# Week 1

# 24 apr. 17 Maandag 10u- 17:30

**Objectives:**

Bulk RNA seq inlezen

Pipelines verwerken en lezen hoe deze werken in genomics core

Single cell RNA seq principes en tools inlezen

Pipeline schrijven voor de VSC HPC

**Documentatie:**

HPC github

NGS github van genomcis core

Gegeven artikelen

**Verworven kennis:**

Waarom is single cell RNAseq zo interessant?

Bij bulk RNAseq is er het probleem dat de expressie van alle cellen worden gegeven in de omgeving van de tumor. Hierdoor wordt soms maar een gemiddelde gegeven van de werkelijke expressiedata.

Het implementeren van single cell RNAseq is een top-notch nieuwe techniek waarbij een cel wordt gebruikt om daaruit de RNAseq te maken. Het grote probleem hierin is dat er wel nog geen pipelines bestaan in de VSC HPC de bedoeling is dus een pipeline te schrijven en maken met de beste tools om interpretatie makkelijker te maken.

**Bulk RNAseq:**

Tools die kunnen worden gebruikt om de read quality te verhogen.

Het trimmen van de basen aangezien ze dichter bij de 3’ end van minder hoge kwaliteit zullen worden. Aan de hand van de software tools:

* FASTX-Toolkit
* Trimmomatic

Read alignment

Er zijn manieren waarbij de uiteindelijk verkregen data zal worden gemapt op een genoom of transcriptoom. Soms kan het zijn dat er uniek gemapt wordt en soms kan het zijn dat er een multiread verschijnt.

Tools voor identificatie:

GRIT zal mappen met de 5’ ends verkregen van CAGE of RAMPAGE.

Lage signaal detectie: Cufflinks, iReckon, SLIDE, StringTie

De novo reconstructie:

Tools gebruikt om een transcriptoom te maken zoals

* SOAPdenovoTrans
* Oases
* Trans-AbySS
* Trinity

Parallel end reads en long reads worden geprefereerd omdat deze meer informatie verzorgen.

Laag expressie reads zorgen voor problemen want deze zorgen voor een misassembly of grotere runtimes.

miRTools is een tool om voorspellingen en profiling van het sRNA species te maken.

BWA, STAR, Bowtie, TopHat zijn aligners gemaakt voor kleine reads te mappen op een referentiegenoom.

2 grote voordelen bij RNAseq vergeleken met array-based technologie om eQTL (expression quantitative trait loci):

1. Het kan varianten opsporen die transcript processing beïnvloeden
2. Reads die heterozygoot SNPs overlappen kan men mappen op het maternaal en/of paternaal chromosoom, quantificatie of allele specifiek expressie binnen individu kan gerealiseerd worden.

**scRNA:**

De keuze van protocol voor single-cell analyse hangt af van welke soort reads er willen bekomen worden in het experiment. Verschillende tools kunnen worden gebruikt in deze analyse.

Typisch gezien worden de libraries van de single-cell experimenten geanalyseerd door mapping op een referentie transcriptoom (bepaalde programma’s zijn hiervoor beschikbaar)

1 package mapt naar het genoom ook (**monocle).** Er is een grotere read-mapping rate als het genoom wordt gebruikt, maar meestal werd een referentie transcriptoom gebruikt om het simpel te houden.

Depending on the chosen single-cell protocol and the aims of the experiment, different bulk RNA-seq pipelines and tools can be used for different stages of the analysis as reviewed by Stegle et al. [[180](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8#CR180)]. Single-cell libraries are typically analyzed by mapping to a reference transcriptome (using a program such as RSEM) without any attempt at new transcript discovery, although at least one package maps to the genome (Monocle [[181](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8#CR181)]). While mapping onto the genome does result in a higher overall read-mapping rate, studies that are focused on gene expression alone with fewer reads per cell tend to use mapping to the reference transcriptome for the sake of simplicity. Other single-cell methods have been developed to measure single-cell DNA methylation [[182](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8#CR182)] and single-cell open chromatin using ATAC-seq [[183](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8#CR183), [184](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8#CR184)]. At present, we can measure only one functional genomic data-type at a time in the same single cell, but we can expect that in the near future we will be able to recover the transcriptome of a single cell simultaneously with additional functional data.

# 25 April 2017 Dinsdag 9u – 17.15

**Log in securen protocol:**

VSC log in securen met SSH

Ssh –keygen

Naam: VSCGenomics

Paswoord: Genomics

Nogmaals: Genomics

Key toevoegen aan de ssh directory:

Mv VSCGenomics ~/.ssh

Ssh-add ~/.ssh/VSCGenomics

Log in de HPC na toevoegen van de public key.

