Coordinated Changes in Gene Expression Kinetics Underlie both Mouse and Human Erythroid Maturation

Melania Barile[1, 2], Ivan Imaz-Rosshandler[1, 2], Isabella Inzani [3], Shila Ghazanfar[4], Jennifer Nichols[2, 5], John C. Marioni[4, 6, 7], Carolina Guibentif[1, 2, 8,*], Berthold Göttgens[1, 2,*]

1. Department of Haematology, University of Cambridge, CB2 0AW Cambridge, UK
2. Wellcome-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, CB2 0AW Cambridge, UK
3. University of Cambridge Metabolic Research Laboratories and MRC Metabolic Diseases Unit, CB2 0QQ Cambridge, UK
4. Cancer Research UK Cambridge Institute, University of Cambridge, CB2 0RE Cambridge, UK
5. Department of Physiology, Development and Neuroscience, University of Cambridge, CB2 3DY Cambridge, UK
6. Wellcome Sanger Institute, Wellcome Genome Campus, CB10 1SA Cambridge, UK
7. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, CB10 1SD Cambridge, UK
8. Sahlgrenska Center for Cancer Research, Department of Microbiology and Immunology, University of Gothenburg, 413 90 Gothenburg, Sweden

* Co-corresponding authors

BG: bg200@cam.ac.uk

CG: carolina.guibentif@gu.se

Abstract

Background: Single cell technologies are transforming biomedical research, including the recent demonstration that unspliced pre-mRNA present in single cell RNA-Seq permits prediction of future expression states. Here we applied this 'RNA velocity concept' to an extended timecourse dataset covering mouse gastrulation and early organogenesis. **Results:** Intriguingly, RNA velocity correctly identified epiblast cells as the starting point, but several trajectory predictions at later stages were inconsistent with both real time ordering and existing knowledge. The most striking discrepancy concerned red blood cell maturation, with velocity-inferred trajectories opposing the true differentiation path. Investigating the underlying causes revealed a group of genes with a coordinated step-change in transcription, thus violating the assumptions behind current velocity analysis suites, which do not accommodate time-dependent changes in expression dynamics. Using scRNA-Seq analysis of chimeric mouse embryos lacking the major erythroid regulator *Gata1*, we show that genes with the step-changes in expression dynamics during erythroid differentiation fail to be up-regulated in the mutant cells, thus underscoring the coordination of modulating transcription rate along a differentiation trajectory. In addition to the expected block in erythroid maturation, the *Gata1*$^{-}$ chimera dataset revealed induction of PU.1 and expansion of megakaryocyte progenitors. Finally, we show that erythropoiesis in human fetal liver is similarly characterized by a coordinated step-change in gene expression. **Conclusions:** By identifying a limitation of the current velocity framework coupled with *in vivo* analysis of mutant cells, we reveal a coordinated step-change in gene expression kinetics during erythropoiesis, with likely implications for many other differentiation processes.

(247 words)

Keywords

RNA velocity; gastrulation; erythropoiesis; Gata1

Background

Cellular differentiation into diverse cell types underpins all metazoan development. Moreover, cellular differentiation processes are also crucial for stem cell-mediated tissue maintenance, and their perturbation has been implicated in ageing-associated regenerative failure as well as malignant transformation (1, 2). Since cellular differentiation decisions are made at the level of individual cells, elucidation of the underlying molecular mechanisms requires the use of single cell approaches. It is no surprise therefore that recent innovations in single cell molecular profiling technologies have been embraced rapidly by developmental and stem cell biologists, with complete single cell gene expression maps now available for developing embryos of several model organisms (3-5, reviewed in 6), as well as large-scale datasets covering adult tissue homeostasis (7-9).

Comprehensive molecular profiling necessarily entails the generation of snapshot data, because cells need to be fixed to examine their molecular content. This in turn represents a major drawback for the study of differentiation processes, which commonly occur over extended timeframes via complex trajectories underpinned by intricate decision-making processes. Much excitement was therefore generated by a recent seminal study (10), which demonstrated that

65 unspliced pre-mRNA present in scRNA-Seq datasets can be exploited to predict likely future
66 expression states. This so-called RNA velocity concept is based on the notion that the ratio
67 between unspliced and spliced RNA differs depending on whether a gene is in the process of
68 being up- or downregulated. During upregulation, there is a relative increase in newly transcribed
69 unspliced RNA, with the converse occurring during downregulation. The RNA velocity framework
70 has rapidly gained traction across the wider single cell community, being applied across multiple
71 experimental systems (11-13), and also extended as part of the scVelo analysis suite (14), which
72 allows inclusion of genes whose transcript levels are not in steady state.

73 One system where the RNA velocity concept has particular potential is erythropoiesis, the
74 process whereby oxygen-transporting red blood cells are generated from multipotent
75 haematopoietic progenitors. Research into the transcriptional control processes of
76 erythropoiesis led to several paradigmatic discoveries, including the dissection of distal
77 transcriptional control elements (15-17), as well as antagonistic transcription factor pairings as
78 executors of lineage choice in multipotent progenitors (18). During embryogenesis, a first so-
79 called primitive wave of erythropoiesis occurs in the yolk sac, followed by a second definitive
80 wave, initiated also in the yolk sac, then predominantly in the fetal liver and later in the adult
81 bone marrow (19). The zinc finger protein Gata1 represents the archetypal erythroid
82 transcription factor, and is required for the maturation of both primitive and definitive erythroid
83 cells (20-23), as well as megakaryocyte maturation (24). However, the precise molecular
84 processes affected by Gata1 deletion in early embryonic erythropoiesis have remained obscure,
85 principally because conventional biochemical methods are unsuitable for the very small number
86 of cells present at these early developmental stages.

87 Here, we have applied RNA velocity to a recently published scRNA-Seq dataset of nine sequential
88 timepoints, spaced 6 hours apart, which encompass mouse gastrulation and early organogenesis
89 (25). We observed that some of the inferred trajectories are incompatible with the existing
90 biological knowledge, as well as with the real time ordering derived from the sequential sampling
91 timepoints. For erythroid differentiation in particular, we show that failure of the velocity
92 framework is due to a concerted increase in transcription rate of a subset of erythroid genes,
93 midway through the red blood cell maturation trajectory. Analysis of *Gata1*$^{-}$ chimeric embryos
94 underscores the concerted nature of this expression boost, consistent with the notion that such
95 concerted upregulation events may be a feature of stabilizing a given differentiated cellular state.

96

97

98 **Results**

99 **Limitations of RNA velocity trajectory inference at organismal scale**

100 To evaluate RNA velocity-based trajectory inference with a complex dataset, we applied the
101 scVelo analysis pipeline (14) to a recently reported timecourse scRNA-Seq dataset covering
102 mouse gastrulation and early organogenesis. This mouse gastrulation atlas contains
103 approximately 120,000 single cell transcriptomes across nine sequential timepoints covering 37
104 major cell types (25). Prior to scVelo analysis, we removed extraembryonic ectoderm and
105 extraembryonic endoderm cells, as they derive from early lineage branching events that are not
106 covered in this dataset. We first applied scVelo to the normalised and batch corrected count
107 matrix across all embryonic stages (Figure 1A). We observed that scVelo correctly identifies the

108 epiblast population as the origin of the global differentiation processes that occur during
109 gastrulation and early organogenesis. In relation to the more differentiated cell types however,
110 there were several instances where scVelo had difficulty in capturing some of the highly complex
111 differentiation events that occur across the entire embryo. For instance, scVelo predicted that
112 E8.0 allantois and mesenchyme cell-types give rise to mesodermal cells from earlier timepoints
113 rather than the E8.25/E8.5 allantoic and mesenchymal cells. Another inconsistency occurred with
114 E8.0-E8.25 endoderm cells, which were predicted to give rise to E6.5-E7 visceral endoderm,
115 rather than the other way round. Most noteworthy, scVelo failed to recapitulate the
116 erythropoiesis branch, where it predicts a backwards differentiation from later to earlier
117 populations. We next repeated this analysis using data from each individual time-point (Figure
118 1B; shown are E7.5 and E8.5). We saw that the pipeline accurately recapitulates known biological
119 trajectories up to E7.5, but observed the same inconsistency from E7.75 to E8.5, with scVelo
120 arrows pointing backwards.

121 Taken together therefore, we have identified that for erythroid development, the output of
122 scVelo is inconsistent with the timecourse information gathered from the experimental design of
123 the gastrulation atlas.

124

125 **Unspliced sequence reads help to discriminate between cell types**

126 We next asked whether this issue is due to a general lack of biologically meaningful information
127 captured in the unspliced reads.

128 To this end, we exploited two variance-based dimensionality reduction methods, Principal
129 Component Analysis (PCA) and Multi-Omics Factor Analysis (MOFA; 26), to interrogate how much
130 inter-population variability is explained by the spliced and unspliced information layers, whether
131 considered separately or together. Upon comparing PC1 and PC2 (or MOFA Factors 1 and 2), in
132 addition to the expected lineage separation obtained using the spliced reads (Figure 2A, left
133 panel), we could also observe a degree of lineage separation when using the unspliced reads
134 alone (Figure 2A, middle panel). In addition, we saw a qualitatively improved separation of the
135 different lineages when spliced and unspliced information is used in combination (Figure 2A, right
136 panel; see Supplementary Figure 1 for further components/factors). Moreover, the MOFA factors
137 account for 16% of variation in the spliced data and 4% of the of variation in unspliced data
138 (Figure 2Bi). Interestingly, a closer look at the MOFA pre-processing and final outcome showed a
139 minor overlap of genes that are highly variable with respect to spliced or unspliced counts (Figure
140 2Bii) and a different weight contributed by the two layers to the final factors (Figure 2Biii).

141 Multiomics factor analysis therefore not only demonstrates that the unspliced reads in the
142 gastrulation atlas dataset contain biologically relevant information, but also suggests that
143 integrated analysis of spliced and unspliced reads may more broadly facilitate the interpretation
144 of complex scRNA-Seq datasets.

145

146 **Analysis of unspliced reads reveals complex expression kinetics**

147 Having confirmed the utility of unspliced reads, we next explored whether the inability to recover
148 real-time progression in whole embryo trajectory inference using scVelo might be related to the

149 assumptions made by the current RNA velocity analysis tools. The derivation of gene-specific
150 expression kinetics underpins the scVelo analysis pipeline, as illustrated by so-called phase plots
151 that depict the amounts of spliced versus unspliced reads within a population of cells (14). If a
152 gene is upregulated during a differentiation timecourse, cells will be placed above the diagonal
153 between no expression and maximum expression due to the relatively larger amount of newly
154 produced pre-mRNA during the gene induction process, while the converse is true for
155 downregulated genes (Figure 3A). Both of these scenarios are readily captured by scVelo, with
156 the predicted vectors of differentiation agreeing with the actual temporal progression. If a given
157 gene however experiences an increase in transcription rate midway through a differentiation
158 timecourse, the sudden increase in unspliced pre-mRNA will result in a phase plot that may be
159 wrongly classified by scVelo, with predicted vectors of differentiation diametrically opposed to
160 the true direction of differentiation (Figure 3A). This is indeed what we observed when inspecting
161 the phase plots of the scVelo driver genes (top-likelyhood genes, Supplementary Table 1), which
162 display a steep increase of unspliced counts in the Erythroid 3 population, leading to a reverse
163 velocity prediction, progressing from Erythroid 3 to earlier populations (Supplementary Figure
164 2A).

