Multiomic single-cell lineage tracing to dissect fate-specific gene regulatory programs

Kunal Jindal[1,2,3], Mohd Tayyab Adil[1,2,3], Naoto Yamaguchi[1,2,3], Xue Yang[1,2,3], Helen C. Wang[4], Kenji Kamimoto[1,2,3], Guillermo C. Rivera-Gonzalez[1,2,3], and Samantha A. Morris[1,2,3].

[1]Department of Developmental Biology; [2]Department of Genetics; [3]Center of Regenerative Medicine. [4]Department of Pediatrics, Division of Hematology and Oncology, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110, USA. *Correspondence: s.morris@wustl.edu

Abstract

Complex gene regulatory mechanisms underlie differentiation and reprogramming. Contemporary single-cell lineage tracing (scLT) methods use expressed, heritable DNA barcodes to combine cell lineage readout with single-cell transcriptomics enabling high-resolution analysis of cell states while preserving lineage relationships. However, reliance on transcriptional profiling limits their adaptation to an ever-expanding tool kit of multiomic single-cell assays. With CellTag-multi, we present a novel approach for independently profiling lineage barcodes with single-cell chromatin accessibility without relying on co-assay of transcriptional state, paving the way for truly multiomic lineage tracing. We validate CellTag-multi in mouse hematopoiesis, characterizing transcriptional and epigenomic lineage priming across progenitor cell populations. In direct reprogramming of fibroblasts to endoderm progenitors, we use CellTag-multi to comprehensively link early cell state with reprogramming outcomes, identifying core regulatory programs underlying on-target and off-target reprogramming. Further, we reveal the Transcription Factor (TF) Zfp281 as a novel regulator of reprogramming outcome, biasing cells towards an off-target mesenchymal fate via its regulation of TGF-β signaling. Together, these results establish CellTag-multi as a novel lineage tracing method compatible with multiple single-cell modalities and demonstrate its utility in revealing fate-specifying gene regulatory changes across diverse paradigms of differentiation and reprogramming.

Keywords: Single-cell lineage tracing; Transcriptomics; Chromatin accessibility; Epigenomics; Multiomics; Hematopoiesis; Reprogramming.

Introduction

The quantification of cell identity is crucial to understanding development, disease, and homeostasis, yet the notion of cell identity remains poorly defined[1]. Single-cell technologies, now tailored to diverse modalities[2], are expanding our understanding of how cell identity is established and maintained[3]. In particular, single-cell lineage tracing (scLT) methods allow cell relationships to be tracked throughout biological processes, revealing cell fate decisions during differentiation and reprogramming[4,5]. Prospective scLT methods label cells with unique genetic 'barcodes' that are expressed as RNA; capturing these barcodes via single-cell RNA-seq (scRNA-seq) allows the parallel capture of lineage information and single-cell transcriptomes[6–13].

These methods to barcode and track cells have been deployed across several *in vitro* differentiation and reprogramming paradigms[5,14]. The accessibility of cells within these systems permits longitudinal sampling and cellular barcoding at precise time points, allowing early progenitor state to be linked to terminal fate (termed 'state-fate analysis'; **Fig. 1a**). Such a strategy has been used to determine how well gene expression state in progenitors reflects eventual cell fate in hematopoiesis[13]. This work demonstrated that subsequent fate could be predicted, albeit with limited accuracy, from progenitor gene expression, indicating the existence of heritable fate determinants that are not captured by scRNA-seq alone. Similarly, viral barcoding, 'CellTagging,' of transcription factor-mediated direct reprogramming of mouse embryonic fibroblasts (MEFs) to induced endoderm progenitors (iEPs), suggested that reprogramming outcome is determined during the early stages of fate conversion[7]. However, the early gene regulatory changes that set cells on their destined path have not been fully characterized. Additional information from epigenomic assays such as single-cell Assay of Transposase Accessible Chromatin by sequencing (scATAC-seq) may be crucial to uncover the heritable properties that play a key role in the establishment and maintenance of cell identity. Previously, natural DNA variation has been used to infer coarse cellular phylogenies with scATAC-seq[15,16]. However, the resolution of such retrospective methods is limited due to their reliance on the accrual of somatic mutations. In contrast, the density of lineage information recorded can be precisely controlled at biologically relevant time points using successive rounds of cellular barcoding[7,17] with prospective methods. This is essential for profiling early, lineage-specific responses in dynamic systems like differentiation and reprogramming.

To enable prospective lineage tracing with chromatin accessibility capture, we have developed 'CellTag-multi.' CellTag-multi is based on our previous CellTagging technology, which uses sequential lentiviral delivery of CellTags (heritable random barcodes) to enable the construction of multi-level lineage trees[7,17]. Here, we introduce a new strategy in which CellTags,

expressed as polyadenylated transcripts, can be captured in both scRNA-seq and scATAC-seq assays allowing for independent tracking of clonal transcriptional and epigenomic state.

We validate this method using *in vitro* hematopoiesis, a well-characterized model of multi-lineage differentiation, and demonstrate highly accurate reconstruction of lineage relationships and capture of lineage-specific progenitor cell states across scRNA-seq and scATAC-seq. Moreover, the addition of chromatin accessibility information to gene expression allows for a significant improvement in the prediction of differentiation outcome from early progenitor state. We also deploy CellTag-multi in the direct lineage reprogramming of fibroblasts to induced endoderm progenitors (iEPs), to characterize early genomic events in rare subpopulations of cells that successfully reprogram. This application reveals how chromatin is remodeled following expression of reprogramming TFs, enabling deeper insight into gene regulatory network reconfiguration. We uncover the TF Foxd2 as a facilitator of on-target reprogramming, increasing the efficiency of MEF to iEP conversion. Conversely, we identify Zfp281 as a TF biasing cells towards an off-target mesenchymal fate via its regulation of TGF-β signaling, which we validate experimentally. We demonstrate that the identification of these TFs as novel reprogramming regulators is only possible via multiomic profiling. Together, these findings highlight the utility of CellTag-multi in defining the molecular regulation of early cell state and its relation to fate across diverse biological applications.

Development and validation of CellTag-multi

CellTagging relies on single-cell capture of CellTags — heritable DNA barcodes expressed as polyadenylated transcripts[7,17,18]. In the standard workflow, CellTags are captured as transcripts and reverse transcribed (RT), along with cellular mRNA, during 3' end scRNA-seq library preparation. In contrast, scATAC-seq directly captures fragments of the accessible genome, omitting capture of CellTag transcripts, rendering CellTagging incompatible with scATAC-seq assays. To enable CellTag profiling with scATAC-seq, we introduced two essential modifications. First, we developed an *in situ* Reverse Transcription (isRT) step to selectively reverse transcribe CellTag barcodes inside intact nuclei. By introducing this additional step after transposition, we omitted the need to RT CellTags during scATAC-seq library construction. Second, we modified the CellTag construct to flank the random barcode with Nextera Read 1 and Read 2 adapters (**Fig. 1b, Ext Fig. 1a, b**).

During scATAC-seq library preparation, nuclei are partitioned into nanoliter droplets along with single-cell barcoding beads and PCR reagents. Each bead contains a barcoded forward primer complementary to the Nextera Read 1 adapter to barcode and linearly amplify all ATAC

fragments during the GEM incubation step. By inserting Nextera Read 1 and Read 2 adapters in the CellTag construct, we enabled single-cell capture of reverse transcribed CellTags along with accessible chromatin during the GEM incubation stage (**Fig. 1c, Ext Fig. 1b**). This strategy improved the CellTag capture rate by >200-fold compared to the unmodified scATAC-seq protocol (**Ext Fig. 1c**). Additionally, we introduced a reverse primer specific to the CellTag cDNA during GEM incubation to exponentially amplify CellTag fragments, while ATAC fragments undergo linear amplification (**Supplementary Table 1, Ext Fig. 1b**). Together, these modifications led to a >50,000-fold increase in CellTag capture (**Ext Fig. 1c**), with CellTags being detected in >96% of cells in scATAC-seq relative to 98% in scRNA-seq (**Ext Fig. 1d**), without negatively impacting scATAC-seq data quality or genome-wide chromatin accessibility signal (**Fig. 1d, e, Ext Fig. 1e, f**).

To support the accurate identification of clonally related cells, it is essential that CellTag signatures from individual cells are captured with high fidelity, minimizing background noise. To assess the fidelity of CellTag signatures captured in scATAC-seq, we performed a species-mixing experiment (**Ext Fig. 2a**). We labeled human (HEK 293T) cells and mouse (expanded iEPs) cells with two different versions of the CellTag-multi library, combined nuclei isolated from both populations in a 1:1 ratio and profiled them using our modified scATAC-seq method. Plotting CellTag reads/cell, we observed that nuclei from each species predominantly consisted of reads from the expected CellTag library, indicating minimal inter-species crosstalk (**Fig. 1f; Ext Fig. 2b, c**).

Finally, to perform large-scale lineage tracing experiments, we synthesized a complex CellTag-multi library containing ~80,000 unique barcodes, as confirmed by sequencing (**Methods**). We applied CellTag-multi to a population of expanded mouse fibroblasts undergoing reprogramming to iEPs and profiled clones with scRNA-seq and scATAC-seq, detecting CellTags in 70% (RNA) and 51% (ATAC) of the cells at an average MOI of 2 (RNA) and 2.5 (ATAC). Filtering, error-correction, and allowlisting of CellTag reads (**Methods**) enabled high-fidelity identification of distinct clones across the two single-cell modalities (**Fig. 1g, h, Ext Fig. 2d-f**). As expected, the correlation between gene expression and accessibility was higher within clones than across clones (**Fig. 1i, j**). These analyses established the efficacy of CellTag-multi for the labeling and capture of clonally related cells across scRNA and scATAC modalities. Next, we leveraged CellTag-multi to link early state with cell fate in diverse cell fate specification and reprogramming paradigms.

Benchmarking CellTag-multi using an *in vitro* model of mouse hematopoiesis

135 To validate lineage analysis across single-cell modalities with CellTag-multi, we applied it to
136 hematopoiesis, a well-characterized paradigm for multi-lineage differentiation. Recently, scLT
137 was used to define the early transcriptional cell states that lead to defined differentiation outcomes
138 in mouse hematopoiesis. However, these analyses suggested that early transcriptional changes
139 alone cannot fully define future cell fate and posited a role for cell states that evade transcriptional
140 profiling, collectively termed hidden state variables[13]. In this context, we aimed to apply CellTag-
141 multi to further refine state-fate linkages in early hematopoiesis by identifying fate-specific
142 changes in both early gene expression and chromatin accessibility.

143 We isolated Lin$^-$, Sca1$^+$, c-Kit$^+$ (LSK) cells from adult mouse bone marrow and cultured
144 them in broad myeloid differentiation media[13]. Upon isolation, we tagged these cells with the
145 CellTag-multi library to track clones across modalities. To capture both early state and fate across
146 clones, we profiled half of the cells 60 hours after initiation of differentiation (Day 2.5; state
147 sample), re-plated the remaining cells across two technical replicates, and collected them for
148 sequencing on Day 5 (fate sample). In the case of both samples, cells were split between scRNA-
149 seq and scATAC-seq (**Fig. 2a**), resulting in the profiling of 9,789 state cells (scRNA-seq: n=5,161;
150 scATAC-seq: n=4,628) and 67,029 fate cells (scRNA-seq: n=56,534; scATAC-seq: n=10,495
151 cells), after quality filtering (**Ext Fig. 3a, b**). We identified cells from all major hematopoietic
152 lineages across single-cell modalities (**Fig. 2b, Ext Fig. 3c**). CellTagging was consistent across
153 single-cell modalities, yielding 83-99% labeled cells.

154 To compare clonal analysis across modalities, we first analyzed the scRNA-seq and
155 scATAC-seq datasets separately and identified clones in each modality independently (**Ext Fig**.
156 **3d**). Lineage hierarchies inferred using clonally related cells (**Methods**) were consistent across
157 scRNA and scATAC despite the chromatin dataset comprising fewer cells, demonstrating the
158 ability of CellTag-multi in defining fate relationships using clonal scATAC-seq data alone (**Fig. 2c,**
159 **d**). Assigning a fate label to each clone, based on the most abundant cell type amongst its Day 5
160 sisters, allowed mapping of coarse fate trajectories on the 2D embeddings (**Fig. 2e, Ext Fig. 3e**).

161 Joint clone calling across both datasets led to an increase in number of cells tracked (**Ext**
162 **Fig. 3f**), likely due to clones that are split across modalities (multiomic clones). We identified a
163 total of 37,441 scRNA-seq cells in 5,973 clones and 6,098 scATAC-seq cells in 3,012 clones,
164 labeled with 4.2 CellTags/cell (in scRNA-seq) and 3.4 CellTags/cell (in scATAC-seq) on average
165 (**Ext Fig. 3g, h**). 2,227 clones spanned both state and fate samples, including 877 multiomic
166 clones. These clones were used for the remainder of the analyses.

167 For visualization, we co-embedded cells from both modalities using Canonical Correlation
168 Analysis (CCA)[19]. Further, we devised a unique clone-cell co-embedding approach to include

169 clones as individual data points in a single-cell embedding, enabling straightforward visualization
170 and assessment of clone-level metadata and global trends across clones (**Ext Fig. 3i**). We first
171 extracted the cell-cell similarity graph, produced as part of standard single-cell analysis workflows.
172 In this graph, each cell is represented by a node and the connection between a pair of cells is
173 weighted based on their phenotypic similarity. Next, we imputed abstract clone nodes and clone-
174 cell edges to this graph based on clonal data. Finally, we used this expanded clone-cell graph as
175 input for dimensionality reduction algorithms such as UMAP[20] or ForceAtlas[21] to produce a single
176 2D-embedding of the data, where both cells and clones are represented by individual points. We
177 applied this visualization to the hematopoiesis data to co-embed RNA and ATAC cells with all
178 clones, with minimal impact on the underlying structure of the data (**Fig. 2f, Ext. Fig. 3j, k**).
179 Clones, now represented as individual data points, faithfully represented their constituent cells
180 (**Ext Fig. 3l**) and can be used to visualize clonal metadata across all cells (**Fig. 2f, right panel**).
181 Consistent with previous reports, we observe continuous transitions from progenitor populations
182 to distinct hematopoietic lineages across modalities, as previously reported[13,22,23] (**Ext Fig. 4a-c**).
183 While CellTag capture was uniform across cell states (**Ext Fig. 4d**), we observed higher clonal
184 expansion along the monocyte lineage, consistent with our myeloid differentiation culture
185 conditions (**Fig. 2f right panel, g**).

186 We linked Day 2.5 cell state with Day 5 fate, by re-assigning each clone, from the joint
187 clone calling results, a fate label based on the most abundant cell type amongst its Day 5 sisters
188 (**Fig. 2h**, **Ext Fig. 4e**). To map early clonal state along the differentiation continuum, we extended
189 our clone-cell embedding approach further and split each clone into sub-clones (up to four) based
190 on the assay and time point capture of each sister (**Ext Fig. 4f**). While Day 5 fate sub-clones
191 localized largely within their respective cell fate clusters (**Ext Fig. 4g**), Day 2.5 state sub-clones
192 associated with each major fate formed distinct groups closer to the undifferentiated progenitors
193 (**Fig. 2i, j**), suggesting early functional priming of immature cells. Moreover, state sub-clones
194 within the same 'fate potential' group overlapped significantly across single-cell modalities (Mann
195 Whitney Wilcoxon test; p-value = 3.76e-5, **Fig. 2j, k**), demonstrating high-fidelity capture of state-
196 fate linkages across transcriptional and epigenomic states with CellTag-multi. Projecting fate bias
197 scores, defined as the fraction of fate sisters belonging to the assigned clonal fate, on to state
198 sub-clones, we observed that low fate bias clones occupied areas closer to the overlapping
199 boundaries of each fate potential region, likely indicating areas of multi-potency (**Fig. 2j, Ext Fig.
200 4h**).

