Laboratório de Imunobiologia

RNA-Seq

Aula 1: Banco de dados

http://bit.ly/2IXMbRn

Edgar Kozlova
Gabriela Luiz

O Curso

Pré requisitos:

Notebook, WiFi, Notepad++, R e Rstudio

Programação das aulas:

1. Banco de dados:
   NCBI/SRA
   NCBI/GEO
2. RStudio e Instalação de pacotes
   edgeR, limma, pheatmap, gplots, ROTS
3. Normalização e Análise Diferencial
   voom, RPKM, FPKM, TPM, CPM, counts
4. Análise Diferencial e Visualização
   Script, MAplot, VolcanoPlot, Heatmap, Venn

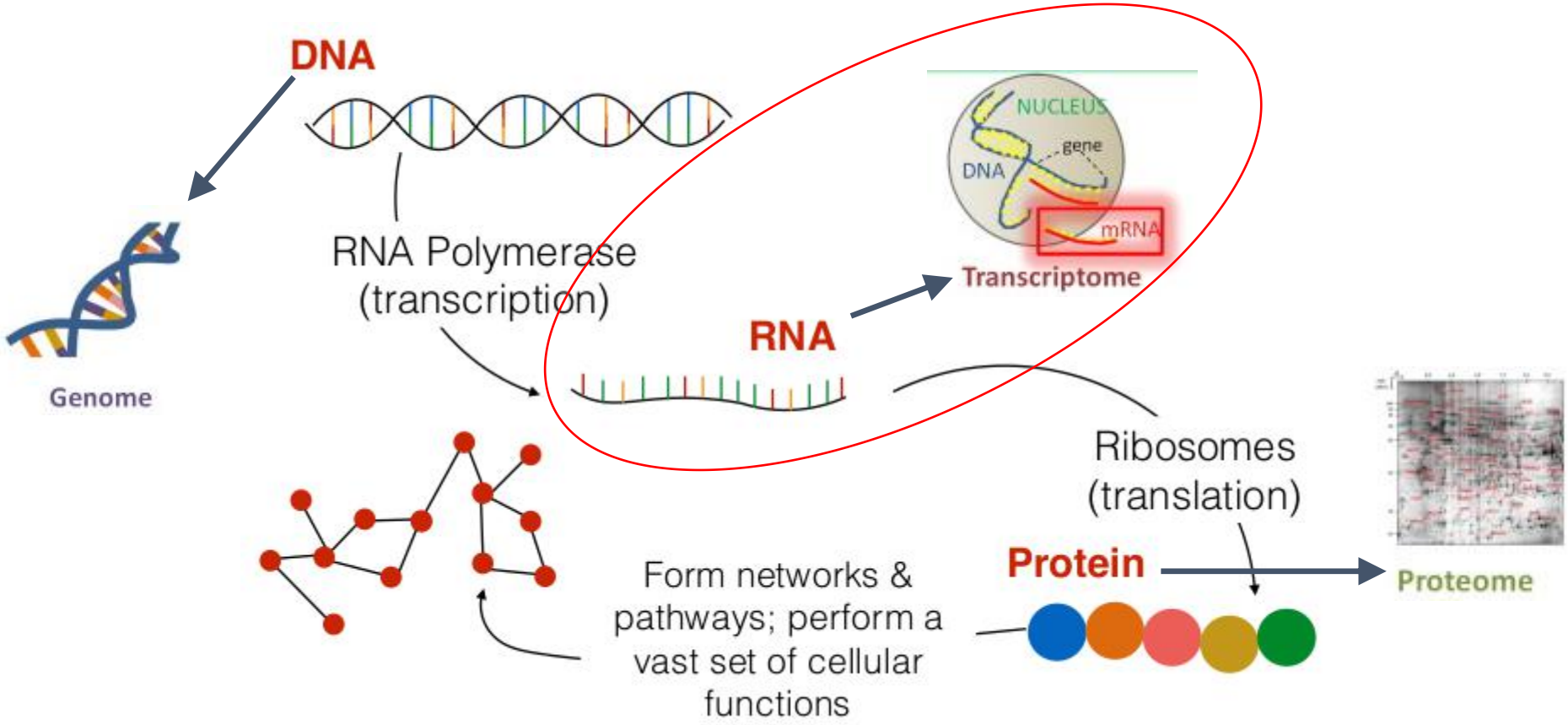
Objetivo

- Introduzir os principais conceitos de RNA-Seq
  - Metodologia/Passos
- Apresentar os bancos de dados
  - Selecionar estudos de interesse
  - Obter os dados de expressão (ou sequenciamento)

