

Análisis transcriptómicos de la expresión génica

Máster Universitario en Bioinformática

Sesión 5



Universidad
Internacional
de Valencia

Dra. Paula Soler Vila
paula.solerv@professor.universidadviu.com

De:
 Planeta Formación y Universidades

Objetivos de la sesión

- 1 Preguntas y respuestas
- 2 Identificar y comprender las principales bases de datos destinadas a la información genómica/transcriptómicos.
- 3 Conocer las características del experimento de RNA-seq a procesar.
- 4 Organización de un proyecto bioinformático: Preparar nuestro entorno de trabajo y nuestra jerarquía de archivos.

Calendario de las sesiones

Julio 2024

LUN	MAR	MIÉ	JUE	VIE	SÁB	DOM
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Agosto 2024

LUN	MAR	MIÉ	JUE	VIE	SÁB	DOM
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Septiembre 2024

LUN	MAR	MIÉ	JUE	VIE	SÁB	DOM
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						



CAMBIO DE CLASE -> El día 30 de julio se trasladará al 6 de septiembre

SESIÓN 8	25/07/2024	Tema 4. Análisis estadístico de la diferencia de expresión.
SESIÓN 9	30/07/2024	
SESIÓN 10	03/09/2024	

PRIMER CONTACTO



BUZZA!

CONCURSO UNIVERSAL



PlayStation® Portable

Preguntas y respuestas

P1. Respecto a la preparación de las bibliotecas de cDNA contesta cuál de las siguientes afirmaciones es CIERTA:

1. El enriquecimiento de mRNA mediante cola de poli-A es recomendable en el caso de procariotas.
2. La fragmentación del cDNA se puede realizar con ARNasas u otros mecanismos como la sonicación.
3. Las secuencias de índice en los adaptadores del inserto son necesarias para distinguir las muestras.

P2. El nacimiento de la transcriptómica se puede considerar que ocurrió:

1. Con el descubrimiento de las enzimas polimerasas.
2. En el año de la invención de la secuenciación de Sanger (1977).
3. Con la aparición de las primeras micromatrices (*microarrays*).

P3. ¿Cuál de las siguientes frases es verdadera?

1. La secuenciación por síntesis (SBS) para lecturas cortas es la tecnología más usada hoy en día para estudios de expresión génica.
2. La secuenciación por síntesis (SBS) para lecturas largas es la tecnología más usada hoy en día para estudios de expresión génica.
3. La transcriptómica espacial es la tecnología más usada hoy en día para estudios de expresión génica.

Preguntas y respuestas

P4. En el protocolo de secuenciación por síntesis (SBS), la reacción en cadena de la polimerasa (PCR) sucede:

1. Durante la preparación de la biblioteca de cDNA.
2. En la amplificación de los haces de copias ("clusters") dentro de la celda de flujo.
3. Durante la secuenciación por síntesis con la adición de nucleótidos de terminación reversible.
4. **Todas las anteriores.**

P5. ¿Cuál de los siguientes procesos NO suceden en la celda de flujo (*flowcell*)?

1. Fragmentación.
2. Transcripción inversa.
3. **A y B.**
4. Unión de nucleótidos de terminación reversible.

P6. La secuenciación de RNA de segunda generación o RNA-seq ...

1. Es la secuenciación directa de moléculas de RNA.
2. Tiene una variabilidad biológica cercana a 0.
3. **Requiere de una integridad del RNA mayor a 7 RIN en mamíferos.**

Preguntas y respuestas

P7. En una secuenciación de ADN se generan 1 millón de lecturas "paired-end" (lecturas pareadas). ¿Qué información se puede inferir de esta cantidad de lecturas?

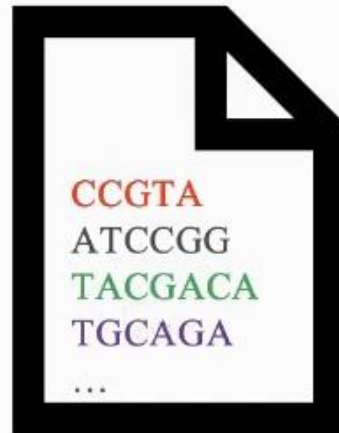
1. Se han secuenciado 1 millón de fragmentos de ADN
2. Se han obtenido 1 millón de lecturas en el R1 y un millón de lecturas en el R2
3. Se han obtenido 1 millón de extremos emparejados.
4. **Todas son correctas**

Sample1_R1.fastq

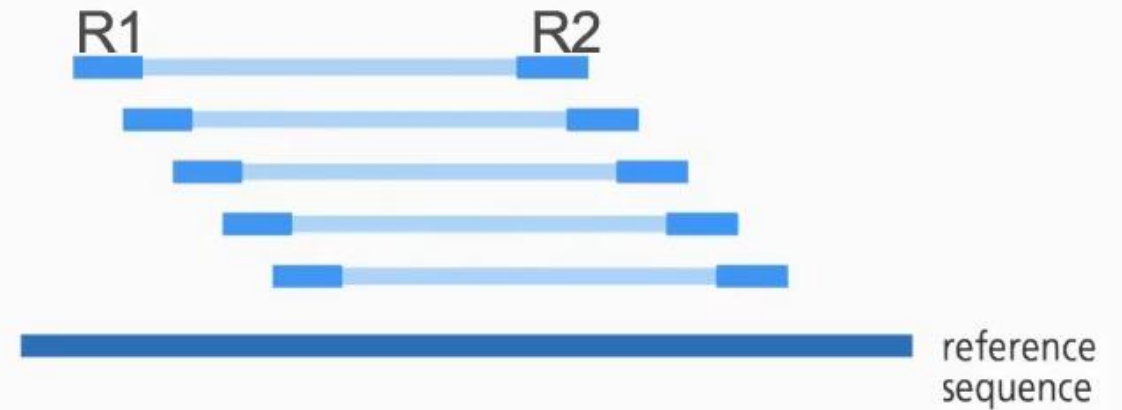


Fragment 1
Fragment 2
Fragment 3
Fragment 4
...


