

# Libro de respuestas

## Taller herramientas de Ensembl y Ensembl Genomes

### Actividad 1.

#### 1. Ir al portal de Ensembl Genomes que corresponda

Como estamos trabajando con *V. vinifera*, debemos ir a *Ensembl plants* (<https://plants.ensembl.org/>)

The screenshot shows the Ensembl Genomes website. At the top, there's a dark blue header with the Ensembl Genomes logo and the tagline "Providing genome data for non-vertebrate species, with tools for the manipulation, analysis and visualisation of that data". Below the header, there's a search bar labeled "Search all genomes" and a "Go" button. On the left, there are two blue boxes: one for "Ensembl COVID-19" with a "SARS-CoV-2" icon and a "Go" button, and another for "Ensembl Rapid Release" with a "2-weekly releases of new assemblies" description and a "Go" button. On the right, there's a white box for "EnsemblPlants" with a "Go to Ensembl Plants" button. Below this, there are three plant species listed: *Triticum aestivum* (IWGSC), *Oryza sativa Japonica Group* (IRGSP-1.0), and *Arabidopsis thaliana* (TAIR10).

#### 2. Registrar la versión actual de Ensembl Genomes

Es la versión 59

##### What's New in Release 59

- New genomes
  - [Aegilops umbellulata](#) : assembly from NCBI [GCA\\_032464435.1](#) and annotation from [KAUST](#).
  - [Vicia faba](#) : assembly from NCBI [GCA\\_948472305.1](#) and annotation from [FBGC](#).
- Updated genomes
  - [Manihot esculenta](#) : assembly from NCBI [GCA\\_001659605.2](#) and annotation from [JGI](#).
  - [Medicago truncatula](#) : assembly from NCBI [GCA\\_003473485.2](#) and annotation from [INRA/CNRS](#).

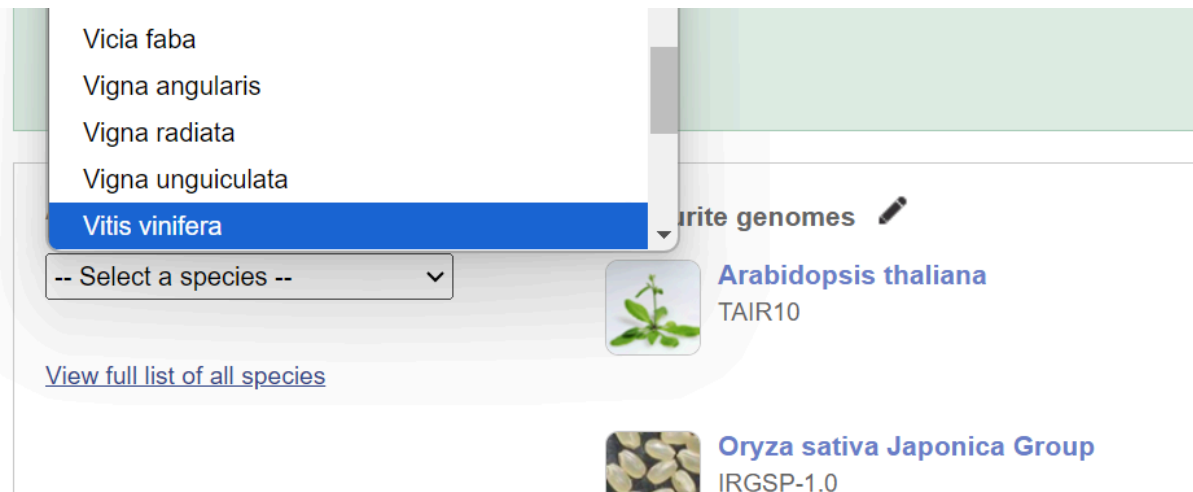
##### Did you know...?

You can search the [Track Hub Registry](#) to find public RNA-Seq studies aligned to plant genomes ([read more](#)).

Track  
Hub  
Registry

Check which features are supported on each species [here](#); if you would like to request a new species please email [Helpdesk](#).

#### 3. Encontrar el sitio para *V. vinifera*



El sitio es [https://plants.ensembl.org/Vitis\\_vinifera/Info/Index](https://plants.ensembl.org/Vitis_vinifera/Info/Index)

#### 4. Registrar el nombre del *assembly* (genoma de referencia), su tamaño en pb, número de cromosomas y cantidad de genes anotados

Assembly PN40024.v4, INSDC Assembly GCA\_910591555.1, May 2021.  
Tamaño: 475,604,073. Coding genes: 35,134

**Vitis vinifera (PN40024.v4)**

**Vitis vinifera Assembly and Gene Annotation**

**About *Vitis vinifera***

*Vitis vinifera* (grape) is the most widely cultivated and economically important grape species used for both eating and wine. It has a diploid genome with haploid chromosome number of 19, and an estimated genome size of ~500 Mb.

**Assembly**

This release is based on a 12x whole genome shotgun sequence assembly and the V1 annotation of the *Vitis vinifera* genome. These data were prepared by a French-Italian [Public Consortium for Grapevine Genome Characterisation](#) under the auspices of the [International Grape Genome Program \(IGGP\)](#). Further details of the sequencing and assembly are available from [Genoscope](#).

**Annotation**

Protein-coding genes were predicted by combining *ab initio* models, *V. vinifera* complementary DNA alignments, and alignments of proteins and genomic DNA from other species. The integration of the data was performed with [GAZE](#).