165 We next set out to identify all genes exhibiting this rapid increase in expression levels in the
166 Erythroid 3 population (Figure 3B). After fitting a linear regression through each population and
167 each gene and testing whether the inferred slopes reflected the expected order based on
168 biological knowledge, we found 89 such genes, which we termed Multiple Rate Kinetics or MURK
169 genes. These genes included *Smim1*, coding for the Vel Blood Group Antigen (27), and *Hba-x*,
170 where we could confirm an increase in expression kinetics using phase plots (Figure 3C).

171 Having identified a set of genes with a coordinated increase in expression rate midway through
172 erythropoiesis, we next asked what function these genes might play in the broader
173 transcriptional program of red blood cell maturation. Visual inspection of the gene list revealed
174 it to contain archetypal red blood cell genes including the globin genes *Hba-x*, *Hbb-a1*, *Hba-a2*,
175 *Hbb-bt*, *Hbb-bh1*, *Hbb-y* (Supplementary table 2). Unsupervised gene ontology analysis
176 confirmed that biological functions essential for red blood cells were highly enriched, including
177 "gas transport" and "heme biosynthetic process" (Figure 3D).

178 We next removed this set of MURK genes and recalculated the RNA velocity inferred trajectories.
179 As can be seen in Figure 3E, inferred vectors of differentiation are now in good agreement with
180 the real time progression of erythropoiesis

181 The scVelo suite also calculates a so-called latent time, which represents the pseudotime
182 ordering hidden in the spliced and unspliced dynamics, and is more powerful than previously
183 described pseudotime inferring approaches since it incorporates both the gene dynamics and the
184 spliced and unspliced information (14). Using the full gene set, the latent time calculation for the
185 erythroid lineage is contrary to the know progression of erythroid differentiation (Figure 3E left
186 panels, Supplementary Figure 2B, left panels). By contrast, removing the MURK genes results in
187 a latent time prediction that is not only consistent with the major axis of erythropoiesis, but also
188 identifies the two sequential inputs described previously (25), namely an early wave directly from
189 posterior mesoderm as well as a second wave coming from yolk sac hemogenic endothelium (see
190 Figure 3E, Supplementary Figure 2B, right panels).

191 Taken together therefore, this analysis shows that inconsistent RNA velocity-inferred trajectories
192 can be remedied by the removal of genes with complex expression kinetics.

193

194 **Erythroid Multiple Rate Kinetics genes are essential for red blood cell function**

195 To corroborate upregulation of our identified MURK genes during erythropoiesis, we
196 interrogated a previously published dataset with transcriptomic analysis of a loss of function
197 model for the erythropoiesis master-regulator *Gata1* (28). *In vitro* differentiation of Gata1 knock-
198 out embryonic stem cells over-expressing human *BCL2* can produce permanently self-renewing
199 immature erythroid progenitor cell lines. One such model, G1ER, contains a tamoxifen-inducible
200 Gata1 transgene, the activation of which triggers erythroid maturation (29, 30; Figure 4A).
201 Microarray-based differential gene expression was performed, comparing the uninduced and
202 induced conditions (28). 76 of our 89 MURK genes overlapped with the genes identified by this
203 microarray-based comparison. Of those, 64 were upregulated, of which 55 showed strong
204 upregulation, 4 were downregulated, and 8 showed no change in expression following induction
205 of Gata1 in the G1ER system, demonstrating a highly significant overlap of our identified MURK
206 genes with the G1ER-induced genes ($p < 10^{-24}$; see Figure 4B).

207 Our newly identified erythropoietic MURK genes therefore perform key roles in red blood cell
208 function, and their upregulation was validated in an independent model of red blood cell
209 maturation.

210

211 **scRNA-Seq of mouse chimeras reveals the early cellular defects in Gata1 loss of function**

212 The G1ER cell line represents an *in vitro* model, and the published differential gene expression
213 data were from bulk microarray profiling, thus precluding any analysis of single-cell gene
214 expression kinetics. We therefore turned to our recently reported Chimaera-Seq approach,
215 whereby scRNA-Seq is coupled with mouse chimeric embryo technology, to define both cellular
216 and molecular consequences of gene knock-outs *in vivo* (25, 31). We used our standard
217 embryonic stem cells (ESCs) expressing a constitutive tdTomato (tdTom) fluorescent marker gene
218 to generate a Gata1 knock-out line (see Methods). *Gata1*$^{-}$ tdTom^{+} cells were injected into tdTom^{-}
219 wild-type blastocyst and transferred into pseudo-pregnant females, resulting in chimeric
220 embryos that we harvested at E8.5. Six chimeric embryos were pooled, dissociated into a single-
221 cell suspension, and tdTom^{+} and tdTom^{-} cell fractions were sorted for scRNA sequencing. We
222 obtained 8420 tdTom^{-} and 7944 tdTom^{+} cells passing quality control and assigned to a cell type,
223 with an average of 4354 genes being detected per cell.

224 We then concatenated the chimera data with the Pijuan-Sala et al. (2019) reference dataset and
225 mapped nearest neighbors (see Methods). We observed an overall homogeneous distribution of
226 both mutant and wild-type fractions throughout the later time-points of the landscape, except
227 for the erythroid branch. Indeed, we observed a block in the erythroid lineage of the mutant cells,
228 which were over-represented in the start of the erythroid differentiation branch, while their wild-
229 type counterparts were present throughout erythroid differentiation (Supplementary Figure 3).
230 Identification of the nearest neighbours of chimeric cells within the reference dataset allowed
231 their quick cell-type annotation, which we used to quantify the differences in the hemato-
232 endothelial cell-type representation within the chimera fractions. This analysis confirmed a

233 severe erythroid differentiation defect of the mutant cells (Figure 4C-E). When examining the
234 reference dataset sampled-time point of the chimera nearest neighbours we also observed a
235 temporal shift within the erythroid lineage, with tdTom$^+$ mutant cells mapping to earlier time-
236 points than their wildtype tdTom$^-$ counterparts, further confirming a developmental block of the
237 mutant cells (Figure 4D, E). In addition, we observed that this erythroid defect was coupled with
238 an over-representation of cells with a megakaryocyte signature (Figure 4C).

239 The newly generated *Gata1*$^-$ Chimaera-Seq data therefore not only recapitulated the expected
240 block in erythroid maturation, but also revealed an expansion of the megakaryocytic lineage in
241 the E8.5 yolk sac.

242

243 **The molecular program affected by Gata1 loss in early embryos**

244 Although the role of Gata1 is well documented in developmental erythropoiesis (21, 23), the early
245 molecular defects of Gata1 loss of function *in vivo* had not been reported. The Gata1 Chimaera-
246 Seq dataset therefore presented an opportunity to dissect the early molecular program
247 controlled by Gata1 *in vivo*. Having registered a defect in erythroid differentiation and an increase
248 in the megakaryocytic lineage population, we performed differential gene expression testing
249 between the chimera mutant and wild-type cells in these clusters (Supplementary Table 3).

250 Regarding the megakaryocytic subset, we observed upregulation of progenitor markers *Kit*,
251 *Gata2* and *Myb* in the *Gata1*$^-$ cells as well as lower expression of maturation genes for the
252 megakaryocyte lineage *Gp5*, *Pf4*, *Mpl* and *Plek* (Figure 5A). Hyper-proliferative megakaryocyte
253 progenitors, detected previously in *Gata1*$^-$ E12.5 fetal livers, led to compromised platelet

254 function, and were suggested to originate in the yolk sac (32). Our results showing over-
255 production of megakaryocytic cells with impaired maturation characteristics in E8.5 *Gata1*$^{-}$
256 chimera yolk sacs support this notion, and importantly place the megakaryocytic defect within
257 the very early phase of megakaryocyte formation.

258 Interestingly, all hemato-endothelial cell subsets displayed up-regulation of *Spi1* (coding for the
259 PU.1 transcription factor) in the *Gata1*$^{-}$ cell fraction compared to wild-type counterpart (FDR <
260 0.01; Figure 5A). Given the previously reported Gata1-PU.1 cross-repression in adult bone
261 marrow (18) and in zebrafish embryonic hematopoiesis (33), we systematically assessed the
262 effect of *Gata1* knockout in the mouse chimera lineages and observed that in *Gata1*$^{-}$ cells, *Spi1*
263 was specifically up-regulated in all hematopoietic sub-clusters, with a stronger effect on Mk and
264 Ery1 subsets. (Supplementary Figure 3).

265 In the early erythroid subset, Ery1, we again noted that the mutant cells displayed increased
266 expression of genes characteristic of a progenitor signature. Conversely, erythroid maturation
267 hallmark genes such as *Hbb-bs* and *Gypa* were downregulated, along with the erythroid Gata1
268 target *Mllt3* (34; Figure 5A). GO-term enrichment analysis of genes downregulated in *Gata1*$^{-}$ Ery1
269 cells revealed biological processes essential to red blood cell function (Figure 5B). Furthermore,
270 we also observed that 48% of the MURK genes identified in Figure 3 overlapped with these genes
271 that fail to up-regulate in *Gata1*$^{-}$ erythroid cells (Figure 5C; $p < 10^{-24}$).

272 In addition to the failure of inducing genes associated with erythroid maturation, single cell
273 resolution molecular analysis also revealed a striking failure to downregulate genes associated
274 with alternative lineage programs such as Pu.1, consistent with the notion that the earliest wave

275 of primitive hematopoiesis produces erythroid cells, megakaryocytes and macrophages, with
276 evidence for at least bipotential progenitor cells (35).

277

278 **The late erythroid increase in expression rate is downstream of Gata1 function**

279 Having generated the Chimaera-Seq single cell data for both wildtype and Gata1 knock-out cells,
280 we next used the ratio of spliced/unspliced reads to explore differences in expression kinetics
281 between the wildtype and mutant cells. As can be seen in Figure 5D, the previously defined MURK
282 genes failed to display the increased rate of expression characteristic for the later stages of
283 erythropoiesis in the mutant cells. The examples shown include the embryonic globin gene *Hbb-*
284 *y*, as well as the *Fam210b* gene, coding for a putative mitochondrial protein recently implicated
285 in erythroid differentiation (36; Figure 5D). This result confirms that the erythroid boost in
286 expression forms part of the transcriptional program downstream of Gata1 function, although it
287 does not demonstrate a direct regulatory role for Gata1.

