CIBER LUNES

RSG URUGUAY

Ciclo Itinerante de Bioinformática y Exploración en Recursos Computacionales

30 DE SETIEMBRE

TALLER HANDS-ON: EXPLORANDO HERRAMIENTAS DE ENSEMBL

Uso de herramientas de Ensembl y Ensembl Genomes

13:30 - 17:30 HS

- Facultad de Agronomía
- Salón de Cómputo (junto a Cantina)

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

ISCB REGIONAL Student GROUP Uruguay

FACULTAD DE AGRONOMÍA

Carla Filippi
cfilippi@fagro.edu.uy

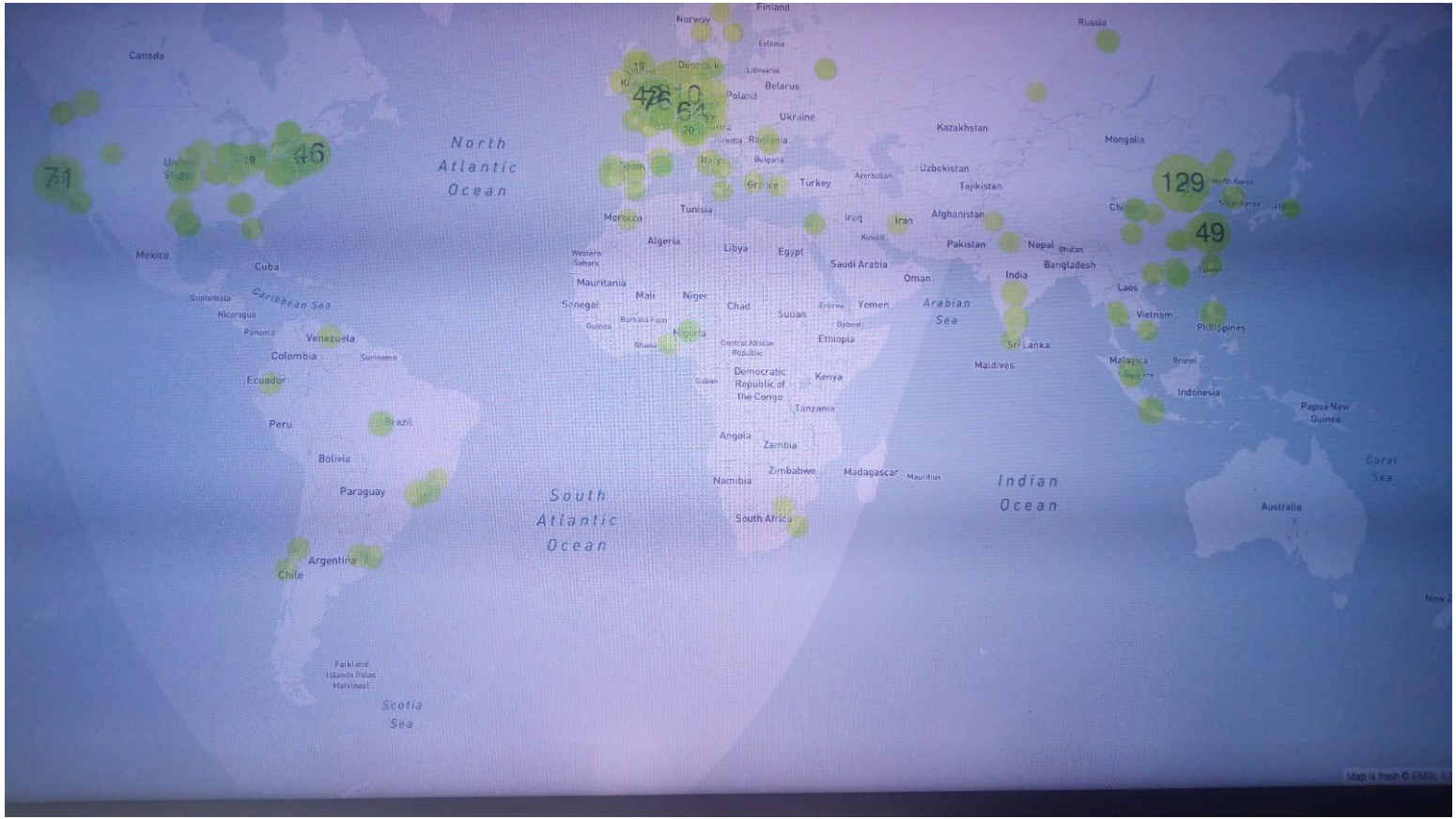
Instituto Bioinformático Europeo (EMBL-EBI, Hinxton UK)

EMBL-EBI

WELLCOME GENOME CAMPUS

wellcome sanger institute

Poniendo a Uruguay en el mapa del EMBL-EBI



Ensembl (https://www.ensembl.org/)

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes

-- Select a species --

Pig breeds

Pig reference genome and 12 additional breeds

View full list of all species

Favourite genomes

Human
GRCh38.p14

Still using GRCh37?

Mouse
GRCm39

Zebrafish
GRCz11

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 112 (May 2024)

- Many new fish genomes have been added to Ensembl
- Population frequency data is available for chicken, dog, goat and sheep through VEP
- Update to our current regulation annotation. The promoters now align with the 5' ends of known transcripts
- VEP will be updated to use the dbNSFP commercial data release

More release news on our blog

Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks.

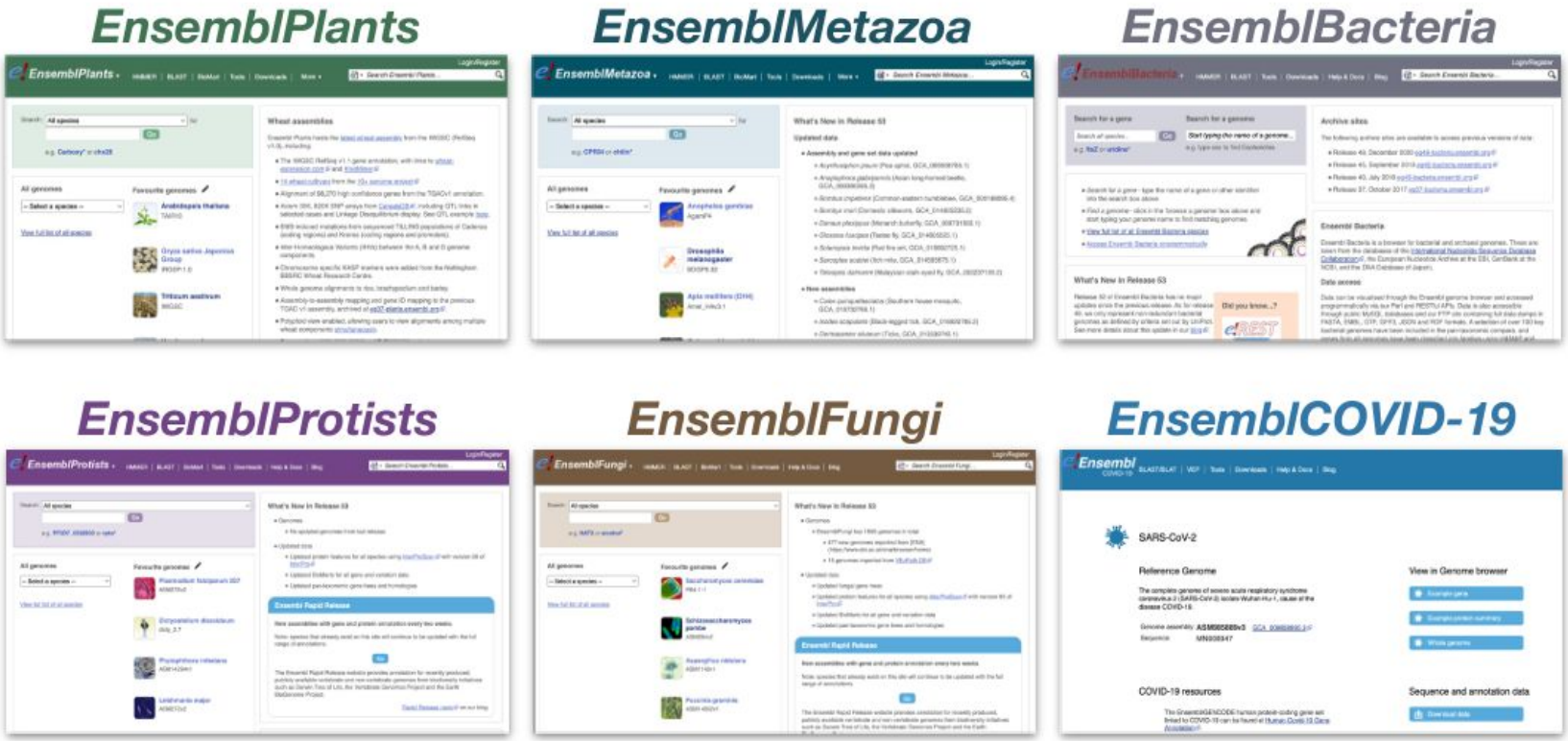
Note: species that already exist on this site will continue to be updated with the full range of annotations.

