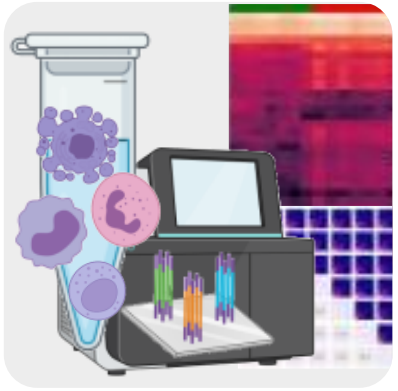
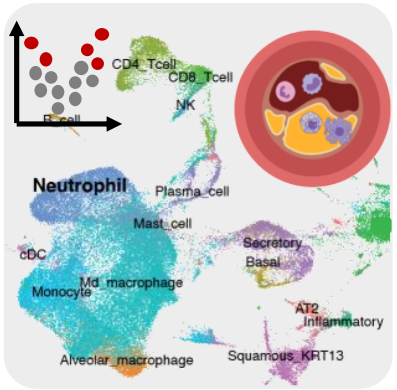




GAME-CIBOG WORKSHOP

RNA-seq
scRNA-seq

CURRICULUM VITAE

Markus Joppich

1989
- Born in Bonn

2012
- B.Sc. Computer Science at RWTH AACHEN UNIVERSITY
- iGEM Aachen 2014 Team Member
  - Best Measurement Project & Best Supporting Software

2015
- M.Sc. Computer Science at RWTH AACHEN UNIVERSITY
- Started PhD at LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN, CRC 1123 - ATHEROSCLEROSIS, MECHANISMS AND NETWORKS OF NOVEL THERAPEUTIC TARGETS

2018
- iGEM Munich 2018 Supervisor
  - Finalist, Best Manufacturing Project, Best Software

2021
- Defended PhD in November 2021
- PostDoc + Habilitation on scRNA-seq, causal networks

mjoppich

@mjoppich

C:\> Linux on Windows? Totally.

Received: 30 September 2020 | Accepted: 10 November 2020
DOI: 10.1111/jth.15179

ORIGINAL ARTICLE

jth

Vascular neutrophilic inflammation and immunothrombosis distinguish severe COVID-19 from influenza pneumonia

Leo Nicolai[1,2,3] | Alexander Leunig[1,2] | Sophia Brambs[1] | Rainer Kaiser[1,2,3] | Markus Joppich[4] | Marie-Louise Hoffknecht[1] | Christoph Gold[1] | Anouk Engel[1] | Vivien Polewka[1] | Maximilian Muenchhoff[3,5,6] | Johannes C. Hellmuth[3,7,8] | Adrian Ruhle[3,5] | Stephan Ledderose[9] | Tobias Weinberger[1,2,3] | Heiko Schulz[9] | Clemens Scherer[1,2,3] | Martina Rudelius[9] | Michael Zoller[10] | Oliver T. Keppler[3,5,6] | Bernhard Zwißler[10] | Michael von Bergwelt-Baildon[3,7,8] | Stefan Kääb[1,2,3] | Ralf Zimmer[4] | Roman D. Bülow[11] | Saskia von Stillfried[11] | Peter Boor[11,12] | Steffen Massberg[1,2,3] | Kami Pekayvaz[1,2,3] | Konstantin Stark[1,2,3]

JCI insight — RESEARCH ARTICLE

Self-sustaining IL-8 loops drive a prothrombotic neutrophil phenotype in severe COVID-19

Rainer Kaiser,[1,2,3] Alexander Leunig,[1,2] Kami Pekayvaz,[1,2,3] Oliver Popp,[4,5] Markus Joppich,[6] Vivien Polewka,[1] Raphael Escaig,[1] Afra Anjum,[1] Marie-Louise Hoffknecht,[1] Christoph Gold,[1] Sophia Brambs,[1] Anouk Engel,[1] Sven Stockhausen,[1,2] Viktoria Knottenberg,[1] Anna Titova,[1] Mohamed Haji,[4,5] Clemens Scherer,[1,2,3] Maximilian Muenchhoff,[3,7,8] Johannes C. Hellmuth,[3,9] Kathrin Saar,[4,5] Benjamin Schubert,[10,11,12] Anne Hilgendorff,[12,13,14] Christian Schulz,[1,2] Stefan Kääb,[1,2,3] Ralf Zimmer,[4] Norbert Hübner,[4,5,15] Steffen Massberg,[1,2,3] Philipp Mertins,[4,5] Leo Nicolai,[1,2,3] and Konstantin Stark[1,2,3]

Title

Protective immune trajectories in early viral containment of non-pneumonic SARS-CoV-2 infection

Author list

Kami Pekayvaz[1,2,#,*], Alexander Leunig[1,2,#], Rainer Kaiser[1,2], Markus Joppich[3], Sophia Brambs[1], , Aleksandar Janjic[4], Oliver Popp[5], Daniel Nixdorf[6,7], Valeria Fumagalli[8, 9], Nora Schmidt[10], Vivien Polewka[1], Afra Anjum[1], Viktoria Knottenberg[1], Luke Eivers[1], Lucas E. Wange[4], Christoph Gold[1], Marieluise Kirchner[5], Maximilian Muenchhoff[11,12,13], Johannes C. Hellmuth[7,13], Clemens Scherer[1,2,13], Raquel Rubio-Acero[12,14], Tabea Eser[12,14], Flora Deák[12,14], Kerstin Puchinger[14], Niklas Kuhl[15,16], Andreas Linder[15,16], Kathrin Saar[5], Lukas Tomas[1,2], Christian Schulz[1,2], Andreas Wieser[12,14], Wolfgang Enard[4], Inge Kroidl[12,14], Christof Geldmacher[12,14], Michael von Bergwelt-Baildon[7,13], Oliver T. Keppler[11,12,13], Mathias Munschauer[10], Matteo Iannacone[8, 9, 17], Ralf Zimmer[3], Philipp Mertins[5], Norbert Hubner[5], Michael Hoelscher[12,14], Steffen Massberg[1,2], Konstantin Stark[1,2,§,*], Leo Nicolai[1,2,§,*]

Using Context-Sensitive Text Mining to Identify miRNAs in Different Stages of Atherosclerosis

Markus Joppich[1] Christian Weber[2] Ralf Zimmer[1]

[1] Department of Informatics, LFE Bioinformatics, Ludwig-Maximilians-Universität München, Munich, Germany
[2] Institute for Cardiovascular Prevention, Ludwig-Maximilians-Universität München, Munich, Germany

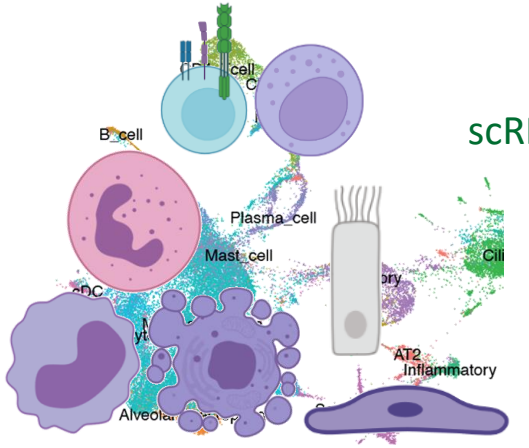
Address for correspondence Markus Joppich, MSc, LFE Bioinformatics, Department of Informatics, Ludwig-Maximilians-Universität München, Amalienstr. 17, Munich, Bavaria 80333, Germany (e-mail: joppich@bio.ifi.lmu.de).

Thromb Haemost 2019;119:1247–1264.

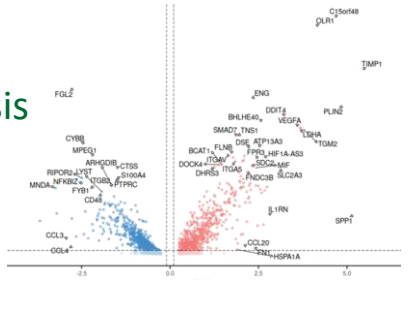
SEQU-INTO: Early detection of impurities, contamination and off-targets (ICOs) in long read/MinION sequencing

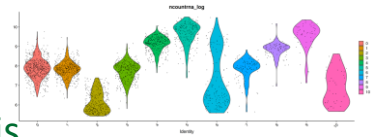
Markus Joppich[a,1], Margaryta Olenchuk[a,1], Julia M. Mayer[a,1], Quirin Emslander[b], Luisa F. Jimenez-Soto[c], Ralf Zimmer[a,*]

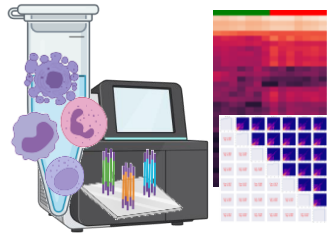
[a] LFE Bioinformatics, Department of Informatics, Ludwig-Maximilians-Universität München, 80333 München, Germany
[b] Physics of Synthetic Biological Systems, Physics Department, Technische Universität München, 85748 Garching, Germany
[c] Walther Straub Institute for Pharmacology and Toxicology, Ludwig-Maximilians-Universität München, Goethestrasse 33, 80336 München, Germany



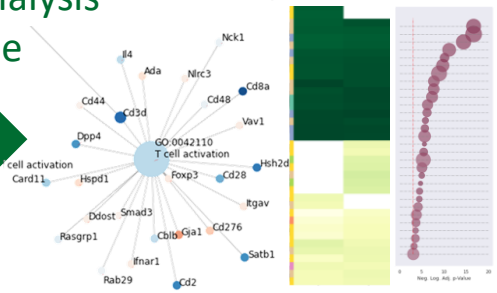
scRNA-seq analysis pipeline





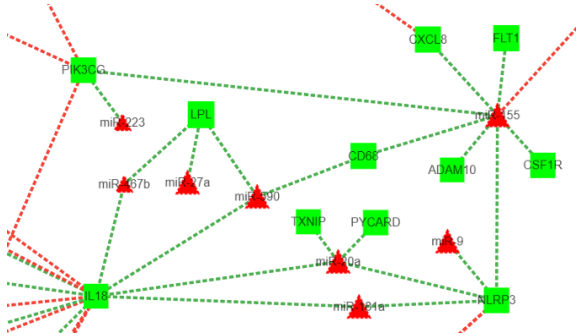


RNA-seq analysis pipeline



Cells & sequencing images: biorender.com

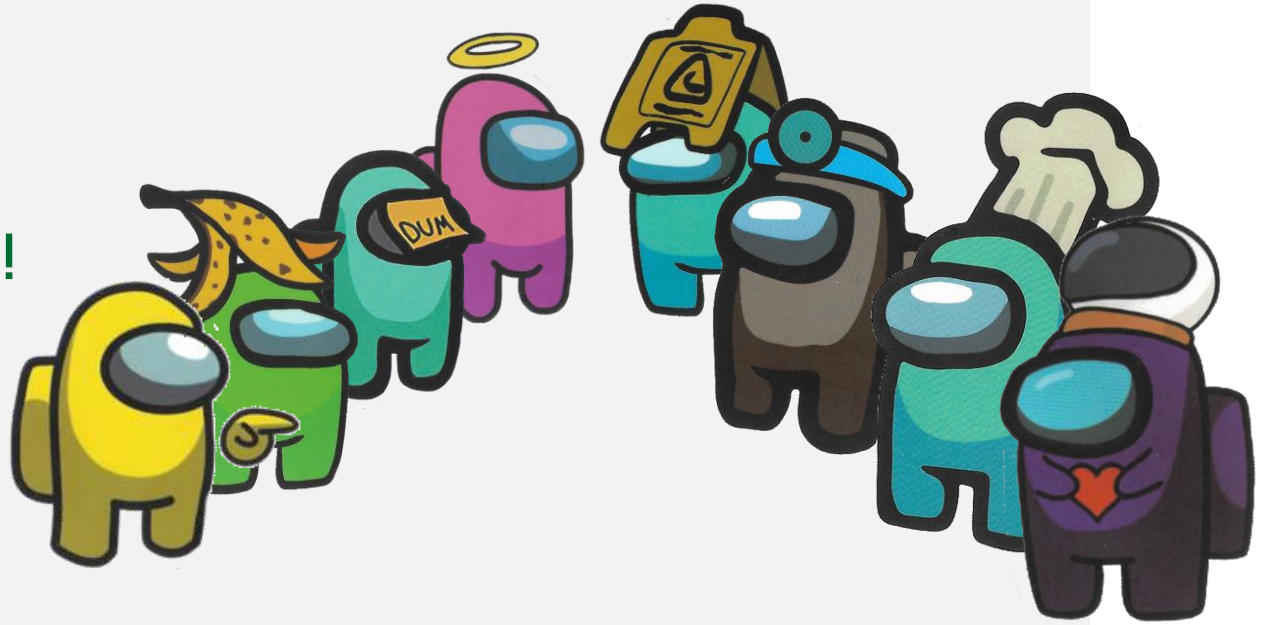
- miRNA ⊣ gene
- Context Information



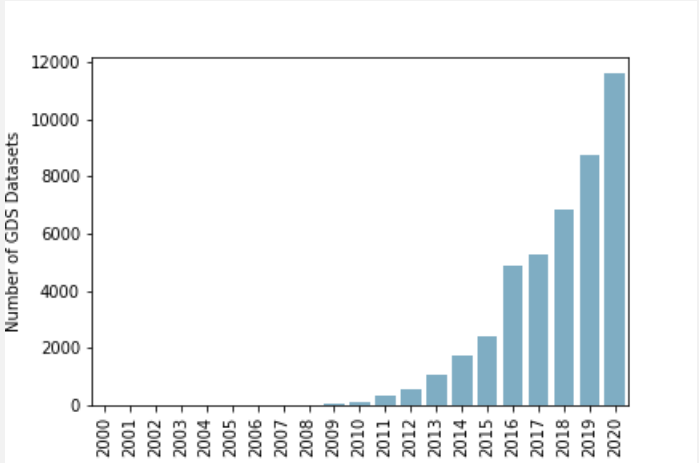
ABOUT YOU

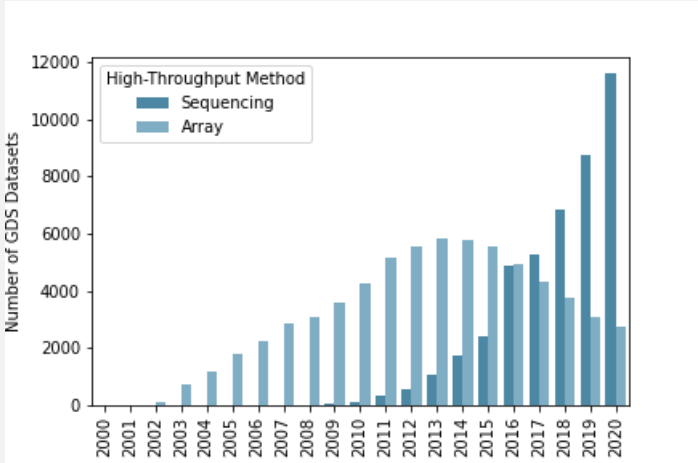
- Who are you?
- What are you studying?
- Why are you here?
- What is your experience?
- What are your expectations?

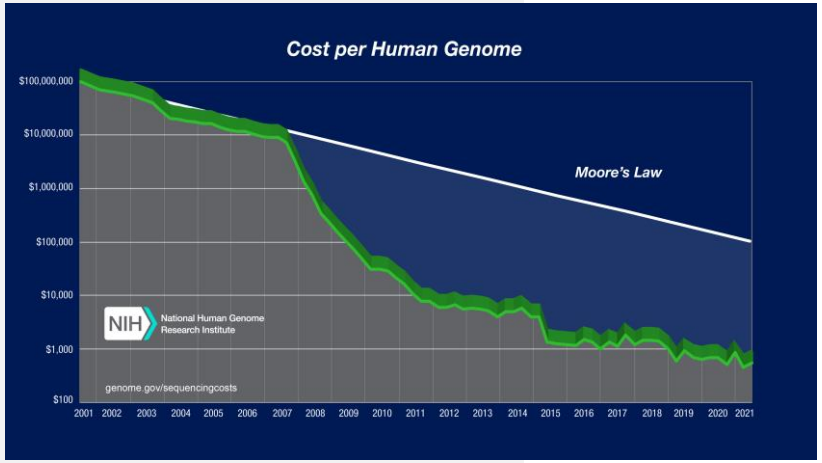
- Collect answers in your team!
  - We'll meet here in ~10min



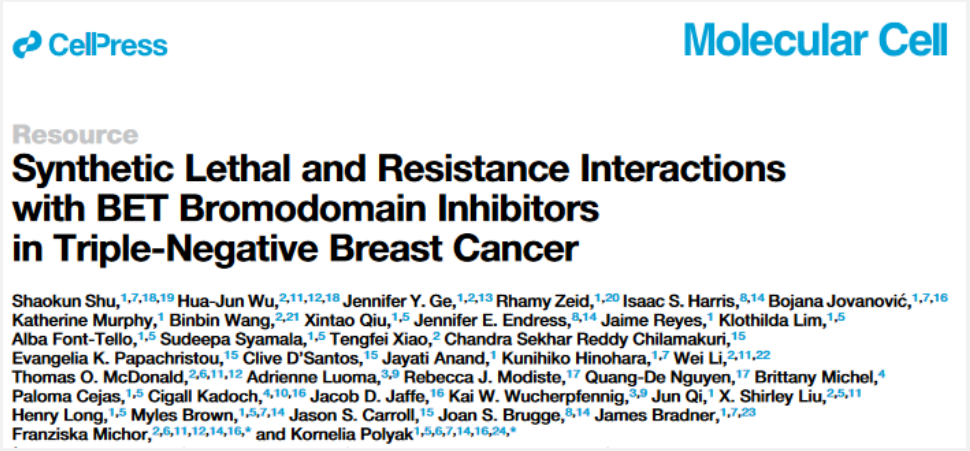
WORKSHOP ON HIGH-THROUGHPUT SEQUENCING







- Sequencing experiments are common and performed on a routinely basis
  - Costs are continuously decreasing
  - Sequencing data is more flexible than array experiments



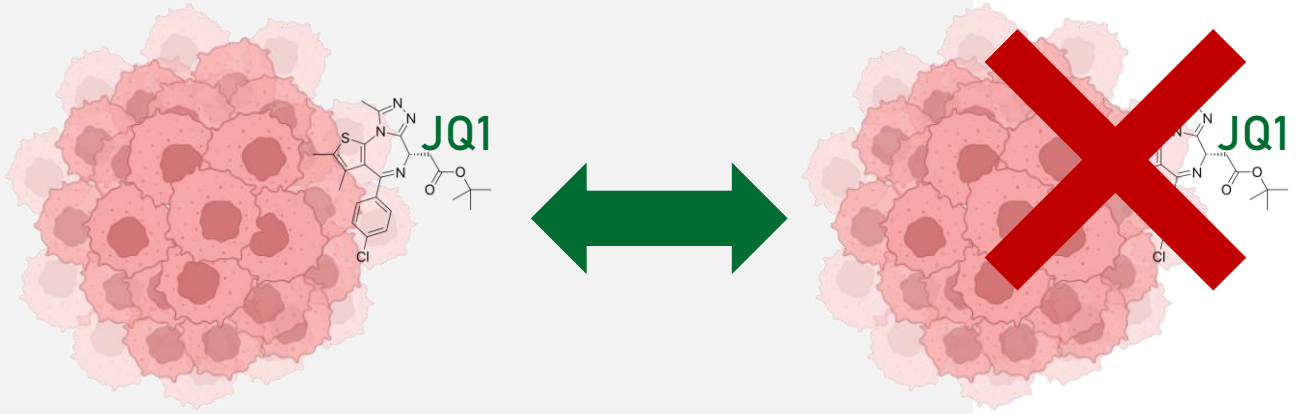
CellPress

Molecular Cell

Resource

Synthetic Lethal and Resistance Interactions with BET Bromodomain Inhibitors in Triple-Negative Breast Cancer

Shaokun Shu,[1,7,18,19] Hua-Jun Wu,[2,11,12,18] Jennifer Y. Ge,[1,2,13] Rhamy Zeid,[1,20] Isaac S. Harris,[8,14] Bojana Jovanović,[1,7,16] Katherine Murphy,[1] Binbin Wang,[2,21] Xintao Qiu,[1,5] Jennifer E. Endress,[8,14] Jaime Reyes,[1] Klothilda Lim,[1,5] Alba Font-Tello,[1,5] Sudeepa Syamala,[1,5] Tengfei Xiao,[2] Chandra Sekhar Reddy Chilamakuri,[15] Evangelia K. Papachristou,[15] Clive D'Santos,[15] Jayati Anand,[1] Kunihiko Hinohara,[1,7] Wei Li,[2,11,22] Thomas O. McDonald,[2,6,11,12] Adrienne Luoma,[3,9] Rebecca J. Modiste,[17] Quang-De Nguyen,[17] Brittany Michel,[4] Paloma Cejas,[1,5] Cigall Kadoch,[4,10,16] Jacob D. Jaffe,[16] Kai W. Wucherpfennig,[3,9] Jun Qi,[1] X. Shirley Liu,[2,5,11] Henry Long,[1,5] Myles Brown,[1,5,7,14] Jason S. Carroll,[15] Joan S. Brugge,[8,14] James Bradner,[1,7,23] Franziska Michor,[2,6,11,12,14,16,*] and Kornelia Polyak[1,5,6,7,14,16,24,*]



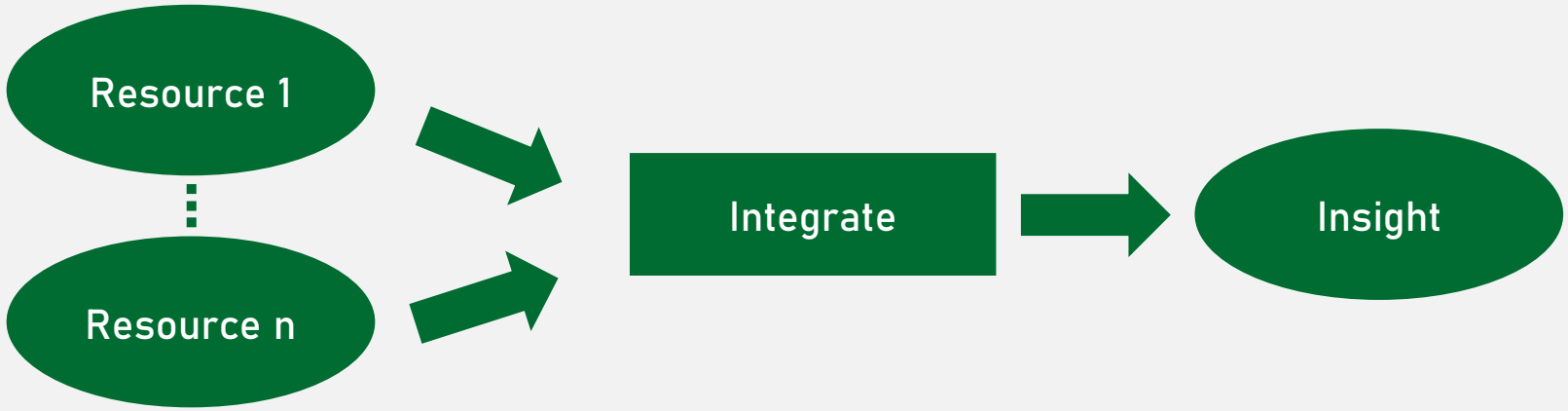
SUM159 SUM159R

- triple-negative breast cancer (TNBC)
  - Does not express estrogen receptor (ESR1), progesterone receptor (PGR), HER2/neu (ERBB2)
  - Cell line SUM159
    - metastatic human breast cancer cell line derived from a patient with anaplastic breast carcinoma
  - Control (SUM159) and BET bromodomain inhibitors (BBDIs) resistant (SUM159R)
    - JQ1 is inhibitor of BET family of bromodomain proteins

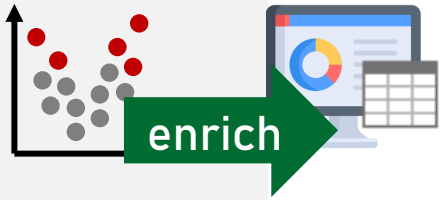
https://www.sciencedirect.com/science/article/pii/S1097276520302690

INTEGRATIVE BIOINFORMATICS

- Integrative Bioinformatics
  - Subdiscipline in Bioinformatics
    - Uses tools of computer science and electronic infrastructure applied to Bio*



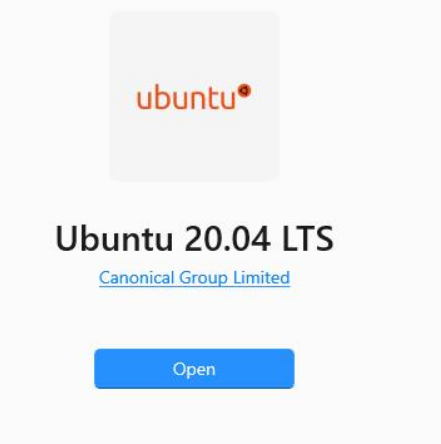
  - Uses existing tools/methods/resources to gain new insights

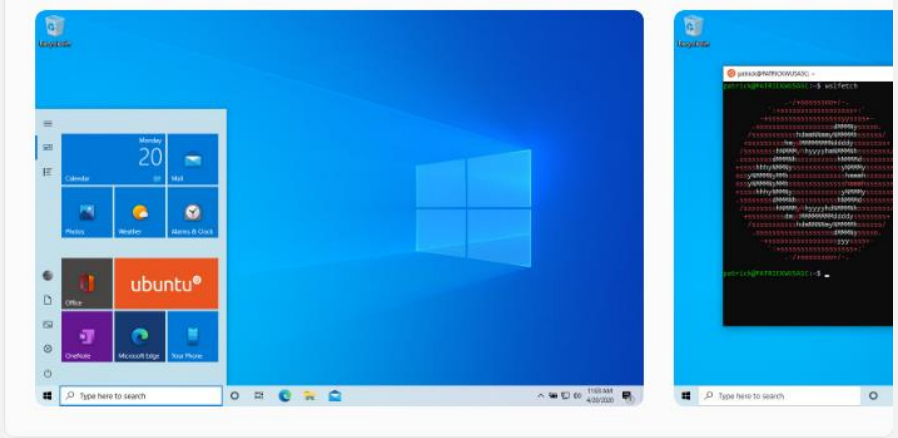


COMPUTER RESOURCES

- Computer

- Some kind of Linux
  - You use Linux? → you're fine (but need Win/Mac to run Loupe browser)
  - You use Windows? → use Windows Subsystem for Linux → you're fine





  - You use Mac OS? → you're mostly fine, but I can't help you
    - My Mac experience dates back to Mac OS X Leopard!

