Diplôme Universitaire en Bioinformatique Intégrative (DU-Bii)

Teaching Module 6 : Integrative Bioinformatics

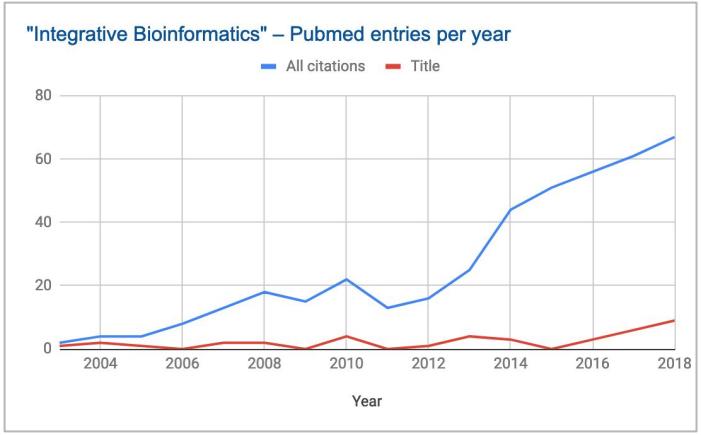
Short link: http://tinyurl.com/dubii19-m6-intro

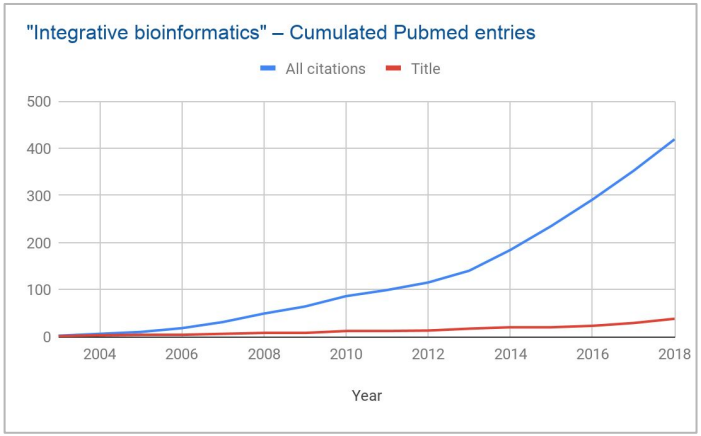
Teachers:

- Anaïs Baudot
- Costas Boulyakis
- Laura Cantini
- Sébastien Déjean
- Jérôme Mariette
- Olivier Sand
- Jacques van Helden

What is this thing called “Integrative bioinformatics” ?

- First occurrence in 2003
  - “Elucidation of ataxin-3 and ataxin-7 function by **integrative bioinformatics**”
    - profile-based sequence analysis + genome-wide functional data (model organisms) => detailed predictions of function of 2 SCA gene products
- Increasing number of citations/year since 2014
- Different connotations
  - Networks
  - NGS-based multi-omics
  - Sometimes used as buzzword to sell a single-molecule-focused study!
  - Frequently associated to medical applications (prognostics, precision medicine, ...)
  - Fashion effect ?
  - ...





What do we mean by Integrative Bioinformatics?

- Data integration
  - Heterogeneous data
  - Multi-omics
  - ...
- Integration of knowledge
  - Annotations from different sources
  - Different levels of the cell (genome, transcriptome, proteome)
  - Requires standardization of the vocabulary (*controlled vocabularies*) and organisation of the concepts (*ontologies*)
- Requires specific analysis methods (see next slides)

