

CLASE 8

Configuración proyectos de genómica aplicada

DBT 792 GENÉTICA Y GENÓMICA EN PRODUCCIÓN ANIMAL

**Profesor
Dr. José Gallardo**

PLAN DE LA CLASE

Introducción

- Investigación reproducible.
- Etapas de un proyecto de genómica.
- Bases de datos genómicos.
- Herramientas computacionales para trabajar en un proyecto de genómica.

Práctica

- a) Iniciar proyecto de genómica aplicada en Github.
- b) Familiarizarse con bases de datos genómicas.

INVESTIGACIÓN REPRODUCIBLE

PREGUNTAS AL CURSO

¿Qué problemas has tenido cuando tienes que elaborar un reporte o documento entre varios autores?

¿Qué problemas has tenido cuando quieres rehacer un análisis de datos a partir de un set de datos antiguo o en el que no has trabajado por mucho tiempo?

PREGUNTAS AL CURSO

Responda de 1 a 5, donde 1 es totalmente en desacuerdo y 5 es totalmente de acuerdo.

A- ¿Están disponibles de forma pública los datos crudos de su tesis de pregrado?.

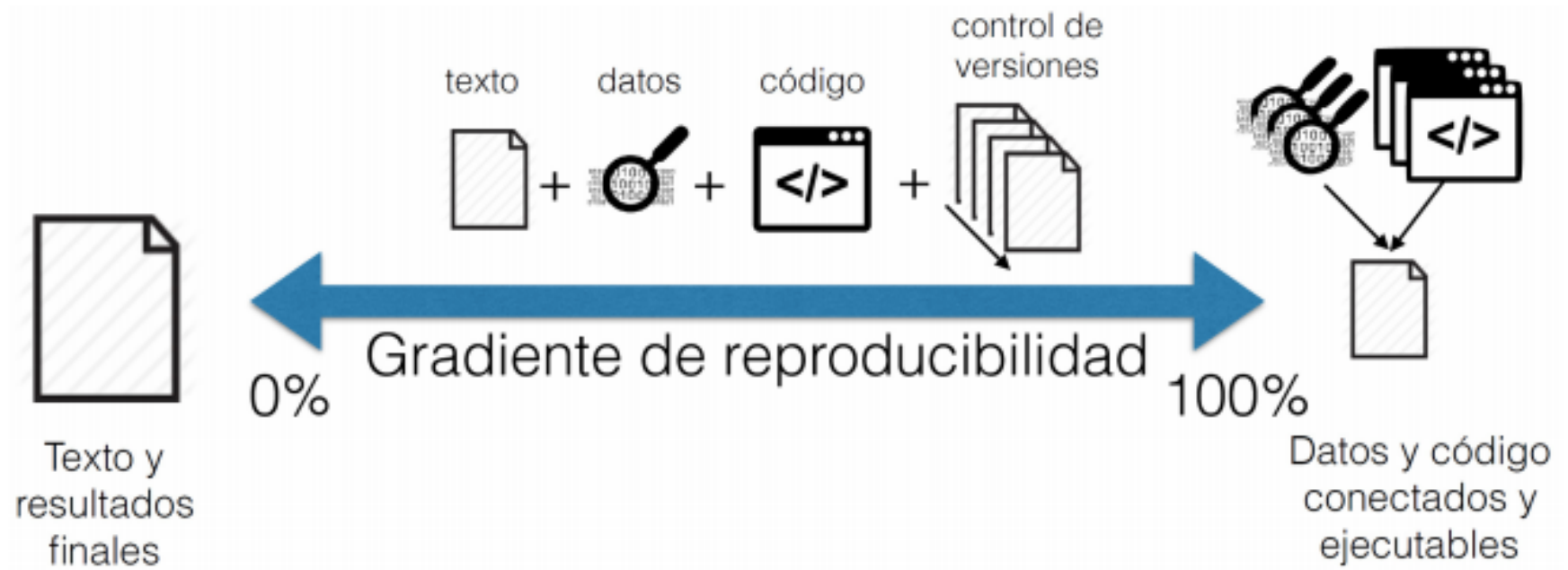
B- ¿Están disponibles de forma pública los métodos de su tesis de pregrado?.

C- ¿Los análisis estadísticos de su tesis de pregrado están codificados con algún lenguaje de programación de código abierto como R, Python u otro similar?.

D- ¿Sería usted capaz de rehacer hoy las tablas y gráficas de su tesis de pregrado a partir de sus propios datos?.

INVESTIGACIÓN REPRODUCIBLE

Investigación reproducible implica que desde los mismos datos y códigos se generarán los mismos resultados.



Peng. 2011. Science 334 (6060). Sánchez et al. 2016 Ecosistemas 25(2): 83-92

ALGUNOS CRITERIOS DE REPRODUCIBILIDAD

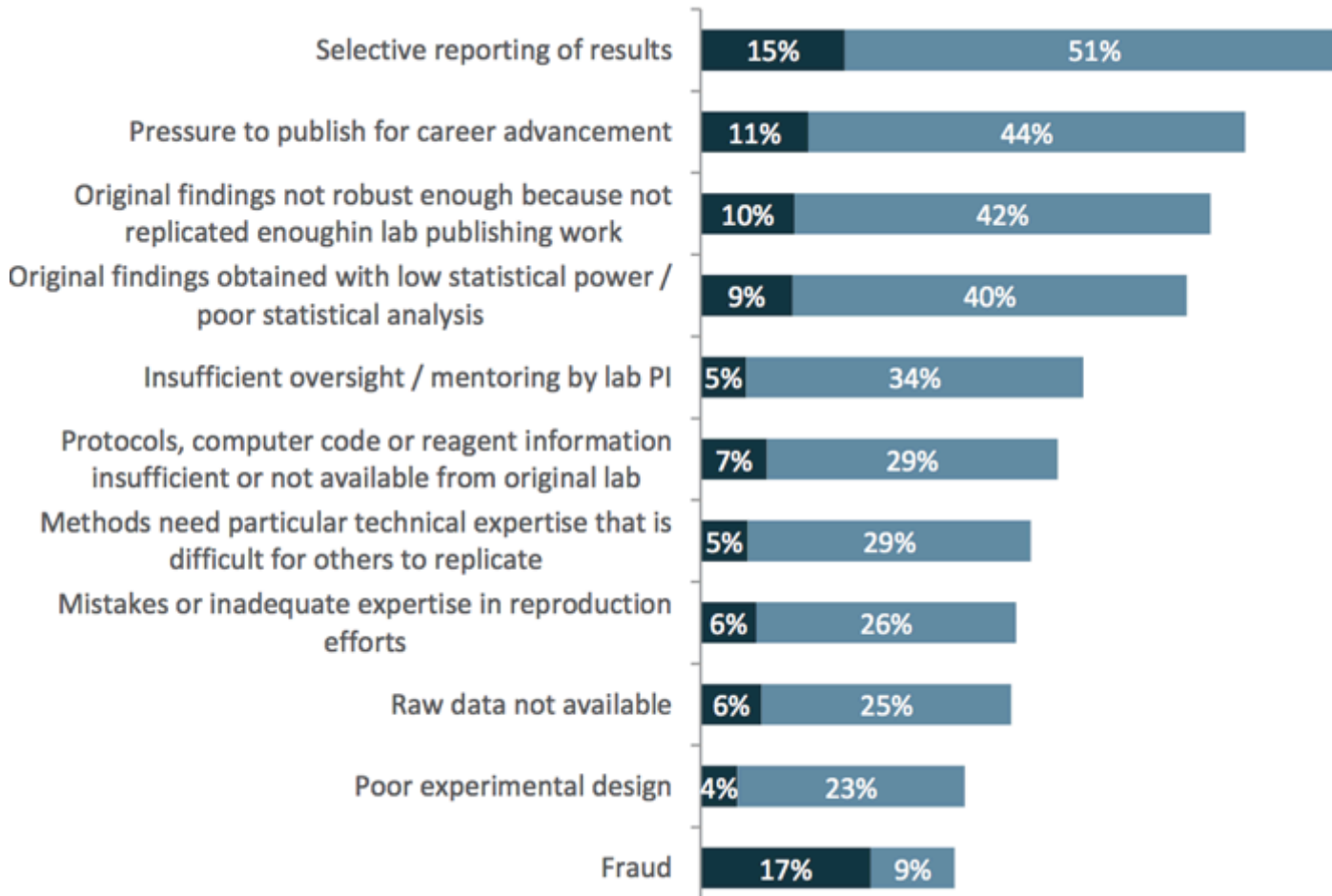
- **Los datos originales están disponibles en la nube.**
 - Los datos están almacenados en formato abierto (texto) .
 - **Todo el análisis y manejo de datos se hace mediante código.**
 - El código genera las tablas y figuras finales.
 - **Los datos brutos están separados de los datos derivados.**
 - Existe un 'script' maestro que ejecuta todos los pasos del análisis ordenadamente.
 - **Existe un documento README que explica los objetivos y organización del proyecto.**
 - Tanto el reporte, como los datos y código son públicos.
-

