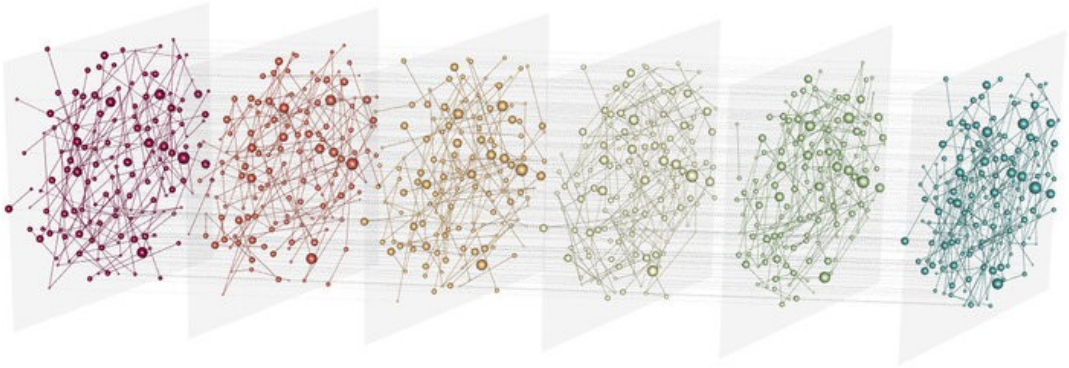
Integrative Genomics: Introduction and application to metagenomics



2024/06

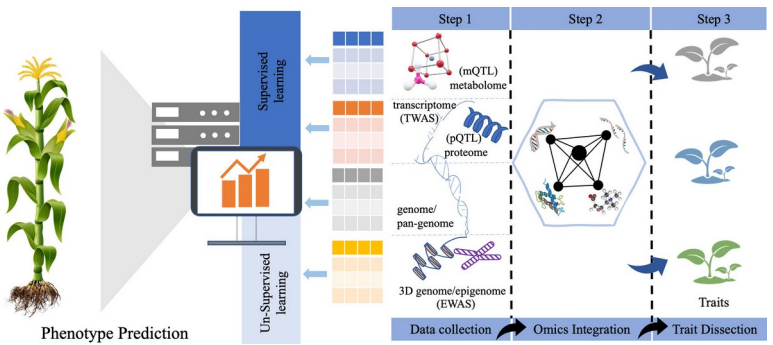
Vincent Manzanilla
UMR Quito

Institut de Recherche pour le Développement FRANCE
French National Research Institute for Sustainable Development

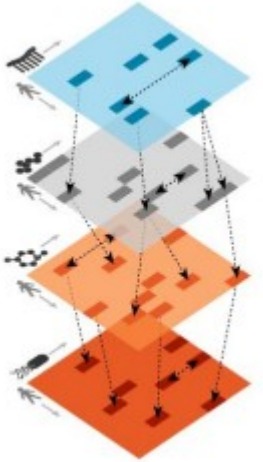
intertryp
Host-vector-parasite-environment interactions in the Neglected Tropical Diseases due to Trypanosomatids

Integration: biological motivation

- Deep insights into complex biological phenomenon
- Subtyping and classification (disease, species, varieties)
- Prediction, diagnostic, identify drivers, selection...



Mahmood et al (2022). Multi-omics revolution to promote plant breeding efficiency. Front Plant Sci.

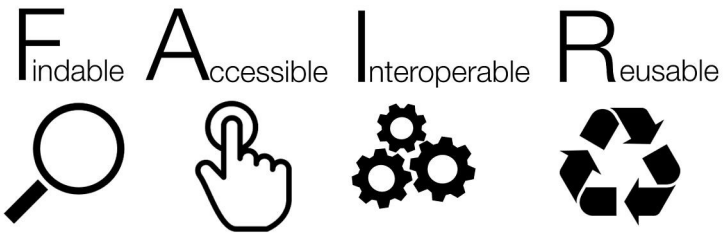


Integration: motivation and challenges

- Integrative genomics is about aggregating heterogeneous data in a statistical fashion.

- **Take advantage of the vast amount of available data**
  - Data access (local/national regulation, infrastructures...)
  - Data representation (structuration, ontologies...)

    -> Need of common representation framework

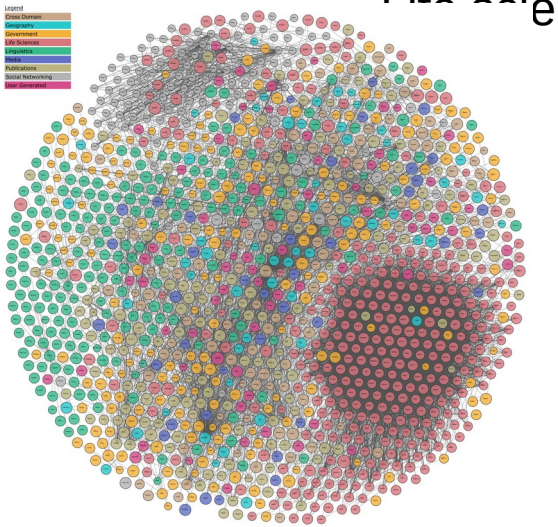


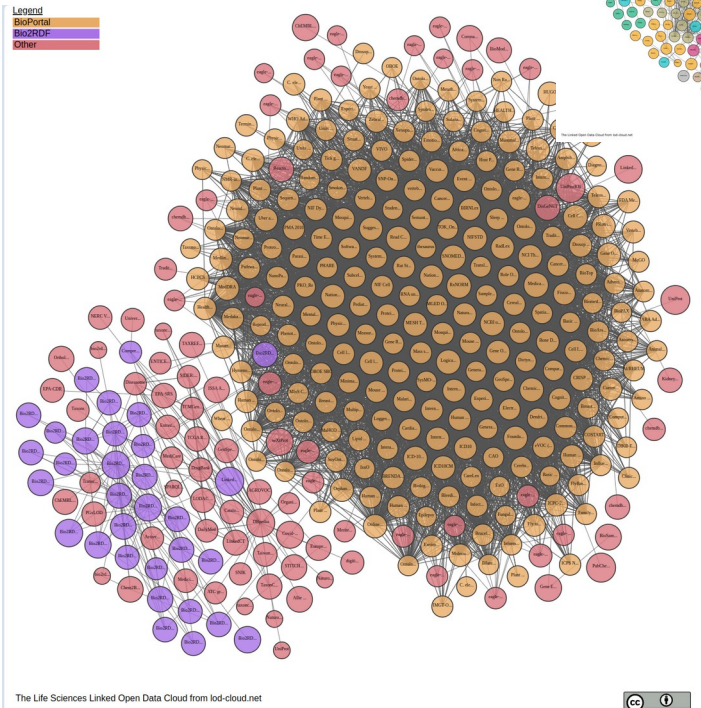
- **Improve our understanding of biological phenomena**
  - Data heterogeneity (technology, format, biological meaning, stat. distribution...)
  - Data complexity (dependances/independances, ad-hoc assumptions...)
  - Amount a data (time/memory consuming)

-> Need of new analysis methods/algorithms

Integration: how?

