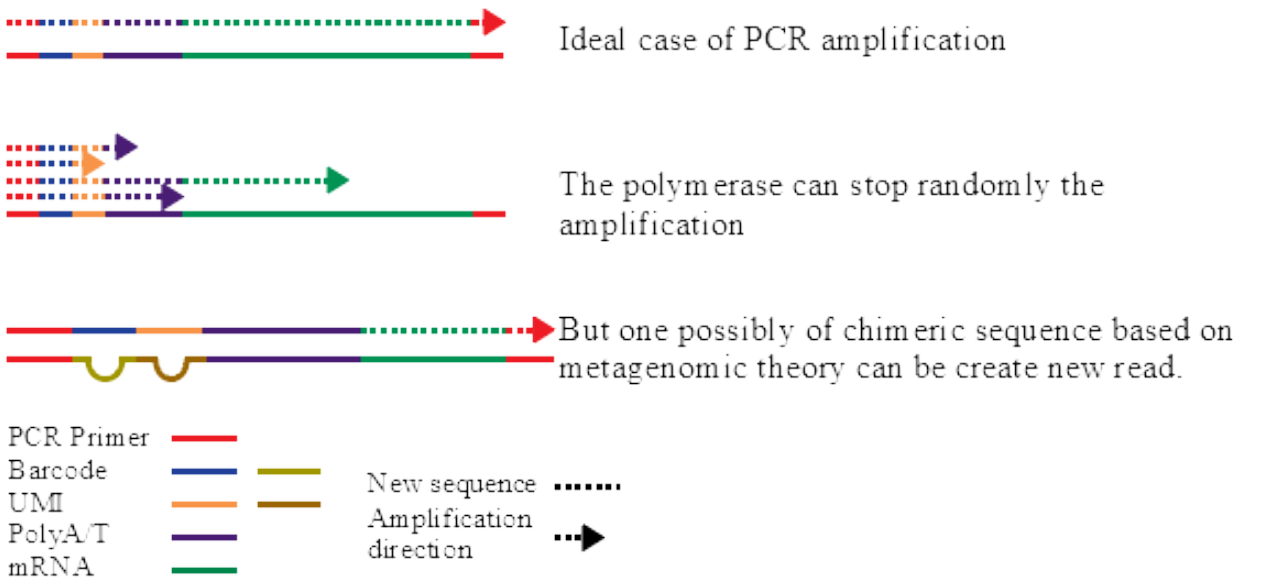
Personal project in single cell RNA-seq

RNA-seq is giving many opportunities to scientists to identify different cells types and alternatives transcripts isoforms[1]. Many tools were were created to identify those cells types. You can extract specific genes[2] by exploiting the differential expression or exploit the data to explain the datasets expression by network[3]. RNA-seq take a part of a tissue composed of multiple cells type. For a better resolution, single cell approach was developed[4]. Genomics, Epigenomics[5], Transcriptomics, Proteomics and Metabolomics are now studied using single-cell approach[6]. Those new study techniques needed new softwares to extract meaningful information from data. For example, new algorithms for the genome assembly in metagenomic can be applied in single cell sequencing[7]. At the same time, G&T-seq, a technique who combines genomic and transcriptomic data extraction from the same cell, was developed[8]. A recent study extracted three data from the same cell[9]. With the data correlation, we can better understand interactions in the cell and improve the data by cross validation.