201 To characterize these fate-specific changes in early cell state on a molecular level, we
202 assessed the enrichment of transcriptional and epigenetic signatures in Day 2.5 sisters for each

203 fate group (**Fig. 2l; Methods**). Using gene expression, we identified several known fate-specific
204 markers in each group, such as *Spp1*[13] and *Ms4a3*[24] in the Monocyte primed group; *Elane* and
205 *Ctsg*[13] in the Neutrophil primed group; *Pf4*[25] and *Gata2*[13] in the Erythroid/Megakaryocyte groups.
206 In the Lymphoid group, we identified *Flt3*, a predominantly Lympho-myeloid gene[26], and several
207 lymphoid fate-specific genes such as *Mef2c*[27] and *Bcl11a*[28]. For epigenetic data, we focused on
208 TF activity scores[29], which estimate the enrichment of TF motifs in single-cell epigenomes[29].
209 Unlike peak accessibility, TF activity feature space is dense and continuous, allowing comparison
210 between small groups of cells, and is easier to interpret relative to individual peak features[29]. TF
211 activity enrichment analysis revealed several expected lineage specifying TFs for each fate[22,30],
212 such as several CEBP TFs enriched in Monocyte and Neutrophil primed groups; GATA1 and
213 GATA2 in the Erythroid/Megakaryocyte and Basophils/Eosinophils/Mast cells groups; Lympho-
214 myeloid TF SFPI1 (also known as PU.1) in the Lymphoid and Dendritic Cells (DC) group, along
215 with BCL factors and MEF2 factors, indicating extensive epigenomic priming in early cells towards
216 their respective cell fate. A complete list of differential gene expression and TF activity enrichment
217 can be found in Supplementary Table 2.
218
219 **Chromatin accessibility and gene expression jointly define fate predictive cell state**
220 Our above state-fate analysis suggests that lineage-specific changes in gene expression are
221 accompanied by extensive epigenetic remodeling, rendering the genome more accessible to fate-
222 specifying TFs. Previous analysis has suggested that cell states hidden from transcriptional
223 profiling play a role in fully defining fate-associated changes in cell state[13]. Changes in chromatin
224 accessibility could account for some of this hidden variance and we tested this hypothesis by
225 assessing whether cell fate can be accurately predicted from early state using our multiomic clonal
226 data.
227 We trained machine learning models to predict clonal cell fate from gene expression or
228 chromatin accessibility profiles of Day 2.5 sisters (**Ext Fig. 5a**). We tested three different
229 architectures: Logistic Regression, Random Forest, and LightGBM, and assessed model
230 performance using prediction accuracy (**Ext Fig 5b**). Overall, Random Forest models performed
231 the best and were used for all downstream analysis. For gene expression, we trained a
232 classification model to predict clonal fate using expression of the three thousand most highly
233 variable genes (HVG) and obtained an accuracy of 75.6% (**Fig. 2m, Ext Fig. 5c**). For chromatin
234 accessibility, we used Day 2.5 imputed TF activity scores (**Methods**) for 884 TF motifs to predict
235 clonal fate and obtained an accuracy of 72.7% (**Fig. 2m**). Notably, an RNA model trained on
236 expression levels of TFs, obtained from the Catalog of Inferred Sequence Binding Preferences

237 (CIS-BP) database, only scored only 63.8% on prediction accuracy (**Fig. 2m**). The significantly
238 lower predictive performance of TF expression compared to TF activity could be attributed to
239 either technical dropout in scRNA-seq or significantly higher lineage specific priming of TF binding
240 sites compared to TF expression, or a combination of both.

241 To assess fate-specific priming in different functional regions of the epigenome, we
242 computed TF activity scores using subsets of accessible peaks and compared fate prediction
243 performance across these feature spaces. Specifically, we computed TF activity scores using
244 only promoter, distal, exonic, or intronic peaks and trained fate prediction models with each. We
245 observed significant variation in performance between different ATAC models, indicating different
246 levels of fate-specific epigenetic priming across functional regions of the genome (**Ext Fig. 5d**).
247 This variation was independent of the number of peaks used to compute each set of TF activity
248 scores (**Ext Fig. 5d**). Distal and Intronic were the best performing models, comparable in
249 performance to the full peak set model ('All'). Promoter and Exonic models performed significantly
250 worse, suggesting that fate-specifying epigenetic changes during these early stages were
251 dominated by changes in distal regulatory regions of the epigenome rather than accessibility of
252 genes themselves. This observation is reinforced by the persistence of TF enrichment trends
253 across state groups in distal and intronic subsets but not in the exonic and promoter subsets (**Ext**
254 **Fig. 5e**). We confirmed these results using SHAP, a game theoretic approach to quantify the
255 contributions of individual input features in explaining the output of a machine learning model[31].
256 Indeed, SHAP analysis showed that in the better-performing models, an increase in CEBP/A motif
257 accessibility and an increase in MECOM motif accessibility were better predictors of Monocyte
258 and Ery/Meg fates, respectively, suggesting a lack of functional priming in the promoter-proximal
259 accessible genome (**Ext Fig. 5f, g**).

260 Finally, we tested whether combining RNA and ATAC features is more predictive of fate
261 than either individual modality. For this, we trained a combined RNA and ATAC model where RNA
262 and ATAC Day 2.5 sister cells within the same clone were paired randomly, and their combined
263 gene expression and TF activity signatures were used to predict clonal fate label. This analysis
264 was limited to multiomic state-fate clones. The combination of both state modalities was
265 significantly better at predicting fate (mean accuracy score = 86.5%) compared to either individual
266 modality or pairs of unrelated RNA and ATAC state cells (**Fig. 2m**). These results show that both
267 gene expression and chromatin accessibility jointly comprise cell states that define future cell fate.
268 Moreover, these modalities consist of non-redundant and highly complementary state information,
269 as a combination of both predicts cell fate much more accurately than each modality in isolation.
270

271 **Dissecting clonal dynamics of direct reprogramming**

272 Our application of CellTag-multi to hematopoiesis demonstrated the method's utility to capture
273 informative gene regulatory dynamics in a well-characterized differentiation paradigm. We next
274 applied CellTag-multi to a less defined system — the direct reprogramming of MEFs to iEPs driven
275 by retroviral overexpression of Hnf4α and Foxa1[7,32,33]. Direct lineage reprogramming presents a
276 unique paradigm of cell identity conversion, with cells often transitioning through progenitor-like
277 states or acquiring off-target identities[34,35]. Such non-linear fate dynamics can be challenging to
278 assess, especially when relying solely on the computational inference of cell fate trajectories[12].
279 Ground truth lineage tracing serves as a crucial resource for dissecting lineage-specific cell state
280 changes during direct reprogramming[7]. Originally reported to yield hepatocyte-like cells[32], we
281 have previously shown that Hnf4α and Foxa1 overexpression in MEFs generates cells with the
282 broader potential to functionally engraft liver and intestine[18,33,36]. This prompted their re-
283 designation as 'induced Endoderm Progenitors' (iEPs). More recently, we have further
284 characterized the similarity of long-term cultured iEPs to regenerating Biliary Epithelial Cells
285 (BECs)[37].

286 Using our original CellTag-based lineage tracing, we identified two distinct iEP
287 reprogramming trajectories: a successful 'reprogrammed' trajectory, characterized by
288 endodermal and hepatic gene expression, and a 'dead-end' trajectory, defined by a failure to
289 extinguish the starting fibroblast identity[7]. Further work demonstrated key functional differences
290 between these fates, with successfully reprogrammed cells harboring the potential to engraft
291 acutely damaged mouse intestine[18]. Our previous lineage tracing suggests that the
292 reprogrammed and dead-end fates are determined in the early stages of fate conversion[7].
293 However, our original CellTagging methodology did not capture any epigenetic information and
294 only sparsely sampled early state clones, limiting mechanistic insight into these initial
295 reprogramming stages.

296 Here, we deployed CellTag-multi in iEP reprogramming, modifying our clonal resampling
297 strategy to optimize state-fate analysis (**Fig. 3a**). First, we transduced MEFs with Hnf4α and
298 Foxa1 for 48 hours to initiate reprogramming, in two independent biological replicates. During the
299 last 12 hours of this 48-hour period, we transduced cells with the complex CellTag-multi library,
300 enabling clonal relationships to be tracked. 72 hours following the final viral transduction
301 (Reprogramming Day 3), we collected two-thirds of the cells for single-cell RNA and ATAC
302 profiling (state sample) and re-plated the remaining cells. Subsequent samples were collected on
303 Days 12 and 21 (fate samples) to assess reprogramming outcome. We also profiled the starting
304 MEF population (scATAC-seq, this study; scRNA-seq from a previous study[7])), resulting in a total

305 of 466,459 single-cells (scATAC-seq: 223,686; scRNA-seq: 242,863) in the final dataset after
306 quality filtering (**Ext Fig. 6a, b**). We identified a total of 8,502 clones, containing 46,438 cells
307 (Replicate 1: 3,068 clones; Replicate 2: 5,416 clones, average clone sizes of 4.8 and 5.9
308 cells/clone, respectively (**Ext Fig. 6c, d**)). We identified 1,428 'state-fate' clones across both
309 replicates, defined as clones that spanned state (Day 3) and at least one fate time point, Day 12
310 or Day 21 (**Ext Fig. 6d**).

311 Following dimensionality reduction and clustering of the co-embedded RNA and ATAC
312 datasets, clone-cell embedding was performed (**Fig. 3b, Ext Fig. 6e, f, g**). We annotated Day 12
313 and 21 fate clusters ('reprogrammed,' 'dead-end,' and 'transition') based on expression and
314 accessibility of known reprogramming associated genes, and unsupervised cell-type classification
315 based on transcriptional state using Capybara[37] (**Fig. 3c; Ext Fig. 7a, b**). In line with our previous
316 reports[7,18,37,38], reprogrammed cells express epithelial and iEP markers, *Cdh1* and *Apoa1*,
317 respectively. Dead-end cells are characterized by the retention of fibroblast gene expression but
318 are still transcriptionally distinct from MEFs, expressing low levels of iEP markers and several
319 dead-end-specific genes such as *Sfrp1,* a Wnt signaling modulator[7] (**Ext Fig. 7b, c**). Transition
320 cells represent states in between MEFs and reprogrammed/dead-end identities. Following cluster
321 annotation, we assigned fate labels to each state-fate clone. As the majority of state-fate clones
322 showed high fate-bias, we assigned clonal fate based on the most abundant cell annotation
323 amongst the fate sisters (**Fig. 3d**), identifying 1,009 reprogrammed, 2,493 dead-end and 1,371
324 transition clones. Dead-end and reprogrammed clones displayed a lineage-specific increase in
325 accessibility of known marker genes (**Fig. 3e**).

326 Using clonal information, we linked each reprogrammed and dead-end clone to its Day 3
327 state sisters, allowing us to track changes in cell identity longitudinally (**Fig. 3f**). These results
328 were consistent when clonal analysis was performed for each modality independently (**Ext Fig.
329 7d-f**). Comparing Capybara transcriptional cell identity scores across lineages, we found that iEP
330 identity scores were consistently higher along the reprogrammed lineage compared to the dead-
331 end lineage. MEF identity scores, while significantly higher along the dead-end lineage, exhibited
332 a steep decline after Day 12 coinciding with an increase in dead-end transcriptional identity score
333 (**Fig. 3g**). This suggested a delayed departure from MEF identity to an alternate cell state. We
334 observed high levels of clonal expansion along both lineages (**Fig. 3h, i**). These observations
335 suggest that despite retaining expression of canonical fibroblast marker genes, dead-end cells
336 are a fundamentally distinct, off-target cell state and reprogramming outcome. Thus, the
337 'reprogrammed' and 'dead-end' fates are better described as 'on-target' and 'off-target'
338 reprogramming, respectively.

339

340 **Linking early state to fate reveals molecular features of off-target reprogramming**

341 Next, to identify early state changes that regulate entry onto distinct fate trajectories, we focused
342 on Day 3 state clones destined to on-target (reprogrammed) or off-target (dead-end)
343 reprogramming fates. From assessing the distribution of Day 3 sisters destined to either of the
344 two fates, it is evident that they are not localized to defined clusters (**Ext Fig. 8a, b**). Further,
345 trajectory inference using CellRank[39] fails to reveal these initial states (**Ext Fig. 8c**),
346 demonstrating the importance of ground truth lineage tracing. We found that both Day 3 gene
347 expression and TF activities were highly predictive of clonal fate. Similar to our analysis of
348 hematopoiesis, fate prediction accuracy was significantly higher when both modalities were
349 considered, as compared to either modality individually. Further, distal and intronic peaks were
350 more predictive of fate than proximal and exonic (**Ext Fig. 8d, e**).

351 To identify early molecular signatures of lineage specification, we compared gene
352 expression, chromatin accessibility, and TF activity scores across MEFs and Day 3 state sisters
353 grouped by fate outcome. Comparing gene expression enrichment across the three groups, 2,116
354 genes were differentially enriched with 1,582 enriched genes uniquely defining each group (**Fig.
355 4a, Ext Fig. 8f**). While some genes displayed transient fate-specific expression, others
356 consistently increased over time in a lineage-specific manner (**Supplementary Table 3**). Early
357 iEP marker genes such as *Apoa1* were enriched in both on- and off-target trajectories on Day 3,
358 consistent with our previous observation that most cells initiate reprogramming[7] (**Ext Fig. 8f, g**).
359 On-target (reprogrammed) enriched genes included *Krt19*, a marker of BECs, Wnt signaling
360 associated genes *Wnt4*, *Anxa8,* and epithelial marker *Ezr* (**Fig. 4b, Supplementary Table 4**).
361 Top off-target (dead-end) related genes included canonical smooth muscle markers *Acta2* and
362 *Tagln* and other mesenchymal genes such as *Ptn,* and *Ncam1*, suggesting broad engagement of
363 mesenchymal programs, in addition to *Sfrp1*, a Wnt signaling pathway inhibitor (**Fig. 4b,
364 Supplementary Table 4**).

365 Comparing genome-wide chromatin accessibility revealed 21,720 Differentially Enriched
366 Regions (DERs) across Day 3 on-target and off-target destined cells and uninduced MEFs,
367 indicating extensive fate-specific epigenetic reconfiguration during early reprogramming (**Fig. 4c,
368 Supplementary Table 5**). DERs were enriched for distal and intergenic peaks, suggesting
369 epigenetic re-patterning of distal regions as a driver of cell fate conversion, consistent with our
370 above observations in hematopoiesis (**Ext Fig. 8h**). Motif analysis revealed enrichment of
371 reprogramming and hepatic TFs in on-target DERs, and several TFs with documented roles in
372 mesenchymal fates[40,41] in off-target DERs (**Ext Fig. 8i, j**). Using our paired RNA and ATAC data,

373 we linked accessible peaks to genes and identified 37,058 putative cis-regulatory elements
374 (CREs)[42] (**Fig 4c**, **Methods**). Gene-linked peaks were enriched for enhancer-like signatures
375 (ELS) from the ENCODE candidate CRE database[43] (**Methods**, **Ext Fig. 8k**). Genes linked to on-
376 target and off-target DERs displayed fate-specific expression patterns (**Fig. 4d, Ext Fig. 8l**). On-
377 target DERs consisted of several CREs linked to endodermal genes, such as *Alb*, *Foxq1,* and
378 *Creb3l2.* In contrast, off-target DERs contained CREs linked to mesenchymal genes such as
379 *Ncam1*, a modulator of Mesenchymal Stromal Cell migration[44], *Fbln2*, a mesenchymal gene
380 associated with embryonic heart development[45], and *Vegfd*, a regulator of angiogenesis[46] and
381 endothelial differentiation of bone marrow-derived mesenchymal stem cells[47] (**Fig. 4c;**
382 **Supplementary Table 5**). In several instances, this analysis captured lineage-specific changes
383 in accessibility of CREs before significant changes in gene expression were detected. For
384 instance, a *Vegfd*-linked CRE overlapping with an ENCODE enhancer displayed enrichment in
385 dead-end destined cells (Day 3), while expression changes were not detectable until Day 12.
386 Similar regulatory changes were observed for *Aox3*[48], a liver-associated aldehyde oxidase, and
387 *Col28a1,* an oligodendrocyte enriched collagen[49], prior to changes in gene expression (**Fig. 4e,**
388 **Supplementary Table 5**).
389 To identify functional changes in chromatin accessibility on a genomic scale, we compared
390 inferred TF activities across on-target and off-target destined cells and uninduced MEFs. To
391 preclude potential false positives, we discarded all TFs with low correlation (< 0.3) with their
392 respective gene activity scores, identifying 47 uniquely enriched TFs (**Fig. 4f, Ext Fig. 8m,**
393 **Supplementary Table 6**). On-target destined cells were highly enriched for the two
394 reprogramming TFs, FOXA1*,* and HNF4A. Other on-target associated TFs included FOXD2,
395 FOXO1*,* and NR1H3, a hepatic fate-specifying TF[50] (**Fig. 4f**). We identified a set of nine TFs
396 uniquely enriched in off-target destined cells (**Fig. 4f (black bar), g**). Several of these TFs
397 (Zfp281, Cebpb, Gata6, Hivep3) have been previously documented to play a role in regulating
398 mesenchymal cell identities[51–54]. Surveying the expression data, none of the off-target TFs display
399 a similar fate-biased enrichment (**Fig. 4g, Ext Fig. 8n**), highlighting the importance of lineage-
400 specific chromatin profiling in identifying these targets. This lack of enrichment could be due to
401 technical dropout during scRNA-seq or due to secondary mechanisms regulating the genomic
402 engagement of these TFs.
403 Altogether, our lineage-specific multiomic assessment of iEP generation demonstrates
404 clear early molecular differences associated with reprogramming outcomes. Indeed, from as early
405 as reprogramming day 3, cells on the dead-end lineage exhibit unique characteristics. Rather
406 than retaining MEF identity, we observe that the dead-end lineage constitutes a highly