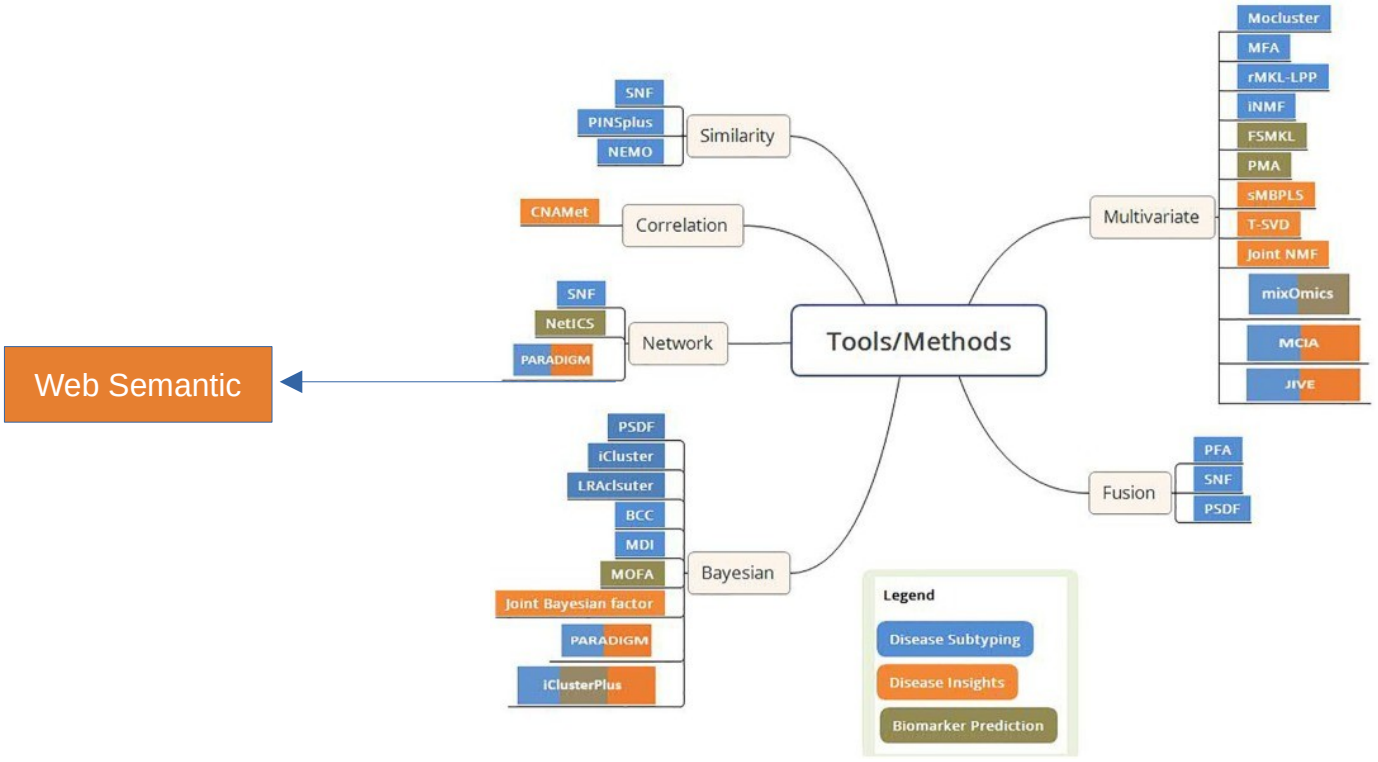
- Improve our understanding of biological phenomena
- Take advantage of the vast amount of available data
- Semantic Web = framework for:
  - integrating data and knowledge
  - querying
  - reasoning





The Linked Open Data Cloud
https://lod-cloud.net

Integration approaches

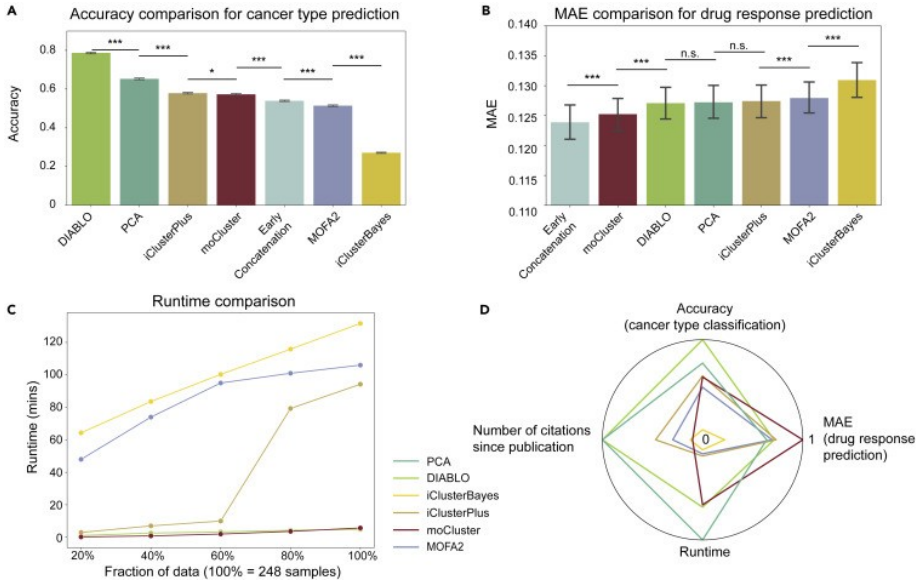


Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020

Integration approaches : the good one ?