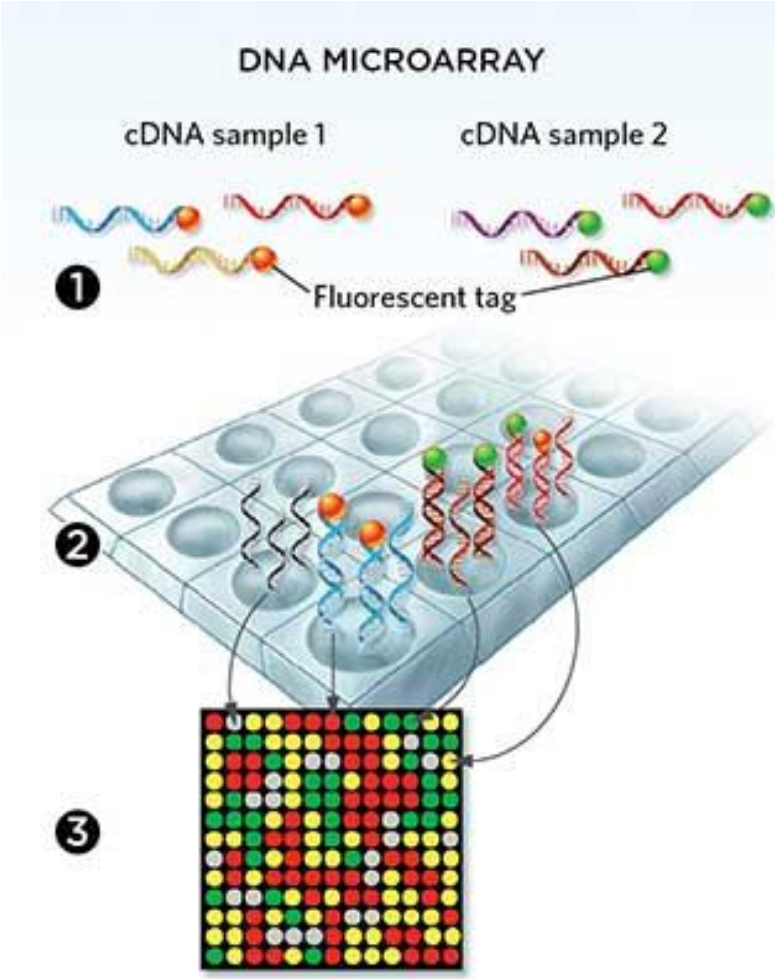
TRANSCRIPTÔMICA

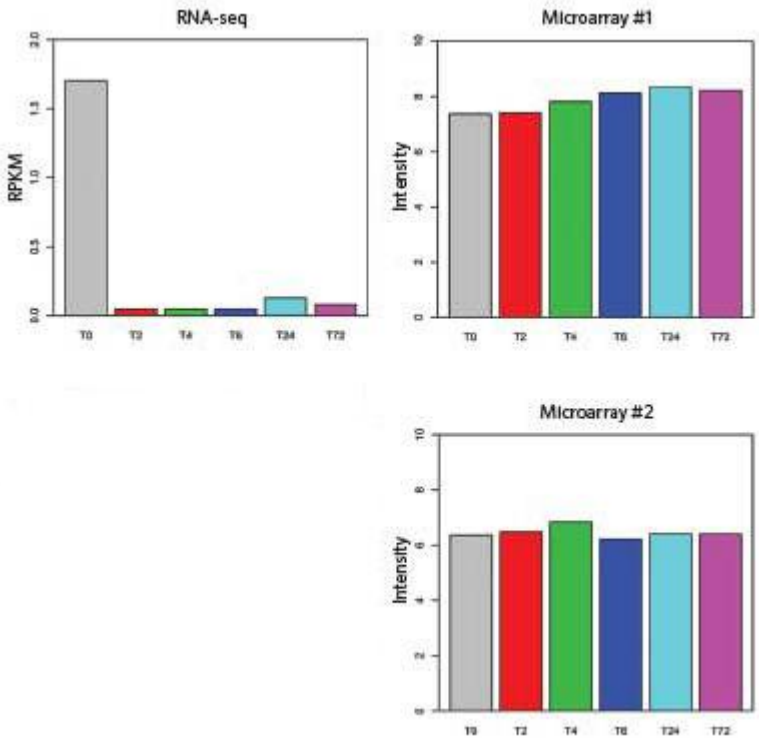
Principais conceitos de RNA-Seq

Transcriptômica



Transcriptômica



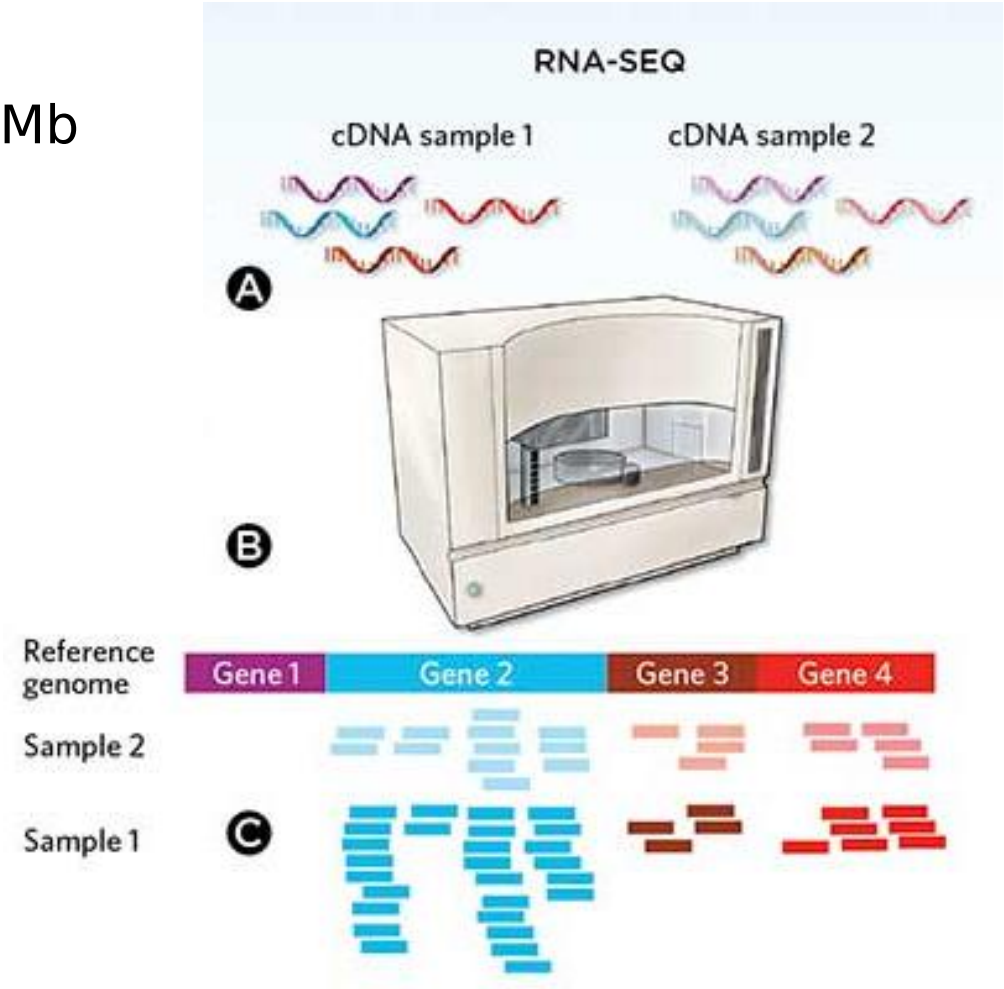


https://www.the-scientist.com/?articles.view/articleNo/43063/title/An-Array-of-Options/