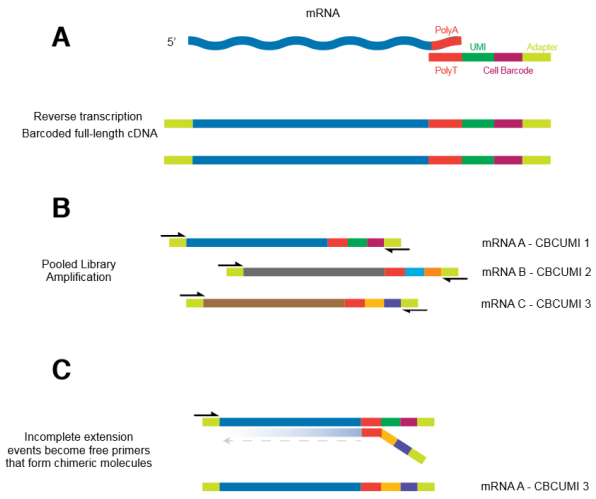
For this project, I concentrate my efforts on scRNA-seq datasets. In 2015, the first experience of the Drop-seq technique has been published[10]. This technique is one of the cheapest, adaptable in an environment Illumina, and allowing the mRNA sequencing of 10,000 cells[6]. In the success of this experience, a new way in transcriptomic analysis emerged. Other techniques using different approach to separate cells and obtain their mRNA also exist. CEL-Seq2[11], STRT[12], sci-RNA[13], Seq-Well[14], all this techniques use Barcode to identify cells and PCR for the DNA amplification. The results are identical, you obtain a cells' expression matrix. You have different way to analyse it[15]: lineage tracing, the algorithm find the genes for combine cell and obtain the dynamic in the development of cell from the dataset[16–18], network construction, this technique use the genes network to reconstruct the different group of cell[19,20], or simply using a tSNE for study the complexity between cells[21,22]. With the explosion in the difference software, scRNA-tools, a database was created[23]. You have many ways to analyse your data and this database[23] could give you the right tool. But before analysing your dataset, you will have to purify your dataset. You can have lot of errors, and generate fake information. How do you know if your cell is not a subset of one cell? Will you combine cells with low genes into the biggest? But, how detect barcode with multiple cell? With the total level of expression. But is this right? If we created errors before the sequencing step. The catch step, and the PCR are problematic. We will take a little proportion of mRNA during the catch step[6], randomly between cells. During PCR, the gene with lot of expression will be more dominant and hide the weakly express gene. In the same time, PCR create difference between cell, in other word cell with certain barcode are more amplified than other[24]. We can use T7 polymerase, but if we increase the error rate[25–27] on all sequence, we increase the barcode and UMI errors. In addition that we know in metagenomics, the PCR create chimerics sequences[28,29] with closes sequences. "Chimeras are artificial recombinants between two or more parental sequences, and they are normally formed when prematurely terminated fragments reanneal to other template DNA during PCR amplification"[30].In metagenocmics, the error can represent be 5%[31] of your dataset. To identify them, the metagenomics specialist developed lot of tools to find a way to identify them. This operation takes a lot time and is computation intensive. The chimeric sequence were deteted[32] in scRNA-seq. In metagenomic, one of those approach use DNA 16S template to detect chimeric sequences. In scRNA-seq we don't have template.

Why a chimeric sequence are a problem in our data?



Personal Source

In this case we can simply detect chimeric product. The sequence have the same UMI from another sequence. But, the problem will be the quantity. A chimeric product can appear multiple times, and we can remove the real sequence if we delete the product. In the same time, A. Dixit speak about one possibility in chimeric sequence creation:



Source : Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments

In concept, only polyA sequence are the problem for the creation in chimeric sequence. And if we have shift between sequence?

You can't just based your correction with UMI or cell barcode sequences.

The Drop-seq developer give us an example in their cookbook[33]:

- CellUMITTT
- Error: AAAAACGTGGG-CAGCGTAATTT
- Fixed: AAAAACGTGGGNCAGCGTAATTT

N

My questions are: How separate good and bad sequences? Can we consider the UMI like a real quantity?

It is why I developed a new approach.

Results

I test the software on data from Quadrato & al.[22]. I used the expression matrix. It have always the barcode in the cell name, each sequencing run have different name, the number of cells in each dataset are different, and each run have a specific quality. I use Cytoscape for the visualisation.

All results are available on : https://github.com/studyfranco/STUDER

But if you have a dataset to suggest, I can test it and add in the result. Send me a link where I can download the expression matrix file.

Finnaly, the project take an another way. I run it with my method. I use two informations:

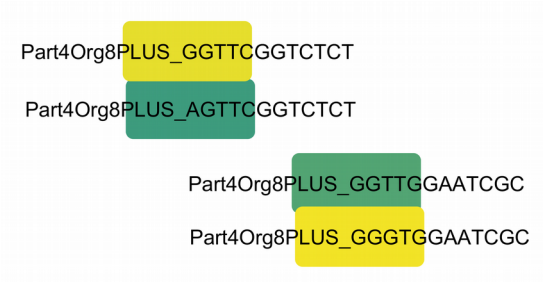
- A tanimoto index based of the presence and absence of one gene in the cell.
- If a cell have a part of the genes of an another cell. I call that internal.

