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

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. 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. 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.

(244 words)

Introduction

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 (Akunuru and Geiger, 2016; Schultz and Sinclair, 2016). 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 (Gerber et al., 2018; Mahadevaiah et al., 2020; Wagner et al., 2018 reviewed

in Ton et al., 2020)), as well as large-scale datasets covering adult tissue homeostasis (Borrett et al., 2020; Dahlin et al., 2018; Weinreb et al., 2020).

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 (La Manno et al., 2018), which demonstrated that unspliced pre-mRNA present in scRNA-Seq datasets can be exploited to predict likely future expression states. This so-called RNA velocity concept is based on the notion that the ratio between unspliced and spliced RNA differs depending on whether a gene is in the process of being up- or downregulated. During upregulation, there is a relative increase in newly transcribed unspliced RNA, with the converse occurring during downregulation. The RNA velocity framework has rapidly gained traction across the wider single cell community, being applied across multiple experimental systems (Kanton et al., 2019; Zhang et al., 2019; Zhou et al., 2019), and also extended as part of the scVelo analysis suite (Bergen et al., 2020), which allows inclusion of genes whose transcript levels are not in steady state.

One system where the RNA velocity concept has particular potential is erythropoiesis, the process whereby oxygen-transporting red blood cells are generated from multipotent haematopoietic progenitors. Research into the transcriptional control processes of erythropoiesis led to several paradigmatic discoveries, including the dissection of distal transcriptional control elements (Grosveld et al., 1987; Higgs et al., 1990; Mettananda et al., 2016), as well as antagonistic transcription factor pairings as executors of lineage choice in multipotent progenitors (Zhang et al., 1999). During embryogenesis, a first so-called primitive wave of erythropoiesis occurs in the yolk sac, followed by a second definitive wave, initiated also in the yolk sac, then predominantly in the fetal liver and later in the adult bone marrow (McGrath and Palis, 2008). The zinc finger protein Gata1 represents the archetypal erythroid transcription factor, and is required for the maturation of both primitive and definitive erythroid cells (Fujiwara et al., 1996; Gutierrez et al., 2008; Pevny et al., 1995; Pevny et al., 1991), as well as megakaryocyte maturation (Shivdasani et al., 1997). However, the precise molecular processes affected by Gata1 deletion in early embryonic erythropoiesis have remained obscure, principally because conventional biochemical methods are unsuitable for the very small number of cells present at these early developmental stages.

Here, we have applied RNA velocity to a recently published scRNA-Seq dataset of nine sequential timepoints, spaced 6 hours apart, which encompass mouse gastrulation and early organogenesis (Pijuan-Sala et al., 2019). We observed that some of the inferred trajectories are incompatible with the existing biological knowledge, as well as the real time ordering derived from the sequential sampling timepoints. For erythroid differentiation in particular, we show that failure of the Velocity framework is due to a concerted increase in transcription rate of a subset of erythroid genes, midway through the red blood cell maturation trajectory. Analysis of *Gata1*$^{-}$ chimeric embryos underscores the concerted nature of this expression boost, consistent with the notion that such concerted upregulation events may be a feature of stabilizing a given differentiated cellular state.

Results

Limitations of RNA velocity trajectory inference at organismal scale

To evaluate RNA velocity-based trajectory inference with a complex dataset, we applied the scVelo analysis pipeline (Bergen et al., 2020) to a recently reported timecourse scRNA-Seq dataset covering mouse gastrulation and early organogenesis. This mouse gastrulation atlas contains approximately 120,000 single cell transcriptomes across nine sequential timepoints covering 37 major cell types (Pijuan-Sala et al., 2019). Prior to scVelo analysis, we removed extraembryonic ectoderm and extraembryonic endoderm cells, as they derive from early lineage branching events that are not covered in this dataset. We first applied scVelo to the normalised and batch corrected count matrix across all embryonic stages (Figure 1A). We observed that scVelo correctly identifies the epiblast population as the origin of the global differentiation processes that occur during gastrulation and early organogenesis. In relation to the more differentiated cell types however, there were several instances where scVelo had difficulty in capturing some of the highly complex differentiation events that occur across the entire embryo. For instance, scVelo predicted that E8.0 allantois and mesenchyme cell-types give rise to mesodermal cells from earlier timepoints rather than the E8.25/E8.5 allantoic and mesenchymal cells. Another inconsistency occurred with E8.0-E8.25 endoderm cells, which were predicted to give rise to E6.5-E7 visceral endoderm, rather than the other way round. Most noteworthy, scVelo failed to recapitulate the erythropoiesis branch, where it predicts a backwards differentiation from later to earlier populations. We next repeated this analysis using data from each individual time-point (Figure 1B; shown are E7.5 and E8.5). We saw that the pipeline accurately

recapitulates known biological trajectories up to E7.5, but observed the same inconsistency from E7.75 to E8.5, with scVelo arrows pointing backwards.

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

Unspliced sequence reads help to discriminate between cell types

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

To this end, we exploited two variance-based dimensionality reduction methods, Principal Component Analysis (PCA) and Multi-Omics Factor Analysis (MOFA; Argelaguet et al., 2020), to interrogate how much inter-population variability is explained by the spliced and unspliced information layers, whether considered separately or together. Upon comparing PC1 and PC2 (or MOFA Factors 1 and 2), in addition to the expected lineage separation obtained using the spliced reads (Figure 2A, left panel), we could also observe a degree of lineage separation when using the unspliced reads alone (Figure 2A, middle panel). In addition, we saw a qualitatively improved separation of the different lineages when spliced and unspliced information is used in combination (Figure 2A, right panel; see Supplementary Figure 1 for further components/factors). Moreover, the MOFA factors account for 16% of variation in the spliced data and 4% of the of variation in unspliced data (Figure 2Bi). Interestingly, a closer look at the MOFA pre-processing and final outcome showed a minor overlap of genes that are highly variable with respect to spliced or unspliced counts (Figure 2Bii) and a different weight contributed by the two layers to the final factors (Figure 2Biii).

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

Analysis of unspliced reads reveals complex expression kinetics

Having confirmed the utility of unspliced reads, we next explored whether the inability to recover real-time progression in whole embryo trajectory inference using scVelo might be related to the assumptions made by the current RNA velocity analysis tools. The derivation of gene-specific

expression kinetics underpins the scVelo analysis pipeline, as illustrated by so-called phase plots that depict the amounts of spliced versus unspliced reads within a population of cells (Bergen et al., 2020). If a gene is upregulated during a differentiation timecourse, cells will be placed above the diagonal between no expression and maximum expression due to the relatively larger amount of newly produced pre-mRNA during the gene induction process, while the converse is true for downregulated genes (Figure 3A). Both of these scenarios are readily captured by scVelo, with the predicted vectors of differentiation agreeing with the actual temporal progression. If a given gene however experiences an increase in transcription rate midway through a differentiation timecourse, the sudden increase in unspliced pre-mRNA will result in a phase plot that may be wrongly classified by scVelo, with predicted vectors of differentiation diametrically opposed to the true direction of differentiation (Figure 3A). This is indeed what we observed when inspecting the phase plots of the scVelo driver genes (top-likelyhood genes, Supplementary Table 1), which display a steep increase of unspliced counts in the Erythroid 3 population, leading to a reverse velocity prediction, progressing from Erythroid 3 to earlier populations (Supplementary Figure 2A).

We next set out to identify all genes exhibiting this rapid increase in expression levels in the Erythroid 3 population (Figure 3B). After fitting a linear regression through each population and each gene and testing whether the inferred slopes reflected the expected order based on biological knowledge, we found 73 such genes, which we termed Multiple Rate Kinetics or MURK genes. These genes included *Smim1* and *Hba-x*, where we could confirm an increase in expression kinetics using phase plots (Figure 3C).

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

We next removed this set of MURK genes and recalculated the RNA velocity inferred trajectories. As can be seen in Figure 3E, inferred vectors of differentiation are now in good agreement with the real time progression of erythropoiesis, with the only discrepancies occurring towards the end, presumably due to the fact there are no future expression states to project into beyond E8.5, the last timepoint sampled in the analysed dataset.

The scVelo suite also calculates a so-called latent time, which represents the pseudotime ordering hidden in the spliced and unspliced dynamics, and is more powerful than previously described pseudotime inferring approaches since it incorporates both the gene dynamics and the spliced and unspliced information (Bergen et al., 2020). Using the full gene set, the latent time calculation for the erythroid lineage is contrary to the know progression of erythroid differentiation (Figure 3E left panels, Supplementary Figure 2C, left panels). By contrast, removing the MURK genes results in a latent time prediction that is not only consistent with the major axis of erythropoiesis, but also identifies the two sequential inputs described previously (Pijuan-Sala et al., 2019), namely an early wave directly from posterior mesoderm as well as a second wave coming from yolk sac hemogenic endothelium (see Figure 3E, Supplementary Figure 2C, right panels).

