

Endeavour: A Web Source for Gene Prioritization in Multiple Species

Sandra Alandes Esteve
Unidad de Bioinformática y Bioestadística

17 octubre 2019



Unidad de
Bioinformática y
Bioestadística



PRINCIPE FELIPE
CENTRO DE INVESTIGACION



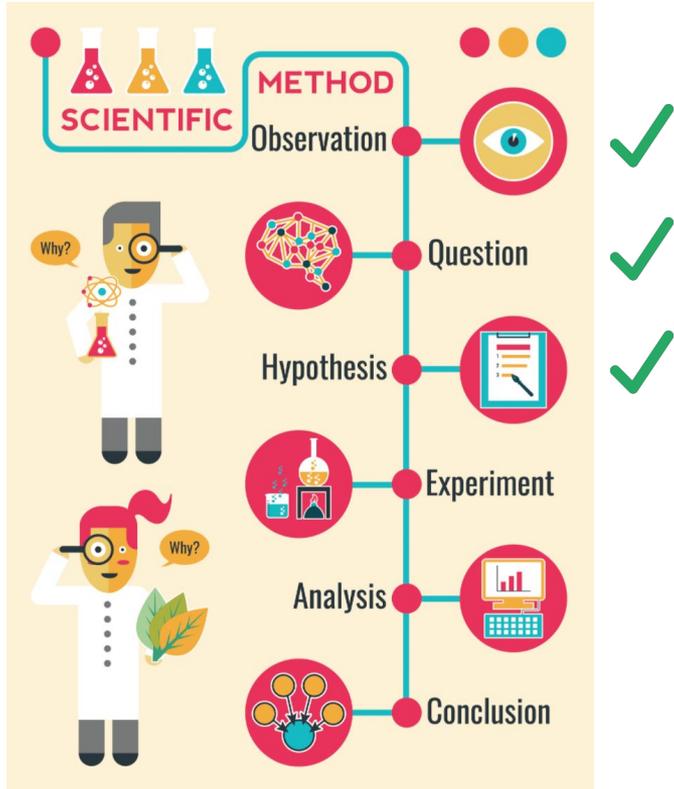
WODA

WEB-BASED OMICS DATA ANALYSIS

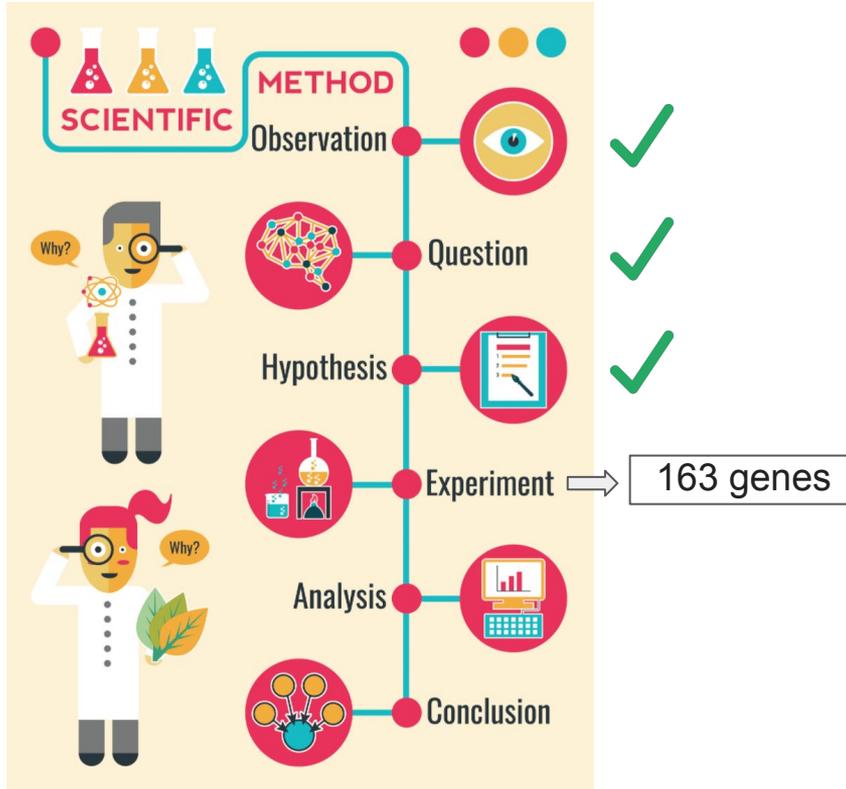
Índice

1. Introducción
2. Algoritmos
3. Selección de la información
4. Interpretación de resultados
5. Consejos
6. Buscador de genes
7. Actividades

