

# Selective silencing rather than targeted activation of gene expression underlies fate choice in human hematopoietic stem cells

Parmentier R<sup>1\*</sup>, Moussy A<sup>1\*</sup>, Chantalat S<sup>2\*</sup>, Racine L<sup>1</sup>, Sudharshan R<sup>3,4</sup>, Papili Gao N<sup>4</sup>, Stockholm D<sup>1</sup>, Corre G<sup>5</sup>, Deleuze JF<sup>2</sup>, Gunawan R<sup>3</sup>, Paldi A<sup>1</sup>.

<sup>1</sup>Ecole Pratique des Hautes Etudes, PSL Research University, St-Antoine Research Center, Inserm U938, 34 rue Crozatier, 75012, Paris, France

<sup>2</sup> Centre National de Recherche en Génomique Humaine, CEA, 91000 Evry, France

<sup>3</sup>Department of Chemical and Biological Engineering, University, Buffalo, NY 14260, USA

<sup>4</sup>Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland

<sup>5</sup>Genethon; 1bis rue de l'International, 91001 Evry, France

\* Equal contribution

## Abstract:

When human cord blood derived CD34+ cells are induced to differentiate *in vitro*, they undergo rapid and dynamic morphological and molecular transformation that are critical for the fate

commitment. Using ATAC-seq and single-cell RNA sequencing, we detected two phases of this process. In the first phase, we observed that a rapid and widespread chromatin opening - that makes most of the gene promoters in the genome accessible - precedes a global upregulation of gene transcription and a concomitant increase in the cell-to-cell variability of gene expression. The second phase is marked by a slow chromatin closure that precedes an overall downregulation of gene transcription and the emergence of coherent expression profiles that characterize distinct cell subpopulations. We further showed that the accessibility of promoters has a crucial effect on whether transcription factor changes will lead to alterations in the expression of their target genes. Our observations are consistent with a model based on the spontaneous probabilistic organization of the cellular process of fate commitment.

## Background

Hematopoietic cells are a widely used model for the study of fate decision and cell differentiation and it is frequently considered as a paradigm of cell differentiation in general. Differentiation is believed to proceed through a series of binary fate decisions under the action of key instructive factors inducing specific changes in the cell that lead to stepwise switches of the expression profiles at critical decision points [1]. The typical representation of this process is a hierarchical decision tree. Such a strict hierarchical process must imply tight regulation of gene expression. The genes involved in the process are well known [2]. But recent single-cell gene expression studies directly contradict the assumption of precise regulation and strictly ordered process. It has been shown that, soon after their stimulation for differentiation, multipotent CD34+ cells go through a phase of disordered gene expression called “multilineage primed” phase characterized by concomitant expression of genes typical for alternative lineages [3–6]. Other

studies demonstrated that hematopoietic stem cells (HSC) gradually acquire lineage characteristics along multiple directions without passing through discrete hierarchically organized progenitor populations [7]. Instead, unilineage-restricted cells emerge directly from a continuum of low-primed undifferentiated hematopoietic stem and progenitor cells [7]. This phase is accompanied by instabilities and fluctuation of the cell transcriptome, morphology and dynamic cell behavior [6,8]. How this quasi-random gene expression pattern is generated, how it is transformed into a defined gene expression profile remains unknown. In order to answer these questions, we determined the order and the timescale of the early chromatin and transcriptional changes that follow the induction of differentiation in CD34+ cells.

To do this, we performed single cell RNA sequencing of human cord blood CD34+ cells at different time points during the 96h period following their stimulation, a period shown to be critical for cell fate decision [6]. The gene expression profiles were correlated to the DNA accessibility changes determined by ATAC-seq at defined time-points during the same period. The experimental strategy is shown in **Fig. 1A**. The data revealed strikingly different dynamics of chromatin accessibility and gene expression that challenges the classical model based on specific stepwise switches.

## Keywords

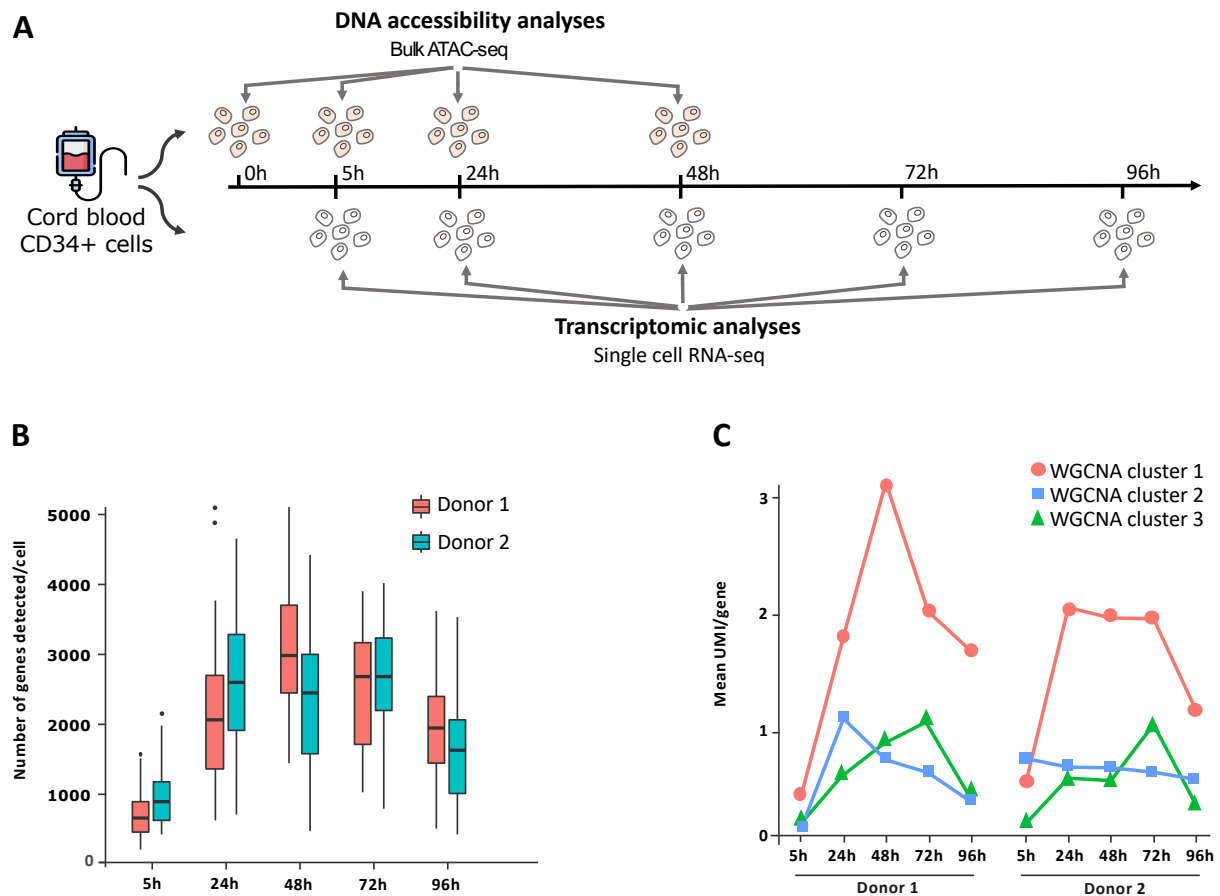
hematopoietic stem cell, fate commitment, single-cell, chromatin remodeling, ATAC-seq, scRNA-seq, transcription factor, promoter accessibility, CD34+

## Results

### Single-cell gene expression analysis using RNA-seq

Human CD34+ cells were isolated from the cord blood of two healthy donors and cultured in the presence of early acting cytokines as described [6]. To identify the transcriptional signatures and estimate their variability at the earliest stages of the differentiation process, we adapted MARS-seq protocol (massively parallel single-cell RNA-sequencing, see Methods) on CD34+ cells randomly sorted at different time points (5h, 24h, 48h, 72h and 96h) after the cells were cultured in the presence of cytokines [9]. A uniform random sampling of a heterogenous population allowed us to evaluate the global changes without any preconceived ideas on the cell categories present in the population. The quantification of gene expression was calibrated using unique molecular identifier (UMI) marked RNAs. Details about quality control of the results are shown in **Additional Table 1**. In order to avoid the potential bias due to batch correction, the results of the two donors were analyzed separately.





**Fig. 1 Gene expression dynamics of cord blood derived CD34+ cells.** **A** CD34+ cells were isolated from human cord blood and cultured in serum-free medium with early acting cytokines. Single-cell RNA sequencing (scRNA-seq) was used to analyze single-cell transcription at 5h, 24h, 48h, 72h and 96h. Concomitantly, at 0h, 5h, 24h and 48h, 5000 living cells were collected to perform ATAC-seq protocol in order to study DNA accessibility dynamics. **B** Number of detected genes per cell with scRNA-seq. Two donors were analyzed separately, both showed similar dynamics. **C** Weighted correlation network analysis (WGCNA) reveals clusters of genes with similar dynamic patterns in the average mRNA expression in Donor1 and Donor2. Note that cluster 1 reproduces the dynamic pattern observed for genes showing detectable expression in single cell in **Fig. 1B**. Cluster 1 = 5194 genes (Donor1) and 5518 genes (Donor2), cluster 2 = 3977 genes (Donor1) and 2602 (Donor2), cluster 3 = 1089 genes (Donor1) and 609 genes (Donor2).

The results revealed important features of the gene expression dynamics (**Fig. 1**). Following stimulation, the transcriptome of the cells underwent rapid and substantial quantitative and qualitative changes. Both the number of expressed genes per cell and the number of mRNA molecules per gene increased substantially (**Fig. 1B and Fig. 1C**). The average number of genes detected per cell at 5h was only 512 $\pm$ 243 in Donor1. This number increased to 1693  $\pm$ 813 at 24h and 2543 $\pm$ 751 at 48h, but then decreased to 2014 $\pm$ 714 at 72h and to 1612  $\pm$ 613 at 96h. The numbers for the cells from Donor 2 were very similar (**Fig. 1C**). The rapid increase of global transcription activity in the cells, that occurred mainly during the first 48h, suggests that cells significantly expand their repertoire of transcribed genes during the initial period of differentiation.

