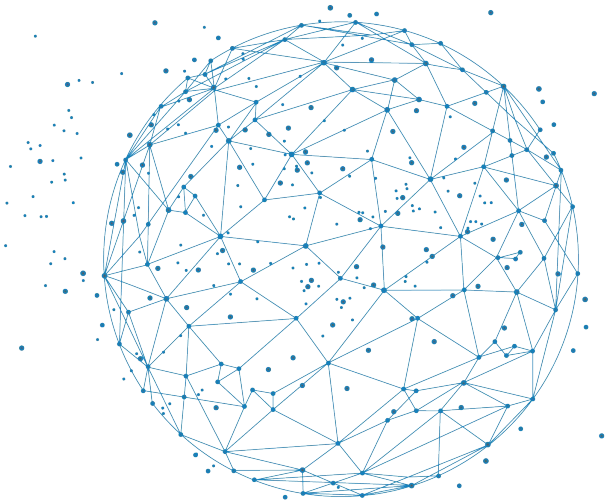
2024 - Fréjus

ÉCOLE THÉMATIQUE ETBII BIOINFORMATIQUE INTÉGRATIVE IFB



École thématique en bioinformatique Intégrative

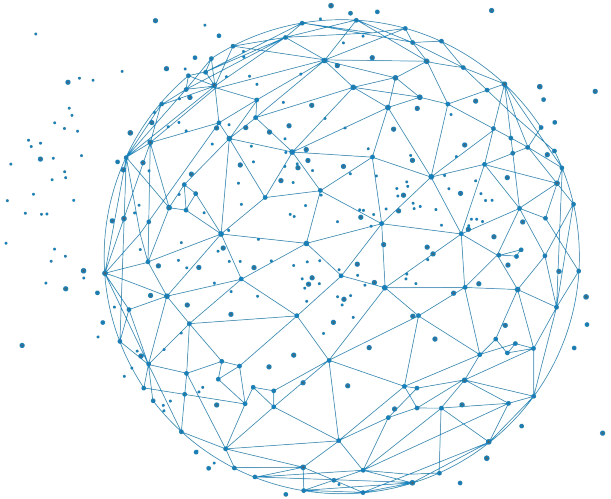
Hélène Chiapello, Olivier Sand
& Lucie Khamvongsa-Charbonnier

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea

Présentation des jeux de données



Breast cancer TCGA from MixOmics

- Human breast cancer is a heterogeneous disease in terms of molecular alterations, cellular composition, and clinical outcome. Breast tumours can be classified into several subtypes, according to levels of mRNA expression (Sørlie et al., 2001).

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

ARTICLE

doi:10.1038/nature11412

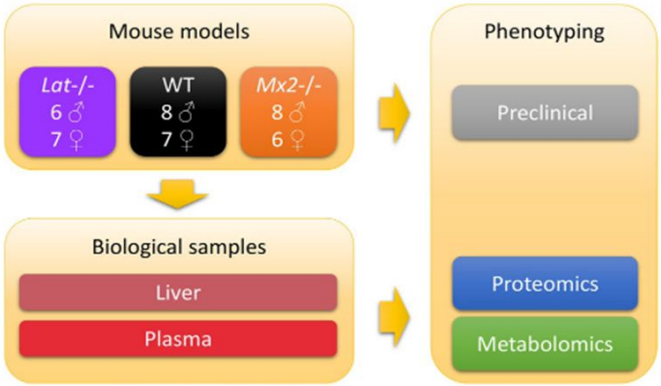
Comprehensive molecular portraits of human breast tumours

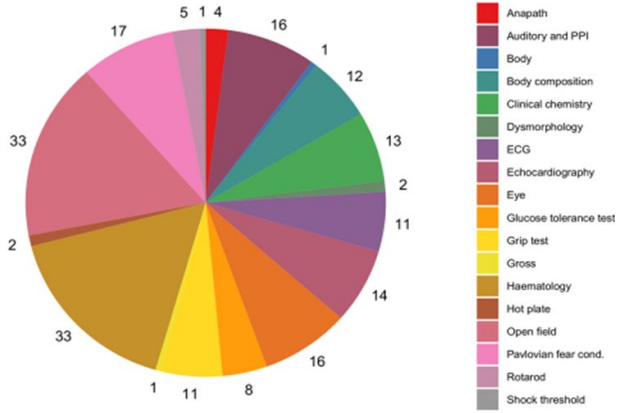
The Cancer Genome Atlas Network*

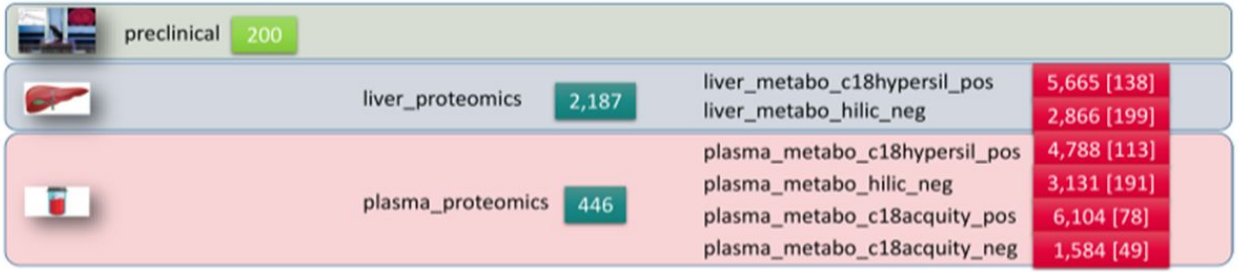
- The mixOmics TCGA dataset is accessed via breast.TCGA' and contains the following:

- breast.TCGA$data.train$mirna (continuous matrix): 150 rows and 184 columns. The expression levels of 184 different sections of miRNA.
- breast.TCGA$data.train$mrna (continuous matrix): 150 rows and 200 columns. The expression levels of 200 different sections of mRNA.
- breast.TCGA$data.train$protein (continuous matrix): 150 rows and 142 columns. The abundance of 142 different proteins
- breast.TCGA$data.train$subtype (categorical vector): length of 150. Indicates the breast cancer subtype of each subject. Includes Basal, Her2 and LumA.

The dataset provides unique molecular information about the physiological role of the *Lat* (involved in neurodevelopmental diseases) and *Mx2 (*modelling Down syndrome in mice*)* genes in mice.

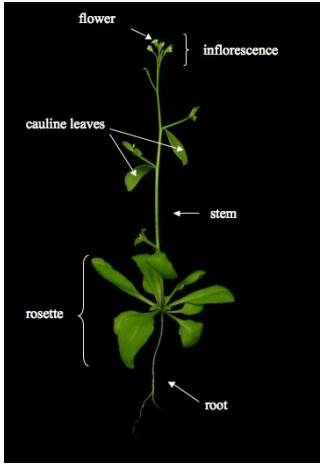






All figures are taken from:
Imbert A, Rompais M, Selloum M, Castelli F, Mouton-Barbosa E, Brandolini-Bunlon M, Chu-Van E, Joly C, Hirschler A, Roger P, Burger T, Leblanc S, Sorg T, Ouzia S, Vandenbrouck Y, Médigue C, Junot C, Ferro M, Pujos-Guillot E, de Peredo AG, Fenaille F, Carapito C, Herault Y, Thévenot EA. ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. Sci Data. 2021 Dec 3;8(1):311. doi: 10.1038/s41597-021-01095-3. PMID: 34862403; PMCID: PMC8642540.

- **Goal:** Understand the mechanisms of plant adaptation to contrasted growth temperature.
- The study focuses on the cell walls (CWs) that represent a dynamic extracellular compartment that contributes to modify the cell and plant shapes at any time during development.
- We will limit ourselves to the following **omic modalities** (n=30):
  - CW **Transcriptomic** on the rosette (p=364) and the floral stem (p=414)
  - CW **Proteomic** on the rosette (p=364) and the floral stem (p=414)
  - **Phenomics** on the rosette (p=5) and the floral stem (p=4)
- Along with :
  - **Altitude Cluster:** the environment height from which is originated a given plant (ref/low/high).
  - **Ecotype:** the genotype specifically designed for a given ecosystem (Col, Grip, Hern, Hosp, Roc).
  - **Temperature:** the temperature at which the studied sample was exposed all along its growth, either 22°C (optimal condition) or 15°C (high altitude condition).
  - **Metadata:** Bioinformatics Annotation and description of all the Cell Wall Proteins (CWPs)
- The R package WallOmicsData allows you to retrieve all the data.



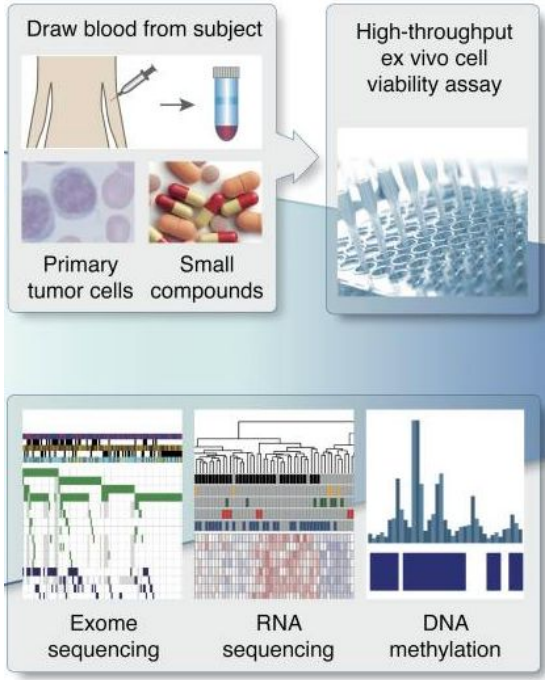
taken from Topics in Animal and Plant Development: From Cell Differentiation to Morphogenesis