Sample1_R2.fastq



Paired-end reads



Pregunta 7



ENA
European Nucleotide Archive

Search 🔍

Examples: histone, BN000085

View 👁

Examples: Taxon:9606, BN000085, PRJEB402

Home

Submit ▾

Search ▾

Rulespace

About ▾

Support ▾

Download All

Scientific Name	Library Layout	Library Source	Library Selection	Read Count	Generated FASTQ files: FTP
Anisolabis maritima	PAIRED	TRANSCRIPTOMIC	PolyA	43,781,456	<div><input type="checkbox"/> SRR1552489_1.fastq.gz</div> <div><input type="checkbox"/> SRR1552489_2.fastq.gz</div>

```
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Downloads]$ zcat SRR1552489_1.fastq.gz | echo $((`wc -l`/4))
43781456
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Downloads]$ zcat SRR1552489_2.fastq.gz | echo $((`wc -l`/4))
43781456
```


Objetivos de la sesión

- 1 Preguntas y respuestas
- 2 Identificar y comprender las principales bases de datos destinadas a la información genómica/transcriptómicos.
- 3 Conocer las características del experimento de RNA-seq a procesar.
- 4 Organización de un proyecto bioinformático: Preparar nuestro entorno de trabajo y nuestra jerarquía de archivos.

¿Qué es y para qué sirve una base de datos?



```
install.packages("wordcloud")  
pip install wordcloud
```


¿Cuántos datos se
recopilaron cada
minuto del día
durante el **2023**?



¿Qué tenemos que tener en cuenta?

- Depósito de los datos y metadatos.
- Comprobación de los datos (“*curators*”). Manual o automatizada.
- Actualización de los datos.
- Tratamiento de errores o de datos inciertos.
- Consistencia con otros datos de la misma naturaleza.
- Formatos.



Data submission



Curators
clean and fix



Standardised
files (IDF, SDRF)



FINDABLE

Persistent
Identifiers (PIDs)

iD

Rich metadata



Indexed data
repositories



PIDs in metadata



ACCESSIBLE

Standard
communications
protocol



Open, free protocol



Authentication,
where necessary



Metadata is always
available



INTER- OPERABLE

Interoperable



Vocabularies



Vocabularies are
FAIR



Linked metadata



REUSABLE

Metadata have
multiple attributes



Usage license



Provenance



Community
standards



**“Tan abierto como sea
posible, tan cerrado como sea
necesario”**

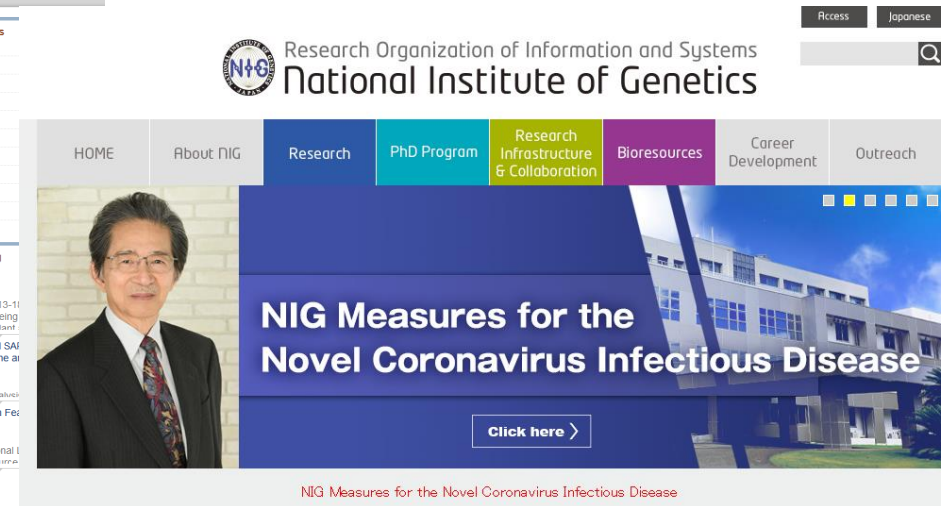
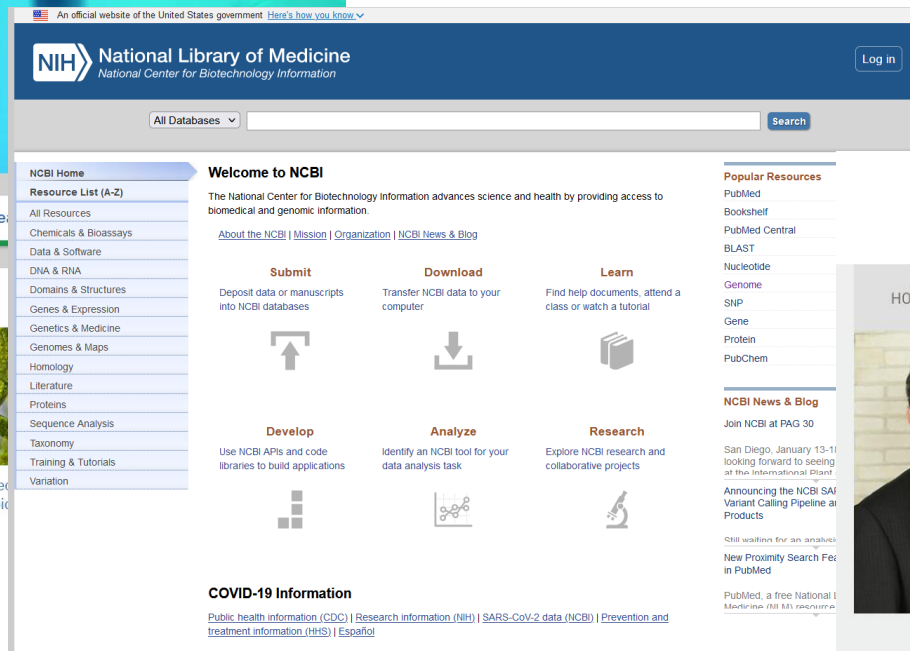
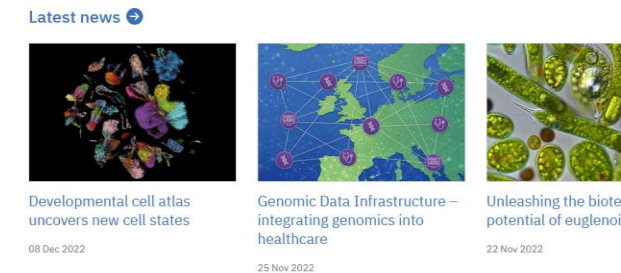
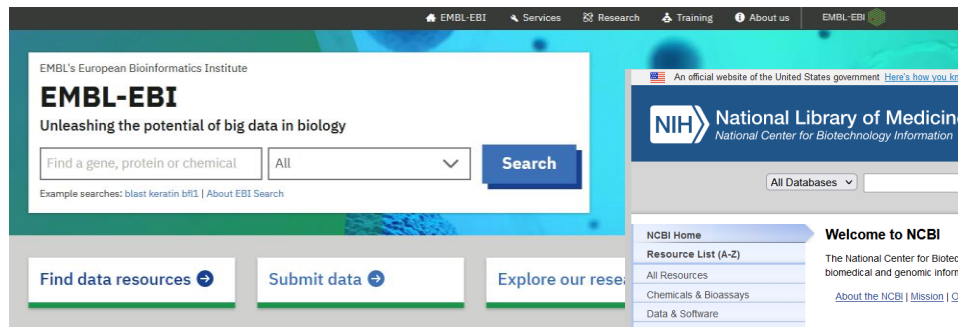
FINDA