Go

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

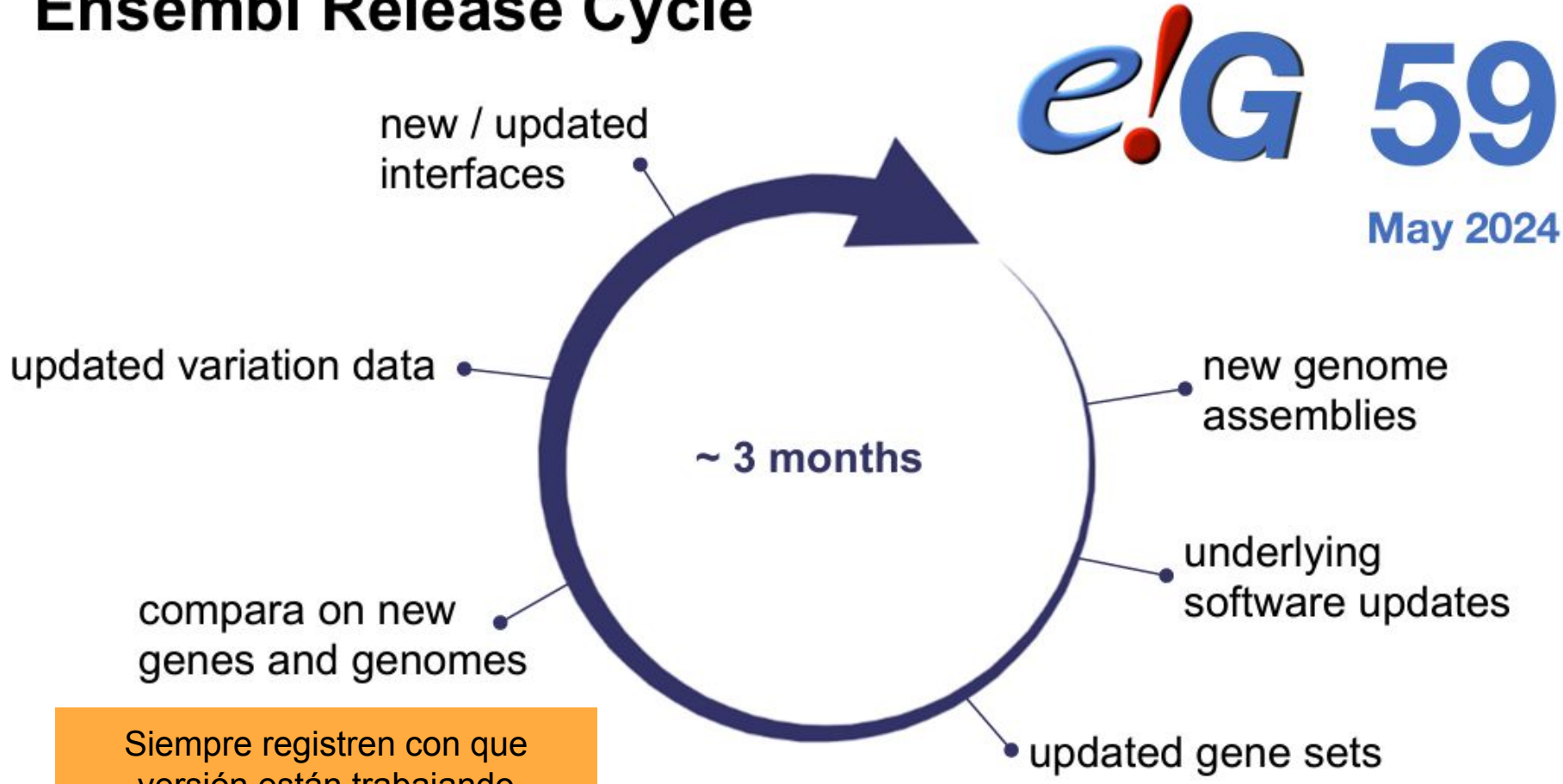
Rapid Release news on our blog

Ensembl Genomes (https://ensemblgenomes.org/)



