Xavier Grand 7 Janvier 2020

Master Sciences Numériques pour la Santé, parcours BCD, 2019-2020.

Analyse de données multi-omiques, la suite logicielle mixOmics.

Tuteur scientifique : Sébastien Déjean, Ingénieur de recherche en calcul scientifique.
Institut de Mathématiques de Toulouse.
sebastien.dejean@math.univ-toulouse.fr

Tuteur pédagogique : Thérèse Commes, Professeur des Universités.
Institute for Regenerative Medicine & Biotherapy, Montpellier.
therese.commes@inserm.fr

Table des matières

Remerciements :

Mes premiers remerciements vont à Sébastien Déjean, qui a su se montrer disponible et pédagogue pour me conseiller dans ce projet bibliographique, tout en me laissant une grande autonomie dans mes choix sur la forme de ce rapport. Je souhaite également le remercier de m'avoir montré qu'un dialogue entre un statisticien et un biologiste peut être très enrichissant. J'envisage avec enthousiasme notre collaboration future, dans l'espoir d'une fin administrative heureuse.

Je souhaite remercier Harold Duruflé, pour nos discussions et ses bons conseils sur les articles à lire pour la réalisation de ce projet.

Je remercie Thérèse Commes pour les commentaires et le regard constructif et bienveillant qu'elle pourrait avoir sur ce travail.

I. Introduction

Vocabulaire :

En préambule, il apparaît nécessaire de définir les termes qui seront abordés dans ce travail de bibliographie. Nous définirons par données « omiques », les données conceptuelles et issues de méthodes d'analyses expérimentales relatives aux différents aspects de la biologie, mesurables par un expérimentateur. Il s'agit, à titre d'exemple, des données du transcriptome, ou transcriptomique, relatives à l'expression des gènes chez un être vivant ; des données du protéome, ou protéomique, relatives à l'accumulation, la présence, la variabilité des protéines chez un être vivant ; des données du métabolome, ou métabolomique, relatives aux métabolites produits par un être vivant. Nous parlerons de cette manière de toutes données biologiques considérées en un ensemble identifiable par l'expérimentateur, le biologiste ou encore l'analyste de ces données. Au terme « omique », peut être accolé le terme technologie. Ainsi, les technologies omiques concernent les technologies biochimiques à haut débit qui permettent de mesurer les ensembles omiques, donc d'obtenir des données omiques.

Contexte biologique :

Prenons pour contexte la biologie végétale. Comme tous les êtres vivants, les végétaux sont soumis aux variations de l'environnement. Encore plus que pour les êtres mobiles, la condition sessile de végétaux leur a imposé une grande capacité d'adaptation. À l'inverse d'un animal qui peut migrer pour la saison froide, la plante doit déployer des stratégies pour résister au froid. Ces changements environnementaux, nommés stress abiotiques, ne sont pas les seuls à avoir un impact les plantes. Les plantes tombent malades elles aussi, et à l'instar des stress abiotiques, elles subissent d'autres stress nommés les stress biotiques. Par ailleurs, au cours de leur développement, les plantes opèrent des changements dans leur propre fonctionnement. Elles ont des périodes de croissance, de floraison, de remplissage de grains ou de fruits, voire de stase. Ces changements macroscopiques sont régis par des changements biochimiques, par l'expression différentielle de gènes, par l'activation ou l'accumulation de protéines, par exemple.

En 1958, Francis Crick (Prix Nobel de médecine 1962) énonçait son Dogme central de la biologie moléculaire : « *L'ADN dirige sa propre réplication en ADN identique, ainsi que sa transcription en ARN, pouvant ou non être traduit en protéines.* » traduit depuis Nature 1970 Central Dogma of Molecular Biology. Nous pouvons extrapoler sur cette base : le génome dirige sa propre réplication, ainsi que le transcriptome, qui régit le protéome. De manière très simplifiée, il est enseigné qu'un gène est transcrit en un ARN messager qui est traduit en une protéine et a pour résultat un impact sur un caractère.

Chez les plantes, il existe quelques gènes qui semblent fonctionner de cette manière. Ce sont les gènes de résistance, nommés ainsi en raison du caractère de résistance à une maladie qu'ils confèrent. Ces gènes, de la famille des NBS-LRR (« Nucleotid-Binding-Site Leucine-Rich-Repeats »), sont des gènes majeurs de la résistance des plantes et leur présence dans le génome suffit à immuniser la plante d'un agent pathogène (Dangl et Jones, 2001, 2006). En revanche, il est bien évident qu'un gène unique ne peut être responsable du caractère de résistance de la plante. Il est nécessaire, mais pas suffisant, il intervient avec un ensemble d'autres gènes qui sont exprimés ou non, et qui, ensemble, auront un effet sur le phénotype.

Une discipline s'attache à décrire les actions synergiques entre les différents éléments (gènes, protéines, métabolite, etc.) : c'est la biologie des systèmes. Cette discipline tend à décrire ou étudier un être vivant d'un point de vue global. Par conséquent, un chercheur en biologie des systèmes étudiera l'ensemble des informations biologiques, à sa disposition ou qu'il aura produites, lui

permettant de répondre à sa question. Par exemple, il ne considèrera pas seulement le transcriptome pour expliquer une différence (ou ressemblance) entre différents individus, mais le génome, son expression et le protéome qui en résulte de manière intégrée.

Motivations scientifiques vis-à-vis de ce sujet de projet bibliographique :

D'autre part, l'expérimentation en biologie est contrainte par des aspects pratiques et financiers qui limitent les possibilités d'un chercheur. Aussi, un plan d'expérience est, la plupart du temps, le résultat d'un compromis entre les moyens à disposition qui déterminent le dimensionnement de l'expérience, et un nombre satisfaisant de conditions à étudier pour répondre à une question. Ainsi, une approche d'analyse appliquée à plusieurs jeux de données issus de différentes expérimentations peut être réalisée pour augmenter la taille de l'échantillonnage, et donc la puissance statistique des résultats qui en découlent. On parle alors de méta-analyse.

Pour ma part, lors d'un précédent projet de post-doctorat, j'ai réalisé une méta-analyse du transcriptome chez les céréales dans le but d'identifier des gènes, conservés entre différentes espèces, potentiellement impliqués dans la réponse à des stress abiotiques tels que la sécheresse. Cette méta-analyse concernait 3 espèces (le riz, le blé et le maïs) et consistait en l'analyse de résultats de puces à ADN de type Affymetrix®, entre 5 et 11 expérimentations par espèce. Les méthodes d'analyses statistiques que j'ai mises en œuvre lors de ce travail m'ont permis d'identifier un certain nombre de gènes différentiellement exprimés en condition de sécheresse chez les trois espèces. Toutefois, à l'issue de ce projet de bibliographie, je fais le constat que les méthodes que j'ai pu utiliser pour cette méta-analyse ne sont pas les meilleures d'un point de vue statistique. Nous verrons, dans la suite de ce rapport, quelles méthodes j'aurais pu ou dû appliquer aux données pour réaliser cette méta-analyse.

Accessibilité des données omiques, émergence de bases de données multi-omiques :

Depuis l'invention de la « PCR » (« Polymerase Chain Reaction », Kary Mullis, prix Nobel de chimie 1993, Mullis K. *et al.* Specific Enzymatic Amplification of DNA *in vitro*: The Polymerase Chain Reaction, 1986), les progrès techniques en biologie moléculaire ont permis de mesurer de plus en plus, et de manière plus précise les changements biochimiques chez les êtres vivants. Les données de séquençages de génomes s'accumulent, de même que les mesures d'expression de gènes.