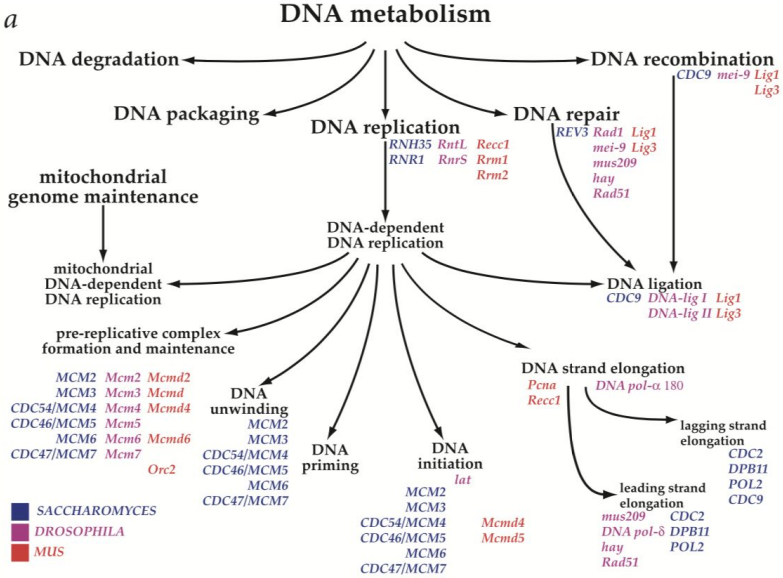
Be FAIR

- **F**indable : identifiers, metadata, search (DOI, URI, BIOSCHEMA)
- **A**ccessible : open access protocols (RDF, LOD, JSON; REST API, web services)
- **I**nteroperable : vocabularies, formal languages (ontologies, semantic web; containers, workflows)
- **R**eusable (MIAME, MIAPPE; Creative Commons licence, EDAM)
- References
  - The **FAIR** Guiding Principles for scientific data management and stewardship. Wilkinson, Dumontier et al., *Nature Scientific Data* **volume 3**, Article number: 160018 (2016).
  - Developing a Framework for Digital Objects in the Big Data To Knowledge (**BD2K**) Commons : report from the Commons Framework Pilots Workshop. Jagodnik KM et al., J. Biomed Inform. 71:49-57 (2017)
  - A design framework and exemplar metrics for FAIRness. Wilkinson et al. *Scientific Data* **volume 5**, Article number: 180118 (2018)
- FAIR principles (in french) : https://www6.inra.fr/datapartage/Produire-des-donnees-FAIR

Biological ontologies - The Gene Ontology (GO)

- Papier fondateur
  - Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). **Gene Ontology: tool for the unification of biology.** Nature Genetics 25, 25–29.
- Utilisations
  - Homogénéisation de l'annotation de bases de données
  - Analyse d'enrichissement (TP)
  - Échanges d'annotations entre bases de données et outils (***interopérabilité***)
- *"Biologists would rather share their toothbrush than share a gene name."*
  - Michael Ashburner, cited in Helen Pearson (2001) Biology's Name Game Nature 411, 631 – 632 (2001) doi:10.1038/35079694, PMID 11395736.

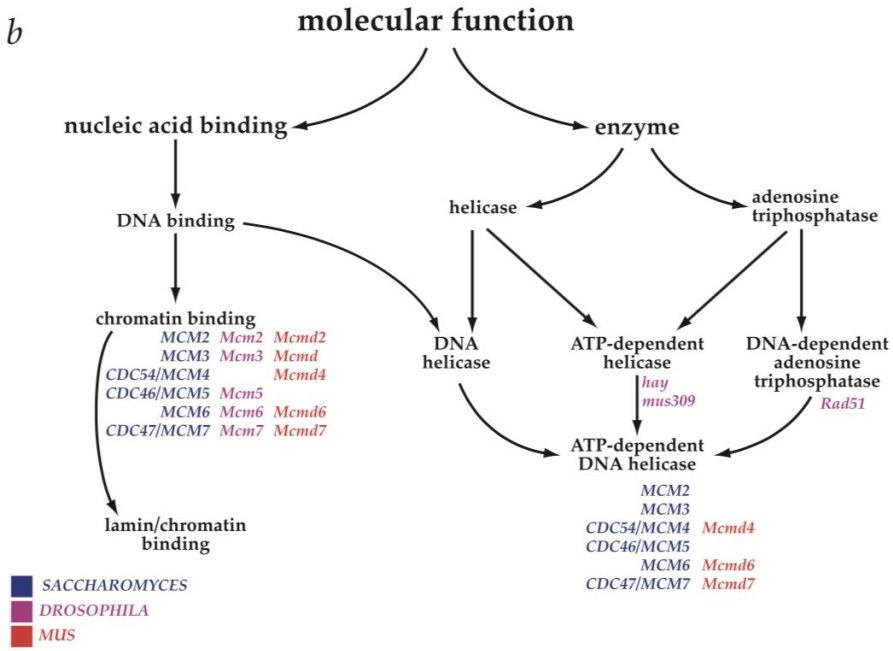
Ontology of biological processes (example)



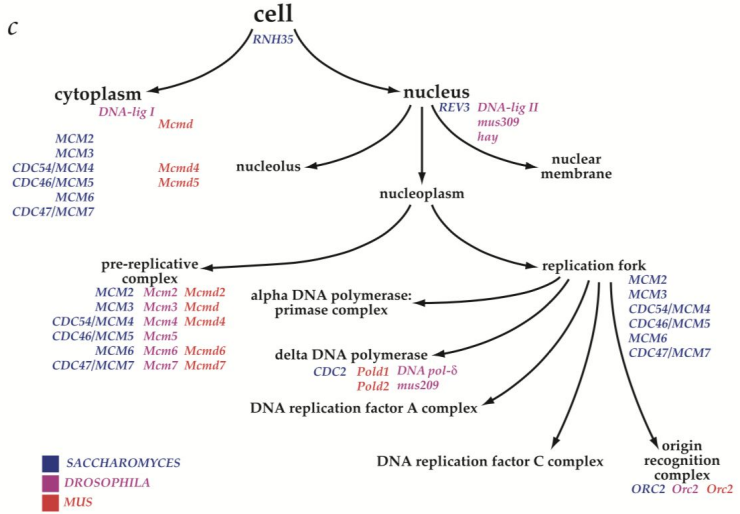
Source: Ashburner, et al. (2000). **Gene Ontology: tool for the unification of biology.** Nature Genetics 25, 25–29.