SABLE

Principales consorcios y centros científicos

- **EMBL-EBI** (Laboratorio Europeo de Biología Molecular - Instituto Europeo de Bioinformática)
- **NCBI** (Centro Nacional para la Información Biotecnológica de los Estados Unidos)
- **NIG** (Instituto Nacional de Genética de Japón)



EMBL-EBI (Laboratorio Europeo de Biología Molecular - Instituto Europeo de Bioinformática)

- **ENA** (*European Nucleotide Archive*)
- **ArrayExpress**
- **Expression atlas**

The screenshot shows the ENA website interface. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search icon. Below this is the ENA logo and the text 'European Nucleotide Archive'. A search bar is present with 'Enter text search terms' and 'Enter accession' fields. A banner below the search bar recommends subscribing to the ENA-announce mailing list. A yellow box contains information about SARS-CoV-2 data submissions, advising users to contact virus-dataflow@ebi.ac.uk for specific advice and mentioning a Drag-and-Drop Data Submission Service. The main content area is titled 'European Nucleotide Archive' and describes the archive's purpose. It lists access methods: browser, search tools, large scale file download, and API. There are four large buttons: Submit, Search, Rulespace, and Support. A 'Latest ENA news' section includes two articles: 'ENA: Improving spatio-temporal annotations Dec 1, 2021, 1:00:00 AM' and 'Retirement of old ENA Browser on 5th August 2020 Jul 16, 2020, 2:00:00 AM'. A 'Tweets from @ENASequences' section is also visible.

The screenshot shows two websites side-by-side. The top website is 'Expression Atlas', which displays 'Gene expression across species and biological conditions'. It has a search bar and a 'Query single cell expression' button. The bottom website is 'Single Cell Expression Atlas', which displays 'Single cell gene expression across species'. It also has a search bar and a 'Query bulk expression' button. Both websites show search results for 'Gene set enrichment' and 'Gene / Gene properties'. The Expression Atlas results show 'Search across 65 species, 4,315 studies, 153,212 assays'. The Single Cell Expression Atlas results show 'Search across 20 species, 304 studies, 10,453,716 cells'. Both websites have a navigation bar with links to Home, Browse experiments, Download, Release notes, FAQ, Help, Licence, About, and Support.

Ejercicio I: Explorando resultados con Expression Atlas

Acceda al **Expression Atlas** (<https://www.ebi.ac.uk/gxa/home>) para explorar los resultados de la expresión génica de un determinado experimento. Para ello, sitúese en la opción *Browse experiments* e indique en el buscador la siguiente información acerca del experimento de interés:


- **Species: Homo Sapiens**
- **Title: Primary Human Airway Epithelial Cultures infected with SARS-CoV-2**

En el expression atlas es recomendable mirar una unidad abajo o arriba de lo que se nos pide, porque muchas veces no lo coge como mayor o igual qué , entoces no entran los valores iguales

- ¿Qué técnica experimental se ha empleado para obtener el valor de expresión de los genes?
- ¿Cuáles son los genes que se encuentran sobreexpresados (*upregulated*) con un valor de *log2-fold change* mayor o igual a 3?.
- ¿Cuáles son los genes que se encuentran infraexpresados (*downregulated*) con un valor de *log2-fold change* menor o igual a -2.5?

Ejercicio II: Extrayendo FASTQ del ENA

¿Cuántos archivos en formato **fastq** podemos encontrar en este estudio (**PRJNA258286**)?



European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

Examples: histone, BN000065

Examples: Taxon:9606, BN000065, PRJEB402

Project: PRJNA258286

To identify genes specifically expressed in lactating mammary glands, the gene expression profiles of luminal and basal cells from different developmental stages were compared. Overall design: Comparison of gene expression in luminal and basal cells harvested from the mammary glands of virgin, 18.5 day pregnant and 2 day lactating mice (2 mice per stage).

