XVIII Congreso de Biometría CEBMADRID

25, 26 y 27 mayo 2022, Madrid

Integrative Analysis of Multi-Omics Data with Addition of Biological Knowledge

Ferran Briansó & Alex Sánchez-Pla
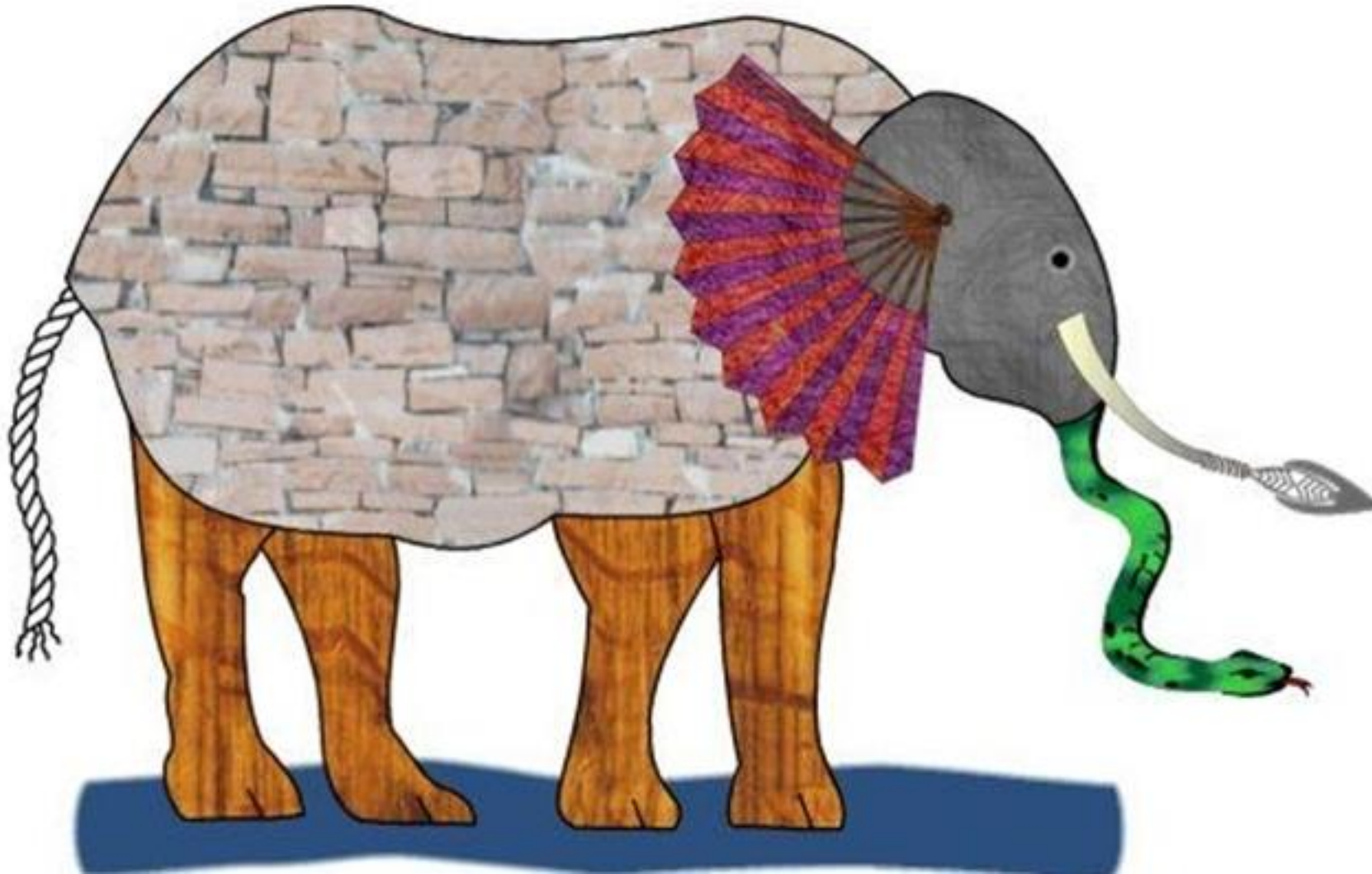**Dpt. Genetics, Microbiology and Statistics**

Vall d'Hebron
Institut de Recerca
UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

