Single Cell RNA-seq DE Analysis

Xin-Qiao Zhang Ph.D

Dec 11, 2020

Part I Background

Part II Public Resources

Part III Example 1

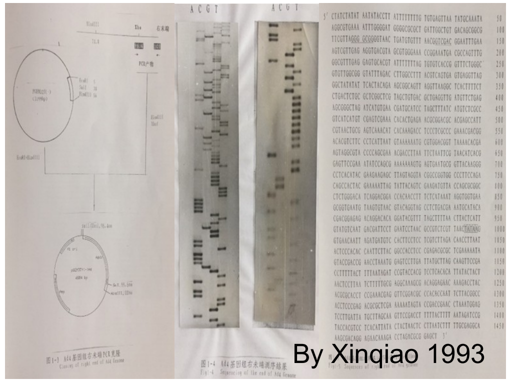
Part IV Example 2

Part I Sequencing Background

Sanger sequencing

Tips

- ➢ In vitro DNA replication
- ➢ 1977 developed, 1986 commercialized
- ➢ Selective incorporation: chain-terminator

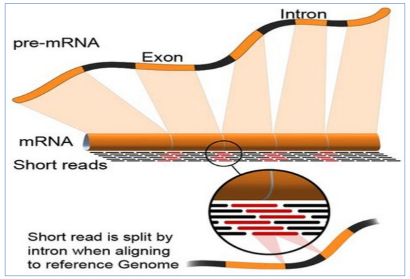


By Xinqiao 1993

Microarray

Tips

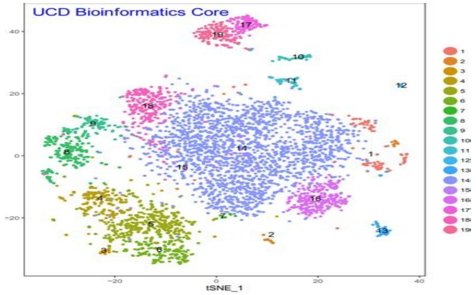
- ➢ Complementary probe hybridization
- ➢ Non radioactive isotope
- ➢ Fragment sequencing


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Next Generation Sequence. RNA-seq

Tips

- ➢ Whole picture of gene expression
- ➢ Splicing, transcript isoform
- ➢ Fusion detection, mutation discorvery
- ➢ For DE: one cell line as one sample, divide samples into two or more groups



scRNA-seq

Tips

- ➢ Identify cell population
- ➢ Uncover novel cell type, cell status, rear cell
- ➢ Discover new marker, gene signatures
- ➢ Profiling heathy and diseased tissues

scRNA-seq

1. Measures the **distribution of expression levels**: each gene across a population of cells

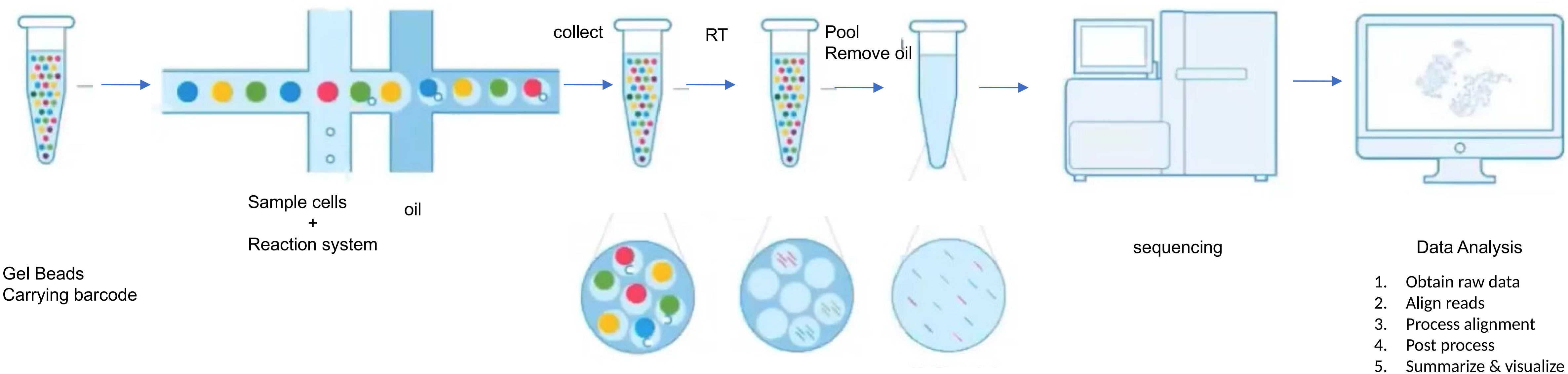
2. Commercial platform: **10x Genomics Chromium**, Fluidigm C1 and Watergen ICELL8

3. Advantage

   - **Cell type specific gene expression pattern**
   - Easy to remove duplicate,
   - Characterize and identify heterogeneous cell population
   - **Discover new cell markers & regulatory pathways**
   - Uncover novel cell types, cell status and rare cell types
   - Study cell-specific changes
   - Comparing distribution

scRNAseq

10xGenomics: Chromium Single Cell Gene Expression: Workflow



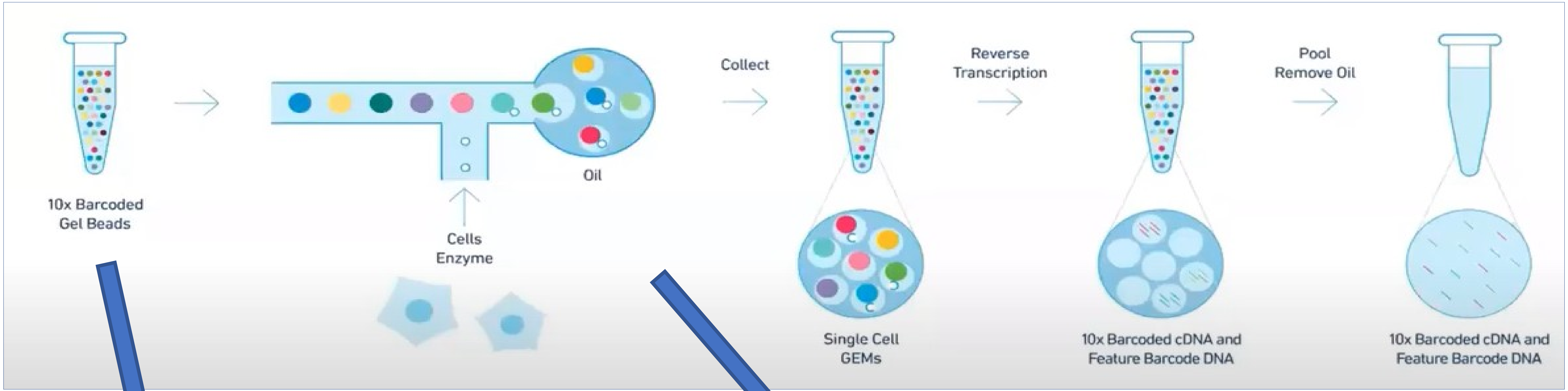
1. Obtain raw data
2. Align reads
3. Process alignment
4. Post process
5. Summarize & visualize

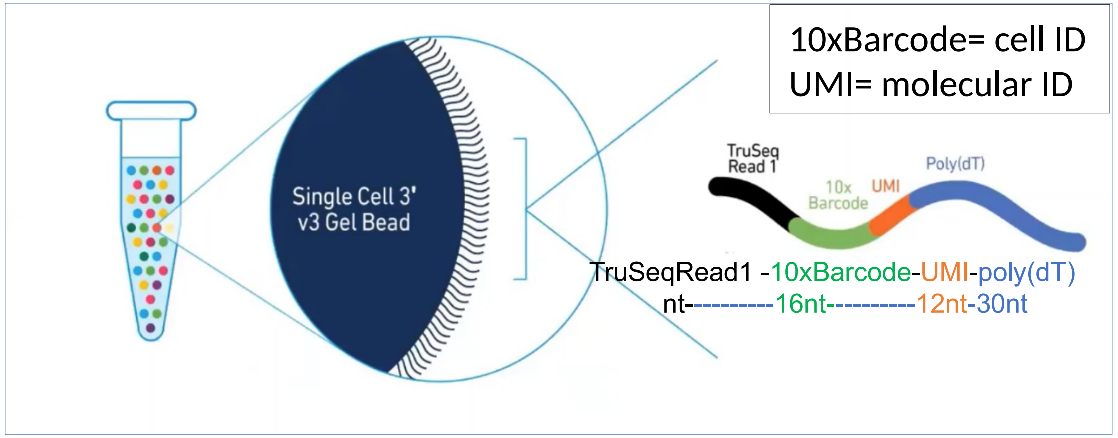
GEMs (Gel Beads-in-Emulsion)

From 10xGenomics

scRNAseq

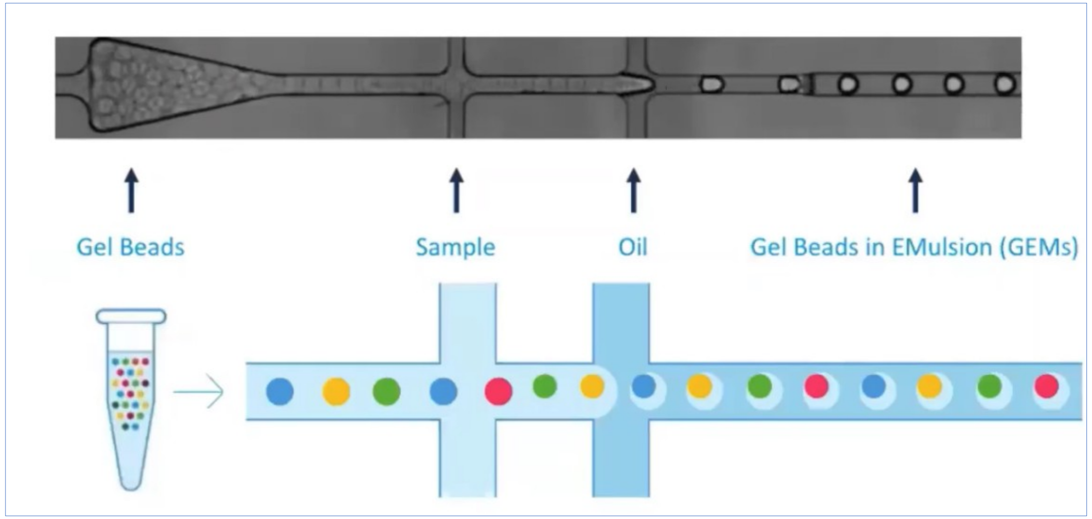
10xGenomics: Chromium Single Cell Gene Expression: Workflow





10xBarcode= cell ID
UMI= molecular ID

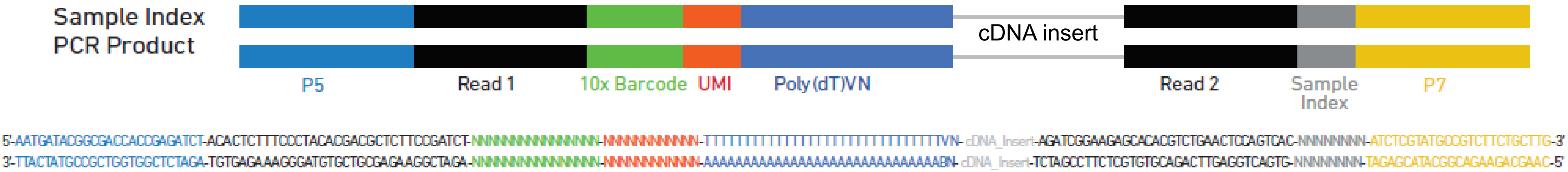
TruSeqRead1 -10xBarcode-UMI-poly(dT)
nt----------16nt----------12nt-30nt