288 However, preliminary modelling analysis suggests that the change observed in MURK gene
289 dynamics is due to altered transcription rates (see Supplementary Note), indicating a close
290 association of the coordinated late erythroid increase in transcription rate with the molecular
291 program downstream of Gata1.

292

293 **A coordinated increase of expression rate during human fetal liver erythropoiesis**

294 Having identified a coordinated increase in transcription rate during mouse yolk sac
295 erythropoiesis, we next wanted to ascertain whether the same phenomenon could also be seen
296 in human cells. Moreover, we were keen to explore an scRNA-Seq dataset generated by a
297 different laboratory, to exclude any potential technical bias caused by our own experimental
298 protocols. We therefore turned to a recently published comprehensive dataset of human fetal
299 liver erythropoiesis (37), and extracted the 49,388 cells annotated to the four clusters
300 encompassing human fetal liver erythropoiesis. When calculating scVelo-based differentiation
301 vectors as well as latent time using the full gene set (see methods), both were reversed (Figure
302 6A, left plots), consistent with the mouse yolk sac results. We therefore again ran our pipeline to
303 discover genes with a potential increase in expression rate along the differentiation pathway.
304 The resulting 97 genes again contained archetypal erythroid genes such as the haemoglobin
305 genes (Figure 6B), with overall gene ontologies demonstrating a functional role in erythropoiesis
306 (Figure 6C, see also Supplementary Table 4). We then recalculated both the scVelo differentiation
307 vectors as well as latent time after removing the fetal liver MURK genes. This revealed scVelo
308 vectors that were consistent with the expected developmental progression (see Figure 6A, right
309 plots). This analysis therefore demonstrates that complex expression kinetics apply broadly to
310 erythropoiesis, and their identification can be used to amend the RNA velocity framework to
311 prevent erroneous predictions.

312

313

314 **Discussion**

315 There is no doubt that single cell molecular profiling constitutes a transformative technology. It
316 suffers however from the major drawback that cells need to be fixed in order to profile them,
317 with the consequence that measurements are by necessity static snapshots. To decipher complex
318 biological processes, however, temporal information is commonly required. The single cell RNA
319 velocity concept raised the prospect of overcoming some of the limitations associated with static
320 measurements, by providing a strategy that can infer future cellular states. The RNA velocity
321 framework is based on an explicit model of transcriptional processes (transcription, splicing,
322 degradation). The notion that physical parameters of gene expression can be deduced from single
323 cell gene expression data had been explored before the single cell RNA velocity concept was
324 introduced (38, 39). However, the scVelo implementation provided an attractive framework for
325 estimating gene-specific expression parameters by taking advantage of the spliced versus
326 unspliced read counts across large cell populations (14). Using erythropoiesis as an example, we
327 show here that this current framework needs to be adapted to accommodate more complex
328 expression kinetics. Importantly, our analysis revealed that sets of genes can show a coordinated
329 increase in transcription rate along a differentiation pathway. Moreover, deletion of the key
330 erythroid regulator Gata1 abrogated this coordinated change in expression dynamics, thus
331 revealing this increase in transcription rate as an important feature of erythropoiesis. Of note,
332 current RNA velocity frameworks consider only a single reason for the presence of introns,
333 namely that a pre mRNA has not been fully processed. However, it is known that other processes
334 such as intron retention can result in the presence of intronic sequences in otherwise fully
335 processed cytoplasmic mRNA molecules (40, 41), thus suggesting that a more granular approach

336 towards both the modelling and experimental analysis of spliced versus unspliced reads
337 represents a promising avenue for future research.

338 Application of the single cell RNA velocity concept has commonly been “confirmatory”, whereby
339 a differentiation path proposed by other means was shown to be consistent with RNA velocity
340 inference. When we applied the RNA velocity framework to the entire mouse gastrulation atlas,
341 some inferred vectors of differentiation agreed with our current understanding of developmental
342 biology, but others disagreed. Deeper interrogation of predictions that conflicted with our
343 current understanding of erythropoiesis showed that the RNA velocity predictions could not be
344 correct, not only because they ran counter to the known expression changes that accompany red
345 blood cell differentiation, but also because they contradicted the real-time sampling of the data.
346 Our results thus highlight certain limitations of the current implementation of this framework for
347 identification of novel trajectories. Importantly however, it is through our observation of the
348 inconsistent predictions that we were led to identify the previously unrecognized dynamic nature
349 of the transcriptional control of erythropoiesis. Moreover, it is plausible that coordinated
350 increases in transcription rate midway through a differentiation process may operate more
351 widely, as a powerful mechanism for stabilising a cell state. Our extension to the scVelo
352 implementation reveals the presence of such time-dependent changes of gene expression
353 parameters and retrieves the concerned MURK genes in developmental trajectories of interest.

354 As to the precise mechanisms, at this stage we can only confidently assert that this process occurs
355 downstream of Gata1 during erythropoiesis. Of note, comprehensive analysis of the G1ER
356 erythroid differentiation model has shown that Gata1-induced maturation triggers increased
357 enhancer/promoter interactions for upregulated genes, and that the most highly enriched motif

358 in the promoters of these genes are GATA sites (42). These observations are therefore consistent
359 with the lineage-determining function of Gata1 involving a coordinated increase in expression
360 kinetics of a set of genes important for red blood cell function.

361 Our observations regarding the Gata1 knock-out phenotype also warrant some discussion. With
362 embryonically lethal phenotypes such as Gata1 knock-out, conventional analysis tends to be
363 somewhat limited, since the embryos are dead because they have no red blood cells. By contrast,
364 the Chimaera-Seq assay enables both quantification of cell numbers as well as characterisation
365 of their molecular profiles. Moreover, there are no secondary effects caused by the dying
366 embryo, because the wildtype host cells rescue overall fetal development, thus allowing a
367 focussed analysis of cell-intrinsic molecular defects. One noteworthy observation from our data
368 is that erythroid differentiation proceeds substantially beyond the stage where *Gata1* expression
369 itself is first initiated, but fails to proceed to the late erythroid phase where expression of
370 canonical red blood cell genes is greatly upregulated. However, gene expression prior to the
371 differentiation block is not normal. In particular, we observed increased Spi1/Pu.1 in the Gata1
372 knock-out cells, consistent with the previously reported (18) but also disputed (43) antagonistic
373 relationship between Gata1 and Pu.1.

374 Within haematopoiesis, Pu.1 is recognised as a key regulator of myeloid and T-cell lineages, but
375 not erythroid cells, even though a role in the proliferation of immature erythroid progenitors has
376 been reported (44, reviewed in 45). Upregulation of Pu.1 in our immature *Gata1* knock-out cells
377 therefore suggests that these cells of the primitive haematopoietic lineage represent progenitors
378 with multilineage potential, rather than being restricted to just the red cell lineage. Further
379 evidence for this notion is provided by our observation that the reduction in erythroid cells in the

380 *Gata1* knock-out is accompanied by an increase in megakaryocyte progenitors, consistent with a
381 model whereby Gata1 levels influence the lineage choice decisions of a multipotent progenitor
382 cell. Live cell tracking studies have suggested that the primary role of Gata1 and Pu.1 may be fate
383 stabilization rather than fate choice (43). The increase in transcription rate of erythroid genes
384 downstream of Gata1 would cohere with stabilizing the erythroid fate, thus suggesting that our
385 results are consistent with roles in both fate choice and fate stabilization.

386 Our observation of an expanded pool of megakaryocyte progenitors may also be of direct
387 relevance to our understanding of the pre-leukaemic transient myeloproliferative disease (TMD)
388 that is prevalent in newborns with trisomy 21 (46). TMD is thought to arise when a fetal specific
389 haematopoietic progenitor cell with trisomy 21 acquires a partial loss of function mutation in
390 *GATA1*, resulting in a short form of GATA1 (GATA1s). TMD is characterized by expansion of
391 immature megakaryocyte progenitors, and in 10 to 20% of cases transforms into malignant acute
392 megakaryoblastic leukaemia (reviewed in 47). Over-expression of GATA1s in mouse models
393 resulted in the identification of mid-gestation fetal liver megakaryocyte progenitors as uniquely
394 sensitive to this mutant GATA1s form compared to their adult bone marrow counterparts (48).
395 The over-represented population of immature megakaryocytic progenitors in our E8.5 *Gata1*$^{-}$
396 chimeras may correspond to the developmental emergence of this transient precursor, TMD-
397 initiating cell, in the yolk sac.

398

399

400 **Conclusions**

401 Taken together, this study reports how the RNA velocity framework can be extended to delve
402 into the transcriptional mechanisms of tissue differentiation, complemented with single cell
403 resolution and *in vivo* analysis of Gata1 function, which revealed a number of previously unknown
404 facets of this canonical regulator of red blood cell development.

405

406

407 **Methods**

408 **scVelo implementation**

409 **Mouse atlas dataset**. To obtain separated count matrices for spliced and unspliced mRNAs, we
410 ran velocyto 0.17.17 (10) on the .bam files from the mouse atlas in Pijuan-Sala et al. 2019 (25;
411 GEO accession number: GSE87038). We kept all cells that passed the QC as described in the
412 original publication, but filtered out from downstream analysis the extraembryonic tissues: ExE
413 endoderm, ExE ectoderm and Parietal endoderm as well as samples with no timepoint allocation
414 (labelled as 'mixed gastrulation'). To select highly variable genes (HVGs) we applied both the
415 scanpy v1.5.1 and the scVelo v0.2.1 (14) pipelines. That is, we removed genes with less than 20
416 shared counts between spliced and unspliced counts, before normalising and log transforming
417 the remaining genes. Then, we selected the top 2500 HVGs from each approach (resulting in a
418 total of 4000, with 1000 overlapping genes) for further calculation of moments; while performing
419 imputation using the top 30 nearest neighbours from the graph connectivities generated with
420 the original UMAP coordinates from Pijuan-Sala et al. 2019. The velocity vectors were computed
421 in dynamical mode rather than steady state.

422 **Human dataset.** We first downloaded raw reads from Popescu et al., 2019 (37; GEO accession
423 number: GSE127980), and aligned them against the human genome hg19-3.0.0 with CellRanger
424 v3.0.2 to generate the .bam files and obtain separated count matrices for spliced and unspliced
425 mRNAs as described above. We filtered out cells with less than 3,550 counts, less than 900 genes
426 and more than 6% mitochondrial counts. Again, we combined scapy and scVelo's pipelines to
427 select 1,500 HVGs to compute PCA coordinates and applied batch correction using the function
428 reducedMNN from the batchelor package v1.4.0 (49), followed by the estimation of velocity
429 vectors in the same way it was done for the mouse dataset.

430

431 **MOFA+ implementation**

432 We ran MOFA+ v1.4.0 (26) using as input the two single cell experiment objects obtained from
433 the spliced and unspliced counts independently. Each object was created in R using the scran
434 v1.16.0 (50) library as follows: we started from the raw counts, normalized them with factor sizes
435 obtained after pre-clustering, log transformed and reduced to 5000 HVG. We then switched to
436 Python v3.7.4, where we regressed out the sample effect and scaled the object to generate a
437 MOFA+ model with standard parameters. Finally, we used reducedMNN to correct the MOFA
438 Factors for batch effects. The same objects used as MOFA input were used for PCA calculation in
439 Figure 2A.