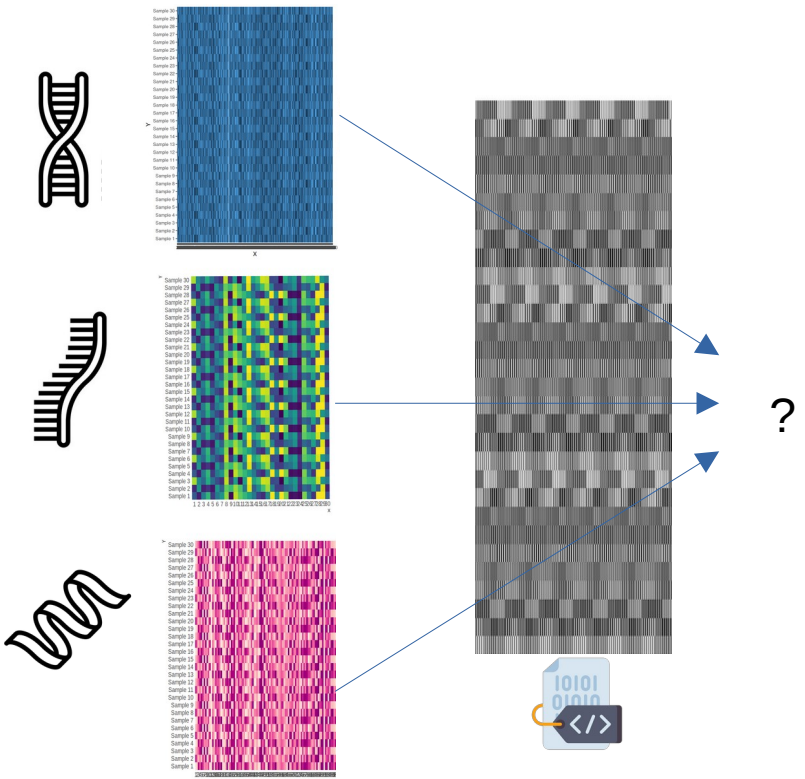
- Integration methods are not unique:
  - comparisons exist... for a given application
  - parametrization need expertise
- Integration methods are not magic!
  - check design and confounding factors
  - perform specific data pre-processing for each omic
  - impute missing values* (different meaning → different strategy)
  - choose the integration strategy based on specific objectives and the data

    (ex. matching between omics) → still no standard pipelines



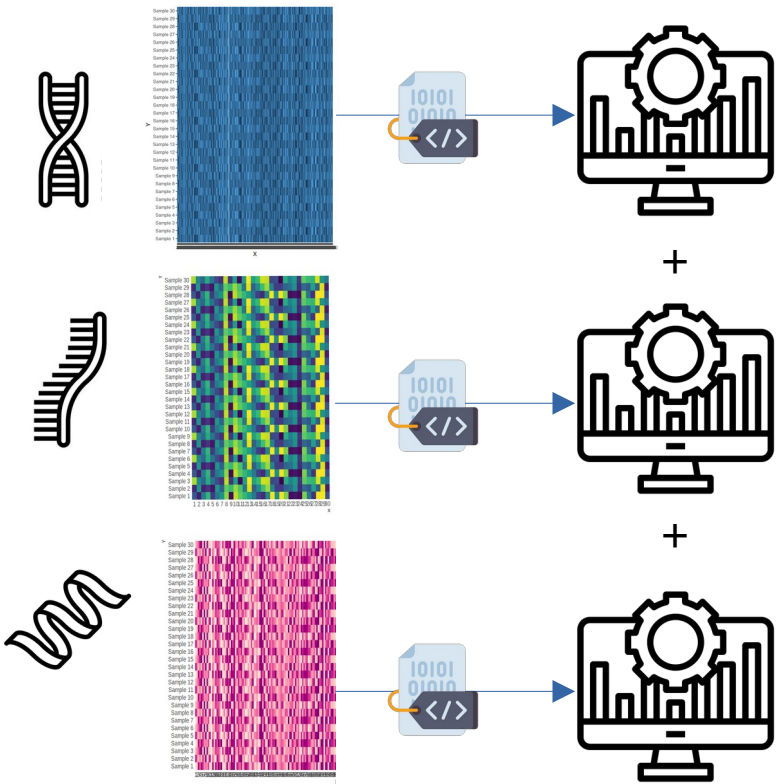
Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022