From 10xGenomics

scRNAseq

Single Cell 3’ Gene Expression Library


Sample Index
PCR Product
cDNA insert
P5
Read 1
10x Barcode
UMI
Poly(dT)VN
Read 2
Sample Index
P7
5'-AATGATACGGCGACCACCGAGATCT-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNNNNNNNNNNNNN-NNNNNNNNNNNN-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN- cDNA_Insert-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNNNN-ATCTCGTATGCCGTCTTCTGCTTG-3'
3'-TTACTATGCCGCTGGTGGCTCTAGA-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA-NNNNNNNNNNNNNNNN-NNNNNNNNNNNN-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAABN- cDNA_Insert-TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG-NNNNNNNN-TAGAGCATACGGCAGAAGACGAAC-5'

UMI (unique molecular identifier): molecular barcode, add during RT before PCR

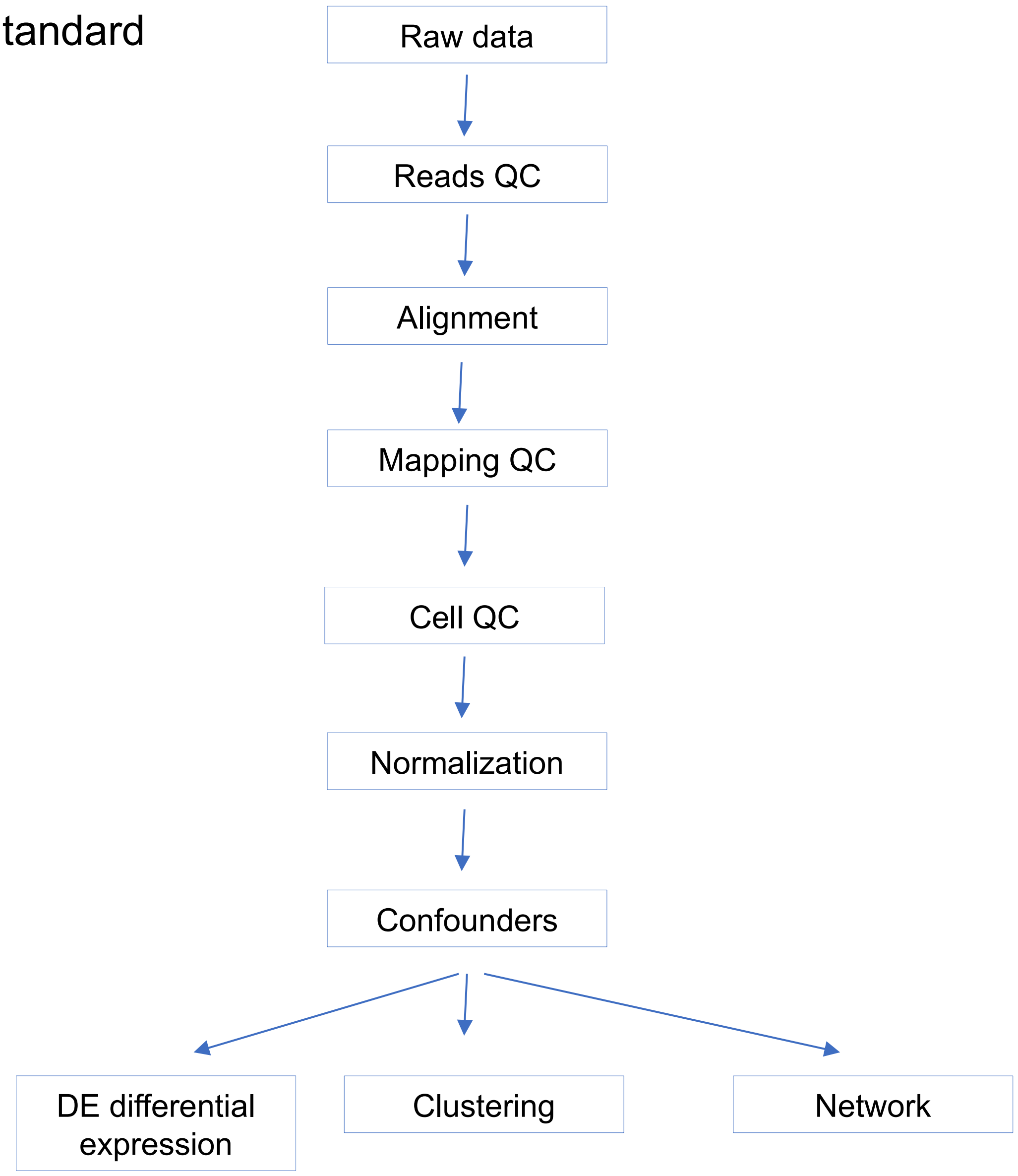
10xBarcode: cell barcode, add during RT before PCR

	Read 1	i7 index	i5 index	Read 2
Purpose	Barcode & UMI	sample Index	n/a	Transcript
Length	28	8	0	91

scRNAseq

scRNAseq Analysis Workflow

Standard



Raw data

Reads QC

Alignment

Mapping QC

Cell QC

Normalization

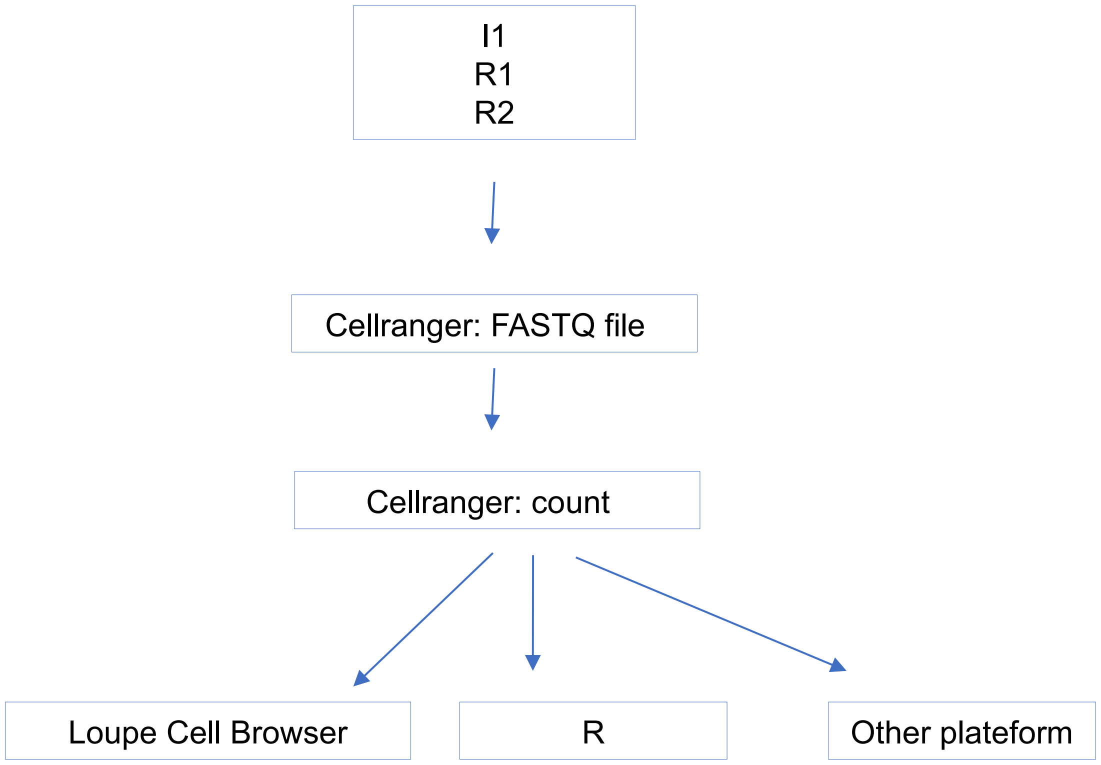
Confounders

DE differential expression

Clustering

Network

10xGenomics



I1
R1
R2

Cellranger: FASTQ file

Cellranger: count

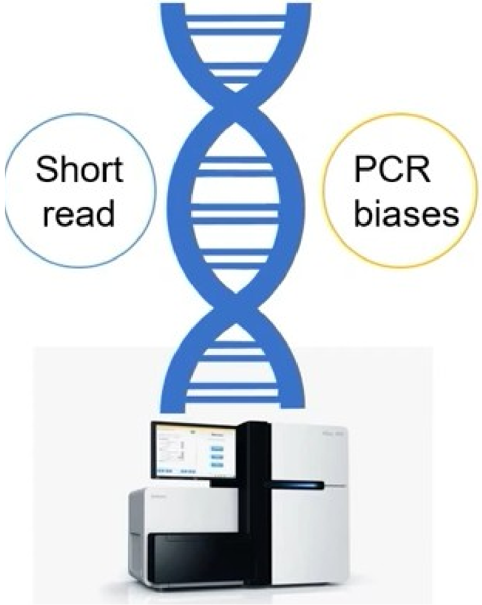
Loupe Cell Browser

R

Other plateform

scRNAseq Issues

1. Remind: garbage in, garbage out
2. Sample: purity, quantity, quality
3. Revers transcription efficiency: < 30%
4. Exons: separated by large introns
5. Gene ‘dropout’:
   - ➢ Low starting amount: since RNA from one cell
   - ➢ Technical factor
   - ➢ Observed zero values:
6. mRNA relative abundance vary wildly
   - ➢ 10e5-10e7 orders of magnitude
   - ➢ Highly expressed genes consumes the majority reads
7. mRNA comes in a wide range of sizes
   - ➢ Small RNAs need be captured separately?
   - ➢ PolyA selection of large RNAs may results in 3’ end bias
8. PCR bias
9. Unwanted variability introduced by batch effects



Single Cell RNA-seq Quality Control

- ➢ Mitochondrial fragment for cell status
- ➢ Library size: total number of reads counts
- ➢ Detected genes
- ➢ ERCCs (external RNA control consortium) and MTs amount
- ➢ Gene QC

