Computational biology laboratory
CHU de Québec – Université Laval

Current Projects

Arnaud Droit, PhD

CHU de Québec

CHAIRE DE RECHERCHE ET D'INNOVATION L'ORÉAL EN BIOLOGIE NUMÉRIQUE
AFFILIÉ À UNIVERSITÉ LAVAL

AD Lab

UNIVERSITÉ LAVAL

Québec


Canada
Ontario
Québec
Terre-Neuve-
et-Labrador
Île-du-Prince-
Édouard
Nouveau-
Brunswick
Nouvelle-Écosse
Rivière-du-Loup
Québec
Montréal
États-Unis

Ville de Québec

Université Laval – Ville de Québec

Research Center - CHUL

The Team (Not complete)

Mickael Leclercq, PhD

Christophe Tav (Doctorant)

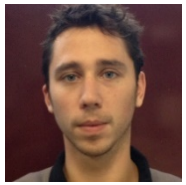
Charles Joly Beauparlant, PhD

Eric Fournier

Julien Prunier, PhD

Kodjovi Dodji Mlaga, PhD

Régis Ongaro (Doctorant)

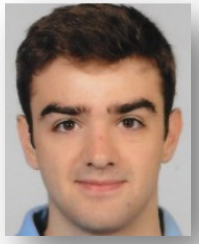


Alban Mathieu

Loic Mangnier

Marie Pier Scott Boyer, PhD

Frédéric Fournier



Antoine Bodein (Doctorant)

Francois Belleau

Afshin Jamshidi, PhD

Kawhla Seddiki (Doctorante)

Gwenaëlle Lemoine (Doctorante)

6

Computational biology laboratory

- 8 research associates (Developers, Bioinformaticians, Biostatisticians, Data scientists…)
- 7 PhD students
- 2 Post-Doc
- some Interns

- Main projects: Multi-OMICS Integrations, Exome sequencing, Epigenetics, Proteomics, Automatic classification

- Linked to Compute Canada infrastructure

Main activities and expertise

- We develop tools and strategies dedicated to
  - Multi-omics data analysis
  - Discovery of biomarker signatures using machine learning approaches
  - Elaboration of deep learning predictive models
  - Integration, exploration and visualization of biological big data
- We offer services to biologists through collaborations to analyse various types of omic data
  - Genomic (WGS, WES), Transcriptomic, Proteomic, Metabolomic, Epigenomic, Meta-genomic (microbiota)
- We have an expertise in the analysis of data in various diseases
  - Alzheimer, Multiple sclerosis, Hormonodependant cancers (Breast, Prostate), Leukemia, Fragile X syndrome, etc.
- L'Oréal Research and Innovation Chair in Digital Biology
  - Characterization of artificial skin reconstruction using Time-dependant multi-omics, including skin microbiota
- Currently in the Bioinformatics laboratory
  - 23 people (1 Professor, 7 Research associates, 6 Phds, 1 master, 2 postdoc, 5 interns, 1 sysadmin)
  - About 20 research projects

Our infrastructure

Computing resources

- **10** servers
- **350 TB** of storage
- **600** cpu cores
- **3 TB** of RAM

Dedicated servers

- **12** Elasticsearch servers for big data management
- **4** virtualisation servers

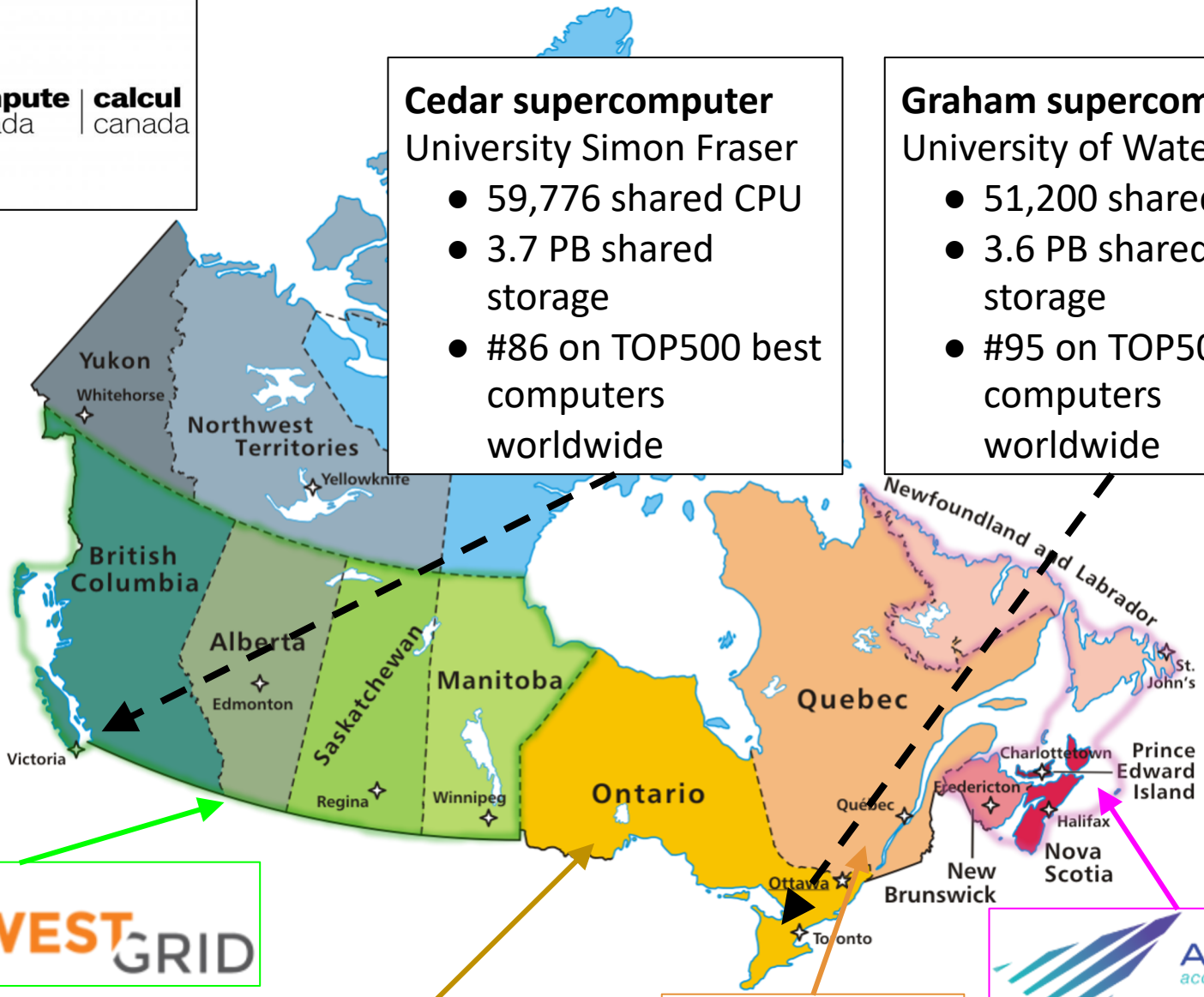
compute canada | calcul canada

Cedar supercomputer
University Simon Fraser

- 59,776 shared CPU
- 3.7 PB shared storage
- #86 on TOP500 best computers worldwide

Graham supercomputer
University of Waterloo

- 51,200 shared CPU
- 3.6 PB shared storage
- #95 on TOP500 best computers worldwide



WESTGRID

compute calcul ONTARIO

Calcul Québec

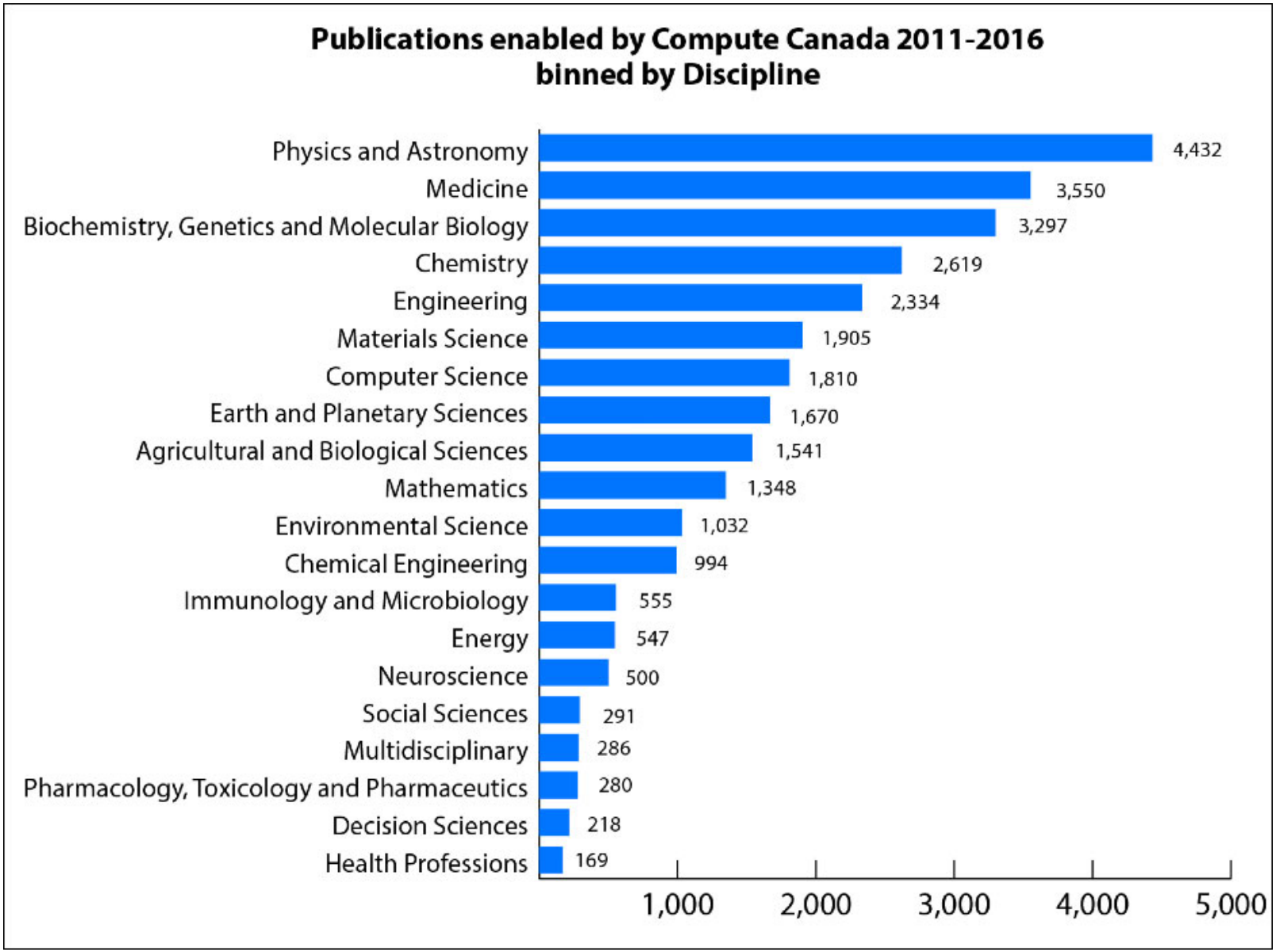
ACENET
accelerate discovery

Compute Canada allocation

- **78 cpu-year** in the beluga-compute cluster
- **300 TB** of PROJECT storage space in the ndc-calculquebec cluster
- **900 TB** of NEARLINE storage space in the ndc-calculquebec cluster

- **98 years-VCPU** in the arbutus-persistent-cloud cluster
- **10 TB** of INFONUAGIC storage space in the arbutus-persistent-cloud cluster
- **375 GB** of RAM in the arbutus-persistent-cloud cluster
- **11** volumes in the arbutus-persistent-cloud cluster
- **2** floating IP addresses in the arbutus-persistent-cloud cluster

Publications enabled by Compute Canada

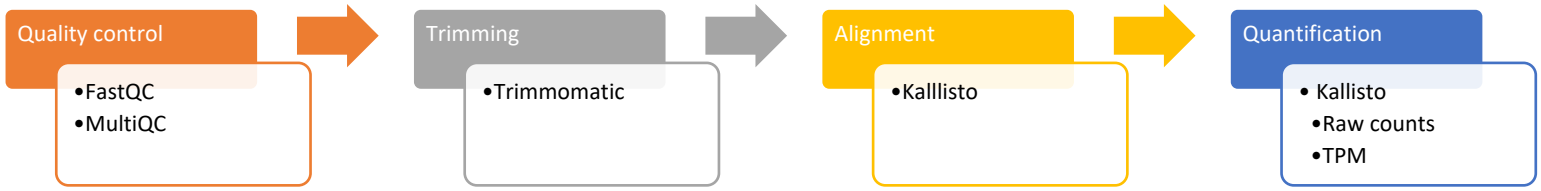


Transcriptomics analyses from the Next Generation Sequencing Lab

Transcriptomics/miRomics pipelines

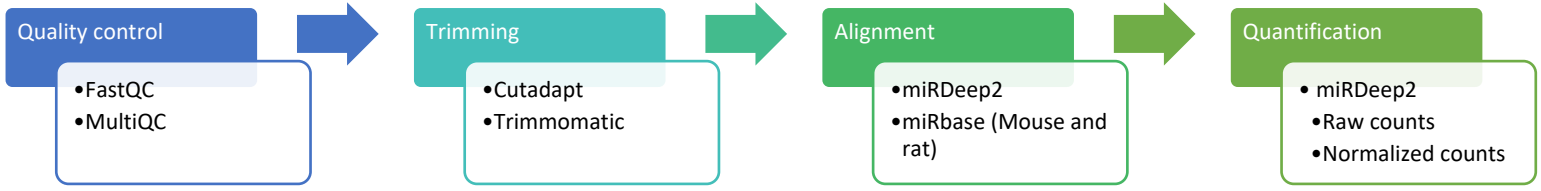
1. mRNA quantification

First, a quality control was performed using FastQC. Then, sequence reads were trimmed using Trimmomatic and aligned using Kallisto, which also produced gene quantification.



2. miRNA quantification

A quality control was performed using FastQC. Then, sequence reads were trimmed using cutadapt and trimmomatic, and were aligned with miRDeep2 against mouse non-coding RNAs. Then, miRNAs were quantified by miRdeep2 using rat and mouse mature miRNAs and pre-miRNAs. Rat and mouse miRNAs are from very close species, then it may be useful to map on these two species.



R packages

Bioconductor (R) packages development

- **Metagene**

Compare the behavior of DNA-interacting proteins

- **Imetagene**

A graphical interface for the metagene package

- **ENCODExplorer**

Compilation of ENCODE metadata

- **similaRpeak**

calculates metrics which assign a level of similarity between ChIP-Seq profiles

- **Nucleosim**

Generate synthetic nucleosome maps

- **consensusSeekeR**

Detection of consensus regions inside a group of experiences using genomic positions and genomic ranges

- **RJMCMC**

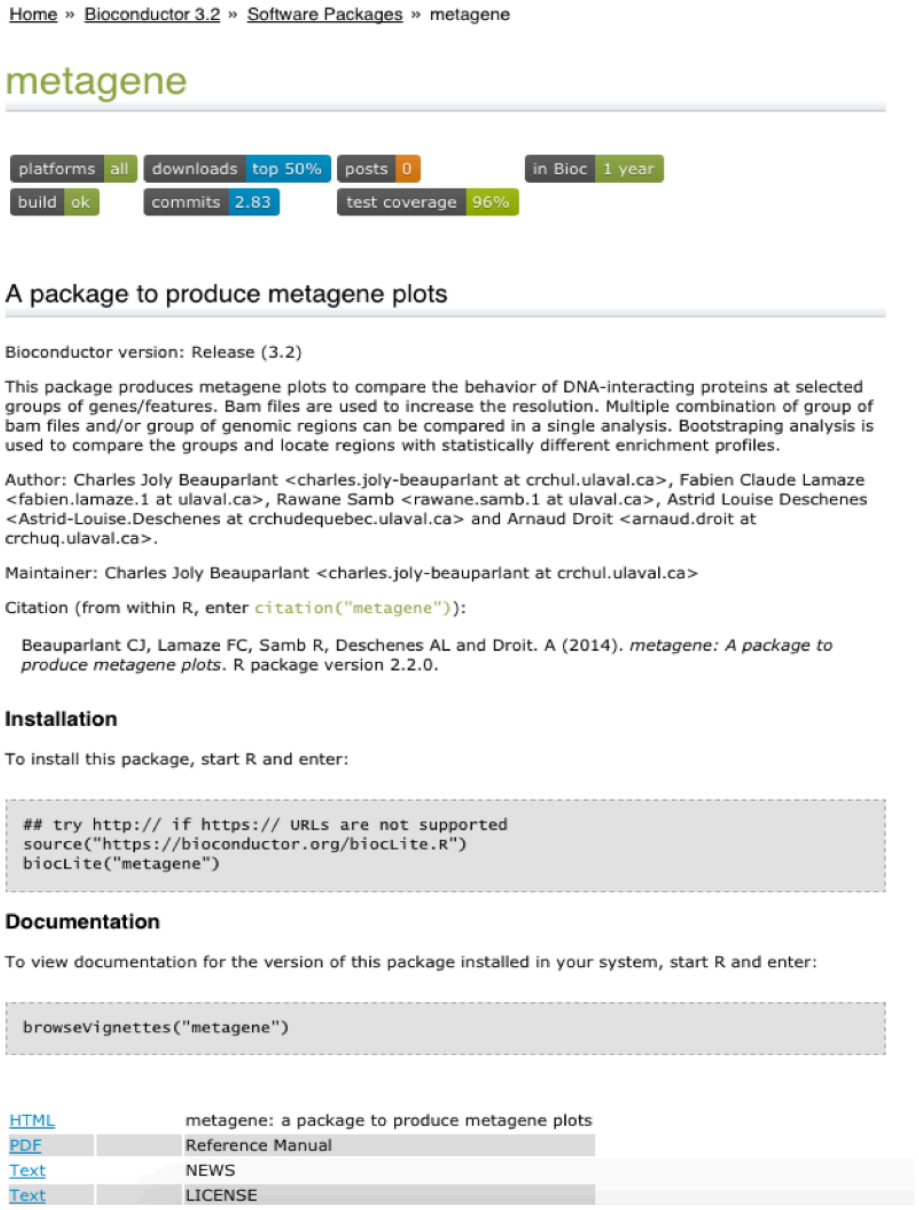
Estimation of nucleosome positions for genome-wide profiling

- **TimeOmics**

Integrate multi-Omics longitudinal data measured on the same biological samples and select key temporal features

- **Gwena**

Gene co-expression network analysis and explore the results in a single pipeline



Home » Bioconductor 3.2 » Software Packages » metagene

metagene

platforms all | downloads top 50% | posts 0 | in Bioc 1 year
build ok | commits 2.83 | test coverage 96%

A package to produce metagene plots

Bioconductor version: Release (3.2)

This package produces metagene plots to compare the behavior of DNA-interacting proteins at selected groups of genes/features. Bam files are used to increase the resolution. Multiple combination of group of bam files and/or group of genomic regions can be compared in a single analysis. Bootstraping analysis is used to compare the groups and locate regions with statistically different enrichment profiles.

Author: Charles Joly Beauparlant <charles.joly-beauparlant at crchul.ulaval.ca>, Fabien Claude Lamaze <fabien.lamaze.1 at ulaval.ca>, Rawane Samb <rawane.samb.1 at ulaval.ca>, Astrid Louise Deschenes <Astrid-Louise.Deschenes at crchudequebec.ulaval.ca> and Arnaud Droit <arnaud.droit at crchuq.ulaval.ca>.

Maintainer: Charles Joly Beauparlant <charles.joly-beauparlant at crchul.ulaval.ca>

Citation (from within R, enter citation("metagene")):

Beauparlant CJ, Lamaze FC, Samb R, Deschenes AL and Droit. A (2014). *metagene: A package to produce metagene plots*. R package version 2.2.0.

Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("metagene")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("metagene")
```

HTML		metagene: a package to produce metagene plots
PDF		Reference Manual
Text		NEWS
Text		LICENSE

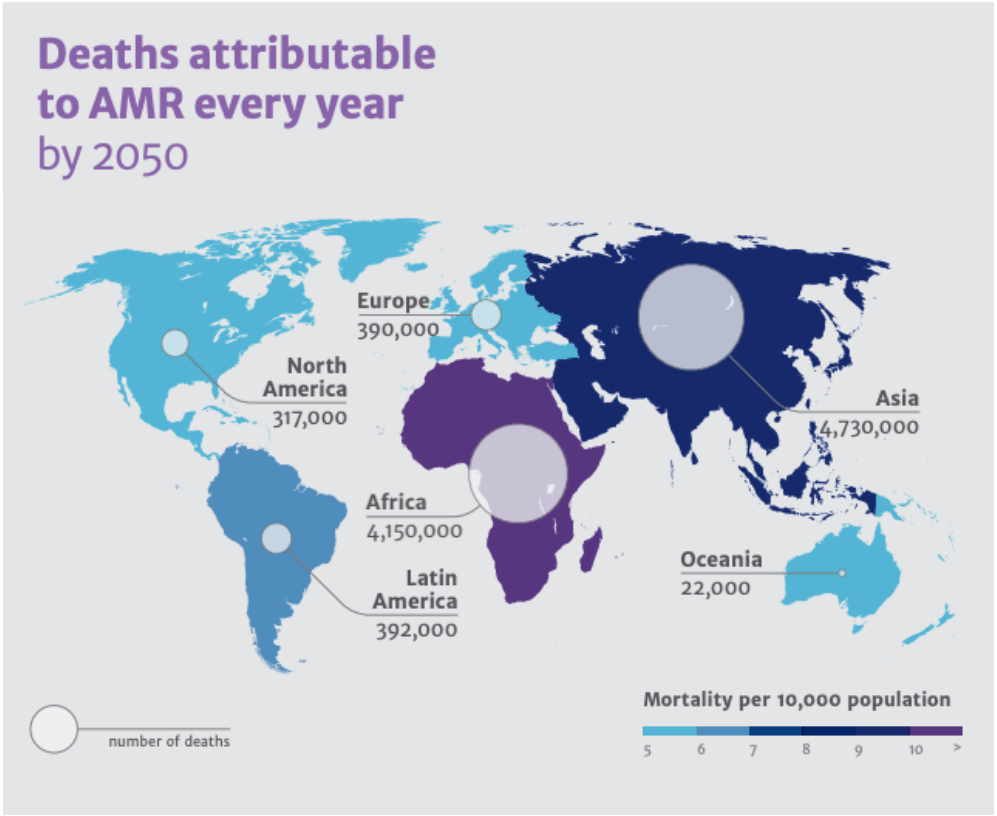
Antibiotics resistance

World Health Organization

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

efsa
European Food Safety Authority

AURA
Antimicrobial Use and Resistance in Australia



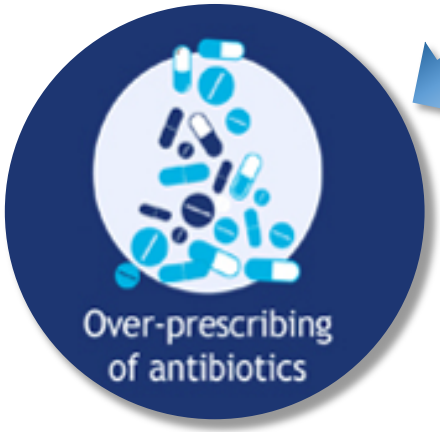
The next global challenge

The causes of antibiotic resistance

The causes of antibiotic resistance

STOP OVERUSE AND MISUSE OF ANTIBIOTICS COMBAT RESISTANCE

HANDLE ANTIBIOTICS WITH CARE



CAUSES OF ANTIBIOTIC RESISTANCE

HANDLE ANTIBIOTICS WITH CARE

Antibiotic resistance happens when bacteria change and become resistant to the antibiotics used to treat the infections they cause.

- Over-prescribing of antibiotics
- Patients not finishing their treatment
- Over-use of antibiotics in livestock and fish farming
- Poor infection control in hospitals and clinics
- Lack of hygiene and poor sanitation
- Lack of new antibiotics being developed

www.who.int/drugresistance

#AntibioticResistance

World Health Organization

Bacterial identification in clinics

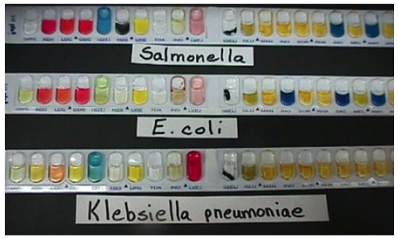
Biological fluid

Bacterial Culture

Biochemical tests

≈ 100h

Bacterial culture on gelose plate



Biochemical tests (API strips)

Suitable antibiotic treatment

Bacterial identification in clinics

Biological fluid

Bacterial Culture → Biochemical tests

≈ 100h

Suitable antibiotic treatment

Broad-spectrum antibiotics

- Sometimes not needed
- Sometimes not efficient
- Increases Bacterial Resistances +++

Bacterial identification in clinics

Biological fluid

Bacterial Culture → Biochemical tests ≈ 100h

Bacterial Culture → MALDI-TOF-MS 24-48h

Suitable antibiotic treatment

Bacterial culture on gelose plate

MALDI-TOF (Biotyper, Vitek-MS)

Bacterial identification in clinics

Biological fluid

Bacterial Culture → Biochemical tests ≈ 100h

Bacterial Culture → MALDI-TOF-MS 24-48h

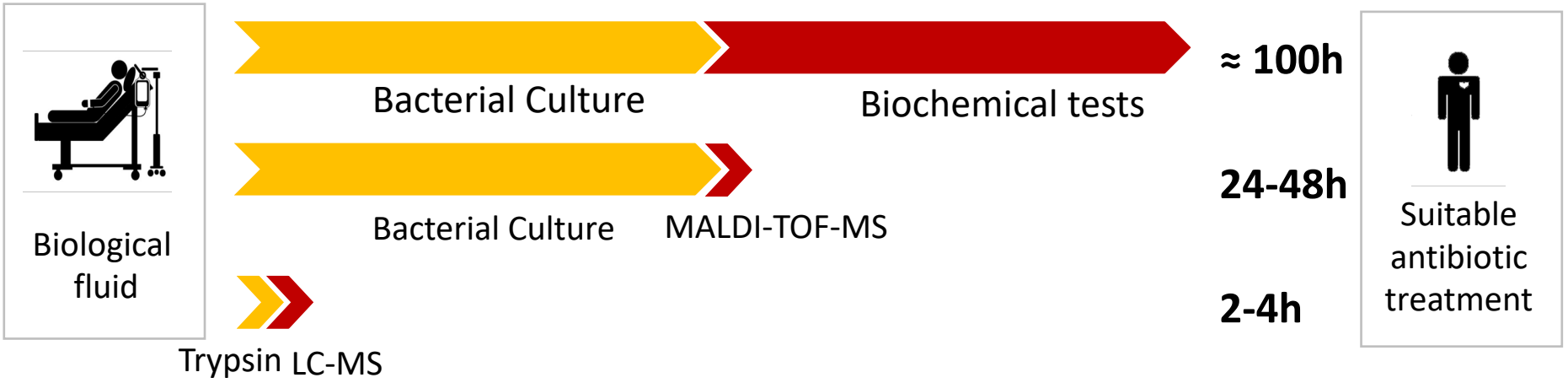
Broad-spectrum antibiotics

Bacterial culture on gelose plate

MALDI-TOF (Biotyper, Vitek-MS)

Suitable antibiotic treatment

Bacterial identification in clinics



Bacterial identification in clinics

Biological fluid

Bacterial Culture → Biochemical tests — ≈ 100h

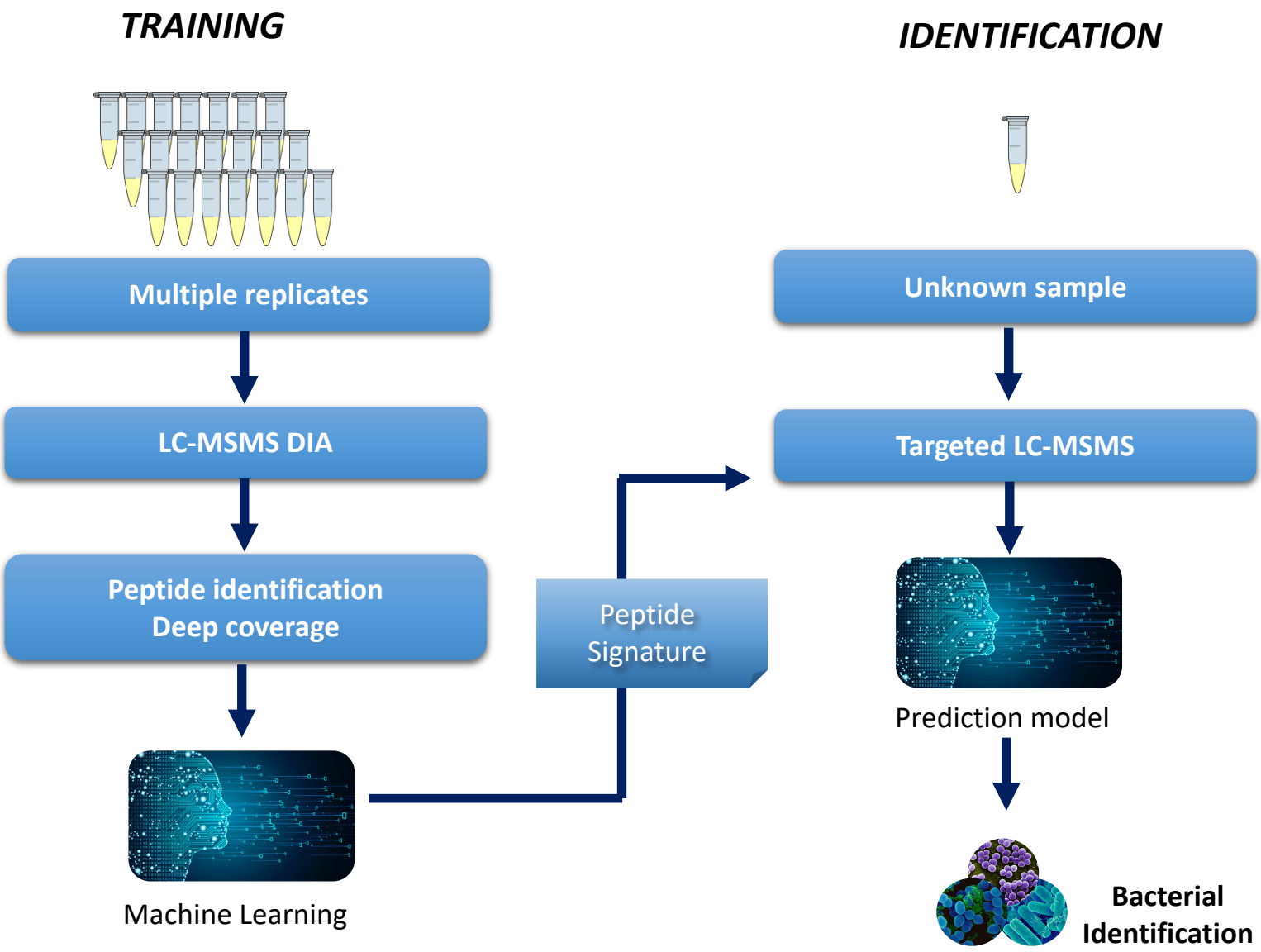
Bacterial Culture → MALDI-TOF-MS — 24-48h

Trypsin → LC-MS — 2-4h

Suitable antibiotic treatment

Broad-spectrum antibiotics ✗

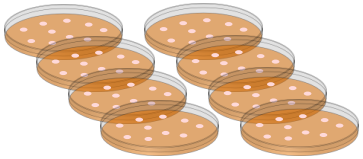
Bacterial Identification by LC-MSMS + ML



Roux-Dalvai *et al.* Mol Cell Proteomics, 2019

TRAINING STEP : Deep Proteome Coverage

Bacterial
15 species

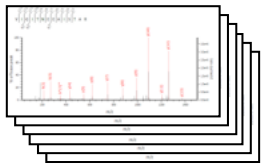


Protein extraction
Trypsin digestion
Peptide fractionation

Orbitrap Fusion

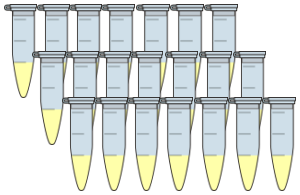
LC-MSMS
DDA

Peptide Identification



Spectral library
31096 peptides

Inoculation replicates
15 species – 190 samples

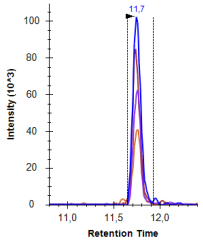


Short sample preparation

Orbitrap Fusion

LC-MSMS
DIA

Skyline



List of observable peptides

4319 peptides

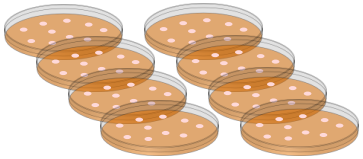
Machine Learning

BioDisc ML

Leclercq M. *et al.*, Front Genet. 2019 May 16;10:452.

TRAINING STEP : Deep Proteome Coverage

Bacterial
15 species

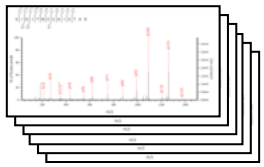


Protein extraction
Trypsin digestion
Peptide fractionation

Orbitrap Fusion

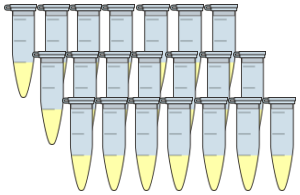
LC-MSMS
DDA

Peptide Identification



Spectral library
31096 peptides

Inoculation replicates
15 species – 190 samples

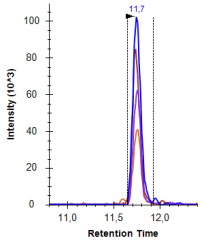


Short sample preparation

Orbitrap Fusion

LC-MSMS
DIA

Skyline



List of observable peptides

4319 peptides

Machine Learning

BioDisc ML

- 80 Machine Learning classifiers
- Hyper-parameters sets
- Feature selection algorithms

≃ 5000 models

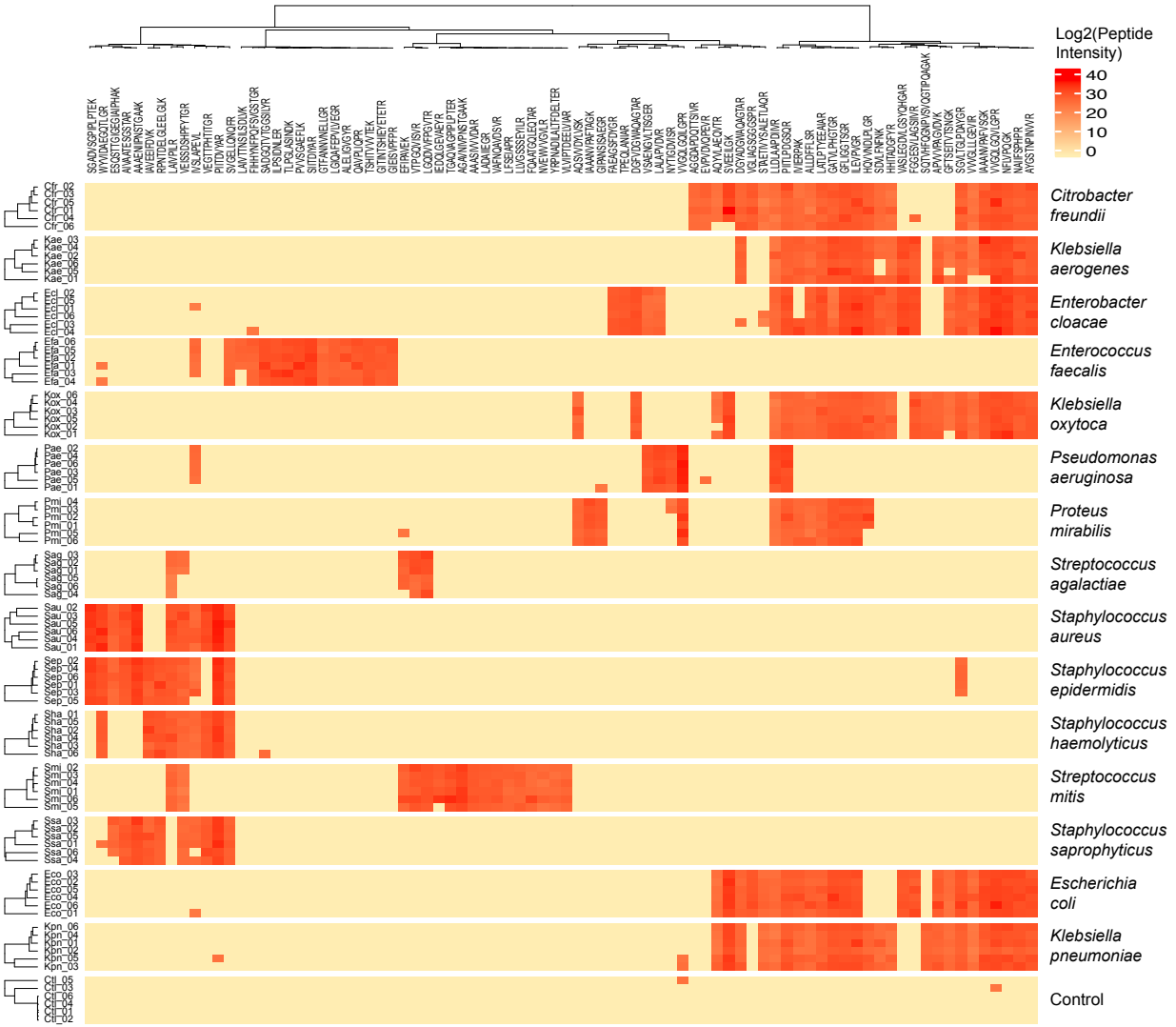
multiple cross-validations

Leclercq M. *et al.*, Front Genet. 2019 May 16;10:452.

TRAINING STEP : Peptide signature

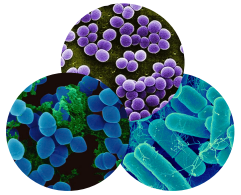
BioDiscML

82 Peptides Signature



- Peptides shared by several species
- Each species has its own pattern

Conclusion and perspectives

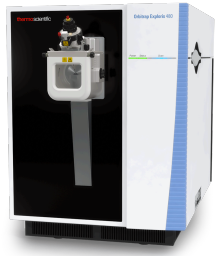


- LC-MS and Machine Learning allow the identification of bacteria in UTI in a short time without culture

- The method can work on triple quadrupole instruments

  - Development of quantification method for UTI on TSQ Altis