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

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

To corroborate upregulation of our identified MURK genes during erythropoiesis, we interrogated a previously published dataset with transcriptomic analysis of a loss of function model for the erythropoiesis master-regulator *Gata1* (Wu et al., 2011). *In vitro* differentiation of Gata1 knock-out embryonic stem cells over-expressing human *BCL2* can produce permanently self-renewing immature erythroid progenitor cell lines. One such model, G1ER, contains a tamoxifen-inducible Gata1 transgene, the activation of which triggers erythroid maturation (Tsang et al., 1997; Weiss et al., 1997; Figure 4A). Microarray-based differential gene expression was performed, comparing the uninduced and induced conditions (Wu et al., 2011). 70 of our 73 MURK genes overlapped with the genes identified by this microarray-based comparison. Of those, 52 were upregulated, of which 45 showed strong upregulation, 4 were downregulated, and 4 showed no change in expression following induction of Gata1 in the G1ER system, demonstrating a highly significant overlap of our identified MURK genes with the G1ER-induced genes ($p < 10^{-11}$; see Figure 4B).

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

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

The G1ER cell line represents an *in vitro* model, and the published differential gene expression data were from bulk microarray profiling, thus precluding any analysis of single-cell gene expression kinetics. We therefore turned to our recently reported Chimaera-Seq approach, whereby scRNA-Seq is coupled with mouse chimeric embryo technology, to define both cellular and molecular consequences of gene knock-outs *in vivo* (Pijuan-Sala et al., 2019). We used our standard embryonic stem cells (ESCs) expressing a constitutive tdTomato (tdTom) fluorescent marker gene to generate a Gata1 knock-out line (see Methods). *Gata1*$^{-}$ tdTom^{+} cells were injected into tdTom^{-} wild-type blastocyst and transferred into pseudo-pregnant females, resulting in chimeric embryos that we harvested at E8.5. Six chimeric embryos were pooled, dissociated into a single-cell suspension, and tdTom^{+} and tdTom^{-} cell fractions were sorted for scRNA sequencing. We obtained 8420 tdTom^{-} and 7944 tdTom^{+} cells passing quality control and assigned to a cell type, with an average of 4354 genes being detected per cell.

We then concatenated the chimera data with the Pijuan-Sala et al. (2019) reference dataset and mapped nearest neighbors (see Methods). We observed an overall homogeneous distribution of both mutant and wild-type fractions throughout the later time-points of the landscape, except for the erythroid branch. Indeed, we observed a block in the erythroid lineage of the mutant cells, which were over-represented in the start of the erythroid differentiation branch, while their wild-type counterparts were present throughout erythroid differentiation (Supplementary Figure 3). Identification of the nearest neighbours of chimeric cells within the reference dataset allowed their quick cell-type annotation, which we used to quantify the differences in the hemato-endothelial cell-type representation within the chimera fractions. This analysis confirmed a severe erythroid differentiation defect of the mutant cells (Figure 4C-E). When examining the reference dataset sampled-time point of the chimera nearest neighbours we also observed a temporal shift within the erythroid lineage, with tdTom^{+} mutant cells mapping to earlier time-points than their wildtype tdTom^{-} counterparts, further confirming a developmental block of the mutant cells (Figure 4D, E). In addition, we observed that this erythroid defect was coupled with an over-representation of cells with a megakaryocyte signature (Figure 4C).

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

The molecular program affected by Gata1 loss in early embryos

Although the role of Gata1 is well documented in developmental erythropoiesis (Fujiwara et al., 1996; Pevny et al., 1995), the early molecular defects of Gata1 loss of function *in vivo* had never been reported. The Gata1 Chimaera-Seq dataset therefore presented an opportunity to dissect the early molecular program controlled by Gata1 *in vivo*. Having registered a defect in erythroid differentiation and an increase in the megakaryocytic lineage population, we performed differential gene expression testing between the chimera mutant and wild-type cells in these clusters (Supplementary Table 3).

Regarding the megakaryocytic subset, we observed upregulation of progenitor markers *Kit*, *Gata2* and *Myb* in the *Gata1*$^{-}$ cells as well as lower expression of maturation genes for the megakaryocyte lineage *Gp5*, *Pf4*, *Mpl* and *Plek* (Figure 5A). Hyper-proliferative megakaryocyte progenitors, detected previously in *Gata1*$^{-}$ E12.5 fetal livers, led to compromised platelet function, and were suggested to originate in the yolk sac (Vyas et al., 1999). Our results showing over-production of megakaryocytic cells with impaired maturation characteristics in E8.5 *Gata1*$^{-}$ chimera yolk sacs support this notion, and importantly place the megakaryocytic defect within the very early phase of megakaryocyte formation.

Interestingly, all hemato-endothelial cell subsets displayed up-regulation of *Spi1* (coding for the PU.1 transcription factor) in the *Gata1*$^{-}$ cell fraction compared to wild-type counterpart (FDR < 0.01; Figure 5A). Given the previously reported Gata1-PU.1 cross-repression in adult bone marrow (Zhang et al., 1999) and in zebrafish embryonic hematopoiesis (Monteiro et al., 2011), we systematically assessed the effect of *Gata1* knockout in the mouse chimera lineages and observed that in *Gata1*$^{-}$ cells, *Spi1* was specifically up-regulated in all hematopoietic sub-clusters, with a stronger effect on Mk and Ery1 subsets. (Supplementary Figure 3).

In the early erythroid subset, Ery1, we again noted that the mutant cells displayed increased expression of genes characteristic of a progenitor signature. Conversely, erythroid maturation hallmark genes such as *Hbb-bs* and *Gypa* were downregulated, along with the erythroid Gata1 target *Mllt3* (Pina et al., 2008; Figure 5A). GO-term enrichment analysis of genes downregulated in *Gata1*$^{-}$ Ery1 cells revealed biological processes essential to red blood cell function (Figure 5B). Furthermore, we also observed that 37% of the MURK genes identified in Figure 3 overlapped with these genes that fail to up-regulate in *Gata1*$^{-}$ erythroid cells (Figure 5C; $p < 10^{-13}$).

In addition to the failure of inducing genes associated with erythroid maturation, single cell resolution molecular analysis also revealed a striking failure to downregulate genes associated with alternative lineage programs such as Pu.1, consistent with the notion that the earliest wave of primitive hematopoiesis produces erythroid cells, megakaryocytes and macrophages, with evidence for at least bipotential progenitor cells (Palis, 2016).

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

Having generated the Chimaera-Seq single cell data for both wildtype and Gata1 knock-out cells, we next used the ratio of spliced/unspliced reads to explore differences in expression kinetics between the wildtype and mutant cells. As can be seen in Figure 5D, the previously defined MURK genes failed to display the increased rate of expression characteristic for the later stages of erythropoiesis in the mutant cells. The examples shown include the embryonic globin gene *Hbb-y*, as well as the *Smim1* gene coding for the Vel Blood Group Antigen (Storry et al., 2013) and the *Fam210b* gene, coding for a putative mitochondrial protein recently implicated in erythroid differentiation (Kondo et al., 2016; Figure 5D). This result confirms that the erythroid boost in expression forms part of the transcriptional program downstream of Gata1 function, although it does not demonstrate a direct regulatory role for Gata1. However, preliminary modelling analysis suggests that the change observed in MURK gene dynamics is due to altered transcription rates (see Supplementary Note), indicating a close association of the coordinated late erythroid increase in transcription rate with the molecular program downstream of Gata1.

A coordinated increase of expression rate during human fetal liver erythropoiesis

Having identified a coordinated increase in transcription rate during mouse yolk sac erythropoiesis, we next wanted to ascertain whether the same phenomenon could also be seen in human cells. Moreover, we were keen to explore an scRNA-Seq dataset generated by a different laboratory, to exclude any potential technical bias caused by our own experimental protocols. We therefore turned to a recently published comprehensive dataset of human fetal liver erythropoiesis (Popescu et al., 2019), and extracted the 49,388 cells annotated to the four clusters encompassing human fetal liver erythropoiesis. When calculating scVelo-based differentiation vectors as well as latent time using the full gene set (see methods), both were reversed (Figure 6A, left plots), consistent with the mouse yolk sac results. We therefore again ran our pipeline to discover genes with a potential increase in expression rate along the differentiation pathway. The resulting 377 genes again contained archetypal erythroid genes such as the hemoglobins (Figure 6B), with overall gene ontologies demonstrating a functional role in erythropoiesis (see Figure 6C); with in addition a high prevalence of cell-cycle related terms. We then recalculated both the scVelo differentiation vectors as well as latent time after removing the fetal liver MURK genes. This revealed scVelo vectors that were consistent with the expected developmental progression (see Figure 6A, right plots). This analysis therefore demonstrates

that complex expression kinetics apply broadly to erythropoiesis, and their identification can be used to amend the RNA velocity framework to prevent erroneous predictions.