Ssh vsc31940@login2.hpc.kuleuven.be

**Lezen literatuur rond scRNA seq:**

<http://ac.els-cdn.com/S1097276517300497/1-s2.0-S1097276517300497-main.pdf?_tid=2faabecc-2988-11e7-a175-00000aacb35f&acdnat=1493105235_66e8c7a191ffeeed4e495d2d16357953>

vele technische variabelen:

1. De sensitiviteit van de scRNAseq methode
2. De accuracy
3. De precisie

De combinatie van deze factoren definieert hoe sterk een bepaalde methode kan zijn.

Smart-seq protocol is geoptimaliseerd volgens de sensitiviteit meer zelfs voor de full-lentgh coverage, accuracy en kost en dit ook voor het smart-seq2 protocol.

Andere protocols hebben full-length coverage opgegeven om de primer ook te kunnen sequencen voor cDNA aanmaak.

Smart-Seq2 heeft de hoogste sensitiviteit met een gemiddelde van 9138 genes per cell.

Overal performance is Smart-Seq2 de beste om te gebruiken in vergelijken met de andere 5 die vergeleken werden.

**Vergadering**

Vergadering rond de stand van zaken zoals: NIPT testen, hoe de SMRT-seq verloopt, herstellen van de pacbio.

RNA-seq methods for transcriptome analysis

<https://www.ncbi.nlm.nih.gov/pubmed/27198714>

R-script and Bioconductor usage voor RNA seq analyse

# 26 April 2017 Woensdag 8.15 – 17.30.

Vergadering om 15u:

Pipeline van Bulk RNA seq bekijken

Onderzoek pipeline van bulk RNAseq

Commando’s HPC

Voor kopieren: rsync

Module load …

Module unload …

Module purge (alle geladen modules worden uitgezet)

**PBS-scripts bekijken.**

(Portable batch systems).

Altijd starten met #!/bin/bash.

Laad de modules nodig voor je script.

Erna de header van je PBS.

Purge je modules en scratch node clean-up

PBS script

1. Include PBS header
2. Load software modules
3. Optimize I/O using a scratch node
4. Check if results are copied/stored in a data or staging directory Scratch is temporary, when the job is finished, data will be lost

Voorbeeld Header:

#!/bin/bash –l

PBS –l walltime 12:00:00

PBS –l mem=100gb

PBS –l nodes=1:ppn=20:ivybridge

PBS –M email adres voor mails te ontvange

PBS –m aeb (mail naar email adres als er iets (b)egint, (a)borted, (e)nds)

PBS – N jobnaam

PBS –A lp\_biogenomics (aan wie moeten de credits worden aangerekend?)

**Stel:**

1 berekening neemt ongeveer een 15 min in beslag van je walltime

stel je wilt 100 keer die berekening dan moet je bvb 8 CPUS nemen dan heb je nog 1500/7 = 215 minuten 🡪 4uur om toch zeer te zijn dus dan neem je ppn=8 en een walltime van 4u om een bepaald proces 100 keer te laten verlopen volgens je data.

Data export kan opgeslagen worden in een CSV formaat en met excel geopend worden voor tekst formats gebruikt GEEN word om dit te laten schrijven.

Gebruik scratch nodes om snel te werken en overbodige data niet in je folder te houden.

Job submission

**ThinKing:**

Qsub jobsnaam.pbs

**Cerebro**

Module load cerebro/2014a

Qsub jobsnaam.pbs

Kosten berekenen

Quotum krijgen hoeveel een proces kost in de supercomputer

Module load accounting

Gquote paths.pbs

Weten welke jobs gesubmit:

Qstat –u vsc31940

Showstart **nummer**

Checkjob **nummer**

Bepaald aantal contigs worden gevonden bij een de novo assembly.