Transcriptômica

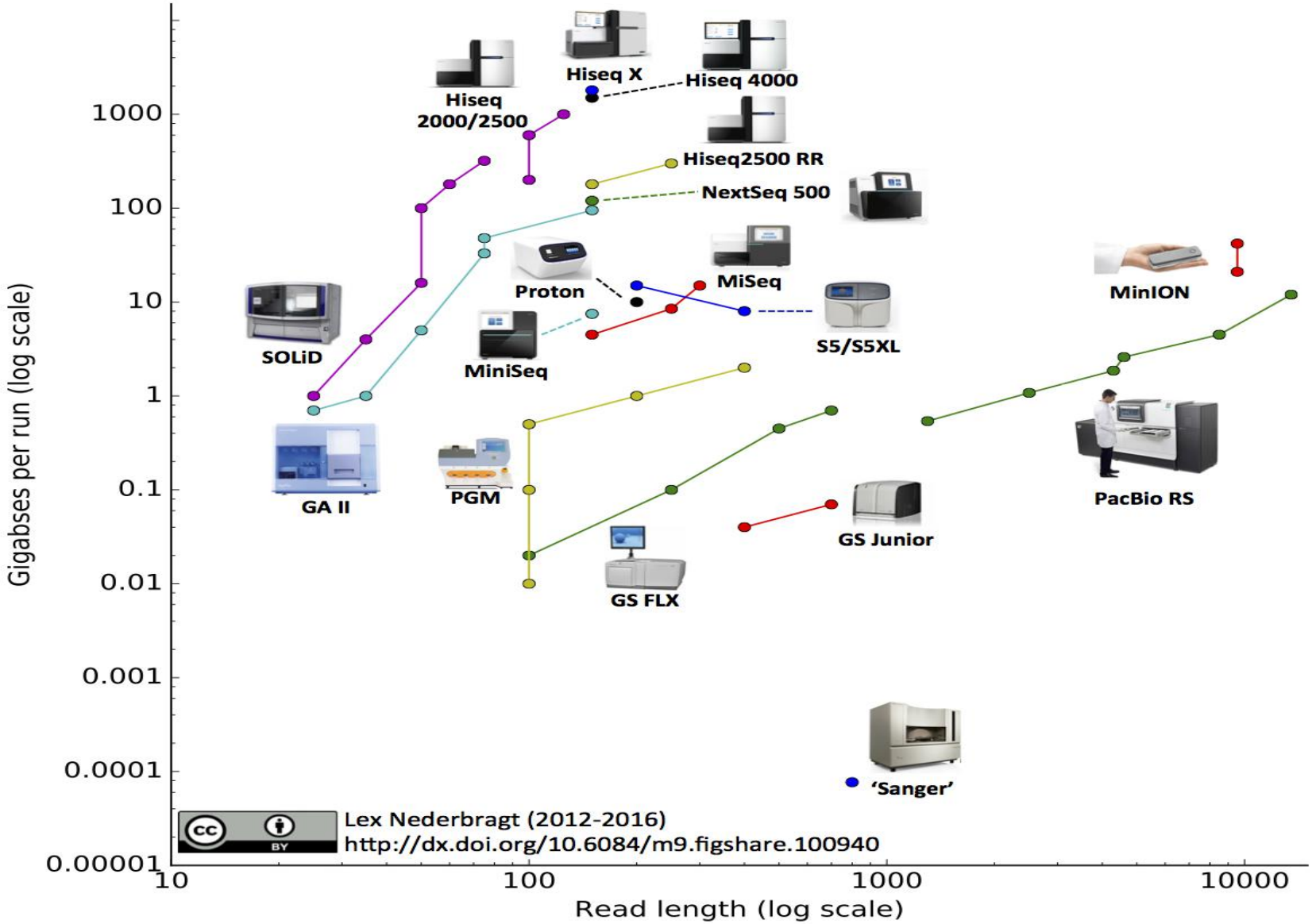
Affymetrix array ~= 5 Mb

RNA-seq ~= 10 Gb

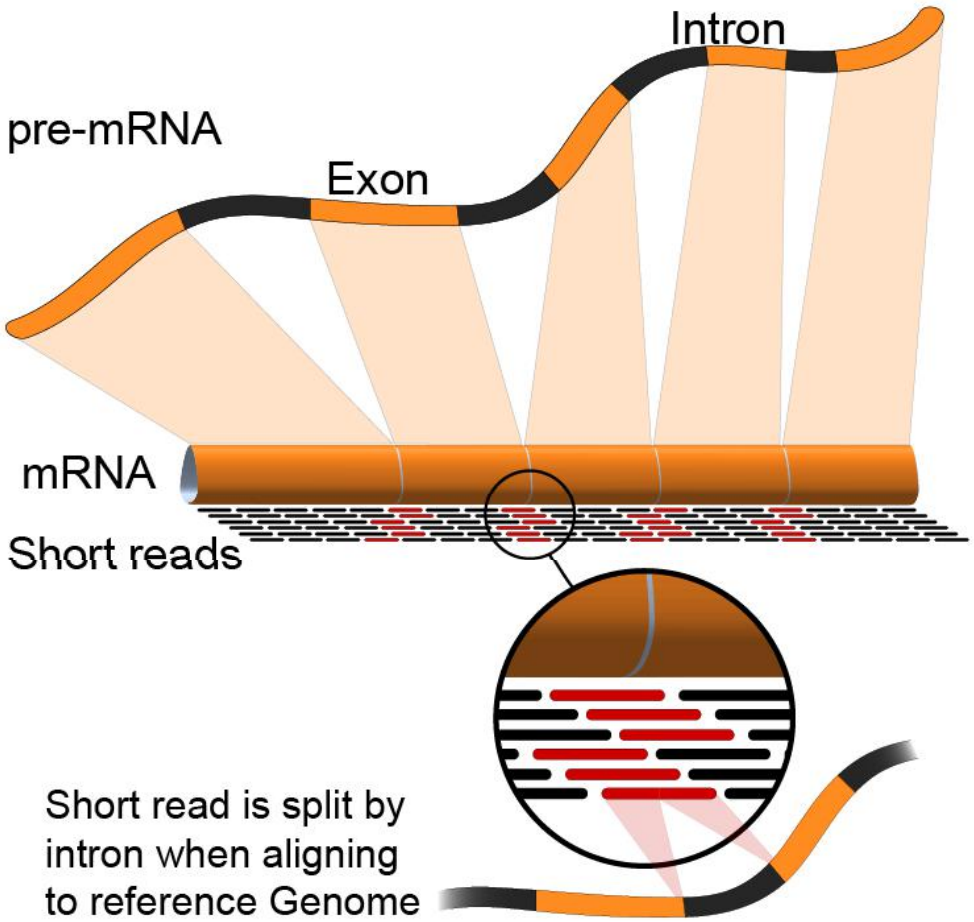


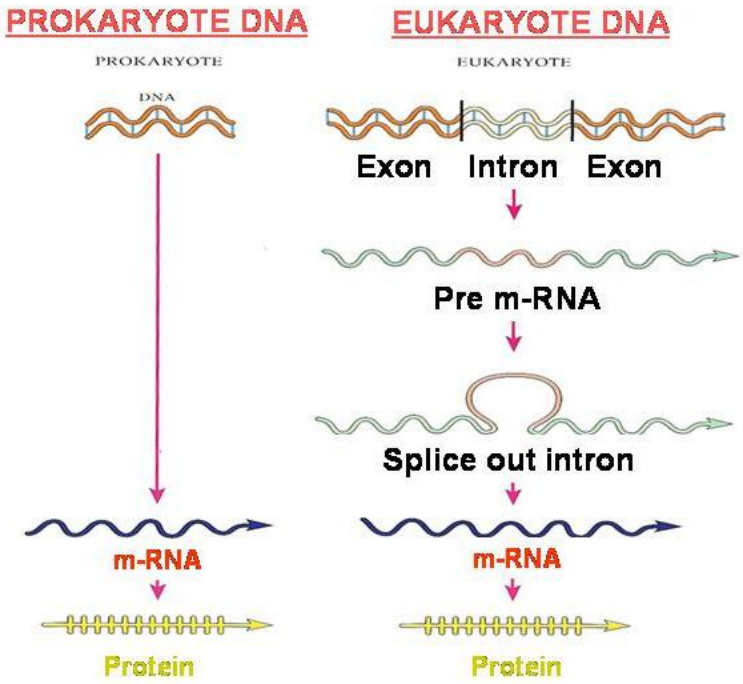
https://www.the-scientist.com/?articles.view/articleNo/43063/title/An-Array-of-Options/

