Functional Analysis

Center for Health Data Science

Health Data Science Sandbox

Overview

1 Rstudio & Rmarkdown

2 Count Matrix & Normalization

3 Exploratory Analysis

4 Differential Expression

5 Functional Analysis

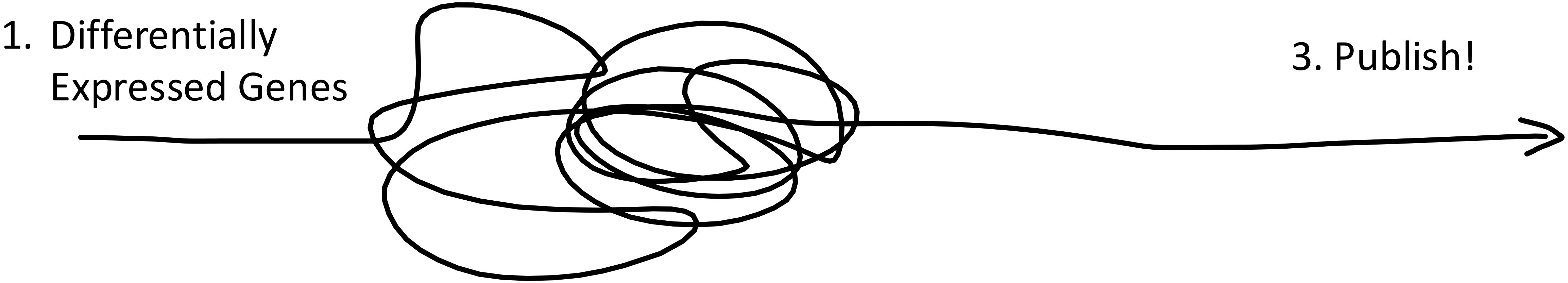
Functional analysis

- Differential Expression analysis
  - Resulted in **up**- and **down-**regulated genes in each comparison

- Define genes of interest:
  - Treatment vs. Control with Log2FC > 1 and adjusted p-value <0.05

- What do these genes do?
  - Do they share a **common function**? e.g. Immuno-processes
  - Part of the **same pathway**? e.g. Nucleotide Metabolism

Functional analysis



1. Differentially Expressed Genes

3. Publish!

2. Anything and everything biologically meaningful & interesting:

- Co-expression and interaction
- Gene set enrichment analysis (pathway, GO terms)
- Disease and drug databases
- …

Annotation

DESeq2 results:

Gene_ID (Ensembl)	LFC	padj	…
ENSG00000223972	1.101	0.001	…
ENSG00000278267	-4.567	0.045	…

Annotation:

EBI (Ensembl) HGNC NCBI UCSC

Gene_ID (Ensembl)	Gene_name	Entrez_ID	RefSeq_ID	Chr	Start	End	Feature
ENSG00000223972	DDX11L1	NA	NR_046018	1	11869	14409	Protein
ENSG00000278267	MIR6859-1	102466751	NR_106918	1	17369	17436	nc-RNA