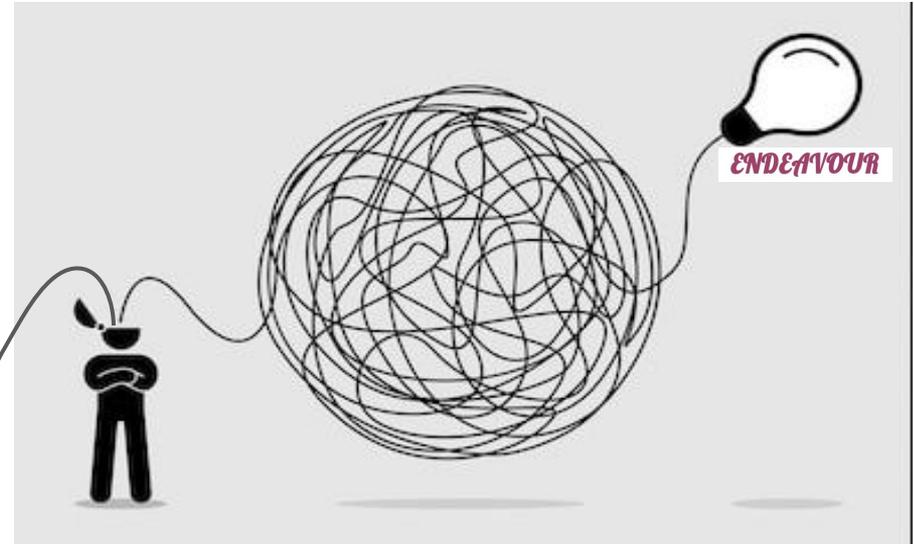
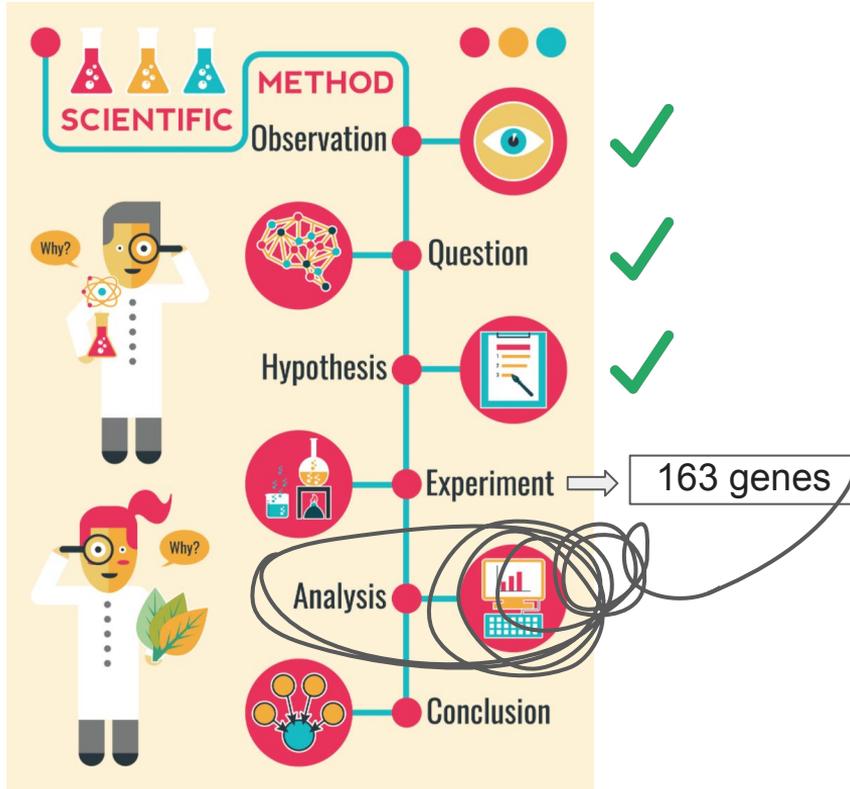
1. Introducción