407 proliferative, mesenchymal cell state with unique markers and regulatory changes, thus
408 representing an 'off-target' reprogrammed state. The early specification of this state is supported
409 by our GRN inference using CellOracle[38], suggesting that network reconfiguration is unique to
410 each trajectory and is established early in the reprogramming process. CellTag-multi has the
411 potential to define the molecular features of these early states, offering deeper mechanistic insight
412 into the reprogramming process.
413
414 **Foxd2 and Zfp281 as drivers of on- and off-target reprogramming**
415 Higher accessibility of both motifs and genomic targets[55] of FOXA1 and HNF4A in on-target cells
416 on Day 3 suggests significant differences in genomic engagement of the reprogramming TFs
417 between the two fate outcomes (**Fig. 5a, Ext Fig. 9a**). This could, at least in part, be explained
418 by differential expression levels of the *Hnf4α-Foxa1* transgene across the two lineages, with off-
419 target destined cells displaying significantly lower transgene expression (**Fig. 5a**; Mann Whitney
420 Wilcoxon test; p-value = 6.5e-42). However, we have also previously described an off-target
421 trajectory expressing high transgene levels, suggesting additional mechanisms influencing
422 genomic engagement by the reprogramming TFs[38].
423 Outside of FOXA1, and HNF4A, we identified FOXD2 as the top on-target fate-specifying
424 TF candidate (**Fig. 5b, Ext Fig. 9b**). Adding Foxd2 to the Foxa1 and Hnf4α reprogramming
425 cocktail led to significantly increased expression of the iEP marker *Cdh1* and decreased
426 expression of mesenchymal marker *Tagln* on reprogramming day 12 (t-test; p-values: *Cdh1* =
427 0.03; *Tagln* = 0.006; 2 biological replicates; 2 technical replicates **Ext Fig. 9c**). In addition, colony
428 formation assays showed a significant increase in the number of CDH1-positive colonies formed
429 with the addition of Foxd2 to the standard iEP reprogramming cocktail (t-test; p-value=0.045; 2
430 biological replicates; **Fig. 5c**), validating its role in improving on-target fate conversion.
431 The top off-target-enriched candidate was ZFP281, a Zinc Finger protein (**Fig. 5d, Ext**
432 **Fig. 9d**). Zfp281 is a known regulator of cell fate in mouse embryonic stem cells[56] and promotes
433 epithelial-to-mesenchymal transitions[57]. To further confirm the inferred enrichment of ZFP281 TF
434 activity in off-target fated cells, we performed Tomtom motif similarity analysis[58] to identify TFs
435 that share a motif similar to ZFP281. We found four other TF motifs that were both significantly
436 similar to the ZFP281 motif (adjusted p-value < 0.05) and were enriched in off-target destined
437 cells. Amongst these TFs, ZFP281 displayed the highest enrichment in the off-target lineage both
438 in terms of gene expression and TF activity (**Ext Fig. 9e**). Additionally, single-cell accessibility of
439 ZFP281 genomic targets[56] was positively correlated with inferred ZFP281 TF activity (Pearson's
440 correlation coefficient = 0.53; **Ext Fig. 9f**) and ZFP281 regulated genes[59] were significantly more

441 predictive of cell fate as compared to a size-matched set of random genes (Mann Whitney
442 Wilcoxon test; p-value = 2.248e-09; **Ext Fig. 9g**), further confirming its role in off-target fate
443 specification during iEP reprogramming. Notably, both *Zfp281* and *Foxd2* failed to show a strong
444 lineage-specific bias in gene expression levels, highlighting the unique insights offered by
445 multiomic lineage tracing in the identification of fate-specifying TFs (**Ext Fig. 9h**).

446 Indeed, inclusion of Zfp281 along with Foxa1 and Hnf4α in the reprogramming cocktail
447 resulted in a moderate but statistically significant reduction in the number of CDH1-positive
448 colonies (t-test; p-value = 0.017; **Fig. 5e**). To further characterize the role of Zfp281 in
449 reprogramming, we performed both overexpression (OE) and shRNA mediated knockdown (KD)
450 of Zfp281, along with respective control samples, and profiled cells with single-cell sequencing on
451 reprogramming day 14 (**Fig. 5f, g, Ext Fig. 10a**). We found that the rate of reprogramming (both
452 on- and off-target) increased with increasing *Zfp281* expression (**Ext Fig 10b**), suggesting a role
453 for Zfp281 in accelerating fate conversion in iEP reprogramming. Moreover, we identified a distinct
454 subpopulation of cells, predominantly consisting of Zfp281 KD cells that were depleted for
455 expression of key markers of both on-target and off-target reprogramming such as *Apoa1* and
456 *Ctla2a* (**Ext Fig 10c-e**). These cells were enriched for MEF and early off-target marker gene
457 expression, depleted for both off-target and on-target markers genes from Day 21 (obtained from
458 our lineage analysis; **Ext Fig. 10f-h**) and thus likely represent a "stalled" cell state due to reduced
459 *Zfp281* expression levels. Despite its acceleration of cell fate conversion broadly, we found that
460 Zfp281 shifted the identity of reprogrammed cells away from an iEP-like state and towards a dead-
461 end/off-target-like state consistently across the OE and KD experiments (**Fig. 5h, i**), confirming a
462 role for Zfp281 in biasing cells towards an off-target fate, as suggested by our lineage tracing
463 analysis. This finding also explains the reduced number of CDH1-positive colonies observed in
464 our colony formation assay, despite the increase in total number of on-target reprogrammed cells
465 upon Zfp281 overexpression.

466 Finally, a key downstream effector of Zfp281 is TGF-β signaling[51,60], an Epithelial-to-
467 Mesenchymal Transition (EMT) associated pathway[61]. Indeed, TGF-β pathway activity, as
468 inferred using PROGENy[62] (**Methods**), increased with *Zfp281* OE and decreased with *Zfp281*
469 KD, suggesting active regulation of TGF-β signaling by Zfp281. Given that on-target
470 reprogramming is characterized by cellular epithelialization and off-target reprogramming is
471 characterized by activation of broad mesenchymal programs, we hypothesized that increased
472 TGF-β signaling mediated via Zfp281 acts as a barrier to on-target reprogramming. Indeed,
473 inhibition of TGF-β signaling during iEP reprogramming using the small molecule SB431542[63] led
474 to a significant increase in expression of reprogramming marker genes *Apoa1* and *Gsta4* and a

475 significant decrease in mesenchymal/off-target genes such as *Serpine1, Snail1, Col1a2* (**Fig. 5k**).
476 This was accompanied by an increase in epithelial/iEP-like morphology as early as day 3 of
477 reprogramming (**Ext Fig 10i**) suggesting a crucial role for TGF-β signaling, downstream of Zfp281,
478 in determining fate outcome during iEP reprogramming.
479
480 **Discussion**
481 Here we have presented CellTag-multi, a method for independent single-cell lineage tracing
482 across scRNA-seq and scATAC-seq assays. In the context of hematopoiesis, we have used
483 CellTag-multi to map transcriptional and epigenomic states of progenitor cells and link them to
484 clonal fate, recapitulating enrichment of known lineage-specific cell state signatures across
485 progenitor populations. With chromatin state, we showed that lineage-specific epigenetic priming
486 is associated with changes in accessibility of known fate-specifying TF motifs and that such
487 changes occur primarily in the regions of the genome distal to promoters. Previous analysis has
488 demonstrated the inability of early transcriptional state alone in predicting cell fate and posited a
489 role for alternate cell state modalities[13]. By exploiting multiomic clonal relationships, we
490 demonstrated that the predictability of cell fate from state is significantly improved when both early
491 transcriptional and epigenomic state are considered, as opposed to either modality individually,
492 suggesting that the RNA and ATAC modalities consist of non-redundant and highly
493 complementary state information.
494 Our application of CellTag-multi to the less characterized paradigm of iEP reprogramming
495 generated similar observations, where multiomic clonal data captured in the early stages of fate
496 conversion is highly predictive of reprogramming outcome. Again, fate-specifying epigenetic
497 changes during early stages of differentiation are dominated by changes in distal regulatory
498 regions of the epigenome. Further, we have been able to molecularly characterize the 'dead-end'
499 state as a highly proliferative, mesenchymal-like cell state, representing an 'off-target'
500 reprogrammed state. Indeed, a similar state has been reported in direct reprogramming of
501 mesenchymal stromal cells to induced hepatocytes, revealing the appearance of *Acta2*-
502 expressing mesenchymal cells during the reprogramming process[64]. Outside of the hepatic
503 lineage, off-target identities have been reported in other reprogramming paradigms[35,65],
504 suggesting that this may be a more general feature of lineage reprogramming.
505 Our multiomic lineage tracing demonstrates the establishment of on- and off-target
506 trajectories from early stages, supported by our earlier transcriptome-based lineage tracing of iEP
507 reprogramming[7] and GRN inference[38]. However, given the single modality capture of relatively
508 few clones in that earlier study, we were not able to comprehensively characterize early molecular

509 states. Here, the collection of ground truth data on lineage, transcriptome, and epigenome has
510 allowed us to better characterize these distinctive early states, enabling novel mechanistic
511 insights into reprogramming. We have shown crucial early differences in gene regulation that lead
512 to distinct reprogramming outcomes. Specifically, we have identified and experimentally validated
513 that Foxd2 promotes successful reprogramming, while Zfp281 activity leads to engagement with
514 an off-target trajectory. Differences in reprogramming TF levels may account for these early
515 differences. However, lower levels of exogenous TF expression do not simply lead to
516 reprogramming failure, as the off-target fate is molecularly unique from fibroblasts and could be
517 considered a reprogramming byproduct in itself. These results suggest that the stoichiometry of
518 TF overexpression in these reprogramming models may offer further insight into how TFs control
519 cell identity. Single-cell analysis of TF binding could provide further insights into the role of
520 differential binding of the two reprogramming TFs in specifying off-target fate.

521 Our recovery of Foxd2 and Zfp281 as novel regulators of early-stage reprogramming was
522 not possible from differential gene expression analysis alone, demonstrating the utility of CellTag-
523 multi. Our data suggests off-target enriched Zfp281 activity from early stages of reprogramming.
524 From our experimental validation, we found that knockdown of Zfp281 expands a population of
525 cells in a 'stalled' state, where they fail to extinguish fibroblast gene expression while upregulating
526 off-target cells. Conversely, overexpression of Zfp281 helps accelerate fate conversion, resulting
527 in a considerable increase in reprogramming efficiency. However, Zfp281 still draws the
528 reprogrammed cells toward an off-target, mesenchymal-like state. A role for this TF in driving
529 broad mesenchymal expression programs, including components of the TGF-β signaling
530 pathway, has recently been described[51]. Here, we demonstrate that the inhibition of TGF-β
531 signaling enhances on-target marker expression while decreasing off-target gene expression.
532 These results suggest a potential strategy to enhance on-target reprogramming, where Zfp281
533 expression can help erase the starting cell identity while blocking downstream TGF-β signaling
534 might prohibit entry onto the off-target trajectory.

535 Altogether, the data we present here across two distinct biological systems demonstrates
536 that lineage-specific capture of gene expression and chromatin accessibility provides rich
537 information on gene regulation, offering unique mechanistic insights into the specification and
538 maintenance of cell identity. More widely, single-cell lineage tracing has revealed distinct, clonally
539 heritable transcriptional states across various biological systems[66–68]. These phenotypic
540 differences, arising from seemingly non-genetic sources, have strong biological implications. For
541 example, clonal variability in cell state has been shown to impact malignant clonal expansion and
542 efficacy of drug treatment in cancer cells[66,68]. Elsewhere, CRISPR-based systems have been

543 used to create mutable barcodes to allow multi-level lineage recording without the need for
544 successive rounds of cell labelling[69,70]. Given its versatility and ease of use, we envision that
545 CellTag-multi can be readily applied to such biological questions and use cases.
546 Finally, we have developed CellTag-multi to work independently with scRNA-seq and
547 scATAC-seq, as existing single-cell methods that co-assay multiple modalities from the same
548 cell[71–74] can suffer from lower data quality compared to methods that profile each modality
549 individually. Further, enabling the capture of lineage in parallel with chromatin accessibility
550 provides users with additional flexibility for experimental design. Advances in single-cell
551 technologies are allowing measurement of an ever-increasing number of cellular modalities. A
552 similar expansion in lineage tracing assays will complement these new methods with multiomic,
553 clonal tracking of cell state and enable deeper mechanistic insight into the regulation of cell
554 identity and clonal heritability of cell state. CellTag-multi, with its cell lineage read out alongside
555 gene expression and chromatin accessibility, paves the way for such multiomic, single-cell lineage
556 tracing methods.
557
558 **Acknowledgments**
559 We thank members of the Morris laboratory for helpful discussions. This work was funded by the
560 National Institute of General Medical Sciences R01 GM126112, and Silicon Valley Community
561 Foundation, Chan Zuckerberg Initiative Grant DAF2021-238797, both to S.A.M. S.A.M. is
562 supported by an Allen Distinguished Investigator Award (through the Paul G. Allen Frontiers
563 Group), a Vallee Scholar Award, a Sloan Research Fellowship, and a New York Stem Cell
564 Foundation Robertson Investigator Award; K.K. is supported by a Japan Society for the Promotion
565 of Science Postdoctoral Fellowship; N.Y. is supported by McDonnell International Scholars
566 Academy fellowship.
567
568 **Author Contributions**
569 Conceptualization, Methodology, K.J., S.A.M.; Software, K.J.; Formal Analysis, K.J., N.Y.;
570 Experimentation, K.J., M.T.A., H.W., X.Y., S.A.M., K.K., G.R.G.; Data Curation, K.J., M.T.A.;
571 Writing – Original Draft, K.J., S.A.M.; Writing – Review & Editing, K.J., M.T.A., H.W., N.Y., K.K.,
572 G.R.G., S.A.M.; Visualization, K.J., S.A.M.; Funding Acquisition, Resources, Supervision, S.A.M.
573
574 **Declaration of Interests**
575 S.A.M. and G.R.G. are cofounders of CapyBio LLC.