+ MANE = Matched Annotation between NCBI and EBI

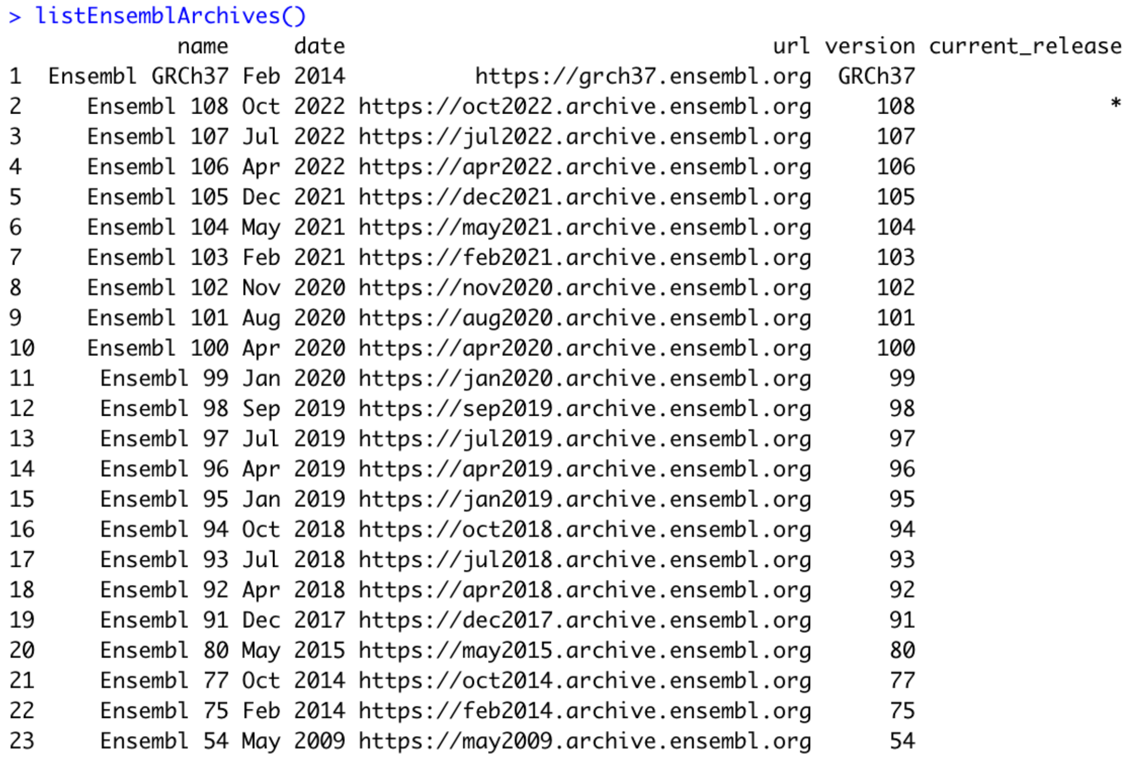
Annotation

Genome assembly updated **every few years**
Annotations updated **every few months**

⚠ Genome assembly versions **MUST** match!

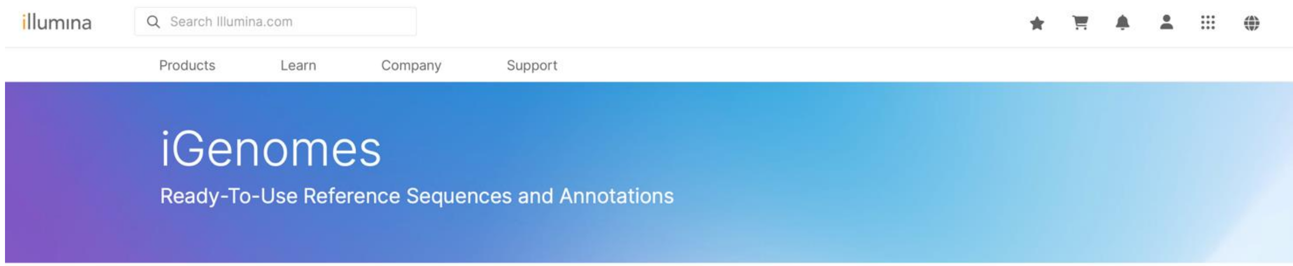
- If you **map to GRCh37**, use **annotations for GRCh37**
- Preferably also use the **same release** for all annotations

Ensemble database annotation releases for GRCh37 build:



```
> listEnsemblArchives()
          name     date                                 url version current_release
1  Ensembl GRCh37 Feb 2014          https://grch37.ensembl.org  GRCh37
2     Ensembl 108 Oct 2022 https://oct2022.archive.ensembl.org     108               *
3     Ensembl 107 Jul 2022 https://jul2022.archive.ensembl.org     107
4     Ensembl 106 Apr 2022 https://apr2022.archive.ensembl.org     106
5     Ensembl 105 Dec 2021 https://dec2021.archive.ensembl.org     105
6     Ensembl 104 May 2021 https://may2021.archive.ensembl.org     104
7     Ensembl 103 Feb 2021 https://feb2021.archive.ensembl.org     103
8     Ensembl 102 Nov 2020 https://nov2020.archive.ensembl.org     102
9     Ensembl 101 Aug 2020 https://aug2020.archive.ensembl.org     101
10    Ensembl 100 Apr 2020 https://apr2020.archive.ensembl.org     100
11     Ensembl 99 Jan 2020 https://jan2020.archive.ensembl.org      99
12     Ensembl 98 Sep 2019 https://sep2019.archive.ensembl.org      98
13     Ensembl 97 Jul 2019 https://jul2019.archive.ensembl.org      97
14     Ensembl 96 Apr 2019 https://apr2019.archive.ensembl.org      96
15     Ensembl 95 Jan 2019 https://jan2019.archive.ensembl.org      95
16     Ensembl 94 Oct 2018 https://oct2018.archive.ensembl.org      94
17     Ensembl 93 Jul 2018 https://jul2018.archive.ensembl.org      93
18     Ensembl 92 Apr 2018 https://apr2018.archive.ensembl.org      92
19     Ensembl 91 Dec 2017 https://dec2017.archive.ensembl.org      91
20     Ensembl 80 May 2015 https://may2015.archive.ensembl.org      80
21     Ensembl 77 Oct 2014 https://oct2014.archive.ensembl.org      77
22     Ensembl 75 Feb 2014 https://feb2014.archive.ensembl.org      75
23     Ensembl 54 May 2009 https://may2009.archive.ensembl.org      54
```

