Overview:

1. From scRNA seq data generated with 10x - cellranger pipeline, create the Seurat object (no idea how to do it, Kerim/Fumio?)
2. Create all the metadata files needed for scVelo
3. Use velocyto to create the loom file (in particular, the file to split spliced and unspliced data)
4. Use the loom file as input for the scVelo analysis pipeline and plotting

This pipeline is largely based on: <https://github.com/basilkhuder/Seurat-to-RNA-Velocity>

# Create Seurat object

No idea how to do this. Kerim could comment?

# Create metadata files

In this part, we use Seurat R package ([1](https://www.sciencedirect.com/science/article/pii/S0092867419305598?via%3Dihub), [2](https://cran.r-project.org/web/packages/Seurat/index.html)) to generate the metadata files.

NOTE: Seurat is exclusively an R package.

write.csv(colnames(GetAssayData(seurat\_object)), file=”cellIS\_obs.csv”)

write.csv(Embeddings(seurat\_object, reduction=”umap”), file=”cell\_embeddings.csv”)

write.csv(seurat\_object@meta.data$seurat\_clusters, file=”clusters.csv”)

# Use velocyto to create loom file

To compute RNA velocity fields, we need to annotate the scRNAseq data to count the number of spliced/unspliced reads for every gene in every cell. This can be done with the original RNA velocity software package, velocyto ([3](https://www.nature.com/articles/s41586-018-0414-6), [4](http://velocyto.org/)).

## Files requirements

To run velocyto we need 4 files, 2 of them from the scRNA seq experiment:

1. the raw bam data (huge file ~30GB)
2. The barcodes file (~25 KB)

According to the [velocyto user guide](http://velocyto.org/velocyto.py/tutorial/cli.html), the other two files are used to split the RNAseq data into spliced/unspliced. These are the gene annotation file and the repeats annotation file. These depend on the model organism and can be downloaded in [GENCODE](https://www.gencodegenes.org/) or [Ensembl](http://www.ensembl.org/info/data/ftp/index.html) and from [USCS](https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=611454127_NtvlaW6xBSIRYJEBI0iRDEWisITa&clade=mammal&org=Mouse&db=mm10&hgta_group=allTracks&hgta_track=rmsk&hgta_table=0&hgta_regionType=genome&position=chr12%3A56694976-56714605&hgta_outputType=primaryTable&hgta_outputType=gff&hgta_outFileName=mm10_rmsk.gtf), respectively.

## Software requirements

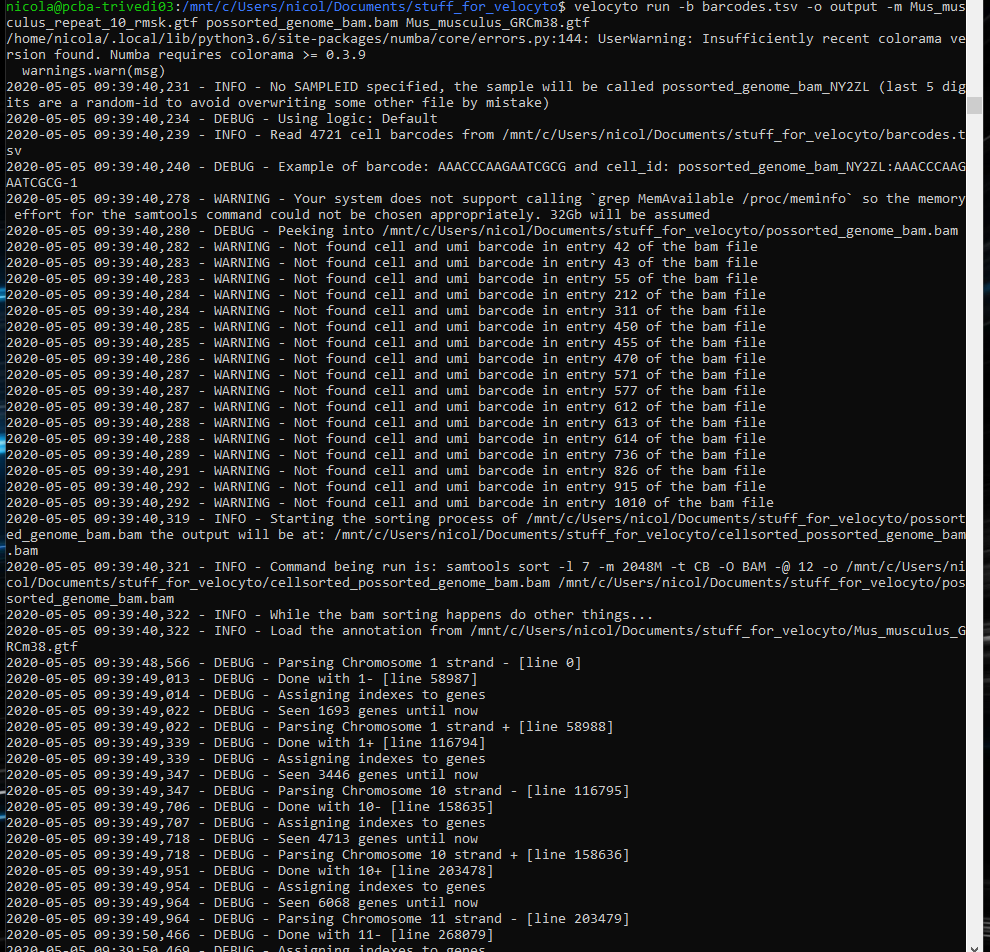
The velocyto software is available for both Python ([5](http://velocyto.org/velocyto.py/install/index.html)) and R ([4](http://velocyto.org/)). Unfortunately, I have not been able to install it on either platform on my Windows laptop. On Windows I got an error due to the “pysam” package. Apparently this is a known issue and it doesn’t seem pysam developers are actively taking care of it: [pysam is not Windows compatible](https://github.com/pysam-developers/pysam/issues/575). The solution provided in the link (WSL) also didn’t work. On the other hand, the R documentation ([4](http://velocyto.org/)) seems to assume the OS is unix based.