BENEFICIOS DE LA CIENCIA REPRODUCIBLE PARA EL ANALISTA DE DATOS

- **Permite la ejecución de tareas de análisis repetitivo sin esfuerzo.**
- **Muy fácil corregir y regenerar resultados, tablas y figuras.**
- **Reducción drástica del riesgo de errores.**
- **Facilitan la colaboración.**
- **Mayor facilidad para escribir reportes.**
- **Facilita el proceso de revisión.**
- **Ahorro de tiempo y esfuerzo al reutilizar código en otros proyectos.**

Adaptado de Sánchez et al. 2016 Ecosistemas 25(2): 83-92

FACTORES QUE CONTRIBUYEN A FALLA DE REPRODUCIBILIDAD



¿QUÉ ES LO PEOR QUE PUEDE PASAR?

RETRACTION

Retraction: Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins *Nat. Med.* 12, 1294–1300 (2006); published online 22 October 2006; corrected online 27 October 2006, 10 May 2007 and 10 October 2007 and corrected after print 21 July 2008; retracted 7 January 2011

We wish to retract this article because we have been unable to reproduce certain crucial experiments showing validation of signatures for predicting response to chemotherapies, including docetaxel and topotecan. Although we believe that the underlying approach to developing predictive signatures is valid, a corruption of several validation data sets precludes conclusions regarding these signatures. As these results are fundamental to the conclusions of the paper, we formally retract the paper. We deeply regret the impact of this action on the work of other investigators.

Nature Medicine would also like to note that several of the earlier correction dates were either omitted or incorrect. The corrigenda published online 10 May 2007, 10 October 2007 and 21 July 2008 mistakenly omitted the earlier correction date of 27 October 2006. The correction in July 2008 went online on 21 July 2008 but was incorrectly noted in the corrigendum as having gone online 18 July 2008.

<https://www.nature.com/articles/nm0111-135>

EDITORIALES QUE PIDEN COMPARTIR DATOS Y CÓDIGOS



SPRINGER NATURE

What is research data?

Raw or processed data files
Software
Code
Models
Algorithms
Protocols
Methods

A condition of publication in a Nature Research journal is that authors **are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.**

EDITORIALES QUE PIDEN COMPARTIR DATOS Y CÓDIGOS



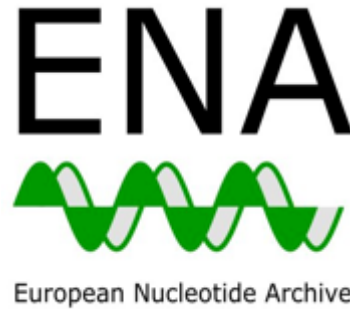
Computer Code and Software

For work where novel computer code was developed, authors should release the code either by depositing in a recognized, public repository such as GitHub or uploading as supplementary information to the publication. The name, version, corporation and location information for all software used should be clearly indicated. Please include all the parameters used to run software/programs analyses.

REPOSITARIOS DE DATOS Y PROYECTOS DE GENÓMICA



ncbi.nlm.nih.gov/genbank/



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



datadryad.org/



<https://github.com/>

REPOSITARIOS DE SOFTWARE PARA GENÓMICA

BIOCONDA[®]

<https://anaconda.org/bioconda/>



<https://biopython.org/>



<https://www.bioconductor.org/>

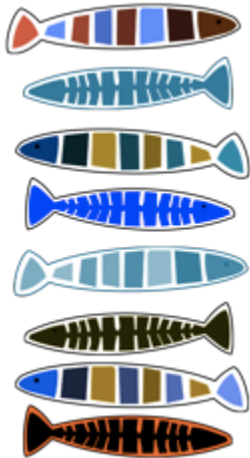
ruta de la investigación reproducible en genómica

Toma de
datos

Manipulación
de datos

Análisis
datos
integrado
con texto

Publicar
resultados



PYTHON



MARKDOWN



GitHub

Adaptado de Sánchez et al. 2016 Ecosistemas 25(2): 83-92

¿CÓMO ELABORAR DOCUMENTOS PARA LA INVESTIGACIÓN REPRODUCIBLE?

