Analysis of long non-coding RNA expression from single cell RNA sequencing datasets.

David Hidalgo Gil

Research project 15 credits.

Wallenberg Neuroscience centre, Lund University

Analysis of long non-coding RNA expression from single cell RNA sequencing datasets.

David Hidalgo Gil[1, *] & Yogita Sharma [1]

[1]Developmental and Regenerative Neurobiology, Wallenberg Neuroscience Center. Lund University

*Corresponding author

E-mail: da4206hi-s@student.lu.se

Index

Abstract

A majority of human genome transcribes into non-coding RNAs. These RNAs have low protein coding potential but have shown to perform wide range of cellular and molecular functions such as regulation of gene expression. They have also shown to regulate pathogenesis of brain tumour and development of neuropsychiatric diseases [1,2,3] . Cell type specific expression analysis needed to better understand the role of long non-coding RNA (lncRNA) in brain development and disease [4].

Introduction

Single cell RNA sequencing has emerged as a powerful technology to study heterogeneity in cell populations[5,6]. However, lncRNA are poorly analysed in these datasets due to two main reasons. Firstly, on technology side, all single cell RNA sequencing methods are dependent on poly A enrichment and all lncRNA do not have a poly A and data resolution is poor. Secondly, bioinformatics pipelines are not subjected to study lncRNA and limited only to study protein-coding genes.

This project aims to design a pipeline to study expression of lncRNA from single cell RNA sequencing dataset. This pipeline can then be further implemented to study broad range of functional and cellular roles of lncRNA in brain development and diseases.

Materials and methods

Single cell RNA sequencing

Human ES cells were differentiated into dopamine progenitor cells across three time points (day 16, 30 and 60) [7]. Single-cell suspensions were loaded onto 10x Genomics Single Cell 3′ Chips along with the reverse transcription (RT) mastermix per the manufacturer's protocol and samples were pooled together for sequencing using in house nextseq machine.

Data pre-processing and Alignment

The raw base calls were demultiplexed and converted to sample specific .fastq files using cellranger mkfastq[1] that uses bcl2fastq program provided by Illumina. For each sample, .fastq files were processed independently using cellranger count version 3.0 pipeline. This pipeline uses splice-aware program STAR[8] to map cDNA reads to the transcriptome (GRCh38). Mapped reads were characterized into exonic, intronic and intergenic if at least 50% of the read intersects with an exon, intronic if it is non-exonic and it intersects with an intron and intergenic otherwise. Only exonic reads that uniquely mapped to transcriptome (and the same strand) were used for the downstream analysis. Aligned reads were filtered for valid barcodes and UMI and observed cell barcodes were retained if they were 1-Hamming-distance away from entering into the whitelist of barcodes.

Downstream analysis

Standard single cell RNA analysis

This analysis was performed using Seurat *(V. 3.1.5)* [9,10] , tidyverse *(V. 1.3.0)* and patchwork *(V. 1.0.0)* in RStudio running R *(V. 4.0).* Code can be found at the project repository.

The standard single cell analysis consisted of quality control (QC), linear analysis, clustering and cell type identification. The main objective for the QC analysis was to remove doublets and poor quality cells. Quality thresholds needs to be determined after inspection of the quality

control plots (see annex Figures 1 A-C). The data was then normalized by a global-scale normalization method that normalizes the expression of each feature across all cells by the total expression, multiplies the result by a scale value of 10000 and log-normalizes the result.

Then we proceeded to identify the 1000 most variable features using Variance-stabilizing transformation (VST) as the selection method (see annex Figures 2 A-C). After that the data was scaled so the mean expression of each feature across cell is 0 and the variance is 1, this is done so that genes with high or low expression level do not have a big impact in the overall analysis [10].

The next step was to perform a linear dimensional reduction, for that we performed a principal component analysis (PCA) to identify correlated feature sets from the most variable features previously determined. To determine how may components to include in further downstream analysis we scored the components using the JackStraw procedure [11]. And selected the features with a Pval < 0.000001. Once the number of selected features was determined for each dataset we proceeded to cluster the cells using graph based approaches [12–14] and a resolution value of 0.1.

