Cell fate decision in erythropoiesis: insights from multi-omics studies.

Steven Tur , Carmen G. Palii , Marjorie Brand

PII: S0301-472X 24)00023-7
DOI: https://doi.org/10.1016/j.exphem.2024.104167
Reference: EXPHEM 104167

To appear in: *xperimental Hematology*

Received date: 15 November 2023
Revised date: 10 January 2024
Accepted date: 13 January 2024

Please cite this article as: Steven Tur , Carmen G. Palii , Marjorie Brand , Cell fate decision in erythropoiesis: insights from multi-omics studies., *xperimental Hematology* 2024), doi: https://doi.org/10.1016/j.exphem.2024.104167

Highlights

- Overview of erythropoiesis and new discoveries using single-cell technologies to understand cell fate decisions.
- Importance of determining the likelihood of cell fate progression and key factors involved in cell fate determination.
- Creating predictive gene regulatory networks is essential for a comprehensive regulatory model of erythropoiesis.
- Tools to identify clusters, trajectories, and create gene regulatory networks for a better understanding of cell fate.

Cell fate decision in erythropoiesis: insights from multi-omics studies.

Steven Tur[1], Carmen G. Palii[1] & Marjorie Brand[1*]

[1]Department of Cell and Regenerative Biology, Wisconsin Blood Cancer Research Institute, Wisconsin Institutes for Medical Research, University of Wisconsin School of Medicine and Public Health, Madison, Carbone Cancer Center, Wisconsin, USA

[*]Correspondence

mbrand3@wisc.edu

Abstract

Every second, the body produces 2 million red blood cells through a process called erythropoiesis. Erythropoiesis is hierarchical in that it results from a series of cell fate decisions whereby hematopoietic stem cells progress towards the erythroid lineage. Single-cell transcriptomic and proteomic approaches have revolutionized the way we understand erythropoiesis, revealing it to be a gradual process that underlies a progressive restriction of fate potential driven by quantitative changes in lineage-specifying transcription factors. Despite these major advances, we still know very little about what cell fate decision entails at the molecular level. Novel approaches that

simultaneously measure additional properties in single cells, including chromatin accessibility, transcription factor binding and/or cell surface proteins are being developed at a fast pace, providing the means to exciting new advances in the near future. In this review, we briefly summarize the main findings obtained from single cell studies of erythropoiesis, highlight outstanding questions, and suggest recent technological advances to address them.

Text

Introduction

Erythropoiesis is an important cellular differentiation process that leads to the formation of red blood cells from hematopoietic stem cells (HSCs) [1, 2]. Owing to sophisticated mouse models [3] and human ex vivo differentiation systems that recapitulate all steps of differentiation [4], erythropoiesis has been comprehensibly analyzed, which makes it an ideal model system to address outstanding questions in biology. For example, enhancers were first characterized through extensive analyses of transcription at the ß-globin locus during erythroid differentiation [5]. More recently, erythropoiesis was among the first complete cellular differentiation systems analyzed by droplet-based single cell RNA sequencing (scRNAseq) [6-8] and single cell proteomics [9]. In this review, we highlight these (and other [10, 11]) studies, which together with novel experimental tools and innovative analysis methods, hold the promise to advance beyond the cellular level towards a mechanistic understanding of cell fate choice in erythropoiesis.

I- Current state of knowledge from single cell studies of erythropoiesis.

One of the main findings from early transcriptomic [6-8, 12] and proteomic [9] single cell profiling was the gradual nature of erythropoiesis whereby HSCs undergo a progressive restriction of fate potential driven by quantitative changes in lineage-specifying transcription factors (LS-TFs) [13]. This concept, which is supported by both transcriptomic and proteomic data provided a very precise description of the early stages of erythropoiesis showing a continuum of differentiation with accumulation of some known, as well as novel, populations at specific points along the erythroid trajectory. Furthermore, these studies revealed for the first-time co-expression of LS-TFs in individual multipotent progenitors at both RNA and protein levels. Importantly, overexpression experiments demonstrated that quantitative changes in the level of a non-erythroid TF is sufficient to deviate progenitors from their preferred erythroid trajectory towards a non-erythroid lineage [9] providing proof-of-principle that quantitative changes in TFs protein levels direct cell fate decision in individual cells.

Another notable finding was the identification of alternative paths to the traditional hematopoietic tree, including an unexpected coupling of the erythroid and the basophil lineages, which again is supported by both transcriptomic and proteomic data [6, 7, 9].

More recently, the “Cellular Indexing of Transcriptomes and Epitopes by sequencing” (CITEseq) approach [14] that provides simultaneous measures of the transcriptome and cell surface proteins in single cells (**Figure 1; Table 1** and see section III below for more details) was used in several studies of erythropoiesis. For example, Doty et al. [11] identified a “death” trajectory that is taken by a significant proportion of erythroid

progenitors at the proerythroblast stage when they express high levels of heme. In this study, the coupling of cell surface proteins to the transcriptome was instrumental in validating the erythroid differentiation trajectory. Furthermore, this study exemplifies the power of single cell approaches to identify small pro-apoptotic cells that could not have been isolated or expanded in vitro without altering their phenotype. Thus, the implementation of single cell technologies presents a viable alternative to the use of fluorescent-activated cell sorting (FACS) for specialized cell populations. Another study using CITEseq led to the identification of granulocytic precursors and macrophages that physically associate with erythroid cells in the bone marrow as part of the erythro-myeloblastic island [10]. Again, cell surface markers coupled to transcriptomic measurements allowed for a more precise definition of those cells, highlighting the usefulness of coupling these two layers of information.

II- Outstanding questions

Single cell transcriptomic has now become a standard approach to examine phenotypes and phenotypic changes during development and disease. However, major questions remain that go beyond the description of lineages and their relationships. In this section, we highlight three questions that we believe have the potential to be addressed by recently developed single cell multi-omics approaches (**Figure 2**).

1. **Infer cell fate probability along the erythropoietic lineage**.