The Gene Ontology (GO)

Ontology of molecular functions (example)



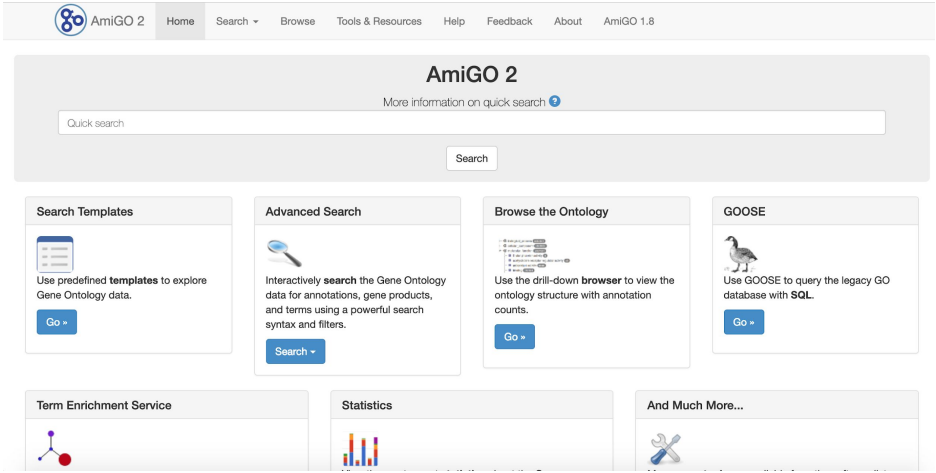
Ontology of cellular components (example)



Source: Ashburner, et al. (2000). **Gene Ontology: tool for the unification of biology.** Nature Genetics 25, 25–29.

Ontologies today

- AmiGO
  - Base de données + outils d'analyse de la Gene Ontology
  - http://amigo.geneontology.org/
-



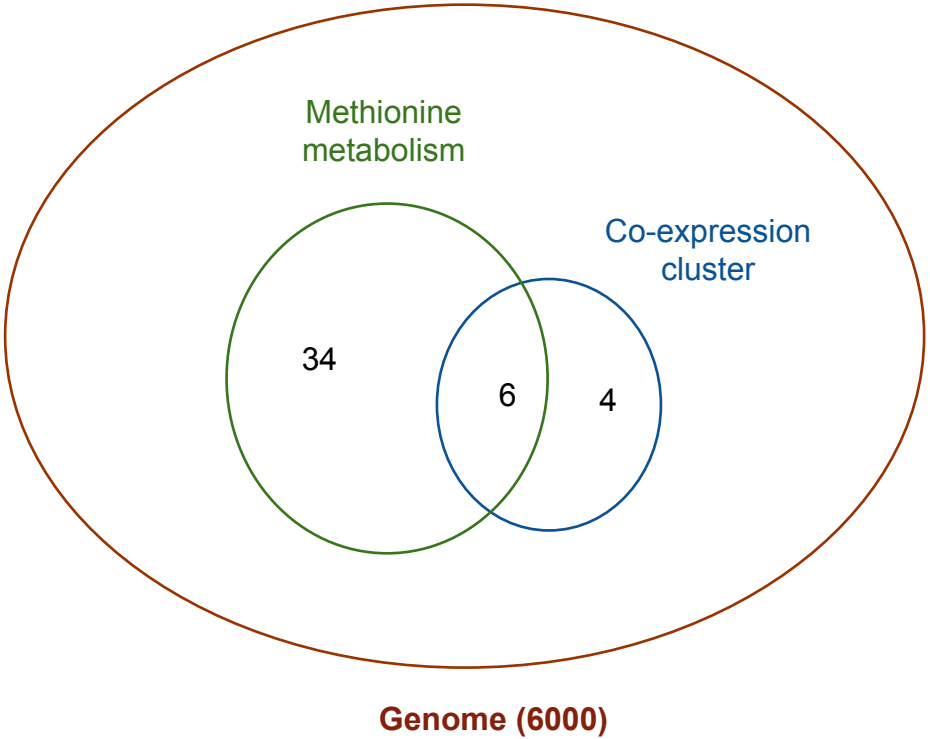
Gene set comparison and enrichment analysis

Two main approaches

- Gene set comparisons
  - Input: a set of functionally related genes
  - Reference: a database of annotated gene functions (GO, pathways, TF targets, …)
  - Approach: evaluate the significance of the intersection (over-represented ?)
  - Stat: hypergeometric test
- Gene Set Enrichment analysis
  - Input: a sorted list of genes
  - Reference: a database of annotated gene functions (GO, pathways, TF targets, …)
  - Approach: evaluate the significance of the rank of the genes belonging to a reference class in the ordered list.
  - Stat: enrichment scores (alternative)