UNIVERSITAT DE
BARCELONA

GRBIO

Integrative analysis allows the combination of distinct omics data

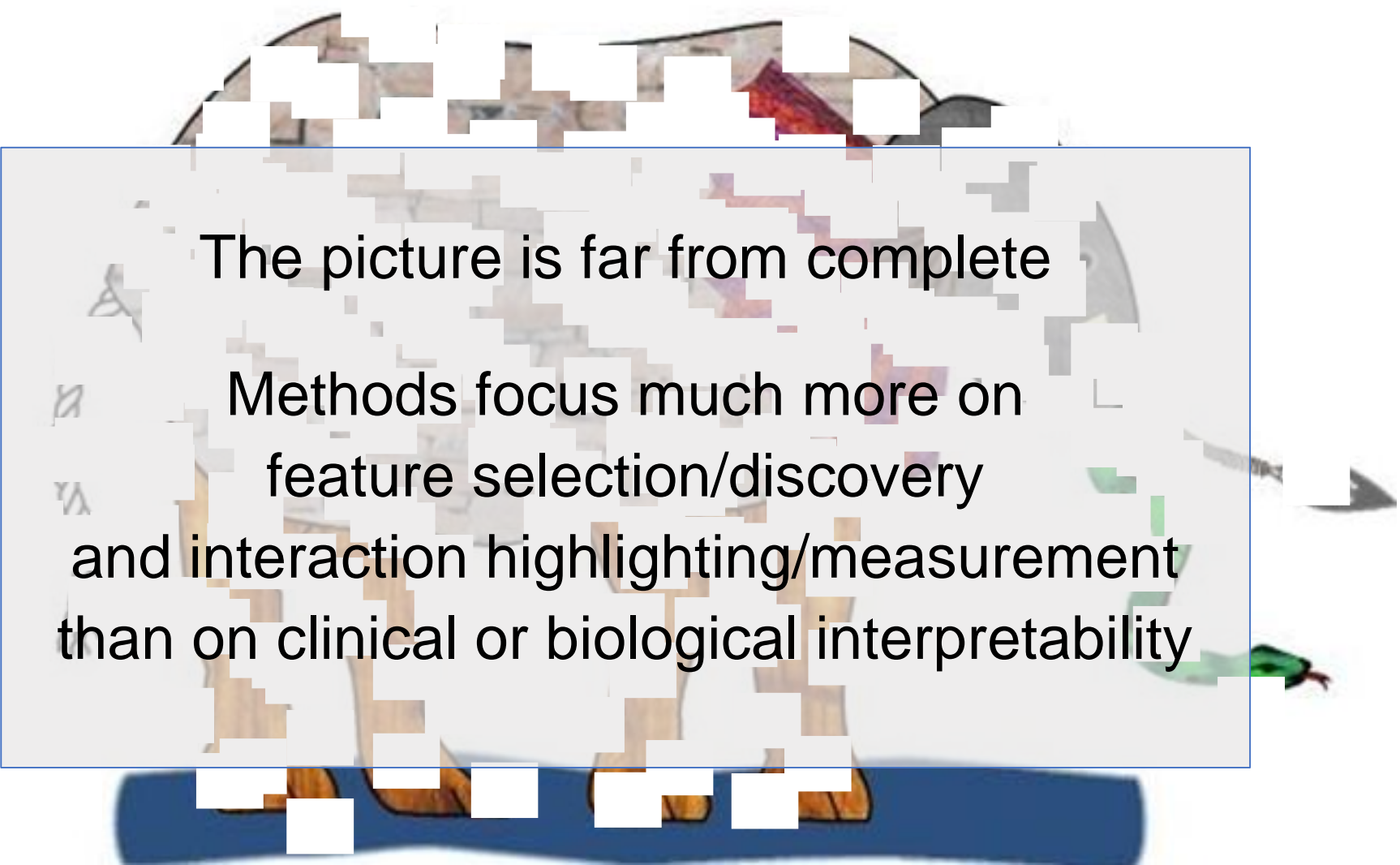# Integrative analysis allows the combination of distinct omics data

But interpretability is a weak point of most multi-omics approaches

But interpretability is a weak point of most multi-omics approaches

The picture is far from complete

Methods focus much more on
feature selection/discovery
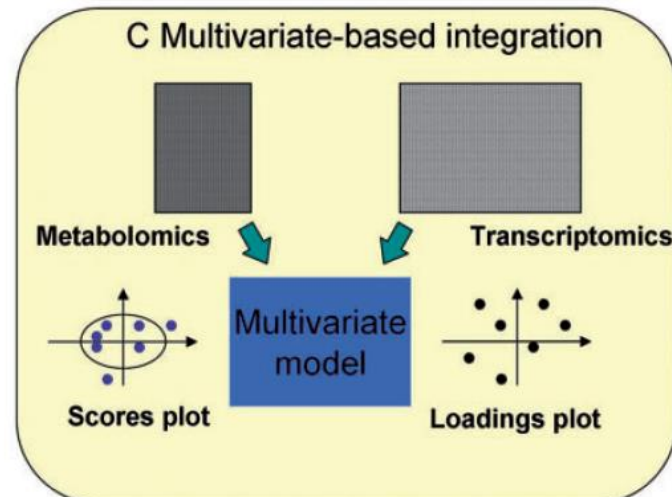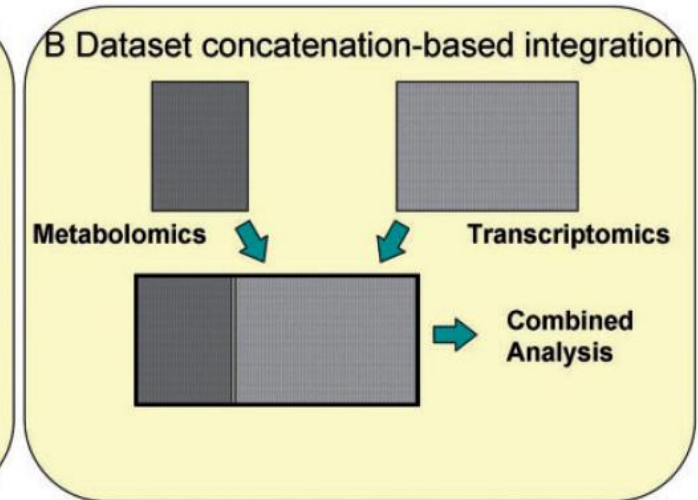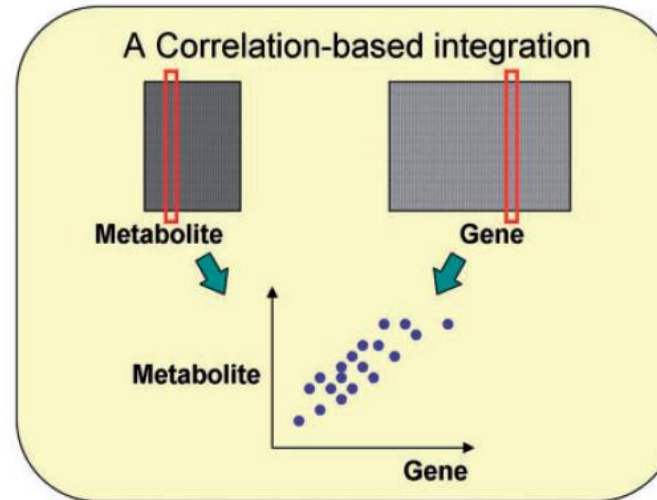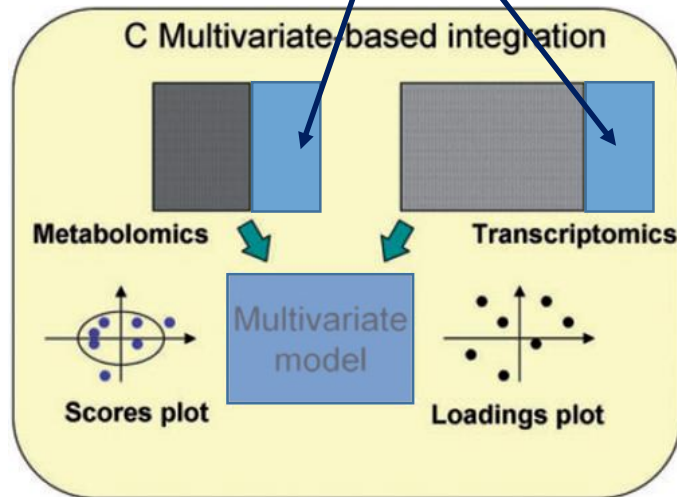and interaction highlighting/measurement
than on clinical or biological interpretability