**Statistics**

**Summary**

Assembly	PN40024.v4, INSDC Assembly <a href="#">GCA_910591555.1</a> , May 2021
Database version	112.4
Golden Path Length	475,604,073
Genebuild by	IGGP
Genebuild method	Curated
Data source	<a href="#">INRAE Grand Est-Colmar</a>


**Gene counts**

Coding genes	35,134
Gene transcripts	41,097

## Actividad 2:

### 1. En Biomart, seleccionar el dataset correspondiente

Debemos ir a BioMart (<https://plants.ensembl.org/biomart/martview/>) y seleccionar como DATABASE: Ensembl Plants Genes 59) y como DATASET: Vitis vinifera genes



New
Count
Results

<b>Dataset</b>	Ensembl Plants Genes 59
Vitis vinifera genes (PN40024 v4)	Vitis vinifera genes (PN40024.v4)

## 2. Ingresar como filtro nuestra lista de genes upregulados a 72h post infección (cuyos IDs están codificados en formato NCBI)

Primero debo establecer los filtros. Puntualmente en esta actividad, acotar la búsqueda a mi lista de genes. Para esto, FILTERS (1)→ GENE (2) → click en Input external references ID list (3) → Desplegable, elegir NCBI gene IDs (4) → En choose file, subir mi lista de genes (unique\_72h\_UP.txt, 5)

<b>Dataset</b>	REGION:		
Vitis vinifera genes (PN40024.v4)	GENE: 2		
<b>Filters</b>	<input type="checkbox"/> Limit to genes (external references)...         With European Nucleotide Archive ID(s) <input checked="" type="radio"/> Only <input type="radio"/> Excluded		
NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] 1	<input checked="" type="checkbox"/> Input external references ID list [Max 500 advised] 3         NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840] 4		
<b>Attributes</b>	Choose File unique_72h_UP.txt 5		
Gene stable ID			
Transcript stable ID			

## 3. Obtener para cada gen (cuyo ID está codificado en formato NCBI) el ID usado por la comunidad científica para esa especie (puntualmente para *Vitis*, se llaman “Vitvi\_XXX”)

Luego obtener la correspondencia entre IDs de NCBI y IDs propios de Vitis. Para esto, ir a Attributes (1), seleccionar Features (2), GENE (3), Gene stable ID (4, así se llama el ID usado por la comunidad), EXTERNAL (5), NCBI gene ID (6)

**Dataset**  
Vitis vinifera genes (PN40024.v4)

**Filters**  
NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified]

**Attributes** 1  
Gene stable ID  
NCBI gene (formerly Entrezgene) ID

Please select columns to be included in the output and hit 'Results' when ready

☒ **Features** 2  
☐ Structures  
☐ Homologues (Max select 6 orthologues)

☐ Variant (Germline)  
☐ Sequences

**GENE:** 3

**Ensembl**

☒ Gene stable ID 4  
☐ Transcript stable ID  
☐ Protein stable ID  
☐ Exon stable ID  
☐ Gene description

☐ Transcript end (bp)  
☐ Transcription start site (TSS)  
☐ Transcript length (including UTRs and CDS)  
☐ Ensembl Canonical  
☐ Transcript count

**EXTERNAL:** 5

**GO**

☐ GO term accession  
☐ GO term name  
☐ GO term definition

☐ GO term evidence code  
☐ GO domain

**GOSlim GOA**

☐ GOSlim GOA Accession(s)  
☐ GOSlim GOA Description

**External References (max 3)**

☐ European Nucleotide Archive ID  
☐ INSDC protein ID  
☐ MEROPS - the Peptidase Database ID  
☐ NCBI gene (formerly Entrezgene) description  
☐ NCBI gene (formerly Entrezgene) accession  
☒ NCBI gene (formerly Entrezgene) ID 6  
☐ PDB ID

☐ RefSeq DNA ID  
☐ RefSeq peptide ID  
☐ STRING ID  
☐ UniParc ID  
☐ UniProtKB/SpliceVariant ID  
☐ UniProtKB/Swiss-Prot ID  
☐ UniProtKB/TrEMBL ID

Luego podemos exportar esos resultados en formato .tsv. Para esto, vamos a Results (1), luego seleccionamos en el desplegable el formato (2), finalmente descargamos con Go (3).

**Dataset**  
Vitis vinifera genes (PN40024.v4)

**Filters**  
NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified]

**Attributes**  
Gene stable ID  
NCBI gene (formerly Entrezgene) ID

Export all results to: File

Email notification to: [input field]

View: 10 rows as HTML

Unique results only: ☐

Go

HTML  
HTML  
CSV  
TSV  
XLS

2

3

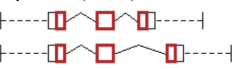
Gene stable ID	NCBI gene (formerly Entrezgene) ID
Vitv18g04132	100232844
Vitv107g00595	100232871
Vitv102g00963	100240765
Vitv108g01371	100240918
Vitv102g00460	100241009
Vitv106g00946	100241104
Vitv109g00300	100241227

Aca vemos la tabla con la correspondencia entre IDs

#### 4. Obtener las secuencias de las proteínas codificadas por estos genes (formato FASTA, aminoácidos).

Luego, para recuperar las secuencias de las proteínas codificadas por estos genes de la lista, ir a Attributes (1), seleccionar Sequences (2), desplegar SEQUENCES (3), seleccionar Peptide (4)