We can now cut the project in 3 main goal.

The first goal are: I want find all same cells informations with different barcode. If you prefers: PCR + Chimeric sequence can create error in the barcode, and if we want find the identical cell in the dataset the work are complicate.

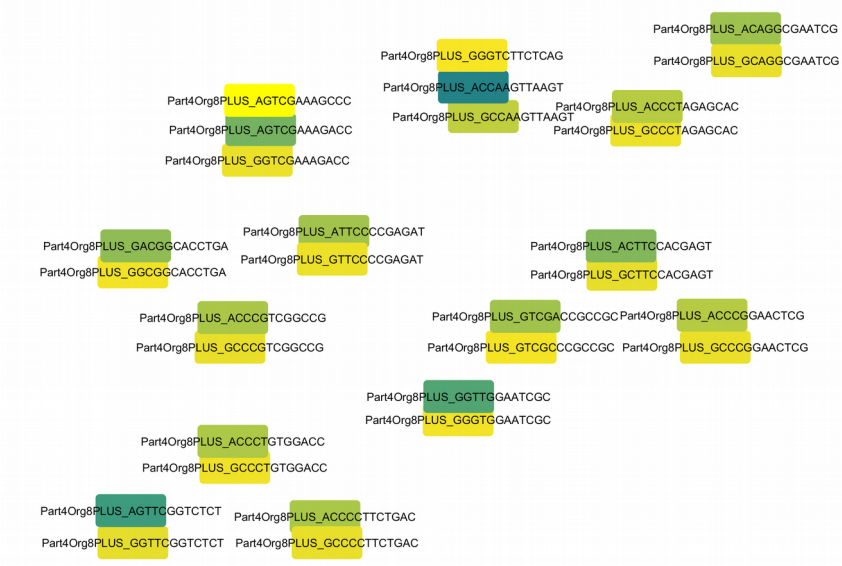
In the ideal case:

If we compare cell in internal with only the basic gene information we obtain for the Organoid4H:



We detect each two cells internal in each two other. We can see 1 base different in the barcode. The yellow cell have less genes express than the green. My first idea: It is possibly to merge the informations or just remove the cell.

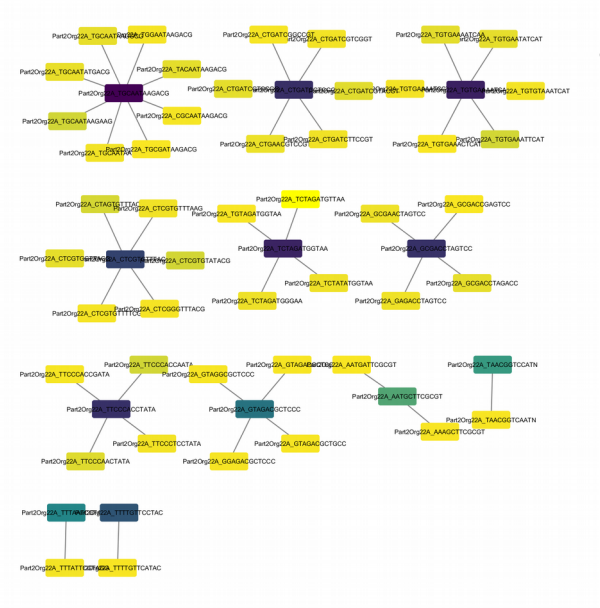
With my software, for the same organoïd:



We identify more internal cells. 16 cells are just a subset of one other cell. 16 cells are 0,5% of our total dataset. This proportion are not so important in our dataset. But possibly have an impact in the algorithm for the analyse.