Annotation

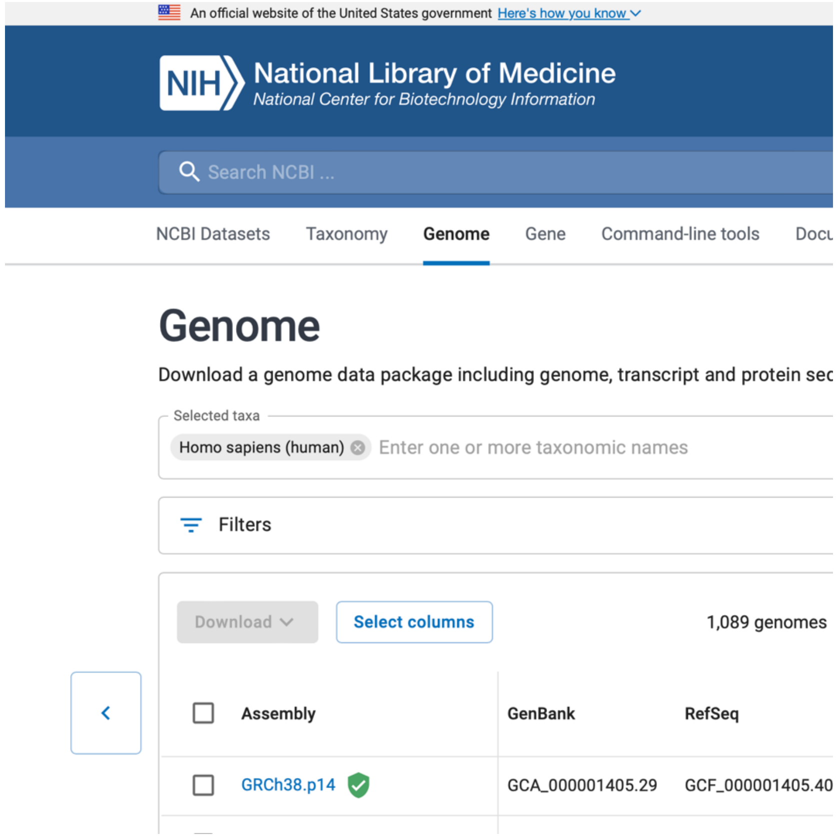


iGenomes

Ready-To-Use Reference Sequences and Annotations

👆 A resource for downloading matching references + annotations

A more updated resource 👉


An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Docu
Filters
Download
Select columns

Genome

Download a genome data package including genome, transcript and protein sec

Selected taxa: Homo sapiens (human) ⊗ Enter one or more taxonomic names

1,089 genomes

□ Assembly	GenBank	RefSeq
□ GRCh38.p14 ✓	GCA_000001405.29	GCF_000001405.40

Annotation

Reference genome = gold standard

Genome assembly GRCh38.p14 reference

- **Chromosome level** assembly
- Contains **chromosomes + scaffolds** with funny names
- Some regions **contain gaps** (telomeres, repetitive regions, centromeres)

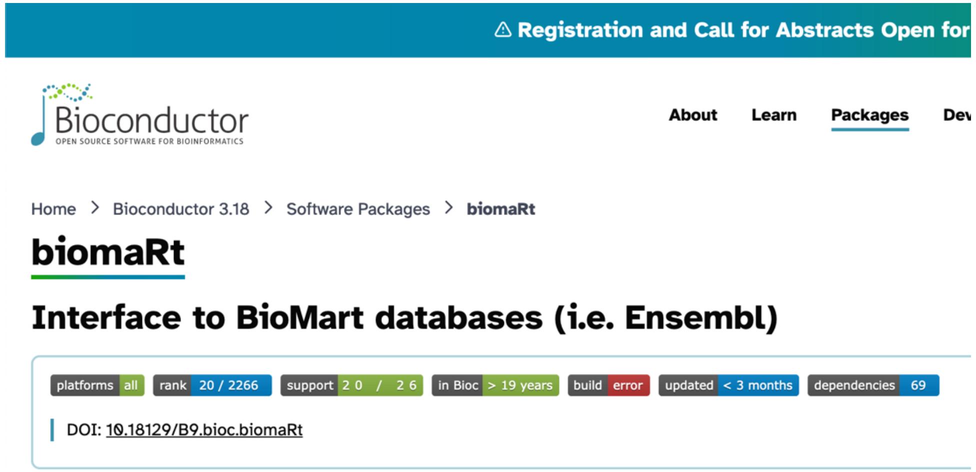
Newer version but less explored

Genome assembly T2T-CHM13v2.0

- **Whole genome** assembly
- Contains **only chromosomes**
- **Complete sequence** from telomere to telomere

Annotation in R

- R packages **biomaRt** or **AnnotationDbi** can be used to convert between identifiers
- You *may* not be able to get the exact release version of a build, but you can likely get one close to it:
  - This could mean some IDs cannot be converted, it is only a few we usually don't care