To reduce the dimensionality of the data to a level we can easily see and interpret we performed UMAP and tSNE reductions of the components previously selected. This allowed us to have a clearer Idea of the data and correlate the clusters to their biological counterparts. To annotate the cell types, present in each cluster we plotted specific cell markers (features) on our data. The list of features and cell types used can be found in “data” at the project repository.

LncRNA expression analysis

After annotating the clusters, the barcodes for the cells in the “Neurons” and “In development” clusters were extracted at each time-point. To avoid poor resolution problems in downstream analysis. Cellbarcodes for each cluster from corresponding samples were used to subset original .bam file for extracting cell type specific aligned reads. This step was performed using

subset-bam function provided by 10X. LncRNA expression for each cell type was calculated using FeatureCounts.

This analysis was performed using DeSeq2 (*V. 1.28.1*), pheatmap (*V 1.0.12)*, and RColorBrewer (*V. 1.1-2)* packages in RStudio running R *(V. 4.0).* Code can be found at the project repository.

The input data for this analysis are the feature counts from all the different clusters. The data was then analyze using DeSeq's 2 standard workflow with count matrix data as input and a sample information table that related each data column with their respective time point and cluster. Before running the analysis the data from day 16 was set as a reference [15]. Features with a Pval < 0.05 were selected as significantly differentially expressed. The similarity between the datasets was analyzed by calculating the Euclidean distance between the different datasets.

Results

Quality control

After mapping and counting the original sequence reads for each time point the data was loaded into their respective Seurat objects and quality plots were generated (see annex Figure 1 A-C). We can see an increase in the number of hypothetical cells with low feature counts at days 30 and 60. This is probably due to genomic fragments suspended in the media by dead cells.

Cells that presented between 1100 and 4000 different features expressed and less than 3% mitochondrial gene contamination were selected for later analysis see Table 1.

Table 1. *Cell counts before and after*

	Before QC	After QC	% Removed
Day 16	8613	8006	7.047
Day 30	4508	3644	19.16
Day 60	6749	4486	33.79

Linear analysis

The linear analysis determined the different components, 24, 26, 28 components were used for future analysis for each respective dataset.

Clustering

UMAP Clustering results (Figure 1) show 4 distinct clusters, these clusters were annotated as "In development", "Neurons", "Oligodendrocytes" and "FBLC" (see annex Table 1 for more details) by plotting specific cell types and matching expression patterns from these markers to our data (see "Cluster identification" in the Image archive). As we can see in Figure 1 the different clusters gain more distance between them as time passes.

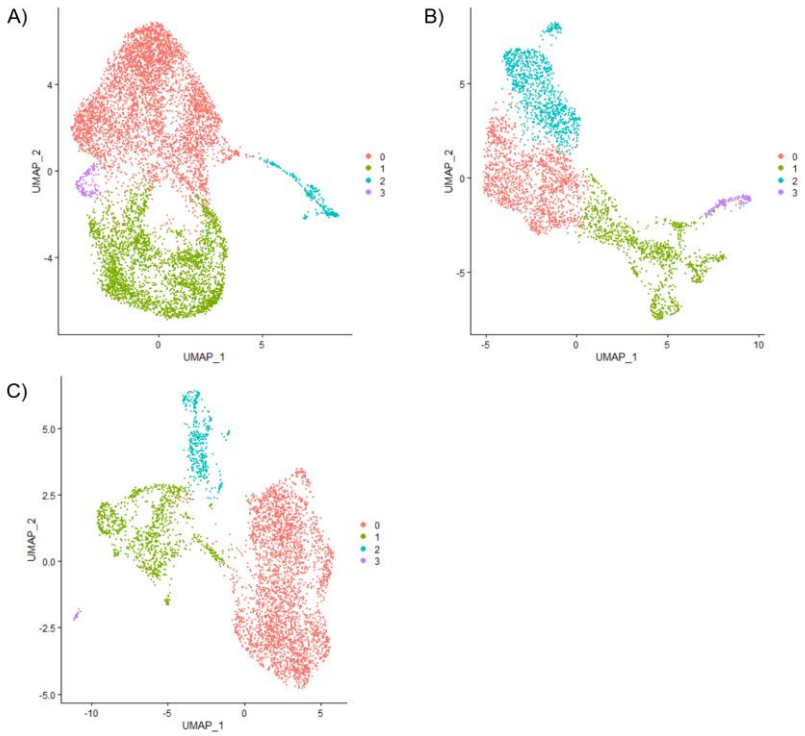


Figure 1*. UMAP clustering results at time-points 16,30 and 60 (A, B and C respectively) using 0.1 resolution and the respective number of dimensions..*

As we can see in Figure 2 the “Neurons” population increases as time passes, at the same time the “In development” population decreases as it is expected but the biggest changes occur at different time-points. The “in development” population shows a decrease by day 30 and the “neurons” populations shows an increase at day 60. This is probably due to the quality of the data as discussed previously.