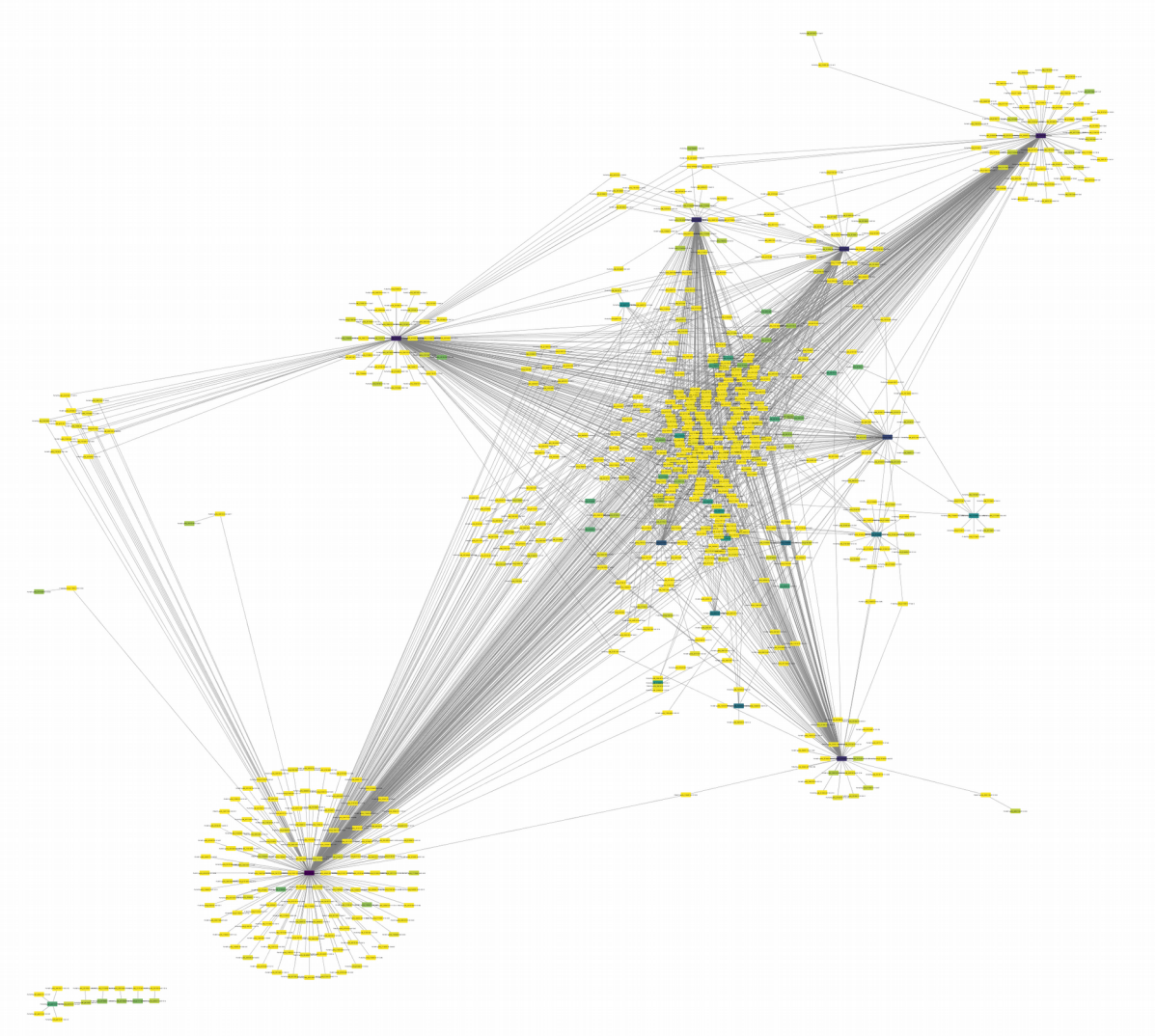
Here, the dataset are good.

Orga1A, with only the basic gene information:



We detect 47 cells, can be a subset in one cell. 1,3% of all cell are not good.

