TRANSPAN

COST ACTION CA21116

Multiomics analyses via integration of various omic data

Sezerman lab

if possible...

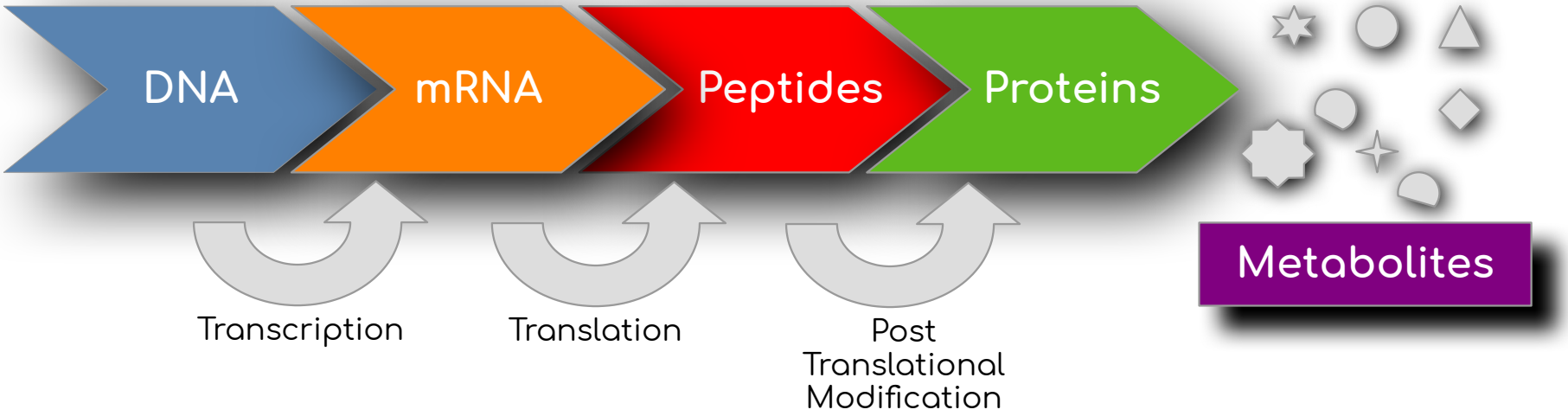
Berkay Y. Ekren, MSc.

Biostatistics and Bioinformatics
PhD student

Contents

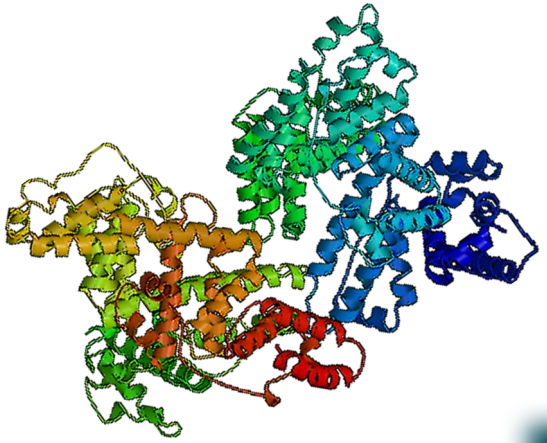
- Omics?
- Different omics
- Different analysis methods
  - Popular Methods
  - Popular Methods – Under a magnifying glass
  - mixOmics
- Case Studies
  - CAKUT
  - Colorectal Cancer