Discussion

There is no doubt that single cell molecular profiling constitutes a transformative technology. It suffers however from the major drawback that cells need to be fixed in order to profile them, with the consequence that measurements are by necessity static snapshots. To decipher complex biological processes, however, temporal information is commonly required. The single cell RNA velocity concept raised the prospect of overcoming some of the limitations associated with static measurements, by providing a strategy that can infer future cellular states. The RNA velocity framework is based on an explicit model of transcriptional processes (transcription, splicing, degradation). The notion that physical parameters of gene expression can be deduced from single cell gene expression data had been explored before the single cell RNA velocity concept was introduced (Ezer et al., 2016; Kim and Marioni, 2013). However, the scVelo implementation provided an attractive framework for estimating gene-specific expression parameters by taking advantage of the spliced versus unspliced read counts across large cell populations (Bergen et al., 2020). Using erythropoiesis as an example, we show here that this current framework needs to be adapted to accommodate more complex expression kinetics. Importantly, our analysis revealed that sets of genes can show a coordinated increase in transcription rate along a differentiation pathway. Moreover, deletion of the key erythroid regulator Gata1 abrogated this coordinated change in expression dynamics, thus revealing this increase in transcription rate as an important feature of erythropoiesis. It remains to be seen how dynamic changes in expression kinetics may play a role in other systems, where splicing and degradation as well as transcription rates may also be time-dependent.

Application of the single cell RNA velocity concept has commonly been “confirmatory”, whereby a differentiation path proposed by other means was shown to be consistent with RNA velocity inference. When we applied the RNA velocity framework to the entire mouse gastrulation atlas, some inferred vectors of differentiation agreed with our current understanding of developmental biology, but others disagreed. Deeper interrogation of predictions that conflicted with our current understanding of erythropoiesis showed that the RNA velocity predictions could not be correct, not only because they ran counter to the known expression changes that accompany red blood cell differentiation, but also because they contradicted the real-time sampling of the data. Our results thus highlight certain limitations of the current implementation of this framework for identification of novel trajectories. Importantly however, it is through our observation of the inconsistent predictions

that we were led to identify the previously unrecognized dynamic nature of thetranscriptional control of erythropoiesis. Moreover, it is plausible that coordinated increases in transcription rate midway through a differentiation process may operate more widely, as a powerful mechanism for stabilising a cell state. Our extension to the scVelo implementation reveals the presence of such time-dependent changes of gene expression parameters and retrieves the concerned MURK genes in developmental trajectories of interest.

As to the precise mechanisms, at this stage we can only confidently assert that this process occurs downstream of Gata1 during erythropoiesis. Of note, comprehensive analysis of the G1ER erythroid differentiation model has shown that Gata1-induced maturation triggers increased enhancer/promoter interactions for upregulated genes, and that the most highly enriched motif in the promoters of these genes are GATA sites (Liu et al., 2020). These observations are therefore consistent with the lineage-determining function of Gata1 involving a coordinted increase in expression kinetics of a set of genes important for red blood cell function.

Our observations regarding the Gata1 knock-out phenotype also warrant some discussion. With embryonically lethal phenotypes such as Gata1 knock-out, conventional analysis tends to be somewhat limited, since the embryos are dead because they have no red blood cells. By contrast, the Chimaera-Seq assay enables both quantification of cell numbers as well as characterisation of their molecular profiles. Moreover, there are no secondary effects caused by the dying embryo, because the wildtype host cells rescue overall fetal development, thus allowing a focussed analysis of cell-intrinsic molecular defects. One noteworthy observation from our data is that erythroid differentiation proceeds substantially beyond the stage where *Gata1* expression itself is first initiated, but fails to proceed to the late erythroid phase where expression of canonical red blood cell genes is greatly upregulated. However, gene expression prior to the differentiation block is not normal. In particular, we observed increased Spi1/Pu.1 in the Gata1 knock-out cells, consistent with the previously reported (Zhang et al., 1999) but also disputed (Hoppe et al., 2016) antagonistic relationship between Gata1 and Pu.1.

Within haematopoiesis, Pu.1 is recognised as a key regulator of myeloid and T-cell lineages, but not erythroid cells, even though a role in the proliferation of immature erythroid progenitors has been reported (Choe et al., 2010; reviewed in Carotta et al., 2010). Upregulation of Pu.1 in our immature *Gata1* knock-out cells therefore suggests that these cells of the primitive haematopoietic lineage represent progenitors with multilineage potential, rather than being restricted to just the red cell lineage. Further evidence for this notion is provided by our observation that the reduction in erythroid cells in the *Gata1* knock-out is accompanied by an increase in megakaryocyte progenitors, consistent

with a model whereby Gata1 levels influence the lineage choice decisions of a multipotent progenitor cell. Live cell tracking studies have suggested that the primary role of Gata1 and Pu.1 may be fate stabilization rather than fate choice (Hoppe et al., 2016). The increase in transcription rate of erythroid genes downstream of Gata1 would cohere with stabilizing the erythroid fate, thus suggesting that our results are consistent with roles in both fate choice and fate stabilization.

Our observation of an expanded pool of megakaryocyte progenitors may also be of direct relevance to our understanding of the pre-leukaemic transient myeloproliferative disease (TMD) that is prevalent in newborns with trisomy 21 (Roberts et al., 2013). TMD is thought to arise when a fetal specific haematopoietic progenitor cell with trisomy 21 acquires a partial loss of function mutation in *GATA1*, resulting in a short form of GATA1 (GATA1s). TMD is characterized by expansion of immature megakaryocyte progenitors, and in 10 to 20% of cases transforms into malignant acute megakaryoblastic leukaemia (reviewed in Bhatnagar et al., 2016)). Over-expression of GATA1s in mouse models resulted in the identification of mid-gestation fetal liver megakaryocyte progenitors as uniquely sensitive to this mutant GATA1s form compared to their adult bone marrow counterparts (Li et al., 2005). The over-represented population of immature megakaryocytic progenitors in our E8.5 *Gata1*$^{-}$ chimeras may correspond to the developmental emergence of this transient precursor, TMD-initiating cell, in the yolk sac.

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

Acknowledgements

We would like to thank Prof. Fabian Theis and Volker Bergen for discussions and valuable input on the scVelo implementation. We thank William Mansfield and the Gurdon Institute animal facility for blastocyst injections, the Flow Cytometry Core Facility at CIMR for cell sorting, Katarzyna Kania and the CRUK-CI genomics core for preparing the 10X libraries and for sequencing. Research in the authors' laboratories is supported by the Wellcome Trust, MRC, CRUK, Blood Cancer UK, NIH-NIDDK, the Sanger-EBI Single Cell Centre; by core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust-MRC Cambridge Stem Cell Institute; and by core funding from Cancer Research UK and the European Molecular Biology Laboratory. C.G. was funded by the Swedish Research Council (2017-06278), I.I. was funded by a British Heart Foundation studentship

(FS/18/56/35177), S.G. was supported by a Royal Society Newton International Fellowship (NIF\R1\181950). This work was funded as part of a Wellcome Strategic Award (105031/D/14/Z) awarded to Wolf Reik, Berthold Göttgens, John Marioni, Jennifer Nichols, Ludovic Vallier, Shankar Srinivas, Benjamin Simons, Sarah Teichmann, and Thierry Voet.

Figure Legends

Figure 1. Inferring Differentiation Trajectories at organismal scale

A. Pijuan-Sala et al. (2019) layout containing single-cell transcriptomes belonging from E6.5 to E8.5, colored by sampled time-point (left) and by cell-type (right). The overlaying arrows result from applying the scVelo pipeline to the whole embryonic dataset and represent inferred developmental trajectories. Arrowheads highlight the erythroid branch, displaying scVelo trajectory predictions that are inconsistent with real-time sampling.

B. Pijuan-Sala et al. (2019) layout highlighting single-cell transcriptomes belonging to E7.5 (left) and E8.5 (right) and colored by cell-type (see legend in A). The overlaying arrows result from applying the scVelo pipeline to these individual time-points and represent inferred developmental trajectories.

