

Des données phénotypiques et « omiques »  
pour la **compréhension** des mécanismes  
et la **prédiction** de caractères de production  
chez les ruminants

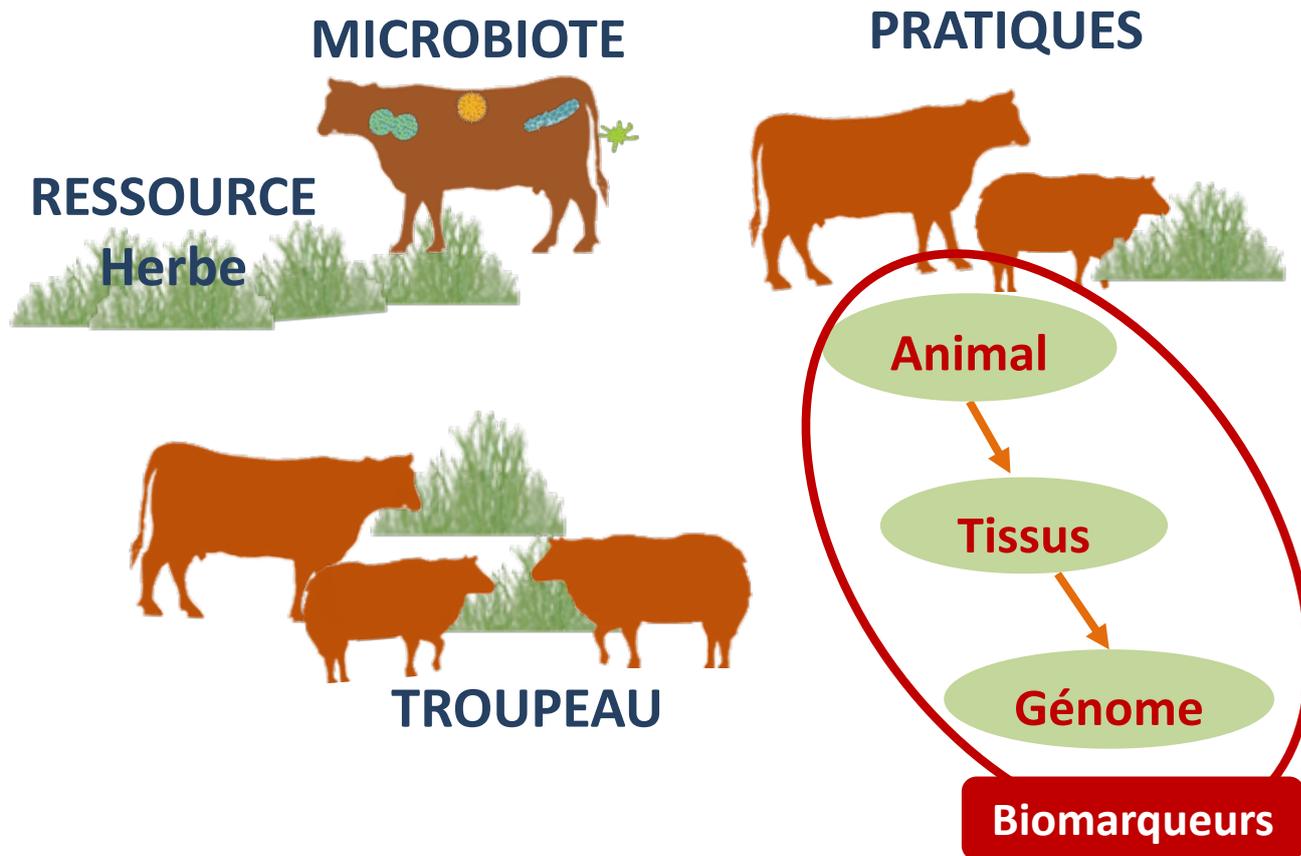
Muriel Bonnet

# ➤ Qui sommes nous ?



**UMR Herbivores** (site *INRAE de Clermont-Ferrand Theix*) 122 agents titulaires, 5 équipes de recherches dont l'équipe **Biomarqueurs des performances, de l'adaptation et des qualités**

# ➤ EQUIPE BIOMARQUEURS dans l'UMR Herbivores



- 15 Biologistes
- 1 Mathématicienne (A. Imbert 1/05/2022)
- 1 Bioinformaticien (1/09/2020)



Qualité  
des produits

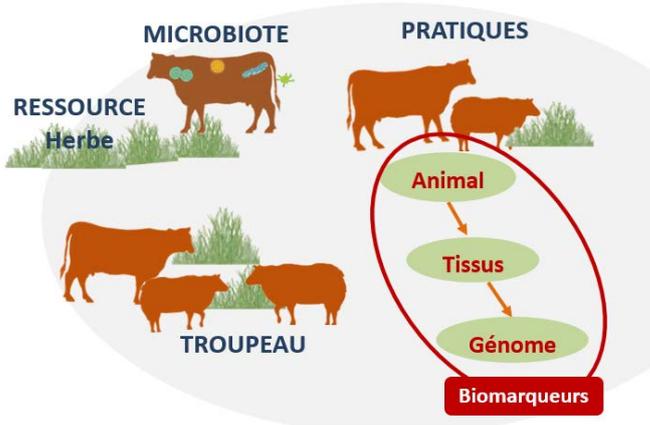


Capacités adaptatives  
robustesse



Efficience  
alimentaire

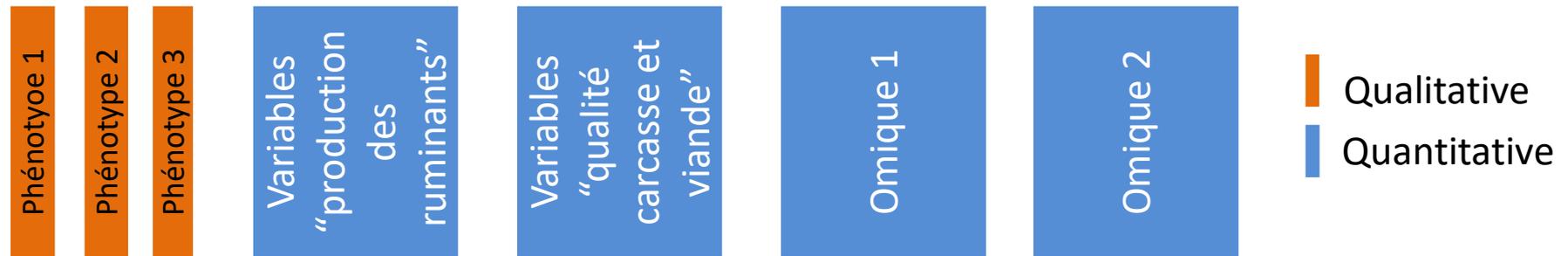
## ➤ Les questions et les objectifs de recherche



Comment des facteurs intrinsèques (âge, race, génétique...) ou extrinsèques (alimentation, environnement, stress...) aux ruminants modifient l'expression du génome dans les tissus ou fluides et en lien avec les productions (lait & viande)?