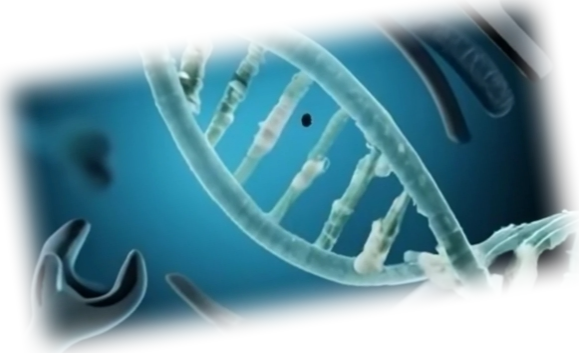
Central Dogma



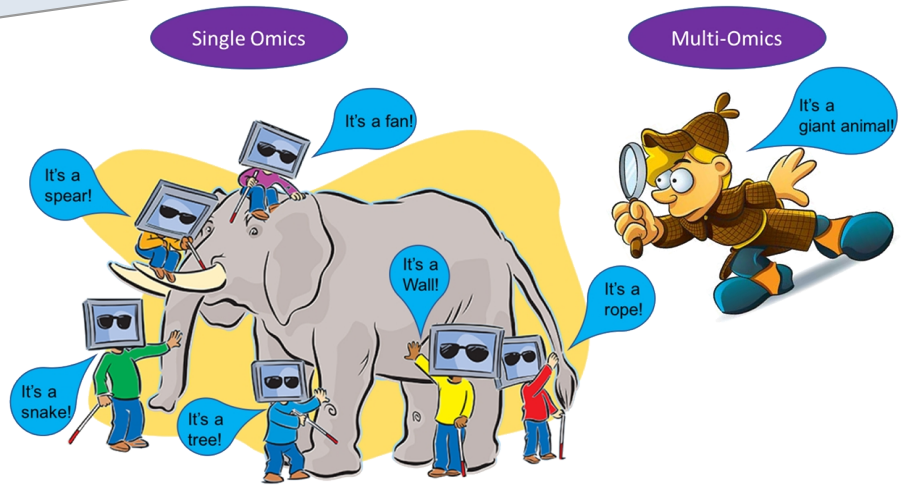
What is omic data?

- Genomics
- Transcriptomics
- Peptidomics
- Proteomics
- Metabolomics
- Metagenomics
- Phenomics
- etc...





Popular Methods



- Supervised
  - (W)SNF – (Weighted) Similarity Network Fusion
  - RF – Random Forest
  - PLS – Partial Least Squares
  - SVM – Support Vector Machine
- Unsupervised
  - RGCCA – Regularized Generalized Canonical Correlation Analysis (~)
  - MOFA – Multi Omics Factor Analysis
  - Clustering (k-means, Bayesian clustering...)
  - JIVE – Joint and Individual Variation Explained
  - CCA – Canonical Correlation Analysis
  - Matrix Factorization Techniques (jDR, NMF, tICA, PCA)

More Popular Methods

MOFA

- Early integration by itself
- Data normalization and transformation
- Matrix creation
- Factor model and latent variable identification
- Bayesian inference model
  - For each omics levels together
- Interpretation and biological validation

SNF

- Can be used as a late integration method
- Normalization and filtering done separately
- Pairwise similarities matrix and network creation for each dataset
- Network merging
- Classification or clustering

jDR

- Created with early integration dimension reduction tools – JIVE, MOFA, GRCCA, PCA...
- Data normalization and filtering with another tool
- Feature selection
- Perform dimension reduction
- Apply clustering
- Interpretation and biological validation

iClusterPlus

- Early integration
- Data normalization and filtering with proper sample identifiers
- Data combined as lists in R
- Clustering
  - Needs each dataset to be normally distributed
  - Gaussian Mixture Model with iterative Bayesian framework
- Determine covariance and includes correlations
- Interpretation and biological validation

mixOmics

Analysis methods:

- Single Omic Data
  - sPCA or independent (sparse Principal Component Analysis)
  - sPLS-DA (sparse Partial Least Squares – Discriminant Analysis)
- N-Integration
  - sPLS (sparse Partial Least Squares)
  - rCCA (regularized Canonical Correlation Analysis)
- P-Integration
  - MINT (Multi-omics Integration using Network-based Tools)
- mixMC

mixOmics (Late Integration)

N Integration

- Normalization
- 2-omic integration
  - PLS
    - Capture maximum covariance
  - sPLS
    - Better feature selection
  - sDA
    - Classification
- Multi-omic integration (blocks)
  - sPLS
    - Feature selection & maximum correlation
  - PLS
    - More condensed data
  - sPLS-DA
    - Classification

mixMC

- Used for integrating microbiome data, specifically
- Multiple omics data
- Separate normalization and feature selection
  - Bayesian network
  - Majority voting
- Data integration via matrix
- Bayesian clustering
- Interpretation and biological validation

P Integration

- MINT workflow
- Multiple omics data
  - Aligning by sample
  - Network construction for each dataset
  - Data integration
  - Dimentionality reduction (sPLS)
  - Feature selection
  - sPLS or PLS
  - Visualization and biological validation