Organism:	Mus musculus (house mouse)
Secondary Study Accession:	SRP045534
Study Title:	Transcriptome analysis of luminal and basal cell subpopulations in the lactating versus pregnant mammary gland Show Less
Center Name:	Smyth, Bioinformatics, Walter and Eliza Hall Institute of Medical Research
Study Name:	Mus musculus
ENA-REFSEQ:	N
PROJECT-ID:	258286
ENA-FIRST-PUBLIC:	2015-01-21
ENA-LAST-UPDATE:	2023-05-19

View:

XML

XML (STUDY)

Download:

XML

XML (STUDY)

Navigation:

Show

Read Files:

Hide

Publications:

Show

Parent Projects:

Show

Related ENA Records:

Show

<https://www.ebi.ac.uk/ena/browser/home>

Obteniendo nuestro dataset de trabajo: PRJNA258286

Download report: JSON TSV

Get download script

Download selected files

Download All

Study Accession	Experiment Accession	Run Accession	Scientific Name	Library Layout	Library Strategy	Library Selection	Generated FASTQ files FTP
PRJNA258286	SRX681988	SRR1552447	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552447.fastq.gz
PRJNA258286	SRX681994	SRR1552453	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552453.fastq.gz
PRJNA258286	SRX681989	SRR1552448	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552448.fastq.gz
PRJNA258286	SRX681991	SRR1552450	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552450.fastq.gz
PRJNA258286	SRX681992	SRR1552451	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552451.fastq.gz
PRJNA258286	SRX681985	SRR1552444	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552444.fastq.gz
PRJNA258286	SRX681986	SRR1552445	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552445.fastq.gz
PRJNA258286	SRX681987	SRR1552446	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552446.fastq.gz
PRJNA258286	SRX681990	SRR1552449	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552449.fastq.gz
PRJNA258286	SRX681993	SRR1552452	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552452.fastq.gz
PRJNA258286	SRX681995	SRR1552454	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552454.fastq.gz
PRJNA258286	SRX681996	SRR1552455	Mus musculus	SINGLE	RNA-Seq	cDNA	<input type="checkbox"/> SRR1552455.fastq.gz

1

SRR1552444

<https://www.ebi.ac.uk/ena/browser/view/PRJNA258286>

Conociendo nuestro *dataset* de trabajo

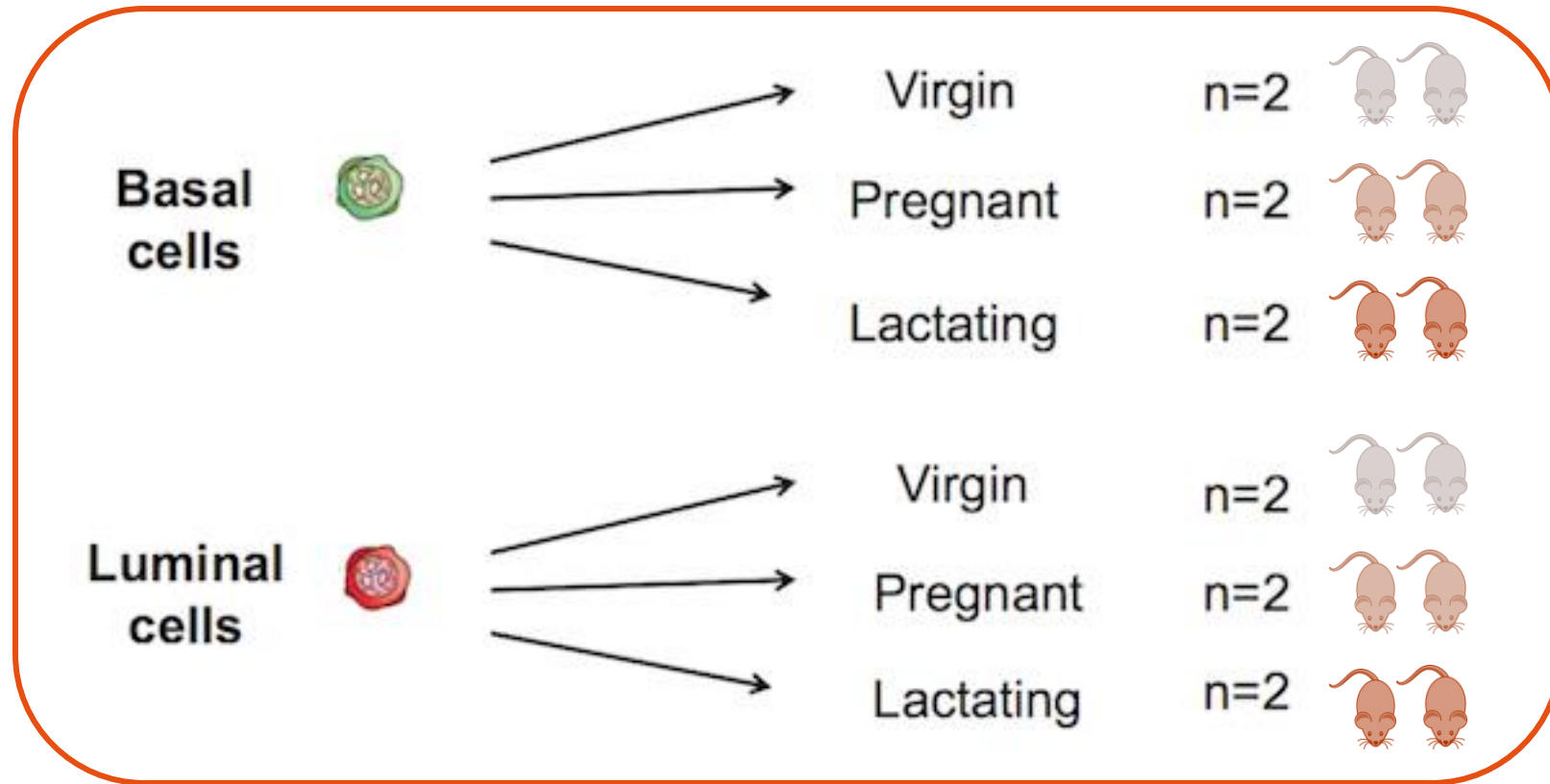
RNA-seq of mouse mammary gland

**BASE DEL ANÁLISIS
COMPUTACIONAL**

**BASE DE LAS ACTIVIDADES DEL
PORTAFOLIO**

Conociendo nuestro *dataset* de trabajo

RNA-seq of mouse mammary gland



Fu, Nai Yang, et al. "EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival." *Nature Cell Biology*