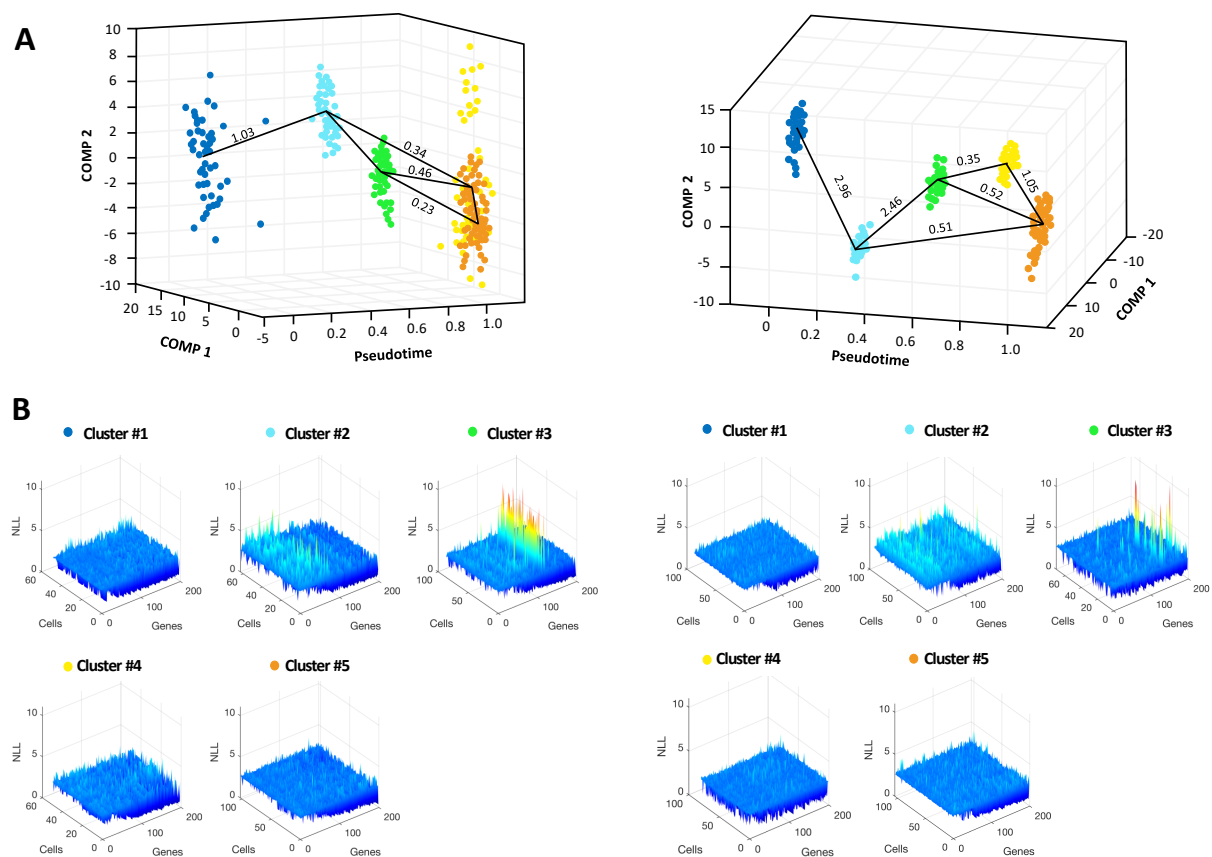
When individual genes were examined, we observed that the corresponding number of mRNA molecules also increased. Using Weighted Correlation Network Analysis (WGCNA), we found clusters of genes with highly correlated mean expression patterns over time (**Fig. 1C**). The three largest clusters together contain more than 8500 genes with mean expressions that generally show a characteristic time profile with an initial increase followed by a subsequent decrease. Thus, the average CD34<sup>+</sup> cell responds to cytokine stimulation by a strong, but transitory, upregulation of transcriptional activity both in terms of the number of genes and the number of transcripts. Due to the very high number of genes in the first cluster, gene ontology (GO) analysis was unsurprisingly irrelevant and showed significant enrichments for all basic cellular functions. Notably, none of the most represented functions were directly related to the hematopoietic lineage (**Additional File. 1**). During the 24h to 48h period after the stimulation, the fraction of the genes transcribed in individual cells raised up to approximately 10-15% of all genes in the genome (**Fig.**

**1B**). After 72h, this number started to decrease (Fig.1B). Importantly, this timing coincides with the period when the first signs of lineage-specific transcriptional changes appear [6].

In order to better characterize cell type specific gene expression patterns, lineage progression, and trajectory of the cells during the fate decision process, we applied a recent method CALISTA (Clustering And Lineage Inference in Single-Cell Transcriptional Analysis) to the single-cell RNA dataset [10]. CALISTA is a likelihood-based method that uses the two-state stochastic model of gene transcription to describe the cell-to-cell variability of gene expression at single-cell level [11]. Here, we employed CALISTA for cell clustering, lineage inference, and calculating single-cell transcriptional uncertainty. In CALISTA, to each cell is assigned a likelihood value, which reflects the joint probability of its gene expression (mRNA counts) based on the mRNA distribution from the two-states model. In order to avoid potential batch effects, we analyzed the single-cell mRNA datasets from two donors independently. For both donors, CALISTA identified five single-cell clusters on the basis of the 200 most variable genes (**Additional Fig. 1** and **Fig. 2A**). In both donors, clusters #1 and #2 were essentially composed of cells isolated at 5h and 24h, respectively (**Additional Fig. 2**). Clusters #3, #4 and #5 contained cells isolated at 48h, 72h and 96h, but with a higher degree of cells from different time points compared to clusters #1 and #2 (**Additional Fig. 1**). CALISTA generated the lineage progression using the clusters based on the distances between each pair of clusters, specifically by adding “transition” edges in the order of increasing distances and cluster pseudotimes – defined by the mode of the sampling time points of the cells in each cluster. The cluster distance between any two clusters gives a measure of dissimilarity in their gene expression distributions and is defined as the maximum difference in the cumulative likelihood values upon reassigning the cells from the original cluster to the other cluster [10]. The

inferred lineage progression graphs for each of the two donors are depicted in **Fig. 2**, showing the emergence of two distinct cell clusters with divergent transcription profiles.

Note that each gene in each individual cell in a cluster can be characterized by a unique likelihood value (see Methods). Here, we use the negative logarithm of the gene likelihood value (NLL) as a metric of transcriptional uncertainty, using which we can probe into the intra-cluster cell-to-cell heterogeneity in the gene expression [10,12]. As shown in **Fig. 2B**, the gene-wise NLLs in the clusters reveal that clusters #2 and #3 are much more heterogenous than the other clusters. Importantly, clusters #2 and #3 contain the cells that display the highest number of expressed genes and of transcripts per gene. In other words, the upregulation of global transcriptional activity in response to cytokine stimulation also causes an increase in transcriptional uncertainty and cell-to-cell heterogeneity in gene expression. Also, the peak of such transcriptional uncertainty precedes the emergence of two distinct gene expression profiles. At 72h and 96h, both the total number of transcribed genes and the number of transcripts per individual gene decrease simultaneously. In a previous study, the gradual emergence of defined expression profiles was observed after 72 hours [6], in agreement with our observations. The analysis of single-cell RNA profiles using CALISTA above demonstrated that starting from 48h, two distinct gene expression profiles start to diverge (**Fig. 2A**).

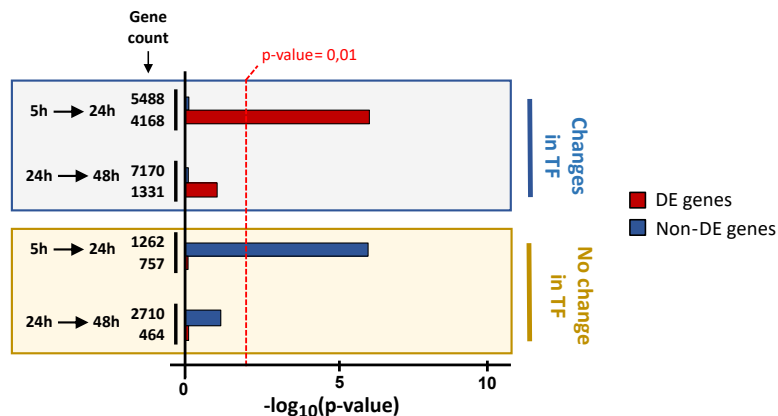


**Fig. 2 Evolution of transcriptome profiles after cell stimulation. A** Transcriptome clusters identified by CALISTA [10]. Each dot corresponds to a cell in the single-cell transcriptomic dataset of cells sampled at 5h, 24h, 48h, 72h and 96h. The x axis corresponds to the pseudotime values and the y-z axes to the first and second principal component (PC) coordinates. The color code for the clusters appears in Fig.2B. The transition edges are represented by black plain lines between the clusters and the numbers are “cluster distances”, a likelihood-based measure of dissimilarity (distance) between cell clusters. **Left panel: Donor1, right panel: Donor2. B** Negative Likelihood matrix for the 200 most variable genes computed by CALISTA for each cluster. Each plot corresponds to a cluster indicated by a color code as in A and the cluster number. **Left panel: Donor1, right panel: Donor2.**

The conclusions drawn on the basis of the general trends in single-cell gene expressions are supported by the expression of genes coding for transcription factors (TFs) essential for hematopoietic differentiation (**Additional Fig. 2**) [2]. Notably, TF-encoding genes showed highly dynamic expression by cells during differentiation. Both the fraction of expressing cells and the number of mRNAs per cell increased and went through a plateau at 48-72h. However, each individual cell expressed a different combination of these genes and no obvious dynamic patterns could be identified. The tendency toward defined hematopoietic transcription profiles can hardly be observed by the end of the time series at 96h.

In order to reveal potentially active regulatory interactions, we identified genes coding for transcription factors (TFs) that showed a change in expression (**Fig. 3**). We used the terminology “change” or “differentially expressed (DE)” to refer to genes that show a statistically significant increase or decrease in the corresponding mRNA level based on the number of UMIs detected in a cell (two-tailed Fisher exact test, see Methods for details). For each TF, we identified its target genes using human transcriptional regulatory networks from the Regulatory Circuits resource [13]. A TF may have multiple target genes, and vice versa, a gene may have several TF regulators. A gene is counted in the “changes in TF” group when at least one of the genes coding for a TF targeting it shows differential expression between 5h and 24h or between 24h and 48h. Note that the genes belonging to this group may or may not themselves be differentially expressed. A gene is counted in the “no change in TF” group when none of its TFs show any differential expression. Genes in the “changes in TF” group between 5h and 24h are significantly over-represented ( $p=1.4e-6$ ) for genes that are differentially expressed (i.e. DE genes), but not for those between 24h and 48h (**Fig. 3**). Genes in the “no change TF” group between 5h and 24h are enriched for genes showing no significant differential expression, but again not for those between 24h and 48h.

