Our application of GESTALT to a vertebrate model organism, zebrafish, demonstrates its
365 potential to yield insights into developmental biology. First, our results suggest that relatively
366 few embryonic progenitor cells give rise to the majority of cells of many adult zebrafish organs,
367 reminiscent of clonal dominance (37, 38). For example, only 5 of the 1,138 alleles observed in
368 ADR1 gave rise to >98% of blood cells, and for all dissected organs, fewer than 7 alleles
369 comprised >50% of cells. There are several mechanisms by which such dominance can emerge,
370 e.g. by uneven starting populations in the embryo, drift, competition, interference, unequal cell
371 proliferation or death, or a combination of these mechanisms (39-42). Controlling the temporal
372 and spatial induction of edits and isolating defined cell types from diverse organs should help
373 resolve the mechanisms by which different embryonic progenitors come to dominate different
374 adult organs.

375
376 Second, we show that GESTALT can inform the lineage relationships amongst thousands of
377 differentiated cells. For example, following the accumulation of edits from ancestral to more

378 complex reveals the progressive restriction of progenitors to germ layers and then organs. Cells
379 within an organ can both share and differ in their alleles, revealing additional information about
380 organ development. Future studies will need to determine whether such lineages reflect distinct
381 cell fates (*e.g.*, blood sub-lineages or neuronal subpopulations), because the anatomical
382 resolution at which we queried alleles was restricted to grossly dissected organs and tissues.
383 Because edited barcodes are expressed as RNA, we envision that combining our system with
384 other platforms will permit much greater levels of anatomical resolution without sacrificing
385 throughput. For example, *in situ* RNA sequencing of barcodes would provide explicit spatial and
386 histological context to lineage reconstructions (19, 20). Also, capturing richly informative
387 lineage markers in single cell RNA-seq or ATAC-seq datasets may inform the interpretation of
388 those molecular phenotypes, while also adding cell type resolution to studies of lineage (43, 44).
389 Such integration may be particularly relevant to efforts to build comprehensive atlases of cell
390 types. Because these single cell methods generate many reads per single cell, this would also
391 facilitate using multiple, unlinked target arrays. In principle, the combined diversity of the
392 barcodes queried from single cells could be engineered to uniquely identify every cell in a
393 complex organism. In addition, orthogonal imaging-based lineage tracing approaches in fixed
394 and live samples (*e.g.*, Brainbow and related methods (16, 29)) and longitudinal whole animal
395 imaging approaches (45, 46) might be leveraged in parallel to validate and complement lineages
396 resolved by GESTALT.

397
398 Although further work is required to optimize GESTALT towards enabling spatiotemporally
399 complete maps of cell lineage, our proof-of-principle experiments show that using multiplex *in*
400 *vivo* genome editing to record lineage information to a compact barcode at an organism-wide
401 scale will be a powerful tool for developmental biology. This approach is not limited to normal
402 development but can also be applied to animal models of developmental disorders, as well as to
403 investigate the origins and progression of cancer. Our study also supports the notion that whereas
404 its most widespread application has been to modify endogenous biological circuits, genome
405 editing can also be used to stably record biological information (47), analogous to recombinase-
406 based memories but with considerably greater flexibility and scalability. For example, coupling
407 editing activity to external stimuli or physiological changes could record the history of exposure
408 to intrinsic or extrinsic signals. In the long term, we envision that rich, systematically generated
409 maps of organismal development, wherein lineage, epigenetic, transcriptional and positional
410 information are concurrently captured at single cell resolution, will advance our understanding of
411 normal development, inherited diseases, and cancer.

412

413 **Acknowledgements**

414

415 For discussion and advice, we thank Joe Felsenstein, Mary Kuhner, Marlies Rossmann, Eva Fast,
416 Caroline Burns, Raz Ben-Yair, Steve Salipante; members of the Shendure Lab, particularly
417 Matthew Snyder; members of the Schier Lab, particularly Jeffrey Farrell, Bushra Raj, Megan

418 Norris, and Nathan Lord; and members of the Horwitz lab, particularly Donovan Anderson. We
419 thank George Church for work on predecessors of this concept with JS in 2000. We thank
420 Michael Desai, Rich Losick, Andrew Murray and Len Zon for comments on the manuscript. We
421 also thank the Bauer Core FACS Facility, the staff of the zebrafish facility, and Lindsey Pieper
422 for technical support. This work was supported by grants from the Paul G. Allen Family
423 Foundation (JS & MSH), an NIH Director's Pioneer Award (JS; DP1HG007811), NIH/NIGMS
424 (AFS; GM056211), NIH/NICHD (AFS; HD085905), and NIH/NIMH (AFS; MH105960). JAG
425 was supported by a fellowship from the American Cancer Society. AHM was supported by a
426 fellowship from the NIH/NHLBI (T32HL007312). JS is an investigator of the Howard Hughes
427 Medical Institute.

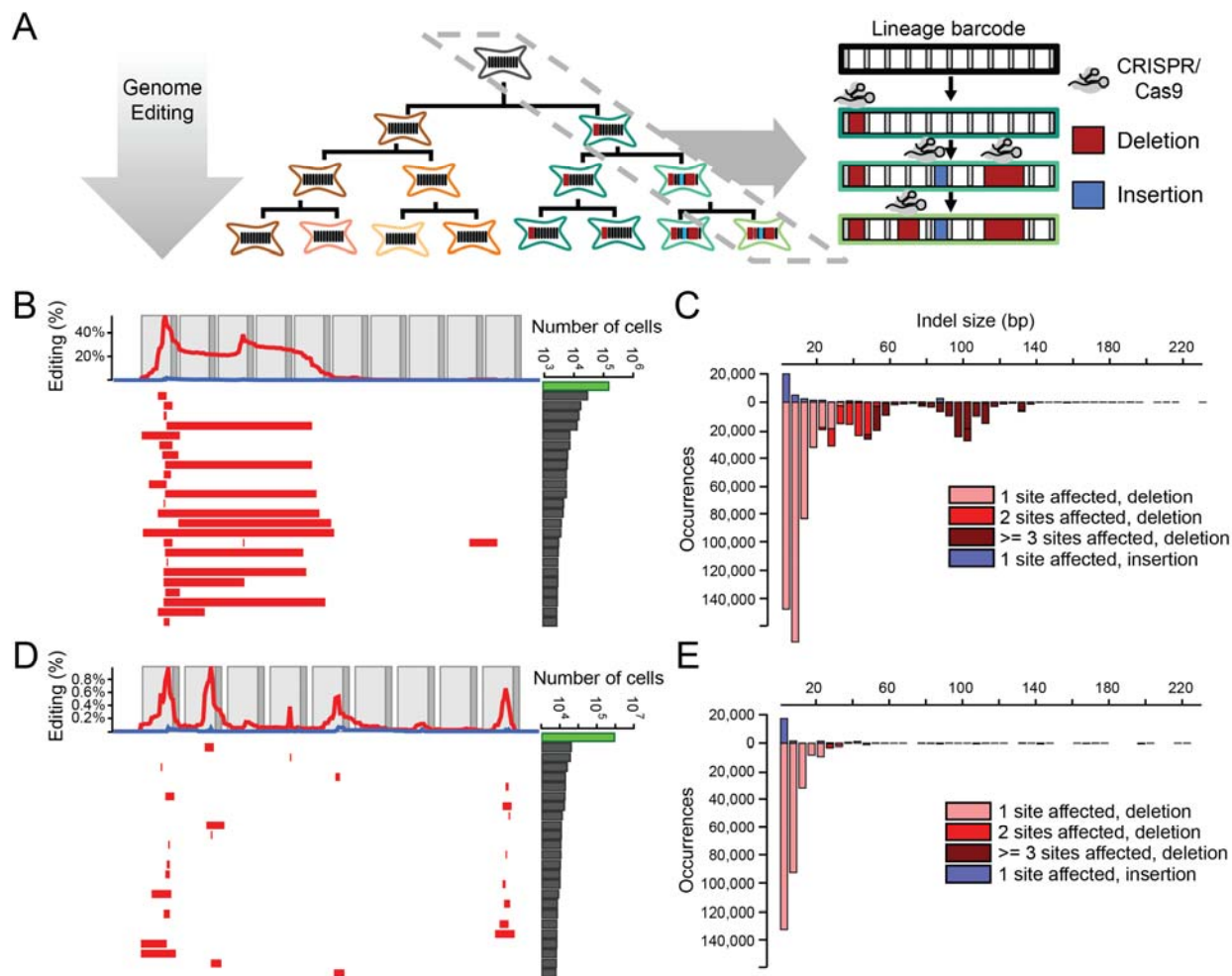
428

429 **Author contributions**

430

431 GF, AM and JS developed the initial concept. GF led the cell culture experiments and developed
432 the UMI protocol, with assistance from AM. JG led the fish experiments, with assistance from
433 AM and GF. AM led development of the analysis pipeline. AM, GF, and JG processed and
434 analyzed the data. AM, GF, JG, AS and JS designed experiments and interpreted the data. MH
435 provided critical early insight. AM, GF, JG, AS, and JS wrote the manuscript.