⚠ Registration and Call for Abstracts Open for

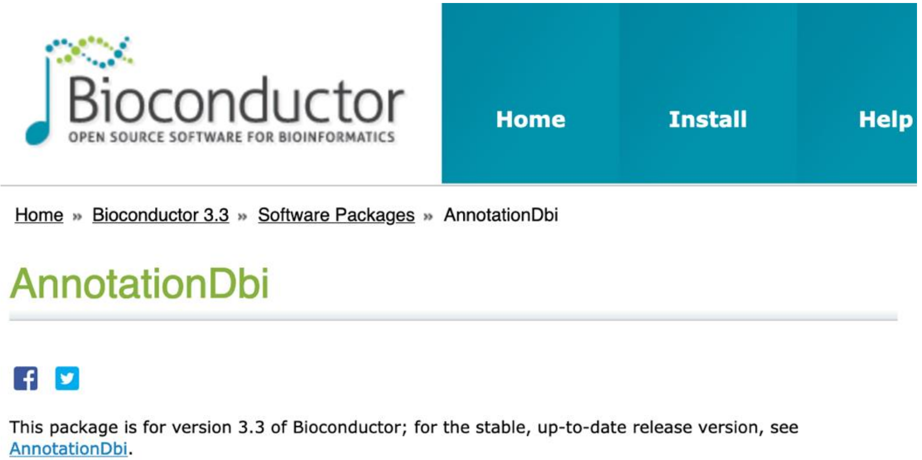
Home > Bioconductor 3.18 > Software Packages > **biomaRt**

biomaRt

Interface to BioMart databases (i.e. Ensembl)

platforms all | rank 20 / 2266 | support 2 0 / 2 6 | in Bioc > 19 years | build error | updated < 3 months | dependencies 69

DOI: 10.18129/B9.bioc.biomaRt



Home » Bioconductor 3.3 » Software Packages » AnnotationDbi

AnnotationDbi

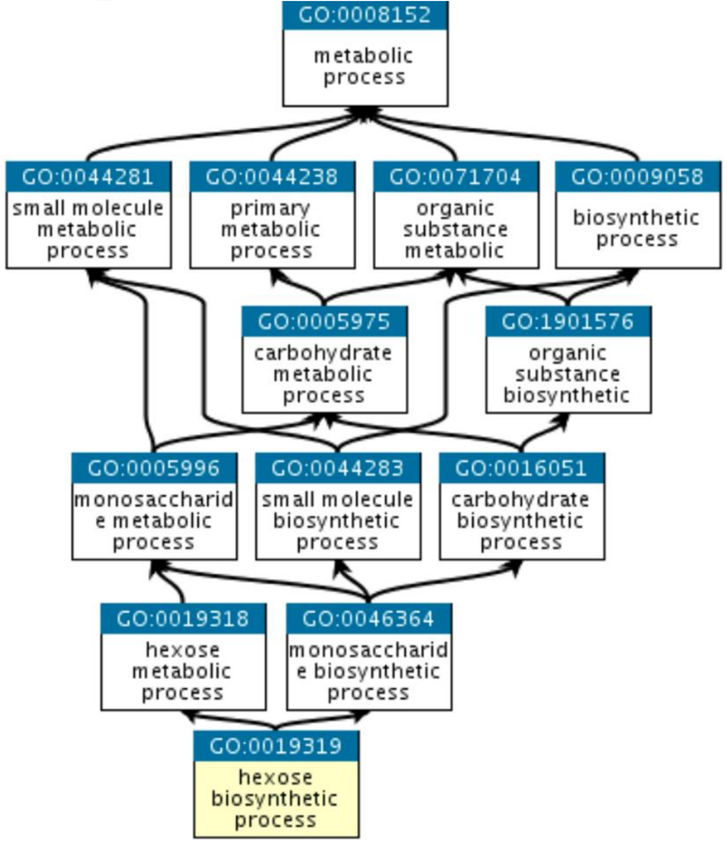
This package is for version 3.3 of Bioconductor; for the stable, up-to-date release version, see AnnotationDbi.

Functional analysis – GO terms

Gene Ontology (GO) Term

- Formal representation of a body of knowledge in the biological domain

- Genes are annotated to different types of knowledge
  - Biological processes: DNA repair, signal transduction, etc.
  - Molecular function: catalysis, transportation, etc.
  - Cellular component: ribosome, nucleus, etc.