440

441 **MURK genes identification**

442 To identify MURK genes, we considered the imputed counts resulting from the scVelo standard
443 pipeline. Then, for each gene and each population among the Erythroid lineage, we calculated
444 the unspliced versus spliced slope with a linear regression, as well as the standard error on the
445 slope. In the mouse dataset we selected all genes for which the slope in Erythroid3 is significantly
446 higher than the slope in Erythoid2 (according to a one-sided t-test p-value < 0.05), the average
447 spliced counts in Erythroid3 is higher than the average spliced counts in every other population,
448 and the slope in Erythroid3 positive. We found 89 genes that respect all these criteria.

449 In the human dataset, in order to obtain erythroid populations more comparable to our mouse
450 data, we re-clustered the erythroid clusters (Figure 6A). We retained the population annotations
451 from the original paper except for the Late Erythroid population, which we defined after
452 performing Leiden clustering on the Umap coordinates. Specifically, we re-allocated a subset of
453 the previously annotated Mid Erythroid population to Late Erythroid, in such a way that they
454 have a similar numbers of cells. We then calculated the unspliced versus spliced slope with linear
455 regression and identified MURK genes where the slope in Late Erythroid is significantly higher
456 than the slope in Mid Erythroid. We found 97 genes respecting these criteria.

457

458 **Gene ontology enrichment analysis**

459 We performed gene ontology enrichment analysis using the http://geneontology.org website
460 comparing the MURK genes against all biological processes, with the default all Mus musculus
461 genes in database as background set (51, 52). We ranked the processes by FDR.

462

463 **Overlap testing**

464 Overlap was tested with Fisher exact test. We calculated the probability of having m = 55 genes
465 of our n = 89 MURK genes mapping to the A = 1022 high response genes (out of N = 4195 genes)
466 in the Wu et al., 2011 publication (GEO accession number: GSE30142) as the probability of
467 randomly picking m elements of a specific type when randomly choosing n elements out of N,
468 where the frequency of the special type is A/N.

469

470 **Gata1^{-} chimera dataset generation and analysis**

471 **Embryo collection.** All procedures were performed in strict accordance to the UK Home Office
472 regulations for animal research under the project license number PPL 70/8406. **Chimaera**
473 **generation.** TdTomato-expressing mouse embryonic stem cells (ESC) were derived as previously
474 described (25). Briefly, ESC lines were derived from E3.5 blastocysts obtained by crossing a male
475 ROSA26tdTomato (Jax Labs – 007905) with a wildtype C57BL/6 female, expanded under the
476 2i+LIF conditions (53) and transiently transfected with a Cre-IRES-GFP plasmid (54) using
477 Lipofectamine 3000 Transfection Reagent (ThermoFisher Scientific, #L3000008) according to
478 manufacturer's instructions. A tdTomato-positive, male, karyotypically normal line, competent
479 for chimaera generation as assessed using morula aggregation assay, was selected for targeting
480 *Gata1*. Two guides were designed using the http://crispr.mit.edu tool (guide 1:
481 CGGCTACTCCACTGTGGCGG; guide 2: CGCTTCTTGGGCCGGATGAG) and were cloned into the
482 pX458 plasmid (Addgene, #48138) as previously described (55). The obtained plasmids were then
483 used to transfect the cells and single transfected clones were expanded and assessed for Cas9-

484 induced mutations. Genomic DNA was isolated by incubating cell pellets in 0.1 mg/ml of
485 Proteinase K (Sigma, #03115828001) in TE buffer at 50°C for 2 hours, followed by 5 min at 99°C.
486 The sequence flanking the guide-targeted sites was amplified from the genomic DNA by
487 polymerase chain reaction (PCR) in a Biometra T3000 Thermocycler (30 sec at 98°C ; 30 cycles of
488 10 sec at 98°C, 20 sec at 58°C, 20 sec at 72°C; and elongation for 7 min at 72°C) using the Phusion
489 High-Fidelity DNA Polymerase (NEB, #M0530S) according to the manufacturer's instructions.
490 Primers including Nextera overhangs were used (F-
491 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTACCCTGCCTCAACTGTG; R-
492 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTTGTCTTGGGCAGGAACA), allowing library
493 preparation with the Nextera XT Kit (Illumina, #15052163), and sequencing was performed using
494 the Illumina MiSeq system according to manufacturer's instructions. An ESC clone showing a 38
495 base-pair frameshift mutation in exon 4 resulting in the functional inactivation of *Gata1* were
496 selected for injection into C57BL/6 E3.5 blastocysts. A total of 6 chimaeric embryos were
497 harvested at E8.5, dissected, and single-cell suspensions were generated by TrypLE Express
498 dissociation reagent (Thermo Fisher Scientific) incubation for 7-10 minutes at 37°C under
499 agitation. Single-cell suspensions were sorted into tdTom+ and tdTom- samples using a BD Influx
500 sorter with DAPI at 1µg/ml (Sigma) as a viability stain for subsequent 10X scRNA-seq library
501 preparation (version 3 chemistry), and sequencing using an S1 flow cell in the Illumina Novaseq
502 platform, which resulted in 8420 $tdTom^-$ and 7944 $tdTom^+$ cells that passed quality control (see
503 "Single-cell RNA sequencing analysis" below).

504 **Single-cell RNA sequencing analysis.** Raw files were processed with Cell Ranger 3.0.2 using
505 default mapping arguments. Reads were mapped to the mm10 genome and counted with

506 GRCm38.92 annotation, including tdTomato sequence for chimera cells. Cell barcodes with
507 expression profiles significantly different to the ambient mRNA expression profile were identified
508 using emptyDrops (56), and cell barcodes with low complexity, i.e. low total mRNA counts and/or
509 high mitochondrial proportion, were identified by fitting four-component bivariate mixture
510 models to the $\log_{10}$-transformed total mRNA counts and percentage of mitochondrial counts, and
511 selecting the components with high total mRNA and low mitochondrial percentage. Gene
512 expression normalization and doublet cell barcodes were identified using the approach taken by
513 Pijuan-Sala et al. (2019). Both spliced and unspliced count matrices were extracted using velocyto
514 0.17.17 (10).

515 **Mapping to the reference dataset.** We mapped the chimaera cells to the mouse atlas following
516 almost exactly the procedure used in the original publication article to map the *Tal1* chimaera.
517 First, we concatenated the mouse atlas and chimaera counts (both previously controlled for
518 quality of the cells), normalized the resulting counts matrix with scran, computed HVGs and then
519 applied multiBatchPCA, and reducedMNN with cosine normalization from batchelor (49) for
520 batch effect correction within samples (where sample refers to a single lane of a 10x Chromium
521 chip) as well as between datasets in order to extract a number of nearest neighbours between
522 the mouse atlas and the chimaera using queryKNN from BiocNeighbors package v1.6.0.

523 **Differential Gene Expression Analysis.** For differential gene expression analysis, we took samples
524 that included at least 7 cells per tdTom status per cell population (eg. Erythroid3). We ran the
525 analysis in scanpy v1.5.1 (57) with Wilcoxon test and choosing 2 as fold change and 0.1 as false
526 discovery rate thresholds.

527

528 **Funding**

529 Research in the authors' laboratories is supported by the Wellcome Trust, MRC, CRUK, Blood
530 Cancer UK, NIH-NIDDK, the Sanger-EBI Single Cell Centre; by core support grants by the Wellcome
531 Trust to the Cambridge Institute for Medical Research and Wellcome Trust-MRC Cambridge Stem
532 Cell Institute; and by core funding from Cancer Research UK and the European Molecular Biology
533 Laboratory. C.G. was funded by the Swedish Research Council (2017-06278), I.I. was funded by a
534 British Heart Foundation studentship (FS/18/56/35177), S.G. was supported by a Royal Society
535 Newton International Fellowship (NIF\R1\181950). This work was funded as part of a Wellcome
536 Strategic Award (105031/D/14/Z) awarded to Wolf Reik, Berthold Göttgens, John Marioni,
537 Jennifer Nichols, Ludovic Vallier, Shankar Srinivas, Benjamin Simons, Sarah Teichmann, and
538 Thierry Voet.

539

540 **Authors' contributions**

541 M.B. performed scVelo implementations in mouse and human datasets, mathematical modelling,
542 and analysis of Gata1 embryonic chimera dataset; I.I-R. assisted on the scVelo implementation in
543 mouse datasets; I.I. performed Gata1 CRISPR/Cas9 targeting and expansion of the resulting
544 mutant lines; S.G. performed quality controls of the Gata1 embryonic chimera dataset; S.G. and
545 C.G. performed initial analysis of the Gata1 embryonic chimera dataset; C.G. designed and
546 optimized the mutant chimera single-cell profiling experiments; B.G. wrote the initial draft of the

547 manuscript; M.B., C.G., J.C.M. edited the manuscript; J.N., J.C.M., C.G. and B.G. supervised the
548 study. All authors read and approved the final manuscript.

549 **Acknowledgements**

550 We would like to thank Prof. Fabian Theis and Volker Bergen for discussions and valuable input
551 on the scVelo implementation. We thank Prof. Ross Hardison for providing the list of Gata1-
552 regulated genes from Wu et al. 2011 (Figure 4B). We thank William Mansfield and the Gurdon
553 Institute animal facility for blastocyst injections, the Flow Cytometry Core Facility at the
554 Cambridge Institute for Medical Research for cell sorting, Katarzyna Kania and the CRUK-CI
555 genomics core for preparing the 10X libraries and for sequencing.

556

557

558 **References**

559 1. Akunuru S, Geiger H. Aging, Clonality, and Rejuvenation of Hematopoietic Stem Cells.
560 Trends Mol Med. 2016;22(8):701-12.
561 2. Schultz MB, Sinclair DA. When stem cells grow old: phenotypes and mechanisms of stem
562 cell aging. Development. 2016;143(1):3-14.
563 3. Mahadevaiah SK, Sangrithi MN, Hirota T, Turner JMA. A single-cell transcriptome atlas of
564 marsupial embryogenesis and X inactivation. Nature. 2020;586(7830):612-7.