1. Comprendre les mécanismes biologiques à l'origine des phénotypes des ruminants producteurs (teneur en lipides des viandes, efficacité alimentaire, la composition du lait...)
2. Prédire les phénotypes à partir de molécules ou autres indicateurs

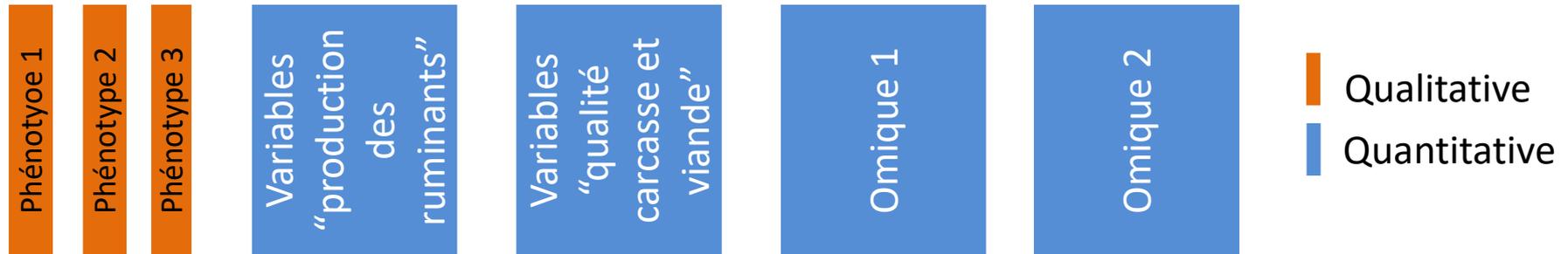
# ➤ Les données et leur traitements



## Les étapes de l'analyse :

1. Préparation des données et statistiques descriptives des données
2. Analyses statistiques et bioinformatiques pour comprendre les bases biologiques des phénotypes
3. Analyses statistiques pour prédire les phénotypes

## ➤ Les données et leur traitements



### Les étapes de l'analyse :

1. Préparation des données et statistiques descriptives des données
2. Analyses statistiques et bioinformatiques pour comprendre les bases biologiques des phénotypes
  - *Analyses univariées et bivariées*
  - *Analyses multivariées*
  - *Fouille de données « bioinformatiques »*
3. Analyses statistiques pour prédire les phénotypes

# ➤ Les analyses univariées et bivariées



**Analyses univariées** Les données : Quantitative

La question : Quelle variance d’une variable selon le phénotype 1 ou x ?

**Analyses bivariées** Les données : 2 Quantitatives ou 1 quantitative et 1 qualitative:

La question : Quelle est la proximité ou l’écart entre 2 variables ?

**Les méthodes**

Une juste adéquation entre le plan expérimental, les données et la question de recherche

**Nos « pratiques »**

- Un package pour identifier un protéome différentiel entre 2 groupes (Bazile et al., 2019)
- Le package Limma pour des analyses différentielles multifactorielles (modèles linéaires; Ritchie et al., 2015; Phipson et al., 2016), à partir de données zootechniques, transcriptomiques ou protéomiques.

# ➤ Développement d'un script pour l'analyse différentielle de protéome entre 2 groupes (Bazile et al., 2019)

**Management of proteomic data to identify differentially abundant proteins according to one discriminant factor and to correlate their abundance with the value of this factor, DOI : 10.5281/zenodo.2539329**

[https://github.com/jane-bzl/Differential abundance and correlation to one factor](https://github.com/jane-bzl/Differential%20abundance%20and%20correlation%20to%20one%20factor)

Test Shapiro  
Homogénéité  
des variances?

Tests de Student  
ou Kruskal-Wallis

Corrélations  
Pearson, Kendall  
ou Spearman  
(table + graphe)

Utilisé dans la publication **Pathways and biomarkers of marbling and carcass fat deposition in bovine revealed by a combination of gel-based and gel-free proteomic analyses. Bazile et al. 2019, Meat science**

## ➤ script R de Bazile et al. : la table de données

Groups made according to the biological conditions

Identification of the samples (number, letters or both)

Quantitative values related to the biological conditions

Values of proteins abundances

group	id	adiposity	Protein1	Protein2	Protein3	Protein4
group1	1	6.3	230795072	72392559	-1308293	54238
group1	6	3.7	267670438	49956496	-1099355	23984
group1	3	4.1	1736366	57744002	-1188752	3574699
group1	4	5.1	72392559	65141951	-1234186	368571
group1	5	6.8	203517615	64672505	-1551479	65781
group2	2	3.0	275028504	64089037	-1125215	5348673
group2	7	2.1	2043358	41658946	-1132811	687253
group2	8	2.2	64089037	50876533	-1049944	68357
group2	9	2.3	231769965	37805488	-1134805	573521
group2	10	2.2	305818430	39682550	-1086689	68967

## ➤ script R de Bazile et al. : la table de résultats

Shapiro test reports the normal (or not) distribution of an abundance (TRUE=normal distribution)

p-values from Student-t-test for normally or Kruskal-Wallis test for non-normally distributed (TRUE= significant pval)

p-values from Student-t-test whatever the distribution of the data (TRUE= significant pval)