RNA-seq

1. Measures the **average expression** level for each gene across a large population of cells

2. Advantage

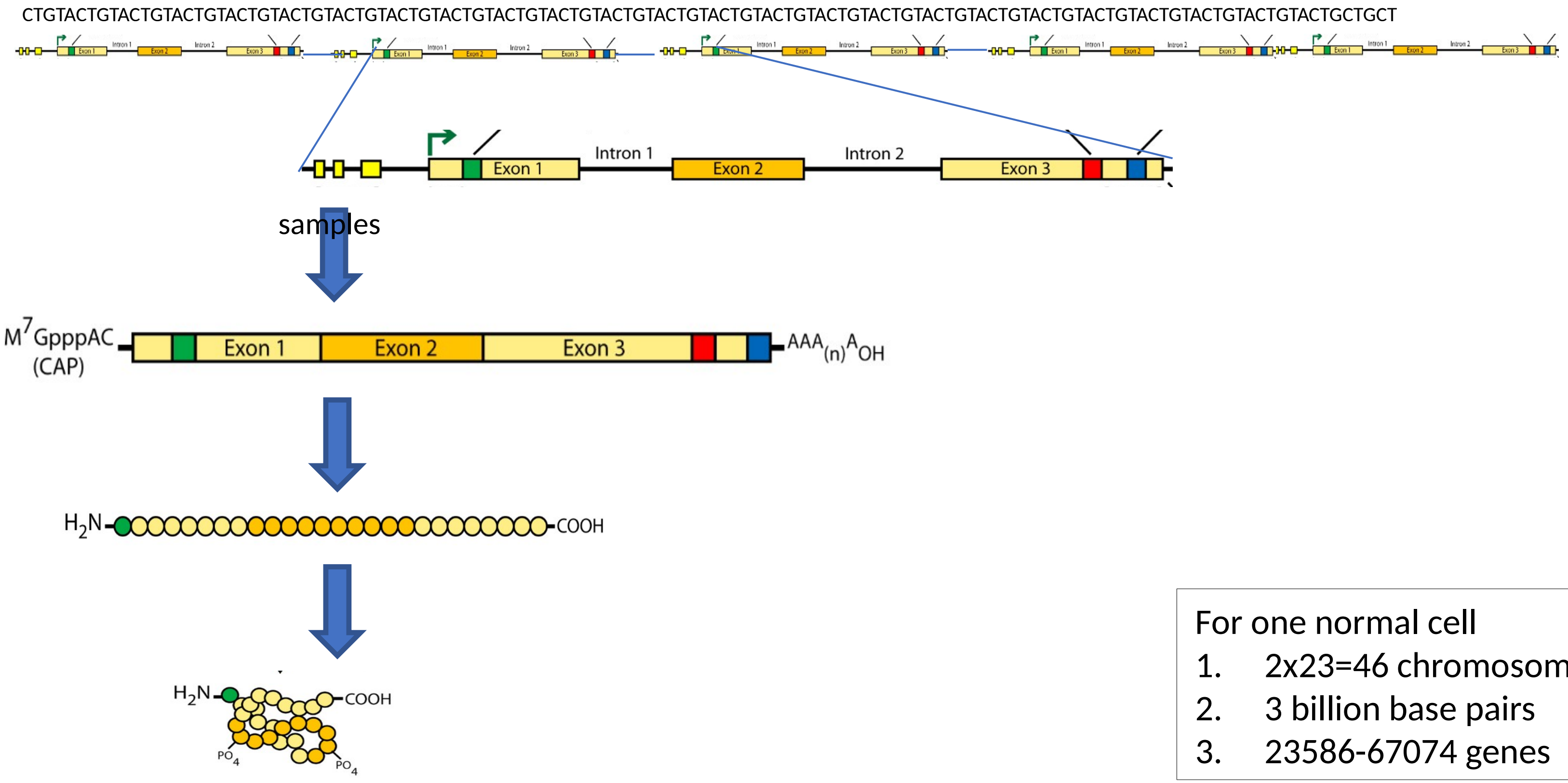
   - **Useful for comparing differential expression**
   - Interpreting mutation
   - Prioritizing protein coding mutation: if no expression, not interesting,
   - Heterozygous mutation expression: wild type allele, lost function; mutant allele, a candidate drug target
   - Useful for quantifying expression signature

3. Disadvantage:

   - Does not provide insights

RNA-seq

Central Dogma of Molecular Biology: concept still important

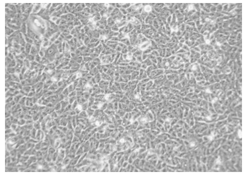

CTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGTACTGCTGCT
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
samples
M7GpppAC
(CAP)
Exon 1
Exon 2
Exon 3
AAA(n)A OH
H2N
COOH
H2N
COOH
PO4
PO4

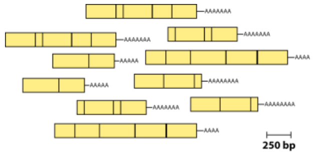
For one normal cell

1. 2x23=46 chromosomes
2. 3 billion base pairs
3. 23586-67074 genes

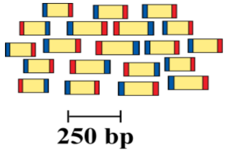
RNA-seq

RNA-seq Methods

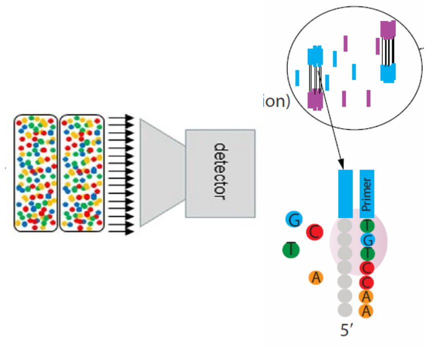




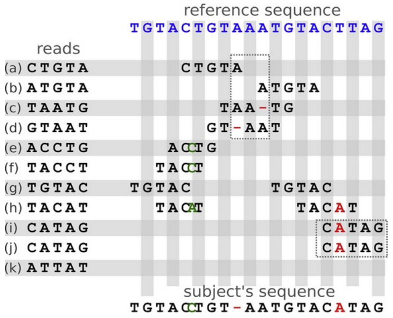
Extract mRNA

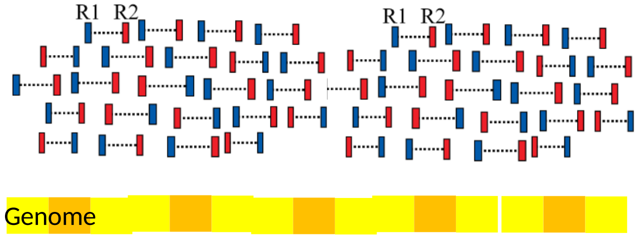


Generate cDNA, fragment,
Size select, add adapters



Sequencing (Illumina HiSeq2000)





Alignment reads,

RNA-seq analysis

1. Alignment and QC
2. Count read for each gene
3. Differential expression (DE) analysis: limma, edgeR, DESeq2
4. Further functional validation: pathway,

Part II Public Resources

1. **scRNA-seq and RNA-seq data from public resource:**
   1) **SRA**: **sequence read archive**, raw sequence data
   2) CCLE:(https://portals.broadinstitute.org/ccle) RNAseq, Expression, fusion….
   3) ExpressionAtlas: EBI, Homo sapiens 1449 experiment, (https://www.ebi.ac.uk/gxa/home)
   4) **GEO dataset/Profiles**: processed data, (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/)
   5) GTEx (https://www.gtexportal.org/home/) less cancer cell lines, mainly for normal cells
   6) COSMIC: (http://cancer.sanger.ac.uk/cell_lines/sample/overview?id=687452) easy search
   7) CELLX (http://cellx.sourceforge.net) not search easily for beginner
   8) BioGPS (http://biogps.org/dataset/)
2. **Drug IC50 from public resource and published paper**
   1) GDSC: IC50 of 518 drug IC50 on 988x cancer cell lines (http://www.cancerrxgene.org/)
   2) PharmacoDB: combined CCLE, GDSC1000, gCSI, GRAY, FIMM, CTRPv2 and UHNBrease
   3) CTRPv2: 481 compounds X 860 cancer cell lines
3. **Public software packages**
   ➢ Bioconductor: containing over 1903 software packages
   ➢ R and RStudio, Python
4. Public reference genome resource: (ENSEMBL) **and gene annotation (gtf, gff)**
5. **Cloud server: Galaxy (also low cost cloud server: AWS-EC2, S3)**

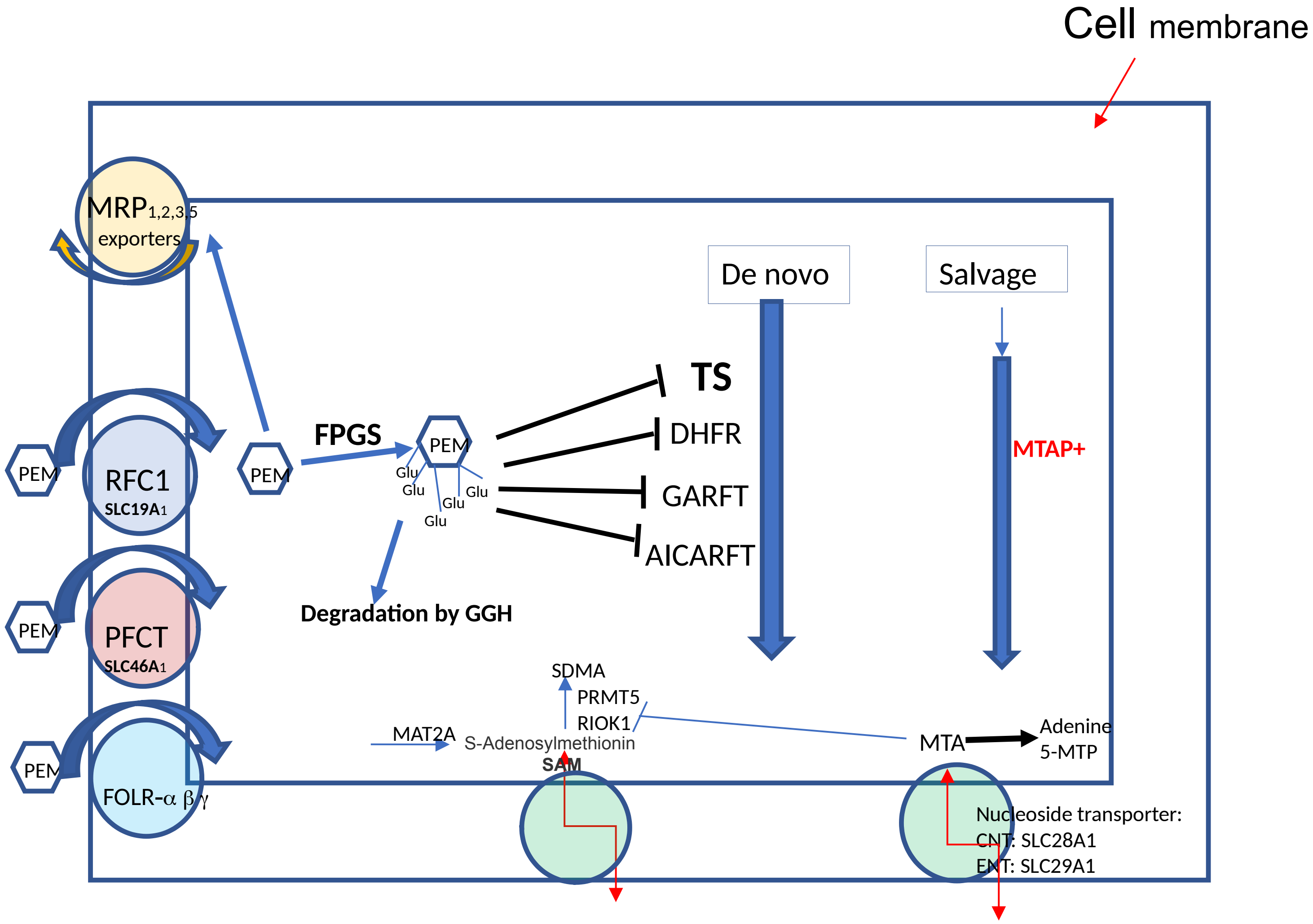
Part III Example 1

RNA-seq Predicting PEM Sensitivity on Bladder Cancer Cells

Example 1: RNA-seq Analysis

Pemetrexed (PEM)

- **PEM**: 1st line drug for NSCLC,
  2nd line drug for bladder cancer
- **MTAP**: S-methyl-5'-thioadenosine phosphorylase
- **Looking for PEM sensitive/resistant gene**
- **One gene vs Gene Signature**