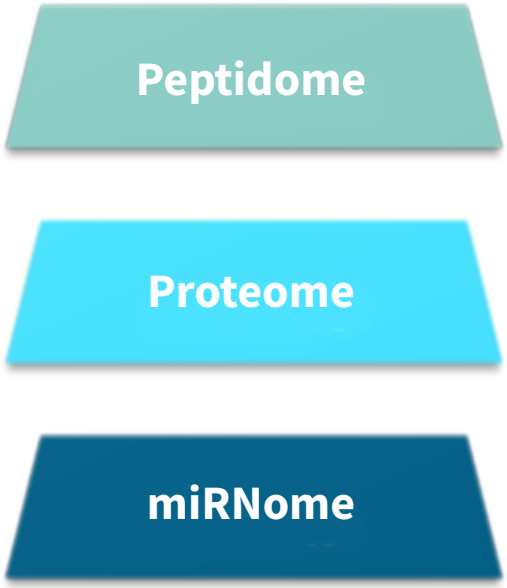
CAKUT – Case Study

- Congenital anomalies of the kidney and urinary tract (CAKUT) is a set of developmental abnormalities consisting of a broad spectrum of renal and urinary tract malformations.
- The disease prevalence is nearly 1 in every 500 live births, and the disease causes death approximately in %0.05.
- According to European Registry for Children on Renal Replacement Therapy, 41.3% of children that receive renal replacement therapy suffer from CAKUT as a major cause of end stage renal disease (ESRD) in pediatric patients.

CAKUT – Case Study

Samples

- Peptidome data:
  - Number of samples: 53
  - Number of peptides: 1008
- Proteome data:
  - Number of samples: 61
  - Number of proteins: 1662
- miRNome data:
  - Number of samples: 150
  - Number of miRNAs: 503

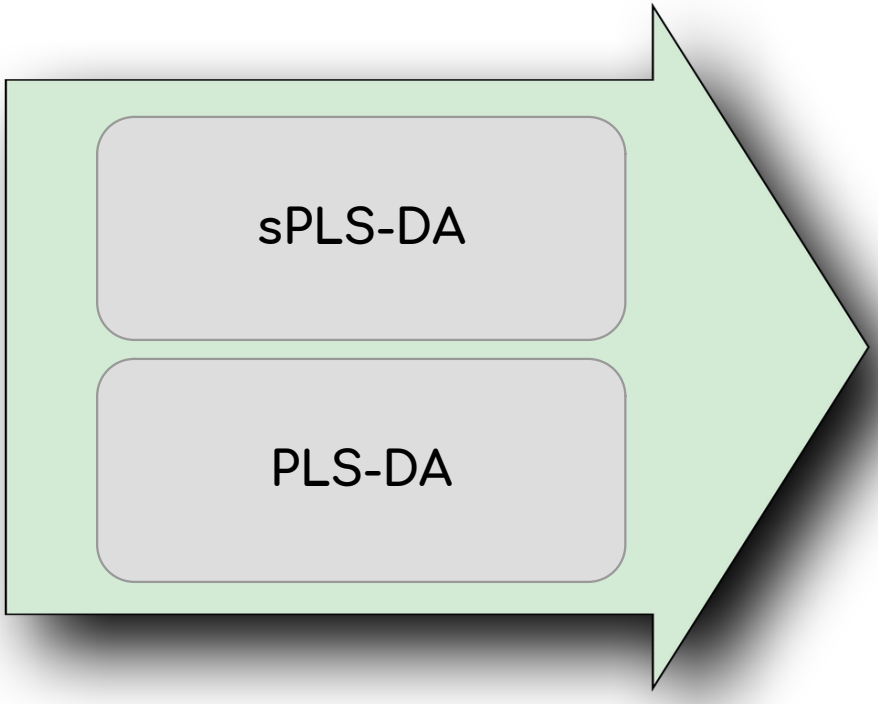


Bayjanov *et al. Nature Scientific Reports.* 2024.