Integration strategies



Inspired from: Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies - Late

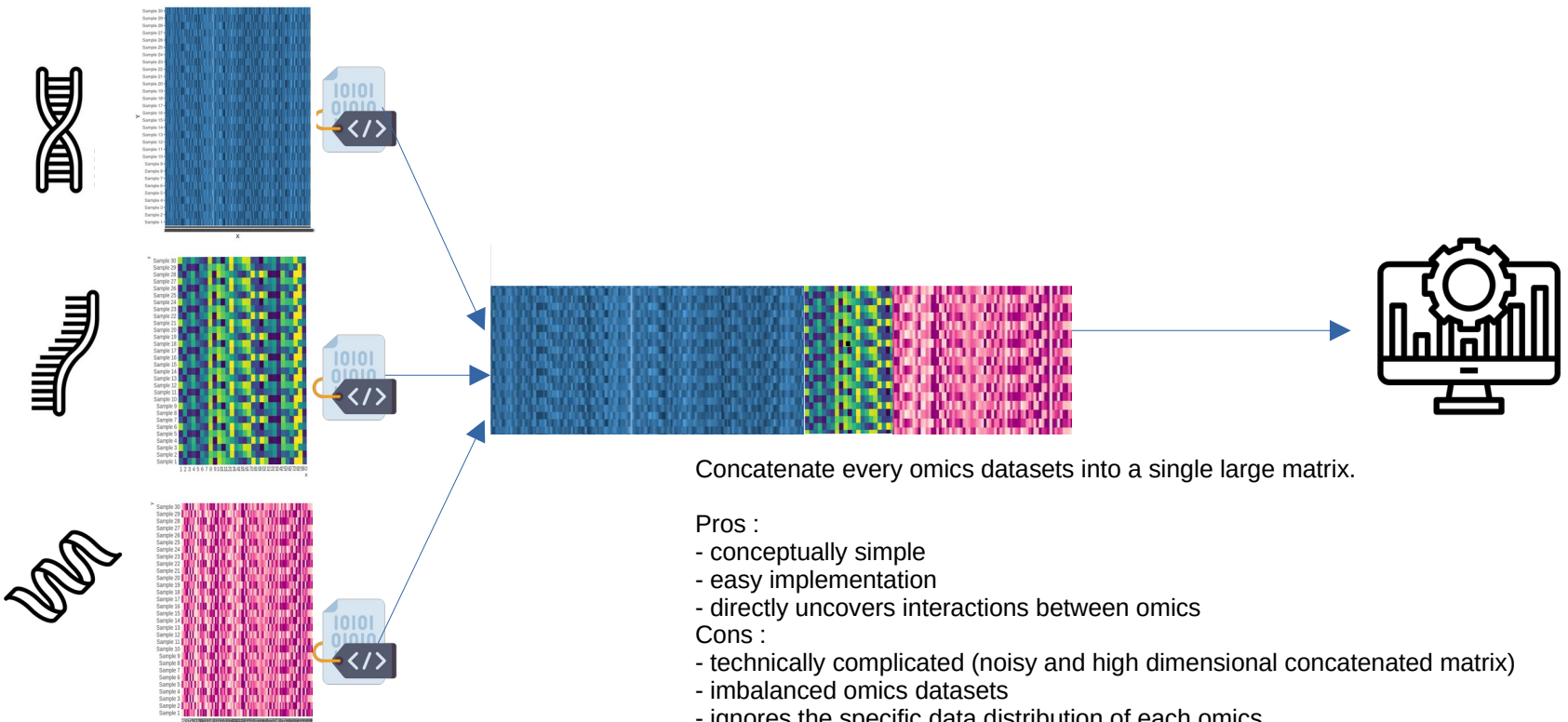


Pros :
- avoid (numerous) challenges of direct omics integration
- use tools designed specifically for each omics
- classical approaches can be used to combine results

Cons :
- cannot capture inter-omics interactions
- complementarity information between omics is not exploited

Inspired from: Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies - Early



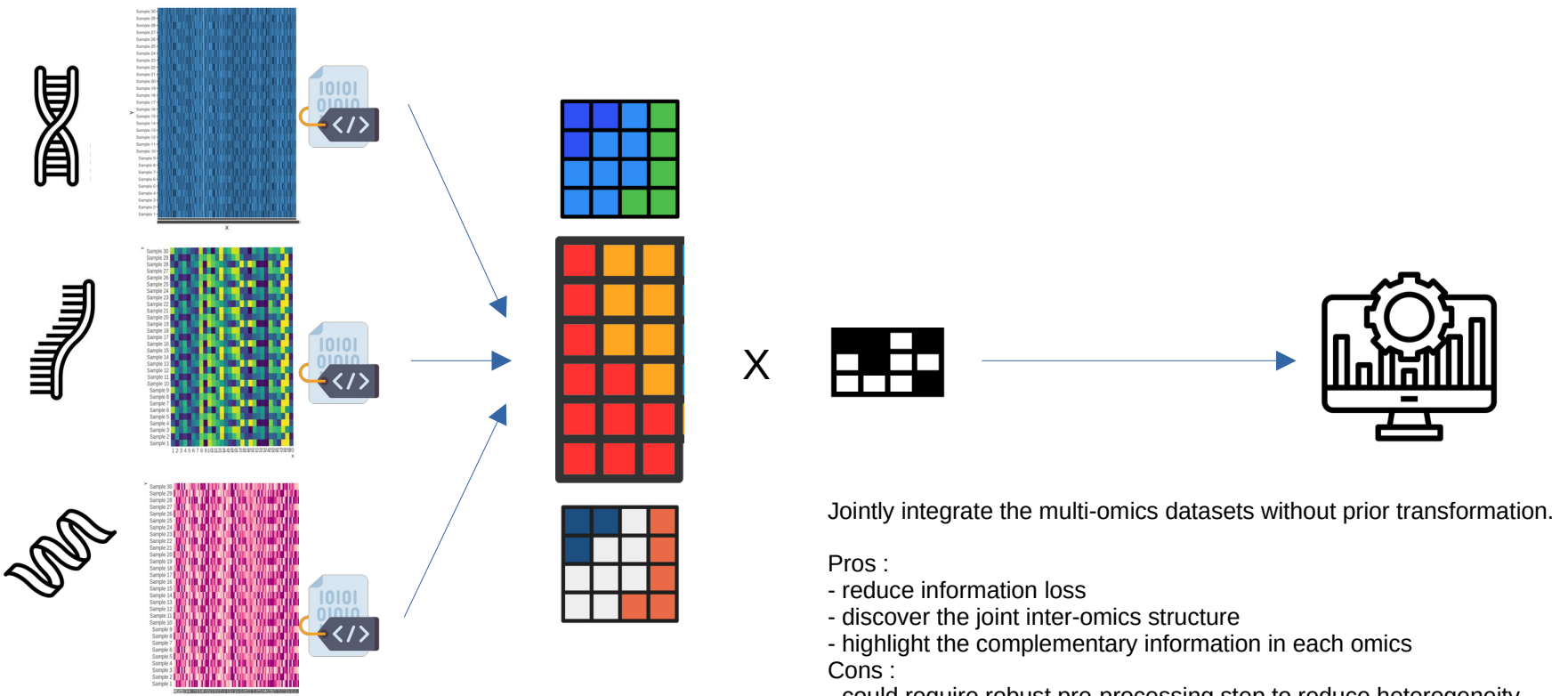
Concatenate every omics datasets into a single large matrix.