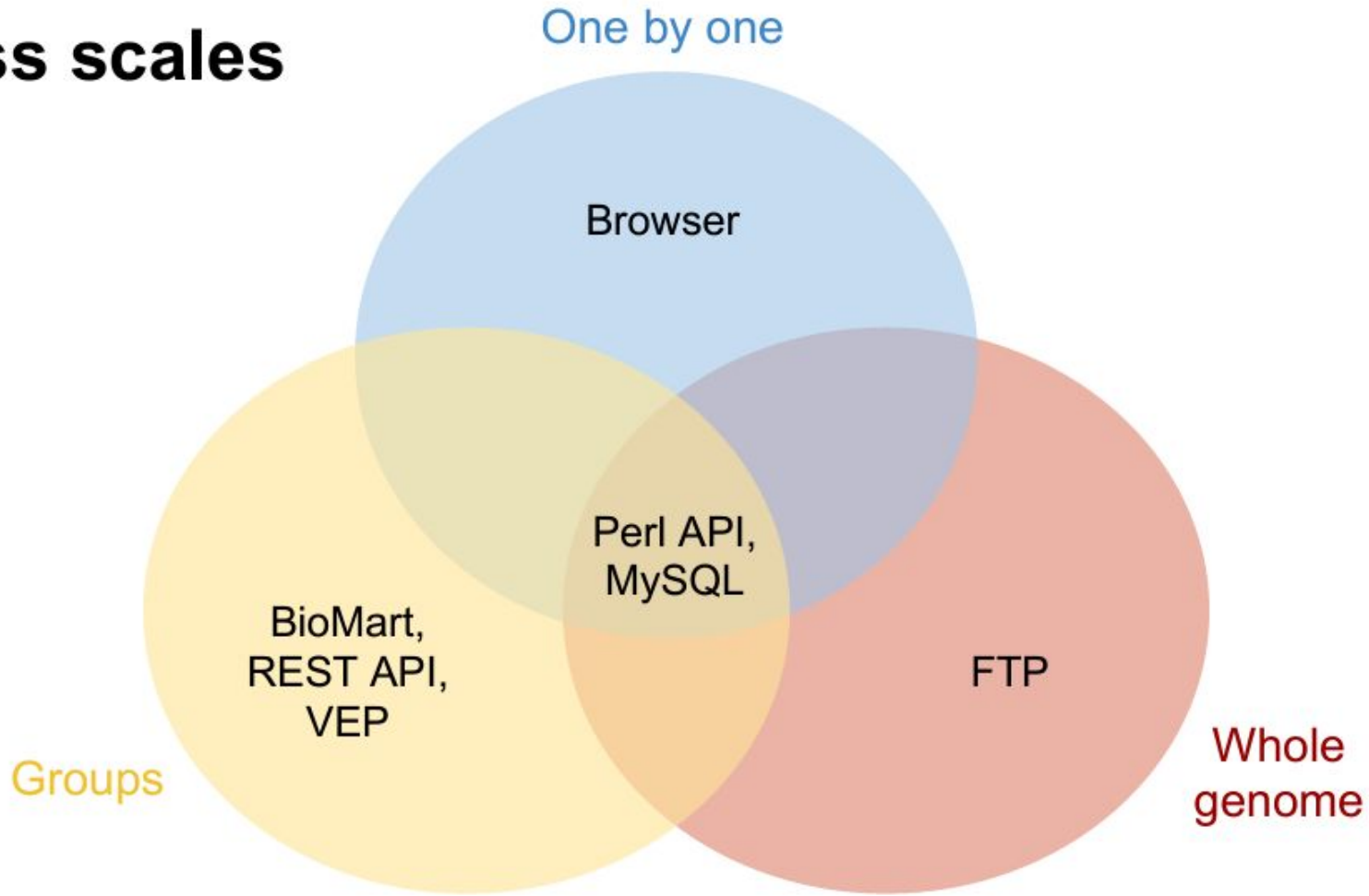
https://doi.org/10.1093/nar/gkab1007

Ensembl Release Cycle

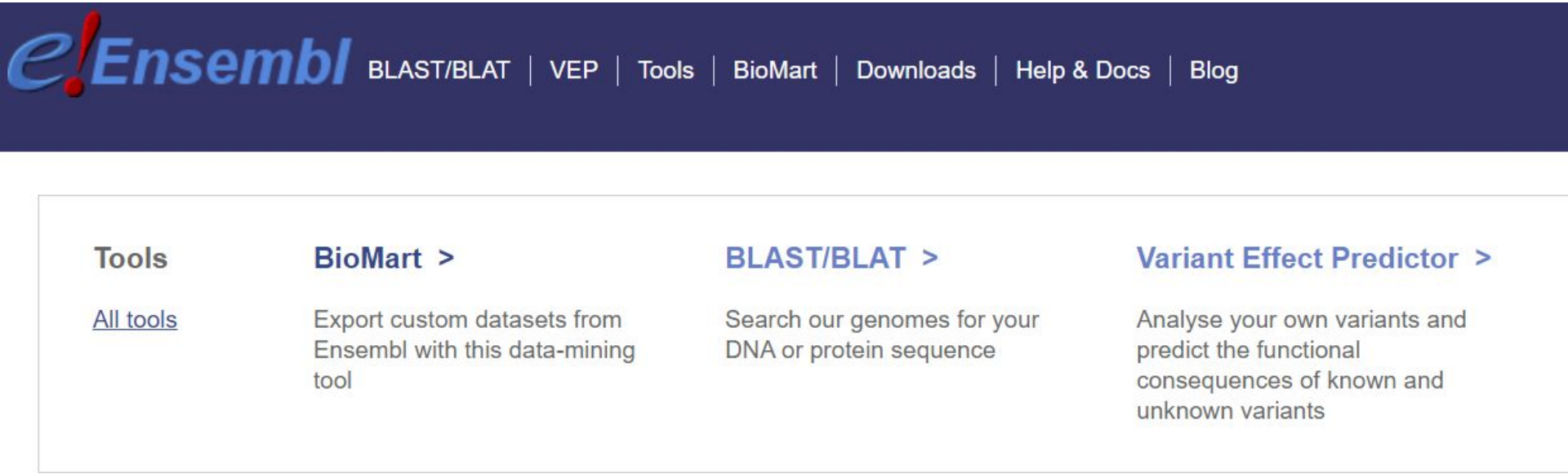


Siempre registren con que versión están trabajando

Access scales



Herramientas de Ensembl (y Ensembl genomes)



Hilo conductor del taller**

**No hay vino, solo café

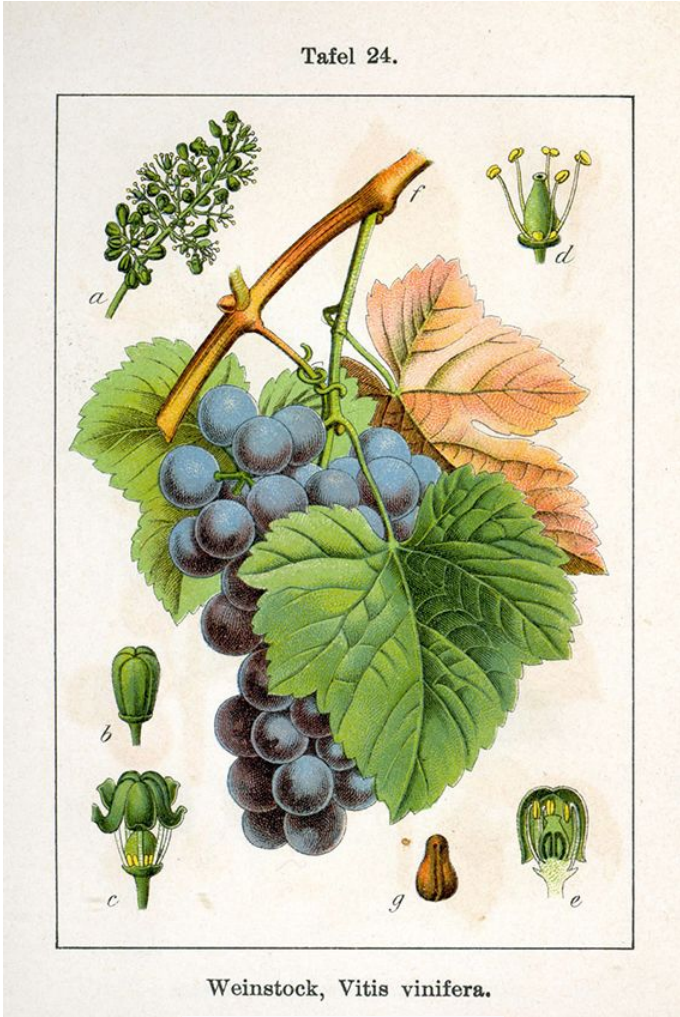
Situación 1

RNAseq de *Vitis vinifera* (vid)

Caracter: respuesta a infección con hongo (grey mold, causado por *Botrytis cinerea*)

Tratamientos: control (0h), 72h y 120h post infección.

Datos (listas de genes diferencialmente expresados, DEG) disponibles en GitHub RSG



Johann Georg Sturm, 1796

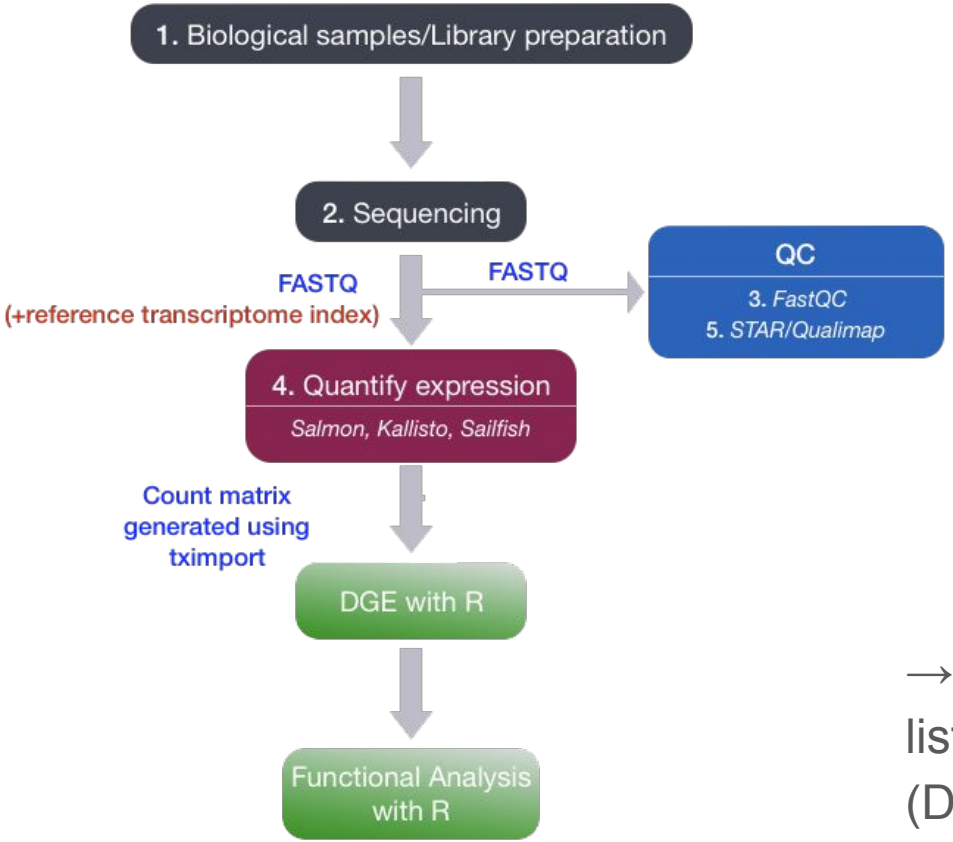
https://doi.org/10.3389/fpls.2023.1127206
SRA ID: PRJNA788159