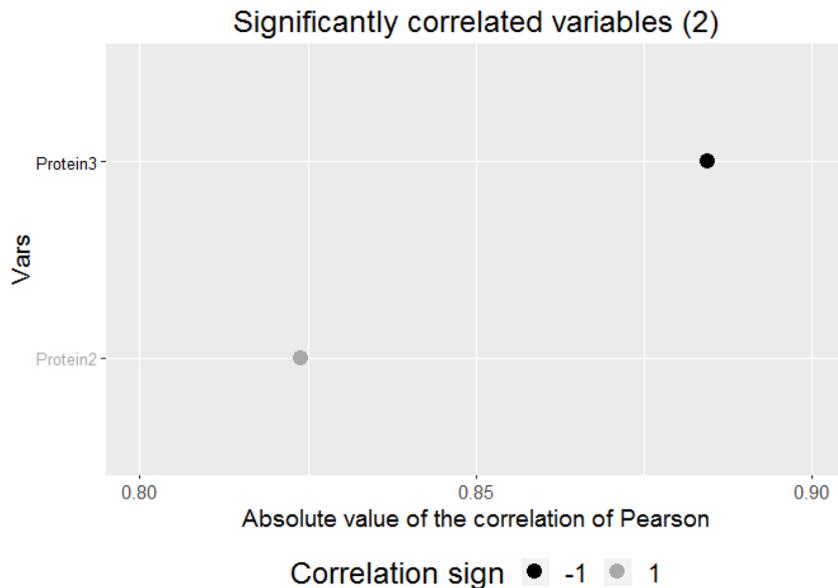
p-values from Kruskal-Wallis - test whatever the distribution of the data (TRUE= significant pval)

	pval.shapiro	normality	pval.testTor KW	signif.testTorK W	pval.Ttest	signif.Ttest	pval.K W	signif.KW
Protein1	0.089	TRUE	0.801	FALSE	0.801	FALSE	0.465	FALSE
Protein2	0.403	TRUE	0.040	TRUE	0.040	TRUE	0.047	TRUE
Protein3	0.018	FALSE	0.047	TRUE	0.061	TRUE	0.047	TRUE
Protein4	0.000	FALSE	0.175	FALSE	0.675	FALSE	0.175	FALSE

- Results saved in your directory as an Excel file (.csv)
- Add or not a correction for multiple tests depending on your objectives

# ➤ Script R de Bazile et al. : les résultats de corrélation

Correlations computed either by the Pearson test (for normally distributed data) or Kendall/Spearman test (for non normally distributed data)



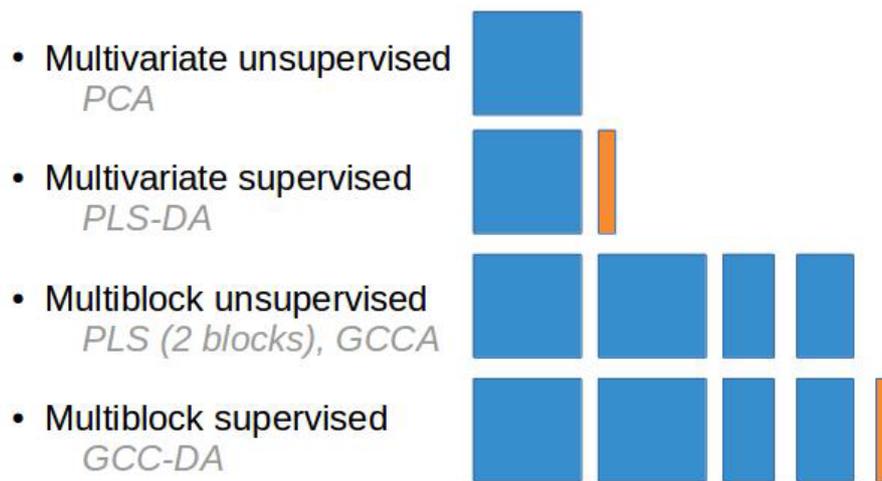
	adiposity	Protein1	Protein2	Protein3	Protein4
adiposity	1	0,06723532	0,82398759	-0,8845269	0,15145677
Protein1	0,06723532	1	-0,0077549	-0,04223508	0,04956437
Protein2	0,82398759	-0,0077549	1	-0,60280322	0,245925
Protein3	-0,8845269	-0,04223508	-0,60280322	1	0,16626722
Protein4	0,15145677	-0,04956437	0,245925	0,16626722	1

- A figure to review the proteins significantly correlated with one parameter
- 2 tables produced and save to your directory :
  - ✓ one with the correlation values
  - ✓ one with the p-value of the correlations

# ➤ Les analyses multivariées: mixOmics



La question : Quelle combinaison de variables contribue à la signature moléculaire (ou combinaison de variables moléculaires ou « animales ») du phénotype ?

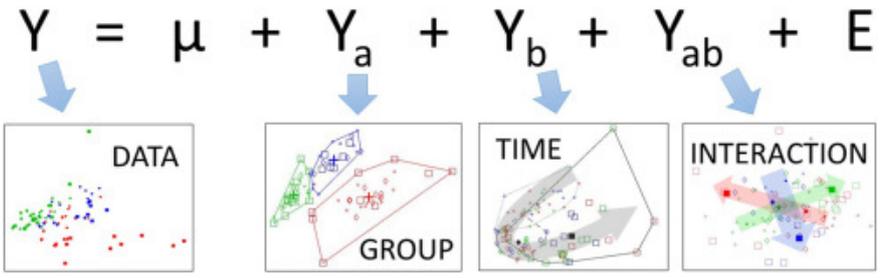


Collaboration avec S. Dejean

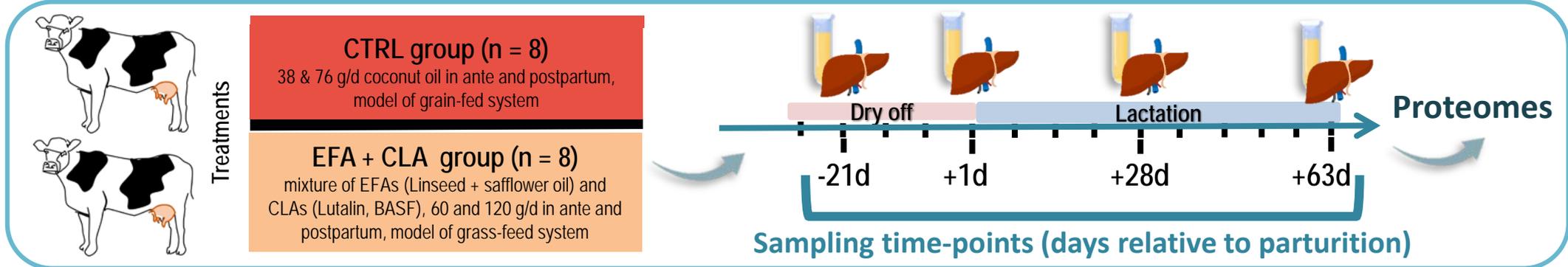
Veshkini et al., 2022 a et b

# ➤ Les analyses multivariées: ANOVA simultaneous component analysis (ASCA) pour des données acquises en cinétique

ASCA : méthode multivariée applicable à des schémas expérimentaux complexes et multifactoriels