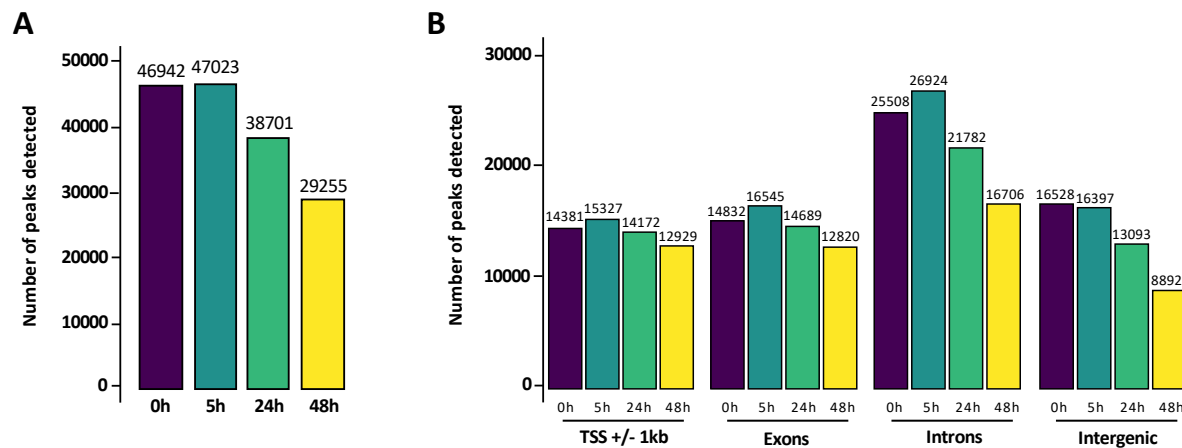
These observations indicate that differential expression of genes between 5h and 24h is connected to changes in the expression of the TF regulators, which reflects the regulatory activity of TFs. However, the regulatory activity of TFs appears to have been restricted after 24h.



**Fig. 3 Global influence of transcription factors on targeted gene expression.** Enrichment analysis of genes in the “change in TF” and “no change in TF” group for differentially expressed (DE) and non-DE genes.

## ATAC-seq analysis of DNA accessibility

DNA accessibility in CD34+ cells was determined using ATAC-seq [14] at four time points (0h, 5h, 24h, and 48h after cell stimulation). We applied a stringent filter to identify accessibility by only retaining peaks that are uniformly detected in the cells of three different donors (see **Additional Table. 2** for donor-related information). Performing ATAC-seq on 5000 cells ensured that the detected accessible DNA regions are present in a substantial fraction of cells. Indeed, accessible sites present in individual or a small number of cells could not be differentiated from the technical noise.



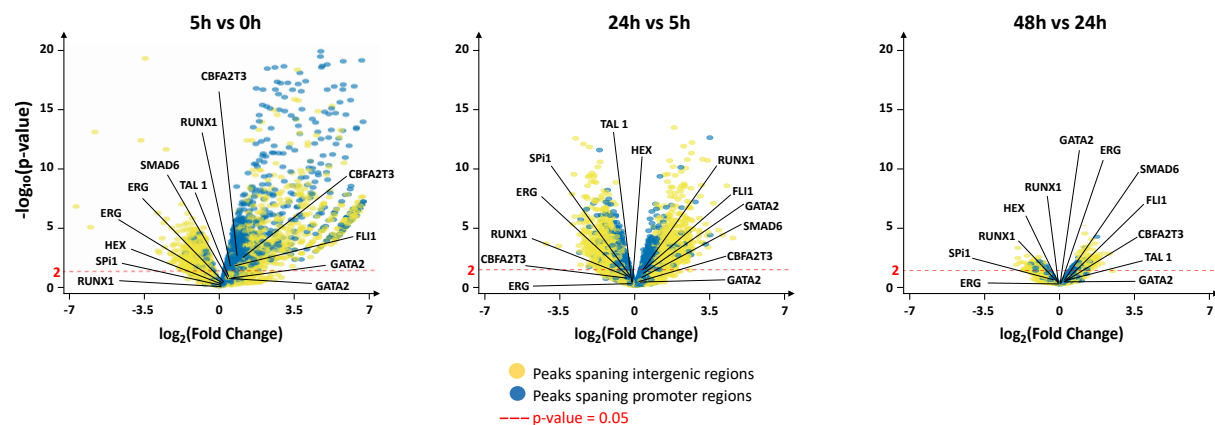
**Fig. 4 Chromatin accessibility dynamics as detected by ATAC-seq. A** Number of accessible regions (peaks) at 4 different time points. **B** Number of peaks in different genomic elements. A single peak may count for two categories if overlapping both. Intergenic category was defined as the exclusion of any other category tested.

Apart from the peaks in intergenic regions, the total number of ATAC-seq peaks first increased rapidly by 10-12% between 0h and 5h in all genomic regions, then decreased gradually at slower rate over the next 48h (**Fig. 4B**). The time-dependent decrease in the number of ATAC-seq peaks varied with their genomic location (**Fig. 4B**). While the number of peaks in distal intergenic regions was halved between 5h and 48h, the decrease in the other locations was less significant (**Fig. 4B**). Particularly, the number of peaks in promoter regions only dropped by 15% between 0h and 48h.

In order to further characterize the dynamics of the ATAC-seq, we also estimated the changes in the size of the peaks present at least at two consecutive time points. As a proxy for the size of a peak, we used the number of sequenced reads that define it. Here we assumed that the normalized number of reads can be used as a rough estimation of the fraction of the cells with at least one of the two copies of the region having accessible DNA. The difference of read counts for the same



ATAC peak detected at two consecutive time points was used to assess the chromatin dynamics. We calculated the log-fold changes of the number of reads of each peak for time intervals and the associated p-values and represented them as volcano plots (Fig. 5). We observed a significant tendency of the peaks already present at 0h to increase in accessibility by 5h (Fig. 5), especially for peaks located in the TSS regions. During this period, a total of 17% (9045 out of 53797 peaks) showed significant changes regarding accessibility. In the same range, between 5h and 24h, 15% (7505 out of 50936 peaks) of the peaks present at both time points changed significantly with an approximately equivalent number of increased and decreased ATAC-seq read counts. However, between 24h and 48h, only 2% (48 out of 40248 peaks) of the peaks showed differential read counts, but again, with roughly equal proportions of increased and decreased peaks (Fig. 5). Overall, most of the changes occurred during the first 24 hours (Fig. 5). First, we observed a rapid increase in peak number and an increase in size (read counts) for the peaks already present. Then the trend was reversed: both the number and size of the peaks decreased between 5h and 24h. This trend was maintained, albeit at a lesser degree, between 24h and 48h. Overall, the ATAC-seq observations indicate an unusually strong wave of chromatin fluctuations during the initial 48 hours long period. The dynamic fluctuations appear to be higher in intergenic regions than in gene-associated regions.



## **Fig 5. Differential analysis of ATAC-seq peaks present at least in two consecutive time**

**points.** The differential analysis is detailed in Methods. Peaks overlapping with promoter regions are highlighted in blue, while those overlapping with intergenic regions are highlighted in yellow. The promoters of the 11 hematopoietic transcription factors are indicated. Only 3 of them (RUNX1, CBFA2T3, TAL1) showed significant differential changes in accessibility between 0h and 5h. After 5h, none of the TF showed significant changes in their TSS region. Note that a TF can be displayed more than once, it is explained by the fact that a TF can have multiple TSSs, also, more than one peak can fit the TSS region.

We explored further the chromatin dynamics in promoter regions of 11 TF-encoding genes known to act as early hematopoiesis regulators [2]. With the exception of GATA1, every gene has accessible promoters with at least one ATAC-seq peak (**Fig. 5** and **Additional Fig. 3**), suggesting that even before cell stimulation with cytokines, the promoter regions of these key regulators genes are fully accessible and remain so along the entire experiment.

To further investigate the gene promoter accessibility, we analyzed the enrichment of various transcription factors binding site (TFBS) motifs among peaks. We observed that many of the TFs of factors known to play a role in hematopoiesis, such as RUNX1, ERG, PU.1 and FLI1 are already highly accessible at 0h and remain detectable at relatively the same level up to 48h (**Additional Fig. 4**). We also note that CTCF (CCCTC-binding factor) binding sites were detected more than five times more frequently among the detected peaks than in other regions, suggesting the implication of chromatin remodeling during this period [15].

## **Combined scRNA-seq and ATAC-seq Analysis**

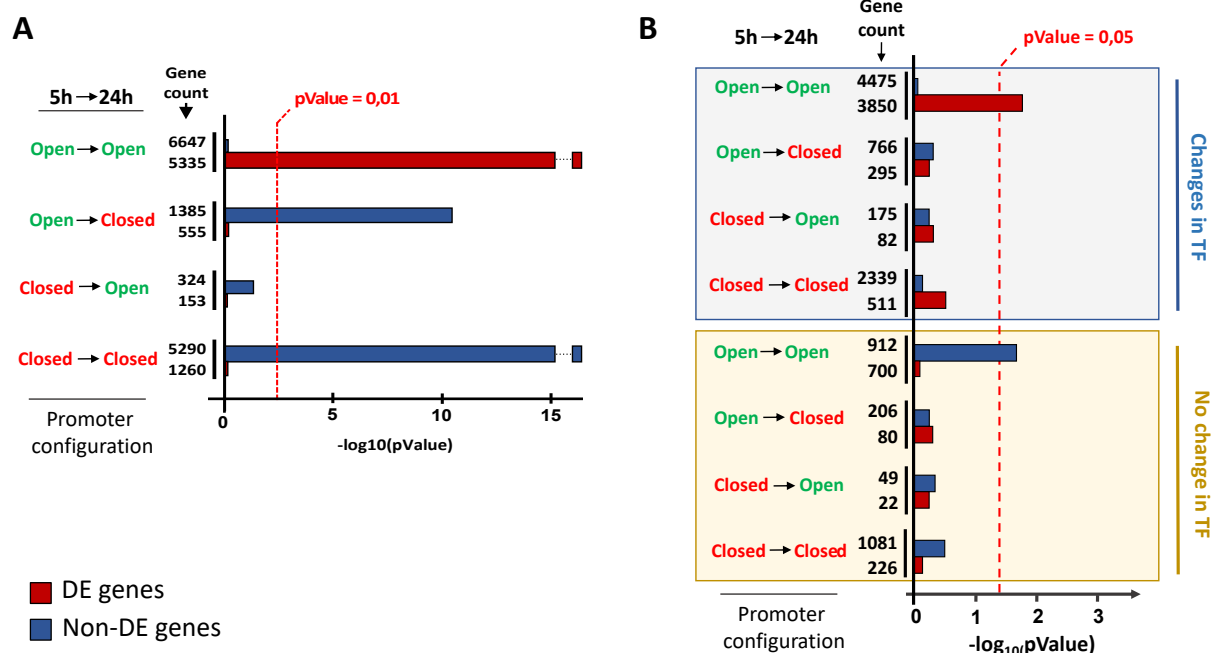
In order to elucidate how the dynamics of chromatin accessibility and the differential gene expressions were related, we combined the scRNA-seq and the ATAC-seq data (see Methods). A careful comparison of scRNA-seq and ATAC-seq analysis in **Fig. 1** and **Fig. 5** shows that the wave of opening and closing of accessible gene promoters/transcription start sites (TSSs) precedes the wave of the increase and decrease of the gene transcription. To make sense of this, first we examined how promoter accessibility of the genes is related to the gene expression. Specifically, we grouped the promoters based on the configuration change of ATAC-seq peaks between 5h and 24h (**Fig. 6A**). By doing so, there are 4 possible combinations of chromatin accessibility state: “open-open”, “open-close”, “close-close” and “close-open”, depending on the presence or absence of ATAC-seq peaks at the given promoter at 5h and 24h, respectively. The period between 5h and 24h is particularly interesting and important, because most of the changes in gene expression and DNA accessibility occur at this stage. We then identified the target genes for each promoter using the Regulatory Circuit resource (see Methods). Note that a promoter may regulate multiple genes and a gene may be regulated by several promoters. Therefore, the total numbers of promoters and genes may be different. Finally, we examined the overrepresentation of DE and non-DE genes among the groups of genes that are the target of each of the four classes of promoter configuration (i.e. open-open, open-close, close-open, and close-close). The analysis showed a significant overrepresentation ( $p\text{-value} < 10e-4$ ) for genes showing DE in the set of genes associated with the “open-open” promoter configuration (**Fig. 6A**). In the “open-close” promoter class, genes showing no DE were overrepresented (**Fig. 6A**). Non-DE genes were also overrepresented in the “close-close” promoter class.