CAKUT – Case Study

1st Method Workflow

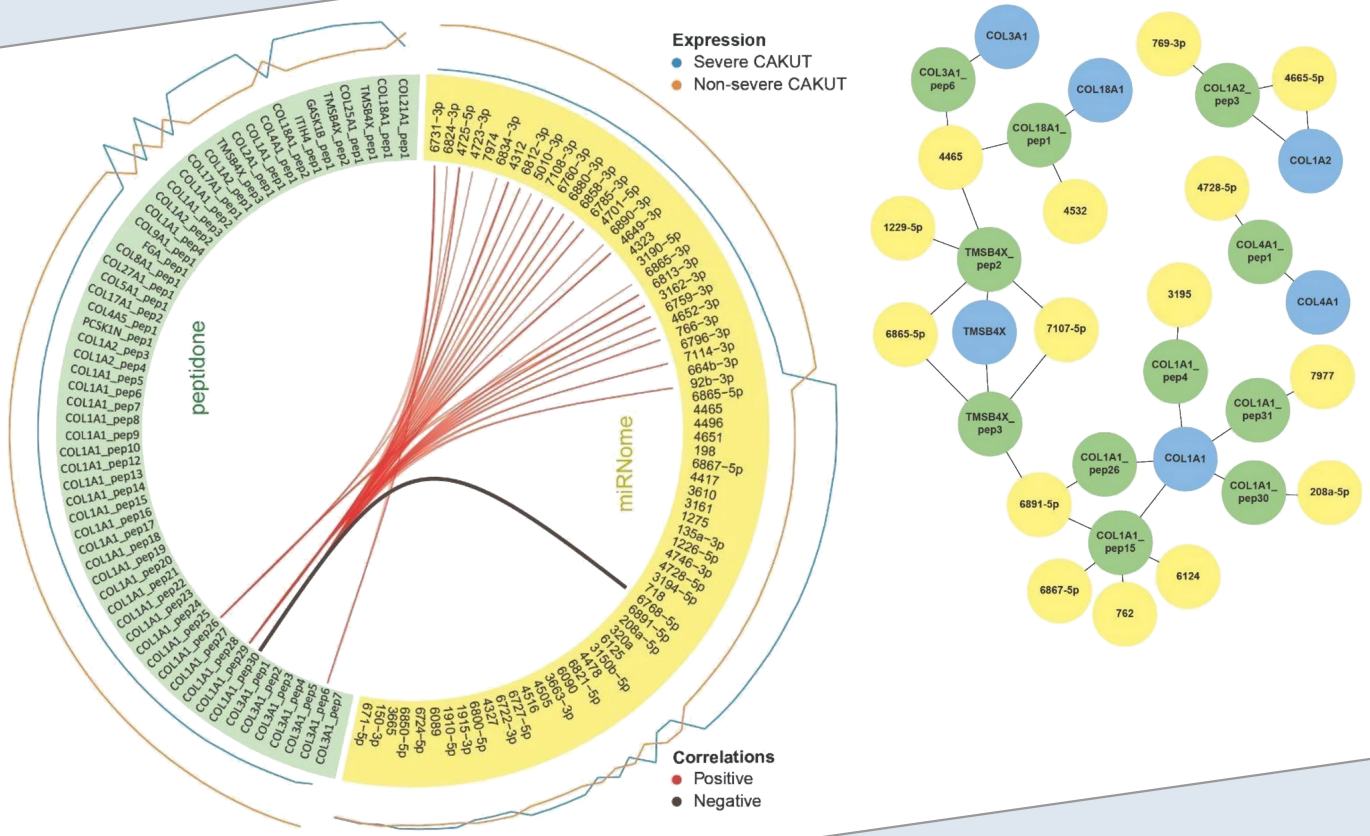
- mixOmics
- Number of CAKUT samples = 46
  - Non-severe cases: 30
  - Severe cases: 16
  - Only Peptidome and miRNome



Bayjanov *et al. Nature Scientific Reports.* 2024.

CAKUT – Case Study

- MixOmics
  - More than 80% correlation
  - Collagens showed high correlation
  - Most differentially expressed genes and peptides showed high correlation