WYSIWYG

What You See Is What You Get



LibreOffice
The Document Foundation



WYSIWYM

What You See Is What You Mean



MARKDOWN



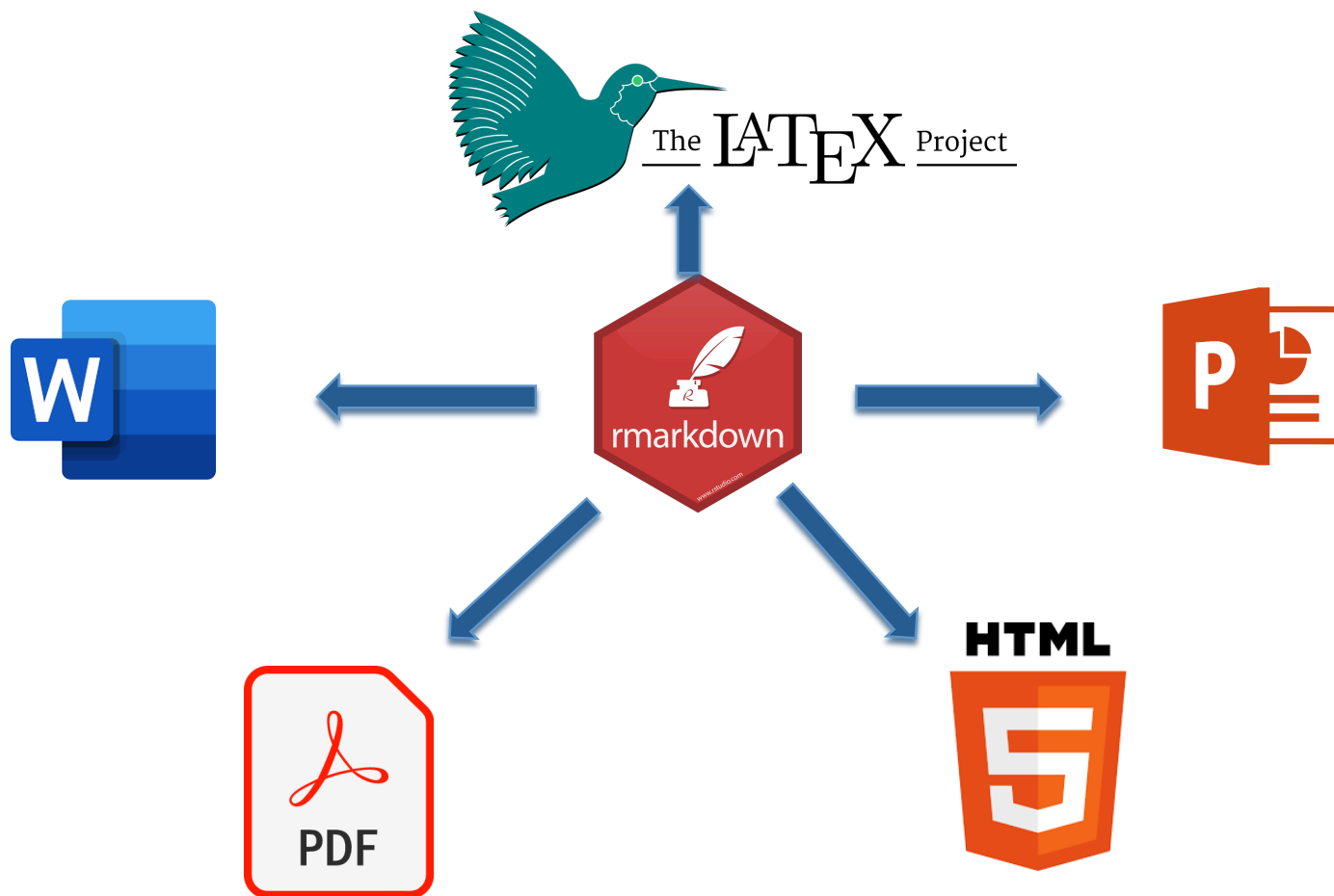
¿QUÉ ES R MARKDOWN?

Rmarkdown es un procesador de texto que permite la creación de reportes de alta calidad para tus clientes.



Learn more about R Markdown at <https://rmarkdown.rstudio.com/>

¿QUÉ TIPOS DE REPORTES PUEDO GENERAR CON RMARKDOWN?



Learn more about R Markdown at <https://rmarkdown.rstudio.com/>

VENTAJAS DE RMARKDOWN

1. Es un software libre y de código abierto, por lo que podemos usarlo sin necesidad de comprar una licencia.
2. Permite trabajar bajo el paradigma de la investigación reproducible (texto sin formato).
3. Cualquiera puede crear reportes, documentos, presentaciones y libros de alta calidad, con poco esfuerzo.
4. Uno de los mejores sistemas para crear reportes colaborativos y mejorar el rendimiento del trabajo de los analistas de datos.

¿QUÉ ES GIT HUB?



GitHub plataforma de trabajo colaborativo y open source que permite elaborar proyectos de programación utilizando el **sistema de control de versiones Git**.