1. Introducción



1. Introducción



PRIORIZACIÓN

ENDEAVOUR

1. Introducción

1. Introducción

Larga lista de genes
candidatos



1. Introducción

Larga lista de genes
candidatos



Falta de dinero



1. Introducción

Larga lista de genes
candidatos



Falta de dinero



Falta de tiempo



1. Introducción

Larga lista de genes
candidatos



Falta de dinero



Falta de tiempo



Hay que empezar



ENDEAVOUR

1. Introducción

Larga lista de genes
candidatos



Falta de dinero



Falta de tiempo



Hay que empezar



PRIORIZACIÓN

ENDEAVOUR

ENDEAVOUR: A web resource for gene prioritization in multiple species

The identification of key genes involved in health and disease remains a formidable challenge. Experimental approaches often produce lists of candidate genes, among which disease causing genes are hidden (i.e., their disease associations are still unknown). These lists of candidate genes can be rather large, and thus experimentally validating each candidate gene would be too expensive and time-consuming. There is therefore the need to predict the most promising candidate genes as to be able to maximise the yield of the experimental validation, which has been defined as 'gene prioritization'.



We have developed a bioinformatics approach to prioritize candidate genes underlying biological processes or diseases, and implemented it into a software application termed 'Endeavour'. Our strategy is based on how similar a candidate gene is to a profile derived from genes already known to be involved in the process of interest. Our approach relies on the integration of multiple heterogeneous sources (e.g., coding sequence, gene expression, functional annotation, literature, regulatory information) that cover what we currently know about these genes.

More precisely, Endeavour consists of three stages: training, scoring and fusion. In the first stage, information about the training genes (genes already known to play a role in the process under study) are retrieved from the genomic data sources in order to build models (one per data source). In the second stage, the models are then used to score the candidate genes and to rank them according to their scores. In the last stage, the rankings (one per data source) are fused into a global ranking using Order Statistics. Endeavour is currently available for human, mouse, rat, fruit fly, zebra fish and worm.

We have successfully used Endeavour to prioritize a DiGeorge syndrome associated region, a congenital heart defects associated region, and to optimize a genetic screen in *Drosophila melanogaster*. Researchers have also used Endeavour to look for genes involved in cleft lip / cleft palate from aCGH data and to analyze the proteome of adipocytes. Please browse our reference section to find a list of Endeavour related publications.

Our Approach



DATA

Data from multiple heterogeneous sources are collected and integrated in our local database, for better performances.

[Read more](#)



ALGORITHMS

Our algorithm uses basic machine learning techniques to model the biological process under study and then to prioritize the candidate genes.

[Read more](#)



ALTERNATIVES

There exists an alternative in case the already known disease genes can not easily be identified.

[Read more](#)