Based on this knowledge, I downloaded and installed Ubuntu 20.4 on a separate machine, and tried with the Python (obviously because Python >> R) velocyto there. I installed [Anaconda](https://docs.anaconda.com/anaconda/install/linux/), and followed the Python installation instructions. This time pysam installed correctly and, after installing the anaconda gcc compiler (conda install gcc\_linux-64), velocyto installed properly. Upon running the command, an error appeared that asked samtools installation. This can be done from [bioconda channel](https://anaconda.org/bioconda/samtools). However, after

velocyto run -b filtered\_barcodes.tsv -o output\_path -m repeat\_msk\_srt.gtf bam\_file.bam annotation.gtf

overnight run, another memory-related (?) error appeared.

Next, I tried installing velocyto Python on WSL on my machine (which has 64GB memory). This needs to be done without anaconda because there is no interface but only command line in WSL. I successfully installed velocyto (pip3 install velocyto) after installing all dependencies as before and “sudo apt-get install samtools” (because in WSL there is no conda so I can’t use the link above...). Now it seems to work with only a few warnings:



## Running velocyto and output

It took ~4 hours to download the 30GB bam file from server. Running

Velocyto run -b barcodes.tsv -o output\_path -m Mus\_musculus\_repeat.gtf possorted\_genome\_bam.bam Mus\_musculus\_GRCm38(\_chr???).gtf

Takes about 4 hours and yields the loom file to be used as input for the scVelo step.

1. Use scVelo to compute RNA velocity and generate beautiful plots

Finally, back to sane programming with Python on my Windows :)

[This is easy.](https://scvelo.readthedocs.io/getting_started.html)

2021-01-28 NEW ATTEMPT - forgot everything!

1. Download gene annotation from GENCODE or ENsembl

<http://velocyto.org/velocyto.py/tutorial/cli.html#download-genome-annotation-file>

1. Install velocyto and dependencies

Follow <http://velocyto.org/velocyto.py/install/index.html>

NOTE: pip install velocyto fails badly, because one if its dependencies (pysam) only works on Unix OS: <https://stackoverflow.com/questions/60197890/why-does-installing-pysam-python-package-fail>

SOLUTION: run in WLS (Windows Subsystem for Linux, to be installed).

>>> pip install numpy scipy cython numba matplotlib scikit-learn h5py click pysam velocyto

>>> sudo apt-get install samtools

1. Run velocyto

From a location where the annotated gtf, the bam and the barcodes are or are close to (in my case: /mnt/f/BAMS\_scRNAseq), run

>>> velocyto run -b 24h/filtered\_features\_bc\_matrix/barcodes.tsv -o 24h 24h/possorted\_genome\_bam.bam Mus\_musculus.GRCm38.102.gtf

NOTE: my folder structure in this case was:

ParentFolder

|\_Mus\_musculus.GRCm38.102.gtf

|\_24h

|\_possorted\_genome\_bam.bam

|\_filtered\_features\_bc\_matrix

|\_barcodes.tsv

NOTE: we are not using the -m option (that is, to mask repetitive elements)

NOTE: this will save the loom file in the same folder as the bam file

1. Run scvelo

Easy python script