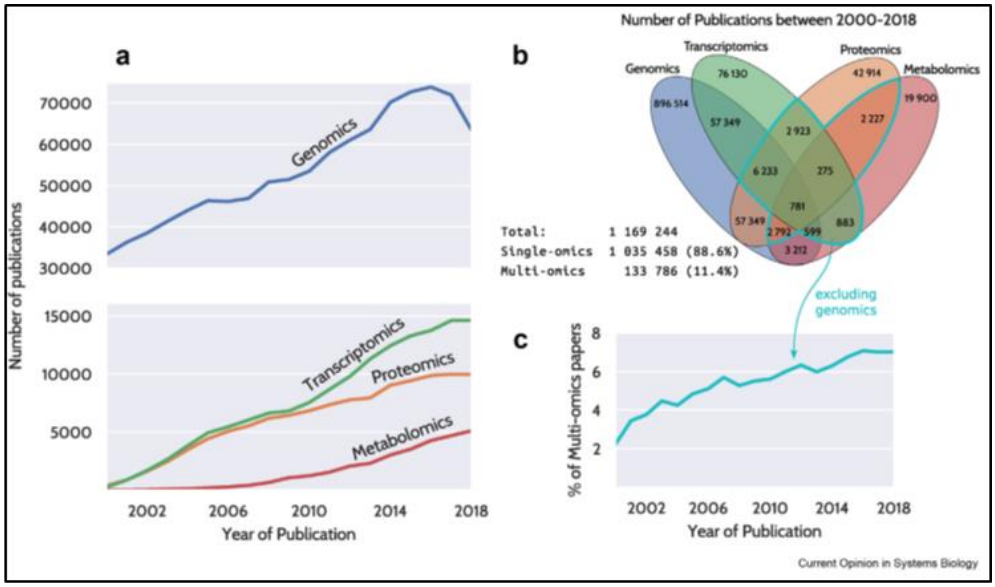

Figure 1 : D'après Noor *et al.*, 2019, Evolution du nombre de publications de données omiques de 2000 à 2018.

En 20 ans, le nombre de publications relatives à l'étude de données omiques a fortement augmenté (Figure 1), particulièrement les données de génomique par exemple, dû à l'essor des nouvelles technologies de séquençage haut débit, les NGS.

Il existe des bases de données qui rassemblent des données omiques d'un même type. Nous pouvons citer la base de données GEO (Genome Expression Omnibus, NCBI), destinée à conserver et à rendre accessibles des données d'expression de génome. Cette base de données contient des données de transcriptomique issues de diverses technologies, telles que des puces à ADN ou de séquençage. La base de données ProteomicsDB (https://www.proteomicsdb.org/) rassemble des données de protéomique chez l'humain et chez l'arabette.

Récemment, des *compendia* de jeux de données ont été créés (Conesa *et al.* 2019) dans le but de réaliser les étapes indispensables à l'analyse, telles que les contrôles de qualité des données, la gestion des données manquantes ou encore la normalisation des données issues de différentes technologies indispensables à leur comparaison. Le constat aujourd'hui fait ressortir que les données issues de chaque technologie omique sont accessibles sur des plateformes différentes et indépendantes, et s'il existe des standards pour le dépôt de chaque donnée omique, il n'en existe pas encore pour des approches multi-omiques.

Intégration de données, une vision holistique :

Les approches multi-omiques sont développées pour répondre à des questions biologiques en prenant en compte, non pas un seul type de données omiques, mais en associant des résultats de génomique, de transcriptomique, de protéomique, de phénomique, etc. Communément les données issues de technologies omiques différentes sont analysées indépendamment à l'aide de méthodes statistiques univariées telles que l'ANOVA, le test de Student, ou encore des analyses de corrélation linéaire de Pearson. Ces analyses unaires des omiques séparément ne permettent pas de décrire la biologie d'un système. Ces relations ne sont pas toujours linéaires, il y a une nécessité de considérer toutes les données dans leur globalité, d'avoir une vision holistique des données omiques. En d'autres termes, il est nécessaire de considérer qu'il existe une relation entre ces données qui reflète le fait que tous les mécanismes qui les sous-tendent agissent de concert pour moduler, influencer, adapter les voies de signalisation biochimiques et la biologie des systèmes.

Problématique :

La problématique que j'ai choisi d'aborder dans ce projet est la suivante : comment comparer et intégrer des données omiques de nature différente, avec quels outils statistiques et informatiques ?

Il existe plusieurs suites logicielles qui permettent de répondre à ces questions. J'ai choisi de m'intéresser à la suite mixOmics. Ce choix a été motivé par le fait que mixOmics a été en partie développée par des équipes de recherches françaises, et notamment par mon tuteur scientifique pour ce projet bibliographique. Elle est gratuite et elle est une référence dans l'analyse multi-omique.

Le rapport est organisé autour de trois parties portant sur le choix d'une méthode statistique pour répondre à une question biologique donnée, la description de la suite logicielle mixOmics, et enfin une conclusion qui ouvre sur les potentielles pistes de recherches pour l'amélioration de mixOmics.

II. Méthodes d'analyses :

1. Préparation des données :

La préparation des données est une étape particulièrement importante à réaliser avant d'en effectuer leur analyse, c'est une étape à ne pas négliger. D'après Duruflé *et al.* 2019, 80% de l'analyse d'un jeu de données consiste en sa préparation. Les données doivent être structurées sous la forme

de matrices avec les échantillons biologiques en ligne et les variables (expression de gènes, accumulation de protéines, hauteur de plantes, etc.) en colonnes. Ainsi, toutes les données omiques en entrée de mixOmics seront organisées de cette manière.

2. À chaque question sa méthode :

De la même façon que l'on ne plante pas une vis avec un marteau, on n'utilise pas n'importe quel outil statistique pour répondre à une question. Ainsi, il est important d'utiliser la bonne méthode statistique adaptée à une question biologique et aux données auxquelles on s'intéresse. Dans cet esprit, l'article Duruflé *et al.* 2019 adresse à chaque question posée à propos de données omiques, ou de biologie en général, une méthode statistique appropriée.

Les méthodes statistiques évoquées dans cette partie ne seront pas détaillées dans ce projet bibliographique, car elles ne constituent pas l'objet de ce travail. Toutefois, il est nécessaire de décrire dans quel cas il est recommandé de les utiliser dans un contexte d'analyse multi-omique. À l'image d'une boîte à outils pour le bio-informaticien, le statisticien recommande l'utilisation de tel ou tel outil pour l'analyse de telles ou telles données omiques.

Ainsi, à chaque question biologique, le statisticien recommande une méthode statistique appropriée :

Étudier une donnée quantitative unique, exemple : quelle est l'expression d'un gène ?

Les statistiques élémentaires telles que la moyenne, la médiane, l'écart-type ou la variance pour évaluer la dispersion sont à utiliser. Il est possible d'illustrer ces statistiques par une représentation graphique, une boîte à moustache par exemple.

Déterminer l'influence d'une variable catégorielle sur une variable quantitative, exemple : la croissance de la plante est-elle la même dans deux conditions différentes ?

Les tests de significativité du type test de Student ou test de rang de Wilcoxon, si l'on compare deux groupes (conditions), et l'ANOVA et le test de Kruskall-Wallis, pour des comparaisons de plus de deux groupes, sont appropriés à cette question. Parmi ces outils, il est nécessaire d'utiliser celui adapté à la structure des données à analyser, un test paramétrique tel que le test de Student nécessite de vérifier que les données suivent une loi Normale et on définit communément qu'il faut des échantillons de minimum 30 valeurs pour pouvoir l'appliquer. En revanche, un test non paramétrique tel que le test de rang de Wilcoxon sera mieux adapté si les échantillons sont de taille inférieure et qu'il n'est pas possible de vérifier leur normalité.

Evaluer les relations entre deux variables quantitatives, exemple : y a-t-il une corrélation entre l'accumulation d'une protéine et le niveau expression du gène qui la code ?