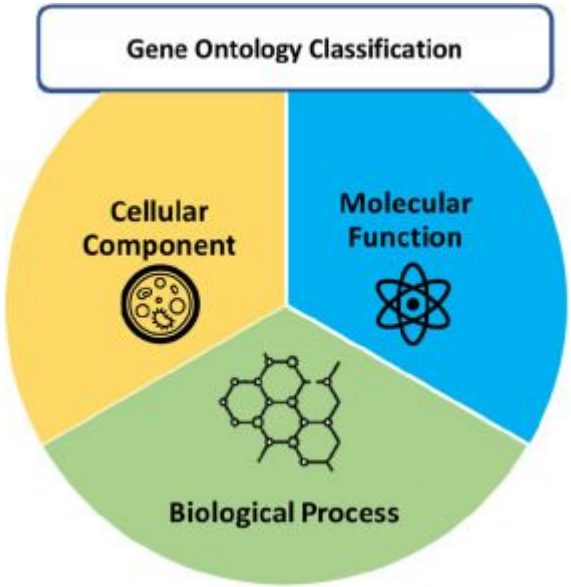
Actividad 1

Vamos a estar trabajando con datos de *Vitis vinifera*

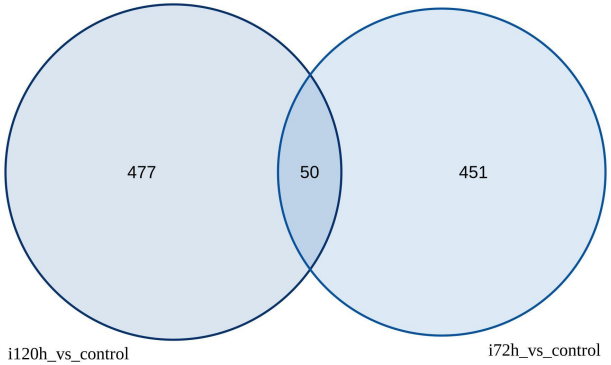
- Vayan al portal de Ensembl Genomes que corresponda
- Registren la versión actual de Ensembl Genomes
- Encuentren el sitio para *V. vinifera*
- Registren el nombre del *assembly* (genoma de referencia), su tamaño en pb, número de cromosomas y cantidad de genes anotados (pista: *More information and statistics*)

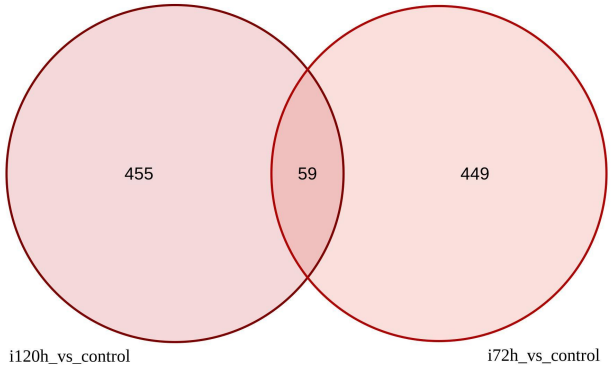
Breves lineamientos RNAseq

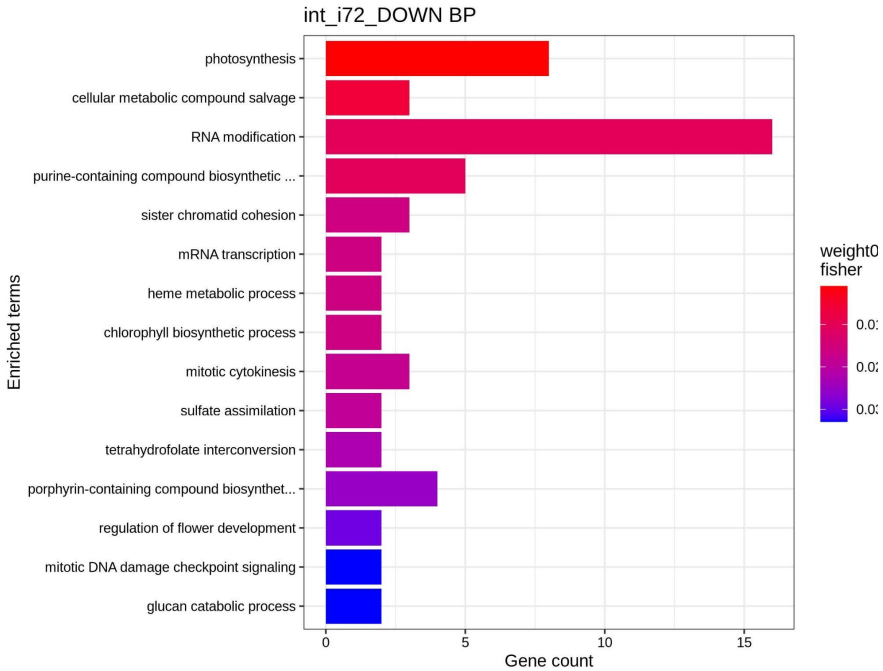


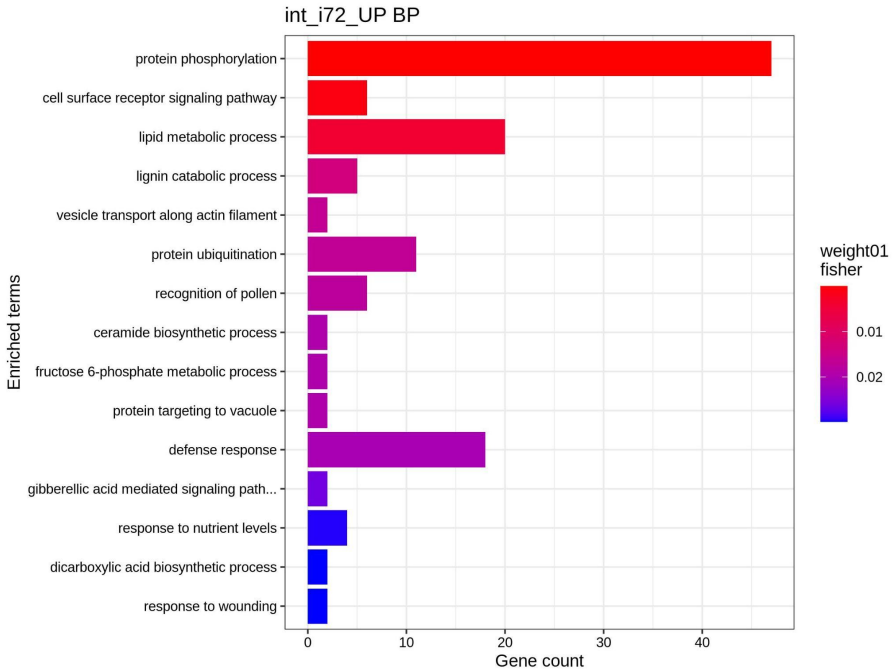


→ Su bioinformático amigo les envía las listas de genes diferencialmente expresados (DEGs) y GOs/KEGG









100268054
100251533
100266412
100256549
100263169
100255092
100262557
100245864
100242603
100263112
109122428
100264549
100242618
100255527
100242997
100268080
100250722
100259464
104880384
104879703
100232844
100251984
104881686
100266147
100254333
100266224
100265715
100260297
100232871
100247488
100255836
100250869
100251128
100257104
100264253
100252448
109123939
100244762
100253658
100264828
100248427
100248541
100257784
100263949
100261601
100257837
100267613
100252322
100262539
100262450
104878982
100265497
100246237
132253938
100267936
100246652
100253138
132252619
100251141
100262891
104882493
100248988


This is fine.