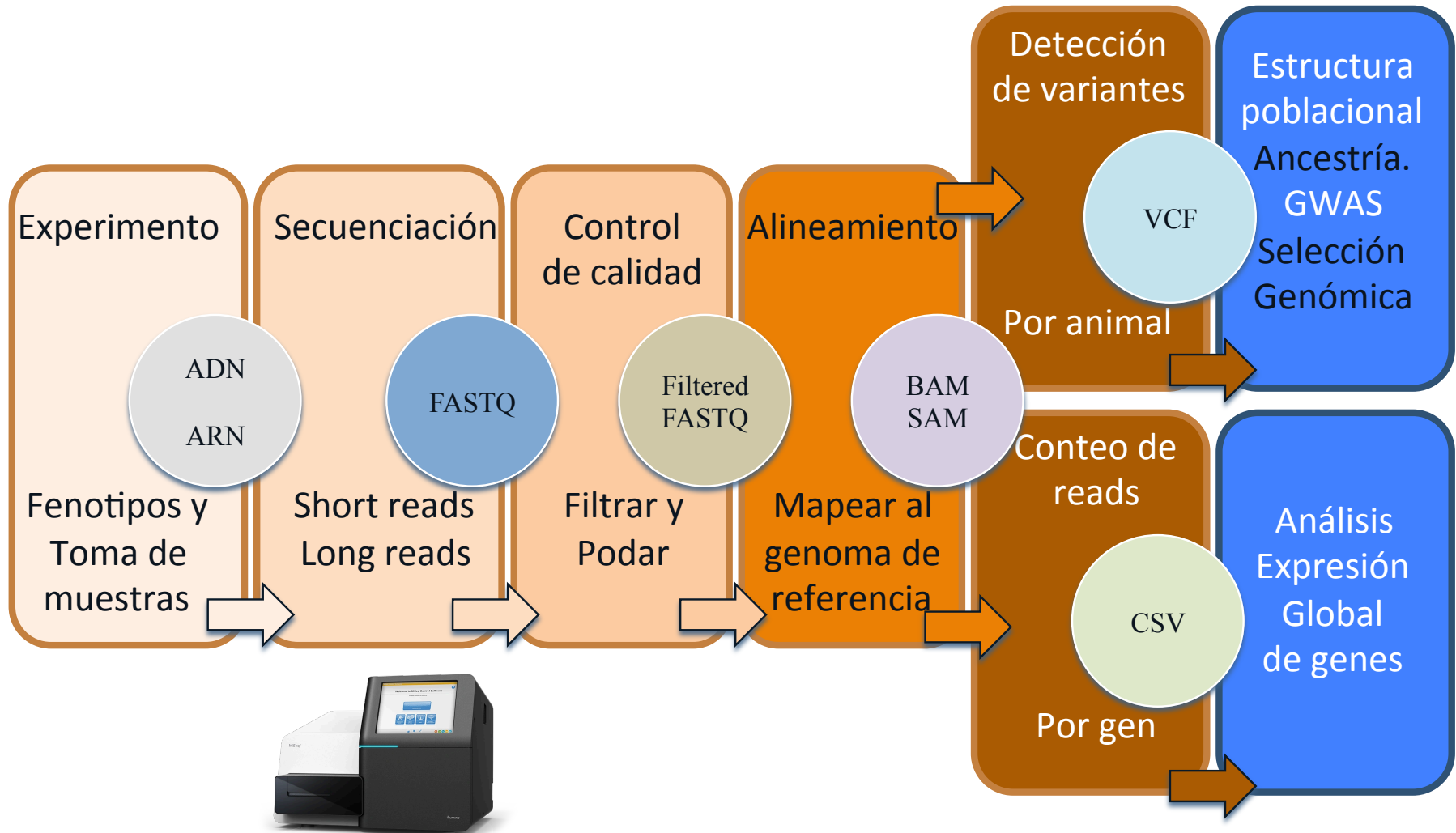
40 millones de usuarios

190 millones de repositories (28 millones publicos)

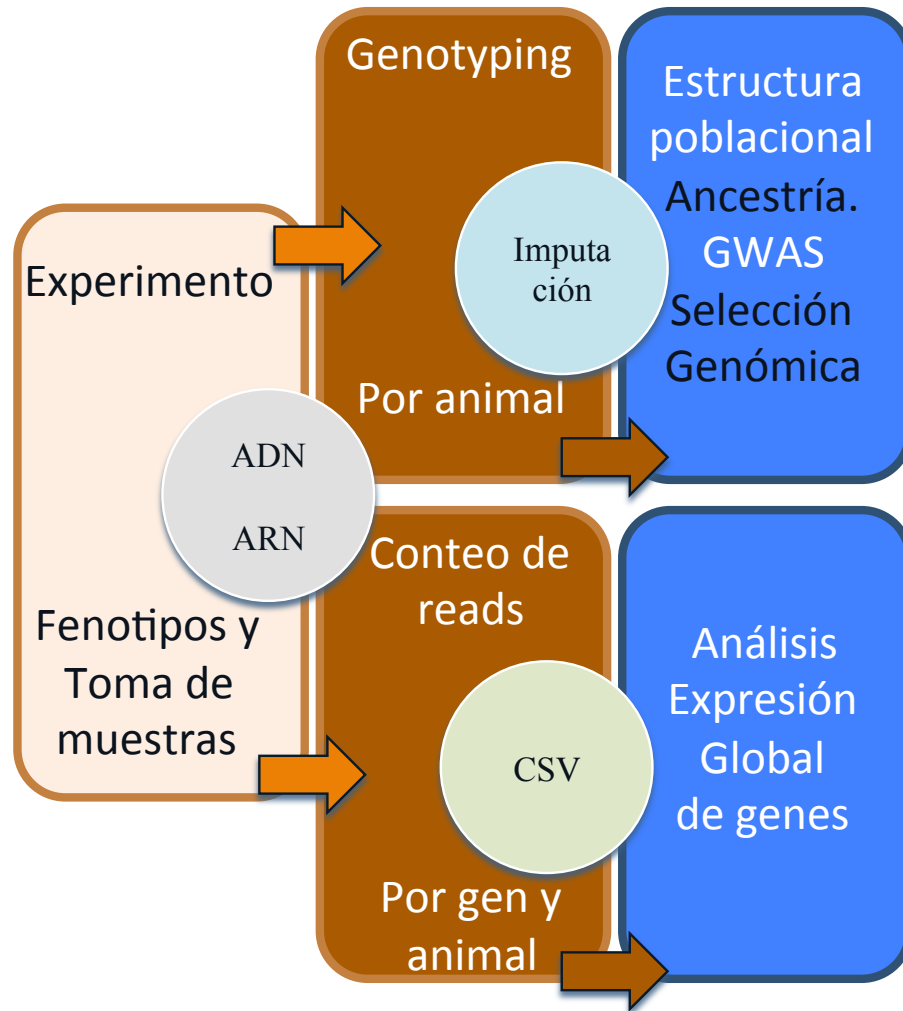
Acepta todos los lenguajes de programación.



FLUJO DE TRABAJO DE UN PROYECTO DE GENÓMICA APLICADA DESDE DATOS DE SECUENCIACIÓN



FLUJO DE TRABAJO DE UN PROYECTO DE GENÓMICA APLICADA DESDE DATOS DE MICROARREGLOS



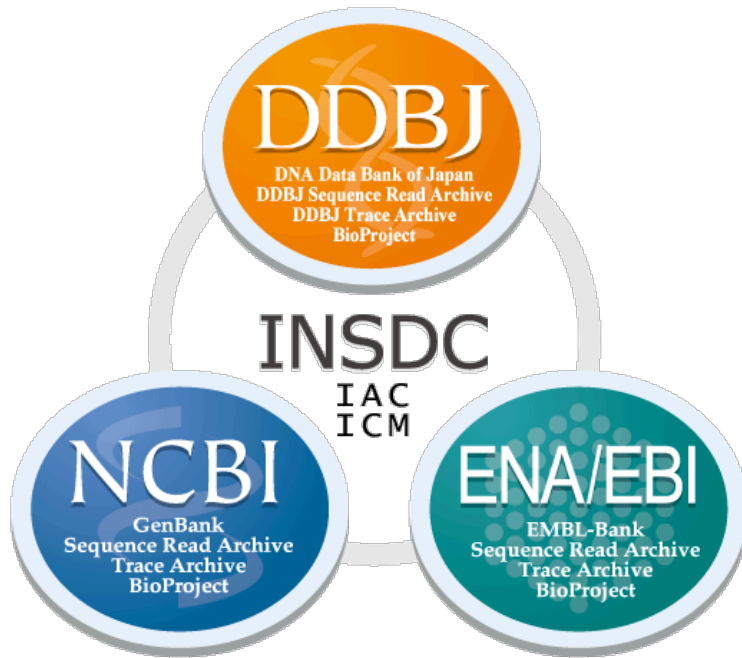
Práctica de investigación reproducible en genómica aplicada

Actividades

- 1.- Crear una cuenta en Github.
- 2.- Explorar repositorios de Github.
- 3.- Crear y compartir repositorio de proyecto genómico con el profesor y los compañeros.
- 4.- En el README elaborar perfil del proyecto incluyendo
 - a) Título: Práctica elaboración proyecto genómica aplicada.
 - b) Autor: Nombre, nacionalidad, profesión.
 - c) Descripción: A completar dependiendo de los datos seleccionados.

INTRODUCCIÓN BASES DE DATOS GENÓMICOS

INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE COLLABORATION



DDBJ: Dna Data Bank of Japan.