Le coefficient de corrélation de Pearson pour des relations linéaires ou le coefficient de Spearman pour des relations monotones sont adaptés à cette question, associés à une représentation graphique de type matrice de corrélations.

Étudier un jeu de données unique, c'est à dire uniquement le transcriptome par exemple, et identifier une tendance ou un profil, un biais d'expérimentation ou encore vérifier que les données se regroupent (« clusterisent ») en fonction des conditions biologiques desquelles elles sont issues, exemple : peut-on observer l'effet des différentes conditions chez les différents génotypes de l'expérience ?

Les analyses factorielles non supervisées, telles que l'ACP (Analyse en Composante Principale), permettent d'accéder à ces informations à propos d'un jeu de données. Au sujet de cette méthode, les auteurs précisent qu'il est généralement recommandé de centrer et réduire les données lorsque l'on analyse des données omiques, afin de fixer la moyenne à 0 et d'unifier la variance ; auquel cas,

pour l'expression du transcriptome par exemple, un gène dont le niveau d'expression est très variable risque de masquer l'effet d'un autre gène dont le niveau d'expression est moins variable, et ce sans justification biologique, biaisant l'interprétation de l'ACP.

Classer des échantillons sur la base d'un jeu de données dans des catégories connues a priori, exemple : peut-on classer les différents génotypes en fonction de leur transcriptome ?

Les méthodes de classification supervisée telles que la PLS-DA (Régression des Moindres Carrées Partiels en version Discriminante) sont utilisées pour déterminer si les données permettent de classer correctement les échantillons en fonction de la catégorie à laquelle ils appartiennent d'une part, mais elles permettent également de prédire la catégorie d'un nouvel échantillon.

Faire ressortir les informations portées par deux jeux de données, où les variables sont mesurées sur les mêmes échantillons, exemple : quelles sont les principales relations entre le protéome et le transcriptome sur les mêmes échantillons ?

Les méthodes dérivées de la PLS permettent de savoir si des informations communes peuvent être extraites des deux ensembles de données ou de mettre en exergue les relations qui peuvent exister entre deux jeux de données.

Faire ressortir les informations portées par plusieurs jeux de données (plus de deux), où les variables sont mesurées sur les mêmes échantillons, exemple : quelles sont les principales relations entre le protéome, le transcriptome et le métabolome sur les mêmes échantillons ?

Les méthodes dérivées des méthodes PLS multi-block ont récemment été développées pour résoudre ce type de question.

Faire ressortir les informations portées par plusieurs jeux de données (plus de deux), où les variables sont mesurées sur les mêmes échantillons dont les catégories sont connues a priori*, exemple : peut-on déterminer une signature multi-omique pour classer des échantillons en fonction de leur génotype ?*

La méthode développée pour répondre à cette question est appelée la PLS-DA multi-block.

3. Les variantes dans les analyses :

Analyses supervisées, non supervisées :

Les groupes d'individus peuvent être connus à l'avance et appartenir à différentes catégories, dans notre contexte. Il peut s'agir de plantes inoculées (malades) et de plantes témoins (saines). Dans d'autres cas, les groupes ne sont pas connus à l'avance, ils ne sont pas définis par des catégories *a priori*. Ainsi, deux types d'analyses sont possibles, les analyses dites supervisées, c'est-à-dire lorsque l'on connait les catégories à l'avance, et non supervisées lorsque l'on ne les connait pas.

La méthode Sparse :

Lorsque l'on considère des données omiques, on considère un ensemble, souvent important, de variables d'un même type sur quelques échantillons : par exemple, l'expression des 25 000 gènes d'un génome chez deux génotypes d'une espèce de plante cultivée dans deux conditions différentes. Or, seulement une partie de ces 25 000 gènes diffère entre les deux génotypes dans les différentes conditions. En d'autres termes, seulement un nombre restreint de variables permettent de différencier les catégories d'échantillons. Ainsi, une proportion de gènes ne varie que très peu, de manière non significative entre les différentes catégories. La méthode Sparse permet de filtrer ces variables peu différentes.

À titre d'exemple, Duruflé *et al.* 2019 décrivent que pour l'ACP, la méthode Sparse permet de ne sélectionner que les variables qui contribuent le plus à la définition des composantes principales. La méthode Sparse est basée sur l'estimateur des moindres carrés pénalisé (LASSO, Least Absolute Shrinkage and Selection Operator penalization).

Les méthodes Sparse sont particulièrement utiles dans un contexte d'analyse de données omiques pour réduire le nombre de variables d'intérêts. On peut considérer que les variables éliminées par les méthodes Sparse sont les variables qui ne sont pas affectées par les différentes conditions ou peu différentes entre les catégories.

Toutes ces méthodes statistiques sont implémentées dans la suite logicielle mixOmics. Bien que cette suite soit destinée à l'intégration et l'analyse de données omiques, les auteurs/concepteurs précisent, à titre de remarque, que ces méthodes d'analyse sont aussi valables pour l'analyse de données autres qu'omiques, non biologiques. Dans la partie suivante, nous développerons le fonctionnement de mixOmics et notamment ses deux principaux outils, DIABLO et MINT.

III. La suite logicielle mixOmics, des outils d'analyse multi-omiques :

Le but d'une analyse multi-omique est d'identifier des sous-ensembles de bio-marqueurs qui expliquent la majorité de la variation entre différentes classes d'échantillons. En biologie des systèmes, il s'agit d'identifier les gènes, les transcrits, les protéines, les métabolites, les caractères phénotypiques et autres données omiques qui permettent de décrire un groupe d'individus par rapport à un autre. La biologie des systèmes, en tant que vision holistique, consiste en l'intégration des données omiques afin de mieux comprendre la réponse physiologique d'un organisme dans sa globalité. Ainsi, l'intégration des données omiques est réalisée pour réduire de grands jeux de données omiques difficiles à manipuler « manuellement » et permet de :

- Avoir une vue d'ensemble du rôle de chaque omiques dans la biologie du système,
- Avoir une meilleure compréhension de la relation qu'il existe entre les différentes omiques,
- Identifier une signature moléculaire d'une réponse d'un organisme à une condition,
- Déduire un modèle prédictif, c'est-à-dire prédire à partir d'un nombre réduit de données le classement d'un individu dans une catégorie (par exemple, un organisme malade ou sain).

Rares sont les méthodes statistiques qui permettent l'intégration de données diverses telles que des données omiques différentes ou issues d'expérimentations indépendantes dans un contexte supervisé. Le package mixOmics fournit une solution « clef en main » pour réaliser ce type d'analyse.

Description du package mixOmics : Rohart et al 2017 mixOmics: An R package for `omics feature selection and multiple data integration Methodology.

MixOmics est un package développé dans le langage de programmation statistique R. Il fait partie des outils d'analyse de données biologiques disponible via Bioconductor. Il est donc disponible gratuitement et le code source est accessible (« open-source »). Il est maintenu des équipes de recherche française et australienne. Les principaux contributeurs actuels sont Sébastien Déjean de l'Institut de Mathématiques de Toulouse pour la France et Kim-Anh Le Cao de l'Université de Melbourne pour l'Australie.

L'analyse réalisée à l'aide de mixOmics est décomposée en trois étapes, le formatage des données d'entrées, l'analyse statistique *stricto sensu* et la représentation graphique des résultats.

Data input :

Les données sont supposées être normalisées avec la méthode de rigueur adaptée au type de données omiques dont il s'agit, ainsi qu'à la technologie dont elles sont issues. Les données peuvent être des données continues ou bien des données de type « comptage » qui sont assimilables à des données continues après normalisation.