SOFTWARE RESOURCES

- You need some R version

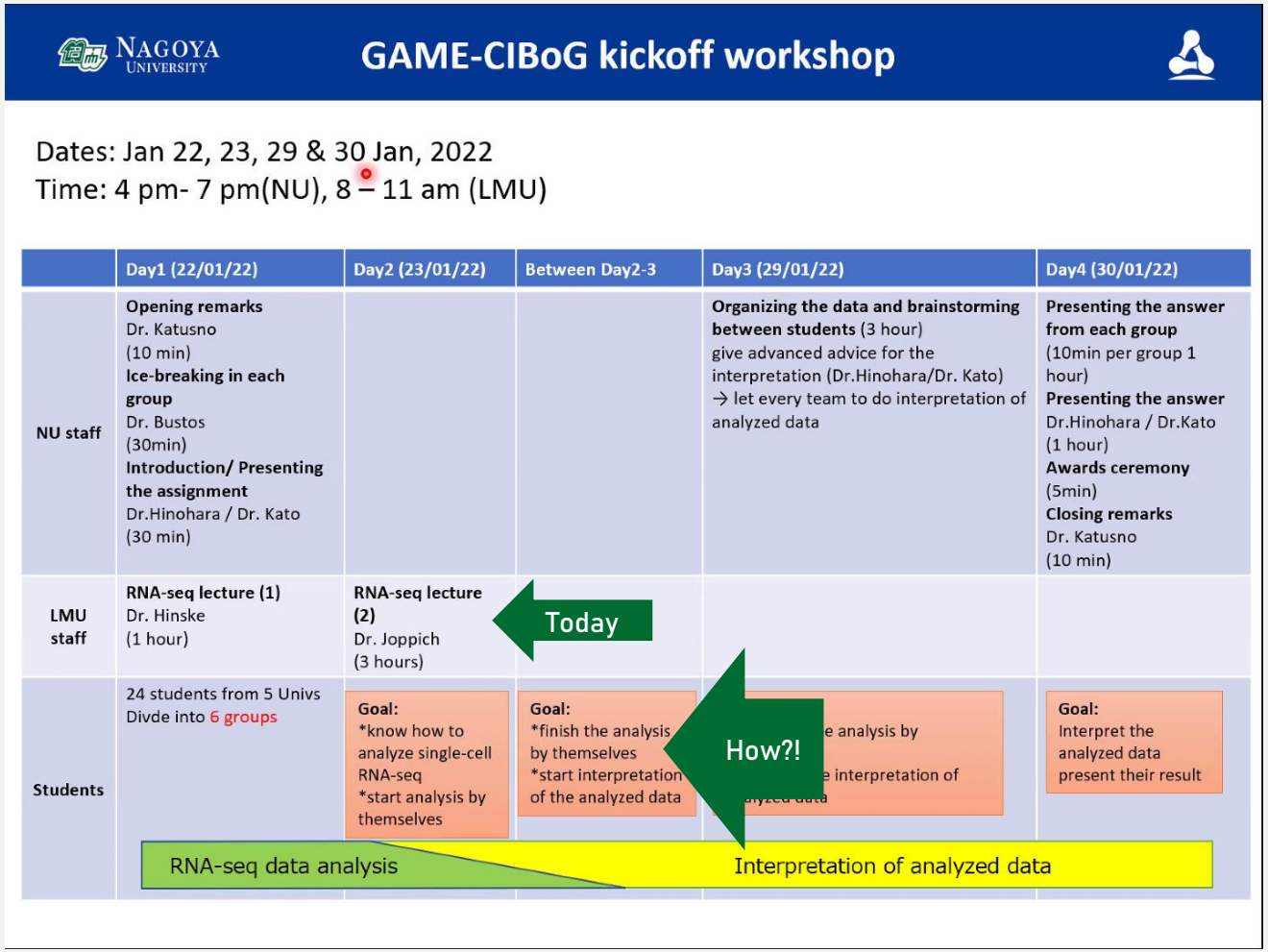
```
mjoppich@spectre3:/mnt/c/Users/mjopp$ R --version
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
```

- You may want to use jupyter+R kernel
  - `install.packages("devtools")`
  - `devtools::install_github("IRkernel/IRkernel")`
  - `IRkernel::installspec()`

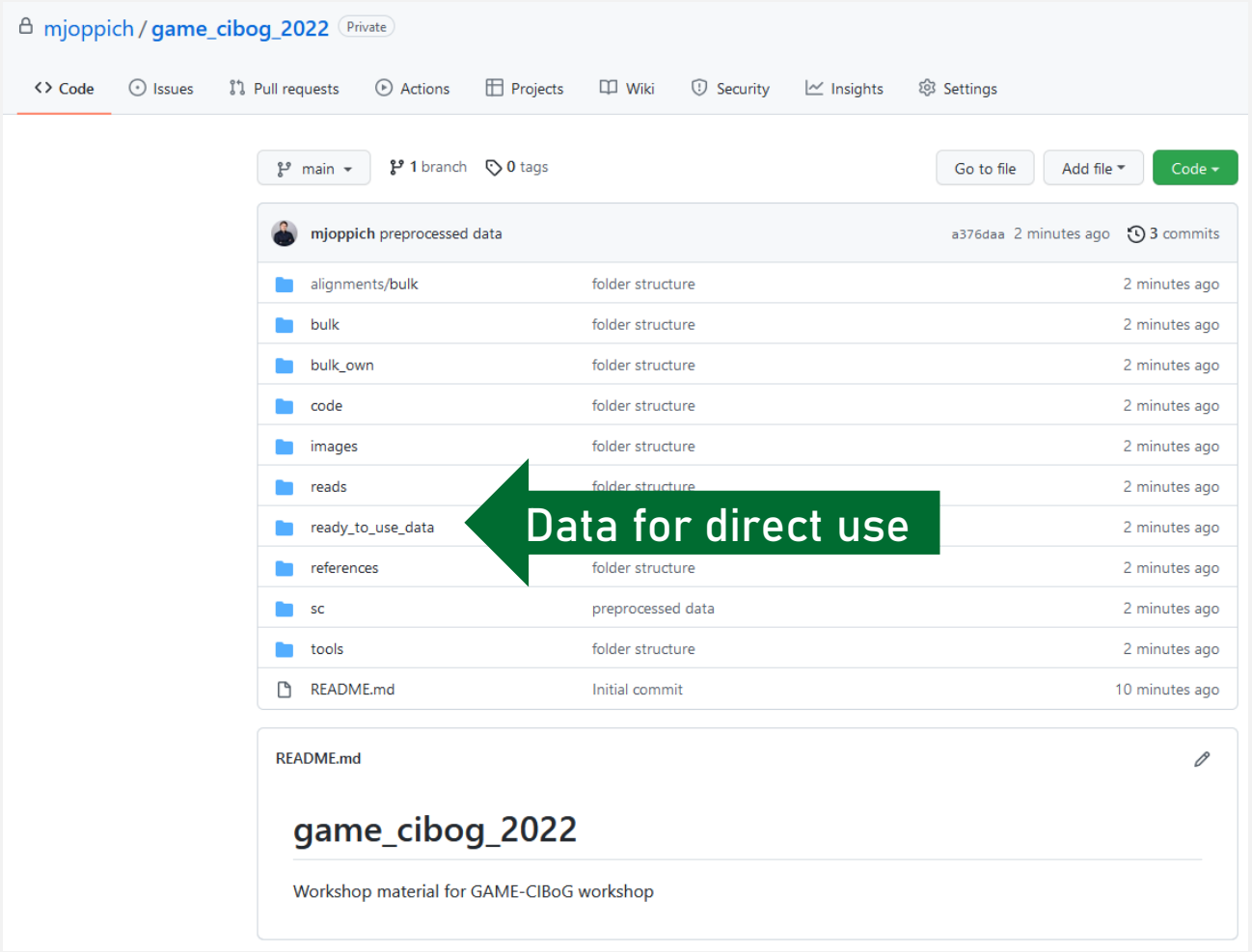
- Some kind of IDE or jupyter-lab in the browser
  - Big fan of Visual Code 
    - View PDF, PNG, JPEG
    - Supports jupyter notebooks and plots

```
options(jupyter.plot_mimetypes = 'image/png')
options(repr.plot.width = 1, repr.plot.height = 0.75, repr.plot.res = 100)
```

TIMELINE


NAGOYA UNIVERSITY
GAME-CIBoG kickoff workshop
Dates: Jan 22, 23, 29 & 30 Jan, 2022
Time: 4 pm- 7 pm(NU), 8 – 11 am (LMU)
Day1 (22/01/22)
Day2 (23/01/22)
Between Day2-3
Day3 (29/01/22)
Day4 (30/01/22)
NU staff
Opening remarks
Dr. Katusno
(10 min)
Ice-breaking in each group
Dr. Bustos
(30min)
Introduction/ Presenting the assignment
Dr.Hinohara / Dr. Kato
(30 min)
Organizing the data and brainstorming between students (3 hour)
give advanced advice for the interpretation (Dr.Hinohara/Dr. Kato)
→ let every team to do interpretation of analyzed data
Presenting the answer from each group
(10min per group 1 hour)
Presenting the answer
Dr.Hinohara / Dr.Kato
(1 hour)
Awards ceremony
(5min)
Closing remarks
Dr. Katusno
(10 min)
LMU staff
RNA-seq lecture (1)
Dr. Hinske
(1 hour)
RNA-seq lecture (2)
Dr. Joppich
(3 hours)
Today
Students
24 students from 5 Univs
Divde into 6 groups
Goal:
*know how to analyze single-cell RNA-seq
*start analysis by themselves
Goal:
*finish the analysis by themselves
*start interpretation of the analyzed data
How?!
Goal:
Interpret the analyzed data
present their result
RNA-seq data analysis
Interpretation of analyzed data

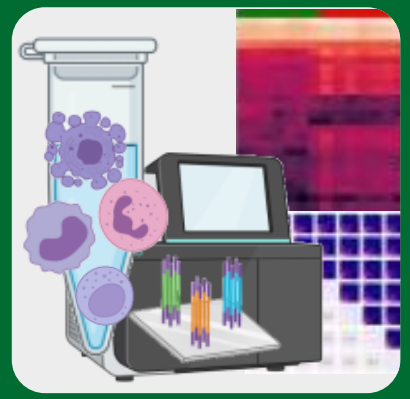
GIT REPOSITORY



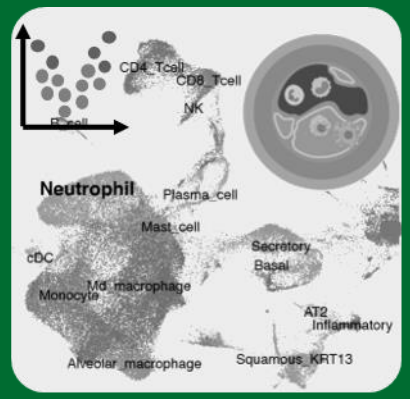
- git clone https://github.com/mjoppich/game_cibog_2022.git
- Folder structure useful (not required though!) for workshop
- Send eMail to

  joppich@bio.ifi.lmu.de

  with GitHub username to get access



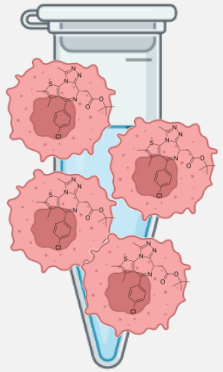
bulk RNA-seq



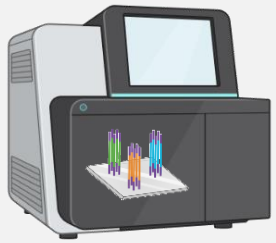
scRNA-seq

BULK RNA-SEQ

SUM159

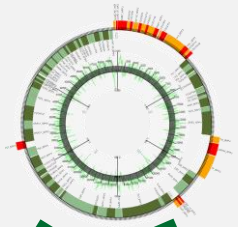


Material Isolation



Sequencing

Genome Annotation



Alignment

Samples

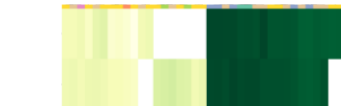
Genes

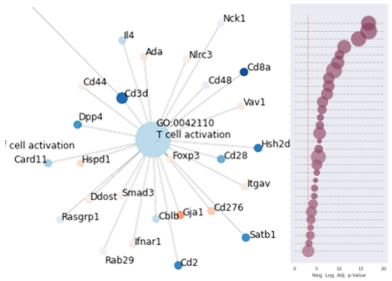
Quantification

Comparison

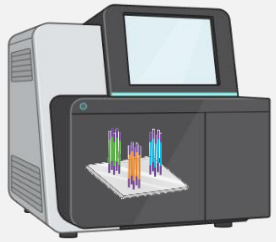
logFC + pVal

Genes





SUM159R



Samples

Genes

Gene Set Enrichment

Data Analysis

Quite some time

Matter of hours*

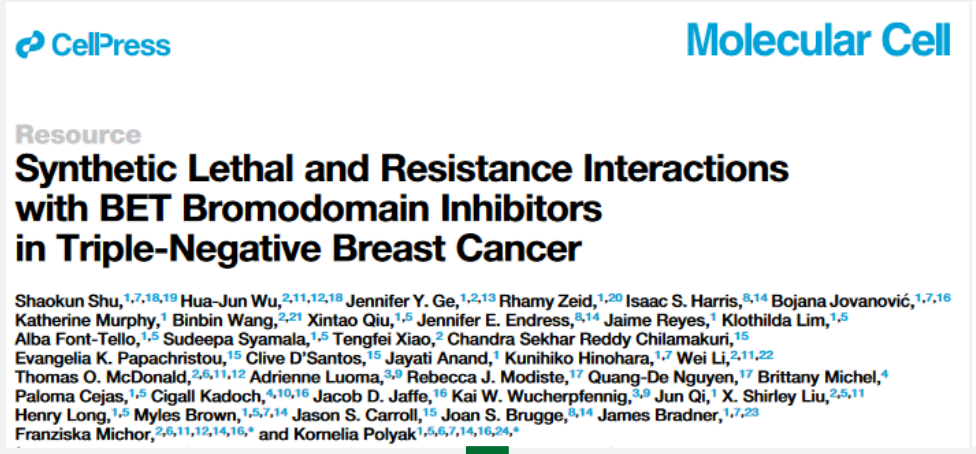
Matter of weeks

Cells & sequencing images: biorender.com

GETTING THE DATA

- Gene expression data are collected in large repositories
  - NCBI Gene Expression Omnibus (GEO)
  - EBI European Nucleotide Archive (ENA)
- For publication whole samples are collected
  - Raw sequencing data (shared with SRA, Sequencing Read Archive)
  - *Processed* expression data
- Here: use processed data
  - Saves time
  - Was probably done right anyhow

SEARCHING THE DATA



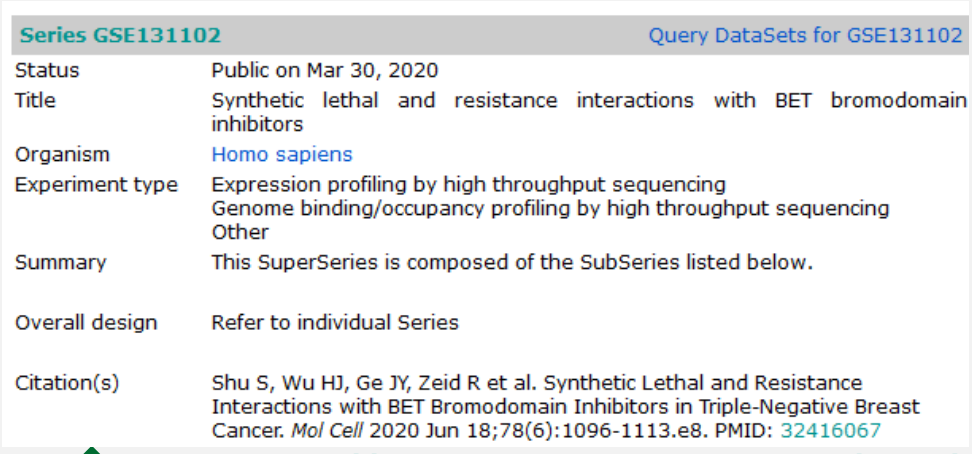
CellPress

Molecular Cell

Resource

Synthetic Lethal and Resistance Interactions with BET Bromodomain Inhibitors in Triple-Negative Breast Cancer

Shaokun Shu,[1,7,18,19] Hua-Jun Wu,[2,11,12,18] Jennifer Y. Ge,[1,2,13] Rhamy Zeid,[1,20] Isaac S. Harris,[8,14] Bojana Jovanović,[1,7,16] Katherine Murphy,[1] Binbin Wang,[2,21] Xintao Qiu,[1,5] Jennifer E. Endress,[8,14] Jaime Reyes,[1] Klothilda Lim,[1,5] Alba Font-Tello,[1,5] Sudeepa Syamala,[1,5] Tengfei Xiao,[2] Chandra Sekhar Reddy Chilamakuri,[15] Evangelia K. Papachristou,[15] Clive D'Santos,[15] Jayati Anand,[1] Kunihiko Hinohara,[1,7] Wei Li,[2,11,22] Thomas O. McDonald,[2,6,11,12] Adrienne Luoma,[3,9] Rebecca J. Modiste,[17] Quang-De Nguyen,[17] Brittany Michel,[4] Paloma Cejas,[1,5] Cigall Kadoch,[4,10,16] Jacob D. Jaffe,[16] Kai W. Wucherpfennig,[3,9] Jun Qi,[1] X. Shirley Liu,[2,5,11] Henry Long,[1,5] Myles Brown,[1,5,7,14] Jason S. Carroll,[15] Joan S. Brugge,[8,14] James Bradner,[1,7,23] Franziska Michor,[2,6,11,12,14,16,*] and Kornelia Polyak[1,5,6,7,14,16,24,*]



Series GSE131102	Query DataSets for GSE131102
Status	Public on Mar 30, 2020
Title	Synthetic lethal and resistance interactions with BET bromodomain inhibitors
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing Genome binding/occupancy profiling by high throughput sequencing Other
Summary	This SuperSeries is composed of the SubSeries listed below.
Overall design	Refer to individual Series
Citation(s)	Shu S, Wu HJ, Ge JY, Zeid R et al. Synthetic Lethal and Resistance Interactions with BET Bromodomain Inhibitors in Triple-Negative Breast Cancer. *Mol Cell* 2020 Jun 18;78(6):1096-1113.e8. PMID: 32416067

https://www.ncbi.nlm.nih.gov/geo/

Deposited Data		
All raw genomic data	GEO	GSE131102
All raw numeric data and image files	Mendeley	https://doi.org/10.17632/p4ypdxmsk5.1

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131102

GETTING THE RIGHT DATA

- Interested in RNA-seq data
  SUM159 cells vs SUM159-JQ1R cells

This SuperSeries is composed of the following SubSeries:
Less...

GSE131026	Synthetic lethal and resistance interactions with BET bromodomain inhibitors [ATAC-Seq]
GSE131030	Synthetic lethal and resistance interactions with BET bromodomain inhibitors [Barcode_seq]
GSE131091	Synthetic lethal and resistance interactions with BET bromodomain inhibitors [CRISPR screen]
GSE131097	Synthetic lethal and resistance interactions with BET bromodomain inhibitors [ChIP-Seq]
GSE131099	Synthetic lethal and resistance interactions with BET bromodomain inhibitors [RNA-Seq]
GSE131135	Synthetic lethal and resistance interactions with BET bromodomain inhibitors [scRNA-Seq]

Samples (62)
Less...

GSM3763447	SUM149_DMSO_12h_Replicate1
GSM3763448	SUM149_DMSO_12h_Replicate2
GSM3763449	SUM149_DMSO_12h_Replicate3
GSM3763450	SUM149_JQ1_10uM_12h_R1
GSM3763451	SUM149_JQ1_10uM_12h_R2
GSM3763452	SUM149_JQ1_10uM_12h_R3
GSM3763453	SUM149_JQ1_2uM_12h_R1
GSM3763454	SUM149_JQ1_2uM_12h_R2
GSM3763455	SUM149_JQ1_2uM_12h_R3
GSM3763456	SUM149R_DMSO_12h_Replicate1
GSM3763457	SUM149R_DMSO_12h_Replicate2
GSM3763458	SUM149R_DMSO_12h_Replicate3
GSM3763459	SUM149R_JQ1_10uM_12h_R1
GSM3763460	SUM149R_JQ1_10uM_12h_R2
GSM3763461	SUM149R_JQ1_10uM_12h_R3
GSM3763462	SUM149R_JQ1_2uM_12h_R1
GSM3763463	SUM149R_JQ1_2uM_12h_R2
GSM3763464	SUM149R_JQ1_2uM_12h_R3
GSM3763465	SUM159_DMSO_12h_R1
GSM3763466	SUM159_DMSO_12h_R2
GSM3763467	SUM159_DMSO_24h_R1
GSM3763468	SUM159_DMSO_24h_R2
GSM3763469	SUM159_DMSO_3h_R1
GSM3763470	SUM159_DMSO_3h_R2
GSM3763471	SUM159_JQ1_12h_R1
GSM3763472	SUM159_JQ1_12h_R2
GSM3763473	SUM159_JQ1_24h_R1
GSM3763474	SUM159_JQ1_24h_R2
GSM3763475	SUM159_JQ1_3h_R1
GSM3763476	SUM159_JQ1_3h_R2
GSM3763477	SUM159_JQ1R_DMSO_24h_R1
GSM3763478	SUM159_JQ1R_DMSO_24h_R2
GSM3763479	SUM159_JQ1R_DMSO_3h_R1
GSM3763480	SUM159_JQ1R_DMSO_3h_R2

SUM159 cells at 3h and 24h

SUM159-JQ1R cells at 3h and 24h

GETTING QUANTIFICATION DATA

Submission date	May 13, 2019
Last update date	Apr 01, 2020
Contact name	Kornelia Polyak
E-mail(s)	kornelia_polyak@dfci.harvard.edu
Phone	617-632-2106
Organization name	Dana-Farber Cancer Institute
Department	Medical Oncology
Lab	Polyak
Street address	450 Brookline Ave
City	Boston
State/province	MA
ZIP/Postal code	02215
Country	USA

Platform ID	GPL11154	
Series (2)	GSE131099	Synthetic lethal and resistance bromodomain inhibitors [RNA-S
	GSE131102	Synthetic lethal and resistance bromodomain inhibitors

Relations

BioSample	SAMN11634886
SRA	SRX5824615

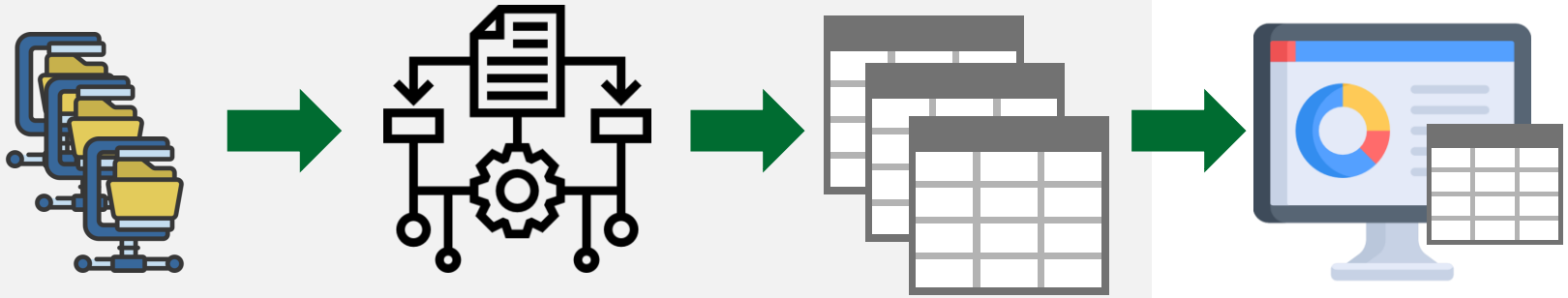
Supplementary file	Size
GSM3763467_SUM159_DMSO_2_R1.counts.txt.gz	94.6 Kb

SRA Run Selector

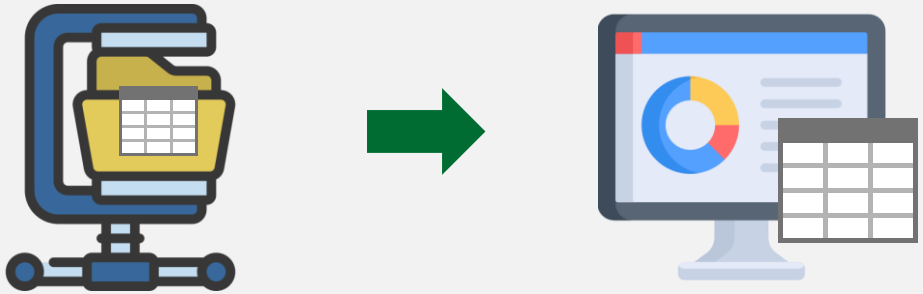
Raw Data

Processed

- Raw Data: you are the boss
  - Much more data, more steps, but most flexible (e.g. new methods)



- Processed Data: much easier (download-read-go)



Pictograms: flaticon.com

BULK PROCESSING PIPELINE

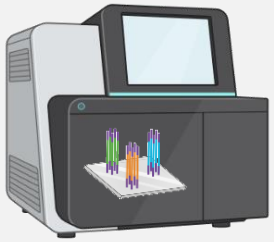
All the steps needed for bulk data processing

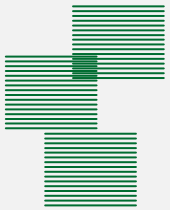
Sample

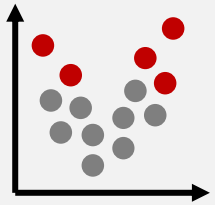
Reads

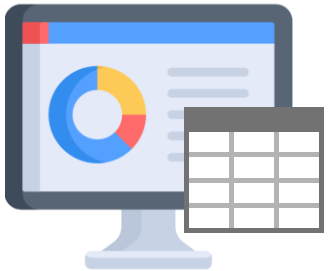
Aligned Reads/ Alignments

Quantification/ Read counts/ Count Matrix

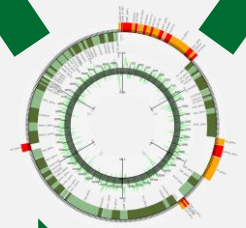








bcl2fastq → hisat2 → featureCounts → DESeq2 → DOSE



Reference Genome

STAR → HTseq

bwa

limma → clusterProfiler

edgeR → Metascape

EmpiRe → DAVID

NOISeq → enrichmentBrowser

Sequencing image: biorender.com, Analysis: flaticon.com

REQUIRED TOOLS FOR DOWNLOADING DATA

Getting the read data

- sratoolkit for downloading files
  - wget --output-document tools/sratoolkit.tar.gz http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz
  - cd tools && tar -vxzf sratoolkit.tar.gz
  - vdb-config –interactive

  - Make path to sratoolkit visible:
  - Add

    export PATH=$PATH:/mnt/t/workshop/tools/sratoolkit.2.11.2-ubuntu64/bin

    to ~/.bashrc and `source ~/.bashrc`

- Then we can call *fastq-dump* as

  `fastq-dump --gzip SRR9048093`

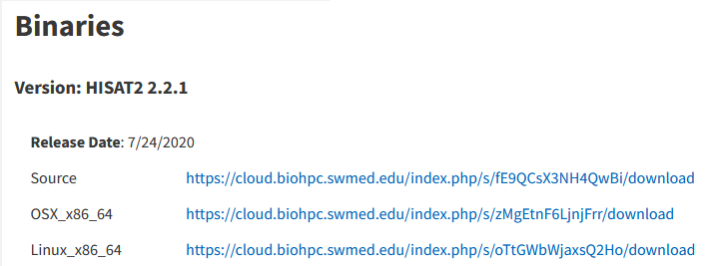
  for downloading specific SRA-file

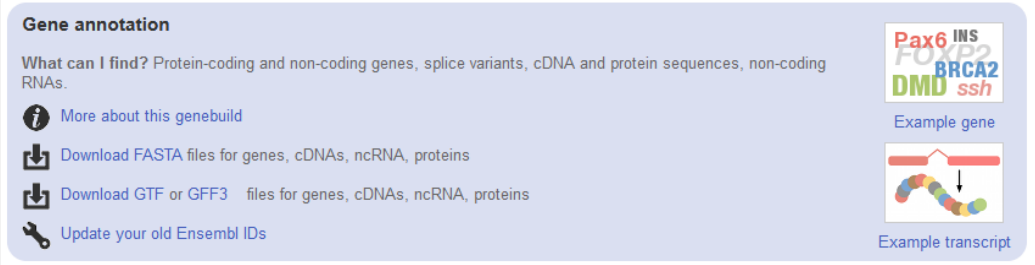
REQUIRED FILES AND REFERENCES

Doing the alignment

- Hisat2
  - `sudo apt install hisat2`
  - http://daehwankimlab.github.io/hisat2/
  - Index/Reference genome: H. sapiens, GRCh38, https://genome-idx.s3.amazonaws.com/hisat/grch38_genome.tar.gz
  - `tar xfz grch38_genome.tar.gz`
- featureCounts / subread
  - `sudo apt install subread`
  - https://sourceforge.net/projects/subread/files/subread-2.0.3/

  - Genome Annotation File:
    - http://ftp.ensembl.org/pub/release-105/gtf/homo_sapiens/Homo_sapiens.GRCh38.105.gtf.gz
    - `gunzip Homo_sapiens.GRCh38.105.gtf.gz`





USING THE TOOLS

How to call the tools?

- General: check websites of tools, `<tool> --help`

```
mjoppich@spectre3:/mnt/w/game_cibog$ hisat2 --help
HISAT2 version 2.1.0 by Daehwan Kim (infphilo@gmail.com, www.ccb.jhu.edu/people/infphilo)
Usage:
  hisat2 [options]* -x <ht2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <ht2-idx>  Index filename prefix (minus trailing .X.ht2).
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.
```

DOWNLOADING BULK RAW DATA

```
#24h R1
! cd reads/bulk/ && fastq-dump --gzip SRR9048093
#24h R1
! cd reads/bulk/ && fastq-dump --gzip SRR9048094
#JQ1 24h R1
! cd reads/bulk/ && fastq-dump --gzip SRR9048103
#JQ1 24h R1
! cd reads/bulk/ && fastq-dump --gzip SRR9048104
```

```
Read 17872256 spots for SRR9048093
Written 17872256 spots for SRR9048093
Read 40740514 spots for SRR9048094
Written 40740514 spots for SRR9048094
Read 43554379 spots for SRR9048103
Written 43554379 spots for SRR9048103
Read 28277603 spots for SRR9048104
Written 28277603 spots for SRR9048104
```

```
! mv reads/bulk/SRR9048093.fastq.gz reads/bulk/SUM159_24h_R1.fastq.gz
! mv reads/bulk/SRR9048094.fastq.gz reads/bulk/SUM159_24h_R2.fastq.gz
! mv reads/bulk/SRR9048103.fastq.gz reads/bulk/SUM159R_24h_R1.fastq.gz
! mv reads/bulk/SRR9048104.fastq.gz reads/bulk/SUM159R_24h_R2.fastq.gz
```

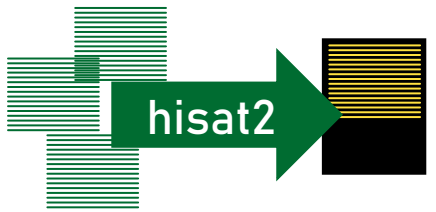
- With the tools available just download all files
  - Rename for better understandability!
  - Downloading took ~1h

ALIGNING READS

Must be done for all samples!

```
! hisat2 -p 4 -x ./references/grch38/genome -U ./reads/bulk/SUM159_24h_R2.fastq.gz |
                    samtools view -@ 4 -bS - |
                    samtools sort -@ 4 -o alignments/bulk/SUM159_24h_R2.bam -

40740514 reads; of these:
  40740514 (100.00%) were unpaired; of these:
    1926278 (4.73%) aligned 0 times
    33800194 (82.96%) aligned exactly 1 time
    5014042 (12.31%) aligned >1 times
95.27% overall alignment rate
[bam_sort_core] merging from 12 files and 4 in-memory blocks...
```

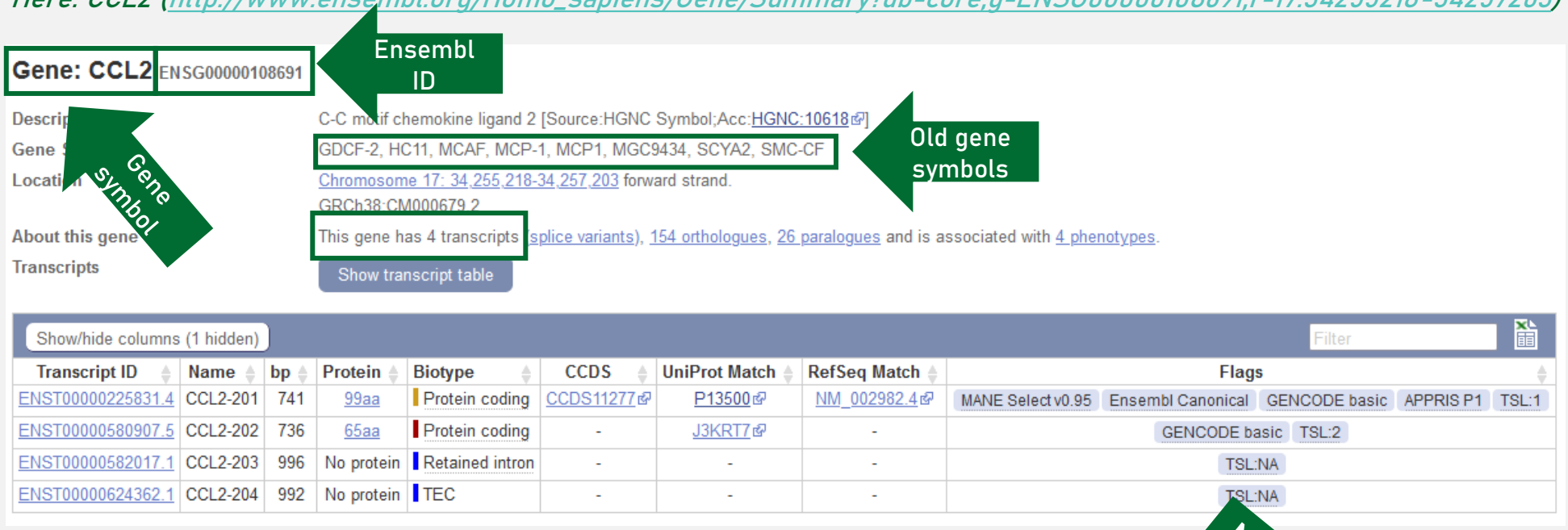
Alignment performancce

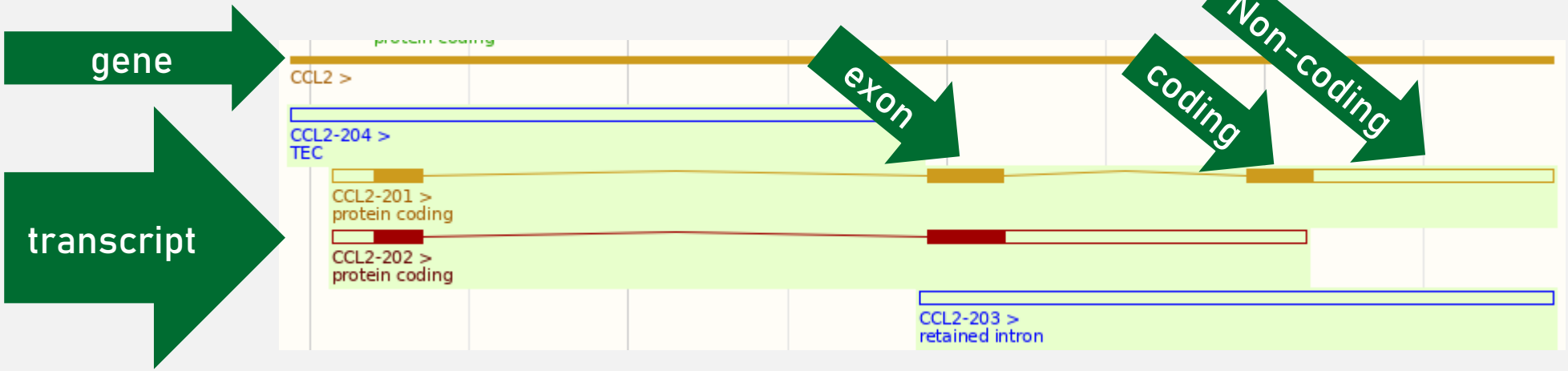


- *hisat2* expects reference file and reads, output: sam format
- *samtools view* converts sam format into compressed bam format
  - Primarily done to save disk space!
- *samtools sort* sorts the reads by genome positions