Pros :
- conceptually simple
- easy implementation
- directly uncovers interactions between omics

Cons :
- technically complicated (noisy and high dimensional concatenated matrix)
- imbalanced omics datasets
- ignores the specific data distribution of each omics
- common definition space (rows or columns → samples or features)

Inspired from: Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies - Mixed



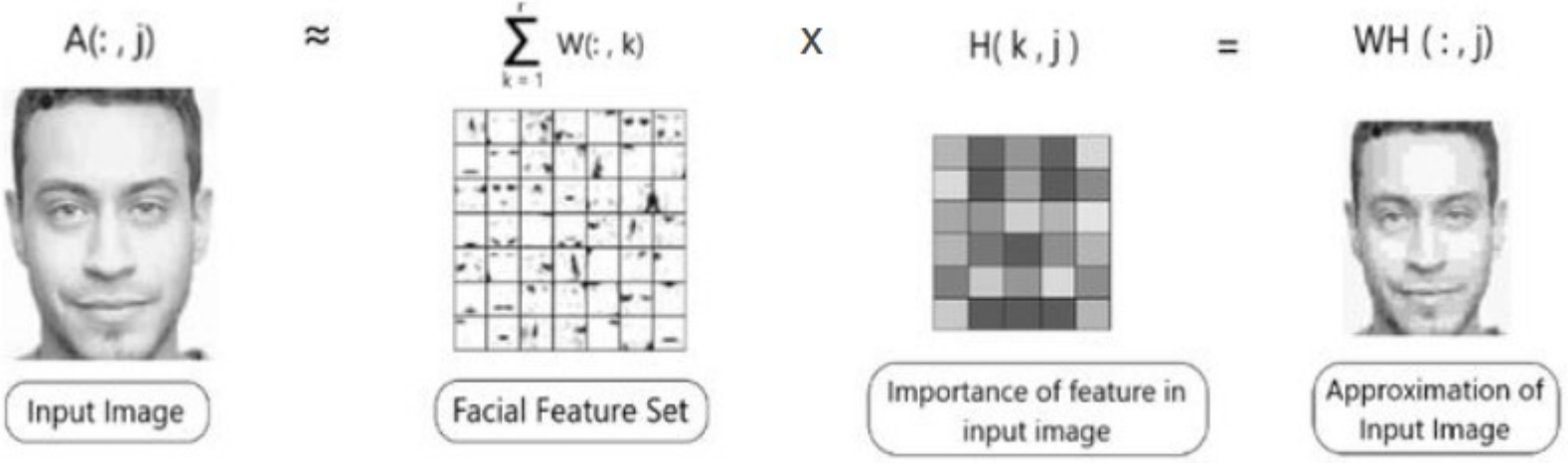
Jointly integrate the multi-omics datasets without prior transformation.

Pros :
- reduce information loss
- discover the joint inter-omics structure
- highlight the complementary information in each omics

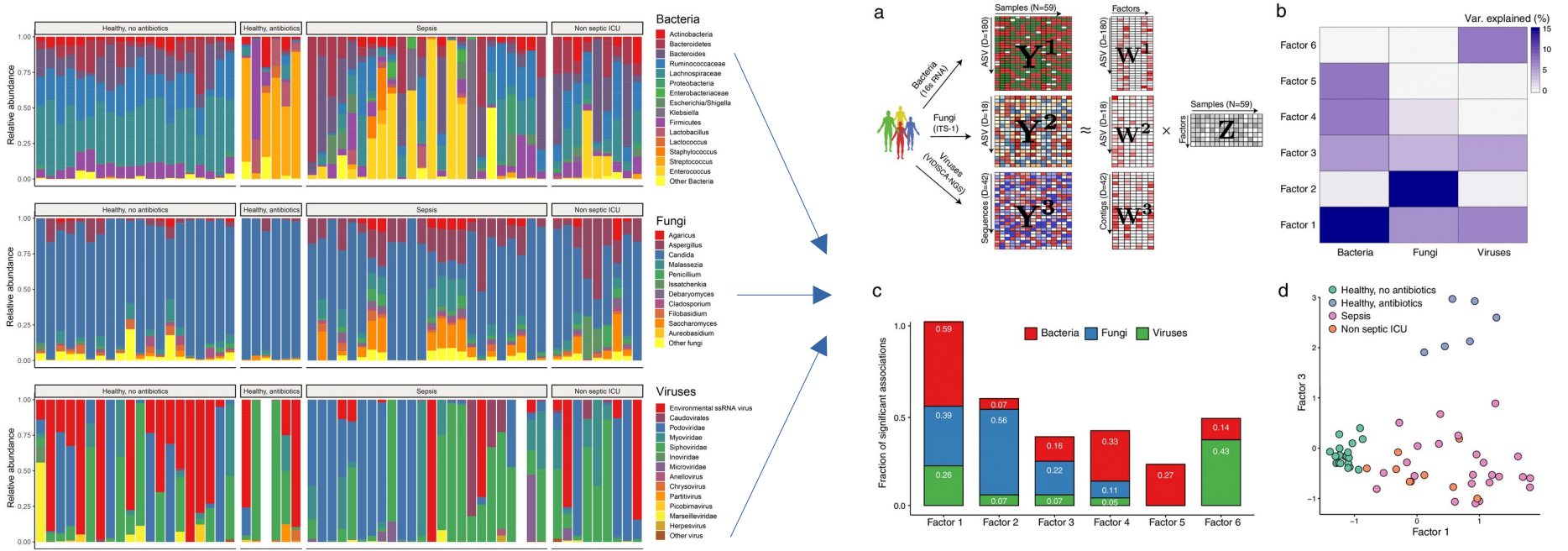
Cons :
- could require robust pre-processing step to reduce heterogeneity
- common latent space assumption

Inspired from: Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies – Mixed: How does that work?

