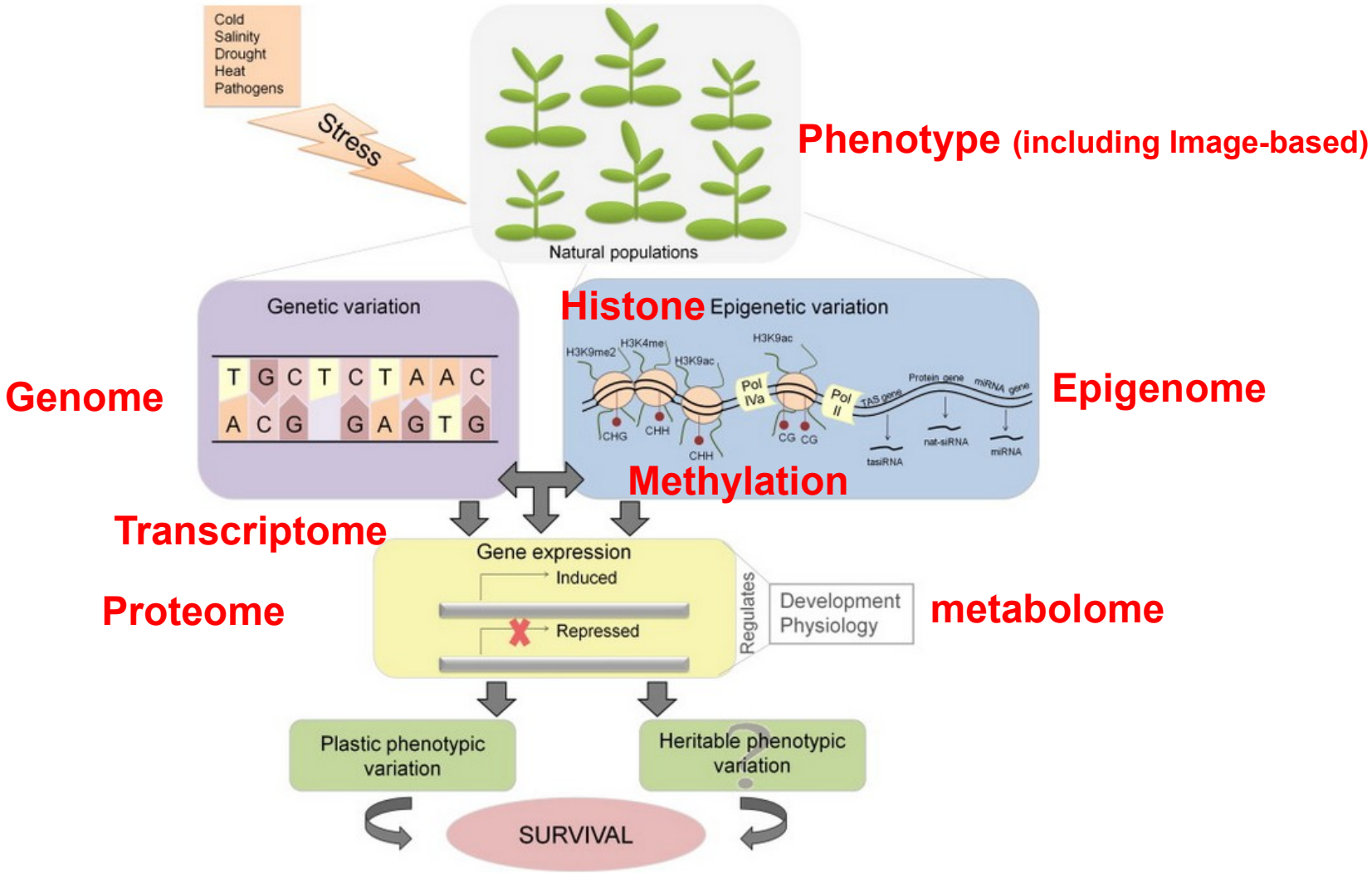
Exploitation des outils statistiques pour l'intégration des données omiques en biologie végétale et animale

Exploiting statistical tools for omics data integration in plant and animal biology

Phd (CIR1-DATA), Emile MARDOC (GDEC/UMRH)
Co-direction: M Bonnet & J Salse

Background – Omics



Toward a System Biology Approach and Tools

Grativol et al. 2012 Bioch Biophys Acta 1819:176-85

Data integration – Two definitions-Concepts

Integration = interconnection

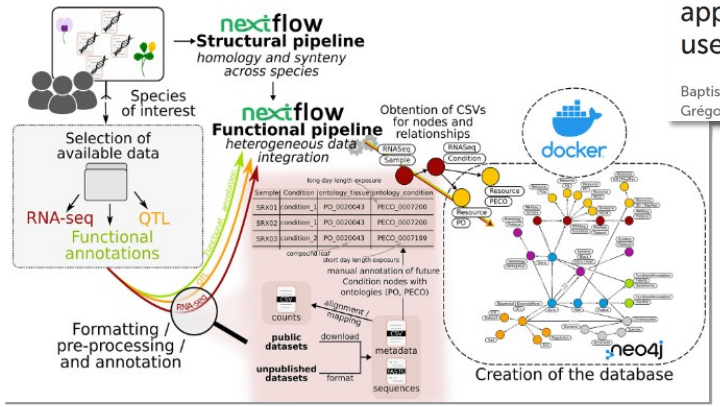
-How to deliver all informations from different databases to address a biological question

-Rely on interoperability

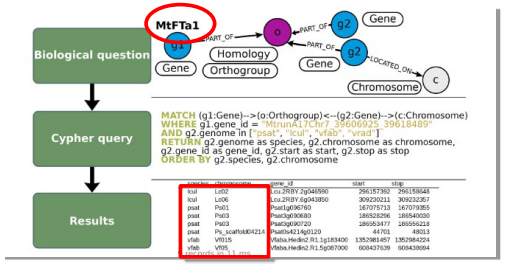
-Several tools and methods available

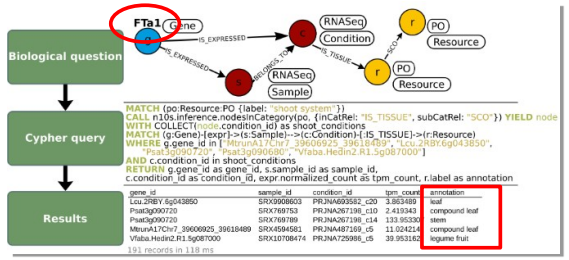
Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes

Baptiste Imbert[1*], Jonathan Kreplak[1], Raphaël-Gauthier Flores[2,3], Grégoire Aubert[1], Judith Burstin[1] and Nadim Tayeh[1*]



OrthoLegKB





Integration = interplay

-Not only interconnection but aiming at delivering novel knowledge

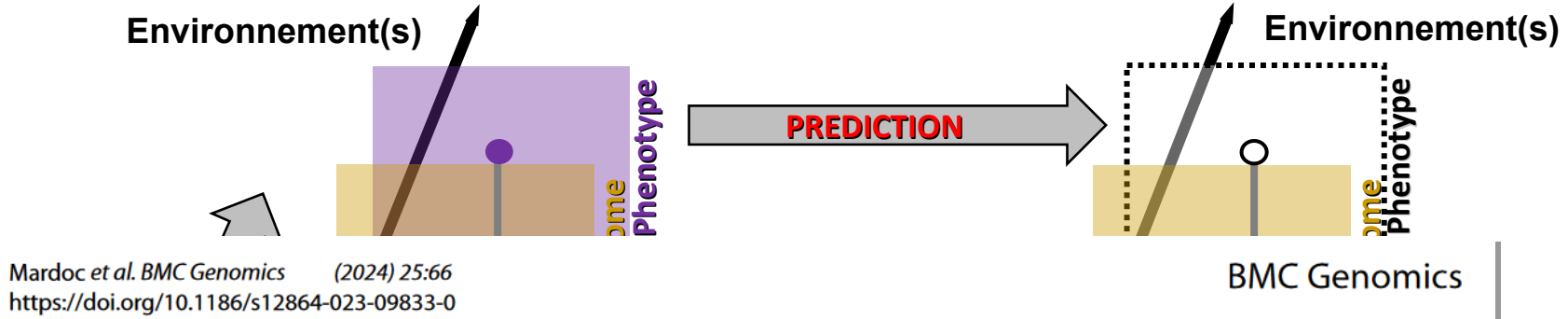
Without *a priori*: Data ✯ Biological process

With *a priori*: Biological process ✯ Data

-Both fundamental and applied science

-Necessary to understand the biology of complex system (biology)

(Gen-)omics data integration – Concepts & Approaches



Mardoc *et al. BMC Genomics* (2024) 25:66
https://doi.org/10.1186/s12864-023-09833-0
BMC Genomics