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

A. Dimensionality reduction with the first two principal components/MOFA factors using spliced reads alone (left), unspliced reads alone (middle) and both spliced and unspliced (right). Single-cell transcriptomes are colored by cell-type annotation; see Figure 1 for full legend.

B. MOFA characterization of spliced and unspliced reads assessing proportion of variance explained (i), overlap in highly variable genes calculating using either spliced or unspliced reads (ii), and factor weight distributions (iii).

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

A. Illustration of phase plot representation in datasets of differentiating cell populations, and associated scVelo predictions

B. Illustration of strategy for MURK gene identification

C. Phase plots of representative MURK genes. X-axis: normalized imputed counts of spliced transcript; y-axis: normalized imputed counts of unspliced transcript.

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

E. Zoomed-in UMAP of the erythroid branch (see Figure 1 for full UMAP) with scVelo calculations, before and after removing MURK genes identified in B.

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

A. Schematic of the G1ER system (Tsang et al., 1997; Weiss et al., 1997)
B. Behaviour of the 73 MURK genes identified in Figure 3 upon Gata1 induction in the G1ER system (Wu et al., 2011). Wu et al. report that upon Gata1 induction they obtained a total of 2769 upregulated genes, 6079 mildly upregulated, 3566 downregulated, and 3445 with no response.
C. UMAPS of *Gata1*$^{-}$ chimera cells allocated a hemato-endothelial identity colored by cell-type (sub-clusters defined in Pijuan-Sala et al. (2019) - BP: Blood Progenitors, EC: Endothelial Cells, Haem: Hemato-endothelial Progenitors, Mk: Megakaryocytes, My: Myeloid cells, Ery: Erythroid cells) and split by genotype.
D. UMAPS of *Gata1*$^{-}$ chimera cells allocated a hemato-endothelial identity colored by sampling timepoint and split by genotype.
E. Barplots with the quantification of chimera cells mapping to each hemato-endothelial lineage of the reference dataset (left) and to sampled time-points of the reference dataset (right).

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

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

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

A. UMAP representation of human fetal liver erythroid cell populations. The overlaying arrows result from applying the scVelo pipeline using all genes (left) or after MURK gene exclusion (right), and represent inferred developmental trajectories. Bottom UMAPs are colored by corresponding scVelo-inferred latent time.

B. Phase plots of representative MURK genes identified in human fetal liver erythropoiesis single-cell RNAseq dataset. See (A) for colour code.

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

Supplementary Figures

1. Dimensionality reduction with the first three principal components/MOFA factors using spliced reads alone (left), unspliced reads alone (middle) and both spliced and unspliced (right). Single-cell transcriptomes are colored by cell-type annotation; see Figure 1 for full legend.
2. Identification of MURK genes along yolk sac erythropoiesis. A. Phase plots of representative scVelo driver genes, with scVelo model prediction overlayed (see also Supplementary Table 1). B. Gene ontology of the 73 MURK genes identified in mouse yolk sac erythropoiesis, including sub-categories. C. Distribution of annotated cell type (top) and sampling time-point (bottom) along scVelo calculated latent time, using all genes (left panels) and after removing the MURK genes identified in Figure 3B-C.
3. Pijuan-Sala et al. (2019) layout highlighting nearest neighbours of *Gata1*$^{-}$ chimeras. In red are nearest neighbours of tdTom+ mutant cells, in black those of tdTom- wildtype cells. To compare with Figure 1A.
4. Impact of Gata1 knockout on *Spi1*/PU.1 expression on the hematoendothelial cell types. X-axis: *Spi1* $\log_2$(fold-change) in *Gata1*$^{-}$ vs WT chimera cells and Atlas nearest neighbours. Y-axis: $\log_{10}$(FDR).

Supplementary Tables

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

Methods

scVelo implementation

Mouse atlas dataset. To obtain separated count matrices for spliced and unspliced mRNAs, we ran velocyto 0.17.17 (La Manno et al., 2018) on the .bam files from the mouse atlas in Pijuan-Sala et al. 2019 (Pijuan-Sala et al., 2019; GEO accession number: GSE87038). We kept all cells that passed the QC as described in the original publication, but filtered out from downstream analysis the extraembryonic tissues: ExE endoderm, ExE ectoderm and Parietal endoderm as well as samples with no timepoint allocation (labelled as 'mixed gastrulation'). Then, we applied scVelo's pipeline v0.2.1 (Bergen et al., 2020). That is, we removed genes with less than 20 shared counts between spliced and unspliced counts, before normalising and log transforming the remaining genes. Then, we selected the top 5,000 highly variable genes (HVGs) for further calculation of moments; while performing imputation using the top 30 nearest neighbours from the graph connectivities generated with the original UMAP coordinates from Pijuan-Sala et al. 2019. The velocity vectors were computed in dynamical mode rather than steady state.

Human dataset. We first downloaded raw reads from Popescu et al., 2019 (Popescu et al., 2019; GEO accession number: GSE127980), and aligned them against the human genome hg19-3.0.0 with CellRanger v3.0.2 to generate the .bam files and obtain separated count matrices for spliced and unspliced mRNAs as described above. We filtered out cells with less than 3,550 counts, less than 900 genes and more than 6% mitochondrial counts. We ran scVelo's pipeline selecting 1,500 HVGs to compute PCA coordinates and applied batch correction using the function reducedMNN from the batchelor package v1.4.0 (Haghverdi et al., 2018), followed by the estimation of velocity vectors in the same way it was done for the mouse atlas.

MOFA+ implementation

We ran MOFA+ v1.4.0 (Argelaguet et al., 2020) using as input the two single cell experiment objects obtained from the spliced and unspliced counts independently. Each object was created in R using the scran v1.16.0 (Lun et al., 2016) library as follows: we started from the raw counts, normalized them with factor sizes obtained after pre-clustering, log transformed and reduced to 5000 HVG. We then switched to Python v3.7.4, where we regressed out the sample effect and scaled the object to generate a MOFA+ model with standard parameters. Finally, we used reducedMNN to correct the MOFA Factors for batch effects. The same objects used as MOFA input were used for PCA calculation in Figure 2A.

535

536 **MURK genes identification**

537 To identify MURK genes, we considered the imputed counts resulting from the scVelo standard
538 pipeline. Then, for each gene and each population among the Erythroid lineage, we calculated the
539 unspliced versus spliced slope with a linear regression, as well as the standard error on the slope. We
540 selected all genes for which the slope in Erythroid3 is significantly higher than the slope in Erythoid2
541 (according to a one-sided t-test p-value < 0.05), the average spliced counts in Erythroid3 is higher than
542 the average spliced counts in every other population, and the slope in Erythroid3 positive. We found
543 73 genes that respect all these criteria in the mouse dataset, 377 in the human dataset.

544

545 **Gene ontology enrichment analysis**

546 We performed gene ontology enrichment analysis using the http://geneontology.org website
547 comparing the MURK genes against all biological processes, with the default all Mus musculus genes
548 in database as background set (Ashburner et al., 2000; The Gene Ontology, 2019). We ranked the
549 processes by FDR.

550

551 **Overlap testing**

552 Overlap was tested with Fisher exact test. We calculated the probability of having m = 45 genes of our
553 n = 73 MURK genes mapping to the A = 1022 high response genes (out of N = 4195 genes) in the Wu
554 et al., 2011 publication (GEO accession number: GSE30142) as the probability of randomly picking m
555 elements of a specific type when randomly choosing n elements out of N, where the frequency of the
556 special type is A/N.