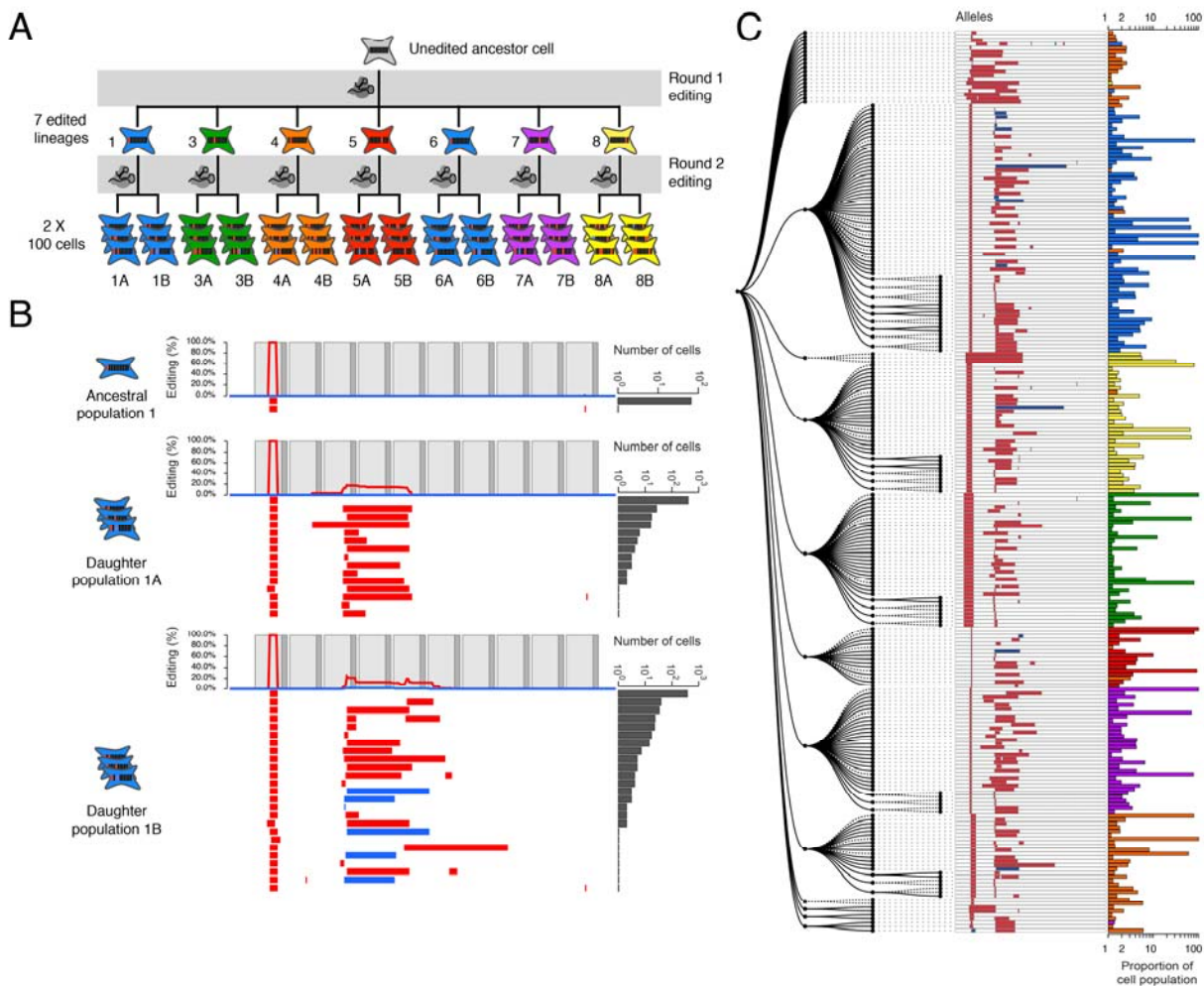
436



437
 438 **Fig. 1. Genome editing of synthetic target arrays for lineage tracing (GESTALT).** (A) An
 439 unmodified array of CRISPR/Cas9 target sites (*i.e.*, a barcode) is engineered into a genome.
 440 Editing reagents are introduced during expansion of cell culture or *in vivo* development of an
 441 organism, such that distinct lineages stably accumulate different edits to the barcode. Because
 442 the resulting edited barcodes are compact, derivative sequences present within a population of
 443 cells can be queried by PCR and sequencing, such that each sequencing read corresponds to the
 444 barcode derived from a single cell. Edited barcodes that are allelic, *i.e.* identical in sequence, but
 445 differ by UMI are inferred to have derived from distinct but closely related cells. The lineage
 446 relationships of alleles that differ in sequence can often be inferred on the basis of editing
 447 patterns. (B) The 25 most frequent alleles from the edited v1 barcode are shown. Each row
 448 corresponds to a unique sequence, with red bars indicating deleted regions and blue bars
 449 indicating insertion positions. Blue bars begin at the insertion site, with their width proportional
 450 to the size of the insertion, which will rarely obscure immediately adjacent deletions. The
 451 number of reads observed for each allele is plotted at the right (log10 scale; the green bar
 452 corresponds to the unedited allele). The frequency at which each base is deleted (red) or flanks
 453 an insertion (blue) is plotted at the top. Light gray boxes indicate the location of CRISPR
 454 protospacers while dark gray boxes indicate PAM sites. For the v1 array, inter-target deletions

455 involving sites 1, 3 and 5, or focal (single target) edits of sites 1 and 3 were observed
456 predominantly. **(C)** A histogram of the size distribution of insertion (top) and deletion (bottom)
457 edits to the v1 array is shown. The colors indicate the number of target sites impacted. Although
458 most edits are short and impact a single target, a substantial proportion of edits are inter-target
459 deletions. **(D)** We tested three array designs in addition to v1, each comprising nine to ten
460 weaker off-target sites for the same sgRNA (v2-v4) (22). Editing of the v2 array is shown with
461 layout as described in panel (B). Editing of the v3 and v4 array are shown in fig. S3, A and B.
462 The weaker sites within these alternative designs exhibit lower rates of editing than the v1 array,
463 but also a much lower proportion of inter-target deletions. **(E)** A histogram of the size
464 distribution of insertion (top) and deletion (bottom) edits to the v2 array is shown. In contrast
465 with the v1 array, almost all edits impact only a single target.
466