  - Development of new signatures (blood, milk, water….)
  - Detection of bacteria having specific resistances or virulences

- Improvement of the strategy using crude DIA signal and Machine Learning



Assemblage de génome du caribou des bois (*Rangifer tarandus caribou*)

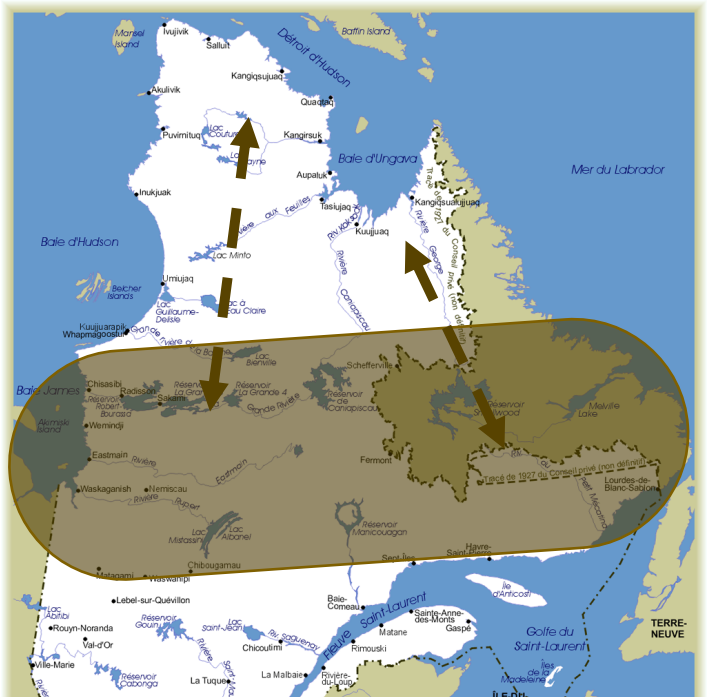
Le Caribou

- Le caribou (*Rangifer tarandus*)
- Plusieurs ss-espèces:
  - De Grant
  - De la toundra
  - Des îles du Nunavut et des TNO
  - De l'archipel de la Reine-Charlotte (éteint)
  - Des bois
- Plusieurs écotypes au sein des "bois"
  - Migrateur (ou toundrique)
  - Forestier
  - Montagnard

Alimentation et migration

- Alimentation: lichen dans des forêts de faible densité

- Migration:
  - Hiver en taiga (forêt)
  - Été en toundra
  - Deux grandes hardes au Québec

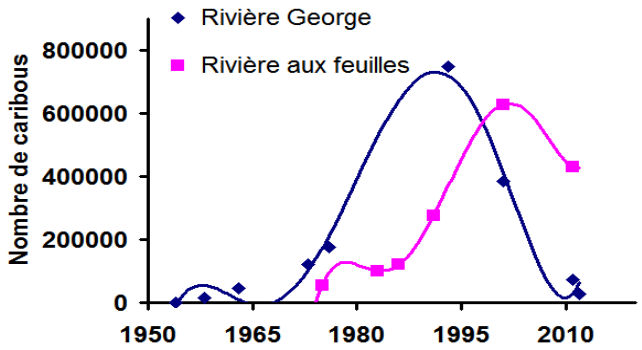


Menaces

- Exploitation minière et forestière

- Changements climatiques

- Baisse des tailles de populations



Menaces

- Exploitation minière et forestière

- Changements climatiques

Nécessités de protection:

> chasse interdite au Québec depuis 01/02/2018

> exception: hardes migratrices par les 1ères nations

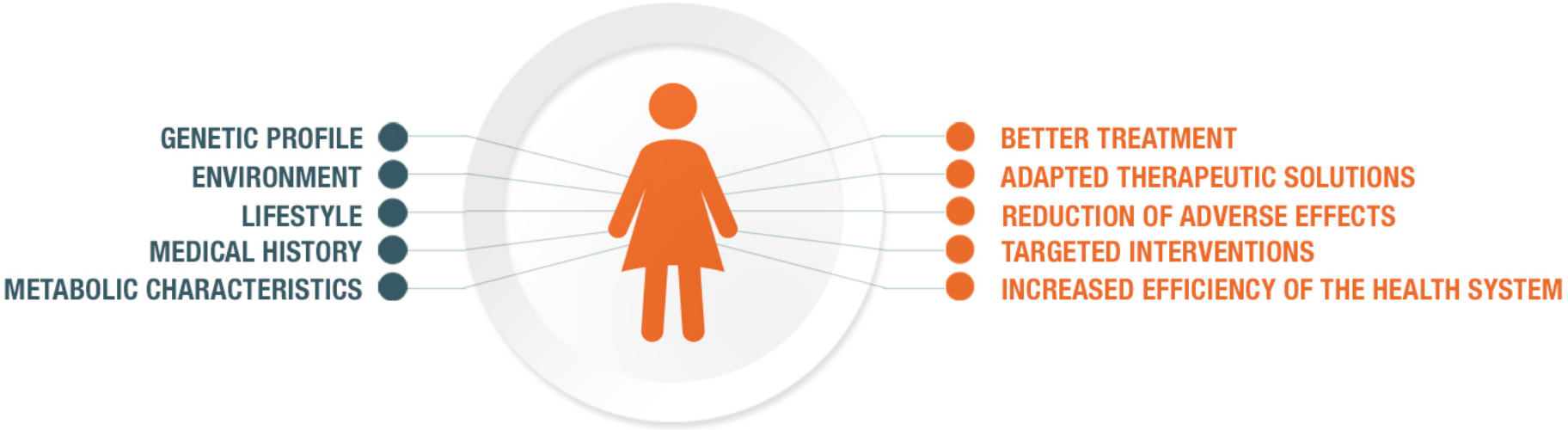
Objectifs du projet

- Designer une biopuce à SNPs
  - Vision à long terme => polyvalence
  - Besoin d'un génome de haute qualité
- Définir les populations au niveau génétique (structure hiérarchique de populations)
- Mettre au point un outil de forensic pour determiner la provenance d'un échantillon en cas de procès pour braconnage.

Personalized medicine
Breast Cancer

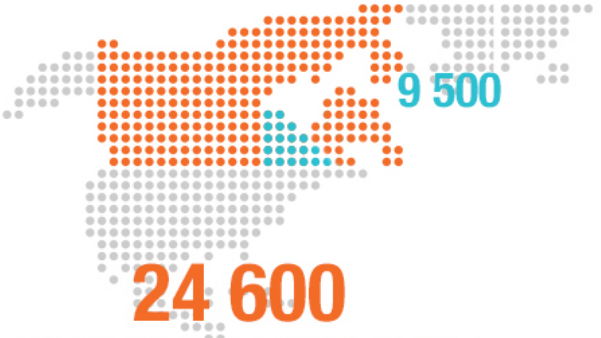
PERSONALISED HEALTHCARE

- PERSONALISED
- PREDICTIVE
- PREVENTIVE
- PARTICIPATIVE



BREAST CANCER BURDEN IN CANADA

1/9 WOMEN WILL DEVELOP
BREAST CANCER
DURING HER LIFE



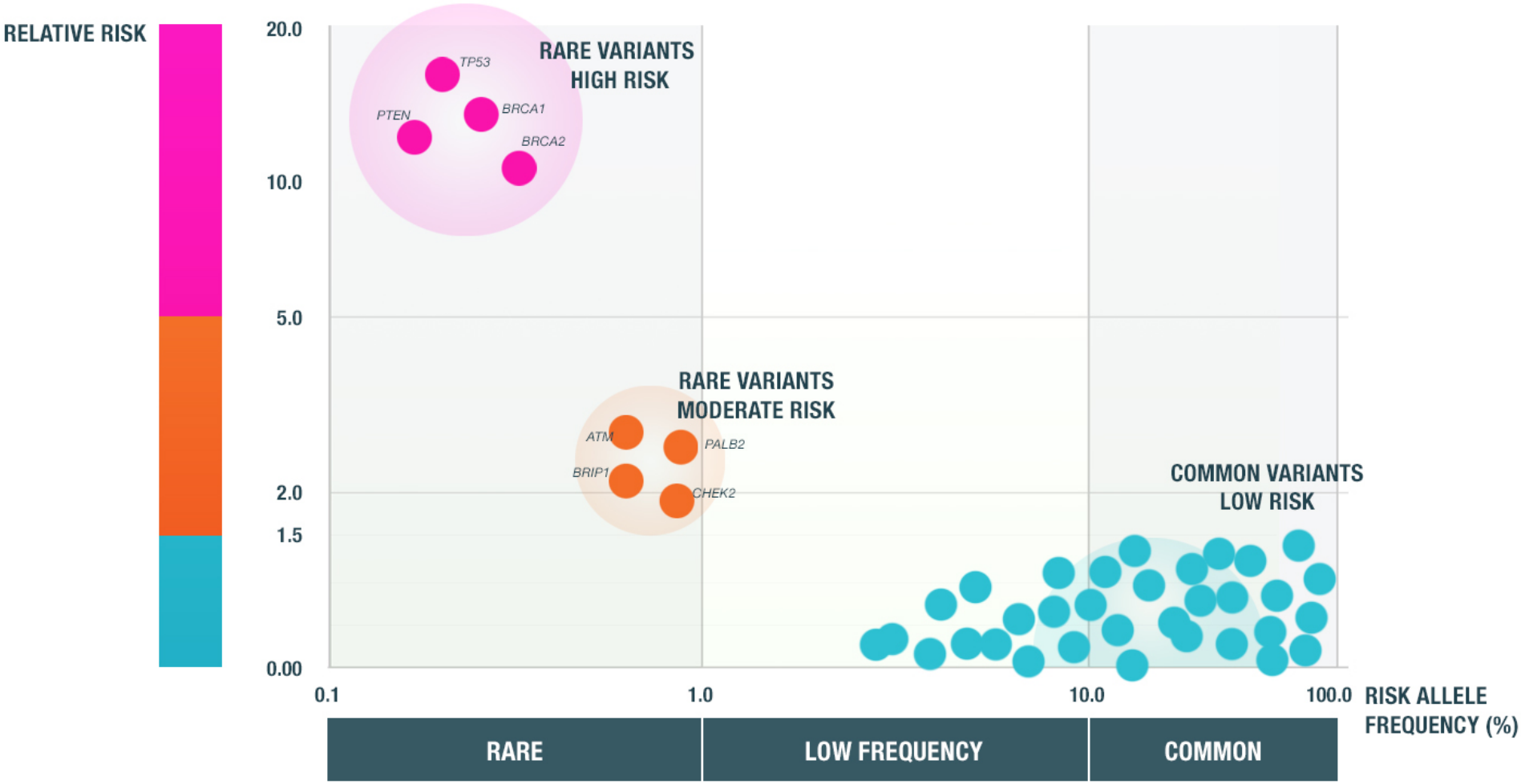
24 600
NEW CASES EACH YEAR

5 000
DEATHS ARE ATTRIBUTED TO
THIS CANCER EACH YEAR

1/5 CASE
OCCUR IN WOMEN < 50 YEARS OLD

36% OF ALL CANCERS DIAGNOSED IN WOMEN
BETWEEN THE AGES OF 30-49

GENETIC VARIATIONS ASSOCIATED WITH BREAST CANCER RISK



GENETIC RISK PROFILE

HUNDRED OF RESEARCH TEAMS FROM

40

COUNTRIES

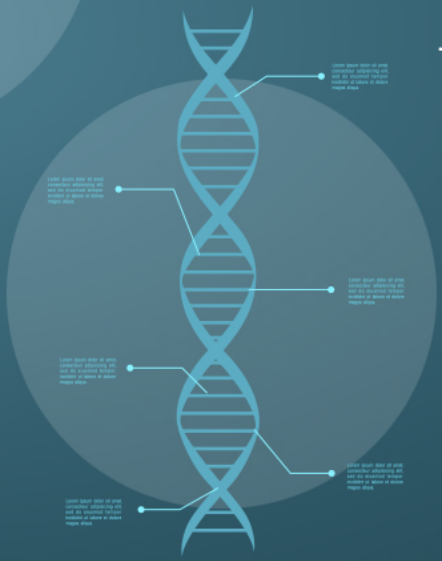
"ONCOARRAY"

150 000

PARTICIPANTS

INTEGRATION OF DATA

MEDICAL HISTORY
TUMOR PATHOLOGY
TREATMENT RESPONSE
ENVIRONMENT
FAMILY HISTORY

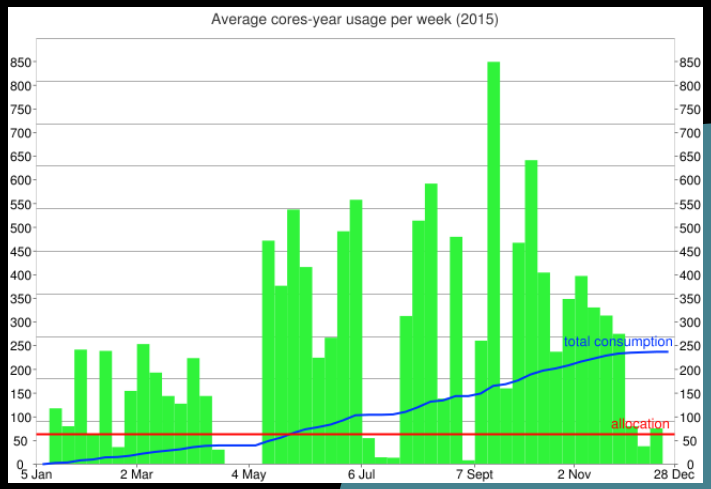


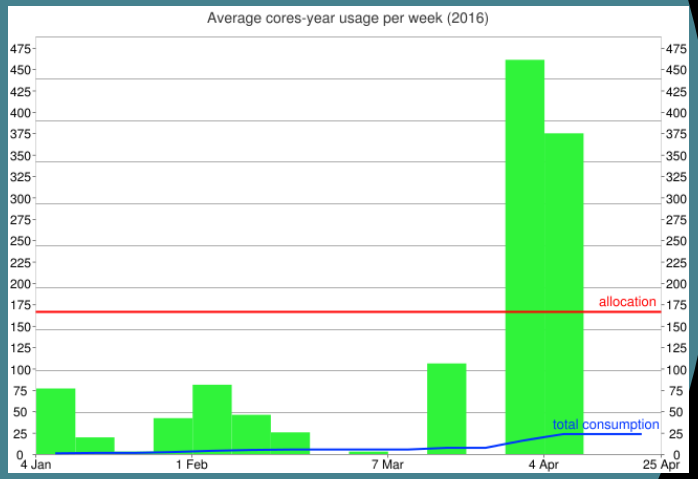
ANALYSIS OF **GENOMIC** DATA

Computational genomics challenges

Royaume-Uni
Pays-Bas
Allemagne
Canada (ON, QC)
États-Unis (UT)

Projet international



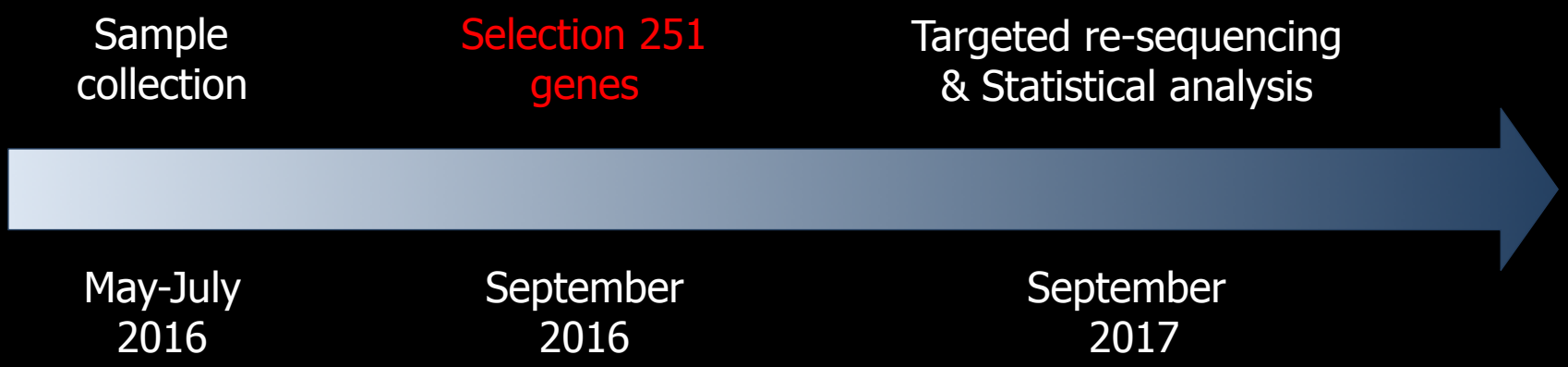


5 050 échantillons

Séquençage Massivement Parallèle

Consommation 2015 - mi 2016 : ~300 coeurs-année
= 150 ans de calcul
sur un ordinateur personnel moyen

Timeline Perspective validation

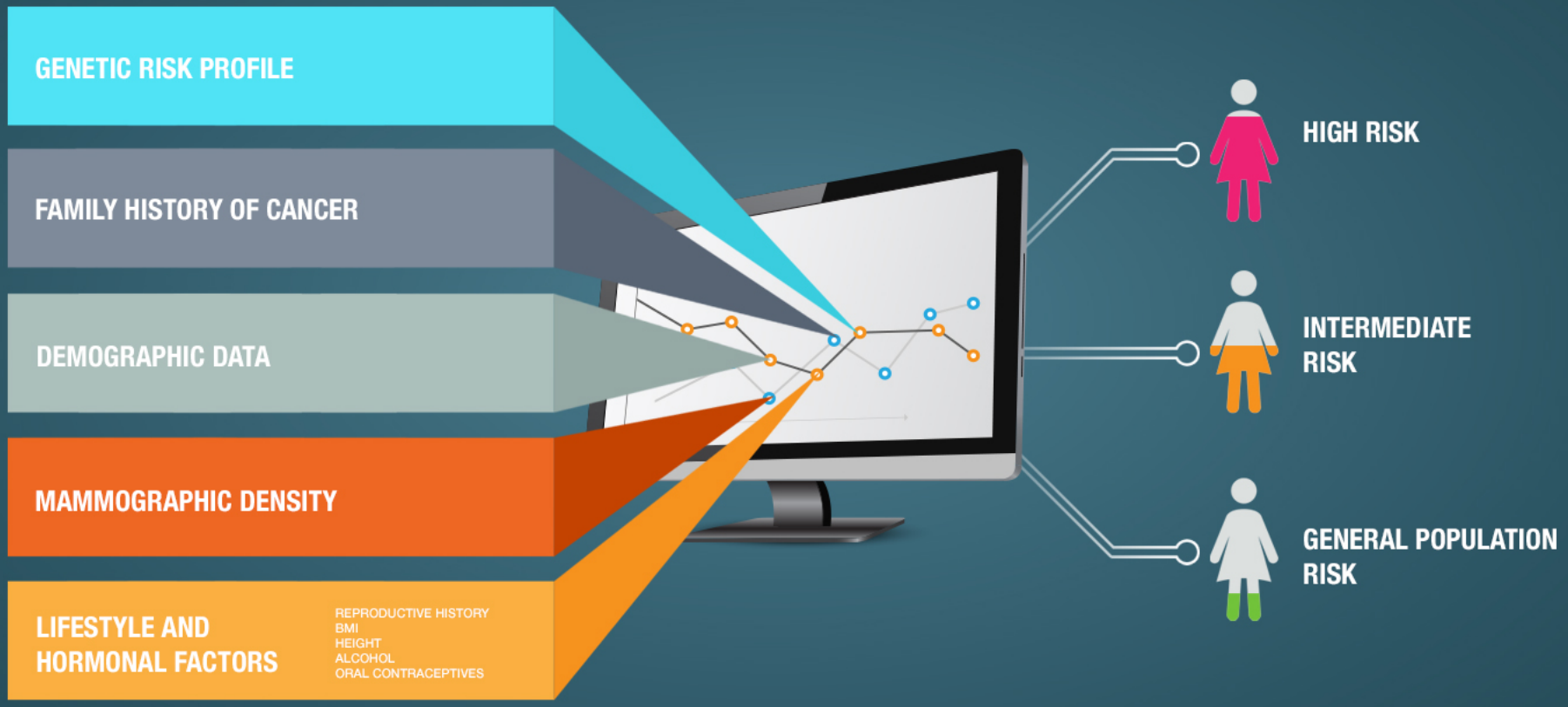


GO Results - 251 genes

Y a-t-il des fonctions, cible d'intérêt dans le cancer ?

	Homo sapiens (REF)	upload 1 (▽ Hierarchy NEW! ?)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
strand displacement	26	13	.33	39.13	+	4.45E-13
↳DNA recombination	226	24	2.89	8.31	+	4.17E-11
↳DNA metabolic process	787	55	10.06	5.47	+	7.83E-21
↳cellular macromolecule metabolic process	6789	124	86.76	1.43	+	1.25E-02
↳macromolecule metabolic process	7494	133	95.77	1.39	+	1.83E-02
↳nucleic acid metabolic process	3987	82	50.95	1.61	+	3.03E-02
↳nucleobase-containing compound metabolic process	4541	93	58.03	1.60	+	5.38E-03
↳heterocycle metabolic process	4678	93	59.78	1.56	+	2.05E-02
↳cellular aromatic compound metabolic process	4735	94	60.51	1.55	+	1.86E-02
DNA double-strand break processing	19	8	.24	32.95	+	1.82E-06
↳double-strand break repair	161	25	2.06	12.15	+	1.90E-15
↳DNA repair	487	48	6.22	7.71	+	6.25E-24
↳cellular response to DNA damage stimulus	735	54	9.39	5.75	+	2.22E-21
↳cellular response to stress	1618	76	20.68	3.68	+	7.94E-20
↳cellular response to stimulus	6211	144	79.37	1.81	+	1.37E-12
↳response to stimulus	7683	153	98.18	1.56	+	6.97E-08
↳response to stress	3365	98	43.00	2.28	+	2.69E-12

RISK PROFILING



STRATÉGIES DE RECHERCHE

NOS OBJECTIFS PRINCIPAUX SONT:

- Améliorer l'évaluation personnalisée du risque afin d'offrir un dépistage et une prévention coût-efficace du cancer du sein aux femmes les plus susceptibles d'en bénéficier.

- Déterminer les approches de mise en œuvre optimales dans le système de santé canadien.

CLINIQUES D'ONCOGÉNÉTIQUE

ACTIVITÉ 1

IDENTIFICATION DE NOUVEAUX GÈNES DE SUSCEPTIBILITÉ AU CANCER DU SEIN

10 000 CAS — 10 000 TÉMOINS

SÉQUENÇAGE D'EXOMES*

50 000 CAS — 50 000 TÉMOINS