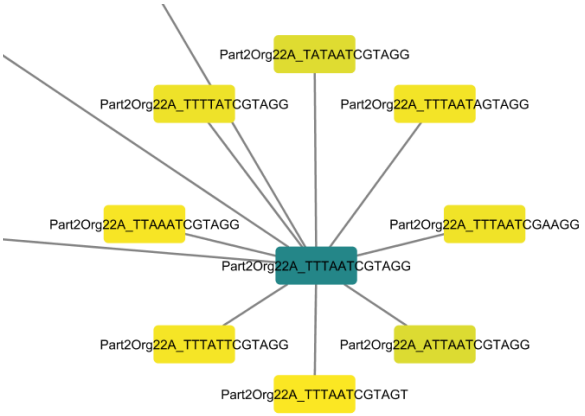
With my software:

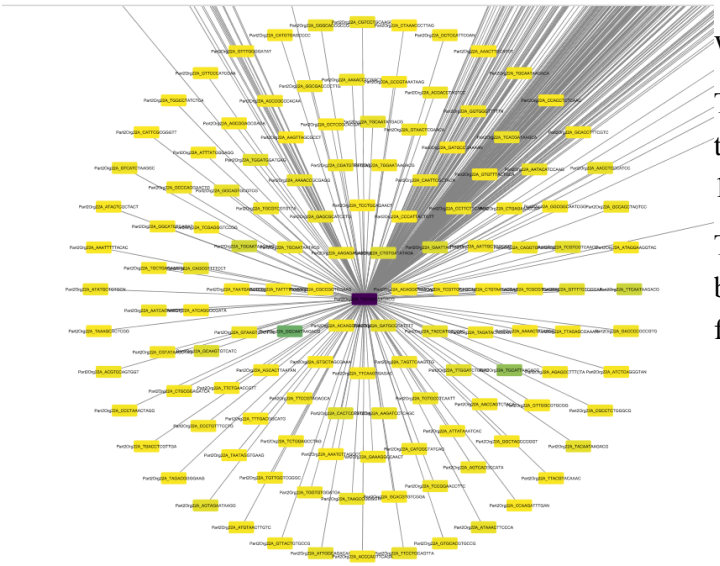


We obtain a new pattern.

We have one part with the precedents results.

This cell are always good.



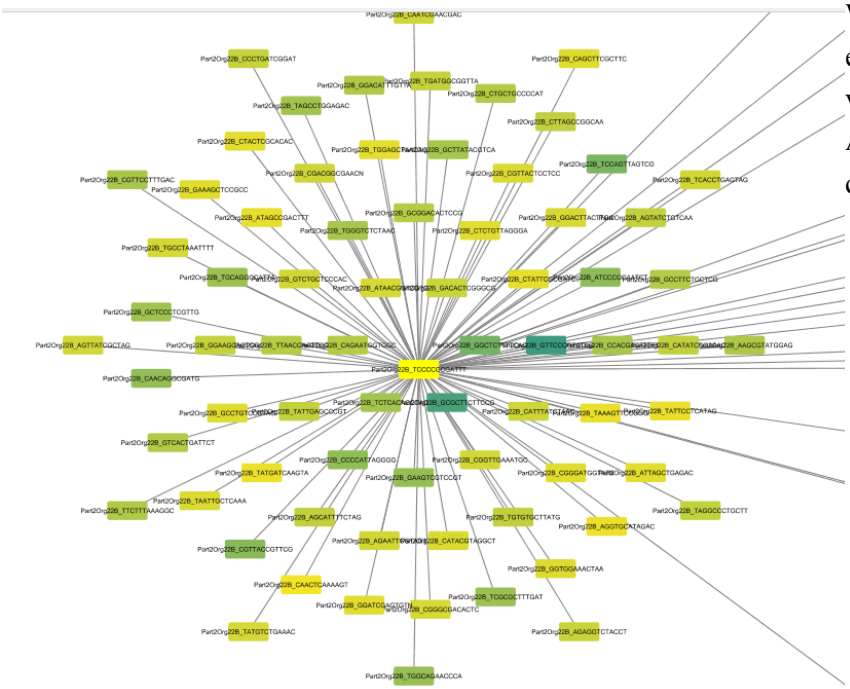


We observe this cell who have lot of internal cells.

The purple cell have lot of gene express and lot of total expression. It can be a drop-let with more than 1 cells.