565 4. Gerber T, Murawala P, Knapp D, Masselink W, Schuez M, Hermann S, et al. Single-cell
566 analysis uncovers convergence of cell identities during axolotl limb regeneration. Science.
567 2018;362(6413).
568 5. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell
569 mapping of gene expression landscapes and lineage in the zebrafish embryo. Science.
570 2018;360(6392):981-7.
571 6. Ton MN, Guibentif C, Gottgens B. Single cell genomics and developmental biology:
572 moving beyond the generation of cell type catalogues. Curr Opin Genet Dev. 2020;64:66-71.
573 7. Borrett MJ, Innes BT, Jeong D, Tahmasian N, Storer MA, Bader GD, et al. Single-Cell
574 Profiling Shows Murine Forebrain Neural Stem Cells Reacquire a Developmental State when
575 Activated for Adult Neurogenesis. Cell Rep. 2020;32(6):108022.
576 8. Weinreb C, Rodriguez-Fraticelli A, Camargo FD, Klein AM. Lineage tracing on
577 transcriptional landscapes links state to fate during differentiation. Science. 2020;367(6479).
578 9. Dahlin JS, Hamey FK, Pijuan-Sala B, Shepherd M, Lau WWY, Nestorowa S, et al. A single-
579 cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice.
580 Blood. 2018;131(21):e1-e11.
581 10. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity
582 of single cells. Nature. 2018;560(7719):494-8.
583 11. Zhang Q, He Y, Luo N, Patel SJ, Han Y, Gao R, et al. Landscape and Dynamics of Single
584 Immune Cells in Hepatocellular Carcinoma. Cell. 2019;179(4):829-45 e20.

585 12. Zhou W, Yui MA, Williams BA, Yun J, Wold BJ, Cai L, et al. Single-Cell Analysis Reveals
586 Regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T Cell
587 Development. Cell Syst. 2019;9(4):321-37 e9.
588 13. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchis-Calleja F, et al. Organoid single-
589 cell genomic atlas uncovers human-specific features of brain development. Nature.
590 2019;574(7778):418-22.
591 14. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell
592 states through dynamical modeling. Nat Biotechnol. 2020.
593 15. Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level
594 expression of the human beta-globin gene in transgenic mice. Cell. 1987;51(6):975-85.
595 16. Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J, Pretorius IM, et al. A major positive
596 regulatory region located far upstream of the human alpha-globin gene locus. Genes Dev.
597 1990;4(9):1588-601.
598 17. Mettananda S, Gibbons RJ, Higgs DR. Understanding alpha-globin gene regulation and
599 implications for the treatment of beta-thalassemia. Ann N Y Acad Sci. 2016;1368(1):16-24.
600 18. Zhang P, Behre G, Pan J, Iwama A, Wara-Aswapati N, Radomska HS, et al. Negative
601 cross-talk between hematopoietic regulators: GATA proteins repress PU.1. Proc Natl Acad Sci U
602 S A. 1999;96(15):8705-10.
603 19. McGrath K, Palis J. Ontogeny of erythropoiesis in the mammalian embryo. Curr Top Dev
604 Biol. 2008;82:1-22.

605 20. Pevny L, Simon MC, Robertson E, Klein WH, Tsai SF, D'Agati V, et al. Erythroid
606 differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription
607 factor GATA-1. Nature. 1991;349(6306):257-60.
608 21. Pevny L, Lin CS, D'Agati V, Simon MC, Orkin SH, Costantini F. Development of
609 hematopoietic cells lacking transcription factor GATA-1. Development. 1995;121(1):163-72.
610 22. Gutierrez L, Tsukamoto S, Suzuki M, Yamamoto-Mukai H, Yamamoto M, Philipsen S, et
611 al. Ablation of Gata1 in adult mice results in aplastic crisis, revealing its essential role in steady-
612 state and stress erythropoiesis. Blood. 2008;111(8):4375-85.
613 23. Fujiwara Y, Browne CP, Cunniff K, Goff SC, Orkin SH. Arrested development of embryonic
614 red cell precursors in mouse embryos lacking transcription factor GATA-1. Proc Natl Acad Sci U
615 S A. 1996;93(22):12355-8.
616 24. Shivdasani RA, Fujiwara Y, McDevitt MA, Orkin SH. A lineage-selective knockout
617 establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet
618 development. EMBO J. 1997;16(13):3965-73.
619 25. Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, et al. A
620 single-cell molecular map of mouse gastrulation and early organogenesis. Nature.
621 2019;566(7745):490-5.
622 26. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a
623 statistical framework for comprehensive integration of multi-modal single-cell data. Genome
624 Biol. 2020;21(1):111.

625 27. Storry JR, Joud M, Christophersen MK, Thuresson B, Akerstrom B, Sojka BN, et al.
626 Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. Nat
627 Genet. 2013;45(5):537-41.
628 28. Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, et al. Dynamics of the epigenetic
629 landscape during erythroid differentiation after GATA1 restoration. Genome Res.
630 2011;21(10):1659-71.
631 29. Tsang AP, Visvader JE, Turner CA, Fujiwara Y, Yu C, Weiss MJ, et al. FOG, a multitype zinc
632 finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and
633 megakaryocytic differentiation. Cell. 1997;90(1):109-19.
634 30. Weiss MJ, Yu C, Orkin SH. Erythroid-cell-specific properties of transcription factor GATA-
635 1 revealed by phenotypic rescue of a gene-targeted cell line. Mol Cell Biol. 1997;17(3):1642-51.
636 31. Guibentif C, Griffiths JA, Imaz-Rosshandler I, Ghazanfar S, Nichols J, Wilson V, et al.
637 Diverse Routes toward Early Somites in the Mouse Embryo. Dev Cell. 2020.
638 32. Vyas P, Ault K, Jackson CW, Orkin SH, Shivdasani RA. Consequences of GATA-1 deficiency
639 in megakaryocytes and platelets. Blood. 1999;93(9):2867-75.
640 33. Monteiro R, Pouget C, Patient R. The gata1/pu.1 lineage fate paradigm varies between
641 blood populations and is modulated by tif1gamma. EMBO J. 2011;30(6):1093-103.
642 34. Pina C, May G, Soneji S, Hong D, Enver T. MLLT3 regulates early human erythroid and
643 megakaryocytic cell fate. Cell Stem Cell. 2008;2(3):264-73.
644 35. Palis J. Hematopoietic stem cell-independent hematopoiesis: emergence of erythroid,
645 megakaryocyte, and myeloid potential in the mammalian embryo. FEBS Lett.
646 2016;590(22):3965-74.

647 36. Kondo A, Fujiwara T, Okitsu Y, Fukuhara N, Onishi Y, Nakamura Y, et al. Identification of
648 a novel putative mitochondrial protein FAM210B associated with erythroid differentiation. Int J
649 Hematol. 2016;103(4):387-95.
650 37. Popescu DM, Botting RA, Stephenson E, Green K, Webb S, Jardine L, et al. Decoding
651 human fetal liver haematopoiesis. Nature. 2019;574(7778):365-71.
652 38. Ezer D, Moignard V, Gottgens B, Adryan B. Determining Physical Mechanisms of Gene
653 Expression Regulation from Single Cell Gene Expression Data. PLoS Comput Biol.
654 2016;12(8):e1005072.
655 39. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell
656 RNA-sequencing data. Genome Biol. 2013;14(1):R7.
657 40. Edwards CR, Ritchie W, Wong JJ, Schmitz U, Middleton R, An X, et al. A dynamic intron
658 retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood.
659 2016;127(17):e24-e34.
660 41. Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. A dynamic intron
661 retention program enriched in RNA processing genes regulates gene expression during terminal
662 erythropoiesis. Nucleic Acids Res. 2016;44(2):838-51.
663 42. Liu X, Chen Y, Zhang Y, Liu Y, Liu N, Botten GA, et al. Multiplexed capture of spatial
664 configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9.
665 Genome Biol. 2020;21(1):59.
666 43. Hoppe PS, Schwarzfischer M, Loeffler D, Kokkaliaris KD, Hilsenbeck O, Moritz N, et al.
667 Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. Nature.
668 2016;535(7611):299-302.

669 44. Choe KS, Ujhelly O, Wontakal SN, Skoultchi AI. PU.1 directly regulates cdk6 gene
670 expression, linking the cell proliferation and differentiation programs in erythroid cells. J Biol
671 Chem. 2010;285(5):3044-52.
672 45. Carotta S, Wu L, Nutt SL. Surprising new roles for PU.1 in the adaptive immune
673 response. Immunol Rev. 2010;238(1):63-75.
674 46. Roberts I, Alford K, Hall G, Juban G, Richmond H, Norton A, et al. GATA1-mutant clones
675 are frequent and often unsuspected in babies with Down syndrome: identification of a
676 population at risk of leukemia. Blood. 2013;122(24):3908-17.
677 47. Bhatnagar N, Nizery L, Tunstall O, Vyas P, Roberts I. Transient Abnormal Myelopoiesis
678 and AML in Down Syndrome: an Update. Curr Hematol Malig Rep. 2016;11(5):333-41.
679 48. Li Z, Godinho FJ, Klusmann JH, Garriga-Canut M, Yu C, Orkin SH. Developmental stage-
680 selective effect of somatically mutated leukemogenic transcription factor GATA1. Nat Genet.
681 2005;37(6):613-9.
682 49. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-
683 sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol.
684 2018;36(5):421-7.
685 50. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of
686 single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5:2122.
687 51. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology:
688 tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
689 52. The Gene Ontology C. The Gene Ontology Resource: 20 years and still GOing strong.
690 Nucleic Acids Res. 2019;47(D1):D330-D8.

691 53. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, et al. The ground state
692 of embryonic stem cell self-renewal. Nature. 2008;453(7194):519-23.
693 54. Wray J, Kalkan T, Gomez-Lopez S, Eckardt D, Cook A, Kemler R, et al. Inhibition of
694 glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases
695 embryonic stem cell resistance to differentiation. Nat Cell Biol. 2011;13(7):838-45.
696 55. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the
697 CRISPR-Cas9 system. Nat Protoc. 2013;8(11):2281-308.
698 56. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell
699 Atlas J, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell
700 RNA sequencing data. Genome Biol. 2019;20(1):63.
701 57. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data
702 analysis. Genome Biol. 2018;19(1):15.

703

704

705 **Figure Legends**

706 **Figure 1. Inferring Differentiation Trajectories at organismal scale**

707 A. Pijuan-Sala et al. (2019) layout containing single-cell transcriptomes belonging from E6.5
708 to E8.5, colored by sampled time-point (left) and by cell-type (right). The overlaying
709 arrows result from applying the scVelo pipeline to the whole embryonic dataset and
710 represent inferred developmental trajectories. Arrowheads highlight the erythroid

711 branch, displaying scVelo trajectory predictions that are inconsistent with real-time
712 sampling.
713 B. Pijuan-Sala et al. (2019) layout highlighting single-cell transcriptomes belonging to E7.5
714 (left) and E8.5 (right) and colored by cell-type (see legend in A). The overlaying arrows
715 result from applying the scVelo pipeline to these individual time-points and represent
716 inferred developmental trajectories. Arrowheads highlight the erythroid branch.

717 **Figure 2. Unspliced counts contribute to explaining the variability among cell types**

718 A. Dimensionality reduction with the first two principal components/MOFA factors using
719 spliced reads alone (left), unspliced reads alone (middle) and both spliced and unspliced
720 (right). Single-cell transcriptomes are colored by cell-type annotation; see Figure 1 for full
721 legend.
722 B. MOFA characterization of spliced and unspliced reads assessing proportion of variance
723 explained (i), overlap in highly variable genes calculating using either spliced or unspliced
724 reads (ii), and factor weight distributions (iii).