REMINDER: GENE STRUCTURE

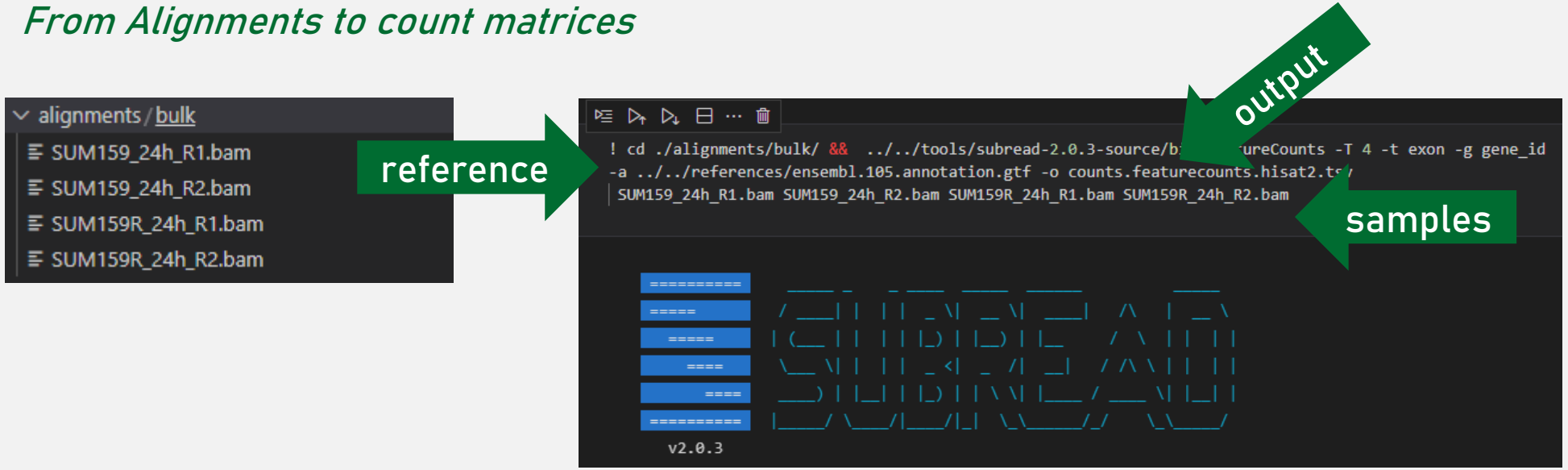
Here: CCL2 (http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000108691;r=17:34255218-34257203)

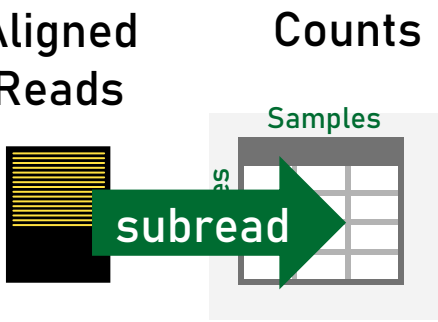




QUANTIFYING GENE COUNTS

From Alignments to count matrices





- Run featureCounts with alignments files and genome annotation
  - Faster with more threads: -T 4
  - Quantify by exons: -t exon
  - Summarize by gene_id: -g gene_id

00_get_data.ipynb | 01_bulk_align_count.ipynb | counts.featurecounts.hisat2.tsv.summary | 02_bulk_preprocessed.ipynb | 02_

alignments > bulk > counts.featurecounts.hisat2.tsv.summary

	Status	SUM159_24h_R1.bam	SUM159_24h_R2.bam	SUM159R_24h_R1.bam	SUM159R_24h_R2.bam
2	Assigned	12730408	29461647	31964377	20473798
3	Unassigned_Unmapped	1101887	1926278	1944297	1520084
4	Unassigned_Read_Type	0	0	0	0
5	Unassigned_Singleton	0	0	0	0
6	Unassigned_MappingQuality	0	0	0	0
7	Unassigned_Chimera	0	0	0	0
8	Unassigned_FragmentLength	0	0	0	0
9	Unassigned_Duplicate	0	0	0	0
10	Unassigned_MultiMapping	6488761	15144678	15243105	10011071
11	Unassigned_Secondary	0	0	0	0
12	Unassigned_NonSplit	0	0	0	0
13	Unassigned_NoFeatures	765478	1736538	1548873	1132298
14	Unassigned_Overlapping_Length	0	0	0	0
15	Unassigned_Ambiguity	1102693	2602009	2799429	1726337

COUNT MATRIX

Final Output

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Geneid	Chr	Start	End	Strand	Leng	SUM159_24h_R1.bam	SUM159_24h_R2.bam	SUM159R_24h_R1.bam	SUM159R_24h_R2.bam	SUM159_SUM	SUM159R_SUM	ALL_SUM
2	ENSG00000198804	MT	5904	7445	+	1542	141955	327668	468323	317305	469623	785628	1255251
3	ENSG00000198886	MT	10760	12137	+	1378	138689	320222	438344	267733	458911	706077	1164988
4	ENSG00000210082	MT	1671	3229	+	1559	76078	178317	247196	142901	254395	390097	644492
5	ENSG00000198712	MT	7586	8269	+	684	64070	149254	241954	153261	213324	395215	608539
6	ENSG00000198938	MT	9207	9990	+	784	62879	149328	305177	187365	212207	492542	704749
7	ENSG00000198727	MT	14747	15887	+	1141	59983	132070	206999	144340	192053	351339	543392
8	ENSG00000074800	1;1;1;1;1;1	8861000;8	8861429;8	-;-;-;-;-;-;	5216	60100	126070	98074	68928	186170	167002	353172
9	ENSG00000111640	12;12;12;1	6534512;6	6534569;6	+;+;+;+;+;+	2396	53008	117725	94460	64355	170733	158815	329548
10	ENSG00000156508	6;6;6;6;6;6	73489308;	73492804;	-;-;-;-;-;-;	10063	52679	115930	89297	63060	168609	152357	320966
11	ENSG00000167658	19;19;19;1	3976056;3	3976747;3	-;-;-;-;-;-;	4021	44637	114492	119551	61234	159129	180785	339914
12	ENSG00000163359	2;2;2;2;2;2	23732400:	23732481	-;-;-;-;-;-;	21516	50449	108069	69778	52632	158518	122410	280928

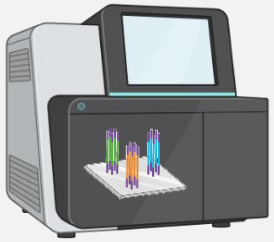
Total Sums			
SUM159_24h_R1.bam	SUM159_24h_R2.bam	SUM159R_24h_R1.bam	SUM159R_24h_R2.bam
12730442	29461460	31964210	20473685

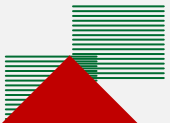
- Large table (>60000 rows!)
- Sums in counts differ between samples and replicates!
  - Absolute values are not comparable!
- Ensembl Geneids are … unique, but not interpretable!

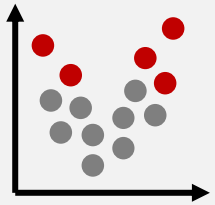
AVOIDING THE HASSLE OF DOING ALIGNMENT+COUNTING

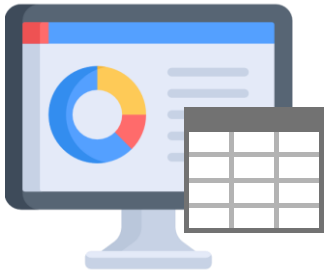
Use preprocessed data

Sample | Reads | Aligned Reads/ Alignments | Quantification/ Read counts/ Count Matrix









bcl2fastq

STAR

bwa

featureCounts

Reference genome

DESeq2

limma

edgeR

EmpiRe

NOISeq

DOSE

clusterProfiler

Metascape

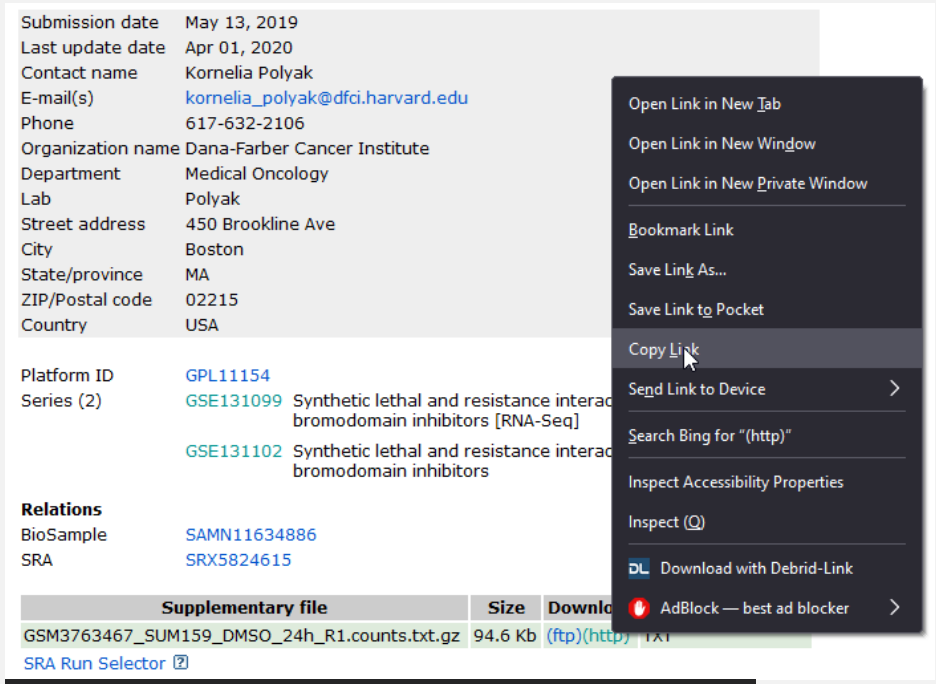
DAVID

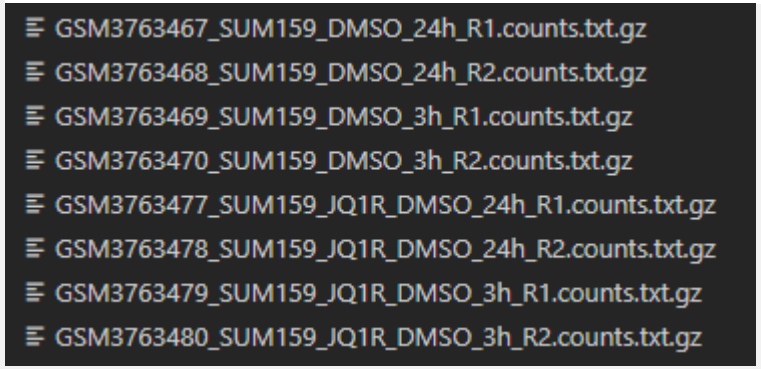
enrichmentBrowser

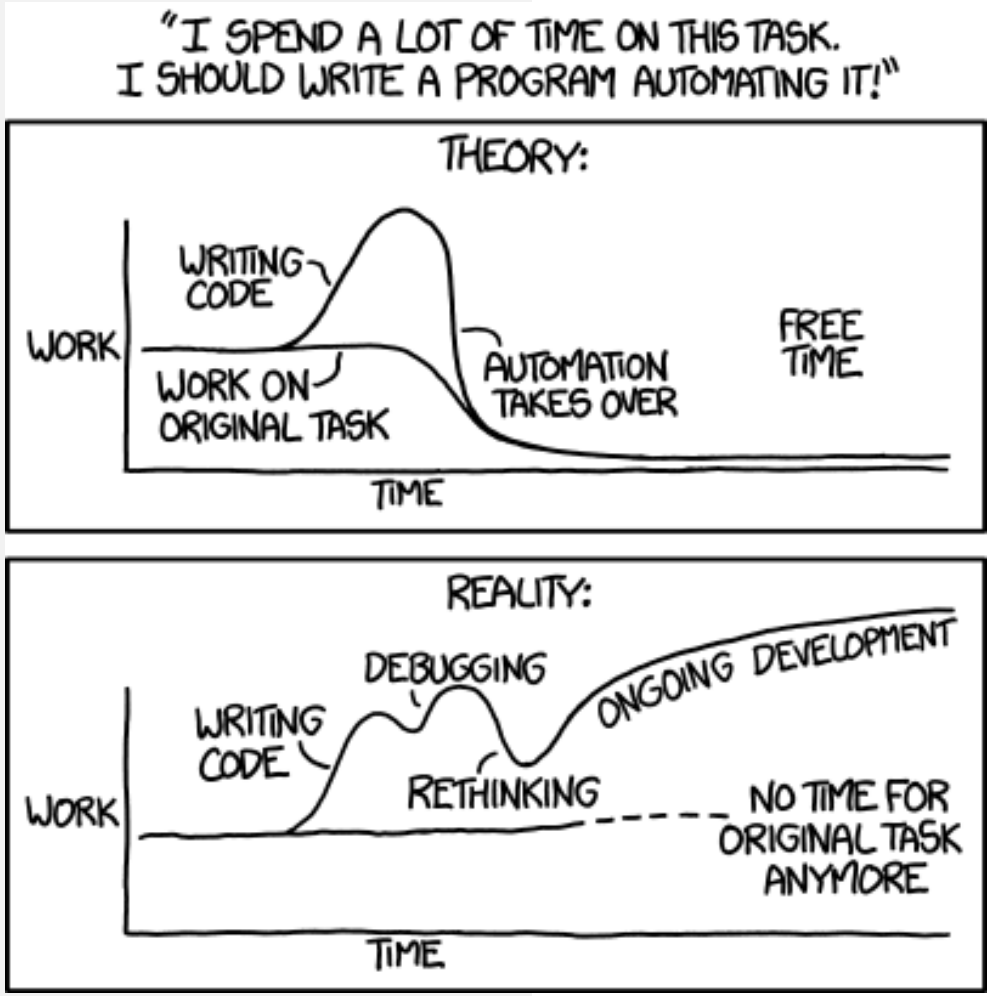
Sequencing image: biorender.com, Analysis: flaticon.com

DOWNLOADING THE PROCESSED DATA

- Interested in 8 samples
  - Manually get the links, download with command-line
  - `wget –O <filename> <URL>`







Xkcd comic: xkcd.com

THE FIRST DE ANALYIS

Reading in the data

- Let's have a peak at the downloaded files: zless

```
mjoppich@spectre3:/mnt/w/game_cibog$ zcat bulk/GSM3763467_SUM159_DMSO_24h_R1.counts.txt.gz | head
A1BG    44
A1BG-AS1        22
A1CF
A2M
A2M-AS1 3
A2ML1   11
A2MP1   0
A3GALT2 0
A4GALT  286
A4GNT   0
```

Gene symbol

Quantification

- Required libraries
  - I/O
  - DE analysis
  - Plotting

```
library(readr)
library('DESeq2')
library(ggplot2)
library(cowplot)
```

THE FIRST DE ANALYIS

Getting all samples loaded

- Read single samples

Let's start with reading in the 24h samples

```
SUM159_24h_r1 = read_tsv("../bulk/GSM3763468_SUM159_DMSO_24h_R1.counts.txt.gz", col_names=F)
SUM159_24h_r2 = read_tsv("../bulk/GSM3763468_SUM159_DMSO_24h_R2.counts.txt.gz", col_names=F)

SUM159jq1r_24h_r1 = read_tsv("../bulk/GSM3763477_SUM159_JQ1R_DMSO_24h_R1.counts.txt.gz", col_names=F)
SUM159jq1r_24h_r2 = read_tsv("../bulk/GSM3763477_SUM159_JQ1R_DMSO_24h_R2.counts.txt.gz", col_names=F)
```

Rows: 23841 Columns: 2

- Merge all samples into one data frame
  - Name the columns

```
colnames(SUM159_24h_r1) = c("gene", "SUM159_24h_r1")
colnames(SUM159_24h_r2) = c("gene", "SUM159_24h_r2")
colnames(SUM159jq1r_24h_r1) = c("gene", "SUM159jq1r_24h_r1")
colnames(SUM159jq1r_24h_r2) = c("gene", "SUM159jq1r_24h_r2")
```

  - Merge

```
sum159_24h.expr_df <- merge(SUM159_24h_r1, SUM159_24h_r2, by = 'gene')
sum159jq1r_24h.expr_df <- merge(SUM159jq1r_24h_r1, SUM159jq1r_24h_r2, by = 'gene')
```

```
merge_24h.expr_df <- merge(sum159_24h.expr_df, sum159jq1r_24h.expr_df, by = 'gene')
```

THE FIRST DE ANALYIS

Getting all samples loaded

- Regularly inspect your objects

```
head(merge_24h.expr_df)
```

A data.frame: 6 × 5

	gene	SUM159_24h_r1	SUM159_24h_r2	SUM159jq1r_24h_r1	SUM159jq1r_24h_r2
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	__alignment_not_unique	1946526	4392993	5571795	3575085
2	__ambiguous	385372	884778	815090	514010
3	__no_feature	2015667	4587061	5254348	3574216
4	__not_aligned	0	0	0	0
5	__too_low_aQual	0	0	0	0
6	A1BG	44	214	171	63

- Remove unwanted “genes”

```
expr_df_24h = merge_24h.expr_df[grep("^__", merge_24h.expr_df$gene, invert=T),]
rownames(expr_df_24h) = expr_df_24h$gene
expr_df_24h$gene = NULL
expr_df_24h
```

A data.frame: 23836 × 4

	SUM159_24h_r1	SUM159_24h_r2	SUM159jq1r_24h_r1	SUM159jq1r_24h_r2
	<dbl>	<dbl>	<dbl>	<dbl>
A1BG	44	214	171	63
A1BG-AS1	22	45	69	58
A1CF	2	4	1	0
A2M	4	28	0	0

DESEQ2

Brief introduction

Love *et al. Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8

Genome Biology

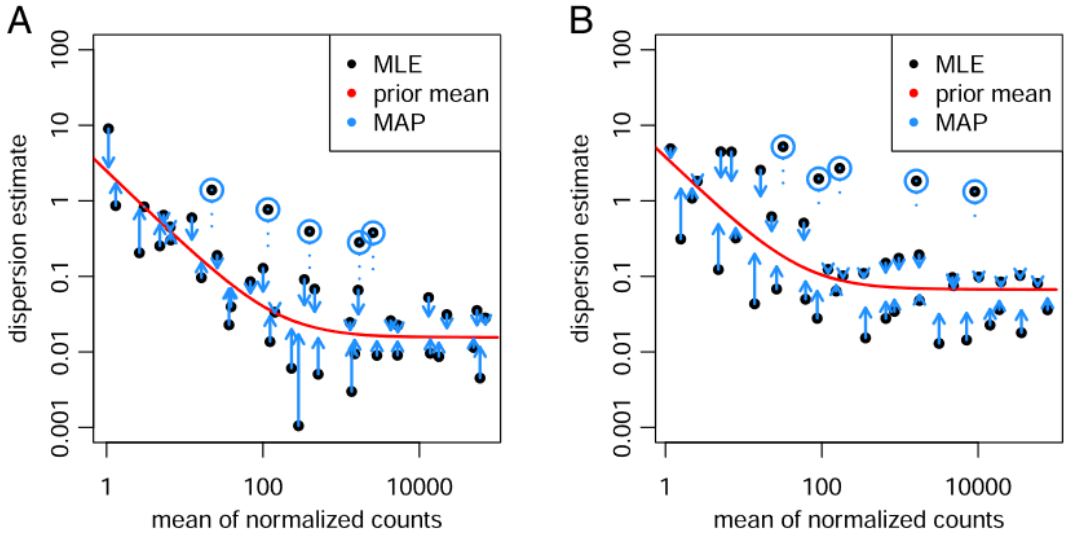
METHOD Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

Abstract

In comparative high-throughput sequencing assays, a fundamental task is the analysis of count data, such as read counts per gene in RNA-seq, for evidence of systematic changes across experimental conditions. Small replicate numbers, discreteness, large dynamic range and the presence of outliers require a suitable statistical approach. We present *DESeq2*, a method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. This enables a more quantitative analysis focused on the strength rather than the mere presence of differential expression. The *DESeq2* package is available at http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html.



If you want to see different examples:
https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

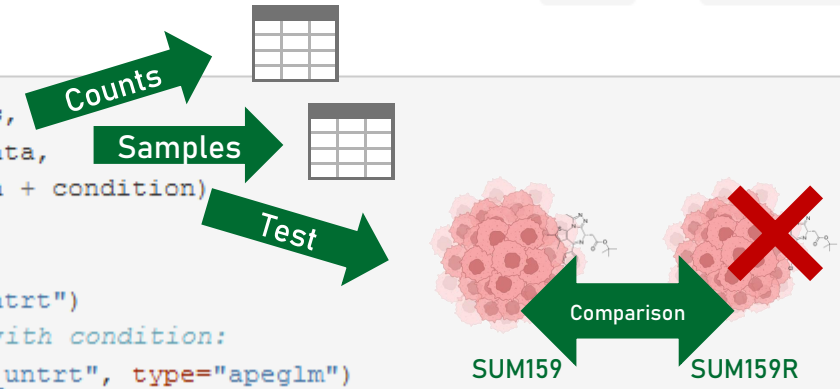
DESEQ2

Quick Start

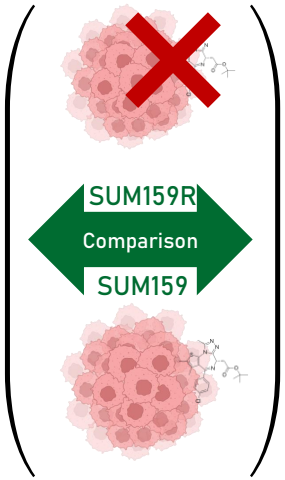
Quick start

Here we show the most basic steps for a differential expression analysis. There are a variety of steps upstream of DESeq2 that result in the generation of counts or estimated counts for each sample, which we will discuss in the sections below. This code chunk assumes that you have a count matrix called `cts` and a table of sample information called `coldata`. The `design` indicates how to model the samples, here, that we want to measure the effect of the condition, controlling for batch differences. The two factor variables `batch` and `condition` should be columns of `coldata`.

```
dds <- DESeqDataSetFromMatrix(countData = cts,
                              colData = coldata,
                              design= ~ batch + condition)
dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
res <- results(dds, name="condition_trt_vs_untrt")
# or to shrink log fold changes association with condition:
res <- lfcShrink(dds, coef="condition_trt_vs_untrt", type="apeglm")
```



$$\text{logFC} = \log_2$$



https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Cells: biorender.com

DESEQ2 PREPARATIONS

- Preparing sample sheet

```
pdat_24h = data.frame( id=c("SUM159_24h_R1","SUM159_24h_R2", "SUM159R_24h_R1","SUM159R_24h_R2"),
                       group=c("SUM159_24h", "SUM159_24h", "SUM159R_24h", "SUM159R_24h") )
pdat_24h
```

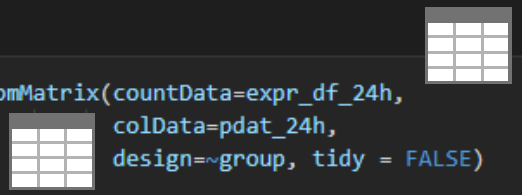
A data.frame: 4 × 2

id	group
<chr>	<chr>
SUM159_24h_R1	SUM159_24h
SUM159_24h_R2	SUM159_24h
SUM159R_24h_R1	SUM159R_24h
SUM159R_24h_R2	SUM159R_24h

- Assigns each sample/replicate a group
  - We will then compare two groups!

DESEQ2

DESeq2 analysis for 24h



```
dds_24h <- DESeqDataSetFromMatrix(countData=expr_df_24h,
                                  colData=pdat_24h,
                                  design=~group, tidy = FALSE)
```

```
converting counts to integer mode

Warning message in DESeqDataSet(se, design = design, ignoreRank):
“some variables in design formula are characters, converting to factors”
```

- Create DESeqDataSet (DDS)
  - Count data/quantifications
  - Sample sheet
  - Design pattern (groups from sample sheet)
- Call DESeq
  - Perform shrinkage*
  - Order result by p-value (BH corrected)