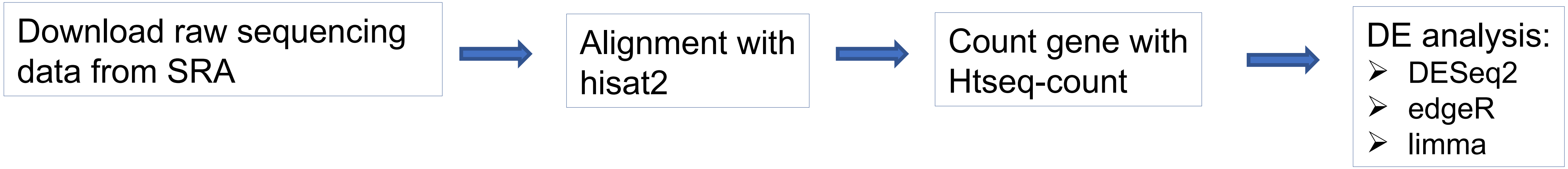


Analyze RNA-seq Data from Public Resource for Predicting Drug Sensitivity

1. **Get RNA-seq raw data from public resource:**
   1) **SRA**: **sequence read archive**, raw sequence data
   2) CCLE:(https://portals.broadinstitute.org/ccle) RNAseq, Expression, fusion….
   3) ExpressionAtlas: EBI, Homo sapiens 1449 experiment, (https://www.ebi.ac.uk/gxa/home)
   4) GEO dataset/Profiles: processed data, (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/)
   5) GTEx (https://www.gtexportal.org/home/) less cancer cell lines, mainly for normal cells
   6) COSMIC: (http://cancer.sanger.ac.uk/cell_lines/sample/overview?id=687452) easy search
   7) CELLX (http://cellx.sourceforge.net) not search easily for beginner
   8) BioGPS (http://biogps.org/dataset/)
2. **Get Drug IC50 from public resource and published paper**
   1) GDSC: IC50 of 518 drug IC50 on 988x cancer cell lines (http://www.cancerrxgene.org/)
   2) PharmacoDB: combined CCLE, GDSC1000, gCSI, GRAY, FIMM, CTRPv2 and UHNBrease
   3) CTRPv2: 481 compounds X 860 cancer cell lines
3. **Public software packages for RNA-seq analysis**
   - ➢ Bioconductor: containing over 1903 software packages
   - ➢ R and RStudio, Python
4. **Public reference genome resource (ENSEMBL) and gene annotation (gtf, gff)**
5. **Local computer and cloud server (AWS-EC2, S3)**

Example 1: RNA-seq Analysis

My analysis workflow



Download raw sequencing data from SRA

Alignment with hisat2

Count gene with Htseq-count

DE analysis:

- ➢ DESeq2
- ➢ edgeR
- ➢ limma

SRAnumber	Cell line
SRR5445845	253JP
SRR5445848	HT1197
SRR5445849	HT1376
SRR5445850	J82
SRR5445851	RT112
SRR5445852	RT4
SRR5445856	T24
SRR5445868	UM-UC3

bam file

table file

dataset

Example 1: RNA-seq Analysis

Compare gene expression difference in bladder cancer cell lines

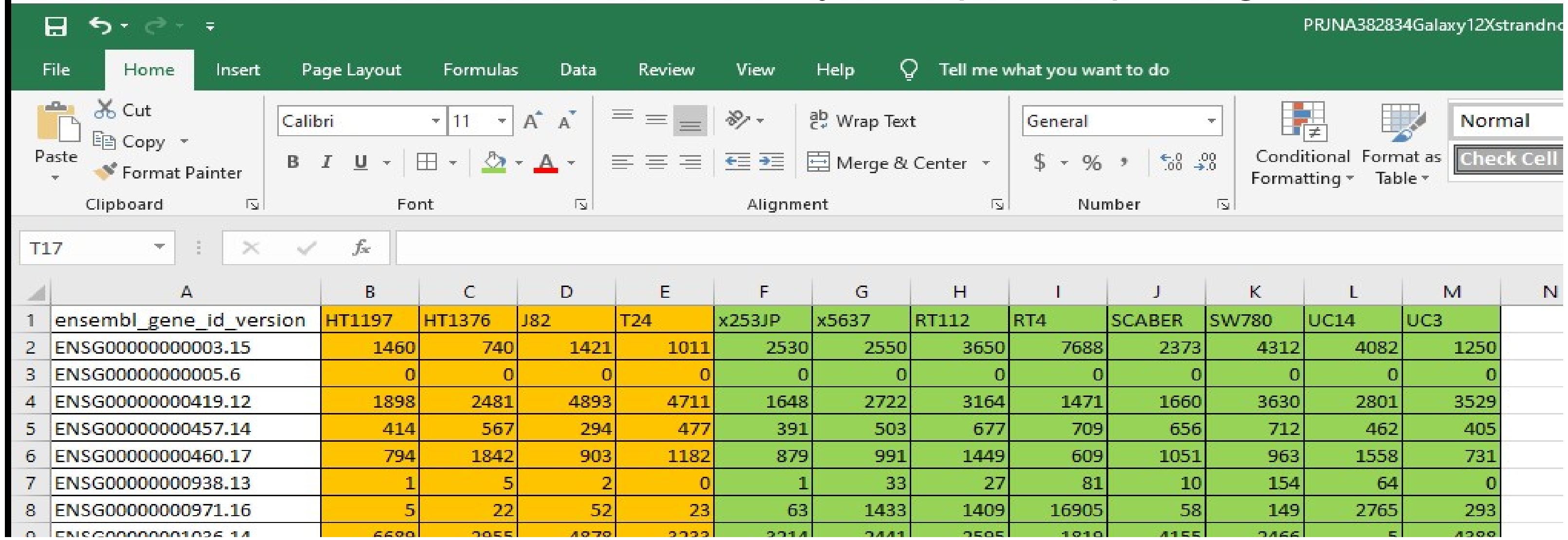
comparison	Cell line#		S	R
Comparison 1	8	R vs S	253, RT112, RT4, UC3	HT1197, HT1376, J82, T24

Interesting genes

1. FBN1, **HS6ST2**, AUTS2, CYP4F11, GPX2, PEA15
2. SLC7A6, NOVA1, OLFML3, MAP3K10, FABP4
3. MOXD1, FN1, FLNC, KRT34, PSG6, PHETA2, TNFSF12, CD99, C4BPB,
4. GALNT6, COL1A2, NLRP10, PSG2
5. IFI27, FILIP1L, MCAM, TGFB2, TIMP4, FBLN2, LINC00899,
6. **MTAP,** MYL9, COL7A1, F3, SECTM1, **CDKN2B, CDKN2A**, TMEM25, UCN2 PTP4A3

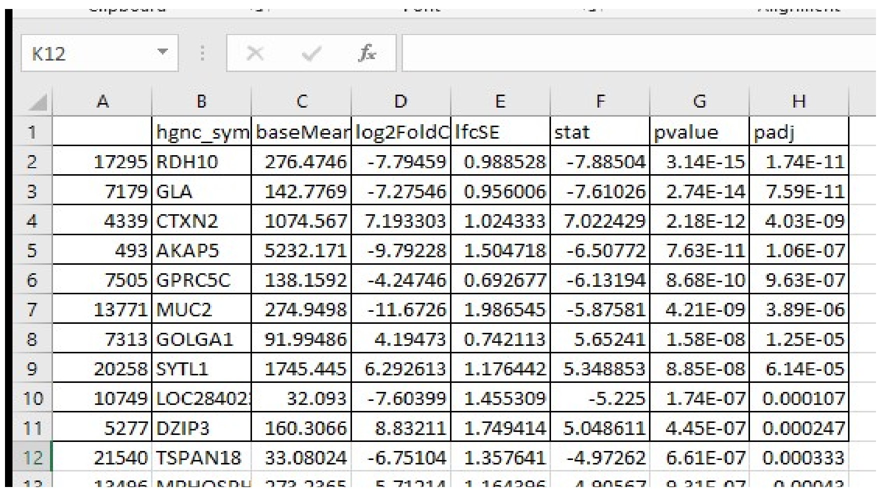
Example 1: RNA-seq Analysis

Gene count result by htseq-count package



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ensembl_gene_id_version	HT1197	HT1376	J82	T24	x253JP	x5637	RT112	RT4	SCABER	SW780	UC14	UC3	
2	ENSG00000000003.15	1460	740	1421	1011	2530	2550	3650	7688	2373	4312	4082	1250	
3	ENSG00000000005.6	0	0	0	0	0	0	0	0	0	0	0	0	
4	ENSG00000000419.12	1898	2481	4893	4711	1648	2722	3164	1471	1660	3630	2801	3529	
5	ENSG00000000457.14	414	567	294	477	391	503	677	709	656	712	462	405	
6	ENSG00000000460.17	794	1842	903	1182	879	991	1449	609	1051	963	1558	731	
7	ENSG00000000938.13	1	5	2	0	1	33	27	81	10	154	64	0	
8	ENSG00000000971.16	5	22	52	23	63	1433	1409	16905	58	149	2765	293	
9	ENSG00000001036.14	6689	2955	4878	3233	3214	2441	2595	1819	4155	2466	5	4388	