Please select columns to be included in the output and hit 'Results' when ready

<b>Dataset</b> Vitis vinifera genes (PN40024.v4) <b>Filters</b> NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] <b>Attributes</b> 1 Gene stable ID Transcript stable ID Peptide <b>Dataset</b> [None Selected]	<input type="radio"/> Features <input type="radio"/> Structures <input type="radio"/> Homologues (Max select 6 orthologues) <input type="radio"/> Variant (Germline) <input checked="" type="radio"/> Sequences 2
<b>SEQUENCES:</b> 3 Sequences (max 1)  <input type="radio"/> Unspliced (Transcript) <input type="radio"/> Unspliced (Gene) <input type="radio"/> Flank (Transcript) <input type="radio"/> Flank (Gene) <input type="radio"/> Flank-coding region (Transcript) <input type="radio"/> Flank-coding region (Gene) <input type="radio"/> 5' UTR <input type="radio"/> 3' UTR <input type="radio"/> Exon sequences <input type="radio"/> cDNA sequences <input type="radio"/> Coding sequence <input checked="" type="radio"/> Peptide 4	

Luego para descargar los datos en formato FASTA, ir a Results (1), Go (2)

**EnsemblPlants**

New Count Results 1

BLAST URL

<b>Dataset</b> Vitis vinifera genes (PN40024.v4) <b>Filters</b> NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] <b>Attributes</b> Gene stable ID	Export all results to: File FASTA Unique results only Go Email notification to: View: 10 rows as FASTA Unique results only >Vitvi01g01848 Vitvi01g01848_t001 Encabezado (> seguido de informacion) MDPEIGGGIAGSAVSFLLKLDVFASREWNLQENIKKAVQNLGRELRSIEALLRDAASKK EHDHQFRVWVQNRDQAYAIEDVLDLFRDLQESVWRRLKMRHSINNLIQDIDRSLSIQSIQQ TKERYHSMASSTNAGNNTDLPVRVAPQFIGNVDTVGLEPTNKLVSUALEPKQRLEVMF VVGMAGLGKTTLVHVSVERVKQHFQGNVITASKSKTKLNILTLLENLGCTITQGADV Secuencia (aminoacidos)
---	---

Pueden mirar cuales son las particularidades del formato FASTA.

## 5. Mejorar los encabezados, para hacerlos informativos: Gene stable ID, cromosoma, posición de inicio, posición final, hebra y protein stable ID

Para esto, volvemos a Attributes (1), HEADER INFORMATION (2), y elegimos las opciones mencionadas (marcadas con \*). Luego volvemos a Results y descargamos nuevamente nuestro FASTA.

<b>Dataset</b> Vitis vinifera genes (PN40024.v4) <b>Filters</b> NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] <b>Attributes</b> 1 Gene stable ID Peptide Protein stable ID Chromosome/scaffold name Gene start (bp) Gene end (bp) Strand	<b>SEQUENCES:</b> <b>HEADER INFORMATION:</b> 2 <b>Gene Information</b> <input checked="" type="checkbox"/> Gene stable ID * <input type="checkbox"/> Gene description <input checked="" type="checkbox"/> Chromosome/scaffold name * <input checked="" type="checkbox"/> Gene start (bp) * <input checked="" type="checkbox"/> Gene end (bp) * <input type="checkbox"/> Gene type <input type="checkbox"/> UniParc ID <input type="checkbox"/> UniProtKB/Swiss-Prot ID <input type="checkbox"/> UniProtKB/TrEMBL ID <b>Transcript Information</b> <input type="checkbox"/> CDS start (within cDNA) <input type="checkbox"/> CDS end (within cDNA) <input type="checkbox"/> 5' UTR start <input type="checkbox"/> 5' UTR end <input type="checkbox"/> 3' UTR start <input type="checkbox"/> 3' UTR end <input type="checkbox"/> Transcript stable ID <input checked="" type="checkbox"/> Protein stable ID * <input type="checkbox"/> Transcript type <input checked="" type="checkbox"/> Strand * <input type="checkbox"/> Transcript start (bp) <input type="checkbox"/> Transcript end (bp) <input type="checkbox"/> Transcript start site (TSS) <input type="checkbox"/> Transcript length (including UTRs and CDS)
---	--

Si exploramos los encabezados, ahora tienen más información (cada campo

separado por pipe, "|")

6. Tenemos este gráfico de enriquecimiento GO, donde se ven los procesos biológicos enriquecidos en los genes de nuestra lista de DEGs *upregulados* a las 72h post infección con el hongo. Vemos que “protein fosforilation” es el término más representado. Averiguar qué genes de mi lista tienen ese GO BP (NOTA: el código es GO:0006468. Para buscar términos GO, ir a <https://www.ebi.ac.uk/QuickGO/>).



Para esto, debemos volver a modificar nuestros filtros. Vamos a filters (1), dejar el filtro de GENE como estaba y sumar GENE ONTOLOGY (2), seleccionar GO term accession (3) y escribir nuestro GO de interés en el casillero (4). Luego descargar los atributos que queramos para esos genes (ejemplo, su Gene stable ID y si NCBI Gene ID).

<b>Dataset</b> Vitis vinifera genes (PN40024.v4) <b>Filters</b> 1 NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] GO Term Accession (e.g. GO:0050789) [Max 500 advised]: [ID-list specified] <b>Attributes</b> Gene stable ID NCBI gene (formerly Entrezgene) ID  <b>Dataset</b> [None Selected]	<b>Please restrict your query using criteria below</b> (If filter values are truncated in any lists, hover over the list item to see the full text)	
	<input type="checkbox"/> REGION: <input type="checkbox"/> GENE: <input type="checkbox"/> GENE ONTOLOGY: 2 <input checked="" type="checkbox"/> GO Term Accession (e.g. GO:0050789) [Max 500 advised] 3 <input type="checkbox"/> GO Term Name [e.g. regulation of biological process] <input type="checkbox"/> GO Evidence code	<div> GO:0006468 4  Choose File No file chosen </div> <div> <input type="text"/>  Choose File No file chosen </div> <div> IBA  IDA  IEA  IEP  IGI </div>