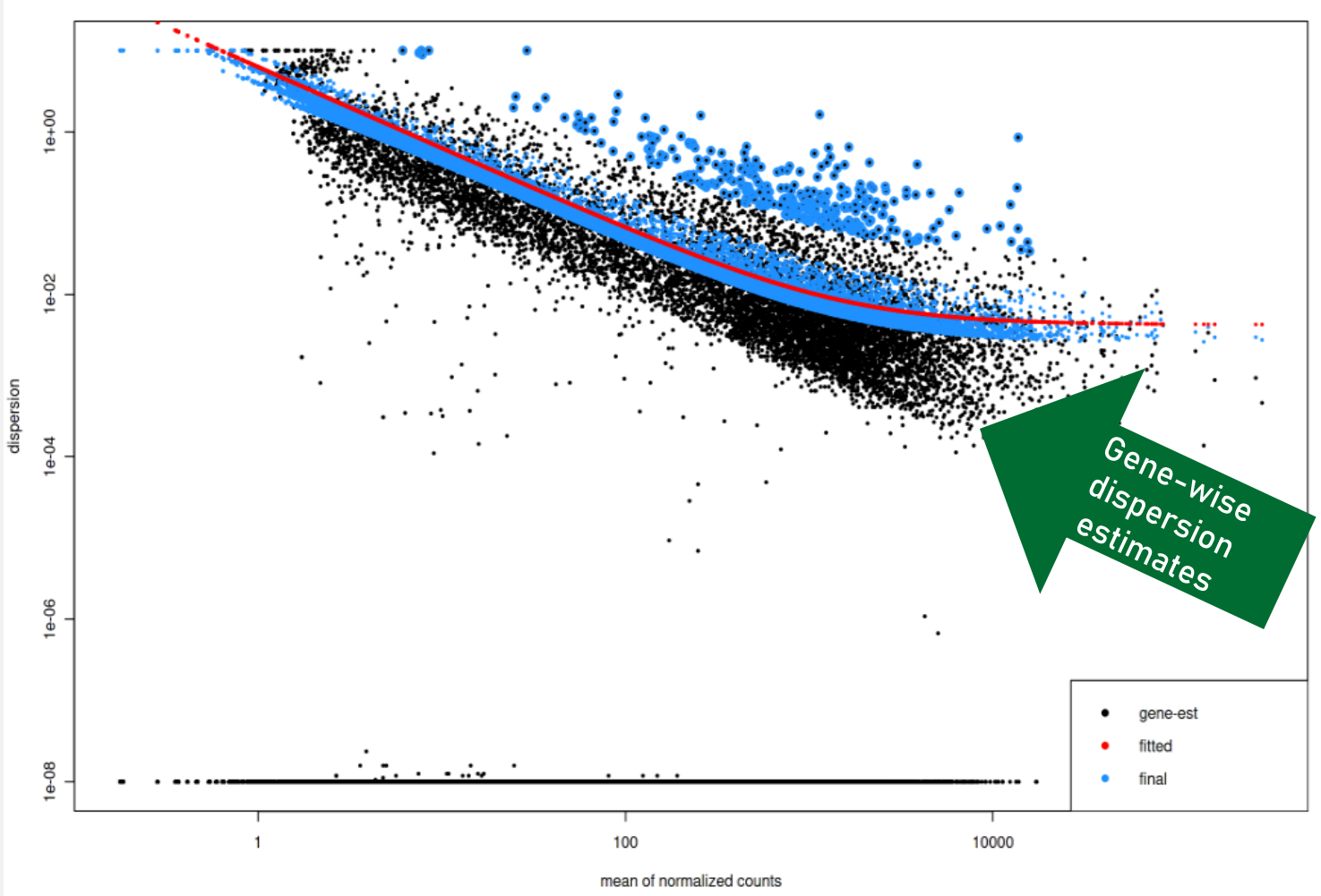
```
dds_24h <- DESeq(dds_24h)
res_lfc_24h <- lfcShrink(dds_24h, coef=resultsNames(dds_24h)[[2]], type="apeglm")
res_lfc_24h <- res_lfc_24h[order(res_lfc_24h$padj, decreasing=FALSE),]


res_24h <- results(dds_24h)
res_24h <- res_24h[order(res_24h$padj, decreasing=FALSE),]
```

```
estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing

using 'apeglm' for LFC shrinkage. If used in published research, please cite:
    Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
    sequence count data: removing the noise and preserving large differences.
    Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

DESEQ2

Interpreting the run

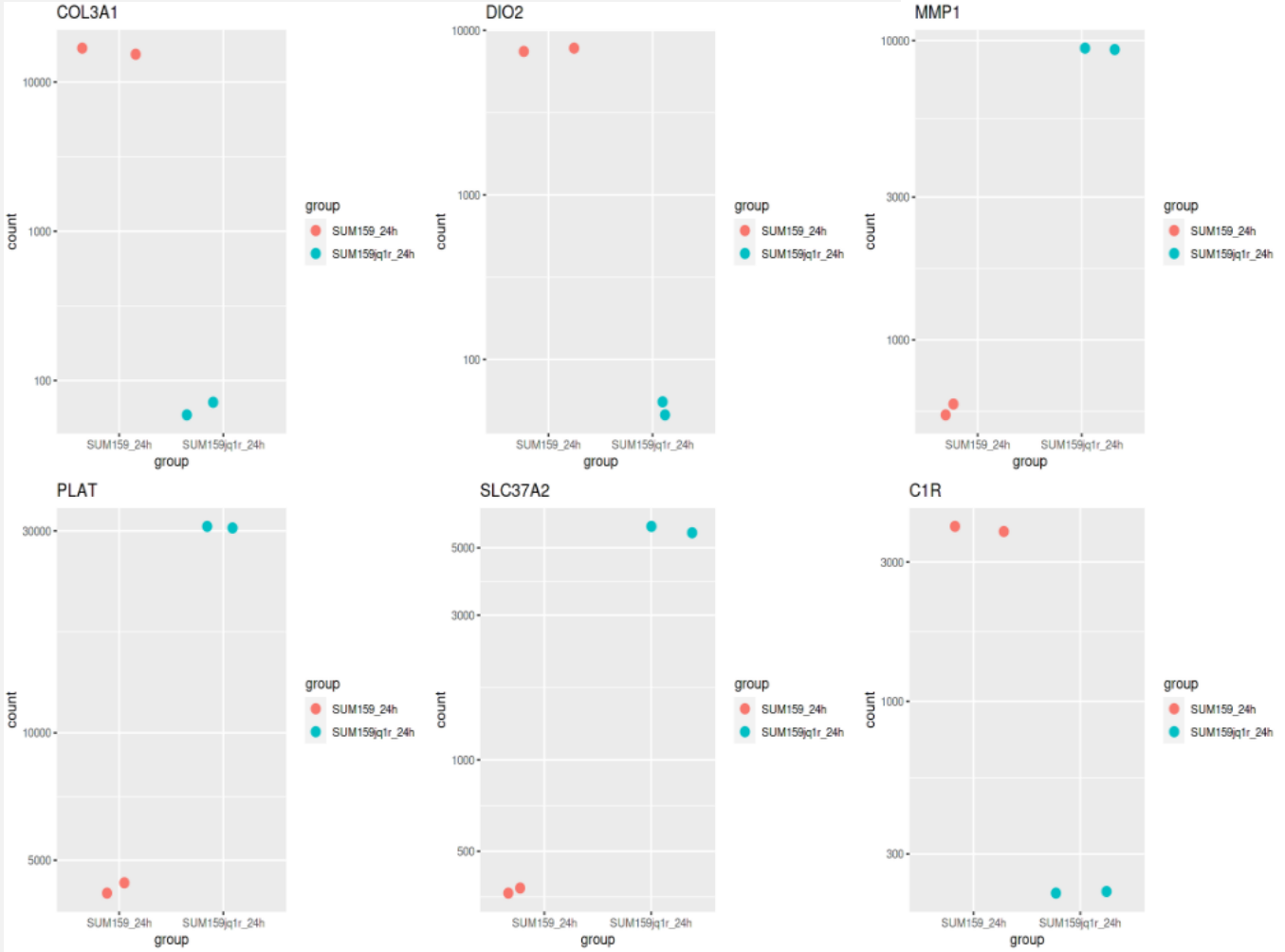
- Dispersion is a measure of spread or variability in the data
  - Here: related to the mean and variance of the data
- Plots expected dispersion value for genes of a given expression strength
- Expectation: generally scattering around the curve
  - dispersion decreasing with increasing mean expression levels



DESEQ2

Interpreting the run

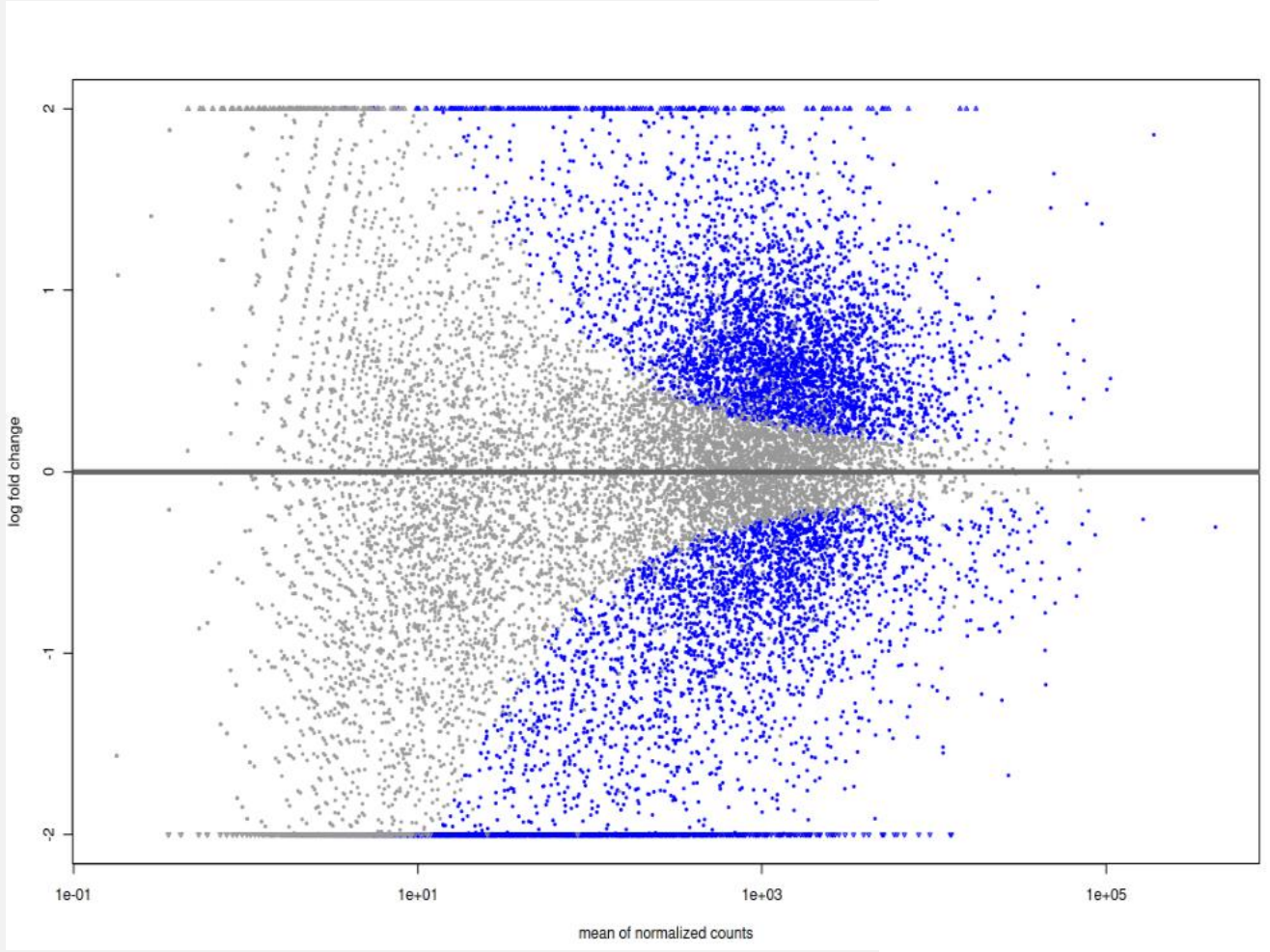
- Take a look at the most differentially expressed genes
  - Normalized! Counts per replicate
  - COL3A1 not highly expressed in SUM159R
    - But in SUM159



DESEQ2

Interpreting the run

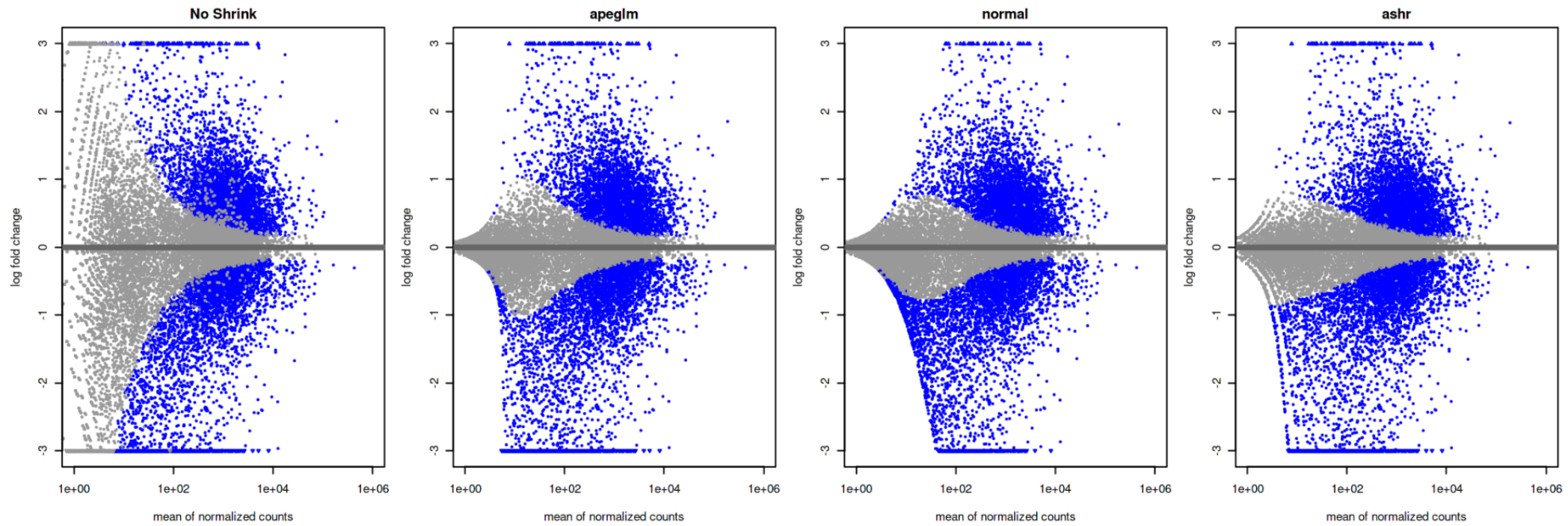
- MA plot
  - Log Fold Changes by mean of expression (in normalized counts)
  - Notice: high fold chances particularly for low counts
    - 12 / 3 => logFC = 2
    - +/- 2 read accuracy
    - 14/1 => logFC = 3.81



DESEQ2

Interpreting the run

- Different shrinkage methods
  - No influence on statistical evaluation!
  - Shrinkage removes the noise associated with log2FCs from low count genes without requiring arbitrary filtering thresholds



ANNOTATING RESULTS

Make it more readable!

```
log2 fold change (MAP): group SUM159R 24h vs SUM159 24h
Wald test p-value: group SUM159R 24h vs SUM159 24h
DataFrame with 6 rows and 5 columns
                 baseMean log2FoldChange     lfcSE       pvalue
                <numeric>      <numeric> <numeric>    <numeric>
ENSG00000168542   7623.80       -8.01821 0.1590666  0.00000e+00
ENSG00000254166   7798.00       -5.74974 0.1165984  0.00000e+00
ENSG00000196611   4798.88        4.00955 0.1011819  0.00000e+00
ENSG00000211448   3935.12       -7.20920 0.1654849  0.00000e+00
ENSG00000104368  17205.64        2.84687 0.0806986 1.03324e-274
ENSG00000134955   2884.91        4.37507 0.1265195 9.71242e-263
```

```
gencodeAnnot = as.data.frame(read_tsv("../references/gencode.grch38.human.35.gtfout.list"))
colnames(gencodeAnnot) = c("gene_stable_id", "gene_name", "biotype")
ncodeAnnot$gene_stable_id = gsub("\\.[0-9]+$", "", gencodeAnnot$gene_stable_id)
ad(gencodeAnnot)
```

A data.frame: 6 × 3

	ne_stable_id	gene_name	biotype
	<chr>	<chr>	<chr>
	ENSG00000000003	TSPAN6	protein_coding
2	ENSG00000000005	TNMD	protein_coding
3	ENSG00000000419	DPM1	protein_coding
4	ENSG00000000457	SCYL3	protein_coding
5	ENSG00000000460	C1orf112	protein_coding
6	ENSG00000000938	FGR	protein_coding

	Row.names	baseMean	log2FoldChange	lfcSE	pvalue	padj	gene_name	biotype
	<l<chr>>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
10002	ENSG00000168542	7623.801	-8.018207	0.15906657	0.000000e+00	0.000000e+00	COL3A1	protein_coding
12941	ENSG00000196611	4798.877	4.009554	0.10118186	0.000000e+00	0.000000e+00	MMP1	protein_coding
13923	ENSG00000211448	3935.122	-7.209199	0.16548490	0.000000e+00	0.000000e+00	DIO2	protein_coding
18508	ENSG00000254166	7797.995	-5.749736	0.11659844	0.000000e+00	0.000000e+00	CASC19	lncRNA
2473	ENSG00000104368	17205.636	2.846869	0.08069864	1.033238e-274	4.615475e-271	PLAT	protein_coding
5773	ENSG00000134955	2884.914	4.375074	0.12651954	9.712416e-263	3.615447e-259	SLC37A2	protein_coding

DESEQ2

Creating outputs

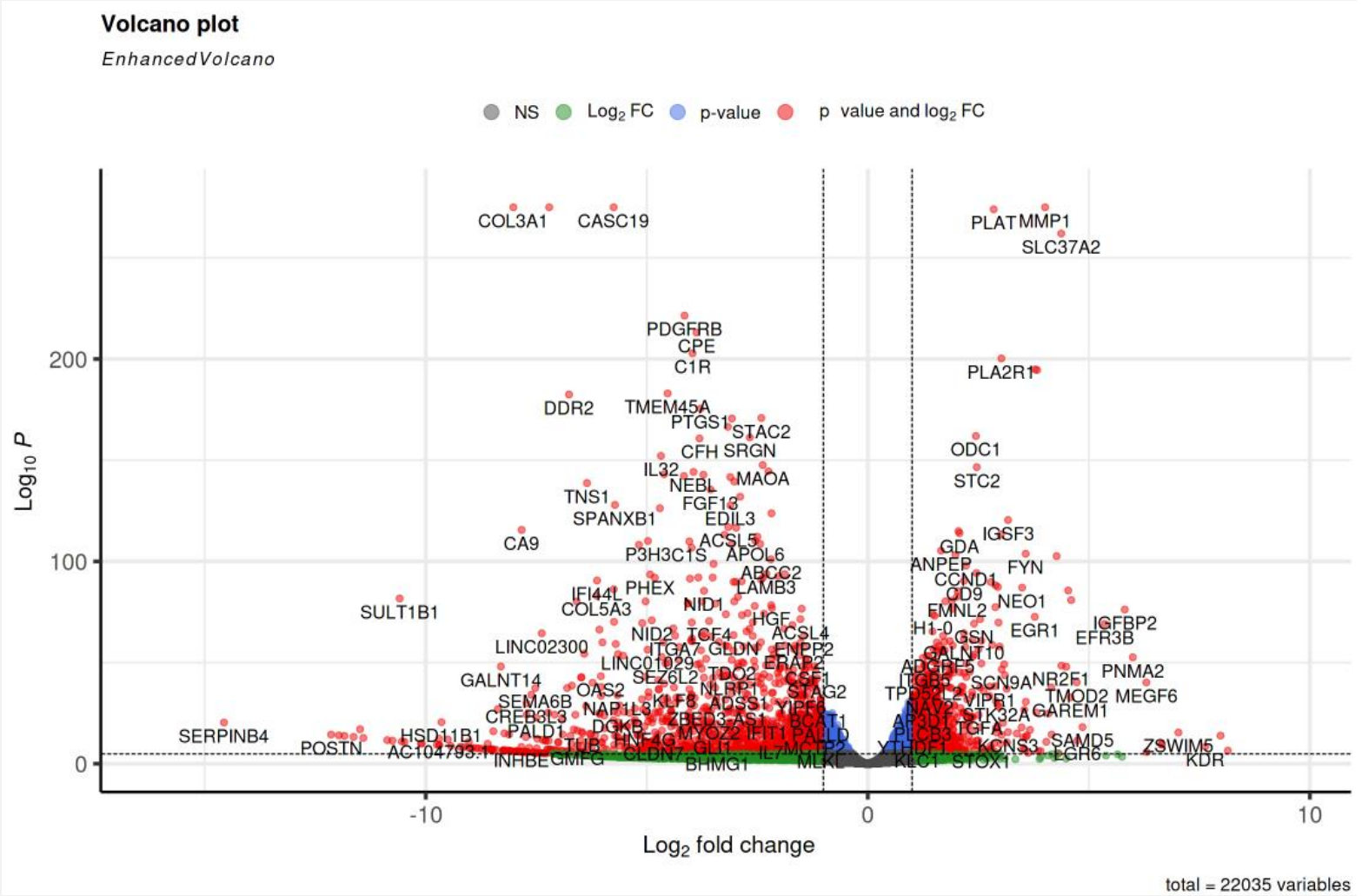
- Filter for only entries with valid adjusted p-values
- Significance only decided on adjusted p-value (BH)
- Often also filtered by logFC
  - Clear signal (p-value)
  - Strong signal (logFC)

```
outres_24h = as.data.frame(res_lfc_24h[! is.na(res_lfc_24h$padj),])
outres_24h = merge(outres_24h, gencodeAnnot, by.x=0, by.y="gene_stable_id")

outres_24h = outres_24h[order(outres_24h$padj), ]

normExpr_24h = counts(dds_24h, normalized=T)
cn_24h = colnames(normExpr_24h)
normExpr2_24h = cbind(rownames(normExpr_24h), normExpr_24h)
colnames(normExpr2_24h) = c("Geneid", cn_24h)


write.table(outres_24h, file="../bulk_own/24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")
write.table(normExpr2_24h, file="../bulk_own/normed_expr.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")

printres_24h = outres_24h[outres_24h$padj < 0.001, ]
write.table(printres_24h, file="../bulk_own/sig.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")

printres_24h.up = printres_24h[printres_24h$log2FoldChange > 1, ]
write.table(printres_24h.up , file="../bulk_own/sig.up.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")

printres_24h.down = printres_24h[printres_24h$log2FoldChange < -1, ]
write.table(printres_24h.down , file="../bulk_own/sig.down.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")
```

```
top24h.de = outres_24h[order(abs(outres_24h$log2FoldChange),decreasing=TRUE),]

top24h.de.500.down = head(top24h.de[top24h.de$log2FoldChange < -1,], n=500)
top24h.de.500.up = head(top24h.de[top24h.de$log2FoldChange > 1,], n=500)

write.table(top24h.de.500.down , file="../bulk_own/top.down.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")
write.table(top24h.de.500.up , file="../bulk_own/top.up.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")

write.table(top24h.de , file="../bulk_own/top.all.24h_sum159_vs_sum159_jq1r.tsv", row.names=F, quote=F, sep="\t")
```

DESEQ2

Results



- Enhanced Vsiualization, e.g. Volcano Plot
  - library(ggrepel)
  - library(EnhancedVolcano)

```
EnhancedVolcano(outres_24h,
    lab = outres_24h$gene_name,
    x = 'log2FoldChange',
    y = 'pvalue')
```

SET ENRICHMENT

Trying to identify common structures

estrogen receptor-negative breast cancer, C4733092

Source: ALL

Result

Gene	Uniprot	Gene Full Name
TNF	P01375	tumor necrosis factor
TP53	P04637	tumor protein p53
IL6	P05231	interleukin 6
VEGFA	P15692	vascular endothelial growth factor A
PIK3CA	P42336	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
BCL2	P10415	BCL2 apoptosis regulator
EGFR	P00533	epidermal growth factor receptor
CTNNB1	P35222	catenin beta 1
PTEN	P60484	phosphatase and tensin homolog
CDKN2A	P42771 Q8N726	cyclin dependent kinase inhibitor 2A
CXCL8	P10145	C-X-C motif chemokine ligand 8
AKT1	P31749	AKT serine/threonine kinase 1
PTGS2	P35354	prostaglandin-endoperoxide synthase 2
IGF1	P05019	insulin like growth factor 1
STAT3	P40763	signal transducer and activator of transcription 3
TLR4	O00206	toll like receptor 4
CCL2	P13500	C-C motif chemokine ligand 2
PIK3CD	O00329	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta

- Gene sets are just lists of genes
  - Sometimes called gene signatures
- Genes were found to be associated with gene set description
  - Here: estrogen receptor-negative breast cancer
  - Association crucial!
    - Automatic retrieval? Evidence? Context?
  - Gene sets/terms not independent!
    - Overlapping genes!
- Command-line tools or web service?
  - disgenet2r+DOSE+clusterProfiler
  - Metascape
- Method?
  - ORA/hypergeometric test
  - GSEA

DISGENET2R: DIRECTLY FINDING

- Follow manual: https://disgenet.org/static/disgenet2r/disgenet2r.html
  - Create account
  - Install disgenet2r

Selected gene symbols

Perform enrichment

```
input_genes = head(top24h.de.500.up$gene_name, n=500)

print(input_genes)

disgenet_input <- gene2disease( gene = input_genes, vocabulary = "HGNC", database = "ALL")

common_genes = intersect(disgenet_input@qresult$gene_symbol, input_genes)

print(paste("Remaining Gene Count: ", length(common_genes), "of", length(input_genes)))

res_enrich <-disease_enrichment( entities =common_genes, vocabulary = "HGNC", database = "ALL")
```

A data.frame: 19 × 7

	ID	Description	source	Ratio	BgRatio	pvalue	FDR
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>
373	C0009402	Colorectal Carcinoma	ALL	165/424	5473/21666	5.346205e-10	2.049735e-06
2247	C0178874	Tumor Progression	ALL	128/424	3865/21666	3.904144e-10	2.049735e-06
4343	C0686619	Secondary malignant neoplasm of lymph node	ALL	101/424	2825/21666	9.763622e-10	2.495582e-06
6593	C2939419	Secondary Neoplasm	ALL	90/424	2492/21666	6.356075e-09	1.218459e-05
1445	C0033578	Prostatic Neoplasms	ALL	69/424	1722/21666	9.714469e-09	1.440937e-05
3706	C0376358	Malignant neoplasm of prostate	ALL	138/424	4502/21666	1.127494e-08	1.440937e-05
5421	C1458155	Mammary Neoplasms	ALL	96/424	2780/21666	1.682902e-08	1.843499e-05
		Schizophrenia	ALL	98/424	2872/21666	2.084294e-08	1.997796e-05
		Prostate carcinoma	ALL	134/424	4388/21666	2.623150e-08	2.141940e-05
		Liver carcinoma	ALL	164/424	5725/21666	3.072684e-08	2.141940e-05
		Oestrogen receptor positive breast cancer		31/424	510/21666	2.989114e-08	2.141940e-05
		Malignant neoplasm of lung				4.860294e-08	3.105728e-05
		Carcinoma				0906e-08	4.046879e-05
		Squamous cell carcinoma of esophagus	ALL	75/424	2053/21666	37174e-08	5.278418e-05
		Primary malignant neoplasm of lung	ALL	120/424	3894/21666	1.197581e-07	6.122033e-05
4870	C0919267	ovarian neoplasm	ALL	87/424	2542/21666	1.407022e-07	6.743155e-05
26	C0001418	Adenocarcinoma	ALL	79/424	2235/21666	1.628121e-07	7.309775e-05
257	C0007103	Malignant neoplasm of endometrium	ALL	52/424	1235/21666	1.754802e-07	7.309775e-05
1045	C0024121	Lung Neoplasms	ALL	59/424	1486/21666	1.811238e-07	7.309775e-05

Breast cancer related

- Performs ORA on DisGeNET gene sets

DOSE/CLUSTERPROFILER

```
sel_df = outres_24h[outres_24h$log2FoldChange > 0,]

nsamples=1500
geneList = head(sel_df$log2FoldChange, n=nsamples)
names(geneList) = head(sel_df$gene_name, n=nsamples)
names(geneList) = mapIds(org.Hs.eg.db, head(sel_df$gene_name, n=nsamples), 'ENTREZID', 'SYMBOL')
geneList = geneList[!is.na(names(geneList))]

geneList = geneList[order(geneList, decreasing = TRUE)]
head(geneList)
```

- Requires list-objects with log-Foldchanges as value, and ENTREZID as name!
  - mapIds transfers gene symbols to entrez ID
    - Conversion is not exact: 1:m mapping, might not find EntrezID for all gene symbols

- Extremely well done tutorial available: https://yulab-smu.top/biomedical-knowledge-mining-book/dose-enrichment.html

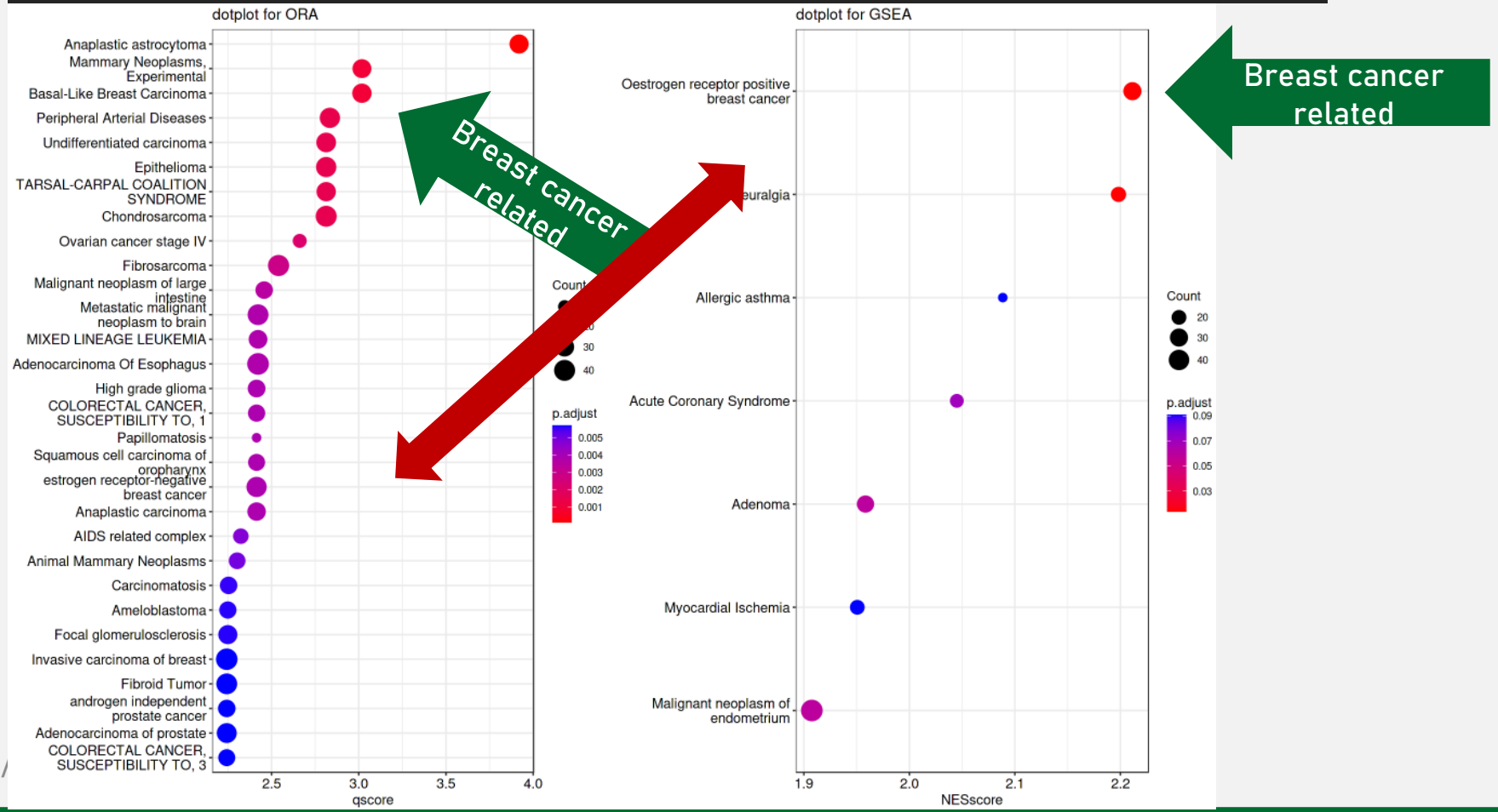
DOSE/CLUSTERPROFILER

Overrepresentation analysis (ORA) → hypergeometric test

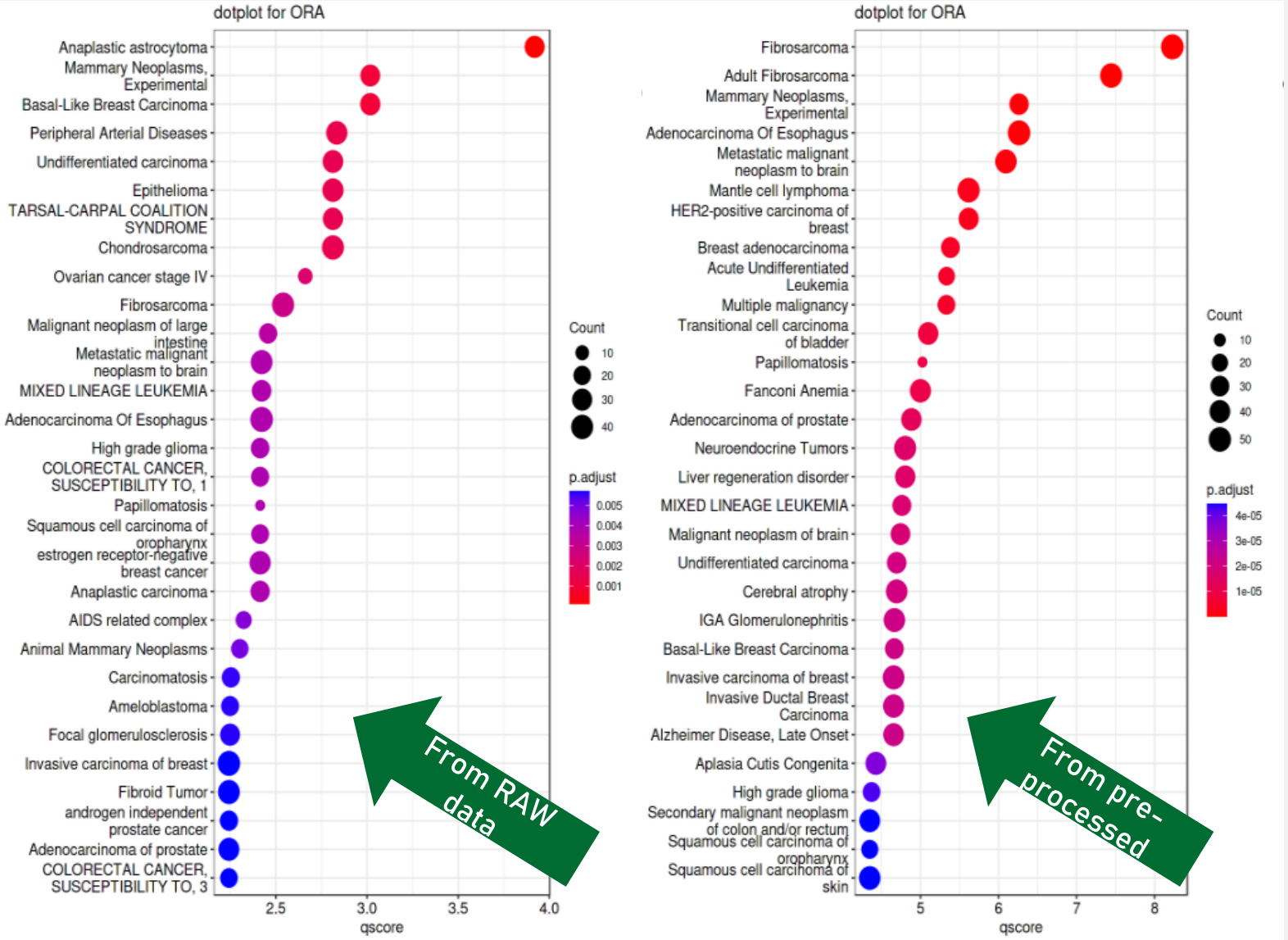
```
dgnEnrich <- enrichDGN(names(geneList),
                       pvalueCutoff  = 0.5,
                       pAdjustMethod = "BH")
dgnEnrich <- setReadable(dgnEnrich, 'org.Hs.eg.db')
```

Gene Set Enrichment Analysis → GSEA

