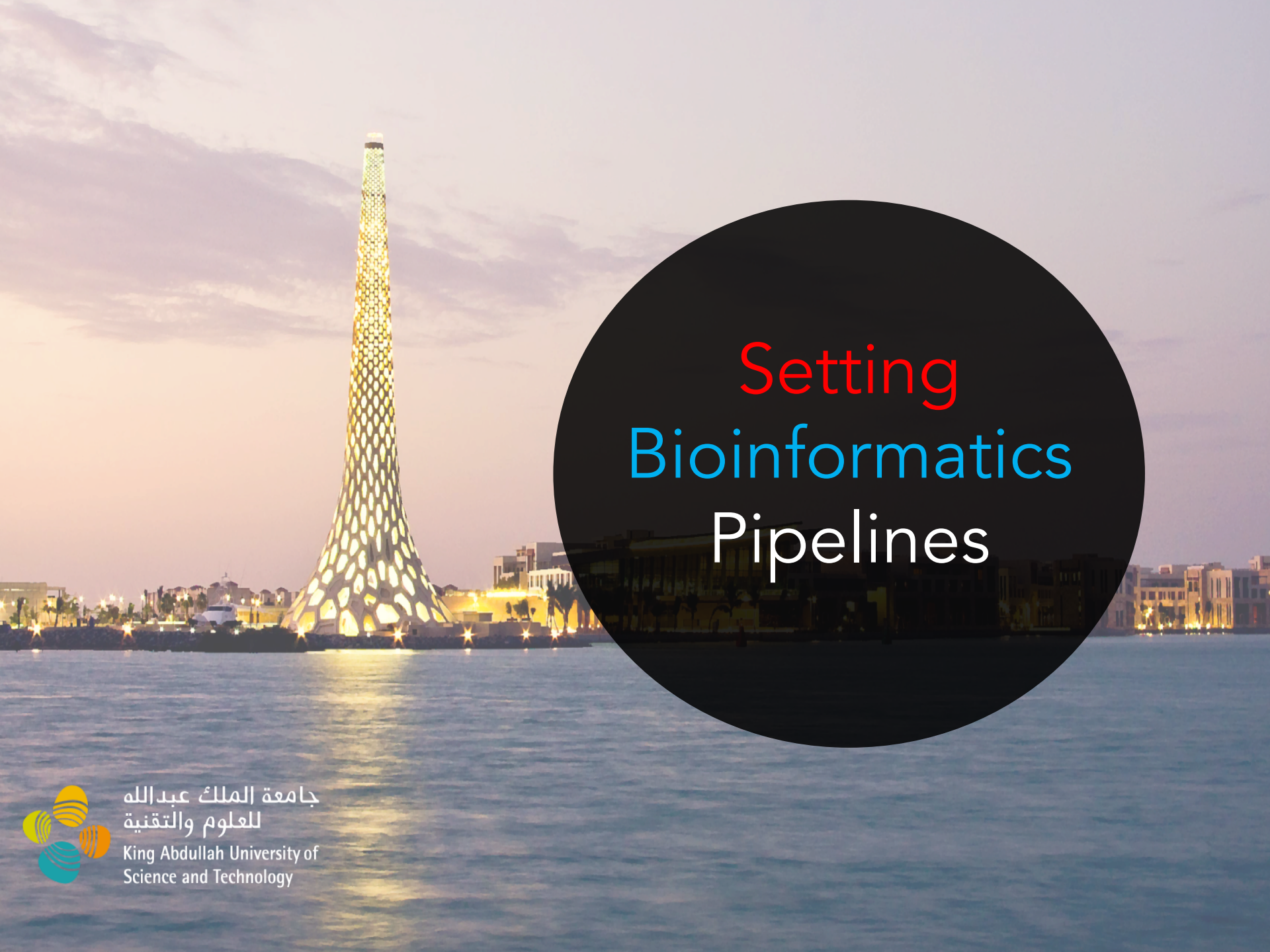

Setting
Bioinformatics
Pipelines
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Setting Bioinformatics Pipelines