Figure 1

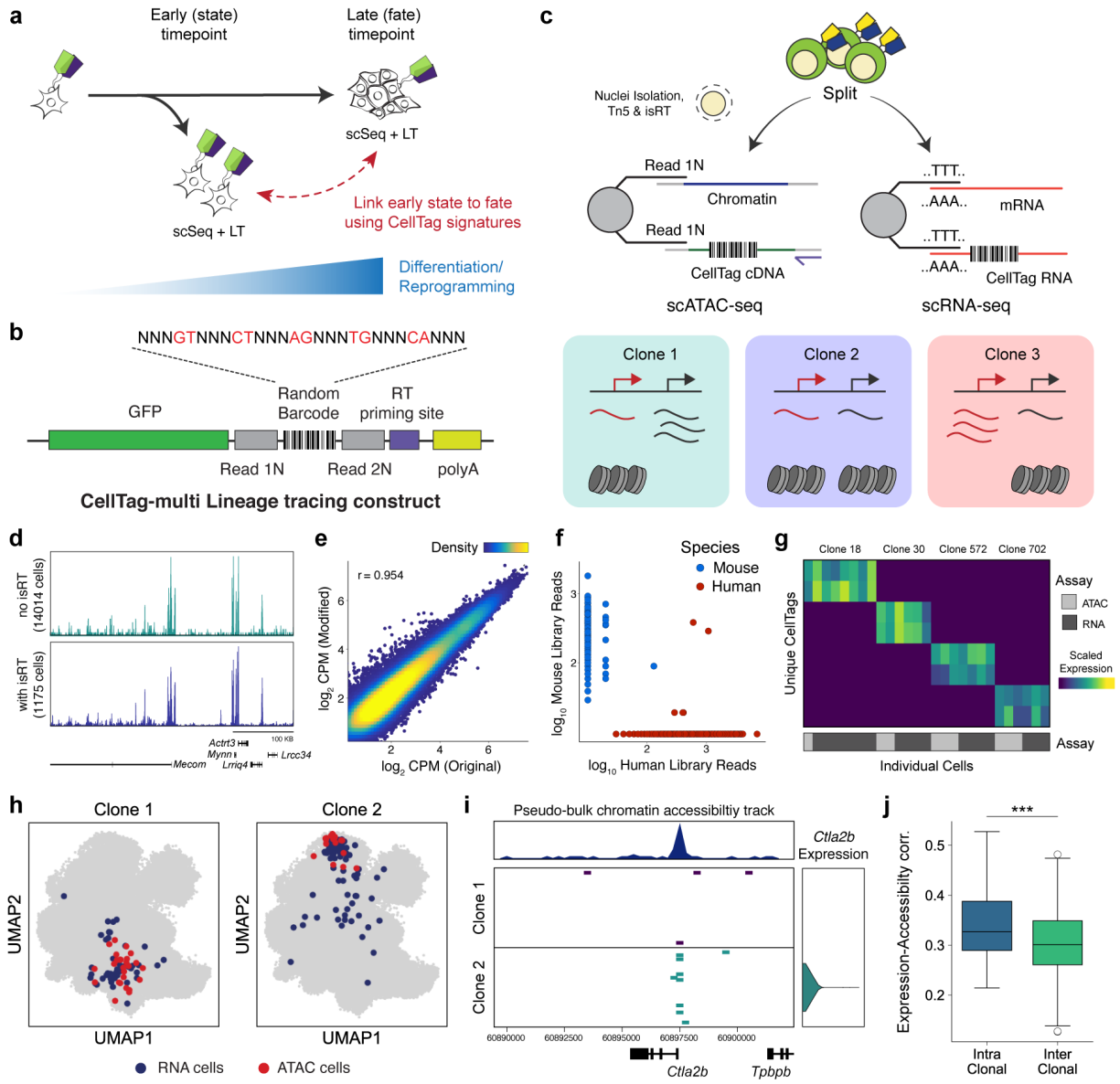
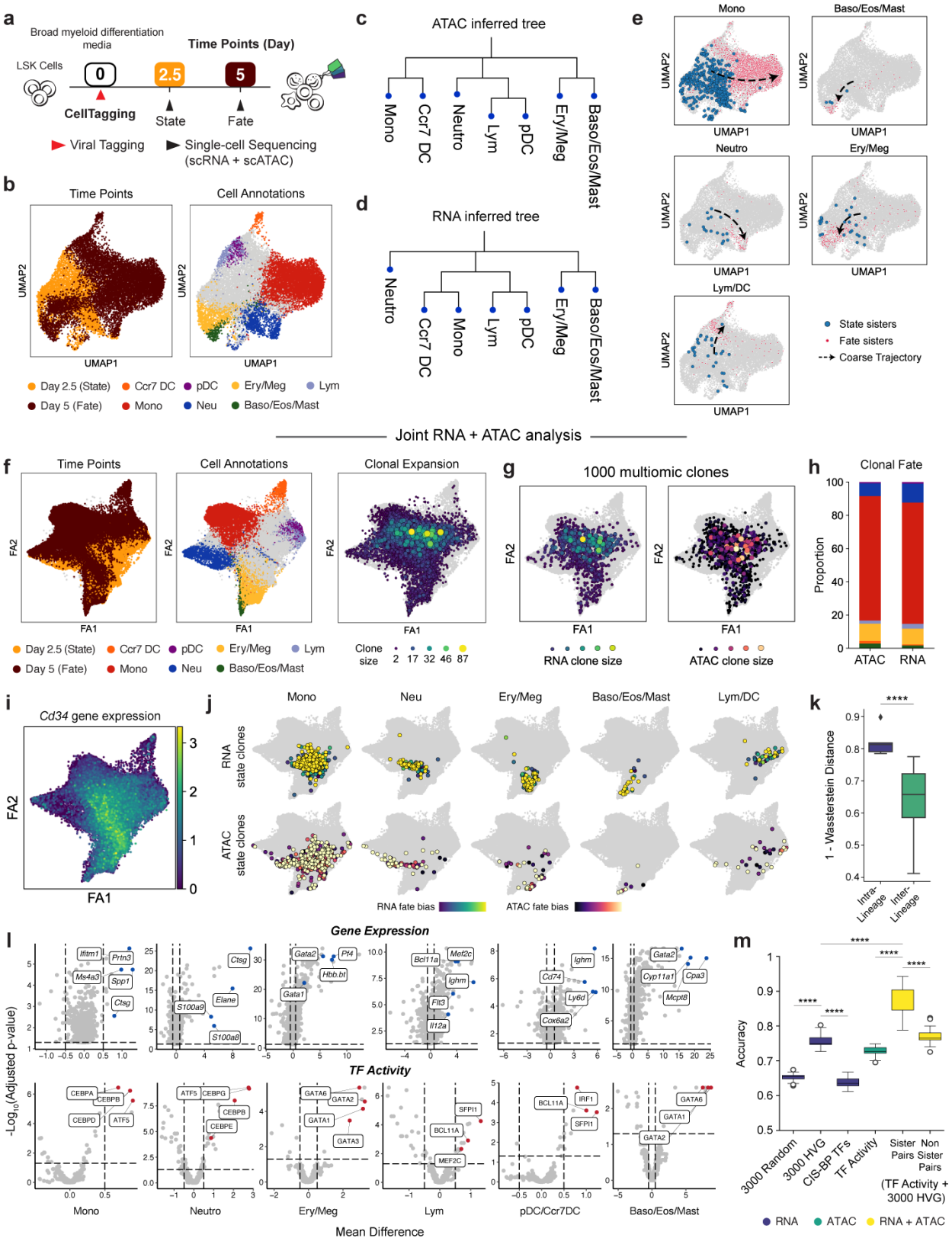


Figure 1. CellTag-multi allows simultaneous capture of lineage information with gene expression and chromatin accessibility. (**a**) A framework for relating early cell state with fate using single-cell lineage tracing. (**b**) Schematic depicting the CellTag-multi lineage tracing construct. (**c**) Schematic detailing parallel capture of CellTags during scRNA-seq and modified scATAC-seq library preparation, using targeted isRT of CellTags in intact nuclei. CellTag-multi enables simultaneous clonal tracking of transcriptional and epigenomic state. (**d**) Browser tracks comparing chromatin accessibility signal across aggregated scATAC-seq profiles generated using the original and modified library preparation methods. (**e**) Scatterplot comparing log

585 normalized reads in ATAC peaks across aggregated scATAC-seq profiles generated with the
586 original and modified library preparation methods. (r = Pearson Correlation Coefficient). (**f**) Plot
587 for the human-mouse species mixing experiment depicting the number of CellTag reads/cell from
588 each CellTag library. (**g**) Heatmap showing scaled CellTag expression in scRNA-seq and
589 scATAC-seq sisters for four multiomic clones identified in a population of expanded
590 reprogramming fibroblasts. (**h**) Joint UMAP of RNA and ATAC cells with two clones (clone 1 and
591 clones 2) cells projected, along with assay information. (**i**) Browser track showing single-cell
592 accessibility at the *Ctla2b* locus and *Ctla2b* gene expression across clones 1 and 2. Top Panel:
593 Pseudo-bulk accessibility signal at the *Ctla2b* locus. (**j**) Box plots comparing intra- and inter-clonal
594 correlation between clonally aggregated gene expression and gene activity scores in the
595 reprogramming dataset (n = 62 clones used; Mann Whitney Wilcoxon test; p-value = 5.39e-4).
596

Figure 2


a
Broad myeloid differentiation media
Time Points (Day)
LSK Cells
0
2.5
5
CellTagging
State
Fate
Viral Tagging
Single-cell Sequencing (scRNA + scATAC)
b
Time Points
Cell Annotations
UMAP2
UMAP1
Day 2.5 (State)
Ccr7 DC
pDC
Ery/Meg
Lym
Day 5 (Fate)
Mono
Neu
Baso/Eos/Mast
c
ATAC inferred tree
Mono
Ccr7 DC
Neutro
Lym
pDC
Ery/Meg
Baso/Eos/Mast
d
RNA inferred tree
Neutro
Ccr7 DC
Mono
Lym
pDC
Ery/Meg
Baso/Eos/Mast
e
Mono
Baso/Eos/Mast
Neutro
Ery/Meg
Lym/DC
State sisters
Fate sisters
Coarse Trajectory
Joint RNA + ATAC analysis
f
Time Points
Cell Annotations
Clonal Expansion
FA2
FA1
Clone size
2
17
32
46
87
g
1000 multiomic clones
RNA clone size
ATAC clone size
h
Clonal Fate
Proportion
ATAC
RNA
i
Cd34 gene expression
j
Mono
Neu
Ery/Meg
Baso/Eos/Mast
Lym/DC
RNA state clones
ATAC state clones
RNA fate bias
ATAC fate bias
k
1 - Wasserstein Distance
Intra-Lineage
Inter-Lineage
l
Gene Expression
TF Activity
-Log10(Adjusted p-value)
Mono
Neutro
Ery/Meg
Lym
pDC/Ccr7DC
Baso/Eos/Mast
Mean Difference
m
Accuracy
3000 Random
3000 HVG
CIS-BP TFs
TF Activity
Sister Pairs
Non Sister Pairs
(TF Activity + 3000 HVG)
RNA
ATAC
RNA + ATAC

597

598 **Figure 2. Application of CellTag-multi to link early hematopoietic cell state with fate.** (**a**)
599 Schematic detailing the experimental design for the *in vitro* hematopoiesis state-fate experiment.
600 (**b**) scATAC-seq UMAPs with time point (left panel) and cell fate information (right panel) projected
601 (Mono: Monocytes; Neu: Neutrophils; Lym: Lymphoids; Ery: Erythroids; Meg: Megakaryocytes;
602 Baso: Basophils; Eos: Eosinophils; Mast: Mast Cells; pDC: plastoid Dendritic Cells). Only major
603 cell fates are highlighted. Hematopoietic lineage hierarchy as inferred from (**c**) scATAC or (**d**)
604 scRNA clone coupling. (**e**) scATAC-seq UMAPs with state and fate sisters for major hematopoietic
605 fates highlighted. (**f**) Clone-cell embedding UMAPs with time point and cell fate information
606 projected onto cells (left and center panels) and clonal expansion information projected onto
607 clones (right panel), detailed cell type annotations are shown in **Ext Fig. 4c**. (**g**) UMAPs with RNA
608 and ATAC clonal expansion information projected onto a thousand random multiomic clones. Both
609 modalities display biased expansion of early myeloid cells, consistent with our differentiation
610 culture conditions. (**h**) Bar plot depicting distribution of cell fates across RNA and ATAC clones
611 (Fates are colored as in **Fig. 2b**). (**i**) UMAP with scaled *Cd34* expression level, a marker of
612 Hematopoietic Stem and Progenitor Cells (HSPCs), projected onto the scRNA cells. (**j**) UMAPs
613 with state (Day 2.5) sub-clones for each major lineage highlighted along the differentiation
614 continuum for both single-cell modalities, with fate bias information projected. (**k**) Box plot
615 comparing overlap between RNA and ATAC state sub-clones within and across cell fates (Mann
616 Whitney Wilcoxon test; p-value = 3.76e-5). (**l**) Volcano plots summarizing the results of differential
617 feature enrichment analysis for each group of state sub-clones across for scRNA (top panel) and
618 scATAC modalities (bottom panel). (**m**) Box plot summarizing accuracy scores of trained state-
619 fate prediction models. Machine learning partially predicts cell fate from Day 2.5 state across both
620 modalities. However, predictive performance increases significantly when both are considered
621 together, highlighting the existence of unique functional priming in both gene expression and
622 chromatin accessibility state (Mann Whitney Wilcoxon test; p-values: **** = $p < 0.0001$, HVG:
623 Highly Variable Genes, n = 25 accuracy values for each model (**Methods**)).

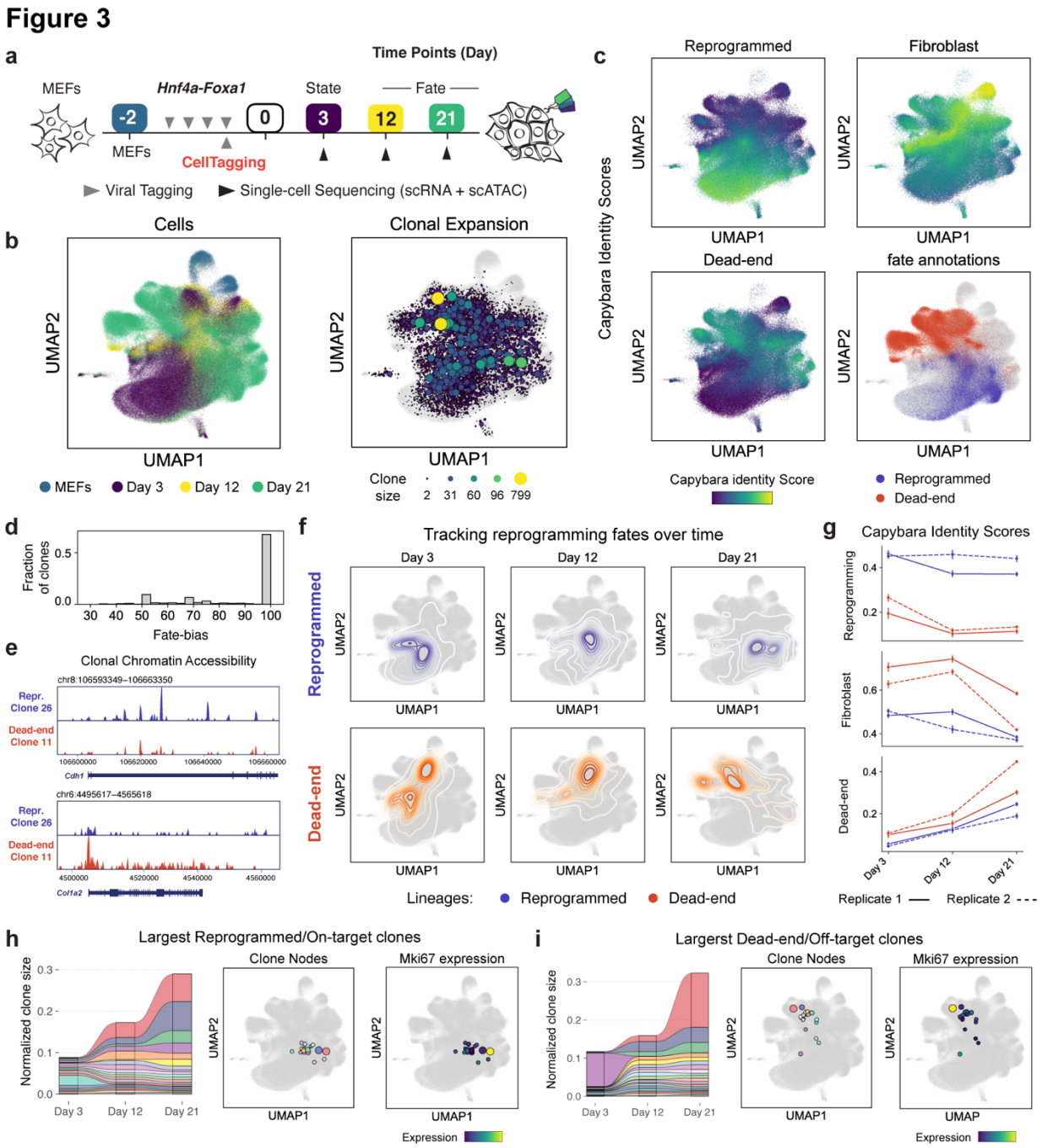


Figure 3. Application of CellTag-multi to dissect clonal fate dynamics in direct reprogramming. (**a**) Experimental design for the direct reprogramming state-fate experiment. (**b**) Cells from both scRNA-seq and scATAC-seq, across all time points, were co-embedded with clones and visualized using a UMAP. (Left Panel) Time point information projected on cells. (Right Panel) Clonal expansion visualized using clone nodes. (**c**) Capybara transcriptional identity scores projected on scRNA-seq cells for reprogrammed, dead-end and fibroblast cell identities,

631 based on a previous lineage tracing dataset[7]. Cell fates were annotated for Days 12 and 21.
632 Reprogrammed and dead-end cell fates are highlighted (Lower Right Panel). (**d**) Histogram of
633 fate bias across all state-fate clones. (**e**) Clonal chromatin accessibility browser tracks for two
634 dead-end and reprogramming clones. (**f**) Contour plots showing longitudinal tracking of cell fates
635 enabled by CellTagging. (**g**) Transcriptional identity dynamics tracked along both lineages. Dead-
636 end cells depart from a MEF like identity and acquire an off-target reprogrammed state. Significant
637 clonal expansion is observed along both lineages, as depicted via alluvial plots, clone nodes and
638 clonal expression levels of *Mki67* (a proliferation marker gene) in the 20 largest (**h**)
639 reprogramming/on-target and (**i**) dead-end/off-target clones.