725 **Figure 3. A set of genes with complex expression kinetics confounds velocity estimation in**
726 **erythropoiesis**

727 A. Illustration of phase plot representation in datasets of differentiating cell populations,
728 and associated scVelo predictions
729 B. Illustration of strategy for MURK gene identification
730 C. Phase plots of representative MURK genes. X-axis: normalized imputed counts of spliced
731 transcript; y-axis: normalized imputed counts of unspliced transcript.

732 D. GO-term enrichment of MURK genes identified in mouse yolk sac erythropoiesis

733 E. Zoomed-in UMAP of the erythroid branch (see Figure 1 for full UMAP) with scVelo
734 calculations, before and after removing MURK genes identified in B. Distinct waves of
735 embryonic erythropoiesis are visible upon MURK gene removal, highlighted with
736 arrowheads.

737 **Figure 4. *In vivo* analysis of Gata1 function using a chimaera assay coupled with scRNA-Seq**

738 A. Schematic of the G1ER system (29, 30)

739 B. Behaviour of the 89 MURK genes identified in Figure 3 upon Gata1 induction in the G1ER
740 system (28). Wu et al. report that upon Gata1 induction they obtained a total of 2769
741 upregulated genes, 6079 mildly upregulated, 3566 downregulated, and 3445 with no
742 response.

743 C. UMAPS of *Gata1*$^{-}$ chimera cells allocated a hemato-endothelial identity colored by cell-
744 type (sub-clusters defined in Pijuan-Sala et al. (2019) - BP: Blood Progenitors, EC:
745 Endothelial Cells, Haem: Hemato-endothelial Progenitors, Mk: Megakaryocytes, My:
746 Myeloid cells, Ery: Erythroid cells) and split by genotype. Orange arrowheads highlight
747 increased population with megakaryocytic signature in Gata1^{-} fraction.

748 D. UMAPS of *Gata1*$^{-}$ chimera cells allocated a hemato-endothelial identity colored by
749 sampling timepoint and split by genotype.

750 E. Barplots with the quantification of chimera cells mapping to each hemato-endothelial
751 lineage of the reference dataset (left) and to sampled time-points of the reference dataset
752 (right).

753 **Figure 5. Gata1 chimaera assay reveals disruption of MURK genes and perturbed yolk sac**
754 **hematopoiesis**

755 A. Violin plots of representative genes differentially regulated in *Gata1*$^{-}$ hematopoietic
756 lineages.
757 B. GO-term enrichment of genes downregulated in *Gata1*$^{-}$ Ery1 cells compared to their WT
758 counterparts in chimeras.
759 C. Venn diagram showing overlap between MURK genes and genes downregulated in *Gata1*$^{-}$
760 Ery1 cells
761 D. Phase plots of MURK genes identified along erythroid differentiation, in E8.5 *Gata1*$^{-}$
762 chimera datasets, colored by tdTom status.

763

764 **Figure 6. Concept of dual kinetics of gene expression is also revealed in human foetal liver**
765 **hematopoiesis**

766 A. UMAP representation of human fetal liver erythroid cell populations. The overlaying
767 arrows result from applying the scVelo pipeline using all genes (left) or after MURK gene
768 exclusion (right). Bottom UMAPs are colored by corresponding scVelo-inferred latent
769 time. In order to facilitate comparison with the mouse data, a new clustering was
770 performed on the erythroid cells, see Methods. MEMP: megakaryocyte-erythroid-mast
771 cell progenitor.
772 B. Phase plots of representative MURK genes identified in human fetal liver erythropoiesis
773 single-cell RNAseq dataset.

774 C. GO-term enrichment of MURK genes identified in human fetal liver erythropoiesis.

775

776 **Supplementary Figures**

777 1. Dimensionality reduction with the first three principal components/MOFA factors using
778 spliced reads alone (left), unspliced reads alone (middle) and both spliced and unspliced
779 (right). Single-cell transcriptomes are colored by cell-type annotation; see Figure 1 for full
780 legend.
781 2. Identification of MURK genes along yolk sac erythropoiesis. A. Phase plots of
782 representative scVelo driver genes, with scVelo model prediction overlayed (see also
783 Supplementary Table 1). B. Distribution of annotated cell type (top) and sampling time-
784 point (bottom) along scVelo calculated latent time, using all genes (left panels) and after
785 removing the MURK genes identified in Figure 3B-C.
786 3. Pijuan-Sala et al. (2019) layout highlighting nearest neighbours of *Gata1*$^{-}$ chimeras. In red
787 are nearest neighbours of tdTom+ mutant cells, in black those of tdTom- wildtype cells.
788 To compare with Figure 1A.
789 4. Impact of Gata1 knockout on *Spi1*/PU.1 expression on the hematoendothelial cell types.
790 X-axis: *Spi1* $\log_2$(fold-change) in *Gata1*$^{-}$ vs WT chimera cells and Atlas nearest neighbours.
791 Y-axis: $\log_{10}$(FDR).

792

793 **Supplementary Tables**

794 1. Driver genes of the scVelo perdictions along erythroid differentiation, ranked by
795 likelihood in the dynamic model.
796 2. List of mouse MURK genes identified in Figure 3B-C, ranked by calculated increase in slope
797 value.
798 3. Differential Expression Analysis of Gata1^{-} tdTom^{+} vs WT tdTom^{-} chimera cells. For the Mk
799 subset, given the low numbers of WT chimera cells present, the nearest neighbors from
800 the reference Atlas dataset were included in the comparison. LFC: log fold change.
801 4. List of human MURK genes identified in Figure 6, ranked by calculated increase in slope
802 value.

Figure 1

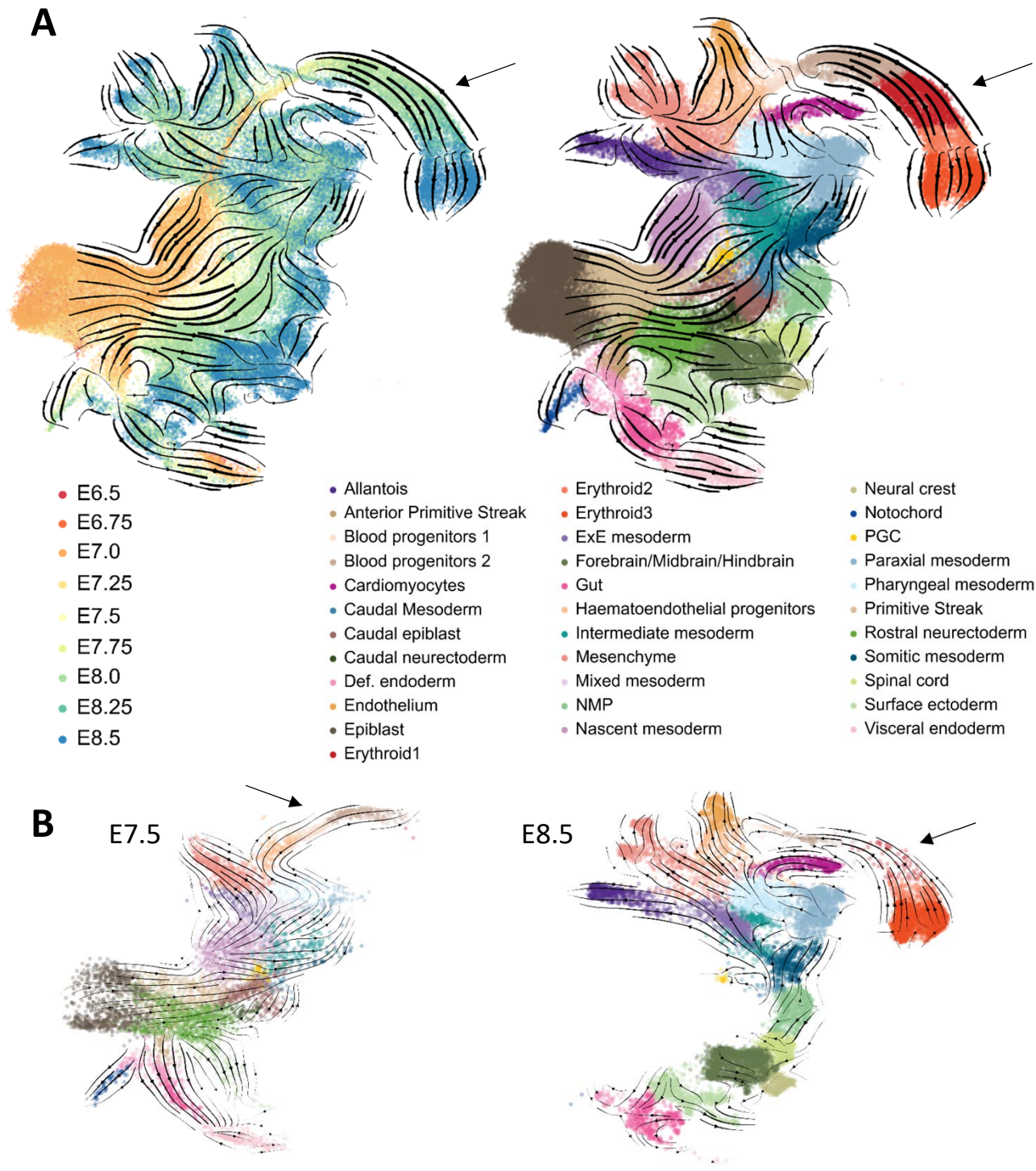

A
E6.5
E6.75
E7.0
E7.25
E7.5
E7.75
E8.0
E8.25
E8.5
Allantois
Anterior Primitive Streak
Blood progenitors 1
Blood progenitors 2
Cardiomyocytes
Caudal Mesoderm
Caudal epiblast
Caudal neurectoderm
Def. endoderm
Endothelium
Epiblast
Erythroid1
Erythroid2
Erythroid3
ExE mesoderm
Forebrain/Midbrain/Hindbrain
Gut
Haematoendothelial progenitors
Intermediate mesoderm
Mesenchyme
Mixed mesoderm
NMP
Nascent mesoderm
Neural crest
Notochord
PGC
Paraxial mesoderm
Pharyngeal mesoderm
Primitive Streak
Rostral neurectoderm
Somitic mesoderm
Spinal cord
Surface ectoderm
Visceral endoderm
B
E7.5
E8.5

