# Individualized multi-omic pathway deviation scores using multiple factor analysis

Andrea Rau, Florence Jaffrezic, Denis Laloë, Paul L. Auer, Regina Manansala, Michael J. Flister, Hallgeir Rui

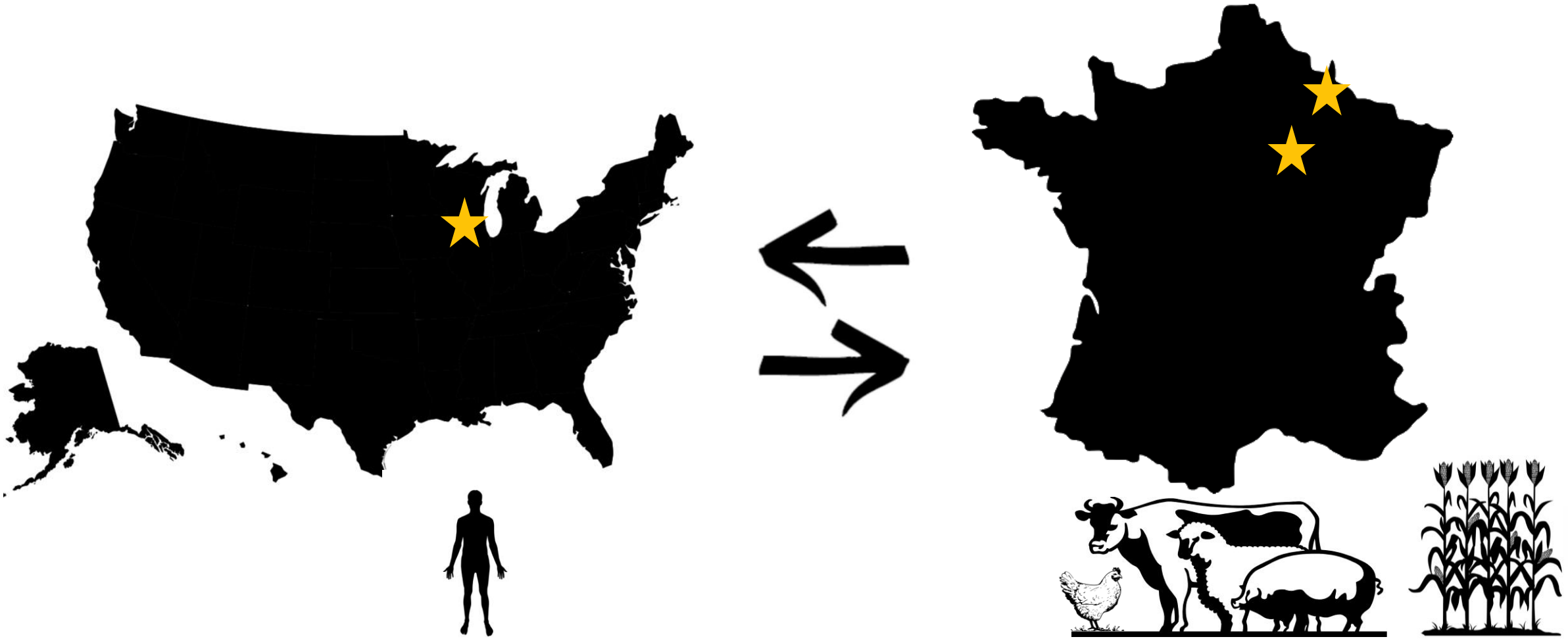
# Individualized multi-omic pathway deviation scores using multiple factor analysis