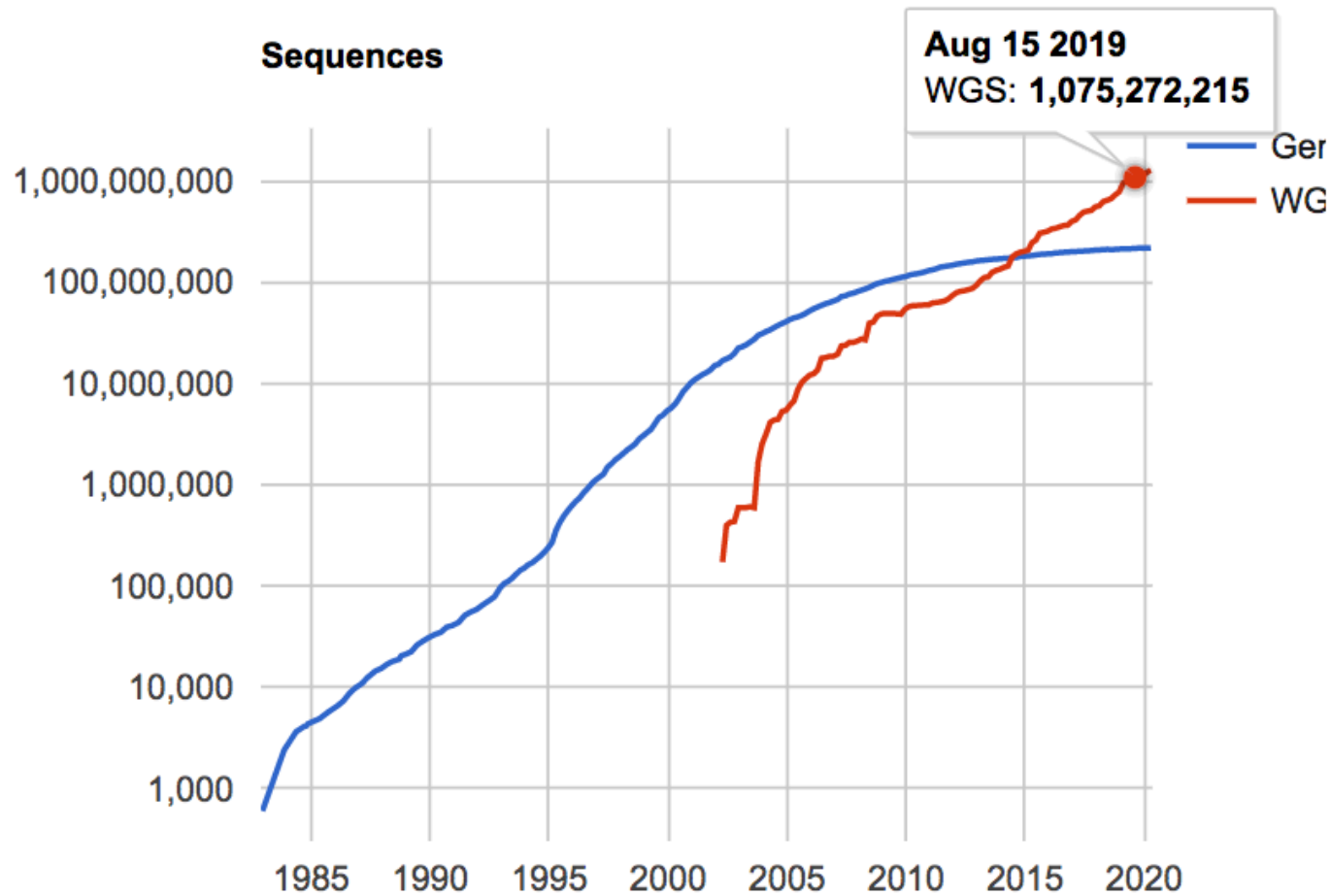
ENA: European Nucleotide Archive.

EVA: European Variation Archive.

NCBI: National center for biotechnology information.



NCBI: EVOLUCIÓN DE SECUENCIAS



RECURSOS NCBI: BASES DE DATOS

Assembly: BD con información sobre la estructura de genomas ensamblados, nombres de ensamblados y otros metadatos (anotación de genes), informes estadísticos y enlaces a datos de secuencias de genómas.

GenBank: BD redundante de secuencias genéticas del National Institutes of Health (NIH), es una colección anotada de todas las secuencias de ADN disponibles públicamente. GenBank es parte del INSDC.

Genome: Contiene secuencia y mapa de genomas completos y en progreso de más de 1000 organismos. Los tres dominios principales de la vida (bacterias, arqueas y eucariotas) están representados, pero también virus, fagos, viroides, plásmidos.

Reference sequence (RefSeq): Colección de secuencias de ADN y ARN (cDNA) no redundantes y curadas producidas por NCBI. RefSeqs proporciona una referencia estable para la anotación del genoma, identificación y caracterización de genes, análisis de mutaciones y polimorfismos, estudios de expresión y análisis comparativos.

RECURSOS NCBI: BASES DE DATOS cont.

SRA (Sequence Read Archive): Almacena datos NGS de plataformas de secuenciación que incluyen Roche 454 GS System®, Illumina Genome Analyzer®, Life Technologies AB SOLiD System®, Helicos Biosciences Heliscope®, Complete Genomics® y Pacific Biosciences SMRT® .

BioProject: Este recurso describe el alcance, el material y los objetivos de proyectos genómicos, proporcionando un mecanismo para recuperar conjuntos de datos almacenados en diferentes bases de datos.

BioSample: Esta base de datos contiene descripciones de materiales de origen biológico utilizados en ensayos experimentales.

ASSEMBLY: GENBANK V/S REFSEQ

History

GenBank Assembly Accession		RefSeq Assembly Accession	Assembly Name	Assembly Level	Status
GCA_000233375.4	≠	GCF_000233375.1	ICSASG v2	Chromosome	Latest GenBank, Latest RefSeq
GCA_000233375.3	n/a	n/a	ICSASG v1	Contig	Replaced GenBank
GCA_000233375.2	n/a	n/a	ICSASG v1	Contig	GenBank suppressed
GCA_000233375.1	n/a	n/a	ASM23337v1	Scaffold	Replaced GenBank

ASSEMBLY: GENBANK V/S REFSEQ

Assembly Definition

Assembly Statistics

Global assembly definition

[Download the full sequence report](#)