Gene set comparisons (over-representation of the intersection

- A given organism has 6,000 genes, 40 of which are involved in methionine metabolism.
- A set of 10 genes were reported as co-regulated in a microarray experiment. Among them, 6 are related to methionine metabolism.
- How significant is this observation ? More precisely, what would be the probability to observe such a correspondence by chance alone ?
-



The hypergeometric test

- Let us define
  - g = 6000 number of genes
  - m = 40 genes involved in methionine metabolism
  - n = 5960 genes not involved in methionine metabolism
  - k = 10 number of genes in the cluster
  - x = 6 number of methionine genes in the cluster

- We calculate the number of possibilities for the following selections
  - **C1**: 10 distinct genes among 6,000
  - **C2**: 6 distinct genes among the 40 involved in methionine
  - **C3**: 4 genes among the 5960 which are not involved in methionine
  - **C4**: 6 methionine and 4 non-methionine genes

- P(X = 6): probability to have exactly 6 methionine genes within a selection of 10
- P(X >= 6) : probability to have at least 6 methionine genes within a selection of 10

$$C1 = C_{m+n}^{k} = \frac{6000!}{10!5990!} = 1.65e^{31}$$

$$C2 = C_{m}^{x} = \frac{40!}{6!34!} = 3.8e^{6}$$

$$C3 = C_{n}^{k-x} = \frac{5960!}{4!5956!} = 5.2e^{13}$$

$$C4 = C_{m}^{x} C_{n}^{k-x} = 2.0e^{20}$$

$$P(X = 6) = \frac{C_{m}^{x} C_{n}^{k-x}}{C_{m+n}^{k}} = 1.219e^{-11}$$

$$P(X \geq 6) = \sum_{i=x}^{k} \frac{C_{m}^{i} C_{n}^{k-i}}{C_{m+n}^{k}} = 1.222e^{-11}$$

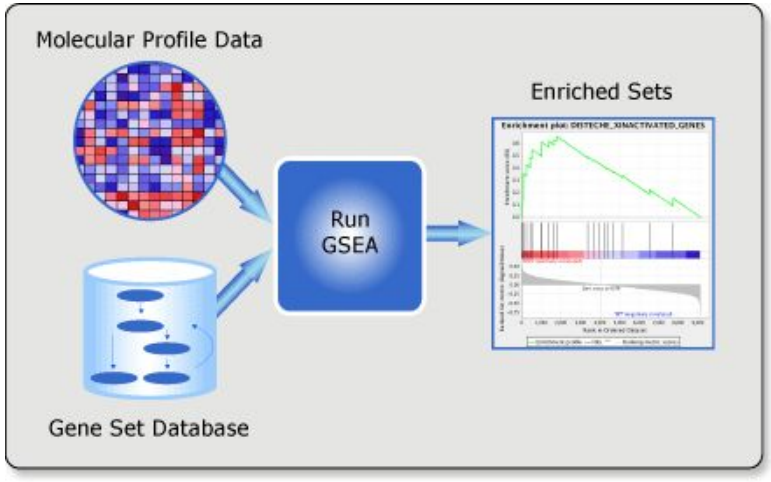
Practical - Annotating a group of DEG

- Differentially expressed genes
  - https://github.com/DU-Bii/module-3-Stat-R/tree/master/seance_5/results
- gProfiler
  - https://biit.cs.ut.ee/gprofiler/
- But:
  - Détecter les fonctions (processus biologiques, pathways, régulation, …) associées au groupe de gènes différentiellement exprimés.
  - Interpréter les réesultats
- Contrôle négatif
- A vous de jouer !

Over-representation analysis : précautions à prendre

- Définissez votre univers (background)
  - Ensemble de tous les gènes susceptibles d'entrer dans votre analyse.
- Pas si simple
  - Tous gènes présents dans les annotations génomiques ?
  - Tous les gènes ayant au moins 1 annotation dans l'ontologie concernée ?
  - Tous les gènes codants ?
  - Gènes représentés sur une biopuce ?
  - Gènes/protéines détectés dans les données expérimentales (RNA-seq, protéomique) ?
  - Les gènes "atteignables" par votre approche
    (ex: gènes-cibles des miRNA, Godard et al. 2015)
- Corrections de tests multiples
  - Choisir sa correction (P-valeur ajustée: Bonferroni, Benjamini-Hochberg, FDR)
  - Corrections pour les dépendances entre tests ? Généralement pas pris en compte dans les outils.