Conociendo nuestro *dataset* de trabajo

nature
cell biology

ARTICLES

EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival

Nai Yang Fu^{1,2}, Anne C. Rios^{1,2,10}, Bhupinder Pal^{1,2,10}, Rina Soetanto³, Aaron T. L. Lun^{2,4}, Kevin Liu^{1,2}, Tamara Beck^{1,2}, Sarah A. Best^{1,2}, François Vaillant^{1,2}, Philippe Bouillet^{2,5}, Andreas Strasser^{2,5}, Thomas Preiss^{3,6}, Gordon K. Smyth^{4,7}, Geoffrey J. Lindeman^{1,8,9,10} and Jane E. Visvader^{1,2,10,11}

Expansion and remodelling of the mammary epithelium requires a tight balance between cellular proliferation, differentiation and death. To explore cell survival versus cell death decisions in this organ, we deleted the pro-survival gene *Mcl-1* in the mammary epithelium. *Mcl-1* was found to be essential at multiple developmental stages including morphogenesis in puberty and alveologenesis in pregnancy. Moreover, *Mcl-1*-deficient basal cells were virtually devoid of repopulating activity, suggesting that this gene is required for stem cell function. Profound upregulation of the Mcl-1 protein was evident in alveolar cells at the switch to lactation, and *Mcl-1* deficiency impaired lactation. Interestingly, *EGF* was identified as one of the most highly upregulated genes on lactogenesis and inhibition of EGF or mTOR signalling markedly impaired lactation, with concomitant decreases in Mcl-1 and phosphorylated ribosomal protein S6. These data demonstrate that Mcl-1 is essential for mammapoiesis and identify EGF as a critical trigger of Mcl-1 translation to ensure survival of milk-producing alveolar cells.

Biblioteca Campus VIU

Referencia

Fu, Nai Yang, et al. "EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival." *Nature Cell Biology*



- **GEO** (*Gene Expression Omnibus*)
- **SRA** (*Sequence Read Archive*)

Fastq-dump

`fastq-dump` is a tool for downloading sequencing reads from [NCBI's Sequence Read Archive \(SRA\)](#). These sequence reads will be downloaded as FASTQ files. How these FASTQ files are formatted depends on the `fastq-dump` options used.




<https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump>

The screenshot shows the NCBI Gene Expression Omnibus (GEO) homepage. The header includes the NCBI logo, navigation links (Resources, How To), and a sign-in button. The main content area is titled "Gene Expression Omnibus" and describes it as a public functional genomics data repository. Below this, there are three main sections: "Getting Started" (Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, How to Download Data), "Tools" (Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, Studies with Genome Data Viewer Tracks, Programmatic Access, FTP Site, ENCODE Data Listings and Tracks), and "Browse Content" (Repository Browser, DataSets: 4348, Series: 193207, Platforms: 24717, Samples: 5521698). At the bottom, there is an "Information for Submitters" section with links for Login to Submit, Submission Guidelines, Update Guidelines, MIAME Standards, Citing and Linking to GEO, Guidelines for Reviewers, and GEO Publications.


The screenshot shows the NIH SRA website. The header includes the NIH logo, the text "National Library of Medicine National Center for Biotechnology Information", and a login button. Below the header, there is a search bar with "SRA" selected and a "Search" button. A banner below the search bar reads "SRA - Now available on the cloud" and describes the SRA as the largest publicly available repository of high throughput sequencing data. Below the banner, there are three main sections: "Getting Started" (Documentation, How to submit, How to search and download, How to use SRA in the cloud, Submit to SRA), "Tools and Software" (Download SRA Toolkit, SRA Toolkit Documentation, SRA-BLAST, SRA Run Browser, SRA Run Selector), and "Related Resources" (Submission Portal, dbGaP Home, BioProject, BioSample).

- **DDBJ**


 [Services](#) [SuperComputer](#) [Statistics](#) [Activities](#) [About Us](#)

DDBJ Web Sites [Terms](#) [Contact](#) [Japanese](#)

[Suspension of the BI-DDBJ activity during the New Year Holidays](#)




Bioinformation and DDBJ Center provides sharing and analysis services for data from life science researches and advances science.




Services

Search, analysis, database services of DDBJ Center



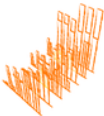
Submission

Navigation for how to submit your data




Super Computer

NIG Supercomputer



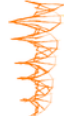
Statistics

Statistics of DDBJ Center services



Activities

Training sessions and achievements of DDBJ Center



About us

About Bioinformation and DDBJ Center

NEWS

DDBJ staff visit to KOBIC in Korea
2022/12/13 [Announcement](#) [DDBJ Center](#)

Read length and direction of paired reads were made optional in the DRA submission
2022/12/12 [Announcement](#) [DRA](#) [DDBJ Center](#)

Suspension of the BI-DDBJ activity during the New Year Holidays
2022/12/12 [Announcement](#) [DDBJ](#) [BioProject](#) [BioSample](#) [DRA](#) [GEA](#) [JGA](#) [AGD](#) [DDBJ Center](#)

Updated tools related to Mass Submission System (MSS)
2022/12/09 [Announcement](#) [DDBJ](#) [DDBJ Center](#)

DDBJ services have been resumed
2022/12/09 [Maintenance](#) [DDBJ](#) [BioProject](#) [BioSample](#) [DRA](#) [GEA](#) [JGA](#) [AGD](#) [DDBJ Center](#)

[more](#)

<https://www.ddbj.nig.ac.jp/index-e.html>

International Nucleotide Sequence Database Collaboration

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between [DDBJ](#), [EMBL-EBI](#) and [NCBI](#).

INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.



Objetivos de la sesión

1

Preguntas y respuestas

2

Identificar y comprender las principales bases de datos destinadas a la información genómica/transcriptómicos.

3

Conocer las características del experimento de RNA-seq a procesar.