467



468

469

470

471

472

473

474

475

476

477

478

479

480

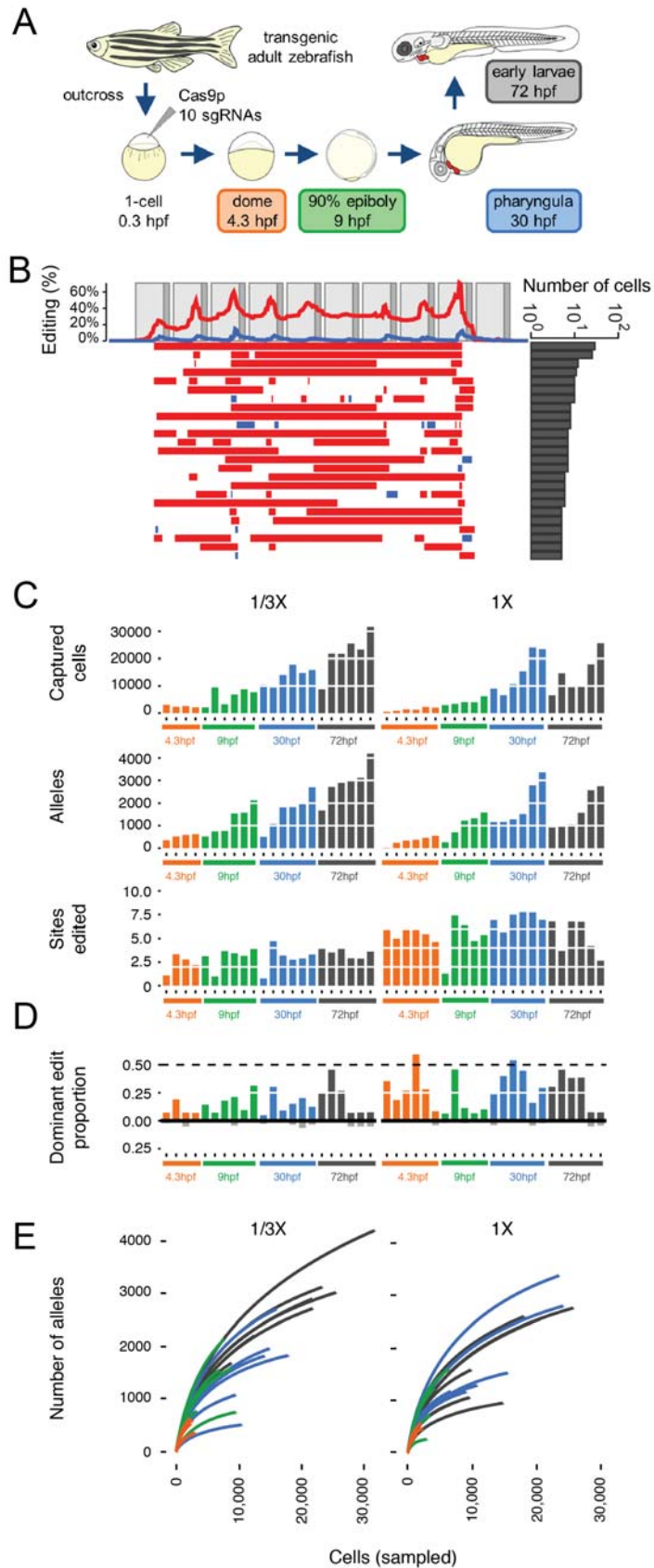
481

482

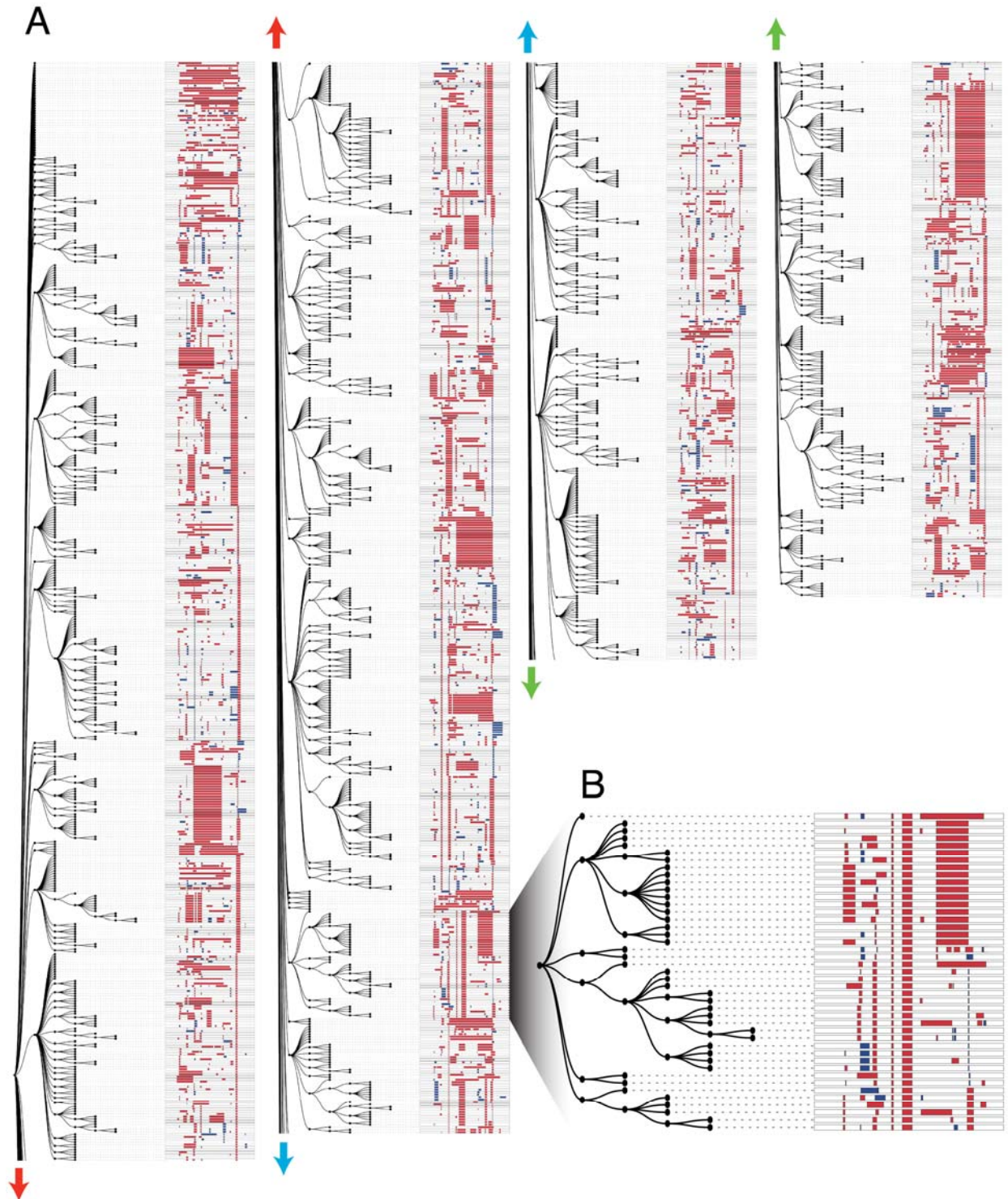
483

Fig. 2. Reconstruction of a synthetic lineage based on genome editing and targeted sequencing of edited barcodes. (A) A monoclonal population of cells was subjected to editing of the v1 array. Single cells were expanded, sampled (#1 to #12), re-transfected to induce a second round of barcode editing, and then expanded and sampled from 100-cell subpopulations (#1a, 1b to #12a, 12b). For clarity, the five clones where the original population was unedited are not shown. **(B)** Alleles observed in the synthetic lineage experiment are shown, with layout as described in the Fig. 1B legend. #1 represents sampling of cells that had been subjected to only the first round of editing; virtually all cells contain a shared edit to the first target. #1a and #1b are derived from #1 but subjected to a second round of editing prior to sampling. These retain the edit to the first target, but subpopulations bear additional edits to other targets. **(C)** Parsimony (fig. S4B) using all alleles from the seven cell lineages represented in panel (A). Clade membership and abundance of each allele are shown on the right. #4 appears to be derived from two cells, one edited and the other wild-type: only 62% of lineage #4 falls into a single clade, consistent with the proportion (64%) of the lineage determined by sequencing to be edited after the first round. It is assumed that cells unedited in the first round either accrued

484 edits matching other lineages (thus causing mixing), or accrued different edits (thus remaining
485 outside the major clades).
486