https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

Some approaches for multi-omics data integration

Addition of Biological Annotations

(GO, GeneSets, pathways…)

before the integrative steps

A Correlation-based integration

Metabolite   Gene

Metabolite

Gene

B Dataset concatenation-based integration

Metabolomics   Transcriptomics

Combined Analysis

C Multivariate-based integration

Metabolomics   Transcriptomics

Multivariate model

Scores plot   Loadings plot

C Multivariate-based integration

Metabolomics   Transcriptomics

Multivariate model

Scores plot   Loadings plot

D Pathway-based integration

Metabolomics   Transcriptomics

Pathways

Significantly altered pathways

Cavill, 2016

Our proposal

1. **Pre-process omics datasets in order to include biological information**
<div align="right"><u>**before**</u> **the joint analysis**</div>

→ **Expanded datasets**

Applying standard biological significance analysis methods
  (e.g. Gene over-representation; Gene Sets; Pathways; Networks… + Custom annotations)

Our proposal

1. **Pre-process omics datasets in order to include biological information**
   <u>before</u> the joint analysis

   → **Expanded datasets**

   Applying standard biological significance analysis methods
   (e.g. Gene over-representation; Gene Sets; Pathways; Networks… + Custom annotations)

2. **Analysis of the expanded datasets by the use of contrasted**
   joint Dimension Reduction techniques

   → MCIA (*made4, omicade*…)
   → RGCCA (*miXomics*…)
   → **Multiple Factor Analysis (*FactoMineR*…)**

Our proposal

1. **Pre-process omics datasets in order to include biological information**
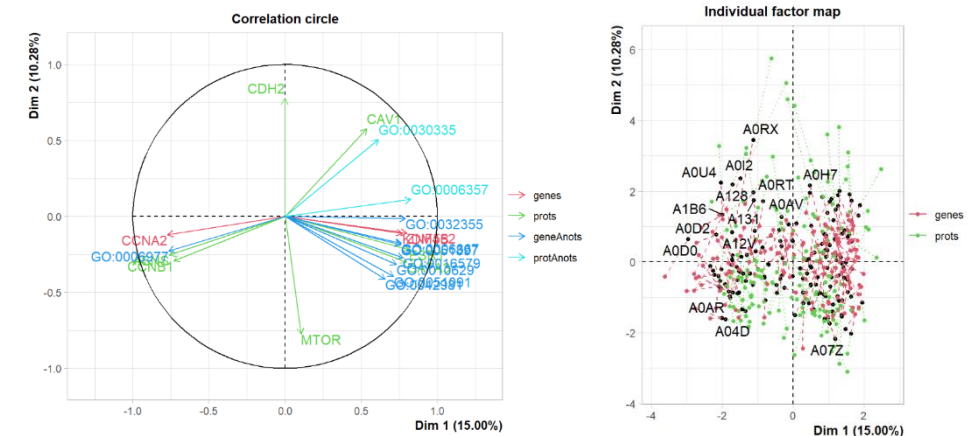   <u>**before**</u> **the joint analysis**

   → **Expanded datasets**

   Applying standard biological significance analysis methods
   (e.g. Gene over-representation; Gene Sets; Pathways; Networks… + Custom annotations)

2. **Analysis of the expanded datasets by the use of contrasted**
   joint Dimension Reduction techniques

   → MCIA (*made4, omicade*…)
   → RGCCA (*miXomics*…)
   → **Multiple Factor Analysis (*FactoMineR*…)**