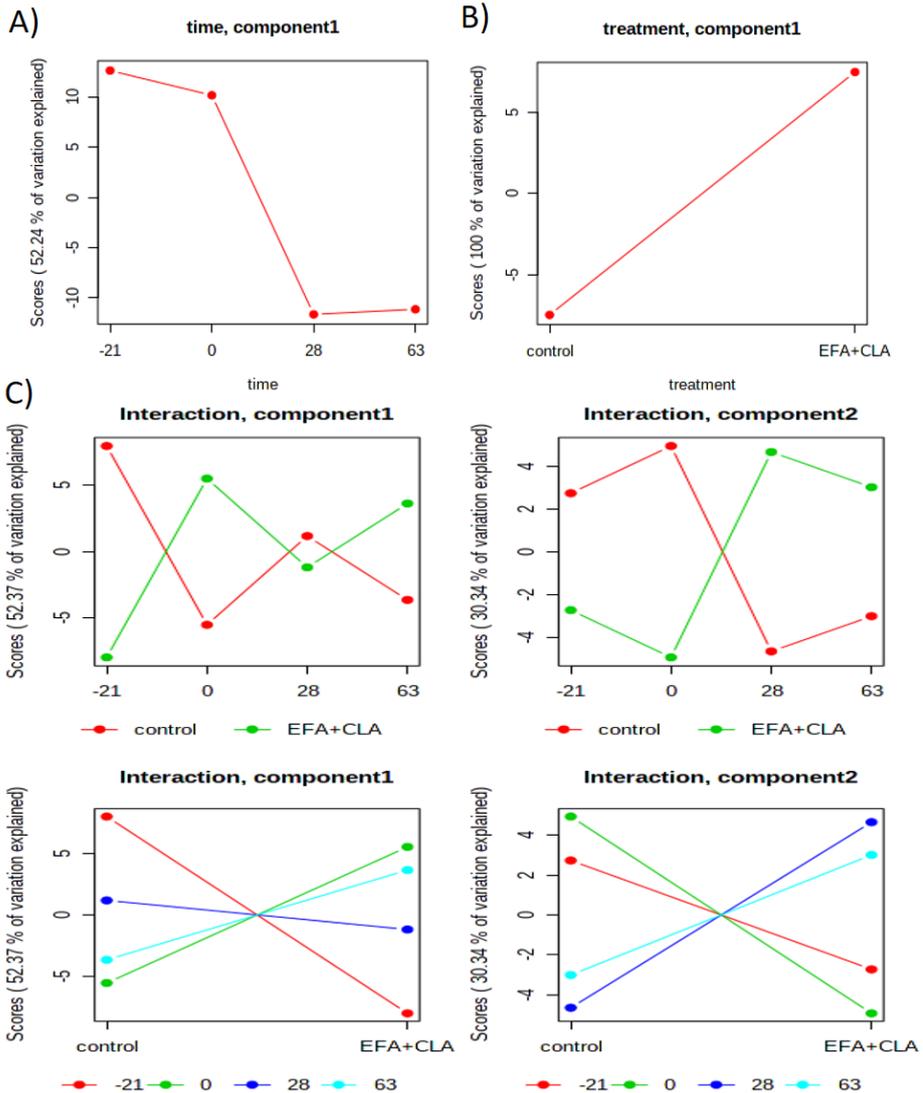


Bertinetto et al., 2020, A tutorial review



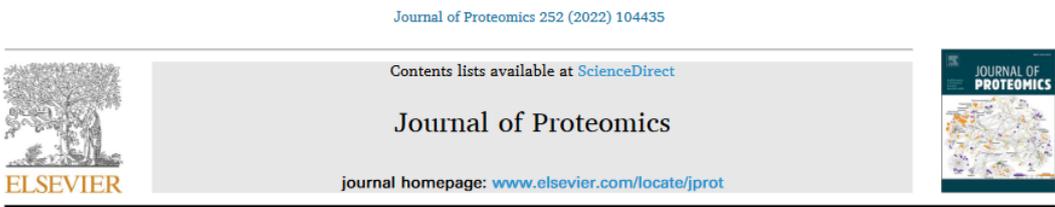
Veshkini et al., 2022 c

# Les analyses multivariées: ASCA appliquée à des données protéomiques acquises en cinétique



Sur 1681 protéines identifiées et quantifiées dans le foie, les abondances sont différentielles selon:

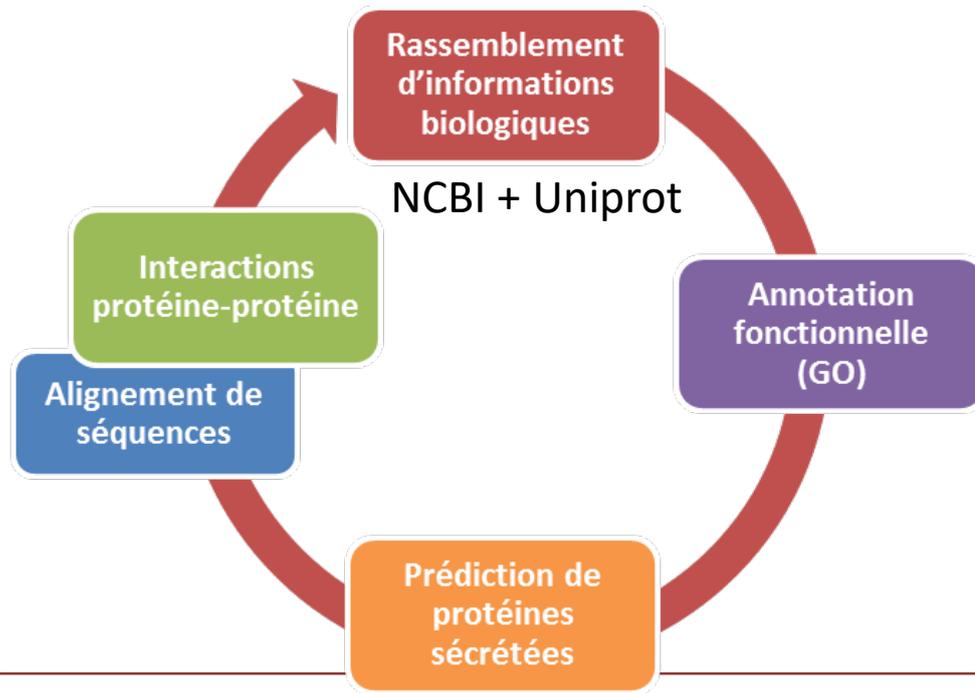
- la date relative au vêlage pour 116 protéines
- le traitement pour 43 protéines
- l'interaction pour 97 protéines