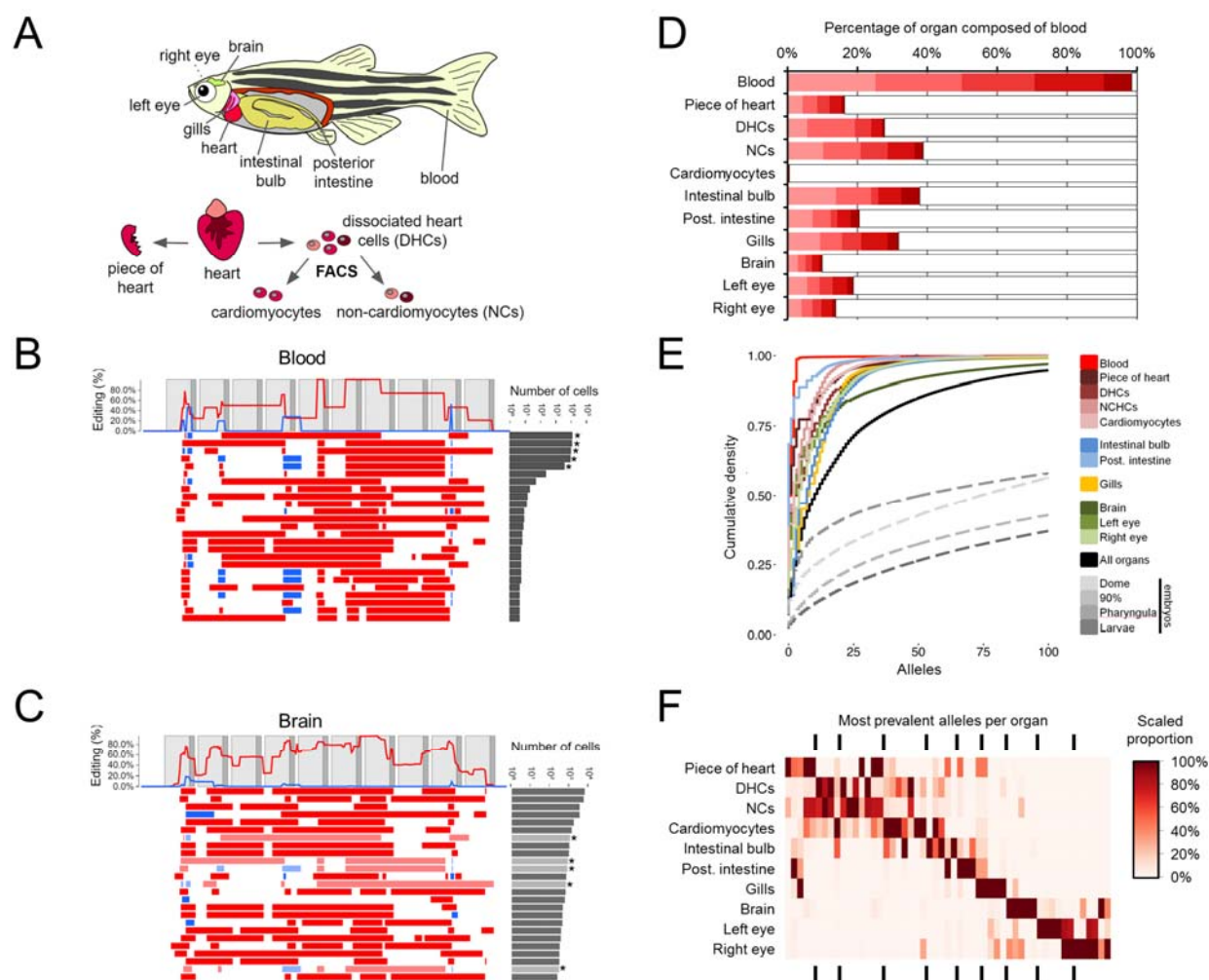


488 **Fig. 3. Generating combinatorial barcode diversity in transgenic zebrafish.** (A) One-cell
489 zebrafish embryos were injected with complexed Cas9 ribonucleoproteins (RNPs) containing
490 sgRNAs that matched each of the 10 targets in the array (v6 or v7). Embryos were collected at
491 time points indicated. UMI-tagged barcodes were amplified and sequenced from genomic DNA.
492 (B) Patterns of editing in alleles recovered from a 30 hpf v6 embryo, with layout as described in
493 the Fig. 1B legend. (C) Bar plots show the number of cells sampled (top), unique alleles
494 observed (middle) and proportion of sites edited (bottom) for 45 v7 embryos collected at four
495 developmental time-points and two levels of Cas9 RNP (1/3x, 1x). Colors correspond to stages
496 shown in panel (A). Although more alleles are observed with sampling of larger numbers of cells
497 at later time points, the proportion of target sites edited remains relatively constant. (D) Bar plots
498 show the proportion of edited barcodes containing the most common editing event in a given
499 embryo. Six of 45 embryos had the most common edit in approximately 50% of cells (dashed
500 line), consistent with this edit having occurred at the two-cell stage (see fig. S8A for example).
501 Colors correspond to stages shown in panel (A). These same edits are rarer or absent in other
502 embryos (black bars below). (E) For each of the 45 v7 embryos, all barcodes observed were
503 sampled without replacement. The cumulative number of unique alleles observed as a function of
504 the number of cells sampled is shown (average of the 100 iterations shown per embryo; two
505 levels of Cas9 RNP: 1/3x on left, 1x on right). The number of unique alleles observed, even in
506 later stages where we are sampling much larger numbers of cells, appears to saturate, and there is
507 no consistent pattern supporting substantially greater diversity in later time-points, consistent
508 with the bottom row of panel (C) in supporting the conclusion that the majority of editing occurs
509 before dome stage.
510



511
512 **Figure 4. Lineage reconstruction of an edited zebrafish embryo.** (A) A lineage reconstruction
513 of 1,323 alleles recovered from the v6 embryo also represented in Fig. 3B, generated by a
514 maximum parsimony approach (fig. S4B). A dendrogram to the left of each column represents
515 the lineage relationships, and the alleles are represented at right. Each row represents a unique
516 allele. Matched colored arrows connect subsections of the tree together. There are many large

517 clades of alleles sharing specific edits, as well as sub-clades defined by ‘dependent’ edits. These
518 dependent edits occur within a clade defined by a more frequent edit but are rare or absent
519 elsewhere in the tree. **(B)** A portion of the tree is shown at higher resolution. A single edit is
520 shared by all alleles in this clade. Six independent edits define descendent sub-clades within this
521 clade, and further edits define additional sub-sub-clades within the clade.