RESEARCH Open Access

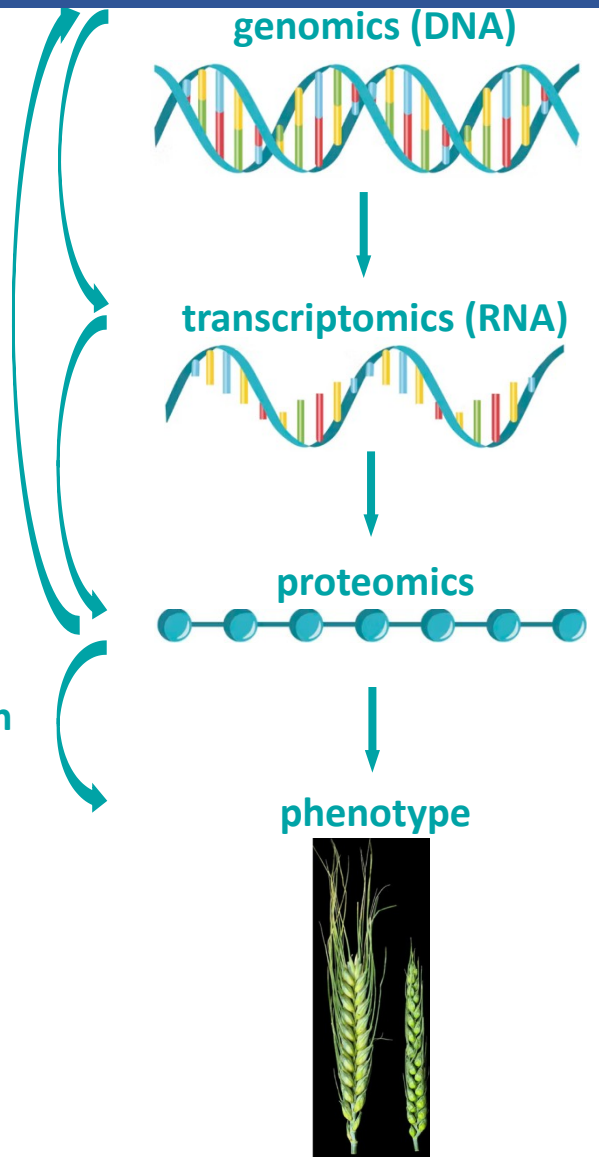
Genomic data integration tutorial, a plant case study

Emile Mardoc[1], Mamadou Dia Sow[1], Sébastien Déjean[2] and Jérôme Salse[1*]

(Gen-)omics data integration – Principles

Omics:
- **heterogeneous data** => difficulty to analyze simultaneously
- **large amount of data** => issues to create, store and analyze data

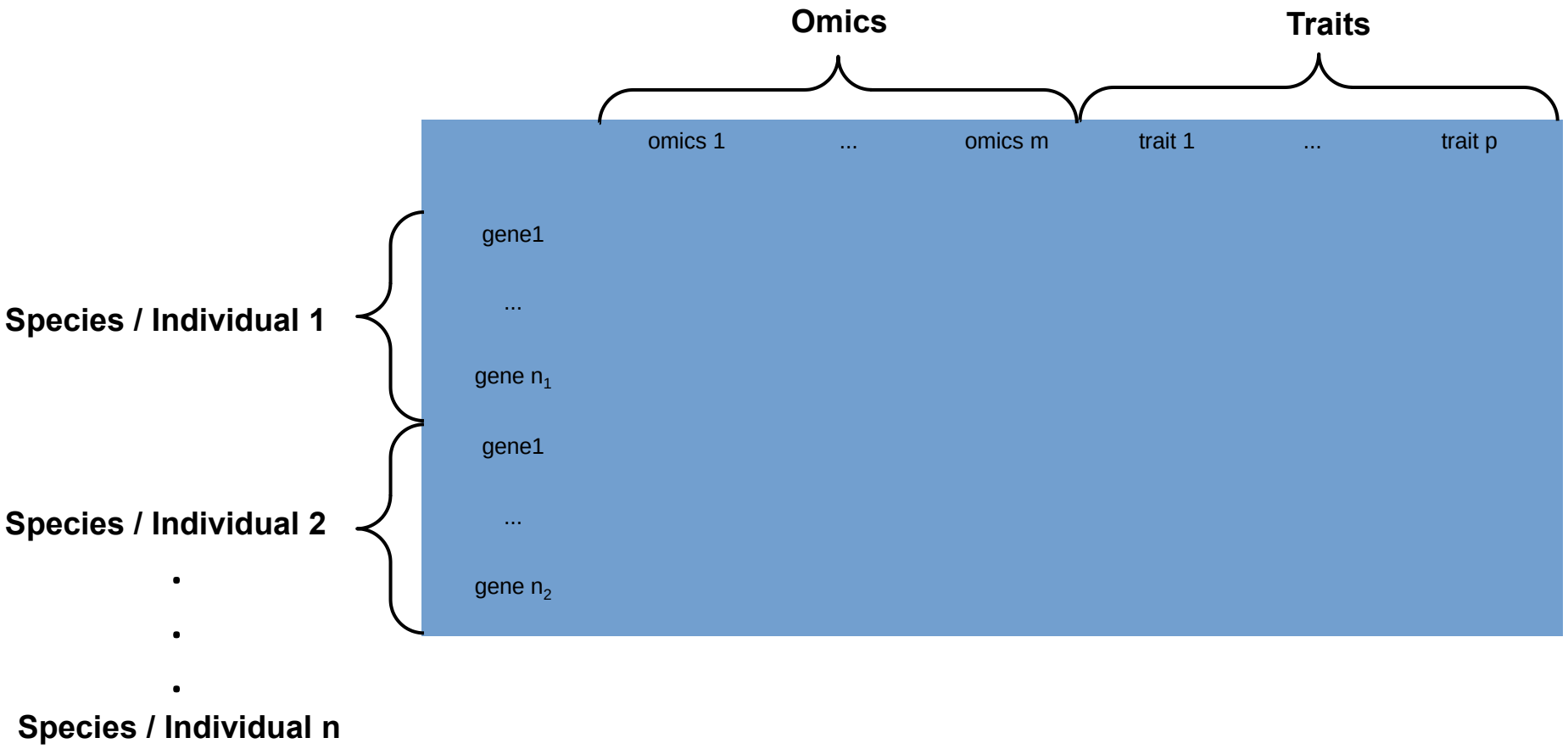
Integration:
- analyze **different data at the same time**, within a single algorithm
- based on different methods: dimension reduction, probabilistic, networks…



