

Laboratorio 03 - Ensamblaje de genomas y predicción de genes

En este laboratorio vamos a familiarizarnos con publicaciones de genomas, y como fueron ensamblados. También vamos a explorar dos métodos de predicción de genes

Parte 1: El artículo genoma

Comencemos por ir a la Genomes on Line Database (GOLD) y escoger un genoma de interés

- Ve a la base de datos [GOLD](#) y busca un genoma eucarionte de interés.

https://gold.jgi.doe.gov

JGI GOLD GENOMES ONLINE DATABASE

JGI HOME LOG IN

Home Search Distribution Graphs Biogeographical Metadata Statistics References Team Help News


Studies	23,562
Biosamples	6,551
Sequencing Projects	82,517
Analysis Projects	63,055
Organisms	74,176

[Download Excel Data file](#)
File last generated: 30 Mar, 2016

Welcome to the Genomes OnLine Database

GOLD Release v.5
GOLD: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.


1. Register



Register your project information and Metadata in the Genomes Online Database

[Register](#)


2. Annotate



Annotate your microbial genome or metagenome with IMG/ER or IMG/MER

[Annotate](#)

3. Publish



Standards in Genomic Sciences

Publish your genome or metagenome in open access standards-supportive journal.

[Publish](#)

Studies Metagenomic 710 Non-Metagenomic 22,852	Biosamples Classification Ecosystems Host-associated 1,934 Engineered 467 Environmental 4,159	Sequencing Projects Complete Projects 8,029 Permanent Drafts 33,499 Incomplete Projects 38,472 Targeted Projects 1,546	Analysis Projects Genome Analysis 46,135 Metagenome Analysis 5,618 Combined Assembly 106 Genome from Metagenome 1,513 Metatranscriptome Analysis 1,923 Single Cell (Screened) 1,633 Single Cell (Unscreened) 792 Transcriptome Analysis 0
Organisms Organisms 74,176 Archaea 1,231 Bacteria 55,908 Eukarya 12,412 Viruses 4,596	Special Projects Type Strain Projects 5,312 GEBA Projects 2,495 HMP Projects 2,916	JGI Projects JGI Studies 1,112 JGI Biosamples 4,012 JGI Sequencing Projects 31,535 JGI Analysis Projects 16,226	Projects with Genbank Data Seq. Projects 42,612 Archaeal Projects 565 Bacterial Projects 35,768

- Navega dentro del portal siguiendo vínculos a, por ejemplo, "Complete projects", "Eukarya" o la casilla de búsqueda.

Responde:

¿Cuántos genomas han sido depositados en GOLD? ¿Son los mismos de GENBANK?
¿Cuál es la distribución de procariontes y eucariontes secuenciados?

En este ejemplo continuaremos con el genoma del Bonobo *Pan paniscus*

The screenshot shows the JGI GOLD Genomes Online Database interface. The browser address bar displays 'gold.jgi.doe.gov'. The header includes the JGI logo, the GOLD logo, and the text 'GENOMES ONLINE DATABASE'. Navigation links include 'JGI HOME' and 'LOG IN'. A secondary navigation bar contains links for 'Home', 'Search', 'Distribution Graphs', 'Biogeographical Metadata', 'Statistics', 'References', 'Team', 'Help', and 'News'.

On the left side, a table provides summary statistics:

Studies ⓘ	23,562
Biosamples ⓘ	6,551
Sequencing Projects ⓘ	82,517
Analysis Projects ⓘ	63,055
Organisms	74,176

The main content area features tabs for 'Organism Information', 'Organism Metadata', 'Isolation Metadata', and 'Environmental Metadata'. The 'Organism Information' tab is active, displaying a table of taxonomic data for 'Pan paniscus':

Organism Name ⓘ	Pan paniscus
Other Names	
Genus ⓘ	Pan
Genus Synonyms	
Species ⓘ	Pan paniscus
Species Synonyms	
Subspecies	
Strain	
Culture Collection	
Domain	EUKARYAL
Phylum	CHORDATA
NCBI Taxonomy ID ⓘ	9597
NCBI Kingdom	Metazoa
NCBI Phylum	Chordata
NCBI Class	Mammalia
NCBI Order	Primates
NCBI Family	Hominidae
NCBI Genus	Pan