DE result by DESeq2 package

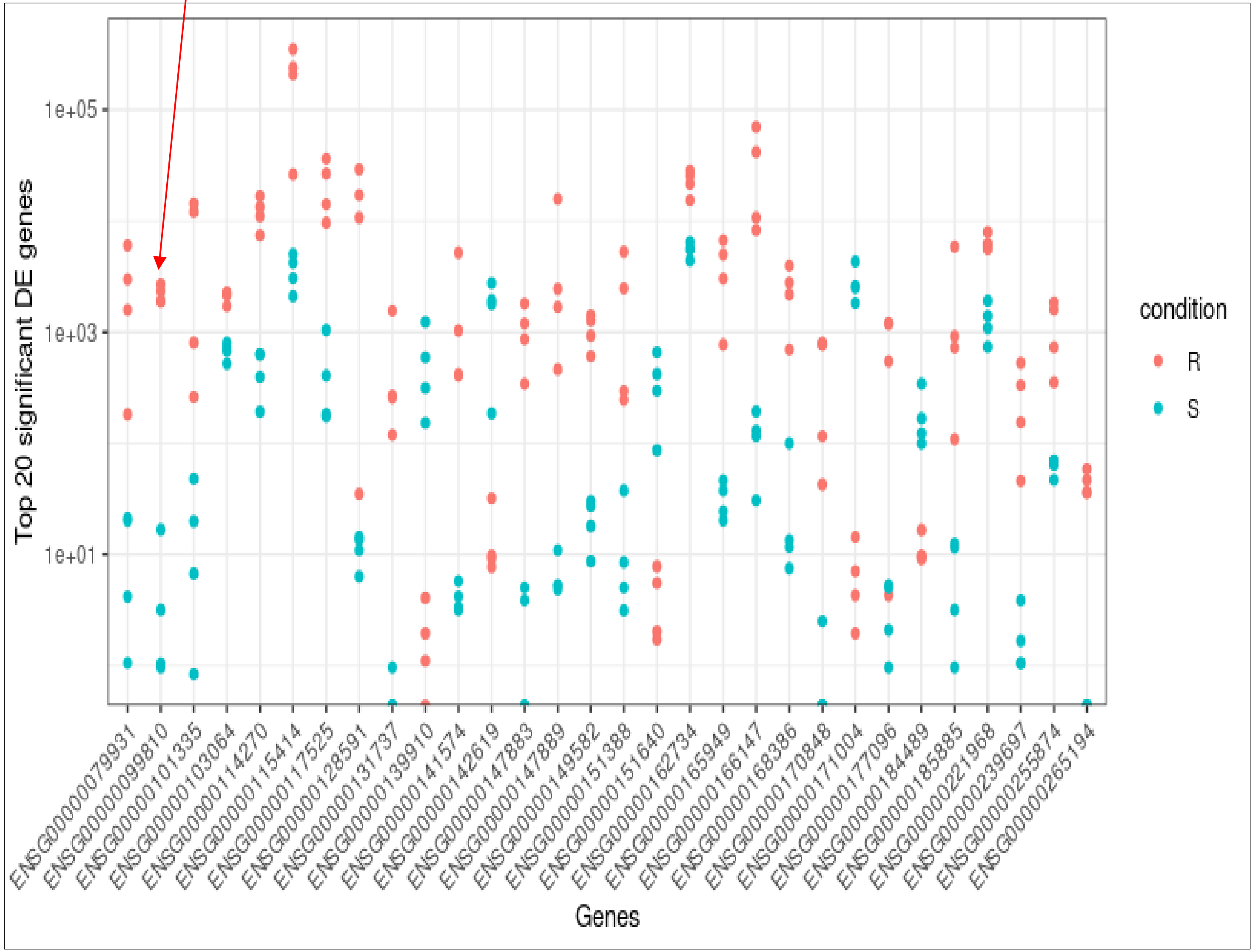


	A	B	C	D	E	F	G	H
1		hgnc_sym	baseMear	log2FoldC	lfcSE	stat	pvalue	padj
2	17295	RDH10	276.4746	-7.79459	0.988528	-7.88504	3.14E-15	1.74E-11
3	7179	GLA	142.7769	-7.27546	0.956006	-7.61026	2.74E-14	7.59E-11
4	4339	CTXN2	1074.567	7.193303	1.024333	7.022429	2.18E-12	4.03E-09
5	493	AKAP5	5232.171	-9.79228	1.504718	-6.50772	7.63E-11	1.06E-07
6	7505	GPRC5C	138.1592	-4.24746	0.692677	-6.13194	8.68E-10	9.63E-07
7	13771	MUC2	274.9498	-11.6726	1.986545	-5.87581	4.21E-09	3.89E-06
8	7313	GOLGA1	91.99486	4.19473	0.742113	5.65241	1.58E-08	1.25E-05
9	20258	SYTL1	1745.445	6.292613	1.176442	5.348853	8.85E-08	6.14E-05
10	10749	LOC28402	32.093	-7.60399	1.455309	-5.225	1.74E-07	0.000107
11	5277	DZIP3	160.3066	8.83211	1.749414	5.048611	4.45E-07	0.000247
12	21540	TSPAN18	33.08024	-6.75104	1.357641	-4.97262	6.61E-07	0.000333
13	13496	MPHOSPH	273.2265	5.71214	1.164296	4.90567	9.31E-07	0.00043