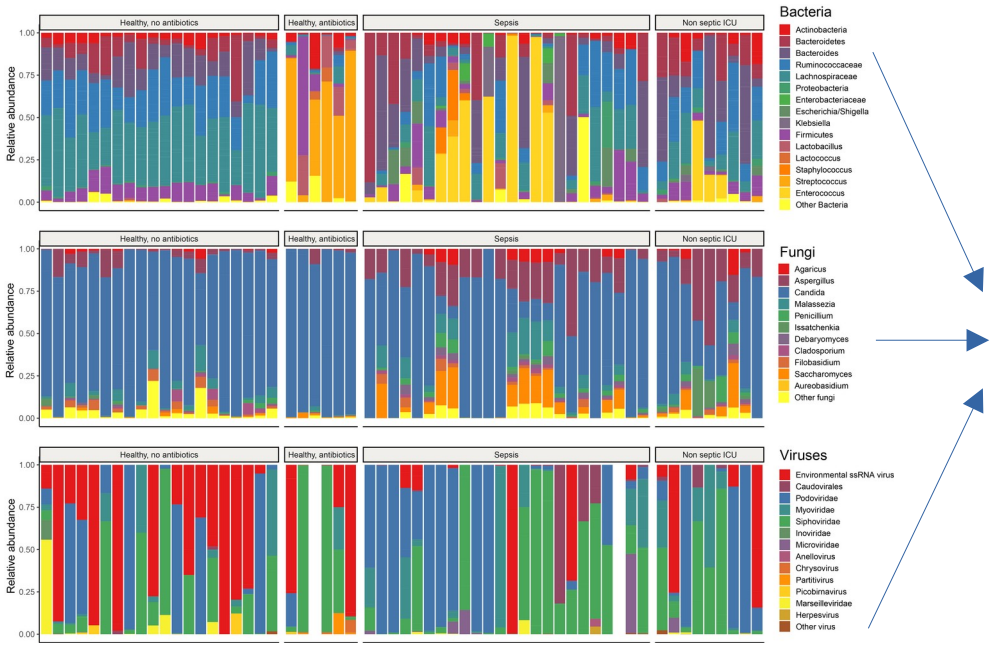


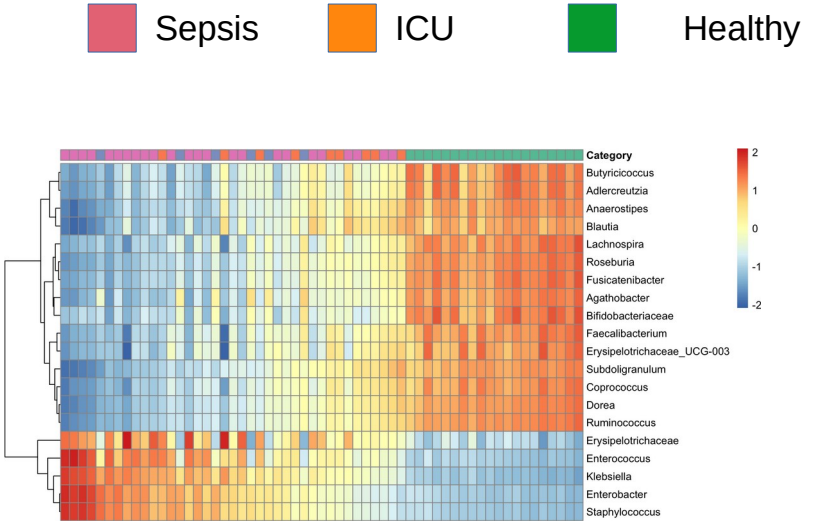
Integration strategies - Mixed



Haak, Bastiaan W., et al. "Intestinal transkingdom analysis on the impact of antibiotic perturbation in health and critical illness." bioRxiv (2020): 2020-06.