Herramienta 1: BioMart

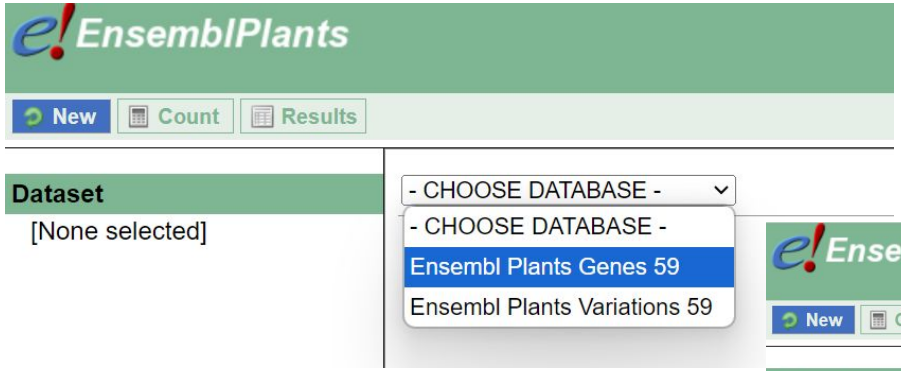
Use this data-mining tool to export custom datasets from Ensembl.

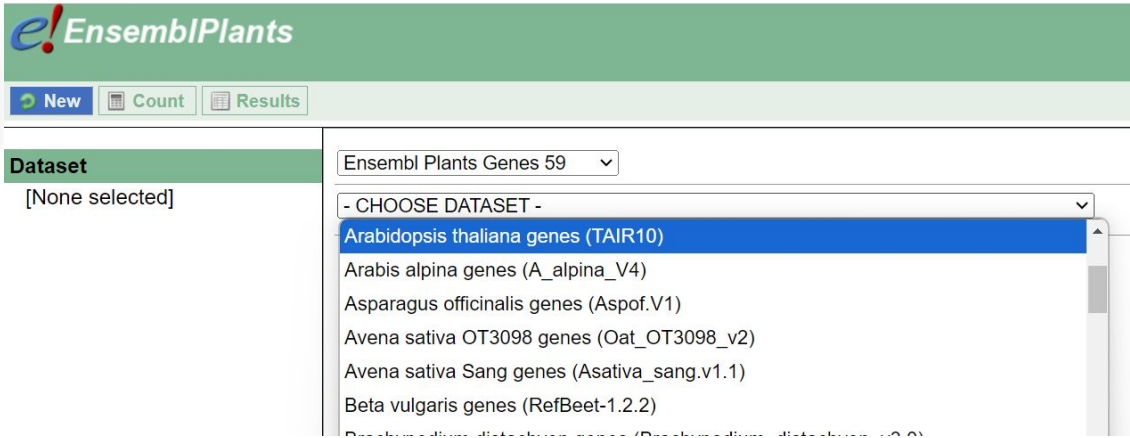
https://plants.ensembl.org/biomart/martview/

Step 1: Dataset	Step 2: Filters	Step 3: Attributes	Step 4: Results
Choose the database and species.	Narrow down the dataset.	Specify your output and what to print on your table.	Choose the format of your results and export.

Herramienta 1: BioMart

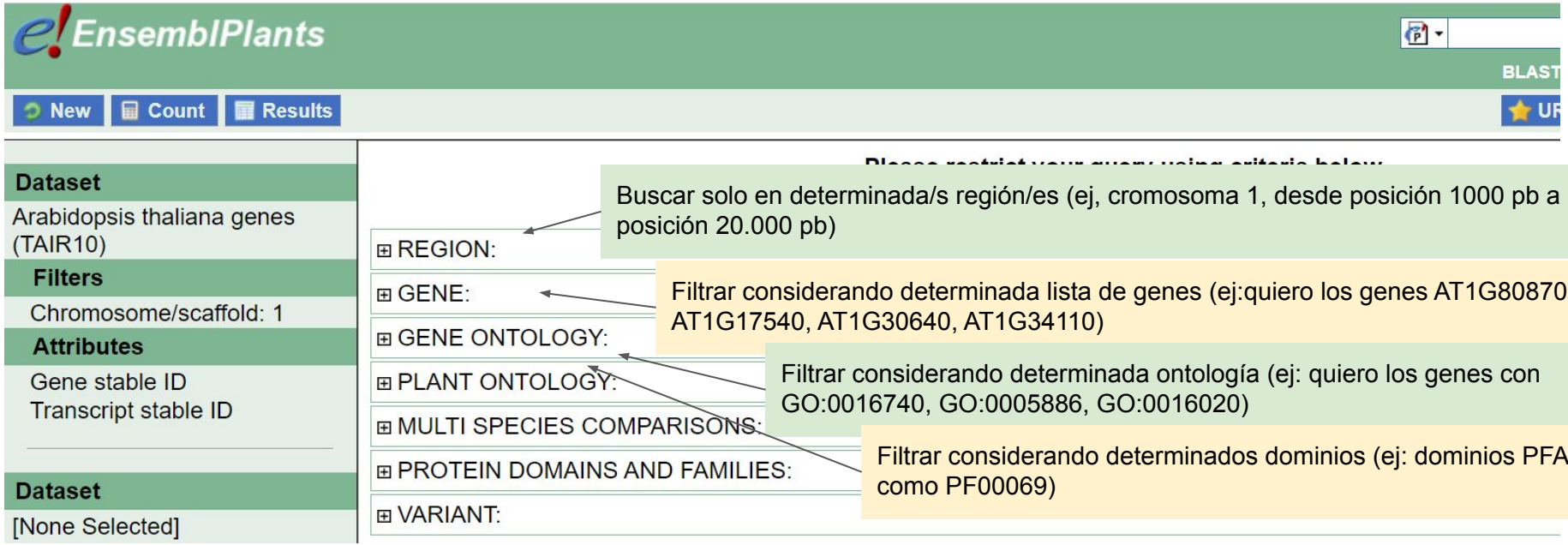
Use this data-mining tool to export custom datasets from Ensembl.