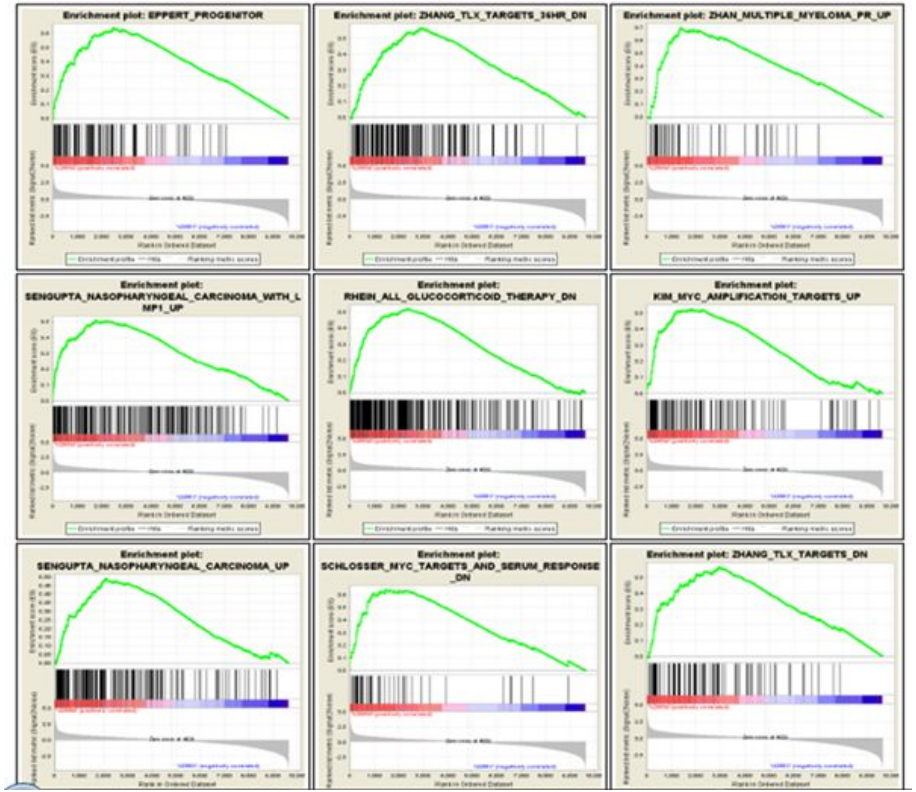
- Gene Set Enrichment Analysis
- Since 2006
- determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states
- http://software.broadinstitute.org/gsea/index.jsp
- MSigDB (The Molecular Signatures Database) : collection of annotated gene sets
- Package R : https://bioconductor.org/packages/release/bioc/html/GSEABase.html



GSEA principle

- All genes are sorted according to some criterion (e.g. differential expression p-value, correlation of expression with other variables, …).
- Each graph compares the ranked gene list with one reference class (e.g. one biological process).
- Black bars denote genes belonging to the reference class.
- The green curve estimates, at each level *i*, the degree of over-representation of the reference genes in the *i* top-ranking genes.

Table: Snapshot of enrichment results



Source:
https://steemit.com/steemstem/@scienceangel/lab-diaries-5-gene-set-enrichment-analysis-gsea-of-a-large-scale-biological-data-part-i

Dedicated journal

- **Journal of Integrative Bioinformatics**
  - https://www.degruyter.com/view/j/jib
  - Launched in 2004
  - CiteScore 2017: **0.77**
  - SCImago Journal Rank (SJR) 2017: **0.336**
  - Topics : tools and databases covering
    - Molecular Databases, Information Systems and Data Warehouses
    - Integration of Data, Methods and Tools
    - Network Analysis, Modeling and Simulation
    - Medical Informatics, Biomedicine and Biotechnology
    - Visualization and animation
    - ...