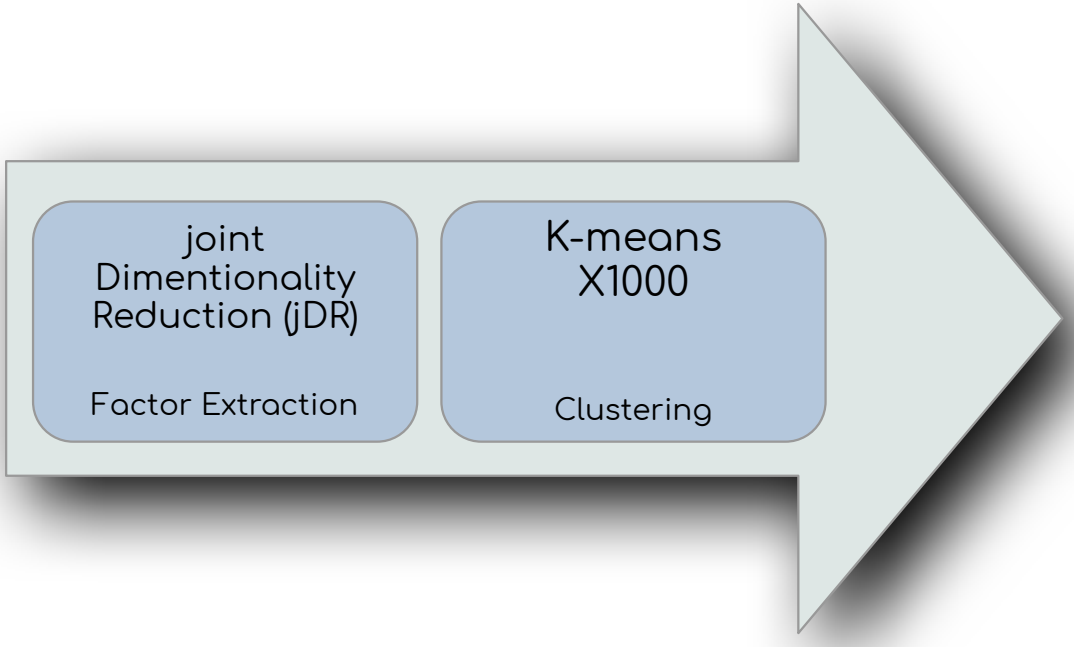


Bayjanov *et al. Nature Scientific Reports.* 2024.

CAKUT – Case Study

2nd Method Workflow

- Momix notebook
- jDR (joint Dimentionality Reduction)
  - iCluster
  - intNMF
  - JIVE
  - MCIA
  - MOFA
  - RGCCA
  - Scikit-fusion
  - tICA
- Number of CAKUT samples = 31
  - Non-severe cases: 18
  - Severe cases: 13



Bayjanov *et al. Nature Scientific Reports.* 2024.

CAKUT – Case Study

- Momix Notebook
  - Various accuracy rates
    - Ranging from 90% to 56%
  - Biological relevance
  - Collagens

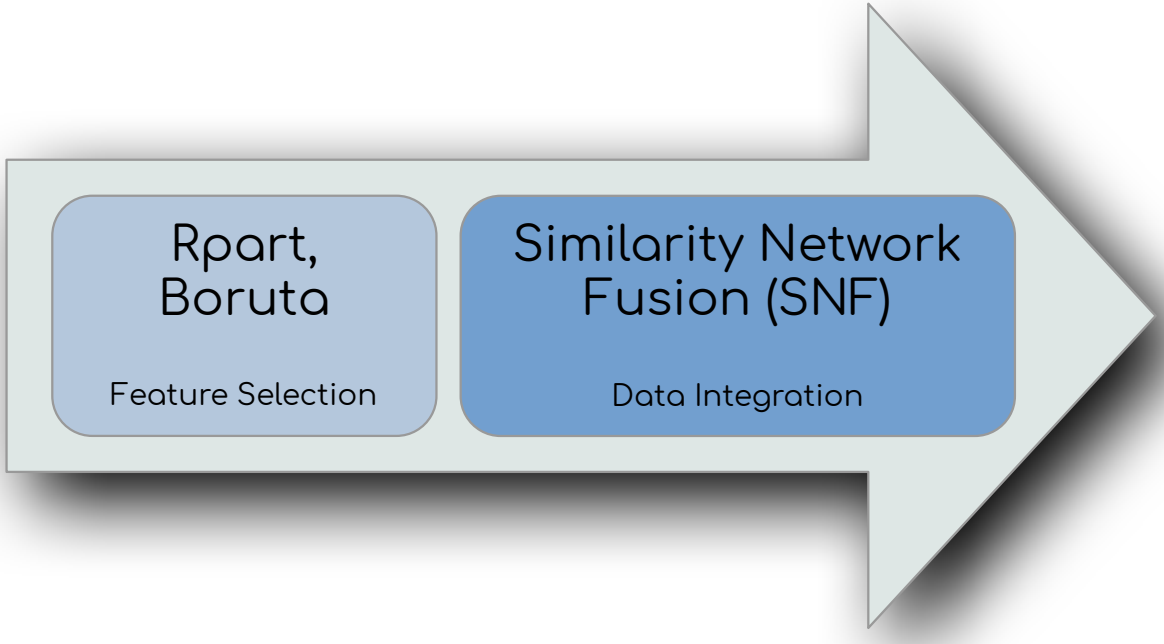
Method	k-means 1st factor (%)	k-means 2nd factor (%)
RGCCA	**90**	58
tICA	65	**87**
MOFA	**87**	54
iCluster	**84**	62
intNMF	74	**81**
JIVE	**77**	74
MCIA	**77**	61
scikit-fusion	56	**65**

Bayjanov *et al. Nature Scientific Reports.* 2024.