557

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

559 **Embryo collection.** All procedures were performed in strict accordance to the UK Home Office
560 regulations for animal research under the project license number PPL 70/8406. **Chimaera generation.**
561 TdTomato-expressing mouse embryonic stem cells (ESC) were derived as previously described (Pijuan-
562 Sala et al., 2019). Briefly, ESC lines were derived from E3.5 blastocysts obtained by crossing a male
563 ROSA26tdTomato (Jax Labs – 007905) with a wildtype C57BL/6 female, expanded under the 2i+LIF
564 conditions (Ying et al., 2008) and transiently transfected with a Cre-IRES-GFP plasmid (Wray et al.,

2011) using Lipofectamine 3000 Transfection Reagent (ThermoFisher Scientific, #L3000008) according to manufacturer's instructions. A tdTomato-positive, male, karyotypically normal line, competent for chimaera generation as assessed using morula aggregation assay, was selected for targeting *Gata1*. Two guides were designed using the http://crispr.mit.edu tool (guide 1: CGGCTACTCCACTGTGGCGG; guide 2: CGCTTCTTGGGCCGGATGAG) and were cloned into the pX458 plasmid (Addgene, #48138) as previously described (Ran et al., 2013). The obtained plasmids were then used to transfect the cells and single transfected clones were expanded and assessed for Cas9-induced mutations. Genomic DNA was isolated by incubating cell pellets in 0.1 mg/ml of Proteinase K (Sigma, #03115828001) in TE buffer at 50°C for 2 hours, followed by 5 min at 99°C. The sequence flanking the guide-targeted sites was amplified from the genomic DNA by polymerase chain reaction (PCR) in a Biometra T3000 Thermocycler (30 sec at 98°C ; 30 cycles of 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 High-Fidelity DNA Polymerase (NEB, #M0530S) according to the manufacturer's instructions. Primers including Nextera overhangs were used (F-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTACCCTGCCTCAACTGTG; R-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTTGTCTTGGGCAGGAACA), allowing library preparation with the Nextera XT Kit (Illumina, #15052163), and sequencing was performed using the Illumina MiSeq system according to manufacturer's instructions. An ESC clone showing a 38 base-pair frameshift mutation in exon 4 resulting in the functional inactivation of *Gata1* were selected for injection into C57BL/6 E3.5 blastocysts. A total of 6 chimaeric embryos were harvested at E8.5, dissected, and single-cell suspensions were generated from three independent pools of embryos by TrypLE Express dissociation reagent (Thermo Fisher Scientific) incubation for 7-10 minutes at 37°C under agitation. Single-cell suspensions were sorted into tdTom+ and tdTom- samples using a BD Influx sorter with DAPI at 1µg/ml (Sigma) as a viability stain for subsequent 10X scRNA-seq library preparation (version 3 chemistry), and sequencing using an S1 flow cell in the Illumina Novaseq platform, which resulted in 8420 $tdTom^-$ and 7944 $tdTom^+$ cells that passed quality control (see "Single-cell RNA sequencing analysis" below).

Single-cell RNA sequencing analysis. Raw files were processed with Cell Ranger 3.0.2 using default mapping arguments. Reads were mapped to the mm10 genome and counted with GRCm38.92 annotation, including tdTomato sequence for chimera cells. Cell barcodes with expression profiles significantly different to the ambient mRNA expression profile were identified using emptyDrops (Lun et al., 2019), and cell barcodes with low complexity, i.e. low total mRNA counts and/or high mitochondrial proportion, were identified by fitting four-component bivariate mixture models to the $\log_{10}$-transformed total mRNA counts and percentage of mitochondrial counts, and selecting the components with high total mRNA and low mitochondrial percentage. Gene expression normalization

and doublet cell barcodes were identified using the approach taken by Pijuan-Sala et al. (2019). Both spliced and unspliced count matrices were extracted using velocyto 0.17.17 (La Manno et al., 2018).

Mapping to the reference dataset. We mapped the chimaera cells to the mouse atlas following almost exactly the procedure used in the original publication article to map the *Tal1* chimaera. First, we concatenated the mouse atlas and chimaera counts (both previously controlled for quality of the cells), normalized the resulting counts matrix with scran, computed HVGs and then applied multiBatchPCA, and reducedMNN with cosine normalization from batchelor (Haghverdi et al., 2018) for batch effect correction within samples (where sample refers to a single lane of a 10x Chromium chip) as well as between datasets in order to extract a number of nearest neighbours between the mouse atlas and the chimaera using queryKNN from BiocNeighbors package v1.6.0.

Differential Gene Expression Analysis. For differential gene expression analysis, we took samples that included at least 7 cells per tdTom status per cell population (eg. Erythroid3). We ran the analysis in scanpy v1.5.1 (Wolf et al., 2018) with Wilcoxon test and choosing 2 as fold change and 0.1 as false discovery rate thresholds.

References

Akunuru, S., and Geiger, H. (2016). Aging, Clonality, and Rejuvenation of Hematopoietic Stem Cells. Trends Mol Med *22*, 701-712.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol *21*, 111.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol.

Bhatnagar, N., Nizery, L., Tunstall, O., Vyas, P., and Roberts, I. (2016). Transient Abnormal Myelopoiesis and AML in Down Syndrome: an Update. Curr Hematol Malig Rep *11*, 333-341.

Borrett, M.J., Innes, B.T., Jeong, D., Tahmasian, N., Storer, M.A., Bader, G.D., Kaplan, D.R., and Miller, F.D. (2020). Single-Cell Profiling Shows Murine Forebrain Neural Stem Cells Reacquire a Developmental State when Activated for Adult Neurogenesis. Cell Rep *32*, 108022.

Carotta, S., Wu, L., and Nutt, S.L. (2010). Surprising new roles for PU.1 in the adaptive immune response. Immunol Rev *238*, 63-75.

633 Choe, K.S., Ujhelly, O., Wontakal, S.N., and Skoultchi, A.I. (2010). PU.1 directly regulates cdk6 gene
634 expression, linking the cell proliferation and differentiation programs in erythroid cells. J Biol Chem
635 *285*, 3044-3052.

636 Dahlin, J.S., Hamey, F.K., Pijuan-Sala, B., Shepherd, M., Lau, W.W.Y., Nestorowa, S., Weinreb, C.,
637 Wolock, S., Hannah, R., Diamanti, E*., et al.* (2018). A single-cell hematopoietic landscape resolves 8
638 lineage trajectories and defects in Kit mutant mice. Blood *131*, e1-e11.

639 Ezer, D., Moignard, V., Gottgens, B., and Adryan, B. (2016). Determining Physical Mechanisms of
640 Gene Expression Regulation from Single Cell Gene Expression Data. PLoS Comput Biol *12*, e1005072.

641 Fujiwara, Y., Browne, C.P., Cunniff, K., Goff, S.C., and Orkin, S.H. (1996). Arrested development of
642 embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. Proc Natl Acad
643 Sci U S A *93*, 12355-12358.

644 Gerber, T., Murawala, P., Knapp, D., Masselink, W., Schuez, M., Hermann, S., Gac-Santel, M.,
645 Nowoshilow, S., Kageyama, J., Khattak, S*., et al.* (2018). Single-cell analysis uncovers convergence of
646 cell identities during axolotl limb regeneration. Science *362*.

647 Grosveld, F., van Assendelft, G.B., Greaves, D.R., and Kollias, G. (1987). Position-independent, high-
648 level expression of the human beta-globin gene in transgenic mice. Cell *51*, 975-985.

649 Gutierrez, L., Tsukamoto, S., Suzuki, M., Yamamoto-Mukai, H., Yamamoto, M., Philipsen, S., and
650 Ohneda, K. (2008). Ablation of Gata1 in adult mice results in aplastic crisis, revealing its essential role
651 in steady-state and stress erythropoiesis. Blood *111*, 4375-4385.

652 Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-
653 sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol *36*, 421-427.

654 Higgs, D.R., Wood, W.G., Jarman, A.P., Sharpe, J., Lida, J., Pretorius, I.M., and Ayyub, H. (1990). A
655 major positive regulatory region located far upstream of the human alpha-globin gene locus. Genes
656 Dev *4*, 1588-1601.

657 Hoppe, P.S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., Moritz, N., Endele, M.,
658 Filipczyk, A., Gambardella, A., Ahmed, N*., et al.* (2016). Early myeloid lineage choice is not initiated
659 by random PU.1 to GATA1 protein ratios. Nature *535*, 299-302.

660 Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchis-Calleja, F., Guijarro, P., Sidow, L., Fleck,
661 J.S., Han, D*., et al.* (2019). Organoid single-cell genomic atlas uncovers human-specific features of
662 brain development. Nature *574*, 418-422.

663 Kim, J.K., and Marioni, J.C. (2013). Inferring the kinetics of stochastic gene expression from single-cell
664 RNA-sequencing data. Genome Biol *14*, R7.

665 Kondo, A., Fujiwara, T., Okitsu, Y., Fukuhara, N., Onishi, Y., Nakamura, Y., Sawada, K., and Harigae, H.
666 (2016). Identification of a novel putative mitochondrial protein FAM210B associated with erythroid
667 differentiation. Int J Hematol *103*, 387-395.

668 La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K.,
669 Kastriti, M.E., Lonnerberg, P., Furlan, A*., et al.* (2018). RNA velocity of single cells. Nature *560*, 494-
670 498.

671 Li, Z., Godinho, F.J., Klusmann, J.H., Garriga-Canut, M., Yu, C., and Orkin, S.H. (2005). Developmental
672 stage-selective effect of somatically mutated leukemogenic transcription factor GATA1. Nat Genet
673 *37*, 613-619.