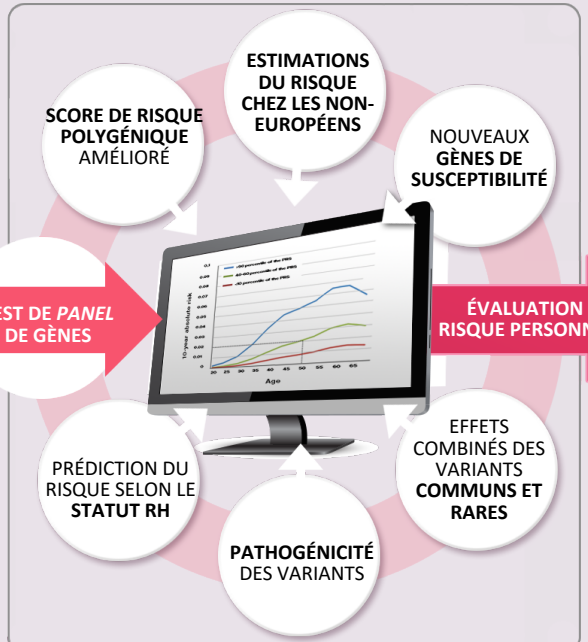
CRIBLAGE DE GÈNES CANDIDATS

ESTIMATIONS PRÉCISES DU RISQUE ASSOCIÉ AUX GÈNES DE SUSCEPTIBILITÉ AU CANCER DU SEIN

* **ÉVALUATION FONCTIONELLE DE VARIANTS RARES**

TEST DE *PANEL* DE GÈNES

ACTIVITÉ 2



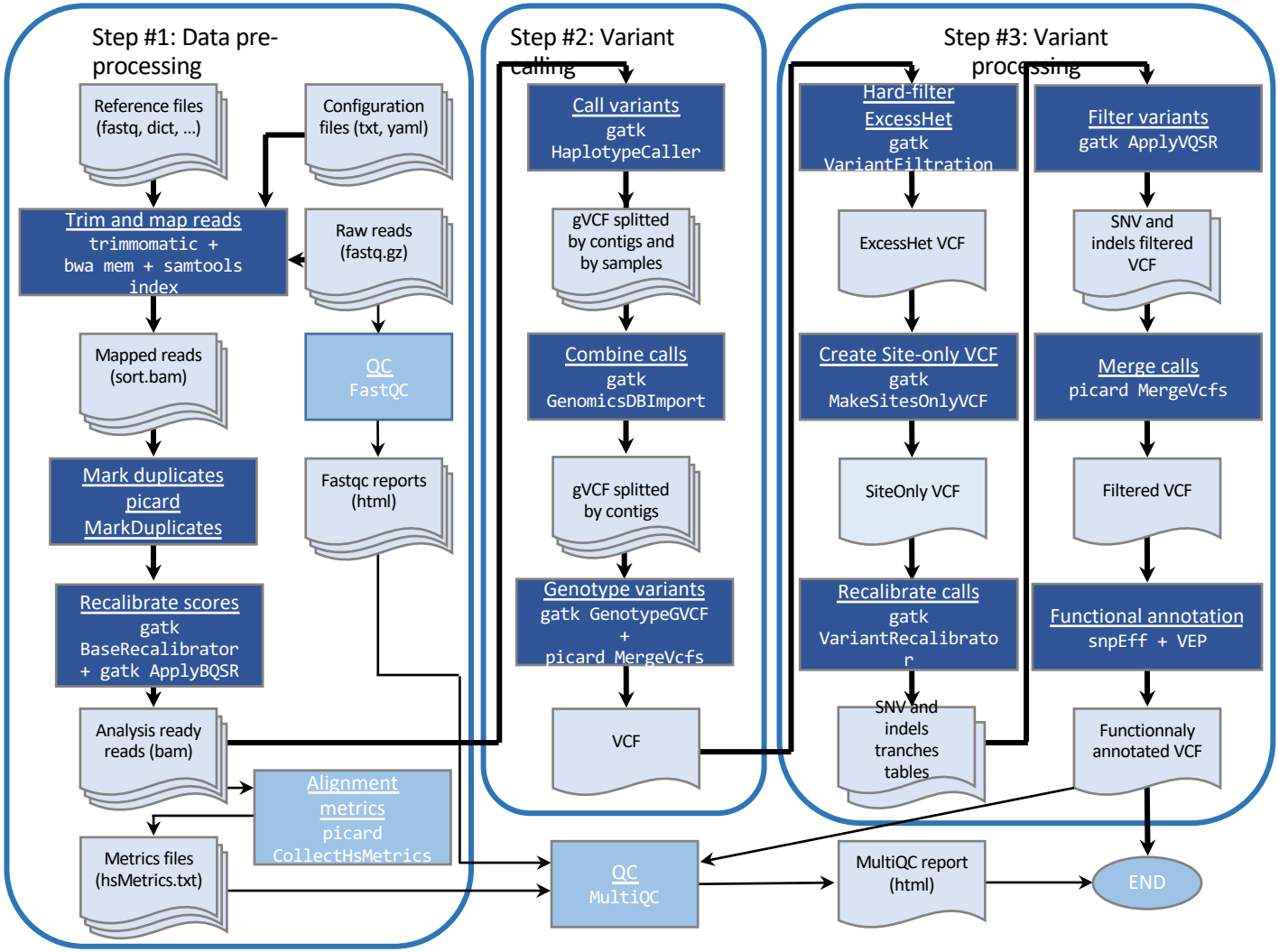
ÉVALUATION DU RISQUE PERSONNALISÉ

PRÉDICTION DU RISQUE **PLUS PRÉCISE**

AMÉLIORER:

- CONSEIL GÉNÉTIQUE
- PROCESSUS DE DÉCISION PARTAGÉE À PROPOS DU **DÉPISTAGE ET DES STRATÉGIES DE RÉDUCTION DU RISQUE**
- ÉFFICACITÉ DE LA **PRÉVENTION PRIMAIRE**

* en collaboration avec le projet international **BRIDGES**


Step #1: Data pre-processing
Reference files (fastq, dict, ...)
Configuration files (txt, yaml)
Trim and map reads
trimmomatic + bwa mem + samtools index
Raw reads (fastq.gz)
Mapped reads (sort.bam)
QC
FastQC
Mark duplicates
picard MarkDuplicates
Fastqc reports (html)
Recalibrate scores
gatk BaseRecalibrator + gatk ApplyBQSR
Analysis ready reads (bam)
Alignment metrics
picard CollectHsMetrics
Metrics files (hsMetrics.txt)
Step #2: Variant calling
Call variants
gatk HaplotypeCaller
gVCF splitted by contigs and by samples
Combine calls
gatk GenomicsDBImport
gVCF splitted by contigs
Genotype variants
gatk GenotypeGVCF + picard MergeVcfs
VCF
Step #3: Variant processing
Hard-filter ExcessHet
gatk VariantFiltration
ExcessHet VCF
Create Site-only VCF
gatk MakeSitesOnlyVCF
SiteOnly VCF
Recalibrate calls
gatk VariantRecalibrator
SNV and indels tranches tables
Filter variants
gatk ApplyVQSR
SNV and indels filtered VCF
Merge calls
picard MergeVcfs
Filtered VCF
Functional annotation
snpEff + VEP
Functionnaly annotated VCF
QC
MultiQC
MultiQC report (html)
END

Multi-omics analyses in Prostate Cancer

Prostate cancer

- Prostate cancer affects about 1 in 7 men
- Cancer most common specific sex in men
- Patients are almost always brought to prostatectomy
- Biochemical recurrence may occur

What therapeutics strategies ?

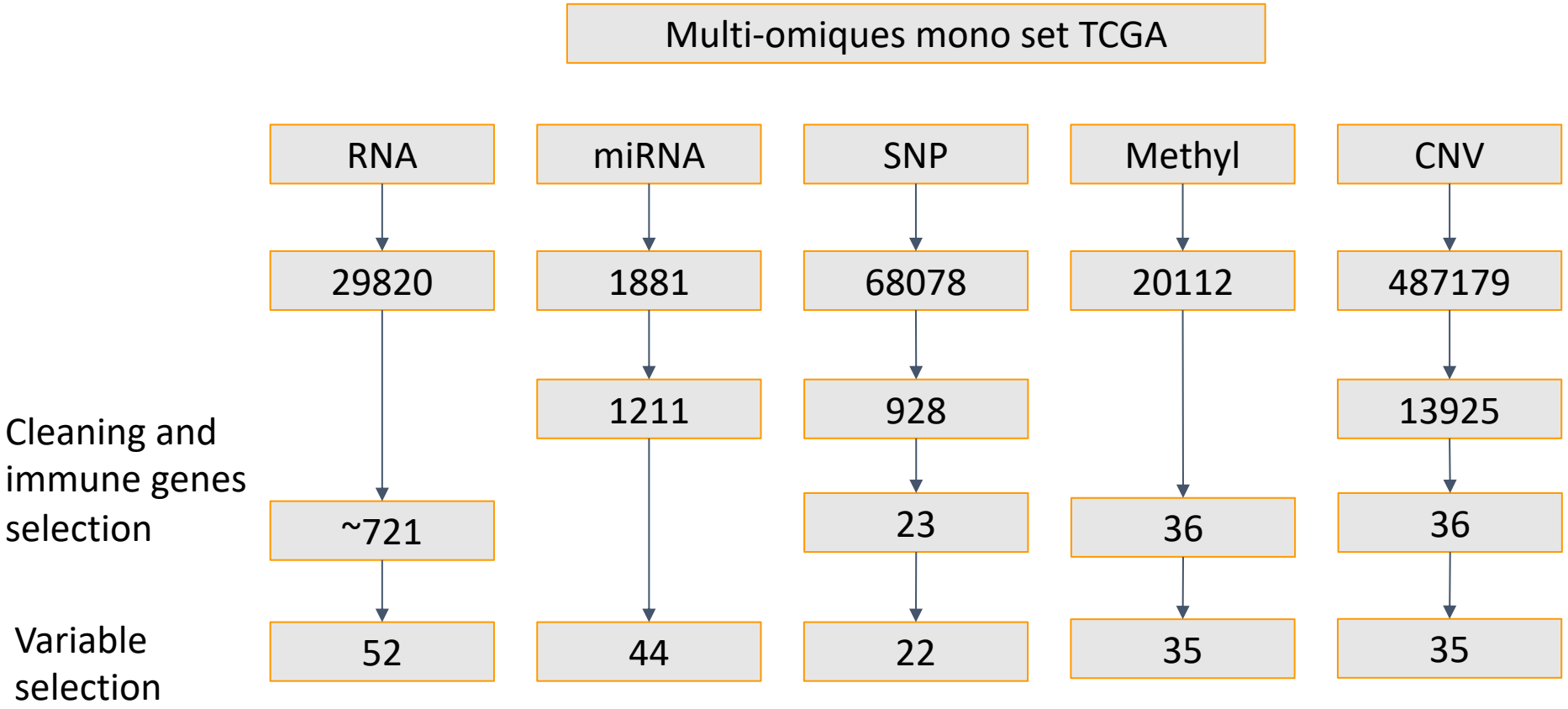
- prostatectomy
- Hormonotherapy
- immunotherapy

Study objectives

Hypothesis: The use of different omics datasets will allow to find recurring and significant biological markers related to the immunity to propose new immunological targets.

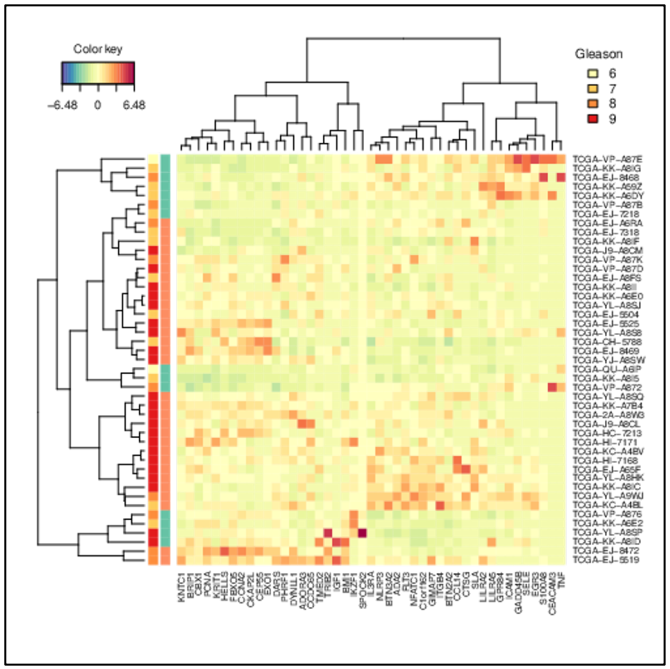
Objectives: 1) Retrieve datasets 2) Treat them uniformly 3) Target biomarkers of immunity 4) Choose appropriate statistical techniques

Data curation and variable selection with MixOmics



mRNA signature

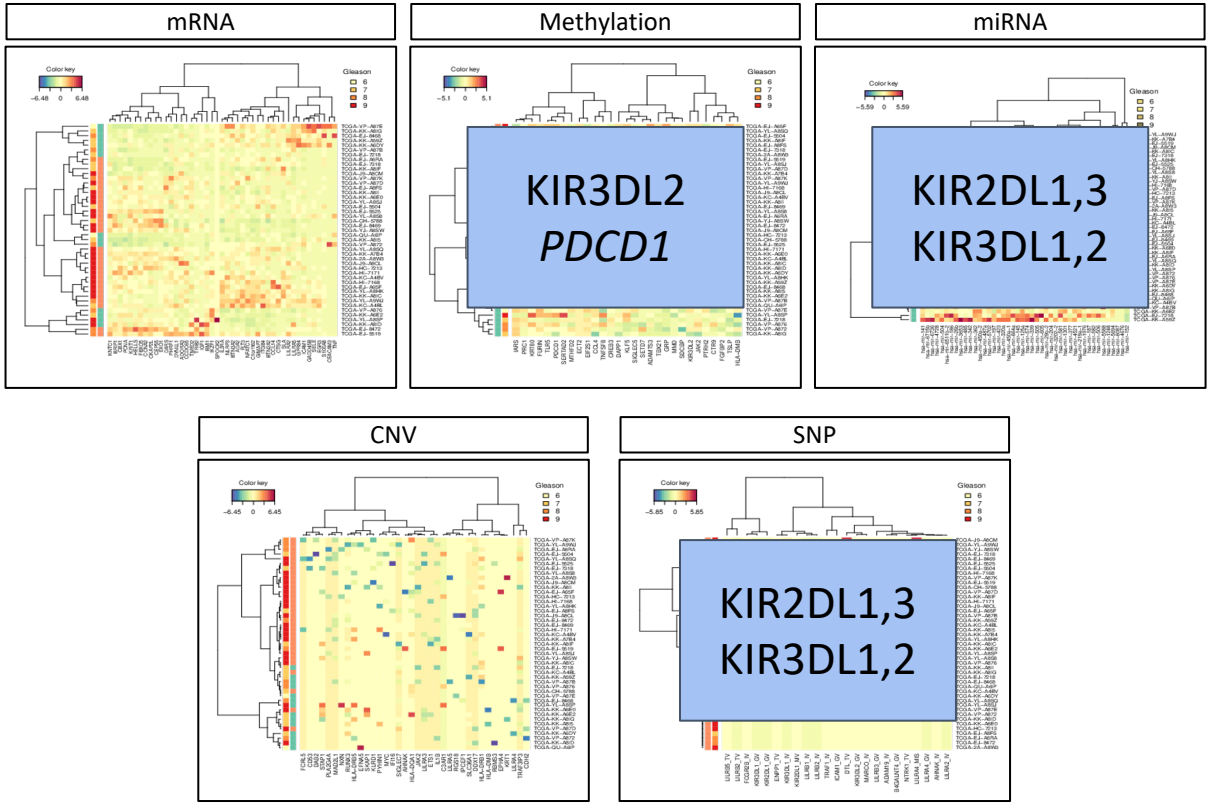
mRNA



BER = 0.20

Most interesting genes families

Gene	Type	Count
ADAM19	intron variant	2
AHNAK	intron variant	2
B4GALNT4	gene variant	3
DTL	transcript variant	2
ENPP1	transcript variant	2
FCGR2B	intron variant	2
ICAM1	gene variant	2
KIR2DL1	gene variant	4
KIR2DL1	missense variant	2
KIR2DL3	intron variant	10
KIR3DL1	gene variant	3
KIR3DL1	intron variant	19
KIR3DL2	gene variant	2
LILRA2	intron variant	3
LILRA4	gene variant	4
LILRA4	missense variant	2
LILRB1	intron variant	6
LILRB2	intron variant	7
LILRB2	transcript variant	3
LILRB3	gene variant	2
LILRB5	transcript variant	3
MARCO	intron variant	2
NTRK1	transcript variant	2
TRAF1	intron variant	2



Development of Multi-OMICS strategies for the study of Human Gut Microbiota

Institut Pasteur

CRI CENTRE DE RECHERCHE EN INFECTIOLOGIE

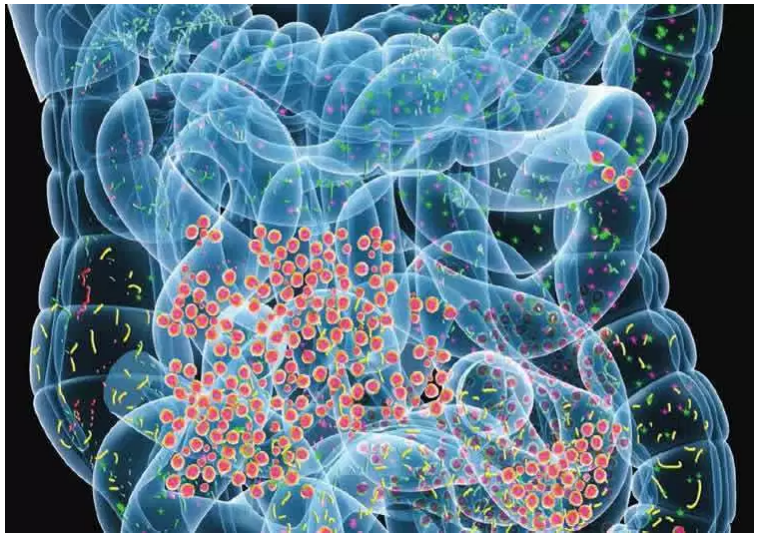
Gut Microbiota

10^{14} microorganisms (ratio 1:1 with human cells)

Hundreds/Thousands of species but 4 major phyla :
93.5% belonged to Bacteroidetes, Firmicutes, Actinobacteria and Proteobacteria

Food

Genetics



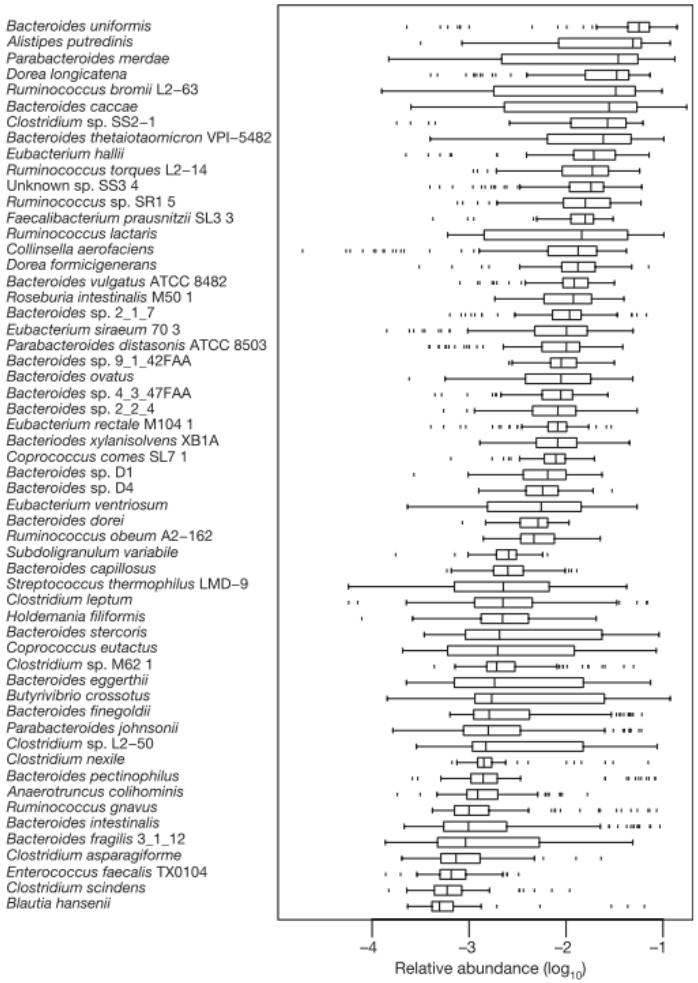
Drugs

Environment

Maintains immune and metabolic homeostasis
Protects against pathogens

Dysbiosis : Inflammatory diseases , diabete, Infection

Model for studying the effect of antibiotics on gut microbiota



from Qin J. et al, 2010

Selection of species most frequently found and most representative of gut microbiota

Artificial gut microbiota
Co-culture of selected species
(from Institut Pasteur and CRI collections)

+ antiobiotics
(doses, time-course)

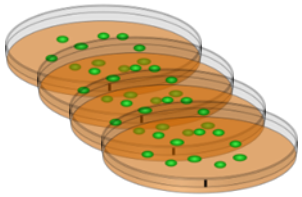
Meta-Multi-Omics Analysis

- **Meta-Genomics**
- **Meta-Transcriptomics**
- **Meta-Proteomics**
- **Meta-Metabolomics**

Data Integration

Methods optimization for Deep coverage of Metaproteome

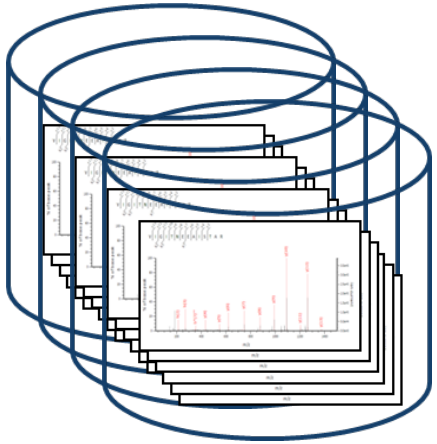
Culture of each species separately



Protein extraction
Trypsin digestion
Intensitive fractionation

DDA analysis
Partial sample coverage
Low complexity spectra

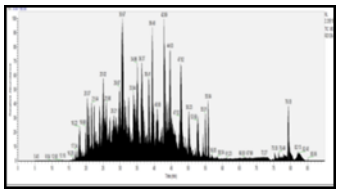
Protein/Peptides Quantification