Example 1: RNA-seq Analysis

Resistant vs Sensitive: top 60 genes 1-30

MTAP



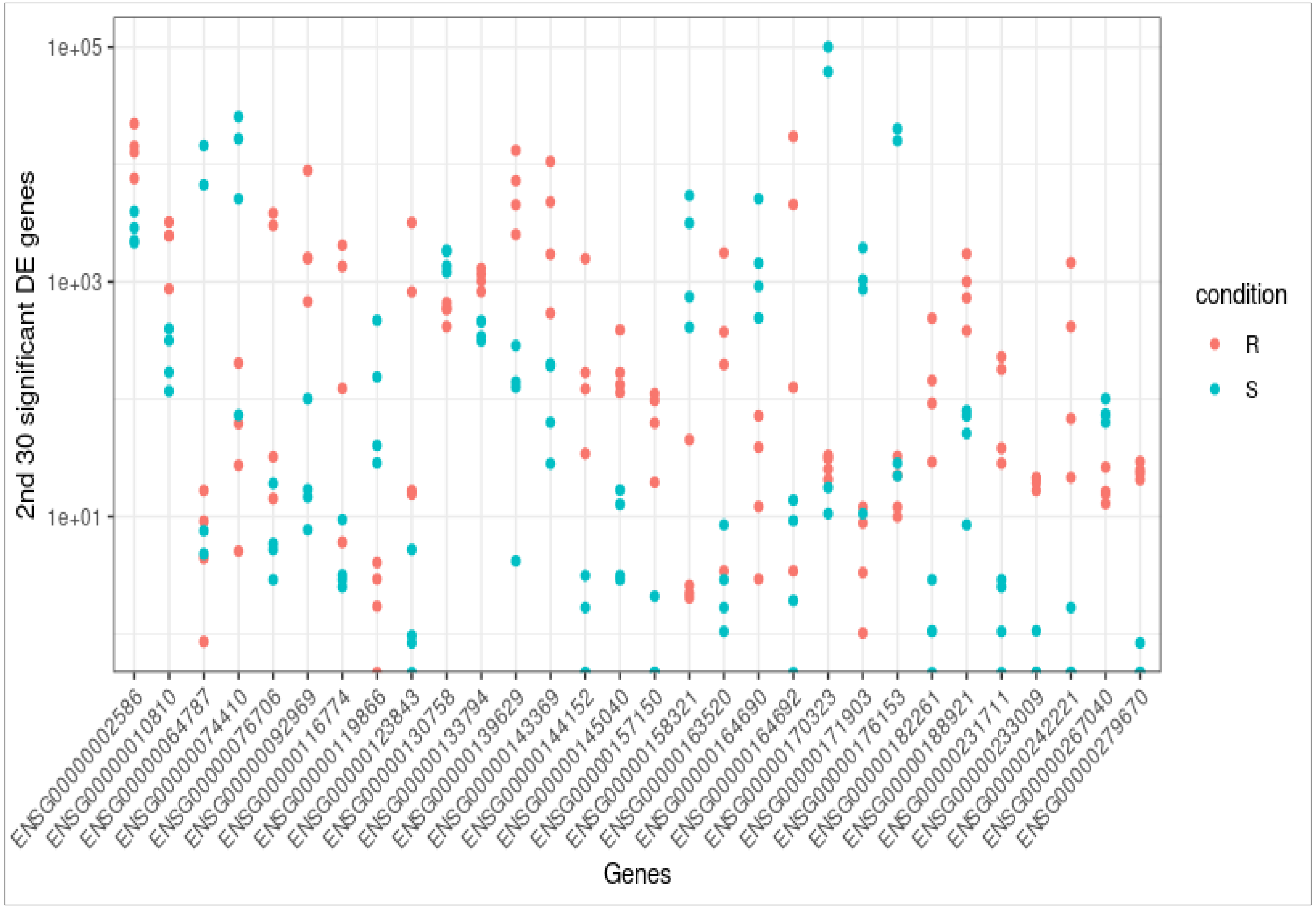
R: PEM resistant

S: PEM sensitive

One point represent one cell line

Example 1: RNA-seq Analysis

Resistant vs Sensitive: top 60 genes 31-60



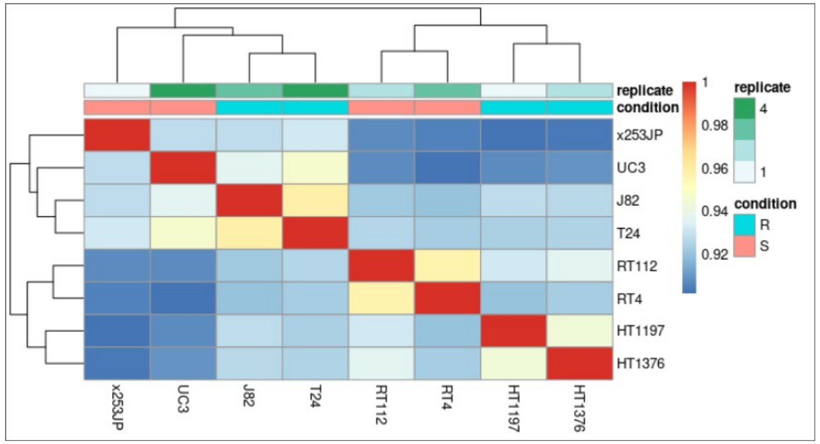
R: PEM resistant

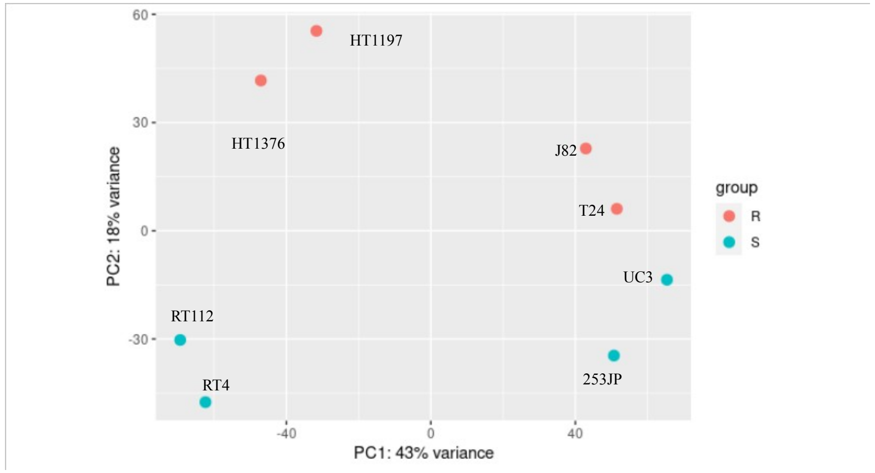
S: PEM sensitive

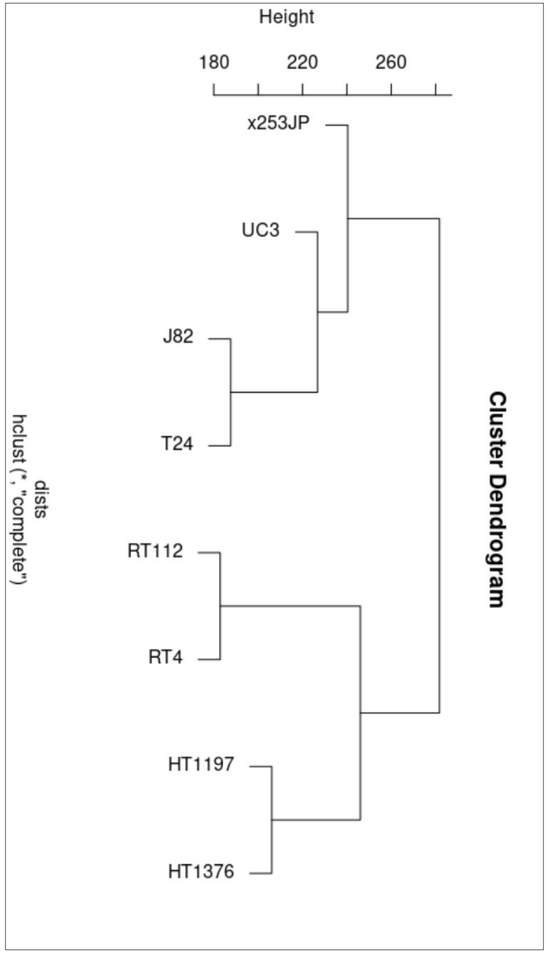
One point represent one cell line

Example 1: RNA-seq Analysis

Resistant vs Sensitive PEM cells: Heatmap and PCA plot

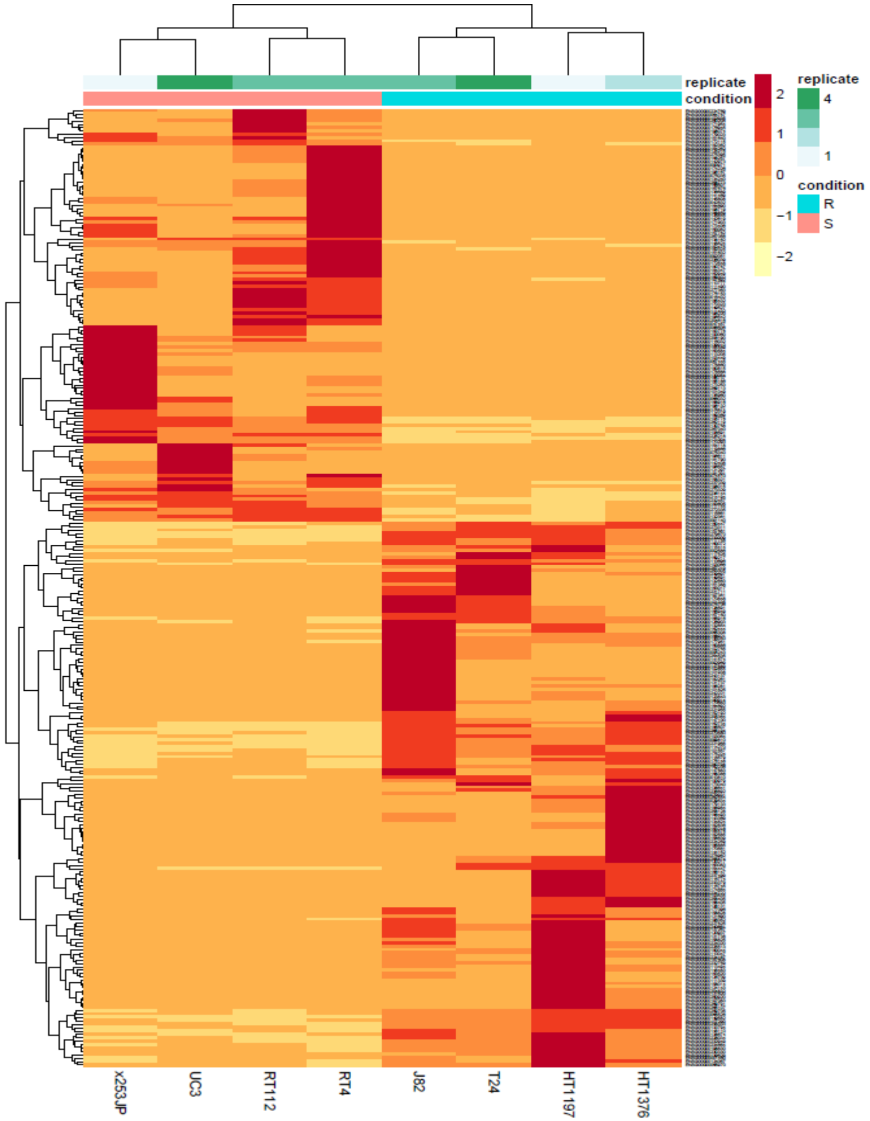






Example 1: RNA-seq Analysis

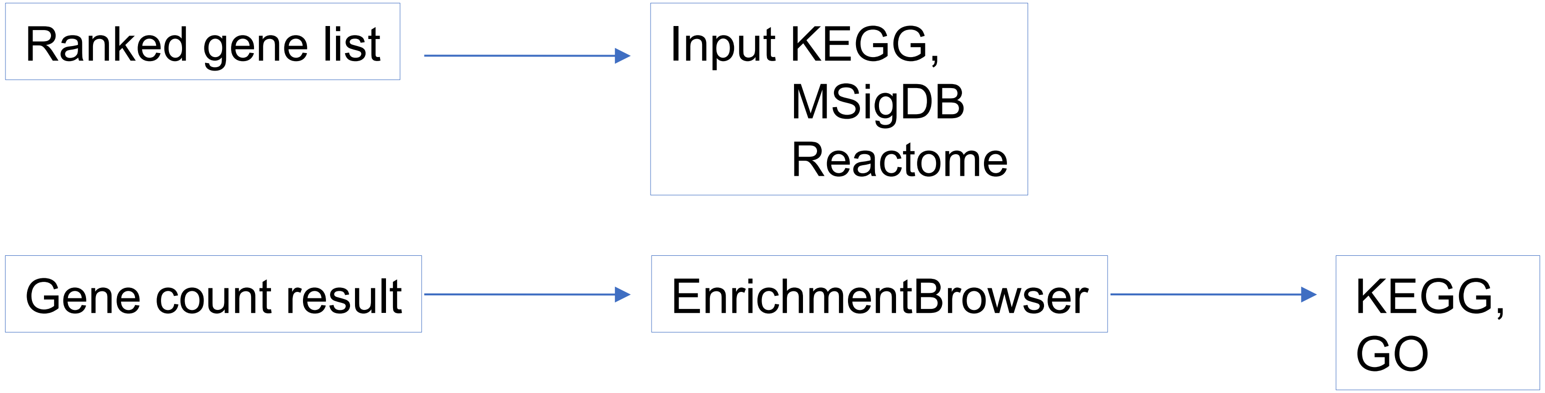
Resistant vs Sensitive PEM cells: Heatmap (p <0.01)



R: PEM resistant

S: PEM sensitive

Pathway/Gene Set Analysis



Ranked gene list → Input KEGG, MSigDB Reactome

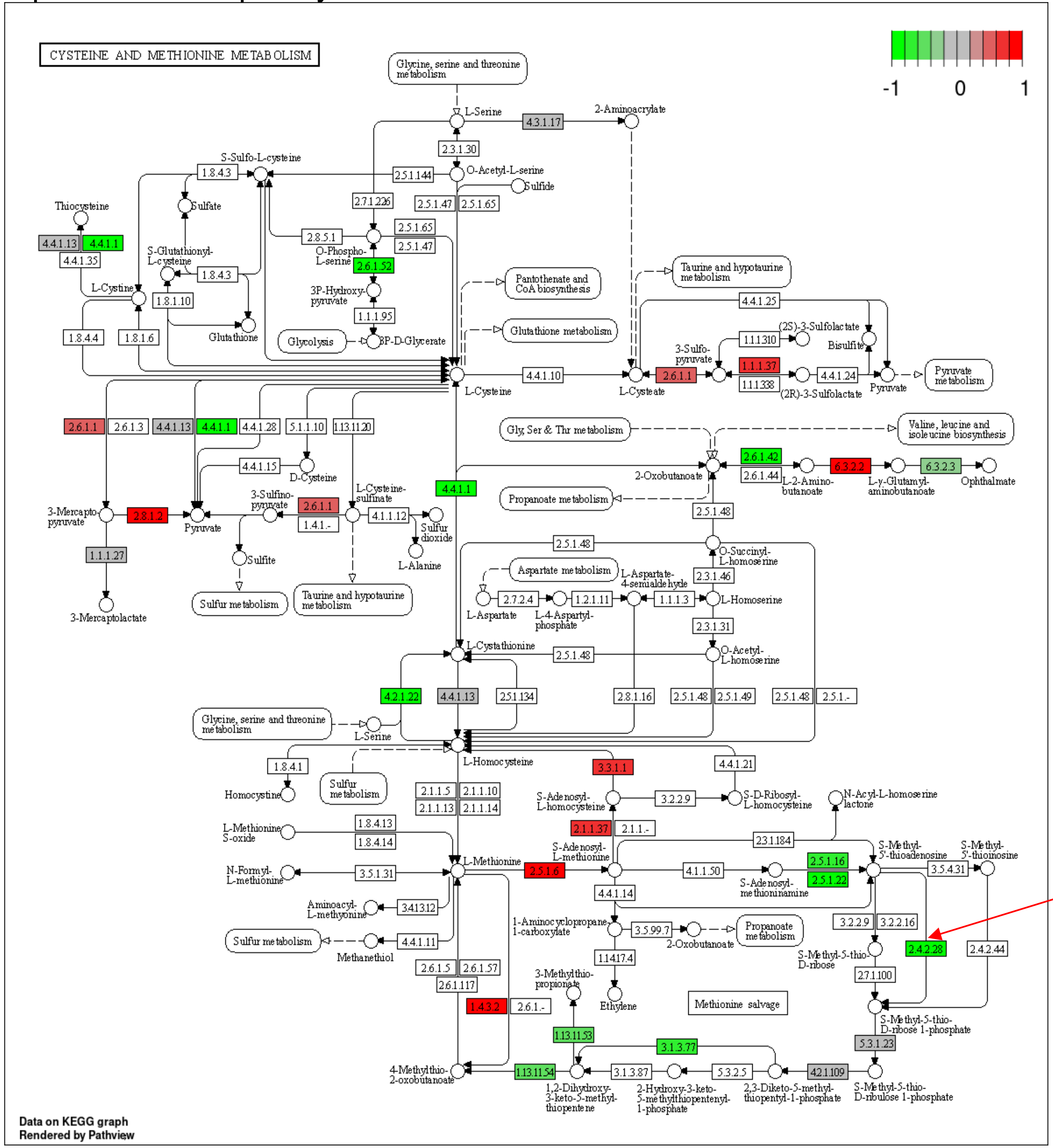
Gene count result → EnrichmentBrowser → KEGG, GO

Pathway: a series of interactions among molecules in a cell, leads to a product or a change.

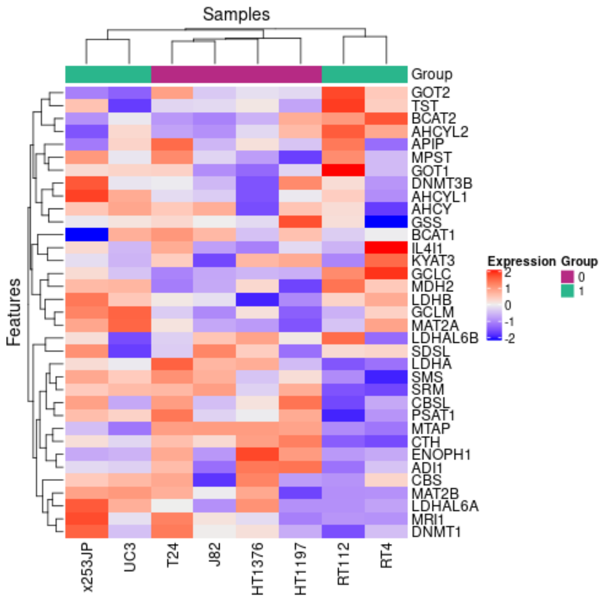
Gene set: an unordered and unstructured collection of genes, can be associated with:

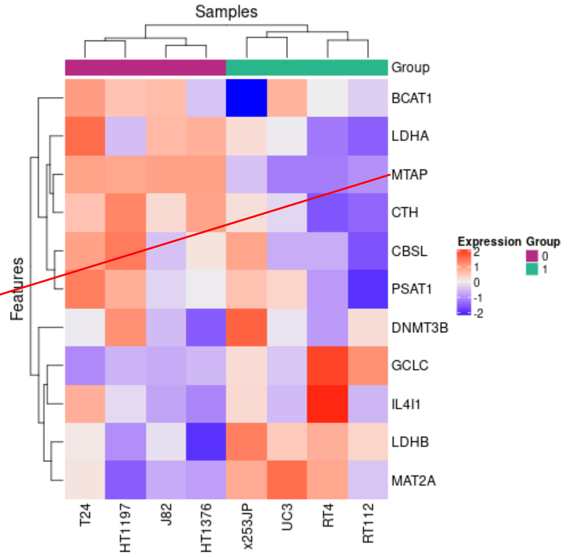
- a specific biological process
- Location
- disease