3. **Process (semi-)automation in ease-to-use tools**

   → **Pipeline for workflow control / visualization**
   → R Package + Interactive application
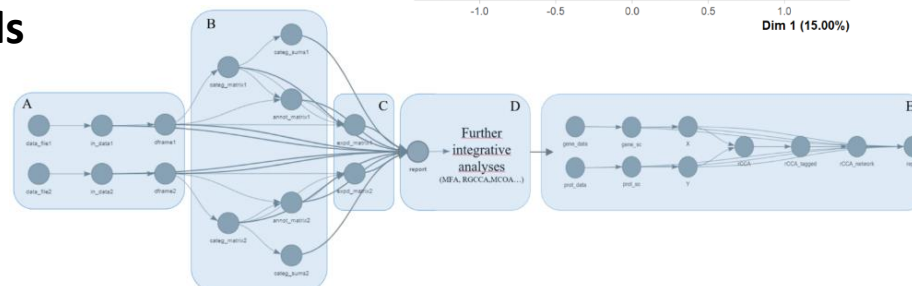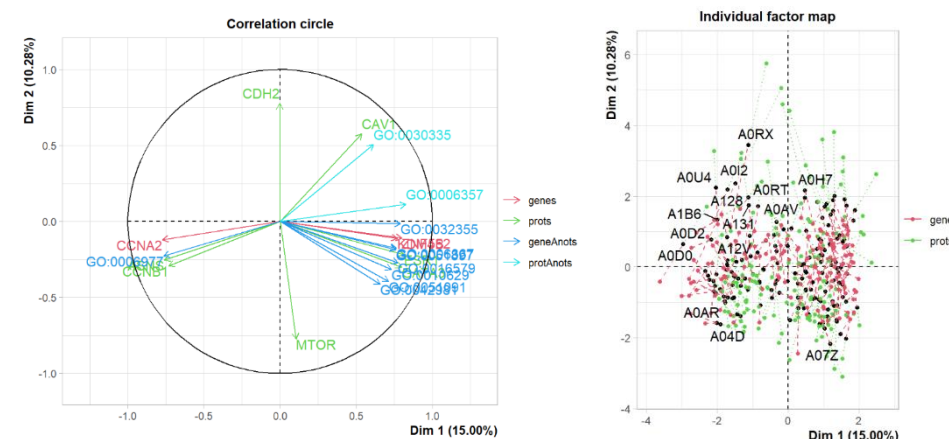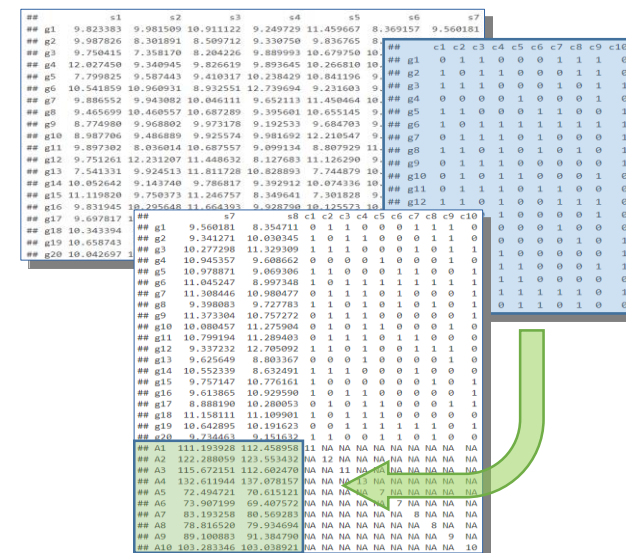
Overview of the methods – Biological Annotations
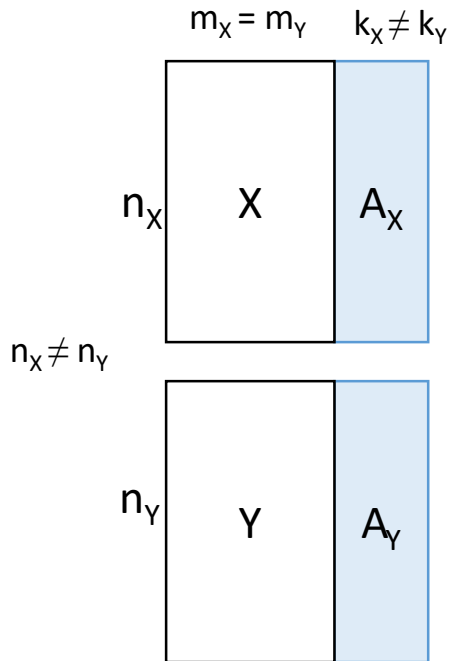
$m_X = m_Y$ $k_X \neq k_Y$

n_X X A_X

$n_X \neq n_Y$

n_Y Y A_Y

n : features (eg. gene ids, different for X and Y)
m : samples (individuals, same ones for X and Y)
k : biological annotations (different for X and Y)

- Gene enrichment (over-representation) | Gene Set Enrichment Analysis

- Annotations to

  - Gene Ontology (GO)

  - Gene Sets (MSigDB)

  - Networks (DisGeNET...)

  - Pathways (KEGG, Reactome)

  - other databases (MeSH...)

and custom annotations are also allowed!

humoral immune response

response mediated by circulating immunoglobulin

complement activation

immunoglobulin mediated immune response

B cell mediated immunity

complement activation, classical pathway

phagocytosis, recognition

Count
50
60
70
80

p.adjust
1.653503e-103
2.275115e-64
4.550230e-64
6.825346e-64
9.100461e-64

0.020.24
GeneRatio

Image created with clusterProfiler

##	c1	c2	c3	c4	c5	c6	c7
## g1	0	1	1	0	0	0	1
## g2	1	0	1	1	0	0	0
## g3	1	1	1	0	0	0	1
## g4	0	0	0	0	1	0	0
## g5	1	1	0	0	0	1	1
## g6	1	0	1	1	1	1	1
## g7	0	1	1	1	0	1	0
## g8	1	1	0	0	0	1	0
## g9	0	1	1	1	0	0	0
## g10	0	1	0	1	1	0	0
## g11	0	1	1	1	0	1	1
## g12	1	1	0	1	0	0	1
## g13	0	0	0	1	0	0	0
## g14	1	1	1	0	0	0	1
## g15	1	0	0	0	0	0	0
## g16	1	0	1	1	0	0	0
## g17	0	1	0	1	1	0	0
## g18	1	0	1	1	1	0	0
## g19	0	0	1	1	1	1	1
## g20	1	1	0	0	1	1	0

Overview of the methods – Data Expansion (I)

$m_X = m_Y$ $k_X \neq k_Y$

$n_X \neq n_Y$

n : features (eg. gene ids, different for X and Y)
m : samples (individuals, same ones for X and Y)
k : biological annotations (different for X and Y)

$$N_G = \phi(G, A_G), \quad G \in X, Y, \ldots,$$

$$N_X = \frac{1}{r_X}(X \times A'_X),$$

$$N_Y = \frac{1}{r_Y}(Y \times A'_Y),$$

Overview of the methods – Data Expansion (II)