Single cell measurements that sample large numbers of cells at multiple stages of differentiation provide an unprecedented opportunity to infer a probability for each cell

derived from HSCs to become erythroid or to diverge and become another cell type. Such fate maps are indispensable to understanding the mechanism of cell fate decision because they allow one to correlate changing molecular properties to the dynamic of cell fate decisions. Several approaches have been developed to infer fate probability along differentiation trajectories, including PBA [15], FateID [16], Palantir [17] and more recently CellRank [18]. One way to estimate the extent to which inferred probability truly reflects cell fate is to combine lineage tracing with scRNAseq at different time-points such that gene expression at one time point can be correlated to fate at a later time point. Such approaches, termed lineage-tracing with single-cell RNA sequencing (LT-scSeq) require the introduction of genetic barcodes that are unique, heritable and detectable by sequencing, and are therefore typically limited to ex vivo differentiation systems, transplanted cells and/or genetically engineered mice [19-23]. Nevertheless, these approaches are very powerful as they revealed for example that cell fate decision occurs earlier than predicted by scRNAseq and that the transcriptome alone (as measured by scRNAseq) is not sufficient to accurately predict cell fate [20]. This suggests that additional heritable properties (e.g. chromatin accessibility) contribute to fate determination. Interestingly, a recently developed inducible Cas9 barcoding mouse line (DARLIN) that combines lineage-tracing with simultaneous measures of transcription, DNA methylation and chromatin accessibility in single cells (using a plate-based approach named Camellia-seq) showed that DNA methylation is strongly associated to clonal memory [24], which highlights the importance of incorporating DNA methylation measurements to cell fate decisions models.

While genomic barcodes provide a practical method for lineage reconstruction, somatic mutations in mitochondrial DNA also allow clonal tracking [25], offering a potential approach for human in vivo lineage-tracing.

2. Measure the key players of cell fate decision and their quantitative changes along the erythroid trajectory.

The realization that transcripts alone are not sufficient to estimate cell fate probabilities highlights the need to measure additional molecular properties in single cells. In this section, we propose a list of molecular players that are likely to be major actors of the cell fate decision process.

Proteins: we and others have shown dramatic discrepancies between transcript levels and protein levels, mostly during dynamic processes like erythropoiesis [26]. Such discrepancies are particularly problematic for ***lineage-specifying TFs*** and ***signaling TFs*** [27] that together represent the main drivers of cell fate decisions and should therefore be measured at the protein level. Furthermore, these proteins often work in a dose-dependent manner [28-30] and should therefore also be quantified, ideally using absolute quantification approaches that provide copy-number measurements [31]. Because they mediate the function of TFs, ***cofactors*** including chromatin-modifying enzymes should also be quantified at the protein level. Finally, ***cell surface proteins*** (not RNAs) should be measured to facilitate purification of prospective populations (**Figure 2**).

Chromatin accessibility: given that chromatin is inherently refractory to transcription, measures of chromatin opening offer critical information on the portions of the genome that have the potential to be transcribed. While regions of opened chromatin are often used to infer TF binding through DNA-binding motifs enrichment, one must keep in mind that these inferences are likely compromised by the complexity of the rules governing TF binding including large redundancies between TFs of the same family [32]. Thus, it is important to measure TF binding directly (**Figure 2**).

TF genomic binding: to facilitate identification of TF target genes, it is necessary to directly measure TF genomic binding in single cells.

Histones and DNA modifications: chromatin modifications provide critical information pertaining to gene expression and should therefore also be measured in single cells.

Spatial transcriptomics: Single-cell transcriptomic approaches such as multiplexed error robust fluorescence in situ hybridization (MERFISH) [33] or sequential fluorescence in situ hybridization (SEQFISH+) [34] provide invaluable information on cell-to-cell interactions, or the position of cells within a tissue. However, those approaches are difficult to combine with the simultaneous measures of other modalities by high-throughput multi-omics approaches (**Figure 1**).

While single cell measurements are often performed in stem and progenitor cells, it may be important to analyze all cells along the erythroid trajectory, including cells that are thought to be committed. Indeed, our data [26] and that of others [35] showed that TFs from non-erythroid lineage are still expressed in late erythroid progenitors. Furthermore,

transgenic mouse experiments combining the knockout of LSD1 with lineage-tracing revealed that late erythroid progenitors can deviate towards the myeloid lineage, which suggests these cells have not completely lost their myeloid potential [35, 36].

3. **Understand cell fate decisions by building predictive gene regulatory networks that integrate the main players of erythropoiesis**

Maybe one of the most challenging aspects of understanding erythropoiesis is to integrate the main players described above into a biologically meaningful model that takes into account all regulatory aspects underlying cell fate decisions. Ideally, such a model should be dynamic, quantitative and predictive. Several attempts have been made at building gene regulatory networks of erythropoiesis [37, 38], including our own temporal model that integrates quantitative changes in protein and mRNA abundances of transcription factors [26]. However, these models have not been built based on single cell measurements. In the next section, we highlight some selected technical and analytical advances that we believe will be key in the progression towards a global regulatory model for erythropoiesis at the single cell level.

III- Single cell multi-omics tools to decipher erythropoiesis

While bulk RNA sequencing has been widely utilized in numerous fields of biology and health research, the primary function of this technique is to measure RNA in many cells within a sample of interest, allowing for the determination of the average expression level of individual genes from the same sample [39]. However, when studying complex

systems such as erythropoiesis and the various pathologies associated with this lineage, it is necessary to have more detailed information on the heterogeneity of the samples of interest and each cell population. To address this issue, single-cell RNA sequencing has gained importance as a means of compensating for the lack of information on the heterogeneity of cell groups. Many research groups are now utilizing this technology to determine the gene expression of each cell providing critical insights into the transcriptional activity and fate decisions of these cell populations [39].

The most widely used single cell approach is based on single cell suspension and gel bead emulsion that creates a fine droplet of oil containing a single cell (**Figure 1**). This process, coupled with next-generation sequencing, has drastically reduced the cost of this technology. Other methods, such as plate-based approaches that provide for deeper sequencing of individual cells or nuclei, are more expensive and challenging [40].