Example 1: RNA-seq Analysis

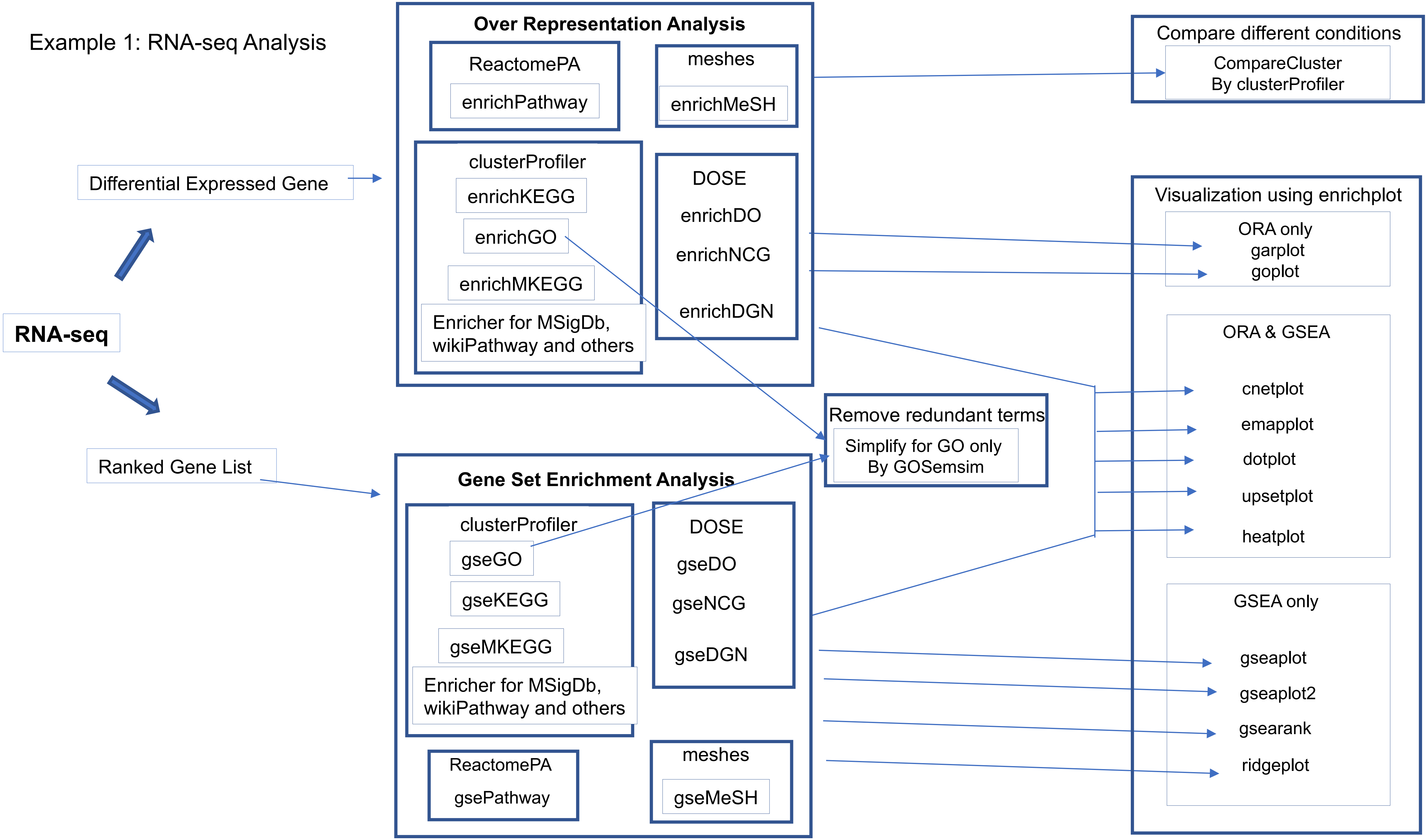


Pathway Results





Example 1: RNA-seq Analysis



RNA-seq

Differential Expressed Gene

Ranked Gene List

Over Representation Analysis

ReactomePA

enrichPathway

meshes

enrichMeSH

clusterProfiler

enrichKEGG

enrichGO

enrichMKEGG

Enricher for MSigDb, wikiPathway and others

DOSE

enrichDO

enrichNCG

enrichDGN

Gene Set Enrichment Analysis

clusterProfiler

gseGO

gseKEGG

gseMKEGG

Enricher for MSigDb, wikiPathway and others

DOSE

gseDO

gseNCG

gseDGN

ReactomePA

gsePathway

meshes

gseMeSH

Remove redundant terms

Simplify for GO only
By GOSemsim

Compare different conditions

CompareCluster
By clusterProfiler

Visualization using enrichplot

ORA only
garplot
goplot

ORA & GSEA

cnetplot

emapplot

dotplot

upsetplot

heatplot

GSEA only

gseaplot

gseaplot2

gsearank

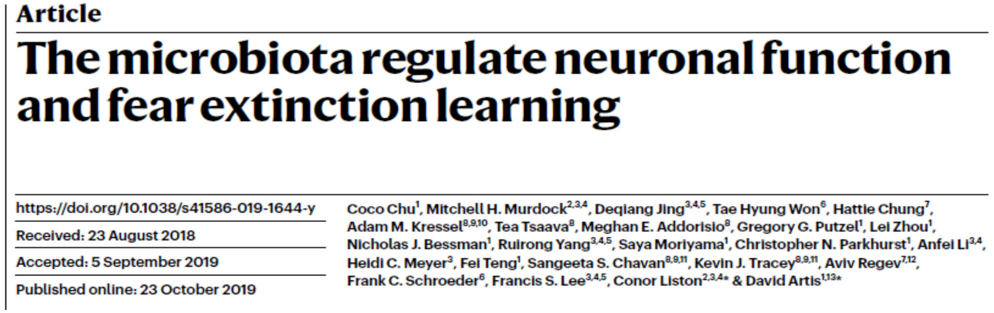
ridgeplot

Part IV Example 2

scRNA-seq Reanalysis on Entorhinal Cortex from Brain

Example 2: scRNA-seq Analysis

Authors' data generation analysis on Cellranger and Loupe Cell Browser:



Article

The microbiota regulate neuronal function and fear extinction learning

https://doi.org/10.1038/s41586-019-1644-y

Received: 23 August 2018

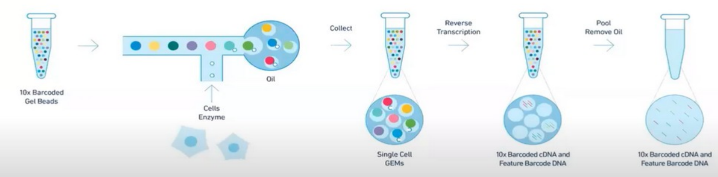
Accepted: 5 September 2019

Published online: 23 October 2019

Coco Chu[1], Mitchell H. Murdock[2,3,4], Deqiang Jing[3,4,5], Tae Hyung Won[6], Hattie Chung[7], Adam M. Kressel[8,9,10], Tea Tsaava[8], Meghan E. Addorisio[8], Gregory G. Putzel[1], Lei Zhou[1], Nicholas J. Bessman[1], Ruirong Yang[3,4,5], Saya Moriyama[1], Christopher N. Parkhurst[1], Anfei Li[3,4], Heidi C. Meyer[3], Fei Teng[1], Sangeeta S. Chavan[8,9,11], Kevin J. Tracey[8,9,11], Aviv Regev[7,12], Frank C. Schroeder[6], Francis S. Lee[3,4,5], Conor Liston[2,3,4*] & David Artis[1,13*]

Nature 574:543-548, 2019

Nuclei extracted from mouse mPFC (medial prefrontal cortex)



Mapping reads to the genome and quality control for expressed matrix

Cell type identification with reference dataset BRETIGEA

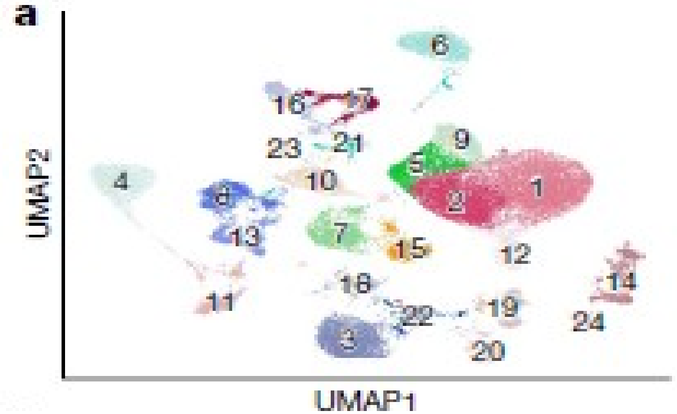
PCA and UMAP; identifying individual and sex-specific gene

Differential expression and gene set enrichment analysis

CellRouter analysis for gene regulatory change (GRN)

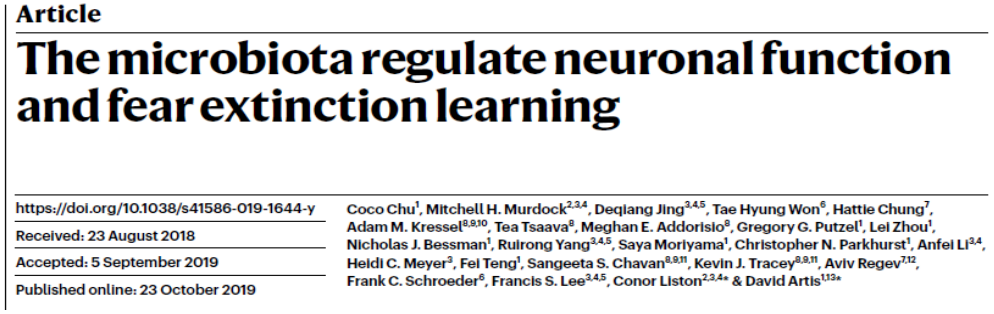
Functional annotation

Comparison with other published data



Example 2: scRNA-seq Analysis

My analysis on Cellranger and Loupe Cell Browser



Article

The microbiota regulate neuronal function and fear extinction learning

https://doi.org/10.1038/s41586-019-1644-y

Received: 23 August 2018

Accepted: 5 September 2019

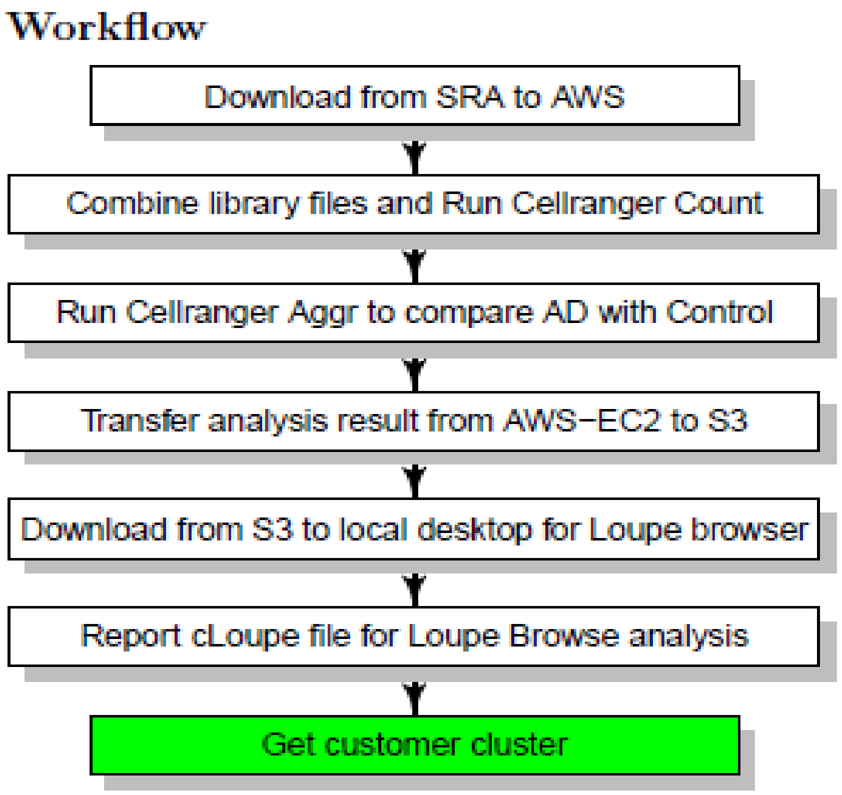
Published online: 23 October 2019

Coco Chu[1], Mitchell H. Murdock[2,3,4], Deqiang Jing[3,4,5], Tae Hyung Won[6], Hattie Chung[7], Adam M. Kressel[8,9,10], Tea Tsaava[8], Meghan E. Addorisio[8], Gregory G. Putzel[1], Lei Zhou[1], Nicholas J. Bessman[1], Ruirong Yang[3,4,5], Saya Moriyama[1], Christopher N. Parkhurst[1], Anfei Li[3,4], Heidi C. Meyer[3], Fei Teng[1], Sangeeta S. Chavan[8,9,11], Kevin J. Tracey[8,9,11], Aviv Regev[7,12], Frank C. Schroeder[6], Francis S. Lee[3,4,5], Conor Liston[2,3,4]* & David Artis[1,13]*