To understand how alterations in DNA accessibility and TF expressions in combination regulated target gene transcription, we classified genes according to whether any of the TFs were differentially expressed between two different time points, 5h and 24h, as we had done earlier – that is, “changes in TF” vs. “no change in TF” grouping (details about the method are explained in **Additional Fig. 5**). We tested the overrepresentation of DE genes among the eight groups of genes based on the different classes of promoter configuration and the DE of the TFs (**Fig. 6B**). We found a significant overrepresentation for DE genes among the genes with at least one of the TFs differentially expressed, but only when the target promoters were accessible at both time points, i.e. in the “open-open” configuration. For the same “open-open” promoter configuration, genes with no change in expression between 5h and 24h were significantly overrepresented in the gene set whose TF expressions were not altered. In other terms, the regulatory action of the TFs can only be observed if and only if the promoters of the target genes are already accessible and remain so between the two time points.

Similar analysis was done between 24h and 48h (**Additional Fig. 6**). During this period, promoters that remain accessible are enriched for DE genes, while inaccessible promoters are enriched for genes showing no DE. Interestingly, the combined enrichment analysis of ATAC-seq and scRNA-seq did not show any significant over-representation for DE or non-DE genes.

Taken together, the integration of gene expression and chromatin accessibility data support the idea that differential expression of genes can be explained by changes in the expression of at least one its TFs. However, the regulatory action of TFs only applies when the promoter remains accessible. Subsequently, the closing of the chromatin on the promoters seems to prohibit the action of TFs.

In order to verify if the observations above hold for specific genes, we compared the gene expression and ATAC-seq profiles of the 11 genes essential for hematopoietic differentiation already considered above (**Additional Fig. 3**) [2]. We saw no correlation between promoter accessibility and transcription. For example, the promoter of the SMAD6 gene is accessible at all time points, still not expressed as shown on the heat-map (**Additional Fig. 2**).



**Fig. 6 Promoter configuration dynamics and transcription influence on gene expression regulation between 5h and 24h.** **A** Enrichment analysis for differentially expressed genes (DE) and non-DE genes depending on the promoter accessibility dynamics. **B** Enrichment analysis differentially expressed genes (DE) and non-DE genes depending on the promoter accessibility dynamics and the changes of the expression of TF-encoding genes that regulate them (two-tailed Fisher exact test, see Methods for details).

These observations clearly reveal the unequal role of the chromatin configuration and TF action on gene expression and explain the chronology of why the initial increase in DNA accessibility precedes the burst in the level and diversity of gene transcription. Since TF action alone is unable to make the promoters accessible, the initial opening of the chromatin has to be a global and non-specific event. This explains why only a subset of genes with initially accessible promoters become transcribed during the later stages.

## Discussion

*In vitro* cultured human cord blood derived CD34+ cells are usually considered as a heterogenous population of cells. Recent studies demonstrated that this heterogeneity is not the result of the mixture of different cell types, but a population of cells with a wide distribution of gene expression patterns [7] that fluctuate, generating morphological instability [6]. The first morphological and molecular signs of the phenotypic diversification appear at the end of the unusually long first cell cycle that follows cytokine stimulation [6]. During the first cell cycle, each cell displays a rather distinct gene expression pattern but is usually morphologically similar. By 48 to 72 hours, one can observe the emergence of two different cellular morphologies and two different characteristic transcription profiles [6]. Such an observation prompted us to investigate the narrow window of time within 48h in more details. The observations reported here reveal the interplay between the dynamic chromatin and gene expression changes.

Using ATAC-seq, we detected at 0h more than 46000 peaks, about 30% of them in gene promoters (TSSs). The number of detected TSSs increased sharply during the first 5 hours of culture (Fig. 4). At the 5h time point, more than 50% of all TSSs promoters in the genome displayed accessible

DNA (**Fig. 4B**). After the rapid initial increase, the number of the peaks started to decrease (**Fig. 4**). There were approximately 16% less accessible TSSs 48 hours later. The number of open intronic and intergenic genomic regions decreased even more rapidly and fell to 50% of the initial number. The tendency to chromatin closing therefore appears as a general feature. Importantly, the wave of chromatin opening and closing is followed by a wave of transcriptional activity. The variety of the transcribed genes and the number of the mRNA molecules per gene was the lowest at 5h – the first time point tested for scRNA-seq – but both increased sharply at 24h, reached a plateau between 48h and 72h and decreased at 96h (**Fig. 1B** and **Fig. 1C**). The 5h-to-48h period corresponds to the multilineage-primed stage of the CD34+ cells that precedes the emergence of the first signs of characteristic gene expression patterns accompanying differentiation [6]. It is a universal feature of the cells during the initial phases of the fate commitment process to progress through a transitional cell state marked by the rise-then-fall in transcriptional uncertainty and a concomitant rise-and-fall of cell-to-cell variability [12]. As reported here, the gene transcription in the CD34+ cells clearly follows the same pattern. The global increase of transcription is preceded by a widespread and non-specific chromatin opening that makes accessible more than 50% of gene promoters in the genome.

Importantly, there is a strong stochastic component in the establishment of the multilineage primed expression state, because the number of gene promoters that are accessible exceeds the number of actually transcribed genes in each cell by 3 to 5 times (**Fig. 1B** and **Fig. 4B**). The emergence of coherent transcription profiles from this heterogeneous transitory state is preceded by chromatin rearrangements. A significant fraction of gene promoters (16%) and intergenic sites (46%) in the genome become inaccessible through chromatin closing between 5h and 48h (**Fig. 4B**). The stabilization of the transcriptome is presumably the consequence of these chromatin

changes. Some promoters gradually become repressed by chromatin closing, while others are stabilized in an open chromatin configuration. The role of TFs appears crucial at this stage. Indeed, the transcription of a gene is changed between 5h and 24h if the expression of TF-encoding genes that regulate them also changes. However, changes of the expression of the TF-encoding genes do not lead to alteration of their target gene expression if the promoters are in “closed” chromatin configuration around the TSS (**Fig. 6B**), indicating that TFs alone are not able to efficiently regulate the gene transcription, and the chromatin accessibility is a pre-requisite for TF action. Since the number of the open promoters is high at the beginning of the process, a competition for the available TFs among accessible promoters may explain the transcriptional and phenotypic fluctuations observed during this period [6]. These fluctuations cease when the transcriptome is stabilized [6].

The proposed scenario of general chromatin destabilization followed by a selective repression of the genes is also supported by the observations showing that the inhibition of chromatin compaction using valproic acid (VPA), a histone deacetylase inhibitor, can maintain the multilineage-primed state with promiscuous transcription profile for a long period [6,8,16]. The removal of VPA allows defined transcriptome profiles to be established [8]. Therefore, chromatin structural changes appear to be causally involved both in the generation of a multilineage-primed state and the stabilization of cell fate choice. In line with this conclusion, a recent study of human fetal hematopoietic cells has also concluded that extensive epigenetic but not transcriptional priming of HSC/MPPs occurs prior to lineage commitment [17].

It will be of particular importance to investigate the process of transcriptome stabilization and the feedback mechanisms that must certainly accompanied it. In this respect, a dynamic positive



feedback loop between permissive chromatin and translational output has been previously reported for embryonic stem- and in CD34+ cells [18]. It is noteworthy that many of the genes with the most variable expression that contribute significantly to the specification of the emerging transcription patterns are ribosomal protein (RP) coding genes (**Additional File. 2**), thus impacting the process of translation [19]. A high degree of RP expression heterogeneity has already been observed in hematopoietic cells, where a small subset of RPs can discriminate cell types belonging to different hematopoietic lineages [20]. Therefore, it is possible that, in addition to the TF and promoter interactions, a feedback action of the translational output may also contribute to the stabilization of the chromatin.

The observed non-specific chromatin opening and the rise of an equally non-specific gene expression as a first step, followed by a slow relaxation toward a defined gene expression pattern and chromatin stabilization, brings a new perspective to our understanding of how cell fate commitment is initiated. According to the conventional view, a switch-like activation of fate-specifying genes, followed by a cascade of activation of specific downstream targets determines cell fate. This view is not compatible with the observations reported here. The alternative possibility is that the typical expression pattern of a committed cell results from the stabilization of a network of interacting set of select genes through a transitory multilineage-primed state that is characterized by stochastic and highly variable expression profile. The transitory stage emerges as a rapid and non-specific answer to a substantial change in the cell's environment that is analogous to the physiological stress response whose role is to prepare the organism to meet new and unforeseen circumstances [21]. Here, we observed a general and non-specific opening of the chromatin that lifts the transcription repression and permits targeted interactions between TFs and gene promoters and enhancers. Put in another way, the quasi-random activation of genes in a cell

under stressful conditions generates a potential of a variety of phenotypic traits in the cell. Some of these traits promote the cell's survival under the new constraints imposed by the evolving microenvironment, and they are selectively stabilized by feedback mechanisms. These mechanisms are not yet identified, but explicit hypotheses have been made [22,23]. Therefore, the process of choice can be viewed as a continuing iterative process of constrained optimization of the cell phenotype over time, a kind of “learning process” that is accomplished by the cell through interactions and cooperation with the surrounding cells and environment. This way to frame the question of fate commitment has been theorized long ago [24–26], and single-cell studies in the recent years have provided more and more experimental support [3,6,12,27,28].

## Conclusions

In the present study we show that chromatin accessibility and gene expression follow different dynamics. Most of the gene promoters become accessible immediately after stimulation of the cells. The non-specific chromatin opening is followed 24 h later by a wave of high and unrestrained gene expression. Each cell has disordered and unique expression profile. However, the DNA accessibility at the gene promoters starts to decrease rapidly. It is followed by the decrease of gene expression and the slow emergence of two distinct profiles by the end of the period. This is likely to be the result of a selective repression process because the evolution of the gene expression profile goes from the general toward more specific. This corresponds to the gradual acquisition of two different morphological forms in the cell population.