- A partir de la ficha GOLD de tu organismo, busca la publicación de su genoma (pista: sigue el Taxonomy ID)
- Busca el vínculo a la base de datos de genomas de NCBI y ubica la referencia al artículo original

ncbi.nlm.nih.gov

Genome [Create alert](#) [Limits](#) [Advanced](#) [Help](#)


Pan paniscus (pygmy chimpanzee)
Representative genome: [Pan paniscus \(assembly panpan1.1\)](#)
Download sequences in FASTA format for [genome](#), [transcript](#), [protein](#)
Download genome annotation in [GFF](#), [GenBank](#) or [tabular](#) format
BLAST against Pan paniscus [genome](#), [transcript](#), [protein](#)

Tools

BLAST Genome

Display Settings: [Overview](#) Send to: [ID: 10729](#)

Organism Overview ; [Organelle Annotation Report \[2\]](#)



Pan paniscus (pygmy chimpanzee)
The bonobo or pygmy chimpanzee

Lineage: [Eukaryota\[1786\]](#); [Metazoa\[646\]](#); [Chordata\[276\]](#); [Craniata\[271\]](#); [Vertebrata\[271\]](#); [Euteleostomi\[267\]](#); [Mammalia\[117\]](#); [Eutheria\[113\]](#); [Euarchontoglires\[51\]](#); [Primates\[26\]](#); [Haplorrhini\[20\]](#); [Catarrhini\[16\]](#); [Hominidae\[5\]](#); [Pan\[2\]](#); [Pan paniscus\[1\]](#)

Although the bonobo, or pygmy chimpanzee, *Pan paniscus* and common chimpanzee *Pan troglodytes* are morphologically similar they are known to have many behavioral differences. Bonobos are endangered in their wild habitat in west-central Africa.

Summary

Submitter: Max-Planck Institute for Evolutionary Anthropology
Assembly level: Chromosome
Assembly: [GCA_000258655.2 panpan1.1 scaffolds: 10,984 contigs: 121,356 N50: 66,676 L50: 11,048](#)
BioProjects: [PRJNA169343](#), [PRJNA49285](#)
Whole Genome Shotgun (WGS): [INSDC: AJFE000000000.2](#)
Statistics: total length (Mb): 3286.64
protein count: 15
GC%: 42.107
NCBI Annotation Release: 102

Publications

1. [The bonobo genome compared with the chimpanzee and human genomes.](#) Prüfer K, et al. Nature 2012 Jun 28
2. [History of the Tfam gene in primates.](#) D'Errico I, et al. Gene 2005 Dec 5
3. [Short KIR haplotypes in pygmy chimpanzee \(Bonobo\) resemble the conserved framework of diverse human KIR haplotypes.](#) Rajal et al. J Exp Med 2001 Jan 1

[More...](#)

Search details

[See more...](#)

Recent activity

[Turn Off](#) [Clear](#)

Pan paniscus Genome

[txid9597\[Organism:noexp\]](#) (1)

- Ahora solo tienes que conseguir el artículo y buscar en los materiales y métodos par responder la siguientes preguntas. Existen una serie de métricas o estadísticas que se usan para evaluar un ensamblaje de genomas.
- Sigue [este vínculo](#) y revisa las siguientes N50, L50, y NG50.