$m_X = m_Y$ $k_X \neq k_Y$

$n_X \neq n_Y$

n_X X A_X

n_Y Y A_Y

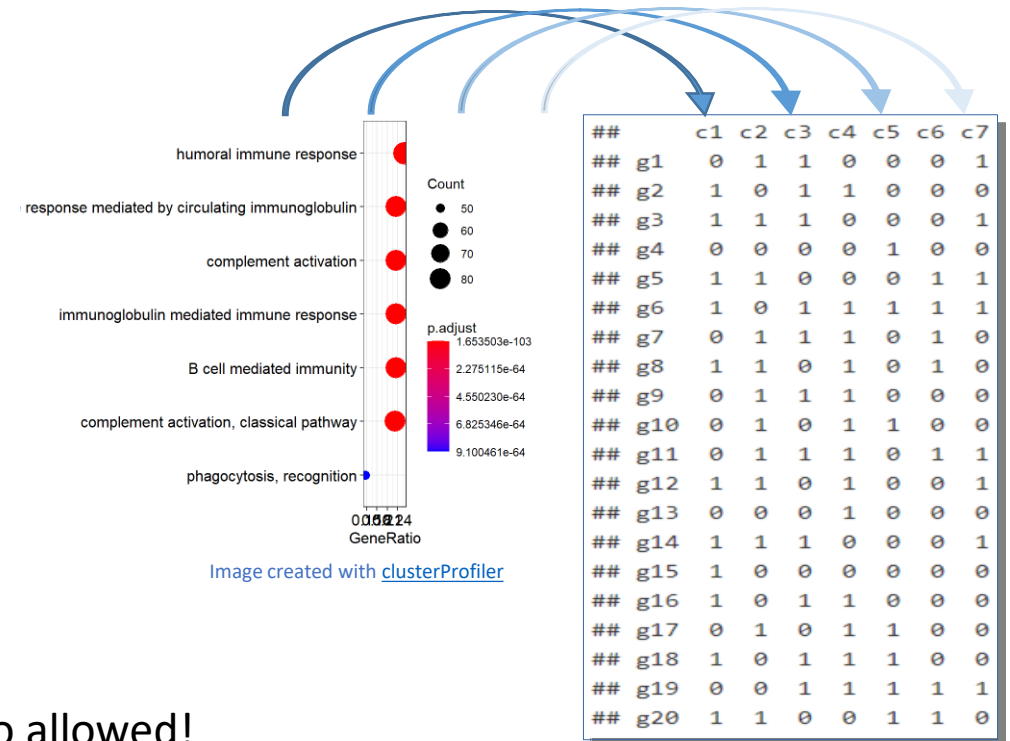n : features (eg. gene ids, different for X and Y)
m : samples (individuals, same ones for X and Y)
k : biological annotations (different for X and Y)

X A_X

N_X

Y A_Y

N_Y

X

Y

$N_{X\,(not\,Y)}$

N_{XY} *

$N_{Y\,(not\,X)}$

$*$ Distinct weights can be applied for factors X and Y of the common N_{XY} expanded matrix (e.g. in the same way as datasets are combined in MFA).

$$N_G = \phi(G, A_G), \quad G \in X, Y, \ldots,$$
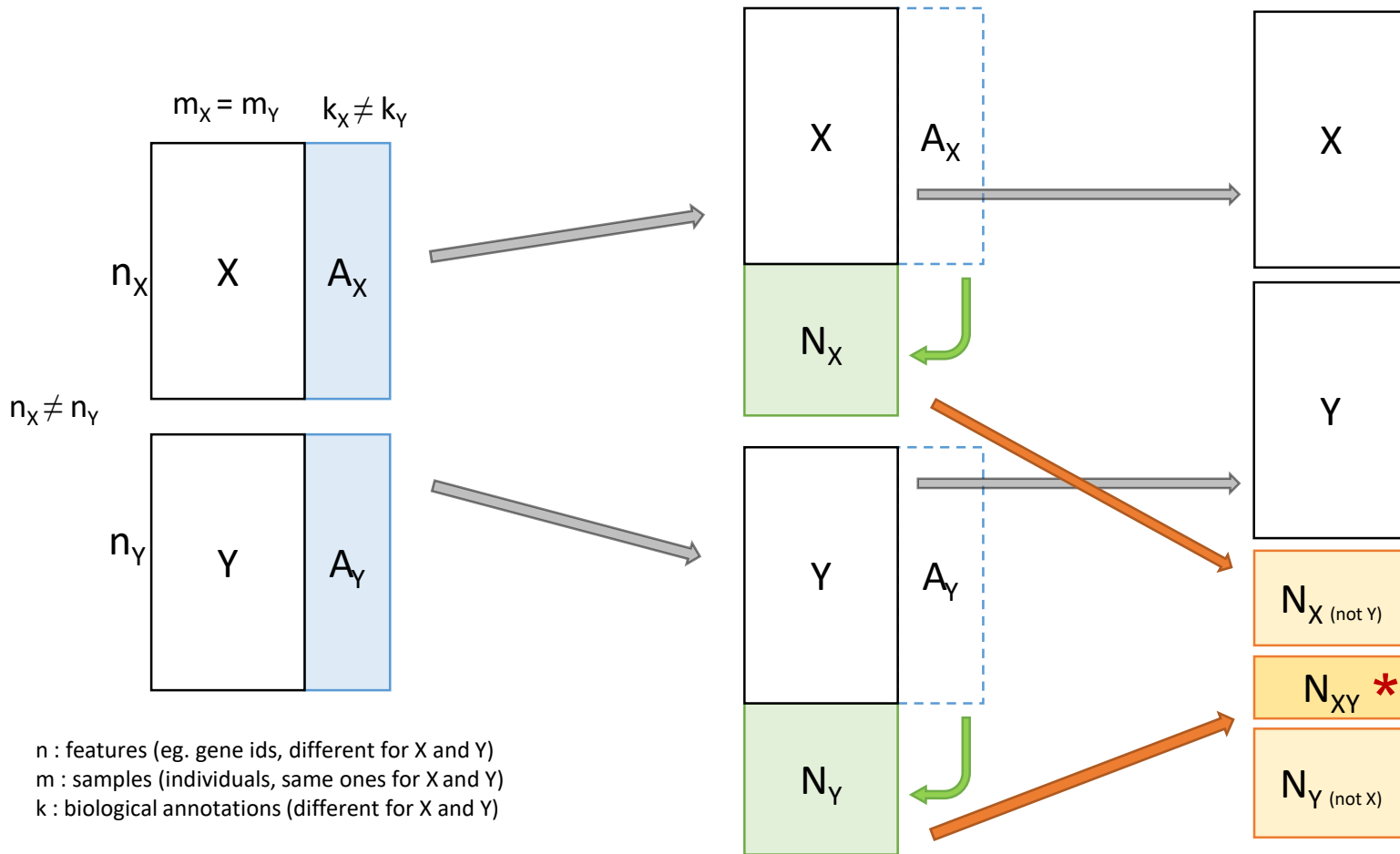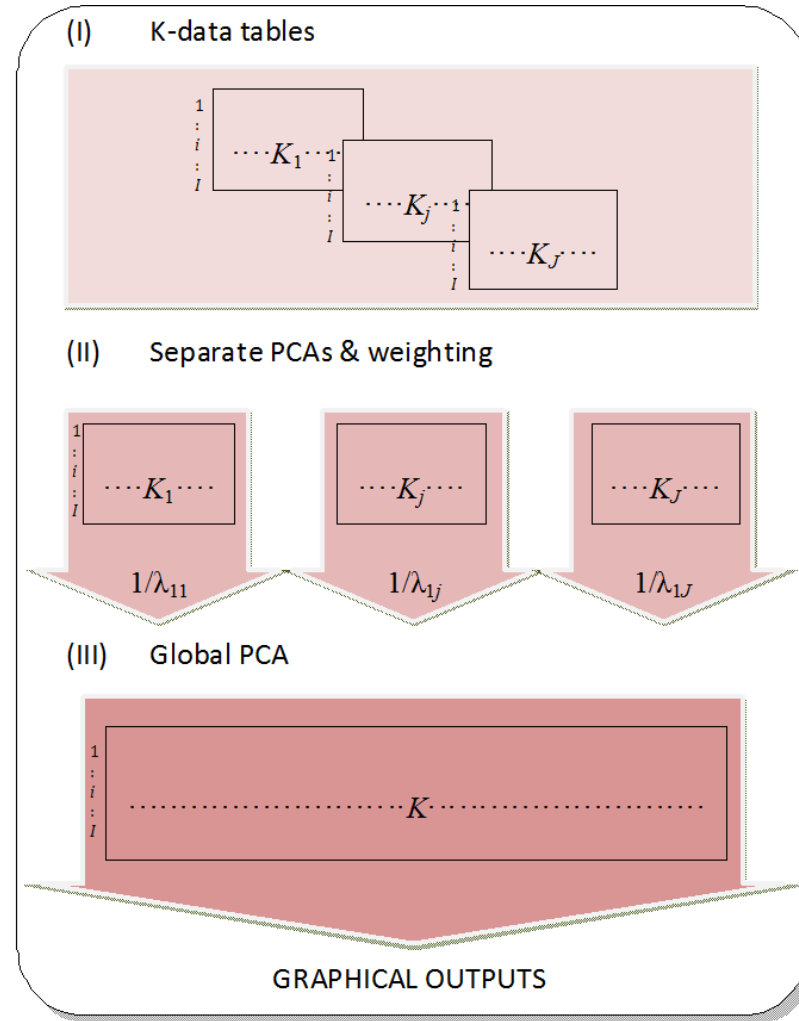
$$N_X = \frac{1}{r_X}(X \times A'_X),$$