Longitudinal liver proteome profiling in dairy cows during the transition from gestation to lactation: Investigating metabolic adaptations and their interactions with fatty acids supplementation via repeated measurements ANOVA-simultaneous component analysis

Arash Veshkini<sup>a,b,c,d</sup>, Harald M. Hammon<sup>b,\*</sup>, Helga Sauerwein<sup>a</sup>, Arnulf Tröscher<sup>e</sup>, Didier Viala<sup>c</sup>, Mylène Delosière<sup>c</sup>, Fabrizio Ceciliani<sup>d</sup>, Sébastien Déjean<sup>f</sup>, Muriel Bonnet<sup>c,\*</sup>

# ➤ La fouille de données avec des outils bioinformatiques



## Tests de ProteINSIDE

Jeu de données test composé de 133 protéines humaines et bien connues :

- 32 protéines de la glycolyse
- 79 hormones
- 1 protéine en double
- d'autres protéines uniques



[www.proteinside.org](http://www.proteinside.org)

Kaspric et al., 2015 a et b

ProteINSIDE fonctionne avec 6 espèces

Espèces cibles :

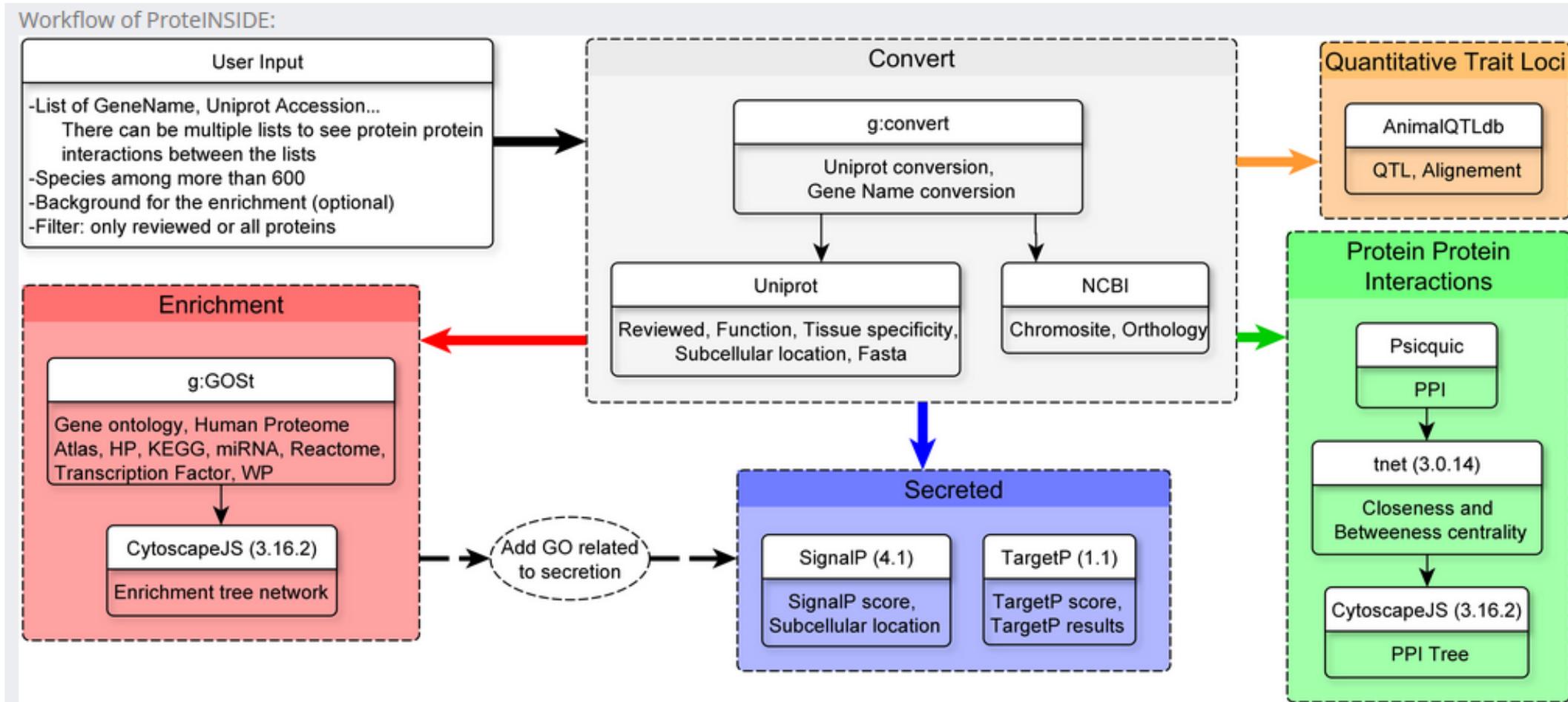
- Bovin 
- Mouton 
- Chèvre 

Espèces modèles :

- Homme 
- Souris 
- Rat 

## ➤ La version 2 de proteINSIDE ([https://umrh-bioinfo.clermont.inrae.fr/ProteINSIDE\\_2/](https://umrh-bioinfo.clermont.inrae.fr/ProteINSIDE_2/))

- Module de conversion des identifiants
- Plus de 600 espèces considérées
- Des résultats visuels pour la Gene Ontology
- La prediction du secrétome et bientôt du surfaceome