Reasoning with data

David Gomez-Cabrero
Jesper N Tegner
Vincenzo Lagani
Robert Lehmann

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Tentative Course Schedule

(Time, topic/emphasis & resources)

Week	Lectures	Topic
1	Tue 01/23/2024	Pipeline analysis: what is it and general guidelines for mastering pipelines. Investigating our first pipeline: RNA-seq. RNA-seq hands-on.
2	Tue 01/30/2024	Setting a pipeline: code, Github, tracing back analysis, Reimplement the RNA-seq pipeline over the new concepts. Review the statistics behind count data I.
3	Tue 02/06/2024	ATAC-seq. Understanding and implementing the pipeline Review the statistics behind count data II.
4	Tue 02/13/2024	Single-cell RNA-seq. Understanding and implementing the pipeline.
5	No schedule	No class
6	Tue 02/27/2024	Single-cell multi-ome. Understanding and implementing the pipeline.
7	Tue 03/05/2024	No class
8	Tue 03/12/2024	Setting an integrative pipeline.
9	Tue 03/19/2024	DNA Methylation. Understanding and implementing the pipeline.
10	Tue 03/26/2024	No class
11	Tue 04/02/2024	No class
12	No schedule	No class
13	Tue 04/16/2024	No class
14	Tue 04/23/2024	No class
15	Tue 04/30/2024	No class
16	Tue 05/07/2024	Per groups a single-pipeline: HiC, CITE-seq, CyToF, etc.

EVALUATION

(1) CLASSES: assistance, participation, hands-on,…

(2) HOMEWORK: provide notes from the classes.

Understanding is most important!!!

- Hands-on + interpretation / exercises + interpretation
- Group exercises.
- Deliver to david.gomezcabrero@kaust.edu.sa
- Subject: BESE394A SBP – “name of student”
- Submit “per day”

(3) FINAL TASK: analysis of a data-set.

- We will provide one set,
- You can provide one of your interest.
- 2 hours duration + 6 days to deliver.
- Data analysis: analysis AND interpretation.

Information updated

Lecture time:

- 8:30 - 11:30 then lunch break.
- 12:30 – 15:30 (*if seminar we will adapt*)
- A homework will be described.

Please join the slack channel

bese394a_setting_bioinfopipelines

Setting
Bioinformatics
Pipelines

Questions?

Outline Day 1

Pipeline analysis: what is it and general guidelines for mastering pipelines.

Investigating our first pipeline: RNA-seq.

RNA-seq hands-on.

WHAT DO WE AIM FOR?

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Reason with (*bio*) data

What are our questions?

Right way of asking the questions.

Setting a plan.