1. Introducción

- Herramienta de priorización de genes

1. Introducción

- Herramienta de priorización de genes

Lista de genes

TOMM40

IDE

APP

APOE

CLU

UBB

KLC1

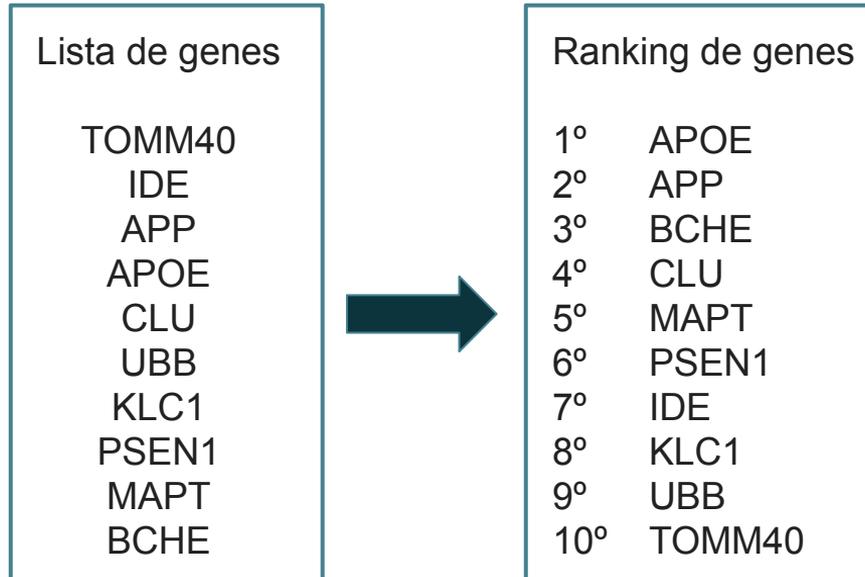
PSEN1

MAPT

BCHE

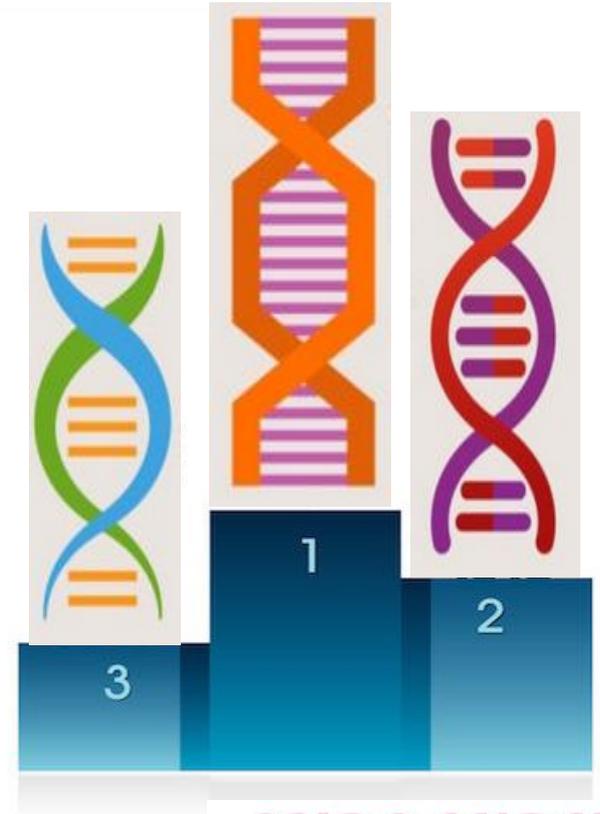
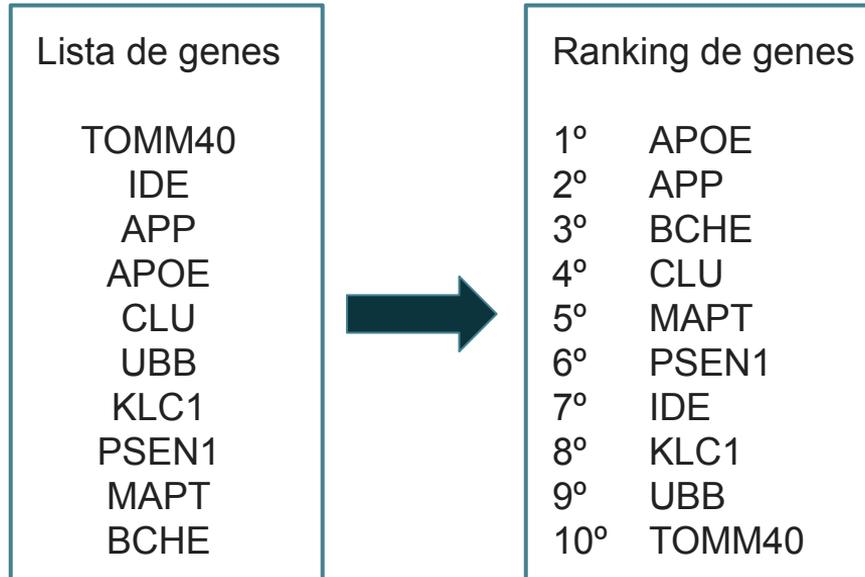
1. Introducción