CAKUT – Case Study

3rd Method Workflow

- Number of CAKUT samples = 31
  - Non-severe cases: 18
  - Severe cases: 13
- Feature selection with rpart
- Omic data integration with SNF



CAKUT – Case Study

Feature Selection

- Performed on each dataset separately by rpart.
- rpart algorithm uses recursive partitioning that creates branches based on the variables that separates the data into homogeneous groups with respect to outcome variable.

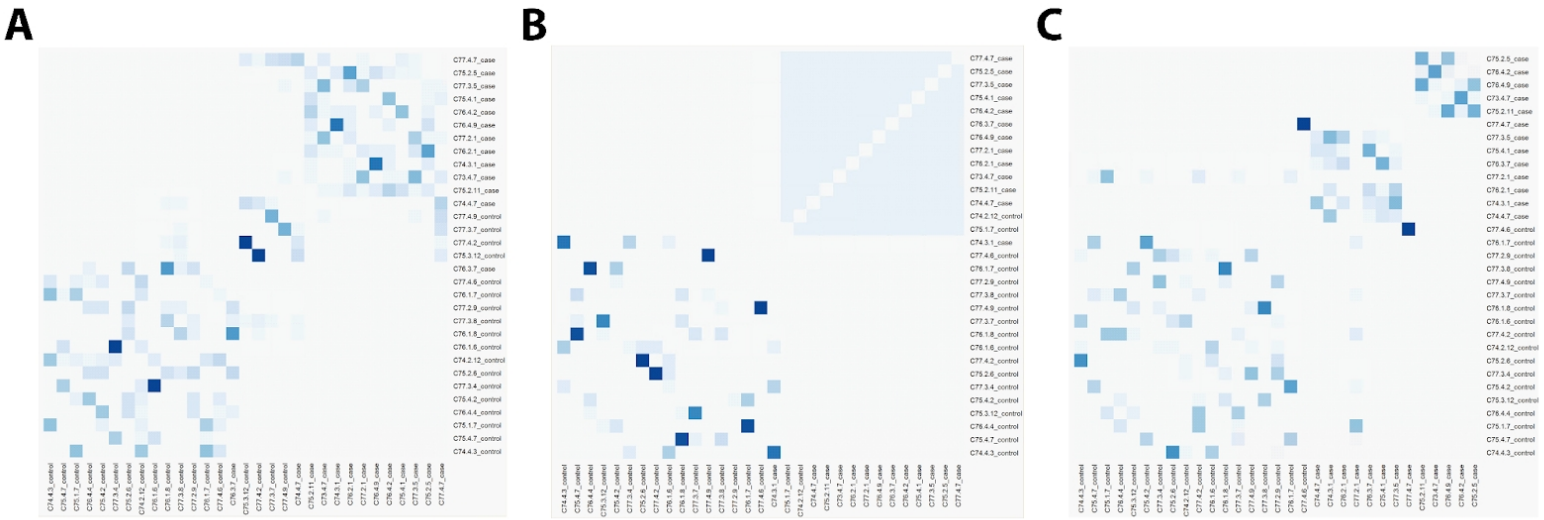
SNF application

- Pairwise correlations calculated and networks were created, where each node represented each sample, and edges pairwise similarities.
- Iterative learning and merging

Type	Name	Importance
miRNA	hsa-miR-1227-3p	98.75
	hsa-miR-129-1-3p	98.00
	hsa-miR-33b-3p	98.1
	hsa-miR-6737-3p	93.1
	hsa-miR-6892-3p	100.00
Proteome	P55287.Q9ULB5.P55289.Q86UP0	100.00
	Q6UVK1	100.00
	Q8WZ75	100.00
Peptide	11868	100.00

Wang *et al*. *Nature Methods*. 2014.