**Tools gebruikt in scRNAseq:**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5030210/>

Pre-processing:

* Cutadapt (removes adapter sequences)
* Trimmomatic (quality-based trimming and removing adaptor sequences)

QC:

* FASTQC
* HTQC
* SolexaQA (alleen voor illumina QC, trimming en quality based filtering)
* SinQC en SCell (specifiek voor snRNA) (SCell doet aan outlier detection)
* Celloline detects low quality cells from expression profiles

Alignment:

* TopHat (capable of detecting novel splices)(widely used in scRNA)
* RSEM (includes an aligner RNA-Seq by Expectation Maximization)(used in scRNA)
* STAR (kan ook met scRNA)
* MapSplice
* GSNAP
* HISAT en Kallisto (nieuwe tools)
* Snelste tools hier zijn TopHat en STAR de nieuwe hebben ook goeie vooruitzichten

# 27 apr. 17 Donderdag 8:45-17.30

Quantification (is converting alignment results in a gene expression profile):

Meestal dezelfde algoritmen gebruikt als in RNA-seq wordt gebruikt

* HTSeq (gebruikt in de verwerking van scRNA)
* FeatureCounts
* RSEM en Cufflinks (maximum likelihood approach)
* Kallisto (blijkbaar snel)

Gene filtering nodig door de hoge noise in RNA-seq datasets eruit te filteren van low quality genes and samples. Gene is pas expressed na een bepaalde threshold FPKM (fragments per kilobase per million reads):

* OEFinder

Normalization

* GRM zorg voor de FPKM, spike ins, RPKM, TPM
* BASICS
* SAMstrt

Differential expression niet zeker welke soort tools een betere performance geeft

* DESeq2
* EdgeR
* MAST
* BASICS

Artikels rond pipeline en scRNA-seq verwerking lezen.

**Presentatie bijwonen:**

rond NIPT testen bij zwangerschap en prevalentie in klinische relevantie rond NRXN1 deletie

**Tutorial VSC HPC**

VSC jobs uitvoeren rond mapping, denovo assembly

Vergadering rond bulk RNA seq pipeline. (welke tools en hoe dit werkt)

MacBook-Air-van-Niels:Documents nielsvanneste$ rsync --progress --copy-links -ahr vsc31940@login.hpc.kuleuven.be:/data/leuven/319/vsc31940/vsc\_ngs\_workshop/results/mapping/test\_ncbi\_sample5.bam ./VSC\_HPC/

Rsync altijd gebruiken vanuit de computer zelf en inloggen in de vsc als je daar bestanden van wilt halen.

Source 🡪 destination laten kopiëren.

# 28 April 2017 Vrijdag 8:45 – 15:45

workflows rond scRNAseq bekijken

FASTQC tool onderzoeken hoe dit werkt.

**Dataset van bulkRNA analyseren (zelf)**

Data van giorgia gebruiken om eventueel scRNA testen te doen of workflow op te maken.

Bulk RNAseq

Online naar scRNAseq data zoeken om de workflow op toe te passen en aan te nemen welke tools best gebruikt worden.

Eventueel scRNAseq data gebruiken van genomics core zelf om te gebruiken.

SCDE is goed voor scRNAseq (check mail Alvaro)

Hemberg lab 🡪 workflow rond scRNAseq

**Grote verschil van scRNA met bulk RNA is dat deze een library maakt van één bepaalde cel en geen populatie aan cellen. Significant meer aandacht moet genomen worden voor de verschillende celexpressie.**

Door normalisatie van de data kunnen de discrepanties ontwekene worden.

FASTQC script maken

MAPPEN van de files RNA-seq

/staging/leuven/stg\_00019/genome

gtf file mus musculus

# Week 2

# 2 mei 2017 Dinsdag 8:45 – 17:30

mapping script debuggen.