Figure 2

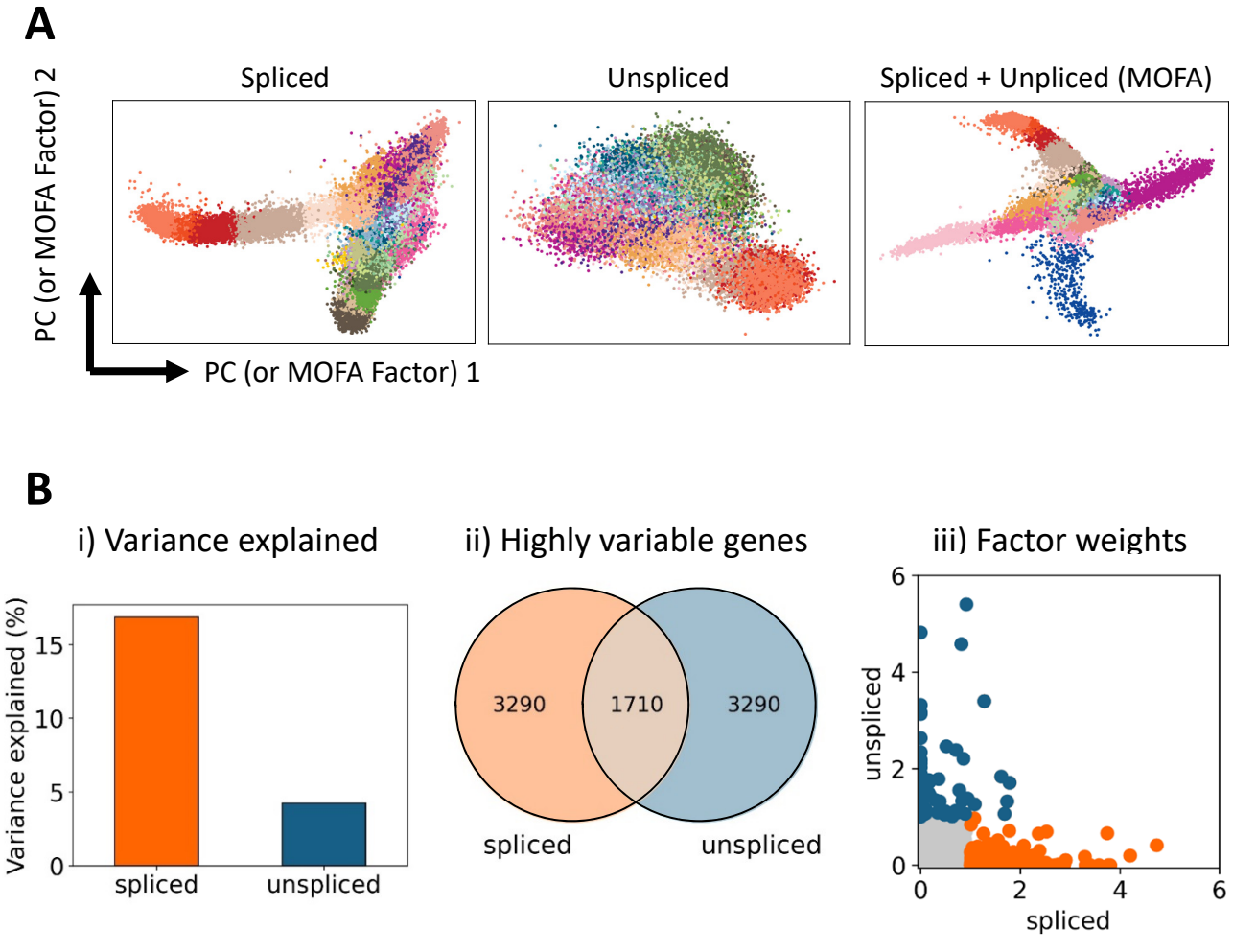

A
Spliced
Unspliced
Spliced + Unpliced (MOFA)
PC (or MOFA Factor) 2
PC (or MOFA Factor) 1
B
i) Variance explained
Variance explained (%)
15
10
5
0
spliced
unspliced
ii) Highly variable genes
3290
1710
3290
spliced
unspliced
iii) Factor weights
unspliced
spliced
0
2
4
6

Figure 3

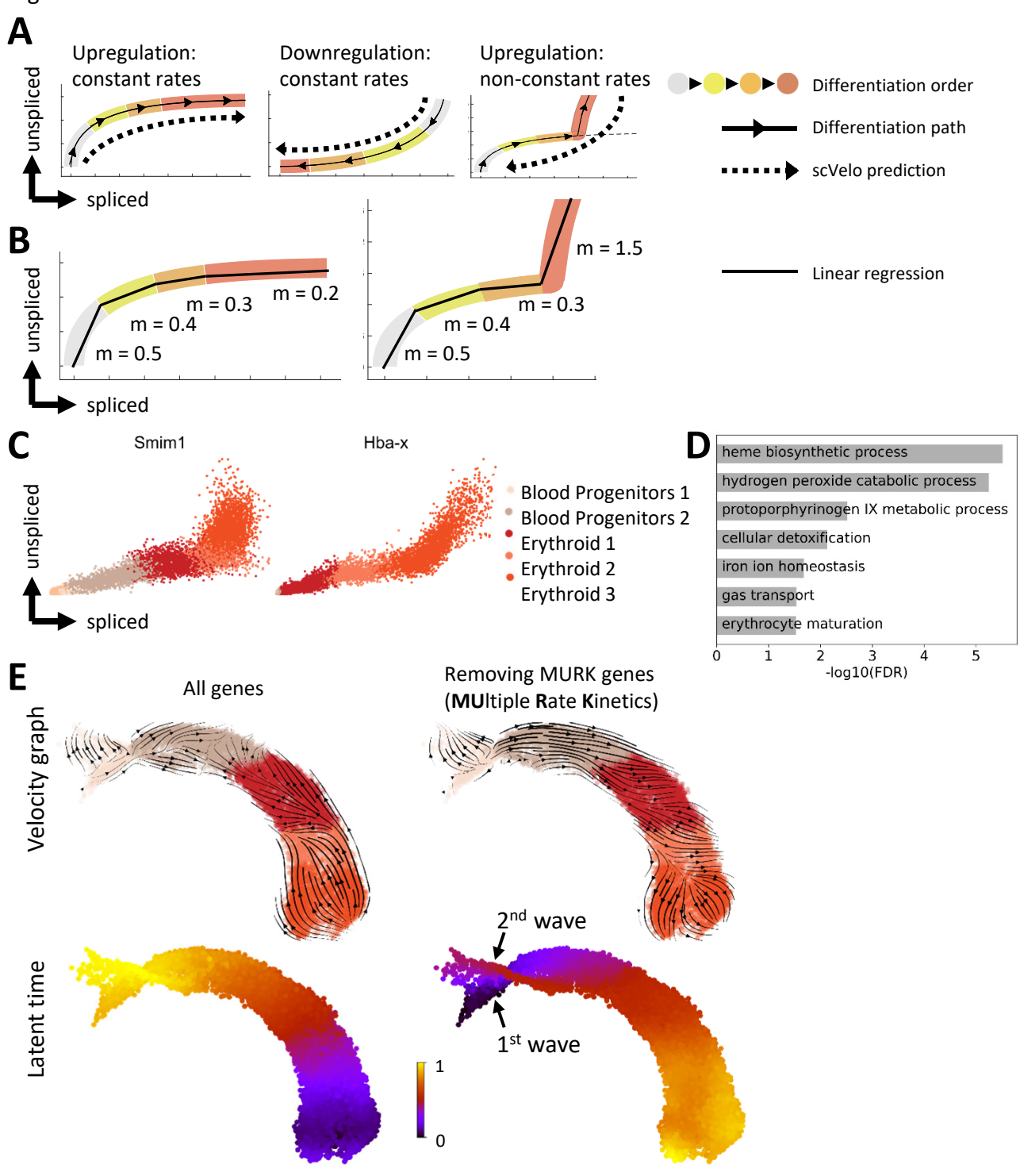

A
Upregulation: constant rates
Downregulation: constant rates
Upregulation: non-constant rates
unspliced
spliced
Differentiation order
Differentiation path
scVelo prediction
B
m = 0.5
m = 0.4
m = 0.3
m = 0.2
m = 1.5
Linear regression
C
Smim1
Hba-x
Blood Progenitors 1
Blood Progenitors 2
Erythroid 1
Erythroid 2
Erythroid 3
D
heme biosynthetic process
hydrogen peroxide catabolic process
protoporphyrinogen IX metabolic process
cellular detoxification
iron ion homeostasis
gas transport
erythrocyte maturation
-log10(FDR)
E
All genes
Removing MURK genes (MUltiple Rate Kinetics)
Velocity graph
Latent time
2nd wave
1st wave