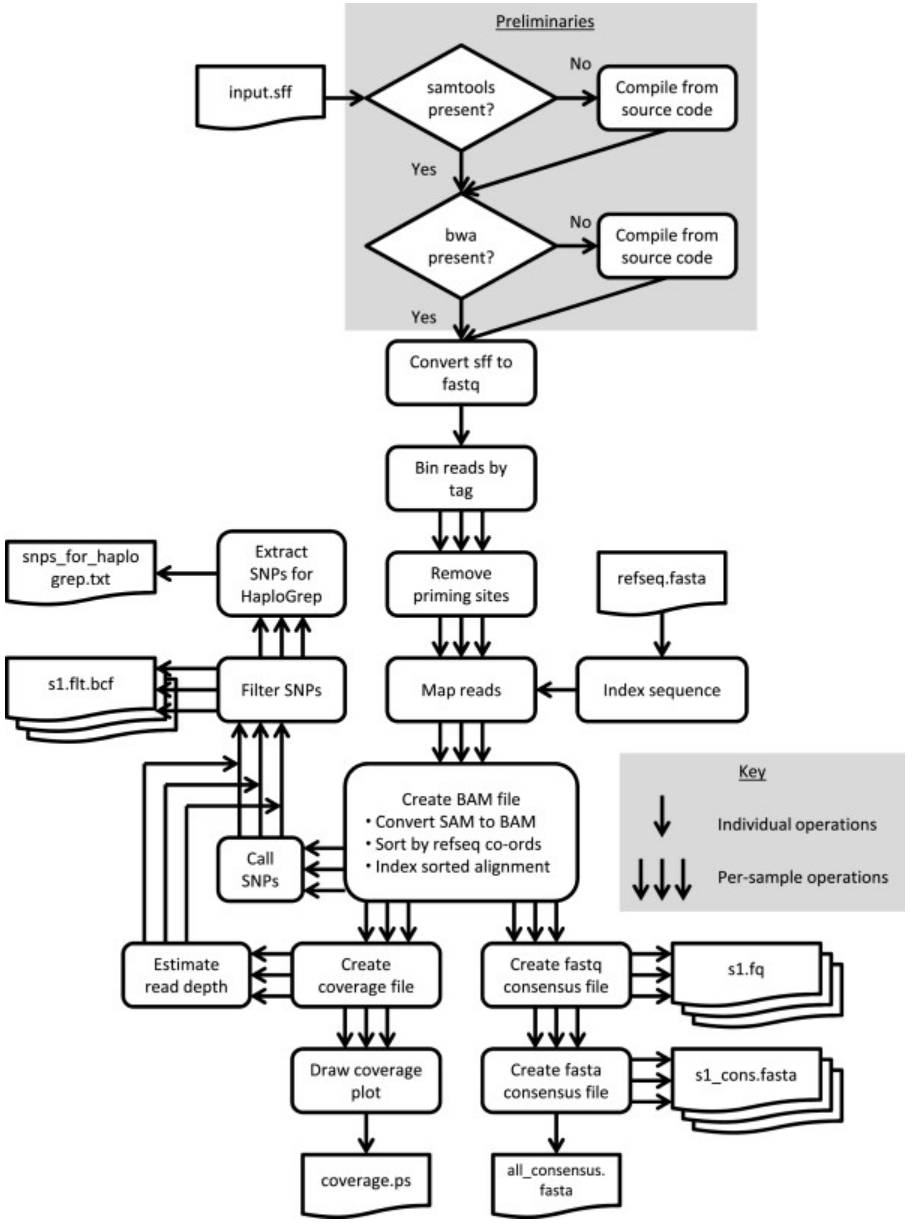
And more…

We want **you** to **think**

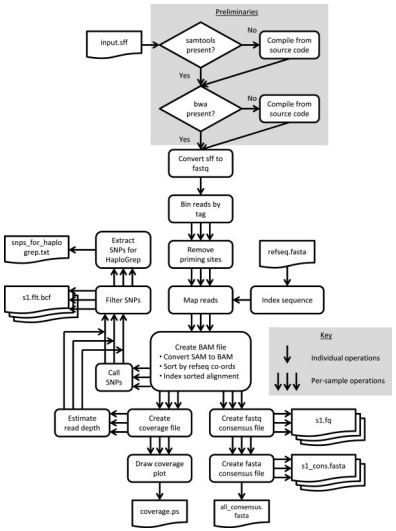
Pipeline analysis: what is it and general guidelines for mastering pipelines.

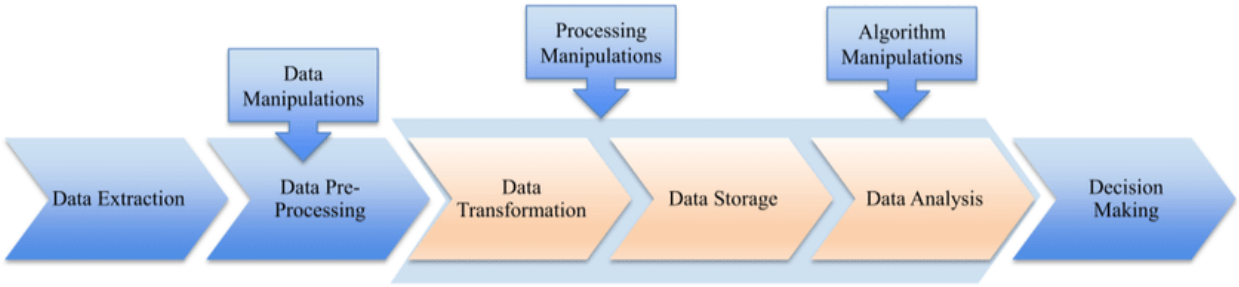
جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