Transcriptômica

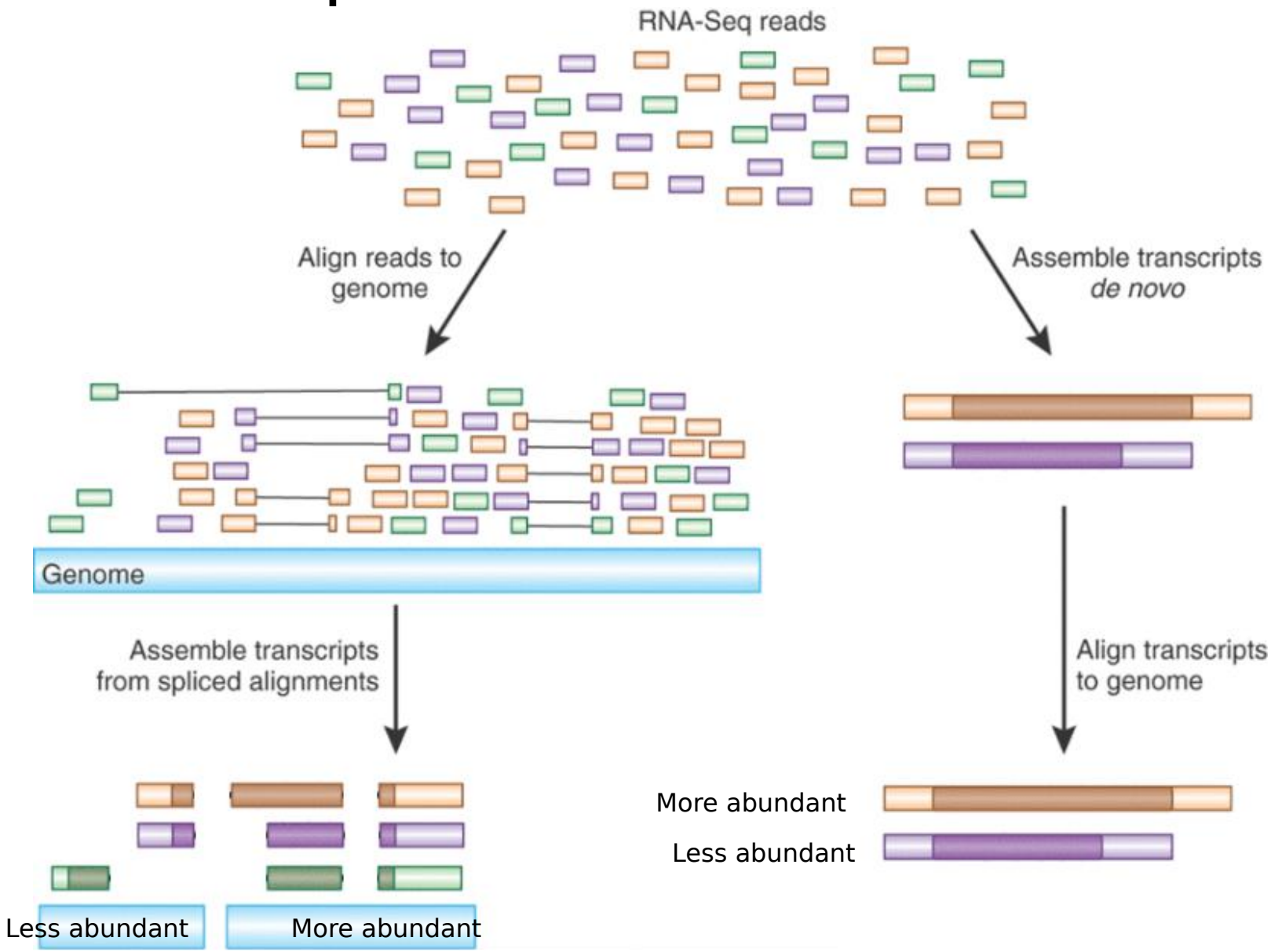


RNA-Seq


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

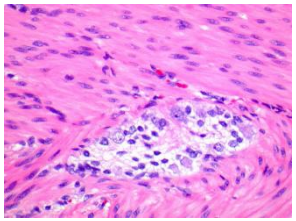

PROKARYOTE DNA
PROKARYOTE
DNA
m-RNA
Protein
EUKARYOTE DNA
EUKARYOTE
Exon
Intron
Exon
Pre m-RNA
Splice out intron
m-RNA
Protein

RNA-Seq

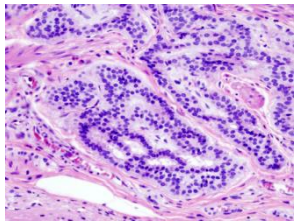

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant
Less abundant
More abundant

RNA-Seq

Amostras de interesse

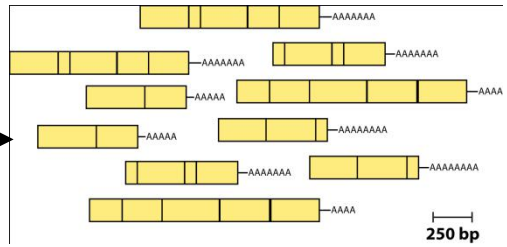


Condição 1

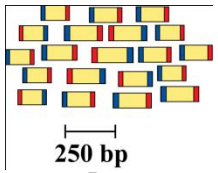


Condição 2

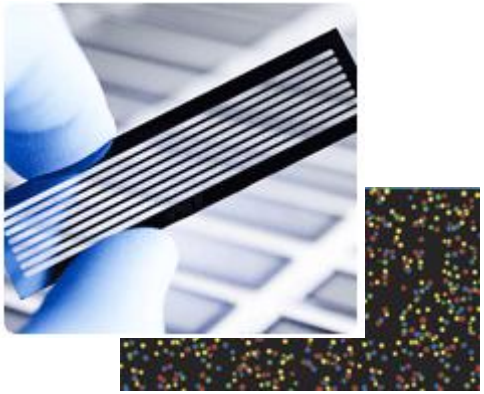
Purificação de RNAm

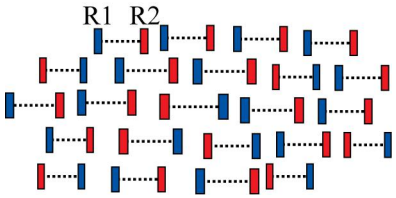


Síntese de cDNA, fragmentação, seleção por tamanho e ligação de adaptadores

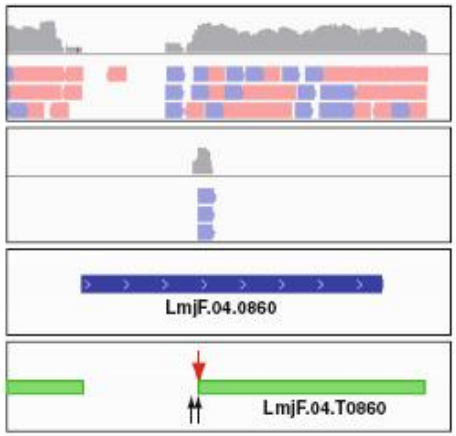


Sequenciamento

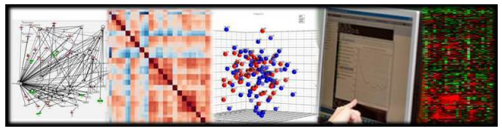




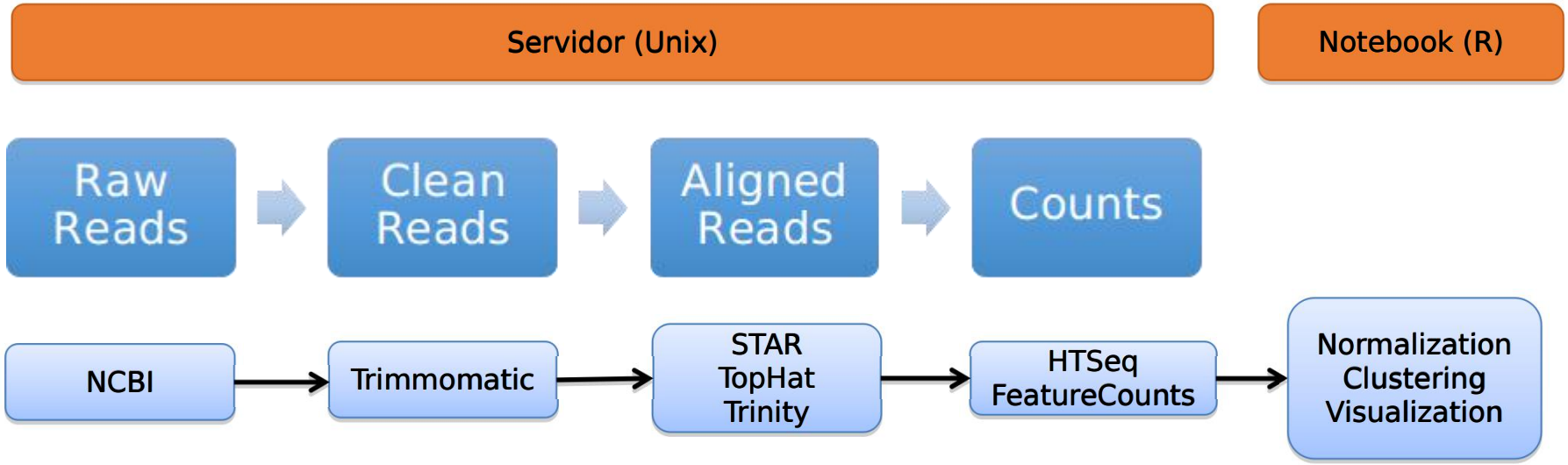
Alinhamento ao genoma e contagem de transcritos