Tournayre et al., 2021 Jobim p. 16

# ➤ WEB DEMO

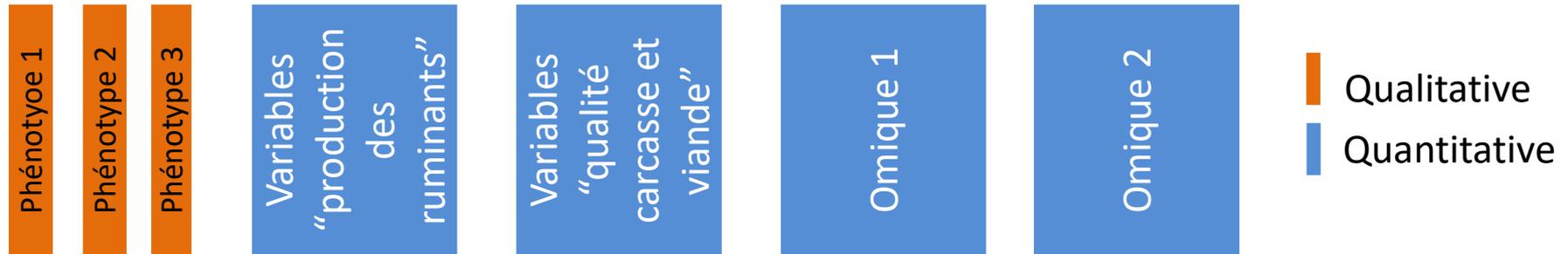
**Jeu de données test composé de 133 protéines humaines et bien connues :**

- **32 protéines de la glycolyse**
- **79 hormones**
- **1 protéine en double**
- **d'autres protéines uniques**



2022\_6\_7\_dEGM9UKXZcRjSLYo1gyCN2pzA

# ➤ Les données et leur traitements



## Les étapes de l'analyse :

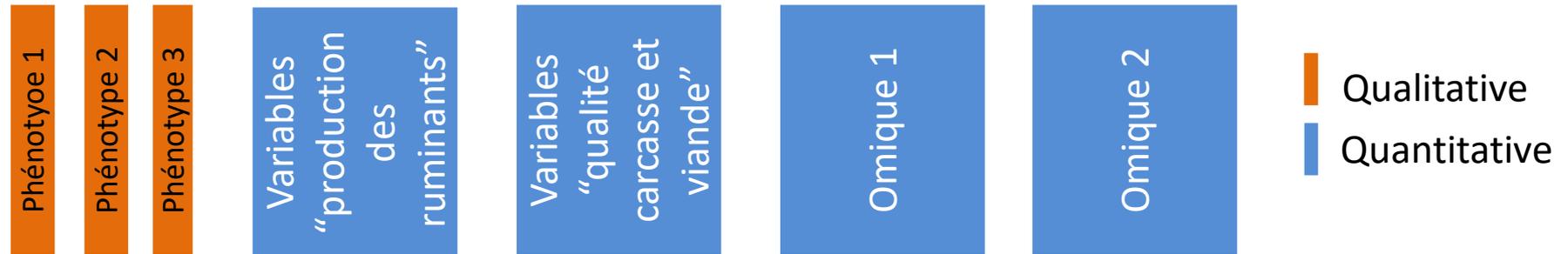
1. Préparation des données et statistiques descriptives des données
2. Analyses statistiques et bioinformatiques pour comprendre les bases biologiques des phénotypes

### 3. Analyses statistiques pour prédire les phénotypes

- *Régression Logistique*
- *Forêt aléatoire*
- *Arbres de décision*
- *(Réseaux de neurone)*
- *Analyse discriminante linéaire*
- *Machine à vecteurs de support....*

(Ellies-Oury et al., 2019; Gagaoua et al., 2019, Soulat et al., 2018, 2019; Bonnet et al., 2020...)

## ➤ Prédiction des phénotypes



La question biologique : Parmi les données, quelles sont celles qui prédisent le phénotype 1 ou x

### Les données :

- des variables qui qualifient/quantifient le phénotype avec une bonne précision
- Une population suffisamment grande pour définir 2 populations : une d'apprentissage et l'autre de validation
- des variables susceptibles de prédire le phénotype dont le nombre est à adapter à la méthode choisie (cf temps de calcul des modèles), à la population étudiée et à l'objectif d'application du résultat

# ➤ ModVarSel (MP. Ellies et coll. Avec M. Chavent, J Saracco, INRIA)

**Une méthode computationnelle pour sélectionner simultanément le meilleur modèle de régression et les variables les plus pertinentes**

Différents modèles de régression (incluant la sélection des variables) :

1- paramétrique (régression linéaire multiple (MLR)),

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \varepsilon$$

2- semi-paramétrique (sliced inverse regression (SIR))

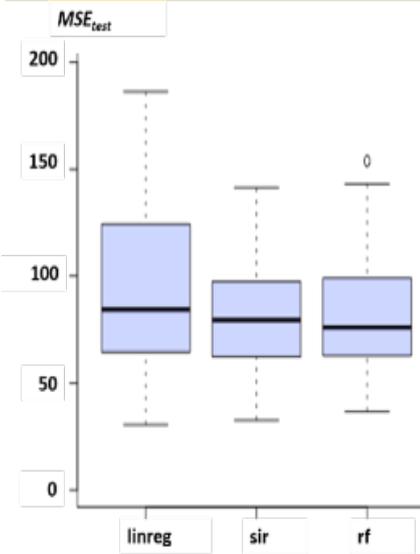
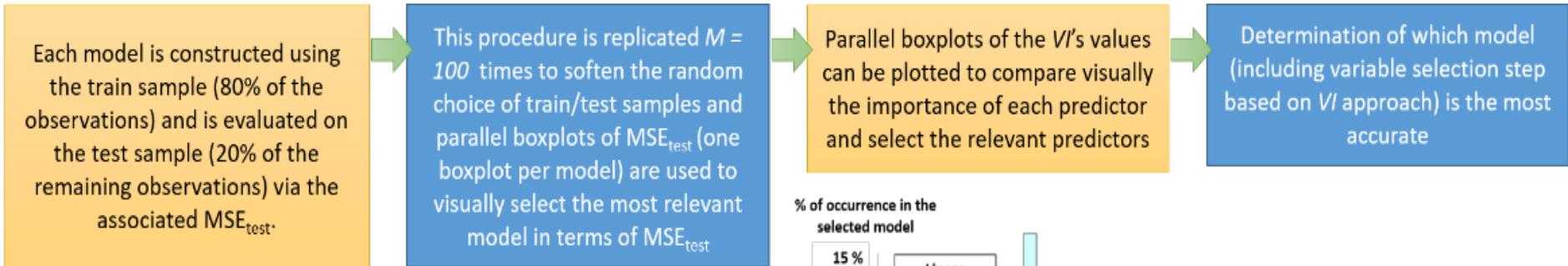
$$Y = f\left(\sum_{j=1}^p \beta_j X_j\right) + \varepsilon$$