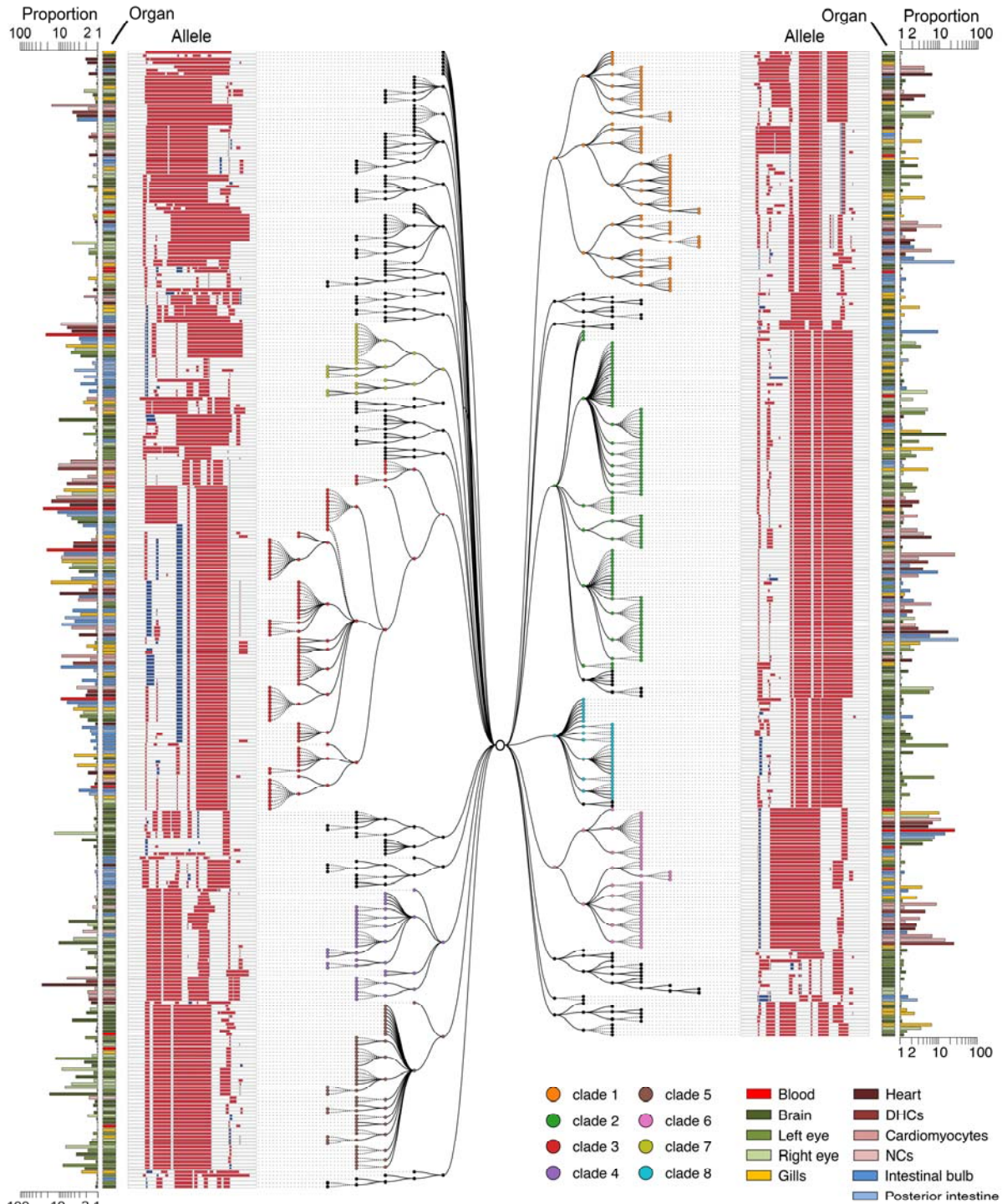


522
 523 **Fig. 5. Organ-specific progenitor cell dominance.** (A) The indicated organs were dissected
 524 from a single adult v7 transgenic edited zebrafish (ADR1). A blood sample was collected as
 525 described in the Methods. The heart was further split into the four samples shown (fig. S10). (B)
 526 Patterns of editing in the most prevalent 25 alleles (out of 135 total) recovered from the blood
 527 sample. Layout as described in the Fig. 1B legend. The most prevalent 5 alleles (indicated by
 528 asterisks) comprise >98% of observed cells. (C) Patterns of editing in the most prevalent 25
 529 alleles (out of 399 total) recovered from brain. Layout as described in the Fig. 1B legend.
 530 Alleles that are identical in sequence to the most prevalent blood alleles are indicated by
 531 asterisks and light shading. (D) The five dominant blood alleles (shades of red) are present in
 532 varying proportions (10-40%) in all intact organs except the FACS-sorted cardiomyocyte
 533 population (0.5%). All other alleles are summed in grey. (E) The cumulative proportion of cells
 534 (y-axis) represented by the most frequent alleles (x-axis) for each adult organ of ADR1 is
 535 shown, as well as the adult organs in aggregate. In all adult organs except blood, the five
 536 dominant blood alleles are excluded. All organs exhibit dominance of sampled cells by a small
 537 number of progenitors, with fewer than 7 alleles comprising the majority of cells. For
 538 comparison, a similar plot for the median embryo (dashed) from each time-point of the

539 developmental time course experiment is also shown. **(F)** The distribution of the most prevalent
540 alleles for each organ, after removal of the five dominant blood alleles, across all organs. The
541 most prevalent alleles were defined as being at >5% abundance in a given organ (median 5
542 alleles, range 4-7). Organ proportions were normalized by column and colored as shown in
543 legend. Underlying data presented in table S2.

544

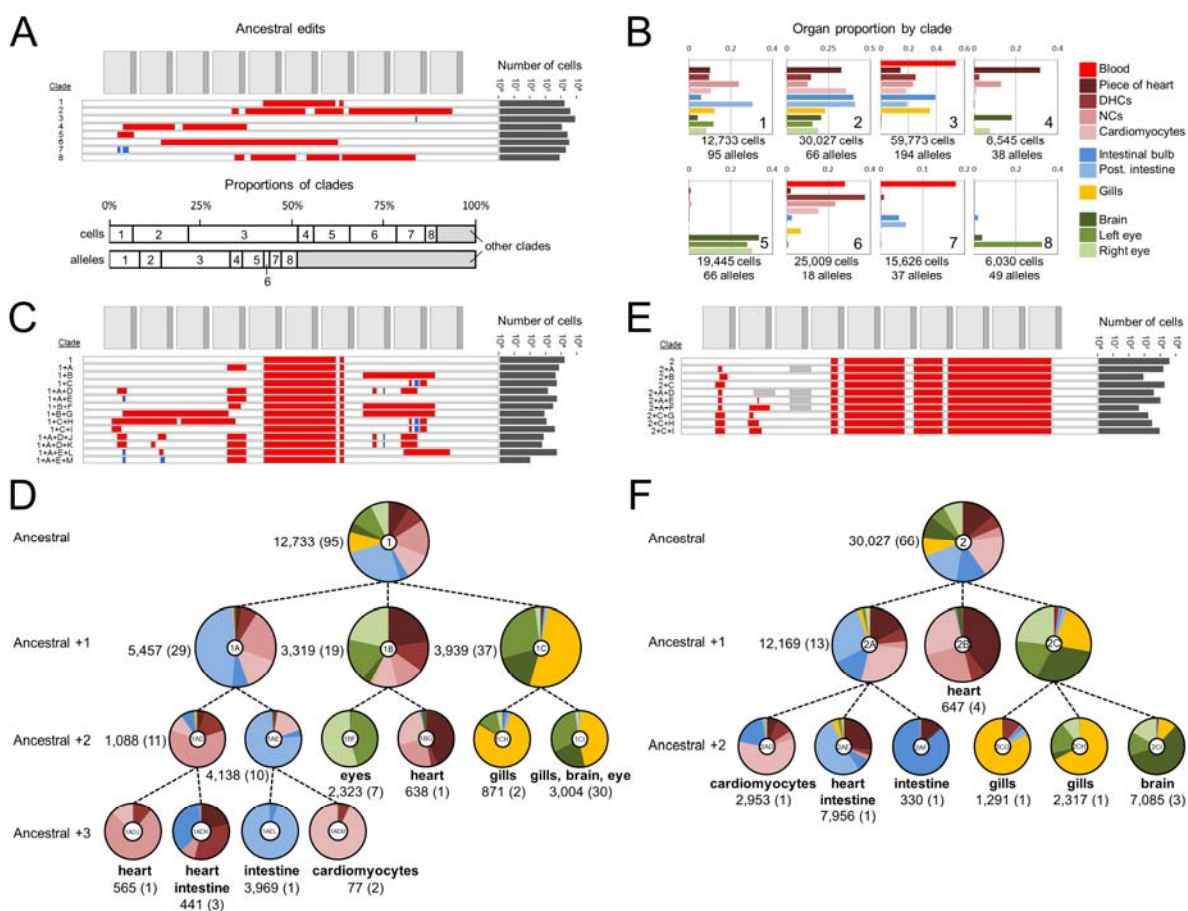
545



546

547 **Fig. 6. Lineage reconstruction for adult zebrafish ADR1.** Unique alleles sequenced from
548 adult zebrafish organs can be related to one another using a maximum parsimony approach into
549 a multifurcating lineage tree. For reasons of space, we show a tree reconstructed from the 611
550 ADR1 alleles observed at least 5 times in individual organs. Eight major clades are displayed
551 with colored nodes, each defined by ‘ancestral’ edits that are shared by all alleles assigned to
552 that clade (shown in Fig. 7A). Editing patterns in individual alleles are represented as shown

553 previously. Alleles in multiple organs are plotted on separate lines per organ and these nodes
554 connected with stippled branches. Two sets of bars outside the alleles identify the organ in
555 which the allele was observed and the proportion of cells in that organ represented by that allele
556 (log scale).
557



558
 559 **Fig. 7. Clades and subclades corresponding to inferred progenitors exhibit increasing**
 560 **levels of organ restriction. (A)** Top panel: The inferred ancestral edits that define eight major
 561 clades of ADR1, as determined by parsimony, are shown, with the total number of cells in
 562 which these are observed indicated on the right. Bottom panel: Contributions of the eight major
 563 clades to all cells or all alleles. 19 alleles (out of 1,138 total) that contained ancestral edits from
 564 more than one clade were excluded from assignment to any clade, and any further lineage
 565 analysis. **(B)** Contributions of each of the eight major clades to each organ, displayed as a
 566 proportion of each organ. To accurately display the contributions of the eight major clades to
 567 each organ, we first re-assigned the five dominant blood alleles from other organs back to the
 568 blood. The total number of cells and alleles within a given major clade are listed below. The
 569 clade contributions of all clades and subclades are presented in table S3. For heart subsamples,
 570 ‘piece of heart’ = a piece of heart tissue, ‘DHCs’ = dissociated unsorted cells; ‘cardiomyocytes’
 571 = FACS-sorted GFP+ cardiomyocytes; and ‘NCs’ = non-cardiomyocyte heart cells. **(C) and (E)**
 572 Edits that define subclades of clade #1 (C) and clade #2 (E), with the total number of cells in
 573 which these are observed indicated on the right. A grey box indicates an unedited site or sites,
 574 distinguishing it from related alleles that contain an edit at this location. **(D) and (F)** Lineage
 575 trees corresponding to subclades of clade #1 (D) and clade #2 (F) that show how dependent
 576 edits are associated with increasing lineage restriction. The pie chart at each node indicates the

577 organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a
578 normalization which accounts for differential depth of sampling between organs. Labels in the
579 center of each pie chart correspond to the subclade labels in (c/e). Alleles present in a clade but
580 not assigned to a descendent subclade (either they have no additional lineage restriction or are at
581 low abundance) are not plotted for clarity. The number of cells (and the number of unique
582 alleles) are also listed, and terminal nodes also list major organ restriction(s), *i.e.* those
583 comprising >25% of a subclade by proportion.