Figure 4

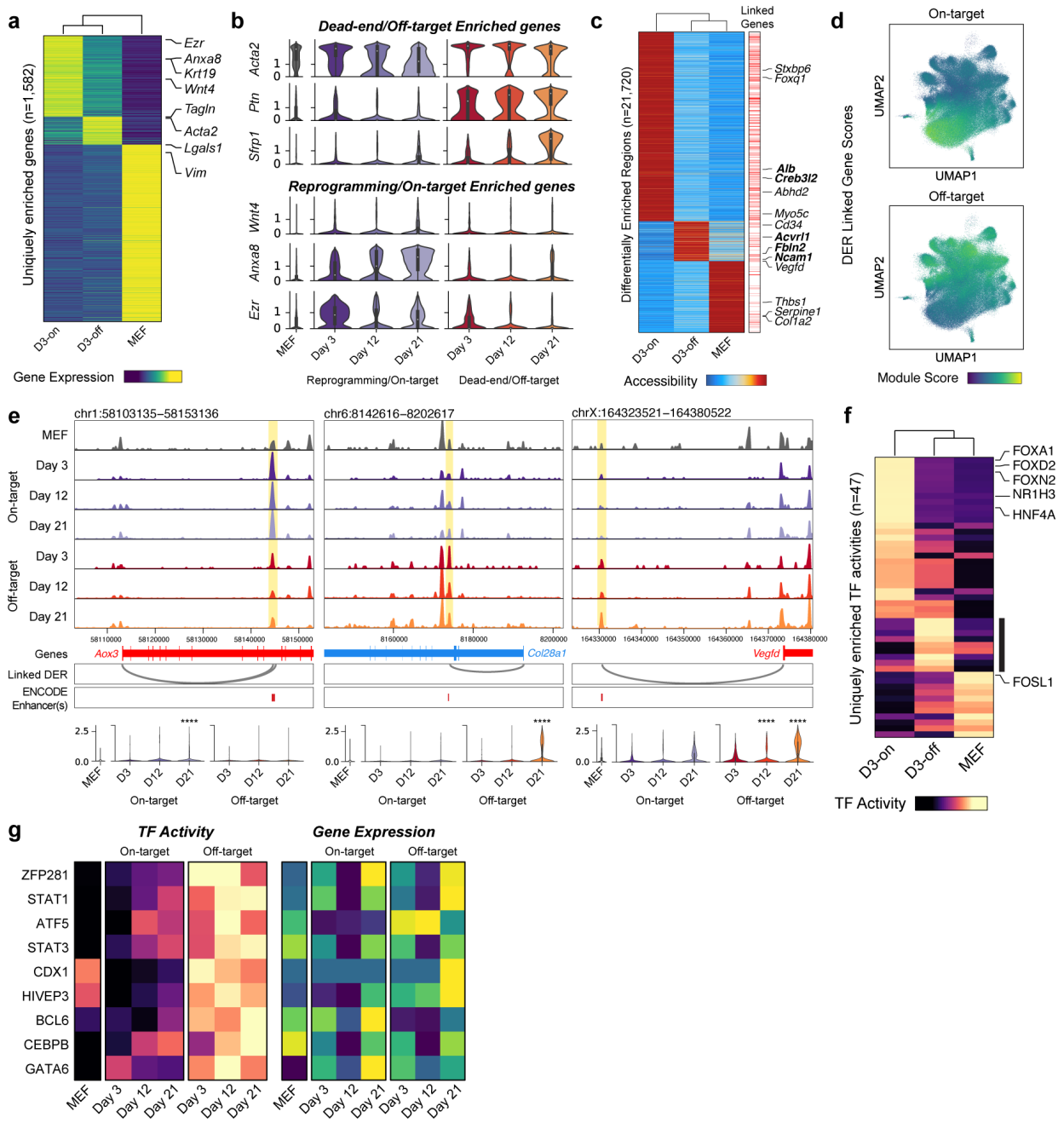


Figure 4. Assessing fate-specific changes in early cell state. (**a**) Heatmap of genes uniquely enriched across uninduced MEFs or one of the two reprogramming fates on Day 3 (FDR threshold: 0.05, log fold change threshold: 0). (**b**) Violin plots of several genes enriched in either off-target (dead-end) destined or on-target (reprogramming) destined cells. (**c**) Heatmap of peaks uniquely enriched across uninduced MEFs or one of the two reprogramming fates on Day 3 (FDR

threshold: 0.05, log fold change threshold: 1). Right panel shows annotation of peaks linked to genes (Methods). (**d**) Module scores for genes linked to either on-target or off-target DERs projected onto the clone-cell embedding. (**e**) Top panel: Accessibility browser tracks for each lineage split by day, highlighting peaks linked to late lineage markers (On-target: *Aox3*; Off-target: *Col28a1* and *Vegfd*) showing lineage specific changes in accessibility on Day 3. The *Aox3* and *Vegfd* linked DERs overlap perfectly with an ENCODE enhancer like element (ELS) while the *Col28a1* linked DER is within 100 bp of an ELS. Bottom panel: Expression levels of the three genes across MEFs and the two reprogramming lineages split by days. The asterisks (*) mark time points and lineage of significant differential enrichment. (**f**) Heatmap of TF activities uniquely enriched across uninduced MEFs or one of the two reprogramming fates on Day 3 (FDR threshold: 0.05, mean difference threshold: 0.5). (**g**) Left Panel: Heatmap showing TF activity (left panel) and gene expression (right panel) levels for off-target associated TFs in MEFs and each reprogramming lineage split by time points. TF activity signatures show a much stronger lineage bias as compared to gene expression values.

Figure 5

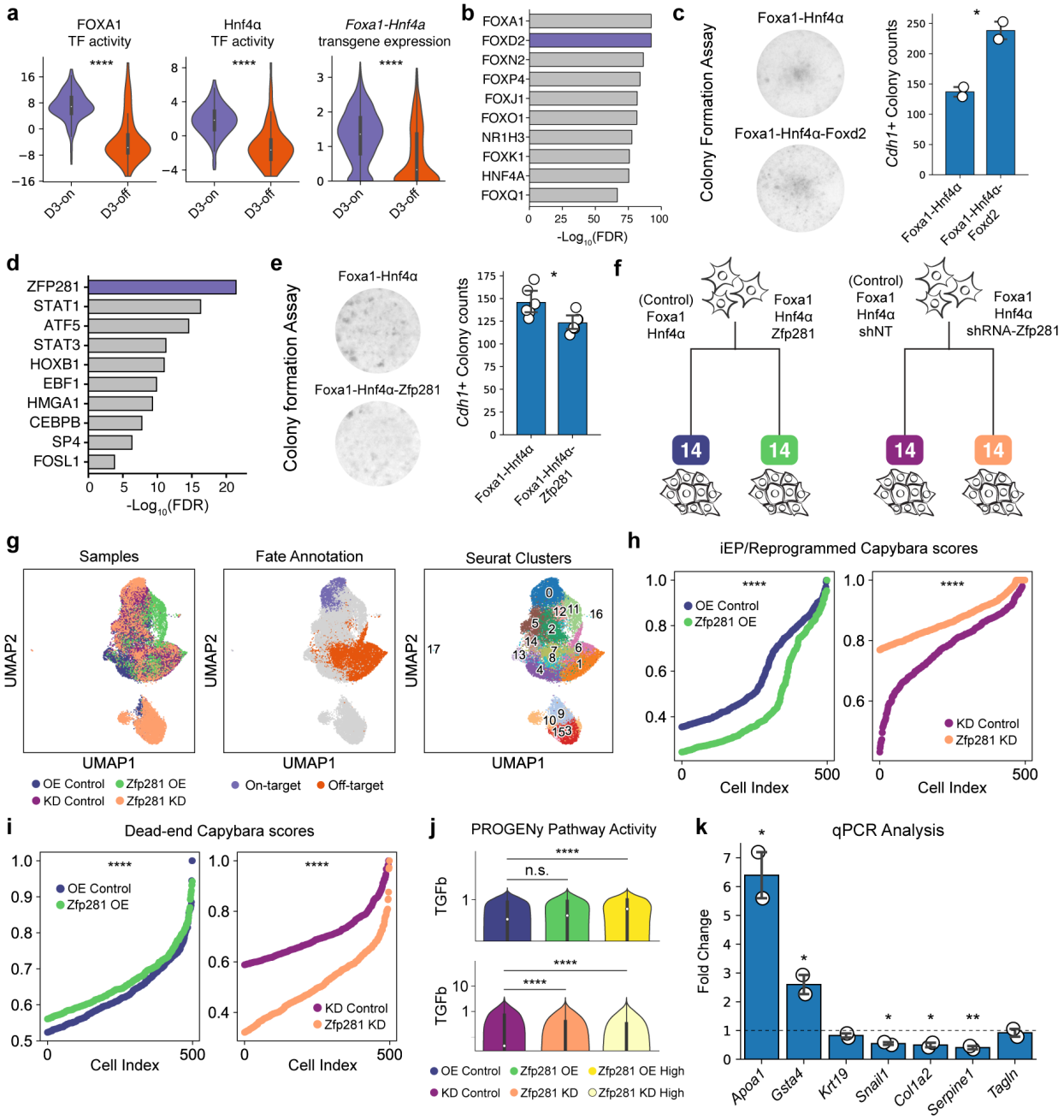


Figure 5. Identification of TF regulators of on-target and off-target reprogramming fate. (**a**) Left and Middle panels: Violin plots comparing enrichment of FOXA1 and HNF4A TF activities across the two reprogramming fates on Day 3 (Mann Whitney Wilcoxon test; p-values: FOXA1 = 9.2e-22, HNF4A = 1.7e-20). Right panel: Violin plot comparing enrichment of the *Hnf4α-Foxa1* transgene expression across the two reprogramming fates on Day 3 (Mann Whitney Wilcoxon

test; p-value = 6.5e-42). (**b**) Top ten TFs enriched in on-target destined cells based on TF activity scores. (**c**) Left Panel: Representative images from the *Foxd2* overexpression colony formation assay; Right Panel: Bar plot showing increase in CDH1-positive colony counts in *Foxd2* overexpressing cells compared to a standard reprogramming experiment (t-test; p-value = 0.045; n = 2 biological replicates). (**d**) Top ten TFs enriched in off-target destined cells based on TF activity scores. (**e**) Left Panel: Representative images from the *Zfp281* overexpression colony formation assay; Right Panel: Bar plot showing decrease in CDH1-positive colony counts in the *Zfp281* overexpressing sample compared to a standard reprogramming experiment. (t-test; p-value = 0.017; n = 6 biological replicates). (**f**) Schematic of the scRNA-seq experiment for *Zfp281* over-expression (OE) and knockdown (KD) during reprogramming. A GFP expression vector and non-target shRNA were used as controls for OE and KD respectively. (**g**) UMAP embedding for all cells profiled in the *Zfp281* OE and KD experiments with sample information (Left), cell fate annotation (Middle) and Seurat cluster information (Right) projected. (**h**) Plot of iEP Capybara identity scores across the KD and OE samples compared to respective controls (Mann Whitney Wilcoxon test; p-values: *Zfp281* OE vs control = 1.07e-53; *Zfp281* KD vs control = 2.19e-53). (**i**) Plot of dead-end Capybara identity scores across the KD and OE samples compared to respective controls (Mann Whitney Wilcoxon test; p-values: *Zfp281* OE vs control = 1.11e-11; *Zfp281* KD vs control = 3.26e-120). (**j**) Violin plots showing variation of TGF-β pathway activity across control vs OE vs OE high cells (upper panel) and control vs KD vs KD high cells (lower panel). OE high cells are defined as the subset OE sample cells with above average *Zfp281* expression. KD high cells are defined as the subset of KD sample cells with below average *Zfp281* expression. (**k**) Bar plots showing fold-change in reprogramming and dead-end marker genes upon small molecule mediated inhibition of TGF-β signaling, compared to a vehicle control, on Day 5 of iEP reprogramming (t-test; p-values: *Apoa1* = 0.02, *Col1a2* = 0.02, *Gsta4* = 0.04, *Serpine1* = 0.009, *Snail1* = 0.01; n=2 technical replicates).

Methods

Tissue culture

Isolation of mouse LSK cells. Lin$^-$ Sca1$^+$ c-Kit$^+$ (LSK) cells were obtained using a previously described protocol[13]. Adult mice were euthanized, bone marrow was extracted from long bone, hips and spine and passed through a 70µm filter. Cells were centrifuged at 300g for 10mins at 4C and the pellet was resuspended in EasySep buffer (STEMCELL, Cat. 20144) at 100 million cells/ml. EasySep lineage depletion kit (STEMCELL, Cat. 19856) was used to remove differentiated cells. Finally, cells were stained for Sca1 (Sca1-AF488; BioLegend clone D7) and cKit (CD117-PE; BioLegend clone 2B8) and sorted using the MoFlo Cell Sorter (Beckman Coulter) with a 130µm nozzle. Isolated LSK cells were counted and used directly for lineage tracing experiments.

Mice and derivation of mouse embryonic fibroblasts. MEFs were derived from embryonic day (E)13.5 C57BL/6J embryos. (The Jackson laboratory: 000664). Heads and visceral organs were removed and the remaining tissue was minced with a razor blade and then dissociated in a mixture of 0.05% trypsin and 0.25% collagenase IV (Life Technologies) at 37 °C for 15 min. After passing the cell slurry through a 70-µM filter to remove debris, cells were washed and then plated on 0.1% gelatin-coated plates, in DMEM supplemented with 10% FBS (Gibco), 2mM l-glutamine and 50mM β-mercaptoethanol (Life Technologies). All animal procedures were based on animal care guidelines approved by the Institutional Animal Care and Use Committee.

General Experimental methods

Lenti- and retro-virus production. Lentiviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pSMAL-CellTag construct (see below), along with packaging constructs pCMV-dR8.2 dvpr (Addgene plasmid 8455), and pCMV-VSVG (Addgene plasmid 8454). Constructs were titered by serial dilution on 293T cells. Hnf4α -T2A-Foxa1 was cloned into the pGCDN-Sam retroviral construct and packaged with pCL-Eco (Novus Biologicals, NBP2-29540), titered on fibroblasts. We opted to generate a bicistronic *Hnf4α-Foxa1* construct, based on the T2A sequence to increase the consistency of reprogramming via maintenance of exogenous transcription factor stoichiometry. Virus was collected 48 h and 72 h after transfection and applied to cells immediately following filtering through a low-protein binding 0.45-µm filter. Wherever applicable, the virus was concentrated using high-speed centrifugation. 20ml of filtered viral supernatant was centrifuged at 50,000g for 2.5 hours at 4ºC, supernatant was removed and the virus was resuspended in 100ul of DMEM. The concentrated virus was stored at -80C.