## 7. Averiguar cuántos genes de mi lista codifican para quinasas.

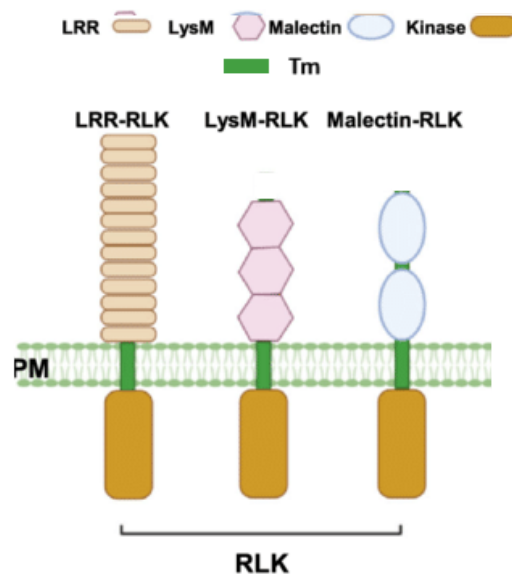
Para esto, tenemos que saber cómo identificar una quinasas. Una forma es conociendo sus dominios, y cómo están codificados en alguna base de datos de dominios. Una de las BD más usadas para esto es PFAM, y el dominio quinasas es PF00069. Podemos filtrar entonces nuestros genes para quedarnos solo con aquellos que tengan ese dominio.

Para esto, ir a Filters (1), dejar el filtro de GENE como estaba y ELIMINAR el de GENE ONTOLOGY del punto anterior, ir a PROTEIN DOMAINS AND FAMILIES (2), seleccionar Limit to genes with these family or domains IDs (3), seleccionar en el desplegable Pfam IDs (4, ojo pueden usar cualquier otra DB, como Interpro, PANTHER, etc, sabiendo el ID del dominio quinasas), escribir en el casillero nuestro Pfam de interés (5, ojo que es sensible a mayúsculas y minúsculas... se escribe PFXXXX).

<b>Dataset</b> Vitis vinifera genes (PN40024.v4) <b>Filters</b> 1 NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] Pfam ID(s) [e.g. PF00004]: [ID-list specified] <b>Attributes</b> Gene stable ID NCBI gene (formerly Entrezgene) ID  <b>Dataset</b> [None Selected]	<b>Please restrict your query using criteria below</b> (If filter values are truncated in any lists, hover over the list item to see the full text)	
	<input type="checkbox"/> REGION: <input type="checkbox"/> GENE: <input type="checkbox"/> GENE ONTOLOGY: <input type="checkbox"/> MULTI SPECIES COMPARISONS: <input type="checkbox"/> PROTEIN DOMAINS AND FAMILIES: 2 <input type="checkbox"/> Limit to genes ... <input checked="" type="checkbox"/> Limit to genes with these family or domain IDs [Max 500 advised] 3	<div> With Pfam ID(s) Only  Excluded </div> <div> Pfam ID(s) [e.g. PF00004] 4  PF00069 5  Choose File No file chosen </div>

## 8. Identificar cuáles de estas quinasas DEG tienen dominios

transmembrana, como posibles RLK (receptor like kinase, proteínas que participan en procesos de defensa a patógenos en plantas).



Para esto, dejar intactos nuestros filtros (i.e. el de GENE y el de GENE ONTOLOGY, e ir a Attributes (1), seleccionar PROTEIN DOMAINS AND FAMILIES (2) y elegir Transmembrane helix (3). Results y Go para descargar.

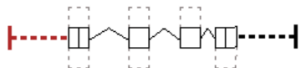
<b>Dataset</b> Vitis vinifera genes (PN40024.v4) <b>Filters</b> NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]; [ID-list specified] Pfam ID(s) [e.g. PF00004]; [ID-list specified] <b>Attributes</b> Gene stable ID NCBI gene (formerly Entrezgene) ID Transmembrane helices <b>Dataset</b> [None Selected]	<b>EXTERNAL</b> <b>PROTEIN DOMAINS AND FAMILIES:</b> 2 <b>Domains</b> <input type="checkbox"/> CDD ID <input type="checkbox"/> CDD start <input type="checkbox"/> CDD end <input type="checkbox"/> Gene3D ID <input type="checkbox"/> Gene3D start <input type="checkbox"/> Gene3D end <input type="checkbox"/> HAMAP ID <input type="checkbox"/> HAMAP start <input type="checkbox"/> HAMAP end <input type="checkbox"/> NCBI/FAM ID <input type="checkbox"/> NCBI/FAM start <input type="checkbox"/> NCBI/FAM end <input type="checkbox"/> PANTHER ID <input type="checkbox"/> PANTHER start <input type="checkbox"/> PANTHER end <input type="checkbox"/> Pfam ID <input type="checkbox"/> Pfam start <input type="checkbox"/> Pfam end <input type="checkbox"/> PIRSF ID <input type="checkbox"/> PIRSF start <input type="checkbox"/> PIRSF end <b>Interpro</b> <input type="checkbox"/> Interpro ID <input type="checkbox"/> Interpro Short Description <input type="checkbox"/> Interpro Description <b>Protein features</b> <input type="checkbox"/> AlphaFold DB import <input type="checkbox"/> AlphaFold DB import start <input type="checkbox"/> AlphaFold DB import end <input type="checkbox"/> MobiDB lite <input type="checkbox"/> MobiDB lite start <input type="checkbox"/> MobiDB lite end <input type="checkbox"/> Coiled-coils (Ncoils) <input type="checkbox"/> Coiled-coils (Ncoils) start <input type="checkbox"/> Coiled-coils (Ncoils) end <input type="checkbox"/> Prints ID <input type="checkbox"/> Prints start <input type="checkbox"/> Prints end <input type="checkbox"/> PROSITE patterns ID <input type="checkbox"/> PROSITE patterns start <input type="checkbox"/> PROSITE patterns end <input type="checkbox"/> PROSITE profiles ID <input type="checkbox"/> PROSITE profiles start <input type="checkbox"/> PROSITE profiles end <input type="checkbox"/> SFLD ID <input type="checkbox"/> SFLD start <input type="checkbox"/> SFLD end <input type="checkbox"/> SMART ID <input type="checkbox"/> SMART start <input type="checkbox"/> SMART end <input type="checkbox"/> Superfamily ID <input type="checkbox"/> Superfamily start <input type="checkbox"/> Superfamily end <input type="checkbox"/> TIGRFAM ID <input type="checkbox"/> TIGRFAM start <input type="checkbox"/> TIGRFAM end <input type="checkbox"/> Low complexity (Seg) <input type="checkbox"/> Low complexity (Seg) start <input type="checkbox"/> Low complexity (Seg) end <input type="checkbox"/> Cleavage site (Signalp) <input type="checkbox"/> Cleavage site (Signalp) start <input type="checkbox"/> Cleavage site (Signalp) end <input checked="" type="checkbox"/> Transmembrane helices 3 <input type="checkbox"/> Transmembrane helices start <input type="checkbox"/> Transmembrane helices end
---	--