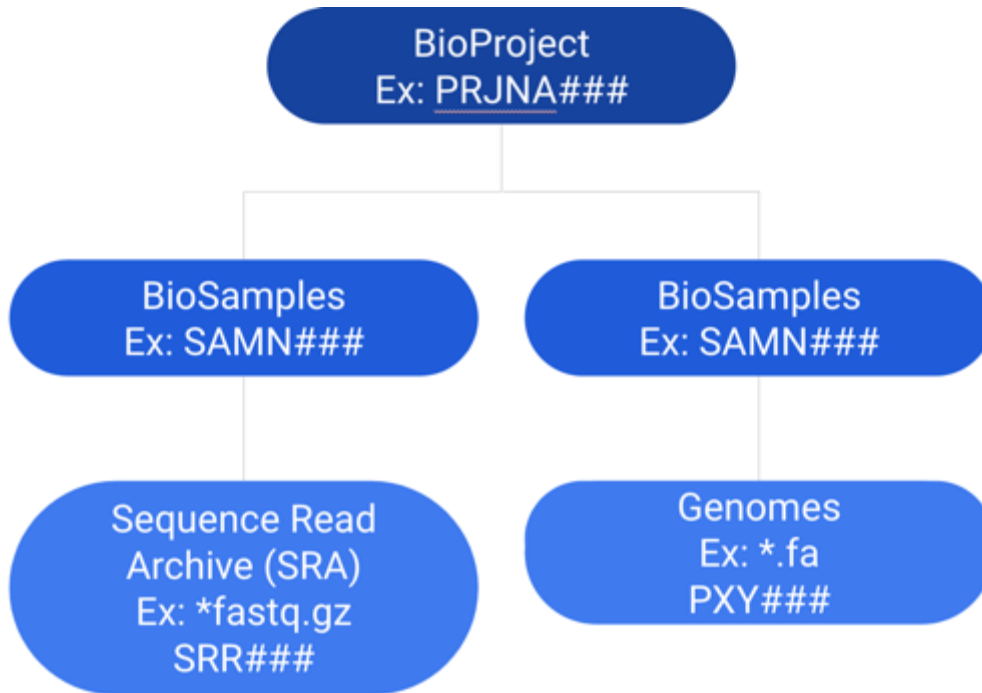
Click on the table row to see sequence details in the table to the right

Assembly Unit: Primary Assembly (GCF_000233385.1)

Assembly Unit Name
Primary Assembly
non-nuclear

Molecule name	GenBank sequence		RefSeq sequence	Unlocalized sequences count
Chromosome ssa01	CM003279.1	=	NC_027300.1	0
Chromosome ssa02	CM003280.1	=	NC_027301.1	0
Chromosome ssa03	CM003281.1	=	NC_027302.1	0
Chromosome ssa04	CM003282.1	=	NC_027303.1	0
Chromosome ssa05	CM003283.1	=	NC_027304.1	0
Chromosome ssa06	CM003284.1	=	NC_027305.1	0
Chromosome ssa07	CM003285.1	=	NC_027306.1	0
Chromosome ssa08	CM003286.1	=	NC_027307.1	0

SRA: ESTRUCTURA DEL ALMACENAMIENTO DE SECUENCIAS NGS.



- 1) Create BioProject.
- 2) Create Biosamples.
- 3) Upload Raw Reads (SRA).
- 4) Upload Genomes.

Práctica bases de datos NCBI:

Assembly y SRA

OBJETIVOS DEL TRABAJO PRÁCTICO

Esta práctica de bases de datos tiene como propósito:

- 1) Reconocer y utilizar las Bases de datos Assembly y SRA del NCBI
- 2) Realizar descarga manual de genomas, secuencias NGS y códigos de verificación.
- 3) Registrar actividades en el proyecto de genómica github.

DESCARGA **MANUAL** / **AUTOMÁTICA** DE GENOMAS

Full Report ▾

ICSASG_v2

Organism name: [Salmo salar \(Atlantic salmon\)](#)

Intraspecific name: Breed: double haploid

Isolate: Sally

Sex: female

BioSample: [SAMN02749551](#)

BioProject: [PRJNA72713](#)

Submitter: International Cooperation to Sequence the Atlantic Salmon Genome

Date: 2015/06/10

Assembly level: Chromosome

Genome representation: full

RefSeq category: representative genome

GenBank assembly accession: GCA_000233375.4 (latest)

RefSeq assembly accession: GCF_000233375.1 (latest)

RefSeq assembly and GenBank assembly identical: no ([hide details](#))

- Only in GenBank: 597817 unplaced scaffolds (in primary assembly-unit)
- Only in RefSeq: chromosome MT (in non-nuclear assembly-unit)
- Unplaced scaffolds shorter than 1,000 bases were omitted from the RefSeq assembly.
- Data displayed for RefSeq version

WGS Project: [AGKD04](#)

Assembly method: MaSuRCA v. 2.0.3

Send to: ▾

↓ **Download
Assembly**

See [Genome](#) Information
for **Salmo salar**

Access the data

[Genome Data Viewer](#)

[RefSeq Annotation Report](#)

[BLAST the assembly](#)

[Full sequence report](#)

[Statistics report](#)

[FTP directory for RefSeq assembly](#)

[FTP directory for GenBank assembly](#)

[NCBI Datasets](#) **NEW**

Assembly Information

[Assembly Help](#)

[Assembly Basics](#)

[NCBI Assembly Data Model](#)

DESCARGA **MANUAL** DE GENOMAS NO LO HAGA

Source database (GenBank or RefSeq) ?

RefSeq

File type ?

Genomic FASTA (.fna)

Estimated size is 701 MB

Download

Genomic Sequence

- ✓ Genomic FASTA (.fna)
- Genomic gaps (.txt)
- RepeatMasker output (.out)
- RepeatMasker run info (text)
- Annotation, Features & Products
 - Genomic GenBank format (.gbff)
 - Genomic GFF (.gff)
 - Genomic GTF (.gtf)
 - Feature count (.txt)
 - Feature table (.txt)
 - CDS from genomic FASTA (.fna)
 - RNA FASTA (.fna)
 - RNA GenBank format (.gbff)
 - RNA from genomic FASTA (.fna)
 - Pseudo without product FASTA (.fna)
 - Protein FASTA (.faa)
 - Protein GenPept format (.gpff)
 - Translated CDS (.faa)
 - RefSeq transcript alignments (.bam)
- NCBI Eukaryotic Genome Annotation
 - Evidence alignments (.gff)
 - Gnomon model GFF (.gff)
 - Gnomon model RNA FASTA (.fna)
 - Gnomon model protein FASTA (.faa)

Send to:

Download
Assembly

Access the data

- Genome Data Viewer
- RefSeq Annotation Report
- BLAST the assembly
- Full sequence report
- Statistics report
- FTP directory for RefSeq assembly
- FTP directory for GenBank assembly
- NCBI Datasets **NEW**

Assembly Information

- Assembly Help
- Assembly Basics
- NCBI Assembly Data Model

DESCARGA MANUAL DE GENOMAS Y CÓDIGO DE VERIFICACIÓN

← → ↻ 🏠 https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/233/375/GCF_000233375.1_ICSASG_v2/

📄 Aplicaciones 🌐 SIAL - Sistema de... 📄 Declaración de Ga... 📄 Affymetrix File Sh... 📄 SQLITE STUDIO... 📄 Fundamentos de... 📄 13. SQLite Databa... 📄 Wiki - CDCB nomi... 📄 tex