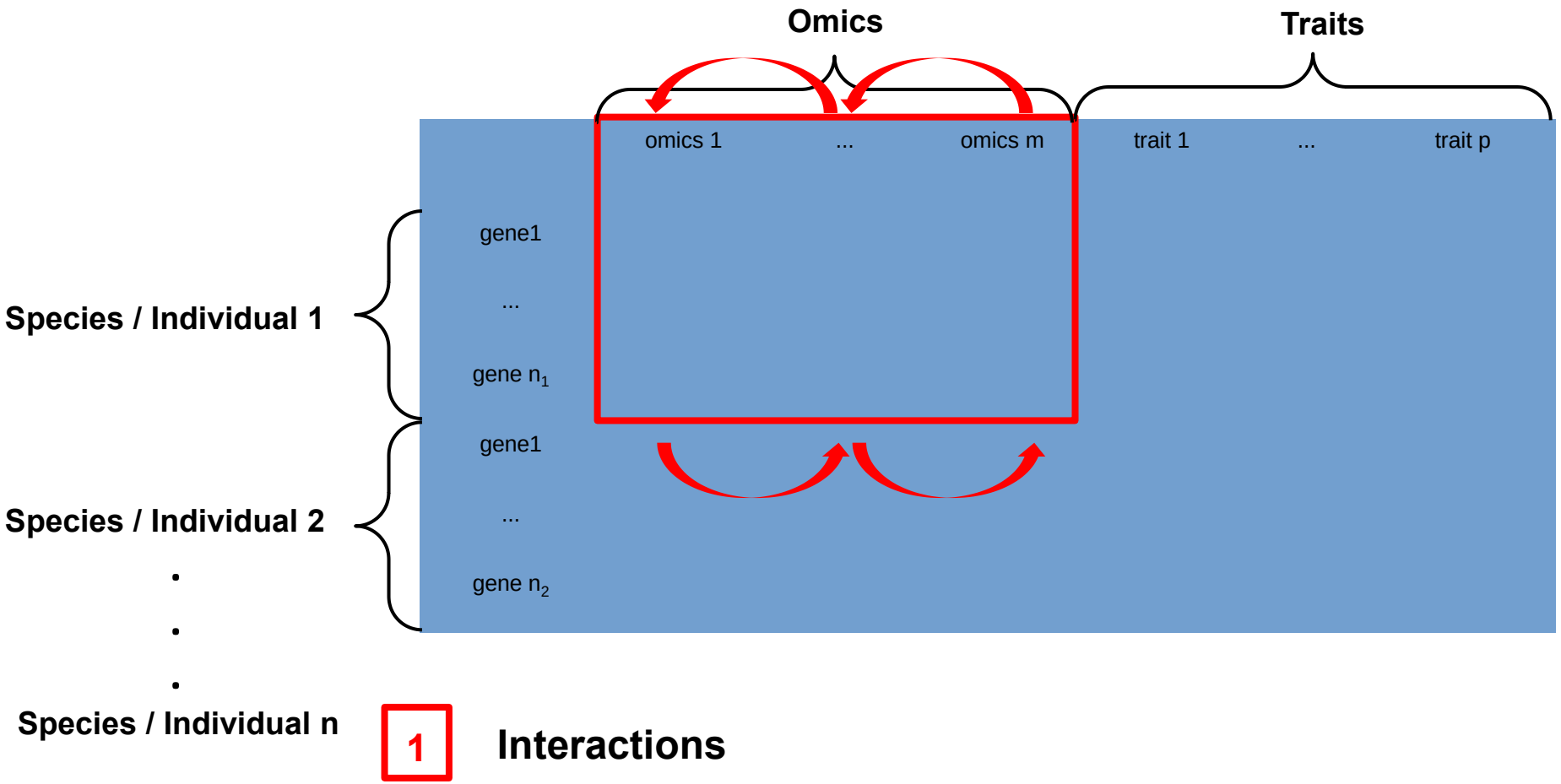
› Biological questions of interest

1. What are the main **interactions-interplay** between (omics)?
2. Which **subsets** are specific across omics for a specific data feature?
3. How to **predict** phenotypes/omics data from other omics data?

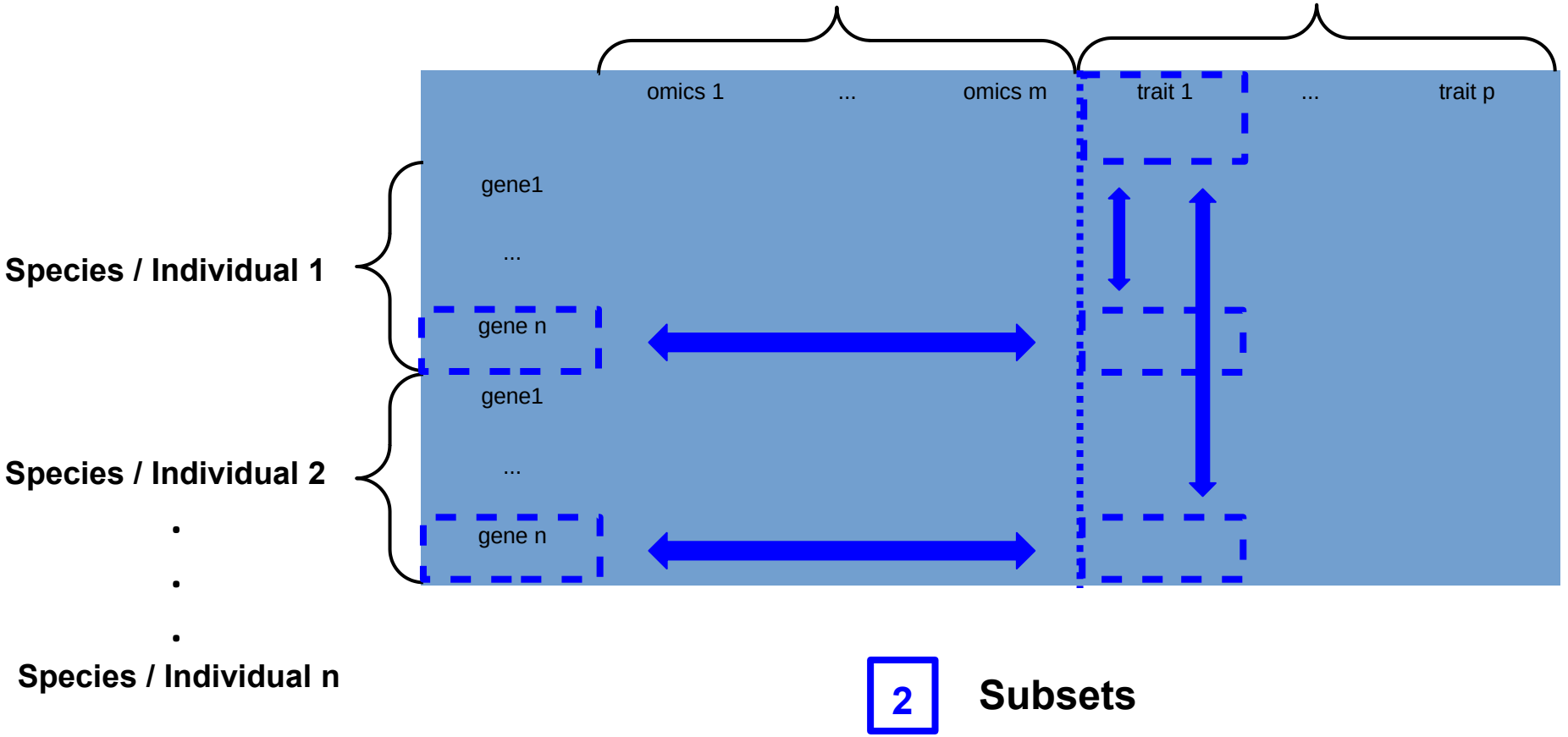
(Gen-)omics data integration – Principles



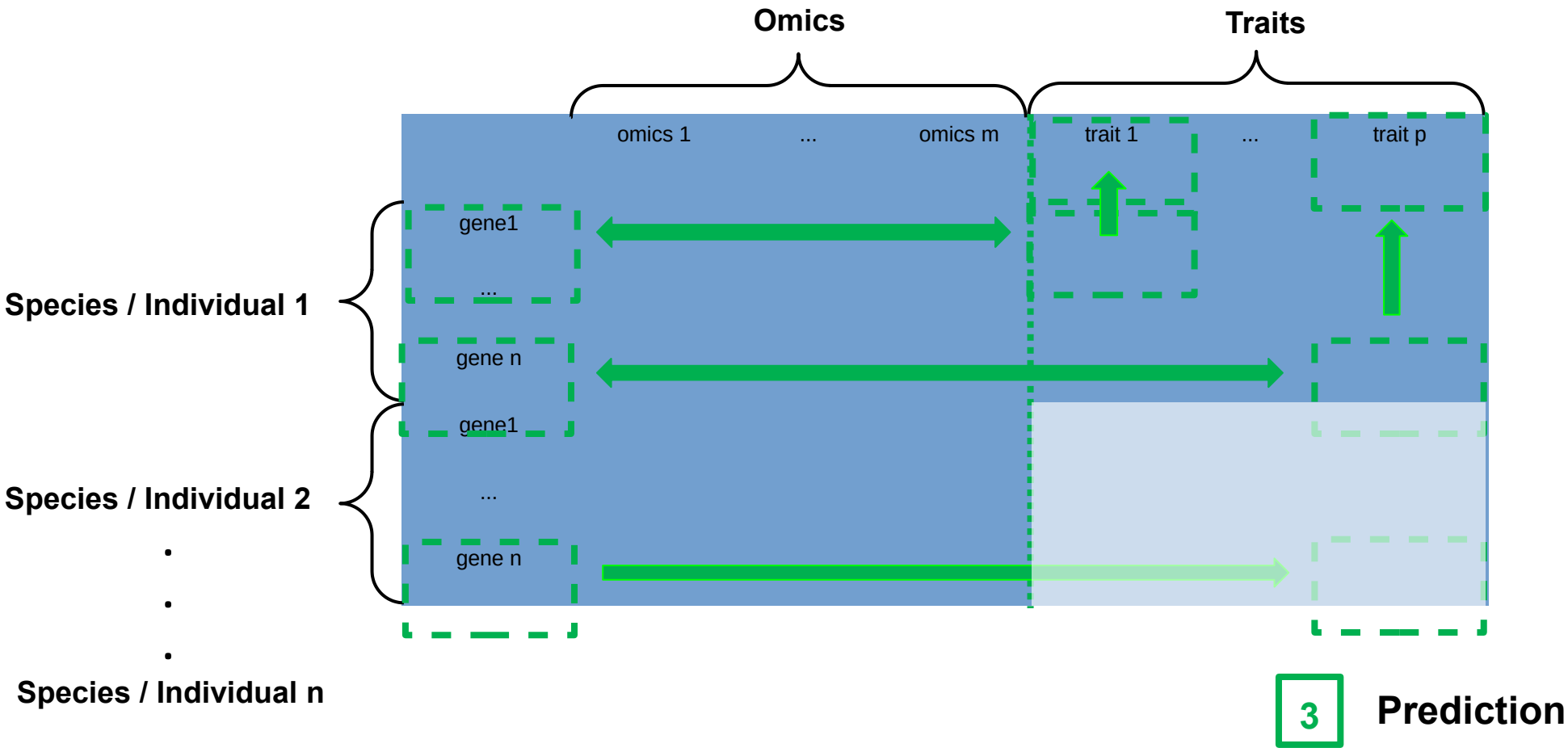
(Gen-)omics data integration – Principles