WHAT IS A PIPELINE?



WHAT IS A PIPELINE?





A pipeline is a **set of processes and tools** used to collect raw data from multiple sources, analyze it and present the results in an understandable format.

Companies use data pipelines to **answer specific business questions** and make strategic decisions based on real data. All available **data sets** (internal or external) are analyzed to obtain this information.

https://datascientest.com/en/all-about-data-pipeline

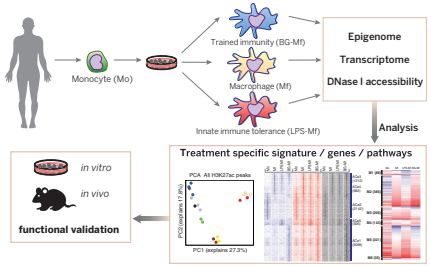
EXERCISE: Identify RNA-seq pipeline

RESEARCH ARTICLE SUMMARY

IMMUNOGENETICS

Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity

DATA SET **GSE58310 (2014)**

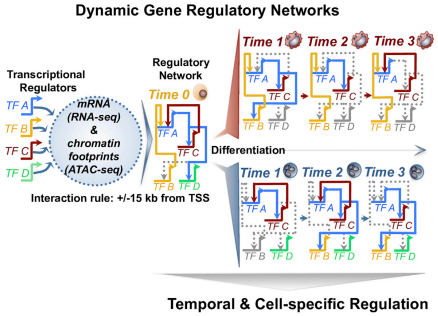


The epigenome, DNase I accessibility, and transcriptome were characterized in purified human circulating monocytes, in vitro differentiated naïve, tolerized (immunosuppression), and trained macrophages (innate immune memory). This allowed the identification of pathways functionally implicated in innate immune memory. This epigenetic signature of human monocyte-to-macrophage differentiation and monocyte training generates hypotheses to understand and manipulate medically relevant immune conditions.

Cell Systems

Dynamic Gene Regulatory Networks of Human Myeloid Differentiation

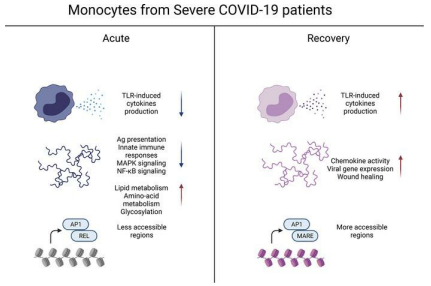
DATA SET **GSE79044 (2017)**



JCI insight

Functional reprogramming of monocytes in patients with acute and convalescent severe COVID-19

DATA SET GSE198256 **(2022)**

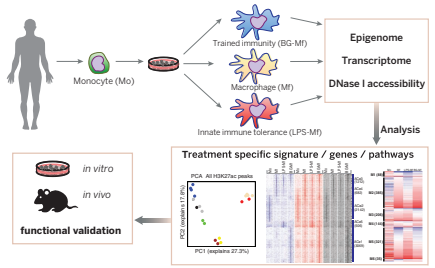


RESEARCH ARTICLE SUMMARY

IMMUNOGENETICS

Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity

DATA SET [GSE58310](https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE58310) **(2014)**



The epigenome, DNase I accessibility, and transcriptome were characterized in purified human circulating monocytes, in vitro differentiated naïve, tolerized (immunosuppression), and trained macrophages (innate immune memory). This allowed the identification of pathways functionally implicated in innate immune memory. This epigenetic signature of human monocyte-to-macrophage differentiation and monocyte training generates hypotheses to understand and manipulate medically relevant immune conditions.