Chaque jeu de données est formaté en une matrice X de taille N observations (en lignes) qui correspondent aux échantillons sur lesquels les données omiques ont été produites, et P

« prédicteurs » (en colonnes) qui représentent les variables mesurées telles que l'expression des gènes du transcriptome, l'accumulation de protéines du protéome, l'accumulation de métabolites du métabolome, etc. Y est une deuxième matrice qui représente la catégorie de chaque observation dans un contexte supervisé, par exemple les catégories « malade » ou « sain ». Y est donc de taille N observations et K catégories (Figure 2).

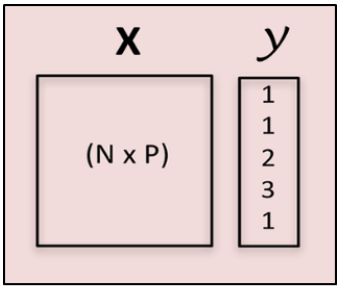

Figure 2 : Représentation schématique des matrices X et Y, illustration du formatage des données pour les analyses de données omiques.

Il est recommandé par les auteurs de réduire la quantité de variables à 10 000 (? 10K) pour diminuer les temps de calcul, en supprimant les variables dont la variance est proche de 0 ou par « Median Absolute Deviation » pour les données de RNAseq par la suppression des gènes très faiblement exprimés.

Analyse statistique des données :

Il est possible de distinguer trois situations définies par les données à analyser. Un premier cas de figure concerne une situation où un seul type de données omiques obtenues lors d'une même expérimentation est disponible. Ce contexte d'étude, dit « Single'Omics » n'est pas à proprement parlé une intégration de données multi-omiques, mais il constitue la brique élémentaire des deux autres situations qui sont l'intégration de données d'un seul type de données omiques obtenues lors de plusieurs expérimentations différentes d'une part (N-intégration), et l'intégration de données omiques de différents types (transcriptomique, protéomique, métabolomique, etc.) sur les mêmes N échantillons d'autre part (P-intégration), (Figure 3).

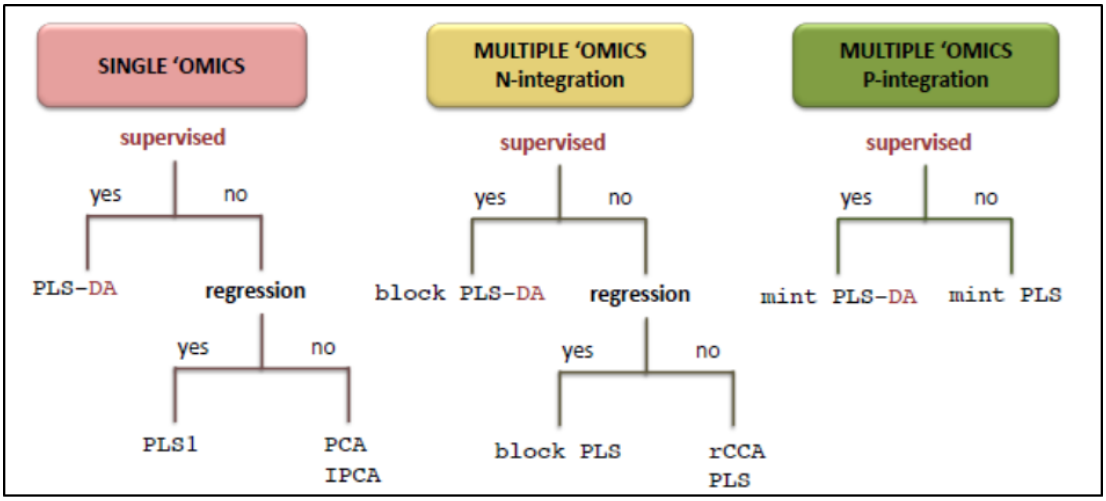

Figure 3 : D'après Lê Cao, 2019, Schéma des différents types d'intégration de données omiques. Les arbres schématisent les différentes possibilités d'analyses (non exhaustives) disponibles avec la suite logicielle mixOmics dans les différents contextes d'intégration de données.

Il est possible de voir l'analyse en Single'Omics comme la brique élémentaire de l'intégration multi-omique. Dans ce cas de figure, le point de départ est une matrice de données à N lignes et P colonnes. Le but est de réduire la dimension de cette matrice par une méthode de factorisation de matrice qui consiste en une agrégation de l'information en une ou plusieurs composantes. MixOmics propose trois méthodes qui sont l'ACP, la PLS-DA et la sPLS-DA (pour Sparse PLS-DA).

L'ACP :

Elle a pour but de définir des composantes principales en maximisant la variance entre les échantillons. Ces composantes principales sont une combinaison linéaire des variables dans un individu. Plusieurs composantes sont calculées, par ordre décroissant de la part de la variance totale entre les échantillons qu'elle explique (Figure 4, PCA). Il est alors possible de représenter dans un espace les échantillons ou les variables en fonction de ces composantes principales et faire ressortir (ou non) des groupes d'échantillons ou de variables.

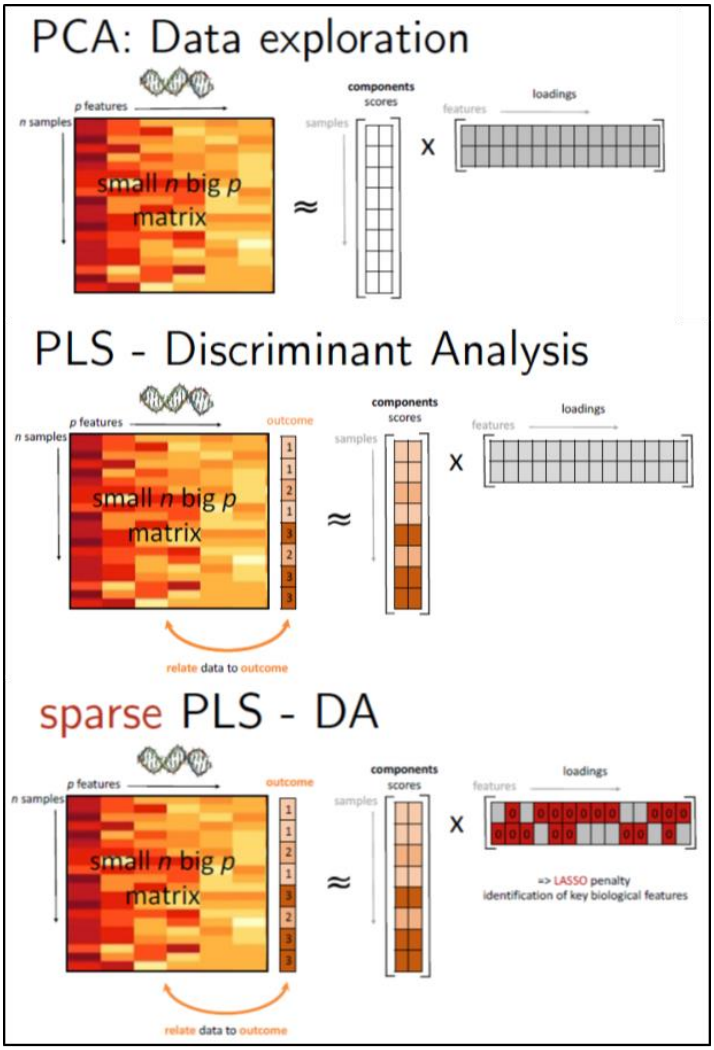

Figure 4 : D'après Lê Cao, 2019, Représentations schématiques des méthodes statistiques ACP (PCA), PLS-DA et sparse PLS-DA. Pour les trois méthodes, une matrice représente les données omiques préparées, une seconde matrice (components) représente les composantes calculées, une troisième matrice (« loadings ») représente une matrice de poids pour chaque variable (non discutée dans ce rapport). Pour les deux analyses en version discriminante, une matrice (« outcome ») représente la matrice catégorielle.