ANDREA RAU

EUROBIOC

DE DUVE INSTITUTE, UCLOUVAIN @ BRUSSELS

DECEMBER 9, 2019

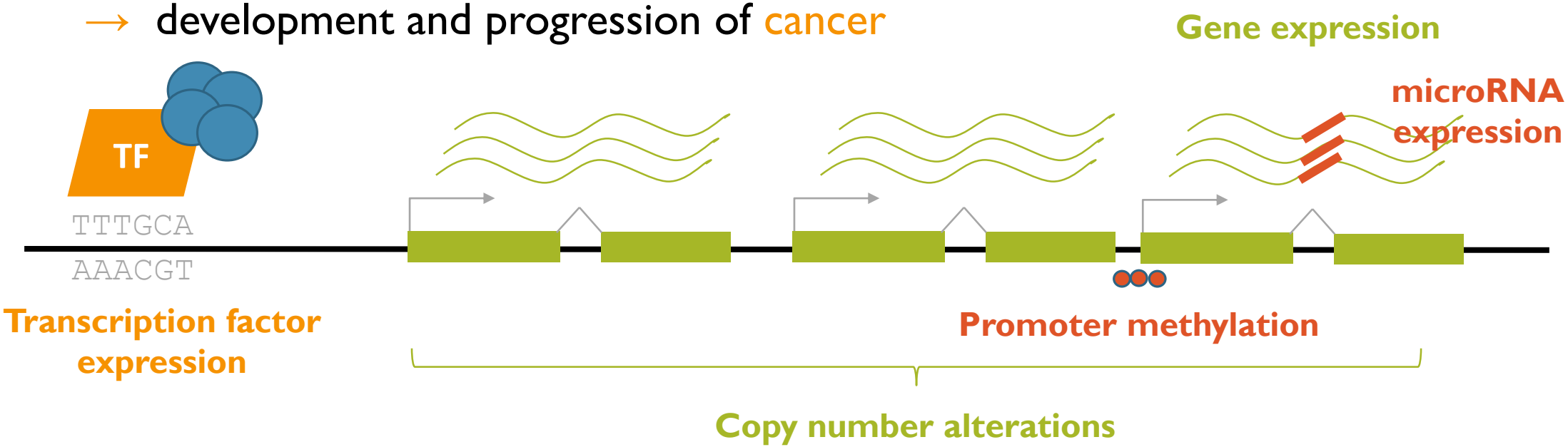


# Transcriptional regulation (in cancer genomes)

Dysregulated genes regulating cell growth/differentiation

→ uncontrolled cell growth

→ development and progression of cancer




...GCAGCGTTCGA...

...GCAACGTTAGA...

→ **Somatic mutations** within tumors, **Germline genetic variation**

… + Chromatin accessibility + RNA processing + RNA stability + Protein activity + …

# The Cancer Genome Atlas (TCGA)

- Comprehensive, **multi-dimensional maps** of key genomic changes in **33 cancer types** from **11k+ individuals**

- Publically available data (multi-tiered data depending on patient identifiability)

- **Widely** used by the research community (1000+ publications by TCGA network + independent researchers)

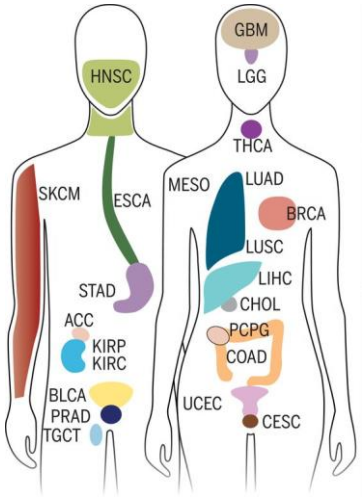
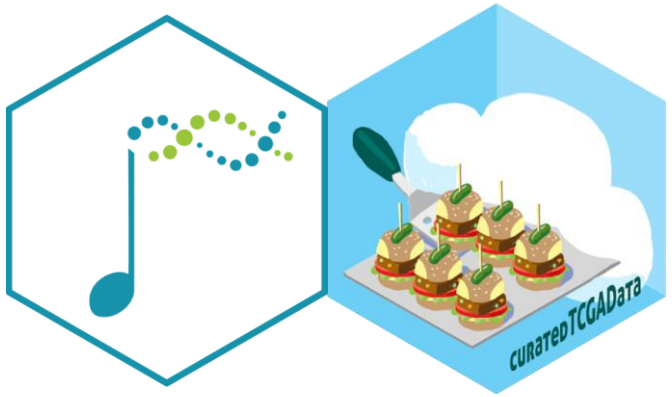



*Image: Corces et al. (2018)*




# Multi-omic data → Multivariate, multi-table methods

- Account for **interdependencies** within and across data types

- (Partially) **matched** omics data across samples or biological entities (e.g., genes)

- In some contexts, limited/incomplete *a priori* knowledge of relevant phenotype groups for comparisons = **unsupervised analysis**