Data analysis generally begins with clustering based on the Louvain or Leiden algorithms [41, 42]. Combined with t-SNE (t-distributed Stochastic Neighbor Embedding), UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) or FA (Force Atlas) graphical representations, the SCANPY, scVI-tools, Bioconductor and Seurat workflows then allow one to categorize all samples by cell type based on RNA expression, thus providing information on sample heterogeneity [43-47] (**Table 1**). In addition, further study can be performed using trajectory analysis tools such as PAGA (Partition-Based Graph Abstraction) [48] or sc-Velo [49] that select or detect a root cell and use RNA information to infer cell trajectories. These tools also

incorporate pseudo-time into their approach to reconstruct differentiation pathways (**Figure 2; Table 1**).

Using scRNA-seq techniques, numerous studies have demonstrated the diversity and complexity of the hematopoietic differentiation process, which involves precise regulation of cell fate with a clear hierarchical structure at different stages [13]. However, scRNA-seq only provides information on transcriptomes and not on proteins. While single-cell protein data can also be obtained using approaches such as Cytometry by Time Of Flight (CyTOF) [9] to measure transcription factors, this approach is limited to 50 proteins that can be measured simultaneously. Finally, scRNAseq does not provide information on other important molecular layers, such as chromatin accessibility and/or cell surface markers that are necessary to purify cell populations of interest. These issues can be addressed at least partly by single cell multi-omics approaches that provide simultaneous information on several layers of information including transcriptome, chromatin opening and/or cell surface proteins. Multi-omics approaches have been developed through barcoding, enabling each cell to be marked with a unique marker identifier (UMI) and each layer of information to be marked with different barcodes. These techniques generally couple two layers of data. Furthermore, methods have recently been developed that simultaneously cover up to three layers [50]. Below, we describe selected single-cell approaches (experimental and analytical) that go beyond sc-RNAseq and that we find the most promising for shining light onto the mechanism underlying cell fate decision mechanisms in erythropoiesis.

1. Experimental advances

CITEseq (Cellular Indexing of Transcriptomes and Epitopes by sequencing) [14] (**Figure 1; Table 1**). The principle of CITEseq is based on incubating cells with a cocktail of barcoded antibodies, that can extend to cover several hundred cell surface proteins. Once the incubation is complete, the cells undergo a gel bead emulsion (GEM) process for single-cell droplet encapsulation, followed by cell lysis **(Figure 1)**. Then, antibody barcodes are hybridized to reverse transcript oligonucleotides bound to beads for future library preparation and sequencing. CITEseq combines information on the transcriptome and the cell surface proteins and allows the construction of precise cell trajectories during different stages of development based on scRNA-seq and the expression of cell surface proteins in single cells [14]. The strength of the CITEseq approach is provided by the information on cell surface proteins which makes it possible to isolate cell populations for further analysis and in vitro or in vivo validation.

Sci-CAR (Single-cell Combinatorial Indexing for Chromatin Accessibility and RNA) [51] (**Figure 1; Table 1)**. Transcriptome and cell surface proteins are not the only possible combinations for multi-omics since techniques can now also integrate chromatin accessibility. Indeed, single-cell multi-omics approaches can integrate scRNA-seq with transposase-accessible chromatin (scATAC-seq). First, the nuclei are separated and spread out in a plate. Then, specific barcodes are added to each well along with RNA and ATAC indexes. The ATAC barcodes also include Tn5 transposase, which cuts at regions of open chromatin. This technique combines scATAC-seq and scRNA-seq on several thousand cells, providing information on the dynamics of chromatin accessibility and gene expression. Furthermore, 10X genomics now offers a commercial kit for the simultaneous measure of scRNA and scATAC in microfluidic systems. Overall, this

technique provides a better understanding of the role of epigenetics in cell fate decisions and memory processes.

TEAseq [52] and **DOGMAseq** [53] (**Figure 1; Table 1)**. These recently developed techniques combine three layers of data: transcripts, cell surface proteins (i.e epitopes) and chromatin accessibility, and are thus termed trimodal (**Figure 1**). They are based on the principle of CITEseq, but the cells are permeabilized during the preparation process after antibody incubation. Once permeabilized, the cells are incubated with the Tn5 transposase, which enters the nucleus and introduces DNA barcodes into open chromatin of each cell. The cells are then isolated by gel bead emulsion in microdroplets containing specific beads. Those beads contain poly-A-tail for the isolation of the transcriptome and cell surface proteins tags. A Tn5 oligo is also present for the ATAC-seq library. This trimodal technique captures transcriptomes, cell surface proteins, and open chromatin to provide a more complete analysis of the differentiation process and the possibility of purifying rare cell groups for further study [52, 53].

In addition, DOGMAseq has been developed to provide the added possibility of measuring a fourth modality i.e. mitochondrial DNA (mtDNA) [53] (**Figure 1; Table 1**). DOGMAseq offers two possible paths to achieve this: one involves cell fixation to preserve mtDNA, while the other involves a slight permeabilization similar to the TEAseq protocol, which allows for the detection of cell surface proteins. In summary, DOGMAseq has demonstrated that mtDNA can be detected as a fourth modality if combined with fixation or slight permeabilization. Fixation is better for detecting mtDNA and permeabilization is better for detecting cell surface proteins. However, it is important to note that TEAseq can also detect mtDNA using the permeabilization with

digitonin approach [52]. Thus, it is possible to use either approach depending on the specific question being asked. Using DOGMAseq, the authors have demonstrated its effectiveness in resolving bone marrow heterogeneity [53].