Análise diferencial



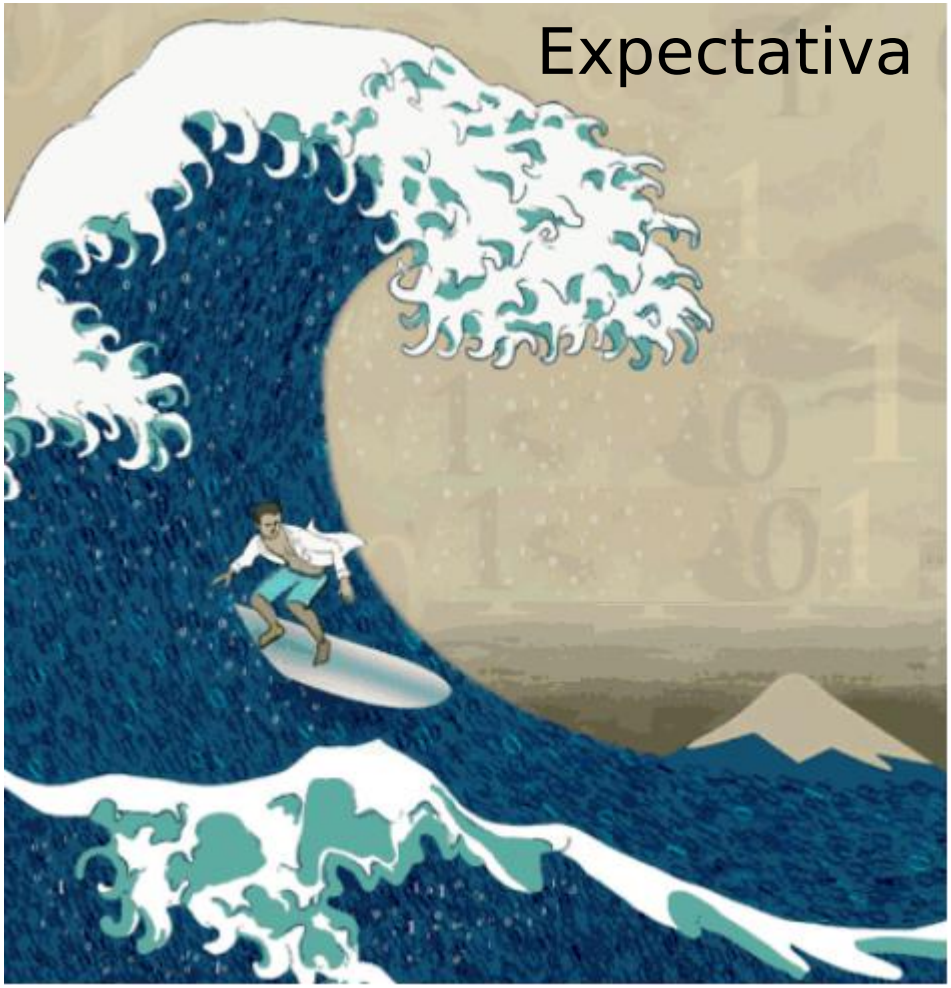
RNA-Seq


Servidor (Unix)
Notebook (R)
Raw Reads
Clean Reads
Aligned Reads
Counts
NCBI
Trimmomatic
STAR
TopHat
Trinity
HTSeq
FeatureCounts
Normalization
Clustering
Visualization

BANCO DE DADOS

Apresentar principais bancos de dados
Selecionar estudos de interesse
Obter os dados de expressão (ou sequenciamento)

Banco de dados

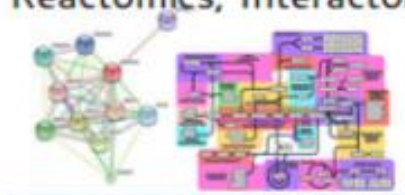


Banco de dados

Realidade

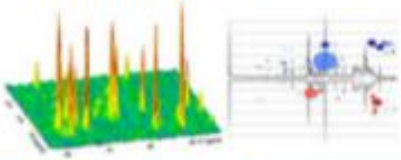
Banco de dados

Reactomics, Interactomics



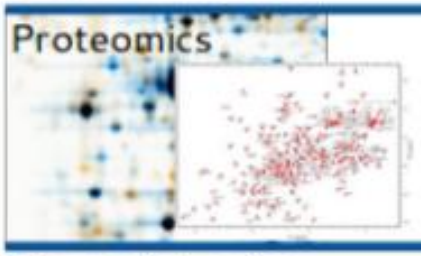
REACTOME IntAct BioCyc KEGG

Metabolomics



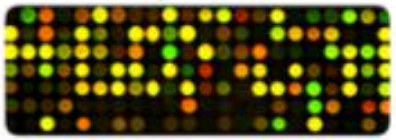
ChEBI PubChem hmp PRIME Platform for RIKEN Metabolomics

Proteomics



UniProt swissprot THE HUMAN PROTEIN ATLAS

Transcriptomics



GEO Gene Expression Omnibus ARRAYEXPRESS NCBI

Genomics

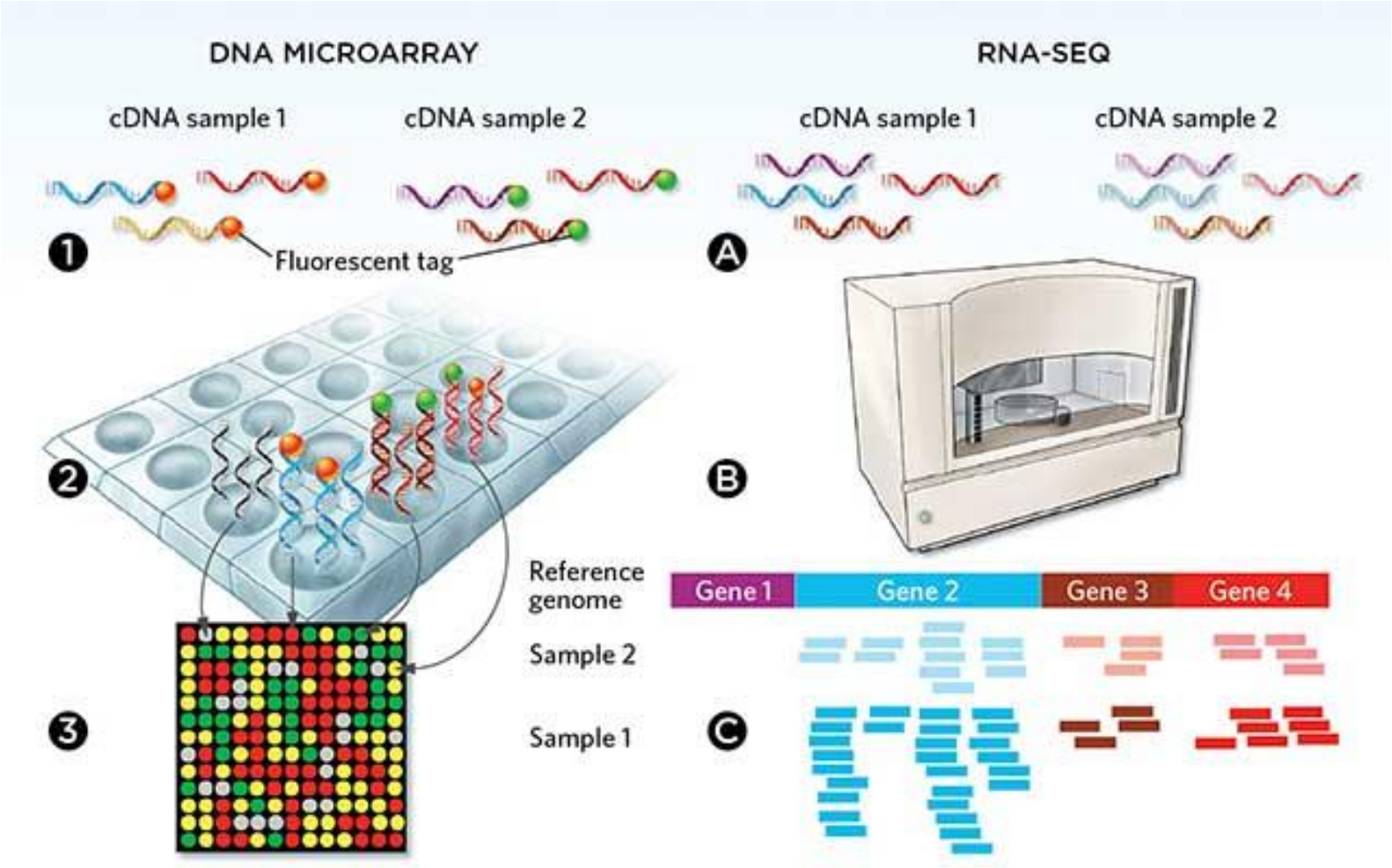
EMBL NUCLEOTIDE SEQUENCE DATABASE ENCODE NCBI

Banco de dados

Sequence Read Archive

Gene Expression Omnibus

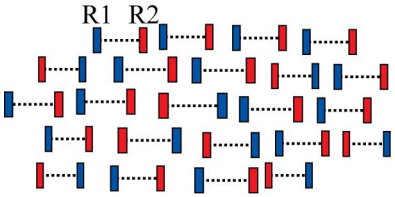
Gene Expression Omnibus



https://www.the-scientist.com/?articles.view/articleNo/43063/title/An-Array-of-Options/