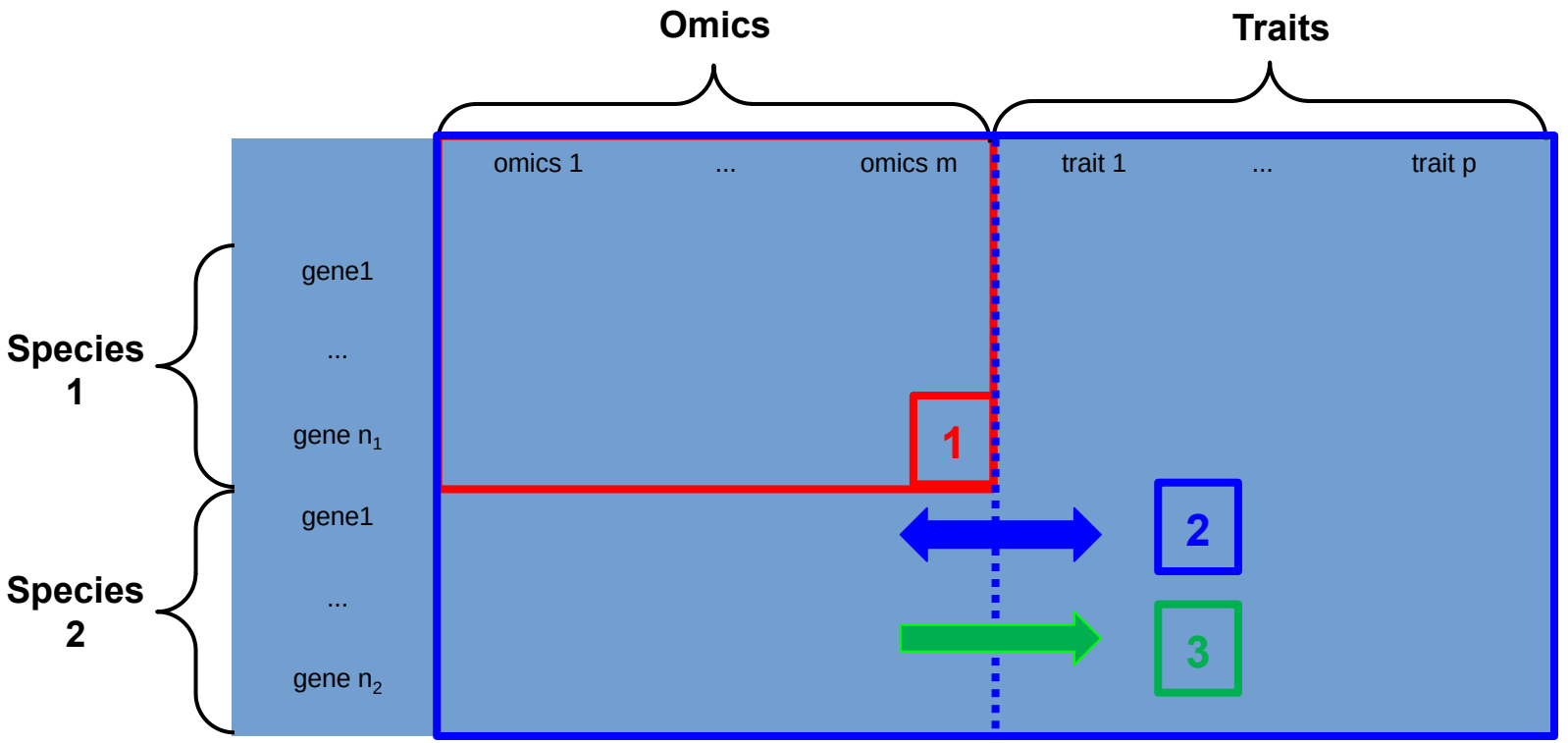
(Gen-)omics data integration – Principles



(Gen-)omics data integration – Principles



(Gen-)omics data integration – Principles

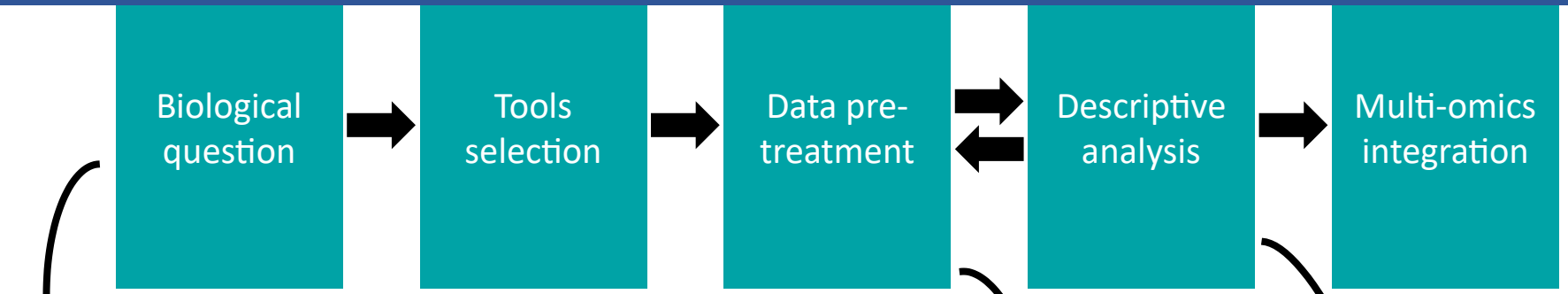


1 **Interactions**

2 **Subsets**

3 **Prediction**

(Gen-)omics data integration – Methods

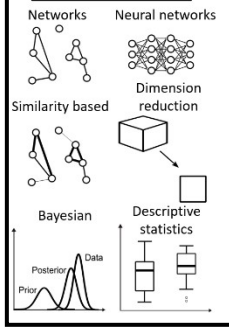


1. What are the main **interactions** between omics data at the whole dataset or subset level?
2. Which **subsets** are specific across omics to a data feature?
3. How to **predict** phenotypes/omics data from other omics data?

TOOLS

Missing values	Outliers	Normalizations	Batch effects
• Deletion • Replacement	• Deletion • Effect reduction	• Normal distribution • Centered / scaled data • Other transformations	• Correction • During the pretreatment or within data analysis algorithms

Types of methods



Data characteristics

Data type:
- quantitative / qualitative

Dimensions:
- number of rows and columns

Hypothesis:
- independence
- normality
- linearity

Tools characteristics

- Coding language
- Open source
- Documentation
- Updates
- Hyperparameters

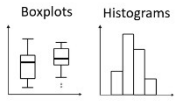
Data import

Dimensions
- number of rows and columns

Data type in the coding language
- characters string
- integer / floating numbers
- dates