674 Liu, X., Chen, Y., Zhang, Y., Liu, Y., Liu, N., Botten, G.A., Cao, H., Orkin, S.H., Zhang, M.Q., and Xu, J.
675 (2020). Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D
676 chromatin by biotinylated dCas9. Genome Biol *21*, 59.

677 Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of
678 single-cell RNA-seq data with Bioconductor. F1000Res *5*, 2122.

679 Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., participants in the 1st Human Cell Atlas,
680 J., and Marioni, J.C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based
681 single-cell RNA sequencing data. Genome Biol *20*, 63.

682 Mahadevaiah, S.K., Sangrithi, M.N., Hirota, T., and Turner, J.M.A. (2020). A single-cell transcriptome
683 atlas of marsupial embryogenesis and X inactivation. Nature *586*, 612-617.

684 McGrath, K., and Palis, J. (2008). Ontogeny of erythropoiesis in the mammalian embryo. Curr Top
685 Dev Biol *82*, 1-22.

686 Mettananda, S., Gibbons, R.J., and Higgs, D.R. (2016). Understanding alpha-globin gene regulation
687 and implications for the treatment of beta-thalassemia. Ann N Y Acad Sci *1368*, 16-24.

688 Monteiro, R., Pouget, C., and Patient, R. (2011). The gata1/pu.1 lineage fate paradigm varies
689 between blood populations and is modulated by tif1gamma. EMBO J *30*, 1093-1103.

690 Palis, J. (2016). Hematopoietic stem cell-independent hematopoiesis: emergence of erythroid,
691 megakaryocyte, and myeloid potential in the mammalian embryo. FEBS Lett *590*, 3965-3974.

692 Pevny, L., Lin, C.S., D'Agati, V., Simon, M.C., Orkin, S.H., and Costantini, F. (1995). Development of
693 hematopoietic cells lacking transcription factor GATA-1. Development *121*, 163-172.

694 Pevny, L., Simon, M.C., Robertson, E., Klein, W.H., Tsai, S.F., D'Agati, V., Orkin, S.H., and Costantini, F.
695 (1991). Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for
696 transcription factor GATA-1. Nature *349*, 257-260.

697 Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C.,
698 Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L*., et al.* (2019). A single-cell molecular map of mouse
699 gastrulation and early organogenesis. Nature *566*, 490-495.

700 Pina, C., May, G., Soneji, S., Hong, D., and Enver, T. (2008). MLLT3 regulates early human erythroid
701 and megakaryocytic cell fate. Cell Stem Cell *2*, 264-273.

702 Popescu, D.M., Botting, R.A., Stephenson, E., Green, K., Webb, S., Jardine, L., Calderbank, E.F.,
703 Polanski, K., Goh, I., Efremova, M*., et al.* (2019). Decoding human fetal liver haematopoiesis. Nature
704 *574*, 365-371.

705 Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering
706 using the CRISPR-Cas9 system. Nat Protoc *8*, 2281-2308.

Roberts, I., Alford, K., Hall, G., Juban, G., Richmond, H., Norton, A., Vallance, G., Perkins, K., Marchi, E., McGowan, S*., et al.* (2013). GATA1-mutant clones are frequent and often unsuspected in babies with Down syndrome: identification of a population at risk of leukemia. Blood *122*, 3908-3917.

Schultz, M.B., and Sinclair, D.A. (2016). When stem cells grow old: phenotypes and mechanisms of stem cell aging. Development *143*, 3-14.

Shivdasani, R.A., Fujiwara, Y., McDevitt, M.A., and Orkin, S.H. (1997). A lineage-selective knockout establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet development. EMBO J *16*, 3965-3973.

Storry, J.R., Joud, M., Christophersen, M.K., Thuresson, B., Akerstrom, B., Sojka, B.N., Nilsson, B., and Olsson, M.L. (2013). Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. Nat Genet *45*, 537-541.

The Gene Ontology, C. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res *47*, D330-D338.

Ton, M.N., Guibentif, C., and Gottgens, B. (2020). Single cell genomics and developmental biology: moving beyond the generation of cell type catalogues. Curr Opin Genet Dev *64*, 66-71.

Tsang, A.P., Visvader, J.E., Turner, C.A., Fujiwara, Y., Yu, C., Weiss, M.J., Crossley, M., and Orkin, S.H. (1997). FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. Cell *90*, 109-119.

Vyas, P., Ault, K., Jackson, C.W., Orkin, S.H., and Shivdasani, R.A. (1999). Consequences of GATA-1 deficiency in megakaryocytes and platelets. Blood *93*, 2867-2875.

Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science *360*, 981-987.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*.

Weiss, M.J., Yu, C., and Orkin, S.H. (1997). Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. Mol Cell Biol *17*, 1642-1651.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol *19*, 15.

Wray, J., Kalkan, T., Gomez-Lopez, S., Eckardt, D., Cook, A., Kemler, R., and Smith, A. (2011). Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. Nat Cell Biol *13*, 838-845.

Wu, W., Cheng, Y., Keller, C.A., Ernst, J., Kumar, S.A., Mishra, T., Morrissey, C., Dorman, C.M., Chen, K.B., Drautz, D*., et al.* (2011). Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res *21*, 1659-1671.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.

744 Zhang, P., Behre, G., Pan, J., Iwama, A., Wara-Aswapati, N., Radomska, H.S., Auron, P.E., Tenen, D.G.,
745 and Sun, Z. (1999). Negative cross-talk between hematopoietic regulators: GATA proteins repress
746 PU.1. Proc Natl Acad Sci U S A *96*, 8705-8710.

747 Zhang, Q., He, Y., Luo, N., Patel, S.J., Han, Y., Gao, R., Modak, M., Carotta, S., Haslinger, C., Kind, D.,
748 *et al.* (2019). Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. Cell *179*,
749 829-845 e820.

750 Zhou, W., Yui, M.A., Williams, B.A., Yun, J., Wold, B.J., Cai, L., and Rothenberg, E.V. (2019). Single-Cell
751 Analysis Reveals Regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T
752 Cell Development. Cell Syst *9*, 321-337 e329.