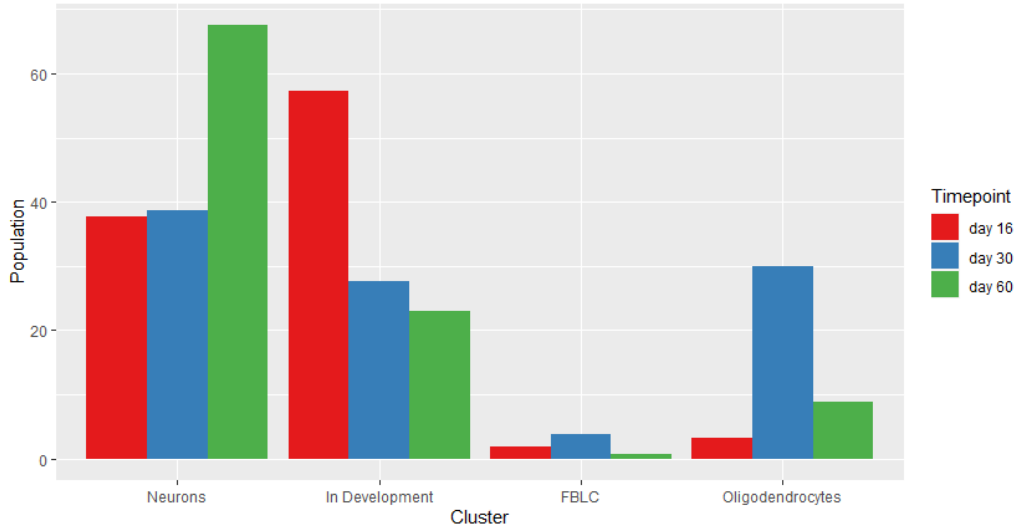


***Figure 2**. Cluster population (%) evolution across all time points. As we can see the “neurons” population increases as the “In development population decreases.*

Differential Expression analysis

After counting the expression of lncRNAs we obtained 16718 different features out of wich 11682 had a total read count higher than 0. The analysis was run with a significance level of 95% (adjusted p-value < 0.05) 666 (5.7%) passed the significance threshold, see Figure 3.

As we can see in the distance matrix shown in Figure 4 the groups are different from each other, at each timepoint the distances between clusters are lower, especially at day 16. This is porbalbly due to the low differentiation level that the cells have at this timepoint. The development stage seems to be the main driver to these differences.

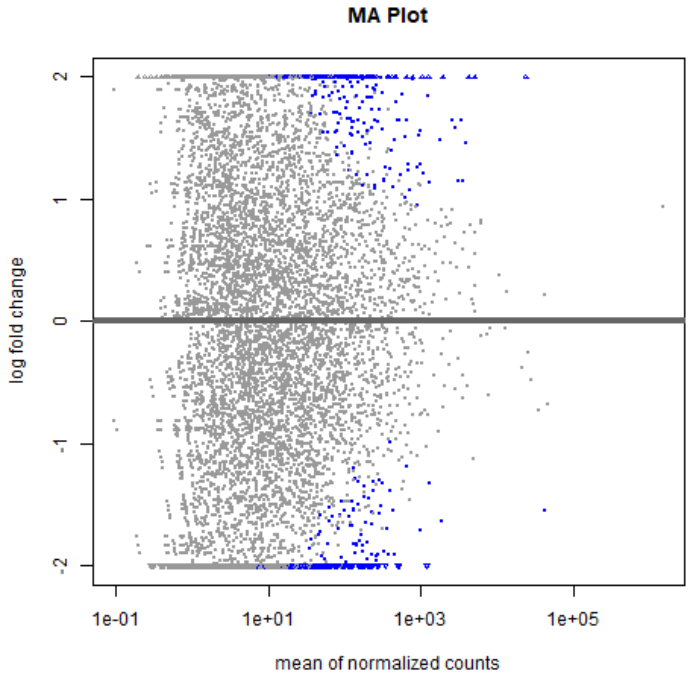


Figure 4. *MA plot of the mean of normalized counts (excluding 0) in the X axis and log fold change in the y axis. Significant features are shown in blue.*