Index of /genomes/all/GCF/000/233/375/GCF_000233375.1_ICSASG_v2

Name	Last modified	Size
Parent Directory		-
GCF_000233375.1_ICSASG_v2_assembly_structure/	2019-12-04 17:34	-
Gnomon_models/	2019-12-04 17:35	-
GCF_000233375.1_ICSASG_v2_assembly_report.txt	2019-12-04 17:34	74M
GCF_000233375.1_ICSASG_v2_assembly_stats.txt	2019-12-04 17:34	29K
GCF_000233375.1_ICSASG_v2_cds_from_genomic.fna.gz	2019-12-04 17:34	35M
GCF_000233375.1_ICSASG_v2_feature_count.txt.gz	2019-12-04 17:34	480
GCF_000233375.1_ICSASG_v2_feature_table.txt.gz	2019-12-04 17:34	6.5M
GCF_000233375.1_ICSASG_v2_genomic.fna.gz	2019-12-04 17:34	669M
GCF_000233375.1_ICSASG_v2_genomic.gbff.gz	2019-12-04 17:35	1.1G
GCF_000233375.1_ICSASG_v2_genomic.gff.gz	2019-12-04 17:35	41M
GCF_000233375.1_ICSASG_v2_genomic.gtf.gz	2019-12-04 17:35	30M
GCF_000233375.1_ICSASG_v2_genomic_gaps.txt.gz	2019-12-04 17:35	1.3M
GCF_000233375.1_ICSASG_v2_protein.faa.gz	2019-12-04 17:35	19M
GCF_000233375.1_ICSASG_v2_protein.gpff.gz	2019-12-04 17:35	50M
GCF_000233375.1_ICSASG_v2_pseudo_without_product.fna.gz	2019-12-04 17:35	6.9M
GCF_000233375.1_ICSASG_v2_rm.out.gz	2019-12-04 17:35	42M
GCF_000233375.1_ICSASG_v2_rm.run	2019-12-04 17:35	864
GCF_000233375.1_ICSASG_v2_rna.fna.gz	2019-12-04 17:35	61M
GCF_000233375.1_ICSASG_v2_rna.gbff.gz	2019-12-04 17:35	173M
GCF_000233375.1_ICSASG_v2_rna_from_genomic.fna.gz	2019-12-04 17:35	64M
GCF_000233375.1_ICSASG_v2_translated_cds.faa.gz	2019-12-04 17:35	24M
README.txt	2019-11-01 14:35	43K
Salmo salar AR100_annotation_release_100	2019-12-04 17:34	747
Salmo salar AR100_annotation_report.xml	2019-12-04 17:34	57K
annotation_hashes.txt	2019-12-04 17:35	410
assembly_statistics.txt	2020-06-03 02:59	14
md5checksums.txt	2019-12-04 17:36	23K

DESCARGA MANUAL / AUTOMÁTICA DE SEQ NGS

Source
DNA (3,392)
RNA (2,438)

Type
genome (592)

Library Layout
paired (2,591)
single (3,421)

Platform
ABI SOLID (31)
Capillary (321)
Illumina (5,553)
LS454 (6)
PacBio SMRT (101)

Strategy
EpiGenomics (171)
Exome (425)
Genome (913)
other (4,503)

Data in Cloud
GS (6,005)
S3 (6,006)

File Type
bam (132)
fastq (4,592)

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

Search results

Items: 1 to 20 of 6012 Selected: 1 << First < Prev Page 1 of 301 Next > Last >>

- ☒ [Illumina NovaSeq 6000 paired end sequencing; Integration of Transcriptome, Gross Morphology and Histopathology in the Gill of Sea Farmed Atlantic Salmon \(*Salmo salar*\): Lessons from Multi-site Sampling](#)
1 ILLUMINA (Illumina NovaSeq 6000) run: 49.6M spots, 5G bases, 1.4Gb downloads
Accession: ERX3980817
- ☐ [Illumina NovaSeq 6000 paired end sequencing; Integration of Transcriptome, Gross Morphology and Histopathology in the Gill of Sea Farmed Atlantic Salmon \(*Salmo salar*\): Lessons from Multi-site Sampling](#)
1 ILLUMINA (Illumina NovaSeq 6000) run: 55.1M spots, 5.5G bases, 1.6Gb downloads
Accession: ERX3980816
- ☐ [Illumina NovaSeq 6000 paired end sequencing; Integration of Transcriptome, Gross Morphology and Histopathology in the Gill of Sea Farmed Atlantic](#)

Search in related databases

Database	Access		all
	public	controlled	
BioSample	4,685		4,685
BioProject	98		98
dbGaP			
GEO Datasets	36		36

Find related data

Database:

Find items

Search details

```
("Salmo salar"[Organism] OR  
salmo salar[All Fields])  
AND "Salmo salar"[orgn]
```


DESCARGA DE METADATOS NGS

Summary ▾ 20 per page ▾


View results as an expanded interactive table using the Run Selector

[Run selector](#)

Search results

Items: 1 to 20 of 592 Selected: 3

<< First < Prev

 Filters activated: genome. [Clear all](#) to show 6012 items.

☒ [Pooled WGS of Atlantic salmon resistant to IPNV](#)

1. 1 ILLUMINA (Illumina HiSeq 2000) run: 231.1M spots, 57.8G bases, 27.9Gb downloads
Accession: SRX7973466

☒ [Pooled WGS of Atlantic salmon susceptible to IPNV](#)

2. 1 ILLUMINA (Illumina HiSeq 2000) run: 208.6M spots, 52.1G bases, 24.3Gb downloads
Accession: SRX7973465

☒ [Pooled WGS of Atlantic salmon resistant to IPNV](#)

3. 1 ILLUMINA (Illumina HiSeq 2000) run: 202.6M spots, 50.6G bases, 22.8Gb downloads
Accession: SRX7973464

Send to: ▾ Filters: [Manage Filters](#)

Choose Destination

- ☐ File
- ☐ Collections
- ☒ Run Selector
- ☐ Clipboard
- ☐ BLAST

Send 3 experiments to Run Selector.

Go

GEO
Datasets

[36](#)

Find related data

Database:

Find items

Search details

("Salmo salar"[Or
salmo salar[All F

BIOPROJECT – BIOSAMPLES - RUNS

SRX7973466: Pooled WGS of Atlantic salmon resistant to IPNV

1 ILLUMINA (Illumina HiSeq 2000) run: 231.1M spots, 57.8G bases, 27.9Gb downloads

Design: 125bp PE

Submitted by: University of Edinburgh

Study: Pooled WGS of IPNV resistant and IPNV susceptible Atlantic salmon

[PRJNA614520](#) • [SRP253762](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: Pool of IPNV resistant Atlantic salmon - year class 2007

[SAMN14429409](#) • [SRS6358235](#) • [All experiments](#) • [All runs](#)

Organism: [Salmo salar](#)

Library:

Name: 150226_D00261_0227_AC6E0MANXX_2_IL-TP-019

Instrument: Illumina HiSeq 2000

Strategy: WGS

Source: GENOMIC

Selection: RANDOM

Layout: PAIRED

Runs: 1 run, 231.1M spots, 57.8G bases, [27.9Gb](#)