Large spectral libraries
=
Deep proteome coverage
(due to fractionation)

Artificial Gut microbiota

DIA analysis
Whole sample coverage
High complexity spectra



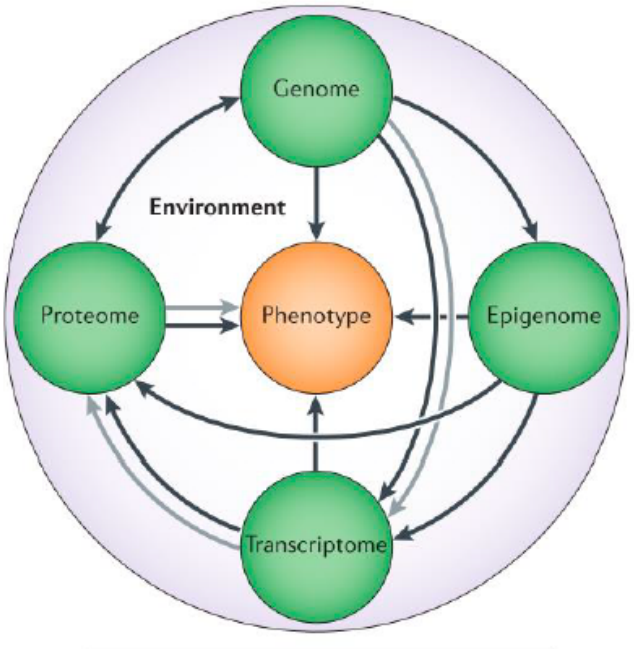
High reproducibility

DIA signal extaction
Protein/Peptides
Quantification

multi-Omics integration

Global view of biological process

Complementary data



Complex and regulated at multiple levels

The more the better?

Scott-Boyer *et al.*, Briefing in Bioinformatics, 2017

Model for studying the effect of antibiotics on gut microbiota

Step1 : Model for studying the effect of antibiotics on gut microbiota

The artificial gut microbiota might be used for :
- Other studies (metabolomic diseases, effect of diets…)
- Improve reference databases for gut microbiota

Step2: Application of developed multi-omics methods on « Culturomics »

Stools form Healthy volunteers treated with antibiotics

Meta-Multi-Omics Analysis

Step3: Direct Analysis from Patient specimens

Stools form Healthy volunteers treated with antibiotics

Meta-Multi-Omics Analysis

Research Chair and Innovation

L'Oréal in Digital Biology

CHAIRE DE RECHERCHE
ET D'INNOVATION
L'ORÉAL
EN BIOLOGIE NUMÉRIQUE
AFFILIÉ À UNIVERSITÉ LAVAL

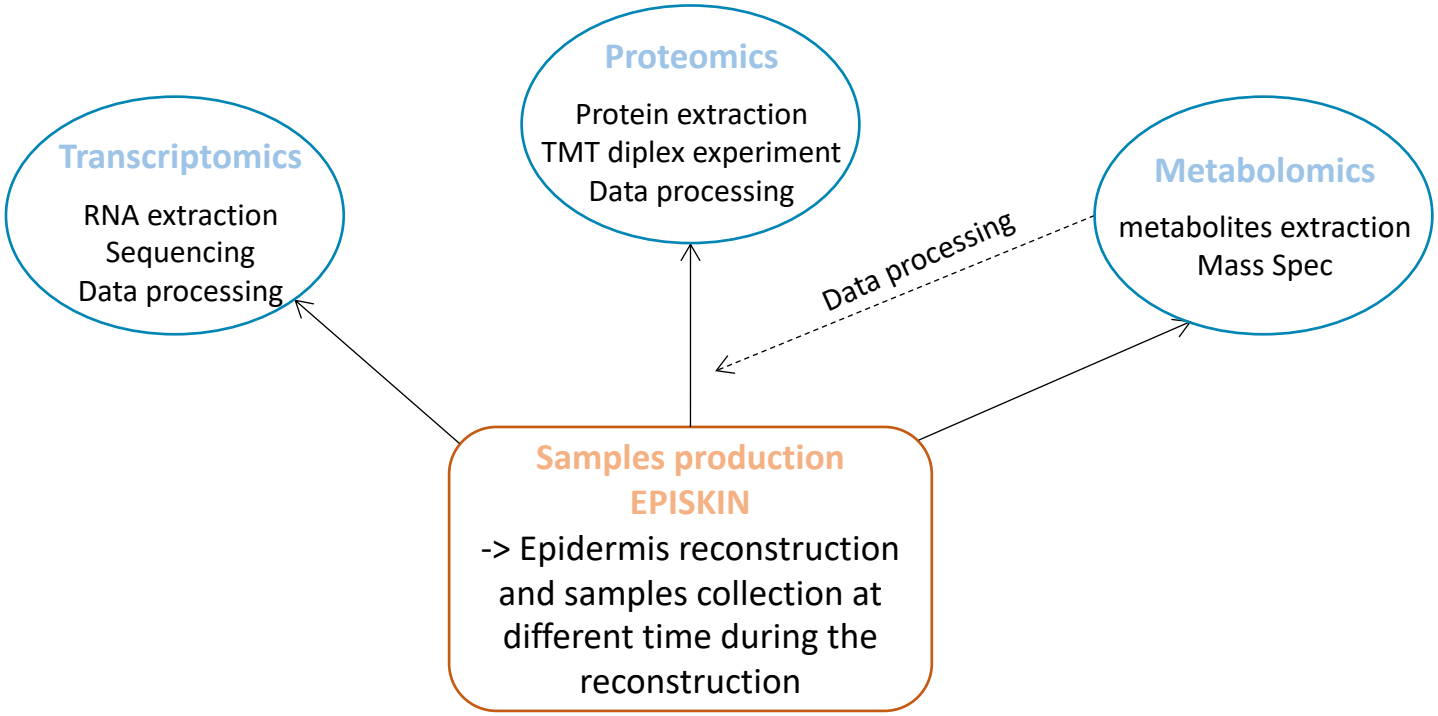
Research Chair

- Epidermis reconstruction (Biological Knowledge)
  - Multi-OMICS analysis (Analytical strategies)
  - Co-Expression (Analytical strategies)
- Microbiome (Biological Knowledge – Analytical Strategies)
- Biomarker discovery by Machine learning
- Kibio.Science (Integrated Platform for big data)

OBJECTIVES

- Understand at the molecular level the epidermal differentiation during the reconstruction process, based on a multi-OMICS approach (transcriptomics, proteomics and metabolomics), from J0 to J28
    - key biological players (genes, proteins, metabolites) involved in each step
    - key and transitional steps of the epidermal differentiation (pathways and biological funtions)
- Integrative biology: develop bioinformatics methods to correlate multi-OMICS signals

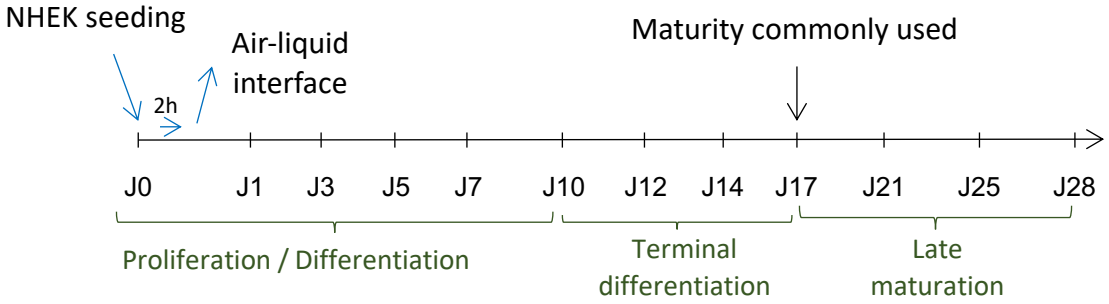
Epidermis reconstruction (Collaboration with HTBD – L'Oréal)



Samples preparation

- **Reconstructed epidermis model *=> RHE (Skinetic)***
  - Cultured in a defined medium, without any medium changes during the reconstruction process

- **Times chosen for the skin samples collection**



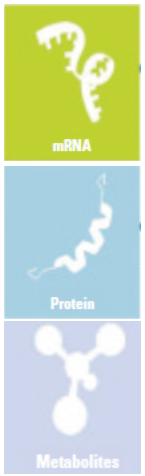
- **Replicates** -> Biological replicates rather than technical replicates => 4 donors

- **Final design**

	J0	J1	J3	J5	J7	J10	J12	J14	J17	J21	J25	J28	Vivo
Transcriptomics	4	4	4	4	4	4	4	4	4	4	4	4	4
Proteomics	4	4	4	4	4	4	4	4	4	4	4	4	4
Metabolomics	4	4	4	4	4	4	4	4	4	4	4	4	4
Hystology			4	4	4	4	4	4	4	4	4	4	4

Data analysis stategy

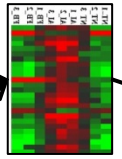
Single Omics Analysis



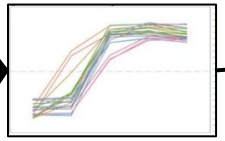
Transcriptomics data

Proteomics data

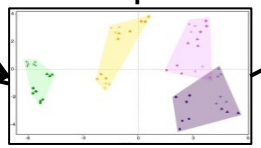
Metabolomics data



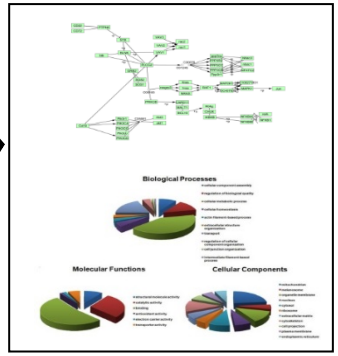
Differential expression Analysis



Kinetic & co-expression Analysis



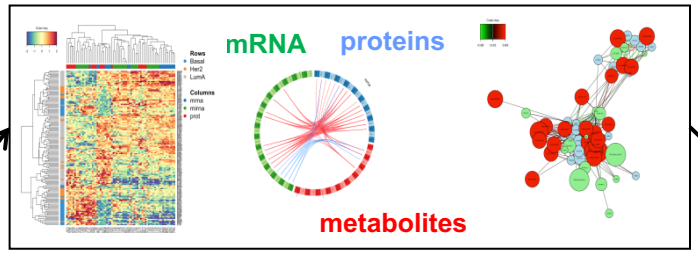
Discriminant Analysis (PCA, PLS-DA, etc…)



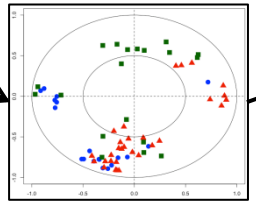
Functional Analysis & Annotations
(Pathway & Go enrichments)

Multi-Omics Analysis

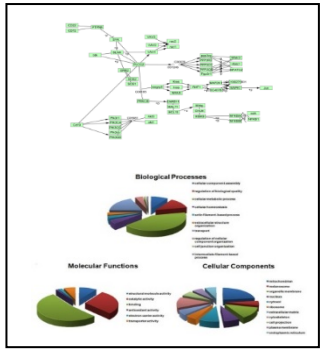
Multi-omics data



Multi-Omics correlation analysis



Multi-Omics discriminant analysis



Functional Analysis & Annotations
(Pathway & GO enrichments)

Analytical Strategies

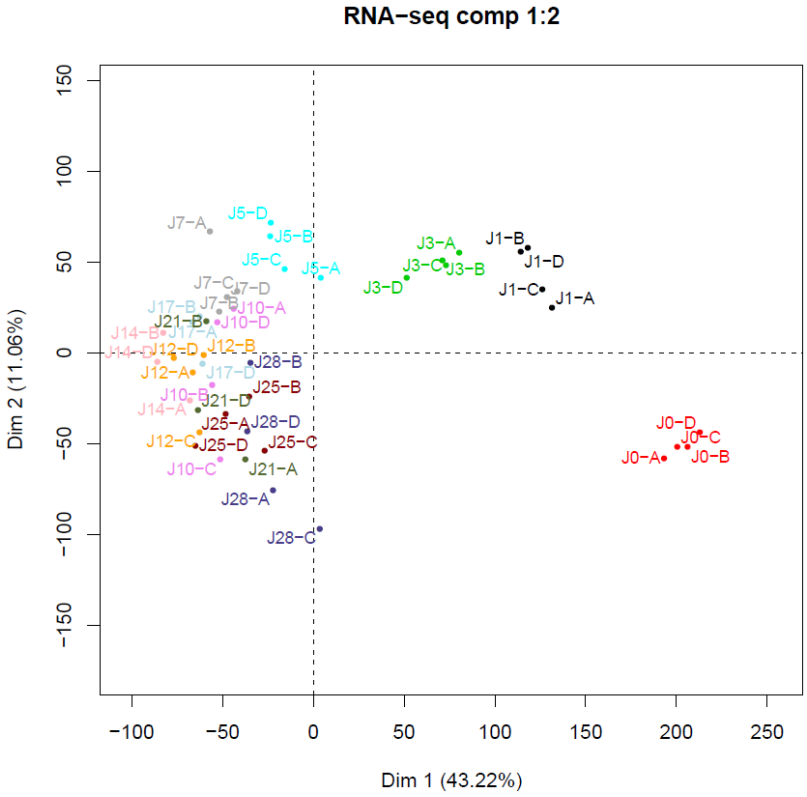
- Principal Component Analysis:
  - →Data quality and samples discrimination
- Kinetics Profiling
  - →Cluster of genes / proteins / metabolites with same kinetics profile
- Functional Analysis of Gene clusters
  - Identification of the main and successive biological functions involved during the epidermis reconstruction
- Transcriptomics Co-expression networks
- R Package MixOmics (exploratory phase)
  - Multi-Omics analysis
  - TimeOmics
- Machine learning

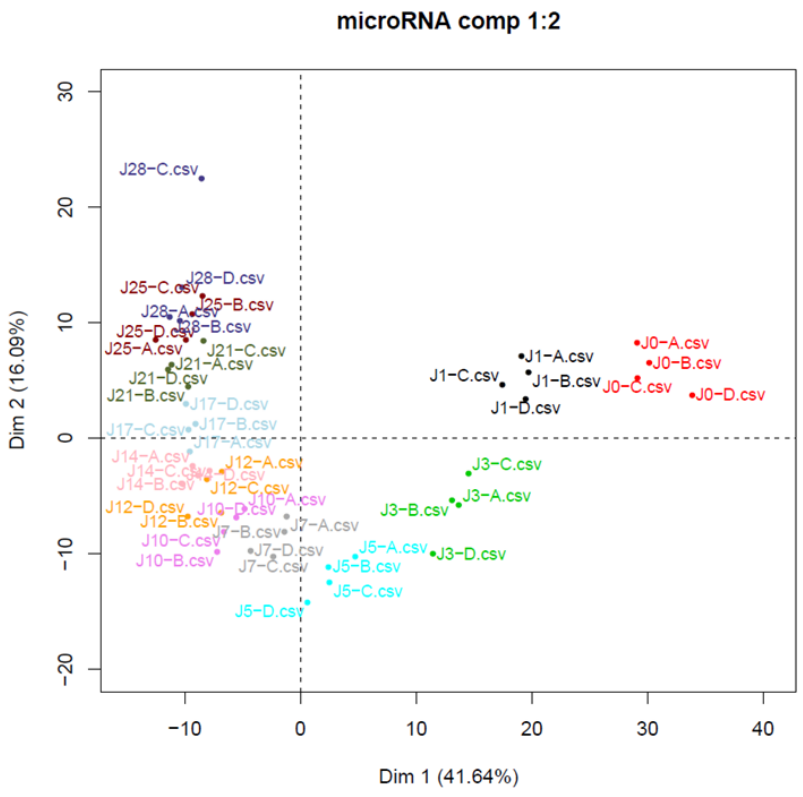
Single omics analysis

miRNA-seq and total RNA-seq and the QC are achieved

43,038 transcripts (15,348 protein coding)

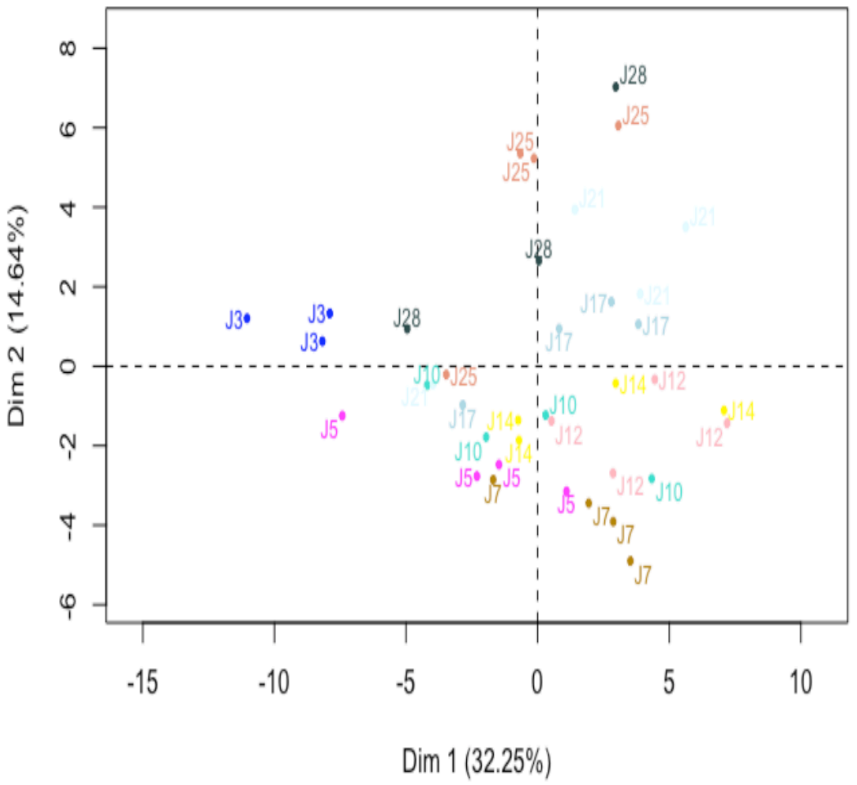
415 miRNA



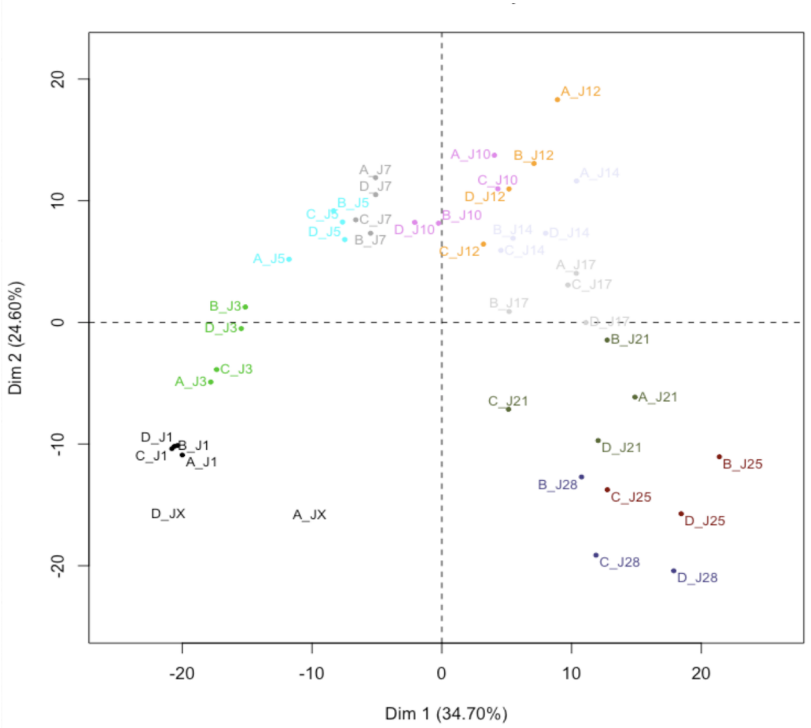


Proteomics and metabolomics

Total of 1,820 quantifiable proteins (TMT)

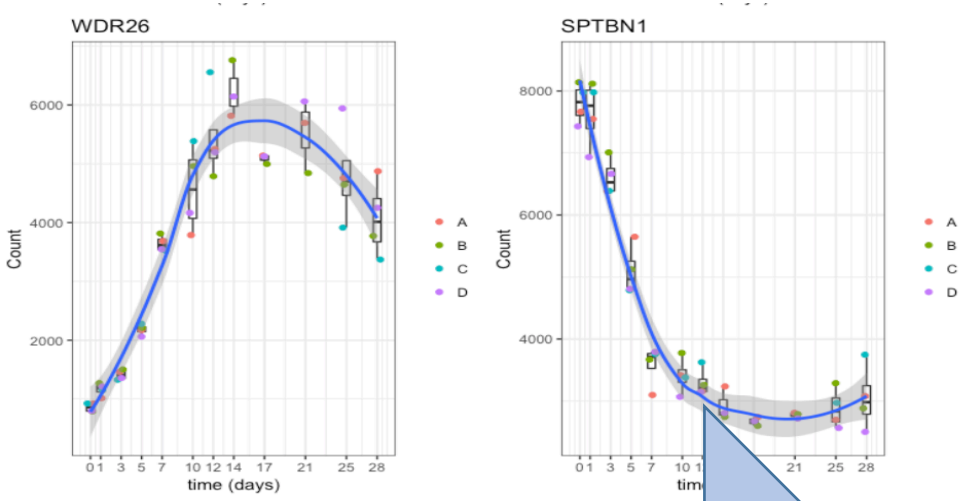


Total of 428 metabolites



Time course analysis: transcriptomics

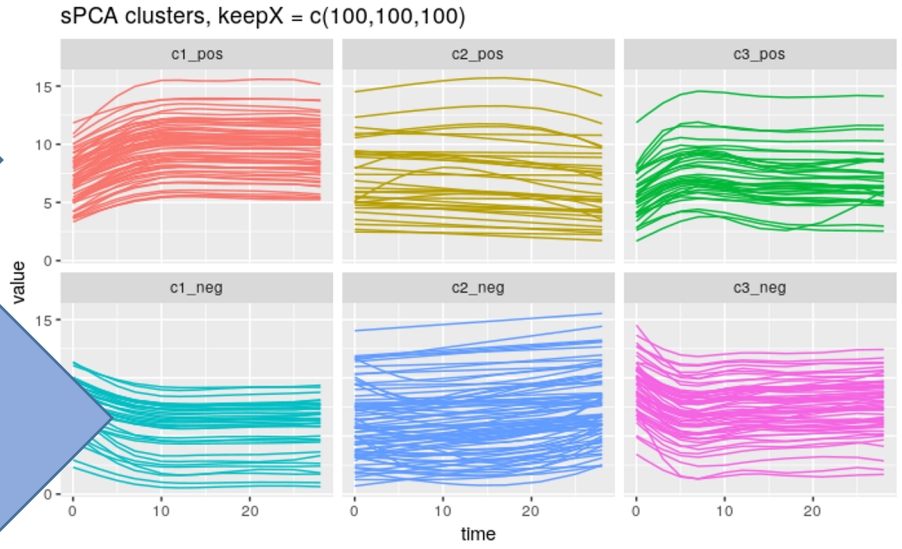
Time course modelisation with splines (LMMS)



GO:0060429~epithelium development