Sc-multi-CUT&Tag (Single-cell multi Cleavage Under Targets and Tagmentation) [54] (**Figure 1, Table 1**). While the above techniques provide a wealth of information on chromatin opening, the transcriptome and cell surface proteins, it is important to realize that motif enrichment as measured by ATACseq data does not necessarily equate to transcription factor binding. Furthermore, the above approaches do not provide information on histone modifications. To obtain information on DNA-protein interactions or histone modifications in single cells it is possible to use sc-multi-CUT&Tag [54], an approach adapted from CUT&Tag [55] that combines antibodies against multiple transcription factors, cofactors and/or histone modifications. The sc-multi-CUT&Tag method provides information on the interactions between multiple proteins with chromatin by combining antibodies directed against the proteins of interest with a protein A-Tn5 (pA-Tn5) transposase fusion protein pre-complexed with barcoded oligonucleotides [54]. A recent variation of this method, named **nano-CUT&Tag** (or nano-CT) proposes to use a nanobody directly fused with Tn5 instead of a secondary antibody [56].

These single-cell approaches will enable characterization of the heterogeneity of several subpopulations in thousands of cells.

2. Analytical advances

As described above, a large number of multi-omics methods have recently emerged that combine several modalities such as transcriptome, chromatin accessibility and cell surface proteins. These approaches have the potential to measure multiple types of data simultaneously in each cell, revealing new information on differentiation processes and cell fate. However, analyzing such data requires powerful tools to extract relevant biological information. We now emphasize some tools that we believe will be the most useful to study erythropoiesis.

Data Integration, Data Transfer and Trajectory Analyses. Multi-omics data analysis can be challenging owing to the complexity of the sequencing information provided by multi-layers of data. Furthermore, some pipelines are designed to handle paired or unpaired data. Paired data refers to measurements of multiple modalities performed simultaneously on the same samples, such as the TEAseq technique [52]. Unpaired data, on the other hand, originates from different techniques and/or different biological samples. It is important to know which computational tools to use for incorporating all layers of data [57]. Among the many available workflows that analyze multi-omic data, the scVI-tools suite provides powerful computing pipelines based on a combination of probabilistic approaches and machine learning [45]. Here we describe several of the scVI-tools (as well as other tools) for multi-omics analyses (**Figure 1; Table 1**).

CITEseq has gained popularity because of its efficiency and the ability to utilize over a hundred antibodies [14]. However, it is important to note that data analysis should not solely focus on the scRNA-seq component. It is also not recommended to rely exclusively on cell surface proteins for validation. **ScVI-tools** offers Total Variational Inference (**TotalVI**), a joint probabilistic analysis that combines both modalities to derive a joined representation [58]. TotalVI has been designed to analyze CITEseq data using both sequencing modalities (**Table 1**). The approach applies mathematical tools and trains a model by machine learning using RNA and protein layers with the option of adding a batch correction [58]. TotalVI also features protein normalization to distinguish foreground from background, joint representation, and differential expression testing. Its efficiency has already been shown studying for example immune cells in mice [59].

The Yosef group also recently developed the **MultiVI** pipeline [60]. MultiVI first focuses on two modalities: transcriptome and open chromatin modalities. It proposes to analyze gene expression and chromatin opening with a deep generative model for probabilistic analysis. The model is suitable for experiments involving simultaneous multimodal measurements and can also integrate a third modality, such as cell surface proteins. Thus, it is ideal for performing analyses on data from TEAseq [52] or DOGMAseq [53] (**Figure 1; Table 1**). Furthermore, the model also offers the possibility of integrating non-paired data, and it can consider technical issues such as background noise and batch effect by integrating batch information, as also proposed by TotalVI [58]. The batch information in TotalVI [58] and MultiVI [60] pipelines enable a correction to be made to incorporate the data correctly in a latent space, considering experimental differences between samples.

Overall, the SCVI-tools pipelines take the best of deep machine learning by combining multi-modalities from different multi-omics techniques, making them a potent tool for analyzing complex datasets [45]. The authors have proven that their model can thoroughly investigate the heterogeneity of samples and lead to a better understanding of cell fate decisions by integrating all modalities.

Machine learning not only integrates paired and unpaired data but can also use multimodal data to perform data transfer on samples that lack one of the modalities. For example, **scArches** (Single-Cell Architecture Surgery) can use a TotalVI model to extract protein measurements from CITEseq combined with gene expression to train the model and then perform "surgery" for data transfer [61] (**Table 1**). Thus, it is possible to take the CITEseq data as a reference model to be trained and matched on a query dataset with scRNA-seq only. Then the models are trained again to perform a data transfer on the query dataset and give in silico protein values on a scRNA-seq dataset while considering batch correction [61]. Using these models, it is possible to analyze large datasets with multimodalities and integrate other data sets from atlases to include more cells or add missing information.

The above-described models are partly used to realize the latent representation with multiple modalities. They can be completed by additional computer tools that perform trajectory analyses and deduce a pseudo time to study cell differentiation (**Figure 2**). Tools like Partition-Based Graph Abstraction (**PAGA**) [48] can analyze latent space sets from TotalVI [58], MultiVI [60] or other pipelines and deduce trajectories starting from a

root cell. PAGA associates each previously identified cluster with a node linked together by weighted edges with the thickness of the ridges showing the degree of connectivity between clusters [48]. The thicker the edges, the greater the connections representing a statistical measure of the connectivity between the identified clusters. Cells are then ordered according to their distance from the root cell. The path established by PAGA then represents the average of all single-cell courses passing through the corresponding cell clusters. It is then possible to deduce the pseudo time from the root cell to track the progression of differentiation (**Figure 2**).

Many pipelines can be used to carry out data integration and trajectory studies. In addition to the methods described above, we note the popular Seurat pipelines that also offer numerous tools for analyzing multi-modality data [46], Seurat uses single modalities first to create a single-modality latent space. The single modalities are then integrated by identifying anchors to propose a new latent space comprising both modalities [46].

From Lineage Trajectories to Gene Regulatory Networks. Trajectory studies are not the only analyses that can result from the sequencing of RNA and other modalities. Indeed, multi-modality sequencing can also be used to decipher the mechanisms underlying gene regulation. We mention **MIRA**, an innovative pipeline based on machine learning, which can analyze a latent space such as MultiVI based on gene expression and chromatin opening modalities [62]. Interestingly, MIRA can perform a complete series of analyses from clustering to latent representation, trajectories, pseudo-time analysis, and critical regulators identification (**Figure 2**; **Table**

1). This is another efficient approach to analyzing multi-modal data. For more details on MIRA, please refer to the paper describing this pipeline [62].