Run	# of Spots	# of Bases	Size	Published
SRR11394646	231,076,335	57.8G	27.9Gb	2020-03-25

METADATA DE UNA “RUN”

[Change](#)

Pooled WGS of Atlantic salmon resistant to IPNV (SRR11394646)

[Metadata](#) [Analysis](#) [Reads](#) [Data access](#)

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR11394646	231.1M	57.8Gbp	29.9G	42.6%	2020-03-25	public

Quality graph ([bigger](#))

This run has 2 reads per spot:

L=125, 100%

L=125, 100%

[Legend](#)

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX7973466	150226_D00261_0227_AC6E0MANXX_2_IL-TP-019	Illumina	WGS	GENOMIC	RANDOM	PAIRED	BLAST

Design:

125bp PE

Biosample	Sample Description	Organism	Links
SAMN14429409 (SRS6358235)		Salmo salar	PRJNA614520 [Pooled WGS of IPNV resistant and IPNV susceptible Atlantic salmon]

ANÁLISIS TAXONÓMICO

Pooled WGS of Atlantic salmon resistant to IPNV (SRR11394646)

Metadata

Analysis

Reads

Data access

Taxonomy Analysis

Unidentified reads: 0.85%

Identified reads: 99.15%

- cellular organisms: 99.15%
 - Eukaryota: 98.03%
 - Opisthokonta: 97.77%
 - Metazoa: 97.75%
 - Salmoninae: 96.29%
 - Salmo: 81.66%
 - Salmo salar: 57.35%
 - Choanoflagellata: < 0.01% (34 Kbp)
 - Fungi: < 0.01% (5 Kbp)
 - Viridiplantae: < 0.01% (133 Kbp)
 - Cryptophyceae: < 0.01% (7 Kbp)
 - Sar: < 0.01% (7 Kbp)
 - Bacteria: 0.35%
 - Viruses: < 0.01% (7 Kbp)

READs = Advanced options

Metadata Analysis **Reads** Data access

Filter: Find Filtered Download [What does it do?](#)

[What can the filter be applied to?](#)

The Run is too big (>1.1G) for searching by sequence substring.

< 1 1 23107634 >

View: ☒ biological reads ☐ technical reads ☒ quality scores [advanced options](#)

Reads (separated)

1. [SRR11394646.1](#) [SRS6358235](#)

name: HWI-D248:227:C6E0MANXX:2:1101:1449:1956,
member: default
x: 1449, y: 1956

2. [SRR11394646.2](#) [SRS6358235](#)

name: HWI-D248:227:C6E0MANXX:2:1101:1430:1966,
member: default
x: 1430, y: 1966

3. [SRR11394646.3](#) [SRS6358235](#)

name: HWI-D248:227:C6E0MANXX:2:1101:1383:1966,
member: default
x: 1383, y: 1966

4. [SRR11394646.4](#) [SRS6358235](#)

>gnIIISRAISRR11394646.1.1 HWI-D248:227:C6E0MANXX:2:1101:1449:1956 forward (Biological)
NTCAATATTGTCATACAGTGCAGCGCTAGACTGAACAAACACAAACAGTGGGNGGCAGTC
AAGGATGATCATCTTCTTCAGGCAGATCCTGACACTCTCACCTGTCTGTTTGA CTCTAAT
ATTAG

One channel quality score

N: 2 T:27 C:27 A:32 A:33 T:37 A:38 T:38 T:38 G:38 T:38 C:36 A:38 T:38 A:38 C:38
A:38 G:38 T:38 G:38 C:38 A:38 G:38 C:38 G:38 C:38 T:38 A:38 G:38 A:38 C:38 T:38
G:38 A:38 A:38 C:38 A:38 A:33 A:38 C:38 A:38 C:38 A:38 A:38 A:38 C:38 A:38 G:36
T:37 G:38 G:38 G:38 N: 2 G:28 G:28 C:37 A:38 G:38 T:38 C:38 A:38 A:38 G:38 G:38
A:31 T:29 G:37 A:38 T:38 C:38 A:38 T:38 C:38 T:38 T:38 C:38 T:38 T:38 C:38 A:38
G:38 G:29 C:38 A:31 G:37 A:38 T:38 C:38 C:38 T:37 G:38 A:38 C:38 A:38 C:38 T:38
C:38 T:38 C:38 A:38 C:38 C:38 T:29 G:15 T:37 C:38 T:35 G:35 T:38 T:38 T:38 G:38
A:38 C:38 T:36 C:38 T:37 A:38 A:38 T:38 A:38 T:36 T:38 A:38 G:38

DESCARGA MANUAL / AUTOMÁTICA DE SEQ NGS

Pooled WGS of Atlantic salmon resistant to IPNV (SRR11394646)

Metadata Analysis Reads Data access

SRA archive data

SRA archive data is normalized by the SRA load process and used by the [SRA Toolkit](#) to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name	Free Egress	Access Type
run	29,204,399 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/sra60/SRR/011127/SRR11394646	worldwide	anonymous
		AWS	s3://sra-pub-run-8/SRR11394646/SRR11394646.1	s3.us-east-1	aws identity
		GCP	gs://sra-pub-run-9/SRR11394646/SRR11394646.1	gs.US	gcp identity

Original format

The original files submitted to SRA. These files may require specific software to open, read and interpret data.

Type	Size	Location	Name	Free Egress	Access Type
fastq	19,139,395 Kb	GCP	https://storage.googleapis.com/sra-pub-src-8/SRR11394646/TP-019_1.fastq.gz.1	worldwide	anonymous
		AWS	https://sra-pub-src-8.s3.amazonaws.com/SRR11394646/TP-019_1.fastq.gz.1	worldwide	anonymous
fastq	19,045,580 Kb	GCP	https://storage.googleapis.com/sra-pub-src-8/SRR11394646/TP-019_2.fastq.gz.1	worldwide	anonymous
		AWS	https://sra-pub-src-8.s3.amazonaws.com/SRR11394646/TP-019_2.fastq.gz.1	worldwide	anonymous

TAREA 1

ACTUALIZAR Y CLONAR TU PROYECTO GITHUB

Debe ser usado como un cuaderno de laboratorio en el cual se registra y almacena la información clave del proyecto.

Registrar información clave del genoma de la especie de interés y del bioproject analizado.

TAREA 2

1.- Instalar software para acceso remoto SSH.

Tu PC es windows: Instala **PuTTY** <https://www.putty.org/>

Tienes una MAC o usas Linux: usaremos la terminal.

2.- Instalar software para transferencia de archivos vis FTP.

Tu PC es windows: Instala **WinSCP**

<https://winscp.net/eng/download.php>

Tienes una MAC: Instala **Cyberduck** <https://cyberduck.io/>

3.- Instalar editor de textos **nano**. <https://www.nano-editor.org/>

RESUMEN DE LA CLASE

- Revisamos la importancia de la investigación reproducible en genómica.
- Iniciamos un proyecto de genómica aplicada en Github.
- Nos familiarizamos con algunas bases de datos del NCBI