Table 1: Bonobo genome assembly characteristics and genomic features compared with the chimpanzee genome (panTro2)

From
The bonobo genome compared with the chimpanzee and human genomes
 Kay Prüfer, Kasper Munch, Ines Hellmann, Keiko Akagi, Jason R. Miller, Brian Walenz, Sergey Koren, Granger Sutton, Chinnappa Kodira, Roger Winer, James R. Knight, James C. Mullikin, Stephen J. Meader, Chris P. Ponting, Gerton Lunter, Saneyuki Higashino, Asger Hobolth, Julien Dutheil, Emre Karakoç, Can Alkan, Saba Sajjadian, Claudia Rita Catacchio, Mario Ventura, Tomas Marques-Bonet, Evan E. Eichler *et al.*
Nature **486**, 527–531 (28 June 2012) | doi:10.1038/nature11128

[back to article](#)

Table 1: Bonobo genome assembly characteristics and genomic features compared with the chimpanzee genome (panTro2)

	Bonobo	Chimpanzee
Bases in contigs	2.7 Gb	3.0 Gb
N50 contigs	67 kb	29 kb
N50 scaffolds	9.6 Mb	9.7 Mb
Human bases covered by alignments	2.74 Gb	2.72 Gb
Lineage-specific substitutions	5.71 million	5.67 million
Indel error rate	0.14 errors kb ⁻¹	0.13 errors kb ⁻¹
Segmental duplication content (>20 kb)	77.2 Mb	76.5 Mb
Lineage-specific retrotransposon integrants	1,445	1,039

See also [Supplementary Information](#), sections 2–4 and 6. kb, kilobase; Mb, megabase; Gb, gigabase.

Responde:

¿Qué es el N50, L50, NG50?
 ¿Cuál es el propósito de calcular estas estadísticas?
 ¿Cuál es el genoma que escogiste? Adjunta la referencia.
 ¿Cuál es el N50 del genoma que escogiste? ¿Y el NG50?
 ¿Qué tipo de tecnología se usó para secuenciar el genoma que escogiste?
 ¿Cuántos cromosomas tiene tu organismo y cuál es su tamaño?

Parte 2: Predicción de genes

En esta parte vamos a utilizar dos programas clásicos para predecir genes [ORFfinder](#) y [GLIMMER](#).

- Tu misión para esta parte es descargar una secuencia y predecir qué genes están presentes usando ambas herramientas.
- Descarga la secuencia problema [aquí](#) o copia y pega desde:

```
ATGAGTATCAAAATCTTATCTGAATCAGAAATTAAACAAGTAGCAAATTCATATCAAGCCCCAGCGGTTTTATTTGCCAAT
CCTAAAAATCTTTACCAACGCAGAGCGAAACGTTTAAAGAGACTTAGCACAAAATCATCCTCTATCTGATTATTTATTATTT
GCTGCAGACATAGTTGAAAGCCAACTTTCCACGTTAGAAAAAATCCTTTACCGCCACAACAGCTTGAACAGTTAAATACT
ATCGAGCCACTAAATGCCAAAACCTTTAAGAGAAACAGTATCTGGCGTGAATACTTAACAGAAATTCTTGATGAAATAAAG
CCCAAAGCTAACGAGCAAATTGCTGCAACAATTGAATTTCTTGAAAAAGCCTCTTCCGCTGAATTAGAAGAAATGGCAAAT
AAACTCTTAGCACAGAATTTAACTTAGTCAGCAGTGATAAAGCCGTCTTTATTTGGGCTGCACCTTCCCTTTATTGGTTA
CAAGCAGCTCAACAAATTCCTCATAATAGCCAAGTTGAAAACGCTGAAAAATTTACATCACTGCCCTGTTTGTGGTTCTTTA
CCTGTGGCAAGTATGGTACAAATTGGTACATCACAAGGTTTACGCTACTTACATTGTAATTTATGTGAAAGTGAATGGAAT
TTGGTACGCGCACAATGCACCAATTGTAATAGTCATGACAACTCGAAATGTGGTCACTAAATGAAGAACTTGCCTTGT
CGTGCCGAAACCTGTGGTAGTTGTGAAAGTTACTTGAAAATGATGTTCCAAGAAAAAGATCCTTACGTAGAACCCTGTAGCC
GATGATTTGGCTTCTATTTTCTTAGATATTGAAATGGAAGAAAAAGGTTTCGCCCCAAGTGGATTAAATCCATTTATTTTT
CCTGCAGAAGAAGCATAAAAAATATAGCCTAGAAATATCTAGGCTAATTATTTAAATCTATAGATAACACGCCATTACCCT
GCATTTTCTCGTCCACCACTGGGTTTCGGATAATAATTTTTCCGAATATCCACTTCATTAACCAATTTTTTCATACAAG
AAACGAGCAGAGTTAGACTCTCTGACTTCTAGCCATAAGGTTTGGACACCTTTTTTCCTTTAATTGAAAGATTAATTTTCCT
AACATAATTTGCCAAATCCACAACCTTGATAAGTAGGCAAAATCGCAATATTAACAAAGTCGCTTCATCCAATACTGTT
TGGCAAATAGCAAAACCGATAATCTGATTATTCTCTATTAATTTTAAATTGAGATAACGCTCCCTTGATTATTTTAAAC
GTACCAAATGACCAAGGCACCAGATGGGCTTGCTGTTTCGATTTTCATACAATCGCTCGAAATCACAGGCTTCAATTGAGAA
ATAATAGACATAATTAAGGCTGCTGAATTTGTTGCCACAACGCTCGTTTGGCTTGATGATTAGATTGAAATTGCTGCCAAC
TTGGCGAGCGATAAACCTGCTCAGCCTGCTTGCAAAATGGCAAAGTGGGTCATTTGGTTCGCTATTTTCTGATAGTAACC
AATAACGAATAGGCTGTTTACATTCCATATGCTGGATTTGATCGTAATTCAAACATAACAATTTCTTTTTTAAGATTAA
GGCTTAACAGCACATCAGCCAACAAAGGCGAGCTACTGATATTTTCATCGGAAACAGTGATAAGGCGAATATTCTCTGCCA
CACTAATTCCTACTGAACCTTGCACTGCTCGGGGCGATATAATCCCACTGGGAAATGCCCATTTCTTGTAAGAAGAT
CGGCTCTGTTTCATAAGTGAATTTTAAAAACGTTGCTTGAAAAATAGTGTTAATTTCTTTTATTAATCTTTGCTAAAATGG
TGCCTAATTTTAAACCAATAACCCACAGGATTCAAAGCG
```