In addition to trajectories, the establishment of gene regulatory networks (**GRNs**) can help to answer critical questions, such as identifying transcription factors that regulate gene expression and better understand the importance of chromatin structure [50] (**Figure 2**). Here, we describe some selected approaches that have been used to establish GRNs**.** First, we note that scRNAseq is, for the most part, sufficient to establish GRNs. However, including other modalities can help to derive GRNs that are more precise and robust. Here we focus on three approaches: Single-cell Multi-Task Network Inference (scMTNI) [63], Dyctis [64] and single-Cell rEgulatory Network Inference and Clustering + (SCENIC+) [65]. These three pipelines can use paired or unpaired scRNA-seq and scATAC-seq to infer GRN (**Figure 2; Table 1**).

The first approach, **sc-MTNI** [63], uses single-cell data from multi-omics considering a cell lineage tree. It is a powerful tool to infer a detailed gene regulatory network for each cell type on a previously defined lineage trajectory. Sc-MTNI can also integrate paired, unpaired, and/or bulk data to establish the final gene regulatory networks. The authors have shown the robustness of their network by applying sc-MTNI to a human hematopoietic dataset. Notably, sc-MTNI was able to identify new regulators linked to hematopoietic regulatory mechanisms and confirm known hematopoietic regulators. Thus, scMTNI is an effective tool for identifying the regulators that steer cells towards a particular path.

The second approach, **Dyctis** [64], also uses scRNAseq and scATACseq data, but infers time-resolved GRNs by using pseudo-time information and context-specific transcription factors footprints. In addition, Dyctis provides a function to compare context-specific networks. Finally, Dyctis can identify TFs with changes in regulatory activities but no change in their levels of expression.

The third approach, **SCENIC+** [65], is the only method that focuses on the inference of gene regulatory networks from enhancers to create enhancer-based GRNs (eGRNs). Moreover, it can detect the presence or absence of enhancers in every cell population identified. Notably, SCENIC+ uses a comprehensive database for TF binding motifs and includes a computational perturbation algorithm to predict the effects of knocking out specific TFs on the GRN.

Thus, these three approaches to GRN inference have different objectives and can help answer complementary biological questions in gene regulation (**Figure 2; Table 1**).

IV- Conclusion

The emergence of single-cell RNA-seq and multi-omics approaches has revolutionized our understanding of cellular and molecular mechanisms at the single-cell level. The limitations of bulk methods have been overcome by the detailed characterization of cellular heterogeneity, particularly in complex processes such as erythropoiesis. Approaches such as CITE-seq, TEAseq, and DOGMA-seq integrate genome,

proteome, and chromatin structure, offering more comprehensive insights into cellular mechanisms (**Figure 1; Table 1**). Advanced analysis pipelines such as Sc-VI, MIRA and scMTNI harness the power of machine learning to integrate and interpret complex multi-omics data, revealing cellular trajectories, pseudo-times, and gene regulatory networks (**Figure 2**). However, much remains to be done to derive molecular information at the gene level, and to address key questions such as the molecular underpinnings of cell fate decision at the gene and chromatin levels in early progenitors. For this, new analysis methods to derive GRNs that incorporate additional information (TF binding, protein levels) as well as new analytical methods that allow comparative analyses between cell trajectories or between experimental conditions are warranted. The future lies in continuously improving these techniques, enabling a deeper exploration of cellular mechanisms and a more precise understanding of complex biological processes.