736

737 **scRNA-seq library preparation.** 3' single-cell RNA library preparation was performed using the
738 Chromium Single Cell Gene Expression Kit from 10x Genomics. Cells were obtained as single-
739 cell suspensions and processed according to the manufacturer's instructions (CG000315).
740

741 **CellTag amplification for scRNA-seq (CellTag-RNA PCR).** An additional PCR step was used
742 to amplify CellTag barcodes from the single-cell cDNA library, obtained after step 2.4 of the 10x
743 Genomics Single Cell Gene Expression Kit user guide (CG000315). 5ul (or at least 60ng) of cDNA
744 was mixed with 2x Q5 HF PCR Master Mix (New England Biolabs) and 500nM of *P5/R1-par* and
745 *P7/SI-R2* primers in a 50ul reaction volume and subjected to the following PCR program: 98 C for
746 30 seconds; N cycles (98ºC for 10 seconds; 54ºC for 30 seconds; 72ºC for 30 seconds); 72ºC for
747 2 minutes. The number of PCR cycles (N) was kept the same as the number of cycles used during
748 sample index PCR of the main scRNA-seq library. CellTag amplicon library was purified using
749 double-sided bead purification (0.4x-0.64x) and quantified on an Agilent TapeStation using the
750 D1000-HS tape. Libraries were either sequenced by themselves (with a 50% Phi-X spike-in) or
751 along with scRNA-seq libraries (preferred). CellTag amplicon libraries were sequenced on an
752 Illumina NextSeq-500 to avoid index hopping-related artifacts. Primer sequences are available in
753 Supplementary Table 1.
754

755 **scATAC-seq library preparation.** Standard scATAC-seq library preparation was performed
756 using the Chromium Single Cell ATAC Kit from 10x Genomics. Cells were obtained as single-cell
757 suspensions, nuclei were isolated using 10x Genomics nuclei isolation protocol (CG000169), and
758 libraries were prepared according to the manufacturer's instructions (CG000209).
759

760 **scATAC-seq library preparation with modifications for CellTag capture.** To prepare single-
761 cell ATAC libraries with CellTag capture, nuclei were isolated using manufacturer's instructions
762 (CG000169), centrifuged to remove supernatant, and lightly fixed in 100ul 0.1% formaldehyde
763 solution for 5 minutes. The reaction was stopped for 5 minutes by adding 30ul of stop buffer
764 (0.625M Glycine, 0.5% BSA, 0.25M ph8 Tris-Cl in PBS). The nuclei suspension was diluted using
765 100ul diluted nuclei buffer (10x Genomics; CG000169) and pelleted using centrifugation. The
766 pellet was subjected to tagmentation for 60 minutes after re-suspension in a 15ul tagmentation
767 reaction (for up to 15k nuclei) according to the manufacturer's instructions (CG000209). After
768 tagmentation, the reaction mixture was diluted with 100ul dilute nuclei buffer, nuclei were pelleted
769 using centrifugation and subjected to targeted in situ reverse transcription in a 100ul reaction

770 volume (20ul of 5x SuperScript IV reaction buffer, 5ul each of dNTPs, DTT, RnaseOUT RNase
771 inhibitor, SuperScript IV Reverse Transcriptase, 1uM of primer *ctac2-rt1*) using the following
772 temperature program: 4ºC for 2 minutes; 10ºC for 2 minutes; 20ºC for 2 minutes; 30ºC for 2
773 minutes; 40ºC for 2 minutes; 45ºC for 10 minutes. After isRT, the reaction mixture was diluted
774 with 100ul dilute nuclei buffer and pelleted using centrifugation. 15ul GEM-nuclei mix was
775 prepared to load nuclei on 10x Genomics Chip E/H by mixing up to 15k nuclei with 6ul of ATAC
776 buffer (from the 10x Genomics scATAC-seq kit) and 3ul of 4uM primer *ctac2-rt1*. Any remaining
777 volume was made up with dilute nuclei buffer. GEM-nuclei mix was loaded onto Chip E/H along
778 with ATAC GEM beads and barcoding enzyme mix, the remaining steps of the scATAC-seq library
779 preparation protocol were performed according to the manufacturer's instructions. Primer
780 sequences are available in Supplementary Table 1. All centrifugation steps were performed at
781 500g for 10 minutes at 4ºC unless stated otherwise.
782
783 **CellTag amplification for scATAC-seq (CellTag-ATAC PCR).** While CellTags can be recovered
784 directly from the sequenced scATAC-seq library with our library preparation, a higher yield can
785 be obtained using an additional targeted PCR step, similar to the scRNA-seq version. For this,
786 5ul of the library is collected after step 3.2 of the user guide (CG000209) and mixed with 2x Q5
787 HF master mix, 500nM of primer *biot-atac2_lin* and water in a 50ul reaction volume, and CellTag
788 containing fragments are linearly amplified using the following PCR program: 98ºC for 30
789 seconds; 20 cycles (98ºC for 10 seconds; 67ºC for 30 seconds; 72ºC for 30 seconds); 72ºC for 2
790 minutes. The CellTag amplicons are purified using streptavidin-coated magnetic bead pulldown
791 (ThermoFisher Scientific; Dynabeads™ MyOne™ Streptavidin C1) and purified fragments are
792 resuspended in 20ul of water. A final sample index PCR is performed to create a sequencible
793 library in presence of 2x Q5 master mix, 500nM each of *partial_p5* and *biot-atac2_e-rev* primers
794 in a 100ul reaction volume using the following PCR program: 98ºC for 30 seconds; 13 cycles
795 (98ºC for 10 seconds; 67ºC for 30 seconds; 72ºC for 30 seconds); 72ºC for 2 minutes and libraries
796 are purified using a double-sided bead cleanup, as described in Step 4.2 of 10x Genomic
797 scATAC-seq user guide (CG000209). Primer sequences are available in Supplementary Table 1.
798
799 **General Computational methods**
800 **Identifying clones.** Clone identification was performed based on our previously described
801 method[7,17]. Reads matching the CellTag-multi barcode sequence pattern
802 $(N)_3GT(N)_3CT(N)_3AG(N)_3TG(N)_3CA(N)_3$ were extracted from single-cell bam files as obtained
803 from CellRanger, filtered to remove false positive transcriptomic/genomic reads and reads

804 originating from non-cell droplets. For scRNA-seq, cell barcode-CellTag-UMI triplets represented
805 by only a single read were discarded. We also provide an estimate of CellTag sequencing
806 saturation to guide users if they require deeper sequencing of their CellTag libraries. CellTags
807 were error-corrected using Starcode[75] to mitigate PCR/sequencing errors and filtered to remove
808 sequences outside of the allowlist. Cell x CellTag read count (ATAC)/ UMI count (RNA) matrices
809 were obtained, binarized and cells with too few or too many tags were removed to obtain the final
810 Cell x CellTag matrices for scRNA-seq and scATAC-seq assays. Cell-cell similarity was computed
811 using the Jaccard similarity metric and clones were identified using graph clustering. Whenever
812 applicable, scRNA-seq and scATAC-seq CellTag matrices were merged before the Jaccard
813 similarity calculation step, to identify clones across single-cell modalities. A detailed pipeline for
814 clone calling can be found at: https://github.com/morris-lab/newCloneCalling
815
816 **Clone cell embedding.** To jointly visualize cells and clones on a single embedding, we developed
817 a unique clone-cell graph embedding approach wherein we impute a cell-cell similarity graph with
818 abstract clone nodes and use it as an input for graph embedding algorithms such as UMAP. For
819 clone-cell embedding, we first obtained our single-cell data as an AnnData object and computed
820 a cell-cell connectivity matrix based on PCA (in case of scRNA-seq) or CCA (in case of joint
821 scRNA-seq scATAC-seq embedding). Next, we created a new AnnData object containing both
822 cells and clones as observations. The connectivity matrix in the .obsm['connectivities'] slot was
823 expanded to introduce clones. Then, clones were connected to their constituent cells by setting
824 the respective entries in the expanded 'connectivities' matrix to 1. Finally, we used this clone-cell
825 AnnData object with the expanded connectivity matrix as an input to graph embedding algorithms
826 such as UMAP or Force Atlas.
827
828 **Section 1**
829 **CellTag-multi library synthesis.** CellTag-multi library was synthesized using Restriction Free
830 (RF) cloning[76]. CellTag-multi barcodes were obtained as a gBlock from IDT DNA (see
831 Supplementary Table 1 for sequence) and cloned into the pSMAL-ctac2 vector. 20ng of the
832 CellTag-multi-v1 gBlock and 100ng of pSMAL-ctac2 vector were mixed with 2x Phusion PCR
833 master mix in a 20ul reaction volume. The reaction mixture was subjected to the following thermal
834 cycling program: 98ºC for 30 seconds; 15 cycles (98ºC for 8 seconds, 60ºC for 20 seconds, and
835 72ºC for 4.5 minutes); 72ºC for 5 minutes. The parental plasmid was digested by adding 2ul of
836 methylation-sensitive restriction enzyme, *DpnI* (New England Biolabs), and incubating the
837 reaction at 37ºC for 2 hours followed by inactivation at 80ºC for 20 minutes. 10ul of the reaction

838 mix was transformed directly into 100ul of Stellar chemically competent cells (Takara Bio), cells
839 were allowed to recover at 37ºC, 250rpm in 1ml of SOC media and plated on a Nunc Square
840 BioAssay plate (Cat. 166508). Plates were incubated overnight at 37ºC. Bacterial colonies were
841 collected using a scraper and allowed to recover in 150ml of LB media supplemented with
842 100ug/ml Ampicillin. CellTag-multi libraries were purified using a Qiagen High speed maxi prep
843 kit (Cat. 12662) and library complexity was assessed as described below. This cloning was
844 performed four times and libraries from each round were pooled to obtain the final high complexity
845 library.
846
847 **Assessing the complexity of CellTag-multi libraries and allowlisting.** A list of allowed CellTag
848 sequences for each CellTag library was created using amplicon sequencing. 50ng of CellTag
849 plasmid library was mixed with 2x Q5 HF Master Mix, 2.5ul each of 0.5uM primers *bATAC_fwd*
850 and *bATAC_rev* in a 25ul reaction volume and subjected to the following PCR program: 98ºC for
851 30 seconds; 10 cycles (98ºC for 10 seconds; 63ºC for 30 seconds; 72ºC for 1 minute). Two
852 amplicon libraries were generated from each CellTag library plasmid preparation in parallel and
853 sequenced on an Illumina Miseq. For each replicate, reads matching the CellTag sequence
854 pattern $(N)_3GT(N)_3CT(N)_3AG(N)_3TG(N)_3CA(N)_3$ were extracted, sequencing/PCR errors were
855 corrected by collapsing tags within 4 edits of each other using starcode[75] and thresholded to retain
856 CellTags containing at least N reads where N = max(10, 90th percentile/10). An allowlist was
857 created by collecting all CellTag sequences retained in thresholded lists from both replicates.
858 Allowlists from the four CellTag libraries were combined to create the master allowlist for the
859 CellTag-multi library (Supplementary Table 7). The detailed analysis code can be found at:
860 https://github.com/morris-lab/newCloneCalling
861
862 **Species mixing experiment**. For the species mixing experiment, mouse iEP-LT cells were
863 tagged with CellTag-multi-v1 library, containing the barcode pattern
864 $(N)_3GT(N)_3CT(N)_3AG(N)_3TG(N)_3CA(N)_3$ and human HEK 293T cells with CellTag-multi-v0 library,
865 containing the barcode pattern $(N)_5GTA(N)_5CCT(N)_5ATC(N)_5GAT(N)_5$. Nuclei were isolated from
866 both species using the Nuclei Isolation for scATAC-seq protocol from 10x Genomics (CG000169)
867 and mixed in a 1:1 ratio. The mixed nuclei sample was processed using the standard scATAC-
868 seq library preparation protocol (v1 kit) from 10x Genomics with modifications to capture CellTags.
869 Single-cell libraries were sequenced on an Illumina Nextseq-500. The resulting sequencing data
870 was aligned to a mixed species reference using CellTag-ATAC v1. The aligned bam file was used
871 for downstream analysis.

872 Reads matching v0 or v1 CellTags were parsed from the mixed species single-cell aligned
873 bam file. Each cell barcode was assigned to one of four categories, based on CellRanger-ATAC
874 species assignments - human, mouse, doublet, non-cell; the distribution of v0 and v1 reads was
875 assessed across the four categories. Cells with fewer than two CellTag reads across both libraries
876 were discarded and the remaining cells were plotted on a barnyard plot. We quantified inter-
877 species cross-talk of CellTags, by calculating the percent of cells, with at least 2 CellTag
878 reads/cell, having less than 95% of CellTag reads originating from the correct, species-specific
879 CellTag library.
880
881 **Assessing the effect of isRT on chromatin accessibility signal.** We compared the effect of
882 introducing an isRT step on scATAC-seq data quality. For this, two single-cell ATAC libraries were
883 prepared with CellTagged HEK 293T cells using either the original 10x Genomics scATAC library
884 preparation protocol (Original) or our modified method (Modified). Sequencing data from both was
885 processed with ArchR[77], dimensionally reduced using LSI, clustered using Louvain clustering, and
886 peaks were identified across samples. Both datasets were compared across several standard
887 scATAC-seq data quality metrics such as fragment size distribution, TSS scores, the number of
888 unique fragments per cell and Fraction reads in Peaks (FRiP) per cell. To compare genome-wide
889 accessibility data across samples, normalized peak counts (Counts Per Million; CPM) were
890 calculated for each sample and plotted on a scatter plot and the Pearson Correlation coefficient
891 was calculated to quantify the similarity between the accessibility signal of the two samples.
892
893 **Analysis of clones in expanded reprogramming fibroblasts.** A subset of the data obtained
894 from our reprogramming dataset (described in section 3) from Days 12 and 21 was used for this
895 analysis. Clones were identified following the standard computational workflow as described
896 above. CellTag abundance was calculated for each CellTag as the percent of metric filtered cells
897 containing that CellTag. Browser tracks depicting single-cell accessibility fragments were plotted
898 using ArchR. Gene expression and gene scores values were averaged on a clonal level.
899 Spearman correlation coefficients were calculated between clonal gene expression and gene
900 score both within (Intra clonal) and across clones (Inter clonal).
901
902 **Section 2**
903 **Lineage tracing during in vitro mouse hematopoiesis.** LSK cells were purified as described
904 above, counted and 5,500 cells were added to a 96-well U-bottom suspension culture plate
905 (GenClone Cat. 25-224) and allowed to recover in broad myeloid differentiation media[13] consisting

906 of SFEM media (STEMCELL), Pen/Strep, IL-3 (20ng/mL; PeproTech Cat. 213-13), FLT3-L
907 (50ng/mL; PeproTech Cat. 250-31L), IL-11 (50ng/mL; PeproTech Cat. 220-11), IL-5 (10ng/mL;
908 PeproTech Cat. 215-15), EPO (3U/mL; PeproTech Cat. 100-64), TPO (50ng/mL; PeproTech Cat.
909 315-14), and mSCF (50ng/mL; R&D Systems Cat. Q78ED8) and IL-6 (10ng/mL; R&D Systems
910 Cat. 406-ML-005) at 37ºC for 2 hours.
911 To allow clone tracking, cells were transduced for 2 days with 10ul of concentrated
912 CellTag-multi virus (~25k unique CellTag sequences) in 100ul differentiation media, in the
913 presence of 6ug/ml DEAE-Dextran after spin-fection at 800g for 90 minutes at 37ºC. 60 hours
914 (2.5 days) after the start of the experiment, 50% of the cells were collected for single-cell profiling
915 and split equally between scRNA-seq and scATAC-seq assays. The remaining cells were split
916 into 2 technical replicates and re-plated in fresh differentiation media. Finally, all the cells were
917 collected on Day 5 and split between scRNA-seq and scATAC-seq profiling.
918
919 **Single-cell library preparation and sequencing.** The v3 single index Gene Expression kit and
920 the v1 scATAC kit from 10x Genomics were used for single-cell library preparation. CellTag-RNA
921 PCR was used to obtain CellTag amplicon libraries as described above. scRNA-seq libraries were
922 sequenced on an Illumina NovaSeq-6000 and the resulting data was computationally dehopped.
923 CellTag amplicon libraries obtained from scRNA-seq libraries were sequenced on an Illumina
924 NextSeq-500. For read alignment, CellTag and transcriptome read files for each sample were
925 processed together using CellRanger, using a custom mm10 reference containing the GFP CDS
926 and UTR, to produce one aligned bam file per sample. scATAC-seq libraries containing both
927 accessible chromatin and CellTag fragments were sequenced on an Illumina NextSeq-500 and
928 processed using CellRanger-ATAC, using the default mm10 reference genome. Aligned bam files
929 from both modalities were used for CellTag processing, other CellRanger and CellRanger-ATAC
930 outputs were used for downstream single-cell analyses.
931
932 **Basic single-cell and clonal analysis of the Hematopoiesis dataset.** CellRanger generated
933 scRNA-seq count matrices were processed using Seurat. Low-quality cells with high
934 mitochondrial reads, low UMIs, and features per cell were removed. Day 2.5 and Day 5 samples
935 were integrated using SCTransform, dimensionally reduced using PCA, and clustered using
936 Louvain clustering. scRNA-seq clusters were annotated using marker gene expression.
937 Fragments files from scATAC-seq samples were processed using ArchR v1.0.1. Valid cell
938 barcodes, as identified by CellRanger-ATAC and passing default ArchR quality filters were
939 retained. Cells were dimensionally reduced using iterative LSI and clustered using Louvain

clustering. Cell type labels were transferred to scATAC-seq clusters using Seurat label transfer and annotations were manually inspected using marker gene scores. For RNA-ATAC co-embedding, scRNA-seq gene expression matrix and scATAC-seq MAGIC imputed[78] Gene Score matrix, as obtained from ArchR, were used as input to the RunCCA function in Seurat. A union set of the top 5000 highly variable genes from each dataset were used for this co-embedding.