$$N_Y = \frac{1}{r_Y}(Y \times A'_Y),$$

$$N_{XY} = \frac{w_X^{*}}{r_{XY}}(X \times A'_{XY}) + \frac{w_Y^{*}}{r_{XY}}(Y \times A'_{XY})$$

Overview of the methods – Multi-Block Analysis
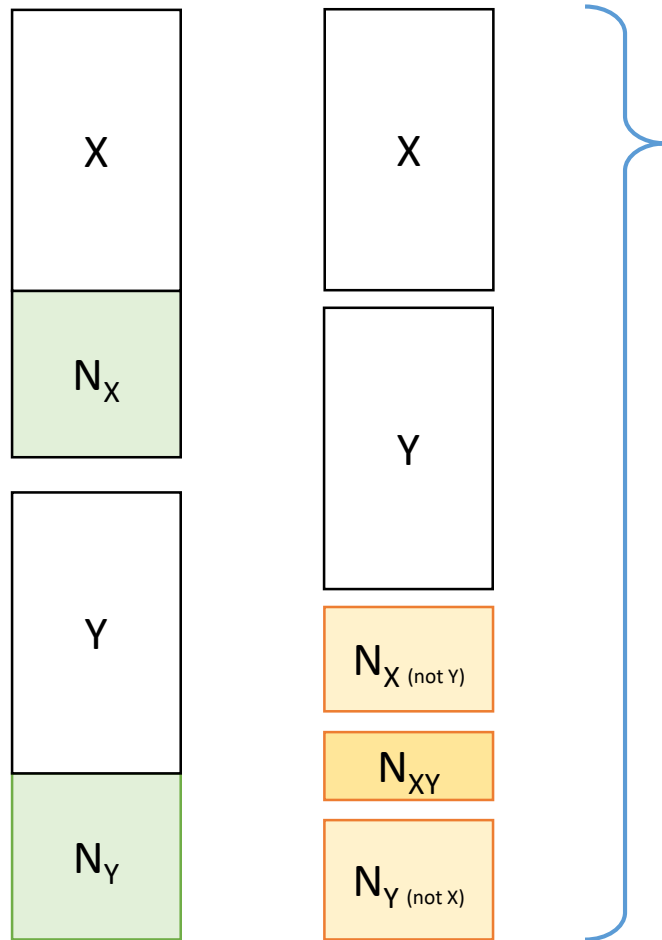
e.g. Multiple Factor Analysis

The core of MFA is a PCA applied to the whole set of variables,

Each group of variables is weighted, rendering possible the analysis of different points of view by taking them equally into account.

MFA allows to look for common factors by providing a representation of each matrix of variables.