* <http://labshare.cshl.edu/shares/gingeraslab/www-data/dobin/STAR/STAR.posix/doc/STARmanual.pdf>
* <http://hemberg-lab.github.io/scRNA.seq.course/construction-of-expression-matrix.html>
* <https://www.nature.com/nmeth/journal/v11/n7/pdf/nmeth.2967.pdf>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531809/>
* [http://download.springer.com/static/pdf/956/art%253A10.1186%252Fs13059-016-0970-8.pdf?originUrl=http%3A%2F%2Fgenomebiology.biomedcentral.com%2Farticle%2F10.1186%2Fs13059-016-0970-8&token2=exp=1493290153~acl=%2Fstatic%2Fpdf%2F956%2Fart%25253A10.1186%25252Fs13059-016-0970-8.pdf\*~hmac=1196e7404f527fab5bfee0d5d3ef3f89d24c82e857070988272e855b5b3d3484](http://download.springer.com/static/pdf/956/art%253A10.1186%252Fs13059-016-0970-8.pdf?originUrl=http%3A%2F%2Fgenomebiology.biomedcentral.com%2Farticle%2F10.1186%2Fs13059-016-0970-8&token2=exp=1493290153~acl=%2Fstatic%2Fpdf%2F956%2Fart%25253A10.1186%25252Fs13059-016-0970-8.pdf*~hmac=1196e7404f527fab5bfee0d5d3ef3f89d24c82e857070988272e855b5b3d3484)
* [http://download.springer.com/static/pdf/187/art%253A10.1186%252Fs13059-016-0881-8.pdf?originUrl=http%3A%2F%2Fgenomebiology.biomedcentral.com%2Farticle%2F10.1186%2Fs13059-016-0881-8&token2=exp=1493280906~acl=%2Fstatic%2Fpdf%2F187%2Fart%25253A10.1186%25252Fs13059-016-0881-8.pdf\*~hmac=bd5648ab224597de05d15b9cee34c8c79390dc4c95bdb9c4ed527ccbd4dd8fd7](http://download.springer.com/static/pdf/187/art%253A10.1186%252Fs13059-016-0881-8.pdf?originUrl=http%3A%2F%2Fgenomebiology.biomedcentral.com%2Farticle%2F10.1186%2Fs13059-016-0881-8&token2=exp=1493280906~acl=%2Fstatic%2Fpdf%2F187%2Fart%25253A10.1186%25252Fs13059-016-0881-8.pdf*~hmac=bd5648ab224597de05d15b9cee34c8c79390dc4c95bdb9c4ed527ccbd4dd8fd7)
* <https://www.nature.com/nrg/journal/v16/n3/pdf/nrg3833.pdf>

debuggen indexing script. FASTA file te hoge permissies

Tekenen dat een scRNA library niet goed is:

* De fractie aan reads die mappen naar het genoom van het organisme, kan via FASTQC direct opgespoord worden. Lage fraction betekent RNA degradatie, contaminatie, ineffecient lysis.
* Door HTSeq. Hoge proportie mapped naar reads van de spike-ins wil zeggen lage kwaliteit aan RNA.
* PCA. Good-quality cells clusteren samen en slechte zie je als outliers.

Mapping PBS script automatisch mappen proberen te maken.

# 3 mei 2017 Woensdag 8:15 – 17:45

debuggen automatische mapping script

aanmaak sorted bam file van mapping die is gelukt

sshfs mounten

met IGV de reads nakijken

alle reads counten aan de hand van HTSeq.count

**alle sample namen verzamelen om dynamische mapper te maken**

**cd fastq\_files**

ll | cut -d' ' -f9 > datafile.txt

**Dynamische mapper script schrijven**

Cd mapping

Bam files in een file opslaan

ll | cut -d' ' -f10 > bamfiles.txt

Door probleem alleen de bam uit de file halen

grep -f bamfiles.txt \*.bam

zorgen dat de grep matches tekst niet worden meegenomen.

less bam.txt | cut -d' ' -f3 > bams.txt

schrijven HTSeq\_count script voor alle data

debuggen HTSeq\_count script.

# 4 mei 2017 Donderdag 8:45 – 17:30

debuggin HTSeq-count script (; was de fout)

tweaken van HTSeq en mapping script

integreren van batch file

uitzoeken verdere verwerking met SCDE, EdgeR, DESeq2

+ debuggen script DESeq

awk moet file “” laten anders veranderd het van naam en problemen bij de csv opmaak

conditions.txt de condities waartegen getest en welke soort materiaal

sample\_condition.txt sample namen en hun conditie worden opgeslagen

script maakt zelf een samples.csv aan

FASTQC script dynamisch maken

scRNA dataset zoeken.

# 5 mei 2017 Vrijdag 9:00 – 16:45

scRNA dataset zoeken

bespreking van verschillende analyse tools voor scRNA

* SCDE
* DESeq2
* Census
* Monocle

**24 mei voorstelling welke beste en makkelijkste te gebruiken tools voor analyze**

[**https://hemberg-lab.github.io/scRNA.seq.datasets/human/pancreas/**](https://hemberg-lab.github.io/scRNA.seq.datasets/human/pancreas/)

Segerstolpe

Normal ok alpha cells

Tegen alpha cells met type II diabetes

Tools die onderzocht moeten worden

<http://cole-trapnell-lab.github.io/monocle-release/getting-started/>

<https://www.nature.com/nmeth/journal/v14/n3/full/nmeth.4150.html>

<http://hms-dbmi.github.io/scde/>

DESeq2

EdgeR

Homo sapiens indexeren met STAR

Fastqc uitvoeren op gekozen cellen van de Segerstolpe diabetes scRNA

Mapping uitvoeren op de RNA

HTSeq count uitvoeren op de RNA

# Week 3

# 8 mei 2017 Maandag 8u – 17:00

**HTSeq count script debuggen**

Count documenten waren leeg door een verkeerde verwijzing naar de HTSeq genome directory

Allerlei soorten reads forward en reverse strand aanwezig in de bam file

* Artikels lezen
* IGV reads bekijken

HTSeq count moet naar de originele GTF file verwijzen voor gene naam en de aantal reads te kunnen tellen.