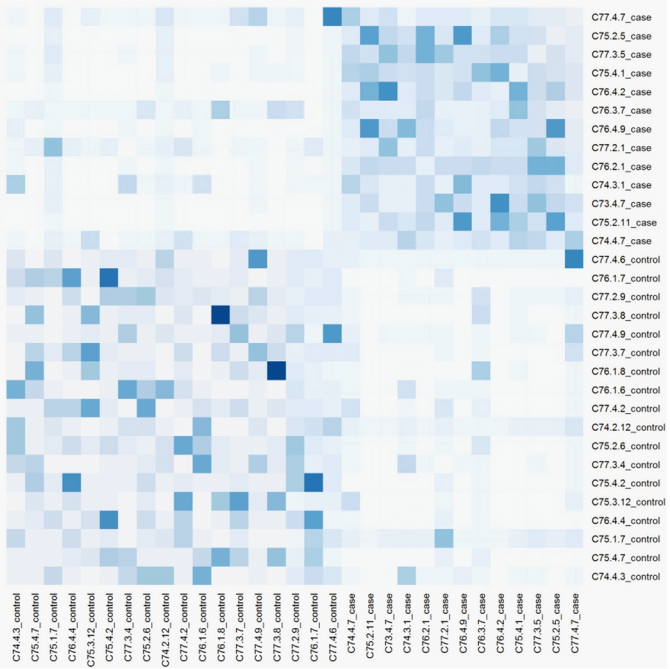
CAKUT – Case Study



miRNome Peptidome Proteome

CAKUT – Case Study

Data Type	Accuracy	Sensitivity	Specificity
miRNome	0.8387	0.9231	0.7778
Proteome	0.7419	0.3846	1
Peptidome	0.9032	0.9231	0.8889
Proteome-miRNome	0.9677	0.9231	1
Proteome-Peptidome	0.2903	0	0.5
MiRNome-Peptidome	0.9677	0.9231	1
Proteome-miRNome-Peptidome	**1**	**1**	**1**



Colorectal Cancer – Case Study

- **Colorectal cancer (CRC)** is the third most common cancer type worldwide, with over 1.9 million new cases diagnosed only in 2020.
- Key risk factors include age, diet, obesity, and family history of the disease.
- Five-year survival rates for colorectal cancer can vary significantly, depending on the stage at diagnosis and access to treatment.

Colorectal Cancer – Case Study

- CRC patients: 34
  - subtype1: 6
  - Subtype2: 13
  - Subtype3: 9
  - Unclassified: 5
- Transcriptome:
  - RNAseq with Illumina HiSeq
  - 125 bp
- Metagenome:
  - 16S with Illumina MiSeq
  - V3-V4 ~250 bp

Transcriptomics analysis

- Fastq filtering
- SolexaQA++ trimming
- STAR for mapping to GRCh38
- Samtools for sam processing
- DeSeq2 and PathfindeR for DE and pathway analysis from KEGG