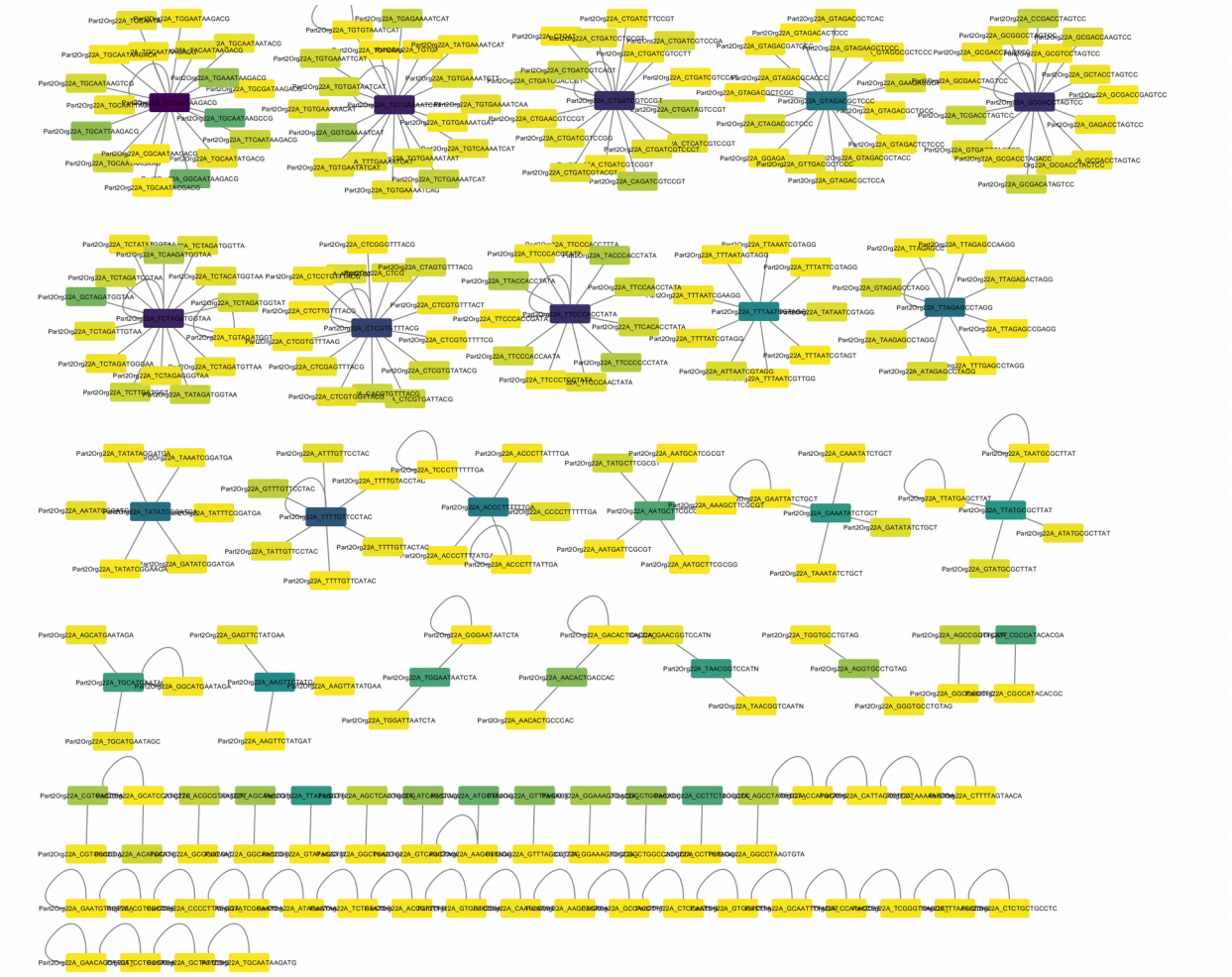
This is the second goal of the software: Detect barcode who contain the expression information from more than 1 cell.

The third goal of my software are in the same time, detect cell with low expression. For this example I will take Orga1B.



We can see one cell, with 409 genes express in the raw dataset. It validate only 93 genes for this cell. And after treatment we detect this cell internal in 92 other cell.

The software give this solution for clean the dataset of Orga1A:



When a cell have an edge who come back to itself, the software advise to remove them. When you have an edge between two cell, the software advise to merge the information between two cell.