- Herramienta de priorización de genes

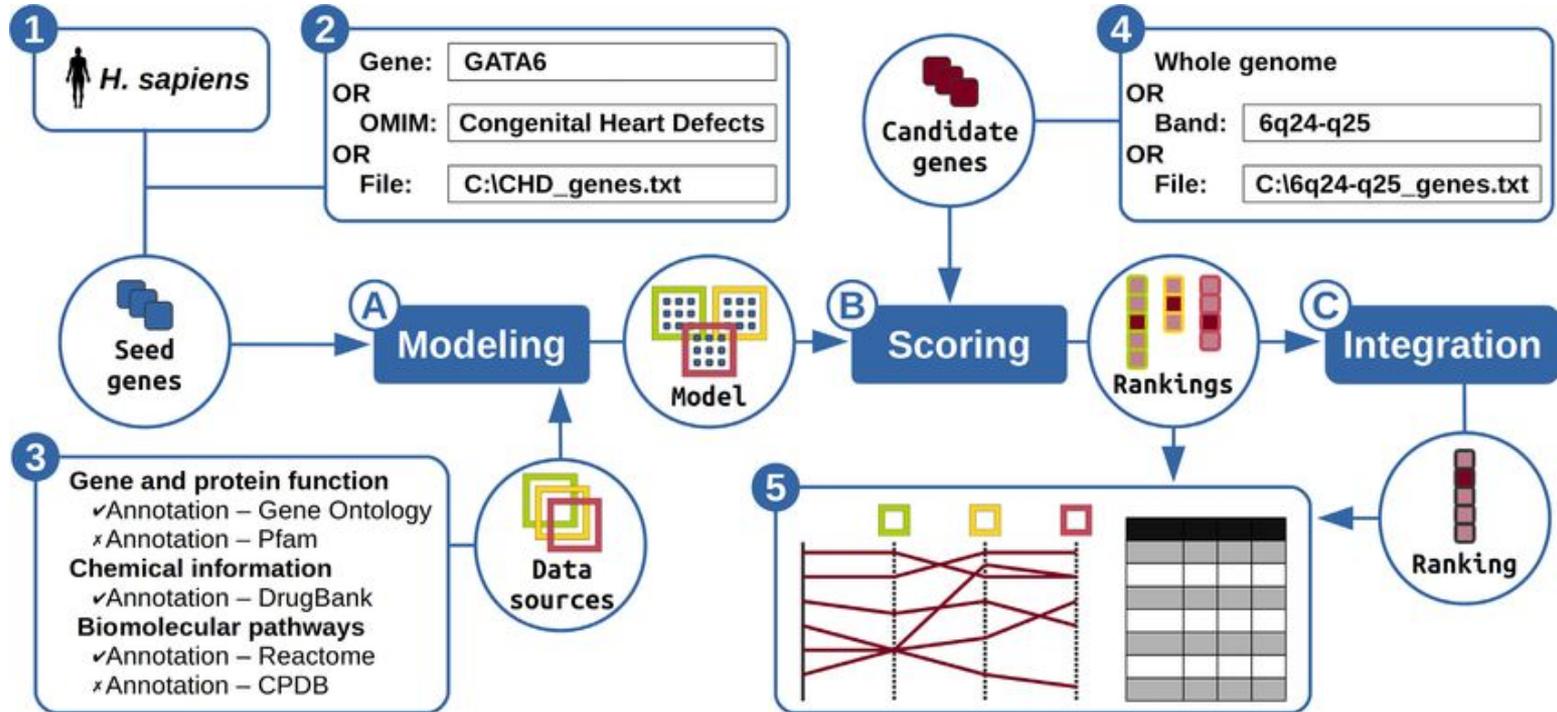


1. Introducción

- Herramienta de priorización de genes



2. Algoritmos



3. Selección de la información

3. Selección de la información

1 Species



Homo sapiens



Mus musculus



Caenorhabditis elegans



Rattus norvegicus



Drosophila melanogaster



Danio rerio

3. Selección de la información

1 Species

2 Training genes



Homo sapiens



Mus musculus



Caenorhabditis elegans



Rattus norvegicus



Drosophila melanogaster



Danio rerio

KCNJ2

Dock5

Nfasc

pan2

clk1

HIST2H2BE

KCNMB4

LRP2

Hist1h2ac

MYOT

FLCN

C10orf128

TF

PEX5L

RELN

AZGP1

Nptx2

Arrdc2

abca8

Tob2

3. Selección de la información

1 Species

2 Training genes

3 Data sources used to build models



Homo sapiens



Mus musculus



Caenorhabditis elegans



Rattus norvegicus



Drosophila melanogaster



Danio rerio

KCNJ2

Dock5

Nfasc

pan2

clk1

HIST2H2BE

KCNMB4

LRP2

Hist1h2ac

MYOT

FLCN

C10orf128

TF

PEX5L

RELN

AZGP1

Nptx2

Arrdc2

abca8

Tob2

Gene and protein function

Annotation – Gene Ontology

Annotation - Pfam

Chemical information

Annotation - DrugBank

Bio-molecular pathways

Annotation - Reactome

Annotation - CPDB

Phenotypic information

Annotation - GAD

Interaction networks

Interaction – String

Interaction – iRefIndex

Interaction – GeneRIF

Expression profiles

Expression – Su et al (2002)

Expression ontologies

Annotation – PaGenBase

Sequence based features

Blast

Annotation – Aura

Annotation – UniProt

Annotation – SIMAP (localization)

Annotation - Stitch

Annotation - WikiPathways