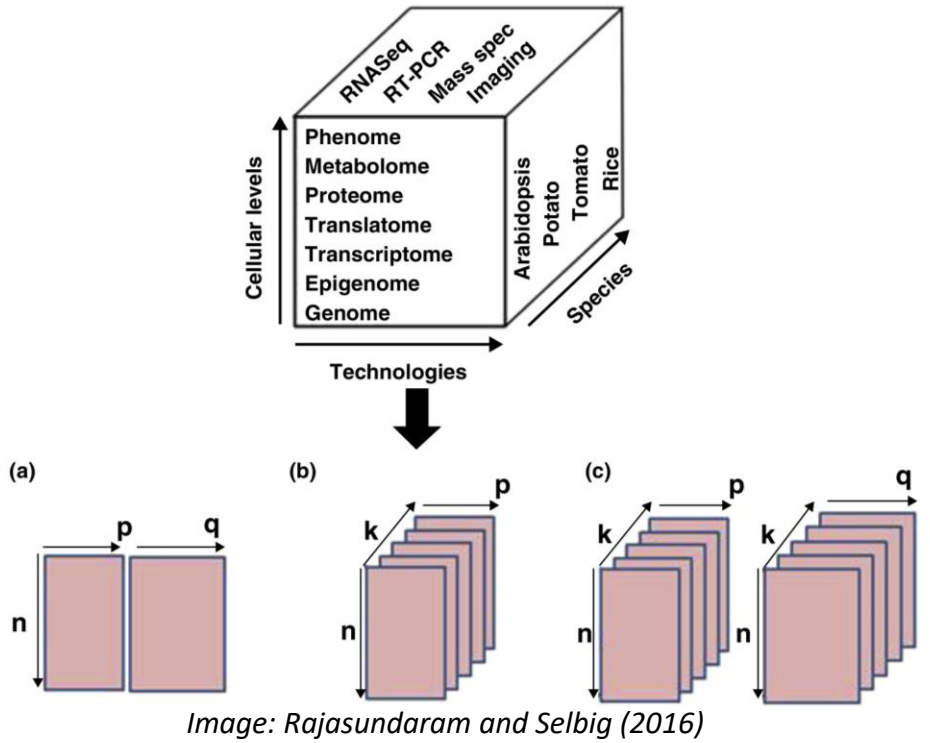



*Image: Rajasundaram and Selbig (2016)*

~~How do we integrate multi-omic data?~~

What question are we specifically addressing? How can we use multi-omic data to answer that question?

# Our focus is specifically on pathway-level inference


reactome

BIOCARTA


For a given pathway of interest, can we **identify** and **quantify** highly **aberrant individuals** in a sample based on **multi-omic data**?

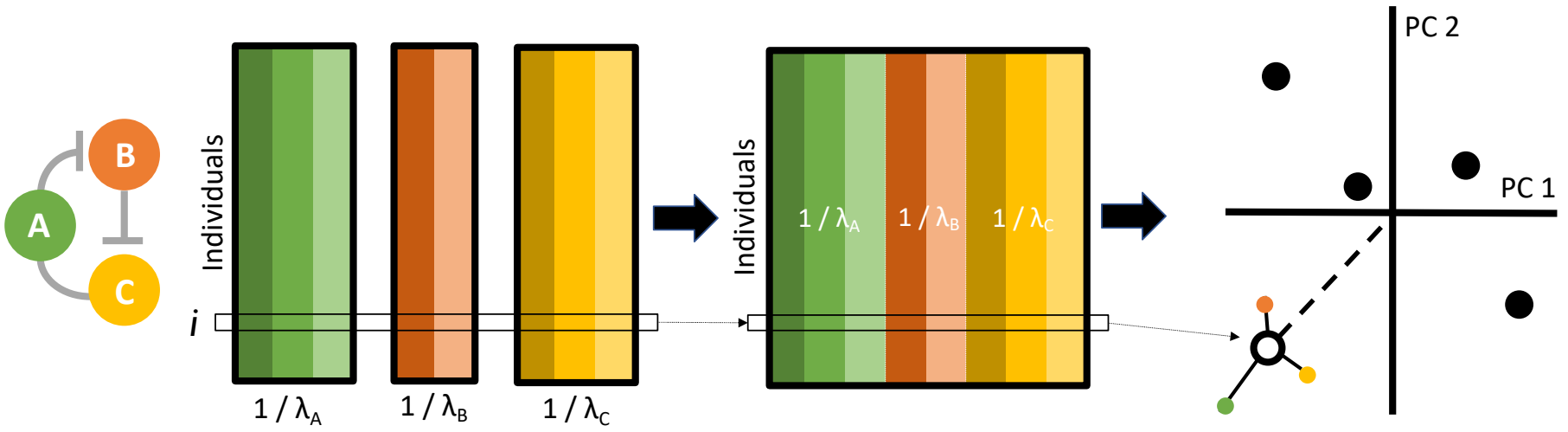
Does patient prognosis correlate with large pathway deviation scores?

Which individuals have the most aberrant profiles for pathways of interest?

Which genes / omic drive these aberrant scores?

# *padma*: Pathway deviation scores using Multiple Factor Analysis

Define an *individualized* pathway-level deregulation score based on multi-omic data using **MFA**



# Individualized pathway and per-gene deviation scores

In the multi-dimensional MFA consensus space, the origin represents the "average" pathway profile across genes, omics, and individuals.