Tomato

- **Goal:** study protein turnover at the global scale in developing tomato (Solanum lycopersicum) fruit.
- **Two omic modalities** were acquired to achieve this goal:
  - **Transcriptomique** (n=27 / p=2375)
  - **Proteomic** (n=27 / p=2375)
- Along with two co-variables:
  - **Days Post Anthesis (DPA):** the count of days that have elapsed since the opening of a flower, serving as a marker to track the temporal progression of plant development.
  - **GRowth stages (GR):** denote specific phases within the tomato plant's lifecycle : germination, leaf development, formation of side shoots, inflorescence emergency, flowering, fruit development, maturity of fruit and senescence
- **Possible question:** find sets of mRNAs and proteins responsible for the discrimination between GR groups or the prediction of the DPA
- The dataset was kindly provided to us by the authors of: Isma Belouah and others, Modeling Protein Destiny in Developing Fruit, Plant Physiology, Volume 180, Issue 3, July 2019, Pages 1709–1724, https://doi.org/10.1104/pp.19.00086

Chronic Lymphocytic Leukemia (CLL)

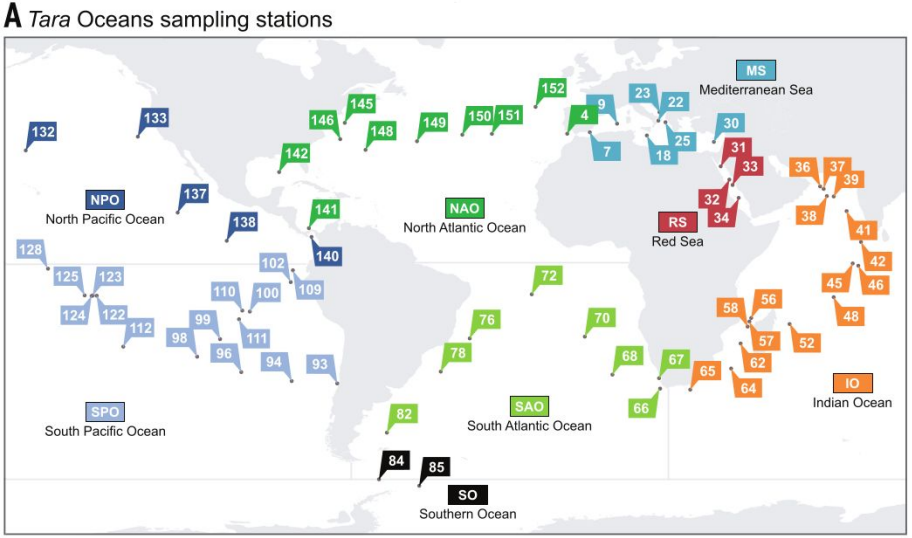
- **Goal**: study ex-vivo response of 243 blood cancer samples (majority of CLL) to 63 drugs with 5 concentrations (+ 3 healthy samples)
- **Five modalities**
  - Copy number variants (n=169) + 6 structural variants (n=125-162)
  - Genomic / WES (n=107) + targeted for 9 genes (n=188-231)
  - Transcriptomic / NGS (n=123)
  - Methylation (n=196)
  - Drug response based on cell survival (n=243 patients + 3 controls)
- Stratification of CLL samples based on drug responses + association to omics
- For CLL sample association between drug response and few mutations/variants (IGHV, TP53, BRAF… and trisomy 12)
- **In CLL dataset (n=200): Variants (p=69), Transcriptomic (p=5000) , Methylation (p=4248) , Drug response (p=310) + metadata (p=9)**



Dietrich S, et al.. Drug-perturbation-based stratification of blood cancer. J Clin Invest. 2018 Jan 2;128(1):427-445.

TARA OCEAN

- Données métagénomiques
  - **NOGS:** relative abundance of orthologous genes (OGs)
  - **Phylo**: counts of S16 rRNA
- 8 points de collectes
  - **SPO**: South Pacific Ocean
  - **NAO**: North Atlantic Ocean
  - **IO**: Indian Ocean
  - **RS**: Red Sea
  - **MS**: Mediterranean Sea
  - **NPO**: North Pacific Ocean
  - **SO**: Southern Ocean
  - **SAO**: South Atlantic Ocean



Sunagawa et al., 2015

- 4 profondeurs
  - **SRF**: Surface Water Layer (0-5 meters)
  - **DCM**: Deep Chlorophyll Maximum (peak of chlorophyll, 0-600 meters)
  - **MIX**: Subsurface epipelagic Mixed Layer
  - **MES(O)**: Mesopelagic zone (from 500/1000 meters)
- Stratification influencée plus par la température que par la géographie ou autres facteurs environnementaux (Sunagawa *et al.*, 2015)
- Source des données : [gitlab MiBiOmics]

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir
FRANCE

cnrs INRAE Inserm cea INVESTIR L'AVENIR