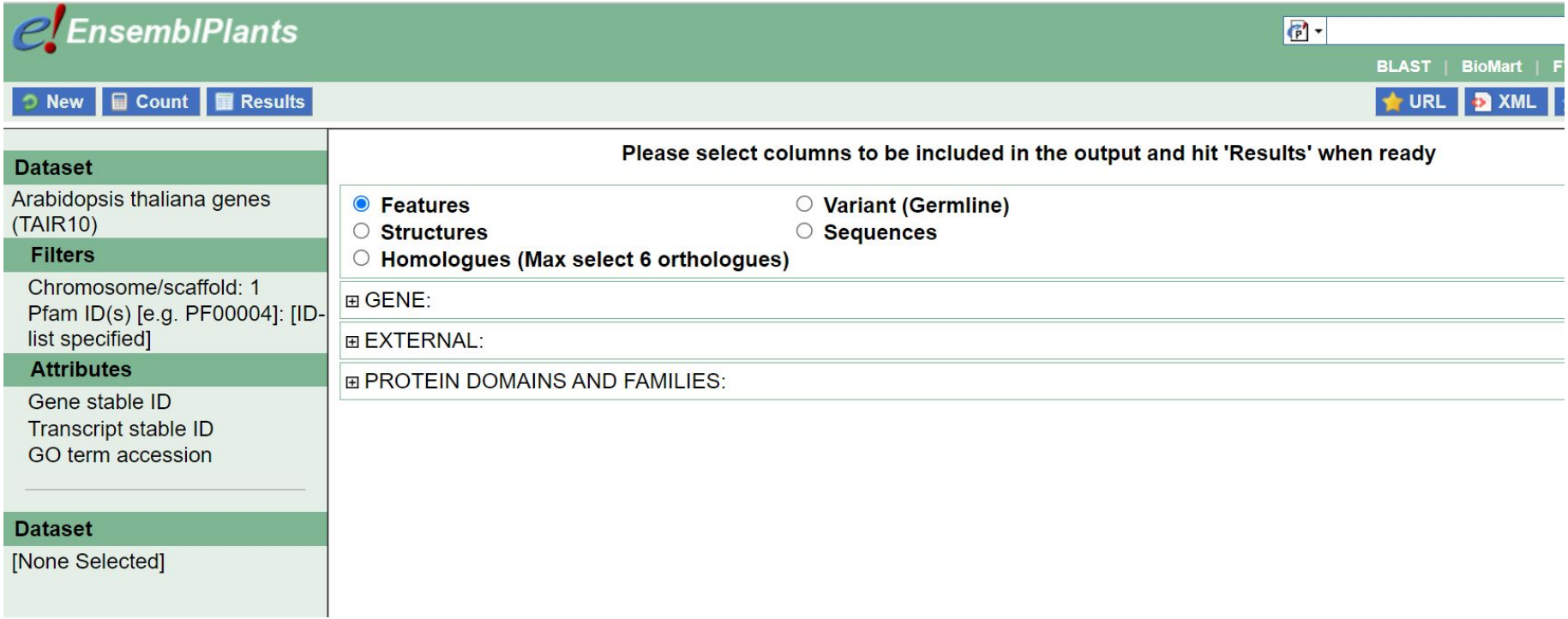
Herramienta 1: BioMart

Filters → que filtros aplicar para extraer datos de interés?



Herramienta 1: BioMart

Attributes → que quiero recuperar? (GO? secuencias? dominios? IDs?)



I NEED

coffee

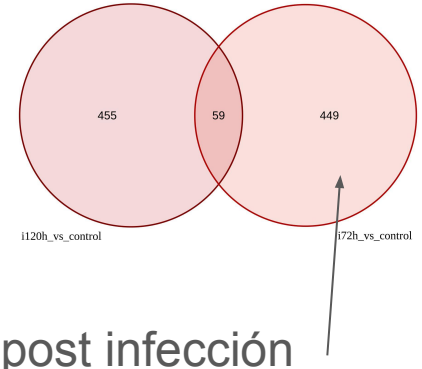
Actividad 2

Para esta actividad tendremos listas de genes diferencialmente expresados (UP: *upregulados* y DOWN, *downregulados*) → descargar de Github

Datos útiles:

- El análisis se hizo con referencia de NCBI, por lo que los IDs de las listas están en formato NCBI (anteriormente Entrez gene ID)
- Vamos a usar Biomart para responder algunas preguntas (pueden inventarse otras)

Actividad 2



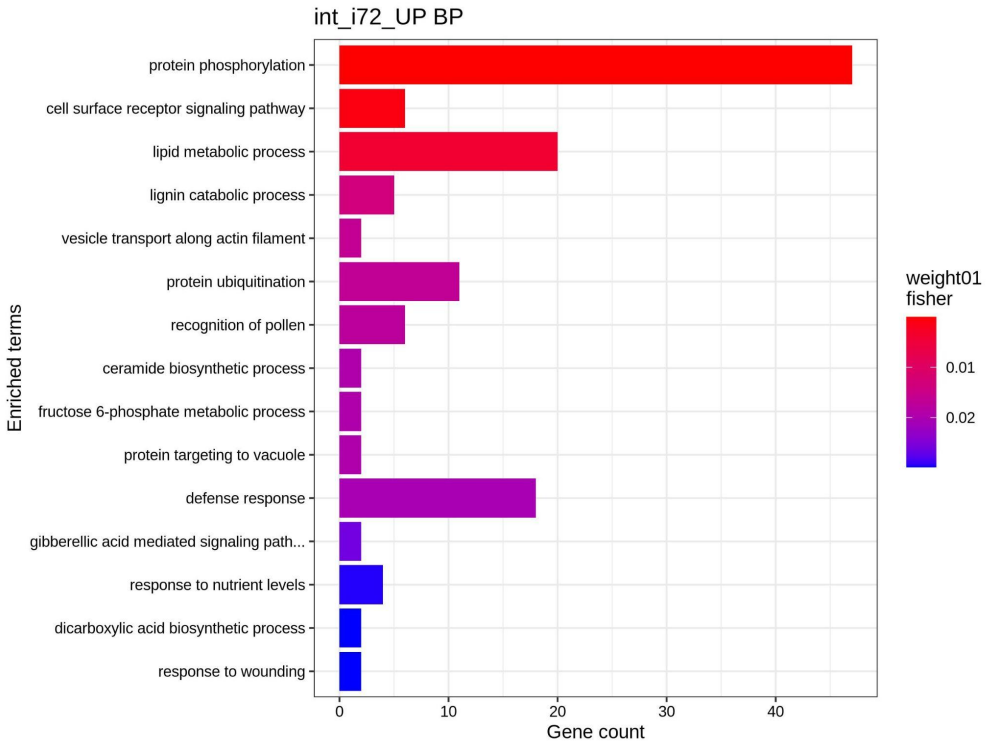
1. En Biomart, seleccionar el dataset correspondiente (i.e. *Vitis*)
2. Ingresar como filtro nuestra lista de genes upregulados a 72h post infección (cuyos IDs están codificados en formato NCBI) **unique_72h_UP.txt**
3. Obtener para cada gen de la lista el ID usado por la comunidad científica para esa especie (puntualmente para Vitis, se llaman “Vitvi_XXX”)
4. Obtener las secuencias de las proteínas codificadas por estos genes (formato FASTA, aminoácidos).

(5. opcional) Pueden mejorar los encabezados de esos FASTA haciéndolos más informativos. Ej, agregar cromosoma, start y end, etc)

Actividad 2

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 fosforilatio*n” 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/).

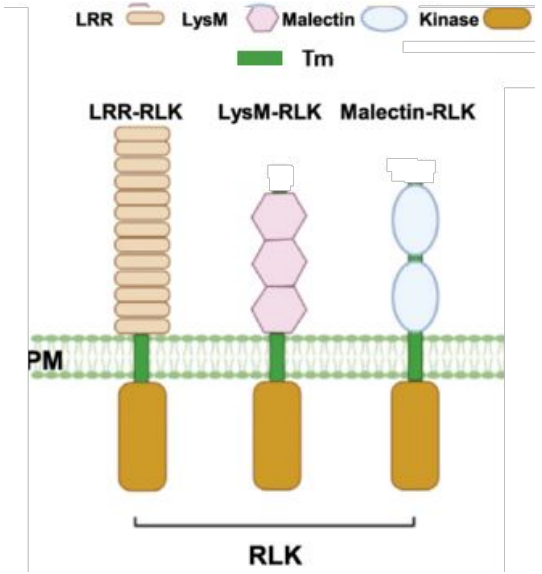


Actividad 2

7. Averiguar cuántos genes de mi lista codifican para quinasas. (pista: ver dominios, ejemplo para la base de datos PFAM, el dominio quinasa es PF00069)

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).