Figure 1: Inferring Differentiation Trajectories at organismal scale

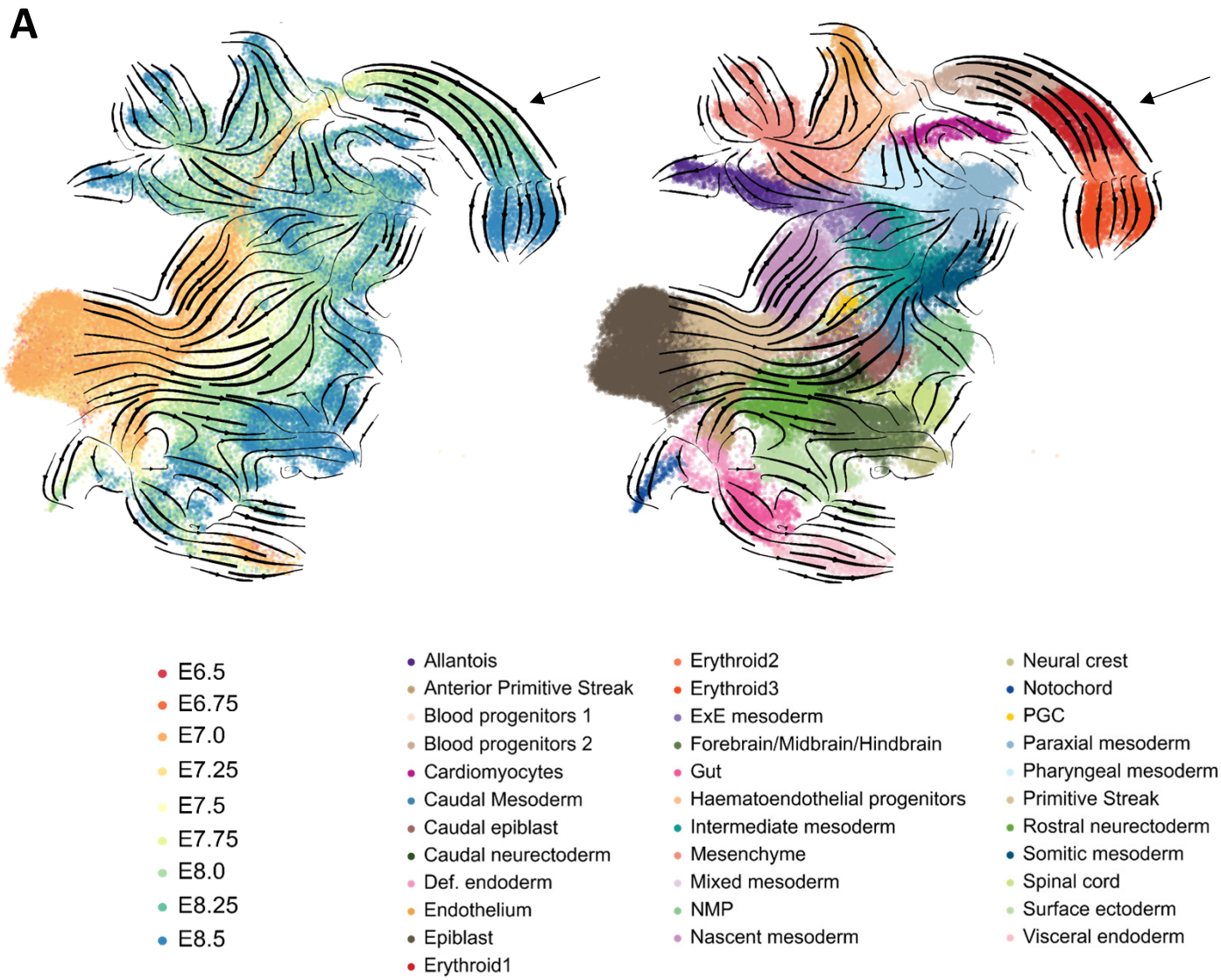
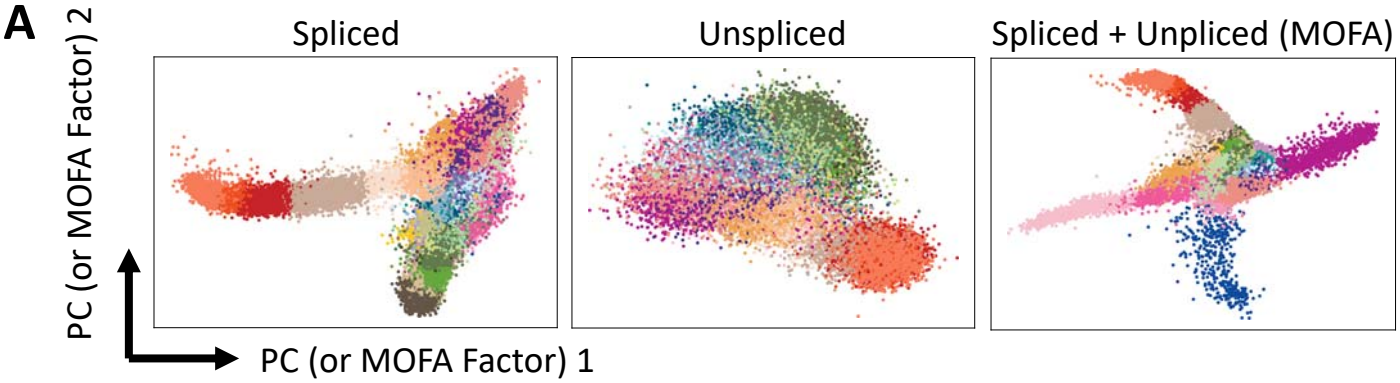


Figure 2: Unspliced counts contribute to explaining the variability among cell types



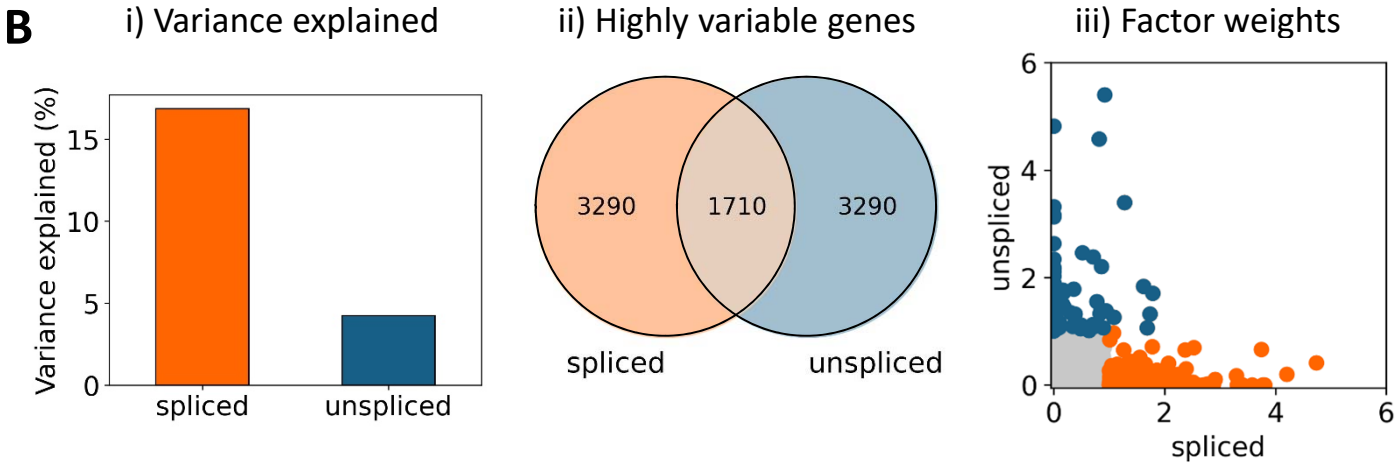


Figure 3. A set of genes with complex expression kinetics confounds velocity estimation in erythropoiesis

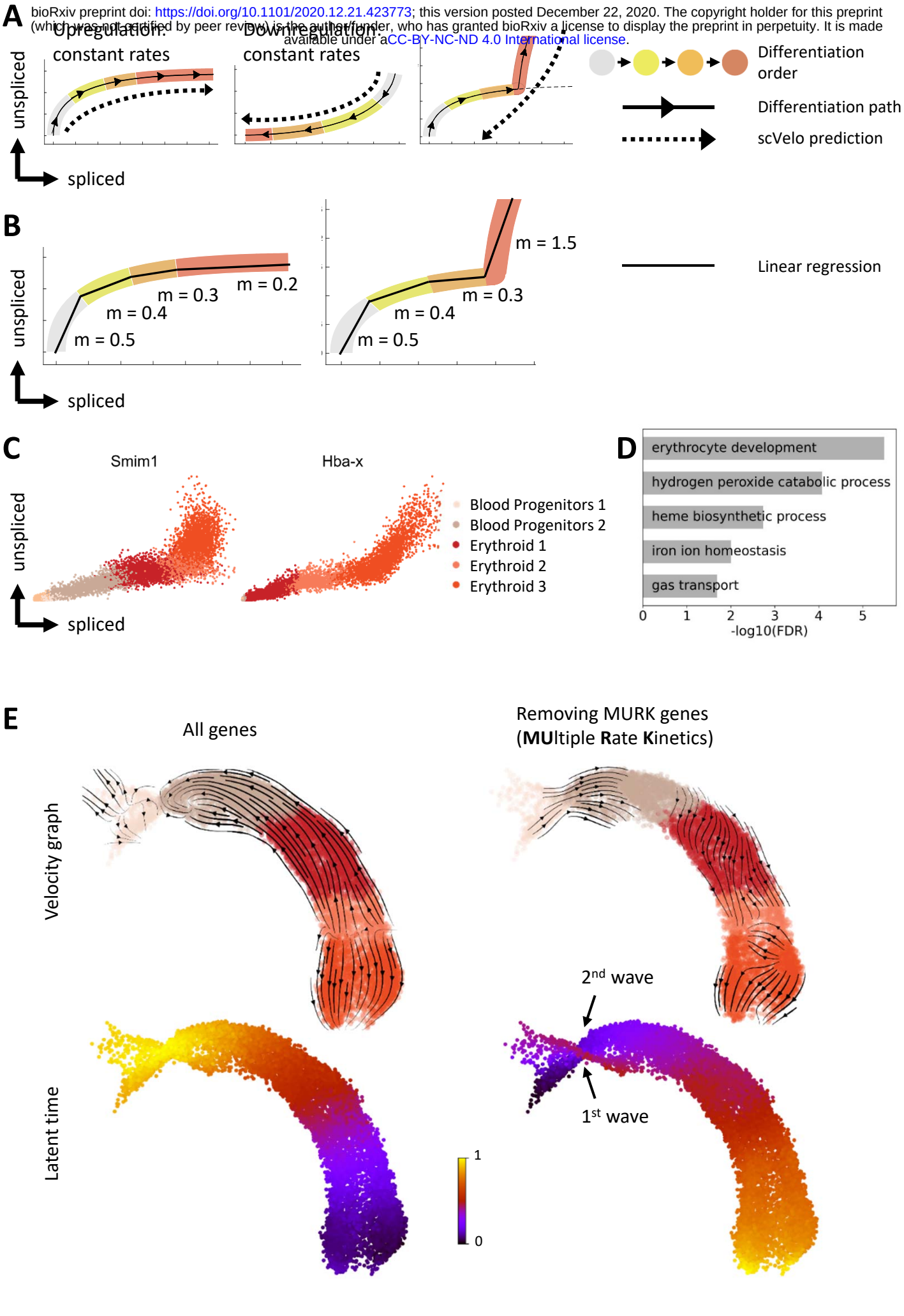


Figure 4: In vivo analysis of Gata1 function using Chimaera assays coupled with scRNA-Seq