**9. Recuperar en formato FASTA (nucleótidos) la región río arriba de cada una de estas quinasas (1000 pb río arriba)**

Para esto, ir Attributes (1), seleccionar Sequences (2), desplegar SEQUENCES (3), seleccionar Flank (Gene, 4) y luego en Upstream flank poner los pb deseados













(1000, 5)

<b>Dataset</b> Vitis vinifera genes (PN40024.v4)	<input type="radio"/> Features <input type="radio"/> Structures <input type="radio"/> Homologues (Max select 6 orthologues)	<input type="radio"/> Variant (Germline) <input checked="" type="radio"/> Sequences 2
<b>Filters</b> NCBI gene (formerly Entrezgene) ID(s) [e.g. 100232840]: [ID-list specified] Pfam ID(s) [e.g. PF00004]: [ID-list specified]	<input type="checkbox"/> SEQUENCES: 3	
<b>Attributes</b> 1 Gene stable ID Protein stable ID Chromosome/scaffold name Gene start (bp) Gene end (bp) Strand Upstream flank [1000]	<b>Sequences (max 1)</b>  <input type="radio"/> Unspliced (Transcript) <input type="radio"/> Unspliced (Gene) <input type="radio"/> Flank (Transcript) <input checked="" type="radio"/> Flank (Gene) 4 <input type="radio"/> Flank-coding region (Transcript) <input type="radio"/> Flank-coding region (Gene) <input type="radio"/> 5' UTR <input type="radio"/> 3' UTR <input type="radio"/> Exon sequences <input type="radio"/> cDNA sequences <input type="radio"/> Coding sequence <input type="radio"/> Peptide	
	<b>Upstream flank</b> <input checked="" type="checkbox"/> Upstream flank 1000 5	

### Actividad 3.

#### 1. Encontrar el sitio para *S. cerevisiae*

 Bombyx mori ASM115162v1 <a href="#">Go to Ensembl Metazoa</a>	 Phytophthora infestans ASM14294v1 <a href="#">Go to Ensembl Protists</a>
  Magnaporthe oryzae MG8  Saccharomyces cerevisiae R64-1-1  Aspergillus nidulans ASM1142v1 <a href="#">Go to Ensembl Fungi</a>	  Streptococcus pneumoniae ASM688v1  Escherichia coli ASM584v2  Bacillus subtilis ASM73511v1 <a href="#">Go to Ensembl Bacteria</a>

Vamos a estar trabajando en Ensembl Fungi (<https://fungi.ensembl.org/index.html>).

El sitio de *S. cerevisiae* es [https://fungi.ensembl.org/Saccharomyces\\_cerevisiae/Info/Index](https://fungi.ensembl.org/Saccharomyces_cerevisiae/Info/Index).

#### 2. Registrar el nombre del *assembly* (genoma de referencia), su tamaño en pb, número de cromosomas y cantidad de genes anotados (pista: More information and statistics)

Assembly R64-1-1, INSDC Assembly GCA\_000146045.2, Sep 2011. Genome Length: 12,157,105. Coding genes: 6600



## Saccharomyces cerevisiae Assembly and Gene Annotation

### About *Saccharomyces cerevisiae*

*Saccharomyces cerevisiae* is a unicellular fungus. It is commonly known as baker's, brewer's or budding yeast. It is used in the production of a number of human foodstuffs, including alcoholic beverages and in the baking industry, and is widely used as a model species in the study of eukaryotic biology. In 1996, the genome of *S. cerevisiae* was the first eukaryotic genome to be completely deciphered.

Image courtesy of American Society for Microbiology

### Assembly

The assembly provided on this site is the R64-1-1 assembly, imported from the Saccharomyces Genome Database ([SGD](#)).

### Annotation

... (#annotation) The protein-coding and non-coding gene model annotation was imported from the Saccharomyces Genome Database ([SGD](#)) in April 2018. This gene set is based on [Liachko et al. 2013](#) [6], and contains 7036 protein-coding genes. Additionally, 91 transposable element genes (TE genes) that were previously annotated as protein-coding genes have now been correctly captured as TE genes. ...