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



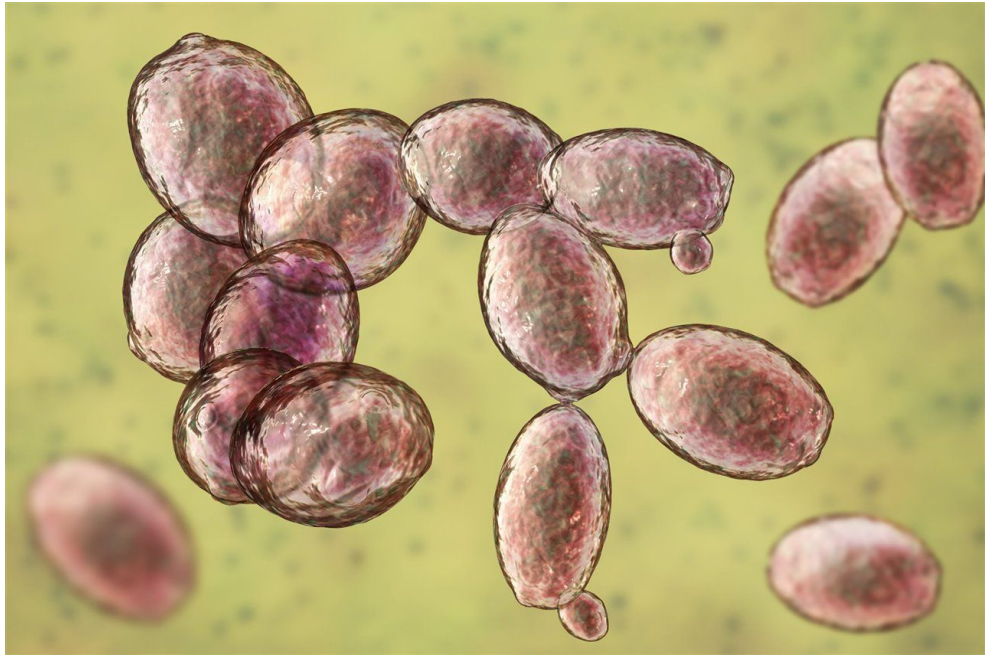

LOW BATTERY
NEED MORE COFFEE

Situación 3

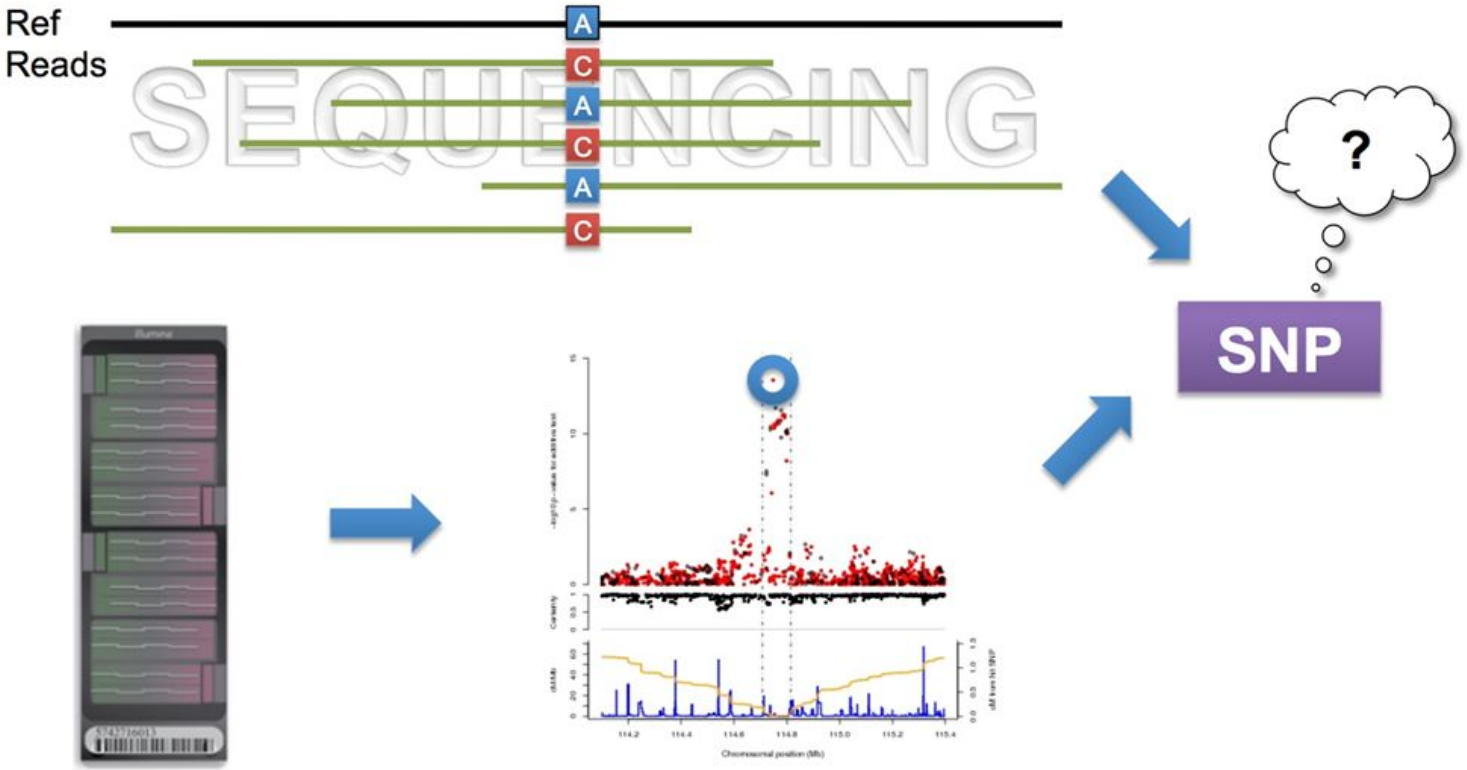
Búsqueda de polimorfismos en *Saccharomyces cerevisiae*

Tenemos datos de polimorfismos de tipo SNP (polimorfismo de nucleótido simple)