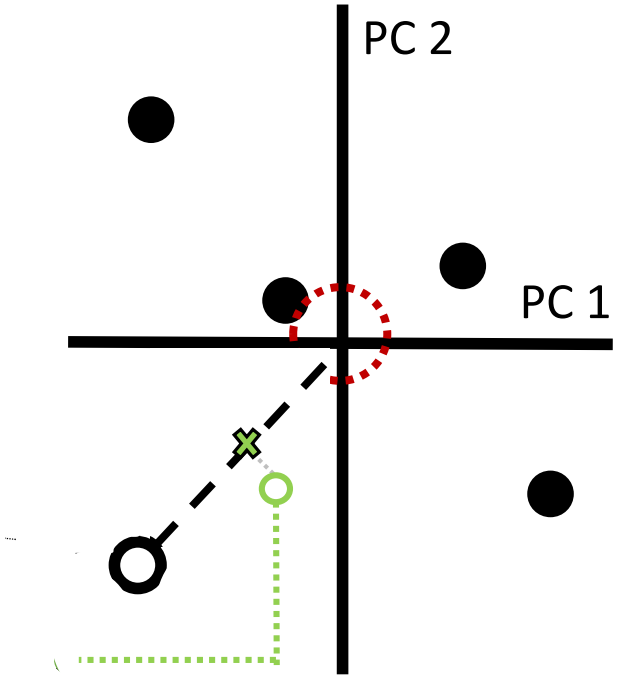
***Pathway deviation score*** = Euclidean distance of MFA factors to the origin for each individual

$$d_i^2 = \sum_{l=1}^{L} f_{i,l}^2$$

Partial MFA factor scores can be computed for each gene

Decompose each pathway deviation score into ***per-gene deviation scores****

$$d_{i,g} = \frac{\sum_{l=1}^{L} f_{i,l}(f_{i,l,g} - f_{i,l})}{\sum_{l=1}^{L} f_{i,l}^2}$$




Richness of additional MFA outputs:

→ Decomposition of the total variance by MFA component

→ % contribution to the inertia of each axis by omic, gene, or individual

FACTOMINER

*We normalize the per-gene deviation scores by each individual's pathway deviation score.

# Applying *padma* to TCGA data

Breast invasive carcinoma (**BRCA**; *n* = 504) and lung adenocarcinoma (**LUAD**; *n* = 144)


- Batch correction performed using removeBatchEffects in *limma*
- RNA-seq + promoter methylation + copy number alterations + miRNA-seq
- miRNA → gene mapping provided by miRTarBase (exact matches, Functional MTI predictions)

- **1136 MSigDB curated canonical pathways** (Biocarta, PID, Reactome, Sigma Aldrich, Signaling Gateway, Signal Transduction Knowledge Environment, Matrisome Project)

Patient prognosis measured using progression-free interval survival times (LUAD) and histological grade (BRCA)

# For which pathways do large deviation scores correlate with poor prognosis?
## Progression-free interval (LUAD)

- **14 pathways** significantly associated with survival (Cox PH*, BH padj < 5%)
- **Higher scores = worse outcome**
- Not linked to tumor mutational burden

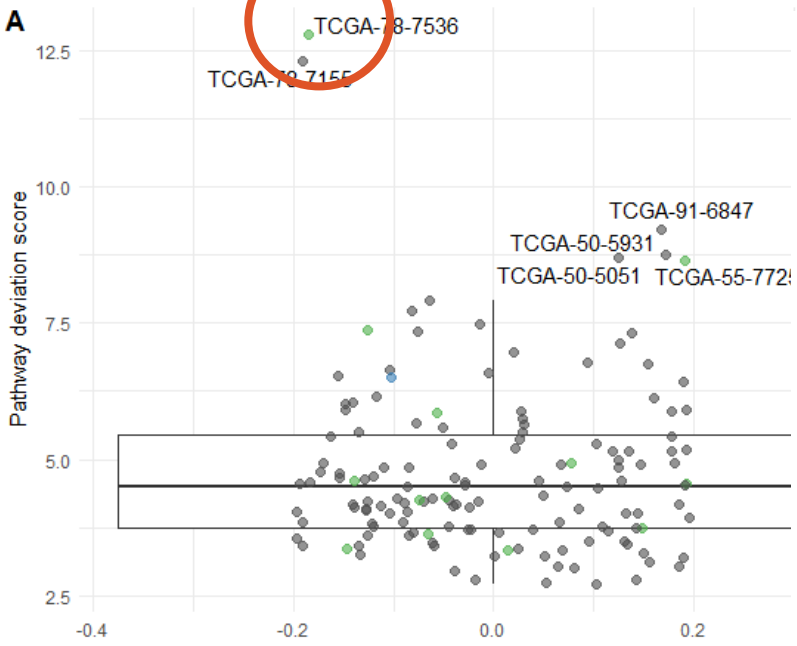
| Pathway name | Database | Adj. p-value | Hazard ratio | # of genes |
|---|---|---|---|---|
| **D4-GDI (GDP dissociation inhibitor) signaling pathway** | Biocarta | 0.0111 | 1.2692 | 13 |
| NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10 | Reactome | 0.0111 | 1.2839 | 12 |
| Class I PI3K signaling events mediated by Akt | PID | 0.0251 | 1.1700 | 35 |
| ATM signaling pathway | Biocarta | 0.0265 | 1.1644 | 20 |
| CARM1 and regulation of the estrogen receptor | Biocarta | 0.0265 | 1.1426 | 35 |
| Homologous recombination repair of replication-independent double-strand breaks | Reactome | 0.0265 | 1.2432 | 16 |
| Role of BRCA1, BRCA2, and ATR in cancer susceptibility | Biocarta | 0.0467 | 1.1823 | 21 |
| … | … | … | … | … |