GO:0045595~regulation of cell differentiation

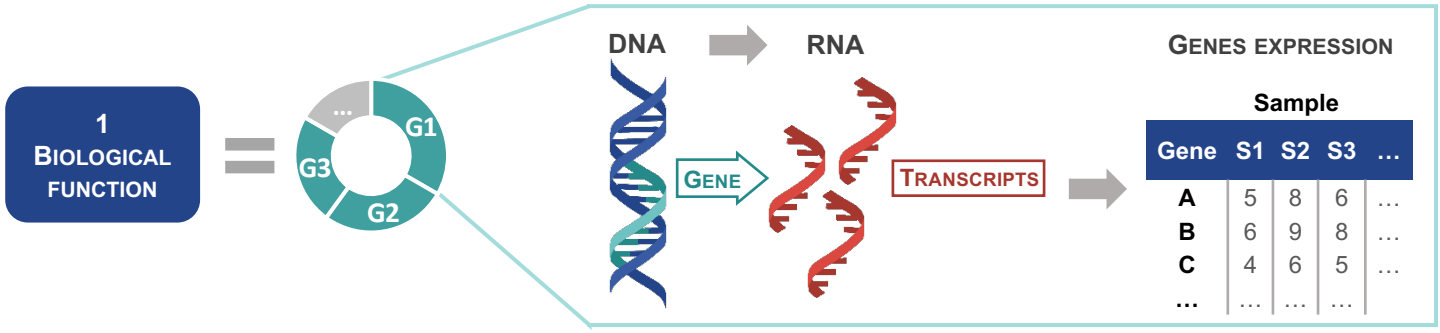
GO:0000278~mitotic cell cycle
GO:0006461~protein complex assembly

Main clusters

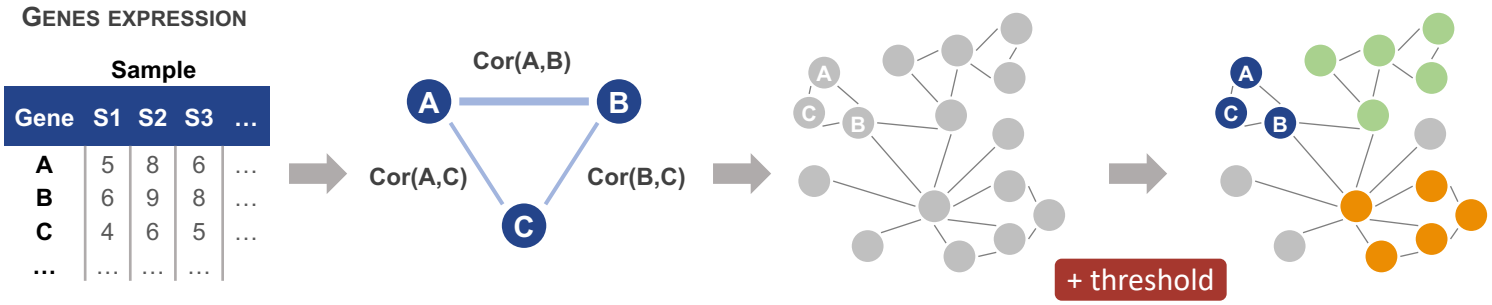


Gene co-expression network

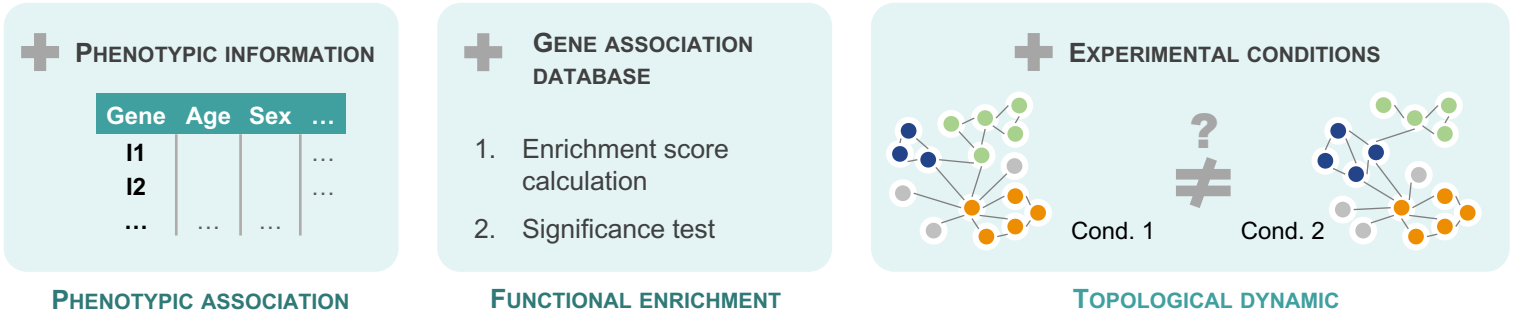
BASE



CO-EXPRESSION METHOD



EXPLOITATION

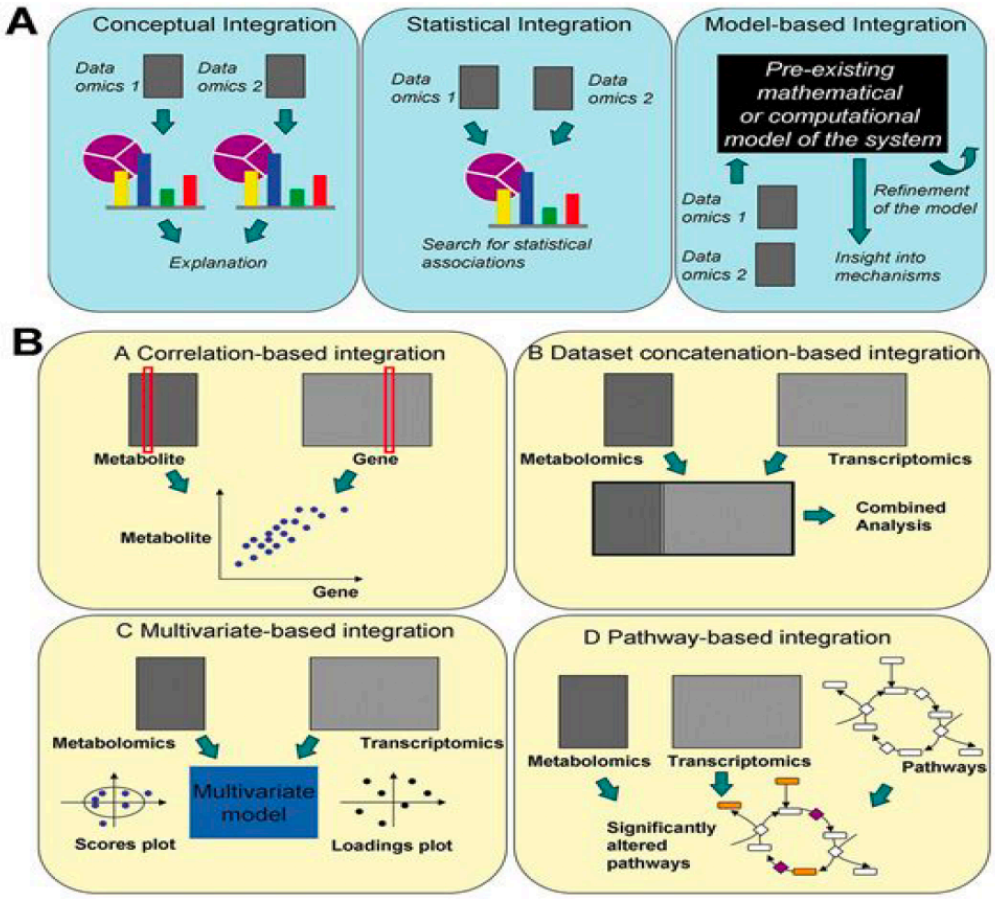


Multi-Omics analysis

Simple multi-Omics Kinetics Profiling

- Multi-Omics Kinetics profiling
- Does genes and corresponding protein have same kinetics profile ?
  → Do we enrich same biological function ?
- mRNA-miRNA target predictions based on experimental experiments

Advanced Multi-Omics approaches Overview



Cavill, R., Jennen, D., Kleinjans, J., & Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Briefings in bioinformatics*, *17*(5), 891-901.

- Static integration
- Time course integration

Next steps & Perspectives

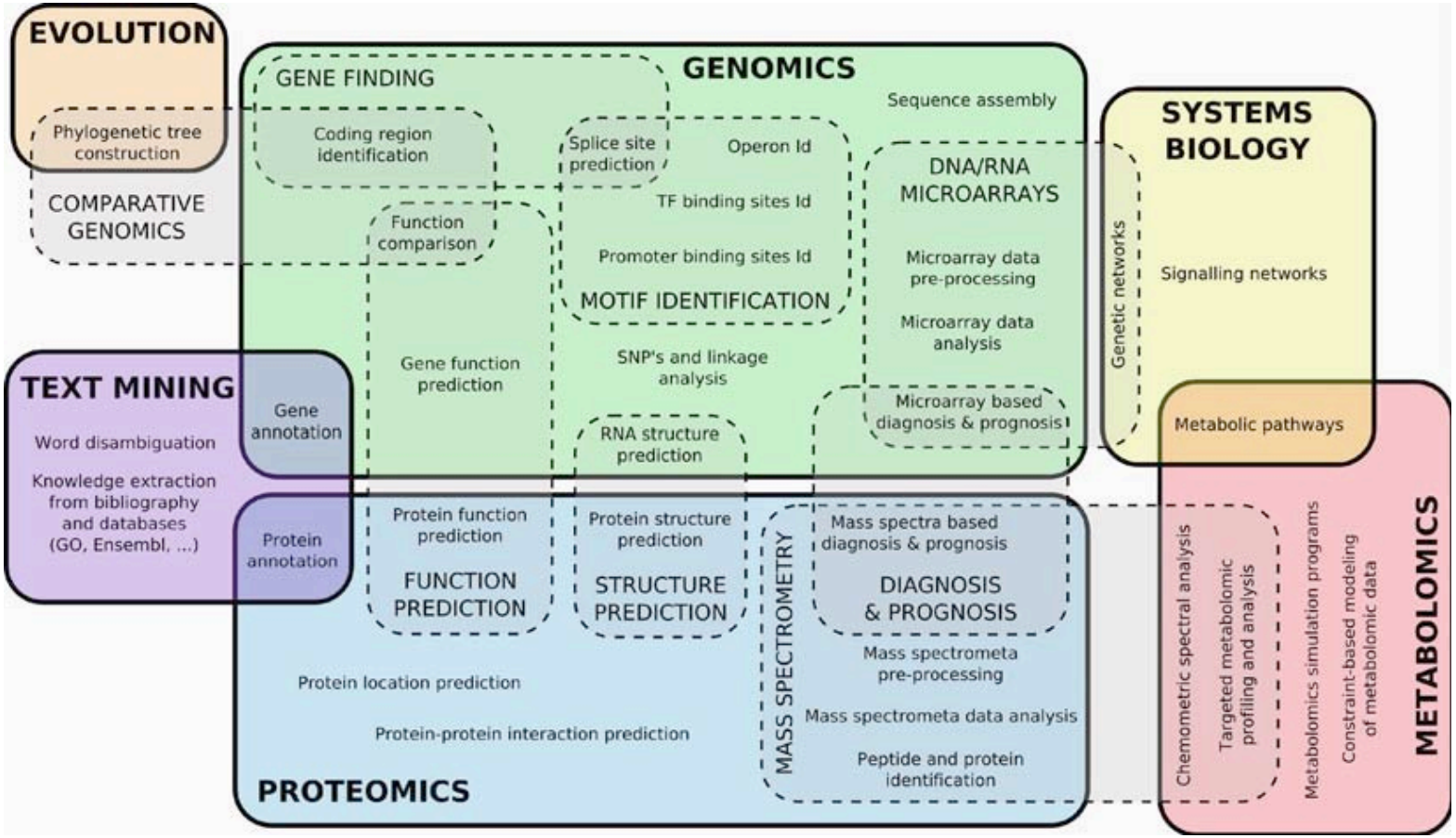
- Continue single omics analysis
  - Kinetics profiling
  - Main biological functions
  - Discriminate molecular entities for each maturation day

- Multi-Omics : new strategies to identifiy innovative biomarkers
    - Improve the robustness of molecular signatures
    - Integrate complexity of biological event
      - Complex regulation between multiple biological entities (mRNA, miRNA, proteins, metabolites)
      - Time lags

- Improve our knowledge of the epidermis formation of in vitro model by deciphering the successive events from transcripts to the metabolites
  - Key events
  - Key molecular entities
  - identify more robust multiple targets for developing new cosmetic concepts

Identify biomarker signatures in OMICs data using machine learning approaches

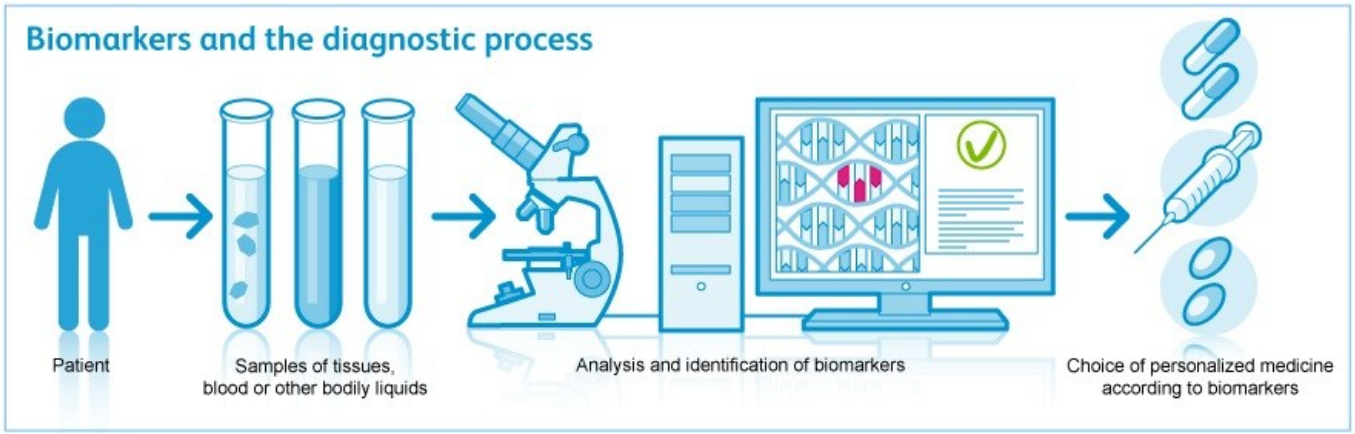
Machine learning applications in biology



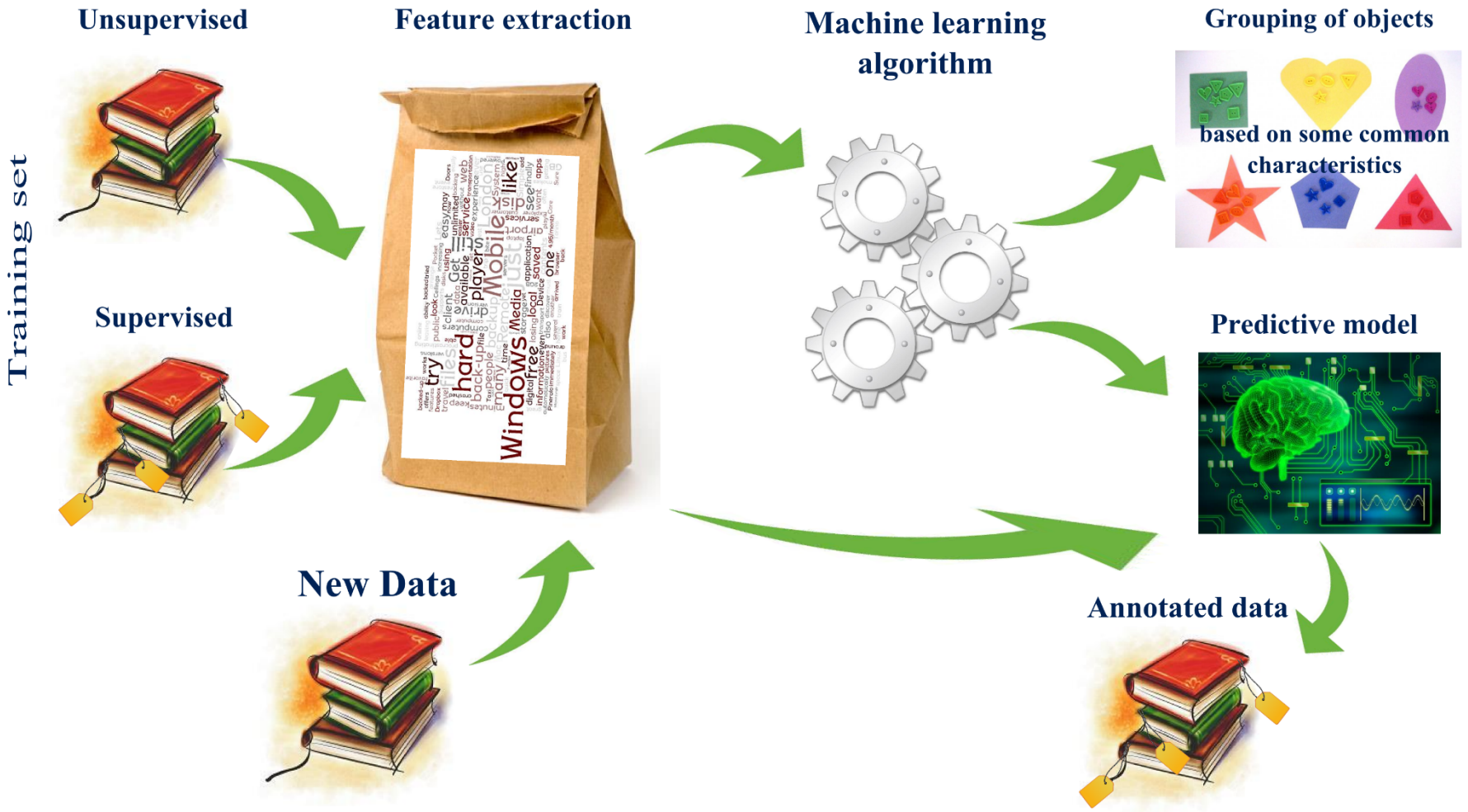
Inza et al, 2010

Biomarker signature discovery

- Biomarker = biological marker
- Substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease
- Indicator of medical state observed from the patient

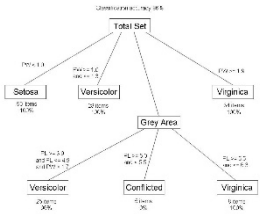


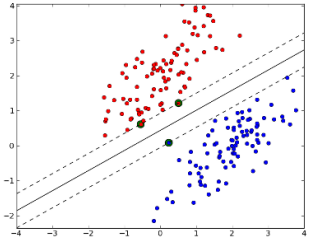
Machine learning workflow

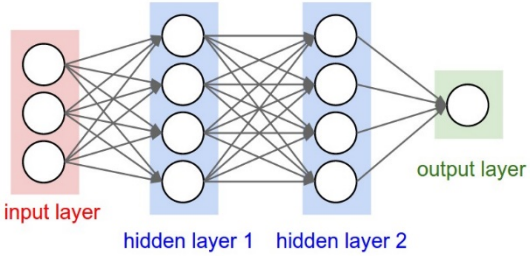


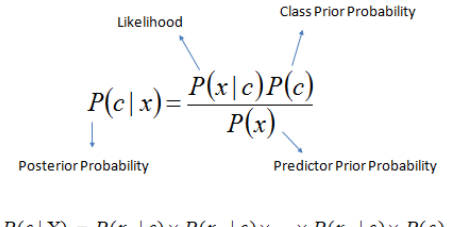
Challenges and Big questions

- OMICs: Tens of thousands of features vs 20~1000 samples
- What feature ranking/selection algorithms to use?
  - Univariate: Information gain, Gain ratio, RELIEFF…
  - Multivariate: Correlation based, Markov blanket filter…
  - What is the number of best features to include in the model?
- What classification/regression algorithms to choose?
  - Decision trees, Random Forest, SVM, NeuralNets, NaiveBayes…
- What hyper-parameters to set?









Proposed approach

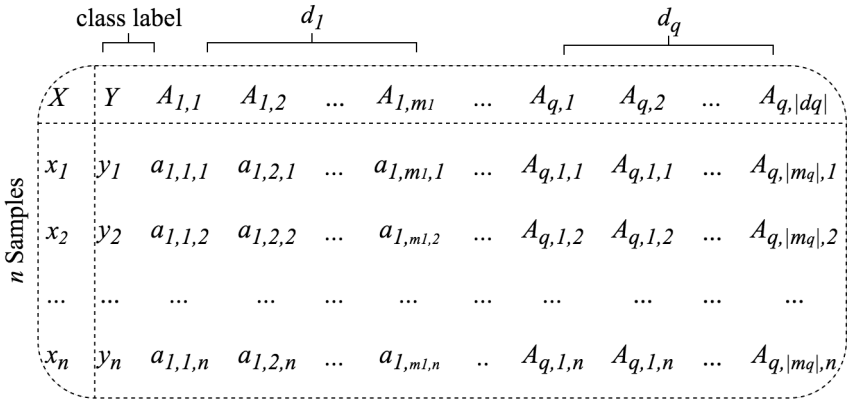
Objective: Help in the choice of the classifier and the best subset of features

- Tests all algorithms available in the weka library
- Multi-threaded supervised forward stepwise feature selection

➢BioDiscML
(Biomarker Discovery by Machine Learning)

Input / output

Input



Samples (e.g. patients)

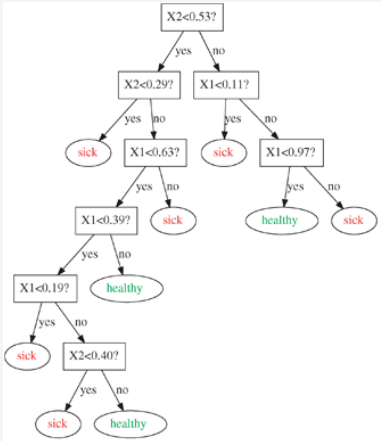
Features

- Gene expression
- Non coding RNA expression
- Protein peptides quantification
- Methylation
- Polymorphisms
- Metabolites quantification
- Clinicopathologic data

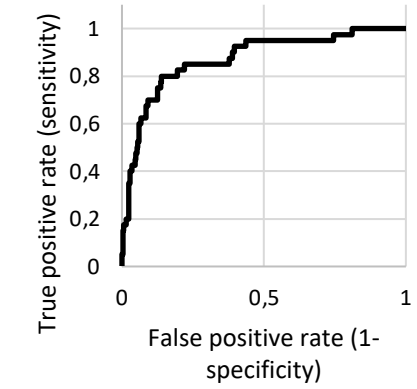
Class

- Outcome(s) (e.g. disease vs normal)
- Category (e.g. bacteria species, cancer stage)
- Level (e.g. product concentration)
- Time (e.g. cancer recurrence)

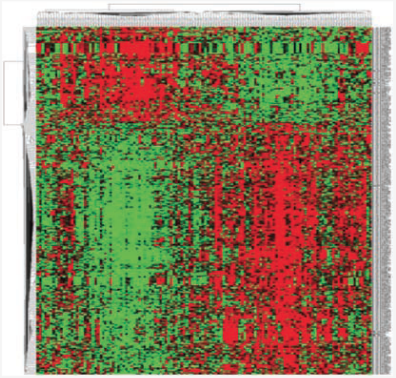
Output



Models



Performances



Signatures

Supported ML classifiers

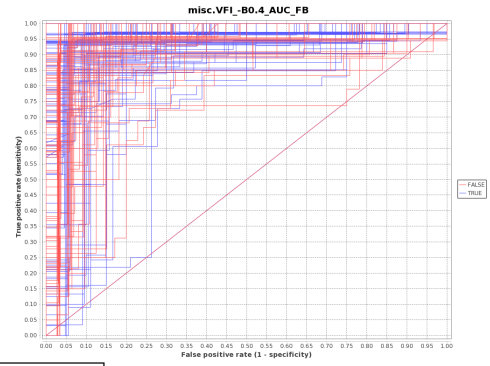
About 85 classifiers, various compatibilities

Category	Classifier	Supported feature types			Supported class types			Ref.
		Nom	Num	MV	Nom	Bin	Num	
Bayes	Averaged N Dependence Estimators (A1DE, A2DE)	x	x	x	x	x		(77)
	Bayes Network	x	x	x	x	x		(78)
	Complement Naive Bayes		x	x	x	x		(79)
	Hidden Markov Model	x	x		x	x		(78, 80)
	Hidden Naive Bayes	x			x	x		(81)
	Naive Bayes	x	x	x	x	x		(82)
	Multinomial Naive Bayes		x		x	x		(83)
	Simple Naive Bayes	x	x	x	x	x		(84)
Functions	ElasticNet		x				x	(85)
	Fisher's Linear Discriminant (FLDA)		x			x		(86)
	Gaussian Processes	x	x	x			x	(87)
	Isotonic Regression		x				x	(88)
	Kernel Logistic Regression	x	x	x		x		(89)
	Latent Dirichlet allocation (LDA)		x		x	x		(90)
	Least Median Squared linear regression	x	x	x			x	(91)
	LibLinear	x	x	x	x	x		(92)
	Support Vector Machine LibSVM	x	x	x	x	x		(93)
	Linear Regression	x	x	x			x	(94)
	Multinomial Logistic Regression	x	x	x	x	x		(95)