Fastq

RNA-seq reads **were aligned using** GSNAP (*65*) using non-default parameters -m 1 - N 1 -n 1 -Q -s Ensembl_splice_68

RNA-seq library data **were initially subjected to a quality control step**, where, based on read distribution over the annotated genome, libraries that are outliers were identified and discarded from further analysis.

Reads were aligned to the Ensembl v68 human transcriptome using Bowtie. Quantification of gene expression was performed using MMSEQ (*31*).

Differential expression was determined using MMDIFF (*32*). A two model comparison was used to identify differentially expressed genes that confer cellular identity Mo/Mf.

Genes with a larger posterior probability for the second model, an RPKM value greater than 2 in any of Mo or Mf and minimally a two-fold expression change were considered as differentially expressed. We calculated the Bayes factor for each model by comparing the differential expression models to the (reference) null-model and applied Bayes' theorem to compute the posterior probability of each model per gene. The expression change directionality of a gene was determined based on the model with the highest posterior probability having to be at least 0.35 for the gene to be considered for downstream analyses.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

SUMMARIZE

1. READS: fastq

2. No QC here (?)

3. MAPPING:

65 -> GSNAP -> ?

QC -> (?)

31 -> MMSEQ ->

https://github.com/eturro/mmseq

4. DIFF EXPRESSION

32 -> MMDIF (?) ->

Larger poster probability 0.35

RPKM > 2

2-fold expression change

5. CLUSTERING

Manual.

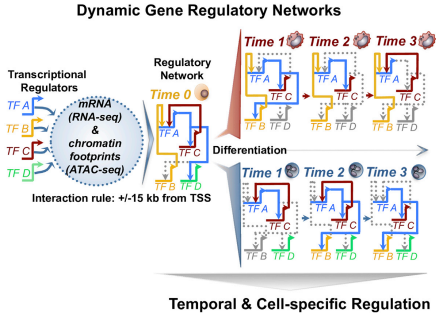
https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE58310

Cell Systems

Dynamic Gene Regulatory Networks of Human Myeloid Differentiation

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

DATA SET **GSE79044 (2017)**



Fastq 1 billion

hg38 reference genome using **STAR** (Dobin et al., 2013) aligner and **mapped to Gencode version 20 gene annotations using Cufflinks** (Trapnell et al., 2010)

Batch effects due to the generation of libraries were considered and cor- rected for using **Combat** (Johnson et al., 2007

Batch-corrected data were **normalized using TMM function in EdgeR** strictly(Robin- son et al., 2010).

maSigPro (Nueda et al., 2014) allows for a two-step regression modeling strategy, which was used in identifying gene expression dynamics across differentiation of all lineages. An alpha of 0.05 for multiple hypothesis testing and a false discovery control of 1% were used, in both gene and transcription factor analysis.

A k-cluster of 13 was selected based on previous analysis using hierarchical clustering and k-means clustering on the entire dataset. Gene ontology enrichments were determined for each cluster using DAVID (Huang et al., 2007). Gene expression heatmaps were generated using Tree View 3.0 (http://bonsai.hgc.jp/mdehoon/software/cluster/software.htm) and using R software.

SUMMARIZE

1. **READS:** fastq

2. **No QC here** (?)

3. **MAPPING:**
   STAR -> hg38
   Cufflinks -> Gencode 20
   QC?

4. **Batch correction:**
   Combat
   QC?