### Variation

The variation data provided by this site were imported from data provided by the Saccharomyces Genome Resequencing Project ([SGRP](#)).

### Statistics

#### Summary

<b>Assembly</b>	R64-1-1 (Saccharomyces cerevisiae S288c assembly from Saccharomyces Genome Database), INSDC Assembly <a href="#">GCA_000146045.2</a> , Sep 2011
<b>Database version</b>	112.4
<b>Golden Path Length</b>	12,157,105
<b>Genebuild by</b>	SGD
<b>Genebuild method</b>	Import
<b>Data source</b>	<a href="#">SGD</a>

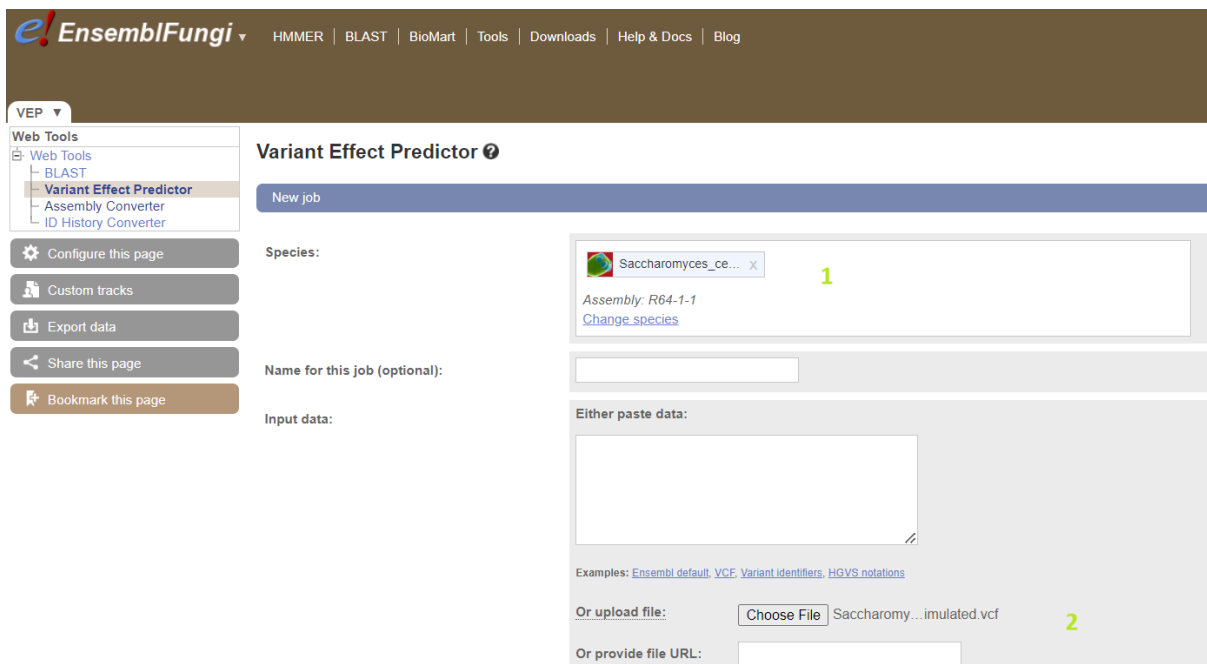
#### Gene counts

<b>Coding genes</b>	6,600
<b>Non coding genes</b>	424
<b>Small non coding genes</b>	424
<b>Pseudogenes</b>	12
<b>Gene transcripts</b>	7,127

## Actividad 4.

### 1. Usando VEP, predecir el efecto de las variantes disponibles en el VCF (descargar el VCF del GitHub de RSG Uruguay)

Para esto, ir a Tools (barra superior), seleccionar VEP. Luego elegir la especie (1, *S. cerevisiae* está por defecto, pero si quisieran elegir otra ir a “change species”). Luego subir los datos en el formato aceptado (uno de ellos es VCF, variant call format, y es el formato de nuestro archivo, así que directamente lo subimos desde Upload file (2)).



**EnsemblFungi** | HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

**VEP** ▼

- Web Tools
  - Web Tools
  - BLAST
  - Variant Effect Predictor**
  - Assembly Converter
  - ID History Converter
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

**Variant Effect Predictor** ?

**New job**

**Species:** Saccharomyces cerevisiae 1

**Assembly:** R64-1-1  
[Change species](#)

**Name for this job (optional):**

**Input data:**

**Either paste data:**

Examples: [Ensembl default VCF](#), [Variant identifiers](#), [HGVS notations](#)

**Or upload file:**   2

**Or provide file URL:**

Si bien podemos ejecutar por defecto, vamos a cambiar algun parametro (Additional configurations, 1). En este caso, vamos a modificar cómo define las relaciones Upstream/ downstream (additional annotations, 2): en vez de usar 5000 pb (defecto) usaremos 1000 pb (3).