La PLS-DA :

DA est destinée à l'analyse discriminante, c'est-à-dire qu'une distinction de catégories est faite dans une matrice Y. Ainsi, la PLS-DA est une régression linéaire qui a pour but de définir des composantes comme pour l'ACP, mais en maximisant la Covariance entre la matrice de données X et la matrice catégorielle Y (Figure 4, PLS-Discriminant Analysis).

La sPLS-DA :

Il s'agit de la variante Sparse de la PLS-DA (Figure 4, sparse PLS-DA). Dans cette analyse, les composantes sont calculées après qu'une étape de pénalisation des données peu variables ait été effectuée. Pour simplifier le discours, les variables qui ne varient pas ou très peu dans l'analyse ne seront pas prises en compte dans le calcul des composantes.

Ainsi, pour la PLS-DA et sa version Sparse, il est également possible de représenter les échantillons ou les variables dans un espace. La représentation graphique des composantes permet une interprétation des résultats.

Pour le biologiste, les composantes correspondent aux variables, dont la variation permet de discriminer les échantillons. Ainsi, dans le cas de l'analyse de données transcriptomiques, il s'agit des gènes différentiellement exprimés qui sous-tendent les différentes catégories des échantillons. Par extrapolation, on parlera d'un ensemble de biomarqueurs qui expliquent une part (calculée) de la variabilité observée entre les différentes catégories d'échantillons.

Dans la partie suivante, nous illustrerons par un exemple les analyses statistiques qui viennent d'être décrites ainsi que leur interprétation. Cela constituera la description de la brique élémentaire des intégrations multi-omiques qui en découlent.

1. Single'Omics :

Afin d'illustrer l'approche Single'Omics, nous allons nous intéresser au travail effectué par Duruflé *et al.* 2019, qui ont réalisé une étude multi-omiques de la croissance de différents écotypes d'*Arabidopsis thaliana* dans deux conditions de températures différentes, 15°C et 22°C. Pour cette partie, nous nous intéresserons aux résultats du transcriptome des cellules pariétales des feuilles de la rosette qu'ils ont obtenus. Ainsi, l'expression de 364 gènes a été mesurée chez 5 écotypes à 2 températures différentes.

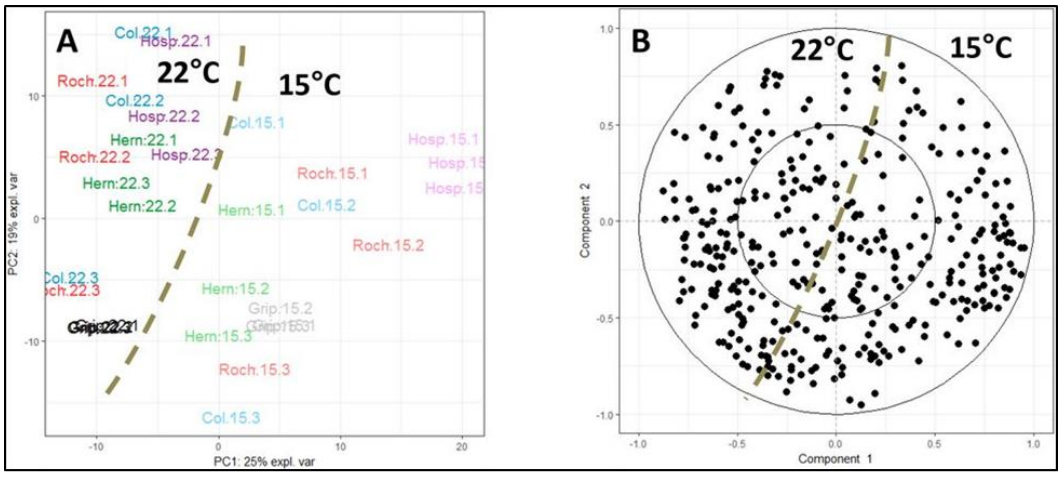

Figure 5 : D'après Duruflé *et al.*, 2019, Représentations graphiques des deux premières composantes de l'ACP (A, B) sur les données de transcriptomiques dans les cellules pariétales des feuilles de la rosette chez *Arabidopsis thaliana* cultivées dans deux conditions de température, 22°C et 15°C. A) Représentation des échantillons, 22°C (couleurs mates) and 15°C (couleurs pastel) et B) la representation des variables.

ACP :

Une ACP est réalisée sur ces données par Duruflé *et al.* 2019. La première composante calculée explique 25% de la variance totale des échantillons, la seconde en explique 19%. L'interprétation graphique de l'ACP (Figure 5) tend à montrer que les échantillons sont groupés en fonction de la température de culture sur la première composante. Ainsi, les auteurs en concluent que les conditions de culture sont probablement responsables de la plus grande part de la variance dans les données de transcriptome. Pour les auteurs, la représentation des variables n'a ici que peu d'intérêt, puisqu'elle ne rend compte que des gènes qui sont sur- ou sous-exprimés en fonction des conditions de culture.

L'ACP est, dans ce contexte, un contrôle de l'expérimentation. En effet, elle permet de se rendre compte de ce que l'on peut observer à travers les données. Ici, la conclusion est que le critère de la température de culture des plantes est responsable de la plus grande part de la variance observée à travers les données de transcriptomique. Cependant, d'après les auteurs, elle ne permet pas de sélectionner des gènes candidats.