	Multilayer Perceptron (multiple implementations) (MLP)	x	x	x	x	x	x	(96)
	Quadratic Discriminant Analysis (QDA)		x		x	x		(97)
	Radial Basis Function (RBF), incl. Network and regressor	x	x	x	x	x	x	(98)
	Stochastic Gradient Descent (SGD)	x	x	x		x		(99, 100)
	Simple Linear Regression		x	x			x	(101)
	Simple Linear Logistic regression	x	x	x	x	x		(102)
	Sequential Minimal Optimization (SMO)	x	x	x	x	x		(103–105)
	Sequential Minimal Optimization regression	x	x	x			x	(106, 107)
	Stochastic Primal Estimated sub-GrAdient SOlver for SVM (Pegasos)	x	x	x		x		(108)
	Voted Perceptron	x	x	x		x		(109)
	Winnow	x		x		x		(110)
	Analogical Modeling	x		x	x	x		(111)
	K nearest neighbours classifier with various distance measures	x	x	x	x	x		(112, 113)
...								

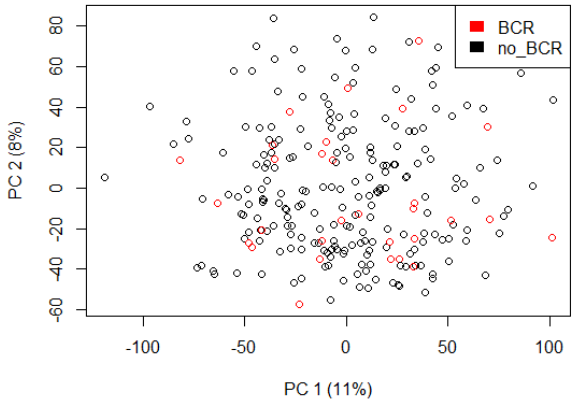
...								
Misc.	CHIRP	x	x	x	x	x		(119)
	Fuzzy Lattice Reasoning (FLR)		x	x	x	x		(120)
	HyperPipes	x	x	x	x	x		(40)
	Ordinal Stochastic Dominance Learner (OSDL)	x			x	x		(121, 122)
	Voting Feature Intervals (VFI)	x	x	x	x	x		(123)
Rules	ConjunctiveRule	x	x	x	x	x	x	(40, 124)
	Decision Table/Naive Bayes hybrid (DTNB)	x	x	x	x	x		(125)
	Fuzzy Unordered Rule Induction Algorithm (FURIA)	x	x	x	x	x		(126)
	Repeated Incremental Pruning to Produce Error Reduction (JRip)	x	x	x	x	x		(127)
	Lazy Associative Classifier (LAC)	x		x	x	x		(128)
	M5 Rules	x	x	x			x	(129–131)
	MODLEM	x		x	x	x		(132)
	Multi Objective Evolutionary Fuzzy	x	x		x	x		(133)
	Nearest-neighbor-like generalized (NNge)	x	x	x	x	x		(134)
	Ordinal Learning Method (OLM)	x	x	x	x	x		(135)
	One Rule (OneR)	x	x	x	x	x		(135, 136)
	PART	x	x	x	x	x		(137)
	PRISM	x			x	x		(133, 138)
	RIpple-DOwn Rule (Ridor)	x	x	x	x	x		(139)
	Rough Set	x	x	x	x	x		(112)
	Zero Rule (ZeroR)	x	x	x	x	x	x	(39)
Trees	Alternating Decision Tree (ADTree)	x	x	x		x		(112, 140)
	Alternating Model Tree	x	x	x			x	(141)
	Best-First Decision Tree (BFTree)	x	x	x	x	x		(142)
	Credal Decision Tree (CDT)	x	x	x	x	x	x	(143)
	Decision Stump	x	x	x	x	x	x	(144)
	ExtraTree		x		x	x	x	(145)
	Functional Trees (FT)	x	x	x	x	x		(146)
	Hoeffding Tree	x	x	x	x	x		(147)
	ID3	x			x	x		(146, 148)
	C4.5 Tree (J48)	x	x	x	x	x		(149)
	C4.5 Consolidated Tree (J48 Consolidated)	x	x	x	x	x		(150, 151)
	LogitBoost Alternating Decision Tree (LADTree)	x	x	x	x	x		(152)
	Logistic Model Trees (LMT)	x	x	x	x	x		(102, 146)
	M5 Model Trees	x	x	x			x	(130, 131)
	Naive Bayes Tree (NBTree)	x	x	x	x	x		(153)
	Random Forest	x	x	x	x	x	x	(154)
...								

Use case : Prostate cancer

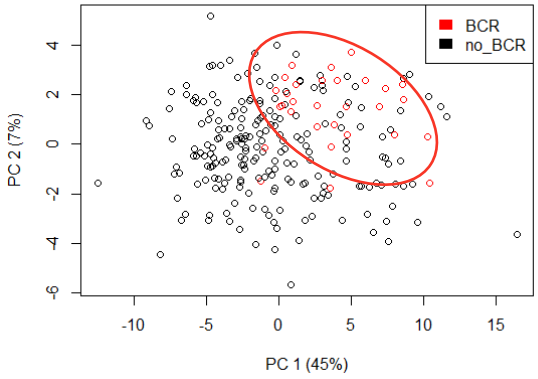
Goal: Predict the biochemical recurrence in prostate cancer
Input: TCGA, 329 patients. Gene expression (RNAseq, 19k genes)
Model: voting features intervals (bias 0.4)
AUC: 0.8 (bootstrapping)

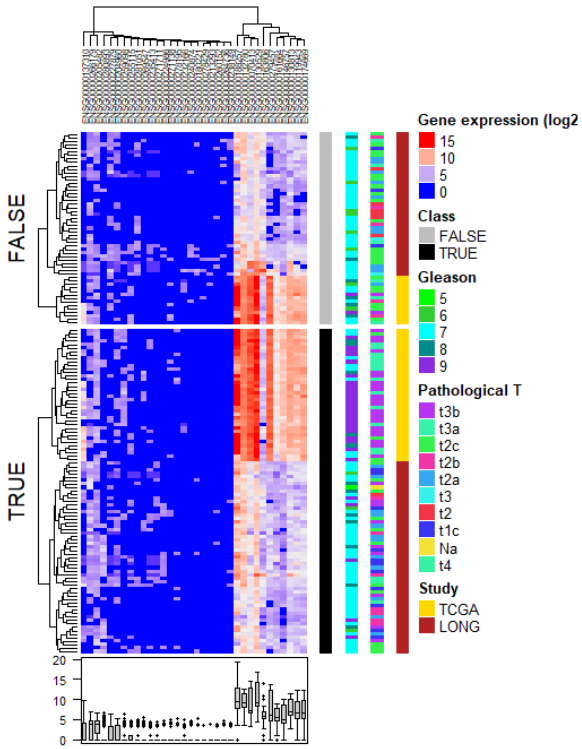


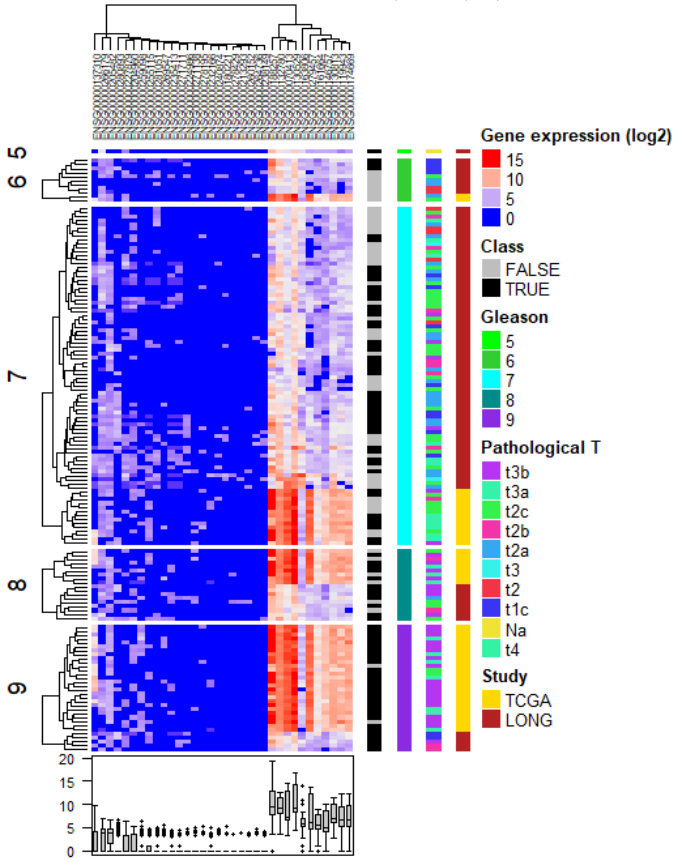
All genes in input



Signature (31 genes)



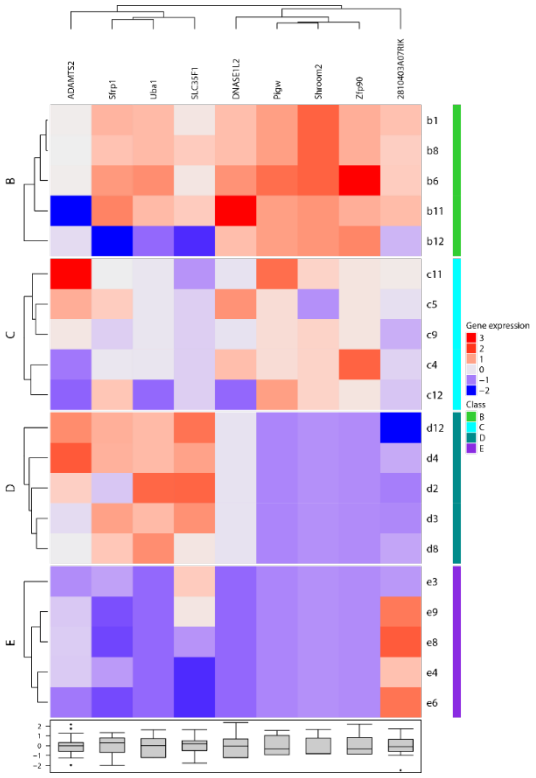




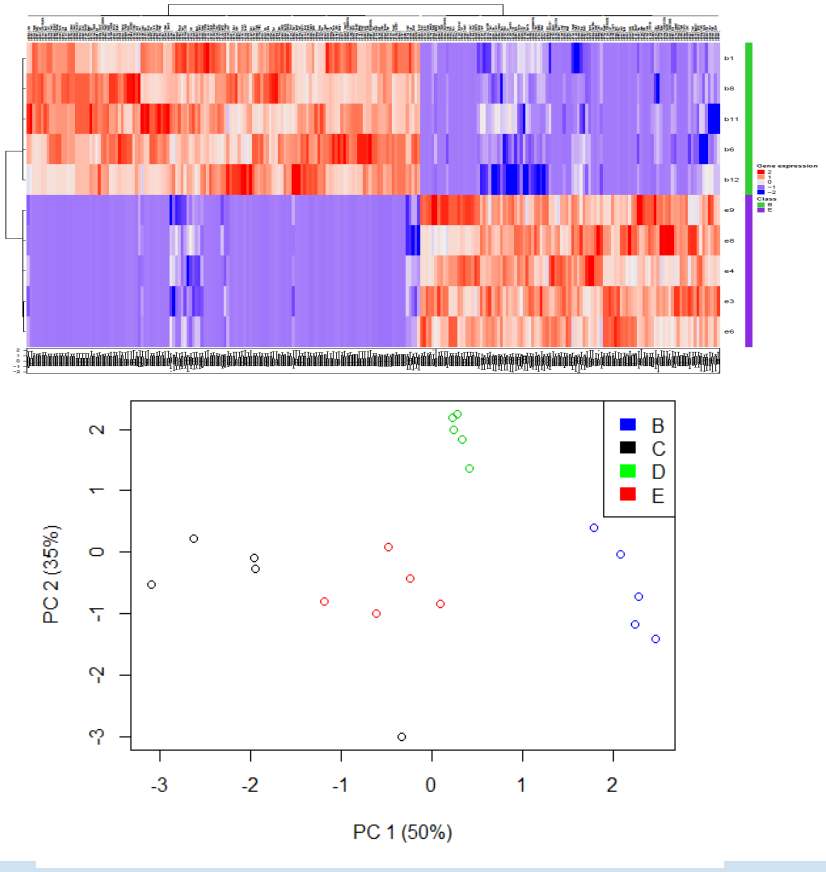
Use case: Mice fed by different fatty acids

- **Study:** 4 groups of mice have been fed different fatty acids
- **Goal:** Find a gene signature that discriminate the mice groups
- **Input:** Gene expression (RNAseq, 22k features)
- **Model found**: K-nearest neighbours classifier
- **AUC**: 0.93

9 genes signature



Correlated genes by spearman correlation (classes B vs E) and 10CV infogain merit >0.9:

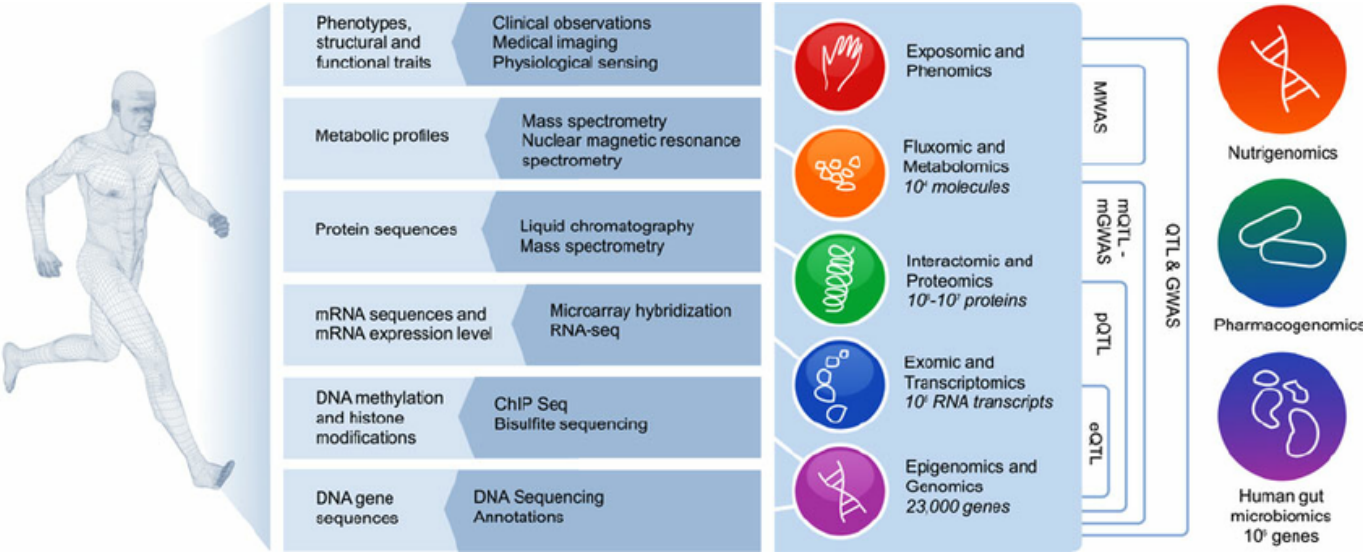


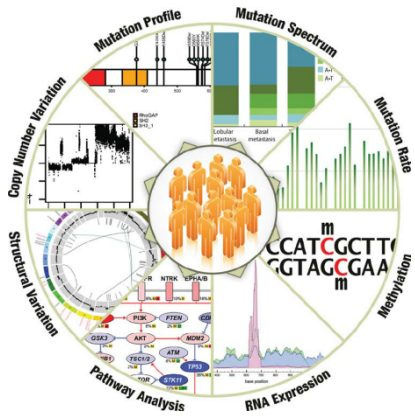
Conclusions

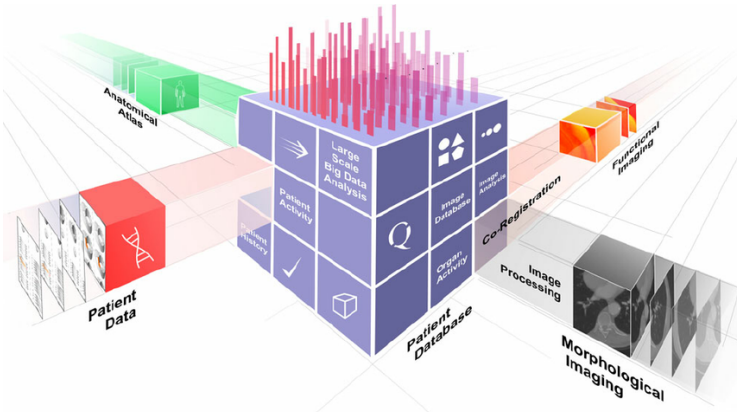
- BioDiscML is an automated approach to search for best ML algorithms which fit input data
- Help use to decide what model to use
- Identifies the best subset of features for prediction
- Report the correlated genes, useful to understand

Kibio.science

From patient to massive data







Andreu-Perez et al. 2015

Evolution for databases

- Evolution in molecular biology
  - From 2 databases in 1980[1] to 1737 databases in 2017[2]
- Databases of biological databases
  - MetaBase[1], NAR[2], NIST[3], OReFiL[4], DBD[5], HSLS[6], and more

[1] D. J. Rigden et al., Nucleic Acids Research, 2012
[2] D. J. Rigden and X. M. Fernández, Nucleic Acids Research, 2018
[3] D. M. Blakeslee et al., NIST Data Gateway, 2016
[4] Y. Yamamoto and T. Takagi, BMC Bioinformatics, 2007
[5] M. V. N. Setty and H. Rao, Jour. of Biochemical Technology, 2009
[6] A. Chattopadhyay et al., J Med Libr Assoc, 2006

Challenges for a researcher

- Many databases & many web services
- Need to…
  - Jump from one website to another to get dataset
  - Use external tools to create relations
  - Use external tools to analyse
  - Visualization