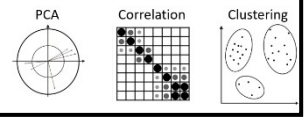
Analysis by variable

- min, max
- mean
- quantiles



Analysis by dataset

Target:
- observations
- variables

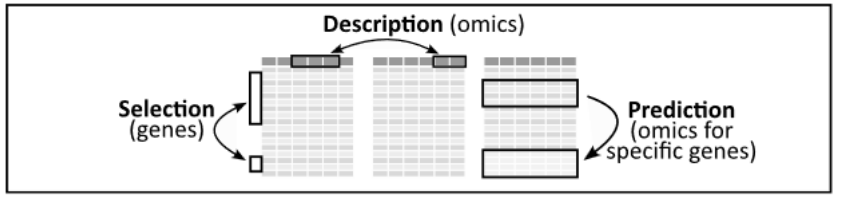


(Gen-)omics data integration – (13) Tools

tool's name	biological question: question(s) of interest	method and tool's characteristics: supervised/unsupervised	methods families	summary	updated	data characteristics: omics	hypothesis
BCC (Bayesian Consensus Clustering)	I) interactions between samples across omics	unsupervised	Statistics	Computes a samples' clustering for each omics dataset by using a probabilistic model, then merges clusters to get a consensus cluster across omics datasets	no	multi-omics (quantitative)	Normal distribution Different omics on the same set of samples
iCluster (iClusterPlus / iClusterBayes)	I) interactions between samples across omics	unsupervised	Statistics / Dimension reduction	Starts with a latent variables regression across datasets by using a probabilistic model, then uses these joint latent variables for samples' clustering	yes	multi-omics (quantitative and qualitative)	Linearity assumption Normal noise distribution Different omics on the same set of samples
JIVE (Joint and Individual Variation Explained)	I) interactions between samples/variables across omics	unsupervised	Dimension reduction	Decomposes each dataset in three terms: a joint effect (across datasets), an individual effect (specific to the dataset) and a noise effect	no	multi-omics (quantitative)	Linearity assumption
LRAcluster (Low-Rank Approximation Cluster)	I) interactions between samples across omics	unsupervised	Statistics / Dimension reduction	Probabilistically computes a common low-dimensional subspace across omics, then uses the K-means algorithm to cluster samples on this subspace	yes	multi-omics (quantitative and qualitative)	Linearity assumption Different omics on the same set of samples
MCIA (Multiple co-inertia analysis) (made4)	I) interactions between samples/variables across omics	unsupervised	Dimension reduction	Projects each dataset on a subspace, then maximizes co-inertia between subspaces to get major information shared by datasets	yes	multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
mixOmics (withPCA, PLS, rCCA, DIABLO...)	I) interactions between samples/variables across omics II) Biomarkers associated to specific trait III) Phenotype prediction from omics data	supervised unsupervised	Dimension reduction	Contains many matrix factorization methods for multivariate analysis and functions for data visualization. The main analysis method for one single dataset is the PCA. For two datasets or more, the main methods are the PLS and rCCA, and their extentions for discriminant analysis, variable selection ('sparse') and multi-blocks analysis.	yes	multi-omics (quantitative and qualitative)	Datasets with the same rows or columns
moCluster (from MOGSA)	I) interactions between samples across omics II) Biomarkers associated to specific trait III) Phenotype prediction from omics data	unsupervised	Statistics / Dimension reduction	Performs a PCA on each dataset, then merges them and finally select the best subtype model	yes	multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
MOFA (Multi-Omics Factor Analysis) (MOFA2)	I) interactions between samples across omics III) Phenotype prediction on latent	unsupervised	Statistics / Dimension reduction	Factorizes datasets with a Bayesian approach to get a small number of latent factors usable for different purposes.	yes	multi-omics (quantitative and qualitative)	Linearity assumption
NEMO (NEighborhood based Multi-Omics clustering)	I) interactions between samples across omics	unsupervised	Similarity-based	Creates one similarity matrix by dataset, then merges them and finally clusters the merged matrix by Spectral clustering	no	multi-omics (quantitative)	Euclidean distance metric
PINS (Perturbation clustering for data INtegration and disease Subtyping) (PINSPlus)	I) interactions between samples across omics	unsupervised	Similarity-based / Network	Does several clustering to identify how often samples are clustered together. Clusterings are made on different datasets, with data perturbed by adding gaussian noise, and different clustering methods are used	yes	multi-omics (quantitative)	Different omics on the same set of samples
RGCCA (Regularized Generalized Canonical Correlation Analysis) (sGCCA)	I) interactions between samples across omics	unsupervised	Dimension reduction	Computes latent variables for each dataset by maximizing correlations within and/or between datasets	yes	multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
SNF (Similarity Network Fusion)	I) interactions between samples across omics	unsupervised	Similarity-based / Network	Creates a similarity matrix then an associated network for each dataset, then iteratively fuses the networks to keep only strong correlations between samples across omics	no	multi-omics (quantitative and qualitative)	Different omics on the same set of samples Euclidean distance metric

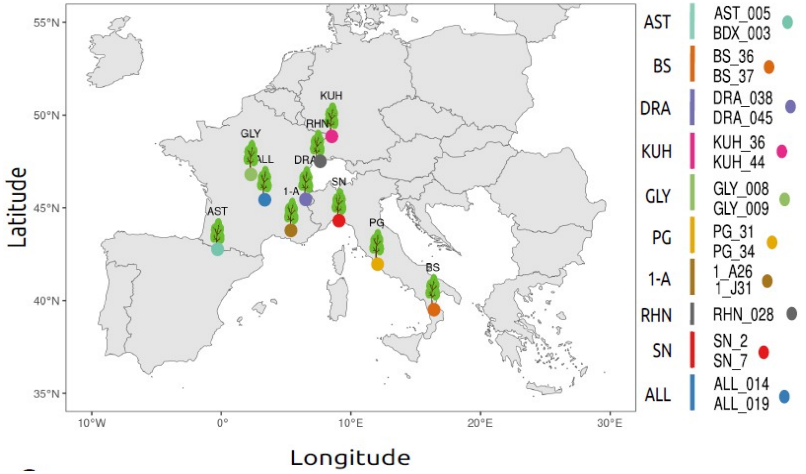
tool's name	biological question: question(s) of interest	method and tool's characteristics: supervised/unsupervised	methods families	summary	updated	data characteristics: omics	hypothesis
mixOmics	I) interactions between samples/variables across omics II) Biomarkers associated to specific trait III) Phenotype prediction from omics data	supervised unsupervised	Dimension reduction	Contains many matrix factorization methods for multivariate analysis and functions for data visualization. The main analysis method for one single dataset is the PCA. For two datasets or more, the main methods are the PLS and rCCA, and their extentions for discriminant analysis, variable selection ('sparse') and multi-blocks analysis.	yes	multi-omics (quantitative and qualitative)	Datasets with the same rows or columns

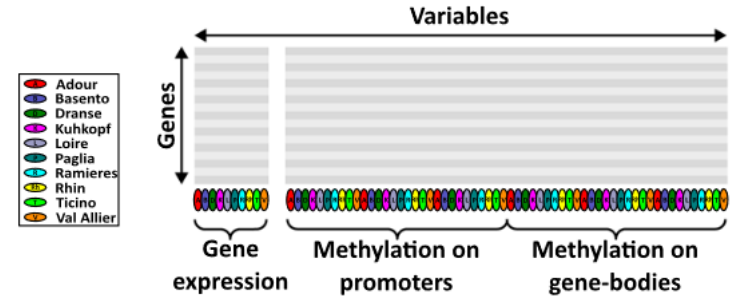
Case study on plants

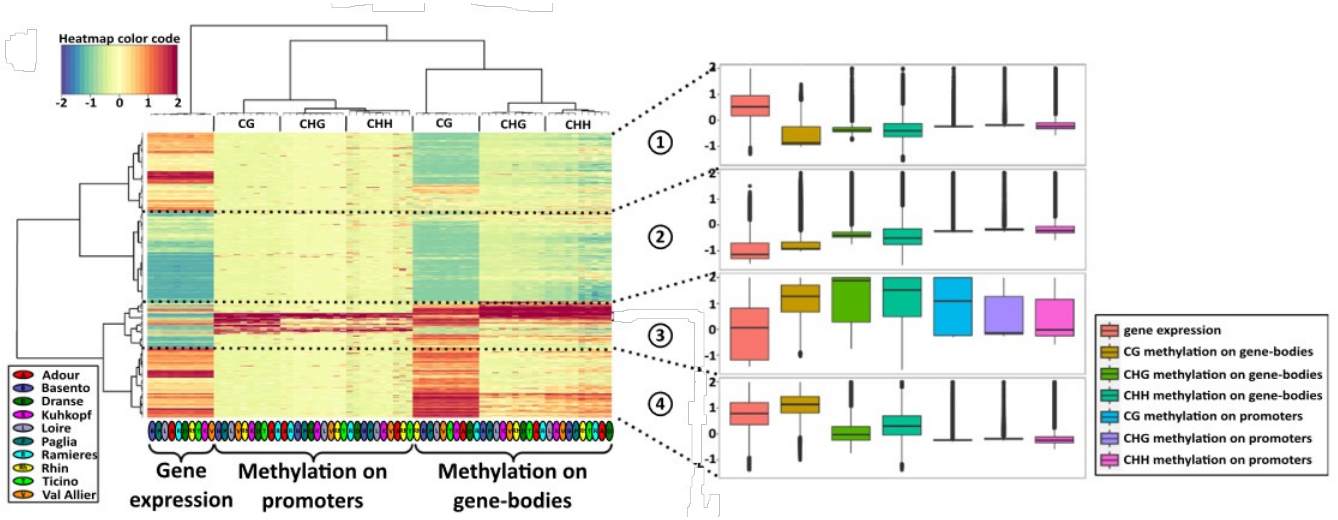


EPITREE

S Maury







(1) **high expression / low methylation** levels for promoters and gene-body in the three methylation contexts