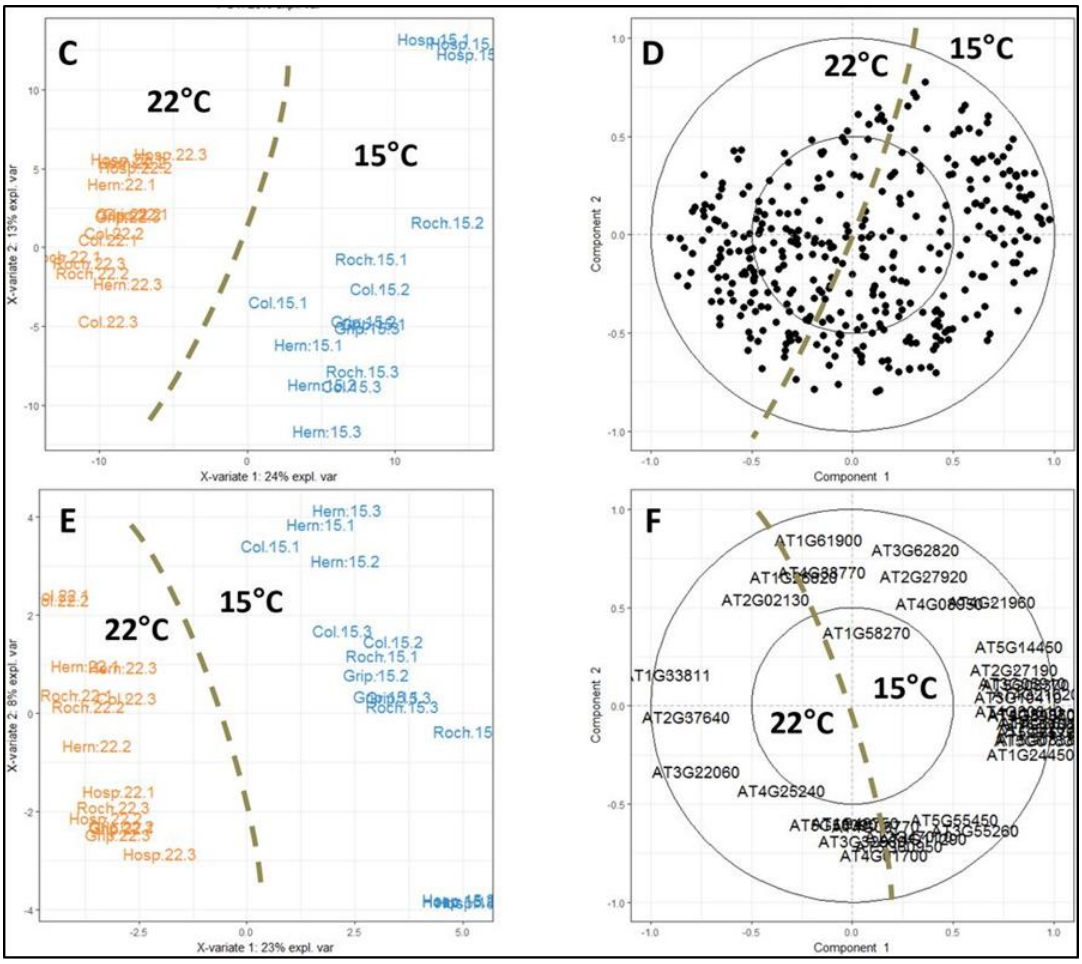

Figure 6 : D'après Duruflé *et al.*, 2019, Représentations graphiques des deux premières composantes de la PLS-DA (C, D) et de la sparse PLS-DA (E, F) sur les données de transcriptomiques dans les cellules pariétales des feuilles de la rosette chez *Arabidopsis thaliana* cultivées dans deux conditions de température, 22°C et 15°C. C) et E) Représentation des échantillons, 22°C (orange) et 15°C (bleu) et D) et F) la representation des variables.

PLS-DA et Sparse PLS-DA:

Les mêmes données sont analysées par les méthodes de PLS-DA et sPLS-DA. Les catégories discriminantes utilisées sont les conditions de températures de culture des plantes. Ainsi, une matrice

X représente les expressions des 364 gènes mesurées, et une matrice Y représente les conditions de culture. Le but de ces deux analyses est de sélectionner les gènes candidats qui sous-tendent la différence entre les deux conditions de culture. Les auteurs précisent que l'utilisation de la version Sparse ou non adresse deux questions différentes. Si le but de l'analyse est d'identifier des gènes candidats pour une analyse fonctionnelle, alors la version Sparse est à privilégier, puisqu'elle va permettre de réduire le nombre de gènes (environ une dizaine).

Afin d'être interprétés, les résultats sont représentés de manière graphique (Figure 6). Pour les deux analyses, les auteurs concluent que les échantillons sont séparés en deux groupes en fonction des températures de cultures. Cependant, il ne semble pas y avoir d'effet important dû aux différents écotypes (Figure 6, C et D).

D'autre part, les auteurs attirent notre attention sur la représentation graphique des gènes pour l'analyse sPLS-DA (Figure 6, E et F), qui permet de se rendre compte que seul un petit nombre de gènes semblent permettre de discriminer les deux conditions de températures.

Ainsi, un nombre réduit de gènes ou biomarqueurs a pu être sélectionné, impliqué dans la différence observée chez ces plantes dans les différentes conditions de températures.

Vers des analyses multi-omiqes :

C'est sur cette base que les analyses intégratives de données omiques sont réalisées à l'aide de la suite logicielle mixOmics. L'intégration de différents types de données omiques est réalisée à l'aide de l'outil DIABLO et peut être vue comme une intégration horizontale des données, et l'intégration de différentes expérimentations d'un même type de données omiques est réalisée avec l'outil MINT et sera quant à elle vue comme une intégration verticale des données (Figure 7).

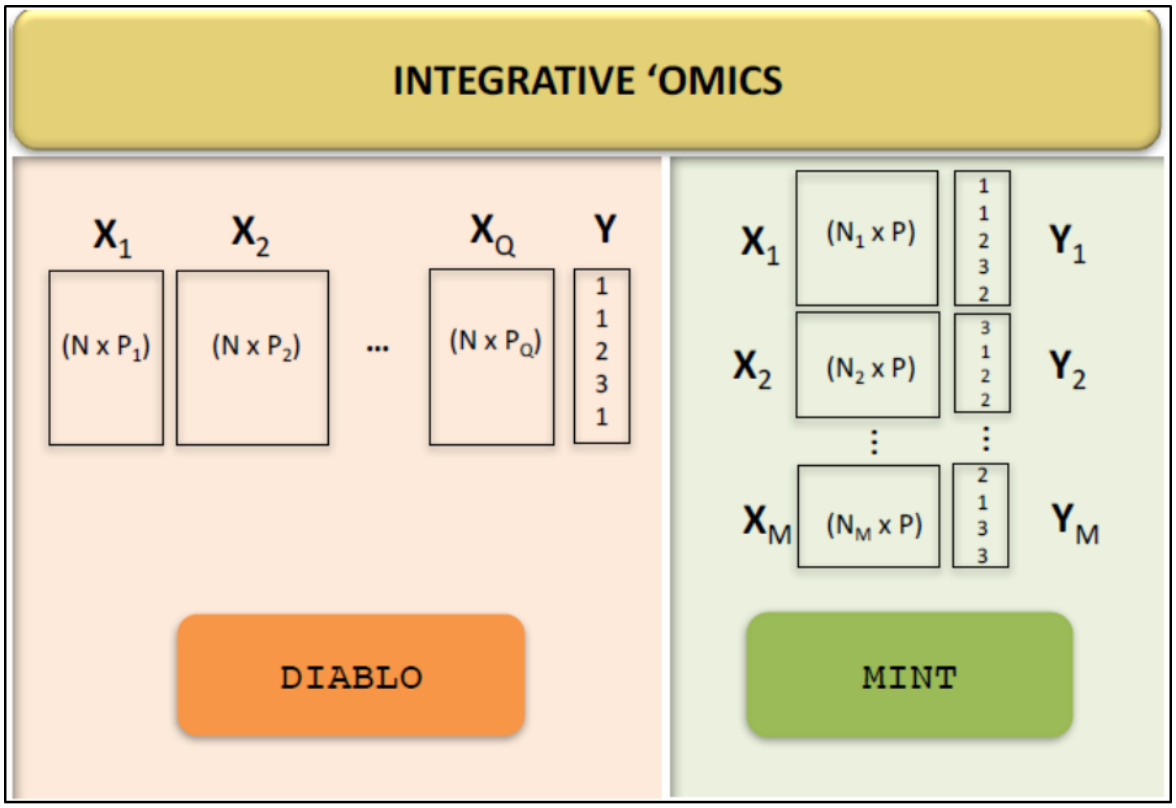

Figure 7 : D'après Lê Cao, 2019, Représentation schématique des outils d'intégration DIABLO et MINT. Illustration de l'intégration horizontale, P-integration réalisée avec DIABLO, et l'intégration verticale, N-integration réalisée avec MINT.

D'un point de vue de l'analyse statistique, alors que dans le cadre d'une analyse en Single'Omics on utilisera les méthodes PLS-DA et sPLS-DA, en analyse multi-omiques, la méthode utilisée est la block-PLS-DA. Il s'agit d'une méthode dérivée de la PLS-DA qui va maximiser la somme des covariances entre les matrices des types de données omiques et de la matrice discriminante deux à deux.

2. Analyse multi-omiques sur les mêmes échantillons : Singh et al 2019 DIABLO :

Nous allons décrire dans cette partie l'intégration de données omiques de différents types de mêmes échantillons biologiques, dite P-integration. Par exemple, l'intégration de données de transcriptomique, de protéomique, de métabolomique obtenues à partir des mêmes individus.

Dans le cadre de ce projet, nous ne nous intéresserons qu'à l'intégration en version discriminante. En biologie, il s'agit du cas de figure le plus fréquent, l'expérimentateur connait les différentes catégories, les différents échantillons et les différentes conditions dans lesquelles ont été produits les échantillons.