```
dgnGSE <- gseDGN(geneList,
                 pvalueCutoff  = 0.5,
                 pAdjustMethod = "BH")
dgnGSE <- setReadable(dgnGSE, 'org.Hs.eg.db')
```



THE INPUT MATTERS



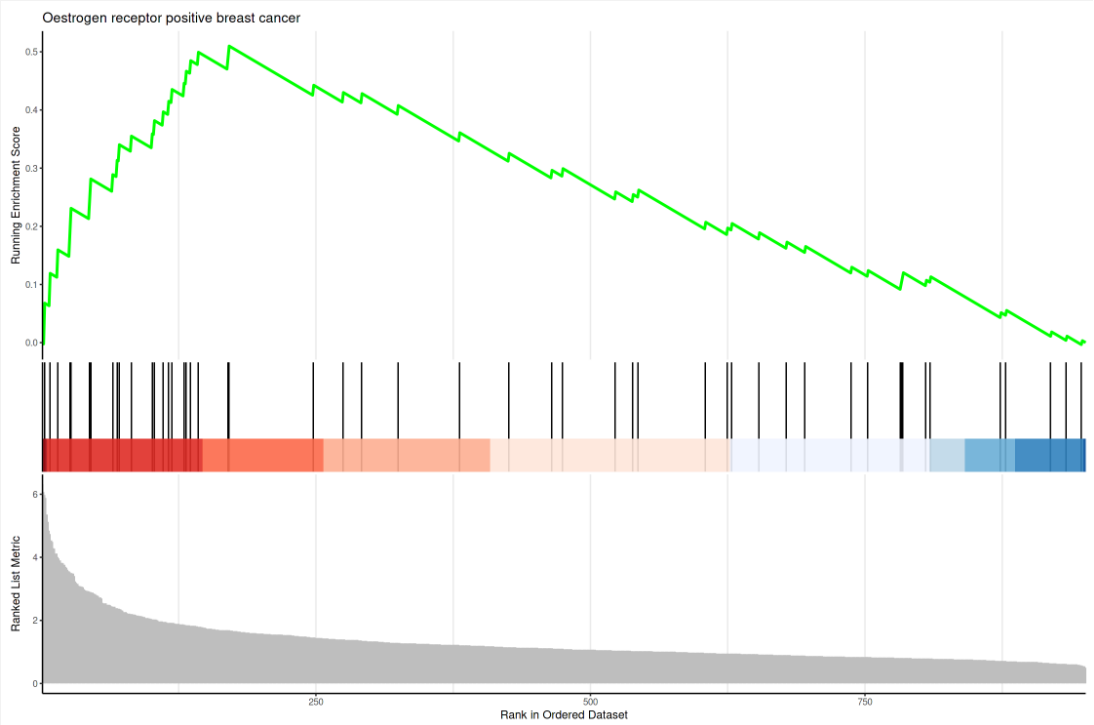
- Same data-set, different results!
- Why?!
  - Different gene annotations?
  - Different alignment?
  - Different quantification strategy?
  - Exclude non-protein-coding genes?

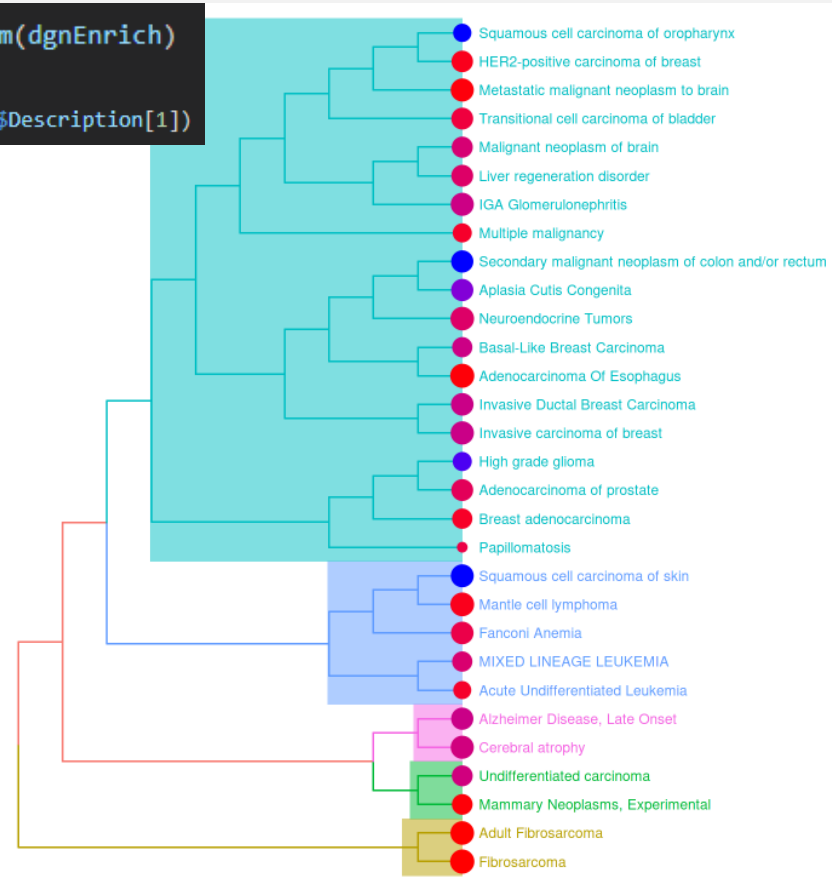
ADVANCED PLOTS

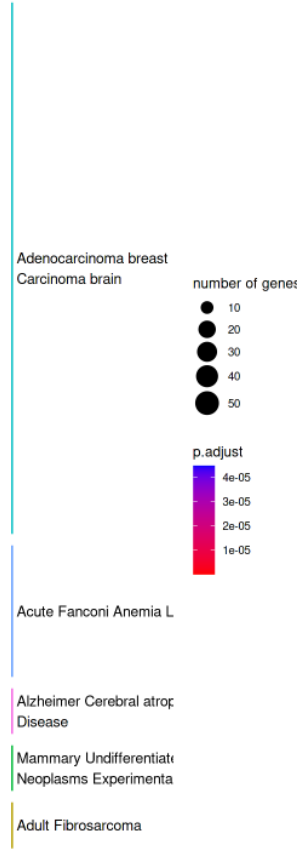
GSEA / Tree clustering

```
dgnEnrich2 <- pairwise_termsim(dgnEnrich)
treeplot(dgnEnrich2)
```

```
gseaplot2(dgnGSE, geneSetID = 1, title = dgnGSE$Description[1])
```







- The treeplot() function performs hierarchical clustering of enriched terms/gene sets
  - Jaccard's similarity index across enriched terms

GENE SET ENRICHMENT

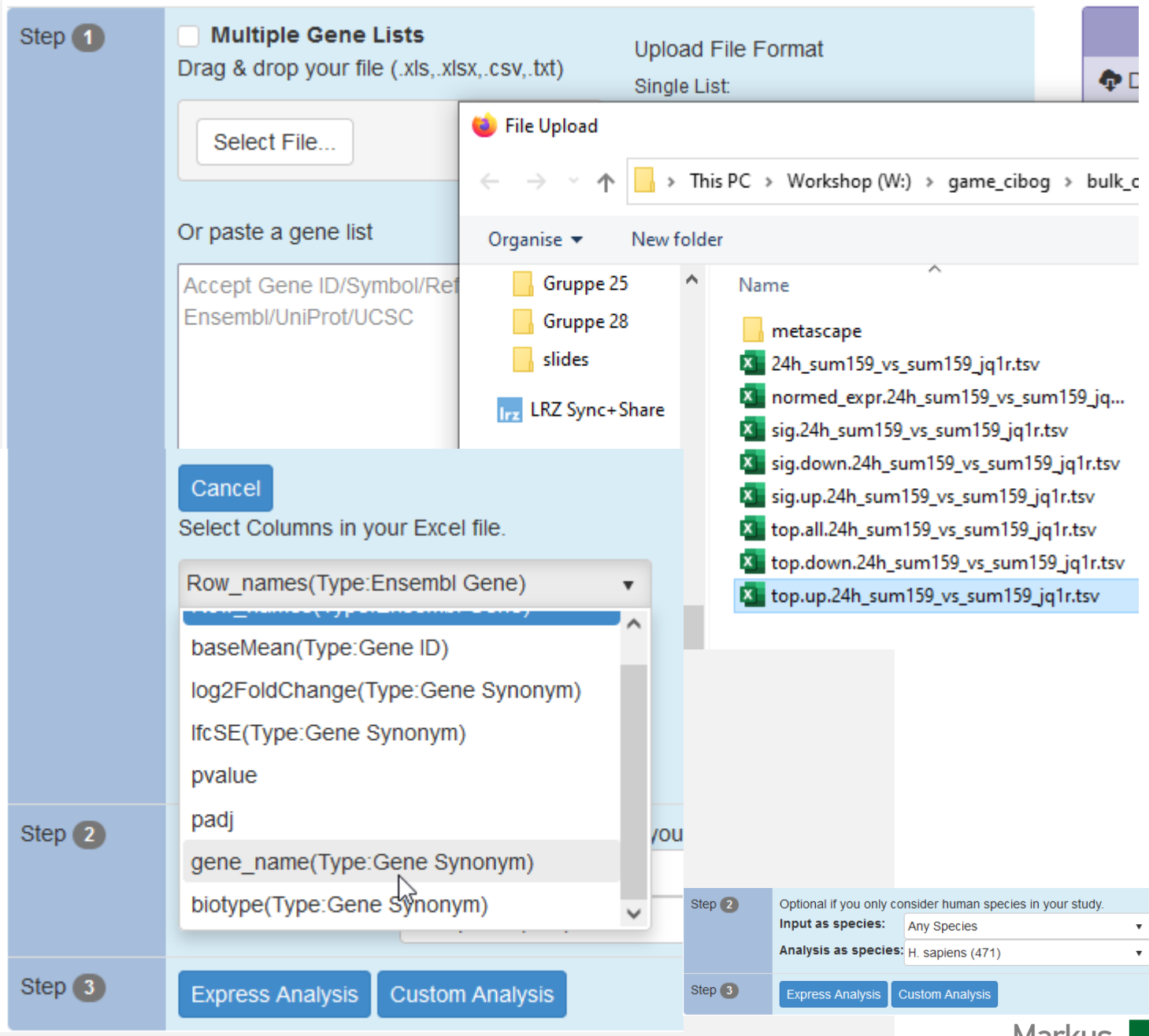
Using Metascape

- http://metascape.org/gp/index.html

- Use one of the results produced before
  - Here: Top500 regulated genes (logFC), up-regulated

- Select gene identifier
  - PROBEID is a gene synonym
  - H. sapiens is correctly recognized

- Start Express Analysis



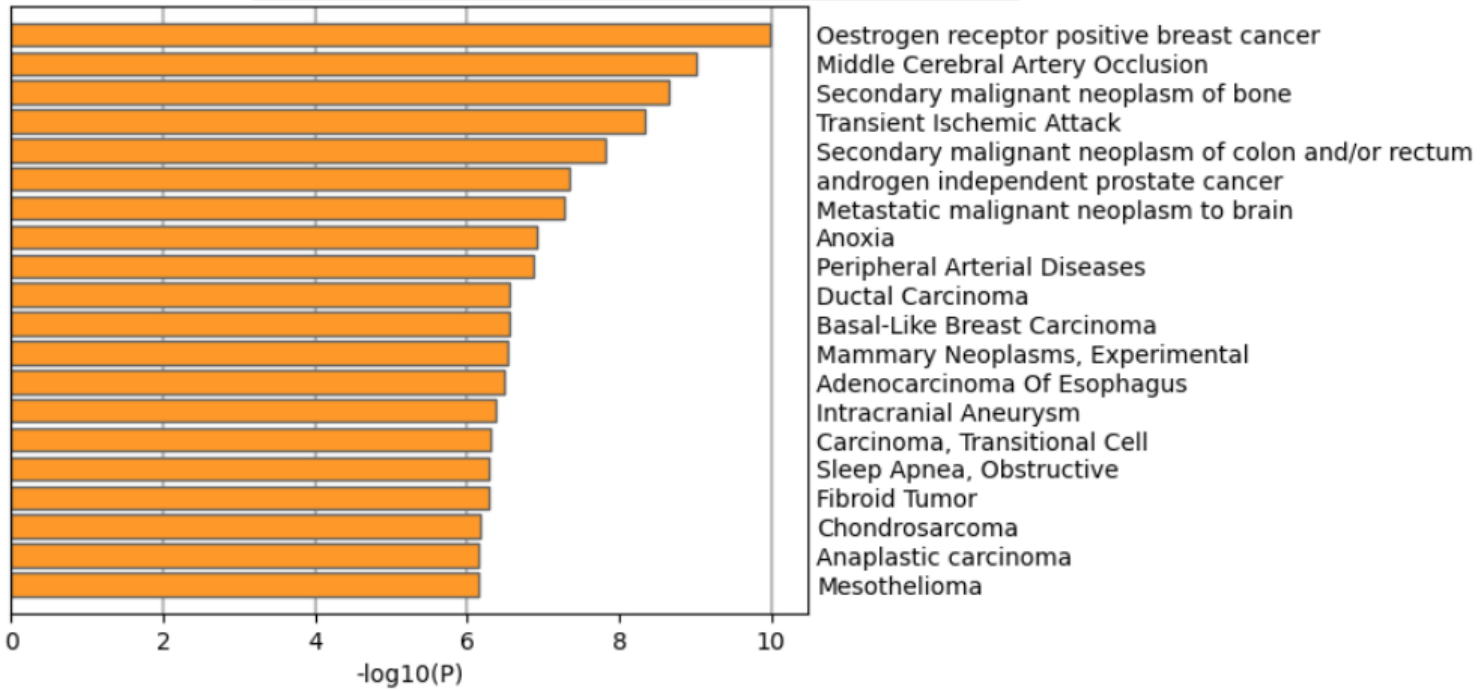
METASCAPE RESULTS

Save them or they are gone

Gene list enrichments are identified in the following ontology categories: COVID, Cell_Type_Signatures, DisGeNET, PaGenBase, TRRUST, Transcription_Factor_Targets. All genes in the genome have been used as the enrichment background. Terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5 (the enrichment factor is the ratio between the observed counts and the counts expected by chance) are collected and grouped into clusters based on their membership similarities. The top few enriched clusters (one term per cluster) are shown in the Figure 4-9. The algorithm used here is the same as that is used for pathway and process enrichment analysis.

PC › Workshop (W:) › game_cibog › bulk_own › metascape_top24h_up

Name	Date modified	Type	Size
CyJS	20/01/2022 12:53	File folder	
Enrichment_GO	20/01/2022 12:53	File folder	
Enrichment_heatmap	20/01/2022 12:53	File folder	
Enrichment_PPI	20/01/2022 12:53	File folder	
Enrichment_QC	20/01/2022 12:53	File folder	
Evidence	20/01/2022 12:53	File folder	
icon	20/01/2022 12:53	File folder	
all.t82vevc8s.zip	20/01/2022 12:53	Compressed (	
AnalysisReport.html	20/01/2022 03:52	Firefox HTML	
AnalysisReport.pptx	20/01/2022 03:52	Microsoft Pov	
metascape_result.xlsx	20/01/2022 03:52	Microsoft Exc	
README.txt	06/03/2021 15:49	Text Documer	



GENE SET ENRICHMENT RESULT COMPARISON

- Within zip-Download all tables available as excel file

Description	PARENT_C	LogP	Enrichmer	Z-score	#TotalGeneInLibrary	#GeneInGO	#GeneInHitList	#GeneInGOAndHitList	%InGO	STDV %In(	GeneID	Hits
Oestrogen receptor positive breast cancer		-10	4	8.4	30211	510	463	31	6.7	1.2	374|595|5	AREG|CCND1|BCL2|KLF5
Middle Cerebral Artery Occlusion		-9	3.4	7.7	30211	626	463	33	7.1	1.2	134|290|3	ADORA1|ANPEP|AREG|
Secondary malignant neoplasm of bone		-8.7	3.3	7.5	30211	647	463	33	7.1	1.2	374|596|1	AREG|BCL2|CCN2|DPYSL

- Does not occur in disgenet2r/DOSE analysis!
  - Different inputs!
  - Adj. p-Value sorted gene list in R
  - FC-sorted gene list for Metascape

- Not only statistical test highly relevant, also selected genes!

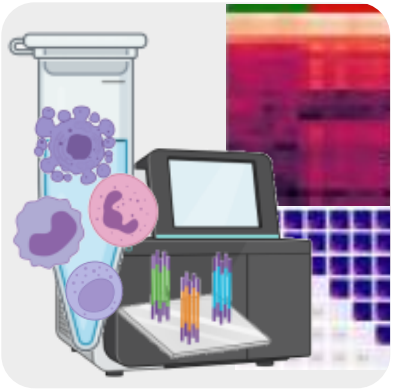
```
metascapeHits = c("AREG","CCND1","BCL2","KLF5","CHKA","EPHB2","EREG","FYN",
"GLI2","GSN","FOXA1","CXCL8","KDR","MAPT","MMP1","RAB27B","SLC20A1","TRPM2",
"FOSL1","STC2","MAP4K4","ADAMTS2","SEMA4D","TRIM29","SGK3","PAG1","LGR6",
"UBASH3B","DNER","E2F7","CD24")
doseInputGenes_pval = head(outres_24h$gene_name, n=nsamples)
pvalSortedOverlap = length(intersect(metascapeHits, doseInputGenes_pval))
doseInputGenes_fc = head(top24h.de.500.up$gene_name, n=nsamples)
fcSortedOverlap = length(intersect(metascapeHits, doseInputGenes_fc))

print(paste("Metascape set size:", length(metascapeHits)))
print(paste("Pval Sorted Overlap:", pvalSortedOverlap))
print(paste("logFC Sorted Overlap:", fcSortedOverlap))
```

✓ 0.5s

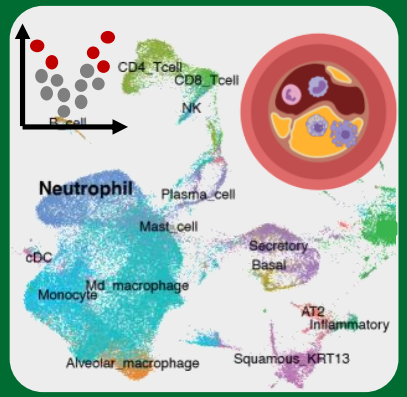
```
[1] "Metascape set size: 31"
[1] "Pval Sorted Overlap: 22"
[1] "logFC Sorted Overlap: 31"
```

SUMMARY

- Analyse public data from raw data (reads)
  - Or from pre-processed data
- Perform differential expression analysis using DESeq2
- Perform Gene Set Enrichment Analyses of significant genes
  - R-based: clusterProfiler, DOSE, disgenet2r, …
  - Web-based: Metascape
- At each “pipeline” step the choice of tools, methods and parameters affects the final result
  - Sorting of lists, number of genes for set enrichment, etc.



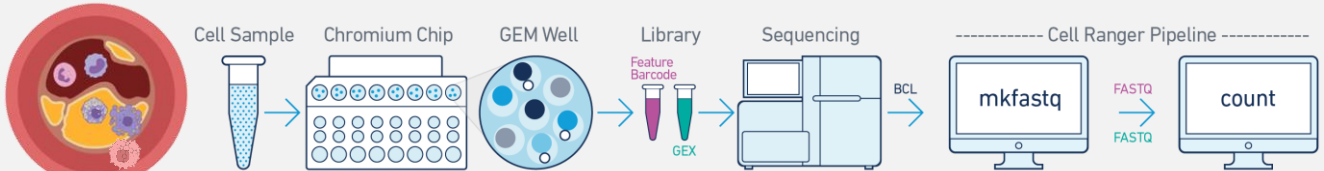
bulk RNA-seq

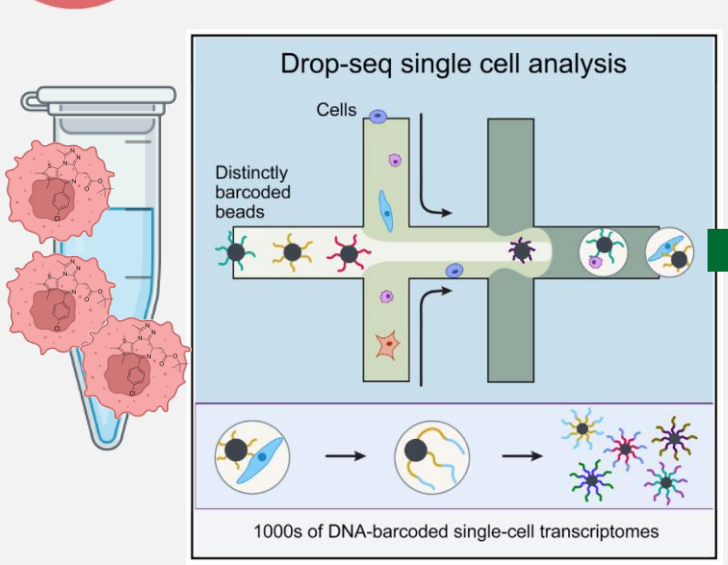


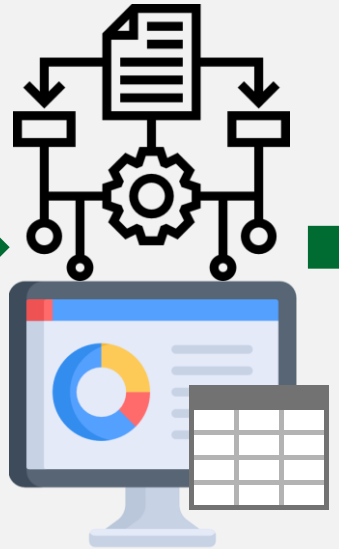
bulk RNA-seq

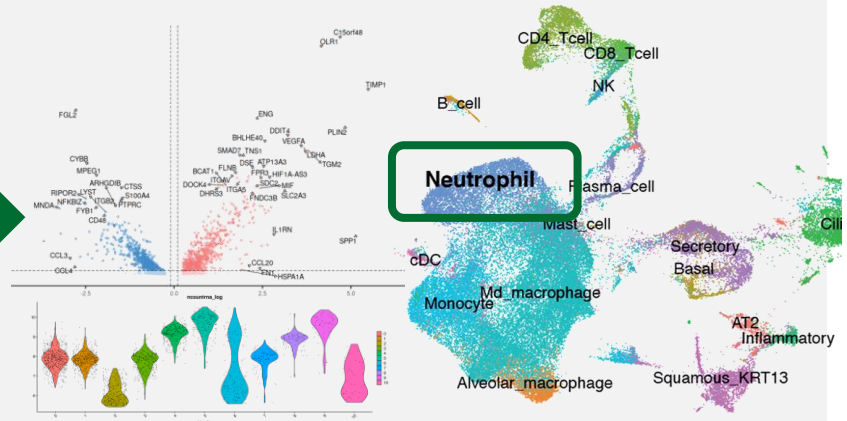
scRNA-seq

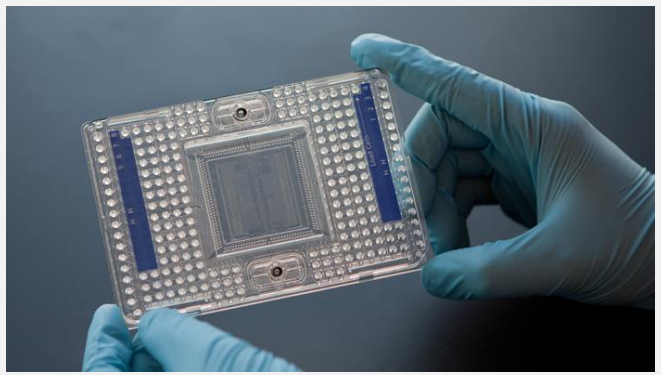
SINGLE CELL TRANSCRIPTOMICS









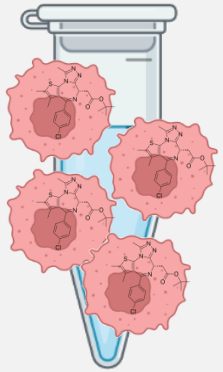


- Drop-seq has replaced well-based single-cell approaches

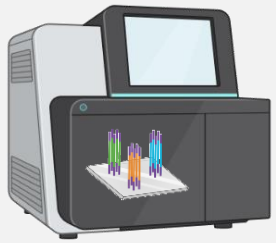
Sequencing image: biorender.com, Analysis: flaticon.com, Workflow: 10X Genomics

SINGLE CELL RNA-SEQ

SUM159

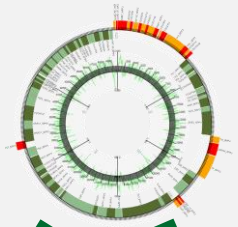


Material Isolation



Sequencing

Genome Annotation



Alignment

Cells

Genes

Quantification

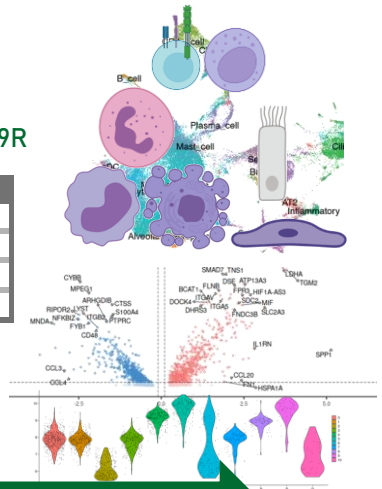
Integrate

SUM159+SUM159R cells

Genes

Cell clustering

Cell identification

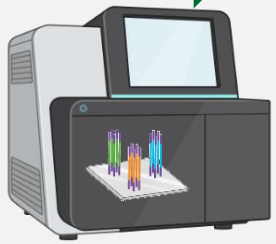


Differential Analysis

Gene Set Enrichment

Data Analysis

SUM159R



Cells

Genes

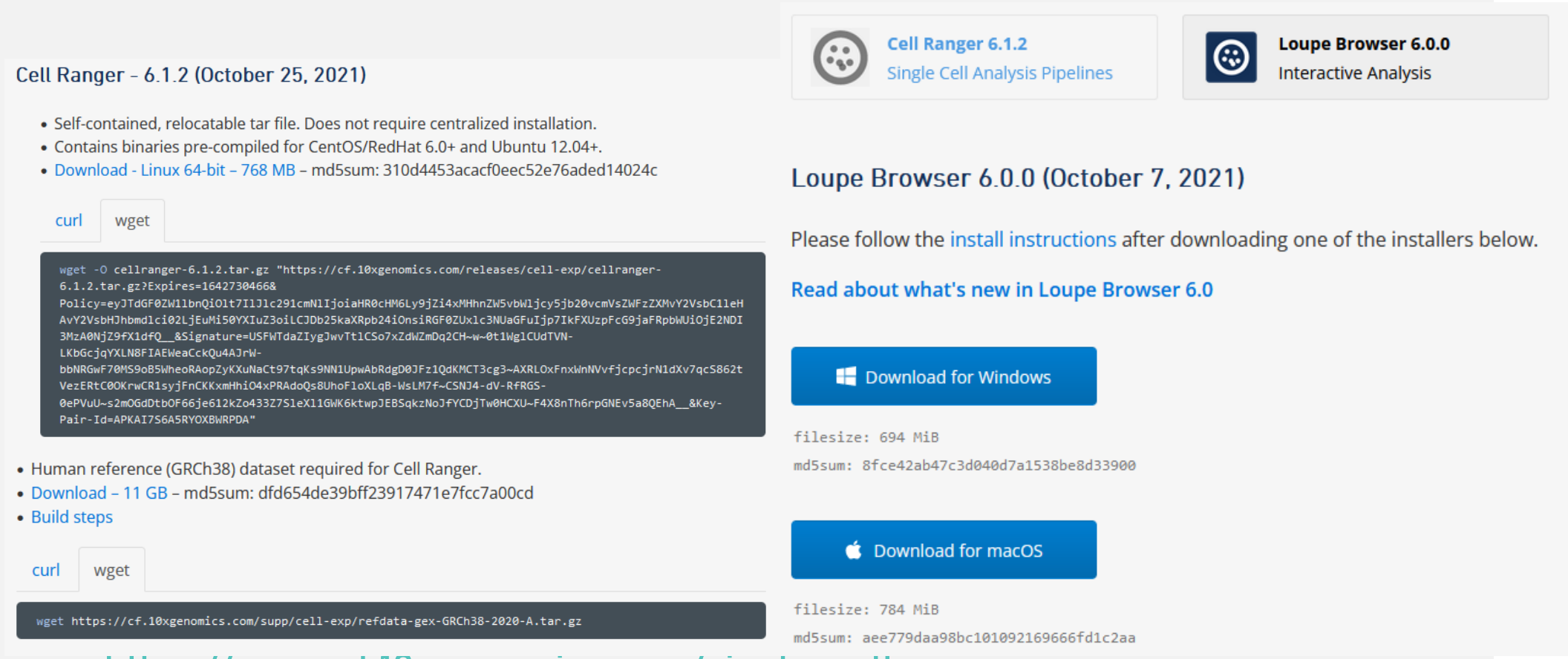
Quite some time

Matter of hours*
~12h

Matter of weeks

Sequencing image: biorender.com, Analysis: flaticon.com, Workflow: 10X Genomics

DOWNLOADING CELLRANGER+LOUPE



Cell Ranger – 6.1.2 (October 25, 2021)

- Self-contained, relocatable tar file. Does not require centralized installation.
- Contains binaries pre-compiled for CentOS/RedHat 6.0+ and Ubuntu 12.04+.
- Download - Linux 64-bit – 768 MB – md5sum: 310d4453acacf0eec52e76aded14024c

curl wget

```
wget -O cellranger-6.1.2.tar.gz "https://cf.10xgenomics.com/releases/cell-exp/cellranger-
6.1.2.tar.gz?Expires=1642730466&
Policy=eyJTdGF0ZW1lbnQiOlt7Il1lc291cmNlIjoiaHR0cHM6Ly9jZi4xMHhnZW5vbW1jcy5jb20vcmVsZWFzZXMvY2VsbC1leH
AvY2VsbHJhbmd1ci02LjEuMi50YXIuZ3oiLCJDb25kaXRpb24iOnsiRGF0ZUx1c3NUaGFuIjp7IkFXUzpFcG9jaFRpbWUiOjE2NDI
3MzA0NjZ9fX1dfQ__&Signature=USFWTdaZIygJwvTt1CSo7xZdWZmDq2CH~w~0t1WglCUdTVN-
LKbGcjqYXLN8FIAEWeaCckQu4AJrW-
bbNRGwF70MS9oB5WheoRAopZyKXuNaCt97tqKs9NN1UpwAbRdgD0JFz1QdKMCT3cg3~AXRLOxFnxWnNVvfjcpcjrN1dXv7qcS862t
VezERtC0OKrwCR1syjFnCKKxmHhiO4xPRAdoQs8UhoFloXLqB-WsLM7f~CSNJ4-dV-RfRGS-
0ePVuU~s2mOGdDtbOF66je612kZo433Z7SleX11GWK6ktwpJEBSqkzNoJfYCDjTw0HCXU~F4X8nTh6rpGNEv5a8QEhA__&Key-
Pair-Id=APKAI7S6A5RYOXBWRPDA"
```

- Human reference (GRCh38) dataset required for Cell Ranger.
- Download – 11 GB – md5sum: dfd654de39bff23917471e7fcc7a00cd
- Build steps

curl wget