3- and non-paramétrique (random forests (RF))

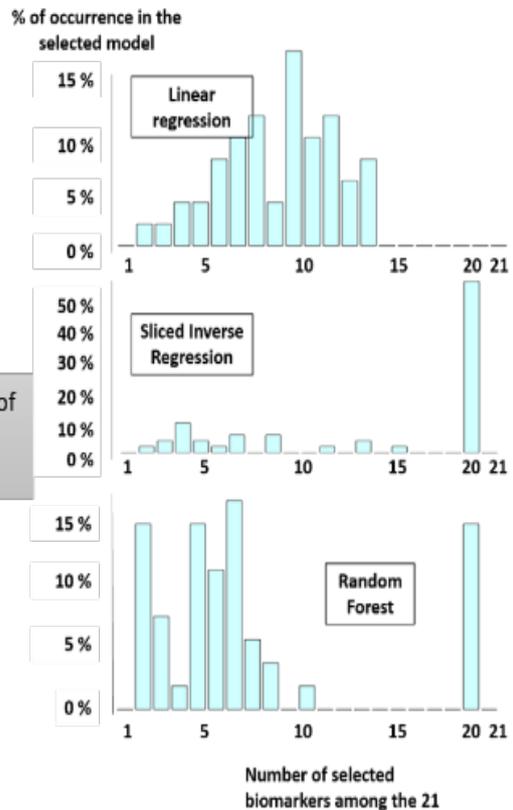
$$Y = f(X) + \varepsilon$$

Les données = une liste retrainte (moins de 50 variables) de molécules, dans la pratique nous utilisons les listes d'ID issus des analyses uni- ou pluri-variées

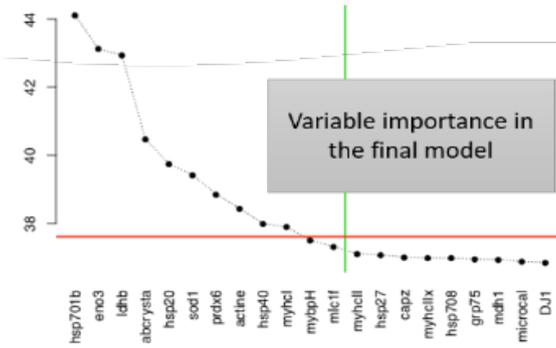
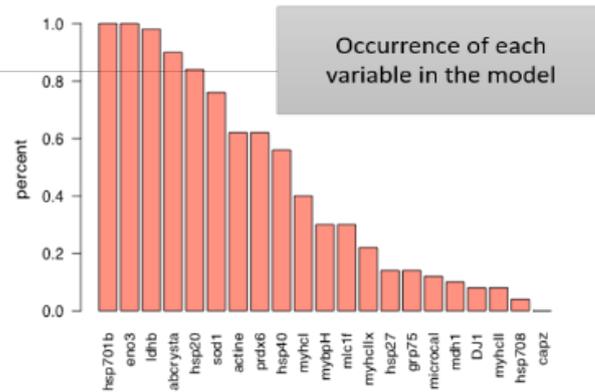
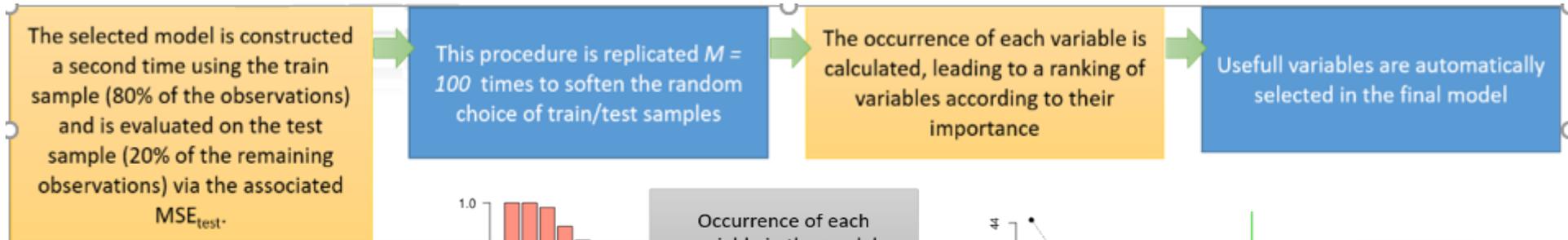
# ➤ ModVarSel : comment sélectionner un modèle?



Information of the number of usefull variables automatically selected

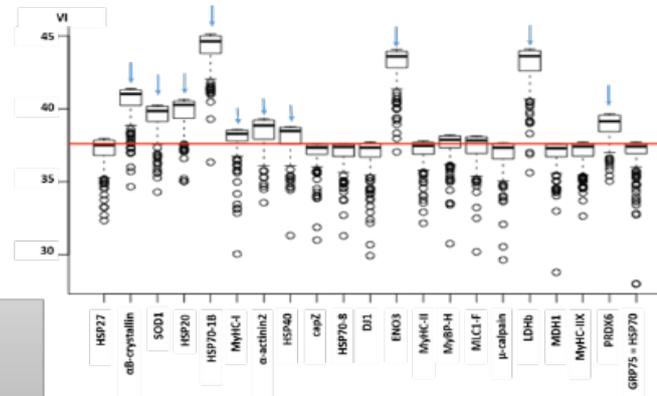


# ➤ ModVarSel : comment sélectionner des variables prédictives ?



Les données :

- une mesure mécanique de la tendreté de la viande
- 21 abondances de protéines
- 71 bovins