Responde:

- ¿Cuántos ORF o genes encontró ORFfinder?
- ¿Cuántos ORF o genes encontró Glimmer?
- ¿Alguno de los genes predichos por estas herramientas coinciden?
- ¿En qué hebra están codificados?
- ¿Qué tipo de programa es GLIMMER? ¿Ab initio o por homología?

Trabajo de laboratorio para la próxima semana

El trabajo de laboratorio para la próxima semana consta de dos partes. La primera parte ya la tienes lista. Simplemente tienes que responder las preguntas que aparecen a través de esta guía y enviar un informe a bioinformatica.unab2016@gmail.com. Envíen a este correo los informes hasta el jueves de la semana siguiente a la realización del práctico. La hora límite es las 23:59. En el Asunto del correo pongan Informe de Laboratorio 0x así nosotros podemos clasificar automáticamente los informes. Para la próxima semana el Asunto del correo debería ser Informe de Laboratorio 03.

Profesor	Nombre	Correo electrónico
Profesor responsable	Dr. Eduardo Castro Nallar	eduardo.castro@unab.cl
Profesor ayudante sección 1	Ingrid Araya Durán	ingrid.araya.duran@gmail.com
Profesor ayudante sección 1	Sandro Valenzuela	sandrolvalenzuelad@gmail.com
Profesor ayudante sección 2	Javier Cáceres	ja.caceresmolina@gmail.com
Profesor ayudante sección 2	Consuelo Bello	consuelobelloz@gmail.com

La segunda parte tiene que ver con leer un artículo científico sobre este tema: Mihai Pop. 2013. *Sequence assembly demystified*. **Nature Reviews Genetics** 14, 157-167 | doi:10.1038/nrg3367.

Puedes acceder al artículo [aquí](#). Recuerda que el contenido de este artículo es el material para el próximo control de entrada.