For clone calling, reads mapping to the CellTag barcode were extracted from single-cell aligned bam files as obtained from CellRanger and CellRanger-ATAC and cell x CellTag UMI matrices were obtained. CellTag data within each modality was merged, retaining sample-of-origin information in the cell barcode, and cell x CellTag UMI (for RNA) and read (for ATAC) count matrices were obtained for each modality. The RNA matrix was binarized at a threshold of more than one UMI count and cells with 2 to 25 CellTags were retained. The ATAC matrix was binarized at a threshold of more than one read count and cells with 1 to 25 CellTags were retained. The two filtered matrices were merged, cell-cell Jaccard similarity matrix was computed and thresholded at 0.6 (for cell pairs within the same modality) and 0.5 (for cell pairs across modalities). The final thresholded matrix was used to identify clones across the entire dataset. Clone-cell embedding was computed as described above, and ForceAtlas2 was used to jointly visualize clones and cells. This embedding was also generated separately for sub-clones where clones were split either by modality or by both, time point and modality. For single-modality clonal analysis, Cell x CellTag matrices for each modality were processed separately with the same thresholds as above. A Jaccard threshold of 0.5 was used for ATAC clone calling and 0.6 was used for RNA clone calling. Lineage hierarchies were obtained using clone coupling as previously described[13]

State-fate linkage in hematopoiesis. To link cell state with fate, we first obtained all clones spanning the two time points (state-fate clones). Each state-fate clone was assigned a fate label, which was the most common fate amongst its Day 5 sisters. Less common lineages were grouped based on similarity, e.g. Erythroid and Megakaryocytes (Ery/Meg); Eosinophils, Basophils, and Mast Cells (Baso/Eos/Mast). Ccr7 DCs and plastoid DCs (DCs). Clones annotated to transition/ progenitor fates were excluded from state-fate analysis unless otherwise specified. Fate bias scores were calculated as percent of Day 5 fate sisters belonging to the annotated fate label.

To map Day 2.5 (state) sub-clones on the clone-cell embedding, we split each clone into sub-clones based on the time point of collection and assay of each sister, to obtain up to four sub-clones RNA/ATAC – state/fate sub-clones. The clone-cell embedding was recomputed using these sub-clones. Overlap between RNA and ATAC sub-clones across the two single-cell

modalities was calculated within each ‘fate potential’ group using the Wasserstein distance metric computed with a 30-dimensional embedding of the sub-clone nodes obtained using the UMAP algorithm. To quantify if state sub-clones closer to the periphery of a 'fate potential’ group were less fate biased, we devised a closeness metric, which is the minimum distance of a state sub-clone from the centroid of an alternative fate potential group. A higher closeness metric would mean that a state sub-clone is farther away from centroids of other fate potential groups. The relationship between the closeness metric and fate bias was plotted using a percentile plot, with percentile rank for the closeness metric on the X-axis and mean fate bias scores for state sub-clones passing the percentile rank on the Y-axis.

To characterize functional priming of cell state, Day 2.5 state sisters in each fate potential group were compared to the rest in gene expression and TF activity space. For scRNA-seq features, we used residuals obtained for the top 3000 highly variable genes after SCTransform normalization in Seurat. For scATAC-seq features, we used TF activity z-scores obtained from chromVAR using the default mouse motif set in ArchR (884 TF motifs). Correction for multiple hypothesis testing was performed using the Benjamini-Hochberg method, setting the FDR threshold for significance at 0.05, unless otherwise specified.

Fate prediction from cell state using machine learning. State-fate machine learning was performed to quantify the predictability of cell fate from early state. A machine learning classifier was tasked to predict the discrete clonal fate label Y as obtained above (possible fate labels: ‘progenitor’, ‘monocyte’, ‘neutrophil’, ‘Lym/pDC/Ccr7-DC’, ‘Ery/Meg’ or ‘Baso/Eos/Mast’), from an input vector of single-cell features X of Day 2.5 cells. For RNA only model, we used residuals of the top 3000 genes for input, for ATAC only model, we used TF activity z-scores (with k-nn imputation where k=20) as input and for the RNA+ATAC model, we randomly paired RNA and ATAC cells within the same sub-clone and concatenated their respective RNA and ATAC feature vectors and used those as input. For training, we used the Repeated Stratified k-fold cross-validation procedure setting both *n_splits* and *n_repeats* to 5. Model performance was evaluated using accuracy and Weighted F1 score.

For each machine learning task, we tested a panel of classifier architectures, logistic regression, LightGBM, and Random Forest. Each was trained and evaluated using the procedure described above. Hyperparameter tuning was performed for each and the following values were tested:

- Random Forest: n_estimators: [100, 300, 1000], max_depth: [10, 50, None], min_samples_leaf: [1, 2, 4], bootstrap: [True, False]

1008 • LightGBM: num_leaves: [7,15,31,80], max_depth: [5,9,30], min_data_in_leaf: [20, 40, 80],
1009 bagging_fraction: [0.8,1], bagging_freq: [3], feature_fraction: [0.1, 0.9]
1010 • Logistic Regression: penalty: ['l2', 'none'], C: np.logspace(-4, 4, 20), solver:
1011 ['lbfgs','newton-cg','saga'], max_iter: [1000]
1012 The python library 'scikit-learn' was used for all machine learning analysis.
1013
1014 **Fate prediction using TF activities derived from distal, intronic, exonic, and promoter peak**
1015 **sets.** ATAC peaks were categorized as intronic, exonic, promoter, or distal using default ArchR
1016 definitions. TF activity scores were calculated for each peak set independently and used for state-
1017 fate prediction analysis as described above. To test if variation in model performance was due to
1018 different numbers of peaks in each set, all peak sets were randomly sub-sampled to 8823 peaks
1019 (number of peaks in the exonic set), TF activity scores were calculated again and state-fate
1020 prediction was performed using these new scores.
1021
1022 **SHapley Additive exPlanations (SHAP) analysis.** The shap python package was used to
1023 perform SHAP analysis and interpret trained machine learning models. The TreeExplainer
1024 function from the 'shap' python package was used to calculate SHAP values for trained random
1025 forest models. For each input feature and fate label, SHAP values were calculated using each
1026 data point in the 25 test sets (*n_splits* x *n_repeats*), resulting in 5 SHAP values per data point per
1027 feature. This helped average out any rare outlier values generated due to a model training artifact.
1028 Feature importance scores were calculated for each input feature for the prediction of each fate
1029 label, by taking the mean of absolute SHAP values for each feature-fate combination. To identify
1030 features positively or negatively correlated with the prediction of a fate label, SHAP correlation
1031 was performed. For each input feature, the Pearson correlation coefficient between its values
1032 (expression/TF activity) and its SHAP values for a given fate was calculated, resulting in one
1033 correlation value per feature per fate.
1034
1035 **Section 3**
1036 **Lineage tracing during iEP reprogramming.** Cryo-preserved P0 MEFs were thawed and
1037 seeded in 0.1% gelatin-coated six-well plates, in DMEM supplemented with 10% FBS, 2 mM l-
1038 glutamine, and 50 mM β-mercaptoethanol (Life Technologies) and penicillin-streptomycin at a
1039 density of 30,000 cells/well. After overnight recovery at 37ºC, cells were transduced every 12
1040 hours for 2 days, with fresh Hnf4α-T2A-Foxa1 retrovirus in the presence of 4 µg/ml protamine
1041 sulfate (Sigma-Aldrich). During the last round of transduction, the retroviral mix was supplemented

1042 with CellTag-multi lentiviral library to initiate clone tracking. On Day 0 of reprogramming, cell
1043 culture media was changed to hepato-medium (DMEM:F-12, supplemented with 10% FBS, 1
1044 μg/ml insulin (Sigma-Aldrich), 100 nM dexamethasone (Sigma-Aldrich), 10 mM nicotinamide
1045 (Sigma-Aldrich), 2 mM l-glutamine, 50 mM β-mercaptoethanol (Life Technologies), and penicillin-
1046 streptomycin, containing 20 ng/ml epidermal growth factor (Sigma-Aldrich)). After 72 hours (Day
1047 3 of reprogramming), cells were dissociated, two-thirds of the cells were collected for single-cell
1048 sequencing and the remaining cells were re-plated on 6-well plates coated with 5 μg/cm^2 Type I
1049 rat collagen (Gibco, A1048301). Two additional samples were collected on Days 11 and 21 for
1050 single-cell sequencing. We used the 10x Genomics v3.1 dual index Gene Expression kit (PN-
1051 1000268) and the v1.1 ATAC kit (PN-1000175) for single-cell profiling. This experiment was
1052 performed in two biological replicates.
1053 CellTag PCR was performed for all scRNA-seq and scATAC-seq libraries, as described
1054 above. scRNA-seq and scATAC-seq libraries were sequenced on an Illumina NovaSeq-6000.
1055 CellTag amplicon libraries were sequenced on an Illumina NextSeq-500 to avoid any index
1056 hopping related artifacts. For read alignment, CellTag and transcriptome/chromatin read files for
1057 each sample were processed together using CellRanger/CellRanger-ATAC, to produce one
1058 aligned bam file per sample. Aligned bam files from both modalities were used for CellTag
1059 processing, other CellRanger and CellRanger-ATAC outputs were used for downstream single-
1060 cell analyses.
1061
1062 **Basic single-cell and clonal analysis of the direct reprogramming dataset.** scRNA-seq count
1063 matrices were processed using Seurat. Quality filtering was performed to remove cells with high
1064 mitochondrial reads and low UMIs and genes per cell. scRNA-seq samples across all time points
1065 and biological replicates were integrated, dimensionally reduced using PCA, and clustered using
1066 Louvain clustering. Cells from Days 12 and 21 were subsetted and re-clustered. Single-cell
1067 identity scores were obtained using Capybara, using Fibroblasts (MEFs), , and reprogrammed,
1068 and dead-end trajectories from a previous dataset[7] as references. Cell clusters were annotated
1069 as 'reprogrammed', 'dead-end', or 'transition' based on these cell identity scores and marker gene
1070 expression. Fragments files from scATAC-seq samples were processed using ArchR. Valid cell
1071 barcodes, as identified by CellRanger-ATAC and passing default ArchR quality filters were
1072 retained. Cells were dimensionally reduced using iterative LSI and clustered using Louvain
1073 clustering. Cells were annotated as 'reprogrammed', 'dead-end', or 'transition' based on marker
1074 gene accessibility. For RNA-ATAC co-embedding, scRNA-seq gene expression matrix and
1075 scATAC-seq MAGIC imputed[78] Gene Score matrix, as obtained from ArchR, were used as input

1076 to the RunCCA function in Seurat. A union set of the top 2000 highly variable genes from each
1077 dataset were used for this co-embedding.
1078 For clone calling, reads mapping to the CellTag barcode were extracted from single-cell
1079 aligned bam files as obtained from CellRanger and CellRanger-ATAC. CellTag data within each
1080 modality was merged, retaining sample-of-origin information in the cell barcode, and cell x CellTag
1081 UMI (for RNA) and read (for ATAC) count matrices were obtained for each modality. The RNA
1082 matrix was binarized at a threshold of more than one UMI count and cells with 2 to 25 CellTags
1083 were retained. The ATAC matrix was binarized at a threshold of more than one read count and
1084 cells with 2 to 25 CellTags were retained. The two filtered matrices were merged, cell-cell Jaccard
1085 similarity matrix was computed and thresholded at 0.6. The final thresholded matrix was used to
1086 identify clones across the entire dataset. Clone-cell embedding was computed as described
1087 above and the UMAP algorithm was used to jointly visualize clones and cells.
1088
1089 **State-fate analysis for the direct reprogramming dataset.** Clones were annotated with one of
1090 three fates – 'reprogrammed', 'transition', or 'dead-end', based on the most abundant cell type
1091 amongst fate sisters. Clonal fate bias scores were calculated as percent of fate sisters (Days 12
1092 and 21) belonging to the annotated fate label. Alluvial plots were constructed using the ggAlluvial
1093 R package. State-fate machine learning analysis was performed exactly as described in the
1094 hematopoiesis section. Classification models were trained to predict either 'reprogrammed' or
1095 'dead-end' fates. Since the frequency distribution of fate labels was less skewed for the
1096 reprogramming dataset, only prediction accuracy scores were used as performance metrics.
1097 CellRank analysis was performed for a 40,000-cell subset of the scRNA-seq dataset, due to
1098 scalability limitations. For feature enrichment analysis, Day 3 sisters in state-fate clones were
1099 grouped by fate. Seurat FindMarkers function was used to identify gene expression markers and
1100 ArchR getMarkerFeatures function was used to identify peak and TF activity markers for each of
1101 the following cell groups - uninduced MEFs, on-target destined cells and, off-target destined cells,
1102 in a series of one versus all comparisons. For peak and TF activity comparisons, both on-target
1103 and off-target cell groups were expanded using k-nearest neighbors (k=5). Uniquely enriched
1104 features (genes/peaks/TFs) were obtained by removing features that were identified as markers
1105 of more than one cell group. TF activity results were further refined by discarding TFs with low
1106 gene score-TF activity correlation (< 0.3). Motif enrichment analysis was performed using the
1107 HOMER package[79] for both on-target and off-target DERs using MEF DERs as background, to
1108 better resolve fate-specific motif enrichment. Mouse ENCODE ELS elements were obtained from
1109 the ENCODE SCREEN database[43]. Only genomic regions annotated as dELS, pELS, dELS,

1110 CTCF-bound, or pELS, CTCF-bound in the SCREEN database were used for enrichment
1111 analysis. The FigR[42] package was used for peak-to-gene linkage analysis. Optimal matching was
1112 used to pair RNA and ATAC cells from the same time points followed by the runGenePeakcorr
1113 function to identify peak-gene pairs. Peak-gene pairs with an adjusted p-value greater than or
1114 equal to 0.05 were discarded. *Foxa1* and *Hnf4α* ChIP-seq peaks from Day 2 of reprogramming
1115 were obtained[55]. These peak sets were added as custom annotations in ArchR and single-cell
1116 accessibility z-scores for each peak set were computed using the addDeviationsMatrix function
1117 in ArchR.
1118
1119 **Computational analysis related to ZFP281 motifs.** Tomtom analysis[58] from the MEME-ChIP
1120 package was used to find highly similar motifs to *Zfp281*. The *Zfp281* position frequency matrix
1121 (PFM) was obtained from ArchR and used as input to the Tomtom web interface. Highly correlated
1122 TF motifs with q-value less than 0.05 were obtained, these were further subsetted for TF activities
1123 enriched in off-target destined cells resulting in a total of four TF motifs for comparison with
1124 *Zfp281*. *Zfp281* ChIP-seq peaks were obtained[56] and single-cell accessibility z-scores were
1125 computed using the addDeviationsMatrix function in ArchR. *Zfp281* gene targets[59] were used as
1126 inputs for a state-fate prediction model, which was trained and evaluated as described above and
1127 compared to a sized-matched set of random genes.
1128
1129 **Plasmid cloning related to Foxd2 and Zfp281 experiments.** Non-targeting shRNA construct
1130 was obtained from Sigma (SHC202; pLKO.5-puro Control Plasmid). Zfp281 targeting shRNA
1131 gene was obtained from Sigma (Clone ID: TRCN0000255746) and cloned into the pLKO.5-puro
1132 lentiviral construct (Sigma SHC201). For over-expression, cDNA fragments were cloned in the
1133 pGCDNsam retroviral construct. *Zfp281* cDNA was obtained from OriGene (Cat. MC205914) and
1134 *Foxd2* cDNA was reverse transcribed from RNA obtained from long-term iEP cells.
1135
1136 **Reprogramming with Foxd2 and Zfp281 perturbations.** Reprogramming was performed as
1137 described above, with the following modifications. For over-expression, cells were transduced
1138 with a 1:1 mixture of *Foxd2/Zfp281* retrovirus and *Hnf4α-Foxa1* reprogramming retrovirus every
1139 12 hours for 2 days. Control cells were transduced with a 1:1 mixture of a GFP control retrovirus
1140 and *Hnf4α-Foxa1* reprogramming retrovirus for the same amount of time. For knockdown, cells
1141 were transduced with the non-targeting control/*Zfp281*-shRNA lentivirus every 12 hours for 1 day
1142 after the 2-day *Hnf4α-Foxa1* retroviral transduction was completed.
1143