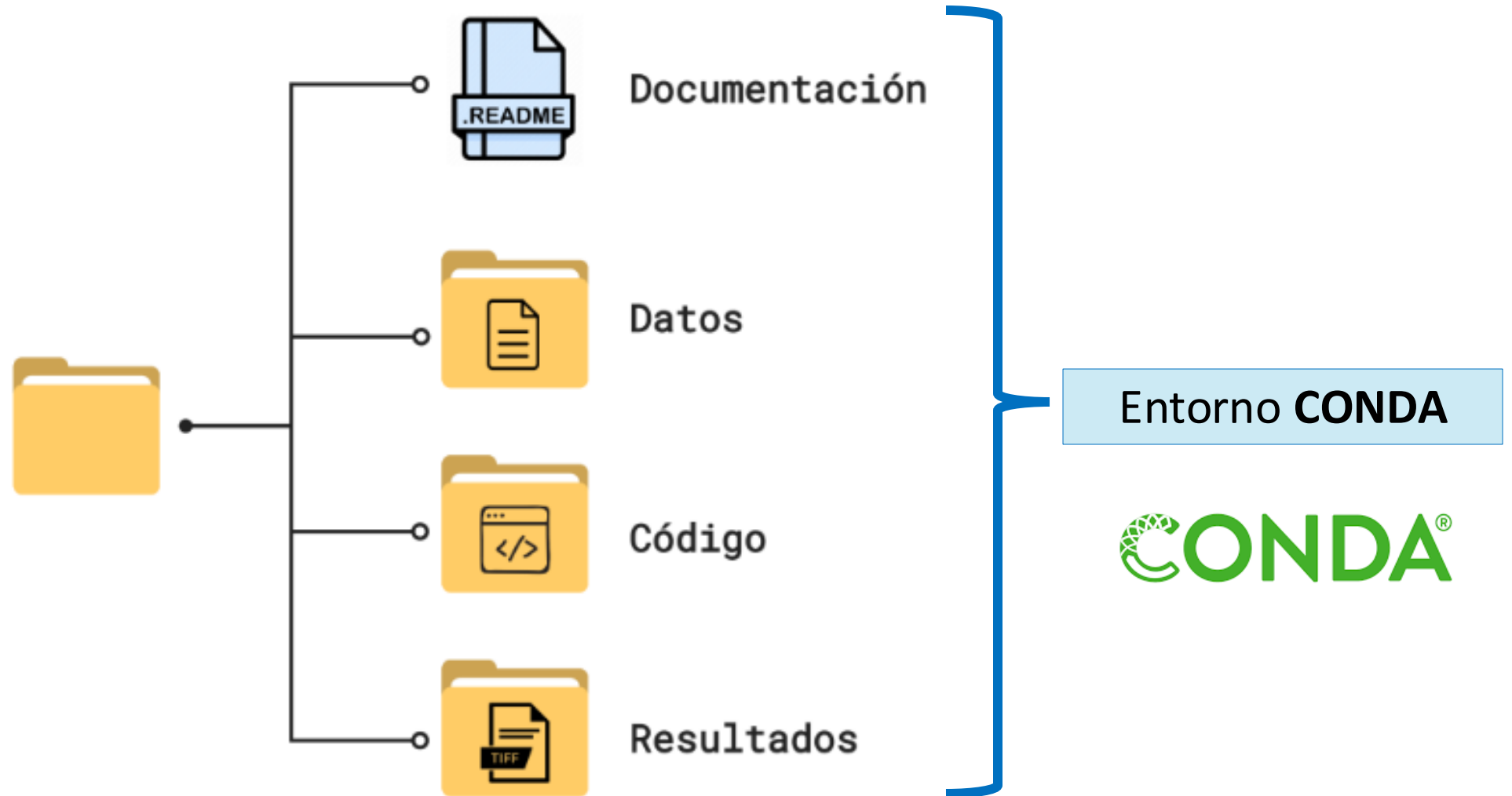
4

Organización de un proyecto bioinformático: Preparar nuestro entorno de trabajo y nuestra jerarquía de archivos.

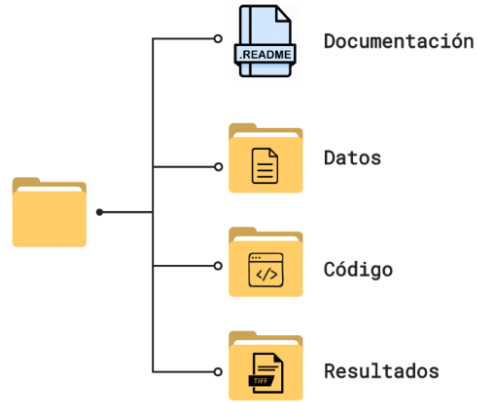


Organización de un proyecto bioinformático

Organización de directorios



Preparación de la jerarquía de archivos



```
[UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Proyecto_JULIO_2024]$ tree -L 3
```

```
├── Code
├── Data
├── Results
└── README
```

Organización de directorios: Documentación



Documentación

* "Leeme".

- Un archivo de texto simple plano (no Word)

- Qué hay dentro del repositorio (y cada uno de sus directorios).
- Qué hacen cada una de las funciones/scripts del repositorio
- Cómo y en qué orden deben ocuparse los scripts para realizar los análisis

```
3Berberis_phylogeog_README.mdown
README
=====

Contains scripts for paralogous loci filtering, output data from the *populations* program
of *Stacks*, as well as R scripts used for analyses and plotting.

Scripts and custom functions
=====

Directory `3Berberis_phylogeog/bin` contains the scripts (numbered) and R functions (not
numbered, called from within the scripts) used for data analysis and plotting.

**1.PopSamples_PostCleaning.r:** filters data to keep only those samples having more than
50% of the mean number of loci per sample, and only those loci present in at least 80% of
the barcoded sample

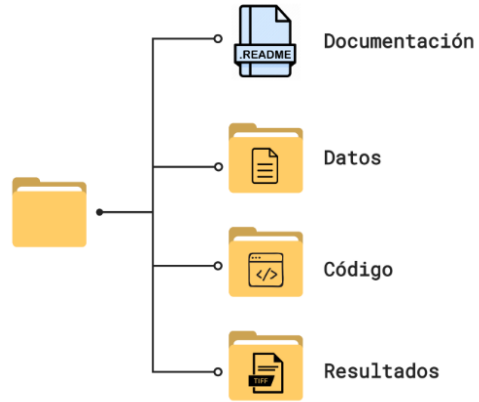
**2.PopSamples_Whitelists-StacksPopulations.script:** produces whitelists and populations
maps to run Stacks populations program including all loci (no paralogous filtering)

**3.PopSamples_excluding_paralogs.r:** uses Stacks populations summary stats output to
identify potential paralog loci. Output arelist of all potential paralogous loci (`.docs/
lociP05`) and potential paralogs within Berberis alpina (`.docs/potentialparalogs`).

**4.StacksPopulations_AllLoci.script:** creates a whitelist file of loci and populations
maps the for subset of samples to analyze. Then runs the *populations* program of *Stacks*
using the lists of putatively paralogous loci and any loci where p=0.5 as blacklists.
Output is in `data.out/PopSamples_m3`.

**4.StacksPopulations_EQsamsize.script:** creates a Poulation Map for a subset of
samples of equal sampling size for B. alpina, Zamorano and B. moranensis and runs the
*populations* program from *Stacks*. Output is in `data.out/PopSamples_m3/`
```