GO:0008152 metabolic process
GO:0044281 small molecule metabolic process
GO:0044238 primary metabolic process
GO:0071704 organic substance metabolic
GO:0009058 biosynthetic process
GO:0005975 carbohydrate metabolic process
GO:1901576 organic substance biosynthetic
GO:0005996 monosaccharide metabolic process
GO:0044283 small molecule biosynthetic process
GO:0016051 carbohydrate biosynthetic process
GO:0019318 hexose metabolic process
GO:0046364 monosaccharide biosynthetic process
GO:0019319 hexose biosynthetic process

Functional analysis – GO terms

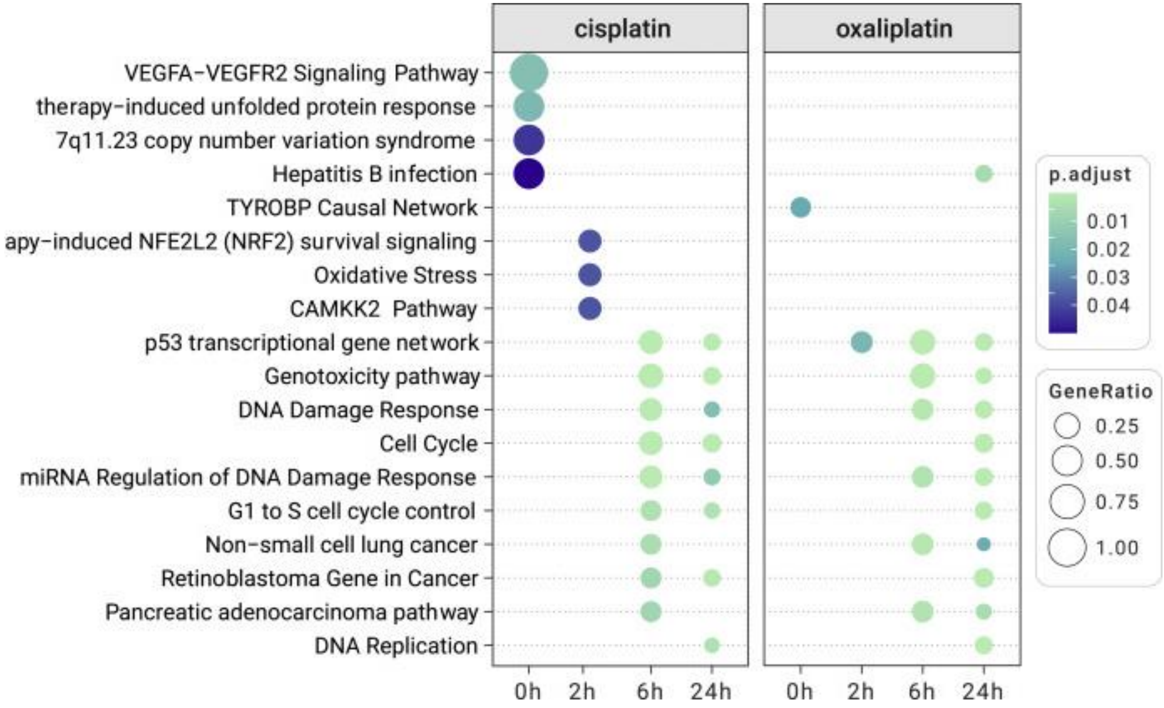
GO terms tend to be redundant, approaches to solve redundancy:

- **DOSE:** disease ontology semantic and enrichment analysis

  R package finds enriched disease pathways

- **GoSemSim:** semantic similarity among GO terms and gene products

  R package shows how similar are the Gene/Disease ontology terms



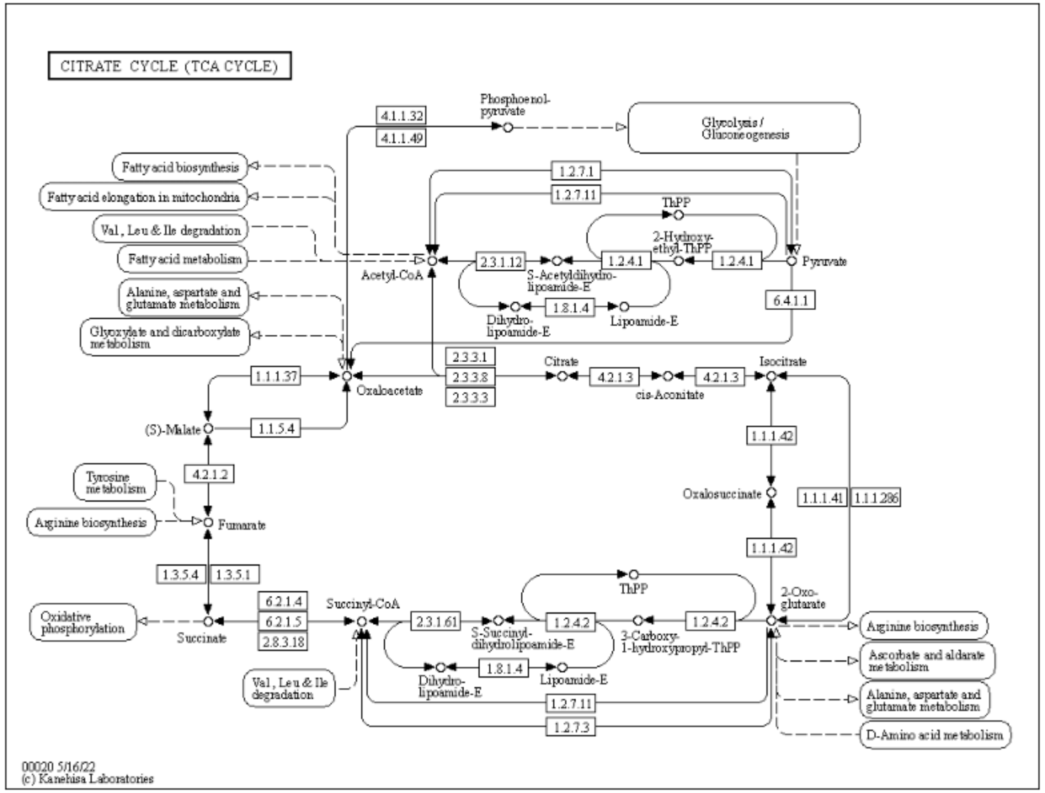
Functional analysis – Pathways

Pathways

Set of genes interacting with each other to perform a specific biological function

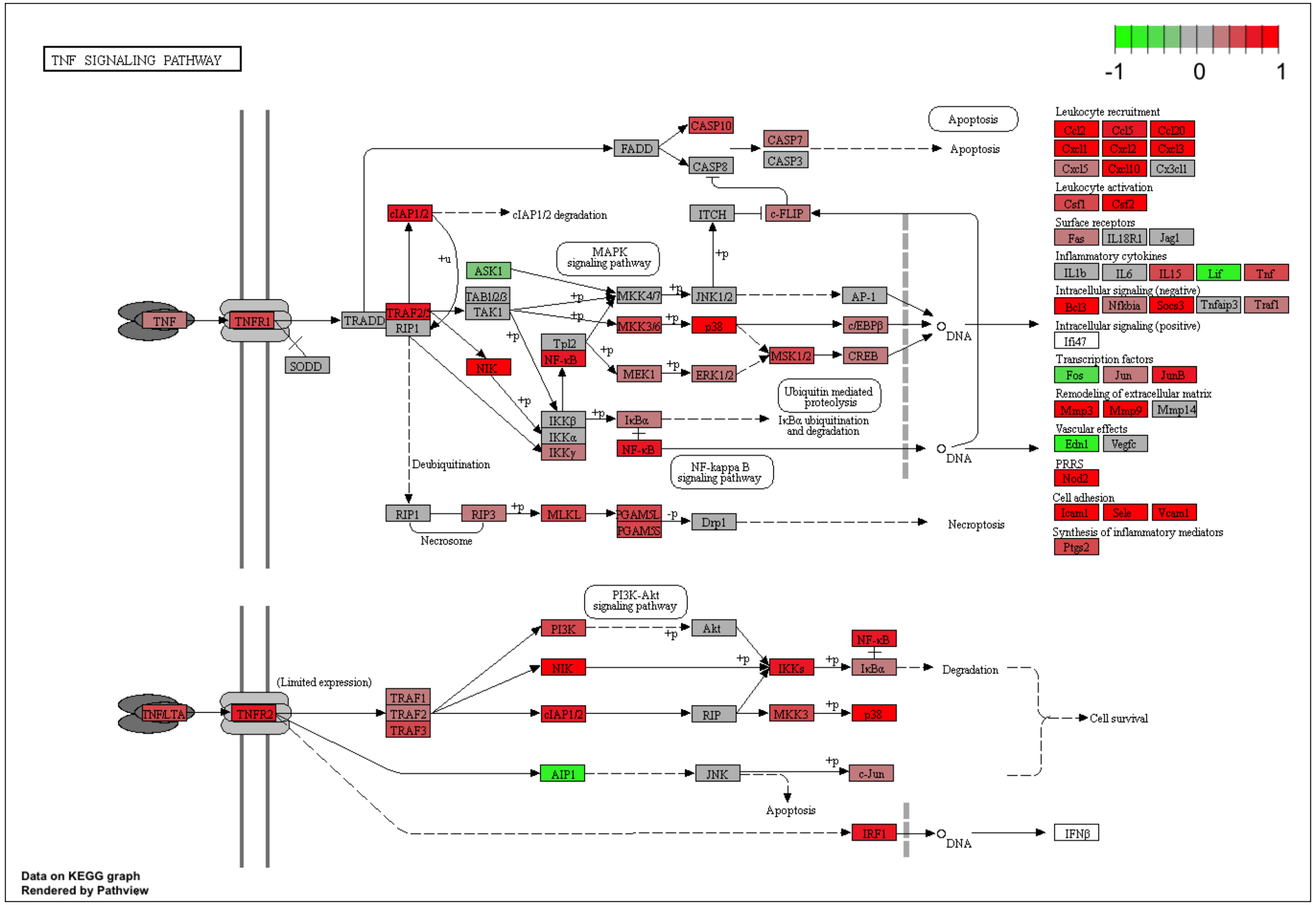
KEGG pathway database

- Metabolism
- Genetic information processing
- Environmental Information Processing
- Cellular processes
- Organismal Systems
- Human Diseases