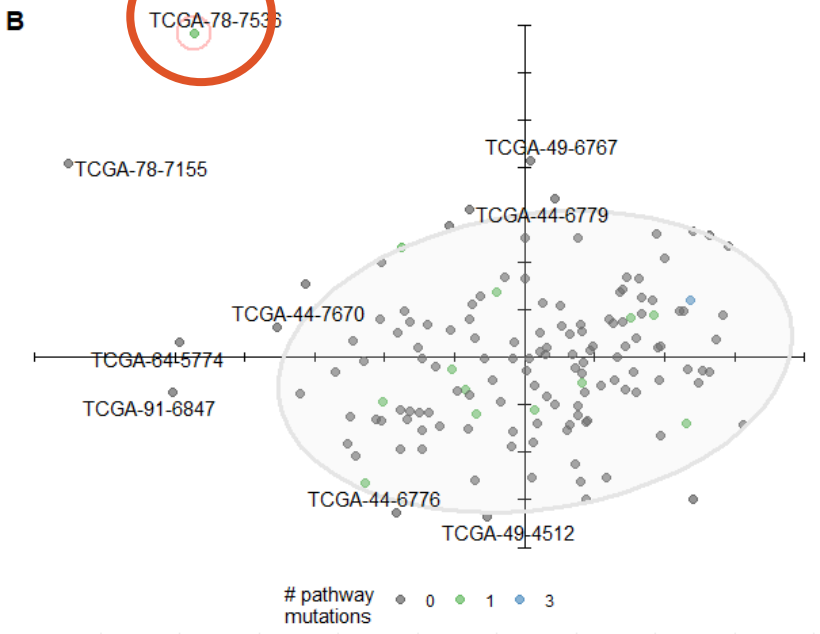
Focus on the D4-GDP dissociation inhibitor signaling pathway…

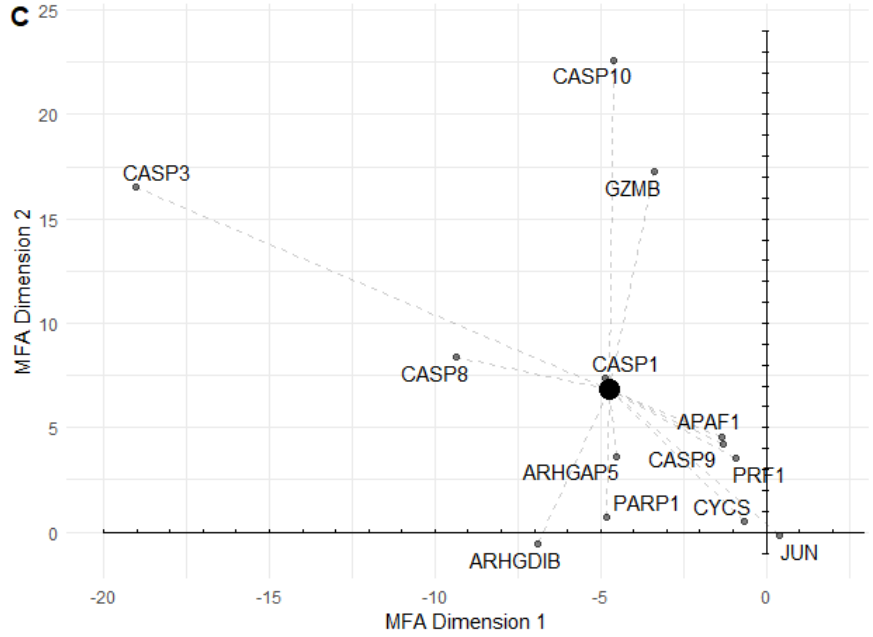
* Test performed on 5% most skewed pathways