5. **DIFF EXPRESSION**

   MaSigPro: alpha of 0.05 for multiple hypothesis testing and a false discovery control of 1%

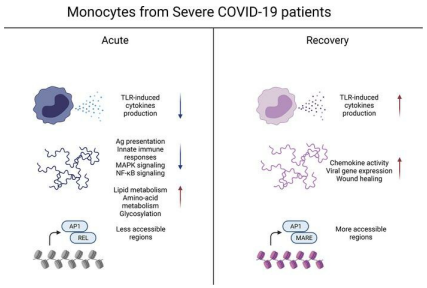
5. **CLUSTERING**

   k-cluster of 13 hierarchical clustering
   k-means clustering on the entire dataset.
   DAVID -> Gene Ontology enrichment

JCI insight
Functional reprogramming of monocytes in patients with acute and convalescent severe COVID-19

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

DATA SET GSE198256 (2022)



Adapters were removed with Trimmomatic-0.36 with the following parame- ters: Truseq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 HEAD-CROP:4.

Reads were then mapped to the reference genome GRCh38 by using STAR_2.5.3a software with default parameters.

Read counts in the alignment BAM files that overlap with the gene features were obtained using **HTSeq-0.9.1 with "--nonunique all" option**

Genes with no raw read count greater or equal to 20 in at least 1 sample were filtered out with an R script

Raw read counts were **normalized**, and a **differential expression analysis** was performed with DESeq2 by applying an **adjusted *P* < 0.05 and an absolute log2 ratio larger than 1**

SUMMARIZE

1. **READS:** fastq
   Adapters removed

2. **No QC here** (?)

3. **MAPPING:**
   STAR 2.5.3a -> hg38
   HTSeq-0.9.1 ""**--nonunique all**" QC?

4. **Batch correction:**

5. **DIFF EXPRESSION**
   Genes: counts read >= 20 in at least 1 sample
   DESeq2: normalization
   DESeq2: differential expression. Adjp <0.05 log2 > 1

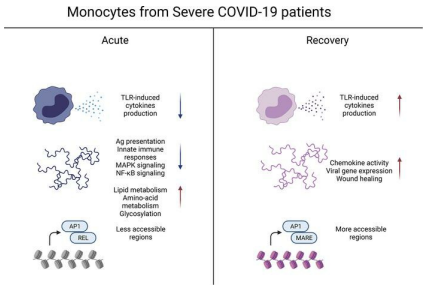
6. **Unclear from here.**
   Not in methods: GSEA BubbleGUM BMC Genomics 2015

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE198256

JCI insight
Functional reprogramming of monocytes in patients with acute and convalescent severe COVID-19

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

DATA SET GSE198256 (2022)



LETS DO OUR OWN ANALYSIS

1. **READS:** fastq
   Adapters removed

2. **No QC here** (?)

3. **MAPPING:**
   STAR 2.5.3a -> hg38
   HTSeq-0.9.1 ““**--nonunique all**” QC?

From here

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE198256

Pipeline analysis: what is it and general guidelines for mastering pipelines.

Relevant aspects for a bioinformatic pipeline:

- **Clarify** the steps.
- **Quality control**: where and how it happens
- **Details on the methods** and tools used.
- **Custom code**: annotated
- WHAT ELSE?

RNA-seq pipeline EXERCISES PER GROUP

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Part 1

1. Explain TMM
2. Explain Voom
3. Explain limma Trend
4. Explain Wald-test in DESeq2
5. Explain LTR in DESeq2

Part 2

1. Generate a tutorial for (beautiful) Heatmaps.
2. Generate a tutorial for PCA in RNA-seq data.
3. Generate a tutorial for ORA
4. Generate a tutorial for GSEA
5. How to select the genes to be included in the analysis based on # reads

Including interpretation!!

RNA-seq pipeline INDIVIDUAL

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Compare “Results from JCI Insight” vs “Results derived from your own analysis”

Identify a data-set related to Monocytes in bulk RNA-seq and run the complete pipeline. Generate a report.

Prepare an organized pipeline for RNA-seq.

- Power point explaining the pipeline.
- Example with R.

Extra mile:

Process the fastq to count table from of the three data-sets. Re-do the analysis and generate a report:

- Code
- Outcomes of the analysis.