Functional analysis – Pathways

Pathview:
R package that visualizes Differentially Expressed (DE) genes with their log2foldChanges (LFC) within a KEGG pathway.



Functional analysis – Pathways

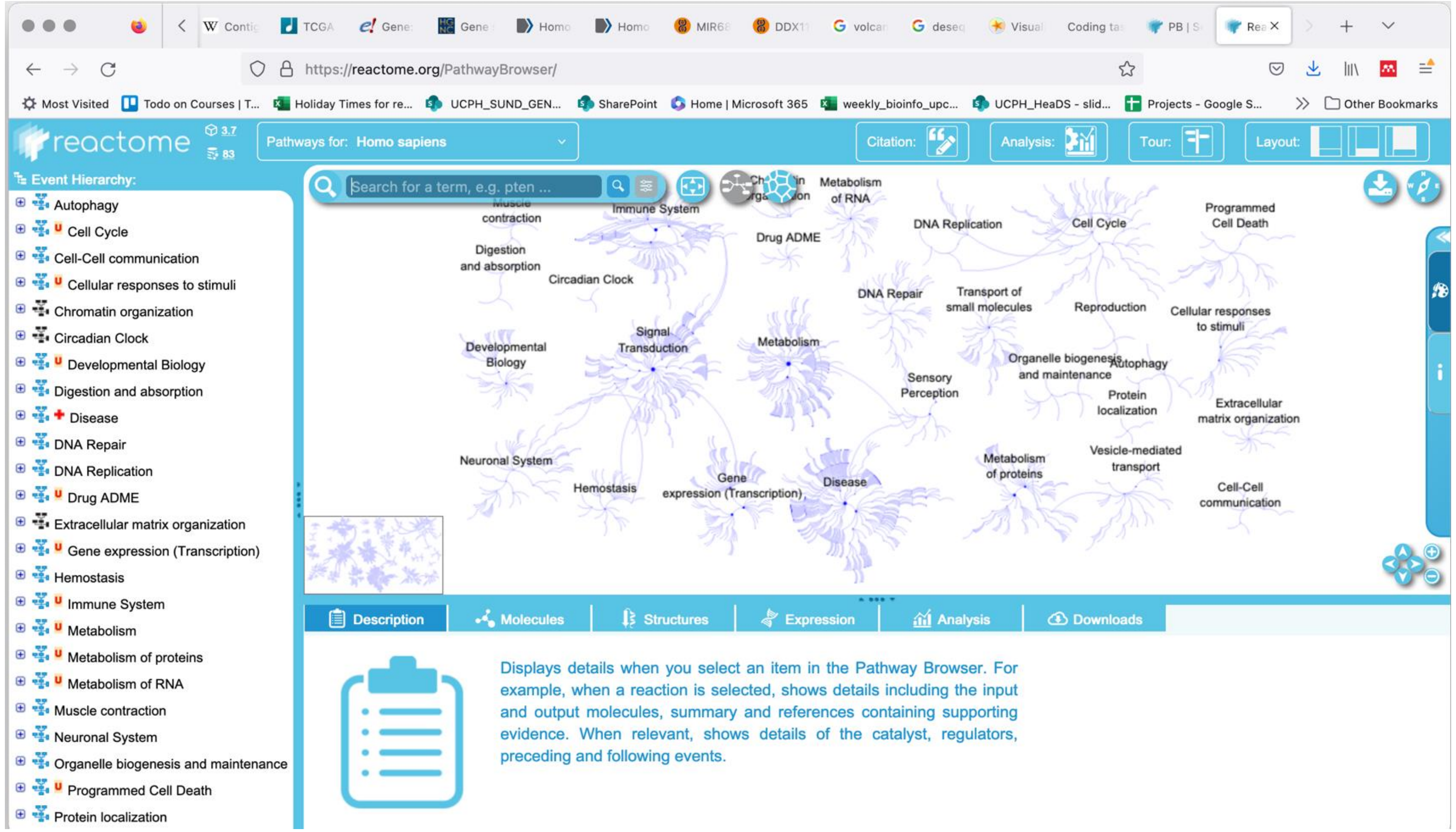
Reactome:

Pathway database:

- open-source
- open access
- manually curated
- peer-reviewed

R-package:

reactomePA



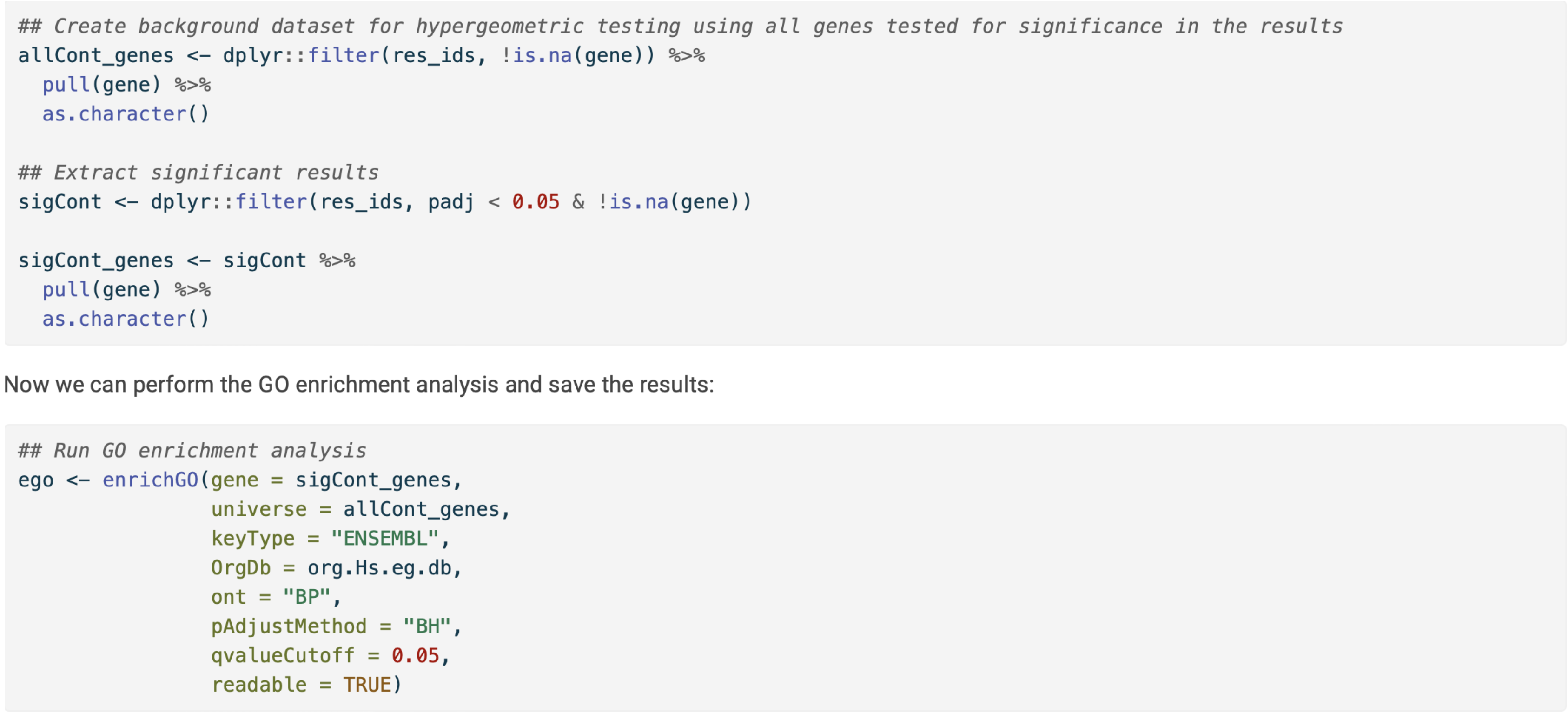
Displays details when you select an item in the Pathway Browser. For example, when a reaction is selected, shows details including the input and output molecules, summary and references containing supporting evidence. When relevant, shows details of the catalyst, regulators, preceding and following events.

Exercise

Let’s annotate some genes!

- Notebook:
  - *08a_FA_genomic_annotation.Rmd*

How to convert gene IDs



```r
## Create background dataset for hypergeometric testing using all genes tested for significance in the results
allCont_genes <- dplyr::filter(res_ids, !is.na(gene)) %>%
  pull(gene) %>%
  as.character()

## Extract significant results
sigCont <- dplyr::filter(res_ids, padj < 0.05 & !is.na(gene))

sigCont_genes <- sigCont %>%
  pull(gene) %>%
  as.character()
```

Now we can perform the GO enrichment analysis and save the results:

```r
## Run GO enrichment analysis
ego <- enrichGO(gene = sigCont_genes,
                universe = allCont_genes,
                keyType = "ENSEMBL",
                OrgDb = org.Hs.eg.db,
                ont = "BP",
                pAdjustMethod = "BH",
                qvalueCutoff = 0.05,
                readable = TRUE)
```