584

585 **Materials and Methods**

586

587 Design of synthetic target arrays

588 Barcodes were designed as arrays of nine to twelve sense-oriented CRISPR/Cas9 target sites (23
589 bp protospacer plus PAM sequences) separated by 3-5 bp linker sequences. Four initial designs
590 (barcodes v1-v4) comprised of target sites for the sgRNA spacer sequence: 5'-
591 GGCCTGCGGCTGGAGGTGG. The v1 barcode was comprised of ten targets arrayed in order
592 of decreasing activity as measured with the GUIDE-seq assay performed in human cells (22),
593 starting with the target perfectly matching the sgRNA spacer sequence. The v2-v4 barcodes
594 comprised of nine to ten non-overlapping target sets, all with activities less than half the
595 perfectly matching target in the GUIDE-seq assay. To reduce repetitive subsequences within
596 each barcode, protospacers were chosen such that no 8 bp sequence was present in more than one
597 protospacer within each barcode. After testing activities of targets in the v1-v4 barcodes in cell
598 culture, the v5 barcode was designed to contain twelve targets that showed greater than ~1%
599 editing activity, including v1 targets 1-6, v3 target 1, v2 targets 1, 2 and 5, and v4 targets 1 and 3.

600 Two new barcodes, v6 and v7, were designed for use in zebrafish, each with ten CRISPR target
601 sites not found in the *D. rerio* genome. Candidate target sequences were screened to remove any
602 homopolymer runs, outside of the NGG of the protospacer, and were selected for editing
603 activity [<http://crispr.mit.edu>]. The v6 and v7 barcodes were constructed as a series of 10
604 protospacer sequences meeting these criteria, with 4 bp linkers.

605 Each barcode was ordered as a gBlock (IDT) with ends compatible for In-Fusion cloning
606 (ClonTech) into the 3' UTR of the EGFP gene in the lentiviral construct pLJM1-EGFP (Addgene
607 #19319).

608 The sequences of all barcodes (v1 through v7) are provided in table S4.

609 Generation of cell lines containing synthetic target arrays

610 To generate cell lines harboring single copies of barcodes, lentiviral particles were produced in
611 HEK 293T cells transfected with lentivirus V2 packaging plasmids and barcode constructs. Viral
612 supernatant harvested three days post transfection was used at low MOI to transduce 293T cells
613 (MOI < 0.2). Successfully transduced cells were selected using puromycin (2 ug/ml), yielding
614 polyclonal, barcode+ populations for barcodes v1-v5. Three monoclonal lines each harboring

615 barcode v1 were generated by single-cell FACS, and used experimentally to compare editing
616 rates across different integration sites. One of these was used as the parent line for cell culture
617 lineages derived using barcode v1.

618 Editing of barcodes in cell lines

619 293T populations bearing barcodes v1-v5 were grown to 50-90% confluency in a 6-well dish.
620 Cells were co-transfected using Lipofectamine 3000 (Life Technologies) according to protocol
621 with 2 μ g pX330-v1 and 0.5 μ g pDsRed in a 6-well dish. One to three days post transfection, the
622 cells were sorted on an Aria III FACS machine for DsRed fluorescence (as a marker
623 transfection). As indicated, either DsRed low, DsRed high, or total DsRed populations were
624 sorted and cultured. At 7 days post-transfection, cells were harvested for gDNA preparation
625 using the Qiagen DNeasy kit.

626 To stably deliver Cas9 and the sgRNA via lentivirus, the spacer sequence was cloned into the
627 plasmid LentiCrispr v2 (Zhang lab, Addgene #52961) and virus was produced in 293T cells in
628 the same manner described above. Wild-type 293T cells were transduced with pLenti-Crispr-V2-
629 HMID.v1 and selected with puromycin, and then transduced with lentivirus bearing barcode v5.
630 To impose a bottleneck, 200 GFP+ cells were sorted from this population and expanded under
631 puromycin selection for two weeks prior to sampling gDNA.

632 Cell culture lineage experiments

633 Twelve lineages were established from a monoclonal barcode v1 293T cell line by transfecting
634 cells as described above, and sorting single DsRed-low cells into a 96-well plate (DsRed low
635 cells were used to limit Cas9 delivery and thus potential saturation of possible edits in this initial
636 editing round). Cell sorting was performed seven days post-transfection, to reduce the likelihood
637 that additional edits would arise after lineages were separated. Single cell-derived populations
638 were expanded in culture for 3 weeks. A sample of cells from each lineage was pelleted and
639 frozen. Next, each of the twelve lineages were transfected a second time, to induce another round
640 of editing. Two 100-cell DsRed-low populations from each lineage were sorted 4-days post-
641 transfection, and cultured to confluence in 96-well plates before harvesting gDNA.

642 Four additional monoclonal populations bearing v5 barcodes edited via transfection of pX330-v1
643 were also isolated by single-cell sorting. Re-editing of each population was achieved by two
644 successive rounds of transfection with pX330-v1 (3 days apart). Cells were harvested for gDNA
645 one week after the second transfection.

646 Barcode amplification and sequencing protocols

647 Kapa High Fidelity Polymerase was used for all barcode amplification steps. Gradient PCRs
648 were performed to optimize annealing temperatures for amplification from gDNA. For
649 experiments performed without UMIs, up to 250 ng of gDNA was loaded into a single 50 μ l
650 PCR reaction and amplified using primers immediately flanking the barcode (see table S4 for
651 oligo sequences). If there was less than 250 ng from a sample, all of it was used in a single
652 reaction. For experiments performed with UMIs, a primer with a sequencing adapter and 10 nt of

653 fully degenerate sequence 5' to the barcode-flanking sequence was used for a single prolonged
654 extension step, in which the temperature was ramped between annealing and extending for five
655 cycles (without a denaturing step to prevent re-sampling of gDNA barcodes). All cell culture
656 experiments and v6 zebrafish embryos received a single extension to incorporate UMIs, whereas
657 v7 embryo time-course experiments and all ADR1 tissues (also v7) received 2 UMI
658 incorporation cycles due to having low gDNA consequent to fewer cells being present in early
659 embryo and sorted heart samples. To minimize repetitive amplification of the same barcode, no
660 reverse primer was included in UMI-tagging reactions. DNA was then purified using AMPure
661 beads (Agencourt), and loaded into a PCR primed from the sequencing adapter flanking each
662 UMI and a site immediately 3' of the barcode.

663 For all experiments, two ensuing qPCRs were performed prior to sequencing to incorporate
664 sequencing adapters, sample indexes, and flow cell adapters. AMPure beads were used to purify
665 PCR products after each reaction.

666 Paired-end sequencing was performed on an Illumina MiSeq using 500- or 600-cycle kits for all
667 cell culture experiments. Zebrafish experiments were sequenced on an Illumina NextSeq using
668 300-cycle kits. All sequencing generated adequate depth to sample each barcode present in a
669 given sample to an average of greater than 10x coverage. To minimize contributions from
670 sequencing error a read threshold was used for calling unique barcodes. This was conservatively
671 set by dividing the number of reads from a sample by the number of expected barcode copies to
672 be present in the amount of gDNA loaded into each PCR based on the assumption that each cell
673 contributed a single barcode.

674 Sequencing data for all samples was processed in a custom pipeline available on GitHub
675 (<https://github.com/shendurelab/Cas9FateMapping>). Briefly, amplicon sequencing reads were
676 first processed with the Trimmomatic software package to remove low quality bases (fig.
677 S4A)(48). The resulting reads were then grouped by their UMI tag. A raw read count threshold
678 was set for each experiment based on sequencing depth, such that only UMIs observed in at least
679 that many reads were analyzed to minimize contributions from sequencing error. For each UMI,
680 a consensus sequence was called by jointly aligning all UMI-matched reads using the MAFFT
681 multiple sequence aligner. These reads were merged using the FLASH (49) read merging tool,
682 and both merged and unmerged reads were aligned to the amplicon reference using the
683 NEEDLEALL (50) aligner with a gap open penalty of 10 and a gap extension penalty of 0.5. To
684 capture read-through, UMI degenerate bases and adapter sequences were included in the
685 reference amplicon sequence, and mismatches to Ns in the degenerate bases were set to a penalty
686 of 0. To eliminate off-target sequencing reads, aligned sequences were required to match greater
687 than 85% of bases at non-indel positions, to have correct PCR primer sequences on both the 5'
688 and 3' ends, and to match at least 50 bases of the reference sequence (including primer
689 sequences). Target sites were deemed edited if there was an insertion or deletion event present
690 within 3 bases of the predicted Cas9 cut site (3 nucleotides 5' of each PAM), or if a deletion
691 spanned the site entirely. Sites were marked as disrupted if there was not perfect alignment of the
692 barcode over the entirety of the reference target sequence. An edited barcode was then defined as
693 the complete list of insertion and deletion events (*i.e.* 'editing events') within the consensus
694 sequence for a given UMI.