Additional configurations:

1

Identifiers

Additional identifiers for genes, transcripts and variants

Variants and frequency data

Co-located variants and frequency data

Additional annotations

Additional transcript, protein and regulatory annotations

Transcript annotation

Transcript biotype:

☒

Exon and intron numbers:

☐

Identify canonical transcripts:

☐

Upstream/Downstream distance (bp):

1000

miRNA structure:

☐

NMD:

☐

UTRAnnotator:

☐

Luego Run (botón verde inferior) y esperamos (va a pasar de estado qued, a running, a done)

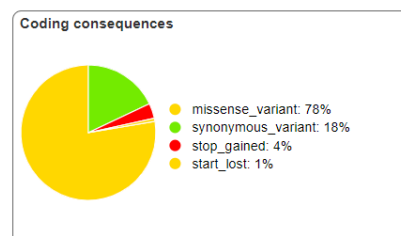
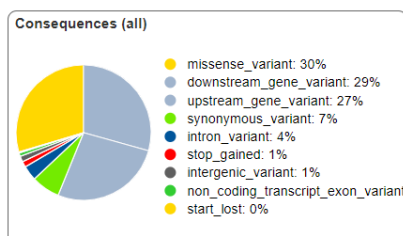
Show/hide columns (1 hidden)		Filter
Analysis	Jobs	Submitted at (GMT)
Variant Effect Predictor	VEP analysis of Saccharomyces_cerevisiae_simulated.vcf in Saccharomyces_cerevisiae <span>Done</span> <a href="#">View results</a>	30/09/2024, 13:41

## 2. ¿Cuántas variantes tiene nuestro archivo?

Teníamos 188 variantes, como se ve en la tabla de estadísticas resumen.

Summary statistics

Category	Count
Variants processed	188
Variants filtered out	0
Novel / existing variants	187 (99.5) / 1 (0.5)
Overlapped genes	328
Overlapped transcripts	328
Overlapped regulatory features	-



## 3. ¿Cuáles son las consecuencias más frecuentes?

Segun el gráfico de torta superior, missense variants (30%), seguido de las downstream (29%), seguido de las upstream (27% de las variantes)

#### 4. ¿Cuántas variantes generan un “missense\_variant”? (pista: usar filtros, consequences)

Podemos usar el filtro por consequences, y pedir missense\_variant

Results preview

Navigation (per variant) | Filters | Download | New job

Page: 1 of 38 | Show: 1 5 10 50 All variants

Consequence is missense\_variant Add

All: VCF VEP TXT

Recuperamos múltiples variantes:

Show/hide columns (20 hidden)									
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Exon	Amino acids	Feature strand	
I-41774	<a href="#">I:41774-41774</a>	G	missense_variant	MODERATE	GPB2	1/1	S/C	Scroll to see more columns »	
I-109268	<a href="#">I:109268-109268</a>	A	missense_variant	MODERATE	FUN26	1/1	G/V		
I-134304	<a href="#">I:134304-134304</a>	G	missense_variant	MODERATE	MDM10	1/1	E/D		
I-183816	<a href="#">I:183816-183816</a>	T	missense_variant	MODERATE	UIP3	1/1	D/V		
II-80410	<a href="#">II:80410-80410</a>	G	missense_variant	MODERATE	ATG8	1/1	T/P		

#### 5. ¿Cuántas variantes son de impacto alto (HIGH)? Cuáles son sus consecuencias? (Ojo que los filtros son sensibles a mayúsculas/minúsculas)

Nuevamente podemos filtrar, ahora por impacto (HIGH). Presionar add

Results preview

Navigation (per variant) | Filters | Download | New job

Page: 1 of 38 | Show: 1 5 10 50 All variants

Impact is HIGH Add

All: VCF VEP TXT

Recuperamos estas seis (6) variantes, dos de stop\_gained y una start\_lost

Show/hide columns (20 hidden)									
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Exon	Amino acids		
IX-153860	<a href="#">IX:153860-153860</a>	T	stop_gained	HIGH	HOS4	1/1	K/*		
IX-394555	<a href="#">IX:394555-394555</a>	G	start_lost	HIGH	-	1/1	M/I		
V-510463	<a href="#">V:510463-510463</a>	T	stop_gained	HIGH	PAB1	1/1	Q/*		
VI-128931	<a href="#">VI:128931-128931</a>	A	stop_gained	HIGH	BLM10	1/1	S/*		
XI-263865	<a href="#">XI:263865-263865</a>	T	stop_gained	HIGH	YJU3	1/1	K/*		
XI-538488	<a href="#">XI:538488-538488</a>	T	stop_gained	HIGH	DYN1	1/1	Y/*		

#### 6. Hay variantes que afectan el gen de la PDC1 (Pyruvate Decarboxylase 1)? (pista: su geneID/Symbol es PDC1, y está en cromosoma XII: posiciones 232390-234081)

Podemos filtrar por Symbol:

Navigation (per variant) Filters Download New job

Show: 1 5 10 50 All variants Symbol is PDC1 Add All: VCF VEP TXT

O por location (XII:232390-234081, se escribe cromosoma:inicio:final, sin espacios).

En cualquiera de los casos es lo mismo, no hay variantes que afecten a ese gen. Podemos extendernos y mirar 1000 pb rio arriba o abajo location (XII:231390-235081) pero da lo mismo.

Results preview

Navigation (per variant) Filters Download New job

Show: 1 5 10 50 All variants Location is XII:231390-235081

Uploaded variant is defined Add All: VCF VEP TXT

Show/hide columns (20 hidden)

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Exon	Amino acids
No data available in table							

Show: 1 5 10 50 All variants

## Actividad 5 (integradora)

1. No encontramos variantes que afectan el gen de la PDC1... pero quizá a alguno de sus parálogos? (Si es que tiene).

Buscar los parálogos de PDC1. Si hay varios, quedarse con el de mayor % de similitud (pista: BioMart) → luego VEP p/ variantes, y sus consecuencias)

En BioMart, elegir como dataset los genes de *S. cerevisiae*

<b>Dataset</b>	Ensembl Fungi Genes 59
Saccharomyces cerevisiae genes (R64-1-1)	Saccharomyces cerevisiae genes (R64-1-1)