# Which individuals have the most highly aberrant multi-omic profiles?

D4-GDI signaling pathway, LUAD

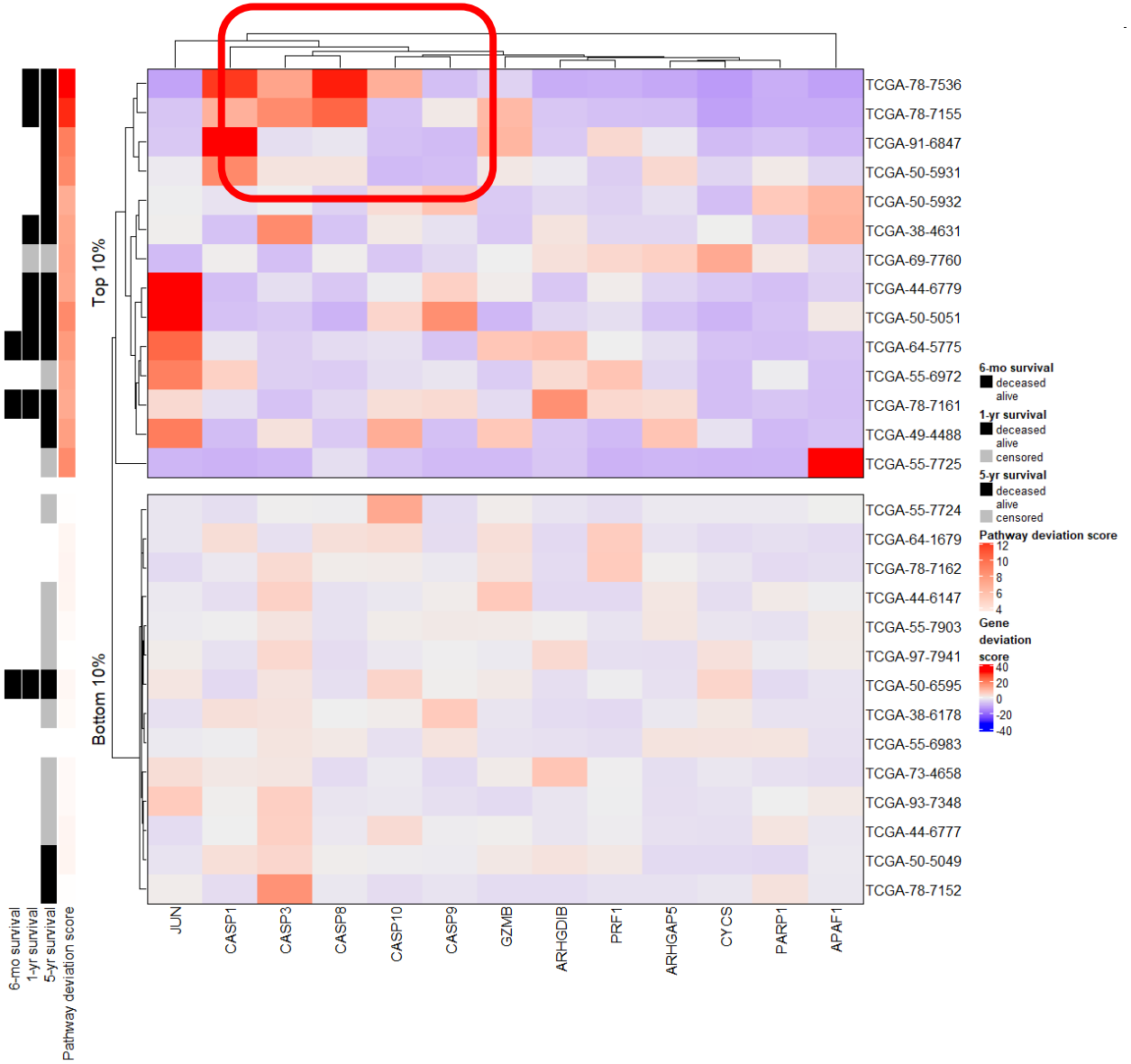


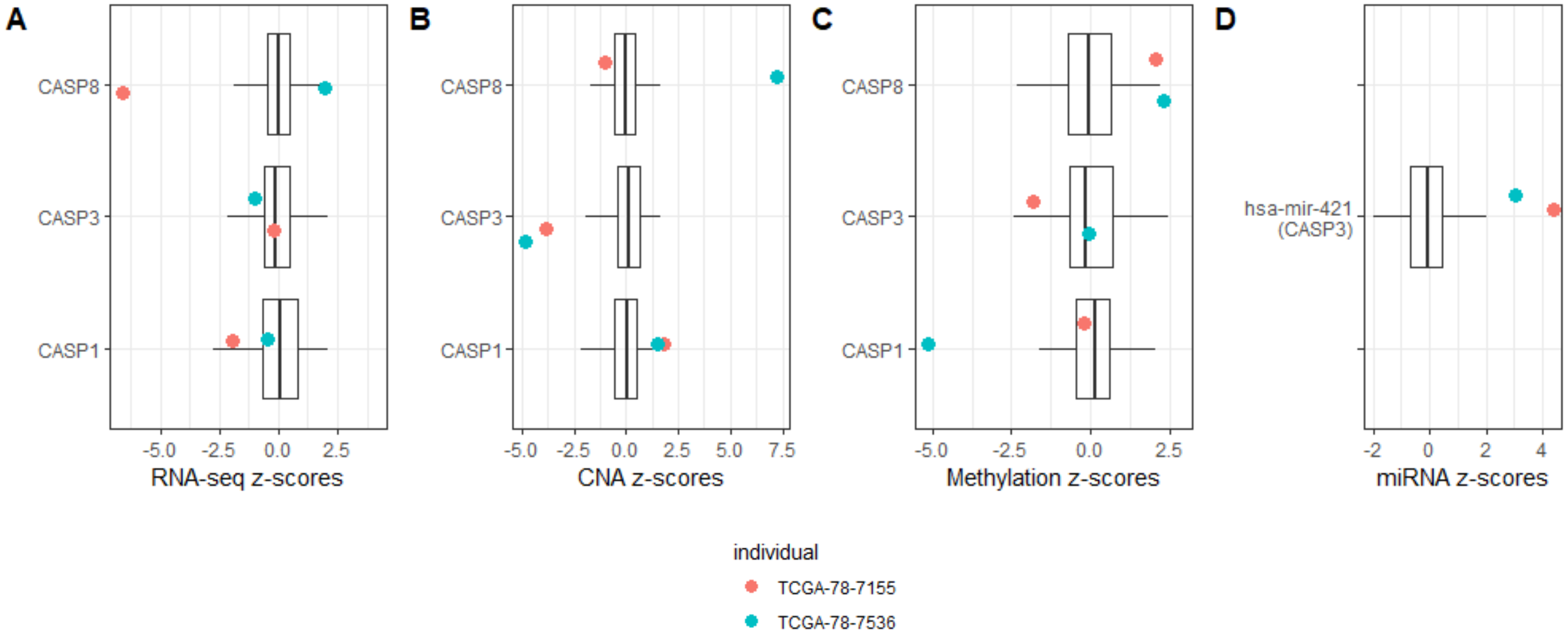

MFA 1: RNA-seq (54.38%)
MFA 2: methylation (42.29%)
MFA 3: CNA (59.18%)