Integration strategies - Mixed

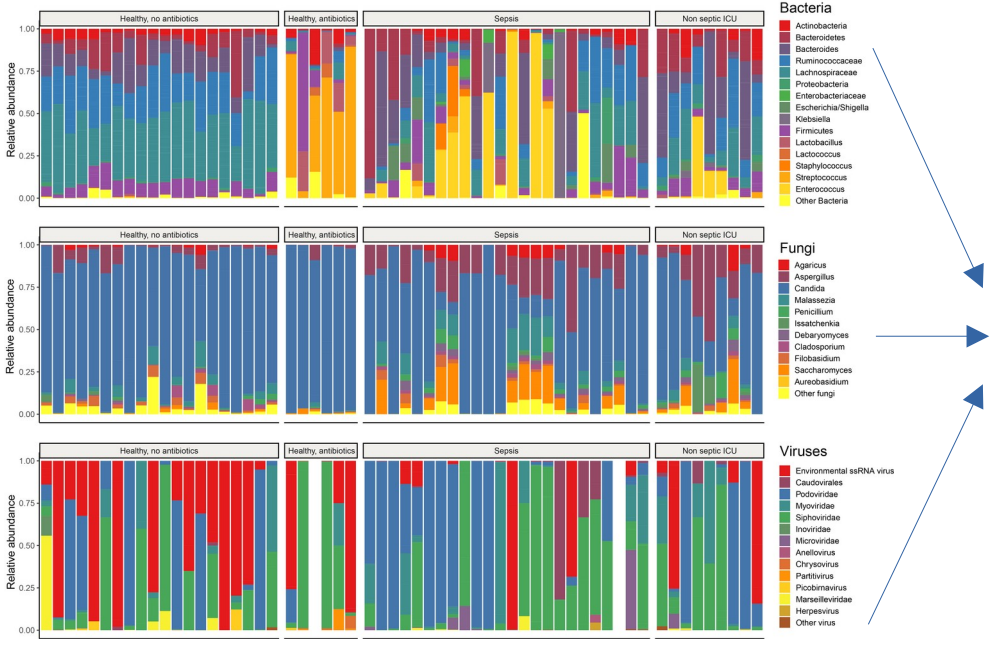


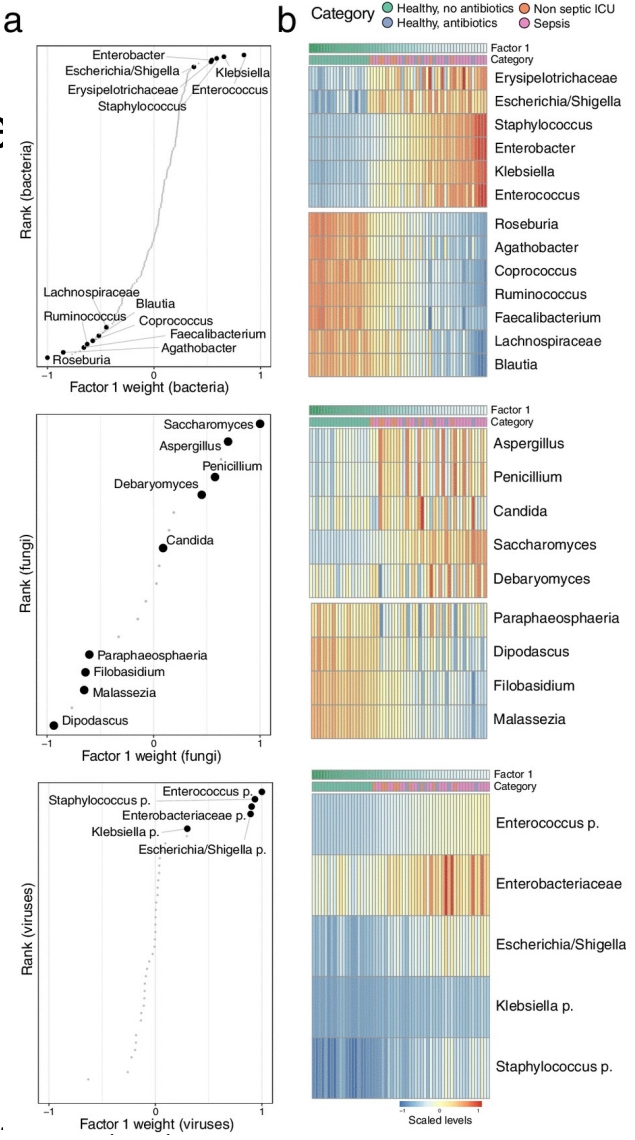


Bacteria
Blue = negative influence on the categories
Red = Positive influence on the categories

Haak, Bastiaan W., et al. "Intestinal transkingdom analysis on the impact of antibiotic perturbation in health and critical illness." bioRxiv (2020): 2020-06.

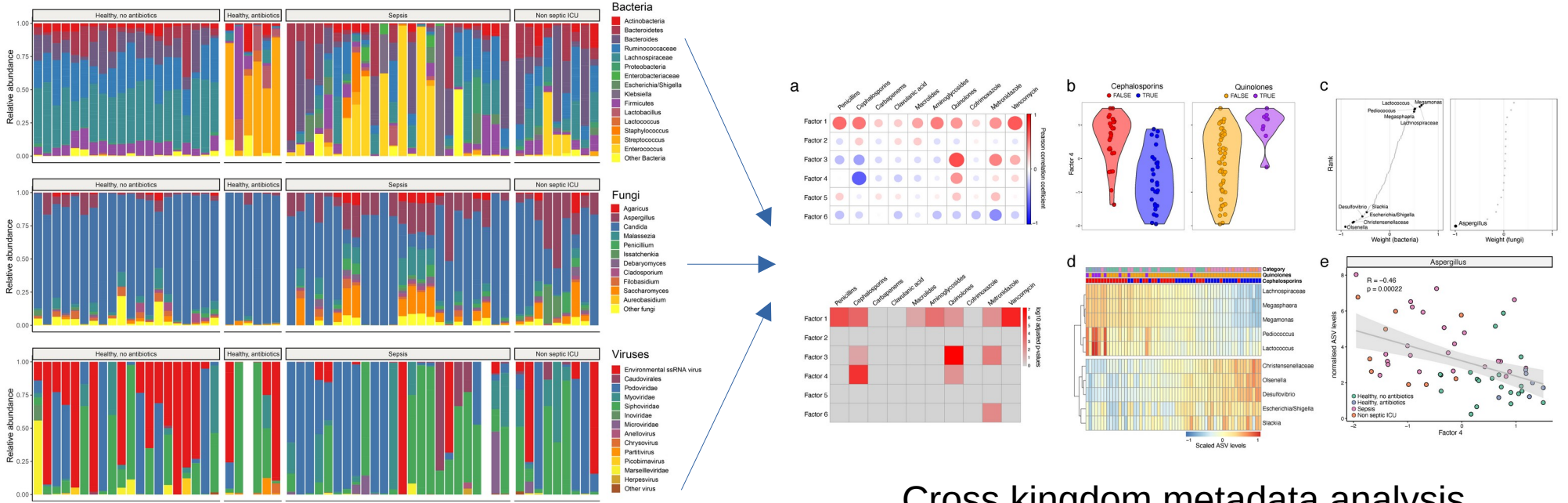
Integration strategies - Mixe





Haak, Bastiaan W., et al. "Intestinal transkingdom analysis on the impact of antibiotic perturbation in health and critical illness.

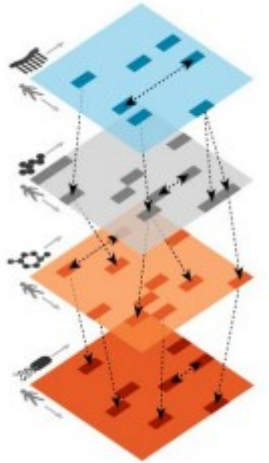
Integration strategies - Mixed

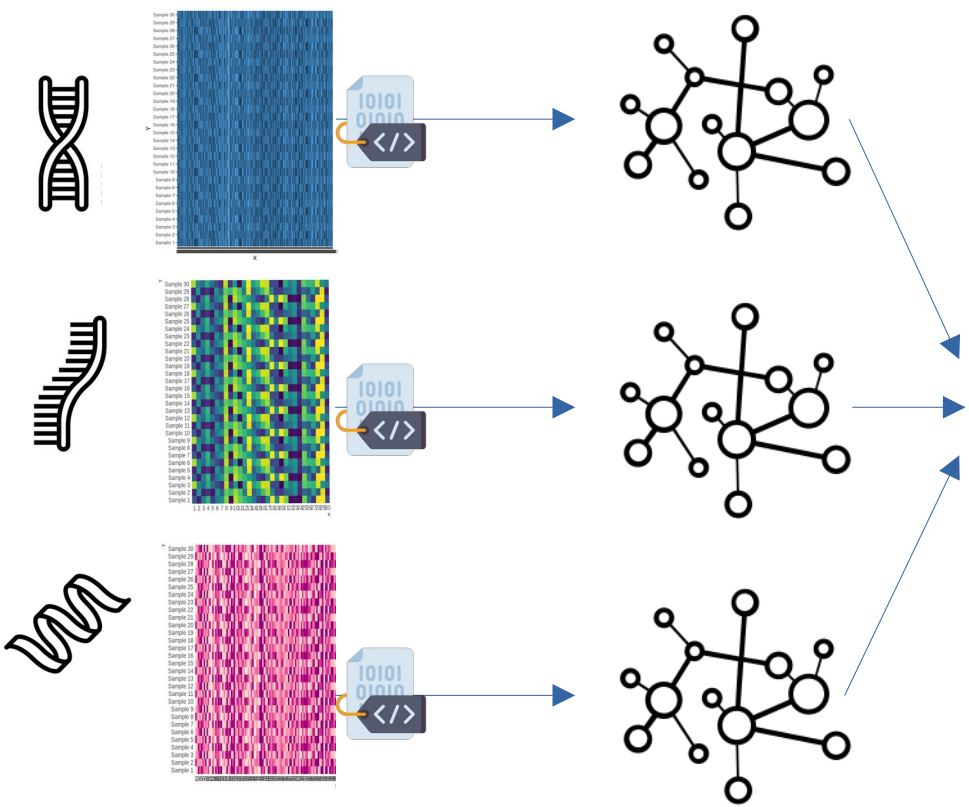


Cross kingdom metadata analysis

Haak, Bastiaan W., et al. "Intestinal transkingdom analysis on the impact of antibiotic perturbation in health and critical illness." bioRxiv (2020): 2020-06.