## Methods

## Cell culture

Umbilical cord blood from anonymous healthy donors was obtained from Centre Hospitalier Sud Francilien, Evry, France or from Etablissement Français du Sang (EFS), Saint Louis Hospital, Paris, France. Mononuclear cells were isolated from cord blood fractions by density centrifugation using Ficoll (Biocoll, Merck Millipore). Human CD34<sup>+</sup> cells were then enriched in the sample by immunomagnetic beads using an AutoMACSpro (Miltenyi Biotec). After collection, enriched CD34<sup>+</sup> cells were frozen in a cryopreservation medium containing 90% of fetal bovine serum (Eurobio) and 10% of dimethylsulfoxide (Sigma) and stored in liquid nitrogen.

After thawing, the CD34<sup>+</sup> cells were cultured in a 96-well plate in a humidified 5% CO<sub>2</sub> incubator at 37°C. Cells were cultured in prestimulation medium made of X-Vivo (Lonza) supplemented with penicillin/streptomycin (respectively 100U/mL and 100ug/mL - Gibco, Thermo Scientific), 50 ng/ml h-FLT3, 25 ng/ml h-SCF, 25 ng/ml h-TPO, 10 ng/ml h-IL3 (Miltenyi) final concentration.

## Fast-ATAC-seq

We used Fast ATAC-seq with minor modifications. This protocol was optimized for blood cells [14]. Prior to transposition, cells were marked with 7AAD and dead cells were removed by FACS (Beckman Coulter). Removing dead cells is an important parameter to ensure clear nucleosome patterns and to improve signal to noise ratio. 5000 living cells were used at each time point. A one-step gentle membrane permeabilization and DNA transposition was performed by adding 50ul transposition mixture (25 uL TD buffer 2X, 2,5uL of transposase TDE1 (Illumina), 0,5 uL

digitonin 0,1% (Promega) and 22 uL water) to the cell pellets and by incubating at 37°C for 30 minutes under agitation. Obtained Transposed DNA were then purified using MinElute PCR Purification Kit (Qiagen) and preamplified using Nextera barcoded primers (Illumina) and NEBNext High-Fidelity 2xPCR Master Mix (New England Biolabs) for 5 cycles. A quantitative PCR amplification was made on 5uL of the sample with SYBR Green to determine the number of additional cycles in order to generate libraries with a minimal number of PCR cycles and to limit PCR bias (according Corces et al [14]). Appropriate number of PCR cycles were applied on the rest of the pre-amplified samples. PCR fragments were purified with MinElute PCR Purification Kit (Qiagen) to get rid of unused primers. A supplemental purification step was performed using Ampure beads kit (Beckman Coulter) to size-select DNA fragments ranging between 100 and 700 pb. ATAC-seq libraries were checked for quality using Bioanalyzer (Agilent) prior to sequencing and sequenced in paired-end mode (2x50bp) on the Illumina HiSeq2500 platform.

## Single-cell RNA sequencing adapted from MARS-seq

To perform scRNA-seq, we adapted MARS-seq protocol (Massively parallel single-cell RNA sequencing) [9]. CD34+ cells were stained with 7AAD to only work living cells and cells were isolated by FACS. Individual cells were sorted into 96-well plates containing 4uL of lysis buffer with specific barcoded RT primers (final concentration: 0,2% Triton, 0,4 U/uL RNaseOUT (Thermofisher Scientific), 400nM idx\_RT\_primers). Idx\_RT\_primers (see **Table. 1**) contain a T7 RNA polymerase promoter for further *in vitro* transcription (IVT), single cell barcodes (**Additional. File. 3**) for subsequent de-multiplexing and unique molecular identifiers (UMIs) allowing correction for amplification biases. After cell sorting, plates were immediately centrifuged and put into dry ice before storage at -80°C preceding the reverse transcription (RT). To open

RNA secondary structure, plates containing single cells were incubated at 72°C for 3 minutes and immediately put in ice. 4uL of RT mix were added in each well (final concentration of RT mix: 20mM DTT, 2mM dNTP, 2X First stranded buffer, 5 U/uL Superscript III RT enzyme, 10% (W/V) PEG 8000). PEG8000 was added in the RT mix because it has been shown that it can increase the cDNA yield in scRNA sequencing [26]. ERCC RNA spike-in mix (Thermo Scientific) was also added to the solution for further amplification quality filtering (dilution 1/40.10e7). The plate was then put into thermocycler (thermocycler program: 42°C-2min, 50°C-50min, 85°C-5min, 4°C hold).

After first retro-transcription, samples were pooled (see Jaitin et al [9]) and ExonucleaseI digestion was performed, followed by 1,2X AMPure beads purification kit (Beckman Coulter) to keep only retro-transcribed single strand cDNA. Samples were eluted in 17uL of 10mM Tris-HCl, pH=7,5. Second strand cDNA synthesis (SSS) using NEBNext mRNA second strand synthesis module kit was then performed (SSS mix: 2uL 10x SSS buffer, 1uL SSS enzyme; thermocycler program: 16°C-150min, 65°C-20min, 4°C hold). Obtained cDNA was linearly amplified by overnight IVT (HighScribe T7 High Yield RNA synthesis, NEB) at 37°C under T7 promoter. The product was purified with 1,3X Ampure beads and eluted in 10uL of 10mM Tris-HCl, 0,1mM EDTA. 9uL of amplified RNA were then enzymatically fragmented with 1uL of 10x RNA fragmentation reagents (Thermofisher Scientific) in 70°C for 3 min. The fragmentation was stopped with 34uL of STOP mix (1,2uL Stop solution, 26,4uL AMPure beads, 9,8uL TE) and samples were purified. Differing from original MARSseq protocol, the second RT was done with primers (P5N6\_XXXX, **Table. 1**) containing random hexamers and specific barcode (**Additional. File. 3**) to distinguish the different plates (final concentration: 5mM DTT, 500uM dNTP, 10uM P5N6\_XXXX, 1X First stranded buffer, 10U/uL Superscript III RT enzyme, 2U/uL RNaseOUT; thermocycler program:

25°C 5min, 55°C 20min, 70°C 15min, 4°C hold). cDNA was purified with 1,2x AMPure beads and eluted in 10uL.

As for ATAC-seq, the appropriate number of PCR cycles was determined using a fraction of the library with SYBR Green based qPCR as described in Zilionis et al [27] (final concentration: 1x Kapa Hifi HotStart PCR mix, 1x SybrGreen, 0,5uM mix primer P5.Rd1/P7.Rd2 (Table.1); Thermocycler program: 95°C 3min – 40cycles: 98°C 20sec, 57°C 30sec, 72°C 40sec – 72°C 5min, 4°C hold). After PCR amplification, libraries were purified with 0,7x AMPure beads. Libraries were checked for quality, using Bioanalyzer HighSensitivity DNA (Agilent) prior to sequencing. Libraries were finally sequenced in paired-end mode (2x50bp) on Illumina HiSeq2500 platform.

**Table 1: structure of primer sequences used in scRNA-seq.**

Primer name	Sequence (5' to 3')
<b>Idx RT primers</b>	5'-CGATTGAGGCCGGTAATACGACTCACTATAGGGGCGACGTGTGCTCTTCCGATCTXXXXXXXXNNNNNTTTTTTTTTTTTTTTN3'
<b>P5N6. XXX</b>	5'-CTACACGACGCTCTTCCGATCTXXXXNNNNNN-3'
<b>P5.Rd1</b>	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
<b>P7.Rd2</b>	5'-CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'

Idx RT primers: TTTTTTTTTTTTTTTTTTTN = poly-T allowing matching with mRNA poly-A tail, NNNN = 4 bases UMI (randomly generated), XXXXXX = 6 bases cell barcode (**Additional. File. 3**). The rest of the sequence consists of a PCR adaptor and a T7 promoter sequence for further IVT amplification. P5N6 XXX: NNNNNN = random hexamer allowing the capture of the fragmented IVT amplified RNA, XXXX = 4 bases “plate barcode” (**Additional. File. 3**). The rest of the sequence consists of a PCR adaptor. P5.Rd1/P7.Rd2 : P5 and P7 Illumina sequencing adaptors.

550

## 551 **Bioinformatic analysis**

552

### 553 **Bulk ATAC-seq**

554

#### 555 Raw data processing

556

557 Tn5 adapters sequences were first trimmed with the following command:

558 `< cutadapt -q 20 -g "AGATGTGTATAAGAGACAG; max_error_rate=0.1; min_overlap = 10" -A`  
 559 `"AGATGTGTATAAGAGACAG; max_error_rate = 0.1; min_overlap = 10" --minimum-length 18 --`  
 560 `times 2 --pair-filter = both >`

561

562 Genome alignment (hg19) was performed using Bowtie2 with the following parameters:

563 `< bowtie2 -x hg19 --no-unal -X 800 >`

564

565 Only Paired-End fragments were kept, considering mapping quality (phred score = 30). Duplicated  
 566 reads were removed using Picard MarkDuplicates tool. In attempt to not bias the signal recovered  
 567 after peak calling due to multiple donors, all paired-end files were randomly downsampled to 16M  
 568 reads (without disrupting pairs of reads) as regard to the smallest number of reads detected in the  
 569 cohort (Donor 1 – 0h, see **Additional Table. 2**).

570

571 ATAC-seq peaks were then called on those downsampled files using:

572 `< macs2 callpeak -f BAMPE -g hs -B --broad --broad-cutoff 0.1 --keep-dup all >`

573

In order to retain only significant accessibility peaks across samples, each list of peaks used in advanced analysis has been defined as the intersection between peaks of the 3 donors tested at the same time point.

#### Peak annotation

Peaks were assigned to genomic regions thanks to a home-made script based on the FindOverlap function from the R package “GenomicRanges” [29]. Genomic elements positions (TSS, exons, introns, CpG islands and CTCF) were retrieved from UCSC database (hg19). Intergenic category was defined as the exclusion of all other defined categories. No priority has been set across the different genomic elements. Therefore, peaks overlapping several genomic features are counted multiple times, resulting in a total number of peaks across elements exceeding the total number of peaks detected at each time point.

#### Peak differential analysis

DEseq2 tool was used to calculate difference in read count between peaks in two consecutive time points [30]. More precisely, the region considered is defined as the interval formed by the union of two overlapping peaks at  $t_2$  and  $t_1$ .

#### Motif enrichment

Peak motif enrichment analysis was conducted with the tool “findMotifsGenome.pl” from the HOMER software tool suite [31]. Background file was generated using an auto-generated list of



random regions across the genome (hg19). Motifs were scanned using the total length of our peaks by providing the option *<size given>*.

## Single-cell RNA-seq (scRNA-seq)

### Raw data processing

Cell and plate barcode demultiplexing steps were accomplished under strict selection criteria with the following command:

```
< cutadapt -q 30 -e 0 -m 30:20 --no-trim --no-indels --pair-filter = any >
```

Fasta files for both barcodes (cells and time) sequences are given in **Additional File. 3**.