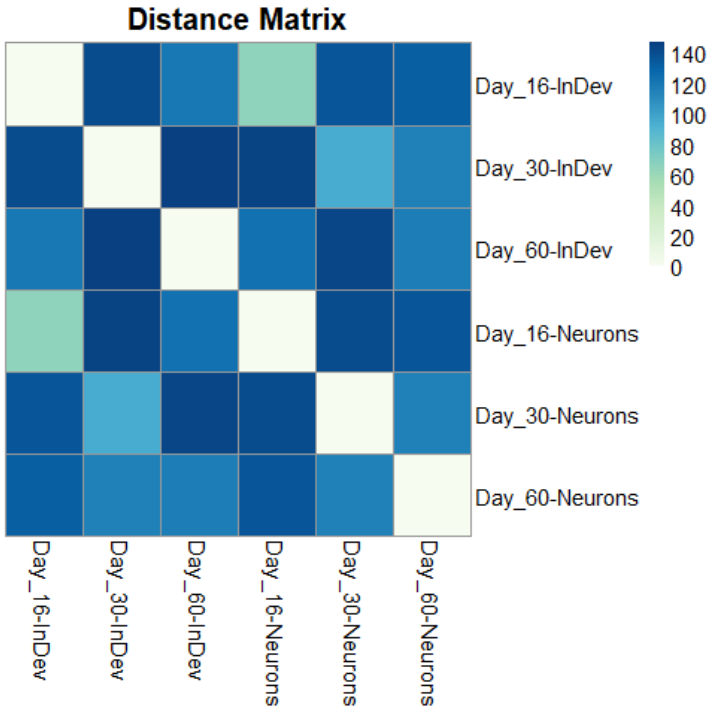


Figure 3. *Distance matrix between groups.*

Conclusions

Here we presented a pipeline that can be implanted further on different datasets to determine long non-coding RNA expression from single cell RNA sequencing datasets. However, we would like to draw the attention towards the fact that, in order to come over poor resolution here we are still looking at long non-coding RNA expression per cluster rather per cell. From the different results obtained we can conclude that lncRNAs are differentially expressed across cell types and time. Also, interestingly, it is not time point but cell type that contributes to major differences in differential expression of long coding RNAs. Further investigation using more replicates is required to establish the role of lncRNAs in defining cell types.

Bibliography

1. Chen, Y. & Zhou, J. LncRNAs: macromolecules with big roles in neurobiology and neurological diseases. *Metab. Brain Dis.* **32**, 281–291 (2017).

2. Wu, P. *et al.* Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res. Bull.* **97**, 69–80 (2013).

3. Ng, S.-Y., Lin, L., Soh, B. S. & Stanton, L. W. Long noncoding RNAs in development and disease of the central nervous system. *Trends Genet.* **29**, 461–468 (2013).

4. Tiklová, K. *et al.* Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. *Nat. Commun.* **10**, 1–12 (2019).

5. Mu, Q., Chen, Y. & Wang, J. Deciphering Brain Complexity Using Single-cell Sequencing. *Genomics, Proteomics Bioinforma.* **17**, 344–366 (2019).

6. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

7. Nolbrant, S., Heuer, A., Parmar, M. & Kirkeby, A. Generation of high-purity human ventral midbrain dopaminergic progenitors for in vitro maturation and intracerebral transplantation. *Nat. Protoc.* **12**, 1962–1979 (2017).

8. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

9. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

10. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data Resource Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

11. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual

Cells Using Nanoliter Droplets Resource Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

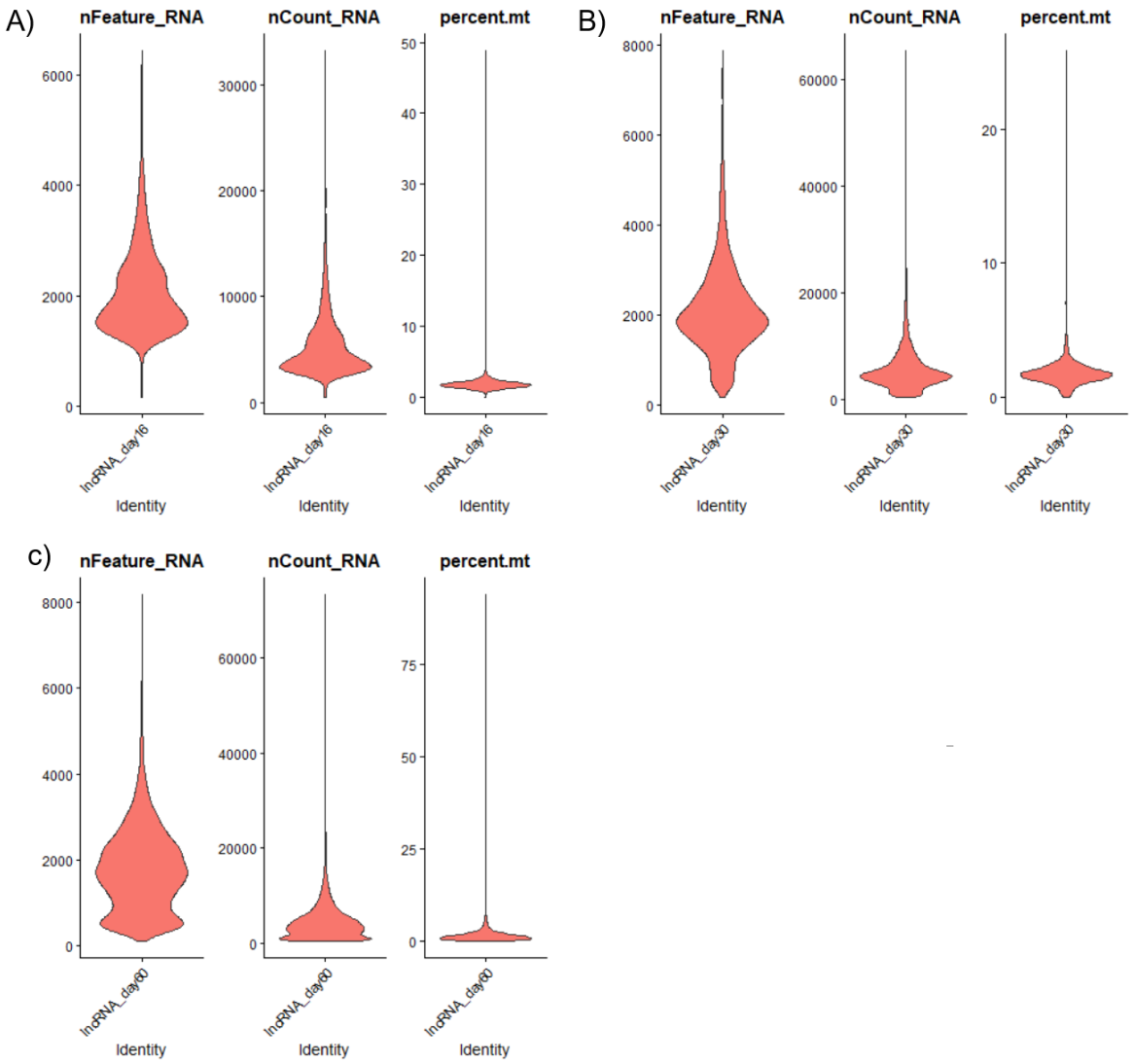
12. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, (2008).

13. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).