```
wget https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz
```

Cell Ranger 6.1.2
Single Cell Analysis Pipelines

Loupe Browser 6.0.0
Interactive Analysis

Loupe Browser 6.0.0 (October 7, 2021)

Please follow the install instructions after downloading one of the installers below.

Read about what's new in Loupe Browser 6.0

Download for Windows

filesize: 694 MiB
md5sum: 8fce42ab47c3d040d7a1538be8d33900

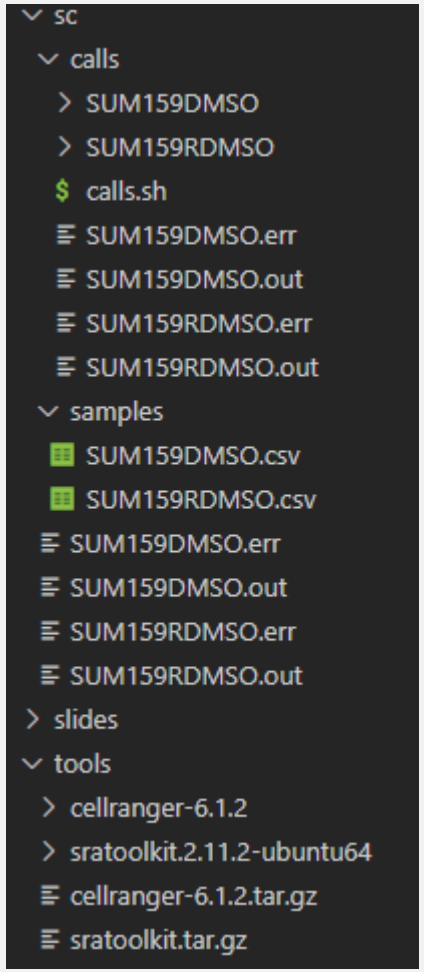
Download for macOS

filesize: 784 MiB
md5sum: aee779daa98bc101092169666fd1c2aa

- https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest
- Might require brief registration

PREPARING SINGLE CELL ANALYSES

Good folder structure is crucial!



- sc-dir
  - Calls subfolder: script to run cellranger+output
  - Samples: sample description files
- Tools
  - Cellranger executable

THE SAMPLE SPECIFICATION

samples > SUM159DMSO.csv

```
1  fastqs, sample, library_type
2  /mnt/w/game_cibog/reads/SUM159DMSO/, SRR9050597, Gene Expression
3  /mnt/w/game_cibog/reads/SUM159DMSO/, SRR9050598, Gene Expression
4  /mnt/w/game_cibog/reads/SUM159DMSO/, SRR9050599, Gene Expression
5  /mnt/w/game_cibog/reads/SUM159DMSO/, SRR9050600, Gene Expression
```

samples > SUM159RDMSO.csv

```
1  fastqs, sample, library_type
2  /mnt/w/game_cibog/reads/SUM159RDMSO/, SRR9050605, Gene Expression
3  /mnt/w/game_cibog/reads/SUM159RDMSO/, SRR9050606, Gene Expression
4  /mnt/w/game_cibog/reads/SUM159RDMSO/, SRR9050607, Gene Expression
5  /mnt/w/game_cibog/reads/SUM159RDMSO/, SRR9050608, Gene Expression
```

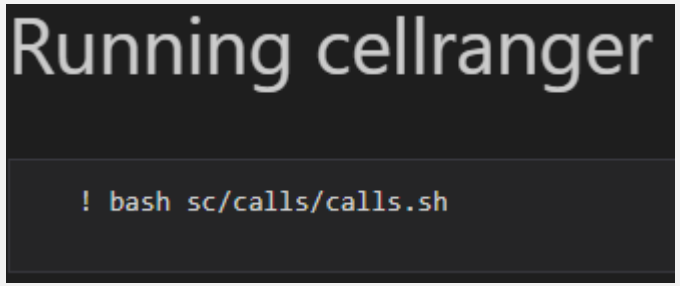
- Contains paths to all reads for a sample
- Specifies whether reads are from gene expression or antibody capture (e.g. CITE-seq)

THE CALLS SCRIPT

```
CURDIR=`pwd`

#path to cellranger binary/executable
CELLRANGER=$CURDIR/tools/cellranger-6.1.2/bin/cellranger
#path to the reference folder
REFFOLDER_HUMAN=$CURDIR/references/refdata-gex-GRCh38-2020-A

#this is for parallelizing cellranger. choose according to your infrastructure!
OPTS="--localcores 6 --localmem 30"

ls $CELLRANGER

BASE=./sc/

#for each sample run cellranger
for SAMPLE in SUM159DMSO SUM159RDMSO;
do

#path to sample file for current sample
SAMPLEFILE=$CURDIR/$BASE/samples/$SAMPLE.csv

#verify samplefile exists
ls $SAMPLEFILE

#if we rerun this sample, delete existing results
rm -rf $BASE/$SAMPLE

# run cellranger for the selected sample, with given reference and given read-files and optional parameters.
# save output in $SAMPLE.out and redirect error-messages to $SAMPLE.err
# wrapping $CELLRANGER into /usr/bin/time --verbose will print how long the run took and how much RAM was used lateron
(cd $BASE/calls/ && /usr/bin/time --verbose $CELLRANGER count --id=$SAMPLE --transcriptome=$REFFOLDER_HUMAN --libraries=$SAMPLEFILE $OPTS > $SAMPLE.out 2> $SAMPLE.err)

done
```

RUNNING CELLRANGER

Organisation matters



CELLRANGER OUTPUT

During the run

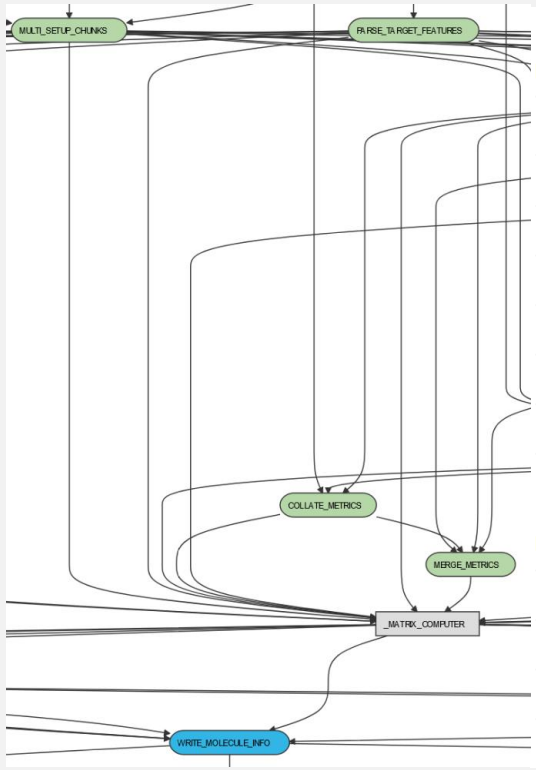
```
Martian Runtime - v4.0.6
Serving UI at http://spectre3:37647?auth=2WXiTSo22fyUhBGH5BoHt-G9wr5ORVX096wuG0F9yCE

Running preflight checks (please wait)...
Checking sample info...
Checking FASTQ folder...
Checking reference...
Checking reference_path (/mnt/w/game_cibog/references/refdata-gex-GRCh38-2020-A) on spectre3...
Checking optional arguments...
mrc: v4.0.6

mrp: v4.0.6

Anaconda: Python 3.8.2

numpy: 1.19.2

scipy: 1.6.2

pysam: 0.16.0.1

h5py: 3.2.1

pandas: 1.2.4

STAR: 2.7.2a

samtools: samtools 1.10
Using htslib 1.10.2
Copyright (C) 2019 Genome Research Ltd.

2022-01-12 10:03:28 [runtime] (ready)           ID.SUM159DMSO.SC_RNA_COUNTER_CS.SC_MULTI_CORE.MULTI_CHEMISTRY_DETECTOR.
2022-01-12 10:03:28 [runtime] (run:local)       ID.SUM159DMSO.SC_RNA_COUNTER_CS.SC_MULTI_CORE.MULTI_CHEMISTRY_DETECTOR.
2022-01-12 10:03:28 [runtime] (ready)           ID.SUM159DMSO.SC_RNA_COUNTER_CS.FULL_COUNT_INPUTS.WRITE_GENE_INDEX
```

Web Interface

Logging output

```
Elapsed (wall clock) time (h:mm:ss or m:ss): 5:10:07
Average shared text size (kbytes): 0
Average unshared data size (kbytes): 0
Average stack size (kbytes): 0
Average total size (kbytes): 0
Maximum resident set size (kbytes): 15888584
```

CELLRANGER PROGRESSION



Pipestance Details

Runtime

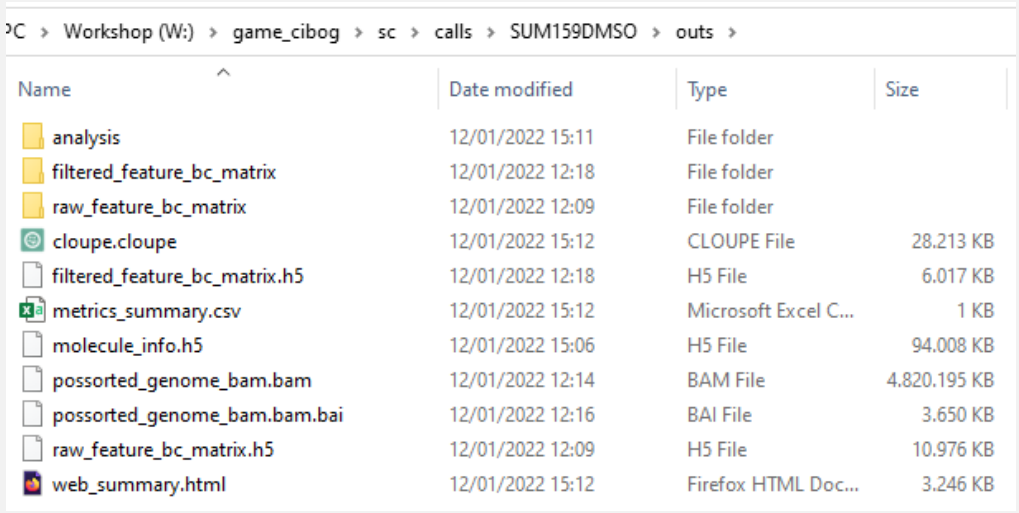
State	running
Cmdline	mrp __SUM159DMSO.mro SUM159DMSO --jobmode local --localcores 6 --localmem 30
User	mjoppich@spectre3, PID=4205
Job Mode	local (6 cores, 30 GB)
Start Time	2022-01-12 10:03:16
Env	MROPORT=37647, MROPROFILE=disable
Versions	martian=v4.0.6, pipelines=cellranger-6.1.2
Logging	log

Paths

Bin	/mnt/w/game_cibog/tools/cellranger-6.1.2/external/martian/bin/mrp
Cwd	/mnt/w/game_cibog/sc/calls
MROPATH	/mnt/w/game_cibog/tools/cellranger-6.1.2/mro
MRO File	__SUM159DMSO.mro

- You can also keep track on progression with web interface
  - But you gotta wait long enough anyway ;)

CELLRANGER OUTPUT



- web_summary.html for quick look at the data
  - Alignment rate, etc.
- cloupe.cloupe file can be used for analysis

- raw_feature_bc_matrix
  - File and folder
  - All barcodes

- Filtered_feature_bc_matrix
  - Cellranger filtered barcodes

WEB SUMMARY

1,135
Estimated Number of Cells

56,468
Mean Reads per Cell

4,676
Median Genes per Cell

Sequencing

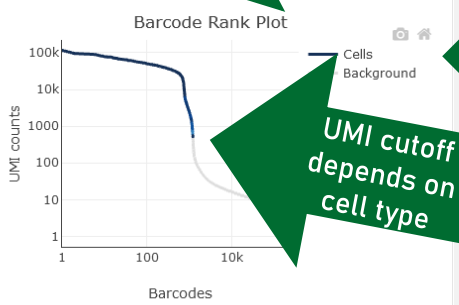
Number of Reads	64,090,665
Number of Short Reads Skipped	0
Valid Barcodes	98.3%
Valid UMIs	100.0%
Sequencing Saturation	31.2%
Q30 Bases in Barcode	97.0%
Q30 Bases in RNA Read	78.8%
Q30 Bases in UMI	96.7%

Mapping

Reads Mapped to Genome	91.4%
Reads Mapped Confidently to Genome	89.3%
Reads Mapped Confidently to Intergenic Regions	3.6%
Reads Mapped Confidently to Intronic Regions	12.5%
Reads Mapped Confidently to Exonic Regions	73.2%
Reads Mapped Confidently to Transcriptome	71.1%
Reads Mapped Antisense to Gene	0.8%

- Estimated number of cells
- Mean reads/genes per cell
- Sequencing information
  - About 25.000-50.000 reads per cell should be planned
- Mapping statistics
  - Most reads should map to exon and transcriptome!

Cells



Kneeplot

Filtered contains blue cells

UMI cutoff depends on cell type

Estimated Number of Cells	1,135
Fraction Reads in Cells	95.9%
Mean Reads per Cell	56,468
Median UMI Counts per Cell	27,135
Median Genes per Cell	4,676
Total Genes Detected	20,548

Sample

Sample ID	SUM159DMSO
Sample Description	
Chemistry	Single Cell 3' v2
Include introns	False
Reference Path	…references/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-6.1.2

For your ref.

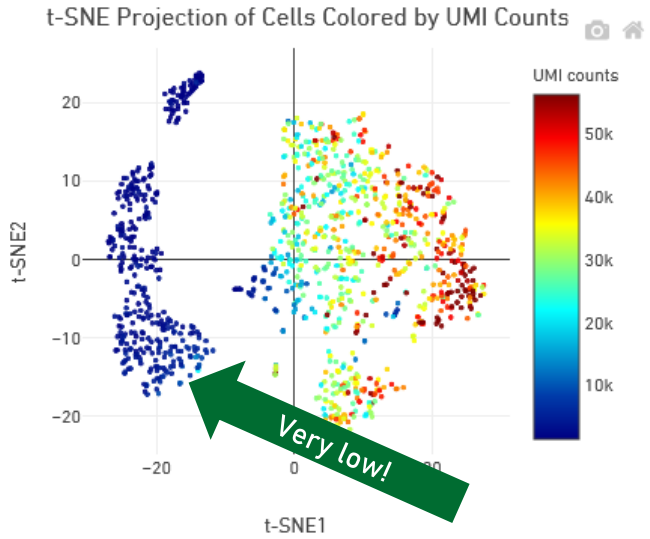
WEB SUMMARY

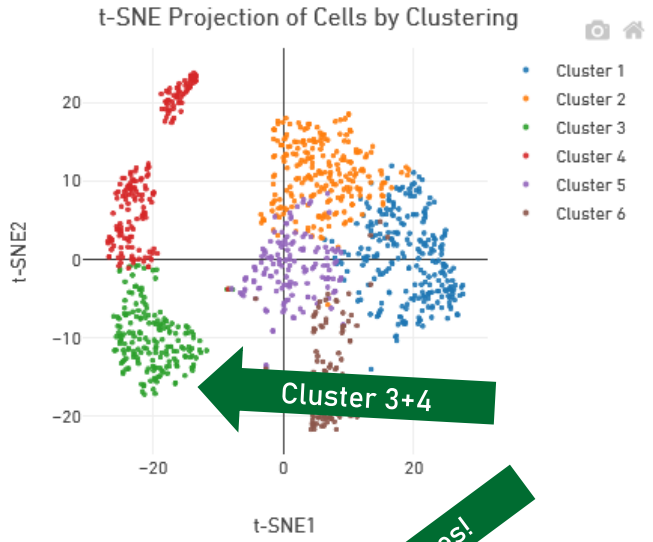
Very brief analysis

- Initial clustering of all filtered cells
- Marker genes per cluster
- Spot strange clusters which might need further attention
  - High MT-content = damaged cells

t-SNE Projection

Clustering Type: Graph-based





MT-Genes!

Top Features by Cluster (Log2 fold-change, p-value)

Feature		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value
ENSG00000198840	MT-ND3	-1.58	9e-25	-0.68	1e-3	3.79	1e-46	4.02	4e-26	-0.58	4e-1	-0.96	8e-5
ENSG00000198899	MT-ATP6	-1.31	7e-18	-0.49	4e-2	3.36	3e-33	3.76	3e-21	-0.60	4e-1	-0.50	1e-1
ENSG00000198938	MT-CO3	-1.33	2e-18	-0.41	1e-1	3.31	2e-32	3.66	6e-20	-0.39	8e-1	-0.57	6e-2
ENSG00000198763	MT-ND2	-1.30	9e-18	-0.40	1e-1	3.31	2e-32	3.63	1e-19	-0.45	7e-1	-0.56	7e-2
ENSG00000198804	MT-CO1	-1.32	2e-18	-0.47	5e-2	3.37	5e-34	3.63	2e-19	-0.52	5e-1	-0.46	2e-1
ENSG00000198886	MT-ND4	-1.21	1e-15	-0.31	3e-1	3.23	1e-30	3.48	2e-17	-0.43	8e-1	-0.58	5e-2
ENSG00000198695	MT-ND6	-1.24	2e-15	-0.29	3e-1	3.30	4e-30	3.58	2e-17	-0.85	9e-2	-0.48	2e-1
ENSG00000198727	MT-CYB	-1.27	4e-17	-0.30	3e-1	3.18	3e-29	3.47	2e-17	-0.37	8e-1	-0.46	2e-1
ENSG00000198712	MT-CO2	-1.27	3e-17	-0.37	1e-1	3.15	1e-28	3.47	2e-17	-0.36	8e-1	-0.32	5e-1
ENSG00000198786	MT-ND5	-1.15	5e-14	-0.25	4e-1	3.22	9e-30	3.39	8e-16	-0.56	5e-1	-0.59	5e-2

Previous | Page 1 of 28 | 10 rows | Next

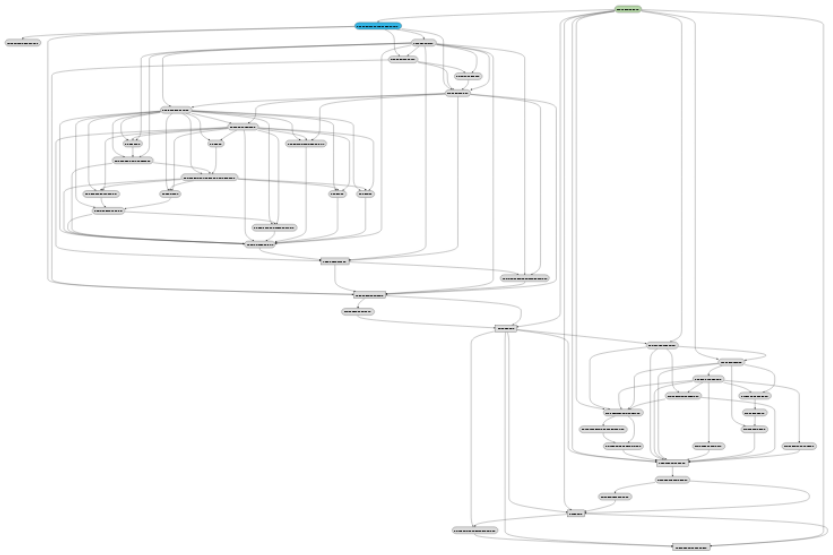
CELLRANGER AGGREGATE

Combines two separate libraries

After having executed cellranger, we must merge both datasets:

```
! cd sc/aggr_calls/ && ../../tools/cellranger-6.1.2/bin/cellranger aggr --localmem 30 --localcores 8 --csv aggr_samples.csv --id SUM159DMSO_RDMSO
```

10X Martian Pipeline Runner / mjoppich / SUM159DMSO_RDMSO / SC_RNA_AGGREGATOR_CS



Pipestance Details

Runtime

State	running
Cmdline	mrp __SUM159DMSO_RDMSO.mro SUM159DMSO_RDMSO --jobmode local --localcores 8 --localmem 30
User	mjoppich@spectre3, PID=20549
Job Mode	local (8 cores, 30 GB)
Start Time	2022-01-12 21:18:27
Env	MROPORT=33385, MROPROFILE=disable
Versions	martian=v4.0.6, pipelines=cellranger-6.1.2
Logging	log

Paths

Bin	/mnt/w/game_cibog/tools/cellranger-6.1.2/external/martian/bin/mrp
Cwd	/mnt/w/game_cibog/sc/aggr_calls
MROPATH	/mnt/w/game_cibog/tools/cellranger-6.1.2/mro
MRO File	__SUM159DMSO_RDMSO.mro

```
@include "rna/sc_rna_aggregator_cs.mro"

call SC_RNA_AGGREGATOR_CS(
    sample_id             = "SUM159DMSO_RDMSO",
    sample_desc           = "",
    pipestance_root       = "/mnt/w/game_cibog/sc/aggr_calls",
    aggregation_csv       = "/mnt/w/game_cibog/sc/aggr_calls/aggr_samples.csv",
    normalization_mode    = "mapped",
    no_secondary_analysis = false,
)
```

sc > aggr_calls > aggr_samples.csv

Absolute pathes