Figure 4

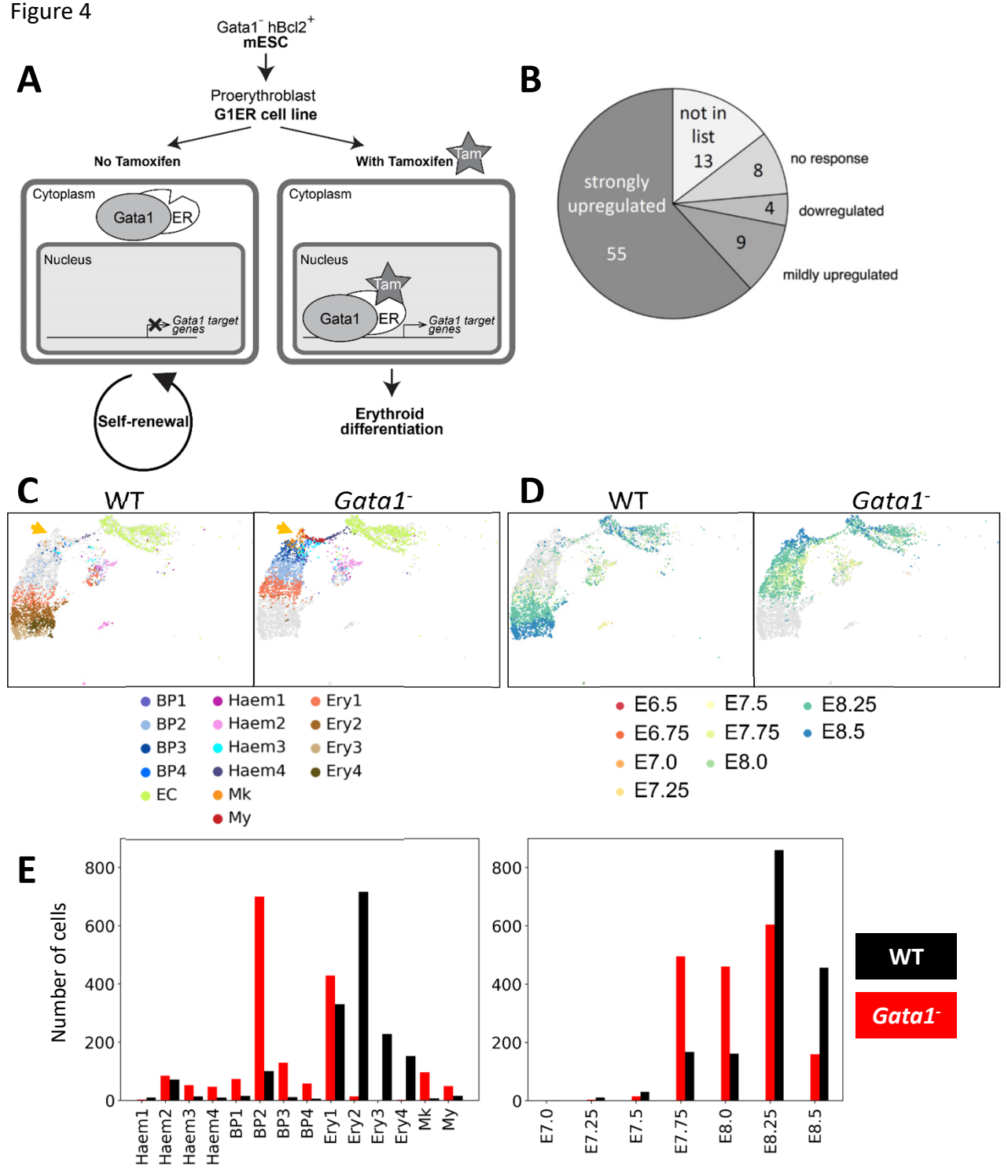

A
Gata1⁻ hBcl2⁺
mESC
Proerythroblast
G1ER cell line
No Tamoxifen
With Tamoxifen
Tam
Cytoplasm
Gata1
ER
Nucleus
Gata1 target genes
Self-renewal
Erythroid differentiation
B
not in list
13
8
4
9
55
strongly upregulated
no response
dowregulated
mildly upregulated
C
WT
Gata1⁻
BP1
BP2
BP3
BP4
EC
Haem1
Haem2
Haem3
Haem4
Mk
My
Ery1
Ery2
Ery3
Ery4
D
WT
Gata1⁻
E6.5
E6.75
E7.0
E7.25
E7.5
E7.75
E8.0
E8.25
E8.5
E
Number of cells
0
200
400
600
800
Haem1
Haem2
Haem3
Haem4
BP1
BP2
BP3
BP4
Ery1
Ery2
Ery3
Ery4
Mk
My
E7.0
E7.25
E7.5
E7.75
E8.0
E8.25
E8.5
WT
Gata1⁻

Figure 5

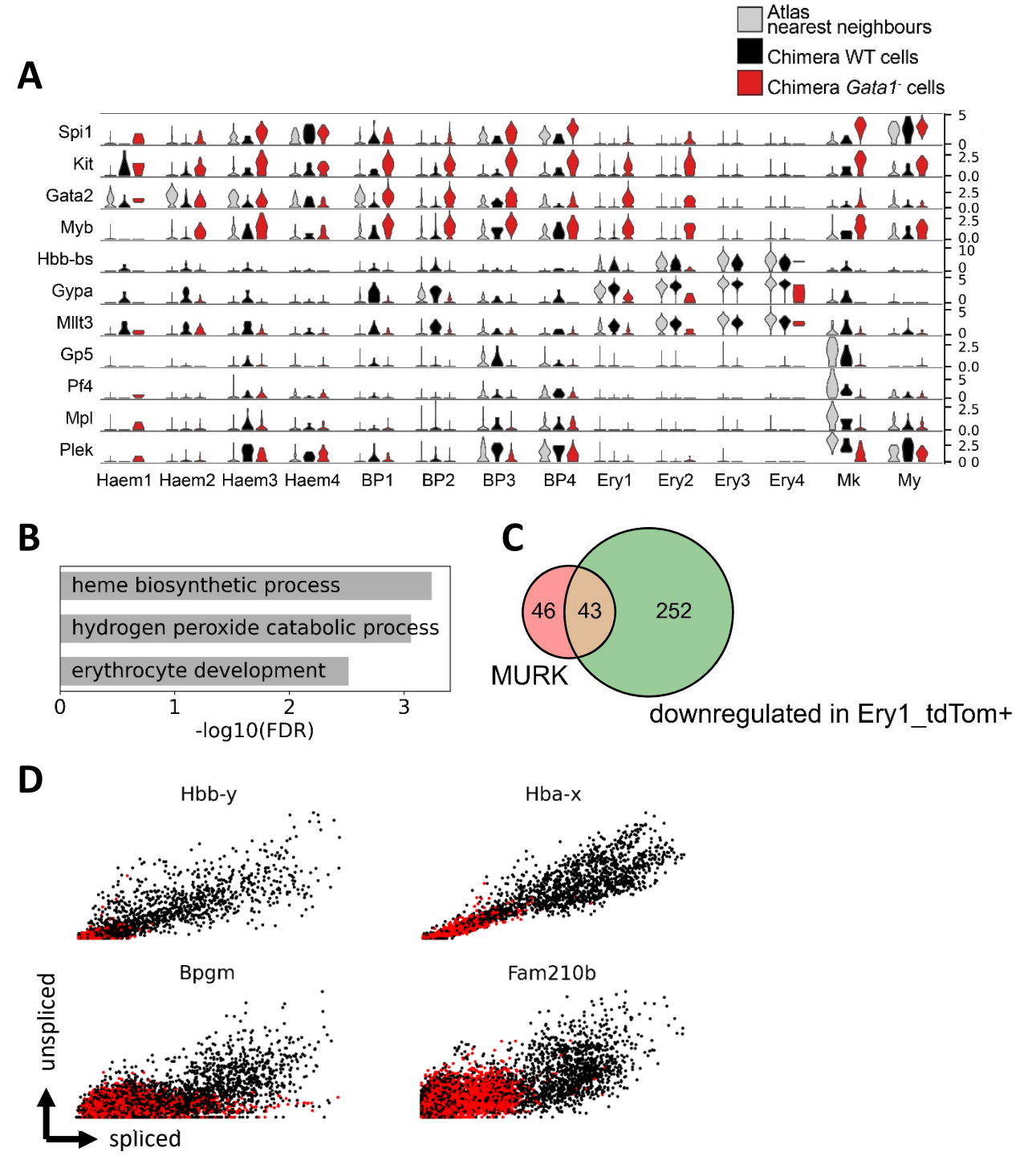
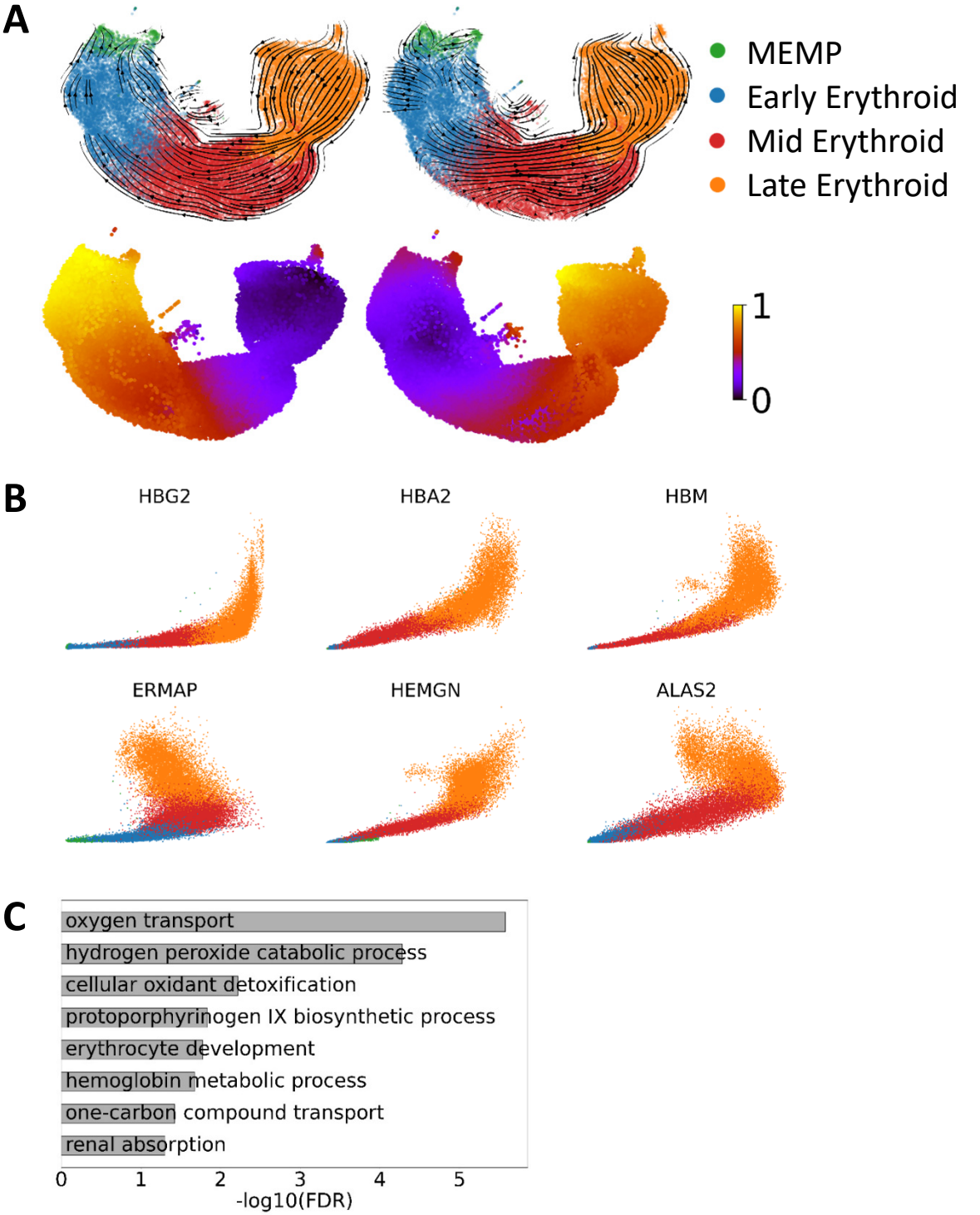

A
Atlas nearest neighbours
Chimera WT cells
Chimera Gata1- cells
Spi1
Kit
Gata2
Myb
Hbb-bs
Gypa
Mllt3
Gp5
Pf4
Mpl
Plek
Haem1
Haem2
Haem3
Haem4
BP1
BP2
BP3
BP4
Ery1
Ery2
Ery3
Ery4
Mk
My
B
heme biosynthetic process
hydrogen peroxide catabolic process
erythrocyte development
0
1
2
3
-log10(FDR)
C
46
43
252
MURK
downregulated in Ery1_tdTom+
D
Hbb-y
Hba-x
Bpgm
Fam210b
unspliced
spliced

Figure 6


A
MEMP
Early Erythroid
Mid Erythroid
Late Erythroid
1
0
B
HBG2
HBA2
HBM
ERMAP
HEMGN
ALAS2
C
oxygen transport
hydrogen peroxide catabolic process
cellular oxidant detoxification
protoporphyrinogen IX biosynthetic process
erythrocyte development
hemoglobin metabolic process
one-carbon compound transport
renal absorption
0
1
2
3
4
5
-log10(FDR)