1/ Hard
2/ Time-consuming
3/ Repeat

Kibio project

- Huge need to connect biological knowledge together
- Link data easily
- New search engine
- Online portal

Challenges to manage the data

- How to efficiently store?
  - Volumes,
  - Data structures,
  - Links,
  - Etc
- How to quickly explore and search?
  - Splitting,
  - Algorithms,
  - Requests,
  - Etc

Interconnected

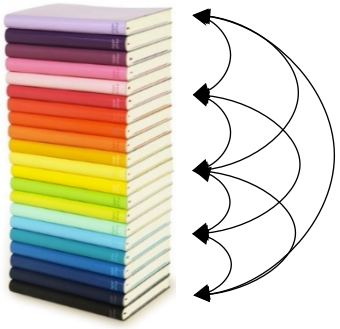
elasticsearch

Objectives

Goal: Create a platform to

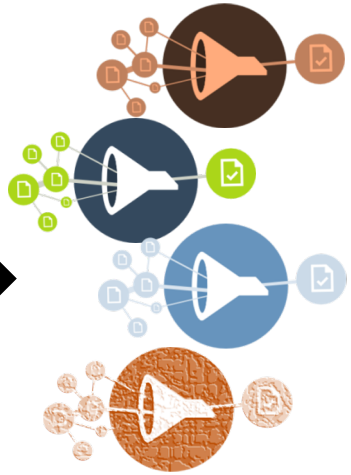
- Store and join together most known bioinformatics databases at one place
- Provide various and flexible plugins to visualize life science data
- Perform instant searches and interact dynamically with the visualizations
- Explore and interpret relations

Biological/ Bioinformatics databases



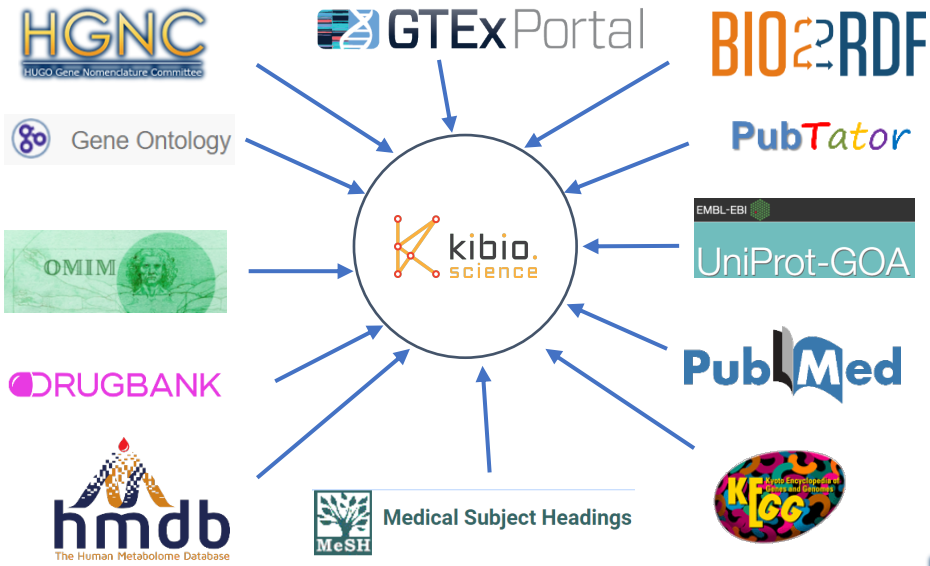
Structured information and Linking

Visualizations

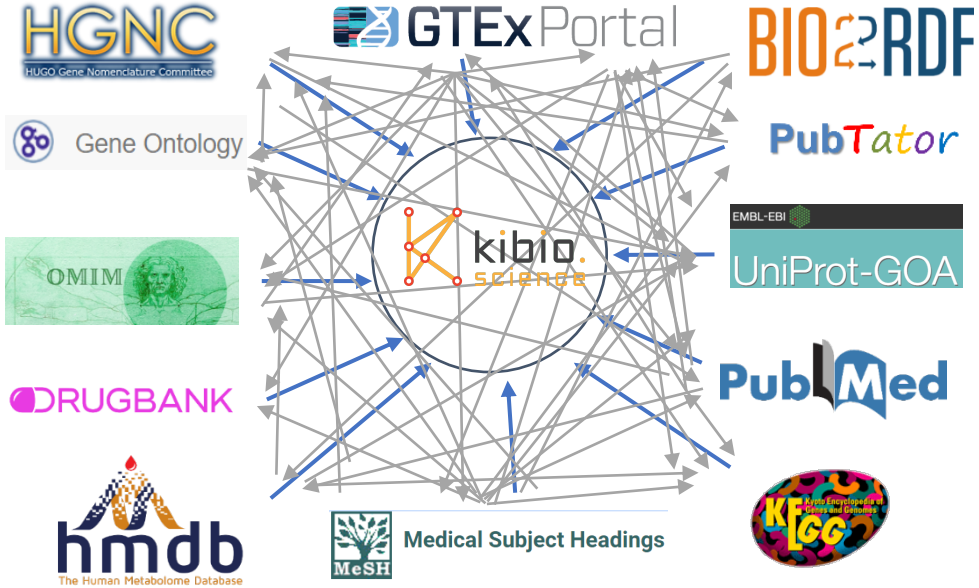


Advanced filtering system

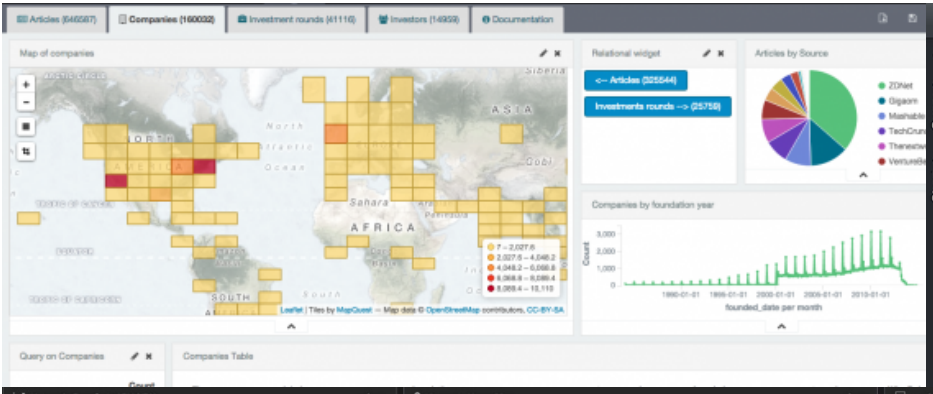
Database integration and linking

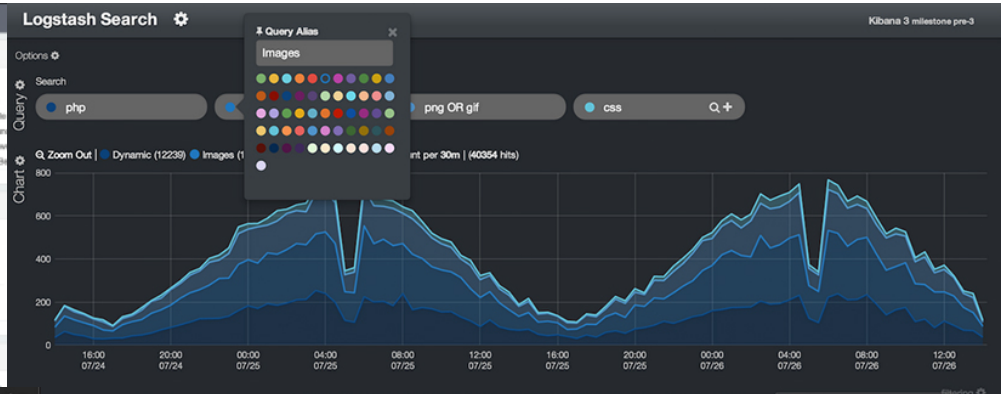


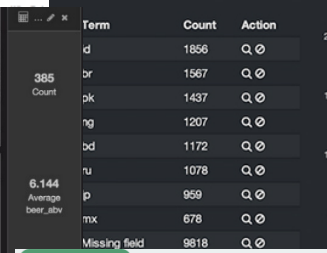
With relations

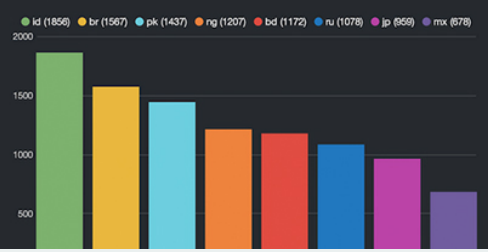


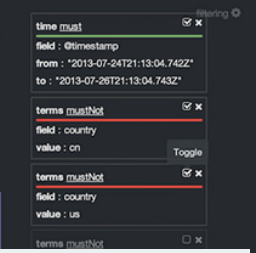
Visualizations

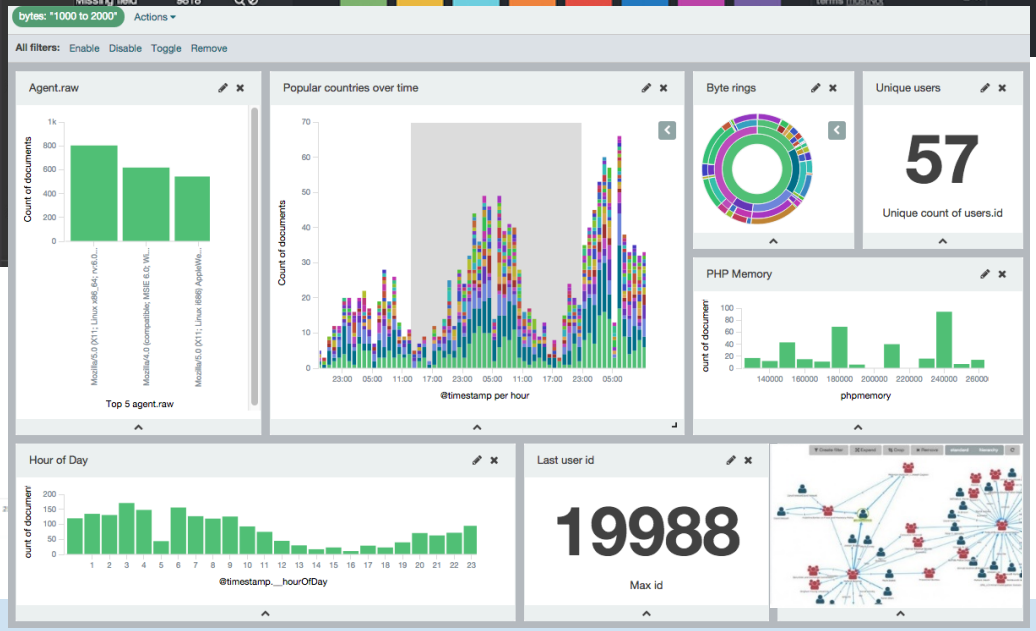


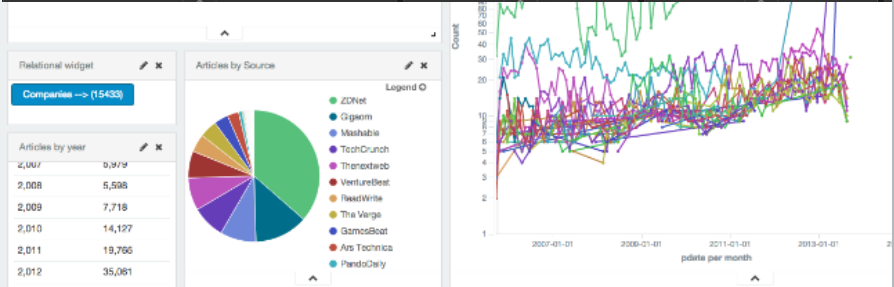






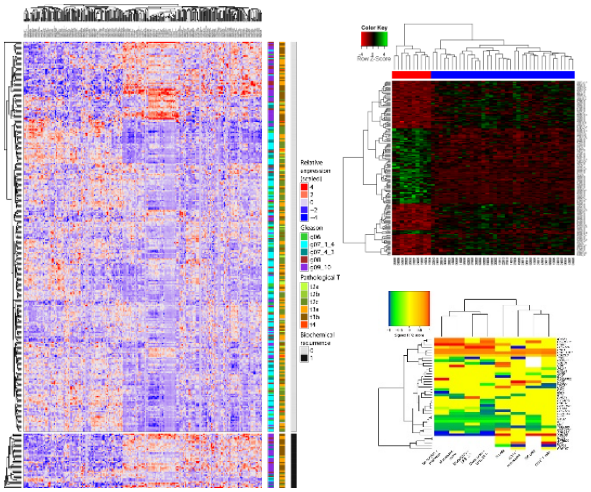




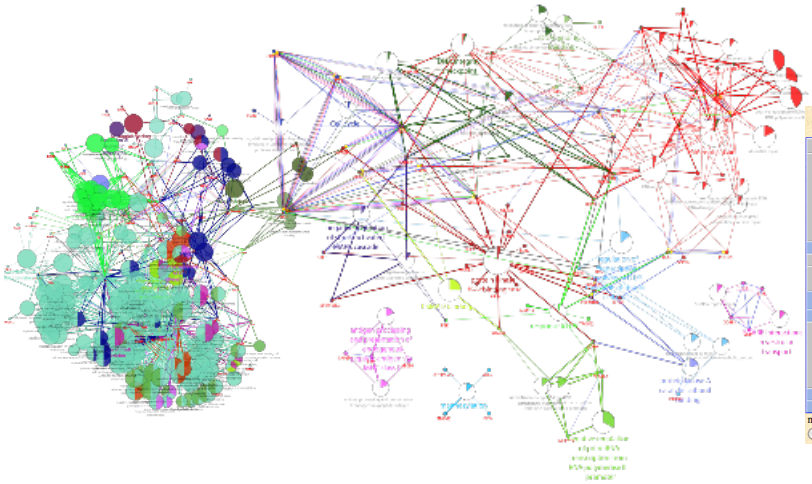


More visualizations

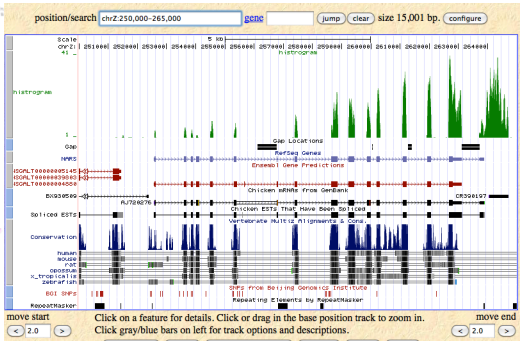
Expression heatmap



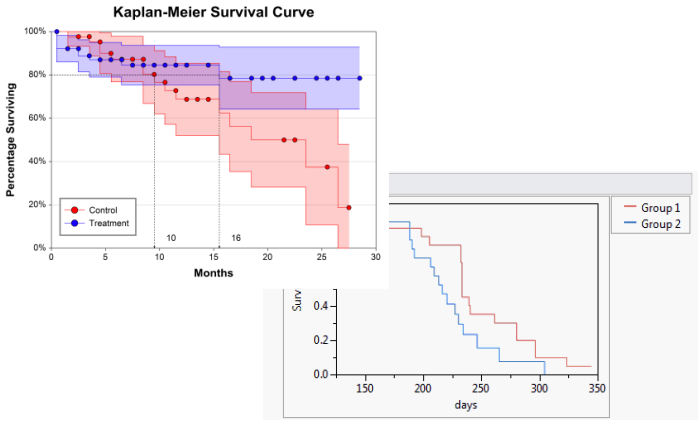
Gene networks



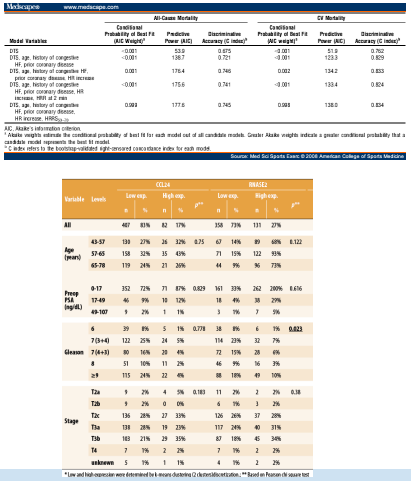
Genome browser



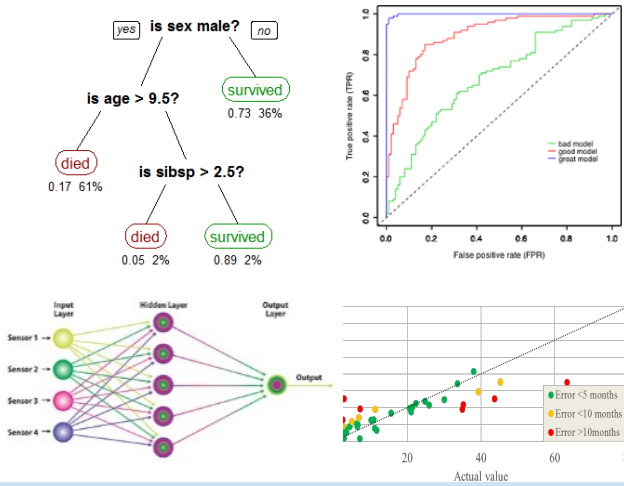
Survival analysis



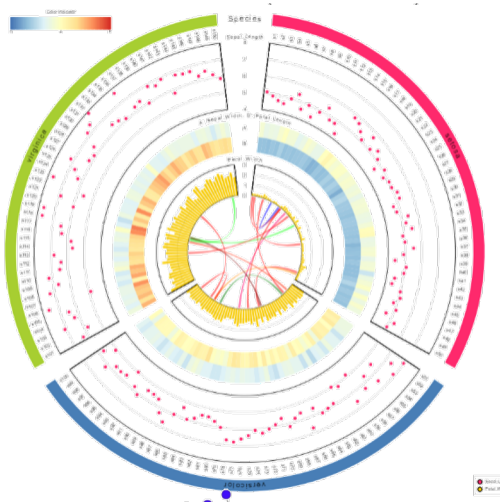
Biostatistics

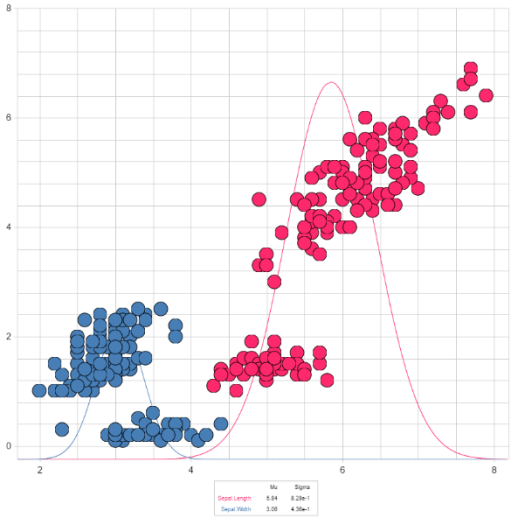


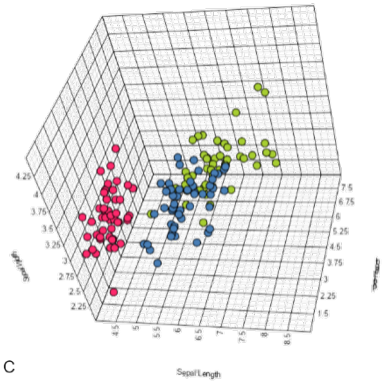
Machine learning

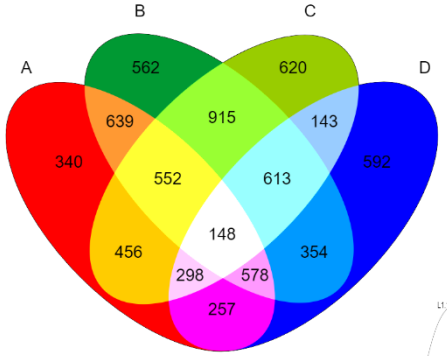


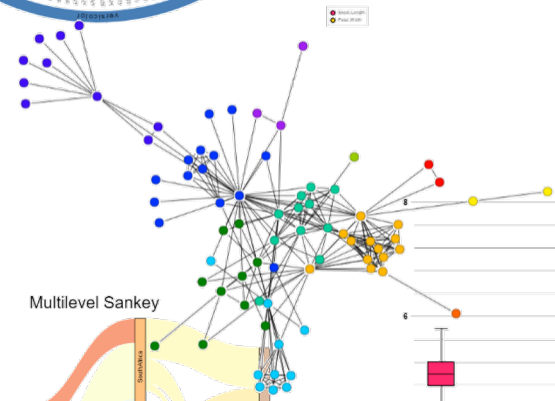
More visualizations



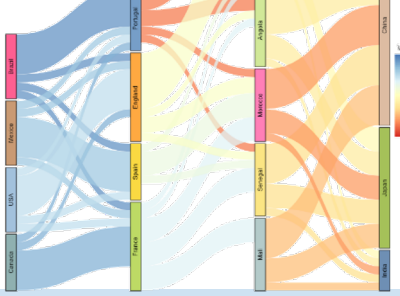


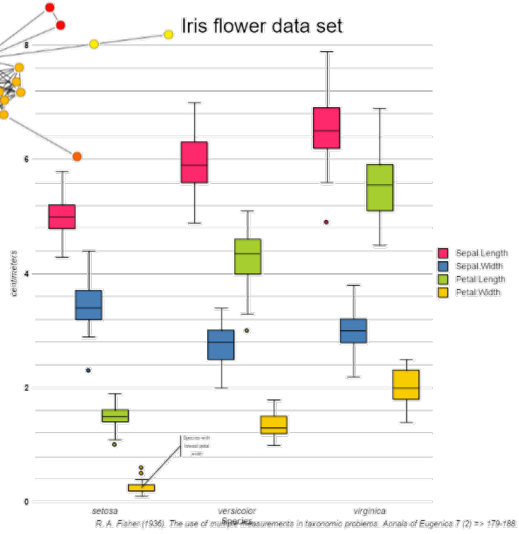


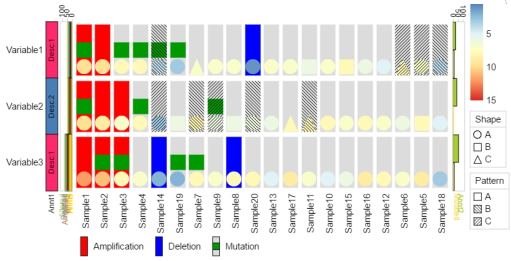




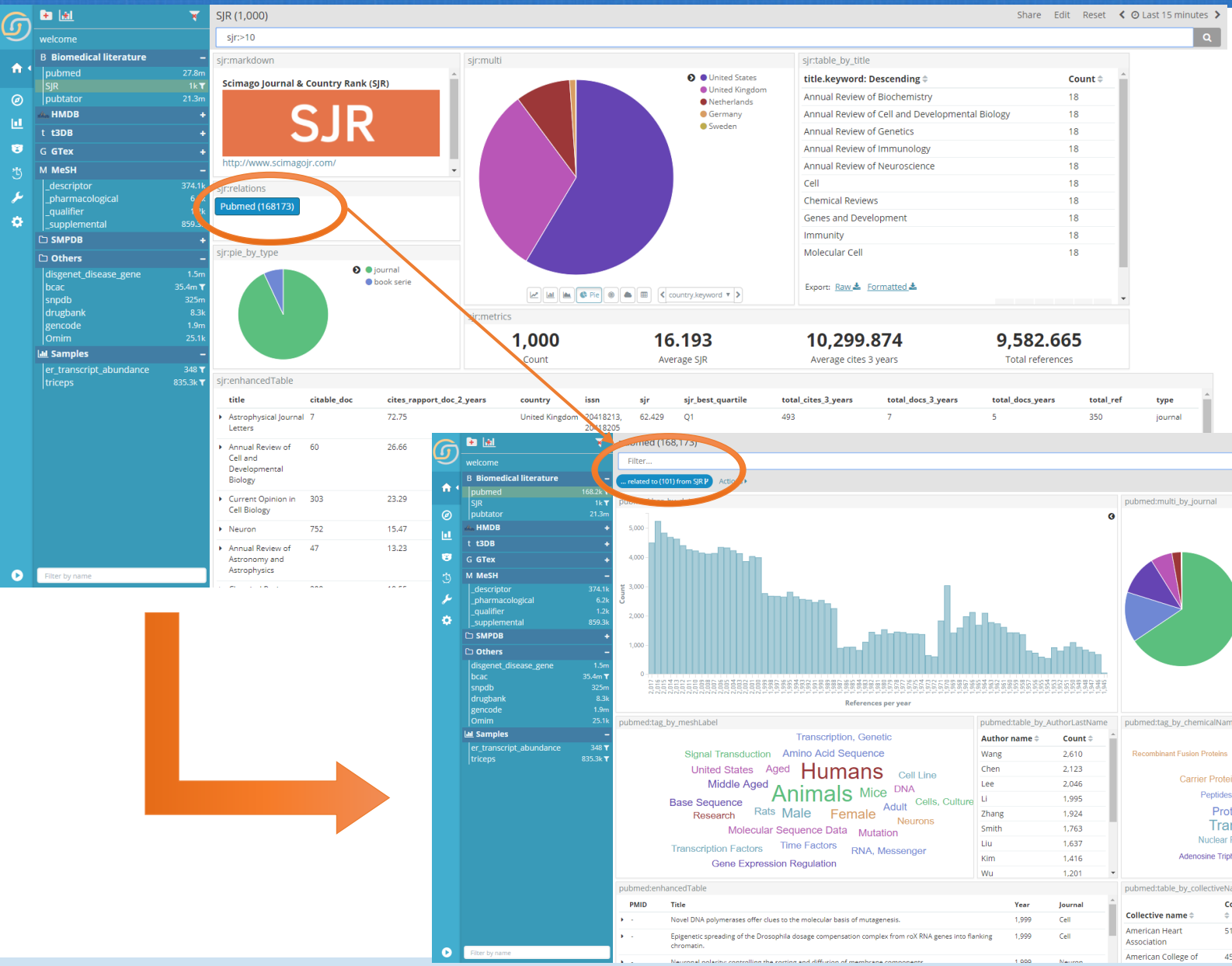
Multilevel Sankey



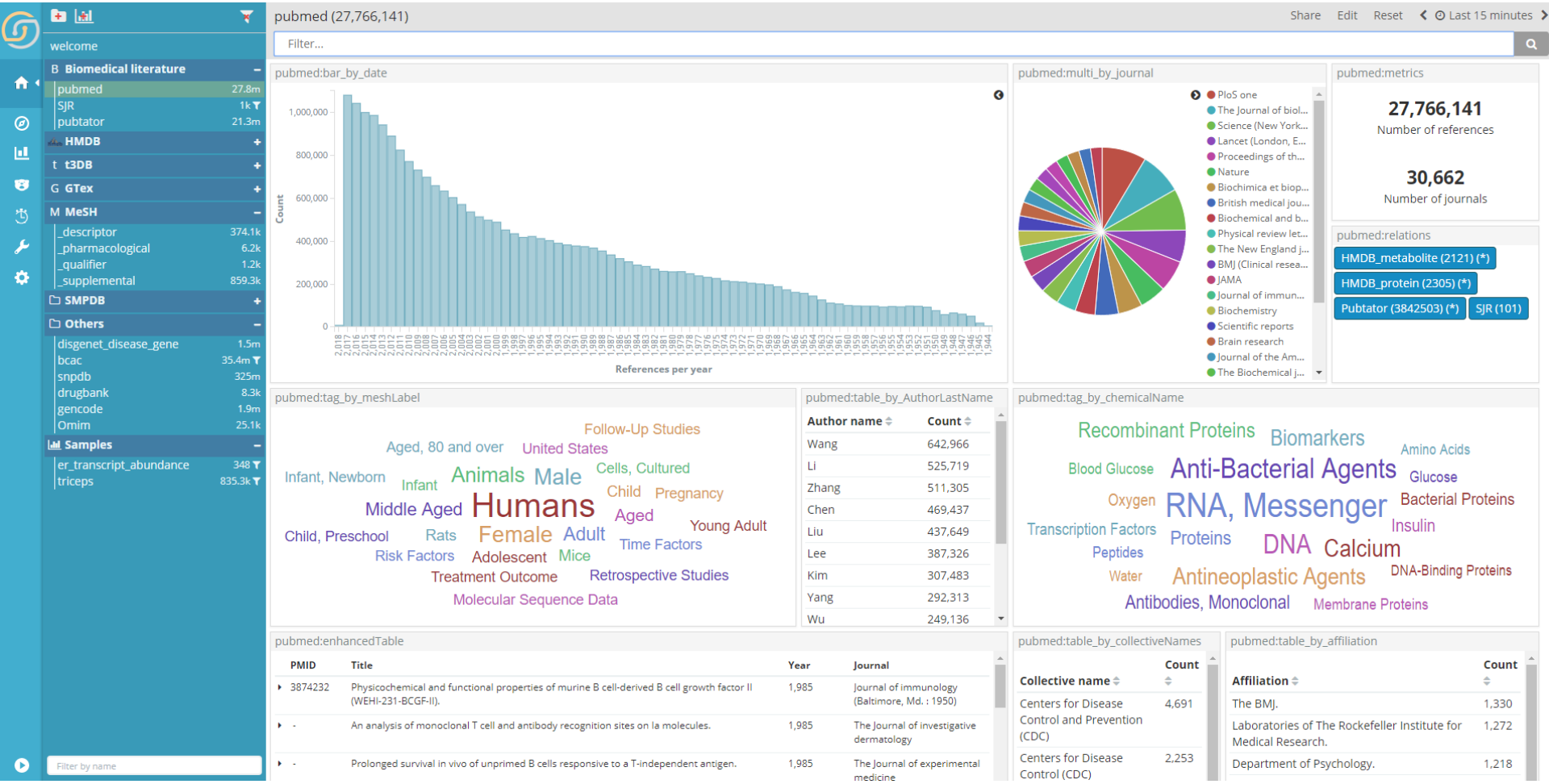




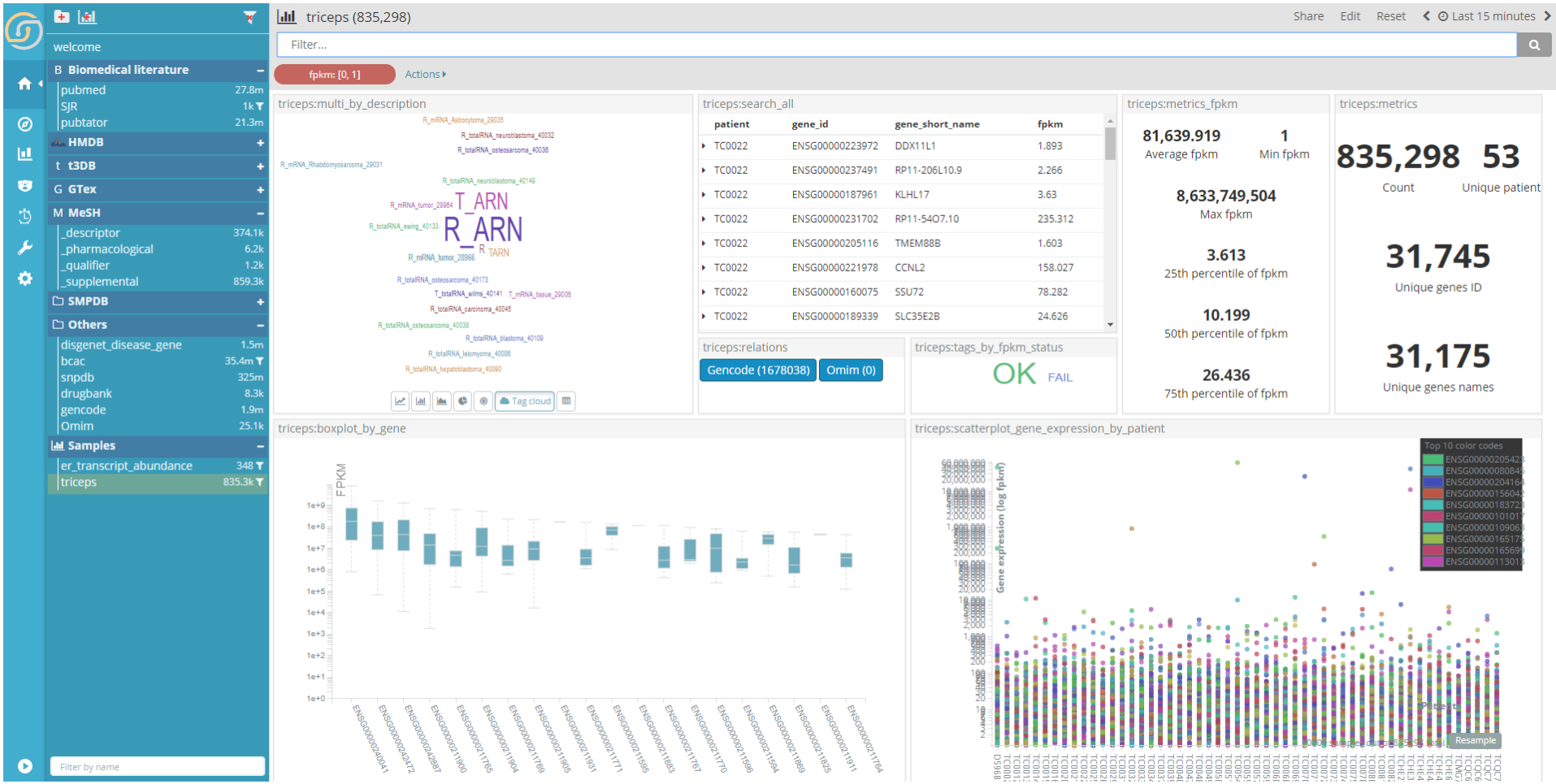
Filter pubmed by impact factors



A little overview of Kibio



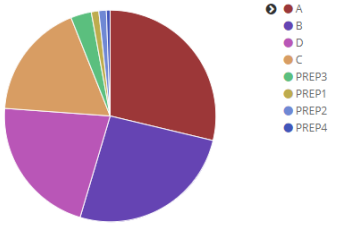
Transcriptomics data



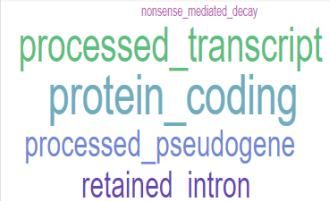
welcome
B Biomedical literature
pubmed 168.2k
SjR 1k
pubtator 21.3m
HMDB
t t3DB
G GTex
M MeSH
_descriptor 374.1k
_pharmacological 6.2k
_qualifier 1.2k
_supplemental 859.3k
SMPDB
Others
disgenet_disease_gene 1.5m
bcac 35.4m
snpdb 325m
drugbank 8.3k
gencode 1.9m
Omim 25.1k
Samples
er_transcript_abundance 348
triceps 835.3k

er_transcript_abundance_v2:pie_by_triplet_type

er_transcript_abundance_v2:pie_by_patient



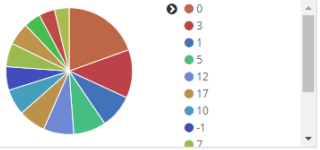
er_transcript_abundance_v2:tags_by_transcript_type



er_transcript_abundance_v2:relations

Gencode (1216) OMIM (13)

er_transcript_abundance_v2:Pie_by_day

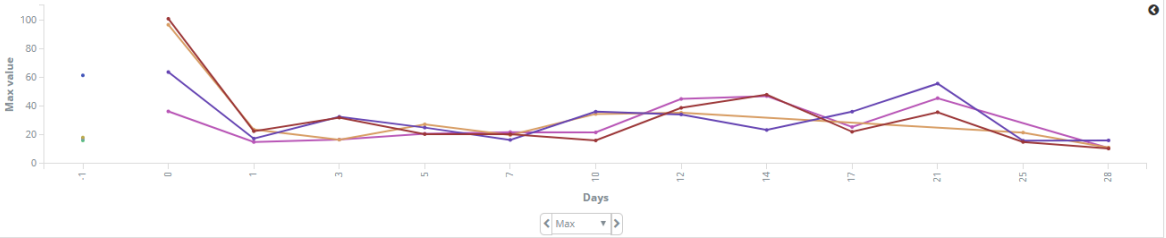


er_transcript_abundance_v2:table_b...

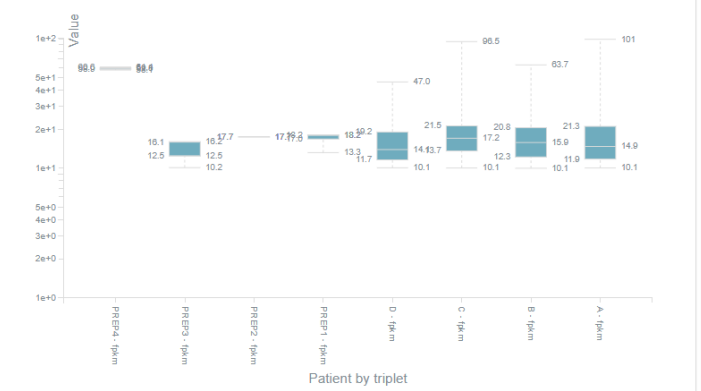
symbol.keyword: Descending	Count
FAM83A	167
NA	66
KRTDAP	22
KRT5	20
KRT14	15
S100A2	15
WDR74	10
EEF1A1	9
KRT6A	8
ANXA2	4

Export: Raw Formatted

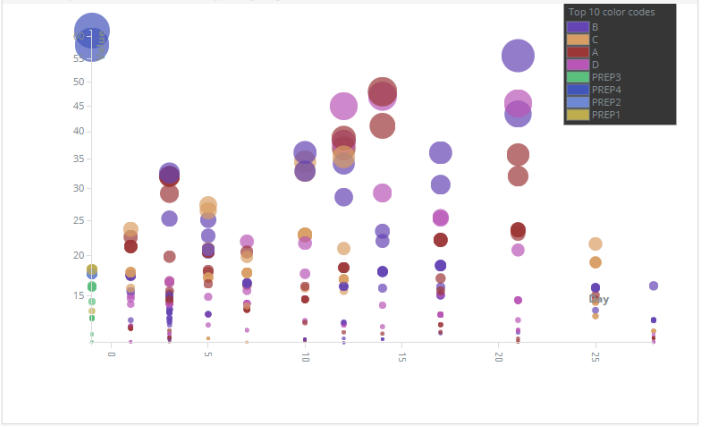
er_transcript_abundance_v2:multi_by_days



er_transcript_abundance_v2:boxplot_by_patient_and_triplet_type



er_transcript_abundance_v2:scatterplot_by_day



er_transcript_abundance_v2:EnhancedTable

patient	day	triplet_type	transcript_type	symbol	ensembl_gene	id	value
A	0	fpkm	processed_pseudogene	NA	ENSG00000226958	ENSG00000226958	100.691
C	0	fpkm	processed_pseudogene	NA	ENSG00000226958	ENSG00000226958	96.538
A	0	fpkm	processed_pseudogene	NA	ENSG00000226958	ENST00000419932	89.206
B	0	fpkm	processed_pseudogene	NA	ENSG00000226958	ENST00000419932	63.666
PREP4	-1	fpkm	processed_pseudogene	NA	ENSG00000226958	ENST00000419932	61.381
PREP4	-1	fpkm	processed_pseudogene	NA	ENSG00000226958	ENSG00000226958	58.121
C	0	fpkm	processed_pseudogene	NA	ENSG00000226958	ENST00000419932	56.917
B	21	fpkm	processed_pseudogene	NA	ENSG00000226958	ENSG00000226958	55.656

Conclusions

- Fast searches within linked databases
- Better exploration
- Intuitive portal
- Import new data
- Future implementations:
  - Analyticals tools (PCA, cohorts comparison)
  - Biomarker discovery tools

Global view


Biomarkers Signature Identification
Analytical Development (MixOmics, Machine Learning)
Big Data Omics (Your Data, Multi-Omics Data, Metagenomics & Open data)