Nature 574:543-548, 2019

All single-cell RNA sequencing data are available from Sequencing Read Archive (SRA), with Download identifiers:
GEO GSE135326, 4x samples (two treated, two control)

sample	Sequence file		
1	Artis_A3_I1_001.fastq.gz Artis_A3_R1_001.fastq.gz Artis_A3_R2_001.fastq.gz.part-aa Artis_A3_R2_001.fastq.gz.part-ab Artis_A3_R2_001.fastq.gz.part-ac	mouse	Treated
2	Artis_A4_I1_001.fastq.gz Artis_A4_R1_001.fastq.gz Artis_A4_R2_001.fastq.gz.part-aa Artis_A4_R2_001.fastq.gz.part-ab Artis_A4_R2_001.fastq.gz.part-ac	mouse	Treated
3	Artis_B3_I1_001.fastq.gz Artis_B3_R1_001.fastq.gz Artis_B3_R2_001.fastq.gz.part-aa Artis_B3_R2_001.fastq.gz.part-ab Artis_B3_R2_001.fastq.gz.part-ac	mouse	Control
4	Artis_B4_I1_001.fastq.gz Artis_B4_R1_001.fastq.gz Artis_B4_R2_001.fastq.gz.part-aa Artis_B4_R2_001.fastq.gz.part-ab Artis_B4_R2_001.fastq.gz.part-ac	mouse	control



Workflow

Download from SRA to AWS

Combine library files and Run Cellranger Count

Run Cellranger Aggr to compare AD with Control

Transfer analysis result from AWS−EC2 to S3

Download from S3 to local desktop for Loupe browser

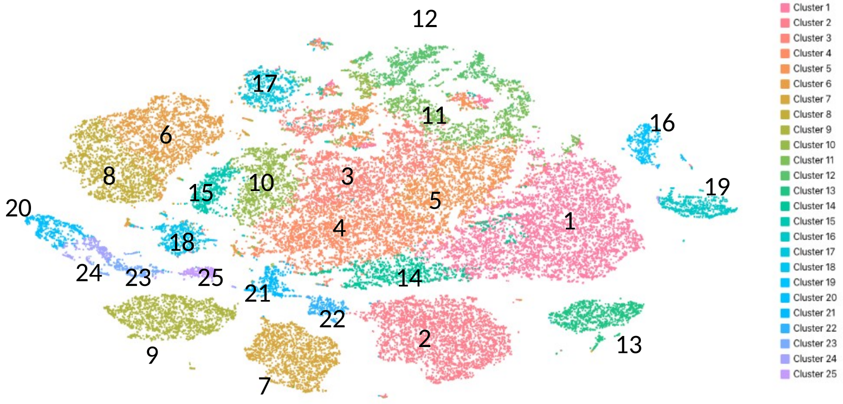
Report cLoupe file for Loupe Browse analysis

Get customer cluster

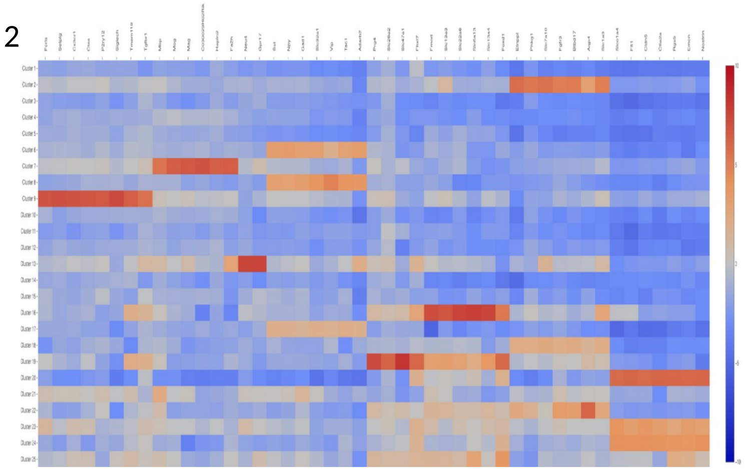
Example 2: scRNA-seq Analysis

My analysis on Cellranger and Loupe Cell Browser: 15x clusters = 8 cell types

XinqiaoFig 1



XinqiaoFig 2



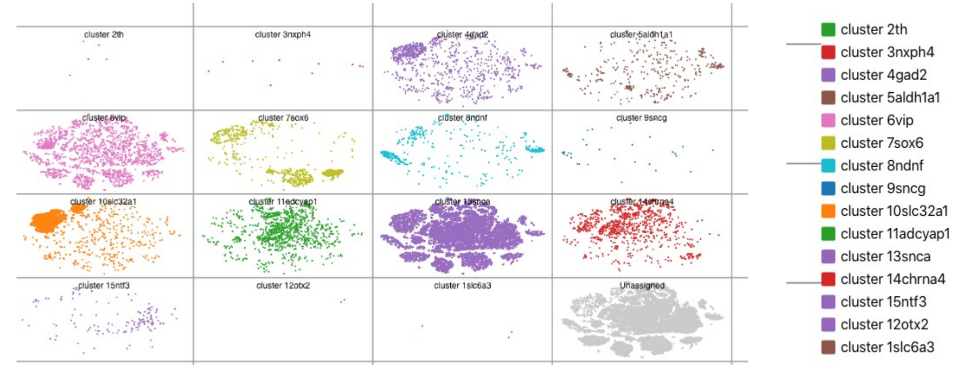
a. Based on the authors’ results, we identified cell types. Our 25 clusters are correlated with authors. Figure1,2.

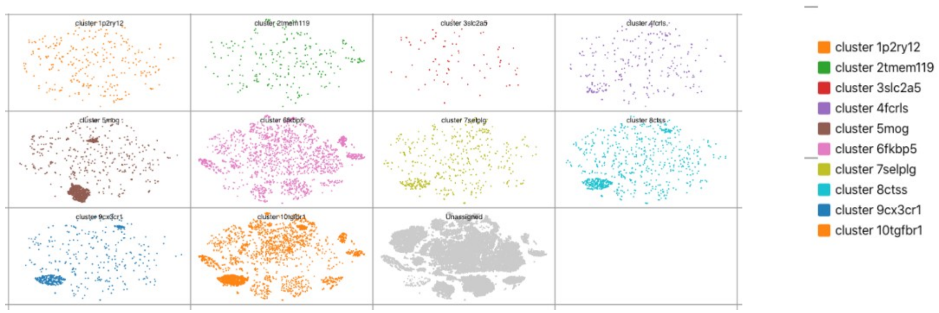
b. Searching signature genes for dopamine cells with Gene/Feature Expression Mode. Using 15 dopamine cell marker genes, we find **Gad2, Sox6, Ndnf and Slc32a1** are neatly clustered and would be good candidates for dopamine cell analysis as figure 3. We further compared treatment mice with control mice with those dopamine genes, and noticed the different distribution.

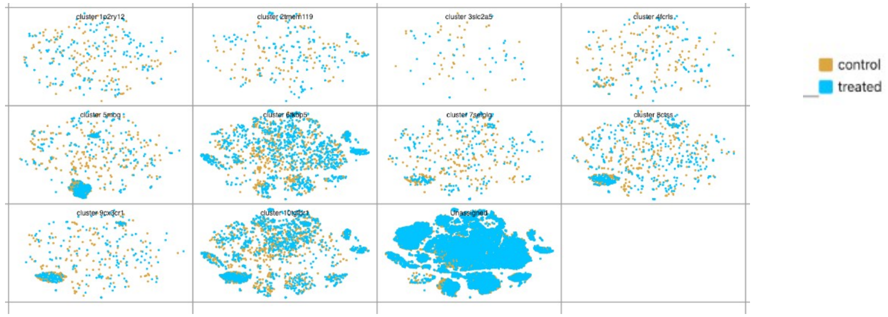
c. Searching signature genes for microglia cells with Gene/Feature Expession Mode. Using 10 microglia cell marker genes, we find **Mog** (for cluster 7); **Ctss, Selplg, Cx3cr1** and **Tgfbr1** (for cluster 9); **Fkbp5** (for cluster 16 and 19) are neatly clustered and would be good candidates for cell analysis as figure 4. We further compared treatment mice with control mice with those microglia cell genes, and noticed upregulated expression for **Fkbp5** gene in our cluster 16 and 19 , consistent with original authors’ volcano result in Extended Data Fig. 6. and Fig. 8 (6-Microglia, 19-Astrocyte 2, 20-Undermined 2, and 21-exPFC/Microglia)

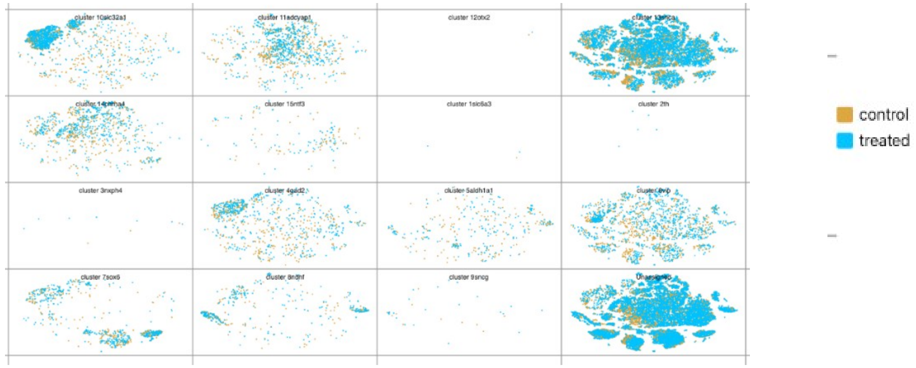
Fig 4 microglia gene cluster

XinqiaoFig 3









Further reading and practice on melanoma scRNAseq

Toward Minimal Residual Disease-Directed Therapy in Melanoma

Florian Rambow,[1,2,15] Aljosja Rogiers,[1,2,15] Oskar Marin-Bejar,[1,2] Sara Aibar,[3,4] Julia Femel,[5] Michael Dewaele,[1,2] Panagiotis Karras,[1,2] Daniel Brown,[6] Young Hwan Chang,[7] Maria Debiec-Rychter,[8] Carmen Adriaens,[1,2] Enrico Radaelli,[9] Pascal Wolter,[10] Oliver Bechter,[10] Reinhard Dummer,[11] Mitchell Levesque,[11] Adriano Piris,[12] Dennie T. Frederick,[12] Genevieve Boland,[12] Keith T. Flaherty,[13] Joost van den Oord,[14] Thierry Voet,[6] Stein Aerts,[3,4] Amanda W. Lund,[5] and Jean-Christophe Marine[1,2,16,17,*]

[1]Laboratory for Molecular Cancer Biology, VIB Center for Cancer Biology, KU Leuven, Leuven, Belgium

[2]Department of Oncology, KU Leuven, Leuven, Belgium

[3]Laboratory of Computational Biology, VIB Center for Brain & Disease Research, KU Leuven, Leuven, Belgium

Cell 2018 Aug 9; 174(4):843-855.e19. PMID:30017245

Smart-seq2 single-cell RNAseq
Raw data: GEO: GSE116237

Acknowledgement

MD Anderson Cancer Center UT

- William Benedict MD
- Monica Spears
- Jianjun Gao MD PhD
- Derek Ng
- Mark Titus PhD
- Jianfeng Cheng MD PhD

HSCSA UT

- Senlin Li MD PhD
- Shujie Zhao MD