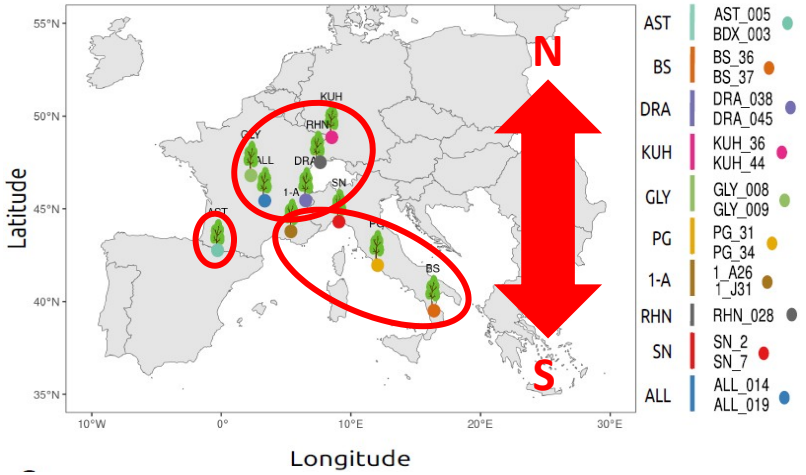
(2) **low expression / low CG gene-body methylation** and moderate non-CG gene-body and promoter methylation levels

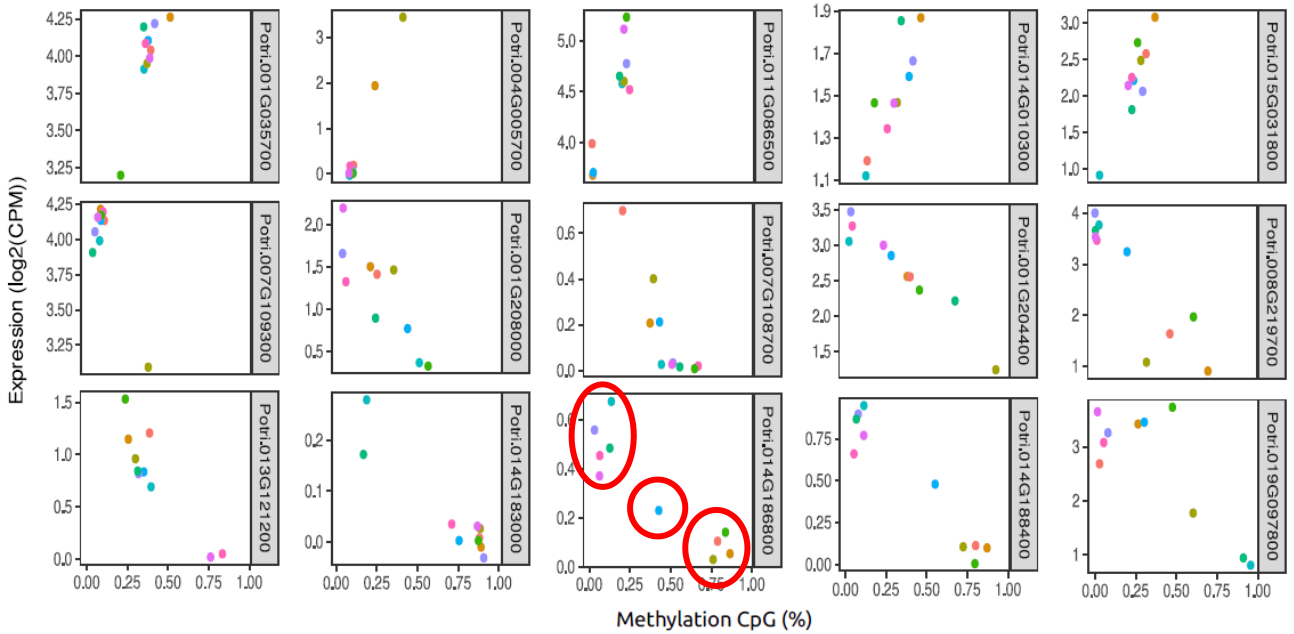
(3) **high methylation /moderate to low expression levels**

(4) **expression and CG gene-body methylation are high** while moderate values are observed for the other methylation features and contexts.

Mardoc et al. 2024 BMC Genomics 25(1):66

Case study on plants





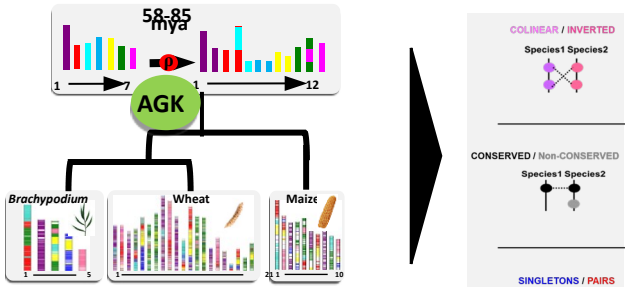
(1) CpG-specific genes (*i.e.* gbM) harbor **positive inter-population correlations** between gene expression and DNA methylation

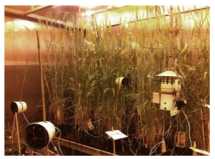
(2) 2 genes related to chromatin and epigenetic regulation (*SAC3* and *PRMT6*), **13 genes involved in stress response and disease resistance** (7 *TIR-NBS-LRR*, 4 NB-ARC, AINTEGUMENTA (*ANT1*, amino-acid transporter) and *LGY1)*

(3) N-S adaptation through expression-methylation gene regulation monitoring-programming ?

Sow et al. Under publication

Case study on plants





DNA-seq (diversity)	✓	✓	✓
RNA-seq	✓	✓	✓
BS-seq	✓	✗	✓

(i) **inverted (*vs* colinear) genes** have higher substitution rates, are less methylated and more expressed (**#1**)
(ii) **conserved (*vs* specific) genes** have lower substitution rates, are less methylated and more expressed (**#2**)
(iii) **Pairs (*vs* singletons)** have more substitutions, are less methylated and less expressed (**#3**)
(iv) WGD derive **Least Fractionated (LF) and Most Fractionated (MF)** duplicated-paralogous blocks, with LF-located genes having less substitutions with no clear consensus on expression or methylation bias between LF/MF-located genes (**#4**)

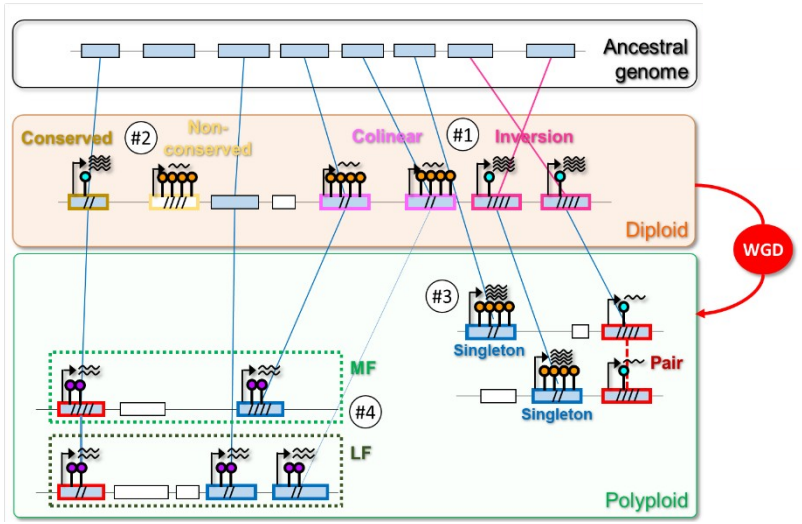
the plant journal SEB

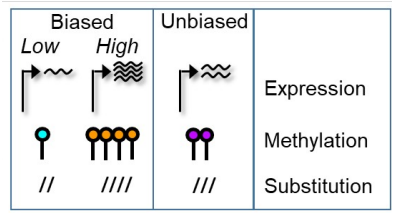
The Plant Journal (2023) **114**, 1243–1266 doi: 10.1111/tpj.16185

Tracing 100 million years of grass genome evolutionary plasticity

Arnaud Bellec[1,†], Mamadou Dia Sow[2,†], Caroline Pont[2], Peter Civan[2], Emile Mardoc[2], Wandrille Duchemin[2], David Armisen[2], Cécile Huneau[2], Johanne Thévenin[3], Vanessa Vernoud[4], Nathalie Depège-Fargeix[4], Laurent Maunas[5], Brigitte Escale[5,6], Bertrand Dubreucq[3], Peter Rogowsky[4], Hélène Bergès[1] and Jerome Salse[2,*]

A Bellec





Bellec et al. 2023 TPJ 114:1243-1266

Case study on animals

M Bonnet

Quelles sont les signatures moléculaires (protéines) de la composition tissulaire ou chimique des carcasses bovines ?