Queremos saber si afectan genes y sus consecuencias



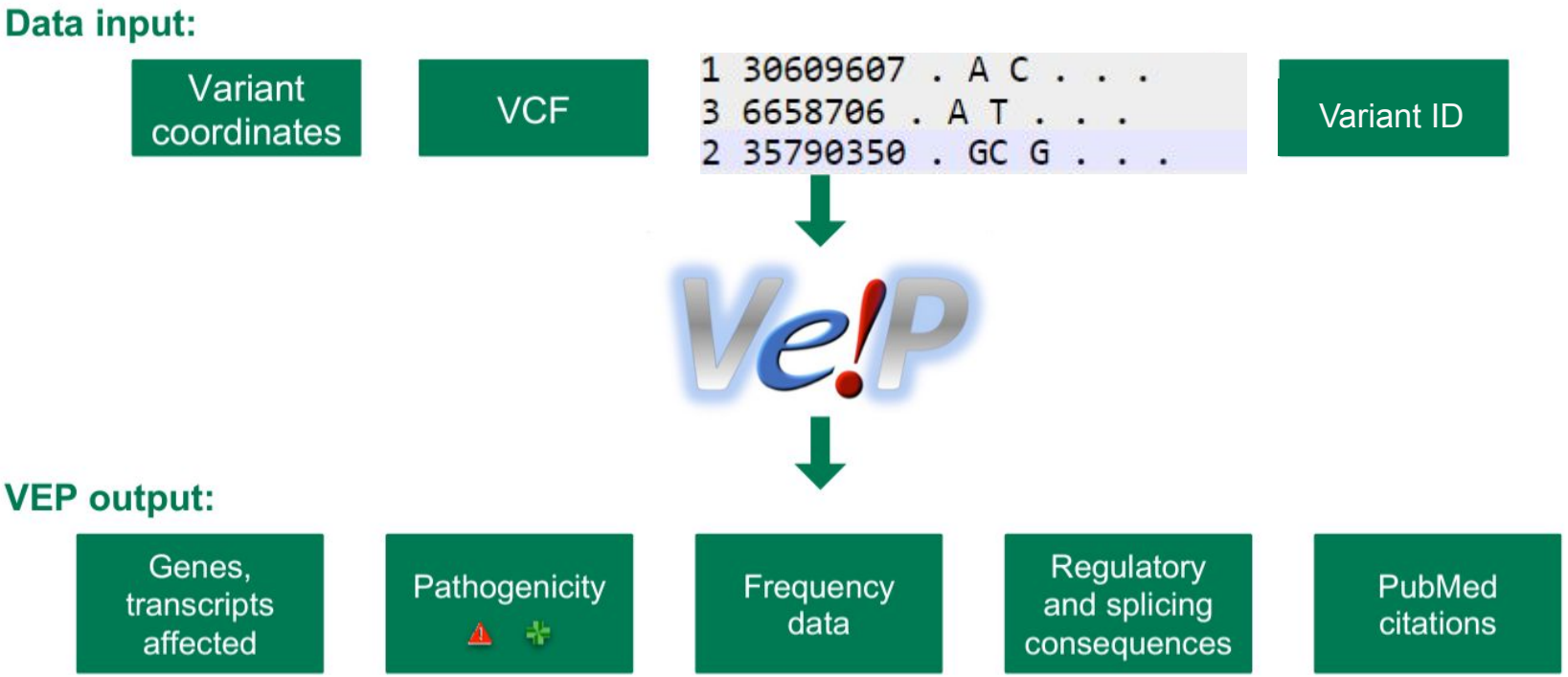
Gráficamente…



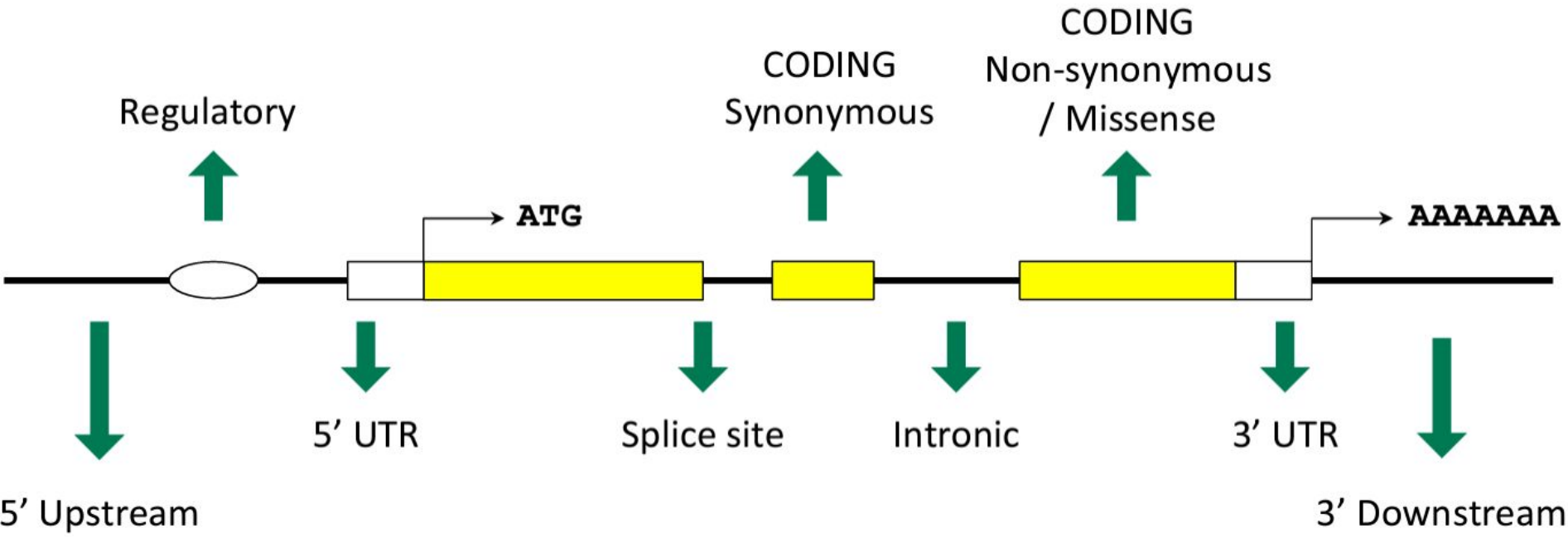
The Variant Effect Predictor

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

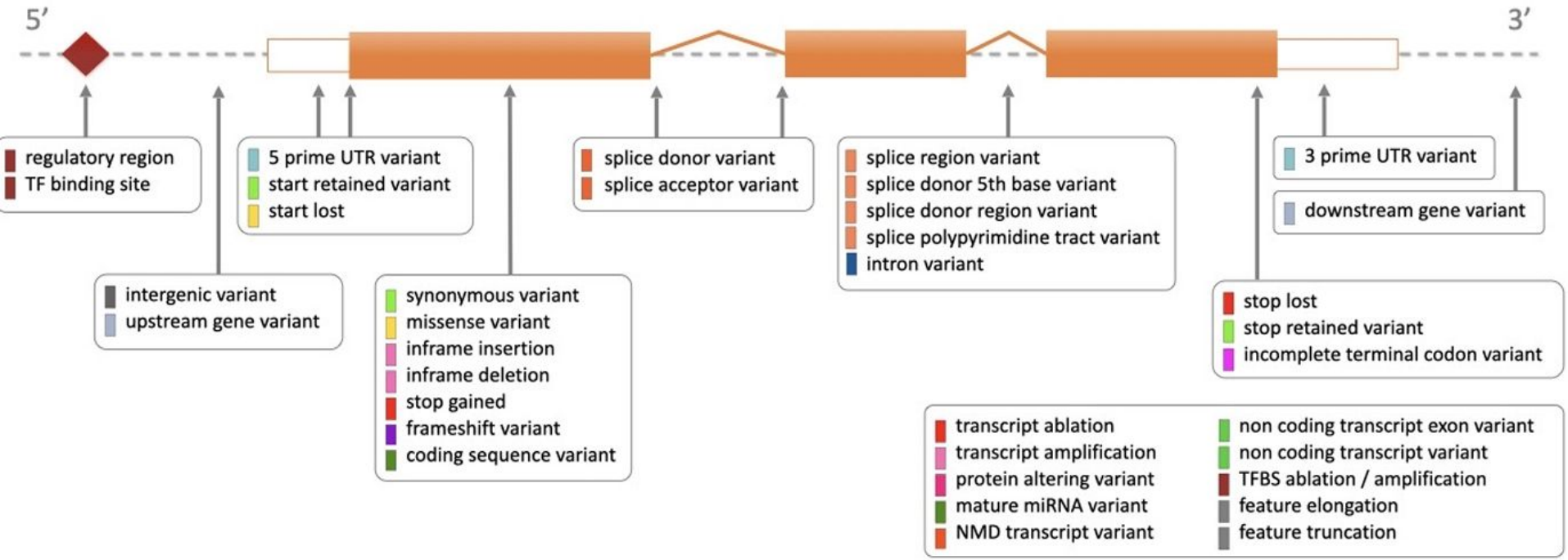
(Similar a SNPeff, 10.4161/fly.19695)



Consecuencias de las variantes



Sequence Ontology (SO)



Actividad 3

Vamos a trabajar con datos de *S. cerevisiae* → vayan al portal del Ensembl Genomes correspondiente.

- Encuentren el sitio para *S. cerevisiae*
- Registren el nombre del *assembly* (genoma de referencia), su tamaño en pb, número de cromosomas y cantidad de genes anotados (pista: *More information and statistics*)

Actividad 4

1. Usando VEP, predecir el efecto de las variantes disponibles en el VCF (descargar de GitHub)
2. Cuantas variantes tiene nuestro archivo?
3. Cuales son las consecuencias más frecuentes?
4. Cuántas variantes generan un “missense_variant”? (pista: usar filtros, consequences)
5. Cuantas variantes son de impacto alto (HIGH)? Cuáles son sus consecuencias? (Ojo que los filtros son sensibles a mayúsculas/ minúsculas)
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)


I'M GONNA
NEED MORE
COFFEE

Actividad 5 (integrando todo lo visto)

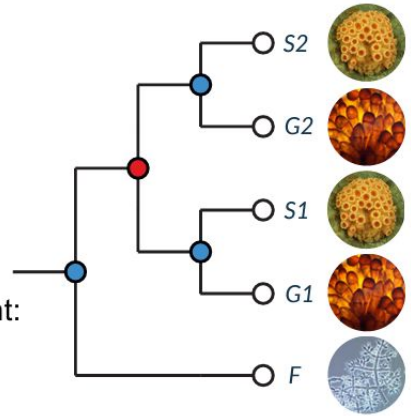
No encontramos variantes que afectan el gen de la PDC1… pero quizá a alguno de sus parálogos? (Si es que tiene)

Orthologues:

- Genes emerged through a **speciation** event:
  - e.g. G1 and S1; G2 and F; S2 and F
- 1-to-1: G1 and S1
- 1-to-many: F and S1, S2, G1, G2

Paralogues:

- Genes emerged through a **duplication** event:
  - e.g. G1 and G2, S1 and S2
- Within species: G1 and G2
- Between species: G1 and S2



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

Actividad 6

Exploración libre de Ensembl Browser (ejemplo, para el parálogo que identificamos recién)