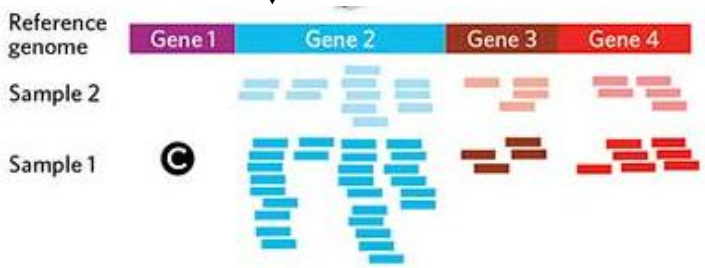
RNA-Seq

Sequence Read Archive

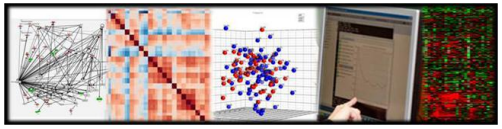


GEO
Gene Expression Omnibus

Alinhamento ao genoma e contagem de transcritos



Análise diferencial



Conhecendo o seu organismo

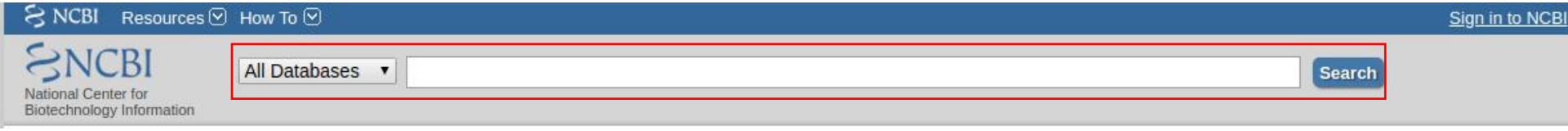
ww.ncbi.nlm.nih.gov/

Escolher 1 organismo

- Homo sapiens

Avaliar dados disponíveis

- Genoma - Genome
- Transcriptoma - SRA, GEO



U.S. National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI databases

Homo sapiens ✖ Search

Results found in 36 databases for **Homo sapiens**

Literature

Books	71,376	books and reports
MeSH	5	ontology used for PubMed indexing
NLM Catalog	36,481	books, journals and more in the NLM Collections
PubMed	17,160,132	scientific and medical abstracts/citations
PubMed Central	2,337,981	full-text journal articles

Health

ClinVar	12	human variations of clinical significance
dbGaP	917	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	6	medical genetics literature and links
OMIM	70	online mendelian inheritance in man

Genes

EST	8,864,000	expressed sequence tag sequences
Gene	549,648	collected information about gene loci
GEO DataSets	1,458,823	functional genomics studies
GEO Profiles	61,958,910	gene expression and molecular abundance profiles
HomoloGene	18,732	homologous gene sets for selected organisms
PopSet	39,141	sequence sets from phylogenetic and population studies
UniGene	0	clusters of expressed transcripts

Proteins

Conserved Domains	108	conserved protein domains
Identical Protein	382,707	protein sequences grouped by

GTR	0	genetic testing registry
MedGen	6	medical genetics literature and links
OMIM	70	online mendelian inheritance in man
PubMed Health	18,279	clinical effectiveness, disease and drug reports

Genomes

Assembly	67	genome assembly information
BioCollections	0	museum, herbaria, and other biorepository collections
BioProject	42,143	biological projects providing data to NCBI
BioSample	4,202,331	descriptions of biological source materials
Clone	17,630,156	genomic and cDNA clones
dbVar	5,227,838	genome structural variation studies
Genome	1	genome sequencing projects by organism
GSS	1,784,160	genome survey sequences
Nucleotide	27,597,516	DNA and RNA sequences
Probe	27,386,764	sequence-based probes and primers
SNP	672,043,185	short genetic variations
SRA	1,418,232	high-throughput DNA and RNA sequence read archive
Taxonomy	1	taxonomic classification and nomenclature catalog

Proteins

Conserved Domains	108	conserved protein domains
Identical Protein Groups	382,707	protein sequences grouped by identity
Protein	85,342,613	protein sequences
Protein Clusters	15	sequence similarity-based protein clusters
Sparcle	608	functional categorization of protein by domain architecture
Structure	39,837	experimentally-determined biomolecular structures

Chemicals

BioSystems	26,329	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	377,174	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	0	deposited substance and chemical information

SRA

NCBI Resources How To Sign in to NCBI

SRA | SRA | Homo sapiens | Search

Create alert Advanced Help

Access
Controlled (301,470)
Public (1,114,808)

Source
DNA (1,108,590)
RNA (293,311)

Type
exome (210,166)
genome (421,546)

Other
aligned data (434,075)

Clear all

Show additional filters

Summary 20 per page Send to: **Filters:** Manage Filters

Search results

Items: 1 to 20 of 1418232

<< First < Prev Page 1 of 70912 Next > Last >>

1. Other Sequencing of Listeria monocytogenes
   1 ILLUMINA (Illumina HiSeq 2500) run: 2.5M spots, 467M bases, 219.7Mb downloads
   Accession: SRX4082325
2. Other Sequencing of Listeria monocytogenes
   1 ILLUMINA (Illumina HiSeq 2500) run: 2.3M spots, 440.8M bases, 194.6Mb downloads
   Accession: SRX4082324
3. Other Sequencing of Listeria monocytogenes
   1 ILLUMINA (Illumina HiSeq 2500) run: 2.5M spots, 507.1M bases, 213Mb downloads
   Accession: SRX4082323
4. Other Sequencing of Listeria monocytogenes
   1 ILLUMINA (Illumina HiSeq 2500) run: 3.2M spots, 617M bases, 282.9Mb downloads
   Accession: SRX4082322
5. Other Sequencing of Listeria monocytogenes
   1 ILLUMINA (Illumina HiSeq 2500) run: 2.3M spots, 450.6M bases, 199.9Mb downloads
   Accession: SRX4082321
6. Other Sequencing of Listeria monocytogenes
   1 ILLUMINA (Illumina HiSeq 2500) run: 2.1M spots, 395.4M bases, 177.8Mb downloads
   Accession: SRX4082320

Results by taxon

Top Organisms [Tree]
- Homo sapiens *(993531)*
- human gut metagenome *(105535)*
- human metagenome *(36917)*
- Salmonella enterica *(33833)*
- Mycobacterium tuberculosis *(30062)*
- All other taxa *(218354)*

More...