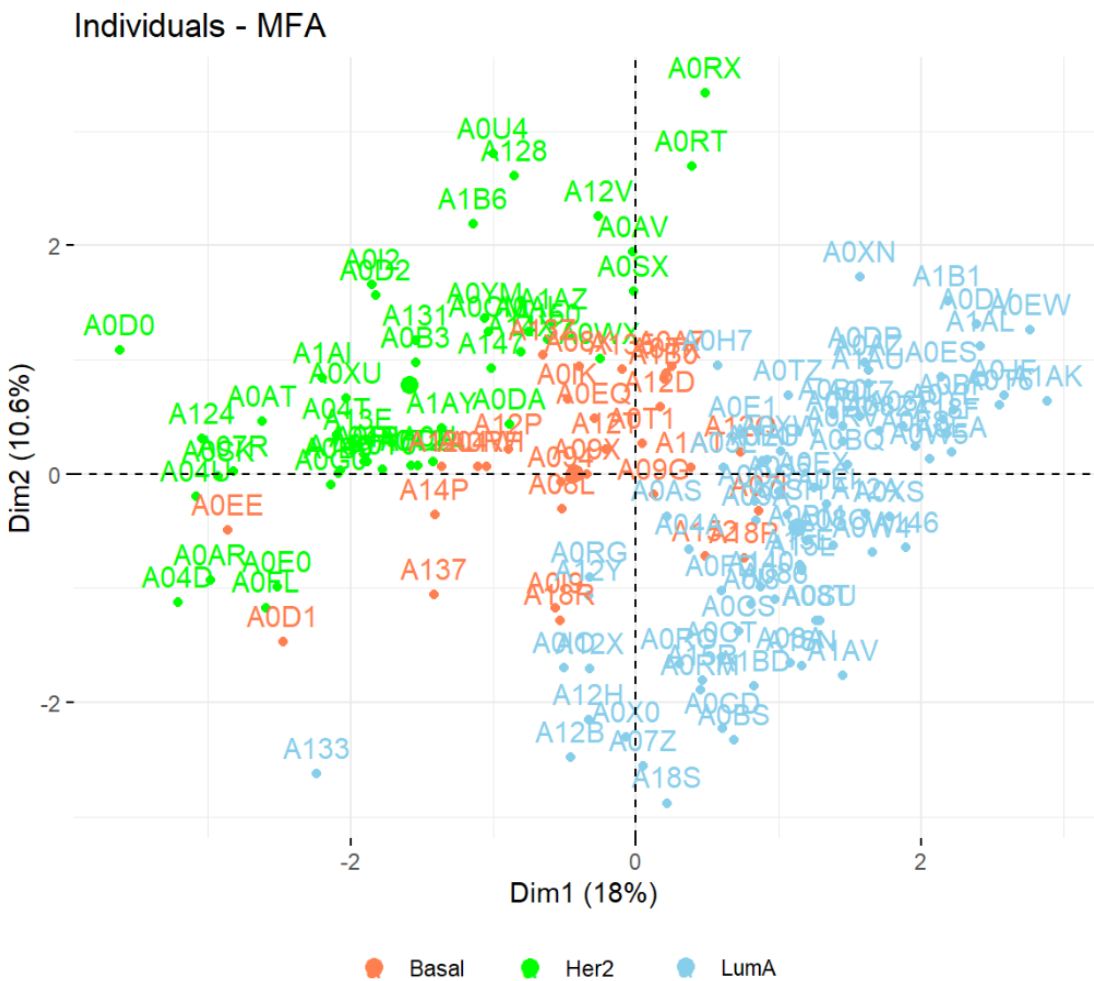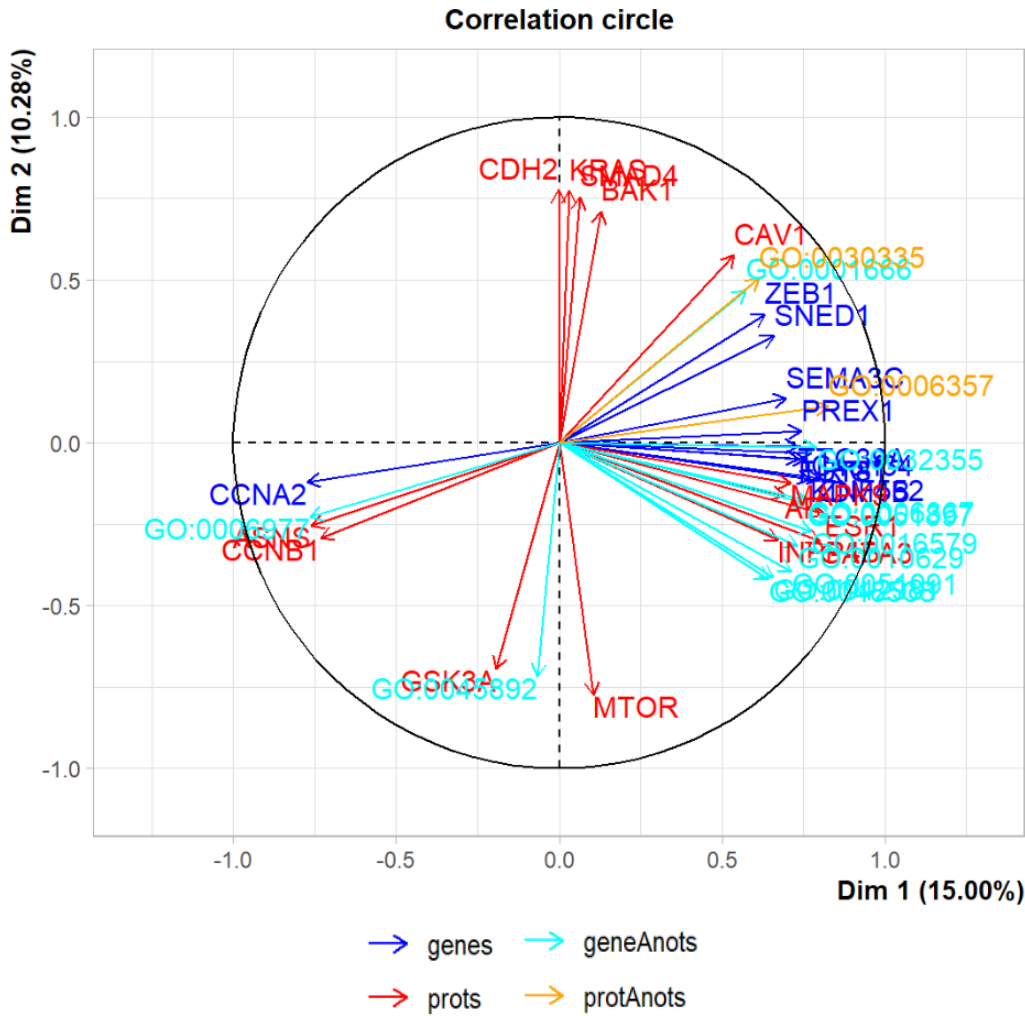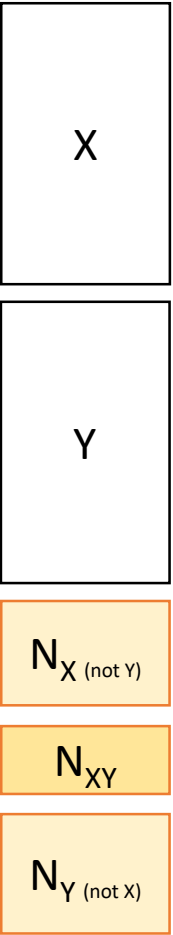# Example results – 150 TCGA breast cancer samples – MFA