Approaches and tools

Two mainstream approaches

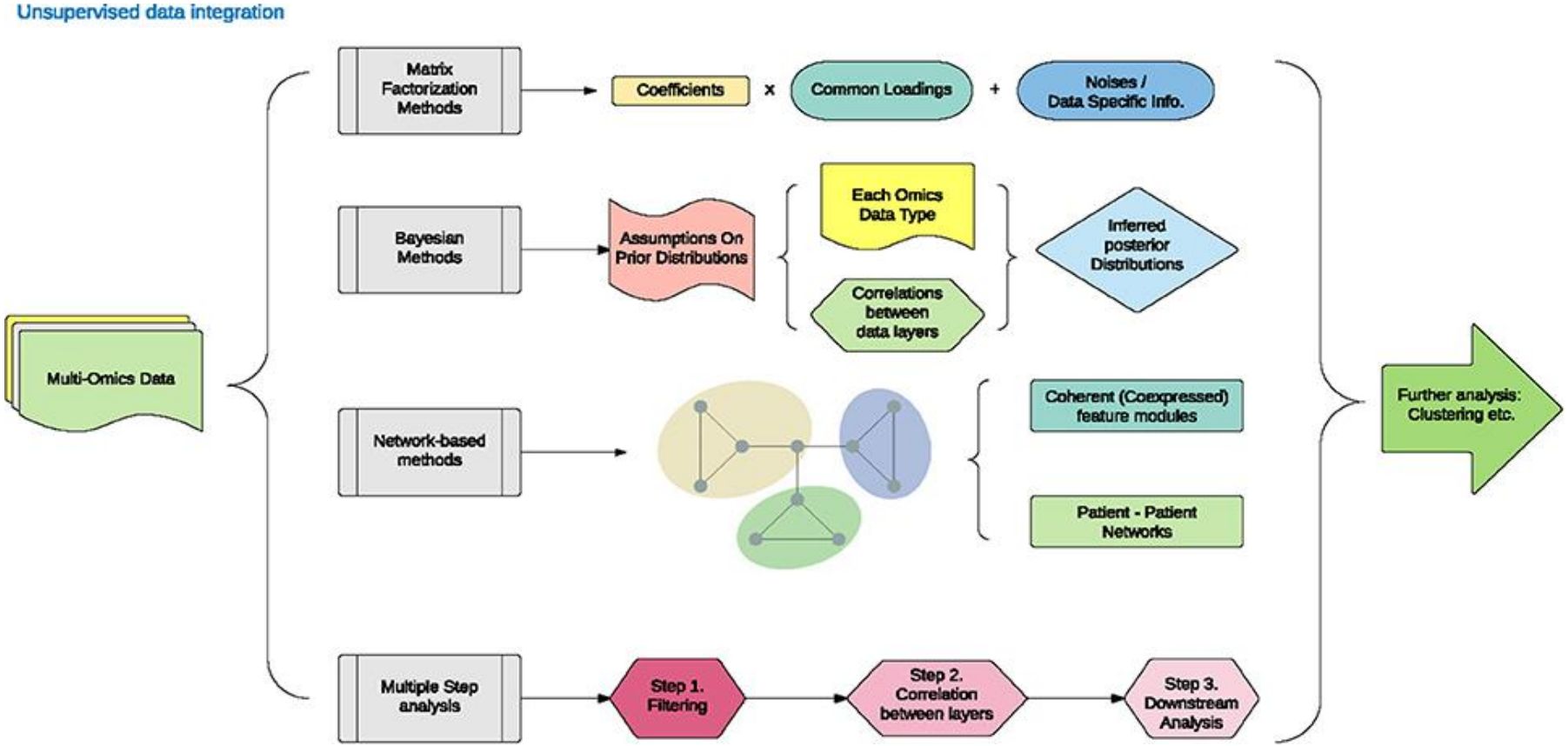
- Multi-level factorization approaches (multivariate statistical analysis)
- Multi-layer (multiplex) networks
- Combining both approaches

Software environments

- Statistical approaches: R package Mixomics (http://mixomics.org/)
- Network approaches: Cytoscape (https://cytoscape.org/)

Panorama of approaches and resources for integrative bioinformatics

Unsupervised data integration



https://www.frontiersin.org/articles/10.3389/fgene.2017.00084/full

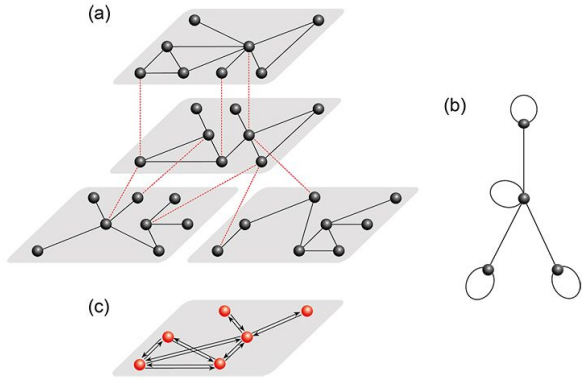
Unsupervised data integration

- draw an inference from input datasets without labeled response variables
- **Matrix factorization methods**
  - projection of variations among data sets onto dimension-reduced space
  - examples
    - Joint Non-negative Matrix Factorization (NMF)
    - iCluster, iCluster+
    - Joint and Individual Variation Explained (JIVE)
    - Joint Bayes Factor
- **Correlation-based analysis**
  - adaptation of Canonical Correlation analysis (multivariate analysis of correlation)
  - penalization and regularization terms added

Unsupervised data integration

- **Bayesian methods**
  - assumptions on different types of data sets w/ various distributions
  - assumptions on correlations among data sets
  - examples
    - Multiple Dataset Integration (MDI)
    - Patient-Specific Data Fusion (PSDF)
    - Bayesian Consensus Clustering (BCC)
    - COpy Number and EXpression In Cancer (CONEXIC)