DESeq uitvoeren op de scRNA data.

Artikel rond dataset lezen.

**Interne audit school BANABA BIT**

# 9 mei ’17 Dinsdag 9:00 – 17:30

**Census lezen tool voor scRNA seq**

<file:///Users/nielsvanneste/Documents/Genomics_core/nmeth.4150.pdf>

census is een tool die verkrijgbaar is uit de monocle2 library.

De nieuwe methode van single-cell RNA-seq zou gen regulatorische veranderingen kunnen detecteren in onderzoek progressie. Deze zijn echter moeilijk af te lezen door de hoge noise in single-cell measurements.

Problemen zijn de hoge ‘drop-outs’

Census zorgt dat transcripts per million (TPM) worden gecount zonder dat spike-in standards of UMI’s nodig zijn. Nieuwe regressiemodel ook aanwezig branch expression analysis modelling (BEAM). Deze tools zijn geimplementeerd in monocle2.

Het Census package kan samenwerken met verschillende soorten scRNA protocol. Census is accuraat zonder dat er spike-in controles nodig zijn.

De tool werd vergeleken met SCDE en monocle zelf.

Vergelijking met DESeq2 gaf toch 62% FDR. SCDE deed het goed, maar toch een paar kleine FP. Census komt niet overeen met de gevonden gen expressie met andere tools (34%).

**Monocle uitzoeken hoe dit werkt**

**Abstract schrijven**

**SCDE artikel lezen en code bekijken**

# 10 mei 2017 Woensdag 8:15 – 17:15

Uitzoeken hoe de counts worden gebundeld.

**Vergadering:**

Abstract voorbeeld van opwekken zoals kanker, diabetes.

Implementeren op HPC moet er ook bij voor industrieel formaat automatiseren van analyse.

Cut commando voor het samenvoegen van de genen eerst cat commando.

**Single cell analysis using GenePattern**

Genepattern geeft een hele serie aan tools 🡪 modules. NGS, microarray, mass spectroscopy,…

Parallel werken is mogelijk en output wordt op de server opgeslagen voor een week. Data kan ook permanent worden gestored.

Genepattern notebook is gebaseerd op jupyter. Modules zijn gewrapped in perl, python, R of java.

Genepattern gebruikt HTSeq-count. Same look as Galaxy, but is easier to install modules

VIB.

**HPC**

SCDE niet installeren op R

Monocle werkt niet

https://unix.stackexchange.com/questions/16443/combine-text-files-column-wise

paste \*count.txt

**Namen van counts opslaan**

ll | cut -d' ' -f14 > names.txt

paste \*\_count > data.txt

**tab delimited**

cut -f1,2,4,6,8,10,12,14,16,18,20 data.txt >> all\_data.txt

more all\_data.txt

read.delim gebruiken

SCDE op laptop testen

edgeR op HPC uitvoeren

# 11 mei 2017 Donderdag 8:45 – 17:15

**SCDE op laptop eens uitvoeren (HPC kan momenteel geen SCDE aan)**

data <- read.delim("all\_data.txt", row.names = "Gene")

**Werking EdgeR uitzoeken**

EdgeR kan analyseren op basis van counts waarin de rij het gen voorstelt en de kolom is je individuele library.

**Belangrijke repository:**

/staging/leuven/stg\_00019/genbas/quantseq/jobs/rnaseq

**module load worker**

EdgeR heeft –wsub <pbs script nodig> -data conditions.txt

Sample\_conditions.csv in de output dir

**Learning pika**

Monocle2 op module load R/3.3.2-foss-2015a-tcl

SCDE, edgeR, DESeq2 op module load R/3.3.1-foss-2015a-tcl

Voor het EdgeR script misschien best eens een DGEList uittesten op de hele set en kijken ofdat deze error er nog steeds is?