FIGURE LEGENDS

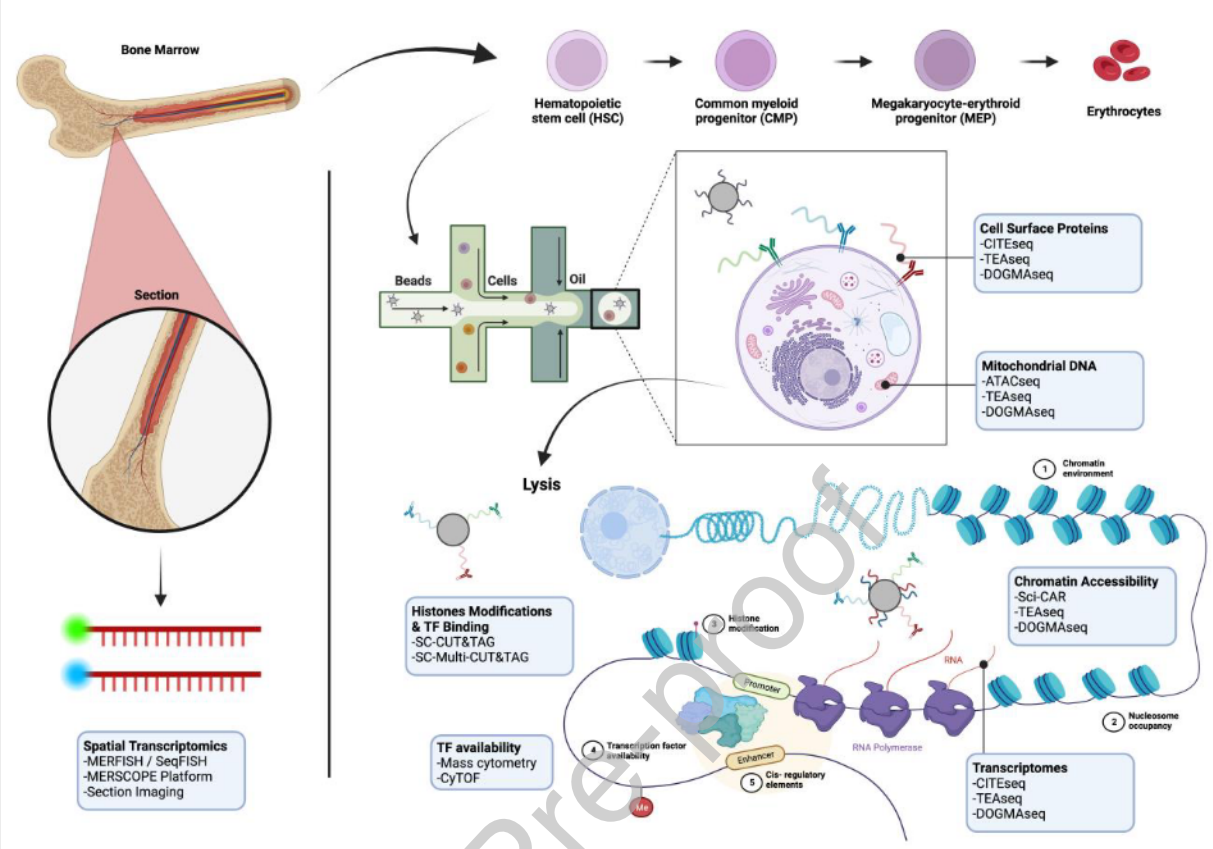

Figure 1: Multi-modal measurements in single cells for erythropoiesis.

In this diagram we first show the possibility of spatial transcriptomics using MERFISH/SeqFISH techniques. This is made possible by the commercial MERSCOPE Platform, which can perform spatial measurement from an organ section such as bone marrow (left panel). The panel on the right shows an example of principal differentiation during erythropoiesis. It is possible to select any stage to perform single cell multi-omics measurements. We then represent the concept of a single droplet encapsulation containing cells and beads. We first zoom in to show that it is possible to perform cell surface markers measurement and/or mitochondrial DNA depending on the technique chosen. We then zoom in to focus on the nucleus. From here, it is possible to perform numerous measurements at the same time or not, depending on the technique chosen.

These include chromatin accessibility, transcriptomes, histone modifications and/or TF binding, and finally the possibility of measuring the availability of transcript factors (right-hand panel).

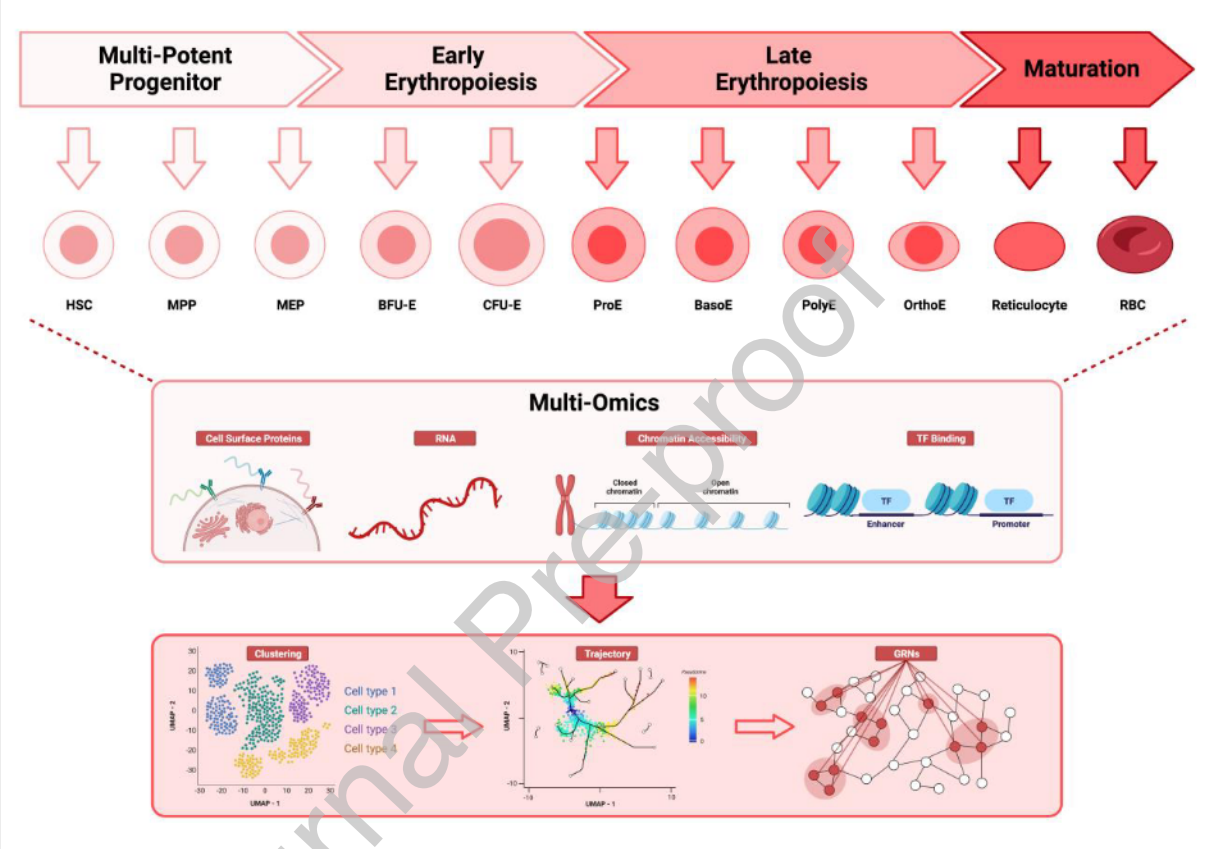

Figure 2: Single cell multi-omics workflow for erythropoiesis.

In this figure we illustrate one of the possible paths of multi-omics experiment at the single cell level for the different stages of differentiation during erythropoiesis. The first part (from top to bottom) shows the 4 main stages of differentiation from multi-potent progenitors to mature cells. We have then represented the different stages of differentiation in more detail for each of the 4 main stages. We then created a first cartoon representation of a possible measurement modality that can be obtained simultaneously or individually. These modalities can include the measurement of cell

surface proteins, RNA, chromatin accessibility, and transcription factor binding. The final diagram explains the main analyses that can be conducted, which are clustering, trajectory, and gene regulatory networks.