Duruflé *et al.* 2019 illustrent la P-integration à travers l'analyse des données de transcriptomique que nous avons décrites pour l'analyse en Single'Omics, auxquelles sont associées des données phénotypiques, de protéomique, et de métabolomique, et la matrice catégorielle est toujours la matrice des températures de culture. Chaque matrice de données est désignée comme un block dans l'analyse. L'analyse est réalisée sur ces données avec l'outil DIABLO du package mixOmics. Ainsi DIABLO va identifier des composantes qui maximisent la somme des covariances entre les blocks deux à deux.

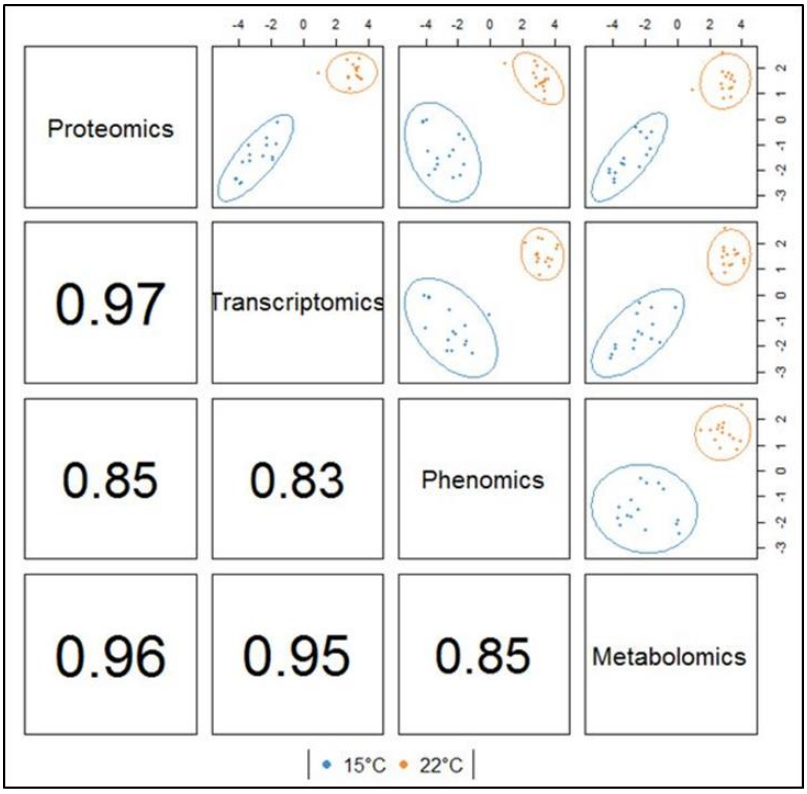

Figure 8 : D'après Duruflé *et al.*, 2019, Représentation graphique de la matrice de corrélation pour les deux premières composantes sur les données de transcriptomique, de protéomique, de métabolomique, phénotypiques et la matrice catégorielle chez *Arabidopsis thaliana* cultivées dans deux conditions de température, 22°C (orange) et 15°C (bleu).

L'interprétation de l'analyse se fait à l'aide de plusieurs représentations graphiques. Une première représentation peut être une matrice de corrélation pour les deux premières composantes entre les différents blocks (Figure 8).

Les coefficients de corrélation entre les différents blocks sont élevés, mais sont dus aux deux groupes de conditions de températures. Il apparaît que la corrélation entre les données de transcriptomique et de protéomique est plus élevée que les autres, ce qui suggère un lien plus important entre ces deux types de données.

Ensuite, les échantillons peuvent être représentés en fonction des deux principales composantes par type de données omiques, de la même façon que nous l'avons vu pour l'analyse en Single'Omics (Figure 9).

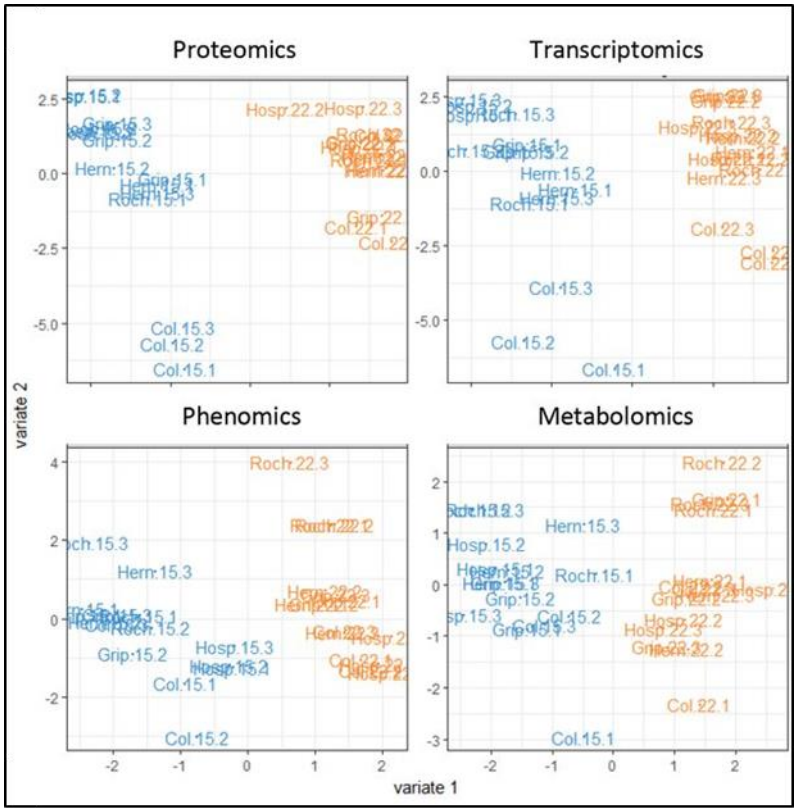

Figure 9 : D'après Duruflé *et al.*, 2019, Représentation graphique de l'analyse DIABLO des échantillons en fonction des deux premières composantes sur les données de transcriptomique, de protéomique, de métabolomique, phénotypiques et la matrice catégorielle chez *Arabidopsis thaliana* cultivées dans deux conditions de température, 22°C (orange) et 15°C (bleu).

L'interprétation de cette représentation suggère que les deux conditions de température sont mieux discriminées par les données de transcriptomiques et de protéomiques.

Les variables peuvent également être représentées graphiquement. Duruflé *et al.* 2019 ont choisi de représenter les résultats dans la version Sparse pour limiter le nombre de bio-marqueurs sur le graphe et ainsi faciliter l'interprétation des résultats (Figure 10).

De cette manière sont représentés les bio-marqueurs qui sont les plus impliqués dans la discrimination des échantillons en fonction de la température de culture.