Top Bioprojects

NIH Epigenomics Roadmap Init... *(1893)*
Production ENCODE epigenomic... *(1493)*
Production ENCODE functional... *(1137)*
Production ENCODE transcript... *(363)*

Search in related databases

Database	Access		all
	public	controlled	
BioSample	825,701	285,519	1,111,220
BioProject	12,952	528	13,480
dbGaP		2	2
GEO Datasets	252,319		252,319

Full

SRX4080450: human monocyte-derived macrophages
1 ILLUMINA (Illumina HiSeq 2000) run: 8.7M spots, 443.8M bases, 355.5Mb downloads

Design: RNA-seq

Submitted by: University of California, Los Angeles

Study: Type I and II IFN Conditioning of Human Macrophage Gene Expression Responses
PRJNA470733 • SRP145599 • All experiments • All runs
show Abstract

Sample:
SAMN09206001 • SRS3297852 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: d1-IFNg-LA-5-5
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: RANDOM
Layout: SINGLE

Runs: 1 run, 8.7M spots, 443.8M bases, 355.5Mb

Run	# of Spots	# of Bases	Size	Published
SRR7161124	8,702,049	443.8M	355.5Mb	2018-05-14

ID: 5566011

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

human monocyte-derived macrophages (SRR7161124)

Metadata | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR7161124	8.7M	443.8Mbp	372.7M	46.6%	2018-05-14	public

Quality graph (bigger)

This run has 1 read per spot:

L=51, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX4080450 to BLAST	d1-IFNg-LA-5-5	Illumina	RNA-Seq	TRANSCRIPTOMIC	RANDOM	SINGLE

Design:

RNA-seq

Biosample	Sample Description	Organism	Links
SAMN09206001 (SRS3297852)		Homo sapiens	PRJNA470733 [Homo sapiens]

Bioproject	SRA Study	Title
PRJNA470733	SRP145599	Type I and II IFN Conditioning of Human Macrophage Gene Expression Responses

Show abstract

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

human monocyte-derived macrophages (SRR7161124)

Metadata | Reads | Download

Filter: Find Filtered Download What does it do?

What can the filter be applied to?

< 1 1 870205 > View: biological reads technic

1. SRR7161124.1 SRS3297852
name: 2:1101:6.10:98.90:N, member: default
2. SRR7161124.2 SRS3297852
name: 2:1101:29.90:99.00:Y, member: default
3. SRR7161124.3 SRS3297852
name: 2:1101:381.00:99.60:Y, member: default
4. SRR7161124.4 SRS3297852
name: 2:1101:423.80:99.70:Y, member: default
5. SRR7161124.5 SRS3297852
name: 2:1101:535.30:100.00:Y, member: default
6. SRR7161124.6 SRS3297852
name: 2:1101:973.70:99.90:Y, member: default
7. SRR7161124.7 SRS3297852
name: 2:1101:15.20:108.60:Y, member: default
8. SRR7161124.8 SRS3297852
name: 2:1101:52.50:102.20:Y, member: default
9. SRR7161124.9 SRS3297852
name: 2:1101:56.90:120.60:Y, member: default

Read

>gnl|SRA|SRR7161124.1 2:1101:6.10:98.90:N *forward (Biological)*

NAAAANNAGNNNNAANCNNNNTGACCAGACANGGACNNCANTTNCNCTACT

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

human monocyte-derived macrophages (SRR7161124)

Metadata | Reads | Download

Download for Experiment SRX4080450

Accession select all	# of bases	# of spots total	# of spots filtered
☑ SRR7161124	443.8M	8.7M	

Filter

Search: Apply

What can the filter be applied to?

Download Format

☐ filtered ☐ clipped ◉ FASTA ○ FASTQ Download

GEO Datasets

NCBI Resources How To Sign in to NCBI

GEO DataSets | GEO DataSets | Homo sapiens | Search

Create alert Advanced Help

Entry type
DataSets (2,173)
Series (47,724)
Samples (1,403,392)
Platforms (5,534)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (504,388)
strain (20,649)
Customize ...

Publication dates
30 days
1 year
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

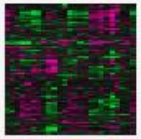
Filters: Manage Filters

Search results

Items: 1 to 20 of 1458823

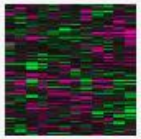
<< First < Prev Page 1 of 72942 Next > Last >>

1. Acute alcohol consumption effect on whole blood (control group): time course



Analysis of blood from subjects administered orange juice w/o alcohol. Blood collected at time points corresponding to collection times for the alcohol group in GDS4938. These results, together with those from GDS4938, provide insight into molecular response of blood during acute ethanol exposure.
Organism: **Homo sapiens**
Type: Expression profiling by array, transformed count, 5 individual, 2 protocol, 5 time sets
Platform: GPL570 Series: GSE20489 25 Samples
Download data: CEL
DataSet Accession: GDS6177 ID: 6177
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

2. MicroRNA-135b overexpression effect on prostate cancer cell line: time course



Analysis of LNCaP prostate cancer (PCa) cells overexpressing miRNA-135b for up to 36 hours. LNCaP cells express the androgen receptor (AR). MiRNA-135b overexpression in AR+ PCa cells results in slower growth compared to AR knockdown. Results provide insight into the basis of this slower growth.
Organism: **Homo sapiens**
Type: Expression profiling by array, transformed count, 2 protocol, 3 time sets
Platform: GPL10558 Series: GSE57820 12 Samples
Download data
DataSet Accession: GDS6100 ID: 6100
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of

Top Organisms [Tree]
Homo sapiens *(1409888)*
Mus musculus *(29458)*
Rattus norvegicus *(8403)*
Plasmodium falciparum *(2292)*
Macaca mulatta *(1552)*
More...

Find related data
Database: Select
Find items

Search details
```
"humans"[MeSH Terms] OR "Homo sapiens"[Organism] OR Homo sapiens[All Fields]
```
Search See more...

Recent activity
Turn Off Clear
Homo sapiens (1458823)

Series GSE107218 Query DataSets for GSE107218

Status	Public on Dec 28, 2017
Title	Developmental differences between neonatal and adult human erythropoiesis
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing

Summary

Studies of human erythropoiesis have relied, for the most part, on the in vitro differentiation of hematopoietic stem and progenitor cells (HSPC) from different sources. Here, we report that despite the common core erythroid program that exists between cord blood- and peripheral blood-HSPC induced towards erythroid differentiation in vitro, significant functional differences exist. We undertook a comparative analysis of human erythropoiesis using these two different sources of HSPC and differentiated them in vitro. We observed that cells derived from cord blood proliferate 4.5 times more than cells derived from peripheral blood. However, these cells present a delay in their differentiation pattern due to increased quantities of progenitors, notably CFU-E. Using our method of immunophenotyping for the study of erythroid progenitors, we document the presence and maintenance of a specific population in peripheral blood-derived erythroid progenitors. This population, defined as IL3R-GPA-CD34+CD36+, has the ability to form both BFU-E and CFU-E colonies in colony-forming assays, reflecting a higher potential. To further understand the differences between cord blood- and peripheral blood- HSPC, we sorted all stages of erythropoiesis from both sources and compared their transcriptome. We document differences at the CD34, BFU-E, poly- and orthochromatic stages. Among the genes presenting the highest differences in expression, many are involved in the regulation of the cell cycle and autophagy. Altogether, our studies provide a qualitative and quantitative comparative analysis of human erythropoiesis and highlight functional differences, critical to our understanding of the impact of the developmental origin of HSPCs on erythroid differentiation.