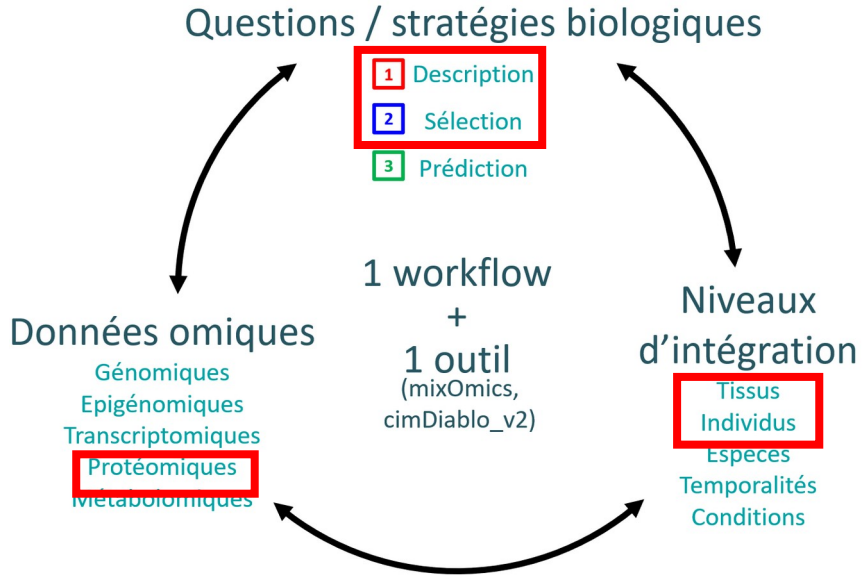
Pourquoi un phénotypage moléculaire du rapport gras/muscle chez le bovin

La composition tissulaire, ou le rapport gras /muscle, influe sur :

- La capacité à faire face à une pénurie alimentaire, un deficit nutritionnel (ex: fourrage de mauvaise qualité lors de période de sécheresse)
- La capacité à transformer un aliment en kg de muscle (**efficience alimentaire**)
- La rendement de carcasse et donc le prix payé à l'éleveur et les kg produits pour l'alimentation humaine (**efficience économique**)

Pas de méthode rapide, peu couteuse et peu invasive pour estimer le rapport gras/muscle.

=> Application du workflow à des données protéomiques de tissus bovins



Case study on animals

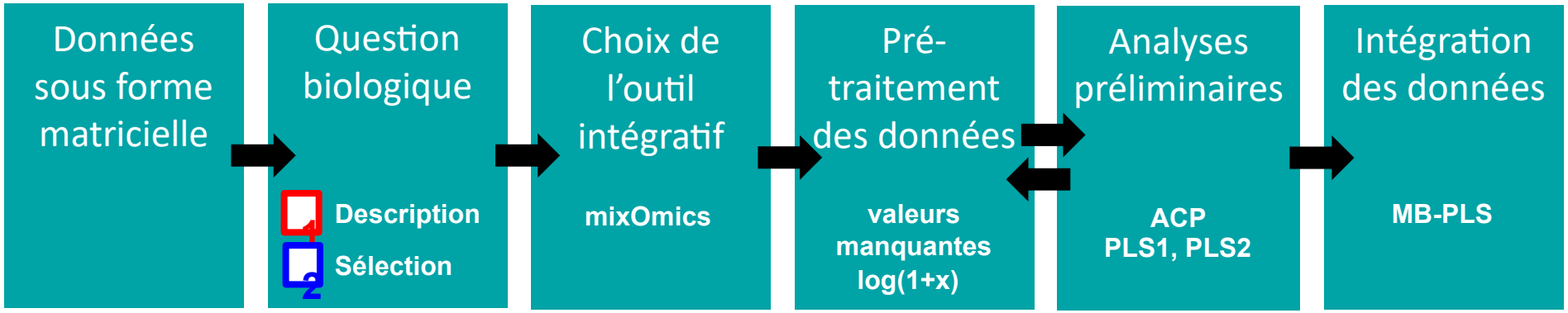
Quelles sont les signatures moléculaires (protéines) de la composition tissulaire ou chimique des carcasses bovines ?

Les données et les méthodes du Workflow

Acquisition de données de protéomiques quantitatives par SWATH-MS dans les tissus de bovins charolais divergent par le rapport gras/muscle :

- Tissu adipeux : 3085 protéines
- Muscle: 2027 protéines
- Foie: 4054 protéines

Le workflow appliqué

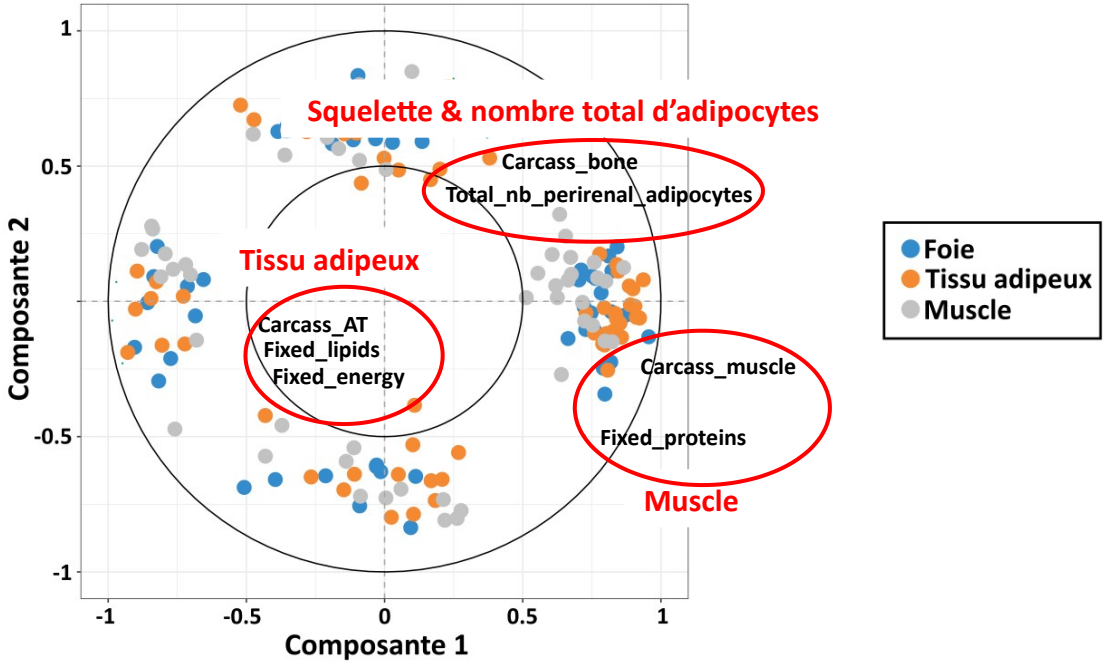


Case study on animals

Quelles sont les signatures moléculaires (protéines) de la composition tissulaire ou chimique des carcasses bovines ?

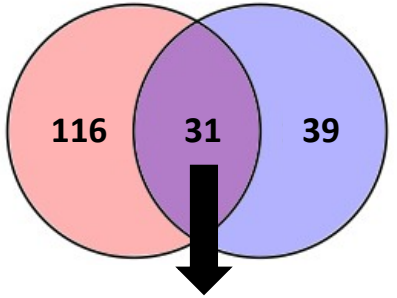
Analyse intégrative multi-tissus: PLS vs. MB-sPLS

MB-PLS, 30 prot. par tissu / composante



- 70 protéines sélectionnées
- 19 liées à la masse de muscle et de tissu adipeux périrénal

PLS1 (147 prot.) MB-PLS (70 prot.)



=> 31 prot. « robustes »
sélectionnées par les PLS1 et MB-PLS

Collaborations avec CSIRO Australie (Joint linkage) et la plateforme Auvergne BioInformatique

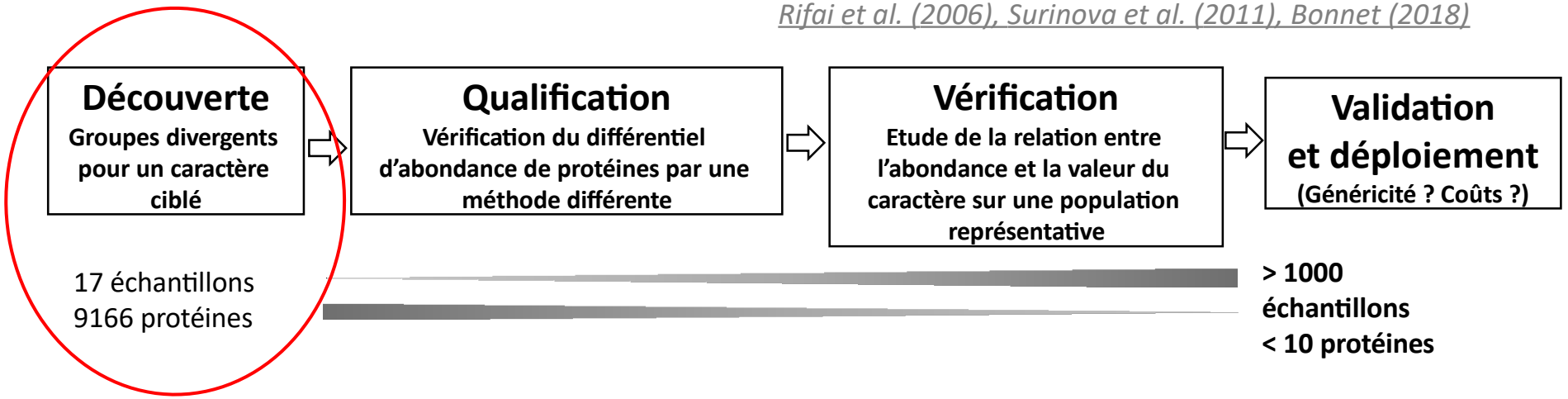
Case study on animals

Quelles sont les signatures moléculaires (protéines) de la composition tissulaire ou chimique des carcasses bovines ?