695 Maximum parsimony lineage reconstruction

696 For lineage reconstruction (fig. S4B), recurrently observed barcode alleles within a single organ
697 or cell population were reduced to a single representative entry. We then used Camin-Sokal
698 maximum parsimony to reconstruct lineages, as implemented in the PHYLIP MIX software
699 package (26). Camin-Sokal maximum parsimony assumes that the initial cell or zygote is
700 unedited, and that editing is irreversible. To run MIX, a matrix was created where each row
701 corresponded to an allele, and each column corresponds to a unique editing event. Each entry in
702 this matrix is an indicator variable of presence or absence of a specific edit in that allele (1 or 0).
703 Events were also weighted by their log-abundance and scaled to the range allowed in MIX (0-Z).
704 MIX was run, and the output was parsed to recover edit patterns in ancestral nodes. Internal
705 parent-child nodes that had identical editing patterns were collapsed using the recovered
706 ancestral alleles and the output tree. When a parent node and child node share the same allele,
707 the grandchildren nodes were transferred to the parent and the child node was removed, creating
708 multifurcating parent nodes. The resulting tree was converted to an annotated JSON tree
709 compatible with our visualization tools. All code is available on the Shendure lab github website:
710 <https://github.com/shendurelab/Cas9FateMapping>.

711 Zebrafish husbandry

712 All vertebrate animal work was performed at the facilities of Harvard University, Faculty of Arts
713 & Sciences (HU/FAS). This study was approved by the Harvard University/Faculty of Arts &
714 Sciences Standing Committee on the Use of Animals in Research & Teaching under Protocol
715 No. 25–08. The HU/FAS animal care and use program maintains full AAALAC accreditation, is
716 assured with OLAW (A3593-01), and is currently registered with the USDA.

717 Cloning transgenesis vector

718 The transgenesis vectors pTol2-DRv6 and pTol2-DRv7 were constructed as follows. The v6 or
719 v7 array was cloned into the 3' UTR of a DsRed coding sequence under control of the ubiquitin
720 promoter (51). This cassette was placed in a Tol2 transgenesis vector containing a cmlc2:GFP
721 marker, which drives expression of GFP in the cardiomyocytes of the heart from 24 hpf to
722 adulthood (52). Plasmids are available from Addgene.

723 Generating transgenic zebrafish

724 To generate founder fish, 1-cell embryos were injected with zebrafish codon optimized Tol2
725 mRNA and pTol2-DR1v6 or pTol2-DR1v7 vector. Potential founder fish were screened for heart
726 GFP expression at 30 hpf and grown to adulthood. Adult founder transgenic fish were identified
727 by outcrossing to wild type and screening clutches of embryos for heart GFP expression at 30
728 hpf.

729 Transgene copy number quantification

730 To identify single copy Tol2 transgenics, copy number was quantified using qPCR (29). Briefly,
731 genomic DNA was extracted from candidate embryos or fin-clips of adult fish using the

732 HotSHOT method (53) and subjected to qPCR using a set of primers targeting DsRed and a set
733 targeting a diploid conserved region of the genome (table S4) and compared to reference non-
734 transgenic, 1-copy and 2-copy transgenic animals using the ddCt method.

735 Generation and delivery of editing reagents

736 sgRNAs specific to each site of the v6 or v7 array were generated as previously described (54),
737 except that sgRNAs were isolated after transcription by column purification (Zymogen). 1-cell
738 embryos resulting from an outcross of a transgenic founder were injected with two different
739 volumes (0.5 nl, 1/3x or 1.5nl, 1x) of Cas9 protein (NEB) and sgRNAs in salt solution (8 μ M
740 Cas9, 100 ng/ μ l pooled sgRNAs, 50 mM KCl, 3 mM MgCl₂, 5 mM Tris HCl pH 8.0, 0.05%
741 phenol red). Transgenic embryos were collected at the time points indicated in the text and
742 genomic DNA extracted as described below. To confirm editing, PCR was conducted on a subset
743 of samples using primers flanking the v6 or v7 array (table S4), and amplicons were loaded on a
744 2% agarose gel for electrophoresis.

745 Imaging

746 Embryos were anaesthetized and manually dechorionated in MS222, mounted in methylcellulose
747 and imaged using a Leica upright fluorescence microscope.

748 Organ Dissection

749 Adult edited single copy transgenic fish were isolated without food for one day to reduce food
750 particles in the gastrointestinal system, then anaesthetized in MS222 and euthanized on ice.
751 Before dissection, blood was collected using a centrifugation method (55). This collection
752 method greatly enriches for blood cells, particularly red blood cells, but also results in
753 contamination from skin or other tissues. The fish were pinned on a silicon mat and surgery was
754 conducted using sterile tools to remove organs as in (56). Organs were washed in PBS and, with
755 the exception of the heart, frozen in tubes on dry ice. A piece of heart tissue was collected before
756 the remainder of the heart was dissociated following manufacturer's instructions (Miltenyi # 130-
757 098-373). After dissociation, a sample of dissociated heart cells was collected (DHCs), and the
758 remaining cells sorted using a Beckman Coulter MoFlo XDP Cell Sorter through a series of three
759 gates to minimize debris and cell doublets, and then split into two additional populations: GFP+
760 cardiomyocytes and GFP- non-cardiomyocyte heart cells (NCs, fig. S10).

761 Genomic DNA preparation from zebrafish embryos and organs

762 Zebrafish embryo and adult organ gDNA was prepared using the Qiagen DNeasy kit. For heart
763 samples from cell sorting experiments, 1 μ l of poly-dT carrier DNA (25 uM) was added prior to
764 gDNA preparation. Digestion with proteinase K at 56° C was performed overnight for intact
765 organs (brain, eyes, gills, intestinal bulb, posterior intestine, and piece of heart) and for 30
766 minutes for blood samples, dissociated heart cells and embryos. gDNA was eluted in 100 μ l,
767 then concentrated using an Eppendorf Vacufuge for samples yielding less than 1 μ g.

768

769

770 **REFERENCES**

771

- 772 1. G. S. Stent, Developmental cell lineage. *Int. J. Dev. Biol.* **42**, 237–241 (1998).
- 773 2. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of
774 the nematode *Caenorhabditis elegans*. *Developmental Biology.* **100**, 64–119 (1983).
- 775 3. K. Kretzschmar, F. M. Watt, Lineage Tracing. *Cell.* **148**, 33–45 (2012).
- 776 4. C. B. Kimmel, R. D. Law, Cell lineage of zebrafish blastomeres. III. Clonal analyses of
777 the blastula and gastrula stages. *Developmental Biology.* **108**, 94–101 (1985).
- 778 5. R. E. Keller, Vital dye mapping of the gastrula and neurula of *Xenopus laevis*. I.
779 Prospective areas and morphogenetic movements of the superficial layer. *Developmental*
780 *Biology.* **42**, 222–241 (1975).
- 781 6. D. A. Weisblat, R. T. Sawyer, G. S. Stent, Cell lineage analysis by intracellular injection
782 of a tracer enzyme. *Science.* **202**, 1295–1298 (1978).
- 783 7. N. M. Le Douarin, M. A. Teillet, Experimental analysis of the migration and
784 differentiation of neuroblasts of the autonomic nervous system and of neurectodermal
785 mesenchymal derivatives, using a biological cell marking technique. *Developmental*
786 *Biology.* **41**, 162–184 (1974).
- 787 8. S. M. Dymecki, H. Tomasiewicz, Using Flp-recombinase to characterize expansion of
788 Wnt1-expressing neural progenitors in the mouse. *Developmental Biology.* **201**, 57–65
789 (1998).
- 790 9. D. L. Zinyk, E. H. Mercer, E. Harris, D. J. Anderson, A. L. Joyner, Fate mapping of the
791 mouse midbrain-hindbrain constriction using a site-specific recombination system.
792 *Current Biology.* **8**, 665–668 (1998).
- 793 10. C. Walsh, C. L. Cepko, Widespread dispersion of neuronal clones across functional
794 regions of the cerebral cortex. *Science.* **255**, 434–440 (1992).
- 795 11. R. Lu, N. F. Neff, S. R. Quake, I. L. Weissman, Tracking single hematopoietic stem cells
796 in vivo using high-throughput sequencing in conjunction with viral genetic barcoding.
797 *Nature Biotechnology.* **29**, 928–933 (2011).
- 798 12. S. N. Porter, L. C. Baker, D. Mittelman, M. H. Porteus, Lentiviral and targeted cellular
799 barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. **15**, 1–14
800 (2014).
- 801 13. S. J. Salipante, M. S. Horwitz, Phylogenetic fate mapping. *Proc Natl Acad Sci USA.* **103**,
802 5448–5453 (2006).