```
1  sample_id,molecule_h5
2  SUM159DMSO,/mnt/w/game_cibog/sc/calls/SUM159DMSO/outs/molecule_info.h5
3  SUM159RDMSO,/mnt/w/game_cibog/sc/calls/SUM159RDMSO/outs/molecule_info.h5
```

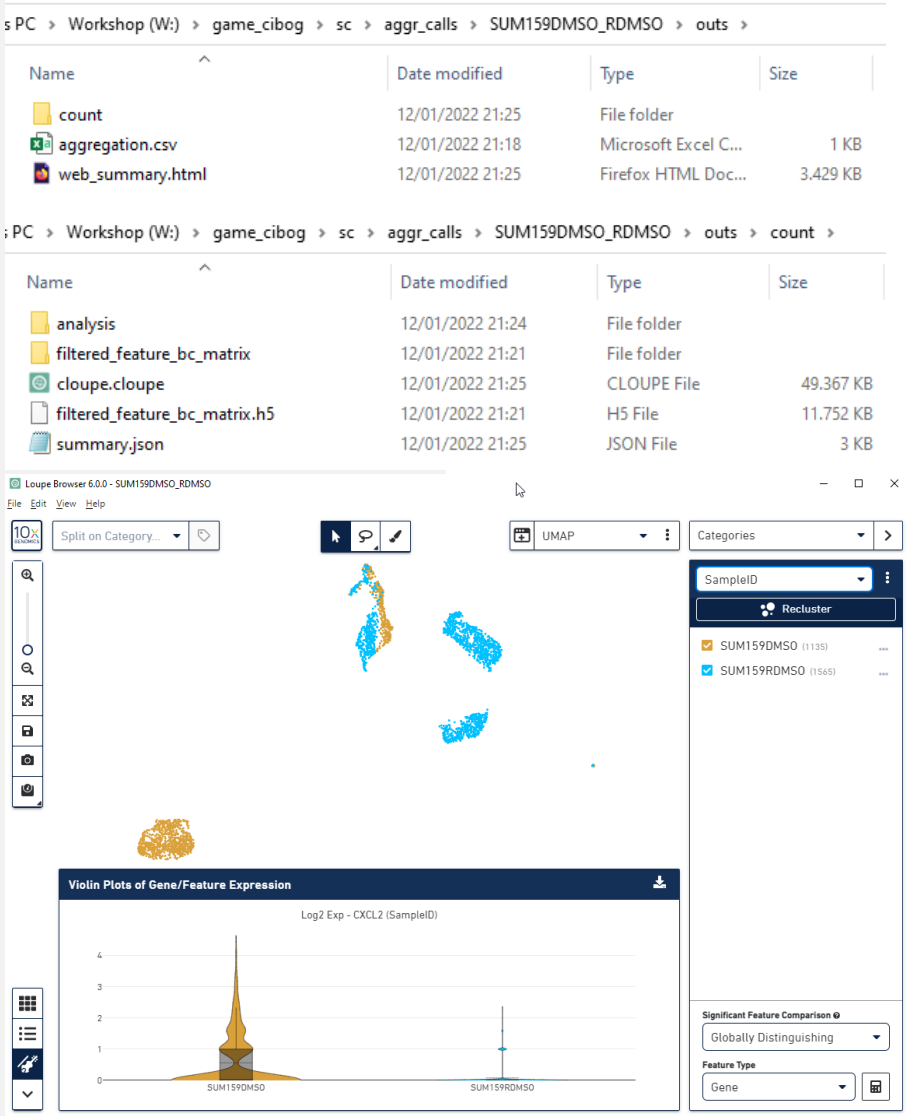
CELLRANGER AGGREGATE OUTPUT

Similar to other outputs

- All “filtered” cells from two cellranger count runs
  - With Seurat or scanpy cell selection possible
    - Maybe the low UMI-count cells are actual cells?

- Integration procedure? Unknown.
  - With Seurat or scanpy multiple ways possible
    - SCTransform, bbknn, rpca, etc.

- cloupe file allows further analysis without using command line tools
  - Usage of scnapy/Seurat strongly encouraged!
    - But worth another workshop!

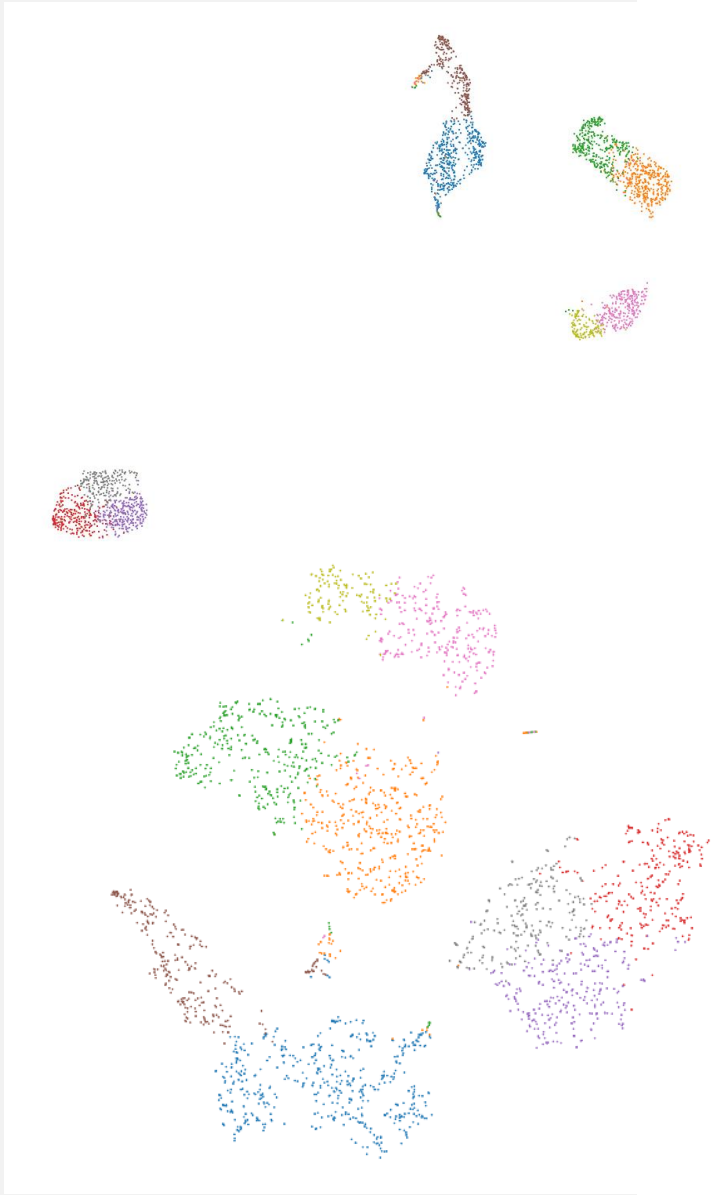


SCRNA-SEQ ANALYSIS WITH LOUPE

Just the most important things

- Complete tutorial:
  https://support.10xgenomics.com/single-cell-gene-expression/software/visualization/latest/tutorial

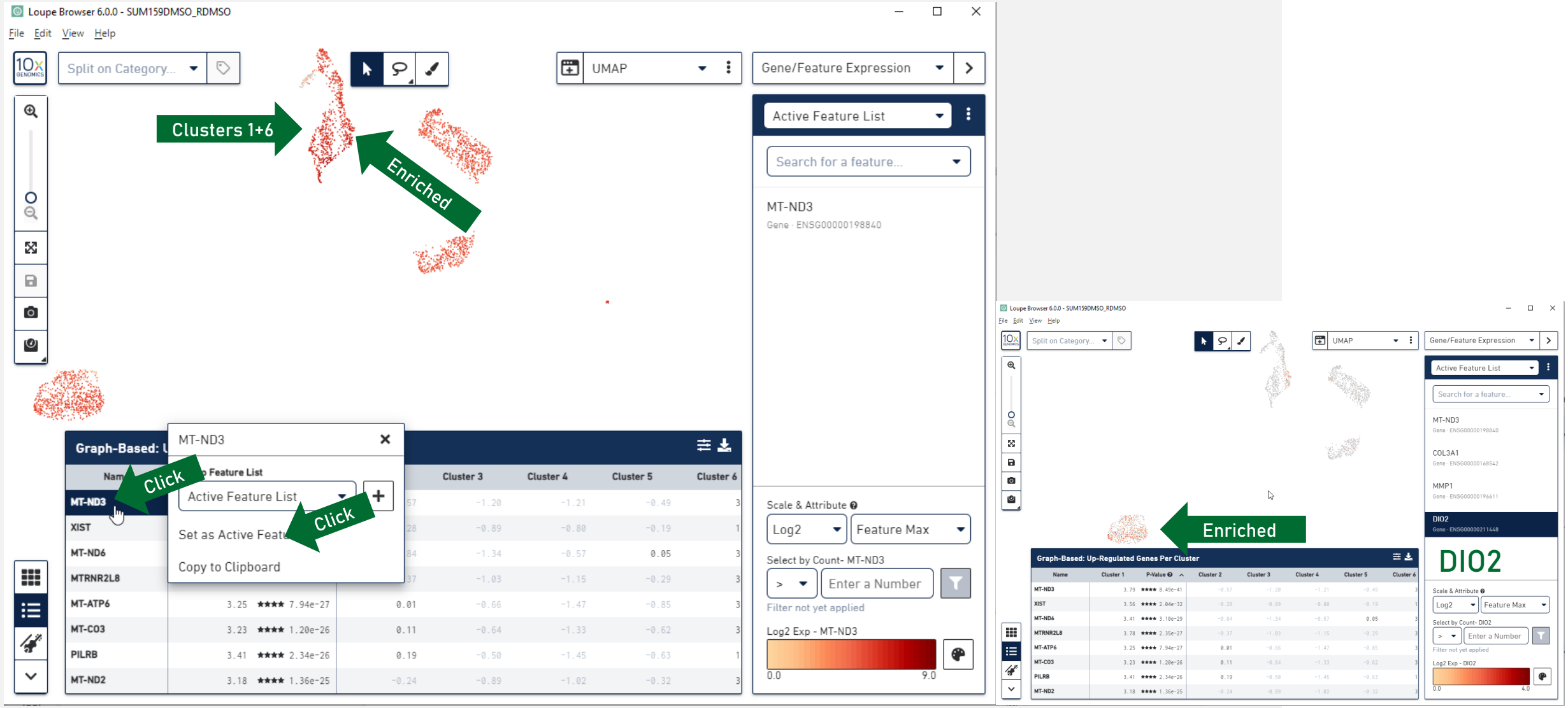
- Loupe allows
  - Visualizations
    - UMAP, t-SNE (whereas UMAP preserves local and global distances better)
    - Violin plots, Heatmaps, Feature plots
  - Re-analysis



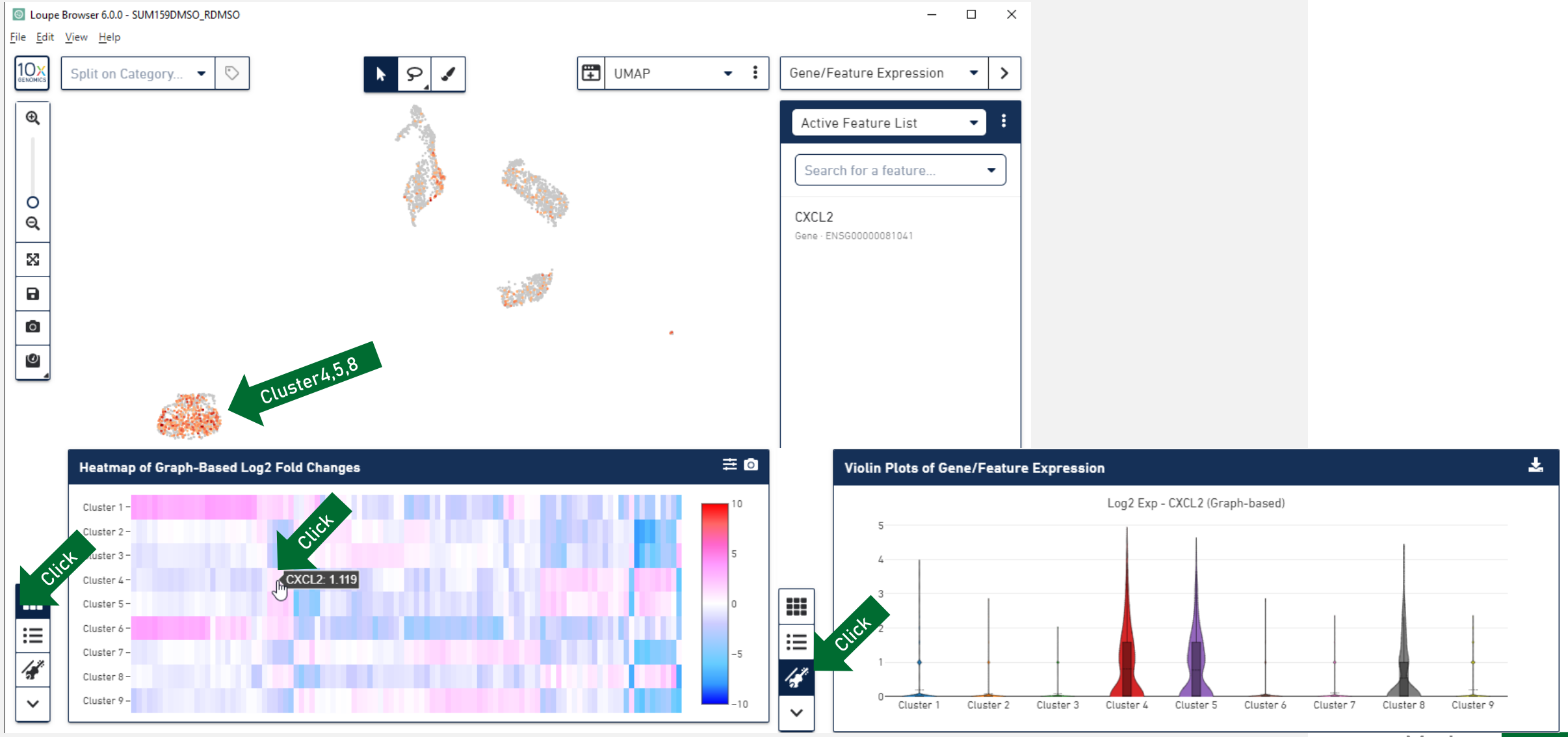
UMAP

t-SNE

LOUPE: INSPECT FEATURES

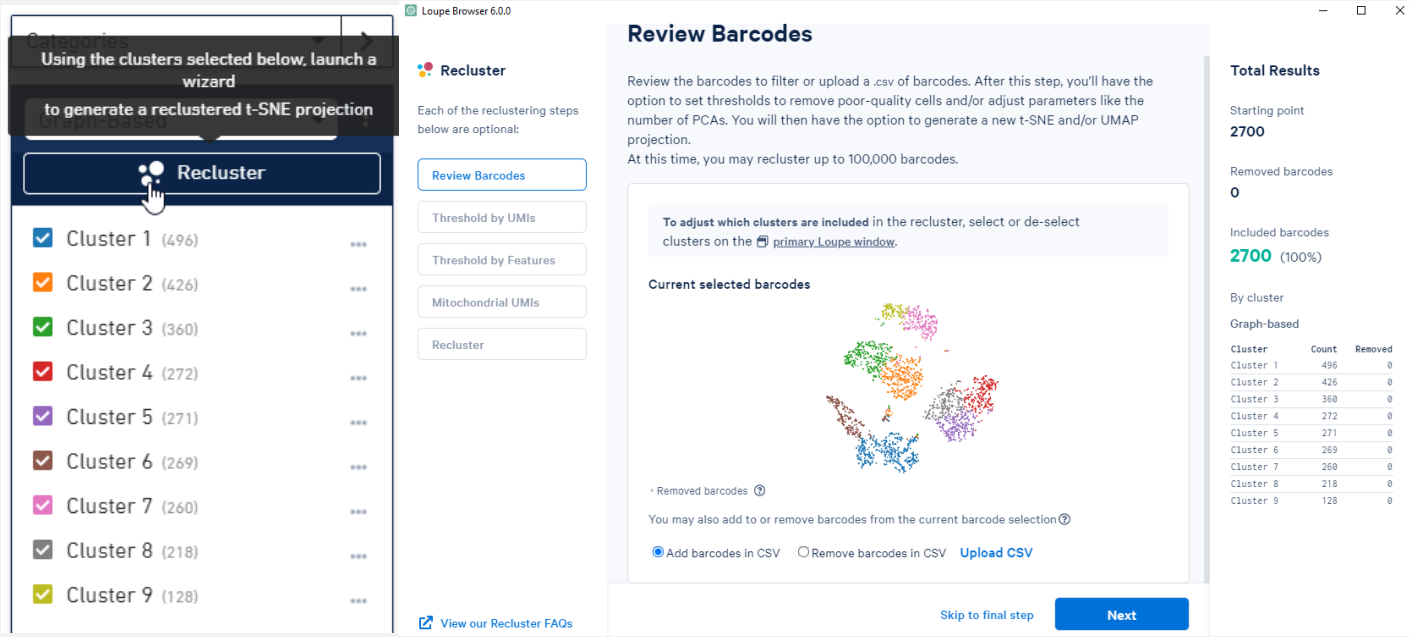


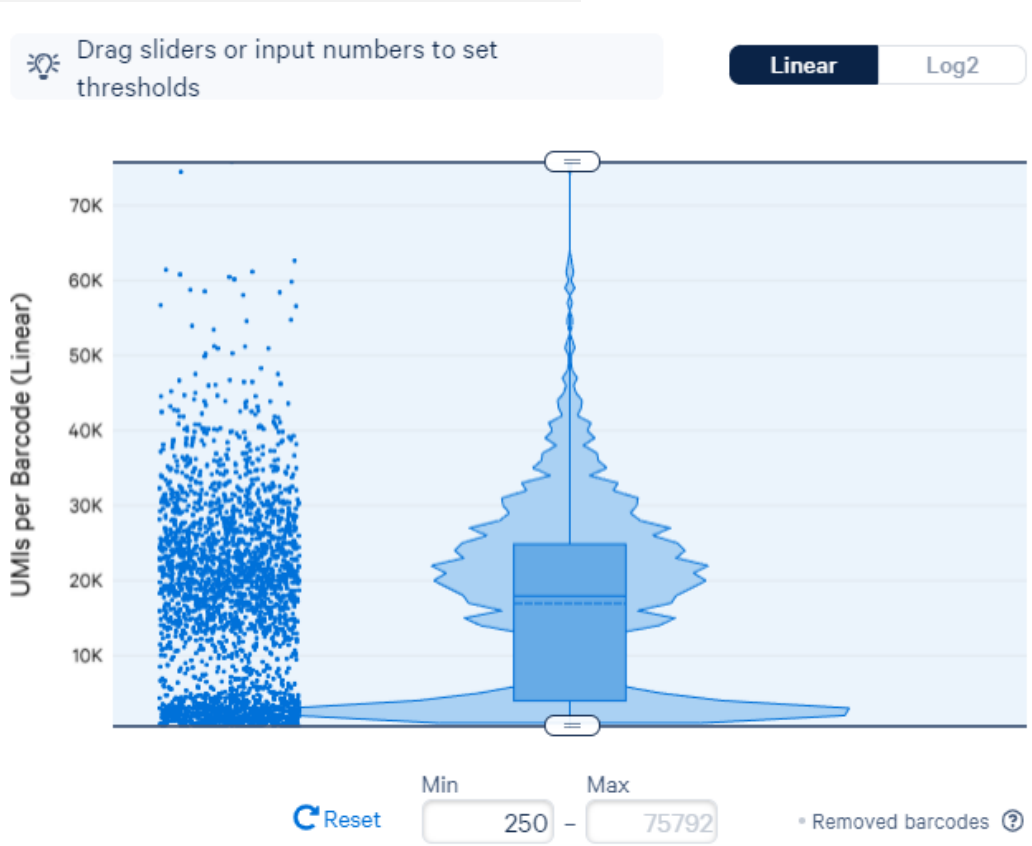
LOUPE: HEATMAP



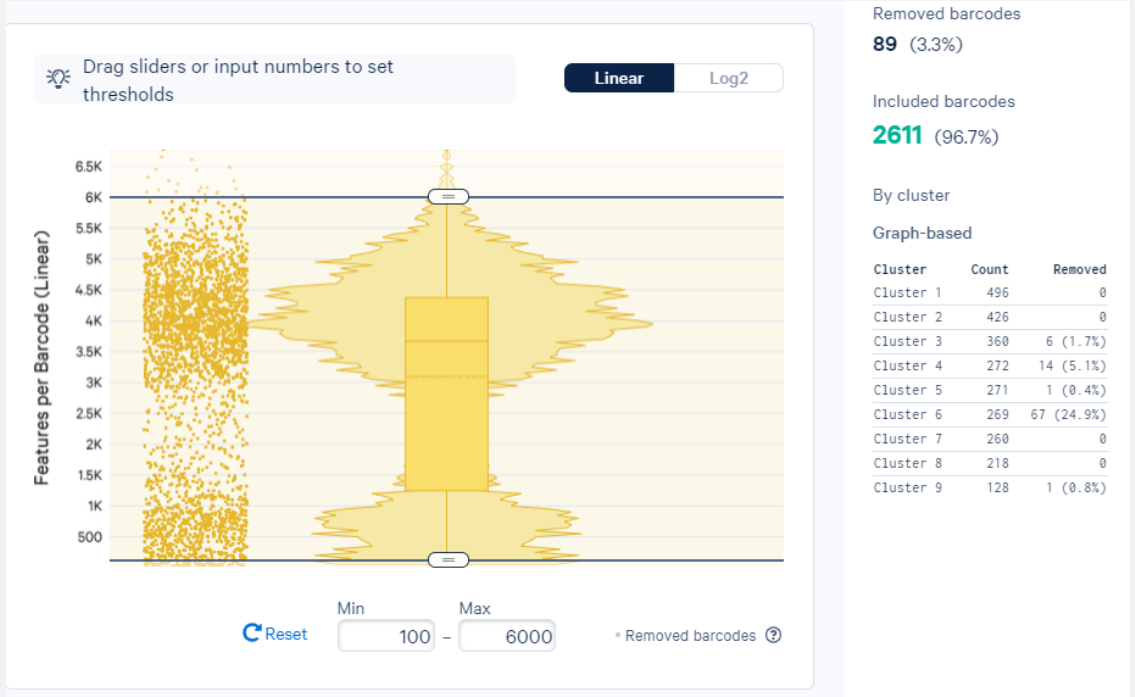
LOUPE: RECLUSTER

- Performs a new analysis
- Let's you remove unwanted cells
  - Such as cells with too high mt-Content





LOUPE: RECLUSTER



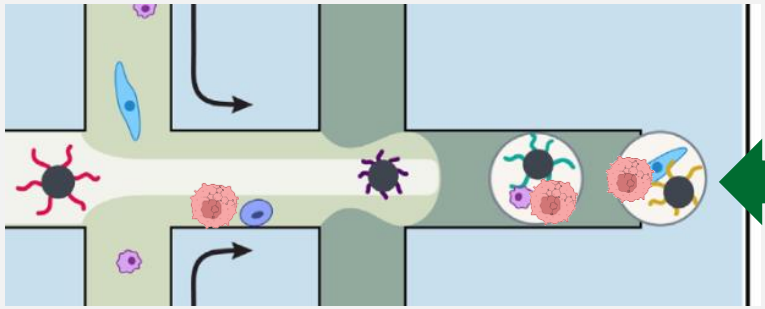
Removed barcodes
89 (3.3%)

Included barcodes
2611 (96.7%)

By cluster

Graph-based

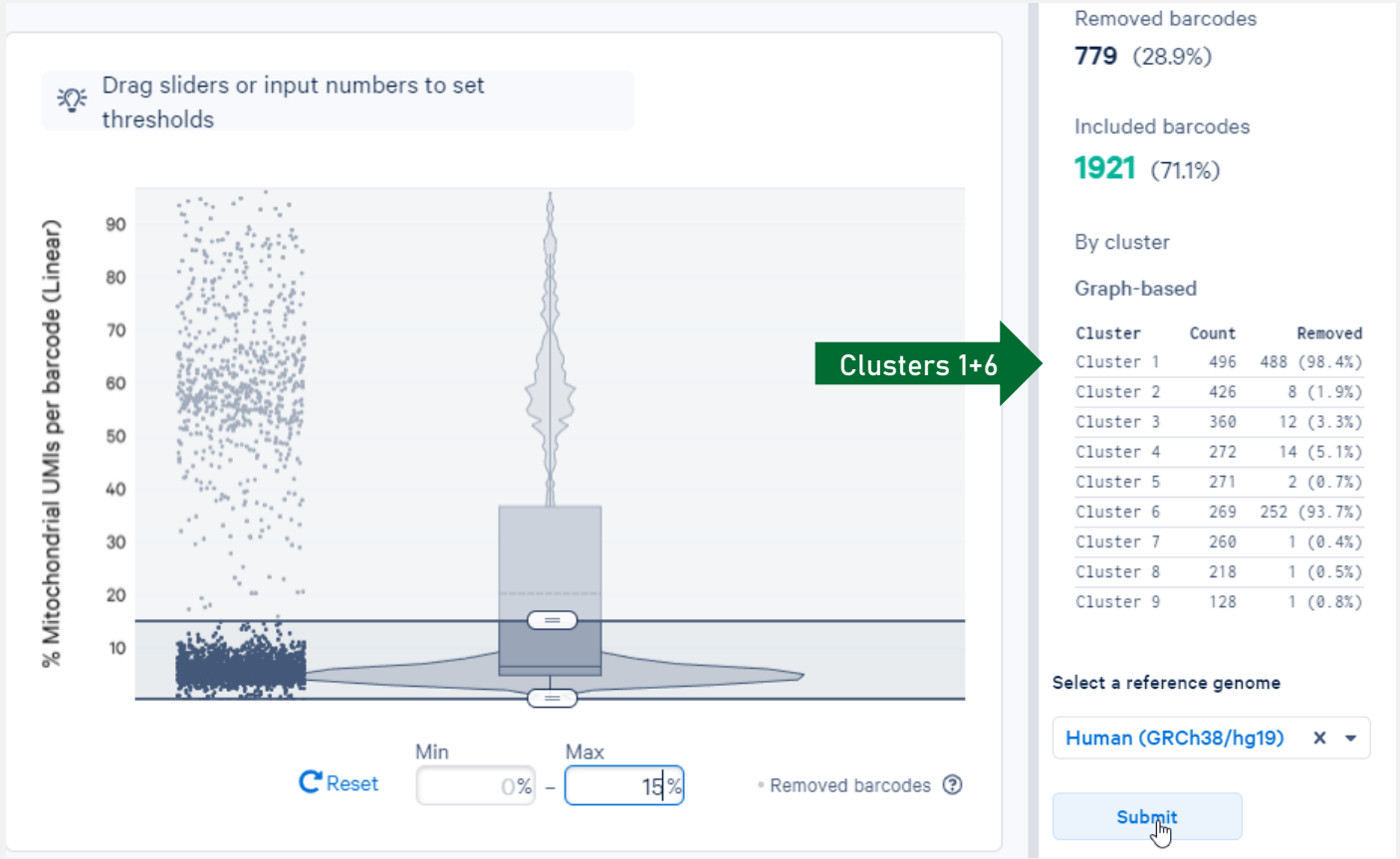
Cluster	Count	Removed
Cluster 1	496	0
Cluster 2	426	0
Cluster 3	360	6 (1.7%)
Cluster 4	272	14 (5.1%)
Cluster 5	271	1 (0.4%)
Cluster 6	269	67 (24.9%)
Cluster 7	260	0
Cluster 8	218	0
Cluster 9	128	1 (0.8%)



- Less than 100 genes might just be fragmented cells
- More than 6000 cells might be doublets
  - 2 cells per droplet

Sequencing image: biorender.com, Analysis: flaticon.com, Workflow: 10X Genomics

LOUPE: RECLUSTER



- 15% mt-Content is maximum
  - But depends on cell type
- Often also just 5%

LOUPE: RECLUSTER

Recluster

Select the type of plot to generate *

☑ t-SNE ☑ UMAP

Select at least one type of plot to generate.

Adjust reanalyze parameters (for advanced users)

Name the recluster *

Give it a good name

250UMI_100_6000FEATURE_MT15

This will generate a new view on the same Loupe file.
You'll still be able to refer back to the original view.

You may click on any of the steps in the left panel to go back and make changes.

Total Results

Starting point
2700

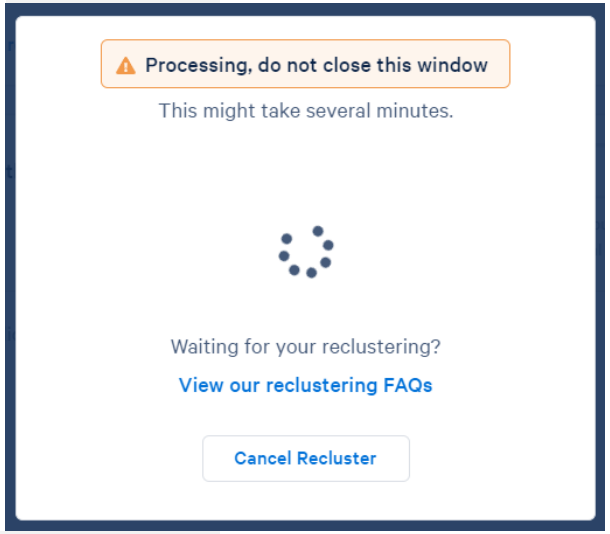
Removed barcodes
779 (28.9%)

Included barcodes
1921 (71.1%)

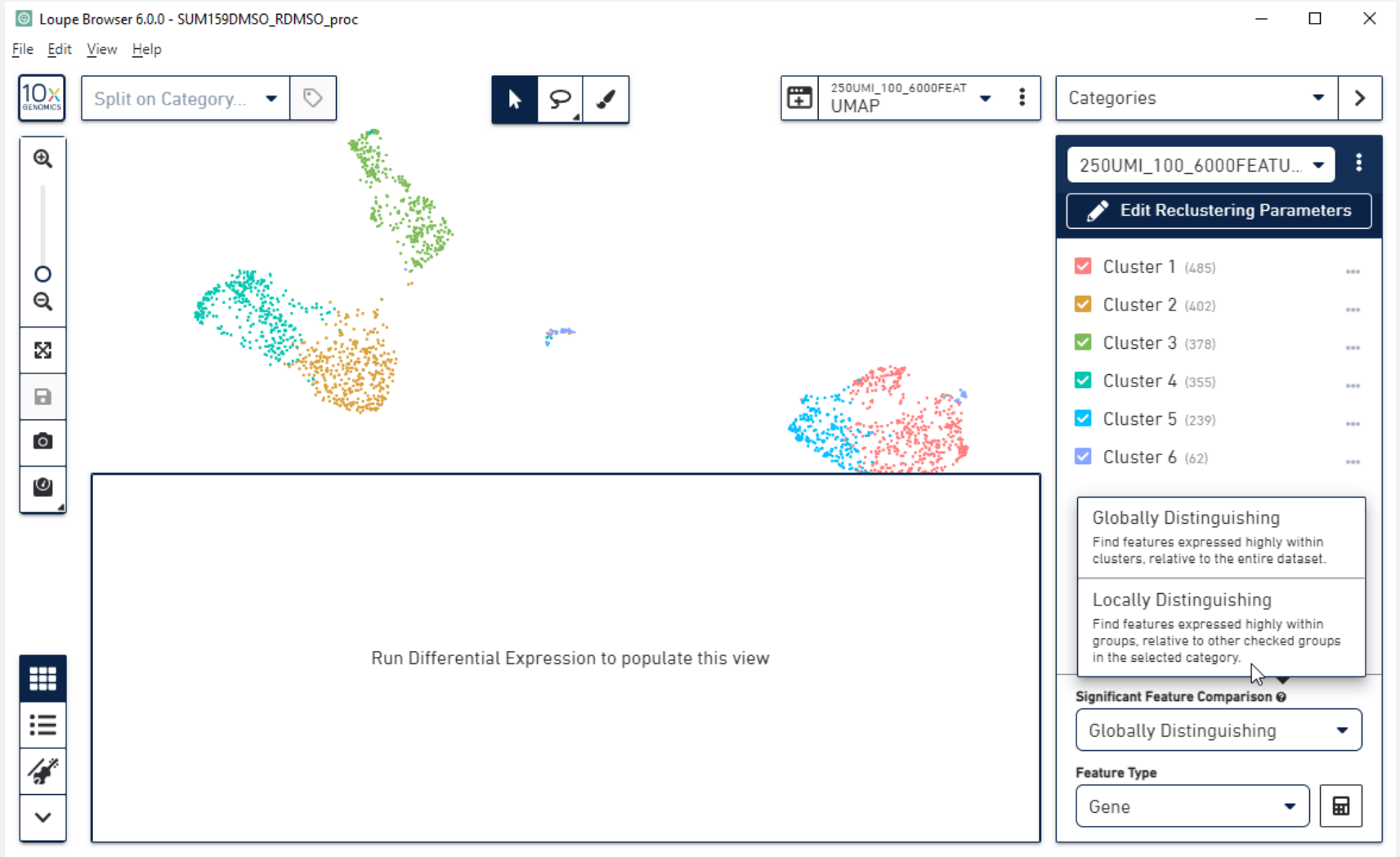
By cluster

Graph-based

Cluster	Count	Removed
Cluster 1	496	488 (98.4%)
Cluster 2	426	8 (1.9%)
Cluster 3	360	12 (3.3%)
Cluster 4	272	14 (5.1%)
Cluster 5	271	2 (0.7%)
Cluster 6	269	252 (93.7%)
Cluster 7	260	1 (0.4%)
Cluster 8	218	1 (0.5%)
Cluster 9	128	1 (0.8%)

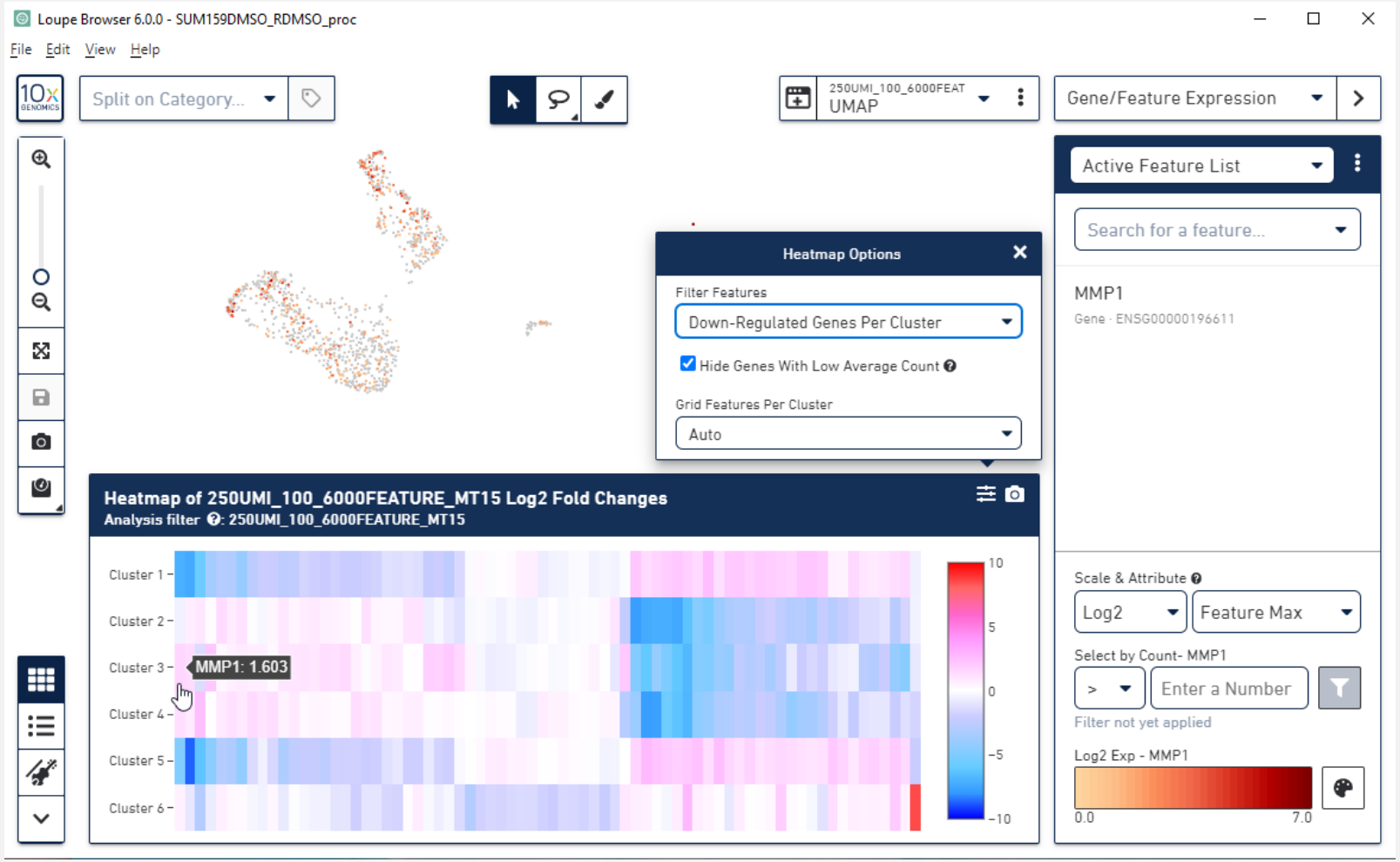


LOUPE: FILTERED DATASET

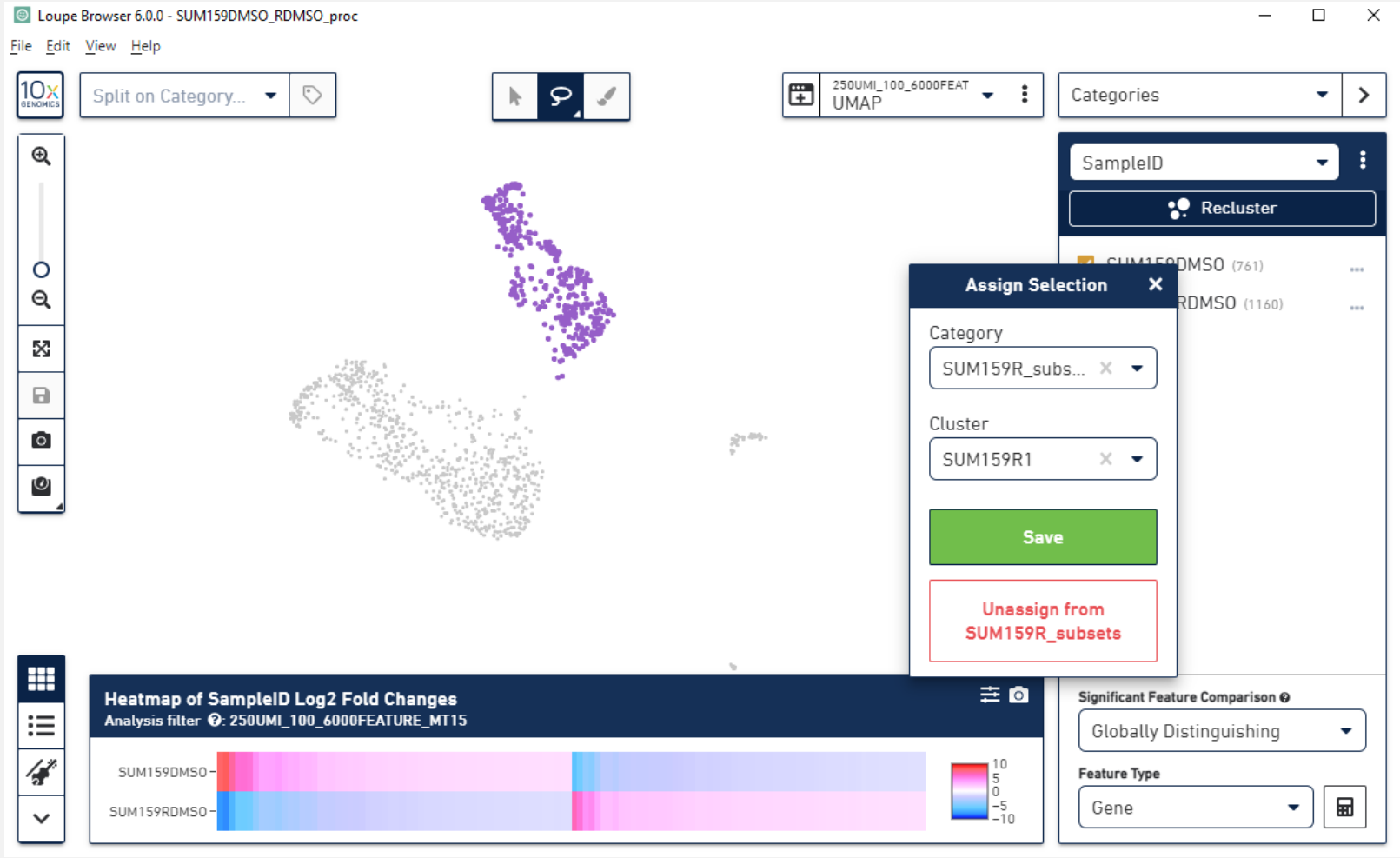


Find Markers

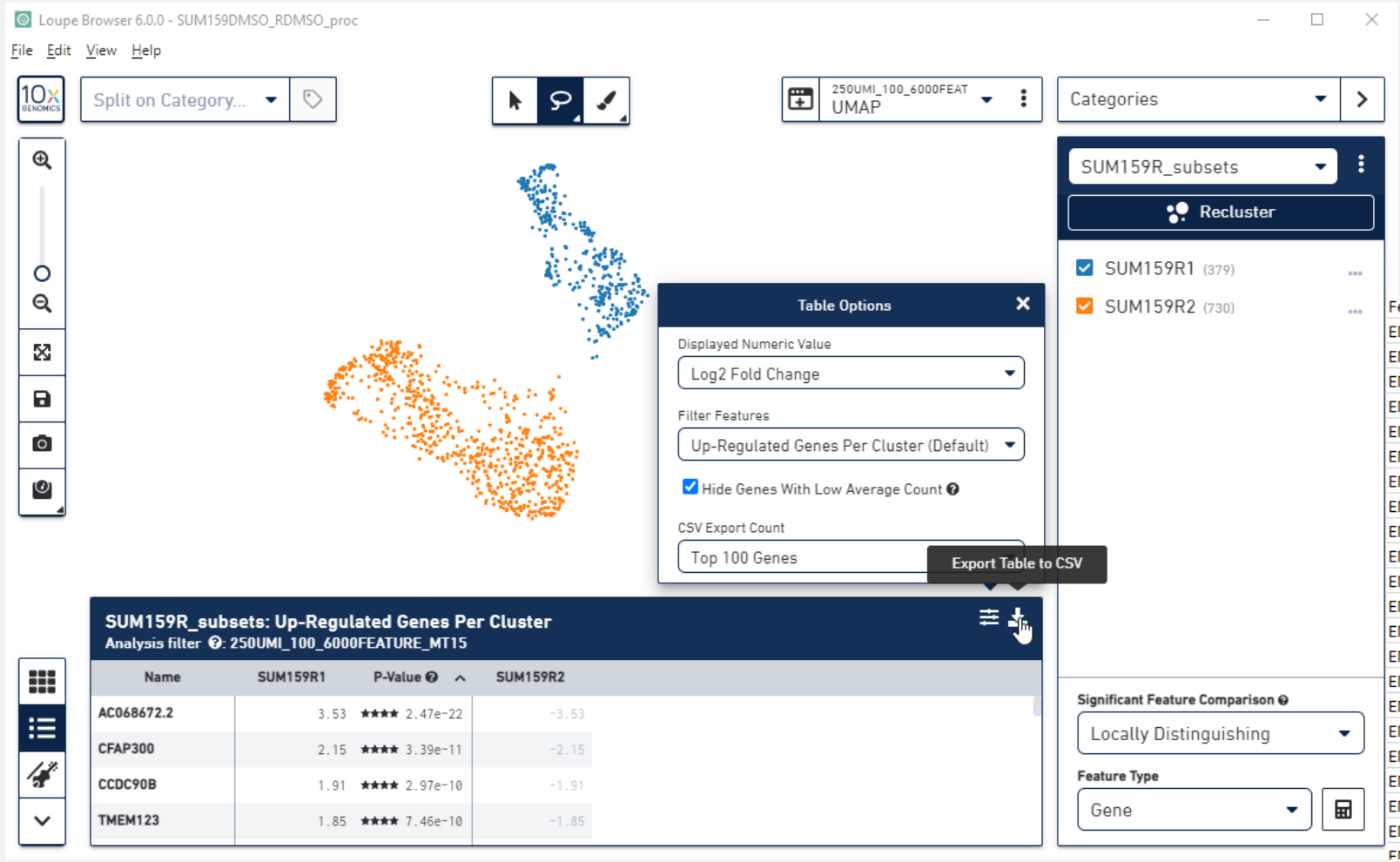
LOUPE: HEATMAP



LOUPE: CREATE SUB-CLUSTERING



LOUPE: DATA EXPORT

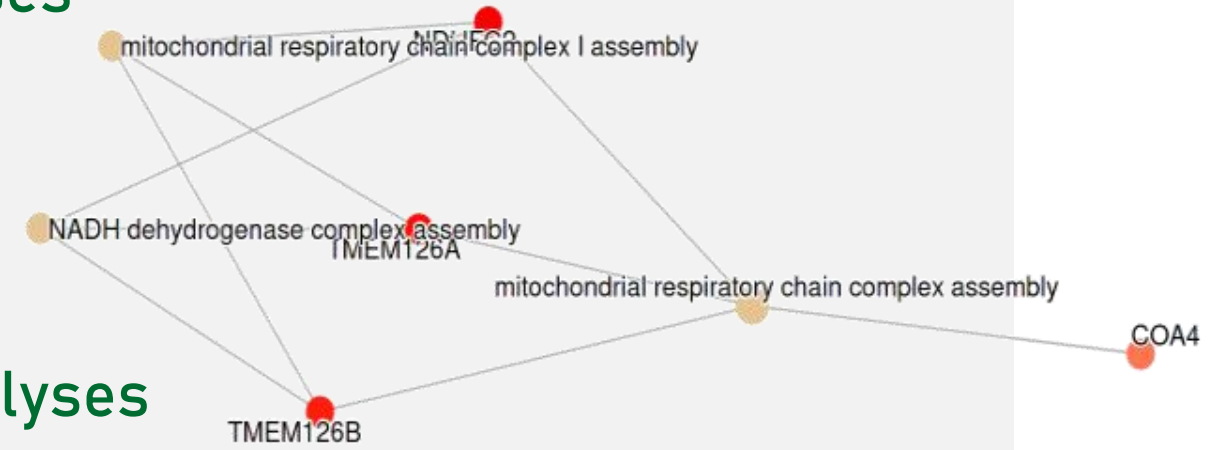


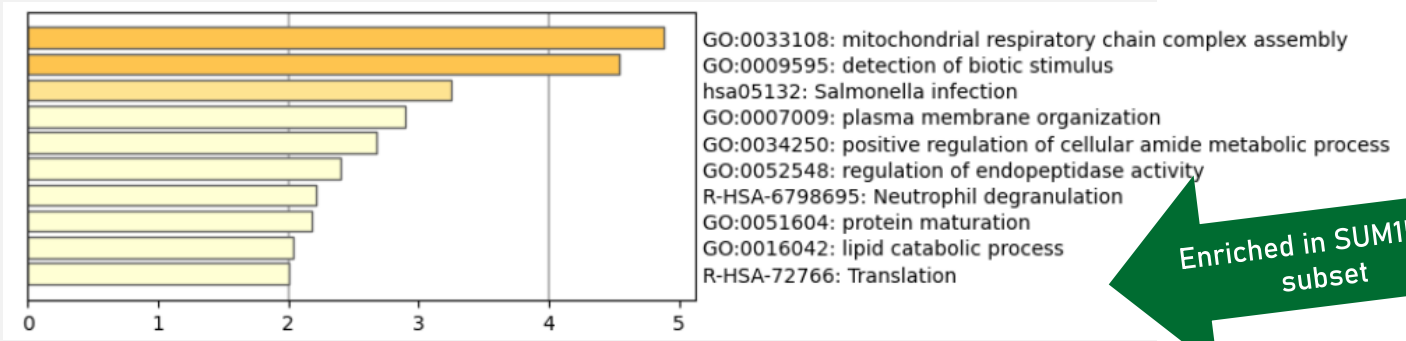
FeatureID	FeatureNa	SUM159R1 Average	SUM159R1 Log2 Fold Change	SUM159R1 P-Value
ENSG00000253377	AC068672.	1.349142878	3.530195793	2.47E-22
ENSG00000137691	CFAP300	1.402430388	2.150597426	3.39E-11
ENSG00000137500	CCDC90B	3.887566102	1.912991489	2.97E-10
ENSG00000152558	TMEM123	11.0402033	1.852589647	7.46E-10
ENSG00000246067	RAB30-DT	4.275111632	1.880286258	1.04E-09
ENSG00000225383	SFTA1P	1.005196219	2.098340267	1.31E-09
ENSG00000137509	PRCP	1.649490664	1.682944272	1.94E-07
ENSG00000223784	LINP1	1.361253676	1.759323228	2.36E-07
ENSG00000078124	ACER3	1.22319058	1.629331154	1.06E-06
ENSG00000137692	DCUN1D5	13.12810484	1.499185498	2.08E-06
ENSG00000151366	NDUFC2	22.81432094	1.482994504	2.87E-06
ENSG00000109321	AREG	1.717311132	1.700556693	3.07E-06
ENSG00000171202	TMEM126	6.		