# Which genes/omics drive large pathway deviation scores?




→ **CASP1**, **CASP3**, and **CASP8** all have high gene-level deviation scores for the two most extreme individuals…

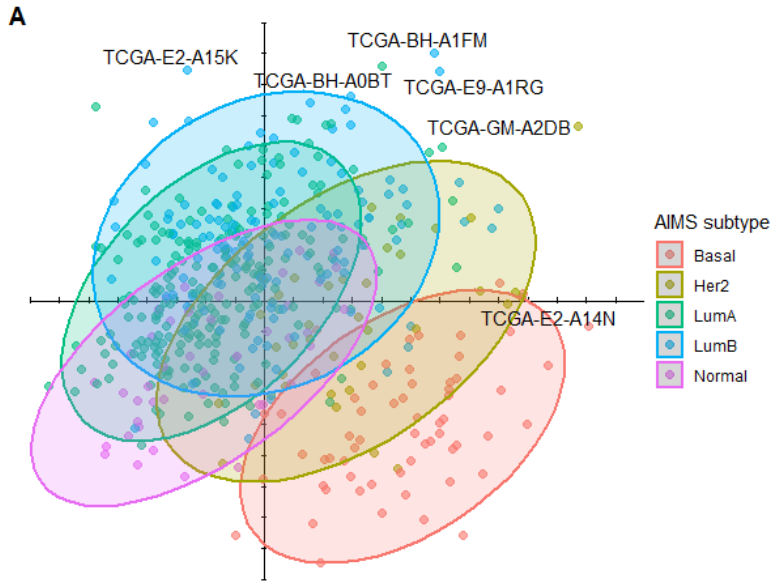
# Which genes/omics drive large pathway deviation scores?