Table 1: Summary of single cell multi-omics approaches.

In this table, we summarize the main multi-omics approaches detailed in this review. For each approach, we specify the different targets and measurements that these different techniques can provide. We then associate with each technique the available software and tools capable of analyzing these data using multi-modalities. References for each of these measurement and analysis techniques are given.

Method	Target	Analysis	References
SC RNA	**Transcriptomes**	**Scanpy** **Seurat** **PAGA** **SCENIC** **SC MTNI**	[44, 46, 48, 63, 66]
CITEseq	**Transcriptomes** **Cell surface proteins**	**Scanpy** **Seurat** **PAGA** **SCENIC** **TotalVI** **scArches**	[14, 44, 46, 48, 58, 61, 63, 66]
Sci-CAR	**Transcriptomes** **Chromatin Accessibility** **Mitochondrial DNA**	**Scanpy** **Seurat** **MultiVI** **Mira** **SC MTNI** **SCENIC/SCENIC+**	[44, 46, 60, 62, 63, 65, 66]
TEAseq	**Transcriptomes** **Cell surface proteins** **Chromatin Accessibility** **Mitochondrial DNA**	**Scanpy** **Seurat** **MultiVI** **TotalVI** **Mira** **SC MTNI** **SCENIC/SCENIC+**	[44, 46, 52, 58, 60, 62, 63, 65, 66]
DOGMAseq	**Transcriptomes** **Cell surface proteins** **Chromatin Accessibility** **Mitochondrial DNA**	**Scanpy** **Seurat** **MultiVI** **TotalVI** **Flowjo** **Mira** **SC MTNI** **SCENIC/SCENIC+**	[44, 46, 53, 58, 60, 62, 63, 65, 66]
SC-Multi-CUT&Tag	**Multi-TFs binding** **Histone modifications** **DNA-Proteins interactions**	**Seurat** **Slingshot** **MEME** **SAMtools**	[46, 54, 55, 67-69]

Acknowledgements

The authors acknowledge F. Jeffrey Dilworth for critically reading the manuscript. This work is supported by grants from the National Institutes of Health, United States (2R01DK098449-06) (to M.B.).

References

1. Palis, J., *Primitive and definitive erythropoiesis in mammals.* Front Physiol, 2014. **5**: p. 3.
2. Schippel, N. and S. Sharma, *Dynamics of human hematopoietic stem and progenitor cell differentiation to the erythroid lineage.* Exp Hematol, 2023. **123**: p. 1-17.
3. Parker, M.P. and K.R. Peterson, *Mouse Models of Erythropoiesis and Associated Diseases.* Methods Mol Biol, 2018. **1698**: p. 37-65.
4. Palii, C.G., R. Pasha, and M. Brand, *Lentiviral-mediated knockdown during ex vivo erythropoiesis of human hematopoietic stem cells.* J Vis Exp, 2011. **doi: 10.3791/2813.**(53).
5. Bender, M.A., et al., *Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region.* Mol Cell, 2000. **5**(2): p. 387-93.
6. Tusi, B.K., et al., *Population snapshots predict early haematopoietic and erythroid hierarchies.* Nature, 2018. **555**(7694): p. 54-60.
7. Pellin, D., et al., *A comprehensive single cell transcriptional landscape of human hematopoietic progenitors.* Nat Commun, 2019. **10**(1): p. 2395.
8. Popescu, D.M., et al., *Decoding human fetal liver haematopoiesis.* Nature, 2019. **574**(7778): p. 365-371.
9. Palii, C.G., et al., *Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate.* Cell Stem Cell, 2019. **24**(5): p. 812-820 e5.
10. Romano, L., et al., *Erythroblastic islands foster granulopoiesis in parallel to terminal erythropoiesis.* Blood, 2022. **140**(14): p. 1621-1634.
11. Doty, R.T., et al., *The transcriptomic landscape of normal and ineffective erythropoiesis at single-cell resolution.* Blood Adv, 2023. **7**(17): p. 4848-4868.
12. Psaila, B., et al., *Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways.* Genome Biol, 2016. **17**: p. 83.
13. Brand, M. and E. Morrissey, *Single-cell fate decisions of bipotential hematopoietic progenitors.* Curr Opin Hematol, 2020. **27**(4): p. 232-240.
14. Stoeckius, M., et al., *Simultaneous epitope and transcriptome measurement in single cells.* Nat Methods, 2017. **14**(9): p. 865-868.
15. Weinreb, C., et al., *Fundamental limits on dynamic inference from single-cell snapshots.* Proc Natl Acad Sci U S A, 2018. **115**(10): p. E2467-E2476.