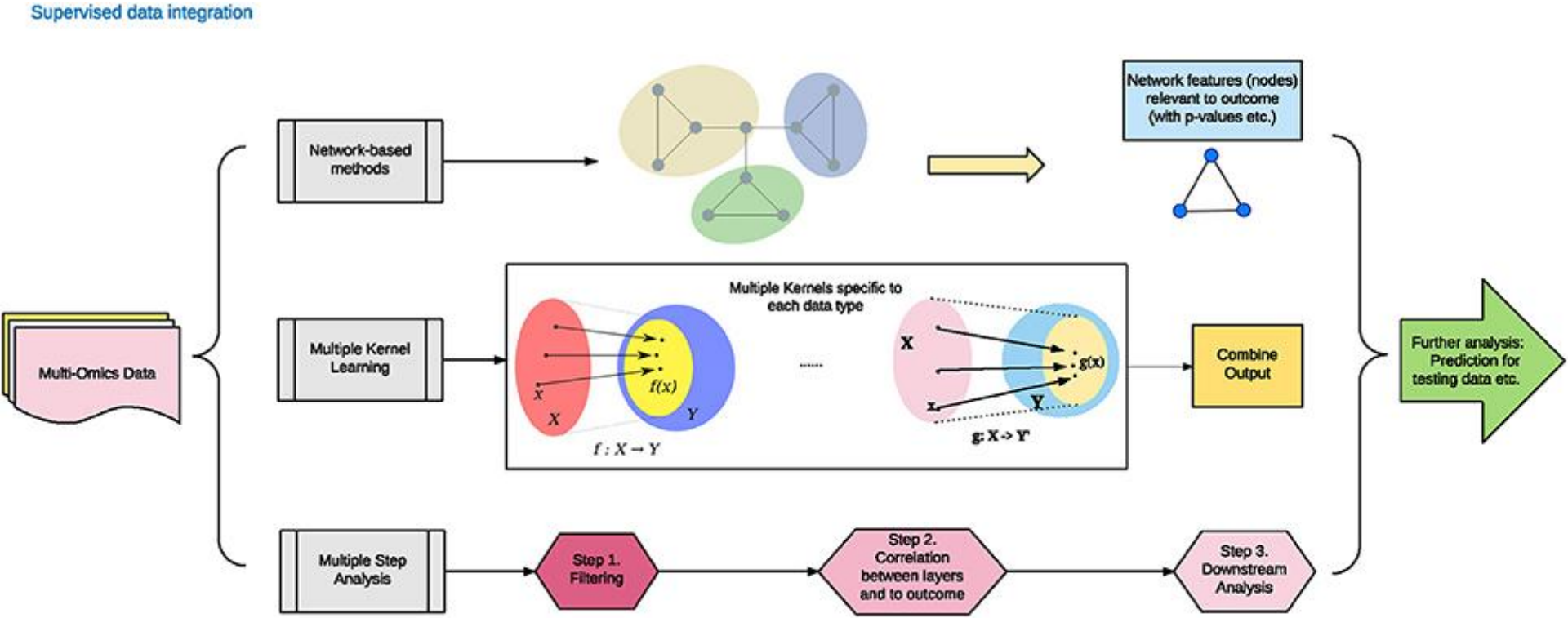
- **Network-based methods**
  - mostly applied for detecting significant genes within pathways, discovering sub-clusters, or finding co-expression network modules
  - examples
    - PAthway Representation and Analysis by DIrect Reference on Graphical Models (PARADIGM)
    - Similarity Network Fusion (SNF)
    - Lemon-Tree
    - Multiplex networks



Unsupervised data integration

- **Multi-Step Analysis**
  - Commonly used to find relationships between the different data types first, and then between the data types and the trait or phenotypes
  - Examples
    - CNAmet
    - In-Trans Process Associated and Cis-Correlated (iPAC)
- **Multiple Kernel Learning**
  - Multi-step
  - Machine learning methods
  - Optimal combination of predefined kernels (weighing factors)
  - Example
    - Regularized Multiple Kernel Learning Locality Preserving Projections (rMKL-LPP)

Supervised data integration



https://www.frontiersin.org/articles/10.3389/fgene.2017.00084/full

Supervised data integration

- Built via information of available known labels from the training omics data
- **Network-based methods**
  - jActiveModules
- **Multiple Kernel Learning**
  - Semidefinite Programming/Support Vector Machine (SDP/SVM)
- **Multi-Step Analysis**
  - Multiple Concerted Disruption (MCD)
  - Anduril

Semi-supervised data integration

- Lies between supervised and unsupervised methods
- Takes both labeled and unlabeled samples to develop learning algorithm
- most of the semi-supervised data integration methods are graph-based
- GeneticInterPred

Other approaches

- **Semantic web approaches**
  - metadata (machine-readable code) defines the data
  - keywords
  - ontologies
- **Data warehousing approaches**
  - data from different sources integrated in a single database
  - needs standardization of formats

- **Atlas - a data warehouse for integrative bioinformatics (2005)**
  - locally stores and integrates biological sequences, molecular interactions, homology information, functional annotations of genes, and biological ontologies
  - Application Programming Interfaces (C++, Java, Perl)
  - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-34
  - Availability : http://bioinformatics.ubc.ca/atlas/ **(not found on the server :-)**

- **DAS : Distributed Annotation System**
  - Integration of biological data (2001)
    - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-2-7
- **Blast2GO**
  - basic version free
  - Blast2GO Pro
  - platform for high-quality functional annotation and analysis of genomic datasets