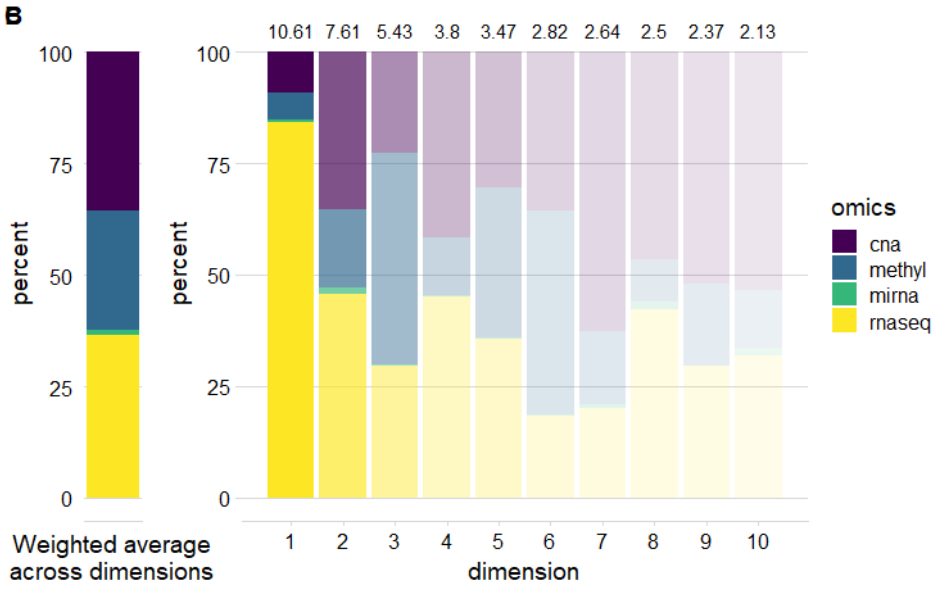


# Pathway deviation scores are associated with other clinically relevant phenotypes

- **Nearly all pathways** are associated with two measures of histological grade
- Higher scores = worse outcome

| Pathway | Database | Ranking | # of genes |
|---|---|---|---|
| **Signaling by Wnt** | Reactome | 3.16 | 63 |
| Apoptotic execution phase | Reactome | 5.00 | 52 |
| APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1 | Reactome | 6.78 | 64 |
| … | … | … | … |




* Mitotic index and nuclear pleomorphism (ANOVA, BH padj < 5%)

# Pathway deviation variability is associated with BRCA subtype

# *padma* results on TCGA breast and lung cancer

(RNA-seq + miRNA-seq + methylation + CNA data, MSigDB canonical pathways)

- Larger *padma* deviation scores = increasingly aberrant pathway variation with significantly worse prognosis (survival, histological grade) in breast and lung cancer
- Potential outlier detection tool

**Innovative use of existing MFA method to calculate and graphically explore individualized multi-omic pathway deviation scores**

## Future work:

- Potential integration into **Bioconductor** ecosystem (notably, for the *MultiAssayExperiment* class)
- Incorporation of known **hierarchical structure** among genes in pathway
- **Interactivity** for result exploration through an integrated Shiny app
- Extensions for **highly structured data** typical in agronomy (e.g., multi-omic data from divergent chicken lines subject to feed/heat stress or maize diversity panels under control/cold conditions)


Rau et al. (2019) Individualized multi-omic pathway deviation scores using multiple factor analysis. *bioRxiv* https://www.biorxiv.org/content/10.1101/827022v2

WISCONSIN BREAST CANCER SHOWHOUSE
NIH NATIONAL CANCER INSTITUTE
MCW MEDICAL SCHOOL | ADVANCING A HEALTHIER WISCONSIN ENDOWMENT
MEDICAL COLLEGE OF WISCONSIN CANCER CENTER
THE MARY KAY FOUNDATION

# Acknowledgements

AgreenSkills plus
Pathways for inventive researchers
Co-funded by the European Union
agreenium





INRA SCIENCE & IMPACT
UWM
MCW

**Individualized multi-omic pathway deviation scores using multiple factor analysis**

Andrea Rau, Regina Manansala, Michael J. Flister, Hallgeir Rui, Florence Jaffrézic, Denis Laloë, Paul L. Auer

**doi:** https://doi.org/10.1101/827022



CSH Cold Spring Harbor Laboratory
bioRxiv THE PREPRINT SERVER FOR BIOLOGY