ERCC mapping was performed using bowtie2 [32] on ERCC known sequences and regular mapping was performed using STAR [33] on the reference genome version hg19 and aligned reads annotated. After quality filtering, reads and UMIs count per gene and ERCC were calculated for expression analysis.

### Cell and gene filtering

Chromosome Y was removed from the analysis to avoid unwanted effects and only protein coding genes were kept for further analysis. Cells with less than 80 000 total reads were removed, as well as cells with more than 10% of reads corresponding to mitochondrial RNA. To reduce undesired effect due to PCR non-linear amplification, ERCC spikes were used to assess the linearity of amplification. Pearson correlation coefficient was calculated for each cell, and only cells above 0,6

were retained. For each cell remaining, genes were defined as detectable if at least two cells contained more than a single UMI (=transcript) and a minimum of 5 reads in total.

#### Single-cell clustering and variability analysis

Clustering analysis was performed with CALISTA (Clustering and Lineage Inference in Single-Cell Transcriptional Analysis), a numerically efficient and highly scalable toolbox for end-to-end analysis of single-cell transcriptomic profiles. This approach includes single-cell mRNA counts in a probabilistic distribution function associated with stochastic gene transcriptional bursts and random technical dropout events. In the data pre-processing, we removed cells with more than 95% of zero expression values and we selected the top 200 most informative genes for further analysis. The optimal number of clusters was chosen to be five based on the eigengap plot (see [10] for more details).

#### WGCNA

We applied Weighted Correlation Network Analysis (WGCNA) [34] to mRNA expression data from each donor, to identify modules of genes with similar gene transcriptional dynamics. We excluded genes without any detectable expression in all samples. In implemented WGCNA, we set the soft-thresholding power for a scale-free topology index of 0,9. For each module, we calculated the mean expression of genes by averaging the UMI counts from the two donors.

#### Enrichment Analysis

We obtained a curated collection of TFs to CAGE-defined promoters to gene isoform mapping for a total of 662 human TFs from the Regulatory Circuits resource [13]. In our analysis, we used only TF – Promoter pairs with moderate confidence scores  $> 0,5$ . We grouped genes based on whether the relevant TFs demonstrated differential expressions. More specifically, a classification of “changes in TF” was given to any gene in which one of its TFs showed a differential expression. Otherwise, a classification of “no change in TF” was assigned. A Fisher exact test was used to perform over- and under-representation analysis [35].

## **ATAC-seq and scRNA-seq combined analysis (accessibility – expression)**

### Identification of Promoters that have configurational changes

In an effort to identify promoter regions that are affected (and not affected) by configurational changes of the chromatin, we used the CAGE-defined promoters to gene isoform mapping from the Regulatory Circuits resource [13] to identify the promoters that overlap with the peaks of ATAC-seq and the corresponding target genes (see **Additional Fig. 5B**). For this purpose, we employed the R Bioconductor package “GenomicRanges” [29]. By comparing the peaks overlapping the promoters between two time points (5h – 24 h and 24h – 48h), we grouped promoters into 4 possible chromatin accessibility configurations: “open-open”, “open-close”, “close-open”, and “close-close”.

### Differential gene expression of single-cell RNA sequencing

We computed Z-scores for every gene in each of the two donors between two different time points using the mean and standard deviation of the UMI counts of approximately 100 single cells.

$$Z_{ij}^{t_2-t_1} = \frac{\text{mean}(UMI_j^{t_2}) - \text{mean}(UMI_j^{t_1})}{\left( \left( \text{sd}(UMI_j^{t_2}) \right)^2 + \left( \text{sd}(UMI_j^{t_1}) \right)^2 \right)^{1/2}}$$

$Z_{ij}^{t_2-t_1}$  denotes the Z-score of the expression change of gene  $j$  in donor  $i$  between time  $t_2$  and  $t_1$ . An average Z-score between the two donors was computed and used to identify the set of differentially expressed genes. We selected a threshold Z-score of 2 and -2 (i.e., two standard deviations of change) to designate upregulated and downregulated genes, respectively. Collectively, they represent the set of differentially expressed genes.

### Enrichment Analysis of Combined ATAC-seq and scRNA-seq

For the combined ATAC- and scRNA-seq analysis, we grouped genes into 8 possible groups based on the chromatin accessibility configurations (i.e., one of the following four configurations: “open-open”, “open-close”, “close-open”, and “close-close”) and whether any one of their TFs showed differential expression (i.e., one of the following two groups: “changes in TF” and “no change in TF”) (see **Additional Fig. 5C**). As with the analysis of scRNA-seq data alone, a gene was assigned to the group “changes in TF” when at least one of its TFs showed differential expression; otherwise, the gene was classified as “no change in TF”. Note that different isoforms of the same gene can have distinct TSSs that are under the control of different promoters. Thus, a gene might

be counted in more than one category in the chromatin accessibility configurations. Consequently, the total sum of the genes in the 8 groups as described above might exceed the total number of genes. A Fisher exact test was used to perform over- and under-representation analysis [35].

## **Declarations**

## **Ethic statement**

Human cord blood (UCB) was collected from placentas and/or umbilical cords obtained from Etablissement Français du Sang (EFS), Saint Louis Hospital, France or from Centre Hospitalier Sud Francilien, Evry, France in accordance with international ethical principles and French national law (bioethics law n°2011-814) under declaration N° DC-201-1655 to the French Ministry of Research and Higher Studies.

## **Availability of data and materials**

Data are available under the NCBI GEO accession number GSE156735.

## **Competing interests**

None

## **Funding**

This work was supported by EPHE (11REC/BIMO), ANR grant ANR-17CE12-0031-01  
«SinCity ».

## Author contribution

AP, AM, RP and JFD designed the study.  
RP, AM, LR and SC conducted the experiments.  
AM, RS, RG and NPG performed CALISTA analysis.  
RP, DS, RS and RG analyzed the ATAC-seq data.  
RP, AM, LR, SC, RS, RG, NPG, DS, GC, JC and AP analyzed the results and performed  
statistical analysis.  
RP, AM and LR prepared the figures.  
AP, RP and LR wrote the paper with the help of their colleagues.

## Acknowledgements

The authors are grateful to Olivier Gandrillon, Camille Fourneaux for helpful discussions and  
Sunil Laxman and Olivier Gandrillon for the critical reading of the manuscript. The authors are  
also grateful to Sophie Foulon for her help in scRNA sequencing protocol design.

## Reference

1. Kawamoto H, Katsura Y. A new paradigm for hematopoietic cell lineages: revision of the  
classical concept of the myeloid-lymphoid dichotomy. Trends Immunol. Trends Immunol; 2009.

735 p. 193–200.

736 2. Sive JI, Göttgens B. Transcriptional network control of normal and leukaemic haematopoiesis.

737 Exp. Cell Res. Academic Press Inc.; 2014. p. 255–64.

738 3. Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C, et al. Multilineage gene

739 expression precedes commitment in the hemopoietic system. Genes Dev. Cold Spring Harbor

740 Laboratory Press; 1997;11:774–85.

741 4. Nimmo RA, May GE, Enver T. Primed and ready: Understanding lineage commitment

742 through single cell analysis. Trends Cell Biol. Elsevier Ltd; 2015. p. 459–67.

743 5. Pina C, Fugazza C, Tipping AJ, Brown J, Soneji S, Teles J, et al. Inferring rules of lineage

744 commitment in haematopoiesis. Nat Cell Biol. Nat Cell Biol; 2012;14:287–94.

745 6. Moussy A, Cosette J, Parmentier R, da Silva C, Corre G, Richard A, et al. Integrated time-lapse

746 and single-cell transcription studies highlight the variable and dynamic nature of human

747 hematopoietic cell fate commitment. PLoS Biol. Public Library of Science; 2017;15.

748 7. Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, et al. Human haematopoietic

749 stem cell lineage commitment is a continuous process. Nat Cell Biol. Nature Publishing Group;

750 2017;19:271–81.

751 8. Moussy A, Papili Gao N, Corre G, Poletti V, Majdoul S, Fenard D, et al. Constraints on

752 Human CD34+ Cell Fate due to Lentiviral Vectors Can Be Relieved by Valproic Acid. Hum

753 Gene Ther [Internet]. Mary Ann Liebert Inc.; 2019 [cited 2020 Jun 1];30:1023–34. Available

754 from: <https://www.liebertpub.com/doi/10.1089/hum.2019.009>

755 9. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively

756 parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science (80-

757 ) [Internet]. American Association for the Advancement of Science; 2014 [cited 2020 Aug

758 21];343:776–9. Available from: <https://science.sciencemag.org/content/343/6172/776>

- 759 10. Papili Gao N, Hartmann T, Fang T, Gunawan R. CALISTA: Clustering and LINEAGE  
760 Inference in Single-Cell Transcriptional Analysis. *Front. Bioeng. Biotechnol. Frontiers Media*  
761 *S.A.*; 2020. p. 18–18.
- 762 11. Peccoud J, Ycart B. Markovian modeling of gene-product synthesis. *Theor Popul Biol.*  
763 *Academic Press*; 1995;48:222–34.
- 764 12. Gao NP, Gandrillon O, Paldi A, Herbach U, Gunawan R. Universality of cell differentiation  
765 trajectories revealed by a reconstruction of transcriptional uncertainty landscapes from single-cell  
766 transcriptomic data. *bioRxiv [Internet]. Cold Spring Harbor Laboratory*; 2020 [cited 2020 Jun  
767 1];2020.04.23.056069. Available from:  
768 <https://www.biorxiv.org/content/10.1101/2020.04.23.056069v1>
- 769 13. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific  
770 regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods.*  
771 *Nature Publishing Group*; 2016;13:366–70.
- 772 14. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-  
773 specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia  
774 evolution. *Nat Genet. Nature Publishing Group*; 2016;48:1193–203.
- 775 15. Ohlsson R, Bartkuhn M, Renkawitz R. CTCF shapes chromatin by multiple mechanisms:  
776 The impact of 20 years of CTCF research on understanding the workings of chromatin  
777 [Internet]. *Chromosoma. Springer*; 2010 [cited 2020 Aug 29]. p. 351–60. Available from:  
778 </pmc/articles/PMC2910314/?report=abstract>
- 779 16. Chaurasia P, Gajzer DC, Schaniel C, D’Souza S, Hoffman R. Epigenetic reprogramming  
780 induces the expansion of cord blood stem cells. *J Clin Invest. American Society for Clinical*  
781 *Investigation*; 2014;124:2378–95.
- 782 17. Ranzoni AM, Tangherloni A, Berest I, Riva SG, Myers B, Strzelecka PM, et al. Integrative



783 Single-cell RNA-Seq and ATAC-Seq Analysis of Human Foetal Liver and Bone Marrow  
784 Haematopoiesis. bioRxiv. Cold Spring Harbor Laboratory; 2020;2020.05.06.080259.

785 18. Bulut-Karslioglu A, Macrae TA, Oses-Prieto JA, Covarrubias S, Percharde M, Ku G, et al.  
786 The Transcriptionally Permissive Chromatin State of Embryonic Stem Cells Is Acutely Tuned to  
787 Translational Output. *Cell Stem Cell*. Cell Press; 2018;22:369-383.e8.