Overall

Supplementary file	Size	Download	File type/resource
GSE107218_CBPB-hg19-counts.txt.gz	3.5 Mb	(ftp)(http)	TXT

Raw data are available in SRA

Processed data is available on Series record

Contributor(s)	Yan H, Hale JP, Jaffray J, Li J, Wang Y, Huang Y, An X, Hillyer C, Wang N, Kinet S, Taylor N, Narla M, Narla A, Blanc L
Citation(s)	Yan H, Hale J, Jaffray J, Li J et al. Developmental differences between neonatal and adult human erythropoiesis. *Am J Hematol* 2018 Aug;93(4):494-503. PMID: 29274096

Submission date	Nov 21, 2017
Last update date	May 08, 2018
Contact name	John Hale
E-mail	jhale@nybloodcenter.org
Organization name	New York Blood Center
Lab	Red Cell Physiology
Street address	310 East 67th Street
City	New York
State/province	NY
ZIP/Postal code	10065
Country	USA

Platforms (1)	GPL16791 Illumina HiSeq 2500 (Homo sapiens)
Samples (24) Less...	GSM2862905 hs_PB_CD34_1
	GSM2862906 hs_PB_CD34_2
	GSM2862907 hs_PB_CD34_3
	GSM2862908 hs_PB_BFU_1
	GSM2862909 hs_PB_BFU_2
	GSM2862910 hs_PB_BFU_3
	GSM2862911 hs_PB_CFU_1
	GSM2862912 hs_PB_CFU_2
	GSM2862913 hs_PB_CFU_3
	GSM2862914 hs_PB_proerythroblast_1
	GSM2862915 hs_PB_proerythroblast_2
	GSM2862916 hs_PB_proerythroblast_3

Supplementary file	Size	Download	File type/resource
GSE107218_CBPB-hg19-counts.txt.gz	3.5 Mb	(ftp)(http)	TXT

Raw data are available in SRA

Processed data is available on Series record

Contributor(s)	Yan H, Hale JP, Jaffray J, Li J, Wang Y, Huang Y, An X, Hillyer C, Wang N, Kinet S, Taylor N, Narla M, Narla A, Blanc L
Citation(s)	Yan H, Hale J, Jaffray J, Li J et al. Developmental differences between neonatal and adult human erythropoiesis. *Am J Hematol* 2018 Aug;93(4):494-503. PMID: 29274096
Submission date	Nov 21, 2017
Last update date	May 08, 2018
Contact name	John Hale
E-mail	jhale@nybloodcenter.org
Organization name	New York Blood Center
Lab	Red Cell Physiology
Street address	310 East 67th Street
City	New York
State/province	NY
ZIP/Postal code	10065
Country	USA

Platforms (1)	GPL16791 Illumina HiSeq 2500 (Homo sapiens)
Samples (24)	GSM2862905 hs_PB_CD34_1
Less...	GSM2862906 hs_PB_CD34_2
	GSM2862907 hs_PB_CD34_3
	GSM2862908 hs_PB_BFU_1
	GSM2862909 hs_PB_BFU_2
	GSM2862910 hs_PB_BFU_3
	GSM2862911 hs_PB_CFU_1
	GSM2862912 hs_PB_CFU_2
	GSM2862913 hs_PB_CFU_3
	GSM2862914 hs_PB_proerythroblast_1
	GSM2862915 hs_PB_proerythroblast_2

Geneid	Chr	Start	End	Strand	Length	Sample1	SampleN
DDX11L1	chr1	11874	12227	+	1652	225	269

GSE107218_CBPB-hg19-counts.txt.gz	3.5 Mb	(ftp)(http)	TXT

Raw data are available in SRA

Processed data is available on Series record

Próxima aula

- Avaliar o estudo GSE107218
  - Baixar o arquivo de contagens
- Procurar estudos de interesse

Organização de dados de coortes

1. Seguimento de Pacientes com TB ativa, TB latente, Vacinados, Controles saudáveis em vários tempos pós-infecção com ou sem tratamento. Coortes de pelo menos dois países

2. Seguimento de pacientes com TB/HIV e controles (TB alone, HIV alone) com e sem tratamento

3. Seguimento de pacientes com Dengue, Dengue grave, controles, com ou sem tratamento

4. Seguimento de pacientes com Flu

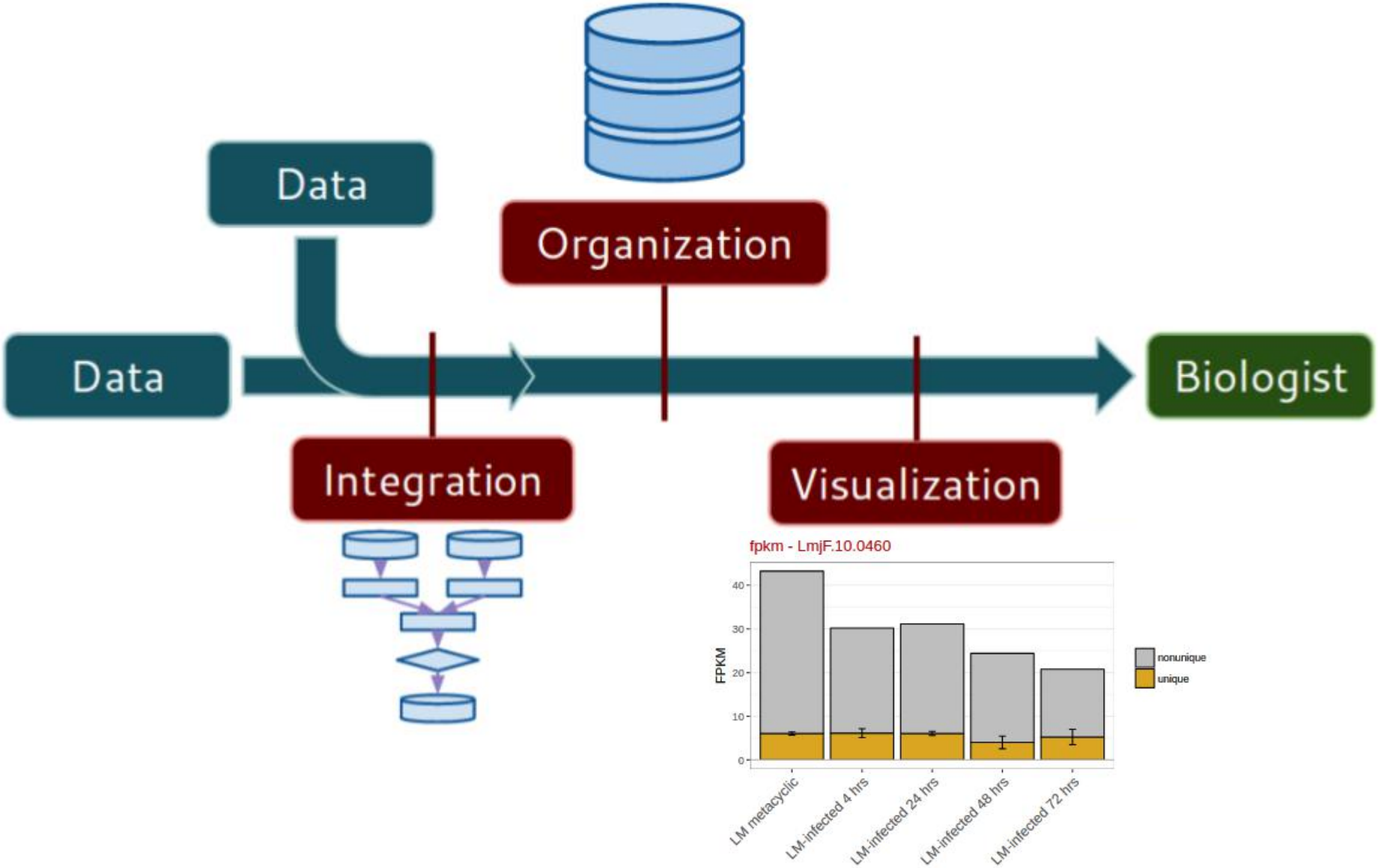
5. Seguimento de pacientes com autoimunidade - artrite reumatóide Lupus.

6. Seguimento de pacientes com Câncer de pulmão

7. Seguimento de pacientes com Sepse

8. Camundongos C57BL/6 infectados com Mtb em diferentes tempos pós infecção

Organização de dados de coortes



DÚVIDAS?