- 803 14. S. Behjati *et al.*, Genome sequencing of normal cells reveals developmental lineages and
804 mutational processes. *Nature*. **513**, 422–425 (2014).
- 805 15. M. A. Lodato *et al.*, Somatic mutation in single human neurons tracks developmental and
806 transcriptional history. *Science*. **350**, 94–98 (2015).
- 807 16. J. Livet *et al.*, Transgenic strategies for combinatorial expression of fluorescent proteins in
808 the nervous system. *Nature*. **450**, 56–62 (2007).
- 809 17. G. M. Church, J. Shendure, Nucleic acid memory device. President And Fellows Of
810 Harvard College, assignee. Patent US20030228611. N.d. Print.
- 811 18. C. A. Carlson *et al.*, Decoding cell lineage from acquired mutations using arbitrary deep
812 sequencing. *Nat Meth*. **9**, 78–80 (2011).
- 813 19. J.-H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing in situ. *Science*. **343**,
814 1360–1363 (2014).
- 815 20. R. Ke *et al.*, In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Meth*.
816 **10**, 857–860 (2013).
- 817 21. J. A. Doudna, E. Charpentier, Genome editing. The new frontier of genome engineering
818 with CRISPR-Cas9. *Science*. **346**, 1258096–1258096 (2014).
- 819 22. S. Q. Tsai *et al.*, GUIDE-seq enables genome-wide profiling of off-target cleavage by
820 CRISPR-Cas nucleases. *Nature Biotechnology*. **33**, 187–197 (2014).
- 821 23. Y. Sancak *et al.*, The Rag GTPases bind raptor and mediate amino acid signaling to
822 mTORC1. *Science*. **320**, 1496–1501 (2008).
- 823 24. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for
824 CRISPR screening. *Nature*. **11**, 783–784 (2014).
- 825 25. B. E. Miner, R. J. Stöger, A. F. Burden, C. D. Laird, R. S. Hansen, Molecular barcodes
826 detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Research*.
827 **32**, e135–e135 (2004).
- 828 26. J. Felsenstein, PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*. **5**, 164–
829 166 (1989).
- 830 27. K. Kawakami, Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol*. **8 Suppl**
831 **1**, S7 (2007).
- 832 28. S. N. Porter, L. C. Baker, D. Mittelman, M. H. Porteus, Lentiviral and targeted cellular
833 barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. *Genome Biol*.
834 **15**, 1–14 (2014).
- 835 29. Y. A. Pan *et al.*, Zebrabow: multispectral cell labeling for cell tracing and lineage analysis

- 836 in zebrafish. *Development*. **140**, 2835–2846 (2013).
- 837 30. C. Thisse, L. I. Zon, Organogenesis--heart and blood formation from the zebrafish point of
838 view. *Science*. **295**, 457–462 (2002).
- 839 31. Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, J. K. Joung, Improving CRISPR-Cas
840 nuclease specificity using truncated guide RNAs. *Nature Biotechnology*. **32**, 279–284
841 (2014).
- 842 32. B. P. Kleinstiver *et al.*, Engineered CRISPR-Cas9 nucleases with altered PAM
843 specificities. *Nature*. **523**, 481–485 (2015).
- 844 33. I. M. Slaymaker *et al.*, Rationally engineered Cas9 nucleases with improved specificity.
845 *Science*. **351**, 84–88 (2016).
- 846 34. R. J. Platt *et al.*, CRISPR-Cas9 knockin mice for genome editing and cancer modeling.
847 *Cell*. **159**, 440–455 (2014).
- 848 35. J. Ablain, E. M. Durand, S. Yang, Y. Zhou, L. I. Zon, A CRISPR/Cas9 vector system for
849 tissue-specific gene disruption in zebrafish. *Developmental Cell*. **32**, 756–764 (2015).
- 850 36. L. Yin *et al.*, Multiplex Conditional Mutagenesis Using Transgenic Expression of Cas9
851 and sgRNAs. *Genetics*. **200**, 431–441 (2015).
- 852 37. V. Gupta, K. D. Poss, Clonally dominant cardiomyocytes direct heart morphogenesis.
853 *Nature*. **484**, 479–484 (2012).
- 854 38. H. J. Snippert *et al.*, Intestinal crypt homeostasis results from neutral competition between
855 symmetrically dividing Lgr5 stem cells. *Cell*. **143**, 134–144 (2010).
- 856 39. A. M. Klein, B. D. Simons, Universal patterns of stem cell fate in cycling adult tissues.
857 *Development*. **138**, 3103–3111 (2011).
- 858 40. C. Blanpain, B. D. Simons, Unravelling stem cell dynamics by lineage tracing. *Nat Rev*
859 *Mol Cell Biol*. **14**, 489–502 (2013).
- 860 41. P. M. Henson, D. A. Hume, Apoptotic cell removal in development and tissue
861 homeostasis. *Trends Immunol*. **27**, 244–250 (2006).
- 862 42. J. Pellettieri, A. Sánchez Alvarado, Cell turnover and adult tissue homeostasis: from
863 humans to planarians. *Annu. Rev. Genet*. **41**, 83–105 (2007).
- 864 43. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of
865 single-cell gene expression data. *Nature Biotechnology*. **33**, 495–502 (2015).
- 866 44. D. A. Cusanovich *et al.*, Multiplex single-cell profiling of chromatin accessibility by
867 combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).
- 868 45. S. G. Megason, S. E. Fraser, Imaging in systems biology. *Cell*. **130**, 784–795 (2007).

- 869 46. Z. Liu, P. J. Keller, Emerging Imaging and Genomic Tools for Developmental Systems
870 Biology. *Developmental Cell*. **36**, 597–610 (2016).
- 871 47. F. Farzadfard, T. K. Lu, Genomically encoded analog memory with precise in vivo DNA
872 writing in living cell populations. *Science*. **346**, 1256272–1256272 (2014).
- 873 48. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina
874 sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).
- 875 49. T. Magoc, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve
876 genome assemblies. *Bioinformatics*. **27**, 2957–2963 (2011).
- 877 50. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open
878 Software Suite. *Trends in Genetics*. **16**, 276–277 (2000).
- 879 51. C. Mosimann *et al.*, Ubiquitous transgene expression and Cre-based recombination driven
880 by the ubiquitin promoter in zebrafish. *Development*. **138**, 169–177 (2011).
- 881 52. C.-J. Huang, C.-T. Tu, C.-D. Hsiao, F.-J. Hsieh, H.-J. Tsai, Germ-line transmission of a
882 myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac
883 myosin light chain 2 promoter of zebrafish. *Dev. Dyn*. **228**, 30–40 (2003).
- 884 53. N. D. Meeker, S. A. Hutchinson, L. Ho, N. S. Trede, Method for isolation of PCR-ready
885 genomic DNA from zebrafish tissues. *Biotech*. **43**, 610–612– 614 (2007).
- 886 54. J. A. Gagnon *et al.*, Efficient mutagenesis by Cas9 protein-mediated oligonucleotide
887 insertion and large-scale assessment of single-guide RNAs. *PLoS ONE*. **9**, e98186 (2014).
- 888 55. F. Babaei *et al.*, Novel blood collection method allows plasma proteome analysis from
889 single zebrafish. *J. Proteome Res*. **12**, 1580–1590 (2013).
- 890 56. T. Gupta, M. C. Mullins, Dissection of organs from the adult zebrafish. *J Vis Exp*, e1717–
891 e1717 (2010).

892