Preparación de la jerarquía de archivos



```
[UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Proyecto_JULIO_2024]$ tree -L 3
```

```
├── Code
├── Data
├── Results
├── README
├── METADATA
```

Organización de directorios: Metadatos

Proyecto_JULIO_2024



Raw data -> SRR1552444.fastq.gz



Metadata



05MBIF.yml



Botón derecho -> Guardar enlace como...

FileName	SampleName	CellType	Status
MCL1.DG_BC2CTUACXX_ACTTGA_L002_R1	MCL1.DG	basal	virgin
MCL1.DH_BC2CTUACXX_CAGATC_L002_R1	MCL1.DH	basal	virgin
MCL1.DI_BC2CTUACXX_ACAGTG_L002_R1	MCL1.DI	basal	pregnant
MCL1.DJ_BC2CTUACXX_CGATGT_L002_R1	MCL1.DJ	basal	pregnant
MCL1.DK_BC2CTUACXX_TTAGGC_L002_R1	MCL1.DK	basal	lactate
MCL1.DL_BC2CTUACXX_ATCACG_L002_R1	MCL1.DL	basal	lactate
MCL1.LA_BC2CTUACXX_GATCAG_L001_R1	MCL1.LA	luminal	virgin
MCL1.LB_BC2CTUACXX_TGACCA_L001_R1	MCL1.LB	luminal	virgin
MCL1.LC_BC2CTUACXX_GCCAAT_L001_R1	MCL1.LC	luminal	pregnant
MCL1.LD_BC2CTUACXX_GGCTAC_L001_R1	MCL1.LD	luminal	pregnant
MCL1.LE_BC2CTUACXX_TAGCTT_L001_R1	MCL1.LE	luminal	lactate
MCL1.LF_BC2CTUACXX_CTTGTA_L001_R1	MCL1.LF	luminal	lactate

Organización de directorios: Metadatos



Documentación

Evitar espacios en blanco, unir palabras con guion bajo

No usar acentos, ni caracteres ajenos a los ingleses (ñ ç)

1	Timestamp	SAMPLE_ID	CELL_TYPE	PRE_TREATMENT	PRE_TREATMENT	TREATMENT_TIME	CONTR	EXPERIMENT_ID	HIC	SEQUE	SEQUE	SEQUENCIN	SEND_FOR_SEQUENCING_ON
2	8/3/2015 11:59:52	eaed0e91e_2a3b69a85	T47D	Untreated	0 Red medium	0 No		1	1	1	1	CRG	8/3/2015
3	8/3/2015 14:51:31	5733f3071_2a3b69a85	T47D	Untreated	0 FBSdes	0 No		1	1	1	1	CRG	8/3/2015
4	8/10/2015 14:13:44	dc3a1e069_51720e9cf	T47D	Untreated	0 Untreated	0 No		1	1	1	1	CRG	8/1/2015
5	8/10/2015 14:22:09	a1e46328c_51720e9cf	T47D	Untreated	0 Progesterone	15 No		1	1	1	1	CRG	8/1/2015
6	8/10/2015 14:25:17	9a7c4a68d_51720e9cf	T47D	Untreated	0 Progesterone	30 No		1	1	1	1	CRG	8/1/2015
7	8/10/2015 14:35:30	b1913e6c1_51720e9cf	T47D	Untreated	0 Progesterone	60 No		1	1	1	1	CRG	8/1/2015
8	8/10/2015 14:37:07	e4aa8d54f_51720e9cf	T47D	Untreated	0 Progesterone	180 No		1	1	1	1	CRG	8/1/2015
9	8/10/2015 14:38:24	dc3a1e069_ec92aa0bb	T47D	Untreated	0 Untreated	0 No		1	2	1	1	CRG	8/1/2015
10	8/10/2015 14:40:10	d4e07ef90_51720e9cf	T47D	Untreated	0 Estrogen	15 No		1	1	1	1	CRG	8/1/2015
11	8/10/2015 14:41:42	7824bad60_51720e9cf	T47D	Untreated	0 Estrogen	30 No		1	1	1	1	CRG	8/1/2015
12	8/10/2015 14:43:16	ab7537068_51720e9cf	T47D	Untreated	0 Estrogen	60 No		1	1	1	1	CRG	8/1/2015
13	8/10/2015 14:44:58	1618f57a8_51720e9cf	T47D	Untreated	0 Estrogen	180 No		1	1	1	1	CRG	8/1/2015
14	9/18/2015 15:32:19	b7fa2d8db_d5f36599c	B-cell	Untreated	0 Untreated	0 No		100	1	1	1	4DGU	9/18/2015
15	9/18/2015 15:33:40	b7fa2d8db_f93db662e	B-cell	Untreated	0 Untreated	0 No		101	1	1	1	4DGU	9/18/2015
16	9/18/2015 15:36:09	fc3e8b36a_95f36c69d	ES-cell	Untreated	0 Untreated	0 No		102	1	1	1	4DGU	9/18/2015
17	9/18/2015 15:40:41	fc3e8b36a_da061e9a0	ES-cell	Untreated	0 Untreated	0 No		103	1	1	1	4DGU	9/18/2015
18	9/18/2015 15:42:49	e6588f786_56e7062b9	Macrophage	Untreated	0 Untreated	0 No		104	1	1	1	4DGU	9/18/2015
19	9/18/2015 15:44:42	e6588f786_3e1166f1f	Macrophage	Untreated	0 Untreated	0 No		105	1	1	1	4DGU	9/18/2015