Example results – *Targets* Pipeline Implementation

Reproducible workflow with a make-like pipeline toolkit *targets* package (https://books.ropensci.org/targets/)
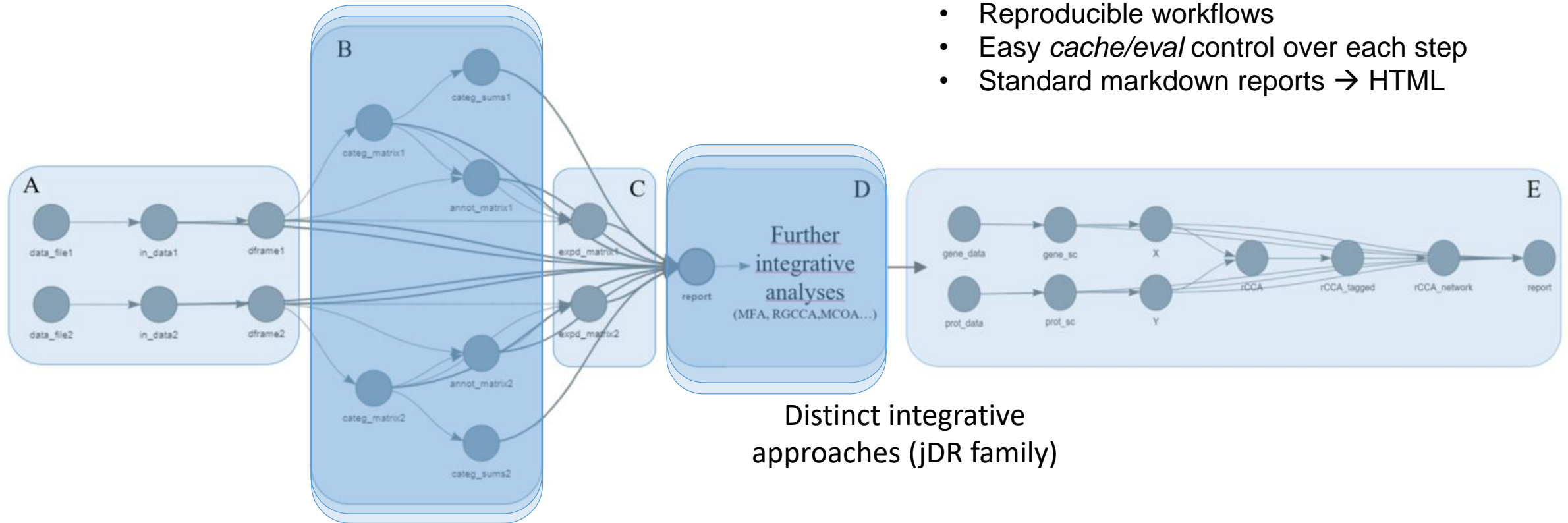
- Reproducible workflows
- Easy *cache/eval* control over each step
- Standard markdown reports → HTML

Flexible annotation
& expansion options

Distinct integrative
approaches (jDR family)

Visit poster #84 **An R pipeline using the "targets" package for Multi-Omics Integrative Analyses**
Poster session 9 (25th 18-20h at Auditorio)

Conclusions

❏ The need for a **better biological interpretation of multi-omics integrative methods** let us to consider the inclusion of biological information during (not after) the analysis process

❏ We propose a method focused on the **expansion of the starting omics datasets, by adding new annotation-derived features to that matrices**, before applying the integrative analysis

❏ This approach allows the inclusion of relevant information from the **main biological annotation tools**, as well as any custom annotation, **combined with the use our preferred Dimension Reduction techniques**

❏ We have implemented a **pipeline for reproducible and easy-to-use execution**, that facilitates the control of each step, the visualization of results and their reporting to PDF/HTML formats

Acknowledgements

Staff of the Statistics and Bioinformatics Unit (UEB, VHIR)
    Special thanks to **Ricardo Gonzalo Sanz**

Neurovascular Diseases Laboratory, VHIR
    Teresa García-Berrocoso, Alba Simats, Joan Montaner

Grup de Recerca en Bioestadística i Bioinformàtica, GRBIO
Universitat de Barcelona & Universitat Politècnica de Catalunya

*Integrative Analysis of Multi-Omics Data
with Addition of Biological Knowledge*

Ferran Briansó & Alex Sánchez-Pla

THANKS FOR YOUR ATTENTION