Integration strategies - Mixed





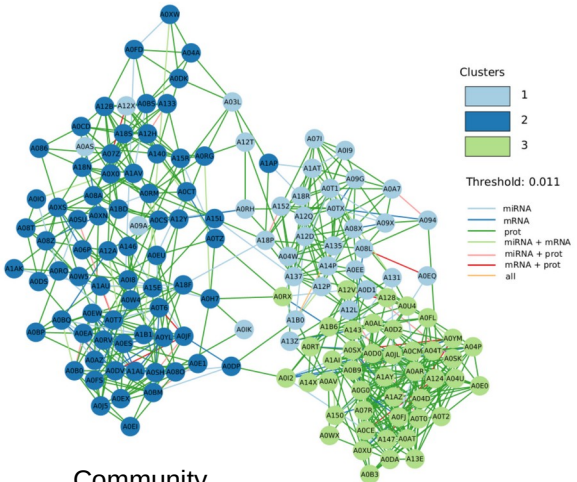
Transform independently each omics dataset into a simpler representation before integration.

Pros :
- new representation is less dimensional and less noisy
- less heterogeneity between omics
- classical approaches can be used on combined representation

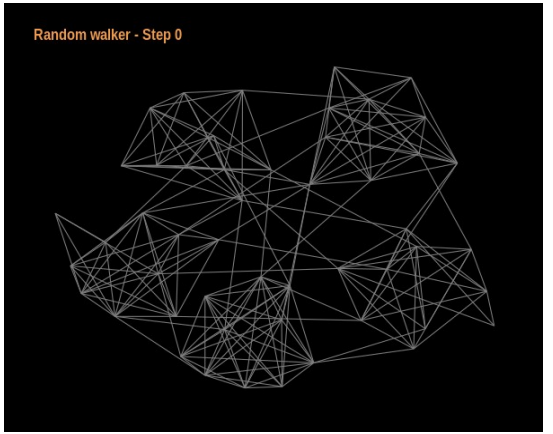
Cons :
- choice of the transformation method is not trivial
- information loss during transformation
- correspondence between omics in the new representation space

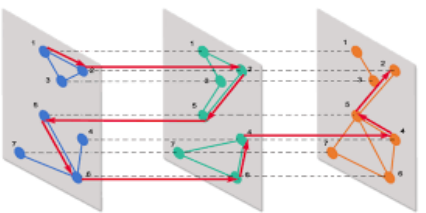
Integration strategies - Mixed

Breast cancer from The Cancer Genome Atlas
(3 modalities)



Community
detection





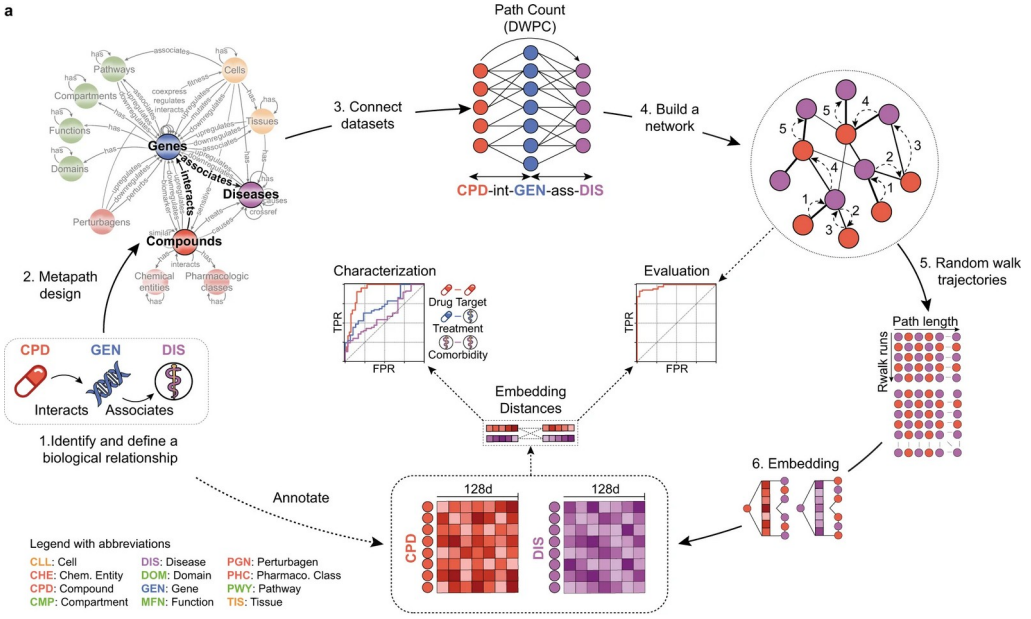
Semantic web / knowledge graph - applications

Scheme of the methodology. First, we define the biological entities (genes/diseases/compounds) to be connected and the specific context to be explored.

Then a source-target network is derived by traversing all the paths available from the source to the target nodes of a given metapath.

The vicinity of each node in the network is then explored by a random walker.

Finally, embeddings are evaluated and characterized.



Fernández-Torras, Adrià, et al. "Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque." Nature Communications 13.1 (2022): 5304.

Take home message

Applications:
Integrative genomics aims to identify patterns, relationships, and interactions between genes, proteins, and other molecular components or organisms.

Design:
Metadata

- document everything for confounding effect
- metadata structure

Design in advance the analysis and the research questions – overview of the downstream analysis.