788 19. Guo H. Specialized ribosomes and the control of translation [Internet]. *Biochem. Soc. Trans.*  
789 Portland Press Ltd; 2018 [cited 2020 Aug 29]. p. 855–69. Available from:  
790 <https://pubmed.ncbi.nlm.nih.gov/29986937/>

791 20. Guimaraes JC, Zavolan M. Patterns of ribosomal protein expression specify normal and  
792 malignant human cells. *Genome Biol* [Internet]. BioMed Central Ltd.; 2016 [cited 2020 Jun  
793 1];17:236. Available from: [http://genomebiology.biomedcentral.com/articles/10.1186/s13059-](http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1104-z)  
794 [016-1104-z](http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1104-z)

795 21. Braun E. The unforeseen challenge: From genotype-to-phenotype in cell populations.  
796 *Reports Prog Phys* [Internet]. Institute of Physics Publishing; 2015 [cited 2020 Aug  
797 29];78:036602. Available from: [https://iopscience.iop.org/article/10.1088/0034-](https://iopscience.iop.org/article/10.1088/0034-4885/78/3/036602)  
798 [4885/78/3/036602](https://iopscience.iop.org/article/10.1088/0034-4885/78/3/036602)

799 22. Paldi A. Stochastic gene expression during cell differentiation: Order from disorder?  
800 [Internet]. *Cell. Mol. Life Sci*. Springer; 2003 [cited 2020 Jul 23]. p. 1775–8. Available from:  
801 <https://link.springer.com/article/10.1007/s00018-003-23147-z>

802 23. Páldi A. Random walk across the epigenetic landscape. *Phenotypic Switch*. Elsevier; 2020. p.  
803 53–76.

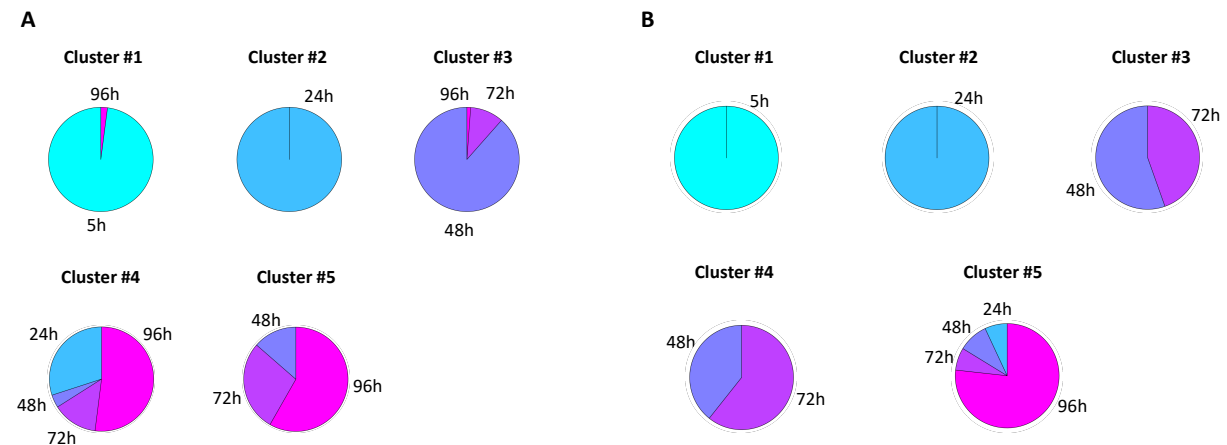
804 24. Kupiec JJ. A chance-selection model for cell differentiation. *Cell Death Differ*. England;  
805 1996;3:385–90.

806 25. Kupiec JJ. A Darwinian theory for the origin of cellular differentiation. *Mol Gen Genet*. Mol

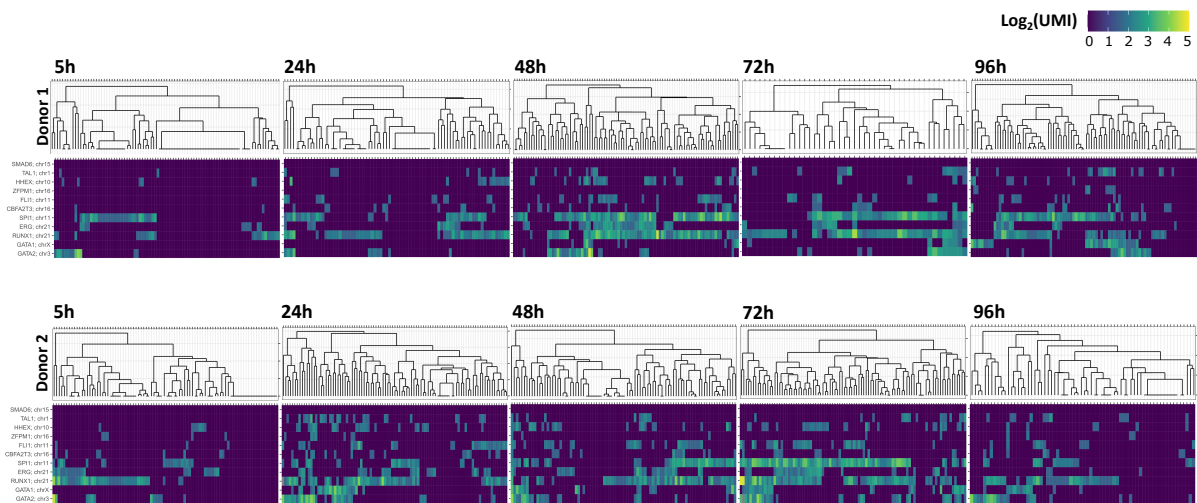
- Gen Genet; 1997;255:201–8.
26. Paldi A. What makes the cell differentiate? Prog Biophys Mol Biol. Pergamon; 2012;110:41–3.
27. Mojtahedi M, Skupin A, Zhou J, Castaño IG, Leong-Quong RYY, Chang H, et al. Cell Fate Decision as High-Dimensional Critical State Transition. PLOS Biol [Internet]. Public Library of Science; 2016 [cited 2020 Jun 4];14:e2000640. Available from: <https://dx.plos.org/10.1371/journal.pbio.2000640>
28. Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, et al. Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. PLoS Biol. Public Library of Science; 2016;14:e1002585.
29. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. PLoS Comput Biol [Internet]. Public Library of Science; 2013 [cited 2020 Jun 1];9:e1003118. Available from: <https://dx.plos.org/10.1371/journal.pcbi.1003118>
30. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. BioMed Central Ltd.; 2014 [cited 2020 Jun 1];15:550. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>
31. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. Mol Cell; 2010;38:576–89.
32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. NIH Public Access; 2012;9:357–9.

- 831 33. STAR: ultrafast universal RNA-seq aligner [Internet]. [cited 2020 Jun 1]. Available from:  
832 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/>
- 833 34. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis.  
834 BMC Bioinformatics [Internet]. BioMed Central; 2008 [cited 2020 Jun 1];9:559. Available from:  
835 <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559>
- 836 35. Agresti A. An Introduction to Categorical Data Analysis Second Edition.  
837

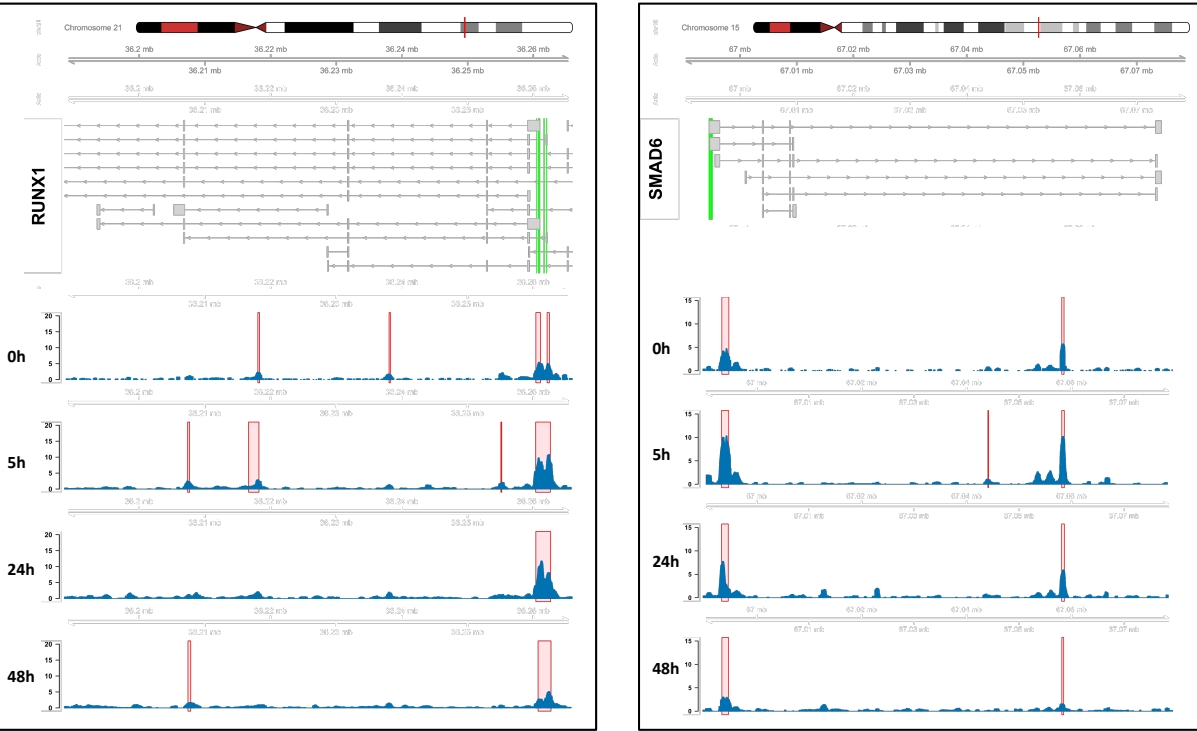
# Supplementary information



**Additional Fig. 1 Repartition of single cells in CALISTA clusters according time of sampling.** For each donor, 5 pie charts represent the proportion of cell corresponding to different time points (5h, 24h, 48h, 72h, 96h) defining the 5 clusters computed by CALISTA (**Fig. 2**). Note that cluster #1 and cluster #2 are almost uniformly composed of cells from 5h and 24h respectively. Starting from 48h, cells tend to fall into several clusters, indicating multiple routes. At 96h, corresponding to the lowest transcriptional level measured, we can observe that cells start to resemble each other again. **Left panel:** Donor 1, **right panel:** Donor 2.








**Additional Fig. 2 Heat-map of 11 TFs known to play a role in hematopoiesis regulatory network.** For each donor, 5 heat maps (one per time point) were drawn on the basis of gene expression intensity for the 11 TFs described in [13]. Gene expression was measured as the sum of UMIs for each gene considered, a log<sub>2</sub> transformation was applied for better visualization. Note that for the gene SMAD6, no transcripts were detected, all time and donor considered.