Annotation - hiPathDB

Annotation - OMIM

Interaction – BioGrid

Interaction – Mint

Expression – Su et al (2004)

Annotation – CGAP

Precalculated – Ouzounis

Annotation – mirZ

Text-mining

Annotation - RGD ChEBI

Annotation – RGD pathways

Annotation – RGD MP

Interaction – I2D

Interaction – HPRD

Expression – CMAP

Annotation - GNF

Precalculated – Prospectr

Annotation – InterPro

Annotation - BioCarta

Annotation – RGD RDO

Interaction – IntAct

Interaction – MIPS

Expression – Lukk et al

Annotation - eGenetics

Precalculated – HaploPred

3. Selección de la información

1 Species

2 Training genes

3 Data sources used to build models

4 Candidates



Homo sapiens



Mus musculus



Caenorhabditis elegans



Rattus norvegicus



Drosophila melanogaster



Danio rerio

KCNJ2

Dock5

Nfasc

pan2

clk1

HIST2H2BE

KCNMB4

LRP2

Hist1h2ac

MYOT

FLCN

C10orf128

TF

PEX5L

RELN

AZGP1

Nptx2

Arrdc2

abca8

Tob2

Gene and protein function

Annotation – Gene Ontology

Annotation - Pfam

Chemical information

Annotation - DrugBank

Bio-molecular pathways

Annotation - Reactome

Annotation - CPDB

Phenotypic information

Annotation - GAD

Interaction networks

Interaction – String

Interaction – iRefIndex

Interaction – GeneRIF

Expression profiles

Expression – Su et al (2002)

Expression ontologies

Annotation – PaGenBase

Sequence based features

Blast

Annotation – Aura

Annotation – UniProt

Annotation – SIMAP (localization)

Annotation - Stitch

Annotation - WikiPathways

Annotation - hiPathDB

Annotation - OMIM

Interaction – BioGrid

Interaction – Mint

Expression – Su et al (2004)