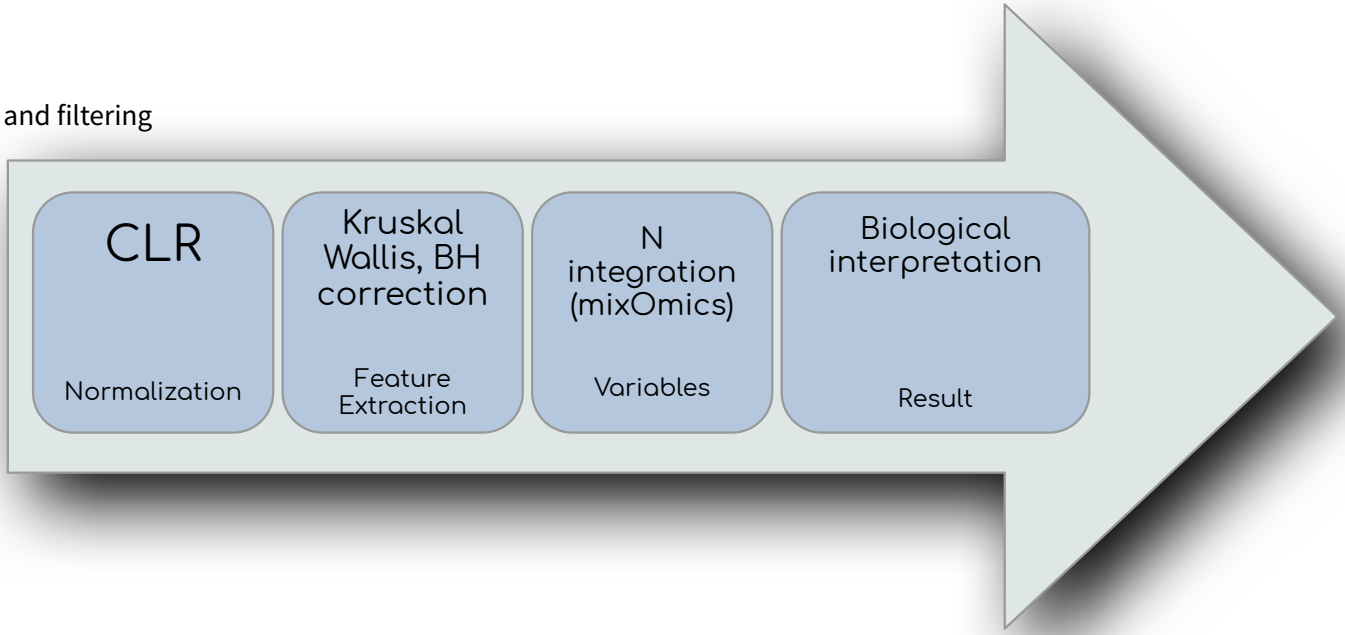
Metagenome analysis

- QIIME2 pipeline
- Picrust2 for pathway abundance values from KEGG

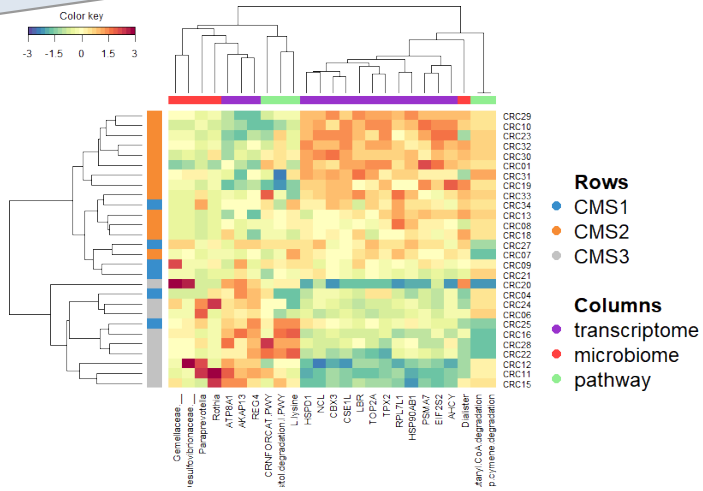
Colorectal Cancer – Case Study

Method Workflow

- CLR normalization (Central Log Transformation) and filtering
- Feature extraction with hypothesis tests
  - P-value correction (BH)
  - Only phylum level of microbiome data
  - Removal of unclassified samples
- sPLS-DA
- MixOmics N-integration



Colorectal Cancer – Case Study

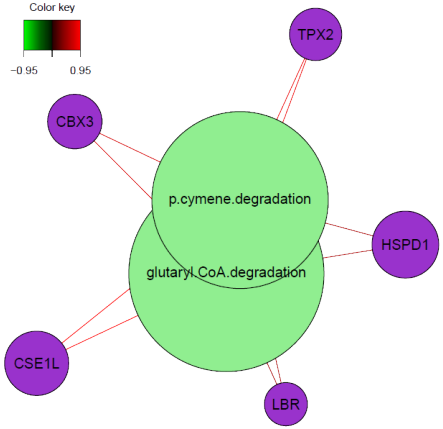


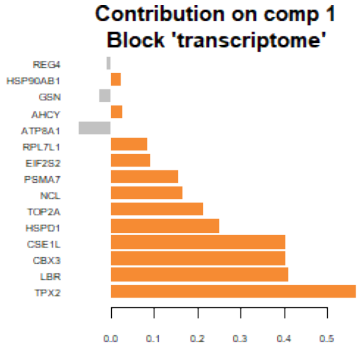
Classification with sPLS-DA

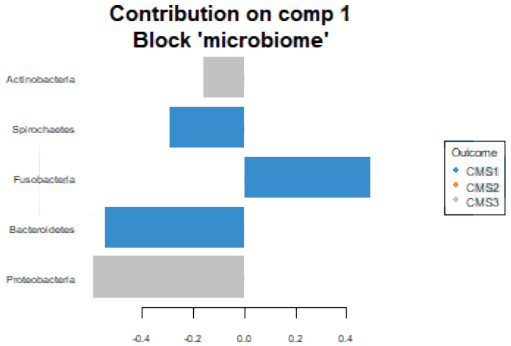
- 87% before preprocessing 100% prediction after preprocessing

N-Integration

- Biological parameter detection
- Affected pathways were detected
  - P.cymene and Glutaryl CoA known to have antioxidant effects.







Sezerman lab

if possible...

Crew!

TRANSPAN

COST ACTION CA21116

Crew!

Thank you...

any questions?