Identificadores únicos de la misma longitud

No unir celdas, repetir los datos las veces que sea necesario

No tener celdas vacías, usar NA para 'No Aplica' o ND para 'No Determinado'

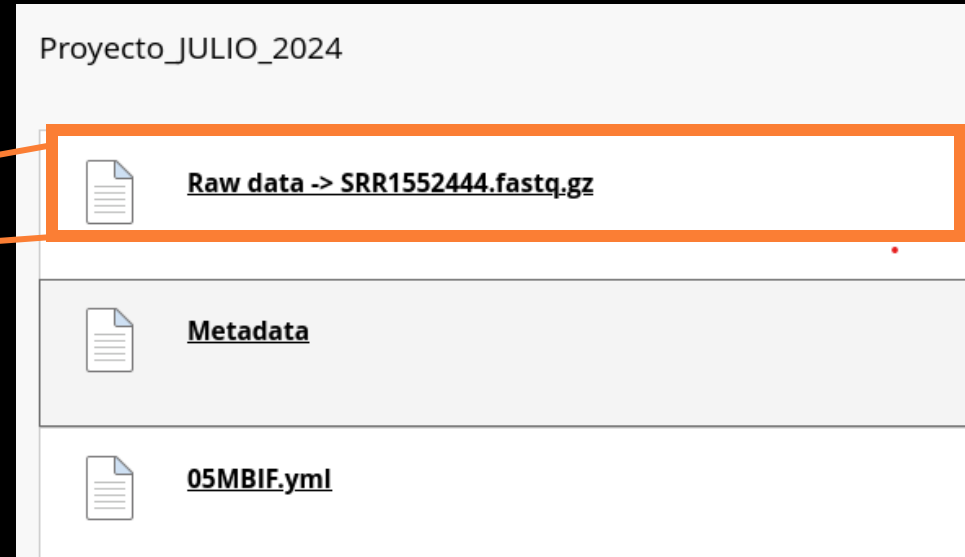
Preparación de la jerarquía de archivos



Datos

```
[UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Proyecto_JULIO_2024]$ tree -L 3
```

```
|— Code
|— Data
|   |— Annotation
|   |— Processed
|   |— Raw
|   |— Reference_genome
|— README
|— METADATA
|— Results
```



Preparación de la jerarquía de archivos



Datos

```
[UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Proyecto_JULIO_2024]$ tree -L 3
```

```
|— Code
|— Data
|   |— Annotation
|   |— Processed
|   |— Raw
|       |— SRR1552444.fastq.gz (27.919.481 lecturas crudas)
|       |— Reference_genome
|— README
|— METADATA
|— Results
```

Organización de directorios: Código



Código

- Enumerar los scripts de acuerdo al orden en que serán ejecutados

```
+++ 1.QualityControl.sh  
+++ 2.Trimming.sh  
+++ 3.ReadAlignment.sh  
+++ 4.DifferentialExpression.R
```

- Utilizar rutas relativas hacia los archivos *input*

```
fastqc ../Data/*.fastqc -o ../Results/
```

- Comentar el script -> facilita que sea entendido por humanos

```
#Change the location of bam files  
mv ../Data/*.bam ../Results/
```

90% of all code comments:



¿Por qué no se comparte el código?



Código

1. Me da vergüenza que vean mi código
2. No quiero que otros saquen provecho de mi código, me pertenece o a mi institución
3. Otros no publican su código, ¿por qué yo sí?
4. Me da pereza pulir mi código para publicarlo
5. Si publico mi código le van a encontrar errores y demandar correcciones o ayuda

Barnes, N. **Publish your computer code: it is good enough**. Nature 467, 753 (2010). <https://doi.org/10.1038/467753a>

Jerarquía de archivos

```
[UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Proyecto_JULIO_2024]$ tree -L 3
```

```
|— Code
|  |— 05MBIF.yml
|— Data
|  |— Annotation
|  |— Processed
|  |— Raw
|      └─ SRR1552444.fastq.gz
|  └─ Reference_genome
|— README
|— METADATA
└─ Results
```

Proyecto_JULIO_2024



Raw data -> SRR1552444.fastq.gz



Metadata



05MBIF.yml

Activación del entorno de trabajo

```
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Code]$ head 05MBIF.yml
```

```
name: 05MBIF
```

```
channels:
```

- bioconda
- conda-forge
- defaults

```
dependencies:
```

- _libgcc_mutex=0.1=main
- _openmp_mutex=4.5=1_gnu
- argon2-cffi=20.1.0=py39h27cfd23_1
- async_generator=1.10=py_0

```
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Code]$ conda env create -f 05MBIF.yml
```

```
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Code]$ conda activate 05MBIF
```

```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Code]$ conda env list
```

```
# conda environments:
```

```
#
```

```
base          /home/paula.soler/miniconda3
```

```
05MBIF        * /home/paula.soler/miniconda3/envs/05MBIF
```



Organización de directorios



Resultados

```
|— Code
|   |— 05MBIF.yml
|— Data
|   |— Annotation
|   |— Processed
|   |— Raw
|       |— SRR1552444.fastq.gz
|   |— Reference_genome
|— README
|— METADATA
|— Results
```



Resultados

Tablas

Gráficos

Figuras

Referencias

Organización del **AWS** (recordad que **AWS** no es un lugar de almacenamiento)



aws



viu

Universidad
Internacional
de Valencia

universidadviu.com

De:
 Planeta Formación y Universidades