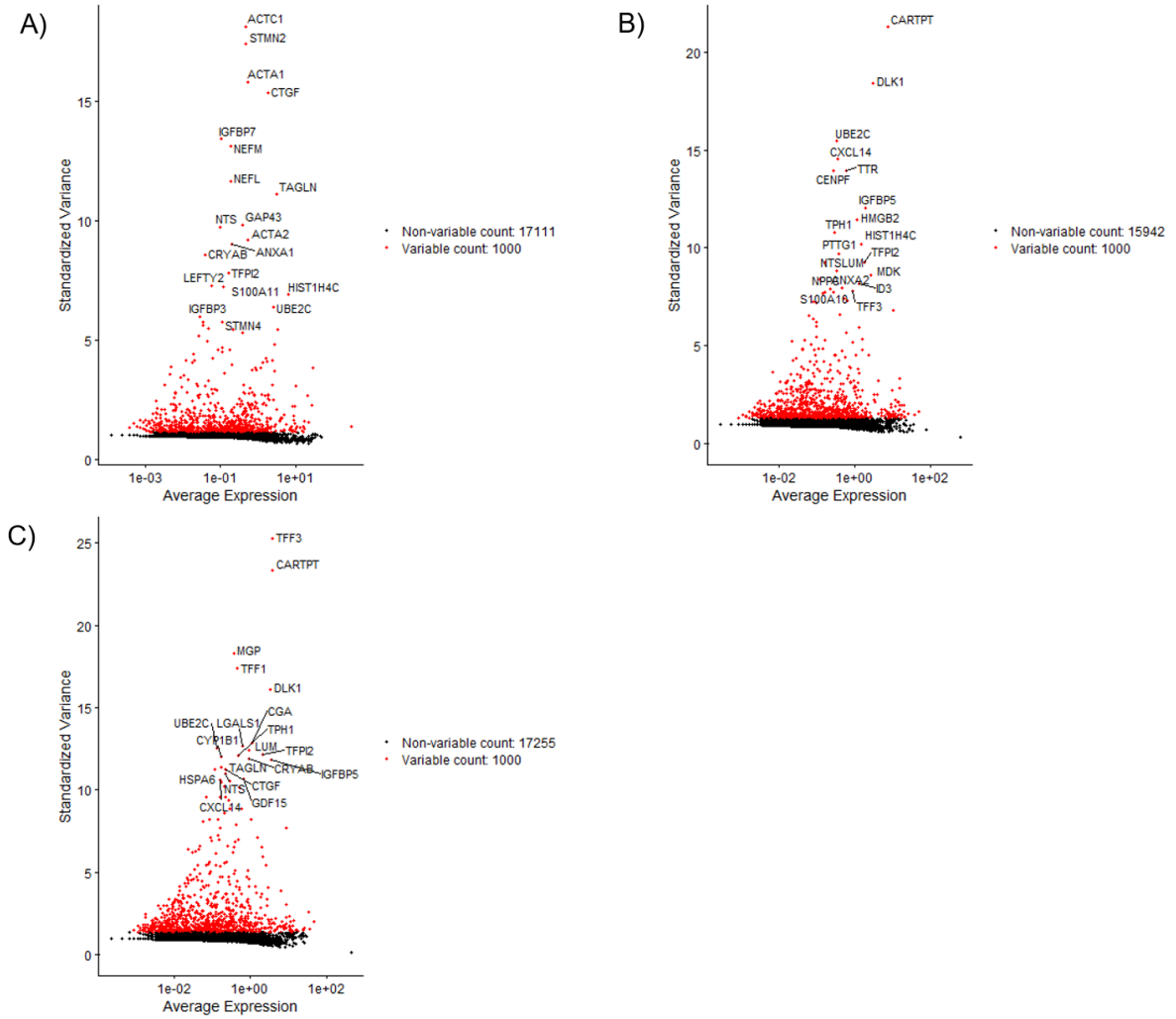
14. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Annex



Annex figure 1. *Quality plot at time point 16, 30 and 60 (A ,B and C respectively).*
nFeature_RNA shows the distribution of the number genes expressed per barcode (hypothetical cell).
nCount_RNA shows the distribution of total RNA counts per cell.
Percent.mt shows the percentage of detected genes of mitochondrial origin.



***Annex figure 2**. Selected top variable genes at timepoint 16, 30 and 60 (A,B and C respectively). Top 20 genes labeled*

Annex Table 1. Clsuter annotation results at each timepoint. “In development”, “Neurons”, “Oligodendrocytes” and “FBLC” were the common terms selected to describe each cluster.

	Day16	Day 30	Day60
0	InDevelopment	Neurons	Neurons
1	Neurons	Oligodendrocytes/ Progenitors	InDvelopment
2	Oligodendrocies/ Pan Neurons/ Extraneuronal	InDevelopment	Oligodendrocytes
3	FBLC	FBLC	FBLC