Annotation – CGAP

Precalculated – Ouzounis

Annotation – mirZ

Text-mining

Annotation - RGD ChEBI

Annotation – RGD pathways

Annotation – RGD MP

Interaction – I2D

Interaction – HPRD

Expression – CMAP

Annotation - GNIF

Precalculated – Prospectr

Annotation – InterPro

Annotation - BioCarta

Annotation – RGD RDO

Interaction – IntAct

Interaction – MIPS

Expression – Lukk et al

Annotation - eGenetics

Precalculated – HaploPred

BEND4

Celf4

ank1

PHLDB2

BTBD11

Kih13

ank1

-

mtch2

znf574

AGK

SLC18A2

-

ddc

SLC10A4

Cnpy2

Dok6

AT4G28910

macrod2

ITPR1

4. Interpretación de resultados

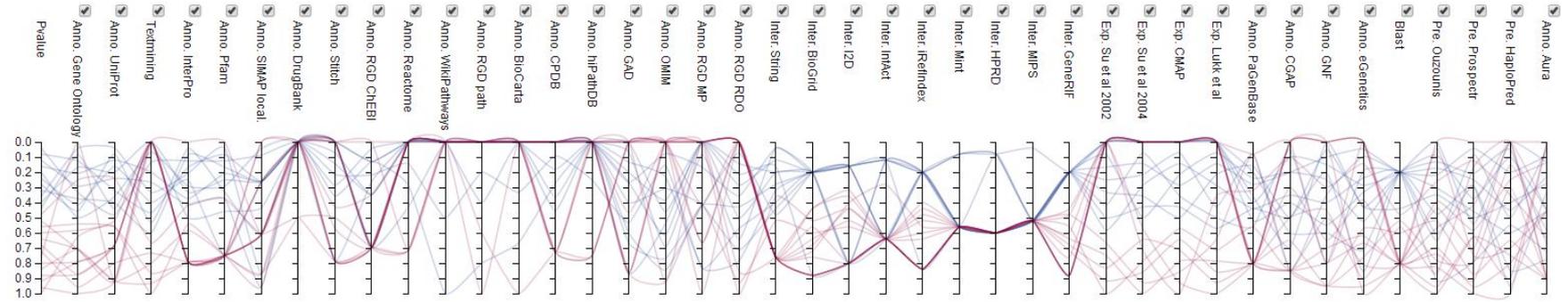
4. Interpretación de resultados

■ Training gene
 ■ Candidate gene
 ■ Training and candidate gene

Group breakdown



Total selected



Gene name	Group	P-value	Annotati...	Annotati...	Text-min...	Annotati...								
CYP7B1	Training gene	0.24	0.4167	0.16	0.4	0.25	0.75	0.2609	0	0.2143	0.2174	0.1429	0.5	0.2
HSPD1	Training gene	0.28	0.4583	0.08	0.5333	0.0417	0.3333	0.6087	0	0.3571	0.1304	0.2857	1	0.8
PNPLA6	Training gene	0.32	0.375	0.4	0.6	0.0833	0.1667	0.2609	0	0.0714	0.3478	0	0	0.4
KIAA0196	Training gene	0.36	0.0417	0.04	0.3333	0.2083	0.0417	0.8696	0	0	0.6957	0	0	0
ZFYVE27	Training gene	0.4	0.1667	0.2	0	0.5	0.4583	0.4348	0	0	0.6957	0	0	0
NIPA1	Training gene	0.44	0.0833	0.44	0	0.3333	0.25	0.2609	0	0	0.1304	0	0	0
KIF1A	Candidate gene	0.48	0.25	0.24	0.6667	0.4167	0.375	0.6087	0.5	0.7857	0.6957	0	0	0
SNED1	Candidate gene	0.52	0.7083	0.92	0	0.7917	0.75	0.6087	0	0.7857				
PASK	Candidate gene	0.56	0.5417	0.56	0.8667	0.7917	0.75	0	0	0.4286				

Showing all 25 rows



5. Consejos

Genes semilla:

- Un conjunto mayor de genes semilla proporciona resultados más fiables. Se recomienda un mínimo de 5 genes y máximo 40 genes.

Fuentes de datos:

- La lista de fuentes de datos disponible depende de la especie seleccionada
- Seleccionar los recursos que mejor se adapten a su problema específico.
- Evitar usar conjuntamente recursos redundantes.
- No hay límite en el número de fuentes de datos, salvo el tiempo de ejecución.

Genes candidatos:

- No hay límite en el número de genes candidatos.

6. Buscador de genes

Beegle

e.g. Alzheimer's disease

SEARCH KNOWN GENES



ENDEAVOUR

Sandra Alandes Esteve

salandes@cipf.es

ENDEAVOUR