- **First software in 2005**
  - **BIAS (Bioinformatics Integrated Application Software)**
  - Environment for carrying out integrative Bioinformatics research requiring **multiple datasets and analysis tools**
  - Follows an object-relational mapping for providing persistent objects
  - Allows third-party tools to be easily incorporated within the system.
  - Supports standards and data-exchange protocols common to Bioinformatics
  - Availability : http://www.mcb.mcgill.ca/~bias/ **(server not found :-))**

KEGG mapper

- First release in 2010
- Collection of tools
  - Pathway mapping
    - Basic search : mapped objects marked in red
    - Coloring options
    - Coloring without search : selected map, color gradation
  - Network mapping
  - Disease mapping
  - …

- since 2010
- free all-in-one online solution composed of interactive tools for metabolic network curation, network exploration and omics data analysis
- https://metexplore.toulouse.inra.fr/index.html/
- collaborative environment
- interactive tables connected to a powerful network visualisation module
- contextualisation of metabolic elements in the network
- calculation of over-representation statistics

- Proteomics Research Environment
- http://www.proteore.org
- galaxy framework (no programming required)
- 15 tools to manipulate, annotate, analyse and visualize **human** data
- share data and workflows

g:profiler

- public web server for characterising and manipulating gene lists
- 400+ species, including mammals, plants, fungi, insects (Ensembl)
- Several tools
  - g:GOSt = statistical enrichment analysis
  - g:Convert = gene identifier conversion tool
  - g:Orth = mapping homologous genes across related organisms
  - g:SNPense = mapping human SNPs to gene names, chromosomal locations and variant consequence terms from Sequence Ontology
- R package

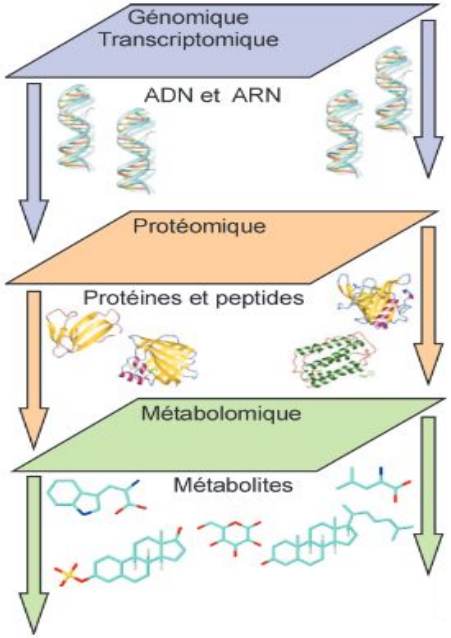
Workflows

- **BioPipe (2003)**
- **BioWBI (2004)**
- **Taverna (2004)**
- **Wildfire (2005)**
- **KDE Bioscience (2006) : Knowledge Discovery Environment of Bioscience**
  - Platform for bioinformatics analysis workflows
  - Integrate data, algorithms, computing resources, and human intelligence
  - More than 60 included programs
  - https://www.sciencedirect.com/science/article/pii/S1532046405000821?via%3Dihub
  - Availability : **? :-)))**
- **Galaxy**
- **Nextflow**
- **SnakeMake**

Actions IFB en bioinformatique intégrative

- Bottleneck: segmentation of the competences and tools for the diverse omics methodologies
- Challenge: ***multi-level integration*** *(beyond multi-omics)*: molecular (genomics, proteomics, metabolomics, …) + structural + cellular / tissular (imaging) + organisms (phenotypes) + health (cohorts, precision medicine) + environment (metagenomics)

- **IFB actions for integrative bioinformatics**
- **Innovation**
  - ***A2.1 Pilot projects*** *:* collaboration with other national infrastructures
    - 8 projects regrouping 11 PIA partners
  - ***A2.2 call to challenges*** (2019-2021): solving technological bottlenecks
  - ***A2.3 Interopérabilité*** and ***integration of bioinfo ressources*** (data and tools)
- **Training**
  - Diplôme Universitaire en Bioinformatique intégrative (1 month course + 1 month personal project on IFB platform)
- **Inserm collaborations**
  - Workshop "***Challenges et perspectives en bioinformatique intégrative***" (2018-10)
  - Transversal call **HuDeCa** (2019-02)



France Médecine Génomique 2025

IFB innovation axis: integrative bioinformatics

- 2018-01 A2.1. Pilot-projects in Integrative Bioinformatics
  - https://www.france-bioinformatique.fr/en/projets-pilotes
  - Call to collaborations with other national research infrastructures
  - 8 projects supported (18 - 24 months FTE) regrouping 17 infrastructures
- 2018-10 Aviesan 1-day workshop
  - ***Challenges and Perspectives in Integrative Bioinformatics***
- 2019-01 Diplôme Universitaire en Bioinformatique Intégrative
  - Paris-Diderot / IFB collaboration
  - https://formation-continue.univ-paris-diderot.fr/dubii
- 2019-01 Participation to Inserm transversal calls (Human development cell atlas, Human data)
- 2019-09 Call to challenges in Integrative Bioinformatics
  - Starting from unsatisfied needs of research teams