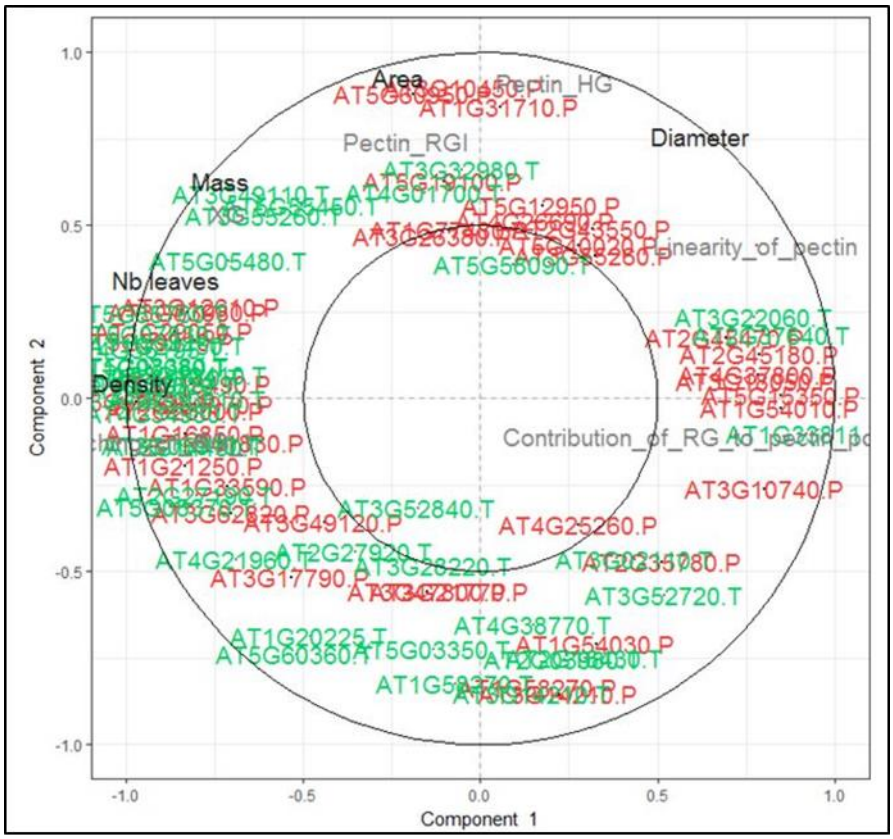

Figure 10 : D'après Duruflé *et al.*, 2019, Représentation graphique de l'analyse DIABLO en version sparse des variables en fonction des deux premières composantes sur les données de transcriptomique, de protéomique, de métabolomique, phénotypiques et la matrice catégorielle chez *Arabidopsis thaliana* cultivée dans deux conditions de température, 22°C et 15°C. En rouge : les protéines, en vert : les gènes, en gris : les métabolites, en noir : les phénotypes.

D'autres types de représentations sont proposés par l'outil DIABLO, telles qu'un réseau entre les différents bio-marqueurs ou encore une représentation des corrélations entre ces bio-marqueurs qui permettent de formuler des hypothèses quant à l'interaction entre ces derniers, dans le but de décrire la biologie du système dans les conditions étudiées. Toutefois, nous ne détaillerons pas ces représentations dans ce rapport, puisqu'il ne s'agit pas de la description des méthodes d'intégrations des données.

L'analyse intégrative multi-omiques en P-intégration permet d'accéder à un petit nombre de bio-marqueurs qui sont les plus impliqués dans la différenciation des échantillons qui ont été cultivés dans les deux conditions de température. Les différentes données ont ainsi été réduites à un nombre réduit de candidats. À partir de ces candidats, il est alors possible de formuler de nouvelles hypothèses.

3. Intégration de données omiques de mêmes types issues de différentes expérimentations indépendantes : Rohart *et al.* 2017 MINT

En complément de l'outil DIABLO, le package mixOmics intègre un outil d'intégration de données omiques d'un même type, mais sur des échantillons provenant d'expérimentations

différentes, l'outil MINT. On parle alors de N-intégration. Par exemple, il peut s'agir de l'intégration de données de transcriptomique obtenues sur plusieurs espèces apparentées pour identifier les gènes communs impliqué dans un caractère, tel que nous l'avons décrit en introduction sur les analyses de la réponse à la sécheresse chez les céréales.

Si la P-integration peut être vue comme l'intégration horizontale des données omiques, la N-integration, quant à elle, est une intégration verticale. Toutefois, les méthodes statistiques utilisées sont les mêmes. Il s'agit de maximiser la somme des covariances entre les différents blocks omiques. L'analyse a pour but de calculer des composantes et d'identifier des bio-marqueurs qui permettent de discriminer les différentes catégories.

Ainsi, les représentations graphiques des résultats sont assez semblables à celles obtenues avec DIABLO, avec les représentations des échantillons ou des variables en fonction des catégories pour chaque block.

L'intégration de données grâce à MINT permet d'identifier des bio-marqueurs communs aux différentes expérimentations intégrées, qui permettent de discriminer les différentes catégories. Sur la base de ces bio-marqueurs, il est possible de générer de nouvelles hypothèses quant à la question biologique initiale.

IV. Conclusion

Nous avons pu voir que l'analyse intégrative de données omiques permet d'accéder à une vision globale, holistique de la biologie du système dans des conditions données. Ce type d'analyse permet de prendre en compte une grande quantité d'information et d'en extraire la partie la plus pertinente pour générer des hypothèses quant à la question biologique posée.

Dans ce rapport bibliographique, l'effort a été mis sur les différentes méthodes d'analyses utilisées dans mixOmics pour réaliser des analyses multi-omiques, plus que pour décrire les analyses en tant que telles. Ce choix s'explique par le fait que les justifications biologiques d'une telle approche ne sont pas en question, mais plutôt la mise en œuvre.

MixOmics est un outil puissant d'intégration de données qui fournit une solution « clef en main » et accessible aux biologistes pour analyser de grands jeux de données hétérogènes. Il adresse à des questions biologiques des outils statistiques adéquats. Il permet de représenter graphiquement les résultats des analyses afin de faciliter l'interprétation des données et fournit un nombre manipulable de bio-marqueurs pertinents.

Nous avons vu à travers de l'outil mixOmics l'intégration de données par des méthodes linéaires. Or, il existe d'autres méthodes en cours d'investigation. En particulier les méthodes dites à Noyaux, qui feront l'objet du stage que nous envisageons.

V. Bibliographie :

Conesa, Ana, et Stephan Beck. « Making Multi-Omics Data Accessible to Researchers ». *Scientific Data* 6, n° 1 (décembre 2019): 251. https://doi.org/10.1038/s41597-019-0258-4.

Crick, Francis. « Central Dogma of Molecular Biology ». *Nature* 227, n° 5258 (août 1970): 561-63. https://doi.org/10.1038/227561a0.

Duruflé, Harold, Vincent Hervé, Philippe Ranocha, Thierry Balliau, Michel Zivy, Josiane Chourré, Hélène San Clemente, *et al.* « Cell Wall Modifications of Two Arabidopsis Thaliana Ecotypes, Col and Sha, in Response to Sub-Optimal Growth Conditions: An Integrative Study ». *Plant Science* 263 (1 octobre 2017): 183-93. https://doi.org/10.1016/j.plantsci.2017.07.015.

Lê Cao, Kim-Anh, Multi-omics statistical integration with mixOmics, 04 SEP 2019, EMBL-ABR. « EMBL-ABR Webinars ». https://www.embl-abr.org.au/webinars/.

Noor, Elad, Sarah Cherkaoui, et Uwe Sauer. « Biological Insights through Omics Data Integration ». *Current Opinion in Systems Biology*, Gene regulation, 15 (1 juin 2019): 39-47. https://doi.org/10.1016/j.coisb.2019.03.007.

Rohart, Florian, Aida Eslami, Nicholas Matigian, Stéphanie Bougeard, et Kim-Anh Lê Cao. « MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms ». *BMC Bioinformatics* 18, n° 1 (27 février 2017): 128. https://doi.org/10.1186/s12859-017-1553-8.

Rohart, Florian, Benoît Gautier, Amrit Singh, et Kim-Anh Lê Cao. « MixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration ». *PLOS Computational Biology* 13, n° 11 (3 novembre 2017): e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Rohart, Florian, mixOmics: An R package for 'omics feature selection and multiple data integration, R Consortium (14 juillet 2018): https://youtu.be/J1JrIJPDfig.

Singh, Amrit, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, et Kim-Anh Lê Cao. « DIABLO: An Integrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays ». *Bioinformatics* 35, n° 17 (1 septembre 2019): 3055-62. https://doi.org/10.1093/bioinformatics/bty1054.