Luego en filter → seleccionar GENE → input external references ID list → en el desplegable, seleccionar gene Names → Escribir en el recuadro el nombre de nuestro gen (PDC1)

<b>Dataset</b>	REGION:	
Saccharomyces cerevisiae genes (R64-1-1)	GENE:	
<b>Filters</b>	<input type="checkbox"/> Limit to genes (external references)... With ChEMBL ID(s) <input checked="" type="radio"/> Only <input type="radio"/> Excluded	
[None selected]	<input checked="" type="checkbox"/> Input external references ID list [Max 500 advised]	
Gene Name(s) [e.g. SCO2]: [ID-list specified]	Gene Name(s) [e.g. SCO2] PDC1 Choose File No file chosen	
<b>Attributes</b>		
Gene stable ID		
Transcript stable ID		

Luego en Attributes (1), seleccionar Homologues (2), PARALOGUES (3) y los datos de los parálogos (ID, posición en el genoma (para luego poder buscar SNPs en esas regiones), % de identidad).

Please select columns to be included in the output and hit 'Results' when ready

☐ Features
 ☐ Variant (Germline)

☐ Structures
 ☐ Sequences

☒ Homologues (Max select 6 orthologues) 2

☐ GENE:

☐ ORTHOLOGUES [A-E]:

☐ ORTHOLOGUES [F-J]:

☐ ORTHOLOGUES [K-O]:

☐ ORTHOLOGUES [P-T]:

☐ ORTHOLOGUES [U-Z]:

☒ PARALOGUES: 3

**Saccharomyces cerevisiae Paralogues**

☒ Saccharomyces cerevisiae paralogue gene stable ID \*
 ☐ Paralogue query protein or transcript ID

☐ Saccharomyces cerevisiae paralogue associated gene name
 ☐ Paralogue last common ancestor with Saccharomyces cerevisiae

☐ Saccharomyces cerevisiae paralogue protein or transcript ID
 ☐ Saccharomyces cerevisiae paralogue homology type

☒ Saccharomyces cerevisiae paralogue chromosome/scaffold name
 ☒ Paralogue %id. target Saccharomyces cerevisiae gene identical to query gene \*

☒ Saccharomyces cerevisiae paralogue chromosome/scaffold start (bp) \*
 ☐ Paralogue %id. query gene identical to target Saccharomyces cerevisiae gene

☒ Saccharomyces cerevisiae paralogue chromosome/scaffold end (bp)

Luego en Results podemos ver que tiene 6 parálogos, y que el que está pintado es el de mayor similitud (posible duplicación más reciente, y mantenimiento de función).

Results

Export all results to: File TSV Unique results only Go

Email notification to:

View: 10 rows as HTML Unique results only

Gene stable ID	Transcript stable ID	Saccharomyces cerevisiae paralogue gene stable ID	Saccharomyces cerevisiae paralogue chromosome/scaffold name	Saccharomyces cerevisiae paralogue chromosome/scaffold start (bp)	Saccharomyces cerevisiae paralogue chromosome/scaffold end (bp)	Paralogue %id. target Saccharomyces cerevisiae gene identical to query gene
YLR044C	YLR044C_mRNA	YDR380W	IV	1234218	1236125	34.9911
YLR044C	YLR044C_mRNA	YEL020C	V	118617	120299	17.5844
YLR044C	YLR044C_mRNA	YMR108W	XIII	484084	486147	23.2682
YLR044C	YLR044C_mRNA	YMR134W	XII	40723	412414	47.7216
YLR044C	YLR044C_mRNA	YGR087C	VII	651290	652981	84.0142
YLR044C	YLR044C_mRNA	YDL080C	IV	310642	312471	51.5098

Para buscar si hay variantes en ese parálogo, volver a la salida de VEP (queda guardada por 10 días) y filtrar por location (XII:409723-413414, con 1000 bp extra rio arriba y rio abajo)

## Exploración libre de Ensembl

Finalmente, podemos explorar ese gen (usando el browser de Ensembl). Para esto, click en el correspondiente Saccharomyces cerevisiae paralogue gene stable ID (YLR134W) y se abrirá la ventana correspondiente a este gen de Ensembl Fungi, para explorarlo:

- Gene-based displays
- Summary

Splice variants

Transcript comparison

Gene alleles

Sequence

Secondary Structure

Gene families

Literature

Fungal Compara

Genomic alignments

Gene tree

Gene gain/loss tree

Orthologues

Paralogues

Pan-taxonomic Compara

Gene Tree

Orthologues

Ontologies

GO: Cellular component

GO: Biological process

GO: Molecular function

PhI: Phibase identifier

Phenotypes

Genetic Variation

Variant table

Variant image

Structural variants

Gene expression

Pathway

Molecular interactions

Regulation

External references

Supporting evidence

ID History

Gene: PDC5 YLR134W

Description

Minor isoform of pyruvate decarboxylase; key enzyme in alcoholic fermentation, decarboxylates pyruvate to acetaldehyde, regulation is glucose- and ethanol-dependent, repressed by thiamine, involved in amino acid catabolism [Source:SGD;Acc:[S000004124](#)]

Location

[Chromosome XII: 410,723-412,414](#) forward strand.  
R64-1-1 BK006945.2

About this gene

This gene has 1 transcript ([splice variant](#)), [437 orthologues](#), [6 paralogues](#) and is a member of [1 Ensembl protein family](#)

Transcripts

Show transcript table

Summary

Name

[PDC5](#) (SGD gene name)

UniProtKB

This gene has proteins that correspond to the following UniProtKB identifiers: [P16467](#)

Gene type

Protein coding

Annotation method

Annotation imported from SGD

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