EdgeR werkt door conditionsfile te implementeren

Userguide van edgeR lezen ter interpretatie

Reads counten fastq files

cat ERR1630037.fastq | echo $(wc -l)

READS VAN scRNA SEQ MOGEN ONDER DE MILJOEN ZIJN

scRNA seq reads gaan van 1-20 miljoen reads

# 12 mei 2017 Vrijdag 9:00 – 15:45

SCDE uitzoeken script verder verwerken. Guide lezen en op de HPC proberen uit te voeren.

R script op de hpc schrijven

PBS script voor scde op HPC schrijven

Debugging SCDE

Error moet nog worden gevonden

Monocle vignettes lezen

Disp\_table aanmaken!!

# Week 4

# 15 mei ’17 Maandag 8:45 – 17:15

Monocle op laptop R script maken en toepassen.

Expressed genes lager dan 10 aangezien ik slecht 9 cellen heb gebruikt in de voorbeeld analyze.

Monocle2 gene annotation file openen

Script aanpassen om cellen te kunnen differentieren.

**SCDE lager pakket qua R gebruiken!!! Misschien werkt dit dan wel (Alvaro werkte dit)**

**R 3.3.2 heeft problemen met matrix voor SCDE heel veel RAM nodig en processoren en de data moet kleiner zijn**

# 16 mei 2017 Dinsdag 8:45 – 17:15

Betekenis van DESeq2 output nagaan

Betekenis van EdgeR output nagaan

Powerpoint van verschillende anayser tools maken

Tussentijds resultaten bespreken.

Data verkregen van een scRNA experiment werd geanalyseerd met SCDE. Acinaire cellen hebben hoge differentiele expressie 🡪 proberen te reproduceren.

Gamma cellen eens uitproberen op differentiele expressie te produceren.

ROC curves opgezocht van EdgeR, monocle, MAST

Hemberg-lab testen van SCDE, EdgeR, monocle op FDR en TPR

**Hoe hoger fold change hoe meer tot expressie in de 2de omstandigheid hoe lager de fold change hoe minder tot expressie in de 2de omstandigheid.**

TPR 🡪 EdgeR 84%

TPR 🡪 monocle 82%

TPR 🡪 SCDE 82% (wel problemen met welke soort R version er wordt gebruikt)

MONOCLE2

DifferentialGeneTest gebruik hiervoor fullmodelformulastr ~CONDITION

**Powerpoint punten!!!**

**Optimale reads**

**Waarom niet deseq2?**

**Verschillende tools presenteren**

**Doelstellingen**

2-10 million reads zijn optimaal (A survey of best practices for RNA-seq data analysis)

MONOCLE2werkt

Debuggen monocle 2 op de hpc

# 17 mei 2017 Woensdag 8:15 –

Aanmaak powerpoint rond single cell analyse tools

Debuggen van monocle2 op HPC

Zorgen dat alle namen overeenkomen in het R script

Andere scRNA data gebruiken op de gemaakte scripts.

Interpretatie van de resultaten en vignettes van de verschillende analyse tools lezen.

Meer data om differentiele gen expressie op te bepalen vinden

Normalizeren gebeurt automatisch van monocle2 als je dit zelf doet kan het algoritme misschien zelfs blokkeren.

**Belangrijke commando’s data samenbrengen voor monocle2:**

* Echo –en Genes \*\_count.txt “\n” > all\_data.txt
* Paste \*\_count.txt | cut –f1,2,4,6,8,10,12,14,16,18,20,… >> all\_data.txt
* PROBLEEM == de header is niet tab separated

# Belangrijke termen

* **Unique molecular identifier (UMI):** korte DNA sequenties die in de interessante DNA moleculen geïncorporeerd zijn. Deze zorgen voor biases.
* **Spike-in:** gekende types en hoeveelheid RNA met die als interne controles worden gebruikt in de experimenten.
* **External RNA Control Consortium (ERCC)**: wijd gebruikt RNA spike-in (artificieel)
* **Read alignment:** short reads die gealigneerd worden op een ref genome of transcriptome.
* **Principal component analysis (PCA)**: een methode om complexe data simpeler te maken. Transformeren naar kleinere data
* **STAR** = spliced transcripts alignment to a reference
* **SAGE** = Serial Analysis of Gene Expression
* **FPKM** (fragments per kilobase per million reads)
* **RPKM** (reads per kilobase per million reads)
* **Confounding factors:** unobserved covariates that affect gene expression and can obscure interpretation if not accounted for
* **Batch effects:** verschil in gene expressie tussen verschillende cellen van de populatie door sample voorbereiding.
* **Drop-outs:** cellen waarbij er genen zijn die heel veel expressie aantonen en cellen waarbij zelfde genen geen expressie geven