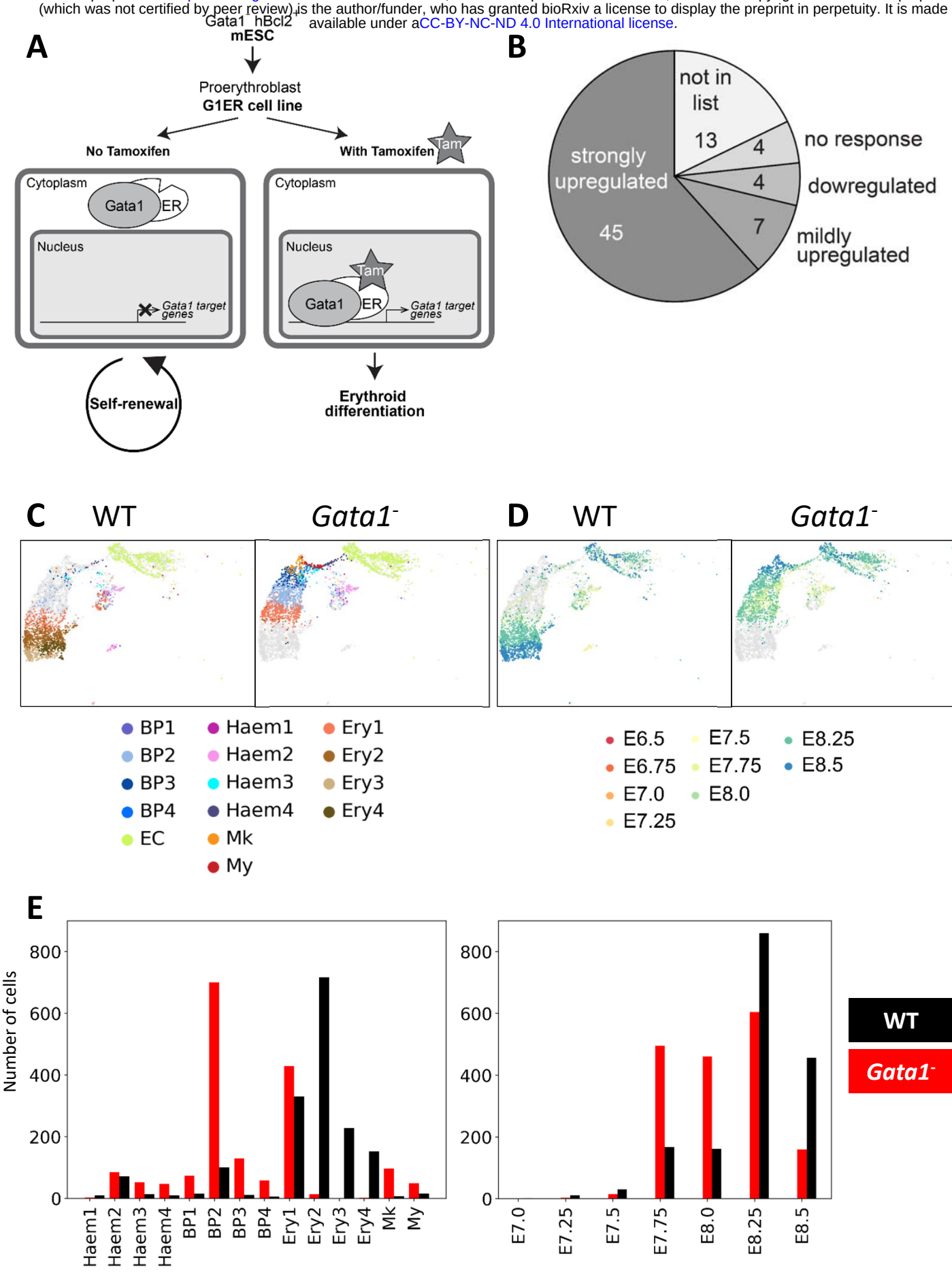


Figure 5. Gata1 chimaera assay reveals disruption of MURK genes and perturbed yolk sac hematopoiesis

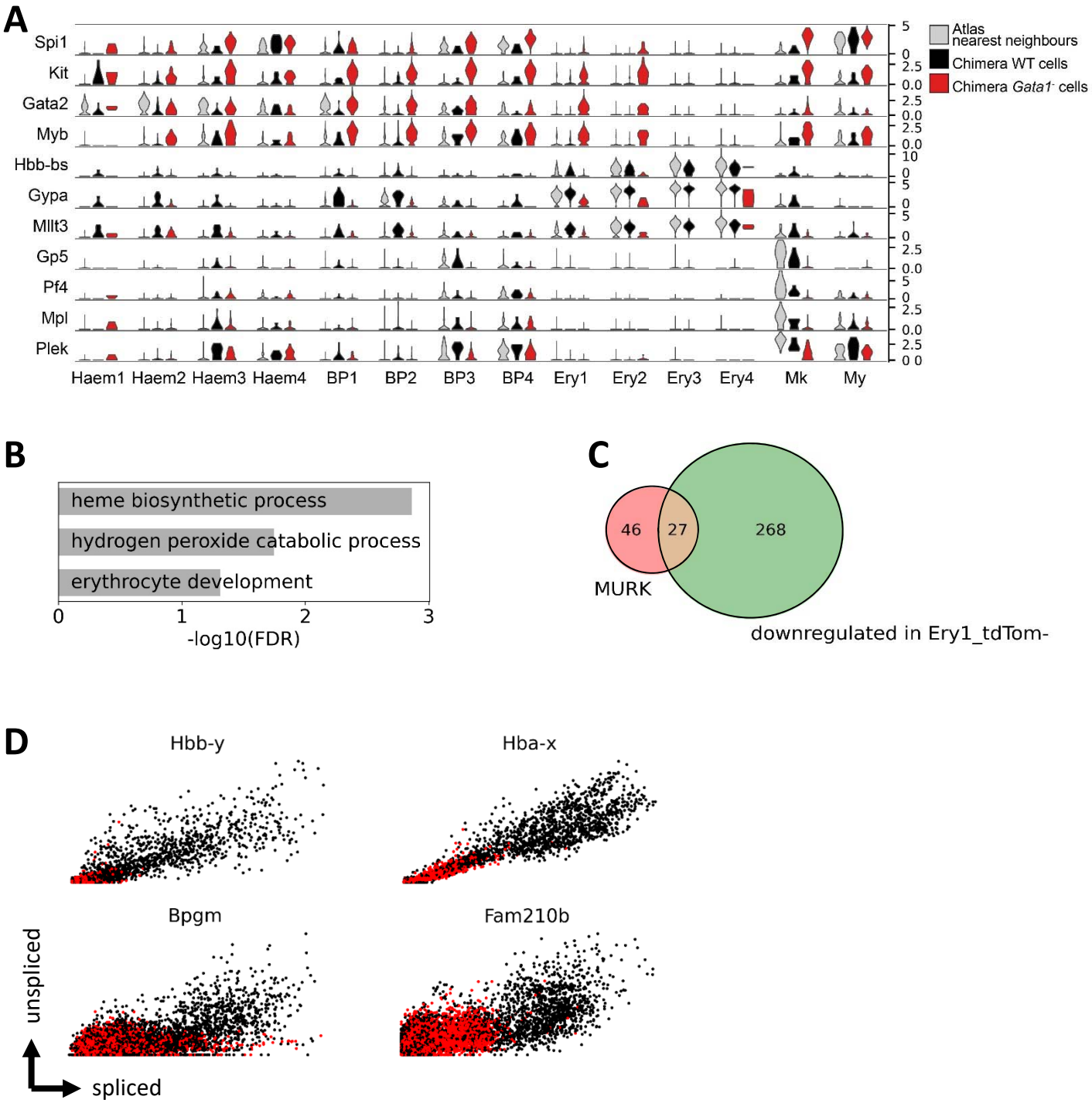


Figure 6. Concept of dual kinetics of gene expression is also revealed in human foetal liver hematopoiesis