In this first example, cell with low coverage are internal in 2 cells. And the possibility cell barcode who contain the expression information from more than 1 cell, are detect when you have 5 cells with different barcode inside them.

This software need 10-30 min for create this Cytoscape representation and clean the dataset. You can find all result on github.

Conclusion

The software can detect more barcode error, multiple cell in the same barcode, and low coverage cell expression. I used only the matrix file in the original software. The threshold to select bad cells, and correct the dataset can be improve. It is a beginning, and I hope improve the results.

Finally, I didn’t find a result who can explain how this method can resolve it. In the next day, I will obtain GOTerm, pathway who detected and test my clean dataset on publish software.

A clean dataset, are the insurance of a good analyse.

Bibliography

1..Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).
2..Chan, C.-K. K. *et al.* A differential k-mer analysis pipeline for comparing RNA-Seq transcriptome and meta-transcriptome datasets without a reference. *Funct. Integr. Genomics* (2018). doi:10.1007/s10142-018-0647-3
3.....Mendoza-Parra, M.-A. *et al.* Reconstructed cell fate-regulatory programs in stem cells reveal hierarchies and key factors of neurogenesis. *Genome Res.* **26**, 1505–1519 (2016).
4.....Wang, W., Gao, D. & Wang, X. Can single-cell RNA sequencing crack the mystery of cells? *Cell Biol. Toxicol.* **34**, 1–6 (2018).
5.…Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nat. Rev. Genet.* **16**, 716–726 (2015).
6.....Mincarelli, L., Lister, A., Lipscombe, J. & Macaulay, I. C. Defining Cell Identity with Single-Cell Omics. *PROTEOMICS* **18**, 1700312 (2018).
7.....Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
8.....Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
9.Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. 19
10.…Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
11.....CEL-Seq2 – sensitive highly-multiplexed single-cell RNA-Seq | RNA-Seq Blog. Available at: https://www.rna-seqblog.com/cel-seq2-sensitive-highly-multiplexed-single-cell-rna-seq/. (Accessed: 19th February 2019)
12.....STRT. Available at: https://emea.illumina.com/science/sequencing-method-explorer/kits-and-arrays/strt.html?langsel=/fr/. (Accessed: 19th February 2019)
13..Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
14..Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
15.....Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, (2016).
16..Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166–179 (2018).
17.…Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702 (2016).
18.....Ding, J. *et al.* Reconstructing differentiation networks and their regulation from time series single-cell expression data. *Genome Res.* **28**, 383–395 (2018).
19.…Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
20.…Woodhouse, S., Piterman, N., Wintersteiger, C. M., Göttgens, B. & Fisher, J. SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol.* **12**, (2018).
21.....Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell Stem Cell* **20**, 858-873.e4 (2017).
22..Quadrato, G. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48–53 (2017).
23.……Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Comput. Biol.* **14**, e1006245 (2018).
24.....Berry, D., Ben Mahfoudh, K., Wagner, M. & Loy, A. Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Appl. Environ. Microbiol.* **77**, 7846–7849 (2011).
25.……McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol. Biol. Int.* **2014**, (2014).
26...Nakano, T. *et al.* T7 RNA Polymerases Backed up by Covalently Trapped Proteins Catalyze Highly Error Prone Transcription. *J. Biol. Chem.* **287**, 6562–6572 (2012).
27.……Ling, L. L., Keohavong, P., Dias, C. & Thilly, W. G. Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and vent DNA polymerases. *Genome Res.* **1**, 63–69 (1991).
28..Gonzalez, J. M., Zimmermann, J. & Saiz-Jimenez, C. Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics* **21**, 333–337 (2005).

29.....Bradley, R. D. & Hillis, D. M. Recombinant DNA sequences generated by PCR amplification. *Mol. Biol. Evol.* **14**, 592–593 (1997).
30..Kim, M. *et al.* Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. *Genomics Inform.* **11**, 102–113 (2013).
31.....Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494–504 (2011).
32.....Dixit, A. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments. *bioRxiv* (2016). doi:10.1101/093237
33. *Java tools for analyzing Drop-seq data. Contribute to broadinstitute/Drop-seq development by creating an account on GitHub*. (Broad Institute, 2019).