› Perspectives : qualifier les 31 biomarqueurs potentiels

Pipeline de recherche de biomarqueurs :

Rifai et al. (2006), Surinova et al. (2011), Bonnet (2018)

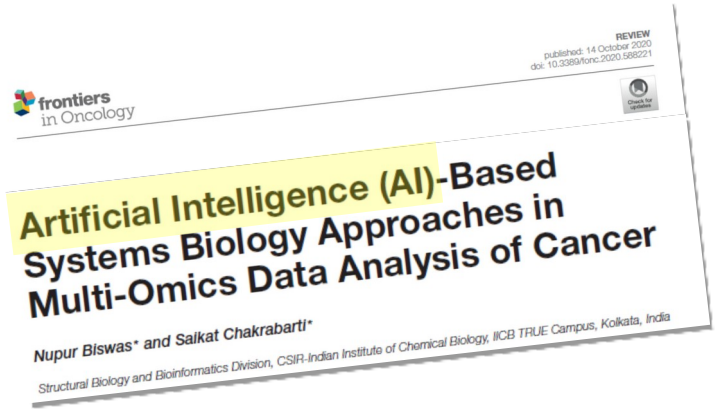


Mardoc et al. → Qualifier et vérifier pour inclure la quantification de quelques protéines à un test rapide par une méthode multiples Elisa en cours de développement.

Biomarqueurs utiles pour :

- phénotypage de caractères complexes (sans méthode destructive)
- sélection génétique pour optimiser le rendement de carcasse

IA for Biological Data Integration


frontiers in Oncology

REVIEW

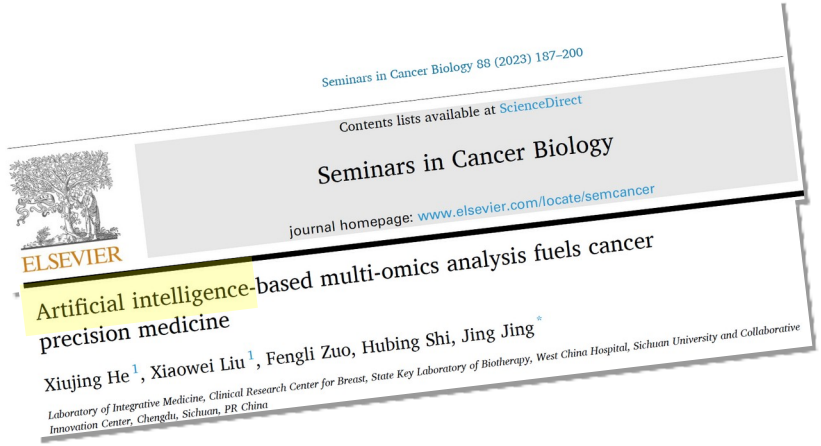
published: 14 October 2020

doi: 10.3389/fonc.2020.586221

Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer

Nupur Biswas* and Saikat Chakrabarti*

Structural Biology and Bioinformatics Division, CSIR-Indian Institute of Chemical Biology, IICB TRUE Campus, Kolkata, India


Seminars in Cancer Biology 88 (2023) 187–200

Contents lists available at ScienceDirect

Seminars in Cancer Biology

journal homepage: www.elsevier.com/locate/semcancer

Artificial intelligence-based multi-omics analysis fuels cancer precision medicine

Xiujing He [1], Xiaowei Liu [1], Fengli Zuo, Hubing Shi, Jing Jing*

Laboratory of Integrative Medicine, Clinical Research Center for Breast, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University and Collaborative Innovation Center, Chengdu, Sichuan, PR China


bioRxiv THE PREPRINT SERVER FOR BIOLOGY

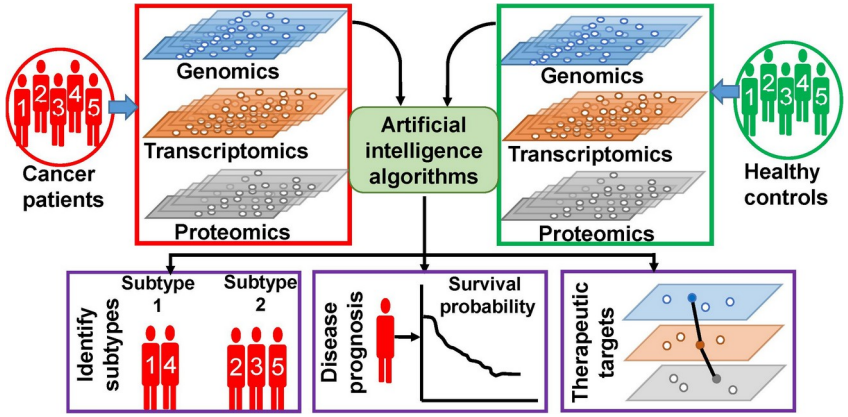
New Results

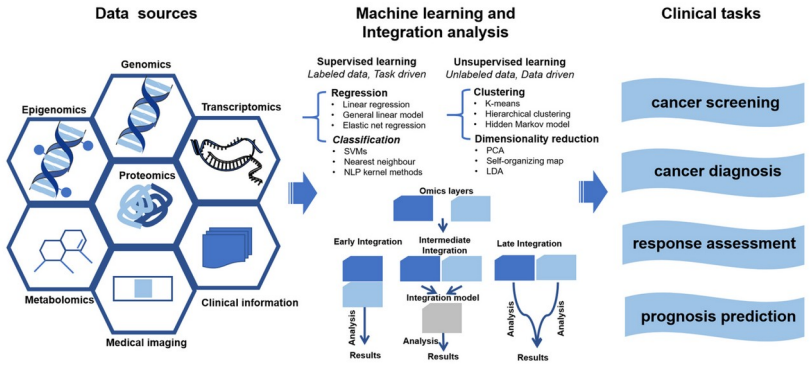
Multi-omics and deep learning provide a multifaceted view of cancer

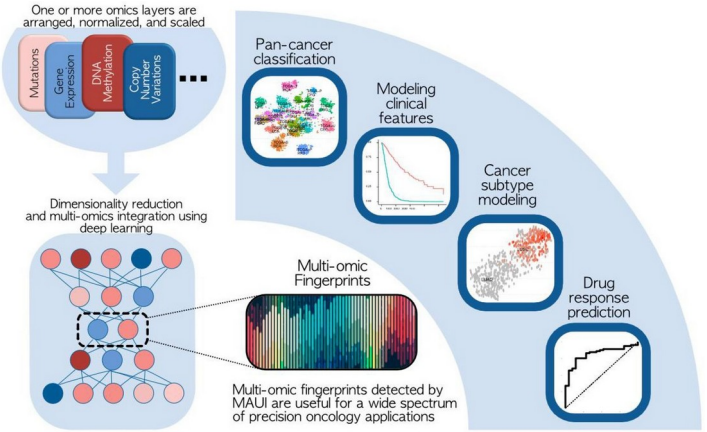
Bora Uyar, Jonathan Ronen, Vedran Franke, Gaetano Gargiulo, Altuna Akalin

doi: https://doi.org/10.1101/2021.09.29.462364

This article is a preprint and has not been certified by peer review [what does this mean?].







WORKSHOP
Master ‘Plant Science’
UE ‘Data Integration’

.L Bonhomme, J Salse

13 December 2024
(Amphi Biologie Végétale)

Morning

Laure TOUGNE (University Lyon-CNRS-INSA, Laboratoire d’InfoRmatique en Image et Systèmes d’information LIRIS) **VISIO**

Benjamin BOUREL (National Institute for Research in Digital Science and Technology, Inria)

Emile MARDOC (INRAE MetaGenoPolis, Jouy-en-Josas)

Afternoon

Romina Paola PEDRESCHI PLASENCIA (University Catolica de Valparaiso, Chile) **VISIO**

Anastase CHARANTONIS (National Institute for Research in Digital Science and Technology, Inria) **VISIO**

☆**Discussion**


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
BÏYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
JUSPAXAR
GOZAIMASHITA
EFCHARISTO
PALDIES
BOLZÏN
MERCI