1144 **Single-cell analysis for Foxd2 and Zfp281 experiments.** Dual indexed v3.1 scRNA-seq
1145 libraries were prepared for all four samples (*Zfp281* OE, OE Control, *Zfp281* KD, KD Control)
1146 according to the manufacturer's instructions (CG000315) and sequenced on a Nextseq-500.
1147 Count matrices were generated and integrated using CellRanger count and aggr commands and
1148 processed using Seurat. Quality filtering was performed to remove cells with high mitochondrial
1149 reads and low UMIs and genes per cell. Cells were dimensionally reduced using PCA, cell cycle
1150 regressed, clustered using Louvain clustering, and visualized using UMAP. Capybara identity
1151 scores were calculated as described in the iEP lineage tracing section above. Markers for each
1152 lineage across time points and uninduced MEFs were obtained (log2 fold change > 0.7, adjusted
1153 p-value < 0.05) and used for gene module scoring for all four samples. Cell clusters displaying
1154 strong enrichment of on-target or off-target markers were annotated with the respective fates.
1155 pROGENY pathway analysis[62] was used to calculate single-cell activity scores for the TGF-β
1156 signaling pathway.
1157
1158 **Colony formation assays.** Colony formation assays were performed as previously described[7].
1159 Reprogramming cells were seeded at low plating density in collagen-coated 6-well plates within
1160 the first 4 days and allowed to form colonies over 2 weeks of reprogramming. Following this, cells
1161 were fixed using 4% paraformaldehyde, permeabilized using 0.1% Triton-X and processed for
1162 CDH1 (E-Cadherin) staining using the VIP peroxidase substrate kit (Vector labs SK4600) and
1163 anti-mouse E-Cadherin primary antibody (1:100, BD Biosciences). Stained colonies were imaged
1164 using a flatbed scanner and quantified using the following script: https://github.com/morris-
1165 lab/Colony-counter
1166
1167 **Quantitative PCR and analysis.** Cells were collected for RNA extraction (RNeasy kit; QIAgen)
1168 on Day 12 of reprogramming and reverse transcribed using the Maxima RT kit (ThermoFisher
1169 K1672). 20ng of reverse transcribed RNA was mixed with TaqMan™ Gene Expression Master
1170 Mix (ThermoFisher Scientific) and gene-specific TaqMan™ probes (Supplementary Table 8) in a
1171 20ul reaction volume and processed according to manufacturer's instructions (4371135) on the
1172 StepOne Plus qPCR system. Per gene fold change for Foxd2 overexpressing cells was calculated
1173 relative to control reprogramming cells (*Hnf4α-Foxa1* and GFP control overexpression) that were
1174 processed in parallel, after normalization to the housekeeping gene, *Actb*.
1175
1176 **Reprogramming with TGF-β inhibition.** Cells were reprogrammed as described above. Cells
1177 were cultured in hepatic media supplemented with 2.6μM of SB431542 (STEM CELL 72232), a

small molecule inhibitor of TGF-b signaling starting on Day 0 of reprogramming. SB431542 containing media was changed every 2 days. Cells were collected for qPCR analysis on Day 5 of reprogramming and processed as described above.

References

1. What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism? *Cell Syst.* **4**, 255–259 (2017).
2. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
3. Morris, S. A. The evolving concept of cell identity in the single cell era. *Dev.* **146**, (2019).
4. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166–179 (2018).
5. VanHorn, S. & Morris, S. A. Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. *Developmental Cell* vol. 56 7–21 (2021).
6. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
7. Biddy, B. A. *et al.* Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
8. Bowling, S. *et al.* An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* **181**, 1410-1422.e27 (2020).
9. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
10. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
11. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
12. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* 1–18 (2020) doi:10.1038/s41576-020-0223-2.
13. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (80-.).* **367**, (2020).
14. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166–179 (2018).

1212 15. LS, L. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-
1213 Cell Genomics. *Cell* **176**, 1325-1339.e22 (2019).
1214 16. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and
1215 chromatin profiling. *Nat. Biotechnol. 2020 394* **39**, 451–461 (2020).
1216 17. Kong, W. *et al.* CellTagging: combinatorial indexing to simultaneously map lineage and
1217 identity at single-cell resolution. *Nat. Protoc.* **15**, 750–772 (2020).
1218 18. Guo, C. *et al.* CellTag Indexing: Genetic barcode-based sample multiplexing for single-
1219 cell genomics. *Genome Biol.* **20**, 1–13 (2019).
1220 19. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21
1221 (2019).
1222 20. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
1223 Projection for Dimension Reduction. (2018) doi:10.48550/arxiv.1802.03426.
1224 21. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph
1225 Layout Algorithm for Handy Network Visualization Designed for the Gephi Software.
1226 *PLoS One* **9**, e98679 (2014).
1227 22. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory
1228 Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
1229 23. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous
1230 process. *Nat. Cell Biol. 2017 194* **19**, 271–281 (2017).
1231 24. Liu, Z. *et al.* Fate Mapping via Ms4a3-Expression History Traces Monocyte-Derived Cells.
1232 *Cell* **178**, 1509-1525.e19 (2019).
1233 25. Tiedt, R., Schomber, T., Hao-Shen, H. & Skoda, R. C. Pf4-Cre transgenic mice allow the
1234 generation of lineage-restricted gene knockouts for studying megakaryocyte and platelet
1235 function in vivo. *Blood* **109**, 1503–1506 (2007).
1236 26. Zriwil, A. *et al.* Direct role of FLT3 in regulation of early lymphoid progenitors. *Br. J.*
1237 *Haematol.* **183**, 588–600 (2018).
1238 27. Stehling-Sun, S., Dade, J., Nutt, S. L., DeKoter, R. P. & Camargo, F. D. Regulation of
1239 lymphoid versus myeloid fate 'choice' by the transcription factor Mef2c. *Nat. Immunol.* **10**,
1240 289–296 (2009).
1241 28. Yu, Y. *et al.* Bcl11a is essential for lymphoid development and negatively regulates p53.
1242 *J. Exp. Med.* **209**, 2467–2483 (2012).
1243 29. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
1244 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat.*
1245 *Methods 2017 1410* **14**, 975–978 (2017).

1246 30. Safi, F. *et al.* Concurrent stem- and lineage-affiliated chromatin programs precede
1247 hematopoietic lineage restriction. *Cell Rep.* **39**, 110798 (2022).
1248 31. Lundberg, S. M. & Lee, S. I. A Unified Approach to Interpreting Model Predictions. *Adv.*
1249 *Neural Inf. Process. Syst.* **2017-December**, 4766–4775 (2017).
1250 32. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by
1251 defined factors. *Nature* **475**, 390–393 (2011).
1252 33. Morris, S. A. *et al.* Dissecting Engineered Cell Types and Enhancing Cell Fate
1253 Conversion via CellNet. 889–902 (2014) doi:10.1016/j.cell.2014.07.021.
1254 34. Wang, H., Yang, Y., Liu, J. & Qian, L. Direct cell reprogramming: approaches,
1255 mechanisms and progress. *Nature Reviews Molecular Cell Biology* vol. 22 410–424
1256 (2021).
1257 35. Treutlein, B. *et al.* Dissecting direct reprogramming from fibroblast to neuron using single-
1258 cell RNA-seq. *Nature* **534**, 391–395 (2016).
1259 36. Cahan, P. *et al.* CellNet: Network biology applied to stem cell engineering. *Cell* **158**, 903–
1260 915 (2014).
1261 37. Kong, W. *et al.* Capybara: A computational tool to measure cell identity and fate
1262 transitions. *Cell Stem Cell* **29**, 635-649.e11 (2022).
1263 38. Kamimoto, K. *et al.* Gene Regulatory Network Reconfiguration in Direct Lineage
1264 Reprogramming (In review). *Stem Cell reports*.
1265 39. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat. Methods 2022 192*
1266 **19**, 159–170 (2022).
1267 40. Liu, Y. *et al.* Increased TEAD4 expression and nuclear localization in colorectal cancer
1268 promote epithelial-mesenchymal transition and metastasis in a YAP-independent
1269 manner. *Oncogene* **35**, 2789–2800 (2016).
1270 41. Khalid, A. B. *et al.* GATA4 regulates mesenchymal stem cells via direct transcriptional
1271 regulation of the WNT signalosome. *Bone* **144**, (2021).
1272 42. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics.
1273 *Cell Genomics* 100166 (2022) doi:10.1016/J.XGEN.2022.100166.
1274 43. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse
1275 genomes. *Nat. 2020 5837818* **583**, 699–710 (2020).
1276 44. Shi, Y. *et al.* Neural cell adhesion molecule modulates mesenchymal stromal cell
1277 migration via activation of MAPK/ERK signaling. *Exp. Cell Res.* **318**, 2257–2267 (2012).
1278 45. Tsuda, T., Wang, H., Timpl, R. & Chu, M. L. Fibulin-2 expression marks transformed
1279 mesenchymal cells in developing cardiac valves, aortic arch vessels, and coronary

1280 vessels. *Dev. Dyn.* **222**, 89–100 (2001).
1281 46. Stacker, S. A. & Achen, M. G. Emerging Roles for VEGF-D in Human Disease.
1282 *Biomolecules* **8**, (2018).
1283 47. Shi, S. *et al.* Sonic hedgehog promotes endothelial differentiation of bone marrow
1284 mesenchymal stem cells via VEGF-D. *J. Thorac. Dis.* **10**, 5476–5488 (2018).
1285 48. Garattini, E., Fratelli, M. & Terao, M. The mammalian aldehyde oxidase gene family.
1286 *Hum. Genomics* **4**, 119–130 (2009).
1287 49. Wang, R. *et al.* Construction of a cross-species cell landscape at single-cell level. *Nucleic*
1288 *Acids Res.* 2022 (2022) doi:10.1093/NAR/GKAC633.
1289 50. Chen, K. T. *et al.* Liver X receptor α (LXRα/NR1H3) regulates differentiation of
1290 hepatocyte-like cells via reciprocal regulation of HNF4α. *J. Hepatol.* **61**, 1276–1286
1291 (2014).
1292 51. Nobre, A. R. *et al.* ZFP281 drives a mesenchymal-like dormancy program in early
1293 disseminated breast cancer cells that prevents metastatic outgrowth in the lung. *Nat.*
1294 *Cancer 2022* 1–16 (2022) doi:10.1038/s43018-022-00424-8.
1295 52. Gharibeh, L. *et al.* GATA6 is a regulator of sinus node development and heart rhythm.
1296 *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2007322118 (2020).
1297 53. Sato, A. *et al.* C/EBPβ Isoforms Regulate Proliferation and Differentiation of
1298 Regenerating Hematopoietic Stem/Progenitor Cells. *Blood* **134**, 3713 (2019).
1299 54. Harsha Krovi, S. *et al.* Thymic iNKT single cell analyses unmask the common
1300 developmental program of mouse innate T cells. *Nat. Commun. 2020 111* **11**, 1–15
1301 (2020).
1302 55. Horisawa, K. *et al.* The Dynamics of Transcriptional Activation by Hepatic
1303 Reprogramming Factors. *Mol. Cell* **79**, 660-676.e8 (2020).
1304 56. Fidalgo, M. *et al.* Zfp281 Coordinates Opposing Functions of Tet1 and Tet2 in Pluripotent
1305 States. *Cell Stem Cell* **19**, 355–369 (2016).
1306 57. Hahn, S., Jackstadt, R., Siemens, H., Hünten, S. & Hermeking, H. SNAIL and miR-34a
1307 feed-forward regulation of ZNF281/ZBP99 promotes epithelial–mesenchymal transition.
1308 *EMBO J.* **32**, 3079 (2013).
1309 58. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity
1310 between motifs. *Genome Biol.* **8**, 1–9 (2007).
1311 59. Wang, Y. *et al.* A permissive chromatin state regulated by ZFP281-AFF3 in controlling the
1312 imprinted Meg3 polycistron. *Nucleic Acids Res.* **45**, 1177–1185 (2017).
1313 60. Ji, W., Mu, Q., Liu, X. Y., Cao, X. C. & Yu, Y. ZNF281-miR-543 Feedback Loop

1314 Regulates Transforming Growth Factor-β-Induced Breast Cancer Metastasis. *Mol. Ther. -*
1315 *Nucleic Acids* **21**, 98–107 (2020).
1316 61. Xu, J., Lamouille, S. & Derynck, R. TGF-beta-induced epithelial to mesenchymal
1317 transition. *Cell Res.* **19**, 156–172 (2009).
1318 62. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer
1319 gene expression. *Nat. Commun. 2017 91* **9**, 1–11 (2018).
1320 63. Inman, G. J. *et al.* SB-431542 Is a Potent and Specific Inhibitor of Transforming Growth
1321 Factor-β Superfamily Type I Activin Receptor-Like Kinase (ALK) Receptors ALK4, ALK5,
1322 and ALK7. *Mol. Pharmacol.* **62**, 65–74 (2002).
1323 64. Orge, I. D. *et al.* Phenotype instability of hepatocyte-like cells produced by direct
1324 reprogramming of mesenchymal stromal cells. *Stem Cell Res. Ther.* **11**, 1–15 (2020).
1325 65. Stone, N. R. *et al.* Context-Specific Transcription Factor Functions Regulate Epigenomic
1326 and Transcriptional Dynamics During Cardiac Reprogramming. *Cell Stem Cell* **25**, 87
1327 (2019).
1328 66. Goyal, Y. *et al.* Pre-determined diversity in resistant fates emerges from homogenous
1329 cells after anti-cancer drug treatment. *bioRxiv* 2021.12.08.471833 (2021)
1330 doi:10.1101/2021.12.08.471833.
1331 67. Mold, J. E. *et al.* Clonally heritable gene expression imparts a layer of diversity within cell
1332 types. *bioRxiv* 2022.02.14.480352 (2022) doi:10.1101/2022.02.14.480352.
1333 68. Fennell, K. A. *et al.* Non-genetic determinants of malignant clonal fitness at single-cell
1334 resolution. *Nature* **601**, 125–131 (2022).
1335 69. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection of
1336 hybrid EMT states. *Cancer Cell* **39**, 1150-1162.e9 (2021).
1337 70. Zhisong, H. *et al.* Lineage recording in human cerebral organoids. *Nat. Methods* (2021).
1338 71. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands
1339 of single cells. *Science (80-.).* **361**, 1380–1385 (2018).
1340 72. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and
1341 chromatin accessibility in the same cell. *Nat. Biotechnol. 2019 3712* **37**, 1452–1457
1342 (2019).
1343 73. Zhu, C. *et al.* An ultra high-throughput method for single-cell joint analysis of open
1344 chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
1345 74. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and
1346 Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
1347 75. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs

search. *Bioinformatics* **31**, 1913–1919 (2015).

76. Bond, S. R. & Naus, C. C. RF-Cloning.org: an online tool for the design of restriction-free cloning projects. *Nucleic Acids Res.* **40**, W209 (2012).

77. JM, G. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

78. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).

79. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).