ENSG00000087884	AAMDC			
ENSG00000166435	XRRA1	1.		
ENSG00000171204	TMEM126	5.		
ENSG00000137693	YAP1	1.		
ENSG00000149273	RPS3			
ENSG00000137492	THAP12			
ENSG00000159063	ALG8	1.		
ENSG00000245694	CRNDE	2.		
ENSG00000149196	HIKESHI	5.		

45	ENSG0000	PGM2L1	1.634958	1.004291	0.01039
46	ENSG0000	VMP1	2.315585	0.9686	0.013977
47	ENSG0000	BEX1	1.2789	1.036113	0.017087
48	ENSG0000	WDR72	2.797594	0.950319	0.017254
49	ENSG0000	JMJD8	1.215924	0.963404	0.018445
50	ENSG0000	TMEM134	1.569559	0.906085	0.029937
51	ENSG0000	PAFAH1B3	1.254679	0.879472	0.04437
52	ENSG0000	ANAPC15	3.875455	0.840493	0.045953
53	ENSG0000	DKK1	1.840841	0.861719	0.063495

LOUPE: SET ENRICHMENT

- Export data from Loupe
  - Warning! Contains also insignificantly regulated genes!
- Run R-based enrichment analyses
  - Gene Ontology
  - Reactome
  - Etc.
- Run web-based enrichment analyses
  - Filter genes for significance before uploading!

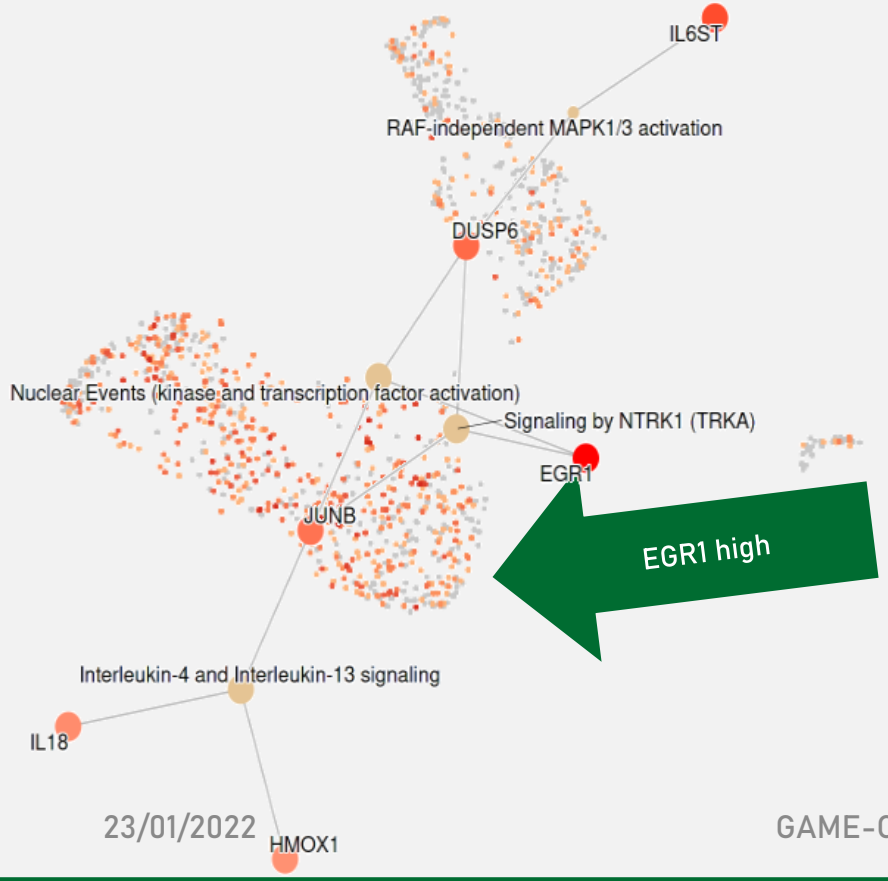


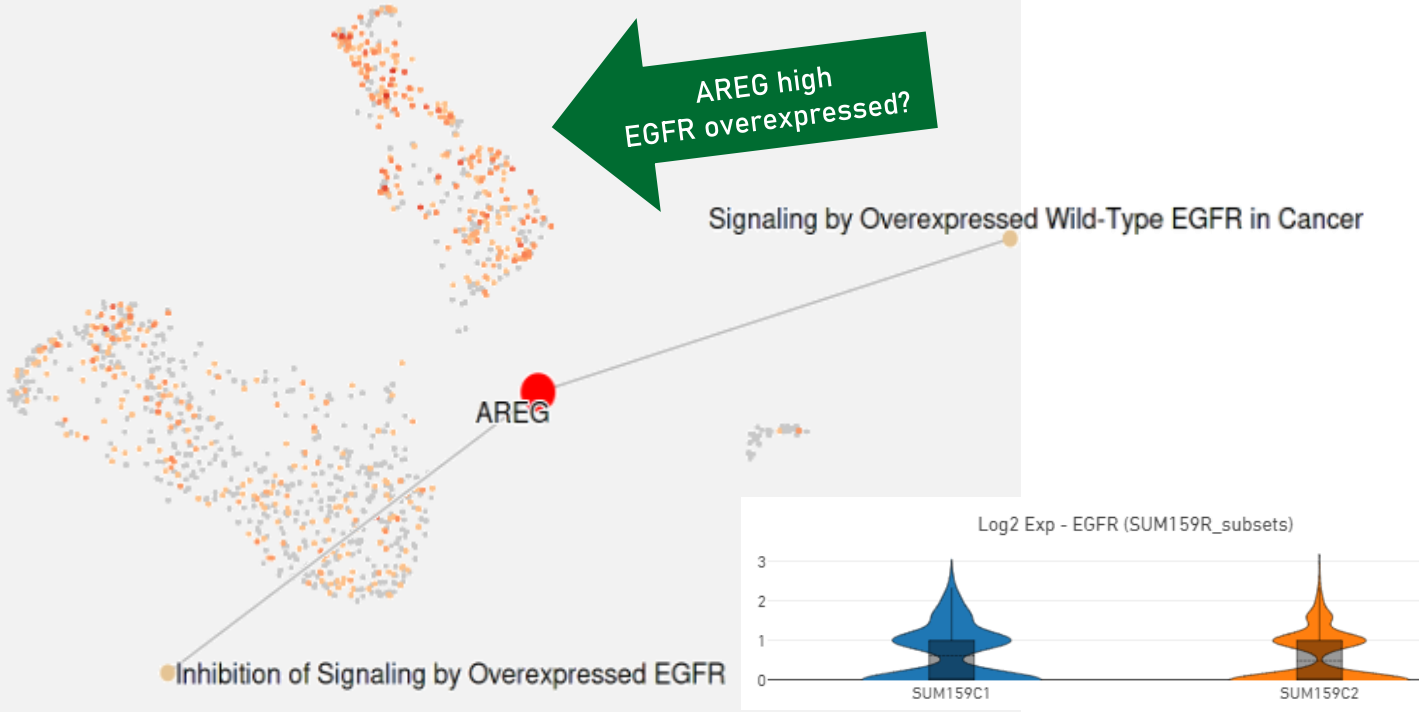


Enriched in SUM159C1 subset

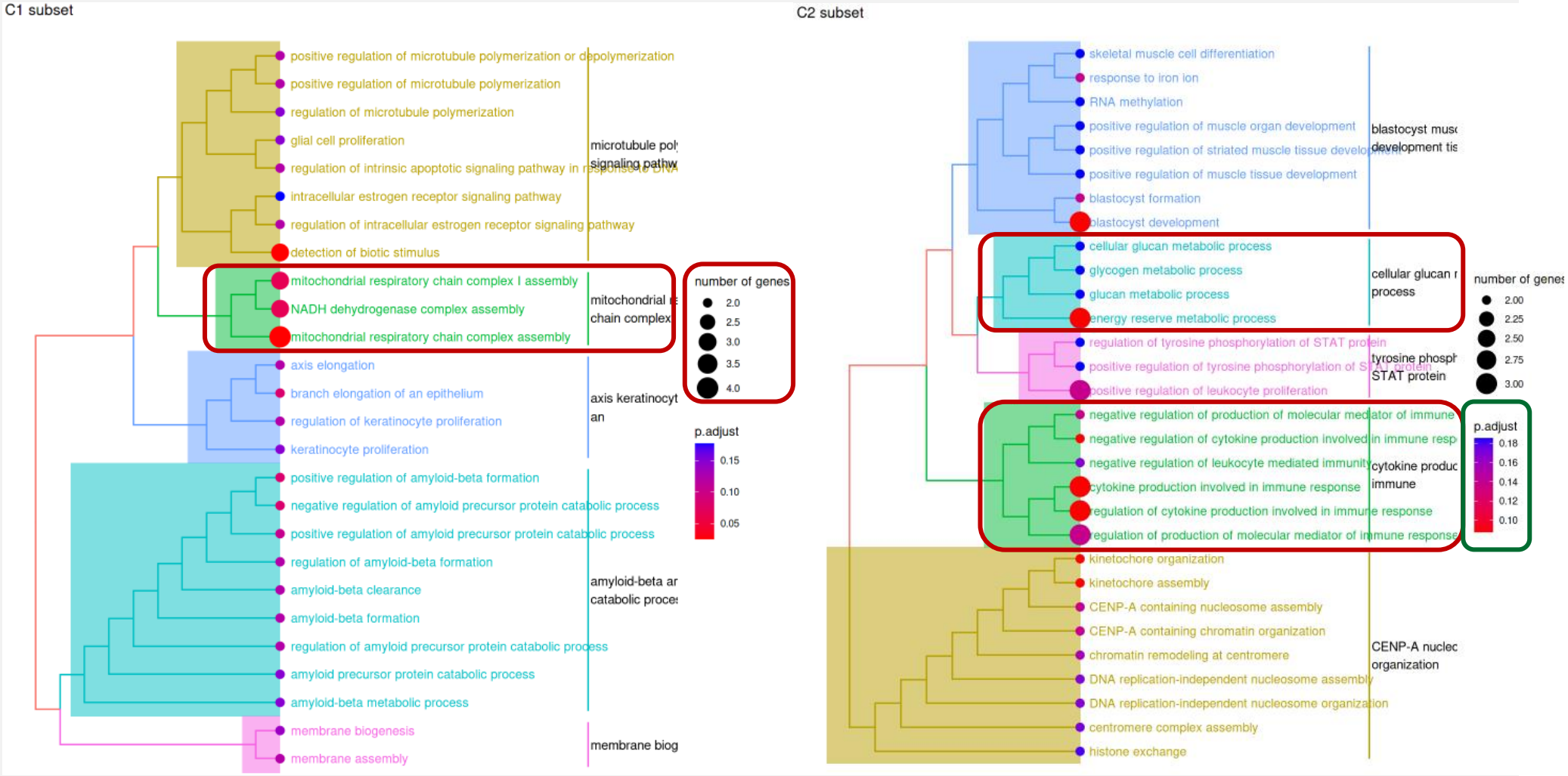
SET ENRICHMENT RESULTS: COMPARING C1+C2

- In C2: EGR1 transcription factor increased
- In C1: AREG increased which is relevant to EGFR/HER1





SET ENRICHMENT RESULTS: COMPARING C1+C2



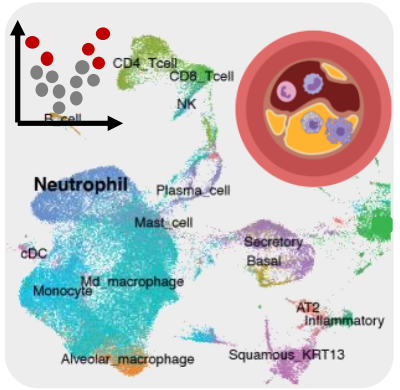
SUMMARY

scRNA-seq analysis

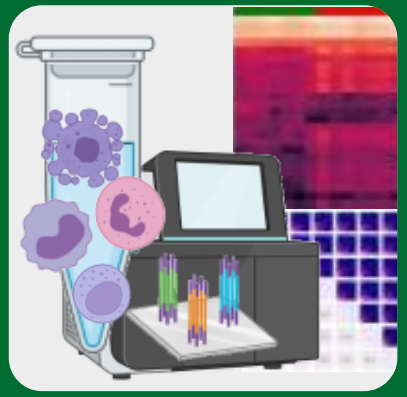
- Perform filtering and clustering of scRNA-seq data from within Loupe

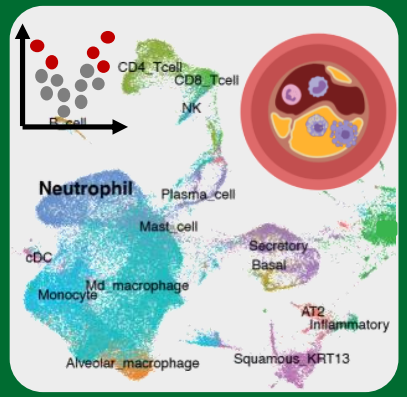
- Analyse scRNA-seq data with Loupe
  - Create new clusters/categories
  - Watch for specific genes
  - Perform Differential Expression analysis

- Perform Gene Set Enrichment Analyses of significant genes
  - R-based: clusterProfiler, DOSE, disgenet2r, …
  - Web-based: Metascape



scRNA-seq

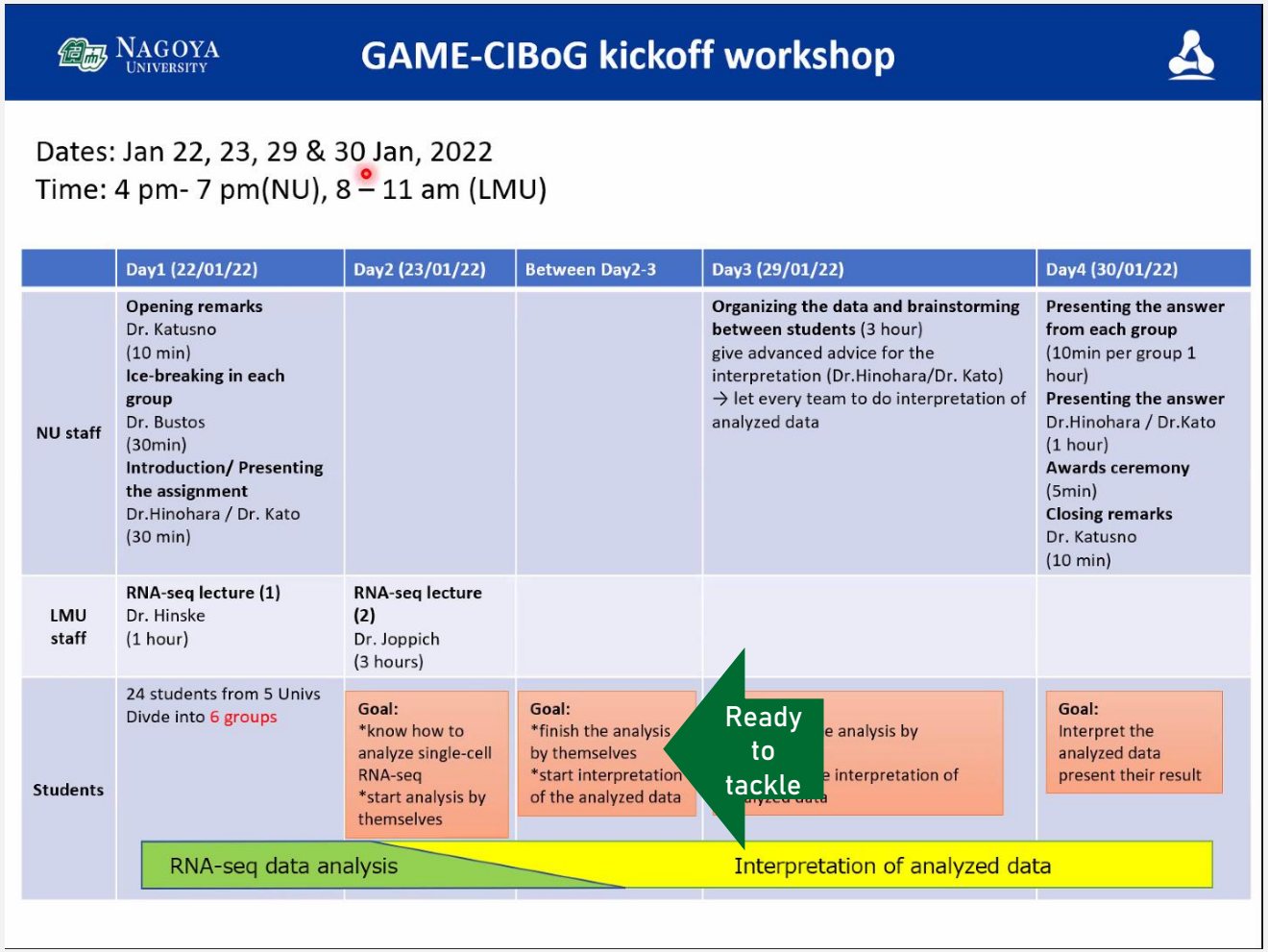




bulk RNA-seq

scRNA-seq

TIMELINE



NAGOYA UNIVERSITY

GAME-CIBoG kickoff workshop

Dates: Jan 22, 23, 29 & 30 Jan, 2022
Time: 4 pm- 7 pm(NU), 8 – 11 am (LMU)

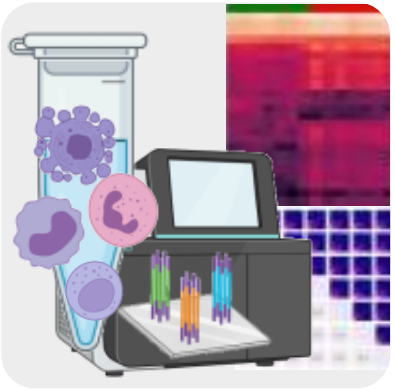
	Day1 (22/01/22)	Day2 (23/01/22)	Between Day2-3	Day3 (29/01/22)	Day4 (30/01/22)
NU staff	**Opening remarks** Dr. Katusno (10 min) **Ice-breaking in each group** Dr. Bustos (30min) **Introduction/ Presenting the assignment** Dr.Hinohara / Dr. Kato (30 min)			**Organizing the data and brainstorming between students** (3 hour) give advanced advice for the interpretation (Dr.Hinohara/Dr. Kato) → let every team to do interpretation of analyzed data	**Presenting the answer from each group** (10min per group 1 hour) **Presenting the answer** Dr.Hinohara / Dr.Kato (1 hour) **Awards ceremony** (5min) **Closing remarks** Dr. Katusno (10 min)
LMU staff	**RNA-seq lecture (1)** Dr. Hinske (1 hour)	**RNA-seq lecture (2)** Dr. Joppich (3 hours)			
Students	24 students from 5 Univs Divde into 6 groups	**Goal:** *know how to analyze single-cell RNA-seq *start analysis by themselves	**Goal:** *finish the analysis by themselves *start interpretation of the analyzed data	e analysis by e interpretation of	**Goal:** Interpret the analyzed data present their result

Ready to tackle

RNA-seq data analysis → Interpretation of analyzed data

SUMMARY

- Analyse public data from raw data (reads)
  - Or from pre-processed data
- Perform differential expression analysis using DESeq2
- Perform Gene Set Enrichment Analyses of significant genes
  - R-based: clusterProfiler, DOSE, disgenet2r, …
  - Web-based: Metascape
- At each “pipeline” step the choice of tools, methods and parameters affects the final result
  - Sorting of lists, number of genes for set enrichment, etc.

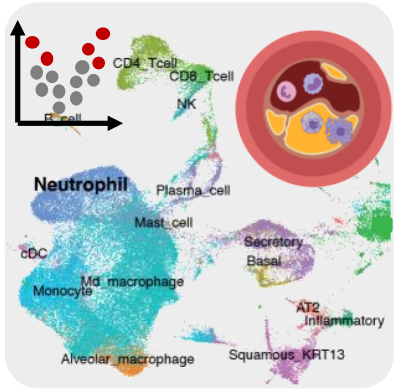


bulk RNA-seq

SUMMARY

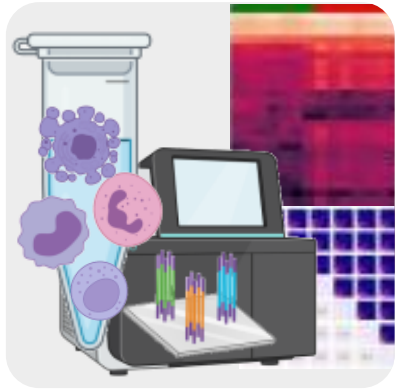
scRNA-seq analysis

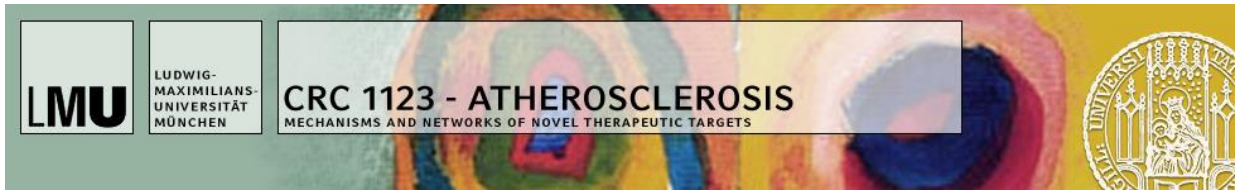
- Analyse public data from raw data (reads)
- Run `cellranger count` and `aggregate` pipelines
- Perform sub-clustering of scRNA-seq data from within Loupe
  - Cell filtering and
  - Differential Expression analysis
- Perform Gene Set Enrichment Analyses of significant genes
  - R-based: clusterProfiler, DOSE, disgenet2r, …
  - Web-based: Metascape

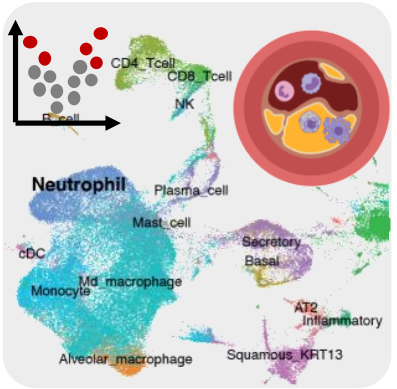


scRNA-seq

THANK YOU







FOR YOUR ATTENTION