16. Herman, J.S., Sagar, and D. Grun, *FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data.* Nat Methods, 2018. **15**(5): p. 379-386.
17. Setty, M., et al., *Characterization of cell fate probabilities in single-cell data with Palantir.* Nat Biotechnol, 2019. **37**(4): p. 451-460.
18. Lange, M., et al., *CellRank for directed single-cell fate mapping.* Nat Methods, 2022. **19**(2): p. 159-170.
19. Biddy, B.A., et al., *Single-cell mapping of lineage and identity in direct reprogramming.* Nature, 2018. **564**(7735): p. 219-224.
20. Weinreb, C., et al., *Lineage tracing on transcriptional landscapes links state to fate during differentiation.* Science, 2020. **367**(6479).
21. Bowling, S., et al., *An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells.* Cell, 2020. **181**(6): p. 1410-1422 e27.
22. Wagner, D.E. and A.M. Klein, *Lineage tracing meets single-cell omics: opportunities and challenges.* Nat Rev Genet, 2020. **21**(7): p. 410-427.
23. Wang, S.W., et al., *CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information.* Nat Biotechnol, 2022. **40**(7): p. 1066-1074.
24. Li, L., et al., *A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells.* Cell, 2023.
25. Miller, T.E., et al., *Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations.* Nat Biotechnol, 2022. **40**(7): p. 1030-1034.
26. Gillespie, M.A., et al., *Absolute Quantification of Transcription Factors Reveals Principles of Gene Regulation in Erythropoiesis.* Mol Cell, 2020. **78**(5): p. 960-974 e11.
27. Choudhuri, A., et al., *Common variants in signaling transcription-factor-binding sites drive phenotypic variability in red blood cell traits.* Nat Genet, 2020. **52**(12): p. 1333-1345.
28. Kulessa, H., J. Frampton, and T. Graf, *GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts.* Genes Dev, 1995. **9**(10): p. 1250-62.
29. Graf, T. and T. Enver, *Forcing cells to change lineages.* Nature, 2009. **462**(7273): p. 587-94.
30. Jost, M., et al., *Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs.* Nat Biotechnol, 2020. **38**(3): p. 355-364.
31. Brand, M. and J.A. Ranish, *Proteomic/transcriptomic analysis of erythropoiesis.* Curr Opin Hematol, 2021. **28**(3): p. 150-157.
32. Suter, D.M., *Transcription Factors and DNA Play Hide and Seek.* Trends Cell Biol, 2020. **30**(6): p. 491-500.
33. Zhang, M., et al., *Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH.* Nature, 2021. **598**(7879): p. 137-143.
34. Eng, C.L., et al., *Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH.* Nature, 2019. **568**(7751): p. 235-239.
35. Yu, L., et al., *An erythroid-to-myeloid cell fate conversion is elicited by LSD1 inactivation.* Blood, 2021. **138**(18): p. 1691-1704.
36. Brand, M., *Epigenetic plasticity of erythroid progenitors.* Blood, 2021. **138**(18): p. 1646-1648.

37. Swiers, G., R. Patient, and M. Loose, *Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification.* Dev Biol, 2006. **294**(2): p. 525-40.
38. Novershtern, N., et al., *Densely interconnected transcriptional circuits control cell states in human hematopoiesis.* Cell, 2011. **144**(2): p. 296-309.
39. Li, X. and C.Y. Wang, *From bulk, single-cell to spatial RNA sequencing.* Int J Oral Sci, 2021. **13**(1): p. 36.
40. Mereu, E., et al., *Benchmarking single-cell RNA-sequencing protocols for cell atlas projects.* Nat Biotechnol, 2020. **38**(6): p. 747-755.
41. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics-Theory and Experiment, 2008.
42. Traag, V.A., L. Waltman, and N.J. van Eck, *From Louvain to Leiden: guaranteeing well-connected communities.* Sci Rep, 2019. **9**(1): p. 5233.
43. Huber, W., et al., *Orchestrating high-throughput genomic analysis with Bioconductor.* Nat Methods, 2015. **12**(2): p. 115-21.
44. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome Biol, 2018. **19**(1): p. 15.
45. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nat Methods, 2018. **15**(12): p. 1053-1058.
46. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587 e29.
47. Gayoso, A., et al., *A Python library for probabilistic analysis of single-cell omics data.* Nat Biotechnol, 2022. **40**(2): p. 163-166.
48. Wolf, F.A., et al., *PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells.* Genome Biol, 2019. **20**(1): p. 59.
49. Bergen, V., et al., *Generalizing RNA velocity to transient cell states through dynamical modeling.* Nat Biotechnol, 2020. **38**(12): p. 1408-1414.
50. Lee, J., D.Y. Hyeon, and D. Hwang, *Single-cell multiomics: technologies and data analysis methods.* Exp Mol Med, 2020. **52**(9): p. 1428-1442.
51. Cao, J., et al., *Joint profiling of chromatin accessibility and gene expression in thousands of single cells.* Science, 2018. **361**(6409): p. 1380-1385.
52. Swanson, E., et al., *Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq.* Elife, 2021. **10**.
53. Mimitou, E.P., et al., *Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells.* Nat Biotechnol, 2021. **39**(10): p. 1246-1258.
54. Gopalan, S., et al., *Simultaneous profiling of multiple chromatin proteins in the same cells.* Mol Cell, 2021. **81**(22): p. 4736-4746 e5.
55. Kaya-Okur, H.S., et al., *CUT&Tag for efficient epigenomic profiling of small samples and single cells.* Nat Commun, 2019. **10**(1): p. 1930.
56. Bartosovic, M. and G. Castelo-Branco, *Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag.* Nat Biotechnol, 2023. **41**(6): p. 794-805.
57. Heumos, L., et al., *Best practices for single-cell analysis across modalities.* Nat Rev Genet, 2023. **24**(8): p. 550-572.
58. Gayoso, A., et al., *Joint probabilistic modeling of single-cell multi-omic data with totalVI.* Nat Methods, 2021. **18**(3): p. 272-282.

59. Thomson, Z., et al., *Trimodal single-cell profiling reveals a novel pediatric CD8alphaalpha(+) T cell subset and broad age-related molecular reprogramming across the T cell compartment.* Nat Immunol, 2023. **24**(11): p. 1947-1959.
60. Ashuach, T., et al., *MultiVI: deep generative model for the integration of multimodal data.* Nat Methods, 2023. **20**(8): p. 1222-1231.
61. Lotfollahi, M., et al., *Mapping single-cell data to reference atlases by transfer learning.* Nat Biotechnol, 2022. **40**(1): p. 121-130.
62. Lynch, A.W., et al., *MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells.* Nat Methods, 2022. **19**(9): p. 1097-1108.
63. Zhang, S., et al., *Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets.* Nat Commun, 2023. **14**(1): p. 3064.
64. Wang, L., et al., *Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics.* Nat Methods, 2023. **20**(9): p. 1368-1378.
65. Bravo Gonzalez-Blas, C., et al., *SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks.* Nat Methods, 2023. **20**(9): p. 1355-1367.
66. Aibar, S., et al., *SCENIC: single-cell regulatory network inference and clustering.* Nat Methods, 2017. **14**(11): p. 1083-1086.
67. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
68. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
69. Street, K., et al., *Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.* BMC Genomics, 2018. **19**(1): p. 477.

Graphical Abstract