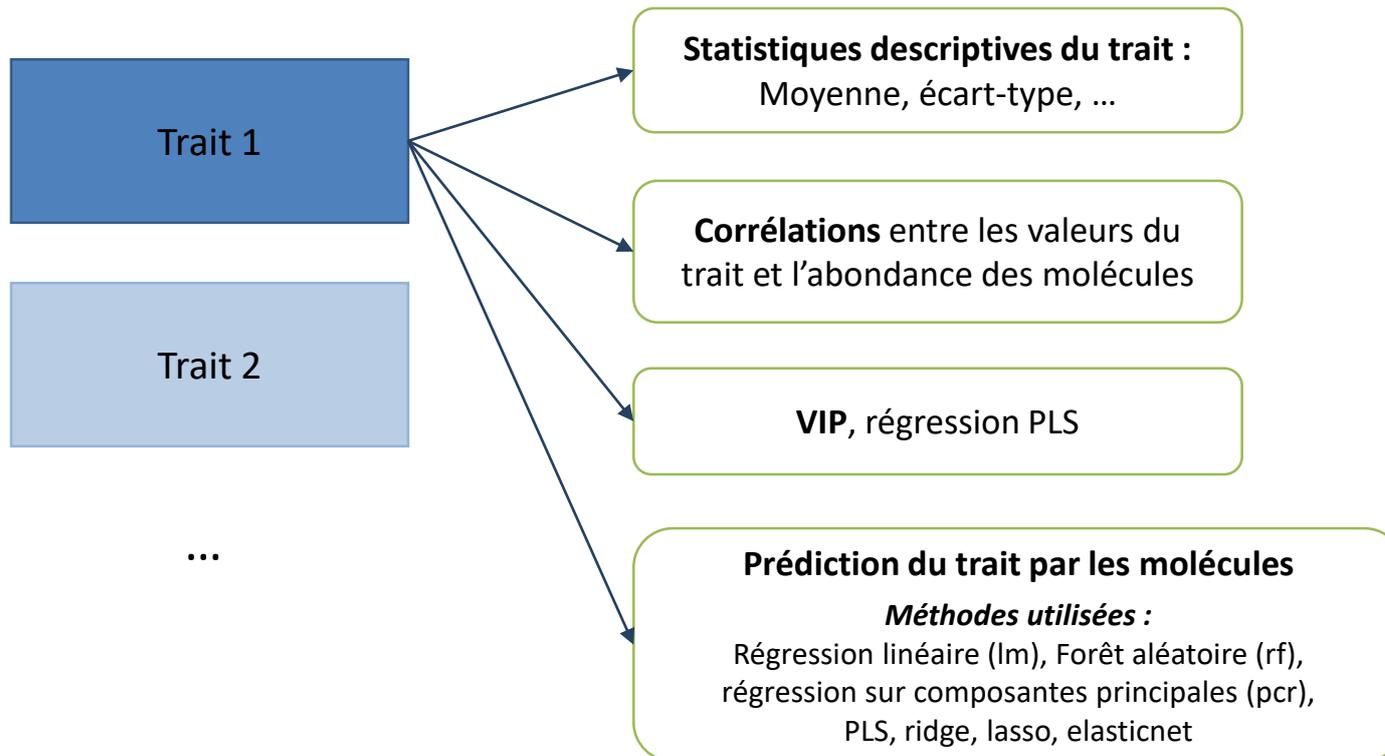


Selected variables

## ➤ Tout savoir sur ModVarSel

- Pour installer le R package modvarsel sur github :  
<https://github.com/chavent/modvarsel>.
- DOI : <https://zenodo.org/record/1445554#.W7YrYPY69PY>
- Voir également :  
<https://chavent.github.io/modvarsel/modvarsel-intro.html>
- Un exemple d'application dans Ellies-Oury et al., 2019 Sci Reports

# ➤ Implémentation d'un script R avec différentes méthodes prédictives



## 1. Sélection des molécules pour chaque méthode

- **lm** : recherche exhaustive des meilleurs modèles à X variables sur critère AIC
- **rf, pcr, PLS** : sélection des biomarqueurs sur critère **VI** (*modvarsel*, Ellies-Oury et al., 2019)
- **lasso, elasticnet** : sélection des biomarqueurs par pénalisation des coefficients du modèle

## 2. Sélection de la meilleure méthode

Estimation de l'erreur de prédiction par validation croisée (RMSE-CV)

Cougoul et al., non publié

# ➤ Un outil tout en un : rapport automatisé en html

**ANALYSE BIOMARQUEURS**  
04 septembre 2020  
Lipivimus Mauron, sensorielle

TG JUT FLAV APP cMQ4 oMQ4 **ALL**

Trait description Correlations ACP VIP (PLS) Predictions

sélection du trait

sélection de l'analyse

résumé de tous les traits

TG JUT FLAV APP cMQ4 oMQ4 **ALL**

Correlations **Predictions**

Variable	Model	R2	RMSECV	nb_var	MLC_1F	Eno3	GRP75	PRDX6	PGM1	HSP70_1B
<b>TG</b>	lm8	0.90	0.7±0.2	7	-0.65	-0.32		-0.16	0.22	-0.99
<b>JUT</b>	lm3	0.44	0.64±0.18	2					0.29	
<b>FLAV</b>	lm6	0.79	0.36±0.12	6	-0.71	-0.50	0.69			
<b>APP</b>	lm4	0.70	0.74±0.25	4	-0.58	-0.47	0.79	-0.27		
<b>cMQ4</b>	lm4	0.78	5.19±1.21	4	-0.62	-0.48	0.79	-0.31		
<b>oMQ4</b>	lm8	0.79	5.45±1.26	4	-0.63	-0.48	0.78	-0.33		

# ➤ Take home messages : le point de vue du biologiste

- Ne pas sous-estimer le temps nécessaire à la préparation et à la vérification de la qualité des données
- Avant toute analyse se poser quelques questions simples : quel objectif scientifique? les données permettent-elles d'atteindre cet objectif? Quel plan d'analyse mettre en place ? Quelles méthodes mobilisées?
- Avant de faire des analyses multi-blocs (intégration multi-omiques), exploiter les analyses et connaissances produites intra-bloc et vérifier la plus-value de faire du multi-blocs
- Intégrer des données phénotypiques et omiques requiert des interactions entre mathématiciens/informatiques/biologistes pour la mise en place d'outils répondant aux questions des biologistes => favoriser des communautés inter-disciplinaires (Biopuce, Digit-Bio...)

# UMR Herbivores

Equipe biomarqueurs des performances, de l'adaptation et des qualités



➤ **Merci pour votre attention**