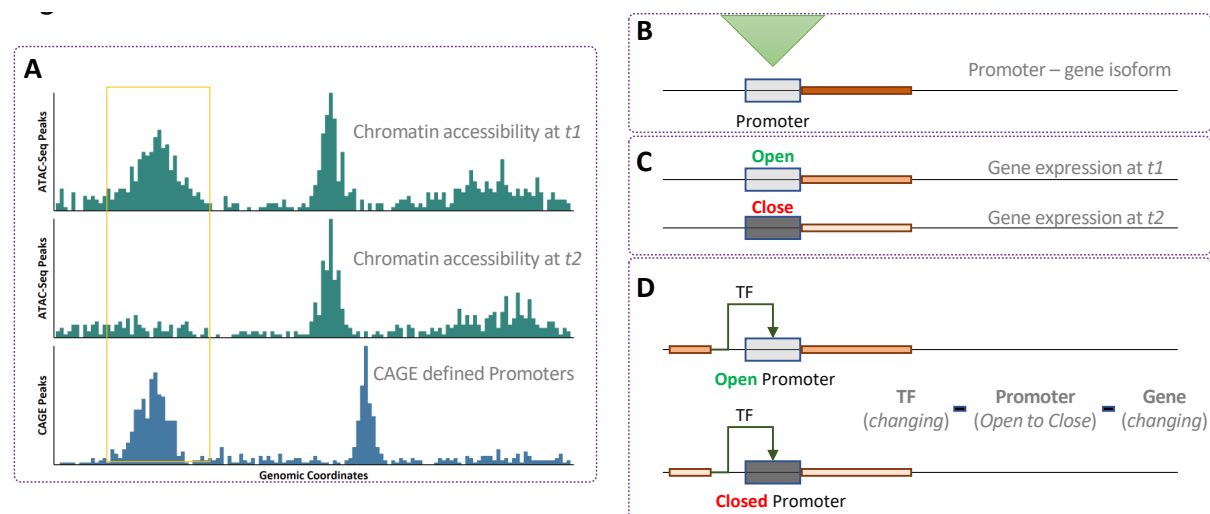


MACS2 detected peak  
Promoter region

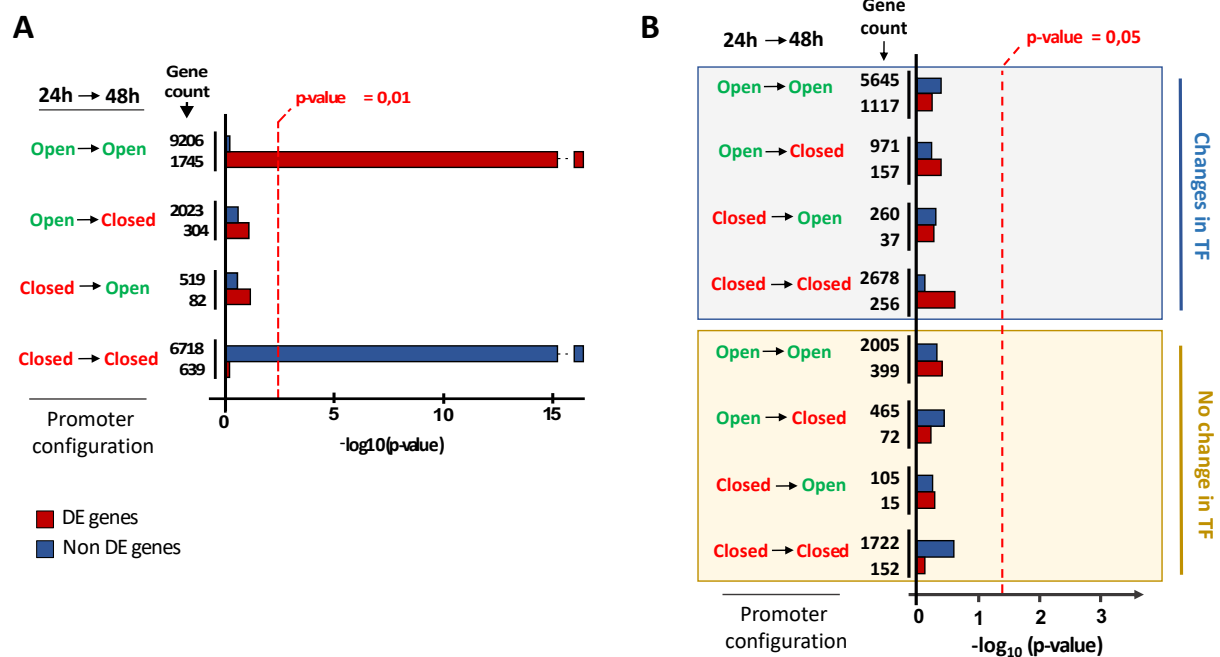
**Additional Fig. 3 Detailed accessibility profiles of 2 TFs known to play a role in hematopoiesis regulatory network.** For the sake of clarity, only two out of the 11 TFs described in [2] are represented here. The rest of the coverage plots can be found in **Additional File.4**. Green regions drawn on gene sequence represent CAGE-defined promoter considered in this study [13]. Red regions represent peaks computed with MACS2 algorithm. Blue “signal” depicts read coverage for one of the three samples available for each time point (the most representative one). Note that except for GATA1, promoter regions always overlap with detected peaks, no matter the time considered. Interestingly, whereas SMAD6 promoter is accessible, no transcript was detected for both Donor 1 and Donor 2, all times considered (**Additional Fig. 2**). This illustrates that accessibility is not sufficient to guarantee transcription.

		p-value				% of detection in peaks				% of detection in random background			
		1e-2126	1e-2418	1e-1180	1e-838	13.10%	15.75%	12.62%	12.60%	2.58%	3.62%	3.69%	3.86%
CTCF		1e-1907	1e-1705	1e-1440	1e-927	46.77%	52.81%	52.66%	52.05%	25.77%	32.94%	32.56%	33.46%
Fli1		1e-1615	1e-1427	1e-1427	1e-637	23.94%	27.60%	26.33%	23.90%	9.68%	13.40%	12.48%	12.37%
SPi1		1e-1654	1e-1375	1e-1169	1e-667	53.49%	60.78%	59.28%	56.41%	33.04%	42.51%	40.75%	40.33%
ERG		1e-697	1e-452	1e-525	1e-327	29.49%	34.28%	32.64%	30.48%	18.26%	24.87%	21.81%	20.83%
RUNX1													
		0h	5h	24h	48h	0h	5h	24h	48h	0h	5h	24h	48h

**Additional Fig. 4 Motif enrichment for selected known hematopoiesis related transcription factors.** At each time point, peak sequences were scanned by HOMER for significantly enriched “known motifs”. The motifs selected here illustrate a significant enrichment (zero or one occurrence per sequence coupled with the hypergeometric enrichment calculations) of motifs associated with hematopoiesis and chromatin remodeling. For an extensive list of tested motifs, see **Additional File.5**.



**Additional Fig. 5 Combined Analysis of scRNA-seq Single cell RNA expression with ATAC-seq chromatin accessibility peaks.** **A** An example showing the changes in the configuration of the chromatin reflected by differences in the peaks of the ATAC-seq and aligning the genomic coordinates with CAGE-defined promoter region (yellow box). **B** The CAGE-defined promoters are linked to gene isoforms. **C** Changes in the promoter accessibility from the ATAC-seq is integrated with the changes in the expression of the gene (RNA-Sequencing of single-cells). In this example, the promoter changes from an open configuration to a close configuration, and the gene isoform in its control changes expression. **D** TFs are then linked to the promoter using moderate to high confidence of motif occurrences in the specific genomic region. TFs themselves could change in their expression. In the example shown, the TF changes in expression, subsequently acting on a promoter that changes from an open to a close configuration and finally regulating a gene that changes in its expression between time  $t2$  and  $t1$ .



**Additional Fig. 6 Promoter configuration dynamics and transcription influence on gene expression regulation between 24h and 48h.** **A** Enrichment analysis for differentially expressed genes (DE) and non-DE genes depending on the promoter accessibility dynamics. **B** Enrichment analysis differentially expressed genes (DE) and non-DE genes depending on the promoter accessibility dynamics and the changes of the expression of TF-encoding genes that regulate them.

**Additional Table. 1. Impact of quality filters on the number of cells retained for final analysis.**

Time	Donor	Initial	Quality filters		
			UMI sum filter	Chr Mito filter	ERCC spike filter
5h	1	96	88	88	88
24h	1	96	79	78	78
48h	1	96	93	93	93
72h	1	96	45	45	45



96h	1	96	89	89	89
5h	2	96	88	88	88
24h	2	96	95	95	95
48h	2	96	92	91	88
72h	2	96	96	96	94
96h	2	96	72	70	68

Initially, we lysed 96 cells for each condition. First, we filtered out cells with less than 80 000 reads in total. Among remaining cells, cells with more than 10% of reads assigned to mitochondrial genome were also not included in the final dataset. Finally, ERCC spikes allowed removing cells with not satisfying linear amplification criterion (Pearson coefficient < 0,6).

**Additional Table. 2 Total number of peaks in ATAC-seq samples and number of common peaks retained for analysis.**

	0h			5h			24h			48h		
	Donor 1	Donor 2	Donor 3	Donor 1	Donor 2	Donor 3	Donor 1	Donor 2	Donor 3	Donor 1	Donor 2	Donor 3
Unique pairs of reads detected	8 391 299	14 310 319	15 701 894	13 372 383	17 359 415	14 998 407	16 460 386	19 605 272	11 588 529	8 569 939	14 513 711	12 665 206
Nb peaks detected after down sampling	66 155	72 075	66 708	60 734	78 486	68 173	54 288	70 713	65 096	50 404	55 249	49 219
Nb common peaks between donors	46 942			47 023			38 711			29 255		

We randomly downsampled each sample to the same level of 16M reads. Peak calling was applied on downsampled bam files. Finally, we considered only the intersection of the 3 peaks dataset available at each time point for further analyses.

**Additional File. 1 GO terms results for scRNA-seq WGCNA clusters**

74 (GO\_results\_WGCNA\_clusters.xlsx).

75

76 Additional File. 2 CALISTA 200 most variable genes matrix and cluster transition genes

77 (CALISTA\_200\_Top\_Variable\_Genes.xlsx).

78

79 Additional File. 3 scRNA-seq barcode sequences used for cell and plate demultiplexing

80 (scRNA-seq\_barcodes.tar.gz).

81

82 Additional File. 4 Detailed accessibility profile of 11 TFs known to play a role in

83 hematopoiesis regulatory network (11\_TF\_Gene\_Coverage.tar.gz).

84

85 Additional File. 5 Motif enrichment obtained with HOMER for 400 known motifs between

86 t = 0h and t = 48h (homer\_results.tar.gz).

87